Application of Machine Learning in Peptide Design

Sepčić, Sabino

Undergraduate thesis / Završni rad

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Rijeka / Sveučilište u Rijeci**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:193:310606

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-04-25**

Repository / Repozitorij:

Repository of the University of Rijeka, Faculty of Biotechnology and Drug Development - BIOTECHRI Repository

UNIVERSITY OF RIJEKA

DEPARTMENT OF BIOTECHNOLOGY

Bachelor's degree

"Biotechnology and Drug Development"

Sabino Sepčić

Application of Machine Learning in Peptide Design

Undergraduate thesis

Rijeka, 2020.

UNIVERSITY OF RIJEKA

DEPARTMENT OF BIOTECHNOLOGY

Bachelor's degree

"Biotechnology and Drug Development"

Sabino Sepčić

Application of Machine Learning in Peptide Design

Undergraduate thesis

Rijeka, 2020.

Mentor: Daniela Kalafatović, PhD, Assistant professor

SVEUČILIŠTE U RIJECI

ODJEL ZA BIOTEHNOLOGIJU

Preddiplomski sveučilišni studij

“Biotehnologija i istraživanje lijekova”

Sabino Sepčić

Primjena strojnog učenja u dizajnu peptida

Završni rad

Rijeka, 2020.

Mentor rada: doc. dr. sc. Daniela Kalafatović

Mentor: Daniela Kalafatović, PhD, Assistant professor

Undergraduate thesis was defended on September 21st before the committee:

1. Karlo Wittine, PhD, Associate professor, Committee Head
2. Christian Reynolds, PhD, Assistant professor, Committee Member
3. Daniela Kalafatović, PhD, Assistant professor, mentor

The thesis has 34 pages, 0 pictures, 1 table and 49 references

Sadržaj

Summary

The main purpose of this thesis is to explore the use of computational techniques such as machine learning and genetic algorithms in peptide design. Peptides, short amino acid chains, have a wide variety of biomedical applications. They are used in drug design and as drug delivery systems. Peptide design is an expensive and time-consuming process because it requires high-throughput screenings together with experimental validation, which require a lot of time and resources. Machine learning and genetic algorithms, which will be reviewed in this thesis, can be used to design peptides faster and cheaper by programming the computer to do the work of designing for us. In addition, peptide properties that are important for their function will be reviewed. Basic physico-chemical properties are being used by computer algorithms to predict and/or optimise a peptide with a wanted trait. Some of these traits are antimicrobial or antiviral activity and can be used in applications such as the design of a new type drug based on peptides. Peptide design with computational techniques is a newly emerging field of study and most of the examples considered are of newer date. For this reason, this thesis aims to spread the knowledge about peptides and peptide design.

Key words: Machine learning, genetic algorithm, peptide, peptide design

Sažetak

Glavna svrha ovog završnog rada je istražiti uporabu računalnih tehnika poput strojnog učenja i genetskog algoritma u dizajnu peptida. Peptidi, kratki lanci aminokiselina, imaju raznolike primjene u polju biomedicine. Koriste se u dizajnu lijekova i kao nosači lijekova. Dizajn peptida je skup i dugotrajan proces jer istraživači provode visoko-protočni probir zajedno sa eksperimentalnom validacijom koji zahtijevaju puno vremena i sredstava. Strojno učenje i genetski algoritmi mogu biti korišteni da dizajniramo peptide brže i jeftinije programiranjem računala da radi proces dizajniranja za nas. Svojstva peptida koja su važna za njihove funkcije će biti objašnjena. Ta fizikalno-kemijska svojstva koriste računalni algoritmi da predviđaju i/ili optimiziraju peptid sa željenom funkcijom. Funkcije, poput antimikrobnog i antiviralnog djelovanja, se mogu primijeniti u dizajnu novih vrsta lijekova baziranih na peptidima. Dizajn peptida uz pomoć računalnih tehnika je novo područje i većina razmotrenih studija su novijeg datuma. Zbog toga, ovaj rad ima za namjeru širenje znanja o peptidima i dizajnu peptida.

Ključne riječi: Strojno učenje, genetski algoritam, peptid, dizajn peptida

1. Introduction

Peptides, coming from Greek language peptós – digested, are short amino acid chains with lengths between two and fifty amino acids connected by peptide bonds. Amino acids are organic molecules that contain two functional groups, amine (-NH2) and carbonyl (-COOH) along with a side chain. There are twenty amino acids that are used as building blocks for proteins, called proteinogenic amino acids, and they are encoded in the human genetic code. The fact that there are twenty different amino acids means that the number of combinatorially different peptides increases by 20 times for each amino acid added to the peptide chain. For instance, a peptide with a sequence containing only 6 amino acids can be built in a 20^6 different i.e. 64 million ways. Peptides with different amino acid composition have different physical and chemical properties. Therefore, it is possible to create peptides with a vast array of different properties and applications.

