Bioinformatics Advance Access published October 6, 2009

Qupe - a Rich Internet Application to take a Step Forward in the Analysis of Mass Spectrometry-Based Quantitative Proteomics Experiments

Stefan P. Albaum [1,4], Heiko Neuweger [1], Benjamin Fränzel [5], Sita Lange [1], Dominik Mertens [1,4], Christian Trötschel [5], Dirk Wolters [5], Jörn Kalinowski [3], Tim W. Nattkemper [4], Alexander Goesmann [1,2]

[1]Computational Genomics, Center for Biotechnology (CeBiTec), Bielefeld University
[2]Bioinformatics Resource Facility, CeBiTec, Bielefeld University
[3]Institute for Genome Research and Systems Biology (IGS), CeBiTec, Bielefeld University
[4]Biodata Mining & Applied Neuroinformatics Group, Faculty of Technology, Bielefeld University
[5]Biomolecular Mass Spectrometry, Department of Analytical Chemistry, Ruhr-University Bochum

Associate Editor: Dr. Trey Ideker

ABSTRACT

Motivation: The goal of present -omics sciences is to understand biological systems as a whole in terms of interactions of the individual cellular components. One of the main building blocks in this field of study is proteomics where tandem mass spectrometry (LC-MS/MS) in combination with isotopic labelling techniques provides a common way to obtain a direct insight into regulation at the protein level. Methods to identify and quantify the peptides contained in a sample are well-established, and their output usually results in lists of identified proteins and calculated relative abundance values. The next step is to move ahead from these abstract lists and apply statistical inference methods to compare measurements, to identify genes that are significantly up- or down-regulated, or to detect clusters of proteins with similar expression profiles.

Results: We introduce the rich internet application Qupe providing comprehensive data management and analysis functions for LC-MS/MS experiments. Starting with the import of mass spectra data the system guides the experimenter through the process of protein identification by database search, the calculation of protein abundance ratios, and, in particular, the statistical evaluation of the quantification results including multivariate analysis methods such as analysis of variance or hierarchical cluster analysis. While a data model to store these results has been developed, a well-defined programming interface facilitates the integration of novel approaches. A compute cluster is utilised to distribute computationally intensive calculations, and a web service allows to interchange information with other -omics software applications. To demonstrate that Qupe represents a step forward in quantitative proteomics analysis an application study on *Corynebacterium glutamicum* has been carried out.

Availability and Implementation: Qupe is implemented in Java utilising Hibernate, Echo2, R and the Spring framework. We encourage the usage of the rich internet application in the sense of the "software as a service" concept, maintained on our servers and accessible at the following location:
http://qupe.cebitec.uni-bielefeld.de

Contact: Stefan.Albaum@CeBiTec.Uni-Bielefeld.DE

1 INTRODUCTION

Present -omics scienes try to understand biological systems as a whole by scrutinising the individual components and their interactions. In this field of study, often referred to as systems biology, proteomics is one of the main building blocks. While a few years ago, two-dimensional gel electrophoresis in combination with single-stage mass spectrometry had been the standard technique to yield information about the proteome in a cell (Hufnagel and Rabus, 2006), recent methods such as liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) provide the possibility to characterise hundreds of peptides in a single sample. A common way to compare the abundance of proteins under two or more conditions is the combination of mass spectrometry with isotopic labelling techniques (Mueller *et al.*, 2008; Zhu *et al.*, 2002; Ong *et al.*, 2002; Wolters *et al.*, 2001), which enables us to obtain a direct insight into regulation at the protein level. Starting from the data recorded by a mass spectrometer instrument, a typical experiment's workflow involves i) a database search to identify proteins contained in a sample, ii) the calculation of peptide abundance ratios, and iii) a following evaluation of the results.

i) The standard method to identify proteins or peptides, respectively, compares the recorded mass spectra with theoretical fragmentation patterns derived from sequence databases, using search engines such as Mascot (TM) (Perkins *et al.*, 1999), Sequest (TM) (Yates *et al.*, 1995), OMSAA (Geer *et al.*, 2004), ProbID (Zhang *et al.*, 2002), or X!Tandem (Craig and Beavis, 2004). An integral element of this "qualitative" part of the workflow is the validation of the reported peptides and proteins. A common strategy therefore is based on the utilisation of decoy databases and the calculation of false discovery rates (Peng *et al.*, 2003; Elias and Gygi, 2007).

A variety of software applications aims to guide through this process of peptide identification and validation, and to provide a standardised way of data management. In general, either specific flat file formats or databases are utilised to store and retrieve e.g. mass spectra data, reported proteins or documentation of the experimental setups. As experiments are often conducted within larger communities and therefore need to be shared between a number of participants, user management and data access

control are vital components of these systems. Examples of such applications are CPAS (Rauch *et al.*, 2006), MASPECTRAS (Hartler *et al.*, 2007), Proteios (ProSE) (Gärdén *et al.*, 2005; Levander *et al.*, 2009), or the command-line based Trans-Proteomics pipeline (TPP) (Keller *et al.*, 2002; Nesvizhskii *et al.*, 2003) and the OpenMS/TOPP framework (Kohlbacher *et al.*, 2007; Sturm *et al.*, 2008). As a recommended standard for proteomics data, the Proteomics Standards Initiative (PSI) at the Human Proteome Organisation (HUPO) (Orchard *et al.*, 2003) specified the MIAPE reporting guidelines (Taylor *et al.*, 2007) - the minimum information about a proteomics experiment.

