RESEARCH ARTICLE

Cleaning of raw peptide MS/MS spectra: Improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise

Nedim Mujezinovic[1], *Günther Raidl*[2], *James R. A. Hutchins*[1], *Jan-Michael Peters*[1], *Karl Mechtler*[1]* *and Frank Eisenhaber*[1]*

[1] Research Institute of Molecular Pathology, Vienna, Austria
[2] Institute of Computer Graphics and Algorithms, Vienna University of Technology, Vienna, Austria

The dominant ions in MS/MS spectra of peptides, which have been fragmented by low-energy CID, are often *b*-, *y*-ions and their derivatives resulting from the cleavage of the peptide bonds. However, MS/MS spectra typically contain many more peaks. These can result not only from isotope variants and multiply charged replicates of the peptide fragmentation products but also from unknown fragmentation pathways, sample-specific or systematic chemical contaminations or from noise generated by the electronic detection system. The presence of this background complicates spectrum interpretation. Besides dramatically prolonged computation time, it can lead to incorrect protein identification, especially in the case of *de novo* sequencing algorithms. Here, we present an algorithm for detection and transformation of multiply charged peaks into singly charged monoisotopic peaks, removal of heavy isotope replicates, and random noise. A quantitative criterion for the recognition of some noninterpretable spectra has been derived as a byproduct. The approach is based on numerical spectral analysis and signal detection methods. The algorithm has been implemented in a stand-alone computer program called MS Cleaner that can be obtained from the authors upon request.

Received: December 23, 2005
Revised: June 8, 2006
Accepted: June 12, 2006

Keywords:
Charge deconvolution / Deisotoping / Noise reduction / Raw mass spectrum processing

1 Introduction

Developments in modern MS have made the large-scale analysis of cellular proteomes possible [1–3]. LC coupled with MS/MS (LC-MS/MS) is the standard technique used for the analysis of complex protein mixtures [4, 5]. Since modern mass spectrometers can generate large datasets with high-throughput, computational analysis of thousands of spectra has become the major bottleneck. Both the accuracy of the computer-generated interpretations (the identity of the proteins and their PTMs) as well as the time and the storage requirements for their computation are a matter of concern.

In many cases, but not always, *b*- and *y*-ions and their derivatives resulting from cleavage at peptide bonds are the most dominant signals in MS/MS spectra of peptides after their fragmentation by low energy CID [5–12]. However, MS/MS spectra typically contain many more peaks than can

Correspondence: Dr. Frank Eisenhaber, Research Institute of Molecular Pathology, Dr. Bohr-Gasse 7, A-1030 Vienna, Austria
E-mail: Frank.Eisenhaber@imp.univie.ac.at
Fax: +43-1-7987-153

Abbreviations: ADH, alcohol dehydrogenase; **IIR**, infinite impulse response; **LC-MS/MS**, LC coupled with MS/MS; **SMC**, structural maintenance of chromosome; **TRF**, transferrin

* These authors contributed equally to this work.

be expected from this fragmentation scheme. Some of them are repeated shifted signals due to the natural isotope distribution [13]. The heavy isotope variants and the monoisotope peak form isotope peak clusters that can be detected with high-resolution instruments. ESI allows measuring the masses of large molecules by producing multiply charged ions, thereby decreasing the *m/z* into detectable ranges [14–18]. If a fragment ion comprises several functional groups capable of acting as a charge carrier, the same isotope peak cluster can be repeated with a different charge state at different *m/z* values in the spectrum. Other signals originate from unknown fragmentation pathways, sample-specific or systematic chemical contaminations, and random noise produced by the electronic detection system.

It is hardly possible to derive any benefit from the above-mentioned additional background peaks that can compose the majority of the spectrum as long as the theoretical understanding of the mechanism of their genesis is scarce. The presence of these peaks not only complicates computer-based spectrum interpretation by increasing the computation time, but also, more critically, false interpretation of high-intensity signals as potential *b*- or *y*-related ions can lead in some cases to incorrect sequence interpretations of proteins or false identification of their PTMs. Particularly, the *de novo* sequencing approach [19–25] is affected by this problem, where each peak is part of a sequence puzzle to be solved, and therefore has initially to be considered as a potential *b*- or *y*-ion. In the case of algorithms based on protein sequence database searches [26–32], the danger of misinterpretation is not so dramatic, especially for protein targets without PTMs, since the space of naturally occurring protein sequences is much smaller than the set of sequences that can be theoretically generated. Usually, a few dominating peaks of the major fragmentation row in the spectrum are sufficient to unambiguously determine the register of a peptide fragment within the original protein sequence. But when the nature of possible PTMs is *a priori* unknown (and, therefore, the mass changes to be anticipated vary widely) or when the database contains many proteins with similar peptides, the background can lead database search methods down a wrong path and result in incorrect protein identification.

Background processing of raw MS/MS spectra from protein samples has not been in the center of interest among the community for a long time, partly due to limitations of measurement accuracy. For example, resolution of isotope clusters requires very precise instruments, which have become available on a broad scale only recently (*e.g.*, the Thermo Finnigan LCQ with close to ~0.5 Da resolution and ~0.3 Da accuracy of mass measurement or the newer LTQ with ~0.3 Da resolution and ~0.2 Da mass determination accuracy). Therefore, some spectrum interpretation algorithms foresee simplified exclusion rules for heavy ion peaks in their scoring or spectra preprocessing schemes [26]. Similarly, deconvolution of multiply charged peaks and deisotoping with procedures described in the literature [33–41] are possible only with very accurate data and resolved isotope clusters. The results are reliable only in the cases of sufficiently large peptide fragments where an isotope peak cluster of the higher charge state is confirmed by respective clusters at the lowest charge state or when the distances between peaks in a cluster accurately match the expected mass differences.

Sometimes, it might be rather advisable to refrain from automatically interpreting very noisy MS/MS spectra instead of generating interpretations that are not justified by the data. The task of unselecting noninterpretable spectra is related to but different from the question of cleaning spectra from noise. Xu *et al.* [42] and Bern *et al.* [43] propose empirical criteria for unselecting bad spectra; *i.e.*, spectra with only few significant peaks over a dense background. For these methods, the relatively high number of false positively unselected (*i.e.*, nevertheless interpretable) spectra remains a problem.

Previous work on raw protein MS/MS spectrum processing has not led to satisfying solutions and, therefore, many currently available MS/MS spectrum analysis packages largely ignore the presence of additional background signals. Most commercial spectrum interpretation software suites contain some noise reduction but the algorithms implemented are not publicly documented. At present, there is only one available program isolatedly dedicated to spectral cleaning, the MASCOT Distiller (see www.matrixscience.com), a commercial software package that optimizes peak location and intensities, given the ideal isotopic distribution of elements contained in peptides. However, the algorithms used in this software are not published and the correctness of peak removal/inclusion has not been evaluated in transparent large-scale tests. In addition, low computation speed and run-time stability issues may create problems in practical laboratory work.

It should be emphasized that, given the incomplete understanding of the chemical process of fragmentation, no automated procedure will match the performance of the experienced eye and the intuition of an MS specialist in the foreseeable future. Nevertheless, the number of mass spectra to be processed in proteomics laboratories is so large that there is no alternative to automated interpretation, may be, augmented by manual inspection of a few selected cases. In this article, we propose fast algorithms for background processing of peptide MS/MS spectra based on numerical spectral analysis and signal recognition approaches. They (i) detect multiply charged replicates and transform them into singly charged monoisotopic peaks, (ii) reduce isotope peak clusters to a single signal, (iii) remove high-frequency and periodic background noise. Finally, as a byproduct, we derive (iv) a spectral criterion for the determination of certain noninterpretable spectra with a very low false-positive rate. The approaches used are robust to mild inaccuracies in the data. We have implemented the algorithms in a software package called MS Cleaner, a program written in the C/C++ language, which can be obtained from the authors upon request. Tests show that noisy MS/MS spectra benefit from the treatment with the background removal procedure.

2 Materials and methods

2.1 Sample preparation

Purified antihuman Smc2 rabbit polyclonal antibody (200 μg) [44], crosslinked to Affi-Gel Protein A beads (100 μL bed-volume, BioRad), was used to immunoprecipitate the condensin complexes from 10 mg of clarified interphase HeLa cell extract. Following extensive washing, immunoprecipitated protein complexes were acid-eluted from the beads, and 10% of the total eluate was analyzed by SDS-PAGE and silver staining. After reduction and acetylation of cysteine residues using DTT and iodoacetamide, respectively, the condensin sample was proteolytically digested using Trypsin Gold (Promega), and the digestion stopped with tetrafluoroacetic acid.

2.2 MS

Tryptic peptides from condensin samples were separated by nano-HPLC [45] on an UltiMate HPLC system and PepMap C18 column (LC Packings, Amsterdam, The Netherlands), with a gradient of 5–75% ACN, in 0.1% formic acid [45, 46]. Eluting peptides were introduced by ESI into an LTQ linear IT mass spectrometer (Thermo Finnigan), where full MS and MS/MS spectra were recorded. In another experiment, a mixture of tryptic peptides from standard, commercially acquired BSA, yeast alcohol dehydrogenase (ADH) or human transferrin (TRF) was used for system optimization and testing. Each protein (100 fmol) was injected into a nano-HPLC device (LC Packings) and MS/MS spectra were acquired using a 3D IT mass spectrometer, model DecaXP (Thermo Finnigan).