Almost all drugs currently in use are either small molecules or proteins. In between two sizes lie peptides, which are currently being under-explored in drug discovery.[1] Some of the possible applications of peptides are in antimicrobial drugs, drug delivery systems and tissue engineering.[2] Peptides have been explored as potential drugs that can help us treat diseases, improve our immune function and even slow down the aging process.[3] Because of their natural biocompatibility, peptides are non-toxic to living tissues, making them a useful asset in biological applications.[4]

There are natural and synthetic peptides. Natural peptides can be found in all domains of life and can be a part of the natural immunity that helps us fight off microbial infections or have a function as hormones and signal molecules.[5] Another approach for the discovery of new peptides is by synthesizing them. Since the invention of solid phase synthesis by Robert Bruce Merrifield, for which he won the Nobel Prize in Chemistry in 1984, peptides and peptide based materials have been increasingly used for a

variety of applications in biomedicine such as drug delivery and regenerative medicine.[6] In addition, it is possible to conduct high-throughput screening to find out which peptides exert sought function and which peptides do not, for example discovering antimicrobial peptides. Computational techniques developed with the increasing computerization of society can be used as a valuable tool in predicting and optimising peptides. Several decades ago, before the advent of cheminformatics, two ways of isolating peptides existed. The first one is related to isolation of peptides from natural sources, such as plants or animals, while the second one consisted in synthesizing peptides in the laboratory.[7]

The field of computational chemistry has brought together biochemistry, physical chemistry, and computer sciences. Consequently, researches from different scientific disciplines are working together to more efficiently use computer power to generate valuable scientific data.[7] Genetic algorithms and machine learning, both widely explored in computer science for a variety of applications, remain underused in the field of peptide chemistry.

Machine learning (ML) is one of the fastest growing fields of computer science. It has gained increased attention in recent years, with the advent of self-driving cars and the Internet. Machine learning algorithms are computer algorithms that are improving automatically with experience. They are used to predict and perform decisions without being explicitly programmed to do so.[8] The algorithms are first trained using a set of training data to create a mathematical model used for predictions. These algorithms can be used in interpretation and integration of high-throughput data.[9] Machine learning is a more economical and faster way of gathering data than unguided experimental evaluation.[10] For example, algorithms can be used to predict if a peptide from a population exhibits certain traits that we are trying to find (e.g. antimicrobial activity) and which physical and chemical properties of that peptide are responsible for that trait.

On the other hand, genetic algorithms (GA) are used to optimise a property (called fitness function) from a population of potential solutions.[11] The

solution in this study is a peptide with a highly optimised property, for instance, a peptide with a better antiviral activity, calculated by a lower minimal inhibitory concentration (MIC). This algorithm reflects the law of natural selection where the most adaptable individual survives and passes on its genes to the next generation. Each individual is characterised with a set of traits called fitness functions, and each generation has their fitness function graded through fitness score. The individual with the best fitness score, for example a peptide with a better trait than other peptides in the population, gets mutated randomly by crossover and is used to make a new population of peptides. Several iterations are used to make the best optimisation of fitness functions. Fitness functions are used to guide the algorithm towards an optimal solution. Genetic algorithms are useful in peptide design because they can be used to optimise and consequently improve a peptide trait of interest.

Most of the publications regarding the theme of peptide design using machine learning and/or genetic algorithms are of newer date and indicate the recent progress in the field of peptide design and optimization. The application of machine learning and genetic algorithms in peptide design is a new area that is gaining interest because of numerous possible applications of peptides. In addition, peptide design has become more reliable. This is due to the fact that we now have an increased understanding of the rules that govern the assembly of proteins and peptides.[4] This results in faster and more accurate strategies for peptide design that could lead to sequences with higher performance *in vitro* and *in vivo*. However, peptides are still not used as much as they could be. Especially, the use of optimisation and prediction techniques, described in this thesis, remains unexplored. There are thousands of studies performed on the theme of *in vitro* antimicrobial peptides, with only a few of those about the application of machine learning or genetic algorithm techniques on peptide design.

Rational design, including biological, structural and physico-chemical descriptors, is important for the discovery of biologically active peptides.[7]

In this work, examples reported in literature will be listed and elucidated to better understand the principle behind algorithms used in peptide design and the potential applications of peptides.

2. Purpose of the thesis

The main focus of this thesis is the application of machine learning and genetic algorithms in peptide design. The objective is to explore the possibility of improving peptide research, making it cheaper, more efficient, and faster. Can machine learning and genetic algorithms be used to identify, create, and/or optimize antimicrobial peptides with potential to become medicines? If the next pandemic comes, could optimisation and prediction techniques be used to gather faster and more accurate data and possibly discover a cure to a new disease? These are the questions that we will aim to answer by reviewing the existing knowledge in this interdisciplinary field.

3. Peptide properties and their importance in peptide design

In this chapter, fundamental peptide properties such as hydrophobicity and net charge and how they influence peptide function will be assessed. Algorithms often use these physico-chemical properties to help us find optimal peptides. For instance, a genetic algorithm that is programmed to optimise a natural antimicrobial peptide, in order to obtain a more potent peptide, is going to generate more positively charged sequences with a higher hydrophobic moment.

3.1. Net charge

Net charge is the total electric charge of a molecule. Peptides can contain neutral, positively, and negatively charged amino acid residues. By summing these charges, we calculate the net charge. The charge depends on the pH of the solution in which the peptide resides. The higher the isoelectric point of a peptide, the higher must be the pH of the solution for that peptide to have a neutral net charge. At physiological pH of 7.4 there are three positively charged and two negatively charged amino acids that are used to form peptides. Positive amino acids are arginine (Arg), histidine (His) and Lysine (Lys) and negative are aspartic acid (Asp) and glutamic acid (Glu). Net charge is an important parameter when designing antimicrobial peptides because of negatively charged bacterial membranes.

