Trends in Food Science & Technology 114 (2021) 176–188

Contents lists available at ScienceDirect

Trends in Food Science & Technology

journal homepage: www.elsevier.com/locate/tifs

Application of quantitative structure-activity relationship to food-derived peptides: Methods, situations, challenges and prospects

Weichen Bo [a,1], Lang Chen [a,1], Dongya Qin [a], Sheng Geng [a], Jiaqi Li [a], Hu Mei [a], Bo Li [b,**], Guizhao Liang [a,*]

[a] *Key Laboratory of Biorheological Science and Technology, Ministry of Education, Bioengineering College, Chongqing University, Chongqing, 400044, China*
[b] *College of Life Sciences, Chongqing Normal University, Chongqing, 401331, China*

ARTICLE INFO

Keywords:
Quantitative structure-activity relationship (QSAR)
Bioinformatics
Peptide
Functional food
Modeling

ABSTRACT

Background: Food-derived bioactive peptides have attracted extensive attention because of their antioxidant, antibacterial, antitumor and antihypertensive effects. The conventional approaches used to acquire bioactive peptides require complicated procedures involving enzymolysis, separation and identification. So far, data-driven computing methods have become an important tool for the screening, design and mechanism exploration of bioactive peptides. The quantitative structure-activity relationship (QSAR), a quantitative method used to describe the structure-activity relationship of compounds, has *been* widely used in drug design, material science, and chemistry; however, there are limited applications in food science.

Scope and approach: Here, we mainly focus on technologies used to perform QSAR modeling in peptides, including dataset collection, structural characterization, variable selection, correlation *methods*, and model validation and evaluation. We also summarize the recent applications, situations, challenges and prospects of the use of QSAR in food-derived bioactive peptides.

Key findings and conclusions: Researchers should make full use of the benefits of QSAR as well as face its challenges. Multiple new methods or combination strategies should be used to achieve QSAR analysis. Much research is needed to improve the knowledge of QSAR in order to discover bioactive peptides. The solution to this task requires multidisciplinary cooperation in multiple fields, including chemistry, computer science, mathematics and, of course, food science.

1. Introduction

Bioactive peptides, as essential substances in maintaining life, are attracting wide attention due to their advantages of high activity and selectivity with few side effects (Mills, Stanton, Hill, & Ross, 2011). To date, a variety of functional peptides have been reported, including antihypertensive, antithrombotic, opioid, antimicrobial, antioxidant, anticancer, and immunomodulatory peptides (Daliri, Lee, & Oh, 2018). Food-derived proteolysis is commonly used to obtain bioactive peptides (Hernandez-Ledesma & Hsieh, 2017), which typically include 3–20 amino acid residues (Pihlanto-Leppälä, 2000). Accordingly, identifying such a large peptide library presents great challenges. It will be time- and consumable-wasting to identify numerous potential peptides by complicated experimental approaches (Udenigwe, 2014). Currently, there is an increasing trend to use *in silico* methods to design, discover and screen bioactive peptides. Compared with conventional methods, *in silico* methods (Fig. 1A), such as mathematical modeling, molecular docking and simulations, chemometrics and quantum-chemical calculations, exhibit significant advantages in obtaining food-derived bioactive peptides (Nongonierma & FitzGerald, 2017).

Herein, it should be mentioned that quantitative structure-activity relationship (QSAR) analysis has attracted attention in food science (FitzGerald, Cermeno, Khalesi, Kleekayai, & Amigo-Benavent, 2020; Nakai & Li-Chan, 1993; Pripp, Isaksson, Stepaniak, Sorhaug, & Ardo, 2005). The concept of QSAR is to quantitatively describe the relationship between the chemical structure and biological activity of the

[*] Corresponding author. Key Laboratory of Biorheological Science and Technology, Ministry of Education, Bioengineering College, Chongqing University, No.174 Shazhengjie, Shapingba, Chongqing, 400044, China.

[**] Corresponding author. College of Life Sciences, Chongqing Normal University, University Town, Shapingba District, Chongqing, 401331, China.

E-mail addresses: libcell@cqnu.edu.cn (B. Li), gzliang@cqu.edu.cn (G. Liang).

[1] These three authors contributed equally to the article.

https://doi.org/10.1016/j.tifs.2021.05.031
Received 21 July 2020; Received in revised form 22 May 2021; Accepted 26 May 2021

studied compounds (Holton, Vijayakumar, & Khaldi, 2013). QSAR methods have been used to design and screen new molecules as well as predict their activities and determine the mechanism of bioactive peptides. In recent years, a variety of methods derived from QSAR have emerged (Baczek & Kaliszan, 2009; Bahadori, Hemmateenejad, & Yousefinejad, 2019; Dearden, Cronin, & Kaiser, 2009; Doytchinova & Flower, 2001; Pal, Jana, Sural, & Chattaraj, 2019).

The earliest origin of QSAR can be traced back to a study on the structure-activity relationship of organic compounds in the 19th century (Kubinyi, 2002). The rapid development of QSAR was first attributed to the establishment of two 2D-QSAR methods in the early 1960s by Hansch-Fujita (Hansch, Maloney, Fujita, & Muir, 1962) and Free-Wilson (Free & Wilson, 1964). Subsequently, a representative 3D-QSAR method, comparative molecular field analysis (CoMFA), was proposed and widely applied in the design and prediction of new molecules (Cramer, Patterson, & Bunce, 1988). Until now, QSAR methods have experienced great progress (Damale, Harke, Kalam Khan, Shinde, & Sangshetti, 2014), in a wide range of research areas outside of traditional QSAR boundaries including synthesis planning, nanotechnology, materials science, biomaterials, and clinical informatics (Muratov, Bajorath, Sheridan, Tetko, Filimonov, Poroikov, et al., 2020). QSAR methods have also been used to explore the structure-activity relationship of food-derived chemicals. For instance, the interaction mechanism of flavonoids and zein in ethanol-water solution was explained by the 3D-QSAR and spectrofluorimetry (Yue et al., 2019). In Bouarab-Chibane's study, QSAR analysis was used to predict the antibacterial properties of polyphenols and other phenolic compounds, such as phenolic acid, flavonoids, hydroquinone, coumarin and quinone (Bouarab-Chibane, Forquet, Lanteri, Clement, Leonard-Akkari, Oulahal, et al., 2019). Another work established a QSAR model to predict hepatic

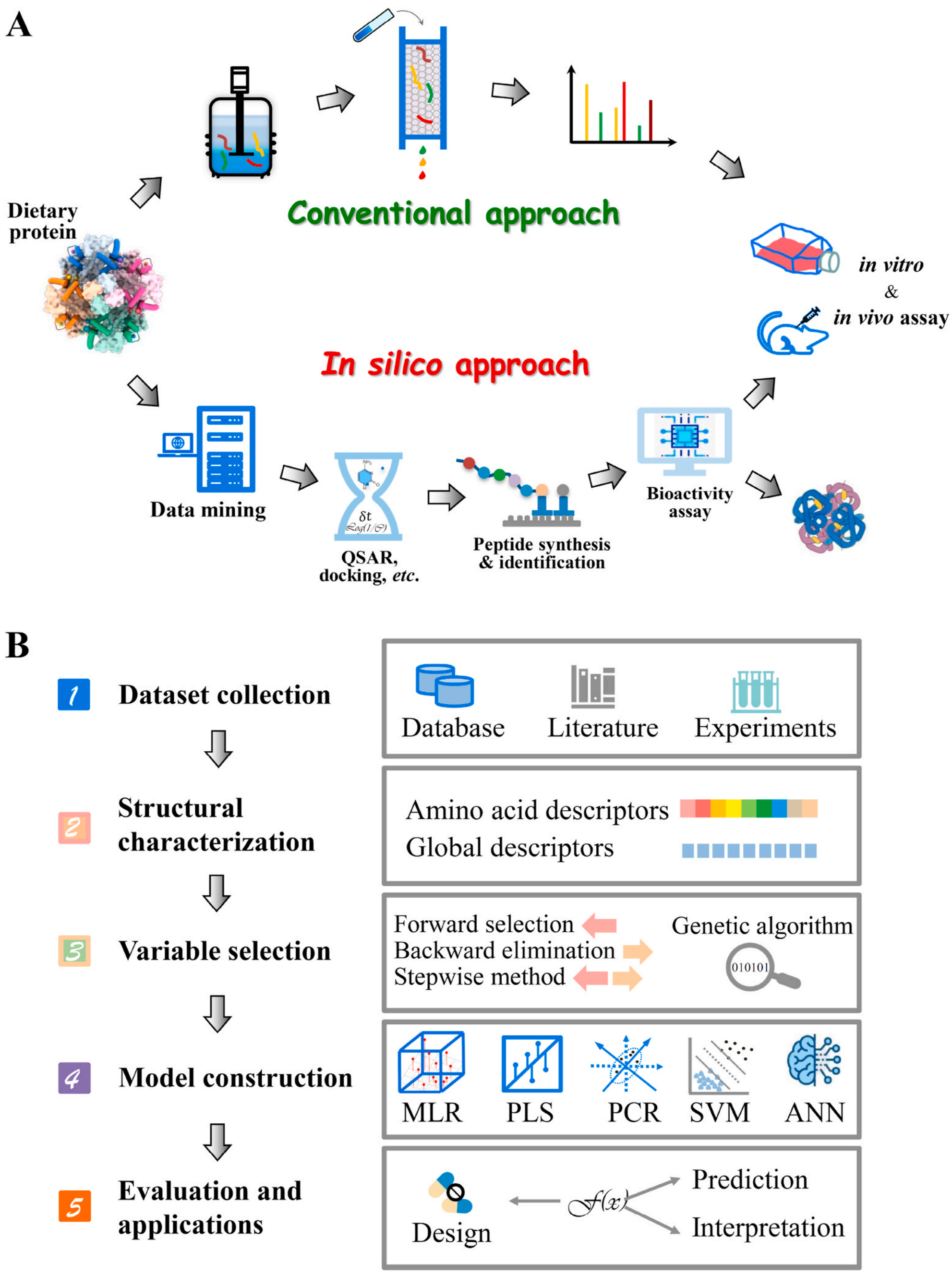

Fig. 1. Active peptide discovery strategy (Fig. 1A) and modeling process (Fig. 1B, take QSAR as an example).

steatosis, and the accuracy of the model reached 70% (Cotterill, Price, Rorije, & Peijnenburg, 2020). In addition, a framework involving the structure-activity relationship and classification model was proposed to identify potential food sweeteners from a vast database of natural molecules (Goel, Gajula, Gupta, & Rai, 2021).

There are some unsolved problems in QSAR modeling. For exmaple, QSAR methods fail to take account of data heterogeneity and define applicability domain in all conditions (Dearden et al., 2009). Moreover, it is not possible to provide a mechanistic interpretation of QSAR model at any time. In addition, conformational flexibility of chemical structures is still a problem for further research (Gasteiger, 2014). However, to be sure, QSAR methods are still of great help to researchers in various fields. In recent years, although QSAR, mainly 2D-QSAR and 3D-QSAR, has been applied in food science to bioactive peptides, extensive attention and a deep understanding of how to apply QSAR to bioactive peptides are still needed. The earlier review on above topic introduced some common cheminformatics approaches (including QSAR) and their application in the analysis of biologically active peptides from food sources (Iwaniak, Minkiewicz, Darewicz, Protasiewicz, & Mogut, 2015). Nongonierma et al. reviewed the strategies for the discovery and identification of milk protein-derived peptides (Nongonierma & FitzGerald, 2016a) and food protein-derived peptides (Nongonierma & FitzGerald, 2017). Next, another review provided an overview of research progress in the bioinformatics methods applied to characterizing, identifying, and elaborating bioactive mechanisms, and producing bioactive peptides, thus presenting an effective workflow as well (Tu, Cheng, Lu, & Du, 2018). After that, Blanco-Miguez et al. clarified that *in silico* prediction revealed the existence of potential bioactive neuropeptides produced by the human gut microbiota (Blanco-Miguez, Fdez-Riverola, Lourenco, & Sanchez, 2019). Also, Barati et al. compiled the recent analytical methods used for food-derived bioactive peptides, focusing on generation *in vitro*, *in vivo*, and *in silico*. For *in silico* methods (including proteome- and genome-based ones), emphasized the web resources and tools that help discover the bioactive peptides from food-derived proteins. (Barati, Javanmardi, Jazayeri, Jabbari, Rahmani, Barati, et al., 2020). These above-mentioned publications offered several new outlooks, for computational methods used for analyzing food-derived peptides. However, few of them comprehensively reviewed the QSAR modeling in peptides, including dataset collection, structural characterization, variable selection, model construction, and model validation and evaluation, and so on. Here, we conduct a comprehensive review to provide exhaustive discussion, systematic summation and rational advice on the application of QSAR in food-derived peptide studies.

