A system for integrative and post-planned analysis of 2-DE/MS centered proteomics data

Juhui WANG[1*], Christophe CARON[2], Xuefeng HE[1], Audrey CARPENTIER[1], Michel-Yves MISTOU[3], Alain TRUBUIL[1], Christophe GITTON[3], Céline HENRY[3] and Alain GUILLOT[3]

[1]INRA, Lab. of Applied Mathematics and Informatics, 78352 Jouy-en-Josas

[2]INRA, Lab. of Mathematics, Informatics and Genomics, 78352 Jouy-en-Josas

[3]INRA, Lab. of Protein Biochemistry and Proteomics Core Facility, 78352 Jouy-en-Josas

Summary

Proteomic analysis is intrinsically an iterative, incremental process. Information is usually acquired gradually by researchers, and in different projects. At the same time, there are relatively few examples of biological data management systems which take into account this reality, most of them usually treat the experiment generated data as static and unchangeable: data are never reconsidered, or seldom, whereas technology becomes more powerful or that other researchers have brought information on data correction. And yet, post-planned analysis [21] which involves multiple iterations and subsequent re-investigations of previously prepared data might bring tremendous benefits.
Named PARIS (Proteomic Analysis and Resources Indexation System), the system we developed here seeks to address this requirement. Compliant with the majority of 2-DE analysis and MALDI-TOF based protein identification softwares, it automatically takes data from them and stores the raw and processed data in a relational database suitable for advanced exploration. Taking into account the standards proposed by PSI (Proteomics Standard Initiative), the system exports the stored data in XML format for data exchange and knowledge sharing. PARIS also manages information about experiments and their biological contexts, and allows the user to search and analyze a large data collection in a global manner. It provides tools for data integration and advanced, cross multi-experiment, multi-experimenter data exploration, and supports visual verification and correction of the analysis results. Implemented in Java, the system is platform independent, accessible to multiple users through Internet. It is also scalable for use for one or many laboratories, and therefore suitable to inter-institute collaborative work.
PARIS can be tested and downloaded at `http://genome.jouy.inra.fr/paris`

1 Introduction

Today, functional genomics and particularly proteomics is becoming a central topic of the bioscience, many large-scale projects are rolled out. All these projects appeal, to a certain extent, to specialized gel image analysis and protein identifi cation softwares such as ImageMaster

*To whom correspondence should be addressed. Tel: (33) 1 34652229; Fax: (33) 1 34652217; Email: Juhui.Wang@jouy.inra.fr

(Amersham Biosciences), Z3 (Compugen), Melanie (Swiss Institute of Bioinformatics), Mascot (Matrix Science), MS-Fit (University of California San Francisco), etc. Although notable improvements continue to be reported on these softwares and new tools allow us to apprehend better the protein contents of a sample, the underlying analysis procedure is still laborious, subjective, incomplete and dependent on the software even if the obtained results are usually acceptable compared to the operator's initial aims. For example, the 2-DE based proteomics analysis is still tiresome and we must manually control and correct many spots and instrument settings. It is subjective because the criteria used for the control and correction are more visual than quantitative and do not always integrate all knowledge available on the technique, in particular when it's a beginner operator who is confronted with this type of data. It is incomplete because major differences between test and reference gels are often observed, and much information is neglected (maybe it is more difficult to accede or to analyze). It is dependent on the used analysis software and its parameter settings because the operator does not always correct nor add omitted spots. Furthermore, the analysis is only worthwhile for a short time since there is no " retro-analysis " facilities for the archived data. Thus data are usually considered "definitive", and never reconsidered, or seldom, whereas the software of analysis became more powerful or that other operators have brought information on the correction of such or such previously obtained conclusions.

To address these problems, we have begun to investigate a new generation system which we describe as "collaborative ". It organizes and represents the data in a way that makes them accessible, verifiable and useful to other researchers, and provides tools to assist the researchers in cross multi-experiment data validation and particular feature findings. Thanks to the data and experience sharing, tremendous profits might be expected from the system.

2 Previous work and background

In contrast to the single activity system, collaborative system might be more useful but more difficult to design. Two major issues are involved: data integration and data exploration:

Data integration addresses data management problems such as data organization, storage and representation, efficient information integration and quick retrieval. A popular solution to this issue is to build a WEB based database. Many such systems in proteomics exist now in academic environment as well as those offered by commercial vendors. Along with experience on creation of proteomic database system accumulated over the last years, this issue begins to be well understood and we encounter less and less difficulties except standard representation and exchange of data [24, 16, 25, 23]. However, most these systems still remain in the stage of data library whose aims are to organize and store the data properly. Failing to review all the work, we cite [2] and [18]. The former introduced the *federated database* concept of creating and sharing distributed databases on the Internet primarily through the World Wide Web. Such federated databases are becoming increasingly popular because it allows effective data sharing without having to independently create, maintain or copy data. This greatly increases data availability and reduces the cost of using data. The later reported a kit of free software package called *Make2ddb II* which allows the organization, distribution and visualization of 2-DE centered proteomics data and of information available on the spots and proteins. An integrated search engine makes it possible to formulate some queries such as to obtain the identity of a spot or a list of gels on which figures a spot. This system can't be considered to be a proteomic resource

integration platform in a strict sense but rather a database management system. First, information managed is hardly sufficient for long-run inter-institute, inter-experiment data comparison and integration, and second the information research and retrieval capabilities, the integrative data analysis tools are somewhat limited. It is very difficult even impossible to store all information about a proteomic experiment such as the sampling mode, the culture conditions of the tissues, the conditions of realization of the electrophoresis, the parameters of gel image sampling and quantification, etc. Moreover, for multi-institute projects which produce significant volume of heterogeneous data, it becomes very difficult to explore the data without completely integrated tools.

Data exploration addresses the data processing issue. It tries to find relationships and patterns contained in the raw data and delivers structured results to the biologist [11, 4, 3]. Proteomic data exploration tools have been developed [19, 20, 15, 10, 7]. Lemkin *et al.* [19, 7] described a distributed gel comparison program based on Internet. It helps visual examination of two gel images located wherever in the World Wide Web. The system operates in an interactive mode. The user selects some landmark points inside both images and the system warps them with using morphing transforms and generates a synthesized image where gray-scale values are interpolated as well as the geometry. This is a pioneering work. It opened up the possibility of multi-experiment collaboration: one can visually compares his own experimental data with data from other laboratories even if the distance which separates them is about thousands kilometers. The project CAROL(*http://gelmatching.inf.fu-berlin.de/Carol.html*) retook the initiative and improved the warping process by integrating a new algorithm based on computational geometry.

