Machine learning approaches for addressing classification problems of four types of immune-peptides

著者	Khatun　MST Shamima
その他のタイトル	４種類の免疫ペプチド分類問題を解決する機械学習アプローチ
学位授与年度	令和2年度
学位授与番号	17104甲情工第358号
URL	http://hdl.handle.net/10228/00008349

Machine learning approaches for addressing classification problems of four types of immune-peptides

４種類の免疫ペプチド分類問題を解決する
機械学習アプローチ

MST SHAMIMA KHATUN

ABSTRACT

Peptides play an important role in all aspects of the immunological reactions to invading cancer and pathogen cells. It has been known for over 40-years that peptides are critical influences in assembling the immune system against foreign invaders. Since then, new knowledge about the generation and function of peptides in immunology has supported efforts to harness the immune system to treat disease. Yet, with little immunological insight, most of the highly productive treatments, including vaccines, have been developed empirically. Nonetheless, increased knowledge of the biology of antigen processing as well as chemistry and pharmacological properties of antigenic and antimicrobial peptides has now permitted to development of drugs and vaccines. Due to advanced technologies, it is vitally important to develop automatic computational methods for rapidly and accurately predicting immune-peptides. In this thesis, the author focuses on the machine learning approaches for addressing classification problems of four types of immune-peptides (anti-inflammatory, proinflammatory, anti-tuberculosis, and linear B-cell peptides).

Numerous inflammatory diseases and autoimmune disorders by therapeutic peptides have received substantial consideration; however, the exploration of anti-inflammatory peptides via biological experiments is often a time consuming and expensive task. The development of novel *in silico* predictors is desired to classify potential anti-inflammatory peptides prior to *in vitro* investigation. Herein, an accurate predictor, called PreAIP (Predictor of Anti-Inflammatory Peptides) was developed by integrating multiple complementary features. We systematically investigated different types of features including primary sequence, evolutionary and structural information through a random forest classifier. The final PreAIP model achieved an AUC value of 0.833 in the training dataset via 10-fold cross-validation test, which was better than that of existing models. Moreover, we assessed the performance of the PreAIP with an AUC value of 0.840 on a test dataset to demonstrate that the proposed method outperformed the two existing methods. These results indicated that the PreAIP is an accurate predictor for identifying anti-inflammatory peptides and contributes to the development of anti-inflammatory peptides therapeutics and biomedical research. The curated datasets and the PreAIP are freely available at http://kurata14.bio.kyutech.ac.jp/PreAIP/.

A proinflammatory peptide (PIP) is a type of signaling molecules that are secreted from immune cells, which contributes to the first line of defense against invading pathogens. Numerous experiments have shown that PIPs play an important role in human physiology such as vaccines and immunotherapeutic drugs. Considering high-throughput laboratory methods that are time consuming and costly, effective computational methods are great demand to timely and accurately identify PIPs. Thus, in this study, we proposed a computational model in conjunction with a multiple feature representation, called ProIn-Fuse, to improve the performance of PIPs identification. Specifically, a feature representation learning model was utilized to generate a set of informative probabilistic features by making the use of random forest models with eight sequence encoding schemes. Finally, the ProIn-Fuse was constructed by the linearly combined models of the informative probabilistic features. The generalization capability of our proposed method evaluated through independent test showed that ProIn-Fuse yielded an accuracy of 0.746, which was over 10% higher than those obtained by the state-of-the-art PIP predictors. Cross-validation and independent results consistently demonstrated that ProIn-Fuse is more precise and promising in the identification of PIPs than existing PIP predictors. The web server, datasets and online instruction are freely accessible at http://kurata14.bio.kyutech.ac.jp/ProIn-Fuse/. We believe that the proposed ProIn-Fuse can facilitate faster and broader applications of PIPs in drug design and development.

Tuberculosis (TB) is a leading killer caused by *Mycobacterium tuberculosis.* Recently anti-TB peptides have provided an alternative approach to combat antibiotic tolerance. Herein, we have developed an effective computational predictor iAntiTB (identification of anti-tubercular peptides) that integrates multiple feature vectors deriving from the amino acid sequences via Random Forest (RF) and Support Vector Machine (SVM) classifiers. The iAntiTB combined the RF and SVM scores via linear regression to enhance the prediction accuracy. To make a robust and accurate predictor we prepared the two datasets with different types of negative samples. The iAntiTB achieved AUC values of 0.896 and 0.946 on the training datasets of the first and second datasets, respectively. The iAntiTB outperformed the other existing predictors. Thus, the iAntiTB is a robust and accurate predictor that is helpful for researchers working on peptide therapeutics and immunotherapy. All the employed datasets and software application are accessible at http://kurata14.bio.kyutech.ac.jp/iAntiTB/.

Linear B-cell peptides are critically important for immunological applications such as vaccine design, immunodiagnostic tests, antibody production, and disease diagnosis and therapy. The accurate identification of linear B-cell peptides remains challenging despite several decades of research. In this work, we have developed a novel predictor, iLBE (Identification of B-Cell Epitope), by integrating evolutionary and sequence-based features. The successive feature vectors were optimized by a Wilcoxon rank-sum test. Then the random forest (RF) algorithm used the optimal consecutive feature vectors to predict linear B-cell epitopes. We combined the RF scores by the logistic regression to enhance the prediction accuracy. The performance of the final iLBE yielded an AUC score of 0.809 on the training dataset. It outperformed other existing prediction models on a comprehensive independent dataset. The iLBE is suggested to be a powerful computational tool to identify the linear B-cell peptides and development of penetrating diagnostic tests. A web application with curated datasets is freely accessible of iLBE at http://kurata14.bio.kyutech.ac.jp/iLBE/.

Taken together, the above results suggest that our proposed predictors (PreAIP, ProIn-Fuse, iAntiTB, and iLBE) would be helpful computational resources for the prediction of anti-inflammatory, pro-inflammatory, tuberculosis, and linear B-cell peptides.

Keywords: Anti-inflammatory peptides, Proinflammatory peptides, Anti-tuberculosis peptides, Linear B-cell epitopes /peptides, Feature encoding, Feature selection, and Machine learning algorithms, Webserver applications

概要

ペプチドは、癌や病原体細胞に対する免疫反応のあらゆる側面で重要な役割を果たす。ペプチドが外来の侵入物に対する免疫系を起動する上で決定的な影響を与えることは 40 年以上前から知られている。それ以来、免疫学におけるペプチドの生成と機能に関する新しい知見は、病気を治療するために免疫系を利用する研究を支えてきた。依然として、免疫学的洞察がほとんどないため、ワクチンを含む効率的治療法のほとんどは、経験的に開発されている。それでもなお、抗原プロセシングの生物学、ならびに抗原性および抗菌性ペプチドの化学・薬理学に関する知見の増加により、現在、薬物およびワクチンの開発が可能になっている。高度な技術により、免疫ペプチドを迅速かつ正確に予測するためのコンピュータ技術を開発することが非常に重要である。この論文では、著者は 4 種類の免疫ペプチド（抗炎症、炎症誘発性、抗結核、および線形 B 細胞エピトープ）の分類問題に対処するための機械学習アプローチに焦点を当てる。

炎症性疾患および自己免疫疾患に対する治療用ペプチドは、多くの検討がなされてきた。しかし、生物学的実験による抗炎症ペプチドの探索は、多くの場合、時間と費用のかかる作業である。新しい*in siloco*予測器の開発は、*in vitro*実験に先立って、潜在的な抗炎症ペプチドを同定するために望まれている。ここでは、PreAIP（抗炎症ペプチドの予測器）と呼ばれる予測器が、複数の補完的機能を統合することによって開発された。一次配列、進化的および構造的情報を含むさまざまなタイプの特徴量を、ランダムフォレスト分類器を介して抽出した。最終的な PreAIP モデルは、10 分割交差検定によるトレーニングデータセットで 0.833 の AUC 値を達成した。これは、既存のモデルよりも優れた値である。さらに、独立の検証用データセットで AUC 値 0.840 を達成し、提案された方法が 2 つの既存の予測器よりも優れていることを示した。これらの結果は、PreAIP が抗炎症ペプチドを同定するための正確な予測器であり、抗炎症ペプチド治療および生物医学研究の開発に貢献した。用いたデータセットと PreAIP は、http：//kurata14.bio.kyutech.ac.jp/PreAIP/から自由に利用できる。

炎症誘発性ペプチド（PIP）は、免疫細胞から分泌されるシグナル伝達分子の一種であり、侵入する病原体に対する防御の第一線を担当する。多くの実験により、PIPはワクチンや免疫療法薬などにおいて重要な役割を果たすことが示されている。ハイスループットな生物実験に時間と費用が掛かることを考えると、効率的なコンピュータ予測は、PIPを短時間にかつ正確に特定するために大きな需要がある。したがって、この研究では、PIP識別性能を向上させるために、ProIn-Fuseと呼ばれる複数の特徴表現を組み合わせた計算モデルを提案した。具体的には、特徴表現学習モデルを利用して、8つのシーケンスエンコーディングスキームを備えたランダムフォレストモデルを利用することにより、確率的予測スコアを計算した。ProIn-Fuseは、確率的予測スコアの線形結合モデルによって構築された。提案手法の汎化性能を独立したテストデータで評価した結果、ProIn-Fuseの精度は0.746であり、これは最新のPIP予測器によって得られた精度よりも10％以上高かった。テストデータによる検証結果は、ProIn-Fuseが既存のPIP予測器よりも正確にPIP識別できることを示した。Webサーバー、データセット、および説明書は、http：//kurata14.bio.kyutech.ac.jp/ProIn-Fuse/から自由にアクセスできる。ProIn-Fuseは、ドラッグデザイン含む幅広いアプリケーションに応用できる。

結核（TB）は、結核菌によって引き起こされる疾患である。最近、抗結核ペプチドは抗生物質耐性に対抗するための代替アプローチを提供している。ここでは、ランダムフォレスト（RF）およびサポートベクターマシン（SVM）分類器を用いてアミノ酸配列に由来する複数の特徴ベクトルを統合する効果的な予測器iAntiTB（抗結核ペプチドの識別）を開発した。 iAntiTBは、線形回帰を介してRFスコアとSVMスコアを組み合わせて、予測精度を向上させた。ロバストで正確な予測器を作成するために、異なるタイプのネガティブサンプルを使用して2つのデータセットを準備した。iAntiTBは、1番目と2番目のデータセットのトレーニングデータセットでそれぞれ0.896と0.946のAUC値を達成した。iAntiTBは、他の既存の予測器の性能を上回った。このように、iAntiTBは、ペプチド治療および免疫療法に取り組んでいる研究者に役立つロバストで正確な予測器である。利用されたすべてのデータセットとソフトウェアアプリケーションは、http：//kurata14.bio.kyutech.ac.jp/iAntiTB/から自由にアクセスできる。

線形 B 細胞エピトープは、ワクチンの設計、免疫診断テスト、抗体産生、疾患の診断や治療などの免疫学的応用に非常に重要である。線形 B 細胞エピトープの正確な同定は、数十年の研究にもかかわらず、依然として挑戦的課題のままである。本研究では、配列の進化的特徴や物理化学的特徴等を統合することにより、新規な線形 B 細胞エピトープ予測モデル（iLBE）を開発した。Wilcoxon 順位和検定によって最適化した特徴ベクトル群をランダムフォレスト（RF）アルゴリズムを用いて学習して、線形 B 細胞エピトープの予測スコアを計算した。ロジスティック回帰を用いて RF スコアを組合せて、予測精度を高めた。 iLBE は、トレーニングデータセットで 0.809 の AUC を達成し、独立のテストデータセットを用いた検定では、既存の予測モデルの性能を超えた。線形 B 細胞エピトープを同定する強力な計算ツールである iLBE は、診断テストの開発に有用である。注釈付きデータセットを備えた iLBE モデルのウエブアプリケーションは自由にアクセスできる http://kurata14.bio.kyutech.ac.jp/iLBE/。

キーワード：抗炎症ペプチド、炎症誘発性ペプチド、抗結核ペプチド、線形 B 細胞エピトープ/ペプチド、特徴符号化、特徴選択、機械学習アルゴリズム、Web サーバーアプリケーション

TABLE OF CONTENTS

LIST OF PUBLICATION

[1] **Mst. Shamima Khatun**, Md. Mehedi Hasan, Watshara Shoombuatong, Hiroyuki Kurata. ProIn-Fuse: improved and robust prediction of proinflammatory peptides by fusing of multiple feature representations. ***Journal of Computer-Aided Molecular Design*** (2020). https://doi.org/10.1007/s10822-020-00343-9.

[2] Md. Mehedi Hasan, Shaherin Basith, **Mst. Shamima Khatun**, Gwang Lee, Balachandran Manavalan, Hiroyuki Kurata. Meta-i6mA: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. ***Briefings in Bioinformatics*** 2020, bbaa202, https://doi.org/10.1093/bib/bbaa202.

[3] **Mst. Shamima Khatun**, Watshara Shoombuatong, Md. Mehedi Hasan, Hiroyuki Kurata. Evolution of sequence-based bioinformatics tools for protein-protein interaction prediction. ***Current Genomics.*** DOI: 10.2174/1389202921999200625103936 (2020).

[4] Md. Mehedi Hasan, Balachandran Manavalan, **Mst. Shamima Khatun**, and Hiroyuki Kurata**.** i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. ***International journal of biological macromolecules*** 157, 752-758 (2020).

[5] Md. Mehedi Hasan, Balachandran Manavalan, Watshara Shoombuatong, **Mst Shamima Khatun**, Hiroyuki Kurata. i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation. ***Plant Molecular Biology***. 103:225–234(2020).

[6] Md Parvez Mosharaf, Md Mehedi Hassan, Fee Faysal Ahmed, **Mst. Shamima Khatun,** Mohammad Ali Moni, Md Nurul Haque Mollah. Computational prediction of protein ubiquitination sites mapping on Arabidopsis thaliana. ***Computational Biology and Chemistry***; 85:107238.doi: 10.1016/j.compbiolchem.2020.107238 (2020)..

[7] Md. Mehedi Hasan, Balachandran Manavalan, Watshara Shoombuatong, **Mst. Shamima Khatun**, and Hiroyuki Kurata. i4mC-Mouse: Improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes. ***Computational and Structural Biotechnology Journal.*** 18:906-912. doi: 10.1016/j.csbj.2020.04.001 (2020).

[8] Md. Mehedi Hasan+, **Mst. Shamima Khatun**+, and Hiroyuki Kurata. Computational

Identification of Linear B-Cell Epitopes by Integrating Sequence and Evolutionary Features. ***Genomics, Proteomics and Bioinformatics***. https://doi.org/10.1016/j.gpb.2019.04.004 **(equal contribution)**

[9] **Mst. Shamima Khatun**, Md. Mehedi Hasan, and Hiroyuki Kurata. PreAIP: Computational Prediction of Anti-inflammatory Peptides by Integrating Multiple Complementary Features, ***Frontiers in Genetics,***10:129**;** doi: 10.3389/fgene.2019.00129 **(2019)**.

[10] Md. Mehedi Hasan, Md. Mamunur Rashid, **Mst. Shamima Khatun**, and Hiroyuki Kurata. Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information, ***Scientific Reports,*** 9(1):8258. doi: 10.1038/s41598-019-44548-x (2019).

[11] Md. Mehedi Hasan, Balachandran Manavalan, **Mst. Shamima Khatun**, and Hiroyuki Kurata. Prediction of S-nitrosylation sites by integrating support vector machine and random forest. ***Molecular Omics***, (2019) 15: 451-458, DOI: 10.1039/c9mo00098d.

[12] **Mst. Shamima Khatun**, Md. Mehedi Hasan, and Hiroyuki Kurata. Efficient Computational Model for Identification of Anti-tubercular Peptides by Integrating Amino Acid Patterns and Properties. ***FEBS letters***, 593(21): 3029-3039.doi: 10.1002/1873-3468.13536. doi: 10.1002/1873-3468.13536 (2019).

[13] Md. Mehedi Hasan*, **Mst Shamima Khatun**, and Hiroyuki Kurata. Large-Scale Assessment of Bioinformatics Tools for Lysine Succinylation Sites. ***Cells***, 8(2), 95, doi.org/10.3390/cells8020095 (2019).

[14] **Mst. Shamima Khatun,** Md. Mehedi Hasan, Md. Nurul Haque Mollah, and Hiroyuki Kurata. "SIPMA: A Systematic Identification of Protein-Protein Interactions in *Zea mays* Using Autocorrelation Features in a Machine-Learning Framework," *IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)*, Taichung, Taiwan, 122-125.doi: 10.1109/BIBE.2018.00030 (2018).

[15] Md. Mehedi Hasan, **Mst. Shamima Khatun**, and Hiroyuki Kurata. A comprehensive review of *in silico* analysis for protein S-sulfenylation sites. ***Protein & Peptide Letters*, *25,* 1-7,** doi: 10.2174/0929866525666180905110619 **(2018)**.

[16] Md. Mehedi Hasan, **Mst. Shamima Khatun**, Md. Nurul Haque Mollah, Cao Yong, and Guo Dianjing. NTyroSite: Computational Identification of Protein Nitrotyrosine Sites Using Sequence Evolutionary Features. ***Molecules***. *23*(7), 1667, doi:10.3390/molecules23071667 (2018).

[17] Md. Mehedi Hasan, **Mst. Shamima Khatun**, and Hiroyuki Kurata. Computational

modeling of lysine post-translational modification: an overview. ***Curr Synthetic Sys Biol*** 6: 137. doi:10.4172/2332-0737.1000137 (2018).

[18] Md. Mehedi Hasan and **Mst. Shamima Khatun**. Prediction of protein post-translational modification sites: an overview. ***Annals Proteom Bioinform.*** 2: 049-057. DOI: 10.29328/journal.apb.1001005 (2020).

[19] Md. Mehedi Hasan and **Mst. Shamima Khatun.** Recent progress and challenges for protein pupylation sites prediction, ***EC Proteomics and Bioinformatics***, 2(1): 36-45, (2017).

[20] Md. Mehedi Hasan, **Mst. Shamima Khatun**, Md. Nurul Haque Mollah, Cao Yong, and Guo Dianjing. A systematic identification of species-specific protein succinylation sites using joint element features information. ***International Journal of Nanomedicine,*** 12:6303-6315. doi: 10.2147/IJN.S140875 (2017).

ACKNOWLEDGEMENTS

This thesis represents not only my work at the keyboard, but it's also a milestone of 3 years of work at Kyushu Institute of Technology. Thanks to almighty Allah who keeps me safe and healthy. First and foremost, I wish to thank my supervisor **Professor Hiroyuki Kurata** who encouraged me to do biomedical research in the area of bioinformatics and taught the art of research scientist. I gratefully acknowledge Kurata Sensei for financial support in my whole PhD journey from the Department of Bioscience and Bioinformatics, Faculty of Computer Science and System Engineering, Kyutech, Japan. I would like to say a very big thanks to Kyushu Institute of Technology for giving such beautiful friends and well-wishers from different countries including Japan. Special thanks from the bottom of my heart to Dr. Mehedi Hasan for his support whole of my 3 years' journey. My sincere thanks go to Dr. Kazuhiro Maeda, Dr. Yu Matsuoka, and Dr. Md. Mamunur Rashid for their generous support in this three years'. I also want to thank the other members of the Kurata Laboratory for creating a wonderful research environment inside of the room. I would like to thank Dr. Nurul Hoque Molla for giving delightful suggestions and encourages me in critical moment of research.

Special thanks to the co-supervisors Prof. Dr. Tetsushi Yada, Prof. Dr. Koichi Hirata, and Prof. Dr. Kosuke Hanada for their insightful suggestions.

Mst. Shamima Khatun
Kyushu Institute of Technology, Japan

LIST OF ABBREVIATIONS

PreAIP	:	Prediction of anti-inflammatory peptide
SF	:	Structural features
PKA	:	Profile-based composition of *k*-spaced amino acid pairs
KSAP	:	*k*-spaced amino acid pairs
PSI-BLAST	:	Position-specific iterative basic local alignment search tool
TPR	:	True positive rate
FPR	:	False positive rate
SVM	:	Support vector machine
RBF	:	Radial basis function
PDB	:	Protein data bank
TB	:	Tuberculosis
RF	:	Random forest
MCC	:	Mathew’s correlation coefficient
Ac	:	Accuracy
Sn	:	Sensitivity
Sp	:	Specificity
TN	:	True negatives
TP	:	True positives
FP	:	False positives
FN	:	False negatives
C5N5-DC	:	N-terminal 5 and C-terminal 5 dipeptides composition
AAindex	:	amino acids index
ROC curve	:	Receiver operating characteristic curve
AUC	:	Area under the curve
AB	:	Adaboost
GB	:	Gradient boosting
KNN	:	K-nearest neighbour
NB	:	Naïve bayes
Tr	:	Training

Ind	:	Independent
MS	:	Mass spectrometry
KW	:	Wilcoxon rank sum test
AIP	:	Anti-inflammatory peptide
pKSAAP	:	Profile-based composition k-space of amino acid pair
HI	:	High indices
ASA	:	Accessible surface area
BTA	:	Backbone torsion angles
SS	:	Secondary structure
IG	:	Information gain
ANN	:	Artificial neural network
KW	:	Kruskal-wallis
AVFS	:	Average value of feature scores
LA	:	Less accurate
MA	:	More accurate
BE	:	Binary encoding
CV	:	Cross-validation
TB	:	Tuberculosis
Mtb	:	Mycobacterium tuberculosis
MDR	:	Multi-drug resistance
iAntiTB	:	Identification of anti-tubercular peptides
DPC	:	Dipeptide composition
TPC	:	Tripeptide composition
LR	:	Linear regression
ProIn-Fuse	:	Prediction of proinflammatory peptides
BCE	:	B-Cell epitope
PSSM	:	Position-specific scoring matrix
PKAF	:	Profile-based amino acids frequency
AFC	:	Amino acid frequency composition
AVAFC	:	Average value of the AFC
ChIP	:	Chromatin immunoprecipitation
DL	:	Deep learning
iLBE	:	Identification of linear B-cell epitope

RNN	:	Recurrent neural networks
DBN	:	Deep belief network
DNN	:	Deep neural networks

LIST OF TABLES

LIST OF TABLES

CHAPTER 1

CHAPTER 1 INTRODUCTION

1.1 Immune-peptide developments

All aspects of the immunological responses, peptides play a critical role to regulate invading pathogens and cancer cells (Alt et al., 2015; De Lorenzi et al., 2017; Hosokawa et al., 2006; Lomash et al., 2010; Margulies et al., 2019; Rosenthal, 2005). Immune-peptides are critical aspects in activating the immune cell against foreign invaders. The function and generation of peptides in immunology have maintained the immune cycle in a cell to treat disease (Gokhale and Satyanarayanajois, 2014; Mojsoska and Jenssen, 2015; Skovbakke and Franzyk, 2017; Teveroni et al., 2016). So far, with little immunological insight, most of the highly effective treatments, including vaccines, have been prepared empirically, with little immunological perception. Nonetheless, improved knowledge about the pharmacological and chemical properties of antigenic and antimicrobial peptides presentation, processing, and acknowledgment by immune cells, has now permitted to development of vaccines and drugs.

Immune peptides can work as immunomodulating agents by either stimulating the immune reaction or blocking the immune reaction. Though the immune-peptide developments are well advanced, autoimmunity arises when autoreactive immune-peptides are triggered to motivate their responses against self-tissues [3]. This occurs due to a lack of breakdown of the mechanism that reins immune tolerance, resulting in miscarriage of the host system to differentiate the cells to self from non-self (Cunningham et al., 2017; Guichard et al., 1994; Kemp, 1990; Kim et al., 2020; Skovbakke and Franzyk, 2017). Different types of organ-specific diseases that may occur via immune-peptides include celiac disease, multiple sclerosis, Type 1 diabetes mellitus, and myasthenia gravis. Immune-peptides suppress or block the immune system as in the case of autoimmune diseases, allergy/asthma, inflammation, and transplantation. A common solid-phase synthesis process of the peptide is shown in **Figure 1-1**.

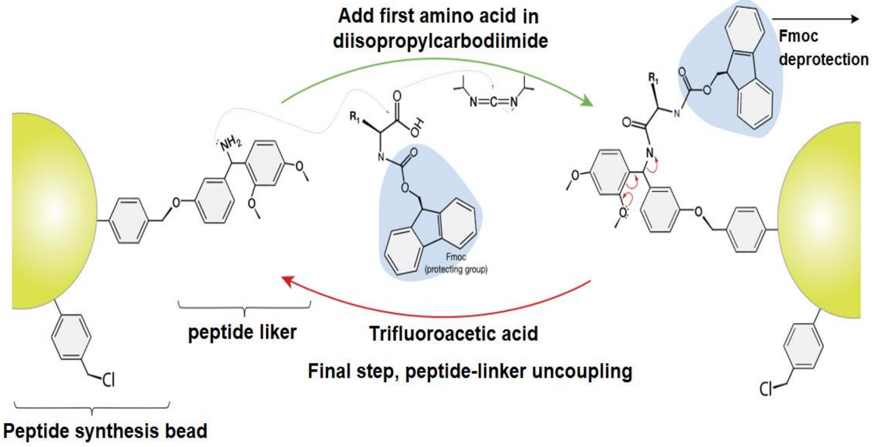

Figure 1-1 A solid-phase peptide synthesis process. Fmoc implies ‘fluorenylmethoxycarbonyl’.

Peptides are categorized or classified according to their functions and sources (Jurczak et al., 2020). We summarized the existing active peptides and their therapeutic agents in **Table 1-1**. Some groups of peptides contain brain peptides, endocrine peptides, incentive peptides, fungal peptides, plant peptides, antibiotic/bacterial peptides, skin peptides/amphibian, venom peptides, anticancer/cancer peptides, vaccine peptides, immune/inflammatory peptides, gastrointestinal peptides, cardiovascular peptides are described in the Handbook of “Biologically Active Peptides” (Kastin, 2017). Overall, peptides are linear but some of the rope structures (Kieber-Emmons et al., 1997). The component of the antioxidant peptides has been defended by the common non-ribosomal peptides. Other non-ribosomal peptides are most common in unicellular organisms, fungi, and plants that are synthesized by modular enzyme developments called nonribosomal peptide synthetises.

Table 1-1 List of different types of peptides.

Peptides	**Description**
Anti- inflammatory	Generally, a peptide was considered as anti-inflammatory (positive sample) if the anti-inflammatory cytokines of peptides induce any one of IL-10, IL-4, IL-13, IL-22, TGFb, and IFN-a/b in T-cell analyses of mouse and human. Numerous endogenous peptides recognized through inflammatory reactions function as anti-inflammatory agents can be employed by new therapies for autoimmune and inflammatory illnesses.
Pro-inflammatory	A proinflammatory cytokine or an inflammatory peptide is referred

	to as a type of signaling molecules, which is secreted from immune cells and certain cell types for promoting the inflammation. Different studies reported that PIPs play an important role in human physiology such as vaccines and immunotherapeutic drugs. Nonetheless, these peptides may cause unwanted immuno-activity in the B or T cell instigation and other proinflammatory events, which belong to any of the proinflammatory cytokines (IL1α, TNFα, IL1β, IL12, IL18, and IL23)
Anti-tubercular	Tuberculosis (TB) peptides is generated by *Mycobacterium tuberculosis* (Mtb), is a type of infective disease, being responsible as a major threat for the human beings. AntiTB peptides with low immunogenicity make them a possible complement for expectable TB drugs.
Anti-cancer	A series of short peptides (~10–60 amino acids) consisted in anti-cancer peptides of which constrain by tumour or migration cell proliferation, or destroy the development of tumour blood vessels.
Dipeptidyl peptidase IV inhibitory	To the treatment of Type 2 diabetes (T2D), the dipeptidyl peptidase IV inhibition is well known as a new possibility drug target. In T2D subjects concentration, these peptides have been revealed by normalizing the blood glucose in cell.
Tumor T cell antigens	Tumor-germline antigens are categorized by their appearance in the testis and on tumors. Epitopes from these antigens are not predictable in the testis, as those cells do not express of major histocompatibility complex, thereby making them attractive targets for T-cell immunotherapy.
Brain peptides	The classical brain peptides are assembled into broad families such as the neurohypophyseal hormones, the hypothalamic-releasing hormones, the opioids, the pituitary peptides, the gastrointestinal peptides, and the tachykinins.
Linear B-cell epitope	Linear B-cell epitopes are critically important for immunological applications, such as vaccine design, immunodiagnostic test,

	antibody production, as well as disease diagnosis and therapy. Nowadays, biopharmaceutical research and development of epitope-based antibodies are growing up due to their high efficiency, biosafety, and acceptability. Thus, the analysis of BCEs is prerequisite for the development of penetrating diagnostic tests and design of the operative vaccines.
Therapeutic peptides	A chain of 40 or less amino acids is regulated and considered as peptide therapeutic. Therapeutic peptides are considered by rational methods with high specificity to bind and modulate a protein interaction of interest.
Cell-penetrating peptides	Cell-penetrating peptides (CPPs) are short peptides that facilitate *cellular* intake and uptake of molecules ranging from nanosize particles. CPPs deliver the cargo into cells. It is commonly use in research and medicine.
Tumor homing peptides	To recognize the tumor cells, tumor homing peptides are linear or cyclic peptides that contained a few amino acids inherent properties. It unambiguously bind to the cell receptors and present on the tumor lymphatic vessels, tumor blood vessels, or tumor cells.
Antiviral peptides	The virus explicit antiviral peptides are known as virucidal. It is directly targeted the viral proteins. In specific regions or components, most of the antivirals have been described to inhibit the development of viruses.
Anti-angiogenic peptides	Preventing the interaction with the receptor via antagonistic peptides could present an effective anti-angiogenic therapy. Most important modulators of angiogenesis is vascular endothelial growth factor.
Host defense peptides	In all complex life forms, host defence peptides (HDPs) are short cationic amphipathic peptides. HDPs have critical roles in the body's reaction to inflammation and infection.
Haemolytic peptides	In clinical trials, therapeutic peptides could be attributed to their toxicity profiles like haemolytic activity that hamper the progress of peptides as drug candidates.

Anti-hypersensitive peptides	Blood hypertension and pressure are massively influenced by the Angiotensin I-converting enzyme. Anti-hypersensitive peptides are occurred during intestinal digestion through the aid of enzymes.
Bitter peptides	In food and pharmaceutical application, bitterness of whey protein hydrolysates (WPH) can negatively affect limit utilization and product quality. The bitter peptides are documented in a commercial WPH using sensory-guided fractionation techniques.
Umami peptides	Food seasoning and healthy eating, umami ingredients are very important. Development and application in food products, umami peptides are found in the natural ingredients with a high demand
Quorum sensing peptides	To activate intracellular response regulators via phosphor-transfer, the *quorum sensing peptides* bind membrane associated receptors. These peptides response the regulators target gene expression.

1.2 Experimental methods for peptide identification

The peptides or epitopes of proteins have been identified by a diversity of experimental techniques including western blotting (Jaffrey et al., 2001), and eastern blotting (Welsch and Nelsestuen, 1988), radioactive chemical method (Slade et al., 2014), mass spectrometry (Agarwal et al., 1969; Medzihradszky, 2005), liquid chromatography (Welsch and Nelsestuen, 1988), and chromatin immunoprecipitation (ChIP)(Umlauf et al., 2004). The MS technique is one of the mainstay routes in detecting peptides in a high-throughput manner. The new MS and capillary liquid chromatography instrumentation have made a revolutionary advance in enrichment strategies in our growing knowledge of many peptides (Doll and Burlingame, 2015). In the last decade of the actual description of many peptides complexity has emerged through diverse technologies and thousands of precise peptides can now be identified with high confidence (Choudhary et al., 2009; Hebert et al., 2014; Hendriks et al., 2014; Imamura et al., 2014; Kim et al., 2011; Masuda et al., 2011; Olsen et al., 2010; Richards et al., 2015; Trinidad et al., 2012). A similar strategy of fragmentation for peptide identification is the beam-type collision-induced dissociation, also called higher-energy collisional dissociation (Syka et al., 2004). These types of fragmentation are characterized by higher activation energy. Most of the fragmentation methods of precursor ions are based on radical anions or thermal electrons

(Myers et al., 2013). These methods are advantageous over collisionally activated dissociation methods for detecting the unstable peptides (e.g., anti-cancer and tumor) because the peptide backbone fragmentation method is virtually independent of the amino acid sequence (Han et al., 2012; Moremen et al., 2012; Ramstrom and Sandberg, 2011).

Notwithstanding the increasing number of experimentally examined AIPs *in vivo*, the molecular mechanism of AIP specificity remains largely unknown. Particularly, the experimental analysis of peptides often requires labor-intensive sample preparations and hazardous or expensive chemical reagents. For instance, in the radioactive assay of protein in the kinase-based methods are often included the radioactive label of ATP as a substrate donor for identifying peptides (Slade et al., 2014). In summary, the identification of peptides by the experimental techniques is laborious, time-consuming, and usually expensive. As an alternative, the machine learning approaches are more efficient for identifying large-scale novel peptides. In the next section, the author will introduce machine learning approaches for different types of peptides prediction.

1.3 Computational approaches for immune-peptides prediction

A large number of computational approaches have been taken toward predicting peptide presentation by different approaches. The last few decades have been remarkable progress in the identification and functional analysis of peptides in proteins for different disease biomarkers. Peptides play a vital role in protein folding, protein function, and interactions with other proteins (DeMartino, 2009; Striebel et al., 2009). Because of critical functions of immune-peptides, it is very important to prediction and analysis the function of diverse peptides. On the other hand, large-scale experimental analysis of immune peptide is time-consuming, laborious, and expensive. An alternative, computational approach that provides an accurate and reliable prediction of immune-peptides is required to complement the experimental efforts and to access the prompt identification of potential immune-peptides prior to their synthesis. In addition, the computational tools can narrow down the number of potential candidates and rapidly generate useful information for investigating further experimental approaches.

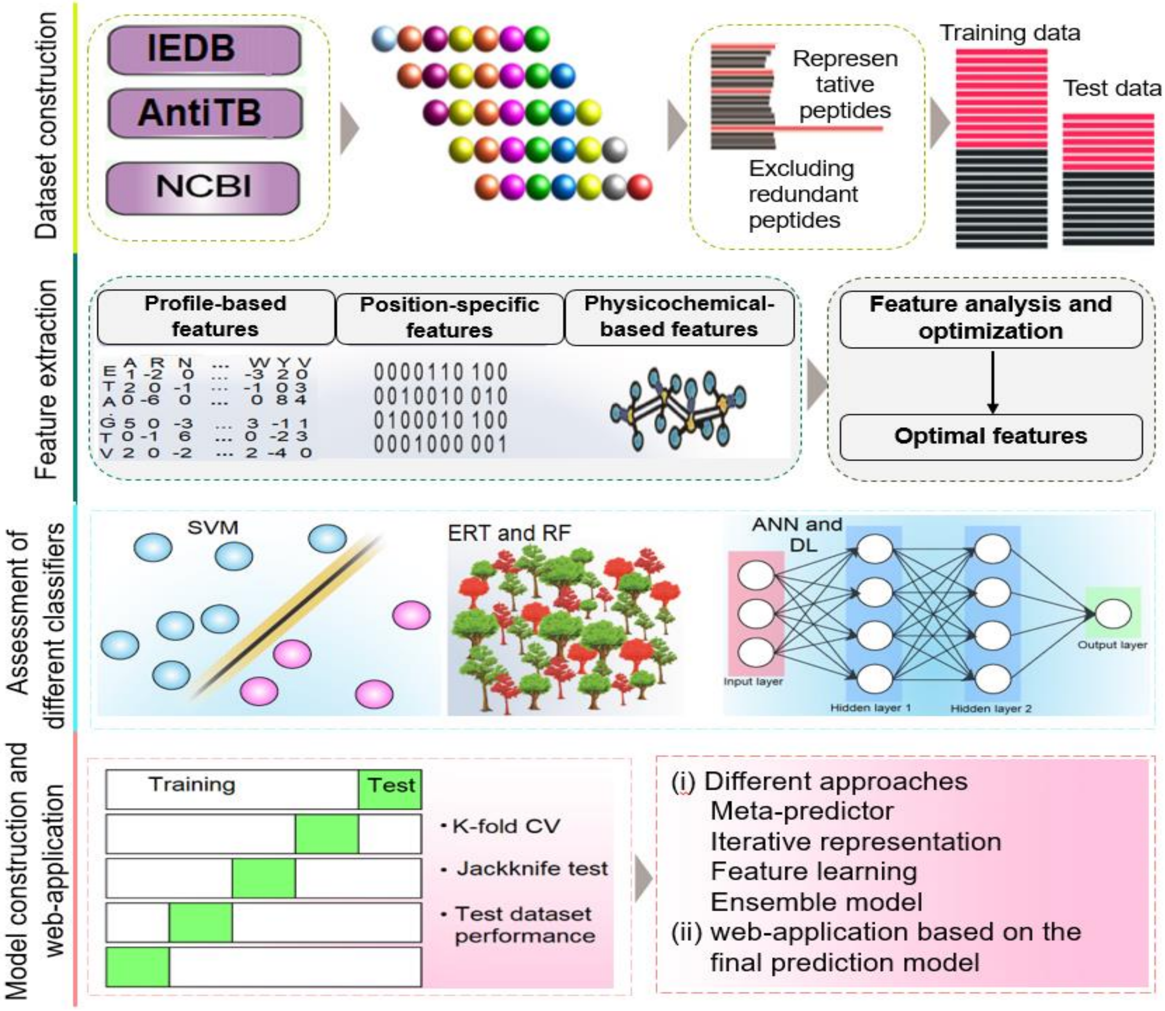

Figure 1-2 A workflow of computational methods for immune-peptide prediction.

