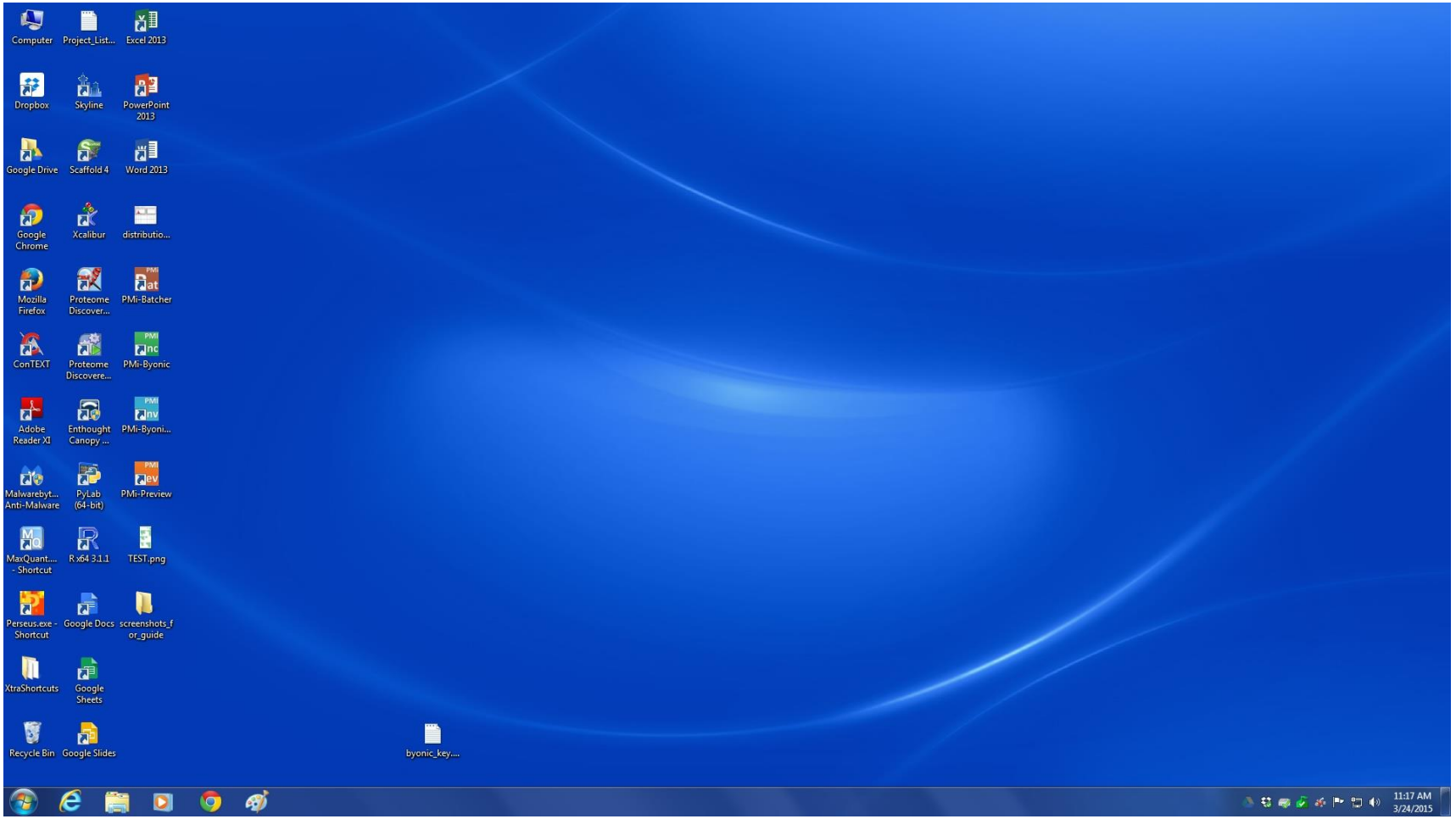
Comet/PAW Pipeline

John Klimek
OHSU Proteomics Shared Resource
Mar 2015

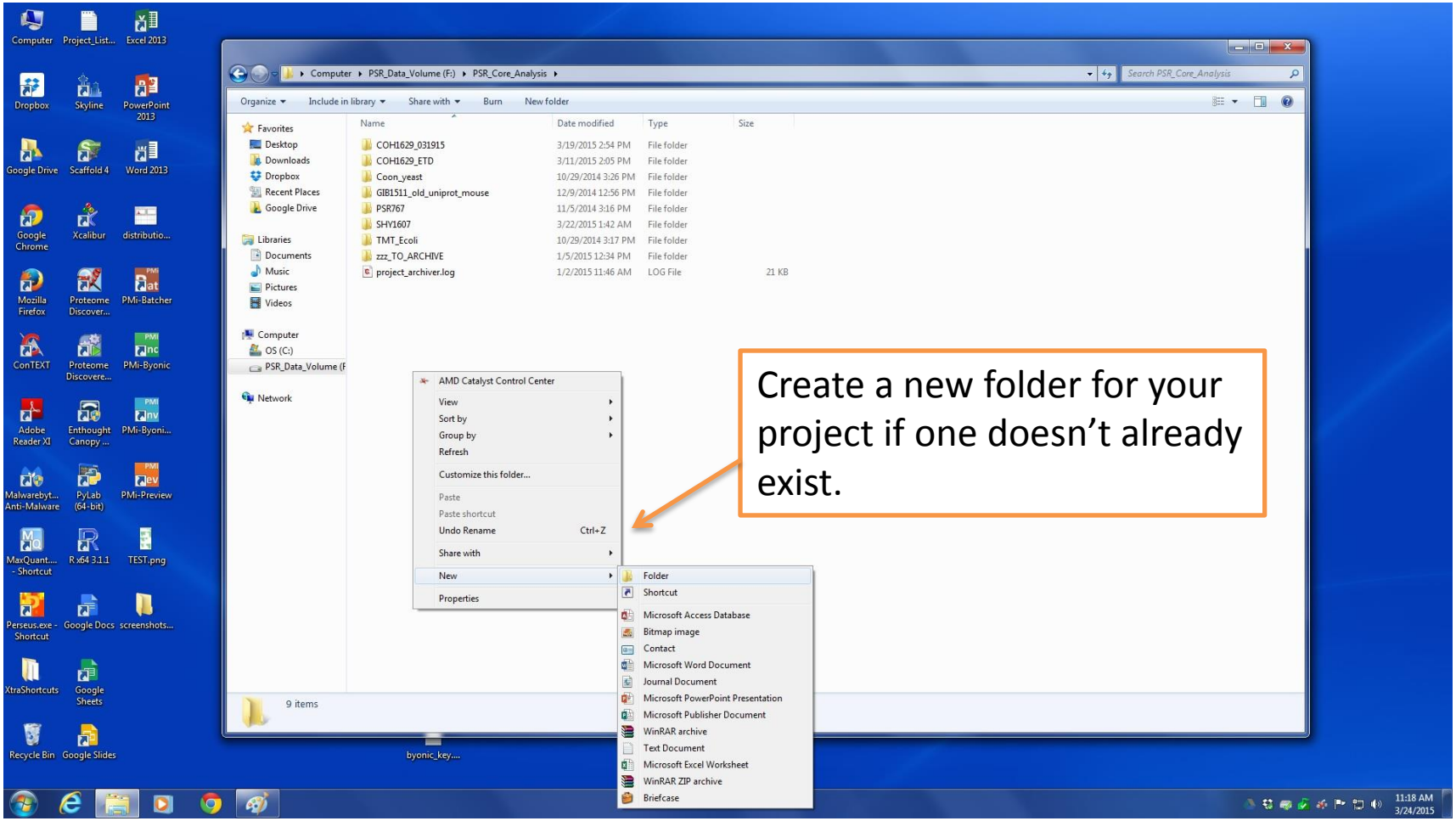
Some notes before you start

- This outline is for analyzing data on the PSR computers. For other computers, you will need to install Python, Comet, and other software to use the workflow. You can contact PSR for a complete list of requirements.
- Three computers in room MRB521 are set up to use this pipeline. They are: MRBA608, MRBA610, & MRBA611.
- A whiteboard on the wall is set up to communicate computer usage. Please try to choose a computer that is less utilized, and note your project on the whiteboard if you plan to step away while your data is processing.
- Several of the programs will show a Windows dialog box upon starting them that will ask if it is okay to allow the program to run. You can press **ok**.
- When closing out a python program running a GUI, remember to close the GUI window first. If you do not, you may see warning when closing other windows.

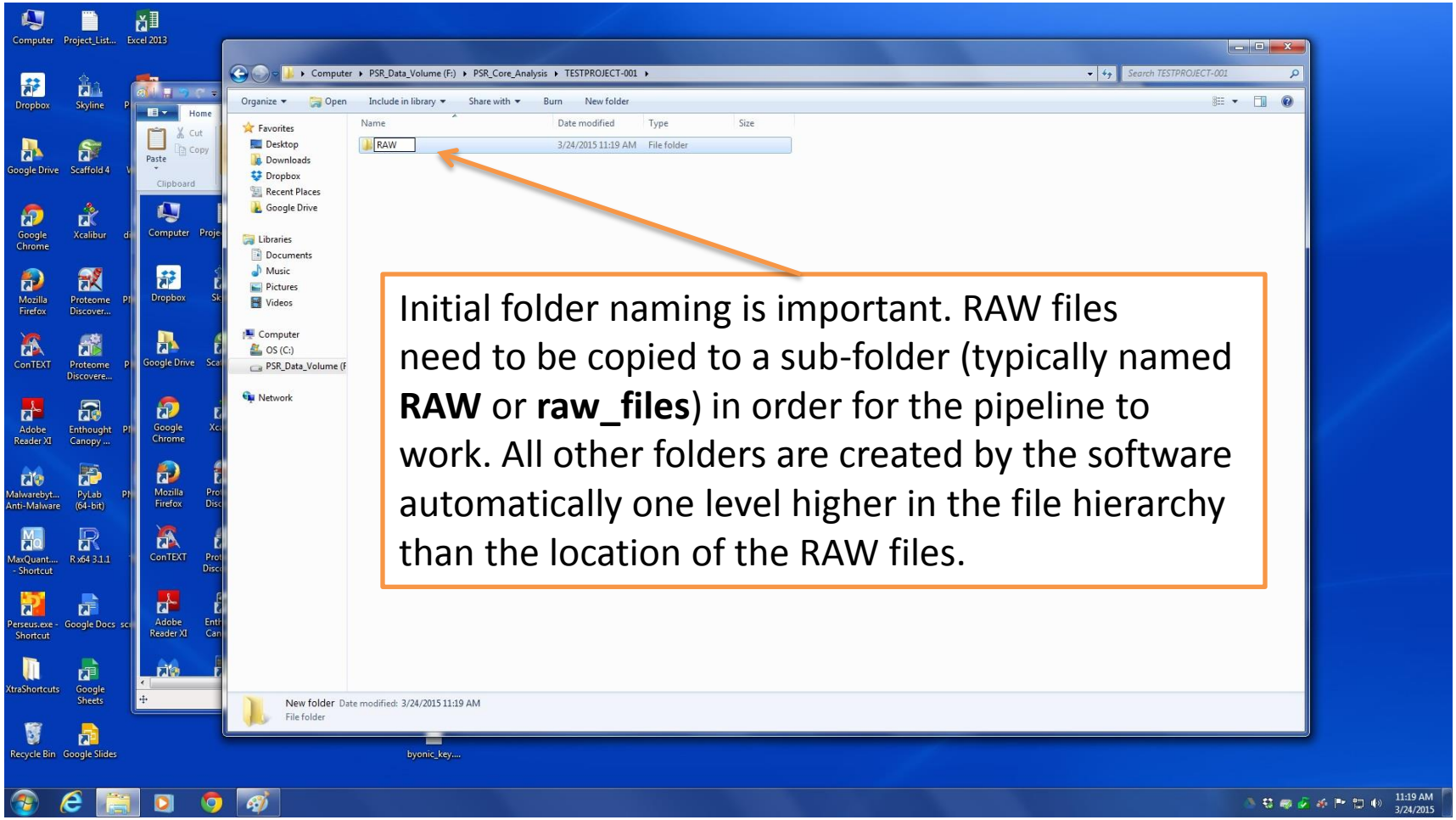
The first step is to get your RAW data from the mass spectrometer onto a flash drive or other device and transfer those files to the computer you'll be doing the analysis on.



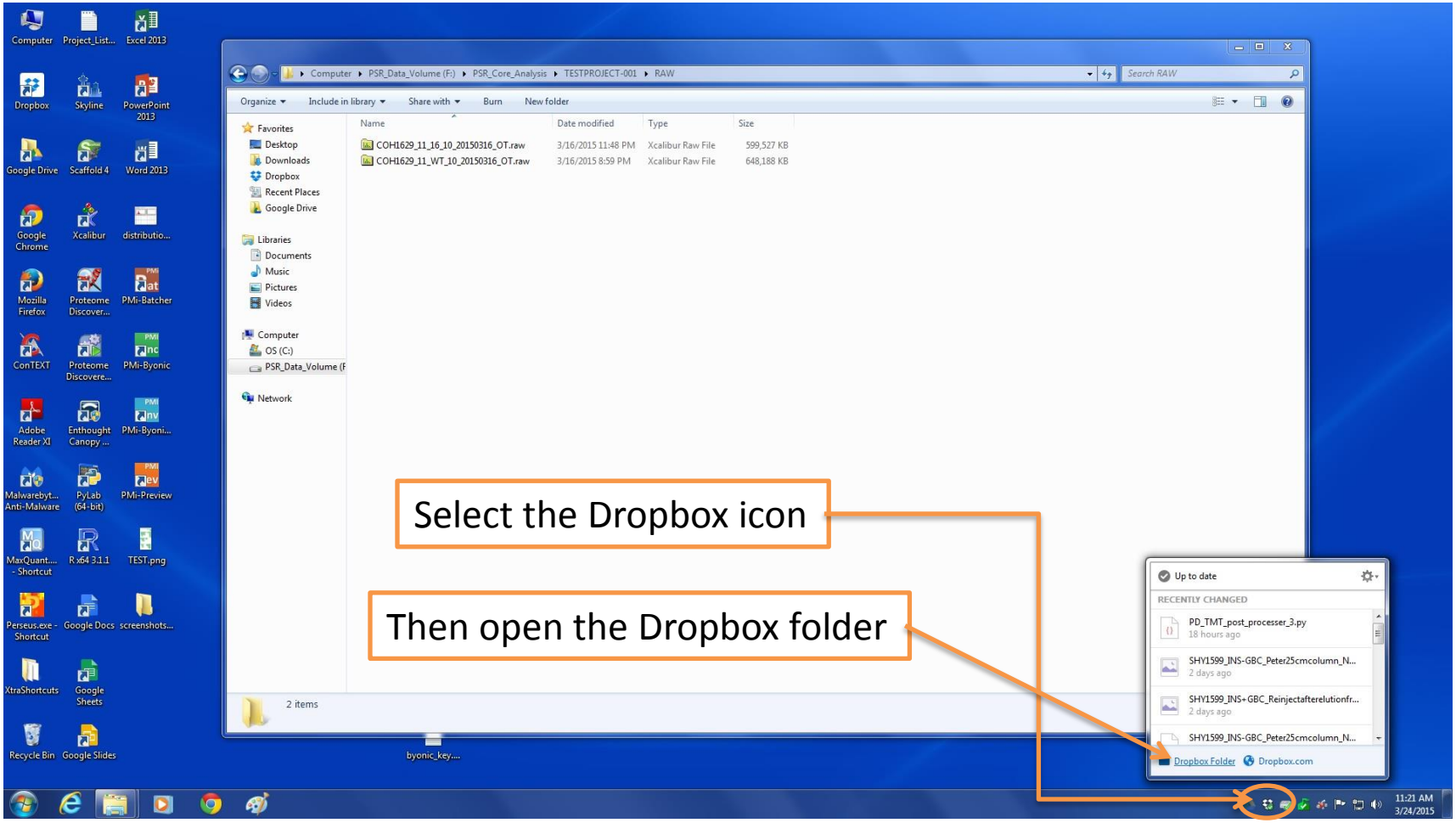
Each computer has a system hard drive (C), and a data analysis drive (F).
You'll want to put the data on the **F:** drive. Data will need to be given a PSR project number and placed in the **F:\PSR_Core_Analysis** folder in order to be archived.



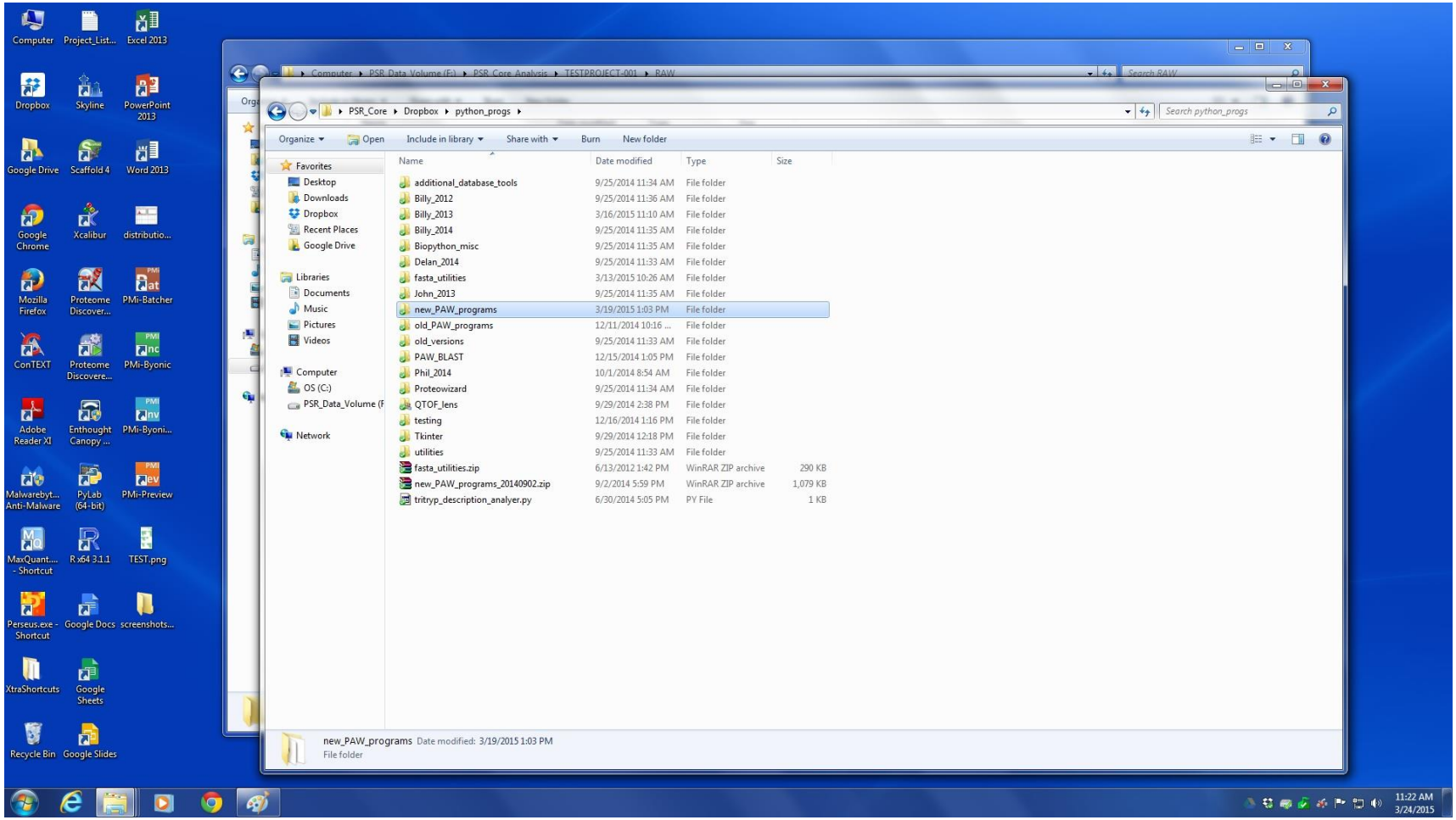
The folder name for a project should match the project ID in either ilab or Project Bubble. In either case, the code will start with the first 3 letters of the PI's last name. Project Bubble projects follow this with a 4-digit number (i.e. **DAV1620**), ilab projects will have a dash then the project number (i.e. **DAV-27**).



