Bioinformatics, 37(24), 2021, 4719–4726
doi: 10.1093/bioinformatics/btab563
Advance Access Publication Date: 29 July 2021
Original Paper

Gene expression

Improved identification and quantification of peptides in mass spectrometry data via chemical and random additive noise elimination (CRANE)

Akila J. Seneviratne, Sean Peters, David Clarke, Michael Dausmann, Michael Hecker, Brett Tully , Peter G. Hains and Qing Zhong *

ProCan®, Children's Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, Westmead, NSW, Australia

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on August 11, 2020; revised on June 15, 2021; editorial decision on July 21, 2021; accepted on July 28, 2021

Abstract

Motivation: The output of electrospray ionization–liquid chromatography mass spectrometry (ESI-LC-MS) is influenced by multiple sources of noise and major contributors can be broadly categorized as baseline, random and chemical noise. Noise has a negative impact on the identification and quantification of peptides, which influences the reliability and reproducibility of MS-based proteomics data. Most attempts at denoising have been made on either spectra or chromatograms independently, thus, important 2D information is lost because the mass-to-charge ratio and retention time dimensions are not considered jointly.

Results: This article presents a novel technique for denoising raw ESI-LC-MS data via 2D undecimated wavelet transform, which is applied to proteomics data acquired by data-independent acquisition MS (DIA-MS). We demonstrate that denoising DIA-MS data results in the improvement of peptide identification and quantification in complex biological samples.

Availability and implementation: The software is available on Github (https://github.com/CMRI-ProCan/CRANE). The datasets were obtained from ProteomeXchange (Identifiers—PXD002952 and PXD008651). Preliminary data and intermediate files are available via ProteomeXchange (Identifiers—PXD020529 and PXD025103).

Contact: qzhong@cmri.org.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

To date, most advances in bioinformatics research have come from the study of DNA (genomics) and RNA (transcriptomics) (Ellis and Perou, 2013; Kong *et al.*, 2020). The technology for high-throughput analysis of proteins (proteomics) has emerged relatively recently (Aebersold and Mann, 2016; Bludau and Aebersold, 2020). One of the major challenges in mass spectrometry (MS) based proteomics is the 'noise' inherent in raw MS data that can obscure peptide signals. These noisy signals interfere with the accurate identification and quantification of peptides, which propagates in downstream analyses.

In micro-channel plate detectors, the collision of an ion with the detector generates a signal. When a host of ions collide with the detector, there could be residual accumulation at the detector, which results in a shift in the baseline (Schneider, 2016). As with most electrical and electronic equipment, MS instruments also produce random noise independent of the sample under study. Random noise depends on the operating conditions, such as gain, temperature and age of the instrument (Chou *et al.*, 2012; Schneider, 2016). The random noise is typically of high frequency and could occur at any point in the mass-to-charge ratio (m/z) and retention time space. In addition, the MS data are influenced by multiple sources of impurities introduced during sample preparation, the reagents used in liquid chromatography, or from particles in the atmosphere. Most of these impurities cause chemical noise, which is distributed over a wide range of retention time, and has a constant value for m/z (Cappadona *et al.*, 2008).

Many attempts have been made to denoise MS data. Most of them denoise spectra (Awan and Saeed, 2016; Coombes *et al.*, 2005; Ding *et al.*, 2009; Du *et al.*, 2008; Kwon *et al.*, 2008; Li *et al.*, 2007; Mujezinovic *et al.*, 2006, 2010; Renard *et al.*, 2009; Shao and Lam, 2013; Yang and Yu, 2011; Zhang *et al.*, 2008), yet important spatial information that is discriminative between the signal of interest and noise is lost when the MS data are summed over retention time. A review of spectrum denoising techniques is given in Yang and Yu

(2011), detailing techniques, such as moving average, Savitzky-Golay filter, Gaussian filter and denoising in wavelet transform domain via L_1 penalized least squares estimation.

PeakSelect (Zhang *et al.*, 2008) explores the fact that ions have isotopes but noise does not in order to differentiate between signal and noise. It uses a Gaussian mixture model and an expectation–maximization algorithm to find the base intensity level (baseline) in a spectrum. MS Cleaner (Mujezinovic *et al.*, 2010) uses a modified Fourier-transform-based criterion to clear the background in the data. MS-REDUCE (Awan and Saeed, 2016) is a low-complexity technique that consists of three steps; spectral classification, peak quantization and weighted random sampling. PeakSelect, MS Cleaner and MS-REDUCE are data dependant acquisition based spectral denoising techniques.

In Coombes *et al.* (2005) and Yang and Yu (2011), the spectra are decomposed into the noiseless signal, baseline noise and random noise. Cappadona *et al.* (2008) claim that the above-mentioned decomposition of the observed signal into three independent terms to be too naive, arguing that a complete description of the data requires consideration of the possible correlation between the terms. Unlike other studies (Coombes *et al.*, 2005; Du *et al.*, 2008; Kwon *et al.*, 2008; Li *et al.*, 2007; Yang and Yu, 2011), denoising is performed on single ion chromatographs in Cappadona *et al.* (2008) and uses a model that takes into account the heteroscedasticity of the stochastic noise. The data heteroscedasticity is discussed in Hundertmark *et al.* (2009), which states that the noise is essentially multiplicative and varies with signal intensity. By means of undecimated wavelet transform (UWT), Cappadona *et al.* (2008) claim to remove both the chemical and the random noise by performing wavelet smoothing and wavelet denoising, respectively. However, careful analysis shows that the algorithm calculates a baseline using the smoothed and denoised signals and performs only a baseline subtraction.

Similar to Cappadona *et al.* (2008), denoising is performed on chromatograms in Ning *et al.* (2014), which presents an algorithm for Baseline Estimation and Denoising With Sparsity. This method explores the sparse derivative nature of the signal and derives a majorization–minimization approach to optimize the penalized criterion.

