Computational Proteomics Analysis System (CPAS): An Extensible, Open-Source Analytic System for Evaluating and Publishing Proteomic Data and High Throughput Biological Experiments

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · FEBRUARY 2006

Impact Factor: 4.25 · DOI: 10.1021/pr0503533 · Source: PubMed

Matthew Bellew
LabKey Software
7 PUBLICATIONS **479** CITATIONS
SEE PROFILE

Jimmy K Eng
University of Washington Seattle
140 PUBLICATIONS **23,323** CITATIONS
SEE PROFILE

Vitor Faca
University of São Paulo
65 PUBLICATIONS **2,312** CITATIONS
SEE PROFILE

David J States
OncProTech LLC
110 PUBLICATIONS **14,357** CITATIONS
SEE PROFILE

research articles Journal of proteome research

Computational Proteomics Analysis System (CPAS): An Extensible, Open-Source Analytic System for Evaluating and Publishing Proteomic Data and High Throughput Biological Experiments

Adam Rauch,[†,‡] Matthew Bellew,[†,‡,||] Jimmy Eng,[†,||] Matthew Fitzgibbon,[†,||] Ted Holzman,[†,||] Peter Hussey,[†,‡,||] Mark Igra,[†,‡,||] Brendan Maclean,[†,‡,||] Chen Wei Lin,[†] Andrea Detter,[†] Ruihua Fang,[†] Vitor Faca,[†] Phil Gafken,[†] Heidi Zhang,[†] Jeffrey Whitaker,[†] David States,[§] Sam Hanash,[†] Amanda Paulovich,[†] and Martin W. McIntosh*,[†]

Fred Hutchinson Cancer Research Center, Seattle, Washington, LabKey Software, Seattle, Washington, University of Michigan, Ann Arbor, Michigan

The open-source Computational Proteomics Analysis System (CPAS) contains an entire data analysis and management pipeline for Liquid Chromatography Tandem Mass Spectrometry (LC−MS/MS) proteomics, including experiment annotation, protein database searching and sequence management, and mining LC−MS/MS peptide and protein identifications. CPAS architecture and features, such as a general experiment annotation component, installation software, and data security management, make it useful for collaborative projects across geographical locations and for proteomics laboratories without substantial computational support.

Keywords: X! Tandem • FuGE • mzXML • pepXML • adenocarcinoma • proteomics

Introduction

The Computational Proteomics Analysis System (CPAS) is an open-source, web-based analysis platform that organizes and annotates general biological experiments and provides capabilities for managing and analyzing LC−MS/MS proteomics data. CPAS's LC−MS/MS features include the X! Tandem[1] open-source search engine, management of rich protein sequence annotations, and an analysis module for mining LC−MS/MS search results. Although CPAS is distributed with X! Tandem, it can be used with other search engines that support the pepXML file format,[2] including Mascot[3] and SEQUEST.[4] CPAS LC−MS/MS architecture is flexible so individual laboratories can customize the LC−MS/MS data pipeline, for instance, by substituting their own local search engines while still using CPAS to submit their searches, annotate their experiments, and mine their data.

In addition to managing the LC−MS/MS experiments of individual laboratories, CPAS was also developed to serve as the basis of cooperative research projects, following the example of the Human Genome Project, which showed that complex, large-scale research efforts can be advanced more quickly by creating shared resources. Implementation of data sharing and management for proteomics is highly challenging given the volume of data and the complex and varied experimental procedures. Several public resources exist for proteomics that follow this model, including the PRIDE (www.ebi.ac.uk/pride),[5] PeptideAtlas (www.peptideatlas.org),[6] and GPM (www.thegpm.org)[7] databases. These systems promote data sharing and integration by serving as repositories for published LC−MS/MS data. For example, PRIDE serves as the repository for all data derived from published research studies and from all of the Human Proteome Organization (HUPO) initiatives. PeptideAtlas and the GPM curate high quality peptide identifications, process all their data in a common manner (X! Tandem for GPM and PeptideProphet[4] for PeptideAtlas), and redisplay the integrated data for mining by the community.

CPAS was designed to complement these existing repositories by creating resources for managing and sharing data *during* the proteomics discovery process, prior to submitting data to the larger public repositories for integration. CPAS is similar to both the Systems Biology Experiment Analysis Management System (SBEAMS)[8] and the Proteome Research Information Management Environment (PRIME)[9] in that they are all based on a relational database back-end, they all collect, store, and analyze data from proteomics experiments, and they each include web-based data mining tools. However, there are many features that make CPAS unique. These include the integration of multiple standard file formats, an Experiment Annotation component that ties together the entire experimental workflow, extensive web-based communication tools to promote collaborative projects, and, most importantly, a simple installer to support broad accessibility by the research community.

CPAS has many features that make it convenient and useful for shared proteomics research. All data and experiment

[*] To whom correspondence should be addressed: Martin W. McIntosh, PhD, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M2-B230, Seattle, WA 98109-1024. Phone: 206.667.4612, Fax: 206.667.7264, E-mail: mmcintos@fhcrc.org.

[†] Fred Hutchinson Cancer Research Center.

[‡] LabKey Software.

[§] University of Michigan.

[||] Listed in alphabetic order.

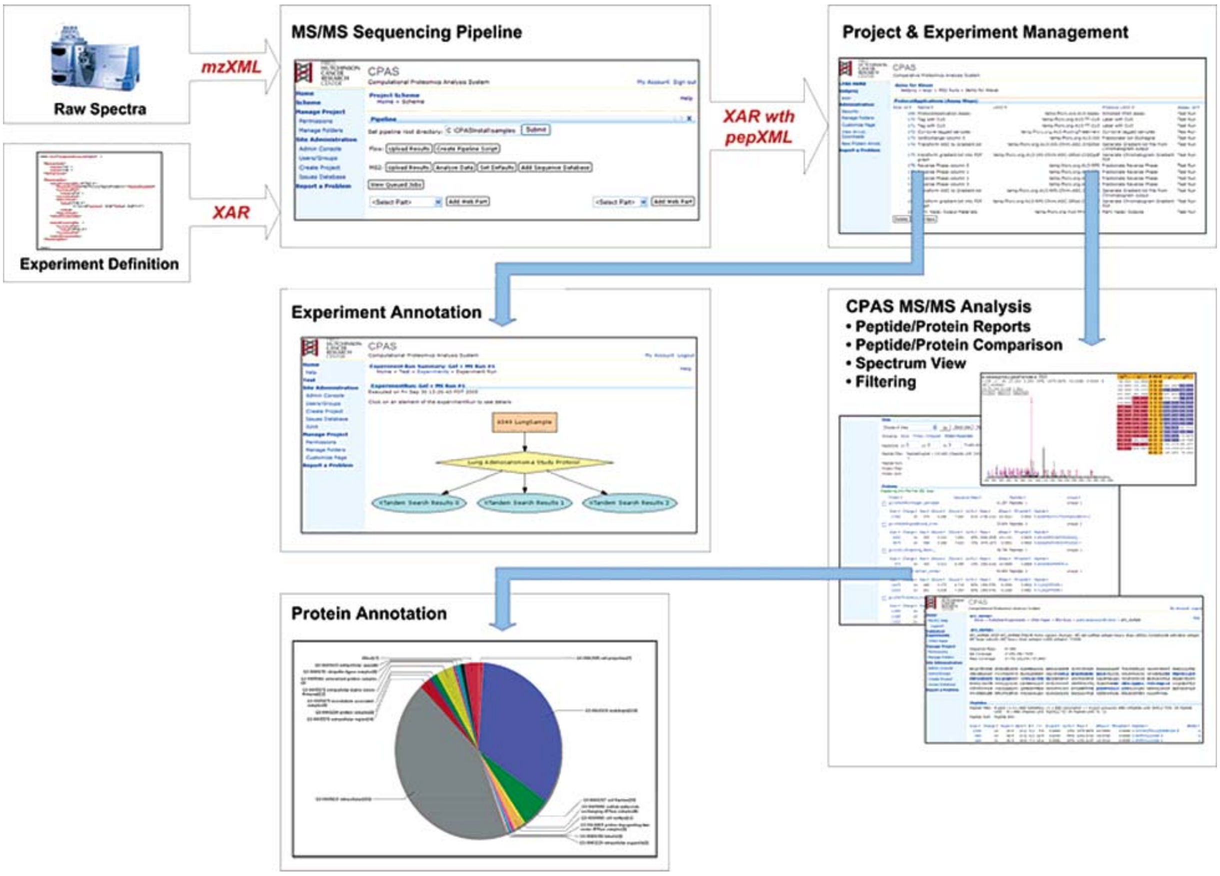

Raw Spectra
mzXML
Experiment Definition
XAR
MS/MS Sequencing Pipeline
CPAS
Computational Proteomics Analysis System
XAR wth pepXML
Project & Experiment Management
Experiment Annotation
CPAS MS/MS Analysis
• Peptide/Protein Reports
• Peptide/Protein Comparison
• Spectrum View
• Filtering
Protein Annotation

Figure 1. Typical MS/MS workflow for a user of the CPAS system.

annotations in CPAS are accessed through a web interface, and a security model allows data to be stored and analyzed securely as well as published to a wider audience; each investigator may publish his or her data to other individuals or to the public by changing access permissions. At the core of CPAS is the Experiment Annotation component, based on the proposed FuGE standard,[10] which allows researchers to carefully document and communicate their entire experimental protocols and include all associated data. Moreover, the analysis module for LC−MS/MS data facilitates data evaluation and interpretation by allowing researchers to publish their protein and peptide identification criteria as well as their results. Visitors can then alter those criteria and reevaluate the experiment. Thus, CPAS provides a flexible approach to publishing and communicating experimental results.