Thus far, the prediction of immune-peptides is an important research topic in the field of immune bioinformatics. Although great progress has been made by employing various machine learning approaches with numerous feature vectors, the problem is still far from being solved. In this work, the author focuses on the machine learning approaches for addressing classification problems of four types of immune-peptides (anti-inflammatory, proinflammatory, anti-tuberculosis, and linear B-cell peptides). A workflow of the prediction pipeline of peptides is shown in **Figure 1-2**. In the next section, the author will discuss the importance of peptides prediction.

1.3.1 Peptide databases

Recently, several peptide databases have been industrialized to maintain and accumulate data on different peptides (Basith et al., 2020b). **Table 1-2** shows the primary databases summarizing data on general and specific functional peptides, such as anti-cancer peptides, cell-penetrating, immune-peptide, anti-inflammatory peptides, Quorum sensing, and Antihypertensive peptides. The establishment of these databases is to generate larger positive and negative samples in regions that are important for peptide drug development and endorse the utility of machine learning approaches. Though, few limitations are associated with these databases that need to be solved. First, most of the established databases focus only on specific bio-peptides like antimalarial peptides, anti-cancer peptides, Anti-tubercular peptides, and so forth. Second, the reported databases cover only positive samples. To overwhelmed these limitations, construct a comprehensive peptide database is essential that integrates more diverse bioactive peptides and includes negative samples for developing effective machine learning models. Additionally, combined efforts of different scientific disciplines will help to compile, link, and develop a large peptide data resource, in which peptide sequences with diverse biological activities could be retrieved from a single large peptide source.

Table 1-2 General and specific databases currently available for the prediction of peptide activities

Peptide database type	Database	Database link	Description	Publication year
General	FeptideDB (Panyayai et al., 2019)	http://www4g.biotec.or.th/FeptideDB/index.php	Database bioactive peptides for foods	2019
	PepBank (Shtatland et al., 2007)	http://pepbank.mgh.harvard.edu/	It's a public database and includes of bioactive peptides with ≤ 20 length of amino acids.	2007
	SATPdb (Singh et al., 2016)	http://crdd.osdd.net/raghava/satpdb/links.php	This database covers peptide 10 categories peptide including toxic peptides, antibacterial peptides, anticancer peptides, anti-viral peptides, antiparasite peptides, and so forth	2015

Anticancer	LAMP (Zhao et al., 2013)	http://biotechlab.fudan.edu.cn/database/lamp/	Database of antibacterial peptides and anticancer peptides	2013
	CancerPPD (Tyagi et al., 2015)	http://crdd.osdd.net/raghava/cancerppd/	Database of anticancer peptides and proteins	2015
	DRAMP (Fan et al., 2016; Kang et al., 2019)	http://dramp.cpu-bioinfor.org/	Database of antimicrobial peptides, anticancer peptides, antibacterial peptides, and so forth	2013
Cell penetrating	CPPsite/CPPsite 2.0 (Agrawal et al., 2016; Gautam et al., 2012)	https://webs.iiitd.edu.in/raghava/cppsite/	Maintains experimentally validated cell penetrating peptides	2012
Quorum sensing	Quorumpeps (Wynendaele et al., 2013)	http://quorumpeps.ugent.be/	Resource for quorum-sensing signaling peptides	2013
Antihypertensive	AHTPDB (Kumar et al., 2015)	http://crdd.osdd.net/raghava/ahtpdb/	Manually curated database of experimentally validated antihypertensive peptides	2015
Antitubercular	AntiTbPdb (Usmani et al., 2018b)	https://webs.iiitd.edu.in/raghava/antitbpdb/	Database of antitubercular or antimycobacterial peptides	2018
Anti-inflammatory	IEDB (Vita et al., 2015)	http://www.iedb.org/	This database contained peptide data on antibody and T-cell epitopes in infectious diseases, transplantation, autoimmunity, and allergy.	2015

1.3.2 Feature descriptors for the prediction of peptides

Feature extraction is one of the most important steps for predicting protein, peptides, DNA, and RNA sequences(Hasan et al., 2020a; Hasan et al., 2018c, 2019a, 2020b; Hasan et al., 2018d; Hasan et al., 2017c; Hasan et al., 2019c, 2020c; Hasan et al., 2020d, e; Hasan et al., 2019d; Khatun et al., 2019a; Khatun et al., 2020a; Khatun et al., 2020b; Mosharaf et al., 2020; Shahjahan et al., 2020). Appropriate features in the prediction model enable the accurate prediction of immune-peptides. In general, these features refer to the characterization of the sequences and local structures around these protein functional sites. Ideally, the features can clearly distinguish peptides from the random features. In the real world, however, the feature of protein functional sites can also exist on the non-functional sites of proteins. In the prediction peptides, this specific problem is particularly prominent due to the sequence diversity. For instance, some motifs are very weak and some are not available without the sequence evolutionary information (Liu et al., 2011; Passerini et al., 2006; Ren et al., 2008; Sharma et al., 2007; Vandermarliere and Martens, 2013; Youn et al., 2007). To address this problem, we can search PSI-BLAST (Altschul et al., 1997) against the NCBI NR database to generate a profile (i.e., position-specific scoring matrix (PSSM)) to generate enhanced features. Such sequence profiles reflect the conservation and variation between protein sequences through evolutionary information (Dekker et al., 2004; Gobel et al., 1994; Lockless and Ranganathan, 1999).

In the prediction of immune-peptides, researchers have made plenty of efforts for mining the different characteristics of peptides. These characteristics might be suitable for a particular peptide classification problem, thus mining new features is always an important task for peptides prediction. The features are mainly obtained from two ways, namely based on the peptide sequences and structures. In addition to the amino acid sequence itself, the physicochemical properties of amino acids have also been widely used in the prediction of peptides (Xu et al., 2015; Zhao et al., 2015). Some of the common physicochemical features include hydrophilicity/hydrophobicity, pKa value of the amino acid residues, the polarity of the amino acid (positively charged residues, residues with negatively charged and uncharged residue), the volume of amino acid side chains, whether it contains benzene, sulfur and so on. At present, most of the physicochemical properties of amino acid residues have been converted into numbers and stored in the famous amino acid index (AIP) database (Kawashima and

Kanehisa, 2000; Kawashima et al., 1999; Kawashima et al., 2008). Until now, the AIP database contains 544 physicochemical properties of amino acid residues, which can be used as a feature set for analyzing immune-peptides.

Recently, several types of pepptide structure features proposed. For example, one can examine the amino acid solvent accessibility of immune-peptides. Analyzing the residue interactions that maintain the stability of protein structures (including hydrophobic interactions, electrostatic interactions, hydrogen bonds, van der Waals interactions, disulfide bonds, and so on) may be also helpful (Halperin et al., 2008; Mooney et al., 2005). Moreover, the residues' structural flexibility information like B-factor and root mean square deviation is sometimes useful, too. Finally, some of the residue contact network parameters (degree, betweenness, closeness, and clustering coefficient) were used as features for peptide prediction (Gupta et al., 2016; Tang et al., 2016). For a real-world prediction task, note that the researchers usually use the integrated feature set to predict the immune-peptides.

1.4 Machine learning approaches for immune-peptides prediction

After determining the appropriate features, the next job is to use an appropriate machine learning algorithm to classify these features for the prediction of immune-peptides. It will improve the accuracy of the prediction if the prediction algorithm is appropriate. In early 1959, Arthur Lee Samuel defined machine learning as "the field of study that gives computers the ability to learn without being explicitly programmed" (Phil, 2013). For the prediction of peptide sequences, some common machine learning algorithms are widely used such as support vector machine (SVM), and random forest (RF), Naïve Bayes (NB), and deep learning (DL). Subsequently, the author will introduce these four common machine learning algorithms.

1.4.1 Support vector machine

To classify the PPI datasets, SVM, or kernel machines are used (Hajisharifi et al., 2014). The SVM maximizes the margins that are related to the inevitability of its classification. The objective of this classifier is likely to have small margins (Hasan et al., 2015) using a label of the training dataset. SVM is very influential and can classify problems with random density information, although it needs large memory requirements and a complex format. The SVM is a little bit slow to train and assess the high dimensional features via radial basis function kernel. Another disadvantage is that the parameters can significantly alter the results. We refer to more details (Hasan et al., 2015; Jia et al., 2015; Kurata, 2018).

1.4.2 Random forest

Random forest is an ensemble and supervised machine learning algorithm (Breiman, 2001). It can integrate multiple classifiers to improve the performances of the prediction (Maclin and Opitz, 1999; Polikar, 2006; Rokach, 2010). The RF algorithm involves numerous ensemble decision trees that can categorize the two-class prediction problem (Liaw, 2002; Schaduangrat et al., 2019; Shoombuatong et al., 2019; Su et al., 2019; Win et al., 2017). On the training model, each decision tree is built using the casual feature vectors that are sampled from a dataset in every node in a tree independently. Then each classification tree is entirely grown via randomly selected variables. To categorize a new entity, the response vector keeps each of the trees in the forest. Allowing the majority voting, one class is allocated to the entity. The RF is an effective algorithm when there exist a large number of features and datasets, and can rank important features for accurate classification (Manavalan et al., 2018c; Manavalan et al., 2018d). The RF is widely used in computational biology research (Boopathi et al., 2019; Hasan et al., 2017b; Hasan et al., 2016; Hasan et al., 2015; Manavalan et al., 2018a; Manavalan et al., 2018c) (M. S. Khatun, 2018).

1.4.3 Naïve Bayes

Naïve Bayes is a predictive algorithm based on the statistical learning theory of the Bayesian theorem. The advantages of this algorithm are very simple and high speed. In the Bayesian theorem, the posterior probability of a random event is the conditional probability, which is assigned after the relevant evidence has been taken into account. Bayesian assumes that a property of a given value is affected in the other values. This assumption is not often established on the model, so its accuracy can be rejected for other properties of the class forecasting models, such as linear regression and logistic regression models. The majority of biologists think that for analyzing the biological data Naïve Bayes is an important algorithm (Rani and Pudi, 2008). Although, these methods are many outliers affected and do not handle the noise model (David J. Hand 2001). In bioinformatics research, Naïve Bayes algorithms are widely used (Shao et al., 2009; Sheppard et al., 2013; Zhang et al., 2006).

1.4.4 Deep learning

Deep learning (DL) consists of several approaches including Recurrent Neural Networks

(RNN), Deep Belief Networks (DBNs), and Deep Neural Networks (DNN), (Chaudhary et al., 2018; Yao et al., 2019; Zhang et al., 2018). Different DL algorithms are suitable for different specific applications. For instance, for the analysis of sequential information, RNNs are appropriate. The DBNs are decent at examining inside associations in high-dimensional data. To predict PPIs, DNN is one of the most suitable ML algorithms (Sun et al., 2017). The DNN input should be the vectors with a fixed dimension. The main parts of the DNN component are to remove highly homologous sample information and eliminate noise, and to decrease data dimensions. DNN architectures are assembled layer-by-layer with a greedy algorithm. DNN helps to pick out unravel features to improve performance.

In summary, the machine learning algorithm is a subfield of computer science and statistics that evolved the study of pattern recognition and computational learning theory in artificial intelligence. For immune-peptides prediction, a machine learning algorithm is an essential step for testing the model performance. In the next two sections, the author will introduce four types of immune-peptides (anti-inflammatory, pro-inflammatory, anti-tubercular, and linear B-cell peptides) prediction by using machine learning approaches.

1.5 Research progress of anti-inflammatory peptides

The present therapy for autoimmune and inflammatory peptides (PIP) involves the use of non-specific anti-inflammatory drugs and other immunosuppressant's (Lowenberger, 2001; Reichhart and Achstetter, 1990; Yi et al., 2019), which are frequently related to different side effects, such as initiation of a higher possibility of infectious diseases and ineffectiveness alongside inflammatory disorders (Tabas and Glass, 2013). Notwithstanding the increasing number of experimentally examined AIPs in vivo, the molecular mechanism of AIP specificity remains largely unknown. On the other hand, large-scale experimental analysis of AIPs is time-consuming, laborious, and expensive. An alternative, computational approach that provides an accurate and reliable prediction of AIPs is required to complement the experimental efforts and to access the prompt identification of potential AIPs prior to their synthesis. To date, three machine learning approaches have been proposed to predict AIPs (Gupta et al., 2017; Manavalan et al., 2018b). In 2017 Gupta et al. employed hybrid features with a SVM classifier to develop the AntiInflam predictor (Gupta et al., 2017). Manavalan et al. developed the AlPpred predictor by using the primary sequence encoding features. Recently, the author proposed a PreAIP predictor by integrating multiple complementary sequence features. Even though the performances of the existing predictor were satisfactory, there is room to advance the prediction performance.

1.6 Research progress of pro-inflammatory peptides

A proinflammatory cytokine or an inflammatory peptide (PIP) is referred to as a type of signaling molecule, which is secreted from immune cells and certain cell types for promoting inflammation (Watkins et al., 1995; Zhang and An, 2007). The importance of PIPs is confirmed through the pathophysiological dealings (Mukhopadhyay et al., 2014; Zhao et al., 2005). For instance, Herpes Simplex Virus-2 produces a glycoprotein G-2 through the gG-2p20 peptide that causes proinflammatory responses in human neutrophils and activates as an effective antineoplastic agent (Bellner et al., 2005; Bylund et al., 2001). Similarly, the C-peptide of PIPs produces proinsulin which is used in peptide-therapeutics but leads to inflammation in vasculature and kidney or long-term deterioration of diseases (Vasic and Walcher, 2012). Those PIP functions are important to analyze. To reduce time and economic cost, a computational identification method of PIPs is needed before experimental verification. There are only a few computational methods developed for PIP identification, e.g., ProInflam (Gupta et al., 2016) and PIP-EL (Manavalan et al., 2018c). In 2016, Gupta et al. firstly introduced a computation method named ProInflam that employed a SVM classifier with different sequence-based features (Gupta et al., 2016). Manavalan et al. developed another computation method named PIP-EL by using several sequence features (Manavalan et al., 2018c). Recently, Khatun et al. develop ProIn-Fuse by fusing multiple feature representations. Existing methods provide good prediction results, but their prediction performances are yet not fully satisfactory and there is still room for further improvement.

1.7 Research progress of anti-tubercular peptides

Tuberculosis (TB) is regulated by *Mycobacterium tuberculosis* (Mtb), which is a type of immune infective disease, being responsible as a major threat for human beings (Hamilton et al., 2015; WHO, 2017b; Zumla et al., 2015). (AlMatar et al., 2018; Jhamb et al., 2014). Many large-scale experimental screenings were carried to explore anti-TB peptides (Padhi et al., 2014; Yount and Yeaman, 2004). Many experimental candidates of anti-TB peptides were found and registered in the AntiTbPdb database (Usmani et al., 2018b). Notwithstanding the increasing number of experimentally validated anti-TB peptides, the mechanisms by which anti-TB peptides affect TB remain largely unknown (Gao et al., 2015; Gavrish et al., 2014; Nikonenko et al., 2004; Usmani et al., 2018b). Since the large-scale experimental identification of anti-TB peptides is laborious and time-consuming, alternative, computational methodologies are required that provide an accurate and robust prediction of anti-TB peptides. Recently, Usmani et al. developed the AntiTBpred, a computational predictor implementing a support vector machine (SVM) classifier (Usmani et al., 2018a). They illustrated that the composition of amino acids and N5C5 binary profiles (i.e., five amino acid residues from the N- and C-terminals) contribute to the enhanced prediction accuracy. Khatun et al. develop iAntiTB by Integrating the Amino Acid Patterns and Properties.

1.8 Research progress of linear B-cell peptides

B-cell peptides or epitope (BCEs) are specific regions of immunoglobulin molecules that can stimulate the immune system, which contributes to a diagnostic test, antibody production, and vaccine design (El-Manzalawy et al., 2008; Tomar and De, 2010; Yang and Yu, 2009). B cells are activated by BCEs to perform a variety of biological functions (Groell et al., 2018; Tomar and De, 2010). Linear BCEs have vast applications in the area of vaccine design, immunodiagnostic test, antibody production, as well as disease diagnosis and therapy (Bryson et al., 2010; Steere et al., 2011; Sweredoski and Baldi, 2009; Wang et al., 2018). Given experimental identification of BCEs is labor-intensive and costly, computational identification of BCEs has gained remarkable interest recently (Balachandran Manavalan1 and Lee, 2018; Gupta et al., 2013; Jespersen et al., 2017; Saha and Raghava, 2006; Wang and Pai, 2014). Several computational approaches have been developed to predict BCEs, which can be categorized into local and global predictors. Local predictors, such as BepiPred (Jespersen et al., 2017), Bcepred (Saha and Raghava, 2007), and COBEpro (Sweredoski and Baldi, 2009), explore some potential BCE encoding sequences from given protein sequences. These local methods aim to identify the regions or stretches of proteins that form BCEs [31], but it is difficult to specify the exact regions. Global predictors, such as iBCE-EL (Balachandran Manavalan1 and Lee, 2018), IgPred (Gupta et al., 2013), ABCpred (Saha and Raghava, 2006), SVMTriP (Yao et al., 2012), and LBtope (Singh et al., 2013), determine whether a given sequence is a BCE or not. Since the number of BCEs has rapidly increased in the immune epitope database (Vita et al., 2018), global methods gain attention as the classifier of BCEs.

Two global methods, LBtope, and iBCE-EL, have recently been developed and publicly available (Balachandran Manavalan1 and Lee, 2018; Singh et al., 2013). These two predictors exclusively investigated primary sequence-based features, such as amino acid composition, binary properties, and physicochemical properties, but did not consider any evolutionary information. Therefore, advanced analytic tools for identifying linear BCEs are still desirable.

1.9 Article description

1.9.1 Development of anti-inflammatory peptides prediction tool

In this thesis, at first, the author develops a bioinformatics tool termed as PreAIP (Predictor of Anti-Inflammatory Peptides) by integrating multiple complementary features. We systematically investigated different types of features including primary sequence, evolutionary and structural information through a random forest classifier. A peptide was considered as an anti-inflammatory (positive sample) if the anti-inflammatory cytokines of peptides induce any one of IL-10, IL-4, IL-13, IL-22, TGFb, and IFN-a/b in T-cell analyses of mouse and human (Jin et al., 2014; Marie et al., 1996). The final PreAIP model achieved an AUC value of 0.833 in the training dataset via 10-fold cross-validation test, which was better than that of existing models.

1.9.2 Development of pro-inflammatory peptides prediction tool

Second, the author develops a novel bioinformatics tool termed ProIn-Fuse, for predicting a pro-inflammatory by using multiple feature representation. The ProIn-Fuse predictor is capable of yielding a high accuracy. Specifically, a feature representation learning model was utilized to generate a set of informative probabilistic features by making the use of random forest models with eight sequence encoding schemes. Then the ProIn-Fuse was constructed by the linearly combined models of the informative probabilistic features. The generalization capability of our proposed method evaluated through independent tests showed that ProIn-Fuse yielded an accuracy of 0.746, which was over 10% higher than those obtained by the state-of-the-art PIP predictors.

1.9.3 Development of anti-tuberculosis peptides prediction tool

Third, the author develops an effective computational predictor iAntiTB (Identification of anti-tubercular Peptides) by the integration of multiple feature vectors deriving from the amino acid sequences via RF and SVM classifiers. The iAntiTB combined the RF and SVM scores via linear regression to enhance the prediction accuracy. To make a robust and accurate predictor we prepared the two datasets with different types of negative samples. The iAntiTB achieved AUC values of 0.896 and 0.946 on the training datasets of the first and second datasets, respectively. The ProIn-Fuse was established by fusing the successive probabilistic scores using a linear regression model.

1.9.4 Development of linear B-cell epitope prediction tool

Fourth, the authors develop a novel predictor, Identification of B-Cell Epitope (iLBE), by integrating evolutionary and sequence-based features for prediction. The successive feature vectors were optimized by a Wilcoxon rank-sum test. Then the random forest (RF) algorithm using the optimal consecutive feature vectors was applied to predict linear B-cell peptides. We combined the RF scores by the logistic regression to enhance the prediction accuracy. iLBE yielded an area under curve (AUC) score of 0.809 on the training dataset and outperformed other prediction models on a comprehensive independent dataset. iLBE is a powerful computational tool to identify the linear B-cell peptides and would help to develop penetrating diagnostic tests.

1.10 Introduction of different sections

In the second, third, fourth, and fifth chapters, the author will report the detailed procedures about the anti-inflammatory, proinflammatory, anti-tuberculosis, linear B-cell peptides prediction approaches, including data collection procedure, feature encoding, feature optimization protocol, model training, performance comparisons, and web servers. Finally, in the sixth chapter, conclusions of this thesis and future research perspectives will also be summarized and discussed, respectively.

CHAPTER 2

CHAPTER 2 PREDICTION OF ANTI-INFLAMMATORY PEPTIDES BY INTEGRATING MULTIPLE COMPLEMENTARY FEATURES

2.1 Introduction

Inflammation responses occur under the normal conditions when tissues are damaged by bacteria, toxins, trauma, heat, or any other reason (Ferrero-Miliani et al., 2007). These responses cause chronic autoimmune and inflammation disorders, including neurodegenerative disease, asthma, psoriasis, cancer, rheumatoid arthritis, diabetes, and multiple sclerosis (Hernandez-Florez and Valor, 2016; Patterson et al., 2014; Steinman et al., 2012; Tabas and Glass, 2013; Zouki et al., 2000). Numerous inflammation mechanisms are crucial for the upkeep of the state of tolerance (Corrigan et al., 2015; Miele et al., 1988). Numerous endogenous peptides recognized through inflammatory reactions function as anti-inflammatory agents can be employed by new therapies for autoimmune and inflammatory illnesses (Delgado and Ganea, 2008; Gonzalez-Rey et al., 2007). The immunotherapeutic aptitude of these anti-inflammatory peptides (AIPs) has various clinical applications such as generation of regulatory T cells and inhibition of antigen-specific T(H)1-driven responses (Delgado and Ganea, 2008). Moreover, certain synthetic AIPs act as effective therapeutic agents for autoimmune and inflammatory disorders (Zhao et al., 2016). For instance, chronic adenoidal direction of human amyloid- peptide causes an Alzheimer's disease. Mice models result in compact deposition of amyloid- peptides, which is a pathological marker of Alzheimer's disease, astrocytosis, microgliosis, and neuritic dystrophy in the brain (Boismenu et al., 2002; Gonzalez et al., 2005; Kempuraj et al., 2017). The present therapy for autoimmune and inflammatory disorders involves the use of non-specific anti-inflammatory drugs and other immunosuppressant's, which are frequently related to different side effects, such as initiation of a higher possibility of infectious diseases and ineffectiveness alongside inflammatory disorders (Tabas and Glass, 2013).

Notwithstanding the increasing number of experimentally examined AIPs *in vivo*, the molecular mechanism of AIP specificity remains largely unknown. On the other hand, large-scale experimental analysis of AIPs is time-consuming, laborious, and expensive. An alternative, computational approach that provides an accurate and reliable prediction of AIPs is required to complement the experimental efforts and to access the prompt identification of potential AIPs prior to their synthesis. To date, two *in silico* methods have been proposed to predict AIPs (Gupta et al., 2017; Manavalan et al., 2018b). In 2017 Gupta et al. employed hybrid features with a support vector machine (SVM) classifier to develop the AntiInflam predictor (Gupta et al., 2017). Manavalan et al. developed the AlPpred predictor by using the primary sequence encoding features with a random forest (RF) classifier (Manavalan et al., 2018b). These two methods used the primary sequence feature information without considering any evolutionary or structural features.

Nonetheless, the performance of the abovementioned existing predictors is not sufficient and remains to be improved. In this study, we have developed an accurate predictor named PreAIP (Predictor of Anti-Inflammatory Peptides) by integrating multiple complementary. We investigated different types sequence features including the primary sequence, evolutionary, and structural through a RF classifier. The PreAIP achieved higher performance on both the training and test datasets than the existing methods. In addition, we obtained valuable insights into the essential sequence patterns of AIPs.

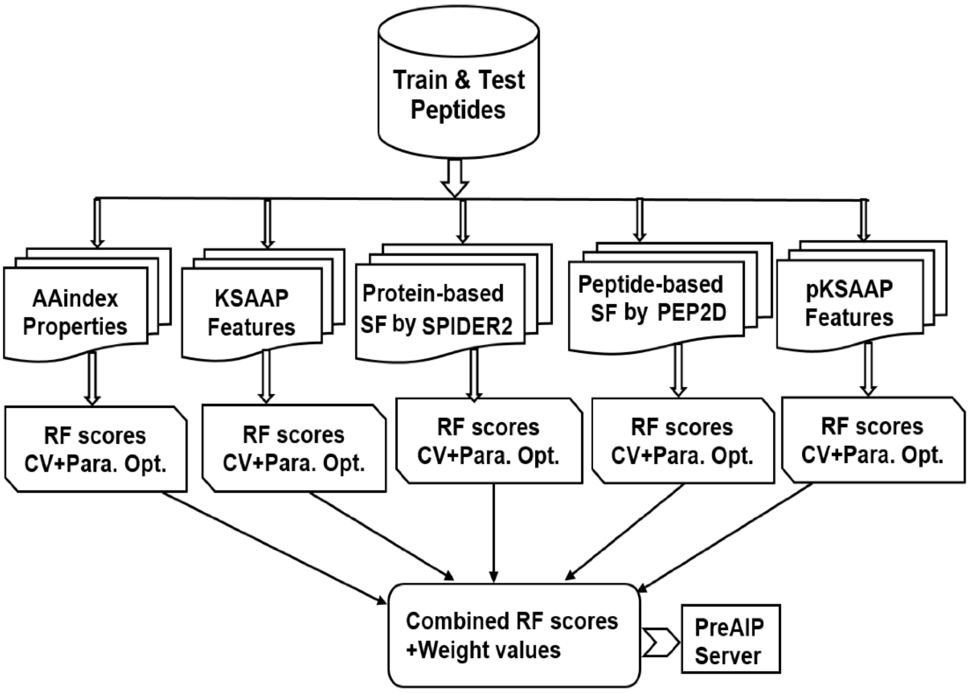

Figure 2-1 Computational framework of PreAIP.

2.2 Materials and methods

2.2.1 Dataset collection

To construct the PreAIP, we collected training and test datasets from a recently published article of the AIPpred (Manavalan et al., 2018b) and the IEDB database (Vita et al., 2018). A peptide was considered as anti-inflammatory (positive sample) if the anti-inflammatory cytokines of peptides induce any one of IL-10, IL-4, IL-13, IL-22, TGFb, and IFN-a/b in T-cell analyses of mouse and human (Jin et al., 2014; Marie et al., 1996). Meanwhile, the linear peptides for anti-inflammatory cytokines were considered non-AIPs (i.e., negative samples). To solve the overfitting problem of the prediction model, CD-HIT was employed with a sequence identity threshold of 0.8 (Huang et al., 2010). After eliminating redundant peptides, the same training and test samples were retrieved from the AIPpred predictor (Manavalan et al., 2018b). More reliable performance would be achieved by using a more stringent criterion of 0.3 or 0.4, as executed in (Hasan et al., 2017a; Hasan et al., 2016). However, this study did not use such a stringent criterion, because the length of the currently available AIPs is between 4 and 25. If we apply a stringent criterion of less than 0.8, the number of the available AIPs is greatly reduced so that we cannot retrieve the datasets employed by the previous predictor (Manavalan et al., 2018b). The collected training dataset results in 1,258 positive and 1,887 negative samples, and the test dataset contains 420 positive and 629 negative samples. All of curated datasets are included in our web server.

2.2.2 Computational framework

An overall computational framework of the proposed PreAIP is shown in **Figure 2-1**. After collecting the positive and negative AIPs from the AIPpred server (Manavalan et al., 2018b), their sequence datasets were transformed into the primary sequence, evolutionary and structural features. We considered polypeptides with 1 to 25 natural amino acids. When the peptide contains less than 25 residues, our scheme provides gaps (-) to the missing residues to compensate a peptide length of 25. To encode the primary sequence features, we employed two encoding methods of the composition of *k*-spaced amino acid pairs (KSAAP) and AAindex properties. An evolutionary feature was encoded by using the position specific encoding matrix, i.e., profile-based composition *k*-space of amino acid pair (pKSAAP). The structural feature (SF) was encoded by using SPIDER2 (Yang et al., 2017) and PEP2D (http://crdd.osdd.net/raghava/pep2d/) bioinformatics tools. The resulting five types of descriptors were independently put into RF models to produce five consecutive, independent RF prediction scores. Those RF scores were linearly combined using the weight coefficients to obtain the final prediction score. A web server was developed to implement the PreAIP.

2.2.3 Feature encoding

The PreAIP was constructed based on a binary classification problem (positive AIPs and negative-AIPs) through RF algorithms. The extraction of a set of relevant features is a crucial step to present a classifier. To keep the generated feature vectors, a high-quality peptide encoding method is necessary. As a substitute of the simple binary representation, we adopted five types of complicated feature encoding methods: AAindex, KSAAP, SPIDER2, PEP2D and pKSAAP, which are briefly described in the following subsections.

Table 2-1. Eight types of high index (HI) of AAindex properties used in this study.

AAindex ID	Index name	Properties Describtion

MIYS990104	HI1	Optimized relative partition energies
BLAM930101	HI2	Alpha helix propensity of position 44 in T4 lysozyme
BIOV880101	HI3	Information value for accessibility
MAXF760101	HI4	Alpha and turn propensities
TSAJ990101	HI5	Volumes including the crystallographic waters using standard radii and volumes.
NAKH920108	HI6	Amino acid composition of MEM of multi-spanning proteins
CEDJ970104	HI7	Amino acid composition and cellular location in proteins.
LIFS790101	HI8	Conformational preference for all beta-strands

2.2.4 Amino acid index properties

Numerical physicochemical properties of amino acids exist in the AAindex database (version 9.1) (Kawashima et al., 2008). After assessing different types of AAindex indices, we selected 8 types of high indices (HI) and ordered them from HI1 to HI8 (**Table 2-1**). In a peptide sequence with length *L*, a (*L*×20) feature vector was generated through the AAindex encoding.

2.2.5 KSAAP encoding

The KSAAP encoding descriptor is widely used in bioinformatics research (Hasan et al., 2018d; Md Mehedi Hasan, 2017). The procedure of KSAAP is briefly described as follows. Peptide sequences contain (20×20) types of amino acid pairs (i.e. AA, AC, AD, … , YY)$_{400}$ for every single *k*, where *k* denotes the space between two amino acids. The optimal k_{max} was set to 0-4 to generate (20×20×5) =2,000 dimensional feature vectors for each corresponding peptide sequence. Details of the KSAAP encoding method are described elsewhere (Hasan et al., 2015).

2.2.6 Structural features

Protein-based SF

The protein-based SF features are generated by the SPIDER2 software that is widely used in bioinformatics research (Lopez et al., 2018; Yang et al., 2017). Three types of features were generated by SPIDER2: accessible surface area (ASA), backbone torsion angles (BTA) and secondary structure (SS). The BTA generated 4-type feature vectors of phi, psi, theta and tau. The SS generated 3-type feature vectors of helix, strand and coil. Totally, 8-type feature vectors were generated SPIDER2. For each peptide sequence, (*L*×8) dimensional feature vectors were generated, where *L* was the length of a given AIP.

Peptide-based SF

We employed PEP2D to generate a peptide structure prediction feature (http://crdd.osdd.net/raghava/pep2d/). The PEP2D generated three types of probability scores: Helix Prob, Sheet Prob, and Coil Prob. For each peptide sequence, (*L*×3) dimensional feature vectors were generated, where *L* was the length of a given AIP.

2.2.7 pKSAAP encoding

In protein or peptide sequence analysis, the PSSM provides useful evolutionary information. This matrix measures the replacement probability of each residue in a protein with all the residues of the genomic code. The PSSM profile was created by using PSI-BLAST (version of 2.2.26+) against the whole Swiss-Prot NR90 database (version of December 2010) with two default parameters, an e-value cutoff of 1.0×10^{-4} and an iteration number of 3 (Hasan et al., 2015). Then, we extracted the feature vectors using the given peptide sequences. After generating the PSSM profile, we generated possible *k*-space pair composition from the PSSM, i.e., pKSAAP, in the same manner as the previous study of protein pupylation site prediction (Hasan et al., 2015). When an optimal *k*-space was between 0 and 4, a (5× 20×20 = 2000) dimensional feature vector was generated.

2.2.8 Feature selection

To find the top ranking features for predicting AIPs, a well-established, supervised method for feature dimensionality reduction, Information Gain (IG) (Thanamani, 2013), was used through a WEKA package (Frank et al., 2004)). A large value of the IG indicates that the corresponding residues have a great impact on prediction performance. The IG processes the decrease in entropy when given information is used to group values of an alternative (class) feature. The entropy of feature *U* is defined as

$$H(U) = -\sum_i P(u_i) log_2(P(u_i)) \qquad (2\text{-}1)$$

where u_i is a set of values of *U* and *P* (u_i) is the prior probability of u_i. Conditional entropy *H*(*U/V*), given another feature *V,* is defined as

$$H(U|V) = -\sum_j P(v_j) \sum_i P(u_i|v_j) log_2(P(u_i|v_j)) \qquad (2\text{-}2)$$

where *P* (u_i | v_j) is the posterior probability of *U* given by the value v_j of *V*. The *IG* is defined as the decreased entropy calculated by subtracting the conditional entropy of *U* given by *V* from the entropy of *U*, as follows.

$$IG(U|V) = H(U) - H(U|V) \qquad (2\text{-}3)$$

2.2.9 Machine learning

The RF is a supervised machine learning algorithm (Breiman, 2001) and is widely used for various biological problems (Bhadra et al., 2018; Hasan MM, 2018; Manavalan et al., 2017; Manavalan et al., 2018b). In brief, the following steps are carried to construct *n* trees of the RF model. Initially, to obtain a new dataset, *N* samples are obtained from the training set by random selection with replacement procedures. To get *n* different datasets this procedure is repeated *n* times and *n* decision trees are built based on the *n* datasets. In this assembling process, for *K* input features, *k* (*k*<<*K*) features are selected randomly, where *k* is the constant during construction of the RF. To split the node, a *gini* impurity criterion is used from the given features. To grow completely, each decision tree is grown without pruning. Afterward getting *n* decision trees, the class with the most votes is the final prediction (Breiman, 2001). An R package was implemented to train the proposed model (https://cran.r-project.org/web/packages/randomForest/). We set *n* to 1000 through the 10-fold cross-validation (CV) test, which is large enough to gain stable prediction.

The performance of the RF was characterized in comparison to three commonly used machine learning algorithms: Naive Bayes (NB) (Lowd, 2005), SVM (Hearst, 1998), and artificial neural network (ANN) (R. S. Michalski 2013). We used the NB and ANN algorithms of the WEKA software (Frank et al., 2004) and the SVM algorithm with a kernel radial basis function (RBF) of the LIBSVM package (http://www.csie.ntu.edu.tw/Bcjlin/libsvm/). In the NB algorithm, we set batch size to 1000 through the 10-fold CV via the WEKA software. For the

ANN algorithm, we considered “MultilayerPerceptron –L 0.3 –M 0.2 –N 500 –V 0 S 0 –E 20 –H a” via the WEKA software. To optimize the parameters of the SVM model, the cost and gamma functions were set to 8 and 0.03125 for KSAAP, respectively, via the LIBSVM package. Similarly, the cost and gamma functions were set to 2 and 0.0123 for AAindex, 32 and 0.0625 for pKSAAP, 16 and 0.125 for SPIDER2, and 8 and 0.015625 for PEP2D.

2.2.10 Combined method

To make an efficient and robust prediction model, optimization of incorporative feature methods is generally essential. We linearly combined the RF scores of the five encoding methods: AAindex, KSAAP, SPIDER2, PEP2D and pKSAAP, using the following formula (Hasan et al., 2017b):

$$\text{Combined} = w_1 \times \text{SPIDER2} + w_2 \times \text{PEP2D} + w_3 \times \text{KSAAP} + w_4 \times \text{AAindex} + w_5 \times \text{pKSAAP} \quad (2-4)$$

where w_1, w_2, w_3, w_4, and w_5 are the weight coefficients indicating the strength of the five descriptors; the sum of w_1, w_2, w_3, w_4, and w_5 is 1. We adjusted each weight from 0 to 1 with an interval of 0.05. When w_1, w_2, w_3, w_4, and w_5 were 0.00, 0.00, 0.15, 0.25, and 0.6, respectively, the AUC value on the CV of training dataset was maximal. Therefore, the linear combination of the three successive RF models of KSAAP, AAindex, and pKSAAP was actually “Combined”.

2.2.11 Performance evaluation matrix

To investigate the performance of the PreAIP, the threshold-dependent and threshold-independent indices were measured. Using the threshold-dependent indices, four widely used statistical measures denoted as accuracy (Ac) specificity (Sp), sensitivity (Sn), and Matthews correlation coefficient (MCC), respectively, were considered. The four outcomes are presented in the following formulas,

$$\text{Ac} = \frac{n(\text{TP}) + n(\text{TN})}{n(\text{TP}) + n(\text{FP}) + n(\text{TN}) + n(\text{FN})} \qquad (2\text{-}5)$$

$$\text{Sn} = \frac{n(\text{TP})}{n(\text{TP}) + n(\text{FN})} \qquad (2\text{-}6)$$

$$\mathrm{Sp}=\frac{\mathrm{n(TN)}}{\mathrm{n(TN)}+\mathrm{n(FP)}} \qquad (7)$$

$$\mathrm{MCC}=\frac{\mathrm{n(TP)}\times\mathrm{n(TN)}-\mathrm{n(FP)}\times\mathrm{n(FN)}}{\sqrt{[\mathrm{n(TN)}+\mathrm{n(FN)}][\mathrm{n}(\mathrm{TP})+\mathrm{n(FP)}][\mathrm{n}(\mathrm{TN})+\mathrm{n(FP)}][\mathrm{n(TN)}+\mathrm{n(FP)}]}} \qquad (2\text{-}8)$$

where n(TP) exemplifies the number of correctly predicted positive samples; n(TN) the number of correctly predicted negative samples; n(FP) the number of incorrectly predicted positive samples, and n(FN) the number of incorrectly predicted negative samples. Furthermore, we used the receiver operating characteristics (ROC) curve (Sn vs. 1-Sp plot) to evaluate the area under the ROC curve (AUC) of the threshold-independent parameter (Centor, 1991; Gribskov and Robinson, 1996).

