SHORT COMMUNICATION

MALDI Efficiency of Metabolites Quantitatively Associated with their Structural Properties: A Quantitative Structure–Property Relationship (QSPR) Approach

Daichi Yukihira,[1] Daisuke Miura,[2] Yoshinori Fujimura,[2] Yoshikatsu Umemura,[3] Shinichi Yamaguchi,[3] Shinji Funatsu,[3] Makoto Yamazaki,[4] Tetsuya Ohta,[4] Hiroaki Inoue,[4] Mitsuru Shindo,[5] Hiroyuki Wariishi[2,6,7]

[1]Graduate School of Bioresource and Bioenvironmental Sciences, Kyushu University, Higashi-ku, Fukuoka, Japan
[2]Innovation Center for Medical Redox Navigation, Kyushu University, Higashi-ku, Fukuoka, Japan
[3]MS Business Unit, Life Science Business Department, Analytical and Measuring Instruments Division, Shimadzu Corporation, Nakagyo-ku, Kyoto, Japan
[4]Advanced Medical Research Laboratories, Research Division, Mitsubishi Tanabe Pharma Corporation, Toda, Saitama, Japan
[5]Institution for Materials Chemistry and Engineering, Kyushu University, Kasuga, Japan
[6]Bioarchitecture Center, Kyushu University, Higashi-ku, Fukuoka, Japan
[7]Faculty of Arts and Science, Kyushu University, Nishi-ku, Fukuoka, Japan
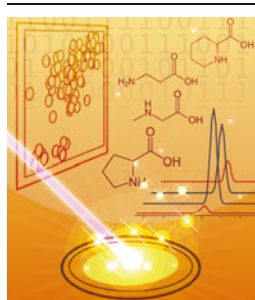
Abstract. Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) experiments require a suitable match of the matrix and target compounds to achieve a selective and sensitive analysis. However, it is still difficult to predict which metabolites are ionizable with a given matrix and which factors lead to an efficient ionization. In the present study, we extracted structural properties of metabolites that contribute to their ionization in MALDI-MS analyses exploiting our experimental data set. The MALDI-MS experiment was performed for 200 standard metabolites using 9-aminoacridine (9-AA) as the matrix. We then developed a prediction model for the ionization profiles (both the ionizability and ionization efficiency) of metabolites using a quantitative structure–property relationship (QSPR) approach. The classification model for the ionizability achieved a 91 % accuracy, and the regression model for the ionization efficiency reached a rank correlation coefficient of 0.77. An analysis of the descriptors contributing to such model construction suggested that the proton affinity is a major determinant of the ionization, whereas some substructures hinder efficient ionization. This study will lead to the development of more rational and predictable MALDI-MS analyses.

Key words: MALDI-MS, Metabolite analysis, QSPR

Introduction

MALDI-MS has come to play a unique role in the analysis of low-molecular-weight biological compounds, principally metabolites [1, 2]. It is well known that the scope of detectable compounds in the MALDI-MS analysis is strongly associated with the molecular species of

Correspondence to: Daisuke Miura; *e-mail:* daipon@agr.kyushu-u.ac.jp, Hiroyuki Wariishi; *e-mail:* hirowari@agr.kyushu-u.ac.jp

the matrix. To date, extensive research has been contributed to elucidate fundamental mechanism of MALDI [3]. However, to clarify whether a target molecular species can be sensitively detected by MALDI-MS, an experimental trial is still required because there is currently no decisive rationale to predict which compounds will be ionizable with which matrices. This problem is largely attributable to the chemical and structural diversity of metabolites, which might hinder the rational understanding of the interrelationships between metabolites and the potential factors affecting their ionization.

In the present study, we aimed to model the relationship between the structural properties of the metabolites and their

ionizability in MALDI. In the targeted analyses, the merit of property modeling lies in the prediction of the probability of the ionization of metabolites yet to be analyzed in MALDI-MS. In the non-targeted analyses, on the other hand, the model would work to screen chemical structures plausibly assigned to a detected peak, even if compounds with similar *m/z* values are not distinguishable. Furthermore, the expected signal response calculated from the ionization efficiency model would provide insights into the abundance of the compound of interest. As a practical case study, we selected 9-AA as the matrix because it is one of the most frequently used matrices for the metabolite analyses MALDI-MS [4]. The MALDI-MS analyses with 9-AA (9-AA-MALDI-MS) have been utilized for various studies, including high-throughput and highly sensitive metabolite analyses [5–8] as well as metabolite MS imaging [9, 10].