2.3 File processing

The MS/MS output, in the form of an Xcalibur raw-file, was converted into dta files using BioWorks software (Thermo Electron, 53 944 spectra in the case of the condensin sample, 2679 for BSA, 2325 for ADH, and 2608 for TRF). The respective dta files were merged to generate a single mgf file (MASCOT generic format) using the merge.pl program (Matrix Science). This original mgf file was then processed using the MS Cleaner program, using the default internal parameters, generating two new mgf files with cleaned and bad spectra, respectively.

2.4 MS/MS data analysis

All three mgf files were used to perform MASCOT MS/MS Ion Searches (Matrix Science). In the case of BSA, ADH, and TRF, the nonredundant protein sequence database was used (as of 15 December, 2005). In the case of the condensin sample, the identification of post-translational phosphorylations was the original task. Therefore, the search was initially performed against a small curated protein database (146 sequences and 68 753 residues), which includes components of the condensin, cohesin, and kinetochore complexes, as well as some common contaminants and trypsin, in the case of the condensin sample. Additionally, we carried out searches against all human as well as against all proteins in the nonredundant database. It should be noted that the MASCOT score for recovering the original proteins tend to be the higher, the smaller the database due to reduced sequence background; thus, the search with the small database of 146 sequences is the more stringent condition compared with searches in the nonredundant database. The MASCOT search parameters were the same in all runs (enzyme: trypsin; fixed modifications: carbamidomethyl (Cys); variable modifications: oxidation (Met); peptide charges: 1+, 2+, and 3+; mass values: monoisotopic; protein mass: unrestricted; peptide mass tolerance: ±2 Da; fragment mass tolerance: ±0.8 Da; max. missed cleavages: 1). The MASCOT search results output html-file was formatted with standard scoring, a significance threshold of $p<0.05$, and an ion score cut-off for each peptide of 30.

3 Results and discussion

For a given raw (but centroided, peak-list transformed) peptide MS/MS spectrum, we propose the application of four separate independent procedures: (i) for detection of multiply charged peaks, (ii) for the removal of latent periodic noise including deisotoping, (iii) for the removal of high-frequency random noise, and (iv) for the detection of noninterpretable spectra. Each algorithm is applied on the same original MS/MS spectrum. First, we describe some illustrative cases to motivate the application of spectral criteria for background removal. After the following description of the four algorithms, we focus on results of testing the MS Cleaner in large-scale application tests.

3.1 Motivation for the application of spectral criteria for background removal

Albeit comprehending the exact mechanism of the genesis of background peaks would allow the construction of an algorithm for their removal, this knowledge is not available and more phenomenological approaches appear necessary. The analogy with electrical signal processing is one possibility; *i.e.*, the series of peaks in the mass spectrogram can be considered as a signal compounded with noise after transfer *via* an information channel, from which the original signal has to be recovered. At the associated website http://mendel.imp.univie.ac.at/mass-spectrometry/ANALYSIS/, we present some case studies with partially designed MS/MS spectra for illustration (see the series of Supplementary Figs. 1–12 at the associated website).

For example, isotope clusters are characterized by equidistant groups of peaks. It should be expected that such clusters are the source of latent periodicity in the signal that should be visible in the form of maxima in the frequency spectrum of the signal. This is indeed the case. From an original peptide MS/MS spectrum (Supplementary Fig. 1), we extracted all peaks relevant for interpretation by MASCOT (Supplementary Fig. 2). A third MS/MS spectrum was created with all MASCOT-interpreted peaks and complemented with artificial isotope clusters (Supplementary Fig. 5). The original MS/MS spectrum exhibits latent periodicity in their Fourier transforms (Supplementary Figs. 3, 4). There is no obvious periodic component in the spectrum with only interpretable peaks (Supplementary Figs. 6, 7). The periodic component reappears in the MS/MS spectrum consisting of MASCOT-interpreted peaks complemented with isotope clusters (Supplementary Figs. 8, 9). Thus, disappearance of isotope clusters correlates with dampening of the prominent periodic spectral component in the Fourier transform.

Similarly, noisy spectra are characterized by large numbers of low intensity peaks. We added artificial random noise to the MS/MS spectrum of MASCOT-interpreted peaks (Supplementary Fig. 10). Application of just a low-pass filter (suppression of the high-frequency part of the frequency spectrum of the signal) leads to the suppression, mainly, of artificially added noise peaks (Supplementary Figs. 11, 12).

Whereas these exemplary cases do not represent a proof of the efficiency for background removal with methods known in numerical signal processing, they show their potential in reasonably modified application settings. Further, it should be said that such procedures can identify true chemical or electronic background but do not aim to identify derivatives of *b*- or *y*-ions. The latter requires algorithmic analysis of chemical decay processes which is not the goal of this work. In the following, we describe the algorithms used in detail. For the convenience of the reader, Supplementary data to this text (dta and mgf files of exemplary mass spectra, additional Supplementary tables and figures) are available at the URL http://mendel.imp.univie.ac.at/mass-spectrometry/.

3.2 Deconvolution of multiply charged peaks

Although ionization techniques such as ESI have the advantage of shifting heavy ions into lower, detectable *m/z* ranges by generating multiply charged fragment ions [33], they can pollute the spectrum by causing replicates of otherwise identical ions at different charge states. In general, these multiply charged signals occur as isotope clusters. For the purpose of spectrum interpretation, peak replicates originating from different charge states have to be unified.

The relative spectral intensities of isotope-variant peaks in a cluster are determined by the natural isotope distributions of carbon, hydrogen, oxygen, nitrogen, and sulfur, the predominant chemical elements in peptide fragments. We use this *a priori* known form of the intensity pattern from multiply charged replicates for searching its reoccurrence in the measured spectrum by correlational analysis. Our algorithm is quite robust relative to inaccuracies in the experimental resolution of isotope clusters due to two artifices in processing the mass spectrum: (i) the removal of small peaks very close to major intensities and (ii) the procedure of interpolated peak densification in the mass range of comparison with the predefined pattern.

The algorithm includes several steps (see also Fig. 1). Prior to spectrum analysis, the general form (the etalon) of isotope cluster patterns is precomputed for double- and triple-charged fragments. The intensity patterns in isotope clusters become complicated with large fragment masses but still can be exactly calculated [47–51]. Given the large number of potential peptide fragment sizes and sequence possibilities, the computational time for taking into account the exact isotopic patterns is too high for a background analysis program. We rely on Wehofsky's polymomial approximation [38, 39], a computational shortcut for the target signal where the relative intensity of the *n*th isotope variant peak (in a pattern of $N \leq 7$ peaks; $k = 6$, the order of expansion) is

$$I(n, M) = A(n) + \sum_{j=1}^{k} B_{\mathrm{j}}(n) M^{\mathrm{j}} \tag{1}$$

where M is the mass corresponding to the first, monoisotopic peak in the cluster ($n = 1$). The relative intensity of this peak is assumed to be 1. $A(n)$ and $B_j(n)$ are fitting parameters taken from Wehofsky's work [38, 39]. Depending on the charge state z, the *m/z* distance between peaks in the pattern is $1/z$ Da and the pattern length is $(N-1)/\mathrm{z}$ Da (Fig. 1E). Finally, the pattern of the etalon is complemented, *i.e.*, densified with totally $20(N-1)/z - N + 1$ additional peaks (with a 0.05 Da *m/z* step) where their intensity is linearly interpolated from the two surrounding pattern-defining peaks with masses $M + (n-1)/\mathrm{z}$ and $M + n/z$ (Fig. 1F). The intensity patterns have been tabulated with an accuracy of 100 Da.

