A New Approach to the Integration of Proteomics Experimental Data

Alessandra Faria-Campos[1], Herbert Fernandes[2], Rodrigo Gomes[1], Breno Rates[1], Adriano Pimenta[1], Glória Franco[1], Sérgio Campos[2]
alessa@icb.ufmg.br

[1]Departamento de Bioquímica e Imunologia
Instituto de Ciências Biológicas
[2]Departamento de Ciência da Computação
Instituto de Ciências Exatas

Universidade Federal de Minas Gerais
Belo Horizonte - Minas Gerais - Brazil

Abstract. Proteomics databases are currently a very important research topic in bioinformatics. Proteomics is the identification and quantitative analysis of proteins expressed in different conditions or life stages of a cell or organism. Proteomic analysis technologies are mainly chromatography, bi-dimensional electrophoresis and mass spectrometry. To identify a protein it is often necessary to perform a series of experiments, comparing the results of such analysis to those found in proteomics databases. Most existing proteomics databases are usually related to only one type of experiment or represent processed results, instead of raw data. Because of this, researchers frequently have to resort to several data repositories in order to be able to perform the identification of the proteins. In this paper we propose an integrated proteomics database that stores raw and processed data, which are indexed allowing them to be retrieved together or individually. The proposed database, named BNDb for Biomolecules Network Database, is implemented using a MySQL server and is being used to store data from the centipede *Scolopendra viridicornis*, the parasite *Schistosoma mansoni*, the scorpion *Tityus serrulatus* and the spider *Phoneutria nigriventer*. The database construction uses a relational approach and data indexes. The proposed data model uses groups of tables for each data subtype, which store details regarding experimental procedures as well as raw data, analysis results and linked publications. BNDb also stores sequence data publicly available which can be associated to newly identified proteins present in the database. BNDb represents a new contribution to proteomics data management providing a useful service for the scientific community.

Keywords: Proteomics database, data model

1 Introduction

One of the great challenges that face science today, and in particular the biological sciences is the management of increasingly large amounts of data generated by high throughput experiments. A research area that has greatly benefited from the development of new and improved technologies for large scale experiments is proteomics. Proteomics is the identification and quantitative analysis of proteins expressed in different conditions or life stages of a cell or organism (Wilkins et al., 1997). Proteomics data are frequently stored in spreadsheets or published in databases, an approach that presents limited integration regarding raw and processed data. Existing proteomics databases are usually related to only one type of data or represent processed results, instead of raw data. Because of this, proteomics researchers frequently have to resort to several different data repositories in order to be able to extract biological information from the experimental data (Yates, 1998; Gras and Muller, 2001). Thus, several initiatives, including those associated to the Human Proteomics Organization (HUPO), have discussed the importance of unified databases harboring raw and processed data and of a system to improve the management of both types of data (Reif et al., 2004; Rohlff, 2004, Martens et al., 2005).

In this work we propose an integrated proteomics database that stores both raw and processed data, which are indexed allowing them to be retrieved together or individually. The proposed database, named BNDb for Biomolecules Network Database, stores proteomics data produced by members of the Minas Gerais Proteomics Network (Brasil) and sequence data publicly available. The database has been implemented using a MySQL server and uses a relational approach and data indexes which speeds up the process of data retrieval and the correlation with sequence data previously identified.

BNDb addresses the problems outlined above in two different ways. First, it imports experimental data directly from the equipment used to perform the experiment. The output files generated by the equipment are parsed by BNDb and the data is inserted in the database without user assistance, eliminating the bottleneck of inserting data manually into the computer. Moreover, an relational database is used to store this data and retrieve it efficiently. The relational database allows us to use standard database techniques to identify relationships between different sets of data and to retrieve these associated data in a straightforward manner. Currently our database uses SQL queries to retrieve the data, meaning

