Proceedings of the 25th Annual International Conference of the IEEE EMBS
Cancun, Mexico • September 17-21, 2003

CellMapBase - An Information System Supporting High-Throughput Proteomics for the Cell Map Project

Zsuzsanna Bencsath-Makkai[1], Alex Bell[3], John Bergeron[3], Daniel Boismenu[1], Robert Funnell[2], Mark Harrison[1], Catherine Mounier[3], Jacques Paiement[1], Line Roy[3], Robert Kearney[2]

[1]Montreal Proteomics Network, Montreal, QC, Canada, [2]Department of Biomedical Engineering, McGill University, Montreal, QC, Canada, [3]Dept. of Anatomy and Cell Biology, McGill University, Montreal, QC, Canada

Abstract– **Information systems supporting the protein discovery process encompass three worlds to model: the world of laboratory experiments, the world of special public libraries, and the world of applicable theories and methods for the derivation, validation and verification of the resulting protein set.**

In this paper we describe the motivation, the design and certain implementation considerations of CellMapBase, a Web-based, database-driven application designed to support the Cell Map project.

Keywords– information systems, proteomics

I. INTRODUCTION

The Cell Map, a Montreal Proteomics Network collaborative effort aims to map comprehensively the proteins in the mammalian cell at the level of the organelle. Proteins will be further categorized and annotated with respect to their location, function, structure, and interactions.

The paper is structured as follows. Section II describes the various technologies involved in the proteomics pipeline whereby biological samples are processed to yield the annotated and categorized protein list. Section III describes the objectives and design requirements underlying CellMapBase (CMB). Section IV provides an overview of the implementation and current status of the project. Section V summarizes the results and provides an indication of direction for further work.

II. THE PROTEOMICS PIPELINE

High-throughput proteomics requires the integration of a series of complex technologies that take the raw biological sample to an abstract list of proteins annotated according to their function, location, and interactions. Fig. 1 shows a simplified block diagram of the information flow in the pipeline.

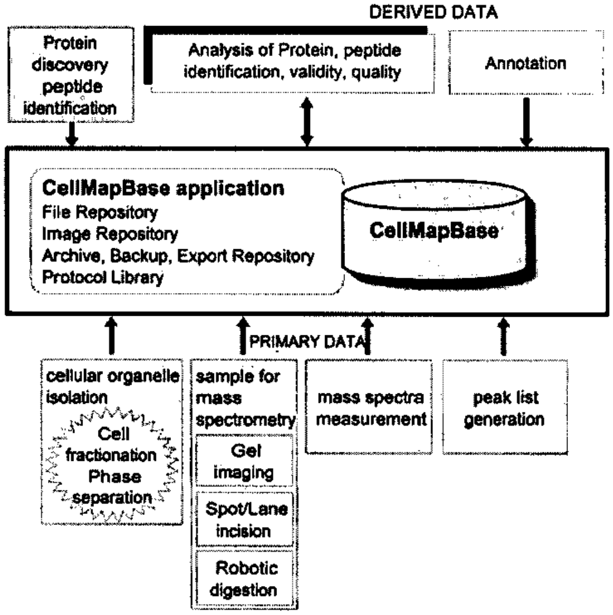

Fig. 1. Interactions between proteomics pipeline and CellMapBase

A. Sample Acquisition

The sample acquisition is the "biological" aspect of the process where the objective is to acquire a sample of biological tissue of interest in such a manner as to preserve its protein makeup.

B. Sample Fractionation

The sample, typically comprising part of an organ, is processed to separate particular organelles.

C. Separation

Proteins within the fractionated sample are separated on the basis of molecular weight and/or iso-electric point using 1D or 2D gels. Gel images provide an empirical measure of the protein distribution.

D. Digestion

Gels are cut into gel samples corresponding to bands (1D gels) or spots (2D gels). The gel samples are then digested with trypsin, a protease that cleaves proteins selectively,

0-7803-7789-3/03/$17.00 ©2003 IEEE

cutting them at lysine and arginine residues. The resulting fragments are small enough to be measured accurately with mass spectrometers.

E. Mass Spectrometry

The digested sample is analyzed by mass spectrometry. Typically the digest is further separated using an HPLC and then subjected to a two-stage analysis in which individual peptides are first identified, and then fragmented to generate the fragmentation spectra used for the identification.

F. Preprocessing and Peak Detection

The raw mass spectra are processed to eliminate noise, deconvolve isotopic effects and correct for different charge states. Peak-detection algorithms are then used to generate a list of peaks corresponding to the peptide fragment.