Every peak of the experimental spectrum is considered as a potential starting point of an isotope cluster pattern (Figs. 1A and B). The mass window with the length of the target signal following each peak is densified with linearly interpolated additional peaks (at 0.05 Da steps) up to the last experimental peak in the window (Fig. 1D). The addition of further peaks (essentially a transformation to a semianalog signal) compensates for possible small inaccuracies in resolving the position of isotope-variant peaks by the instrument's software. The correlation coefficient of the observed intensities with those from the precomputed pattern is calculated (Figs. 1G and H). Very high correlation (above 0.95 or even 0.99 in the case of very accurate data) indicates reoccurrence of the target signal in the pattern. Detected multiply charged peak clusters are removed and converted into a singly charged monoisotopic peak that is added to the spectrum. In rare cases when the same piece of spectrum is interpreted both as triply and as doubly charged clusters with high correlation coefficients, the charge state with the higher coeffi-

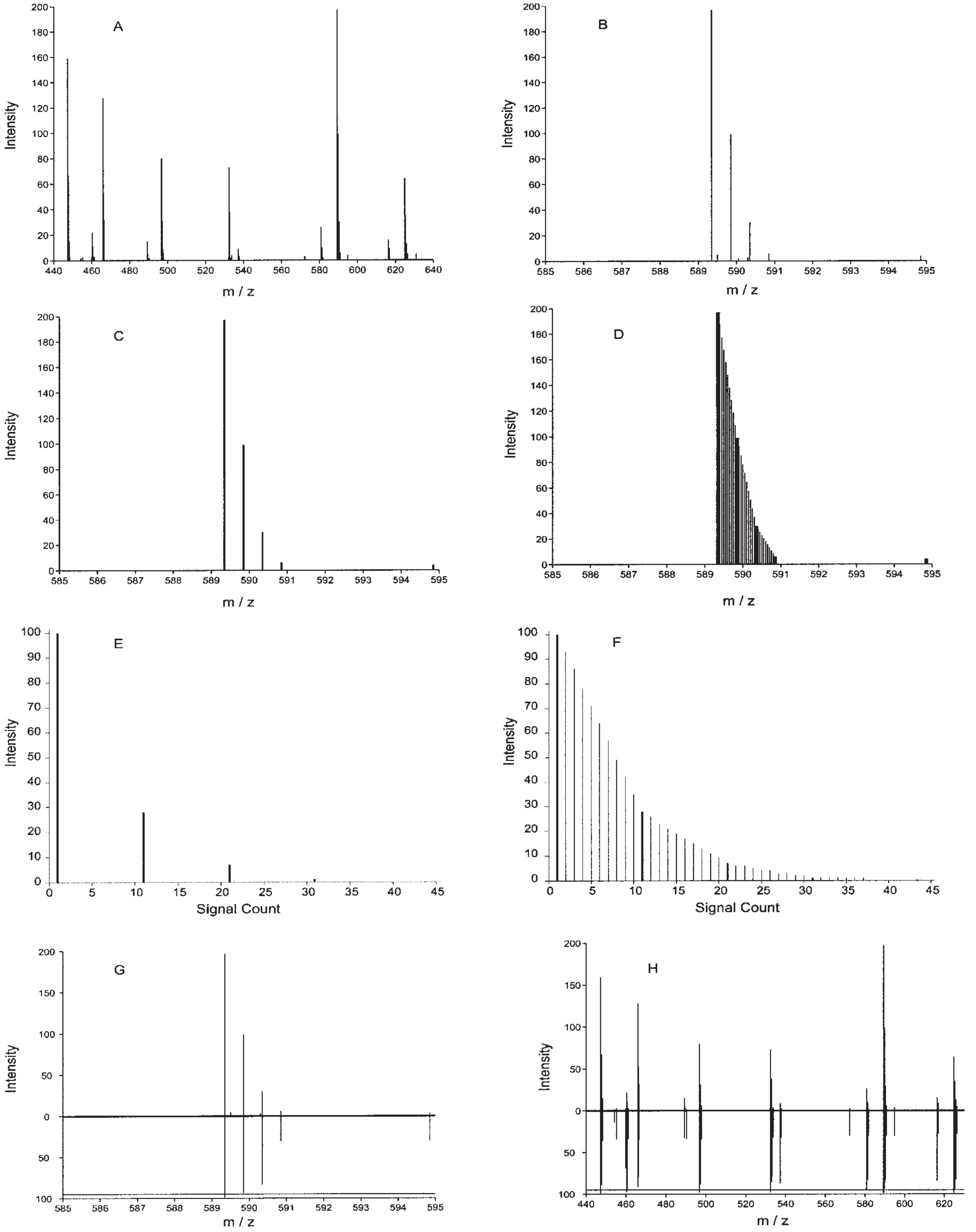


Figure 1. Determination of multiply charged replicates with correlation analysis. This series of diagrams illustrates the process of removing multiply charged replicates. The abscissa represents the *m/z* (the signal count in 0.1 Da/charge unit in E and F); the ordinate axis shows peak intensity in relative units. To the order of diagrams: (A), (B) are in the first row, (C), (D) in the second, *etc.* (A) Piece of raw MS/MS spectrum. (B) Peak cluster from raw spectrum at greater magnification. (C) The same peak cluster after removal of small peaks. (D) The same peak cluster after densification. (E) Precomputed pattern of isotope peak cluster. Here and in (F), only the relative abscissa value is important (with an undefined additive constant). (F) The same pattern after densification. (G) Peak cluster from raw spectrum together with coefficients of correlation with the precomputed pattern (in the lower part of the graph, multiplied by 100%; the horizontal line corresponding to 95% is shown). (H) The same but for the whole piece of raw spectrum.

cient is taken. It should be noted that our algorithm does not look after correlations between the occurrences of singly and multiply charged ions from the same chemical species.

This procedure works adequately as long as no very low-intensity peaks close to major intensities of an isotope cluster interfere (distance below ~0.25 Da, a user-defined measure below the machine accuracy). These peaks are typically artifacts that can arise from random noise or from the transformation of the continuous MS/MS spectrum into the centroid form as a discrete signal. Prior to spectrum densification, the small interfering peaks between the main isotope cluster peaks have to be merged with the closest main peak in the cluster; *i.e.*, this is essentially a procedure of reversal of small peak creation. For the peak-merging algorithm, a weight directed graph $G(V, E)$ is constructed. The set of vertices (V) is the set of all m/z values in the window. A directed edge $e_{I,j} \in E$ is added between two vertices v_i, $v_j \in V$ if the distance d between peaks v_i, v_j is less than the user-defined accuracy value. The direction of the edge is defined to be from v_i to v_j if intensity (v_i) < intensity (v_j). The weight w_i of an edge $e_{i,j}$ is defined as the distance between two vertices v_i and v_j (in 0.01 Da units). If a vertex v_i giving origin to the edge $e_{i,j}$ is actively removed from the graph (and its intensity is added to the vertex v_j), then edges to other vertices can also vanish. *Via* systematic enumeration (for example with topological ordering), an edge-free subgraph can be computed without large computational cost that fulfills the condition that the sum of weights of actively removed edges is minimal.

It should be noted that, with the procedure for finding multiply charged isotope clusters that uses the criterion of high correlation with an etalon, not all such clusters will probably be detected. Most importantly, our algorithm relies on the resolution of isotope clusters for multiply charged replicates. For the typical mode of an MS/MS device without Fourier-transformation capability, the fast scanning precludes the detection of isotope clusters in many instances and the respective multiply charged ions will not be detected. It also possible that sections of MS/MS spectra with very dense noise let our algorithm believe the existence of a cluster; thus, the removal of the original noise cluster will lead to the creation of a single noise peak with higher m/z. It is also not excluded that, in some other instances, our algorithm might generate a few false-positive predictions. Especially, problems will appear in the following cases: (i) aggressive baseline suppression used in some instruments will affect the relative intensities of peaks of an isotope cluster and might reduce the correlation with the predefined etalon. This problem can be addressed during spectrum recording. (ii) True interfering peaks overlaying an isotope cluster will either result in low correlation and prevent the detection of the multiply charged cluster (if they are of high intensity and lead to a low correlation coefficient) or disappear in the peak-merging algorithm (if they are of low intensity). (iii) Some interpretable ions are very close (*e.g.*, y-NH_3 and y-H_2O differ only by 1 Da) and might create the false impression of an isotope cluster if their relative intensities are commensurate with the etalon. The results of real-life applications (see below) show that the cases listed are rare in real applications.

3.3 Removal of latent periodic noise including deisotoping of the spectrum

Correlation of the measured MS/MS spectrum with pre-calculated isotopic intensity distributions is efficient only for multiply charged peak clusters since the probability of finding additional, unrelated peaks in the spectrum with distance of 1 Da is high. Therefore, correlation analysis with predefined patterns is not really useful for deisotoping. But if we treat an MS/MS spectrum as a set of signals in the time domain where the m/z axis is the analog of time and the intensity of each peak in the MS/MS spectrum as the intensity of a signal at a certain time, we can consider the single-charged peak signals as a periodical function (with periodicity of ~1 Da for singly charged peaks). This periodical function in the time domain results in a power spectrum in the frequency domain where the reoccurring elements can be more easily recognized.

Besides isotope variants, there can be other sources of spectral contamination with latent periodicity, for example from the electronic detection system or from the accompanying chemical polymer contaminants such as silanes, *etc.* Reoccurring signals at quasi-constant mass shifts can be seen in the frequency domain, *i.e.*, as characteristic reoccurrences of high amplitudes at multiples of a base frequency f_B in the Fourier transform of the tandem mass spectrum. Yet another Fourier transformation applied at the frequency domain level can be used to determine this base frequency f_B. As we have seen above, suppression of periodically reoccurring intensity maxima in the power spectrum can effectively remove latent periodical noise including minor isotope variant peaks (Fig. 2). When writing this manuscript, we noticed that periodicity analysis has been previously proposed for the detection of chemical background in MS fingerprints of small organic or inorganic compounds [52].

