HyPep 1.0 Manual

Updated 9/28/2022

Contents

About Hypep

What is HyPep?

HyPep is an open-source Python based algorithm designed for hybrid analysis of neuropeptide mass spectral data through sequence homology searching and accurate mass matching.

License

Using HyPep

HyPep is freely available for download from GitHub (https://github.com/lingjunli-research/HyPep-v1.0) and has an included user interface for increased accessibility.

Computational requirements

HyPep requires Python 3 installation (https://www.python.org/downloads/). The Python version for confirmed use at time of release is Python 3.10.6. HyPep is only supported for use with Windows OS, though limited functionality may be available with MacOS and Linux/UNIX. HyPep is designed for functionality with .RAW mass spectra files, a file type proprietary to Thermo Scientific instruments.

Input files are also required from companion software: PEAKS (Bioinformatics Solutions, Inc.), TopFD (Xiaowen Liu Research Group), and RawConverter (John Yates Research Group). TopFD requires .mzML input; .RAW files can be converted to .mzML file type through

MSConvert (Proteowizard) and can be converted to .MS2 file type through RawConverter or MSConvert.

HyPep can be run using the Windows command prompt, but it is generally a good practice to run through an Anaconda environment.

Software package	Version verified for use	Download location
Python	3.10.6	https://www.python.org/downloads/
PEAKS	PEAKS Studio Xpro	https://www.bioinfor.com/peaks-studio/
RawConverter	1.2.0.1	http://fields.scripps.edu/rawconv/
TopFD	1.5	https://www.toppic.org/software/toppic/register.html
MSConvert	3	https://proteowizard.sourceforge.io/download.html
Anaconda	3.9	https://www.anaconda.com/products/distribution

Installation and start up

1. Download HyPep from GitHub (https://github.com/lingjunli-research/HyPep-v1.0), save to any location in the C: drive.

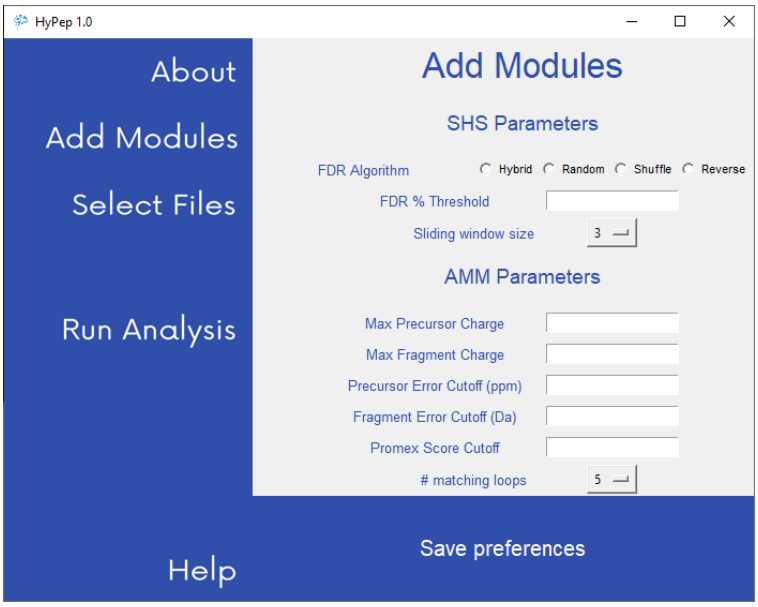
2. Open command prompt (either Anaconda or built-in)

