Journal of Bioinformatics and Computational Biology
Vol. 12, No. 5 (2014) 1440001 (15 pages)
© The Authors
DOI: 10.1142/S0219720014400010

Imperial College Press
www.icpress.co.uk

Identification of best indicators of peptide-spectrum match using a permutation resampling approach

Malik N. Akhtar[*,**], Bruce R. Southey[*,**], Per E. Andrén[†],
Jonathan V. Sweedler[‡] and Sandra L. Rodriguez-Zas[*,§,¶,||]

[*]*Department of Animal Sciences*
University of Illinois Urbana-Champaign
Urbana, IL 61801, USA

[†]*Department of Pharmaceutical Biosciences*
Uppsala University, Uppsala 75124, Sweden

[‡]*Department of Chemistry, University of Illinois Urbana-Champaign*
Urbana, IL 61801, USA

[§]*Department of Statistics, University of Illinois Urbana-Champaign*
Urbana, IL 61801, USA

[¶]*Institute for Genomic Biology*
University of Illinois Urbana-Champaign
Urbana, IL 61801, USA
[||]*rodrgzzs@illinois.edu*

Received 17 June 2014
Revised 20 August 2014
Accepted 17 September 2014
Published 31 October 2014

Various indicators of observed-theoretical spectrum matches were compared and the resulting statistical significance was characterized using permutation resampling. Novel decoy databases built by resampling the terminal positions of peptide sequences were evaluated to identify the conditions for accurate computation of peptide match significance levels. The methodology was tested on real and manually curated tandem mass spectra from peptides across a wide range of sizes. Spectra match indicators from complementary database search programs were profiled and optimal indicators were identified. The combination of the optimal indicator and permuted decoy databases improved the calculation of the peptide match significance compared to the approaches currently implemented in the database search programs that rely on distributional assumptions. Permutation tests using *p-values* obtained from software-dependent matching scores and *E-values* outperformed permutation tests using all other indicators. The higher overlap in matches between the database search programs when using end permutation compared to existing approaches confirmed the superiority of the end permutation method to identify peptides. The combination of effective match indicators and the end permutation method is recommended for accurate detection of peptides.

[||]Corresponding author.

[**]These authors contributed equally to this manuscript.

Keywords: Neuropeptides; database search programs; tandem MS; *p-value*; significance levels; Ends permuted decoy database; permutations.

1. Introduction

Mass spectrometry discovery has revolutionized proteomic research enabling the characterization and quantification of hundreds of peptides from samples ranging in size and complexity.[1–6] In tandem mass spectrometry (MS/MS) experiments, the peptides present in the sample can be identified by sequence database search programs.[7,8] These programs attempt to match the fragment ions from the observed spectra with the fragment ions from theoretical spectra generated from the known or predicted peptide sequences in the target database. Based on the number of matched fragment ions between observed and theoretical spectra the database search programs calculate scores that reflect the quality of the match between both spectra. Subsequently, these scores are converted into a measure of the statistical evidence supporting the match.[9,10]

Two related components, the match score and the statistical significance assigned to the score (e.g. cross-correlation score and Weibull *p-value* in Crux; and hyperscore and *E-value* in X! Tandem), influence the capability to detect peptides. Database search software differ in the algorithms and assumptions to assess the observed-theoretical spectra match leading to different matching score indicators (e.g. number of matched fragment ions, cross-correlation) and different methods to assess statistical significance of the match. The comparative effectiveness of the scores to capture the match has not been evaluated.

One commonly used approach to convert a specific observed-theoretical spectra match score into a statistical significance value encompasses fitting a specific parametric distribution to all the match scores attained from the target database[11,12] or from decoy peptides generated from the target database matches.[13] Alternatively, significance values can be obtained in a nonparametric fashion from the decoy peptides.[14] A previous comparative study of the database search programs demonstrated that, for some peptides, detection using significance value estimation approaches implemented in the database search programs remains challenging.[7] This situation can be traced back to the low significance levels obtained with existing approaches particularly for short peptides under 15 amino acids in length.[7]

The challenges of peptide identification using existing approaches include false negatives due to match significance levels that do not surpass the minimum detection threshold, false positives due to incorrectly spectra match surpassing the minimum threshold, and missed peptides due to sample complexity leading to multiple peptides present in the single tandem spectrum (also known as chimeric spectra).[7] The bias introduced by the existing approach has major impact in small peptides. These peptides are unlikely to be identified at high significance levels by most database search programs due to limited number of fragment ions to accumulate high matching scores.[7,15,16] The programs assign low significance levels to tandem match spectra that contain incomplete fragmentation (i.e. missing signal peaks) and noise peaks. This is because these spectra can result in peptide matches with low scores that cannot be differentiated from other random matches.[7,17] Likewise, increases in the effective search database size (such as those rising from the consideration of post-translational modifications) can reduce the sensitivity of the algorithms to detect peptides at accurate significance levels.[15]

In the target–decoy approach, observed spectra are matched to theoretical spectra from reverted or reshuffled sequences from the target database together with the original target sequences.[18] The target–decoy approach aims at avoiding stringent significant threshold to control for multiple testing across peptides.[19,20] However, for small peptides, most decoy database construction methods produce few spectra that have more extreme matches, artificially inflating the significance levels. Other decoy databases construction methods that exploit the capability of resampling approaches to generate null hypothesis while controlling the experiment-wise error rate should be evaluated.

