DOI 10.1002/pmic.200500378 *Proteomics* 2006, *6,* 1758–1769

RESEARCH ARTICLE

SPLASH: Systematic proteomics laboratory analysis and storage hub

Siaw Ling Lo[1], Tao You[1]*, Qingsong Lin[1], Shashikant B. Joshi[1], Maxey C. M. Chung[2] and Choy Leong Hew[1]*

[1] Department of Biological Sciences, Faculty of Science, National University of Singapore, Singapore
[2] Department of Biochemistry, Faculty of Medicine, National University of Singapore, Singapore

In the field of proteomics, the increasing difficulty to unify the data format, due to the different platforms/instrumentation and laboratory documentation systems, greatly hinders experimental data verification, exchange, and comparison. Therefore, it is essential to establish standard formats for every necessary aspect of proteomics data. One of the recently published data models is the proteomics experiment data repository [Taylor, C. F., Paton, N. W., Garwood, K. L., Kirby, P. D. *et al.*, *Nat. Biotechnol.* 2003, *21*, 247–254]. Compliant with this format, we developed the systematic proteomics laboratory analysis and storage hub (SPLASH) database system as an informatics infrastructure to support proteomics studies. It consists of three modules and provides proteomics researchers a common platform to store, manage, search, analyze, and exchange their data. (i) Data maintenance includes experimental data entry and update, uploading of experimental results in batch mode, and data exchange in the original PEDRo format. (ii) The data search module provides several means to search the database, to view either the protein information or the differential expression display by clicking on a gel image. (iii) The data mining module contains tools that perform biochemical pathway, statistics-associated gene ontology, and other comparative analyses for all the sample sets to interpret its biological meaning. These features make SPLASH a practical and powerful tool for the proteomics community.

Received: May 27, 2005
Revised: August 12, 2005
Accepted: September 1, 2005

Keywords:
Database / PEDRo / Proteomics data exchange / Proteomics data mining / XML

Correspondence: Professor Dr. Maxey C. M. Chung, Department of Biochemistry, Faculty of Medicine, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, Singapore
E-mail: bchcm@nus.edu.sg
Fax: +65-6779-1453

Abbreviations: AGML, annotated gel markup language; **GO**, gene ontology; **HUP-ML**, human proteome markup language; **ICAT**, isotope-coded affinity tag; **KEGG**, Kyoto Encyclopedia of Genes and Genomes; **MIAPE**, minimum information about a proteomics experiment; **PEDRo**, proteomics experiment data repository; **SPLASH**, systematic proteomics laboratory analysis and storage hub; **SVG**, scalable vector graphics; **XML**, extensible markup language; **XSL**, extensible stylesheet language; **XSP**, extensible server pages

1 Introduction

In the postgenomic era, proteomics research has become a major platform for the studies of biological processes, as well as for biomarker and drug discovery [1]. The rapid growth of the proteomics field is largely attributed to the developments in protein separation technologies and MS [2, 3]. 2-DE continues to be the most prevalent approach for traditional proteomics [4]. Its long-standing popularity is primarily due to developments of narrow-range IPG strips [5], DIGE [6], and gel image analysis software [7], all of which have contributed

* These authors contributed equally to this work.

to the dramatic improvements in the resolution of complex proteomes, and our ability to quantitate protein levels over a wide range of concentrations.

However, 2-DE remains to be a challenging experimental approach with complex and laborious workflows. Over the past few years, various nongel-based technologies have been developed as alternatives to 2-DE. These include liquid-phase electrophoresis, CE, and various combinations of LC [2]. Quantitative methods such as isotope-coded affinity tag (ICAT) [8], iTRAQ [9], and stable isotope labeling by amino acids in cell culture (SILAC) [10] have also been established to facilitate protein profiling and to study protein expression levels.

MS and related technologies also play critical roles in proteomics, and continue to undergo rapid developments [3]. Combinations of different ion sources with different mass analyzers have led to the introduction of several new types of mass spectrometers that possess improved sensitivity, mass accuracy, and resolution, and the ability to perform analyses in a high-throughput fashion, all of which are necessary for accurate and reliable protein identification. However, data analysis software and storage methods vary among instruments, even for different models of mass spectrometers from the same manufacturer.

Often, variability in data acquisition and data interpretation leads to incompatibility for sharing information among different laboratories. Also, in the published literature, important information, such as sample extraction and preparation procedures, and analytical methods, is often absent. Despite the on-going efforts in standardizing the guidelines for publishing in the scientific literature, it is often difficult for different proteomics laboratories to share, compare, and verify their experimental data due to the wide variety of protein separation technologies, mass spectrometric data acquisition, and processing methods.

This issue was first addressed by the Consortium for the Functional Genomics of Microbial Eukaryotes group, who subsequently developed a standard data model for describing proteomics experiments called proteomics experiment data repository, also known as PEDRo [11]. It systematically captures and stores important aspects of a proteomics experiment.

Based on PEDRo, we have recently developed systematic proteomics laboratory analysis and storage hub (SPLASH), a web-interfaced extensible markup language (XML)-based system, to establish the necessary informatics infrastructure that is required to support proteomics studies in terms of systematic data management and analysis, and comprehensive data searching. This paper is targeted at bioinformaticians who work on similar projects and proteomics researchers who have a need to use SPLASH for their own research. Therefore, Section 2 of this paper delve into the implementation details, while the functionality of this system is demonstrated in Sections 3 and 4.

To accommodate for the workflow of human diseases research in our laboratory as closely as possible, some modifications have been applied to the original PEDRo data model. While changes have been made in several entity relationships, all the information contained in the original model has been maintained in the system to ensure data exchange and dissemination in the original PEDRo format. In addition to the basic data maintenance module (data entry, update, and tracking functions), the system also consists of data search and mining modules to analyze the high-throughput data. The data search module provides several means to search the database, including protein database accession numbers, protein name, 2-DE spot ID, protein p*I* and MW, and 2-DE dates. There are additional options to view protein information and differential expression display (of the protein spots) among the 2-DE gels for the same sample by clicking on the gel image. To enable efficient and systematic analyses and verification of the data, SPLASH incorporates the data mining module including gene ontologies (GOs), Kyoto Encyclopedia of Genes and Genomes (KEGG) biochemical pathway analysis, and comparative sample analysis tools. The online demonstration for the most up-to-date progress on SPLASH is at http://oncoproteomics.nus.edu.sg/splash.

2 Materials and methods

SPLASH is developed on the Apache Cocoon® platform (Fig. 1), with various kinds of XML-related technologies involved. In order to cater for in-house usage and sharing of data with international research community, SPLASH is designed to provide individual users restricted access to different portals. Currently, there are three portals available, *viz*., search portal, mining portal, and maintenance portal, for data search, data mining, and data management, respectively. Overall system design and implementations for every functional portal are detailed in this section.