2. Workflow of QSAR modeling of peptides

QSAR modeling involves multiple steps, such as dataset collection, structural characterization, variable selection, model construction, model validation and evaluation. Fig. 1B displays the detailed workflow of QSAR modeling applied to peptides. We will introduce the key issues in each step according to this workflow.

2.1. Dataset collection

The selection of the dataset is the first key step for QSAR modeling. A suitable dataset can largely ensure the application scope and prediction ability of a QSAR model. The source, structure or sequence, and endpoint values of the compounds used for modeling should be seriously considered. The datasets can be acquired from literature, databases, or one's own experimental results. It is recommended to try to avoid datasets from a different source or datasets that were determined by different protocols, which often leads to unconvincing modeling results (Dearden et al., 2009). If two peptides have the same sequences but different endpoint values, or if there are duplicate samples in the dataset, the modeling results will be adversely affected. Another key issue is the unit of the endpoint values. It is recommended to use molar concentration rather than weight concentration or dose to represent the endpoints (Gedeck, Rohde, & Bartels, 2006).

As one of the most crucial steps, data collection should take the subsequent modeling into account, too. For instance, to avoid the overfitting in QSAR modeling, the balance of sample size between the positive and negative groups should be considered. Moreover, there are several issues to be thought in this process, such as data integrity and heterogeneity.

2.2. Structural characterization

Molecular structural characterization is one of the key points in QSAR modeling. Here, structural characterization is used to describe the structural characteristics of the compounds using chemometric methods. The resulting parameters are called descriptors. Commonly, global descriptors and local descriptors (also known as amino acid descriptors) are used to characterize the structure of peptides. Global descriptors are molecular descriptors that characterize a compound as a whole. For instance, the descriptors like volume and polar surface area characterize a compound in a wholistic fashion, thus they are called global descriptors of molecules. Global descriptors have been utilized to computationally predict the potential of given compounds, such as hERG channel inhibitors (Sinha & Sen, 2011), and bioactivity scores and chemical reactivity properties (Flores-Holguin, Frau, & Glossman-Mitnik, 2019). Meanwhile, some programs have used global descriptors for prediction of bioactive peptides, such as ADMETSar (Yang, Lou, Sun, Li, Cai, Wang, et al., 2019) and ADMETLab (Jie, Ning-Ning, Zhi-Jiang, Lin, Yan, Defang, et al., 2018). Both above programs used the structural parameters of compounds annotated in SMILES code, but not amino acid sequences.

The basic idea of amino acid descriptors is to quantitatively describe the residues in peptides, thereby converting the amino acid sequence into a matrix-vector of structural descriptors. Table 1 summarizes the representative amino acid descriptors (the meaning of the descriptors can be found in Tables S1–S16 in the supplementary data). We highlight a few common descriptors as follows. A set of commonly used descriptors of peptides in QSAR are the "z-scales", which are scales of hydrophilicity and bulk and electronic properties determined by principal component analysis (PCA) based on 29 physicochemical variables of 20 coded amino acids (Hellberg, Sjostrom, Skagerberg, & Wold, 1987). A *web of science* search reveals that the z-scales-based method has been cited more than 400 times. Later, Sandberg et al. characterized the structures of 87 amino acids using 5z-scales (Sandberg, Eriksson, Jonsson, Sjostrom, & Wold, 1998), which incorporate lipophilic, steric, electronic properties and other properties derived from PCA, with 26 physicochemical variables. The ISA-ECI descriptor was proposed to characterize the isotropic surface area (ISA) of the amino acid side chain and the electronic charge index (ECI) of all the atoms in the side chain, which made it easy to interpret the QSAR of the studied peptides (Collantes & Dunn, 1995). We performed factor analysis (FA) on 335 amino acid indices to acquire the factor analysis scales of generalized amino acid information (FASGAI) (Liang, Yang, Kang, Mei, & Li, 2009), involving hydrophobicity, alpha and turn propensities, bulk properties, compositional characteristics, local flexibility and electronic properties. FASGAI has been used for the QSAR of bitter dipeptides, ACE inhibitors, and cationic antimicrobial peptides (Liang & Li, 2007; Liang et al., 2009).

As we know, peptides are molecules with unique 3D structures. From this perspective, the structural characterization of peptides should reasonably reflect their 3D-characteristics. In this sense, global descriptors exhibit more advantages than amino acid descriptors. Here, we specifically wish to mention a 3D-QSAR method called CoMFA (Cramer et al., 1988), which uses molecular fields to study properties or activities at the molecular level. The 3D-QSAR model based on CoMFA can provide some insights into the key structural factors affecting the biological activity of the studied peptides through a contour map, which is helpful

Table 1
The representative amino acid descriptors.

No.	Descriptors	Acquisition method	Meaning
1	z-scales (Hellberg et al., 1991)	PCA	Hydrophilicity, bulk, and electronic properties
2	GRID (Cocchi & Johansson, 1993)	PCA	Interaction energies of 20 coded amino acids with six different probes
3	ISA-ECI (Collantes & Dunn, 1995)	Calculation	Isotropic surface area and Electronic charge index
4	MS-WHIM (Zaliani & Gancia, 1999)	PCA	3D electrostatic potential
5	VHSE (Mei, Liao, Zhou, & Li, 2005)	PCA	Principle components scores of hydrophobic, steric, and electronic properties
6	BLOSUM62 (Henikoff & Henikoff, 1992)	VARIMAX	Physicochemical and substitution matrix
7	FASGAI (Liang et al., 2009)	FA	Factor analysis scale of generalized amino acid information
8	SZOTT (Liang, Zhou, Zhou, Zhang, & Li, 2006)	PCA	A matrix of 1369 0D-3D structural variables for 20 coded amino acids
9	ST-SCALE (Yang et al., 2010)	PCA	Structural topological descriptors
10	NNAAIndex (Liang, Liu, Shi, Zhao, & Zheng, 2013)	FA	Index of natural and nonnatural amino acids
11	VSW (Tong, Liu, Zhou, Wu, & Li, 2008)	PCA	Principle components scores for weighted holistic invariant molecular index
12	HESH (Shu et al., 2009)	PCA	Hydrophobic, electronic, steric, and hydrogen
13	DPPS (Tian, Yang, Lv, Yang, & Zhou, 2009)	PCA	Divided physiochemical property scores
14	QTMS (Hemmateenejad, Yousefinejad, & Mehdipour, 2011)	PCA	Quantum topological molecular similarity
15	Lin's scales (Lin, Long, Bo, Wang, & Wu, 2008)	Calculation	van der Waal's volume, net charge index and hydrophobic parameter
16	ProtFP (van Westen et al., 2013)	PCA	Physicochemical parameters

when designing new peptides (Wu et al., 2014). However, for the 3D-CoMFA method, in the absence of a reliable 3D receptor-bound structure, a multitude of ligand alignment protocols with different parameter settings are required to yield different sets of lead compounds for increasing the prediction uncertainties (Cramer, 2003). A 3D modeling method similar to CoMFA, CoMSIA (comparative molecular similarity indices analysis) (Klebe, Abraham, & Mietzner, 1994), has also been used for 3D-QSAR analysis of peptides (Qi, Lin, Zhang, & Wu, 2017; Wu et al., 2014). The results obtained from CoMSIA were often more robust than those obtained from CoMFA (Hou, Li, Li, Liu, & Xu, 2000).

Due to relative complexity and high flexibility of the whole peptide molecule, the calculation or characterization of these 3D-features to obtain an X-matrix will face greater challenges. The activities of the peptides are related to the position, constitution and physicochemical properties of amino acids. Therefore, amino acid descriptors are generally used to describe the structural characteristics of the peptides. Amino acid descriptors can avoid the defects of global descriptors, which is the reason that various amino acid descriptors have been proposed in recent years. More importantly, amino acid descriptors can reflect how an amino acid at a certain site affects the structure and activity of the peptide, providing a reference for obtaining novel functional peptides by modifying specific sites.

Amino acid descriptors will generate variables with different numbers depending on the length of the studied peptides. For example, when amino acid descriptors, such as z-scales and FASGAI, are used to characterize the structures of the peptides, variables with different numbers will be generated, making modeling impossible. A data transformation method, auto-cross covariance (ACC) (Andersson, Sjöström, & Lundstedt, 1998), involving the autocovariance and cross-covariance calculated from sequential data, was proposed to generate variables with the same number for each sample, thus enabling the application of conventional modeling methods. ACC transformation has been adopted for QSAR models for a set of ACE inhibitors derived from milk-driven peptides (Bahadori et al., 2019).

2.3. *Variable selection*

The next key step after molecular characterization is variable selection, which is essential to ensure the robustness and adequate interpretation of a QSAR model. Currently, several variable selection methods have appeared in QSAR modeling. The detailed principles of the methods can be found in a recent review (Shahlaei, 2013). Here, we provide a brief overview of several representative methods involving forward selection, backward elimination, the stepwise method, and the genetic algorithm (GA). In forward selection, which is an "in but not out" algorithm, a variable with a significant impact on dependent variables will be introduced until a new variable cannot be introduced. Backward elimination is an "out but not in" algorithm, i.e., in the equation containing all variables, each variable without significant impact on the dependent variables is eliminated until none of the independent variables can be eliminated. The stepwise method simultaneously carries out a forward selection process and backward elimination (Smith, Gill, & Hammond, 1985) and is therefore an effective way to find the optimal subspace.

GA is a popular method for variable selection, which imitates the natural selection and natural genetic mechanism of biology (Niazi & Leardi, 2012). According to the principles of "survival competition" and "survival of the fittest", it gradually approaches the optimal solution after generations of evolution (Leardi & Gonzalez, 1998). The GA scheme for variable selection includes three basic goals (Hasegawa, Miyashita, & Funatsu, 1997): (1) to create an initial population of chromosomes corresponding to a set of variables, (2) to evaluate the fitness of each chromosome in the population by the predictivity of the model, and (3) to reproduce the population of chromosomes in the next generation resulting from three operations: selection, crossover and mutation. Other variable selection methods (Shahlaei, 2013; Tropsha, Gramatica, & Gombar, 2003), such as y-scrambling, artificial neural network (ANN), simulated annealing, particle swarm optimization, automatic relevance determination and k Nearest Neighbors, will not be reviewed here due to the limited text space.

2.4. *Model construction*

When the selected variables as the independent variables and the correct response values of the dataset as the dependent variables are prepared, the next step is to take scientific methods to establish the model. This step is called model construction. The approaches used for modeling can be basically divided into linear approaches, such as multiple linear regression (MLR) and partial least square (PLS), and nonlinear approaches, such as support vector machine (SVM) and ANN, as shown in Table 2. Next, we will discuss the principle and characteristics of each method.