ii) Common experimental strategies for relative quantification are based on the incorporation of stable isotopes. In 2003, RelEx (MacCoss *et al.*, 2003) and ASAPRatio (Li *et al.*, 2003) were introduced to calculate relative abundance ratios from samples that are metabolically labelled using e.g. heavy stable nitrogen isotopes. ProRata (Pan *et al.*, 2006) and Census (Park *et al.*, 2008), the successor of RelEx, are further examples of quantification tools, while other labelling approaches encompass ICAT (Gygi *et al.*, 1999) or SILAC Ong *et al.* (2002). In general, these tools are standalone software applications that have solely been designed for the process of quantification. The recently introduced MaxQuant (Cox and Mann, 2008) supports the SILAC approach, and is the first tool that additionally integrates protein identification using the Mascot (TM) search engine.

iii) While the aforementioned applications allow to identify and quantify an organism's proteome, their end product is usually a list of calculated abundance ratios or expression values for the identified proteins. As a typical experimental setup includes more than one condition, the resulting values need to be combined to form e.g. a data matrix (Kumar and Mann, 2009). At this point of analysis, proteomics researchers are somehow left out in the cold since existing software solutions as listed above lack support of advanced data analysis. Moreover, in many workflows it is often not yet clear what the best analysis methodology is, whether to identify up-down protein regulation, for comparative studies with varying conditions, to detect protein clusters with similar expression profiles, or to fuse information with external databases such as KEGG (Kanehisa and Goto, 2000). Software such as spreadsheet programs or statistical programming languages, albeit generally usable for this purpose, demand a high level of background knowledge and training, or do not adapt to the complexity of proteomics data. In addition, data and associated meta-data are not found connected in a single place.

A software application that provides a comprehensive set of statistical methods for various -omics data sources is the tool DAnTE (Polpitiya *et al.*, 2008). This application, however, relies on the import of measurements in form of the aforementioned data matrix or spreadsheet data, and does neither integrate peptide or protein identification and quantification nor implement data management functions to organise experiments or projects. Ramos *et al.* (2008) are following a different approach with their protein information and property explorer (PIPE) that does not aim at the statistical evaluation but at a functional analysis of identified peptides. The application allows for server-side data storage and provides for example functionality to associate Gene Ontology (Ashburner *et al.*, 2000) information with identified proteins.

We have developed Qupe with two aims. First, we wanted to design a software package that integrates all aspects of the mass spectrometry-based proteome analysis workflow discussed above, from identification to multivariate statistical analysis. Second, we wanted to move forward in bringing algorithms closer to the biologists and developed Qupe as a so called rich internet application. As such, it addresses the limitations in "the richness of the application interfaces, media and content" (Allaire, 2002, p. 1) of classical web applications and offers an interface that behaves similar to standalone software applications running on a user's desktop. Qupe is independent from any operating system and the need for installation on individual workstations is omitted. Hence, data stored in the system such as mass spectra, or analysis results may be accessed on any computer connected to the internet.

2 IMPLEMENTATION AND METHODS

Qupe is based on the Spring framework (Johnson, 2003; Interface21, 2008). It is compliant to the Java Platform Enterprise Edition (Java EE) specification, and thereby portable across all compatible application servers. Following the three tier architecture model, the system is separated into data access, logic, and presentation layer (see Figure 1). Data stored in the system is protected by a number of security measures. In the first place, Qupe incorporates a generalised project management system (GPMS). On this level, security is based on discrete grants on databases and associated tables. The system has already successfully been used in other software packages hosting hundreds of international -omics projects (Neuweger *et al.*, 2008; Dondrup *et al.*, 2009). A second level of application-based security has been implemented utilising access control list (ACL) directives on selected database objects. In addition, Qupe uses HTTP over Secure Sockets Layer (SSL) to secure all web communications.

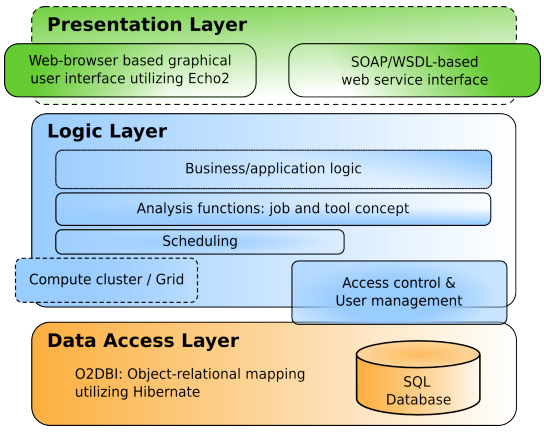

Fig. 1. This diagram depicts the three tier architecture model of Qupe. The data access layer provides an object-relational mapping utilising Hibernate. The implementation of the application or business logic is located in the second layer, including the framework for the execution of computationally intensive tasks. The presentation layer is separated in two distinct components. A graphical user interface allows the interaction with the system through a standard web browser, and a SOAP/WSDL-based web service can be utilised by other applications for data exchange.

