Machine Learning Screening of Bile Acid-Binding Peptides in APeptide Database Derived From Food Proteins

Kento Imai

Department of Biomolecular Engineering, Graduate School of Engineering, Nagoya University, Nagoya 464-8603, Japan

Kazunori Shimizu

Department of Biomolecular Engineering, Graduate School of Engineering, Nagoya University, Nagoya 464-8603, Japan

Hiroyuki Honda (✉ honda@chembio.nagoya-u.ac.jp)

Department of Biomolecular Engineering, Graduate School of Engineering, Nagoya University, Nagoya 464-8603, Japan

Research Article

Keywords: Edible protein, Peptide database, Peptide screening, Informatics, Bioactive peptides

Posted Date: February 19th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-209326/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Scientific Reports on August 9th, 2021. See the published version at <https://doi.org/10.1038/s41598-021-95461-1>.

Abstract

In the present study, a new bioactive peptide screening method was developed using BIOPEP-UWM and machine learning. Training data was initially obtained using high-throughput techniques, and positive and negative datasets were generated. The predictive model was generated by calculating the explanatory variables of the peptides. To understand both site-specific and global characteristics, amino acid features (for site-specific characteristics), and peptide features (for global characteristics) were generated. The constructed models were applied to the peptide database generated using BIOPEP-UWM, and bioactivity was predicted to explore candidate bile acid-binding peptides. Using this strategy, seven novel bile acid-binding peptides (VFWM, QRIFW, RVWVQ, LIRYTK, NGDEPL, PTFTRKL, and KISQRYQ) were identified. Our novel screening method could be easily applied to industrial applications using whole edible proteins. The proposed approach would be useful for identifying bile acid-binding peptides, as well as other bioactive peptides, as long as a large amount of training data can be obtained.

Introduction

Bioactive peptides are protein fragments that have positive benefits in humans¹, and are therefore promising candidates for cosmetics and health food. Many bioactive peptides have been identified to date. For example, the alpha-casein-derived peptides RYLGY, AYFYPEL, and YQKFPQY have angiotensin-converting enzyme (ACE)-inhibitory activity², and the beta-lactoglobulin-derived peptides VAGTWY, AASDISLLDAQSAPLR, IPAVFK, and VLVLDTDYK have bactericidal activity³.

The current classical approach to peptide screening has remained unchanged for many years. In the classical approach, source proteins are hydrolyzed with proteases, then separated and purified. Candidate bioactive peptides are identified by liquid chromatography-mass spectrometry (LC-MS), and the bioactivity of the candidate peptides is confirmed using synthetic peptides⁴. This approach is time consuming, and often results in extremely low yields of peptides due to the requirement for fractionation and concentration prior to LC-MS. In addition, in many cases the initial proteolytic mixture is prepared based on the specific interests of researchers and industries, with no *a priori* knowledge, and no guarantee that the desired bioactive peptides are present. The classical approach therefore has a significant 'trial and error' element, potentially leading to wasted time and money.

To circumvent these limitations and accelerate this process, *in silico* approaches for identifying novel bioactive peptides have been proposed⁵. *In silico* approaches make use of peptide databases containing sequences derived from proteins of interest, and implement bioinformatic tools to predict bioactivity. Many peptide databases have been developed, including the useful database BIOPEP-UWM⁶, which stores bioactive peptides along with edible proteins, allergenic proteins with their epitopes and sensory peptides, and amino acids. In addition, it implements some predictive tools, including the theoretical degree of hydrolysis and bioactivity prediction. Using the BIOPEP-UWM database, the appropriate fraction of DPP-4 (dipeptidyl peptidase-4) inhibiting peptides derived from mealworms (*Tenebrio molitor*) was selected⁷. This approach has also been adopted in the pigeon pea (*Cajanus cajan*)⁸. Recent studies have

shown that the combination of databases with advanced machine learning-based bioinformatics tools is a promising approach for screening and developing novel bioactive peptides. For example, Meher *et al.*⁹ created an antimicrobial peptide by using predictive models with support vector machine (SVM) algorithms and antimicrobial databases CAMP¹⁰, APD3¹¹ and AntiBP2¹². Gautam *et al.* predicted cell penetrating activity by using SVM and novel databases¹³, and achieved a maximum accuracy of 97.4 %.

