Published in final edited form as:

J Proteome Res. 2010 August 6; 9(8): 4152–4160. doi:10.1021/pr1003856.

Quantifying the impact of chimera MS/MS spectra on peptide identification in large scale proteomics studies

Stephane Houe^{†,§}, Robert Abernathy[†], Kutralanathan Renganathan[†], Karen Meyer-Arendt[†], Natalie G. Ahn^{†,§}, and William M. Old^{†,‡}

[†]Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309

[§]Howard Hughes Medical Institute, University of Colorado, Boulder, CO 80309

SUMMARY

A complicating factor for protein identification within complex mixtures by LC/MS/MS is the problem of “chimera” spectra, where two or more precursor ions with similar mass and retention time are co-sequenced by MS/MS. Chimera spectra show reduced scores due to unidentifiable fragment ions derived from contaminating parents. However, the extent of chimeras in LC/MS/MS datasets and their impact on protein identification workflows are incompletely understood. We report ChimeraCounter, a software program which detects chimeras in datasets collected on an Orbitrap/LTQ instrument. Evaluation of synthetic chimeras created from pairs of well-defined peptide MS/MS spectra reveal that chimeras reduce database search scores most significantly when contaminating fragment ion intensities exceed 20% of the targeted fragment ion intensities. In large scale datasets, the identification rate for chimera MS/MS is 2-fold lower compared to non-chimera spectra. Importantly, this occurs in a manner which depends not on absolute precursor ion intensity, but on intensity relative to the median precursor intensity distribution. We further show that chimeras reduce the number of accepted peptide identifications by increasing false negatives while showing little increase in false positives. The results provide a framework for identifying chimeras and characterizing their contribution to the poorly understood false negative class of MS/MS.

INTRODUCTION

A predominant method for large scale identification of proteins in complex mixtures is “bottom-up” proteomics, where proteins are proteolyzed into peptides and separated by reversed-phase chromatography prior to mass analysis by electrospray ionization mass spectrometry (LC/MS/MS). As peptide ions are detected, they are targeted and sequenced by intensity dependent selection and gas phase fragmentation. The resulting fragmentation spectra (MS/MS) are searched against a sequence database to identify peptide sequences and infer the proteins in the sample.

Low MS/MS identification rates and low discrimination by search programs result in poor reproducibility and under-sampling of proteins present in complex samples, and thus remain a major impediment to complete proteome sampling. Although modern hybrid ion trap mass spectrometers can acquire data at speeds up to 5 Hz, only a fraction of the collected MS/MS spectra can be successfully matched to peptide sequences with high confidence (usually in

[‡]Corresponding author: William M. Old, Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado 80309, Phone: 303-492-5519, Fax: 303-492-2439, william.old@colorado.edu.

Supporting Information Available: Supplementary Fig. S1 shows the comparison of ion injection time versus precursor intensity for chimeras and non-chimeras. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

the range of 10–30%). Many factors contribute to this effect. Complex gas phase fragmentation chemistry may result in MS/MS spectra with non-canonical fragment ions that are not considered by database search algorithms¹. The sequenced peptides may not be present in the database or may have unanticipated post-translational modifications². Data collection might also yield MS/MS spectra of poor quality due to low signal to noise, low proton mobility, or suboptimal collision energies.

Another complication arises when peptides with similar m/z ratios co-elute, generating spectra that we refer to as “chimera” MS/MS. Peptide fragmentation is achieved in an LTQ ion trap instrument by a two step process. First the precursor ion is isolated by ejecting all ions outside of an isolation m/z range of 2–3 Da. The trapped peptide ions are then dissociated using resonance activation and the resulting fragment ions are then detected using mass selective instability³. Chimeras result from the isolation and simultaneous fragmentation of two or more distinct molecular ions within the isolation m/z range. Fragments from multiple parent ions will be present in the MS/MS spectrum, increasing the number of unidentified fragments in sequence database searches. Commonly used search programs such as MASCOT and SEQUEST are dramatically affected by the presence of unidentified ions, leading to reduced search scores and poor discrimination of the co-fragmented peptides. Acquisition of chimeras in some methods is intentional, such as LC/MS^E sequencing⁴ where precursors over a large mass range are co-fragmented to circumvent the sampling issues inherent in data-dependent acquisition. These methods have been demonstrated for low complexity samples (e.g. prokaryotic proteins⁵). A data independent method using selection of 10 m/z windows was used to identify and profile *C. elegans* proteins⁶. A simulation study demonstrated the ability to identify chimeras acquired in high resolution mass analyzers⁷.

