Byonic™: Wildcard Search™
Marshall Bern, Ph.D., Protein Metrics Inc. (www.proteinmetrics.com)
October 2014

Benefits summary
- Understand how Byonic’s Wildcard search can identify unexpected modifications
- Best practices for how and when to use a Wildcard search

Method

Protein Metrics’ Byonic™ software allows users to customize proteomics searches to the types and prevalences of modifications in the sample using a capability called Wildcard search that is not found in other search engines. Wildcard search enables the user to find sequence variants, known but unanticipated modifications, and completely unknown modifications, that is, unexplained mass deltas. Wildcard search is especially useful in studying protein samples with insufficient databases (antibodies, environmental samples, venom, etc.) and highly modified proteins (eye lens, oxidative footprinting, etc.), but it is also useful for more routine proteomics to guard against false negatives—high quality spectra going unidentified due to unforeseen modifications.

As shown in Figure 1, the user sets Byonic’s Wildcard search on the Modifications tab by checking the box and inputting a mass range. Byonic will then consider any one modification within the mass range on any amino acid residue in each peptide. A mass range of −30 to +60 Da covers most low-mass modifications: cation adducts, methylation, dimethylation, acetylation, over-alkylation, and so forth. A mass range of −30 to +210 Da covers larger mass modifications (glycation, O-GlcNAc, DTT artifact, etc.) but takes roughly 2.7 times as long.

The exact modification masses considered by a Wildcard search depend upon the user-supplied precursor mass tolerance. For high-resolution MS1, defined to be precursor tolerance < 0.1 Da or 100
ppm, Byonic obtains the modification mass from the precursor mass; for example, the peptide AEFVEVT2K^{2+} has monoisotopic mass 924.504, so Byonic will apply a modification of mass of 27.996 to each residue in turn to match this peptide against a spectrum with a measured precursor mass of 952.5 (= 924.504+27.996). For low-resolution MS1, defined to be precursor tolerance $\geq 0.1$ Da or 100 ppm, Byonic uses exact integers from −99 to +99 Da, a mass defect of 0.05 for losses or gains in the range of 100 to 200 Da, a mass defect of 0.10 for losses or gains in the range 200 to 300 Da, and so forth.

Wildcard searches are slow, so it is best to use a focused database and also cut down the list of known modifications as much as possible. In Figure 1, we set Total common max to zero, so that only the rare modifications apply. Another way to reduce the search space is to restrict the Wildcard to certain residues. The Restrict to residues box accepts single-letter amino acid abbreviations, along with n and c for peptide N- and C-terminus.

A Wildcard search with a range of −100 to +100 Da is not the same as setting the precursor mass tolerance to 100 Da. A Wildcard search sets a value for the mass delta and then places it on each residue in each candidate (and hence run 20 times longer for 20-residue peptides), but a wide precursor window simply admits more unmodified candidates. In the case that a spectrum contains a peptide with a novel modification, the Wildcard search will give a good estimate of the mass delta and site localization, but the wide-window search will require subsequent analysis to determine the same information. For non-labile modifications, the Wildcard search will be more sensitive as well, because it will look for fragment ions carrying the modification.

We use Wildcard searches early in data analysis in a role similar to Preview, to determine which known modifications to enable in the “main” search. In this case, individual spectrum identifications are unimportant; the user is looking for prevalent PTMs and artifacts. We use Wildcard searches late in data analysis to identify publication-quality individual spectra that contain sequence variants and unusual PTMs.

**Example**

We obtained a data set from a Thermo Fisher Q-Exactive instrument running a tryptic digest of mouse brain tissue. The data set (the same one used for the application note on mass tolerances) contains ~60,000 scans over a 180-minute LC gradient. Figure 1 shows the modification settings for a Wildcard search with a mass range of −30 to +210 Da; the other settings were fully specific digestion, a maximum of two missed cleavages, 8 ppm precursor tolerance and 15 ppm fragment tolerance. (The fragment tolerance for a Wildcard search should be set a little looser than normal for high-resolution MS2, because Wildcards take their exact masses from the precursor measurement.) The search against a protein database with ~2600 target proteins and an equal number of decoys took ~2 hours; with Total common max set to 1 instead of 0 the search would have taken at least 10x longer.
Figures 2 – 4 show some example spectrum assignments from this search along with our interpretations of the Wildcard masses. Most Wildcard masses are interpretable, but it takes some skill and experience to map the Wildcard to the most likely sequence variant, known modification, or combination of sequence variants and modifications. Wildcard search occasionally turns up new and interpretable modifications, not yet in the literature; we have published D/E [+15.011] (hydroxamic acid) and M [+33.969] (homocysteic acid). And surprisingly often, Wildcard search turns up “mystery modifications” that resist interpretation.

Figure 2. Three Wildcard modifications on methionine. Top is clearly oxidation (true mass 15.9949); middle is carbamidomethylation (true mass 57.0215) on an unexpected residue, not listed in Unimod; bottom is carboxymethylation on protein N-terminus (true mass 58.0055), which is most likely an in vivo modification because iodoacetic acid was not used in sample preparation.
Figure 3. Top shows N-terminal propionaldehyde modification (true mass 40.0313). Bottom is most likely AFVH[+57.0215]W[+15.9949]YVGEGMEEGFSEAR, over-alkylation and oxidation. There is a limit of one Wildcard modification per peptide, so some Wildcards are really combinations of two or more modifications. High mass accuracy and manual analysis can often tease apart the combinations.
Figure 4. Top shows HexNAc (true mass 203.0794). In this case, due to the close mass match, Byonic “recognized” the Wildcard as glycosylation and then annotated oxonium ions and ~y11 to ~y15, the y-ions with neutral loss of HexNAc. Bottom shows an A → V substitution (true mass delta 28.0313). MSVFIGNSTAIQELFK matches tubulins from other organisms (e.g., schistosomes), but not mouse. This same sample also includes the usual mouse sequence MSATFIGN...