In the present study, a novel strategy to screen bioactive peptides derived from edible proteins was developed using BIOPEP-UWM and machine learning. This strategy allows for the prediction of all bioactive peptides derived from edible proteins with a suitable positive and negative training dataset. The experimental workflow is summarized in Fig. 1. We made use of training data obtained with a high-throughput peptide array to generate positive and negative datasets. The predictive model was generated using the explanatory variables of peptides in these datasets. Finally, this model was applied to a peptide database derived from edible proteins using BIOPEP-UWM. We tested our peptide screening tool by searching for bile acid-binding peptides. In humans, cholesterol absorption occurs in the proximal jejunum of the small intestine, where both dietary cholesterol and biliary cholesterol are available for uptake from the intestinal lumen via bile acid micelles^{14, 15}. Bile acid-binding peptides interact with bile acids that form the micelles, and subsequently disrupt the micelles, contributing to the suppression of intestinal cholesterol absorption. We have previously designed bile acid-binding peptides with informatics approach^{16–19}. However, the designed peptides were not found in storage proteins or protein sources and proteases were selected based on our interests. Bile acid-binding peptides work on the intestinal tract, and we therefore do not need to consider their absorption from the small intestine when developing novel health foods. Using our novel approach, we have established the framework for rapid and cost-effective screening for bioactive peptides, which may be applied to the development of new health-promoting products.

Results And Discussion

Measurement of bile acid binding in a synthetic peptide array

Training data is essential for the construction of the classification model. To generate training data, 460 4-, 5-, 6-, and 7-mer peptides were generated in a peptide array, and their bile acid binding activities were evaluated. The sequences and fluorescent intensities are shown in Table S2 and Figure S1. The fluorescence intensity of 4-mers was relatively lower than that of longer peptides (Fig. S1(A)). The observed low intensity of 4-mers in the training data may be due to the relatively low hydrophobicity of 4-mer peptides. Using the peptide array data, 150 peptides with the highest fluorescent intensities were defined as the 'positive' dataset, and 150 peptides with the lowest fluorescent intensities were defined as a 'negative' dataset for bile acid binding activity. The average fluorescence intensities of the positive and negative datasets are shown in Table 1. Since there was a significant difference between the two datasets ($P < 0.001$), the randomly designed peptide library contained peptides with different bile acid binding bioactivities.

Construction of predictive model and evaluation of model performance

To construct the predictive model, the peptide features of 300 peptides (positive = 150, negative = 150) were calculated. For each 4-mer, 56 features were generated (28 amino acid features and 28 global features), for each 5-mer 63 (35 amino acid features and 28 global features), 6-mer 70 (42 amino acid features and 28 global features), and 7-mer 77 (49 amino acid features and 28 global features), and used as explanatory variables. Three algorithms were used to construct the predictive model (SVM, RF, LR), and the model performance was evaluated by comparing accuracy, precision, and recall. The peptides with a probability of >0.5 were designated as positive, and those with a probability of <0.5 were designated as negative for bile acid binding ability. Except for the precision scores of 5- and 7-mers, all RF scores were the highest out of the 3 tested algorithms (Table 2). RF was therefore selected for the predictive algorithm.

The scores 4-mer peptides were lower than the scores of longer peptides (Table 2). The ratio of the average fluorescence intensity of positive the dataset and that of the negative dataset was defined as the P/N intensity ratio. In Table 1, the P/N intensity ratio of 4-mers (2.67) was lower than that of longer peptides (3.63 for 5-mers, 4.11 for 6-mers, 3.87 for 7-mers). This is caused by the relatively lower overall fluorescence intensity of the 4-mer training data. The model performance was roughly correlated with the P/N intensity ratio. The reason for the poor performance is the relatively large number of FPs and FNs predicted by the acquired model when the P/N intensity ratio is low.