The technique given in Cappadona *et al.* (2008) perform denoising of single ion chromatograms, allowing partial exploration of the 2D information of MS data. However, it is not strictly a 2D denoising technique, because m/z and retention time dimensions are not considered jointly.

Here, we present chemical and random additive noise elimination (CRANE), a technique for removing three major sources of noise from raw electrospray ionization–liquid chromatography mass spectrometry (ESI-LC-MS) data via 2D UWT (Starck *et al.*, 2007), inspired by image denoising techniques. It is applied to raw (unprocessed) ESI-LC-MS data and explores 2D information compared with previous 1D approaches in the field. Wavelet transformation is a technique widely used to process signals in the frequency domain, which facilitates the separation of data into different components, thus isolating undesirable components, such as noise. For these reasons, image processing techniques (Cai and Harrington, 1998; Yang and Fei, 2011) based on wavelet transform are modified for denoising MS data in this study.

The aim of this study is to develop a denoising algorithm for MS data that can add value to proteomic studies by enhancing high quality identification and quantification of peptides and proteins in a single sample. A description of the CRANE algorithm is given in Section 2, which details how the wavelet coefficients are manipulated to remove the baseline, random and chemical noise. The effectiveness of CRANE is demonstrated by denoising the MS1 (first stage of MS) and all the MS2 (second stage of MS) windows of the data-independent acquisition (DIA) MS datasets from Navarro *et al.* (2016) and Krasny *et al.* (2018). The original files and the denoised files are processed using OpenSWATH (Röst *et al.*, 2014) and Pyprophet (Rosenberger *et al.*, 2017). The output is compared using various performance measures to demonstrate the benefits of CRANE.

2 Crane algorithm

Under standard ESI-LC-MS conditions, m/z and the retention time are discretely sampled from two orthogonal dimensions. Therefore, when all scans of an MS window are drawn adjacent to each other, the acquired data can be represented as a 2D image (Supplementary Fig. S1). The remainder of this section describes each step of the CRANE algorithm (Fig. 1) and the importance of each step (Figs 2 and 3). Pseudocode of the algorithm is given in the Supplementary Section S1.2.

2.1 Index space conversion

Wavelet transform assumes that the measurements are made equi-spaced from each other. If we were applying wavelet transform to a time series, then, the time lapse between two measurements would need to be the same. Thus, before wavelet transform can be applied to MS data, it must be transformed into appropriate coordinates. For the retention time coordinate, if the cycle time is t_c and the first scan is at t_0, then the retention time vector of each window can be described as

$$t(i_R) = t_c i_R + t_0, \tag{1}$$

where i_R is the retention time index. For the m/z coordinate, the mass analyser in a time-of-flight (TOF) mass spectrometer measures the time taken for a charged ion, accelerated by an electric field of known strength (U), to move a known distance (d). It is a digital sensor that samples at a constant frequency. Thus, measurements are made at constant time intervals

$$t(i_{\mathrm{MZ}}) = t_p\ (i_{\mathrm{MZ}} + \gamma), \tag{2}$$

where i_{MZ} is an integer value, t_p is constant and γ is a constant offset, and this time is related to the m/z of the ion such that

$$t(i_{\mathrm{MZ}}) = \frac{d}{\sqrt{2U}}\sqrt{\frac{m_i}{z_i}} + \mathrm{O}, \tag{3}$$

where m_i and z_i are the mass and the charge of the ith peptide and O represents higher-order terms. With this in mind, using Equations (2) and (3) it becomes possible to convert m/z to an integer index through the following relationship (Chernushevich *et al.*, 2001)

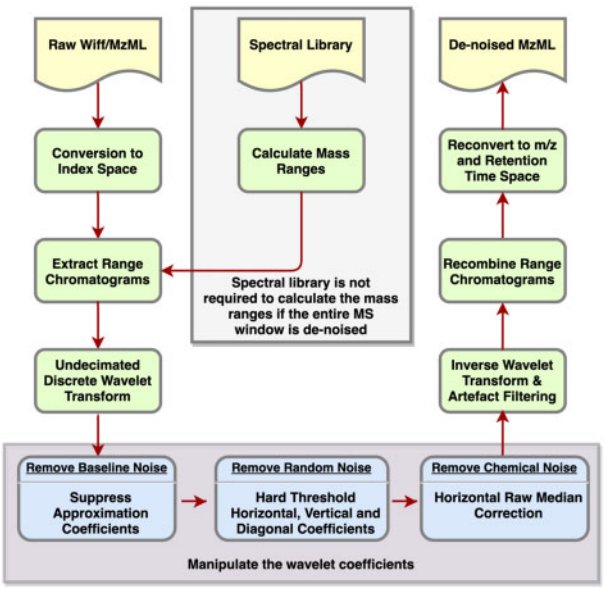

Fig. 1. CRANE algorithm flow chart, indicating the steps that are taken to convert a raw MS data file into the wavelet domain, how wavelet coefficients are manipulated to remove the three major sources of noise, and how data are converted back to m/z and retention time domain to create a denoised data file