that we can identify sets of data in different experiments that have common attributes directly. This enables us to relate experiments that have the same data values or that may have been previously unknown to be related. It is important to notice that the focus of BNDb is to store and analyze proteomics experimental data. It is possible to use tools that offer a level of flexibility in the design of the database such as workflow based tools and object oriented databases that is not present in BNDb. Workflow based systems allow users to easily change the protocols and data being stored; object oriented databases make it simpler to change the structure of the database. However, this flexibility is not needed in proteomics, since the types of experiments is fixed for this type of analysis, and the flexibility offered by those tools does not add to the task at hand. Moreover, flexibility often comes at a price, often the ease of modification of tables and experiment steps make the system less efficient, and the fact that these changes are made through a user interface, makes it impossible to implement more complex analyses that cannot be directly modeled through the user interface. The objective of BNDb rather than enabling its final user to make a restricted set of changes to the model, is to make it possible to implement very sophisticated analysis algorithms and tools which can be easily added directly to the database, but cannot be described in terms of an end-user interface.

Some other approaches have been used to store proteomics experimental data. PEDRo is a data schema for how to store and share proteome data (Taylor et al., 2003; Garwood et al., 2004). PRIDE is a database that stores protein and peptide identifications, another attempt that aims primarily to disseminate and make data publicly available (Martens et al., 2005b). Another implementation, 2DDB aims to store and analyze quantitative proteomics data (Malmstrom et al., 2006). None of the mentioned systems however, proposes to integrate project managing, dissemination of raw and analyzed data and cross-referencing of proteomics results with available sequence data as proposed by BNDb.

2 The Data Model

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 such as bi-dimensional electrophoresis (2D-PAGE), liquid chromatography (LC) and mass spectrometry (MS). It is important to notice that these experiments are usually related, since a sample separation by 2D-PAGE or LC usually precedes

the MS analysis. The data model has to take this into consideration in order to represent the order and relationship of the experiments correctly.

The major components of the data model of the BNDb database can be seen in Figure 1, where the main entity is the *experiment*. Experiments are associated with projects, and those are associated with researchers that belong to them. Known DNA,EST or protein sequences are also associated to experiments. Experiments can be of different types: *Liquid Chromatography,1D/2D Gels* and *Mass Spectrometry* and each experiment can have raw data and results stored in the database.

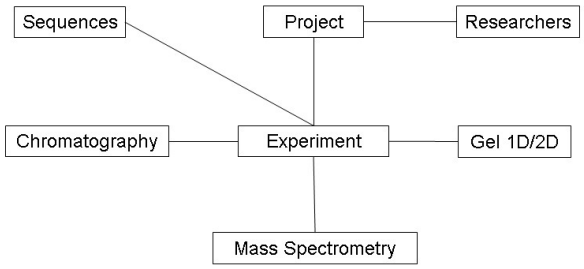


Fig. 1. Main components of the data model.

A simplification of BNDb data model can be seen in Figure 2 (some auxiliary tables are not shown for clarity). In the database, an experiment can be of type *chromatography*, *gel electrophoresis*, or *mass spectrometry*. In each case the corresponding table has an entry associated with the experiment. Each type of experiment has one or more specific results, either chromatography peaks, gel spots or m/z values. These results are also stored in separate tables, and are associated with the experiment. Image and chromatogram files are stored outside of the database, but a link to these files is stored in the database along with numeric results.

The experiment is the main component of the model because BNDb is aimed primarily at assisting researchers in cross-linking their experi-