G. Peptide Identification

The fragmentation spectra are compared with theoretical spectra determined by analytically fragmenting the proteins in an appropriate protein database. In our case, we use MASCOT (MatrixScience) [1], which implements a probability-algorithm to score peptide identifications.

H. Protein Identification.

Proteins linked to peptides are identified with measurable certainty and sorted. The identification is based on the number of peptides linked to them and their scores. This will yield a redundant list of identifications in which the same peptides point to multiple proteins. Typically, this identification is performed using peptides from the entire experiment.

I. Clustering

Proteins are clustered to reduce the redundancy by generating disjoint sets of peptides from the identifications. Each cluster will contain all proteins that link to the set of peptides within the cluster. In effect, the clusters define the minimum number of proteins that can fully explain the peptides identified within the experiment.

J. Annotation

The clusters and proteins are annotated according to location, function, structure, and interactions.

III. CMB DESIGN REQUIREMENTS

The Cell Map project generates and analyses large amounts of experimental data; and requires comprehensive reporting on results and supporting evidence. Fig. 2 shows the information flow of the discovery process.

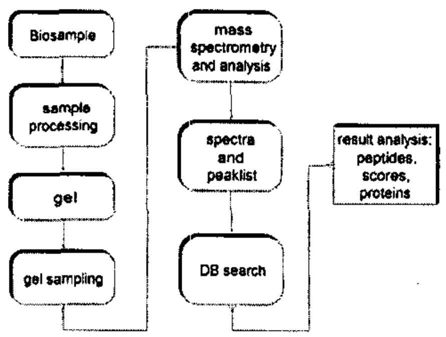
Biosample
sample processing
gel
gel sampling
mass spectrometry and analysis
spectra and peaklist
DB search
result analysis: peptides, scores, proteins

Fig. 2. Protein discovery process

The major functional requirements of the CellMapBase system were defined to be:

- Track a sample through the experiments and analyses.
- Collect and manage the experimental and derived data, and the documentation of the derivation methods.
- Build a protocol library from the protocols used.
- Build the PROFUSE non-redundant protein database storing and linking the qualified results from public and private databases along with their reference data.
- Provide support for annotation, for the comparison of annotations from different groups and for the integration of an annotation server.
- Provide support for different computational methods in result analysis, result quantification and validation.
- Provide on-line and ad-hoc reports and publishable results.
- Assist in project management.
- Enforce selective data retrieval, user and user access management. Authenticate users, authorize and control access to system modules, to operations and to data records.
- Provide for and administer the access-controlled hierarchical application menu system, and the interface system generation.

In other words, CMB should provide the mechanisms to support the Cell Map project, and dynamically define interface and functionality.

IV. IMPLEMENTATION

A. Structural modules

Fig. 3 shows the main modules of CMB. In a fairly natural way the main system modules were defined along the stations/processes of the pipeline:

- Laboratory Information Management System (LIMS) for data tracking of samples and experiments

- Data Miner and Publisher
- Protein Annotation Information Database of pathways, known and inferred functions, organelle and protein location, families
- PROFUSE including a Reference Database of retrieved and used references (as indexes)
- Tools and program resources such as visualization tools, vocabulary management subsystem to manage the Proteomics Vocabulary, Glossary, Dictionary and project-specific code list
- CMB Administrator for people and project management; database-driven, authorized-access-controlled on-line interface generation; and project-membership-dependent selective data retrieval

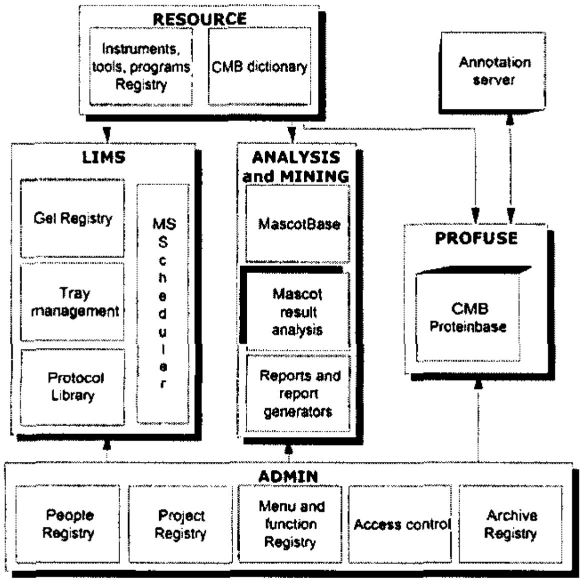

Fig. 3. The schematic, functional modules of CMB

B. Underlying Technologies

The most important considerations in selecting the technologies for this system were the ease of access for users and administrators, and a robust, reliable DBMS. The application has a 3-tier architecture: a thin client (a Web-browser), an application server and the DBMS. All interfaces are Web-based.