Although these tools are very useful for multi-experiment and multi-experimenter data manipulation, they expose several major drawbacks. In particular, they focus on raw data manipulation and visualization, and offer no possibility for advanced concept manipulation like intrinsic biological pattern identification and protein relationship discovery in proteomic data. Since the nature of proteomic data is highly interrelated [22, 14, 17], such functions should be necessarily beneficial. For example, spots detected on different gels can correspond to a same protein, or its allelic forms, while different spots detected on the same gel can relate to a protein in different phosphorylation states etc. This information is acquired gradually by researchers, and usually in different projects. Building a system able to integrate these individual and disparate results could make it possible to benefit from all information available in the whole data. The system is not a data library any more but a resource and knowledge ware exploring the intrinsic relationships which exist in the proteomic data. Given a spot, we will have access to the list of the spots with which it was connected in other experiments, thus summarizing work of many people and produced from multiple projects. This also makes it possible to distribute new knowledge quickly, one can submit his results to the ware as soon as new information is discovered.

3 System design and deployment

PARIS exists as two software components (see Fig.1), each of which may run separately. The first component called *2-DE centered component* is responsible for managing data from 2-DE based analysis and the second is known as *MS centered component*, it manages data for protein identification based on mass spectrometry analysis. The two components are liked by using the

underlying biological context information, and share some common codebase resources.

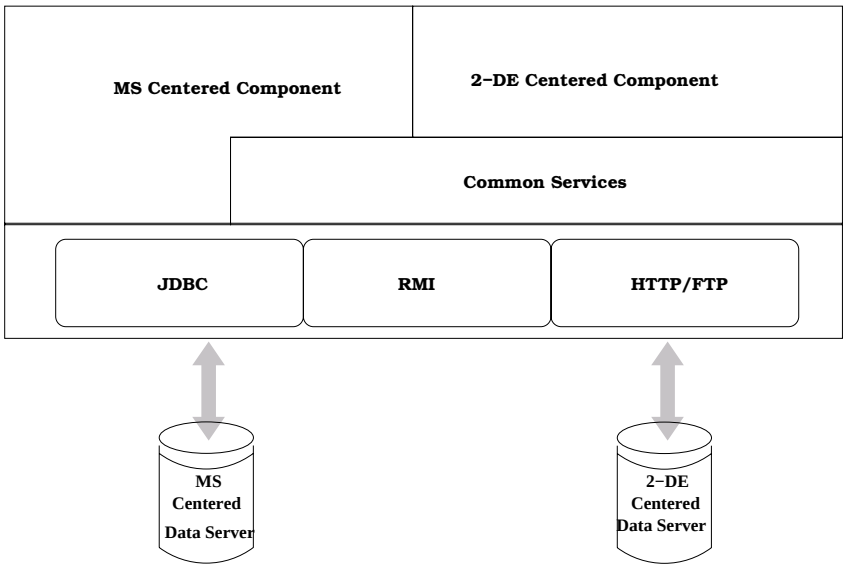


Figure 1: An overview of the PARIS system.

3.1 The 2-DE centered component

As shown in Fig.2, the 2-DE centered component has a 3-tier architecture and accordingly constitutes three major parts: a data server, a search and analysis engine and a graphical user interface. The data server was developed with the famous and free database management system PostgreSQL (http://www.postgresql.org/), it manages data with regard to the 2-DE experimentation including sample preparation, experiment design, parameter settings of the analysis procedures and result interpretation. It also stores information necessary to connect with external proteomic resources such as KEGG, UniPROT, trEMBL, etc. A table enables to describe the URL (Uniform Resources Locator) localization of external sources, their structure and relationships with information contained in the database, as well as the way we can use the information. For example, given the source KEGG, the description contained in PARIS states that it contains information about the metabolic pathways of protein and claims the constraints to respect when using such information as well as the relationships between that metabolic information and the data model implemented in system PARIS.

The search and analysis engine analyzes the queries formulated by biologist, provides fast, highly selective access to internal and external data sources, and structures the analysis results. From information provided by the biologist, it defi nes a search strategy taking into account the availability of local data and external sources, starts appropriate processing, and formats the results in order to facilitate its interpretation.

The 2-DE centered component includes also an advanced graphical user interface. Based on pure Java solution, this interface enables users to visualize, compare, annotate and correct or validate the data and results of analysis. It provides also most of the basic functionalities of an image manipulation tool such as zooming, scrolling, region-of-interest (ROI) selection and image contrast enhancement as well as spot emboss and gradient computing.

The system is particularly designed for maximizing the system usefulness. It exploits multi-resolution image concept and optimize information delivery over Internet. This prevents users from delay due to remote information fetch. On the other hand, PARIS also assists the users to formulate their queries . Users can either select queries from an established list or compose queries according to standard SQL syntax. Complex queries can also be formulated from elements directly selected on the images and be previewed before being submitted to the search engine.

