JLab Proteogenomics KNIME Tutorial

jLabProteogenomicsKNIME package contains nodes of exact string matching algorithms WuManber and SuffixTree. These nodes search peptide sequences through a nucleotide or amino acid sequence database. In addition to them EnzymeChecker node is available to filter mapped peptides according to enzymatic cleavage rule.

These nodes are available at: http://bioinformatics.iyte.edu.tr/Proteogenomics We can install the jLabProteogenomics KNIME nodes from `Help > Install New Software`

Here we enter the above link to `work with` field, then press `add`. After restarting KNIME, nodes will appear under `Community Nodes/Proteogenomics`

The nodes are dependent on KNIME File Handling plugin. Therefore, before installing the proteogenomics nodes, first KNIME File Handling plugin must be installed.

Installation of KNIME File Handling nodes can be done in same way via http://www.knime.org/update/2.8/

The needed KNIME File Handling plugin must be equal and greater than version 2.9.2.

How to upload sequence files:

Under proteogenomics directory, Input File node is used to upload sequence database files and peptide query file in .fasta or .fa format. For WuManber node, peptide query file can be in .mzid format.

After execution of Input File node, first 50 lines of the file can be viewed upon right-click on the node.

How to execute WuManber or Suffix Tree:

Input File nodes are connected to WuManber or Suffix Tree node. First in-port must be sequence file and second in-port must be sequence file. Right-click on node will show list of options and by clicking configure, output file directory can be selected. Execute command will generate output in .gff or .gff3 format.

How to execute Enzyme Checker:

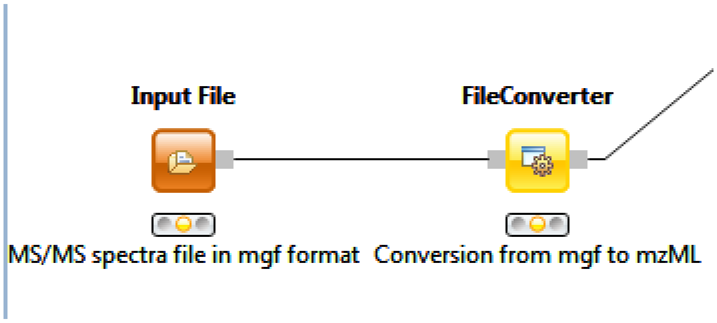
Enzyme Checker takes one sequence file and one gff/gff3 file as input. First in-port is sequence file which can be the same file used as sequence database in WuManber or Suffix Tree node execution. Second in-port is either can be out-port of WuManber or Suffix Tree nodes or user can upload a gff/gff3 file. Node configuration can be done via right-click on Enzyme Checker node. Defaults are for Trypsin enzyme.

How to use Proteogenomics nodes with OpenMS nodes available at KNIME Analytics Full version:

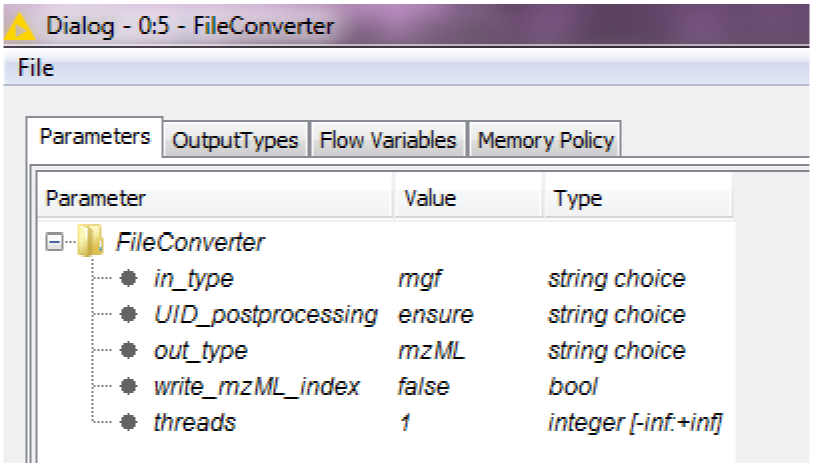
Proteogenomics nodes can be used with OpenMS Mass spectrometry analysis platform. Peptide query file can be generated by peptide identification tools such as OMSSA, XTandem, MSGFDB. Input, output file and settings requirements for OpenMS nodes are given in the node documentation of OpenMS.

Here, as an example OMSSA algorithm is used with Proteogenomics Nodes WuManber and EnzymeChecker. The workflow is available under bioinformatics.iyte.edu.tr/lelantos and the data used in this workflow can be downloaded from same url as Data.zip. test.mgf and testDB-Rebuttal.fasta are the input spectra and database files.

