EuPA OPEN PROTEOMICS XXX (2015) XXX–XXX

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: http://www.elsevier.com/locate/euprot

Proteopathogen2, a database and web tool to store and display proteomics identification results in the mzIdentML standard☆

Vital Vialas[a,b,*], *Concha Gil*[a,b]

[a] *Departamento de Microbiología II, Universidad Complutense Madrid (UCM), Spain*
[b] *Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain*

ARTICLE INFO

Article history:
Available online xxx

We want to dedicate this work to Juan Pablo's memory, for his dedication to standardization in proteomics and inspiring work in this field.

Keywords:
mzIdentML
Candida albicans
Web application
Database
Proteopathogen

ABSTRACT

The Proteopathogen database was the first proteomics online resource focused on experiments related to *Candida albicans* and other fungal pathogens and their interaction with the host. Since then, the HUPO-PSI standards were implemented and settled, and the first large scale *C. albicans* proteomics resource appeared as a *C. albicans* PeptideAltas. This has enabled the remodeling of Proteopathogen to take advantage and benefit from the use of the HUPO-PSI adopted format for peptide and protein identification mzIdentML and continue offering a centralized resource for *C. albicans*, other fungal pathogens and different cell lines proteomics data.

1. Introduction

The opportunist pathogenic fungus *Candida albicans*, under usual circumstances, is a harmless resident commensal in human mucous membranes of a large percentage of the population. However, taking advantage of weakened host immune defenses, for instance in immunocompromised cancer or AIDS patients, it may switch to its pathogenic status, overproliferating and becoming thus the main etiological agent of candidiasis, one of the most prevalent and costly types of fungal infections in global terms.

Proteomics studies have been addressed to study this commensal to pathogenic transition by approaching the dimorphic, yeast form to hyphal form switch [1,2], by specifically aiming at the study of some other clinically relevant biological processes such as apoptosis [3–5] or biofilm formation [6]; or targeting sets of proteins that interact first with the host like surface exposed and secreted proteins [6,7].

However, until recently, the resulting proteomics identification datasets were sparse and disseminated. The Proteopathogen database [8] was the first public online proteomics data repository specifically focused on experiments aimed at the study of *C. albicans* and other fungal species

☆ This new Proteopathogen database and web tool is public online at http://proteopathogen2.dacya.ucm.es.

* *Corresponding author at*: Departamento de Microbiología II, Universidad Complutense Madrid (UCM), Spain. Tel.: +34 913941743.
E-mail address: vvialasf@ucm.es (V. Vialas).

http://dx.doi.org/10.1016/j.euprot.2015.04.002

pathogenic traits. Since no standard format for peptide and protein identification results was available, Proteopathogen was developed to compile and display identification lists in different tabulated text formats depending on the software used to generate and process the results.

At that time, the HUPO – Proteomics Standards Initiative (PSI) already had a trajectory striving to highlight the importance of standardization and providing formats that would comply with MIAPE (*Minimum Information About a Proteomics Experiment*) guidelines as reviewed in Ref. [9]. Some *de facto* standard formats existed like mzXML and pepXML [10], but the advent, years later, of the HUPO-PSI approved formats for mass spectrometry output data [11] and for identification results [12] among others, surfaced the efforts and claims by the community to finally adopt formats to facilitate data comparison, exchange and verification. This also inspired and boosted the development of an assortment of format conversion tools and libraries [13,14], and stand-alone software for visualization of the content of the files in standard formats [15] but, most importantly for the purpose of this work, enabled the possibility for Proteopathogen to benefit from the mzIdentML adopted standard for identification results, incorporating it as the input data format and using it as inspiration for information display.

More recently, the most comprehensive, up to the current date, online *C. albicans* proteomics data repository was developed and integrated in PeptideAtlas [16]. These publicly available *C. albicans* results have been used to establish a new version of Proteopathogen with a solid foundation.

In this background, we present here a revisited Proteopathogen database and web based tool adapted to read and display peptide and protein identification data based upon the mzIdentML format. It is the first online database specifically developed to map and store the contents of files in mzIdentML, it has been initially populated with the *C. albicans* PeptideAtlas identification results and it is publicly accessible at http://proteopathogen2.dacya.ucm.es/.

2. Materials and methods

The original identification result files were obtained from PeptideAtlas repository datasets PAe001976, PAe001977, PAe001978, PAe001979, PAe001980, PAe001981, PAe001982, PAe001983, PAe001984, PAe001985, PAe001986, PAe001987, PAe001988, PAe001989, PAe002110, and PAe002111.

