EuPA OPEN PROTEOMICS 8 (2015) 22–27

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 6 May 2015

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

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

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.

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