Evaluating Reproducibility and Similarity of Mass and Intensity Data in Complex Spectra—Applications to Tubulin

Matthew T. Olson,[a] Paul S. Blank,[a] Dan L. Sackett,[b] and Alfred L. Yergey[a]

[a] Laboratory of Cellular and Molecular Biophysics, National Institute of Child Health and Human Development, NIH, Bethesda, Maryland, USA
[b] Laboratory of Integrative and Medical Biophysics, National Institute of Child Health and Human Development, NIH, Bethesda, Maryland, USA

We present a data processing approach based on the spectral dot product for evaluating spectral similarity and reproducibility. The method introduces 95% confidence intervals on the spectral dot product to evaluate the strength of spectral correlation; it is the only calculation described to date that accounts for both the non-normal sampling distribution of the dot product and the number of peaks the spectra have in common. These measures of spectral similarity allow for the recursive generation of a consensus spectrum, which incorporates the most consistent features from statistically similar replicate spectra. Taking the spectral dot product and 95% confidence intervals between consensus spectra from different samples yields the similarity between these samples. Applying the data analysis scheme to replicates of brain tubulin CNBr peptides enables a robust comparison of tubulin isotype expression and post-translational modification patterns in rat and cow brains. (J Am Soc Mass Spectrom 2008, 19, 367–374) © 2008 American Society for Mass Spectrometry

It is widely recognized that multiple acquisitions of MALDI spectra have extensive variability. Addressing this problem in an automated and robust way becomes vital to the analysis of complex spectra because such spectra often contain peaks that are both inconsistent in their presence or absence and show great variability in intensity. Evaluating differences between replicate spectra manually can be tedious, if not impossible, and automatic tests for discriminating similar and dissimilar spectra have several caveats that have only been partially elaborated to date. Our calculations use the spectral dot product to evaluate spectral similarity. While several papers [1–4] have used the spectral dot product for evaluating spectral similarity, particularly the similarity of a query spectrum against library references [1, 3, 4], only one paper [4] has addressed the critical issue of how an algorithm or investigator attributes significance to a spectral dot product—a nontrivial detail because the dot product alone can yield equivocal information. Dot product ambiguity arises both from the number of peaks, as has been described [4], and from the different sampling distribution of the dot product at different values. We introduce the 95% confidence intervals on the spectral dot product to address both these sources of ambiguity. To accommodate for spectral variance, multiple acquisitions of the same sample are used to generate a consensus spectrum which incorporates the consistent features from all the replicate spectra. The dot product and its 95% confidence intervals allow for validation of the consensus spectrum by evaluating the similarity of each replicate to the consensus spectrum. The same method is also used to compare and evaluate similarity or dissimilarity between consensus spectra from different sources.

Microtubules comprise one of the major fiber systems of the cytoskeleton and are polymers of the heterodimeric protein tubulin. Microtubules and tubulin are present in all eukaryotic cells and are responsible for cellular structure, chromosomal motion during mitosis, and intracellular trafficking, including axonal transport. Tubulin is the most abundant protein in mammalian brains. Multiple genes produce isotypes for each subunit of the tubulin dimer [5, 6] (Table 1 presents the isotypes used in these experiments), and further diversity is introduced when the expression products undergo extensive post-translational modification (PTM). While the genes are highly conserved across species, differences in isotype expression and PTM reflect key biological processes and have also been implicated in disease states.

As shown in Figure 1, extensive genetic variability in tubulin isotypes occurs at the C-terminus, making the

Address reprint requests to Dr. A. L. Yergey, Laboratory of Cellular and Molecular Biophysics, National Institute of Child Health and Human Development, NIH, Building 10, Room 9D52, Bethesda, MD 20892, USA. E-mail: aly@helix.nih.gov

1044-0305/08/$32.00
doi:10.1016/j.jasms.2007.11.012

Published online November 23, 2007
Received June 19, 2007
Revised November 6, 2007
Accepted November 7, 2007

Table 1. GenBank accession numbers for the tubulin subunits and isotypes used in these experiments

Tubulin subunit	Isotype	Accession
Alpha	K-a1	K00558
	B-a1	X01703
	a3	K03460
	a4	X06956
Beta	b2	X79535
	b4	P05217
	b5	A26561
	b6	U47634

C-terminus a practical way to distinguish isotypes. While isotype-specific antibodies are available for some of these types, mass spectrometry (MS) is particularly well suited to such differentiation. Further, most tubulin PTM occurs at the C-terminus. These modifications include polyglutamylation and polyglycylation, the addition of glutamate or glycine residue(s) by formation of a peptide bond on the γ-carboxy group of a glutamate in the sequence, and detyrosination, removal of C-terminal tyrosine and possibly the penultimate glutamate. As a consequence of both the numbers of isotypes and their PTMs, tubulin samples may yield complex spectra since the same tubulin C-terminus may contain up to 12 added glutamylate residues and/or up to 12 added glycine residues. When these possibilities are combined with the additional possibility of detyrosination and loss of penultimate glutamate in the three isotypes that have a C-terminal tyrosine, there are 3006 possible C-termini for mammalian tubulin. While spectra from living tissues do not exhibit all of these possibilities, they are nonetheless complex. Accurately defining this complexity requires careful sample preparation and development of data analysis methods to evaluate spectrum variability and compare differing spectra. The methods presented here demonstrate a general methodology for evaluation of spectral reproducibility and similarity with the spectral dot product and introduce a helpful peak assignment strategy with demonstrated ability to assign the peaks in complex spectra.