Data access layer

Our data model is strongly adapted to the suggestions made by the Proteomics Standards Initiative (PSI) at the HUPO (Orchard *et al.*, 2003). Storage of mass spectra data follows the open source format mzData (Orchard *et al.*, 2004) developed by the PSI. Further aspects of the data model, which are realised in accordance to the PSI recommendations,

concern the stored data about reported peptides and proteins, which are nowadays described in the recently introduced analysisXML (Proteomics Informatics Standards Group, 2008). Particular emphasis was placed on the storage of analysis results such as calculated abundance ratios, visualisations or the output of statistical tests. To cope with future requirements for the data model and facilitate the addition of further attributes or classes, the development followed the model driven architecture (MDA) approach (Object Management Group, 2008) using the model designer O2DBI (Linke, B., unpublished data). The implementation utilises the Hibernate library (Red Hat Middleware, 2008).

Logic layer

Qupe includes several analysis functions for datasets such as those resulting from time series experiments. Furthermore, a well-defined programming interface (API) allows an easy development of new functions, supporting the retrieval and processing of data as well as the storing and visualisation of results of an analysis such as new datasets or graphics. The API supports the integration of routines written in R (R Development Core Team, 2008; Chair for computeroriented statistics and data analysis, 2008) allowing developers to resort to a wealth of established data analysis methods. A Sun Grid Engine/DRMAA binding (Sun Microsystems, 2009) has been incorporated, that enables computationally intensive tasks to benefit from the advantages of a distributed computing solution.

Presentation layer

A graphical user interface, implemented using the Echo2 web framework (NextApp, Inc., 2008), allows the interaction with the system through a standard web browser. At second, Qupe provides a web service interface based on SOAP and the web service description language (WSDL) (Gudgin *et al.*, 2008), which can be utilised by other applications to exchange analysis results as for example to retrieve complete datasets of calculated abundance ratios.

3 RESULTS

In the following important aspects and parts of Qupe are described in detail. We propose a workflow to quantitatively analyse isotopically labelled data from LC-MS/MS experiments as depicted in Figure 2.

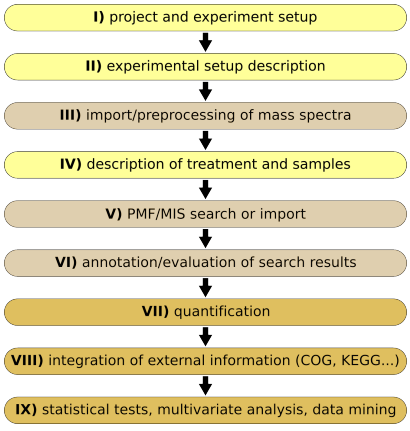

Fig. 2. The diagram depicts a proposed workflow to quantitatively analyse isotopically labelled data from LC-MS/MS experiments: from experiment and project setup (I, II and IV) to mass spectra import and database search (III,V and VI) to quantification and further analysis (VII-IX).

I) Project and experiment setup The web browser-based application provides extensive capabilities to group and integrate all data relevant to a particular experiment. This comprises a description of the experimental setup as well as mass spectra data and analysis results. Database access is firstly secured by a generalised project management system (GPMS), and secondly, fine-grained privileges may be assigned to individual experiments and projects.

II) Experimental setup description Qupe supports the description of experimental setups to allow for future retrieval of information about an experiment such as treatment of individual samples. Therefore, a number of predefined worksteps are provided that may be enhanced with additional details. Several worksteps may then be combined to describe the complete workflow of an experiment. A sample workstep would for example describe the cultivation of organisms including parameters such as optical density or growth medium.

III) Data acquisition: import/preprocessing of mass spectra Qupe currently allows the import of mass spectra data in the open source formats mzXML (Pedrioli *et al.*, 2004) and mzData (Orchard *et al.*, 2004). The system primarily targets at the analysis of LC-MS/MS data, but has also been designed to handle other types of data. As such a proprietary format by Bruker (Bruker Daltonics, Billerica, MA) for single-stage mass spectrometry data recorded by a Maldi-ToF instrument is already supported. Imported mass spectra can be visualised (see Figure 3A), and currently implemented tools support the preprocessing of MS/MS spectra, for example to filter mass spectra having a total ion current value below a certain threshold.

IV) Description of treatment and samples For an experiment one or more types of treatment, such as temperature or concentration of a substance, may be defined and furthermore divided into levels, e.g. 10 and 20 degrees celcius for the type temperature. To support the user in finding an appropriate terminology the ontology lookup service of the EBI may be queried (Côté *et al.*, 2006; Martens *et al.*, 2005). Individual samples (datasets) of an experiment can then be assigned to the defined levels and handled accordingly in further analysis. If for example samples were taken in distinct time intervals, therefrom calculated abundance ratios will be grouped in separate datasets that can then be compared to each other using statistical inference methods.

V) PMF/MIS search or import Peptide mass fingerprinting or MS/MS ion search can be carried out by an integrated Mascot (TM) search engine (Perkins *et al.*, 1999). Searches of the same set of mass spectra may be batch processed for example by means of the definition of ranges for peptide tolerance values or by querying several databases at once. Additionally, Qupe supports the import of DTASelect-filter files (Tabb *et al.*, 2002), so that further analysis can be based for example on Sequest (TM) (Yates *et al.*, 1995) results.