In protein samples of high complexity, the frequency of chimeras can be high enough to significantly affect identification rates. Hoopmann et al. examined the frequency of chimeras with a software tool, Hardklor, and estimated that 11% of MS/MS are chimeras, with an additional 29% of MS/MS with parent isotope distributions inconsistent with peptide analytes⁸. The effect of chimera sequencing on identification rates has not been explored to date, although it has been suggested to account for the suppression of reporter ion ratios in isobaric isotope labeling methods such as ITRAQ^{9,10}. Previous approaches for identifying chimeras used iterative database searching, or probabilistic approaches such as ProbIDtree¹¹, but such methods suffer from low sensitivity due to the extremely large number of combinatorial possibilities when considering mixtures of ions from two or more peptides.

Here, we explore the impact of chimeras on the process of peptide identification in data dependent analysis of complex samples, demonstrating a method for detecting and quantifying their effect on search engine discrimination. We describe ChimeraCounter, a software tool for predicting chimera MS/MS spectra from precursor isotope patterns, and use it to analyze the frequency of chimeras in shotgun proteomics experiments and to assess their effect on identification rates. Our results show that in a typical data-dependent LTQ-Orbitrap profiling analysis of complex samples, the percentage of chimeras may reach as high as 50% of total spectra, and that the rate of successful identification is 2-fold lower for chimeras compared to non-chimera MS/MS. Additionally, we analyze a medium complexity sample of known composition, and show that chimeras increase the false negative rate of peptide identification by suppressing search scores.

EXPERIMENTAL PROCEDURES

Data collection

LC-MS/MS was carried out using a Thermo LTQ-Orbitrap mass spectrometer interfaced with a Waters nanoAcquity UPLC, outfitted with a BEH C18 reversed phase column (25 cm $\times$ 75 μ m i.d., 1.7 μ m, 100Å, Waters). Peptide mixtures (5 μ L, 0.2–20 μ g) were loaded and separated by a linear gradient from 95% Buffer A (0.1% formic acid) to 40% Buffer B (0.1% formic acid, 80% acetonitrile) over 120 min at flow rate 300 nL/min.

MS/MS were collected enabling monoisotopic precursor and charge selection settings. Ions with unassigned charge state or charge state = 1 were excluded. For each MS scan, the 10 most intense ions were targeted with dynamic exclusion 30 s, 1 Da exclusion width, and repeat count = 1. The maximum injection time for Orbitrap parent scans was 500 ms, allowing 1 microscan and AGC = 1×10^6 . The maximal injection time for the LTQ MS/MS was 250 ms, with 1 microscan and AGC 1×10^4 . The normalized collision energy was 35%, with activation Q = 0.25 for 30 ms.

Samples

Tryptic peptides derived from human leukemia cells (K562) were used as a standard sample. The initial concentration of K562 digests was 4 μ g/ μ L. K562 was sequentially diluted to 0.4 μ g/ μ L and 0.04 μ g/ μ L with Buffer A, and 5 μ L of each solution was run in triplicate. The Sigma universal protein standard (UPS1, Sigma Aldrich) was used as the defined protein mixture standard, containing 48 purified human recombinant proteins present in equimolar ratios¹².

Search programs

MS/MS were searched against a human protein database (IPI v.3.27) using MASCOT, with ions score thresholds set to false discovery rate (FDR) = 0.01 using inverted database searching¹³. MASCOT ions score thresholds for MH_2^{+2} , MH_3^{+3} , and MH_4^{+4} and above were 32.2, 23.7 and 25.3, respectively. Parent ion tolerances were set to 50 ppm on the monoisotopic peak (A_0) and the first isotopic peak (A_1) and the fragment ion tolerance was set to 0.8 Da, allowing 1 missed cleavage. False discovery rates shown in receiver operating characteristic (ROC) curves were calculated as q-values, which are the minimum FDR at which a given MS/MS assignment would be accepted^{13,14}.

ChimeraCounter

Chimera spectra were recorded by developing a software program, ChimeraCounter, written in Python. ChimeraCounter examines Orbitrap precursor scans in order to identify MS/MS attempts with more than one isotopic distribution ("family") of ions (Fig. 1). An MS/MS was deemed to be a chimera when peaks that were distinct from the isotopic peaks of the targeted precursor were present within ± 1 Da of the targeted precursor, and had peak height greater than a specified percentage of the precursor peak height. We call this metric the "percent chimera intensity (PCI)". RAW data files from the LTQ-Orbitrap were used to generate mzXML files using the ReAdW software¹⁵. The m/z for the precursor ion was used to find the nearest peaks within the parent scan of each MS/MS attempt. The charge of the precursor was used to identify all peaks within the isotopic family of the precursor, using m/z tolerance of 0.01 Da. This was done to compensate for errors in centroid peak locations. The peaks in the targeted precursor ion series were then removed from consideration. Intensities of remaining peaks within the $[-1.0, 1.0]$ m/z window, centered on the precursor ion m/z location, were then evaluated. When the ratio of peak height to precursor ion peak height exceeded a user-defined value, the MS/MS was scored as a chimera and peak m/z were reported.

