Search Compare Cache Files and the Raw Data Extraction Daemon Improve Quantification Analysis Support and Ease of Protein Prospector Installation

Peter R. Baker[1], Juan A. Oses[1], Bing Gao[1] and Robert J. Chalkley[1]

[1]Mass Spectrometry Facility, Dept. of Pharmaceutical Chemistry, University of California, San Francisco, USA

Introduction

Previous to v6 Protein Prospector's Search Compare analyses which included quantitation were forwarded by the web server to a Windows server with a full Prospector installation. The problems with this approach were:

1). Search Compare is considerably slower on Windows;

2). No job control was available, so the server could become overloaded;

3). There was no way to parallelize the quantitation requests;

4). Search Compare needed to be running during the raw data extract requests which could take several hours;

5). A full Windows Prospector installation needed to be installed and maintained.

6). If extra report columns or a tab delimited report were required the quantitation had to be rerun.

7) Normal Search Compare runs of large datasets are also often quite slow. If extra columns are added or you want to switch from say a protein to a peptide report it takes the same amount of time to run the program on subsequent occasions. This processing also took place if you click on a protein to get a peptide report.

8). It was difficult to extract the information from a Protein Prospector report as part of a processing pipeline.

Methods

The Windows Prospector installation has been replaced by a raw data extraction daemon installed as a service which waits for new jobs appearing in a directory shared between the Windows and LINUX nodes. This serves requests for both quantitation data sets and single spectrum display. A setup file can set the maximum number of concurrent quantitation extractions with remaining ones being placed in a queue. Vendor software such as the Thermo MSFileReader or Sciex Analyst needs to be installed as required. Search Compare request are now fully run on the LINUX server and a cache file option in JSON format is available which supports a wide range of different reports once created. A cache file can be created which corresponds to a Search Compare report from multiple combined data sets.

Conclusions

- The new cache file option allows reports that previously would have taken several hours to be completed in a few seconds. The makes the package much more interactive.
- The new features for supporting raw data access, quantitation and JSON cache are now available in Protein Prospector versions on the web and for local installation.
- We will shortly support quantification of Bruker timsTOF data.
- A parallelized version of the raw data extraction method is also under development.
- One potential use of the JSON cache files is to allow customized post processing modules. We are investigating using this for labile crosslinkers such as DSSO.

Acknowledgements

•This work was supported by the Dr Miriam and Sheldon G. Adelson Medical Research Foundation

Search Compare Cache File Options

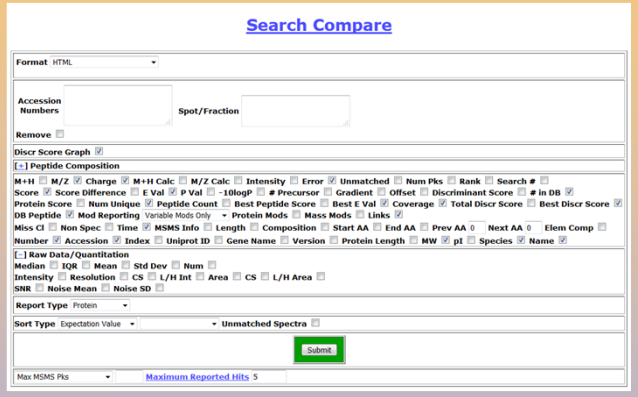

Fig. 1 Search Compare options still available after cache file created.

After a cache file has been created most of Search Compare options are still available. The options that are not longer available are:

- Items related to score and FDR thresholds.
- Quantitation type or peak fitting options.
- A Protein cache file can only generate a protein report.
- A Peptide cache file can generate a Protein, Peptide or Modifications report.
- A Time cache file can generate a Protein or a Time report.
- A Time cache file without unmatched spectra can't show them.
- A Crosslinking report can generate a Protein, Peptide, Modifications or Crosslinking report

Selecting and Deleting Cache Files

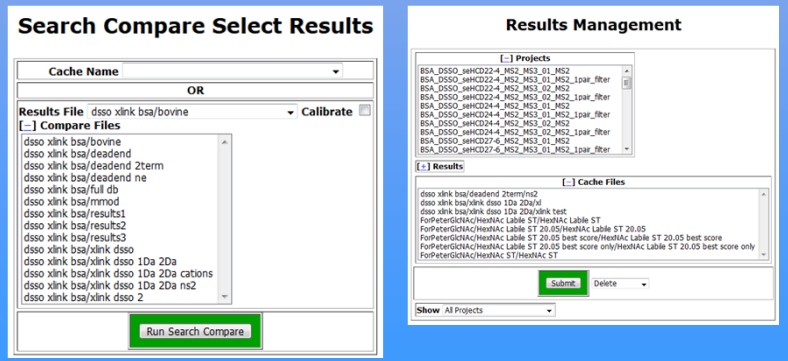

Fig. 2 Selecting and deleting cache files.

JSON Cache File

There was previously no Search Compare output option which comprehensively recorded search result s in a manner which was convenient for use in subsequent processing.

- HTML contains display formatting information and is thus difficult to parse. It only contains the information you asked for on the Search Compare form.
- Tab Delimited output is highly redundant. It also only contains the information you asked for and doesn't contain things like peak list information or search parameters.
- There are also library formats, ouput options for MS-Viewer and for pepXML. These also only contain what is necessary to fulfil the standards.

JSON is a simple, human readable, compact, text based data interface format which uses key/value pairs. Arrays are held in square brackets.

A Protein Prospector JSON cache file can fully recreate a Search Compare report in a small fraction of the time it took to generate the original report. As it also stores the data for columns you have the option to view these after creating a cache file.


```
"project_info": {
        "dsso xlink bsa": {
                "base_dir": "G:/prospector/repository",
                "centroid_paths":
["x/b/xbxhwdgdd2/batchtag/data/2020_04/VsAP1PG7GU5DWbcW/0815_07.mgf","x/b/xbxhwdgdd2/batchtag/data/2020_04/VsAP1PG7GU5DWbcW/0815_08.mgf"],
                "fraction_names": ["0815_07","0815_08"],
                "num_spectra": {
                        "id": 45131
                }
        }
},
```


Fig. 3 Part of a JSON cache file.

Examples of Improved Performance

	SILAC	2plex Neucode
Number of fractions	1	41
Number of spectra	67816	658994
Raw file size	1.5 GB	71.3 GB
Peak list file size	253 MB	1.3 GB
Number of proteins (1%/1% FDR)	1559	3752
Number of peptides (1%/1% FDR)	8861	24182
Number of quantitation peptides	6017	17473
Initial search time	46m	3h 40m
Quantitation time	1hr 35m	16hr 59m
Report time with no quantitation	21s	3m 50s
Cache file display time peptide quan	3s	5s
Results File size	27.8 MB	304 MB
Temporary raw exclusion file size	285 MB	627 MB
Cache file size/Compressed file size	12.1 MB/1.3 MB	36 MB/4.4 MB

Fig. 4 Some performance statistics for 2 example quantitation data sets.