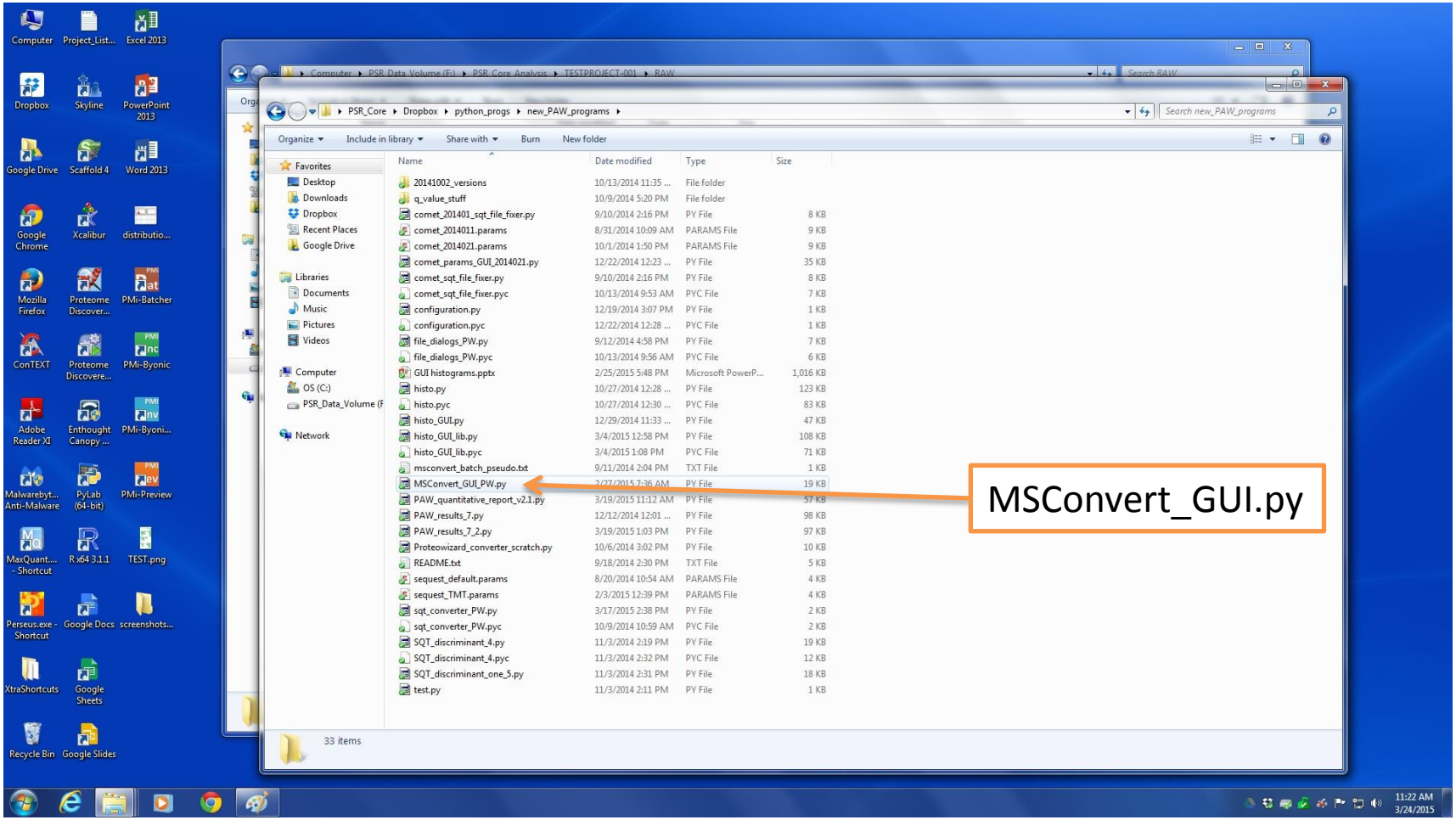
Once the RAW files are copied into the correct directory, then it's time to start analyzing the data. PAW Programs can be found in the **Dropbox** folder indicated below.



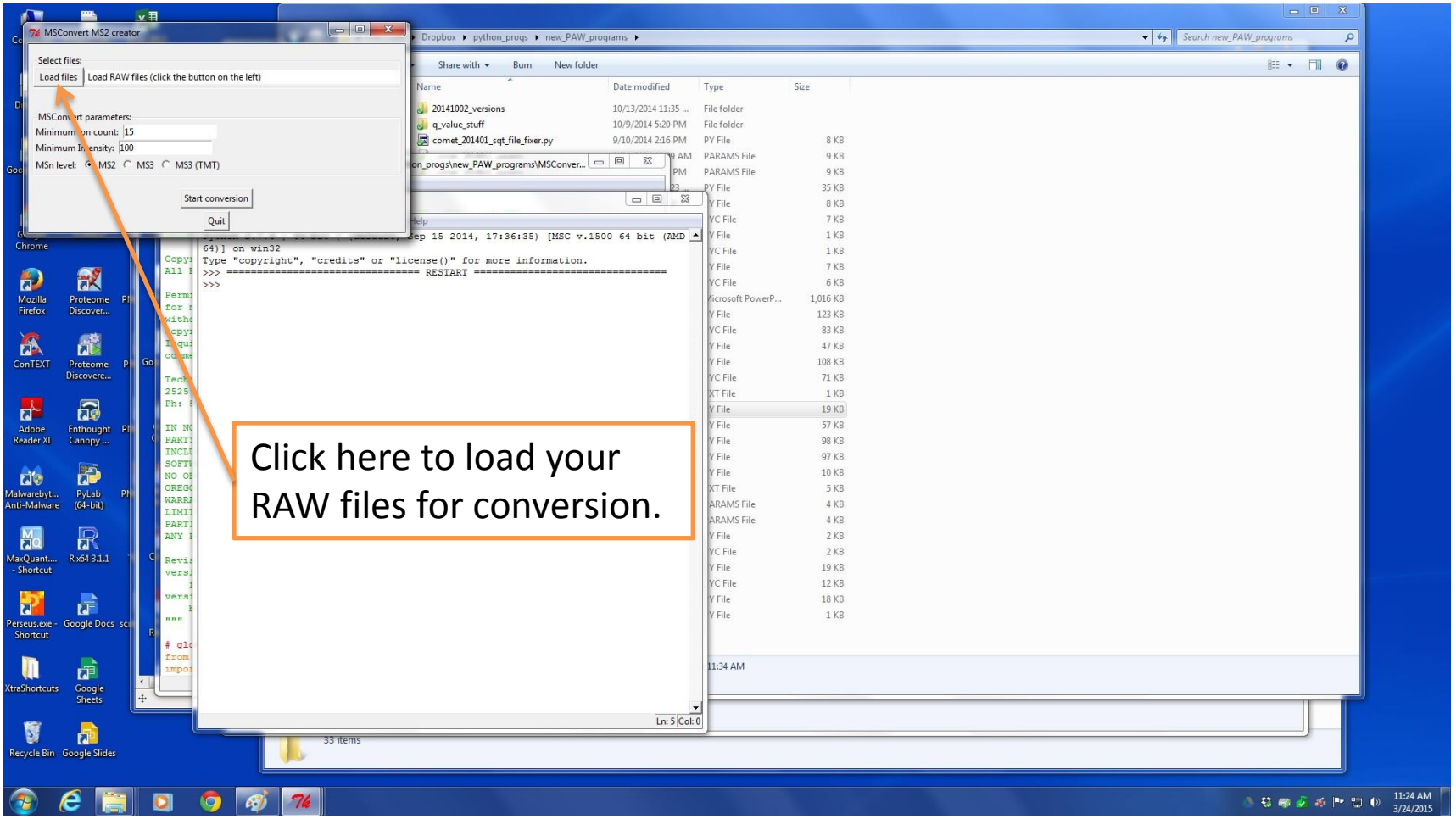
Once you open the **Dropbox** folder, open the **python_progs** folder, and then the **new_PAW_programs** folder as shown below.



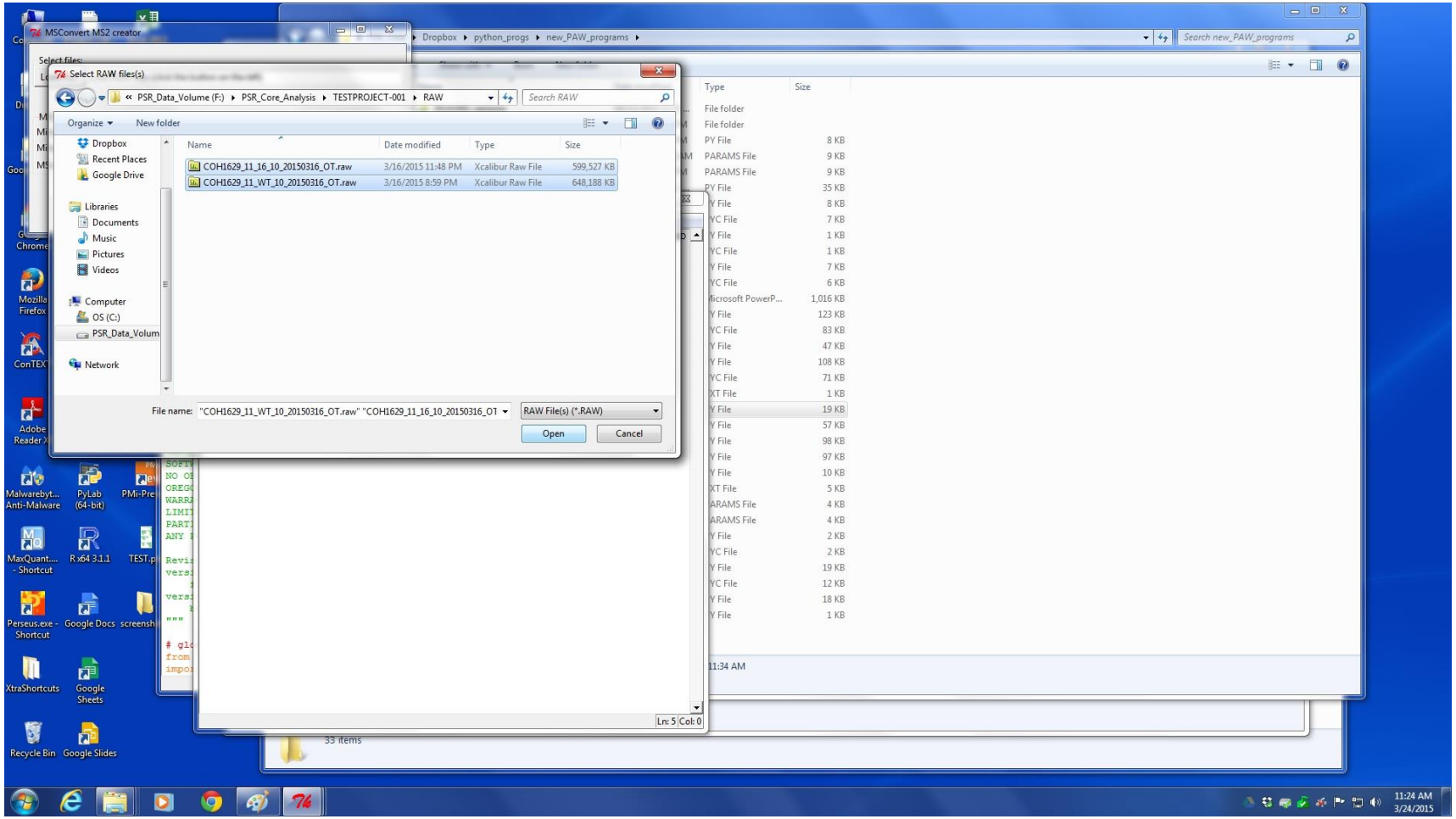
The first program to run is **MSConvert_GUI.py**. This program will create ms2 files to be used by COMET, and place them in a **ms2_files** folder. Double-click on the program to start it.



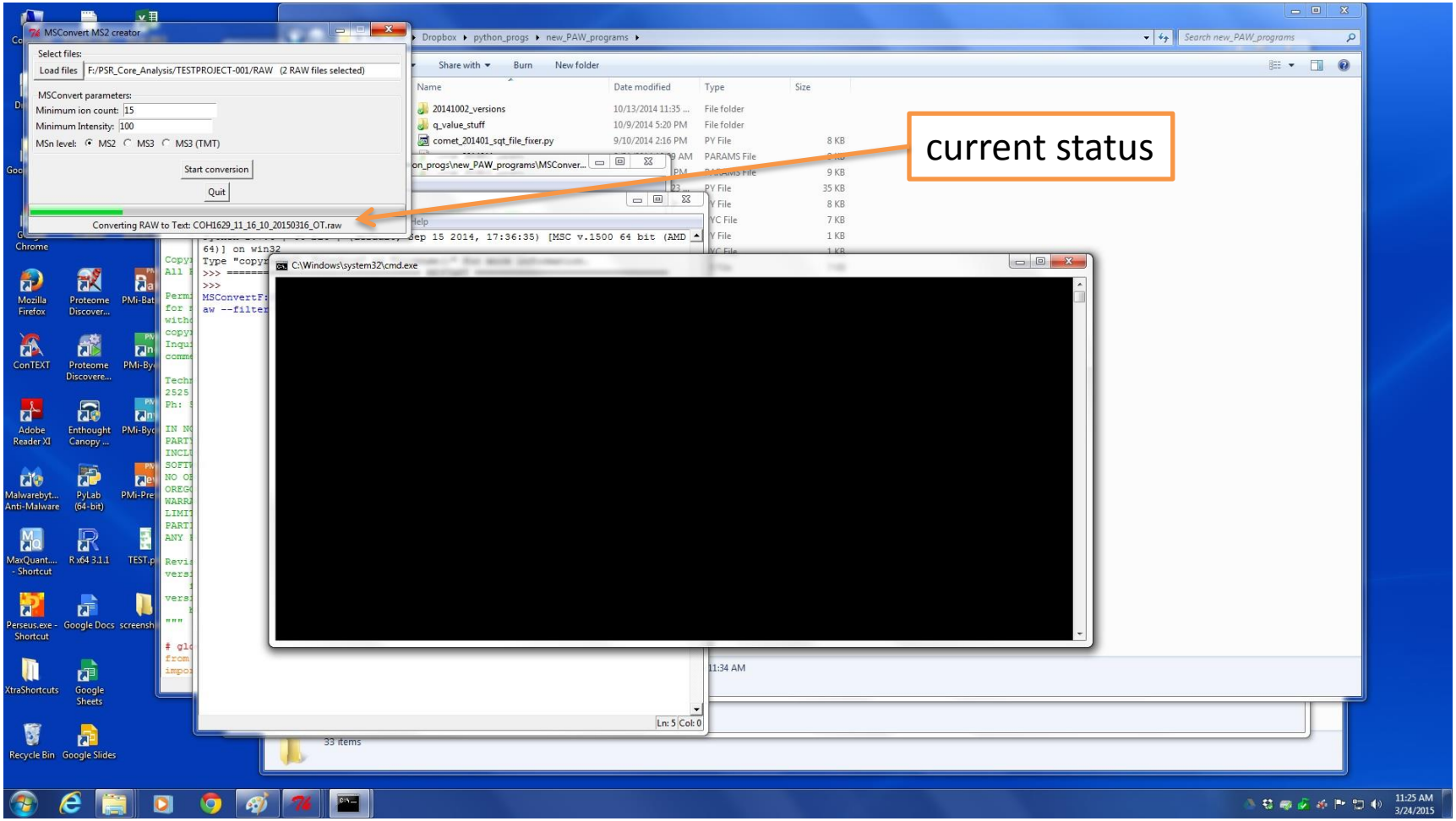
Once the program opens press **F5** to run (you shouldn't need to edit the source code). The GUI will pop open at this point. The MSConvert parameters should be OK in most cases, so you'll only be using the **Load files** and the **Start conversion** buttons.



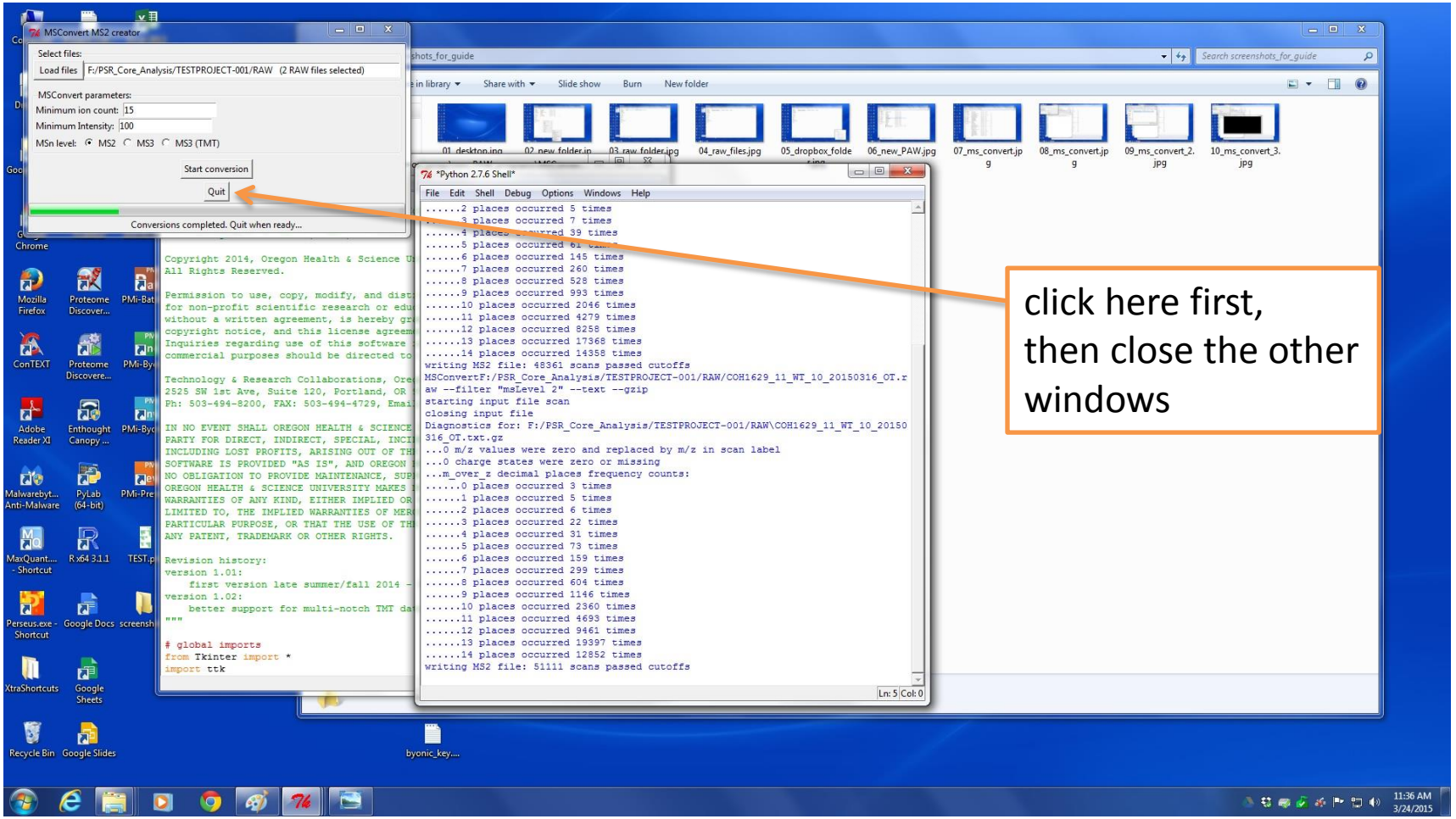
When the dialog window opens, navigate to your directory and select all the RAW files you wish to analyze. Then press **Open** to get back to the main GUI, and then press **Start Conversion** to process the files.



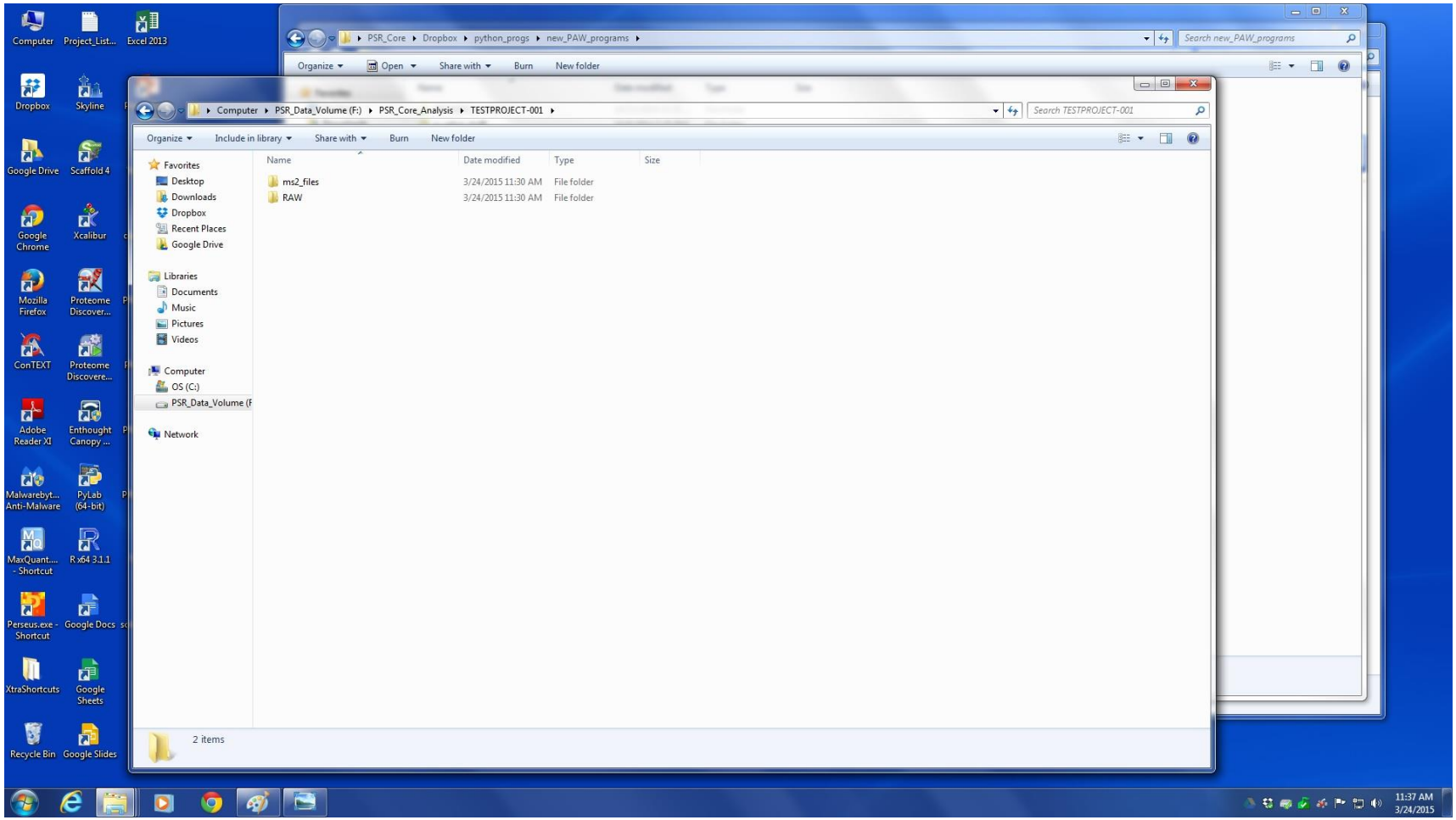
Once the conversion is started, a command window will pop up and the progress bar in the GUI will appear. The status is also listed at the bottom of the GUI window. The conversion can take several minutes for each RAW file, so you may want to step away for a few minutes while the data processes.