3.2. Hydrophobicity

Water, being a polar molecule, attracts other polar molecules. Amino acid hydrophobicity is a degree of affinity between water and amino acid side

chain residues. The more hydrophobic a molecule is, the less it attracts water molecules. Amino acids with non-polar side chains are more hydrophobic than amino acids with polar side chains.[12] In water solutions, peptides and proteins are taking a conformation in which hydrophilic amino acid residues are residing at the layer closest to the water molecules, and hydrophobic amino acid residues reside inside of formed structures. Hydrophobicity is another parameter used in prediction and optimisation of antimicrobial peptides[7,11,13] because hydrophobic peptide side chains are more effectively attached and inserted into bacterial membranes.[13]

3.3. Secondary structures and the importance of alpha helix

Secondary structures are formed by hydrogen bonds between amino and carbonyl groups of different residues within the same polypeptide molecule. Those bonds can cause the peptide to form two most common distinctive structures: alpha helices and beta sheets. Alpha helix is a right hand-helix conformation in which backbone N-H and C=0 groups form hydrogen bonds between every fourth amino acid. This causes the peptide to curl into a rod-like structure whose inner section consists of a straight coiled main chain with the side chain extending outward.[14] In beta sheet secondary structures, chains line up in parallel one next to the other, and these beta strands connect by several hydrogen bonds from amino acid residues. In this thesis, focus will be put on alpha helical structures rather than beta sheets because there are more examples in the literature that describe them. In a work by Yoshida and co-workers, optimised peptides had their conformation changed from random coil to an alpha helical form which increased their antimicrobial activity.[13] The peptide in an α-helix conformation must have a minimum length of 20 amino acids to span the phospholipid bilayer (~30 Å thick),[15] which disrupts bacterial membranes and is an important mechanism of action. Amphipathic helical structures facilitate the attachment and insertion into bacterial membranes.[16]

4. Application of machine learning and genetic algorithms in peptide design

In this section, the application of ML and GA for the design of antimicrobial, antiviral, anticancer, anti-inflammatory and self-assembling peptides will be reviewed as summarized in table 1. Some studies did not specify the algorithm used in their research, which could hinder the ability of researchers in replicating those studies.

TABLE 1. APPROACHES, ALGORITHMS AND FUNCTIONS IN PEPTIDE DESIGN

	Algorithm	Peptide function
APPROACH MACHINE LEARNING	Random forest	Antiviral[17]
	Support vector machine, random forest, instance-based classifier and k-star	Antiviral[18]
	Counter-propagation artificial neural network	Anticancer[19]
	Random forest, gradient boosting and logistic regression	Self-assembling[2]
	Multiple linear regression	Self-assembling[20]
	Random forest	Anti-inflammatory[21]
GENETIC ALGORITHM	Custom genetic algorithm	Antimicrobial[11]

COMBINED APPROACH (ML+GA)	Unspecified	Antimicrobial[13]
COMBINED APPROACH (ML+EVOLUTIONARY ALGORITHM)	Machine learning: support vector machine, Optimised by evolutionary algorithm: simulated molecular evolution	Anticancer[22]

4.1. Antimicrobial peptides

With the rising threat of antibiotic resistant bacteria, there is now an underlying need to develop effective and cheap medication to combat bacterial infections.[7] The widespread use of antibiotics in 20th Century helped to increase our global life expectancy and quality, but the misuse of antibiotics in humans and animals is accelerating the development of antibiotic resistance. This leads to higher medical costs and increased mortality rates.[23]

There are several methods alternative to the use of conventional antibiotics that can be used to treat microbial infections. Phage therapy, treatment of bacterial infections with the use of bacteriophages, is an example.[24] Another method for treating microbial infections are antimicrobial peptides (AMPs). Antimicrobial peptides act through a diverse set of mechanisms such as membrane destabilization (barrel stave, wormhole and carpet mechanisms) and inhibition of intracellular components.[25] By using a combination of peptides together with antibiotics we may be able to enhance antibiotic activity against multi-drug resistant bacteria.[7] For example, synthetic peptides DJK-6 (VQWRRIRVWVIR-NH2) and DJK-5 (VQWRAIRVRVIR-NH2) in combination with β-lactams, such as meropenem, led to a 16-fold

decrease in antibiotic concentration needed to kill a multidrug resistant carbapenemase-producing *K. pneumoniae.*[26]

Several computer techniques have been developed to make the design process of antimicrobial peptides cheaper and faster in order to obtain a second generation of selective antimicrobials through physico-chemically guided peptide design.[7] The next-generation of antimicrobials should be designed to result in higher selectivity towards pathogens.[7]

Despite AMPs tend to present different lengths and secondary structures in hydrophobic environments, most of them share amphipathic and cationic character at neutral pH.[27] Several studies have implemented genetic algorithms to increase the antimicrobial activity of peptides found in nature (e.g. guavanin).[11,13] The use of computational approaches in manipulation of natural AMP sequences could lead to the discovery of AMPs with distinct mechanisms of action.[11]

