The Benefits of Recycling – Protein Prospector: the Eco-Friendly Search Engine

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

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

Introduction

Users would often like to search for a wider range of amino acid modifications than is usually practical in a typical database search program. Protein Prospector can now restrict searches to either proteins, peptides or sites in which particular modifications have been previously identified. The sites are stored in an sqlite database which is associated with a particular FASTA database. The database can be populated from publically available site databases (eg UniProt dat files, peff format databases[1], published site databases[2]) or from previous search results. The site database options can be combined with other variable modification parameters such as motifs. Using such a database has the potential to greatly decrease search times and improve false discovery rates.

A further new feature is that each user can save multiple different sets of database search form settings. Thus complex sets of search parameters, such as may be used for glycosylation searches, only need to be specified once.

Populating a Site Database

The FA-Index program can be used to create a site database from UniProt dat files (see panel) or peff files. The modifications of interest are selected from a menu.

Modifications can be added to the site database from search results by creating a tab delimited report with columns containing the required information.

For example the Search Compare program now has a modifications report (Fig. 1) which reports the best spectrum associated with a particular modified site assignment. The spectrum reported is the one with the best SLIP score[3] for a given site modification of the hits that pass the various score thresholds. The report can be filtered so that it only includes the modifications of interest. The sites that are already in the database are listed.

Site Database Constrained Searches

A large (2,187,704 MSMS spectra) unenriched, SILAC labelled data set was chosen to test the new search options (Fig. 2). Murine ES cells were left untreated (K0R0) or treated with 100ng/mL of cycloheximide (CHX), for 1 h (K4R6) or 3h (K8R10). Cells were harvested, lysed in 8M urea and protein concentrations were estimated using BCA. Equal amounts of the 3 SILAC labeled samples were combined, treated with DTT and iodoacetic acid, then digested o/n with trypsin (2% of total protein) at 37C. Peptides were extracted using SepPak cartridges, and 200 ug of tryptic peptides was resuspended in 20 mM ammonium formate pH 10.3 and separated in a 2 to 30 % MeCN gradient in the presence of 20 mM ammonium formiate pH 10.3. 70 fractions were collected, and 18 of these fractions were analyzed by LCMSMS using 6 h gradients in a C18 RP monolithic column, interfaced with an EasySpray emitter to a QExactive plus.

3 database searches were performed. The parameters common to the 3 searches were:

Database=SwissProt.2015.12.1, taxonomy = MUS MUSCULUS (16740 entries), max missed cleavages=2
Tolerance precursor=20ppm, fragment=30ppm, Constant mod=Carboxymethyl (C)
Max variable mods=2
Acetyl (Protein N-term),Acetyl+Oxidation (Protein N-term M),Gln->pyro-Glu (N-term Q),Met-loss (Protein N-term M), Met-loss+Acetyl (Protein N-term M),Oxidation (M), Label:13C(6) (R) - Label 1, Label:13C(6)15N(2) (K) - Label 2, Label:13C(6)15N(4) (R) - Label 2, Label:2H(4) (K) - Label 1

The additional variable mods used for the site database search were:

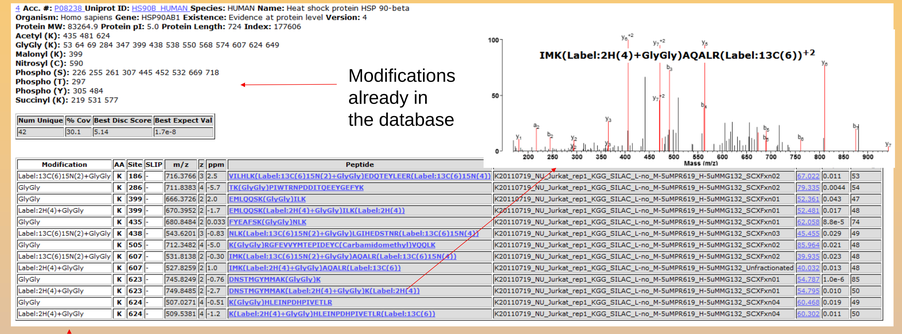
Acetyl (K) – Peptide, Dimethyl (K) – Peptide, Dimethyl (Uncleaved R) - Peptide, Methyl (K) - Peptide, Methyl (R) – Peptide, Nitro (Y) – Peptide, Oxidation (P) – Peptide, Phospho (STY) – Peptide, Succinyl (Uncleaved K) – Peptide, Sulfo (Y) – Peptide, TriMethyl (Uncleaved K) – Peptide

The additional variable mods used for the standard multiple mod search were:

Acetyl (K), Dimethyl (K), Dimethyl (Uncleaved R), Methyl (K), Methyl (R), Nitro (Y) , Oxidation (P), Phospho (STY) , Succinyl (Uncleaved K), Sulfo (Y) , TriMethyl (Uncleaved K)

The results are shown in the panel to the top right.