Converting to the frequency domain, the discrete Fourier transform Y of the MS/MS spectrum (S) is found by taking the N-point fast Fourier transform $Y - \mathrm{FFT}(S,N)$. The value N is calculated as $N = 2^n + 1$, where n is the smallest integer larger than $\log_2[(x_{\max} - x_{\min})/0.05]$. The values $x_{\max}$ and $x_{\min}$ are the largest and the smallest m/z values in the spectrum, respectively. The power spectrum, a measurement of the power at various frequencies, is $PS = Y \cdot Y^*/N$ (see example in Fig. 2A, called PS-graph below). Typically, the power spectrum of a good MS/MS spectrum is quasi-periodic. The length of this period (the base frequency f_B) is determined with another Fourier-transformation, where we consider the power spectrum as a signal in the time domain (Fig. 2B, called PSPS graph below). In order to remove the reoccurring elements from the power spectrum, a multiband reject filter has to be introduced for each MS/MS spectrum. The filter is created by the Yulewalk method of auto

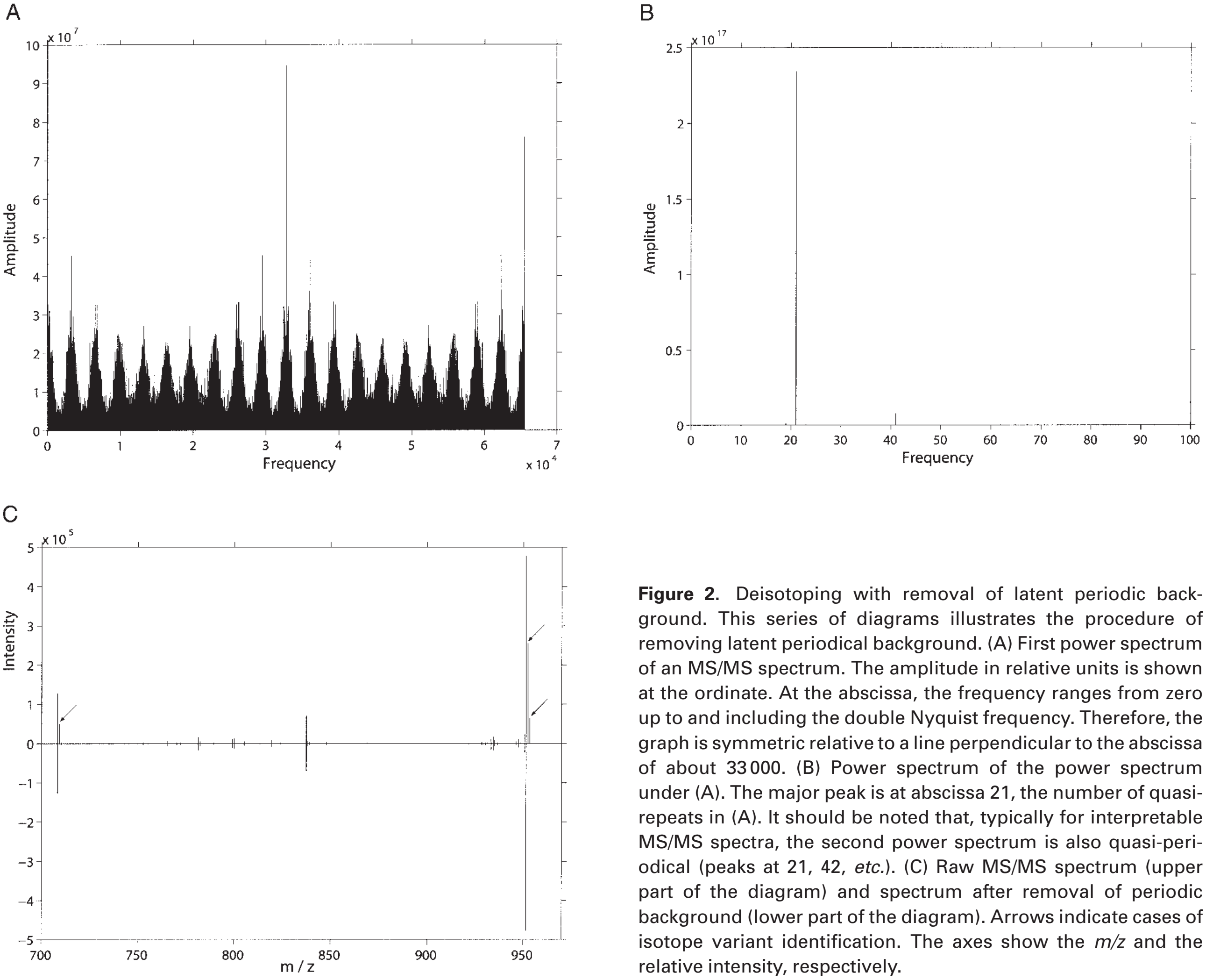


Figure 2. Deisotoping with removal of latent periodic background. This series of diagrams illustrates the procedure of removing latent periodical background. (A) First power spectrum of an MS/MS spectrum. The amplitude in relative units is shown at the ordinate. At the abscissa, the frequency ranges from zero up to and including the double Nyquist frequency. Therefore, the graph is symmetric relative to a line perpendicular to the abscissa of about 33 000. (B) Power spectrum of the power spectrum under (A). The major peak is at abscissa 21, the number of quasi-repeats in (A). It should be noted that, typically for interpretable MS/MS spectra, the second power spectrum is also quasi-periodical (peaks at 21, 42, *etc.*). (C) Raw MS/MS spectrum (upper part of the diagram) and spectrum after removal of periodic background (lower part of the diagram). Arrows indicate cases of isotope variant identification. The axes show the *m/z* and the relative intensity, respectively.

regressive moving average (ARMA) spectral estimation [53]. Yulewalk designs recursive infinite impulse response (IIR) digital filters using a least-squares fit to a specified frequency response. Frequencies required by the Yulewalk method are calculated by applying a median filter to the power spectrum (over 300–500 discrete data points) and by computing a second power spectrum (PSPS-graph) in order to obtain the most prominent frequency of the first power spectrum (PS-graph). The created IIR filter is used to filter the MS/MS spectrum in the time domain [54]. After filtering, the recovered MS/MS spectrum might contain some signals with negative intensity or some new signals with positive intensity. Additionally, some signals occurring with positive intensities both in the original raw spectrum and the recovered spectrum have lost considerable intensity in the latter (threshold of 95%; this number should be higher for very clean and regular spectra). All three types of signals are corrected to zero in a final step. Examination of exemplary spectra has shown that suppression of latent periodicities in the MS/MS spectrum effectively also removes low-intensity peaks originating from higher mass isotopes in isotope clusters (Fig. 2C).

In some cases, PS-graphs of dta files display several, overlaying modes of periodicities. The respective PSPS-graphs have several maxima with similar intensities. If the numerically largest maximum is at very low base frequencies f_B (*e.g.*, there are only a few maxima in the PS-graph), the application of the periodical multiband filter with this f_B can lead to severe damage of the MS/MS spectrum. To avoid this problem, we routinely set intensities in the PSPS-graph to zero for low frequencies (up to and including the threshold f_{BT} = 14 abscissa units).

3.4 Removal of high-frequency random noise

Noisy MS/MS spectra suffer from many superfluous peaks densely distributed over the whole *m/z* range. Assuming that the random noise in an MS/MS spectrum exists as signals of high frequency of occurrence, a Butterworth IIR low-pass filter [55] is applied to the spectrum in the time domain. The normalized stop frequency of the filter is in the range from 0.5 to 0.9 (the best result was obtained with stop frequency 0.8). An empirical threshold of 99.99% is applied to remove all signals, which have lost intensity above this threshold, from the raw spectrum.

3.5 Recognition of noninterpretable spectra

Our experience of power spectrum analysis of MS/MS spectra also indicates a criterion that can be used for the identification of bad spectra that are not useful for further study. We observed two types of irregularities that coincide with hard-to-interpret protein MS/MS spectra: (i) The first power spectrum can exhibit very low amplitudes for low frequencies. (ii) Finding the most prominent frequency in the second power spectrum can be ambiguous (several similarly high peaks). In both cases, our procedures for background removal cannot be straightforwardly applied and, therefore, each mass spectrum is subjected to a routine check during analysis.

With the base frequency derived from the second power spectrum, it is possible to compute the position of expected maxima and minima in the first power spectrum (Figs. 3A and B). We determine whether the real minima and maxima within periods are, on average, closer to the expected positions or closer to the positions with the shift of half a period. If the spectrum is shifted (*i.e.*, if the sum of distances of real maxima and minima from their expected positions is larger than the positions with a shift of half a period) away from the expected position of minima/maxima, the procedure for deisotoping is halted.

Unfortunately, large shifts in the power spectrum away from expected minima/maxima often indicate bad spectra. For making an appropriate decision, the periodicity of the spectrum is also tested with a similar elementary criterion as the shift. We rely on the coefficient of dispersion (C_d) of peak distances in the power spectrum, calculated as the ratio of the SD of peak distances (s) to the mean value of peak distances ($\overline{X}$).

$$C_d = \frac{s}{\overline{X}} \qquad (2)$$

A C_d close to zero indicates good coincidence of distances between maxima (and, respectively, minima) of consecutive periods with the expected distance (equal to the period length). Large values of C_d signal distorted periodicity in the power spectrum and a periodicity model appears not applicable. Such spectra are returned to further processing without removal of the latent periodic noise.

