InsPecT Tutorial

Sam Payne

Winter 2007

InsPecT is a tool for interpreting peptide tandem mass spectra. Information about the program and its authors can be found at http://peptide.ucsd.edu/. The purpose of this tutorial is to help guide you through the steps necessary to complete a successful run of the program. This tutorial is divided into four parts:

- Preparing your computer
- Setting up the InsPecT run
- The InsPecT run
- Analysis of the InsPecT results

The easiest operating system to use is Windows. However, the download includes the source code, so you can compile it wherever you please. The only officially supported operating systems are Windows and Linux. Some technical vocabulary is explained in Appendix A.

This tutorial covers only installing and setting up a basic InsPecT run. It does not cover most options in our basic workflow. Please consult the InsPecT Advanced Tutorial and the Unrestrictive Search Tutorial. Both of these introduce features that we routinely use. I cannot overemphasize this: please read the next tutorials to understand proper use of InsPecT.

Part 1. Preparing Your Computer

To begin, download the InsPecT package from http://peptide.ucsd.edu/. Click through to “Software” and find the InsPecT download. Save this zip file to your desktop, unzip it into a folder named something clever like “InsPecT”. The zip includes the source code for InsPecT as well as an analysis toolkit.

One distinction to make clear at the beginning is this: InsPecT is a program that you “run” and not one that you “open.” There is no graphical interface; all your work will be by typing commands into the command line. If you are unfamiliar with this, please read Appendix C. There is a web interface at http://proteomics.bioprojects.org/MassSpec, which we encourage you to use.

The installation of InsPecT and the toolkit is much more complicated on Mac/Linux than on Windows, so it is covered separately in appendix B. This section will only include relevant information for Windows users.

The Python programming language

To help you analyze the InsPecT results, the authors included several Python scripts with the InsPecT download. Using these scripts requires Python on your machine. The newest Python release can be found at http://www.python.org/download/releases/. Download and install Python 2.5. Next, add Python to your PATH, in the form of “C:\Python25”. See http://www.xmission.com/~comphope/issues/ch000549.htm for help with the PATH variable. Next, install the Python Image Library. Go to http://www.pythonware.com/products/pil/#pil115 and download and install the file for python 2.5 for windows. Finally, install the Numerical Python library “NumPy.” Go to http://sourceforge.net/project/showfiles.php?group_id=1369 and download and install the file numpy-1.0.win32-py2.5.exe.

Part2. Setting Up The InsPecT Run

InsPecT expects your data to be in the right format prior to the run: MS/MS spectra in a common format, a trie database, and an input file. In this section I introduce the idea of a command prompt and interfacing with programs on the command line. If these topics are unfamiliar to you, please read relevant entries in appendix A and C.

MS/MS Spectra

The spectra files for InsPecT must be in a common, non-proprietary format. The preferred format is mzXML, but the mgf format is acceptable. The lab with an MS/MS machine should have tools to convert to one of these formats. (If you are that lab, then ask the manufacturer.) See also the Sashimi project page (http://sashimi.sourceforge.net/software_glossolalia.html) for mzXML conversion tools.

Database Setup

For its super-fast speed, InsPecT requires a special kind of database called a “trie”. The trie is made from a fasta file of protein sequences. (The human proteome can be downloaded from NCBI at ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/H_sapiens/protein). Put the fasta file in the InsPecT\Database directory and then go to the command line (in the InsPecT directory) and type

```
> python PrepDB.py FASTA Database\myDB.fasta
```

where “myDB.fasta” is the name of your database. This creates two files which will be used by InsPecT: myDB.trie and myDB.index. Once these files are created, they can be reused for later InsPecT runs. Therefore, it is advisable to create a whole-proteome database for your organism of interest. If the script fails, please read Appendix C.

Input file

A simple input file feeds arguments to InsPecT. Here I introduce basic and necessary commands only. More options are covered in the advanced tutorial. An input file, which you create for each InsPecT run, looks like this:

```
spectra,Fraction01.mzxml
instrument,ESI-ION-TRAP
protease,Trypsin
DB,myDB.trie
# Protecting group on cysteine:
mod,57,C,fix
```

Each line contains a command and the argument, separated by a comma. The following explanation of the arguments is abridged from the documentation page for searching.
spectra,[FILENAME] - Specifies a spectrum file to search.
instrument,[TYPE] - Options are ESI-ION-TRAP, QTOF, and FT-Hybrid.
db,[FILENAME] - Specifies the name of a database (.trie file) to search.
protease,[NAME] - Specifies the name of a protease.
mod,[MASS],[RESIDUES],[TYPE],[NAME] - Specifies an amino acid modification. The example above reflects the addition of CAM (carbamidomethylation, done by adding iodoacetamide) to cysteine, inhibiting disulfide bonds. Most searches should include this line.