2.1 Overall system design

XForms-XSP-XML-XSL-SVG (XSP, extensible server pages; XSL, extensible stylesheet language; SVG, scalable vector graphics) have been adopted for their intrinsic modularity and ability to dynamically create server-side web forms for extensive user interactivity such as data entry and validation, presentation of information-rich graphics, for example, 2-DE images.

Various JAVA programs are developed for complex data manipulation, such as two-sample *t*-test (in data mining module), batch data uploading and data exchange (in data management module). Several JavaScripts are used to create interactive web pages. Perl scripts are written to handle the fast sequence retrieval from FASTA files for further sequence analyses.

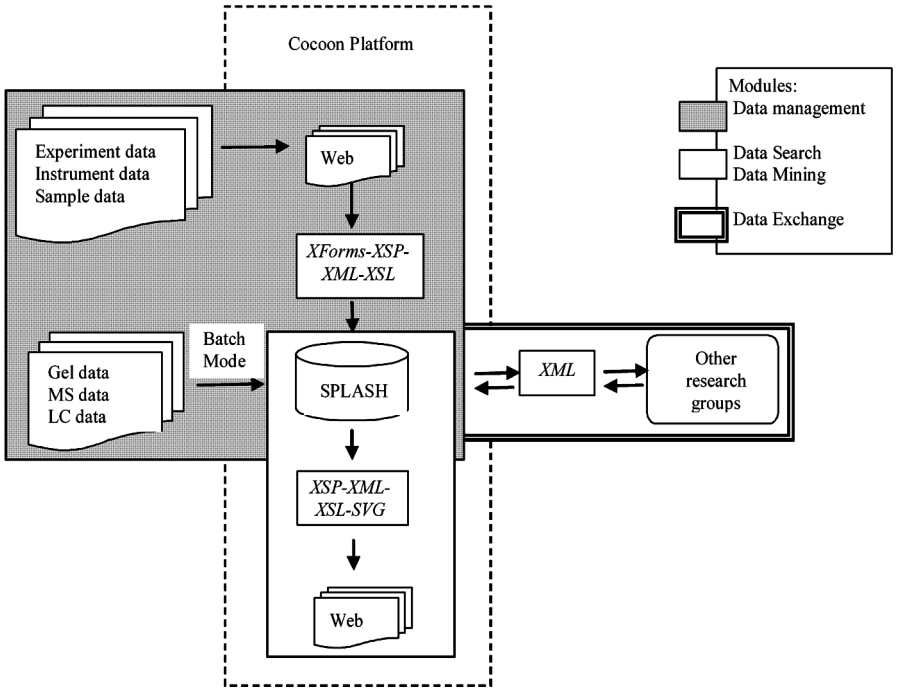

Figure 1. Overall system design. XForms-XSP-XML-XSLT is adopted in the data management module. XForms allows various kinds of constraints to be applied on the input fields. Its good compatibility with XSP technology enables the combination of XForms-XSP-XML-XSL to dynamically create forms for data entry and updating. XSP-XML-XSL-SVG is used in the data search and data mining modules. XSP is intensively used for interfacing between the web request and data retrieval for the data search module. Its ESQL logic sheet is heavily used to perform SQL queries before serializing the results in XML. This XML is then transformed by XSL Transformation (XSLT) into an SVG image for interactive visualizations, including descriptions of 2-D graphics and graphical applications in XML.

2.2 Database structure design

The SPLASH database structure is customized from the original PEDRo data model [11] to simplify data queries. In the same framework as PEDRo, changes are mainly made to the relationships among different classes of the four sections (sample generation, sample processing, MS, and MS result analysis) (Fig. 2). All sections are connected by a central relationship-keeping class named "Analyte Processing Step".

In addition to these, new classes (N1, N2 in Fig. 2) are added to connect sample processing information with its protein identification information. This greatly reduces the complexity required to develop data mining tools. At the same time, these customizations are interconvertible with the original PEDRo for data exchange and dissemination. The GO and KEGG biochemical pathway information (N3, Fig. 2) have been incorporated for data analysis and annotation. Both of these are downloaded regularly and integrated into the database.

2.3 System requirements

Apart from the Apache Cocoon, the current SPLASH system relies on several other technologies. Both SVG and XForms require plug-ins which are freely available for some of the most popular web browsers, including Internet Explorer and the Mozilla Firefox.

2.4 Search portal – Gel image linkage

SPLASH enables the linking between gel image and its relevant data, such as the linkage for the differential expression display for a particular protein spot in all the physical gels in the same sample with its protein details and *vice versa*. A combination of XSP-XML-XSL-SVG is used to implement this feature. First, an XSP file is used to extract information of all the detected spots of one physical gel from the database into an XML file. This XML subsequently undergoes XSL transformation to generate an SVG image for viewing.

2.5 Mining portal

2.5.1 Sample comparison

In order to facilitate comparative proteomics analyses, a sample comparison module has been built. This module enables us to evaluate the statistical significance of the quantitative differences between the test and control samples (or in our context, paired samples of tumor and normal tissues), in both the 2-DE and ICAT experiments.

The comparative methods include a two-sample *t*-test for a single sample followed by sample-to-sample comparisons for 2-DE datasets and multisample comparisons for ICAT datasets. Sample-to-sample comparisons can reveal the common trends among different sample sources and hence reduce the chances of false discoveries.

2.5.2 GO implementation

GO sets a paradigm for systematic categorization of high-throughput dataset based on three structured and controlled ontologies (biological processes, cellular components, and molecular functions) [12]. To assign GO annotation for the