As described in Ref. [16] the data sets come from a range of experiments including yeast to hypha transition assays, membrane protein extractions and a set of phosphoprotein enrichment approaches. In all cases, cells from the clinical isolates SC5314 were grown in YPD medium. For obtaining cells in hyphal form, either heat-inactivated fetal bovine serum or Lee medium pH 6.7 was used. As for the mass spectrometry, spectra were acquired in different set ups and platforms in a data-dependent manner. A summary of the experiments set ups and conditions is shown in Table 1.

Consistently with the PeptideAtlas project principles, the MS output files were processed through the Trans Proteomic Pipeline. The steps involved, first, sequence database searching using *X! Tandem* with *k-score* [18] and a custom sequence database obtained from Candida Genome Database [19] with appended decoy counterparts and common contaminants for peptide-to-spectrum matching and FDR assessment. Then the post-processing validation tools *PeptideProphet* [20], *ProteinProphet* [21] and *iProphet* [22] provided filtered lists of peptides and proteins with high probabilities. And finally FDR was computed for different probability thresholds.

Table 1 – Summary of experiments, MS output files, instrument and PeptideAtlas datasets.

Type of dataset	Number of MS output files	Instrument	Peptide Altas datasets
Candida albicans culture with SILAC labeling, digested protein extracts enriched in phosphopeptides IMAC/TiO2	57	Orbitrap XL, Orbitrap Velos	PAe001976 PAe001977 PAe001978 PAe001979 PAe001980 PAe001984 PAe001985 PAe001986 PAe001987 PAe001988 PAe001989
Candida albicans total protein extract, 2 Triple-TOF runs, 2 μg and 4 μg	2	Triple-TOF	PAe001983
HYPHAL form and yeast form total protein extracts	8	Orbitrap Velos	PAe002110 PAe002111
LTQ membrane proteins [17]	3	LTQ	PAe001981
LTQ proteins from acidic subproteome [1]	8	LTQ	PAe001982

Each of the PeptideAtlas repository datasets consists on the MS output spectra files and a set of pepXML and protXML files with lists of high confidence peptide and proteins respectively. These were combined, independently for each dataset, by means of a custom script written in the Ruby scripting language (available in supplemental data) to create mzIdentML files (mzIdentML version 1.1.0) with the merged information. In order to check the files were generated correctly and ensure data quality they were all validated (semantic and MIAPE-compliant validation) with mzidValidator [15].

A completely new MySQL relational database was implemented *ad hoc* to map elements in the mzIdentML files as depicted in Fig. 1 (schema available in supplemental data). Then, using the Ruby scripting language (version 2.0.0) and the Rails web application development framework (version 4.0.0) a script was created to parse the data in the mzIdentML files, store the relevant elements in the corresponding tables (available in supplemental data) and eventually create the web application to display the data.

3. Results and discussion

A total number of sixteen mzIdentML files, corresponding to each of the PeptideAtlas repository datasets, grouped into five different experiments were compiled and used to initially populate the Proteopathogen database. These account

Please cite this article in press as: Vialas V, Gil C. Proteopathogen2, a database and web tool to store and display proteomics identification results in the mzIdentML standard. EuPA Open Proteomics (2015), http://dx.doi.org/10.1016/j.euprot.2015.04.002