Since the balance between the correctly predicted AIPs and non-AIPs is critically responsible for accurate prediction, Sp and Sn are intuitive, intelligible measures. Typically, high Sp decreases Sn. In this study, the prediction performance of the PreAIP for the training dataset was evaluated with a stepwise change in Sp. We calculated Sn, Ac and MCC at high (0.903), moderate (0.801) and low (0.709) levels of Sp. These three levels of Sp were given by setting the high (0.468), moderate (0.388) and low (0.342) thresholds of the RF score. In the same manner, we measured the performance of the individual encoding scheme of KSAAP, AAindex, SPIDER2, PEP2D, and pKSAAP at each level of Sp. When the same threshold values of the RF score were applied to prediction of the test dataset, the high, moderate and low levels of Sp were calculated as 0.871, 0.747, and 0.636, respectively.

To assess the performance of the PreAIP using the measures of Ac, Sp, Sn, MCC, and AUC, a 10-fold CV test was used. For the 10-fold CV, original training samples were randomly and equally picked up into 10 subclasses. Among 10 subclasses, one subclass was singled out as the test sample, and the remaining 9 subclasses were considered as the training sample. Then we computed all performance measures for each predictor. We repeated this procedure 10 times by changing the training and test samples. Eventually, we calculated the average value of each performance measure for each predictor.

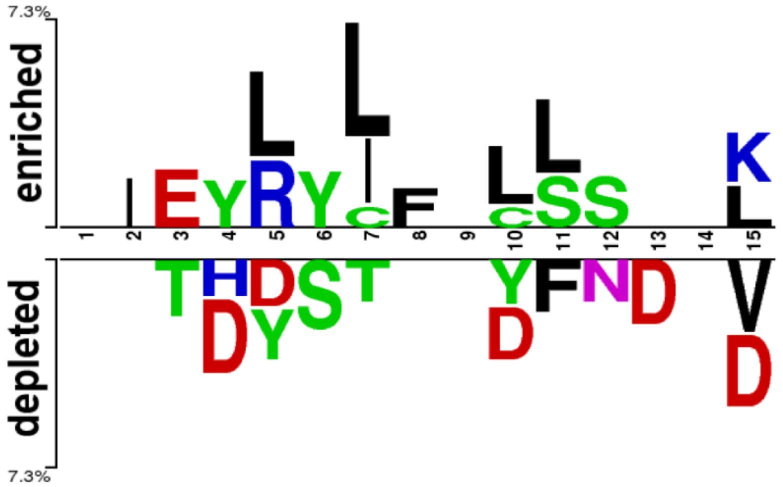

Figure 2-2. Sequence logo representation of positive and negative AIPs. The upper portion (enriched) is represented by positive AIPs, while lower portion (depleted) negative AIPs. The statistically significant local sequence within the N-terminal 15-residues of AIPs was plotted with $p < 0.05$ by Welch's *t*-test.

2.3 Results and discussion

2.3.1 Sequence preference analysis of AIPs

To investigate the amino acid preference of positive and negative AIPs, we performed sequence compositional preference analysis using the amino acids from the 1 to 15 N-terminal residues of training sets. The length of the AIPs ranged between 4 and 25 amino acid residues in this study. The average length of AIPs was 15 amino acids. Since Ialenti et al. suggested that the AIP activity is located in the N-terminal region of the molecule (Ialenti et al., 2001), we investigated the 1 to 15 N-terminal amino acids by the sequence compositional preference analysis. A non-existing residue was coded by "O" to fill the corresponding position of the AIPs.

At first, we submitted the 1 to 15 N-terminal amino acids of positive and negative AIPs to the sample logo online server (http://www.twosamplelogo.org/) to generate the sequence logo representations (**Figure 2-2**). The height for each amino acid was in proportion to the percentage of positive (over-represented) or negative (under-represented) peptides. The logos were scaled according to their statistical significance threshold of $p < 0.05$ by Welch's *t*-test. Leucine (L) at positions 5, 7, 10, 11, and 15, cysteine (C) at position 7 and 10, isoleucine (I) at positions 2 and 7, arginine (R) at position 5, phenylalanine (F) at position 8, and lysine (K) at position 15 were significantly overrepresented compared with other amino acids, while aspartic acid (D) at positions 4, 5, 10, 13 and 15, threonine (T) at positions 3 and 7, valine (V) at position15 were significantly underrepresented. In addition, tyrosine (Y) at positions 4 and 5 was overrepresented, while Y at positions 5 and 10 underrepresented. These results suggested that positive and negative AIPs are significantly different.

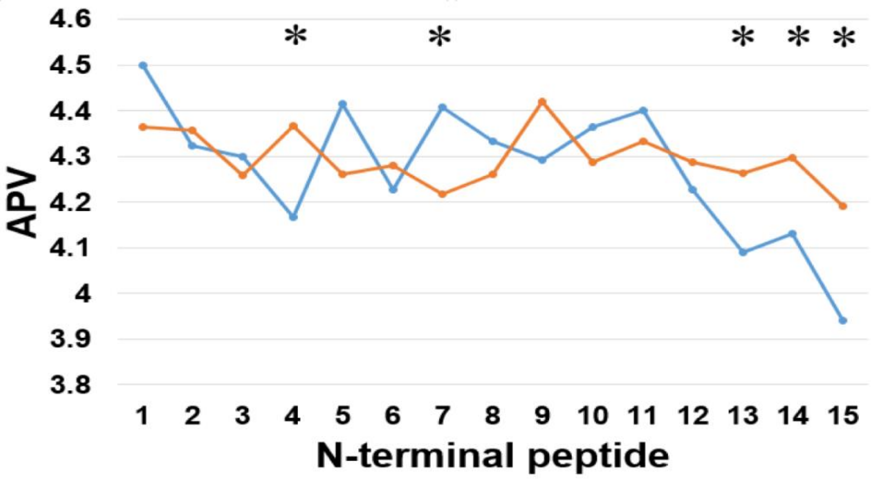

Figure 2-3. Comparison of evolutionary information of positive and negative AIPs. Blue lines represent the positive AIP, while orange lines the negative AIPs. "*" represents that the APV is statistically different between both the AIPs, with $p < 0.05$ by the KW test.

Table 2-2 Statistical difference in the APVs between the positive and negative AIPs. The *p*-values were calculated using the KW test and corrected by the Bonferroni test. '*' represents *p*-values < 0.05.

N-terminal positive	*p*-value
1	5.41E-01
2	1.00
3	1.00
4	1.01E-02*
5	9.64E-01
6	1.00
7	4.64E-02*
8	1.00
9	1.00

10	1.00
11	1.00
12	1.00
13	3.64E-02*
14	2.11E-02*
15	1.79E-02*

Secondly, we examined the evolutionary conservation features of the PreAIP using the average PSSM value (APV) for each amino acid within 1 to15 N-terminal amino acids of AIPs. The evolutionary conservation information of APV of both the positive and negative AIPs is illustrated in **Figure 2-3**. Some of amino acid positions of positive and negative AIPs showed significantly different scores. Furthermore, a nonparametric Kruskal-Wallis (KW) test was used to examine whether positive and negative AIPs were significantly dissimilar. The *p*-values were calculated and corrected by the Bonferroni test (**Table 2-2**).

Thirdly, we examined the AAindex encoding features of PreAIP. Eight types of informative amino acid indices were used and named HI1 to HI8 as the input feature vectors from the AAindex database. We examined these HI amino acid properties of both the positive and negative AIPs. As illustrated in **Figure 2-4**, the average values of the eight indices were renamed as AVHI1 to AVHI8. These indices represented the amino acid compositions of intracellular proteins. Some of the AIPs had distinct amino acid compositions in the eight high-quality amino acid indices between two samples of AIPs (**Figure 2-4**). The KW test was used to examine whether two samples of AIPs were significantly dissimilar with respect to the eight HI properties. The *p*-values were calculated and corrected by the Bonferroni test (**Table 2-3**). Significantly different AAindex values with *p*-value <0.05 appeared at some positions of AIPs, as marked with ‘*’ in **Figure 2-4**.

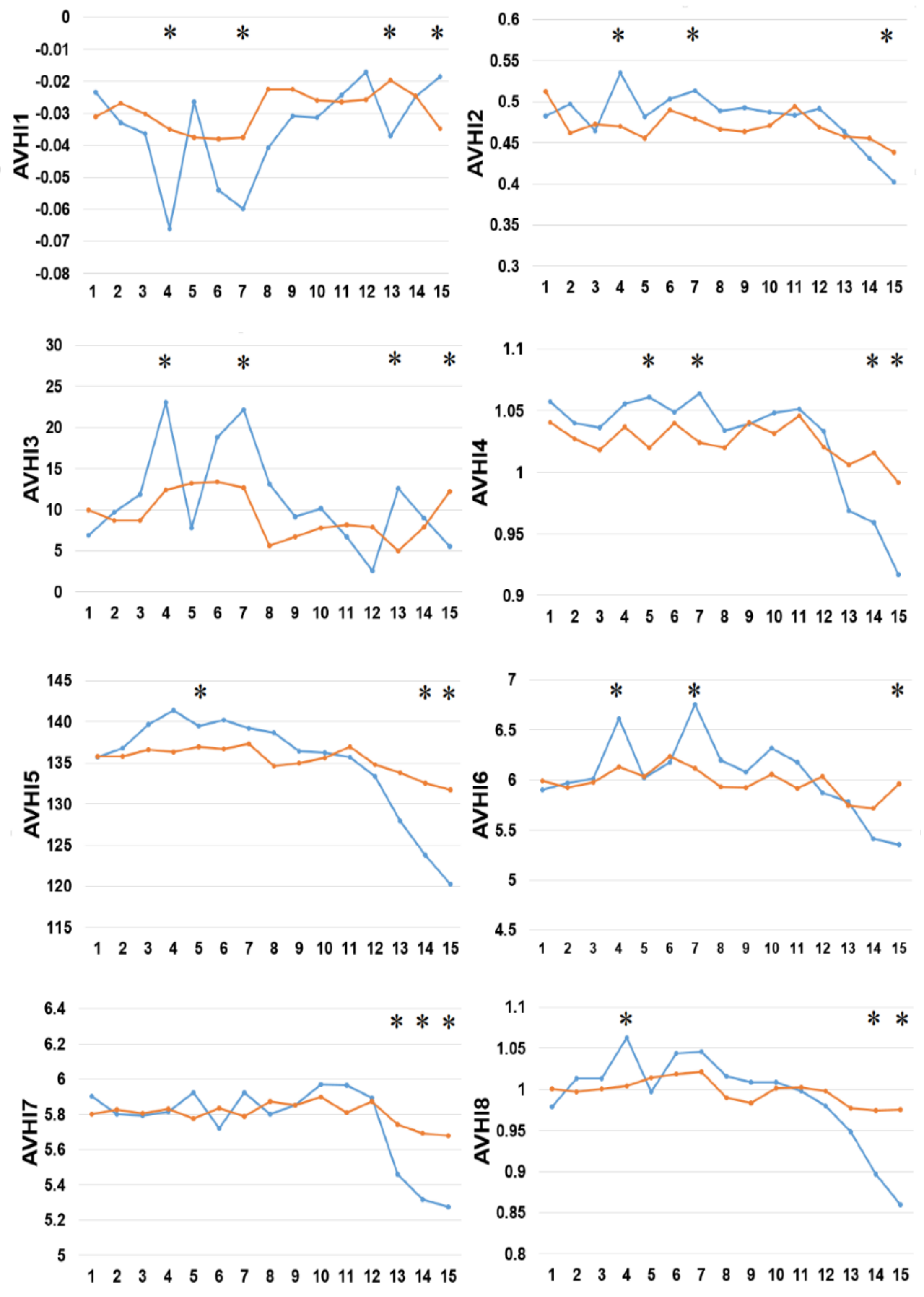

Figure 2-4. Comparison of eight high-quality amino acid indices between two samples of AIPs. The eight high-quality amino acid indices from HI1 to HI8 are placed at the centers of eight amino acid index clusters, which indicate high residue propensities of AAindex. The row represents the N-terminal peptide, while the blue lines signify the positive AIP and the orange lines the negative AIPs. "*" represents that the amino acid indices are statistically different between both the samples with $p < 0.05$ by the KW test.

Table 2-3 Statistical difference in the high index of AAindex properties between the positive and negative AIPs. The *p*-values were calculated using the KW test and corrected by the Bonferroni test. ‘*’ represents *p*-values < 0.05.

N-terminal positive	AVHI1	AVHI2	AVHI3	AVHI4	AVHI5	AVHI6	AVHI7	AVHI8
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	2.91E-02*	1.15E-03*	1.01E-02*	1.00	1.00	4.01E-02*	1.00	3.01E-02*
5	1.00	1.00	5.98E-01	2.69E-02*	3.64E-02*	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7	2.45E-02*	3.86E-02*	4.64E-02*	2.39E-02*	1.00	3.37E-02*	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
13	1.99E-02*	1.00	4.87E-02*	1.00	1.00	1.00	4.07E-02*	1.00
14	1.00	1.00	1.00	2.19E-02*	3.01E-02*	1.00	3.01E-02*	3.99E-02*
15	4.65E-02*	3.65E-02*	4.08E-02*	3.79E-02*	1.29E-02*	2.11E-02*	1.33E-02*	2.38E-02

Table 2-4 Statistical difference in the 8 types of SFs by SPIDER2 between the positive and negative AIPs. The *p*-values were calculated using the KW test and corrected by the Bonferroni test. ‘*’ represents *p* values < 0.05.

N-terminal positive	AAS	Phi	Psi	The	Tau	Coil	Stand	Helix
1	1.00	1.00	3.78E-01	1.00	7.78E-01	1.00	1.00	7.08E-01
2	1.00	1.00	8.67E-02	1.00	4.67E-02*	1.00	1.00	6.67E-01
3	1.00	1.00	1.69E-01	1.00	3.69E-02*	1.00	1.00	6.54E-02
4	1.00	1.00	1.91E-02*	1.00	3.11E-02*	1.00	1.00	4.14E-02*
5	1.00	1.00	4.98E-02*	1.00	5.98E-02	1.00	1.00	4.08E-02*
6	1.00	1.00	8.88E-02	1.00	1.00	1.00	1.00	1.00
7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
12	1.00	1.00	1.00	1.00	1.00	1.00	4.98E-02*	1.00
13	5.99E-02	3.75E-02*	2.87E-02*	2.87E-02*	3.67E-02*	1.00	6.56E-01	7.67E-02
14	4.40E-02*	3.13E-02*	1.00E-02*	4.13E-03*	2.97E-03*	1.00	1.00	4.97E-02*
15	3.78E-02*	2.39E-02*	1.08E-03*	1.39E-03*	1.65E-03*	4.53E-02*	1.00	1.99E-02*

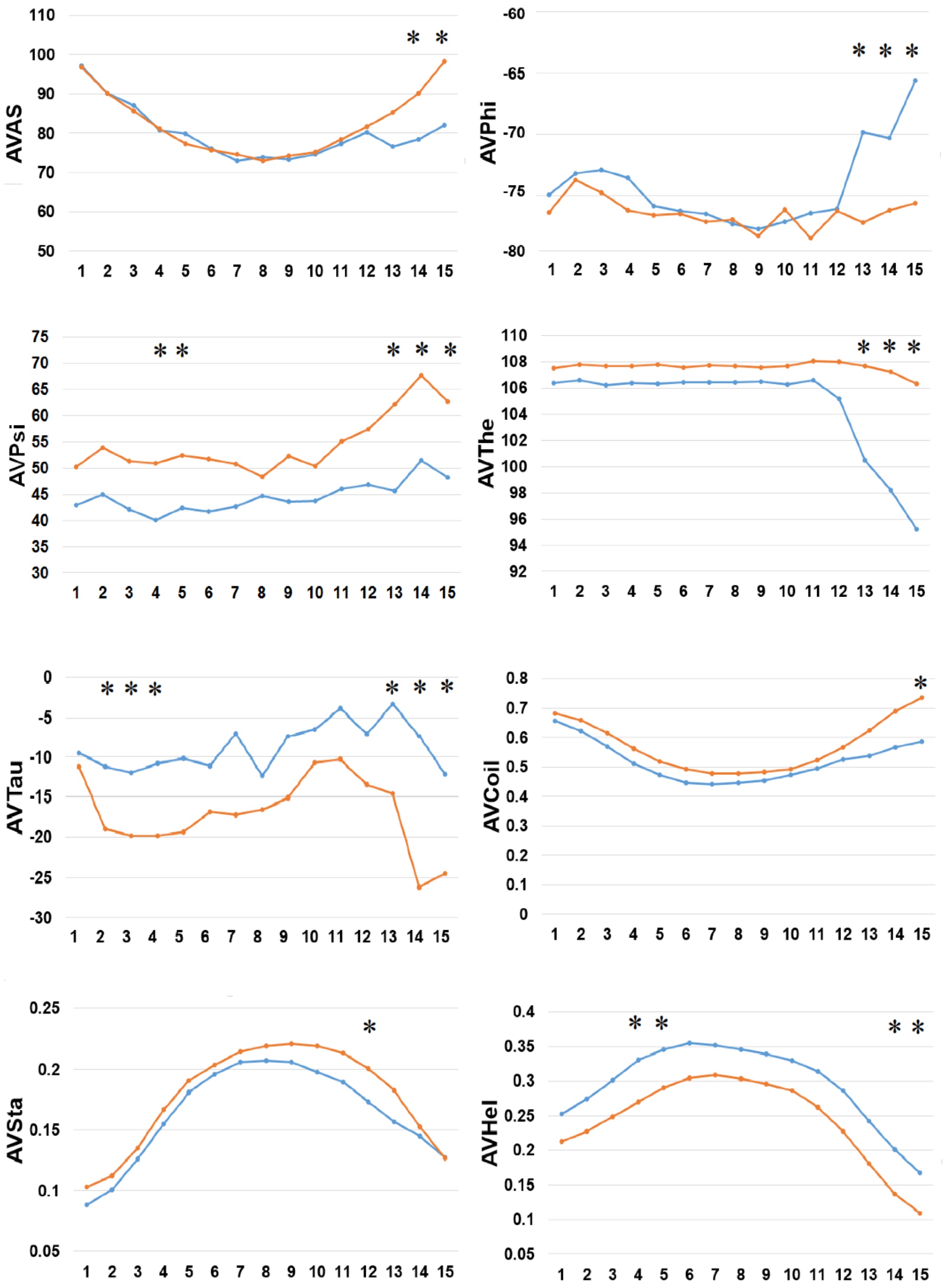

Figure 2-5. Comparison of 8 types of the SFs by SPIDER2 between positive and negative AIPs. The row represents the N-terminal peptide, while the blue lines signify the positive AIPs and the orange lines the negative AIPs. “*” represents that the SFs are statistically different between both the samples with $p < 0.05$ by the KW test.

Finally, we examined the difference in 8 types of SFs by SPIDER2 between the positive and negative AIPs, as shown in **Figure 2-5**. We calculated the average value of 8 types of SFs for SPIDER2: ASA, phi, psi, theta, tau, coil, stand, and helix of both the positive and negative AIPs. The average features were represented as AVAS, AVPhi, AVPsi, AVThe, AVTau, AVCoil, AVSta, and AVHel (**Figure 2**-5). We plotted these average values of SFs with respect to the 1 to 15 N-terminal AIPs. Distinguished differences were observed between the positive and negative samples of AIPs. The KW test was employed to examine whether two sample of AIPs were significantly dissimilar among the eight SFs. The *p*-values were calculated and corrected by the Bonferroni test (**Table 2-4**). Significantly different SFs were perceived at some positions of AIPs, with a *p*-value <0.05, as indicated with '*' in **Figure 2-5**.

The above analysis of residue preference between the positive and negative AIPs suggested that the combination of the primary sequence, evolutionary and structural amino acid occurrences achieves a precise prediction.

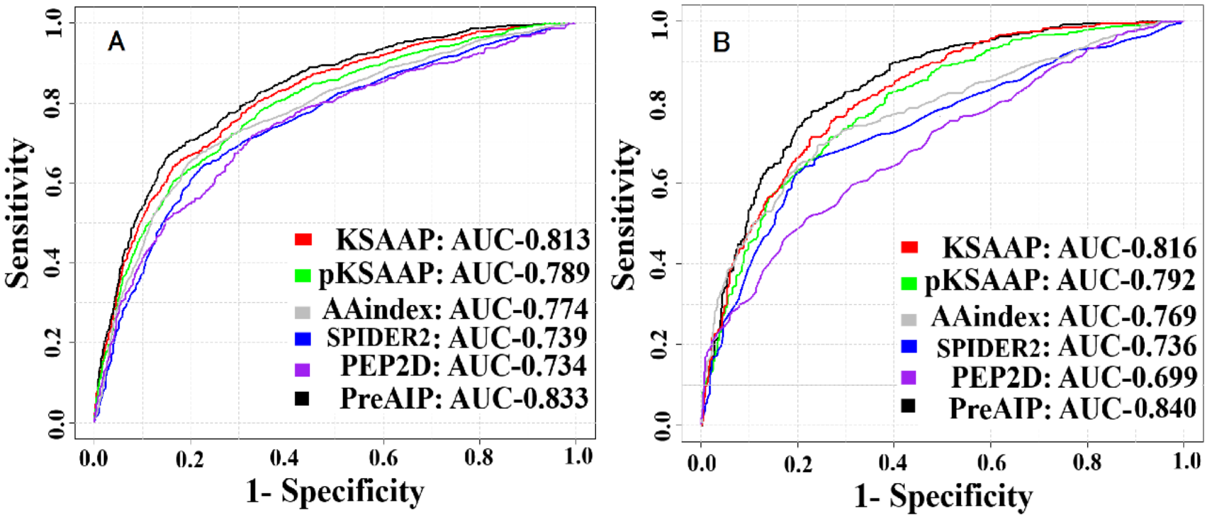

Figure 2-6. ROC curves of the various prediction models. **(A)** 10-fold CV test on a training dataset and **(B)** test dataset. The PreAIP combined the KSAAP, pKSAAP, and AAindex methods. High AUC values show accurate performance.

2.3.2 Overall prediction performance of PreAIP

The selected five descriptors (AAindex, KSAAP, SPIDER2, PEP2D, and pKSAAP) were separately used for prediction of AIPs. Optimization of multiple encoded features is generally essential in the training model to reduce dimensionality while retaining the significant feature. To achieve this, we performed multiple rounds of experiments to select appropriate feature vectors using the IG feature selection via 10-fold CV test on training set; however, it turned out that the IG feature selection did not improve prediction performance. Thus, the IG feature was used to collect significant features and for interpreting a superiority of KSAAP encoding.

Table 2-5. AUC values for prediction performance of the training dataset by 10-fold CV test

Methods	Sp	Sn	Ac	MCC	AUC	***p*-value**
pKSAAP	0.798	0.647	0.738	0.450	0.789	0.017
AAindex	0.795	0.644	0.735	0.448	0.774	0.012
SPIDER2	0.765	0.434	0.633	0.235	0.739	0.004
PEP2D	0.769	0.411	0.629	0.219	0.734	0.004
KSAAP	0.805	0.656	0.745	0.463	0.813	0.118
PreAIP*	0.806	0.709	0.767	0.508	0.833	

* PreAIP is the combined method of SPIDER2, PEP2D, KSAAP, AAindex and pKSAAP encoding schemes and their weight coefficients are 0.00, 0.00, 0.15, 0.25, and 0.6, respectively via RF scores. A *p*-value was computed based on the final model of AUC values by using a Wilcoxson matched-pair signed test.

We accessed the performances of the training model of five successive encoding methods of AAindex, KSAAP, SPIDER2, PEP2D, and pKSAAP through a 10-fold CV test using the RF classifier. The prediction results by each of five encoding features and the 'Combined features' are shown in **Figure 2-6A**. The AUCs of AAindex, KSAAP, SPIDER2, PEP2D, and pKSAAP were 0.774, 0.813, 0.739, 0.734, and 0.789, respectively. The KSAAP performed best for the 5 single encoding approaches in terms of Sn, MCC and AUC (**Table 2-5**). The "Combined features" (PreAIP) showed better performance with an AUC of 0.833 than any other single feature. It is noted that "Combined features" means a linear combination of the RF scores (Materials and Methods). Moreover, the PreAIP presented the highest AUC value (0.840) in the test dataset (**Figure 2-6B**). The performance of PreAIP was effective and reasonable for all the tested cases (**Figure 2-6**) and was best in the AIP prediction.

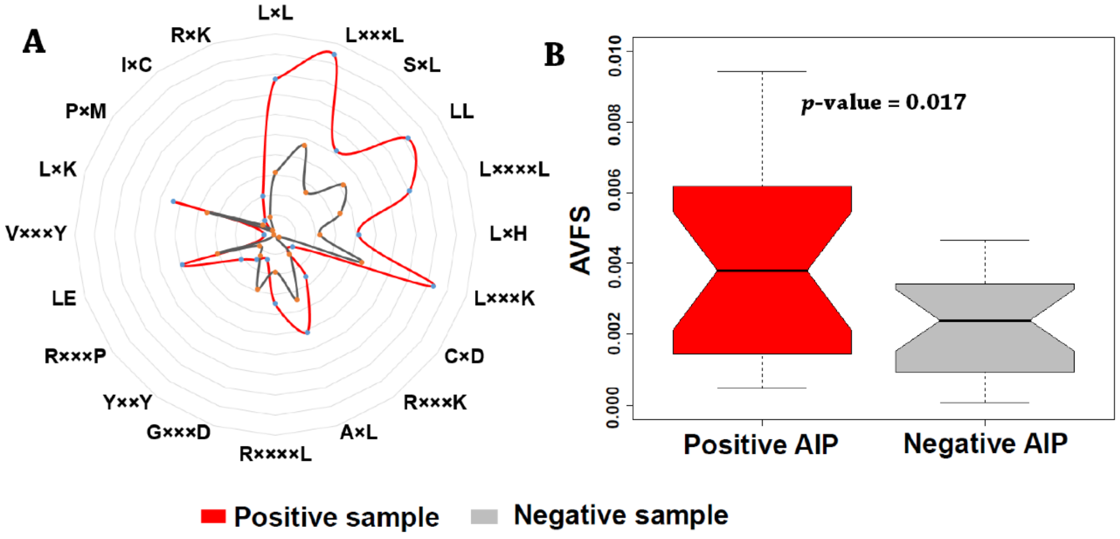

Figure 2-7 Top 20 amino acid pairs selected by the IG feature of the KSAAP method. **(A)** The radar diagram is represented by the composition of each amino acid pair whose length is proportional to the composition of KSAAP features. **(B)** Box plot shows the top 20 average value of feature scores (AVFS) by the IG. Red color denotes the positive AIPs, while gray color denotes the negative AIPs. The *p*-value is computed by two-sample *t*-test.

Table 2-6 Top 20 IG features of KSAAP encoding with corresponding amino acid pair positions.

Ser. No	IG features	Amino acid pairs
1	0.01145	L×L
2	0.01077	L×××L
3	0.00909	S×L
4	0.00807	LL
5	0.00562	L××××L
6	0.00475	L×H
7	0.00446	L×××K
8	0.0044	C×D
9	0.0041	R×××K
10	0.00368	A×L
11	0.00368	R××××L
12	0.00356	G×××D
13	0.00346	Y××Y
14	0.00345	R×××P
15	0.00332	LE
16	0.00331	V×××Y
17	0.0033	L×K

18	0.0033	P×M
19	0.00328	I×C
20	0.00327	R×K

In addition, we found that KSAAP performed best for all the five single encoding methods. To investigate the most significant residue of the KSAAP method, the top 20 amino acid pairs of AIPs were examined through the IG feature selection. The top 20 significant residue pair scores and their corresponding positions are listed in **Table 2-6**. These significant features are also presented using a radar diagram (**Figure 2-7A**). For example, the feature sequence motif 'L×L', which is represented by 1-spaced residue pair of 'LL', is the most important residue pair, where '×' stands for any amino acid. The feature ‘L×××L’ represented the second enriched motif surrounding positive samples of AIPs. Similarly, the feature 'LL', which represents a 0-spaced residue pair of 'LL', is important and enriched in the negative samples AIPs. Similarly, to keep other *k*-space amino acid pairs from KSAAP, the same exemplification was employed. Residue preference analysis demonstrated that "L", "Y", "C", "D", and "I" residues frequently appear for AIPs (**Figures 2-2** and **2-7A**). These residues are expected to play a key role in the recognition of AIPs. To characterize the top 20 KSAAP-specific features, we compared the numbers of positive and negative AIPs. **Figure 2-7B** showed the top 20 average value of feature scores (AVFS) by the IG. The average of top 20 features was significantly different between two samples of AIPs with *p*-value <0.05, suggesting the effectiveness of the KSAAP encoding. The significant residue pair scores are listed in **Table 2-6**, which provides some insights into the sequence patterns of the AIPs. They deserve further experimental validation.

Table 2-7 Performance comparison with exiting predictors using test dataset

Predictor	Threshold	Sp	Sn	Ac	MCC	AUC	*p*-value
AntiInflam (LA)	-0.3	0.892	0.258	0.638	0.197	0.647	<0.001
AntiInflam (MA)	0.5	0.417	0.786	0.565	0.210	0.706	<0.001
AIPpred	Server	0.746	0.741	0.744	0.479	0.813	0.039
PreAIP	High	0.871	0.618	0.770	0.512	0.840	
	Moderate	0.747	0.784	0.762	0.522	0.840	
	Low	0.636	0.863	0.727	0.492	0.840	

A *p*-value was computed based on AUC values by using a Wilcoxson matched-pair signed test and *p*<0.05 indicates a statistically significant difference between the proposed PreAIP and each selected method. The performances of AntiInflam LA and MA methods were computed using default threshold (server) values of -0.3 and 0.5, respectively. The AIPpred threshold was the same as given by its server.

2.3.3 Comparison of PreAIP with existing predictors using test dataset

We evaluated the performances of PreAIP along with that of existing predictors on the test dataset. We submitted the test set to the AIPpred (Manavalan et al., 2018b) and AntiInflam (Gupta et al., 2017) servers to assess the performance. It is noted that AntiInflam server provides different thresholds values. We used two threshold values of -0.3 and 0.5 and renamed as less accurate (LA) and more accurate (MA) models (Gupta et al., 2017), respectively. The AIPpred represents the state-of-the-art predictor available. The average performances of the LA, MA, AIPpred and PreAIP are illustrated in the **Table 2-7**. The LA showed the highest Sp (0.892) with the lowest Sn (0.258), MCC (0.197) and AUC (0.647) for all the predictors. The PreAIP with the high threshold presented much higher Sn (0.618) Ac (0.770), MCC (0.512) and AUC (0.840) than LA, while it provided Sp (0.871) comparable to LA. The PreAIP with the low threshold showed the highest Sn (0.863), while keeping Sp, Ac, MCC and AUC at a high level. While the AIPpred presented considerably high values to all the measures of Sp, Sn, Ac, MCC and AUC, the PreAIP with the moderate threshold outperformed the AIPpred, presenting well-balanced, high prediction performances. The PreAIP performance improvement was found distinct on the test dataset by the Wilcoxson matched-pair signed test, demonstrating its ability to predict unseen peptides.

Table 2-8 Performance comparison of PreAIP with AIPpred using training dataset.

Methods	Threshold	Sp	Sn	Ac	MCC	AUC	*p*-value
AIPpred	Default	0.711	0.758	0.730	0.460	0.801	0.034
	High	0.903	0.632	0.795	0.566	0.833	
PreAIP	Moderate	0.801	0.719	0.768	0.520	0.833	
	Low	0.709	0.784	0.739	0.484	0.833	

A *p*-value was computed based on AUC values by using a Wilcoxson matched-pair signed test and $p<0.05$ indicates a statistically significant difference between the proposed PreAIP and AIPpred.

2.3.4 Comparison of PreAIP with AIPpred using training dataset

We compared the performance of the proposed PreAIP with the AIPpred using the same training dataset. In this study, the same dataset as the AIPpred set was used to make a fair comparison for prediction performance of AIPs. As shown in **Table 2-8**, the PreAIP achieved a better performance than the AIPpred in terms of Ac, Sp, Sn, MCC and AUC. The AUC value was nearly 3% higher than the AIPpred predictor. The PreAIP performance (AUC) improvement over the AIPpred was demonstrated on the training set by the Wilcoxson matched-pair signed test (**Table 2-8**).

Table 2-9 AUC values of AIP prediction by different machine learning algorithms based on a 10-fold CV test

Algorithms	SPIDER2	PEP2D	AAindex	KSAAP	pKSAAP	Combined
RF	0.739	0.734	0.774	0.813	0.789	0.833
NB	0.659	0.655	0.707	0.729	0.717	0.736
SVM	0.698	0.677	0.738	0.766	0.749	0.779
ANN	0.662	0.649	0.716	0.741	0.736	0.753

"Combined" indicates that the performance of the optimized combined features. The combined score of RF was given as the sum of the five SPIDER2, PEP2D, AAindex, KSAAP, and pKSAAP features with weight values of 0.00, 0.00, 0.15, 0.25, and 0.6 respectively. In the same way, the weight values of NB, SVM, and ANN were given as (0.00, 0.00, 0.10, 0.35, and 0.55), (0.00, 0.00, 0.22, 0.45, and 0.33), and (0.00, 0.00, 0.18, 0.5, and 0.32), respectively.

Table 2-10 AUC values with 60% peptide redundancy on the training dataset by 10-fold CV test

Methods	Sp	Sn	Ac	MCC	AUC
pKSAAP	0.802	0.627	0.719	0.413	0.768
AAindex	0.786	0.613	0.704	0.388	0.753
SPIDER2	0.755	0.414	0.594	0.235	0.739
PEP2D	0.761	0.365	0.574	0.199	0.693
KSAAP	0.801	0.652	0.731	0.443	0.806
PreAIP*	0.806	0.709	0.761	0.486	0.821

* PreAIP is the combined method of SPIDER2, PEP2D, KSAAP, AAindex and pKSAAP encoding schemes and their weight coefficients are 0.00, 0.00, 0.10, 0.35, and 0.55, respectively via RF scores

2.3.5 Comparison of different machine learning algorithms

The performance of the RF was compared to the three widely used machine learning algorithms, NB, SVM, and ANN by using the same training datasets and features, as shown in **Table 2-10**. The AUC values of the prediction by the five algorithms were calculated by a 10-fold CV test, while using the SPIDER2, PEP2D, AAindex, KSAAP, and pKSAAP encodings and their combined method. The RF provided higher AUC than any other algorithms for all the encoding methods and their combined method.

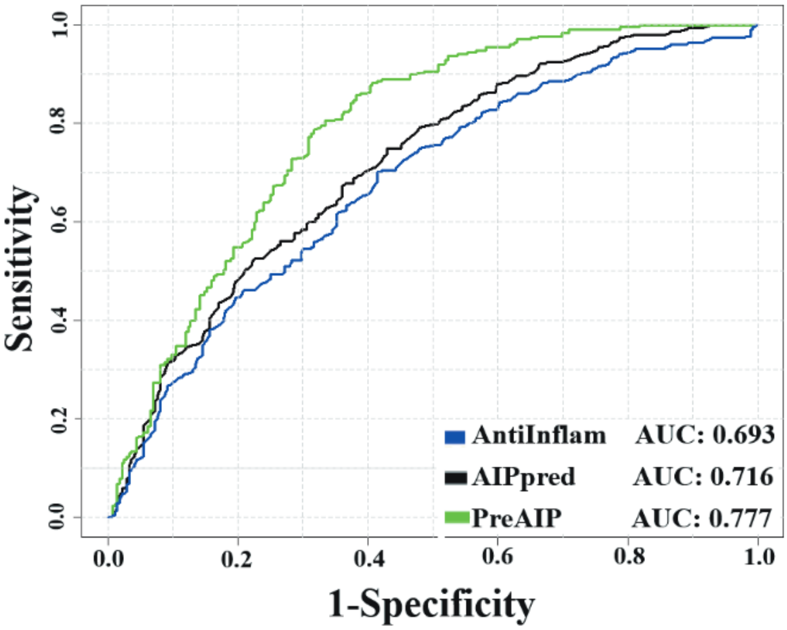

Figure 2-8: AUC values with 60% peptide redundancy removal on the test dataset.

2.3.6 The effect of peptide redundancy on the predictive model

The peptide redundancy might lead to the overestimation on the predictive performance. Therefore, we adopted 60% identity cut-off at the peptide level by CD-HIT (Huang et al., 2010). After removing the 60% sequence redundancy, we re-assembled a training dataset that contained 1,098 positive and 1,226 negative samples, and the test dataset contains 308 positive and 275 negative samples. After removal of the 60% peptide redundancy, the overall performance of PreAIP in the 10-fold CV decreased slightly (AUC = 0.821) as shown in **Table 2-10**. Moreover, PreAIP could still achieve the best performance on the independent testing dataset (**Figure 2-8**). For instance, when compared with AIPpred, PreAIP achieved AUC values of approximately 6% higher. The PreAIP also achieved at least an 8% AUC improvement compared with AntiInflam. These performance comparison results prove that PreAIP predictor provides a stable or competitive performance compared with the other predictors on the test dataset, even after 60% peptide redundancy.