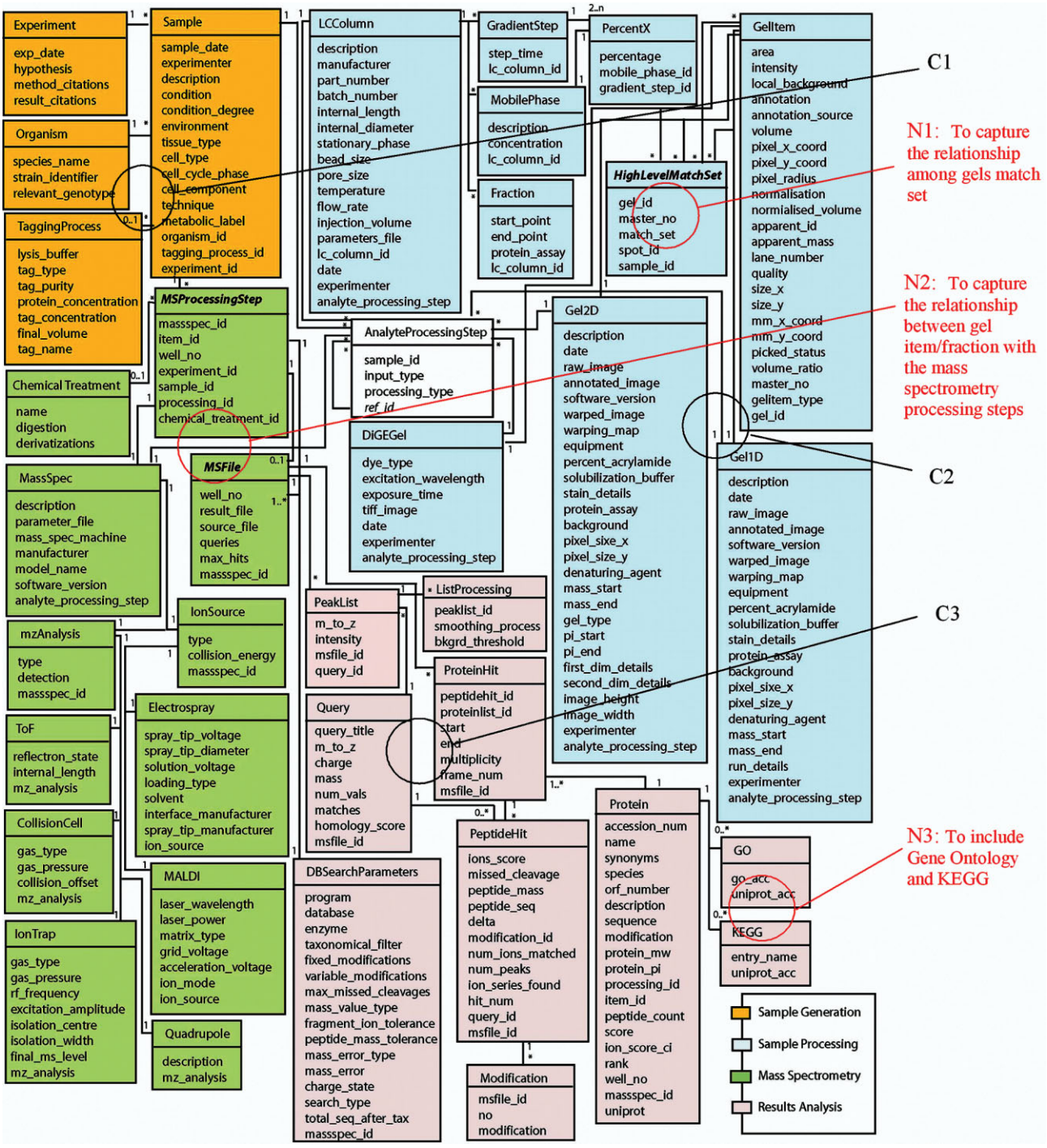

Figure 2. SPLASH database structure. There are three customizations on the PEDRo data model (C1, C2, C3) and three new areas are introduced (N1, N2, N3). C1: the PEDRo "Sample Origin" class has been merged into the Sample class. C2: the PEDRo "Gel" class has been split into "Gel 1D", "Gel 2D", and "DIGE Gel" classes. Original Band and Spot classes are merged into the "Gel Item" class to make it universal for various types of gels. C3: a more result-oriented data structure has been designed in favor of result handling and viewing. Details of corresponding sections of the original PEDRo data model can be found in the Supplementary Material. N1: "High Level Match Set" new class is introduced to capture the relationship of matched gel spot from image analysis of multiple gels with its original gel spot in each individual gel. N2: a new class called "MS Processing Step" links the MS processing information directly to their corresponding Gel Item (gel-based method) and "Fraction" (LC based method) classes. "Chemical Treatment" class is added to document the chemical process before MS. Another new class "MS File" is added to store the information of MS raw files, in order to maintain links to the "Peak List" and "Protein Hit" classes in the MS result analysis section. N3: Classes of GO and KEGG pathway databases.

proteins, the proteins' accession numbers are used to retrieve the respective GO terms from UniProt, a GO annotated database [13]. If no exact match is found in UniProt, a BLAST [14] search will be carried out to find a sequence homologue with a minimum coverage of 75% and identity of 35% [15]. Iterative queries are used to find the hierarchical relationship among GO terms within a proteomics dataset. The results can be presented in either a categorical or tree view *via* different XSL transformations. This analysis can be applied to single sample or multiple samples.

In order to identify the GO categories of interest for ICAT and 2-DE data, *i.e.*, GO category that is enriched or depleted significantly with respect to a reference group, two-sided Fisher's exact test is conducted for each GO category [16]. In the context of human disease research, the reference group can be considered as either the whole experimental (ICAT or *t*-test) dataset, or all the human proteins carrying GO annotations.

When the reference group is chosen as the whole experimental dataset, the null hypothesis for a certain GO category is the proportion of changed proteins in this category among all the proteins of this category is equal to the proportion of changed proteins falling out of this category among all the proteins falling out of this category in this experiment (Table 1). Fisher's exact test provides the exact probability value, even when there are very small numbers in the contingency table. Since no prior knowledge about where the observations tend to lie in the contingency table is known, the two-sided Fisher's exact test is performed to detect significant differences in both directions across the table. The test is conducted at the significant level of 0.05.

Table 1. 2×2 contingency table for changed proteins and unchanged proteins in a particular GO category. N_c and n_c are replaced by N_{up} and n_{up} or N_{down} and n_{down} when analyzing up-regulated or down-regulated proteins, respectively

	Changed proteins	Unchanged proteins	Total
In category	n_c	$n - n_c$	n
Not in category	$N_c - n_c$	$(N - n) - (N_c - n_c)$	$N - n$
Total	N_c	$N - N_c$	N

2.5.3 KEGG pathway implementation

KEGG provides a suite of databases [17] integrating the current knowledge on molecular interaction networks in biological processes, and the information about genes, proteins, chemical compounds, and biochemical reactions. A JAVA program has been developed to upload KEGG Markup Language (KGML) files describing biochemical pathways to SPLASH. Biochemical pathway analysis can be applied to single or multiple LC-MS datasets, ICAT datasets, or a set of 2-DE proteins corresponding to two-sample *t*-test's results. For the latter two cases, protein entries which are differentially expressed are highlighted by different colors on the report lists.

2.5.4 Overview and progress tracking

In order to provide an intuitive view of each individual project's progress and its experimental workflow, the database structure has been designed to maintain the relative orders of each analytical step by introducing a new "ref_id" in Analyte Processing Step table. In addition, a unique ID (massspec_id) is assigned to each MS experiment to link the initial analytical steps (such as 2-DE and LC) with the corresponding identified protein records for ease of tracing the experimental details.

2.6 Maintenance portal

2.6.1 Data entry

Data input forms are designed to manually enter the experimental records. Various templates have been designed for instrument data, such as MALDI machine setting, to represent standardized experimental protocol within an organization and to reduce the data entry effort. For batch experimental results, several JAVA programs are developed to import them from GPS Explorer (Applied Biosystems), MASCOT (Matrix Science), PDQuest (BioRad), and DeCyder (GE Healthcare) software into the database.

2.6.2 Data exchange

JAVA programs have been developed to import and export XML data in PEDRo format. For both data import and export, an intuitive configuration file, containing the data fields mapping between PEDRo and SPLASH database, is used. The XML file is converted to sets of "field and value" pairs before it is input into SPLASH with reference to the configuration file. By changing the configuration file, this program can keep up with the newest standard readily. Data export to an XML file is done in the same way with additional formatting using the hierarchical relationship of the PEDRo XML structure.

