Bioinformatics, 34(24), 2018, 4300–4301
doi: 10.1093/bioinformatics/bty581
Advance Access Publication Date: 9 July 2018
Applications Note

Structural bioinformatics

Noise peak filtering in multi-dimensional NMR spectra using convolutional neural networks

Naohiro Kobayashi[1,*], Yoshikazu Hattori[2], Takashi Nagata[3,4], Shoko Shinya[1], Peter Güntert[5,6,7], Chojiro Kojima[8] and Toshimichi Fujiwara[1]

[1]Institute for Protein Research, Osaka University, Osaka 565-0871, Japan, [2]Faculty of Pharmaceutical Sciences, Tokushima Bunri University, Tokushima 770-8514, Japan, [3]Institute of Advanced Energy and [4]Graduate School of Energy Science, Kyoto University, Kyoto 611-0011, Japan, [5]Institute of Biophysical Chemistry, Goethe-University, 60438 Frankfurt am Main, Germany, [6]Department of Chemistry, Tokyo Metropolitan University, 192-0397, Japan, [7]Laboratory of Physical Chemistry, ETH Zürich, 8093 Zürich, Switzerland and [8]College of Engineering Science, Yokohama National University, Yokohama 240-0801, Tokyo, Japan

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on October 31, 2017; revised on May 10, 2018; editorial decision on July 3, 2018; accepted on July 6, 2018

Abstract

Motivation: Multi-dimensional NMR spectra are generally used for NMR signal assignment and structure analysis. There are several programs that can achieve highly automated NMR signal assignments and structure analysis. On the other hand, NMR spectra tend to have a large number of noise peaks even for data acquired with good sample and machine conditions, and it is still difficult to eliminate these noise peaks.

Results: We have developed a method to eliminate noise peaks using convolutional neural networks, implemented in the program package Filt_Robot. The filtering accuracy of Filt_Robot was around 90–95% when applied to 2D and 3D NMR spectra, and the numbers of resulting non-noise peaks were close to those in corresponding manually prepared peaks lists. The filtering can strongly enhance automated NMR spectra analysis.

Availability and implementation: The full package of the program, documents and example data are available from http://bmrbdep.pdbj.org/en/nmr_tool_box/Filt_Robot.html.

Contact: naohiro@protein.osaka-u.ac.jp

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

1 Introduction

The modern techniques for NMR signal assignment and structure determination for small proteins in solution have been established using multi-dimensional spectra with isotope-labeled proteins (Sugiki *et al.*, 2017). There are a number of programs to automate these tasks, among which for instance UNIO and FLYA are capable of both NMR signal assignment and structure calculation (Schmidt and Güntert, 2012; Serrano *et al.*, 2012). Such automated programs normally require peak tables from a number of spectra that are obtained using a spectrum viewer program or some other external tools. The biggest burden for the preparation of peak tables is correctness of the identified peak lists. On the other hand, NMR spectra may contain noise signals, for instance from incorrect phasing and sinc type truncation artifacts. The bulk water signal may give severe baseline distortions near the important $^{1}H^{\alpha}$ and $^{1}H^{\beta}$ signals. Noise tends to be a more serious problem in the case of samples with low protein concentration due to a low signal-to-noise ratio. In the last few years, several new peak picking programs have been released (e.g. Klukowski *et al.*, 2015; Würz and Güntert, 2017). However, the concept of our program differs substantially from these since our tool is designed for strong noise elimination. Here, we present a first approach for fully automated noise filtration using

Table 1. Results of Filt_Robot noise filter applied to bmr16647 and Lamin-G465D spectra

NMR experiment	Number of peaks			Accuracy		
	Initial	HSQC filter	CNN filter	Recall[a] (%)	Precision[b] (%)	*F*-value[c] (%)
bmr16647						
2D ^{1}H–^{15}N HSQC	323	n.a.	87	100.0	90.8	95.2
2D ^{1}H–^{13}C HSQC[d]	1 332	n.a.	334	98.8	73.4	84.8
3D CBCA(CO)NH	1023	871	140	97.1	97.1	98.6
3D HNCACB	723	681	230	97.0	99	98.3
3D HCCH-TOCSY[d]	17 760	4484	845	95.1	85.6	90.1
3D ^{15}N-edited NOESY	2 880	1683	806	99.7	95.4	97.5
3D ^{13}C-edited NOESY[d]	20 243	6935	1997	99.0	84.1	91
Lamin-G465D						
2D ^{1}H–^{15}N HSQC	364	n.a.	173	95.5	98.8	97.2
2D ^{1}H–^{13}C HSQC[d]	2144	n.a.	815	99.4	86.7	92.7
3D CBCA(CO)NH	1862	1441	292	100.0	94.9	97.4
3D HNCACB	1700	1285	490	98.9	92.7	95.7
3D HCCH-TOCSY[d]	35 280	6320	1 290	97.1	85.6	93.8
3D ^{15}N-edited NOESY	5 776	4965	2 497	99.2	97.9	98.5
3D ^{13}C-edited NOESY[d]	54 016	17 056	4 273	96.2	97	96.6

[a]$Recall = TP/(TP + FN)$.

[b]$Precision = TP/(TP \pm FP)$, with FP = number of noise peaks identified as real peaks, FN = eliminated real peaks, TP = correctly identified real peaks.

