BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 1 2004, pages 133–135
DOI: 10.1093/bioinformatics/btg389

PARIS: a proteomic analysis and resources indexation system

Juhui Wang[1,*], Christophe Caron[1], Michel-Yves Mistou[2], Christophe Gitton[2] and Alain Trubuil[1]

[1]INRA, Unit of Biometrics and [2]INRA, Unit of Biochemistry and Proteins, 78352 Jouy-en-Josas, France

Received on March 6, 2003; revised on May 28, 2003; accepted on June 11, 2003

Downloaded from https://academic.oup.com/bioinformatics/article/20/1/133/229125 by guest on 18 April 2025

ABSTRACT

Summary: We developed a system for managing data from two-dimensional electrophoresis-based proteomic experiments. Named PARIS, the system stores gel image and information about experiments and analysis procedures, allows the user to search and navigate in genomic and proteomic data, supports visual verification and validation of the analysis results, and provides tools for cross multi-experiment and multi-experimenter data validation and exploration.

Availability: The software is freely available from http://www.inra.fr/bia/J/imaste/Projets/PARIS/index.html

Contact: juhui.wang@jouy.inra.fr

Computer-assisted processing is essential for two-dimensional electrophoresis (2DE)-based proteomic analysis because of the complexity of information present in 2DE images. Even if notable improvements continue to be reported on automatic tools, proteomic data analysis is still laborious and requires manual intervention of biologists for correction and validation of the provided results. Because of the complexity of information to be analysed and the eventual subjectivity related to the context of a particular project, it seems that an approach based on only one operator–one workflow model operating independently of the other existing platforms is not a very efficient nor effective operating mode compared with the massive amount of knowledge and information contained in the data. We can thus expect tremendous benefits from the advantages of an approach, which we describe as 'collaborative'. It organizes and represents the data in a way that makes it accessible, verifiable and useful to other researchers, and provides tools to assist the researchers in cross multi-experiment data validation and particular feature discovery.

In contrast to single activity systems, collaborative systems might be more useful but more difficult to design. Two major issues are involved: data sharing and data exploration. Data sharing addresses data management problems, such as data structure, storage and representation, efficient information integration and quick retrieval. A popular solution to these problems is to build a WEB-based proteomic database. Many such systems now exist, e.g. SWISS-2DPAGE from http://www.expasy.org/ch2d/, dbEngine from http://www-lecb.ncifcrf.gov/Software/dbEngine.html and EBP from http://www.mpiib-berlin.mpg.de/2D-PAGE/. 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 (Helfrich, 2002). Although these systems are very useful for multi-experiment and multi-experimenter data manipulation, they have major drawbacks. In particular, they focus on raw data retrieval and visualization, and offer no possibility for advanced concept manipulation like intrinsic biological pattern identification and protein relationship exploration in proteomic data. Since the nature of proteomic data is highly interrelated, such functions should be necessarily beneficial.

Relying on these observations, we have developed a new proteomic data management system. Like PEDRo (Taylor *et al.*, 2003) or YPRC-PDB (Cho *et al.*, 2002), this new generation system goes beyond simple data management and is expected to become a foothold for knowledge sharing and exchange between researchers. It organizes numerous data emanating from a set of proteomic experiments like sampling, culture condition, image quantification, protein expression, genomic and metabolic data, and provides tools for data navigation and exploration, such as identification of expression level changes, image-based information query, cross multi-experiment data validation and particular feature discovery.

Figure 1 shows the overall architecture of the system, it has a 3-tier architecture and comprises three major parts: a PostgreSQL-based database server, an integrated research engine and a graphical user interface. The database server manages data regarding experimentation, including sample preparation, parameter settings of analysis procedures, through to the biological interpretations of the results. A notable feature of our system is that we seek to describe the experiments as precisely as possible. Indeed, if we compare data from different physiological situations, it is essential to have a good knowledge of the experimental conditions on all levels. It is the availability of this information which may

*To whom correspondence should be addressed.

