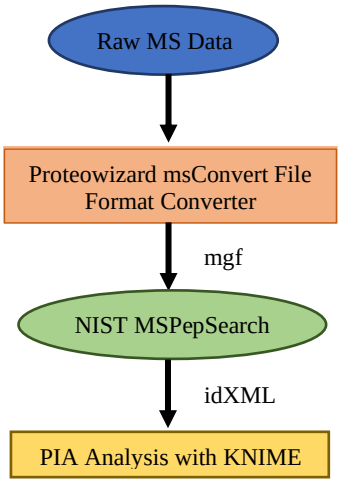
Tutorial: Spectral Library Search with NIST MSPepSearch and Error Rate Estimation



Contents:

Introduction

Mass spectral library searching is a valuable method for assigning peptide sequences to experimental tandem mass spectra. Advantages of using mass spectral libraries include a reduced search space and peak-intensity based scoring. Moreover, for well characterized samples, one may not expect to identify peptides that have not previously been observed. In fact, identifying a peptide that has previously been observed is expected for certain experiments. In addition to directly identifying peptide sequences, peptide mass spectral libraries may also be used to confirm peptide sequences identified by sequence database search algorithms and to identify recurring, unidentified mass spectra.

In this demonstration we will search experimental query mass spectra against a NIST MS Search binary format of the Consensus Human Orbitrap-HCD library (human_hcd_tryp_best, 398 373 mass spectra, 257 122 unique peptide sequences, 27.9% proteome coverage) as well as the corresponding reverse decoy mass spectral library[1]. The Consensus Human Orbitrap-HCD library was recently made publicly available and contains 86% more peptides than the previous version. The methods used to build the consensus libraries were presented at the ASMS Reboot 2020[2]. Freely available NIST peptide mass spectral libraries also include alternative ionization modes (i.e., ion trap), multiple species, modifications such as phosphorylation, and derivatization with iTRAQ-4 or TMT-10.

In this tutorial, we will analyze a raw data file from PXD014415[3], a ProteomeXchange dataset submitted by Joao Paulo. The raw data file we will process is a human HCT116 cell sample in which the cysteine residues in the extracted proteins have been alkylated using iodoacetamide

[1] https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:humanhcd20160503

[2] https://chemdata.nist.gov/dokuwiki/lib/exe/fetch.php?media=peptidew:sergey_sheetlin_asms2020.pdf

[3] Lim, M. Y., Paulo, J. A., & Gygi, S. P. (2019). Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. *Journal of proteome research*, *18*(11), 4020-4026.

followed by digestion by Lys-C and subsequently trypsin. The resulting peptides were analyzed on an Orbitrap Fusion Lumos mass spectrometer.

Preliminaries

Please download the following items from https://chemdata.nist.gov/:

- Search for the "ASMS 2021" page and download zipped file (6.6 GB in size) containing:
  - proteoNIST_v2.2.6 folder
  - proteoNIST shortcut
  - "ExampleFiles" folder containing:
    - TrypticHighQuality folder
    - Uniprot-all.fasta
    - a11853_human_90min_hrMS2.raw

Please note that this demonstration requires Microsoft Windows 10. For protein assembly, we will use the basic PIA (Protein Inference Algorithms) analysis[4,5] with KNIME (1.6 GB), an analytics platform. Briefly, PIA is an open source algorithm for protein inference that supports all current Proteomics Standards Initiative (PSI) standard formats and thereby input from almost all search engines. Here, we will import the OpenMS intermediate file format, idXML, generated by proteoNIST and use PIA analysis to view the relationships between PSMs, peptides and protein groups. To do so, please download and install KNIME. The instructions to install the PIA nodes have been adapted from a more detailed tutorial. These steps are:

1. To install the PIA nodes from the Community Node Repository, go to "Help">"Install New Software".
2. Under the "Work With" drop down menu, select "KNIME Community Extensions (Trusted) - http://update.knime.com/community-contributions/trusted/4.4".
3. The PIA nodes can be found in the "Bioinformatics & NGS" group or you may search for them. Next, select the PIA nodes. After accepting the license, restart KNIME.