First, 200 metabolite standard compounds were selected to cover a wide range of structural diversity and biological importance, and their ionization profiles in MALDI-MS with 9-AA were examined. Second, a quantitative structure–property relationship (QSPR) analysis was performed to model the experimental evaluation using molecular descriptors of the compounds. As there were hundreds of descriptors available, the Random Forest method was employed because of its robust applicability to large multivariate data and unbiased modeling performance [11]. The importance of the descriptors was estimated and discussed with regard to the relevance to the ionizability and ionization efficiency of the compounds.

Methods

The detailed methods for the MALDI-MS analysis and QSPR analysis are described in the Supplemental Materials.

MALDI-MS Analysis of Metabolite Standards

The ionizability and ionization efficiency in MALDI-TOF-MS (AXIMA Confidence, Shimadzu, Japan) analysis for each standard compound was assessed using 9-AA as the matrix. Ionization efficiency was represented as limit of detection (LOD) value in ppm.

Summary of the QSPR Analysis

MDL Molfiles of individual metabolites were acquired from the PubChem website (http://pubchem.ncbi.nlm.nih.gov), using a list of PubChem Compound IDs (CIDs) as the query. The acquired MDL Molfiles were applied for the calculation of the molecular descriptors by the PaDEL-Descriptor software program [12]. The types of molecular descriptors included 1-2D and 3D type descriptors and fingerprints. Descriptors with zero variance or 95 % identical values (including NAs) were excluded from the subsequent analysis.

The LOD was used as the response variable, which could be considered as an inverse measure of the ionization efficiency. In the classification model, the responsive variable was converted to a categorical value denoted as *ionized* or *not ionized*, corresponding to whether the LOD value could be evaluated or not. In the regression model, where *not ionized* observations were eliminated, the LOD values were used in the molar concentrations. Modeling of the inter-relationships between the descriptors and the ionization profiles of metabolites was conducted using the Random Forest method [11]. The importance of variables for constructing a model was evaluated as the mean decrease in accuracy. All of the analyses were performed using the R language [13]. Random Forest and decision tree models were constructed by the party package [14]. The accuracy of the prediction model was evaluated based on the correct rate given as a fraction of the number of correct predictions to the number of the examined metabolites. The performance of a regression model was evaluated by Spearman's rank correlation coefficients between the measured LODs and the predicted values.

Results and Discussion

First, we investigated the ionizability and ionization efficiency of 200 compounds to clarify the coverage of 9-AA-MALDI-MS for the metabolite analysis (Table 1). As a result of the test analysis, 104 out of 200 compounds were detected as deprotonated peaks. The LOD value ranged from 0.00125 to 100 ppm. As the chemical diversity defines the applicability of models constructed using the dataset, the taxonomy superclass of the metabolites in the sample set was summarized in Table 1 (see the Supplemental Materials for the details of the experimental result). Interestingly, distinct ionization profile was observed even in compounds with a similar structure (e.g., alanine and β-alanine, or leucine and isoleucine, Figure 1a, b). In these cases, β-alanine and isoleucine exhibited concentration-dependent peak intensity in MALDI-MS analysis, whereas alanine and leucine were not detected. Generally, structural similarity of low-molecular-weight compounds should give similar physicochemical properties. In contrast, these observations strongly indicated that apparent properties of the molecule, such as the presence of functional groups, are insufficient to explain the diverse ionization profiles of the compounds.

The physicochemical factors of the metabolites that influenced the ionization profiles were of interest. To address these factors, we performed non-hypothesis-based statistical modeling, where the source of efficient MALDI was sought by molecular descriptors of target compounds. First, we constructed a Random Forest QSPR model for the ionizability prediction (*ionized* or *not ionized*) using the whole descriptor provided by the PaDEL-Descriptor (*Global* model). The overall accuracy of the prediction was 86.0 %, and there were no significant biases with regard to the

Table 1. The Ionization Profiles of Metabolites in the 9-AA-MALDI-MS Analysis and the Predictive Accuracy of the Random Forest Ionizability Models