Part 3. The InsPecT Run

Now that you have gotten this far, the actual InsPecT run is very simple. Type the following at the command prompt

```
> InsPecT.exe –i InputFile.txt –o OutputFile.txt                (windows)
> ./InsPecT –i InputFile.txt –o OutputFile.txt                     (unix)
```

Depending on the size of the database and the number of spectra, InsPecT may take only a few minutes, or several hours. After processing every 100 spectra, the program will print progress to the screen, so you won't be totally lost on the time investment.

Part 4. Analysis of the Results

The InsPecT output file contains an annotation for every MS/MS spectrum, most of which are not statistically significant. Basic filtering and analysis can be done with the toolkit, i.e. python scripts included in the distribution. It is *essential* that you do post-processing. At the very least you **must** run the PValue.py script (see below and Advanced tutorial).

PValue.py

The purpose of this script is to weed out insignificant results. The method is derived from a statistical model of the results distribution (Keller et al 2002 Analytical Chemistry 74:5383-5392). See appendix C for a walk through. This step is crucial if you have multiple InsPecT results files from a single experiment (perhaps stemming from multiple mzxml files for the multiple fractions of your LC run). It creates a unified pvalue across all files.

Summary.py

So a big text file is not what you want? Well, we kind of figured that. The summary script creates a webpage with peptides grouped into proteins. It uses the Python Image Library to make pictures of annotated spectra. Again, the walk through in appendix C shows you how to use this program.

Appendix A – Vocabulary

Arguments – Input options to a computer program, eg. spectrum file, database, etc.

Command prompt – A text input window for giving commands to the program. Windows users should click Start/Run and type “cmd”. This will pop up a command prompt window, which is what we want. See terminal, appendix C.

Compile – Computer programs are written in a humanish language, like C or Java. The computer cannot directly run these files. They must be converted to machine language (which is all ones and zeros: 1000100101110101 01). Compiling is the process of converting a program to machine language. Commercial software comes pre-compiled, which is why you normally don’t worry about this step.

Directory Navigation – When a command prompt is open, it is living (or working) in a particular directory. All commands are interpreted in the context of that directory. For the purpose of this tutorial, you always want to be in the “InsPecT” directory. Only there will the commands outlined in the tutorial be understood. After opening a command prompt, you see something like this

C:\Documents and Settings\Sam>

where “Sam” is your own user name and is the current working directory. If you put the InsPecT folder on your desktop, you navigate to it by typing “cd Desktop\InsPecT” See Appendix C.

Directory Structure – When using the command line, it is necessary to correctly reference folders and files. All folders in a computer are stored internally in something analogous to a family tree. If I add two folders to my InsPecT folder called “Spectra” and “Results”, these would be siblings and the parent would be “InsPecT”. The Spectra directory may have a folder like “nfKB”. e.g.

```
Desktop → InsPecT → Spectra → nfKB → 0001.mzXML
                  → Results → nfKB → RawResults → 0001.txt
                  → Results → nfKB →Pvalued → 0001.txt
```

Executable – The actual file that you run for a given program.

Terminal – What unix systems (including Mac OS X) call a command prompt. On the Mac, it can be found in Applications/Utilities folder.

Appendix B – The Apple Install

This appendix is to help the Mac OS X user install all the necessary components for running InsPecT. Here at the beginning, note that Mac is not a supported platform. You can get it to work, but it may take some effort. I don't have full instructions for this, but some helpful stuff is written out below.

For the average Mac user, this may involve new terminology and territory. Please read appendix A to become familiar all the defined terms. If at any time you have trouble with this protocol, I recommend that you find the geekiest person that you can, show them this tutorial and ask for help.

C Compiler

I assume that you've already downloaded and unzipped the InsPecT file from the website. InsPecT is written in the computer language C. Non-windows users will have to compile the program into an executable for their machine. Apple's Xcode will do this for us. Go to http://developer.apple.com/tools/xcode. Over on the right side click "Download Tools." Walk through the installation (click the .dmg file that is on your desktop for starters). After the installation is finished, open a terminal and type

```
> which gcc
```

The teminal should return something like "/usr/bin/gcc" telling you the location of the compiler (named gcc). If instead it returns something like "no gcc in (/usr/local/bin:/usr/bin)" then you have a problem.

