DMT-(S)-Pro-OSu derivatization UPLC/ESI-MS/MS is a powerful method for the comprehensive analysis of dipeptides

Keisuke Ito (✉ sukeito@u-shizuoka-ken.ac.jp)  
University of Shizuoka

Yuko Terada  
University of Shizuoka

Ryoka Nakayama  
University of Shizuoka

Takashi Shigeta  
University of Shizuoka

Takumi Hisada  
Nagoya University

Masaya Fujitani  
Nagoya University

Kei Kanie  
Nagoya University

Ryuji Kato  
Nagoya University

Serina Fukui  
University of Shizuoka

Eiji Sugiyama  
University of Shizuoka

Hajime Mizuno  
University of Shizuoka

Kenichiro Todoroki  
University of Shizuoka

Article

Keywords: dipeptide, UPLC/ESI-MS/MS, cocoa bean, machine learning model, analytical behavior

Posted Date: February 2nd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2414378/v1
Abstract

Dipeptides are important components that contribute to the characteristics of various foods; however, the dipeptide composition of foods is poorly understood. In the present study, a new, comprehensive analysis method for dipeptides, i.e., \((\mathcal{S})\)-2,5-dioxopyrrolidin-1-yl-1-(4,6-dimethoxy-1,3,5-triazin-2-yl)pyrrolidine-2-carboxylate (DMT-(\mathcal{S})-Pro-OSu) derivatization ultra-performance liquid chromatography–electrospray ionization tandem mass spectrometry (UPLC/ESI-MS/MS), was developed. Using the method, the analytical behavior of 313 dipeptides in the UPLC/ESI-MS/MS was collected. The obtained comprehensive dataset elucidated the dipeptide profiles in fermented cocoa beans. Furthermore, the machine learning on the dataset provided the quantitative understanding of physicochemical molecular descriptors that affect the elution time of dipeptides on the reversed-phase LC analysis, namely, an \textit{in silico} elution time prediction model was successfully constructed. The DMT-(\mathcal{S})-Pro-OSu derivatization UPLC/ESI-MS/MS method is a powerful tool for the comprehensive analysis of dipeptides.

Introduction

Dipeptides contribute to the characteristics of a wide variety of foods\textsuperscript{1–3}. They play important roles in flavor formation in fermented foods (e.g., alcoholic beverages, soy sauce, cheese, and bread), meat, and broth\textsuperscript{3–5}. Furthermore, they are involved in various physiological functions of foods\textsuperscript{1,2,6,7}; for example, the inhibitory effects of human dipeptidyl peptidase-IV (hDPP-IV) (anti-type 2 diabetic effect) and angiotensin-converting enzyme (antihypertensive effect) have been reported\textsuperscript{8–10}. Because dipeptides are more rapidly and efficiently absorbed into the human body than amino acids and proteins\textsuperscript{11,12}, they are used for enteric nutritional supplements and foods for sports. As aforementioned, dipeptides are important components that contribute to various characteristics of foods; however, their composition is poorly understood.

A limited number of reports have simultaneously analyzed and quantified multiple types of dipeptides, because there are as many as 400 types of dipeptides composed of 20 different amino acids, and there are multiple dipeptides with the same molecular weight (e.g., Arg–Thr and Thr–Arg) that are difficult to distinguish. Takahashi et al. quantified 32 dipeptides in alcoholic beverages \textit{via} ultrahigh-pressure liquid chromatography–tandem mass spectrometry (UHPLC-MS/MS) and UHPLC–quadrupole time-of-flight mass spectrometry (UHPLC-Q-TOF/MS)\textsuperscript{13}. Yamamoto et al. detected 237 dipeptides in soy sauces \textit{via} liquid chromatography–tandem mass spectrometry (LC-MS/MS) in multiple reaction monitoring mode. In addition, they analyzed the correlations between the dipeptide profile and taste differences in soy sauces\textsuperscript{14}. Dipeptides have also been analyzed as biomarkers. Wu et al. evaluated 361 dipeptides in lung cancer tissue \textit{via} a dansyl [5-(dimethylamino)-1-naphthalene sulfonamide] derivatization in conjunction with LC-MS. In addition to detecting 257 dipeptides in the tumor tissue, they reported 90 dipeptides that were significantly increased in cancer tissue compared with normal lung tissue\textsuperscript{15}. Ozawa et al. identified 236 dipeptides from liver cancer tissue \textit{via} LC-MS/MS and capillary electrophoresis-MS/MS\textsuperscript{16}.  