Technologies used are

- Web
- Commercial database management system (Oracle)
- Commercial application server (Cold Fusion Server of Macromedia)
- Public Protein databases, databanks
- Commercial mass spectra analyzer, peak detector and search engine (MassLynx of MatrixScience)
- Commercial mathematical programming and visualization software (Matlab)
- XML for data and report exchange
- barcode systems for unique sample and gel image identification

The choice of the application server was driven by the available expertise within the project developers.

We handle database crashing, backup and restore requests through the selection of a powerful commercial database, and through file, backup and archiving software.

Our system will need to interface with all public protein databanks and databases.

We foresee that the following trends will influence the CMB project:

- Access and analysis mode of the public libraries and databases
- Newly acquired hardware and software in the laboratories

C. Access Control

The system requires complex access control and user authentication. Data privacy options are implemented as multi-layer authentication and access-control mechanisms. The UsersRolesAccess module serves to

- selectively control user access to menu groups and application functions
- selectively control access to database records and structures
- create automatic log data of operations and operators

D. Development

There is a dynamic interaction between schema development, application development, and process improvement with newly devised methods and tools developed in-house.

CMB system

System development was driven by the needs of the scientific endeavour. We identified the bottlenecks along the pipeline and concentrated the development efforts on them. In parallell we identified those processes where we could, immediately, put into use data-collection forms conforming to the CMB schema. The data collected in these digital forms, are for example, historical data for the LIMS module, and were used to populate the People and Project Registries.

The main bottleneck on the pipeline was the complex, multi-thread process of capturing and validating the results of the discovery process.

Standards for annotations, even the definition of what comprises annotation, is in development by the international scientific community [2]. The CMB annotation of a protein sequence includes the classification of proteins into groups according to homology, specifying splice variants and fragments, links to Pub-Med leading to the relevant literature, and an assignment of at least one functional category and function features (EMBL)[3]. The automation of the Annotation process and its quality control are a major focus of in-house development.

The Menu system is mapped into the CMB schema. It is designed to be managed within CMB and be implemented as a perpetual query to the Menu Registry with the selection-condition being the result of a query to the UsersRolesAccess module that identifies the user, the usertype, the associated project set and the role played in those projects.

Reporting and Mining provides controlled access to the protein identification results. This module is planned to become the basis of the CMB warehouse to be located on a separate node.

Methods and tools developed in-house
The analysis of the parsed MASCOT data feeds the database with accountably characterized data making it possible to include a quantification of the quality of identification, of motifs, exact matches, etc. This is the base data-populating mechanism of the non-redundant, integrated protein base PROFUSE.

Another focus of in-house development is to automate the querying of public databases in a datafile-driven batch-mode, and to automate the extraction and feeding of the results into CMB.

V. SUMMARY

"Most scientists would agree that it is important to document your work so that another scientist can replicate it, a tradition begun by Louis Pasteur." [4] To have control over the described process, to support repeatable, reliable discoveries and to accommodate additional information needs for annotation, we designed and we iteratively developing CellMapBase (CMB), a system of interdependent, collaborative tools, an information system/application that acts as data and application manager.

We set out to define CMB with the above citation being our guiding design principle. The result is an audit-ready data repository where we can trace a protein to the biological sample it was derived from; retrieve protocols, parameters and associated files related to the experiment and find references to methods, programs, instruments and public databases used in the discovery and identification process.

We restricted the scope of this paper to highlight the problem space, indicate the design requirements, list the used technologies and sketch out the implementation considerations.

Repeatability and reliability imply, however, automated procedures producing quantifiable results with reproducible methods. This requirement is not necessarily attainable using commercial, proprietary programs such as Mascot. As our understanding of the problem-solution space grew we identified a rich world of embedded research problems both in the 'experiment' world and in the 'library' world.

REFERENCES

[1] MASCOT (MatrixScience) http://www.matrixscience.com
[2] L. Stein, R. Dowell, DAS and BioDA http://www.biodas.org
[3] EMBL Features and qualifiers http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html#overview
[4] K. Robison, Documenting your search Available at: http://arep.med.harvard.edu/seqanal/documenting.html