EXCLI Journal 2018;17:734-752 – ISSN 1611-2156
Received: June 20, 2018, accepted: July 10, 2018, published: July 25, 2018

Review article:

UNRAVELING THE BIOACTIVITY OF ANTICANCER PEPTIDES AS DEDUCED FROM MACHINE LEARNING

Watshara Shoombuatong, Nalini Schaduangrat, Chanin Nantasenamat*

Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology,
Mahidol University, Bangkok 10700, Thailand

* Corresponding author: E-mail: chanin.nan@mahidol.edu (C.N.);
Phone: +66 2 441 4371; Fax: +66 2 441 4380

http://dx.doi.org/10.17179/excli2018-1447

ABSTRACT

Cancer imposes a global health burden as it represents one of the leading causes of morbidity and mortality while also giving rise to significant economic burden owing to the associated expenditures for its monitoring and treatment. In spite of advancements in cancer therapy, the low success rate and recurrence of tumor has necessitated the ongoing search for new therapeutic agents. Aside from drugs based on small molecules and protein-based biopharmaceuticals, there has been an intense effort geared towards the development of peptide-based therapeutics owing to its favorable and intrinsic properties of being relatively small, highly selective, potent, safe and low in production costs. In spite of these advantages, there are several inherent weaknesses that are in need of attention in the design and development of therapeutic peptides. An abundance of data on bioactive and therapeutic peptides have been accumulated over the years and the burgeoning area of artificial intelligence has set the stage for the lucrative utilization of machine learning to make sense of these large and high-dimensional data. This review summarizes the current state-of-the-art on the application of machine learning for studying the bioactivity of anticancer peptides along with future outlook of the field. Data and R codes used in the analysis herein are available on GitHub at https://github.com/Shoombuatong2527/anticancer-peptides-review.

Keywords: cancer, anticancer, antitumor, anticancer peptides, host defense peptides, bioactivity, machine learning, QSAR

INTRODUCTION

Cancer is now regarded as the second leading cause of death, and remains a major cause of morbidity throughout the world (Arnold et al., 2015) with lung, liver, colorectal, stomach and breast cancer representing the most common types of cancers occurring worldwide (WHO, 2018a). Estimates from GLOBOCAN indicate that about 14.1 million new cancer cases encompassed approximately 8.8 million deaths in 2015 (Ferlay et al., 2015; WHO, 2018a). In addition, the main mechanisms by which cancers are formed include abnormal, uncontrollable cell growth that leads to the formation of tumors which can then undergo angiogenesis and continue to become metastatic (Felício et al., 2017). Despite recent advances in cancer treatments, such as radiation therapy, targeted therapy or chemotherapeutic agents (Thundimadathil 2012), they have a relatively low success rate and present a risk of recurrence. For instance, the process of killing cancer using chemotherapeutic agents is often associated with deleterious effects, including damages to normal cells and tissues, and lead to chemical resistance whereby adaptation mutations of cancer cells may occur (Hoskin and Ramamoorthy 2008). Therefore, the discovery and development of a new class of anticancer drugs has become crucial. Furthermore, the situation has become worse due to the fact that many new cancers arise from bacterial and viral etiological agents (Vedham et al., 2014). This fact coupled with the increase in antimicrobial resistance (AMR), especially the multidrug resistant variants, has raised concern. To this effect, the WHO has emphasized an urgency for the discovery of new therapeutic agents (WHO, 2018b).

Peptide therapeutics have attracted great interest for development as drug candidates as they are regarded to be safe, efficacious, highly selective with good tolerability as well as exhibit attractive pharmacological profiles (Craik et al., 2013; Vlieghe et al., 2010; Lau and Dunn, 2018; Fosgerau and Hoffmann, 2015). A summary on the strengths and weaknesses of therapeutic peptides are provided in Figure 1. Owing to their intrinsically smaller size as compared to protein-based biopharmaceuticals, peptides are therefore more economical to produce due to lesser production complexity (Fosgerau and Hoffmann, 2015) while at the same time possess more agility in their pharmacokinetics. The aforementioned properties are distinguishing features that set them apart from small molecules-based drugs and protein-based therapeutics. Thus far, there are more than 7,000 naturally occurring peptides in existence that have been shown to afford a wide range of bioactivities (e.g. tumor homing, antihypertensive, antiparasitic, antiviral, antiangiogenic, antibiofilm, antimicrobial, anticancer, etc.) that can consequently be applied to target various diseases such as cancer, diabetes, cardiovascular diseases, etc. (O'Brien-Simpson et al., 2018; Jin and Weinberg 2018; Karpiński and Adamczak 2018). As of now, 60 peptide-based drugs have been FDA-approved (Usmani et al., 2017) while another 150 peptides (Lau and Dunn 2018) are currently in the pipeline of preclinical and clinical studies.

The breakthrough discovery of cecropin, the first antimicrobial peptide (AMP) (i.e. isolated from injecting silk moth, *Hyalophora cecropia*, with bacteria) was reported by Steiner et al. (1981). In another landmark study conducted by Zasloff (1987), AMPs from the African clawed frog, *Xenopus laevis*, were isolated and characterized for their role in the immune defense and is known as magainins. Since then, thousands of AMPs have been found in almost all living organisms such as plants, bacteria, fungi, animals etc. (Li et al., 2012). Over the past decade, the use of AMPs as therapeutic agents for treating diseases have increased constantly.

The process of understanding the importance of AMPs might be useful for the discovery of new and resistance-free therapies for infectious as well as non-infectious diseases. Antimicrobial peptides constitute a mechanism of immune defense of the innate immune system with low antigenicity (Iwasaki et al., 2009; Pasupuleti et al., 2012) that can be found in numerous eukaryotic organisms of different species (Reddy et al., 2004). More recently, research on AMPs have elucidated that these peptides also provide anticancer activity and thus termed anticancer peptides (ACPs). ACPs have been found to exhibit short time-frame of interaction (i.e. decreases the probability of resistance), low toxicity (i.e. not devoid of side effects as it may harm normal cells), specificity, good solubility as well as good tumor penetration thereby indicating the great potential for the use of ACPs in cancer therapy (Domalaon et al., 2016; Gaspar et al., 2015; Figueiredo et al., 2014; Riedl et al., 2011).

Since they are not traditional drugs, the clinical development of therapeutic peptides face numerous challenges owing to their weaknesses as summarized in Figure 1. Stability of peptides (i.e. lack of correlation between *in vitro* experiments and its efficacy in *in vivo* models) is a challenging issue. In spite of this, promising results have been sparingly been demonstrated in some animal studies (Makobongo et al., 2012; Deslouches et al., 2007; Berge et al., 2010; Camilio et al., 2014;

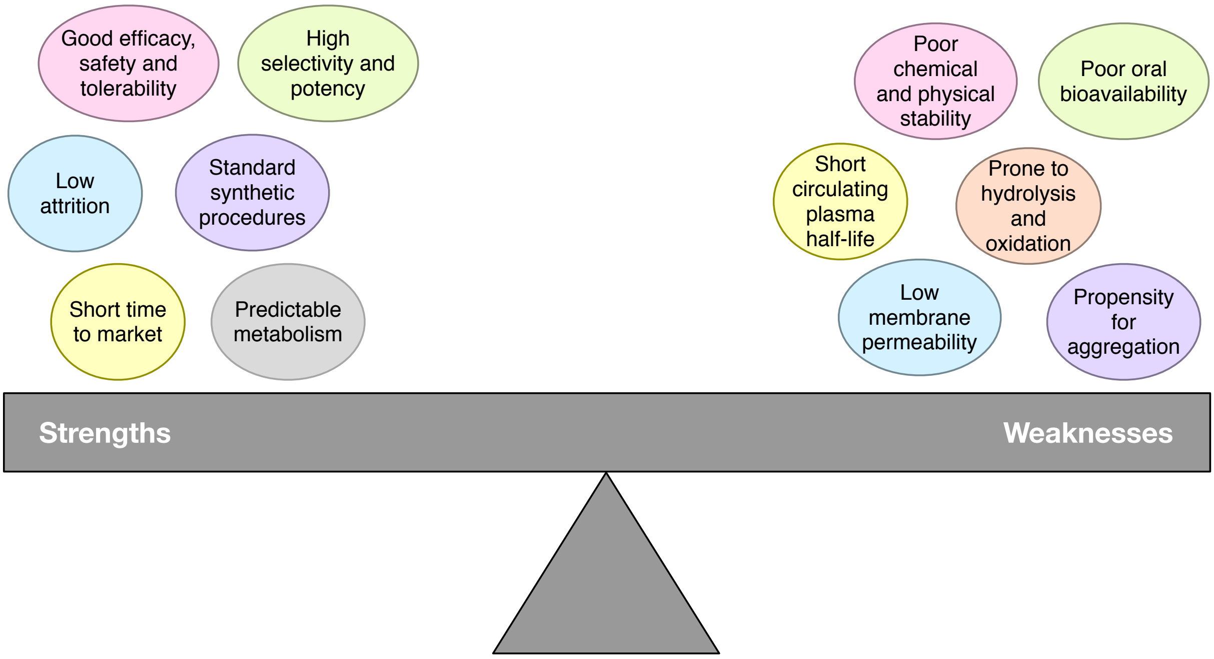
Good efficacy, safety and tolerability
High selectivity and potency
Low attrition
Standard synthetic procedures
Short time to market
Predictable metabolism
Poor chemical and physical stability
Poor oral bioavailability
Short circulating plasma half-life
Prone to hydrolysis and oxidation
Low membrane permeability
Propensity for aggregation
Strengths
Weaknesses

Figure 1: Strengths and weaknesses of therapeutic peptides. Concepts summarized from Fosgerau and Hoffmann, 2015.