2.3.7 Advantages of PreAIP

In theoretical viewpoints, comparison of the proposed PreAIP with existing predictors is summarized: (1) The PreAIP investigated the primary sequence, physicochemical properties, structural and evolutionary features, although the AIPpred and AntiInflam predictors used only primary sequence encoding method. For instance, in AntiInflam method (Gupta et al., 2017), studied hybrid features based on primary sequence encoding schemes such as amino acid composition (AAC), dipeptide composition (DPC), and tripeptide composition with SVM algorithm. The AIPpred (Manavalan et al., 2018b) studied individual composition (AAC, AAindex, DPC, and chain-transition-composition) through multiple machine learning algorithms. (2) Since existing prediction tools did not control the Sp level, users cannot understand which AIP is highly positive or negative from their servers. On the other hand, the PreAIP controlled Sp at high, moderate and low levels by changing the threshold of the RF scores, based on 10- fold CV test results. A limitation of the PreAIP is that the employed dataset is still small, but we believe that the dataset will grow to enable intensive identification of AIPs.

2.3.8 Development of PreAIP Server

A web server of the PreAIP has been developed and publically accessible at http://kurata14.bio.kyutech.ac.jp/PreAIP/. The web application was implemented by programming languages of Java scripts, Perl, R, CGI scripts, PHP and HTML. After submitting a query sequence to the server, it generates consecutive feature vectors. Then, the server optimizes the performances through RFs. After completing the submission job, the server returns the result in the output webpage which consists of the job ID and probability scores of the predicted AIPs in a tabular form. A user gets a job ID like "2018032900067" and can save this ID for a future query. The server stores this job ID for one month. The input peptide sequence must be in the FASTA format. Each of the 20 types of standard amino acids must be written as one uppercase letter. See the test example on the server. The length of AIP sequence was limited from 1 to 25. If users submit 200 amino acids, the PreAIP takes first 1 to 25 residues to analyze. When the peptide contains less than 25 residues, the PreAIP provides gaps (-) to the missing residues to compensate a peptide length of 25.

2.4 Summary of chapter 2

We have designed an accurate and efficient computational predictor for identifying potential AIPs. It outperforms the existing methods and is effective in understanding some mechanisms of AIP identification. An IG-based feature selection method was carried out to suggest sequence motifs of AIPs from KSAAP encoding. A user-friendly web-server was developed and freely available for academic users.

CHAPTER 3

CHAPTER 3 PREDICTION OF PROINFLAMMATORY PEPTIDES BY FUSING OF MULTIPLE FEATURE REPRESENTATIONS

3.1 Introduction

A proinflammatory cytokine or an inflammatory peptide (PIP) is referred to as a type of signaling molecules, which is secreted from immune cells and certain cell types for promoting the inflammation (Watkins et al., 1995; Zhang and An, 2007). The PIPs contain interleukin-1 (IL-1), IL-12, and IL-18, interferon-gamma, tumor necrosis factors, and granulocyte-macrophage association motivating factors, which contribute to the first line of defense against invading pathogens (Scarpioni et al., 2016). Different studies reported that PIPs play an important role in human physiology such as vaccines and immunotherapeutic drugs (Cavaillon, 2001; Pinho-Ribeiro et al., 2015; Zhang and An, 2007). Nonetheless, these peptides may cause unwanted immuno-activity in the B or T cell instigation and other proinflammatory events (Gordon et al., 2005; Gustafsson et al., 2010; Shi et al., 2015). Diverse transferrable agents were found in proteins with immunomodulatory properties, which can assist in the evolution and instigation of diseases (Desmet, 1987; Hsu et al., 2001).

The importance of PIPs is confirmed through the pathophysiological dealings (Mukhopadhyay et al., 2014; Zhao et al., 2005). For instance, Herpes Simplex Virus-2 produces a glycoprotein G-2 through the gG-2p20 peptide that causes proinflammatory responses in human neutrophils and activates as an effective antineoplastic agent (Bellner et al., 2005; Bylund et al., 2001). Similarly, the C-peptide of PIPs produces proinsulin which is used in peptide-therapeutics but leads to inflammation in vasculature and kidney or long-term deterioration of diseases (Vasic and Walcher, 2012). Those PIP functions are important to analyze. To reduce time and economic cost, a computational identification method of PIPs is needed before experimental verification. There are only a few computational methods developed for PIP identification, e.g., ProInflam (Gupta et al., 2016) and PIP-EL (Manavalan et al., 2018c). In 2016, Gupta et al. firstly introduced a computation method named ProInflam that employed a support vector machine (SVM) classifier with different sequence-based features (Gupta et al., 2016). Recently, Manavalan et al. developed another computation method named PIP-EL by using several sequence features (Manavalan et al., 2018c). All in all, existing methods provide good prediction results, but their prediction performances are yet not fully satisfactory and there is still room for further improvement.

Motivated by these considerations, we thus propose a computational predictor named ProIn-Fuse (Prediction of Proinflammatory Peptides) to accurately predict PIPs by making the use of multiple feature representations. The overall framework of the ProIn-Fuse is depicted in **Figure 3-1**. Firstly, we collected up-to-date PIPs and non-PIPs from the IEDB database and constructed the benchmark dataset containing samples with low similarity. Secondly, we calculated the probabilistic scores by employing random forest (RF) algorithm in conjunction with various encoding schemes, i.e., the k-mer composition from profile (kmer-pr), profile-based composition of the amino acid (PKA), k-mer composition of the amino acid (kmer-ac), k-space amino acid pairs (KSAP), binary encoding (BE), amino acid index properties (AIP), N-terminal 5 and C-terminal 5 dipeptides composition (C5N5-DC), and structural features (SF). Thirdly, the ProIn-Fuse was established by fusing the successive probabilistic scores using a linear regression (LR) model. Cross-validation (CV) and independent results showed that the ProIn-Fuse yielded better performance than those obtained by existing predictors and other well-known machine learning (ML) models, signifying that it has a great advantage as an auxiliary tool for PIP identification.

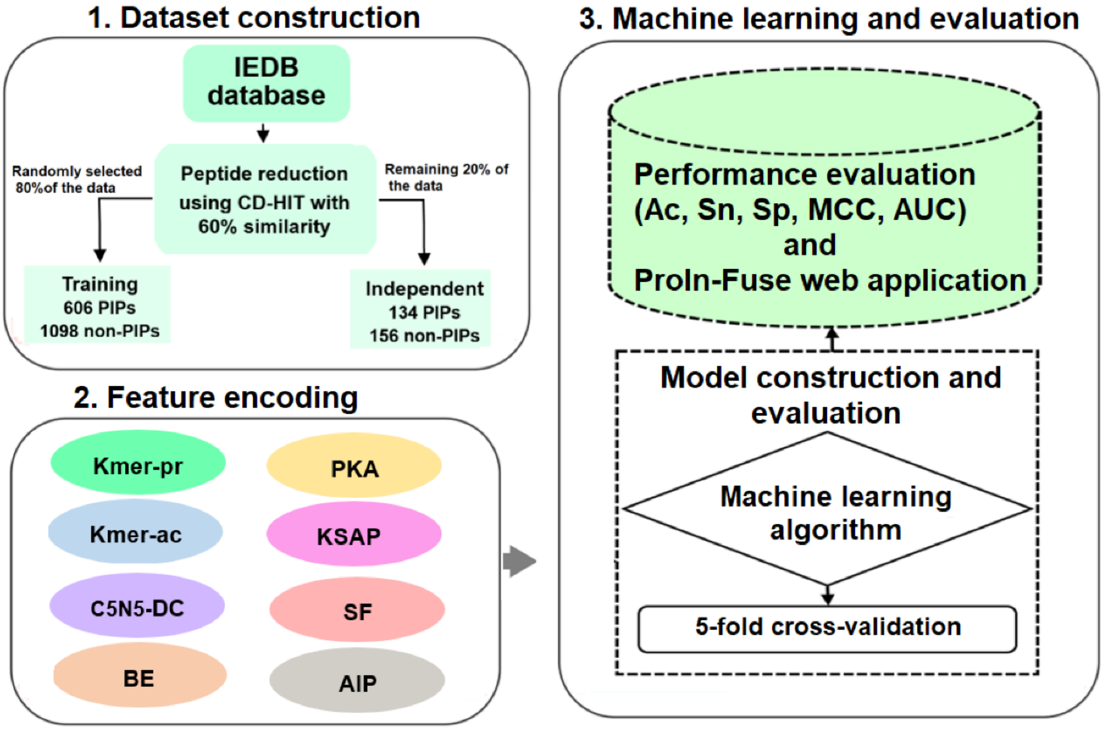
1. Dataset construction
IEDB database
Randomly selected 80%of the data
Peptide reduction using CD-HIT with 60% similarity
Remaining 20% of the data
Training 606 PIPs 1098 non-PIPs
Independent 134 PIPs 156 non-PIPs
2. Feature encoding
Kmer-pr
PKA
Kmer-ac
KSAP
C5N5-DC
SF
BE
AIP
3. Machine learning and evaluation
Performance evaluation (Ac, Sn, Sp, MCC, AUC) and ProIn-Fuse web application
Model construction and evaluation
Machine learning algorithm
5-fold cross-validation

Figure 3-1 Computational workflow for the prediction of proinflammatory peptides

3.2 Materials and Methods

3.2.1 Dataset preparation

To develop an ML-based predictor, we collected experimentally validated positive datasets from the IEDB database (1,505 PIPs), which belong to any of the proinflammatory cytokines (IL1α, TNFα, IL1β, IL12, IL18, and IL23) (Fleri et al., 2017; Vita et al., 2015). Negative samples (3,350 non-PIPs) that are excluded from the proinflammatory cytokines were collected from the IEDB database. The PIPs or non-PIPs whose amino acid residue length is greater than or equal to 5 and less than or equal to 25 were considered. Although previous studies have provided their benchmark datasets (Manavalan et al., 2018c), these datasets still contain many redundant samples leading to overestimated accuracy. Here, to avoid overestimation caused by the homology biases, the sequence identity between both positive and negative datasets was reduced to 0.60 using CD-HIT (Huang et al., 2010). After such a screening process, the benchmark dataset consists of 741 PIPs and 1254 non-PIPs. The benchmark dataset was randomly divided into the training and independent sets with a ratio of 8:2. Finally, the training dataset consists of 607 PIPs and 1098 non-PIPs, while the independent dataset consists of 134 PIPs and 156 non-PIPs. The training and independent datasets used in this study are publicly accessible at http://kurata14.bio.kyutech.ac.jp/ProIn-Fuse/download.php.

3.2.2 Feature encodings

To develop a sequence-based predictor, the critical step is to represent a peptide sequence by a fixed-length feature vector (Hasan et al., 2019b; Hasan et al., 2020f; Hasan et al., 2020h). To encode PIP and non-PIP sequences, eight types of encoding schemes were used: Kmer-pr, PKA, Kmer-ac, KSAP, BE, AIP, C5N5-DC, and SF. We summarized each encoding as follows.

3.2.3 Kmer-pr encoding

The Kmer-pr was generated by the PSI-BLAST (version 2.2.26+) with two restrictions of iteration times of 3 and e-value of 1.0×10^{-4} from the Swiss-Prot database, respectively (Chen et al., 2009; Dong et al., 2013). The Kmer-pr generated a PSSM profile for the PIP and non-PIP sequences and encoded the composition-based features of the profile. At K=0 and 3, an 8020 (20+20×20×20)-dimensional (D) feature vector was generated.

3.2.4 PKA encoding

The PKA encoding measured the possible *k*-space composition from the PSSM profile, in the same way as the earlier study of anti-inflammatory peptides identification (Khatun et al., 2019a). A 1200 (3×20×20)D feature vector was generated when the optimal K-space was 0, 1, and 2.

3.2.5 Kmer-ac encoding

The Kmer-ac encoded the amino acid residue sequences with a fixed length of amino acids. At K=1, the Kmer-ac encodes monopeptides into a 20D feature vector. At K=2 it encodes dipeptides into a 400(20×20)D feature vector; at K=3 it encodes tripeptides into an 8000(20×20×20)D feature vector. In this study, K=1 and 3 are considered, which generates an 8,020D feature vector.

3.2.6 KSAP encoding

The KSAP encoding is successively used in many bioinformatics prediction tasks (Hasan M.M., 2018; Hasan MM, 2017; Md. Mehedi Hasan, 2017). Possible *k*-space amino acid pairs were collected from the curated peptides. At K=0, a 400(20×20)D feature vector was generated. At K=0, 1, and 2, it generates a 1200D feature vector.

3.2.7 Binary encoding

A 20-dimensional binary vector is used to encode an amino acid residue of a peptide sequence (Hasan et al., 2016). The BE encoding generates a 420 (21×20)D feature vector for a peptide sequence window.

3.2.8 AIP encoding

The AIP encoding uses amino acid properties (Kawashima et al., 2008). We selected 10 instructive amino acid indices by assessing the diverse types of properties (**Table 3-3)**. In a 22 peptide length of sequences, the AIP encoding generates a 220(22×10)D feature vector.

3.2.9 C5N5-DC encoding

To employ C5N5-DC, we extracted 5 amino acids from C- and N-terminal. Then we encode a new sequence window as dipeptides. When 20 natural amino acids are considered, the C5N5-DC scheme generates a 400(20×20)D feature vector.

3.2.10 Structure feature encoding

We employed SF to represent the PIP and non-PIP sequences. We used SPIDER2 (Yang et al., 2017) that considers the backbone torsion angles, accessible surface area, and secondary structure. The SF that consists of 8 types of feature vectors generated a 176 (22×8)D feature vector for a sequence window.

3.2.11 Machine learning algorithms

Different ML-based algorithms were employed to classify the PIP and non-PIP sequences, including RF, SVM, AB, and NB. The RF is a supervised ML algorithm (Hasan et al., 2017c; Hasan and Kurata, 2018; Hasan et al., 2016; Tahir M, 2019), which works as a group of decision trees (Basith et al., 2020a; Charoenkwan et al., 2020; Charoenkwan et al., 2013; Hasan et al., 2019d; Khatun et al., 2019b; Maclin and Opitz, 1999; Manavalan et al., 2019b; Polikar, 2006; Rokach, 2010). An R package of ‘randomForest’ was employed (Hasan et al., 2019d; Khatun et al., 2019a). The SVM has been widely applied to address binary class prediction problems (Hasan et al., 2015). We used a radial basis function of the LIBSVM package to optimize their parameters (cost function and gamma) by a simple grid search. The AB is an adaptive boosting ML algorithm. To improve the method performance, The AB classifiers are linearly combined with the weight coefficients that characterize the final output of the boosted classifier. The R package (https://cran.r-project.org/web/packages/adabag) was employed for AB. The NB is a simple ML algorithm based on applying Bayes' theorem with robust naive assumptions (Hasan et al., 2018a; Hasan et al., 2017a). We used an R package of NB to classify PIP and non-PIP samples at (https://cran.r-project.org).

3.2.12 Performance evaluation matrixes

To evaluate the performance of our prediction models, we used the five statistical measures: accuracy, sensitivity, specificity, Matthews’ Correlation Coefficient (MCC) (Liaw, 2002; Manavalan et al., 2018c; Schaduangrat et al., 2019; Shoombuatong et al., 2019; Su et al., 2019; Win et al., 2017), and Area Under the Curve (AUC). The following formulas are used to calculate these measures:

$$\mathrm{Ac} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \qquad (1)$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \qquad (2)$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \qquad (3)$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FP})}} \qquad (4)$$

where TP, FP, TN, and FN respectively characterize the numbers of the correctly predicted samples as PIP, incorrectly predicted ones as PIP, correctly predicted ones as non-PIP, and incorrectly predicted ones as non-PIP. To evaluate the AUC values, the ROC curves were measured by the pROC package of the R language (Centor, 1991; Gribskov and Robinson, 1996).

3.2.13 Fusion model

To enhance the prediction performance, we fused the curated ML probability scores from the Kmer-pr, PKA, Kmer-ac, KSAP, BE, AIP, C5N5-DC, and SF encodings via an LR model as follows:

$$\text{Fusion} = \sum_{i=1}^{8} W_i\, M_i \tag{5}$$

where W_i, (i= 1, 2, 3, …, 8*)* are the weight coefficients and M_i the ML score of each single-encoding based model under a constraint: $\sum_{i=1}^{8} W_i = 1$. The linearly combined models of the ML scores assessed by the eight encodings are denoted as the fusion model.

3.2.14 Hybrid model

To examine the effectiveness of the ProIn-Fuse, we compared it with a hybrid model (i.e., hybrid model is the conjunction of all curated features in plain as a row) for the prediction of PIPs. In the hybrid model, the eight feature vectors (F) generated by the Kmer-pr, PKA, Kmer-ac, KSAP, BE, AIP, C5N5-DC, and SF were lined up in a row as follows:

$$\mathrm{H} = (\mathrm{F(Kmer - pr), F(PKA), F(Kmer - ac), F(KSAP), F(AIP), F(C5N5 - DC), F(SF)}) \tag{6}$$

where H is the hybridized feature vector consisting of the eight feature vectors. The total dimension of hybrid model was 19,656D. Details are described elsewhere [35, 38]. To reduce the dimension from the hybrid model, we considered a feature ranking method of Wilcoxon rank sum (WR) test [30, 40].

3.3 Results and Discussion

3.3.1 Preference of PIP sequence

It is of importance to examine the sequence preference of PIPs and non-PIPs. **Figure 3-2** compares the difference of amino acid frequencies between PIP and non-PIP sequences on the first 20 N-terminal residues (http://www.twosamplelogo.org/). We found that the PIP and non-PIP samples have significantly different sequence preferences. The arginine (R) at positions 1 and 4, serine (S) at positions 2, 11, and 12, leucine (L) at positions 5, 11, and 14-16, glycine (G) at positions 16, 18, and 19, asparagine (N) at positions 5 and 6, Glutamine (Q) at positions 1 and 10 were enriched in the upstream of the PIPs. For the non-PIPs, aspartic acid (D) at positions 1, 5, 8, 11, 12, 15, and 17 and G at positions 7, 10, and 14 were depleted. However, no significant residue was enriched at positions of 3, 8, 9, 13, and 20, and was depleted at positions of 3, 9, 13, 16, and 18-20. These observations suggested that the PIP and non-PIP samples have distinct location-specific preferences of amino acid residues. This information is important to identify PIPs.

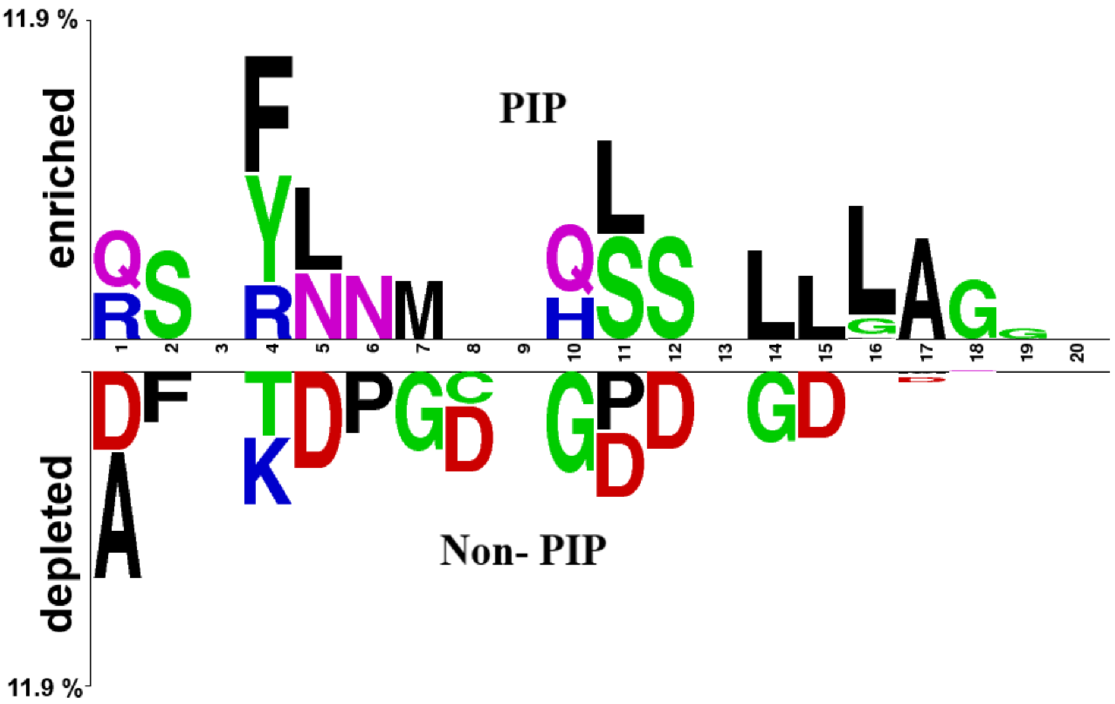

Figure 3-2 Two-Sample-Logos were visualized for PIP and non-PIP samples (https://www.twosamplelogo.org/). On the 20N-terminal amino acid sequence, the position-wise residues significantly enriched or depleted (*t*-test, $P < 0.05$) are presented

3.3.2 Evaluation of ProIn-Fuse on the training dataset

PIPs and non-PIPs sequences of the benchmark dataset were encoded into the eight feature vectors by using the Kmer-pr, PKA, Kmer-ac, KSAP, BE, AIP, C5N5-DC, and SF. The resultant feature vectors were inputted into the RF model to construct eight, single encoding-based RF models. The prediction performances of them were evaluated using 5-fold CV tests as shown in **Figure 3A**. The Kmer-pr and PKA encodings achieved a similar performance with AUC of ~0.78. As seen, their AUCs were 2.6-8.6% higher than the AUCs obtained from the other six encodings. As suggested by many studies, there were a number of ways to incorporate multiple prediction models, including meta-predictors (Boopathi et al., 2019; Manavalan et al., 2019a, b, c), hybrid models and fusion methods (Hasan et al., 2019b, c). Here, we employed the fusion method that linearly combines the eight, single encoding-based RF models, named as ProIn-Fuse. The weight coefficients of them were optimized to maximize the AUC. In the ProIn-Fuse, the optimal weight coefficients of Kmer-pr, PKA, Kmer-ac, KSAP, BE, AAindex, C5N5-DC and SF are 0.35, 0.45, 0.10, 0.00, 0.00, 0.10, 0.00 and 0.00, respectively, indicating the Kmer-pr, PKA, Kmer and AAindex contributed 45%, 35%, 1%, and 1% to the final prediction, respectively, while the remaining encodings did not contribute to the final prediction. In the combined model, the Kmer-pr and PKA-based models significantly contributed to the prediction, compared to the other encoding models. The AUCs of all the single encoding-based models and the ProIn-Fuse model are assessed by 5-fold CV test in **Figure 3-3A**. Remarkably, the ProIn-Fuse yielded the highest AUC of 0.817 with the values of Sn, Sp, Ac, and MCC of 0.596, 0.866, 0.784, and 0.506, respectively (**Table 3-1**). All in all, the ProIn-Fuse significantly outperformed all the single encoding-based models with two-sample *t*-test at the level of *p*-value< 0.05.

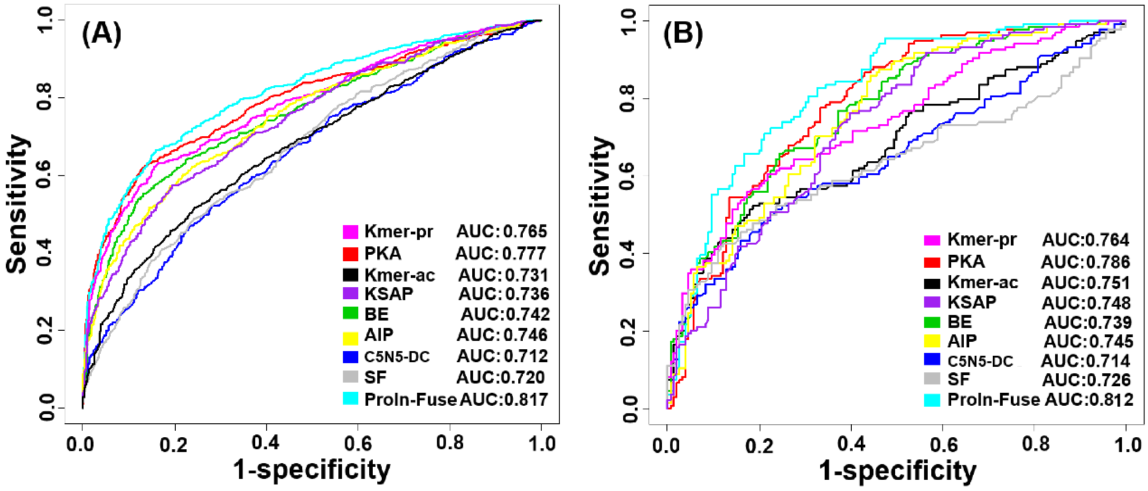

Figure 3-3 ROC curves of ProIn-Fuse and eight, single encoding-based models on **A)** training and **B)** independent datasets.

Table 3-1 Performance of ProIn-Fuse and eight, single encoding-based models on the training dataset.

Method	Sn	Sp	Ac	MCC	AUC	P-value
Kmer-pr	0.519	0.887	0.750	0.448	0.765	0.021
PKA	0.534	0.892	0.754	0.456	0.777	0.028
Kmer-ac	0.478	0.842	0.711	0.413	0.731	<0.001
KSAP	0.484	0.852	0.713	0.422	0.736	<0.001
BE	0.504	0.889	0.752	0.434	0.742	0.012
AIP	0.508	0.890	0.751	0.445	0.746	0.017
C5N5-DC	0.442	0.829	0.693	0.293	0.712	<0.001
SF	0.454	0.822	0.695	0.312	0.720	<0.001
ProIn-Fuse	0.596	0.866	0.784	0.506	0.817	-

For the ProIn-Fuse model, the optimal weights for Kmer-pr, PKA, Kmer-ac, KSAP, BE, AIP, C5N5-DC, and SF are 0.35, 0.45, 0.10, 0.00, 0.00, 0.10, 0.00 and 0.00, respectively.

3.3.3 Comparison of RF with other well-known MLs on training dataset

To demonstrate the effectiveness of the RF algorithm employed by the ProIn-Fuse, the ProIn-Fuse was compared with the fusion models that linearly combine the SVM-, AB- and NB-evaluated scores with the eight, single encoding schemes, which are named SVM-, AB- and NB-Fuse models, respectively. By CV test, we compared the ProIn-Fuse with the SVM-, AB- and NB-Fuse models, as shown in **Table 3-4** and **Figure 3-4**. The ProIn-Fuse achieved higher performances than any other ML-fusion models, while SVM-Fuse was comparable to the ProIn-Fuse. Moreover, the AUC of ProIn-Fuse was 2-7% higher than those obtained by SVM-, AB- and NB-Fuse models, indicating the superiority of the RF over the other well-known ML algorithms.

Table 3-2 Comparison with existing predictors

Method	Sn	Sp	Ac	MCC
ProInflam	0.666	0.596	0.628	0.264
PIP-EL	0.542	0.741	0.649	0.299
ProIn-Fuse	0.666	0.814	0.746	0.488

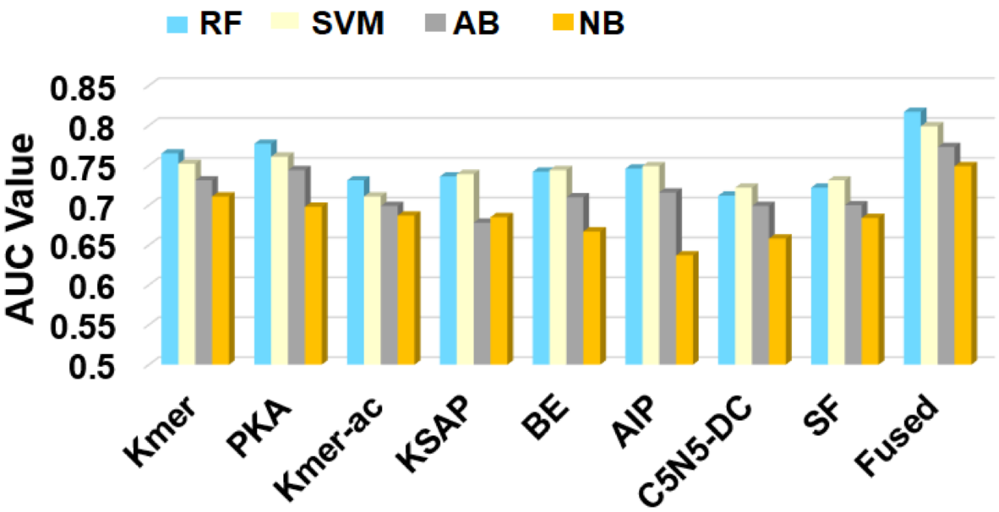

Figure 3-4 Performance comparison of different machine learning algorithms

3.3.4 Comparison of ProIn-Fuse with a hybrid model on training dataset

As mentioned above, there were various ways to incorporate multiple prediction models. In this section, we compared the performance of the ProIn-Fuse against the hybrid model, a sequential combination model, on the same training dataset. The hybrid model lined up all of the eight feature vectors in a row, and then feed these feature vectors into four different classifiers (i.e., RF, SVM, AB, and NB). As shown in **Figure 3-5**, the hybrid models implementing RF, SVM, AB, and NB yielded AUC values of 0.789, 0.763, 0.757, and 0.744. Furthermore, the WR test was employed to select the important features from the hybrid model. According to the relevance to the redundancy between the features, the WR test can rank all the features themselves. Based on the WR test, we selected the top 1450D features from the total 19,656D and inputted them to RF, SVM, AB, and NB, respectively, and evaluated the resultant prediction models using the 5-fold CV test. As shown in **Table 3-5**, with feature selection the hybrid models implementing RF, SVM, AB, and NB yielded AUC values of 0.794, 0.779, 0.763, and 0.746, respectively, while the ProIn-Fuse provided an AUC value of 0.817. Thus, the AUC value of the ProIn-Fuse was ~2 to 6% higher than that of any hybrid models.

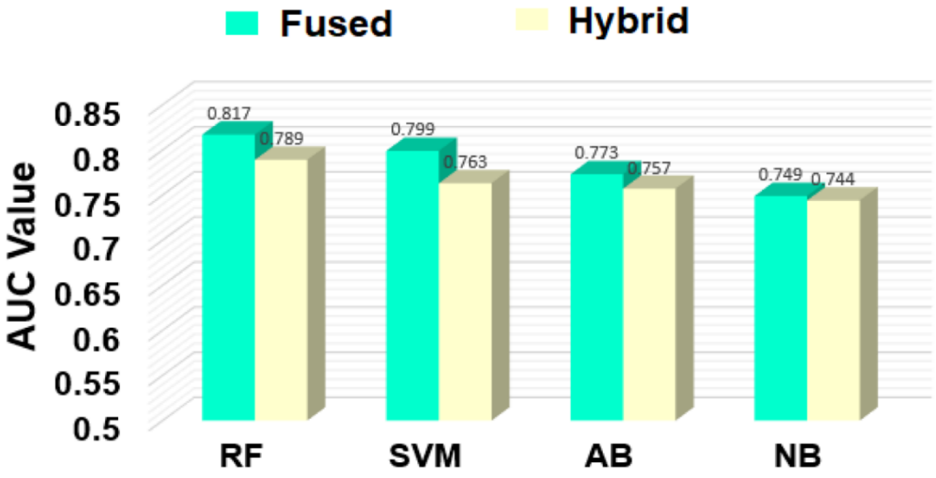

Figure 3-5 Performance comparison of the fused and hybrid models using diverse machine learning algorithms

3.3.5 Performance of ProIn-Fuse on independent datasets

In this section, we validated the generalization capability of the ProIn-Fuse by evaluating its performance on the independent dataset. The performance of the ProIn-Fuse was compared with the eight, single encoding-based RF models, as shown in **Figure 3-3B**. For all the single encoding-based RF models, the PKA encoding achieved the highest AUC of 0.786, while the Kmer-pr and Kmer-ac encodings yielded the second and third highest AUCs of 0.764 and 0.751, respectively. These results were well consistent with the weight coefficients of the ProIn-Fuse, where the weight coefficients of Kmer-pr, PKA and Kmer-ac are 0.35, 0.45 and 0.10, respectively. The ProIn-Fuse achieved the best AUC of 0.822, which was 3-11% higher than the AUCs of the eight, single encoding-based RF models.

Table 3-3 Selected 10 types of AIP properties used in this study.

Properties	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
MIYS990104	-0.04	0.07	0.13	0.19	-0.38	0.14	0.23	0.09	-0.04	-0.34	-0.37	0.33	-0.30	-0.38	0.19	0.12	0.03	-0.33	-0.29	-0.29
BLAM930101	0.96	0.77	0.39	0.42	0.42	0.80	0.53	0.00	0.57	0.84	0.92	0.73	0.86	0.59	-2.50	0.53	0.54	0.58	0.72	0.63
MAXF760101	1.43	1.18	0.64	0.92	0.94	1.22	1.67	0.46	0.98	1.04	1.36	1.27	1.53	1.19	0.49	0.70	0.78	1.01	0.69	0.98
CEDJ970104	7.9	4.9	4.0	5.5	1.9	4.4	7.1	7.1	2.1	5.2	8.6	6.7	2.4	3.9	5.3	6.6	5.3	1.2	3.1	6.8
LIFS790101	0.92	0.93	0.60	0.48	1.16	0.95	0.61	0.61	0.93	1.81	1.30	0.70	1.19	1.25	0.40	0.82	1.12	1.54	1.53	1.81
ARGP820101	0.61	0.60	0.06	0.46	1.07	0.	0.47	0.07	0.61	2.22	1.53	1.15	1.18	2.02	1.95	0.05	0.05	2.65	1.88	1.32
ARGP820102	1.18	0.20	0.23	0.05	1.89	0.72	0.11	0.49	0.31	1.45	3.23	0.06	2.67	1.96	0.76	0.97	0.84	0.77	0.39	1.08
BHAR880101	0.357	0.529	0.463	0.511	0.346	0.493	0.497	0.544	0.323	0.462	0.365	0.466	0.295	0.314	0.509	0.507	0.444	0.305	0.420	0.386
ARGP820103	1.56	0.45	0.27	0.14	1.23	0.51	0.23	0.62	0.29	1.67	2.93	0.15	2.96	2.03	0.76	0.81	0.91	1.08	0.68	1.14
ISOY800107	1.34	2.78	0.92	1.77	1.44	0.79	2.54	0.95	0.00	0.52	1.05	0.79	0.00	0.43	0.37	0.87	1.14	1.79	0.73	0.00

Table 3-4 AUC values of different machine learning algorithms on the training dataset.

Method	**RF**	**SVM**	**AB**	**NB**
Kmer-pr	0.765	0.751	0.743	0.721
PKA	0.777	0.759	0.746	0.691
Kmer-ac	0.731	0.728	0.695	0.688
KSAP	0.736	0.738	0.673	0.681
BE	0.742	0.746	0.714	0.663

AIP	0.746	0.747	0.716	0.646
C5N5-DC	0.712	0.722	0.694	0.653
SF	0.720	0.726	0.692	0.672
Combined	0.817	0.799	0.773	0.749

“Combined” specifies that the performance of the optimized fused features. For the RF model, the optimal weights for Kmer-pr, PKA, Kmer-ac, KSAP, BE, AIP, C5N5-DC, and SF are 0.35, 0.45, 0.10, 0.00, 0.00, 0.10, 0.00 and 0.00, respectively. In the same way, the weight values of SVM, AB, and NB were given as (0.30, 0.35, 0.00, 0.15, 0.00, 0.2, 0.00, and 0.00), (0.30, 0.55, 0.00, 0.1, 0.00, 0.05, 0.00, and 0.00), and (0.35, 0.40, 0.00, 0.15, 0.00).

3.3.6 Comparison of ProIn-Fuse with existing predictors

To investigate the superiority of the ProIn-Fuse, we compared its performance with the existing two PIP predictors, i.e., ProInflam(Gupta et al., 2016) and PIP-EL(Manavalan et al., 2018c), using the same independent dataset. **Table 3-2** compares the prediction results among the ProIn-Fuse, ProInflam and PIP-EL. Note that the prediction results of ProInflam and PIP-EL were obtained by feeding the protein sequences to their web servers. The ProIn-Fuse achieved better performances (i.e., Sn of 0.666, Sp of 0.814, Ac of 0.746, and MCC of 0.488) than PIP-EL and ProInflam for all the four statistical measures. Furthermore, the MCC value of the ProIn-Fuse was 19% and 22% higher than the PIP-EL and ProInflam, respectively. Considering that the independent test is a rigorous CV method, we thus claim that the proposed ProIn-Fuse is superior to the existing PIP predictors.

Table 3-5 AUC values Performance of hybrid model on the training dataset using WR-based feature selection approach.

Method	Sn	Sp	Ac	MCC	AUC
RF	0.558	0.890	0.771	0.467	0.794
SVM	0.542	0.890	0.766	0.459	0.779
AB	0.477	0.891	0.743	0.439	0.763
NB	0.452	0.890	0.734	0.431	0.746

3.4 Summary of chapter 3

We have developed an efficient and accurate computational predictor named ProIn-Fuse for PIPs identification. The ProIn-Fuse linearly combined the eight probability scores evaluated by the single encoding-based RF models. The ProIn-Fuse more effectively identified PIPs than any single encoding-based RF models and the other fusion/hybrid models. To validate the superiority of the ProIn-Fuse, we have compared it with ProInflam and PIP-EL using the independent test. The ProIn-Fuse outperformed the existing predictors. To help potential users, a user-friendly web-application of the ProIn-Fuse was provided for public use at http://kurata14.bio.kyutech.ac.jp/ProIn-Fuse/. It is highly anticipated that the proposed ProIn-Fuse can be instrumental in facilitating the identification of novel PIPs for drug design and discovery.