The spectral dot product has an established use as the means by which mass spectra are similar to spectra of known substances in a reference library [1, 3, 4, 7]. Monigiatti and Berndt took a different approach and used consensus spectra for the purpose of database searching with peptide mass fingerprints by using a Spearman's rank correlation coefficient [8]. The calculations of this work focus on the problem of spectral similarity and reproducibility, not on database searching. For reproducibility calculations alone, as considered in the present work, it is necessary to make this distinction since one assumes that all the spectra should be from the same molecules whereas in database searching one cannot assume this.

Methods

Isolation of Microtubules and Cyanogen Bromide Digestion

Tubulins were extracted from rat and bovine brains and bovine testicles by an established method [9–11]. The extracted tubulins were cleaved by an optimized protocol involving 0.1M CNBr, which is roughly a 2-fold to 10,000-fold excess of cyanogen bromide (CNBr) relative to methionine, in 20% trifluoroacetic acid (TFA) solution. Cleavage proceeded overnight at room temperature in a light tight vessel. After CNBr cleavage, the sample was quenched with an equivalent volume of deionized water, concentrated on a reverse phase spin column (Vivascience, Edgewood, NY), and eluted in matrix. The matrix consisted of 5 mg/mL a-cyano hydroxycinnamic acid in 50% acetonitrile. Sodium adduction to the highly glutamylated C-termini was minimized with the addition of sufficient 100 mM ammonium sulfate to achieve a 10 mM concentration in the matrix [12].

Mass Spectrometry

Each sample was spotted in triplicate on the MALDI target, and 30,000 shots—10 spectra of 1000 shots per spot—were acquired in batch mode for each sample on

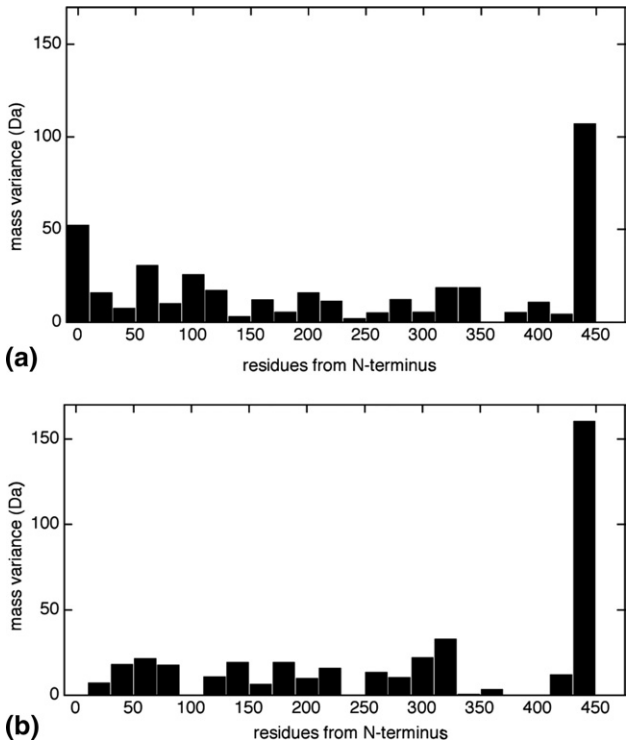

Figure 1. The standard deviation of mass among alpha (**a**) and (**b**) isotypes binned in groups of 20 amino acids. Minor differences among isotypes contribute only minor variance, while the C-terminus contains much more dramatic heterogeneity. These graphs do not include any post-translational modification, which further contributes to C-terminal variance.

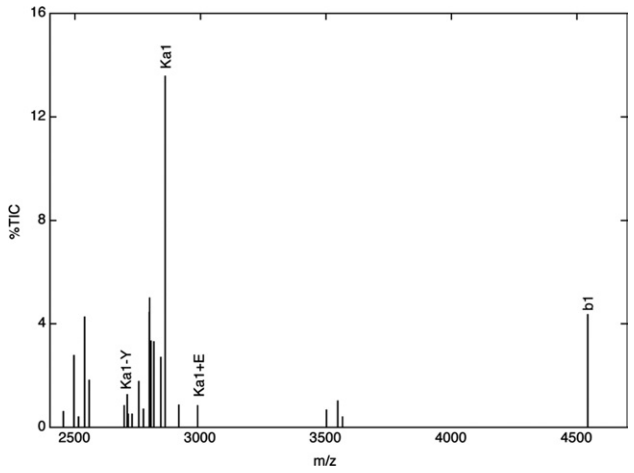

Figure 2. The labeled consensus spectrum of CNBr cleavage products obtained from chicken erythrocyte tubulin. This spectrum makes an exceptional calibrant in negative ion mode for 2 to 5 kDa because it contains the peak from b1 and because it is relatively simple among tubulin spectra. The species labeled in this figure are used in the external calibration of every tubulin spectrum presented here.

an Applied Biosystems (ABI) 4700 (PerSeptive Biosystems, Framingham, MA) operated in negative reflector mode. The spectra were calibrated externally with the use of CNBr cleaved microtubules from chicken erythrocytes spotted in a proximal location to the sample locations and acquired immediately before the acquisition of the sample spectra. The chicken erythrocyte tubulin spectrum (Figure 2) was chosen for its robust signal in negative ion mode, its simplicity and the mass range of its signal.

