Supplementary Materials

MS-GFDB Search Parameters For 6-protein Data

Raw MS/MS spectra from all digests except CNBr were searched with MS-GFDB[1] v7747 using the following parameters: carbamidomethylation (+57 Da) protecting group on cysteins, one allowed ^{13}C, one allowed non-enzymatic termini, specific protease (Tyrp, chymotrypsin, LysC, GluC, AspN, or ArgC), high accuracy LTQ instrument ID, 30 ppm parent mass tolerance, and target-decoy used for FDR calculation. Considered modifications were: M/W+16 (Oxidation), N-terminal Q-17 (Pyroglutamate formation), and N+1 (Deamidation). For the CNBr digest, the same parameters were used except the non-specific protease option was specified along with 2 allowed non-enzymatic termini. CNBr searches also optionally considered the known modifications M/V-30 (homoserine) and M-48 (homoserine lactone). The appropriate fragmentation methods were also set (pair and triplet searches invoked the “Merge spectra from the same precursor” parameter).

To accurately compute FDR for spectrum IDs, a larger decoy database was appended to the small 6-protein database. The merged database consisted of the reference sequences for the six target proteins, common laboratory contaminants, proteolytic enzymes used, and the full proteome of Drosophila melanogaster (including isoforms and excluding fragments, total size of 17MB) downloaded from the UniProt web site on July 16, 2010[2]. Spectra were then separated by precursor charge (2, 3, and 4 or greater) and filtered by 1% spectrum-level FDR where any hit to a Drosophila sequence was considered a false positive.

PepNovo+ Training Procedures

High resolution ETD and HCD data were generated in conjunction with the 2011 ABRF-iPRG study on SCX fractions from the Lys-C digest of a yeast lysate. Aliquots from fractions 9-12 were used to train PepNovo$^+$ models. An aliquot of fraction 10 was used in the 2011 ABRF-iPRG study. The data were obtained on the same instrument as the 6 protein mixture data set. The iPRG dataset was collected with the same mass spectrometer parameters described in the main methods section except CID was not performed, the top 6 precursors were fragmented, the dynamic exclusion time was 50 sec, and the instrument had not yet been upgraded to the generation 2 ion optics. The LC gradient was also shallower. The 133 min run time 0.1% formic acid/acetonitrile gradient eluted peptides with ~10-20 sec chromatographic peak widths using the following steps: load at 3%B; elute with 5-35%B in 90 min, elute with 35-90%B in 10 min; wash at 90%B for 9 min. The original data files and a table detailing the LC gradient mass spectrometer acquisition may be downloaded from ftp://ftp.broadinstitute.org/distribution/proteomics/public_datasets/Guthals_JPR_2013.

Data used to train PepNovo$^+$ models:

The undigested Saccharomyces cerevisiae lysate, Reference Material (RM) 8323, was obtained from the National Institute of Standards and Technology (NIST) and is described at https://www-s.nist.gov/srmors/view_report.cfm?srm=8323. The S. cerevisiae, strain BY4741, was grown at Boston Biochem Inc. (Cambridge, MA) in rich (yeast peptone dextrose) medium and harvested by continuous-flow centrifugation. The cell pellet was then washed twice with ice-cold water, and lysed by incubation with ice-cold trichloroacetic acid (10 mL/L) in water for 1 h at 4 °C. The precipitate was collected by centrifugation, washed twice with 100 mL/L water in acetone, and pelleted again. The lyophilized yeast lysate was homogenized at NIST through manual grinding. The ground yeast lysate powder was suspended in 50 mM ammonium bicarbonate containing 6 M urea, pH 7.85. After gently stirring at 5 ºC overnight, the yeast lysate solution was filtered through a 0.22 μm cellulose acetate filter. To remove urea from the yeast lysate solution, the solution was thoroughly dialyzed (6,000 Da to 8,000 Da cutoff) at 5 °C using 50 mM ammonium bicarbonate as the dialysis buffer.

The iPRG started with 200 ug of frozen yeast lysate and vacuum centrifuged to almost dryness, then resuspended in 6M Guanidine HCl / 25mM ammonium bicarbonate (AmBic), reduced with 2mM TCEP, then alkylated with 5mM iodoacetamide. After diluting to 3M Guanidine HCl/25mM AmBic, 8ug LysC (Wako) was added for overnight digestion, then further diluted to 2M Guanidine HCl/25mM AmBic with an additional 4ug LysC (Roche) and digested overnight again.

The Lys-C digest was desalted and fractionated by strong cation exchange (SCX) chromatography. The SCX separation was done using a Polysulfoethyl A column from PolyLC (200 x 2.1 mm, 5µm, pore size 200A), 13 peptide containing fractions were collected as the salt gradient went from 1-50% B in 40 min at 200 ul/min. Buffer A: 10mM ammonium formate, 25% acetonitrile, pH 3. Buffer B: 500mM ammonium formate, 25% acetonitrile, pH 6.8. A total of 9,857 charge 2 PSMs (8,616 unique peptides) and 5,506 charge 3 PSMs (4,907 unique peptides) were used to train the CID model. 58,892 charge 2 tryptic PSMs (12,704 unique peptides) and 5,980 charge 3 Lys-C PSMs (3,016 unique peptides) were used to train the HCD model. Finally, 5,849 charge 2 tryptic PSMs (5,252 unique peptides), 7,127 charge 3 Lys-C PSMs (3,741 unique peptides), and 4,203 charge 4 Lys-C PSMs (2,095 unique peptides) were used to train the ETD model. All MS/MS spectra were deconvoluted[3] prior to training. After training the models, performance was measured on the 6-protein data set for all enzymes in terms the percentage of observed breaks and explained score (metrics defined in manuscript). It was then determined that spectra from tryptic peptides yielded better scored spectra when training HCD and ETD charge 2 models while spectra from Lys-C peptides yielded better scored spectra when training HCD charge ≥ 3 and ETD charge ≥ 3 models.

