The PSI Universal Spectrum Identifier (USI)

Eric Deutsch[1], Juan Antonio Vizcaíno[2], Yasset Perez-Riverol[2], Jeremy Carver[3], Benjamin Pullman[3], Shin Kawano[4], Zhi Sun[1], Luis Mendoza[1], Pierre-Alain Binz[5], Gerben Menschaert[6], and Nuno Bandeira[3]

[1] Institute for Systems Biology, Seattle; [2] European Bioinformatics Institute, Hinxton (EMBL-EBI); [3] University of California, San Diego; [4] Database Center for Life Science, Kashiwa, Japan; [5] CHUV Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland; [6] Biobix, Ghent University, Ghent, Belgium

Introduction

The Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI) (Deutsch et al.[1]) is in the process of developing the Universal Spectrum Identifier (USI) standard.

The USI will provide a standardized format for referring to a single publicly released spectrum from a dataset or from a spectral library.

The USI will enable communication of important mass spectral evidence both in publications and in software implementations.

The USI is still in development and we welcome your feedback!

Motivation

For many applications, there is a need to refer to a specific spectrum in a global way, especially when that spectrum provides evidence for an important new result.

Many journals require annotated spectra for identification claims. The Human Proteome Project (HPP) guidelines require annotated spectra as evidence for novel detection claims.

Reviewers and readers need to see spectra to make sure that the spectral evidence really supports the conclusions.

But all too often, the evidence is this:

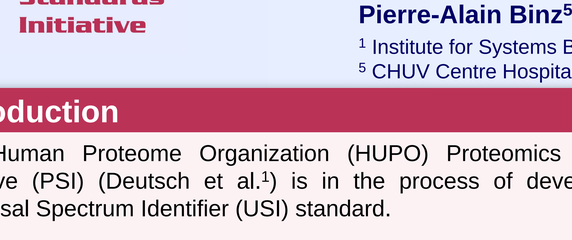


Figure 1. Example claimed evidence for peptide LNEEADGAQKLLVK.
Source: somewhere in PXD005940, but good luck finding which spectrum

Furthermore, there is a need to refer to specific spectra in articles or supplementary information.

There is a need to refer to specific spectra in a global way in software and in application programming interfaces (APIs)

There is a need to refer to spectra in specific datasets and also in specific spectral libraries.

There is a need to refer to origin spectra in spectral libraries themselves.

A spectral hash (SPLASH) (Wohlgemuth et al.[2]) mechanism has been proposed and is in use by the metabolomics community to refer to specific compound reference spectra, but for various reasons, this will not scale well to the billions of spectra available in proteomics data repositories.

USI Design

The USI is a multi-part key identifier of the form:

mzspec:<collection>:<msRun>:<indexType>:<indexNumber>:<optional interpretation>

Dataset spectrum example using native scan number:

mzspec:PXD002437:00261_A06_P001564_B00E_A00_R1:scan:10951

standard prefix	collection identifier	MS run identifier (fileroot of .raw, .mzML)	index flag	scan index

Dataset spectrum example using native scan number with optional interpretation:

mzspec:PXD002437:00261_A06_P001564_B00E_A00_R1:scan:10951:DLGNM[oxidation]EENK/2

spectrum interpretation

Implementations

The USI has a partial implementation at PeptideAtlas and MassIVE. The resource spectrum viewer will display the USI for any displayed spectrum if there is an associated PXD identifier.

Users can enter any valid USI, and if PeptideAtlas or MassIVE contains that PX dataset (it doesn't have all of them), then the spectrum will be fetched and displayed, irrespective of whether that spectrum passes quality filters.

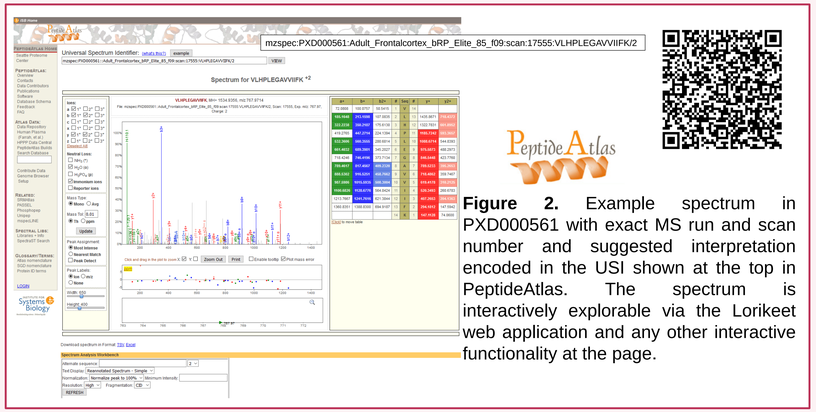


Figure 2. Example spectrum in PXD000561 with exact MS run and scan number and suggested interpretation encoded in the USI shown at the top in PeptideAtlas. The spectrum is interactively explorable via the Lorikeet web application and any other interactive functionality at the page.