Simulated chimeras were constructed from 686 pairs of MS/MS spectra from LC-MS/MS datasets of tryptic digests of total cellular protein from K562 erythroleukemia cells, corresponding to 150 pairs of distinct sequences and 536 pairs of identical sequences (Table 1, 20 μ g). The MS/MS spectra were selected from high confidence assignments, by first removing all MS/MS that had been identified as chimera, then filtering the remaining spectra for MASCOT ions score between 32 and 60. The score range used for selecting spectra was high enough to pass the threshold for confident identification, but low enough to avoid spectra dominated by very high intensity fragment ions, which would be insensitive to variations introduced by added spectra. Spectra in each pair were selected to have the same charge state, a precursor mass difference no greater than 50 ppm, and similar precursor ion intensities that varied by no more than 3-fold. The spectra within each pair were each normalized to their base peak intensity. The second spectrum's intensities were scaled by multiplying by a given fraction ("mixing ratio") before merging it with the first spectrum to yield each synthetic chimera spectrum.

The resulting chimera spectra were then "recentroided" to reproduce the peak spacing generated by the centroid algorithm used by LTQ acquisition software to record MS/MS spectra. First, fragments ions were sorted by descending intensity. Second, every peak within ± 0.7 Da of the highest intensity peak was centroided using a center-of-mass calculation for the m/z locations, and a centroid peak intensity was calculated from the sum of peak intensities. Third, each of the peaks in the window was removed from the pool of starting peaks. The process was repeated until all peaks were combined in the windowing/centroiding process. The parent ion m/z was taken as the monoisotopic m/z from spectrum A. The resulting spectra were then written to a MGF file for submission to MASCOT.

RESULTS AND DISCUSSION

Detection of chimera MS/MS

Chimera MS/MS result from sequencing two or more distinct molecular ions which are similar in mass and co-elute within a small time window (Fig. 1). Precursor isolation is usually performed with a broad mass window (2–3 Da) to maximize sensitivity of the MS/MS scan. This however occurs at the expense of increasing the frequency of chimera MS/MS when analyzing complex samples, because the chances that co-eluting ions fall within the isolation window become significant.

In order to detect chimera MS/MS, we developed the ChimeraCounter software program, which examines the isotopic signatures in the full scan MS preceding the MS/MS (outlined in Fig. 2). When two or more peptide ions are close in elution time and m/z , the isotopic peaks of these parent ions should overlap in the preceding high resolution (Orbitrap) MS scan (Fig. 1A). Any peaks not consistent with those of the targeted parent ion indicate the presence of a chimera MS/MS. In the high resolution MS scan, a targeted peak is defined as that peak nearest the targeted m/z reported in the MS/MS header, and contaminating peaks are defined as those which are unrelated to the isotopes of the targeted ion, and within a given m/z tolerance (see Methods). We define the percent chimera intensity (PCI) as the intensity of the highest contaminating peak expressed as a percentage of the targeted peak intensity. The PCI estimates the relative abundances of the co-sequenced precursors and thus their associated total fragment ion intensities, assuming equal isolation and fragmentation efficiency of the multiple precursors. The PCI threshold may thus be used to define which MS/MS spectra are likely chimeras. Above a low PCI threshold of 5%, nearly 90% of the MS/MS in an LC/MS/MS dataset would be labeled as chimeras, yet only a fraction of these spectra would show contaminating fragment ions at a level which significantly affected identification scores. Thus, it is critical to establish the threshold for the PCI that predicts

those chimeras that lead to suppressed database search scores and/or false positive identifications.

We evaluated the PCI at which a co-sequenced contaminating peptide significantly affects peptide identification scores. To do this, we created simulated chimera MS/MS by combining spectra confidently assigned to distinct peptide sequences together in different pair wise combinations. By varying the ratio of base peak intensity from each of two spectra, we simulated variable contributions of the contaminating ion, and used this “fragment ion mixing ratio” to estimate the critical PCI. After searching simulated chimeras against the human IPI database using MASCOT, we measured the effect of different fragment ion mixing ratios on MASCOT scores. As fragment ions from the contaminating spectra were added in increasing amounts, we expect that scores would be reduced compared to a homogeneous peptide MS/MS.

Synthetic chimera MS/MS were created by combining fragments from pairs of spectra drawn from a set of confidently identified peptide MS/MS, each representing a single peptide-spectrum match. A single LTQ/Orbitrap LC/MS/MS dataset was collected on a tryptic digest of human K562 cell proteins (20 µg), and peptide-spectrum matches were selected according to criteria indicated in Fig. 3. The matches were filtered by MASCOT score, retaining those with ions scores between 32 and 60, to avoid spectra dominated by very high signal-to-noise fragment ions. Any MS/MS that were scored as chimera by ChimeraCounter were removed from consideration, using $PCI \geq 10\%$. This low threshold stringently eliminated large numbers of MS/MS, and ensured that the remaining spectra reflected single peptides. Peptide-spectrum matches were then added in pair wise combinations when their precursor masses were within 50 ppm of each other and their precursor ion intensities were less than 3-fold apart, in order to minimize the effect of varying intensities and signal to noise on subsequent searching.

