Quick Start to data analysis using EuGenoSuite

EuGenoSuite is an automated pipeline for proteogenomic analysis from mass spectrometry proteomics data using open source peptide identification algorithms. By applying proteomic search by multiple algorithms, EuGenoSuite provides better coverage of proteome at fixed FDR. Proteomic analysis features relevant to eukaryotic gene models like protein inference (important to analyse transcriptomic data where isoforms will have shared peptides) are added on top of the peptide searches and result integration. Features like protein level FDR calculation and result filtering is key feature of this application.

This command line application can be used as an independent proteomic data search tool for peptide and protein identification using OMSSA and X!Tandem algorithms. It can also be integrated to Eukaryotic proteogenomic pipelines to enable peptide identifications from extra-large genomic and transcriptomic databases.

Executables will be provided for windows and linux platforms and requires OMSSA and X!Tandem to be already installed.

A README file provides information about setup and usage.

Download and Setup EuGenoSuite

To use EuGenoSuite you need OMSSA and X!Tandem to be installed, A protein fasta to search against, MGF spectra files in a directory, a config file to define path for input files and a parameter file to define parameters for the search.

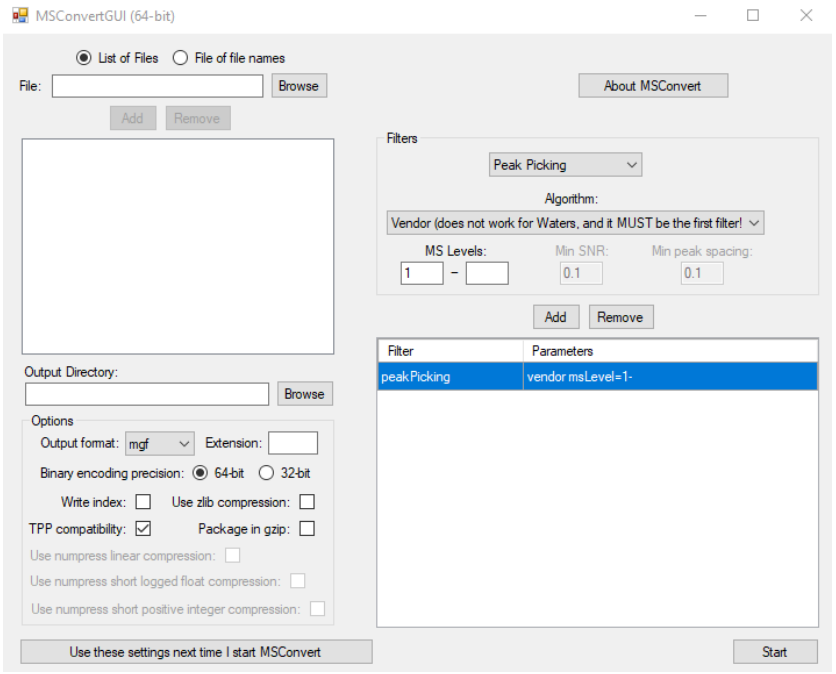
No separate installation step is required for the EuGenoSuite itself. Simple unarchiving the .zip should enable the executables to be used. In linux you might need to provide permissions and probably dos2unix of fasta and MGF files.

For Quick Setup, two zip folders for Windows and Linux containing search tools are made available with this document. Just unarchive the folder.

Convert raw data to the mgf format

The proprietary raw data needs to be converted to open data format. For this, MSconvert tool can be used which is a part of ProteoWizard tool and can be downloaded from this link: http://proteowizard.sourceforge.net/index.html

EuGenoSuite can read mgf file format. The tutorial for the tool is available on the website. The parameters required to select is shown in the screenshot.



Setting up search database

A search database is required to setup for searching MS/MS datasets. A protein sequence database can be used for this or alternatively, a three frame translation of transcriptome database can be used. The database file can be in FASTA format containing protein sequences. This file is need to be provided in DATABASE directory provided in EuGenoSuite directory.

All 115 common contaminants defined in cRAP fasta file were included in target database present in DATABASE directory.

Setting up configuration file

A config file is required to setup. A sample Test.conf is provided in the tool directory.

```
#path for file containing protein sequences in FASTA format.
-g=./DATABASE/Protein.fasta;
#OMSSA directory path. Program tries to find omssacl executable in this
directory.
-O=./OMSSA;
#TANDEM directory path. Program tries to find /bin/Tandem.exe in this
directory.
-T=./Tandem;
#Directory path containing mgf files
-s=./INPUT;
#parameter file(Please see Test_params.txt)
-r=params.txt;
#formatdb.exe directory path. program tries to find formatdb executable in
this directory.
-f=./DATABASE/;
#Decoy type. 1 for Separate decoy. 0 for concatenated decoy.
-d=1;
#A fasta file containing probable contaminants.
-c=./DATABASE/crap.fasta
```

Setting up search parameters

A parameters file is required to setup. A sample Test_params.txt is provided in tool directory.

```
Enzyme=0;  #Protease enzyme ID.0 for trypsin. For enzyme ids please refer ./CONFIG/enzyme_list.txt

MC=1;      #Missed Clevages

PreTol=6;  #Precursor matching tolerance

PreUnit=ppm;     #Unit for precursor tolerance. Available options: ppm or Da.

ProTol=0.8;      #Product ion matching tolerance.

ProUnit=Da;      #Unit for product ion tolerance.Available options:ppm or Da.Omssa supports only Da for product tolerance.

FixedMod=3;      #Fixed modification ID/IDs. Please refer ./CONFIG/mod_list.txt for modification ids.

VarMod=1;  #variable modification ID/IDs. Please refer ./CONFIG/mod_list.txt for modification ids.

SpectraFile=;    #Spectra file path. Leave blank as file names will be read from the dierctory path provided automatically.

Database=; #Database file. Leave blank as database is created by translating the genome file.

Decoy_Database=;#Decoy database file. Leave blank as decoy database is created by reversing the translated genome file.

FDR_Value=1;     #FDR cutoff value. Separate FDR is calculated.

FDR_Type=1;      #FDR Type. 1 for Separate target decoy Search. 0 for Concatenated target decoy search. Default 1.
```

Running EuGenoSuite

After complete setup, the EuGenoSuite can be run through command line. To run it, open the tool directory in cmd or terminal and type

`EuGenoSuite_w.exe <Test.conf>`.

The results are generated in OUTPUT directory for each samples.