J. Integrative Natural Sci.
Vol.17, No.1 (2024) pp.21-30
https://doi.org/10.13160/ricns.2024.17.1.21

Machine learning Anti-inflammatory Peptides Role in Recent Drug Discovery

Subathra Selvam[1†]

[1†] *Computational Biology Laboratory, Department of Genetic Engineering,*
SRM Institute of Science and Technology, SRM Nagar,
Kattankulathur-603203, Tamil Nadu, India

Abstract

Several anti-inflammatory small molecules have been found in the process of the inflammatory response, and these small molecules have been used to treat some inflammatory and autoimmune diseases. Numerous tools for predicting anti-inflammatory peptides (AIPs) have emerged in recent years. However, conducting experimental validations in the lab is both resource-intensive and time-consuming. Current therapies for inflammatory and autoimmune disorders often involve nonspecific anti-inflammatory drugs and immunosuppressants, often with potential side effects. AIPs have been used in treating inflammatory illnesses like Alzheimer's disease and can limit the expression of inflammatory promoters. Recent advances in adverse incident predictions (AIPs) have been made, but it is crucial to acknowledge limitations and imperfections in existing methodologies.

Keywords: Anti-inflammatory, Peptides, Machine learning.

(**Recceived** January 19, 2024; **Revised** February 20, 2024; **Accepted** March 11, 2024)

1. Introduction

Peptides are naturally occurring biological agents with lengths ranging from 2 to 50 amino acids that serve vital activities as anti-infectives, growth factors, hormones, biological messengers, and neurotransmitters. The discovery of peptide hormones including insulin, vasopressin, oxytocin, and gonadotropin-releasing hormones inspired significant discoveries in biology, chemistry, pharmacology, and other cutting-edge drug development technologies. Antiangiogenic peptides (AAPs), antibacterial peptides (ABPs), anti-inflammatory peptides (AIPs), anticancer peptides (ACPs), antifungal peptides (AFPs), and other peptides have been found to have a variety of biological functions. The chemical and biological diversity of peptides makes them appealing for the creation of novel therapies[1]. Almost 7000 peptides with diverse properties have been identified in the past, including cell-penetrating peptides

[†] Corresponding author: supsysubi98@gmail.com

(CPP), anticancer peptides (ACP), antiviral peptides (AVP), and anti-inflammatory peptides (AIP)[2].

Inflammation is a multiple process, which is mediated by activated inflammatory or immune cells and it is caused by the release of chemicals from tissues and migrating cells. Its responses occur under normal conditions when tissues are damaged by bacteria, toxins, trauma, heat, or any other reason. These responses cause chronic autoimmune and inflammation disorders, including neurodegenerative disease, asthma, psoriasis, cancer, rheumatoid arthritis, diabetes, and multiple sclerosis[3]. Hence, the current therapy for inflammatory and autoimmune disorders involves the use of nonspecific anti-inflammatory drugs and other immunosuppressants, which are often accompanied by potential side effects[4]-[5]. AIPs have lately been employed in the treatment of numerous inflammatory illnesses such as Alzheimer's disease. Similarly, both natural and synthesized peptides have the capacity to limit the expression of inflammatory promoters[2]. Though wet-lab production of peptide-based drugs is very costly and time-consuming[6]. Hence it is essential to develop an in-silico based classification model for prediction. Several anti-inflammatory predictions have been developed in recent years and this review paper speaks about the recent advances in AIP and its future perspectives.

In this present review, we have provided a comprehensive overview of recent advancements in AIPs (Anti-Inflammatory Peptide). Despite the numerous studies that have been conducted in this field, it remains crucial to acknowledge the limitations and imperfections inherent in the existing methodologies. It is imperative to strive towards devising an improved approach for the anticipation of AIPs, addressing the drawbacks observed in the current methods.