3 Result

In this section, functionality of each portal is exemplified through search and analysis of several sets of human colorectal cancer data. The statistical approach successfully detects some consistent trends among the datasets, while the hypothesis-driven GO analysis provides insights into the functional and subcellular localization distribution with statistical significance.

3.1 Search portal

SPLASH provides various search options (Fig. 3), namely, "Accession No", "Protein Name", "Spot Number", "p*I* and MW", and "2-D Gel Date". The first three search options have very flexible selections. Queries can be limited to all records, a particular experiment, a specific sample, or a physical gel (or an LC-MS set, or an ICAT dataset where appropriate). Users can also browse for useful information by clicking on a gel image, in the two options known as 2-D Gel Image and "Gel Differential Expression".

3.2 Mining portal

3.2.1 Sample comparison

The sample comparison module handles two types of data. First, comparisons can be made for 2-DE experiments, *via* "Two-sample *t*-test" (Fig. 4A) and "Sample Comparison" (Fig. 4B). The former function analyzes gel image annotations of one sample, and fishes out significant up- and down-regulated spots. In order to find significant and consistent differences between multiple paired samples, the latter function applies two-sample *t*-test on each individual sample, followed by a consensus finding process where consistently up- and down-regulated spots are found. The consistencies among multiple samples manifest the fundamental change of tumor and may have a higher chance to represent biological significance. Both tools allow average normalized intensity ratio restriction and significance level selection.

Second, "ICAT Differential Exp" function (Fig. 4C) finds consistencies among multiple ICAT experiments according to the ICAT threshold range selected.

3.2.2 GO

In SPLASH, GO analysis can be carried out for one protein, multiple LC-MS sets, multiple ICAT datasets, or a set of proteins corresponding to the two-sample *t*-test results of a paired sample (Fig. 5). The system also allows analysis of multiple datasets from different analytical methods, for example, combining LC-MS sets and ICAT datasets, but the generated result does not have any statistical significance.

Prior to the analysis of an LC-MS dataset, the user can set the ion score's confidence interval and top rankings from MASCOT MS results to select the subset of data to analyze.

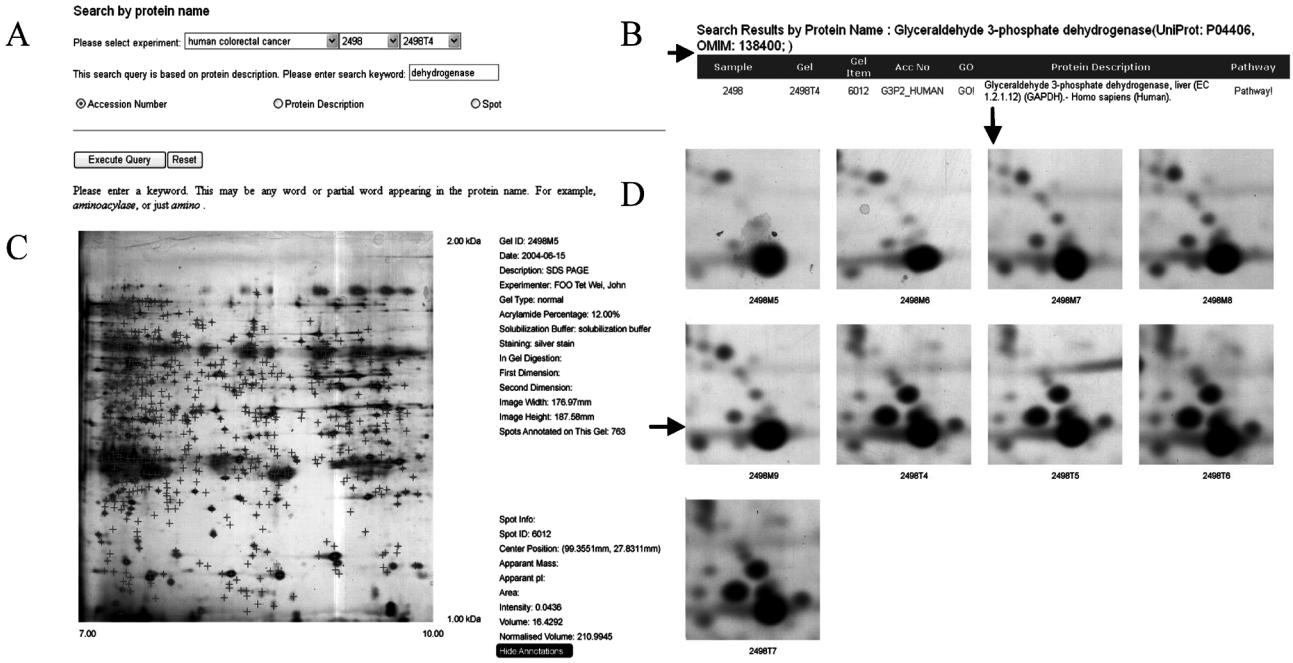

Figure 3. Search portal design. (A) This screenshot shows the selection page for the "Protein Name" option where multilevel selection is allowed. (B) On the result page, a record (glyceraldehyde 3-phosphate dehydrogenase) is found. Hyperlinks to its sample, gel, and spot details, GO analysis results, associated biochemical pathways, and related OMIM record are found in the table. Spot detail page further points to the differential expression display (D). (C) Both the "2-DE Image" and the Gel Differential Expression options allow the user to look for information by clicking on a gel. They provide annotated full gel image as shown. All the detected spots are labeled with small red crosses which can also be turned off. Basic information regarding this gel is lined up at the top right portion of the image. When the mouse cursor hovers on a cross, it flips into green with corresponding spot details displayed on the bottom left. User can click on the cross to view corresponding protein details or the differential expression display of this particular spot on all the gels of the same sample, depending on the chosen tool. (D) Differential expression display of the spot 6012 (centralized) in human colorectal cancer sample 2498.