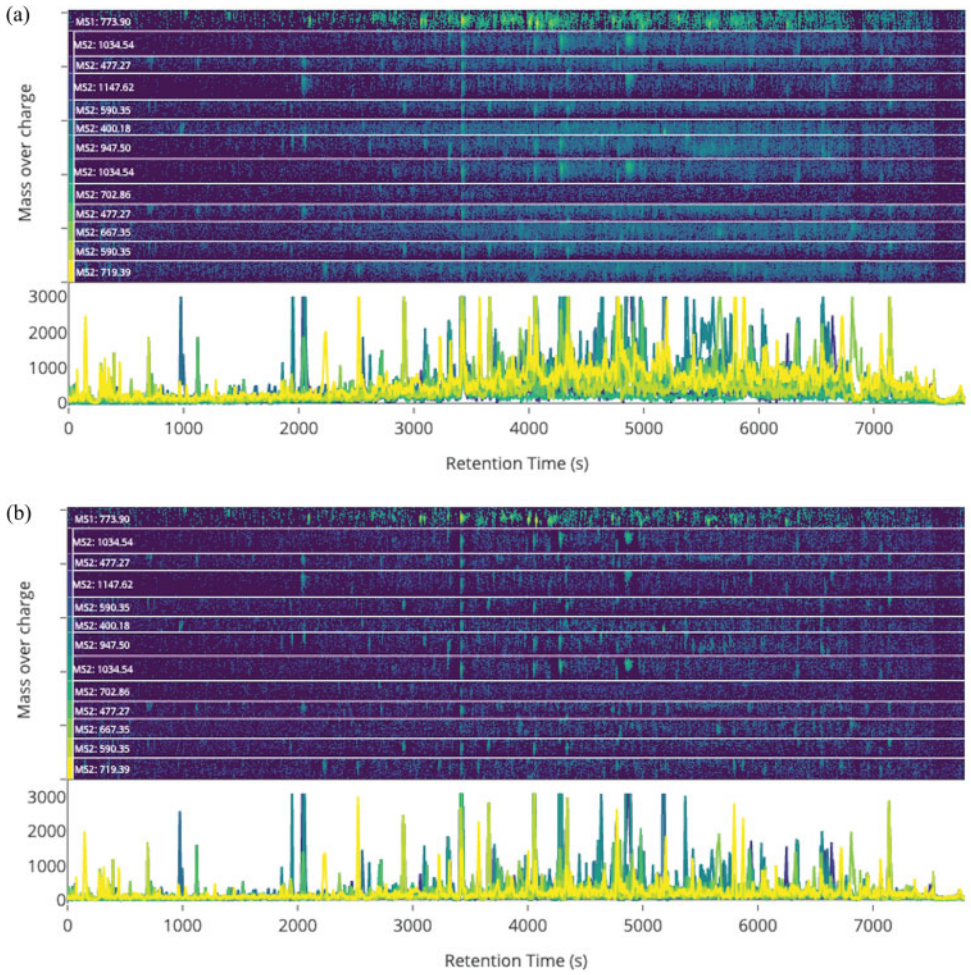

Fig. 2. Illustration of baseline noise removal using a fragment plot of AAADALSDLEIKDSK peptide extracted from the hye124_ttof5600_32fix_lgillet_l150206_001 file of the multicentre dataset from Navarro *et al.* (2016). A fragment plot consists of two sections. The top half shows intensity plots of the extracted ion envelops of a precursor and its fragments stacked one after the other. The bottom half of the fragment plot shows the corresponding chromatograms. All the extracted ion envelops have 100 ppm width. (**a**) Raw fragment plot. Between 2500 and 7500 s the chromatogram peaks originate at a level higher than zero showing a shift in the baseline. (**b**) CRANE denoised fragment plot

$$i_{\mathrm{MZ}} + \gamma = \frac{d}{t_p\sqrt{2U}}\sqrt{\frac{m_i}{z_i}} + O \Rightarrow \frac{m_i}{z_i} = (\alpha[i_{\mathrm{MZ}} + \gamma] + \beta)^2, \quad (4)$$

where $\alpha = \frac{t_p\sqrt{2U}}{d}$ and β represents higher-order terms.

The current implementation of CRANE accepts as input MS data in the Toffee file format (Tully, 2020). The Toffee file format is used as it converts the MS data into index space and enables fast subsampling of the data. The index conversion in Toffee supports TOF instrument data and further work needs to be done before it can be effectively used for Orbitrap data.

2.2 Mass range calculation

A vast amount of computer memory is required to denoise an entire MS window at once. Therefore, strips of data are extracted from the MS window and processed independently. There are two different methods for data extraction, and the first is to split the entire window into strips. However, this method is computationally expensive. Depending on the available memory and the number of levels of wavelet decomposition required, a suitable mass index width is defined. Then the entire MS window is subdivided along the m/z axis into strips with an equal width. Each strip contains data along the entire retention time span of the experiment, which we name as an extracted ion envelop (Fig. 3a and Supplementary Fig. S1) following the nomenclature given in Smith *et al.* (2014). The reason for splitting along the m/z axis instead of the retention time axis is to facilitate the identification and removal of chemical noise.

The second method of data extraction is to identify the m/z ranges of interest based on the spectral library. Pipelines, such as OpenSWATH (Röst *et al.*, 2014), require a spectral library, which has information about the expected m/z of the peptides and fragments of interest. Based on the m/z values listed in the spectral library and data used by downstream proteomic pipelines, the m/z ranges that require denoising can be identified. Similar to the first method, strips of data are extracted along the entire retention time range. Since only the areas of interest are denoised, this method is computationally efficient. However, the drawback is that the files may need to be denoised again if the spectral library or the pipeline changes.

2.3 Wavelet transform

The extracted raw data strips are transformed to the wavelet domain (Starck *et al.*, 2007) and the current implementation of CRANE uses the PyWavelets library (Lee *et al.*, 2019). This module accepts as input an intensity matrix of the extracted data with each row having a fixed m/z index and each column having a fixed retention time index. At each level of 2D UWT, the signal is decomposed into approximation, horizontal, vertical and diagonal components. The algorithm given in Cappadona *et al.* (2008) decomposes single ion chromatograms up to six levels using Coifmann wavelet of order one. Empirically, we have found that Daubechies wavelet of order two works better for CRANE. An adaptive wavelet level selection technique is described in the Supplementary Section S1.3.

