PO-CON1857E

SHIMADZU
Excellence in Science

Deep learning methods applied to the analysis of metabolomics data

ASMS 2019 WP 389

Shinji Kanazawa[1,3,4], Yohei Yamada[1], Hiroyuki Yasuda[1], Akihiro Kunisawa[1,3], Toru Shiohama[1], Shigeki Kajihara[1], Norio Mukai[1], Masaki Kakisako[2], Go Fujisawa[2], Yuzuru Yamakage[2], Junko Iida[1,3], Eiichiro Fukusaki[5], Fumio Matsuda[4]
1 Shimadzu Corporation, Kyoto, Japan,
2 Fujitsu Limited, Tokyo, Japan,
3 Osaka University Shimadzu Analytical Innovation Research Laboratory, Osaka University, Osaka, Japan,
4 Graduate School of Information Science and Technology, Osaka University, Osaka,
5 Graduate School of Engineering, Osaka University, Osaka, Japan

Deep learning methods applied to the analysis of metabolomics data

Overview

In this study a deep learning algorithm was applied to automating peak integration for metabolomic samples. The objective was not only to accelerate reliable and dependable peak integration but also to eliminate manual parameters selection.

Introduction

A number of machine learning methods have been applied to bioinformatics and metabolite analyses including self-organizing maps, support vector machines, kernel machines, Bayesian networks or fuzzy logic. Advanced machine learning algorithms have been also applied to in silico MS chromatogram annotation for metabolite identification. As a general approach in peak integration algorithms are designed to determine the start point and end point of a chromatographic peak to enable a calculation of peak area [1, 2].

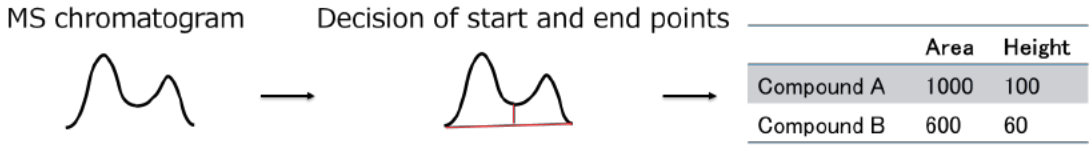


Figure 1. General procedure for peak integration

Metabolomics samples are complex mixtures of compounds representing a diverse array of chemical space and idiosyncratic chromatographic behavior. A simple global model for peak detection and integration in LC-MS/MS methods commonly results in the need to manually change the baseline for several compounds in the sample.

For example, for the test data it takes several hours for an 'expert' operator to manually check, reintegrate and reprocess peak integration for 30 samples each containing 100 compounds (that is 3000 peak integration reviews) on food analysis. Consequently, there is a clear need for automation. The purpose of this study was as follows;

- **to create a unified standard on peak integration**
- **to achieve correct peak integration higher than 90%**

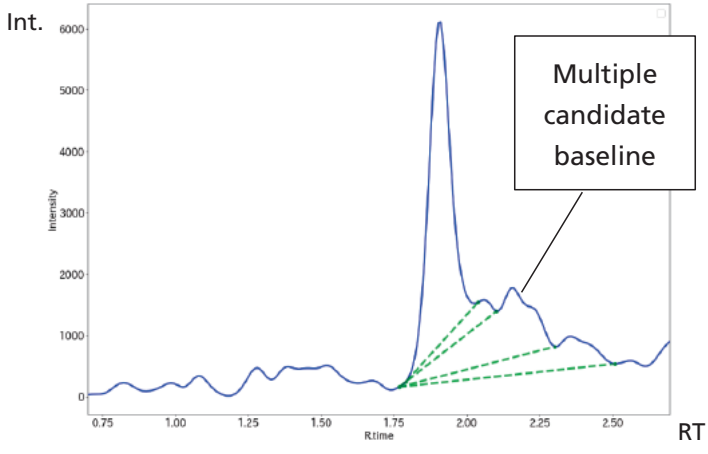


Figure 2. Problem on manual peak integration

Methods

Transformation of input and output

In this study, the task of developing a robust peak integration for idiosyncratic chromatography behavior was formulated as general object recognition and Single Shot MultiBox Detector (SSD [3]) was used as a method for detecting objects.

