Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

BNDb - Biomolecules Nucleus Database: an integrated proteomics and transcriptomics database

A.C. Faria-Campos[1], R.R. Gomes[1], F.S. Moratelli[1], H. Rausch-Fernandes[2], G.R. Franco[1] and S.V.A. Campos[2]

[1]Laboratório de Genética Bioquímica, Instituto de Ciências Biológicas, [2]Laboratório de Universalização de Acesso, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
Corresponding author: A.C. Faria-Campos
E-mail: alessa@icb.ufmg.br

Genet. Mol. Res. 6 (4): 937-945 (2007)
Received August 03, 2007
Accepted September 25, 2007
Published October 05, 2007

ABSTRACT. Proteomics correspond to the identification and quantitative analysis of proteins expressed in different conditions or life stages of a cell or organism. Methods used in proteomics analysis include mainly chromatography, two-dimensional electrophoresis and mass spectrometry. Data generated in proteomics analysis vary significantly, and to identify a protein it is often necessary to perform a series of experiments, comparing its results to those found in proteomics databases. Existing proteomics databases are usually related to only one type of experiment or represent processed results, not raw data. Therefore, proteomics researchers frequently have to resort to several data repositories in order to be able to perform the identification. In this paper, we propose an integrated proteomics and transcriptomics database that stores raw and

processed data, which are indexed allowing them to be retrieved together or individually. The proposed database, dubbed BNDb for Biomolecules Nucleus Database, is implemented using an MySQL server and is being used to store data from the parasite *Schistosoma mansoni*, the scorpion *Tityus serrulatus* and the spider *Phoneutria nigriventer*. The database construction uses a relational approach and data indexes. The data model proposed uses groups of tables for each data subtype, which store details regarding the experimental procedure as well as raw data, analysis results and associated publications. BNDb also stores transcriptomics data publicly available which are associated with identifications performed on new samples. By using BNDb, we expect not only to contribute to proteomics research but also to provide a useful service for the scientific community.

Key words: Databases, Proteomics, Transcriptomics, *Schistosoma*, Data integration, MySQL

INTRODUCTION

An unprecedented technological revolution has been recently witnessed by the scientific community in the methods for the acquisition of biological data ranging from the production of DNA sequences to the evaluation of mRNA expression patterns. This era started with the genomic revolution and is being followed by an even larger one, the proteomics revolution (Panisko et al., 2002; Kalia and Gupta, 2005).

Proteomics corresponds to the identification and quantitative analysis of proteins expressed in different conditions or life stages of a cell or organism (Bensmail and Haoudi, 2003). Methods used by proteomics researchers include mainly two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), high-performance liquid chromatography (HPLC) and mass spectrometry (MS) analysis (Panisko et al., 2002; Kalia and Gupta, 2005). The data generated by these experiments vary significantly depending on the experiment type and conditions used. Moreover, to uniquely identify a protein, it is often necessary to perform a series of experiments and compare its results to other results found in proteomics databases, making it complex to manage experimental data and reducing the efficiency of the protein identification process.

The number of proteomics databases currently available is considerable and it keeps growing (http://www.proteomicworld.orgDatabasePage.html). However, there is no well-standardized representation for proteomics data. This is mainly due to two reasons: the field is young and rapidly evolving and its dynamics makes it difficult to define the key data in a set of results (Taylor et al., 2003). Moreover, existing proteomics databases are usually related to only one type of data or represent already processed results, not raw data. Therefore, proteomics researchers frequently have to resort to several different data repositories in order to be able to perform the identification (Yates, 1998; Gras and Muller, 2001). Because there is no integration between experimentation, analysis and comparison with existing data, the experimental raw data produced are usually stored manually without any means to associate these data auto-

matically with results of analyses and resulting publications. Several initiatives including those associated with the Human Proteomics Organization have discussed the importance of raw data availability in association with processed data and the need for unified databases harboring both types of data (Reif et al., 2004; Rohlff, 2004; Martens et al., 2005a).

Some attempts have been made to achieve this purpose. PEDRo is a data schema for how to store and share proteomics data (Taylor et al., 2003; Garwood et al., 2004). PRIDE is another attempt that aims primarily to disseminate and make data publicly available (Martens et al., 2005b). In another application, 2DDB aims to store and analyze quantitative proteomics data (Malmström et al., 2006). None of the mentioned systems, however, proposes to integrate project managing dissemination of raw and analyzed data and to cross-reference proteomics results with available transcriptome data.

