International Journal of
Molecular Sciences

MDPI

Article

Signal Deconvolution and Noise Factor Analysis Based on a Combination of Time–Frequency Analysis and Probabilistic Sparse Matrix Factorization

Shunji Yamada [1,2], Atsushi Kurotani [2], Eisuke Chikayama [2,3] and Jun Kikuchi [1,2,4,*]

[1] Graduate School of Bioagricultural Sciences, Nagoya University, Furo-cho, Nagoya 464-8601, Chikusa-ku, Japan; shunji.yamada@riken.jp

[2] RIKEN Center for Sustainable Resource Science, 1-7-22 Suehiro-cho, Yokohama 230-0045, Tsurumi-ku, Japan; atsushi.kurotani@riken.jp (A.K.); chikaya@nuis.ac.jp (E.C.)

[3] Department of Information Systems, Niigata University of International and Information Studies, 3-1-1 Mizukino, Niigata 950-2292, Nishi-ku, Japan

[4] Graduate School of Medical Life Science, Yokohama City University, 1-7-29 Suehiro-cho, Yokohama 230-0045, Tsurumi-ku, Japan

* Correspondence: jun.kikuchi@riken.jp; +81-45-508-9439

Received: 27 March 2020; Accepted: 19 April 2020; Published: 21 April 2020

Abstract: Nuclear magnetic resonance (NMR) spectroscopy is commonly used to characterize molecular complexity because it produces informative atomic-resolution data on the chemical structure and molecular mobility of samples non-invasively by means of various acquisition parameters and pulse programs. However, analyzing the accumulated NMR data of mixtures is challenging due to noise and signal overlap. Therefore, data-cleansing steps, such as quality checking, noise reduction, and signal deconvolution, are important processes before spectrum analysis. Here, we have developed an NMR measurement informatics tool for data cleansing that combines short-time Fourier transform (STFT; a time–frequency analytical method) and probabilistic sparse matrix factorization (PSMF) for signal deconvolution and noise factor analysis. Our tool can be applied to the original free induction decay (FID) signals of a one-dimensional NMR spectrum. We show that the signal deconvolution method reduces the noise of FID signals, increasing the signal-to-noise ratio (SNR) about tenfold, and its application to diffusion-edited spectra allows signals of macromolecules and unsuppressed small molecules to be separated by the length of the T_2* relaxation time. Noise factor analysis of NMR datasets identified correlations between SNR and acquisition parameters, identifying major experimental factors that can lower SNR.

Keywords: NMR; molecular complexity; FID; short-time Fourier transform; matrix factorization; T_2* relaxation time; diffusion-edited spectrum; signal-to-noise ratio; acquisition parameters; correlation network analysis

1. Introduction

NMR spectroscopy is one of the most powerful tools available for molecular characterization at the atomic level [1]. Because it is non-invasive, NMR has been applied to data-driven analyses of molecular complexity in many areas of health [2], food [3], materials [4], and the environment [5]. In measuring NMR signals, the main challenges are the sensitivity and resolution of the NMR spectrum [6]. On the one hand, various techniques and devices for improving sensitivity have been developed, such as high-field magnets [7], cryogenic detection systems [8], shimming and locking to adjust the magnetic field [9], and dynamic nuclear polarization [10]. In addition, pulsed field gradient (PFG), nonuniform sampling [11] and magnetization transfer techniques such as cross-polarization [12] and INEPT (Insensitive nuclei enhanced by polarization transfer) [13] have been developed to enhance the sensitivity per unit time. On the other hand, compact and benchtop NMR instruments with lower resolution have become highly cost-effective owing to marked progress in the materials used for the permanent magnet [14].

Regarding spectral resolution, many pulse sequences for the measurement of one-dimensional (1D)-NMR with selective signal suppression, including pre-saturation, Carr–Purcell–Meiboom–Gill (CPMG) [15], WATER suppression by GrAdient Tailored Excitation (WATERGATE) [16], diffusion-editing [17], double quantum filter [18], and pure shift NMR [19], have been developed to reduce signal overlap. However, the spectra have remaining overlapping signals, or the overlapping peaks themselves contain part of the information of the sample. In this regard, overlapping signals can be separated by two-dimensional (2D)-NMR, in which multiple free induction decays (FIDs) are measured over a small change in evolution time, but this approach is time consuming [20].

Conventionally, methods for improving the sensitivity and resolution of FIDs are adjusted by pre-processing steps, such as zero filling and apodization, before Fourier transformation (FT) is carried out [21]. Other methods for reducing mathematical noise from FID signals focus on the region of interest (ROI), such as reference deconvolution [22], harmonic inversion noise removal (HINR) [23], and complete reduction to amplitude frequency table (CRAFT) [24]. In addition, STFT and wavelet transform [25] have been developed as alternative transformation methods to FT for analyzing the relationship between the time and frequency of FIDs. In principle, the exponential decay constant of the FID obtained by applying a 90° pulse to create transverse magnetization is the T_2 relaxation time, a physical parameter independent of field inhomogeneity. In reality, however, because of the effect of magnetic field inhomogeneity, the decay constant of the FID is defined as T_2*, an instrument-dependent parameter, rather than T_2. STFT has the ability to extract time-varying behavior from FIDs, allowing for the analysis of dynamic chemical shifts of atoms in flexible proteins [26]. In addition, it has been reported that STFT can extract T_2* information from FIDs and improve the results of discriminant analysis [27]. Applying the same idea to covariance NMR [28], T_2*-weighted covariance NMR improves the sensitivity and resolution of signals based on the difference in T_2*, determined by dividing each FID in the t_1 dimension of 2D-NMR to create a series of sub-FIDs [29]. In an alternative approach, matrix factorization (MF) is commonly used to extract signal components and separate peaks in spectra [30]. For example, a noise reduction method using principal component analysis (PCA), which is one of the most commonly used multivariate analysis methods for extracting features of data, has been applied to solid CP-MAS NMR data measured by various parameters [31]. Therefore, the quality and amount of information from FIDs can be maximized by applying corrections based on different characteristics. Nevertheless, all these methods require multiple FIDs obtained by adding either spectral dimensions or multiple conditions of samples or parameters. There is also a computational approach such as CORE (COmponent-REsolved; a multi-component spectral separation approach previously introduced method). It focuses on diffusion coefficients to separate the NMR signals of different compounds in PFG-NMR [32–34]. However, this technique requires a specific NMR probe with a coil for generating PFG.

In the current move toward a digital innovation society, tools for NMR measurement informatics are becoming increasingly important [35]. Alongside this, the value of raw NMR datasets for reuse in research studies is rising [36]. Although the quality of raw data influences the value of knowledge obtained in terms of both insight and prediction [37], data cleansing methods for utilizing various kinds of NMR data accumulated over many years, such as data quality checks, noise reduction, and signal deconvolution, have not been established.

