PROTEOME-3D: An Interactive Bioinformatics Tool for Large-Scale Data Exploration and Knowledge Discovery*

Deborah H. Lundgren‡, Jimmy Eng§, Michael E. Wright§, and David K. Han‡¶

Comprehensive understanding of biological systems requires efficient and systematic assimilation of high-throughput datasets in the context of the existing knowledge base. A major limitation in the field of proteomics is the lack of an appropriate software platform that can synthesize a large number of experimental datasets in the context of the existing knowledge base. Here, we describe a software platform, termed PROTEOME-3D, that utilizes three essential features for systematic analysis of proteomics data: creation of a scalable, queryable, customized database for identified proteins from published literature; graphical tools for displaying proteome landscapes and trends from multiple large-scale experiments; and interactive data analysis that facilitates identification of crucial networks and pathways. Thus, PROTEOME-3D offers a standardized platform to analyze high-throughput experimental datasets for the identification of crucial players in co-regulated pathways and cellular processes. Molecular & Cellular Proteomics 2:1164–1176, 2003.

Experimental methodologies, fine-tuned in recent years to allow high-throughput protein and cDNA analyses, have resulted in exponential growth of protein and cDNA expression profiles and interaction datasets. A number of large-scale analyses, such as the two-hybrid interaction maps and cDNA microarray technology, now allow interaction and expression datasets from large numbers of genes to be analyzed quickly and efficiently in a single experiment (1, 2). Protein profiling arrays for the comparable large-scale analysis of protein expression patterns are under active development as well (3, 4). When perfected, their output should be equally prolific. Finally, mass spectrometry, possibly the most important proteomics tool to date (5, 6), generates vast quantities of data through large-scale liquid chromatography (LC) tandem mass spectrometry (MS/MS) identification of expressed proteins in complex mixtures.

Predictably, technological advances enabling high-throughput analysis have resulted in an accumulation of experimental data at a rate far exceeding the current ability to assimilate that data. Transforming the rapidly proliferating quantities of experimental data into a usable form in order to facilitate data analysis is a challenging task. Numerous specialized databases and graphical tools have been described to organize the growing collection of large-scale experimental datasets (7–16). These tools have made significant contributions toward functional data organization and the display of protein complexes and hierarchical relationships. Yet the initial interpretation of experimental datasets in an interactive and intuitive way remains a challenge. Important functional information can only be determined through careful and detailed analysis of experimentally identified and quantified data in the context of the current knowledge base. Functional analysis, which is requisite to an exhaustive understanding of cellular networks and pathways, represents a major bottleneck in proteomics today. It is recognized that bridging the expansive gap between the current state of knowledge and the ultimate goal of understanding whole cellular networks requires a global discovery phase to pinpoint pivotal proteins in cellular networks (17). Tools that integrate diverse experimental results with the current knowledge base would undoubtedly facilitate the understanding of biological networks and pathways. Visualization of biological data is an important component of such applications (18).

We describe here a Web-based data exploration and knowledge discovery tool called PROTEOME-3D that utilizes three essential features for effective assimilation and analysis of large-scale experimental datasets: 1) automated construction of a customized database of expressed proteins/mRNAs from the public knowledge base using user-defined criteria; 2) graphical tools for displaying and comparing experimental results in the form of proteomic landscapes; 3) an interactive user interface for in-depth analysis of experimental results. Sample applications are provided to demonstrate how this tool can facilitate the evaluation of experimental results. (For information on how to obtain a copy of PROTEOME-3D, contact David K. Han at han@nso.uchc.edu.)

From the ‡Center for Vascular Biology, Department of Physiology, University of Connecticut School of Medicine, Farmington, CT 06030, and §Institute for Systems Biology, Seattle, WA 98103. Received, June 24, 2003, and in revised form, August 6, 2003. Published, MCP Papers in Press, August 7, 2003, DOI 10.1074/mcp.M300059-MCP200

© 2003 by The American Society for Biochemistry and Molecular Biology, Inc.

This paper is available on line at http://www.mcponline.org
Experimental procedures

Information flow—The general flow of information through PROTEOME-3D is outlined in Fig. 1. Experimental results generated from isotope-coded affinity tag (ICAT) analysis or from cDNA microarrays are pre-processed to create an input file of protein identities (ids) and abundance ratios (see “Database” subsection below for more detail). Protein ids are then used to generate a customized, user-defined dataset from public databases, and the combined experimental and retrieved data are stored in a local database. The PROTEOME-3D graphical interface is accessed through Internet Explorer. Three-dimensional (3D) display and protein page screens are linked for easy navigation, and each screen communicates with the local database through a servlet stored on the server (19). The protein page provides user-selectable links to public and/or proprietary databases and the capability to construct additional customized links.

