Supporting Information

Combinatorial Libraries of Synthetic Peptides as a Model for Shotgun Proteomics

Brian C. Bohrer[1], Yong Fuga Li, [2] James P. Reilly,[1] David E. Clemmer,[1] Richard D. DiMarchi,[1] Predrag Radivojac,[2] Haixu Tang,[2] and Randy J. Arnold[*,1]

[1] Department of Chemistry, Indiana University, Bloomington, IN 47405
[2] Department of Informatics and Computing, Indiana University, Bloomington, IN 47408
[*] To whom correspondence should be addressed. e-mail: rarnold@indiana.edu

Estimation of False Discovery Rate (FDR)

Searches of the *D. radiodurans* data against the reverse *D. radiodurans* database as a decoy resulted in a FDR of 3.9% with a Mascot score threshold of 25. The same type of decoy search for the peptide libraries is insufficient to accurately estimate false discovery rate. Consider the two scenarios for biological versus synthetic peptides. Low quality spectra from biological peptides can potentially match to any peptide (within the precursor tolerance) from any protein in the organism database. For these peptides, when searched against a sufficiently large database, random matches are reasonably likely at low score thresholds and decoy (i.e. reverse) databases are sufficient to model these random matches. Alternatively, spectra from synthetic peptides can only match to a considerably smaller set of known sequences, and so even at very low score thresholds, there are not incorrect ‘random’ sequences in the database to which poor quality spectra can match. For these reasons, one reasonable estimate of the FDR for the peptide libraries is no higher than the 3.9% we calculate for the *D. radiodurans* data.

Two additional approaches can be used to estimate FDR for the peptide libraries. The first involves searching the peptide library data against the *D. radiodurans* database as a decoy. Table S-1 below illustrates the results from applying this approach and results in an estimated FDR similar to or less than the 3.9% reported above at a Mascot score threshold of 25. As noted in the table, the database has a similar (slightly smaller) number of peptides of the same length as the corresponding library, but in the search, the peptide length for a match is not restricted such that a peptide of any length can match to the spectra from each peptide library. As expected, increasing the score threshold reduces the FDR at the expense of peptide identifications. Results appear to be similar for searches of BB12A data against both forward and reverse *D.*

radiodurans databases as decoys, indicating that the forward database is a sufficient decoy for the FDR estimation.

Table S-1. Average FDR estimates as percentages based on Mascot searches against *D. radiodurans* forward and reverse databases based on ten replicate analyses.

Search	**Score threshold**			Peptides in library	Same length peptides
	25	**30**	**35**		
D. radio vs. rev. *D. radio*	3.9 ±0.6				
BB12A vs. fwd. *D. radio*	2.2 ±0.3	0.9 ±0.2	0.5 ±0.2	4096	2645
BB12A vs. rev. *D. radio*	2.1 ±0.2	1.0 ±0.2	0.5 ±0.1		
BB11A vs. fwd. *D. radio*	1.9 ±0.2	0.6 ±0.2	0.1 ±0.1	4608	3018
BB10A vs. fwd. *D. radio*	2.7 ±0.3	0.9 ±0.2	0.4 ±0.3	5184	3338
BB9A vs. fwd. *D. radio*	4.4 ±0.6	1.8 ±0.2	0.7 ±0.3	3888	3779

Unfortunately, the use of a biological database as a decoy for synthetic peptides does not present the same type of similarity of sequences that exists among the different sequences in these four libraries. Thus, another type of decoy search was performed using additional database entries that include all the possible one amino acid deletions, two consecutive amino acid deletions, and one amino acid insertions (repeating an amino acid, except when the residue was the C-terminal lysine or arginine). These database entries also covered the most likely synthetic errors in generating the peptide libraries. (We suspect that deletion of a single amino acid is the most likely to occur and that a single insertion cannot happen synthetically but would have a similar decoy effect for peptide-spectrum matching.) The results for these searches using a single replicate analysis of each library are shown in Table S-2.

Table S-2. Peptide identifications by Mascot using a score threshold of 25 for peptide libraries searching either the library alone or searching the library in a database containing three decoy libraries for each.