Expat XML Parser

InsPecT supports a new database format using XML. For this reason, it requires an XML parser called expat. Download expat from http://sourceforge.net/projects/expat/. The installation of this requires you to be the root user (administrator) for you computer. Most OS X machines have this disabled. No fear, you can follow the instructions on this page to enable it. www.spy-hill.com/~myers/help/apple/EnableRoot.html. After conquering the root user issue, extract the zipped expat download and run the following three commands from the unzipped directory.

```
> ./configure
> make
> make install
```

Building InsPecT

Open a terminal, navigate to the InsPecT directory and type "make". A lot of stuff will get printed to the screen; don't worry. In the end, you should get something like this

```
cc -g -DDEBUG -D_CONSOLE -O3 -funroll-loops -lm -o InsPecT main.o   FreeMod.o
Trie.o Utils.o DeNovo.o Mods.o Score.o Tagger.o SVM.o Run.o ChargeState.o
```

```
Scorpion.o Spliced.o SpliceDB.o BN.o Spectrum.o SpliceScan.o SNP.o ORFDB.o
ExonGraphAlign.o ParseXML.o base64.o PValue.o
```

and you should find the executable, called “InsPecT,” in the directory.

Python Libraries

Python, another programming language, is used for post-processing of the InsPecT results. Mac OSX comes with python already installed. You will, however need to install two auxiliary libraries: NumPy and PIL. Lets get the NumPy first. Go to http://sourceforge.net/project/showfiles.php?group_id=1369 and download the file called numpy-1.0.1.tar.gz (possibly bigger numbers, but still the .tar.gz). Then do the following commands

```
> gunzip numpy-1.0.1.tar.gz
> tar –xvf numpy-1.0.1.tar
> cd numpy
> sudo python setup.py install
```

Finally let’s tackle the python image library (PIL). You can find the most current version of the PIL online at http://www.pythonware.com/products/pil/#pil116. You should download the version that says “Source Kit” as it is the source code. Run the following commands

```
>gunzip Imaging-1.1.6.tar.gz
>tar –xvf Imaging-1.1.6.tar
> cd directory of untarred thing
>./configure
>make
>make install
```

If getting the PIL and NumPy does not work, you could also try DarwinPorts or Fink. I’ve never used these, but they claim to help you install software on your mac, and make it painless.

http://fink.sourceforge.net/download/index.php
http://darwinports.opendarwin.org/

PyInsPecT

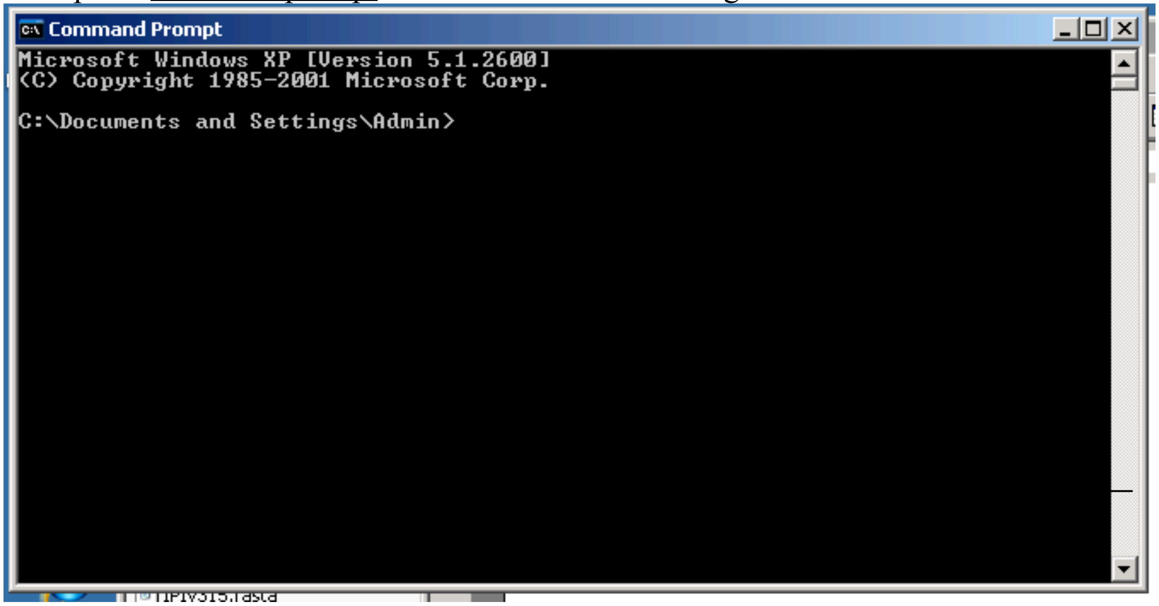
One last note, many of the post-processing scripts use PyInspect, the python interface into the C code of InsPecT. You must create a file for this to work correctly. Do the following:

```
> Python ReleasePyInspect.py build
```

At the end of execution, this command will print the path to PyInspect.so. Copy that file to the InsPecT directory (from some random temp directory). Then things should work.