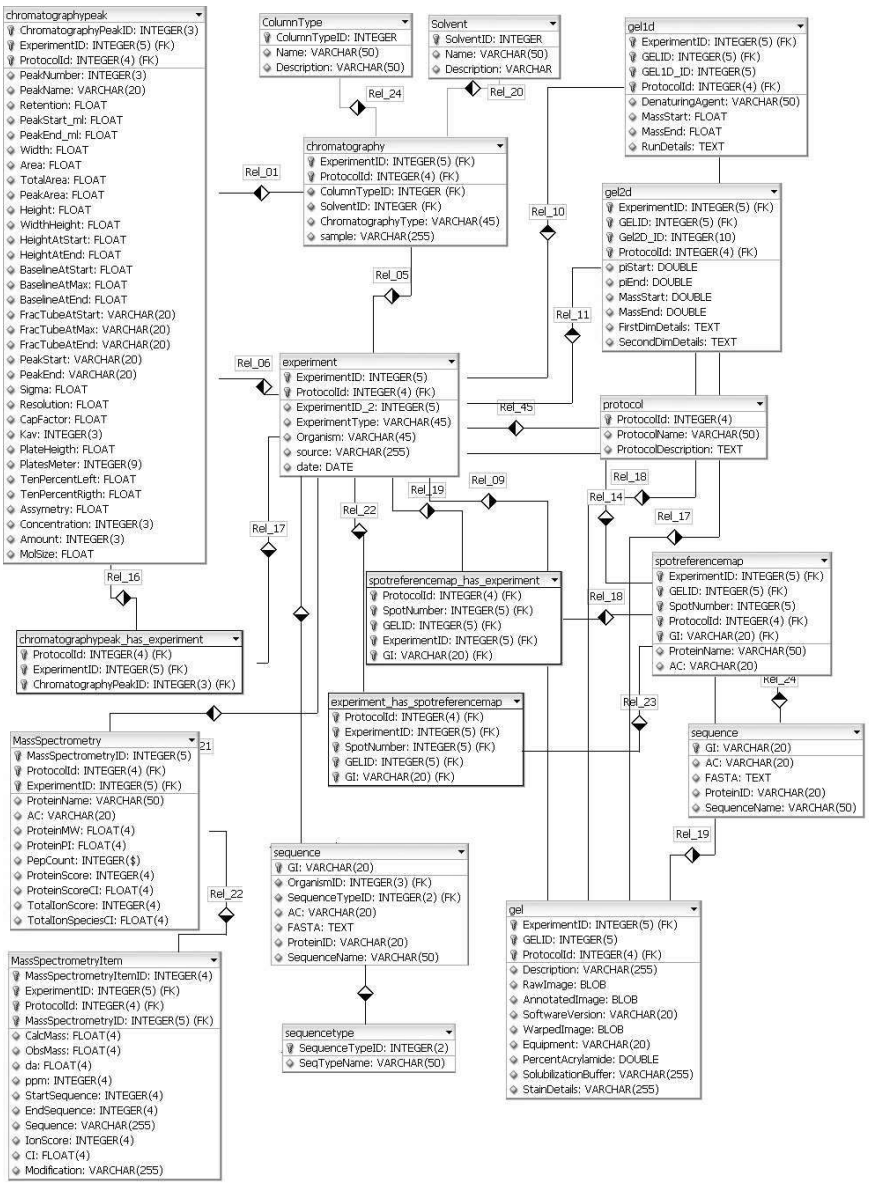
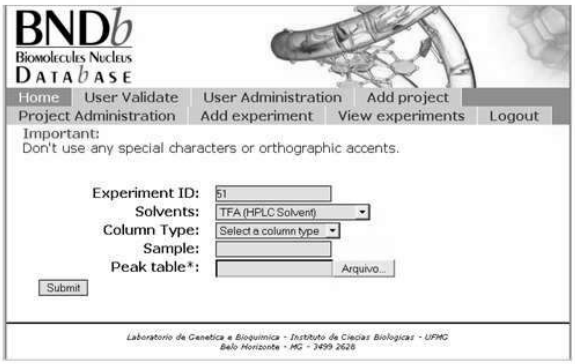