Makovitzki et al., 2009) in which good efficacy of peptides were able to establish *in vitro* stability with bioavailability in animal models. Another major drawback of therapeutic peptides is their poor oral bioavailability. This can be addressed by conjugating the peptide with a delivery system that allows it to bypass the digestive system and thus, enhance the pharmacokinetic properties of such peptides. Several studies have been conducted on the modifications and/or conjugations (e.g. substitution with non-canonical amino acid, peptide-peptide hybridization, target or polymer modification, PEGylation etc.) of therapeutic peptides (Narayana et al., 2015; Braunstein et al., 2004; Papo and Shai, 2003; Hu et al., 2016; Spinks et al., 2017; Kelly et al., 2016; Li et al., 2016). Another concern is the short half-life of peptides. However, it should be noted that it is this particular characteristic of therapeutic peptides that allows it to escape resistance unlike other oncogenic therapies. However, research on improving the half-life of peptides without compromising their potency is currently an active area of research (Hao et al., 2015; Podust et al., 2013; Schellenberger et al., 2009; Garay et al., 2012; Penchala et al., 2015). Despite some limitations, no other class of peptides have been able to surpass the multi-functionality of bioactive/therapeutic peptides and thus, these peptides possess high potential for use in many avenues of clinical applications.

The post-genomic era has brought about the birth of several omics (e.g. peptidomics, proteomics, glycomics, transcriptomics, interactomics, etc.) in our attempts to understand the fundamentals of life and how we can contribute to sustainability and the improvement of the quality of life (i.e. development of new diagnostics, therapeutics, etc.). These data are amassing at an exponential rate with no slowing down in hindsight, which sets the stage for the utilization of machine learning in making sense of these data and translating them into useful and actionable insights. There have been extensive reports on the utilization of machine learning approaches for correlating the sequences of therapeutic peptides with their biological activity (Shi et al., 1998; Nagarajan et al., 2006; Alam and Khan, 2014; Mohseni Bababdani and Mousavi, 2013; Tong et al., 2014; Li et al., 2017). A review of the literature indicated that there are currently no review articles concerning the use of machine learning and quantitative

EXCLI Journal 2018;17:734-752 – ISSN 1611-2156
Received: June 20, 2018, accepted: July 10, 2018, published: July 25, 2018

structure-activity relationship (QSAR) as applied to therapeutic peptides. However, there are a few review articles examining the use of QSAR for studying the biological activity of peptides at the general level particularly with emphasis on food protein-derived bioactive peptides (Nongonierma and FitzGerald 2016), peptides in general (as well as proteins and nucleic acids) (Zhou et al., 2008), peptides in general (as well as chemical molecules and proteins) (Du et al., 2008). In a series of recent articles, Lee et al. (2016, 2017, 2018) examined another facet on the use of machine learning (i.e. particularly support vector machine) together with targeted experiments (i.e. killing assays and small-angle X-ray scattering (SAXS) experiments) to explore the membrane activity in undiscovered peptide sequence space in which the aim was not on the antimicrobial activity but on the membrane curvature that is necessary for the activity and the subsequent relationship to sequence homology.

To the best of our knowledge, this review article represents the first systematic review on the utilization of machine learning for studying the bioactivity of anticancer peptides. It is hoped that this review would help contribute to further growth and expansion of the field by providing readers with the current state-of-the-art of the field as well as expected future trends and outlook.

ANTICANCER PEPTIDES

ACPs are small peptides that usually contain 5 to 50 amino acid residues while possessing high hydrophobicity and a positive net charge (i.e. cationic in nature) (Melo et al., 2011). Thus, ACPs can interact with anionic cell membrane components of cancer cells and then selectively kill cancer cells. Additionally, ACPs can interfere with cancer cells by causing apoptosis mediated via mitochondrial disruption (Chen et al., 2001), triggering necrosis via cell lysis (Papo et al., 2006), stimulate the immune system of the host and prevent tumor angiogenesis (Al-Benna et al., 2011). Being a subset of AMPs, the characteristics of ACPs are very similar. However, the physicochemical properties that drive some AMPs to possess anticancer activity is still unclear and more research is needed to understand these differences and help drive specific designs of ACPs. There have been a number of AMPs encountered in nature that possess anticancer activity, such as Aurein 1.2 (GLFDIIKKIAESF) a peptide isolated from a frog species (*Litoria aurea*), represents an AMP with antibacterial activity which was also highly active towards 55 different cancer cell lines *in vitro*, without any significant cytotoxicity (Rozek et al., 2000; Dennison et al., 2007; Giacometti et al., 2007). In addition, the human neutrophil peptide-1 (HNP-1, ACYCRIPACIAGERRYGTCIYQGALWAFCC), represents an intrinsic AMP found in the innate immune system that plays a fundamental role in the defense against pathogens. The full mechanism of action of this peptide against cancer cells has not yet been established, but the activity has already been confirmed for different cancer cell lines, with very low cytotoxicity against healthy cells (McKeown et al., 2006; Gaspar et al., 2015). Furthermore, in terms of their structure, ACPs are mainly categorized as adopting either an α-helix (i.e cecropin, magainin, melittin, and buforin II) or β-sheet (i.e defensins (HNP-1, HNP-2 and HNP-3), lactoferricin B and tachyplesin) conformation due to their inability of fold into a well-defined structure in solution (Hoskin and Ramamoorthy, 2008).

In the more recent years, a lot of focus has been placed on research into ACPs with the increase in AMP databases. One such database, the antimicrobial peptide database (APD3) (Wang et al., 2016) (Available at http://aps.unmc.edu/AP/main.php) recorded as of May 10, 2018, a total of 2,981 AMPs, out of which, 215 have been classified as ACPs from various sources (animals, plants, bacteria, fungi and synthetic) (Figure 2). It should however, be noted that the different categories of the peptides (i.e. antibacterial, antiviral, antiparasitic, anticancer etc.) will

EXCLI Journal 2018;17:734-752 – ISSN 1611-2156
Received: June 20, 2018, accepted: July 10, 2018, published: July 25, 2018

contain peptides that overlap due to some exhibiting dual properties. In addition, upon further analysis of the peptide length determining anticancer activity, it was observed that (Figure 3), out of the 214 ACPs in the database (1 peptide "AP02769" contained a non-canonical amino acid and was excluded from the analysis) 73 (34.11 %) and 60 (28.04 %) were 21-30 and 11-20 amino acids in length, respectively. Furthermore, peptides of length between 21-30 amino acids exhibiting antibacterial, antifungal, antiparasitic and antiviral activities were observed at 746 (29.83 %), 358 (33.58 %), 35 (33.98 %) and 58 (32.22 %), respectively. Therefore, the most optimal peptide length for AMPs, especially for ACPs is 21-30 and hence, it is of great value to optimize the peptide length. Moreover, upon comparison of the most frequently observed amino acid residues constituting each category of AMPs (Figure 4), it can be seen that for ACP functioning, G (Gly at 10.88 %), K (Lys at 10.25 %) and L (Leu at 11.23 %) are the most predominant. Keeping with this tread, the most frequently observed amino acid for all categories of AMPs was Gly which was found at 10.98 %, 10.88 %, 10.79 %, 10.77 % and 11.82 % for ABPs, ACPs, AFPs, APPs and AVPs, respectively. Lysine was also abundantly found in most AMP categories, as indicated in Figure 4, as it is a positively charged residue which could provide improvement in the cell and tissue penetrating properties of peptides (Li and Cho, 2012). In addition, the hydrophobic residue Leucine was also predominant in all AMPs (10.88 %, 11.28 %, 10.66 %, 10.13 % and 8.11 % for ABPs, ACPs, AFPs, APPs and AVPs, respectively) which infers its importance in the structure and function of proteins (Jayaraj et al., 2009) especially since therapeutic peptides usually contain around 50 % hydrophobic residues (Mansour et al., 2014). Furthermore, the anticancer activity of human AMPs have only been evaluated for six peptide classes (Wang et al., 2016) such as HNP-1, HNP-2, HNP-3, hBD-1, LL-37, and granulysin whose structures are shown in Figure 5.

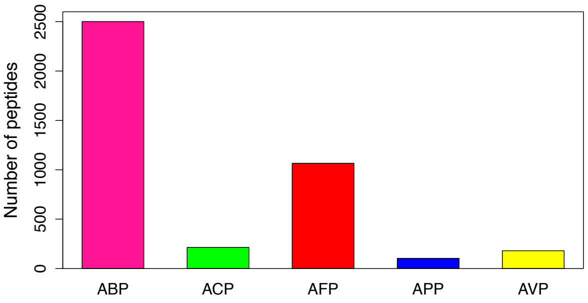

Figure 2: Bar plot of the number of antibacterial peptides (ABP), anticancer peptides (ACP), antifungal peptides (AFP), antiparasitic peptides (APP) and antiviral peptides (AVP). Data is collected from the antimicrobial peptide database (APD3) (Wang et al., 2016).

