Rapid Proteomics Analysis Algorithm Development Using Proteome Discoverer: A Development Environment for Proteomics Scientists

Iman Mohtashemi; Kai Fritzemeier; Hans Grensemann; Bernard Delanghe; David Horn; Torsten Ueckert

Thermo Fisher Scientific, San Jose, CA

Thermo SCIENTIFIC

Overview

Purpose: Here we present the Thermo Scientific Proteome Discoverer software, a node-based framework for proteomics data processing. We will explain the general structure of the software and provide multiple examples of creation of nodes for use in this system.

Methods: All data were acquired on a hybrid ion trap-Orbitrap^{TM} mass spectrometer system equipped with electron transfer dissociation (ETD). New nodes were developed in C# using Visual Studio 2010 and their utility were demonstrated by incorporating them into Proteome Discoverer^{TM} workflows.

Results: Development, deployment and analysis of various node implementations are discussed.

Introduction

The speed and ultra high resolution acquisition of Orbitrap-based technologies have made the characterization of increasingly larger number of previously undetectable features in a single run challenging. Data analysis bottlenecks are emerging both with respect to the speed of analysis and the depth of converting large data sets into meaningful information. While traditional database searching techniques still remain the mainstream of traditional proteomics analysis, other analysis tools are continuously improving in various areas The breadth and depth of algorithms from academic labs to commercial organizations are daunting yet an infrastructure does not exist to rapidly develop, deploy and share unique algorithms amongst proteomics scientists. A node-based development environment is proposed to test and deploy analysis algorithms without unnecessary overhead

Methods

Deploying New Nodes Into the Proteome Discoverer framework

Proteome Discoverer software is written in the Microsoft C#/.NET environment and runs under Microsoft Windows. It consists of a server component for processing the data with user defined processing workflows and a client component for creating and scheduling processing workflows, and creating and displaying the final reports. The workflows consist of processing nodes that perform in a specific task in the data analysis pipeline. These processing nodes are implemented as plug-ins, making it easy to extend the data processing with new algorithms or functionality.

To create a new node for use in the Proteome Discoverer framework, a developer will start with a Microsoft Visual Studio 2010 solution that contains a template project that performs similar functionality to the node to be added. In the solution, the developer defines the node name, the input and output data types, whether or not the algorithm is tied to the standard license or a separate one, and the algorithm parameters and their acceptable limits. Figure 1 shows the code created for the neutral loss filter example, where the input and output for the node are MS/MS spectra and the appearance of the node is tied to the standard license in Proteome Discoverer software and thus will be available for all users. The code in Figure 1 shows the parameters that are available for users in the Proteome Discoverer Workflow Editor user interface for that node and their acceptable limits. If the user sets any value outside these limits, the workflow will exit with an immediate error.

FIGURE 1. The code below defines the name of the node, the input and output data structures, and the parameters available for the users.