Generation of the Consensus Spectrum

Peak lists from the replicate spectra, unfiltered and unsmoothed, were exported as flat text files via the ABI 4700 Series (ver. 3.0) software. In addition to *m/z* values, the software reports peak intensity as signal versus noise (S/N); for these experiments, the peak lists contained all peaks above an S/N value of 5. A Perl script binned the peaks between 2400 and 4300 Da into 1 Da bins, which were centered on the nominal mass (rather than ± 0.5 Da) to account for the mass defect of peptides in this range. The algorithm calculated average intensity and standard deviation (σ) for each bin and generated a consensus spectrum. Peaks were allowed into the consensus spectrum if the relative standard deviation (RSD = σ/average intensity) of their intensities was less than an arbitrary cutoff of 1.7; this corresponds to the requirement that the peak be present with a reproducible intensity in roughly one third or more of the replicate spectra. While nominal mass bins were used to bin replicate spectra, the calibrated mass values within those bins were retained. Calibrated masses within the same nominal mass bin in all the replicate spectra were averaged to yield the calibrated masses in the consensus spectrum.

Evaluation of Spectral Reproducibility and Refinement of the Consensus Spectrum

Consensus and replicate spectra were converted into a zero-filled format and then normalized into a unit vector by dividing each bin in the spectrum by the magnitude of the spectrum (square root of the sum of squared intensities). The spectra were mean centered, which rendered a spectrum of intensities that varied around the mean; intensity values greater than the mean were positive and those less than the mean were negative. After mean centering, the spectra were re-normalized. As a measure of spectral reproducibility, spectral dot products, r, between the mean centered consensus spectra and the mean centered replicates used to generate them were calculated by eq 1, where $I_{\text{consensus j}}$ and $I_{\text{replicate j}}$ are the normalized intensities in the jth mass bin for each spectrum.

$$r = \sum_j I_{consensus,\, j} I_{replicate,\, j} \tag{1}$$

Dot products of mean centered spectra can have values from −1 to 1; dot products close to 1 indicate spectral similarity, dot products close to zero indicate little similarity, and dot products close to −1 indicate inverse relationship. Because the dot product, which can be equated algebraically to the Pearson's correlation coefficient, does not have a normal sampling distribution, calculation of confidence intervals to gauge the strength of correlation requires a Fisher's transformation to convert the dot product, r, to a normally distributed variable, z′ [13].

The standard error, σ, on the normal distribution of z′ relies on N, the number of nominal mass bins the two spectra have in common.

$$\sigma = \frac{1}{\sqrt{N-3}} \tag{2}$$

The confidence on z′ was calculated by adding and subtracting the product of the standard error and the z value for 95% confidence interval (CI), i.e., z = 1.96:

$$\text{CI} = z' \pm z\sigma \tag{3}$$

Taking the inverse of Fisher's transformation for each of the confidence intervals yields the CI of the dot product.

The 95% CI on dot products between replicate and consensus spectra was used to determine which spectra exhibited reproducible characteristics and which ones were primarily noise. A filtered set of superior replicates was generated by selecting only the replicate spectra with a positive dot product and positive 95% CI with the consensus spectrum. A new consensus spectrum was then generated from the filtered set of replicates.

Peak Assignment and Comparison of Consensus Spectra

Consensus spectra from different sources were evaluated for similarity using the dot product and 95% CI as described in the previous section. Additionally, a separate Perl script found the best match between the calibrated masses in the consensus spectrum and a list of monoisotopic masses for all possible tubulin species. The list compiles predictions of CNBr cleavage products, with up to 1 missed cleavage, for all the tubulin genes. The list also contains predicted C-terminal modifications: up to 12 glutamylations, 12 glycylations, and detyrosination of α subunits with and without the removal of the penultimate glutamate.

Results and Discussion

Reliable acquisition of peaks in the complex spectra generated by CNBr-generated peptides of brain tubulin requires an optimized sample preparation. Cleavage with CNBr has been described elsewhere [14–16] as a desirable method for the analysis of tubulin C-termini because the resulting fragments have the least overlap with non-C-terminal fragments. The non-C-terminal fragments, as shown in Figure 1, are generally isobaric between isotypes and so are not amenable to discrimination by MS analysis.

Efficient cleavage of tubulin with CNBr yields the tubulin C-termini with masses ranging from 1.5 to 4 kDa. Because of the intrinsic negative charge and the extensive glutamylation on the tubulin C-terminus, these peptides ionize better in negative ion mode than in positive ion mode. Since standard calibration mixtures have variable ionization in negative ion mode, a tubulin sample was required for adequate external calibration. The CNBr fragments generated from chicken erythrocyte tubulins (Figure 2) were used because of spectral simplicity, owing to limited isotype heterogeneity in this sample, and because of the broad mass range, owing to the presence of the exceptionally long C-terminus of b1 [17], a β-tubulin isotype that is exclusive to hematologic tissues.