Library	BBxA only	BBxA	1 deletion	2 deletion	1 insertion	Low FDR %	High FDR %
BB12A	1306	1278	84	0	0	2.14	6.57
BB11A	1154	1132	73	6	0	1.91	6.45
BB10A	1077	1056	43	5	0	1.95	4.07
BB9A	1220	1197	113	8	90	1.89	9.44

Notes: Low FDR % = (BBxA only – BBxA)*100/BBxA only; High FDR % = 1 deletion*100/BBxA

Notice that FDR can be estimated in two different ways using the data from these searches. The first, less conservative approach is labeled as "Low FDR %" in Table S-2 and is based on the idea that including an appropriate decoy database (which appears to be the 1 deletion sequences) will reduce the number of spectra matched to the expected BBxA sequences, and the reduction should estimate the false discovery rate. In some calculations, this type of estimate would be multiplied by two to account for the false positives still present as matches to the expected sequences. Thus this FDR estimate is approximately 2% (or, if you prefer 4%) which is on par with the previous estimates shown in Table S1. The second, more conservative approach is labeled as "High FDR %" in Table S2 and is based on the idea that the number of matches to an appropriate decoy database (in this case the 1 deletion sequences) is a good estimate of false identifications in the matches to expected sequences. This estimate provides values from 4 to 9.5% for the different libraries, although it should be noted that the decoy library is as much as 5-fold larger in size than the expected library, which inflates the estimated FDR. Also, it should be noted that all but 19 of the 313 "1 deletion" matches could be determined by manual examination to be incorrect peptide-spectrum matches based on one or more of three main criteria: noisy spectra or peaks matching to noise, unmatched significant peaks in the spectrum, and incomplete fragmentation. While subjective, the manual inspection also revealed that for libraries BB12A, BB11A and BB10A, as many as a dozen different sequence tag mass redundancies could explain the incorrect identifications. Interestingly, similar mass redundancies occurred with the BB9A 1-insertion sequences, but not for the 1-insertion sequences of the other three libraries (thus zero identifications to these decoy sequences).

Hydrophobicity Distributions

The numbers in Table S-3 below correspond to the distributions plotted in Fig. 2 in the main text.

Table S-3. Quartile values of Eisenberg hydrophobicity values for three distributions corresponding to each of the four synthetic peptide libraries. The values for Human correspond to 6000 randomly selected peptides of each length from the human proteome in Swiss-Prot.

File	**Quartile**		
	0.25	**0.5**	**0.75**
BB9A – library	-1.38	0.01	1.35
BB9A – identifications	-1.04	0.27	1.53
Human length 9	-1.95	-0.39	1.08
BB10A – library	-1.05	0.23	1.51
BB10A – identifications	-0.53	0.73	2.04
Human length 10	-1.73	-0.17	1.4
BB11A – library	-0.87	0.46	1.81
BB11A – identifications	-0.70	0.69	2.07
Human length 11	-1.69	-0.03	1.63
BB12A – library	-0.63	0.74	2.10
BB12A – identifications	-0.42	0.99	2.33

Human length 12	-1.67	0.21	1.90

Peptide Identification Saturation Curves

The curves shown in Figure S-1 below are the same data as shown in Figure 3 in the text normalized to the number of unique sequences in each library.