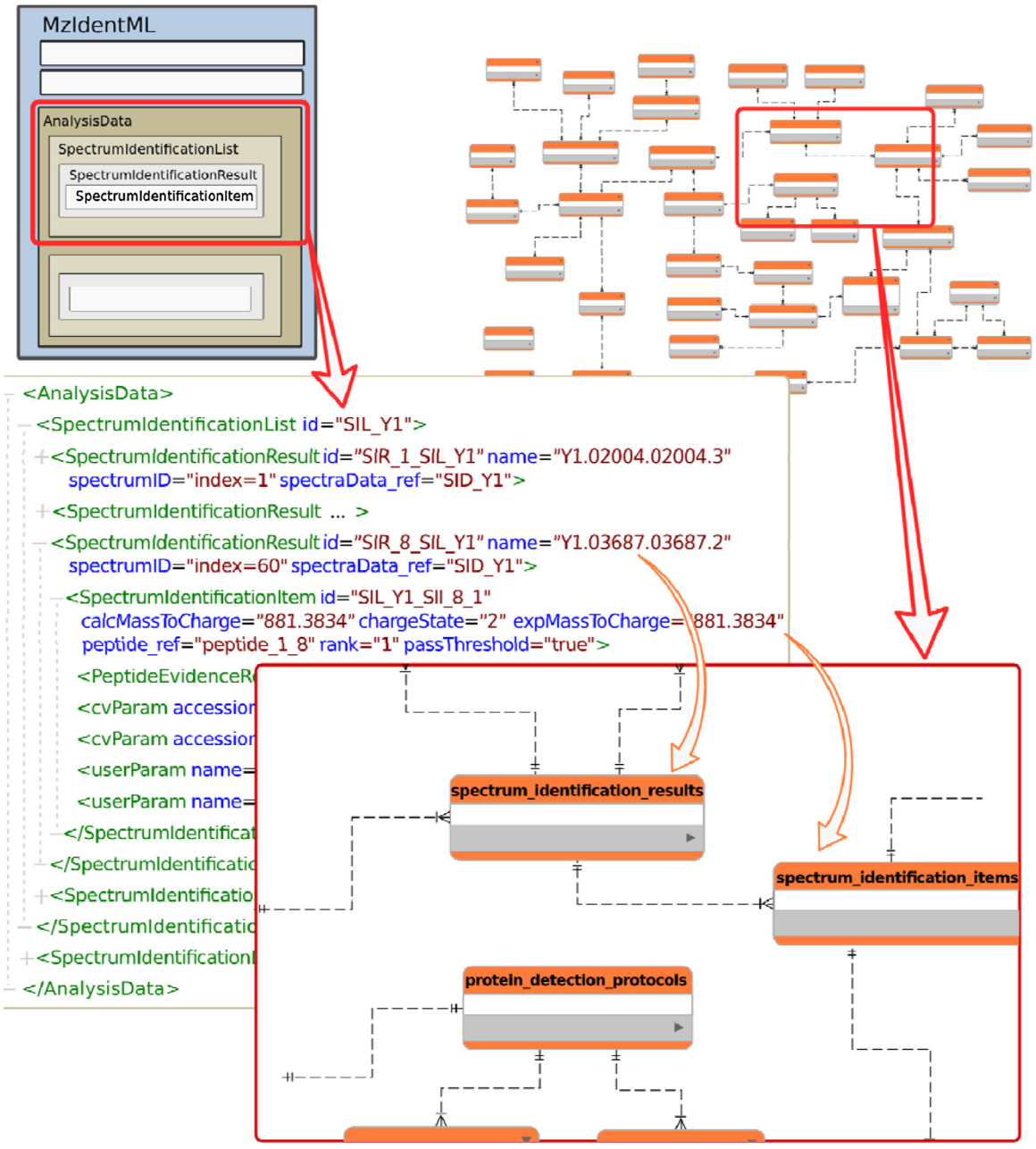

Fig. 1 – mzIdentML to database mapping. The MySQL schema was specifically designed to accommodate elements from the mzIdentML format. Figure shows the *one-to-many* relationship between the <*SpectrumIdentificationResult*> and <*SpectrumIdentificationItem*> elements.

for approximately 22,000 distinct peptides and 2600 different proteins that can be queried and viewed through the web interface.

Precisely, a stringent FDR cut-off at the PSM level set at 0.005, yields 21,883 peptides with 0.0024 FDR (peptide level) and 2577 proteins with 0.0170 FDR (protein level) as computed with Mayu, a software specifically designed to estimate accurate protein level error rates in large datasets [23] (see supplemental Table 1).

The mzIdentML contents can be browsed for each file in Proteopathogen in a means inspired by the structure in the format, particularly that under the <*AnalysisData*> element containing the datasets generated by the analyses. That is, for each mzIdentML file, shown in its experimental context, a user can select either the *spectrum identification* information (corresponding to the <*SpectrumIdentification*> element) and view its related information, the search protocol, search database and the list of every peptide to spectrum assignment; or the *protein detection* (corresponding to the <*ProteinDetection*> element) showing the list of peptides grouped into the inferred original proteins (Fig. 2).