2. Overview of ML model prediction:

Data processing and cleaning is the first critical step involved in the construction of successful ML-based models. Because there are no automated data collection procedures, researchers have to perform extensive literature searches, and then scrutinize, curate, and assemble the data into a data set or database. For these reasons, data processing and cleaning is a time-consuming and laborious process[7]. Figure 1 illustrates the complete workflow of AIPs model generation. Both positive and negative compounds should be presented. And the redundant compounds should be removed in order to enhance the model prediction. Different types of splits are known, mostly it is advisable to consider 70% to 80% of original data as training for the prediction model development, and 20% to 30% as independent data, which can be used to assess the transferability of the developed model. Feature encoding methods carry most significant role in model developing and some of the feature encoding methods were described. AAI is a numerical representation of the properties of amino acids in a protein sequence. ASDPC is a method to represent protein sequences by considering their amphiphilic properties. AAC calculate percentage of each amino acid present in a protein sequence. Am-PseAAC is a feature that combines the information from amino acid composition and physicochemical properties. ATC describes the correlation between the occurrences of tripeptides at different positions in a protein sequence. BPF is a representation of protein sequences based on their physicochemical properties. CKSAAP describes the occurrence frequencies of amino acid pairs with a specific gap. CTD represents proteins using three descriptors: amino acid composition, transition, and distribution. DPC describes the frequency of occurrence of all possible pairs of amino acids. DDE characterizes the distribution of distances between amino acids in a protein sequence. EAAC considers the electrostatic attraction between amino acids. EGAAC combines evolutionary information and amino acid composition. GAAC considers amino acids grouped based on their physicochemical properties. GGDPC incorporates gene ontology information with dipeptide composition. GDPC combines gene ontology information and dipeptide composition. GTPC integrates gene ontology information with tripeptide composition. PseAAC is a method that extends amino acid composition to include additional information.

PCP describes proteins based on their physicochemical properties. QSO represents the sequence order information of proteins. RAAC is a simplified version of amino acid composition. TPC describes the frequency of occurrence of all possible tripeptides in a protein sequence. Various feature encoding methods are available, some are illustrated in **Table. 1.** Further model training and evaluation is carried out.

One of the main goals of any ML algorithm is to rigorously train the model for the accurate classification of any unseen data. During model training, feature descriptors generated from the training data set, along with the class (response variable: positive or negative), are inputted to a ML classifier, where it learns the relationship between feature descriptors (x) and response variable (y), and makes subsequent predictions for any newly provided data sets. The main objective of a good ML model is to generalize from training data to independent data. The classifiers commonly used in computational biology and bioinformatics, include AdaBoost (AB), ANN, deep learning (DL), extreme learning machine (ELM), extremely randomized tree (ERT), gradient boosting (GB), k-nearest neighbor (KNN), RF, SVM, and extreme gradient boosting (XGB) [8]-12]. Each classifier has its advantages and disadvantages with respect to data quantity, training speed, and feature encodings.

Instead of randomly selecting a specific ML classifier, it is essential to explore one or more sets of classifiers on the same training data set, and then choose the most appropriate one. Specifically, each classifier has a set of parameters that separates signal from noise; this set of parameters must be optimized during training. Further, cross-validation technique (CV) is used to limit the overfitting of the model. Three types of CV techniques employed in ML protocols include n-fold CV, Stratified K-fold CV and leave-one-out CV (LOOCV)[13]-[14]. Four sets of metrics, commonly used to assess prediction performance of various methods, include sensitivity (Sn), specificity (Sp), accuracy (ACC), and MCC[15].

3. Recent advances in AIP prediction

In 2017, Gupta and colleagues presented a novel approach for developing an anti-inflammatory predictor utilizing a support vector machine ("SVM") classifier. Among the various hybrid models they explored, the TPC_HYB model, which combined TPC-based features with motif features, demonstrated remarkable performance and emerged as the top-performing one. The TPC_HYB model achieved an impressive accuracy of 78.1%, indicating its proficiency in accurately classifying anti-inflammatory properties of compounds. Additionally, it was found to have a Matthews's correlation coefficient (MCC) of 0.58, signifying a good balance between sensitivity and specificity. Moreover, the area under the curve (AUC) value of 0.86 further validated its strong discriminatory power[5].