Performing a Mass Spectral Library Search

To perform a mass spectral library search, we will begin by launching the proteoNIST_v2.2.6 shortcut provided. You should immediately be prompted to "Select NIST formatted spectral library". Please select the "TrypticHighQuality" folder, which includes the reference and reverse decoy mass spectral libraries described earlier. Note that this step simply directs the program to the directory containing "human_hcd_tryp_best" and "human_hcd_tryp_best_decoy"; therefore, it is not necessary to go into either of the library folders. After selecting the library, you should see the path selected at the top of the window as highlighted in red in Figure 1.

---

[4] Uszkoreit, J., Maerkens, A., Perez-Riverol, Y., Meyer, H. E., Marcus, K., Stephan, C., Kohlbacher, O., Eisenacher, M. (2015). PIA: an intuitive protein inference engine with a web-based user interface. *Journal of proteome research*, *14*(7), 2988-2997.

[5] Uszkoreit, J., Perez-Riverol, Y., Eggers, B., Marcus, K., & Eisenacher, M. (2018). Protein inference using PIA workflows and PSI standard file formats. *Journal of proteome research*, *18*(2), 741-747.

Next, select “File”>”Analyze”>”Process Thermo RAW file” followed by selection of the raw file we will be processing in this demonstration “a11853_human_90min_hrMS2.raw” located in the “ExampleFiles” folder. You will first see “MSConvert Raw extraction in progress” during which the vendor-specific format has been converted to mascot generic format (mgf) using ProteoWizard[6]. Next, you will notice “Processing results…” which indicates that the program has begun searching the tandem mass spectra provided (Figure 1).
At this time, a multiprocess implementation of NIST MSPepSearch is being performed. Both NIST MSPepSearch (batch processing) and NIST MS Search (single spectrum search) are also available for download separately[7]. The parameters used here include a 20 ppm precursor mass tolerance and 50 ppm fragment ion tolerance. You may notice that we have not specified any fixed or variable modifications. If so, you are very observant! In fact, with mass spectral library searching the modifications are part of the mass spectral annotation and are therefore not specified in the search.

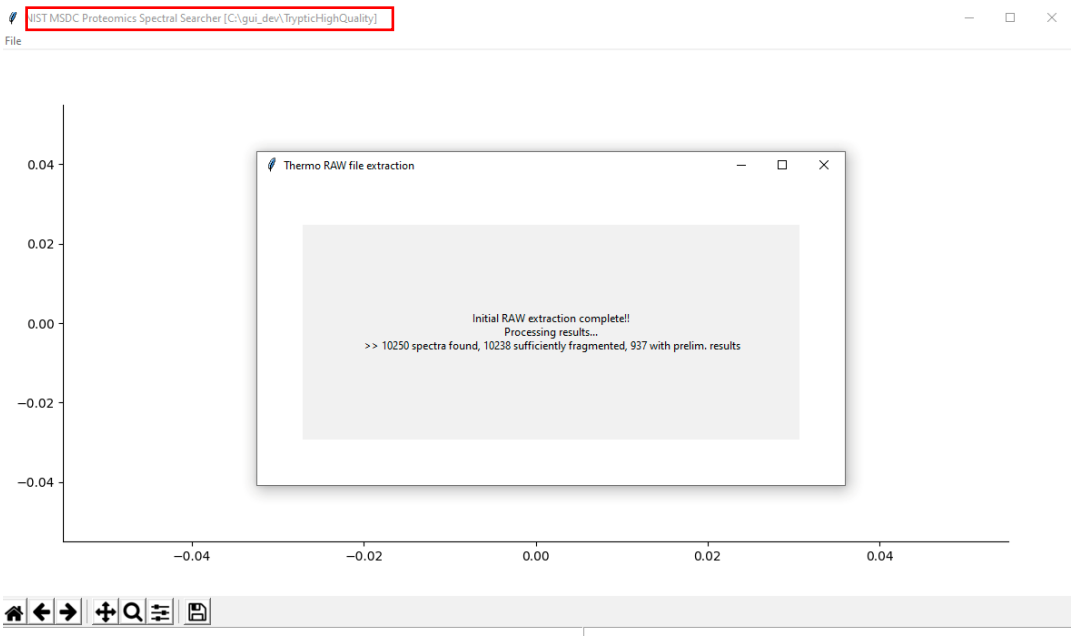


Figure 1: View of the proteoNIST GUI after conversion of a Thermo RAW file to mgf format has completed and processing with NIST MSPepSearch has begun.