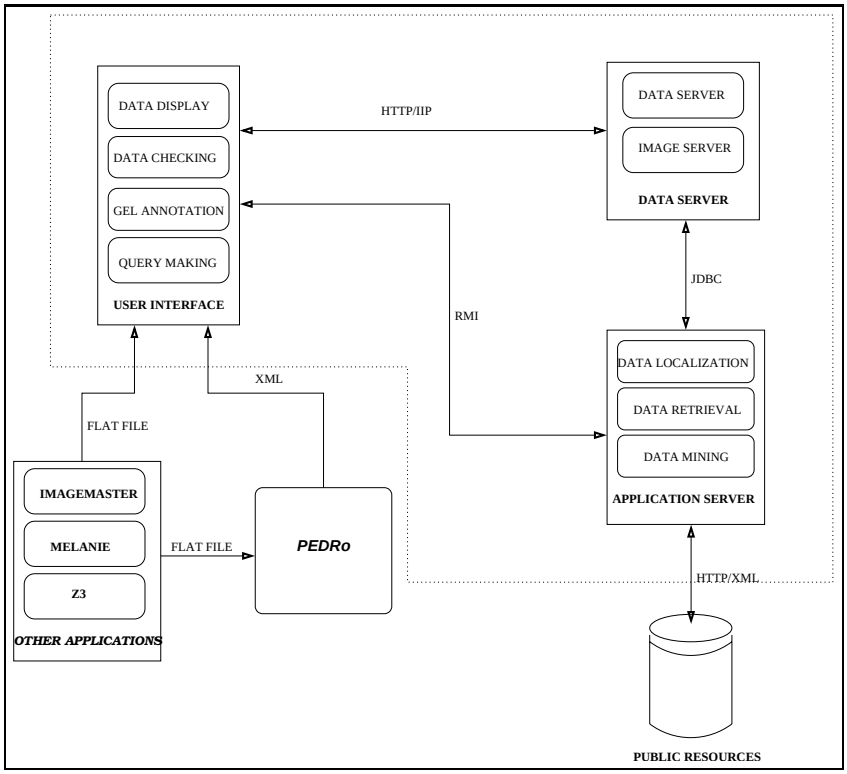


Figure 2: The architecture of the 2-DE centered component.

3.2 The MS centered component

The MS centered component consists of three mains parts too(see Fig.3) : a data manager, a data loading sub-module and a data analysis sub-module. Implemented in PostgreSQL, the data manager is responsible for MS related data organization and retrieval such as the raw spectra generated by the machine, the biological context information, and the parameters and procedures used for spectra filtering, protein identification and interpretation.

The data loading module contains a variety of data import and verification functions. It was designed to allow user (i) to convert internal legacy data to open standard format; (ii) to import the MS based analysis results into the PARIS system; (iii) to correct, validate and annotate the data and results. For the moment, only the data generated by our local proteomics core facilities are supported. These include a protein sequencer, a MALDI-TOF based analyzer, and some peptide mass fingerprinting tools (MS-Fit and MSCOT). The development of data import functions is being pursued to support most data formats known in proteomics laboratory.

There are two ways to place legacy data into PARIS. User can place manually his data and results piece by piece by filling out structured forms and fields. He can also batch the loading tasks into an automatic module. This module can extract the data and results either from flat files or from a MS-ACCESS database generated by the core platform, reorganize and upload them into the data servers of the PARIS system.

The data analysis module is an end-user module from where biologists can consult and interpret their data. They also use this module to export the data by generating standard compliant[1] XML files. Once integrated with the 2-DE centered module, the system enables us to make more integrative analysis.

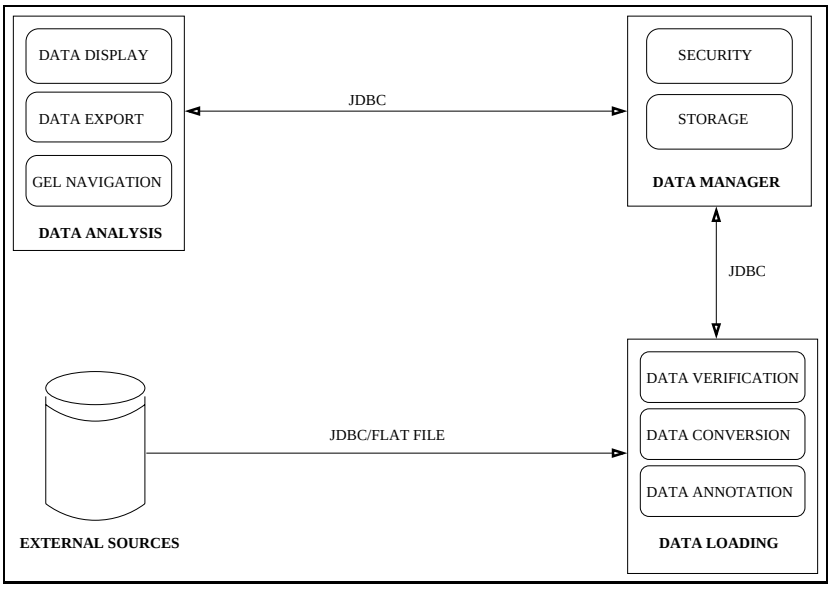


Figure 3: The architecture of the MS centered component.

3.3 Collaborative deployment

Both components of PARIS have been designed based on a client-server architecture. This makes very easy the deployment for inter-laboratory and inter-institute collaboration. As shown in Fig.4, we can build a network of collaborative systems from PARIS. The servers of this network gather and manage data, and meanwhile, the clients of PARIS consult, modify and interpret the data. This gives us a way to make benefit from the resources and experience of others. Scientists who use PARIS might integrate the newest knowledge into their experience and to communicate their results as soon as discovered.

PARIS is also a high scalable system. By replicating and distributing the PARIS servers as many times as necessary, each server can be administrated separately. a server does not need to manage all the available data, but only the local data and the communication with other servers. This makes it easy to accommodate heavier and higher loads. In fact, load growth means to add more server nodes, could have hardly impact on the efficiency of the global system.

[1] Most of the proteomics standardization efforts are still in progress (get the recent development of PSI from http://psidev.sourceforge.net), we will always try to integrate the last version of these standards whenever considered to be relatively stable.