MLR. MLR is a multivariate statistical technique for examining the linear correlation between two or more independent variables and a single dependent variable. As we know, MLR is the earliest and most basic modeling method used in 2D-QSAR by Hansch-Fujita (Hansch et al., 1962). The quality of the MLR model is related to the number of independent variables in the regression equation (*M*) and the number of observations (*N*). It is generally accepted that *N* is at least 5 times greater than *M* to eliminate an accidental correlation (Topliss & Costello, 1972). Multicollinearity of independent variables is another important issue in MLR modeling. To effectively solve the multicollinearity problems, it is essential to screen the independent variables by the variable selection

Table 2
The properties, usage scenarios and examples in food science and technology field for multivariate statistical and machine learning methods.

Method	Property	Usage scenario	Example
MLR	1) Lead to a more precise understanding of the relative influence of one or more predictor factors to the criterion value 2) The ability to identify outliers	For examining the linear correlation between two or more independent variables and a single dependent variable	Hansch et al. (1962) Topliss and Costello (1972) Le Maux et al. (2015)
PLS	1) Handle more descriptor variables than compounds 2) Handle nonorthogonal descriptors and multiple results 3) Provide predictive accuracy and low risk of chance correlation	For constructing predictive models, particularly if these variables are highly co-linear	Casal et al. (1996) Put et al., 2006 Wu et al. (2006)
SVM	1) Works well for a clear margin of separation between classes 2) More effective in high dimensional spaces 3) Relatively memory efficient	1) For small sample sizes and nonlinearity 2) For high dimensionality 3) For local minimization	Noble (2006) Charoenkwan et al. (2020) Yang (2004)
ANN	1) Large-scale parallel processing 2) good self-adaptive ability 3) strong learning and fault-tolerance function	For prediction of the outcome where complex nonlinearity or/and important interactions can be found in the data set	Petritis, Kangas, Ferguson, Anderson, Pasa-Tolic, Lipton, et al. (2003) Guan et al. (2019) Yan, Bhadra, Li, Sethiya, Qin, Tai, et al. (2020)

methods mentioned above. The variables used for MLR can be derived from PCA or FA and can thereby perform principal component regression (PCR) or factor regression, respectively. PCA condenses and synthesizes raw data to simplify the data matrix and reduces the number of dimensions to reveal the internal structural characteristics of the data (Iwaniak et al., 2015). These variables named principal components from PCA are not related to each other but also comprehensively reflect information from the original multiple variables by a linear combination manner. FA studies how to condense many original variables into a few factor variables to make the factor variables more easily interpreted. The common factors by FA are independent of each other when they are orthogonal and related when they are oblique.

PLS. PLS, as a multivariate statistical method, is particularly suitable for regression modeling when the sample size is smaller than the number of variables (Wold, Sjöström, & Eriksson, 2001). PLS draws on the idea of component extraction in PCR but is fundamentally different from PCR. The components extracted by PLS can allow for a good interpretation of the dependent variables while providing a summary of information on the independent variables and eliminating noise interference to some extent. As a result, the PLS-based calculations are more reliable than PCR, and the resulting model is more easily interpreted. PLS not only reduces the dimensionality of the data but also effectively solves the multicollinearity problem among variables (Wold, Trygg, Berglund, & Antti, 2001).

SVM. As a machine learning algorithm based on the principle of structural risk minimization (Cortes & Vapnik, 1995), SVM can solve the problems of small sample sizes, nonlinearity, high dimensionality and local minimization (Yang, 2004). The core idea of SVM for solving the classification problem of two types of samples in cases able to be separated linearly is to find an optimally separating hyperplane that makes the maximum-margin between the two types of samples (Fig. 2A). The optimal hyperplane is determined by a few samples called support vectors. For nonlinear cases, SVM utilizes nonlinear mapping to map data into high-dimensional space, which is expressed as a linear combination of data points (Noble, 2006).

SVM solves the regression problem in a similar way as it solves the classification problem, but unlike the classification problem, the regression seeks a hyperplane that minimizes the distance from all sample points to the hyperplane (Fig. 2B). The SVM transforms nonlinear problems into linear problems through spatial transformations, thus the implementation of mathematical operations in the feature space becomes key. However, feature spaces tend to have high or even infinite dimensions, so they face problems such as dimensional disaster due to the huge increase in computational amount after spatial transformation. Solving nonlinear problems by introducing a kernel function (Fig. 2C) is a core technique that effectively resolves the contradiction between high dimension and computational complexity.

ANN. ANN, which is inspired by the operations of neurons in the brain, utilizes the combination of nonlinear units to learn the highly complex functions of the inputs and is characterized by several distinct features, including large-scale parallel processing, good self-adaptive ability, strong learning function, and fault-tolerance function (Lo, Rensi, Torng, & Altman, 2018). There are many kinds of neural networks that can be roughly divided into two types: supervised and unsupervised. The supervised ANN first trains the known samples and then forecasts the unknown samples. The unsupervised method does not need to train the known samples but can be directly used for prediction or classification of the compounds.

Back propagation ANN (BP ANN), as the typical representative of the supervised ANN, consists of three parts: the input layer, the hidden layer and the output layer (Fig. 3). In addition to BP ANN, several other ANNs, such as probabilistic neural networks, Bayesian neural networks and Kohonen self-organizing map neural networks, are commonly used. The advantages and disadvantages of these methods have been described in detail (Niculescu, 2003). Here, we focus on the convolutional neural network (CNN) and long-short-term-memory (LSTM) neural network that have emerged in recent years. As one of the most successful implementations in deep learning algorithms stemming from ANNs, CNN (Fig. 3) is a kind of feedforward neural network. In CNN, local filters scan through input space to search for recurring local spatial patterns, thereby automatically extracting features from raw data, and solving complicated nonlinear problems. CNN has been widely applied in computer vision, natural language processing, speech recognition and, more recently, in cheminformatics and computer-aided drug design (Lo et al., 2018). A variant of CNN, a graph convolution network (GCN), uses the similar concept of a local spatial filter, thereby operating on the graph to learn the features from its neighborhood. GCNs have been applied in QSAR analysis (Baskin, Palyulin, & Zefirov, 1997). We have used different GCN structures to learn the representation of small molecules (Shi et al., 2019). As a subclass of recurrent neural networks, LSTM networks (Fig. 3) are another major family of deep neural networks that use gated units and memory cells to capture long- and short-term temporal dependencies within input sequences (Lo et al., 2018). LSTM networks have been used for the structural characterization of polypeptides with different lengths (Muller, Hiss, & Schneider, 2018).

One of the main parameters of ANNs is the number of hidden neurons. This number depends on the structure and number of dependent and independent variables, as well as the number of compounds measured by these variables. Cross-validation is a common method to determine the number of neurons, i.e., the whole samples are divided into two groups, a training set for training the network and a test set for evaluating the prediction ability (Salt, Yildiz, Livingstone, & Tinsley,

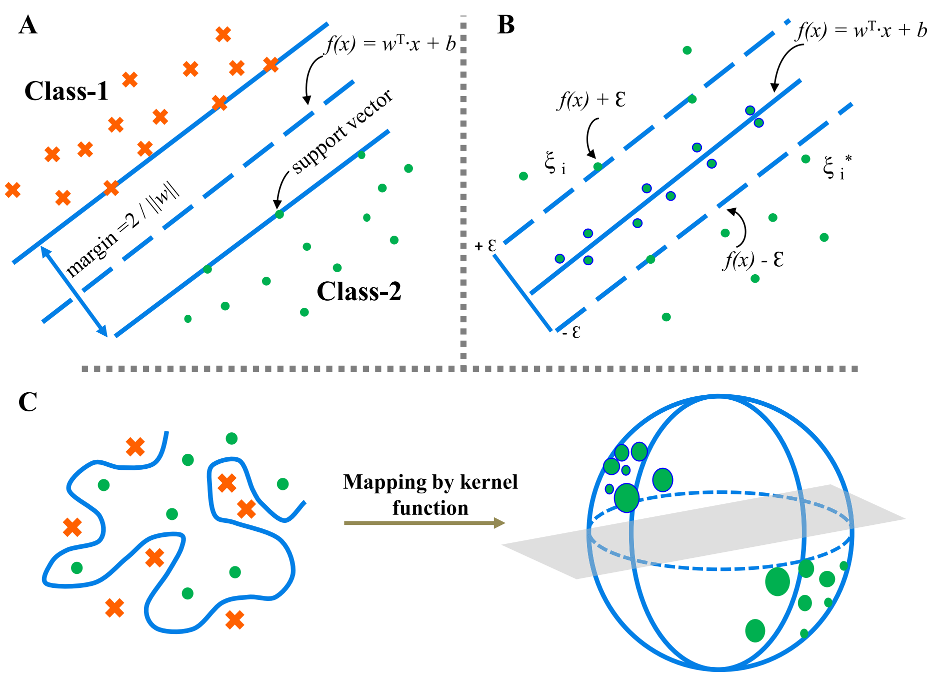

Fig. 2. The illustrations of SVM. In classification (Fig. 2A), SVM learns a decision function $f(x) = w^T x + b$, where w is known as weight vector; there is one function for each feature, whose linear combination predicts the value of y. x is the input feature vector. T is the transpose operator. If $f(x) > 0$, then x is labeled as +1 (Class-1), otherwise it is labeled as −1 (Class-2). In regression (Fig. 2B), a margin of tolerance (ε) is set in approximation to the SVM, which would have already requested it from the problem. In addition, ξ_i and ξ_i* are the introduced nonnegative relaxation variables, which are used to allow regression error. Fig. 2C shows that, by using the kernel function, the SVM-based discrimination model maps the low dimensional input data to the high dimensional space, and finally, the classification hyperplane is solved in the high dimensional space.

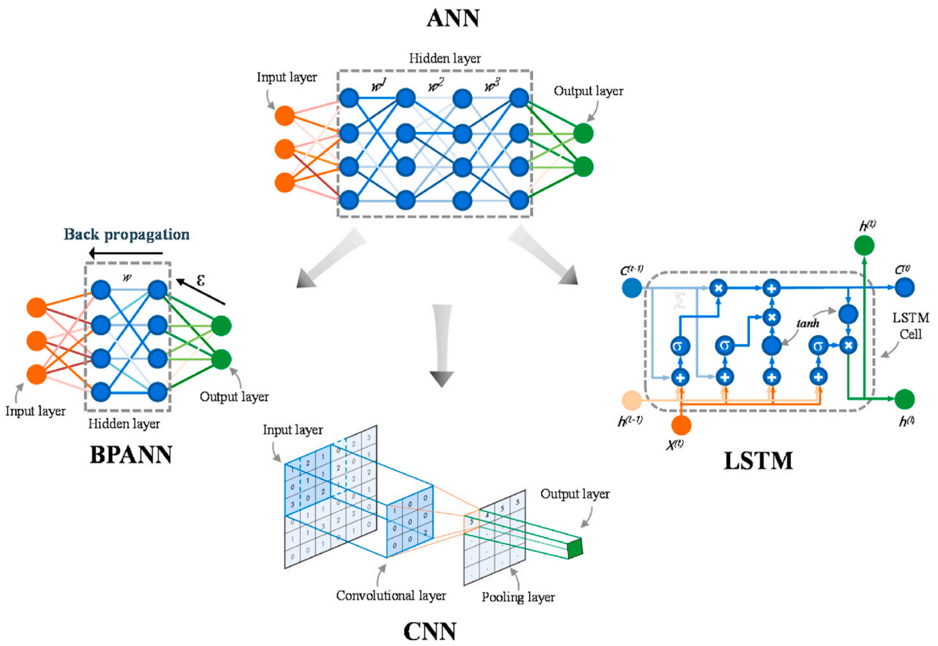

Fig. 3. The illustrations of ANNs. w is the weight of the link between the neurons in hidden layer. In BPANN, ϵ is the propagated error from the actual output to the predicted output, which is used to improve the weights of the previous layer. In LSTM, $^{(t-1)}$ and $h^{(t-1)}$ indicate the previous cell state and hidden state respectively, while c^t and h^t are the current cell state and hidden state, respectively. In addition, $X^{(t)}$ is the input feature vector, σ represents the sigmoid activation function (between 0 and 1 for Sigmoid), and *tanh* is defined as the activation function, whose value is between −1 and 1.