In this study, by focusing on acquisition parameters [38–42] and noise [25], we have developed an NMR measurement informatics tool for data cleansing based on FID signal deconvolution and noise factor analysis. Our method for deconvoluting signals and noise factor analysis can be applied to original single FIDs from 1D-NMR and is based on STFT [43] and PMSF [44]. It differs from conventional noise reduction using multivariate analysis [34] because it does not require multiple

1D-NMR data that are measured on many samples or acquired with several acquisition parameters. The difference in T_2* on the time axis determined by performing STFT for each frequency component is useful to separate signals based on MF instead of ROI [22–24]. Our method that focuses on the relaxation time utilizes the attenuation behavior of the FID signal without any hardware upgrade for NMR research field. Lastly, we have developed a function for collecting acquisition parameters as a measurement of experimental factors from a directory of NMR data, and investigated the relationship between signal-to-noise ratio (SNR) and acquisition parameters. A researcher performing NMR must select parameters for each experiment, and normally chooses a reasonable set of parameters based on their experience. We show that these parameters can be characterized in terms of their correlation with SNR by a statistical analysis of accumulated NMR datasets. Therefore, this method will be useful to determine the optimal conditions of acquisition parameters.

2. Results and Discussion

2.1. Signal Deconvolution Method

In this study, signal deconvolution, based on the combined method of STFT and PSMF, was applied to FIDs of 1D-NMR to separate the components and improve SNR. The theory behind the signal deconvolution method is described in detail in the Supplementary Materials. In brief, in FT NMR spectroscopy, the FID is the NMR signal generated by non-equilibrium nuclear spin magnetization processing along the magnetic field. In general, this non-equilibrium magnetization can be generated by applying a pulse of resonant radiofrequency close to the Larmor frequency of the nuclear spins of the sample. Each FID is commonly a sum of multiple decayed oscillatory signals. These signals return to equilibrium at different rates or relaxation time constants. Thus, analysis of the relaxation times of an FID for a sample gives significant insight into the chemical composition, structure, and mobility of the sample. FIDs acquired by NMR measurement are composed of many signals derived from the sample, in addition to several types of noise, such as external noise, physical vibration, power supply, and internal noise from the spectrometer due to thermal noise. Therefore, an FID can be modeled as:

$$S(t) = S_{signal}(t) + S_{noise}(t) \tag{1}$$

where $S(t)$ is the measured signal, and $S_{signal}(t)$ and $S_{noise}(t)$ are sets of ideal signals and signals from different types of noise, respectively (Equation (1) and Supplementary Equation (S1)) [45]. The relaxation process can then be described as the exponential decay of the transverse magnetization $S(t)$ (Supplementary Equation (S2)) [46]. The shorter the relaxation time T_2*, the more rapid the decay. If an FID has more than one component, it will be the sum of contributions from each component (Supplementary Equation (S3)).

Whereas standard FT (Supplementary Equation (S4)) contains only the frequency domain, STFT contains both frequency and time domains. Because the FID signal decays exponentially with time, for STFT, it needs to be divided into several small time intervals (segments) to analyze the time–frequency feature accurately, and FT is used to determine the frequency feature of each segment, thereby increasing the accuracy of signal feature extraction. STFT uses a window function to obtain each weighted segment on the time axis, and then applies FT to each segment. STFT of $S(t)$ can be written as:

$$STFT_S(\tau, \omega) = \int_{-\infty}^{\infty} S(t) g(t - \tau) \exp(-i\omega t) dt \tag{2}$$

where the window function g is first used to intercept the progress of FT on $S(t)$ around $t = \tau$ locally, and then FT of the segment is performed on t (Equation (2) and Supplementary Equation (S5)) [43]. By moving the center position of the window function g sequentially, all the FTs at different times can be obtained.

$STFT_S(\tau, \omega)$ is a complex-valued function (Supplementary Equations (S6)–(S9)) composed of two types of signal: real (*Re*, Supplementary Equation (S7)) and imaginary (*Im*, Supplementary

Equation (S8)), whose phases differ from each other by 90° (Supplementary Figure S1). To change the complex value into an absolute value, the following equation is applied:

$$|z| = \sqrt{Re^2 + Im^2} = \sqrt{\left(\gamma \cos \omega t \exp\left(-\frac{t}{T_2^*}\right)\right)^2 + \left(\gamma \sin \omega t \exp\left(-\frac{t}{T_2^*}\right)\right)^2} \quad (3)$$

For the matrix factorization method PSMF [47], positive-valued matrices are needed, and the original signal values must be converted to their logarithmic form for optimal analysis. To convert the absolute value in Equation (3) to a positive logarithmic form, the following Equation (4) (Supplementary Equation (S10)) is applied:

$$V = \log_{10}(|z| + 1) \quad (4)$$

Signal deconvolution can be then formulated as finding the factorization of the data matrix V (Supplementary Equations (S11) and (S12)):

$$V = W \cdot H + residuals = W_{signal} \cdot H_{signal} + W_{noise} \cdot H_{noise} + residuals \quad (5)$$

In this method using PSMF, we focus on sparse factorizations and on properly accounting for uncertainties while computing the factorization. Equation (5) estimates that the signal component ($W_{signal} \cdot H_{signal}$) decays exponentially with time, while the noise component ($W_{noise} \cdot H_{noise}$) is a random or flat value. To reconstruct the FIDs, the absolute value within each component is converted back to a complex value (Supplementary Equations (S13) and (S14)). The inverse STFT is computed by overlap-adding the inverse fast FT signals in each segment of the STFT spectrogram (Supplementary Equation (S15)).

To evaluate SNR, both noise-removed and noise-only FIDs are converted to signal and noise spectra, respectively, by applying standard FT. SNR is calculated as the ratio of the signal peak intensity to the noise value by using the method of Mnova (Supplementary Equation (S16)) [48]. The noise value is calculated by using the standard deviation of the signals-free region (Supplementary Equation (S17)). Finally, the relative SNR is the ratio of the SNR after denoising ($SNR_{denoised}$) to the original SNR ($SNR_{original}$), which is calculated as follows (Equation (6) and Supplementary Equation (S18)):

$$Relative\ SNR = \frac{SNR_{denoised}}{SNR_{original}} \quad (6)$$

Figure 1 shows an example of application of our signal deconvolution process to sucrose ^{1}H-NMR. STFT of the original FID adds a time axis to the frequency axis of the conventional FT spectrum (Figure 1a). The STFT spectrogram is three-dimensional, showing the frequency, time, and intensity of signal and noise. The matrix of the spectrogram was separated into signal and noise components based on the patterns of relaxation time using PSMF (Figure 1b). Each component was then converted into a signal FID and time-domain noise data by using inverse STFT (Figure 1c). Lastly, the time-region data were converted into the denoised spectrum and noise by using standard FT (Figure 1d). Regarding the noise reduction of the sucrose data, SNR of the denoised spectrum was improved about tenfold relative to the original data. In other words, for the sucrose sample, a 100-fold longer acquisition time would be required to obtain the same SNR without denoising. We compared signal and spectral quality between the original FT and noise reduction data (Supplementary Figure S2 and Table S1). There was almost no difference between them.