To investigate the importance of the input features, the variable importance was estimated according to the increase in the predictive error due to the permutation of out-of-bag data for the given variable. The importance of each of the input variables is shown in Table S3. Most of the top 10 selected features referred to global features of peptides, namely av, sd, min, max, with the exception of two specific features: residue2_Molecular_weight for 4-mers and residue1_Isoelectric_point for 7-mers. In addition, two features for 4-mers, four features for 5-mers, four features for 6-mers, and five features for 7-mers were related to the peptide isoelectric point. Similarly, five features for 4-mers, three features for 5-mers, two features for 6-mers, and two features for 7-mers were related to molecular weight. This suggests that the global peptide features are more important than the site-specific features for bile acid binding activity in peptides of 4-7 amino acids. Bile acid molecules are amphiphilic, with a hydrophobic steroid core and hydrophilic hydroxyl groups, and therefore have strong surfactant action. Since peptide binding can occur in either direction with bile acids, site-specific peptide features may be less important.

Features referring to isoelectric point and molecular weight were among the most important in Table S3. This suggests that peptides with high isoelectric points or high molecular weights bind to strongly to bile acid. The five amino acids with the highest isoelectric points are R, K, H, P, and I²⁰, and the top five for molecular weight are W, Y, R, F, and H²¹. The basic or aromatic peptides therefore have higher binding activity against bile acids. Some studies have investigated the binding mechanisms between bile acids and other compounds, such as sterols and nisin^{22, 23, 24, 25}, and revealed that hydrophobic amino acids,

especially aromatic amino acids, interact with bile acid micelles. These findings are in agreement with the top 10 features identified in Table S3.

Construction of edible peptide database and prediction of bile acid binding activities

A set of 710 edible proteins were obtained from BIOPEP-UWM and digested using all available predicted protease binding sites (Table 3), resulting in 199568 4-mers, 198808 5-mers, 198055 6-mers, and 197310 7-mers. After removing duplicate sequences, the dataset contained 56171 4-mers, 89663 5-mers, 98387 6-mers, and 102805 7-mers. A total dataset of approximately 350000 peptide sequences was thus generated.

The RF model was applied to the peptide datasets, and the results are shown in Table S4. To verify the RF model, the results were sorted by the probability of being classified as positive for bile acid binding activity. Fifty peptides from the top and 50 peptides from the bottom of the probability list were synthesized and their bile acid binding activities were determined using a peptide array. The synthesized sequences are listed in Table S5, and their fluorescence intensities are shown in Figure 2. The average fluorescence intensity of positive peptides was higher than that of negative peptides ($P < 0.001$), indicating that the RF model could successfully predict bile acid binding activity. The details of the peptides are shown in Table S6.

Novel bile acid binding peptides from edible proteins

The top five peptides, ranked by fluorescence intensity in a peptide array for bile acid binding, are shown in Table 4. Seven of the peptides with the highest scores for bile acid binding activity mapped to storage proteins in the database: VFWM from legumin A (*Pisum sativum*)²⁶, QRIFW from high molecular weight glutenin (*Triticum aestivum*)²⁷, RVWVQ from profilin-1 (*Hordeum vulgare*)²⁸, LIRYTK from serum albumin (*Gallus gallus*)²⁸, NGDEPL from legumin chain B fragment (*Vicia faba*)²⁹, PTFTRKL from chicken connectin (titin) fragment (*Gallus gallus*)²⁸, and KISQRYQ from alpha-S2-casein (*Bos taurus*)²⁸. NGDEPL was predicted to have low affinity for bile acid; however, it had a high bile acid binding activity according to the peptide array. The mechanisms underlying this apparent contradiction are unclear, but this peptide might bind stereospecifically to bile acids. Since storage proteins are favorable for the manufacture of health foods and cosmetics, these protein sources are expected to contain novel bioactive components.