To facilitate wider adoption, CPAS has been designed to be platform- and database-independent (e.g., Windows, Linux, Macintosh, PostgreSQL, MS SQL Server). It is distributed under a liberal open-source license (Apache 2.0) to promote unrestricted community development. Another key feature of CPAS is its integration of multiple standardized file formats for raw mass spectra (mzXML[11]), database search results (pepXML), and an Experiment Annotation component (FuGE). Moreover, to make it more accessible to laboratories that do not have large informatics staff or expertise to devote to an open-source project, we provide a simple installer and user documentation. Installation provides the complete data pipeline including search engine, database server, web services, and all other core CPAS components, including workflow management for LC−MS/MS analysis.

This paper describes the overall design and methods of the CPAS implementation and demonstrates some of the system's capabilities as applied to a LC−MS/MS-based proteomics experiment. All data and Supporting Information described in this paper may be accessed, and in many cases reanalyzed, through the CPAS installation at the Fred Hutchinson Cancer Research Center's (FHCRC) Computational Proteomics Laboratory (http://proteomics.fhcrc.org/CPAS). The CPAS source code, installer, and user documentation are available at http://cpas.fhcrc.org.

Experimental Section

We provide a brief overview here of the CPAS LC−MS/MS data workflow (see Figure 1) and the system's capabilities. More detailed descriptions are provided below. Functions to manage all phases of data acquisition and analysis are accessible through web interfaces. The user begins at the Data Pipeline module, which combines raw mass spectra (in mzXML format) and experimental descriptions in an eXperiment ARchive (XAR). The XAR includes a manifest of all the archive contents, including all experiment annotations and results. If the user requests database searching, the Data Pipeline will submit the mzXML data to the search algorithm and manage the specification of search parameters and FASTA files. Users can monitor the progress of their searches via the web interface. Once uploaded, the CPAS system offers graphical and tabular views of the experimental steps and their inputs and output. MS/MS search results can be accessed through the Experiment Annotation component and evaluated using the MS/MS analytic module, which allows proteins and peptides to be sorted and filtered by various criteria and individual fragment ion assignments to be examined. Integrated protein annotations, which are automatically linked following parsing of the FASTA file, allow access to a variety of up-to-date external sources as well as synthesis of higher-level views of collections of protein identifications, such as the frequency of occurrence of certain Gene Ontology (GO) terms.

CPAS Core Components. CPAS consists of a core system of services that provide underlying system functionality. Modules, which provide most data handling and analytical support (such as LC−MS/MS mining), plug into the core. This design means the platform is easily extensible: the architecture allows new analytic modules to be added and integrated without having to modify the core system. All analysis modules developed for CPAS have access to the following core components.

Experiment Annotation Component and Data Sharing. The Experiment Annotation component allows laboratories to track and organize biological experiments and view the workflow of those experiments in both textual and graphical formats. The graphical interface displays summary information about the materials, protocols, and data items in the experiment. Users can select (i.e., click) an item—for example, a gel image or FASTA file—to view more detailed information about the data element. When a LC−MS/MS experiment result is selected the user will invoke the LC−MS/MS analysis module.

The CPAS Experiment Annotation descriptors are based on the FuGE object model standard.[10] A manifest file (with a .xar.xml extension) stores the information describing the experiments, including materials, protocols, and types of data involved in the experiment. Files produced in the process of running assays (e.g., raw MS data represented in an mzXML file) and the results of data analysis procedures (e.g., a pepXML file from a search engine result) are stored in external files pointed to by this manifest file. These external files can be located on the local file server or referenced and accessed over the web. The experiment files and constituent data elements may be placed together in a XAR and loaded into CPAS or to any other compatible system.

Sample Management. CPAS helps manage information about biological samples via the Sample Management service, which stores information about any biological sample type. This includes terms describing the sample or the individual that supplied it. By integrating Sample Management with the Experiment Annotation component, the sample information can be added automatically and samples can be linked to the data generated from assays performed on them.

Protein Services. Protein Services manage protein sequence annotations and databases to help investigators cope with the ever-accelerating growth of new information about proteins and their properties. Sequence annotations are automatically updated; however, updates to the system are stored incrementally so that any previous version of a database annotation can be retrieved at any time. Protein Services interact closely with the MS/MS analysis module (see below) to allow users to view up-to-date descriptions of protein sequences they have identified. Protein annotations may be loaded from richly annotated sources such as UniProt XML files[12,13] or from the FASTA files used in the analysis of the MS/MS runs.

Project Management. CPAS provides services to manage research projects, including a security mechanism that allows only those individuals with proper permission to access specific data elements. The CPAS organizational and security structure accommodates multiple unrelated projects, such as different projects within a laboratory, different laboratories within an institute, or different investigators across institutes. Data in the system is stored in projects and subfolders. One or more administrators can assign permissions to groups of users for each project folder. This gives project administrators flexible control over which users have read or write permissions to data, meeting notes, and other materials.

CPAS also provides tools for project communication in order to facilitate collaboration. The tools include: (a) a Message Board for posting announcements, meeting minutes, presentation slides, and documentation; (b) a Contacts List that stores contact information for all users with permission to the project; (c) an Issues tracking system, modeled on the software development process, for managing tasks, decisions, and workflow; and (d) Wikis for posting formatted text and links.

CPAS LC−MS/MS Data Analytic Module. A key design element of CPAS is the ability to generate analytic modules that plug into and use the core CPAS services and act on the data elements stored as part of the Experiment Annotation component. CPAS currently has an analytic module for LC−MS/MS proteomics data. The MS/MS analytic module stores, shares, analyzes, mines, and publishes high-throughput tandem MS data. Users can mine data using a variety of search engines because CPAS supports pepXML, a file format for representing MS/MS analysis results that has converters/exporters available for SEQUEST, X! Tandem, and Mascot (http://sourceforge.net/projects/sashimi). The module also supports custom analytic fields such as a field to store PeptideProphet[4] analysis results or X! Tandem Expectation values. Users can examine individual LC−MS/MS runs and groups of runs using complex customizable analytic filters for peptides and proteins. These filters can be named and saved for later use, and they can be applied across runs for consistent analysis. They can also be marked "shared" so other CPAS users can apply the same named filters to other data in the folder simply by choosing the filter's name from a list. The module interacts with Protein Services (described above) to display rich annotations for putative protein identifications, and it also links back to the raw data (mzXML) stored on the file server. Users can export data and results to other formats including Excel, TSV, PKL, and DTA for additional analysis using other tools.

Data Pipeline. The Data Pipeline module supports data submission and workflow control. It is the conduit through which users enter LC−MS/MS data and experimental descriptions into CPAS. An administrator may specify a pipeline root location on a disk, which can be shared with other project members using network file sharing or via FTP. Search results already processed externally and represented in the pepXML format may be loaded here, and new analyses can be run internally using X! Tandem. When using X! Tandem, the interface allows users to specify a set of mzXML files and search criteria. In all configurations, whether the search is done externally or internally, users may specify experimental descriptions and information to associate samples with results. Following processing, the annotations and results are loaded automatically into CPAS for viewing through the Experiment Annotation component.