The following parameters were used by MS-GFDB: carbamidomethylation (+57 Da) protecting group, one allowed ^{13}C, specific protease (Tryp or LysC), one allowed non-enzymatic termini, high accuracy LTQ instrument ID, 7 ppm (for Frese et al spectra) or 30 ppm (for remaining spectra) parent mass tolerance, target-decoy used for FDR calculation, CID, HCD, or ETD fragmentation method (depending on type of spectra), and the optional modifications M/W+16 (Oxidation), N-terminal Q-17 (Pyroglutamate formation), and N+1 (Deamidation). The IPI human database v3.87 was used for spectra from human samples and the UniProt yeast database was used for remaining spectra.

Clustering

After merging CID/HCD/ETD spectra from the same precursor, merged spectra from the same peptide were clustered as well. Clustering of CID MS/MS spectra is typically done by MS-Cluster[4] before PepNovo^{+} scoring, but here a different approach was adopted to handle any combination of CID/HCD/ETD PRM spectra, using a simplified version of the hierarchical clustering algorithm described by Frank et al[4]. Initially, every set of CID, HCD, and/or ETD PRM spectra from the same precursor represented its own cluster C with its merged PRM spectrum S (as described in the Methods section in the paper). Given a pair of cluster PRM spectra (S_1, S_2) from a pair of clusters (C_1, C_2), similarity between the clusters was evaluated as $matchScore(C_1, C_2)$ = $min($ matched PRM score in S_1/total PRM score in S_1 , matched PRM score in S_2/total PRM score in S_2). Given a threshold η, the clustering procedure iterated over the following steps:

```
[1] Find all pairs $(S_1, S_2)$ with equal parent masses $M_1 = M_2$ (within 20 ppm parent mass tolerance) and $matchScore(C_1, C_2) \geq$ η. If no pairs were found then exit.
[2] Initialize the set $visited \leftarrow \emptyset$
[3] For each pair $(C_1, C_2)$ in order of decreasing score $matchScore(C_1, C_2)$:
    a. if $C_1 \in visited \vee\ C_2 \in visited$, then continue to next pair
    b. $visited \leftarrow visited \cup \{C_1, C_2\}$
    c. $C_1 \leftarrow C_1 \cup C_2$
    d. Remove $C_2$ and $S_2$
    e. Use steps 1-6 from the CID/HCD/ETD merging procedure to re-compute $S_1$ from all unique CID/ETD and/or HCD/ETD pairs in $C_1$
[4] Repeat from [1]
```

η was chosen to be conservative (0.72) to avoid clustering spectra from different peptides, which could have introduced de novo sequencing errors in down-stream alignment/assembly steps (>99% of all clusters containing two or more triplets that were separately identified by MS-GFDB had all matching peptide IDs). When propagating MS-GFDB PSMs to compute sequencing accuracy, the representative peptide match for a given cluster was the most common over all identified spectra (CID, ETD or HCD) in the cluster (ties were broken by selecting the PSM with the lowest FDR per peptide match).

Meta-SPS Parameters

The following parameters were used when sequencing CID/HCD/ETD, CID/ETD, and HCD/ETD: 0.05 maximum allowable p-value of PRM spectral alignments (SPS fits the distribution of alignment scores to a Gaussian curve to approximate p-values), 3.35 minimum overlap score between SPS contigs during the second stage of alignment/assembly[5], and 0.04/0.1 peak/parent mass Da tolerance. Since MS/MS deconvolution[3] supports ppm tolerances, spectra were deconvoluted with 10 ppm peak tolerance (all remaining steps used a fixed Da tolerance of 0.04). Remaining parameters were set to their default values. For paired CID/HCD, we used the same parameters except for 0.045 max p-value (as was done previously for separate CID and HCD[3]).

References

1. Kim S, Mischerikow N, Bandeira N, Navarro JD, Wich L, Mohammed S, Heck AJR, Pevzner PA (2010) The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Molecular & Cellular Proteomics* 9:2840–2852.

2. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh L-SL (2008) The Universal Protein Resource (UniProt). *Nucleic Acids Research* 35:D190–D195.

3. Guthals A, Clauser KR, Bandeira N (2012) Shotgun protein sequencing with meta-contig assembly. *Molecular & Cellular Proteomics* 10:1084–96.

4. Frank AM, Bandeira N, Shen Z, Tanner S, Briggs SP, Smith RD, Pevzner PA (2008) Clustering millions of tandem mass spectra. *Journal of Proteome Research* 7:113–122.

5. Bandeira N, Clauser KR, Pevzner PA (2007) Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins. *Molecular & Cellular Proteomics* 6:1123–34.