In this study, we propose an integrated proteomics and transcriptomics database that stores raw and processed data, which are indexed allowing them to be retrieved together or individually. The proposed database, dubbed BNDb for Biomolecules Nucleus Database, stores proteomics data produced by members of the Minas Gerais Proteomics Network and transcriptomics data publicly available. The database is implemented using an MySQL server and uses a relational approach and data indexes, which speeds up the process of protein identification and the correlation with transcriptomics data previously identified.

MATERIAL AND METHODS

The database BNDb was implemented using an MySQL server version 4.0.21 running on a Pentium 2.5 GHZ machine using Linux Suse distribution 9.2. The database construction uses a relational approach and data indexes to link experiments to each other and to the results and those to projects. The software DBDesigner 4.5.6 was used for the data model project. The data model proposed uses groups of tables for each data subtype, which store all details regarding the experimental procedure as well as raw data, analysis results and associated publications resulting from a specific experiment. The data model proposed has been designed to store data for the parasite *Schistosoma mansoni*, the scorpion *Tityus serrulatus* and the spider *Phoneutria nigriventer*.

The proposed database also stores transcriptomics data from these organisms, so that diverse information regarding an organism can be retrieved automatically. It contains all the data from transcriptomics analysis publicly available which are associated with identifications performed on new samples. The transcriptome data are stored in FASTA format along with identifications as gi or accession numbers.

Proteomics analysis and data modeling

The final objective of the proteomics analysis is to uniquely identify the set of proteins that are present in a given sample. In order to do so, several experiments are typically performed. The most common types of experiments are 2D-PAGE, HPLC and MS. It is important to note that these experiments are usually inter-related. That is, they depend on one another, since a sample separation by 2D-PAGE or HPLC usually precedes MS analysis. The data model has to take this into consideration in order to represent the sequence of experiments correctly. In order

Table 1. Proteomics data specification.

Type of experiment	Type of data	Data specification
Bi-dimensional polyacrylamide gel electrophoresis (2D-PAGE)	Spots in the gel	Position in the reference map in the gel, presence in different samples, isoelectric point, apparent molecular weight, intensity, area, volume
High-performance liquid chromatography (HPLC)	Chromatogram peaks	Number of peaks, peak area, total area, peak height, baseline, asymmetry concentration
Mass spectrometry (MS)	Molecular weight/peptide sequence	Computed mass, observed mass, sequence, ion score, modifications, identified protein, gi, accession number

to develop such representation, researchers working with proteomics data have been asked to highlight the main information for each kind of experiment which is shown in Table 1.

RESULTS

The data model

The data model consists of the understanding of which types of data will be stored in the database and how it will be stored. This is an essential task, because it defines not only which information will be stored and retrieved but also the relationship between these data, i.e., which attributes are related to one another. It is through these relationships that the system can maintain system consistency and identify situations that require users' attention. For example, chromatography experiments generate chromatograms with several peaks that will be analyzed individually. However, the chromatography experiment must be also stored in the database. The experiment data include the number of peaks produced, the total peak area, and the position of each peak in the chromatogram. The data model establishes this relation and guarantees that peaks will be stored only after the chromatogram has been as well, guaranteeing that all necessary data for the analysis will be present.

The major components of the data model of the database BNDb can be seen in Figure 1, where the main entity is the experiment. Experiments are associated with projects, and those are associated with the researchers who belong to them. Experiments can be of three different types, and each can have raw data and results stored in the database. Each experiment can also have associated with it specific characteristics and protocols.

The complete entity relationship model can be seen in Figure 2, where the major database tables can be seen together with the relations between them. The constructed model uses a relational database with tables linked through foreign keys. That is, fields in one table are directly linked to fields in another table. The tables are divided according to the model in Figure 1. Administration tables store information about projects and their members. Experiment tables are divided into three subgroups, 2D-PAGE, HPLC and MS. These store experimental data. Image and chromatogram files are stored outside of the database, but a link to these files is stored in the database. Transcriptomic tables are also included in this model. These tables include information about nucleotide and protein sequences as well as the sequence in FASTA format.