1992). ANN has been used for QSAR since the 1990s, however, their shortcomings cannot be ignored, such as randomness for the setting of network topology and learning parameters (Huang, Kangas, & Rasco, 2007), overtraining, lack of interpretability, and the black box effect (Iwaniak et al., 2015). In spite of these shortcomings, ANNs are still considered valuable tools for modeling in QSAR and will be used in many areas including food science (Dobchev & Karelson, 2016; Huang et al., 2007; Niculescu, 2003).

Table 2 presents the properties, usage scenarios and examples in food science and technology field for multivariate statistical and machine learning methods. As a commonly used multivariate statistical method, MLR, which is applied for examining the linear correlation between two or more independent variables and a single dependent variable (Le Maux, Nongonierma, & FitzGerald, 2015), can result in a precise understanding of the relative influence of one or more predictor factors to the criterion value, with the ability to identify outliers. Another linear modeling method is that PLS, which can handle more variables than compounds and nonorthogonal descriptors and multiple results, thus providing predictive accuracy and low risk of chance correlation for constructing the models, particularly if the variables are highly co-linear. For machine learning methods, SVM can work well for a clear margin of separation between classes and are more effective in high dimensional spaces. As mentioned above, SVM models are especially suitable for small sample sizes, nonlinearity, high dimensionality and local minimization (Yang, 2004). The ANN method is a large-scale parallel processing method, with good self-adaptive ability, strong learning and fault-tolerance function. ANN models can be applied for predicting the outcome where complex nonlinearity or/and important interactions can be found in the data set (Yan, Bhadra, Li, Sethiya, Qin, Tai, et al., 2020).

Moreover, it has to be mentioned that, the scoring card method (SCM) is a popular category of methods utilized in predicting and characterizing the functions of peptides and proteins. SCM only relies on the amino acid sequences but not known structure information of these biomolecules. On the basis of SCM, several web-based servers and programs were released currently, such as iBitter-SCM (Charoenkwan et al., 2020), iUmami-SCM (Charoenkwan et al., 2020), and iDPPIV-SCM (Charoenkwan et al., 2020). Using only the weighted sum between components and propensity scores, these tools can help biologists easily identify the desired peptides and proteins.

Generally, linear models are built using interpretable physicochemical descriptors, while nonlinear models are generally regarded as a 'black box' that is inexplicable (Cumming, Davis, Sorel, Markus, & Hongming, 2013). In the QSAR study of peptides, the choice of the linear or nonlinear method to establish the model should be determined according to the structure-activity relationship of the peptides studied. In general, if there is a complicated nonlinear relationship between the structure parameters and the activity of peptides, it is suggested to choose the nonlinear methods. Of course, if a linear method can be used to build a QSAR model with a good prediction and generalization ability, we should first choose the linear method instead of the more complex nonlinear method, because the model based on the nonlinear method is more difficult to explain.

2.5. Model evaluation and validation

As shown in Fig. 4, the metrics used for QSAR model evaluation and validation are divided into two types, including classification and regression. The performance of one classification model can be evaluated by many metrics, such as Receiver Operating Curve-Area under Curve (ROC-AUC) (Van Erkel & Peter, 1998), accuracy, precision, sensitivity and specificity. The ROC-AUC value close to one indicates the more authentic the method. Accuracy refers to how the model correctly recognizes the samples. Precision describes the number of true positives divided by the number of false positives and true positives. Sensitivity describes the true positive rate, which is the number of positive compounds that are determined as positive. Specificity depicts the true negative rate, referring to the number of negative compounds that are correctly identified as negative.

The metrics referring to regression models are involved some indicators that measure the magnitude of the difference between the true value and the predicted value. The coefficient of determination (R^2), *F*-value and error are important statistical parameters to evaluate the performance of a regression-like QSAR model (Fig. 4). R^2 reflects the degree of correlation between the independent variable(s) and the dependent variable(s). *F*-value is defined as the explained mean square to the residual mean square. The common errors in the QSAR model are characterized by the mean squared error (*MSE*) or root mean square error (*RMSE*). The error of each variable in the model also needs to be considered, and in general, if the error of a variable coefficient is close to, or greater than, the magnitude of the coefficient of the descriptor, the smaller the contribution of the descriptor to the model.

One of the purposes for QSAR modeling is to predict the activity of unknown compounds. Thus, it is particularly important to evaluate and validate the prediction ability of the model. Cross-validation (CV), including leave-one-out (*LOO*) and leave-many-out (*LMO,* such as k-fold CV and nested CV) and hold-out are widely used internal validation methods. *LOO* takes one sample out of all the samples, in turn, and uses the remaining *n*-1 samples to build a model and predict the activity of the sample until each sample is taken out and predicted. The principle of *LMO* is similar to that of *LOO*. The predictive ability of cross-validation can be evaluated by a cross-validated correlation coefficient, Q^2_{cv} (Fig. 4).

It has been reported that a QSAR model with a high predictive ability should meet the following requirements: $Q^2_{\mathrm{cv}} > 0.500$ and $R^2 > 0.600$ (Tropsha et al., 2003). A good internal validation result does not ensure a high prediction ability of the model (Aptula, Jeliazkova, Schultz, & Cronin, 2005). External validation is recommended to evaluate the external prediction ability of the model. Good practice of the external validation is splitting the total data set into a training set for establishing a model and a test set for evaluating the performance of the model obtained. The external prediction power of a model can be evaluated by an external Q^2_{cv} named Q^2_{ext} (Tropsha et al., 2003).

3. Applications of QSAR in peptides

In recent years, QSAR methods have been applied in the screening of new molecules, prediction of physicochemical properties, identification of functional activities, and elucidation of mechanisms for bioactive peptides. These studies involve not only modeling methodology but also new applications of QSAR in bioactive peptides. The following part will focus on the application of QSAR in various peptides.

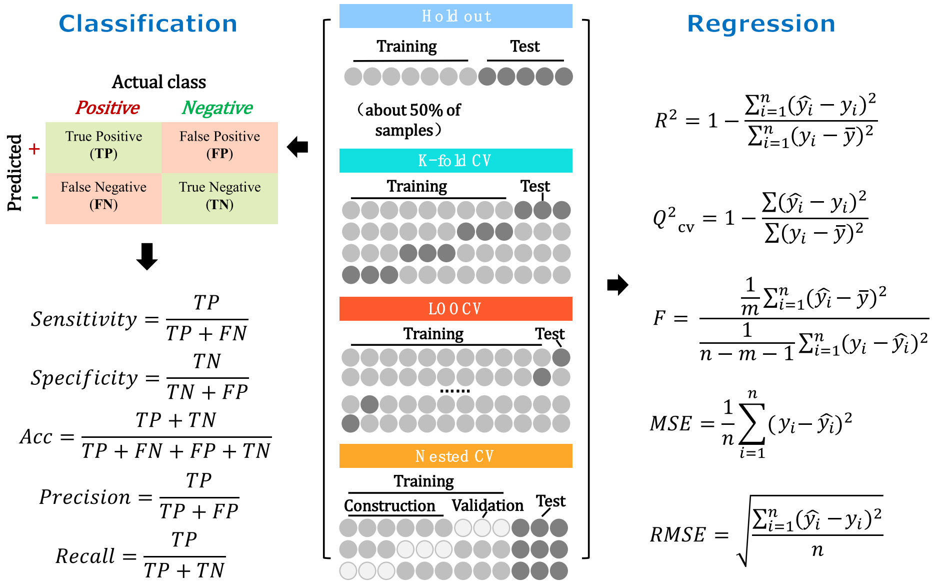
Classification
Actual class
Positive
Negative
Predicted
+
-
True Positive (TP)
False Positive (FP)
False Negative (FN)
True Negative (TN)
$Sensitivity = \frac{TP}{TP+FN}$
$Specificity = \frac{TN}{TN+FP}$
$Acc = \frac{TP+TN}{TP+FN+FP+TN}$
$Precision = \frac{TP}{TP+FP}$
$Recall = \frac{TP}{TP+TN}$
Hold out
Training
Test
（about 50% of samples）
K-fold CV
Training
Test
LOOCV
Training
Test
......
Nested CV
Training
Construction
Validation
Test
Regression
$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
$Q^2{}_{\mathrm{cv}} = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2}$
$F = \frac{\frac{1}{m}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n-m-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$

Fig. 4. Model evaluation and validation methods.

3.1. Retention prediction of peptides

Prediction of peptide retention behavior, combined with routine MS/MS data analysis, is helpful to enhance the credibility of peptide identification in proteomic analysis. Apart from the classic methods used to predict peptide retention (such as sequence-specific retention calculator proposed by *Krokhin* (Krokhin, 2006), recently, the quantitative structure-retention relationship (QSRR) (Baczek & Kaliszan, 2009), evolving from QSAR, has become a popular technique to predict the retention time of analytes. The modeling process of QSRR (Kaliszan, 2007; Taraji et al., 2018) is similar to that of QSAR.

Based on the amino acid composition, two linear QSRR models for predicting the peptide retention time in four RP-HPLC columns were constructed using MLR ($R^2 = 0.993$) and PLS ($R^2 = 0.999$) models. The authors found that the retention of small peptides in RP-HPLC was highly related to the contribution sum of amino acid residues to the chromatographic retention; moreover, under the same elution conditions, the inherent hydrophobicity of the peptides and the properties of the stationary phase played a determinative role on the retention of in RP-HPLC (Casal, MartinAlvarez, & Herraiz, 1996). An MLR-based QSRR model ($R^2 = 0.927$) was proposed to predict the gradient RP-HPLC retention of 101 peptides at fixed separation conditions based on three descriptors, i.e., logarithm of the sum of retention times of the amino acids, logarithm of the van der Waals volume and logarithm of the calculated *n*-octanol-water partition coefficient (Kaliszan et al., 2005). An MLR model was utilized to predict the retention time to attempt to improve peptide identification based on HILIC–MS/MS (Le Maux et al., 2015). One of the main characteristics for this model is that it considered hydrophilicity and length of short peptides. Put, Daszykowski, Baczek, and Vander Heyden (2006) used 128 descriptors selected by uninformative variable elimination to construct a PLS-based QSRR model to predict the retention of the peptides measured by gradient elution RPLC.

Machine learning methods have also been used to predict retention of peptides. A BPANN model consisting of 20 input nodes, 2 hidden nodes, and 1 output node was used to predict the retention times for the peptides identified by MS/MS (Petritis, Kangas, Ferguson, Anderson, Pasa-Tolic, Lipton, et al., 2003). The authors also proposed an improved ANN-based method consisting of 1052 input nodes, 24 hidden nodes, and 1 output node to predict the peptide retention times in RPLC (Petritis, Kangas, Yan, Monroe, Strittmatter, Qian, et al., 2006). The reliable predictive capability of this model was attributed to the descriptors used, involving amino acid composition, peptide conformation and configuration information, and the fitting ability for the descriptors and the dependent values by ANN. Recently, Guan, Moran, and Ma (2019) predicted LC-MS/MS properties of peptides from sequences by bidirectional LSTM recurrent neural networks models.

3.2. ACE peptide inhibitors

A variety of QSAR models for food-derived ACE inhibitory peptides have been reported. One commonly used ACE inhibitory dataset for QSAR modeling is a dataset that includes 58 dipeptides (Hellberg, Eriksson, Jonsson, Lindgren, Sjostrom, Skagerberg, et al., 1991). Many authors have used this dataset for QSAR modeling (Table S17). The modeling results showed that the HESH-based model exhibited the highest Q^2_{cv} (0.838) of all the models. The structure-activity relationship on this dataset given by four QSAR models, i.e., HESH, z-scales, ISA-ECI and FASGAI, revealed that the hydrophobicity, bulky properties, electronic characteristics at the corresponding sites were the main factors affecting the activities of ACE dipeptide inhibitors.