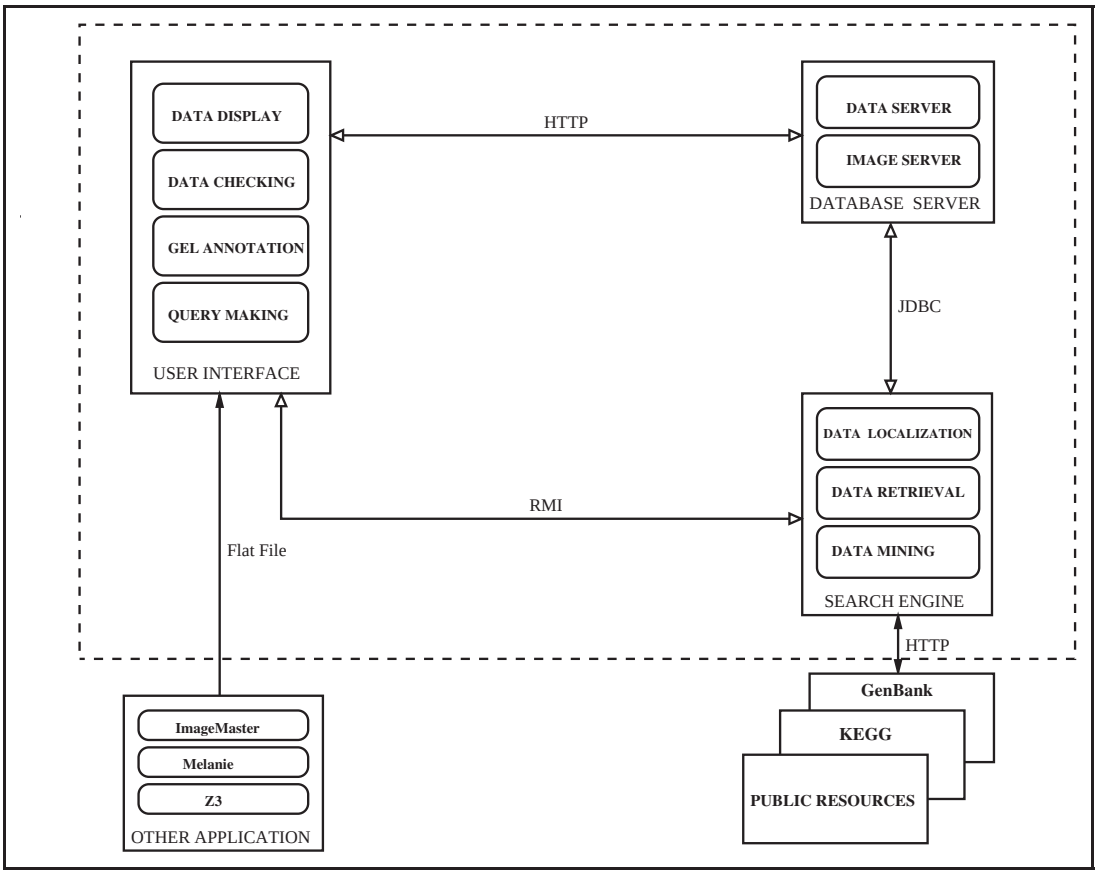

Fig. 1. System overview.

make the comparison effective. We should only compare data obtained under comparable conditions because the validity of the biological interpretations depends on it closely.

The search engine analyses the queries formulated by the biologist, provides fast, highly selective access to internal and external data sources, and structures the found results. From information provided by the biologist, it defines a search strategy taking into account the availability of local data and external sources, transforms the information into SQL or HTTP statements and forwards them to appropriate data servers (data server, image server or externals resources). Results found by the different servers are first collected and then structured by the search engine in order to facilitate their reading and interpretation. To take on this role of hub, the search engine holds a meta-data model about the structure of data managed by PARIS. Currently, we manage only information necessary to connect with external genomic and proteomic resources, such as GenBank (http://www.ncbi.nlm.nih.gov/Genbank), SWISS-PROT (http://www.expasy.org/sprot) and Kegg (http://www.genome.ad.jp/kegg). A table describes the geographic localization of the external sources, their structure and the relationships with information contained in the database, as well as the way we can use the information. For example, given a source *S*, the description could state that *S* contains information about the metabolic pathways of some proteins and depicts the constraints to respect when using this information as well as the relationships between that metabolic information and the data model implemented in the PARIS system.

An important feature that distinguishes our system from others is that queries implemented in PARIS are not limited to access the needed entries in the database, relations that were not explicitly noticed in the data may be suggested to the biologists for validation. Such functions make it possible to draw the biologist's attention to eventual inconsistencies contained in the data. For example, a typical query frequently run by our biologist partners concerns metabolic pathway integration: given the name of a protein and an organism, PARIS is able to find the gels that are not compatible with one or any pathway defined in Kegg, and to highlight the virtual position of missing proteins on the gels. This function makes the time-consuming and tiresome gel correction and annotation tasks easier.

PARIS also includes an advanced graphical interface. Based on Java Advanced Imaging (JAI), this interface enables users to visualize, compare and correct or validate the results of

Downloaded from https://academic.oup.com/bioinformatics/article/20/1/133/229125 by guest on 18 April 2025

analysis. It provides 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. Other functions help us to understand the relationships that exist in the proteomic data. The spot over which the mouse cursor is situated is highlighted, and the associated genomic and proteomic information is displayed. Given a term such as spot, protein, physico-chemical characteristics, etc., we can retrieve all gels having relationships with this term, and display them in a set of windows. Furthermore, particular relationships such as spot matching, deduced spot matching and matched spot ramification are highlighted by using pseudo-colours. These functions are simple but meet the practical requirements of the biologist. Note that PARIS also imports results obtained from most gel image analysis software found in biology laboratories, e.g. Z3, ImageMaster, Melanie. The import is carried out via the user interface. This solution gives users the possibility to correct and validate the data before being submitted to the data servers.

PARIS is particularly designed for maximizing the system's usefulness. It exploits tilled image concept promoted by the Internet Imaging Protocol (IIP) initiative and optimizes information delivery over Internet. This avoids users having delays due to remote information fetch. On the other hand, PARIS also assists the users in formulating their queries. Users can either select queries from an established list or compose queries by combining previous queries. Complex queries can be formulated from elements directly selected on the images and be previewed before being submitted to the search engine.

REFERENCES

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

Helfrich,J.P. (2002) Raw data to knowledge warehouse in proteomic based drug discovery: a scientific data management issue. *Comput. Proteomics Suppl.*, **32**, 48–53.

Taylor,C.F., Paton,N.W., Garwood,K.L., Kirby,P.D., Stead,D.A., Yin,Z., Deutsch,E.W., Selway,L., Walker,J., Riba-Garcia,I. *et al*. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat. Biotechnol.*, **21**, 247–254.

Downloaded from https://academic.oup.com/bioinformatics/article/20/1/133/229125 by guest on 18 April 2025