HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2011 April 01.

Published in final edited form as:

Nat Biotechnol. 2010 October ; 28(10): 1015–1017. doi:10.1038/nbt1010-1015.

ProHits: an integrated software platform for mass spectrometry-based interaction proteomics

Guomin Liu[1], **Jianping Zhang**[1], **Brett Larsen**[1], **Chris Stark**[1], **Ashton Breitkreutz**[1], **Zhen-Yuan Lin**[1], **Bobby-Joe Breitkreutz**[1], **Yongmei Ding**[1], **Karen Colwill**[1], **Adrian Pasculescu**[1], **Tony Pawson**[1,2], **Jeffrey L. Wrana**[1,2], **Alexey I. Nesvizhskii**[3], **Brian Raught**[4], **Mike Tyers**[1,2,5,*], and **Anne-Claude Gingras**[1,2,*]

[1] Centre for Systems Biology, Samuel Lunenfeld Research Institute, 600 University Avenue, Toronto, Ontario, M5G 1X5, Canada

[2] Department of Molecular Genetics, University of Toronto, 1 Kings College Circle, Toronto, Ontario, M5S 1A8 Canada

[3] Departments of Pathology and Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan USA, 48109-0602

[4] Ontario Cancer Institute and McLaughlin Centre for Molecular Medicine, 101 College St, Toronto, ON M5G 1L7 Canada

[5] Wellcome Trust Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Mayfield Road, Edinburgh, EH9 3JR Scotland UK

Affinity purification coupled with mass spectrometric identification (AP-MS) is now a method of choice for charting novel protein-protein interactions, and has been applied to a large number of both small scale and high-throughput studies1. However, general and intuitive computational tools for sample tracking, AP-MS data analysis, and annotation have not kept pace with rapid methodological and instrument improvements.

To address this need, we developed the ProHits LIMS platform. ProHits is a complete open source software solution for MS-based interaction proteomics that manages the entire pipeline from raw MS data files to fully annotated protein-protein interaction datasets. ProHits was designed to provide an intuitive user interface from the biologist's perspective, and can accommodate multiple instruments within a facility, multiple user groups, multiple laboratory locations, and any number of parallel projects. ProHits can manage all project scales, and supports common experimental pipelines, including those utilizing gel-based separation, gel-free analysis, and multi-dimensional protein or peptide separation.

[*]Correspondence should be addressed to M.T. (m.tyers@ed.ac.uk) or A.-C.G. (gingras@lunenfeld.ca).

Author contribution:

GL and JPZ devised and coded all aspects of the platform; CS and BJB implemented protein annotation and provided advice on database architecture; YD wrote the Mascot parser; BL, AB, ZL, KC, AIN, TP, JW and BR provided suggestions on software features; MT conceived and guided the project; ACG, BR and GL wrote the instruction manuals; MT and ACG co-directed project development; ACG wrote the manuscript with input from BR and MT.

ProHits is a client-based HTML program written in PHP that runs a MySQL database on a dedicated server. The complete ProHits software solution consists of two main components: a Data Management module, and an Analyst module (Fig. 1a; see Supplementary Fig. 1 for data structure tables). These modules are supported by an Admin Office module, in which projects, instruments, user permissions and protein databases are managed (Supplementary Fig. 2). A simplified version of the software suite ("ProHits Lite"), consisting only of the Analyst module and Admin Office, is also available for users with pre-existing data management solutions or who receive pre-computed search results from analyses performed in a core MS facility (Supplementary Fig. 3). A step-by-step installation package, installation guide and user manual (see Supplementary Information) are available on the ProHits website (www.prohitsMS.com).