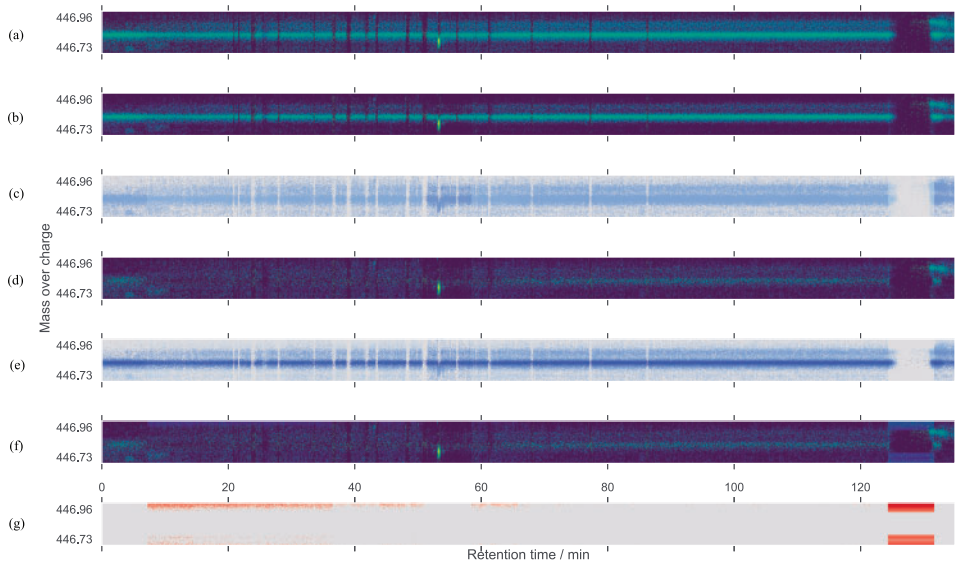

Fig. 3. Illustration of the importance of CRANE algorithmic steps using an extracted ion envelop. (**a**) The raw extracted ion envelop shows a section of the MS1 window of a DIA-MS data file as an intensity map with retention time as the horizontal axis and m/z as the vertical axis. The bright dot at ~55 min is a peptide of interest and the horizontal line is chemical noise. (**b**) The extracted ion envelop with horizontal, vertical and diagonal components VisuShrink hard thresholded and approximation coefficients suppressed. This shows the output if the chemical noise removal step is skipped. (**c**) Noise removed by the hard thresholding step, difference between a and b. (**d**) CRANE denoised extracted ion envelop with baseline, random and chemical noise removed, and artefact filtered after inverse transform to spatial domain. (**e**) Noise removed by CRANE. (**f**) CRANE denoised extracted ion envelop without artefact filtering. (**g**) Artefacts in Figure 3f showing the importance of artefact filtering

The properties of the wavelet transform that help in the image denoising are sparseness, clustering and correlation between neighbouring wavelet coefficients (Donoho and Johnstone, 1995). All these properties help in differentiating the noise from the signal, thus enabling its removal.

Unlike the discrete wavelet transform, which down-samples the approximation and detail coefficients at each decomposition level, the UWT does not incorporate the down-sampling operations. Therefore, the approximation and detail coefficients at each level are the same length as the original signal. Consequently, UWT is an inherently redundant scheme. This redundancy also aids in denoising and is thus used in CRANE.

2.4 Baseline noise removal

When many analytes hit the detector in quick succession (e.g. analysing a complex tissue sample with many peptides), there is a baseline shift (Fig. 2a). This is a very slow-moving variation; hence, it is captured in the low frequency components. The baseline noise can be removed by suppressing the approximation coefficients, which capture the low frequency components (Fig. 2b).

2.5 Random noise removal

The wavelet transform's energy compactness helps greatly in denoising. Energy compactness refers to the fact that most of the signal energy is contained in a few large wavelet coefficients, whereas a small portion of the energy is spread across a large number of small wavelet coefficients. These coefficients represent details as well as high frequency noise in the image. By appropriately thresholding these wavelet coefficients, image denoising is achieved while preserving fine structures in the image.

There are two decisions that must be made when performing denoising via wavelet coefficient thresholding, namely threshold selection and thresholding technique.

2.5.1 Threshold selection

There are many techniques including VisuShrink (Donoho and Johnstone, 1994), SUREShrink (Donoho and Johnstone, 1995) and BayesShrink (Chipman *et al.*, 1997) for the selection of the threshold. From empirical results, VisuShrink is optimal for CRANE. VisuShrink threshold is calculated as,

$$T = \sigma\sqrt{2\log M}, \tag{5}$$

where σ is the SD of noise and M is the number of pixels in the image. When the noise variance is unknown, it can be estimated via the median absolute deviation (MAD) of the lowest level of diagonal coefficients (D_1). If we assume that the noise can be modelled as a zero mean Gaussian, then the following relationship holds, $\text{MAD} = \text{median}(|d_{ij}^1 - \text{median}(D_1)|) \approx 0.67449\,\sigma$, where $d_{ij}^1 \in D_1$, i and j are the column and row numbers of the D_1 coefficient matrix. Since the threshold is selected per extracted ion envelop, we cater for the heteroscedasticity of noise.

2.5.2 Thresholding technique

There are many methods of applying the threshold to wavelet coefficients. Hard thresholding sets all the coefficients lower than the threshold to zero and leave the other coefficients unaffected. The potential limitation of this technique is that it could lead to artefacts after inverse wavelet transform. Soft thresholding, on the other hand, shrinks all coefficients towards zero by the threshold value (Donoho and Johnstone, 1994), which addresses the issue of artefacts. However, in this application, it could be undesirable as this could lead to a drop in the value of peptide intensities. CRANE uses hard thresholding with an artefact filter.

The effects of various steps of the CRANE algorithm are illustrated based on a raw extracted ion envelop (Fig. 3a). The effects of VisuShrink with hard thresholding and approximation coefficient suppression are given in Figure 3b and the noise removed is given in Figure 3c.

2.6 Chemical noise removal

Figure 3b demonstrates that chemical noise is unaffected by random and baseline noise removal. Image denoising techniques try to preserve line structures, thus removing chemical noise is not covered by image denoising techniques. In many cases, chemical noise has a

similar m/z to that of the peptides of interest. Therefore, thresholding in spatial domain will affect the intensities of the peptides.