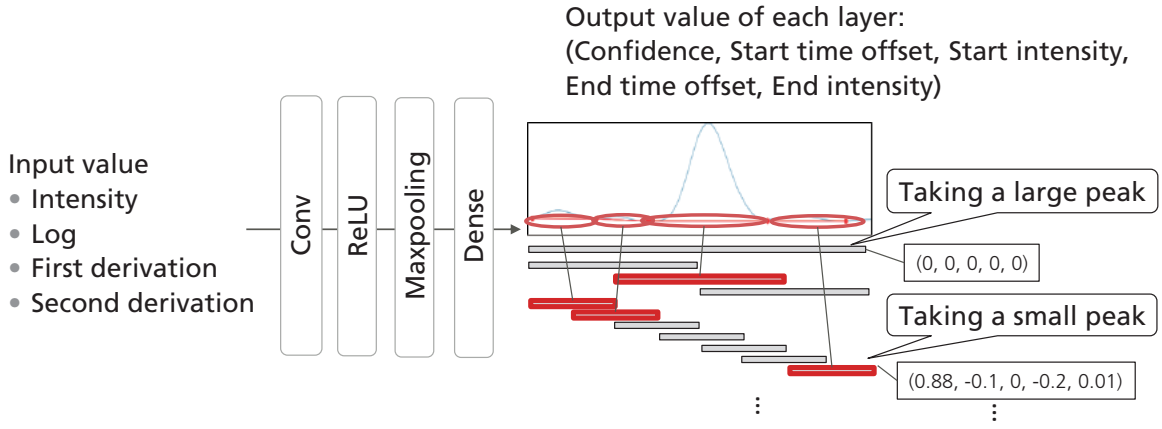


Figure 3. Transformation of input and output

Data augmentation

As a first step, unseparated peaks having known area value were generated. Second, correct peak integration (label) was added. Third, a peak insertion point to a blank chromatogram was determined. Fourth, it is merged with the peak and baseline were merged to generate new chromatogram. Using this approach, a learning pattern was applied to chromatographic behavior.

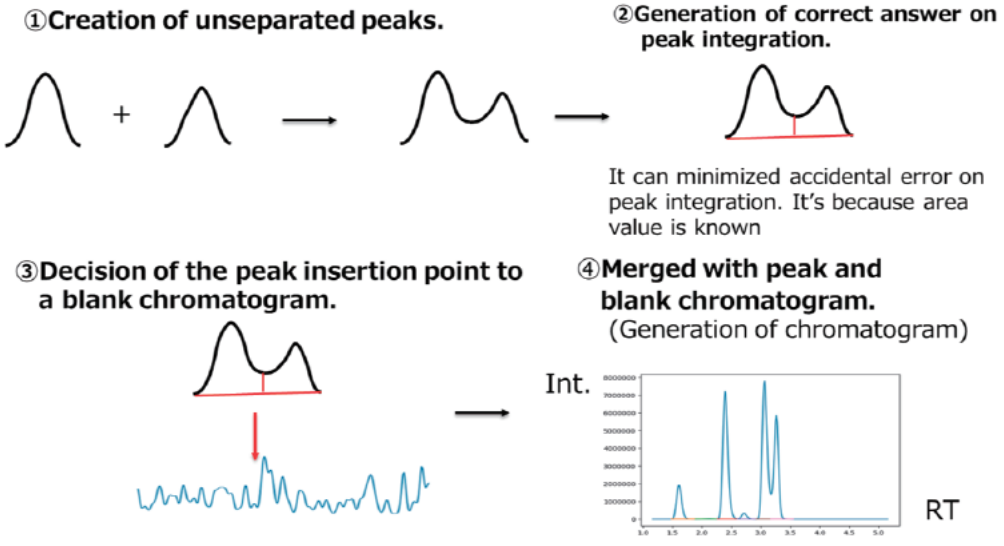


Figure 4. Data augmentation

Evaluation

We defined "true positive" as the predicted and correct peak ranges overlap by 50% or more.