Appendix C: Working with the Command Line

As noted in part 1, the InsPecT program cannot be opened. The only way to work with the program is at the command line, a concept I'm going to explain. So let's start out and open a command prompt. You should see something like this.



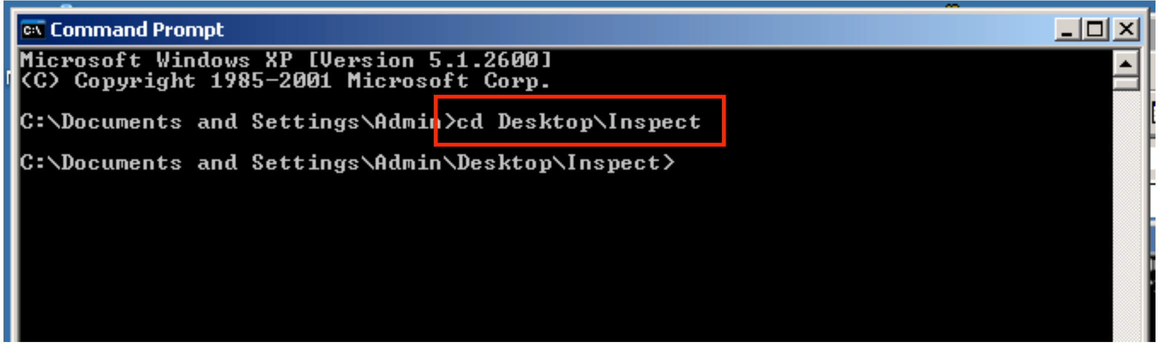
```
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

C:\Documents and Settings\Admin>
```

The cursor will probably be blipping just to the right of the ">". It is here that you type in commands and "run" the program. In the tutorial, I will often use ">" as an indicator that you are supposed to type something in at the command line. Note that the ">" is just a placeholder and should NOT be typed.

Our first task is to navigate to the correct directory. We do this by changing our current directory. If you stored InsPecT in a folder on the Desktop then navigate to it by typing

```
> cd Desktop\InsPecT
```

Here "cd" is a command to Change Directories. Your should now see



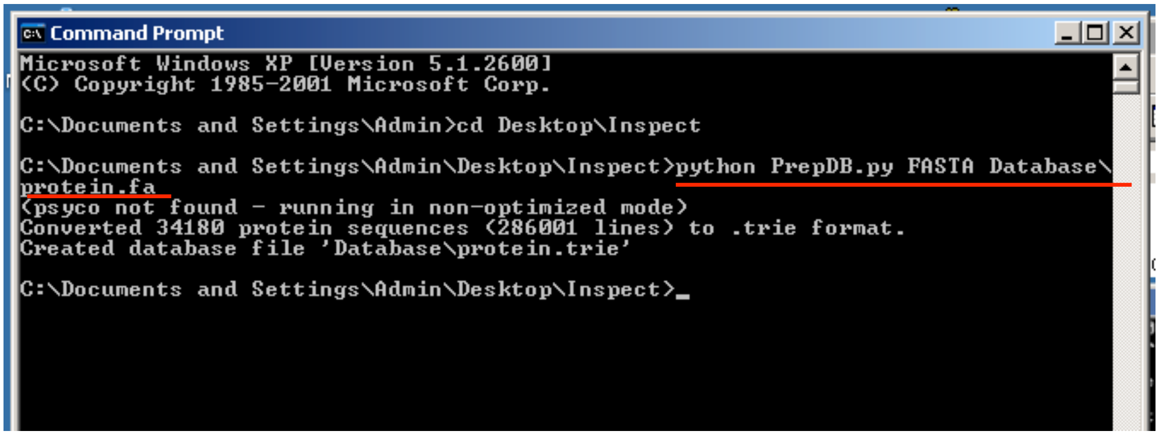
```
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

C:\Documents and Settings\Admin>cd Desktop\Inspect

C:\Documents and Settings\Admin\Desktop\Inspect>
```

Only after you have moved into this directory will the computer understand InsPecT commands. Let's do the first step of the tutorial, creating a database. After that, I think that you'll have the hang of it. From the Database Setup portion of Part 2, you see the command

```
> python PrepDB.py FASTA Database\myDB.fasta
```