Synthetic chimera MS/MS were created by mixing the paired spectra with higher and lower intensity, referred to as Spectrum A and Spectrum B respectively, and scaling the fragment ions in Spectrum B between 0–50% of their original intensities. After searching the synthetic chimera using MASCOT, we examined cases where the search program correctly identified the peptide corresponding to Spectrum A. The ions scores varied widely for the peptide-spectrum matches used in this experiment (from 32 to 60), therefore, we calculated the difference in ions score between the chimera and the homogeneous Spectrum A ($\Delta IonsScore = IonsScore_{Spectrum A} - IonsScore_{chimera}$).

Fig. 4A shows the cumulative histogram of spectra versus $\Delta IonsScore$, for Spectra A and for the synthetic chimera composed of Spectra A and different ratios of Spectra B. As expected, the scores fell to lower values with chimeras, and increasing the contribution of Spectra B led to a further reduction in ions score. When Spectra B were added at 5% intensity of Spectrum A, 98% of the synthetic spectra showed $\Delta IonsScore \leq 10$, indicating that chimeras with low amounts of contaminating ions have little effect on ions scores. In contrast, when Spectra B were 50% of Spectra A, 65% of the chimeras showed $\Delta IonsScore \geq 10$. These findings revealed a significant deterioration in score when MS/MS are comprised of fragment ions from more than one peptide.

These experiments were performed by combining spectra assigned to distinct peptide sequences (i.e., different scan, different peptide ID), which could not control for the possibility that summing spectra from the same peptides together might also affect scoring. To test this possibility, we mixed MS/MS spectra corresponding to the same peptide sequence (i.e., different scan, same peptide ID), and plotted the cumulative histogram against $\Delta IonsScore$ (Fig. 4B). When spectra were summed, most $\Delta IonsScore$ values were

negative, indicating that ions scores increased when spectra were added together, compared to single spectra. This indicated improved spectral quality following summation of most spectral pairs, most likely due to increased signal to noise as seen in signal averaged spectra. Nevertheless, the effect was relatively small; regardless of the fragment ion mixing ratio, 85% of cases showed $\Delta\text{IonsScore}$ greater than -10 and less than $+10$, indicating minimal effects of summation on ions scores.

Next, we determined the threshold for the fragment ion mixing ratio at which identification of the most abundant peptide (peptide A, corresponding to Spectrum A) deteriorated. This was done by measuring the percentages of MS/MS where MASCOT correctly identified peptide A as the top-ranked assignment (Fig. 4C). When peptide B was present at 20%, peptide A was the top ranked assignment in 98.6 % of the synthetic chimeras. Above 20% peptide B, this percentage fell continuously, such that peptide A was correctly identified in only 87% of cases containing 50% of Spectrum B mixed with Spectrum A. This experiment shows that the presence of fragment ions from contaminating non-targeted peptides interferes with the identification of targeted ions by decreasing MASCOT scores proportional to the ratio of mixing. Fig. 4C shows that top ranking of the correct peptide is most dramatically reduced at fragment ion ratios above 20% peptide B. Thus, we chose $\text{PCI} \geq 20\%$ as the threshold for designating chimera MS/MS throughout the study.

Effect of sample loading on chimeras and success rate

The simulation results (Fig. 4) indicated that fragments from contaminating co-sequenced peptides suppress MASCOT scores. This would predict lower rates of successful identifications for chimera MS/MS, as the scores fall below the score threshold for stringent acceptance. We examined this in LC/MS/MS datasets of a human K562 tryptic digest, performed at three different sample loadings (20, 2 and 0.2 μg), each analyzed in three technical replicates. Table 1 shows that the number of MS/MS attempts ranged from 11,000/run at the lowest loading up to 20,000/run at the highest loading. The number of spectra that could be matched to peptides with high confidence ($\text{FDR} \leq 0.01$) ranged from 20–28% of all MS/MS attempts.

Precursor scans were then inspected using ChimeraCounter to identify those with significant contributions from contaminating ions. The analysis revealed that 49–55% of all MS/MS attempts were chimera spectra, using the threshold established above ($\text{PCI} \geq 20\%$). Over the 100-fold decrease in sample loading, the number of detectable features decreased by two-fold, but the percentages of chimeras remained the same. Of the spectra labeled as chimera, only 11–18% were successfully matched with peptides; this was significantly lower than non-chimera spectra, where 30–40% could be identified with high confidence. Thus, for chimera spectra, the rate of successful identifications was lower by more than two-fold compared to non-chimeras.