Database—Experimental results, together with a customized dataset retrieved from public databases, are stored locally in a relational database (Oracle 9i). For each experiment loaded in the database, a list of MS/MS-identified proteins and their calculated abundance ratios is initially read from an INTERACT summary web page, which contains one row of data for each peptide scan conclusively identified by SEQUEST and quantified by XPRESS (20, 21). Alternately, microarray output identified by gene ids and stored in a tab-delimited file is read in a pre-processing step, and a file of corresponding protein ids and abundance ratios is produced. A series of Java application programs are then executed, resulting in population of the local database with the experimental results and desired user-defined data. At a minimum, the experiment table contains, for each experiment, a nested table of entries comprising protein accession number, proportion, and standard deviation of the proportion. The proportion attribute is a mean value derived from the abundance ratio as follows: for each identified peptide in a given experiment, the peptide’s abundance ratio A/B is used to calculate a proportion B/(A + B), where B is the relative abundance of the peptide in the experimentally perturbed sample and A is its relative abundance in the control sample. Then, from all the peptide samples for a given protein, an average proportion and standard deviation are computed and stored in the nested table entry. This value serves as a normalized representation of abundance ratio, providing comparable values for comparisons between up- and down-regulated proteins, with values ranging from 0 for maximum down-regulation to 1 for maximum up-regulation. Additional information pertinent to each experiment can be stored in the experiment table as well.

Detailed information retrieved from public databases for each protein is stored in the protein table. An example of a set of stored attributes includes National Center for Biotechnology Information (NCBI) Protein Data Bank accession number, molecular weight, pI, cross-references to NCBI nucleotide and Online Mendelian Inheritance in Man (OMIM) databases, maploc, and a variety of descriptive fields such as keywords, definition, function, disease, subcellular location, and pathway. To avoid database redundancies and changing accession numbers, to retrieve latest annotations, and to ensure that the local database accurately reflects latest updates, it is necessary to routinely download the latest publicly available databases and subsequently update the local Oracle database.

Results

Data Presentation

3D Graphic Display—An interactive Java3D graphic display represents a given experiment as a set of cone objects (Fig. 2, upper left). Each cone depicts a protein, uniquely identified by its mass (x-axis) and pI (z-axis). The abundance ratio (converted to a proportion, as described above) is graphed on the y-axis and corresponds to the height of the cone. The base of each cone sits on the plane of y = 0.5 (green), which represents a 1:1 abundance ratio. Cones depicting up-regulated and down-regulated proteins project above and below the reference plane, respectively. Cone color is mapped to the interval (0, 1) on the y-axis, with blue, green and red representing 0.0, 0.5, and 1.0, respectively. Bright red cones, then, denote highly up-regulated proteins, and bright blue cones denote those proteins that are highly down-regulated. For specific information on an individual protein, its cone can be selected by mouse click, which highlights the cone and displays a subset of the protein’s attributes in a text field along the bottom of the screen. Next to the text field is a button linking the highlighted protein to the protein page screen, which interfaces with the local database and a number of external data sources. Additional buttons at screen bottom allow selection of alternate experimental displays. Global scene manipulation (zooming in or out, rotation, and scene translation up, down, left, or right) is executed through mouse buttons and the Alt key.

To the right of the graphic display is the content pane, which provides the functionality for experiment-to-experiment
Interactive Proteome Landscape Facilitates Data Exploration

3D Graphic Display

Exp/Exp Comparison

Detail of Intersecting Cones

Protein Page

Local Database

Public/Proprietary Databases
comparisons. A second experiment is chosen for comparison with the current experiment from a drop-down list under the heading “Select 2nd Exp.” Optionally, a separate color can be chosen to represent each experiment (“Exp Colors”), enabling cones of the two experiments to be easily distinguished in the graphic display. Desired comparisons are executed through radio buttons under the heading “Select Function,” with options including intersection, union, and complements. An intersection between two experiments is illustrated in Fig. 2, middle and lower left, where Exp1 cones are displayed in violet and Exp2 cones are displayed in green. The detail image of the intersection (Fig. 2, lower left) shows how the inner cone is visible through the transparent outer cone, allowing a visual comparison of the protein’s expression in the two experiments.