Software Architecture and Implementation. CPAS was designed to be cross-platform, easily scalable, maintainable, and completely open source. To meet these requirements, we made extensive use of Java from Sun Microsystems along with other well-supported, broadly used, stable, open-source subcomponents including: Tomcat, the standard Java web server;

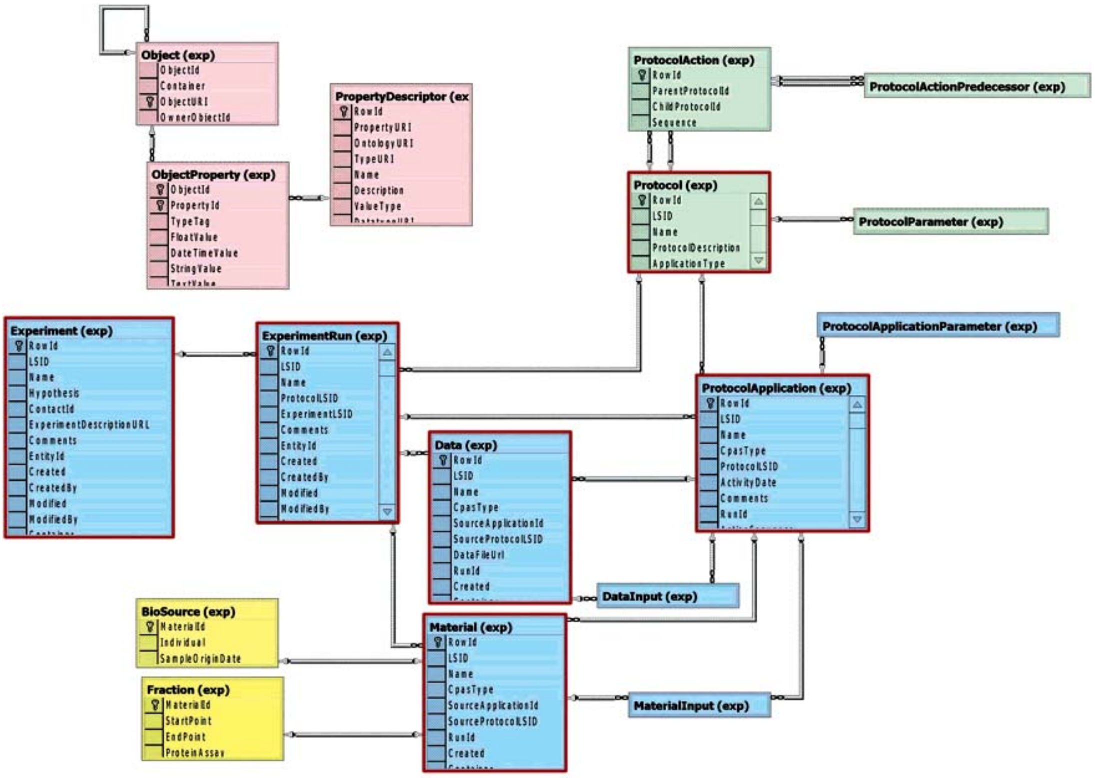


Figure 2. CPAS Database Schema for Experiment Annotation. These tables implement the core of the FuGE proposed standard for describing experiments. The four "Protocols" tables (green) define protocols as an ordered sequence of steps. The "Experiment Run" tables (blue) record series of protocols acting on material or data inputs and producing material or data outputs. Data items in the tables outlined in red are identified as globally unique Life Sciences Identifiers (LSID).[14] Detailed information about each experimental procedure is attached to these main data items via properties defined in ontologies and stored in the "Properties" tables (red). The "Extend Types" tables (yellow) store additional information specific to particular types of data and materials.

Struts, a standard web application framework; and Jakarta Commons, a standard low-level utility library. CPAS uses an object-oriented design that enables teams of developers to subdivide and work on functional elements of the system relatively independently.

CPAS is implemented as a Tomcat web application and is deployed as a single exploded WAR (Web ARchive) directory. CPAS requires access to a relational database with which it communicates though an abstraction layer that isolates the system from subtle differences between database implementations. For each supported database, a small dialect translation layer must be written. We have implemented dialects for Microsoft SQL Server and PostgreSQL; support for Oracle and MySQL is under development. See the Supporting Information for more detailed information on CPAS's architecture.

Security. CPAS users are authenticated against an LDAP provider, such as the institution's network name server, if one exists, or against a list of known user accounts and encrypted password pairs stored in the CPAS database. Experimental data and other materials are stored in projects and their sub-folders, much like a file system. Each project has one or more groups of users associated with it, and each group can have a distinct set of permissions (e.g., read only, read and write) to each of the project's folders. Project administrators manage both group membership and the permissions those groups have to the data stored within their project and its subfolders. When users log in, the authorization system determines what data they have permission to view, edit, and/or delete and provides access accordingly.

All administration functions for managing users, groups, projects, folders, and permissions are available through the web-based user interface. The person who installs CPAS at their site becomes the fist site administrator and has the authority to invite new users into the system. New users receive an email with a link back to the system where they set their own password (if they are not on the organization's LDAP domain) and log into the system for the first time. Site administrators can promote other users to become site administrators as well.

Database Schema. Figure 2 summarizes the database schema for the Experiment Annotation component and highlights major components including Protocols, Experimental Run, Properties, and Extended Types. See Technical Notes in the Supporting Information for additional database schemas.

The "Protocols" set of tables (shown in green in Figure 2) stores definitions of experimental procedures. A protocol can be atomic or can consist of nested action steps that are themselves protocols. CPAS currently supports only one level of nesting: a parent protocol describing an "ExperimentRun" and the child protocol steps within that run. Child protocol steps ("ProtocolActions") have an ordering described in the ProtocolActionPredecessor table, the structure of which allows for both branching out (e.g., fractionation) and branching in (e.g., pooling). Protocols are defined at the folder level and can be shared by all ExperimentRuns within that folder.

The "Experiment Run" set of tables (shown in blue in Figure 2) stores a record of the execution of a protocol, acting on specific material and/or data inputs and producing specific material and/or data outputs. An ExperimentRun can be understood as a cycle: Material and/or Data (e.g., tissue sample, raw LC−MS/MS data file) are input; a ProtocolApplication acts on the input and produces material and/or data outputs; these outputs usually become inputs into the next ProtocolAction step in the ExperimentRun, and the cycle continues. A ProtocolApplication is best understood as the instance of a specific Protocol that is a ProtocolAction step within the overall run. The inputs, parameters, and outputs of the ProtocolApplcation are all specific to the instance. One ProtocolAction step may correspond to multiple ProtocolApplications within the run, corresponding to running the same experimental procedure on different inputs or applying different parameter values. The Experiment object allows the grouping of multiple runs for the purpose of comparing and/or publishing results. CPAS currently allows for a run to belong to only one experiment.

The "Properties" tables (shown in red in Figure 2) allow researchers to attach custom properties (i.e., annotations) to the primary objects in an experiment. The three-table structure is designed to store: (a) unique, reusable property identifiers and their labels (PropertyDescriptors); (b) the specific objects to which these properties apply (Objects), such as a Protocol or Data object, or a defined sub-group of properties; and (c) individual property values (ObjectProperties). The bold red lines in Figure 2 are drawn around the top-level objects that may have properties attached to them.

The "Extended Types" tables (shown in yellow in Figure 2) reduce the burden of storing all the useful information about Material or Data as rows in the ObjectProperties table. CPAS allows for the addition of extended types that are implemented as separate tables joined to the base object table. CPAS currently makes only minimal use of extended types as a proof of concept. The BioSource table, for example, can be used to store sample information. A nonkey column in an extended type table is the equivalent of a record in the ObjectProperty table. While extended types can be more efficient or easier to query than generic properties, they are also more difficult to populate and maintain.

Protein Services Design. CPAS maintains protein sequence information in a schema that helps keep queries fast and flexible (see Technical Notes in the Supporting Information). Rather than trying to maintain a local copy of all information relevant to each protein sequence, we store "identifiers" (i.e., accession codes to be inserted into URLs) that point to external web-based information sources. We also store small text "annotations" such as GO descriptions, descriptions of functional or structural regions within the sequence, and information about associated diseases and biological pathways. While the identifiers serve as links to external databases and web pages, the annotations are human readable and database-searchable. The system also supports entry of local protein sequences, local annotations, and pointers to local databases. A sequence or annotation marked as "defunct" will not automatically be deleted from the database, which means old FASTA files can be reanalyzed with new annotations even if their records have been deleted or replaced by subsequent information in the primary source. We use only the formal Linnean genus and species to determine sequence uniqueness; information associating a sequence to taxa lower than species is inserted into the database as an annotation. When an MS/MS analysis is loaded into CPAS, the associated FASTA file is examined. If the file has not already been loaded, its sequences and annotations are parsed, and sequences that have not yet been added to the local database are entered. The FASTA header line is also parsed and new annotations are added to the CPAS database.