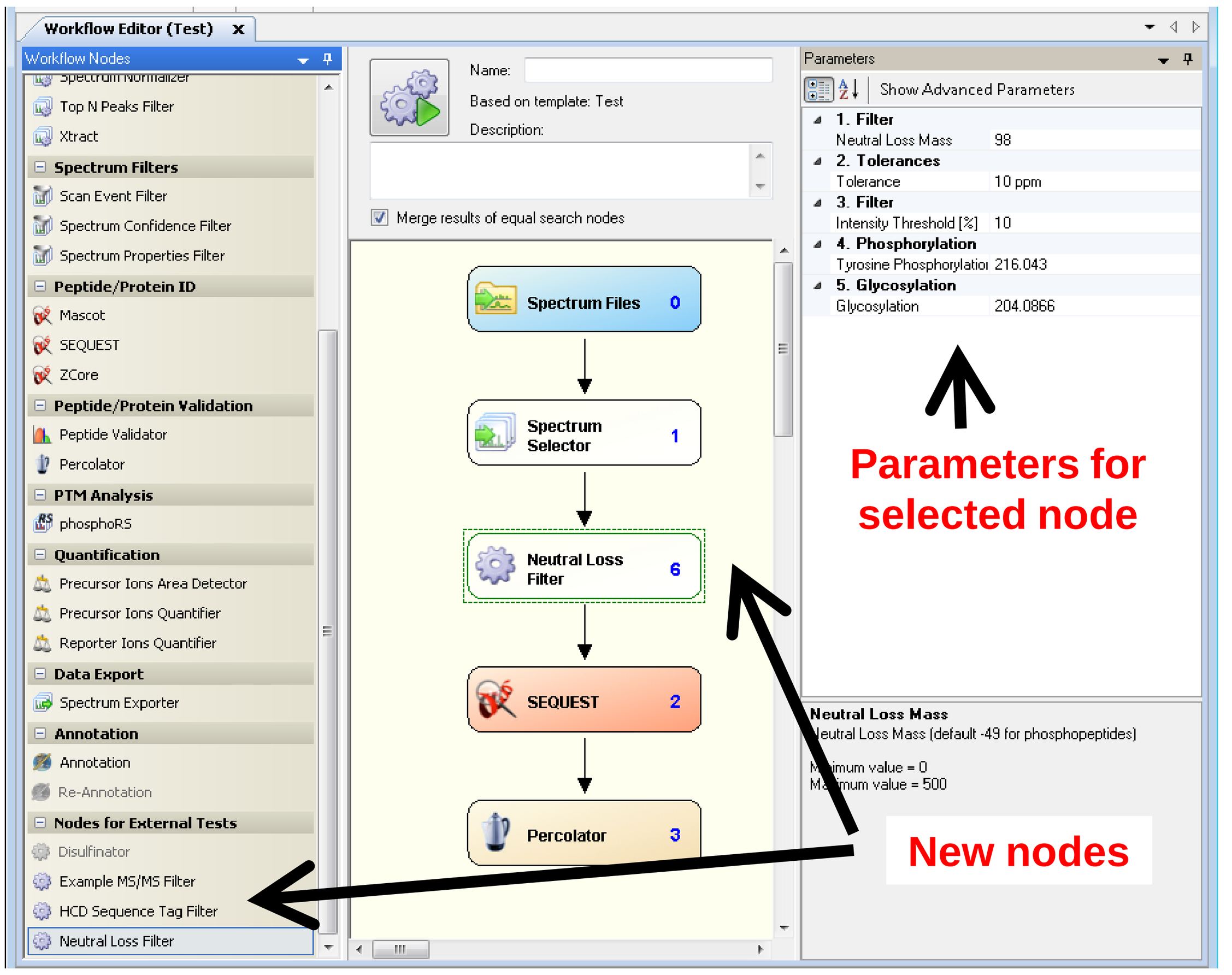
```
[ProcessingNode("F1F0FCBE-2274-41E9-ADD3-15BEC1006E41",
    Category = ProcessingNodeCategories.ExternalTest,
    DisplayName = "Neutral Loss Filter",
    Description = "Filters out non-phosphorylated peptides",
    MainVersion = 0,
    MinorVersion = 4)]
[ProcessingNodeConstraints(UsageConstraint = UsageConstraint.Unristricted)]
[ProcessingNodeConnectionPoint("ISpectrumSink",
    DataTypeSpecialization = StandardDataTypeSpecializations.AcceptAny,
    ConnectionDirection = ConnectionDirection.Incoming,
    ConnectionMultiplicity = ConnectionMultiplicity.Single,
    ConnectionMode = ConnectionMode.Manual,
    ConnectionRequirement = ConnectionRequirement.RequiredAtDesignTime,
    ConnectionDisplayName = ProcessingNodeCategories.SpectrumAndFeatureRetrieval)
[ProcessingNodeConnectionPoint("ISpectrumSource",
    DataTypeSpecialization = StandardDataTypeSpecializations.AcceptAny,
    ConnectionDirection = ConnectionDirection.Outgoing,
    ConnectionMultiplicity = ConnectionMultiplicity.Multiple,
    ConnectionMode = ConnectionMode.Manual,
    ConnectionRequirement = ConnectionRequirement.Optional)]
[LicenseFeature(
    Feature = "Discoverer_Base",
    Description = "Discoverer base license.",
    ShowIfNotAvailable = false)]
```

```
[ParameterInfo(
  Category = "1. Filter",
  DisplayName = "Neutral Loss Mass",
  Description = "Neutral Loss Mass (default -49 for phosphopeptides)",
  DefaultValue = "98",
  MinimumValue = "0",
  MaximumValue = "500",
  Position = 1)]
public DoubleParameter NeutralLoss;

[ParameterInfo(Category = "2. Tolerances",
    DisplayName = "Tolerance",
    Description = "Tolerance.",
    Subset = "Da|mmu|ppm",
    DefaultValue = "10.0 ppm", MinimumValue = "0.0001 Da | 0.01 ppm", Max
    //IntendedPurpose = ParameterPurpose.DynamicModification,
    Position = 1)]
public MassToleranceParameter PeptideTolerance;

[ParameterInfo(
    Category = "3. Filter",
    DisplayName = "Intensity Threshold [%]",
    DefaultValue = "10",
    MinimumValue = "0",
    MaximumValue = "100",
    Position = 1)]
```