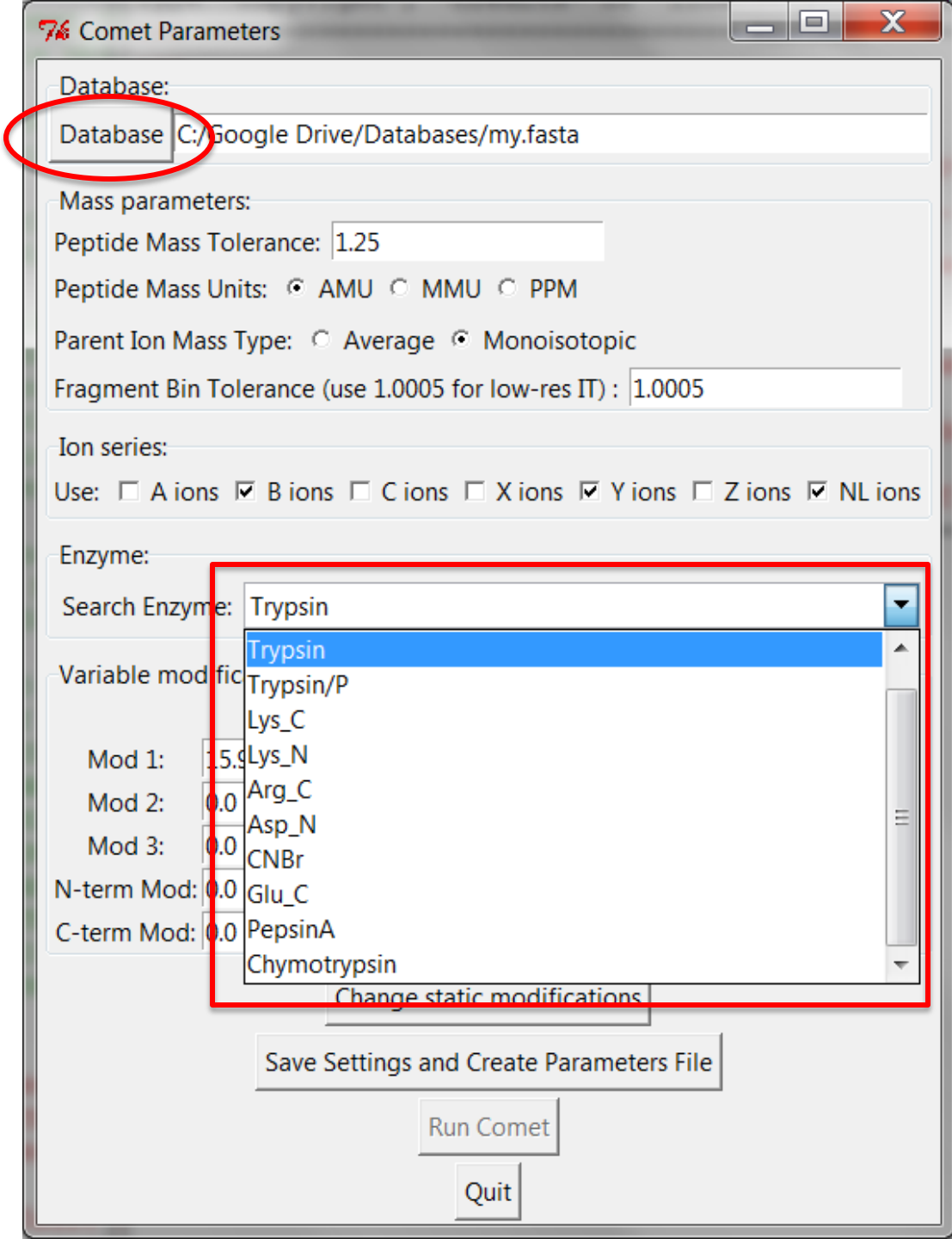
Once the conversion is complete, the windows will appear similar to below with some information about the conversion displayed, and the status bar in the GUI will read **Conversions completed. Quit when ready…** At this point, you can click the **Quit** button then close the other python windows.



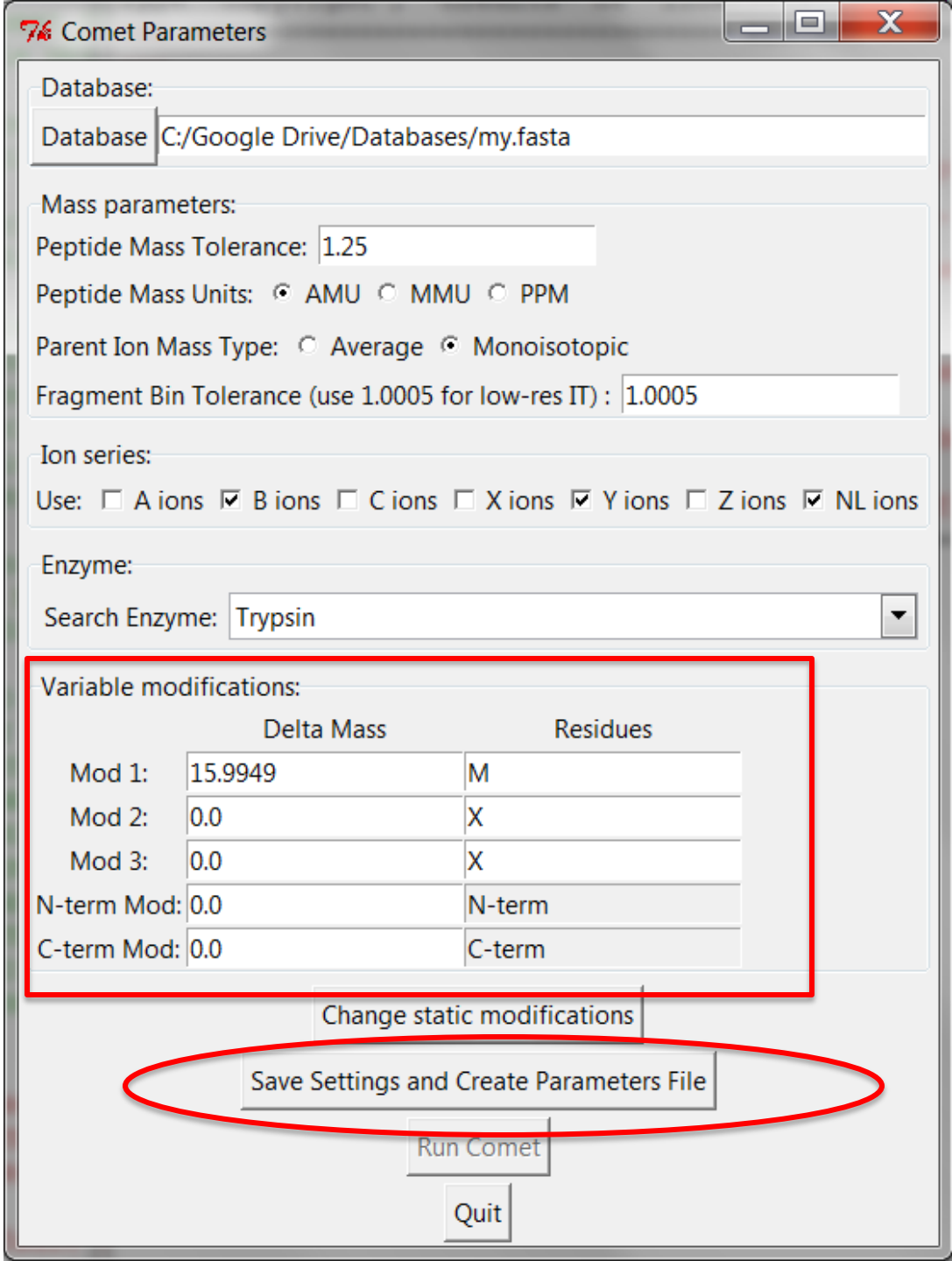
At this point, the project folder should look similar to below with two sub-folders: **ms2_files** and **RAW**.



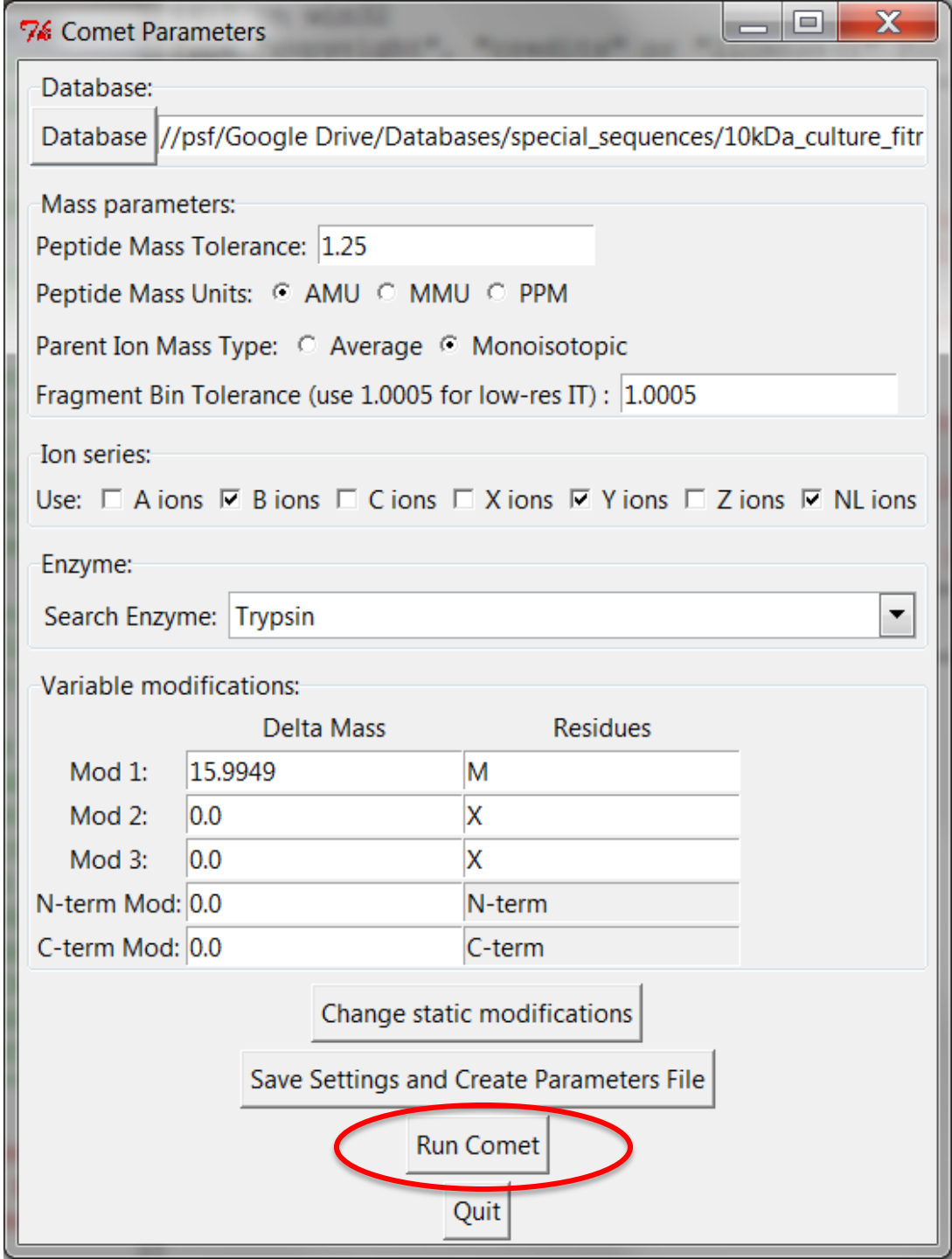
After ms2 file creation, Comet database searches can be run. Open the **comet_GUI.py** program. It starts the same way as the others, double-click to open the program, and then press **F5** to run. A GUI window (see right) will open. Comet has many options. The GUI presents the most commonly used ones. Click the **Database** button and browse to the FASTA protein database location (usually on Google Drive). The Mass parameters pane is set for Orbitraps. Use **2.50**, **AMU**, and **Average** for low-resolution traps. Ion series is set for CID or HCD. Use **C** and **Z** ions instead of B and Y ions for ETD (neutral loss ions are OK). Select the enzyme used to digest your sample (usually **Trypsin**).



Variable modifications are specified next. The fewer mods the better. Oxidized Met is a common mod that happens during sample processing and is set by default. Accurate modification mass is important for Orbitraps. Consult with PSR staff for help. Static modifications are done through a separate pop up window (**Change static modifications** button). The default is alkylation of Cys (C+57). Once parameters are set, press the **Save Settings and Create Parameters File** button. A dialog box will pop up and ask you to located the folder with the ms2 files created above. A **comet.params** file will be written in the folder with your ms2 files.



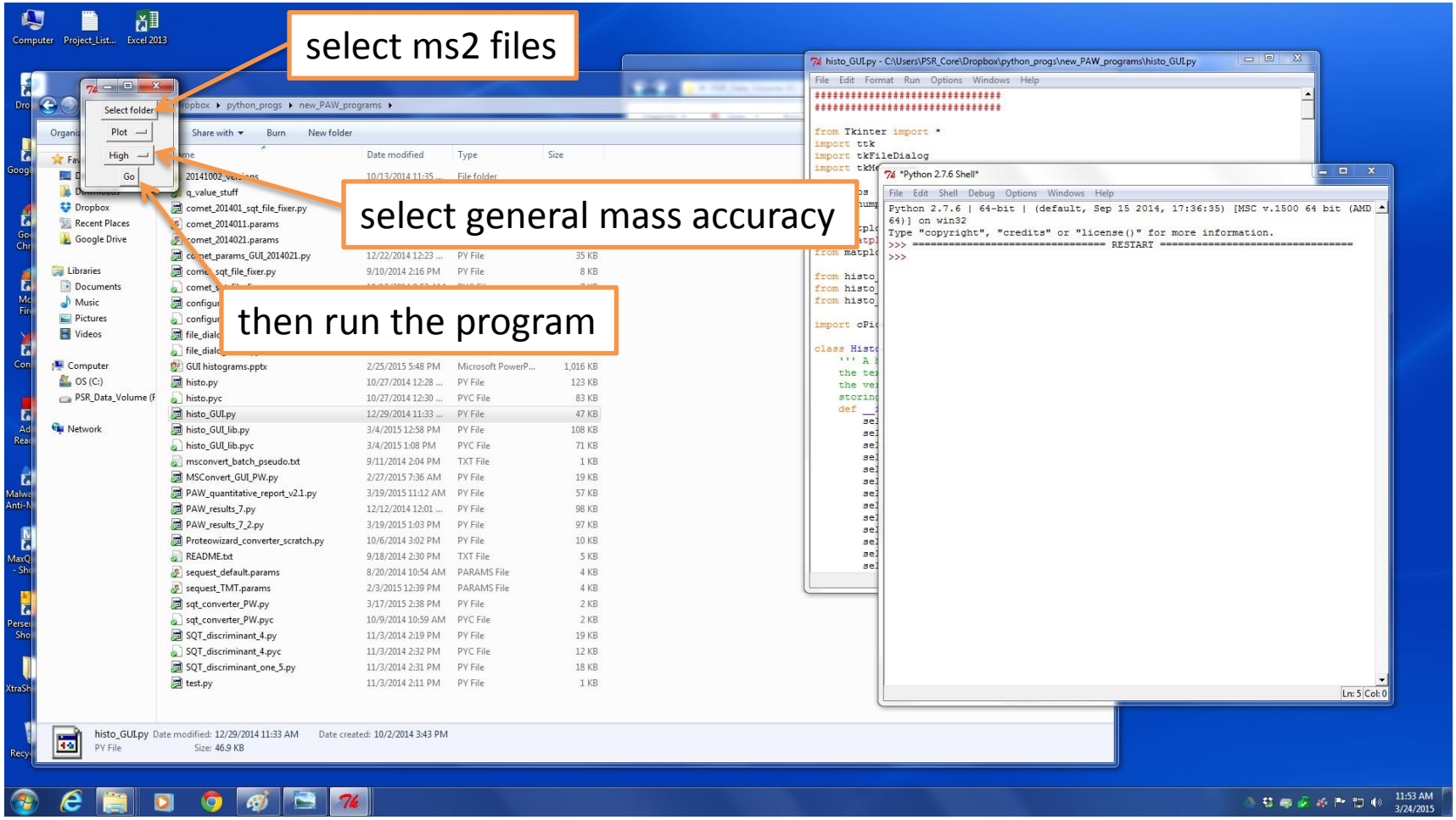
After the parameters file has been written, Comet searches can be launched by clicking the **Run Comet** button. The GUI window will lengthen to show a progress bar and status message. The console window will also display information during search progress. Black command windows will also come and go as the searches run (similar to the ms2 creation). After Comet searches are done, some post-processing of results will take place. The status message will let you know when everything is finished. Be patient, Comet searches and the post-processing can take several minutes to hours to complete depending on search parameters and database size/complexity.



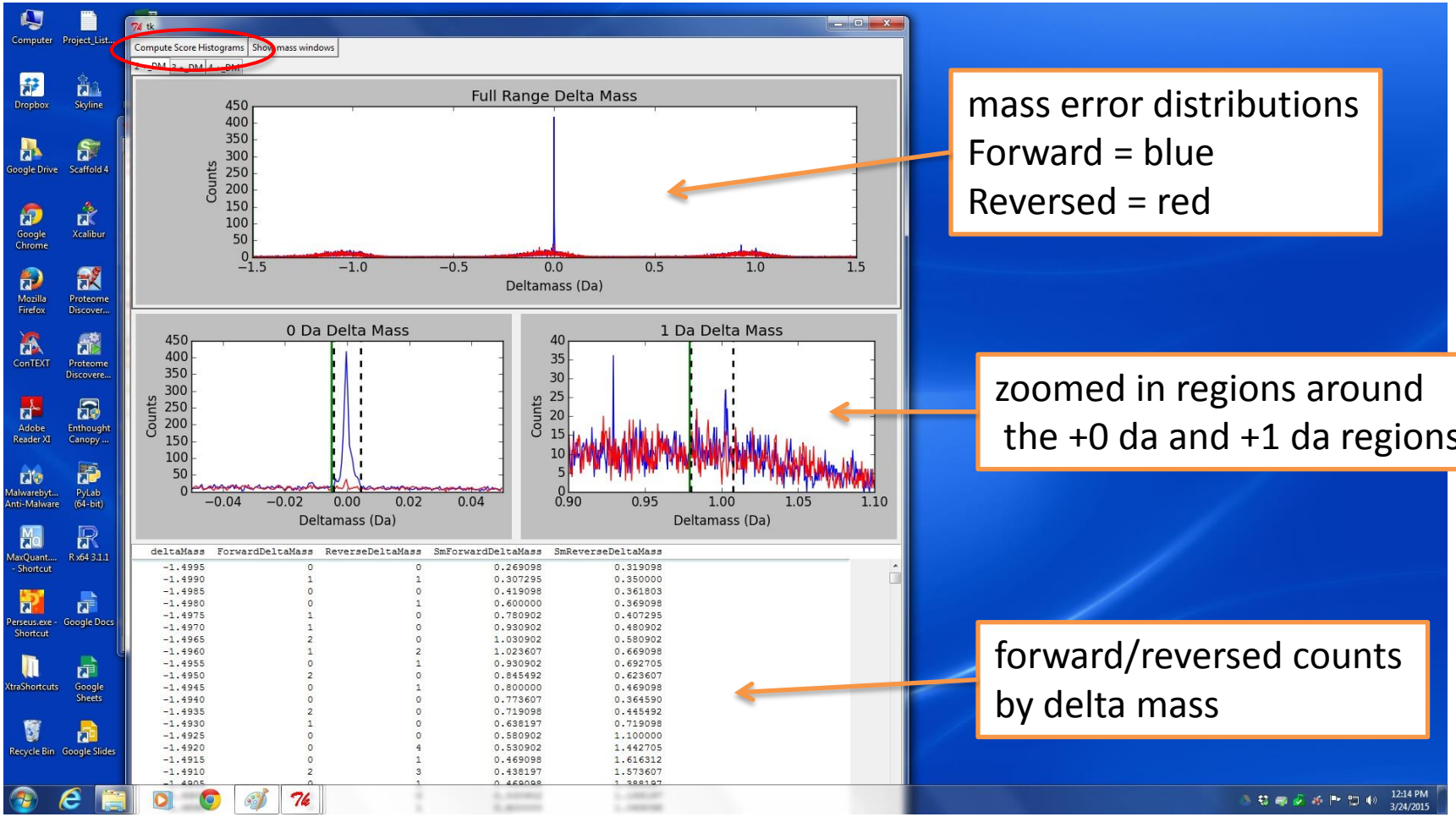
NOTE: The histogram viewing GUI program described next was written for SEQUEST searches (which are very similar to Comet searches) and requires a **sequest.params** file. A mock **sequest.params** file must be created at this point in time. Parsing of **comet.params** files is planned for the near future. Consult PSR staff for help with this step.

The **comet_GUI.py** program makes configuring and running basic Comet searches pretty easy. It just executes some Windows command line calls and runs a couple of other Python programs. These steps can also be done manually and are described in slides at the end of the main tutorial. This can be handy when repeating a search with different parameters, or if more advanced Comet parameters are needed.

The next pipeline step is **histo_GUI.py**. It starts the same way as the others, double-click to open the program, and then press **F5** to run. A small GUI window will open. You can use the top button to select the **ms2_files** folder you wish to analyze, and the bottom button is for toggling between high mass accuracy (Orbitrap) data, and low mass accuracy data (everything else).