Page 3/19
Although there are several reports analyzing dipeptide composition, comprehensive analysis methods have not been established. In this study, a new, comprehensive analysis method for dipeptides, i.e., DMT-(S)-Pro-OSu derivatization ultra-performance liquid chromatography–electrospray ionization tandem mass spectrometry (UPLC/ESI-MS/MS), was developed. The analytical behavior of 313 dipeptides via UPLC/ESI-MS/MS was successfully collected using this method. The obtained comprehensive dataset elucidated dipeptide composition in food. Moreover, the machine learning model constructed with the dataset (comprising the physicochemical molecular descriptors of peptide and their elution time) afforded a quantitative understanding of relationship between the combination of the physicochemical effect of dipeptides and their elution time in the reversed-phase LC analysis. This DMT-(S)-Pro-OSu derivatization UPLC/ESI-MS/MS method can be a powerful tool for the comprehensive analysis of dipeptides.

Results And Discussion

Dipeptide analysis using (S)-2,5-dioxopyrrolidin-1-yl-1-(4,6-dimethoxy-1,3,5-triazin-2-yl)pyrrolidine-2-carboxylate (DMT-(S)-Pro-OSu) derivatization reagent

The derivatization reagent, DMT-(S)-Pro-OSu, was developed for the UPLC-MS/MS analysis of amines and amino acids in our previous work\textsuperscript{17}. A succinimidyl ester moiety of DMT-(S)-Pro-OSu reacts with amines under mild conditions and generates the reagent-specific product ion. Furthermore, its triazine moiety possessing a high proton affinity enables highly sensitive detection in the ESI mode. At the beginning of this study, it was determined whether the derivatization reagent could be used for dipeptide analysis. Based on our work analyzing amino acids\textsuperscript{17}, the derivatization reaction of dipeptides and their cleavage were assumed as presented in Fig. 1a. In Fig. 1b, the MS/MS spectra of Ala–His, used as a representative dipeptide, is presented. Specific and characteristic cleavages were observed from the collision-induced dissociation of the protonated molecule [M + H]\textsuperscript{+}. The most characteristic MS transitions for the derivative of DMT-(S)-Pro-OSu were m/z 463.2→209.1. The fragment with m/z 209.1 is an ion derived from the DMT-(S)-Pro-OSu moiety. In addition, m/z 280.1 is an a-type immonium ion fragmented between the α-carbon and the carboxyl group of the amino acid on the N-terminal side, as presented in Fig. 1b. The signal intensity of the a\textsubscript{1} fragment ion (m/z 280.1) was one-fifth of that of the reagent ion (m/z 209.1). These results indicated that the DMT-(S)-Pro-OSu derivatization reagent was applicable for the analysis of not only amines but also dipeptides.

Optimization of precursor ion m/z and elution time for each dipeptide

The 313 available dipeptide standards were labeled with DMT-(S)-Pro-OSu and analyzed via UPLC/ESI-MS/MS to determine their optimized precursor ions (m/z). A series of analyses determined the optimized precursor ion (m/z) for the 313 dipeptide derivatives (Fig. 2a).

Subsequently, we determined the elution time of the dipeptide derivatives in the reversed-phase LC analysis. In the analysis, the precursor ion (m/z) was adapted for each dipeptide, as presented in Fig. 2a,
and the product ion (m/z) was set to 209.1, the reagent ion. The elution times obtained for the 313 dipeptide derivatives are presented in Fig. 2b. There was a wide range of elution times, from 2.65 min (Lys–His) to 10.73 min (Phe–Phe). X–Ile forms were eluted approximately 0.20 min earlier than X–Leu forms when the same amino acids were at the N-terminus. For example, the elution times of Ala–Ile and Ala–Leu were 7.34 and 7.53 min, respectively. In addition, there was a trend that Ile–X was eluted approximately 0.20 min earlier than Leu–X when the same amino acids were present at the C-terminus. Accordingly, the elution time was suggested to reflect the physicochemical characteristics of each dipeptide.