An example is presented by Porto et al. in a study published in 2018 that constructed a custom genetic algorithm to optimise a plant based guava peptide.[11] The authors used an initial population of glycine rich guava peptide fragments from peptide Pg-AMP1 as a template for the creation of guavanins, synthetic antimicrobial peptides. Fitness functions were used to optimise peptide properties such as hydrophobic moment and alpha helix propensity. Two kinds of amino acids got preferentially selected in the optimisation process: the positively charged arginine and the hydrophobic residues, leucine and isoleucine. The sequences with the best fitness score got subjected to a Basic Local Alignment Search Tool (BLAST) search, which locates similarities between regions in biological sequences,[28] against a CAMP database[29] to measure their antimicrobial activity. CAMP database is a collection of sequences and structures of antimicrobial peptides which is available online.[29] Random mutations were performed on the template guavanin peptide and subsequent best peptides per population (calculated with fitness score).[11] The best peptide was used in the next process of mutation. Every 50 iterations, peptides had their antimicrobial activity

measured and more peptide hits got acquired at higher iterations. The resulting guavanin peptides had their antimicrobial activity increased with their lower MIC value and toxicity reduced, as well as having different sequences than those in AMP database. Alpha helical structure was found in guavanins and they had more positive net charge as well as being more hydrophobic. Guavanin 2, the most potent guavanin peptide from the optimisation process, has a mechanism of action that involves disruption and hyperpolarization of membranes in bacterial cells (tested against *E. coli* cells).[11]

Another example was presented by Yoshida et al. in 2018, that consisted in the optimization of an amphibian peptide Temporin-Ali[13] isolated from a frog.[30] Frog skin is one of the richest natural sources of AMPs.[31] Yoshida et al. developed a new approach to AMP design, combining machine learning prediction, genetic algorithm optimisation with *in vitro* bacterial assays, which more efficiently provided predictions when generating next generations.[13] Machine learning and genetic algorithms used in this study were not clearly specified. Their method differs from other methods that combine machine learning and evolutionary algorithm by the following: previous algorithms perform virtual screening of potential candidates using an existing database or knowledge (for instance, structure-activity relationship). On the other hand, their custom combination of algorithms optimises a desired feature from a set of targets, in this instance peptides, even with no database or information previously available. This is because evolutionary algorithm constructs its own database as it performs iterations from a starting population, and Yoshida et. al successfully guided those optimisations with predictions by machine learning that allowed them to navigate such a large peptide space. From a starting 13 residue peptide Temporin-Ali which was used as a starting wild type (WT) peptide, a library of 93 sequences was created using Position-Specific-Iterative-BLAST (PSI-BLAST), which iteratively searches for sequences in databases.[28] The peptide library was used as a starting population for the iteration process

of new peptide generation through random amino acid mutations, and after only three iterations a 162-fold increased antimicrobial activity was achieved.[13] The most potent peptides were tested *in vitro* in bacterial assays with *E. coli* and 44 out of 96 had 20 times lower IC50 values than the WT peptide. Net charge and hydrophobic moment were shown to have a high correlation with IC50 values and were increased through optimisation process. More positive net charge suggests that the peptide better binds to negatively charged bacterial membrane, and higher hydrophobic moment suggests that the peptides in the optimised generation formed aliphatic helical structures with hydrophilic and hydrophobic amino acid residues at each side of the helix.[32] All the peptides in the third generation were also predicted to be non-toxic partly because they originated from a natural peptide.[30] Combining *in silico* algorithms with experimental validation enabled them to rapidly optimise antimicrobial features of a peptide, and a normally high cost of experimental validation got mitigated by the predicting power of machine learning.[13]

4.2. Antiviral peptides

Viruses are sub-microscopic pathogens that replicate inside living cells. They are not metabolically active and do not present a cell structure, which is generally considered to be a criteria necessary for life. History has shown that viral infections can be fatal for our species, and nowadays is not an exception. The novel Coronavirus (SARS-CoV-2), that causes the COVID-19 disease, has marked our species with its high death toll and a negative economic impact. Currently, a high amount of resources is being used to try and find a cure for this new infection. Vaccines would be the best solution for long-term immunity against viruses and would constitute an effective shield against coronavirus rapid transmission. Drugs, such as antibodies, are also being developed that could cure the viral infection and decrease the symptoms. A possible approach to treat viral infections are antiviral

peptides (AVPs).[17] Those peptides can be found in nature or artificially synthesised.[33] A list of peptides known to have antiviral activity has been compiled into an online database AVPpred.[34] For instance, lactoferrin is an iron-binding glycoprotein that shows major antiviral activity against viruses as an immunity enhancer.[35] The glycoprotein contains various antimicrobial peptides that are released after hydrolysis by proteases.[36] The first AVP approved by the FDA in 2003 against HIV is called Enfuvirtide (Fuzeon, T-20),[37] and it was a reason for a significant increase of life quality in patients affected with AIDS.[38] Antiviral peptides have several mechanisms in their fight against viruses.[17] They are known to interrupt the signaling process of viruses, block the fusion of viruses to host membranes and inhibit replication in host cells which may involve integrase, DNA polymerase, reverse transcriptase or protease.[39]

Several studies have been made on the theme of optimization and prediction of antiviral peptides. The goal of studies conducted by Chang and Yang in 2013[17] and Qureshi et al. published in 2015[18] was to predict antiviral peptides and peptide antiviral activity using machine learning.