Protein Page—The protein page screen (Fig. 2, lower right) provides the functionality for in-depth analysis of a selected protein. It comprises an interface to the local Oracle database, a customizable set of links to user-selected external Web sites, and a query-building tool for use with the local database. The protein page is accessible from the graphic display screen by a button click, as described in the previous paragraph. The screen is functionally divided into two sections. The top section displays attributes of the current protein that are stored in the local database, including comments that can be selected by name from a drop-down list. Additional information about the protein is accessible through customized links, which are defined on a pop-up window where the user can select from default sites or input his/her own URLs. The bottom section of the protein page screen comprises a query-builder tool, with a button linking back to the display page, allowing results of database queries to be viewed as a 3D graphic display. Fields, operators, and connectors for a query are selected from drop-down lists; the search key is typed into a text box. Subqueries can be combined, using an appropriate connector, until the desired query is complete. After execution, query results can be viewed in a text box or as a 3D graphic display. Additional clauses can be added to the existing query, facilitating a natural progression to increasingly more specific subsets of data, or, alternately, the user can begin again with a new query.

Sample Applications—PROTEOME-3D provides a visual summary of potentially massive sets or selected subsets of experimental data, including experiment/experiment comparisons; an advanced querying capability on locally stored data; and direct exploration of experimental results in the context of the public knowledge base. This combination of features allows efficient navigation through vast quantities of data and extraction of subsets of experimental proteins with unique properties relevant to the area of research. The chosen subset can then be viewed graphically and further analyzed by accessing the public databases, or by experiment/experiment comparison, in which common or unique proteins from a second profiling experiment are jointly displayed in the 3D graphical format. The following examples, using data from ICAT proteomics profiling experiments (20, 25),2 demonstrate specifically how this tool can help analyze large-scale protein profiling experiments systematically and efficiently.

Example 1: Analysis of Membrane Proteins That Are Differentially Regulated During Apoptosis—Apoptotic cells are rapidly recognized and engulfed by neighboring cells and macrophages. This recognition process is theoretically a summation of changes in adhesive and repulsive signals presented on the apoptotic cell surface (22). In an effort to identify adhesive engulfment ligands, we profiled membrane proteins from control Jurkat cells and anti-Fas immunoglobulin M-treated apoptotic Jurkat cells, using a previously described method (20). We then manually explored the experimental results, together with pertinent public databases, searching for candidate proteins for further study. In fact, the impetus for creating an automated discovery tool came from this and similar time-consuming, manual searches of databases and experimental results. Here we explore the utility of PROTEOME-3D in the analysis of this dataset. Briefly, proteins positively identified (with \( p > 0.9 \)) and experimentally quantified from the INTER-ACT page (Fig. 3A) were loaded into the PROTEOME-3D local database, automatically providing three inter-related tools for systematic and efficient analysis: 1) the local database itself, storing information retrieved from a variety of public databases, together with experimental data; 2) the interactive 3D graphical display of experimental results; and 3) a protein page linked to the 3D display, providing an interactive, queryable interface to the local database as well as user-defined links to publicly available network and pathway databases. These features allow rapid analysis of proteins for putative engulfment ligands. Among the 114 proteins identified and quantified from the Jurkat experiment, one stands out immediately in the 3D graphic display of experimental results (Fig. 3B); highly up-regulated human annexin I protein (ANX1_HUMAN). Selecting this protein by mouse click, we

2 D. K. Han, unpublished data.

Fig. 2. Interactive PROTEOME-3D screens. Graphical display of protein up- and down-regulation for single experiment is shown in upper left. The height/color of each cone represents the proportion of its experimental quantity to its total (control + experimental) quantity. Cones above the translucent green plane are up-regulated; those below are down-regulated; the green plane represents no change. In experiment/experiment comparison (middle and lower left), proteins common to both experiments (intersect) are shown. For proteins up-regulated or down-regulated in both experiments, inner cone is visible within translucent outer cone (as indicated by arrow). Protein page (right) provides an interactive interface, with text fields displaying locally stored protein attributes; links to public or proprietary databases of interest; and a query-building tool for searching the local database. Graphical Display and Protein Page screens are linked via a button to allow easy navigation between them.
Fig. 3. Utilizing PROTEOME-3D to evaluate ICAT profiling proteomics experimental results. A, Interact file (partially shown) contains a record for each MS peptide scan. B, Graphic display reveals highly up-regulated (red) and highly down-regulated (blue) proteins at a glance. C, Up-regulated protein, ANX1_HUMAN, has been selected for investigation via Protein Page.
switch to the protein page for further exploration of this protein (Fig. 3C). Its keyword annotation lists calcium/phospholipid-binding as a known function, and recent literature cited in the OMIM link implicates annexin 1 in anti-inflammatory activity (23, 24). These features suggest a possible role for human annexin 1 in the apoptotic cell engulfment process and anti-inflammatory activity. Due to its significant up-regulation in the graphic display and its membrane-binding activities, this protein is quickly recognized as an intriguing candidate for further study. In fact, additional experiments were performed on annexin 1 (as a result of the original manual discovery process), and its function as an endogenous engulfment ligand was discovered (25). Thus, the ability to graphically visualize the proteome landscape of a profiling experiment and quickly access known biological activity allows the user to focus on candidate proteins that are most relevant to the particular biological system.