In STFT, the size of a window function $g(t-\tau)$ is important. We define the percentage of the time width as the percentage of the window size to FID length. After examining different percentages of the time widths, we found that signal components could be properly extracted in 1.5% and 3.1% (512 and 1024 points for 33,280 points), but not 6.2% (2048 points for 33,280 points) (Supplementary Figure S3). This is because the larger time width does not improve spectra since STFT becomes standard FT. Consequently, the percentage against the effective average region of FIDs is important for this method.

Based on this result, the percentage of the time width was set to 3.1% for data analyzed in Figure 1. When using this method for data with short effective regions (fast relaxation systems such as solid-state NMR and quadrupole nucleus), data processing must be adjusted to maintain the shorter percentage of the time width. In addition, if an FID consists of a number of signals with differing T_2*, it will not be possible to choose an optimal filter for all lines simultaneously by applying commonly used apodization. The apodization such as exponential filtering decreases both signal and noise. In contrast, the method that we propose enables signal and noise to be extracted from an FID based on each pattern of T_2* relaxation time.

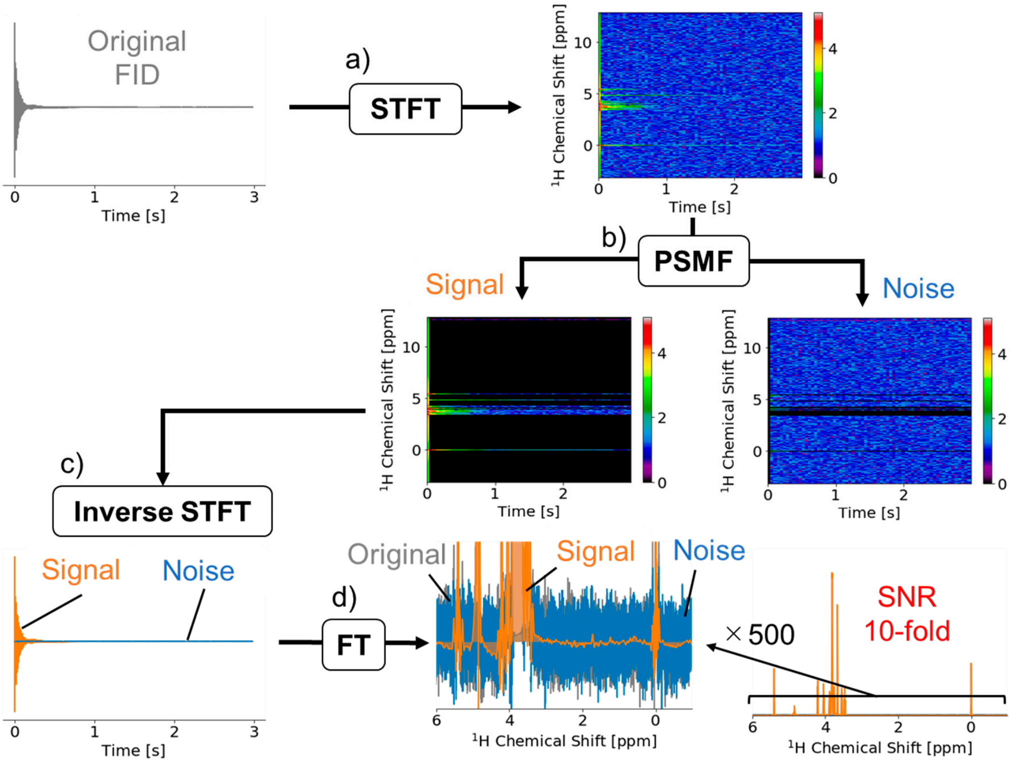

Figure 1. The free induction decay (FID) signal deconvolution method and its application to ^{1}H-NMR data for sucrose. (**a**) The spectrogram was obtained by applying short-time Fourier transform (STFT) to the original FID. (**b**) The matrix obtained after STFT was applied to probabilistic sparse matrix factorization (PSMF), which separated it into signal and noise components. (**c**) The signal and noise components were converted into a noise-removed FID signal (orange) and a time-domain noise signal (blue) by using inverse short-time Fourier transform. (**d**) Finally, the noise-removed FID and the time-domain noise signal were converted to a frequency-domain spectrum by applying standard Fourier transform. As compared with the original FID, the signal-to-noise ratio of the denoised FID was improved about tenfold.

We compared the performance of PSMF with that of three other MF methods, namely standard nonnegative matrix factorization (NMF), sparse nonnegative matrix factorization (SNMF), and probabilistic nonnegative matrix factorization (PMF) (Figure 2). For PSMF, the noise region was successfully removed from the signal component (Figure 2a). For the other three methods, by contrast, the noise component remained in the signal component (Figure 2b–d). Regarding the PSMF time-varying coefficients, the signal component attenuated gradually over time, whereas the noise component attenuated sharply in the first segment and then became flat from the second segment (Supplementary Figure S4a). This observation suggests that part of the signal component may be included in the initial stage of the noise component. Therefore, for the optimal result in Figure 1, the initial value of the noise component is added as a signal component. The time-varying coefficients of the other three methods were characterized by containing mostly noisy components in the signal components, suggesting that the signal components were not properly extracted (Supplementary Figure S4b–d). The signal component is theoretically considered to be sparse data that comprise only specific frequency components. PSMF is a method that considers noise and uncertainty under the sparseness constraint, which suggests that it is suitable for removing noise from ^{1}H-NMR data. We also examined the effect of the number of components in PMSF on signal deconvolution, which showed that it was possible to properly extract signal components when there were two components (Supplementary Figure S5). When the number of components was increased, only noise components were separated more finely. Based on this result, the number of components was set to 2 in the signal deconvolution method for noise reduction. In the case of more complex data, such as the NMR signal of a mixture, it may be possible to apply the method to the characterization of multiple components by separating them with an arbitrary number of components.