The GUI will open after a short loading time, and look something like below. The first step is to set the mass windows. If you are using low mass accuracy data you can skip this step and simply press the **Compute Score Histograms** button to jump ahead to the next window.



mass error distributions
Forward = blue
Reversed = red

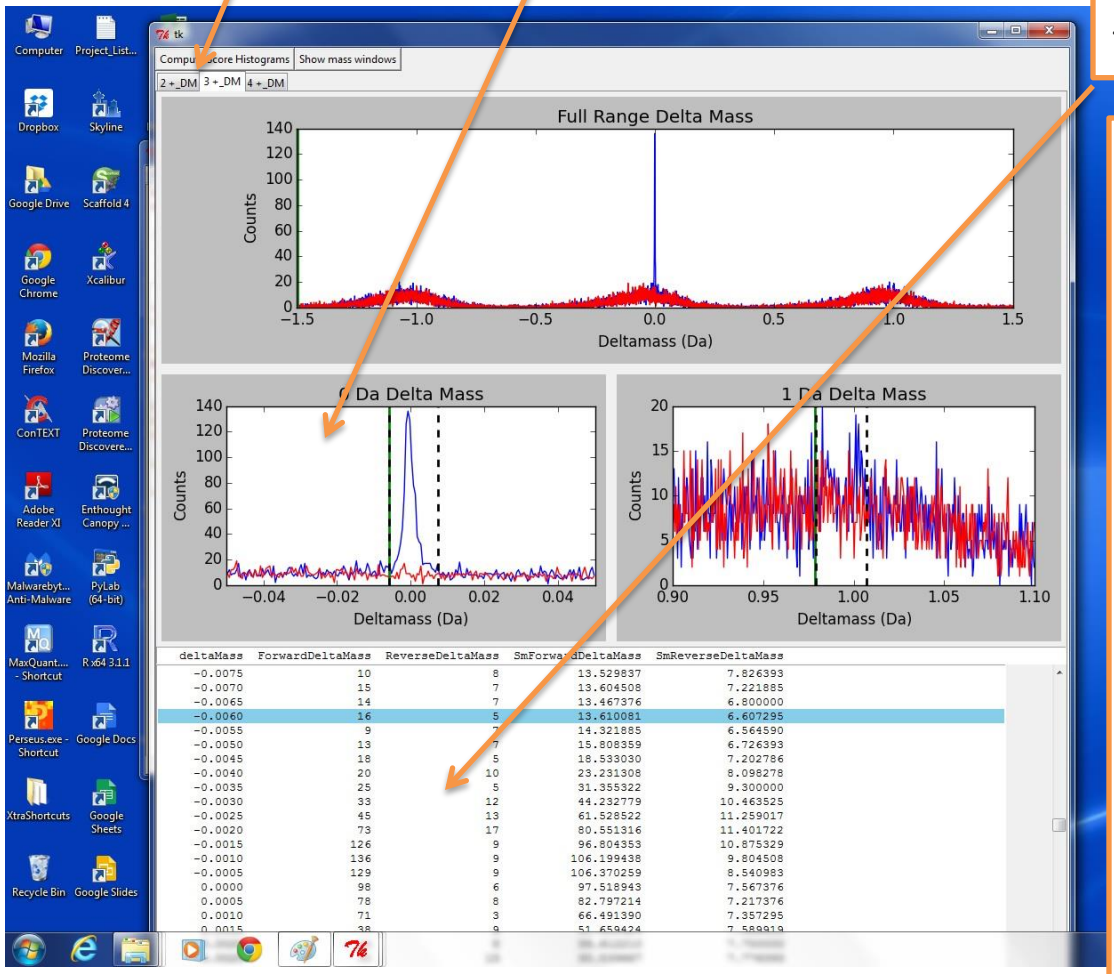
zoomed in regions around the +0 da and +1 da regions

forward/reversed counts by delta mass

The first step in selecting the mass windows is to click on the tab for the charge state you wish to set (thresholds should be checked/adjusted for all charge states).

Next click on the window you want to edit (0 Da or 1 Da).

Then click on the table at the bottom.

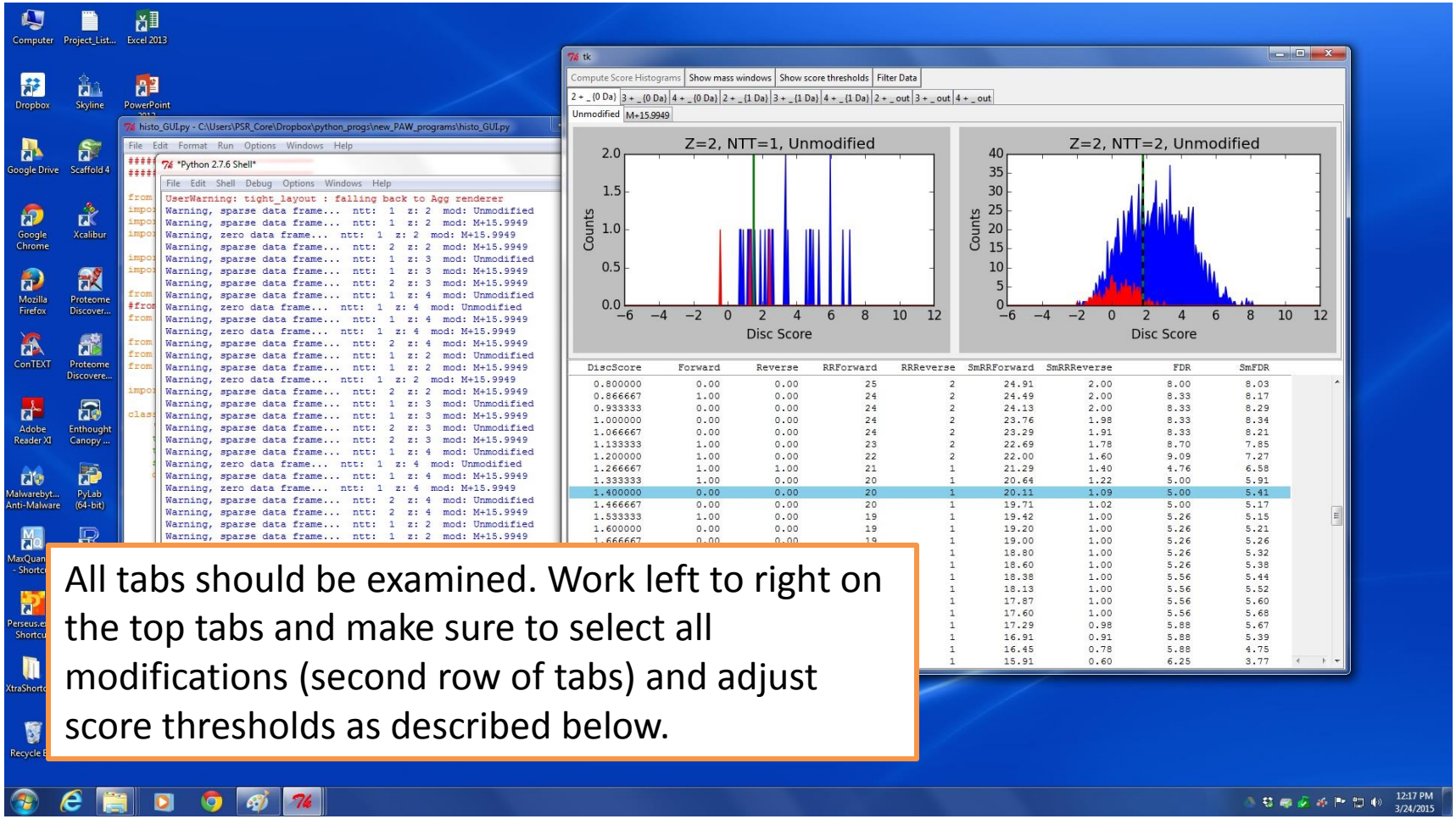


Once the table is selected, the left and right arrow keys toggle between the upper and lower thresholds; the up and down arrow keys adjust the threshold higher or lower.

The **enter** key updates the threshold to the current dotted line location. You want to capture just the blue peaks associated with the forward distributions between the thresholds. The default windows may be OK.

Once all the thresholds are set, click The **Compute Scores Histograms** button to create the search score histograms.

After the mass accuracy filtering, next is filtering by discriminate scores. The tabs and windows work in similar ways as the mass accuracy window, with the exception of there only being a single minimum threshold for each plot. There are two rows of tabs above the score histograms. The top are different charge states and delta mass windows. The bottom tabs are for different modification states. There will be multiple score histograms (plots) for different enzymatic cleavage states.

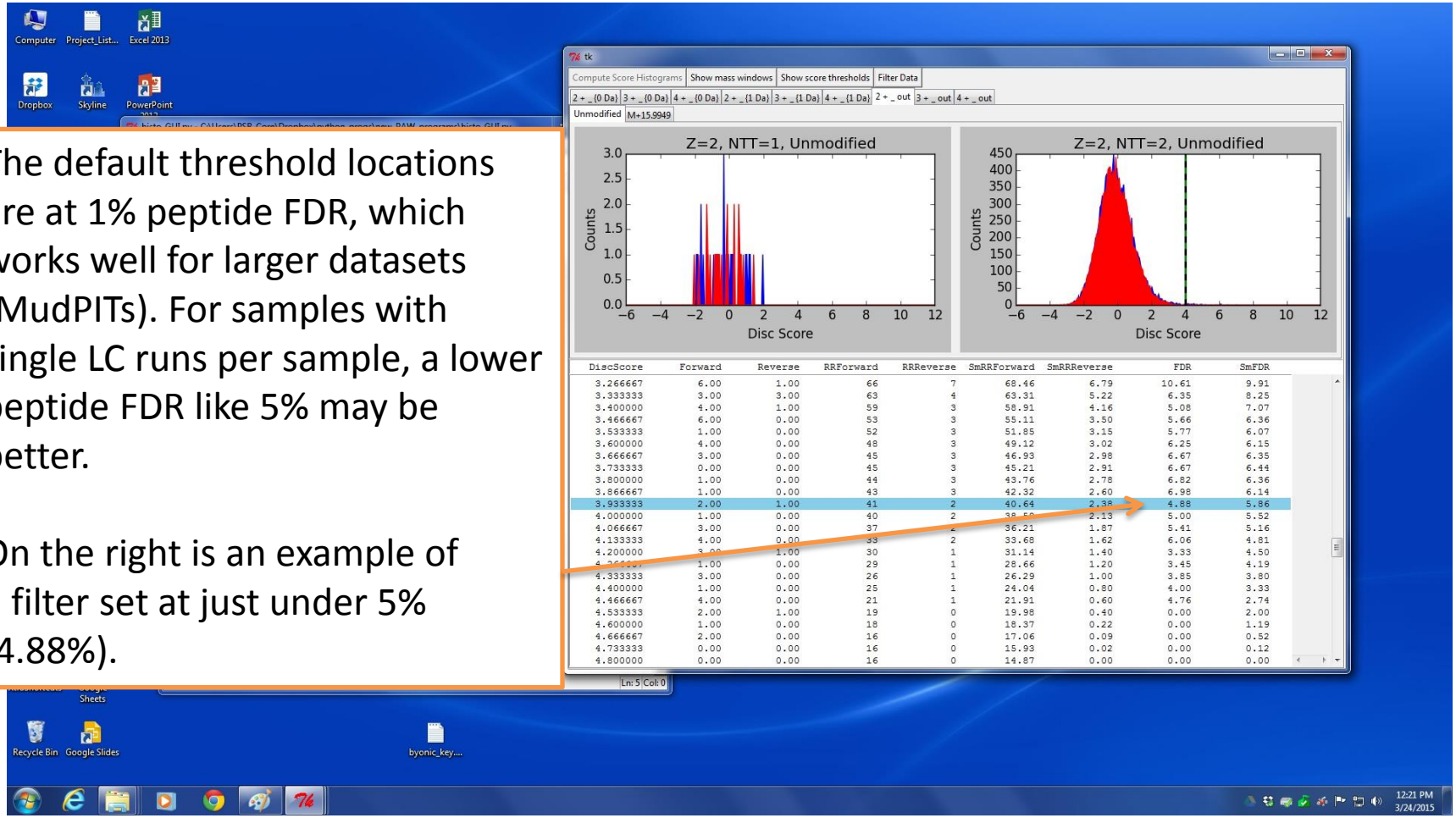


All tabs should be examined. Work left to right on the top tabs and make sure to select all modifications (second row of tabs) and adjust score thresholds as described below.

The goal when setting peptide thresholds is to end up with a reasonable final protein false discovery rate (FDR). How a peptide FDR translates to a protein-level FDR will vary depending on the size of the database and the complexity of your samples. What follows are some basic guidelines.

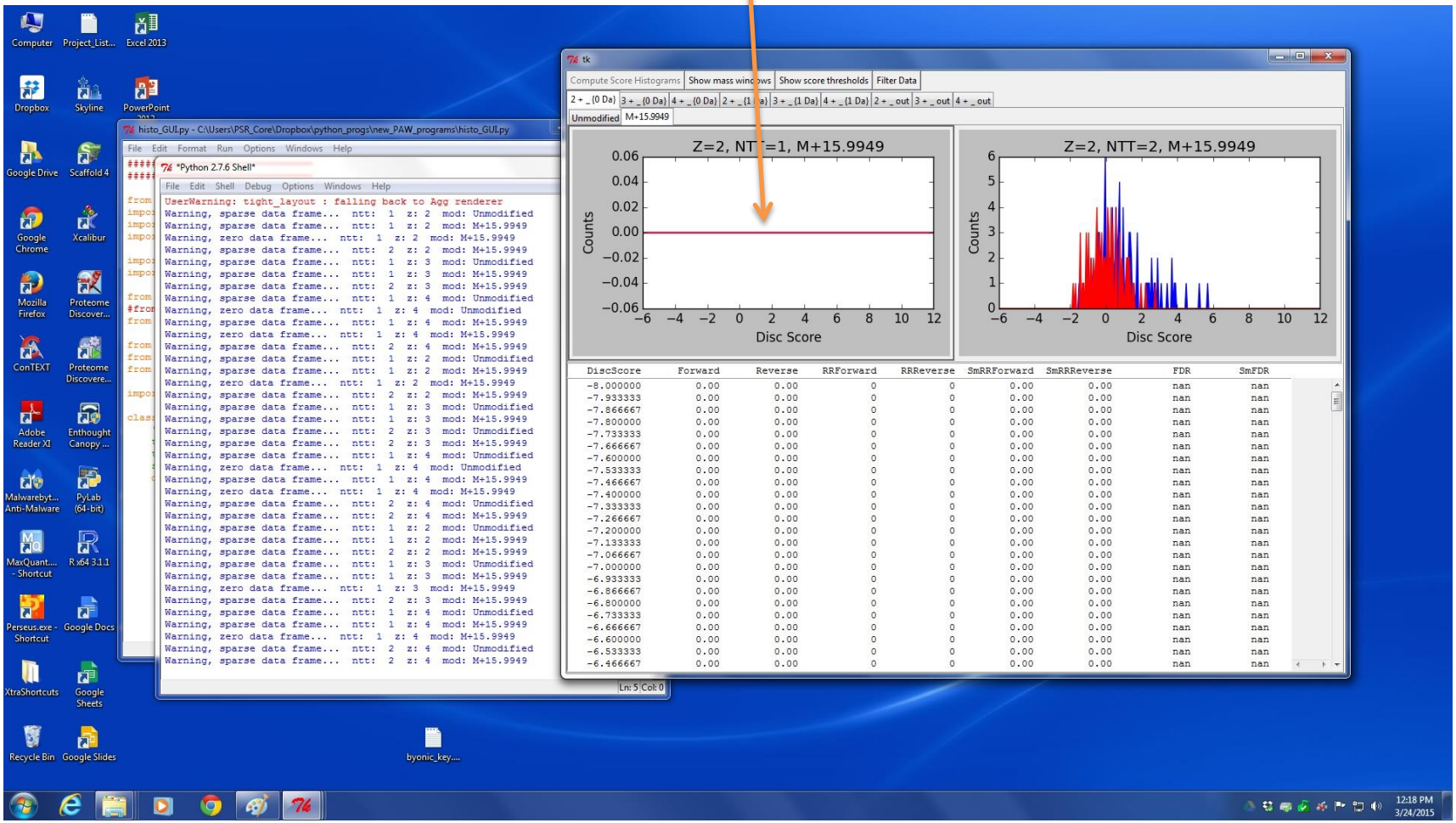
The default threshold locations are at 1% peptide FDR, which works well for larger datasets (MudPITs). For samples with single LC runs per sample, a lower peptide FDR like 5% may be better.

On the right is an example of a filter set at just under 5% (4.88%).



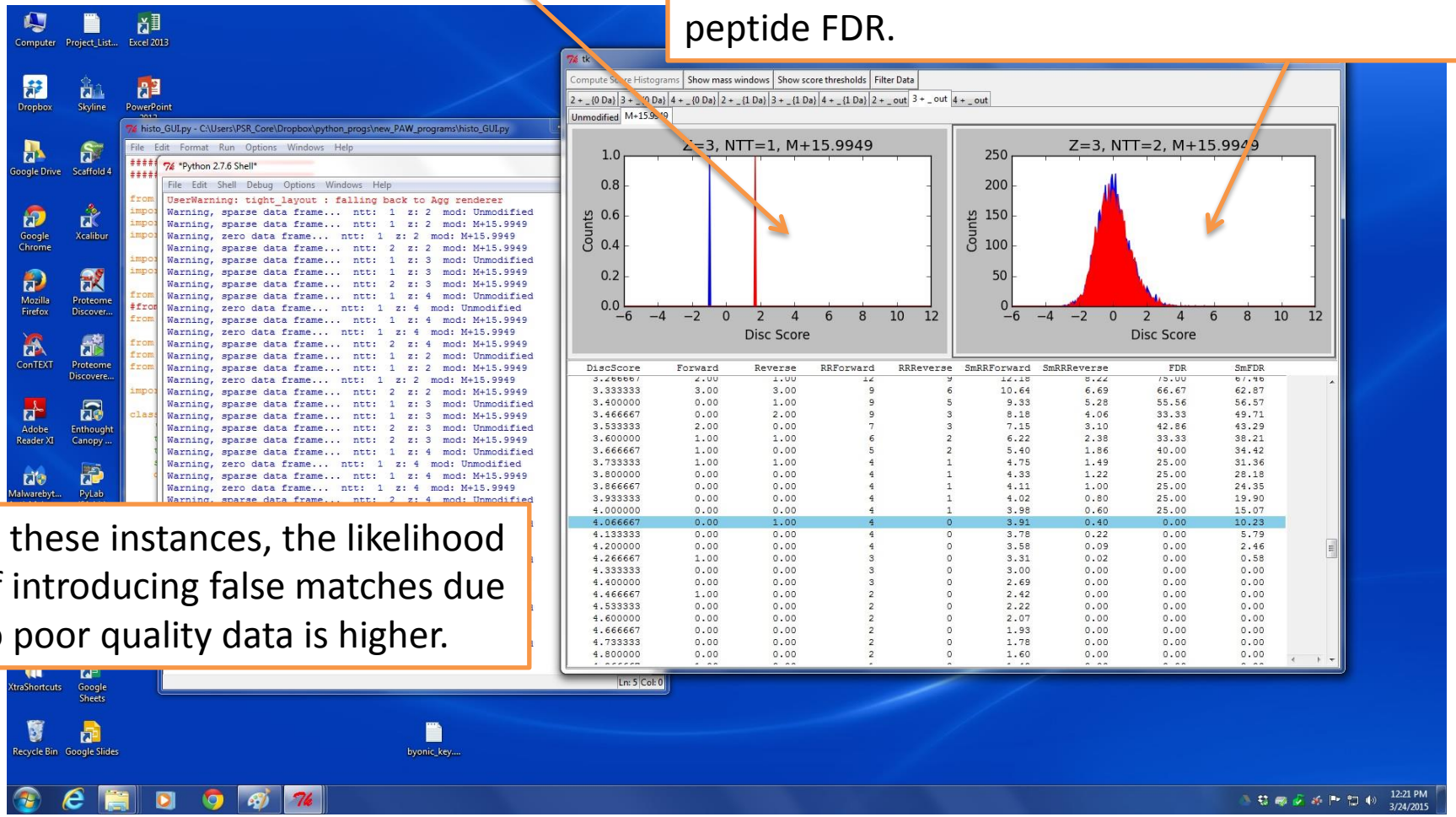
DiscScore	Forward	Reverse	RRForward	RRReverse	SmRRForward	SmRRReverse	FDR	SmFDR
3.266667	6.00	1.00	66	7	68.46	6.79	10.61	9.91
3.333333	3.00	3.00	63	4	63.31	5.22	6.35	8.25
3.400000	4.00	1.00	59	3	58.91	4.16	5.08	7.07
3.466667	6.00	0.00	53	3	55.11	3.50	5.66	6.36
3.533333	1.00	0.00	52	3	51.85	3.15	5.77	6.07
3.600000	4.00	0.00	48	3	49.12	3.02	6.25	6.15
3.666667	3.00	0.00	45	3	46.93	2.98	6.67	6.35
3.733333	0.00	0.00	45	3	45.21	2.91	6.67	6.44
3.800000	1.00	0.00	44	3	43.76	2.78	6.82	6.36
3.866667	1.00	0.00	43	3	42.32	2.60	6.98	6.14
3.933333	2.00	1.00	41	2	40.64	2.38	4.88	5.86
4.000000	1.00	0.00	40	2	38.58	2.13	5.00	5.52
4.066667	3.00	0.00	37	2	36.21	1.87	5.41	5.16
4.133333	4.00	0.00	33	2	33.68	1.62	6.06	4.81
4.200000	3.00	1.00	30	1	31.14	1.40	3.33	4.50
4.266667	1.00	0.00	29	1	28.66	1.20	3.45	4.19
4.333333	3.00	0.00	26	1	26.29	1.00	3.85	3.80
4.400000	1.00	0.00	25	1	24.04	0.80	4.00	3.33
4.466667	4.00	0.00	21	1	21.91	0.60	4.76	2.74
4.533333	2.00	1.00	19	0	19.98	0.40	0.00	2.00
4.600000	1.00	0.00	18	0	18.37	0.22	0.00	1.19
4.666667	2.00	0.00	16	0	17.06	0.09	0.00	0.52
4.733333	0.00	0.00	16	0	15.93	0.02	0.00	0.12
4.800000	0.00	0.00	16	0	14.87	0.00	0.00	0.00

To accurately control peptide sequence errors, peptides are divided into many classes depending on accurate mass, search parameters, and charge states. In many datasets, several of these classes may be empty or have very few estimated correct identifications. There will be no dotted threshold lines for these classes.