Notably, the information Proteopathogen displays will depend on how complete the original mzIdentML files are. For instance, for files including the optional <*Fragmentation*>

Please cite this article in press as: Vialas V, Gil C. Proteopathogen2, a database and web tool to store and display proteomics identification results in the mzIdentML standard. EuPA Open Proteomics (2015), http://dx.doi.org/10.1016/j.euprot.2015.04.002

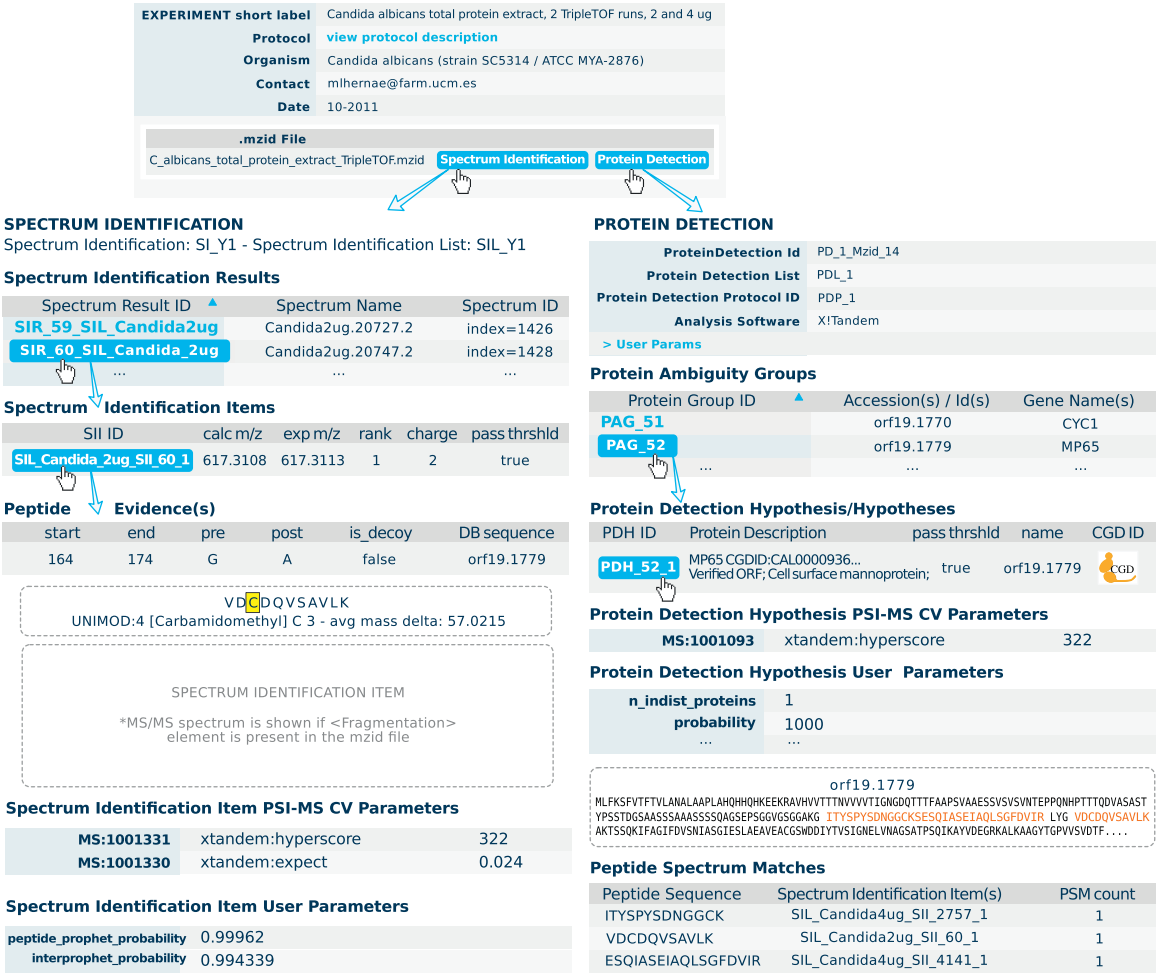

Fig. 2 – Information displayed in the web interface. Proteopathogen displays two main sets of information for the selected mzIdentML file. The *spectrum identification* section shows how for each spectrum there is a list of possible identification results, each having its *peptide evidence*, *i.e.* a sequence at a particular position in a protein sequence. The particular selected peptide is shown in its protein context in the *protein detection* section, which displays the complete list of the inferred proteins for the selected mzIdentML file with links to Candida Genome Database (CGD).

element under <*SpectrumIdentificationItem*>, Proteopathogen will display an annotated and interactive MS/MS spectrum. In addition, the optional <*cvParam*> and <*userParam*> elements, that describe and annotate with controlled vocabularies and user-defined information respectively different elements throughout the file, might be more or less profuse depending on the software that created them.