Several calculations, the consensus spectrum and calculations for spectral correlation, were performed on the spectra obtained from the samples described above. All of these calculations rely on binning the spectrum. Before discussion of results, it is important to address two fundamental parameters of the bin—size and center—because these parameters have an intricate relationship to spectral correlation. Bin size matters because excessively small mass bins suppress spectral correlation while excessively large mass bins lead to spurious correlation. As a first step to determining the appropriate bin size, the predicted mass difference between nearest neighbors in the list of possible tubulin species was calculated. These calculations show that 1 Da or more separates about 70% of nonisobaric tubulin species from their nearest neighbors. Nevertheless, the possibility remains that two peaks differing by <1 Da could be binned together and averaged. To evaluate if this is occurring, the data processing methods used here conserve the calibrated mass information and calculate the variance of peaks assigned to the same bin, and peaks with too much mass variance (>25 ppm) were discarded from the consensus spectrum. For the data presented, the average standard deviation of masses in a nominal mass bin is 9.7 ppm (median: 8.8 ppm; range: 2.0–23.6 ppm), which is well within the expected instrumental variability for the ABI 4700 in this mass range.

Mass bins for these calculations were centered at the nominal mass rather than the half mass. This is done because extensive tubulin C-terminal glutamylation leads to the possibility of a substantial negative mass defect in the tubulin C-terminal peptides, i.e., the mean mass defect is about 0 Da with a standard deviation of 0.4 Da compared with the expected defect of 0.8 Da for an average distribution of peptides in this mass range. The standard deviation for masses in the nominal mass bin (mentioned above as 9.7 ppm) could only be achieved for tubulin with mass bins centered at nominal mass. Centering anywhere else in the nominal mass bin leads to improper averaging of peaks and at least a 2-fold increase in the error associated with masses in the consensus spectrum. Two figures illustrating these points on spectra binning are included in the supplementary materials (which can be found in the electronic version of this article).

Spectral similarity between two mean centered and normalized spectra is evaluated here by the use of the dot product. Beyond the geometric interpretation of the dot product in two- or three-dimensional vectors, for more complex data the dot product can be understood as the algebraic equivalent of the Pearson's correlation coefficient when the data are mean-centered. In a mean centered spectrum, the mean is subtracted from every non-zero intensity value so that each value can be interpreted with respect to the mean. Consequently, the product of overlapping peaks can have a negative value if a peak is large with respect to the mean in one spectrum and small with respect to the mean in the other spectrum. Thus, as shown in Figure 3, values for the spectral dot product between mean centered spectra tend to be lower than those calculated on uncentered spectra. The lower dot product values carry two advantages: First, because mean centered dot products span values from -1 to 1, as opposed to uncentered dot products that range from 0 to 1, the values for mean centered dot products have a clearer means of interpretation because 0 is the point at which noncorrelation cannot be distinguished from correlation. Second, mean centering the dot product equates it to the Pearson's correlation coefficient, which enables the use of Fisher's transformation to calculate confidence intervals. Confidence intervals complement information from the dot product because they indicate the strength of the correlation. Most pertinently, the confidence intervals identify which positive values of the dot product have signif-