The aims of this study were: (1) to compare indicators of observed-theoretical spectra matches and characterize the accuracy of the resulting statistical significance using permutation testing, (2) to develop novel decoy databases including resampling of terminal positions in the peptide sequence and identify the conditions for accurate computation of match significance levels, and (3) to demonstrate the application of the novel decoy approach using popular database search programs.

1.1. *Theoretical-observed spectra match indicators*

Table 1 lists the observed-theoretical spectrum match indicators evaluated and corresponding database search programs: Crux (version 1.37),[21] OMSSA (version 2.1.8),[12] and X! Tandem (version 2013.02.01.1).[11] These programs were selected because their open source nature allowed the retrieval of intermediate match indicators through modification of the source code.

Table 1. Crux, X! Tandem, and OMSSA match indicators used.

Programs	Indicators
Crux	Number of matched *b*- and *y*-fragment ions SEQUEST preliminary (Sp) score Cross-correlation (XCorr) score DeltaCn (ΔCn) score *p-value*: computed from the Weibull distribution using 10^3 XCorr scores
X! Tandem	Number of matched *b*- and *y*-fragment ions Convolution score Hyperscore *E-value*: computed assuming hypergeometric distribution for hyperscores
OMSSA	Number of matched *b*- and *y*-fragment ions Lambda or Poisson mean Poisson *p-value* *E-value*: Poisson *p-value* multiplied by effective database size

In X! Tandem the hyperscore is computed by multiplying the factorial of the number of matched b- and y-fragment ions with the convolution score (dot product of the intensities of the fragment ions common between observed and theoretical spectra). The X! Tandem *E-value* is estimated from the distribution of hyperscores from all the matches of a spectrum in the database. OMSSA uses a Poisson distribution with a mean that is function of the fragment ion tolerance, number of matched fragment ions, and neutral mass of the precursor. The Poisson probability is calculated using the number of matched ions and Poisson mean. This probability is then used to estimate the *E-value* by multiplying the Poisson probability by the effective database size for each spectrum. For the Sp score, Crux takes into account the intensities of the shared fragment ions between the observed and theoretical spectra and the consecutive number of matched b- and y-ions. For the XCorr score, the intensities of the matched ions between observed-theoretical spectra are summed and adjusted using the XCorr scores calculated from a range of shifts in m/z values.

Database search specifications were: (1) mass type: monoisotopic; (2) fragment ion charge: default values; "mz-bin-width": 0.3 (Crux); (3) no post-translational modifications; (4) enzyme: "whole protein" (OMSSA) or custom cleavage site to avoid cleavage of the provided neuropeptide database (Crux and X! Tandem); (5) precursor ion tolerance: 1.5 Da; (6) fragment ion tolerance: 0.3 Da (OMSSA and X! Tandem); and (7) OMSSA "ht": 8 to consider only those database peptides that had one or more fragment ion matching including one of top 8 highest fragment ion peaks in the observed spectrum. The selected specifications follow program settings previously used to evaluate the ability of the database search programs to identify peptides.[7]

1.2. *Observed spectra, target and decoy databases*

The performance of alternative indicators to assign the statistical significance to spectra matches was investigated on a murine linear ion trap (LTQ) tandem spectra dataset.[22] Spectra and peptide identification were obtained from the SwePep database (http://www.swepep.org).[22] The tandem spectra dataset consisted of 80 observed tandem spectra from neuropeptides without post-translational modifications. The majority of the peptides (92%) had precursor charge states +2 or +3. The target database included the 80 peptides with observed spectra studied and all other peptides that could have been produced from the known 95 mouse prohormones including those that produced the 80 peptides studied. The exhaustive list of target peptides was obtained from the PepShop[23] database (http://stagbeetle.animal.uiuc.edu/pepshop) including information from the SwePep, UniProt,[24] and NeuroPred.[25]

To understand the performance of the software under best conditions, optimal spectra (that contains all expected b- and y-fragment ions) were simulated for the peptides in the target database using corresponding precursor charge states. For each spectrum, all b- and y-fragment ions with +1 charge state were simulated with uniform intensity. Additional peaks due to loss of a single ammonia or water molecule were simulated when the *b*- or *y*-ion sequence contained water or ammonia losing amino acids.[7] Due to the presence of all expected fragment ions, the optimal spectra should be detected by the database search programs with high confidence.