The peptide spectral match (PSM) output to the GUI will include those whose spectral match factor (MF) falls above the threshold corresponding to an estimated 1% false discovery rate (FDR), based on the target decoy approach[8]; this score threshold is shown when the search is complete. The query tandem mass spectra are listed in the top left window and may be selected for single spectrum searching (Figure 2).

---

[6] Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology*, *30*(10), 918-920.

[7] https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:pepsoftware

[8] Elias, J. E., & Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, *4*(3), 207-214.

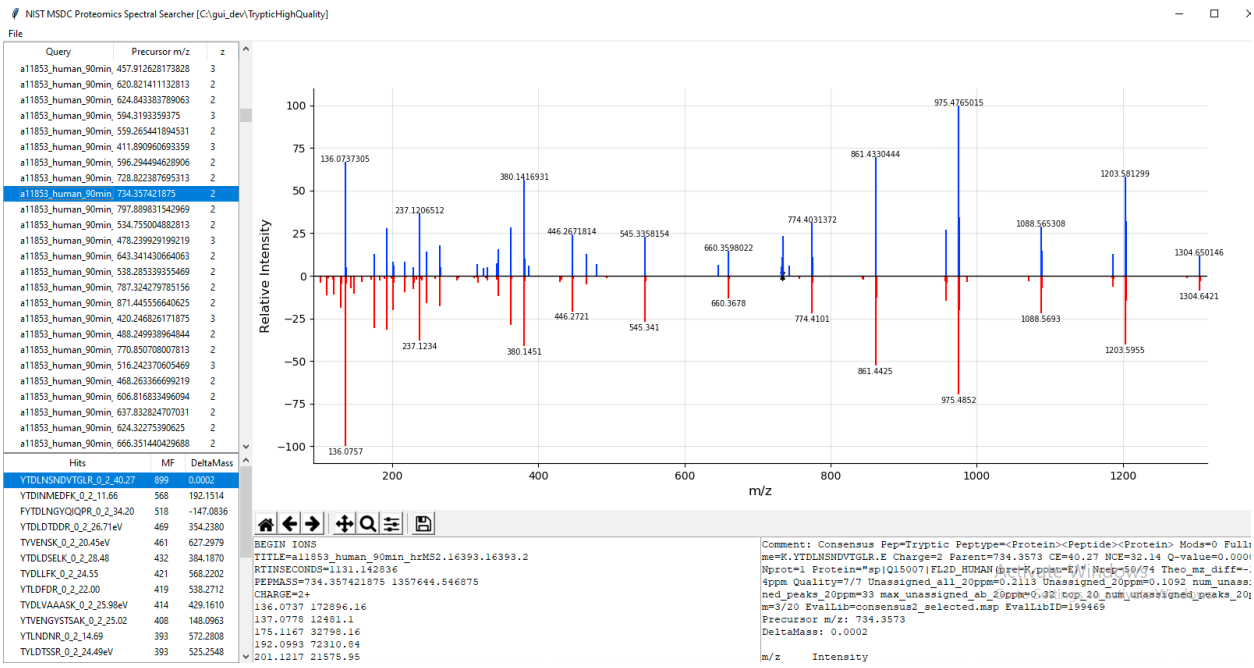


Figure 2: Single spectrum viewer for reference mass spectral library matches to query mass spectra for which there is a direct spectral match that meets the estimated 1% FDR threshold.

Inspecting Mass Spectral Library-derived Peptide Spectral Matches

With the direct search of the tandem mass spectra against the reference and decoy mass spectral libraries complete, we can now inspect mass spectral library matches to experimental mass spectra. It is important to note here that each time we click on a different query spectrum in the GUI, an on-the-fly direct and hybrid mass spectral library search is performed. The NIST MSPepSearch hybrid search is similar to a so-called open search in which the precursor mass tolerance is not restricted. Furthermore, each fragment ion in the experimental mass spectrum is allowed to match a fragment ion in the reference spectrum directly or after shifting by the difference in precursor mass (DeltaMass).