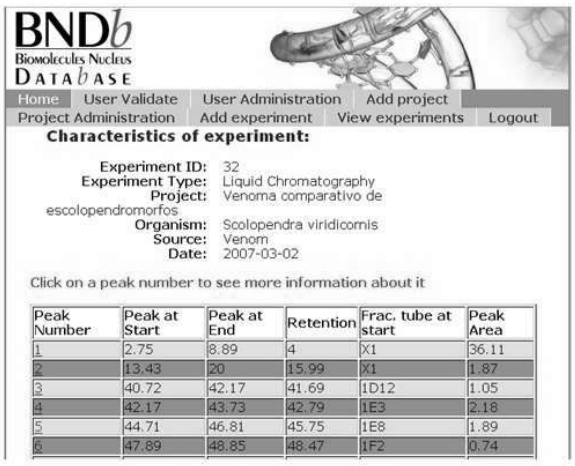
Fig. 2. The data model.

mental data. The common data flow starts when a new experiment is performed and its data is entered in the *experiment* table. Each experiment receives an internal ID that identifies it uniquely. According to the type of experiment the tables related to that type of experiment are also filled

(A)



(B)



(C)

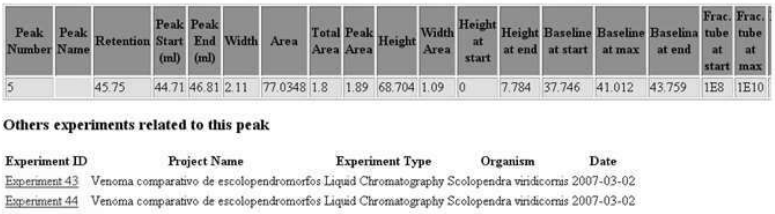


Fig. 3. User interface to upload data and retrieve it from the database. (A) Data upload screen; (B) Experiment analysis screen; (C) Peak analysis screen including cross-links to related experiments.

and associated to the internal ID. Experiments are often related to other experiments, such as when a particular result prompts the researcher to perform a more detailed analysis on a peak or spot. BNDb then stores the associated experiments by cross-linking the internal IDs which allows

it to establish not only a relational, but also a temporal link, so that the sequence of experiments performed can be followed later. Figure 3 shows how this complex set of operations can be visualized by the user. Through the user web interface the researcher can visualize the relationships of his experiment without worrying about the details of database construction. The BNDb database is accessed through a web based interface that uses php scripts to communicate with the database. Experimental data, however, is not directly inserted in the database using the web interface. Instead, the users uploads the experiment results file using the web interface and parsers are used to interpret this data and insert it in the database. This makes importing data faster and less error prone, since the files generated from the equipments that perform the experiments are read automatically by the parsers.

Through the association of the tables in the database it is possible to identify exactly how experiments are related to one another, assisting the researchers in controlling the flow of experiments performed, and also in explaining the results and how they were obtained. It is important to notice that this is possible in BNDb because of the use of a relational database, since it enables the user to search for related items regardless of where they have been stored. Hierarchical storage methods such as PEDRo may have difficulty following some of the relationships due to the nature of their model.

The *project* component allows BNDb to store data of different projects and researchers, enabling each to work independently with exclusive access to their data. BNDb stores not only projects and project members, but also the researcher that has performed each experiment, making possible a fine control of the experiment flow and using security restriction to garantee acess to specific data.

Sequence tables are also included in this model. These tables contain information about nucleotide and protein sequences as well as the sequence in FASTA format. This sequence can be exported from the database to be used in other bioinformatics tools such as similarity searches (using BLAST) or evolutive analysis. The sequence data have been obtained directly from the National Center for Biotechnology Information (NCBI). Sequence data stored in BNDb is an important aspect of this model, because it allows researchers to access all data available on a certain protein from its nucleotide sequence to the proteomic analysis, making it simpler and faster to perform the analysis. To our knowledge this feature is not available in other proteomic databases.

We are currently finishing the implementation of the web interface, as well as the parsers for importing experiment data. The BNDb will then be populated by proteomics data generated by experiments from several laboratories at UFMG and Fiocruz. We are also currently working on importing sequence data from NCBI and to have it inserted in the database.