**Example 2: Analysis of Membrane Proteins from Prostate Cancer Cells That Are Regulated by Treatment with Androgen Homolog R1881**—Selection of a smaller subset of proteins from a large number of experimentally identified and quantified proteins or cDNAs requires efficient and systematic analysis of the experimental dataset. Here, we describe the use of PROTEOME-3D to analyze experimental ICAT profiling data from a prostate cancer cell line (LNCaP) treated with or without androgen analog R1881 for 24 h. The entire experimentally determined dataset includes 4052 proteins displayed in Fig. 4A. It is well documented that multi-step carcinogenesis requires sequential aberrations in chromosomal DNA, and a large number of chromosomal changes associated with prostate cancer have been described (Table I). The ability of PROTEOME-3D to automatically retrieve user-relevant data for local storage makes chromosomal localization of all of the identified and quantified proteins a queryable feature. Querying all proteins mapped to any of the loci listed in Table I results in the selection of 921 proteins. This subset is then further refined using the functional term “Tumor suppressor,” resulting in a subset of 31 proteins. Adding to the query the more specific term “Tumor suppressor” results in the two down-regulated and three up-regulated proteins responsive to androgen treatment composing the final subset. This progressive refinement of the search criteria, along with a graphic display of the resultant subset, is depicted in Fig. 4B. Now a much more manageable set of candidate proteins can be further analyzed via the protein page (Fig. 4C), allowing for additional queries and/or in-depth analysis of pathway and network databases. For example, one of the up-regulated proteins, ATM_HUMAN, is selected from the graphic display by mouse click and its links are followed to GENEGO and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway databases (Fig. 4C, lower left and right, respectively). Thus, interactions and networks associated with this experimentally identified and quantified tumor suppressor protein can be easily explored from the PROTEOME-3D platform, allowing easy integration of experimental data with the current knowledge base.

**Example 3: Analysis of Coregulation of Cellular Processes**—Many biological processes are controlled by multi-protein complexes whose properties, such as abundance and subcellular location, are integral to that control (26). Exploring coregulation of proteins categorically—proteins associated with particular functional groups or particular subcellular locations, for example—provides crucial biological information. This type of analysis can be easily done with PROTEOME-3D, using the query builder tool and displaying the results graphically (Fig. 5). We functionally categorized proteins from the LNCaP profiling experiment as mitochondrial, glycolytic, housekeeping, and structural by querying locally stored comment and keyword fields. The resulting graphical displays, which summarize the profiles of proteins associated with these functional groups, are shown in Fig. 5A. In Fig. 5B, the experimental results are categorized by subcellular location into the four broad groups of extracellular (secreted), cytoplasm, membrane, and nucleus. In each of the displays, highly up- or down-regulated proteins are easily distinguished by their corresponding bright red or blue color and can be easily selected for further analysis.

**Example 4: Analysis of Multiple Profiling Experiments**—The efficient evaluation of multiple, large-scale expression analyses, time-dependent changes in expression, and protein profiles generated by diverse methodologies requires a software platform that can simultaneously display multiple experiments for analysis. We have implemented a unique component of PROTEOME-3D, termed “Multi-Experiment Comparison,” where multiple proteome, cDNA microarray, or combinations of profiling experiments can be efficiently analyzed. The first example (Fig. 6, A–C), comparing datasets extracted from Jurkat and LNCaP cell lines described above (Example 1 and Example 2), demonstrates salient features of this tool. Fig. 6A illustrates the intersection of proteins identified in each experiment, allowing a visual comparison of expression patterns for proteins common to the two cell lines. Zooming in, as shown previously in Fig. 2 (lower left), allows the experimentalist to discern similar or divergent patterns of regulation of a particular protein between the two experiments. Fig. 6B illustrates the Complement1 function, depicting proteins identified in Exp1 (Jurkat cell line) but not in Exp2 (LNCaP). Additional functions (not shown) include a display of proteins uniquely identified in Exp2 (Complement2), a combination of Complements 1 and 2 (Complement), and all proteins identified in either experiment (Union). Subsets of the data can also be easily extracted for comparison, as in Fig. 6C, where expression profiles of cytoplasmic proteins common to both experiments (Intersection) are displayed.

PROTEOME-3D’s multi-experiment comparison can also be used to analyze a cDNA microarray dataset concurrently with an ICAT proteome profiling experiment, as shown in Fig.
Interactive Proteome Landscape Facilitates Data Exploration

A. 