Several QSAR models of ACE inhibitory peptides with different lengths (penta-, hexa-, hepta- and octa-peptides) were established by a combination of PLS regression and 5z-scales characterization. The authors used the models to predict the potential peptides from Qula casein hydrolysates, and determined three new peptides (PFPGPIPN, KYIPIQ and LPLPLL) (Lin, Zhang, Han, & Cheng, 2017). Another similar work showed that a QSAR model was applied to predict the ACE inhibitory peptides from hydrolysates of Qula casein by a two-enzyme combination approach (Lin et al., 2018). Accordingly, four o peptides (KFPQY, MPFPKYP, MFPPQ and QWQVL) were synthetized and evaluated. Therein, KFPQY exhibited the highest ACE inhibitory activity ($IC_{50} = 12.37 \pm 0.43$ μM). One work (Sagardia, Roa-Ureta, & Bald, 2013) that must be mentioned was a QSAR analysis for 263 ACE inhibitory peptides with more than 5 amino acids, which were described according to the structural characteristics of the C-terminal five-residue sequences by 38 physicochemical descriptors. The resulting model based on a generalized linear method ($R^2 = 0.944$) could be used to predict and design potential ACE inhibitory peptides. Another work by Gu et al. focused on characterizing of ACE inhibitory peptides with LC-MS/MS and QSAR approaches (Gu & Wu, 2013). Additionally, Wu, et al. analyzed ACE-inhibitory, antimicrobial and bitter-tasting peptides with QSAR models (Wu et al., 2014).

Food-derived proteins are an important source of ACE peptide inhibitors. It makes sense to work out how many potential ACE peptide inhibitors are present in common food-derived proteins. Gu, Majumder, and Wu (2011) used QSAR models to evaluate the potential ACE inhibitory peptides from major food-derived proteins by *in silico* digestion. A total of 5709 peptides ranging from 2 to 6 amino acid residues were predicted, showing that the proteins from livestock meat, milk, egg, soybean and canola were good sources of ACE inhibitory peptides. Wu, Aluko, and Nakai (2006) established two PLS-based QSAR models of ACE inhibitory peptides consisting of 168 dipeptides and 140 tripeptides characterized by z-scales. The authors showed that amino acids at many specific positions or their characteristics were beneficial to the improvement of the activity, for instance, the bulky and hydrophobic properties of amino acid residues for dipeptides, while the aromatic amino acids at C-terminal, the positively charged amino acids at the intermediate site, and the hydrophobic amino acids at N-terminal for tripeptides. They used the models to identify new peptides within the sequences of pea protein, bovine milk protein and soybean protein. Finally, three peptides (LRW, IKP and FW) exhibited strong antihypertensive activities.

Most of the above QSAR studies on ACE peptide inhibitors were based on amino acid descriptors. Here, we specifically mention two modeling works based on CoMFA and CoMSIA for ACE inhibitory peptides. Two models by CoMFA and CoMSIA were proposed to explore the 3D-QSAR of ACE inhibitory peptides with a phenylalanine C-terminal (Qi et al., 2017). According to the models, the authors determined IC_{50} values of four ACE inhibitory tripeptides, including GEF (13 mM), VEF (23 mM), VRF (5 mM) and VKF (11 mM). Wu et al. carried out CoMFA- and CoMSIA-based 3D-QSAR modeling on 113 ACE inhibitory peptides (2–6 amino acid residues) (S. F. Wu et al., 2014). The two models described above could well correlate the activity and structural characteristics of ACE inhibitory peptides; however, as previously mentioned, molecular alignment based on CoMFA and CoMSIA models is a difficult task. Particularly for peptide molecules, due to their inconsistent length, the requirement to find a suitable common structure in the process of achieving alignment remains a problem to be solved (Qi et al., 2017).

3.3. Bitter peptides

Studying the structure-bitterness relationship of peptides will be beneficial to eliminating or masking their bitter taste, thereby expanding their applications (Iwaniak, Minkiewicz, Darewicz, & Hrynkiewicz, 2016). Recently, there have been many QSAR modeling studies on bitter peptides. It should be noted that a dataset consisting of 48 bitter dipeptides (Hellberg et al., 1991) is commonly used for verifying the characterization ability of a set of descriptors in QSAR modeling. The PLS modeling results for this dataset in recent literature (Table S17) showed that the HESH-based model (Shu, Mei, Yang, Liao, & Li, 2009) derived from our group generated a highest Q^2_{cv} (0.865) in all the models. Through analysis of the structure-activity relationship for the 48 bitter dataset, we found that three representative QSAR models (HESH, ISA-ECI and FASGAI) did not reach the same conclusions for the structure-activity relationship of the bitter dipeptides, mainly because of the inconsistent structural characterization methods used. These inconsistent findings need to be evaluated using other computational or experimental means. Of course, this also enlightens us to the fact that the choice of structural characterization method is important in QSAR modeling, which generally requires the selection of features closely related to the activity of the studied peptides.

Three bitter peptide datasets containing 48 dipeptides, 52 tripeptides and 23 tetrapeptides were characterized by 87 amino acid descriptors, followed by variable selection using a bootstrapping soft shrinkage approach. The resulting PLS models qualitied for di-, tri- and tetrapeptides with Q^2_{cv} at 0.941, 0.742, and 0.956, respectively (Xu & Chung, 2019). Features of Fourier Transform Raman Spectroscopy as independent variables were used to predict the bitterness of more than 200 bitter peptides (2–14 amino acids) by a PLS model (Kim & Li-Chan, 2006a). The relationship between the spectral features and the bitterness revealed that the bulky and hydrophobic amino acids at the C- and N-terminals were the main factors affecting the bitterness of the studied peptides. A QSAR model, based on PLS regression and the independent variables, i.e., the *z*-scores and/or total hydrophobicity, peptide length, and log mass values, was proposed to predict the bitterness of a dataset consisting of 224 bitter peptides and five amino acids. The structure-activity analysis demonstrated that for large peptides (tetrapeptides), the bulky hydrophobic amino acids at C-terminal and the bulky basic amino acids at N-terminal exhibited decisive impacts on the bitterness of the peptides (Kim & Li-Chan, 2006b).

3.4. Antioxidant peptides

QSAR methods have been applied in the study of the structure-activity relationship of antioxidant peptides. One work explored the structure-activity relationship of the antioxidant dipeptides containing Tyr and Trp by four sets of amino acid descriptors (Zheng, Zhao, Dong, Su, & Zhao, 2016). As a result, the PLS model by DPPS generated a higher predictive performance ($R^2 = 0.72$ and $Q^2_{cv} = 0.66$) on 24 Tyr-containing dipeptides than other models. Meanwhile, the authors showed that the amino acids at the C-terminal for Tyr-containing dipeptides preferred to scavenge $ABTS^{\bullet+}$ if they had a low steric property, hydrophobicity, hydrogen-bond ability and electronic property, while the bulky hydrophobic amino acids with hydrogen-bonding property at N-terminal were suitable for the scavenging of $ABTS^{\bullet+}$. QSAR of 55 tripeptides, which were characterized by the hydrobiology scale π_α and electronic and steric parameters of amino acid side chains, was explored (Uno, Kodama, Yukawa, Shidara, & Akamatsu, 2020). The results suggested that Cys at any position, the aromatic amino acids at C-terminal, the hydrophobic residues at N-terminal, and the small HOMO-LUMO (Highest occupied molecular orbital-Lowest unoccupied molecular orbital) gap of intermediate residues could favorably improve the antioxidant activity of the peptides.

Ren's group performed QSAR analysis of antioxidant tripeptides from β-lactoglobulin using divided physicochemical property score descriptors (Tian et al., 2015). They constructed several MLR-based QSAR models, with $R^2 = 0.643$ and $Q^2_{cv} = 0.541$ on 108 samples as a training set, and $Q^2_{ext} = 0.635$ on 46 samples as a test set. Their results showed that Cys and Trp generally contributed the positive antioxidant activity to the tripeptides; moreover, the electronic and hydrogen-bonding properties of the amino acids at any site, as well as the steric properties of amino acids at the C- and N-terminals, played an important role in the antioxidant activity of the tripeptides. PLS-based QSAR was carried out on three antioxidant peptide datasets ranging from 3 to 20 amino acid residues, characterizing the structures by 17 kinds of amino acid descriptors (Li & Li, 2013). The results indicated that the amino acids at C-terminal regions were more important for the antioxidant activity than those at N-terminal regions; in addition, the electronic property of amino acids, as well as the bulky and hydrophobic amino acids at C-terminal made significant contribution to the antioxidant activity as determined by three free radical systems.

3.5. Antimicrobial peptides

Antibacterial peptides have high efficiency and a wide range of antibacterial effects on bacteria, fungi, viruses, protozoa, etc., when applied in the food field. At present, many QSAR models have been used to predict the activity and to explore the structure-activity relationship of antimicrobial peptides. We carried out QSAR analysis on a set of antimicrobial peptides derived from Bac2A against *P. aeruginosa*, using GA-PLS combined with 20 descriptor sets. Of all the models, our FASGAI-based model exhibited the highest prediction performance, and the hydrophobic, bulk and electronic properties of amino acids were closely related to the antimicrobial activities (Qian, Liang, Liu, & Liang, 2017). Wu et al. constructed 3D QSAR models based on CoMFA and CoMSIA on 24 antimicrobial peptides (Wu et al., 2014). They showed that for nine-residue antimicrobial peptides, high activities were observed for the samples with Lys or Arg at the second and fifth sites, providing valuable design suggestions for potential peptides. Another work was that Fjell et al. identified novel antibacterial peptides by chemoinformatics and machine learning (Fjell et al., 2009).

3.6. Other bioactive peptides

Prediction of bioactive peptides is an extremely important task. Currently, there exist several bioinformatics approaches applied to this field, such as AHTPin for antihypertensive peptides (Kumar, Chaudhary, Chauhan, Nagpal, Kumar, Sharma, et al., 2015), AntiCP for anticancer Peptides (Tyagi, Kapoor, Kumar, Chaudhary, Gautam, & Raghava, 2013) and AVPpred for antiviral peptides (Thakur, Qureshi, & Kumar, 2012). Among them, because of the significant advantages QSAR can be used not only for prediction of above functional peptides, but also not for other food-derived peptides. In Keska & Stadnik's study, the authors characterized a total of 252 dipeptides as DPP-IV inhibitors using 16 physicochemical descriptors, followed by PCA, then constructed an MLR model to predict the biological activity of DPP-IV inhibitors (Keska & Stadnik, 2020). Another QSAR modeling of DPP-IV inhibitors was conducted on 56 milk protein-derived peptides (2–5 amino acid residues) (Nongonierma & FitzGerald, 2016b). Pripp (2006) developed a PLS-based QSAR model of the prolyl oligopeptidase inhibitory peptides derived from *α*-casein by three amino acid descriptors, i.e., hydrophobicity, molecular bulkiness and isoelectric charge. 3D-QSAR studies of 38 rubiscolin analogues as *δ* opioid peptides using CoMFA and CoMSIA indicated that the local hydrophobic and hydrophilic characteristics of amino acids at positions 3, 4, 5, and 6 were closely related to the *δ* opioid activity of rubiscolin analogues (Caballero, Saavedra, Fernandez, & Gonzalez-Nilo, 2007). Kumar, Chaudhary, Singh Chauhan, et al. (2015) developed four SVM-based models for antihypertensive peptides including four categories (i) tiny peptides, (ii) small peptides, (iii) medium peptides and (iv) large peptides characterized by chemical descriptors combined with amino acid composition. Finally, a web-based platform was constructed for screening and designing antihypertensive peptides. All the studies mentioned above, but not limited to, contributed to the development of functional foods based on functional peptides.