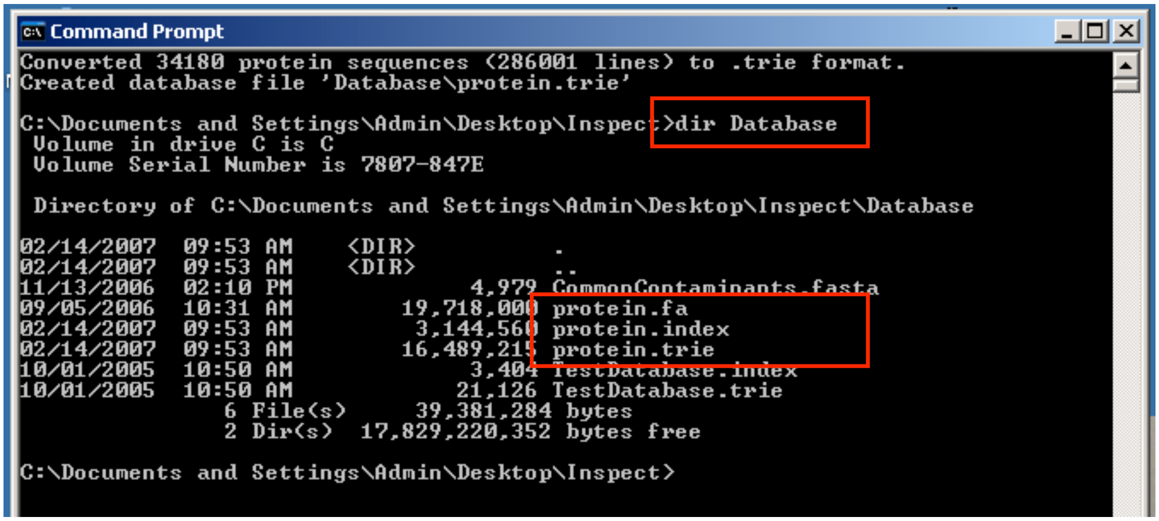
Let's do that. The text "myDB.fasta" is a placeholder for whatever you happened to name your database. Download the human proteome from the NCBI link in the tutorial (protein.fa.gz). Unzip the file and put it into the InsPecT\Database folder. Your Database folder should now have a file called "protein.fa". To convert that fasta file to the correct format, type

```
> python PrepDB.py FASTA Database\protein.fa
```



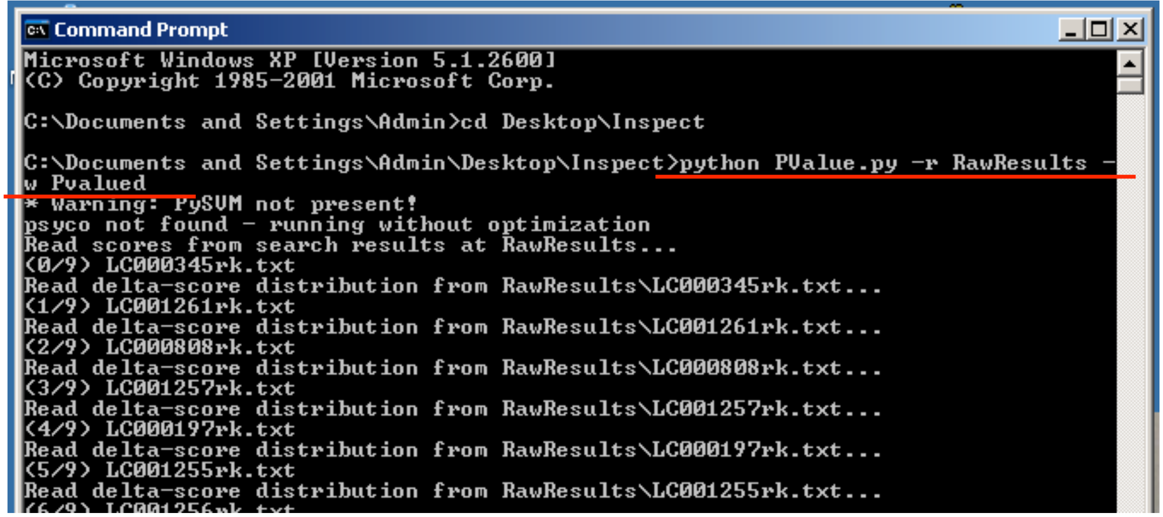
```
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

C:\Documents and Settings\Admin>cd Desktop\Inspect

C:\Documents and Settings\Admin\Desktop\Inspect>python PrepDB.py FASTA Database\
protein.fa
(psyco not found - running in non-optimized mode)
Converted 34180 protein sequences (286001 lines) to .trie format.
Created database file 'Database\protein.trie'

C:\Documents and Settings\Admin\Desktop\Inspect>_
```

Notice that you've created new files protein.trie and protein.index. That's wonderful. We can check that the file exists by listing the directory ("dir" command).