The database has currently been filled with data from the proteomic analysis of the venom from the arthropod *Scolopendra viridicornis*, a common Brazilian centipede. In order to assess the complexity of the venom of *Scolopendra viridicornis*, a pooled venom sample (1 mg) was subjected to bi-dimensional liquid chromatography. This technique consists of the sequential use of ion-exchange fractionation (first dimension), followed by further purification by reversed phase (RP) chromatography (second dimension) of the fractions obtained in the first step. After the RP step, the fractions were analyzed by electrospray ionization quadrupole/time-of-flight mass spectrometry (ESI-Q-TOF/MS). The fractions which contained proteins and peptides purified to a homogeneous state were subjected to N-terminal sequencing by automated Edman's degradation. Then, similarity searches were performed by the Fasta3 tool against the Uniprot and Swiss-Prot data Bank. Details on the methodology were provided by Rates and co-workers (2007). In Figure 4 we can see the initial chromatogram (4A) and the subsequent ones produced from a second analysis of the individual resulting peaks(4B). The user can then choose one peak from individual experiments and visualize the spectrum corresponding to the MS analysis (4C).

3 Implementation

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

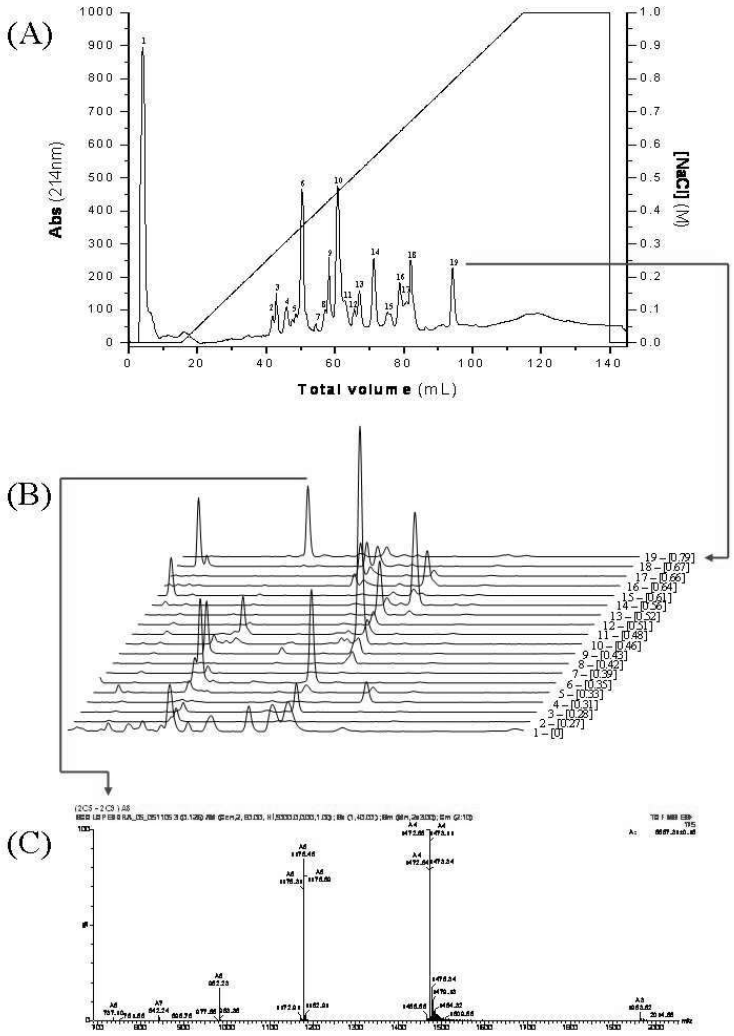


Fig. 4. Example of cross-linked experiments in a typical proteomics analysis. (A) Initial chromatogram from the analysis of *Scolopendra viricornis* venom; (B) Subsequent chromatograms produced from the analysis of the individual peaks in A; (C) Spectrum resulting from analysis of a peak selected in B.

analysis of the venom from the scorpion *Tittyus serrulatus* and the spider *Phoneutria nigriventer*.