4. The databases used for QSAR modeling of bioactive peptides

As mentioned above, QSAR modeling of peptides requires high-quality datasets. The heterogeneity and the response values of datasets used for modeling can directly affect the generalization and prediction ability of the obtained QSAR models. In recent years, a variety of bioinformatics databases have provided an important basis for the QSAR study of peptides. This is because a large portion of the dataset used for QSAR modeling was derived from these databases. We summarized the databases that are commonly used for QSAR modeling of bioactive peptides (Table 3). The nine databases, BIOPEP-UWM, DBAASP, AHTPDB, APD, PeptideDB, MBPDB, EROP-Moscow, PepBank and SATPdb, provide basic data on food-derived bioactive peptides, including length, activity, and physicochemical properties. Here, a database named AAindex (https://www.genome.jp/aaindex/) should be mentioned. This database integrates numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids. The amino acid index in AAindex is an important source for amino acid descriptors such as FASGAI. We have recently established a new Database of Food-derived Bioactive Peptides (DFBP, http://www.cqudfbp.net/, unpublished), including comprehensive information on food-derived bioactive peptides and food-derived proteins. Moreover, DFBP kindly provides some online analysis tools for *in silico* hydrolysis, structural characterization for amino acid descriptors, and prediction of toxicity and bitterness.

5. Conclusions and prospects

QSAR approaches have been used to screen, design and identify food-derived bioactive peptides. On the one hand, we need to further reveal the benefits of QSAR. QSAR greatly reduces the time and expenses involved in the identification and evaluation of food-derived peptides, thereby more accurately acquiring peptides. It is helpful in probing the mechanism(s) of action by quantitatively describing the relationship between the structure and the activity/property of a peptide and in solving problems that are difficult to address in experiments. On the other hand, we should face the challenges of QSAR that arise with application to peptides. For example, the limited availability of high-quality datasets makes it difficult to meet the requirement for QSAR modeling. As the saying goes “making bricks without straw”, without a high-quality dataset, it is difficult to obtain a QSAR model with reliable prediction ability.

Table 3
The databases used for QSAR modeling of bioactive peptides.

Name	Link	Description	Reference
BIOPEP-UWM	http://www.uwm.edu.pl/biochemia/index.php/en/biopep	Database of bioactive peptides	Minkiewicz, Iwaniak, and Darewicz (2019)
DBAASP	https://dbaasp.org/	Database of antimicrobial activity and structure of peptides	Malak, Andrei, Phillip, Griggs, Burke, Hurt, et al. (2016)
AHTPDB	http://crdd.osdd.net/raghava/ahtpdb/	Database of antihypertensive peptides	Kumar, Chaudhary, Sharma, Nagpal, Chauhan, Singh, et al. (2015)
AAindex	https://www.genome.jp/aaindex/	Database of amino acid indices, substitution matrices and pairwise contact potentials	Kawashima et al. (2008)
APD	http://aps.unmc.edu/AP/	Antimicrobial peptide database	Wang, Xia, and Zhe (2016)
PeptideDB	http://www.peptides.be/?p=home	Signaling peptides from animal source	
MBPDB	http://mbpdb.nws.oregonstate.edu/	Milk bioactive peptide database	Nielsen, Beverly, Qu, and Dallas (2017)
DFBP	http://www.cqudfbp.net/	Database of food-derived bioactive peptides	
EROP-Moscow	http://erop.inbi.ras.ru/	Database of curated oligopeptide sequences	Zamyatnin, Borchikov, Vladimirov, and Voronina (2006)
PepBank	http://pepbank.mgh.harvard.edu	Database of biologically active peptides	Shtatland, Guettler, Kossodo, Pivovarov, and Weissleder (2007)
SATPdb	http://crdd.osdd.net/raghava/satpdb/links.php	Metabase of bioactive peptides	Singh, Chaudhary, Dhanda, Bhalla, Usmani, Gautam, et al. (2016)

Another emphasis here is that there are limited descriptors used to build QSAR models; thus, structural characterization is still the focus of QSAR modeling for peptides, and the structural characteristics related to the activity of peptides that are scientifically extracted will largely determine the success or failure of QSAR modeling. At this point, it is still difficult to characterize the structures of peptides of different lengths. We call for new methods of scientific structural characterization related to the activity of peptides and expect to integrate new descriptors to realize the structural characterization of peptides (Wang, Yuan, Wu, Lin, & Yang, 2017). In any case, an excellent descriptor should reasonably characterize the structural characteristics of peptides and extract important structural information related to the functions of peptides. Most importantly, the extracted information can provide some insight into the relationship between the structures and functions of the studied peptides.

The method of model construction is another important issue in QSAR modeling of peptides. We should not only take advantage of traditional modeling approaches but also purposefully introduce new modeling approaches or integrate modeling techniques, such as the parameter selection algorithm and the sample grouping method. We should also pay special attention to applications of machine learning algorithms and artificial intelligence in QSAR. It should be noted that due to the differences in the selected sample quality, structural parameters and numbers, it is difficult to recommend a specific method as the best and only method for QSAR modeling of peptides (Iwaniak et al., 2015). Therefore, it is necessary to try multiple methods or combination strategies to achieve QSAR analysis.

So far, available QSAR studies of bioactive peptides have been limited; moreover, few promising peptide sequences have been evaluated *in vivo*. This means that much research is needed to improve the knowledge of using QSAR to discover bioactive peptides (Nongonierma & FitzGerald, 2016c). It is first noted that, QSAR approaches depend exclusively on the physicochemical features of the ligands (molecular descriptors) when no information is available concerning the 3D structure of the target. (Consonni, Todeschini, & Pavan, 2002; Swamidass et al., 2009). Second, inappropriate endpoint units, confounded descriptors and non-interpretable descriptors may reduce the application scope of the model. Third, it is quite common for samples to be replicated in the sets, which falsely improves the accuracy of the QSAR model. Addionly, inadequate or missing statistic measures make the assessment of model difficult. In the future, how to establish accurate and feasible QSAR models for prediction of the potential activity, screening of new molecules, and determination of the corresponding mechanism(s) of action will be a daunting task. Certainly, the solution to this task also requires multidisciplinary cooperation in multiple fields, including pharmacy, chemistry, computer science, mathematics and, of course, food science.

Author contributions

Bo Weichen: Conceptualization, Data Curation, Writing-Original Draft. Chen Lang: Conceptualization, Writing-Reviewing and Editing. Qin Dongya: Resources, Writing-Reviewing and Editing. Geng Sheng: Resources, Visualization. Li Jiaqi: Resources, Visualization. Mei Hu: Resources, Visualization. Li Bo: Supervision, Funding Acquisition. Liang Guizhao: Supervision, Funding Acquisition.

Declaration of competing interest

None.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 31771975 and 31571782), and the Natural Science Foundation of Chongqing CSTC (No. cstc2018jcyjAX0765).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tifs.2021.05.031.

References

Andersson, P. M., Sjöström, M., & Lundstedt, T. (1998). Preprocessing peptide sequences for multivariate sequence-property analysis. *Chemometrics and Intelligent Laboratory Systems, 42*(1–2), 41–50.

Aptula, A. O., Jeliazkova, N. G., Schultz, T. W., & Cronin, M. T. D. (2005). The better predictive model: High q(2) for the training set or low root mean square error of prediction for the test set? *QSAR & Combinatorial Science, 24*(3), 385–396.

Baczek, T., & Kaliszan, R. (2009). Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics. *Proteomics, 9*(4), 835–847.

Bahadori, M., Hemmateenejad, B., & Yousefinejad, S. (2019). Quantitative sequence-activity modeling of ACE peptide originated from milk using ACC-QTMS amino acid indices. *Amino Acids, 51*(8), 1209–1220.

Barati, M., Javanmardi, F., Jazayeri, S. M. H. M., Jabbari, M., Rahmani, J., Barati, F., et al. (2020). Techniques, perspectives, and challenges of bioactive peptide generation: A comprehensive systematic review. *Comprehensive Reviews in Food Science and Food Safety, 19*(4), 1488–1520.

Baskin, I. I., Palyulin, V. A., & Zefirov, N. S. (1997). A neural device for searching direct correlations between structures and properties of chemical compounds. *Journal of Chemical Information and Computer Sciences, 37*(4), 715–721.

Blanco-Miguez, A., Fdez-Riverola, F., Lourenco, A., & Sanchez, B. (2019). *In silico* prediction reveals the existence of potential bioactive neuropeptides produced by the human gut microbiota. *Food Research International, 119*, 221–226.

Bouarab-Chibane, L., Forquet, V., Lanteri, P., Clement, Y., Leonard-Akkari, L., Oulahal, N., et al. (2019). Antibacterial properties of polyphenols: Characterization and QSAR (quantitative structure activity relationship) models. *Frontiers in Microbiology, 10*, 829.

Caballero, J., Saavedra, M., Fernandez, M., & Gonzalez-Nilo, F. D. (2007). Quantitative structure-activity relationship of rubiscolin analogues as delta opioid peptides using comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA). *Journal of Agricultural and Food Chemistry, 55*(20), 8101–8104.

Casal, V., MartinAlvarez, P. J., & Herraiz, T. (1996). Comparative prediction of the retention behaviour of small peptides in several reversed-phase high-performance liquid chromatography columns by using partial least squares and multiple linear regression. *Analytica Chimica Acta, 326*(1–3), 77–84.

Charoenkwan, P., Kanthawong, S., Nantasenamat, C., Hasan, M., & Shoombuatong, W. (2020). iDPPIV-SCM: A sequence-based predictor for identifying and analyzing dipeptidyl peptidase IV (DPP-IV) inhibitory peptides using a scoring card method. *Journal of Proteome Research, 19*(10), 4125–4136.

Charoenkwan, P., Yana, J., Nantasenamat, C., Hasan, M., & Shoombuatong, W. (2020). iUmami-SCM: A novel sequence-based predictor for prediction and analysis of umami peptides using a scoring card method with propensity scores of dipeptides. *Journal of Chemical Information and Modeling, 60*(12), 6666–6678.

Charoenkwan, P., Yana, J., Schaduangrat, N., Nantasenamat, C., Hasan, M. M., & Shoombuatong, W. (2020). iBitter-SCM: Identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. *Genomics, 112*(4), 2813–2822.

Cocchi, M., & Johansson, E. (1993). Amino acids characterization by GRID and multivariate data analysis. *Quantitative Structure-Activity Relationships, 12*(1), 1–8.

Collantes, E. R., & Dunn, W. J., 3rd (1995). Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogues. *Journal of Medicinal Chemistry, 38*(14), 2705–2713.

Consonni, V., Todeschini, R., & Pavan, M. (2002). Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *Journal of Chemical Information and Computer Sciences, 42*(3), 682–692.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Cotterill, J., Price, N., Rorije, E., & Peijnenburg, A. (2020). Development of a QSAR model to predict hepatic steatosis using freely available machine learning tools. *Food and Chemical Toxicology, 142*, 111494.

Cramer, R. D. (2003). Topomer CoMFA: A design methodology for rapid lead optimization. *Journal of Medicinal Chemistry, 46*(3), 374–388.

Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society, 110*(18), 5959–5967.

Cumming, J. G., Davis, A. M., Sorel, M., Markus, H., & Hongming, C. (2013). Chemical predictive modelling to improve compound quality. *Nature Reviews Drug Discovery, 12*(12), 948–962.

Daliri, E. B., Lee, B. H., & Oh, D. H. (2018). Current trends and perspectives of bioactive peptides. *Critical Reviews in Food Science and Nutrition, 58*(13), 2273–2284.

Damale, M. G., Harke, S. N., Kalam Khan, F. A., Shinde, D. B., & Sangshetti, J. N. (2014). Recent advances in multidimensional QSAR (4D-6D): A critical review. *Mini Reviews in Medicinal Chemistry, 14*(1), 35–55.