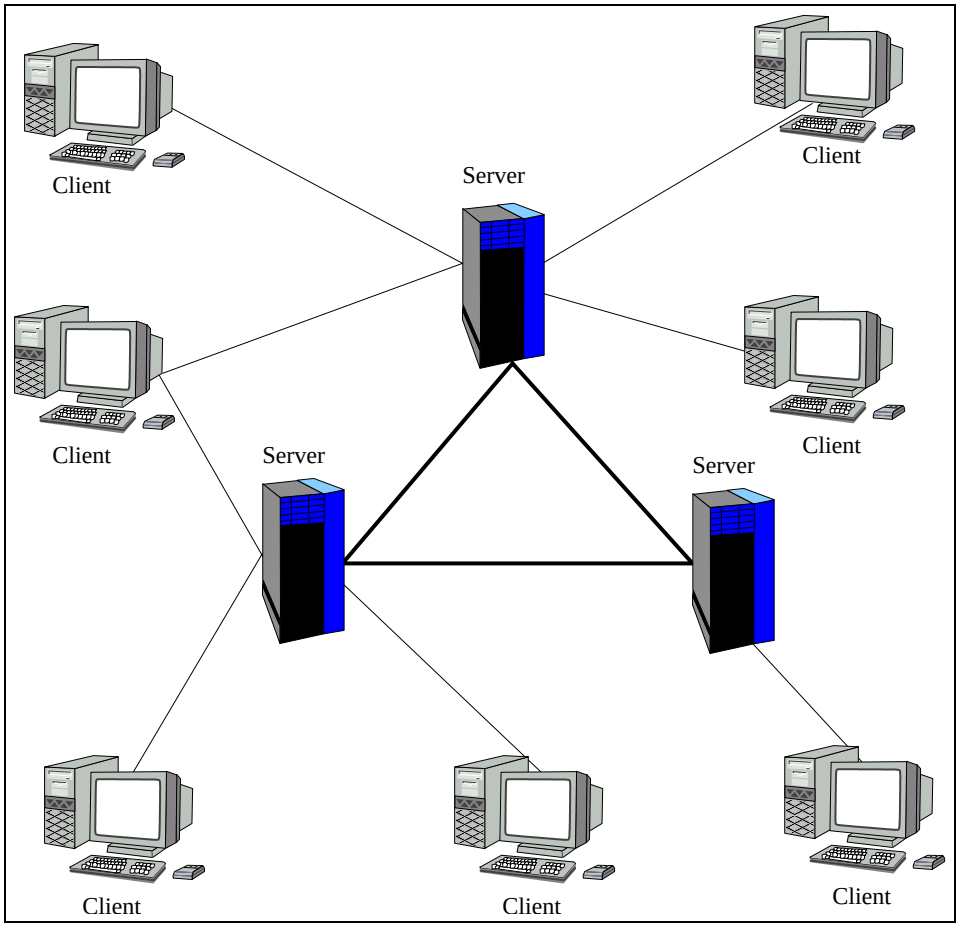


Figure 4: Network of collaborative systems build from PARIS. This makes it possible for us to exchange and share the data and experience rapidly.

4 Data modeling

Several proteomics data models have been published recently [25, 16, 24, 23]. Among them, PEDRo (Proteomics Experiment Data Repository) [25] recommended by the PSI (Proteomics Standard Initiative) is one of the most accepted by the proteomics community. It consists of fours mains blocks: biological context; separation techniques including 2-DE; mass spectrometry laboratory protocols; and mass spectrometry data analysis. Although PEDRo is designed to cover the different steps of the work-flow and a variety of techniques used in a proteomics experiment, it's fundamentally focused on the general aspects of the proteomics and therefor diffi cult to cope with the challenge of integrative and post-planned proteomic data analysis.

Fig.5 shows the data model implemented in our system. It provides a consistent framework for representing and handling information that is useful for integrative and post-planned analysis. We have focused on the description of the relevant data and their relationships, and seek to describe the experiments as precise as possible. Indeed, if we would analyze the data in long-run manners, it is essential to have a good knowledge of the experimental conditions on all the levels: conditions of culture of the tissues, protocols of extraction of proteins, parameters and procedures used in data processing, techniques and methods used in expression quantifi - cation and protein identifi cation, etc. It is the availability of these informations which may

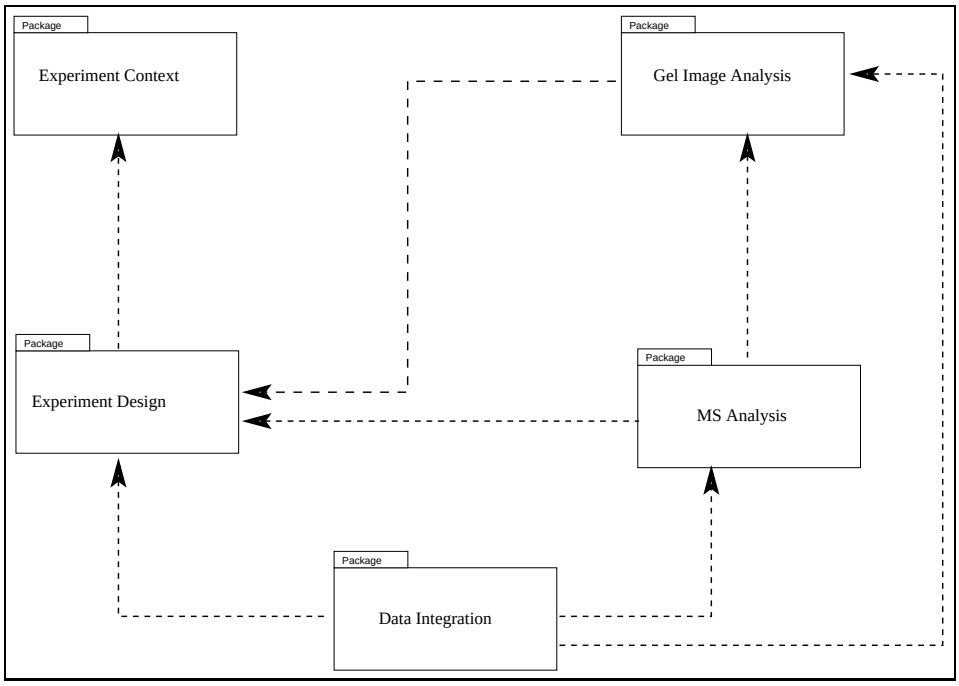
make the global analysis effective [5]. We should compare only data obtained under compatible conditions because the validity of biological interpretation depends on it closely. However, proteomics is known for its complexity in measurement and for its diversity in techniques, methods and protocols. Detailed description of the proteomics work-flow seems to be hardly compatible with the universality of a model. We thus limited our investigation to the two absolutely central techniques for proteomics analysis [1, 12]: the two-dimensional polyacrylamide gel electrophoresis (2-DE) for protein separation coupled with mass spectrometry (MS) for the protein identification.

On the other hand, proteomics data standardization is becoming an increasingly significant issue. Data standardization is the trend and seems to be one of the most important conditions necessary to ensure the durability of a system for proteomics data integration. PARIS is a standard compliant system in the sense that we can export our data by generating XML files compliant with the PSI recommended standards such as PEDRo, mzData and in the future mzIdent.