B. 
4052 proteins in total

Query 1

Subset of 921 proteins

Query 2

Subset of 31 proteins

Query 3

Subset of 5 proteins

C. 

Protein Data

Exp/Protein: UNCAP ATM_HUMAN Change

Involved in signal transduction, cell cycle control and DNA repair. May function as a tumor suppressor. Necessary for activation of AELi and SAPK. Phosphorylates p53, NFkBIA, BRCA1, CTIP, NIE1N (NEB1), TEF1, and RAD9. May play a role in vesicle and/or protein transport. Inhibited by wortmannin. Could play a role in T-cell development, gonad and neurological function.

*Database Links: pdb red uniprot genpep keep uni

Copyright © Ganego, 2000-2003

Molecular & Cellular Proteomics 2.11
TABLE I
Known chromosomal aberrations documented in human prostate cancer cells

| Chromosome Location | Gain | Loss | Reference Source |
|---------------------|------|------|-----------------|
| 1q                  | x    | x    | 33, 39          |
| 2p or q             | x    | x    | 38              |
| 3p or q             | x    | x    | 38              |
| 4p or q             | x    | x    | 33              |
| 5q                  | x    | x    | 33, 35, 38      |
| 6p or q             | x    | x    | 28–30, 33, 37–38 |
| 7p or q             | x    | x    | 35, 37–39       |
| 8p or q             | x    | x    | 28, 29, 33–39   |
| 9p or q             | x    | x    | 35, 38          |
| 10p or q            | x    | x    | 28, 37–39       |
| 11p or q            | x    | x    | 33, 38          |
| 12q                 | x    |      | 33              |
| 13q                 | x    | x    | 28, 31, 33–35, 37–39 |
| 14q                 | x    |      | 33              |
| 15q                 | x    |      | 33              |
| 16p or q            | x    | x    | 28, 29, 33, 37–39 |
| 17p or q            | x    | x    | 28, 29, 33, 35, 37–38 |
| 18p or q            | x    | x    | 29, 35, 37–38   |
| 19                  | x    | x    | 27, 28          |
| 20p or q            | x    |      | 33              |
| 21p or q            | x    |      | 33              |
| 22p or q            | x    | x    | 33              |
| Xp or q             | x    | x    | 28, 37–39       |
| Yp or q             | x    | x    | 27, 32, 37      |

6, D–F. LNCaP cells treated with androgen analog R1881 were used to isolate mRNAs and proteins for comparative analysis. The regulation of 56 genes is displayed in Fig. 6D, where specific gene accession numbers have been converted to Swiss-Prot loci in order to compute molecular weight and pl for the 3D graphical display page. Subsequent comparison of 56 genes with 2294 proteins from the ICAT experiment, shown in Fig. 6E, results in 25 common proteins. Using this feature of PROTEOME-3D, the investigator can easily explore and analyze co-regulated and differentially regulated mRNAs and proteins (Fig. 6F).

DISCUSSION

Systematic and efficient analysis of vast genomic and proteomic data sets is a major challenge for researchers today. Crucial biological advances in the study of model organisms are made daily, and new information is continuously deposited in publicly available databases. Thus, for each protein or mRNA that is identified and quantified from expression profiling experiments, a wealth of biologically relevant information, such as associated biological networks and pathways, protein interaction partners, biochemical activities, pathological/disease association, subcellular and tissue-specific expression, domain structure and function, may exist in publicly available databases. These datasets can be efficiently utilized by experimentalists to decipher complex functions of proteins if the crucial information is easily accessible. Yet due to the profound differences among the various biological databases housing such diverse data, information retrieval is an overwhelmingly manual, rate-limiting step in the researcher’s analysis of experimental results, making integration of existing biological data a critical problem (40). Thus, to systematically and efficiently evaluate large-scale experimental results in the context of existing biologically relevant data, at least four crucial features are required: 1) automatic retrieval of user-defined information to construct a customized, queryable database; 2) an intuitive graphical and query platform to display and analyze experimental data in the context of the customized database; 3) efficient utilization of web-based bioinformatics software tools for data interpretation, prediction of function, and modeling; and 4) scalability and reconstruction of the database in response to changing user needs and an ever-expanding base of knowledge and bioinformatics tools.