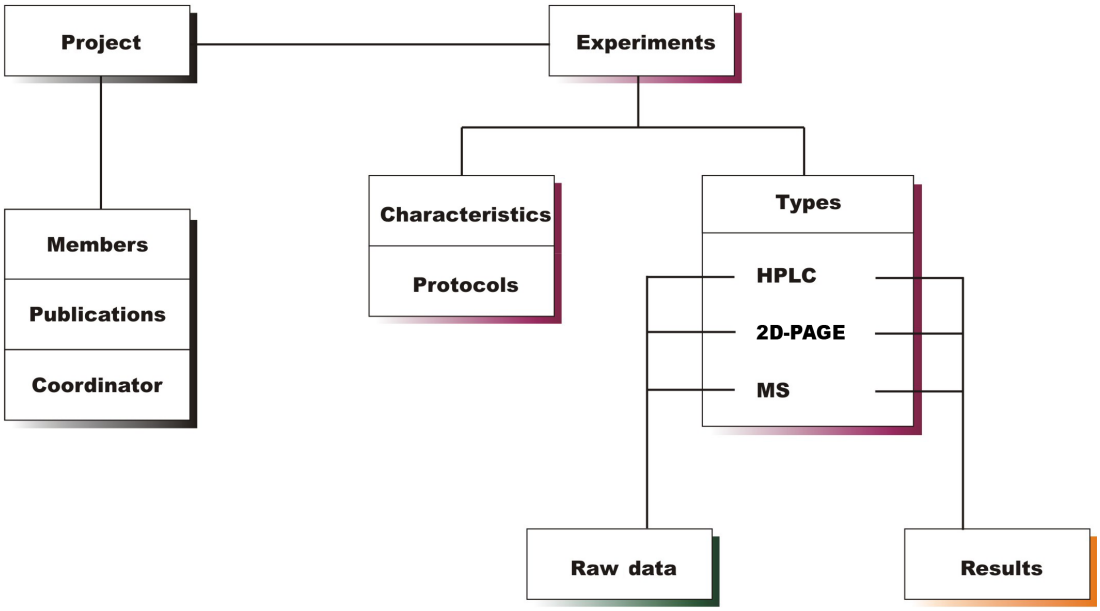

Figure 1. Task executed by the database BNDb. In blue are administrative tasks. In red are experiment-related tasks. The database stores raw (in green) and processed (yellow) data. HPLC = high-performance liquid chromatography; 2D-PAGE = two-dimensional polyacrylamide gel electrophoresis, MS = mass spectrometry.

This sequence can be exported from the database to be used in other bioinformatics tools such as similarity searches (using BLAST) or evolutionary analysis. The transcriptomic data have been obtained directly from the National Center for Biotechnology Information (NCBI).

It is important to note that the transcriptome data stored in the BNDb is an important aspect of this model, because it allows researchers to access all data available about a certain protein from its nucleotide sequence to the proteomics analysis, making it simpler and faster to perform the analysis. This feature is not available in other proteomics databases to our knowledge.

BNDb is accessed through a web-based interface that uses *php* scripts to communicate with the database. An example of an access page can be seen in Figure 3. Experimental data are not inserted in the database using the web interface. Instead, parsers are used to interpret these data directly from the experiment results. This makes importing data faster and less error prone, since the files generated from the instruments that perform the experiments are read automatically by the parsers.

We are currently finishing the implementation of the web interface, as well as the parsers for importing experiment data. BNDb will then be populated by proteomics data generated by experiments from several laboratories at UFMG and Fiocruz. We have also obtained transcriptome data from NCBI and have inserted those in the database. We have obtained 203,395 sequences form *S. mansoni*, 17 sequences from *T. serrulatus* and 89 sequences from *P. nigriventer*, which include nucleotide and protein sequences.