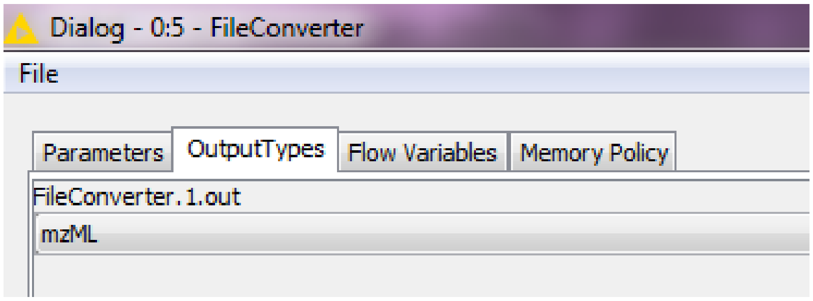
1- File type conversion:
   OpenMS Omssa adapter takes .mzML format as spectra input format. Therefore, .MGF format is converted to .mzML by using FileConverter under FileHandling category.



FileConverter node is configured via right-click -> configure. Input type is specified as mgf and output type is specified mzML. In addition to that output file type, .mzML, must be set under OutputTypes tab.



Parameter	Value	Type
FileConverter		
in_type	mgf	string choice
UID_postprocessing	ensure	string choice
out_type	mzML	string choice
write_mzML_index	false	bool
threads	1	integer [-inf:+inf]

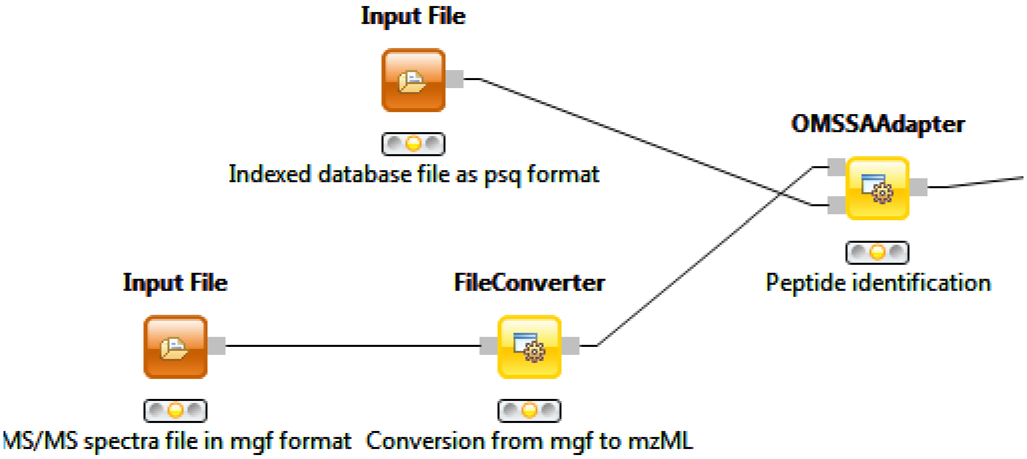


FileConverter.1.out

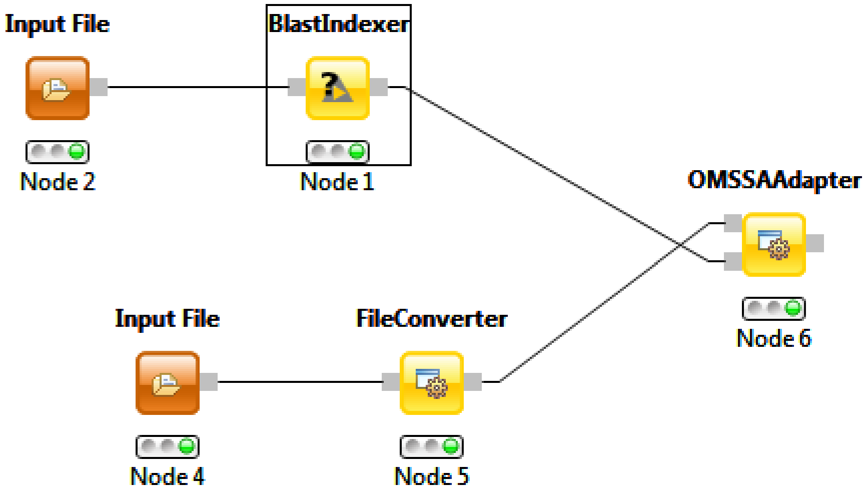
mzML

2- Identification:

OMSSAAdapter node is under Identification category of OpenMS. It takes input spectra as .mzML file to first port and input indexed database file with .psq extension to second port. The indexing of database file in .FASTA format must be done by using makeblastdb or formatdb tools of NCBI.