VI) Annotation/evaluation of search results To ensure that further analysis rests on a solid ground of verified peptide or protein identifications, it is necessary to assess the reported hits produced by database search tools. In Qupe, this can be based upon the calculation of false discovery rates (FDR) as suggested by Reidegeld *et al.* (2008). The preconditions for this are that concatenated decoy databases (Peng *et al.*, 2003; Elias and Gygi, 2007) have been employed. In the first instance all peptide or protein hits that were

either imported or reported by the integrated Mascot (TM) search engine are stored in database. Based on user-defined parameters such as the exclusion of specific charge states, a certain FDR-threshold, or, alternatively, a minimal score value, reported hits are filtered to gain the set of proteins and peptides that will be included in further analysis.

VII) Quantification Isotopic labelling techniques allow the measurement of relative abundances of several hundreds of proteins or peptides. Qupe supports the import of ProRata quantification results, and provides own implementations of quantification algorithms (see supplementary data for a description of an algorithm integrated in Qupe).

VIII) Integration of external information To extend the knowledge about identified proteins information from external resources such as Uniprot (UniProt Consortium, 2008) or KEGG (Kanehisa and Goto, 2000) can be integrated. This comprises COG or KOG (Tatusov *et al.*, 2003) classes and numbers, or EC numbers and pathway information. If protein identifiers have been derived from the GenDB annotation system (Meyer *et al.*, 2003), a mapping onto regions via BRIDGE (Goesmann *et al.*, 2003) is also available. This information can then be used for example to calculate the distribution of COG categories. Another function, which is integrated in Qupe, allows to map identified proteins and their calculated abundance ratios on KEGG pathways (see Figure 3B).

IX) Statistical tests, multivariate analysis, data mining In many proteomics workflows, it has not been elucidated yet, which statistical analysis methods are suitable for the analysis of quantitative data. Qupe provides a number of analysis functions and guides an experimenter through the process of the statistical evaluation of abundance ratios. Currently, the software builds on established and well-known statistical methods, while it additionally eases the development of novel approaches utilising a well-defined API. The one-sample t-test, the analysis of variance and the non-parametric Kruskal-Wallis test have been adapted to quantitative proteomics data. To account for the multiple testing situation and to give control of the family wise error rate, resulting p-values can be adjusted using e.g. the methods of Bonferroni or Holm. Other functions that Qupe suggests for data analysis are the principal component analysis (PCA) and hierarchical clustering algorithms using Ward's method, complete and average linkage and Euclidean as well as correlation based distances. The PCA is used to analyse covariances, and may thereby reveal the intrinsic dimensionality of the data, while the hierarchical cluster analysis seeks to identify groups of co-regulated proteins. According to the defined type(s) of treatment and their levels (see VI) similarly (by means of a distance function) expressed proteins are grouped into clusters. Using colour-codes for the calculated ratios, results of such an analysis can be evaluated in form of a heatmap as shown in Figure 3C. A further aim of cluster analysis is to find an optimal number of clusters. For this purpose, Qupe provides cluster indices such as Calinski-Harabasz (Calinski and Harabasz, 1974), Index-I (Maulik and Bandyopadhyay, 2002) or Davies-Bouldin (Davies and Bouldin, 1979) (see Figure 3D).

4 APPLICATION STUDY

In this application study we want to demonstrate the capabilities of the rich internet application Qupe with the analysis of a MudPIT experiment conducted at the University of Bochum. Proteins from the gram-positive bacterium *Corynebacterium glutamicum* were scrutinised on hyperosmotic conditions - a stress stimulus the biotechnologically relevant organism may be exposed to during fermentation. Utilising the stable isotope labelling approach, bacteria were cultivated in media containing either ^{14}N or ^{15}N. Samples were taken before the osmotic shock, that was induced by adding sodium chloride, and after 15, 60, and 180 minutes. Each sample was analysed in an 8-step MudPIT experiment. Using Xcalibur mass spectra were recorded on a LTQ XL Orbitrap (Thermo Fisher Scientific Inc., Waltham, MA). Further details of this analysis are published elsewhere (Fränzel *et al.*, ????).
The resulting 38 datasets were converted into the open source format mzXML with the tool "ReAdW" (Keller *et al.*, 2002; Nesvizhskii *et al.*, 2003). Using the web interface of the software, these datasets were then imported into Qupe running on a server at Bielefeld University. Therefore, a new experiment with appropriate read and write permissions for the participating experimenters was created to hold all (further) information and data. Subsequentially, spectra were preprocessed to filter for low overall intensities or insufficient numbers of peaks in the data, and afterwards submitted to the Mascot (TM) search engine. The composite target decoy database of *C. glutamicum* was derived from the corresponding GenDB genome annotation project (Kalinowski *et al.*, 2003). Afterwards false discovery rates were calculated and used to filter the observed peptide hits. The automatic annotation tool retained 7258 peptide hits for further analysis, which in summary corresponded to 715 identified proteins. Information about the identified proteins was enriched by querying external resources for COG class names or EC numbers, finding for example more than 13 percent of all identified proteins in the functional category "Translation, ribosomal structure and biogenesis". Before peptide quantification took place the experimental factor "time" was set up and the imported samples were assigned to the four different values 0, 15, 60, and 180 minutes according to the timespan after shock. A univariate analysis of variance with the factor "time" revealed 39 proteins as significant differentially regulated regarding the four distinct timepoints. This includes some temperature shock proteins, a putative transcriptional regulator and a phosphoglycerate dehydrogenase. The hierarchical cluster analysis seeks to identify groups or clusters, respectively, of co-regulated proteins. A result of such an analysis can be a heatmap as shown in Figure 3C, or a division of all proteins in a number of clusters. Utilising the cluster index "Calinski-Harabasz" this optimal number of groups was for example found at 13 clusters for Euclidean distances and the average linkage method (see Figure 3D) in our application study.