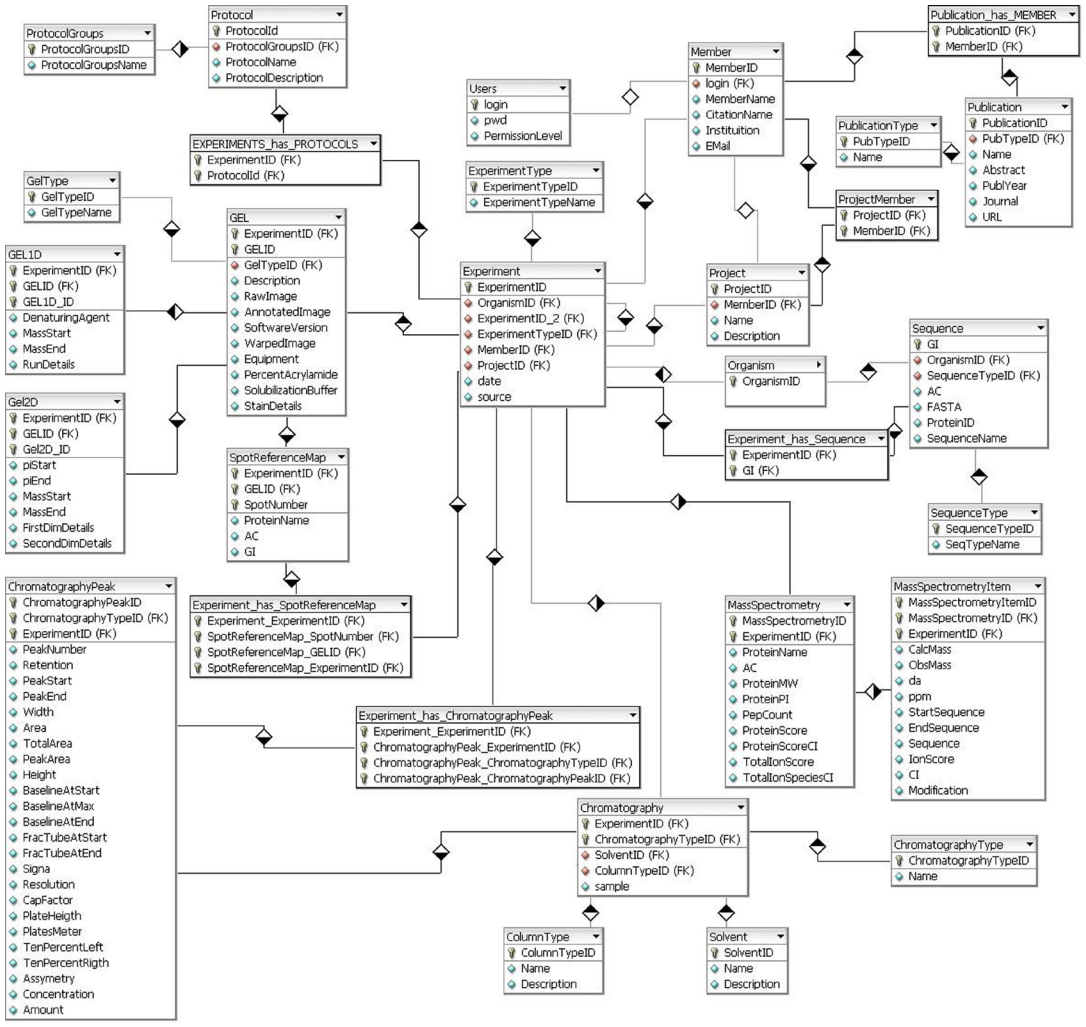

Figure 2. BNDb complete entity relationship model showing all tables in the model and the relationships.

DISCUSSION AND CONCLUSIONS

In this paper, we have presented BNDb, the Biomolecules Nucleus Database, a proteomics and transcriptomics database that stores data from proteomics experiments. BNDb stores not only results from proteomics analysis but also its raw data. It also connects the information produced by different types of experiments as well as transcriptome data in order to be able to present to the researcher a complete picture of the experimental process, making it easier to access all data related to specific experiments. This speeds up the proteomics analysis and increases its reliability.

