

# PepCache User Manual

Franz Roos, 25 February 2007

## Typical usage example (Linux console)

```
(new console)
mkdir job0003
cd job0003
./start.sh >logbatch.log
```

## Required input

- A copy of your current \*.param files needs to be in the directory where you start the search.
- FASTA files: specify their path in in\_fastafiles.param  
wrong paths are a frequent source of errors!
- mgf files: specify their path when starting the program; you can use wildcards to read many files at once
- posttranslational modifications: specify them in in\_AAmodifications.param

## Supervise memory and CPU usage of running jobs

Search only as many spectra at once as fit into your RAM. Typically, 100'000 spectra per Gigabyte fit well, or even more. You can use wildcards when specifying spectra to be loaded.

Windows NT, 2000, XP: CTRL-ALT-DELETE

Linux: top

## Parameter files \*.param

in_fastafiles.param	Decides which databases are used, and after which database results are written.
in_modifications.param	Specifies posttranslational modifications to be searched. _ internal modifications ( N-terminal peptide [ N-terminal protein ) C-terminal peptide ] C-terminal protein e.g. _S

## Result files

All output files will be written to your current directory.

*peptides*	Gives one line per spectrum, with plenty of information
*proteins*	Summarizes identifications by proteins; simple confidence calculation only
*scores*	Confidence distributions

## Penalty values (PepCache1.0, Services.cpp; compile if you change the values)

```
penalty_methox=0.3;
penalty_misscleav=0.3;
penalty_ptm=1.2;
penalty_tryptic=1.2;
penalty_genomic=1.2;
penalty_snp=2.0; // genomic + snp = 3.2
penalty_spliced=2.0; //genomic + spliced = 3.2
```

## Typical problems on Linux

- **dos2unix**; Symptoms: does not find paths, does not parse modifications properly. If you edit text files on Windows (source code, fasta files or parameter files) and transfer them to Linux, please use the dos2unix command in Linux to adjust the text file format to Linux. Otherwise, proteins in fasta files may be torn into pieces by wrong line breaks, and file paths are not recognized any more.

```
dos2unix *.param
```

- **permissions**: If a \*.sh file is not working, try `chmod 755 *.sh`

## Parameters

	Default	
aa	10	Spectrum preprocessing, number of highest peaks to retain per 100 Da of parent mass, e.g. parent mass 1850 Da, aa10 -> 185 peaks retained
ab		Spectrum preprocessing, sliding window size
ac		Spectrum preprocessing, number of highest peaks to retain within sliding window
ad		Spectrum preprocessing, sliding window size, isotope based, ...
ae		Spectrum preprocessing, isotope based, ...
bb		Hypergeometric model, additionally blocked bins
bm		How many of the best matches should be retained. If there are >5 or so, it takes time at the beginning of the search to sort the list several times. And each match uses memory during the search.
bn		Result reporting, confidence values, in how many bins the distributions are divided in the file out*scores.txt.
bs		Score development, write a file that shows the offsets around the b- and y-ions based on search results?
bw		Result reporting, how many of the best matches should be shown (usually, only the best is shown)
ch		Size of the processor L2 cache
cs		Only search doubly charged spectra (default = 1)
dl		Discrimination score: 0=best_score, 1=p-value (almost equal to best_score), 2=delta between best and 2 <sup>nd</sup> best score; 3=delta between best and 3 <sup>rd</sup> best score; ...
el		Speed performance timing, calculate scores more than once to measure time requirement CPU time / loading time
<b>gn</b>		<b>Splicing, gap length minimum</b>
<b>gx</b>		<b>Splicing, gap length maximum</b>

hs			Load a subset of spectra only out of the argument files, requires files in_hotspec*
ht	10000		Genomic search around "hotspots" only, e.g. already identified regions only, to save time; file containing list must be given in in_fastfiles.param
ma	<b>5.0</b>		<b>Mass tolerance above monoisotopic mass (default 5 Da) e.g. experimental mass 1838, theoretical monoisotopic 1834, is within</b>
mb	<b>1.0</b>		<b>Mass tolerance below monoisotopic mass</b>
md	1 true	bool	Search posttranslational modifications
mp	<b>500</b>		<b>Result reporting, minimum peptide confidence; multiply by 1000, if you want 0.99, give 990 as argument</b>
nk			Score development, hypergeometric score
np	<b>0</b>	<b>Bool</b>	<b>Splicing, allow for any prefix end (increases search space by 16!)</b>
ns	<b>0</b>	<b>Bool</b>	<b>Splicing, allow for any prefix end (increases search space by 16!)</b>
pn			<b>Allowed penalty for peptide modifications of all kind, sum of: PTM, oxidized methionines, missed cleavages, non-tryptic ends, genomic hits, ...</b>
pp			Penalty attributed to PTMs (except oxidized methionine)
rs			Sort spectra in the memory at the beginning, costs a little bit of time first, saves some time for long searches
rz			
sc			Scoring to be used
sa	<b>0</b>	<b>Bool</b>	<b>Search for potential SNPs based on protein/peptide sequence</b>
sn	<b>0</b>	<b>Bool</b>	<b>Search for SNPs based on nucleotide sequence</b>
sp	0	Bool	Search for splice sites
Ta	1	Bool	Cache optimization, let tuple buffer adapt automatically
Tb	10000		Cache optimization, set initial tuple buffer size
te	<b>2</b>		<b>Tryptic ends required; 2 = fully tryptic, 1 = semi-tryptic, 0 = non-tryptic allowed</b>
wd			Write subset of spectra as dta files
wg	<b>1</b>	<b>Bool</b>	<b>Search whole genome</b>
ws	0	Bool	Write subset of spectra as mgf files
wt	0	Bool	Write theoretical spectra
xs	1		Use only every xth spectrum that is loaded for the search, e.g. 58322 spectra are given as argument and -xs10, 5832 spectra will be used. Saves time for feasibility checks.