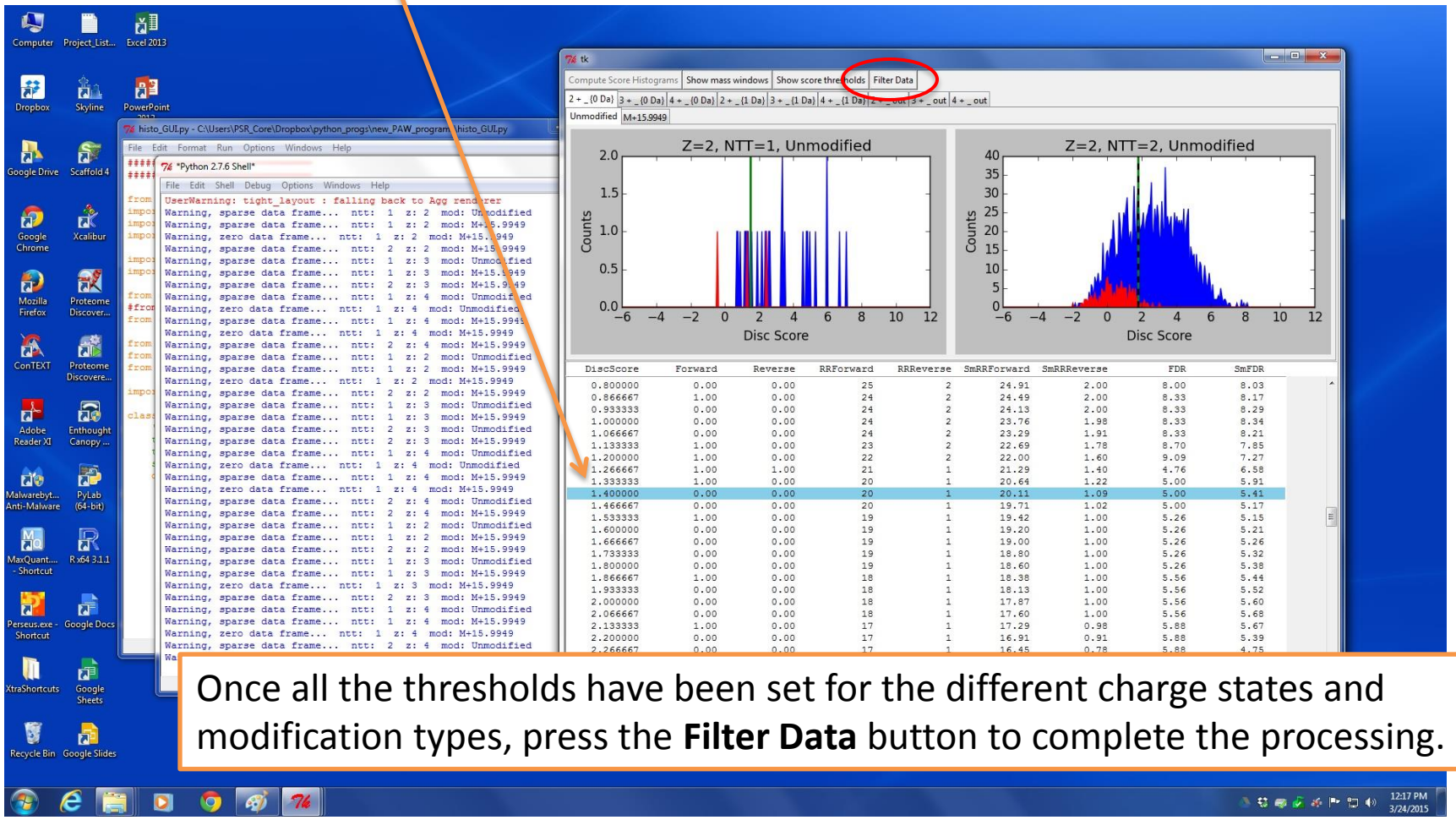
In addition to empty peptide classes, you can ignore sparse (nearly empty) classes such as the one on the left.

It is also a good idea to exclude peptide classes with few forward matches above the desired peptide FDR.



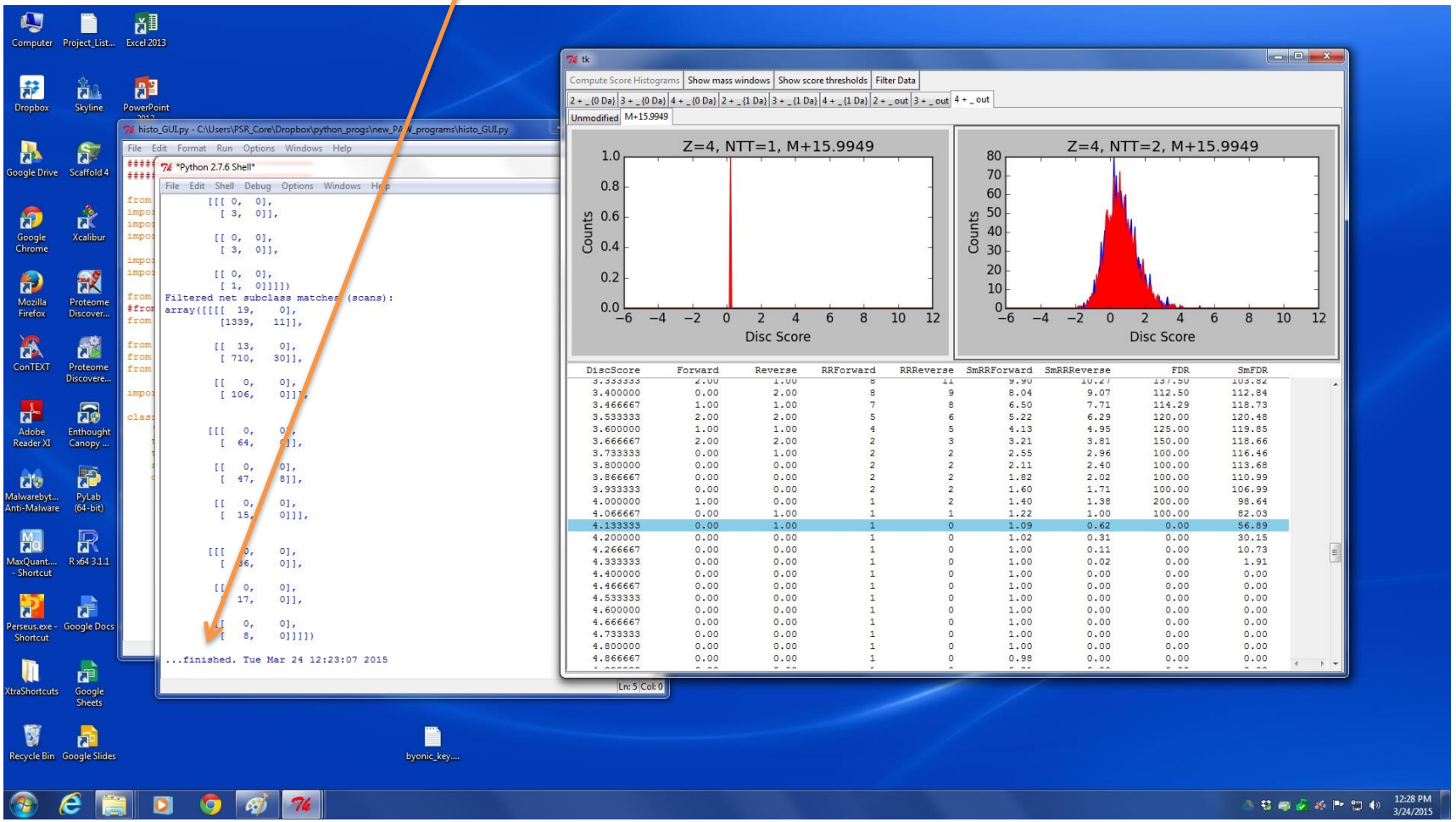
In these instances, the likelihood of introducing false matches due to poor quality data is higher.

Finally, setting a minimum threshold for the discriminate score (like the 1.4 that is used below) is also a good idea. These very low scores would likely fail manual inspection, and wouldn't meet accepted proteomics publication standards.

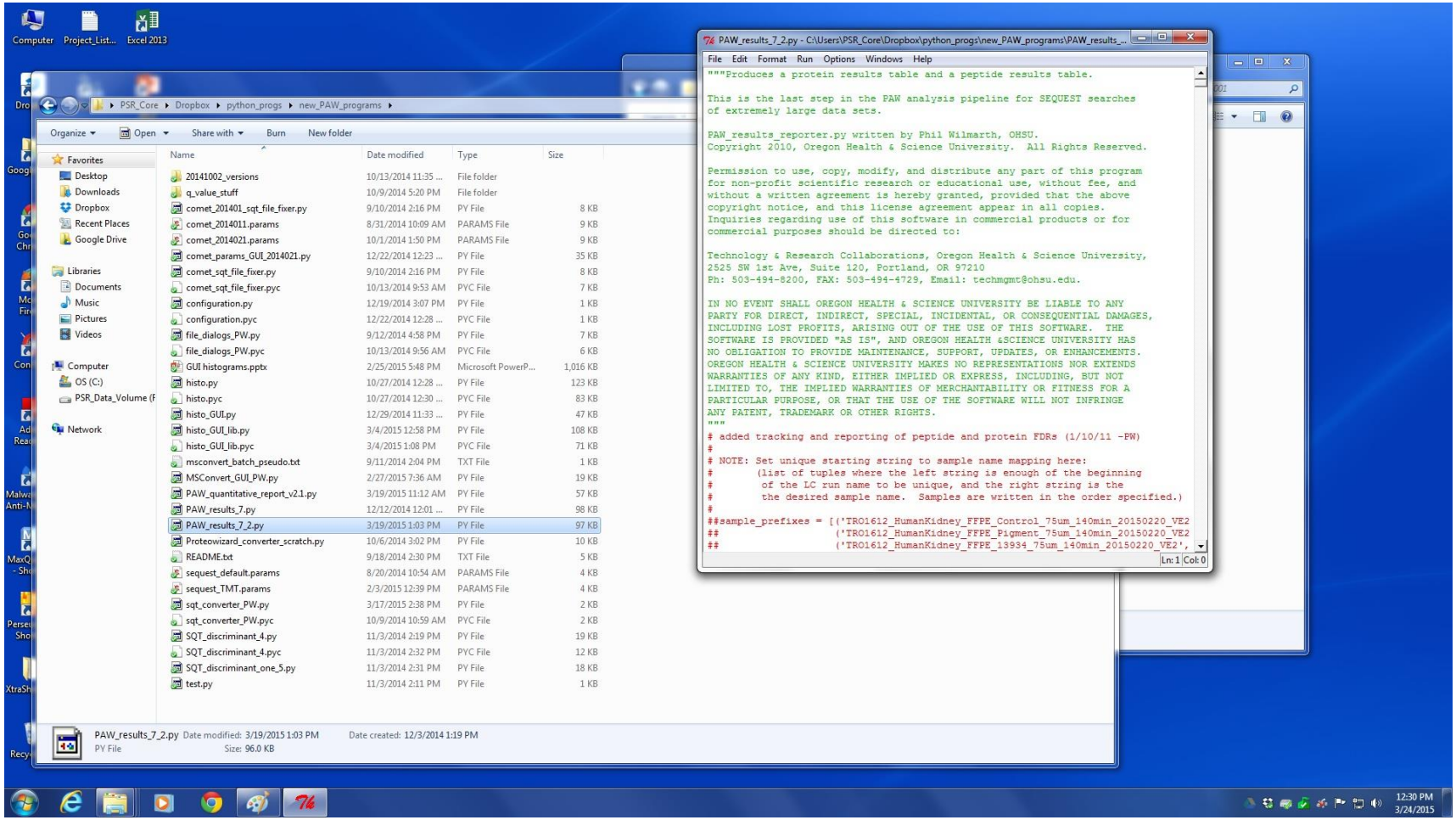


Once all the thresholds have been set for the different charge states and modification types, press the **Filter Data** button to complete the processing.

Filtering will usually take about 20-30 seconds per file (or longer) to process and write out the filtered files. When it is finished, the python shell will display **…finished**. You can close the GUI and python windows to exit.



The final step in the analysis pipeline in the **PAW_results_7.py** program (note: the number at the end may change with new versions). Unlike with previous programs you'll need to edit the source code after opening the program file. If you are unfamiliar with Python coding, consult with the PSR staff for assistance.

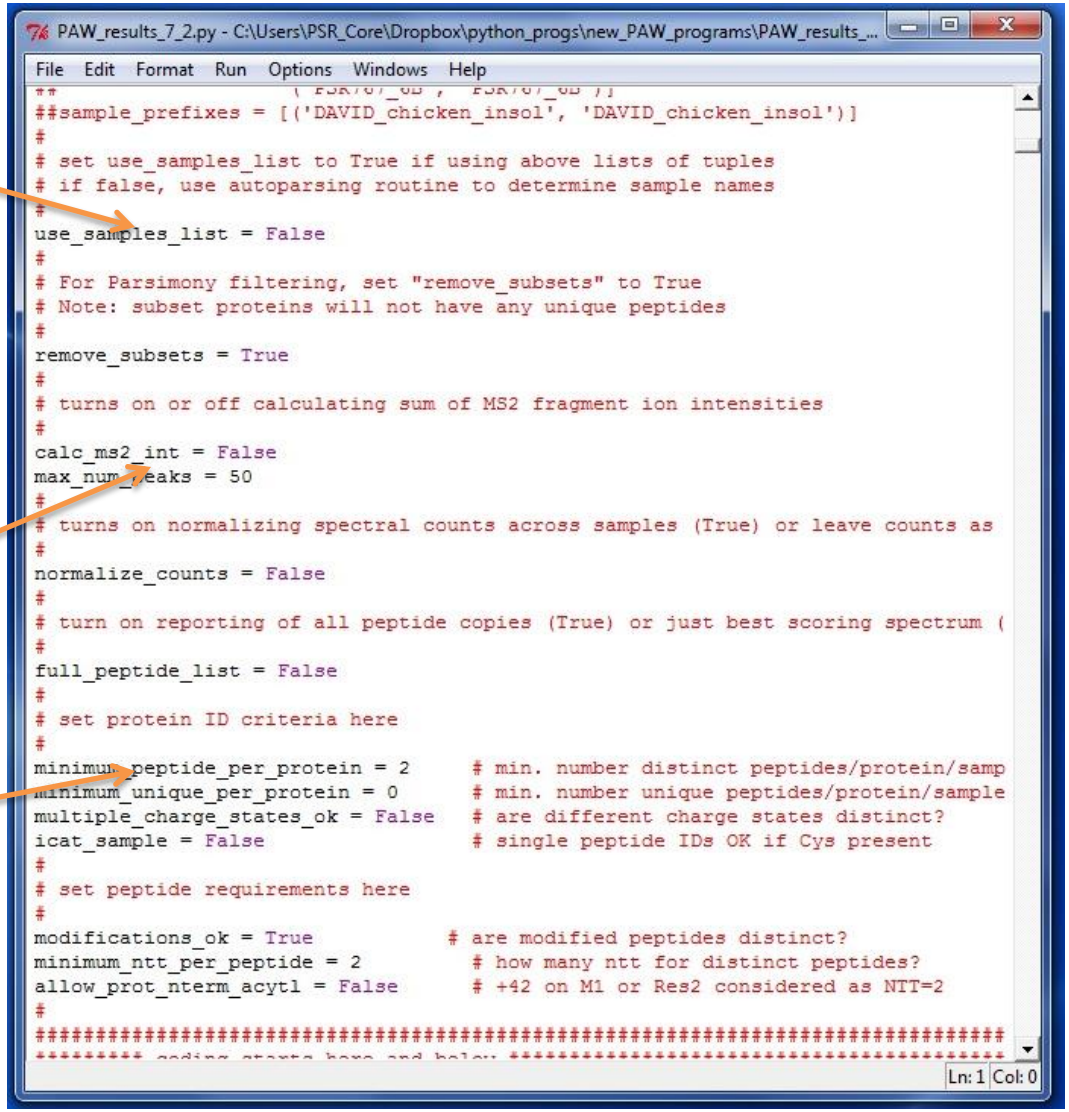


Below the initial list of sample names (which will be covered momentarily), there's a couple of potential flags that are commonly changed and should be checked and set properly before running the program.

The program will automatically try and determine how to combine the different files into sample sets. If you need to bypass that logic and define your sample names manually, set this to **True**.

This should be set to **False** unless you're specifically doing an ms2 intensity quantitative project.

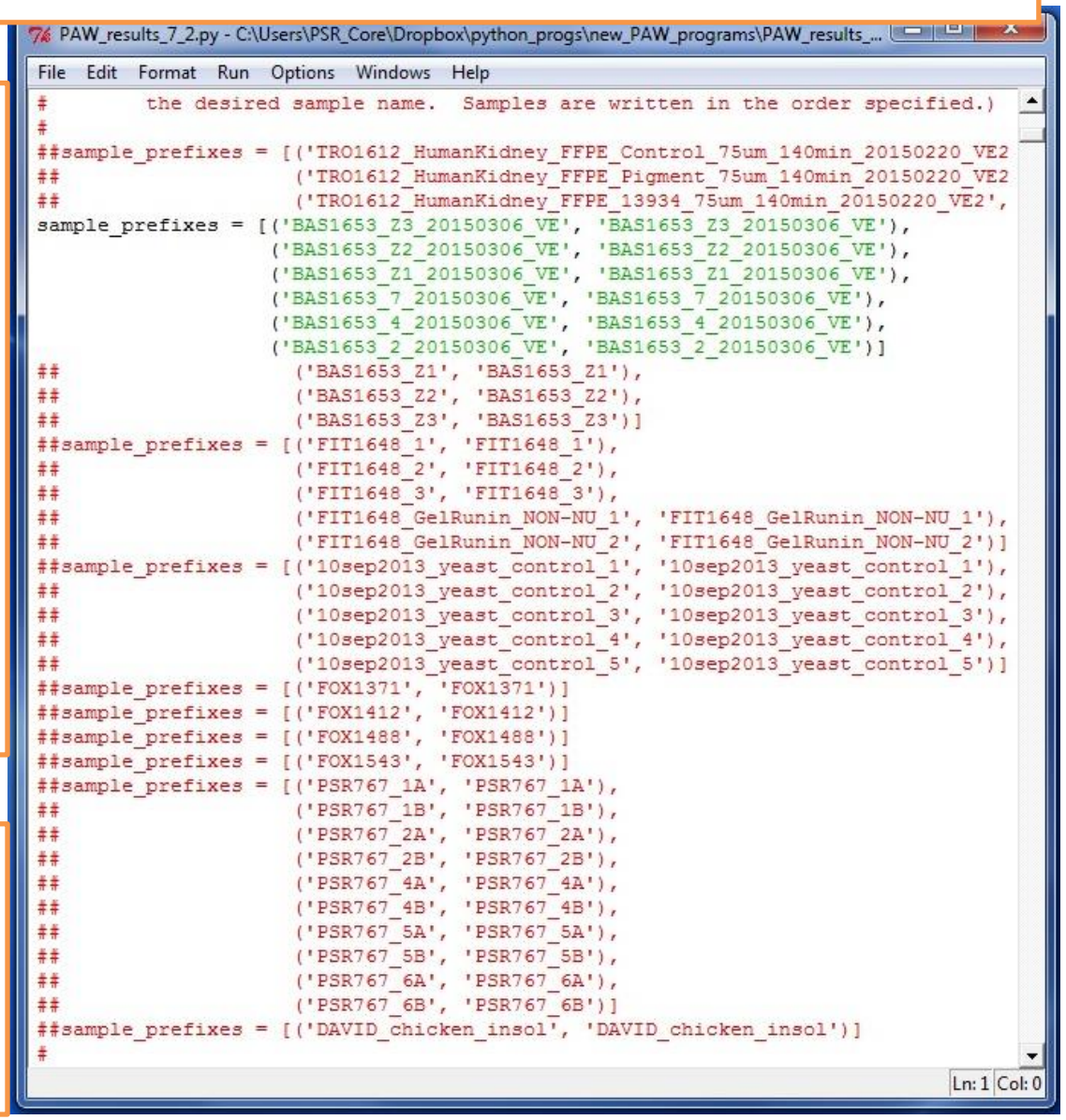
This should be set to **2** (peptides per protein) for most analyses.



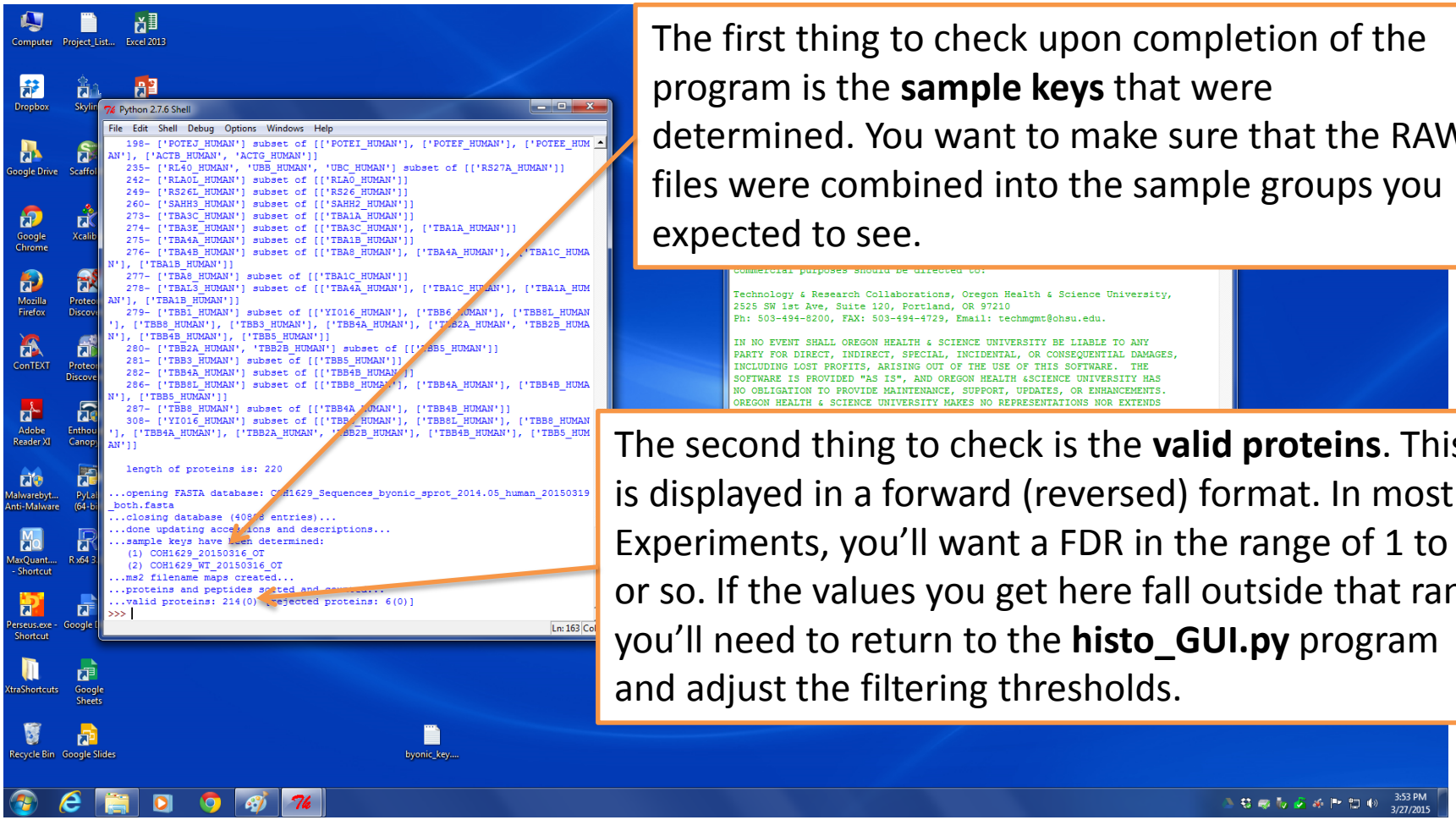
The program's logic for combining different instrument runs can be overridden by setting **use_samples_list** to **True**. In this case, you'll have to define how to combine the runs into samples.