Aspects of data acquisition were compared for the chimera and non-chimera MS/MS. Modulation of ion injection time enables an ion trap with finite capacity to efficiently trap and sequence ions over very large dynamic ranges, while avoiding the deleterious effects of space charging. This relies on automatic gain control software which estimates the flux of ions entering the instrument within a 2 Da mass window around each precursor ion of interest. Ideally, the targeted precursor is isolated to the exclusion of all other ions, resulting in relatively pure fragmentation spectra. However, the presence of contaminating ions in chimeras complicates the estimation of ion flux, leading to systematic differences in ion injection time between chimeras and non-chimeras as a function of the targeted precursor intensity (Suppl. Fig S1). At any precursor ion intensity, ion injection times for chimera were systematically lower, as expected if the actual number of ions trapped was higher than indicated by the targeted precursor intensity.

Thus, the reduced numbers of chimera identifications could be explained by the trend towards reduced scoring due to unidentified fragment ions, revealed by the simulation studies. Alternatively, reduced identifications might be explained by poorer quality spectra for chimera MS/MS with weaker fragment ion intensities for targeted peptides. To distinguish between these possibilities, we partitioned chimera and non-chimera MS/MS into 11 ranges of precursor ion intensity and evaluated the number of chimeras within each range. Fig. 5A–C shows histograms of spectra *versus* precursor ion intensity, where the expected shift of MS/MS spectra to lower intensity at reduced sample loading was obvious. As precursor ion intensity decreased, the percentage of MS/MS that were scored as chimera (PCI $\geq 20\%$) increased. Thus, the ions of lowest intensity were enriched in chimera MS/MS, as expected due to the higher density of ions with comparable intensities in this range. Interestingly, the distribution of chimeras *versus* intensity showed bimodality at 20 and 2 μg loadings. This implied that two factors contribute to chimeras. At very low intensity, the chimera frequency reached 90% of spectra, at all loading amounts. We attribute these chimeras to noise from the LC/MS/MS system, which would be constant at any sample loading. At higher intensities, chimeras increased in a manner which tracked the precursor ion intensities, reaching 65% at intensities ~ 5 -fold below the median intensity values of precursor ions, and decreasing at lower intensities. We attribute chimeras in this peak to the presence of contaminating peptide ions where the likelihood of co-elution with peptides of similar m/z is much higher. The results reveal that it is not simply precursor ion intensity, but rather intensity relative to the median value for all ions, which determines the frequency of occurrence of chimeras. This explains why the sample loading had little effect on the frequency of chimeras.

We next evaluated the “success rate” of peptide matches, defined as the percentage of MS/MS that were successfully identified with high confidence. Fig. 5D–F plots success rates against the intensities of precursor ions. As expected, the overall success rate for all ions increases with intensity, most likely due to improved MS/MS signal to noise. Importantly, the success rates for chimeras were systematically lower than for non-chimera MS/MS within each intensity bin. Thus, the difference in success rate between non-chimeras and chimeras is independent of absolute intensity, which indicates that it is not the enrichment in chimera spectra that determines the lower success rate of lower intensity ions. From the success rates for chimera and non-chimera MS/MS, along with the number of chimera in each intensity range, we calculated the overall impact of chimeras on peptide identifications. We estimate that the number of identified peptides would increase by more than 30% without chimera MS/MS, regardless of sample loading.

Additional insight emerged when analyzing trends in success rate *versus* the precursor ion intensity distribution, where the success rate increased as sample loadings decreased from 20 to 0.2 μg (Fig. 5D–F). For example, for the 20, 2 and 0.2 μg loadings, the success rates for precursor ions with intensity 1×10^6 were 32%, 35% and 50% for non-chimeras, and 15%, 20%, and 40% for chimeras. This was counterintuitive, because it showed that success rate is not dependent on absolute precursor ion intensity. Instead, success rate depends on the precursor ion intensity relative to the median intensity distribution (Fig. 5A–C). Thus, there are fewer ions with intensity of 1×10^6 or greater in experiments at sample loadings of 0.2 μg , compared to 20 μg .

We next plotted the success rate *versus* the percentages of chimeras, using measurements within each intensity range. Fig. 6 shows that the frequency of chimeras was inversely correlated with success rate, revealing a strong linear correlation ($R^2 = 0.96$). Surprisingly, the slopes and intercepts were similar at each sample loading. This indicates that the correlations between chimeras and success rate depend not on absolute precursor ion intensity, but rather on intensity relative to the median distribution. Other sample types

analyzed in the same manner showed similar correlations. The lower success rate for chimera spectra can be explained by simulation results showing suppression of search scores due to co-sequenced ions in chimera MS/MS.