In the Data Management module, raw data from all mass spectrometers in a facility or user group are copied to a single secure storage location in a scheduled manner. Data are organized in an instrument-specific manner, with folder and file organization mirroring the organization on the acquisition computer. ProHits also assigns unique identifiers to each folder and file. Log files and visual indicators of current connection status assist in monitoring the entire system. The Data Management module monitors the use of each instrument for reporting purposes (Supplementary Fig. 4–5). Raw MS files can be automatically converted to appropriate file formats using the open source ProteoWizard converters (http://proteowizard.sourceforge.net/). Converted files may be subjected to manual or automated database searches, followed by statistical analysis of the search results, according to any user-defined schedule; search engine parameters are also recorded to facilitate reporting and compliance with MIAPE guidelines2. Mascot3, X!Tandem4 and the TransProteomics Pipeline (TPP5) are fully integrated with ProHits via linked search engine servers (Supplementary Fig. 6–7).

The Analyst module organizes data by project, bait, experiment and/or sample, for gel-based or gel-free approaches (Fig. 1a; for description of a gel-based project, see Supplementary Fig. 8). To create and analyze a gel-free affinity purification sample, the user specifies the bait gene name and species. ProHits automatically retrieves the amino acid sequence and other annotation from its associated database. Bait annotation may then be modified as necessary, for example to specify the presence of an epitope tag or mutation (Supplementary Fig. 9). A comprehensive annotation page tracks experimental details (Supplementary Fig. 10), including descriptions of the Sample, Affinity Purification protocol, Peptide Preparation methodology, and LC-MS/MS procedures. Controlled vocabulary lists for experimental descriptions can be added via drop-down menus to facilitate compliance with annotation guidelines such as MIAPE6 and MIMIx7, and to facilitate the organization and retrieval of data files. Free text notes for cross-referencing laboratory notebook pages, adding experimental details not captured in other sections, describing deviations from reference protocols and links to gel images or other file types may be added in the Experimental Detail page. Once an experiment is created, multiple samples may be linked to it, for example technical replicates of the same sample, or chromatographic fractions derived from the same preparation. All baits, experiments, samples and protocols are assigned unique identifiers.

Once a sample is created, it is linked to both the relevant raw files and database search results. For multiple samples in HTP projects, automatic sample annotation may be established by using a standardized file naming system (Supplementary Fig. 11), or files may be manually linked. Alternatively, search results obtained outside of ProHits (with the X!Tandem or Mascot search engines) can be manually imported into the Analyst module (Supplementary Fig. 12). The ProHits Lite version enables uploading of external search results for users with an established MS data management system.

In the Analyst module, mass spectrometry data can be explored in an intuitive manner, and results from individual samples, experiments or baits can be viewed and filtered (Supplementary Fig. 13–14). A user interface enables alignment of data from multiple baits or MS analyses using the Comparison tool. Data from individual MS runs, or derived from any user-defined sample group, are selected for visualization in a tabular format, for side-by-side comparisons (Fig. 1b; Supplementary Fig. 15–17). In the Comparison view, control groups and individual baits, experiments or samples are displayed by column. Proteins identified in each MS run or group of runs are displayed by row, and each cell corresponds to a putative protein hit, according to user-specified database search score cutoff. Cells display spectral count number, unique peptides, scores from search engines, and/or protein coverage information; a mouse-over function reveals all associated data for each cell in the table. For each protein displayed in the Comparison view, an associated Peptide link (Fig. 1b) may also be selected to reveal information such as sequence, location, spectral counts, and score, for each associated peptide. Importantly, all search results can be filtered. For example, ProHits allows for the removal of non-specific background proteins from the hit list, as defined by negative controls, search engine score thresholds, or contaminant lists. Links to the external NCBI and BioGRID8 databases are provided for each hit to facilitate data interpretation. Overlap with published interaction data housed in the BioGRID database8 can be displayed to allow immediate identification of new interaction partners. A flexible export function enables visualization in a graphical format with Cytoscape9, in which spectral counts, unique peptides, and search engine scores can be visualized as interaction edge attributes. The Analyst module also includes advanced search functions, bulk export functions for filtered or unfiltered data, and management of experimental protocols and background lists (e.g. Supplementary Fig. 18–20).