The characterization of the spectral match significance is based on various indicators (such as the number of matched ions, Sp score, and *E-values*) obtained from a decoy database generated using permutation.[26] A single target database was created for all database search programs by selecting all peptides within 12 Da (corresponding to 3 m/z ion tolerance with a +4 charge state) of the precursor mass for each tandem spectrum. This mass limit results from the database search programs preselecting candidate peptides based on peptide mass and user-defined mass tolerances. Permutations of each target candidate sequence residues at the N- and C-terminal ends were used to populate the decoy database. The N- and C-terminal ends (one, two, or three positions on both peptide ends) in the target sequences were exhaustively substituted with all mono-, di-, or tri-mer combinations of the 19 standard amino acids to generate decoy peptides. Leucine and isoleucine were treated as the same amino acid in all permutations and comparisons between candidate and permuted sequence. The substitutions only at the terminal ends kept the internal amino acid composition of the target peptides unchanged in the resulting decoy peptides. This terminal permutation strategy generated decoy peptides that were more similar to their target peptides yet disrupted the pattern of *b*- and *y*-fragment ions that are used in matching the observed and theoretical spectra. The terminal regions were selected because the ions from the terminal regions had better sensitivity than the ions from the central region of peptide.

For the accurate assessment of significance levels, the terminal permutation strategy generates informative reference null distributions that are constituted by truly random peptides (different from target peptides). The exact permutation test controls the probability of type I error below a selected alpha level due to the consideration of all random sequences for a target peptide of given amino acid length. However, an exact test can generate sizeable decoy databases and handling such large databases remains challenging due to limitation of the current database search programs.[26] The terminal permutations offer an alternative and computationally feasible approach to generate an exhaustive set of decoy peptides. These decoys, that are used to generate null distributions, are based on the permutation of few selected positions that disrupt the *b*- and *y*-ion patterns of the target peptides.

From the termini permutation strategy, three decoy databases: Ends1, Ends2, and Ends3 were evaluated. Ends1 encompasses $236 * (19$ N-terminal amino acids$)$ $* (19$ C-terminal amino acids$) = 236 * 360 = 84{,}960$ decoy peptides; Ends2 encompasses $236 * (19 * 19$ N-terminal amino acids$) * (19 * 19$ C-terminal amino acids$) =$ $236 * x130{,}320 = 30{,}755{,}520$ decoy peptides; and Ends3 encompasses $236 * (19 *$ $19 * 19$ N-terminal amino acids$) * (19 * 19 * 19$ C-terminal amino acids$) = 236 *$ $47{,}045{,}880 = 1{,}120{,}027{,}680$ decoy peptides. Separate permuted databases were created for each observed spectra in Ends3 due to inability of the database search programs to adequately handle the size of the permuted decoy database.

The target database was appended to each of the Ends decoy databases for the combined target-decoy search strategy. The merging of the target and decoy databases provided unbiased *p-value* estimates and avoided zero *p-values*.[26]

For each observed-theoretical spectra match indicator, the permutation *p-values* were computed as the relative frequency of the sum of the matches in the target-decoy database that had indicator values equal or better than the observed-target spectra matches. A Bonferroni adjusted threshold *p-value* $< 1 \times 10^{-4}$ based on a 1% experiment-wise error rate ($0.01/80 \approx 1 \times 10^{-4}$) was used to compare performance of the different indicators. A sensitivity analysis enabled the assessment of the impact of the *p-value* threshold on the capability of match indicators to detect the peptides. The limited number of observed and annotated spectra prevented unbiased analysis using receiver operating characteristic (ROC) curve.

2. Results and Discussion

A threefold-strategy was used to characterize the performance of spectra match indicators from database search programs to detect peptides. First, optimal simulated spectra were searched against the target database to obtain a baseline performance in the absence of data quality issues such as presence of noise peaks, missing signal peaks, and low signal-to-noise ratio. Second, real spectra were searched against the target database to study the influence of data quality issues on peptide detection significance levels relative to the baseline performance. Third, the performance of the match indicators to detect peptides in realistic scenarios using End-permuted decoy databases was demonstrated.

2.1. *Peptide detection benchmarks using optimal and real spectra against the target database*

Table 2 summarizes the number of peptides detected by the three database search programs at various significance *E*- or *p-value* thresholds when optimal uniform simulated spectra and real tandem mass spectra were searched against the target database.

For the optimal simulated spectra, the three programs accurately detected all peptides at *E*- or *p-value* $< 2 \times 10^{-1}$. At *E*- or *p-value* $< 1 \times 10^{-4}$, the Crux, OMSSA, and X! Tandem detected 9 (11.25%), 80 (100%), and 72 (90.0%) target peptides, respectively. The significance levels of the X! Tandem *E-values* increased linearly with the increase in peptide length and only peptides greater than 8 amino acids in length (hyperscore > 40) reached a significance level of *E-value* $< 1 \times 10^{-4}$. OMSSA *E-values* were less correlated with peptide length or number of matched *b*- and *y*-ions. The minimum *E-value* was 1×10^{-6} and corresponded to an 11 amino acid-long peptide that had a +2 precursor charge state spectrum. The lower significance level of Crux peptide matches, relative to the OMSSA and X! Tandem, have been confirmed previously.[7] At a less stringent threshold *p-value* $< 1 \times 10^{-2}$, Crux

Table 2. Number of peptides matched at various significance levels of the $\log_{10}$-transformed *E*- or *p-values* (rounded down to the nearest integer) when the optimal simulated spectra and real tandem spectra were searched against the standard target database.