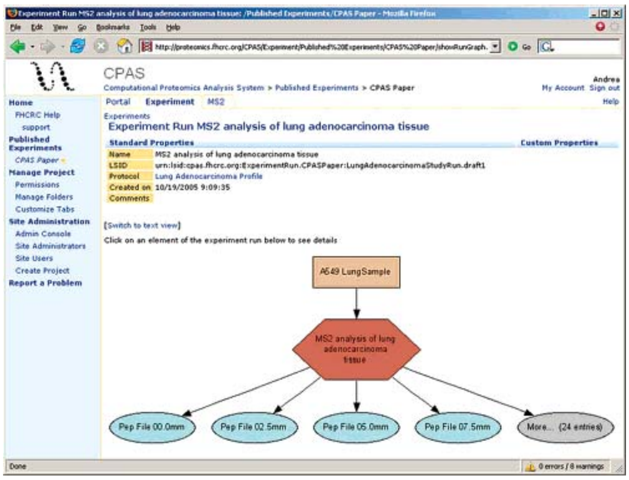
Proteomics Experiment Analysis Using CPAS. To illustrate the use of CPAS for MS-based proteomics research, we describe here an experiment for profiling membrane surface proteins of the lung adenocarcinoma A549 cell line using an intact protein separation method. This experimental measurement was part of a larger and more comprehensive set of experimental procedures for evaluating the A549 lung cancer cell surface.[15,16] Briefly, lung adenocarcinoma A549 cells were cultured in 150 × 25 mm dishes as adherent monolayers in DMEM media supplemented with 10% FBS and 1% penicillin/streptomycin, at 37 °C in a 6% CO_2 humidified incubator. Each cell dish was biotinylated with 10 mL of Sulfo-NHS-LC-biotin 0.25 mg/mL in D-PBS at room temperature for 10 min. The biotinylated proteins extracted from 3×10^8 cells (32 dishes) were isolated using an UltraLink Immobilized Monomeric Avidin affinity column.[17] Eluted proteins were run in a precast Ready Gel (8.6 × 6.8 cm, 12%) using Tris-glycine buffer system. Gels were stained with Gel Code blue solution. A duplicate gel was electroblotted to PVDF to visualize biotinylated proteins. Two lanes loaded with 20 μg of biotinylated proteins were divided into 28 2.5 mm-wide slices and submitted to in situ digestion. The extracted peptides were analyzed by LC−MS/MS in an LTQ-FTICR mass spectrometer (ThermoFinnigan) using data-dependent acquisition.

Data acquired from the LTQ-FTICR were converted to the mzXML data format and submitted to the LC−MS/MS Data Pipeline described above. The pipeline used X! Tandem to search the spectra against a human subset of the NCI's nonredundant protein sequence database, release 20040928, downloaded from the NCI's Advanced Biomedical Computing Center (ftp://ftp.ncifcrf.gov/pub/nonredun/protein.nrdb.Z). Results were uploaded automatically to the CPAS site as part of a XAR.

Results and Discussion

We illustrate here the use of CPAS to evaluate the lung adenocarcinoma experiment, emphasizing LC−MS/MS analysis. Except where indicated, all screen shots can be reproduced by visiting the Published Experiments area of the CPAS website (http://proteomics.fhcrc.org/CPAS/Project/Published%20Experiments/begin.view?). Data may be interactively mined at that site as well. Additional screen shots of the analysis can be found in Supporting Information. (Note: CPAS currently supports Internet Explorer and Firefox browsers on Windows and the Firefox browser on Macintosh.)

Accessing and Navigating Experiments. Experiment descriptions are accessible via the "Experimental Navigator" section of any folder. To access our example experiment, we used the list of projects on the left side of the CPAS home page (http://proteomics.fhcrc.org/CPAS/) to navigate to the "CPAS Paper" subfolder of the "Published Experiments" project. The initial view presented the experiments and MS/MS run measurements stored in this folder as well as a message board with Supporting Information. After choosing Lung Adenocarcinoma Study from the "Experiment Navigator" list, we selected the "MS2 analysis of lung..." Experiment Run to access a graphical representation of the experiment workflow (Figure 3). Each block in the graphical flowchart links to more detailed information.



Experiment Run MS2 analysis of lung adenocarcinoma tissue

Standard Properties		Custom Properties
Name	MS2 analysis of lung adenocarcinoma tissue	
LSID	urn:lsid:cpas.fhcrc.org:ExperimentRun.CPASPaper:LungAdenocarcinomaStudyRun.draft1	
Protocol	Lung Adenocarcinoma Profile	
Created on	10/19/2005 9:09:35	
Comments		

[Switch to text view]

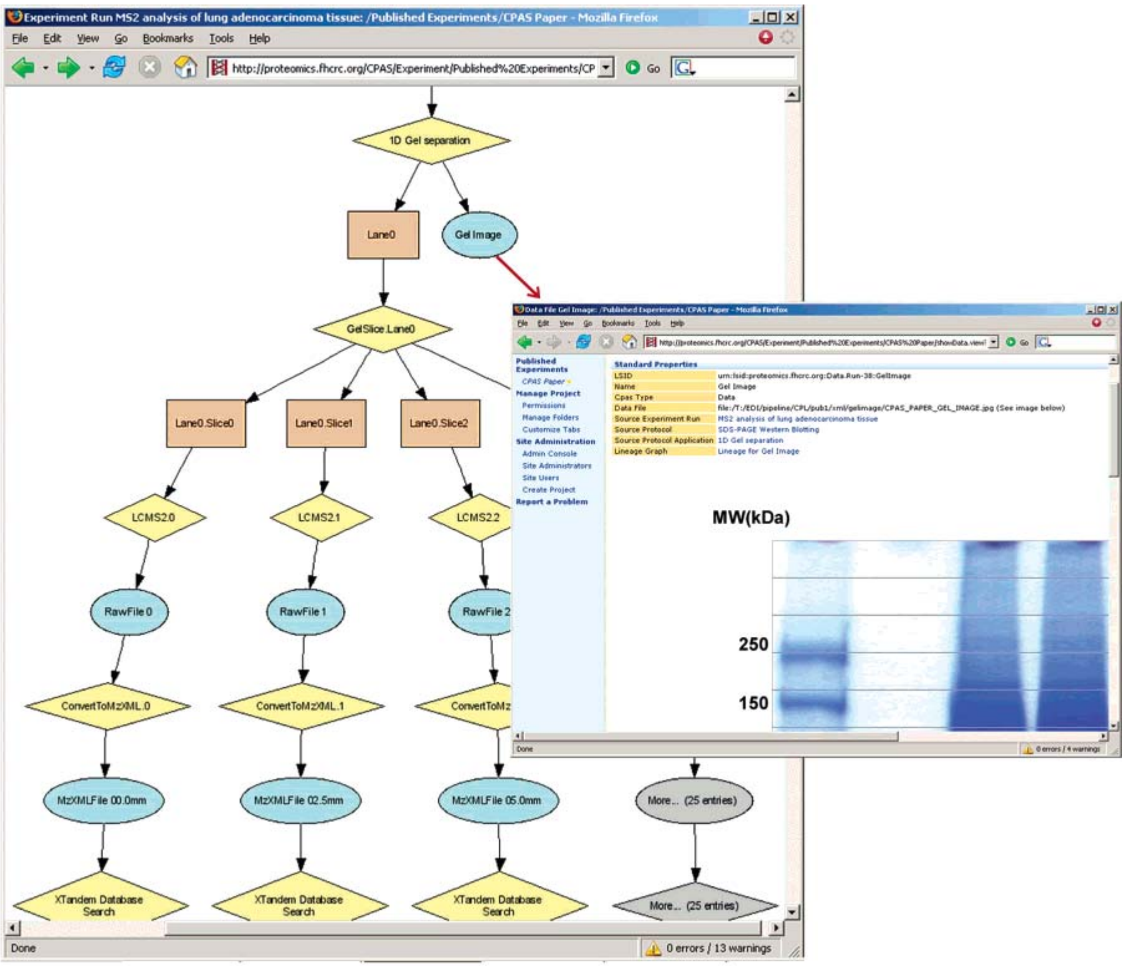
Click on an element of the experiment run below to see details

Figure 3. Graphical representation of the overall Lung A549 experiment workflow.

We selected the MS2 analysis of lung adenocarcinoma tissue block to view a more detailed flowchart of the protocol steps in this experiment (Figure 4). We inspected an image of the original gel by selecting the "Gel Image" oval. The gel image, which was loaded as part of the original XAR, gives an example of what occurs when a user accesses a data type for which no analysis module exists: CPAS provides the data back to the user in its original format (e.g., image, Excel file, URL).