There are many ways of approaching this problem. Extracting the line structure using ridgelet transform (Candès, 1998; Fadili and Starck, 2012) is one such approach. However, given the dimension of the raw data files, a simple and computationally efficient method is desirable.

If the extracted raw data are plotted as a heatmap with retention time as the *x* axis and m/z as the *y* axis, the chemical noise appears parallel to the *x* axis (Fig. 3a and Supplementary Fig. S1). Consequently, it appears only in the horizontal components in the wavelet decomposition. Therefore, by performing row median correction of the horizontal components, we are able to successfully remove the chemical noise. The vertical and diagonal components carry sufficient information to restore isotropic structures, such as the peptide of interest (Fig. 3d and e).

2.7 Inverse transform and artefact filtering

In the current implementation of CRANE, the inverse wavelet transform is performed using the PyWavelet library (Lee *et al.*, 2019). Horizontal row median correction and VisuShrink hard thresholding could result in artefacts when inverse-transformed to the spatial domain. Therefore, the following artefact filters are used. First, any negative values in the spatial domain in the denoised data are set to zero because MS intensities are always positive. Second, if the denoised signal is higher than the original at any point, the denoised pixels are replaced by the raw data, since CRANE assumes that noise is additive.

Figure 3f shows the denoised data if only the negative correction is applied to the denoised data. Figure 3g shows the artefacts that have been introduced. Figure 3d shows that artefact filtering resolves this issue.

3 Results

We have chosen the multicentre and matrisome datasets released by Navarro *et al.* (2016) and Krasny *et al.* (2018), respectively, to demonstrate the performance of CRANE. Raw files and the CRANE denoised files were processed with and without the in-built background subtraction in OpenSWATH, which we call OSW1 and OSW0, respectively. OpenSWATH has two background subtraction settings: original and exact. Since the results were comparable for both settings, we will only show the results of the original setting in this article as OSW1. In OSW1, the background is computed as the average of the left and right hand edges of the peak, which is then multiplied by the width of the peak. This value is then subtracted from the peak intensity to obtain the denoised intensity. Raw files processed with OSW0 are considered as the reference since none of the denoising techniques is applied to that set of data.

3.1 Multicentre benchmarking dataset

Multicentre dataset was selected because it comprises DIA-MS data acquired with different instrument platforms and acquisition methods. Raw data files were obtained from ProteomeXchange (dataset PXD002952). This dataset includes two experiments, each with hybrid proteome samples, consisting of tryptic digests of human, yeast and *Escherichia coli* proteins. The proportions of the hybrid samples A and B of experiment HYE124 were set so as to yield expected peptide and protein ratios of 1:1 (A/B) for human, 2:1 for yeast and 1:4 for *E.coli* proteins. The samples of experiment HYE110 were prepared so the peptide and protein ratios were 1:1 (A/B) for human, 10:1 for yeast and 1:10 for *E.coli* proteins. All the samples of experiments HYE124 and HYE110 have 65% and 67% of human proteins, respectively. Data have been collected using SCIEX TripleTOF (TTOF) 5600 and TTOF 6600 instruments, using four different Sequential Window Acquisition of All Theoretical Mass Spectra acquisition methods by varying the window number (32 or 64) and window size (fixed or variable) (Navarro *et al.*, 2016). Details of how the peptide and protein data were generated are given in the Supplementary Section S2.

3.1.1 Technical variance

The dispersion of quantitative values is calculated as coefficients of variation (CV) for each identified peptide among technical replicates of each sample. The peptide CV distribution of CRANE denoised files with OSW0 is similar to that of raw files processed with OSW0. However, when OpenSWATH background subtraction feature is enabled (OSW1), the CVs increase both in terms of median and variance (Supplementary Fig. S2a and b and Tables S2 and S3). The dispersion of the quantified protein values among technical replicates of each sample shows a similar pattern (Supplementary Fig. S2c and d and Tables S2 and S3).

3.1.2 Identification statistics

Denoising via CRANE increases the peptide and protein identifications when 32 windows are used. There is no significant change in the number of identifications on CRANE denoised files when 64 windows are used. OSW1 shows a drop in the number of peptide and protein identifications compared with OSW0 when the TTOF 6600 is used (Fig. 4 and Supplementary Figs S3–S5 and Tables S4–S9).

Sample A has a higher proportion of yeast proteins than sample B in both HYE124 and HYE110 experiments. In experiment HYE110, CRANE_OSW0 with the 32-window scheme show a 3–5% improvement in the number of yeast peptide and protein identifications in sample A compared to a 7–16% improvement in sample B (Fig. 4 and Supplementary Fig. S4) compared with RAW_OSW0. In experiment HYE124, CRANE denoised files show a similar pattern of a higher improvement in the number of peptide and protein identifications in sample B compared with sample A when the 32-window scheme is used (Supplementary Figs S3 and S5). Therefore, denoising with CRANE has assisted with the identifications of low abundance yeast peptides and proteins in the 32-window scheme.

In contrast, when the TTOF 6600 is used, OSW1 shows a larger drop in the number of yeast peptide and protein identifications in sample B compared with sample A. In experiment HYE110, RAW_OSW1 shows a 6–11% drop in the number of yeast peptide and protein identifications in sample B compared with about 2% drop in sample A (Fig. 4 and Supplementary Fig. S4). OSW1 also shows a larger drop for the 32-window scheme compared with that of the 64 windows when the TTOF 6600 is used (Supplementary Tables S5 and S8).