Dearden, J. C., Cronin, M. T. D., & Kaiser, K. L. E. (2009). How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research, 20*(3–4), 241–266.

Dobchev, D., & Karelson, M. (2016). Have artificial neural networks met expectations in drug discovery as implemented in QSAR framework? *Expert Opinion on Drug Discovery, 11*(7), 627–639.

Doytchinova, I. A., & Flower, D. R. (2001). Toward the quantitative prediction of T-cell epitopes: CoMFA and CoMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. *Journal of Medicinal Chemistry, 44*(22), 3572–3581.

FitzGerald, R. J., Cermeno, M., Khalesi, M., Kleekayai, T., & Amigo-Benavent, M. (2020). Application of *in silico* approaches for the generation of milk protein-derived bioactive peptides. *Journal of Functional Foods, 64*, 103636.

Fjell, C. D., Jenssen, H., Hilpert, K., Cheung, W. A., Pante, N., Hancock, R. E. W., et al. (2009). Identification of novel antibacterial peptides by chemoinformatics and machine Learning. *Journal of Medicinal Chemistry, 52*(7), 2006–2015.

Flores-Holguin, N., Frau, J., & Glossman-Mitnik, D. (2019). Computational prediction of bioactivity scores and chemical reactivity properties of the Parasin I therapeutic peptide of marine origin through the calculation of global and local conceptual DFT descriptors. *Theoretical Chemistry Accounts, 138*, 78.

Free, S. M., & Wilson, J. W. (1964). A mathematical contribution to structure-activity studies. *Journal of Medicinal Chemistry, 7*(4), 395–399.

Gasteiger, J. (2014). Some solved and unsolved problems of chemoinformatics. *SAR and QSAR in Environmental Research, 25*(6), 443–455.

Gedeck, P., Rohde, B., & Bartels, C. (2006). Qsar - how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *Journal of Chemical Information and Modeling, 46*(5), 1924–1936.

Goel, A., Gajula, K., Gupta, R., & Rai, B. (2021). *In-silico* screening of database for finding potential sweet molecules: A combined data and structure based modeling approach. *Food Chemistry, 343*, 128538.

Guan, S. H., Moran, M. F., & Ma, B. (2019). Prediction of LC-MS/MS properties of peptides from sequence by deep learning. *Molecular & Cellular Proteomics, 18*(10), 2099–2107.

Gu, Y. C., Majumder, K., & Wu, J. P. (2011). QSAR-aided *in silico* approach in evaluation of food proteins as precursors of ACE inhibitory peptides. *Food Research International, 44*(8), 2465–2474.

Gu, Y. C., & Wu, J. P. (2013). LC-MS/MS coupled with QSAR modeling in characterising of angiotensin I-converting enzyme inhibitory peptides from soybean proteins. *Food Chemistry, 141*(3), 2682–2690.

Hansch, C., Maloney, P. P., Fujita, T., & Muir, R. M. (1962). Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. *Nature, 194*(4824), 178–180.

Hasegawa, K., Miyashita, Y., & Funatsu, K. (1997). GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *Journal of Chemical Information and Modeling, 37*(2), 306–310.

Hellberg, S., Eriksson, L., Jonsson, J., Lindgren, F., Sjostrom, M., Skagerberg, B., et al. (1991). Minimum analogue peptide sets (MAPS) for quantitative structure-activity relationships. *International Journal of Peptide & Protein Research, 37*(5), 414–424.

Hellberg, S., Sjostrom, M., Skagerberg, B., & Wold, S. (1987). Peptide quantitative structure-activity-relationships, a multivariate approach. *Journal of Medicinal Chemistry, 30*(7), 1126–1135.

Hemmateenejad, B., Yousefinejad, S., & Mehdipour, A. R. (2011). Novel amino acids indices based on quantum topological molecular similarity and their application to QSAR study of peptides. *Amino Acids, 40*(4), 1169–1183.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences, 89*(22), 10915–10919.

Hernandez-Ledesma, B., & Hsieh, C. C. (2017). Chemopreventive role of food-derived proteins and peptides: A review. *Critical Reviews in Food Science and Nutrition, 57*(11), 2358–2376.

Holton, T. A., Vijayakumar, V., & Khaldi, N. (2013). Bioinformatics: Current perspectives and future directions for food and nutritional research facilitated by a Food-Wiki database. *Trends in Food Science & Technology, 34*(1), 5–17.

Hou, T. J., Li, Z. M., Li, Z., Liu, J., & Xu, X. J. (2000). Three-dimensional quantitative structure-activity relationship analysis of the new potent sulfonylureas using comparative molecular similarity indices analysis. *Journal of Chemical Information and Modeling, 40*(4), 1002–1009.

Huang, Y., Kangas, L. J., & Rasco, B. A. (2007). Applications of artificial neural networks (ANNs) in food science. *Critical Reviews in Food Science and Nutrition, 47*(2), 113–126.

Iwaniak, A., Minkiewicz, P., Darewicz, M., & Hrynkiewicz, M. (2016). Food protein-originating peptides as tastants - physiological, technological, sensory, and bioinformatic approaches. *Food Research International, 89*, 27–38.

Iwaniak, A., Minkiewicz, P., Darewicz, M., Protasiewicz, M., & Mogut, D. (2015). Chemometrics and cheminformatics in the analysis of biologically active peptides from food sources. *Journal of Functional Foods, 16*, 334–351.

Jie, D., Ning-Ning, W., Zhi-Jiang, Y., Lin, Z., Yan, C., Defang, O., et al. (2018). ADMETlab: A platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. *Journal of Cheminformatics, 10*(1), 29.

Kaliszan, R. (2007). QSRR: Quantitative structure-(chromatographic) retention relationships. *Chemical Reviews, 107*(7), 3212–3246.

Kaliszan, R., Baczek, T., Cimochowska, A., Juszczyk, P., Wisniewska, K., & Grzonka, Z. (2005). Prediction of high-performance liquid chromatography retention of peptides with the use of quantitative structure-retention relationships. *Proteomics, 5*(2), 409–415.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M. (2008). AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Research, 36*(D1), D202–D205.

Keska, P., & Stadnik, J. (2020). Structure-activity relationships study on biological activity of peptides as dipeptidyl peptidase IV inhibitors by chemometric modeling. *Chemical Biology & Drug Design, 95*(2), 291–301.

Kim, H. O., & Li-Chan, E. C. Y. (2006a). Application of Fourier transform Raman spectroscopy for prediction of bitterness of peptides. *Applied Spectroscopy, 60*(11), 1297–1306.

Kim, H. O., & Li-Chan, E. C. Y. (2006b). Quantitative structure-activity relationship study of bitter peptides. *Journal of Agricultural and Food Chemistry, 54*(26), 10102–10111.

Klebe, G., Abraham, U., & Mietzner, T. (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *Journal of Medicinal Chemistry, 37*(24), 4130–4146.

Krokhin, O. V. (2006). Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: Application to 300-and 100-angstrom pore size C18 sorbents. *Analytical Chemistry, 78*(22), 7785–7795.

Kubinyi, H. (2002). From narcosis to hyperspace: The history of QSAR. *Quantitative Structure-Activity Relationships, 21*(4), 348–356.

Kumar, R., Chaudhary, K., Sharma, M., Nagpal, G., Chauhan, J. S., Singh, S., et al. (2015a). AHTPDB: A comprehensive platform for analysis and presentation of antihypertensive peptides. *Nucleic Acids Research, 43*(D1), D956–D962.

Kumar, R., Chaudhary, K., Singh Chauhan, J., Nagpal, G., Kumar, R., Sharma, M., et al. (2015b). An *in silico* platform for predicting, screening and designing of antihypertensive peptides. *Scientific Reports, 5*, 12512.

Le Maux, S., Nongonierma, A. B., & FitzGerald, R. J. (2015). Improved short peptide identification using HILIC-MS/MS: Retention time prediction model based on the impact of amino acid position in the peptide sequence. *Food Chemistry, 173*, 847–854.

Leardi, R., & Gonzalez, A. L. (1998). Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemometrics and Intelligent Laboratory Systems, 41*(2), 195–207.

Liang, G., & Li, Z. (2007). Factor analysis scale of generalized amino acid information as the source of a new set of descriptors for elucidating the structure and activity relationships of cationic antimicrobial peptides. *QSAR & Combinatorial Science, 26*(6), 754–763.

Liang, G., Liu, Y., Shi, B., Zhao, J., & Zheng, J. (2013). An index for characterization of natural and non-natural amino acids for peptidomimetics. *PloS One, 8*(7), Article e67844.

Liang, G. Z., Yang, L., Kang, L. F., Mei, H., & Li, Z. L. (2009). Using multidimensional patterns of amino acid attributes for QSAR analysis of peptides. *Amino Acids, 37*(4), 583–591.

Liang, G. Z., Zhou, P., Zhou, Y., Zhang, Q. X., & Zl, L. (2006). New descriptors of aminoacids and their applications to peptide quantitative structure-activity relationship. *Acta Chimica Sinica, 64*(5), 393–396.

Li, Y. W., & Li, B. (2013). Characterization of structure-antioxidant activity relationship of peptides in free radical systems using QSAR models: Key sequence positions and their amino acid properties. *Journal of Theoretical Biology, 318*, 29–43.

Lin, Z. H., Long, H. X., Bo, Z., Wang, Y. Q., & Wu, Y. Z. (2008). New descriptors of amino acids and their application to peptide QSAR study. *Peptides, 29*(10), 1798–1805.

Lin, K., Zhang, L. W., Han, X., & Cheng, D. Y. (2017). Novel angiotensin I-converting enzyme inhibitory peptides from protease hydrolysates of Qula casein: Quantitative structure-activity relationship modeling and molecular docking study. *Journal of Functional Foods, 32*, 266–277.

Lin, K., Zhang, L., Han, X., Meng, Z., Zhang, J., Wu, Y., et al. (2018). Quantitative structure-activity relationship modeling coupled with molecular docking analysis in screening of angiotensin I-converting enzyme inhibitory peptides from Qula casein hydrolysates obtained by two-enzyme combination hydrolysis. *Journal of Agricultural and Food Chemistry, 66*(12), 3221–3228.

Lo, Y. C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today, 23*(8), 1538–1546.

Malak, P., Andrei, G., Phillip, C., Griggs, H. L., Burke, S. R., Hurt, D. E., et al. (2016). DBAASP v.2: An enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Research, 44*(D1), D1104–D1112.

Mei, H., Liao, Z. H., Zhou, Y., & Li, S. Z. (2005). A new set of amino acid descriptors and its application in peptide QSARs. *Biopolymers, 80*(6), 775–786.

Mills, S., Stanton, C., Hill, C., & Ross, R. P. (2011). New developments and applications of bacteriocins and peptides in foods. *Annual Review of Food Science and Technology, 2*, 299–329.

Minkiewicz, P., Iwaniak, A., & Darewicz, M. (2019). BIOPEP-UWM database of bioactive peptides: Current opportunities. *International Journal of Molecular Sciences, 20*(23), 5978.

Muller, A. T., Hiss, J. A., & Schneider, G. (2018). Recurrent neural network model for constructive peptide design. *Journal of Chemical Information and Modeling, 58*(2), 472–479.

Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., et al. (2020). QSAR without borders. *Chemical Society Reviews, 49*(11), 3525–3564.

Nakai, S., & Li-Chan, E. (1993). Recent advances in structure and function of food proteins: QSAR approach. *Critical Reviews in Food Science and Nutrition, 33*(6), 477–499.

Niazi, A., & Leardi, R. (2012). Genetic algorithms in chemometrics. *Journal of Chemometrics, 26*(6), 345–351.

Niculescu, S. P. (2003). Artificial neural networks and genetic algorithms in QSAR. *Journal of Molecular Structure-Theochem, 622*(1–2), 71–83.