		Log_{10}-transformed *p-values*							
Program	Spectra	0[a]	1	2	3	4	5	≥6	Peptides (%) at $<1 \times 10^{-4}$
Crux	Optimal	2	5	12	52	3	1	5	11.3
	Real	9	8	9	44	1	0	9	12.5
OMSSA	Optimal	0	0	0	0	0	0	80	100.0
	Real	0	0	1	2	1	3	73	96.3
X! Tandem	Optimal	0	0	4	4	2	6	64	90.0
	Real	1	8	11	15	16	11	18	56.3

[a] Significance threshold (t) for matches to be significant at *p-value* $<1 \times 10^{-t}$.

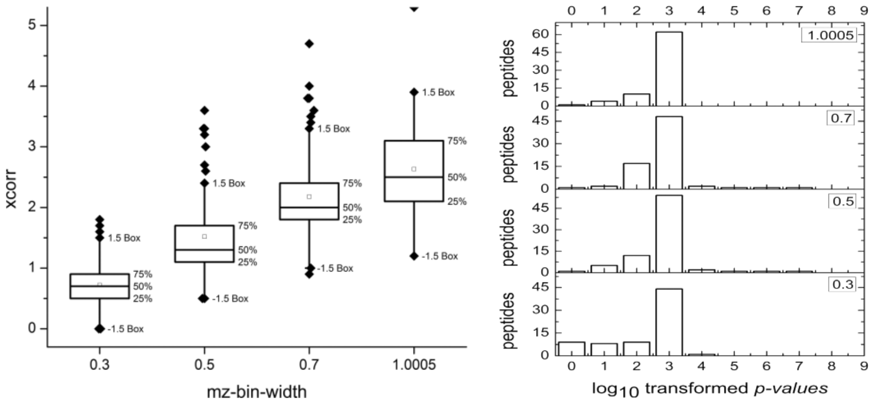
identified 73 (91.25%) peptides with seven peptides between 7 to 14 amino acids in length undetected.

Crux, OMSSA, and X! Tandem correctly matched 10 (12.5%), 77 (96.35%), and 45 (56.3%) real spectra, respectively, at *E*- or *p-value* $<1 \times 10^{-4}$. A large number of peptides (44) were detected with a *p-value* $<10^{-3}$ indicating the previously noted difficulty of obtaining significant matched with Crux.[7] The spectra quality features such as missing peaks, noise peaks, and low intensity peaks tended to reduce the positive correlation that was observed between peptide length and *E-value* in the optimal simulated scenario.

Higher number of Weibull points (XCorr scores) were correlated with more significant *p-values* in Crux.[7] Consistent with prior work, the increase in the number of Weibull points from 10^3 to 10^4, and 10^5 resulted in 24 and 10 more peptides that reached *p-value* $<1 \times 10^{-4}$ relative to the 10^3 scenario, respectively. However, 17 and 40 more peptides had *p-value* $>1 \times 10^{-2}$ with 10^4 and 10^5 Weibull points, respectively, than with 10^3 points (data not shown). Further investigation uncovered that peptides that did not reach the significance threshold were affected by the "mz-bin-width" (fragment ion tolerance) parameter. Increasing the "mz-bin-width" values from 0.3 to 1.0005 increased XCorr scores, and consequently, reduced the number of peptides that had *p-value* $>1 \times 10^{-2}$ (Fig. 1). Thus, the 0.3 specification appears to provide more conservative results. However, to use comparable search specification for the three database search programs, from this point onwards, all Crux results were calculated using the more conservative 0.3 "mz-bin-width".

2.2. *Peptide detection using real spectra against the End decoy database*

The detection of peptides from observed real spectra when matched against the End-permuted decoy database improved relative to the standard comparison against a target database. Figure 2 depicts the distribution of the effective database size

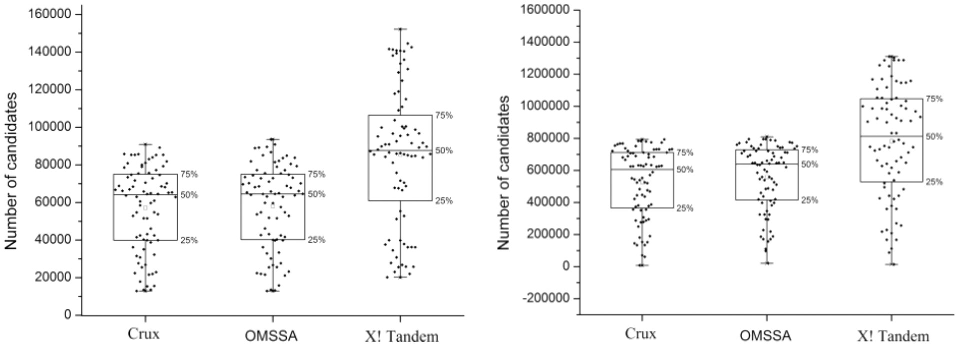


(a) (b)

Fig. 1. Box plots of Crux XCorr scores (a) and number of peptides correctly identified at different $-1 * \log_{10}$-transformed Weibull *p-values* (b) using "mz-bin-width" values of 0.3, 0.5, 0.7, and 1.0005.

corresponding to each observed spectra for the three database search programs when two (Ends2) or three (Ends3) terminal residues were permuted. The patterns in these box plots showed that X! Tandem evaluated more decoy sequences than the Crux and OMSSA.