Superclass	#	Ionized	Correct rate of prediction model					LOD (ppm)	
			Global	2D	3D	KRFP [c]	MACCSFP [c]	Minimum	Max
Aliphatic acyclic compounds	10	1 (10 %)	1.000	1.000	1.000	1.000	1.000	0.00125	0.00125
Amino acids, peptides, and analogues	59	49 (83 %)	0.847[a] / 0.831[b]	0.833[a] / 0.983[b]	0.915[a] / 1.000[b]	0.864[a] / 0.763[b]	0.750[a] / 0.864[b]	0.0125	10
Aromatic heteromonocyclic compounds	12	6 (50 %)	0.917	0.833	1.000	0.833	0.750	0.157	2.50
Aromatic heteropolycyclic Compounds	8	3 (38 %)	0.750	0.875	1.000	0.750	0.750	0.0313	50.0
Aromatic homomonocyclic Compounds	12	8 (67 %)	0.750	0.667	0.833	0.667	0.583	0.050	100
Carbohydrates and their conjugates	13	6 (46 %)	0.769	0.846	0.846	0.846	0.692	0.0250	25.0
Lipids	36	2 (6 %)	0.972	0.944	0.917	0.917	0.944	0.157	0.312
Nucleosides, nucleotides, and analogues	27	19 (70 %)	0.926	0.963	0.889	0.889	0.778	0.00625	6.25
Organic acids and derivatives	14	8 (57 %)	0.571	0.857	0.929	0.571	0.429	0.0500	12.5
Others	9	2 (22 %)	0.889	0.889	0.778	0.778	0.889	0.625	100
Whole compounds	200	104 (52 %)	0.860	0.885	0.910	0.840	0.795	0.00125	100

[a] Predicted by whole-data model
[b] Predicted by amino acid model
[c] The other available types of fingerprint were omitted because of their moderate performance
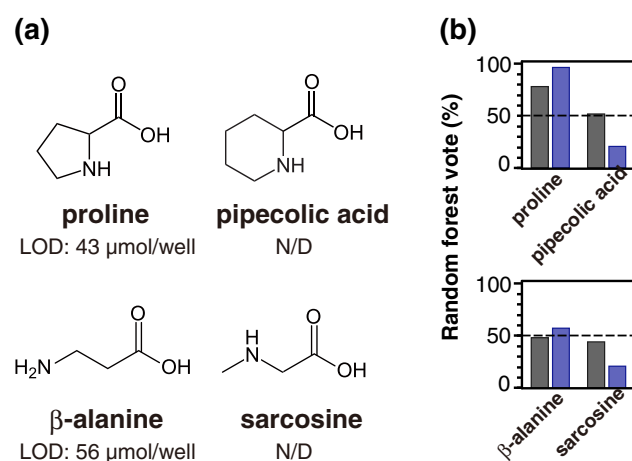
Figure 1. Distinct ionization profiles of structurally similar compounds in MALDI-MS analysis and their Random forest prediction. **(a)** Structural formulas and LODs of four representative compounds with similar structures but distinct ionization profiles. **(b)** The prediction of ionizability by the 3D descriptor model for whole compounds (*gray bar*) and by the 3D descriptor amino-acid-specific model (*blue bar*) represented as the votes of the ensemble trees. When the ratio of positive vote (ionizable) exceeds 50 %, the corresponding compound is predicted to be ionizable

estimation error and the metabolite class (Table 1, *Global* model for whole compounds).

The prediction model was then investigated to estimate the prerequisite properties for the ionization of a compound in a 9-AA-MALDI-MS analysis. In the *Global* model, the descriptors with higher importance indicated the electrotopological state of strength for potential hydrogen bonds and the area of the negatively charged surface (Supplemental Figure S-1a and Supplemental Table S-1). These descriptors belong to the 2D and 3D descriptors, respectively. The electrotopological state value (E-state value) is a kind of 2D descriptor that combines both the electronic characteristics and the topological environment of each skeletal atom in the molecule [15]. The importance of the E-state value indicated that the strength of possible hydrogen bonds positively correlated with the ionizability in MALDI. It was clear that the ionization profiles were strongly influenced by the interaction between molecules. In addition to the global model, which incorporated all the type of descriptors available, the respective types of descriptors were applied to construct Random Forest prediction models to investigate the relevance of each descriptor types to the prediction performance (Table 1). As the result, 3D model exhibited the highest performance followed by 2D model (91.0 % and 88.5 % accuracy rate for whole compounds, respectively). Considering the variable importance of these models (Supplemental Figure S-1b, c), although the strength of hydrogen bonds well represented the ionization profile, the information of charged surface

area led to a better ionizability model. This result was reasonable because the charged surface area indicated the electron distribution within the molecules that should cover the effect of hydrogen bond acceptors. The further functioning of the negatively charged surface area could be the effectiveness of proton abstraction in the interaction with matrix molecule, 9-AA.