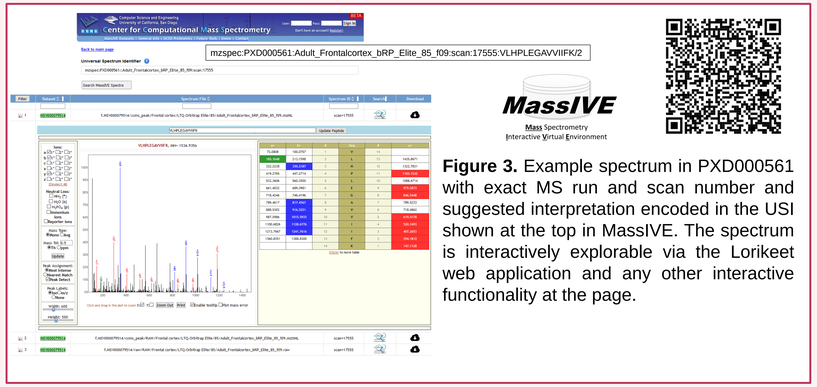


Figure 3. Example spectrum in PXD000561 with exact MS run and scan number and suggested interpretation encoded in the USI shown at the top in MassIVE. The spectrum is interactively explorable via the Lorikeet web application and any other interactive functionality at the page.

Next Steps

Early draft specification is available. Next, create a full PSI-format specification for PSI Document Process.

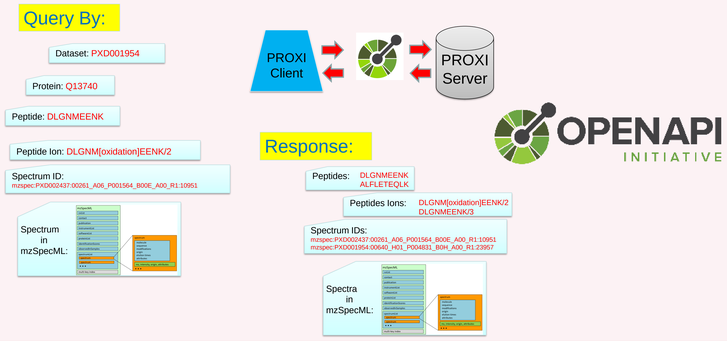
Create additional implementations at PRIDE and other resources (jPOST, sORFs.org are planning already).

Add to Human Proteome Project (HPP) Guidelines for next year's manuscripts

We welcome your feedback! Learn more at: **http://psidev.info/usi**

Related – PROXI

The in-development PROXI API will provide a mechanism to query for information in any of the ProteomeXchange repositories and potentially other resources using the same OpenAPI YAML API definition.



Related – PSI Spectral Library Format

The PSI is also developing a next-generation spectral library format, which will encode far more metadata about each entry, including USI references for all spectra in a cluster or consensus group.

Join the development!

References

1. Deutsch EW, Orchard S, Binz P-A, Bittremieux W, Eisenacher M, Hermjakob H, Kawano S, Lam H, Mayer G, Menschaert G, Perez-Riverol Y, Salek RM, Tabb DL, Tenzer S, Vizcaíno JA, Walzer M, Jones AR. Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. J Proteome Res. 2017 Dec 1;16(12):4288–4298. PMCID: PMC5715286
2. Wohlgemuth G, Mehta SS, Mejia RF, Neumann S, Pedrosa D, Pluskal T, Schymanski EL, Willighagen EL, Wilson M, Wishart DS, Arita M, Dorrestein PC, Bandeira N, Wang M, Schulze T, Salek RM, Steinbeck C, Nainala VC, Mistrik R, Nishioka T, Fiehn O. SPLASH, a hashed identifier for mass spectra. Nat Biotechnol. 2016 Nov 8;34(11):1099–1101. PMCID: PMC5515539

Acknowledgements

This work was supported by the National Institutes of Health NIGMS grants R24 GM127667 and R01 GM087221 and NIBIB grant U54 ES017885.