Most of the predicted bioactive peptides in the present dataset were obtained by proteolysis by enzymes from plants or microorganisms, and proteolysis by gastrointestinal enzymes³⁰. Therefore, to evaluate the utility of these peptides at the industrial scale, we examined whether the seven peptides derived from storage proteins could be generated using peptidases or proteases. As a result, KISQRYQ was predicted to be generated from alpha-S2-casein (*Bos taurus*) with peptidyl-Lys metalloendopeptidase (*Armillaria mellea* neutral proteinase). Gutiez *et al.* had previously investigated the relationship between the autolysis caused by lactic acid bacteria and the production of angiotensin-converting enzyme (ACE)-inhibitory peptides, and reported that KISQRYQ was generated from skimmed milk (alpha-S2-casein) by

Lactococcus lactis subsp. lactis IL 1403³¹. Taken together, this suggests that KISQRYQ could be a candidate bioactive peptide for health food.

In the present study, a new bioactive peptide screening method was developed based on a synthetic peptide library for bile acid binding and machine learning. A database containing peptide sequences derived from edible proteins was developed to identify peptides with features that are associated with bile acid binding. Combining these two tools, novel bile acid-binding candidate peptides have been discovered. Among the peptides with the highest predicted scores for bile acid binding activity, seven (VFWM, QRIFW, RVWVQ, LIRYTK, NGDEPL, PTFTRKL, and KISQRYQ) were derived from storage proteins. Among them, KISQRYQ was predicted to be generated from alpha-S2-casein (*Bos taurus*) with peptidyl-Lys metalloendopeptidase (*Armillaria mellea* neutral proteinase) or from skim milk with *Lactococcus lactis subsp. lactis* IL 1403. Our novel method could successfully screen bioactive peptides, and can easily be applied to industrial applications based on whole edible proteins. The proposed approach would be useful for bile acid-binding peptides, as well for other bioactive peptides, as long as a large amount of training data could be obtained.

Materials And Methods

Materials

Fmoc amino acid OH was purchased from Watanabe Chemical Industries, Ltd. (Japan). BSA was purchased from Fujifilm Wako Pure Chemical Corporation (Japan). Taurocholic acid (T-4009) was purchased from Sigma-Aldrich (USA). Anti-cholic acid antibody (FKA502) was purchased from Cosmo Bio (Japan). Anti-rabbit IgG conjugated Alexa 488 (ab150077) antibody was purchased from Abcam (UK).

Synthetic peptide array generation and bile acid binding assay

To generate positive and negative peptide training datasets for our machine learning algorithm, we synthesized 460 4, 5, 6, 7-mer peptides that were randomly generated using R software (version 3.5.3) (R development Core Team, <https://www.r-project.org/>). All peptides were synthesized on a cellulose membrane with a spot synthesizer (Intervis, ASP222, Cologne, Germany) as previously reported³². Fmoc-aund-OH was introduced at the C-terminal end of the peptides as a spacer. After synthesizing, the side-chain-protecting groups of the Fmoc amino acid were removed with trifluoroacetic acid. The membrane was washed thoroughly with diethyl ether and methanol and dried. The membrane was soaked in PBS for 24 h, then transferred into 1% BSA in PBS solution at 37°C at 12 h before the commencement of the assay. A bile acid binding assay was conducted according to a previous study¹⁶. After washing with PBS, 10 µg/mL taurocholic acid dissolved in PBS was added to the arrays and incubated for 1 h. After washing with PBS, anti-cholic acid antibody dissolved in 0.25% BSA was added to the array and incubated for 1 h at 37°C. After washing with TBS containing 0.05 % Tween 20, 2 µg/mL of anti-rabbit IgG conjugated Alexa 488 dissolved in PBS was added and incubated for 1 h at 37°C. After washing with

TBS, peptide spots were fluorescently detected with a fluorescent imager (Typhoon FLA-7000, GE Healthcare Japan Life Sciences, Tokyo, Japan). The scanned images were quantified using Image Quant TL (GE Healthcare Japan Life Sciences, Tokyo, Japan). Average fluorescence intensities were calculated by subtracting the peptide array treated only with the secondary antibody from the triplicate fluorescence intensities of the same peptide sequence.

