Incorporating Support Vector Machine With Sequential Minimal Optimization to Identify Anticancer Peptides

Yu Wan
Chinese university of Hong Kong, Shenzhen
Zhuo Wang
Chinese university of Hong Kong, shenzhen https://orcid.org/0000-0002-7076-8432
Tzong-Yi Lee (✉ francislee0215@gmail.com)
The Chinese University of Hong Kong

Research article

Keywords: Anticancer peptides, support vector machine, SVM, sequential minimal optimization

Posted Date: September 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-78243/v1

Abstract

Background

Cancer is a major cause of death worldwide. To treat cancer, the use of anticancer peptides (ACPs) has received increasing attention in recent years. ACPs are a unique group of small molecules that can target and kill cancer cells fast and directly. However, identifying ACPs by wet-lab experiments is time-consuming and labor-intensive. Therefore, it is significant to develop computational tools for ACPs prediction.

Results

This study chose amino acid composition (AAC), N5C5, k-space, position-specific scoring matrix (PSSM) as features, and analyzed them by machine learning methods, including support vector machine (SVM) and sequential minimal optimization (SMO) to build a model (model 2) distinguishing ACPs from non-ACPs. Since a growing number of studies have shown that some antimicrobial peptides (AMPs) exhibit anticancer function, a model (model 1) to distinguish ACPs from AMPs is also been developed. Comparing to previous models, models developed in this research show better performance (accuracy: 82.5% for model 1 and 93.5% for model 2).

Conclusions

This work utilizes a new feature, PSSM, which contributes to better performance than other features. In addition to SVM, SMO is used in this research for optimizing SVM and the SMO-models show better performance than unoptimized models. Last but not least, this work provides two different functions, including distinguishing ACPs from AMPs and distinguishing ACPs from all peptides. The second SMO-optimized model, which utilizes PSSM as feature, performs better than all other existing tools.

Background

Cancer is a leading cause of death and the most important barrier to increasing life expectancy worldwide in this century [1]. This disease is caused by the growth and uncontrolled proliferation of abnormal cells. Conventional cancer treatments, including radiation therapy and chemotherapy, often have adverse effects on normal cells and thus not effective enough [2]. Moreover, some mechanisms also lead to drug resistance from the cancerous cells [3]. Therefore, a novel treatment which lacks adverse effects, targets specifically to cancer cells, and with a low possibility of drug resistance is in need urgently.

In recent years, a new group of small peptides, anticancer peptides (ACPs), has been discovered that can target and kill cancer cells specifically while not affecting healthy cells [4]. The high selectivity and low toxicity of ACPs depend on multiple differences between cancer cells and normal cells, including membrane net charge and unique molecules on membrane [5]. Due to the specificity and low toxicity of

ACPs, they have been receiving growing attention as a novel cancer treatment and be considered as promising [6]. In order to promote its application, it is of great significance to distinguish ACPs from all peptides. Nevertheless, finding anticancer peptides by experiments could be both time-consuming and labor-intensive [7]. To deal with this problem, computational identification prior to wet-lab experiments is necessary. Machine learning methods could be of great help to classify and predict those special peptides. Moreover, some characteristics of cancer cells, such as the negative surface charge of their membrane, also shared by bacterial cells [8]. Under consideration of this fact, a hypothesis is proposed that ACPs share similar features with another group of small molecules that can specifically target and kill microbes, called antimicrobial peptides (AMPs) [9]. Indeed, some antimicrobial peptides (AMPs) are discovered to exhibit anticancer function according to recent studies [4]. Thus, distinguishing ACPs from AMPs may help the discovery of ACPs more accurate, more convenient and faster.

In order to identify and predict ACPs, many computational tools for predicting ACPs have been designed, including Hajisharifi's model [10], AntiCP [11] developed by Tyagi and his colleagues, iACP [12] designed by Chen *et al*, and MLACP [13] developed by Manavalan and his colleagues. Hajisharifi *et al* use physicochemical properties, PseAAC as characteristics of peptide sequences and support vector machine (SVM) as a machine learning method to identify ACPs. Their method is claimed to perform with an accuracy of 83.82% [10]. By analyzing amino acid composition (AAC) of peptides and using SVM as a machine learning method, AntiCP offers two models that can distinguish ACPs from either AMPs or non-ACPs based on different datasets [11]. In MLACP, they analyze the amino acid composition, dipeptide composition, atomic composition and physiochemical properties separately and hybridlike. Then apply two machine learning methods: SVM and random forest to build models basing on peptide characteristics. The performance of MLACP is claimed to be better than any other existing methods, with an accuracy of 87.5%. The deficiency of the MLACP study is that it does not offer a model that can distinguish ACPs from AMPs [13].