Let's begin by looking a good quality peptide spectral match. Please click on the query title "SCAN_6020". We will begin by drawing your attention to the match factor (MF) score listed as the second column in the hit list. The MF is calculated using a modified cosine of the angle between the query and reference library tandem mass spectra and therefore reflects the degree of spectral similarity. The MF can range from 0 to 999, with 999 being identical. The top hit for this query spectrum has a MF of 845 (Figure 3). If you click on the top-ranking sequence in the hit list, IGQQPQQPGAPPQQDYTK, we can see a head-to-tail comparison of the two tandem mass spectra. You will notice that the two spectra are indeed highly similar. Furthermore, you will notice that the second sequence in the hit list has a lower MF of 759 with a DeltaMass that corresponds to loss of the initial lysine (a modification that results in the same sequence as the top hit). This example is intended to highlight that inspecting the lower scoring reference library matches can be informative in judging whether an assignment is likely correct.

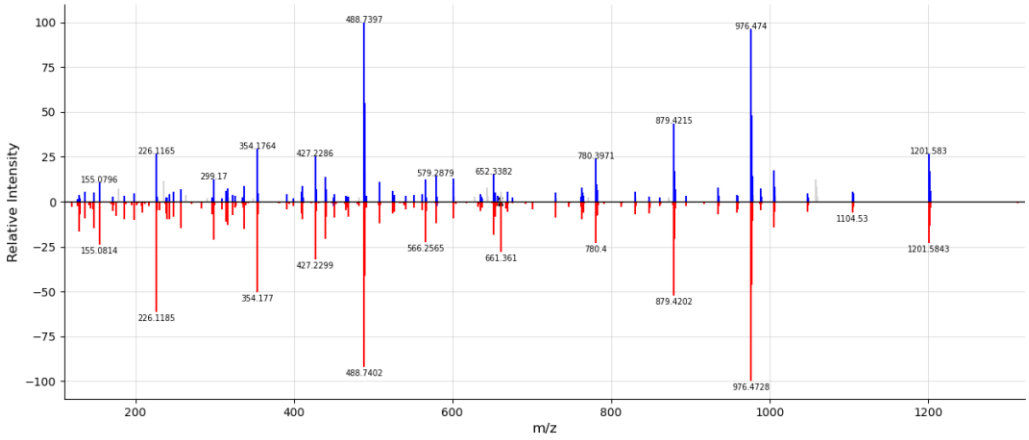


Figure 3: Head-to-tail plot of query spectrum SCAN_6020 to the reference library spectrum corresponding to IGQQPQQPGAPPQQDYTK (z=3) with MF of 845.

Now that we have inspected high quality peptide spectral matches, let's look at an example of a questionable identification. If you click on query "SCAN_9060", the top hit is a hybrid match to IMEEFFR with a MF of 504 and a DeltaMass of 138.0558 Da (Figure 4). The 11th ranking identification is a direct match to sequence LMEEIMSEK with MF of 446. The difference in MF of 58 between the 1st and 11th hits is not large. Moreover, we can see that there are overlapping fragment ions between the reference library spectra. Through this comparison, we can see that the direct reference library assignment is an ambiguous identification due to incomplete fragment ion coverage. You may also have noticed that the assigned peptide sequence is short, only 8 amino acids in length, and is therefore susceptible to homology errors. While the reference mass spectral library contains short sequences, these should be removed in post processing steps.