Effect of chimeras on false negative identifications

The results above suggested that chimeras will exert their greatest influence on peptide identifications by increasing false negative assignments (i.e., peptides rejected due to poor scores), rather than by increasing false positive assignments (i.e., by increasing score thresholds at a given false discovery rate). However, the effect of chimeras on false negative rates is difficult to assess in shotgun datasets of complex mixtures because the number of true identifications is unknown. We therefore examined chimeras in a dataset of a standard protein mixture whose composition is completely known. LC/MS/MS was performed on a tryptic digest of the Sigma UPS1 standard, which contains 48 purified human recombinant proteins present in equimolar ratios¹². MS/MS spectra were searched against the human IPI 3.27 protein database concatenated with the 48 proteins in this defined mixture with additional proteins identified by the ABRF Proteins Standards Research Group Bioinformatics Committee (104 sequences in total)¹⁶. Because the proteins present are known, MS/MS assignments to peptides to protein standards can be assumed true, while assignments to non-standard proteins can be assumed false. False negatives (FN) are estimated by the number of MS/MS assignments which were true but rejected due to low scores, and false positives (FP) are estimated by the number of assignments accepted but false. The false negative rate was calculated as FN divided by the total number of class true assignments ($FNR = FN/True$). In this way we can differentiate the effects of chimeras on FPs and FNR in a typical LC/MS/MS experiment.

Chimeras were assigned using ChimeraCounter with $PCI \geq 20\%$. We found that the FNR for chimeras (53%) was 2-fold higher than for non-chimeras (28%), presumably due to suppression of MASCOT scores below thresholds corresponding to $FDR=0.01$ (Table 2). Similarly, at $FDR = 0.05$, the FNR was 2.5 fold higher for chimeras vs. non-chimeras (36% *versus* 14%). Receiver operating characteristic (ROC) curves illustrated this effect over the entire range of scores^{13,14} (Fig. 7). The difference in sensitivity ($1-FNR$) was large, especially at low FDR values (< 0.02) typically used in large scale proteomics. The dramatic difference in discrimination between chimeras and non-chimeras was due to higher rejection of true assignments among chimeras. We did not observe increased false positives among chimeras, which would have required an increase in score thresholds (Table 2). Thus, the largest impact of chimeras on peptide identification is due to the suppression of search scores for true assignments, rather than increased random sequence matches and false positive identifications.

CONCLUSIONS

Chimera spectra significantly complicate the analysis of complex protein mixtures, yet their effects on peptide and protein identification in complex mixtures is poorly understood. Here we present ChimeraCounter software for automated detection and quantification of chimera spectra, which identifies chimeras based on isotopic signatures in full scan MS spectra, and allows quantitative assessment of their effect on searching and success rate at varying precursor ion intensities and levels of ion contamination. Our analyses of large scale datasets of soluble lysate proteins as well as defined protein mixtures leads to four novel and important conclusions. First, the impact of chimeras on search scores is low until the intensity of the contaminating ion within a chimera reaches 20% of the intensity of the major ion; above this threshold, MASCOT scores deteriorate rapidly. Second, more than 50% of MS/MS spectra in large scale datasets represent chimeras using a threshold of $\geq 20\%$ contaminating ion, which is higher than previously estimated. Third, the success rate for

identifications among chimeras is more than 2-fold lower than among non-chimera spectra, due to the effect of chimeras in suppressing search scores for true assignments; thus, chimeras mainly increase false negative identifications, with less effect on false positive identifications. Fourth, the frequency of chimeras unexpectedly tracked median precursor ion intensities, rather than dominate precursor ions with low intensities. Thus, the success rate is independent of absolute intensity, and sample loading has little effect on the frequency of chimeras. Together, these findings yield new insights into the behavior of chimeras which are important to consider in large scale LC/MS/MS experiments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ABBREVIATIONS

LC/MS/MS	liquid chromatography and mass spectrometric peptide sequencing
UPLC	ultra-high pressure liquid chromatography
AGC	automatic gain control
UPS1	universal proteomics standard-1 sample
IPi	international protein index
ROC	receiver-operator characteristic
TP	true positive
FP	false positive
TN	true negative
FN	false negative
FDR	false discovery rate ($=FP/(TP+FP)$)
FNR	false negative rate ($=FN/(TP+TN)$)
MGF	MASCOT generic file
PCI	percent chimera intensity

Acknowledgments

This work was supported by National Cancer Institute grants R01 CA126240 and R01 CA125291, part of NCI Clinical Proteomic Technologies for Cancer (<http://proteomics.cancer.gov>) initiative, and NIH grant R01 CA118972 (NGA).