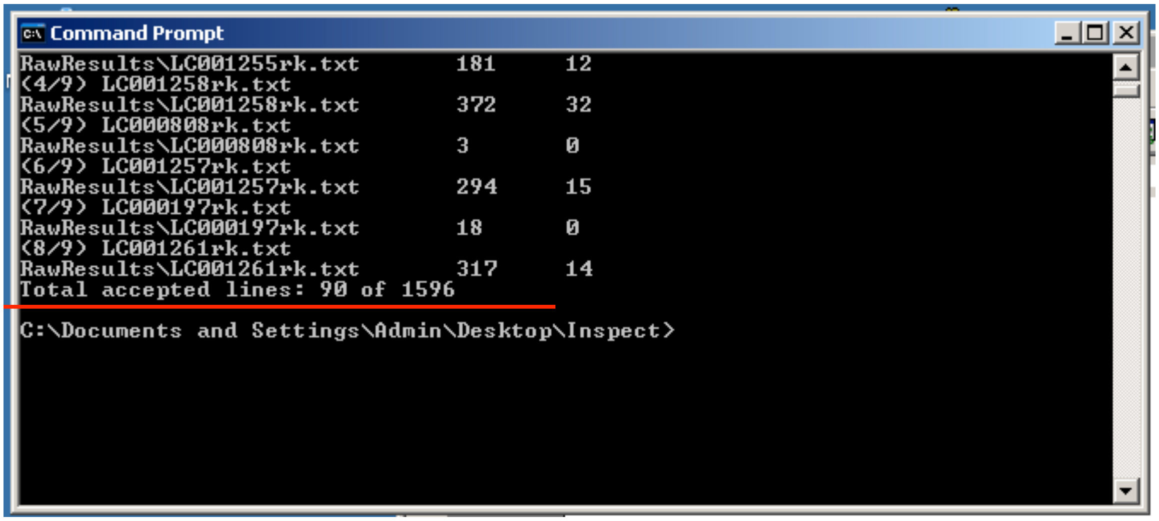
```
Converted 34180 protein sequences (286001 lines) to .trie format.
Created database file 'Database\protein.trie'

C:\Documents and Settings\Admin\Desktop\Inspect>dir Database
 Volume in drive C is C
 Volume Serial Number is 7807-847E

 Directory of C:\Documents and Settings\Admin\Desktop\Inspect\Database

02/14/2007  09:53 AM    <DIR>          .
02/14/2007  09:53 AM    <DIR>          ..
11/13/2006  02:10 PM             4,979 CommonContaminants.fasta
09/05/2006  10:31 AM        19,718,000 protein.fa
02/14/2007  09:53 AM         3,144,560 protein.index
02/14/2007  09:53 AM        16,489,215 protein.trie
10/01/2005  10:50 AM             3,404 TestDatabase.index
10/01/2005  10:50 AM            21,126 TestDatabase.trie
               6 File(s)     39,381,284 bytes
               2 Dir(s)  17,829,220,352 bytes free

C:\Documents and Settings\Admin\Desktop\Inspect>
```

Now, supposing that you've finished the InsPecT run, let's filter the output with PValue.py. I am assuming a directory structure, which you ought to mirror. I've put my raw results from InsPecT in a folder called "RawResults" within my InsPecT directory. I want to put the filtered results into a folder called "Pvalued". Here is what I input