Figure 2. Proteome Discoverer Workflow Editor with the new nodes

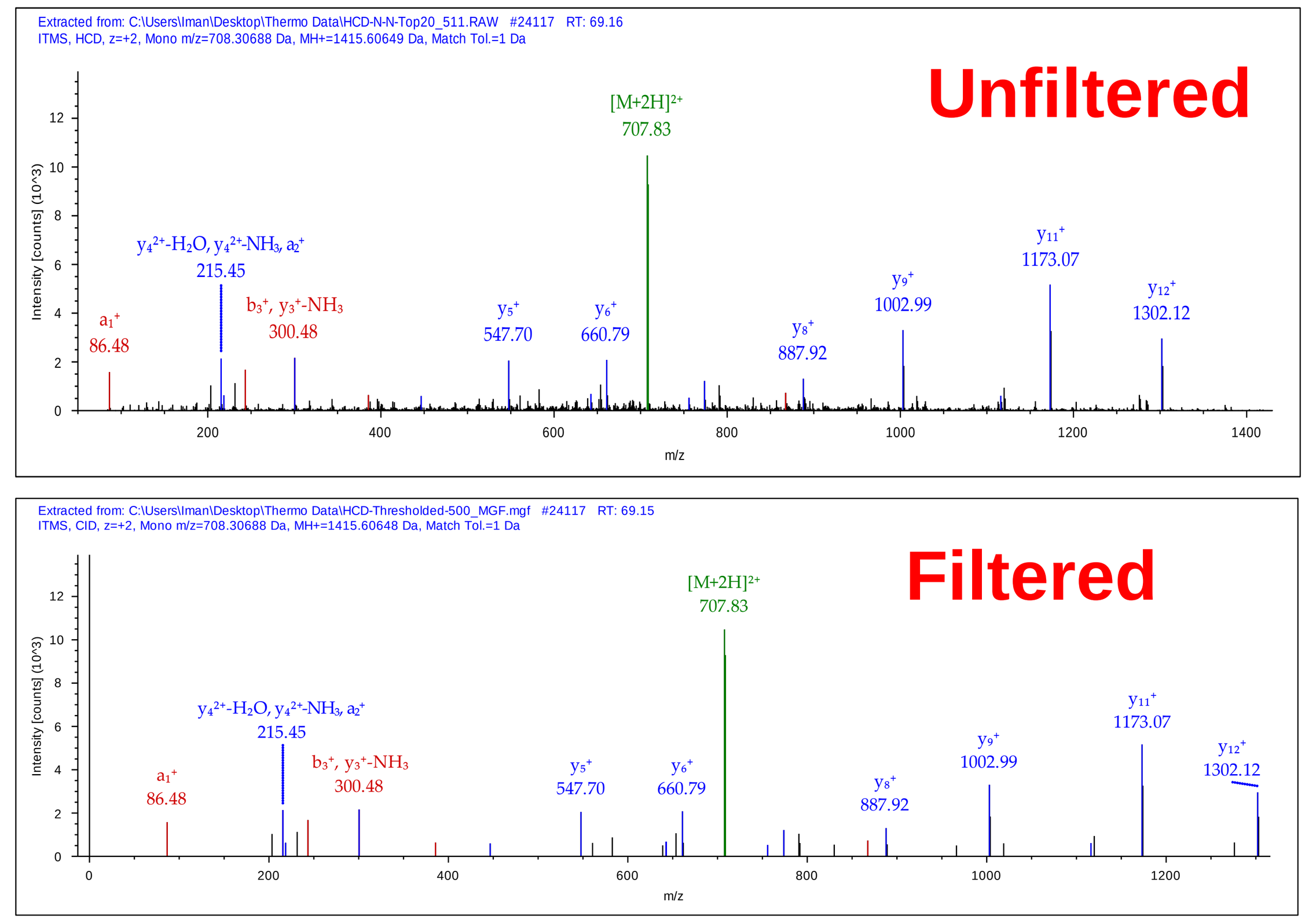


Results

Normalized ms^2 filter node

Each ms/ms spectra was normalized to the most intense peak within the spectra. A threshold was then applied to achieve maximum identification rate with minimal search input(fig 3). Display of annotated spectra is seamlessly integrated without affecting any other components in Discoverer