Creating a software tool to encompass the four crucial features outlined above is a challenging and ongoing task, particularly with respect to the ever-expanding publicly available base of knowledge and bioinformatics tools. PROTEOME-3D represents an initial attempt to automate the laborious and time-consuming process of experimental data analysis in order to efficiently identify the most salient features and evaluate those features in the context of the existing knowledge base. It is a platform built upon the integration of three related components: 1) a scalable, queryable, customized relational database for local storage of user-defined biologically relevant data; 2) an intuitive and interactive 3D display for evaluating and comparing experimental results; and 3) an interactive user interface for systematic analysis of experimental data. We have demonstrated with specific examples how PROTEOME-3D can be effectively utilized for large-scale experimental analysis. The interactive 3D proteomic landscape provides a striking visual overview of experimental results whose notable features can then be further analyzed through integration of the display with the local, queryable database of biologically relevant data. A customizable set of links to user-selected external sources expedites concurrent utilization of those web-based tools of particular relevance to the analysis. This ability to link PROTEOME-3D with a number of web resources is especially critical for in-
Interactive Proteome Landscape Facilitates Data Exploration

A. Mitochondrial  Glycolytic

Housekeeping  Structural

B. Extracellular  Cytoplasm

Membrane  Nucleus
vestigation of protein families, splice isoforms, and multiple redundant entries in the databases. For example, exploration of a regulated protein/cDNA may require detailed analysis of multiple splice isoforms expressed in that tissue. Thus, the ability to create customized links and retrieve user-defined data is essential for detailed data exploration and knowledge discovery.

We intend to expand the capabilities of PROTEOME-3D in several areas. One critical need is the development of a customizable database interface, implemented as a Web-based form, allowing users to specify attributes of interest from a defined set of public databases for inclusion in their local, queryable database. In this way the software can be adapted to the specific and evolving needs of individual laboratories. In our initial implementation, we retrieved data from GenBank and OMIM databases. We intend to expand that list to include AmiGO, the Gene Ontology database, to enable queries on a standardized set of annotations (16); Alliance for Cellular Signaling’s Molecule Pages database, to obtain qualitative and quantitative data requisite to network modeling (41, 42); and additional databases, such as the Biomolecular Interaction Network Database (12), Human Protein Reference Database (www.hprd.org), liveDIP (7), and KEGG (www.genome.ad.jp/kegg/kegg2.html), to provide pertinent data on protein activities, networks, pathways, and interactions.

We also envision expanding our software to automatically interface with key bioinformatics resources. For example, the Virtual Cell project developed at University of Connecticut Health Center (43) provides tools for building testable models based on experimental results. It is an invaluable aid in understanding biological systems, fine-tuning theoretical models, and guiding the direction of future research. Virtual Cell is already accessible from our protein page interface via a customized link. We are currently working toward automating at least part of the model-building process, using data retrieved from public databases and supplementing with experimental data generated in the laboratory in order to automatically create the initial compartments and species of the Virtual Cell. As reference databases mentioned above become more densely populated with both qualitative and quantitative protein interaction data, more of the model-building process will be adaptable to automation. Also, integrating output from the Virtual Cell simulation with PROTEOME-3D’s graphical display will allow a visual comparison of Virtual Cell’s predicted cell protein profiles with the corresponding experimentally determined abundance ratios. This automation will expedite the detailed analysis of large-scale datasets by virtual-experimental tools.

PROTEOME-3D’s multiple-experiment comparison feature is a flexible tool that can be used effectively in a number of diverse applications. For instance, as quantitative protein profiling studies become more prevalent (44), the experiment/experiment comparison can be used to simultaneously display pathological versus normal tissue samples from multiple patients in order to visualize how the profile patterns differ. Alternately, this feature provides the means for simultaneous comparison of protein profiles generated by two different methodologies, such as the microarray-generated differential expression data and ICAT-LC/MS/MS-generated abundance ratios displayed in Fig. 6, to see how well their protein compositions and abundance ratios coincide under a given set of experimental conditions.

Although a number of graphical programs have been recently introduced for use in the analysis of microarray datasets—for example, Gene MicroArray Pathway Profiler (45), Onto-Express (46), MAPFinder (47), and GoMiner (48)—PROTEOME-3D represents a novel approach to initial data exploration and knowledge discovery. Rather than the text-based graphical files that are output by the programs mentioned above, PROTEOME-3D displays experimental results in the form of interactive 3D proteomic landscapes, which are easily interpreted visually and easily superimposed to provide a visual experiment/experiment comparison as well. In addition to its flexible graphical display, PROTEOME-3D constructs a customized, queryable local database of expressed proteins/mRNAs from the public knowledge base and provides an interactive user interface for efficient and systematic analysis of experimental results. This paper has detailed specific features of PROTEOME-3D and described numerous sample applications, although the tool has a general applicability that extends beyond the specific examples presented here. Its utility in the analysis of large-scale experimental datasets makes it an invaluable tool in multiple biological and computational research environments.