The constructed prediction models for amino acids (*"Amino Acids, Peptides, and Analogues"* class) exhibited relatively poor accuracy, even though they were a major class in our data set. Our models were effective for a broad spectrum of metabolites, but they still lacked the ability to model rather faint structural differences of amino acids. The reason of this defect could be strongly attributed to the relevance of hydrogen bonds. As both amines and carboxyl groups in amino acids can form hydrogen bonds, the ionizabilities of amino acids could be overestimated. To address these issues, we attempted to improve the prediction performance for amino acids because they are one of the most important classes in the metabolite analysis because of their significant metabolic and regulatory versatility [16]. We thus developed new models specific for amino acids to improve the predictive accuracy and investigate the relevant structural properties. Again, the models were constructed using the whole or the individual types of descriptors. As a result, the accuracy of model prediction improved for all types of descriptors (Table 1). Especially, the 3D model achieved a perfect prediction of the ionizability, even for the above-mentioned pairs of structurally similar amino acids (Figure 1c). Fingerprinting descriptors provided still a moderate accuracy (86.4 % correct rate for the highest value by the MACCSFP model), indicating that the presence of substructures was insufficient to fully represent the ionizability of amino acids. Unlike the class-independent model (whole-data model), the relevant 3D descriptors were not involved with the charged surface areas, but Weighted Holistic Invariant Molecular (WHIM) descriptors [17] (Supplemental Figure S-1d). WHIM descriptors provide information about the whole 3D-molecular structure in terms of the size, shape, symmetry, and atom distribution. This result was intriguing because the shape of the molecules itself was relevant rather than electronic properties. It has been reported that cation affinities of amino acids were associated with degree of linearity [18], which is a direct index of the flexibility of molecule [19]. It was thus suggested that the shape properties of target compounds affect their interaction with other molecules to promote or inhibit their ionization.

The Random Forest method is applicable to a regression, averaging the output of decision trees [11]. The experimentally evaluated ionization efficiency, indicated by LOD values, was also modeled by the Random Forest method using individual types of descriptors. While the *Global* and 3D ionization efficiency models both reached $\rho = 0.77$ (Supplemental Figure S-2a, b, and the variable importance for *Global* model was shown in Supplemental
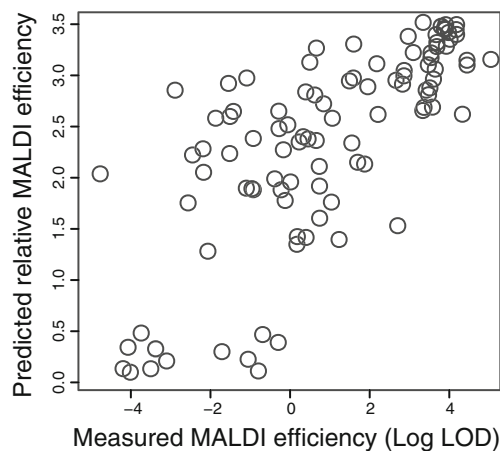

Figure 2. The Random forest regression model for the ionization efficiency in 9-AA-MALDI. The 2D model showed the best performance in terms of the regression for ionization efficiency. The rank correlation coefficient for the plot was indicated as ρ. The models of other types (Global, MACCSFP, and 3D) can be found in Supplemental Figure S-2

Figure S-1e), the best predictive performance was achieved with 2D descriptors, evaluated as $\rho = 0.78$ (2D model, Figure 2, and the variable importance was shown in Supplemental Figure S-1f). The MACCSFP also provided a highly accurate model compared to the 2D, 3D, and *Global* models ($\rho = 0.69$, Supplemental Figure S-2b). It was supposed that the fundamental trend of the ionization efficiency was reasonably modeled. The 2D model indicated that the quantitative extent of ionization was mainly associated with E-state index of double-bonded oxygen and the strength of the potential hydrogen bonds (Supplemental Figure S-1f). Hence, overall results indicated that the partial negative charge in the molecule could be a prerequisite for ionization, and that the richness of carbonyl oxygen should be preferable for efficient negative MALDI because of the basic condition brought by 9-AA. However, Sun et al. showed that pH condition altered the ionized metabolite profiles specifically to analyzed molecular species [20]. They also reported that multiplexed solvent could be used for optimization of analyte-matrix interaction during co-crystallization [8]. The formation of hydrogen bonds, which could be affected by pH condition, might result in specific crystal structures with the advantage of binding energy, leading to distinct MALDI efficiencies. Noteworthy, structural flexibility of the target compounds might play a special role to specific interaction with other molecules, presumably the matrix molecules to reduce ionization energies [21], which determine the fate of their ionization profiles.