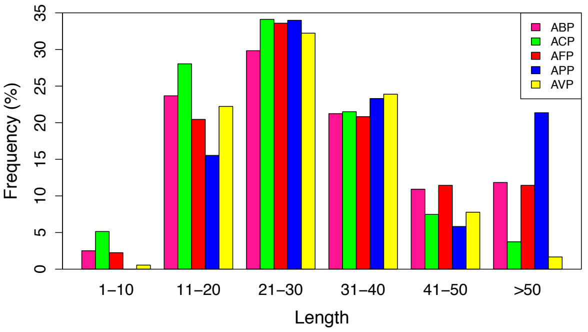

Figure 3: Bar plot showing the peptide length distribution in percentage for antibacterial peptides (ABP), anticancer peptides (ACP), antifungal peptides (AFP), antiparasitic peptides (APP) and antiviral peptides (AVP) collected from the Antimicrobial Peptide Database (APD3) (Wang et al., 2016).

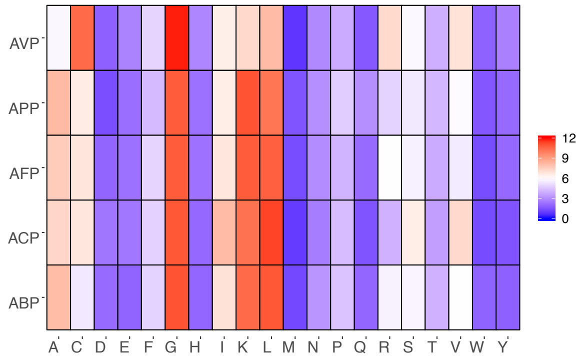

Figure 4: Heat map showing the amino acid compositions in percentage for antibacterial peptides (ABP), anticancer peptides (ACP), antifungal peptides (AFP), antiparasitic peptides (APP) and antiviral peptides (AVP). Data was collected from the Antimicrobial Peptide Database (APD3) (Wang et al., 2016).

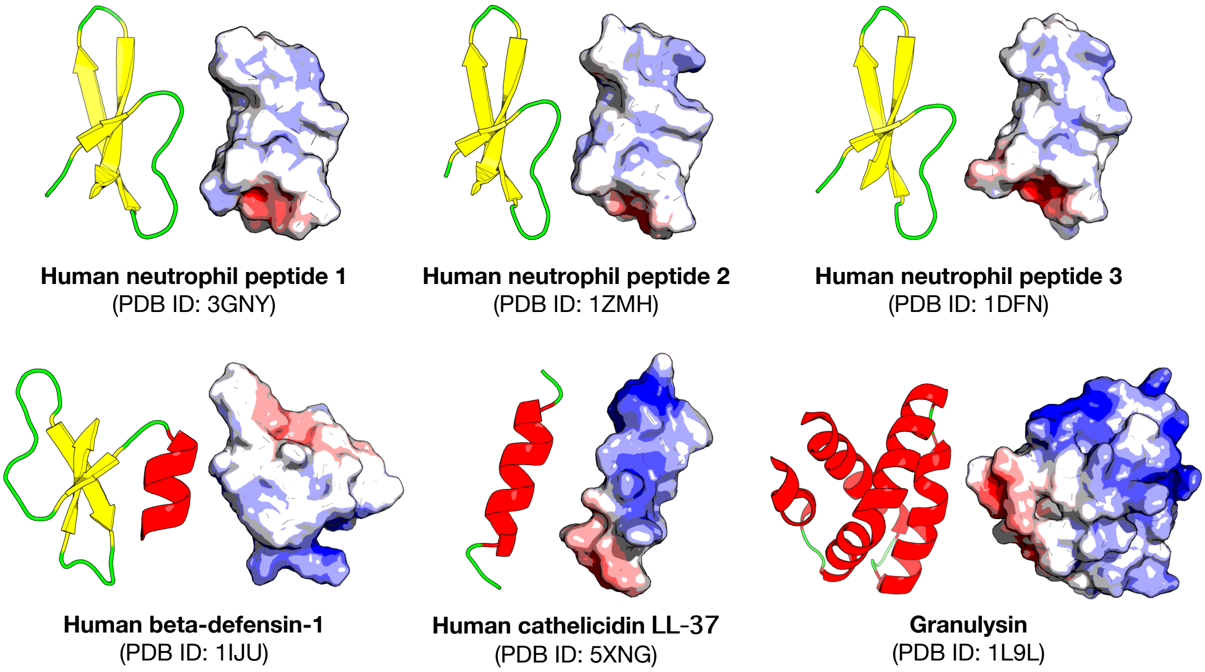

Figure 5: Structures of human-derived anticancer peptides

From thousands of available AMPs and many more that can be synthetically created, only a few have managed to reach clinical trials. Presently, only ten therapeutic peptides to treat various tumor types are currently being evaluated in various phases of preclinical and clinical trials (Felício et al., 2017). This may be due to challenging developmental processes for turning these peptides into potent pharmaceutical drugs (e.g. cost of synthesis, peptide size, charge, and solubility) (Tørfoss et al., 2012). However, with the increase in ACP research, more peptides may reach clinical trials in the future. With the help of synthetic approaches, peptide sequences could be altered so as to enhance their anticancer properties. But, the effect of these structural modifications on the physicochemical properties will need to be elucidated. Recently, these types of studies have increasingly made use of computational approaches (Prada-Gracia et al., 2016; Maccari et al., 2015; Kliger, 2010; Tyagi et al., 2013; Simeon et al., 2017). In addition, several databases exist which have pooled the data of existing sequences that pertain to bioactive or therapeutic peptides. Some of the selected databases are described in Table 1, where, out of all of the individual databases, only one named CancerPPD (Tyagi et al., 2015) is available for ACPs. Five of the databases, are the most comprehensive databases for AMPs that have been combined from various organisms. In addition, it is noteworthy that only one database, THPdb (Usmani et al., 2017) exists whereby data from FDA-approved peptides and proteins are available. Besides the databases mentioned in Table 1, there is another database, TumorHoPe (Kapoor et al., 2012), a database that provides information regarding experimentally characterized tumor targeting/homing peptides. These peptides recognize tumor tissues and tumor-associated micro-environments, including tumor metastasis. Thus, they can be used to deliver drugs selectively in tumors. In addition, a database catering to cell-penetrating peptides, CPPsite (Gautam et al., 2012) that could also be advantageous for recognizing tumors as they exhibit similar properties such as short length (10-30 amino acids), are cationic or amphipathic (containing Arg and Lys residues), and

Table 1: List of selected major databases available for bioactive and therapeutic peptides.

Dataset	Description	No. of entries	URL	Reference
APD3	Database of natural AMPs with defined sequence and activity	2,981	http://aps.unmc.edu/AP/	Wang et al., 2016
BIOPEP	Database of biologically active peptide sequences	3,681	http://www.uwm.edu.pl/biochemia/index.php/en/biopep	Minkiewicz et al., 2008
CAMP	Database of sequences, structures and family-specific signatures of prokaryotic and eukaryotic AMPs	10,247	http://www.camp.bicnirrh.res.in/	Waghu et al., 2016
CancerPPD	Database of experimentally verified anticancer peptides (ACPs) and proteins	3,491	http://crdd.osdd.net/raghava/cancerppd/	Tyagi et al., 2015
DRAMP	Database created with the objective of providing a useful resource for sequence- and structure-activity studies on AMPs	17,349	http://dramp.cpu-bioinfor.org	Fan et al., 2016
LAMP	Tool to aid the discovery and design of AMPs as new antimicrobial agents	5,547	http://biotechlab.fudan.edu.cn/database/lamp	Zhao et al., 2013
PeptideDB	Database of naturally occurring signalling peptides such as cytokines, growth factors, AMPs, peptide hormones etc.	20,027	http://www.peptides.be/	Liu et al., 2008
SATPdb	Database of structurally annotated therapeutic peptides with unique, experimentally validated sequences	19,192	http://crdd.osdd.net/raghava/satpdb/	Singh et al., 2016
THPdb	Database of FDA approved therapeutic peptides and proteins	852	http://crdd.osdd.net/raghava/thpdb/index.html	Usmani et al., 2017

high lipophilicity. More recently, a database dedicated to compiling structural information of bioactive peptides named StraPep (Wang et al., 2018), which currently displays structures for 3,791 peptides as well as provides detailed information for each one (i.e. experimental structure, secondary structure, post-translational modification, etc.).

MACHINE LEARNING

Machine learning is a natural outgrowth of the integration of computer science, mathematics and statistics that allows software application to become accurate in prediction without prior known information (Nasrabadi 2007). The basic application of machine learning is to build algorithms that can formulate a data (a matrix X_{ij}, where each row $\boldsymbol{x}$=(x_{i1}, x_{i2}, x_{i3},..., x_{ii}) is a sample composed of j features) with its proper form and use a prediction model to elucidate an output.