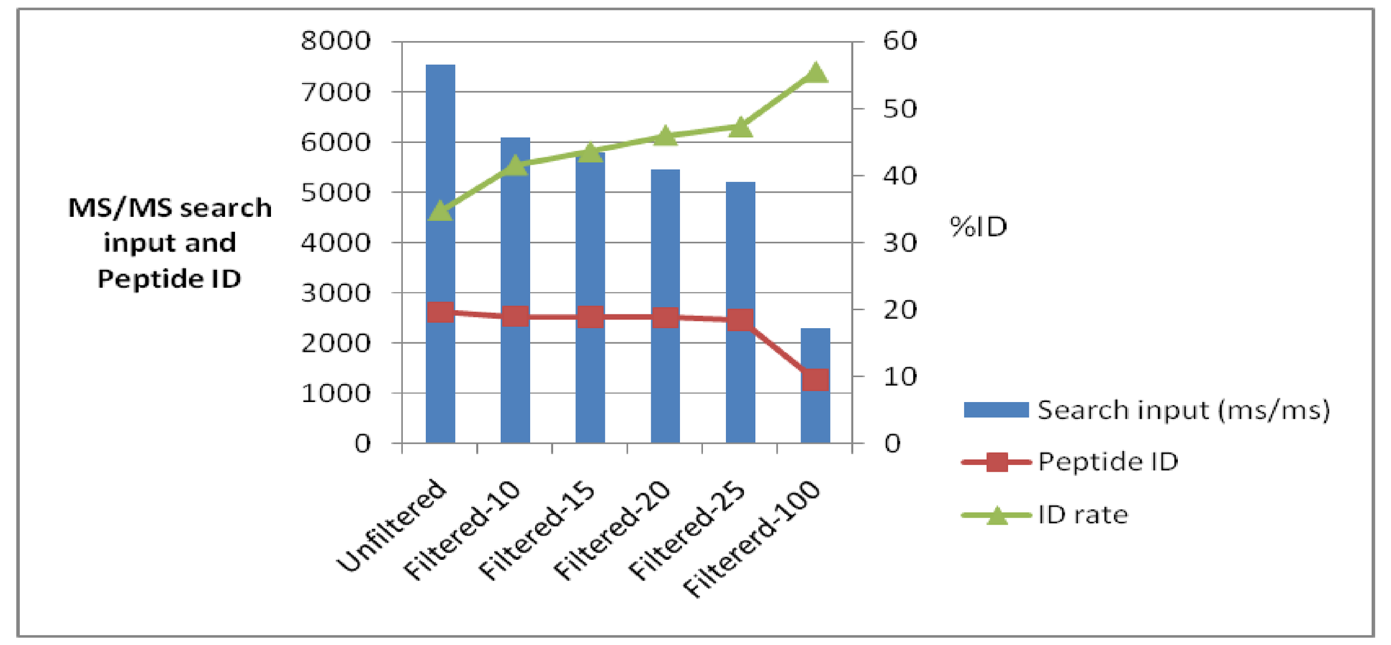
Figure 3. Filtered ms2 spectrum

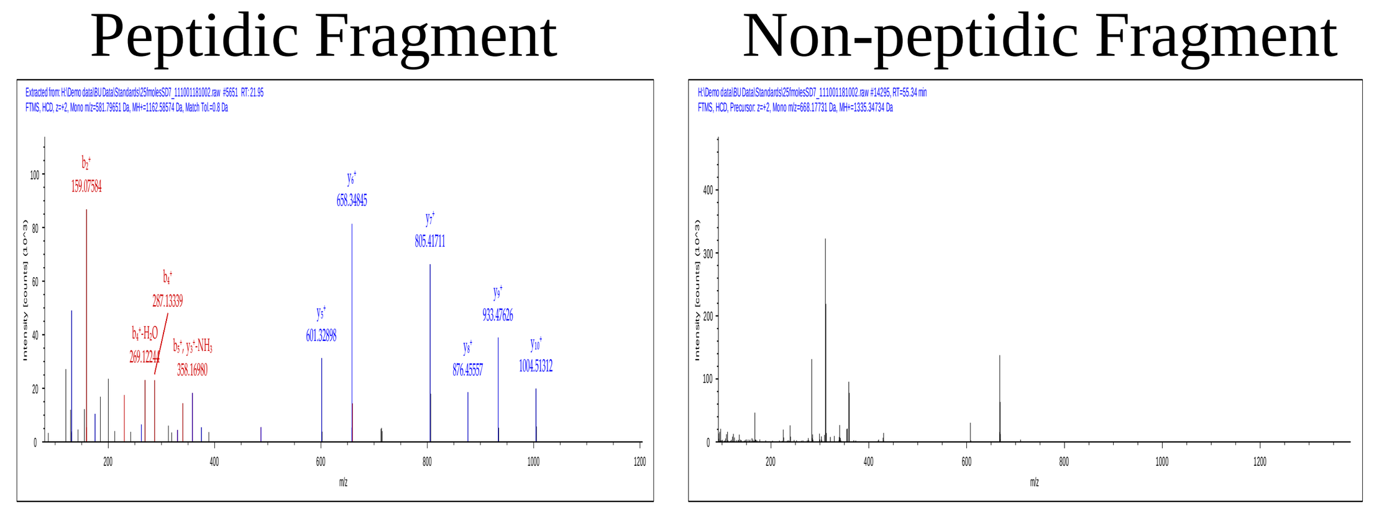


HCD Sequence tag filter

A peptide sequence tag filter was applied to HCD ms2 data to minimize search input for non-peptidic fragments (fig4)

