OpenSlice: Radical Data Sharing for Proteogenomics

Manor Askenazi[1], Kelly Ruggles[2], Jennifer Teubl[2], David Fenyo[2]

[1]Biomedical Hosting LLC, Arlington, M [2]NYU Langone Medical Center, New York

Introduction

The CPTAC project is providing unprecedented access to very large and systematically acquired clinical mass spectrometry data sets. These data sets, in combination with the matched, TCGA-derived genomic characterization, constitute a milestone for proteogenomics. In particular, the CPTAC colon dataset constitutes the largest genomically characterized mass spectrometry dataset acquired in full discovery mode. As such, the resulting data must be made amenable to facile and open-ended re-analysis to leverage our growing understanding of colon cancer and its biochemistry. To support such open-ended discovery work, in a distributed fashion and on global scale, we have developed a data system that can interactively host the complete raw data-set independent of any particular interpretation (by e.g. peptide identification software). The system is indexed to enable complex queries effectively opening up the data to remote mass informatics work, including both qualitative and quantitative queries.

The OpenSlice Mass Informatics Server indexes mass spectrometry raw files and exposes the resulting data through a RESTful web API, enabling the user to remotely access all aspects of the original files using a simple web browser, including access to MS1 scans which are typically ignored by most proteomics data-analytic pipelines. The system provides web-viewers for both raw MS1 and MS2 spectra as well as an overlay view interpreting the MS2 data based on a user provided peptide hypothesis. **The indexing process supports quantitative queries on a repository-wide scale, in particular the creation, in real-time of extracted ion chromatograms across all hosted clinical samples (potentially thousands of Raw files, i.e. thousands of acquisition hours).** This capability in turn enables hypothesis driven slicing of previously acquired data without the need to rerun a full identification pipeline.

A publicly accessible OpenSlice server (http://openslice.fenyolab.org/) has been implemented specifically to host the CPTAC colon dataset. It hosts all the colon Orbitrap-class data released to date, which constitutes 1425 raw files in turn representing thousands of instrument acquisition hours. The user then has the ability to investigate the files in their entirety, including both MS1 and MS2 data. The type of questions that can be asked using OpenSlice include: Which samples contain evidence for a variant peptide? What total signal strength (ion current) is associated with these observations? What are the clinical characteristics of the underlying samples/patients?

Information Architecture

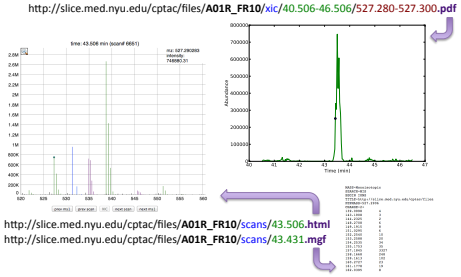
Some Design Principles:

Mass Spectra + URLs = HyperPeaks

1. Hyperlink or it didn't happen!!!
2. If it isn't quantitative, it isn't science.
3. Don't register, don't track, don't customize.
4. No downloads required: neither data nor software.
5. If the datapoints are static, *why are you using a database*?

The Slice API for Remote "QualBrowsing":

http://slice.med.nyu.edu/cptac/files/**A01R_FR10**/xic/40.506-46.506/527.280-527.300.**pdf**



http://slice.med.nyu.edu/cptac/files/**A01R_FR10**/scans/43.506.**html**

http://slice.med.nyu.edu/cptac/files/**A01R_FR10**/scans/43.431.**mgf**

Technological Enablers

The NoDB file format:

The premise of the NoDB data-layout is that if the data elements being shared are static and can be enumerated such that every element corresponds to a unique natural number (i.e. a natural index corresponding to a perfect minimal hash can be generated with ease), then there is no need for the typical machinery of the general purpose database server. Instead, a very simple indexing scheme can be implemented which enables answering an arbitrary object lookup request in just 2 disk accesses. The size of the object repository is limited only by disk space and, in principle, the system can be hosted on an unmodified web server (i.e., there is no need for an app-server layer).

For more information see: http://www.nodb.net

The mzdb file format:

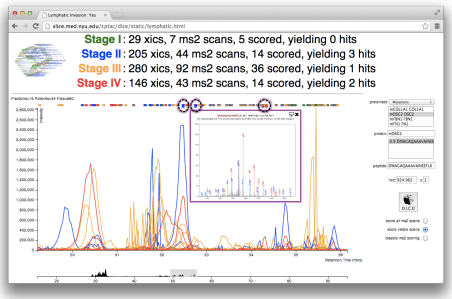
The mzdb format is a simple, fast and standard (i.e. sqlite-based, similar in spirit to PNNL's YAFMS) representation of an individual Raw file. The format is optimized for portability and speed; it is the representation used by the server when responding to Slice RESTful requests. For the D.I.C.E. interface (see below) the NoDB format is used as concurrent access to hundreds of files/XICs may be required. The mzdb format is implemented as part of the mz library of Python scripts. The "examples" folder contains an extractor for Thermo Rawfiles called liberate.py; the tool is also distributed as a Windows executable (no Python installation required).

For more information see: https://github.com/manor/mz

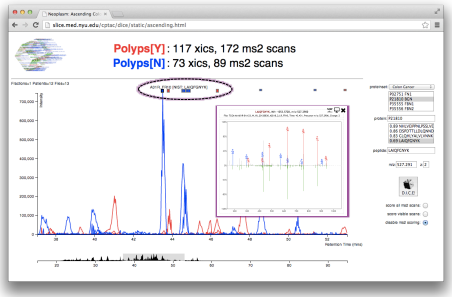
Total Ion Awareness with D.I.C.E.!

The **D**istributed **I**on **C**hromatogram **E**xtractor demonstrates the ability to extract arbitrary masses from a very large collection of Rawfiles interactively. The system annotates the XIC with any available MS2 data, highlighting the spectra that are readily identifiable using the NIST peptide library. Furthermore, the system will attempt (at the user's request) to annotate all unidentified spectra according to the user's hypothesis. Consequently the user can post ad-hoc requests without needing to re-analyze the entire corpus of acquired data.

Example Observations



Example1: On-the-fly identification of 5 patients with a DSG2 mutation, from a 44 patient sample pre-filtered by lymphatic invasion. The resulting XICs are colored by disease stage.



Example2: Interactive validation of peptide identifications using the NIST spectral library, within a quantitative context involving 12 patients with an Ascending Colon Neoplasm. The 13 resulting XICs are colored by history of Colon Polyps.

Acknowledgements:
This work has utilized computing resources at the High Performance Computing Facility of the Center for Health Informatics and Bioinformatics and New York University Langone Medical Center.