Feature generation

Seven features were considered for the prediction of bile acid binding activity (Table S1). General physicochemical features of peptides were described by pI^{20} , polarity²⁰, hydrophobicity³³, and molecular weight²¹, while structural features were described by Ph (the index about helix) and Pt (the index about turn). Xia *et al.* investigated the existence of amino acids in secondary structures and defined the new indices, Ph, Ps (the index about sheet), and Pt³⁴. The correlation coefficient between Ph and Ps is $> |0.98|$; therefore, Ps was excluded from the feature index in this research. In addition, previous research has revealed that hydrophobic amino acids, especially aromatic ones, interact with bile acid micelles^{15, 22, 23, 35}, so the number of aromatic amino acids was included as a peptide feature. Based on these features, the global features of the library peptides were generated. For example, in the case of 4-mer peptides, each amino acid has seven features (Table S1), so 28 amino acid features were generated for each 4-mer peptide. In addition, four global values, the maximum, minimum, average, and standard deviation (sd) were generated for each peptide. This means that a total of 56 features (28 amino acid features and 28 global features) were generated and used as explanatory variables for each 4-mer peptide. All features were calculated in R.

Construction of prediction models

To construct the prediction model, three algorithms were used in Python: support vector machine (SVM), random forest (RF)³⁶, and logistic regression (LR). Scikit-learn libraries³⁷ were adopted and leave-one-out cross validation (LOOCV) was imported into Python. The parameters for the algorithms were set as follows: In the SVM (linear) model, the default value of the parameter cost ($C = 1$) was used. In the RF model, the number of trees to grow (n_{tree}) were set at 100 or 500, and the number of variables randomly sampled as candidates at each split (m_{try}) was set to "auto". In the LR model, the penalty was set to "lasso," C was set to 10 or 50, and the maximum number of iterations taken for the solvers to converge (max_iter) was set to 100. The probability of binding to bile acid was calculated for all peptides and classified on the basis of the score = 0.5.

The performance of all three machine learning models was evaluated using 3 metrics:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{Precision} = (TP) / (TP + FP)$$

$$\text{Recall} = (TP) / (TP + FN)$$

TP, true positive; TN, true negative; FP, false positive; FN, false negative.

Generation of peptide database for edible proteins

A total of 710 protein sequences were obtained from BIOPEP-UWM, available at <http://www.uwm.edu.pl/biochemia/index.php/pL/biopep> (accessed in October 2018)⁶. Peptides were generated based on predicted enzyme cleavage sites obtained from PeptideCutter, available at https://web.expasy.org/peptide_cutter/. All possible enzymes were used to generate theoretical bioactive peptides based on the database. The resulting peptide sequences were sorted into 4-, 5-, 6-, and 7-mer peptides with 1 residue shift from the N-terminal amino acid using R software. The peptide database was generated in csv format.

Statistical analysis

Data are presented as the mean $\pm$ standard deviation (SD). Student's t-test was used for between-group comparisons. A p value <0.05 was considered to be statistically significant.

Declarations

Corresponding Author

*Hiroyuki Honda

Acknowledgements

We would like to thank Editage for English language editing.

Funding Sources

This work was supported by JSPS KAKENHI (Grant Numbers: JP19H00837, JP20J10655).

Notes

The authors declare no competing interests.

Author Contributions

K.I. designed and performed the experiments. K.I, K.S. and H.H. conceived the experiments and wrote the manuscript.

References

1. Bhandari, D. *et al.* A Review on Bioactive Peptides: Physiological Functions, Bioavailability and Safety. *Int. J. Pept. Res. Ther.* **26**, 139–150 (2020).