The application of machine learning for correlating the relationship that exists between structures of biological and chemical entities (i.e. peptides and proteins for the former while small molecules for the latter) with their observed or experimentally measured biological activity gives rise to an exciting field of research known as quantitative structure-activity relationship (QSAR). The formulation of a QSAR model entails the generation of quantitative and/or qualitative description of the biological or chemical entities (i.e. known as descriptors) and their subsequent correlation with the biological activity (e.g. IC_{50}, EC_{50}, % activity, etc.) through the use of machine learning algorithms.

Details on the best practices for the development of QSAR models is beyond the scope of this review and readers are directed to previous literature (Nantasenamat and Prachayasittikul, 2015; Tropsha, 2010; Shoombuatong et al., 2017a, b). Briefly, characteristics of a robust QSAR model is best summarized by the OECD principles (OECD, 2014) as outlined in Table 2. In a nutshell, it can be clear that a robust QSAR model should be properly prepared and curated, afford good performance as well as being interpretable so as to facilitate the utilization of the model for gaining insights into the underlying biological activity.

MODEL SET-UP FOR PREDICTING ANTICANCER PEPTIDES

Based on the prior knowledge of peptide sequence analysis, anticancer peptide prediction should be tackled in two associated ways: discriminating ACPs from non-ACPs and then predicting the anticancer activity of such ACPs. Due to the limitation of the experimental approach (e.g. slow and laborious process, expensive, difficulty in peptide purification etc.) for identifying the anticancer activity, computational tools for discriminating ACPs from non-ACPs is an essential way for saving the time-consuming and expensive cost.

Typically, the computational tool construction based on machine learning algorithm consists of four main elements, e.g. data collection, feature representation, model construction and model evaluation (Shoombuatong et al., 2012, 2015a, b, 2016, 2017a, b; Win et al., 2017; Pratiwi et al., 2017; Nantasenamat et al., 2015). In the point of view of machine learning, the use of reliable dataset plays a crucial role to obtain an efficient and generalized model. Previously, there have been many datasets that were used for developing various prediction models as shown in Table 3. Meanwhile, the remaining important elements are listed in Tables 4 and 5. In the following section, a comprehensive summary of previous works in this field are highlighted.

MACHINE LEARNING MODELS FOR THE PREDICTION OF ANTICANCER PEPTIDE

Previously, a variety of computational approaches, including AntiCP (Tyagi et al., 2013), Hajisharifi et al.,'s method (2014), ACPP (Vijayakumar and Ptv, 2015), iACP (Chen et al., 2016), Feng et al.,'s method (Li and Wang, 2016), iACP-GAEnsC (Akbar et al., 2017), Fazlullah et al.,'s method (Khan et al., 2017) and SAP (Xu et al., 2018), have been developed, which will be discussed in the following section. Almost all of the existing methods were developed by using support vector machine (SVM) cooperating with various types of peptide features, except for iACP-GAEnsC (Akbar et al., 2017) that was based on the ensemble approach. The overview of their datasets, type of features, machine learning algorithms and validation methods are shown in Table 5. Meanwhile, Table 6 lists the performance comparison among the existing methods as evaluated by 5-fold CV, 10-fold CV and jackknife test.

Tyagi et al. (2013) first addressed this problem by using SVM-based predictor named AntiCP, in which TY1, TY2 and TY3 datasets were implemented with AAC, DPC and binary profile. The research however, does not specifically state the type of kernel function used. SVM model with ACC/DPC yielded prediction accuracies of 85.52 %/ 85.29 % and 75.70 %/75.20 % respectively, evaluated

Table 2: Summary of the OECD principles for the development of robust QSAR models.

#	OECD principle	Description
1	Defined endpoint	To ensure that the dataset is of high quality; particularly that all endpoint values are free from error
2	Unambiguous algorithm	To ensure the transparency and reproducibility of the QSAR model
3	Defined domain of applicability	To define the boundaries for which the QSAR model is capable of making predictions for query compounds such that they are not too structurally different than those used to train the model
4	Appropriate measures of goodness-of-fit, robustness and predictivity	To rigorously evaluate the performance of the QSAR model
5	Mechanistic interpretation	To ensure that the model can be mechanistically interpreted

Table 3: Summary of all datasets used in this research for evaluating anticancer peptide prediction.

Dataset	Sequence identity (%)[a]	No. of ACP	No. of Non-ACP	Total number	Reference
TY1	100	225	2,250	2,475	Tyagi et al., 2013
TY2	100	225	1,372	1,597	Tyagi et al., 2013
TY3	100	225	225	450	Tyagi et al., 2013
TY_IND	100	50	50	100	Tyagi et al., 2013
ZOH	90	138	206	344	Hajisharifi et al., 2014
SA_TRAIN	100	217	3,979	4,196	Vijayakumar and Ptv, 2015
SA_IND	100	40	40	80	Vijayakumar and Ptv, 2015
SA_RAND	100	-	2,000	2,000	Vijayakumar and Ptv, 2015
WC_IND	100	150	150	300	Chen et al., 2016
LEE	100	422	422	844	Manavalan et al., 2017

[a] Peptides having more than 90 % or 100 % pairwise sequence identity were removed from the dataset.

Table 4: Summary of all peptide features and their feature groups in this research.

Feature name	CS	ATC	PCP	PseCOM	SM
Amino acid composition (AAC)	✓				
Atomic composition (ATC)			✓		
Auto covariance of the average chemical shift (acACS)		✓			
Amphiphilic pseudo amino acid composition (Am-PseAAC)				✓	
Binary profile (BP)	✓				
Dipeptide composition (DPC)	✓				
G-Gap dipeptide composition (g-gap DPC)	✓				
Local alignment kernel					✓
Physicochemical properties (PCP)			✓		
Pseudo amino acid composition (PseACC)				✓	
Pseudo G-Gap dipeptide composition (Pse-g-gap DPC)				✓	
Protein relatedness measure (PRM)			✓		
Reduce amino acid composition (RACC)			✓		
Split amino acid composition (SAAC)	✓				

CS: Composition, ATC: Autocorrelation, PCP: Physicochemical properties, PseCOM: Pseudo Composition, SM: Similarity measure

EXCLI Journal 2018;17:734-752 – ISSN 1611-2156
Received: June 20, 2018, accepted: July 10, 2018, published: July 25, 2018

Table 5: Summary of existing methods for predicting anticancer peptides.

Method	Classifier[a]	Sequence feature (No.)[b]	Testing method[c]	Web-server	Reference
AntiCP	SVM	AAC, DPC, BP (200)	10-fold CV and independent test	✓	Tyagi et al., 2013
Hajisharifi et al.,'s method	SVM	PseACC, LAK (>200)	5-fold CV and independent test		Hajisharifi et al., 2014
ACPP	SVM, AdaBoost	PRM (60)	10-fold CV and independent test	✓[d]	Vijayakumar and Ptv, 2015
iACP	SVM	g-gap DPC (400)	5-fold CV, Jackknife and independent test	✓	Chen et al., 2016
Li and Wang's method	SVM	AAC, RACC, acACS (80)	5-fold CV, Jackknife		Li and Wang, 2016
Khan et al.,'s method	SVM, *k*-NN	SAAC, DPC, PseAAC (552)	Jackknife		Khan et al., 2017
iACP-GAEnsC	Ensemble, SVM, *k*-NN, PNN, RF, GRNN	Pse-g-gap DPC, Am-Pse-AAC, RACC (588)	Jackknife		Akbar et al., 2017
MLACP	SVM, RF	AAC, ATC, DPC, PCP (436)	10-fold CV	✓	Manavalan et al., 2017
SAP	SVM, RF, LibD3C	g-gap DPC (400)	5-fold CV		Xu et al., 2018

[a] *k*-NN: *k*-nearest Neighbor, GRNN: generalized neural network, LibD3C: hybrid model of ensemble pruning, PNN: probabilistic neural network, RF: random forest, SVM: support vector machine.
[b] AAC: amino acid composition, ATC: atomic composition , acACS: auto covariance of the average chemical shift, Am-PseAAC: amphiphilic pseudo amino acid composition, BP: binary profile, DPC: dipeptide composition, g-gap DPC: G-Gap dipeptide composition, LAK: local alignment kernel, PCP: Physicochemical properties, PseACC: Pseudo amino acid composition, Pse-g-gap DPC: Pseudo G-Gap dipeptide composition, PRM: protein relatedness measure, RACC: reduce amino acid composition, SAAC: split amino acid composition.
[c] 5-fold CV: 5-fold cross-validation, 10-fold CV: 10-fold cross-validation
[d] The webserver version is currently unavailable.

Table 6: Performance benchmark comparing various computational methods evaluated by 5- and 10-fold cross-validation and jackknife test.