AntiFlamPred was developed by Alotaibi et al. The proposed model showcased exceptional performance by incorporating deep learning techniques and underwent rigorous testing and validation. Various evaluation methods, such as cross-validation, self-consistency, jackknife, and independent set testing, were meticulously applied to ensure the model's reliability and accuracy. The results were truly impressive, with the proposed model achieving an outstanding AUC (Area under the Curve) value of 0.919 and a Mathew's correlation coefficient (MCC) of 0.735. These high scores demonstrate the effectiveness and stability of the model in predicting AIPs accurately. Notably, the researchers conducted thorough testing on both the benchmark dataset and an independent test set. During the 10-fold cross-validation test on the benchmark dataset, the model achieved an AUC of 0.919 and an MCC of 0.735, reaffirming its robustness and superior performance. Furthermore, when applied to the independent test set, the model still achieved remarkable results with an AUC of 0.907 and an MCC of 0.681. This indicates that the proposed classification model is not only highly effective but also cost-effective, making it a powerful tool for AIP prediction. Its ability to maintain excellent performance on an unseen dataset demonstrates its generalizability and real-world applicability, crucial qualities for any practical classification model[6].

AIPpred, an innovative AIP predictor introduced by Manavalan et al., has garnered attention for its remarkable prediction performance using a random forest (RF) classifier and sequence encoding features. The method achieved an impressive AUC (Area under the Curve) of 0.814 and an MCC (Matthews Correlation Coefficient) of 0.479. What sets AIPpred apart is the extensive and diverse benchmarking dataset it utilized during its development, allowing for robust and reliable predictions. In their pursuit of the most effective prediction model, the researchers explored four different machine learning-based algorithms, among which SVM (Support Vector Machine) was one. However, it was the RF-based approach that demonstrated superior performance, making AIPpred the pioneering application of an RF-based method in the field of AIP prediction. A key factor contributing to

AIPpred's success is the use of optimal DPC (Discrete Positional Code) features, identified through a meticulous feature selection protocol. By selecting the most relevant features, AIPpred ensures that the predictive model focuses on the most critical aspects of the input sequences, enhancing its accuracy. Another noteworthy aspect of AIPpred is its unique parameter-optimization procedure. The researchers employed ten independent 5-fold cross-validations to fine-tune the machine learning parameters. This thorough and rigorous optimization process helps to avoid overfitting and ensures the model's generalizability, making AIPpred a robust predictor even on unseen data. The introduction of AIPpred into the realm of AIP prediction has opened new avenues for leveraging RF-based methods, demonstrating the potential of this approach in addressing complex biological problems. With its strong predictive performance and reliable results, AIPpred holds promise in advancing our understanding of AIPs and their implications in various biological processes[16].

PreAIP was developed by Khatun et al. through a random forest classifier incorporating manifold features like primary sequence and structural information. The performance evaluation showed an AUC value of 0.840 and MCC of 0.512 on the test dataset[17].

Wei et al used hybrid sequence-based features which were further optimized to select widely discriminative features and trained 8 random-forest models to predict 8 functionally different peptides yielding an AUC value of 0.75[18].

Deng et al explores various algorithms and encoding schemes for AIP identification, finding that six out of eight existing methods only used RF for their models. The RF algorithm in conjunction with DDE and CKSAAP achieved good performance, while the ET algorithm was not employed in any existing methods. The effectiveness of feature fusion was demonstrated by evaluating the performance of ET-based and RF-based models. The final AIPStack model achieved an average AUC of 0.808 on the training set, representing an improvement of AUC of 1.4% −26.9% compared to the three constituent models. The study also found that tree-based models generally performed better, so the authors chose the two best tree-based models as the base-classifiers. The AIPStack performed well on all three independent sets, demonstrating the stability and reliability of the method. The SHAP algorithm was applied for model interpretation, revealing essential features for AIP optimization, such as LS.gap0, LE, SL, and LL. The composition analysis revealed significant differences between AIPs and non-AIPs in dipeptide composition[19].