For each peptide, some matches of the observed spectrum against the decoy database spectra were indistinguishable from each other in terms of all indicators (e.g. the number of matched fragment ions, XCorr score, and Sp score). This is because for each peptide, the Ends2 and Ends3 decoy databases had di-mer and tri-mer residue combinations with similar total monoisotopic masses. These numerically



(a) (b)

Fig. 2. Box plots depicting the distribution of number of candidate decoy peptides within precursor mass tolerance per queried observed peptide considered by Crux, OMSSA, and X! Tandem for the (a) Ends2 and (b) Ends3 permuted decoy databases.

indistinguishable matches were counted as one when calculating the permutation *p-values* to avoid biases toward any one database search program.

Table 3 summarizes the number of peptides matched at different $\log_{10}$-transformed permuted *p-value* significant levels across match indicators and database search programs for the Ends1, Ends2, and Ends3 decoy databases. The searches against Ends1 decoy database resulted in lower significance levels for all peptide matches from the three database search programs across various match indicators. The lower number of permuted sequences available in the Ends1 decoy database resulted in permutation *p-values* that were not significant at the Bonferroni adjusted threshold of $< 1 \times 10^{-4}$.

2.2.1. *X! Tandem*

The level of significance of the matches to the decoy databases increased from Ends1 to Ends2 and stabilized between Ends2 and Ends3 decoy databases (Table 3). The Ends2 and Ends3 decoy databases enabled the detection of 34.95% to 38.70% more peptides than the target database. Overall, the X! Tandem indicator convolution score had the lowest detection rate among all indicators suggesting that the convolution score alone is inadequate to discriminate between true target and false decoy matches. Detections and significance levels were similar for the hyperscore and *E-value* indicators. Furthermore, detection rate was comparable between hyperscore and the number of matched ions across the three End decoy databases. End decoy databases improved peptides detection relative to the target database for number of matched ions, hyperscore and *E-value* indicators.

The peptides that were not detected by the hyperscore were also not detected by the number of matched ion indicator. The decoy database size was not correlated with the significance level or capability to detect the peptide. Of the undetected peptides, two peptides were not detected with the Ends2 and Ends3 databases. Meanwhile five undetected peptides in the Ends2 database were significant with the Ends3 database, four other peptides that were significant in the Ends2 database were not detected (became nonsignificant) in the Ends3 decoy database. The nonsignificant peptides in the Ends3 database were either nonsignificant or marginally significant in the target database.

Table 4 summarizes the number of peptides detected in the target and Ends3 decoy databases, target only, Ends3 only, and missed by both databases when the number of matched ions and hyperscore indicators are considered. The Ends3 decoy database enabled the detection most peptides (42 out of 45) that were significant in the target database in addition to the 32 peptides that were missed by the standard target database. The performance of the number of matched ions and hyperscore was comparable. The higher significance of the matches resulting from the consideration of the hyperscore relative to all other X! Tandem indicators can be attributed to the use of peak intensity in the scoring and the theoretical spectrum synthesis process.[15]

Table 3. Number of peptides detected by different spectra match indicators within database search programs across $\log_{10}$-transformed *p-values* levels (rounded down to the nearest integer) using End decoy databases.

			Log_{10}-transformed *p-values*							
Programs	Database[a]	Indicators	0[b]	1	2	3	4	5	≥ 6	Pep. $< 1 \times 10^{-4}$[c]
X! Tandem	Ends1	# of ions	0	8	72	0	0	0	0	0
		Convolution	0	25	55	0	0	0	0	0
		Hyperscore/*E-value*	0	9	71	0	0	0	0	0
	Ends2	# of ions	0	0	0	7	65	8	0	73
		Convolution	0	2	20	41	17	0	0	17
		Hyperscore/*E-value*	0	0	0	4	67	9	0	76
	Ends3	# of ions	0	0	0	6	29	44	1	74
		Convolution	0	0	1	26	31	22	0	53
		Hyperscore/*E-value*	0	0	0	5	20	51	4	75
Crux	Ends1	# of ions	0	20	60	0	0	0	0	0
		Sp	0	19	61	0	0	0	0	0
		XCorr/ΔCn	4	30	46	0	0	0	0	0
	Ends2	# of ions	0	0	0	15	65	0	0	65
		Sp	0	0	0	13	67	0	0	67
		XCorr/ΔCn	1	6	12	28	33	0	0	33
	Ends3	# of ions	0	1	1	24	27	27	0	54
		Sp	0	1	1	20	28	30	0	58
		XCorr/ΔCn	0	3	17	25	23	12	0	35
OMSSA	Ends1	# of ions	0	16	64	0	0	0	0	0
		Lambda	2	29	49	0	0	0	0	0
		p-value/*E-value*	0	14	66	0	0	0	0	0
	Ends2	# of ions	0	0	0	22	58	0	0	58
		Lambda	0	6	15	25	34	0	0	34
		p-value/*E-value*	0	0	0	11	69	0	0	69
	Ends3	# of ions	0	0	0	10	51	19	0	70
		Lambda	0	0	0	17	43	20	0	63
		p-value/*E-value*	0	0	0	7	33	40	0	73

[a] Ends1: the last one N- and C-terminal amino acids were permuted (decoy peptides: $236 \times 360 = 84,960$); Ends2: the last two N- and C-terminal amino acids were permuted (decoy peptides: $236 \times 130,320 = 30,755,520$); Ends3: the last three N- and C-terminal amino acids were permuted (decoy peptides: 47,045,880).