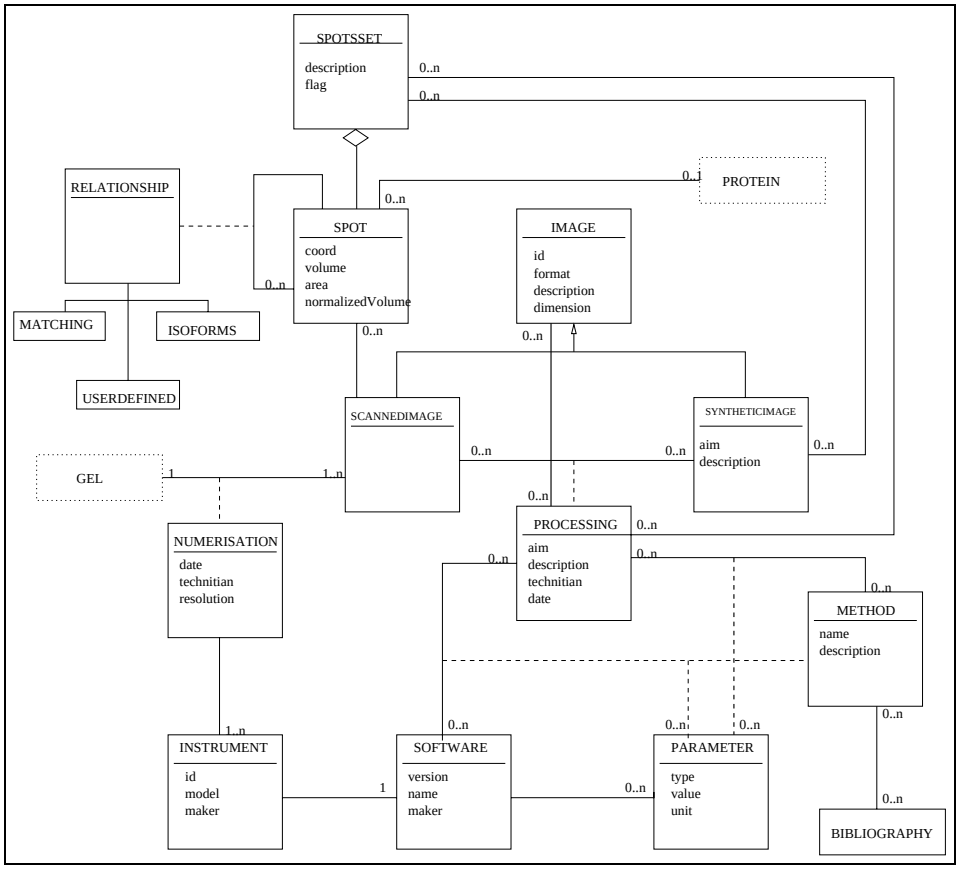
To keep the model maintainable, we have divided it into packages. Each package describes a set of closely related information that can be shared with other packages. The experiment context package captures data about the laboratory information management such as who, where, when and why is an experiment carried out? The experiment design package manages data about the experimentation itself: with which material, according to which protocol, and for which hypothesis is the experiment performed? This package is shared by the two main components of PARIS (2-DE centered and MS centered components) and it ensures the linkage between these two components. The gel image analysis package captures the parameters and procedures used for gel image analysis and protein expression interpretation. A series of gels may be produced by repeating the same experiment to ensure the reliability of the data, a gel may be re-scanned as many times as necessary and the produced image may be analyzed with different parameters and procedures. Beyond the model defined in PEDRo, our model allows detailed description of used gel image analysis softwares, procedures and parameters settings for data processing. The MS package manages data about MALDI-TOF based spectra acquisition and peptide fingerprinting based protein identification. Once more, our model captures more details about the procedures used in data processing and protein interpretation. The last package is the data integration package thanks to which the system interconnects with other public resources such as GENBANK, UniPROT, and KEGG, etc. Given a resource, we store information about its URL localization and entrance code as well as the rules we must obey for its exploration.

5 Functions and features

The most important feature of PARIS is its “collaborative” nature. The system consists of two main parts and affects two types of users: data provider and data consumer. Users who wish to make share their data and experience are data providers. They will run the server application that controls access to the resources for which it is responsible and mediates connections with other servers. The server is up at all times, creating a permanent location on the network. Thus, it becomes a persistent presence on the net for information storage and knowledge exchange. On the other hand, users who wish to use the data of others are data consumers. They just need to run a client application. Automatically deployed by Java Web Start technology, this application is accessible just by a click on a hyper-link within a Web page. There is no need for the



(a)



(b)

Figure 5: A partial view of the conceptual data model implemented in PARIS. The model is represented with the industry-standard UML conventions (a) the package diagram. The system PARIS encompasses five packages, and each package describes a set of closely related information. (b) A detailed description of the gel image analysis package.

users to manually download and install the application. Data are available to any scientist having Internet connection. In addition, sophisticated tools taking advantage of technology shifts were and will continue to be integrated seamlessly into the system. It gives us the possibilities to analyze new data by using archived data or to analyze private data by using public data. The benefits are clear [9, 17]. Scientists using PARIS will be able to integrate the newest knowledge into their experiments and to communicate their results as soon as discovered.

For knowledge sharing system, data access control is a critical issue. It is ensured here by using a UNIX-like user group management policy in which users are divided into three groups: knowledge "owner", owner's group members and other users not in the previous groups. Thanks to this policy, the data providers can decide whether to keep their data private or on the contrary to open their data to other communities.

PARIS can not only process data stored in its distributed servers, but it can also query on the fly other public resources (KEGG, UniPROT, trEMBL) in order to fetch the data that we need. For example, given the name of a protein and an organism, the system can query the KEGG database and find the metabolic pathways in which the protein is implicated. For each pathway found, PARIS searches the public databases UniPROT and trEMBL in order to compute the theoretical pI and Mw values of the proteins co-regulated in this pathway, and finally projects them on an experiment generated gel. This procedure enables us to highlight which pathway is activated on the gel, show where the missing proteins are virtually located and eventually their curated expression (calculated from cordon usage). It provides tools for the biologists to confront their experiment generated data with the theoretically derived ones.

An equally important feature of PARIS is its ability for data mining. The queries implemented in PARIS are not limited in means for accessing the needed entries in the database, they also discover relations that were not explicitly noticed in the data. Examples include: isoform proteins searching, spots matching deduction, matched spots ramification highlighting, etc. All these queries need specialized algorithms.