This research offers more functions and better performance. First of all, sequences of examined ACPs, non-ACPs, and AMPs without anti-cancer functions are collected. With these data, two different groups of datasets are constructed: (1) inspected ACPs as positive data and AMPs without anti-cancer function as negative data; (2) examined ACPs as positive data, simple non-ACPs as negative data. Then characteristics of those peptide sequences are analyzed, considering four features, amino acids composition (AAC), N5C5, k-space and position-specific scoring matrix (PSSM), separately and also hybridized. It is the first time that PSSM is used as a feature in ACP prediction studies. Based on the analysis of those features, several models are built with the help of two machine learning methods: SVM [14] and Sequential Minimal Optimization (SMO) [15]. Comparing the performance of those models, two best ones are chosen: SMO-1, which utilizes SMO to analyze PSSM feature of a dataset (1), and SMO-2, which uses SMO as well and is based on analysis of PSSM of dataset (2). At last, the same testing dataset is applied to test the performance of SMO-1, SMO-2, AntiCP-1, AntiCP-2, RFACP (from MLACP), and SVMACP (from MLACP) [13]. As for results, comparing to AntiCP-1, which is also designed to distinguish ACPs from AMPs, SMO-1 shows higher accuracy, specificity and MCC. In addition, the performance of SMO-1 is of better balance. As for SMO-2, identifying ACPs from all kinds of peptides, performs better with consideration of accuracy, sensitivity and MCC, and shows relatively more balanced performance than AntiCP-2, RFACP, and SVMACP does. In general, this research built two models with different functions: one is for predicting ACPs from AMPs, which share some similarities to ACPs, and another one is used to distinguish ACPs from all peptides. The second SMO-optimized model shows better performance than the unoptimized model and other existing tools.

Results

Characterization of the sequence-based features of ACPs

The general AAC analysis results are shown as **Figure 1**. The frequency of each kind of amino acid in different datasets is shown in different colors. Comparing to AMPs but non-ACPs (peptides in negative dataset 1), K, L, A are much more frequent in ACPs, whereas N, Y, Q are dominant in negative dataset 1(with the lowest p-values). Similarly, comparing to non-ACPs in negative dataset 2, L, W, A are dominant in ACPs, whereas M, R, Q are dominant in non-ACPs (with the lowest p-values). Those significant differences of frequency of each amino acid in different datasets contribute greatly to later classification.

In **Figure 2**, N5C5 results of positive dataset, negative 1 dataset and negative 2 dataset are shown separately in different colors. Moreover, it also shows comparison between positive and negative 1 dataset, and comparison between positive and negative 2 dataset in below, based on the difference values. According to the result of positive dataset, K, L are the two most dominant amino acids in N5C5 of ACPs. Taking position under consideration, K is dominant in the third position of C-terminal end, L is dominant in the first position of C-terminal end, and G is dominant in first position of N-terminal end. G is also dominant in the first position of N-terminal end of negative 1 group (AMPs but non-ACPs). Contrarily, M is the most frequent one in the first position of N-terminal end of non-ACPs in negative 2 dataset. Comparing positive dataset to negative 1 dataset, significant differences can be found: A, L, F, K are more major in positive data while C is more major in negative 1 data. On the contrary, comparing positive dataset to negative 2 dataset, distributions of each amino acid in each position are more divergent, and less contrasts could be extracted.

The result of k-space analysis is shown as **Figure 3**. With X representing spacings between amino acids, the ten most diverse k-space pairs comparing positive data to negative 1 data are KXXXK, KXL, KXXK, LXK, LXXXXK, AK, KK, LXXXXXK, AXXXXK, KXXXXL. The ten most different k-space pairs comparing positive data to negative 2 dataset are KXXXK, LXK, KXL, LXXXXK, KXXK, LXXXXXK, LXXXL, KXA, AK, KXXXXA. It should be noted that these results are roughly correspondent to previous AAC and N5C5 results.

Model Performance

Characteristics of peptide data are then utilized to build models, using machine learning methods such as SVM and SMO. **Table 1, 2, 3, 4** show the model performance of analysis of both single and hybrid features using SVM and SMO method. In general, AAC, N5C5, k-space range from 0 to 4 and PSSM are used to build model separately. Then, AAC, N5C5, k-space = 0 are hybridized in pairs and altogether to build some other models. It should be noted that in SVM models, shown as **Table 1, 2**, the weight of each model is tried to be adjusted from 0.1 to 10 and the one which could perform the best accuracy is chosen as final weight. **Table 1** demonstrates the model performance using SVM to analyze positive dataset and negative dataset 1. As mentioned above, in order to improve performance of those models, the weight of each model is adjusted. For each feature, the first row lists the performance of model with original weight 1.0, and the second row lists the performance of model with optimum weight, which can maximize accuracy. However, the table omits the second row of each model if the original weight 1.0 maximize accuracy. Among all those models, the one used PSSM as feature with a weight of 0.8 performs the best, whose accuracy is 83.26%.