In addition to browsing through the contents of the stored mzIdentML files, Proteopathogen implements a query system yielding global results. That is, for a specific queried protein name, as found in the <*ProteinDetectionHypothesis*> name attribute, the search results display all the distinct peptide sequences found mapped into the protein sequence, regardless of the experiment in which they were identified, and the supporting spectra for each peptide sequence, while keeping track of the original <*SpectrumIdentification*> and mzIdentML file. A peptide sequence may also be searched, obtaining, when found, the corresponding protein, or group of proteins, and the set of supporting spectra, again in global scope.

The use of the Ruby scripting language, unlike other compiled languages (Java, C/C++) that are commonly used in other software used to visualize proteomics file formats, enables a quick, easy to implement, flexible manner of parsing complex XML files, and creation and manipulation of objects that have to be stored in a very precise order in a database. In addition, the argument of speed in computationally intensive tasks in favor of compiled languages is getting blurry nowadays with the array of *xml* parsing libraries that are continuously developed and improved for scripting languages. The type of solution implemented in Proteopathogen is a DOM (Document Object Model) parser, that creates an in-memory tree representation of the whole XML hierarchy. Arguably, a parser of the type SAX (Simple API for XML parsing) would perform better in terms of speed for large files but as trade-off, leaping back and forth in search of cross-referenced elements, as is the case in mzIdentML, would be difficult or even impossible to implement. Nevertheless, future work in the direction of a SAX implementation of the parser and a comparison in

Please cite this article in press as: Vialas V, Gil C. Proteopathogen2, a database and web tool to store and display proteomics identification results in the mzIdentML standard. EuPA Open Proteomics (2015), http://dx.doi.org/10.1016/j.euprot.2015.04.002

performance with respect to the current one, would be of great interest.

Proteopathogen will greatly benefit from the adoption of mzIdentML as input data format. Any proteomics experiment on *C. albicans*, or any fungal pathogen–host interaction, as long as they are provided in valid (semantically valid and MIAPE-compliant) mzIdentML (version 1.1.0), will be welcome to be integrated in the database. To that purpose, users provided with login credentials may submit their files either through a simple upload form in the web application or transfer them using a specifically set up FTP server. Finally, the Rails framework for web application development will take care of any scalability issues with ease and allow for any kind of visualization improvements.

4. Conclusions

The Proteopathogen web application and database has been completely rebuilt to accommodate and display *C. albicans*, or any fungal pathogen for that matter, proteomics identification results in the HUPO-PSI adopted format for peptide and protein identification mzIdentML. This makes it the first public online database specifically designed to store the information contained in these types of files and display its contents following an analogous structure.

Transparency document

The Transparency document associated with this article can be found in the online version.

Acknowledgements

This work has been financially supported by project BIO2012-31767 Ministerio de Economía y Competitividad, Spain, PROPMT (S2010/BMD-2414) from the Comunidad Autónoma de Madrid, REIPI, Spanish Network for the Research in Infectious Diseases (RD12/0015/0004), and PRB2 (PT13/0001/0004) from the ISCIII. VV held a research contract associated to project BIO2012-31767. The authors are grateful to Daniel Tabas-Madrid and Alberto Pascual Montano from CNB-CSIC for outstanding support and assistance in setting up the web application.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.euprot.2015.04.002.

REFERENCES

[1] Monteoliva L, Martinez-Lopez R. Quantitative proteome and acidic subproteome profiling of *Candida albicans* yeast-to-hypha transition. J Proteome Res 2010;10:502–17, http://dx.doi.org/10.1021/pr100710g.

[2] Gow N, van de Veerdonk F. *Candida albicans* morphogenesis and host defence: discriminating invasion from colonization. Nat Rev 2011;10:112–22, http://dx.doi.org/10.1038/nrmicro2711.

[3] Madeo F, Herker E, Wissing S. Apoptosis in yeast. Curr Opin Microbiol 2004;7:655–60, http://dx.doi.org/10.1016/j.mib.2004.10.012.

[4] Fernández-Arenas E, Cabezón V. Integrated proteomics and genomics strategies bring new insight into *Candida albicans* response upon macrophage interaction. Mol Cell Proteomics 2007;6:460–78, http://dx.doi.org/10.1074/mcp.M600210-MCP200.

[5] Ramsdale M. Programmed cell death in pathogenic fungi. Biochim Biophys Acta 2008;1783:1369–80, http://dx.doi.org/10.1016/j.bbamcr.2008.01.021.