Deposition of all mass spectrometry-associated data in public repositories is likely to become mandatory for publication of proteomics experiments2, 7, 10. Open access to raw files is essential for data reanalysis and cross-platform comparison; however, data submission to public repositories can be laborious due to strict formatting requirements. ProHits facilitates extraction of the necessary details in compliance with current standards, and generates Proteomic Standard Initiative (PSI) v2.5 compliant reports11, either in the MITAB format for BioGRID8 or in XML format for submission to IMEx consortium databases12, including IntAct13 (Supplementary Fig. 21). MS raw files associated with a given project can also be easily retrieved and grouped for submission to data repositories such as Tranche14.

ProHits has developed to manage many large-scale in-house projects, including a systematic analysis of kinase and phosphatase interactions in yeast, consisting of 986 affinity

purifications15. Smaller-scale projects from individual laboratories are readily handled in a similar manner. Examples of AP-MS data from both yeast and mammalian projects are provided in a demonstration version of ProHits at www.prohitsMS.com, and in Supplementary documents.

The modular architecture of ProHits will accommodate additional new features, as dictated by future experimental and analytical needs. Although ProHits has been designed to handle protein interaction data, simple modifications of the open source code will enable straightforward adaptation to other proteomics workflows.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank G. Bader, H. Hermjakob, S. Orchard, J.A. Vizcaíno, C. Le Roy, R. Beavis and members of the Tyers and Gingras laboratories for helpful discussions. We are grateful to D. Figeys, S. Angers, D. Fermin, T. LeBihan, F. Ellisma, C. Poitras and B. Coulombe for testing beta versions of ProHits. We thank W. Dunham, E. Deutsch, D. Fermin, T. Glatter, M. Goudreault, L. D'Ambrosio and R. Ewing for critical reading of the manuscript and instruction manual and L. Ng, J. Wei and N. Mohammad for IT support. Supported by grants from the CIHR (MOP-84314 to A.-C.G., MOP-12246 to M.T., MOP-81268 to B.R., GSP-36651 to T.P, J.L.W. and M.T., FRN 82940 to M.T., and a resource grant to T.P., A.-C.G, J.L.W. and M.T.), the NIH (5R01RR024031 to M.T. and CA-126239 to A.I.N.), MRI-ORF (T.P., J.L.W. and A.-C.G.), the Canada Foundation for Innovation (T.P., J.L.W., A.-C.G. and M.T.), and Genome Canada through Ontario Genomics Institute (T.P. and J.L.W.). We wish to acknowledge support from the Mount Sinai Hospital Foundation; Canada Research Chairs in Functional Genomics and Bioinformatics to M.T, in Proteomics and Molecular Medicine to B.R., and in Functional Proteomics to A.-C.G.; the Lea Reichmann Chair in Cancer Proteomics to A.-C.G. and a Scottish Universities Life Sciences Alliance Research Professorship and a Royal Society Wolfson Research Merit Award to M.T.

References

1. Gingras AC, Gstaiger M, Raught B, Aebersold R. Analysis of protein complexes using mass spectrometry. Nat Rev Mol Cell Biol. 2007; 8:645–654. [PubMed: 17593931]
2. Taylor CF, et al. The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol. 2007; 25:887–893. [PubMed: 17687369]
3. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999; 20:3551–3567. [PubMed: 10612281]
4. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004; 20:1466–1467. [PubMed: 14976030]
5. Keller A, Eng J, Zhang N, Li XJ, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol. 2005; 1:0017. [PubMed: 16729052]
6. Ewing RM, et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol Syst Biol. 2007; 3:89. [PubMed: 17353931]
7. Orchard S, et al. The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat Biotechnol. 2007; 25:894–898. [PubMed: 17687370]
8. Breitkreutz BJ, et al. The BioGRID Interaction Database: 2008 update. Nucleic Acids Res. 2008; 36:D637–640. [PubMed: 18000002]
9. Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–2504. [PubMed: 14597658]
10. Cottingham K. MCP ups the ante by mandating raw-data deposition. J Proteome Res. 2009; 8:4887–4888. [PubMed: 19891508]