A

B

Sample comparison result table for match set 2004-10-07-13-32-29

Spots with Type I Occurrencies over 4 times (altogether 3 spots)

Master No	Type I Occurrence	Type II Occurrence	Type III Occurrence	average Ratio (Type I spots)
2123	4	0	0	39.543
2107	4	0	0	16.73
2111	4	0	0	4.406

Spots with Type I Occurrencies less than 4 times (altogether 414 spots)

Master No	Type I Occurrence	Type II Occurrence	Type III Occurrence	average Ratio (Type I spots)
3109	3	0	1	7.824
6002	3	0	0	60.408
3114	3	0	0	31.668
3102	3	0	0	14.451
1404	3	0	0	13.016
4102	3	0	0	11.725
3002	3	0	0	6.362

C

Differential Expression Mining Result with ICAT ratio >3 for human colorectal cancer_2446_2446-ICAT and human colorectal cancer_2361_2361-ICAT and human colorectal cancer_3428_3428-ICAT

Acc No	Name	MW	pI	UniprotFreq	2446-ICAT ratio	2361-ICAT ratio	3428-ICAT ratio
IPI00472102	**60 kDa heat shock protein, mitochondrial precursor**	61174.44	5.70	P10809	3	4.910	3.114	3.129
IPI00005611	PHD finger-like domain protein 5A	12356.95	8.78	Q7RTV0	2	3.560	3.405	
IPI00012074	HNRPR protein	71170.39	8.23	Q9BV64	2	3.346	3.357	**2.643**
IPI00029574	Putative S100 calcium-binding protein A11 pseudogene	11279.74	7.77	Q60417	2	3.687	3.703	
IPI00218993	Splice Isoform 2 Of Heat-shock protein 105 kDa	92057.43	5.42	Q92598	2		3.852	3.830
IPI00298860	Lactotransferrin precursor	78288.00	8.56	P02788	2	5.611	1.753	8.278

Figure 4. (A) Two-sample *t*-test is conducted for the human colorectal cancer sample 2498 in the whole ratio range. After the test, statistically significant spots are fished out and further classified into three groups when reporting. "Type I" group only contains spots present on all the physical gels for this paired sample; "Type II" contains spots consistently present on all the gels of one type (such as normal) but are consistently missing on all the other type of gels (such as tumor). The rest are lumped together as "Type III". They are inconsistent but passed the *t*-test. Apart from the table, results can be shown more intuitively as an SVG image with four heat maps, containing all spots or only one type of spots. Each matched spot is represented by a single colored band in the order of either spot serial number (ascending) or the ratio of average normalized intensity of that spot (descending) from the top. Ratio is defined as

$$\text{Ratio} = \frac{\bar{x}_{\text{test}}}{\bar{x}_{\text{control}}}. \text{ (in our context, Ratio} = \frac{\bar{x}_{\text{tumor}}}{\bar{x}_{\text{normal}}})$$

Red is for up-regulation, green for down, and black for comparable cases. When the mouse cursor hovers on a band in the latter three columns, the band itself, and its corresponding band in the first column flips into blue with the spot information displayed on the right. Clicking on the band leads to the differential expression display mentioned before. (B) Sample-to-sample comparison is performed for four human colorectal cancer samples, 2443, 2466, 2446, and 2498. A very stringent selection criterion (ratio > 2, significance level = 0.05, Type I for all four samples) is applied. Consistent records are listed separately from the rest. Master number is the unified spot number across the four samples. Clicking on it leads to the spot details for each sample. Original spot matching was done in PDQuest. (C) Consistencies are found for three human colorectal cancer ICAT datasets. Consistent records are bolded.

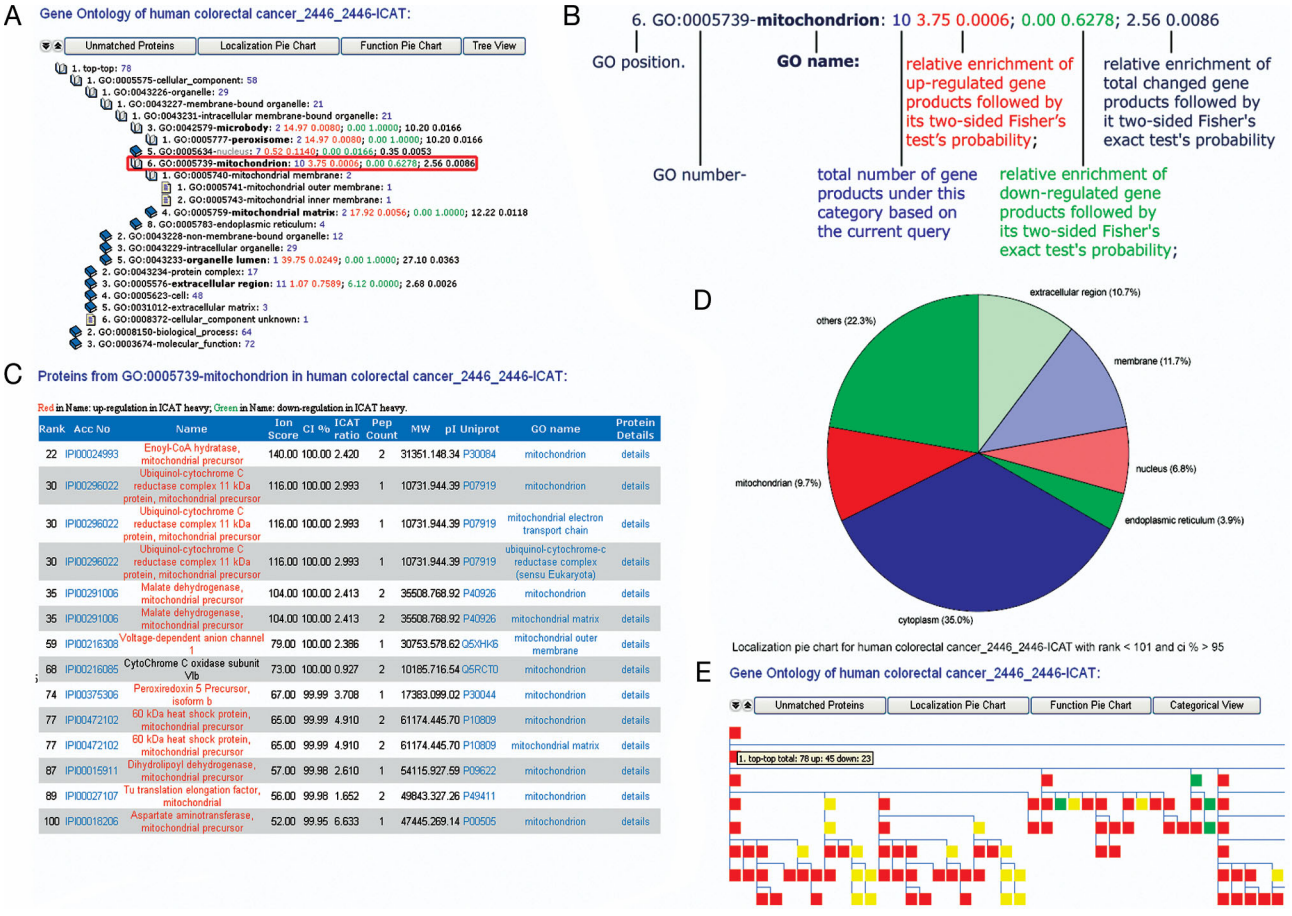

Figure 5. GO analysis is applied to a subset of human colorectal cancer sample's ICAT dataset 2446. Proteins with top 100 MASCOT ranks and confident CIs above 95% are selected for GO annotations until GO level 7. All human proteins carrying GO annotations are selected as the reference group. (A) Tool provides a comprehensive categorical view of the GO structure. There are 78 proteins, within this set, carrying GO annotations. (B) Statistical results regarding a specific GO category. (C) Proteins under a specific GO category are color-coded according to their ICAT threshold. Further analyses on each individual protein can be carried out by following links in the result table. (D) Pie chart showing GO terms' cellular components makeup. (E) Tree view of the GO structure. Red indicates predominant up-regulation $\left(\frac{\text{Proteins with threshold} > 1}{\text{All proteins}} > 0.5\right)$ under the GO category; green points out predominant down-regulations; and yellow is for the rest.