Nielsen, S. D., Beverly, R. L., Qu, Y., & Dallas, D. C. (2017). Milk bioactive peptide database: A comprehensive database of milk protein-derived bioactive peptides and novel visualization. *Food Chemistry, 232*, 673–682.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology, 24*(12), 1565–1567.

Nongonierma, A. B., & FitzGerald, R. J. (2016a). Strategies for the discovery, identification and validation of milk protein-derived bioactive peptides. *Trends in Food Science & Technology, 50*, 26–43.

Nongonierma, A. B., & FitzGerald, R. J. (2016b). Structure activity relationship modelling of milk protein-derived peptides with dipeptidyl peptidase IV (DPP-IV) inhibitory activity. *Peptides, 79*, 1–7.

Nongonierma, A. B., & FitzGerald, R. J. (2016c). Learnings from quantitative structure-activity relationship (QSAR) studies with respect to food protein-derived bioactive peptides: A review. *RSC Advances, 6*(79), 75400–75413.

Nongonierma, A. B., & FitzGerald, R. J. (2017). Strategies for the discovery and identification of food protein-derived biologically active peptides. *Trends in Food Science & Technology, 69*, 289–305.

Pal, R., Jana, G., Sural, S., & Chattaraj, P. K. (2019). Hydrophobicity versus electrophilicity: A new protocol toward quantitative structure-toxicity relationship. *Chemical Biology & Drug Design, 93*(6), 1083–1095.

Petritis, K., Kangas, L. J., Ferguson, P. L., Anderson, G. A., Pasa-Tolic, L., Lipton, M. S., et al. (2003). Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Analytical Chemistry, 75*(5), 1039–1048.

Petritis, K., Kangas, L. J., Yan, B., Monroe, M. E., Strittmatter, E. F., Qian, W. J., et al. (2006). Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Analytical Chemistry, 78*(14), 5026–5039.

Pihlanto-Leppälä, A. (2000). Bioactive peptides derived from bovine whey proteins: Opioid and ace-inhibitory peptides. *Trends in Food Science & Technology, 11*(9–10), 347–356.

Pripp, A. H. (2006). Quantitative structure-activity relationship of prolyl oligopeptidase inhibitory peptides derived from beta-casein using simple amino acid descriptors. *Journal of Agricultural and Food Chemistry, 54*(1), 224–228.

Pripp, A. H., Isaksson, T., Stepaniak, L., Sorhaug, T., & Ardo, Y. (2005). Quantitative structure activity relationship modelling of peptides and proteins as a tool in food science. *Trends in Food Science & Technology, 16*(11), 484–494.

Put, R., Daszykowski, M., Baczek, T., & Vander Heyden, Y. (2006). Retention prediction of peptides based on uninformative variable elimination by partial least squares. *Journal of Proteome Research, 5*(7), 1618–1625.

Qian, Y., Liang, Y. N., Liu, W. Q., & Liang, G. Z. (2017). Comprehensive comparison of twenty structural characterization scales applied as QSAM of antimicrobial dodecapeptides derived from Bac2A against P-aeruginosa. *Journal of Molecular Graphics and Modelling, 71*, 88–95.

Qi, C., Lin, G., Zhang, R., & Wu, W. (2017). Studies on the bioactivities of ACE-inhibitory peptides with phenylalanine C-terminus using 3D-QSAR, molecular docking and in vitro evaluation. *Molecular Informatics, 36*(9). Article 1600157.

Sagardia, I., Roa-Ureta, R. H., & Bald, C. (2013). A new QSAR model, for angiotensin I-converting enzyme inhibitory oligopeptides. *Food Chemistry, 136*(3–4), 1370–1376.

Salt, D. W., Yildiz, N., Livingstone, D. J., & Tinsley, C. J. (1992). The use of artificial neural networks in QSAR. *Pesticide Science, 36*(2), 161–170.

Sandberg, M., Eriksson, L., Jonsson, J., Sjostrom, M., & Wold, S. (1998). New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *Journal of Medicinal Chemistry, 41*(14), 2481–2491.

Shahlaei, M. (2013). Descriptor selection methods in quantitative structure-activity relationship studies: A review study. *Chemical Reviews, 113*(10), 8093–8103.

Shi, T. T., Yang, Y. W., Huang, S. H., Chen, L. X., Kuang, Z. Y., Heng, Y., et al. (2019). Molecular image-based convolutional neural network for the prediction of ADMET properties. *Chemometrics and Intelligent Laboratory Systems, 194*, 103853.

Shtatland, T., Guettler, D., Kossodo, M., Pivovarov, M., & Weissleder, R. (2007). PepBank - a database of peptides based on sequence text mining and public peptide data sources. *BMC Bioinformatics, 8*(1), 280-280.

Shu, M., Mei, H., Yang, S. B., Liao, L. M., & Li, Z. L. (2009). Structural parameter characterization and bioactivity simulation based on peptide sequence. *QSAR & Combinatorial Science, 28*(1), 27–35.

Singh, S., Chaudhary, K., Dhanda, S. K., Bhalla, S., Usmani, S. S., Gautam, A., et al. (2016). SATPdb: A database of structurally annotated therapeutic peptides. *Nucleic Acids Research, 44*(D1), D1119–D1126.

Sinha, N., & Sen, S. (2011). Predicting hERG activities of compounds from their 3D structures: Development and evaluation of a global descriptors based QSAR model. *European Journal of Medicinal Chemistry, 46*(2), 618–630.

Smith, D. W., Gill, D. S., & Hammond, J. J. (1985). Variable selection in multivariate multiple-regression. *Journal of Statistical Computation and Simulation, 22*(3–4), 217–227.

Swamidass, S. J., Azencott, C. A., Lin, T. W., Gramajo, H., Tsai, S. C., & Baldi, P. (2009). Influence relevance voting: An accurate and interpretable virtual high throughput screening method. *Journal of Chemical Information and Modeling, 49*(4), 756–766.

Taraji, M., Haddad, P. R., Amos, R. I. J., Talebi, M., Szucs, R., Dolan, J. W., et al. (2018). Chemometric-assisted method development in hydrophilic interaction liquid chromatography: A review. *Analytica Chimica Acta, 1000*, 20–40.

Thakur, N., Qureshi, A., & Kumar, M. (2012). AVPpred: Collection and prediction of highly effective antiviral peptides. *Nucleic Acids Research, 40*(W1), W199–W204.

Tian, M., Fang, B., Jiang, L., Guo, H. Y., Cui, J. Y., & Ren, F. Z. (2015). Structure-activity relationship of a series of antioxidant tripeptides derived from beta-Lactoglobulin using QSAR modeling. *Dairy Science & Technology, 95*(4), 451–463.

Tian, F., Yang, L., Lv, F., Yang, Q., & Zhou, P. (2009). *In silico* quantitative prediction of peptides binding affinity to human MHC molecule: An intuitive quantitative structure-activity relationship approach. *Amino Acids, 36*(3), 535–554.

Tong, J., Liu, S., Zhou, P., Wu, B., & Li, Z. (2008). A novel descriptor of amino acids and its application in peptide QSAR. *Journal of Theoretical Biology, 253*(1), 90–97.

Topliss, J. G., & Costello, R. J. (1972). Chance correlations in structure-activity studies using multiple regression-analysis. *Journal of Medicinal Chemistry, 15*(10), 1066–1068.

Tropsha, A., Gramatica, P., & Gombar, V. K. (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science, 22*(1), 69–77.

Tu, M. L., Cheng, S. Z., Lu, W. H., & Du, M. (2018). Advancement and prospects of bioinformatics analysis for studying bioactive peptides from food-derived protein: Sequence, structure, and functions. *Trac-Trends in Analytical Chemistry, 105*, 7–17.

Tyagi, A., Kapoor, P., Kumar, R., Cmhaudhary, K., Gautam, A., & Raghava, G. P. S. (2013). *In Silico* odels for designing and discovering novel anticancer peptides. *Scientific Reports, 3*, 2984.

Udenigwe, C. C. (2014). Bioinformatics approaches, prospects and challenges of food bioactive peptide research. *Trends in Food Science & Technology, 36*(2), 137–143.

Uno, S., Kodama, D., Yukawa, H., Shidara, H., & Akamatsu, M. (2020). Quantitative analysis of the relationship between structure and antioxidant activity of tripeptides. *Journal of Peptide Science, 26*(3), e3238.

Van Erkel, A. R., & Peter, M. (1998). Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology. *European Journal of Radiology, 27*(2), 88–94.

Wang, G., Xia, L., & Zhe, W. (2016). APD3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research, 44*(D1), D1087–D1093.

Wang, T., Yuan, X. S., Wu, M. B., Lin, J. P., & Yang, L. R. (2017). The advancement of multidimensional QSAR for novel drug discovery - where are we headed? *Expert Opinion on Drug Discovery, 12*(8), 769–784.

van Westen, G. J., Swier, R. F., Wegner, J. K., Ijzerman, A. P., van Vlijmen, H. W., & Bender, A. (2013). Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): Comparative study of 13 amino acid descriptor sets. *Journal of Cheminformatics, 5*(1), 41.

Wold, S., Sjöström, M., & Eriksson, L. (2001a). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*(2), 109–130.

Wold, S., Trygg, J., Berglund, A., & Antti, H. (2001b). Some recent developments in PLS modeling. *Chemometrics and Intelligent Laboratory Systems, 58*(2), 131–150.

Wu, J. P., Aluko, R. E., & Nakai, S. (2006). Structural requirements of angiotensin I-converting enzyme inhibitory peptides: Quantitative structure-activity relationship study of di- and tripeptides. *Journal of Agricultural and Food Chemistry, 54*(3), 732–738.

Wu, S. F., Qi, W., Su, R. X., Li, T. H., Lu, D., & He, Z. M. (2014). CoMFA and CoMSIA analysis of ACE-inhibitory, antimicrobial and bitter-tasting peptides. *European Journal of Medicinal Chemistry, 84*, 100–106.

Xu, B. Y., & Chung, H. Y. (2019). Quantitative structure-activity relationship study of bitter di-, tri- and tetrapeptides using integrated descriptors. *Molecules, 24*(15), 2846.

Yan, J. L., Bhadra, P., Li, A., Sethiya, P., Qin, L. G., Tai, H. K., et al. (2020). Deep-AmPEP30: Improve short antimicrobial peptides prediction with deep learning. *Molecular Therapy - Nucleic Acids, 20*, 882–894.

Yang, Z. R. (2004). Biological applications of support vector machines. *Briefings in Bioinformatics, 5*(4), 328–338.

Yang, H. B., Lou, C. F., Sun, L. X., Li, J., Cai, Y. C., Wang, Z., et al. (2019). AdmetSAR 2.0: Web-service for prediction and optimization of chemical ADMET properties. *Bioinformatics, 35*(6), 1067–1069.

Yang, L., Shu, M., Ma, K., Mei, H., Jiang, Y., & Li, Z. (2010). ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. *Amino Acids, 38*(3), 805–816.

Yue, Y. K., Geng, S., Shi, Y., Liang, G. Z., Wang, J. S., & Liu, B. G. (2019). Interaction mechanism of flavonoids and zein in ethanol-water solution based on 3D-QSAR and spectrofluorimetry. *Food Chemistry, 276*, 776–781.

Zaliani, A., & Gancia, E. (1999). MS-WHIM scores for amino acids: A new 3D-description for peptide QSAR and QSPR studies. *Journal of Chemical Information and Computer Sciences, 39*(3), 525–533.

Zamyatnin, A. A., Borchikov, A. S., Vladimirov, M. G., & Voronina, O. L. (2006). The EROP-Moscow oligopeptide database. *Nucleic Acids Research, 34*(D1), D261–D266.

Zheng, L., Zhao, Y. J., Dong, H. Z., Su, G. W., & Zhao, M. M. (2016). Structure-activity relationship of antioxidant dipeptides: Dominant role of Tyr, Trp, Cys and met residues. *Journal of Functional Foods, 21*, 485–496.