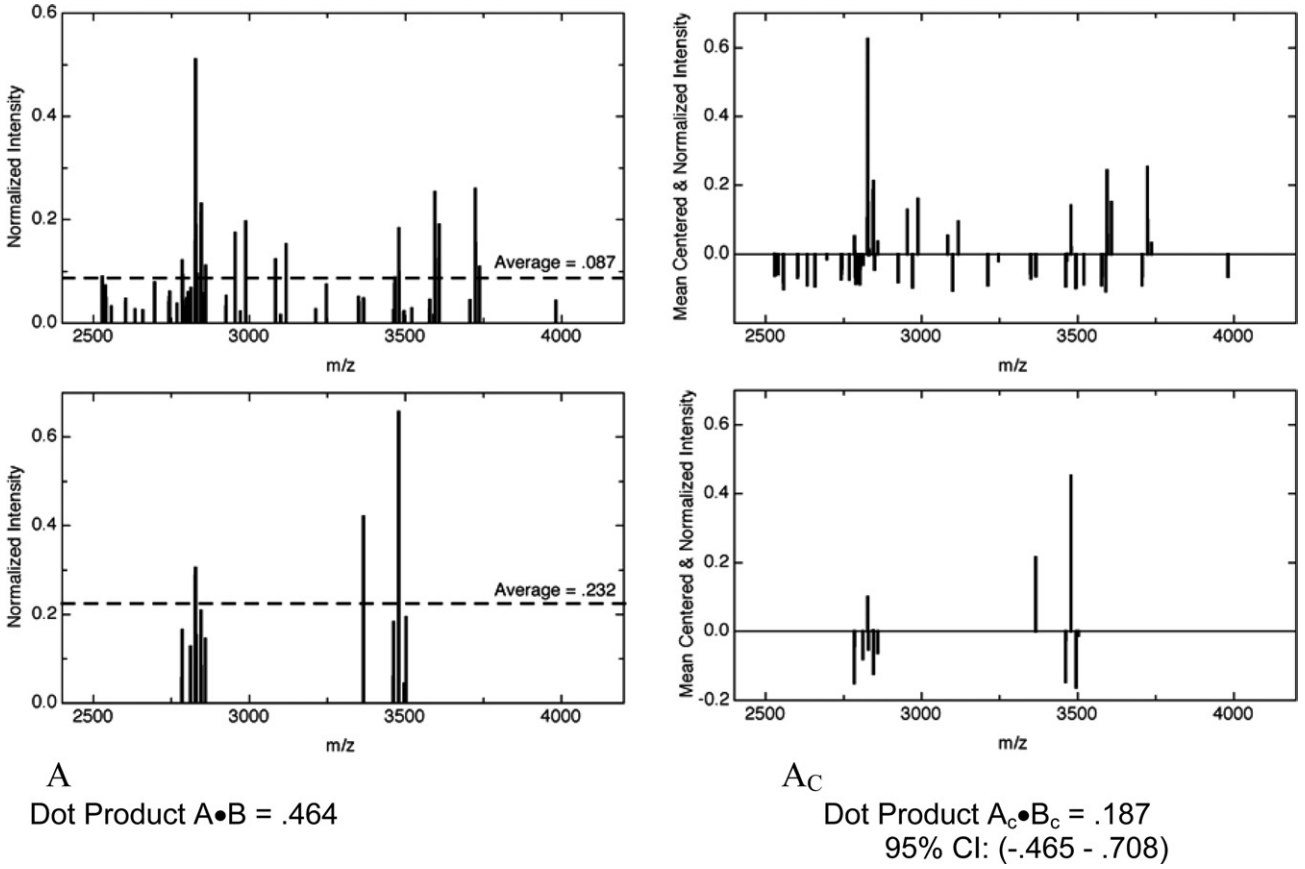

Figure 3. Normalized raw spectra from bovine brain (BBT) and bovine testes (BTT) are depicted in (**a**) and (**b**), respectively, with the average values of intensities plotted as dashed horizontal lines. When the average value of intensities for each spectrum is subtracted from each peak in spectrum A and B, Ac, and Bc result. A·B and $A_c \cdot B_c$ are calculated and shown for comparison. Because the sampling distribution for centered dot products is the same as that of Pearson's correlation coefficient, $A_c \cdot B_c$ is given along with its 95% confidence interval.

icance and which do not. When the dot product is positive but the confidence interval includes zero, correlation cannot be differentiated from noncorrelation with statistical significance, so the similarity is void statistically.

Because the dot product is the sum of products for all the overlapping peaks in normalized spectra, it cannot be interpreted as a continuous variable with rigid cutoff values. Rather, the dot product, like the correlation coefficient, is an index of linearity between the intensity for a mass in one spectrum and the intensity for the same mass in the other spectrum. As with any correlation coefficient, the dot product has variance that is inversely related to the number of overlapping bins, by eq 3. Calculating the variance on the dot product itself would be meaningless because the dot product has a non-normal sampling distribution; the farther the dot product from zero the greater its skew towards zero. To consider the variance in spectral comparisons with the dot product, we transform the dot product into a normal variable by Fisher's transformation, which is the same transformation used on the Pearson's correlation coefficient. This allows for the calculation of 95% CI on a normal variable. Taking the inverse of Fisher's transformation for each of the confidence intervals yields the CI of the dot product.

The 95% CI can be used to filter the set of replicates. Thirty replicate spectra were acquired from CNBr cleavage products of bovine testicular tubulin (BTT). Replicates from this sample varied widely as shown in Figure 4. With the replicate spectra binned in the nominal mass bins centered on the nominal mass, a consensus spectrum was formulated by incorporating only the peaks whose intensities had an RSD ≤ 1.7. The dot products and 95% CI for the replicate versus consensus comparisons is shown by the black points in Figure 5. The poor reproducibility for this sample is evident from this Figure because most of the replicate spectra have insignificant correlation with the consensus, so only 10 peaks matriculate into the consensus. When the black error bars in Figure 5 that have a negative 95% CI are removed from the set, and the filtered set—the spectra in the figure with red bars on them—is used to calculate the consensus spectrum, the number of peaks in the consensus spectrum increases to 15, and the average difference between the upper and

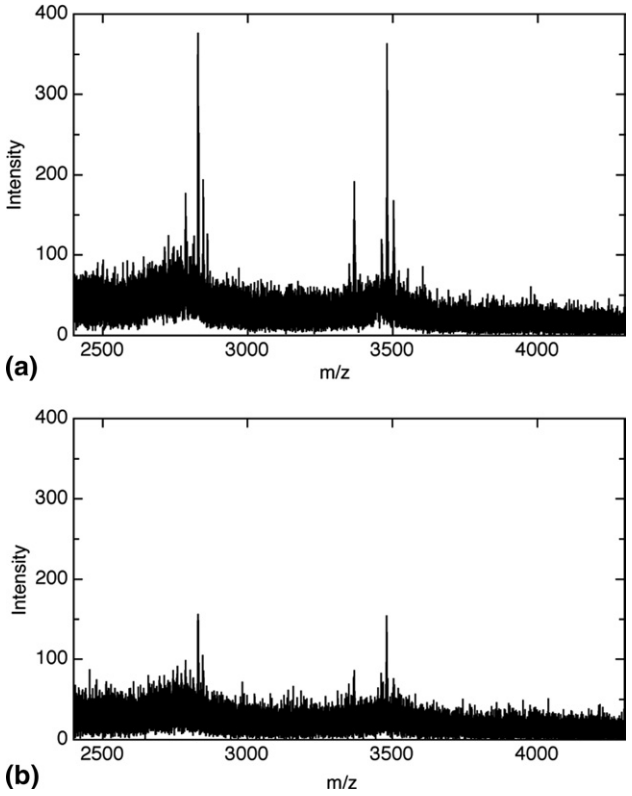

Figure 4. Raw spectra showing the range between quality (**a**) bovine testicular tubulin (BTT) and noisy (**b**) BTT replicate spectra. Intensity values are given as counts. With multiple replicate spectra and the use of the dot product, noisy replicates can be removed from the sample set.

lower confidence intervals decreases from 1.32 (±0.30) to 0.71 (±0.17).