The number of proteomics databases available nowadays is large and keeps growing. However, most of them have been designed to store processed and curated data. This means that the proteomics data are only stored in the database after the experiments have been com-

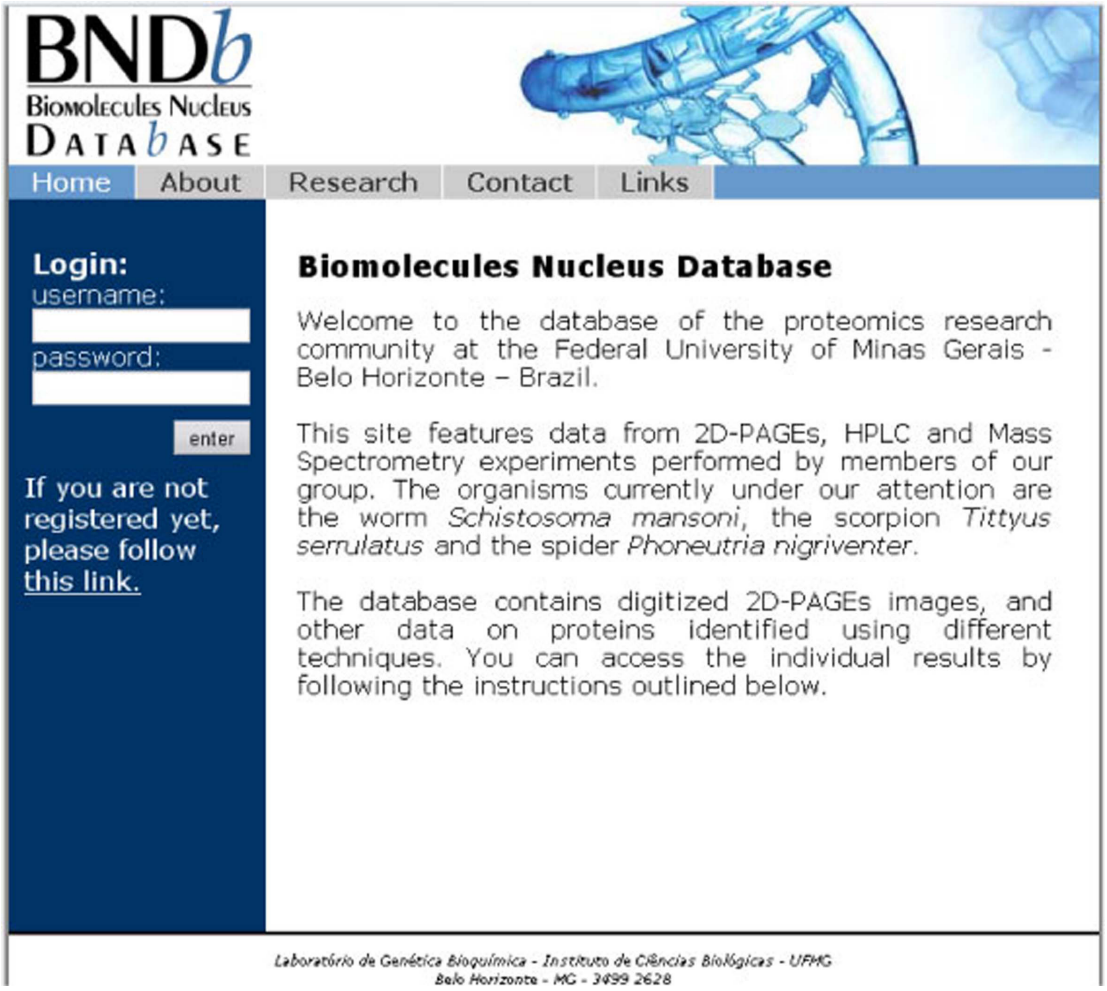

Figure 3. BNDb web interface screenshot. Users access the database through this page. Users who are not registered are directed to register through a form open in the link on this page. 2D-PAGE = two-dimensional polyacrylamide gel electrophoresis; HPLC = high-performance liquid chromatography.

pleted and their results completely analyzed and curated. These databases assist researchers in comparing their experiments with others that have already been completed. Consequently, these databases perform a different task than BNDb, and in fact are complementary to it. The focus of BNDb is not in storing the final results (even though this is also done), but rather to help researchers in keeping track of the data generated by individual experiments and understanding how these data relate to other experimental data even before the experiment results have been processed.