Similarly, **Table 2** shows performance of models that use SVM to analyze positive dataset and negative dataset 2. After adjusting weight of each model, the one with the highest accuracy hybridizes AAC and N5C5 for analysis. With a weight of 1.4, the accuracy of that model reaches 93.74%. **Table 3** displays performance of models that are built by SMO algorithm and comparing positive dataset to negative dataset 1. Using PSSM as representative characteristic of peptides, the accuracy of that model ranks the first among all models in this group at 85.31%. **Table 4** displays statistics of models which are constructed using SMO as machine learning method and negative dataset 2 as negative data. After evaluation of performance, the model which analyzes PSSM shows the highest accuracy of 94.92%.

Features	Weight	Sensitivity	Specificity	Accuracy
AAC	1.0	0.762	0.827	80.13
	0.8	0.717	0.904	81.10
N5C5	1.0	0.751	0.717	73.97
	0.6	0.624	0.833	75.38
k-space = 0	1.0	0.447	0.600	52.38
k-space = 1	1.0	0.445	0.598	52.16
k-space = 2	1.0	0.434	0.600	51.73
k-space = 3	1.0	0.432	0.598	51.51
k-space = 4	1.0	0.423	0.596	50.54
AAC + k-space=0	1.0	0.536	0.613	57.45
AAC+N5C5	1.0	0.749	0.873	81.10
	0.9	0.739	0.888	81.32
N5C5 + k-space = 0	1.0	0.521	0.616	56.80
AAC + N5C5 + k-space = 0	1.0	0.546	0.631	58.86
	1.1	0.730	0.475	60.26
PSSM	1.0	0.782	0.855	81.86
	0.8	0.767	0.898	83.26

Table 1. performance of models based on positive dataset and negative 1 dataset using SVM as classifier

Features	SN	SP	ACC	Weight	SN	SP	ACC
AAC	0.892	0.918	90.50	1.5	0.920	0.896	90.82
N5C5	0.877	0.929	90.28	3.8	0.957	0.896	92.66
k-space = 0	0.767	0.691	72,89	1.2	0.940	0.553	74.62
k-space = 1	0.743	0.663	70.30	1.2	0.935	0.509	72.14
k-space = 2	0.838	0.667	75.27	1.1	0.940	0.624	78.19
k-space = 3	0.823	0.663	74.30	1.1	0.931	0.611	77.11
k-space = 4	0.793	0.663	72.79	1.2	0.978	0.525	75.16
AAC + k-space=0	0.916	0.771	84.34	1.1	0.944	0.754	84.88
AAC+N5C5	0.942	0.927	93.41	1.4	0.955	0.920	93.74
N5C5 + k-space = 0	0.981	0.769	87.47	0.8	0.864	0.823	88.01
AAC + N5C5 + k-space = 0	0.974	0.853	91.36	0.9	0.965	0.877	92.12
PSSM	0.927	0.918	92.22	1.4	0.948	0.905	92.66

Table 2. performance of models based on positive dataset and negative 2 dataset using SVM as classifier

Features	Sensitivity	Specificity	Accuracy	MCC
N5C5	0.700	0.808	0.754	0.511
AAC	0.719	0.862	0.791	0.587
K-space=0	0.790	0.838	0.814	0.629
K-space=1	0.834	0.868	0.851	0.702
K-space=2	0.812	0.879	0.846	0.693
K-space=3	0.795	0.825	0.810	0.620
K-space=4	0.780	0.849	0.814	0.630
AAC+K-space=0	0.793	0.840	0.816	0.634
AAC+N5C5	0.728	0.873	0.800	0.607
N5C5+K-space=0	0.784	0.834	0.809	0.618
AAC+N5C5+K-space=0	0.793	0.849	0.821	0.642
PSSM	0.844	0.862	0.853	0.706

Table 3. performance of models based on positive dataset and negative 1 dataset using SVM as classifier