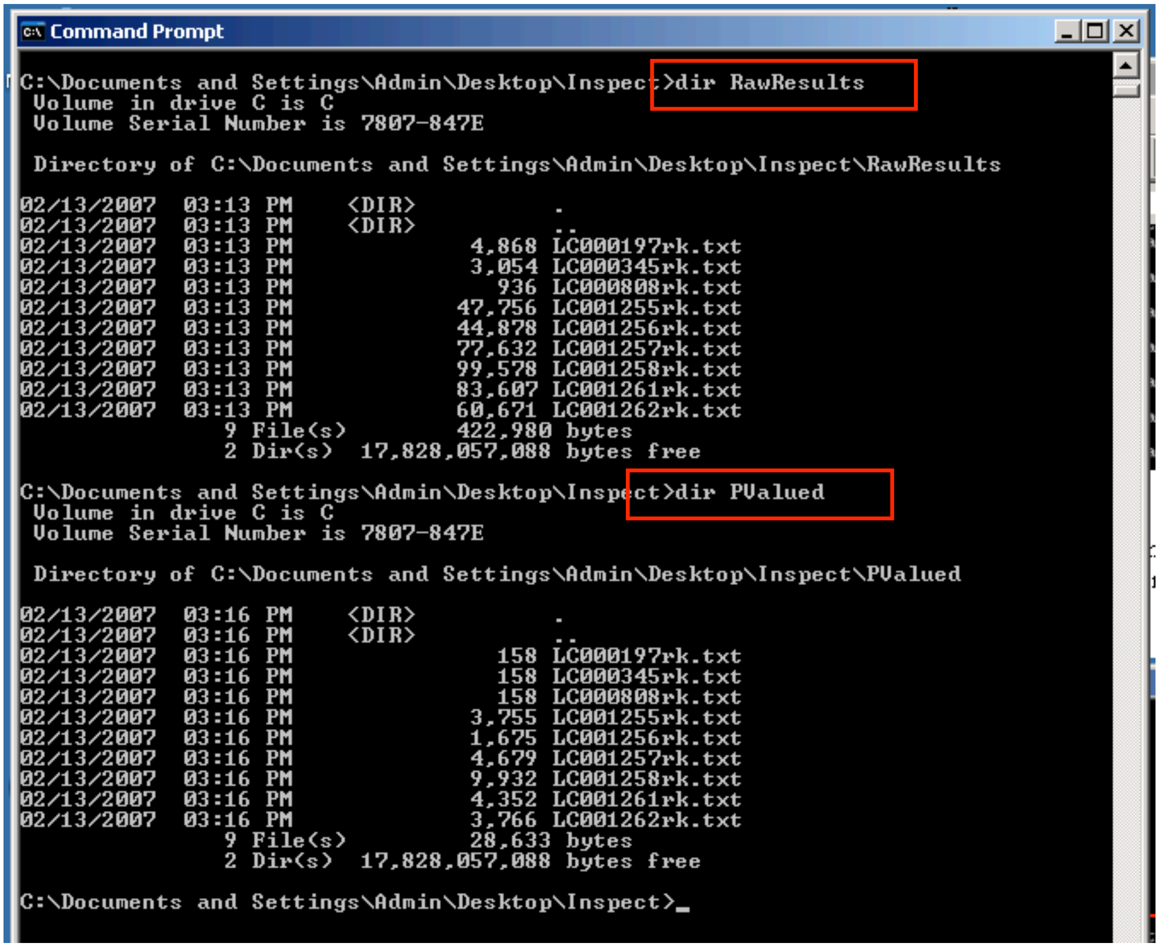
```
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

C:\Documents and Settings\Admin>cd Desktop\Inspect

C:\Documents and Settings\Admin\Desktop\Inspect>python PValue.py -r RawResults -
w Pvalued
* Warning: PySVM not present!
psyco not found - running without optimization
Read scores from search results at RawResults...
(0/9) LC000345rk.txt
Read delta-score distribution from RawResults\LC000345rk.txt...
(1/9) LC001261rk.txt
Read delta-score distribution from RawResults\LC001261rk.txt...
(2/9) LC000808rk.txt
Read delta-score distribution from RawResults\LC000808rk.txt...
(3/9) LC001257rk.txt
Read delta-score distribution from RawResults\LC001257rk.txt...
(4/9) LC000197rk.txt
Read delta-score distribution from RawResults\LC000197rk.txt...
(5/9) LC001255rk.txt
Read delta-score distribution from RawResults\LC001255rk.txt...
(6/9) LC001256rk.txt
```

You can see that it is telling me that I am missing some valuable libraries (which you could install if you want), and then it starts to read in files and do stuff. It finally ends with –



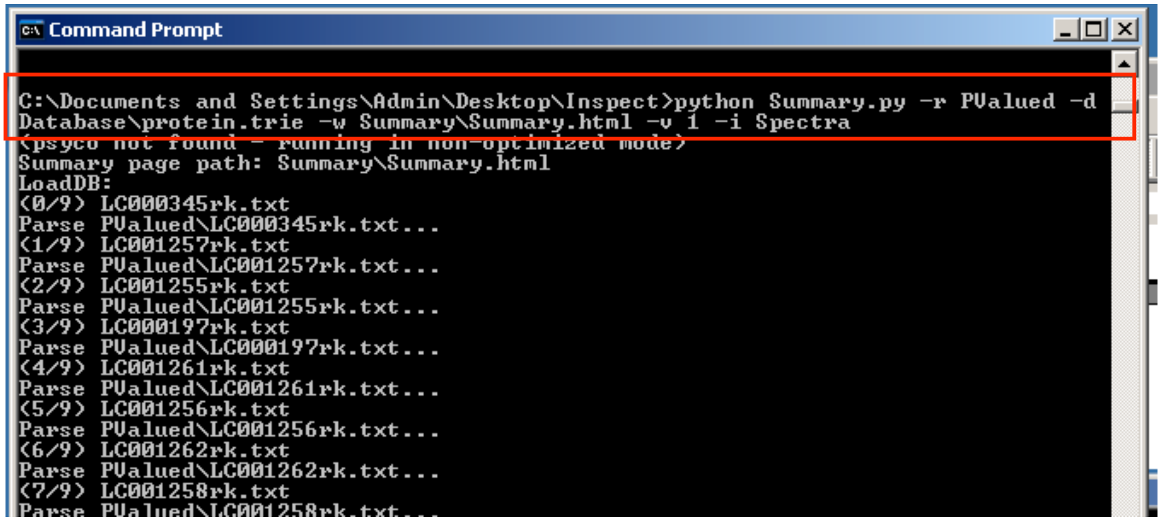
```
RawResults\LC001255rk.txt       181     12
(4/9) LC001258rk.txt
RawResults\LC001258rk.txt       372     32
(5/9) LC000808rk.txt
RawResults\LC000808rk.txt       3       0
(6/9) LC001257rk.txt
RawResults\LC001257rk.txt       294     15
(7/9) LC000197rk.txt
RawResults\LC000197rk.txt       18      0
(8/9) LC001261rk.txt
RawResults\LC001261rk.txt       317     14
Total accepted lines: 90 of 1596

C:\Documents and Settings\Admin\Desktop\Inspect>
```