Conclusions

This study was primarily intended to lead to more rational and predictive MALDI-MS analyses. In contrast to empirical

approaches, this study employed a systematic analysis of the ionization profile in 9-AA-MALDI-MS for the first time. In the MALDI-MS analysis, the ionizability prediction model evaluates the likelihood of peak identification. On the other hand, the ionization efficiency model would help to estimate the abundance of the metabolite based on the observed signal intensity. The relevant descriptors found in this study can be interpreted as the structural preference specific to 9-AA and/or negative mode MALDI-MS analysis. The QSPR approach should also be applicable for other MALDI matrices to characterize the structural properties of target compounds for preferred ionization. Such information will play an indispensable role in the strategic development of MALDI-MS-based studies.

Acknowledgment

Open Access

References

1. Dally, J.E., Gorniak, J., Bowie, R., Bentzley, C.M.: Quantitation of underivatized free amino acids in mammalian cell culture media using matrix assisted laser desorption ionization time-of-flight mass spectrometry. Anal. Chem. **75**, 5046–5053 (2003)
2. Edwards, J.L., Kennedy, R.T.: Metabolomic analysis of eukaryotic tissue and prokaryotes using negative mode MALDI time-of-flight mass spectrometry. Anal. Chem. **77**, 2201–2209 (2005)
3. Knochenmuss, R.: Ion formation mechanisms in UV-MALDI. Analyst **131**, 966–986 (2006)
4. Vermillion-Salsbury, R.L., Hercules, D.M.: 9-Aminoacridine as a matrix for negative mode matrix-assisted laser desorption/ionization. Rapid Commun. Mass Spectrom. **16**, 1575–1581 (2002)
5. Amantonico, A., Oh, J., Sobek, J., Heinemann, M., Zenobi, R.: Mass spectrometric method for analyzing metabolites in yeast with single cell sensitivity. Angew. Chem. Int. Ed. **47**, 5382–5385 (2008)
6. Miura, D., Fujimura, Y., Tachibana, H., Wariishi, H.: Highly sensitive matrix-assisted laser desorption ionization-mass spectrometry for high-throughput metabolic profiling. Anal. Chem. **82**, 498–504 (2010)
7. Yukihira, D., Miura, D., Saito, K., Takahashi, K., Wariishi, H.: MALDI-MS-based high-throughput metabolite analysis for intracellular metabolic dynamics. Anal. Chem. **82**, 4278–4282 (2010)
8. Sun, G., Yang, K., Zhao, Z., Guan, S., Han, X., Gross, R.W.: Matrix-assisted laser desorption/ionization time-of-flight mass spectrometric analysis of cellular glycerophospholipids enabled by multiplexed solvent dependent analyte–matrix interactions. Anal. Chem. **80**, 7576–7585 (2008)
9. Miura, D., Fujimura, Y., Yamato, M., Hyodo, F., Utsumi, H., Tachibana, H., Wariishi, H.: Ultrahighly sensitive in situ metabolomic imaging for visualizing spatiotemporal metabolic behaviors. Anal. Chem. **82**, 9789–9796 (2010)
10. Miura, D., Fujimura, Y., Wariishi, H.: In situ metabolomic mass spectrometry imaging: recent advances and difficulties. J. Proteome. **75**, 5052–5060 (2012)
11. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
12. Yap, C.W.: PaDEL-Descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. **32**, 1466–1474 (2011)
13. R Core Team: R, a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2012)
14. Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., Van, D.L.: Survival ensembles. Biostatistics **7**, 355–373 (2006)
15. Hall, L.H., Kier, L.B.: Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. J. Chem. Inf. Comput. Sci. **35**, 1039–1045 (1995)
16. Wu, G.: Amino acids: metabolism, functions, and nutrition. Amino Acids **37**, 1–17 (2009)
17. Todeschini, R., Lasagni, M., Marengo, E.: New molecular descriptors for 2D and 3D structures. Theory J. Chemometr. **8**, 263–272 (1994)
18. Siu, F., Che, C.: Quantitative structure–activity (affinity) relationship (QSAR) study on protonation and cationization of α-amino acids. J. Phys. Chem. A **110**, 12348–12354 (2006)
19. Devillers, J., Balaban, A.T.: Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Amsterdam (1999)
20. Sun, G., Yang, K., Zhao, Z., Guan, S., Han, X., Gross, R.W.: Shotgun metabolomics approach for the analysis of negatively charged water-soluble cellular metabolites from mouse heart tissue. Anal. Chem. **79**, 6629–6640 (2007)
21. Kinsel, G.R., Knochenmuss, R., Setz, P., Land, C.M., Goh, S., Archibong, E.F., Hardesty, J.H., Marynick, D.S.: Ionization energy reductions in small 2,5-dihydroxybenzoic acid–proline clusters. J. Mass Spectrom. **37**, 1131–1140 (2002)