Features	Sensitivity	Specificity	Accuracy (%)	MCC
N5C5	0.905	0.931	0.918	0.836
AAC	0.896	0.931	0.914	0.828
K-space=0	0.890	0.929	0.909	0.819
K-space=1	0.933	0.950	0.942	0.884
K-space=2	0.924	0.942	0.933	0.866
K-space=3	0.898	0.935	0.917	0.834
K-space=4	0.898	0.933	0.916	0.832
AAC+K-space=0	0.935	0.955	0.945	0.890
AAC+N5C5	0.942	0.948	0.945	0.890
N5C5+K-space=0	0.922	0.948	0.935	0.871
AAC+N5C5+K-space=0	0.927	0.950	0.934	0.877
PSSM	0.950	0.948	0.949	0.898

Table 4. performance of models based on positive dataset and negative 2 dataset using SVM as classifier

Among all models, two models with the best performance are chosen as the final models of this research: using SMO method to analyze PSSM feature of positive dataset against negative 1 dataset (named as SMO-1) and positive dataset against negative 2 dataset (named as SMO-2). In order to prevent overfitting, testing sets are utilized to test the performance of those two models constructed in this research.

Comparison with existing ACPs prediction tools in terms of performance

To show significance and success of those two models, the testing dataset is also applied to test existing models, including two AntiCP models and two MLACP models. Testing results are shown in **Table 5**. Testing data from positive and negative 1 dataset are applied on SMO-1, AntiCP-1, RFACP and SVMACP. The model constructed in this work, SMO-1, shows the highest specificity. And the RFACP performs well with the highest accuracy and MCC. Similarly, testing data from positive and negative 2 dataset are applied on SMO-2, AntiCP-2, RFACP and SVMACP**,** which are all models utilized to distinguish ACPs from all kinds of peptides. Considering accuracy, sensitivity and MCC, SMO-2 shows the best performance among all models. Although the specificity value of SMO-2 is slightly lower than MLACP models, SMO-2 shows more balanced performance in each performance evaluation, indicating that SMO-2 is more abundant for both positive prediction and negative prediction. In general, SMO-2 show the best performance among all existing models.

Datasets	Tool	Sensitivity	Specificity	Accuracy	MCC
Positive + Negative 1	SMO-1	0.86	0.84	0.825	0.706
	AntiCP-1	1	0	0.500	/
	RFACP (MLACP)	0.96	0.78	0.850	0.721
	SVMACP (MLACP)	0.75	0.74	0.760	0.500
Positive + Negative 2	SMO-2	0.95	0.95	0.935	0.898
	AntiCP-2	0.91	0.88	0.895	0.790
	RFACP （MLACP）	0.76	0.99	0.860	0.773
	SVMACP （MLACP)	0.74	0.98	0.875	0.739

Table 5. comparison of my models and existing tools

Discussion

Some problems of conventional anticancer treatments, such as drug resistance and toxicity to other normal cells, make it necessary and urgent to discover other novel anticancer treatments [2, 3]. Among those promising treatments, anticancer peptides have received broad attention and interest. Due to the special structure of ACPs and its specific interaction with cancer cells, this special group of molecules can target and kill cancer cells without destroying other normal cells [4]. Prior to the wet-lab experiment, a computational predictive tool will definitely be helpful for the identification of ACPs. Moreover, in consideration of the similarity between ACPs and AMPs [9], it is regarded as a more efficient way of ACPs identification by searching from AMPs, because there are more examined sequence data of AMPs could be obtained.

Nevertheless, most of the existing tools only provide a function that identifies ACPs from all kinds of peptides [10-13]. Therefore, this work creates a tool with more functions and better performance of prediction. In order to achieve this goal, several innovative efforts or improvements have been made. First of all, more data are collected comparing to previous studies. In total, 1492 positive and 7068 negative (4433 for negative 1 and 2635 for negative 2) data are gathered from seven different sources. Then, balanced datasets with 463 sequences in each training dataset and 100 sequences, which are independent of training data, in each testing dataset are constructed. Another improvement in this research is that new features are chosen for characterization, including N5C5, k-space and PSSM. The innovative utilization of PSSM in anticancer prediction enhances the performance of model and shows significant benefit, indicating that PSSM should also be regarded as a key factor when separating ACPs and non-ACPs.

In model construction step, a better machine learning algorithm, SMO [15], is chosen and applied for classification, and increases the accuracy by 2.46% and 1.26% comparing to the SVM models. This performance suggests that SMO is a better choice than SVM in this case, and shows the success of SMO in text classification, proteomics projects, and analysis of high-dimensional data. Models built in this research is further compared with previous ACP prediction tools using an independent testing dataset. SMO-1 performs better than AntiCP-1 considering accuracy, specificity and MCC value. As for SMO-2, it performs better than all other tools in general, except the specificity of SMO-2 model is slightly lower than that of MLACP models. However, considering the fact that the sensitivity and MCC of MLACP models are significantly lower than those of SMO-2, which shows the imbalance of MLACP models, SMO-2 is still better.