CHAPTER 4

CHAPTER 4 PREDICTION OF ANTI-TUBERCULAR PEPTIDES BY INTEGRATING THE AMINO ACID PATTERNS AND PROPERTIES

4.1 Introduction

Tuberculosis (TB) is regulated by *Mycobacterium tuberculosis* (Mtb), is a type of infective disease, being responsible as a major threat for the human beings (Hamilton et al., 2015; WHO, 2017b; Zumla et al., 2015). Among 10 reasons for human deaths, TB is the foremost cause, mentioned by the ‘Global TB report 2018’ issued by the World Health Organization (WHO) (WHO, 2017a). In 2017, TB killed 1.6 million people. There were ten million people newly affected by TB with 5.8 million males, 3.2 million females, and 1.0 million kids (WHO, 2017a). TB is the universal health anxiety, mostly in developing countries. It is assessed that 44% TB covered by only three high-risk countries such as India, China, and Indonesia (WHO, 2017a). Nearly 90-95% of infected people induce immune responses against Mtb, thus they do not get ill. It is called to be latently infected, because TB-causing bacteria harbor somewhere in human bodies, especially in lungs. The remaining 5-10% of infected people become sick with active TB when bacteria replicates are unrelieved in spite of all-out efforts by the immune system.

Treatment of TB typically leads to complication and shows high mortality rate (nearly 15%) due to the widespread of multi-drug resistance (MDR) strains (Wilson and Tsukayama, 2016). The TB treatment is far from satisfactory at present. General treatment requires a long-term, daily administration of drugs (Wang et al., 2015), which are less effective and toxic due to severe side effects. MDR is resistant to most influential first-line anti-TB drugs, such as rifampicin and isoniazid. It needs treatment with the second-line medications including fluoroquinolones and aminoglycosides [7], which in general are more side effects, less effective and much more expensive than the first-line drugs. The MDR is an urgent priority for developing anti-TB new drugs, mentioned by WHO (Arbex et al., 2010). Different complex mechanisms are involved in the expansion of MDR acquired by the *Mycobacterium*. The MDR is related to the diverse iatrogenic factors, thus the MDR rates are increasing in highly populated cities and low-income countries (Kim and Yang, 2017; Silva et al., 2016). Due to the technical limitations of *in vitro* drug vulnerability testing, the drug-resistance mechanisms of TB have not clearly been defined by WHO (Silva et al., 2016), notwithstanding several initiatives of TB treatment. Novel medicines are still desirable to control this severe disease (Aggerbeck et al., 2018; Chaurasiya, 2018).

Nowadays, a peptide-based therapy appears as a potential alternative to therapies of anti-mycobacterial drugs (Zasloff, 2006). Anti-TB peptides with low immunogenicity make them a possible complement for expectable TB drugs (AlMatar et al., 2018; Jhamb et al., 2014). Large-scale experimental screenings were carried to explore anti-TB peptides

(Padhi et al., 2014; Yount and Yeaman, 2004). Many experimental candidates of anti-TB peptides were found and registered in the AntiTbPdb database (Usmani et al., 2018b). Notwithstanding the increasing number of experimentally validated anti-TB peptides, the mechanisms by which anti-TB peptides affect TB remain largely unknown (Gao et al., 2015; Gavrish et al., 2014; Nikonenko et al., 2004; Usmani et al., 2018b). Since the large-scale experimental identification of anti-TB peptides is laborious and time-consuming, alternative, computational methodologies are required that provide an accurate and robust prediction of anti-TB peptides. Recently, Usmani et al. developed the AntiTBpred, a computational predictor implementing a support vector machine (SVM) classifier (Usmani et al., 2018a). They illustrated that the composition of amino acids and N5C5 binary profiles (i.e., five amino acid residues from the N- and C-terminals) contribute to the enhanced prediction accuracy. However, the exact performance of AntiTBpred was not assessed, because they did not separate the training and independent samples.

In this research, we have established a computational predictor termed iAntiTB (Identification of Anti-tubercular Peptides) through integration of amino acid patterns and properties, as shown in **Figure 4-1**. We classified four sequential feature vectors through the Random Forest (RF) and Support Vector Machine (SVM) and then combined the RF and SVM scores via a linear regression model. The resulting iAntiTB outperformed the existing predictors.

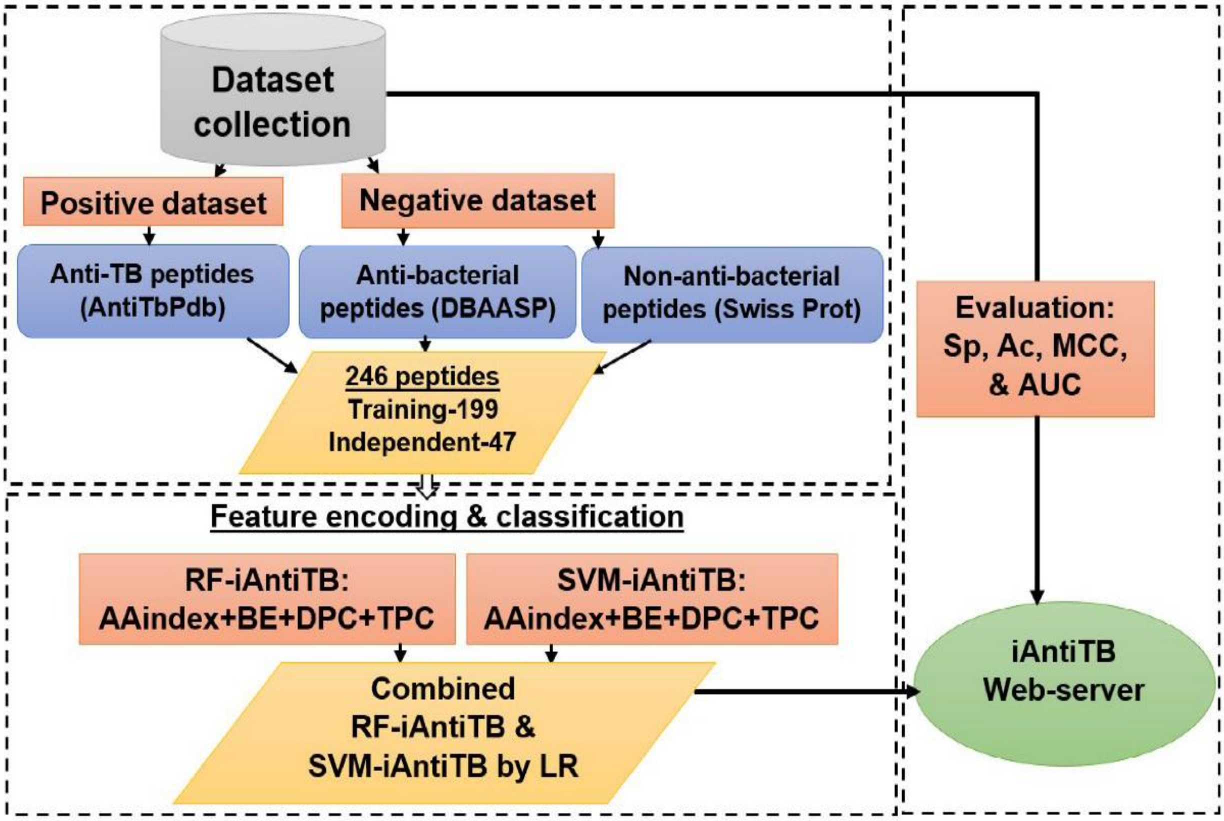
Dataset collection
Positive dataset
Negative dataset
Anti-TB peptides (AntiTbPdb)
Anti-bacterial peptides (DBAASP)
Non-anti-bacterial peptides (Swiss Prot)
246 peptides
Training-199
Independent-47
Evaluation: Sp, Ac, MCC, & AUC
Feature encoding & classification
RF-iAntiTB: AAindex+BE+DPC+TPC
SVM-iAntiTB: AAindex+BE+DPC+TPC
Combined RF-iAntiTB & SVM-iAntiTB by LR
iAntiTB Web-server

Figure 4-1 An overview of iAntiTB for predicting anti-TB peptides.

4.2 Materials and Methods

4.2.1 Data construction

To construct an efficient computational model, we collected the positive samples of anti-TB peptides from the AntiTbPdb database (Usmani et al., 2018b). After eliminating the duplicate peptides, 246 positive unique peptides were selected that are effective against Mycobacterium. The length of the peptide varies from 5 to 61. Next, we collected the two sets of negative samples as same as a recently published article [11]. The negative samples were from the non-anti-bacterial peptides of the Swiss-Prot database (Bairoch and Apweiler, 2000) and anti-bacterial peptides from the DBAASP database (Pirtskhalava et al., 2016). The collected negative samples are blind from positive ones. From the Swiss-Prot database, 246 non-anti-bacterial peptides were collected, while removing the positive samples and anti-bacterial peptides. They were named the "first negative samples".

From DBAASP database [27], we have selected anti-bacterial peptides containing natural residues and are operative against Gram negative and Gram positive microbes. After eliminating the peptide redundancy (i.e., remove the identical dataset as same positive ones), 4,192 distinctive peptides were left. From this, we have kept one of our second negative samples, containing 246 anti-bacterial peptides as same [11]. Then we used the same strategies as a recently published article to divide the positive and negative samples with a ratio of 1:1 (Usmani et al., 2018a). Consequently, the training dataset of 199 positive and 199 first-negative samples and the independent datasets of 47 positive and 47 first-negative samples were named as the "first dataset". The training dataset of 199 positive and 199 second-negative samples and the independent datasets of 47 positive and 47 second-negative samples were named as the "second dataset". The two different datasets were employed to investigate the robustness of the proposed predictor. Note that the positive samples are common between the first and second datasets and the range of peptide length was kept the same for all the datasets.

4.2.2 Sequence preference analysis

The sequence preference logos were generated using an online two-sample-logo web server (Vacic et al., 2006). The graphical logos are the representative residues in the multiple peptide/window fragment sequences, which provides the position specific preference of amino acids. The length was limited from 1 to 20 in this study. Therefore, we submitted the curated datasets to the two sample logos server at (http://www.twosamplelogo.org/) and generated sequence logos. Amino acid preference at each position is signified by a stack of symbols, where large symbols denote repeatedly detected residues or conserved residues.

4.2.3 Feature descriptors

To encode positive and negative peptide samples, four types of descriptors were employed: amino acids index (AAindex) properties, binary encoding (BE), dipeptide composition (DPC), and tripeptide composition (TPC). We summarized each descriptor as follows.

4.2.4 Amino acid index properties

The database of AAindex (a version of 9.1) registers numerical indices of biochemical and physicochemical properties (Kawashima et al., 2008). After evaluating various kinds of properties, we selected 8 types of topmost informative indexes: TSAJ990101, NOZY710101, NAKH920108, CEDJ970104, LIFS790101, BLAM930101, MAXF760101, and KLEP840101. The anti-TB peptide samples were transformed into the feature vectors of the AAindex properties. An (L×8) dimensional vector was generated where L was the peptide sequence length.

4.2.5 Binary encoding

The BE scheme represents the positional wise amino acid information. The BE summarizes the compositional information as well as the order of positional information (Hasan et al., 2017a; Hasan et al., 2018c). The BE was generated for each peptide, where each amino acid is represented by a 20-dimensional vector. For example, Alanine (A) is represented by a vector of (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0). The peptide with a length of 20 was characterized by a 400 (20 × 20) dimensional vector.

4.2.6 Dipeptide composition

The DPC values were calculated based on the composition of amino acid pairs (e.g., AA, AC, AD)$_{400}$ (Usmani et al., 2018a). To each peptide sample, a 400 (20 × 20) dimensional vector was generated. The DPC was given by:

$$\text{DPC} = [f_1, f_2, \dots, f_{400}] \qquad (1)$$

where f_i signifies the dipeptide composition of the i-th residue pair in {AA, AC, AD, … , YY}.

4.2.7 Tripeptide composition

The TPC values were calculated based on the composition of three residues (e.g., AAA, AAC)$_{8,000}$ that were connected in a sequence. For each peptide sample, an 8,000 (20 × 20 × 20) dimensional vector was generated. The TPC value of each tripeptide was given by:

$$\text{TPC} = [f_1, f_2, \dots, f_{8000}] \qquad (2)$$

where f_i represents the tripeptide composition of the i-th residue pair in {AAA, AAC, AAD, … YYY}.

4.2.8 Machine learning algorithms

Two well-known supervised machine learning classifiers of SVM (Hearst, 1998) and RF (Breiman, 2001) were employed in this study. The RF algorithm has been widely used in medicine and computational biology fields (Hasan M.M., 2018; Hasan MM, 2018). RF works on a large ensemble of classifiers and regression trees. The RF models of the 'randomForest' package (https://www.r-project.org/) were optimized with 1,000 trees via 10 fold cross-validation (CV) test. The SVM^{light} software (version 6.02, http://www.cs.cornell.edu/People/tj/svm_light/) was used with default parameters. The SVM^{light} is an intelligible software that allows researchers to implement various kernels such as linear, polynomial, radial and sigmoid kernels (Usmani et al., 2018a).

4.2.9 Feature selection

The optimization of the encoded features is a crucial step in the sequence analyses [43-46]. In this study, a well-established feature dimensionality reduction, GainRatioAttributeEval (GA) of a WEKA software (https://www.cs.waikato.ac.nz/ml/weka/) was used. The GA evaluated the contribution of each feature by measuring the gain ratio with respect to the positive and negative samples. The attribute with a large value of GA is critically responsible for prediction. To select effective feature vectors, we executed multiple rounds of the GA with 10-fold CV test on the training dataset. In this study, it turned out that the GA scheme hardly increased the prediction performance. Therefore, the GA was applied to select the vital features and to deduce the supremacy of the DPC and TPC encoding schemes.

4.2.10 Combined model

To enhance the performance of the predictor, a linear regression model was used to combine the RF and SVM scores for AAindex, BE, DPC, and TPC, respectively. The models that combine the RF and SVM scores for the four descriptors were named RF-iAntiTB and SVM-iAntiTB, respectively, as follows.

$$\mathrm{RF} - \mathrm{iAntiTB} = \mathrm{AAindex}_{\mathrm{RF}} * w_1 + \mathrm{BE}_{\mathrm{RF}} * w_2 + \mathrm{DPC}_{\mathrm{RF}} * w_3 + \mathrm{TPC}_{\mathrm{RF}} * w_4 \qquad (3)$$

$$\mathrm{SVM} - \mathrm{iAntiTB} = \mathrm{AAindex}_{\mathrm{svm}} * w_5 + \mathrm{BE}_{\mathrm{svm}} * w_6 + \mathrm{DPC}_{\mathrm{svm}} * w_7 + \mathrm{TPC}_{\mathrm{svm}} w_8 \qquad (4)$$

The iAntiTB linearly combined the scores by the RF-iAntiTB and the RF-iAntiTB, as follows.

$$\mathrm{iAntiTB} = \mathrm{RF} - \mathrm{iAntiTB} * w_9 + \ \mathrm{SVM} - \mathrm{iAntiTB} * w_{10} \qquad (5)$$

Weight coefficients: w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9, and w_{10} are adjusted from 0 to 1 with an interval of 0.05.

4.2.11 Performance measurement

To measure the prediction performances, accuracy (Ac), sensitivity (Sn), specificity (Sp), and Matthews's correlation coefficient (MCC) were employed as follows:

$$\mathrm{Ac} = \frac{\mathrm{n(TP)} + \mathrm{n(TN)}}{\mathrm{n(TP)} + \mathrm{n(FP)} + \mathrm{n(TN)} + \mathrm{n(FN)}} \qquad (6)$$

$$\mathrm{Sn} = \frac{\mathrm{n(TP)}}{\mathrm{n(TP)} + \mathrm{n(FN)}} \qquad (7)$$

$$\mathrm{Sp} = \frac{\mathrm{n(TN)}}{\mathrm{n(TN)} + \mathrm{n(FP)}} \qquad (8)$$

$$\mathrm{MCC} = \frac{\mathrm{n(TP)} \times \mathrm{n(TN)} - \mathrm{n(FP)} \times \mathrm{n(FN)}}{\sqrt{[\mathrm{n(TN)} + \mathrm{n(FN)}][\mathrm{n}(\mathrm{TP}) + \mathrm{n}(\mathrm{FP})][\mathrm{n}(\mathrm{TN}) + \mathrm{n}(\mathrm{FP})][\mathrm{n(TP)} + \mathrm{n(FN)}]}} \qquad (9)$$

where n(TP) characterizes the number of accurately anticipated anti-TB peptides, n(TN) represents that of accurately anticipated non-anti-TB peptides, n(FP) that of incorrectly predicted anti-TB peptides, and n(FN) that of incorrectly anticipated non-anti-TB peptides. Moreover, we measured the area under the ROC curve (AUC).

Meanwhile, the equilibrium between the anticipated anti-TB and non-anti-TB are analytically liable for precise estimation, Sn and Sp are inherent, comprehensible procedures. Generally, Sp increases with a decrease in Sn. The result of the predictors was assessed with a stepwise adjustment by Sp on the training dataset. We changed an Sp threshold level to understand how accurately anti-TB peptides are identified. In the first dataset, the Sp was set to 0.913, 0.851, and 0.756 for the high, moderate, and low levels by using threshold values of 0.3, 0.25, and 0.19, respectively. In the second dataset, the Sp was set to 0.960, 0.869, and 0.814 for the high, moderate, and low levels by using the threshold values of -0.05, -0.08, and -0.13, respectively. Details in threshold selection strategies are described in our previous study (Khatun, 2019).

4.2.12 iAntiTB web implementation

A user-friendly and publicly accessible web-server was established to implement the iAntiTB. Users submit an anti-TB peptide of interest to a query box, then the server returns the prediction consequence to the output webpage that contains the combined probability scores, job ID, and prediction decisions. Users retrieve the job ID for a next inquiry. The server keeps this ID for 30 days.

4.3 Results and Discussion

4.3.1 Analysis of anti-TB peptides

It is important to examine the sequence preference of positive and negative samples. First, we generated the two sample sequence logos with respect to the first and second datasets

[28]. As shown in **Figure 4-2A**, we found a difference in sequences between the positive and negative samples in the first dataset. In the positive peptide samples, the Lysine (K) residues at positions 1 and 7, Tryptophan (W) at positions 2, and Leucine (L) at positions 5, 6, 8, 12, and 20 were enriched. On the other hand, there were no significantly enriched amino acids at positions 3, 9, 11, 13, 15, 17, and 18; there were no depleted amino acids at positions 2, 8, 9, and 12. These observations suggested that positive and negative samples have distinct location-specific differences. In **Figure 4-2B**, there was different sequence information between the positive and negative samples in the second dataset. The K at position 1, 7, 9, 10, 14, and 19, Phenylalanine (F) at position 5 and 9, Cysteine (C) at position 2 and 13, Histidine (H) at position 13 and 18, and W at positions 2 and 11 were significantly enriched. There were no enriched amino acids at positions 4, 8, 12, 15, and 20 and no depleted amino acids at positions 1, 3, 13, 14, and 19. These analyses suggested that the positive and negative samples have distinct location-specific differences. The positional and frequency-wise methods were found important to identify anti-TB peptides. The amino acid residues of K, W, F, L, and K were enriched at the same positions 1, 2, 5, 6, and 7 between the first and second datasets, but the positions of the depleted amino acid residues were not consistent between them. It suggests that the frequently occurring amino acid residues of positive samples are robust with respect to changes in negative samples.

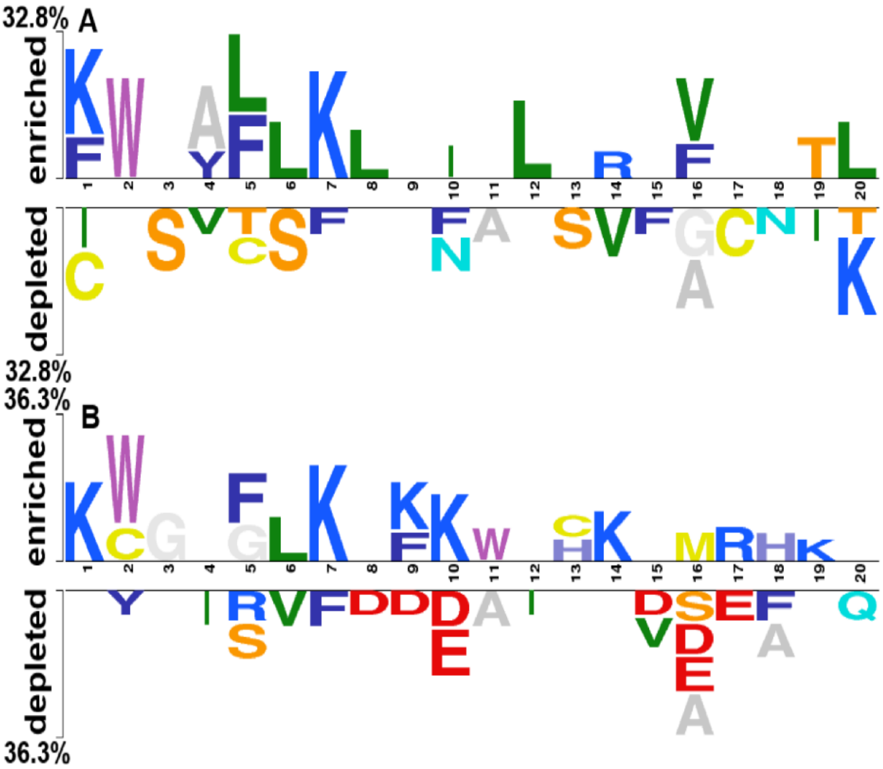

Figure 4-2 Sequence logo representations of anti-TB peptides. The amino acid occurrences are shown for the positive and negative samples. (**A**) The two-sample logos for the first dataset. (**B**) Two-sample logos of the second dataset.

4.3.2 Optimization of peptide length

The peptide length is an important factor of the prediction performance [47,48]. To assess the influence of the adjacent residues, the peptide lengths were optimized using the AUC values. The peptide length was increased from 4 to 24 and encoded by the four consecutive methods of AAindex, BE, DPC, and TPC for the first and second datasets. Then we classified the encoded feature vectors by RF algorithm via 10-fold CV test (**Figure 4-3**). The optimal peptide length 20 was finally selected after several trials of developing the iAntiTB predictor.

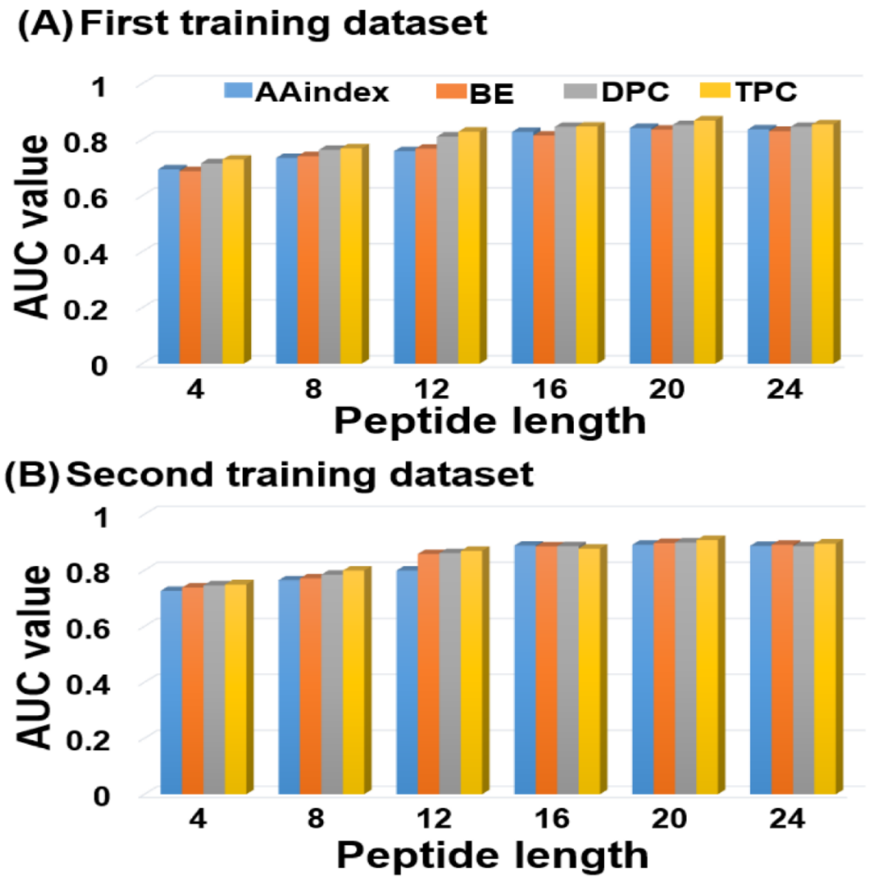

Figure 4-3 AUC value by optimizing of peptide length using the RF classifier via 10-fold CV. (**A**) First training dataset and (**B**) second training dataset.

4.3.3 Evaluation of iAntiTB using the first dataset

To train a predictor, we used the first dataset. We selected four descriptors of AAindex, BE, DPC, and TPC to characterize the samples. The RF and SVM algorithms were employed to explore the features that are correlated with anti-TB in the training dataset. Then, the performance results were assessed by using a 10-fold CV and an independent test via RF and SVM algorithms. In both of the training and independent datasets, **Figure 4-4** depicted the ROC curves for the four single descriptor models and the combined models with the four descriptors (RF-iAntiTB and SVM-iAntiTB). The combined model showed higher AUC than any single descriptor model. For the RF-iAntiTB the AUC value was highest on the training dataset, when the weight coefficient for AAindex, BE, DPC, and TPC were 0.35, 0.15, 0.3, and 0.2, respectively. The AUC was maximal for the SVM-iAntiTB when the weight coefficient for AAindex, BE, DPC, and TPC was 0.05, 0.25, 0.3, and 0.4, respectively. **Table 4-1** shows the prediction performance of the RF-iAntiTB and SVM-iAntiTB on the training and independent datasets. The RF-iAntiTB and SVM-iAntiTB provided high AUC values of 0.887 and 0.849 on the training dataset and 0.882 and 0.871 on the independent one, respectively.

Table 4-1 Prediction performance for the training and independent datasets. T* indicates the training dataset; I* the independent dataset.

Measure	First dataset				Second dataset			
	RF		SVM		RF		SVM	
	T*	I*	T*	I*	T*	I*	T*	I*
Sp	0.915	0.872	0.903	0.875	0.955	0.914	0.947	0.937
Sn	0.699	0.729	0.595	0.660	0.766	0.801	0.708	0.787
Ac	0.807	0.801	0.749	0.768	0.861	0.858	0.828	0.862
MCC	0.628	0.607	0.531	0.548	0.733	0.718	0.678	0.733
AUC	0.887	0.882	0.849	0.871	0.934	0.953	0.907	0.955

The final model of iAntiTB linearly combined the scores of the RF-iAntiTB and SVM-iAntiTB, where the weight coefficient for them were 0.75 and 0.25, respectively. As the iAntiTB with a high threshold of Sp (0.913) showed Sn of 0.707, Ac of 0.810, and MCC of 0.636. A moderate threshold of Sp (0.851) provided Sn of 0.759, Ac of 0.804, and MCC of 0.599; a low threshold of Sp (0.756) showed Sn of 0.793, Ac of 0.775, and MCC of 0.492. The iAntiTB presented high values to all the measures on the independent dataset (**Table 4-2**). In summary, the iAntiTB presented high prediction performance on both the training and independent samples of the first dataset.

4.3.4 Evaluation of iAntiTB using the second dataset

By using the second dataset, we evaluated the robustness of the iAntiTB with respect to change in negative samples. The ROC curves for the four single descriptor models and the combined models with the four descriptors (RF-iAntiTB and SVM-iAntiTB) were plotted on both the training and independent datasets (**Figure 4-5)**. The combined models showed higher AUC than any single descriptor model. For the RF-iAntiTB the AUC value was highest on the training dataset, when the weight coefficient of AAindex, BE, DPC, and TPC were 0.15, 0.2, 0.3, and 0.35, respectively. The AUC was maximal for the SVM-iAntiTB when the weight coefficient of AAindex, BE, DPC, and TPC was 0.2, 0.3, 0.4, and 0.1, respectively. The AUC values of the RF-iAntiTB and SVM-iAntiTB on the training dataset were 0.934 and 0.907, respectively (**Figure 4-5AC)**. The AUC values of the RF-iAntiTB and SVM-iAntiTB on the independent dataset were 0.953 and 0.955, respectively (**Figure 4-5BD)**. The detailed performances of the RF-iAntiTB and SVM-iAntiTB are shown in **Table 4-1**.

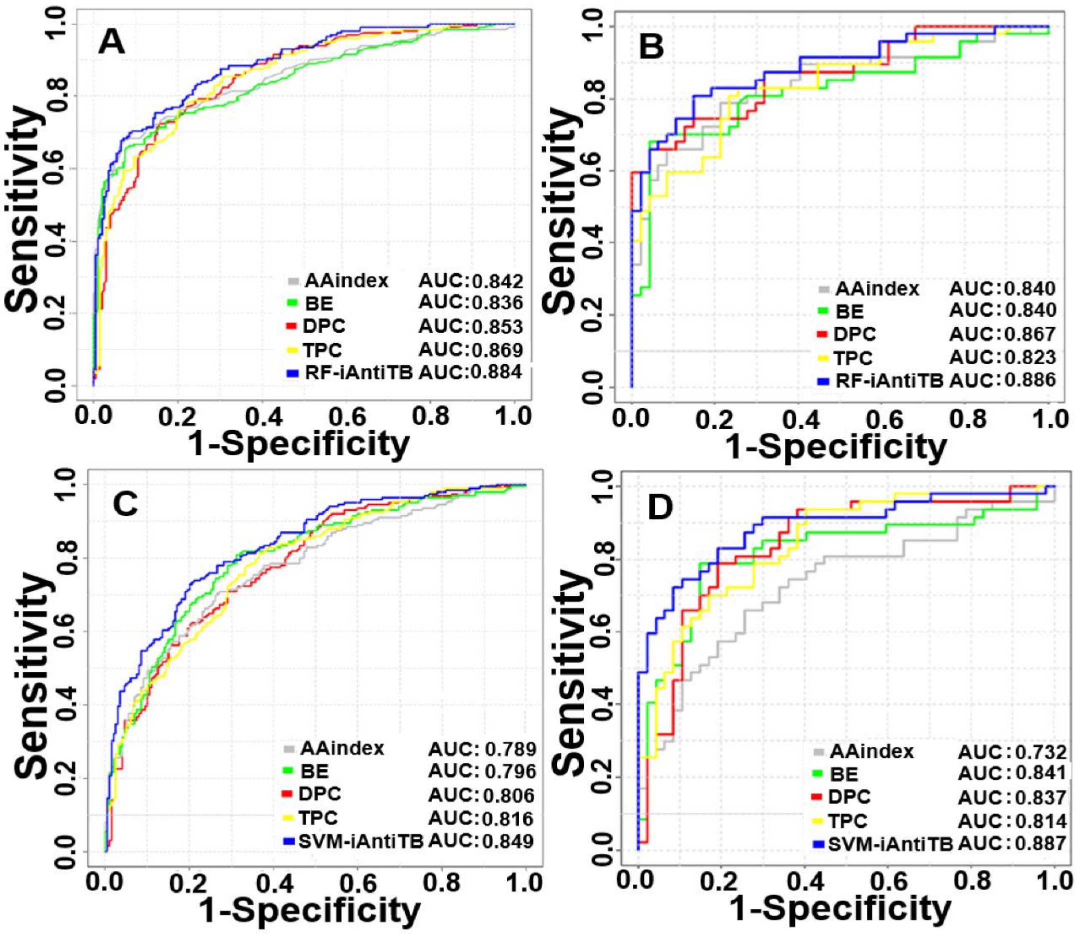

Figure 4-4 ROC curves of anti-TB peptide prediction on the first dataset. (**A**) The RF classifier is applied to the training dataset. (**B**) The RF classifier is applied to the independent dataset. (**C**) The SVM classifier is applied to the training dataset. (**D**) The SVM classifier is applied to the independent dataset.

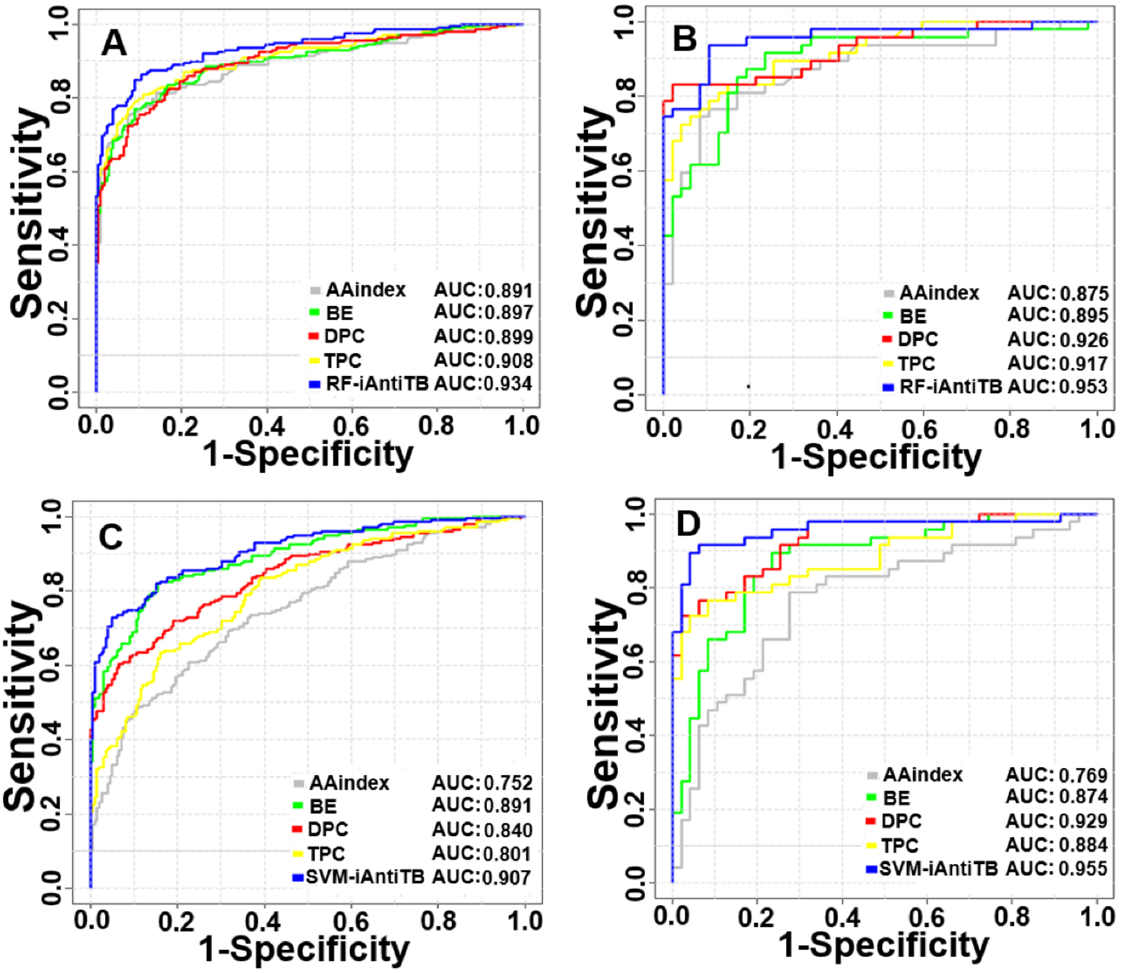

Figure 4-5 ROC curve of anti-TB peptide prediction on the second dataset. (**A**) The RF classifier is applied to the training dataset. (**B**) The RF classifier is applied to the independent dataset. (**C**) The SVM classifier is applied to the training dataset. (**D**) The SVM classifier is applied to the independent dataset.

Table 4-2 Frequently occurring DPCs and TPCs and their corresponding feature selection scores on the training data.

Top 20 features	DPC for first dataset		DPC for second dataset		TPC for first dataset		TPC for second dataset	
	Score	DPC	Score	DPC	Score	TPC	Score	TPC
1	0.1703	IW	0.2094	VM	0.1612	LKK	0.192	KKL

2	0.158	RT	0.1791	IE	0.1511	WWK	0.176	LKK
3	0.1511	AI	0.1703	IW	0.1436	KKW	0.1732	RWF
4	0.1436	KC	0.1673	VY	0.1436	ALA	0.1643	HWR
5	0.1395	LG	0.1673	VT	0.1436	VKG	0.1643	RRW
6	0.1395	II	0.1643	CA	0.1395	AGK	0.1612	RWR
7	0.1395	WY	0.1633	HT	0.1395	RVC	0.1546	KWW
8	0.1351	MK	0.1511	AI	0.1351	KRW	0.1511	WKW
9	0.1351	LS	0.1511	HY	0.1351	KWW	0.1511	KWL
10	0.1351	HS	0.1511	DE	0.1351	QKL	0.1511	KCK
11	0.1351	CA	0.1511	NW	0.1351	RIK	0.1475	WRR
12	0.1334	ND	0.1475	DN	0.1351	KFK	0.1475	VDY
13	0.1302	LW	0.1475	WY	0.1351	KWL	0.1436	KKW
14	0.1302	KM	0.1475	KY	0.1351	VNY	0.1436	LRG
15	0.1302	IC	0.1475	NG	0.1334	WWW	0.1436	VCR
16	0.1302	CW	0.1475	NC	0.1302	RWR	0.1395	IKK
17	0.1302	VT	0.1436	TY	0.1302	RWF	0.1395	WRK
18	0.1302	CE	0.1436	II	0.1302	RRK	0.1395	WRW
19	0.1302	TR	0.1436	YG	0.1302	YQG	0.1395	KFK
20	0.1302	CL	0.1395	CE	0.1302	QFG	0.1351	LAK

For example, the feature 'NxxE' represents a 2-spaced residue (any amino acid) pair of 'NE', where x stands for any amino acid. The same representation was applied to other k-spaced residue pairs.

