

## 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 MScCluster, 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_MINSORE	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/aBT
LA_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_3
0min_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

# CSPA parameters
SPA_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 CSPA reference protein (index is 1-based)
# FORCE_REFERENCE=1

# Parameters for clustalw sequence alignments
CLUSTALW_MINSORE=250

```