In 2021, Zhang and colleagues introduced a novel feature representation strategy for their anti-inflammatory peptide predictor, which resulted in significant performance improvements. Their method achieved an accuracy (ACC) of 0.762 and a Matthews’s correlation coefficient (MCC) of 0.495, representing a notable two-percentage-point increase over the AIPpred model. Moreover, their approach demonstrated a similar accuracy level to the best model of PreAIP[20].

According to Lin et al.'s research in 2021, their predictive model, PREDAIP, demonstrated superior performance compared to existing tools in the field. PREDAIP achieved an impressive accuracy (ACC) of 85.6%, a Matthews's correlation coefficient (MCC) of 0.739, and an area under the curve (AUC) of 0.938. These values were notably higher, by 2.9% to 29.6%, 9.6% to 47.3%, and 2.6% to 28.5%, respectively, when compared to three other established tools, namely AntiInflam, AIPpred, and PreAIP. The significant performance improvements suggest that their predictor surpasses the capabilities of existing methods .

Guo et al and Yan et al, states the utilization of an ensemble learning strategy might improve the performance of AIP identification. Recently, two ensemble learning methods, namely PreTP-EL and PreTP-Stack, were reported, but their performance in AIP prediction might be still limited. Accordingly, it is essential to develop a new prediction model with higher accuracy, which could not only help improve our understanding of the association between peptide sequence and anti-inflammatory activity but also provide a reference for the rational design of AIPs based on the important features given by the model explanation[21].

4. Conclusion

The anticipation of biopeptides plays a pivotal role in uncovering and creating effective peptide-oriented medications. Nevertheless, the experimental techniques employed in identifying and producing biopeptides come with a high cost, demand significant labor and time, and frequently involve extended periods of uncertain trial and error. To counteract these challenges, data-centric computational approaches, notably Machine Learning (ML), have been devised to swiftly and efficiently forecast therapeutic peptides. The capacity to prognosticate the therapeutic attributes of peptides based solely on their sequence information has inspired computational biologists to formulate various ML-driven prediction utilities.

In general, Machine Learning (ML) driven prediction utilities offer a robust structure for tackling a range of challenges within the domain of peptide-oriented investigations. Despite the multitude of ML tools that have been created, there remains an opportunity for further enhancement and augmentation of this collection. As the volume of accessible data continues to expand and computer capabilities escalate, ML algorithms are poised to bring about a transformative impact on the realm of peptide therapeutics in the coming times[22]. Designing vaccines involves the intricate task of identifying peptides capable of inducing anti-inflammatory cytokines, a challenge in the field. Computational prediction of Anti-Inflammatory Peptide (AIP) candidates proves crucial in streamlining labor-intensive experimental processes. This task stands out as more intricate compared to other peptide-based prediction methods like those for anticancer, antiviral, and cell-penetrating peptides. Typically, methods developed with robust datasets verified through experiments find broad applications in modern biology.

Systematic effort delves into comprehending the characteristics of anti-inflammatory-inducing peptides and aims to construct a reliable prediction model. Several studies advocate for an AIP model accompanied by a webserver, enhancing user accessibility and contributing to reproducibility. Such a webserver allows users to predict AIPs, fostering ease of use and broadening its applicability. The development of a user-friendly tool aligns with the trend of leveraging computational approaches to expedite and augment biological research. Given the multitude of studies conducted in previous years, there arises a necessity to anticipate a sophisticated model in the realm of predictive modeling for this field in terms of accuracy and efficiency.

Reference

[1] Basith S, Manavalan B, Hwan Shin T, Lee G., “Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening.”, Med Res Rev Vol. 40, pp. 1276-314, 2020. https://doi.org/10.1002/med.21658.

[2] Attique M, Farooq MS, Khelifi A, Abid A., “Prediction of Therapeutic Peptides Using Machine Learning: Computational Models, Datasets, and Feature Encodings.”, IEEE Access Vol. 8, pp. 148570-94, 2020.

[3] Khatun MS, Hasan MM, Kurata H., “PreAIP: Computational Prediction of Anti-inflammatory Peptides by Integrating Multiple Complementary Features.”, Front Genet Vol. 10, p. 129, 2019. https://doi.org/10.3389/fgene.2019.00129.