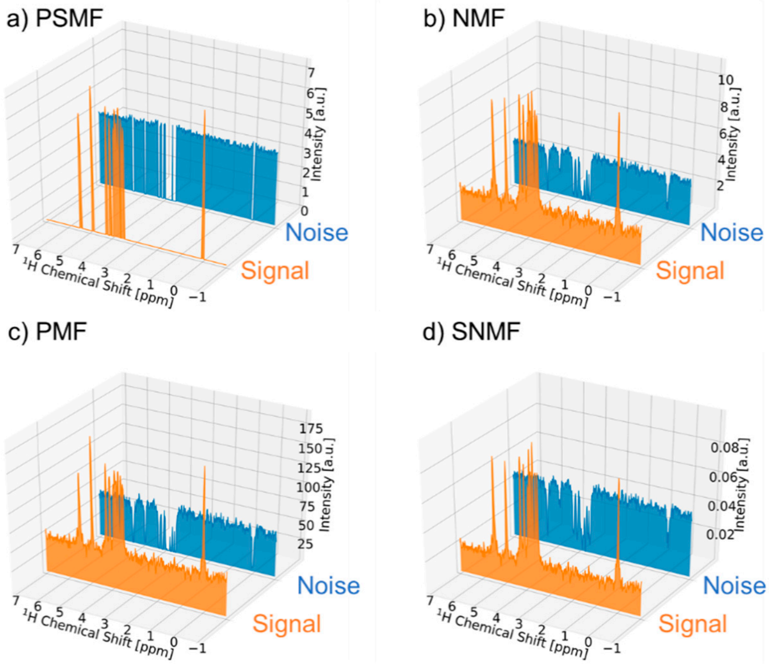

Figure 2. Comparison of four matrix factorization (MF) methods in signal deconvolution. Shown are spectral patterns of signal deconvolution for sucrose ^{1}H-NMR data using (**a**) PSMF, (**b**) NMF, (**c**) PMF, and (**d**) SNMF. The signal components are shown in orange and the noise components are shown in blue.

2.2. Noise Reduction in NMR Data Measured by Various Pulse Sequences

The improvement in the relative SNR achieved by the noise reduction method was investigated by using large-scale data measured by various pulse sequences (Figure 3). Here, we analyzed the following three pulse sequences, which are generally used depending on the target of analysis: CPMG, which detects small molecules with long T_2*, diffusion-edited, which detects proteins and lipids with relatively short T_2*, and WATERGATE, which detects both of these. For the analysis of extensive data, percentages of the time width to FID lengths were set to 6.3% for CPMG and WATERGATE, 12.5% for diffusion-edited (1024 points for 16384 and 8192 points), and the initial three values of the noise component were added as a signal component. For CPMG and WATERGATE, the improvement rate was 3.7-fold and 3.3-fold, respectively. On the other hand, it was only 2.2-fold for diffusion-edited NMR data (Figure 3a). As a result of comparing the relative SNRs of three typical pulse sequences for 10 representative samples, the data of diffusion-edited tended to be lower than those of CPMG and WATERGATE as in the case of large-scale data (Figure 3b, Supplementary Table S2) since the time width for diffusion-edited (12.5%) is higher than that of the other two pulse sequences (6.3%). The SNR of any NMR data set is related to the acquisition parameters (Supplementary Figures S6–8). In NMR data using CPMG and WATERGATE, the SNR is related to several acquisition parameters, such as receiver gain (RG), number of scans (NS), relaxation delay time (D1), spectral width (SW), and offset of the transmitter frequency (O1), whereas in diffusion-edited NMR, the SNR is particularly related to the gradient pulse in the *z*-axis (GPZ). In diffusion-edited NMR, signals from small molecules with long T_2* relaxation times are suppressed. We therefore considered that, if the GPZ setting was insufficient, signals of small molecules would remain, resulting in a difference in relative SNR. As expressed, the peak SNR depends on T_2* because an FID with large T_2* yields a sharp line with higher SNR at the peak [38]. Thus, it seems likely that the diffusion-edited NMR data contain a lot of broad signals derived from macromolecules, resulting in less improvement as compared with CPMG and WATERGATE which have many sharp signals.

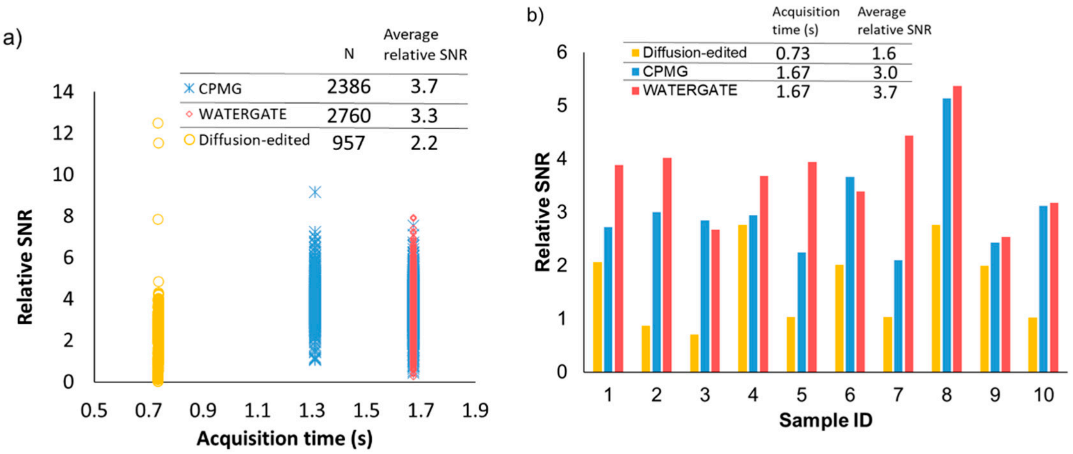

Figure 3. Relative SNR in data measured by three pulse sequences. (**a**) Shown is the relationship between the relative SNR after application of the noise reduction method to large-scale data measured by three pulse sequences: CPMG (blue), WATERGATE (red), and diffusion-edited (yellow), and its acquisition time. The upper part of the figure shows the number of spectra and the average relative SNR for each pulse sequence. (**b**) Comparison of the efficiency for improvement of the SNR measured by three pulse sequences: CPMG (blue), WATERGATE (red), and diffusion-edited (yellow), among NMR spectra derived from sample ID of 1 to 10. The acquisition time and the average relative SNR for each pulse sequence are shown in the upper part of the figure.

2.3. *Application of Signal Deconvolution Method in Diffusion-Edited NMR*

We further examined the application of our signal deconvolution method to diffusion-edited NMR data. For the optimal analysis of these data, the percentage of the time width to FID length was set to 6.3% (512 points for 8192 points), and the initial value of the noise component was added as a signal component. The original FID was separated into three components, including noise and the long and short components of T_2* (Figure 4a,b). By extracting each component and performing standard FT, the SNR of the denoised spectrum was improved about threefold as compared with the original data. In addition, we obtained individual spectra for the short and long components of T_2* (Figure 4c,d). Thus, the diffusion-edited spectrum was separated into signals from macromolecules and small molecules by the length of the T_2* relaxation time. The composition of molecules in these signals is related to the GPZ value of the acquisition parameters (Supplementary Figures S8 and S9). We consider that insufficient GPZ is the main factor affecting the relative SNR of diffusion-edited NMR data because, if GPZ is insufficient, relatively more signals from small molecules are contained in the measured signals. Knowing this composition will help to evaluate the data quality of diffusion-edited NMR.