The **sample_prefixes** section of the code does this. Each line is a pairing. The first item is the beginning string the program will look for. Every file name that starts with what is listed in the first set of quotes (only enough to be unique is required) will be combined into a single column in the results spreadsheet with the column name given by the right item string.

Once everything is ready to go press **F5** to run the program as usual, and press **ok** if the program asks you if you'd like to save any changes.



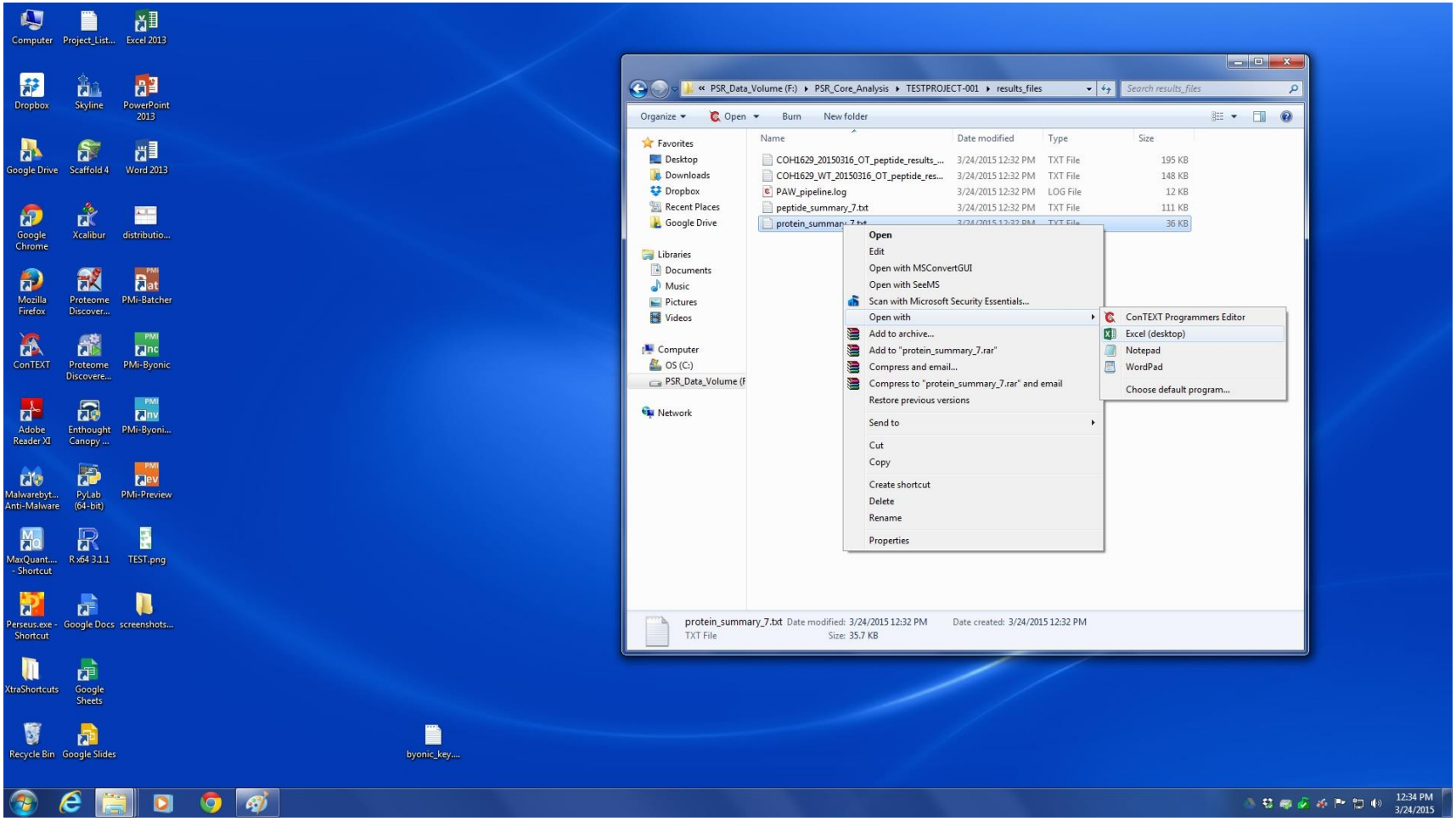
After navigating to the **filtered_ms2_files** folder you wish to analyze, the program will churn through the list of potential proteins and create the final results spreadsheets, as well as providing you with two important pieces of information.



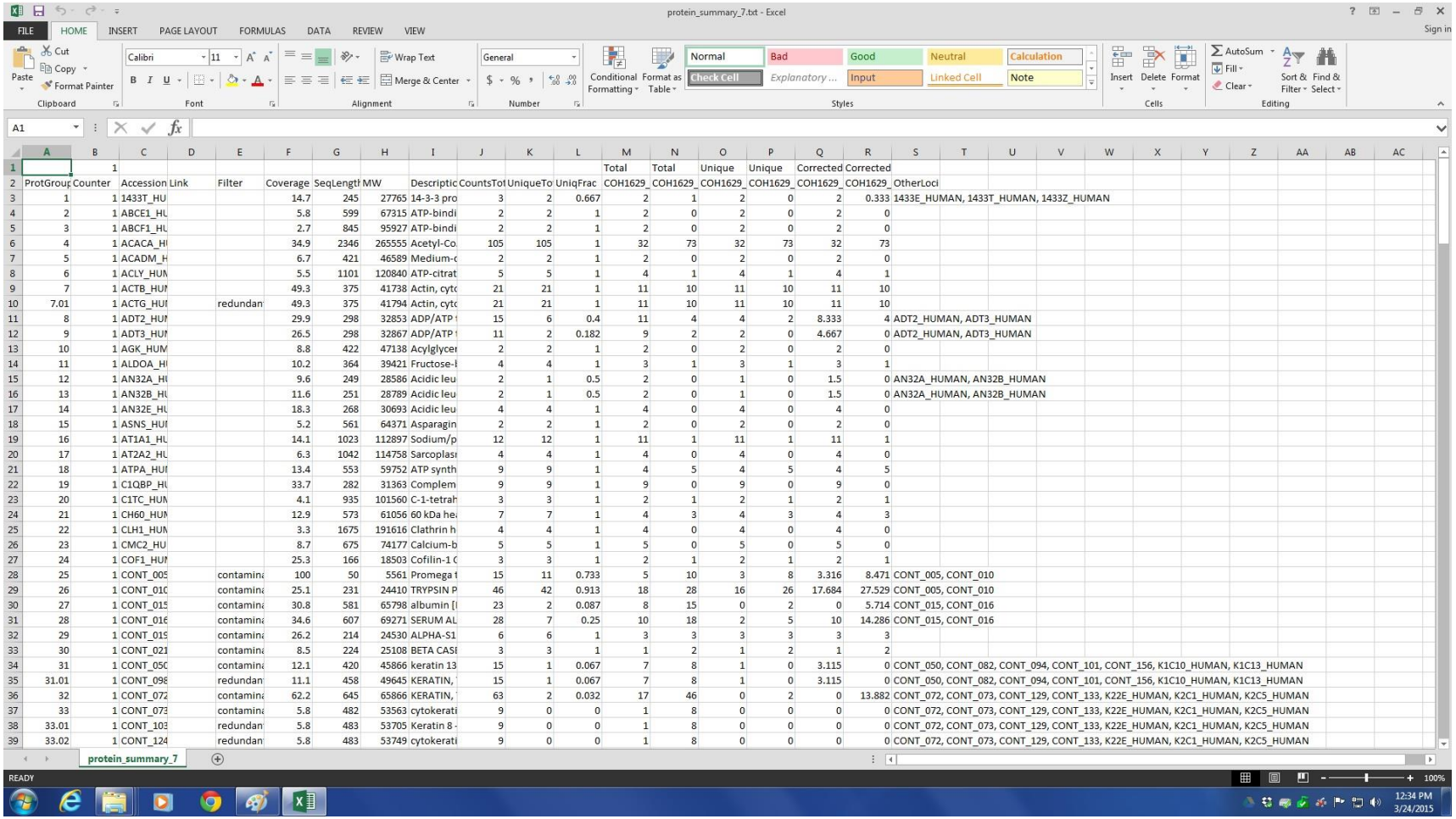
The first thing to check upon completion of the program is the **sample keys** that were determined. You want to make sure that the RAW files were combined into the sample groups you expected to see.

The second thing to check is the **valid proteins**. This is displayed in a forward (reversed) format. In most Experiments, you'll want a FDR in the range of 1 to 5% or so. If the values you get here fall outside that range, you'll need to return to the **histo_GUI.py** program and adjust the filtering thresholds.

The final output from the pipeline will be a **protein_summary_7.txt** file and a **peptide_summary_7.txt** file. They can be found in the **results_files** folder that's created by the **PAW_Results_7.py** program. To view these files, right click on them and choose to open them in Microsoft Excel (or OpenOffice).



The **protein_summary** file will contain the proteins identified, and the **peptide_summary** file will list the identified amino acid sequences. Information regarding the formatting and layout of the spreadsheet can be found in the “PAW Results Guide” in the Educational Links section of the PSR Website.



protein_summary_7.txt - Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1		1											Total	Total	Unique	Unique	Corrected	Corrected											
2	ProtGroup	Counter	Accession	Link	Filter	Coverage	SeqLength	MW	Descriptic	CountsTot	UniqueTo	UniqFrac	COH1629_	COH1629_	COH1629_	COH1629_	COH1629_	COH1629_	OtherLoci										
3	1	1	1433T_HU			14.7	245	27765	14-3-3 pro	3	2	0.667	2	1	2	0	2	0.333	1433E_HUMAN, 1433T_HUMAN, 1433Z_HUMAN										
4	2	1	ABCE1_HU			5.8	599	67315	ATP-bindi	2	2	1	2	0	2	0	2	0											
5	3	1	ABCF1_HU			2.7	845	95927	ATP-bindi	2	2	1	2	0	2	0	2	0											
6	4	1	ACACA_HI			34.9	2346	265555	Acetyl-Co	105	105	1	32	73	32	73	32	73											
7	5	1	ACADM_H			6.7	421	46589	Medium-c	2	2	1	2	0	2	0	2	0											
8	6	1	ACLY_HUM			5.5	1101	120840	ATP-citrat	5	5	1	4	1	4	1	4	1											
9	7	1	ACTB_HUM			49.3	375	41738	Actin, cyto	21	21	1	11	10	11	10	11	10											
10	7.01	1	ACTG_HUI		redundan	49.3	375	41794	Actin, cyto	21	21	1	11	10	11	10	11	10											
11	8	1	ADT2_HUI			29.9	298	32853	ADP/ATP t	15	6	0.4	11	4	4	2	8.333	4	ADT2_HUMAN, ADT3_HUMAN										
12	9	1	ADT3_HUI			26.5	298	32867	ADP/ATP t	11	2	0.182	9	2	2	0	4.667	0	ADT2_HUMAN, ADT3_HUMAN										
13	10	1	AGK_HUM			8.8	422	47138	Acylglycer	2	2	1	2	0	2	0	2	0											
14	11	1	ALDOA_HI			10.2	364	39421	Fructose-I	4	4	1	3	1	3	1	3	1											
15	12	1	AN32A_HU			9.6	249	28586	Acidic leu	2	1	0.5	2	0	1	0	1.5	0	AN32A_HUMAN, AN32B_HUMAN										
16	13	1	AN32B_HU			11.6	251	28789	Acidic leu	2	1	0.5	2	0	1	0	1.5	0	AN32A_HUMAN, AN32B_HUMAN										
17	14	1	AN32E_HU			18.3	268	30693	Acidic leu	4	4	1	4	0	4	0	4	0											
18	15	1	ASNS_HUI			5.2	561	64371	Asparagin	2	2	1	2	0	2	0	2	0											
19	16	1	AT1A1_HU			14.1	1023	112897	Sodium/p	12	12	1	11	1	11	1	11	1											
20	17	1	AT2A2_HU			6.3	1042	114758	Sarcoplas	4	4	1	4	0	4	0	4	0											
21	18	1	ATPA_HUI			13.4	553	59752	ATP synth	9	9	1	4	5	4	5	4	5											
22	19	1	C1QBP_HU			33.7	282	31363	Complem	9	9	1	9	0	9	0	9	0											
23	20	1	C1TC_HUM			4.1	935	101560	C-1-tetrah	3	3	1	2	1	2	1	2	1											
24	21	1	CH60_HUM			12.9	573	61056	60 kDa hea	7	7	1	4	3	4	3	4	3											
25	22	1	CLH1_HUM			3.3	1675	191616	Clathrin h	4	4	1	4	0	4	0	4	0											
26	23	1	CMC2_HU			8.7	675	74177	Calcium-b	5	5	1	5	0	5	0	5	0											
27	24	1	COF1_HUM			25.3	166	18503	Cofilin-1 (	3	3	1	2	1	2	1	2	1											
28	25	1	CONT_005		contamina	100	50	5561	Promega t	15	11	0.733	5	10	3	8	3.316	8.471	CONT_005, CONT_010										
29	26	1	CONT_010		contamina	25.1	231	24410	TRYPSIN P	46	42	0.913	18	28	16	26	17.684	27.529	CONT_005, CONT_010										
30	27	1	CONT_015		contamina	30.8	581	65798	albumin [I	23	2	0.087	8	15	0	2	0	5.714	CONT_015, CONT_016										
31	28	1	CONT_016		contamina	34.6	607	69271	SERUM AL	28	7	0.25	10	18	2	5	10	14.286	CONT_015, CONT_016										
32	29	1	CONT_019		contamina	26.2	214	24530	ALPHA-S1	6	6	1	3	3	3	3	3	3											
33	30	1	CONT_021		contamina	8.5	224	25108	BETA CASI	3	3	1	1	2	1	2	1	2											
34	31	1	CONT_050		contamina	12.1	420	45866	keratin 13	15	1	0.067	7	8	1	0	3.115	0	CONT_050, CONT_082, CONT_094, CONT_101, CONT_156, K1C10_HUMAN, K1C13_HUMAN										
35	31.01	1	CONT_098		redundan	11.1	458	49645	KERATIN,	15	1	0.067	7	8	1	0	3.115	0	CONT_050, CONT_082, CONT_094, CONT_101, CONT_156, K1C10_HUMAN, K1C13_HUMAN										
36	32	1	CONT_072		contamina	62.2	645	65866	KERATIN,	63	2	0.032	17	46	0	2	0	13.882	CONT_072, CONT_073, CONT_129, CONT_133, K22E_HUMAN, K2C1_HUMAN, K2C5_HUMAN										
37	33	1	CONT_073		contamina	5.8	482	53563	cytokerati	9	0	0	1	8	0	0	0	0	CONT_072, CONT_073, CONT_129, CONT_133, K22E_HUMAN, K2C1_HUMAN, K2C5_HUMAN										
38	33.01	1	CONT_103		redundan	5.8	483	53705	Keratin 8 -	9	0	0	1	8	0	0	0	0	CONT_072, CONT_073, CONT_129, CONT_133, K22E_HUMAN, K2C1_HUMAN, K2C5_HUMAN										
39	33.02	1	CONT_124		redundan	5.8	483	53749	cytokerati	9	0	0	1	8	0	0	0	0	CONT_072, CONT_073, CONT_129, CONT_133, K22E_HUMAN, K2C1_HUMAN, K2C5_HUMAN										

protein_summary_7

READY

12:34 PM
3/24/2015

The End

The following slides describe running Comet searches via command line and performing the comet results post-processing steps manually.

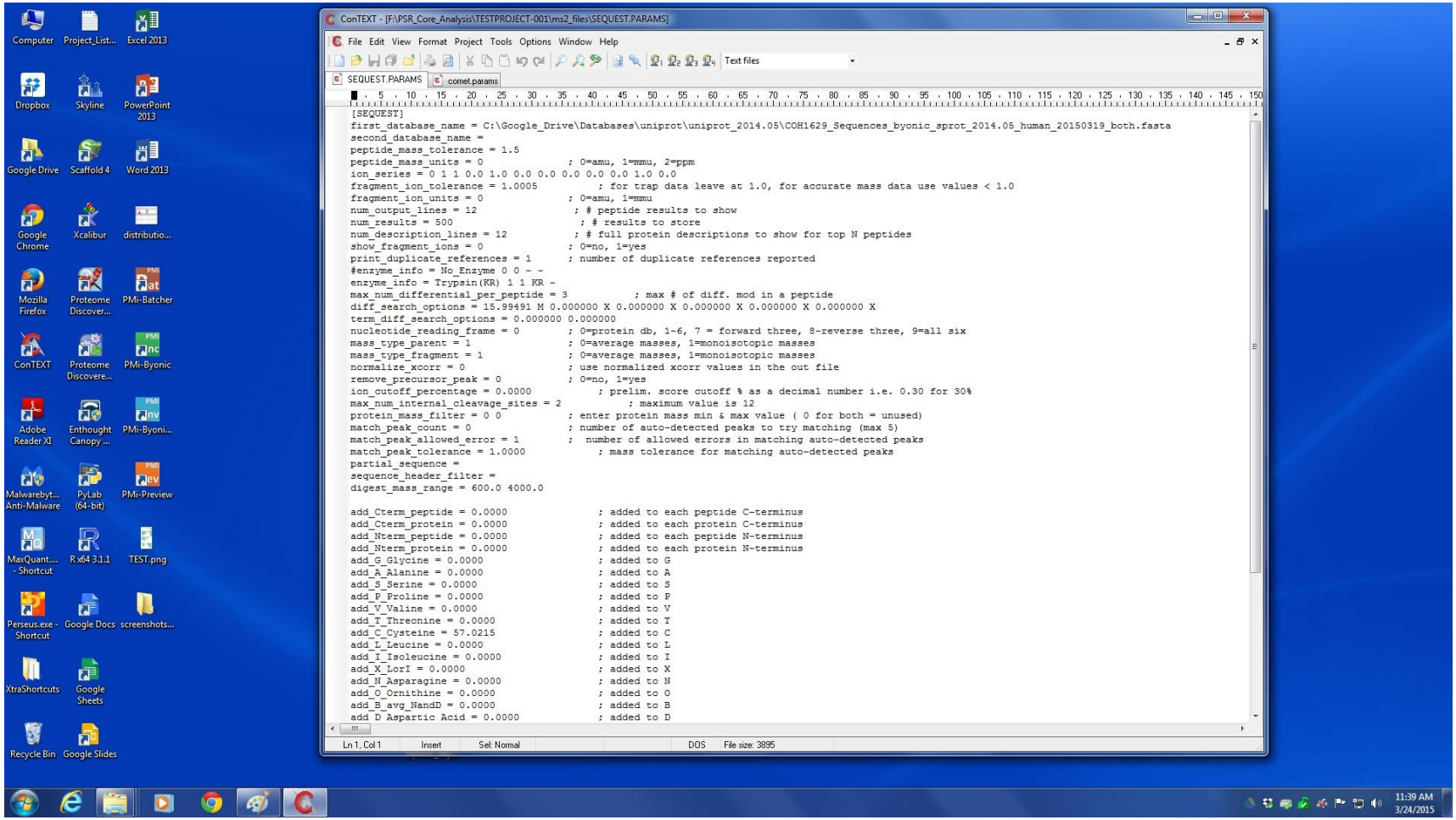
Situations where this might be useful are:

Doing a search using the same parameters as in a previous analysis where the **comet.params** file can be copied to the new **ms2_files** location.