Chang and Yang predicted the features of highly effective antiviral peptides based on the random forest machine learning algorithm.[17] In this study, random forest model that was based on physico-chemical properties of peptides was useful in predicting AVPs. The physico-chemical properties were hydrophobicity, secondary structure, instability, net charge, size, amino acid composition and aliphacity[17] (calculated as aliphatic index, relative volume that is occupied by polypeptide side chains).[40] AVPs had a higher abundance of lysine compared to random peptides, and lysine is also an important amino acid in distinguishing AMPs.[17] In addition to a different amino acid composition, AVPs had a higher tendency to aggregate *in vivo* and had an abundance of alpha helix secondary structures.[17] Aggregation has been suggested to also be an important feature of AMPs, because peptides might have a higher availability while being clumped together.[41] It is not yet clear in which way the alpha helix conformations interact with the

virus. It has been suggested that alpha helical structure could interact with enzymes important for virus replication or have another similar impact on host cell membranes.[17]

In another work by Qureshi et al., multiple machine learning techniques predicted peptide antiviral activity.[18] Their algorithm, AVP-IC50Pred, is a regression-based algorithm that predicts IC50 values of antiviral peptides. Four different machine learning techniques were used to develop their regression based algorithm using experimentally proven datasets: support vector machine, random forest, instance-based classifier and k-star of which random forest and support vector machine were proven to be most effective. Peptide features used in model development are amino acid composition, physico-chemical properties, solvent accessibility, secondary structure, binary profiles and their hybrids for model development.[18] Predicted peptides were not tested in any biological assay so it is possible that not all peptides predicted exert antiviral features.

There has currently not been any research on the topic of computational optimisation or genetic algorithm in AVP design.

4.3. Anticancer peptides

Cancer is a disease that is able to change normal cell function and growth with a possibility to spread to other tissues in the organism. It affects almost all animal species and has been decreasing human life quality through our entire history. Cancer is one of the leading causes of death globally, and massive effort is being conducted to improve cancer diagnosis, treatment, and prevention. The conventional therapy involves chemotherapy and radiotherapy, which often have serious side effects and can make patients feel even weaker than without medication. The mechanism of some of the common anticancer drugs involves the inhibition of mitosis (cell division) or induction of DNA damage. These mechanisms can be fatal to humans if

agents are not specific to cancer cells. The other problem that common chemotherapeutics and receptor-targeted anticancer agents face is the problem of resistance[22]. Several cell resistance mechanisms have been identified, such as DNA damage repair, signalling cascade alteration, drug inactivation or efflux and target protein alteration.[42] Receptor independent mechanism of anticancer peptides (ACPs) may hinder the development of resistance in cancer cells.[43] There are not many known ACPs, and therefore new methods to discover efficient ACPs are needed. High throughput screening of potential peptides together with experimental validation is expensive and time consuming. Because of that, computer techniques are a clever way to bypass those obstacles in the design of ACPs.

Anticancer peptides have properties that have a large impact on their potency, and those are: amphipathicity, positive net charge and moderate overall hydrophobicity.[22] The cationic nature of ACPs is thought to be the reason of their selectivity towards cancer cells, which often have higher negative membrane surface charge compared to non-transformed cells.[19] These properties are used to predict ACPs from a vast data set or to optimise a starting peptide.

Several machine learning techniques have been applied in ACP design, being k-nearest neighbor, support vector machine, random forest and generalized/probabilistic neural network.[19] Small training sizes and lack of experimental validation prevented validity checks of those studies. A recent study by Grisoni et al. published in 2019, developed a counter-propagation artificial neural network (CPANN) machine learning algorithm that was used to design experimentally validated and potent ACPs.[19] The algorithm was trained on two manually assembled databases of peptides with known activities against lung and breast cancer taken from CancerPPD.[44] The descriptors used for the prediction were encoded as peptides' side-chain functionalities, giving each amino acid residue a pharmacophore label (aromatic, lipophilic, hydrogen-bond donor/acceptor and positively/negatively charged).[19] They calibrated four CPANN models on

different peptides, architecture and descriptors. Peptides were given ensemble scores to measure their anticancer abilities, and out of 21 chosen peptides (15 positive, 3 negative and 3 random) 5 were active both against MCF7 and A549 in experimental evaluation.[19] This proved that machine learning can be used in anticancer peptide design, and that the predicted peptides were more likely to have anticancer capabilities than randomly selected or high-throughput screened peptides.