5 DISCUSSION

A variety of desktop and web applications that aim at a similar set of functionality compared to Qupe are already available. In terms of data management this includes MASPECTRAS (Hartler *et al.*, 2007), a web application that supports the import of the results from several search engines, provides peptide validation, and quantification based on ASAPRatio. A unique feature of the system is an integrated algorithm to map identified peptides to proteins. This accounts for the problem that a single peptide is often shared by a group of proteins. Proteios (ProSE) (Gärdén *et al.*, 2005; Levander *et al.*, 2009) is another web application that

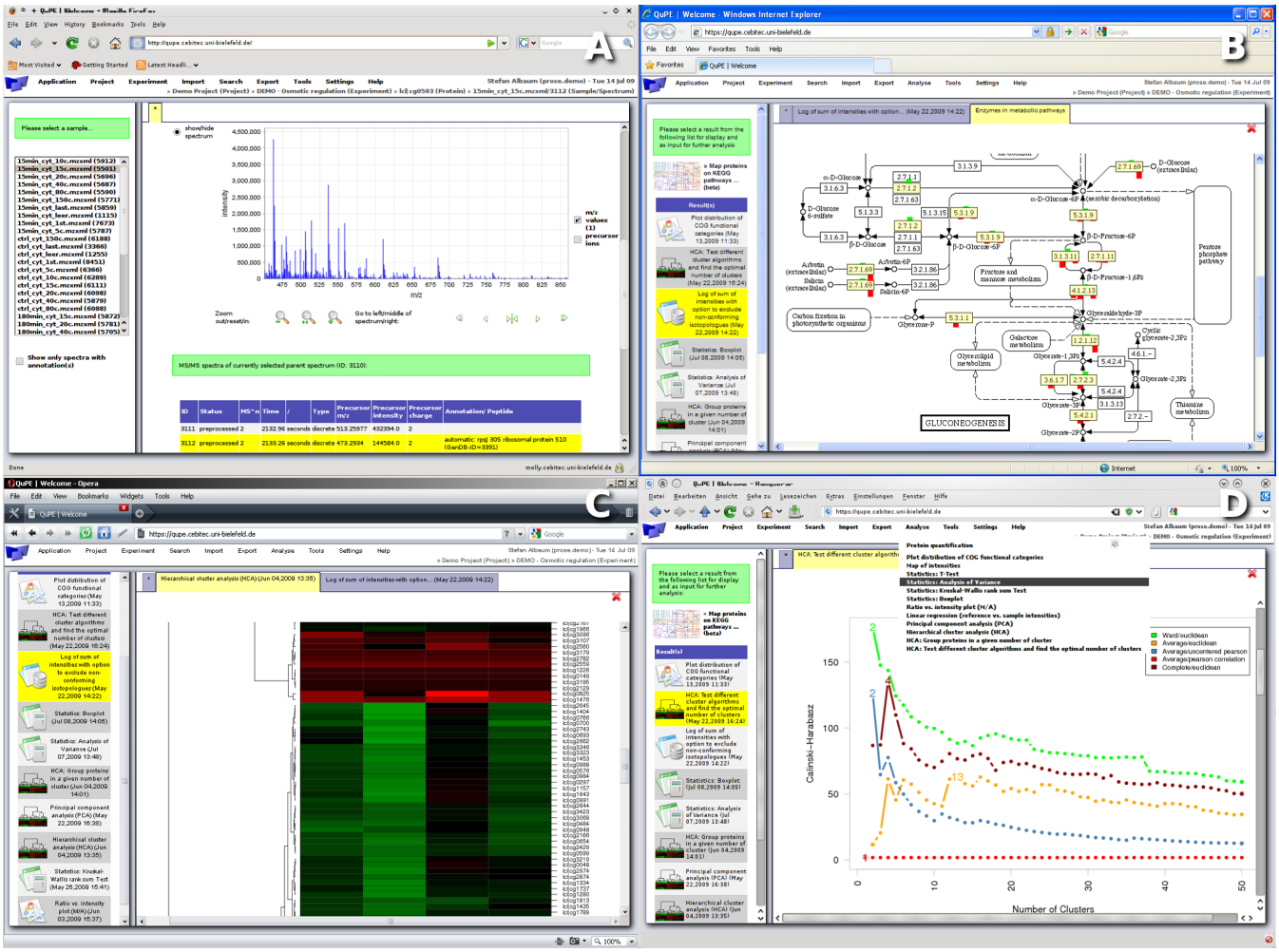

Fig. 3. Qupe provides a highly user-friendly web-interface. The screenshot in Figure 3A shows a view that allows to browse and analyse imported mass spectrometry data of a selected experiment. The titlebar includes the application's menu and informs about currently selected experiments and datasets. Each imported sample or run, respectively, can be found in the list on the left hand side of this view. If an entry is selected, the corresponding spectra information as for example zoomable mass-to-charge vs. intensity plots will show up in the right part of the view. Figure 3B shows a mapping of identified proteins (marked yellow) and calculated abundance ratios (indicated as red and green bars) on KEGG pathways. In this case, data of the described application study on *C. glutamicum* is projected on the pathway "Glycolysis/Gluconeogenesis". Figure 3C shows the results of a hierarchical cluster analysis using wards linkage method and Euclidean distances in form of a heatmap. The columns of the heatmap indicate the four different timepoints, that the samples have been taken at, while each rows stands for a protein. A better way to interpret the results of a cluster analysis is to utilise cluster indices. In the example shown in Figure 3D such an index (Calinski-Harabasz) has been computed. As one possible conclusion drawn from this index, the hierarchical cluster algorithm using average-linkage and Euclidean distances may be investigated further with the quantified proteins grouped into 13 clusters (see supplementary data for a high-resolution version of this Figure).

offers a comparable set of features like MASPECTRAS concerning data management, documentation of experimental processes, and search engine integration. Similar to Qupe, it furthermore provides a programming interface, that allows for further extensions of the system, and integrates a web service for database access. Another example of such systems is CPAS (Rauch *et al.*, 2006), which again features comprehensive data management functionalities, and a pipeline for protein identification and validation including the search engines X!Tandem, Mascot and Sequest. A further, detailed discussion and comparison of desktop and web applications including for example the Trans-Proteomics pipeline (TPP) (Keller *et al.*, 2002; Nesvizhskii *et al.*, 2003) can be found for example in Nesvizhskii *et al.* (2007), Mueller *et al.* (2008), and Hartler *et al.* (2007).

In direct comparison, it has to be considered that, particulary, Proteios and MASPECTRAS support more data formats and furthermore integrate additional search engines. However, while these applications focus on data management and the identification and evaluation of proteins from mass spectrometry data, Qupe goes one step further, and explores new frontiers of data analysis with the adaption of multivariate statistical methods to quantitative proteomics data. Qupe is highly extensible and eases the integration