For ICAT dataset and 2-DE *t*-test protein set, two-sided Fisher's exact test is applied to each GO category to find out the GO categories of interest compared to the reference group. A probability value lower than 0.05 indicates a significant enrichment or depletion of the proteins under this GO category and therefore implies a further research target.

The GO analysis tool provides an expandable and collapsible categorical view of the GO structure. Proteins under each GO category are viewable. Proteins without GO annotations are also compiled as a standalone list. Sequence-based domain/motif and localizations predictions are viewable by clicking the link on their accession numbers. Pie charts showing the GO terms' distribution in terms of their cellular component ontology and molecular function ontology can be dynamically created. A tree view of the three ontologies is also available.

3.2.3 KEGG biochemical pathway analysis

This tool (Fig. 6) finds relevant biochemical pathways for a set of proteins, including multiple LC-MS sets, multiple ICAT dataset, or a set of differentially expressed proteins passing the two-sample *t*-test identified by the gel-based method. LC-MS results can also be selected together with ICAT datasets.

3.2.4 Overview

SPLASH provides a panorama view of the existing data, which is also a simple way to access information (Fig. 7).

Biochemical pathway(s) mining results for sample 2446 :

Red in Name: up-regulation in ICAT heavy or tumor where appropriate; Green in Name: down-regulation in ICAT heavy or tumor where appropriate.

Glycolysis / Gluconeogenesis (path:hsa00010)

Rank	UniProt	Name	CI %	ICAT ratio	KEGG Entry Name	Type	Reaction	OMIM
87	P09622	Dihydrolipoyl dehydrogenase, mitochondrial precursor	99.98	2.610	hsa:1738	gene	rn:R01698	246900
28	P13929	EnolasE 3	100.00	3.873	hsa:2023 hsa:2026 hsa:2027	gene	rn:R00658	131370
14	P04075	Fructose-bisphosphate aldolase	100.00	2.836	hsa:226 hsa:229 hsa:230	gene		103850
14	P04075	Fructose-bisphosphate aldolase	100.00	2.836	hsa:226 hsa:229 hsa:230	gene		103850
14	P04075	Fructose-bisphosphate aldolase	100.00	2.836	hsa:226 hsa:229 hsa:230	gene		103850
14	P04075	Fructose-bisphosphate aldolase	100.00	2.836	hsa:226 hsa:229 hsa:230	gene	rn:R01070	103850
24	P04406	Glyceraldehyde-3-phosphate dehydroGenase	100.00	2.029	hsa:2597 hsa:26330	gene	rn:R01061; rn:R01063	138400

Figure 6. Seventy-two out of the total 128 *homo sapiens*-related KEGG biochemical pathways (HSA v0.6) are found correlated with the aforementioned ICAT subset. Related KEGG entries are listed for each pathway. Some KEGG entries, such as fructose-bisphosphate aldolase, may have multiple presences in one pathway. Some (glyceraldehyde-3-phosphate dehydrogenase) may be involved in multiple reactions. Each KEGG entry's basic information is listed. User may click on the UniProt number to view protein details, or click on the entry ID to see its KEGG online annotation. Corresponding OMIM online record and KEGG reaction information are also accessible.

Analyte Methods Overview

Expand Collapse

- 1. Experiment- Acute Promyelocytic Leukaemia: To investigate the difference in protein profile between treated and normal leukaemic cell
- 2. Experiment- Blood: blood LC experiment
- 3. Experiment- HCC: To identify interesting candidate protein through differential expression proteomics study of HCC
- 4. Experiment- HCT: proteome study via 2D-LC
- 5. Experiment- human colorectal cancer: reasons to cause cancer can be traced back through protein differential expression comparison
  - 1. Sample- 2361: human colon cancer tissue sample, obtained from SGH
  - 2. Sample- 2443: human colon cancer tissue sample, obtained from SGH
  - 3. Sample- 2446: human colon cancer tissue sample, obtained from SGH
    - 1. Gel2D- 2446T1: SDS PAGE
    - 2. Gel2D- 2446T2: SDS PAGE
    - 3. Gel2D- 2446T3: SDS PAGE
    - 4. Gel2D- 2446T4: SDS PAGE
    - 5. Gel2D- 2446M1: SDS PAGE
      - MassSpec- [Results] MALDI_ToF: Applied Biosystems 4700 proteomics analyzer MALDI-TOFTOF
    - 6. Gel2D- 2446M2: SDS PAGE
    - 7. Gel2D- 2446M3: SDS PAGE
    - 8. Gel2D- 2446M4: SDS PAGE
    - 9. Gel2D- 2446M5: SDS PAGE
    - 10. LCColumn- 2446-ICAT: human colon cancer tissue 2446M-ICATlight; 2446T-ICATheavy study
      - MassSpec- [Results] MALDI_ToF: Applied Biosystems 4700 proteomics analyzer MALDI-TOFTOF
  - 4. Sample- 2466: human colon cancer tissue sample, obtained from SGH
  - 5. Sample- 2498: human colon cancer tissue sample, obtained from SGH
  - 6. Sample- 3428: human colon cancer tissue sample, obtained from SGH

Figure 7. All experimental documentation and results for gel analysis, LC, and MS experiments are displayed in a collapsible list.

3.3 Maintenance portal

Proteomics data maintenance can be done *via* the web interface in this module. Experimental information including samples, analytical methods, and MS results are entered manually or uploaded in batch mode. These data can also be updated when necessary (Fig. 8).

The sample generation, sample processing, and MS sections of PEDRo can be readily imported into the database and the data from SPLASH can also be directly exported as PEDRo-compliant XML file. The data exchange of the MS result analysis section is currently designed to be compatible with the MS standard format mzData (http://psidev.sourceforge.net/ms/#mzdata).

4 Discussion

Active efforts are being made to update SPLASH's underlying structure to accommodate for the newest HUPO guidelines and other popular data formats to ensure its compatibilities with the latest technologies. Besides, limitations and the potential solutions for the hypothesis-driven GO analysis methods, annotation data reliability issues, and MS data validation practices are discussed in Section 4.1.