**Separation and identification of dipeptides with the same precursor ions**

An example of dipeptides with the same precursor ions is presented in Fig. 3. Although Val–Ile and Val–Leu have the same precursor ion (m/z 467.2), they were successfully separated as they have different elution times, 8.66 min for Val–Ile and 8.81 min for Val–Leu. Collectively, the 313 available dipeptides were analyzed individually, and their optimized precursor ion (m/z) and elution time were determined. In the simultaneous UPLC-MS/MS analysis of the 313 dipeptides, the identification of 87 was difficult due to changes in elution time, probably caused by matrix effects and other interferences; however, another 226 dipeptides were successfully separated and identified by adapting their elution time, precursor ion, and product ion (reagent ion and a₁ fragment ion).

**Calibration curves of dipeptides**

To prepare calibration curves, the 226 dipeptide standards detectable in the simultaneous analysis were mixed and diluted to 0.125–2 µM. From the analysis of the mixed standard, calibration curves were successfully obtained for 126 of the dipeptides, although the quantification of another 100 dipeptides was difficult in the simultaneous analysis. Because the precursor ion and elution time of these 100 dipeptides were not specific for each dipeptide, quantification using the peak area of their reagent ion (m/z 209.1) was complex. We attempted to quantify those dipeptides using the peak area of their a₁ fragment ion, but they did not exhibit concentration dependence. In contrast, the calibration curves for the 126 dipeptides were successfully obtained, as the peak area of the reagent ion (m/z 209.1) and/or a₁ fragment ion changed in a concentration-dependent manner at 0.5–2 µM (Supplementary Table 1). Among them, dipeptides with specific precursor ions and elution times were quantified using the peak area of the reagent ion (m/z 209.1). Some dipeptides with the same precursor ion and elution time were difficult to separate, but the adaptation of their a₁ fragment ion enabled their separation and quantification. Ozawa et al. quantified 335 of dipeptides with a limit of quantification (LOQ) of 0.1–10 µM via capillary electrophoresis-MS/MS and LC-MS/MS¹⁸. Takahashi et al. detected 32 dipeptides with a limit of detection (LOD) of 0.8–100 nM via UHPLC-MS/MS and UHPLC-Q-TOF/MS¹³, and Hanh et al. analyzed five dipeptides with LOD at 0.13–0.54 µg/mL (approximately 0.4–2 µM) via LC-MS¹⁹. The sensitivity of our analysis method was comparable with that of previous studies, as we quantified 126 dipeptides with an LOQ of 0.5 µM.
detected in Japanese sake at 1–97 µM, and 23 dipeptides were contained in parmesan cheese at 3–2807 µmol/kg\textsuperscript{13,20}. Therefore, our analytical method is able to quantify dipeptides in foods.