[b] Significance threshold (t) for matched to be considered significant at *p-value* $< 1 \times 10^{-t}$.

[c] The number of peptides detected at *p-value* $< 1 \times 10^{-4}$.

2.2.2. *Crux*

Peptide detection and significance levels were similar for the XCorr and ΔCn across Ends2 and Ends3 decoy databases. The XCorr and ΔCn detected 33 (41.25%) and 35 (43.75%) peptides in the Ends2 and Ends3 decoy databases, respectively (Table 3). The lower peptide detection rate of XCorr and ΔCn with decoy databases indicates that XCorr and ΔCn are less suitable than the other indicators (Sp and number of ions). Overall, the Sp indicator identified two and four more peptides (*p-value* $< 1 \times 10^{-4}$) than the number of matched ions indicator in Ends2 and Ends3, respectively (Table 3).

Table 4. Number of peptides detected by spectra match indicators from database search programs using the target and Ends3 decoy databases.

		Number of peptides detected in Ends3 permuted and target databases			
Program	Indicators	PT[a]	P	T	None
Crux	# of ions	7	47	3	23
	Sp	7	51	3	19
OMSSA	# of ions	67	3	10	0
	E-value	70	3	7	0
X! Tandem	# of ions	42	32	3	3
	Hyperscore	43	32	2	3

[a] PT: peptides detected at *p-value* $<1 \times 10^{-4}$ in both target and Ends3 databases; P: peptides detected at *p-value* $<1 \times 10^{-4}$ in Ends3 database only; T: peptides detected at *p-value* $<1 \times 10^{-4}$ in the target database only; None: missed peptides (*p-value* $>1 \times 10^{-4}$) in both databases.

Combining the number of matched ions or Sp indicators with the End decoy databases improved the peptide detection relative to the target database alone. The Ends2 and Ends3 databases had 67.5% to 83.75% peptide detection rate compared to 12.50% with the target database with both indicators. The number of matched ion indicator missed more peptides (23) than the Sp indicator (19). The Ends3 permuted database detected 51 peptides missed by the standard target database using Sp indicator (Table 4).

2.2.3. *OMSSA*

Table 3 summarizes the $\log_{10}$-transformed *p-values* for the OMSSA match indicators: number of matched ions, lambda, Poisson *p-value*, and *E-value*. The Poisson *p-value* and *E-value* indicators provided similar peptide detection rate and significance levels. Therefore, results from the *E-value* indicator will be further discussed. The lambda indicator overall detected lower number of peptides than the number of matched ion and *E-value* indicators suggesting that the lambda alone is inadequate to discriminate between target and decoy matches. The Ends2 and Ends3 decoy databases provided further discrimination between the number of matched ions and *E-value* indicators, with significance levels and peptide detection rate in the decoy database higher than the target database when the *E-value* indicators was considered. The *E-value* indicator provided more true detections across significance thresholds than the number of matched ions and lambda indicators.

2.2.4. *Comparison among database search indicators*

Table 4 lists the number of peptides identified by the target and Ends3 decoy, target only, Ends3 decoy only, and not identified by either database when the number of matched ions and *E-value* indicators are considered. Meanwhile the number of ions and *E-value* indicators detected three peptides using the Ends3 decoy database that were missed by the target database, these indicators detected 10 and 7 peptides, respectively using the target database that were missed by the decoy database.

Approximately, 88% peptide detections were shared by the target and Ends3 databases using the *E-value* indicator.

2.3. *Comparison of spectra match indicators and database search software*

Figure 3 depicts the number of peptides detected one, two or all three database search programs when the number of matched ion and best score indicator from each of the three programs was used to compute the permutation *p-value*. The best score indicator was defined as the indicator that exhibited the highest difference between the target and decoy peptides. The best spectra match indicators were *E-value* for OMSSA, hyperscore for X! Tandem, and Sp for Crux.

The Ends decoy databases supported higher consensus among the three programs when compared to the target database. For the Ends3 decoy database, all three programs detected slightly less peptides together when considering the number of matched ions compared to the best indicator (50 versus 56). A similar number of peptides were detected by any two programs using the number of matched ions than the best score indicator (72 versus 73). OMSSA and Crux detected more peptides with the best indicator than the number of matched ion indicator and X! Tandem detected similar number of peptides with the number of matched ions and the hyperscore. Using either the number of matched ions or best score indicator, X! Tandem detected more peptides than OMSSA and Crux and OMSSA detected more peptides than Crux.