3. Navigate to downloaded HyPep folder within the command prompt

4. Launch HyPep GUI using the input: `python gui_flattened.py`

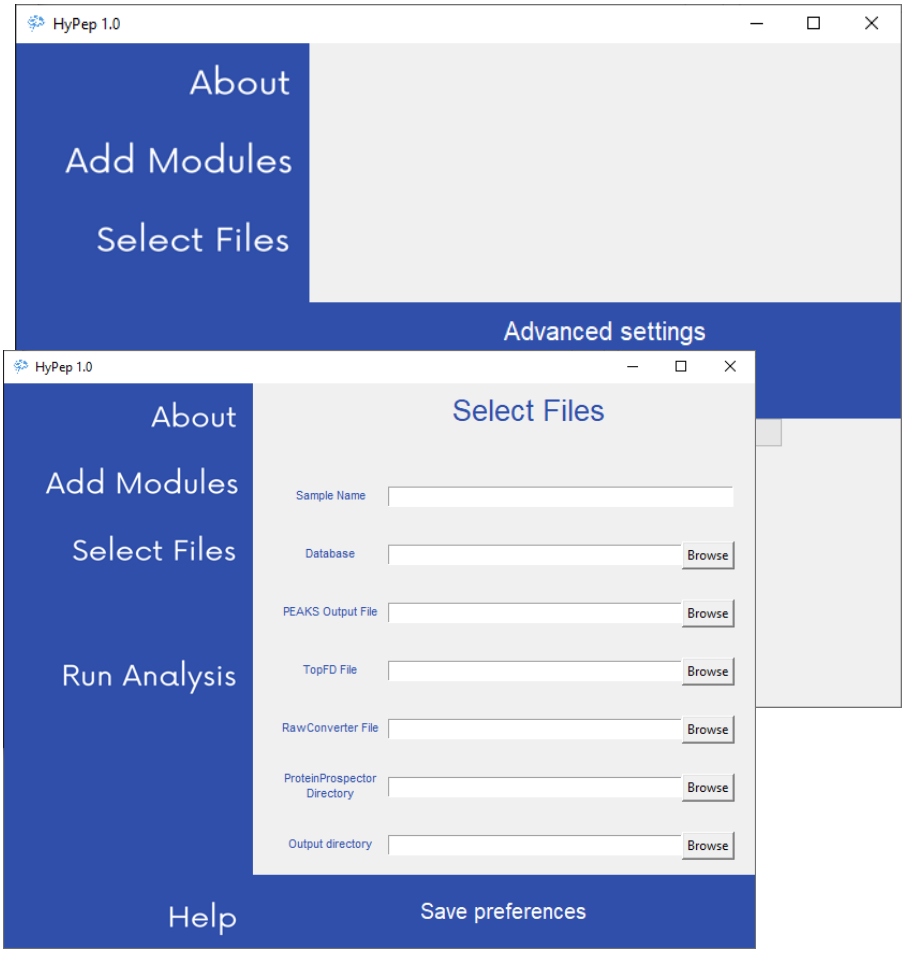
HyPep user interface

The HyPep user interface has three primary pages: Add Modules, Select Files, and Run Analysis, designed to access in this order.



Each parameter of Add Modules is described in later sections. It should be noted that prior to moving to the Select Files page, “Save Preferences” must be selected.

Each selection of Select Files is described in later sections. It should be noted that prior to moving to the Run Analysis page, “Save Preferences” must be selected



After the Add Modules and Select Files pages have been completed, “Run Analysis” within the Run Analysis page can be selected. Analysis has begun when the green bar begins to move. All progress of analysis will be reported in the command prompt.

Loading data

In selecting data, simply click the “browse” button to the right of the input field within the Select Files page. Unless noted, all .csv/.txt file headers must match the template files.

Sample name

Input a sample name within this field, no spaces or special characters are permitted.

Database

The database is input as a .csv file with 3 columns: Accession, Sequence, Monoisotopic $[M+H]^+$ *m/z* value.

	A	B	C	D
1	Accession	Sequence	Monoisotopic Mass	
2	1	LRVAPEESPVL	1209.68376	
3	2	LRVAPEEHPVLL	1372.7947	
4	3	YAFGL(Amidated)	569.30821	
5	4	YSFGL(Amidated)	585.30312	
6	5	YTFGL(Amidated)	599.3188	
7	6	YDFGL(Amidated)	613.298	
8	7	SYAFGL(Amidated)	656.3402	