The proposed database also stores sequence data from these organisms, so that diverse information regarding an organism can be retrieved

automatically. It contains sequence data publicly available which will be associated to identifications performed in new samples. This data is stored in FASTA format along with identifications as gi or accession numbers.

4 Discussion and Conclusions

In this paper we have presented BNDb, the Biomolecules Network Database, a proteomics integration database. BNDb stores raw and analyzed data along with project management attributes. It also connects the information produced by different types of experiments as well as sequence 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 proteins. This speeds up the proteomics analysis and increases its reliability.

The number of proteomics databases available nowadays is high and is still increasing. However, most of them have been designed to store processed and curated data or store one type of experiment only. This means that the proteomics data is only stored in the database after the experiments have been completed 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 tracking of the data generated by individual experiments and how this data relates 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, 2DDB and PEDRo. Both store proteomics raw experimental data in a similar way as BNDb. 2DDB, however, focus on the protein identification and how to identify experiments that relate to a given protein. 2DDB is based on a core data model describing fundamentals such as experimental description and identified proteins (Malmstrom et al., 2006). It is efficient in determining the path through which a protein has been identified but raw data are not the focus of the project, and in fact this data is not part of the core data model of 2DDB. As a consequence, 2DDB helps researchers after the protein identification has been performed but is less helpful during the experimental phase, when it is necessary to store raw data and experiments attributes independent of which protein it relates to since this is not known at the

time. 2DDB also does not store sequence data. PEDRo stores proteomics data based on the experiment order in which it was generated (Garwood et al., 2004). It is not, however, a relational database system as are BNDb and 2DDB. Instead, it stores *and processes* data only in a XML format file. As a consequence, it is not so efficient in storing large volumes of data. Besides, the XML storage imposes a natural indexing of the data, since these files are read and stored sequentially. As a consequence, it is very simple to retrieve information in the same order as it was stored, but if one needs to correlate data in a different order using an XML file becomes inefficient, since all information must be reordered in main memory to allow a different indexing. In our case, data is cross-referenced in different ways depending on the analysis being performed, and since the order changes frequently, no predetermined order would be efficient. We use a relational database to store data, because relational databases are designed to allow information retrieval in multiple orders efficiently. So, if a researcher using the PEDRo system wants to access data based on other criteria than the established order, access is not efficient, particularly for large databases, because PEDRo uses the XML format not only for storage, but also for processing the data.

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 which 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 this data directly out of 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 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 on the database.

BNDb has been designed to store data from experiments of several different organisms. At the moment data from *S. vidicornis*, *S. mansoni*, *T. serrulatus* and *P. nigriventer* proteomic studies are being collected to

feed the database. These experiments have been performed in different laboratories by different research groups, demonstrating the usefulness of the BNDb, which can provide assistance in the proteomics research for a large scientific community. Moreover, it demonstrates the capability of the database to store data from different formats and research groups, emphasizing also its flexibility.

The construction of a new data model for proteomics data importing and storing represents an important contribution for proteomics and bioinformatics. We have developed a tool that combines a powerful storage engine (the relational database) and a friendly access interface, aiming to assist proteomics researches directly at data handling and storage.

5 References

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 3:526-32.

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

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

Martens L, Nesvizhskii AI, Hermjakob H, Adamski M et al. (2005). 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 5:3537-45.

Rates, B., M. P. Bemquerer, M. Richardson, M. H. Borges, R. A. Morales, M. E. De Lima e A. M. Pimenta (2007). Venomic analyses of Scolopendra viridicornis nigra and Scolopendra angulata (Centipede, Scolopendromorpha): Shedding light on venoms from a neglected group. Toxicon 49:810-26.

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.

Wilkins MR, Williams KL, Appel, RD and Hochstrasser (1997). Proteome Research: New Frontiers in Functional Genomics. Springer -Verlag, Berlim.