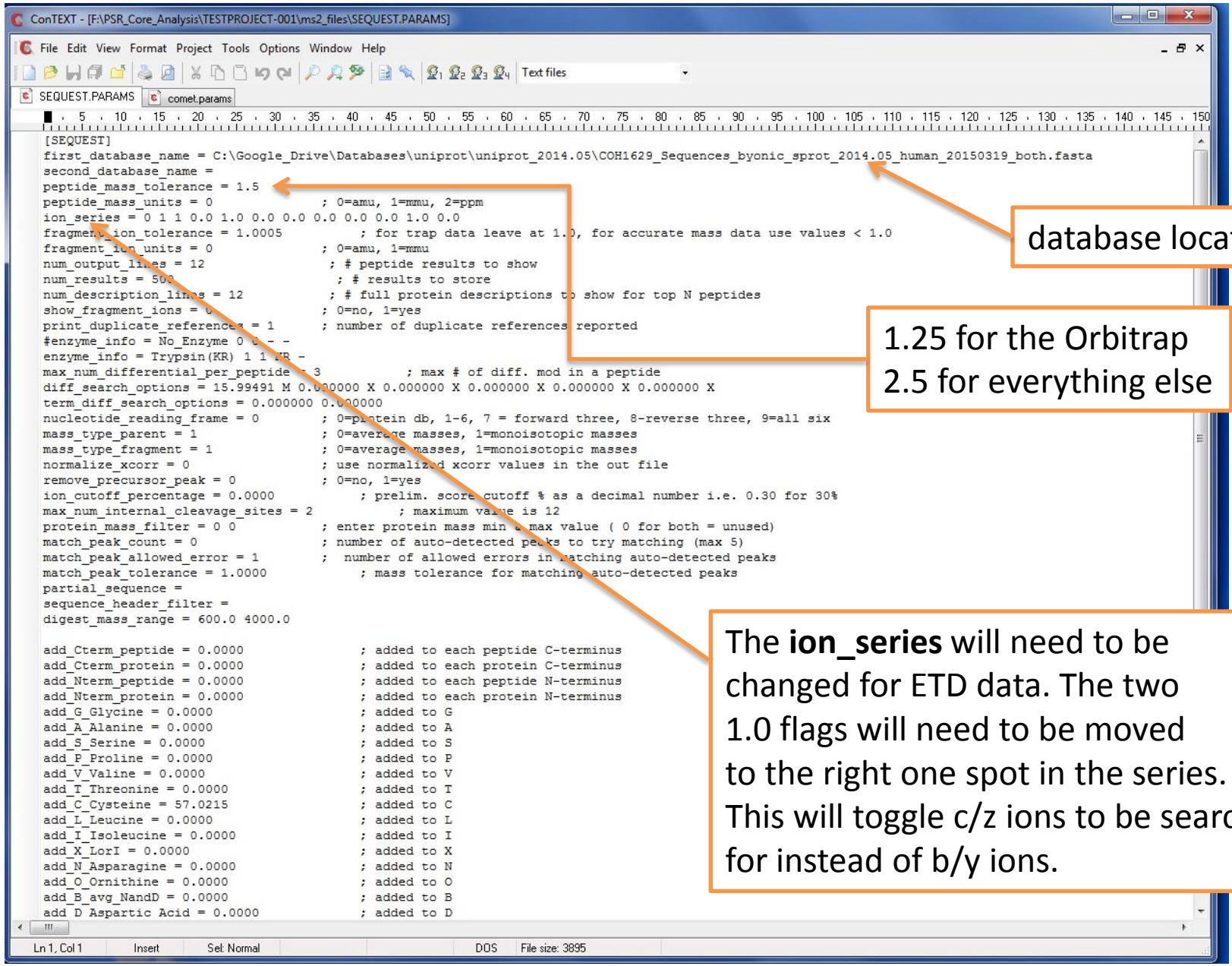
Making changes to the **comet.params** file that are beyond the basic parameters available in the GUI program.

NOTES: sequest.params files are discussed. The need for these files may change in the near future. Also, **comet.params** files can change format with different versions of Comet and contain a version key. If the version key does not match the Comet version, Comet will issue an error message and terminate.

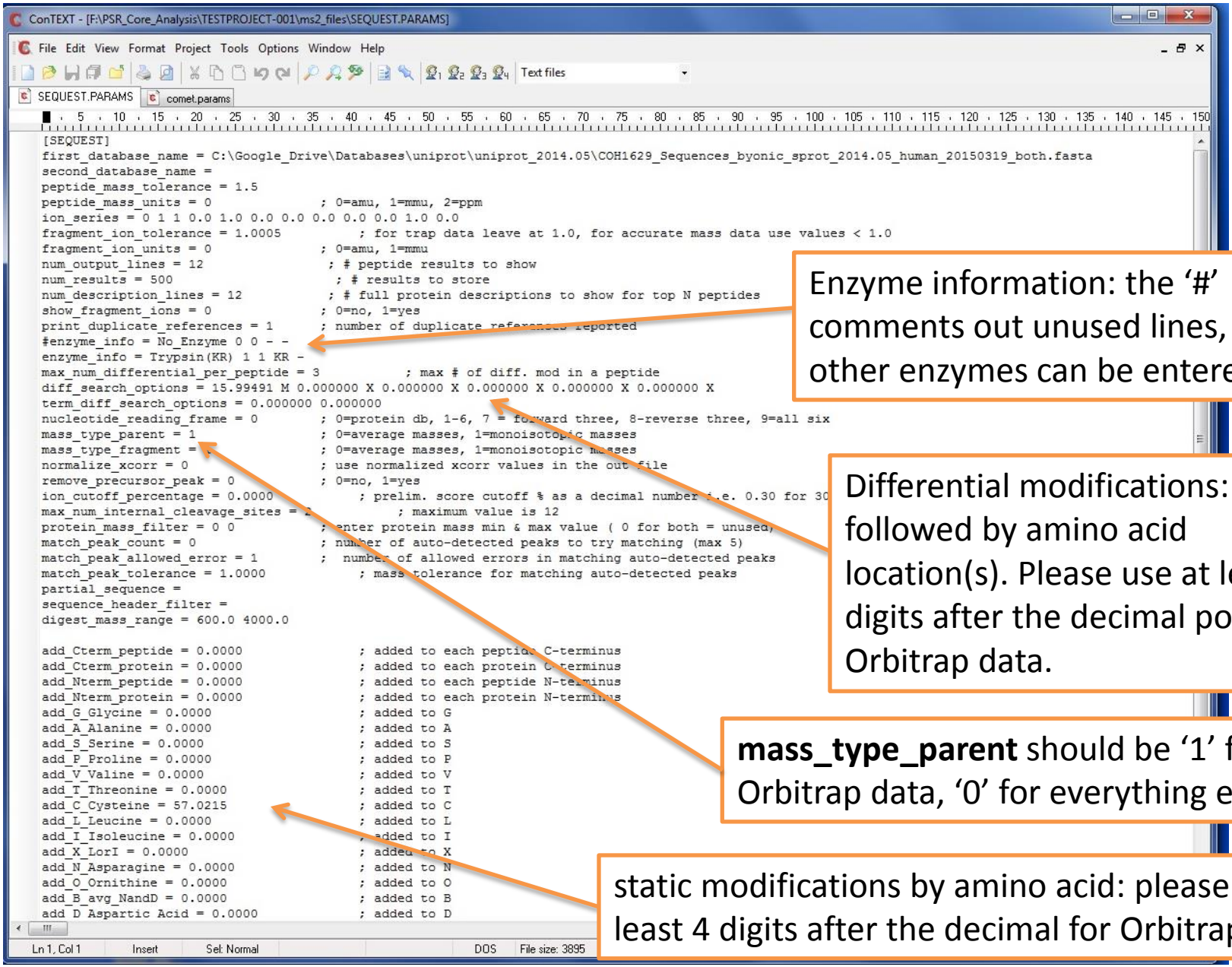
Once the parameter files have been copied over, they can be opened in any text editor. There are a few different places you'll likely need to make changes: database, fragment and parent ion mass tolerances, parent ion type, differential and static modifications, enzyme info, and ion series.



SEQUEST Parameters



SEQUEST Parameters



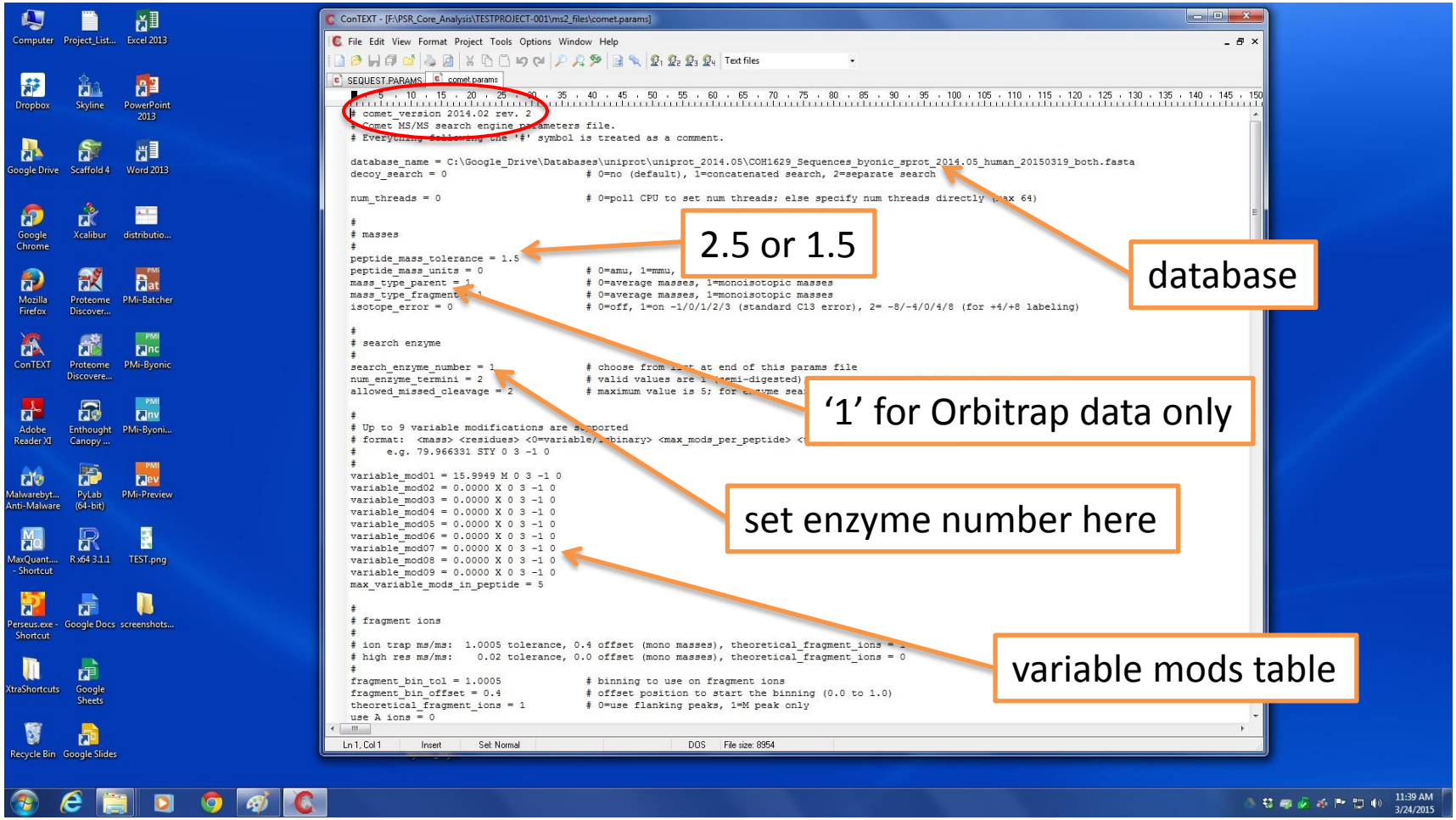
Enzyme information: the '#' comments out unused lines, and other enzymes can be entered here.

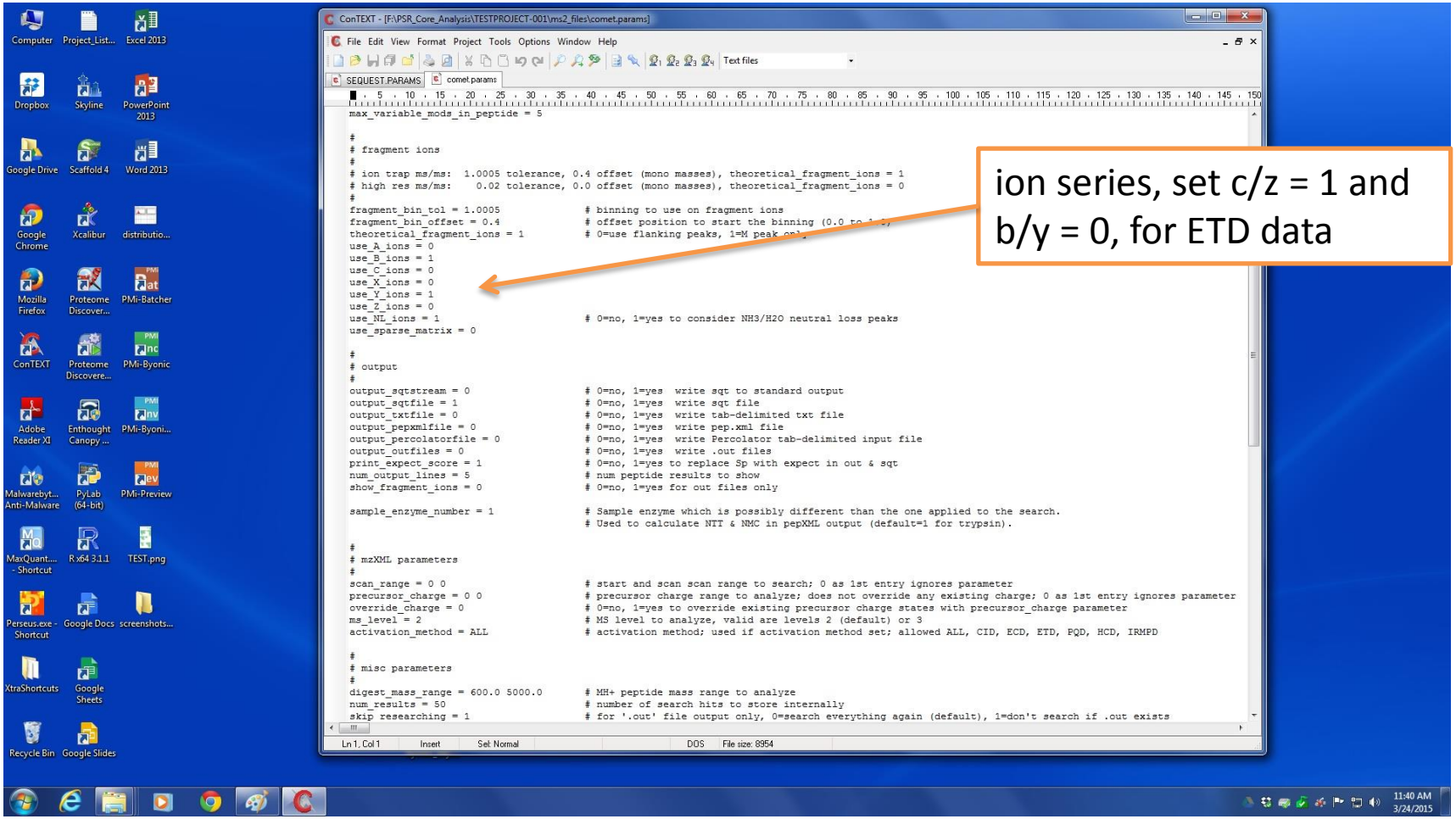
Differential modifications: mass followed by amino acid location(s). Please use at least 4 digits after the decimal point for Orbitrap data.

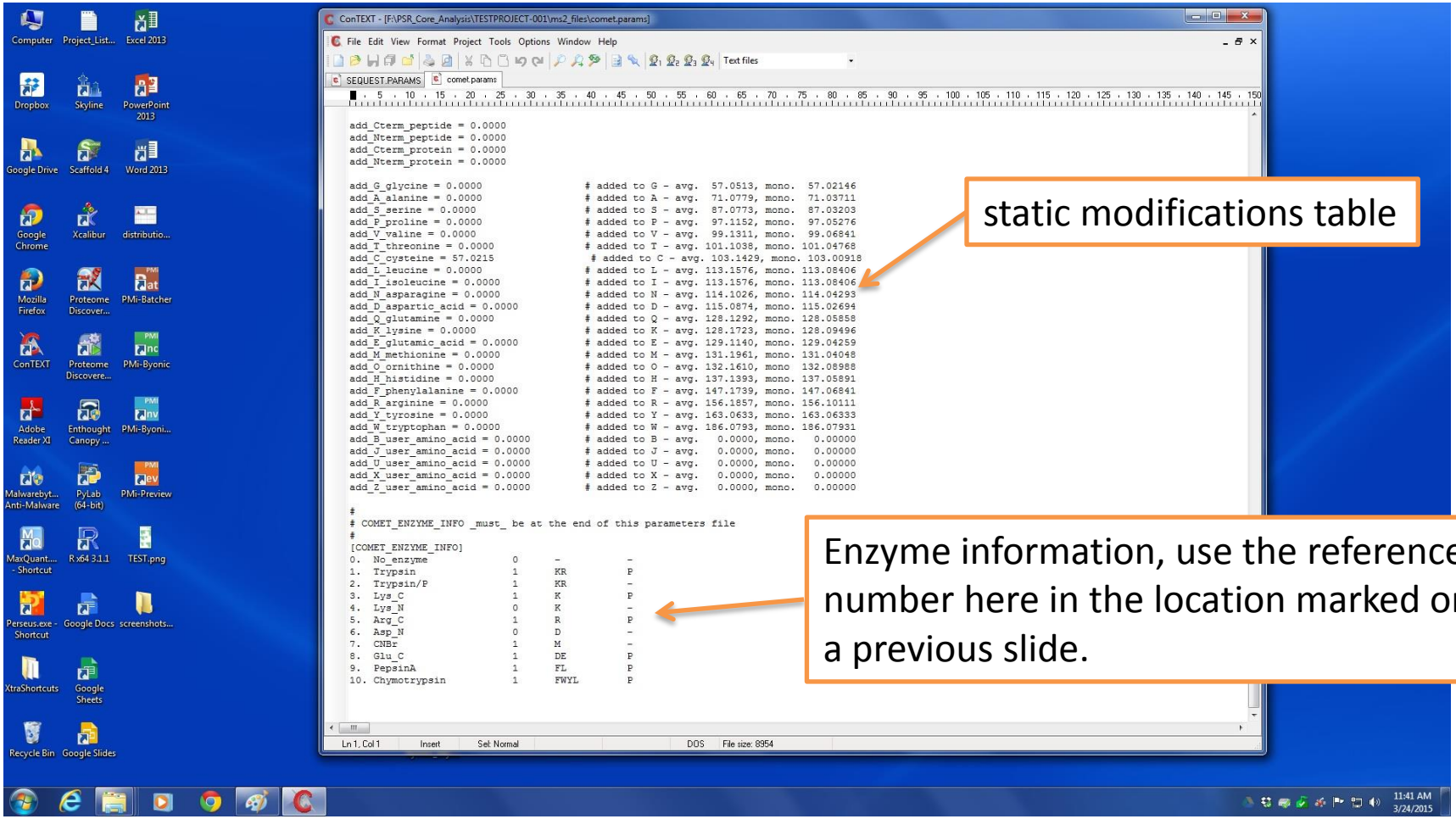
mass_type_parent should be '1' for Orbitrap data, '0' for everything else.

static modifications by amino acid: please use at least 4 digits after the decimal for Orbitrap data.

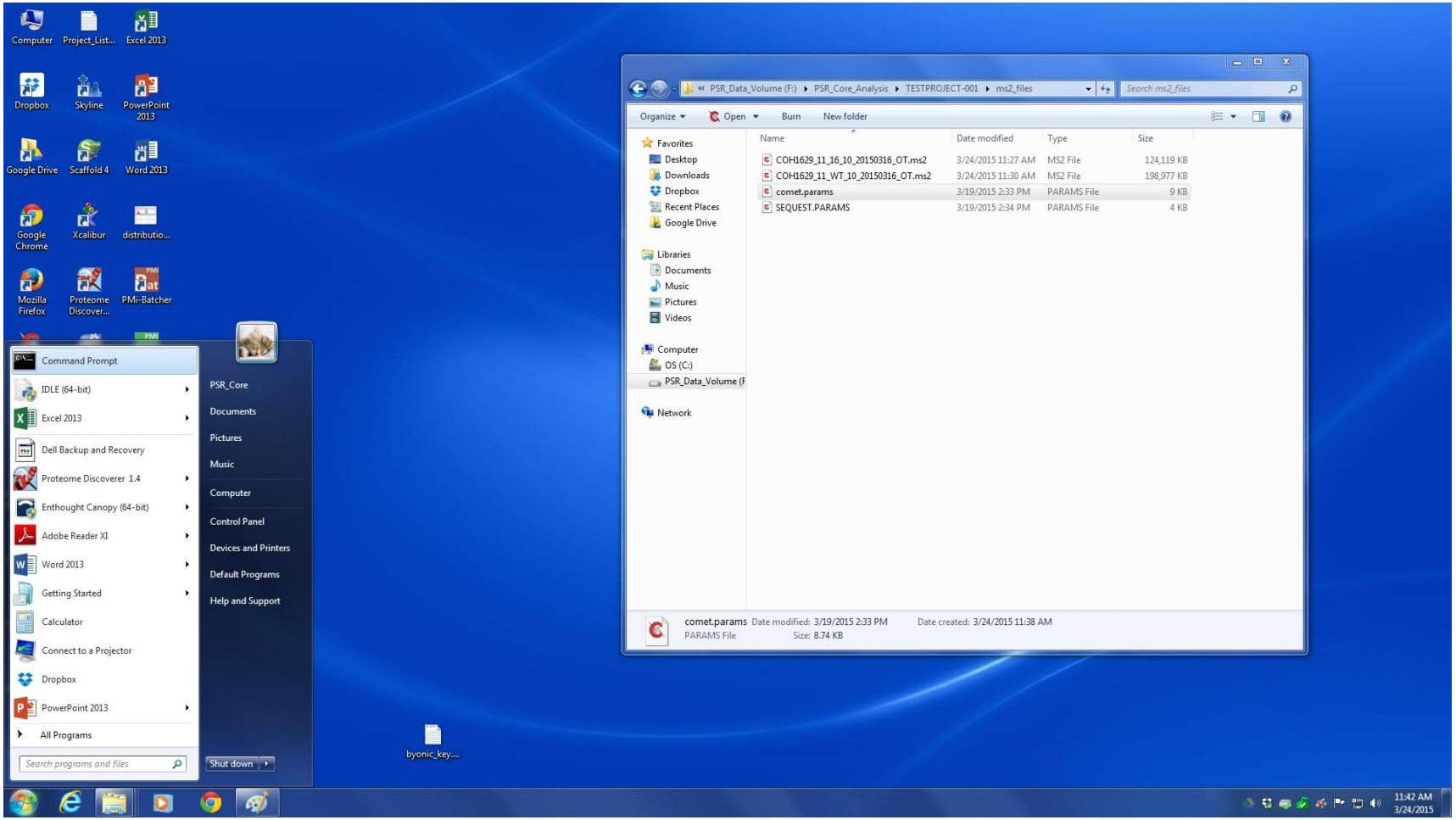
The **comet.params** file is very similar to the **sequest.params** file; however, it is longer and better annotated. All of the same options are there just in a different format. The two files should be checked to make sure the parameters are identical prior to analyzing the data. The various considerations for the parameters are the same as with the **sequest.params** file.