These observations are mirrored in *E.coli* peptide and protein identifications as well. CRANE_OSW0 increases low abundance peptide and protein identifications and therefore displays a larger increase of identifications in sample A compared with sample B when 32-window scheme is used. While OSW1 seem to have the opposite effect when the TTOF 6600 is used by displaying a higher drop in the

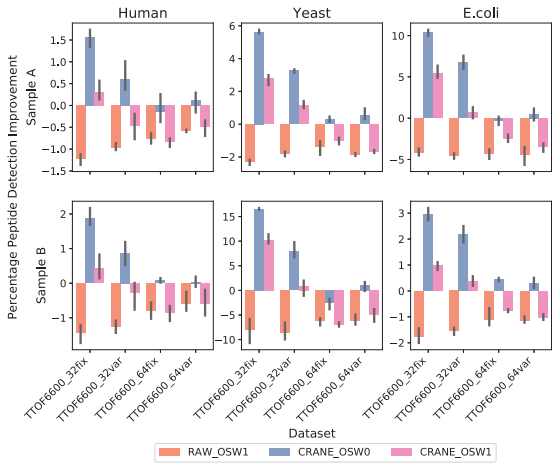

Fig. 4. Percentage peptide detection improvements of RAW_OSW1, CRANE_OSW0 and CRANE_OSW1 compared to the reference (RAW_OSW0) of samples A and B of the four datasets of experiment HYE1110

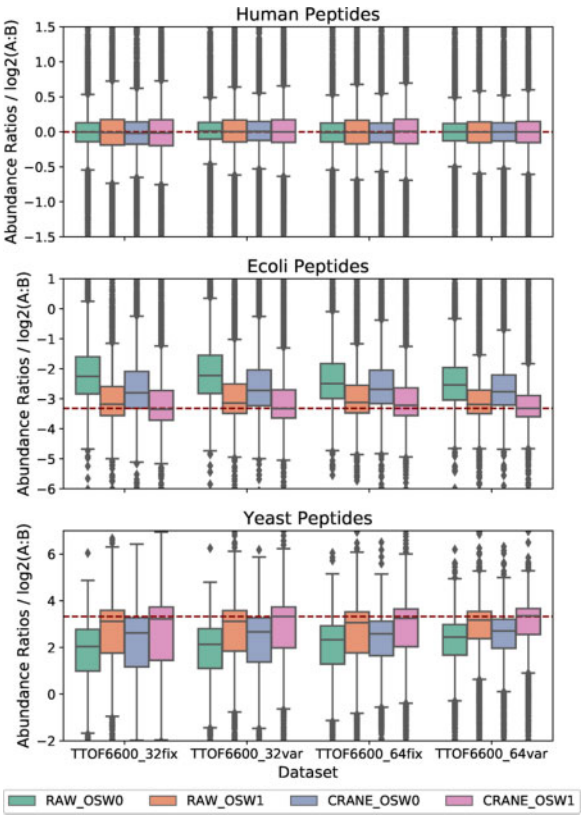

Fig. 5. Peptide abundance ratio distribution comparison for multicentre experiment HYE110 of human, yeast and *E.coli* peptides for RAW_OSW0, RAW_OSW1, CRANE_OSW0 and CRANE_OSW1. The red dashed lines show the expected abundance ratios

number of *E.coli* identifications in sample A compared with that of sample B (Fig. 4 and Supplementary Figs S3–S5 and Tables S6 and S9).

3.1.3 Quantification performance

Logarithmic (log2) ratios of quantified amounts are calculated for each identification and each sample pair in the dataset [log2(A/B)] (Fig. 5 and Supplementary Figs S6–S8 and Tables S10–S17). Since median normalization uses human peptides as the reference, the absolute median deviation from the expected abundance ratio is artificially adjusted to zero for human peptides and proteins (Fig. 5 and Supplementary Figs S6–S8). In order to compare on an equal footing, only the common peptides and proteins across all the tools are considered when calculating the median deviation from the expected and SD of the abundance ratio distributions.

The Levene's statistic is calculated to test the null hypothesis that the variance of the abundance ratio distributions of RAW_OSW1, CARNE_OSW0 and CRANE_OSW1 is that same as that of RAW_OSW0 for each species (Supplementary Tables S10–S13). CRANE_OSW0 has the lowest abundance ratio SD and with $P \gg 0.01$, its variance is the same as that of the reference (RAW_OSW0). The abundance ratio SD of RAW_OSW1 and CRANE_OSW1 is significantly higher.

The absolute median deviation from the expected abundance ratio for yeast and *E.coli* peptides and proteins of CRANE_OSW0, RAW_OSW1 and CRANE_OSW1 is notably smaller than that of RAW_OSW0 (Fig. 5 and Supplementary Figs S6–S8). The absolute median deviation from the expected abundance ratio for each tertile [first: lowest intensity (0–33.3%), second: medium intensity (33.3–66.7%) and third: highest intensity (66.7–100%)] show a similar pattern (Supplementary Tables S14–S17). Therefore, both CRANE and OSW1 denoising techniques improve the peptide and protein quantification accuracy.

We consider two abundance ratio validity ranges, defined as three times and five times of SD from the average log ratio, respectively (Supplementary Tables S18–S21). Following the pattern of identification statistics, CRANE_OSW0 has a higher number of peptides and proteins within the validity range when the 32-window scheme is used and has the same number of identifications as OSW-BS0 for the 64-window scheme. OSW1 has fewer identifications within the validity range compared to RAW_OSW0 when TTOF 6600 data are used. The increase in the number of yeast and *E.coli* peptides and proteins within the validity range for CRANE_OSW0 when the 32-window scheme is used can vary between 5% and 16%. The decrease in the number of yeast and *E.coli* peptides and proteins within the validity range for RAW_OSW1 can vary between 2% and 11%.

3.1.4 Species separation

The abundance ratio of the identifications of each species is expected to have a certain value that corresponds to the way the samples A and B are composed for experiments HYE124 and HYE110. Therefore, to measure the species overlap, we calculated the area under the curve (AUC) of the receiver operator characteristic curve constructed by varying an abundance ratio threshold. The inverse hyperbolic tangent (arctanh) transform of the AUC is listed in Supplementary Tables S22 and S23. In order to compare on an equal footing, only the common peptides and proteins across all the tools are considered for species separation calculations. CRANE_OSW0 is the best in species separation.