PEAKS Output File

The HyPep input file from PEAKS is the *de novo peptides*.csv file.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Fraction	Source File	Feature	Peptide	Scan	Tag Length	Denovo Score	ALC (%)	length	m/z	z	RT	Predict RT	Area	Mass	ppm	PTM	local confidence (%)	tag (>=0%)	mode
2	27	2021_0817_Brain_1.raw	F27:48539	ESDAAEVK	F27:16840	14	99	99	14	753.3717	2	34.24	-	1.17E+07	1504.726	2.2		100 100 100 100 100 1(	ESDAAEVKV	HCD
3	27	2021_0817_Brain_1.raw	F27:48996	ESDAAEVK	F27:25238	27	99	99	27	759.1259	4	52.92	-	3.89E+07	3032.467	2.4		100 100 100 100 100 1(	ESDAAEVKV	HCD
4	27	2021_0817_Brain_1.raw	F27:43669	ESDAAEVK	F27:19220	19	99	99	19	697.6764	3	39.64	-	1.46E+07	2090.002	2.8		100 100 100 100 100 1(	ESDAAEVKV	HCD
5	27	2021_0817_Brain_1.raw	F27:16596	EVGKLLLK	F27:28002	8	99	99	8	450.3006	2	58.71	-	1.92E+06	898.5851	1.7		100 100 100 100 100 1(	EVGKLLLK	HCD
6	27	2021_0817_Brain_1.raw	F27:20935	KSVFDQLK	F27:21488	8	99	99	8	482.7776	2	44.63	-	4.12E+06	963.5389	1.9		100 100 100 100 100 1(	KSVFDQLK	HCD
7	27	2021_0817_Brain_1.raw	F27:8402	EAEATLK	F27:9260	7	99	99	7	381.2062	2	17.68	-	4.27E+06	760.3967	1.5		100 100 100 100 100 1(	EAEATLK	HCD
8	27	2021_0817_Brain_1.raw	F27:8402	EAEATLK	F27:9266	7	99	99	7	381.2062	2	17.68	-	4.27E+06	760.3967	1.5		100 100 100 100 100 1(	EAEATLK	HCD
9	27	2021_0817_Brain_1.raw	F27:35872	TMDVKVY	F27:12322	10	99	99	10	611.7941	2	24.15	-	1.72E+06	1221.57	2.9		100 99 100 100 100 10(	TMDVKVYH	HCD
10	27	2021_0817_Brain_1.raw	F27:20935	KSVFDQLK	F27:21493	8	99	99	8	482.7776	2	44.63	-	4.12E+06	963.5389	1.9		100 100 100 100 100 1(	KSVFDQLK	HCD
11	27	2021_0817_Brain_1.raw	F27:16983	KFDPSALK	F27:13180	8	99	99	8	453.2583	2	25.93	-	7.02E+06	904.5018	0.3		100 100 100 100 99 10(	KFDPSALK	HCD
12	27	2021_0817_Brain_1.raw	F27:10230	RVEVSYER	F27:6092	9	99	99	9	398.5516	3	11.75	-	3.39E+06	1192.631	1.6		99 99 100 100 100 100	RVEVSYERR	HCD
13	27	2021_0817_Brain_1.raw	F27:48539	ESDAAEVK	F27:16852	14	99	99	14	753.3717	2	34.24	-	1.17E+07	1504.726	2.2		99 98 100 100 100 100	ESDAAEVKV	HCD
14	27	2021_0817_Brain_1.raw	F27:16983	KFDPSALK	F27:13122	8	99	99	8	453.2583	2	25.93	-	7.02E+06	904.5018	0.3		100 100 100 100 99 10(	KFDPSALK	HCD
15	27	2021_0817_Brain_1.raw	F27:9899	LGPMMKL	F27:33200	7	99	99	7	395.2219	2	69.73	-	3.34E+06	788.4288	0.5		99 100 100 100 100 10(	LGPMMKL	HCD
16	27	2021_0817_Brain_1.raw	F27:35409	EAEATLKD	F27:30924	11	99	99	11	607.8369	2	65.09	-	2.27E+08	1213.655	3.3		99 99 100 100 100 100	EAEATLKDL	HCD