The third important feature of PARIS is its capability to analyze data in an integrated and post-planned manner. From a set of new data, we can query the data servers (local or remote) for particular findings (genomic information, experimental conditions, etc.) and compare the findings with the new data. We can also annotate and analyze together the archived and new data. This helps us to confirm or on the contrary, cancel conclusions obtained previously. By grouping results from different investigations, we can expect to generate new findings.

An other important feature implemented in the system concerns the proteomic data standardization issue. Based on the framework defined by PSI, PARIS provides a pure XML solution for data exchange. Data emanating from different gel image analysis and protein identification softwares are edited by using tools which explores an adapted XML schema and generates a XML data file [25, 23, 13]. Finally, it automatically parses the generated XML data file and uploads the data into its servers. Of course, PARIS can also export its data in XML format compliant partially with Pedro, mzData and mzIdent. We must notice that, for the foreseeable future, integrated proteomic information systems will have to cope with a variety of different data formats and its standardization. XML based models will be eminently the natural choice. By envisioning an approach based on XML, we will ensure the evolution and thus relevance of the PARIS system.

6 Applications

Post-planned analysis of *L. lactis* 2-DE data. One of the major applications of PARIS is post-planned analysis of 2-DE data. Proteomics analysis is well known for its iterative and incremental nature. Data and knowledge are usually acquired gradually by researchers, and in different projects. Previously accumulated data should be re-investigated with new knowledge to guarantee their validity, and new data should be always analyzed with previously prepared data to ensure their consistency.

Fig.6 shows an example of post-planned analysis of 2-DE data in *L. lactis* proteome profiling during its development in milk [8]. A simple visual re-examination of previously prepared data with recently acquired information makes us find some important spots (proteins) omitted in previous analyses.

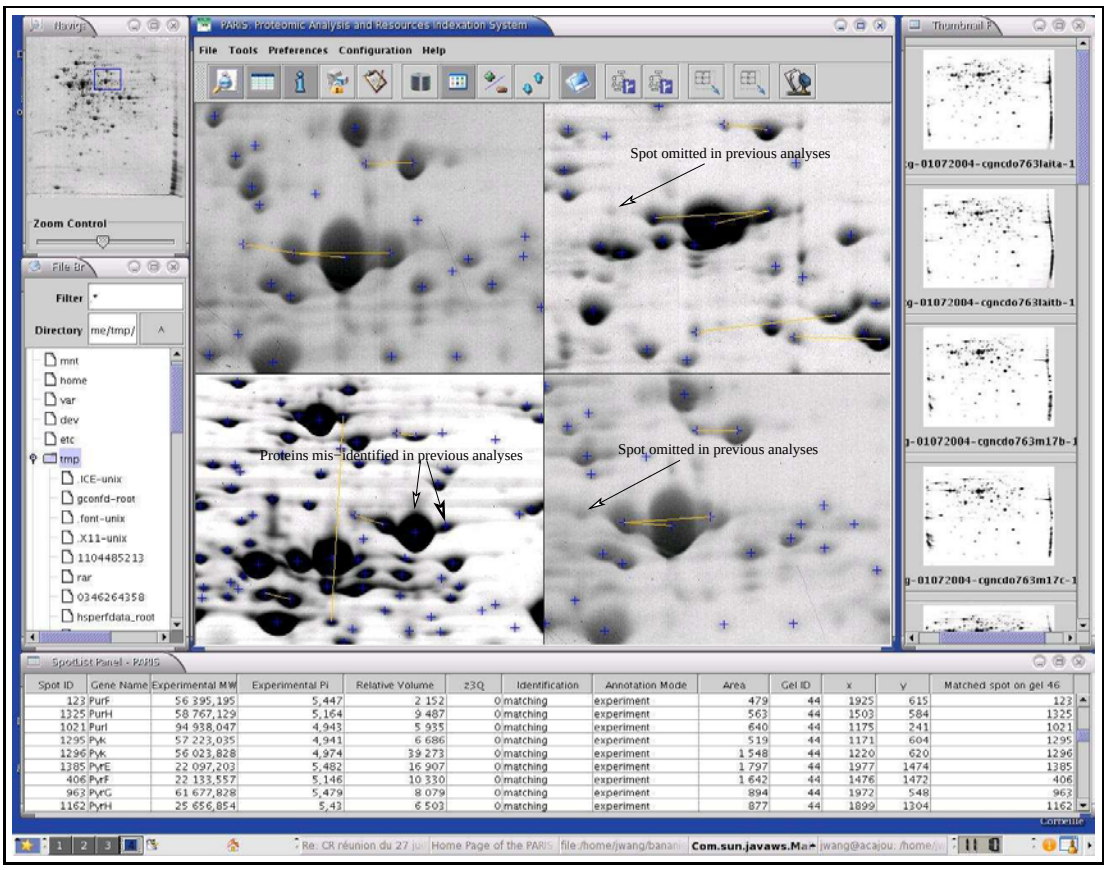


Figure 6: Post-planned analysis of the *L. lactis* polyacrylamide gel electrophoresis data with PARIS. In this example, we used new data to re-analyze the previously acquired ones and found some important spots omitted and/or proteins mis-identified in previous analyses. The relationships illustrated here on the gels correspond to protein iso-form relationship and the found omitted spots and mis-identified proteins are related to the *pyk* gene products.

Integrative analysis of 2-DE data with metabolic pathway information. This function consists in presenting features which can be useful, but does not have real existence in the database and can be derived in association with other public resources. It is about suggestions made from configurations or relationships discovered by synthesizing different types of information found in private and public resources. Fig.7 illustrates a situation of pathway integration in which,

given a protein p and a gel g, we synthesize an image to compare the *in silico* and experimentally found positions of the co-regulated proteins implicated in a metabolic pathway. This gives us a way to control our experimental data with some theoretical ones, if bias found, to make deeper investigation on where and why the *in silico* and experimental data are not consistent, and therefor to make benefi t from other proteomics data for experiment design.

This function is achieved by a series of basic database queries. We fi rst search from KEGG all pathways containing protein p. For each found pathway, we then identify the co-regulated proteins, and calculate their theoretical pI and Mw values, and fi nally project these values on to the original gel image. The achievement of this process requires to explore data contained in several public data resources such as GENBANK, UniPROT and trEMBL.