Acknowledgments—We thank R. Aebersold, J. Glomset, and R. Berlin for their comments and support, Les Loew, DongGuk Shin, Ann Cowan, Hsin-wei Wang, and Winfred Kruger for helpful discussion and technical assistance, and members of Han laboratory for discussions. We also thank Andrej Bugrim for giving us a trial license for the Genego software and for his help in network and pathway analysis.

* This work was supported by National Institutes of Health Grants HL67569, GM65764, and HL 70694. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

† To whom correspondence should be addressed: David K. Han, Center for Vascular Biology, Department of Physiology, MC 3501, E 5041, University of Connecticut Health Center, 263 Farmington Ave.

Fig. 5. Filtering experimental results by category. A, Multiple functional groups are selected and displayed using the 3D graph: mitochondria, energy—87 proteins; fatty acid, glucose, glycolysis—62 proteins; cell proliferation, growth, maintenance—109 proteins; cell adhesion, cytoskeleton—129 proteins. B, Proteins from distinct subcellular compartments are selected and displayed: extracellular, secreted—76 proteins; cytoplasm—173 proteins; membrane—368 proteins; nucleus—218 proteins.
Fig. 6. Multiple experiment comparisons. Jurkat (Exp1) and LNCaP (Exp2) ICAT experiments are compared in A–C. Proteins identified and quantified are displayed in violet (Exp1) and gold (Exp2). A, Proteins commonly identified in both Exp1 and Exp2 (intersection) are shown. B, Proteins unique to Exp1 (complement1) are shown. C, Intersection of cytoplasmic proteins in Exps 1 and 2 is shown. cDNA microarray results (Exp1, D), ICAT profiling results (Exp2, E), and their intersection (F) are shown. More highly up- or down-regulated genes/proteins are displayed with translucent colors to allow visualization of the inner cones.
REFERENCES

1. Auerbach, D., Thaminy, S., Hottinger, M. O., and Stagljar, I. (2002) The post-genomic era of interactive proteomics: Facts and perspectives. Proteomics 2, 511–623.

2. Shoemaker, D. D., Shoemaker, E. E., Armour, C. D., He, Y. D., Garrett-Engle, P., McDonagh, F. D., Loerch, P. M., Leonardson, A., Lum, Y. P., Cavet, G., Wu, L. F., Altschuler, S. J., Edwards, S., King, J., Tsang, J. S., Schimmack, G., Schelter, J. M., Koch, J., Ziman, M., Martin, M. L., Li, B., Cundiff, P., Ward, T., Castle, J., Krolevski, M., Meyer, R. M., Yao, A., Burchard, J., Kiid, M. J., Dai, H., Phillips, J. W., Linsley, P. S., Stoughton, R., Scherer, S., and Boguski, M. S. (2001) Experimental annotation of the human genome using microarray technology. Nature 409, 922–927.

3. Jenkins, R. E., and Pennington, S. R. (2001) Arrays for protein expression profiling: Towards a viable alternative to two-dimensional gel electrophoresis? Proteomics 1, 13–29.

4. MacBeath, G. (2002) Protein microarrays and proteomics. Nat. Genet. Suppl. 32, 526–532.

5. Aebersold, R., and Goodlett, D. R. (2001) Mass Spectrometry in Proteomics. Chem. Rev. 101, 269–295.

6. Hochstrasser, D. F., Sanchez, J., and Appel, R. D. (2002) Proteomics and its trends facing nature’s complexity. Proteomics 2, 807–812.

7. Xiaoqun, J. D., Xaniarios, I., and Eisenberg, D. (2002) Describing biological protein interactions in terms of protein states and state transitions. Mol. Cell. Proteomics 1, 104–116.

8. Sirava, M., Schafer, T., Eiglsperger, M., Kaufmann, M., Kohlbacher, O., Bornberg-Bauer, E., and Lengh, H. P. (2002) BioMiner—Modeling, analyzing, and visualizing biochemical pathways and networks. Bioinformatics 18, S219–S230.

9. Karp, P. D. (2002) Pathway databases: A case study in computational symbolic theories. Science 293, 2040–2044.

10. Salamonson, W., Mok, K. Y. C., Kolatkar, P., and Subbiah, S. (1999) BioJAKE: A tool for the creation, visualization and manipulation of metabolic pathways. Proceedings of the Pacific Symposium on Biocomputing 1999, 392–400.

11. Karp, P. D. (1999) Metabolic databases. Trends Biochem. Sci. 23, 114–116.

12. Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F. F., Pawson, T., and Schaff, J. V. (2001) Loss of the short arm of the Y chromosome in human prostate cancer. Eur. Urol. 39, 275–282.

13. Bae, V. L., Jackson-Cook, C. K., Maygard, S. J., Plymate, S. R., Chen, J., and Ware, J. L. (1998) Metastatic sublines of an SV40 Large T antigen immortalized human prostate epithelial cell line. The Prostate 34, 275–282.