A study by Gabernet et al. published in 2019 developed a machine learning model that predicted ACPs and then utilised an evolutionary molecular design algorithm to optimise the selectivity of predicted peptides towards cancer cells.[22] Support vector machine algorithm (SVM) was used as a predictive machine learning algorithm that was trained on positive (alpha-helical ACPs from CancerPPD database)[44] and negative (non-ACPs from non-transmembrane proteins in PDB database).[45] A library of ACPs was created with their SVM model and peptides had a higher SVM score than randomly generated peptides, meaning that the algorithm enriched the libraries with potentially active anticancer peptides.[22] Four peptides were correctly predicted by testing on cancer cells, and one peptide, AmphiArc2 (KIFKKFKTIIKKVWRIFGRF), was used as a parent peptide in the process of further optimisation. The model used for optimisation was simulated molecular evolution (SME), a class belonging to evolutionary algorithms, which genetic algorithm is also a part of. After three iterations, the activity of peptides towards noncancer cells was decreased by 20 times, while having reduced anticancer potency. This result suggests that the optimisation of peptide selectivity towards noncancer cell types results in the reduction of peptide potency. More work is required to further prove this hypothesis and this study is a basis for future research.[22]

4.4. Self-assembling peptides

Self-assembling peptides are a class of peptides that spontaneously assemble into nanostructures. Governed by physico-chemical characteristics, these peptides can form various structures such as nano fibers, tubes, particles, vesicles, micelles, suspensions and hydrogels.[46] The assembly can be controlled by properties, such as pH, salt concentration and primary structure.[20] These structures have attracted interest in the field of nanotechnology, and when assembled, can be used as vehicles for drug delivery, scaffolds for cell and tissue regeneration and as means of reducing side effects of cytotoxic drugs.[47] The more expensive production of longer, helical peptides also turned interest towards shorter self-assembling peptides. They are designed to resemble natural lipids, with an extended uncharged "tail" with several neutral amino acids and one or two charged amino acids at the C-terminus of the peptide which are resembling lipid headgroups.[4]

Machine learning has been applied in the design of self-assembling peptides. A study presented by Li et al. published in 2019 applied machine learning to the design of self-assembling hydrogels.[2] Hydrogels are networks of hydrophilic polymer chains which can maintain large amounts of water, similar to natural tissues. As such, hydrogels have a great potential for biomedical applications and because of their biocompatibility can be used in cell encapsulation, drug delivery and tissue engineering. Li et al. successfully utilized a combinatorial approach to create a peptide library and screen for peptides with hydrogel assembling function. Machine learning algorithms: random forest, gradient boosting and logistic regression were selected from an extensive list of linear and nonlinear classification algorithms because they have shown best prediction abilities. Algorithms studied the correlation between the ability to form hydrogels and chemical features of peptides. The results showed that the quantum chemistry based descriptors SpMax1_Bh (largest absolute of Burden

modified eigen-value), SpMin1_Bhi (smallest absolute of Burden modified eigen-value) and monomer1 (Fmoc-amino acids) showed the highest correlation with gel forming behaviour and peptide features. To test the ability of peptides in hydrogel assembling, two peptides with predicted hydrogel-forming ability were tested on TM4 cells. Both hydrogels supported the proliferation of TM4 mouse cells, which indicates that predicted peptides are biocompatible and can be used in cell cultures.[2] This is the first study that has applied machine learning techniques in peptide hydrogel design, and further research could increase our understanding of hydrogel design.

Another recent study by Thurston and Ferguson published in 2018 applied machine learning techniques in designing self-assembling π-conjugated oligopeptides.[20] These oligopeptides can be used as a biocompatible supramolecular optoelectronic material. To endorse these self-assembled peptides with photophysical and optoelectronic activity, synthetic oligopeptides can be inserted with polymeric π-conjugated inserts in addition to standard amino acids. Π-inserts are conjugated aromatic cores that promote hydrophobic stacking and delocalisation of electrons within the nanostructure. In this work, a perylenediimide (PDI) and naphthalenediimide (NDI) conjugated core were used. The machine learning algorithm, multiple linear regression, was used to train a quantitative structure-activity relationship model (QSPR) over training data. Dimerisation and trimerization free energies were the key physico-chemical determinants of trained QSPR model, and they were used in prediction of π-conjugated oligopeptides with the best oligomerisation thermodynamics.[20]

Overall, there are not many studies about self-assembling peptides and the application of machine learning or genetic algorithms in their design. To my knowledge, there are currently no applications of optimisation or genetic algorithms in their design. They can have many applications, and more research is needed to have a better understanding of these nanomaterials.

4.5. Anti-inflammatory peptides

Inflammation is a natural biological response to harmful agents. The signs of inflammation: swelling, pain, loss of function, increased heat and redness can leave a larger impact on one's health than an infection. The current way of treating inflammation involves the use of non-specific nonsteroidal or steroidal anti-inflammatory drugs, and those can potentially have side effects such as a higher risk of infectious diseases.

In one of the studies of machine learning in anti-inflammatory peptides (AIPs) design by Manavalan et al. published in 2018, random forest was used in sequence-based prediction.[21] Four different machine learning algorithms were explored (random forest, extremely randomised tree, support vector machine and k-nearest neighbour) including the compositional features: amino acid index, amino acid composition, dipeptide composition, chain-transition-composition and physico-chemical properties for predicting AIPs and non-AIPs. Random forest algorithm with dipeptide composition based model had the best performance among other combinations of algorithms and features. Feature selection protocol on dipeptide composition was used in prediction. The protocol excluded most of Met, Cys and Trp containing dipeptides which indicates that the arrangement of local dipeptides is an important feature in classification of AIPs and non-AIPs. The proposed predictor has been shown as promising and it is available at a web server. Compared to experimental approaches and other available tools, it provides a cost-effective and reliable approach in AIP prediction.[21]

To my knowledge, there have currently been no applications of genetic algorithms or optimisation techniques in AIP design.