From the other proteomic databases available, two data models relate more closely to the BNDb model, namely, 2DDB and PEDRo. Both store proteomics raw experimental data in a similar way compared to BNDb. 2DDB, however, focuses on protein identification and how to identify experiments that relate to a given protein. 2DDB is efficient in determining the

path through which a protein has been identified. The raw data generated by the experiments, however, are not the focus of the project, and in fact these data are not part of the core data model of 2DDB. As a consequence, 2DDB can help researchers after the experimental process has identified the protein, but potentially not as much during the experimental phase, when it is necessary to store raw data independent of which protein it relates to since this is not known at the time. 2DDB also does not store data related to transcriptomics. PEDRo stores proteomics data based on the experiment order in which the data were generated. It is not, however, a relational database system as are BNDb and 2DDB. Instead, it stores data in an XML format file. As a consequence, it is not so efficient in storing large volumes of data. Besides, XML storage imposes a natural indexing of the data, making it possible to access the data in one order (the order in which it was stored, in PEDRo's case the experimental order), but not in a different order. Therefore, if a researcher using the PEDRo system wants to access data based on criteria other than the established order, access is not efficient, particularly for large databases. PEDRo also does not have the ability to connect to other databases, either external ones, or extended ones, such as the transcriptome database included in BNDb.

For data capture, PEDRo database makes extensive use of XML for capturing, transmitting, storing, and searching proteomics data. The data-capture process uses a software tool that prompts users for values for different fields, and includes facilities for importing substantial data files, such as those representing peak lists. The tool constructs data-entry forms from the XML Schema definition of the PEDRo model. The result of the data-capture process is thus an XML file that corresponds to the PEDRo schema. BNDb, on the other hand, uses a web server as the interface for data capture. Simple forms constructed in *php* language are made available for data entry. The researcher uploads the result files generated directly by the experiment using this interface. The files are processed through parsers that are used to interpret these data directly from the experiment results. Using a web-based interface gives an increased portability to data capture since users do not need to install any specific software to have access to the database for data importing and exporting. Also, this system makes data importing and storing faster and less error prone, since the files are imported automatically, processed by the parsers and inserted directly in the database.

BNDb has been designed to store data from experiments of three different organisms, *S. mansoni*, *T. serrulatus* and *P. nigriventer*. These experiments have been performed in different laboratories by different research groups, demonstrating the usefulness of the BNDb, which will provide assistance in proteomics research for a large scientific community. Moreover, it demonstrates the capability of the database to store data from different formats and research groups, demonstrating also its flexibility.

The construction of a new data model for importing and storing proteomics data represents an important contribution to proteomics and bioinformatics, young and growing fields that represent the new frontier in biological sciences.

REFERENCES

Bensmail H and Haoudi A (2003). Postgenomics: proteomics and bioinformatics in cancer research. *J. Biomed. Biotechnol.* 2003: 217-230.

Garwood K, McLaughlin T, Garwood C, Joens S, et al. (2004). PEDRo: a database for storing, searching and disseminating experimental proteomics data. *BMC Genomics* 5: 68.

Gras R and Muller M (2001). Computational aspects of protein identification by mass spectrometry. *Curr. Opin. Mol. Ther.* 3: 526-532.

Kalia A and Gupta RP (2005). Proteomics: a paradigm shift. *Crit. Rev. Biotechnol.* 25: 173-198.

Malmström L, Marko-Varga G, Westergren-Thorsson G, Laurell T, et al. (2006). 2DDB - A bioinformatics solution for analysis of quantitative proteomics data. *BMC Bioinformatics* 7: 158.

Martens L, Nesvizhskii AI, Hermjakob H, Adamski M, et al. (2005a). Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. *Proteomics* 5: 3501-3505.

Martens L, Hermjakob H, Jones P, Adamski M, et al. (2005b). PRIDE: the proteomics identifications database. *Proteomics* 5: 3537-3545.

Panisko EA, Conrads TP, Goshe MB and Veenstra TD (2002). The postgenomic age: characterization of proteomes. *Exp. Hematol.* 30: 97-107.

Reif DM, White BC and Moore JH (2004). Integrated analysis of genetic, genomic and proteomic data. *Expert. Rev. Proteomics* 1: 67-75.

Rohlff C (2004). New approaches towards integrated proteomic databases and depositories. *Expert. Rev. Proteomics* 1: 267-274.

Taylor CF, Paton NW, Garwood KL, Kirby PD, et al. (2003). A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat. Biotechnol.* 21: 247-254.

Yates JR (1998). Database searching using mass spectrometry data. *Electrophoresis* 19: 893-900.