[4] Zhao D, Teng Z, Li Y, Chen D., “iAIPs: Identifying Anti-Inflammatory Peptides Using Random Forest.”, Front Genet Vol. 12, pp. 1-9, 2021. https://doi.org/10.3389/fgene.2021.773202.

[5] Gupta S, Sharma AK, Shastri V, Madhu MK, Sharma VK., “Prediction of anti-inflammatory proteins/peptides: An insilico approach.”, J Transl Med, Vol. 15, pp. 1-11, 2017. https://doi.org/10.1186/s12967-016-1103-6.

[6] Fahad Alotaibi Muhammad Attique YDK., “AntiFlamPred: An Anti-Inflammatory Peptide Predictor for Drug Selection Strategies.”, Comput Mater \& Contin Vol. 69, pp. 1039-55, 2021. https://doi.org/10.32604/cmc.2021.017297.

[7] Shaker B, Yu M-S, Song JS, Ahn S, Ryu JY, Oh K-S, et al., “LightBBB: computational prediction model of blood-brain-barrier penetration based on LightGBM.”, Bioinformatics, Vol. 37, pp. 1135-9, 2021. https://doi.org/10.1093/bioinformatics/btaa918.

[8] Breiman L, Friedman JH, Olshen RA, Stone CJ., “Classification and Regression Trees”, 1984.

[9] Bansal M, Goyal A, Choudhary A., “A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning.”, Decis Anal J, Vol. 3, p. 100071, 2021. https://doi.org/https://doi.org/10.1016/j.dajour.2022.100071.

[10] Chen T, Guestrin C., “XGBoost: A scalable tree boosting system.”, Proc

ACM SIGKDD Int Conf Knowl Discov Data Min 2016;13-17-Augu:785-94. https://doi.org/10.1145/2939672.2939785.

[11] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al., "LightGBM: A highly efficient gradient boosting decision tree.", Adv Neural Inf Process Syst 2017;2017-Decem:3147-55.

[12] Kabiraj S, Raihan M, Alvi N, Afrin M, Akter L, Sohagi SA, et al. Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm. 2020 11th Int. Conf. Comput. Commun. Netw. Technol., 2020, p. 1-4. https://doi.org/10.1109/ICCCNT49239.2020.9225451.

[13] Prusty S, Patnaik S, Dash SK., "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer.", Front Nanotechnol, Vol. 4, pp. 1-12, 2022. https://doi.org/10.3389/fnano.2022.972421.

[14] Refaeilzadeh P, Tang L, Liu H., "Cross-Validation. In: LIU L, ÖZSU MT, editors.", Encycl. Database Syst., Boston, MA: Springer US; pp. 532-8, 2009. https://doi.org/10.1007/978-0-387-39940-9_565.

[15] Wahi D, Jamal S, Goyal S, Singh A, Jain R, Rana P, et al., "Cheminformatics models based on machine learning approaches for design of USP1/UAF1 abrogators as anticancer agents.", Syst Synth Biol, Vol. 9, pp. 33-43, 2015. https://doi.org/10.1007/s11693-015-9162-1.

[16] Manavalan B, Shin TH, Kim MO, Lee G., "AIPpred: Sequence-Based Prediction of Anti-inflammatory Peptides Using Random Forest.", Front Pharmacol, Vol. 9, p. 276, 2018. https://doi.org/10.3389/fphar.2018.00276.

[17] Khatun MS, Hasan MM, Kurata H., "PreAIP: Computational prediction of anti-inflammatory peptides by integrating multiple complementary features.", Front Genet, Vol. 10, 2019. https://doi.org/10.3389/fgene.2019.00129.

[18] Wei L, Zhou C, Chen H, Song J, Su R., "ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides.", Bioinformatics, Vol. 34, pp. 4007-16, 2018. https://doi.org/10.1093/bioinformatics/bty451.