The case of quasi-periodic but shifted spectra is more complicated. In such a situation, if the coefficient of dispersion is not larger than 3.3 (an empirically derived threshold), the algorithm predicts that the respective MS/MS spectra cannot be reliably analyzed with the interpretation software [30]. As will be shown below, spectra flagged with this criter-

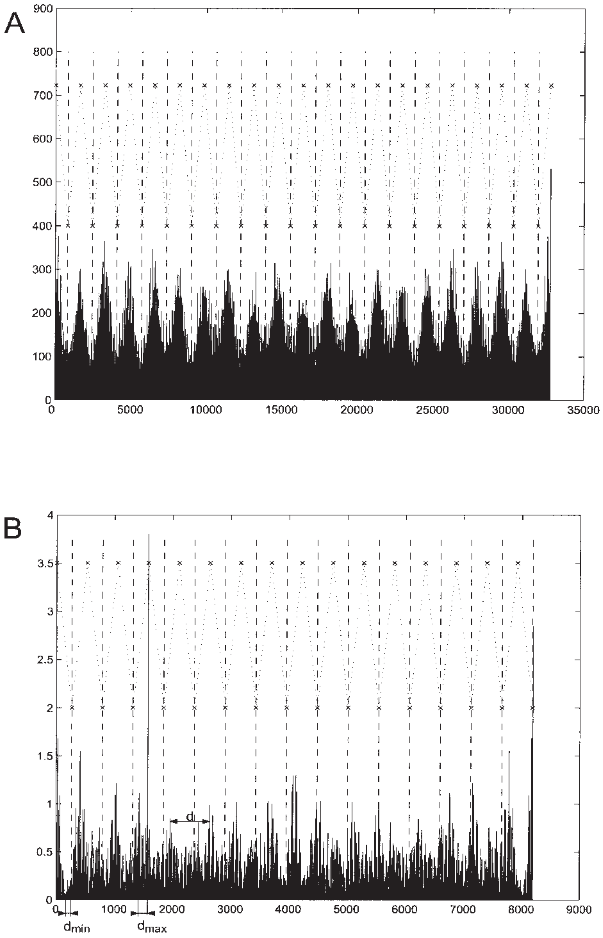


Figure 3. Determination of quality of the first power spectrum of an MS/MS spectrum. (A) We observed that the power spectrum derived with Fourier transformation from an easily interpretable MS/MS spectrum is typically quasi-periodic without phase shift as in this example. The original dta file and the view of the original spectrum in jpg- and tif-format are available at website http://mendel.imp.univie.ac.at/mass-spectrometry/. We show the power spectrum from zero to the doubled Nyquist frequency. Having the number of periods determined from the second power spectrum, the expected positions of minima and maxima in the first power spectrum can be calculated. With dashed lines, the abscissa positions of expected minima of intensity are indicated. Both expected minima and maxima positions are emphasized at the respective abscissa values with markers (crosses), which are interconnected *via* a dotted line for visual guidance. Obviously, the true minima and maxima of the power spectrum coincide well with their expected positions. (B) Example of a difficult to interpret spectrum (see the website for dta file and spectrum view). The true maxima and minima of the respective periods are irregularly shifted with respect to the expected positions. The expression d_{min} denotes the distance between the true and the expected position of a minimum within a period, d_{max} measures the deviation for the maximum (a thin continuous line denotes the expected position of the respective maximum). The peak distance d is the difference of abscissa positions between maxima of consecutive periods (similarly for the minima). The SD s and the mean value X are calculated from the set of all peak distances.

ion are indeed not well interpretable even with database search-based software (*i.e.*, no protein hits are found or only hits with very low reliability).

In rare cases, the suppression of very low frequencies in the PSPS-graph leads to incorrect base frequency determination (to f_B that is too high) and, consequently, to apparently shifted spectra. These few spectra marked as non-interpretable are false-positively rejected and represent part of the price for automatically cleaning large-scale MS/MS measurements from background with spectral methods as described here.

3.6 General considerations for testing procedures for background removal in tandem mass spectra

In the ideal world, background removal algorithms would be parametrized and tested against a large library of MS/MS spectra where the different types of all noise (*e.g.*, multiply charged peaks, isotope clusters, random noise, *etc.*) are explicitly annotated in electronically readable form and the rates of true- and false-positive detection of various noise types can be directly computed. Unfortunately, such a library was not available to us during this work and its creation is beyond the scope of our effort. We validated our background removal algorithm implicitly. The automated interpretation of MS/MS spectra with MASCOT has become a virtual standard in proteomics laboratories; therefore, we compared the MASCOT-generated interpretations both for the original MS/MS spectra and the spectrum versions after the application of our background removal procedure. Discrepancies between both interpretations can be automatically detected in large-scale tests of real datasets and summarized by computer programs. We used this approach also for parametrizing the MS Cleaner. The parameters described above have been selected to achieve a minimum of cases of accidental removal of peaks that are relevant for interpretation by MASCOT in large-scale tests. Finally, we tested the MASCOT Distiller in the same setting.

We wish to emphasize, to the best of our knowledge, that possibly existing internal procedures for background masking in MASCOT have not been described in the public literature. If there are any, they are the same in all test applications in this work and the results are independent of them.

3.7 Results of background removal in MS/MS spectra obtained with 100 fmol BSA, ADH, and TRF

To test the MS Cleaner in large-scale practical applications, we used MS/MS spectra from protein samples with known composition. In our setup, such spectra are regularly produced for the purpose of quality control of MS instrumentation with low concentrations (100 fmol) of BSA, ADH, or TRF. Original and cleaned spectra as well as Supplementary tables that show changes of scores of leading peptide hits are available at the associated website. The results of applying the background removal procedure are summarized in Table 1. First, it is evident that protein hits are found from the cleaned MS/MS spectra with considerably increased scores. This is evident for the total protein score (between 10 and 15%, see Table 1A). Scores improve for the majority of all leading peptide hits (about 70%, see Table 1B), a decrease is observed for about 10% of cases but did not affect the interpretation except of one case (see below). In general, the likelihood of retrieving the sample protein and the sequence coverage improve (see Table 1A). This conclusion is in line with the logics of MS/MS spectra interpretation schemes such as MASCOT: The MS Cleaner-based background

Table 1. Influence of background removal on the recovery of BSA, ADH, and TRF in MS/MS spectra of 100 fmol test samples

(A)

Search	dta files	Score	Match	Cov. (%)
BSA				
Raw spectra	2679	1844	65	51
Cleaned spectra	2484	2094	70	56
Bad spectra	195	n/a	n/a	n/a
Yeast ADH				
Raw spectra	2325	536	24	29
Cleaned spectra	2060	594	25	29
Bad spectra	265	n/a	n/a	n/a
Human TRF				
Raw spectra	2608	1643	61	41
Cleaned spectra	2442	1846	65	44
Bad spectra	166	64	1	2

(B)

	BSA	ADH	TRF
Total peptide hits	70	25	68
Scores increased	47	18	48
Scores unchanged	5	4	3
Scores decreased	13	2	6
Hits only after cleaning	5	1	8
Hits lost after cleaning	0	0	3

The MS/MS spectra were interpreted with MASCOT directly ("raw spectra") and after processing with the background removal procedure ("cleaned spectra") described in this article. (A) The "score" is the MASCOT score from all successful searches; "match" is the number of searches that recover the peptides from the protein used. "cov (%)" reports the sequence coverage. The line "bad spectra" reports the number of files that are considered not "interpretable" by the criterion described in the text (n/a – not applicable). Only in one case could MASCOT recognize a peptide from the original protein in a bad spectrum that is visually also of low quality. (B) Changes of scores of leading peptides in MASCOT searches as a result of background cleaning (summary digest of Supplementary tables at the website).

removal decreases the number of peaks considerably. Therefore, the number of alternative (including false-positively hit) protein sequences that might fit a given spectrum reduces and the scores of the top hits against the alternatives naturally improve.