4.1 Standardization of data model

PEDRo [11] was published as a data model, which can be implemented as a database, for the standardization of prote-

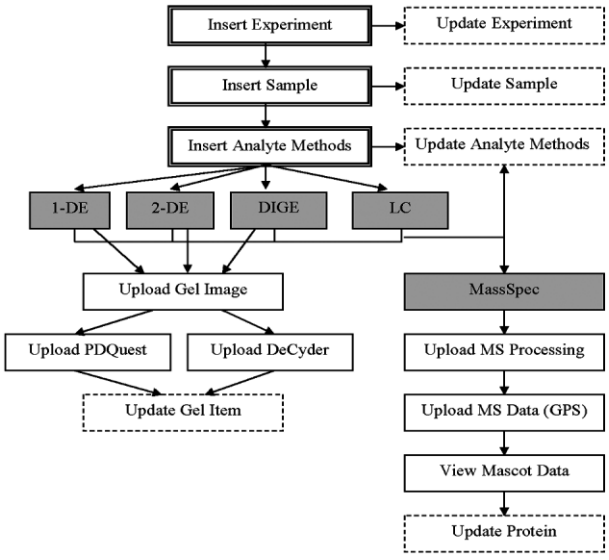

Figure 8. This is the workflow for data maintenance (data entry and update). Starting with a completely new experiment, experimental records are entered using the Insert modules (double-line box). User can choose from different processing methods (gray box) to enter. Mass spectrometer's information is keyed in afterwards. Gel image and analysis data from third-party software (PDQuest and DeCyder) are uploaded into SPLASH in batch mode (white box). Similarly, MS data and analysis results from third party tool (GPS) are also uploaded in correspondence with their spots or LC fractions. To amend data, use the Update tools (dash line box).

omics data. Recently, a new set of guidelines, minimum information about a proteomics experiment (MIAPE) [18], is introduced to define the minimal information needed. The aim of these efforts is to standardize minimum reporting requirements for proteomics experiments by introducing various modules and subdomains with specific data and metadata. Besides the MIAPE guidelines, the MS data such as mzData and mzIdent from HUPO proteomics standards initiative (PSI) are also under active development. To comply with future standards, SPLASH database structure has been designed to be robust and versatile to allow for future additional implementations. For example, a new processing step or instrument can easily be incorporated by linking to the central class Analyte Processing Step. In addition, the implementation on data exchange using a configuration file allows easy customization for future standards. SPLASH is upgraded on an ongoing basis to keep in compliance with MIAPE so that it can fulfill its role in proteomics data exchange and dissemination.

4.2 Sources of bias for the two-sided Fisher's exact test

The results of the two-sided Fisher's exact test which is applied in the GO mining module are susceptible to several sources of error.

Firstly, the ICAT experimental errors encountered during both the sample preparation and protein identification processes can never be avoided. The ICAT reagents selectively label the cysteinyl residues of proteins, covering approximately 85–90% of a proteome [8], and consequently, 10–15% of proteins which do not contain cysteines are not detected by this method. Identification of ICAT-labeled peptides is largely determined by the sensitivity of the mass spectrometer used. Less abundant proteins thus have fewer chances to be identified. Therefore, the experiment's accuracy limits the precision of the parameters that define the two-by-two contingency table. The ICAT quantitation generally has an SD of 10–20% [19]. Therefore, we have arbitrarily chosen 50% change in expression levels as the statistically significant cutoff threshold to differentiate changed and unchanged protein levels. However, false-positive and false-negative assignments cannot be avoided as they are dependent on experimental limitations. For example, in the case of altered expressions with respect to the whole ICAT dataset, all parameters essential to construct the contingency table, including n_c, $N - n$, N_c, and $N - N_c$, are affected. Knowledge of this source of bias would be beneficial in the evaluation of the results.

Secondly, comparing the subset of proteins with all the human proteins carrying GO annotations also causes bias. All proteins that are not detected by the ICAT experiment are counted into n and N, but not into N_c. Consequently the contingency table tends to exhibit bias on the right. In addition, the GO database itself causes further bias. In the known human proteome, out of 48 953 proteins (International Protein Index (IPI) [20] database released in February, 2005), only 26 444 of them carry GO annotation. This significant difference further entails bias.

Thirdly, the protein identification process may generate false discoveries due to data reliability issue as discussed in Section 4.3.

4.3 Corrections for the two-sided Fisher's exact test

When a large number of proteins have been detected by a high-throughput method, the issue of the false positives accompanying the two-sided Fisher's exact test becomes significant. Several statistical techniques of controlling the Type I error rate are available [21, 22]. Initially, a number of ways associated with the family-wise Type I error rate (FWER) were developed to control the probability of committing even one error in the family of hypotheses. However, these methods are often very stringent and subsequently make the true positive discovery even harder. The false discovery rate (FDR) is the expected proportion of erroneously rejected null hypotheses among the rejected ones. Controlling FDR is less conservative and sufficient for the GO result analysis. As a result, it is a potentially more powerful technique for data analysis. A statistical tool applying this method is now under development.

4.4 Evaluation of GO and KEGG data reliability

UniProt is currently the main linking point to GO and KEGG pathway in SPLASH as it is a high quality and extensively cross-referenced information hub for protein sequences. While UniProt accession number is used for retrieving and integrating all known information about a protein, we use IPI sequence database for protein identification by MS and MS/MS experiments, as this database also includes annotated gene splice variants, thus ensuring protein identification with greater confidence. The identified protein is then assigned its UniProt accession number (if available) through cross-referencing between the two databases. However, for those proteins that are not in UniProt or have not been annotated with GO or KEGG data, our current approach is to find the sequence homologue of the unmatched protein and extract its UniProt accession number for further analysis. Although sequence homologues can be inferred to have similar function, it is possible that the protein identified by BLAST may not be the "correct" protein. Thus it may be annotated with incorrect GO or KEGG information. In addition, it is important to point out that some of the GO annotation is based on inferred from electronic annotation (IEA), known to be a source of errors as no curator has checked the annotation to verify its accuracy. Taking all these factors into consideration, we are in an active process of devising a reliability index for GO and KEGG annotation in order to improve data accuracy.

4.5 MS data validation

In addition to storing and managing the research data, it is equally important to have tools to validate the research data, especially the MS data. Since the data analysis and verification can be subjected to user's judgment, it is beneficial to have additional tools to analyze the spectra so as to evaluate and validate (a) the potential of false-positive matches, *i.e.*, protein not in database and (b) observed fragmentation trends which may not be incorporated into current MS/MS search algorithms. One such available tool is Trans-Proteomic-Pipeline (TPP) (http://tools.proteomecenter.org/TPP.php). It can be used to validate peptides, to quantitate peptides and proteins, and to identify proteins by MS/MS analysis. It also allows visualization of LC-MS results in 2-D formats for profiling purposes. Hence, we plan to implement TPP or comparable software to improve the quality and efficiency of proteomics data collection.

Recently many scientific journals have started to request researchers to statistically validate their research data, especially the MS derived data, so as to ensure that only high-quality and statistically significant data can be published [23]. In order to make sure that our MS data is comprehensive enough for peer-review and validation, we have taken steps to ensure that a specific type of information such as the type of MS, search engine, or a sequence database used, how peptide and protein assignments were made using the MS software, and the number of peptides (and their sequences) matched when identifying a protein, are captured for every MS experiment.

4.6 Comparisons with other software and systems

In addition to PEDRo, several efforts have been developed to address the proteomics data standardization issues. These include proteois [24], human proteome markup language (HUP-ML) [25], and annotated gel markup language (AGML) [26]. SPLASH can similarly be made compliant with these standards.