Even though most of accessible data of examined ACPs are collected in this study, the amount is still not adequate. As a result, this research may have some limitation, and could be improved in the future with more sequence data.

Conclusions

This research presents a new scheme for the identification of ACPs, including utilizing a new important feature, PSSM, and a new helpful algorithm, SMO, for optimizing SVM for classification. In addition, this work offers two functions: (1) distinguishing ACPs from AMPs and (2) distinguishing ACPs from all kinds of peptides. With the help of SMO, optimized models perform better than ordinary models and other existing tools.

Methods

The process of this research is extracted and shown as a flowchart in **Figure 4**. Details of the process will be explained in following sections.

Dataset preparation

In this research, three datasets are constructed: positive dataset, negative dataset 1 and negative dataset 2. Positive dataset refers to anticancer peptides which are examined by experiments. They are collected form LEE dataset (total: 422) [13], Tyagi dataset (total: 450) [11], APD (total: 225)[16] and CancerPPD (total: 422) [17]. Negative dataset 1 is a collection of peptides that are anti-microbial but not anti-cancer. They are adapted from dbAMP dataset (total: 4057)[18] and Tyagi dataset (total: 1372). Peptides in negative dataset 2 are non-ACP, which are collected from Uniprot (total: 2635). Since anticancer peptides have been prove to be effective small molecules (<50 amino acids) [19], peptides longer than 50 amino acids are removed out of datasets. In addition, peptide contains artificial amino acids are removed. After this filtration step, 1492 peptide sequences in positive dataset, 4433 peptide sequences in negative dataset 1 and 2635 peptide sequences in negative dataset 2 are obtained. To reduce identical or similar peptides sequence, CD-HIT program [20] is utilized in this research. Results are shown as **Table 6**.

100% sequence-identity cut-off is applied on all of those three datasets. Then the processed positive dataset is compared with processed negative dataset 1 and processed negative dataset 2 separately using CD-HIT-2D [20]. It identifies and removes sequences in negative datasets that are similar to ones in positive dataset above a threshold of 40%. In addition, peptides that contained non-natural amino acids are removed. To balance datasets, some of peptide sequences are removed from negative datasets randomly. Ultimately, as shown in **Table 7**, each of those three datasets has 563 peptide sequences. Each dataset is then divided randomly into two subsets, the one contained 463 peptides is utilized as training dataset and the other one which contained 100 peptides is used as testing dataset.

	Positive	Negative 1	N1-P	Negative 2	N2-P
Original	1492	4433	/	2635	/
1.0	565	2753	2697	1585	/
0.9	398	2055	2559	1178	/
0.8	306	1664	2426	892	/
0.7	249	1358	2290	724	/
0.6	201	1097	2091	624	/
0.5	159	765	1667	531	/
0.4	107	439	1101	399	1494

Table 6. CD-HIT results of datasets

	Positive	Negative 1	Negative 2
Training set	463	463	463
Testing set	100	100	100

Table 7. Number of peptides in each dataset

Features Investigation

In order to utilize machine learning methods analyzing peptide sequences, features of sequences have to be extracted. In this research, 4 features are considered: amino acids composition (AAC), N5C5, k-space and position-specific scoring matrix (PSSM).

AAC

The AAC is the proportion of each amino acid in a given peptide sequence. It summarizes the peptide information in a vector of 20 dimensions. The AAC method has been successfully and widely applied in sequence-based classifications[21].

N5C5

Five amino acids from both N-terminal and C-terminal end of a given peptide are cut off and then connected as a novel sequence. Then the proportion of each amino acid in those new N5C5 sequences is calculated. Furthermore, to better analyze N5C5 sequences and visualize analysis results, heatmaps that show frequencies of each amino acid in each position are drawn.

K-space

K-space method extracts pairs of amino acids which have k (k = 0, 1, …) spacings form a given peptide sequence. In total, (N-k-1) pairs are selected from a peptide sequence which consists of N amino acids. After gathering all amino-acid-pairs, the frequency of each kind of pair is counted. In order to explore k-space diversity between positive dataset and those two negative datasets, the difference value of k-space frequency in positive dataset and that in negative datasets is then calculated. At last, those difference values of amino-acid-pairs are sorted, and ten pairs with the highest difference values are listed.

PSSM

PSSM is generated from a group of sequences previously aligned according to structural or sequence similarity. A PSSM for a given protein is an N 20 matrix P = {Pij : i = 1… N and j = 1 . . . 20 } , where N is the length of the protein sequence. It assigns a score Pij for the jth amino acid in the ith position of the query sequence. A large value indicates a highly conserved position while a small value indicates a weakly conserved position[22].