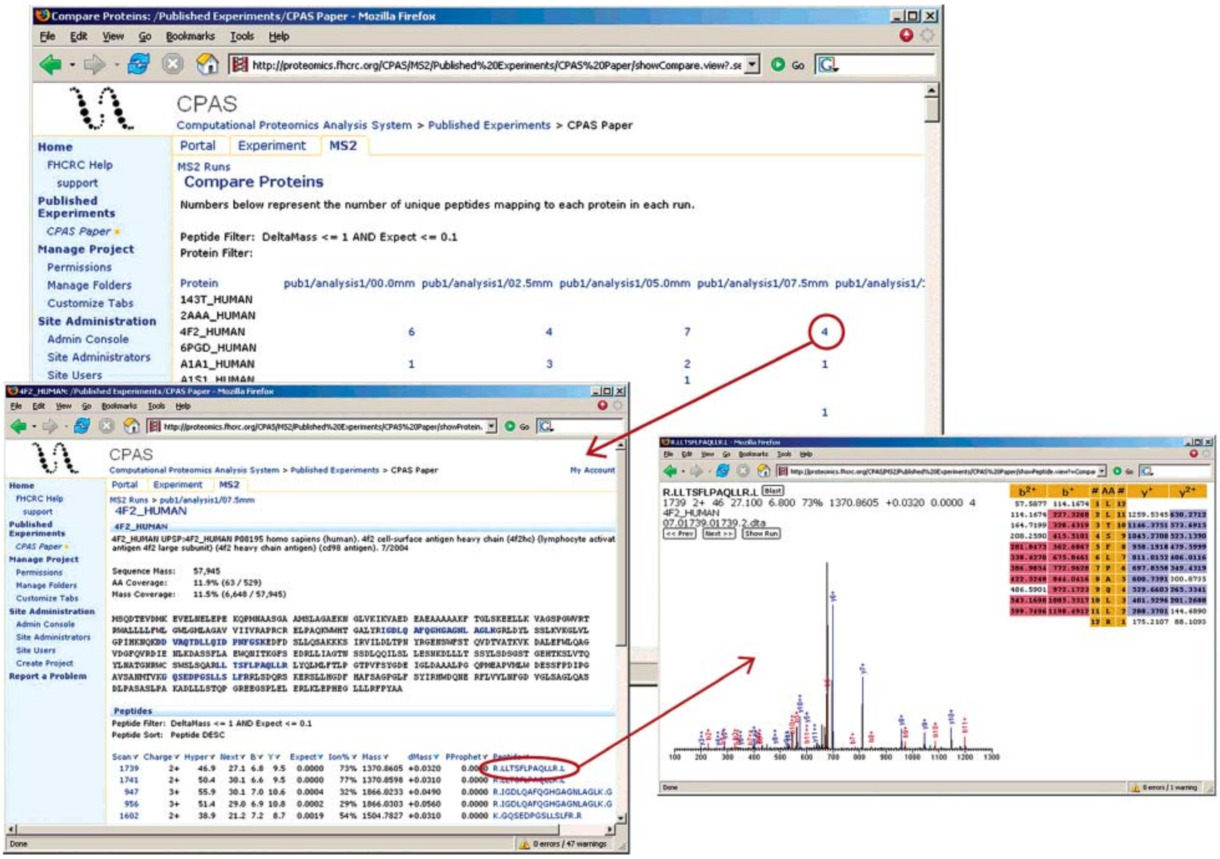
Accessing Experiment Results. Access to LC−MS/MS results is available via the "MS2 Runs" section of a project. For our example, we selected a single MS/MS run from the "MS2 Runs" section of the main CPAS Paper web page. From the single run page, we used the extensive protein and peptide filtering capabilities of CPAS to build a protein filter that included the criteria for determining a correct identification. This filter, named "pub1 (shared)," retains only those proteins with peptides that have Expectation values ≤ 0.1 and Delta Mass ≤ 1.0.

We then accessed "MS2 Runs" from the main CPAS Paper web page to view the set of LC−MS/MS measurements that comprised this experiment. We "selected all" 28 LC−MS/MS runs and chose "Compare Proteins," applying the "pub1 (shared)" filter. This produced a list of 1578 peptides representing 727 different sequences across all runs. The compare proteins results page lists all proteins identified across all runs that passed the selected filtering criteria (Figure 5). The numbers in the grid indicate the number of peptides found for that protein sequence (row) in the corresponding LC−MS/MS run (column). Each number links to a web page for the protein sequence and corresponding peptides found in that run. This page allows detailed examination of the findings, including individual spectra, peptide sequences, and sequence annotations (not shown here).



Standard Properties	
LSID	urn:lsid:proteomics.fhcrc.org:Data.Run-38:GelImage
Name	Gel Image
Cpas Type	Data
Data File	file:/T:/EDI/pipeline/CPL/pub1/xml/gelimage/CPAS_PAPER_GEL_IMAGE.jpg (See image below)
Source Experiment Run	MS2 analysis of lung adenocarcinoma tissue
Source Protocol	SDS-PAGE Western Blotting
Source Protocol Application	1D Gel separation
Lineage Graph	Lineage for Gel Image

Figure 4. Detailed graphical view of the experiment protocol that results following selection of the Lung Adenocarcinoma Study Protocol link from Figure 3. Each block links to data elements such gel images (right) and sample data (not shown).



Compare Proteins: /Published Experiments/CPAS Paper - Mozilla Firefox

http://proteomics.fhcrc.org/CPAS/MS2/Published%20Experiments/CPAS%20Paper/showCompare.view?.se

CPAS

Computational Proteomics Analysis System > Published Experiments > CPAS Paper

Home
FHCRC Help
support
Published Experiments
CPAS Paper
Manage Project
Permissions
Manage Folders
Customize Tabs
Site Administration
Admin Console
Site Administrators
Site Users

Portal | Experiment | MS2

MS2 Runs

Compare Proteins

Numbers below represent the number of unique peptides mapping to each protein in each run.

Peptide Filter: DeltaMass <= 1 AND Expect <= 0.1
Protein Filter:

Protein	pub1/analysis1/00.0mm	pub1/analysis1/02.5mm	pub1/analysis1/05.0mm	pub1/analysis1/07.5mm	pub1/analysis1/:
143T_HUMAN					
2AAA_HUMAN					
4F2_HUMAN	6	4	7	4	
6PGD_HUMAN					
A1A1_HUMAN	1	3	2	1	
A1S1_HUMAN			1		
				1	

4F2_HUMAN: /Published Experiments/CPAS Paper - Mozilla Firefox

http://proteomics.fhcrc.org/CPAS/MS2/Published%20Experiments/CPAS%20Paper/showProtein.

CPAS

Computational Proteomics Analysis System > Published Experiments > CPAS Paper

My Account

Home
FHCRC Help
support
Published Experiments
CPAS Paper
Manage Project
Permissions
Manage Folders
Customize Tabs
Site Administration
Admin Console
Site Administrators
Site Users
Create Project
Report a Problem

Portal | Experiment | MS2

MS2 Runs > pub1/analysis1/07.5mm

4F2_HUMAN

4F2_HUMAN

4F2_HUMAN UPSP:4F2_HUMAN P08195 homo sapiens (human). 4f2 cell-surface antigen heavy chain (4f2hc) (lymphocyte activat antigen 4f2 large subunit) (4f2 heavy chain antigen) (cd98 antigen). 7/2004

Sequence Mass: 57,945
AA Coverage: 11.9% (63 / 529)
Mass Coverage: 11.5% (6,648 / 57,945)

```
MSQDTEVDMK EVELNELEPE KQPMNAASGA AMSLAGAEKN GLVKIKVAED EAEAAAAAKF TGLSKEELLK VAGSPGWVRT
RWALLLLFWL GWLGMLAGAV VIIVRAPRCR ELPAQKWWHT GALYRIGDLQ AFQGHGAGNL AGLKGRLDYL SSLKVKGLVL
GPIHKNQKDD VAQTDLLQID PNFGSKEDFD SLLQSAKKKS IRVILDLTPN YRGENSWFST QVDTVATKVK DALEFWLQAG
VDGFQVRDIE NLKDASSFLA EWQNITKGFS EDRLLIAGTN SSDLQQILSL LESNKDLLLT SSYLSDSGST GEHTKSLVTQ
YLNATGNRWC SWSLSQARLL TSFLPAQLLR LYQLMLFTLP GTPVFSYGDE IGLDAAALPG QPMEAPVMLW DESSFPDIPG
AVSANMTVKG QSEDPGSLLS LFRRLSDQRS KERSLLHGDF HAFSAGPGLF SYIRHWDQNE RFLVVLNFGD VGLSAGLQAS
DLPASASLPA KADLLLSTQP GREEGSPLEL ERLKLEPHEG LLLRFPYAA
```

Peptides

Peptide Filter: DeltaMass <= 1 AND Expect <= 0.1
Peptide Sort: Peptide DESC

Scan	Charge	Hyper	Next	B	Y	Expect	Ion%	Mass	dMass	PProphet	Peptide
1739	2+	46.9	27.1	6.8	9.5	0.0000	73%	1370.8605	+0.0320	0.0000	R.LLTSFLPAQLLR.L
1741	2+	50.4	30.1	6.6	9.5	0.0000	77%	1370.8598	+0.0310	0.0000	R.LLTSFLPAQLLR.L
947	3+	55.9	30.1	7.0	10.6	0.0004	32%	1866.0233	+0.0490	0.0000	R.IGDLQAFQGHGAGNLAGLK.G
956	3+	51.4	29.0	6.9	10.8	0.0002	29%	1866.0303	+0.0560	0.0000	R.IGDLQAFQGHGAGNLAGLK.G
1602	2+	38.9	21.2	7.2	8.7	0.0019	54%	1504.7827	+0.0310	0.0000	K.GQSEDPGSLLSLFR.R

Done | 0 errors / 47 warnings

R.LLTSFLPAQLLR.L - Mozilla Firefox

http://proteomics.fhcrc.org/CPAS/MS2/Published%20Experiments/CPAS%20Paper/showPeptide.view?+Compa