The computational time of the searches was calculated on a computer with 3.40 GHz Intel Core i7-3770 processor. Searching the target database only using Crux (using 1,000 Weibull points), X! Tandem and OMSSA averaged 1.14, 0.013, and 0.14 s per spectrum, respectively. Crux averaged 0.04, 3.54, and 40.65 s for

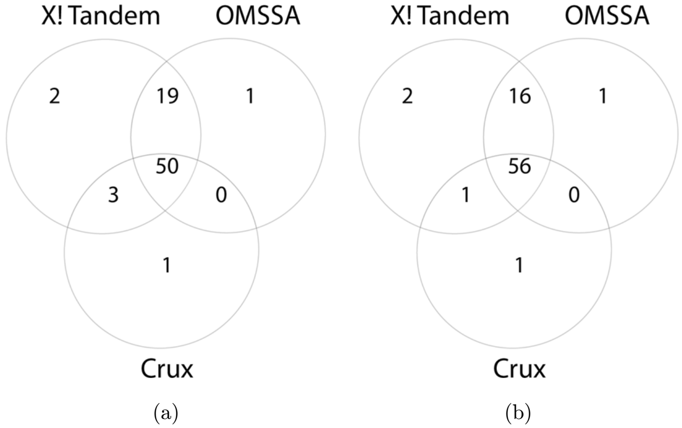


Fig. 3. Distinct and shared number of peptide detected in the Ends3 decoy database using (a) the number of matched ions or (b) the best indicator for each database search program (OMSSA *E-value*, Crux Sp score, and X! Tandem hyperscore).

Ends1, Ends2, and Ends3 decoy databases, respectively. X! Tandem averaged 0.15, 6.54, and 116.26 s per spectrum for Ends1, Ends2, and Ends3 decoy databases, respectively. OMSSA averaged 0.34, 21.72, and 604.00 s per spectrum for Ends1, Ends2, and Ends3 decoy databases, respectively. The longer search time for the X! Tandem and OMSSA using the Ends3 decoy database relative to Ends2 database could be due to the searching of separate decoy databases for each spectrum in addition to the larger database size of the Ends3 decoy database. Furthermore, the comparisons of the peptide detection rate between the Ends2 and Ends3 database suggest that detection performance similar to the Ends3 database could be obtained using a smaller random sample of the decoys in the Ends3 database. Overall, the dramatic improvement in the peptide identification highlights the efficacy of the terminal residue permutation decoy database.

3. Conclusions

The present study demonstrated that the spectra match indicators Sp (Crux), hyperscore (X! Tandem) and *E-value* (OMSSA) with a terminal residue permutation decoy database enabled effective detection of peptides compared to target database. The Ends decoy databases improved the consensus among database search programs to identify peptides. The End decoy databases can be integrated to other database search programs. The new candidate decoy peptides resulting from the permutation can also be used to discover novel peptides.

In the present study, Ends decoy databases were generated from subset of target database peptides that were within 12 Da of the observed spectra precursor masses since database search programs initially filter candidate peptides based on precursor mass. The approach can be extended to any number of peptides, types of peptides and other database search programs. This could be accomplished by generating the required number of permuted peptides from peptide-spectrum matches obtained by searching observed spectra against the target database using the desired database search program.

Acknowledgments

The support of NIH (Grant Numbers: R21 DA027548, P30 DA018310 and R21 MH096030) is greatly appreciated. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

1. Hummon AB, Amare A, Sweedler JV, Discovering new invertebrate neuropeptides using mass spectrometry, *Mass Spectrom Rev* **25**(1):77–98, 2006.
2. Zamdborg L, LeDuc RD, Glowacz KJ *et al.*, ProSight PTM 2.0: Improved protein identification and characterization for top down mass spectrometry, *Nucleic Acids Res* **35** (Web Server issue):W701–W706, 2007.

