Creating a data analysis pipeline for producing high-quality glycopeptide mass spectral libraries

Sergey L. Sheetlin, Yuri A. Mirokhin, Guanghui Wang, Xiaoyu Yang, Concepcion A. Remoroza, Yi Liu, Xinjian Yan,
Dmitrii V. Tchekhovskoi, Zachary C. Goecker, Meghan C. Burke, Stephen E. Stein
Biomolecular Measurement Division, National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, USA

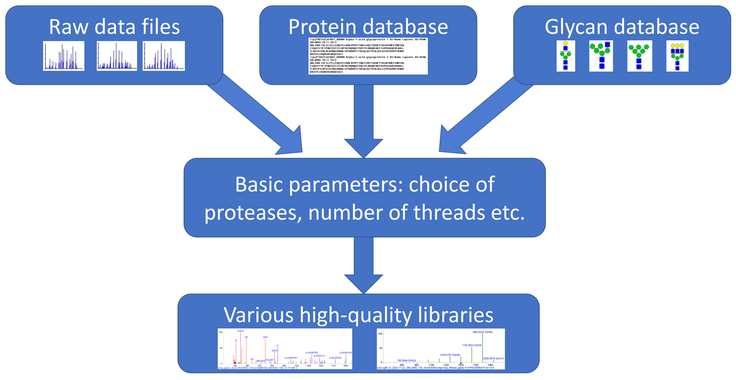
NIST National Institute of Standards and Technology U.S. Department of Commerce | Mass Spectrometry Data Center

chemdata.nist.gov

Introduction

- Previously[1], multiple tools developed at the NIST Mass Spectrometry Data Center were integrated into a single graphical user interface (GUI)
- This allowed creating series of sequentially operating programs to be configured and executed through a pre-defined 'pipeline'
- This poster describes extension of this pipeline for the specific task of comprehensive glycopeptide analysis and building high-quality glycopeptide mass spectral libraries from raw data
- The pipeline is implemented as a GUI with minimum required input parameters
- The generated libraries can be used with an updated version (not currently available for download) of the NIST Search Software[2]
- The NIST Search Software can be used for examining and comparing spectra in the resulting glycopeptide libraries

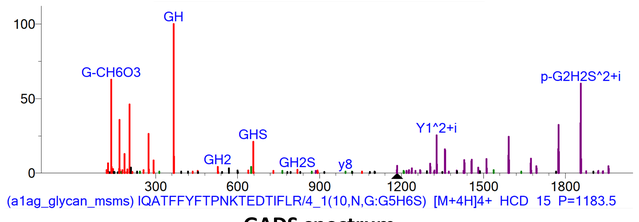
Input parameters



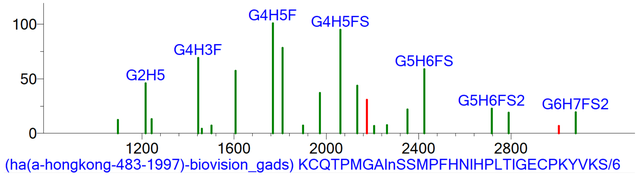
Types of generated libraries

(1) MS/MS glycopeptide libraries; (2) Glycopeptide abundance distribution spectra (GADS[3]) libraries; (3) Glycan libraries; (4) MS/MS peptide libraries.

Annotated glycopeptide spectrum



GADS spectrum



MSFragger[4,5] pipeline (for glycopeptide analysis)

ReAdw4Mascot2 (Dmitrii V. Tchekhovskoi): Input .RAW files are converted into the mzXML and MGF formats;
ProMS[6] (Xinjian Yan): analyzes search results (optional; run in parallel)

↓

FragPipe[4,5]: searching input RAW files, finding and validating glycopeptide matches.

↓

Format4gads (Guanghui Wang): combines and formats FragPipe output

↓

XICAnalyzer (Guanghui Wang): identifies peaks in an extracted ion chromatogram (XIC), calculates XIC theoretical isotope envelope, validates assignments of the search engine

↓

Validating spectra (Yuri A. Mirokhin): filtering spectra and creating libraries in the MSP text format

↓

MS_Piano[7] (Xiaoyu Yang): annotates peaks of peptide and glycopeptide spectra

↓

Make_GADS (Stephen E. Stein): builds GADS libraries and generates various summary statistics

↓

Building libraries in the NIST binary format (Dmitrii V. Tchekhovskoi): building libraries using the NIST tool Lib2NIST[2]

↓

Building Glycan libraries (Yuri A. Mirokhin)