of additional formats or tools as well as the development of novel methodologies. A well-defined API not only provides access to data stored in the system, but also unifies both configuration and execution of analysis functions and presentation of the results. We could already show the expandability through the integration of Maldi-ToF data and peptide mass fingerprinting. Furthermore, Qupe gives the opportunity to retrieve the data analysed within the system using a SOAP/WSDL-based webservice. The service has already been used to couple Qupe to ProMeTra, a web application to map expression values on biological pathways (Neuweger *et al.*, 2009).

6 CONCLUSIONS

We have designed and implemented the rich internet application Qupe with the first aim to provide a software package that supports the complete workflow of a proteomics experiment based on tandem mass spectrometry and stable isotopic labelling of proteins. This includes standardised data management, data integration, documentation of experimental processes, and, in particular, a guidance on applicable analysis methods. With the presented range of methods for statistical evaluation experimenters may draw reliable and meaningful conclusions from their data. Utilising comprehensive approaches such as cluster analysis algorithms, experimenters may identify co-regulated proteins, and thereby gain new insights into the mechanisms of protein biosynthesis. As a second aim, we wanted to bring algorithms closer to the biologists, and developed the software as a so called rich internet application. Qupe is accessible from any place where an internet connection is available. This enables sharing of information and data not only between different departments such as a laboratory and an office but also between different universities or institutions. Following the concept of software as a service any installation or requirement of maintenance is omitted while data integrity and security are conserved.

The range of functions of Qupe will be extended in the near future, where for instance other quantification algorithms will be supported, or new data format specifications will be regarded covering the recently released mzML (Mass Spectrometry Standards Working Group, 2008), and the analysisXML data format (Proteomics Informatics Standards Group, 2008).

AUTHORS CONTRIBUTIONS

SPA has designed the current version of Qupe. Together with SL he implemented most parts of the application. BF and CT devised and performed the use case experiment. HN contributed to mass spectrometry data handling and preprocessing. DM supported the implementation of the quantification algorithm. JK and DW contributed to the biological background, TWN and AG initiated, supervised, and directed the whole project. All authors have read and approved the manuscript.

ACKNOWLEDGMENTS

SPA and SL received financial support from the BMBF in the frame of the QuantPro initiative [grant 0313812]. HN would like to thank the International Graduate School in Bioinformatics and Genome Research for providing financial support. The authors further wish to thank the BRF system administrators for expert technical support. We would especially like to thank the workgroups of D. Becher (Greifswald University) and A. Poetsch (Bochum University) who kindly provided datasets and material.

REFERENCES

Allaire, J. (2002). Macromedia flash mx - a next-generation rich client. Technical report, Macromedia white paper.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, **25**(1), 25–29.

Calinski, R. B. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, **3**, 1–27.

Chair for computeroriented statistics and data analysis (2008). rjava - low-level r to java interface. `http://rosuda.org/rJava/`.

Côté, R. G., Jones, P., Apweiler, R., and Hermjakob, H. (2006). The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, **7**, 97.