2. Contreras, M. M., Carrón, R., Montero, M. J., Ramos, M. & Recio, I. Novel casein-derived peptides with antihypertensive activity. *Int. Dairy J.* **19**, 566–573 (2009).
3. Pellegrini, A., Dettling, C., Thomas, U. & Hunziker, P. Isolation and characterization of four bactericidal domains in the bovine β -lactoglobulin. *Biochim. Biophys. Acta.* **1526**, 131–140 (2001).
4. Udenigwe, C. C. Bioinformatics approaches, prospects and challenges of food bioactive peptide research. *Trends Food Sci. Technol.* **36**, 137–143 (2014).
5. Agyei, D., Tsopmo, A. & Udenigwe, C. C. Bioinformatics and peptidomics approaches to the discovery and analysis of food-derived bioactive peptides. *Anal. Bioanal. Chem.* **410**, 3463–3472 (2018).
6. Minkiewicz, I. & Darewicz BIOPEP-UWM Database of Bioactive Peptides: Current Opportunities. *Int. J. Mol. Sci.* **20**, 5978 (2019).
7. Dávalos Terán, I., Imai, K., Lacroix, I. M. E., Fogliano, V. & Udenigwe, C. C. Bioinformatics of edible yellow mealworm (*Tenebrio molitor*) proteome reveal the cuticular proteins as promising precursors of dipeptidyl peptidase-IV inhibitors. *J. Food Biochem.* **44**, e13121 (2020).
8. Boachie, R. T. *et al.* Enzymatic release of dipeptidyl peptidase-4 inhibitors (gliptins) from pigeon pea (*Cajanus cajan*) nutrient reservoir proteins: In silico and in vitro assessments. *J. Food Biochem.* **43**, e13071 (2019).
9. Meher, P. K., Sahu, T. K., Saini, V. & Rao, A. R. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci. Rep.* **7**, 42362 (2017).
10. Yeaman, M. R. & Yount, N. Y. Mechanisms of antimicrobial peptide action and resistance. *Pharmacol. Rev.* **55**, 27–55 (2003).
11. Wang, G., Li, X. & Wang, Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **44**, D1087–D1093 (2016).
12. Lata, S., Mishra, N. K. & Raghava, G. P. S. AntiBP2: Improved version of antibacterial peptide prediction. *BMC Bioinformatics.* **11**, S19 (2010).
13. Gautam, A. *et al.* In silico approaches for designing highly effective cell penetrating peptides. *J. Transl. Med.* **11**, 74 (2013).
14. Altmann, S. W. *et al.* Niemann-Pick C1 Like 1 Protein Is Critical for Intestinal Cholesterol Absorption. *Science (80-)*. **303**, 1201–1204 (2004).
15. Boachie, R., Yao, S. & Udenigwe, C. C. Molecular mechanisms of cholesterol-lowering peptides derived from food proteins. *Curr. Opin. Food Sci.* **20**, 58–63 (2018).
16. Imai, K., Shimizu, K. & Honda, H. Predictive selection and evaluation of appropriate functional peptides for intestinal delivery with a porous silica gel. *J. Biosci. Bioeng.* **128**, 44–49 (2019).
17. Ito, M., Shimizu, K. & Honda, H. Searching for high-binding peptides to bile acid for inhibition of intestinal cholesterol absorption using principal component analysis. *J. Biosci. Bioeng.* **127**, 366–371 (2019).