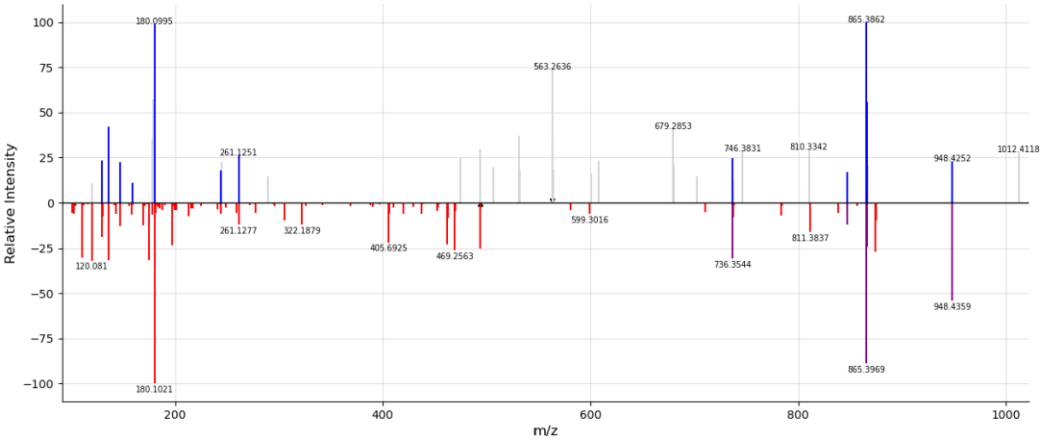


Figure 4: Head-to-tail plot of query spectrum SCAN_9060 to the reference library spectrum corresponding to IMEEFFR (z=2) with MF of 504.

For a small fraction of identifiable tandem mass spectra, the hybrid search may recover peptide identifications with precursor errors. This can include errors in monoisotopic peak assignment, charge state assignment as well as precursor mass that falls outside of the direct or narrow search tolerances. One example can be seen for the query spectrum "SCAN_79953" (Figure 5). Here, the top-ranking hit corresponds to a hybrid seach identification with a DeltaMass of -911.4550 Da. The mass defect for this DeltaMass may cause you to scratch your head, as it's very difficult to arrive at such a DeltaMass with modifications listed in UniMod. Such a DeltaMass often suggests a precursor error. In this example, the precursor, assigned a charge state of 2+, is actually a charge state 3+ precursor.

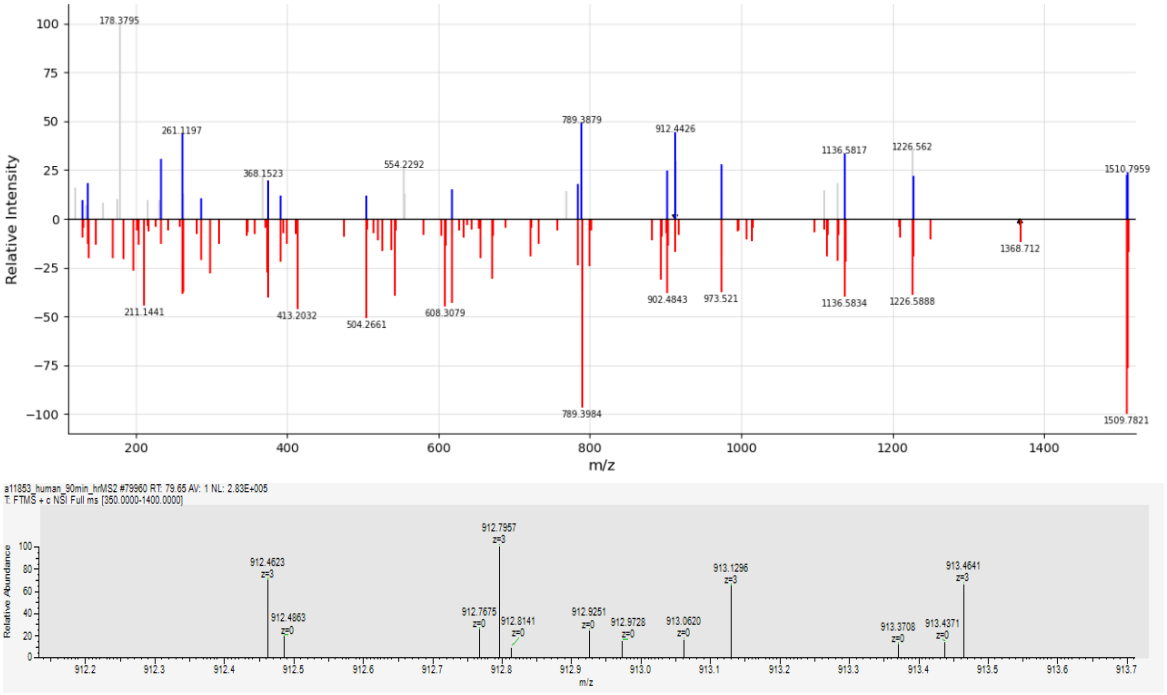


Figure 5: Head-to-tail plot of query spectrum SCAN_79953 to the reference library spectrum corresponding to IPVTDEEQTNVPYIYAIGDILEDK (z=2) with MF of 656 and DeltaMass of -911.4550 Da.