The final model of the iAntiTB linearly combined the RF-iAntiTB and SVM-iAntiTB scores with weight coefficients of 0.35 and 0.65, respectively. The iAntiTB with a high threshold of Sp (0.960) showed Sn of 0.745, Ac of 0.853, and MCC of 0.721 (**Table 4-3**). A moderate threshold of Sp (0.869) provided Sn of 0.835, Ac of 0.851, and MCC of 0.705, while a low Sp threshold of 0.814 showed Sn of 0.880, Ac of 0.847, and MCC of 0.696. In the iAntiTB, the AUC value was 0.946, while the AUCs of the RF-iAntiTB and SVM-iAntiTB were 0.934 and 0.907, respectively. The Sn, Sp, and MCC for high, moderate, and low values were also evaluated on the independent dataset in **Table 4-3**. Altogether, the iAntiTB presented robust performances to the second dataset.

Table 4-3 Prediction performances of the iAntiTB at high, moderate, and low thresholds

Dataset	Training dataset						Independent dataset				
	Threshold	Sp	Sn	Ac	MCC	AUC	Sp	Sn	Ac	MCC	AUC
First dataset	High	0.913	0.707	0.810	0.636	0.896	0.851	0.750	0.800	0.604	0.913
	Moderate	0.851	0.759	0.804	0.599	0.896	0.809	0.771	0.789	0.580	0.913
	Low	0.756	0.793	0.775	0.492	0.896	0.745	0.813	0.779	0.559	0.913
Second dataset	High	0.960	0.745	0.853	0.721	0.946	0.875	0.936	0.905	0.812	0.959
	Moderate	0.869	0.835	0.851	0.705	0.946	0.833	0.957	0.895	0.796	0.959
	Low	0.814	0.880	0.847	0.696	0.946	0.771	0.964	0.868	0.740	0.959

The performances of the iAntiTB were computed using threshold values of 0.3, 0.25, and 0.19 for the first dataset and -0.05, -0.08, and -0.13 for the second dataset. A 10-fold CV of training test was employed.

4.3.5 Significant features of DPC and TPC

Firstly, to explore the most significant residues of the DPC and TPC encoding schemes, the top 20 features were collected by the GA feature selection scheme from the first dataset. The significant residue sequences and the scores with their corresponding positions are listed in **Table 4-2**. A radar diagram shows the significant residue sequences as shown in **Figure 4-6A**. For the DPC methods on the first dataset, 'IW' represented the most important residue pair and was enriched in the positive samples. The pair of ‘RT’ signified the second enhanced motif adjoining negative samples of anti-TB peptides. Likewise, the top 20 important features were collected from the TPC method from the first dataset. The amino acid residues of ‘LKK’ was most enriched in the radar diagram (**Figure 4-6C**). The average of the top 20 features between two samples was found significant (by a two-sample paired *t*-test with a *p*-value of <0.05) (**Table 4-4**), signifying the efficiency of the DPC and TPC encodings on the first dataset.

Table 4-4 Statistical significance (p-value) of the top 20 features by using the two-sample paired *t*-test between positive and negative samples on the training data.

Top 20 features	DPC for first dataset	DPC for second dataset	TPC for first dataset	TPC for second dataset
1	2.25E-03	8.92E-03	1.27E-04	1.94E-06
2	6.91E-04	9.01E-03	1.41E-03	2.90E-05
3	2.09E-03	4.16E-02	4.41E-03	2.67E-05
4	4.41E-03	7.35E-05	1.01E-02	1.47E-04
5	7.82E-03	3.91E-04	1.72E-02	2.27E-04
6	4.45E-02	2.26E-03	7.82E-03	2.59E-04
7	7.02E-03	2.82E-02	2.78E-02	8.04E-04
8	3.70E-02	4.28E-02	1.39E-02	1.41e-03
9	1.39E-02	1.85E-02	3.93E-02	2.09E-03
10	1.11E-02	1.25E-03	4.29E-02	1.01E-02
11	2.49E-03	2.09E-02	2.93E-03	2.50E-03
12	4.52E-02	1.13E-08	1.42E-02	3.61E-03
13	5.22E-02	5.75E-06	3.42E-02	4.41E-03
14	2.25E-03	4.07E-02	3.11E-02	6.41E-03
15	1.22E-02	1.67E-02	1.29E-02	2.63E-04
16	2.26E-05	1.24E-02	1.02E-02	7.82E-03
17	8.32E-03	9.62E-04	1.41E-03	1.11E-02
18	3.26E-02	2.61E-02	7.29E-02	7.82E-03
19	6.05E-03	3.17E-03	4.24E-03	3.12E-03
20	5.64E-03	8.92E-03	2.49E-02	1.92E-02

Secondly, the top 20 significance features were collected from the second dataset by using the DPC and TPC schemes (**Table 4-2**). The DPC of ‘VM’ was most enriched in the positive samples of the first dataset (**Figure 4-6B**). The TPC of ‘KKL’ was most enriched in the positive samples. The collected features by using the DPC and TPC schemes were significantly validated by a two-sample pair *t*-test for the second dataset (**Table 4-4**), suggesting the effectiveness of DPC and TPC encoding schemes. The enriched residues were estimated to play an important role in identifying anti-TB peptides. Moreover, we observed several common DPCs of ‘IW’, ‘CA’, ‘II’, ‘AI’, ‘VT’, and ‘CE’ in both the first and second datasets. In the TPC scheme, we found ‘LKK’, ‘RVC’, ‘KRW’, ‘WWK’, and ‘KKW’ are common to both the datasets. The above measurement suggests that the common DPCs and TPCs are involved in the prediction of anti-TB peptides.

Table 4-5 Performances of different anti-TB peptide predictors on the training dataset.

Predictor	**First dataset**					**Second dataset**				
	Sp	Sn	Ac	MCC	AUC	Sp	Sn	Ac	MCC	AUC
RF-iAntiTB	0.915	0.699	0.806	0.628	0.887	0.995	0.766	0.860	0.733	0.934
SVM-iAntiTB	0.903	0.595	0.749	0.531	0.849	0.947	0.708	0.828	0.678	0.907
iAntiTB	0.913	0.707	0.808	0.636	0.896	0.960	0.745	0.852	0.721	0.946
AntiTBpred	0.729	0.802	0.766	0.530	0.830	0.862	0.688	0.775	0.560	0.850

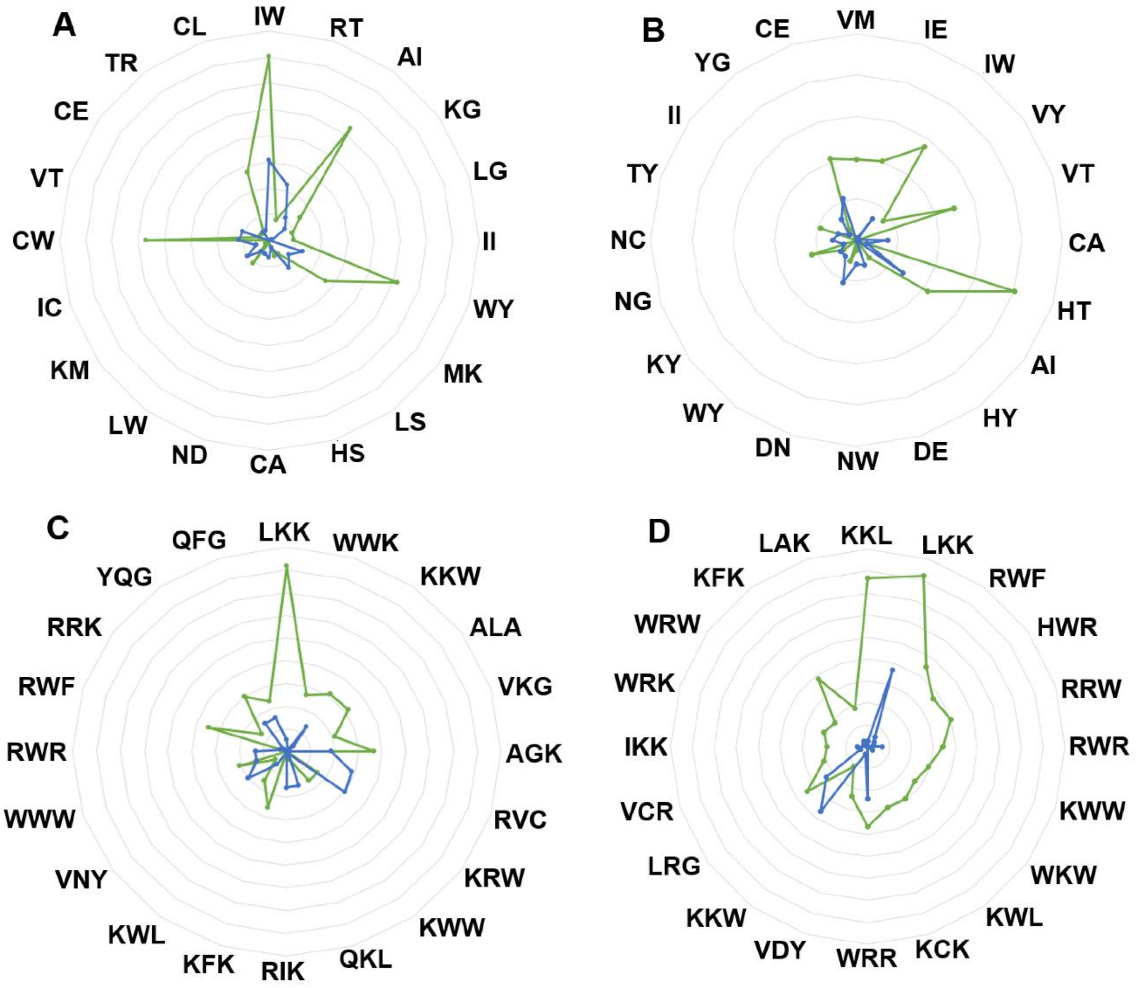

Figure 4-6 Top 20 amino acid residues selected by the GA feature selection method. Green color denotes anti-TB peptides, while blue color denotes nonanti-TB peptides. The radar diagrams of AB and CD are represented with respect to the DPCs and TPCs, respectively. (**A**) Frequently occurring DPCs in the first dataset. (**B**) Frequently occurring DPCs in the second dataset. (**C**) Frequently occurring TPCs in the first dataset. (**D**) Frequently occurring TPCs in the second dataset.

4.3.6 Comparison performance of iAntiTB with AntiTBpred

To make a fair comparison with the existing predictor AntiTBpred, we used two types of training samples (Materials and Methods). AntiTBpred predictor reserved all samples as a positive and negative samples without considering training and independent sets. Since the developers of the AntiTBpred did not separate the training and independent samples, the performance comparison on the independent set was not reasonable. We directly assessed the AntiTBpred performance according to their original literature. As shown in **Table 4-5**, the proposed iAntiTB predictors achieved much higher AUCs on the training sets of the first and second datasets than the AntiTBpred.

4.3.7 Advantages of iAntiTB

Assessment of the iAntiTB in comparison to the current predictor antiTBpred is abridged in a theoretical viewpoint. Firstly, the iAntiTB employed the AAindex, BE, DPC, and TPC, while the antiTBpred used the amino acid composition, DPC, and binary profiles via N5C5 encodings. Secondly, the iAntiTB combined the RF-iAntiTB and SVM-iAntiTB scores via a linear regression model, while the antiTBpred did not consider any combined one. Thirdly, the iAntiTB controlled a threshold value of Sp to understand which peptides contribute to prediction of anti-TB peptides, while the antiTBpred did not control the Sp level. Finally, the iAntiTB investigated the residues critically responsible for the anti-TB peptide prediction, while the antiTBpred did not illustrate them.

4.4 Summary of chapter 4

In this chapter, author develop developed the iAntiTB to accurately predict anti-TB peptides by integrating the four descriptors through RF and SVM algorithms. To characterize the significant features, a feature selection analysis was carried out to facilitate the explaining and understanding of our prediction model. The iAntiTB is a promising computational predictor that outperforms the existing one. A web-application of the iAntiTB is presented for the public to facilitate drug discovery.

CHAPTER 5

CHAPTER 5 PREDICTION OF LINEAR B-CELL PEPTIDES BY INTEGRATING SEQUENCE AND EVOLUTIONARY FEATURES

5.1 Introduction

B-cell peptide or epitopes (BCEs) are specific regions of immunoglobulin molecules that can stimulate the immune system(Wang, 2020; Wang et al., 2020a; Wang et al., 2020b; Wang et al., 2020c; Yan et al., 2020; Yang et al., 2020; Yao et al., 2020; Yi et al., 2020; Zhang et al., 2020), which contributes to diagnostic test, antibody production, and vaccine design (El-Manzalawy et al., 2008; Tomar and De, 2010; Yang and Yu, 2009). B cells are activated by BCEs to perform a variety of biological functions (Groell et al., 2018; Tomar and De, 2010). Identification of BCEs is challenging but crucial for immunotherapy and immunodiagnostics (Guedes et al., 2018; Ma et al., 2018; Mangsbo et al., 2018; Yi et al., 2017). Nowadays, biopharmaceutical research and development of peptide-based antibodies are growing up due to their high efficiency, biosafety, and acceptability (Kang et al., 2019; Kozlova et al., 2018; Olvera et al., 2020; Peng et al., 2020; Poretsky et al., 2020; Rahman Kh et al., 2016; Rao et al., 2020; Usmani et al., 2018b). Thus, the analysis of BCEs is prerequisite for the development of penetrating diagnostic tests and design of the operative vaccines.

BCEs are categorized into two groups: continuous and discontinuous ones (Barlow et al., 1986; El-Manzalawy et al., 2008; Langeveld et al., 2001). Peptides in the continuous group, called linear BCEs, consists of consecutive amino acids. Discontinuous peptides are provided in the form of spatially folded polypeptides and their antigen-binding residues are scattered in their amino acid sequences, making it hard to find them from the primary sequences [21]. To identify the discontinuous peptides, it is necessary to consider many factors such as biochemical properties and structural proximity (Gao et al., 2012; Liang et al., 2009; Sweredoski and Baldi, 2008). Despite the complex form of the discontinuous peptides, they are less effective diagnostic/treatment tools than continuous ones (Kozlova et al., 2018). Linear BCEs have vast application in the area of vaccine design, immunodiagnostic test, antibody production, as well as disease diagnosis and therapy (Bi et al., 2017; Bryson et al., 2010; Chen and Chang, 2017; El-Manzalawy et al., 2017; Khairy et al., 2017; Steere et al., 2011; Sweredoski and Baldi, 2009; Wang et al., 2018; Yu et al., 2016). Given experimental identification of BCEs is labor intensive and costly, computational identification of BCEs has gained remarkable interest recently (Balachandran Manavalan1 and Lee, 2018; Gupta et al., 2013; Jespersen et al., 2017; Saha and Raghava, 2006; Wang and Pai, 2014). Several computational approaches have been developed to predict BCEs, which can be categorized into local and global predictors. Local predictors, such as BepiPred (Jespersen et al., 2017), Bcepred (Saha and Raghava, 2007), and COBEpro (Sweredoski and Baldi, 2009), explore some potential BCE encoding sequences from given protein sequences. These local methods aim to identify the regions or stretchs of proteins that form BCEs [31], but it is difficult to specify the exact regions. Global predictors, such as iBCE-EL (Balachandran Manavalan1 and Lee, 2018), IgPred (Gupta et al., 2013), ABCpred (Saha and Raghava, 2006), SVMTriP (Yao et al., 2012), and LBtope (Singh et al., 2013), determine whether a given sequence is a BCE or not. Since the number of BCEs have rapidly increased in the immune peptide database (Vita et al., 2018), global methods gain attention as the classifier of BCEs. Two global methods, LBtope and iBCE-EL, have recently been developed and publicly available (Balachandran Manavalan1 and Lee, 2018; Singh et al., 2013). These two predictors exclusively investigated primary sequence-based features, such as amino acid composition, binary properties, and physicochemical properties, but did not consider any evolutionary information. Therefore, advanced analytic tools for identifying linear BCEs are still desirable.

In this work, we have established a computational, global predictor named Identification of Linear B-cell Peptide (iLBE) by integrating sequence and evolutionary features. For evolutionary features, we considered the position-specific scoring matrix

(PSSM) and composition of profile-based amino acids frequency (PKAF) encoding descriptors. For primary sequence features, we considered amino-acid index property (AIP) and amino acid frequency composition (AFC). To optimize the consecutive feature vectors, a non-parametric Wilcoxon-rank sum (WR) test was employed. Then the random forest (RF) algorithm using the optimal consecutive feature vectors was used to identify linear BCEs. By the combination of the RF scores through logistic regression (LR), the iLBE yielded better performance than other predictors. Finally, we implemented iLBE as a user-friendly web application. The computational outline of the iLBE is shown in **Figure 5-1**.

5.2 Materials and Methods

5.2.1 Dataset preparation

Experimentally well-characterized datasets of BCEs are needed to develop an accurate machine learning (ML) classifier. We pulled an experimental dataset of linear peptides from the Immune Peptide Database (IEDB), which consists of the verified positive samples (BCEs) and negative samples (non-BCEs) (Schisler and Palmer, 2000; Vita et al., 2015). The IEDB integrates multi-species datasets derived from virus, bacteria, and fungi. We removed homolog sequences from these collected datasets. To evaluate the potential over-fitting problem in the prediction model, a 70% sequence homology reduction method of CD-HIT was performed (Huang et al., 2010). To make a fair comparison with other methods available, the same training and independent samples were retrieved from a recent study (Balachandran Manavalan1 and Lee, 2018). The training model contained 4440 BCEs and 5485 non-BCEs, whereas the independent dataset consisted of 1110 BCEs and 1408 non-BCEs. To avoid the prediction biases, a none-redundant dataset of experimentally validated BCEs and non-BCEs was used, and the samples with more than 70% sequence similarity were excluded. In this study, the peptide length of BCEs and non-BCEs was set to 24. When the length of positive and negative peptide samples was < 24, the null residues (gaps) were added downstream. The curated datasets are shown in our web server and a statistics of the curated dataset is included in **Table 5-1**.

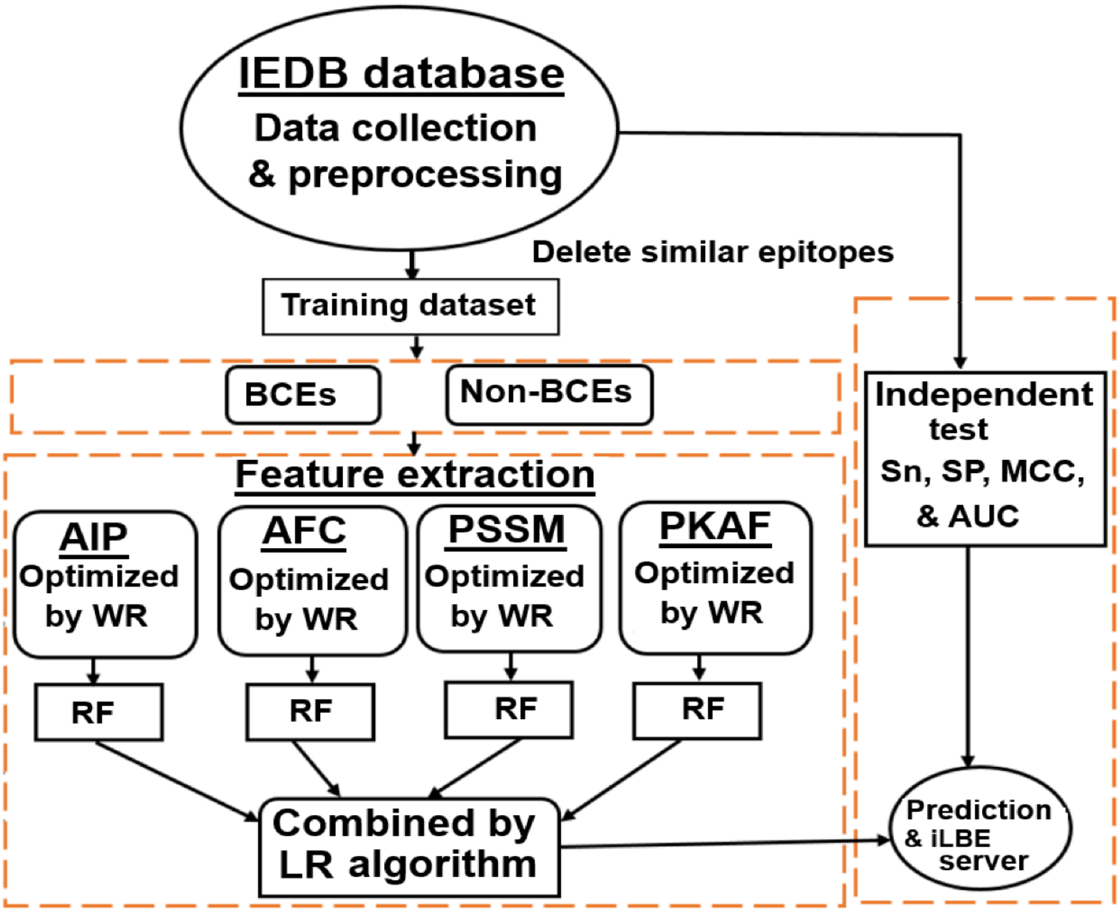

Figure 5-1 Overview of the iLBE.

Table 5-1 Statistics of the datasets used in this study

Length of epitope	**Training set**		**Independent set**	
	BCE	Non-BCE	BCE	Non-BCE
7-12aa	478 (10.77%)	372 (6.78%)	129 (11.62%)	115 (8.17%)
13-20aa	3,465 (78.04%)	4,910 (89.52%)	870(78.38%)	1,215 (86.29%)
21~	497 (11.19%)	203 (3.7%)	111 (10%)	78 (5.54%)
Total	4,440 (100%)	5,485 (100%)	1,110 (100%)	1,408 (100%)

The parentheses represent the percentages of epitopes.

5.2.2 Feature encoding strategies

PSSM profile

The PSSM profile was generated using the PSI-BLAST (a version of 2.2.26+) with the whole Swiss-Prot non-redundant-protein database (a version of December 2010). We used two onset parameters: an iteration times of 3 and e-value cutoff of 0.0001 (Hasan et al., 2017b; Hasan et al., 2018d). The feature vectors were extracted based on the sequence of BCEs and non-BCEs. For each peptide sequence with length 24, an (24 × 20) dimensional vector was generated via the PSSM encoding. When the query peptide length is < 24, zero was added downstream of each PSSM to neutralize the null residues.

PKAF encoding

After generating the PSSM profile, we generated PKAF feature vectors (Dong et al., 2013; Hasan et al., 2015). In brief, if the residue pair appears between *m* and *m+k+1*, the composition scores were measured or standardized by the following formula:

$$S_{ij} = \frac{\sum_{i,j=1}^{T} \max[\min\{\text{PSSM}(m, x_i), \text{PSSM}(m+k+1, x_j)\}, 0]}{W-1} \tag{1}$$

where *W* is the peptide length of BCEs, a *k*-spaced residues characterized as $x_i\{k\}x_j$ (*i, j*= 1, 2, …, 20) represent 20 types of common residues, and *T* means that $x_i\{k\}x_j$ performs *T* times for the positive /negative samples. PSSM (m, x_i) signifies the score of amino acid x_i at m^{th} row in $x_i\{k\}x_j$, and PSSM (m+k+1, x_j) indicates the score of residue x_j at the row of $(m+k+1)^{\text{th}}$. An optimum value of *k* is 0 or 1, and the dimension of PKAF was 800.

In addition, we employed a similarity-search-based tool of BLAST (version of ncbi-blast-2.2.25+) to examine whether a query peptide belongs to BCEs or not (Altschul et al., 1997; Whelan et al., 2013). An e-value of 0.01 via BLASTP was used for the whole Swiss-Prot non-redundant90 database (version of December 2010).

AIP encoding

The AIP database (a version of 9.1) contained numerical indices of biochemical and physicochemical properties of amino acids (Kawashima et al., 2008). With assessing various types of indices, we measured 8 types of high informative indices, including

NAKH920108, CEDJ970104, LIFS790101, BLAM930101, MAXF760101, TSAJ990101, NOZY710101, and KLEP840101. To produce the feature vectors, the selected AIP properties were transformed into the BCEs and non-BCEs. A null residue was used to fill the gap and pseudo residues. In a peptide sequence with length W, an (W × 8) dimensional vector was generated via the AIP encoding.

AFC encoding

The AFC encoding is widely used for representing short sequence peptide motifs [21,24]. The procedure of AFC is briefly described as follows. When a peptide is composed of 20 types of common residues, it contains $(\mathrm{AA, AC, AD, \ldots, YY})_{400}$ types of residue pairs. An optimal value of k, which signifies the frequency of any two-amino acid pairs, was set to 0 or 1. Consequently, 20 × (k+1) × 20 = 800 distinguished residue pairs were generated. The feature vector was then calculated and standardized by the following formula:

$$\left(\frac{N_{AA}}{N_{total}}, \frac{N_{AC}}{N_{total}}, \;\cdots\;, \frac{N_{YY}}{N_{total}}\right)_{400} \qquad (2)$$

where N_{total} is the length of peptide in the total composition residues. If peptide length W is 24 and k is 0 or 1, then $N_{\mathrm{total}} = W{-}k{-}1$ is 23 or 22, respectively. ($N_{AA}, N_{AC}, \ldots, N_{YY}$) represents the frequency vector of amino acid pairs within the BCEs and non-BCEs.

5.2.3 Feature selection

Uncorrelated and redundant features may exist in the generated feature vectors, which can affect the accuracy of a prediction model (Hasan et al., 2017b). Hence, feature selection approaches are important to collect the informative features and to characterize the intrinsic properties of BCEs. To characterize the features important for predicting BCEs, a well-established reduction method of feature dimensionality, WR, was used. A large value of the WR specifies that the corresponding residues have a great impact on the prediction performance. Details in the WR scheme are described elsewhere (Hasan et al., 2018d).

5.2.4 Model training and evaluation

To construct a prediction model, an RF classifier was used. It is a supervised ML algorithm and widely used in bioinformatics research (Hasan et al., 2018b; Hasan et al., 2017c; Hasan et al., 2016; Li et al., 2012; Md. Mehedi Hasan, 2017; Pan et al., 2014; Zhao et al., 2018). In brief, the RF is an ensemble of a number of decision trees, H = $\{H_1(S), H_2(S), \ldots, H_N(S)\}$, which are built on *N* random subcategories of the training samples. This forest was trained with the bagging method to build an ensemble of decision trees. The general idea of the bagging method is that learning models are assembled to increase the global performance. Details in the RF algorithm were provided in previous studies (Hasan et al., 2018d; Hasan et al., 2016). The R package was employed to implement the RF into the proposed iLBE (https://cran.r-project.org/web/packages/randomForest/).

Three commonly used ML algorithms, naive Bayes (NB) (Lowd, 2005), support vector machine (SVM) (Hearst, 1998), and artificial neural network (ANN) (R. S. Michalski 2013), were compared with the RF algorithm. The WEKA software (Frank et al., 2004) was used for the NB and ANN algorithms and the LIBSVM software (https://www.csie.ntu.edu.tw/~cjlin/libsvm/) was used for the SVM algorithm

To construct the final model of iLBE, the respective RF scores evaluated from the four features (PSSM, PKAF, AIP, and AFC) were combined using a LR algorithm. The LR algorithm was effectively used in ubiquitination site prediction (Chen et al., 2015). After examining the performance of the resulting S-prediction models (S is the number of the encoding schemes) the final prediction score P was calculated by:

$$\log\left(\frac{\mathrm{P}}{1-\mathrm{P}}\right) = \sum_{\mathrm{n}=1}^{\mathrm{S}} \beta_n R_n + \alpha \qquad (3)$$

where β_n is the regression coefficient, R_n is the RF score of each feature, and β is the regression constant. The R software package (https://cran.r-project.org/) was employed for a generalized model of LR.

5.2.5 Performance evaluation matrixes

To examine the performance of iLBE, four widely-used statistical measures, represented as sensitivity (Sn), specificity (Sp), accuracy (Ac), and Matthews correlation coefficient (MCC), were defined as:

$$\mathrm{Sn} = \frac{\mathrm{n(TP)}}{\mathrm{n(TP)} + \mathrm{n(FN)}} \tag{4}$$

$$\mathrm{Sp} = \frac{\mathrm{n(TN)}}{\mathrm{n(TN)} + \mathrm{n(FP)}} \tag{5}$$

$$\mathrm{Ac} = \frac{\mathrm{n(TP)} + \mathrm{n(TN)}}{\mathrm{n(TP)} + \mathrm{n(FN)} + \mathrm{n(FP)} + \mathrm{n(TN)}} \tag{6}$$

$$\mathrm{MCC} = \frac{\mathrm{n(TP)} \times \mathrm{n(TN)} - \mathrm{n(FP)} \times \mathrm{n(FN)}}{\sqrt{[\mathrm{n(TN)} + \mathrm{n(FN)}][\mathrm{n}(\mathrm{TP}) + \mathrm{n(FP}\,)][\mathrm{n}(\mathrm{TN}) + \mathrm{n(FP}\,)][\mathrm{n(TP)} + \mathrm{n(FN)}]}} \tag{7}$$

where n(TP), n(TN), n(FP), and n(FN) demonstrate the number of anticipated positive, anticipated negative, unexpected positive, and unexpected negative samples, respectively. Furthermore, we depicted the receiver operating characteristic (ROC) curve (Sn vs. 1-Sp) and measured the area under curve (AUC) values (Centor, 1991; Gribskov and Robinson, 1996).

The prediction performance was assessed using 10-fold cross-validation (CV) test on the training model until no further improvement occurred after each round of optimization parameters. The training dataset was separated into 10 groups, where 9 of the groups were used for training and the remaining one for test. This selection process was repeated 10 times to assess the average performance of the 10 models.

5.2.6 Model development

To develop the prediction model, we first compiled the training and independent datasets in the same manner as described by Manavalan et al. (see Dataset preparation section) (Balachandran Manavalan1 and Lee, 2018). The prediction result was evaluated based on the criterion of whether the indication measure (Sp, Sn, MCC, Ac, or AUC) exceeds a threshold value. The AUC value of the ROC curve was evaluated, with the threshold value of the RF score changed to classify a BCE or non-BCE. The threshold value determines the desirable balance to successfully detect positive and negative BCEs. The true positive rate (Sn) and the false positive rate (1-Sp) were calculated for each threshold value of the RF scores. The high, moderate, and low-level thresholds were determined based on RF scores of 0.485, 0.410, and 0.360, respectively, which corresponded to Sp levels of 0.866, 0.747, and 0.636 in the training set results, respectively.

5.2.7 Web application and implementation

To provide a prediction service of potential BCEs to the scientific community, an accessible web page of the iLBE was established at http://kurata14.bio.kyutech.ac.jp/iLBE/. The web application was written in various programming languages including Perl, R, CGI scripts, HTML, and PHP. The server takes antigen peptides written with 20 types of common amino acids in the FASTA format. When the submission job is completed, the server returns the prediction results with a combined RF score of the predicted BCEs in a tabular form to the output webpage with the job ID and a query peptide. Users can save the ID for a future query and the iLBE server stores this ID for a month.

5.3 Results and discussion

5.3.1 Analysis of positional amino acids

To investigate the sequence preference of BCEs and non-BCEs, we performed amino acid positional analysis using the iceLogo software (Colaert et al., 2009). In the training datasets, 1 to 15 residues were employed to create iceLogos. The average length of the BCE and non-BCEs was set to 15. Significant differences in the surrounding BCEs and non-BCEs were observed by Welch's *t*-test with $p < 0.05$ (**Figure 5-2**). The neutral amino acids P, N, and Y showed a strong preference on BCEs at positions 3, 4, 6, 7, 8, and 10, while amino acids A, H, L, M, and V showed a strong preference for non-BCEs. This analysis supports the idea that different residues are targeted by distinct BCEs, suggesting that combination of different features is critical for accurate prediction of BCEs.

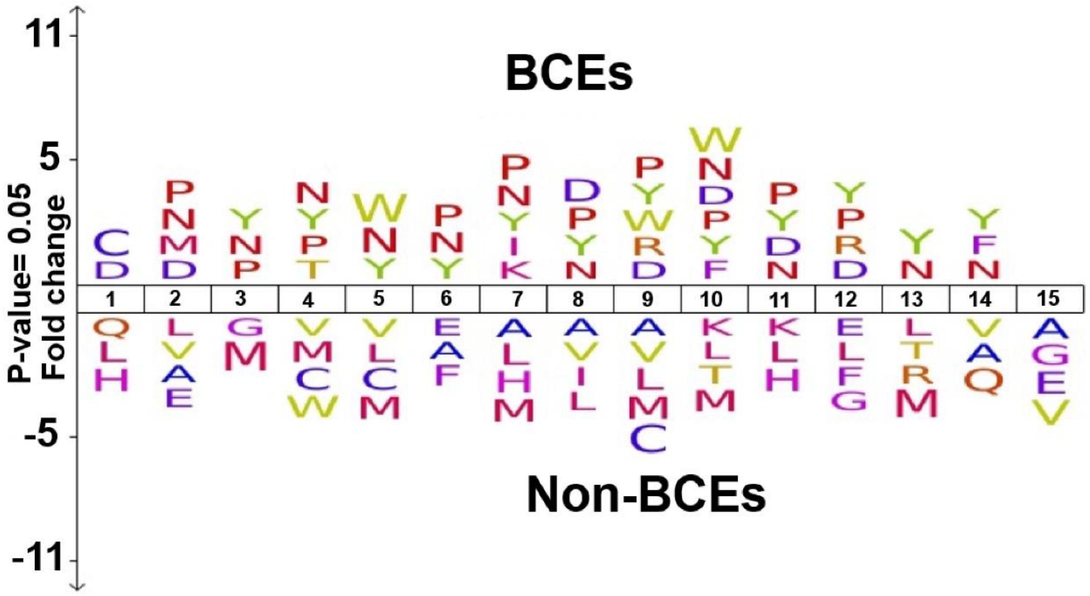

Figure 5-2 Distribution of amino acids of BCEs. The iceLogo software (https://iomics.ugent.be/icelogoserver/) is used. The amino acids show a significantly different distribution between the BCE and non-BCEs (p<0.05).

Table 5-2 Performance comparison among four single feature methods and the combined feature method (iLBE)

Method	Sp	Sn	Ac	MCC	AUC	*p*-value
PSSM	0.703	0.714	0.708	0.368	0.746	0.006
AIP	0.704	0.689	0.697	0.369	0.742	0.006
PKAF	0.705	0.737	0.719	0.429	0.774	0.033
AFC	0.703	0.739	0.719	0.432	0.775	0.038
iLBE	0.747	0.759	0.752	0.496	0.809	

A10-fold CV test was applied to the training dataset. The 2-6 columns represent the prediction performances of the single feature method and the combined method (iLBE). The last column signifies a statistical test based on the AUC measures by a two-tailed t-test, where $p \leq 0.05$ indicates a statistically meaningful difference between the iLBE and each single feature method.

5.3.2 Selection of the optimal model

To inspect the performance of the iLBE, the curated BCE datasets were first coded as mathematical feature vectors based on the four successive encodings of AIP, AFC, PSSM, and PKAF. Given prediction performance may be impaired by uncorrelated and redundant evidence in the curated features, we used the WR method to optimize the feature vectors. After several trials, top 170, 510, 320, and 490 feature vectors were selected from the AIP, AFC, PSSM, and PKAF descriptors, respectively. Then the selected feature vectors were rearranged in the ascending order of WR values. The RF classifiers were trained by using the final four encoding feature vectors. The decision trees of RF were optimized over the training dataset by a 10-fold CV test. Then the RF scores by the PSSM, AIP, PKAF, and AFC encoding methods were combined by the LR scheme with regression coefficients of 0.435, 0.102, 1.337, and 0.465, respectively. As shown in **Table 5-2**, AFC presented a higher performance than any other single encoding approach in terms of Ac, Sn, MCC, and AUC in the training dataset. The combined model of iLBE outperformed all the four single encoding approaches in terms of Sn, MCC, Ac, and AUC. The superiority of iLBE was confirmed to be significant by two-tailed *t*-test.

Table 5-3 Performance comparison of iLBE with existing predictors

Predictors	Threshold	Sp	Sn	Ac	MCC	AUC
LBtope	-	0.672	0.660	0.667	0.330	0.730
iBCE-EL	-	0.739	0.716	0.729	0.454	0.782
	High	0.866	0.568	0.733	0.452	0.809
iLBE	Moderate	0.747	0.759	0.752	0.496	0.809
	Low	0.636	0.838	0.726	0.475	0.809

A10-fold CV test was applied to the training dataset. The performances of the LBtope and iBCE-EL methods were collected according to their published studies. In the proposed iLBE, the high, moderate and low-level thresholds were determined based on the RF scores of 0.485, 0.410 and 0.360, respectively, which corresponded to the Sp levels of 0.866, 0.747 and 0.636 in the training dataset results.

The performances of each single feature vector-trained model and the combined model were evaluated in the training and independent datasets, as shown in **Figure 5-3**. AUCs obtained using iLBE were higher than those obtained using any single feature model for both training and independent datasets, demonstrating the robustness of the iLBE model. Moreover, we also measured the predictive performance based on either sequence or evolutionary features alone for the training and independent datasets (**Table 5-5**). The AUC values of the sequence feature-based methods were at most 0.791 and 0.798 for the training and independent sets, respectively (**Table 5-5**)). Similarly, the AUC values of the evolutionary feature-based methods were at most 0.789 and 0.786 for the training and independent sets, respectively. Neither the sequence nor evolutionary feature-based methods outperformed iLBE, indicating that the combination of the sequence and evolutionary features in iLBE is effective for enhanced prediction accuracy.

