**Analysis of MS/MS data with the SPS**

The SPS analysis of a dataset starts by executing "**main\_specnets sps.params"** from the command line.

Example command lines, using "**<sps\_dir>**" to denote the path to SPS/SpecNets binaries:

* Run **main\_specnets** on the current node: "**<sps\_dir>/bin/main\_specnets sps.params**"
* Run **main\_specnets** on an SGE compute node: "**qsub -l h\_vmem=1G <sps\_dir>/bin/main\_specnets sps.params -g**"

**Parameters**

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Description** |
| -g | – | Run on SGE |
| -ll | Integer (0-9) | Log level (9 for less information) |
| -lf | File name | Log file name |

**Examples**

# Run project logging only errors, using parameters file 'sps.params'

~/sps/main\_specnets -ll 9 -lf log.txt sps.params

# Run project on sge grid logging only errors, using parameters file 'sps34.params'

~/sps/main\_specnets -ll 9 -lf log.txt sps34.params -g

# Run project logging errors and warnings, using parameters file 'sps.params'

~/sps/main\_specnets -ll 5 -lf log.txt sps.params -s

**Parameter files**

All parameter values, including the name of the file(s) containing the MS/MS spectra, are specified in the parameters file **sps.params**. Of course, you can choose any file name for the parameters file and multiple parameters files can coexist in the same directory.

The parameters file is a text file where comment lines start with '**#**', empty lines are ignored and parameters are specified using the format **PARAMETER\_NAME=PARAMETER\_VALUE**. The valid parameter names and ranges of values are given below.

**Main parameters (required)**

|  |  |  |
| --- | --- | --- |
| **Parameter name** | **Valid values** | **Description** |
| INPUT\_SPECS\_MS | Any valid file name | Names of the files containing the MS/MS spectra. Valid file formats are MGF, mzXML, ms2 and multi-spectra pkl. Multiple file names should be separated by ';'. |
| FASTA\_DATABASE | Any valid file name | Database of protein sequences in FASTA format. |
| EXE\_DIR | Any valid path | The directory containing the SPS / Spectral Networks binaries and configuration files. e.g.: **<install directory>/bin** |
| AMINO\_ACID\_MASSES | Any valid amino acid masses file | Used to select amino acid masses by fixed Cysteine blocking group: No blocking (set to AA\_standard.txt), blocked with IAA (set to AA\_cys\_iaa.txt) or blocked with NIPIA (set to AA\_cys\_nipia.txt) |
| REPORT\_DIR | Any absolute path | Output directory for report files (it will be created if non-existent). Should be accessible by the spsplot CGI at the given absolute path to enable interactive HTML reports allowing for user visualization and annotation of contigs and spectra. |
| REPORT\_TITLE | String | Title of report pages. Report title string cannot include spaces. |
| GRID\_EXE\_DIR | Any valid path | Path to SPS/SpecNets binaries on SGE compute nodes. Default value is "" (empty) |
| GRID\_NUMNODES | Any integer >= 0 | Number of SGE jobs to launch per SPS job. Default value is zero (no SGE grid node available) |
| GRID\_PARAMS | String | Parameters to be passed directly to SGE. Default is "-l h\_vmem=1G", which specifies the memory quota per SPS/SpecNets SGE job of 1 gigabyte. |
| GRID\_SGE\_EXE\_DIR | Any valid path | Path to SGE binaries (e.g., qsub) on SGE compute nodes |