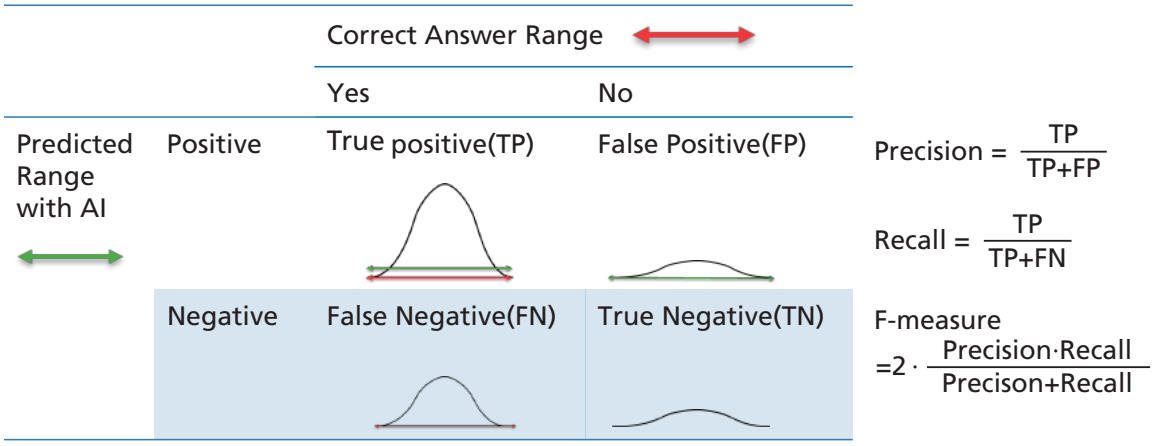


Figure 5. Precision, Recall and F-measure

We also defined the difference between the peak area predicted by the parameter-free deep learning method and the correct peak area given by the expert operator.

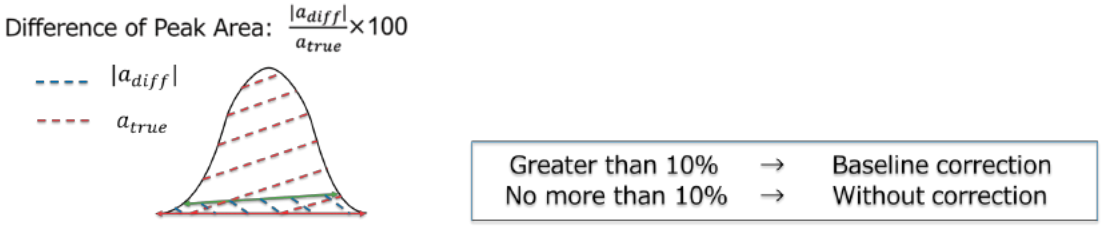


Figure 6. Baseline correction

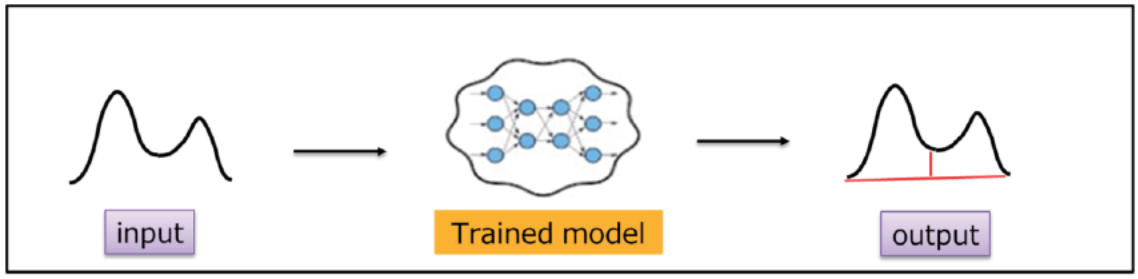
We defined a performance metric "correction time" as $w_1 \cdot FP + w_2 \cdot FN + w_3 \cdot S + w_4 \cdot C$, where S is the number of peaks manual baseline correction were needed, C is the number of matching baseline and w_i (i=1, 2, 3, 4) are the time taken to confirm and correct FP, FN, S and C. The values of FP, FN, S, C and correction time were measured for 8 data sets. From these results, w_i (i=1, 2, 3, 4) were estimated by solving the simultaneous equations. The following results were obtained:

$$\boldsymbol{w_1 = 3.922, w_2 = 21.990, w_3 = 18.533, w_4 = 0.140.}$$

Results

Peak integration using deep learning

The input of the deep learning is a chromatogram which consists of retention time and intensity. The output is the start and end-point of peak integration.



Retention time	Intensity
1.8532	9482
1.8624	9502
1.8712	9632
...	...

Time at the start point	Intensity at the start point	Time at the end point	Intensity at the end point
1.8532	9482	2.21427	7497
1.70527	10296	1.8532	9482
...	...	...	...