In addition, we used BLAST to determine the sequence profile information of BCEs and non-BCEs in the training dataset [40]. In total 1038 BCE and 597 non-BCE samples were selected out of 4440 BCE and 5485 non-BCE samples via the BLASTP with an e-value of 0.01. Then the BLAST performance was evaluated through 10-fold CV test. The Sp, Sn, Ac, MCC, and AUC were 0.811, 0.214, 0.544, 0.042, and 0.569, respectively, which are lower than those of iLBE. Therefore, BLAST was not considered for the final prediction.

Table 5-4 Performance comparison with existing predictors on the independent dataset

Predictors	**Threshold**	**Sp**	**Sn**	**Ac**	**MCC**	**AUC**	***p*-value**
LBtope	-	0.567	0.759	0.615	0.328	0.730	**<0.01**
iBCE-EL	-	0.724	0.742	0.732	0.463	0.786	**<0.05**
	High	0.861	0.554	0.726	0.440	0.813	
iLBE	Moderate	0.745	0.752	0.748	0.494	0.813	
	Low	0.635	0.830	0.721	0.467	0.813	

The high, moderate and low thresholds in the 2nd column were considered based on the training dataset performances. The 8th column represents the statistically significant difference (*p*<0.05) by a paired two-sample *t*-test based on the AUC values between the iLBE and each existing method.

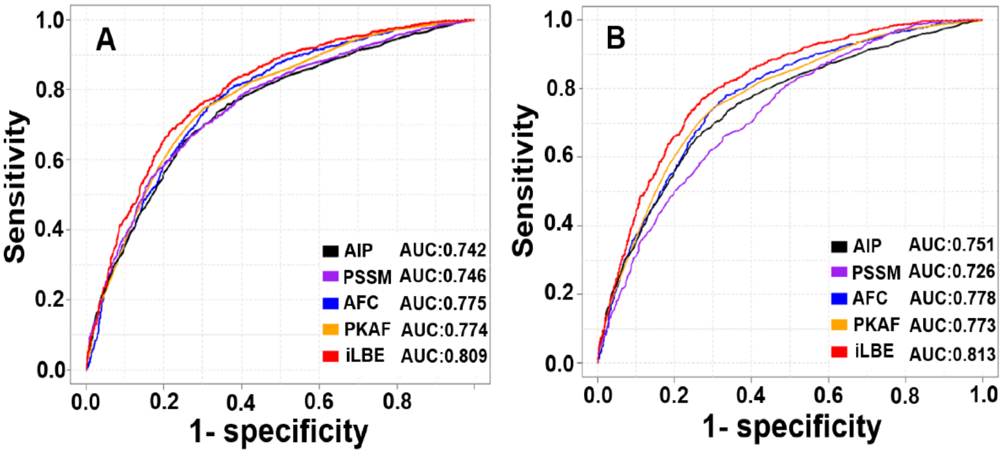

Figure 5-3 ROC curves of the various prediction models. (A) The training dataset. (B) The independent data set. The iLBE is the LR-combined model of the PSSM, AIP, PKAF and AFC encoding schemes. Their LR coefficients are 0.435, 0.102, 1.337, and 0.465, respectively.

We found that the AFC scheme presented the highest AUC, Sp, Sn, Ac, and MCC for all four single encoding methods (**Table 5-2**). To investigate significant residues estimated by the AFC method, the top 25 amino acid pairs were examined through the WR feature selection. The top 25 significant residue pairs and the WR scores were listed in **Table 5-6**. As shown in **Figure 5-4**, the average value of the AFC was measured for the BCE and non-BCE peptides. The selected feature of **LxT** (where 'x' signifies any amino acid) was the most significant residue pair and depleted around non-BCE (*P* = 3.112E–12, *t*-test, **Table 5-6**). Likewise, the feature SP that characterizes a 0-spaced (i.e., there is no space in this case) pair of residues SP is important and enriched in BCEs (**Figure 5-4**; *P* =

2.88E–09, *t*-test, **Table 5-6**). The above similar concept was applied to other selected pairs of residues (**Figure 5-4**). Importantly, the top 25 features contained P, N, and Y residues, which showed strong preference in positional residue analysis (**Figure 5-2**). These residues would play an important role in the recognition of BCEs. Moreover, as shown in **Table 5-6**, the average AFC of top 25 features was significantly different between BCEs and non-BCEs ($P < 0.05$; paired two-sample *t*-test).

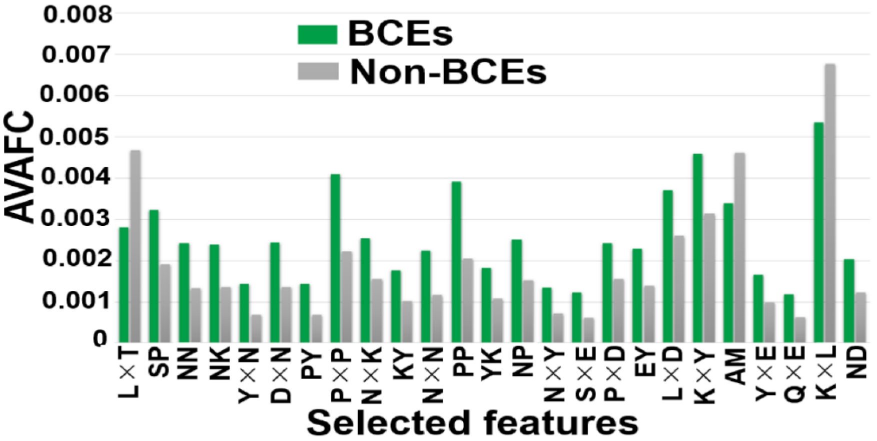

Figure 5-4 The distribution of the top 25 significant features deriving from the AFC scheme. The Y-axis represents the average value of the AFCs for BCEs and non-BCEs. The X-axis represents the selected features.

5.3.3 Optimal length of peptides

To optimize the length of short peptides, we investigated the different lengths (5, 10, 15, 20, or 25 amino acids) of BCEs using the four encoding schemes of AIP, PSSM, AFC, and PKAF and their combined scheme (iLBE) (**Table 5-7**). The RF algorithm without any feature selection approach was used to evaluate prediction performance on the training data via 10-fold CV test. The prediction performance increased with an increase in sequence length, and was saturated for lengths of 20 and 25 (**Table 5-7**). Therefore, a 24 sequence length of 24 was determined for iLBE.

Table 5-5 AUC values of prediction based on sequence or evolutionary methods

Methods		Training dataset	Independent dataset
Sequence-based method	AFC	0.742	0.751
	AIP	0.775	0.778
	AFC+AIP	0.791	0.798
Evolutionary-based method	PSSM	0.746	0.726
	PKAF	0.774	0.773
	PSSM+PKAF	0.789	0.786

5.3.4 Comparison with different ML algorithms

The RF algorithm was characterized in comparison with the widely-used ML algorithms of NB, SVM, and ANN on the same training dataset. AUC values of predictions using the four algorithms without any feature selection were evaluated by 10-fold CV test. As shown in **Table 5-8**, the RF algorithm provided a higher AUC than any other algorithms. Accordingly, we implement the RF algorithm in iLBE.

Table 5-6 Top 25 AFC features ranked by a WR-based selection method

No. of feature	WR feature	*p*-value
1	L×T	3.112E-12
2	SP	2.88E-09
3	NN	4.76E-08
4	NK	1.29E-08
5	Y×N	2.91E-08
6	D×N	9.39E-09
7	PY	9.18E-08
8	P×P	1.28E-08
9	N×K	2.82E-07
10	KY	1.03E-06
11	N×N	6.77E-08

12	PP	1.76E-07
13	YK	2.51E-06
14	NP	6.09E-06
15	N×Y	7.01E-06
16	S×E	4.04E-06
17	P×D	4.17E-06
18	EY	1.28E-06
19	L×D	3.34E-06
20	K×Y	3.068E-06
21	AM	7.75E-06
22	Y×E	9.32E-06
23	Q×E	1.21E-05
24	K×L	1.78E-05
25	ND	3.39E-04

The *p*-values were calculated using a paired *t*-test for the top 25 significant BCEs and non-BCEs.

5.3.5 Comparison of iLBE with existing methodologies

We evaluated the prediction performance of the proposed iLBE with existing approaches on the same dataset. First, we employed the training dataset to compare the performance of iLBE with those of the LBtope and iBCE-EL models, which are the state-of-the-art predictors and publicly accessible. As shown in **Table 5-3**, an increase in Sp decreased Sn for iLBE. iLBE with the moderate threshold showed higher Sp, Sn, MCC, Ac, and AUC than LBtope and iBCE-EL, demonstrating that iLBE outperforms the existing pioneering predictors. Furthermore, we compared the performance of iLBE with that of LBtope and iBCE-EL in the independent dataset (see Method). As shown in **Table 5-4**, an increase in the Sp also decreased the Sn for iLBE in the independent dataset. iLBE with the moderate threshold outperformed the two existing methods in terms of Sp, MCC, and AUC, while it presented almost the same Sn as LBtope. The superiority of iLBE to the existing methods was confirmed to be significant ($P < 0.05$, two sample *t*-test).

Table 5-7 AUC values for different lengths of epitopes

Methods	AIP	PSSM	AFC	PKAF	iLBE
5aa	0.526	0.538	0.546	0.555	0.563
10aa	0.557	0.559	0.579	0.576	0.598
15aa	0.589	0.588	0.663	0.687	0.718
20aa	0.703	0.737	0.765	0.761	0.781
25aa	0.716	0.729	0.758	0.763	0.786

A 10-fold CV test was applied to the training dataset

5.3.6 Effect of combination methods

To investigate the effects of combination methods on the prediction performance, we built a competitive model of iLBE, which arranges the four encoding vectors of AFC, AIP, PSSM, and PKAF in a row, instead of the use of LR. It is named as the sequential combination model. The resultant total dimension was 2192. The top 380 feature vectors were collected and rearranged in the ascending order of WR values. The WR-optimized feature vectors were used to train the RF classifier via 10-fold CV test. The sequential combination model with and without feature collection approaches yielded AUC values of 0.778 and 0.767 on the training dataset, respectively (**Figure 5-5A**), and presented 0.798 and 0.781 on the independent dataset, respectively **(Figure 5-5B**). The LR-based combination of iLBE outperformed the sequential combination model (**Figure 5-3**) and was found to be the best in this study.

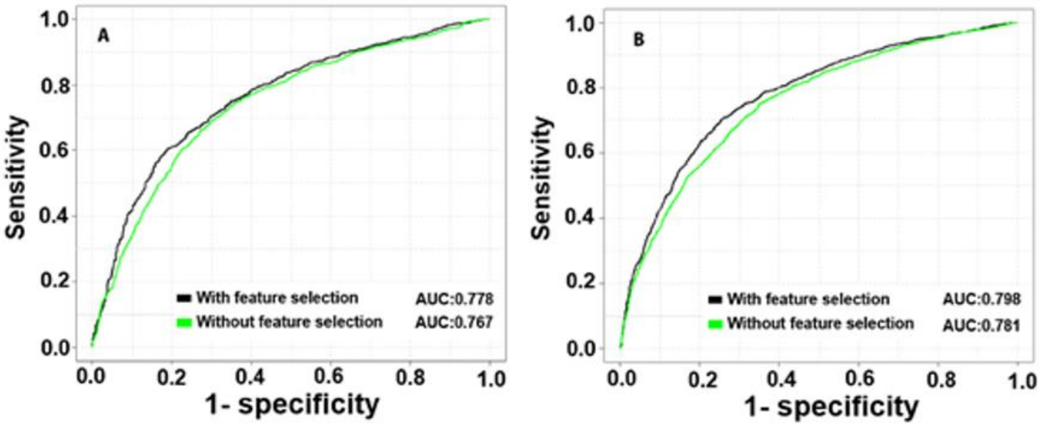

Figure 5-5 ROC curve of the sequential combination model that integrates the feature vectors with and without feature selections approach. A) Training and B) Independent datasets.

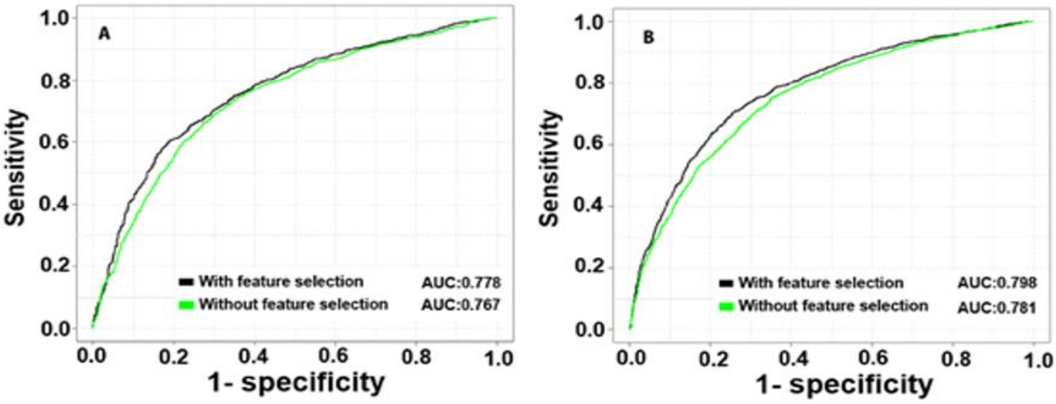

Figure 5-5 ROC curve of the sequential combination model that integrates the feature vectors with and without feature selections approach. A) Training and B) Independent datasets

Table 5-8 AUC values for different ML algorithms

Algorithms	PSSM	AIP	PKAF	AFC	iLBE
NB	0.682	0.717	0.736	0.747	0.756
ANN	0.699	0.711	0.732	0.739	0.743
SVM	0.733	0.721	0.753	0.766	0.774
RF	0.738	0.739	0.768	0.767	0.788

A 10-fold CV test was applied to the training dataset

5.4 Summary of chapter 5

We have developed a novel computational predictor, iLBE, that accurately predicts BCEs for both the training and independent datasets. iLBE outperformed existing state-of-the-art predictors LBtope and iBCE-EL. The iLBE model combined the sequence-based features and evolutionary information, while the LBtope and iBCE-EL predictors only used sequence-based encoding methods. iLBE employed the LR-based combined model of the RF-based classifiers, while LBtope and iBCE-EL used SVM and an ensemble ML model, respectively. Importantly, iLBE allows the use of various threshold values at high, moderate, and low levels to demonstrate whether a BCE is highly positive or negative, which is not available in the existing prediction tools. As a complementary to the experimental strategies, iLBE provides insight into the functional and significant characteristics of BCEs. A user-friendly web-application was also developed for easy use by the immunological research community.

Availability

A web application with curated datasets for iLBE is freely accessible at http://kurata14.bio.kyutech.ac.jp/iLBE/.

CHAPTER 6

CHAPTER 6 CONCLUSIONS AND PERSPECTIVES

6.1 Conclusions

High-throughput omics-based bioinformatics methods have been widely used in the study of biology, resulting in that the need for rigorous, computational analysis of biological data has never been so greater. This thesis focuses on the prediction of four types of protein four types of immune-peptides (anti-inflammatory, pro-inflammatory, anti-tuberculosis, and linear B-cell peptides). At first, a novel predictor termed as PreAIP has been developed for the prediction of pro-inflammatory peptides. The prediction result suggests that the integrating multiple encoding is able to capture important sequence evolutionary information, which plays an important role in the performance improvement. Moreover, a feature selection experiment was performed to characterize the contributive features and facilitate better understanding and interpretation of prediction model. These analyses also demonstrate that the proposed method can be used as a powerful tool for understanding the mechanism of pro-inflammatory peptides. Taken together, these findings suggest that the novel software PreAIP can be served as a powerful tool to help the identification of pro-inflammatory peptides. The web server and curated datasets in this study are freely available at http://kurata14.bio.kyutech.ac.jp/PreAIP/.

Secondly, a novel predictor called ProIn-Fuse has been developed through the integration of different sequence features. The ProIn-Fuse predictor is capable of yielding a high accuracy. Specifically, a feature representation learning model was utilized to generate a set of informative probabilistic features by making the use of random forest models with eight sequence encoding schemes. Then the ProIn-Fuse was constructed by the linearly combined models of the informative probabilistic features. The web server and curated datasets are freely available at http://kurata14.bio.kyutech.ac.jp/ProIn-Fuse/.

Thirdly, an effective computational predictor iAntiTB (Identification of anti-tubercular Peptides) has been developed by the integration of multiple feature vectors deriving from the amino acid sequences via RF and SVM classifiers. The iAntiTB combined the RF and SVM scores via linear regression to enhance the prediction accuracy. To make a robust and accurate predictor we prepared the two datasets with different types of negative samples. The iAntiTB achieved AUC values of 0.896 and 0.946 on the training datasets of the first and second datasets, respectively. The ProIn-Fuse was established by fusing the successive probabilistic scores using a linear regression model. For user community, a free accessible web application of iAntiTB is available at http://kurata14.bio.kyutech.ac.jp/iAntiTB/.

Finally, authors develop a novel predictor, Identification of B-Cell Epitope (iLBE), by integrating evolutionary and sequence-based features for prediction. The successive feature vectors were optimized by a Wilcoxon rank-sum test. Then the RF algorithm using the optimal consecutive feature vectors was applied to predict linear B-cell peptides. We combined the RF scores by the logistic regression to enhance the prediction accuracy. iLBE is a powerful computational tool to identify the linear B-cell peptides and would help to develop penetrating diagnostic tests. A web application of iLBE predictor is available at http://kurata14.bio.kyutech.ac.jp/iLBE/.

6.2 Perspectives

In this thesis the author discussed the machine learning approaches for addressing classification problems of four types of immune-peptides (anti-inflammatory, pro-inflammatory, anti-tuberculosis, and linear B-cell peptides). To assist knowledge discoveries through intensive analysis of vast amounts of immune-peptides, further improvements are required. The followings are important perspectives for immune-peptides prediction.

To further improve the prediction performance, we have the following suggestions. First, to decrease bias in the training dataset, excluding highly homologous sequences is needed. Such a dataset will be helpful for developing more reliable and powerfully trained models. Second, based on our analysis, we observed that feature-encoding approaches converging on position specific information and profile-based information may be very suitable for classifying immune-peptides. Third, several performance improvement protocols have recently been developed (Hasan et al., 2020c; Hasan et al., 2020d; Hasan et al., 2020g; Manavalan, 2020), including adaptive feature learning, iterative representation feature, meta-classifier representation, and fusing with multi-view evidence. Applying multiple approaches on the same dataset and selecting the most suitable one may evolution model robustness. Exploring different classifiers on the same dataset and selecting an appropriate one are recommended.

Finally, web servers should be developed while considering the capabilities of researchers. Current immune-peptide-based predictors are developed based merely on sequence information. With the increase of dataset whose structures are known, researcher might take structural-based immune-peptides analyses and forecasts into account for more comprehensive understanding of immune-peptide patterns.

References

Agarwal, K.L., Kenner, G.W., and Sheppard, R.C. (1969). Feline gastrin. An example of peptide sequence analysis by mass spectrometry. Journal of the American Chemical Society *91*, 3096-3097.

Aggerbeck, H., Ruhwald, M., Hoff, S.T., Borregaard, B., Hellstrom, E., Malahleha, M., Siebert, M., Gani, M., Seopela, V., Diacon, A.*, et al.* (2018). C-Tb skin test to diagnose Mycobacterium tuberculosis infection in children and HIV-infected adults: A phase 3 trial. PloS one *13*, e0204554.

Agrawal, P., Bhalla, S., Usmani, S.S., Singh, S., Chaudhary, K., Raghava, G.P., and Gautam, A. (2016). CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. Nucleic Acids Res *44*, D1098-1103.

AlMatar, M., Makky, E.A., Yakici, G., Var, I., Kayar, B., and Koksal, F. (2018). Antimicrobial peptides as an alternative to anti-tuberculosis drugs. Pharmacological research *128*, 288-305.

Alt, J.A., Qin, X., Pulsipher, A., Orb, Q., Orlandi, R.R., Zhang, J., Schults, A., Jia, W., Presson, A.P., Prestwich, G.D.*, et al.* (2015). Topical cathelicidin (LL-37) an innate immune peptide induces acute olfactory epithelium inflammation in a mouse model. Int Forum Allergy Rhinol *5*, 1141-1150.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research *25*, 3389-3402.

Arbex, M.A., Varella Mde, C., Siqueira, H.R., and Mello, F.A. (2010). Antituberculosis drugs: drug interactions, adverse effects, and use in special situations. Part 2: second line drugs. Jornal brasileiro de pneumologia : publicacao oficial da Sociedade Brasileira de Pneumologia e Tisilogia *36*, 641-656.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic acids research *28*, 45-48.

Balachandran Manavalan1, R.G.G., Tae Hwan Shin1,3, Myeong Ok Kim4, and Lee, a.G. (2018). iBCe-eL: A New ensemble Learning Framework for Improved Linear B-Cell epitope Prediction. Frontiers in Immunology *9*, 1695.

Barlow, D.J., Edwards, M.S., and Thornton, J.M. (1986). Continuous and discontinuous protein antigenic determinants. Nature *322*, 747-748.

Basith, S., Manavalan, B., Hwan Shin, T., and Lee, G. (2020a). Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. Medicinal research reviews.

Basith, S., Manavalan, B., Hwan Shin, T., and Lee, G. (2020b). Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. Med Res Rev *40*, 1276-1314.

Bellner, L., Thoren, F., Nygren, E., Liljeqvist, J.A., Karlsson, A., and Eriksson, K. (2005). A proinflammatory peptide from herpes simplex virus type 2 glycoprotein G affects neutrophil, monocyte, and NK cell functions. Journal of immunology *174*, 2235-2241.

Bhadra, P., Yan, J., Li, J., Fong, S., and Siu, S.W.I. (2018). AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. Scientific reports *8*, 1697.

Bi, C., Shao, Z., Zhang, Y., Hu, L., Li, J., Huang, L., and Weng, C. (2017). Identification of a linear B-cell epitope on non-structural protein 12 of porcine reproductive and respiratory syndrome virus, using a monoclonal antibody. Arch Virol *162*, 2239-2246.

Boismenu, R., Chen, Y., Chou, K., El-Sheikh, A., and Buelow, R. (2002). Orally administered RDP58 reduces the severity of dextran sodium sulphate induced colitis. Annals of the rheumatic diseases *61 Suppl 2*, ii19-24.

Boopathi, V., Subramaniyam, S., Malik, A., Lee, G., Manavalan, B., and Yang, D.C. (2019). mACPpred: A Support Vector Machine-Based Meta-Predictor for Identification of Anticancer Peptides. Int J Mol Sci *20*.

Breiman, L. (2001). Random Forests. Machine Learning *45*, 5-32.

Bryson, C.J., Jones, T.D., and Baker, M.P. (2010). Prediction of immunogenicity of therapeutic proteins: validity of computational tools. BioDrugs : clinical immunotherapeutics, biopharmaceuticals and gene therapy *24*, 1-8.

Bylund, J., Christophe, T., Boulay, F., Nystrom, T., Karlsson, A., and Dahlgren, C. (2001). Proinflammatory activity of a cecropin-like antibacterial peptide from Helicobacter pylori. Antimicrobial agents and chemotherapy *45*, 1700-1704.

Cavaillon, J.M. (2001). Pro- versus anti-inflammatory cytokines: myth or reality. Cellular and molecular biology *47*, 695-702.

Centor, R.M. (1991). Signal Detectability - the Use of Roc Curves and Their Analyses. Medical Decision Making *11*, 102-106.

Charoenkwan, P., Kanthawong, S., Schaduangrat, N., Yana, J., and Shoombuatong, W. (2020). PVPred-SCM: Improved Prediction and Analysis of Phage Virion Proteins Using a Scoring Card Method. Cells *9*.

Charoenkwan, P., Shoombuatong, W., Lee, H.C., Chaijaruwanich, J., Huang, H.L., and Ho, S.Y. (2013). SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. PloS one *8*, e72368.

Chaudhary, K., Poirion, O.B., Lu, L., and Garmire, L.X. (2018). Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clinical cancer research : an official journal of the American Association for Cancer Research *24*, 1248-1259.

Chaurasiya, S.K. (2018). Tuberculosis: Smart manipulation of a lethal host. Microbiology and immunology *62*, 361-379.

Chen, C.W., and Chang, C.Y. (2017). Peptide Scanning-assisted Identification of a Monoclonal Antibody-recognized Linear B-cell Epitope. J Vis Exp.

Chen, K., Jiang, Y., Du, L., and Kurgan, L. (2009). Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. Journal of computational chemistry *30*, 163-172.

Chen, Z., Zhou, Y., Zhang, Z., and Song, J. (2015). Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features. Briefings in bioinformatics *16*, 640-657.

Choudhary, C., Kumar, C., Gnad, F., Nielsen, M.L., Rehman, M., Walther, T.C., Olsen, J.V., and Mann, M. (2009). Lysine acetylation targets protein complexes and co-regulates major cellular functions. Science *325*, 834-840.

Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J., and Gevaert, K. (2009). Improved visualization of protein consensus sequences by iceLogo. Nature methods *6*, 786-787.

Corrigan, M., Hirschfield, G.M., Oo, Y.H., and Adams, D.H. (2015). Autoimmune hepatitis: an approach to disease understanding and management. British medical bulletin *114*, 181-191.

Cunningham, A.D., Qvit, N., and Mochly-Rosen, D. (2017). Peptides and peptidomimetics as regulators of protein-protein interactions. Curr Opin Struct Biol *44*, 59-66.

David J. Hand , K.Y. (2001). Idiot's Bayes: Not So Stupid after All? International Statistical Review / Revue Internationale de Statistique *69*, 385-398.

De Lorenzi, E., Chiari, M., Colombo, R., Cretich, M., Sola, L., Vanna, R., Gagni, P., Bisceglia, F., Morasso, C., Lin, J.S.*, et al.* (2017). Evidence that the Human Innate Immune Peptide LL-

37 may be a Binding Partner of Amyloid-beta and Inhibitor of Fibril Assembly. J Alzheimers Dis *59*, 1213-1226.

Dekker, J.P., Fodor, A., Aldrich, R.W., and Yellen, G. (2004). A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. Bioinformatics *20*, 1565-1572.

Delgado, M., and Ganea, D. (2008). Anti-inflammatory neuropeptides: a new class of endogenous immunoregulatory agents. Brain, behavior, and immunity *22*, 1146-1151.

DeMartino, G.N. (2009). PUPylation: something old, something new, something borrowed, something Glu. Trends in biochemical sciences *34*, 155-158.

Desmet, V.J. (1987). Cholangiopathies: past, present, and future. Seminars in liver disease *7*, 67-76.

Dong, X., Zhang, Y.J., and Zhang, Z. (2013). Using weakly conserved motifs hidden in secretion signals to identify type-III effectors from bacterial pathogen genomes. PloS one *8*, e56632.

El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2008). Predicting flexible length linear B-cell epitopes. Computational systems bioinformatics Computational Systems Bioinformatics Conference *7*, 121-132.

El-Manzalawy, Y., Dobbs, D., and Honavar, V.G. (2017). In Silico Prediction of Linear B-Cell Epitopes on Proteins. Methods Mol Biol *1484*, 255-264.

Fan, L., Sun, J., Zhou, M., Zhou, J., Lao, X., Zheng, H., and Xu, H. (2016). DRAMP: a comprehensive data repository of antimicrobial peptides. Sci Rep *6*, 24482.

Ferrero-Miliani, L., Nielsen, O.H., Andersen, P.S., and Girardin, S.E. (2007). Chronic inflammation: importance of NOD2 and NALP3 in interleukin-1beta generation. Clinical and experimental immunology *147*, 227-235.

Fleri, W., Vaughan, K., Salimi, N., Vita, R., Peters, B., and Sette, A. (2017). The Immune Epitope Database: How Data Are Entered and Retrieved. Journal of immunology research *2017*, 5974574.

Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I.H. (2004). Data mining in bioinformatics using Weka. Bioinformatics *20*, 2479-2481.

Gao, J., Faraggi, E., Zhou, Y., Ruan, J., and Kurgan, L. (2012). BEST: improved prediction of B-cell epitopes from antigen sequences. PloS one *7*, e40104.

Gao, W., Kim, J.Y., Anderson, J.R., Akopian, T., Hong, S., Jin, Y.Y., Kandror, O., Kim, J.W., Lee, I.A., Lee, S.Y.*, et al.* (2015). The cyclic peptide ecumicin targeting ClpC1 is active against Mycobacterium tuberculosis in vivo. Antimicrobial agents and chemotherapy *59*, 880-889.

Gautam, A., Singh, H., Tyagi, A., Chaudhary, K., Kumar, R., Kapoor, P., and Raghava, G.P. (2012). CPPsite: a curated database of cell penetrating peptides. Database (Oxford) *2012*, bas015.

Gavrish, E., Sit, C.S., Cao, S., Kandror, O., Spoering, A., Peoples, A., Ling, L., Fetterman, A., Hughes, D., Bissell, A.*, et al.* (2014). Lassomycin, a ribosomally synthesized cyclic peptide, kills mycobacterium tuberculosis by targeting the ATP-dependent protease ClpC1P1P2. Chemistry & biology *21*, 509-518.

Gobel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. Proteins *18*, 309-317.

Gokhale, A.S., and Satyanarayanajois, S. (2014). Peptides and peptidomimetics as immunomodulators. Immunotherapy *6*, 755-774.

Gonzalez-Rey, E., Anderson, P., and Delgado, M. (2007). Emerging roles of vasoactive intestinal peptide: a new approach for autoimmune therapy. Annals of the rheumatic diseases *66 Suppl 3*, iii70-76.

Gonzalez, R.R., Fong, T., Belmar, N., Saban, M., Felsen, D., and Te, A. (2005). Modulating bladder neuro-inflammation: RDP58, a novel anti-inflammatory peptide, decreases inflammation and nerve growth factor production in experimental cystitis. The Journal of urology *173*, 630-634.

Gordon, Y.J., Romanowski, E.G., and McDermott, A.M. (2005). A review of antimicrobial peptides and their therapeutic potential as anti-infective drugs. Current eye research *30*, 505-515.

Gribskov, M., and Robinson, N.L. (1996). Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput Chem *20*, 25-33.

Groell, F., Jordan, O., and Borchard, G. (2018). In vitro models for immunogenicity prediction of therapeutic proteins. European journal of pharmaceutics and biopharmaceutics : official journal of Arbeitsgemeinschaft fur Pharmazeutische Verfahrenstechnik eV.

Guedes, R.L.M., Rodrigues, C.M.F., Coatnoan, N., Cosson, A., Cadioli, F.A., Garcia, H.A., Gerber, A.L., Machado, R.Z., Minoprio, P.M.C., Teixeira, M.M.G.*, et al.* (2018). A comparative in silico linear B-cell epitope prediction and characterization for South American and African Trypanosoma vivax strains. Genomics.

Guichard, G., Benkirane, N., Zeder-Lutz, G., van Regenmortel, M.H., Briand, J.P., and Muller, S. (1994). Antigenic mimicry of natural L-peptides with retro-inverso-peptidomimetics. Proc Natl Acad Sci U S A *91*, 9765-9769.

Gupta, S., Ansari, H.R., Gautam, A., Open Source Drug Discovery, C., and Raghava, G.P. (2013). Identification of B-cell epitopes in an antigen for inducing specific class of antibodies. Biology direct *8*, 27.

Gupta, S., Madhu, M.K., Sharma, A.K., and Sharma, V.K. (2016). ProInflam: a webserver for the prediction of proinflammatory antigenicity of peptides and proteins. J Transl Med *14*, 178.

Gupta, S., Sharma, A.K., Shastri, V., Madhu, M.K., and Sharma, V.K. (2017). Prediction of anti-inflammatory proteins/peptides: an insilico approach. Journal of translational medicine *15*, 7.

Gustafsson, A., Sigel, S., and Ljunggren, L. (2010). The antimicrobial peptide LL37 and its truncated derivatives potentiates proinflammatory cytokine induction by lipoteichoic acid in whole blood. Scandinavian journal of clinical and laboratory investigation *70*, 512-518.

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

Halperin, I., Glazer, D.S., Wu, S., and Altman, R.B. (2008). The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. BMC genomics *9 Suppl 2*, S2.

Hamilton, C.D., Swaminathan, S., Christopher, D.J., Ellner, J., Gupta, A., Sterling, T.R., Rolla, V., Srinivasan, S., Karyana, M., Siddiqui, S*., et al.* (2015). RePORT International: Advancing Tuberculosis Biomarker Research Through Global Collaboration. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America *61Suppl 3*, S155-159.

Han, X., Yang, K., and Gross, R.W. (2012). Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses. Mass spectrometry reviews *31*, 134-178.

Hasan, M., Khatun, S., and Kurata, H. (2018a). A comprehensive review of in silico analysis for protein S-sulfenylation sites. Protein and peptide letters.

Hasan M.M., K.H. (2018). GPSuc: Global Prediction of Generic and Species-specific Succinylation Sites by Aggregating Multiple Sequence Features. PloS one.

Hasan, M.M., Basith, S., Khatun, M.S., Lee, G., Manavalan, B., and Kurata, H. (2020a). Meta-i6mA: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. Brief Bioinform.

Hasan, M.M., Guo, D., and Kurata, H. (2017a). Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information. Molecular bioSystems *13*, 2545-2550.

Hasan, M.M., Khatun, M., S., Mollah, M.N., Cao, Y., and Guo, D. (2017b). A systematic identification of species-specific protein succinylation sites using joint element features information. International Journal of Nanomedicine *In Press*.

Hasan, M.M., Khatun, M.S., and H., K. (2018b). Computational Modeling of Lysine Post-Translational Modification: An Overview. Curr Synthetic Sys Biol *6*, 137.

Hasan, M.M., Khatun, M.S., and Kurata, H. (2018c). A Comprehensive Review of In silico Analysis for Protein S-sulfenylation Sites. Protein Pept Lett *25*, 815-821.

Hasan, M.M., Khatun, M.S., and Kurata, H. (2019a). Large-Scale Assessment of Bioinformatics Tools for Lysine Succinylation Sites. Cells *8*.

Hasan, M.M., Khatun, M.S., and Kurata, H. (2020b). iLBE for Computational Identification of Linear B-cell Epitopes by Integrating Sequence and Evolutionary Features. Genomics Proteomics Bioinformatics.

Hasan, M.M., Khatun, M.S., Mollah, M.N.H., Yong, C., and Dianjing, G. (2018d). NTyroSite: Computational Identification of Protein Nitrotyrosine Sites Using Sequence Evolutionary Features. Molecules *23*.

Hasan, M.M., Khatun, M.S., Mollah, M.N.H., Yong, C., and Guo, D. (2017c). A systematic identification of species-specific protein succinylation sites using joint element features information. Int J Nanomedicine *12*, 6303-6315.

Hasan MM, K.M. (2017). Recent Progress and Challenges for Protein Pupylation Sites Prediction. EC PROTEOMICS AND BIOINFORMATICS *2*, 36-45.

Hasan MM, K.M., Mollah MNH, Yong C, and Guo D (2018). NTyroSite: Computational Identification of Protein Nitrotyrosine Sites Using Sequence Evolutionary Features. Molecules.

Hasan, M.M., and Kurata, H. (2018). GPSuc: Global Prediction of Generic and Species-specific Succinylation Sites by aggregating multiple sequence features. PloS one *13*, e0200283.

Manavalan B, Hasan MM, Basith S, Gosu V, Shin TH, Lee G: Empirical Comparison and Analysis of Web-Based DNA N (4)-Methylcytosine Site Prediction Tools. Mol Ther Nucleic Acids 2020, 22:406-420.

Hasan, M.M., Manavalan, B., Khatun, M.S., and Kurata, H. (2019c). Prediction of S-nitrosylation sites by integrating support vector machines and random forest. Mol Omics *15*, 451-458.

Hasan, M.M., Manavalan, B., Khatun, M.S., and Kurata, H. (2020c). i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. Int J Biol Macromol *157*, 752-758.

Hasan, M.M., Manavalan, B., Shoombuatong, W., Khatun, M.S., and Kurata, H. (2020d). i4mC-Mouse: Improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes. Comput Struct Biotechnol J *18*, 906-912.

Hasan MM SW, Kurata H, Manavalan B: Critical evaluation of web-based DNA N6-methyladenine site prediction tools. Briefings in Functional Genomics 2021(DOI: 10.1093/bfgp/elaa028).

Hasan, M.M., Manavalan, B., Shoombuatong, W., Khatun, M.S., and Kurata, H. (2020f). i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation. Plant molecular biology.

Hasan, M.M., Rashid, M.M., Khatun, M.S., and Kurata, H. (2019d). Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information. Sci Rep *9*, 8258.

Khatun MS, Hasan MM, Shoombuatong W, Kurata H: ProIn-Fuse: improved and robust prediction of proinflammatory peptides by fusing of multiple feature representations. J Comput Aided Mol Des 2020, 34(12):1229-1236.Hasan, M.M., Schaduangrat, N., Basith, S., Lee, G., Shoombuatong, W., and Manavalan, B. (2020h). HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. Bioinformatics.

Hasan, M.M., Yang, S., Zhou, Y., and Mollah, M.N. (2016). SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. Molecular bioSystems *12*, 786-795.

Hasan, M.M., Zhou, Y., Lu, X., Li, J., Song, J., and Zhang, Z. (2015). Computational Identification of Protein Pupylation Sites by Using Profile-Based Composition of k-Spaced Amino Acid Pairs. PloS one *10*, e0129635.

Hearst, M.A. (1998). Support Vector Machines. IEEE Intelligent Systems, 18-28.

Hebert, A.S., Richards, A.L., Bailey, D.J., Ulbrich, A., Coughlin, E.E., Westphall, M.S., and Coon, J.J. (2014). The one hour yeast proteome. Molecular & cellular proteomics : MCP *13*, 339-347.