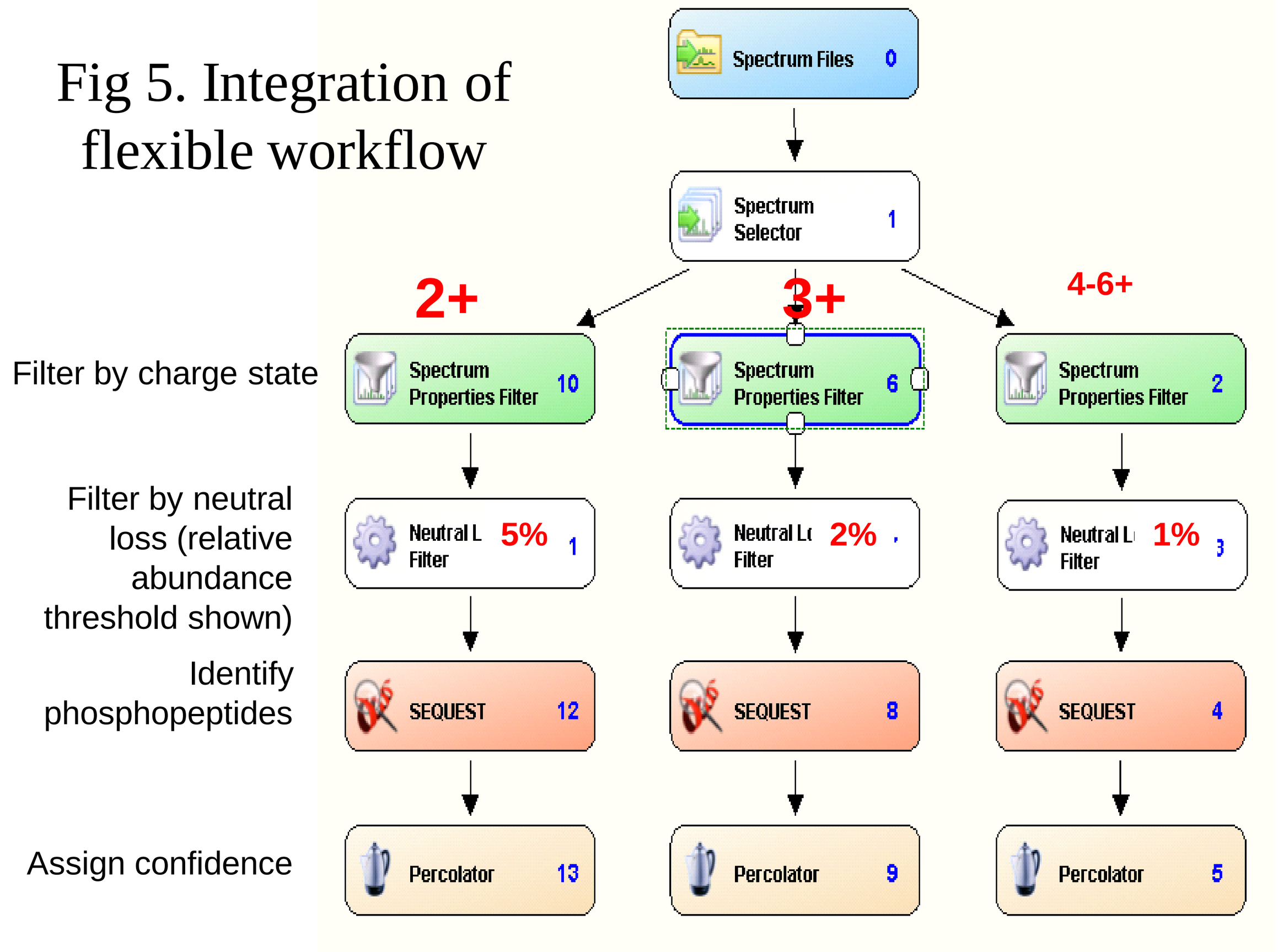
Figure 4. HCD sequence tag filter ID vs input





CID ms^2 neutral loss

CID ms2 screened for phosphrylation via neutral loss monitoring (fig5)



ETD triggered ms^3 for Disulfide Mapping

Custom data fields for data display for disulfide mapping (fig 6).Each new field is assigned a new GUID