To show that the spectra with positive correlation to the original consensus represent the spectra with the most reproducible peak patterns and to rule out the possibility that these positively correlating spectra were in some way random selections rather than the best spectra on which to base the consensus, we investigated consensus spectra generated from random samples of five spectra from among the set of noncorrelated spectra, which excluded the spectra with the five lowest dot products. These consensus spectra demonstrated variable and inconsistent correlation to each other and to the original consensus spectrum. Additionally, the randomly selected spectra had universally less correlation to their consensus than was seen with the consensus spectrum generated from only positively correlating replicates. Finally, a consensus spectrum generated from the spectra with the least correlation to the 30-replicate consensus spectrum (BTT replicates 26–30) demonstrated the least correlation to the original consensus, and the replicates had no correlation to this new consensus spectrum. These observations demonstrate the reliability of the 95% confidence interval to select for the reproducible spectra in a particularly noisy set of replicates. The example here represents the most dramatic case of improvement encountered with the use of this filtering schema. The other spectra in these experiments demonstrated far greater reproducibility, and the effect of replicate selection for consensus generation was negligible.

Generating a consensus spectrum and attenuating it by its correlation with the replicates yields a final spectrum with a high level of statistical confidence. This confidence derives from the mass precision and intensity reproducibility requirements involved in the selection of peaks that matriculate into the consensus spectrum and from the ability of the dot product to select the spectra which exhibit the most reproducible patterns. The final spectrum yielded by this exercise is a spectrum in which every feature, including the spectral pattern, has statistical confidence. As a result, the consensus spectrum comprises a more reliable and accurate tool for evaluating biological questions.

The consensus spectra for tubulins from rat (RBT) and bovine (BBT) brains are shown in Figure 6a and b, respectively. These spectra represent the most consistent features in the 30 replicate spectra of 1000 shots each. The estimated amount of tubulin protein depos-

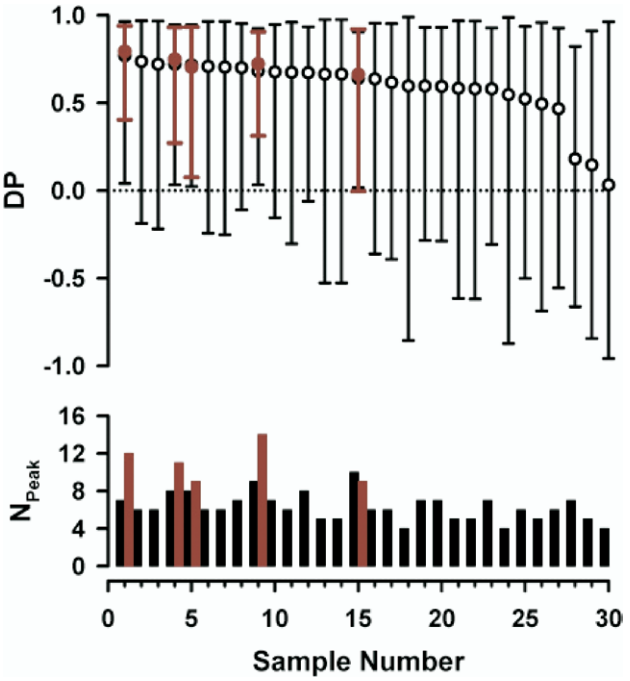

Figure 5. Spectral dot products (DP) with their asymmetric 95% confidence intervals (CI) along with the number of overlapping peaks (Npeak) for the consensus spectrum of 30 bovine testicle tubulin (BTT) replicate spectra versus each replicate spectrum. The spectral numbers are arranged in order of descending dot product. The dark bars show the dot product of the replicates versus the consensus of 30 spectra. The red bars show the dot product between the replicates that have a positive 95% CI with the original consensus spectrum and a consensus spectrum made from only those replicates. When only the spectra with positive 95% CI are used for the formulation of the consensus spectrum, the number of peaks in the consensus spectrum increases, all but one of the replicates maintain confident correlation with the consensus, and the confidence intervals are tighter.

ited onto the MALDI target for these consensus spectra was 7.5 pmol/sample. Given the existence of ~25 C-terminal peaks in each spectrum, the minimum amount of substance per peak is at most 300 fmol, which agrees with the expected sensitivity for the instrument.

While the dot product yields instant and generally meaningful information about spectral similarity, the ultimate significance of spectral comparison lies in peak assignment and comparison of biological properties. A Perl script that matches the calibrated masses of the peaks in the consensus spectrum with the best possible monoisotopic mass match from a list of possible tubulin species is able to assign 72.9% and 71.8% of BBT and RBT spectra, respectively. The average error for these assignments is 23 (±16) ppm. The calibrated masses in the consensus spectrum were used to make these assignments because, as discussed above, ~30% of the peptides have more than one possible match assignment within a 1 Da window. These cases required the accuracy of the calibrated mass. Comparison of the mass assignments made between the bovine brain (BBT) and testicular (BTT) tubulin samples reveals some interesting differences. Only three peaks are found in common between the samples, and all of them are of unmodified C-terminal peptides: K-a1 is observed as 2.1% of total ion current (TIC) in BBT and 4.7% of TIC in BTT, b-5 (1.0% of TIC in BBT and 13.7% of TIC in BTT), and b4a (4.5% of TIC in BBT and 21.3% of TIC in BTT).