Method	Testing method	Benchmarking dataset	Accuracy (%)	MCC	Reference
AntiCP	10-fold CV	TY3	91.44	0.83	Tyagi et al., 2013
Hajishaifi et al.,'s method	5-fold CV	ZOH	89.70	0.78	Hajisharifi et al., 2014
ACPP	10-fold CV	SA_TRAIN	97.70	0.92	Vijayakumar and Ptv 2015
iACP	5-fold CV	ZOH	94.77	0.89	Chen et al., 2016
Li and Wang's method	Jackknife	ZOH	93.61	0.88	Li and Wang 2016
Khan et al.,'s method	Jackknife	ZOH	93.31	0.86	Khan et al., 2017
iACP-GAEnsC	Jackknife	ZOH	96.45	0.91	Akbar et al., 2017
MLACP	10-fold CV	LEE	96.40	0.89	Manavalan et al., 2017
SAP	5-fold CV	ZOH	91.86	0.83	Xu et al., 2018

5-fold CV: 5-fold cross-validation, 10-fold CV: 10-fold cross-validation

Received: June 20, 2018, accepted: July 10, 2018, published: July 25, 2018

by a 10-fold CV method on TY1 and TY2 datasets. These results revealed that the importance of ACC feature for enhancing ACP prediction was not quite different from DPC feature. But, when binary (NT10) based models were applied, where NT10 was the first 10 residues and each amino acid was represented by (20*10)-dimensional vector, the accuracy improved to 91.44 %. Finally, SVM based on the NT10 models performed well with 89 % accuracy and 0.78 MCC on TY_IND dataset. Finally, a web-server (Available at http://crdd.osdd.net/raghava/anticp/) was developed to help experimental scientists in predicting minimum mutations required for improving anticancer potency, virtual screening of peptides for discovering novel anticancer peptides and scanning natural proteins for identification of ACPs.

Hajisharifi et al., (2014) took advantage of PseAAC feature and local alignment kernels for improving the prediction performance of the model. In the study, the benchmark ZOH dataset was firstly created by collecting data from the antimicrobial peptide database (APD2) (Wang et al., 2009, available at http://aps.unmc.edu/AP/.) The ZOH dataset consisted of 192 ACPs and 215 non-ACPs and then, to prevent an overestimation of prediction results due to highly similar sequences, peptides with more than 90 % similarity were removed from the initial ZOH dataset using CD-HIT (Li and Godzik, 2006). Finally, a total of 138 ACPs and 206 non-ACPs were gained as summarized in Table 2. SVM model conjunction with PseACC feature showed the values of accuracy, sensitivity, specificity and MCC of 83.82 %, 81.84 %, 85.36 % and 0.66, respectively, evaluated by a 5-fold CV procedure. Meanwhile, using a local alignment kernel yielded better prediction results than PseACC feature with improvements of > 6 % and 10 % on both Ac and MCC, respectively.

Only one year later, Vijayakumar and Ptv (2015) utilized two powerful SVM and AdaBoost models cooperating with the protein relatedness measure (PRM) parameters called ACPP. The PRM feature represents each peptide with the degree distribution of amino acids deviating from a theoretical protein/peptide. To build a prediction model, SVM model with radial basis function (RBF) kernel and the tuning *cost* and *gamma* parameters of 2 and 0.0078, respectively, were used, while AdaBoost model based on the linear combination of simple weak classifiers with the tuning number of 10 iterations was applied. In this study, ACPP was evaluated with a 10-fold cross-validation method and independent test. SVM and AdaBoost were first carried out on the imbalanced dataset containing 217 ACPs and 3,979 non-ACPs as summarized in Table 2. The prediction results showed that SVM and AdaBoost yielded MCC values as low as 0.59 and 0.57, while, the balanced dataset (217 peptide sequences on both ACPs and non-ACPs), yielded increased accuracies for SVM and AdaBoost of 0.92 and 0.88, respectively. Based on these results, the authors stated that the PRM feature adopted to classify ACPs from non-ACPs was effective. Although, in this study, a web-server was established at http://acpp.bicpu.edu.in/predict.php, however, it is currently unavailable.

In 2016, there were two different research groups that made efforts to develop ACP predictors, i.e. iACP (Chen et al., 2016) and Feng et al.,'s method (Li and Wang, 2016). Chen et al. (2016) proposed an approach to take advantage of SVM model in conjunction with g-gap dipeptide compositions (g-gap DPC), where g = 0, 1, 2, 3 or 4 and g =0 is DPC, as well as working together with ANOVA (analysis of variance). Herein, SVM model with radial basis function (RBF) kernel and their optimal parameter of *cost* = 2 and *gamma* = 0.125 were used. The ANOVA approach via the incremental feature selection (IFS) was used for selecting informative features among g-gap DPCs. The process of determining the optimal number of features was conducted according to the following steps: (1) the feature with the highest F-score was selected as the input of SVM and the prediction performance assessed with 5-fold CV was calculated to

Received: June 20, 2018, accepted: July 10, 2018, published: July 25, 2018

evaluate the performance of this feature; (2) the feature with the second highest F-score was then combined with the first feature to form a new feature subset and the prediction performance with the criteria was still used to estimate the performance of the new feature subset; (3) this process was done when the prediction performance of 400 features were calculated. The highest accuracy of 94.77 % can be achieved by using g=1 and the 126 top-ranked informative features. Li and Wang (2016) attempted to improve the prediction performance by using SVM model with hybrid composition, i.e. AAC, auto covariance of the average chemical shift (acACS) and reduced amino acid composition (RAAC). The parameters of RBF kernel used were tuned using the grid search method. Initial prediction results for their model using AAC on the ZOH dataset showed the value of accuracy and MCC of 91.86 % and 0.83, respectively. The second and third highest accuracies were obtained from using RACC (84.01 %) and asACS (82.56 %), respectively. Meanwhile, the combination features of AAC, RAAC and acACS performed best with 93.61 % accuracy and 0.87 MCC. The authors of this paper suggested that these combination features were helpful to the prediction of ACPs.

In 2017, three different ACP predictors were developed with various types of machine learning algorithms and peptide features, i.e. Khan et al.,'s method (2017), iACP-GAEnsC (Akbar et al., 2017) and MLACP (Manavalan et al., 2017). Khan et al. (2017) utilized SVM and *k*-nearest Neighbor (*k*-NN) models with a variety of peptide features, i.e. split amino acid composition (SAAC), DPC and PseAAC, to find the suitable feature for discriminating ACPs from non-ACPs. The total number of feature spaces of SAAC, DPC and PseAAC were 400, 62 and 60, respectively. To build prediction models, authors used RBF kernel to create SVM model, while euclidian distance was used to compute the distance among the peptide sequences. The optimum parameters of these two models were obtained during the training phase. The performance comparison evaluated by jackknife test demonstrated that SVM and *k*-NN models using SAAC outperformed the other two features with an accuracy of 93.31 % and 90.17 %. Akbar et al. (2017) examined the ability of a variety of machine learning algorithms, i.e. SVM, random forest (RF), *k*-nearest Neighbor (*k*-NN), generalized neural network (GRNN), and probabilistic neural network (PNN). In this study, each peptide was represented by three different feature extraction schemes using RAAP, Pse-g-Gap dipeptide composition (Pse-g-gap DPC) and amphiphilic PseAAC (Am-PseACC). Finally, the evolution genetic algorithm was used to measure the diversity and optimum outcome or prediction results of the different methods called iACP-GAEnsC. Initial prediction results showed that using Am-PseACC with jackknife test achieved accuracies of 93.60 %, 90.41 %, 91.28 %, 86.33 % and 93.89 % for SVM, *k*-NN, PNN, RF, GRNN and GAEnsC, respectively. Their best accuracy of 94.45 % was achieved by using an ensemble approach with the merging of SVM, *k*-NN, PNN, RF and GRNN associated with a hybrid feature of RAAP, Pse-g-gap DPC and Am-PseACC. Manavalan et al. (2017) developed machine learning-based methods (SVM and RF), named SVMACP, RFACP and MLACP using a combination of features, including ACC, DPC, PCP and ATC. The number of dimensions for ACC, DPC, PCP and ATC features were 20, 400, 11, 5, respectively. For each model, authors optimized the RF (*ntree* and *mtry*) and SVM (*cost* and *gamma*) parameters by using 10-fold CV on the TY3 dataset. In the case of using a single feature, RFACP and SVMACP yielded accuracies ranging from 81.4 %-86.8 % and 75.9 %-85.8 %, respectively. The best accuracy and MCC of 87.2 and 0.70, respectively, was achieved by using RF model with the combination feature of ACC, DPC, PCP and ATC.

Recently, Xu et al. (2018) developed the MRMD method to select important features from g-gap DPC. The selected, informative feature was used as an input feature to train the

the SVM model called SAP. The paper does not specifically state the type of kernel function used. For a 5-fold CV, SAP using all 400 features yielded 91.86 % accuracy and 0.83 MCC, while using selected features offered a 90.70 % accuracy and 0.81 MCC. Furthermore, SAP was also compared with RF and LibD3D, where LibD3D is a selective ensemble model. The overall accuracy comparison showed that SAP (91.78 %) was quite comparable with RF (91.88 %) and LibD3D (89.24 %) models.