R.LLTSFLPAQLLR.L Blast
1739 2+ 46 27.100 6.800 73% 1370.8605 +0.0320 0.0000 4
4F2_HUMAN
07.01739.01739.2.dta
<< Prev | Next >> | Show Run

b^{2+}	b^{+}	#	AA	#	y^{+}	y^{2+}
57.5877	114.1674	1	L	12		
114.1674	227.3268	2	L	11	1259.5345	630.2712
164.7199	328.4319	3	T	10	1146.3751	573.6915
208.2590	415.5101	4	S	9	1045.2700	523.1390
281.8473	562.6867	5	F	8	958.1918	479.5999
338.4270	675.8461	6	L	7	811.0152	406.0116
386.9854	772.9628	7	P	6	697.8558	349.4319
422.5248	844.0416	8	A	5	600.7391	300.8735
486.5901	972.1723	9	Q	4	529.6603	265.3341
543.1698	1085.3317	10	L	3	401.5296	201.2688
599.7496	1198.4912	11	L	2	288.3701	144.6890
		12	R	1	175.2107	88.1093

Done | 0 errors / 1 warning

Figure 5. Comparing Multiple MS/MS Runs — Proteins and peptides from two or more runs, either from the same or different experiments, can be compared. The results can be exported to Excel for distribution and further analysis. Here proteins from all 28 runs in the experiment were compared using the "pub1" filter. Numbers in the grid link to more information on identified protein sequences (bottom left), including links to individual peptide fragment spectra (bottom right).

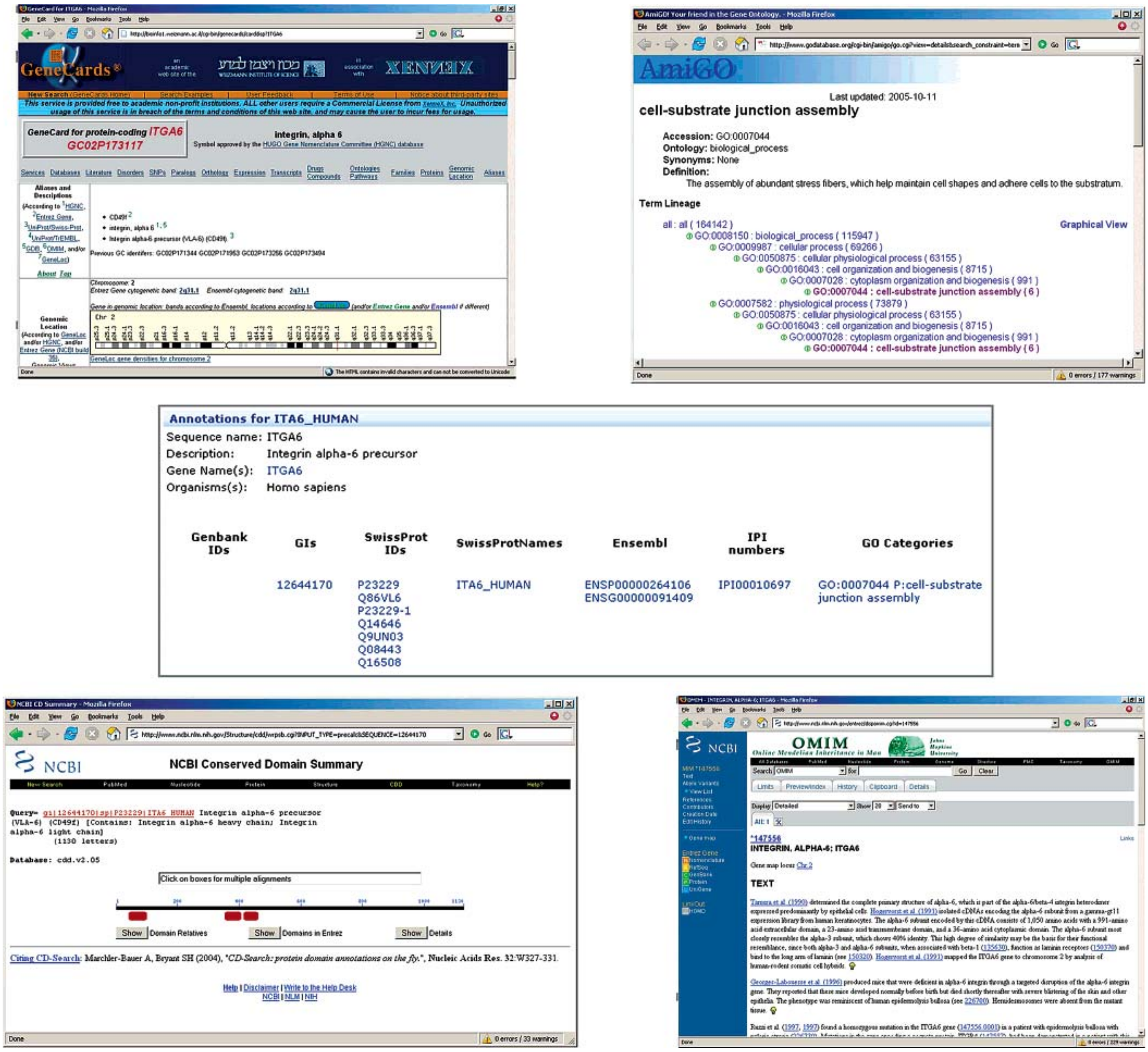
To confirm the overall experimental protocol, which attempted to enrich for membrane bound proteins, we categorized the 727 identified protein sequences by their GO cellular component categories.[18] To determine the categories, CPAS first looked in Protein Services to determine the set of all GO cellular component annotations. Since this list is too fine-grained, CPAS looked in its local copy of the GO database to find a parent category three steps from the root (the 3rd level classifications). Here it found 186 classifications for "cell membrane". This summary suggests that the overall experimental design of enriching for this class of proteins succeeded.

Interrogating Findings Using Protein Services. One advantage of this experimental design was the ability to identify proteins that may exist in the sample in two native forms or molecular weights. Because each LC−MS/MS measurement corresponds to a 2.5 mm-wide strip cut from two lanes of the 1D SDS gel, peptides found in runs from nonadjacent gel slices that are matched to the same protein sequence could indicate such a protein. The molecular weight (MW) range for each gel slice was estimated based on the linear regression plot of the log MW of the molecular weight standards versus migration distance. To better fit this wide range of data points, linear regressions were calculated separately for two regions of the plot: 15−75 kDa and 75−250 kDa (regression and fit was performed outside of CPAS).

We can look for such proteins in the protein list either online using the compare proteins page (Figure 5) or by exporting the list to Excel. Visual inspection of the list after sorting by protein sequence name revealed many protein sequences identified in multiple contiguous fractions; however, we also observed some proteins in discontiguous gel strips corresponding to large differences in molecular weight. Among such proteins are several members of the Integrin alpha family, each found at two distinct molecular weights, including ITGA3, ITGAV, and ITGA6.

Using the links provided by Protein Services (Figure 6) we confirmed that Integrins are integral cell-surface proteins, known to be involved in signaling and cell-adhesion. By clicking the protein name, we accessed links to GenBank, GeneCards, NiceProt, Ensembl, and other major information providers. According to GeneCards, the alpha subunits form a family with possibly more than 500 members, several of which, including ITGA3 and ITGAV, are known to be cleaved, post-translationally, and then re-linked with a disulfide bridge. The Swiss-Prot entry for ITGA6 (*P*23229) suggests that the full transcript may be post-translationally cleaved into a heavy chain and a light chain, but there has been no direct experimental evidence supporting the existence of the light chain.

We next evaluated whether our findings were consistent with the existence of both chains. Using the LC−MS/MS mining capability, we examined the amino acid coverage for ITGA6 by the peptides in the strip beginning at 12.5 mm and by the peptides in the strip beginning at 42.5 mm. Although it does not provide conclusive proof of their existence, the sequence shown in CPAS for ITA6_Human (black) includes both the hypothesized heavy and light chains (Figure 7). The six unique



GeneCard for protein-coding ITGA6
GC02P173117

integrin, alpha 6
Symbol approved by the HUGO Gene Nomenclature Committee (HGNC) database

This service is provided free to academic non-profit institutions. ALL other users require a Commercial License from XenneX, Inc. Unauthorized usage of this service is in breach of the terms and conditions of this web site, and may cause the user to incur fees for usage.

Aliases and Descriptions (According to [1]HGNC, [2]Entrez Gene, [3]UniProt/Swiss-Prot, [4]UniProt/TrEMBL, [5]GDB, [6]OMIM, and/or [7]GeneLoc)

- CD49f [2]
- integrin, alpha 6 [1, 5]
- Integrin alpha-6 precursor (VLA-6) (CD49f) [3]

Previous GC identifiers: GC02P171344 GC02P171953 GC02P173256 GC02P173494

Genomic Location (According to GeneLoc and/or HGNC, and/or Entrez Gene (NCBI build 35))

Chromosome: 2
Entrez Gene cytogenetic band: 2q31.1 Ensembl cytogenetic band: 2q31.1

GeneLoc gene densities for chromosome 2

The HTML contains invalid characters and can not be converted to Unicode

Last updated: 2005-10-11

cell-substrate junction assembly

Accession: GO:0007044
Ontology: biological_process
Synonyms: None
Definition:
The assembly of abundant stress fibers, which help maintain cell shapes and adhere cells to the substratum.

