Hscorer : User guide

This script was written for python2.5 and above and has been tested on a LINUX Red Hat 5, Windows XP and windows 7 desktop operating systems.

The requirements are as follows:
python interpreter & source code may be downloaded here: http://www.python.org/
Required python packages:
numpy1.4.1 http://www.scipy.org/
path : http://pypi.python.org/pypi/path.py/2.2

Renaming downloaded files:

The three downloaded files need to be renamed, the fourth mascotmasses.txt does not need to be renamed.

Hscorer.txt → Hscorer.py
Examplemgf.txt → Example.mgf
Exampledata.txt → Example.data

Overview:

This single script performs both the MGF file filtering and the Hscoring described in the Manuscript. The reason for this is that it's the simplest method to parse out the mz/intensity data for each spectrum and then match it to an extract of the corresponding mascot search results. One may wish to create the filtered MGF file first and put this forward for Mascot searching. Later, one can simply take the resulting mgf file and an extract of the Mascot .dat file and run the script again, this time the hscore file will contain the expected cleavage patterns. It must be kept in mind that the filtered MGF file bears the extension .hmgf and would NOT be picked up by the Hscorer script, unless renamed to the .mgf extension.

Command line synopsis:

The following parameters may be supplied. When not supplied, default settings are used. All parameters must be prefixed with '--' and supplied after the command 'python Hscorer.py' (see example)

--myDir = directory from which to pick up mgf file. If no directory is supplied here then the current directory (where script is running) is used instead. The programme expects to find (and will process) all files with the extension '.mgf'. The file containing data on which the Hscore should be performed (essentially an extract of the Mascot.dat file) should have this filename stem but with the extension '.data'

For example 'myfile.mgf' will be filtered and deconvoluted /deisotoped and the filtered data written to 'myfile.hmgf', the mascot extract would have the name myfile.data.

see 'file formats' section for more information

--tolerance = tolerance for ion to theoretical fragment matching, default is 20.0
--units = unit of tolerance given above (ppm or da only), default is ppm
--quantmeth = quantification method used (tmt/itraq only); the reporter ions corresponding to hard-coded values are removed from the MGF file. Default is NONE (ie no ions removed from MGF)

--massfile = file containing masses of amino acids & any modifications (loosely based on Mascot's dat file 'mods' section –see file format section for a detailed description.

File formats

MGF file

It is expected that the MGF file adheres to the conventions described in the Mascot documentation (http://www.matrixscience.com/help/data_file_help.html), This means that the TITLE line contains an identifier which can be parsed out and used to match the spectrum information to the Mascot file extract (see below)

Example:

BEGIN IONS
TITLE=**spectrum** *< F000260>*
PEPMASS=391.2837308
CHARGE=2+
118.577477 127.963500
137.466110 134.210400
140.857254 112.668200
145.897202 141.760600
147.062927 182.860600
149.023056 867.354600
...
652.524719 125.630200
673.542114 161.847800
END IONS

BEGIN IONS
TITLE=spectrum *<F00065>*
...

It is expected that the TITLE line contains the word 'spectrum' followed by a space, followed by the identifier. In this example we have used F000260

Mascot extract (.data file)

An extract of a mascot output on which to perform the Hscoring should take the format: 'id, peptide, numeric modstring', which may be found in the mascot 'peptides' section. The first ID corresponds exactly to the spectrum id given in the .MGF file (see above)

For our example the line in the dat file might look thus:
1_p1=0,748.386780,-0.000404,4,**THELHL**,36,**00000000**,10.10,0001002010000000000,0,0;"IPI00013212.1":0:445:450:1,"IPI00910540.1":0:394:399:1
would equal

F000260	THELHL	00000000

in the extract file (where each column is tab-separated)

Hscore file (*_hscore.data file)

The results of the H-score procedure are concatenated to the first 3 columns of the .data file (and written to a new file). For the example above, the outcome would be something like, again tab-separated

F000260	THELHL	00000000	TH|E|L|H|L	5

Masses file

The format of the file containing the masses extracted from the Mascot .dat file. It should be text (tab-delimited) where the first column is the AA residue; delta value from Mascot for the given modification or element name and the second column is the mass. Any further columns will be ignored. It is worth noting that the masses of fixed modifications should be added to the respective amino acids (as is the case in the Mascot .dat file) ie for 'C': 160.030649, carbamidomethylated cysteine. And 'K':357.257895 for TMT6plex modified lysine.

ie part of the massfile will have the following format (each column is tab-separated)

A	71.037114
B	114.53494
4	229.162933

where the mascot.dat file 'masses section' will have this format:

A=71.037114
B=114.534940
delta4=229.162933,TMT6plex
NeutralLoss4=0.000000

Example

Using the supplied files the user may test the script works by running it in the following way (with all cmd line arguments)

myprompt$>python Hscorer.py --massfile mascotmasses.txt --tolerance 20.0 --units ppm

```
running with following parameters:
on directory .
with tolerance 20.000000
units ppm
quantmeth None
massfile mascotmasses.txt
reporterions to remove

[path(u'.\\example.mgf')]
There are 461 records for which to create H-Score
```

The files created are then:

example.hmgf (newly synthesized filtered MGF file)
example_hscore.data (h-score data file)

if one wants to filter out the TMT reporter ions from the .mgf file the command line would be:

myprompt$>python Hscorer.py --massfile mascotmasses.txt --tolerance 20.0 --units ppm – quantmeth tmt