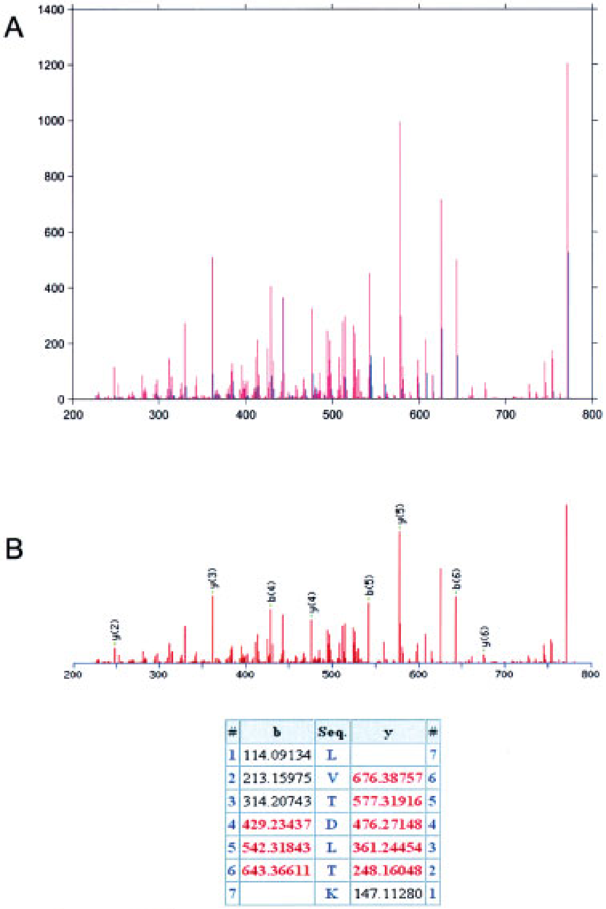
MS/MS spectra considered noninterpretable by our procedure are indeed bad spectra. In only one out of 626 cases was the original protein recovered by MASCOT. Here, MASCOT assigned a score of 64 (see Table 1 and also data and figures at http://mendel.imp.ac.at/mass-spectrometry/false positive-partA.html). Visual inspection of the spectrum revealed almost no significant peaks above background. We found that this single artifact of rejection by MS Cleaner is a result of the suppression of low frequencies in the PSPS-graph and would disappear with a slightly reduced threshold f_{BT} = 12. In contrast, there are a considerable number of spectra (about 10%) that become interpretable for MASCOT only after background removal with our procedures (five for BSA, one for ADH, and eight for TRF, see Table 1B). An example is shown in Fig. 4. Out of the 373 peaks in the spectrum, 83 are recognized as background and are removed. As a result, MASCOT was no longer confused and was able to assign a full *y*-series and many *b*-ions.

Although all procedures described in this work are essential for various aspects of background reduction, they contribute differently from the quantitative point of view. As can be seen from the data in Table 2, the spectral-analytic criteria (removal of latent periodic and high-frequency noise) are most efficient in reducing the background since their share among the removed peaks is above 90%. In the BSA, ADH, and TRF applications, about 15% of all peaks in the original spectra get removed by our program and the file storage requirement is reduced by the same amount. We tested the computational performance of MS Cleaner on a stand-alone PC (under the Windows XP operating system). For the BSA case, 2679 dta files were cleaned in 4:52 min (0.11 s *per* spectrum). The MASCOT time on the same machine reduced from 64 min (for the untreated data) to 57 min (cleaned files). The respective numbers for ADH (2325 files) and TRF (2608 files) are 5:36 (0.14 s *per* file), 75, 64 and 4:15 (0.10 s *per* file), 58, 50 (all values in minutes). Thus, savings of computational costs are considerable under the condition of increased reliability of spectrum interpretation.

Table 2. Contribution of different procedures in the background removal to the experiment for recovery of BSA, ADH, and TRF in MS/MS spectra of 100 fmol test samples

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
BSA	4293	20 749	1248	32 570	326 627	50 523 (58 860)	15.47
ADH	1041	12 353	1402	18 208	215 499	27 940 (33 004)	12.97
TRF	3123	19 297	1483	28 779	294 546	44 710 (52 682)	15.18

Four sources contribute to the peak removal: (i) At the start, all peaks with a spacing smaller than the user-defined accuracy are merged (default: 0.25 Da). (ii) Number of peaks removed by the periodic noise detection procedure (including deisotoping). (iii) Number of peaks identified by the deconvolution of multiply charged replicates. (iv) Number of peaks found by the routine for high-frequency noise removal. It can be seen that the spectral-analytic criteria are most efficient in background reduction. In the last three columns, we list the total number of peaks in the original spectra. (v) The number of peaks removed and the percentage from the total number of peaks. Some procedures identify the same peaks as noise. To assess this effect, we present the arithmetic sum of the numbers from all noise reduction procedures (1–4) in parentheses (in the penultimate column). Apparently, 10–20% of all identified background peaks is found by multiple criteria.



#	b	Seq.	y	#
1	114.09134	L		7
2	213.15975	V	676.38757	6
3	314.20743	T	577.31916	5
4	429.23437	D	476.27148	4
5	542.31843	L	361.24454	3
6	643.36611	T	248.16048	2
7		K	147.11280	1

Figure 4. Example of a spectrum that was only interpretable after background removal. (A) Original MS/MS spectrum of 100 fmol BSA (abscissa: *m/z* in Da, ordinate: relative intensity; totally 373 peaks). Background peaks that have been removed by MS Cleaner are shown in blue (83), other peaks are shown in red (290). (B) MASCOT interpretation of the cleaned spectrum (as peptide sequence LVTDLTK). We show the spectrum with assignment of *b*- and *y*-ions and the table representing the sequence ladder. Both the original dta file as well as the cleaned version are available at the website.

3.8 Detailed analysis of MS Cleaner's removal of multiply charged peaks in the dta files of the BSA set

We attempted to check whether the multiply charged peaks assigned by MASCOT are detected by our program MS cleaner. After having manually analyzed the whole BSA dataset, we found only two peaks interpreted as doubly charged by MASCOT that had also a remnant isotope cluster (in the dta file 369.369.2, see Supplementary data at http://mendel.imp.ac.at/mass-spectrometry/beforeafterBSA.htm). For this spectrum, MS Cleaner revealed seven doubly charged clusters. Two of them (at $m/z = 315.70$ and 320.30) include the two doubly charged peaks found by MASCOT. The other five are composed of noise peaks. It should be noted that spectral procedures (as a rule, the algorithm for high-frequency noise removal) mark many low intensity peak clusters (comparable with the five latter ones) as noise, too. As discussed above, MS/MS measurement accuracy and scanning speed on many instruments prevent the detection of isotope clusters in many cases. We think that the algorithm for detecting multiply charged clusters will work the better, the more accurate the spectra are recorded (as in the new generation of Fourier-transformation instruments) and the more complete isotope clusters are represented in the data.

3.9 Application of the background removal to the condensin dataset

It should be noted that, in the latter example, low concentrations of proteins are intentionally applied to achieve limiting cases of mass spectra. The analysis of the condensin complex mass spectra is a more biologically relevant application. For this purpose, we decided to purify and analyze condensin complexes from cultured human HeLa cells. Human cells contain two distinct condensin complexes, called condensin I and condensin II, which bind chromosomes specifically in mitosis and contribute to their condensation and structural integrity [44, 56–58]. Both complexes are hetero-oligomers composed of five subunits. Two ATPase subunits of the structural maintenance of chromosome (SMC) family, called Smc2 and Smc4, are shared between condensin I and condensin II. In addition, each complex contains a set of distinct non-SMC subunits, called kleisin-γ [57], CAP-G, and CAP-D2 in the case of condensin I, and kleisin-β [57], CAP-G2, and CAP-D3 in the case of condensin II. We immunopurified both complexes simultaneously using antibodies to their common Smc2 subunit and analyzed the resulting sample both by SDS-PAGE and silver staining (Fig. 5) and by in-solution digest followed by LC-MS/MS. Silver staining revealed bands that correspond to Smc2, Smc4, and to all six non-SMC subunits that are present in condensin I and condensin II. The MS/MS spectra were processed using the MS Cleaner. All three datasets, the original, the cleaned, and the bad spectra, were used to perform a MASCOT MS/MS Ions Searches against a small and curated protein database as well as against the nonredundant protein database (all proteins and all human proteins).

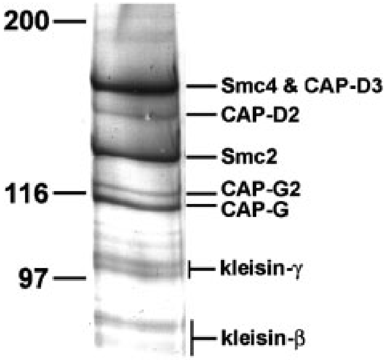


Figure 5. Quality of the condensin complex purification. SDS-PAGE silver-stained gel of the purified human condensin complexes. The bands were previously identified by Yeong *et al.* [58]. This result confirms the purity of the complex obtained in the experiment.

A summary of the MASCOT search results for this experiment is shown in Table 3. First, we consider the case of searching the small database consisting of 146 sequences. Each of the eight condensin subunits showed an increase in MASCOT score (mean increase of 8.2%), and number of peptide matches (mean increase of 4.8%) following the cleaning procedure. As a rule, the percentage of sequence coverage obtained was the same or higher for searches using the cleaned spectra than for those using the original spectra. The only exception from this list was kleisin-β, which showed a 2% reduction in the sequence coverage after cleaning. Closer inspection revealed that this reduction was due to a single peptide match generated by a single MS/MS spectrum that visually appears of low quality (see data and figures at http://mendel.imp.ac.at/mass-spectrometry/false positive-partB.html). This MS/MS spectrum has very few significant peaks above the baseline, and is classified as "noninterpretable" by the MS Cleaner. We found out that this artifact is a result of low frequency suppression in the PSPS-graph and could be avoided with a slightly reduced threshold $f_{BT} = 12$. However, the MASCOT program generated a match between this spectrum and the peptide QGEVLASR (within kleisin-β). It was classified as a hit with a MASCOT score of 45, although the majority of the peaks that contributed to the assignment are very small and the most significant peaks do not contribute to this interpretation. Thus in this case, the removal of just a single nonreliable peptide during the cleaning process resulted in a small reduction in sequence coverage, although the MASCOT score for the protein as a whole was increased as a result of background removal.