LITERATURE CITED

1. Paizs B, Suhai S. Fragmentation pathways of protonated peptides. *Mass Spectrom Rev.* 2005; 24:508–548. [PubMed: 15389847]
2. Picotti P, Aebersold R, Domon B. The implications of proteolytic background for shotgun proteomics. *Mol Cell Proteomics.* 2007; 6:1589–1598. [PubMed: 17533221]
3. Schwartz JC, Senko MW, Syka JEP. A two-dimensional quadrupole ion trap mass spectrometer. *Journal of the American Society for Mass Spectrometry.* 2002; 13:659–669. [PubMed: 12056566]
4. Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics.* 2006; 5:144–156. [PubMed: 16219938]

5. Silva JC, Denny R, Dorschel C, Gorenstein MV, Li GZ, Richardson K, Wall D, Geromanos SJ. Simultaneous qualitative and quantitative analysis of the *Escherichia coli* proteome: a sweet tale. *Mol Cell Proteomics*. 2006; 5:589–607. [PubMed: 16399765]
6. Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods*. 2004; 1:39–45. [PubMed: 15782151]
7. Masselon C, Pasa-Tolic L, Lee SW, Li L, Anderson GA, Harkewicz R, Smith RD. Identification of tryptic peptides from large databases using multiplexed tandem mass spectrometry: simulations and experimental results. *Proteomics*. 2003; 3:1279–1286. [PubMed: 12872228]
8. Hoopmann MR, Finney GL, MacCoss MJ. High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal Chem*. 2007; 79:5620–5632. [PubMed: 17580982]
9. Ow SY, Salim M, Noirel J, Evans C, Rehman I, Wright PC. iTRAQ Underestimation in Simple and Complex Mixtures: The Good, the Bad and the Ugly. *Journal of Proteome Research*. 2009; 8:5347–5355. [PubMed: 19754192]
10. Karp NA, Huber W, Sadowski PG, Charles PD, Hester SV, Lilley KS. Addressing accuracy and precision issues in iTRAQ quantitation. *Mol Cell Proteomics*. 2010 [Epub ahead of print].
11. Zhang N, Li XJ, Ye M, Pan S, Schwikowski B, Aebersold R. ProbiDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics*. 2005; 5:4096–4106. [PubMed: 16196091]
12. Andrews, PC.; Arnott, DP.; Gawinowicz, MA.; Kowalak, JA.; Lane, WS.; Lilley, KS.; Martin, LT.; Stein, S. ABRF 2006. Long Beach, CA: 2006. ABRF-sPRG 2006 study: a proteomics standard. Poster available at <http://www.abrf.org/ResearchGroups/ProteomicsStandardsResearchGroup/EPosters/ABRFsPRGStudy2006poster.pdf>.
13. Kall L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res*. 2008; 7:29–34. [PubMed: 18067246]
14. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*. 2003; 100:9440–9445. [PubMed: 12883005]
15. Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK, Kapp E, McComb ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu W, Aebersold R. A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol*. 2004; 22:1459–1466. [PubMed: 15529173]
16. Lane, WS.; Nesvizhskii, AI.; Searle, B.; Tabb, DL.; Kowalak, JA.; Seymour, SL. ABRF 2007. Tampa, FL: 2007. Bioinformatic Evaluation of Datasets Derived from the ABRF sPRG Proteomics Standard. Poster available at <http://www.abrf.org/ResearchGroups/ProteomicsInformaticsResearchGroup/Studies/sPRG-BIC2007poster.pdf>.

Figure 1. Spectral chimera MS/MS

(**A**) Examination of a high resolution Orbitrap scan shows more than one precursor ion within the isolation window for MS/MS. In this example, A_0 , A_1 , and A_2 are isotope peaks from the targeted precursor ion, and B_0 , B_1 , and B_2 are isotope peaks from a contaminating precursor ion. (**B**) The MS/MS spectrum contains fragment ions from A (blue) and B (red) precursor ions. Observed fragment ions are annotated on each peptide sequence.

Figure 2. Process diagram for ChimeraCounter

Orbitrap scans of precursor ion peaks are processed to remove isotope peaks, after which ChimeraCounter records m/z and intensities of all precursor ions within each isolation window. Chimeras are scored when more than one peak is present and the ratio of peak intensity normalized to the highest peak in the window is greater than a specified percent chimera intensity (PCI) value.

Figure 3. Strategy for constructing chimera MS/MS with varying PCIs

Synthetic chimera MS/MS were constructed from pairs of MS/MS spectra from high confidence assignments mixed in varying ratios and recentroided. The ratios of base peak intensity between the precursor ions for the two spectra were used to calculate PCI as $Intensity_B \div Intensity_A \times 100$.

Figure 4. Chimeras lead to reduced search scores

(A) MS/MS spectra (A & B) were mixed in varying ratios, and cumulative percentages are plotted versus $\Delta\text{IonScore}$, calculated as $\text{IonScore}_{\text{chimera}}$ subtracted from $\text{IonScore}_{\text{Spectrum A}}$. For most chimeras, the ions score decreases as Spectrum B increases.