**Optional parameters**

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter name** | **Valid values** | **Default** | **Description** |
| TOLERANCE\_PEAK | 0.0 - 0.4 | 0.4 | Peak mass tolerance (in Daltons) used for de novo sequencing. |
| TOLERANCE\_PM | 0.0 - 3.0 | 1.5 | Parent mass tolerance (in Daltons) used for de novo sequencing. |
| TOLERANCE\_PEAK\_PPM | Any number >0 | Use Da tolerance | Peak mass tolerance (in PPM) used for charge deconvolution and PRM clustering. |
| TOLERANCE\_PM\_PPM | Any number >0 | Use Da tolerance | Parent mass tolerance (in PPM) used for charge deconvolution and PRM clustering. |
| DECONV\_MS2 | 0/1 | 0 | Enables MS/MS charge deconvolution for input spectra prior to PepNovo scoring, should only be used with high-accuracy fragment masses |
| ACTIVATION | CID/HCD | CID | Specifies the activation method of input MS/MS spectra to choose the appropriate PepNovo scoring model |
| INSTRUMENT\_TYPE | FT/IT | IT | Whether or not fragment masses are high accuracy (FT means 0.05 Da tolerance or lower) – used for choosing the PepNovo scoring model |
| CORRECT\_PM | yes/no | no | Correct MS/MS spectra parent mass.  Should be no for high-accuracy parent masses. |
| GUESS\_CHARGE | yes/no | no | Guess MS/MS spectra precursor charge.  Should be no for high-accuracy parent masses. |
| MIN\_SPECTRUM\_QUALITY | 0.0 - 1.0 | 0.15 | MS/MS spectra with inferior quality scores are discarded. |
| CLUSTER\_MIN\_SIZE | Any integer >=0 | 1 | Minimum number of spectra per cluster to retain cluster-consensus spectrum for further analysis. Set to zero to disable clustering. |
| CLUSTER\_TOOL | PrmClust/  MSCluster | MSCluster | Which clustering tool to use. Besides MSCluster, one can cluster the PRM spectra after PepNovo scoring |
| MAX\_MOD\_MASS | Any number >0 | 100 | Maximum mass for a post-translational modification (in Daltons). Use absolute values for negative mass offsets (e.g. loss of water). |
| MIN\_OVERLAP\_AREA | 0.0 - 1.0 | 0.45 | Minimum percentage of overlapping mass between two spectra to compute spectral alignments. Lower values allow for the detection of small overlaps but lead to longer run times; usually not set to less than 0.4. |
| PARTIAL\_OVERLAPS | 0/1 | 1 | If 0, only allow spectral alignments where the endpoints align. |
| MIN\_RATIO | 0.0 - 1.0 | 0.35 | Minimum percentage of matched peak scores in a spectral alignment. |
| MIN\_MATCHED\_PEAKS | Any integer >0 | 4 | Minimum number of matched peaks in a spectral alignment. |
| MAX\_PVALUE | 0.0 - 1.0 | 0.05 | Maximum p-value to accept spectrum/spectrum alignment. Default value is 0.05 (may be too strict for datasets with small number of spectra). |
| FILTER\_TRIGS | yes/no | yes | Determines whether spectral alignments need to be confirmed by transitive closure. If set to "yes" then a spectral alignment between spectra A,B is only accepted if there are at least two other alignments A,C and B,C with consistent alignment offsets. Default is "yes", should be set to "no" for spectral networks projects. |
| TAG\_LEN | Any integer >=3 | 6 | Length of the sequence tags used for matching spectra/contigs against the FASTA database. |
| MIN\_MATCHED\_PEAKS\_DB | Any integer >=4 | 6 | Minimum number of matched peaks when aligning contig sequences against the FASTA database. |
| CLUSTALW\_MINSCORE | Any number >0 | 250 | Minimum ClustalW score to transfer contig/database alignments between database proteins using ClustalW protein/protein alignments (see Bandeira et al., Nature Biotechnology 2008 for details). Set to 10000 to disable cSPS homology assembly. |
| MIN\_METACONTIG\_SIZE | Any integer >=0 | 0 | Any value > 0 enables MetaSPS. Minimum allowable number of assembled contigs per meta-contigs (1 includes unmerged SPS contigs with meta-contigs for maximum coverage, 2 and higher include only meta-contigs of increasing size, which increases the average sequence length/accuracy of reported sequences with decreasing coverage). |
| MIN\_METACONTIG\_SCORE | 0.0 – 10.0 | No default | If enabling MetaSPS, must specify the minimum allowable overlap score between aligned contigs |

**Example parameters file**

# System parameters

INSTALLDIR=~/sps

REPORT\_DIR=./report

EXE\_DIR=$INSTALLDIR/bin

# SGE parameters

GRID\_NUMNODES=100

GRID\_NUMCPUS=1

GRID\_SGE\_EXE\_DIR=/opt/sge62/bin/lx24-amd64

GRID\_EXE\_DIR=$INSTALLDIR/bin

# Input files

REPORT\_TITLE=Test\_project

FASTA\_DATABASE=./data/homolog\_prots\_LC.fasta

AMINO\_ACID\_MASSES=../bin/AA\_cys\_iaa.txt

INPUT\_SPECS\_MS=./data/aBTLA\_LC\_AspN\_042707.mgf;./data/aBTLA\_LC\_chymotrypsin\_042707.mgf;./data/aBTLA\_LC\_pepsin\_30min\_042707.mgf;./data/aBTLA\_LC\_pepsin\_3h\_042707.mgf;./data/aBTLA\_LC\_trypsin\_042707.mgf;./data/aBTLA\_hybrid\_LC\_DTT\_IAA\_AspN\_ON\_100407.mgf;./data/aBTLA\_hybrid\_LC\_DTT\_IAA\_chymotryp\_30min\_100407.mgf;./data/aBTLA\_hybrid\_LC\_DTT\_IAA\_chymotryp\_3h\_100407.mgf;./data/aBTLA\_hybrid\_LC\_DTT\_IAA\_tryp\_30m\_100407.mgf;./data/aBTLA\_hybrid\_LC\_DTT\_IAA\_tryp\_ON\_100407.mgf

# Main parameters

TOLERANCE\_PEAK=0.4

TOLERANCE\_PM=1.0

# Preprocessing parameters

CLUSTER\_MIN\_SIZE=1

CLUSTER\_MODEL=LTQ\_TRYP

MIN\_SPECTRUM\_QUALITY=0.1

CORRECT\_PM=no

GUESS\_CHARGE=no

# Alignment parameters

MIN\_OVERLAP\_AREA=0.45

RESOLUTION=0.1

FILTER\_TRIGS=yes

MIN\_MOD\_MASS=-100

MAX\_MOD\_MASS=100

MIN\_RATIO=0.4

MAX\_PVALUE=0.05

MIN\_MATCHED\_PEAKS=4

PARTIAL\_OVERLAPS=1

# CSPS parameters

SPS\_PROJECTS=sps\_projects.txt

# Parameters for tag-based selection of homologous proteins

TAG\_LEN=6

MAX\_AA\_JUMP=2

DOUBLE\_AA\_JUMPS=1

MATCH\_TAG\_FLANKING\_MASSES=0

MAX\_NUM\_MODS=2

MIN\_MATCHED\_PEAKS\_DB=7

# Use line below to force a specific CSPS reference protein (index is 1-based)

# FORCE\_REFERENCE=1

# Parameters for clustalw sequence alignments

CLUSTALW\_MINSCORE=250