Protein Assembly via PIA analysis with KNIME

In order to inspect the proteins inferred from the peptide identifications, we will perform a PIA analysis with KNIME. If you have already installed KNIME and the PIA nodes from the community contributions repository, open KNIME and follow the steps outlined below.

1. To add the required nodes for analysis, type the name of each node in the Node Repository search field, select the node and drag to your workspace. The nodes we will use are (a) Input File, (b) PeptideIndexer, (c) Port to URL, (d) PIA Compiler and (c) PIA Analysis.
2. Connect the nodes as shown in Figure 6.

3. Before starting the analysis, we will set the appropriate parameters for each node beginning with “Input File”. Right click on the “Input File” node and select “Configure”.
    a. The Input File node, labeled as Node 4 in Figure 6, will be used to configure the idXML input file, an intermediate file format generated by OpenMS. Under the “Options” tab, use “Browse” to select the idXML file generated by the proteoNIST. Next, click “OK”.
    b. The second Input File node, labeled as Node 5 in Figure 6, will be used to configure the FASTA file input. Under the “Options” tab, select the “uniprot-all.fasta” located in the ExampleFiles folder. Next, click “OK”.
    c. Next, we will configure the “PeptideIndexer” node.
        i. First, specify the “decoy_string” by typing “DECOY_” (without quotation marks) in the value field.
        ii. Confirm that “decoy_string_position” is prefix, “missing_decoy_action” and “unmatched_action” are warn, and “keep_unreferenced_proteins” is false.
        iii. Confirm that “write_protein_description” and “IL_equivalent” and are true.
        iv. When true, “write_protein_sequence” computes the sequence coverage; however, this adds considerable time to the analysis. Here, you may select true or false.
        v. Confirm that “aaa_max” is set to a value of 3 and “mismatches_max value” is set to 0.
        vi. To improve the run time, you may increase the number of threads to an integer suitable with your setup.
        vii. Lastly, confirm that the enzyme name is “Trypsin/P” and the “specificity” is set to full.
4. We will use all of the default settings for the PIA Compiler node. This node structures the data for PIA Analysis.
5. Next, we will configure the PIA Analysis node. Here, we will be prompted to execute the upstream nodes before selecting the appropriate parameters.
    a. Under the “general” tab, deselect “Create PSM sets”, as we are only using a single search engine in this demonstration.
    b. Under the “PSMs” tab, change “FileID for PSM output” to 1 as we are only processing a specific file.
    c. Deselect “Calculate FDR for all files”.
    d. Next, we will change the decoy pattern to “DECOY_.*”
    e. Under the “proteins” tab, select the Occam’s razor protein inference method. Both the Occam’s razor and Spectrum Extractor protein inference methods apply the principle of maximal parsimony; however, the Spectrum Extractor also assigns a spectrum to only one peptide. Because we have only included the rank 1 identifications from NIST MSPepSearch, we will use the Occam’s razor method.
    f. In the field “Filters for protein level”, we will select “# unique peptides (Protein)” and require greater than or equal to 2 and select “Add”.

g. Next, select “OK” and execute the node. When finished, select “View: PIA Result Analysis” and let’s take a look!

6. Here, we can sort the protein accessions by the number of peptides and PSMs.
7. When you click on a row in the table, you will see the list of associated peptide identifications as well as a diagram illustrating the protein parsimony results.
    a. Here, we can see the peptide evidence used to assign a single accession to a protein group (Figure 7).
    b. Continue to browse the protein groups to see just how complex the protein assembly can be!