telling me that it has accepted 90 lines from the original input. Now in my InsPecT directory there is a folder called "Pvalued" which contains a filtered file for every original file. We can see this by listing the directories as so.



```
C:\Documents and Settings\Admin\Desktop\Inspect>dir RawResults
 Volume in drive C is C
 Volume Serial Number is 7807-847E

 Directory of C:\Documents and Settings\Admin\Desktop\Inspect\RawResults

02/13/2007  03:13 PM    <DIR>          .
02/13/2007  03:13 PM    <DIR>          ..
02/13/2007  03:13 PM             4,868 LC000197rk.txt
02/13/2007  03:13 PM             3,054 LC000345rk.txt
02/13/2007  03:13 PM               936 LC000808rk.txt
02/13/2007  03:13 PM            47,756 LC001255rk.txt
02/13/2007  03:13 PM            44,878 LC001256rk.txt
02/13/2007  03:13 PM            77,632 LC001257rk.txt
02/13/2007  03:13 PM            99,578 LC001258rk.txt
02/13/2007  03:13 PM            83,607 LC001261rk.txt
02/13/2007  03:13 PM            60,671 LC001262rk.txt
               9 File(s)        422,980 bytes
               2 Dir(s)  17,828,057,088 bytes free

C:\Documents and Settings\Admin\Desktop\Inspect>dir PValued
 Volume in drive C is C
 Volume Serial Number is 7807-847E

 Directory of C:\Documents and Settings\Admin\Desktop\Inspect\PValued

02/13/2007  03:16 PM    <DIR>          .
02/13/2007  03:16 PM    <DIR>          ..
02/13/2007  03:16 PM               158 LC000197rk.txt
02/13/2007  03:16 PM               158 LC000345rk.txt
02/13/2007  03:16 PM               158 LC000808rk.txt
02/13/2007  03:16 PM             3,755 LC001255rk.txt
02/13/2007  03:16 PM             1,675 LC001256rk.txt
02/13/2007  03:16 PM             4,679 LC001257rk.txt
02/13/2007  03:16 PM             9,932 LC001258rk.txt
02/13/2007  03:16 PM             4,352 LC001261rk.txt
02/13/2007  03:16 PM             3,766 LC001262rk.txt
               9 File(s)         28,633 bytes
               2 Dir(s)  17,828,057,088 bytes free

C:\Documents and Settings\Admin\Desktop\Inspect>_
```

Now to finish off the analysis, let's do a Summary script on the Pvalued results. If I've stored my spectra in a "Spectra" folder inside the InsPecT folder, the following command will create a web page style output with images for all of the annotated peptides. All the output will be in a folder called "Summary" and the file "Summary.html" will be what you look at in the web browser.



```
C:\Documents and Settings\Admin\Desktop\Inspect>python Summary.py -r PValued -d
Database\protein.trie -w Summary\Summary.html -v 1 -i Spectra
(psyco not found - running in non-optimized mode)
Summary page path: Summary\Summary.html
LoadDB:
(0/9) LC000345rk.txt
Parse PValued\LC000345rk.txt...
(1/9) LC001257rk.txt
Parse PValued\LC001257rk.txt...
(2/9) LC001255rk.txt
Parse PValued\LC001255rk.txt...
(3/9) LC000197rk.txt
Parse PValued\LC000197rk.txt...
(4/9) LC001261rk.txt
Parse PValued\LC001261rk.txt...
(5/9) LC001256rk.txt
Parse PValued\LC001256rk.txt...
(6/9) LC001262rk.txt
Parse PValued\LC001262rk.txt...
(7/9) LC001258rk.txt
Parse PValued\LC001258rk.txt...
```