[6] Vialás V, Perumal P, Gutierrez D, Ximénez-Embún P, Nombela C, Gil C, et al. Cell surface shaving of *Candida albicans* biofilms, hyphae and yeast form cells. Proteomics 2012, http://dx.doi.org/10.1002/pmic.201100588.

[7] Gil-Bona A, Llama-Palacios A, Parra CM, Vivanco F, Nombela C, Monteoliva L, et al. Proteomics unravels extracellular vesicles as carriers of classical cytoplasmic proteins in *Candida albicans*. J Proteome Res 2014, http://dx.doi.org/10.1021/pr5007944.

[8] Vialás V, Nogales-Cadenas R, Nombela C, Pascual-Montano A, Gil C. Proteopathogen, a protein database for studying *Candida albicans*–host interaction. Proteomics 2009;9:4664–8, http://dx.doi.org/10.1002/pmic.200900023.

[9] Martínez-Bartolomé S, Binz P-A, Albar JP. The Minimal Information about a Proteomics Experiment (MIAPE) from the Proteomics Standards Initiative. Methods Mol Biol 2014;1072:765–80, http://dx.doi.org/10.1007/978-1-62703-631-3_53.

[10] Deutsch E. File formats commonly used in mass spectrometry proteomics. Mol Cell Proteomics 2012;11:1612–21, http://dx.doi.org/10.1074/mcp.R112.019695.

[11] Martens L, Chambers M, Sturm M. mzML—a community standard for mass spectrometry data. Mol Cell Proteomics 2011;10, http://dx.doi.org/10.1074/mcp.R110.000133. R110.000133.

[12] Jones AR, Eisenacher M, Mayer G, Kohlbacher O, Siepen J, Hubbard SJ, et al. The mzIdentML data standard for mass spectrometry-based proteomics results. Mol Cell Proteomics 2012;11, http://dx.doi.org/10.1074/mcp.M111.014381. M111.014381.

[13] Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 2012;30:918–20, http://dx.doi.org/10.1038/nbt.2377.

[14] Griss J, Reisinger F, Hermjakob H, Vizcaíno JA. jmzReader: a Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats. Proteomics 2012;12:795–8, http://dx.doi.org/10.1002/pmic.201100578.

[15] Ghali F, Krishna R, Lukasse P, Martínez-Bartolomé S, Reisinger F, Hermjakob H, et al. Tools (Viewer, Library and Validator) that facilitate use of the peptide and protein identification standard format, termed mzIdentML. Mol Cell Proteomics 2013;12:3026–35, http://dx.doi.org/10.1074/mcp.O113.029777.

[16] Vialas V, Sun Z, Loureiro Y, Penha CV, Carrascal M, Abián J, et al. A *Candida albicans* PeptideAtlas. J Proteomics 2013, http://dx.doi.org/10.1016/j.jprot.2013.06.020.

[17] Cabezón V, Llama-Palacios A, Nombela C, Monteoliva L, Gil C. Analysis of *Candida albicans* plasma membrane proteome. Proteomics 2009;9:4770–86, http://dx.doi.org/10.1002/pmic.200800988.

[18] MacLean B, Eng J, Beavis R, McIntosh M. General framework for developing and evaluating database scoring algorithms

using the TANDEM search engine. Bioinformatics 2006;22:2830–2, http://dx.doi.org/10.1093/bioinformatics/btl379.
[19] Costanzo MC, Arnaud MB, Skrzypek MS, Binkley G, Lane C, Miyasato SR, et al. The Candida Genome Database: facilitating research on *Candida albicans* molecular biology. FEMS Yeast Res 2006;6:671–84, http://dx.doi.org/10.1111/j.1567-1364.2006.00074.x.
[20] Choi H, Nesvizhskii AI. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. J Proteome Res 2008;7:254–65, http://dx.doi.org/10.1021/pr070542g.
[21] Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 2003;75:4646–58.
[22] Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics 2011;10, http://dx.doi.org/10.1074/mcp.M111.007690. M111.007690-M111.007690.
[23] Reiter L, Claassen M, Schrimpf SSP, Jovanovic M, Schmidt A, Buhmann JM, et al. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics 2009;8:2405–17, http://dx.doi.org/10.1074/mcp.M900317-MCP200.

Please cite this article in press as: Vialas V, Gil C. Proteopathogen2, a database and web tool to store and display proteomics identification results in the mzIdentML standard. EuPA Open Proteomics (2015), http://dx.doi.org/10.1016/j.euprot.2015.04.002