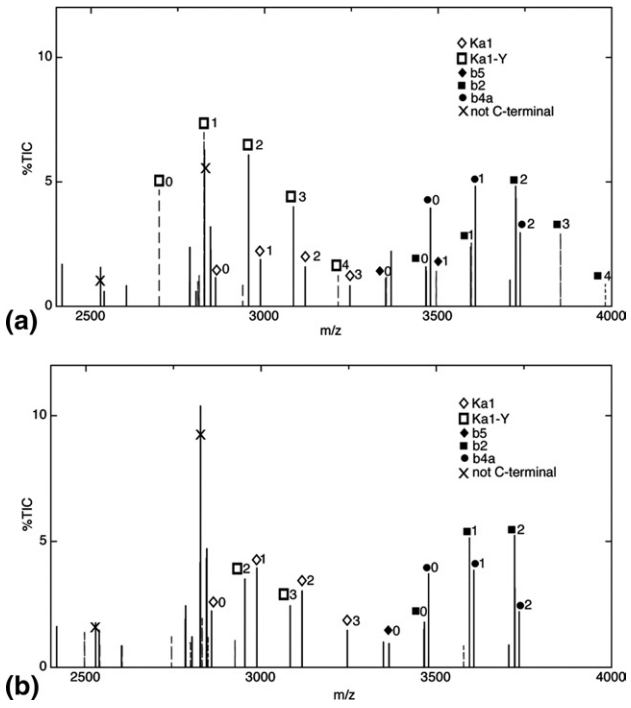

Figure 6. Consensus spectra for CNBr cleavage products from rat (**a**) and bovine (**b**) brain tubulin. Solid lines indicate peaks the two spectra share, and dash lines indicate peaks specific to only one spectrum. Peaks with "X" on them are assigned as non-C-terminal fragments.

Table 2. Inter-consensus dot products with 95% CI in parenthesis for bovine brain (BBT), rat brain (RBT), and bovine testicular (BTT) tubulin

	BBT	BTT
RBT	0.645 (0.460–0.776)	0.230 (−0.368–0.693)
BTT	0.187 (−0.465–0.708)	

Similar results were obtained if rat brain rather than bovine brain tubulin was used for the comparison.

Because the tubulin C-terminal species exhibit key differences in acidity that affects ionization efficiency [18], the spectra are not quantitative in a strict sense. This is particularly evident in the disparity between the sum of signal from α and β C-termini, which should be equal because α and β are equimolar. However, given the extensive reproducibility that has been demonstrated in this dataset, it is reasonable to consider that the trend shapes among the peaks reflect the population trends of tubulin C-terminal PTM in the tissue.

Close inspection of Figure 6a and b reveal subtle differences between them. The relative intensities of peaks the spectra have in common differ significantly, and some peaks exist in only one spectrum. The spectral dot product and 95% CI between these two consensus spectra (Table 2) show a high level of correlation while the dot products between RBT and BBT and that of BTT are low, and the 95% CI cross zero. As with replicate versus consensus comparisons, the 95% CI aids in the use of the spectral dot product because these intervals factor in the number of overlapping peaks to put the dot product into perspective. The finding of no statistical correlation between BTT and the brain spectra is expected; brain tubulin complexity is known to persist across species [19], but tubulin expression and modification patterns differ across tissues within the same species.

The example of similarities and differences between BTT and brain spectra highlights the point that spectral correlation both by dot product and 95% CI should be followed by investigation of similarities and differences between spectra with potential biological implication. Beyond yielding a concise parameter for spectral correlation, use of the dot product involves a perception of the spectrum, as a binned and centered vector, in ways that facilitate elucidation of these spectral details, and these benefits form the basis for further automation of queries into the nuances of complex spectra.

Conclusion

The experiments reported here have demonstrated improvements in the methods of data treatment that are likely to be generally applicable for interpretation of complex spectra. These include a rational approach to determining bin size and center for spectral correlation, the generation of a consensus spectrum, and attenuation of the replicate dataset through reproducibility monitoring with the spectral dot product and 95% CI,

the calculation and interpretation of inter-consensus dot products, and the assignment of peaks with an expected peak list. All of these techniques enhance the investigation of the complex spectra by tagging relationships and aiding significance determination. The methods presented here will form the basis for the analysis of more samples in the future, particularly those that are involved in human disease.

The data associated with this manuscript are available as part of the Supplementary Materials and may also be downloaded from the Proteome Commons Tranche using the hash key[§]: oXe+yy1vwgbBUFpdcy747Rl8rrJJcyc92AQVQfEUyAaHVHaDXb1sIO3UkIqoJi6G1exzhq8yxSlswZ4N5H7kh2vIvCUAAAAAAAAB8yQ = =

Or this link: http://www.proteomecommons.org/datadownloader.jsp?fileName = oXe+yy1vwgbBUFpdcy747Rl8rrJJcyc92AQVQfEUyAaHVHaDXb1sIO3UkIqoJi6G1exzhq8yxSlswZ4N5H7kh2vIvCUAAAAAAAAB8yQ = =

Acknowledgments

The authors thank Dr. Yana Krisemann and Vera Tran, summer students who contributed to this work in its early stages. This research was supported by the Intramural Research Program of the National Institute of Child Health and Human Development.