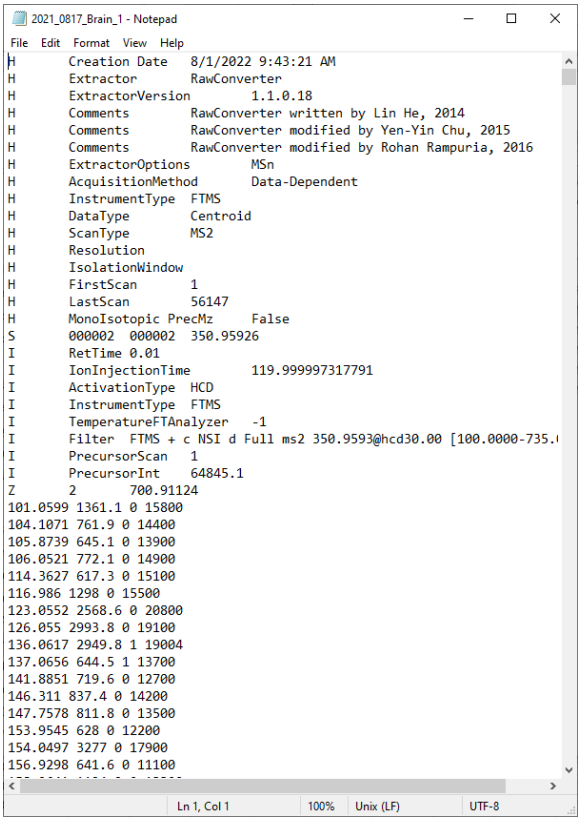
TopFD file

The HyPep input file from TopFD file is a .csv file with a title ending in “frac.mzrt” after output from TopFD.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ID	Fraction_I	Envelope_	Mass	MonoMz	Charge	Intensity	mzLo	mzHi	rtLo	rtHi	color	opacity	promex_score	
2	0	0	14	418.259	419.266	1	1.75E+10	419.166	421.371	99.2908	99.4825	#FF0000	0.1	-339.146	
3	1	0	6	408.276	409.284	1	6.94E+09	409.184	411.389	99.0657	99.1578	#FF0000	0.1	-339.146	
4	2	0	3	5608.91	1403.23	4	1.52E+08	1403.13	1405.59	98.488	98.5062	#FF0000	0.1	53.0046	
5	2	0	11	5608.91	1122.79	5	2.85E+09	1122.69	1124.69	98.4687	98.5265	#FF0000	0.1	53.0046	
6	2	0	16	5608.91	935.826	6	6.04E+09	935.726	937.429	98.4298	98.5448	#FF0000	0.1	53.0046	
7	2	0	16	5608.91	802.28	7	1.08E+10	802.18	803.669	98.4388	98.5448	#FF0000	0.1	53.0046	
8	2	0	12	5608.91	702.121	8	3.17E+09	702.021	703.349	98.4298	98.5665	#FF0000	0.1	53.0046	
9	3	0	11	1182.77	592.392	2	7.59E+08	592.292	593.996	9.91483	10.0148	#FF0000	0.1	54.4977	
10	3	0	18	1182.77	395.264	3	9.58E+09	395.164	396.367	9.893	10.132	#FF0000	0.1	54.4977	
11	3	0	11	1182.77	296.7	4	2.78E+09	296.6	297.552	9.91483	10.0488	#FF0000	0.1	54.4977	
12	3	0	1	1182.77	237.561	5	4.74E+06	237.461	238.263	9.927	9.927	#FF0000	0.1	54.4977	