Term Lineage

all : all (164142)
GO:0008150 : biological_process (115947)
GO:0009987 : cellular process (69266)
GO:0050875 : cellular physiological process (63155)
GO:0016043 : cell organization and biogenesis (8715)
GO:0007028 : cytoplasm organization and biogenesis (991)
GO:0007044 : cell-substrate junction assembly (6)
GO:0007582 : physiological process (73879)
GO:0050875 : cellular physiological process (63155)
GO:0016043 : cell organization and biogenesis (8715)
GO:0007028 : cytoplasm organization and biogenesis (991)
GO:0007044 : cell-substrate junction assembly (6)

0 errors / 177 warnings

Annotations for ITA6_HUMAN

Sequence name: ITGA6
Description: Integrin alpha-6 precursor
Gene Name(s): ITGA6
Organisms(s): Homo sapiens

Genbank IDs	GIs	SwissProt IDs	SwissProtNames	Ensembl	IPI numbers	GO Categories
	12644170	P23229 Q86VL6 P23229-1 Q14646 Q9UN03 Q08443 Q16508	ITA6_HUMAN	ENSP00000264106 ENSG00000091409	IPI00010697	GO:0007044 P:cell-substrate junction assembly

NCBI Conserved Domain Summary

```
Query= gi|12644170|sp|P23229|ITA6_HUMAN Integrin alpha-6 precursor
(VLA-6) (CD49f) [Contains: Integrin alpha-6 heavy chain; Integrin
alpha-6 light chain]
        (1130 letters)

Database: cdd.v2.05
```

Click on boxes for multiple alignments

Citing CD-Search: Marchler-Bauer A, Bryant SH (2004), "*CD-Search: protein domain annotations on the fly.*", Nucleic Acids Res. 32:W327-331.

0 errors / 33 warnings

*147556
INTEGRIN, ALPHA-6; ITGA6

Gene map locus Chr.2

TEXT

Tamura et al. (1990) determined the complete primary structure of alpha-6, which is part of the alpha-6/beta-4 integrin heterodimer expressed predominantly by epithelial cells. Hogervorst et al. (1991) isolated cDNAs encoding the alpha-6 subunit from a gamma-gt11 expression library from human keratinocytes. The alpha-6 subunit encoded by this cDNA consists of 1,050 amino acids with a 991-amino acid extracellular domain, a 23-amino acid transmembrane domain, and a 36-amino acid cytoplasmic domain. The alpha-6 subunit most closely resembles the alpha-3 subunit, which shows 40% identity. This high degree of similarity may be the basis for their functional resemblance, since both alpha-3 and alpha-6 subunits, when associated with beta-1 (135630), function as laminin receptors (150320) and bind to the long arm of laminin (see 150320). Hogervorst et al. (1991) mapped the ITGA6 gene to chromosome 2 by analysis of human-rodent somatic cell hybrids.

Georges-Labouesse et al. (1996) produced mice that were deficient in alpha-6 integrin through a targeted disruption of the alpha-6 integrin gene. They reported that these mice developed normally before birth but died shortly thereafter with severe blistering of the skin and other epithelia. The phenotype was reminiscent of human epidermolysis bullosa (see 226700). Hemidesmosomes were absent from the mutant tissue.

Ruzzi et al. (1997, 1997) found a homozygous mutation in the ITGA6 gene (147556.0001) in a patient with epidermolysis bullosa with

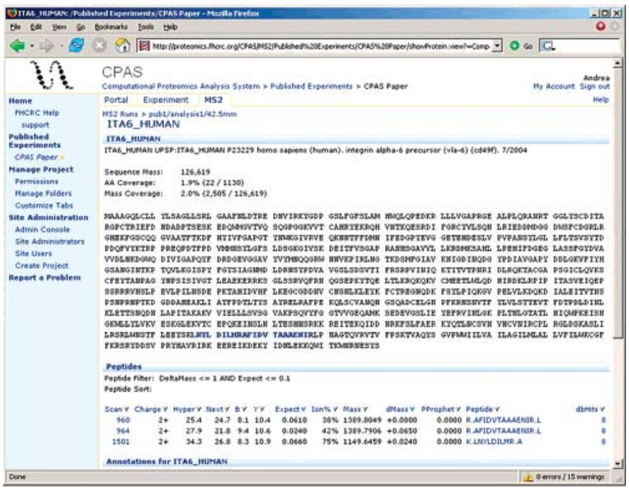
Figure 6. Protein Annotations for ITGA6 link to (clockwise from top left): Human GeneCards – summarizes many gene features, including chromosomal position; AmiGO – classifies the function, location, and reactions of the protein; OMIM – lists genetic diseases associated with different polymorphic forms; and NCBI Conserved Domain Summary – displays schematics of functional regions of the protein.

peptides identified in the "12.5 mm" run (not shown) covered the amino acid sequence corresponding to a hypothesized heavy chain (amino acids 24–938), and the two unique peptides identified in the "42.5 mm" run (shown in blue in Figure 7) covered the sequence corresponding to the hypothesized light chain (amino acids 942–1130)

Conclusion

For the experiment presented here, CPAS allowed us to effectively describe a complex proteomics experiment and analyze the LC–MS/MS measurements. Using CPAS's data mining and informatics capabilities, we investigated whether the experimental method was reasonably selective in capturing cell surface proteins, and we evaluated the ability of intact protein-based approaches to permit the identification of protein isoforms. Though not demonstrated here, a more comprehensive analysis would include use of CPAS's additional LC–MS/MS mining tools such as: (a) visual inspection of individual LC–MS/MS spectra; (b) application of different filtering criteria; (c) additional X! Tandem searches using different search parameters; or (d) additional searches using different search algorithms.

Although the CPAS distribution includes X! Tandem, implementation of CPAS in a high throughput production facility will require customization for a local environment that may use different search engines, perhaps installed on a cluster external to CPAS. Such customization, however, should not be difficult, and the Data Pipeline was designed to accommodate such modifications. For example, the CPAS installations at FHCRC and the University of Michigan connect a large and



http://proteomics.fhcrc.org/CPAS/MS2/Published%20Experiments/CPAS%20Paper/showProtein.view?=Comp

CPAS

Computational Proteomics Analysis System > Published Experiments > CPAS Paper

MS2 Runs > pub1/analysis1/42.5mm

ITA6_HUMAN

ITA6_HUMAN

ITA6_HUMAN UPSP:ITA6_HUMAN P23229 homo sapiens (human). integrin alpha-6 precursor (vla-6) (cd49f). 7/2004

Sequence Mass:	126,619
AA Coverage:	1.9% (22 / 1130)
Mass Coverage:	2.0% (2,505 / 126,619)

```
MAAAGQLCLL YLSAGLLSRL GAAFNLDTRE DNVIRKYGDP GSLFGFSLAM HWQLQPEDKR LLLVGAPRGE ALPLQRANRT GGLYSCDITA
RGPCTRIEFD NDADPTSESK EDQWMGVTVQ SQGPGGKVVT CAHRYEKRQH VNTKQESRDI FGRCYVLSQN LRIEDDMDGG DWSFCDGRLR
GHEKFGSCQQ GVAATFTKDF HYIVFGAPGT YNWKGIVRVE QKNNTFFDMN IFEDGPYEVG GETEHDESLV PVPANSYLGL LFLTSVSYTD
PDQFVYKTRP PREQPDTFPD VMMNSYLGFS LDSGKGIVSK DEITFVSGAP RANHSGAVVL LKRDMKSAHL LPEHIFDGEG LASSFGYDVA
VVDLNKDGWQ DIVIGAPQYF DRDGEVGGAV YVYMNQQGRW NNVKPIRLNG TKDSMFGIAV KNIGDINQDG YPDIAVGAPY DDLGKVFIYH
GSANGINTKP TQVLKGISPY FGYSIAGNMD LDRNSYPDVA VGSLSDSVTI FRSRPVINIQ KTITVTPNRI DLRQKTACGA PSGICLQVKS
CFEYTANPAG YNPSISIVGT LEAEKERRKS GLSSRVQFRN QGSEPKYTQE LTLKRQKQKV CMEETLWLQD NIRDKLRPIP ITASVEIQEP
SSRRRVNSLP EVLPILNSDE PKTAHIDVHF LKEGCGDDNV CNSNLKLEYK FCTREGNQDK FSYLPIQKGV PELVLKDQKD IALEITVTNS
PSNPRNPTKD GDDAHEAKLI ATFPDTLTYS AYRELRAFPE KQLSCVANQN GSQADCELGN PFKRNSNVTF YLVLSTTEVT FDTPDLDINL
KLETTSNQDN LAPITAKAKV VIELLLSVSG VAKPSQVYFG GTVVGEQAMK SEDEVGSLIE YEFRVINLGK PLTNLGTATL NIQWPKEISN
GKWLLYLVKV ESKGLEKVTC EPQKEINSLN LTESHNSRKK REITEKQIDD NRKFSLFAER KYQTLNCSVN VNCVNIRCPL RGLDSKASLI
LRSRLWNSTF LEEYSKLNYL DILMRAFIDV TAAAENIRLP NAGTQVRVTV FPSKTVAQYS GVPWWIILVA ILAGILMLAL LVFILWKCGF
FKRSRYDDSV PRYHAVRIRK EEREIKDEKY IDNLEKKQWI TKWNRNESYS
```