Model construction by machine learning techniques

In this study, supervised learning technique should be applied on text data for classification. Therefore, support vector machine (SVM)[14] is utilized in cooperation with sequential minimal optimization (SMO) [15]. For model construction, WEKA software (version 3.8.4) [23], and packages including LIBSVM (version 3.24) [24] and SMO package (using default parameters) within WEKA are utilized.

SVM is a data-driven supervised algorithm which constructs separating hyperplanes in high-dimensional space and selects the maximum-margin one for classification[25]. Based on its solid theoretical foundations, SVM has been successfully applied in various recognition and classification studies, including text classification[26], which is utilized in this research. SVM has also been successfully and widely used for high-dimensional biological data, including examination of gene expression profiles[27], mass spectra and genomics projects[28]. Comparing to other classifiers, such as artificial neural networks, SVM shows higher accuracy, particularly when numbers of features are large[28]. Furthermore, to improve the performance of SVM model, a program is designed to determine the optimum value of weight vector for each model in this research. As for adjusting gamma and cost value, a program in LIBSVM package[24] is applied to each model.

However, SVM does have some problems, including complexity and slow training speed for large-scale data. In order to solve these problems, another algorithm, SMO, is also applied for classification and shows both faster speed and better performance. SMO is a new algorithm for training SVMs, which breaks large quadratic programming (QP) optimization problem, a significant obstacle in original SVM algorithm, into a series smallest possible QP problem. By solving those smaller QP problems analytically, a time-consuming numerical QP optimization as an inner loop could be circumvented, and thus the computational time is shortened. The SVM maximization problem is as:

$$\max_{\lambda} \sum_{j=1}^{m} \lambda_j - \frac{1}{2} \sum_{j=1}^{m} \sum_{k=1}^{n} \lambda_j \lambda_k y_j y_k x_j x_k , 0 \leq \lambda_j \leq C, \forall_j, \sum_{j=1}^{m} y_j \lambda_j = 0$$

where, λ is Lagrange multiplier, x is input data and y represent class label. In SMO, two Lagrange multipliers are optimized while all other multipliers are kept constant using this equation[29]:

$$\lambda_1 y_1 + \lambda_2 y_2 = -\sum_{j=3}^{m} \lambda_j y_j = c.$$

Moreover, since SMO only utilizes linear amount of memory, it can handle very large training sets[15], which is perfectly aligned with the need in biological data analysis. To compute a linear SVM, only one weight vector needs to be stored. The stored weight vector can be easily updated to reflect new Lagrange multiplier values by:

$$\vec{w}^{\text{new}} = \vec{w} + y_1(\alpha_1^{new} - \alpha_1)\overrightarrow{x_1} + y_2(\alpha_2^{new,clipped} - \alpha_2)\overrightarrow{x_2}.[15]$$

This algorithm has shown success in some biological applications, such as metabolism studies[30], genomics[31] and molecular studies[32].

Performance evaluation

To evaluate performance of machine learning models, four indexes are calculated: accuracy, specificity (SP), sensitivity (SN) and Matthews correlation coefficient (MCC). Details of these metrics are shown as following equations:

$$\mathrm{SP} = \frac{TN}{TN+FP}$$

$$\mathrm{SN} = \frac{TP}{TP+FN}$$

$$\mathrm{ACC} = \frac{TP+TN}{TP+TN+FP+FN}$$

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

where TP-true positive-represents the number of correctly predicted positive labels, TN-true negative-refers to the number of corrected predict negative labels, FP-false positive-represents the number of positive labels that are wrongly predicted as negative, and FN-false negative-refers to the number of negative labels that are wrongly predicted as positive by the classifier. In addition to those evaluation metrics, receiver operating characteristic (ROC) curve is also generated in the step of weight adjustment to visualize the relationship of true positive rate and false positive rate, and used for comparison of performance.

Cross-validation and independent testing sets

In order to enhance robustness of the prediction model, ten-fold cross-validation is applied in model training step. In addition, to evaluate model built in this research and compare its performance with that of other existing tools, independent testing datasets are constructed in the dataset preparation step.

Abbreviations

AAC: amino acid composition

ACP: anticancer peptides

AMP: antimicrobial peptides

PSSM: position specific scoring matrix

SMO: sequential minimal optimization

SVM: support vector machine

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The dataset generated and analyzed during the current study are available at: https://github.com/georginawan/proteomics_project

Funding

This work was supported by the Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, China.

Authors' contributions

TYL contributed to the study concept and design. YW carried out the data analysis and drafted the manuscript. TYL and YW helped conceive the study. All authors contributed to the data interpretation and revised the manuscript.