Cox, J. and Mann, M. (2008). Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*, **26**(12), 1367–1372.

Craig, R. and Beavis, R. C. (2004). Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, **20**(9), 1466–1467.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**, 224–227.

Dondrup, M., Albaum, S. P., Griebel, T., Henckel, K., Jünemann, S., Kahlke, T., Kleindt, C. K., Küster, H., Linke, B., Mertens, D., Mittard-Runte, V., Neuweger, H., Runte, K. J., Tauch, A., Tille, F., Pühler, A., and Goesmann, A. (2009). Emma 2–a mage-compliant system for the collaborative analysis and integration of microarray data. *BMC Bioinformatics*, **10**, 50.

Elias, J. E. and Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, **4**(3), 207–214.

Fränzel, B., Trötschel, C., Rückert, C., Kalinowski, J., Poetsch, A., and Wolters, D. A. (????). Adaptation of corynebacterium glutamicum to salt-stress conditions. Manuscript submitted for publication in Proteomics.

Gärdén, P., Alm, R., and Häkkinen, J. (2005). Proteios: an open source proteomics initiative. *Bioinformatics*, **21**(9), 2085–2087.

Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004). Open mass spectrometry search algorithm. *J Proteome Res*, **3**(5), 958–964.

Goesmann, A., Linke, B., Rupp, O., Krause, L., Bartels, D., Dondrup, M., McHardy, A. C., Wilke, A., Pühler, A., and Meyer, F. (2003). Building a bridge for the integration of heterogeneous data from functional genomics into a platform for systems biology. *J Biotechnol*, **106**(2-3), 157–167.

Gudgin, M., Hadley, M., Mendelsohn, N., Moreau, J.-J., Nielsen, H. F., Karmarkar, A., and Lafon, Y. (2008). Soap version 1.2. `http://www.w3.org/TR/soap12-part1/`.

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*, **17**(10), 994–999.

Hartler, J., Thallinger, G. G., Stocker, G., Sturn, A., Burkard, T. R., Körner, E., Rader, R., Schmidt, A., Mechtler, K., and Trajanoski, Z. (2007). Maspectras: a platform for management and analysis of proteomics lc-ms/ms data. *BMC Bioinformatics*, **8**, 197.

Hufnagel, P. and Rabus, R. (2006). Mass spectrometric identification of proteins in complex post-genomic projects. soluble proteins of the metabolically versatile, denitrifying 'aromatoleum' sp. strain ebn1. *J Mol Microbiol Biotechnol*, **11**(1-2), 53–81.

Interface21 (2008). Spring framework. `http://www.springframework.org`.

Johnson, R. (2003). *Expert One-on-One J2EE Design and Development*. Wiley Publishing, Inc.

Kalinowski, J., Bathe, B., Bartels, D., Bischoff, N., Bott, M., Burkovski, A., Dusch, N., Eggeling, L., Eikmanns, B. J., Gaigalat, L., Goesmann, A., Hartmann, M., Huthmacher, K., Krämer, R., Linke, B., McHardy, A. C., Meyer, F., Möckel, B., Pfefferle, W., Pühler, A., Rey, D. A., Rückert, C., Rupp, O., Sahm, H., Wendisch, V. F., Wiegräbe, I., and Tauch, A. (2003). The complete corynebacterium glutamicum atcc 13032 genome sequence and its impact on the production of l-aspartate-derived amino acids and vitamins. *J Biotechnol*, **104**(1-3), 5–25.

Kanehisa, M. and Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **28**(1), 27–30.

Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal Chem*, **74**(20), 5383–5392.

Kohlbacher, O., Reinert, K., Gröpl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., and Sturm, M. (2007). Topp–the openms proteomics pipeline. *Bioinformatics*, **23**(2), e191–e197.

Kumar, C. and Mann, M. (2009). Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett*.

Levander, F., Hakkinen, J., Vincic, G., Mansson, O., and Warell, K. (2009). The proteios software environment - an extensible multi-user platform for management and analysis of proteomics data. *J Proteome Res*.

Li, X. J., Zhang, H., Ranish, A., and Aebersold, R. (2003). Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal. Chem.*, **75**(23), 6648–6657.

MacCoss, M. J., Wu, C. C., Liu, H., Sadygov, R., and Yates, J. R. (2003). A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal. Chem.*, **75**(24), 6912–6921.

Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J., and Apweiler, R. (2005). Pride: the proteomics identifications database. *Proteomics*, **5**(13), 3537–3545.

Mass Spectrometry Standards Working Group (2008). mzml 1.0.0 specification. `http://psidev.info/index.php?q=node/257`.

Maulik, U. and Bandyopadhyay, S. (2002). Performace evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(12), 1650–1654.

Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R., and Pühler, A. (2003). Gendb–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res*, **31**(8), 2187–2195.

Mueller, L. N., Brusniak, M.-Y., Mani, D. R., and Aebersold, R. (2008). An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J Proteome Res*, **7**(1), 51–61.

Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, **75**(17), 4646–4658.

Nesvizhskii, A. I., Vitek, O., and Aebersold, R. (2007). Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods*, **4**(10), 787–797.