Proteios is a client-server proteomics application which aims to provide public data storage and retrieval. Therefore, it has a graphical user interface (GUI) similar to PEDRo for proteomics data collection and validation, and requires supporting databases for data storage. Currently proteios is capable of importing and exporting in PEDRo and mzData formats. Its focus on developing a data repository infrastructure will be beneficial to the proteomics community. SPLASH, on the other hand, is developed with data storage and exchange in mind, together with functions to query and analyze.

HUP-ML describes 2-DE experimental conditions and protein identification in detail. It provides a client-server-based editor to convert the data into XML format for storage and query. Similar to HUP-ML, AGML also focuses on 2-DE and MS analysis results. It gives a web-interfaced solution for data importing and presentation. The efforts of both HUP-ML and AGML in capturing the essence of a 2-DE experiment and its MS data is useful in finalizing a 2-DE/MS standard for proteomics experiment. However, for a comprehensive and systematic proteomics data analysis, it is still essential to include other approach, such as LC, and to incorporate tools for further analysis.

4.7 Supported data, software, and instruments

Since SPLASH has been developed to support the current experimental setting in our laboratory, it is currently designed to support data input from: (a) gel imaging software, PDQuest (BioRad), and DeCyder (GE HealthScience); and (b) MS software, MASCOT (Matrix Science), GPS Explorer (Applied Biosystems), and PS1 (Applied Biosystems). As shown in Fig. 8, the proteomics workflows that are currently supported are 1-DE, 2-DE, LC with or without ICAT labeling, but we are in the process of incorporating the details of multiple dimension LC separations and iTRAQ labeling into SPLASH. With the component-based application framework and modularity of our system design, SPLASH has additional capability to accommodate emerging tools and software.

In addition to GO and KEGG pathway datasets, there is also a plan to integrate the datasets of Reactome [27] and IntAct [28], both compatible with the PSI XML interchange standard so as to allow future molecular interaction information data exchange. There have been active efforts to

make SPLASH an open source project as well. All the source codes will be documented and released to the research community for free installation and customization in the near future. Eventually, through the integration of various instruments and datasets, the SPLASH users will be able to access the wealth of publicly available human proteome knowledge in a systematic, well-structured manner, thus providing a solid basis for new discovery and research.

In conclusion, here we present SPLASH, a web-interfaced database system designed to be compliant with the PEDRo data model. This system aims to provide researchers a means to arrive at a systems-level understanding of their proteomics research data, especially for those data generated from high-throughput methods. SPLASH is sufficiently robust and versatile to accommodate new guidelines from MIAPE and mzData/mzIdent, thus fulfilling its role to systematically store, manage, search, analyze, and disseminate proteomics information for the proteomics community.

The authors would like to acknowledge the financial support of Lee Hiok Kwee Foundation, Office of Life Sciences of National University of Singapore and Singapore Cancer Syndicate (SCS) for this project. We also wish to express our thanks to all the DBS Onco-proteomics Centre staff/students (alphabetically, Xuezhi Bi, John Foo, Cynthia Liang, Justin Lim, Teck Kwang Lim, Mavis Low, Jason Neo, Gek San Tan, Hwee Tong Tan, Sandra Tan, Wee Wee Tan, Zubaidah Binte Mohamed Ramdzan, Shi Hui Wu) for their excellent experimental work and kind advice on SPLASH. Lastly but not least, we would like to thank Dr. Jin-Hua Han for her constructive suggestions on the manuscript.

5 References

[1] Tyers, M., Mann, M., *Nature* 2003, *422*, 193–197.

[2] Issaq, H. J., *Electrophoresis* 2001, *22*, 3629–3638.

[3] Aebersold, R., Mann, M., *Nature* 2003, *422*, 198–207.

[4] Herbert, B. R., Pedersen, S. K., Harry, J. L., Sebastian, L. *et al.*, *PharmaGenomics* 2003, *3*, 3–10.

[5] Fey, S. J., Larsen, P. M., *Curr. Opin. Chem. Biol.* 2001, *5*, 26–33.

[6] Gharbi, S., Gaffney, P., Yang, A., Zvelebil, M. J. *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 91–98.

[7] Marengo, E., Robotti, E., Antonucci, F., Cecconi, D. *et al.*, *Proteomics* 2005, *5*, 654–666.

[8] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F. *et al.*, *Nat. Biotechnol.* 1999, *17*, 994–999.

[9] Desouza, L., Diehl, G., Rodrigues, M. J., Guo, J. *et al.*, *J. Proteome Res.* 2005, *11*, 377–386.

[10] Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B. *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 376–386.

[11] Taylor, C. F., Paton, N. W., Garwood, K. L., Kirby, P. D. *et al.*, *Nat. Biotechnol.* 2003, *21*, 247–254.

[12] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D. *et al.*, *Nat. Genet.* 2000, *25*, 25–29.

[13] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C. *et al.*, *Nucleic Acids Res.* 2004, *32*, D115–D119.

[14] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. *et al.*, *Nucleic Acids Res.* 1997, *25*, 3389–402.

[15] Brenner, S. E., Chothia, C., Hubbard, T. J., *Proc. Natl. Acad. Sci. USA* 1998, *95*, 6073–6078.

[16] Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D. *et al.*, *Genome Biol.* 2003, *4*, R28.

[17] Kanehisa, M., Goto, S., *Nucleic Acids Res.* 2000, *28*, 27–30.

[18] Orchard, S., Hermjakob, H., Julian, R. K. Jr., Runte, K. *et al.*, *Proteomics* 2004, *4*, 490–491.

[19] Hansen, K. C., Schmitt-Ulms, G., Chalkley, R. J., Hirsch, J. *et al.*, *Mol. Cell. Proteomics* 2003, *2*, 299–314.

[20] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y. *et al.*, *Proteomics* 2004, *4*, 1985–1988.

[21] Ge, Y., Dudoit, S., Speed, T. P., *Resampling-Based Multiple Testing for Microarray Data Analysis*, Technical Report (#633), Department of Statistics, UC Berkeley, 2003.

[22] Al-Shahrour, F., Diaz-Uriarte, R., Dopazo, J., *Bioinformatics* 2004, *20*, 578–580.

[23] Carr, S., Aebersold., R., Baldwin, M., Burlingame, A. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 531–533.

[24] Garden, P., Alm, R., Hakkinen, J., *Bioinformatics* 2005, *21*, 2085–2087.

[25] Yoshida, Y., Miyazaki, K., Kamiie, J., Sato, M. *et al.*, *Proteomics* 2005, *5*, 1083–1096.

[26] Stanislaus, R., Jiang, L. H., Swartz, M., Arthur, J., Almeida, J. S., *BMC Bioinformatics* 2004, *5*, 9.

[27] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P. *et al.*, *Nucleic Acids Res.* 2005, *33*, D428–D432.

[28] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S. *et al.*, *Nucleic Acids Res.* 2004, *32*, D452–D455.