4.6. Other applications of peptides

Peptides can be used as biosensors, where the affinity of AMPs is used to detect microbials. For example, a nanostructured biosensor with AMP clavanin A was used to detect Gram-negative bacteria.[48] Peptides can be predicted to have high selectivity towards certain bacteria, which could be useful in areas such as hospitals, where peptides could possibly detect Staphylococcus aureus in the environment before the infection occurs. Additionally, self-assembling peptides have been shown to have a useful application in tissue regeneration, but peptide aggregates ability of drug delivery has not been fully researched, especially regarding the use of computer techniques in their design. Another application of peptides are immunomodulatory peptides, where an engineered immunomodulatory peptide clavanin-MO modulated innate immunity by means of producing immune mediators and recruitment of leukocytes to the infection site.[49]

5. Conclusion

The versatility of peptides makes them attractive for potential antibacterial, antiviral, anticancer, anti-inflammatory, and optoelectrical/tissue engineering candidates, and other applications. Computer processing power is greatly increasing year after year, and different types of algorithms are becoming more efficient at performing complex tasks. Computational chemistry and its use in chemical design (example being peptide design) is being improved on daily basis. I believe that we will see more peptides in various medical applications. Peptides are still far from being widely used in medicine, but the new rise of antibiotic resistant bacteria and virus infections prove that a new type of medicine that is not based only on classic antibiotics is needed. Certain peptides, such as the antiviral peptide Enfuvirtide (AIDS treatment)[37], have already found their uses outside clinical trials. Future approaches in drug design will probably see greater implementation of computational techniques, which will more easily connect researches from various parts of the globe into a single network to make researching more efficient. All the possible applications of peptides would hardly fit in a short thesis, and more research should be conducted to better our understanding of these small amino acid chains.

6. References

1. Bhardwaj, G. *et al.* Accurate de novo design of hyperstable constrained peptides. *Nature* **538**, 329–335 (2016).

2. Li, F. *et al.* Design of self-assembly dipeptide hydrogels and machine learning via their chemical features. *Proc. Natl. Acad. Sci. U. S. A.* **166**, 11259–11264 (2019).

3. Proksch, E. *et al.* Oral intake of specific bioactive collagen peptides reduces skin wrinkles and increases dermal matrix synthesis. *Skin Pharmacol. Physiol.* **27**, 113–119 (2014).

4. Boyle, A. L. Applications of de novo designed peptides. in *Peptide Applications in Biomedicine, Biotechnology and Bioengineering* 51–86 (Elsevier Inc., 2018). doi:10.1016/B978-0-08-100736-5.00003-X.

5. Daffre, S. *et al.* Bioactive natural peptides. in *Studies in Natural Products Chemistry* vol. 35 597–691 (Elsevier, 2008).

6. Mitchell, A. R. Bruce Merrifield and solid-phase peptide synthesis: A historical assessment. *Biopolymers* **90**, 175–184 (2008).

7. Der Torossian Torres, M. & De La Fuente-Nunez, C. Reprogramming biological peptides to combat infectious diseases. *Chem. Commun.* **55**, 15020–15032 (2019).

8. Koza, J. R., Bennett, F. H., Andre, D. & Keane, M. A. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. in *Artificial Intelligence in Design '96* 151–170 (Springer Netherlands, 1996). doi:10.1007/978-94-009-0279-4_9.

9. Lima, A. N. *et al.* Use of machine learning approaches for novel drug discovery. *Expert Opinion on Drug Discovery* vol. 11 225–239 (2016).

10. Tallorin, L. *et al.* Discovering de novo peptide substrates for enzymes using machine learning. *Nat. Commun.* **9**, 1–10 (2018).

11. Porto, W. F. *et al.* In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design. *Nat. Commun.* **9**, 1–12 (2018).

12. Hydrophobicity - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/hydrophobicity.

13. Yoshida, M. *et al.* Using Evolutionary Algorithms and Machine Learning to Explore Sequence Space for the Discovery of Antimicrobial Peptides. *Chem* **4**, 533–543 (2018).

14. Alpha Helix - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/alpha-helix.

15. Casciaro, B., Cappiello, F., Cacciafesta, M. & Mangoni, M. L. Promising approaches to optimize the biological properties of the antimicrobial peptide esculentin-1a(1-21)NH2: Amino acids substitution and conjugation to nanoparticles. *Frontiers in Chemistry* vol. 5 26 (2017).

16. Brogden, K. A. Antimicrobial peptides: Pore formers or metabolic inhibitors in bacteria? *Nature Reviews Microbiology* vol. 3 238–250 (2005).

17. Chang, K. Y. & Yang, J. R. Analysis and Prediction of Highly Effective Antiviral Peptides Based on Random Forests. *PLoS One* **8**, (2013).

18. Qureshi, A., Tandon, H. & Kumar, M. AVP-IC $_{50}$ Pred: Multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (IC $_{50}$). *Biopolymers* **104**, 753–763 (2015).