Input File
Indexed database file as psq format
OMSSAAdapter
Peptide identification
Input File
FileConverter
MS/MS spectra file in mgf format
Conversion from mgf to mzML

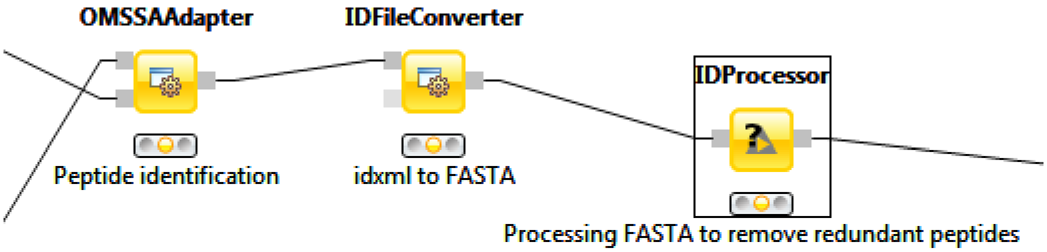
OpenMS does not provide NCBI indexed FASTA file databases. Therefore, if you do not have pre-indexed NCBI formatted database file, you can install SequenceAnalysis nodes available at http://bioinformatics.iyte.edu.tr/SequenceAnalysis as described above. BlastIndexer node will index the input FASTA database file by using makeblastdb executable.



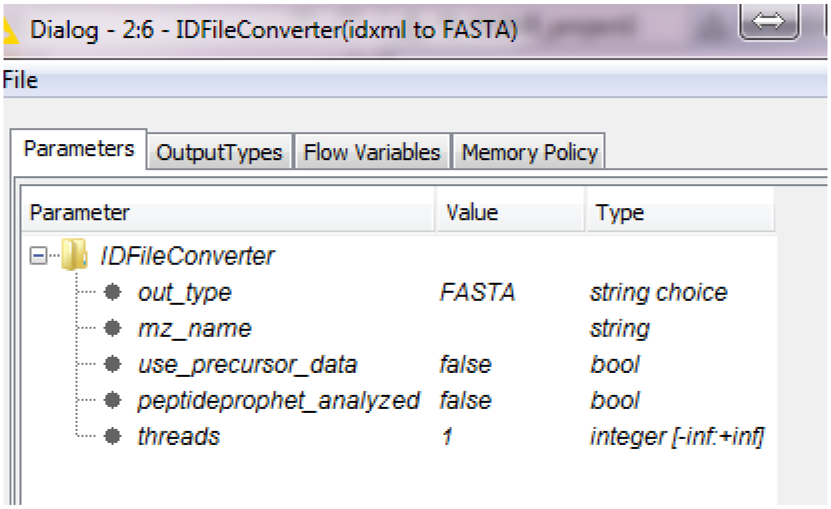
Configuration of BlastIndexer node can be done by right-click->configure.

3- Peptide query file generation

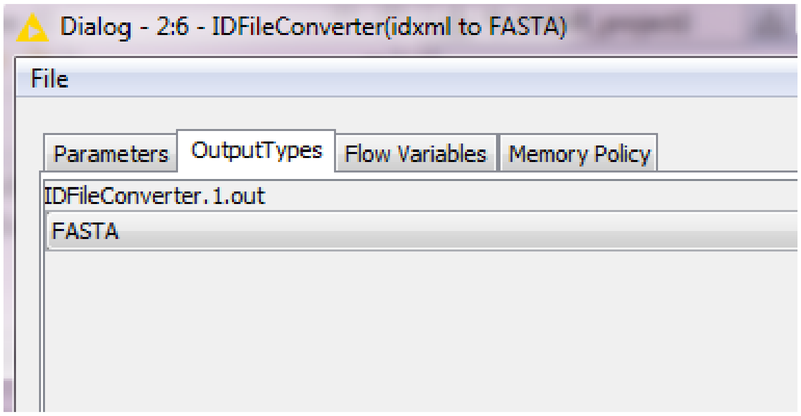
OMSSAAdapter returns idXML file output which cannot be parsed by WuManber or SuffixTree nodes. These nodes only accept FASTA format for peptide queries. Therefore, IDFileConverter node of OpenMS must be used. However, query FASTA file created by IDFileConverter has redundant peptide sequences. In order to remove redundant sequences from query file, IDProcessor node under Proteogenomics category should be used.