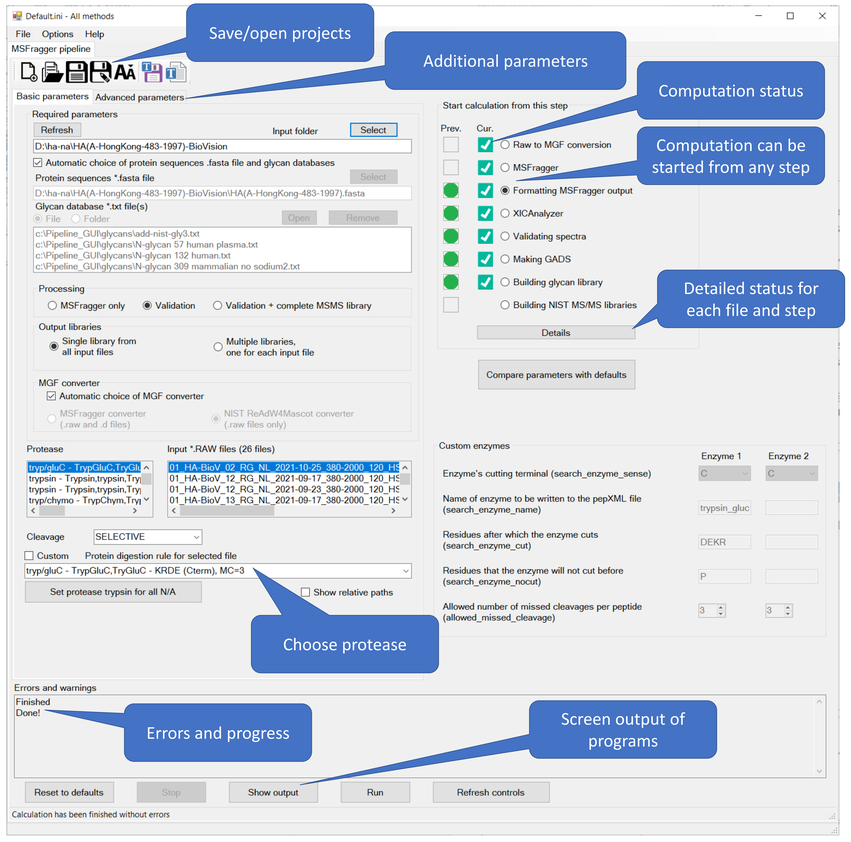
↓

MSPepSearch[2] (Dmitrii V. Tchekhovskoi): searches mass spectra in libraries using Hybrid Search[8] (optional)

- The GUI is implemented by Sergey Sheetlin
- FragPipe parameters are attuned by Guanghui Wang
- Extensive pipeline testing and conducting experiments are done by Stephen E. Stein, Yi Liu, Concepcion Remoroza, Zachary C. Goecker, Meghan C. Burke

Main features

- MSFragger and other components of FragPipe are used in the pipeline
- The computation is parallelized among computer cores
- There is an option for either N- or O-glycans
- A detailed description of the GUI is available through the menu "Help"
- Computation can be started from any step using the button "Run"
- The pipeline GUI analyzes the input settings and outputs warnings if some problems are detected during execution
- A batch software for processing multiple input data is available



Requirements

- The GUI is compatible with Windows 7, 10 and 11.
- **FragPipe** tools required licensing prior to use.

References

([1]) Sheetlin, Sergey L., at al, *Graphical User Interface for Integration of NIST Mass Spectrometry Data Center Library Building Tools.* Poster on ASMS 2021 Conference on Mass Spectrometry and Allied Topics.
([2]) NIST Search Software, NIST MSPepSearch, Lib2NIST: https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:nist17
([3]) Remoroza, C.A., et al., *Representing and Comparing Site-Specific Glycan Abundance Distributions of Glycoproteins.* J. Proteome Res. 2021, 20, 9, 4475–4486.
([4]) Kong, A. T., et al., *MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics.* Nature Methods 2017, 14, 5, 513-520.
([5]) FragPipe: https://fragpipe.nesvilab.org/
([6]) Rudnick, P.A., et al., *A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline.* J Proteome Res, 2016. **15**(3): p. 1023-32.
([7]) Yang, X., et al., *MS_Piano: A Software Tool for Annotating Peaks in CID Tandem Mass Spectra of Peptides and N-Glycopeptides.* J. Proteome Res. 2021, 20, 9, 4603–4609.
([8]) Meghan C. Burke, et al., *The Hybrid Search: A Mass Spectral Library Search Method for Discovery of Modifications in Proteomics.* J. Proteome Res. 2017, 16, 5, 1924–1935.

Restrictions and Disclaimers