RawConverter File

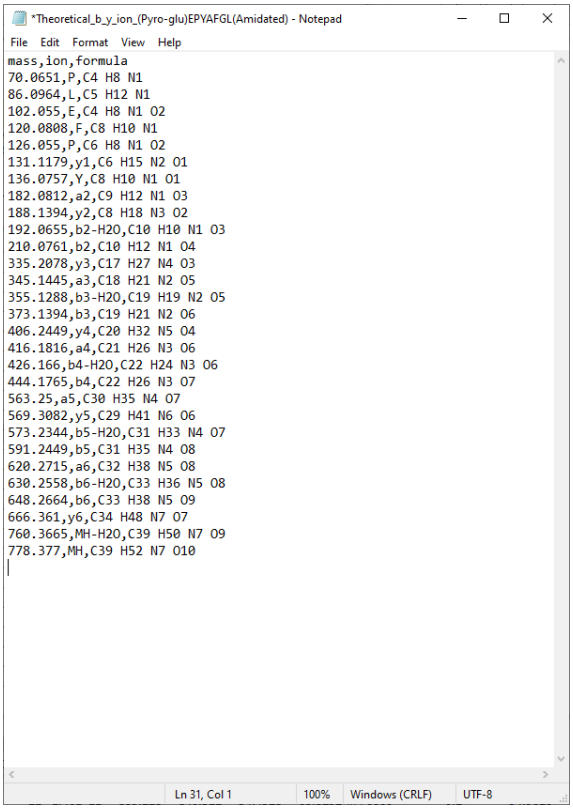
The HyPep input file from RawConverter or MSConvert is a .MS2 file.



ProteinProspector Input

A comma-delimited .txt file is required for each database sequence, which contains the ion mass, ion identity, and the ion chemical formula. If ion chemical formula is not known, keep the column header and use this column to repeat the ion identity. Column headers should be: mass, ion, formula. The file name should for each sequence should begin with “Theoretical_b_y_ion_” followed by the peptide name, written exactly as it is written in the database. These ion lists can be generated from ProteinProspector (https://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msproduct), or another program of choice.

*An alpha version of HyPep 2.0 is available upon request, which has this function automated.



Output Directory

The output directory can be located in any drive and is an empty folder where all results will be exported. If folder is not empty, previous results will be written over. The output folder must not be a sub-directory of the input folder.

Search options

False discovery rate algorithm selection

Reverse

Reverses each database sequence in original target database.

Shuffle

Randomly shuffles amino acid residues within each sequence in original target database.

Random

All amino acids in original target database are concatenated into one sequence, shuffled, and new sequences are formed while maintaining the sequence length frequency from the original target database.

Hybrid

A novel decoy database generation method which is a hybrid of shuffle and random decoy database methods. All amino acids in original target database are concatenated into one sequence, shuffled, and new sequences are formed while 1) maintaining the sequence length frequency from the original target database and 2) the amino acid frequency pattern of each sequence from the original target database is conserved in each corresponding decoy sequence.

False discovery rate threshold
The false discovery rate corresponds to the expected number of theoretical false positive identifications.

Sliding window size
The sliding window size controls the size of the amino acid window during alignment between a *de novo* sequenced peptide query and database sequence. The sliding window size ranges from 1-10. Must be a whole number. Recommended value is 2.

Maximum precursor charge
The maximum precursor charge is the maximum charge value that will be considered for a precursor peak. All charges from 1-*n*, where *n* is the maximum precursor charge, will be analyzed. Must be whole number. Recommended value is 8.

Maximum fragment charge
The maximum fragment charge is the maximum charge value that will be considered for a fragment peak. All charges from 1-*n*, where *n* is the maximum fragment charge, will be analyzed. Must be whole number. Recommended value is 4.

Precursor error cutoff
Error margin for a precursor match.

Fragment error cutoff
Error margin for a fragment match.

Promex score cutoff
Promex is a likeliness score exported from TopFD, where higher values represent greater likeliness. Recommended value is -10.

Number of matching loops
The number of loops that will be conducted to make PSMs while maximizing the number of unique peptide identifications (*i.e.*, more loops will result in more identifications to an extent). Recommended value is 5.

Interpreting Results
All results will be located in the specified output folder, in the sub-folder named “final_results”. Results from the SHS module are called “SampleName_SHS_final_report”, the results from the AMM module are called “SampleName_AMM_final_report”, and the final HyPep output containing the combined SHS and AMM module results is labeled “SampleName_combined_AMM_SHS_final_report”. Additionally, all parameters selected prior to analysis are reported as a parameter .txt file.

Troubleshooting
With any issues, please either submit a public issue on the HyPep GitHub (https://github.com/lingjunli-research/HyPep-v1.0/issues), or email lawashburn@wisc.edu with questions. Future releases of HyPep will be announced at https://www.lilabs.org/resources.