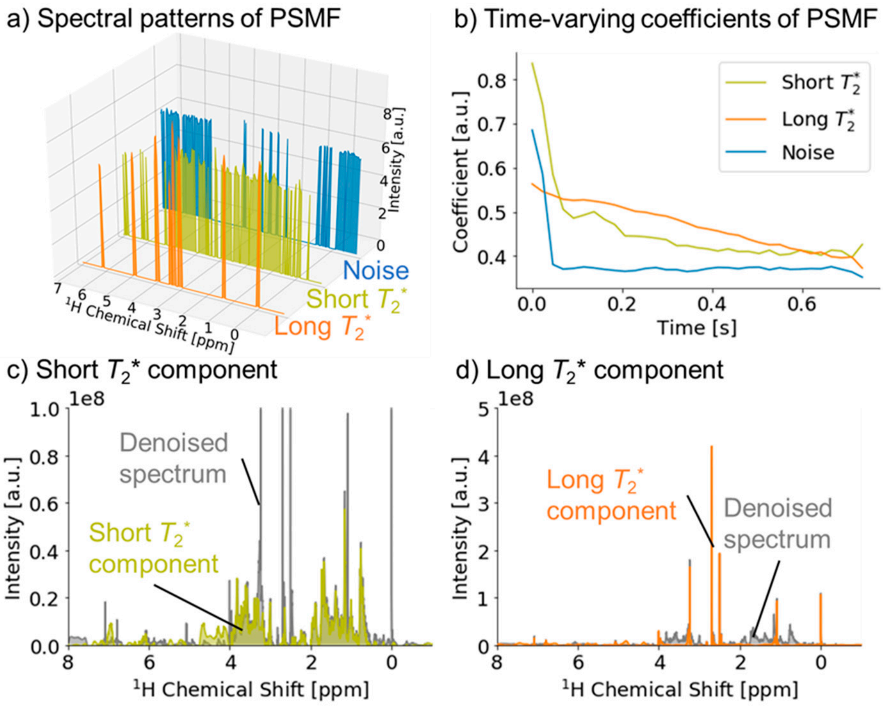

Figure 4. Application of the signal deconvolution method to diffusion-edited spectra. (**a**) Spectral patterns showing signals from small molecules (orange) and macromolecules (olive) separated by the length of the T_2* relaxation time, and noise (blue). (**b**) Time-varying coefficients of each component in MF. (**c**) Denoised spectrum (gray), and spectrum of the short T_2* component (olive). (**d**) Denoised spectrum (gray), and spectrum of the long T_2* component (orange).

2.4. Noise Factor Analysis in Data Measured by Low- and High-Field NMR at Multiple Institutions

To investigate the comprehensive relationship between noise and several acquisition parameters, we analyzed noise factors in data acquired by low- and high-field NMR at multiple institutions. We collected NMR data for four compounds (glucose, sucrose, citric acid, and lactic acid) measured by benchtop NMR (60 MHz) and high-field NMR (500, 600, and 700 MHz) from five institutions/data repositories (RIKEN, NUIS (Niigata University of International and Information Studies), BMRB [49], BML [50], and HMDB [51]) (Supplementary Table S3). The results of correlation analysis between noise and experimental parameters were first summarized as a heatmap (Supplementary Figure S10). With a specific focus on the experimental parameters that affect the SNR, we then derived a network of experimental factors affecting noise based on the correlation coefficients between SNR and experimental parameters (Figure 5). Here, in addition to the SNR calculated using Mnova, we calculated a theoretical SNR value (calcSNR) using a previously described SNR formula (Supplementary Equation (S19)) [52] in order to obtain a theoretical SNR index based on acquisition parameters. Figure 5 shows that, based on the correlation between SNR and, for example, number of scans (NS) and signal intensity (e.g., standard, sample, and solvent), the integration of strong signals will increase noise and reduce SNR. Therefore, the suppression of water signals and sample concentration will be important factors to obtain NMR data with a good SNR.

In situations where longer NMR measurements are needed owing to poor signals (e.g., for nuclei of low sensitivity and/or low natural abundance, and samples of low concentration), paying attention to the certain factors, as discussed here, may provide significant improvements in SNR [38], or even more marked savings in measurement time for a given SNR. For example, too long an acquisition time is not beneficial for SNR. An FID of the time constant T_2* gives, on Fourier transformation, a line width of $1/\pi T_2$* or approximately $1/3T_2$*. Thus, data acquisition beyond about $3T_2$* provides little gain in resolution, but causes a considerable deterioration in SNR. In addition, the spectral width may be set high enough to prevent aliasing of NMR signals. If not, there may be still other signals that fold, namely noise, meaning that the final SNR in the spectrum deteriorates.

Receiving efficiency (*R*) has been proposed as a way to characterize how efficiently the NMR signal can be observed after a unit transverse magnetization in a sample under optimal probe tuning and matching conditions [39]. In that study, the NMR signal amplitude was described as a function of the instrument constant, receiver gain, excitation angle θ, inhomogeneity factor $I(\theta)$, concentration of the observed nucleus, and sample volume. Modern NMR spectrometers require receivers to work within their linear ranges to maintain high-fidelity line shapes and peak integration [40]. The NMR receiver gain is a parameter that is often chosen to maximize SNR. For example, for optimal sensitivity, a dilute analyte needs to be observed with high NMR receiver gain, while the strong, interfering solvent signal must be suppressed [41]. In this case, the dependence of $I(\theta)$ on θ becomes more significant because homogeneity is typically lower for a cryoprobe than for its conventional counterpart [42], and failing to recognize the dependence of $I(\theta)$ on θ alone may potentially lead to errors in quantification as large as 5%. Other factors that we have discussed have less effect on SNR, but are significant in terms of line shape.