Figure 6. Custom data fields for disulfinator node

```
CustomDataField crosslinkedPeptides = ProcessingServices.CustomDataService.GetOrCreateCustomDataField(workflow.
ID, new Guid("6C382FDA-C6AA-4CDB-A390-AB4D9B0011C8"), "Crosslinked Peptide(s)", ProcessingNodeNumber,
searchNodeInfo.ProcessingNodeNumber, CustomDataTarget.Peptide, CustomDataType.String, CustomDataAccessMode.Read,
 DataVisibility.Visible);
CustomDataField massCrosslinkedPeptide = ProcessingServices.CustomDataService.GetOrCreateCustomDataField(
workflow.ID, new Guid("4866912F-E945-4AD1-872C-88D934402AEF"), "Mass Crosslinked Peptide", ProcessingNodeNumber,
 searchNodeInfo.ProcessingNodeNumber, CustomDataTarget.Peptide, CustomDataType.Double, CustomDataAccessMode.Read
, DataVisibility.Visible, plotType: PlotType.Numeric, format: "0.000");
CustomDataField calculatedMassCrosslinkedPeptide = ProcessingServices.CustomDataService.
GetOrCreateCustomDataField(workflow.ID, new Guid("87C0A83E-7CBD-4554-A995-C77E3CDF8986"), "Calculated Mass
Crosslinked Peptide", ProcessingNodeNumber, searchNodeInfo.ProcessingNodeNumber, CustomDataTarget.Peptide,
CustomDataType.Double, CustomDataAccessMode.Read, DataVisibility.Visible, plotType: PlotType.Numeric, format:
"0.000");
CustomDataField deltaMassCrosslinkedPeptide = ProcessingServices.CustomDataService.GetOrCreateCustomDataField(
workflow.ID, new Guid("D80C2033-7C3D-4B66-B4CD-CE3227E27897"), "Delta Mass Crosslinked Peptide [PPM]",
ProcessingNodeNumber, searchNodeInfo.ProcessingNodeNumber, CustomDataTarget.Peptide, CustomDataType.Double,
CustomDataAccessMode.Read, DataVisibility.Visible, plotType: PlotType.Numeric, format: "0.000");
```

Figure 7. Custom data fields for disulfinator

Sequence	m/z [Da]	MH+ [Da]	ΔM [ppm]	Crossl	XCorr	Crosslinked Peptide(s)	Mass Cross	Calculated	Delta Mas	Peptide 1
MPCTEDYLSLILNR	834.41071	1667.81413	0.55	337	2.59	MPCTEDYLSLILNR---EACFAVEGPK	2715.292	2715.283	-3.602	EACFAVEGPK
LFTFHADICTLPDTEK	925.95300	1850.89873	-0.36	335	0.65	MPCTEDYLSLILNR---LFTFHADICTLPDTEK	3515.702	3515.690	-3.565	MPCTEDYLSLILNR
LFTFHADICTLPDTEK	925.95258	1850.89787	-0.82	333	1.88	MPCTEDYLSLILNR---LFTFHADICTLPDTEK	3515.702	3515.690	-3.565	MPCTEDYLSLILNR
MPCTEDYLSLILNR	834.40985	1667.81243	-0.47	333	2.49	MPCTEDYLSLILNR---LFTFHADICTLPDTEK	3515.702	3515.690	-3.565	LFTFHADICTLPD...