Node configurations for IDFileConverter and IDProcessor are given below.

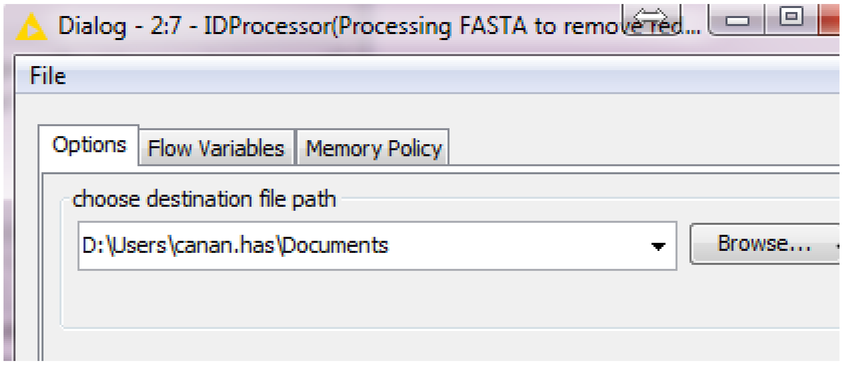


Parameter	Value	Type
IDFileConverter		
out_type	FASTA	string choice
mz_name		string
use_precursor_data	false	bool
peptideprophet_analyzed	false	bool
threads	1	integer [-inf:+inf]



IDFileConverter.1.out

FASTA


Dialog - 2:7 - IDProcessor(Processing FASTA to remove red...
File
Options
Flow Variables
Memory Policy
Browse...

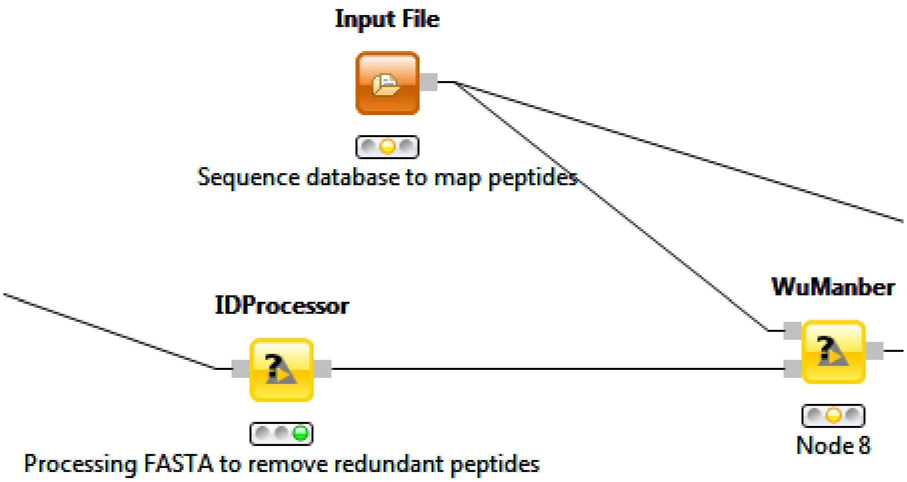
choose destination file path

D:\Users\canan.has\Documents

4- Peptide mapping

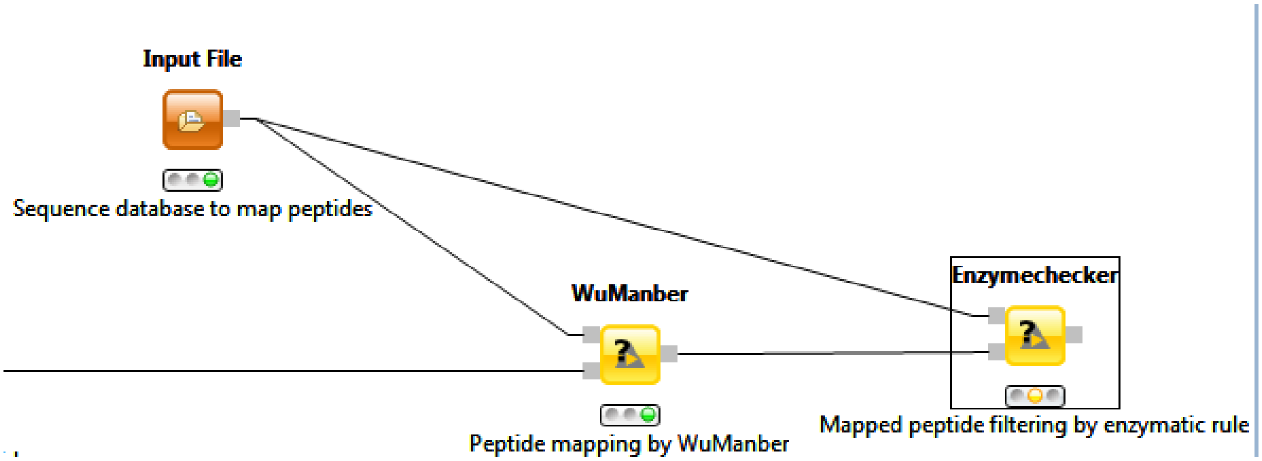
Database search algorithms do not return all peptide locations in database entries, for instance proteins. Therefore, WuManber or SuffixTree node can be used to map identified peptides to their locations.