3. Xie F, London SE, Southey BR, Annangudi SP, Amare A, Rodriguez-Zas SL, Clayton DF, Sweedler JV, The zebra finch neuropeptidome: Prediction, detection and expression, *BMC Biol* **8**:28-7007-8-28, 2010.
4. Zhang X, Petruzziello F, Zani F, Fouillen L, Andren PE, Solinas G, Rainer G, High identification rates of endogenous neuropeptides from mouse brain, *J Proteome Res* **11**(5):2819–2827, 2012.
5. Jia C, Lietz CB, Ye H, Hui L, Yu Q, Yoo S, Li L, A multi-scale strategy for discovery of novel endogenous neuropeptides in the crustacean nervous system, *J Proteomics* **91**:1–12, 2013.
6. Southey BR, Lee JE, Zamdborg L *et al.*, Comparing label-free quantitative peptidomics approaches to characterize diurnal variation of peptides in the rat suprachiasmatic nucleus, *Anal Chem* **86**(1):443–452, 2014.
7. Akhtar MN, Southey BR, Andren PE, Sweedler JV, Rodriguez-Zas SL, Evaluation of database search programs for accurate detection of neuropeptides in tandem mass spectrometry experiments, *J Proteome Res* **11**(12):6044–6055, 2012.
8. Akhtar MN, Southey BR, Andren PE, Sweedler JV, Rodriguez-Zas SL, Evaluation of significance level assignment of database search programs using monte carlo permutation approach, *6th Int Conf Bioinformatics and Computational Biology*, Las Vegas, Nevada, USA, March 24–26, 2014.
9. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS, Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis* **20**(18):3551–3567, 1999.
10. Nesvizhskii AI, A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics, *J Proteomics* **73**(11):2092–2123, 2010.
11. Craig R, Beavis RC, TANDEM: Matching proteins with tandem mass spectra, *Bioinformatics* **20**(9):1466–1467, 2004.
12. Geer LY, Markey SP, Kowalak JA *et al.*, Open mass spectrometry search algorithm, *J Proteome Res* **3**(5):958–964, 2004.
13. Klammer AA, Park CY, Noble WS, Statistical calibration of the SEQUEST XCorr function, *J Proteome Res* **8**(4):2106–2113, 2009.
14. Higdon R, Hogan JM, Van Belle G, Kolker E, Randomized sequence databases for tandem mass spectrometry peptide and protein identification, *OMICS* **9**(4):364–379, 2005.
15. Kapp EA, Schutz F, Connolly LM *et al.*, An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis, *Proteomics* **5**(13):3475–3490, 2005.
16. Frese CK, Boender AJ, Mohammed S, Heck AJ, Adan RA, Altelaar AF, Profiling of diet-induced neuropeptide changes in rat brain by quantitative mass spectrometry, *Anal Chem* **85**(9):4594–4604, 2013.
17. Sadygov RG, Yates JR, A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases, *Anal Chem* **75**(15):3792–3798, 2003.
18. Elias JE, Gygi SP, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, *Nat Methods* **4**(3):207–214, 2007.
19. Lai Y, Conservative adjustment of permutation *p-values* when the number of permutations is limited, *Int J Bioinform Res Appl* **3**(4):536–546, 2007.
20. Knijnenburg TA, Wessels LF, Reinders MJ, Shmulevich I, Fewer permutations, more accurate *p-values*, *Bioinformatics* **25**(12):i161–i168, 2009.
21. Park CY, Klammer AA, Kall L, MacCoss MJ, Noble WS, Rapid and accurate peptide identification from tandem mass spectra, *J Proteome Res* **7**(7):3022–3027, 2008.

22. Falth M, Skold K, Norrman M, Svensson M, Fenyo D, Andren PE, SwePep, a database designed for endogenous peptides and mass spectrometry, *Mol Cell Proteomics* **5**(6):998–1005, 2006.
23. Southey BR, Akhtar MN, Andrén PE, Sweedler JV, Rodriguez-Zas SL, A comprehensive resource in support of sequence-based studies of neuropeptides, **6**:144, 2013.
24. UniProt Consortium, The universal protein resource (UniProt) in 2010, *Nucleic Acids Res* **38**(Database issue):D142–D148, 2010.
25. Southey BR, Amare A, Zimmerman TA, Rodriguez-Zas SL, Sweedler JV, NeuroPred: A tool to predict cleavage sites in neuropeptide precursors and provide the masses of the resulting peptides, *Nucleic Acids Res* **34**(Web Server issue):W267–W272, 2006.
26. Ernst MD, Permutation methods: A basis for exact inference, *Stat Sci* **19**:676–685, 2004.

Malik N. Akhtar is a PhD student of Bioinformatics in the Department of Animal Sciences at the University of Illinois, Urbana-Champaign. He received his BSc degree in Bioinformatics from COMSATS Institute of information technology, Pakistan and MSc in Bioinformatics from the University of Illinois, Urbana-Champaign.

Bruce R. Southey is a research assistant professor of Bioinformatics in the Department of Animal Sciences at the University of Illinois, Urbana-Champaign. He received his MSc degree from Massey University, New Zealand and PhD from the University of Wisconsin-Madison. Southey is the lead statistician at the Bioinformatics Core of the Proteomics for Cell–Cell Signaling at the University of Illinois.

Per E. Andrén is a senior lecturer and researcher in the Department of Pharmaceutical Sciences in the University of Uppsala, Sweden.

Jonathan V. Sweedler is a Professor of Chemistry in the Department of Chemistry at the University of Illinois, Urbana-Champaign. He received his BS degree in Chemistry from the University of California at Davis and PhD from the University of Arizona. He leads the Proteomics Center for Cell–Cell Signaling at the University of Illinois. He is also associated with the Beckman Institute, Biotechnology Center, Neuroscience Program, and Bioengineering Program in the University of Illinois, Urbana-Champaign.

Sandra L. Rodriguez-Zas is a Professor of Bioinformatics in the Department of Animal Sciences and Statistics at the University of Illinois, Urbana-Champaign. She received her MSc and PhD in Quantitative Genetics from the University of Wisconsin-Madison. She is the director of the Bioinformatics Core of the Proteomics Center for Cell–Cell Signaling at the University of Illinois.