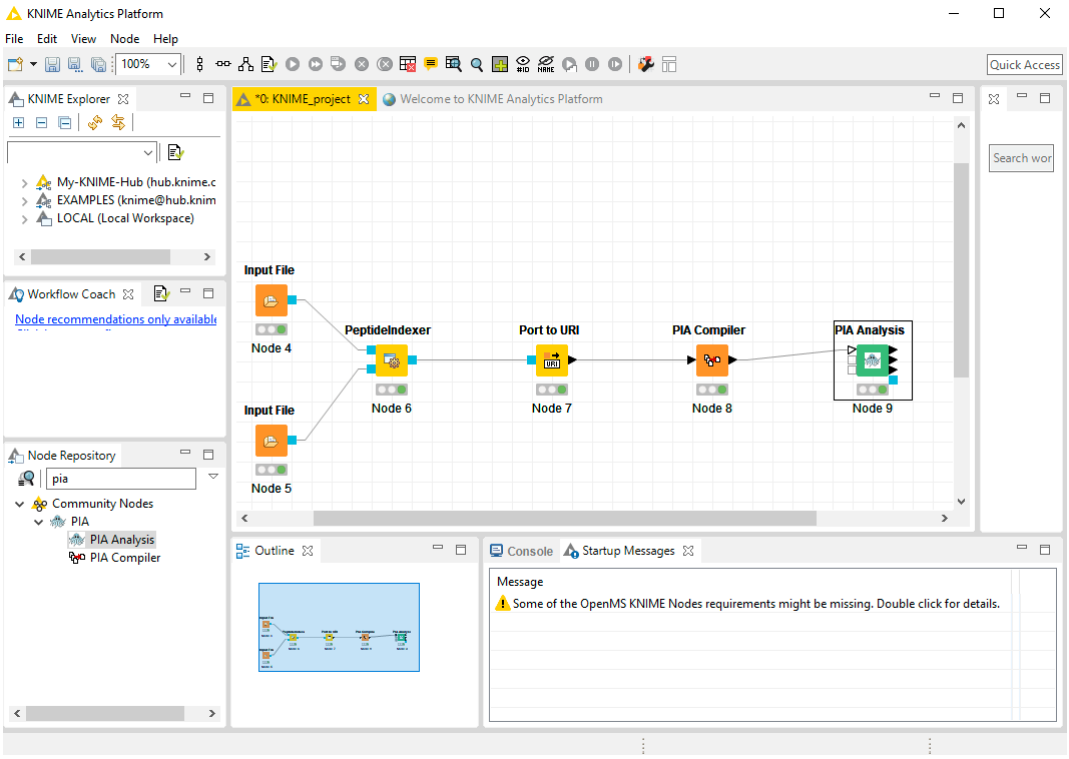


Figure 6: KNIME workspace for PIA analysis.

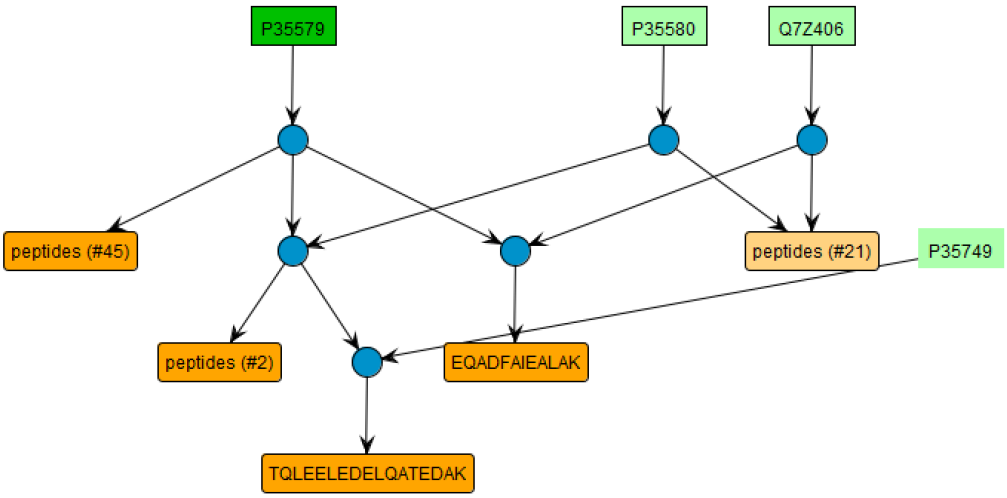


Figure 7: View of PIA Analysis protein inference in which 53 peptide spectral matches, corresponding to 49 distinct peptide sequences, were assigned to the protein accession P35579 (Myosin-9).