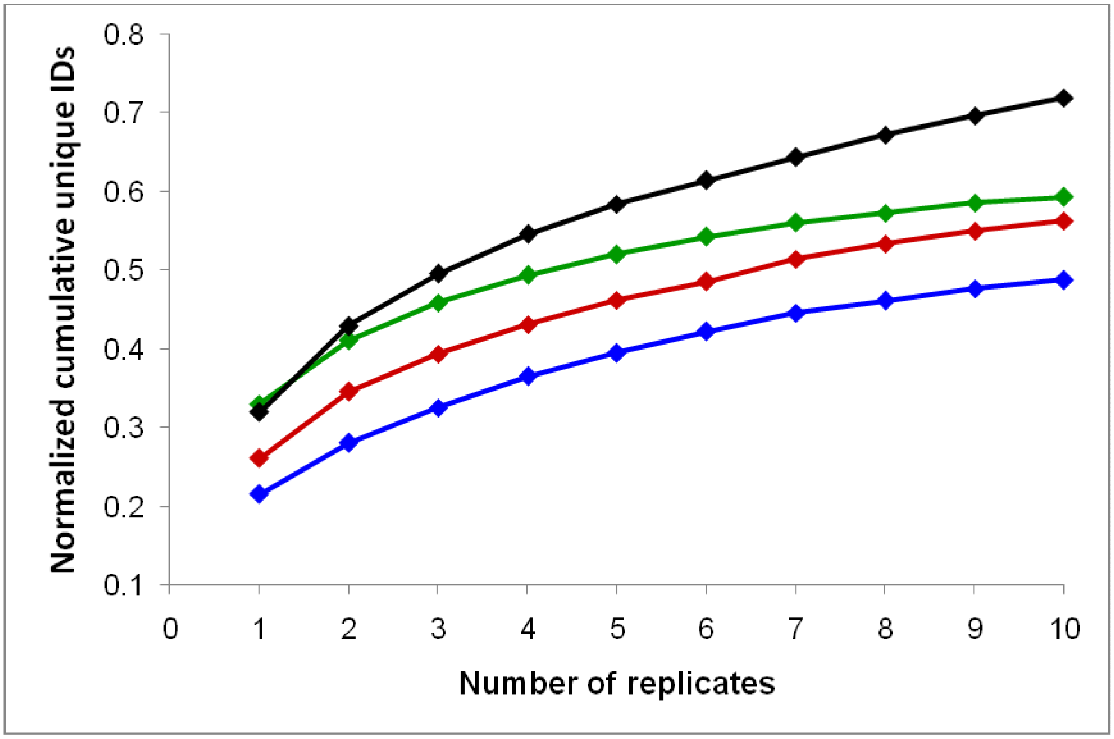

Figure S-1. Normalized cumulative unique peptide identifications obtained upon subsequent analyses up to ten replicates for BB12A, BB11A, BB10A, and BB9A (black, red, blue, and green diamonds, respectively).

Peptide Repeatability and Precursor Signal Strength

Peptide precursor ion signal intensity is an important factor in determining whether a peptide is identified in LC-MS/MS shotgun experiments. The data shown below in Figure S-2 demonstrate that higher peak intensities (and to a greater extent peak areas) are correlated with repeatability.