18. Ito, M., Shimizu, K. & Honda, H. Bile acid micelle disruption activity of short-chain peptides from tryptic hydrolyzate of edible proteins. *J. Biosci. Bioeng.* **130**, 514–519 (2020).
19. Takeshita, T. *et al.* Screening of peptides with a high affinity to bile acids using peptide arrays and a computational analysis. *J. Biosci. Bioeng.* **112**, 92–97 (2011).
20. Zimmerman, J. M., Eliezer, N. & Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **21**, 170–201 (1968).
21. Zamyatnin, A. A. Protein volume in solution. *Prog. Biophys. Mol. Biol.* **24**, 107–123 (1972).
22. Matsuoka, K. *et al.* NMR Study on Solubilization of Sterols and Aromatic Compounds in Sodium Taurodeoxycholate Micelles. *Bull. Chem. Soc. Jpn.* **80**, 2334–2341 (2007).
23. Dominguez, C. *et al.* Interactions of bile salt micelles and colipase studied through intermolecular nOes. *FEBS Lett.* **482**, 109–112 (2000).
24. Acquah, C., Di Stefano, E. & Udenigwe, C. C. Role of hydrophobicity in food peptide functionality and bioactivity. *J. Food Bioact.* **4**, 88–98 (2018).
25. Gough, R. *et al.* Simulated gastrointestinal digestion of nisin and interaction between nisin and bile. *LWT.* **86**, 530–537 (2017).
26. Lycett, G. W., Croy, R. R. D., Shirsat, A. H. & Boulter, D. The complete nucleotide sequence of a legumin gene from pea (*Pisum sativum* L.). *Nucleic Acids Res.* **12**, 4493–4506 (1984).
27. Anderson, O. D. *et al.* Nucleotide sequences of the two high-molecular-weight glutenin genes from the D-genome of a hexaploid bread wheat, *Triticum aestivum* L. cv cheyenne. *Nucleic Acids Res.* **17**, 461–462 (1989).
28. UniProt. <https://www.uniprot.org/> accessed 03/02/2021
29. Heim, U., Schubert, R., Bäumllein, H. & Wobus, U. The legumin gene family: structure and evolutionary implications of *Vicia faba* B-type genes and pseudogenes. *Plant Mol. Biol.* **13**, 653–663 (1989).
30. Agyei, D., Ongkudon, C. M., Wei, C. Y., Chan, A. S. & Danquah, M. K. Bioprocess challenges to the isolation and purification of bioactive peptides. *Food Bioprod. Process.* **98**, 244–256 (2016).
31. Gútiérrez, L. *et al.* Controlled enterolysin A-mediated lysis and production of angiotensin converting enzyme-inhibitory bovine skim milk hydrolysates by recombinant *Lactococcus lactis*. *Int. Dairy J.* **34**, 100–103 (2014).
32. Kozaki, I., Shimizu, K. & Honda, H. Effective modification of cell death-inducing intracellular peptides by means of a photo-cleavable peptide array-based screening system. *J. Biosci. Bioeng.* **124**, 209–214 (2017).
33. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
34. Xia, X., Xie, Z. P. & Structure Neighbor Effect, and a New Index of Amino Acid Dissimilarities. *Mol. Biol. Evol.* **19**, 58–67 (2002).
35. Hermoso, J. *et al.* Neutron crystallographic evidence of lipase-colipase complex activation by a micelle. *EMBO J.* **16**, 5531–5536 (1997).

36. Leo, B. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
37. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Tables

Due to technical limitations, table 1,2,3,4 is only available as a download in the Supplemental Files section.

Supplemental Information

Figure S1: The fluorescent intensities for the training datasets according to a peptide array to detect acid bile binding activity.

The intensities are shown for 4-mer (A), 5-mer (B), 6-mer (C), and 7-mer (D) synthetic peptides.

Table S1: Features used to construct the model.

Table S2: All florescence intensities used to generate training datasets.

Table S3: Selected features of the RF prediction model for bile acid binding activity.

Bold type: features related to molecular weight, underlined: features related to isoelectric point, **both bold type and underlined**: features related to ‘aromatic amino acids’.

Table S4: The number of peptides that were predicted as positive or negative for bile acid binding activity in the database.

Table S5: Sequences of peptides that were synthesized for the evaluation of the model.

P-positive and N-negative

Table S6: The details of the peptides synthesized in Table S5

Figures

Figure 1

Schematic showing experimental workflow Positive and negative datasets were generated from the training data. Subsequently, explanatory variables were generated with amino acid features (site-specific features) and peptide features (global features). Predictive models were constructed using a combination of the training data and explanatory variables. A new database containing peptides found in edible proteins was created using the edible protein database BIOPEP-UWM. Lastly, constructed models were applied to the edible protein database and the bioactivity of each peptide was predicted.

Figure 2

Bile acid binding activity of the top and bottom 50 peptides. In order to evaluate the 4-mer (A), 5-mer (B), 6-mer (C), and 7-mer (D) models, the 50 peptides predicted to have the most and least bile acid binding activity were synthesized, and their bile acid-binding activity was evaluated using a peptide array. In each group, 50 peptides were selected, with the 50 peptides with the most predicted activity designated as 'positive', and the 50 peptides with the least predicted activity designated as 'negative'.

Supplementary Files

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

- [Supportinginformation.docx](#)

- [Supportingtable.xlsx](#)
- [Table1.jpg](#)
- [Table2.jpg](#)
- [Table3.jpg](#)
- [Table4.jpg](#)