Additional Resources

This tutorial is intended to be an introduction to mass spectral library searching and interpreting mass spectral library derived PSMs. For users that are interested in more advanced searching options, the NIST MSPepSearch and MS Search are available for download as well as mass spectral libraries for different species, derivatization, modifications and ionization modes. Detail-oriented users may also wish to learn more about software for MS1 level analysis, NIST-ProMS, which is part of the NCI Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline[9].

Also, for experiments in which multiple raw files are to be processed en masse, and therefore FDR control at the protein-level is required, using MSPepSearch within alternate search engines such as Mascot (Matrix Science, example shown in Figure 7)[10,11] or Proteome Discoverer[12] (Thermo Scientific) may be of use.

---

[9] Rudnick, P. A., Markey, S. P., Roth, J., Mirokhin, Y., Yan, X., Tchekhovskoi, D. V., Edwards, N.J., Thangudu, R.R., Ketchum, K.A., Kinsinger, C.R., Mesri, M. (2016). A description of the clinical proteomic tumor analysis consortium (CPTAC) common data analysis pipeline. *Journal of proteome research*, *15*(3), 1023-1032.

[10] Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS: An International Journal*, *20*(18), 3551-3567.

[11] http://www.matrixscience.com/help/spectral_library.html

[12] https://assets.thermofisher.com/TFS-Assets/CMD/manuals/Man-XCALI-97808-Proteome-Discoverer-User-ManXCALI97808-EN.pdf

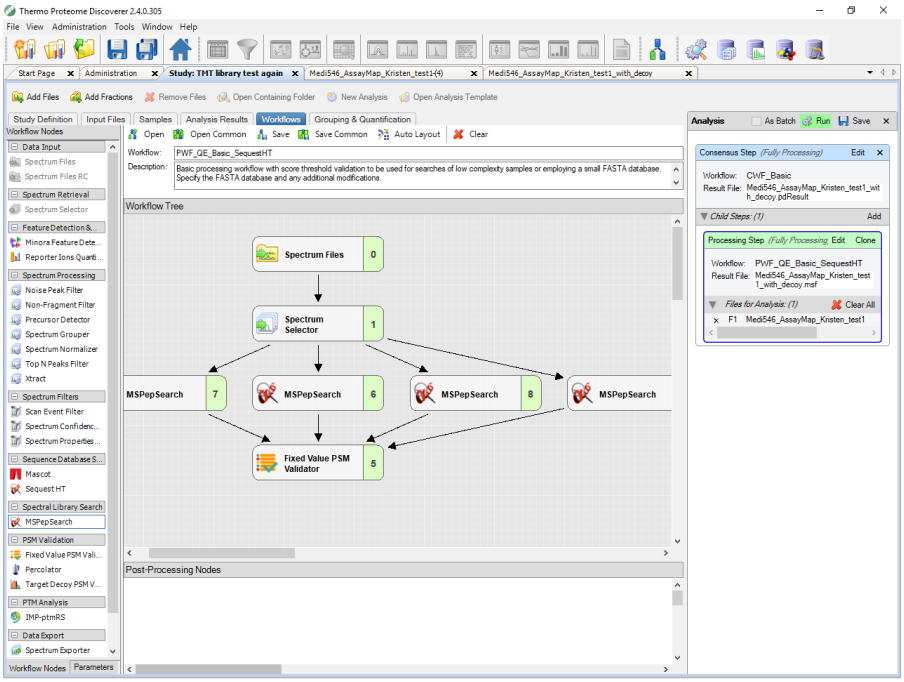


Figure 7: Sample workflow in which NIST MSPepSearch is used within Proteome Discoverer.

If any questions arise after the tutorial, please feel free to contact massspec@nist.gov.

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this tutorial in order to specify the experimental procedure identified adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.