ProHits: integrated software for mass spectrometry–based interaction proteomics

To the Editor:
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 studies. 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 have developed the ProHits laboratory information management system 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 data sets. It 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 using gel-based separation, gel-free analysis and multidimensional protein or peptide separation.

This software platform is a client-based HTML program written in PHP (PHP: Hypertext Preprocessor) 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 preexisting data management solutions or who receive precomputed search results from analyses performed in a core MS facility (Supplementary Fig. 3). A step-by-step installation package, installation guide and user manual (Supplementary Data) are available on the ProHits website (http://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 Figs. 4 and 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 (Minimum Information about a Proteomics Experiment) guidelines. Mascot, X!Tandem and the TransProteomics Pipeline (TPP) are fully integrated with ProHits via linked search engine servers (Supplementary Figs. 6 and 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 liquid chromatography-tandem MS (LC-MS/MS) procedures. Controlled vocabulary lists for experimental descriptions can be added by drop-down menus to facilitate compliance with annotation guidelines, such as MIAPE® and MIMIx (Minimum Information about a Molecular Interaction Experiment)®, 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 (e.g., technical replicates of the same sample or chromatographic fractions derived from the same preparation). All samples, experiments 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 high-throughput 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, MS data can be explored in an intuitive manner, and results from individual samples, experiments or baits can be viewed and filtered (Supplementary Figs. 13 and 14). A user interface enables alignment of data from multiple baits or MS analyses using the ‘Comparison’ viewing 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 and Supplementary Figs. 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 the removal of nonspecific background proteins from the hit list, as defined by negative controls, search engine score thresholds or contaminant lists. Links to the external US National Center for Biotechnology Information (NCBI) and the Biological General Repository for Interaction Datasets (BioGRID) databases are provided for each hit to facilitate data interpretation. Overlap with published interaction data housed in the BioGRID database can be displayed to allow immediate identification of new interaction partners. A flexible export function enables visualization in a graphical format with Cytoscape, 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 (Supplementary Figs. 18–20).

Deposition of all MS-associated data in public repositories is likely to become mandatory for publication of proteomics experiments. 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 reports, either in the MITAB (MapInfo. TAB binary) format for BioGRID or in XML format for submission to International Molecular Exchange (IMEx) consortium databases, including IntAct (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 Tranche.

ProHits was developed to manage many large-scale in-house projects, including a systematic analysis of kinase and phosphatase interactions in yeast, consisting of 986 affinity purifications. Smaller-scale projects from individual laboratories are readily handled in a similar manner. Examples of AP-MS data from

Figure 1. Overview of ProHits. (a) Modular organization of ProHits. The Data Management module backs up all raw MS 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. On the left are shown filtered Comparison results for four human baits and one negative control (see Supplementary Fig. 17 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 on the property selected for visualization, shown for this example as total peptide numbers; mouseovers 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 (EntrezGene, BioGRID or NCBI Protein) or to the list of identified peptides. The top right shows the visualization of data in Cytoscape with MS 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 gray. At the bottom right is an example of the Peptide view for the protein WIPF3 in the WASL AP-MS experiment.
both yeast and mammalian projects are provided in a demonstration version of ProHits (http://www.prohitsMS.com/) and in Supplementary Data.

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.

Note: Supplementary information is available on the Nature Biotechnology website.

AUTHOR CONTRIBUTIONS
G.L. and J.Z. devised and coded all aspects of the platform; C.S. and B.-J.B. implemented protein annotation and provided advice on database architecture; Y.D. wrote the Mascot parser; B.L., A.B., Z.-Y.L., K.C., A.P., A.I.N., T.P., J.L.W. and B.R. provided suggestions on software features; M.T. conceived and guided the project; A.-C.G., B.R. and G.L. wrote the instruction manuals; M.T. and A.-C.G. co-directed project development; A.-C.G. wrote the manuscript with input from B.R. and M.T.