3.2 Matrisome dataset

The matrisome DIA-MS dataset is designed to analyse extracellular matrix (ECM) and ECM-associated proteins. Proteomic analysis of ECM proteins is difficult because many components of the matrisome, in particular the core matrisomal proteins, are highly insoluble and ECM-associated proteins are found in low abundance. Therefore, ECM enrichment strategies that are capable of solubilizing matrisomal components have been developed. Majority of ECM enrichment techniques lead to an enrichment in core ECM components but results in the decrease in the soluble ECM-associated proteins. Thus, ECM enrichment methods create a distorted view of the actual matrisomal content of the system under study (Krasny *et al.*, 2018). Raw data files were obtained from ProteomeXchange (dataset PXD008651). Data have been collected using SCIEX TTOF 5600 with 31 precursor isolation windows. Mouse liver and lung samples with and without ECM enrichment have been considered each with six biological replicates and three technical replicates. Details of how the peptide and protein data were generated are given in the Supplementary Section S4.

3.2.1 Technical variance

The peptide and protein CV data display the same trend as the multicentre data. CV distribution of CRANE denoised files with OSW0 is similar to that of raw files processed with OSW0. However, when OpenSWATH background subtraction feature is enabled (OSW1), the CV increases (Supplementary Figs S9 and S10).

3.2.2 Identification statistics

The ECM non-enriched samples have about three times as much peptide and protein identifications compared with the ECM enriched samples. In the liver samples, the ECM enriched samples have more matrisomal peptides and proteins compared with the non-enriched. However, in the lung samples, the ECM enrichment does not seem to improve ECM peptide and protein detections. There are fewer ECM-associated proteins detected in the ECM enriched samples. All the denoising techniques have improved the peptide and protein identifications. CRANE_OSW1 displays the highest peptide and protein detections closely followed by CRANE_OSW0. CRANE denoised files show 15–20%

improvement in peptide and protein detections with the highest improvement shown in non-enriched liver samples. Similar to the multicentre dataset, we observe that CRANE aids the detection of low abundance ECM-associated proteins (Supplementary Figs S11–S14 and Tables S24–S27).

3.2.3 Differential analysis

All the intensities are log2 transformed and the difference in protein expression is compared between the six biological replicates of liver and lung samples. ECM enriched and non-enriched samples are analysed separately. The log2 transform of the average intensity ratio (fold change) of lung versus liver is used in conjunction with the Welch's *t*-test *P*-value with threshold 0.05 to identify statistically significant differential expression (Supplementary Figs S15–S18). ECM non-enriched samples have a higher number of differentially expressed proteins. All the denoising techniques improve the number of differentially expressed proteins with CRANE_OSW1 having the highest improvement.

4 Future work

The current implementation of the CRANE algorithm is written in the Python programming language and is not optimized in terms of computation efficiency. The denoising of each file of the multicentre and matrisome datasets on an Intel Platinum 8168 2.7 GHz CPU took on average 36 and 31 h, respectively. An implementation using a language such as C++ would result in a denoising tool with shorter processing time.

5 Discussion

The step-by-step development of a novel denoising technique CRANE based on 2D UWT is presented. Results show CRANE is an effective technique for removing chemical, baseline and random noise.

The importance of each algorithmic step and the impact of denoising are illustrated using a benchmarking dataset and a biological cohort. CRANE is also compared with the in-built denoising option of OpenSWATH (OSW1). Denoising files with CRANE identifies more peptides and proteins when fewer MS2 windows are used. This is most likely due to the increase in MS2 noise with a wider isolation window width. Furthermore, OpenSWATH is heavily dependent on MS2 scores. Denoising with CRANE further promotes identification of low abundance peptides. This is encouraging as low abundance peptides appear close to the noise level and it is challenging for any denoising technique to remove noise without affecting them.

The OSW1 background subtraction results in fewer identified peptides and proteins when data from a TTOF 6600 is used. A great proportion of the lost peptides is represented by low abundance ions. This could be due to OpenSWATH using a spatial domain approximation of the background that is based on the intensities of the edges of the peaks. This simplistic approximation does not seem to work well for low abundance peptides. In contrast, CRANE removes baseline, random and chemical noise by manipulating wavelet coefficients.

On quantification alone, OSW1 performs better than CRANE while CRANE is superior in identifications. The benefits of both CRANE and OSW1 can be harnessed by using the methods in conjunction. However, irrespective of whether OSW1 is applied to raw data or to CRANE denoised data, it results in higher CV and reduced low abundance identifications on TripleTOF 6600 data. Denoising with CRANE results in a simultaneous increase in the number of identifications and quantitative accuracy while maintaining a low CV for both TripleTOF 5600 and TripleTOF 6600 instruments and all acquisition window schemes tested.

Acknowledgements

The authors thank the other members of the ProCan® team for support, and Roger Reddel, Phillip J. Robinson, Rosemary Balleine, Rohan Shah, Stefan Tenzer and Hannes Röst for feedback, comments and assistance with the manuscript.

Funding

This work was supported by Cancer Council NSW (IG-18-01 to B.T. and Q.Z.). ProCan® is supported by the Australian Cancer Research Foundation, Cancer Institute New South Wales (NSW) (2017/TPG001, REG171150); NSW Ministry of Health (CMP-01); The University of Sydney, Cancer Council NSW (IG 18-01); Ian Potter Foundation, the Medical Research Futures Fund (MRFF-PD), National Health and Medical Research Council (NHMRC) of Australia European Union grant (GNT1170739, a companion grant to support the European Commission's Horizon 2020 Program, H2020-SC1-DTH-2018-1, 'iPC—individualizedPaediatricCure' [ref. 826121]); and National Breast Cancer Foundation (IIRS-18-164). This work was done under the auspices of a Memorandum of Understanding between Children's Medical Research Institute and the U.S. National Cancer Institute's International Cancer Proteogenome Consortium (ICPC).