Hendriks, I.A., D'Souza, R.C., Yang, B., Verlaan-de Vries, M., Mann, M., and Vertegaal, A.C. (2014). Uncovering global SUMOylation signaling networks in a site-specific manner. Nature structural & molecular biology *21*, 927-936.

Hernandez-Florez, D., and Valor, L. (2016). Protein-kinase inhibitors: A new treatment pathway for autoimmune and inflammatory diseases? Reumatologia clinica *12*, 91-99.

Hosokawa, I., Hosokawa, Y., Komatsuzawa, H., Goncalves, R.B., Karimbux, N., Napimoga, M.H., Seki, M., Ouhara, K., Sugai, M., Taubman, M.A.*, et al.* (2006). Innate immune peptide LL-37 displays distinct expression pattern from beta-defensins in inflamed gingival tissue. Clin Exp Immunol *146*, 218-225.

Hsu, H.Y., Chang, M.H., Ni, Y.H., and Huang, S.F. (2001). Cytomegalovirus infection and proinflammatory cytokine activation modulate the surface immune determinant expression and immunogenicity of cultured murine extrahepatic bile duct epithelial cells. Clinical and experimental immunology *126*, 84-91.

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics *26*, 680-682.

Ialenti, A., Santagada, V., Caliendo, G., Severino, B., Fiorino, F., Maffia, P., Ianaro, A., Morelli, F., Di Micco, B., Carteni, M.*, et al.* (2001). Synthesis of novel anti-inflammatory peptides derived from the amino-acid sequence of the bioactive protein SV-IV. European journal of biochemistry *268*, 3399-3406.

Imamura, H., Sugiyama, N., Wakabayashi, M., and Ishihama, Y. (2014). Large-scale identification of phosphorylation sites for profiling protein kinase selectivity. Journal of proteome research *13*, 3410-3419.

Jaffrey, S.R., Erdjument-Bromage, H., Ferris, C.D., Tempst, P., and Snyder, S.H. (2001). Protein S-nitrosylation: a physiological signal for neuronal nitric oxide. Nature cell biology *3*, 193-197.

Jespersen, M.C., Peters, B., Nielsen, M., and Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic acids research *45*, W24-W29.

Jhamb, S.S., Goyal, A., and Singh, P.P. (2014). Determination of the activity of standard anti-tuberculosis drugs against intramacrophage Mycobacterium tuberculosis, in vitro: MGIT 960 as a viable alternative for BACTEC 460. The Brazilian journal of infectious diseases : an official publication of the Brazilian Society of Infectious Diseases *18*, 336-340.

Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2015). iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. Journal of theoretical biology *377*, 47-56.

Jin, Y., Wi, H.J., Choi, M.H., Hong, S.T., and Bae, Y.M. (2014). Regulation of anti-inflammatory cytokines IL-10 and TGF-beta in mouse dendritic cells through treatment with Clonorchis sinensis crude antigen. Experimental & molecular medicine *46*, e74.

Jurczak, P., Witkowska, J., Rodziewicz-Motowidlo, S., and Lach, S. (2020). Proteins, peptides and peptidomimetics as active agents in implant surface functionalization. Adv Colloid Interface Sci *276*, 102083.

Kang, X., Dong, F., Shi, C., Liu, S., Sun, J., Chen, J., Li, H., Xu, H., Lao, X., and Zheng, H. (2019). DRAMP 2.0, an updated data repository of antimicrobial peptides. Sci Data *6*, 148.

Kastin, A. (2017). Handbook of Biologically Active Peptides. ELSEVIWER *23*, 14394-14409.

Kawashima, S., and Kanehisa, M. (2000). AAindex: amino acid index database. Nucleic acids research *28*, 374.

Kawashima, S., Ogata, H., and Kanehisa, M. (1999). AAindex: Amino Acid Index Database. Nucleic acids research *27*, 368-369.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. Nucleic acids research *36*, D202-205.

Kemp, D.S. (1990). Peptidomimetics and the template approach to nucleation of beta-sheets and alpha-helices in peptides. Trends Biotechnol *8*, 249-255.

Kempuraj, D., Selvakumar, G.P., Thangavel, R., Ahmed, M.E., Zaheer, S., Raikwar, S.P., Iyer, S.S., Bhagavan, S.M., Beladakere-Ramaswamy, S., and Zaheer, A. (2017). Mast Cell Activation in Brain Injury, Stress, and Post-traumatic Stress Disorder and Alzheimer's Disease Pathogenesis. Frontiers in neuroscience *11*, 703.

Khairy, W.O.A., Wang, L., Tian, X., Ye, J., Qian, K., Shao, H., and Qin, A. (2017). Identification of a novel linear B-cell epitope in the p27 of Avian leukosis virus. Virus Res *238*, 253-257.

Khatun, M., S. Hasan, M.M. and Kurara H (2019). PreAIP: Computational prediction of anti-inflammatory peptides by integrating multiple complementary features. Frontiers in Genetics.

Islam MM, Alam MJ, Ahmed FF, Hasan MM, Mollah MNH: Improved Prediction of Protein-Protein Interaction Mapping on Homo Sapiens by Using Amino Acid Sequence Features in a Supervised Learning Framework. Protein Pept Lett 2020.

Khatun, M.S., Hasan, M.M., Shoombuatong, W., and Kurata, H. (2020a). ProIn-Fuse: improved and robust prediction of proinflammatory peptides by fusing of multiple feature representations. J Comput Aided Mol Des.

Khatun, M.S., Shoombuatong, W., Hasan, M.M., and Kurata, H. (2020b). Evolution of Sequence-based Bioinformatics Tools for Protein-protein Interaction Prediction. Curr Genomics *21*, 454-463.

Khatun, S., Hasan, M., and Kurata, H. (2019b). Efficient computational model for identification of antitubercular peptides by integrating amino acid patterns and properties. FEBS letters.

Kieber-Emmons, T., Murali, R., and Greene, M.I. (1997). Therapeutic peptides and peptidomimetics. Curr Opin Biotechnol *8*, 435-441.

Kim, S., Nam, H.Y., Lee, J., and Seo, J. (2020). Mitochondrion-Targeting Peptides and Peptidomimetics: Recent Progress and Design Principles. Biochemistry *59*, 270-284.

Kim, W., Bennett, E.J., Huttlin, E.L., Guo, A., Li, J., Possemato, A., Sowa, M.E., Rad, R., Rush, J., Comb, M.J*., et al.* (2011). Systematic and quantitative assessment of the ubiquitin-modified proteome. Molecular cell *44*, 325-340.

Kim, Y.R., and Yang, C.S. (2017). Host-Directed Therapeutics as a Novel Approach for Tuberculosis Treatment. Journal of microbiology and biotechnology *27*, 1549-1558.

Kozlova, E.E.G., Cerf, L., Schneider, F.S., Viart, B.T., C, N.G., Steiner, B.T., de Almeida Lima, S., Molina, F., Duarte, C.G., Felicori, L*., et al.* (2018). Computational B-cell epitope identification and production of neutralizing murine antibodies against Atroxlysin-I. Scientific reports *8*, 14904.

Kumar, R., Chaudhary, K., Sharma, M., Nagpal, G., Chauhan, J.S., Singh, S., Gautam, A., and Raghava, G.P. (2015). AHTPDB: a comprehensive platform for analysis and presentation of antihypertensive peptides. Nucleic Acids Res *43*, D956-962.

Hasanm MM., Kurata, H. (2018). iLMS, Computational Identification of Lysine-Malonylation Sites by Combining Multiple Sequence Features. 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), Taichung, Taiwan, 356-359.

Langeveld, J.P., Martinez-Torrecuadrada, J., Boshuizen, R.S., Meloen, R.H., and Ignacio Casal, J. (2001). Characterisation of a protective linear B cell epitope against feline parvoviruses. Vaccine *19*, 2352-2360.

Li, B.Q., Cai, Y.D., Feng, K.Y., and Zhao, G.J. (2012). Prediction of protein cleavage site with feature selection by random forest. PloS one *7*, e45854.

Liang, S., Zheng, D., Zhang, C., and Zacharias, M. (2009). Prediction of antigenic epitopes on protein surfaces by consensus scoring. BMC bioinformatics *10*, 302.

Liaw, A., Wiener (2002). Classification and regression by random forest. R news 2, 18–22.

Liu, Z., Cao, J., Ma, Q., Gao, X., Ren, J., and Xue, Y. (2011). GPS-YNO2: computational prediction of tyrosine nitration sites in proteins. Molecular bioSystems *7*, 1197-1204.

Lockless, S.W., and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science *286*, 295-299.

Lomash, S., Nagpal, S., and Salunke, D.M. (2010). An antibody as surrogate receptor reveals determinants of activity of an innate immune peptide antibiotic. J Biol Chem *285*, 35750-35758.

Lopez, Y., Sharma, A., Dehzangi, A., Lal, S.P., Taherzadeh, G., Sattar, A., and Tsunoda, T. (2018). Success: evolutionary and structural properties of amino acids prove effective for succinylation site prediction. BMC genomics *19*, 923.

Lowd, D. (2005). Naive Bayes models for probability estimation. 05 Proceedings of the 22nd international conference on Machine learning *New York, USA*, 529 - 536.

Lowenberger, C.A. (2001). Form, function and phylogenetic relationships of mosquito immune peptides. Adv Exp Med Biol *484*, 113-129.

M. S. Khatun, M.M.H., M. N. H. Mollah and H. Kurata (2018). SIPMA: A Systematic Identification of Protein-Protein Interactions in Zea mays Using Autocorrelation Features in a Machine-Learning Framework 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), Taichung, Taiwan, 122-125.

Ma, J., Wei, Y., Zhang, L., Wang, X., Yao, D., Liu, D., Liu, W., Yu, S., Yu, Y., Wu, Z*., et al.* (2018). Identification of a novel linear B-cell epitope as a vaccine candidate in the N2N3 subdomain of Staphylococcus aureus fibronectin-binding protein A. Journal of medical microbiology *67*, 423-431.

Maclin, R., and Opitz, D. (1999). Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research.

Manavalan, B., Basith, S., Shin, T.H., Choi, S., Kim, M.O., and Lee, G. (2017). MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget *8*, 77121-77136.

Rashid MM, Shatabda S, Hasan MM, Kurata H: Recent Development of Machine Learning Methods in Microbial Phosphorylation Sites. Curr Genomics 2020, 21(3):194-203.

Manavalan, B., Basith, S., Shin, T.H., Wei, L., and Lee, G. (2019a). AtbPpred: A Robust Sequence-Based Prediction of Anti-Tubercular Peptides Using Extremely Randomized Trees. Comput Struct Biotechnol J *17*, 972-981.

Manavalan, B., Basith, S., Shin, T.H., Wei, L., and Lee, G. (2019b). mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics *35*, 2757-2765.

Manavalan, B., Basith, S., Shin, T.H., Wei, L., and Lee, G. (2019c). Meta-4mCpred: A Sequence-Based Meta-Predictor for Accurate DNA 4mC Site Prediction Using Effective Feature Representation. Mol Ther Nucleic Acids *16*, 733-744.

Manavalan, B., Shin, T.H., Kim, M.O., and Lee, G. (2018b). AIPpred: Sequence-Based Prediction of Anti-inflammatory Peptides Using Random Forest. Front Pharmacol *9*, 276.

Manavalan, B., Shin, T.H., Kim, M.O., and Lee, G. (2018c). PIP-EL: A New Ensemble Learning Method for Improved Proinflammatory Peptide Predictions. Front Immunol *9*, 1783.

Manavalan, B., Subramaniyam, S., Shin, T.H., Kim, M.O., and Lee, G. (2018d). Machine-Learning-Based Prediction of Cell-Penetrating Peptides and Their Uptake Efficiency with Improved Accuracy. J Proteome Res *17*, 2715-2726.

Manavalan, B.H., MM; Basith S; Vijayakumar Gosu, Tae-Hwan Shin, Gwang Lee (2020). Empirical Comparison and Analysis of Web-based DNA N4-methylcytosine Site Prediction Tools. Molecular Therapy-Nucleic Acids.

Mangsbo, S.M., Fletcher, E.A.K., van Maren, W.W.C., Redeker, A., Cordfunke, R.A., Dillmann, I., Dinkelaar, J., Ouchaou, K., Codee, J.D.C., van der Marel, G.A.*, et al.* (2018). Linking T cell epitopes to a common linear B cell epitope: A targeting and adjuvant strategy to improve T cell responses. Molecular immunology *93*, 115-124.

Margulies, D.H., Jiang, J., and Natarajan, K. (2019). Structure and Function of Molecular Chaperones that Govern Immune Peptide Loading. Subcell Biochem *93*, 321-337.

Marie, C., Pitton, C., Fitting, C., and Cavaillon, J.M. (1996). Regulation by anti-inflammatory cytokines (IL-4, IL-10, IL-13, TGFbeta)of interleukin-8 production by LPS- and/ or TNFalpha-activated human polymorphonuclear cells. Mediators of inflammation *5*, 334-340.

Masuda, T., Sugiyama, N., Tomita, M., and Ishihama, Y. (2011). Microscale phosphoproteome analysis of 10,000 cells from human cancer cell lines. Analytical chemistry *83*, 7698-7703.

Md Mehedi Hasan, M.S.K., Md Nurul Haque Mollah, Cao Yong and Guo Dianjing (2017). A systematic identification of species-specific protein succinylation sites using joint element features information. International Journal of Nanomedicine *12*, 6303—6315.

Md. Mehedi Hasan, D.G.a.H.K. (2017). Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information Molecular BioSystms.

Medzihradszky, K.F. (2005). Peptide sequence analysis. Methods in enzymology *402*, 209-244.

Miele, L., Cordella-Miele, E., Facchiano, A., and Mukherjee, A.B. (1988). Novel anti-inflammatory peptides from the region of highest similarity between uteroglobin and lipocortin I. Nature *335*, 726-730.

Mojsoska, B., and Jenssen, H. (2015). Peptides and Peptidomimetics for Antimicrobial Drug Design. Pharmaceuticals (Basel) *8*, 366-415.

Mooney, S.D., Liang, M.H., DeConde, R., and Altman, R.B. (2005). Structural characterization of proteins using residue environments. Proteins *61*, 741-747.

Moremen, K.W., Tiemeyer, M., and Nairn, A.V. (2012). Vertebrate protein glycosylation: diversity, synthesis and function. Nature reviews Molecular cell biology *13*, 448-462.

Mosharaf, M.P., Hassan, M.M., Ahmed, F.F., Khatun, M.S., Moni, M.A., and Mollah, M.N.H. (2020). Computational prediction of protein ubiquitination sites mapping on Arabidopsis thaliana. Comput Biol Chem *85*, 107238.

Mukhopadhyay, S., Mondal, S.A., Kumar, M., and Dutta, D. (2014). Proinflammatory and antiinflammatory attributes of fetuin-a: a novel hepatokine modulating cardiovascular and glycemic outcomes in metabolic syndrome. Endocrine practice : official journal of the American College of Endocrinology and the American Association of Clinical Endocrinologists *20*, 1345-1351.

Myers, S.A., Daou, S., Affar el, B., and Burlingame, A. (2013). Electron transfer dissociation (ETD): the mass spectrometric breakthrough essential for O-GlcNAc protein site assignments-a study of the O-GlcNAcylated protein host cell factor C1. Proteomics *13*, 982-991.

Nikonenko, B.V., Samala, R., Einck, L., and Nacy, C.A. (2004). Rapid, simple in vivo screen for new drugs active against Mycobacterium tuberculosis. Antimicrobial agents and chemotherapy *48*, 4550-4555.

Olsen, J.V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M.L., Jensen, L.J., Gnad, F., Cox, J., Jensen, T.S., Nigg, E.A.*, et al.* (2010). Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. Science signaling *3*, ra3.

Olvera, A., Noguera-Julian, M., Kilpelainen, A., Romero-Martin, L., Prado, J.G., and Brander, C. (2020). SARS-CoV-2 Consensus-Sequence and Matching Overlapping Peptides Design for COVID19 Immune Studies and Vaccine Development. Vaccines (Basel) *8*.

Padhi, A., Sengupta, M., Sengupta, S., Roehm, K.H., and Sonawane, A. (2014). Antimicrobial peptides and proteins in mycobacterial therapy: current status and future prospects. Tuberculosis *94*, 363-373.

Pan, X., Zhu, L., Fan, Y.X., and Yan, J. (2014). Predicting protein-RNA interaction amino acids using random forest based on submodularity subset selection. Computational biology and chemistry *53PB*, 324-330.

Panyayai, T., Ngamphiw, C., Tongsima, S., Mhuantong, W., Limsripraphan, W., Choowongkomon, K., and Sawatdichaikul, O. (2019). FeptideDB: A web application for new bioactive peptides from food protein. Heliyon *5*, e02076.

Passerini, A., Punta, M., Ceroni, A., Rost, B., and Frasconi, P. (2006). Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. Proteins *65*, 305-316.

Patterson, H., Nibbs, R., McInnes, I., and Siebert, S. (2014). Protein kinase inhibitors in the treatment of inflammatory and autoimmune diseases. Clinical and experimental immunology *176*, 1-10.

Peng, J., Xiao, Y., Wan, X., Chen, Q., Wang, H., Li, J., Chen, J., and Gao, R. (2020). Enhancement of Immune Response and Anti-Infection of Mice by Porcine Antimicrobial Peptides and Interleukin-4/6 Fusion Gene Encapsulated in Chitosan Nanoparticles. Vaccines (Basel) *8*.

Pinho-Ribeiro, F.A., Hohmann, M.S., Borghi, S.M., Zarpelon, A.C., Guazelli, C.F., Manchope, M.F., Casagrande, R., and Verri, W.A., Jr. (2015). Protective effects of the flavonoid hesperidin methyl chalcone in inflammation and pain in mice: role of TRPV1, oxidative stress, cytokines and NF-kappaB. Chemico-biological interactions *228*, 88-99.

Pirtskhalava, M., Gabrielian, A., Cruz, P., Griggs, H.L., Squires, R.B., Hurt, D.E., Grigolava, M., Chubinidze, M., Gogoladze, G., Vishnepolsky, B*., et al.* (2016). DBAASP v.2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. Nucleic acids research *44*, 6503.

Polikar, R. (2006). Ensemble based systems in decision making. Circuits and systems magazine, IEEE *6*, 21-45.

Poretsky, E., Dressano, K., Weckwerth, P., Ruiz, M., Char, S.N., Da, S., Abagyan, R., Yang, B., and Huffaker, A. (2020). Differential activities of maize Plant Elicitor Peptides as mediators of immune signaling and herbivore resistance. Plant J.

R. S. Michalski , J.G.C.a.T.M.M. (2013). Machine Learning: An Artificial Intelligence Approach. Springer Publishing Company.

Rahman Kh, S., Chowdhury, E.U., Sachse, K., and Kaltenboeck, B. (2016). Inadequate Reference Datasets Biased toward Short Non-epitopes Confound B-cell Epitope Prediction. The Journal of biological chemistry *291*, 14585-14599.

Ramstrom, M., and Sandberg, H. (2011). Characterization of gamma-carboxylated tryptic peptides by collision-induced dissociation and electron transfer dissociation mass spectrometry. European journal of mass spectrometry *17*, 497-506.

Rani, P., and Pudi, V. (2008). RBNBC: Repeat Based Naive Bayes Classifier for Biological Sequences. Ieee Data Mining, 989-994.

Rao, B., Zhou, C., Zhang, G., Su, R., and Wei, L. (2020). ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides. Brief Bioinform *21*, 1846-1855.

Reichhart, J.M., and Achstetter, T. (1990). Expression and secretion of insect immune peptides in yeast. Res Immunol *141*, 943-946.

Ren, J., Wen, L., Gao, X., Jin, C., Xue, Y., and Yao, X. (2008). CSS-Palm 2.0: an updated software for palmitoylation sites prediction. Protein engineering, design & selection : PEDS *21*, 639-644.

Richards, A.L., Hebert, A.S., Ulbrich, A., Bailey, D.J., Coughlin, E.E., Westphall, M.S., and Coon, J.J. (2015). One-hour proteome analysis in yeast. Nature protocols *10*, 701-714.

Rokach, L. (2010). Ensemble-based classifiers. Artificial Intelligence Review *33*, 1-39.

Rosenthal, K.S. (2005). Immune peptide enhancement of peptide based vaccines. Front Biosci *10*, 478-482.

Hasan MM, Alam MA, Shoombuatong W, Kurata H: IRC-Fuse: improved and robust prediction of redox-sensitive cysteine by fusing of multiple feature representations. J Comput Aided Mol Des 2021.

Saha, S., and Raghava, G.P. (2007). Prediction methods for B-cell epitopes. Methods in molecular biology *409*, 387-394.

Scarpioni, R., Ricardi, M., and Albertazzi, V. (2016). Secondary amyloidosis in autoinflammatory diseases and the role of inflammation in renal damage. World journal of nephrology *5*, 66-75.

Schaduangrat, N., Nantasenamat, C., Prachayasittikul, V., and Shoombuatong, W. (2019). Meta-iAVP: A Sequence-Based Meta-Predictor for Improving the Prediction of Antiviral Peptides Using Effective Feature Representation. International journal of molecular sciences *20*.

Schisler, N.J., and Palmer, J.D. (2000). The IDB and IEDB: intron sequence and evolution databases. Nucleic acids research *28*, 181-184.

Shahjahan, M., Khatun, M.S., Mun, M.M., Islam, S.M.M., Uddin, M.H., Badruzzaman, M., and Khan, S. (2020). Nuclear and Cellular Abnormalities of Erythrocytes in Response to Thermal Stress in Common Carp Cyprinus carpio. Front Physiol *11*, 543.

Shao, J., Xu, D., Tsai, S.N., Wang, Y., and Ngai, S.M. (2009). Computational identification of protein methylation sites through bi-profile Bayes feature extraction. PloS one *4*, e4920.

Sharma, A., Rastogi, T., Bhartiya, M., Shasany, A.K., and Khanuja, S.P. (2007). Type 2 diabetes mellitus: phylogenetic motifs for predicting protein functional sites. Journal of biosciences *32*, 999-1004.

Sheppard, S., Lawson, N.D., and Zhu, L.J. (2013). Accurate identification of polyadenylation sites from 3' end deep sequencing using a naive Bayes classifier. Bioinformatics *29*, 2564-2571.

Shi, J., Liu, Y., Wang, Y., Zhang, J., Zhao, S., and Yang, G. (2015). Biological and immunotoxicity evaluation of antimicrobial peptide-loaded coatings using a layer-by-layer process on titanium. Scientific reports *5*, 16336.

Shoombuatong, W., Schaduangrat, N., Pratiwi, R., and Nantasenamat, C. (2019). THPep: A machine learning-based approach for predicting tumor homing peptides. Computational biology and chemistry *80*, 441-451.

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

Silva, J.P., Appelberg, R., and Gama, F.M. (2016). Antimicrobial peptides as novel anti-tuberculosis therapeutics. Biotechnology advances *34*, 924-940.

Singh, H., Ansari, H.R., and Raghava, G.P. (2013). Improved method for linear B-cell epitope prediction using antigen's primary sequence. PloS one *8*, e62216.

Singh, S., Chaudhary, K., Dhanda, S.K., Bhalla, S., Usmani, S.S., Gautam, A., Tuknait, A., Agrawal, P., Mathur, D., and Raghava, G.P. (2016). SATPdb: a database of structurally annotated therapeutic peptides. Nucleic Acids Res *44*, D1119-1126.

Skovbakke, S.L., and Franzyk, H. (2017). Anti-inflammatory Properties of Antimicrobial Peptides and Peptidomimetics: LPS and LTA Neutralization. Methods Mol Biol *1548*, 369-386.

Slade, D.J., Subramanian, V., Fuhrmann, J., and Thompson, P.R. (2014). Chemical and biological methods to detect post-translational modifications of arginine. Biopolymers *101*, 133-143.

Steere, A.C., Drouin, E.E., and Glickstein, L.J. (2011). Relationship between immunity to Borrelia burgdorferi outer-surface protein A (OspA) and Lyme arthritis. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America *52 Suppl 3*, s259-265.

Steinman, L., Merrill, J.T., McInnes, I.B., and Peakman, M. (2012). Optimization of current and future therapy for autoimmune diseases. Nature medicine *18*, 59-65.

Striebel, F., Imkamp, F., Sutter, M., Steiner, M., Mamedov, A., and Weber-Ban, E. (2009). Bacterial ubiquitin-like modifier Pup is deamidated and conjugated to substrates by distinct but homologous enzymes. Nature structural & molecular biology *16*, 647-651.

Su, R., Hu, J., Zou, Q., Manavalan, B., and Wei, L. (2019). Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. Briefings in bioinformatics.

Sun, T., Zhou, B., Lai, L., and Pei, J. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC bioinformatics *18*, 277.

Sweredoski, M.J., and Baldi, P. (2008). PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. Bioinformatics *24*, 1459-1460.

Sweredoski, M.J., and Baldi, P. (2009). COBEpro: a novel system for predicting continuous B-cell epitopes. Protein engineering, design & selection : PEDS *22*, 113-120.

Syka, J.E., Coon, J.J., Schroeder, M.J., Shabanowitz, J., and Hunt, D.F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America *101*, 9528-9533.

Tabas, I., and Glass, C.K. (2013). Anti-inflammatory therapy in chronic disease: challenges and opportunities. Science *339*, 166-172.

Tahir M, T.H., Chong KT (2019). iDNA6mA (5-step rule): Identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule. Chemometrics and Intelligent Laboratory Systems.

Tang, H., Su, Z.D., Wei, H.H., Chen, W., and Lin, H. (2016). Prediction of cell-penetrating peptides with feature selection techniques. Biochem Biophys Res Commun *477*, 150-154.

Teveroni, E., Luca, R., Pellegrino, M., Ciolli, G., Pontecorvi, A., and Moretti, F. (2016). Peptides and peptidomimetics in the p53/MDM2/MDM4 circuitry - a patent review. Expert Opin Ther Pat *26*, 1417-1429.

Thanamani, B.A.a.A.S. (2013). Feature Selection based on Information Gain. International Journal of Innovative Technology and Exploring Engineering *2*, 2278-3075.

Tomar, N., and De, R.K. (2010). Immunoinformatics: an integrated scenario. Immunology *131*, 153-168.

Trinidad, J.C., Barkan, D.T., Gulledge, B.F., Thalhammer, A., Sali, A., Schoepfer, R., and Burlingame, A.L. (2012). Global identification and characterization of both O-GlcNAcylation and phosphorylation at the murine synapse. Molecular & cellular proteomics : MCP *11*, 215-229.

Tyagi, A., Tuknait, A., Anand, P., Gupta, S., Sharma, M., Mathur, D., Joshi, A., Singh, S., Gautam, A., and Raghava, G.P. (2015). CancerPPD: a database of anticancer peptides and proteins. Nucleic Acids Res *43*, D837-843.

Umlauf, D., Goto, Y., and Feil, R. (2004). Site-specific analysis of histone methylation and acetylation. Methods in molecular biology *287*, 99-120.

Usmani, S.S., Bhalla, S., and Raghava, G.P.S. (2018a). Prediction of Antitubercular Peptides From Sequence Information Using Ensemble Classifier and Hybrid Features. Frontiers in pharmacology *9*, 954.

Usmani, S.S., Kumar, R., Kumar, V., Singh, S., and Raghava, G.P.S. (2018b). AntiTbPdb: a knowledgebase of anti-tubercular peptides. Database (Oxford) *2018*.

Vacic, V., Iakoucheva, L.M., and Radivojac, P. (2006). Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics *22*, 1536-1537.

Vandermarliere, E., and Martens, L. (2013). Protein structure as a means to triage proposed PTM sites. Proteomics *13*, 1028-1035.

Vasic, D., and Walcher, D. (2012). Proinflammatory effects of C-Peptide in different tissues. International journal of inflammation *2012*, 932725.

Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., and Peters, B. (2018). The Immune Epitope Database (IEDB): 2018 update. Nucleic acids research.

Vita, R., Overton, J.A., Greenbaum, J.A., Ponomarenko, J., Clark, J.D., Cantrell, J.R., Wheeler, D.K., Gabbard, J.L., Hix, D., Sette, A*., et al.* (2015). The immune epitope database (IEDB) 3.0. Nucleic Acids Res *43*, D405-412.

Wang, G. (2020). Bioinformatic Analysis of 1000 Amphibian Antimicrobial Peptides Uncovers Multiple Length-Dependent Correlations for Peptide Design and Prediction. Antibiotics (Basel) *9*.

Wang, H.W., and Pai, T.W. (2014). Machine learning-based methods for prediction of linear B-cell epitopes. Methods Mol Biol *1184*, 217-236.

Wang, J., Dong, R., Zou, P., Chen, Y., Li, N., Wang, Y., Zhang, T., and Pan, X. (2020a). Identification of a Novel Linear B Cell Epitope on the Sao Protein of Streptococcus suis Serotype 2. Front Immunol *11*, 1492.

Wang, J.Y., Sun, H.Y., Wang, J.T., Hung, C.C., Yu, M.C., Lee, C.H., and Lee, L.N. (2015). Nine- to Twelve-Month Anti-Tuberculosis Treatment Is Associated with a Lower Recurrence

Rate than 6-9-Month Treatment in Human Immunodeficiency Virus-Infected Patients: A Retrospective Population-Based Cohort Study in Taiwan. PloS one *10*, e0144136.

Wang, M., Wei, Y., Yu, W., Wang, L., Zhai, L., Li, X., Wang, X., Zhang, H., Feng, Z., Yu, L., *et al.* (2018). Identification of a conserved linear B-cell epitope in the Staphylococcus aureus GapC protein. Microbial pathogenesis *118*, 39-47.

Wang, R., Wang, Y., Edrington, T.C., Liu, Z., Lee, T.C., Silvanovich, A., Moon, H.S., Liu, Z.L., and Li, B. (2020b). Presence of small resistant peptides from new in vitro digestion assays detected by liquid chromatography tandem mass spectrometry: An implication of allergenicity prediction of novel proteins? PLoS One *15*, e0233745.

Wang, X., Fang, L., Zhan, J., Shi, X., Liu, Q., Lu, Q., Bai, J., Li, Y., and Jiang, P. (2020c). Identification and characterization of linear B cell epitopes on the nucleocapsid protein of porcine epidemic diarrhea virus using monoclonal antibodies. Virus Res *281*, 197912.

Watkins, L.R., Maier, S.F., and Goehler, L.E. (1995). Immune activation: the role of pro-inflammatory cytokines in inflammation, illness responses and pathological pain states. Pain *63*, 289-302.

Welsch, D.J., and Nelsestuen, G.L. (1988). Amino-terminal alanine functions in a calcium-specific process essential for membrane binding by prothrombin fragment 1. Biochemistry *27*, 4939-4945.

Whelan, F.J., Yap, N.V., Surette, M.G., Golding, G.B., and Bowdish, D.M. (2013). A guide to bioinformatics for immunologists. Frontiers in immunology *4*, 416.

WHO (2017a). Global Tuberculosis Report. Geneva: World Health Organization, 1-262.

Charoenkwan P, Yana J, Nantasenamat C, Hasan MM, Shoombuatong W: iUmami-SCM: A Novel Sequence-Based Predictor for Prediction and Analysis of Umami Peptides Using a Scoring Card Method with Propensity Scores of Dipeptides. J Chem Inf Model 2020.

Wilson, J.W., and Tsukayama, D.T. (2016). Extensively Drug-Resistant Tuberculosis: Principles of Resistance, Diagnosis, and Management. Mayo Clinic proceedings *91*, 482-495.

Win, T.S., Malik, A.A., Prachayasittikul, V., JE, S.W., Nantasenamat, C., and Shoombuatong, W. (2017). HemoPred: a web server for predicting the hemolytic activity of peptides. Future medicinal chemistry *9*, 275-291.

Wynendaele, E., Bronselaer, A., Nielandt, J., D'Hondt, M., Stalmans, S., Bracke, N., Verbeke, F., Van De Wiele, C., De Tre, G., and De Spiegeleer, B. (2013). Quorumpeps database: chemical space, microbial origin and functionality of quorum sensing peptides. Nucleic Acids Res *41*, D655-659.

Xu, H.D., Shi, S.P., Wen, P.P., and Qiu, J.D. (2015). SuccFind: a novel succinylation sites online prediction tool via enhanced characteristic strategy. Bioinformatics.

Yan, J., Bhadra, P., Li, A., Sethiya, P., Qin, L., Tai, H.K., Wong, K.H., and Siu, S.W.I. (2020). Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning. Mol Ther Nucleic Acids *20*, 882-894.

Yang, H., Zhou, H., Huang, Z., Tao, K., Huang, N., Peng, Z., and Feng, W. (2020). Induction of CML-specific immune response through cross-presentation triggered by CTP-mediated BCR-ABL-derived peptides. Cancer Lett *482*, 44-55.

Yang, X., and Yu, X. (2009). An introduction to epitope prediction methods and software. Reviews in medical virology *19*, 77-96.

Yang, Y., Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., and Zhou, Y. (2017). SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks. Methods in molecular biology *1484*, 55-63.

Yao, B., Zhang, L., Liang, S., and Zhang, C. (2012). SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. PloS one *7*, e45152.

Yao, B., Zheng, D., Liang, S., and Zhang, C. (2020). SVMTriP: A Method to Predict B-Cell Linear Antigenic Epitopes. Methods Mol Biol *2131*, 299-307.

Yao, Y., Du, X., Diao, Y., and Zhu, H. (2019). An integration of deep learning with feature embedding for protein-protein interaction prediction. PeerJ *7*, e7126.

Yi, L., Cao, Z., Tong, M., Cheng, Y., Yang, Y., Li, S., Wang, J., Lin, P., Sun, Y., Zhang, M.*, et al.* (2017). Identification of a novel linear B-cell epitope using a monoclonal antibody against the carboxy terminus of the canine distemper virus nucleoprotein and sequence analysis of the identified epitope in different CDV isolates. Virology journal *14*, 187.

Yi, Y., Lv, Y., You, X., Chen, J., Bian, C., Huang, Y., Xu, J., Deng, L., and Shi, Q. (2019). High throughput screening of small immune peptides and antimicrobial peptides from the Fish-T1K database. Genomics *111*, 215-221.

Yi, Z., Ling, Y., Zhang, X., Chen, J., Hu, K., Wang, Y., Song, W., Ying, T., Zhang, R., Lu, H.*, et al.* (2020). Functional mapping of B-cell linear epitopes of SARS-CoV-2 in COVID-19 convalescent population. Emerg Microbes Infect *9*, 1988-1996.

Youn, E., Peters, B., Radivojac, P., and Mooney, S.D. (2007). Evaluation of features for catalytic residue prediction in novel folds. Protein science : a publication of the Protein Society *16*, 216-226.

Yount, N.Y., and Yeaman, M.R. (2004). Multidimensional signatures in antimicrobial peptides. Proceedings of the National Academy of Sciences of the United States of America *101*, 7363-7368.

Yu, T.F., Ma, B., and Wang, J.W. (2016). Identification of linear B-cell epitopes on goose parvovirus non-structural protein. Vet Immunol Immunopathol *179*, 85-88.

Zasloff, M. (2006). Inducing endogenous antimicrobial peptides to battle infections. Proceedings of the National Academy of Sciences of the United States of America *103*, 8913-8914.

Zhang, J.M., and An, J. (2007). Cytokines, inflammation, and pain. International anesthesiology clinics *45*, 27-37.

Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., Yuan, D., Tao, Y., Guo, Y., Ni, X., and Shi, T. (2018). Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. Frontiers in genetics *9*, 477.

Zhang, S., Chen, J., Hong, P., Li, J., Tian, Y., Wu, Y., and Wang, S. (2020). PromPDD, a web-based tool for the prediction, deciphering and design of promiscuous peptides that bind to HLA class I molecules. J Immunol Methods *476*, 112685.

Zhang, S.W., Pan, Q., Zhang, H.C., Shao, Z.C., and Shi, J.Y. (2006). Prediction of protein homo-oligomer types by pseudo amino acid composition: Approached with an improved feature extraction and Naive Bayes Feature Fusion. Amino acids *30*, 461-468.

Zhao, L., Wang, X., Zhang, X.L., and Xie, Q.F. (2016). Purification and identification of anti-inflammatory peptides derived from simulated gastrointestinal digests of velvet antler protein (Cervus elaphus Linnaeus). Journal of food and drug analysis *24*, 376-384.

Zhao, X., Chen, L., and Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. Mathematical biosciences *306*, 136-144.

Zhao, X., Koshiba, T., Fujimoto, Y., Pirenne, J., Yoshizawa, A., Ito, T., Kamei, H., Jobara, K., Ogawa, K., Uryuhara, K.*, et al.* (2005). Proinflammatory and antiinflammatory cytokine production during ischemia-reperfusion injury in a case of identical twin living donor liver transplantation using no immunosuppression. Transplantation proceedings *37*, 392-394.

Zhao, X., Ning, Q., Chai, H., and Ma, Z. (2015). Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique. Journal of theoretical biology *374*, 60-65.

Zhao, X., Wu, H., Lu, H., Li, G., and Huang, Q. (2013). LAMP: A Database Linking Antimicrobial Peptides. PLoS One *8*, e66557.

Zouki, C., Ouellet, S., and Filep, J.G. (2000). The anti-inflammatory peptides, antiflammins, regulate the expression of adhesion molecules on human leukocytes and prevent neutrophil adhesion to endothelial cells. FASEB journal : official publication of the Federation of American Societies for Experimental Biology *14*, 572-580.

Zumla, A., George, A., Sharma, V., Herbert, R.H., Baroness Masham of, I., Oxley, A., and Oliver, M. (2015). The WHO 2014 global tuberculosis report--further to go. The Lancet Global health *3*, e10-12.