The aforementioned articles showed promising results in the use of various types of machine learning algorithms and peptides features as summarized in Tables 5 and 6. As seen in Table 3, the ZOH is known as the valid benchmark dataset used for developing various prediction models (Hajisharifi et al., 2014; Chen et al., 2016; Xu et al., 2018; Khan et al., 2017; Akbar et al., 2017). Amongst these methods, iACP (Chen et al., 2016) and iACP-GAEnsC (Akbar et al., 2017) showed their best predictive accuracies of 94.77 % and 96.45 % as evaluated by 5-fold CV and jackknife test procedures, respectively. In addition, iACP revealed its efficiency by carrying out an independent WC_IND data achieving an accuracy and MCC of 92.67 % and 0.85, respectively. Considering that the independent test is the most rigorous cross-validation method, it might be stated that iACP (Chen et al., 2016) was superior to other prediction methods as demonstrated in Table 5. Amongst the existing methods, some of them (Tyagi et al., 2013; Chen et al., 2016; Manavalan et al., 2017) determined the important amino acids and dipeptide that were enriched in anticancer peptides using componential analysis (Manavalan et al., 2017; Tyagi et al., 2013) and F-score (Chen et al., 2016).

BIOLOGICAL INSIGHTS FROM PREDICTIVE MODELS

Feature importance analysis from existing models (Tyagi et al., 2013; Chen et al., 2016; Manavalan et al., 2017 indicated that in general anticancer peptides are abundant in Cys, Glu, Phe, Gly, lle, Lys and Phe when compared to non-anticancer peptides (Chen et al., 2016). Particularly, Tyagi et al. (2013) reported that Gly, Leu, Ala and Phe were preferential residues at the N-terminus of anticancer peptides while Val, Cys, Leu and Lys were likely to be found at the C-terminus. Furthermore, Manavalan et al. (2017) revealed that the 10 top-ranking features in anticancer peptides were comprised of dipeptides rich in positively charged and aromatic residues (e.g. KK, AK, KL, AL, KA, KW, LA, LK, FA and LF). Moreover, it should also be noted that desirable trait for anticancer peptides is their cell penetrating ability such that they can specifically neutralize their target while maintaining low toxicity.

LIMITATIONS OF CURRENT MACHINE LEARNING MODELS

The use of machine learning algorithm is one of the important factors in the steady growth of the field of anticancer drug discovery and development. Most of the reported anticancer peptide prediction methods were mainly developed in order to enhance the prediction accuracy by taking advantage of the complexity of prediction methods and the number of feature types. Overall, most research articles showed encouraging results with having satisfied accuracies of more than 90 %. Nevertheless, there is still room for development to improve the existing methods as useful and interpretable models for facilitating experimental scientists and related researchers as demonstrated by a series of recent publications (Shoombuatong et al., 2012, 2015a, b, 2016, 2017a, b; Win et al., 2017; Pratiwi et al., 2017; Nantasenamat et al., 2015) and summarized in comprehensive reviews (Nantasenamat et al., 2015; Shoombuatong et al., 2017a, b).

In addition, the most commonly used benchmark dataset ZOH, (Hajisharifi et al., 2014) consisted of 138 ACPs and 206 non-ACPs in which only ACPs were derived from the experimental verification method. It could be stated that existing methods developed by

Received: June 20, 2018, accepted: July 10, 2018, published: July 25, 2018

ZOH dataset might not be completely suited to accurately filter experimentally verified non-ACPs from ACPs. Furthermore, peptide features were intrinsically heterogeneous, noisy and multi-dimensional, but only a few existing methods (Chen et al., 2016; Xu et al., 2018) took advantage of feature selection techniques to qualify and rank the importance and the contributions of the features for the model performance. Thus, these method has utilized only partial information of the biological activity of ACP. It could be stated that the role of different types of peptide features contributing to the biological activity of anticancer peptide are still poorly understood. Additionally, a variety of methods were used to evaluate the prediction performance of ACP predictions as listed in Table 5, including N-fold cross-validation, where N is 5 or 10, jackknife test and independent test. The independent test is an effective way to test the performance of a model in real-world applications and verify the generalization of a model, but only few existing methods (Tyagi et al., 2013; Vijayakumar and Ptv, 2015; Chen et al., 2016; Li and Wang, 2016) were assessed with this method. Finally, according to the fifth principle of OEC which states that, it is necessary and significant of an interpretable QSAR model to provide important factors that can enhance the biological activity of peptides or compounds. Amongst the existing methods, some of them (Tyagi et al., 2013; Chen et al., 2016; Manavalan et al., 2017) provided the results of feature importance analysis by using componential analysis (Manavalan et al., 2017; Tyagi et al., 2013) and F-score (Chen et al., 2016). However, they did not clearly mention which features contributed most to prediction performance. Moreover, the SVM model was not straight-forward enough to interpret the underlying biological implications of anticancer peptides.

CONCLUSION

The success story of therapeutic peptides is starting to gain moment with more than 60 approved by the FDA and more than 150 peptides have reached pre-clinical and clinical stages. In addition, a literature review have indicated that there is a large volume of on-going studies being carried out in the field. In spite of the large sum of papers on the utilization of machine learning approaches for the development of QSAR models of bioactive and therapeutic peptides, however there are few review articles that examine the field in a systematic manner. It is the intent of this review article to fill this gap by providing readers with the current advancements pertaining to the current state-of-the-art on the prediction of anticancer peptides via the use of machine learning approaches.

A survey of existing QSAR models against anticancer peptides suggested that almost all provided reasonably high prediction accuracies and in spite of this, there are limiting factors that may hinder their full potential for application as follows:

(i) Absence of experimentally verified non-anticancer peptides.
(ii) Inclusion of trivial and non-informative features during the model building process.
(iii) Lack of comprehensive evaluation method and failure to make use of interpretable learning methods.

In efforts to augment the robustness of the predictive model, herein are recommendations:

(i) Increase the size of the peptide dataset by combining all data sources together as to capture as much as possible of the pattern of dataset for alleviating uncertainties in the prediction system.
(ii) Familiarize oneself with the background and details of the descriptors being used such that the resulting models could be interpreted in a meaningful manner as to gain biological insights for guiding further experiments.
(iii) Use interpretable learning algorithms as to allow the interpretation of important features responsible for the biological activity.

(iv) Ensure that the model is externally validated on an independent test set as well as defining the applicability domain of the model as to verify the ability of the model for extrapolation to future unknown data.

(v) Ensure the reproducibility of constructed models such that interested users could extend the model further.

(vi) If possible, constructed models should be made publicly available in the form of webservers so as to facilitate easy access to the model's prediction capability.

Conflict of interests

The authors declare that no competing interests exist.

Acknowledgements

This work is supported by the Office of Higher Education Commission and the Thailand Research Fund (No. MRG6180226); the New Researcher Grant (A31/2561) from Mahidol University; and the Center of Excellence on Medical Biotechnology (CEMB), S&T Postgraduate Education and Research Development Office (PERDO), Office of Higher Education Commission (OHEC), Thailand. Partial support from the annual budget grant (B.E. 2557-2559) of Mahidol University is also acknowledged.

REFERENCES

Akbar S, Hayat M, Iqbal M, Jan MA. iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. Artif Intell Med. 2017;79:62–70.

Alam S, Khan F. QSAR and docking studies on xanthone derivatives for anticancer activity targeting DNA topoisomerase IIα. Drug Des Devel Ther. 2014;8:183–95.

Al-Benna S, Shai Y, Jacobsen F, Steinstraesser L. Oncolytic activities of host defense peptides. Int J Mol Sci. 2011;12:8027–51.

Arnold M, Karim-Kos HE, Coebergh JW, Byrnes G, Antilla A, Ferlay J, et al. Recent trends in incidence of five common cancers in 26 European countries since 1988: Analysis of the European Cancer Observatory. Eur J Cancer. 2015;51:1164–87.

Berge G, Eliassen LT, Camilio KA, Bartnes K, Sveinbjørnsson B, Rekdal O. Therapeutic vaccination against a murine lymphoma by intratumoral injection of a cationic anticancer peptide. Cancer Immunol Immunother. 2010;59:1285–94.

Braunstein A, Papo N, Shai Y. *In vitro* activity and potency of an intravenously injected antimicrobial peptide and its DL amino acid analog in mice infected with bacteria. Antimicrob Agents Chemother. 2004;48: 3127–9.

Camilio KA, Berge G, Ravuri CS, Rekdal O, Sveinbjørnsson B. Complete regression and systemic protective immune responses obtained in B16 melanomas after treatment with LTX-315. Cancer Immunol Immunother. 2014;63:601–13.

Chen W, Ding H, Feng P, Lin H, Chou K-C. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016;7:16895–909.

Chen Y, Xu X, Hong S, Chen J, Liu N, Underhill CB, et al. RGD-Tachyplesin inhibits tumor growth. Cancer Res. 2001;61:2434–8.

Craik DJ, Fairlie DP, Liras S, Price D. The future of peptide-based drugs. Chem Biol Drug Des. 2013;81: 136–47.

Dennison SR, Harris F, Phoenix DA. The interactions of aurein 1.2 with cancer cell membranes. Biophys Chem. 2007;127:78–83.

Deslouches B, Gonzalez IA, DeAlmeida D, Islam K, Steele C, Montelaro RC, et al. De novo-derived cationic antimicrobial peptide activity in a murine model of Pseudomonas aeruginosa bacteraemia. J Antimicrob Chemother. 2007;60:669–72.

Domalaon R, Findlay B, Ogunsina M, Arthur G, Schweizer F. Ultrashort cationic lipopeptides and lipopeptoids: Evaluation and mechanistic insights against epithelial cancer cells. Peptides. 2016;84:58–67.