Acknowledgements

I would like to thank the support from Undergraduate Research Award, CUHKSZ.

References

1. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. **68**(6): p. 394-424.
2. Palumbo, M.O., et al., *Systemic cancer therapy: achievements and challenges that lie ahead.* Front Pharmacol, 2013. **4**: p. 57.
3. Gatti, L. and F. Zunino, *Overview of tumor cell chemoresistance mechanisms.* Methods Mol Med, 2005. **111**: p. 127-48.
4. Gaspar, D., A.S. Veiga, and M.A. Castanho, *From antimicrobial to anticancer peptides. A review.* Front Microbiol, 2013. **4**: p. 294.
5. Schweizer, F., *Cationic amphiphilic peptides with cancer-selective toxicity.* Eur J Pharmacol, 2009. **625**(1-3): p. 190-4.
6. Riedl, S., D. Zweytick, and K. Lohner, *Membrane-active host defense peptides--challenges and perspectives for the development of novel anticancer drugs.* Chem Phys Lipids, 2011. **164**(8): p. 766-81.

7. Thundimadathil, J., *Cancer treatment using peptides: current therapies and future prospects.* J Amino Acids, 2012. **2012**: p. 967347.
8. Hoskin, D.W. and A. Ramamoorthy, *Studies on anticancer activities of antimicrobial peptides.* Biochim Biophys Acta, 2008. **1778**(2): p. 357-75.
9. van Zoggel, H., et al., *Antitumor and angiostatic activities of the antimicrobial peptide dermaseptin B2.* PLoS One, 2012. **7**(9): p. e44351.
10. Hajisharifi, Z., et al., *Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test.* J Theor Biol, 2014. **341**: p. 34-40.
11. Tyagi, A., et al., *In silico models for designing and discovering novel anticancer peptides.* Sci Rep, 2013. **3**: p. 2984.
12. Chen, W., et al., *iACP: a sequence-based tool for identifying anticancer peptides.* Oncotarget, 2016. **7**(13): p. 16895-909.
13. Manavalan, B., et al., *MLACP: machine-learning-based prediction of anticancer peptides.* Oncotarget, 2017. **8**(44): p. 77121-77136.
14. Osuna, E.E., *Support vector machines : training and applications*. 1998, Massachusetts Institute of Technology, Sloan School of Management. p. 202 p.
15. Platt, J.C., *Fast Training of Support Vector Machines using Sequential Minimal Optimization.* Microsoft Research, 2000.
16. Wang, G., X. Li, and Z. Wang, *APD3: the antimicrobial peptide database as a tool for research and education.* Nucleic Acids Res, 2016. **44**(D1): p. D1087-93.
17. Tyagi, A., et al., *CancerPPD: a database of anticancer peptides and proteins.* Nucleic Acids Res, 2015. **43**(Database issue): p. D837-43.
18. Jhong, J.H., et al., *dbAMP: an integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data.* Nucleic Acids Res, 2019. **47**(D1): p. D285-D297.
19. Li, F.M. and X.Q. Wang, *Identifying anticancer peptides by using improved hybrid compositions.* Sci Rep, 2016. **6**: p. 33910.
20. Huang, Y., et al., *CD-HIT Suite: a web server for clustering and comparing biological sequences.* Bioinformatics, 2010. **26**(5): p. 680-2.
21. Usmani, S.S., S. Bhalla, and G.P.S. Raghava, *Prediction of Antitubercular Peptides From Sequence Information Using Ensemble Classifier and Hybrid Features.* Front Pharmacol, 2018. **9**: p. 954.
22. Gribskov, M., A.D. McLachlan, and D. Eisenberg, *Profile analysis: detection of distantly related proteins.* Proc Natl Acad Sci U S A, 1987. **84**(13): p. 4355-8.
23. Eibe Frank, Mark A. Hall, and Ian H. Witten. The WEKA Workbench. *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann, Fourth Edition, 2016.
24. Chih-Chung Chang, C.-J.L., *LIBSVM : a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, 2011.

25. Noble, W.S., *What is a support vector machine?* Nat Biotechnol, 2006. **24**(12): p. 1565-7.

26. Tong, S. and D. Koller, *Support vector machine active learning with applications to text classification.* Journal of Machine Learning Research, 2002. **2**(1): p. 45-66.

27. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.

28. Byvatov, E. and G. Schneider, *Support vector machine applications in bioinformatics.* Appl Bioinformatics, 2003. **2**(2): p. 67-77.