[c]$F\text{-}value = 2 \times Recall \times Precision/(Recall + Precision)$. See Supplementary Material for the other spectrum data and details.

[d]Acquired for aliphatic region only.

convolutional neural networks (CNN), and we demonstrate its robustness in identifying peaks which can be used for the automated assignment program FLYA to obtain NMR signal assignments and 3D structures with high accuracy.

2 Materials and methods

Convolutional neural networks are now available in many neural networks tool kits. For this study, we have chosen CNTK (Cognitive Neural network Tool Kit) version 2.0 developed by Microsoft: https://www.microsoft.com/en-us/research/publication/an-introduction-to-computational-networks-and-the-computational-network-toolkit/. The input peak data, comprising noise and real peaks, are collected from 2D and 3D spectra as described in the Supplementary Material. The peak positions are detected by searching the maximal data point in a square for 2D or cubic for 3D sub-matrix, by quadratic interpolation over the closest three points in each dimension. The submatrix around the detected peak center in the *x*–*y* (and *y*–*z* for 3D spectra) planes are extracted and interpolated to generate 40×40 images. The data intensities are normalized by linearly transforming the intensity of the center point in the submatrix into 0–127 for negative values and 128–255 for positive values. A CNN training data including 2800 noise and 2800 real peaks was collected from 2D and 3D spectra of uniformly ^{13}C/^{15}N-labeled ubiquitin and extended by rotation, mirroring and sign changes to generate ~58 000 images. The graphical user interface and the network structure of the CNN filter are shown in Figure 1. More details on the preparation of the training data and the network structure can be found in the Supplementary Material. The CNTK script, training/test data and demo toolkits are available from our web-site.

3 Results

Two benchmarks were performed using 2D and 3D spectra for a uniformly isotope labeled SH3 domain, which is available from the BMRB archive with accession code bmr16647 and Lamin-G465D (a mutant of human lamin A, 147a.a.) The 3D peak tables were roughly filtered by applying a position mask based on the peak positions in the automatically prepared peak tables for 2D ^{1}H–^{15}N HSQC and ^{1}H–^{13}C HSQC before applying the CNN filter. The noise filtration of the peak lists for spectra took about 20–30 min on a standard PC. Statistics are listed in Table 1 and Supplementary Tables S1 and S2. The peak lists were submitted to the FLYA algorithm in CYANA version 3.98 for automated signal assignment, followed by structure calculation with CYANA using the chemical shift table from FLYA, the CNN-filtered NOESY peak lists, and backbone dihedral angle restraints from TALOS+ (Shen *et al.*, 2009) as input. The structures are very close to the deposited NMR structure with PDB-ID 2KRS and 1IFR, respectively (Supplementary Figs S11 and S12).

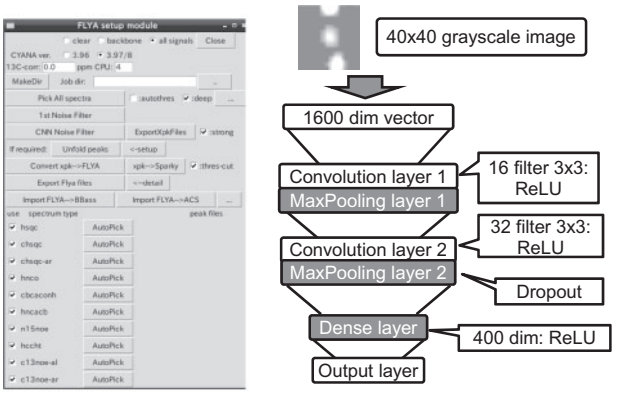

Fig. 1. Filt_Robot module (left) and schematic representation of the CNN filter (right)

In conclusion, our tool can be applied to noise elimination in NMR peak lists for obtaining accurate chemical shifts and NMR structures as well as providing quality factors for the identified peaks. The feasibility of our tool strongly depends on the quality of the sample and the spectrum data. The obtained assignments and structures can be assessed with validation tools such as RPF (Huang *et al.*, 2012).

Funding

This work was supported by the Platform Project for Supporting Drug Discovery and Life Science Research from AMED and JSPS KAKENHI grants 15K06970.

Conflict of Interest: none declared.

References

Huang,Y.J. *et al.* (2012) RPF: a quality assessment tool for protein NMR structures. *Nucleic Acids Res.*, **40**, W542–W546.

Klukowski,P. *et al.* (2015) Computer vision-based automated peak picking applied to protein NMR spectra. *Bioinformatics*, **31**, 2981–2988.

Schmidt,E. and Güntert,P. (2012) A new algorithm for reliable and general NMR resonance assignment. *J. Am. Chem. Soc.*, **134**, 12817–12829.

Serrano,P. *et al.* (2012) The J-UNIO protocol for automated protein structure determination by NMR in solution. *J. Biomol. NMR*, **53**, 341–354.

Shen,Y. *et al.* (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR*, **44**, 213–223.

Sugiki,T. *et al.* (2017) Modern technologies of solution nuclear magnetic resonance spectroscopy for three-dimensional structure determination of proteins open avenues for life scientists. *Comput. Struct. Biotechnol. J.*, **15**, 328–339.

Würz,J.M. and Güntert,P. (2017) Peak picking multidimensional NMR spectra with the contour geometry based algorithm CYPICK. *J. Biomol. NMR*, **67**, 63–76.