**Dipeptide profile of fermented cocoa beans from different geographical origins**

By using the UPLC-MS/MS method established in this study, we indeed investigated the dipeptide profile of four fermented cocoa beans from different geographical origins: Brazil, Ghana, Grenada, and Haiti. Fermented cocoa beans are ingredients of chocolate and cocoa liquor\textsuperscript{21}. Dipeptides generated during cocoa bean fermentation are important as precursors of the development of cocoa flavor in cocoa bean products\textsuperscript{22-25}. As the flavor of cocoa bean products varies depending on the geographical origin of fermented cocoa beans, it is suggested that the dipeptide constituents of the beans are different\textsuperscript{26,27}. However, as far as we know, there are no comprehensive studies of the content and composition of dipeptides in fermented cocoa beans. Thus, extracts of these beans were DMT-(S)-Pro-OSu-derivatized, and 126 dipeptides were quantified in a simultaneous analysis. To the best of our knowledge, there have been several reports that simultaneously analyze the dipeptides contained in foods. Takahashi et al. quantified 32 dipeptides in alcoholic beverages via UHPLC-MS/MS and UHPLC-Q-TOF/MS\textsuperscript{13}. Yamamoto et al. detected 237 dipeptides in soy sauce via LC-MS/MS; however, the quantification was insufficient\textsuperscript{14}. Compared with these reports, our analytical method has simultaneously quantified the largest number of dipeptides (126) in food research. As a result of the analysis, a total of 42 kinds of dipeptides were determined (Fig. 4). The numbers of dipeptides detected from each sample were as follows: Brazil, 25; Ghana, 29; Grenada, 30; and Haiti, 32. A total of 17 common dipeptides were identified: Trp–Pro, Phe–Lys, Lys–Trp, Phe–Gly, Phe–Phe, Val–Phe, Ala–Trp, Lys–Pro, Asn–Trp, Phe–Val, Glu–Trp, Tyr–Tyr, Ser–Pro, Ala–Phe, Gly–Phe, Ile–Pro, and Glu–Glu. Certain dipeptides were detected from only one sample: Brazil, Glu–Gln; Ghana, Tyr–His, Ala–Val, Lys–Met, and Tyr–Arg; Grenada, Pro–Trp; and Haiti, Ile–Val, Asp–Leu, Thr–Trp, Met–Tyr, Trp–Trp, and Glu–Lys. Collectively, some dipeptides, e.g., Trp–Pro, Phe–Lys, Lys–Trp, and Phe–Gly, are common in different geographical origins, whereas some dipeptides specific to each geographical origin would diversify the characteristics of each cocoa bean. There are reports analyzing oligopeptides in fermented cocoa beans\textsuperscript{23,25,27}. However, the separation of peptides with the same molecular weight is not complete in those reports. In the comparison of these reports and our results, six dipeptides—Trp–Pro, Phe–Gly, Gly–Phe, Phe–Val, Val–Phe, and Ala–Trp—were identified in previous works and the current work. Accordingly, these six dipeptides would distribute commonly in fermented cocoa beans regardless of their geographical origin and fermentation condition.

**Construction of *in silico* elution time prediction model from the dipeptide analysis dataset**

Commonly, the major interaction between the column and eluate has been considered as a hydrophobic interaction. We first examined to construct a machine learning model with the peptides’ residual hydrophobicity scores, practically N-terminal, and C-terminal hydropathy indices (two descriptors) (Supplementary Fig. 1). However, the elution time prediction performance did not exceed the simple correlation between the total hydropathy score of dipeptides and their elution time (correlation coefficient
0.65, \( R^2 = 0.451 \). This result strongly indicates that the hydrophobic interaction information is certainly a major effect, although not sufficient for enabling high performance elution time prediction.

Therefore, we finally tried to understand the combinatorial rules hidden in the physicochemical properties of dipeptides and column interaction by constructing a machine learning model using multiple physicochemical molecular descriptors to predict their elution time. As the size of our UPLC-MS/MS analysis dataset for dipeptides was significantly large and the accuracy of each dipeptide determination was high, it was expected that our data can extract the universal combinatorial rule between the complex multiple physicochemical effects and their elution times in dipeptides. Using the 313 datasets of 23 physicochemical molecular descriptors describing dipeptide sequences together with their elution times (Supplementary Table 2), the elution time prediction model was constructed. The prediction performance of the obtained model was high (\( R^2 = 0.832 \), root mean squared error (RMSE) = 0.753) (Fig. 5) and significantly exceeded the model trained only with hydrophobicity score (\( R^2 = 0.443 \), RMSE = 1.37) (Supplementary Fig. 1). Such a result clearly indicates that the dipeptide–column interaction has more complex rules; therefore, a machine learning model that explores the best combination of multiple molecular descriptors is more effective for prediction.

By interpreting the weighted descriptors in the multiple-descriptor prediction model (Fig. 5) “Molecular Weight,” “Bond Polarizabilities,” and “Atomic Polarizabilities” were found to be the three major indices suggested to contribute to the elution time prediction (Table 1). In other words, this multiple parameter combination can better explain the prediction of elution time. It was also suggested that “Mannhold LogP” and “XLogP,” which reflect the hydrophobicity of molecule, were also weighted in the model but were not the most heavily weighted. These results also indicated that the dipeptide–column interaction was affected by complex combination with hydrophobicity and other physicochemical properties.

Our data also suggests that the dipeptide elution time can be predicted with simple linear regression model and does not require deep neural network models that have a higher risk of overfitting. This finding encourages further investigations for other short peptide prediction models. By following our concept of peptide sequence conversion combined with simple regression model, peptide elution model trials can escape tuning too many hyperparameters in deep neural network models with limited data, which cannot be easily enlarged by their cost.