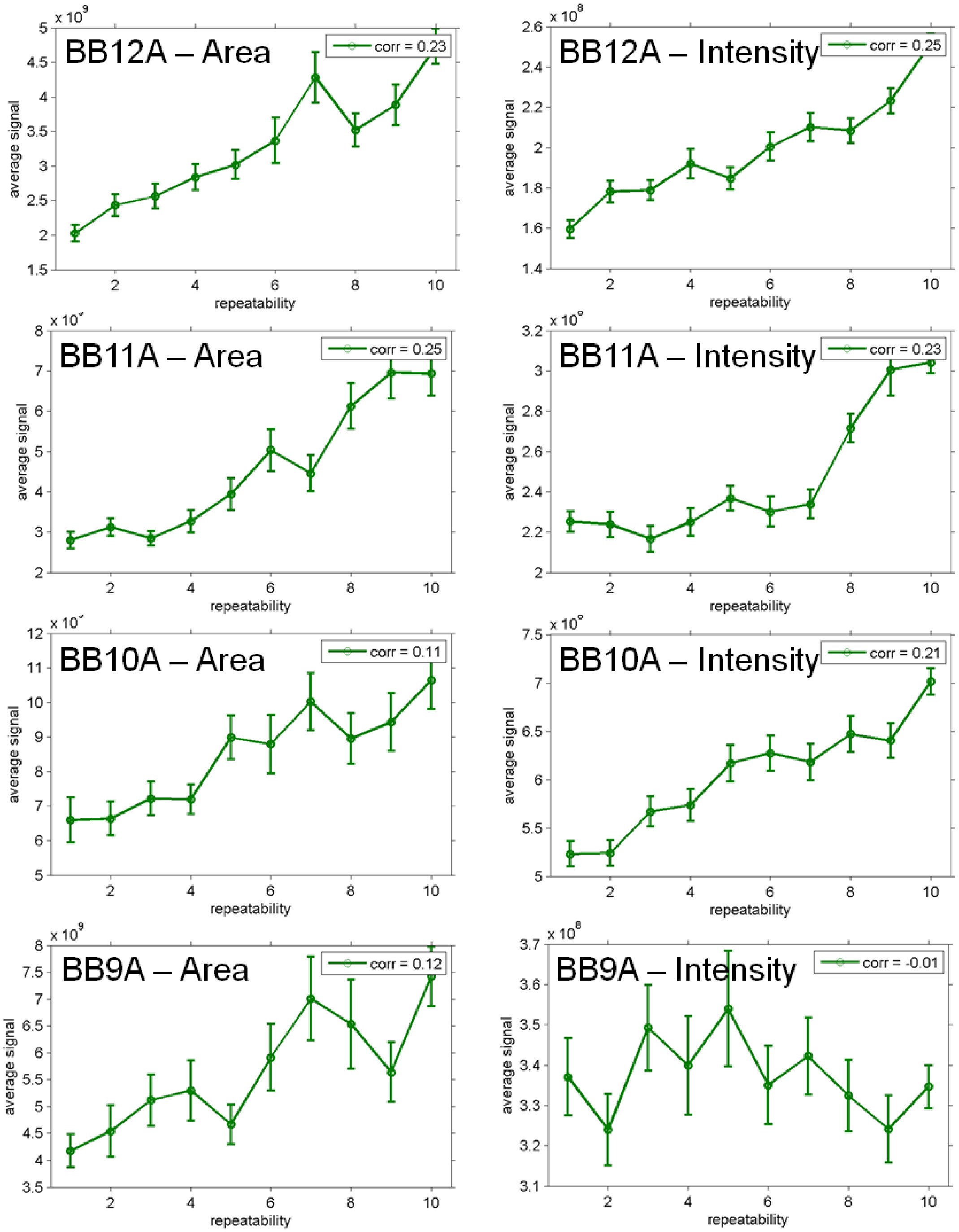

Figure S-2. Mean peptide precursor signal peak areas (left) and apex peak intensities for peptides versus the number of replicates (out of ten) in which the peptide was identified for peptide libraries BB12A (top), BB11A, BB10A, and BB9A (bottom).

Peptide Library Mixtures and Length Bias

As demonstrated in Figure 5 in the text, there is a distinct bias in favor of identifying longer peptides when the synthetic libraries BB9A, BB10A, BB11A, and BB12A are mixed, regardless of the molar ratio. One explanation for this bias is related to the potential for longer peptides to more readily accommodate multiple charges during the electrospray ionization process. Figure S-3 part A below illustrates this bias as the number of peptide identifications for precursor ions of a particular bin of peak intensities. The mean peak intensities, noted in the figure caption, correlate well with the number of peptide identifications for that library (represented by the area under each curve in Figure S-3). Notice that the bias can be alleviated by altering the relative amount of each library in the mixture (part B).

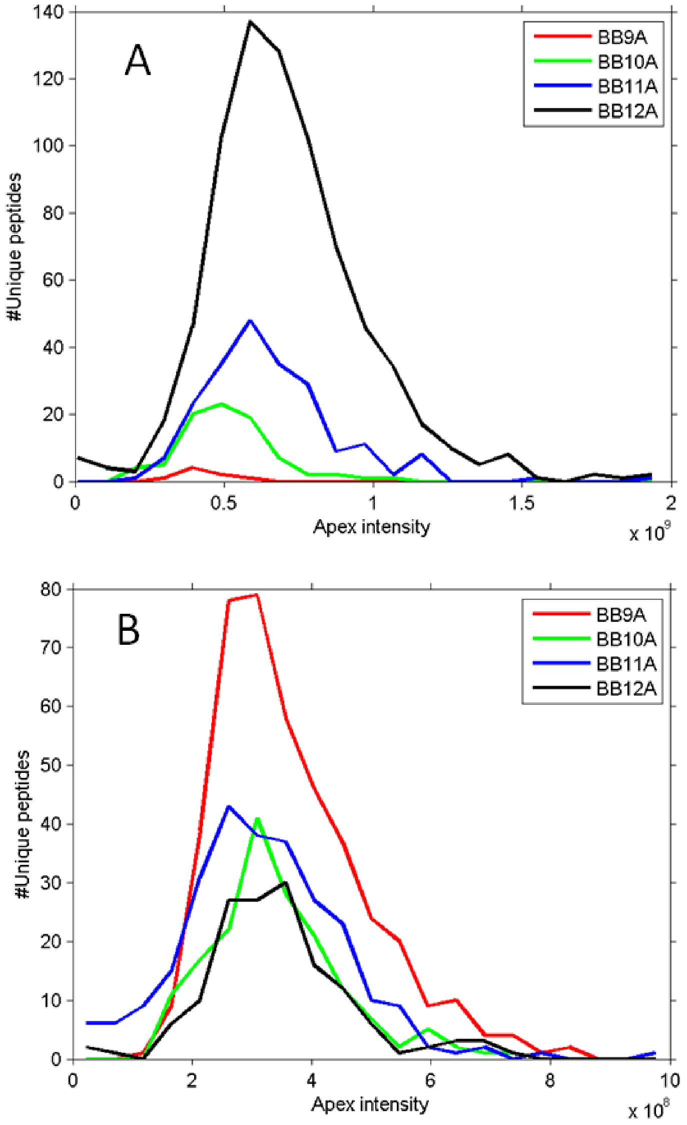

Figure S-3. Peptide identifications plotted versus binned precursor peak intensity for A) the 50:50:50:50 femtomole and B) 200:80:60:40 BB9A:BB10A:BB11A:BB12A library mixtures. Mean precursor peak intensities of A) 4.5, 5.08, 6.52, and 7.09 and B) 3.64, 3.44, 3.18, and 3.43 x10^{8} for libraries BB9A, BB10A, BB11A, and BB12A, respectively, were observed.