Du Q-S, Huang R-B, Chou K-C. Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. Curr Protein Pept Sci. 2008;9:248–60.

Fan L, Sun J, Zhou M, Zhou J, Lao X, Zheng H, et al. DRAMP: a comprehensive data repository of antimicrobial peptides. Sci Rep. 2016;6:24482.

Felício MR, Silva ON, Gonçalves S, Santos NC, Franco OL. Peptides with dual antimicrobial and anticancer activities. Front Chem. 2017;5:5.

Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136:E359-86.

Figueiredo CR, Matsuo AL, Massaoka MH, Polonelli L, Travassos LR. Anti-tumor activities of peptides corresponding to conserved complementary determining regions from different immunoglobulins. Peptides. 2014;59:14–9.

Fosgerau K, Hoffmann T. Peptide therapeutics: current status and future directions. Drug Discov Today. 2015; 20:122–8.

Garay RP, El-Gewely R, Armstrong JK, Garratty G, Richette P. Antibodies against polyethylene glycol in healthy subjects and in patients treated with PEG-conjugated agents. Expert Opin Drug Deliv. 2012;9:1319–23.

Gaspar D, Freire JM, Pacheco TR, Barata JT, Castanho MARB. Apoptotic human neutrophil peptide-1 anti-tumor activity revealed by cellular biomechanics. Biochim Biophys Acta. 2015;1853:308–16.

Gautam A, Singh H, Tyagi A, Chaudhary K, Kumar R, Kapoor P, et al. CPPsite: a curated database of cell penetrating peptides. Database. 2012;2012: bas015.

Giacometti A, Cirioni O, Riva A, Kamysz W, Silvestri C, Nadolski P, et al. *In vitro* activity of aurein 1.2 alone and in combination with antibiotics against gram-positive nosocomial cocci. Antimicrob Agents Chemother. 2007;51:1494–6.

Hajisharifi Z, Piryaiee M, Mohammad Beigi M, Behbahani M, Mohabatkar H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. J Theor Biol. 2014;341:34–40.

Hao X, Yan Q, Zhao J, Wang W, Huang Y, Chen Y. TAT modification of alpha-helical anticancer peptides to improve specificity and efficacy. PLoS ONE. 2015; 10:e0138911.

Hoskin DW, Ramamoorthy A. Studies on anticancer activities of antimicrobial peptides. Biochim Biophys Acta. 2008;1778:357–75.

Hu C, Chen X, Zhao W, Chen Y, Huang Y. Design and modification of anticancer peptides. Drug Des. 2016;5: 138.

Iwasaki T, Ishibashi J, Tanaka H, Sato M, Asaoka A, Taylor D, et al. Selective cancer cell cytotoxicity of enantiomeric 9-mer peptides derived from beetle defensins depends on negatively charged phosphatidylserine on the cell surface. Peptides. 2009;30:660–8.

Jayaraj V, Suhanya R, Vijayasarathy M, Anandagopu P, Rajasekaran E. Role of large hydrophobic residues in proteins. Bioinformation. 2009;3:409–12.

Jin G, Weinberg A. Human antimicrobial peptides and cancer. Semin Cell Dev Biol. 2018; In Press. DOI: 10.1016/j.semcdb.2018.04.006.

Kapoor P, Singh H, Gautam A, Chaudhary K, Kumar R, Raghava GPS. TumorHoPe: a database of tumor homing peptides. PLoS ONE. 2012;7:e35187.

Karpiński TM, Adamczak A. Anticancer activity of bacterial proteins and peptides. Pharmaceutics. 2018; 10(2):54.

Kelly GJ, Kia AF-A, Hassan F, O'Grady S, Morgan MP, Creaven BS, et al. Polymeric prodrug combination to exploit the therapeutic potential of antimicrobial peptides against cancer cells. Org Biomol Chem. 2016; 14:9278–86.

Khan F, Akbar S, Basit A, Khan I, Akhlaq H. Identification of anticancer peptides using optimal feature space of chou's split amino acid composition and support vector machine. In: Proceedings of the 2017 4th International Conference on Biomedical and Bioinformatics Engineering - ICBBE 2017;2017:91–6.

Kliger Y. Computational approaches to therapeutic peptide discovery. Biopolymers. 2010;94:701–10.

Lau JL, Dunn MK. Therapeutic peptides: Historical perspectives, current development trends, and future directions. Bioorg Med Chem. 2018;26:2700–7.

Lee EY, Fulan BM, Wong GCL, Ferguson AL. Mapping membrane activity in undiscovered peptide sequence space using machine learning. Proc Natl Acad Sci USA. 2016;113:13588–93.

Lee EY, Lee MW, Fulan BM, Ferguson AL, Wong GCL. What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? Interface Focus. 2017;7:20160153.

Lee EY, Wong GCL, Ferguson AL. Machine learning-enabled discovery and design of membrane-active peptides. Bioorg Med Chem. 2018;26: 2708-18.

Li F-M, Wang X-Q. Identifying anticancer peptides by using improved hybrid compositions. Sci Rep. 2016;6: 33910.

Li H, Anuwongcharoen N, Malik AA, Prachayasittikul V, Wikberg JES, Nantasenamat C. Roles of D-amino acids on the bioactivity of host defense peptides. Int J Mol Sci. 2016;17(7):1023.

Li H, Schaduangrat N, Simeon S, Nantasenamat C. Computational study on the origin of the cancer immunotherapeutic potential of B and T cell epitope peptides. Mol Biosyst. 2017;13:2310–22.

Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.

Li Y, Xiang Q, Zhang Q, Huang Y, Su Z. Overview on the recent study of antimicrobial peptides: origins, functions, relative mechanisms and application. Peptides. 2012;37:207–15.

Li ZJ, Cho CH. Peptides as targeting probes against tumor vasculature for diagnosis and drug delivery. J Transl Med. 2012;10(Suppl 1):S1.

Liu F, Baggerman G, Schoofs L, Wets G. The construction of a bioactive peptide database in Metazoa. J Proteome Res. 2008;7:4119–31.

Maccari G, Di Luca M, Nifosì R. In silico design of antimicrobial peptides. Methods Mol Biol. 2015;1268: 195–219.

Makobongo MO, Gancz H, Carpenter BM, McDaniel DP, Merrell DS. The oligo-acyl lysyl antimicrobial peptide $C_{12}K-2\beta_{12}$ exhibits a dual mechanism of action and demonstrates strong *in vivo* efficacy against *Helicobacter pylori*. Antimicrob Agents Chemother. 2012; 56:378–90.

Makovitzki A, Fink A, Shai Y. Suppression of human solid tumor growth in mice by intratumor and systemic inoculation of histidine-rich and pH-dependent host defense-like lytic peptides. Cancer Res. 2009;69: 3458–63.

Manavalan B, Basith S, Shin TH, Choi S, Kim MO, Lee G. MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget. 2017;8:77121–36.

Mansour SC, Pena OM, Hancock REW. Host defense peptides: front-line immunomodulators. Trends Immunol. 2014;35:443–50.

McKeown STW, Lundy FT, Nelson J, Lockhart D, Irwin CR, Cowan CG, et al. The cytotoxic effects of human neutrophil peptide-1 (HNP1 and lactoferrin on oral squamous cell carcinoma (OSCC *in vitro*. Oral Oncol. 2006;42:685–90.

Melo MN, Ferre R, Feliu L, Bardají E, Planas M, Castanho MARB. Prediction of antibacterial activity from physicochemical properties of antimicrobial peptides. PLoS ONE. 2011;6:e28549.

Minkiewicz P, Dziuba J, Iwaniak A, Dziuba M, Darewicz M. BIOPEP database and other programs for processing bioactive peptide sequences. J AOAC Int. 2008;91:965–80.

Mohseni Bababdani B, Mousavi M. Gravitational search algorithm: A new feature selection method for QSAR study of anticancer potency of imidazo[4,5-b]pyridine derivatives. Chemometr Intell Lab Syst. 2013;122:1–11.

Nagarajan V, Kaushik N, Murali B, Zhang C, Lakhera S, Elasri MO, et al. A Fourier transformation based method to mine peptide space for antimicrobial activity. BMC Bioinformatics. 2006;7(Suppl 2):S2.

Nantasenamat C, Prachayasittikul V. Maximizing computational tools for successful drug discovery. Expert Opin Drug Discov. 2015;10:321–9.

Nantasenamat C, Worachartcheewan A, Jamsak S, Preeyanon L, Shoombuatong W, Simeon S, et al. AutoWeka: toward an automated data mining software for QSAR and QSPR studies. Methods Mol Biol. 2015; 1260:119–47.

Narayana JL, Huang H-N, Wu C-J, Chen J-Y. Efficacy of the antimicrobial peptide TP4 against *Helicobacter pylori* infection: *in vitro* membrane perturbation via micellization and *in vivo* suppression of host immune responses in a mouse model. Oncotarget. 2015;6: 12936–54.

Nasrabadi NM. Pattern recognition and machine learning. J Electron Imaging. 2007;16:049901.