14. Trapman, J. (2002) Molecular genetics of prostate cancer. (www.eur.nl/fgg/pathol/research/trapman/pc_genetics.htm).

15. Steiner, T., Junker, K., Burkhartd, F., Braunsdorf, A., Janitzky, V., and Schubert, J. (2002) Gain in chromosome 8q correlates with early progression in hormonal treated prostate cancer. Eur. Urol. 41, 167–171.

16. Verhagen, P. C., Hermans, K. G., Brok, M. O., Van Weerden, W. M., Timans, M. G., de Weger, R. A., Boon, T. A., and Trapman, J. (2002) Deletion of chromosome region 8q14–16 in prostate cancer. Int. J. Cancer 102, 142–147.

17. Latli, A., Morant, P., Fournier, G., Mangin, P., Berthon, P., and Cussenot, O. (2002) CHC1-L, a candidate gene for prostate carcinogenesis at 13q14.2, is frequently affected by loss of heterozygosity and underexpressed in human prostate cancer. Int. J. Cancer 99, 689–696.

18. Jordan, J. J., Hanlon, A. L., Ale-Salem, T., Greenberg, R. E., and Tricoli, J. V. (2001) Loss of the short arm of the Y chromosome in human prostate carcinoma. Cancer Genet. Cytogenet. 124, 122–126.

19. Kasahara, K., Taguchi, T., Yamazaki, I., Kamada, M., Yuki, K., and Shuin, T. (2002) Detection of genetic alterations in advanced prostate cancer by comparative genomic hybridization. Cancer Genet. Cytogenet. 137, 59–63.

20. Schulz, W. A., Elo, J. P., Flori, A. R., Penmanen, S., Santourilids, S., Engers, R., Buchardt, M., Seift, H. H., and Visakorpi, T. (2002) Genomewide DNA hypomethylation is associated with alterations on chromosome 8 in prostate carcinoma. Genes Chromosomes Cancer 35, 58–65.

21. Wolter, H. & Trijic, D., Gottfried, H. W., and Mattfeldt, T. (2002) Chromosomal changes in incidental prostatic carcinomas detected by comparative genomic hybridization. Eur. Urol. 41, 328–334.

22. Tsuchiya, N., Slezk, J. M., Liber, M. M., Bergstralh, E. J., and Jenkins, R. B. (2002) Clinical significance of alterations of chromosome 8 detected by fluorescence in situ hybridization analysis in pathologic organ-confined prostate cancer. Genes Chromosomes Cancer 34, 363–371.

23. Brothman, A. R., Maxwell, T. M., Cui J., Deubler, D. A., and Zhu, X. L. (1999) Chromosomal clues to the development of prostate tumors. Prostate 38, 303–312.

24. Bova, G. S., and Isaacs, W. B. (1996) Review of allelic loss and gain in prostate carcinoma. World J. Urol. 14, 338–346.

25. Nuopponen, N., and Visakorpi, T. (1999) Molecular biology of progression of prostate cancer. Eur. Urol. 35, 351–354.

26. Stein, L. D. (2003) Integrating Biological Databases. Nature Rev. Gen. 4, 337–345.

27. Tausig, R., Ranganathan, R., Ross, E. M., and Gilman, A. G. (2002) Overview of the Alliance for Cellular Signaling. Nature 420, 703–706.

28. Liu, J., Ning, Y., Hedley, W., Saunders, B., Chen, Y., Tindill, N., Hannay, T., and Subramaniam, S. (2002) The Molecule Pages database. Nature 420, 716–717.

29. Schaff, J., and Loew L. M. (1999) The virtual cell, in Biocomputing: Proceedings of the 1999 Pacific Symposium (Altmann, R. B., Dunker, A. K., Hunter, L., Klein, T. E., and Lauderdale, K., eds), pp. 228–239, World Sci, Singapore.

30. Aebersold, R. (2003) Constellations in a cellular universe. Nature 422, 115–116.

31. Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C., and Conklin, B. R. (2002) GenMAPP, a new tool for viewing and analyzing microarray
46. Khatri, P., Draghici, S., G. C. Ostermeier, and Krawetz, S. A. (2002) Profiling gene expression using Onto-Express. Genomics \textbf{79}, 266–270
47. Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., and Conklin, B. R. (2002) MAPFinder: Using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data.
48. Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Labadidi, S., Bussey, K. J., Riss, J., Barrett, J. C., and Weinstein, J. N. (2003) GoMiner: A resource for biological interpretation of genomic and proteomic data. \textit{Genome Biol.} \textbf{4}, R28