As our constructed elution time prediction model showed good performance, the elution profiles of the dipeptides that were not in our data can be predicted \textit{in silico} with high reliability. Such a prediction of the elution time profile will support the estimation of unknown dipeptides found in the UPLC-MS/MS analysis. Without the need to prepare all standard dipeptide UPLC-MS/MS profiles, our data-based elution model estimations will provide clues to understand the molecular candidates for unknown extracts analyzed \textit{via} UPLC-MS/MS. Moreover, as our prediction model recognizes dipeptides by their total molecular profile, rather than their exact sequence, it should have the potential to predict other molecules of similar size to dipeptides, as in our previous work\textsuperscript{28}. Our data also suggest that for such a large volume of peptide data and its UPLC-MS/MS profile, machine learning models can complement the
metabolomics analysis of food ingredients, which commonly consist of many unknown small molecules that cannot always be clearly identified by the “standard substance.”

Although dipeptides are important components that contribute to food properties, e.g., flavor and physiological functions, the dipeptide composition of foods is poorly understood. This study developed a new, comprehensive analysis method for dipeptides, namely, the DMT-(S)-Pro-OSu derivatization UPLC/ESI-MS/MS method. The obtained comprehensive dataset elucidated not only the dipeptide profiles of the fermented cocoa beans but also the combinatorial effect of physicochemical molecular descriptors that affect the elution time of dipeptides in reversed-phase LC analysis. In particular, our approach of constructing the in silico elution time prediction model should provide a new viewpoint on LC analysis and contribute to the development of new analytical methods using LC. DMT-(S)-Pro-OSu derivatization UPLC/ESI-MS/MS is a powerful method for the comprehensive analysis of dipeptides.

**Methods**

**Materials and chemicals**

Formic acid, acetonitrile, and triethylamine were purchased from FUJIFILM Wako Pure Chemical Industries (Osaka, Japan). Methanol was obtained from Kanto Kagaku Co., Inc. (Tokyo, Japan). A dipeptide library was purchased from AnaSpec (California, USA). Available 313 dipeptides were used in this study. Fermented cocoa beans cultivated in Brazil, Ghana, Haiti, and Grenada were obtained from Conche (Shizuoka, Japan), a bean-to-bar chocolate shop.

**Ultra-performance liquid chromatography–electrospray ionization tandem mass spectrometry (UPLC/ESI-MS/MS) with selected reaction monitoring (SRM) mode**

Based on our previous work analyzing amino acids\textsuperscript{17,29}, the following UPLC/ESI-MS/MS conditions were set: ACQUITY UPLC I-class (Waters, Milford, MA) connected to a Xevo™ TQ-S triple quadrupole mass spectrometer (Waters). The following UPLC conditions were adopted: column, ACQUITY UPLC BEH C18 column (1.7 µm, 150 x 2.1 mm i.d., Waters); column temperature, 40°C; flow rate: 0.3 mL/min; and injection volume, 1 µL. The LC gradient using mobile phases A and B (A: 0.1% formic acid in H$_2$O, B: 0.1% formic acid in acetonitrile) was as follows: 0–1 min = 10% B; 1–15 min = 70% B; and 15–16 min = 95% B. The separation and detection conditions of UPLC/ESI-MS/MS were as follows: ion mode, positive-ion mode; capillary voltage, 3.00 kV; cone voltage, 30 V; cone gas flow, 150 L/h; nebulizer gas flow, 7.0 L/h; collision gas flow, 0.15 mL/min; collision energy, 30 eV; collision cell exit potential, 5 V; desolvation temp, 200°C; and desolvation gas flow, 800 L/h. MS transitions (precursor and product ions) were optimized for each dipeptide derivative. The $m/z$ of the precursor ion in selected reaction monitoring (SRM) mode was set based on the predicted structure of the derivatized dipeptides. An analytical software (MassLynx, version 4.1) was used for the system control and data processing.
(S)-2,5-Dioxopyrrolidin-1-yl-1-(4,6-dimethoxy-1,3,5-triazin-2-yl)pyrrolidine-2-carboxylate (DMT-(S)-Pro-OSu) derivatization of dipeptides