Detailed information on WuManber and SuffixTree node configurations are given in above. In this example, output of IDProcessor is given as input peptide query.

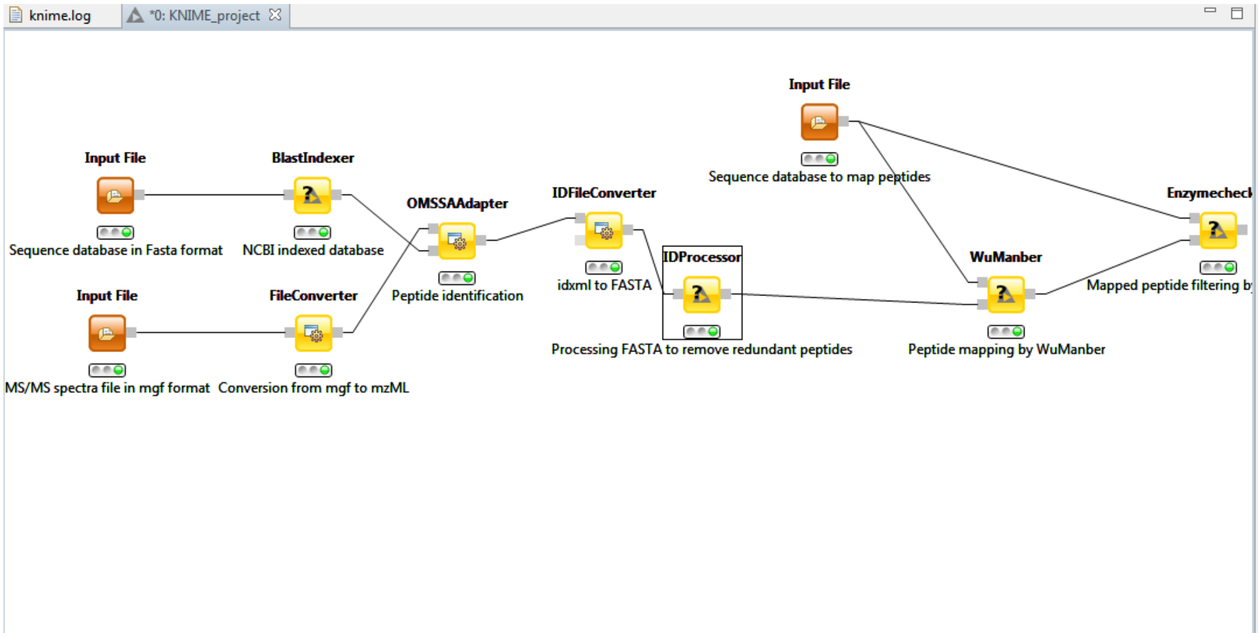

Input File
Sequence database to map peptides
IDProcessor
Processing FASTA to remove redundant peptides
WuManber
Node 8

5- Location filtering

WuManber and SuffixTree map peptides to database entries without considering enzymatic digestion rules. Therefore, EnzymeChecker node as explained above is needed to be used to filter mapping results which are not compatible to cleavage rules.


Input File
Sequence database to map peptides
WuManber
Peptide mapping by WuManber
Enzymechecker
Mapped peptide filtering by enzymatic rule

When all nodes are executed, the triple buttons under nodes will return to green. Please check configuration of each node by right-click->configure to ensure correct location settings.


knime.log
*0: KNIME_project
Input File
Sequence database in Fasta format
BlastIndexer
NCBI indexed database
Input File
MS/MS spectra file in mgf format
FileConverter
Conversion from mgf to mzML
OMSSAAdapter
Peptide identification
IDFileConverter
idxml to FASTA
IDProcessor
Processing FASTA to remove redundant peptides
Input File
Sequence database to map peptides
WuManber
Peptide mapping by WuManber
Enzymecheck
Mapped peptide filtering b