It should be noted that all cases of peptide detection by MASCOT in spectra classified as noninterpretable by MS Cleaner (14 out 1318 dta files) lead to low scores with marginal sequence coverage by MASCOT when there are very

Table 3. Influence of background removal on the recovery of condensin subunits in MS/MS data

Protein	Raw			Cleaned			Increment			Bad		
	Score	Match	Cov. (%)	Score	Match	Cov. (%)	Score %	Match %	Cov. (%)	Score %	Match %	Cov. (%)
(A)												
Smc4	3768	329	57	4125	341	64	9.5	3.6	12.3	98	2	1
CAP-D2	3637	182	65	4038	195	69	11.0	7.1	6.2	33	1	1
Smc2	2957	219	55	3239	231	57	9.5	5.5	3.6	201	4	4
CAP-D3	2627	104	42	2772	108	43	5.5	3.8	2.4	n/a	n/a	n/a
CAP-G	2554	106	55	2678	110	55	4.9	3.8	0.0	200	3	3
CAP-G2	1992	82	44	2255	86	50	13.2	4.9	13.6	154	3	6
Kleisin-γ	1843	78	61	1979	84	63	7.4	7.7	3.3	n/a	n/a	n/a
Kleisin-β	1245	45	69	1306	46	67	4.9	2.2	−2.9	45	1	1
(B)												
Smc4	4829	416	62	5188	424	64	7.4	1.9	3.2			
CAP-D2	4411	229	66	4818	241	68	9.2	5.2	3.0			
Smc2	4054	300	61	4436	312	64	9.4	4.0	3.8			
CAP-D3	3134	118	43	3329	125	45	6.2	5.9	3.9			
CAP-G	2850	117	51	3014	120	52	5.8	2.6	1.5			
CAP-G2	2553	106	50	2760	110	51	8.1	3.8	1.8			
Kleisin-γ	2158	94	61	2300	96	61	6.6	2.1	0.7			
Kleisin-β	1446	48	65	1573	49	65	8.8	2.1	−0.8			
(C)												
Smc4	4502	321	59.860	4865	328	62	8.1	2.2	3.4			
CAP-D2	4176	192	64.954	4590	204	67	9.9	6.3	2.5			
Smc2	3747	246	59.733	4137	255	62	10.4	3.7	3.4			
CAP-D3	2862	100	53.695	3060	104	54	6.9	4.0	1.5			
CAP-G	2453	76	24.860	2627	81	25	7.1	6.6	2.5			
CAP-G2	2239	163	39.463	2500	165	41	11.7	1.2	3.4			
Kleisin-γ	1892	146	34.005	2167	149	36	14.5	2.1	5.9			
Kleisin-β	1043	31	45.785	1104	31	46	5.9	0.0	1.4			

The MS/MS spectra were interpreted with MASCOT directly (″raw spectra″ from 53 944 dta files, total size 460 MB) or after processing with the background removal procedure (″cleaned spectra″ from 52 626 dta files, total size 284 MB) described in this article. The "score" is the MASCOT score from successful searches; "match" is the number of searches that recover the peptides from the protein used. "cov (%)" reports the sequence coverage. We present the results of three searches: (A) against the database of 146 proteins, (B) against the human proteins in the nonredundant database and (C) against all proteins in the nonredundant database.
The columns "bad spectra" report cases of files (among 1318 dta files, total size 7 MB) that are considered not interpretable by the criterion described in the text (n/a – not applicable) where MASCOT could, nevertheless, recognize the original protein in a database of 146 proteins but with a low score.
Cov., Coverage.

few significant peaks above an apparent noise. Changing to MASCOT searches against larger databases leads, as a trend, to even more dramatic improvements of scores and sequence matches (Table 3). In the case of the full nonredundant protein sequence database, there is even an increase of sequence coverage for kleisin-β after background removal with our procedure because MASCOT was unable to assign a match to several noisy spectra against the extensive sequence background of the largest database.

In a practical setup, the computational efficiency is also important. MS Cleaner processed the 53 944 spectra from the condensin experiment in less than 4 h on a single standard PC; *i.e.*, in 0.25 s *per* file. However, the application of our background removal procedure reduces the pure MASCOT computing time for the body of 53 944 dta files in the condensin complex case by about 25%, even in the case of a small database of 146 sequences; the size of the cleaned mgf file is decreased by 39%. Therefore, application of the MS Cleaner significantly reduces computing time and storage.

3.10 Comparison between MASCOT Distiller and MS Cleaner

There are no tools for background removal in peptide MS/MS spectra readily available in the public domain. Among commercial programs, only MASCOT Distiller is explicitly devoted to this task. From the scientific point of view, a correct comparison of MASCOT Distiller with our tool is not

Table 4. Comparison between MASCOT Distiller and MS Cleaner

Protein	Raw		MASCOT Distiller			MS Cleaner		
	Score	Match	Score	Match	Time	Score	Match	Time
BSA	1844	65	1565	44	7:40	2094	70	3:58
ADH	36	24	612	15	6:48	594	25	2:34
TRF	1643	61	1532	38	5:48	1846	65	3:23

The MS/MS spectra for BSA, ADH, and TRF were interpreted with MASCOT directly ("raw spectra") and after processing with MASCOT Distiller and with the background removal procedure described in this article ("MS Cleaner"). The "score" is the MASCOT score from all successful searches; "match" is the number of searches that recover the peptides from the protein used. The processing time is presented in min:sec. The performance of the procedure described in this article is superior compared with that of MASCOT Distiller with respect to score, and number of correct sequence matches. In addition, it consumes only 50% time on an identical computer with the same operating system environment.

possible, because the algorithms used in commercial MASCOT Distiller have not been properly described in public and the reasons for differential performance of the two programs cannot be causally interpreted. In Table 4, we present the results of application of the two programs on the BSA-, ADH-, and TRF-datasets. Whereas MASCOT Distiller produces mixed results with respect to the score and sequence matches (one increase and two decreases), our program increases the score and the number of matches in all three cases. At the same time, the computation time is only about 50% of that from MASCOT Distiller. In the case of the larger condensin dataset, MASCOT Distiller did not complete computation regularly and interrupted with a run-time error. As was shown above, application of our software improved the interpretability of the condensin dataset.

3.11 Future developments

It should be noted that possibilities for further improvement of background removal and of computation costs reduction are evident. Unfortunately, most spectra do not contain useful peptide information. At the same time, the currently proposed mechanism for finding noninterpretable spectra detects only a minor fraction of them. As the data for the BSA example in Table 1 show, only 70 spectra out of 2679 (2.6%) are interpretable by MASCOT but only 195 out of 2679 (6.7%) have been unselected by our algorithm as "bad." Similar results have been found for other protein targets (Tables 1A and 3). Therefore, identification of noninterpretable spectra early in the workflow is critical for reducing the computational load [42, 43]. Sequence ladder testing and entropic criteria are simple and efficient alternatives with virtually no false positives (manuscript in preparation).

4 Concluding remarks

The background from multiply charged replicates, isotope variants, sample-specific and systematic contaminations, and the noise from the electronic detection system create considerable problems during mass spectrum interpretation. Computation time is wasted for noninterpretable spectra, and background peaks occupy a significant share of the storage capacity for mass-spectrometric data.

The procedures described in this article are able to remove a considerable part of these problems. The data show that background removal following our recipes improves reliability of hit assignments by database search-based methods (as tested by interpretability with MASCOT) considerably (as measured by scores and, in part, also by peptide match and sequence coverage). On the technical side, both the storage requirement for datasets and the computation time with MS/MS spectra interpretation software is reduced by 25–40% as a result of noise reduction with our tool.

Our tool is designed for applications in a proteomics context where lots of spectra need to be automatically interpreted. It does not aim to compete with manual noise identification by experts. The efficiency of the multiply charged isotope cluster recognition procedure depends on measurement accuracy and scanning speed; the better the isotope clusters are resolved (for example, with Fourier-transformation instruments in contrast to LCQ/LTQ in this work), the better will be their determination with the correlation analysis approach. Similarly, the spectral removal criteria for latent periodic noise and high-frequency noise might require new parameterization if more accurate instruments are applied.

In the future, we will analyze how *de novo* sequencing with MS/MS data will benefit from this type of background removal.

The authors are grateful to Werner Kubina, Georg Schneider, and Michael Wildpaner for advise in software design and for maintaining the local MASCOT implementation, to Ines Steinmacher for support in carrying out MS measurements, to Christian Gieffers for the Smc2 antibody, to Kerstin Wendt (all from IMP, Vienna) for advice on purifying condensin complexes, and *to Kurt Varmuza (Technical University, Vienna) for advise. This work has been supported by Boehringer Ingelheim, Gen-AU BIN (to F. E.), Gen-AU APP (to K. M.), and the European Commission FP6 Integrated Project "MitoCheck" (to J.-M. P.).*

5 References

[1] Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F. *et al.*, *Proc. Natl. Acad. Sci. USA* 1996, *93*, 14440–14445.