DMT-(S)-Pro-OSu was synthesized as previously described\(^\text{17}\). Based on our report analyzing amino acids, dipeptides were derivatized as follows. Forty microliters of dipeptides (1.25–20 µM) in H\(_2\)O, 40 µL of 10-mM DMT-(S)-Pro-OSu in acetonitrile, and 120 µL of 100-mM triethylamine in acetonitrile were mixed and reacted at room temperature for 40 min. The reaction was quenched with 200 µL of H\(_2\)O/methanol/acetonitrile = 3/1/1 (v/v/v). The reacted solution was filtered through a 0.22 µm PVDF membrane (Merck Millipore, Cork, Ireland), and 1 µL was analyzed using the UPLC/ESI-MS/MS system.

### Calibration curves of dipeptides

The 226 dipeptides that were simultaneously detected in the analysis using the UPLC/ESI-MS/MS system. The reacted solution was diluted with H\(_2\)O to contain 0.125, 0.25, 0.5, 1, or 2 µM of each dipeptide. The solution (1 µL) was then analyzed using the UPLC/ESI-MS/MS system. The calibration curves were constructed by plotting the peak area of the DMT-(S)-Pro-OSu derivative against the injected amount of dipeptide (n = 3).

### Analysis of dipeptides in the fermented cocoa beans

Cocoa beans that were cultivated and fermented in four different countries (Brazil, Ghana, Haiti, and Grenada) were used. The beans were frozen at −30°C and ground using mortar and pestle. The ground bean powder (5 g) was extracted with 30 mL of H\(_2\)O for 1 h at 4°C. The supernatant (1 mL) was isolated and centrifuged (21,500 g, 15 min, 4°C). The obtained supernatant was vacuum-dried and redissolved in 200 µL of H\(_2\)O. The solution was deproteinated with 800 µL of acetonitrile (~30°C, 15 min) and centrifuged (21,500 g, 20 min, −10°C). The supernatant was vacuum-dried, and the resulting residue was dissolved in 40 µL of H\(_2\)O. After the DMT-(S)-Pro-OSu derivatization, the reacted solution was purified via solid-phase extraction (ZipTip C18, Merck Millipore) using a 10% acetonitrile–water mixture containing 0.1% (v/v) formic acid as an eluent. After vacuum-drying, the resulting residue was dissolved in 100 µL of 10% acetonitrile–water mixture containing 0.1% (v/v) formic acid and filtered through a 0.22-µm PVDF membrane (Merck Millipore). The solution (1 µL) was then subjected to the UPLC/ESI-MS/MS system. For the quantification of dipeptides in the cocoa bean extracts, the standard curves of area versus concentration (µM) were prepared for each standard material. The UPLC/ESI-MS/MS analysis of the extracts was repeated three times, and the dipeptide content was expressed as mean ± SE.

### Construction of machine learning model for predicting elution time from dipeptide sequence

Based on the 313 dipeptide sequence dataset (Supplementary Table 2) combined with their individual elution time data determined via UPLC/ESI-MS/MS, a LASSO regression model was constructed as the “elution time predictor from amino acid sequence”. First, all the dipeptide sequence data were converted into their molecular descriptors, with 23 physicochemical characteristics applied to describe one
dipeptide as a single molecule, using the RDKit (http://www.rdkit.org/) and CDK Toolkit (Chemistry Development Kit) (https://cdk.github.io/index.html) in KNIME (version 3.7: https://www.knime.com/). From the total KNIME molecular descriptors (46 descriptors from RDKit and 128 descriptors from CDK), we statistically analyzed their multicollinearity and selected 23 descriptors that were independent and the most interpretable. The list of descriptors is presented in Supplementary Table 3. Using such physicochemical molecular descriptors as the explanatory valuables, the elution time of each dipeptide was set as the objective valuable for the construction of the machine learning model. By the nature of the LASSO model, meaningless descriptors are excluded from the model, and meaningful descriptors are weighted to make the best combination effect. Therefore, such an objectively extracted combination rule of descriptors from the elution time prediction model is worth interpreting to understand the complex combination of multiple physicochemical effects. For an input dipeptide sequence, the final prediction model can predict its elution time as a continuous value. In the comparative analysis for understanding the molecular rule in elution time prediction, dipeptides were also analyzed only by their hydrophobicity scores, known to be the primary physicochemical effect in column separation. First, the total hydrophobicity score for each dipeptide was calculated using the hydropathy index (A: 0.73; C: 0.96; D: −1.03; E: −1.03; F: 1.06; G: 0.63; H: −0.93; I: 1.62; K: −1.16; L: 1.39; M: 0.76; N: −1.03; P: −0.4; Q: −1.03; R: −1.36; S: −0.13; T: −0.1; V: 1.52; W: −0.17; Y: −0.3)30 from AAindex (https://www.genome.jp/aaindex/). To investigate the explanatory effect of the total hydrophobicity score of dipeptides, the LASSO model was constructed using hydropathy indices of N- and C-terminal amino acids (two descriptors) as explanatory variables, and dipeptide elution time was used as the objective valuable. For comparison, the simple correlation coefficient to the dipeptide elution time was also analyzed. The performances of the LASSO model were evaluated via leave-one-out cross-validation. The prediction performances of the LASSO models were evaluated by two indices: the coefficient of determination (R²) and the RMSE. All data handling and modeling were programmed by R (version 3.6.2) (https://www.r-project.org/).