The diffi culty related to this process relays primarily on the fact that signifi cant distortion might be observed between the gel and the theoretically calculated pI and Mw values. Furthermore, we work in multi-experiment context. Given a biological sample, the associated proteins might be found in images scanned using different resolution or derived from gels associated with different pH gradients. All these factors complicate the virtual spots positioning. The positioning procedure must take into account the local and global deformations observed in gel image g. We have used a method of tessellation: given a pair of pI and Mw values, we fi rst delimit a zone containing the protein and then deform it according to an affi ne transformation so that the protein is embodied perfectly into the real gel image.

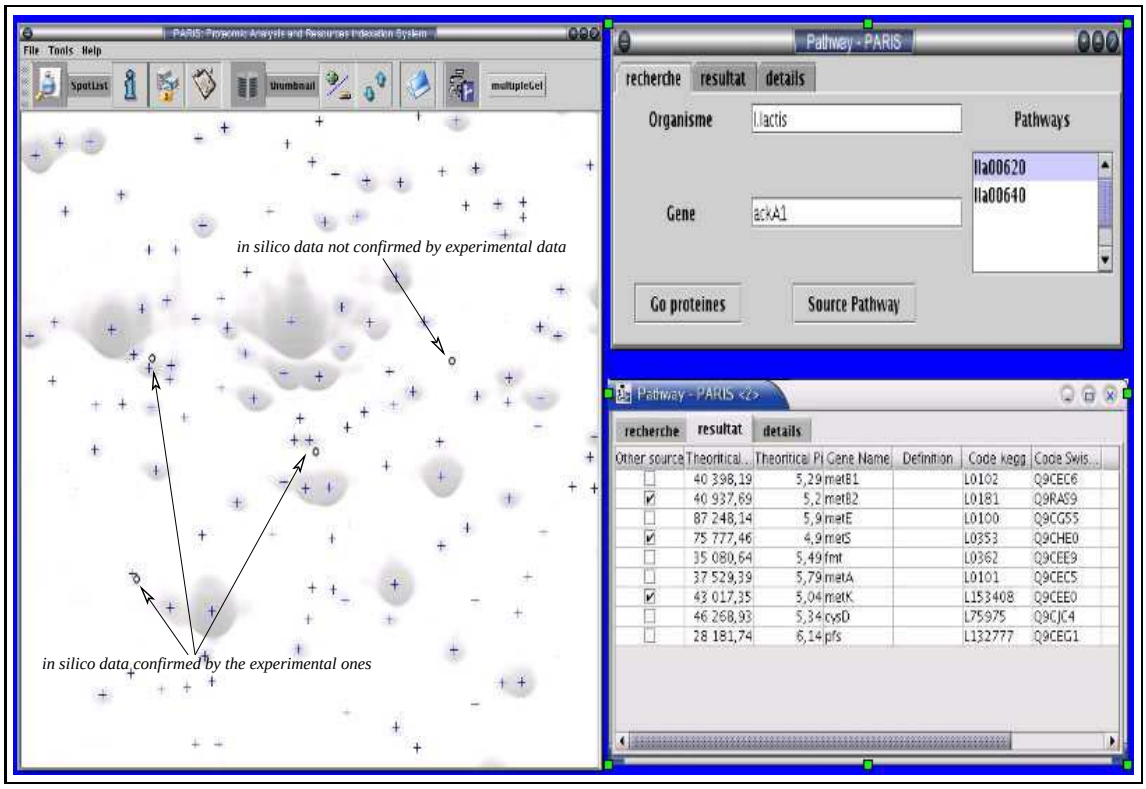


Figure 7: Data integration with PARIS. We illustrate here an example of how to use pathway information found in KEGG to analyze the gel centered data. In this example, the experiment generated data are marked by a cross and the *in silico* generated data are marked by a circle.

7 Conclusions

Proteomic data integration which involves unifying data that are scientifically related, but originate from unrelated sources, is a central topic of the post-genomics era. Herein, we have developed a network based system which makes it beneficial to re-use as much as possible the information emanating from multiple experiments and multiple experimenters.

Like system YPRC-PDB presented in [6], our system gone beyond the raw data management and visualization. It organizes numerous data emanating from a proteomic experiment like sampling, protein quantification and identification, etc. We seek to describe the experiments as precise as possible. Indeed, if we would compare the expression level of certain proteins in different physiological situations, it is essential to have a good knowledge of the experimental conditions on all the levels: conditions of culture, protocols of extraction, parameters of the electrophoresis, techniques of quantification, spectra data processing and protein identification, etc. It is the availability of these informations which may make the comparison effective. We should compare only the data obtained under compatible conditions because the validity of the biological interpretation depends on it closely.

Centered at such a biological context and data relationship focused database, we furthermore equipped our system with some capacities for data exploration. It supports interface with specialized gel image analysis and protein identification softwares, facilitates visual verification and correction of the analysis results, and provides tools for advanced concept manipulation like virtual gel synthesis, protein expression comparison, implicit spot matching deduction, and cross multi-experiment, multi-experimenter data validation. These functions are vital for proteomics. In fact, on the scope of a collaborative analysis, information contained in a gel isolated from others becomes no significant, important is that contained in a group or a set of gels. The mathematical formulation of these interrelated concepts makes it possible to cluster proteomics information according to some criteria like geometrical, photometric (spot profiling and patterns), physiochemical (molecular polarity and weight), and functional, etc. Consequently, it becomes possible and even primordial for the biologist to be able to exploit these relationships in order to predict new relations or feature findings.

PARIS has been designed and was being implemented using an object-oriented approach in pure Java. Other than standard Java libraries and Swing, PARIS uses JAI (Java Advanced Imaging) for the basic image manipulation and analysis, and JDBC (Java Database Connection) for database management. Communication between the client and the server is handled via Java's RMI (Remote Method Invocation) technology while the communication with external servers is based on the standard HTTP protocol. This makes it easy to deploy the application on standardized station equipped with a simple HTML browser, and hence minimizes the deployment and maintenance effort.

With this system, we can easily control and cross-validate data stemming from different specialized analysis softwares, test and evaluate their consistency, make possible the re-examination of data analyzed at a former date or by different experimenters. The permanent enrichment of the database could then allow us to benefit from the whole data. This constitutes our first stage towards an effective data integration system.