11. Hermjakob H, et al. The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat Biotechnol. 2004; 22:177–183. [PubMed: 14755292]
12. Orchard S, et al. Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. Proteomics. 2007; 7 (Suppl 1):28–34. [PubMed: 17893861]
13. Kerrien S, et al. IntAct--open source resource for molecular interaction data. Nucleic Acids Res. 2007; 35:D561–565. [PubMed: 17145710]
14. Falkner JA, Hill JA, Andrews PC. Proteomics FASTA archive and reference resource. Proteomics. 2008; 8:1756–1757. [PubMed: 18442177]
15. Breitkreutz A, et al. A global protein kinase and phosphatase interaction network in yeast. Science. 2010; 328:1043–1046. [PubMed: 20489023]

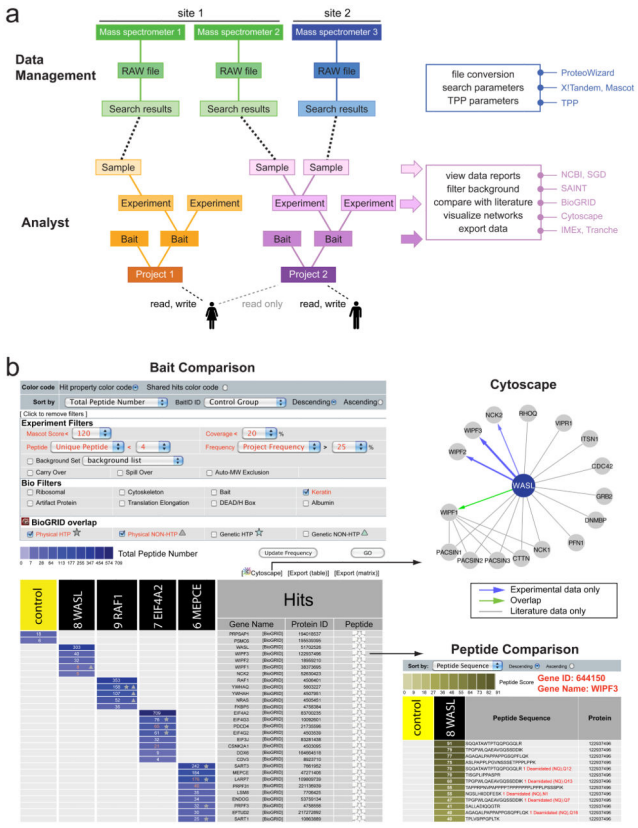

Figure 1.

Overview of ProHits. (**a**) Modular organisation of ProHits. The Data Management module backs up all raw mass spectrometry data from acquisition computers, and handles data conversion and database searches. The Analyst module organizes data by project, bait, experiment and sample (gel-free project shown; see Supplementary Fig. 8 for gel-based organization). Search results from the Data Management module are parsed to individual samples defined within the Analyst module. ProHits can handle large collaborative projects, and offers several security layers. In the Analyst module, several view, filter and export functions enable data analysis. Functions provided by external software are listed on the right. (**b**) ProHits Comparison page. *Left*: Filtered Comparison results for four human baits and one negative control (see Supplementary Fig. 17 online for unfiltered data). Display, sort, filter and literature overlap options are listed on the top; selected options in this example are shown in red. Filtered results are displayed at the bottom of the page. Columns represent individual baits. Comparison at the Experiment or Sample levels is also possible. Rows list the hits that pass selected filters. Color-coding and intensity in each cell is based in the property selected for visualization, shown for this example as total peptide numbers; mouse-overs of each cell will list all properties. A star or triangle inside the cell indicates an interaction identified in previous high-throughput (star) or low-throughput (triangle) studies in BioGRID. Each term in the hits column is hyperlinked to external databases (Entrez Gene, BioGRID or NCBI Protein) or to the list of identified peptides. *Right, top*: Visualization of data in Cytoscape with mass spectrometric information encoded as an edge attribute. Interactions detected for the example bait protein WASL that are not reported in

BioGRID are shown as blue edges with color intensity mapped spectral counts and thickness mapped to number of unique peptides; overlap interactions detected in both the experiment and in BioGRID are shown in green; interactions detected only in BioGRID are shown in grey. *Right, bottom*: Example of the Peptide View for the protein WIPF3 in the WASL AP-MS experiment.