MPCTEDYLSLILNR	834.41162	1667.81597	1.65	329	1.16	MPCTEDYLSLILNR---RPCFSALTPDETYVPK	3488.705	3488.690	-4.346	RPCFSALTPDETY...
MPCTEDYLSLILNR	834.41138	1667.81548	1.36	327	0.91	MPCTEDYLSLILNR---RPCFSALTPDETYVPK	3488.703	3488.690	-3.647	RPCFSALTPDETY...
RPCFSALTPDETYVPK	912.45618	1823.90508	2.94	327	0.89	MPCTEDYLSLILNR---RPCFSALTPDETYVPK	3488.703	3488.690	-3.647	MPCTEDYLSLILNR
MPCTEDYLSLILNR	834.41022	1667.81316	-0.03	323	3.27	MPCTEDYLSLILNR---LCVLHEK	2506.256	2506.250	-2.412	LCVLHEK
MPCTEDYLSLILNR	834.41205	1667.81682	2.16	315	2.19	QNCDQFEK---MPCTEDYLSLILNR	2676.219	2676.210	-3.431	QNCDQFEK
MPCTEDYLSLILNR	834.41046	1667.81365	0.26	311	2.17	QNCDQFEK---MPCTEDYLSLILNR	2676.217	2676.210	-2.427	QNCDQFEK
DAIPENLPPLTADFAEDKDVCK	1201.09058	2401.17388	6.07	305	0.81	DAIPENLPPLTADFAEDKDVCK---LFTFHADICTLPDTEK	4249.044	4249.036	-1.932	LFTFHADICTLPD...
DAIPENLPPLTADFAEDKDVCK	1201.09058	2401.17388	6.07	301	0.62	DAIPENLPPLTADFAEDKDVCK---LCVLHEK	3239.602	3239.596	-1.940	LCVLHEK
DAIPENLPPLTADFAEDKDVCK	1201.08899	2401.17070	4.75	299	0.69	DAIPENLPPLTADFAEDKDVCK---QNCDQFEK	3409.563	3409.556	-2.029	QNCDQFEK
EYEATLEECCAKDDPHACYSTVFDK	1434.08496	2867.16265	-7.65	283	0.42	DAIPENLPPLTADFAEDKDVCK---EYEATLEECCAKDDPHACYSTVFDK	5263.313	5263.305	-1.485	DAIPENLPPLTAD...
DAIPENLPPLTADFAEDKDVCK	1201.07312	2401.13896	-8.47	281	0.90	DAIPENLPPLTADFAEDKDVCK---EYEATLEECCAKDDPHACYSTVFDK	5263.313	5263.305	-1.485	EYEATLEECCAK...
SHCIAEVEK	508.24765	1015.48802	0.27	267	1.44	SHCIAEVEK---EACFAVEGPK	2062.959	2062.957	-1.053	EACFAVEGPK
EACFAVEGPK	525.75061	1050.49394	1.35	267	1.44	SHCIAEVEK---EACFAVEGPK	2062.959	2062.957	-1.053	SHCIAEVEK
RPCFSALTPDETYVPK	912.45319	1823.89910	-0.34	263	1.21	SHCIAEVEK---RPCFSALTPDETYVPK	2836.370	2836.364	-1.894	SHCIAEVEK
SHCIAEVEK	508.24637	1015.48546	-2.26	261	1.13	SHCIAEVEK---LCVLHEK	1853.928	1853.925	-1.582	LCVLHEK
LCVLHEK	841.46179	841.46179	2.04	259	1.38	SHCIAEVEK---LCVLHEK	1853.928	1853.925	-1.582	SHCIAEVEK
SHCIAEVEK	508.24704	1015.48680	-0.93	259	0.68	SHCIAEVEK---LCVLHEK	1853.928	1853.925	-1.582	LCVLHEK
MPCTEDYLSLILNR	834.41034	1667.81340	0.11	255	2.53	SHCIAEVEK---MPCTEDYLSLILNR	2680.284	2680.278	-2.433	SHCIAEVEK
MPCTEDYLSLILNR	834.41144	1667.81560	1.43	251	3.22	SHCIAEVEK---MPCTEDYLSLILNR	2680.287	2680.278	-3.527	SHCIAEVEK
QNCDQFEK	506.21542	1011.42357	3.47	247	0.71	SHCIAEVEK---QNCDQFEK	2023.886	2023.885	-0.766	SHCIAEVEK