Peptides

Peptide Filter: DeltaMass <= 1 AND Expect <= 0.1

Peptide Sort:

Scan	Charge	Hyper	Next	B	Y	Expect	Ion%	Mass	dMass	PProphet	Peptide	dbHits
960	2+	25.4	24.7	8.1	10.4	0.0610	38%	1389.8049	+0.0800	0.0000	R.AFIDVTAAAENIR.L	8
964	2+	27.9	21.8	9.4	10.6	0.0240	42%	1389.7906	+0.0650	0.0000	R.AFIDVTAAAENIR.L	8
1501	2+	34.3	26.8	8.3	10.9	0.0660	75%	1149.6459	+0.0240	0.0000	K.LNYLDILMR.A	8

Annotations for ITA6_HUMAN

Figure 7. Protein Detail page for ITGA6 from the LC−MS/MS run of the gel strip beginning at 42.5 mm. The sequence information appears at the top of the screen, and the identified Peptides and Annotations are listed below.

growing number of mass spectrometry instruments to CPAS, with database searching performed using X! Tandem or other search algorithms located on clusters external to CPAS. Full use of the Protein Services component will also require configuration to support the download of rich protein annotations.

CPAS was developed to facilitate data sharing and exchange in cooperative biological research projects, including, but not limited to, proteomics. Because CPAS implements the FuGE object model, it includes the capacity to represent any type of experiment or data format. We introduce the XAR format to capture those experiment annotations along with constituent experimental data. CPAS can accomplish data sharing in two ways. Individual investigators can grant access permissions for online interrogation of their data, or they can simply export all data and annotations by providing a XAR. For example, the FHCRC CPAS implementation is currently using XAR formats to exchange data and annotations from a large number of proteomics consortia. Moreover, GenoLogics[19] has also adopted FuGE and the XAR format in ProteusLIMS (a commercial and proprietary laboratory information management system) and has demonstrated the feasibility of using FuGE and XAR for data sharing by exporting complex data types and experiment annotations directly into CPAS (see the “Ovarian Cancer 2” experiment on the CPAS Paper web page to view the exported experiment description).

Several extensions to the CPAS MS/MS module are needed to make online analysis more complete and suitable to a wider range of laboratories and investigators. Most obvious is that CPAS currently lacks the ability to perform relative quantitation using isotopic labeling (e.g., ICAT,[20] SILAC[21]); although, plans are underway to add that capacity. Presently, CPAS does not support the mzData file format;[22] however, representatives of the mzXML and mzData communities recently agreed to merge these formats, and further CPAS development will support that process. CPAS will also integrate other emerging standards proposed by the Proteomics Standards Initiative. In addition, as CPAS is adopted by users with additional needs, more functionality will need to be added to the user interface in order to support more complex queries within and across experiments.

While basic tools are provided to generate experiment annotations and XARs, the system would benefit most from more flexible open-source client tools for creating and editing XAR files. Open-source tools that make it more convenient to load generic data into CPAS or exchange data between FuGE-compatible systems would also enhance the system. At present, CPAS automatically generates a XAR only for those simple experiments accommodated by the Data Pipeline: simple specimen processing protocols and LC−MS/MS search parameters and data output. General tools for creating more flexible, generic experimental protocols have recently been implemented commercially,[19] but open-source tools are not yet available. To help researchers make use of CPAS’s general Experiment Annotation in the meantime, the Published Experiments site at http://proteomics.fhcrc.org/CPAS includes a folder for distributing Experiment.xml template files (Experiment Examples). Over time, new experiment annotation examples will be deposited there so CPAS users can download experimental descriptions for their own use.

Many extensions to CPAS are currently planned and underway, including the addition of new analysis modules (e.g., flow cytometry). However, the primary strategy to improve its performance is to develop an open-source development community to build on the system’s architecture and add analytic modules. By developing a rich and active open-source community CPAS will expand its capabilities by inviting the development of new modules and by integrating with other existing platforms. The ultimate goals of CPAS development will allow federation of multiple CPAS (or XAR-supporting) installations to share data and queries without the need to exchange raw data between systems.

Acknowledgment. This work was funded by NCI subcontract 23XS144A and a grant from the Canary Fund.

Supporting Information Available: Interactive analysis of the data presented here is available at http://proteomics.fhcrc.org/CPAS. Select Published Experiments from the menu on the left, and then select CPAS Paper. TechnicalNotes.ppt − set of slides describing CPAS architecture and implementation. QuickStart_LoadData.ppt − set of slides showing sample data being loaded into CPAS via the Data Pipeline. AnalysisDemo.ppt − set of slides showing screenshots of each of the analysis steps described in the Results and Discussion section. Gel image from the lung adenocarcinoma experiment. This material is available for free at http://proteomics.fhcrc.org/CPAS and at http://pubs.acs.org.

References

(1) Craig, R.; Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466−1467.

(2) Keller, A.; Eng, J. K.; Zhang, N.; Li, X. J.; Aebersold, R., A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **2005**, EPub, (August).

(3) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551−3567.

(4) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74* (20), 5383−5392.

(5) Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R., PRIDE: the proteomics identifications database. *Proteomics* **2005**, *5* (13), 3537−3545.

(6) Desiere, F.; Deutsch, E. W.; Nesvizhskii, A. I.; Mallick, P.; King, N. L.; Eng, J. K.; Aderem, A.; Boyle, R.; Brunner, E.; Donohoe, S.; Fausto, N.; Hafen, E.; Hood, L.; Katze, M. G.; Kennedy, K. A.; Kregenow, F.; Lee, H.; Lin, B.; Martin, D.; Ranish, J. A.; Rawlings, D. J.; Samelson, L. E.; Shiio, Y.; Watts, J. D.; Wollscheid, B.; Wright, M. E.; Yan, W.; Yang, L.; Yi, E. C.; Zhang, H.; Aebersold, R., Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **2005,** *6* (1), R9.

(7) Craig, R.; Cortens, J. P.; Beavis, R. C., Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **2004**, *3* (6), 1234−1242.

(8) Systems Biology Experiment Analysis System (SBEAMS). http://www.sbeams.org.

(9) Proteome Research Information Management Environment (PRIME). http://prime.proteome.med.umich.edu.

(10) Jones, A.; Hunt, E.; Wastling, J. M.; Pizarro, A.; Stoeckert, C. J., Jr., An object model and database for functional genomics. *Bioinformatics* **2004**, *20* (10), 1583−1590.

(11) Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R., A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **2004**, *22* (11), 1459−1466.

(12) Bairoch, A.; Apweiler, R.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S., The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2005**, *33*, (Database issue), D154−159.

(13) Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S., UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, (Database issue), D115−119.

(14) Clark, T.; Martin, S.; Liefeld, T., Globally distributed object identification for biological knowledgebases. *Brief Bioinform.* **2004**, *5* (1), 59−70.

(15) Faca, V.; Deng, B.; Phanstiel, D.; Newcomb, L.; Hanash, S. In *Identification of glycoproteins in lung adenocarcinoma cell surface* HUPO 4th Annual World Congress, Munich, Germany, Aug 29−Sep 1, 2005; Molecular & Cellular Proteomics: Munich, Germany, 2005; S190.

(16) Shin, B. K.; Wang, H.; Yim, A. M.; Le Naour, F.; Brichory, F.; Jang, J. H.; Zhao, R.; Puravs, E.; Tra, J.; Michael, C. W.; Misek, D. E.; Hanash, S. M., Global profiling of the cell surface proteome of cancer cells uncovers an abundance of proteins with chaperone function. *J. Biol. Chem.* **2003**, *278* (9), 7607−7616.

(17) Jang, J. H.; Hanash, S., Profiling of the cell surface proteome. *Proteomics* **2003**, *3* (10), 1947−1954.

(18) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25* (1), 25−29.

(19) GenoLogics http://www.genologics.com/cpas.

(20) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **1999**, *17* (10), 994−999.

(21) Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M., Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **2002**, *1* (5), 376−386.

(22) Proteomics Standards Initiative http://psidev.sourceforge.net.

PR0503533