References

1. Stein, S. Estimating Probabilities of Correct Identification from Results of Mass Spectral Library Searches. *J. Am. Soc. Mass Spectrom.* **1994,** *5*, 316–323.
2. Wan, K. X.; Vidavsky, I.; Gross, M. L. Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle. *J. Am. Soc. Mass Spectrom.* **2001,** *13*, 85–88.
3. Frewen, B.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J. Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries. *Anal. Chem.* **2006,** *78*, 5678–5684.
4. Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J.; King, N.; Stein, S.; Aebersold, R. Development and Validation of a Spectral Library Searching Method for Peptide Identification from MS/MS. *Proteomics* **2007,** *7*, 655–667.
5. Stanchi, F.; Corso, V.; Scannapieco, P.; Ievolella, C.; Negrisolo, E.; Tiso, N.; Lanfranchi, G.; Valle, G. TUBA8: A New Tissue Specific Isoform of α-Tubulin that is Highly Conserved in Human and Mouse. *Biochem. Biophys. Res. Commun.* **2000,** *270*, 1111–1118.
6. Tuszynski, J.; Carpenter, E.; Huzil, J.; Malinski, W.; Luchko, T.; Ludena, R. The Evolution of the Structure of Tubulin and Its Potential Consequences for the Role and Function of Microtubules in Cells and Embryos. *Int. J. Dev. Biol.* **2006,** *50*, 341–358.
7. Kuster, B.; Schirle, M.; Mallick, P.; Aebersold, R. Scoring Proteomes with Proteotypic Peptide Probes. *Nat. Rev. Mol. Cell. Biol.* **2005,** *6*, 577–583.
8. Monigiatti, F.; Berndt, P. Algorithm for Accurate Similarity Measurements of Peptide Mass Fingerprints and Its Application. *J. Am. Soc. Mass Spectrom.* **2005,** *16*, 13–21.
9. Sackett, D.; Knipling, L.; Wolff, J. Isolation of Microtubule Protein from Mammalian Brain Frozen for Extended Periods of Time. *Protein Expr. Purif.* **1991,** *2*, 390–393.
10. Sackett, D. Rapid Purification of Tubulin from Tissue and Tissue Culture Cells Using Solid-Phase Ion Exchange. *Anal. Biochem.* **1995,** *228*, 343–348.
11. Wolff, J.; Sackett, D.; Knipling, L. Cation Selective Promotion of Tubulin Polymerization by Alkali Metal Chlorides. *Protein Sci.* **1996,** *5*(10), 2020–2028.
12. Woods, A.; Huang, A.; Cotter, R. J.; Pasternack, G.; Pardoll, D.; Jaffee, E. Simplified High-Sensitivity Sequencing of a Major Histocompatibility Complex Class I–Associated Immunoreactive Peptide Using Matrix Assisted Laser Desorption Ionization Mass Spectrometry. *Anal. Biochem.* **1995,** *226*(1), 15–25.
13. Zar, J. *Biostatistical Analysis*, 2nd ed.; Prentice-Hall: Englewood Cliffs, NJ, 1974; pp. 311–313.
14. Rao, S.; Åberg, F.; Nieves, E.; Horwitz, S. B.; Orr, G. A. Identification by Mass Spectrometry of a New α-Tubulin Isotype Expressed in Human Breast and Lung Carcinoma Cell Lines. *Biochemistry* **2001,** *40*, 2096–2103.
15. Verdier-Pinard, P.; Wan, F.; Burd, B.; Angeletti, R. H.; Horwitz, S. B.; Orr, G. A. Direct Analysis of Tubulin Expression in Cancer Cell Lines by Electrospray Ionization Mass Spectrometry. *Biochemistry* **2003,** *42*(41), 12019–12027.
16. Verdier-Pinard, P.; Shahabi, S.; Wang, F.; Burd, B.; Xiao, H.; Goldberg, G.; Orr, G. A.; Horwitz, S. B. Detection of Human β-Tubulin Expression in Epithelial Cancer Cell Lines by Tubulin Proteomics. *Biochemistry* **2005,** *44*(48), 15858–15870.
17. Rudiger, M.; Weber, K. Characterization of the post-translational modifications in tubulin from the marginal band of avian erythrocytes. *Eur. J. Biochem.* **1993,** *218*(107),116.
18. Lecchi, P.; Olson, M. T.; Brancia, F. L. The Role of Esterification on Detection of Protonated and Deprotonated Peptide Ions in Matrix Assisted Laser Desorption/Ionization (MALDI) Mass Spectrometry (MS). *J. Am. Soc. Mass Spectrom.* **2005,** *16*, 1269–1274.
19. Redeker, V.; Frankfurter, Parker, S.; Rossier, J.; Detrich, W. Post-translational Modification of Brain Tubulins from the Antarctic Fish *Notothenia coriiceps*: Reduced C-terminal Glutamylation Correlates with Efficient Microtubule Assembly at Low Temperature. *Biochemistry* **2004,** *43*, 12265–12274.

§ A hash key is a compressed encoding of a string x which, while not guaranteed to be unique, is statistically unlikely to "collide" with a hash computed from a different string y. This hash was generated using the MD5 hash algorithm and it may also be used to check that the data has not been changed since publication.