Figure 7. Peak integration using deep learning

Training, validation and test sets

To determine major features or combinations of features that can result in robust, dependable peak integration, a training set of primary metabolites including a panel of amino acid, organic acid and nucleotide extracted from a wide variety of samples were used as matrix. Compounds were detected using high performance LC-MS/MS analysis (Shimadzu Corporation). The data was curated by the expert operator and was split into training set of 11,011 data, validation set of 1,400 data, and test set of 1,400 data for each compound. The training set was used to develop and train the model. However, much larger dataset is generally required to train deep learning models. Therefore, we performed the data augmentation to increase training data from 11,011 data to 73,303 data.

Deep learning methods applied to the analysis of metabolomics data

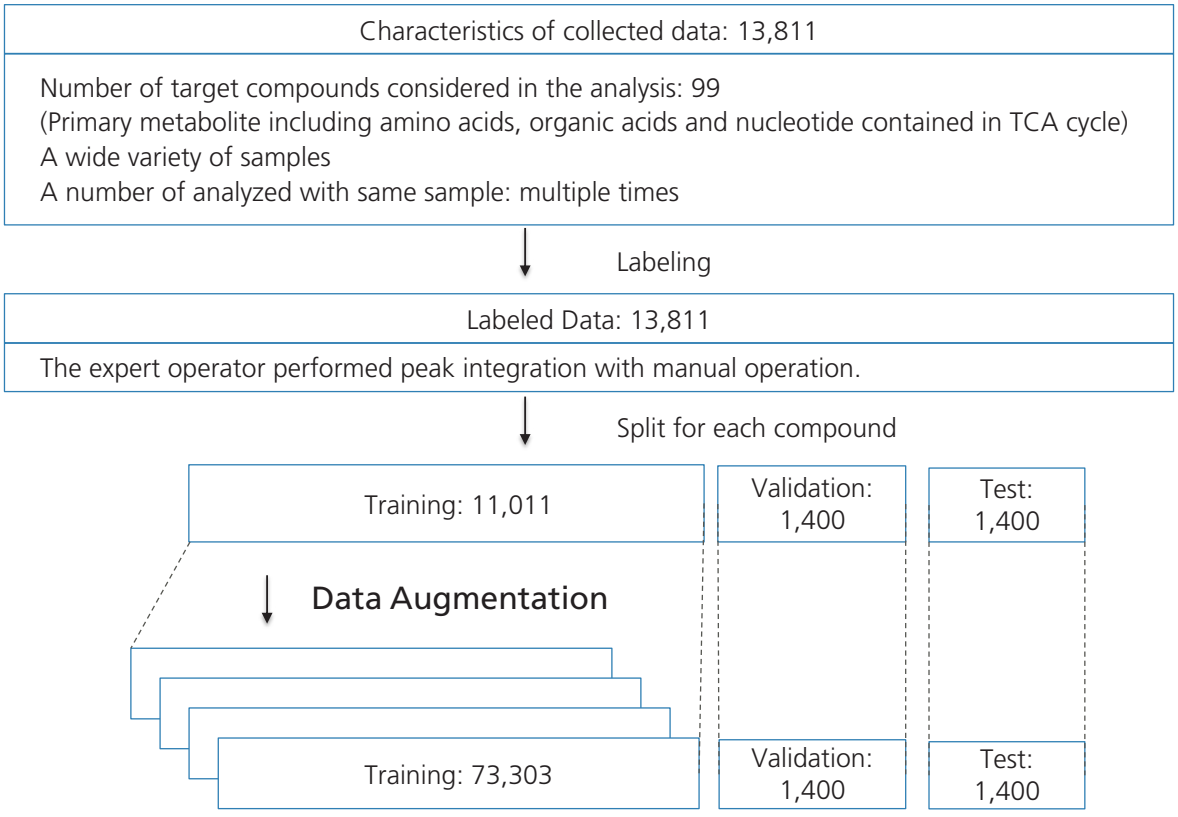


Figure 8. Training, validation and test sets

Evaluation of algorithms

We evaluated precision, recall and F-measure of the deep learning method, i-PeakFinder [1] and Chromatopac™ [2]. The parameter-free deep learning method achieves F-measure of 0.910 (precision of 0.883 and recall of 0.938).

Table 1. Peak detection results

Algorithm name	True positive(TP) + False positive(FP)	True positive(TP) + False negative(FN)	Precision	Recall	F-measure
Deep learning method	1860	1751	**0.883**	**0.938**	0.910
i-PeakFinder [1]*	1469	1751	0.864	0.725	0.788
Chromatopac [2]*	4364	1751	0.349	0.870	0.498

*Peak integration parameters are optimized before comparison for i-PeakFinder and Chromatopac.