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. Figyeys, S. Angers, D. Fermin, T. LeBihan, F. Ellisma, C. Poitras and B. Coulome 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 N. Ng, J. Wei and N. Mohammad for IT support. Supported by grants from the CIHR (MOP-83414 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 82390 to M.T. and a resource grant to T.P., A.-C.G., J.L.W. and M.T.), the NIH (5R01RO24301 to M.T., R01GM094231-01 to A.I.N. and A.-C.G., and CA-126239 to A.I.N.), MRC-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.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Guomin Liu1, Jianping Zhang1, Brett Larsen1, Chris Stark1, Ashton Breitkreutz1, Zhen-Yuan Lin1, Bobby-Joe Breitkreutz1, Yongmei Ding1, Karen Colwill1, Adrian Pascalescu1, Tony Pawson1,2, Jeffrey L Wrana1,2, Alexey I Nesvizhskii3, Brian Raught4, Mike Tyers2,5 & Anne-Claude Gingras1,2
1Centre for Systems Biology, Samuel Lunenfeld Research Institute, Toronto, Ontario, Canada. 2Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. 3Departments of Pathology and Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA. 4Ontario Cancer Institute and McLaughlin Centre for Molecular Medicine, Toronto, Ontario, Canada. 5Wellcome Trust Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, UK.
e-mail: m.tyers@ed.ac.uk or gingras@lunenfeld.ca

1. Gingras, A.C., Gstaiger, M., Raught, B. & Aebersold, R. Nat. Rev. Mol. Cell Biol. 8, 645–654 (2007).
2. Taylor, C.F. et al. (MIAPE). Nat. Biotechnol. 25, 887–893 (2007).
3. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Electrophoresis 20, 3551–3567 (1999).
4. Craig, R. & Beavis, R.C. Bioinformatics 20, 1466–1467 (2004).
5. Keller, A., Eng, J., Zhang, N., Li, X.J. & Aebersold, R. Mol. Syst. Biol. 1, 0017 (2005).
6. Ewing, R.M. et al. Mol. Syst. Biol. 3, 89 (2007).
7. Orchard, S. et al. Nat. Biotechnol. 25, 894–898 (2007).
8. Breitkreutz, B.J. et al. Nucleic Acids Res. 36, D637–640 (2008).
9. Shannon, P. et al. Genome Res. 13, 2498–2504 (2003).
10. Cottingham, K. J. Proteome Res. 8, 4887–4888 (2009).
11. Hermjakob, H. et al. Nat. Biotechnol. 22, 177–183 (2004).
12. Orchard, S. et al. Proteomics 7 Suppl 1, 28–34 (2007).
13. Kerrien, S. et al. Nucleic Acids Res. 35, D561–565 (2007).
14. Falkner, J.A., Hill, J.A. & Andrews, P.C. Proteomics 8, 1756–1757 (2008).
15. Breitkreutz, A. et al. Science 328, 1043–1046 (2010).

More sizzle thanizzle

To the Editor:
In an echo of Mark Twain’s response when reading his own published obituary that “The report of my death has been exaggerated,” I should like to correct an inaccuracy about GlaxoSmithKline’s (Brentford, UK) EpiNova Discovery Performance Unit (DPU), which was mentioned in Catherine Shaffer’s news article entitled “Pfizer explores rare disease path” from the September issue.

The article suggested that EpiNova had “fizzled out.” As vice president and head of the EpiNova DPU, I can confirm that, on the contrary, this early drug discovery unit continues to research and build alliances in our search to apply the knowledge of epigenetics to the quest for new medicines for patients. In fact, with our own first-class science, innovation and entrepreneurial spirit, and our external alliances with leading epigenetics research units, including the biotech company Cellzome (Heidelberg, Germany), and Cambridge (UK), Harvard (Cambridge, MA), Oxford (UK) and Rockefeller (New York) universities, we are in glowing health.

Many of your readers who have heard our presentations at the American Chemical Society meeting this August in Boston, as well as Miptec 2010 and the Society for Medicines Research Epigenetics Meeting held last month in Basel and London, respectively, will already know this. We are also sponsoring the “Epigenetics of Chromatin Modifications in Inflammation” meeting, taking place in Oxford this December.

We look forward to building on the excellent collaborations that are already in place and advancing our work in the area of immune-inflammation long into the future, with the aim of bringing out new innovative medicines for immuno-inflammatory diseases.

COMPETING FINANCIAL INTERESTS
The author declares competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Kevin Lee
EpiNova DPU, iCEDD, GlaxoSmithKline, Stevenage, UK.
e-mail: Kevin.2.lee@gsk.com