Search Compare Modifications Report



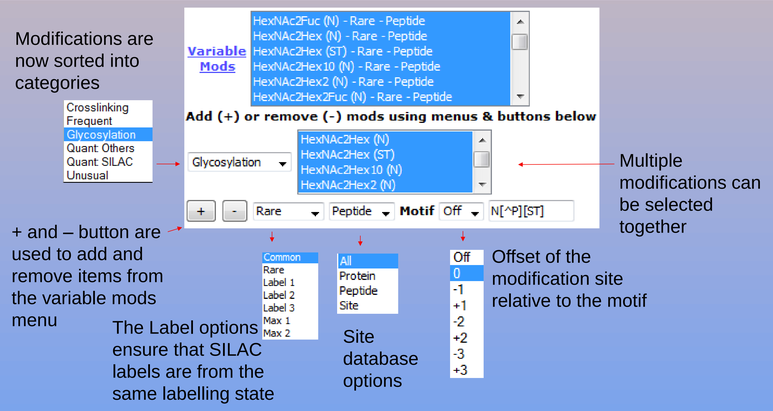
The report is sorted by modification and then by site. Modifications may be grouped to deal with eg Glycosylation or labelling.

The best spectrum supporting a given modification based on SLIP score and subject to score thresholds. Multiple sites can be represented by the same spectrum

The modifications of interest can be selected from a menu on the Search Compare form

Fig. 1 The Search Compare Modification Report (data from Udeshi et al 2012 [4]).

Variable Modification Options



There are 4 options that constrain database searches for a particular modification to those in the site database (Fig 2).
1). **All** – All potential sites are considered. This is the standard setting.
2). **Protein** – Modification only considered for proteins in which it occurs in the site database.
3). **Peptide** - Modification only considered for peptides in which it occurs in the site database.
4). **Site** - Modification only considered for sites in which it occurs in the site database.

Fig. 2 Batch-Tag Variable Modification Menu Items

What's in a UniProt dat File

- 549832 protein entries
- 213211 sites listed from 178 different modifications (excluding glycosylations)
- Some common modifications: Phospho S – 101088, Acetyl K – 22690, Phospho T – 22136, Phospho Y 9281, Succinyl K – 6905, PyridoxalPhosphate K – 6544, Oxidation P – 2308
- Modification names (eg (3R,4R)-3,4-dihydroxyproline) need converting to Unimod names (eg Dioxidation P)
- Some modifications such as GlyGly have no entries. Some entries correspond to common artifact modifications
- Acetylated N-termini are listed as modified amino acids
- The size of the SwissProt dat file 2.73 GB, Size of corresponding sqlite site database 12.6 MB
- Full UniProt dat file (including TrEMBL) is 170 GB

Results

	Standard Mods + 4 SILAC Labels	Multiple Mods + 4 SILAC Labels	
		With Site Database	Without Site Database
Proteins	6382	6389	5680
Search Time (min)	99.5	170	502
Search Space Factor*	1	1.03	7.01
Acetyl K		25	33
Dimethyl K		3	49
Dimethyl R		14	35
Methyl K		3	126
Methyl R		9	59
Nitro Y		3	39
Oxidation P		1	230
Phospho S		492	496
Phospho T		54	79
Phospho Y		1	2
Succinyl K		0	24
Sulfo Y		0	6
Trimethyl K		0	13

* The Search Space Factor is the effect the search parameters have on expectation value relative to a search which just looks for standard modifications. It is calculated by comparing the number of precursor hits for peptides found in all the searches.

Conclusions

- The site database search enabled the detection of a large number of biologically significant PTMs without significantly effecting the protein detection rate
- The search time was doubled. The standard multiple mod search was 5x longer.
- On inspection many of the hits from the standard multi mod search were found to be hits to the correct peptide but the wrong modification state. For example some of the Oxidized P hits were actually Oxidized M or W or peptides with 2 Lys8 SILAC modifications
- The Protein variable modification option is likely to be useful for modifications that are specific to particular proteins and where it is suspected that all the sites have yet to be found
- One problem of using Site rather than Peptide in a search is that it requires all the modifications of a particular type to be detected unambiguously
- There were problems with using SQLite database on systems with nfs mounted disks. These need to be resolved before the software can be released
- There are likely to be some database maintenance issues if user results are used to build site database. This is because site positions can change if the database entries are edited
- See poster 365 on the use of a site data for glycopeptide analysis

References

1. Binz P. et al, A Common Sequence Database Format in Proteomics, **J Proteomics & Bioinformatics**, 01/2008; DOI: 10.4172/jpb.s1000170
2. Keegan S, Cortens JP, Beavis RC, Fenyö D, g2pDB: A Database Mapping Protein Post-Translational Modifications to Genomic Coordinates. **J Proteome Res**. 2016 Mar 4;15(3):983-90.
3. Baker PR, Trinidad JC and Chalkley RJ, Modification Site Localization Scoring Integrated into a Search Engine. **Mol Cell Proteomics**, Vol. 10, No. 7 (2011)
4. Udeshi, et al. Methods for quantification of in vivo changes in protein ubiquitination following proteasome and deubiquitinase inhibition. **Mol Cell Proteomics**. 11(5):148-59 (2012).

Acknowledgements

This work was supported by grant number NIGMS 8P41GM103481 and the Howard Hughes Medical Institute.