Nongonierma AB, FitzGerald RJ. Learnings from quantitative structure–activity relationship (QSAR studies with respect to food protein-derived bioactive peptides: a review. RSC Adv. 2016;6:75400–13.

O'Brien-Simpson NM, Hoffmann R, Chia CSB, Wade JD. Antimicrobial and anticancer peptides (Editorial). Front Chem. 2018;6:13.

OECD (Organization For Economic Co-operation and Development). Guidance Document on the Validation of (Quantitative Structure-Activity Relationship [(QSAR] Models. Paris: OECD, 2014 (OECD Series on Testing and Assessment, No. 69).

Papo N, Shai Y. New lytic peptides based on the D,L-amphipathic helix motif preferentially kill tumor cells compared to normal cells. Biochemistry. 2003;42: 9346–54.

Papo N, Seger D, Makovitzki A, Kalchenko V, Eshhar Z, Degani H, et al. Inhibition of tumor growth and elimination of multiple metastases in human prostate and breast xenografts by systemic inoculation of a host defense-like lytic peptide. Cancer Res. 2006;66:5371–8.

Pasupuleti M, Schmidtchen A, Malmsten M. Antimicrobial peptides: key components of the innate immune system. Crit Rev Biotechnol. 2012;32:143–71.

Penchala SC, Miller MR, Pal A, Dong J, Madadi NR, Xie J, et al. A biomimetic approach for enhancing the *in vivo* half-life of peptides. Nat Chem Biol. 2015;11: 793–8.

Podust VN, Sim B-C, Kothari D, Henthorn L, Gu C, Wang C, et al. Extension of *in vivo* half-life of biologically active peptides via chemical conjugation to XTEN protein polymer. Protein Eng Des Sel. 2013;26: 743–53.

Prada-Gracia D, Huerta-Yépez S, Moreno-Vargas LM. Application of computational methods for anticancer drug discovery, design, and optimization. Bol Méd Hosp Infant México. 2016;73:411–23.

Pratiwi R, Malik AA, Schaduangrat N, Prachayasittikul V, Wikberg JES, Nantasenamat C, et al. CryoProtect: a web server for classifying antifreeze proteins from nonantifreeze proteins. J Chem. 2017;2017: 9861752.

Reddy KVR, Yedery RD, Aranha C. Antimicrobial peptides: premises and promises. Int J Antimicrob Agents. 2004;24:536–47.

Riedl S, Zweytick D, Lohner K. Membrane-active host defense peptides--challenges and perspectives for the development of novel anticancer drugs. Chem Phys Lipids. 2011;164:766–81.

Rozek T, Wegener KL, Bowie JH, Olver IN, Carver JA, Wallace JC, et al. The antibiotic and anticancer active aurein peptides from the Australian bell frogs Litoria aurea and Litoria raniformis. Eur J Biochem. 2000;267:5330–41.

Schellenberger V, Wang C-W, Geething NC, Spink BJ, Campbell A, To W, et al. A recombinant polypeptide extends the *in vivo* half-life of peptides and proteins in a tunable manner. Nat Biotechnol. 2009;27:1186–90.

Shi LM, Fan Y, Myers TG, O'Connor PM, Paull KD, Friend SH, et al. Mining the NCI anticancer drug discovery databases: genetic function approximation for the QSAR study of anticancer ellipticine analogues. J Chem Inf Comput Sci. 1998;38:189–99.

Shoombuatong W, Hongjaisee S, Barin F, Chaijaruwanich J, Samleerat T. HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. Comput Biol Med. 2012;42:885–9.

Shoombuatong W, Prachayasittikul V, Anuwongcharoen N, Songtawee N, Monnor T, Prachayasittikul S, et al. Navigating the chemical space of dipeptidyl peptidase-4 inhibitors. Drug Des Devel Ther. 2015a;9: 4515–49.

Shoombuatong W, Prachayasittikul V, Prachayasittikul V, Nantasenamat C. Prediction of aromatase inhibitory activity using the efficient linear method (ELM). EXCLI J. 2015b;14:452–64.

Shoombuatong W, Nabu S, Simeon S, Prachayasittikul V, Lapins M, Wikberg JES, et al. Extending proteochemometric modeling for unraveling the sorption behavior of compound–soil interaction. Chemometr Intell Lab Syst. 2016;151:219–27.

Shoombuatong W, Prathipati P, Owasirikul W, Worachartcheewan A, Simeon S, Anuwongcharoen N, et al. Towards the revival of interpretable QSAR models. In: Kunal R (ed): Advances in QSAR modeling (pp 3-55). Basel: Springer International Publ., 2017a. (Challenges and Advances in Computational Chemistry and Physics, Vol. 24).

Shoombuatong W, Prathipati P, Prachayasittikul V, Schaduangrat N, Malik AA, Pratiwi R, et al. Towards predicting the cytochrome P450 modulation: from QSAR to proteochemometric modeling. Curr Drug Metab. 2017b;18:540–55.

Simeon S, Li H, Win TS, Malik AA, Kandhro AH, Piacham T, et al. PepBio: predicting the bioactivity of host defense peptides. RSC Adv. 2017;7:35119–34.

Singh S, Chaudhary K, Dhanda SK, Bhalla S, Usmani SS, Gautam A, et al. SATPdb: a database of structurally annotated therapeutic peptides. Nucleic Acids Res. 2016;44:1119-26.

Spinks CB, Zidan AS, Khan MA, Habib MJ, Faustino PJ. Pharmaceutical characterization of novel tenofovir liposomal formulations for enhanced oral drug delivery: *in vitro* pharmaceutics and Caco-2 permeability investigations. Clin Pharmacol. 2017;9:29–38.

Steiner H, Hultmark D, Engström A, Bennich H, Boman HG. Sequence and specificity of two antibacterial proteins involved in insect immunity. Nature. 1981; 292:246–8.

Thundimadathil J. Cancer treatment using peptides: current therapies and future prospects. J Amino Acids. 2012;2012:967347.

Tong J, Zhao X, Zhong L. QSAR studies of imidazo[4,5-b]pyridine derivatives as anticancer drugs using RASMS method. Med Chem Res. 2014;23:4883–92.

Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 2010; 29:476–88.

Tyagi A, Kapoor P, Kumar R, Chaudhary K, Gautam A, Raghava GPS. In silico models for designing and discovering novel anticancer peptides. Sci Rep. 2013; 3:2984.

Tyagi A, Tuknait A, Anand P, Gupta S, Sharma M, Mathur D, et al. CancerPPD: a database of anticancer peptides and proteins. Nucleic Acids Res. 2015;43: D837-43.

Tørfoss V, Ausbacher D, Cavalcanti-Jacobsen C de A, Hansen T, Brandsdal B-O, Havelkova M, et al. Synthesis of anticancer heptapeptides containing a unique lipophilic $\beta^{2,2}$-amino acid building block. J Pept Sci. 2012;18:170–6.

Usmani SS, Bedi G, Samuel JS, Singh S, Kalra S, Kumar P, et al. THPdb: Database of FDA-approved peptide and protein therapeutics. PLoS ONE. 2017;12: e0181748.

Vedham V, Divi RL, Starks VL, Verma M. Multiple infections and cancer: implications in epidemiology. Technol. Cancer Res Treat. 2014;13:177–94.

Vijayakumar S, Ptv L. ACPP: A web server for prediction and design of anti-cancer peptides. Int J Pept Res Ther. 2015;21:99–106.

Vlieghe P, Lisowski V, Martinez J, Khrestchatisky M. Synthetic therapeutic peptides: science and market. Drug Discov Today. 2010;15:40–56.

Waghu FH, Barai RS, Gurung P, Idicula-Thomas S. CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. Nucleic Acids Res. 2016;44:D1094-7.

Wang G, Li X, Wang Z. APD2: the updated antimicrobial peptide database and its application in peptide design. Nucleic Acids Res. 2009;37:933-7.

Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res. 2016;44:1087-93.

Wang J, Yin T, Xiao X, He D, Xue Z, Jiang X, et al. StraPep: a structure database of bioactive peptides. Database. 2018;2018:bay038.

WHO. Cancer. 2018a. Available from: http://www.who.int/news-room/fact-sheets/detail/cancer.

WHO. Antimicrobial resistance. 2018b. Available from: http://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance.

Win TS, Malik AA, Prachayasittikul V, Wikberg JES, Nantasenamat C, Shoombuatong W. HemoPred: a web server for predicting the hemolytic activity of peptides. Future Med Chem. 2017;9:275–91.

Xu L, Liang G, Wang L, Liao C. A novel hybrid sequence-based model for identifying anticancer peptides. Genes. 2018;9(3).

Zasloff M. Magainins, a class of antimicrobial peptides from Xenopus skin: isolation, characterization of two active forms, and partial cDNA sequence of a precursor. Proc Natl Acad Sci USA. 1987;84:5449–53.

Zhao X, Wu H, Lu H, Li G, Huang Q. LAMP: A database linking antimicrobial peptides. PLoS ONE. 2013; 8:e66557.

Zhou P, Tian F, Wu Y, Li Z, Shang Z. Quantitative Sequence-Activity Model (QSAM): applying QSAR strategy to model and predict bioactivity and function of peptides, proteins and nucleic acids. Curr Comput Aid Drug Des. 2008;4:311–21.