19. Grisoni, F. *et al.* De novo design of anticancer peptides by ensemble

artificial neural networks. *J. Mol. Model.* **25**, 1–10 (2019).

20. Thurston, B. A. & Ferguson, A. L. Machine learning and molecular design of self-assembling -conjugated oligopeptides. *Mol. Simul.* **44**, 930–945 (2018).

21. Manavalan, B., Shin, T. H., Kim, M. O. & Lee, G. AIPpred: Sequence-based prediction of anti-inflammatory peptides using random forest. *Front. Pharmacol.* **9**, (2018).

22. Gabernet, G. *et al.* In silico design and optimization of selective membranolytic anticancer peptides. *Sci. Rep.* **9**, 11282 (2019).

23. Antibiotic resistance. https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance.

24. Bacteriophage: A solution to our antibiotics problem? How we can us a virus to fight bacterial infection. http://sitn.hms.harvard.edu/flash/2018/bacteriophage-solution-antibiotics-problem/.

25. Corrêa, J. A. F., Evangelista, A. G., Nazareth, T. de M. & Luciano, F. B. Fundamentals on the molecular mechanism of action of antimicrobial peptides. *Materialia* vol. 8 100494 (2019).

26. Ribeiro, S. M. *et al.* Antibiofilm peptides increase the susceptibility of carbapenemase-producing Klebsiella pneumoniae clinical isolates to β-lactam antibiotics. *Antimicrob. Agents Chemother.* **59**, 3906–3912 (2015).

27. Powers, J. P. S. & Hancock, R. E. W. The relationship between peptide structure and antibacterial activity. *Peptides* **24**, 1681–1691 (2003).

28. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* vol. 25 3389–3402 (1997).

29. Waghu, F. H. *et al.* CAMP: Collection of sequences and structures of antimicrobial peptides. *Nucleic Acids Res.* **42**, D1154 (2014).

30. Wang, M. *et al.* Five novel antimicrobial peptides from skin secretions of the frog, Amolops loloensis. *Comp. Biochem. Physiol. - B Biochem. Mol. Biol.* **155**, 72–76 (2010).

31. Conlon, J. M. Structural diversity and species distribution of host-defense peptides in frog skin secretions. *Cellular and Molecular Life Sciences* vol. 68 2303–2315 (2011).

32. Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. The helical hydrophobic moment: A measure of the amphiphilicity of a helix. *Nature* **299**, 371–374 (1982).

33. Vilas Boas, L. C. P., Campos, M. L., Berlanda, R. L. A., de Carvalho Neves, N. & Franco, O. L. Antiviral peptides as promising therapeutic drugs. *Cellular and Molecular Life Sciences* vol. 76 3525–3542 (2019).

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

35. Elnagdy, S. & AlKhazindar, M. The Potential of Antimicrobial Peptides as an Antiviral Therapy against COVID-19. *ACS Pharmacol. Transl. Sci.* (2020) doi:10.1021/acsptsci.0c00059.

36. Sinha, M., Kaushik, S., Kaur, P., Sharma, S. & Singh, T. P. Antimicrobial lactoferrin peptides: The hidden players in the protective function of a multifunctional protein. *International Journal of Peptides* vol. 2013 12 (2013).

37. Drugs@FDA: FDA-Approved Drugs. https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm?event=overview.process&ApplNo=021481.

38. Ashkenazi, A., Wexler-Cohen, Y. & Shai, Y. Multifaceted action of

Fuzeon as virus-cell membrane fusion inhibitor. *Biochimica et Biophysica Acta - Biomembranes* vol. 1808 2352–2358 (2011).

39. Castel, G., Chtéoui, M., Heyd, B. & Tordo, N. Phage Display of Combinatorial Peptide Libraries: Application to Antiviral Research. *Molecules* **16**, 3499–3518 (2011).

40. ExPASy - ProtParam documentation. https://web.expasy.org/protparam/protparam-doc.html.

41. Torrent, M., Andreu, D., Nogués, V. M. & Boix, E. Connecting Peptide Physicochemical and Antimicrobial Properties by a Rational Prediction Model. *PLoS One* **6**, e16968 (2011).

42. Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: An evolving paradigm. *Nature Reviews Cancer* vol. 13 714–726 (2013).

43. Papo, N. & Shai, Y. Visions & Reflections Host defense peptides as new weapons in cancer treatment. doi:10.1007/s00018-005-4560-2.

44. Tyagi, A. *et al.* CancerPPD: A database of anticancer peptides and proteins. *Nucleic Acids Res.* **43**, D837–D843 (2015).

45. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* vol. 28 235–242 (2000).

46. Lee, S. *et al.* Self-assembling peptides and their application in the treatment of diseases. *International Journal of Molecular Sciences* vol. 20 (2019).

47. Lee, S. *et al.* Molecular Sciences Self-Assembling Peptides and Their Application in the Treatment of Diseases. doi:10.3390/ijms20235850.

48. de Miranda, J. L. *et al.* A simple nanostructured biosensor based on clavanin A antimicrobial peptide for gram-negative bacteria detection. *Biochem. Eng. J.* **124**, 108–114 (2017).

49. Silva, O. N. *et al.* An anti-infective synthetic peptide with dual antimicrobial and immunomodulatory activities. *Sci. Rep.* **6**, 1–11 (2016).