[19] Deng H, Lou C, Wu Z, Li W, Liu G, Tang Y., "Prediction of anti-inflammatory peptides by a sequence-based stacking ensemble model named AIPStack.", IScience, Vol. 25, p. 104967, 2022. https://doi.org/https://doi.org/10.1016/j.isci.2022.104967.

[20] Zhang J, Zhang Z, Pu L, Tang J, Guo F., "AIEpred: An Ensemble Predictive Model of Classifier Chain to Identify Anti-Inflammatory Peptides.", IEEE/ACM Trans Comput Biol Bioinforma, Vol. 18, pp. 1831-40, 2021. https://doi.org/10.1109/TCBB.2020.2968419.

[21] Guo Y, Yan K, LV H, Liu B., "PreTP-EL: prediction of therapeutic peptides based on ensemble learning." Brief Bioinform, Vol. 22, bbab358, 2021. https://doi.org/10.1093/bib/bbab358.

[22] Shaker B, Yu MS, Song JS, Ahn S, Ryu JY, Oh KS, et al., "LightBBB: Computational prediction model of blood-brain-barrier penetration based on LightGBM.", Bioinformatics, Vol. 37, pp. 1135-9, 2021.

https://doi.org/10.1093/bioinformatics/btaa918.

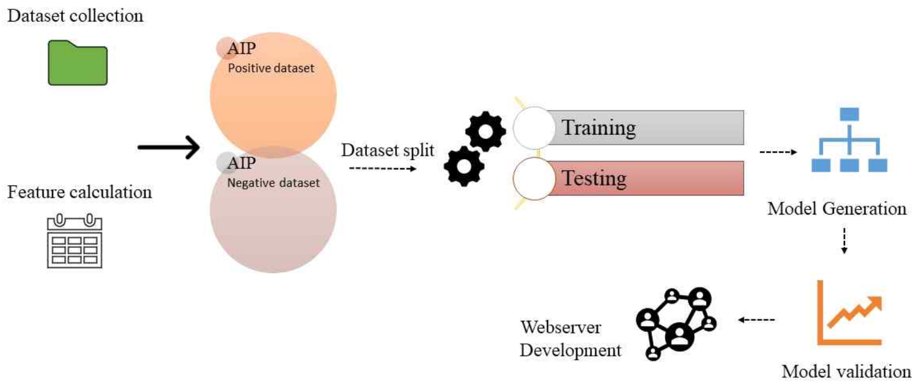

Fig. 1. Workflow of AIP’s model generation

Table. 1. Feature encoding methods

Feature encoding	
AAI	Amino acid index
ASDPC	Adaptive skip dipeptide composition
AAC	Amino acid composition
Am-PseAAC	Amphiphilic pseudo amino acid composition
ATC	Atomic composition
BPF	Binary profile
CKSAAP	Composition of K-spaced amino acid pairs
CTD	Composition-transition-distribution
DPC	Dipeptide composition
DDE	Dipeptide deviation from expected mean
EAAC	Enhanced amino acid composition
EGAAC	Enhanced grouped amino acid composition
GAAC	Grouped amino acid composition
GGDPC	G-gap dipeptide composition
GDPC	Grouped dipeptide composition
GTPC	Grouped tripeptide composition
PseAAC	Pseudo amino acid composition
PCP	Physicochemical properties
QSO	Quasi sequence order
RAAC	Reduced amino acid composition
TPC	Tripeptide composition

Table. 2. List of currently available AIPs methods

Methods	Author	Feature encoding	Model	Accuracy	Web Server
Anti-inflammatory	Gupta el al	TPC_HYB	SVM	78.1%	Yes
AntiFlamPred	Alotaibi et al	FIV	DNN	84%	No
AIPpred	Manavalan et al	DPC	RF	74%	Yes
PreAIP	Khatun et al	KSAAP & Combined	RF	83%	Yes
PEPred-Suite	Wei et al	89 class features	RF	72%	Yes
AIEpred	Zhang et al	AAC, PSSM, PP	RF	76%	No
PREDAIP	Lin et al, 2021	emRMR-SFS	ERT	85.6%	No
AIPStack	Deng et al	CKSAAP, DDE, and the hybrid features	ERT & RF	75%	No