Conflict of Interest: none declared.

References

Aebersold,R. and Mann,M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature*, **537**, 347–355.

Awan,M.G. and Saeed,F. (2016) MS-REDUCE: an ultrafast technique for reduction of big mass spectrometry data for high-throughput processing. *Bioinformatics*, **32**, 1518–1526.

Bludau,I. and Aebersold,R. (2020) Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Nat. Rev. Mol. Cell Biol.*, **21**, 327–340.

Cai,C. and Harrington,PdB. (1998) Different discrete wavelet transforms applied to denoising analytical data. *J. Chem. Inform. Comput. Sci.*, **38**, 1161–1170.

Candès,E.J. (1998) Ridgelets: theory and applications. PhD Thesis, Stanford University.

Cappadona,S. *et al.* (2008) Wavelet-based method for noise characterization and rejection in high-performance liquid chromatography coupled to mass spectrometry. *Anal. Chem.*, **80**, 4960–4968.

Chernushevich,I.V. *et al.* (2001) An introduction to quadrupole–time-of-flight mass spectrometry. *J. Mass Spectrom.*, **36**, 849–865.

Chipman,H.A. *et al.* (1997) Adaptive Bayesian wavelet shrinkage. *J. Am. Stat. Assoc.*, **92**, 1413–1421.

Chou,S.-W. *et al.* (2012) Wavelet-based method for time-domain noise analysis and reduction in a frequency-scan ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.*, **23**, 1855–1864.

Coombes,K.R. *et al.* (2005) Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, **5**, 4107–4117.

Ding,J. *et al.* (2009) A novel approach to denoising ion trap tandem mass spectra. *Proteome Sci.*, **7**, 9.

Donoho,D.L. and Johnstone,I.M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Donoho,D.L. and Johnstone,I.M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.*, **90**, 1200–1224.

Du,P. *et al.* (2008) A noise model for mass spectrometry based proteomics. *Bioinformatics*, **24**, 1070–1077.

Ellis,M.J. and Perou,C.M. (2013) The genomic landscape of breast cancer as a therapeutic roadmap. *Cancer Discov.*, **3**, 27–34.

Fadili,J. and Starck,J.-L. (2012) Curvelets and ridgelets. In: Meyers,R.A. (ed.), *Computational Complexity: Theory, Techniques, and Applications*. Springer, New York, NY, pp. 754–773.

Hundertmark,C. *et al.* (2009) MS-specific noise model reveals the potential of iTRAQ in quantitative proteomics. *Bioinformatics*, **25**, 1004–1011.

Kong,L. *et al.* (2020) Multi-omics analysis based on integrated genomics, epigenomics and transcriptomics in pancreatic cancer. *Epigenomics*, **12**, 507–524.

Krasny,L. *et al.* (2018) SWATH mass spectrometry as a tool for quantitative profiling of the matrisome. *J. Proteom.*, **189**, 11–22.

Kwon,D. *et al.* (2008) A novel wavelet-based thresholding method for the pre-processing of mass spectrometry data that accounts for heterogeneous noise. *Proteomics*, **8**, 3019–3029.

Lee,G.R. *et al.* (2019) PyWavelets: a Python package for wavelet analysis. *J. Open Source Softw.*, **4**, 1237.

Li,X. *et al.* (2007) A wavelet-based data pre-processing analysis approach in mass spectrometry. *Comput. Biol. Med.*, **37**, 509–516.

Mujezinovic,N. *et al.* (2006) Cleaning of raw peptide MS/MS spectra: improved protein identification following deconvolution of multiply

charged peaks, isotope clusters, and removal of background noise. *Proteomics*, **6**, 5117–5131.

Mujezinovic,N. *et al.* (2010) Reducing the haystack to find the needle: improved protein identification after fast elimination of non-interpretable peptide MS/MS spectra and noise reduction. *BMC Genomics*, **11** (**Suppl. 1**), S13.

Navarro,P. *et al.* (2016) A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.*, **34**, 1130–1136.

Ning,X. *et al.* (2014) Chromatogram baseline estimation and denoising using sparsity (BEADS). *Chemometr. Intell. Lab. Syst.*, **139**, 156–167.

Renard,B.Y. *et al.* (2009) When less can yield more - Computational preprocessing of MS/MS spectra for peptide identification. *Proteomics*, **9**, 4978–4984.

Rosenberger,G. *et al.* (2017) Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods*, **14**, 921–927.

Röst,H.L. *et al.* (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.*, **32**, 219–223.

Schneider,L.V. (2016) Mass spectral data processing. Tech. Rep., Veritomyx. 10.13140/RG.2.2.26279.75684.

Shao,W. and Lam,H. (2013) Denoising peptide tandem mass spectra for spectral libraries: a Bayesian approach. *J. Proteome Res.*, **12**, 3223–3232.

Smith,R. *et al.* (2014) Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. *BMC Bioinformatics*, **15** (**Suppl. 7**), S9.

Starck,J.-L. *et al.* (2007) The undecimated wavelet decomposition and its reconstruction. *IEEE Trans. Image Process.*, **16**, 297–309.

Tully,B. (2020) Toffee – a highly efficient, lossless file format for DIA-MS. *Sci. Rep.*, **10**, 8939.

Yang,C. and Yu,W. (2011) A brief review of signal processing issues in mass spectrometry-based proteomics studies. In: Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), California, USA. pp. 1036–1040.

Yang,X. and Fei,B. (2011) A wavelet multiscale denoising algorithm for magnetic resonance (MR) images. *Meas. Sci. Technol.*, **22**, 25803–25803.

Zhang,J. *et al.* (2008) PeakSelect: preprocessing tandem mass spectra for better peptide identification. *Rapid Commun. Mass Spectrom.*, **22**, 1203–1212.