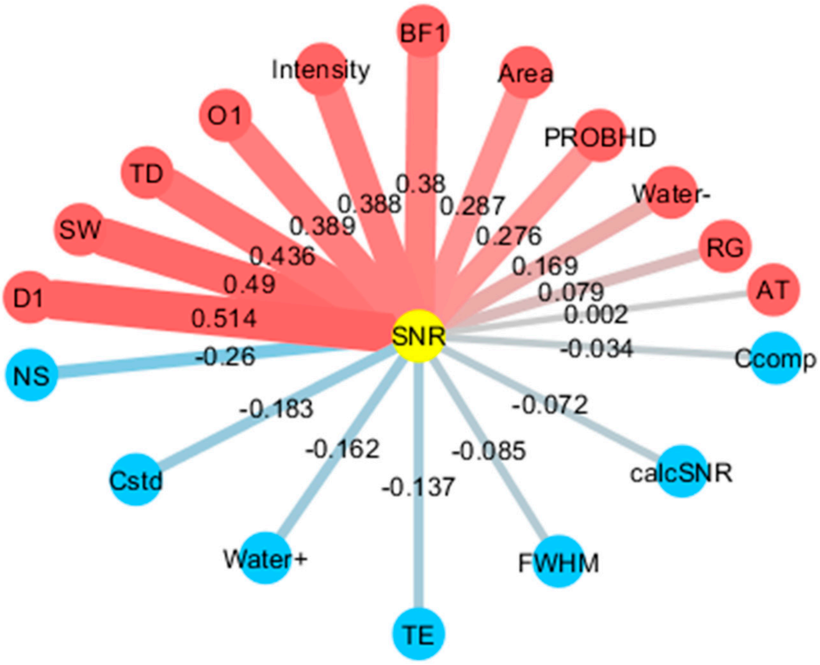

Figure 5. Analysis of experimental factors based on a correlation network of SNR and experimental parameters. The network diagram was drawn by setting positive correlations to red, negative correlations to blue, and the magnitude of the correlation coefficient to the edge thickness. Abbreviations: SNR, signal-to-noise ratio; calcSNR, calculated SNR; Cstd, concentration of standard compound; Ccomp, concentration of compound; Water+, positive intensity of water signal peak to standard peak; Water–, negative intensity of water signal peak to standard peak; Intensity, intensity of standard signal; FWHM, full width at half maximum; Area, area of standard signal; RG, receiver gain; NS, number of scans; D1, relaxation delay time; SW, spectral width; AT, acquisition time; TD, time-domain data size; O1, offset of transmitter frequency; TE, temperature; BF1, basic transmitter frequency for channel F1 in Hertz; PROBHD, if cryoprobe, value is 4, if not, value is 0.

3. Materials and Methods

3.1. Signal Deconvolution Method

The signal deconvolution method was developed in python 3, and built as a graphical user interface (GUI) tool using Tkinter. The tool is available on http://dmar.riken.jp/NMRinformatics/.

The processing of NMR data was implemented by using the nmrglue [53] package in Python. PSMF [47], PMF [54], SNMF [55], and standard NMF [56] were calculated based on the NIMFA Python library for nonnegative matrix factorization [44].

3.2. Noise Factor Analysis Method

The noise factor analysis consisted of four steps implemented in python 3, namely: (1) Collecting acquisition parameters of NMR data: FID and acquisition parameters were searched from the selected NMR data directory and written to CSV files. (2) Calculating SNR: each FID was usually processed to an FT spectrum and denoised spectrum, and the SNR and its improvement ratio were calculated. In the noise factor analysis of data collected from multiple databases, SNR was calculated by using Mnova. (3) Calculating the correlation coefficient between SNR and each parameter by Pearson's correlation coefficient. (4) Visualizing experimental factors: the nodes, edges, and widths of networks based on the correlation coefficient were transformed in GML format by using the Networkx package in Python. Lastly, the network figure was drawn by using Cytoscape [57].

3.3. NMR Data Acquisition

Briefly, ^{1}H-NMR data were by recorded using an Avance II 700 Bruker spectrometer equipped with a 5-mm inverse CryoProbe operating at 700.153 MHz for ^{1}H. In the ^{1}H -NMR data, the number of data using CPMG pulse sequence was 2386, the number of data using WATERGATE pulse sequence was 2760, and the number of data in the 1D LED experiment using bipolar gradients (diffusion-edited) pulse sequence was 975 [58–61]. Regarding these large data sets, a summary of information on the sample and acquisition parameters (the sample title, solvent, acquisition time, acquisition point, and the original SNR) is available at http://dmar.riken.jp/NMRinformatics/. Data sets for comparing the relative SNRs of three typical pulse sequences for 10 representative samples are shown in Supplementary Table S2. To demonstrate the denoising method, data for sucrose and citric acid were acquired by using the presaturation (program name; "zgpr") pulse sequence. To demonstrate the method of separating signals in the diffusion-edited spectrum, ^{1}H-NMR data for fish muscle were measured by a diffusion-edited pulse sequence. Lastly, 48 sets of ^{1}H-NMR data (glucose, sucrose, citric acid, and lactic acid) were collected from the following five sites; RIKEN, NUIS, BMRB, BML, and HMDB. The data were measured with NMR spectrometers of 60, 500, 600, and 700 MHz manufactured by Bruker, Varian, and Nanalysis (Supplementary Table S3).

4. Conclusions

We have developed a measurement informatics tool for NMR signal deconvolution and noise factor analysis and used it to investigate the relationship between noise and acquisition parameters in accumulated NMR datasets. This method enables 1D-NMR spectra to be evaluated with a high SNR, and residual signals from small molecules to be removed from diffusion-edited spectra. This method can be adjustable to any T_2* length, recycle delay, sample molecular weight, or measurement temperature. The percentage of the time width against the effective average signal region of FIDs must be adjusted according to T_2* length. Therefore, when using this method for fast relaxation systems such as solid-state NMR and quadrupole nucleus, additional efforts are needed. In the case of 2D-NMR, it is necessary to use this method by splitting each t_1-dimensional FID and creating a series of sub-FIDs. Noise factor analysis of accumulated NMR datasets might facilitate the investigation of experimental factors related to a lower SNR. Therefore, these methods will help to determine optimal acquisition parameters, to cleanse data, including data management and noise reduction in accumulated NMR datasets, and to promote data-driven studies of molecular complexity using NMR.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/8/2978/s1. An NMR measurement informatics tool for data cleansing and information on the sample and acquisition parameters are available at http://dmar.riken.jp/NMRinformatics/.

Author Contributions: All authors conceived and designed this study. S.Y. implemented a measurement informatics tool with great help from A.K., E.C. and J.K. The manuscript was written through the contributions of all authors. All authors have read and agreed to the published version of the manuscript.

Funding: S.Y. was supported by the RIKEN Junior Research Associate Program and a travel grant from the 8th Asia-Pacific NMR Symposium (APNMR) 2019 during the period of this research. This work was partially supported by a grant from the Agriculture, Forestry, and Fisheries Research Council, as well as Strategic Innovation Program (SIP) from CAO (to J.K.).