Deep learning methods applied to the analysis of metabolomics data

We also calculated the difference between the peak area predicted by the deep learning method and the correct peak area given by the expert operator. The difference of the peak area was used to determine if baseline correction was necessary. Furthermore, the estimated correction time of each algorithm was evaluated on the condition that an expert operator manually integrates peaks. The estimated correction time of proposed method is 43% smaller than i-PeakFinder [1] and 65% smaller than the Chromatopac [2].

Table 2. Estimated correction time results

Algorithm name	FP	FN	TP		Estimated correction time (hour)
			number of corrected baseline	number of matching baseline	
Deep learning method	217	**108**	293	**1350**	**2.46**
i-PeakFinder [1]	**200**	482	**211**	1058	4.29 - 43%
Chromatopac [2]	2842	229	483	1039	7.02 - 65%

Figure 9 shows the result of the expert operator and the result of the AI. They show the parameter-free deep learning method is consistent with peak integration with expert operator.

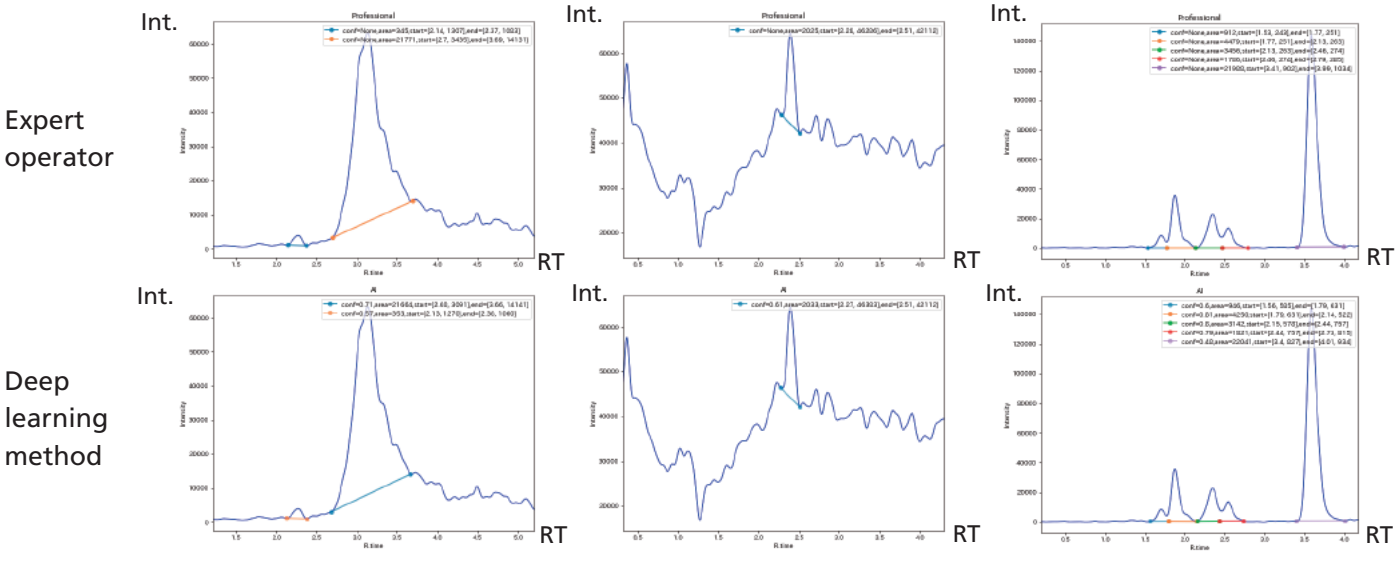


Figure 9. Comparison of the expert operator and deep learning method

Conclusions

- We developed the parameter-free deep learning method.
- We developed technology to produce extra data to compensate for the lack of training data.
- We proposed the new performance metric to measure correction time.
- The parameter-free deep learning method correctly integrates 91% of all peaks.

Deep learning methods applied to the analysis of metabolomics data

References

1. https://www.shimadzu.com/an/data-net/labsolutions/data_integrity.html
2. https://www.shimadzu.com/an/hplc/support/lib/lctalk/23/23lab.html
3. Liu, Wei, et al. "SSD: Single Shot MultiBox Detector." arXiv preprint arXiv:1512.02325 (2015). Link

First Edition: June, 2019

SHIMADZU

Shimadzu Corporation
www.shimadzu.com/an/