References

[1] N Leigh Anderson, Alastair D Matheson, and Sandra Steiner. Proteomics: applications in basic and applied biology. *Current Opinion in Biotechnology*, 11:408–412, 2000.

[2] R. D. Appel, A. Bairoch, J. C. Sanchez, J. R. Vargas, O. Golaz, C. Pasquali, and D. F. Hochstrasser. Federated 2-de database: a simple means of publishing 2-de data. *Electrophoresis*, 17:540–546, 1996.

[3] Pierre Auger and Christophe Lett. Integrative biology: linking levels of organization. *Comptes Rendues Biologies*, 326:517–522, 2003.

[4] Raji Balasubramanian, Thomas LaFramboise, Denise Scholtens, and Robert Gentleman. A graph-theoretic approach to testing associations between disparate sources of functional genomics data. *Bioinformatics*, 20(18):3353–3362, 2004.

[5] Mark S. Boguski and Martin W. Mclntosh. Biomedical informatics for proteomics. *Nature*, 422:233–237, 2003.

[6] S.Y. Cho, K.S. Park, J.E. Shim, M.S. Kwon, K.H. Joo, W.S. Lee, J. Chang, H. Kim, H.C. Chung, H.O. Kim, and Y.K. Paik. An integrated proteome database for two-dimensional electrophoresis data analysis and laboratory information management system. *Proteomics*, 2(9):1104–1113, September 2002.

[7] Lemkin P. F., Thornwall G., and Evans J. *Protein Protocols Handbook*, chapter Comparing 2D Electrophoretic Gels Across Internet Databases, pages pp 279–305. Humana Press, 3nd edition, 2005.

[8] Christophe Gitton, Mickael Meyrand, Juhui Wang, Christophe Caron, Alain Trubuil, Alain Guillot, and Michel Yves Mistou. Proteomic signature of Lactococcus lactis NCDO763 cultivated in milk. *Applied and Environmental Microbiology*, 71(11):7152–7163, November 2005.

[9] C. S. Goh, N. Lan, N. Echols, S. M. Douglas, D. Milburn, P. Bertone, R. Xiao, L. C. Ma, D. Zheng, Z. Wunderlich, T. Acton, G. T. Montelione, and M. Gerstein. Spine 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Res.*, 31(11):2833–8, June 2003.

[10] Bensmail H, Golek J, Moody M, Semmes JO, and Haoudi A. A novel approach for clustering proteomics data using bayesian fast fourier transform. *Bioinformatics*, 21(10):2210–2224, 2005.

[11] John P. Helfrich. Raw data to knowledge warehouse in proteomic based drug discovery: a scientifi c data management issue. *Computational Proteomics Supplement*, 32:48–53, March 2002.

[12] Bent Honoré, Morten Ostergaard, and Henrik Vorum. Functional genomics studied by proteomics. *BioEssays*, 26(8):901–915, 2004.

[13] Proteomics Standard Initiative. mzdata xml schema documentation, version 1.05 (http://psidev.sourceforge.net/ms/xml/mzdata/mzdata.html), 2005.

[14] Mellor J., Yanai I., Clodfelter K., Mintseris J., and DeLisi C. Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.*, 30(1):306–309, 2002.

[15] Donald J. Johann, Michael D. McGuigan, Amit R. Patel, Stanimire Tomov, Sally Ross, Thomas P. Conrads, Timothy D. Veenstra, David A. Fishman, Gordon R. Whiteley, Emanuel F. Petricoin, and Lance A. Liotta. Clinical proteomics and biomarker discovery. *Annals of the New York Academy of Sciences*, 1022:295–306, 2004.

[16] Andrew Jones, Jonathan Wastling, and Ela Hunt. Proposal for a standard representation of two-dimensional gel electrophoresis data. *Comparative and Functional Genomics*, 4:492–501, 2003.

[17] Prince J.T., M.W. Carlson, R. Wang, P. Lu, and E.M. Marcotte. The need for a public proteomics repository. *Nature Biotechnology*, 22:471–472, 2004.

[18] Mostaguir K., Hoogland C., Binz P. A., and Appel R. D. The Make 2D-DB II package: conversion of federated two-dimensional gel electrophoresis databases into a relational format and interconnection of distributed databases. *Proteomics*, 8:1441–1444, August 2003.

[19] P. F. Lemkin. The 2DWG meta-database of two-dimensional electrophoretic gel images on the internet. *Electrophoresis*, 18:2759–2773, 1997.

[20] D.H. Lundgren, J. Eng, M.E. Wright, and D.K. Han. Proteome-3D: an interactive bioinformatics tool for large-scale data exploration and knowledge discovery. *Mol. Cell. Proteomics*, 3:1374–1376, 2003.

[21] Jane M. C. Oh, Samir M. Hanash, and Daniel Teichroew. Mining protein data from two-dimensional gels: Tools for systematic post-planned analyses. *Electrophoresis*, 20(4-5):766–774, May 1999.

[22] Kemmeren P., van Berkum N., Vilo J., Bijma T., Donders R., Brazma A., and Holstege F. Protein interaction verifi cation and functional annotation by integrated analysis of geneme scale data. *Mol. Cell*, 9:1133–1143, 2002.

[23] Per Gärdén, and Rikard Alm and Jari Häkkinen. PROTEIOS: an open source proteomics initiative. *Bioinformatics*, 21(9):2085–2087, 2005.

[24] Romesh Stanislaus, Liu Hong Jiang, Martha Swartz, John Arthur, and Jonas S. Almeida. An XML standard for the dissemination of annotated 2D gel electrophoresis data complemented with mass spectrometry results. *BMC Bioinformatics*, 5(9):1471–2105, January 2004.

[25] C. F. Taylor, N. W. Paton, K. L. Garwood, P. D. Kirby, D. A. Stead, Z. Yin, E. W. Deutsch, L. Selway, J. Walker, I. Riba-Garcia, S. Mohammed, M. J. Deery, J. A. Howard, T. Dunkley, R. Aebersold, D. B. Kell, K. S. Lilley, P. Roepstorff, J. R. Yates III, A. Brass, A. J.P. Brown, Ph. Cash, S. J. Gaskell, S. J. Hubbard, and S. G. Oliver. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotechnology*, 21(3):247 – 254, March 2003.