Acknowledgments: The authors thank Feifei Wei, Seiji Yoshida, Minoru Fukuchi and Yuuri Tsuboi (RIKEN) for their support with NMR data acquisition.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Takeuchi, K.; Baskaran, K.; Arthanari, H. Structure determination using solution NMR: Is it worth the effort? *J. Magn. Reson.* **2019**, *306*, 195–201. [CrossRef] [PubMed]
2. Jimenez, B.; Holmes, E.; Heude, C.; Tolson, R.F.M.; Harvey, N.; Lodge, S.L.; Chetwynd, A.J.; Cannet, C.; Fang, F.; Pearce, J.T.M.; et al. Quantitative Lipoprotein Subclass and Low Molecular Weight Metabolite Analysis in Human Serum and Plasma by ^{1}H NMR Spectroscopy in a Multilaboratory Trial. *Anal. Chem.* **2018**, *90*, 11962–11971. [CrossRef] [PubMed]
3. Chikayama, E.; Yamashina, R.; Komatsu, K.; Tsuboi, Y.; Sakata, K.; Kikuchi, J.; Sekiyama, Y. FoodPro: A Web-Based Tool for Evaluating Covariance and Correlation NMR Spectra Associated with Food Processes. *Metabolites* **2016**, *6*, 36. [CrossRef] [PubMed]
4. Singh, K.; Kumar, S.P.; Blümich, B. Monitoring the mechanism and kinetics of a transesterification reaction for the biodiesel production with low field ^{1}H NMR spectroscopy. *Fuel* **2019**, *243*, 192–201. [CrossRef]
5. Kikuchi, J.; Ito, K.; Date, Y. Environmental metabolomics with data science for investigating ecosystem homeostasis. *Prog. Nucl. Magn. Reson. Spectrosc.* **2017**, *104*, 56–88. [CrossRef] [PubMed]
6. Wishart, D.S. NMR metabolomics: A look ahead. *J. Magn. Reson.* **2019**, *306*, 155–161. [CrossRef] [PubMed]
7. Maeda, H.; Yanagisawa, Y. Future prospects for NMR magnets: A perspective. *J. Magn. Reson.* **2019**, *306*, 80–85. [CrossRef]
8. Kovacs, H.; Moskau, D.; Spraul, M. Cryogenically cooled probes—A leap in NMR technology. *Prog. Nucl. Magn. Reson. Spectrosc.* **2005**, *46*, 131–155. [CrossRef]
9. Clos, L.J., II; Jofre, M.F.; Ellinger, J.; Westler, W.M.; Markley, J.L. NMRbot: Python scripts enable high-throughput data collection on current Bruker BioSpin NMR spectrometers. *Metabolomics* **2013**, *9*, 558–563. [CrossRef]
10. Ardenkjær-Larsen, J.H.; Fridlund, B.; Gram, A.; Hansson, G.; Hansson, L.; Lerche, M.H.; Servin, R.; Thaning, M.; Golman, K. Increase in signal-to-noise ratio of >10,000 times in liquid-state NMR. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 10158–10163. [CrossRef]
11. Kazimierczuk, K.; Orekhov, V. Non-uniform sampling: Post-Fourier era of NMR data collection and processing. *Magn. Reson. Chem.* **2015**, *53*, 921–926. [CrossRef] [PubMed]
12. Pines, A.; Gibby, M.G.; Waugh, J.S. Proton-Enhanced Nuclear Induction Spectroscopy. A Method for High Resolution NMR of Dilute Spins in Solids. *J. Chem. Phys.* **1972**, *56*, 1776–1777. [CrossRef]
13. Morris, G.A.; Freeman, R. Enhancement of nuclear magnetic resonance signals by polarization transfer. *J. Am. Chem. Soc.* **1979**, *101*, 760–762. [CrossRef]
14. Blümich, B. Low-field and benchtop NMR. *J. Magn. Reson.* **2019**, *306*, 27–35. [CrossRef]
15. Meiboom, S.; Gill, D. Modified Spin-Echo Method for Measuring Nuclear Relaxation Times. *Rev. Sci. Instrum.* **1958**, *29*, 688. [CrossRef]
16. Piotto, M.; Saudek, V.; Sklenar, V. Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR* **1992**, *2*, 661–665. [CrossRef]
17. Vilén, E.M.; Klinger, M.; Sandström, C. Application of diffusion-edited NMR spectroscopy for selective suppression of water signal in the determination of monomer composition in alginates. *Magn. Reson. Chem.* **2011**, *49*, 584–591. [CrossRef]
18. Chandrakumar, N. Chapter 3 1D Double Quantum Filter NMR Studies. *Annu. Rep. NMR Spectrosc.* **2009**, *67*, 265–329. [CrossRef]

19. Lopez, J.; Cabrera, R.; Maruenda, H. Ultra-Clean Pure Shift ^{1}H-NMR applied to metabolomics profiling. *Sci. Rep.* **2019**, *9*, 6900. [CrossRef]
20. Gouilleux, B.; Rouger, L.; Giraudeau, P. Ultrafast 2D NMR: Methods and Applications. *Annu. Rep. NMR Spectrosc.* **2018**, *93*, 75–144. [CrossRef]
21. Castañar, L.; Poggetto, G.D.; Colbourne, A.; Morris, G.A.; Nilsson, M. The GNAT: A new tool for processing NMR data. *Magn. Reson. Chem.* **2018**, *56*, 546–558. [CrossRef] [PubMed]
22. Morris, G.A.; Barjat, H.; Home, T.J. Reference deconvolution methods. *Prog. Nucl. Magn. Reson. Spectrosc.* **1997**, *31*, 197–257. [CrossRef]
23. Taylor, H.S.; Haiges, R.; Kershaw, A. Increasing Sensitivity in Determining Chemical Shifts in One Dimensional Lorentzian NMR Spectra. *J. Phys. Chem. A* **2013**, *117*, 3319–3331. [CrossRef]
24. Krishnamurthy, K. CRAFT (complete reduction to amplitude frequency table)—Robust and time-efficient Bayesian approach for quantitative mixture analysis by NMR. *Magn. Reson. Chem.* **2013**, *51*, 821–829. [CrossRef] [PubMed]
25. Ibrahim, M.; Pardi, C.; Brown, T.; McDonald, P.J. Active elimination of radio frequency interference for improved signal-to-noise ratio for in-situ NMR experiments in strong magnetic field gradients. *J. Magn. Reson.* **2018**, *287*, 99–109. [CrossRef] [PubMed]
26. Langmead, C.J.; Donald, B.R. Extracting structural information using time-frequency analysis of protein NMR data. In Proceedings of the Fifth Annual International Conference on Computing Machinery, Montreal, QC, Canada, 22–25 April 2001; pp. 164–175.
27. Hirakawa, K.; Koike, K.; Kanawaku, Y.; Moriyama, T.; Sato, N.; Suzuki, T.; Furihata, K.; Ohno, Y. Short-time Fourier Transform of Free Induction Decays for the Analysis of Serum Using Proton Nuclear Magnetic Resonance. *J. Oleo Sci.* **2019**, *68*, 369–378. [CrossRef]
28. Short, T.; Alzapiedi, L.; Brüschweiler, R.; Snyder, D. A covariance NMR toolbox for MATLAB and OCTAVE. *J. Magn. Reson.* **2010**, *209*, 75–78. [CrossRef]
29. Manu, V.; Gopinath, T.; Wang, S.; Veglia, G. T_2* weighted Deconvolution of NMR Spectra: Application to 2D Homonuclear MAS Solid-State NMR of Membrane Proteins. *Sci. Rep.* **2019**, *9*, 8225. [CrossRef]
30. Yamada, S.; Ito, K.; Kurotani, A.; Yamada, Y.; Chikayama, E.; Kikuchi, J. InterSpin: Integrated Supportive Webtools for Low- and High-Field NMR Analyses Toward Molecular Complexity. *ACS Omega* **2019**, *4*, 3361–3369. [CrossRef]
31. Kusaka, Y.; Hasegawa, T.; Kaji, H. Noise Reduction in Solid-State NMR Spectra Using Principal Component Analysis. *J. Phys. Chem. A* **2019**, *123*, 10333–10338. [CrossRef]
32. Stilbs, P. Automated CORE, RECORD, and GRECORD processing of multi-component PGSE NMR diffusometry data. *Eur. Biophys. J.* **2012**, *42*, 25–32. [CrossRef] [PubMed]
33. Stilbs, P. RECORD processing–A robust pathway to component-resolved HR-PGSE NMR diffusometry. *J. Magn. Reson.* **2010**, *207*, 332–336. [CrossRef]
34. Stilbs, P.; Paulsen, K.; Griffiths, P. Global Least-Squares Analysis of Large, Correlated Spectral Data Sets: Application to Component-Resolved FT-PGSE NMR Spectroscopy. *J. Phys. Chem.* **1996**, *100*, 8180–8189. [CrossRef]
35. Kikuchi, J.; Yamada, S. NMR window of molecular complexity showing homeostasis in superorganisms. *Analyst* **2017**, *142*, 4161–4172. [CrossRef] [PubMed]
36. Pupier, M.; Nuzillard, J.-M.; Wist, J.; Schlörer, N.E.; Kuhn, S.; Erdélyi, M.; Steinbeck, C.; Williams, A.; Butts, C.P.; Claridge, T.D.W.; et al. NMReDATA, a standard to report the NMR assignment and parameters of organic compounds. *Magn. Reson. Chem.* **2018**, *56*, 703–715. [CrossRef]
37. Halouska, S.; Powers, R. Negative impact of noise on the principal component analysis of NMR data. *J. Magn. Reson.* **2006**, *178*, 88–95. [CrossRef]
38. Becker, E.D.; Ferretti, J.A.; Gambhir, P.N. Selection of optimum parameters for pulse Fourier transform nuclear magnetic resonance. *Anal. Chem.* **1979**, *51*, 1413–1420. [CrossRef]
39. Mo, H.; Harwood, J.; Zhang, S.; Xue, Y.; Santini, R.; Raftery, D. A quantitative measure of NMR signal receiving efficiency. *J. Magn. Reson.* **2009**, *200*, 239–244. [CrossRef] [PubMed]
40. Mo, H.; Harwood, J.S.; Raftery, D. A quick diagnostic test for NMR receiver gain compression. *Magn. Reson. Chem.* **2010**, *48*, 782–786. [CrossRef]
41. Mo, H.; Harwood, J.S.; Raftery, D. Receiver gain function: The actual NMR receiver gain. *Magn. Reson. Chem.* **2010**, *48*, 235–238. [CrossRef]