(B) Controls for spectral summation are performed by pairing different MS/MS spectra corresponding to the same peptide sequence. Adding two spectra together increases the ions score, suggesting that in panel A, the small percentage of chimeras with ions score higher than Spectrum A are an effect of summing spectra. (C) Fractions of chimeras matched correctly to peptide A are plotted versus PCI. Above PCI=20%, incorrect matches occur in 2% or more cases.

Figure 6. Chimeras and success rate are inversely correlated

Plots of success rate *versus* chimeras within each intensity range show a linear inverse correlation, which is invariant with sample loading.

Figure 7. Chimeras show lower discrimination between true and false assignments during automated searching

ROC curves plot Sensitivity (1-false negative rate) versus False Discovery Rate (FDR, q-value corrected) for chimera (PCI $\geq 20\%$) and non-chimera spectra. FNR and FDR values are determined from searches of datasets collected on protein standards, assuming that matches to protein standards are true and all other matches are false.

Table 1

Quantifying chimeras in LC/MS/MS analyses of complex proteolytic digests.

Sample ^a		MS/MS ^b		High Confidence IDs ^c		Chimeras ^d	Success Rate (%) ^e		
Loading	Run	Total	Peptides	Unique Peptides	Proteins	Total (% of all MS/MS)	All Spectra	Chimeras	Non-Chimeras
20 µg	Rep1	20,683	5,799	3,594	1,220	10,909 (53)	28	18	40
	Rep2	20,727	5,794	3,571	1,253	10,576 (51)	28	17	39
	Rep3	19,359	5,364	3,368	1,195	10,430 (54)	28	18	39
	Average	20,256	5,652	3,511	1,223	10,638 (53)	28	18	39
2 µg	Rep1	18,392	4,607	3,440	1,224	9,159 (50)	25	15	35
	Rep2	19,912	4,955	3,720	1,271	9,856 (49)	25	15	35
	Rep3	19,525	4,744	3,476	1,218	9,847 (50)	24	14	34
	Average	19,276	4,769	3,545	1,238	9,621 (50)	25	15	35
0.2 µg	Rep1	11,984	2,737	2,279	896	6,055 (51)	23	13	33
	Rep2	10,860	2,229	1,883	823	5,916 (54)	21	12	31
	Rep3	12,259	2,874	2,355	918	5,963 (49)	23	14	33
	Average	11,701	2,613	2,172	879	5,924 (50)	22	13	32

^a Human K562 cytosolic proteins were proteolyzed with trypsin and examined by 1D-LC/MS/MS at varying sample loadings.

^b Total MS/MS attempts.

^c Peptides and proteins were identified using MASCOT, searched with tolerances 50 ppm parent ion mass and 0.7 Da fragment ion mass. MASCOT ions score cutoffs were set at FDR=0.01 as determined by a separate reversed database search.

^d Chimeras were quantified using ChimeraCounter as described in Methods.

^e Success Rate calculated as (Number of peptides identified)÷(Total MS/MS attempts)

Table 2
False negative rates for peptide identifications differ between chimera and non-chimera MS/MS.^a

Matches to Standard Proteins (True) ^b				Matches to Other Proteins (False) ^b				
All Matches	Accepted ^c	Rejected ^c	FNR ^d	All Matches	Accepted ^c	Rejected ^c	FDR ^e	
FDR=0.01:								
Total	913	575	338	0.37	2,296	5	2,291	0.009
Chimera	331	155	176	0.53	1,053	0	1,053	0.00
Non-chimera	582	420	162	0.28	1,243	5	1,238	0.012
FDR=0.05:								
Total	913	709	204	0.22	2,296	36	2,260	0.048
Chimera	331	211	120	0.36	1,053	9	1,044	0.041
Non-chimera	582	498	84	0.14	1,243	27	1,216	0.051

^a LC/MS/MS was performed on a tryptic digest of a protein standard mixture containing 48 purified human recombinant proteins. Peptides were identified using MASCOT by searching against the human protein database (IPI v.3.27) concatenated with the 48 proteins in this defined mixture with additional proteins identified by the ABRF Proteins Standards Research Group Bioinformatics Committee (104 sequences in total)¹⁶. Chimeras and non-chimeras were identified and quantified using ChimeraCounter.

^b Identified peptides are scored True when they matched one of the 104 proteins within the protein standard mix, and scored False when they matched proteins not contained within the standard mix.

^c Peptides were Accepted or Rejected when their peptide MASCOT ions scores were respectively above or below thresholds of FDR=0.01 or 0.05.

^d False negative rates (FNR) were calculated as $[FN (Rejected, True)]/[Class True (Accepted, True + Rejected, True)]$

^e False discovery rates (FDR) were calculated as $[FP (Accepted, False)]/[TP+FP (Accepted, True + Accepted, False)]$