ConTEXT - [F:\PSR_Core_Analysis\TESTPROJECT-001\ms2_files\comet.params]
File Edit View Format Project Tools Options Window Help
SEQUEST.PARAMS
comet.params
max_variable_mods_in_peptide = 5
#
fragment ions
#
ion trap ms/ms: 1.0005 tolerance, 0.4 offset (mono masses), theoretical_fragment_ions = 1
high res ms/ms: 0.02 tolerance, 0.0 offset (mono masses), theoretical_fragment_ions = 0
#
fragment_bin_tol = 1.0005 # binning to use on fragment ions
fragment_bin_offset = 0.4 # offset position to start the binning (0.0 to 1.0)
theoretical_fragment_ions = 1 # 0=use flanking peaks, 1=M peak only
use_A_ions = 0
use_B_ions = 1
use_C_ions = 0
use_X_ions = 0
use_Y_ions = 1
use_Z_ions = 0
use_NL_ions = 1 # 0=no, 1=yes to consider NH3/H2O neutral loss peaks
use_sparse_matrix = 0
#
output
#
output_sqtstream = 0 # 0=no, 1=yes write sqt to standard output
output_sqtfile = 1 # 0=no, 1=yes write sqt file
output_txtfile = 0 # 0=no, 1=yes write tab-delimited txt file
output_pepxmlfile = 0 # 0=no, 1=yes write pep.xml file
output_percolatorfile = 0 # 0=no, 1=yes write Percolator tab-delimited input file
output_outfiles = 0 # 0=no, 1=yes write .out files
print_expect_score = 1 # 0=no, 1=yes to replace Sp with expect in out & sqt
num_output_lines = 5 # num peptide results to show
show_fragment_ions = 0 # 0=no, 1=yes for out files only
sample_enzyme_number = 1 # Sample enzyme which is possibly different than the one applied to the search.
# Used to calculate NTT & NMC in pepXML output (default=1 for trypsin).
#
mzXML parameters
#
scan_range = 0 0 # start and scan scan range to search; 0 as 1st entry ignores parameter
precursor_charge = 0 0 # precursor charge range to analyze; does not override any existing charge; 0 as 1st entry ignores parameter
override_charge = 0 # 0=no, 1=yes to override existing precursor charge states with precursor_charge parameter
ms_level = 2 # MS level to analyze, valid are levels 2 (default) or 3
activation_method = ALL # activation method; used if activation method set; allowed ALL, CID, ECD, ETD, PQD, HCD, IRMPD
#
misc parameters
#
digest_mass_range = 600.0 5000.0 # MH+ peptide mass range to analyze
num_results = 50 # number of search hits to store internally
skip_researching = 1 # for '.out' file output only, 0=search everything again (default), 1=don't search if .out exists
Ln 1, Col 1
Insert
Sel: Normal
DOS
File size: 8954
ion series, set c/z = 1 and b/y = 0, for ETD data

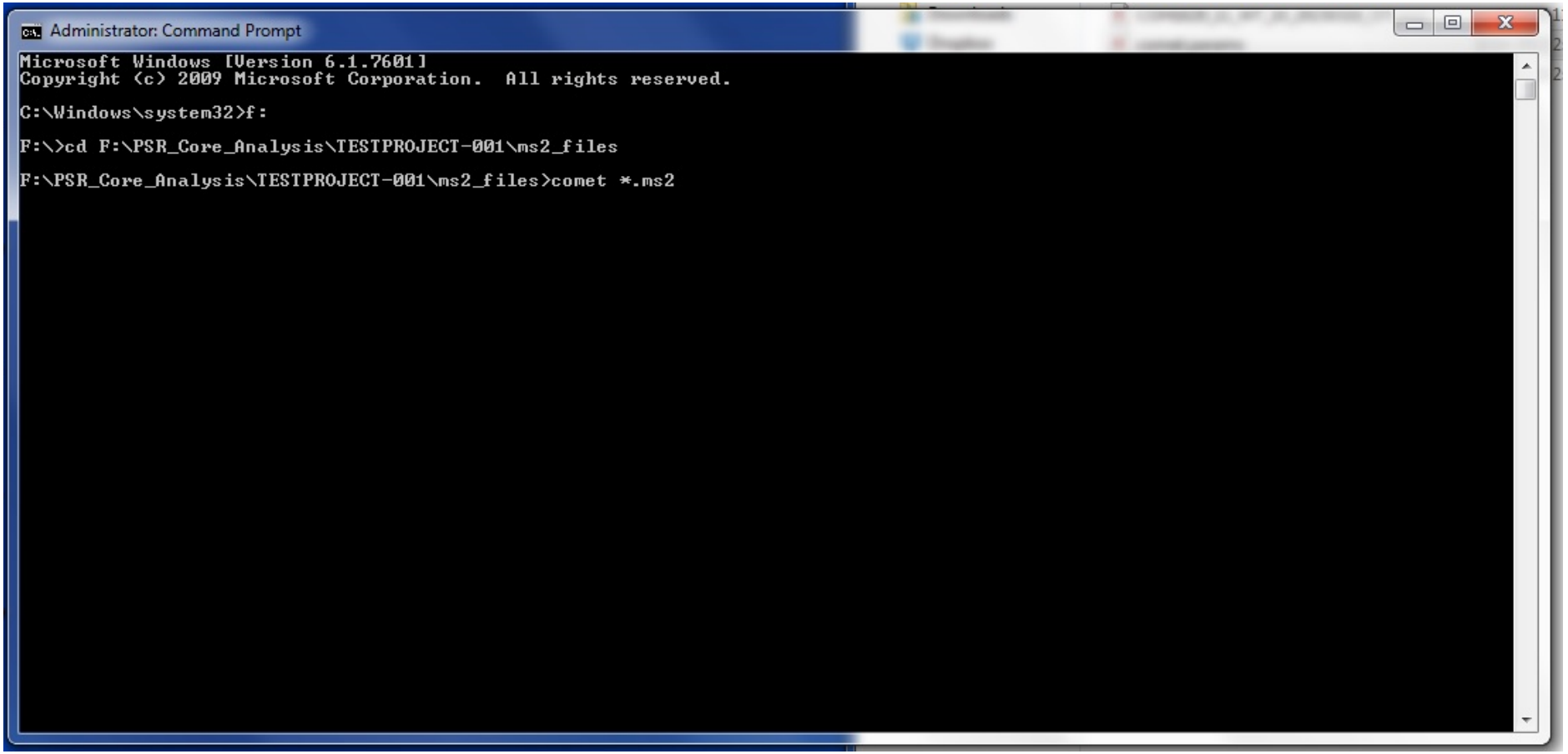

ConTEXT - [F:\PSR_Core_Analysis\TESTPROJECT-001\ms2_files\comet.params]
File Edit View Format Project Tools Options Window Help
SEQUEST.PARAMS
comet.params
add_Cterm_peptide = 0.0000
add_Nterm_peptide = 0.0000
add_Cterm_protein = 0.0000
add_Nterm_protein = 0.0000
add_G_glycine = 0.0000 # added to G - avg. 57.0513, mono. 57.02146
add_A_alanine = 0.0000 # added to A - avg. 71.0779, mono. 71.03711
add_S_serine = 0.0000 # added to S - avg. 87.0773, mono. 87.03203
add_P_proline = 0.0000 # added to P - avg. 97.1152, mono. 97.05276
add_V_valine = 0.0000 # added to V - avg. 99.1311, mono. 99.06841
add_T_threonine = 0.0000 # added to T - avg. 101.1038, mono. 101.04768
add_C_cysteine = 57.0215 # added to C - avg. 103.1429, mono. 103.00918
add_L_leucine = 0.0000 # added to L - avg. 113.1576, mono. 113.08406
add_I_isoleucine = 0.0000 # added to I - avg. 113.1576, mono. 113.08406
add_N_asparagine = 0.0000 # added to N - avg. 114.1026, mono. 114.04293
add_D_aspartic_acid = 0.0000 # added to D - avg. 115.0874, mono. 115.02694
add_Q_glutamine = 0.0000 # added to Q - avg. 128.1292, mono. 128.05858
add_K_lysine = 0.0000 # added to K - avg. 128.1723, mono. 128.09496
add_E_glutamic_acid = 0.0000 # added to E - avg. 129.1140, mono. 129.04259
add_M_methionine = 0.0000 # added to M - avg. 131.1961, mono. 131.04048
add_O_ornithine = 0.0000 # added to O - avg. 132.1610, mono 132.08988
add_H_histidine = 0.0000 # added to H - avg. 137.1393, mono. 137.05891
add_F_phenylalanine = 0.0000 # added to F - avg. 147.1739, mono. 147.06841
add_R_arginine = 0.0000 # added to R - avg. 156.1857, mono. 156.10111
add_Y_tyrosine = 0.0000 # added to Y - avg. 163.0633, mono. 163.06333
add_W_tryptophan = 0.0000 # added to W - avg. 186.0793, mono. 186.07931
add_B_user_amino_acid = 0.0000 # added to B - avg. 0.0000, mono. 0.00000
add_J_user_amino_acid = 0.0000 # added to J - avg. 0.0000, mono. 0.00000
add_U_user_amino_acid = 0.0000 # added to U - avg. 0.0000, mono. 0.00000
add_X_user_amino_acid = 0.0000 # added to X - avg. 0.0000, mono. 0.00000
add_Z_user_amino_acid = 0.0000 # added to Z - avg. 0.0000, mono. 0.00000
#
COMET_ENZYME_INFO _must_ be at the end of this parameters file
#
[COMET_ENZYME_INFO]
0. No_enzyme 0 - -
1. Trypsin 1 KR P
2. Trypsin/P 1 KR -
3. Lys_C 1 K P
4. Lys_N 0 K -
5. Arg_C 1 R P
6. Asp_N 0 D -
7. CNBr 1 M -
8. Glu_C 1 DE P
9. PepsinA 1 FL P
10. Chymotrypsin 1 FWYL P
Ln 1, Col 1
Insert
Sel: Normal
DOS
File size: 8954
static modifications table
Enzyme information, use the reference number here in the location marked on a previous slide.

After the parameters files are in place, the next step is to open up the command window and launch the Comet searches. The command prompt shortcut can be found on the PSR computers as indicated in the picture below. A **<shift> right-mouse-click** on a folder will also open a command window at that location.

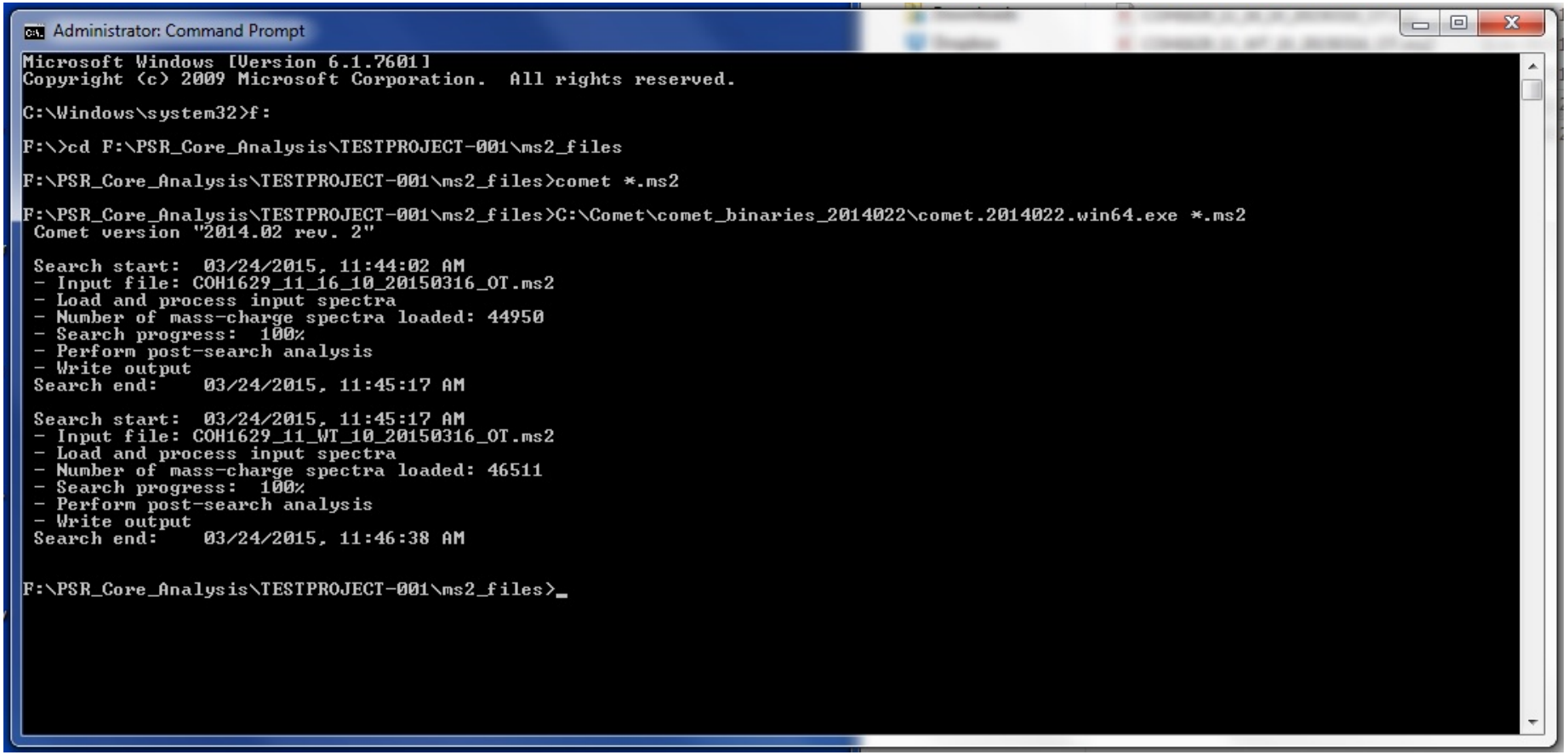


The three commands you'll need to use to run your search are:

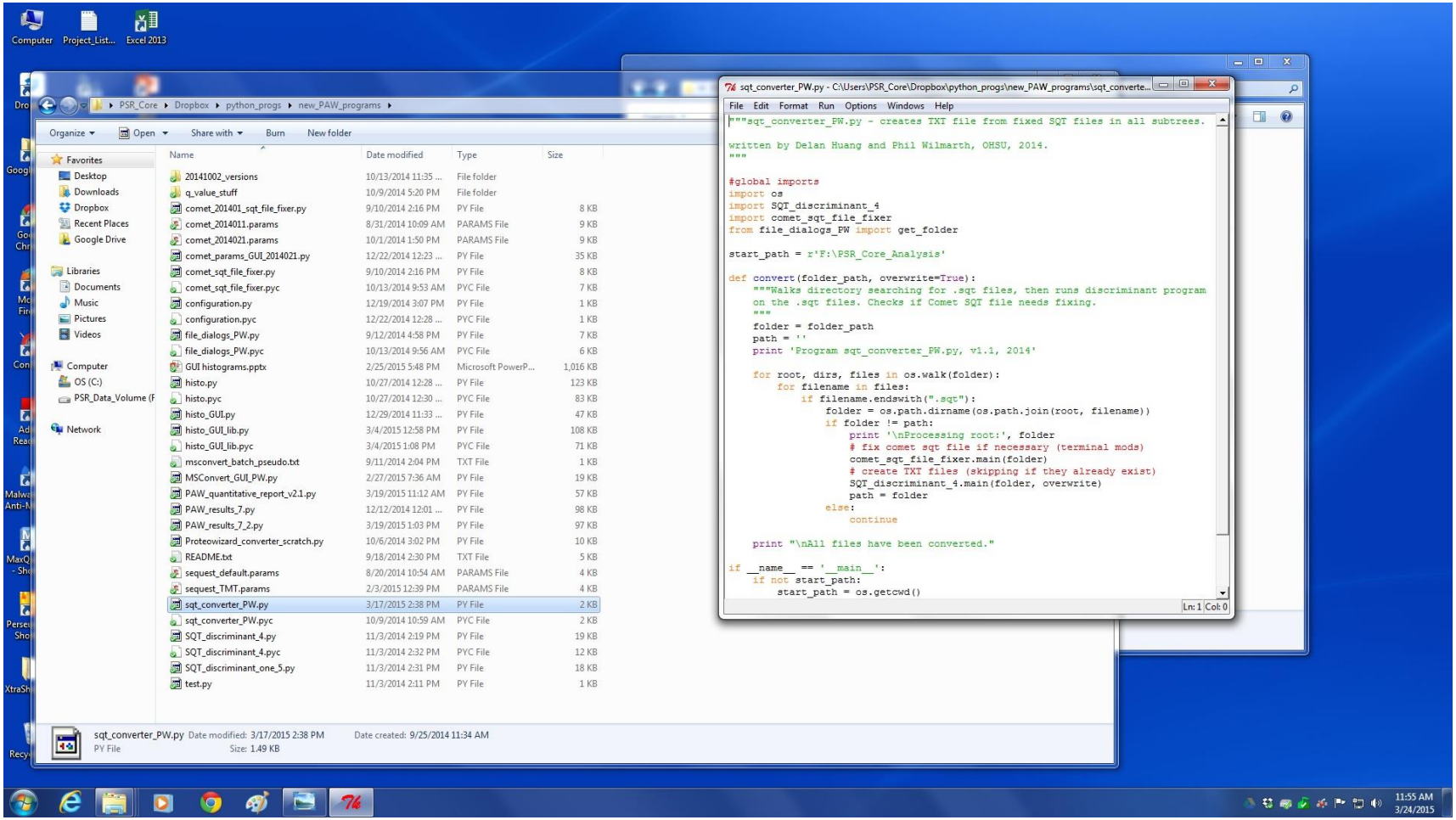
1) switch to the **F:** drive
2) **cd** (change directory) to your file location. This path can be pasted in with a right click menu. The tab key can also be used to auto-complete the path name.
3) **comet *.ms2** searches all the ms2 files in the folder; files can also be searched individually.



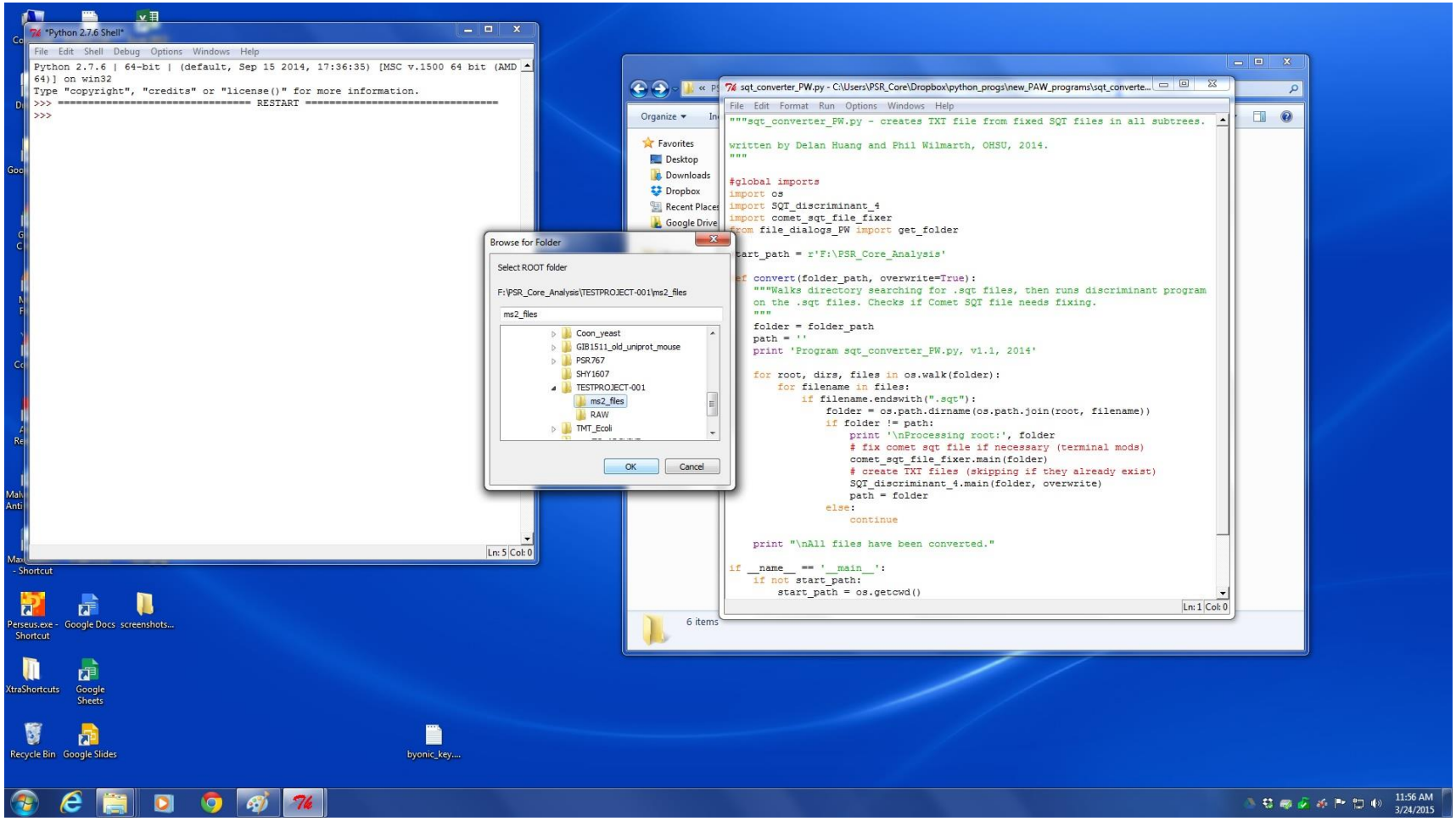
A completed Comet search will display as seen below. Comet often runs much faster than SEQUEST searches used to. A search with a single differential modification against a typical database should only take about one minute or so. Once the searches have completed you can close the command prompt window.



The next step is to run the **sqt_converter.py** program. The program takes the sqt files Comet creates and, after some processing of the data, it creates txt files that the rest of the PAW Pipeline will process. The program can launched with a double-click and run by pressing **F5** once the python window opens.



After launch, an open folder dialog box will ask you to browse to the **ms2_files** folder containing the files you wish to analyze. After selecting the proper folder and pressing okay, the program will run. This is probably the most time-intensive part of the pipeline at the moment, and can easily take 15-20 minutes per file to do the conversion.



When the processing is complete, the output window will display **All files have been converted**. At this point, it's safe to close both windows and move on to the next step. The output will also display scans with missing data. There are commonly a couple of these scans in each analysis; however, lots of missing data could indicate a problem.