**Declarations**

**DATA AVAILABILITY**

The data underlying this article are available in the article and in its online supplementary material.

**ACKNOWLEDGEMENTS**

The fermented cocoa beans were kindly provided from Mr. Katsunori Tanaka (Conche, bean-to-bar chocolate shop). This research was funded by the Japan Society for the Promotion of Science (KAKENHI grant JP 21H02144A0 and 21K19092 to K.I., and 19K15792 and 22492279 to Y.T.) and Fuji Foundation for Protein Research (K.I.).

**AUTHOR CONTRIBUTIONS**

N., S. F., T. S., E. S., H. M., K. T., and K. I.; In silico analysis, T. H., M. F., K. K., R. K., and K. I. All authors have read and agreed to the published version of the manuscript.

COMPETING INTERESTS

The authors declare no conflict of interest.

References


**Table 1**

Table 1 is available in the Supplementary Files section.

**Figures**
Figure 1

The analytical procedure for dipeptides. **a** Derivatization reaction of dipeptides with DMT-\((S)\)-Pro-OSu and their cleavage in the UPLC-ESI/MS/MS analysis. A reagent ion \((m/z\) 209.1\) was generated by dissociating the DMT-\((S)\)-Pro-OSu moiety. **b** The MS/MS spectra of Ala–His was used as a representative dipeptide. \(m/z\) 209.1 was the ion derived from the DMT-\((S)\)-Pro-OSu moiety. \(m/z\) 280.1 is an a-type immonium ion fragmented between an a-carbon and the carboxyl group of the amino acid on the N-terminal side.
Figure 2

**Precursor ion and elution time for 313 dipeptides.** Precursor ion m/z (a) and elution time (b) were optimized for each of the 313 dipeptides. Gray: not analyzed.
Figure 3

Separation and identification of dipeptides with the same precursor ion. Example of dipeptides with the same precursor ion, Val–Ile and Val–Leu. Cleavage in UPLC-MS/MS and the extracted SRM chromatograph are presented.
Figure 4

Dipeptide content in the fermented cocoa beans from different geographical origins. Dipeptide content (nmol/g) in the fermented cocoa beans cultivated in Brazil (sky blue), Ghana (orange), Grenada (green), and Haiti (pink). Each data is expressed as mean ± SE (n = 3).
Figure 5

**Prediction performance of the elution time prediction model.** Correlation plot of the actual elution time and the predicted elution time from the elution time prediction model trained with 23 molecular descriptors. X-axis, experimentally determined elution time; Y-axis, predicted elution time from the dipeptide sequence. Each plot represents single dipeptide prediction. In total, data for 313 dipeptides are plotted. The gray zone indicates the distance of the standard deviation value of experimentally determined elution times of 313 dipeptides to determine the level of prediction error compared with the deviations in the experimental data.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- SupplFigurePDF.pdf
- SupplTable1.xlsx
- SupplTable2.xlsx
- SupplTable3.xlsx
- SupplTable4.xlsx
• Table.xlsx