42. Mo, H.; Harwood, J.; Raftery, D. NMR quantitation: Influence of RF inhomogeneity. *Magn. Reson. Chem.* **2011**, *49*, 655–658. [CrossRef] [PubMed]
43. Liu, H.; Dong, H.; Ge, J.; Bai, B.; Yuan, Z.; Zhao, Z. Research on a secondary tuning algorithm based on SVD & STFT for FID signal. *Meas. Sci. Technol.* **2016**, *27*, 105006. [CrossRef]
44. Zitnik, M.; Zupan, B. NIMFA: A python library for nonnegative matrix factorization. *J. Mach. Learn. Res.* **2012**, *13*, 849–853.
45. Liu, H.; Dong, H.; Ge, J.; Liu, Z.; Yuan, Z.; Zhu, J.; Zhang, H. A fusion of principal component analysis and singular value decomposition based multivariate denoising algorithm for free induction decay transversal data. *Rev. Sci. Instrum.* **2019**, *90*, 035116. [CrossRef] [PubMed]
46. Keeler, J. *Understanding NMR Spectroscopy*; Appollo—University of Cambridge Repository: Cambridge, UK, 2004. [CrossRef]
47. Dueck, D.; Morris, Q.; Frey, B.J. Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics* **2005**, *21*, 144–151. [CrossRef] [PubMed]
48. Claridge, T. MNova: NMR data processing, analysis, and prediction software. *J. Chem. Inf. Model.* **2009**, *49*, 1136–1137. [CrossRef]
49. Ulrich, E.L.; Akutsu, H.; Doreleijers, J.F.; Harano, Y.; Ioannidis, Y.E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; et al. BioMagResBank. *Nucleic Acids Res.* **2007**, *36*, D402–D408. [CrossRef]
50. Ludwig, C.; Easton, J.; Lodi, A.; Tiziani, S.; Manzoor, S.E.; Southam, A.; Byrne, J.J.; Bishop, L.M.; He, S.; Arvanitis, T.N.; et al. Birmingham Metabolite Library: A publicly accessible database of 1-D ^{1}H and 2-D ^{1}H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics* **2011**, *8*, 8–18. [CrossRef]
51. Wishart, D.S.; Feunang, Y.D.; Marcu, A.; Guo, A.C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617. [CrossRef]
52. Larive, C.K.; Jayawickrama, D.; Őrfi, L. Quantitative Analysis of Peptides with NMR Spectroscopy. *Appl. Spectrosc.* **1997**, *51*, 1531–1536. [CrossRef]
53. Helmus, J.J.; Jaroniec, C.P. Nmrglue: An open source Python package for the analysis of multidimensional NMR data. *J. Biomol. NMR* **2013**, *55*, 355–367. [CrossRef] [PubMed]
54. Laurberg, H.; Christensen, M.G.; Plumbley, M.; Hansen, L.K.; Jensen, S.H. Theorems on Positive Data: On the Uniqueness of NMF. *Comput. Intell. Neurosci.* **2008**, *2008*, 1–9. [CrossRef] [PubMed]
55. Kim, H.; Park, H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* **2007**, *23*, 1495–1502. [CrossRef] [PubMed]
56. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [CrossRef]
57. Demchak, B.; Hull, T.; Reich, M.; Liefeld, T.; Smoot, M.; Ideker, T.; Mesirov, J.P. Cytoscape: The network visualization tool for GenomeSpace workflows. *F1000Research* **2014**, *3*, 151. [CrossRef] [PubMed]
58. Yoshida, S.; Date, Y.; Akama, M.; Kikuchi, J. Comparative metabolomic and ionomic approach for abundant fishes in estuarine environments of Japan. *Sci. Rep.* **2014**, *4*. [CrossRef]
59. Misawa, T.; Wei, F.; Kikuchi, J. Application of Two-Dimensional Nuclear Magnetic Resonance for Signal Enhancement by Spectral Integration Using a Large Data Set of Metabolic Mixtures. *Anal. Chem.* **2016**, *88*, 6130–6134. [CrossRef]
60. Asakura, T.; Sakata, K.; Date, Y.; Kikuchi, J. Regional feature extraction of various fishes based on chemical and microbial variable selection using machine learning. *Anal. Methods* **2018**, *10*, 2160–2168. [CrossRef]
61. Wei, F.; Fukuchi, M.; Ito, K.; Sakata, K.; Asakura, T.; Date, Y.; Kikuchi, J. Large-Scale Evaluation of Major Soluble Macromolecular Components of Fish Muscle from Conventional ^{1}H NMR Spectral Database. *Molecules* **2020**, *25*, 1966. [CrossRef]