Custom date Fields

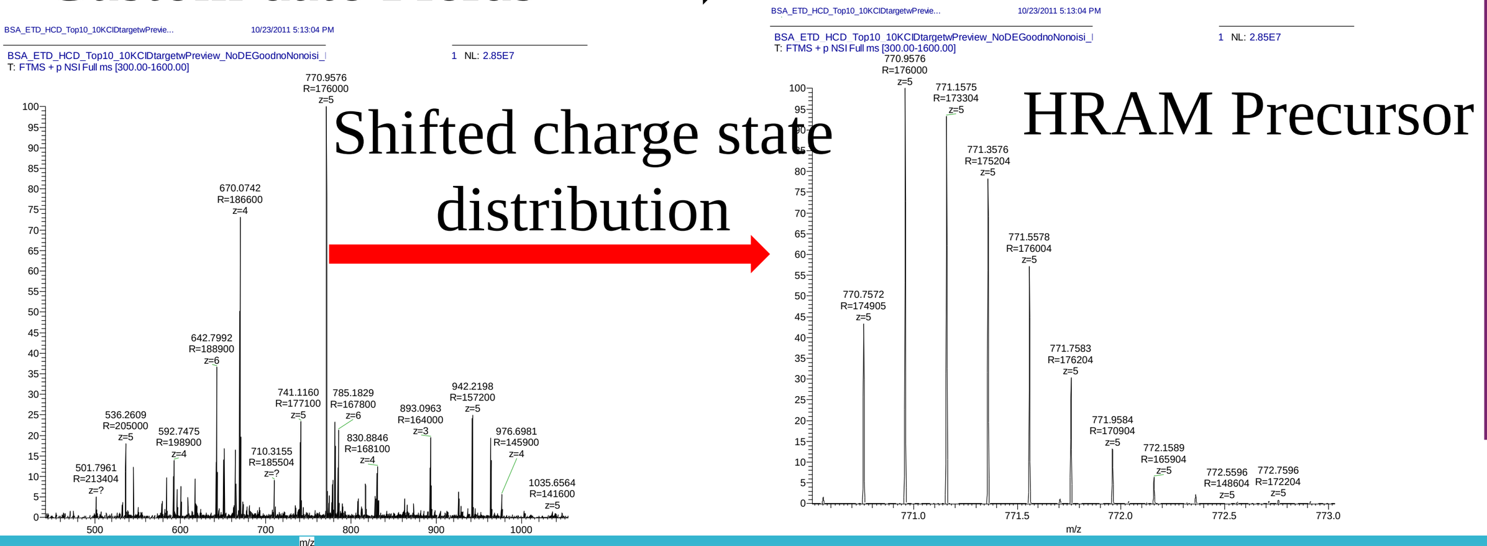
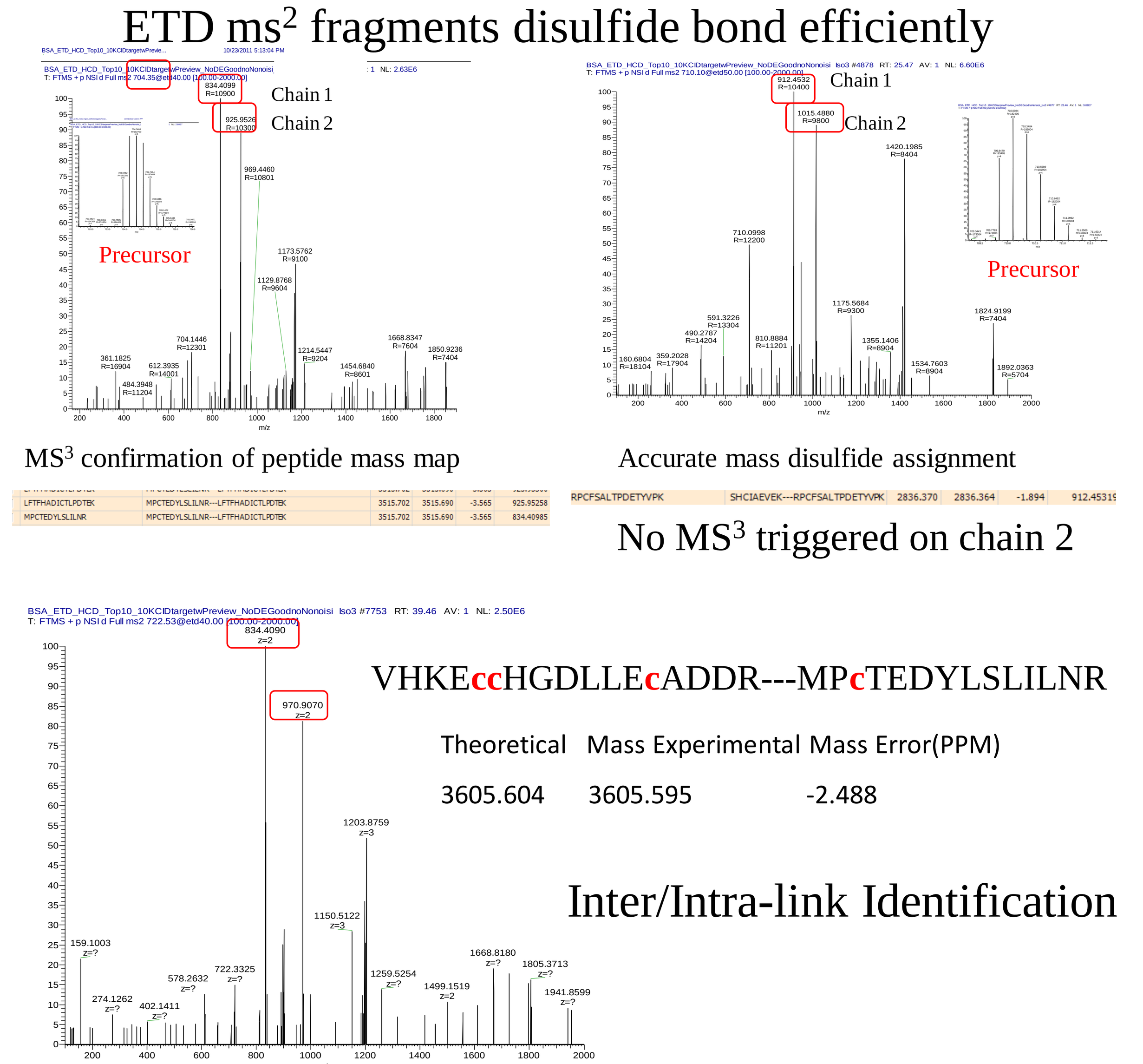


FIGURE 8. Disulfide Map.



Conclusion

- A node development environment is implemented in Proteome Discoverer that allows rapid deployment of custom algorithms without the need to develop graphical user interfaces and other input/output infrastructure. This allows for algorithms to be deployed and shared among the mass spectrometry community
- Multiple nodes were implemented in PD 1.3 and easily integrated into the workflows. For example, the neutral loss filter (fig 5) was integrated with multiple node connections I/O to search under variable parameter settings without affecting any other part of Discoverer
- The external nodes developed for this poster can be used as templates for other development.
- Non-traditional proteomics workflows such as disulfide mapping presented here can also be easily integrated and the data can be displayed via the custom data fields properties of Proteome Discoverer
- Nodes, source code and templates will be available at www.PD-Nodes.org

References

1. Wu et al. 2010. Identification of the Unpaired Cysteine Status and Complete Mapping the 17 Disulfides of Recombinant Tissue Plasminogen Activator Using LC-MS with ETD/CID