Neuweger, H., Albaum, S. P., Dondrup, M., Persicke, M., Watt, T., Niehaus, K., Stoye, J., and Goesmann, A. (2008). Meltdb: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics*, **24**(23), 2726–2732.

Neuweger, H., Persicke, M., Albaum, S., Bekel, T., Dondrup, M., Huser, A., Winnebald, J., Schneider, J., Kalinowski, J., and Goesmann, A. (2009). Visualizing post genomics data-sets on customized pathway maps by prometra - aeration-dependent gene expression and metabolism of corynebacterium glutamicum as an example. *BMC Syst Biol*, **3**(1), 82.

NextApp, Inc. (2008). Echo web framework. `http://echo.nextapp.com`.

Object Management Group, I. (2008). Omg model driven architecture. `http://www.omg.org/mda/`.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, **1**(5), 376–386.

Orchard, S., Hermjakob, H., and Apweiler, R. (2003). The proteomics standards initiative. *Proteomics*, **3**(7), 1374–1376.

Orchard, S., Hermjakob, H., Julian, R. K., Runte, K., Sherman, D., Wojcik, J., Zhu, W., and Apweiler, R. (2004). Common interchange standards for proteomics data: Public availability of tools and schema. *Proteomics*, **4**(2), 490–491.

Pan, C., Kora, G., Tabb, D. L., Pelletier, D. A., McDonald, W. H., Hurst, G. B., Hettich, R. L., and Samatova, N. F. (2006). Robust estimation of peptide abundance ratios and rigorous scoring of their variability and bias in quantitative shotgun proteomics. *Anal Chem*, **78**(20), 7110–7120.

Park, S. K., Venable, J. D., Xu, T., and Yates, J. R. (2008). A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat Methods*, **5**(4), 319–322.

Pedrioli, P. G. A., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004). A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol*, **22**(11), 1459–1466.

Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003). Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (lc/lc-ms/ms) for large-scale protein analysis: the yeast proteome. *J Proteome Res*, **2**(1), 43–50.

Perkins, D., Pappin, D., Creasy, D., and Cottrell, J. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**(18), 3551–3567.

Polpitiya, A. D., Qian, W.-J., Jaitly, N., Petyuk, V. A., Adkins, J. N., Camp, D. G., Anderson, G. A., and Smith, R. D. (2008). Dante: a statistical tool for quantitative analysis of -omics data. *Bioinformatics*, **24**(13), 1556–1558.

Proteomics Informatics Standards Group (2008). analysisxml. `http://psidev.info/index.php?q=node/319`.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramos, H., Shannon, P., and Aebersold, R. (2008). The protein information and property explorer: an easy-to-use, rich-client web application for the management and functional analysis of proteomic data. *Bioinformatics*, **24**(18), 2110–2111.

Rauch, A., Bellew, M., Eng, J., Fitzgibbon, M., Holzman, T., Hussey, P., Igra, M., Maclean, B., Lin, C. W., Detter, A., Fang, R., Faca, V., Gafken, P., Zhang, H., Whiteaker, J., Whitaker, J., States, D., Hanash, S., Paulovich, A., and McIntosh, M. W. (2006). Computational proteomics analysis system (cpas): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J Proteome Res*, **5**(1), 112–121.

Red Hat Middleware, L. (2008). Hibernate. `http://www.hibernate.org`.

Reidegeld, K. A., Eisenacher, M., Kohl, M., Chamrad, D., Körting, G., Blüggel, M., Meyer, H. E., and Stephan, C. (2008). An easy-to-use decoy database builder software tool, implementing different decoy strategies for false discovery rate calculation in automated ms/ms protein identifications. *Proteomics*, **8**(6), 1129–1137.

Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., and Kohlbacher, O. (2008). Openms - an open-source software framework for mass spectrometry. *BMC Bioinformatics*, **9**, 163.

Sun Microsystems (2009). Sun grid engine. `http://gridengine.sunsource.net`.

Tabb, D. L., McDonald, W. H., and Yates, J. R. (2002). Dtaselect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res*, **1**(1), 21–26.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. (2003). The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**(41), 1–14.

Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P.-A., Julian, R. K., Jones, A. R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E. W., Dunn, M. J., Heck, A. J. R., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T. A., Patterson, S. D., Ping, P., Seymour, S. L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T. M., Whitelegge, J. P., Wilkins, M. R., Xenarios, I., Yates, J. R., and Hermjakob, H. (2007). The minimum information about a proteomics experiment (miape). *Nat Biotechnol*, **25**(8), 887–893.

UniProt Consortium (2008). The universal protein resource (uniprot). *Nucleic Acids Res*, **36**(Database issue), D190–D195.

Wolters, D., Washburn, M., and Yates, J. (2001). An automated multidimensional protein identification technology for shotgun proteomic. *Anal. Chem.*, **73**(23), 5683–5690.

Yates, J. R., Eng, J. K., McCormack, A. L., and Schieltz, D. (1995). Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem*, **67**(8), 1426–1436.

Zhang, N., Aebersold, R., and Schwikowski, B. (2002). Probid: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, **2**(10), 1406–1412.

Zhu, H., Pan, S., Gu, S., Bradbury, E. M., and Chen, X. (2002). Amino acid residue specific stable isotope labeling for quantitative proteomics. *Rapid Commun Mass Spectrom*, **16**(22), 2115–2123.