[2] Pandey, A., Mann, M., *Nature* 2000, *405*, 837–846.

[3] McCormack, A. L., Schieltz, D. M., Goode, B., Yang, S. *et al.*, *Anal. Chem.* 1997, *69*, 767–776.

[4] Washburn, M. P., Wolters, D., Yates, J. R., *Nat. Biotechnol.* 2001, *19*, 242–247.

[5] Wysocki, V. H., Tsaprailis, G., Smith, L. L., Breci, L. A., *J. Mass Spectrom.* 2000, *35*, 1399–1406.

[6] Hunt, D. F., Yates, R., Shabanowitz, J., Winston, S., Hauer, C. R., *Proc. Natl. Acad. Sci. USA* 1986, *83*, 6233–6237.

[7] Poulter, L., Tylor, L. C., *Int. J. Mass Spectrom. Ion Process.* 1989, *91*, 183–197.

[8] Alexander, A. J., Thibault, P., Boyd, R. K., Curtis, J. M., Rinehart, K. L., *Int. J. Mass Spectrom. Ion Process.* 1990, *98*, 107–134.

[9] Somogyi, A., Wysocki, V. H., Mayer, I., *J. Am. Soc. Mass Spectrom.* 1994, *5*, 704–717.

[10] Papayannopoulos, I. A., *Mass Spectrom. Rev.* 1995, *14*, 49–73.

[11] Cox, K. A., Gaskell, S. J., Morris, M., Whiting, A., *J. Am. Soc. Mass Spectrom.* 1996, *7*, 522–531.

[12] Dongre, A. R., Jones, J. L., Somogyi, A., Wysocki, V. H., *J. Am. Soc. Mass Spectrom.* 1996, *118*, 8365–8374.

[13] Yergey, J., Heller, D., Hansen, G., Cotter, R. J., Fenselau, C., *Anal. Chem.* 1983, *55*, 353–356.

[14] Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., Whitehouse, C. M., *Science* 1989, *246*, 64–71.

[15] Mann, M., *Org. Mass Spectrom.* 1990, *25*, 575–587.

[16] Smith, R. D., Loo, J. A., Ogorzalek Loo, R. R., Busman, M., Udseth, H. R., *Mass Spectrom. Rev.* 1991, *10*, 359–451.

[17] Kebarle, P., Tang, L., *Anal. Chem.* 1993, *65*, 972A–986A.

[18] McLafferty, F. W., *Acc. Chem. Res.* 1994, *27*, 379–386.

[19] Scoble, H. A., Biller, J. E., Biemann, K., *Fresenius Z. Anal. Chem.* 1987, *327*, 239–245.

[20] Bartels, C., *Biomed. Environ. Mass Spectrom.* 1990, *19*, 363–368.

[21] Johnson, R. S., Taylor, J. A., *Mol. Biotechnol.* 2002, *22*, 301–315.

[22] Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E., Pevzner, P. A., *J. Comput. Biol.* 1999, *6*, 327–342.

[23] Zhang, Z., McElvain, J. S., *Anal. Chem.* 2000, *72*, 2337–2350.

[24] Horn, D. M., Zubarev, R. A., McLafferty, F. W., *PNAS* 1994, *97*, 10313–10317.

[25] Taylor, J. A., Johnson, R. S., *Anal. Chem.* 2001, *73*, 2594–2604.

[26] Eng, J. K., McCormack, A. L., Yates, J. R., *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[27] Yates, J. R., Eng, J., McCormack, A. L., Schieltz, D. M., *Anal. Chem.* 1995, *67*, 1426–1436.

[28] Yates, J. R. III, McCormack, A. L., Eng, J., *Anal. Chem.* 1996, *68*, 534A–540A.

[29] Yates, J. R. III, Eng, J. K., McCormack, A. L., *Anal. Chem.* 1995, *67*, 3202–3210.

[30] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[31] Sadygov, R. G., Eng, J., Durr, E., Saraf, A. *et al.*, *J. Proteome. Res.* 2002, *1*, 211–215.

[32] Zhang, N., Aebersold, R., Schwikowski, B., *Proteomics* 2002, *2*, 1406–1412.

[33] Mann, M., Meng, C. K., Fenn, J. B., *Anal. Chem.* 1989, *61*, 1702–1708.

[34] Ferrige, A. G., Seddon, M. J., *Rapid Commun. Mass Spectrom.* 1991, *5*, 374–379.

[35] Reinhold, B. B., Reinhold, V. N., *J. Am. Soc. Mass Spectrom.* 1992, *3*, 207–215.

[36] Zhang, Z., Marshall, A., *J. Am. Soc. Mass Spectrom.* 1998, *9*, 225–233.

[37] Gentzel, M., Kocher, T., Ponnusamy, S., Wilm, M., *Proteomics* 2003, *3*, 1597–1610.

[38] Wehofsky, M., *Thesis*, Justus-Liebig-Universität Giessen, Germany, 2001.

[39] Wehofsky, M., Hoffmann, R., *J. Mass Spectrom.* 2002, *37*, 223–229.

[40] Jaitly, D., Page-Belanger, R., Faubert, D., Thibault, P., Kebarle, P., *MSMS Peak Identification and its Applications*, ISMB/ECCB 2004, 2004, http,//www.ismb.org/ismbeccb2004/accepted_papers (Communication 46), 1–3.

[41] Horn, D. M., Zubarev, R. A., McLafferty, F. W., *J. Am. Soc. Mass Spectrom.* 2000, *11*, 320–332.

[42] Xu, M., Geer, L. Y., Bryant, S. H., Roth, J. S. *et al.*, *J. Proteome. Res.* 2005, *4*, 300–305.

[43] Bern, M., Goldberg, D., McDonald, W. H., Yates, J. R. III, *Bioinformatics* 2004, *20*, I49–I54.

[44] Hirota, T., Gerlich, D., Koch, B., Ellenberg, J., Peters, J. M., *J. Cell Sci.* 2004, *117*, 6435–6445.

[45] Mitulovic, G., Smoluch, M., Chervet, J. P., Steinmacher, I. *et al.*, *Anal. Bioanal. Chem.* 2003, *376*, 946–951.

[46] Mitulovic, G., Stingl, C., Smoluch, M., Swart, R. *et al.*, *Proteomics* 2004, *4*, 2545–2557.

[47] Blom, K. F., *Org. Mass Spectrom.* 1988, *23*, 194–203.

[48] She, J., McKinney, M., Petreas, M., Stephens, R., *Organohalogen Compd.* 1995, *23*, 171–174.

[49] Rockwood, A. L., *Rapid Commun. Mass Spectrom.* 1995, *9*, 103–105.

[50] Rockwood, A. L., VanOrden, S. L., *Anal. Chem.* 1996, *68*, 2027–2030.

[51] Rockwood, A. L., VanOrden, S. L., Smith, R. D., *Rapid Commun. Mass Spectrom.* 1996, *10*, 54–59.

[52] Baranov, V., US Patent 6590 204, 2003.

[53] Friedlander, B., Porat, B., *IEEE Trans. Aerosp. Electron. Syst.* 1984, *AES-20*, 158–173.

[54] Oppenheim, A. V., Schafer, R. W., *Discrete-Time Signal Processing*, Englewood Cliffs, Prentice-Hall, NJ 1989.

[55] Parks, T. W., Burrus, C. S., *Digital Filter Design*, John Wiley & Sons, New York 1987.

[56] Ono, T., Losada, A., Hirano, M., Myers, M. P. *et al., Cell* 2003, *115*, 109–121.

[57] Schleiffer, A., Kaitna, S., Maurer-Stroh, S., Glotzer, M. *et al., Mol. Cell* 2003, *11*, 571–575.

[58] Yeong, F. M., Hombauer, H., Wendt, K. S., Hirota, T. *et al., Curr. Biol.* 2003, *13*, 2058–2064.

6 Addendum: Web supplement

At the website http://mendel.imp.univie.ac.at/mass-spectrometry/, supplementary resources are available: (i) the webpage with illustrative examples that motivate the application of frequency-spectral criteria for background removal in MS/MS spectra at the link http://mendel.imp.univie.ac.at/mass-spectrometry/ANALYSIS/; (ii) the raw MS/MS data (mgf format) and the respective background-cleaned version for the BSA, ADH, and TRF samples together with Supplementary tables showing the changes of scores of leading peptide hits; (iii) views of the original spectra used in Fig. 3; (iv) the dta files and MS/MS-spectral views of the data used in Fig. 4; (v) details about the single false-positively rejected spectrum from the TRF dataset (http://mendel.imp.ac.at/mass-spectrometry/falsepositive-partA.html); (vi) result listings for the condensin dataset of the MASCOT search against the nonredundant protein sequence database both with restriction to human proteins and without any taxonomic restriction; (vii) details about a false-positively rejected spectrum from the condensin dataset (http://mendel.imp.ac.at/mass-spectrometry/falsepositive-partB.html), and (viii) a demonstration version of the MS Cleaner with user manual (Windows XP edition).