29. Naveed, H., et al., *Human activity recognition using mixture of heterogeneous features and sequential minimal optimization.* International Journal of Machine Learning and Cybernetics, 2019. **10**(9): p. 2329-2340.

30. Chen, L., C. Chu, and K. Feng, *Predicting the types of metabolic pathway of compounds using molecular fragments and sequential minimal optimization.* Comb Chem High Throughput Screen, 2016. **19**(2): p. 136-43.

31. Huang, T., Y. Shu, and Y.D. Cai, *Genetic differences among ethnic groups.* Bmc Genomics, 2015. **16**.

32. Periwal, V., Rajappan, J.K., Jaleel, A.U. et al, *Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets.* BMC Res Notes 4, 2011.

Figures

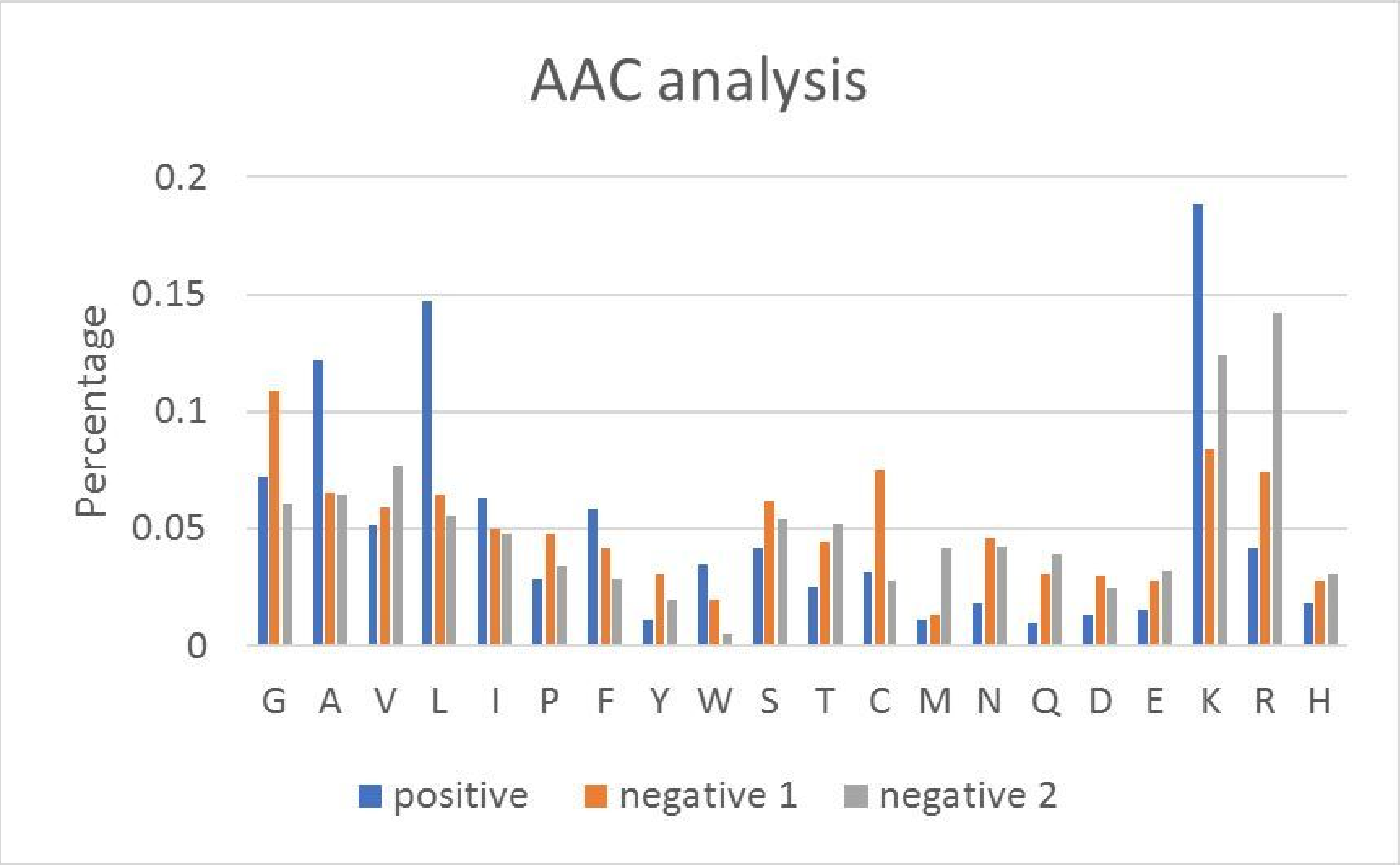


Figure 1

AAC analysis of positive, negative 1, negative 2

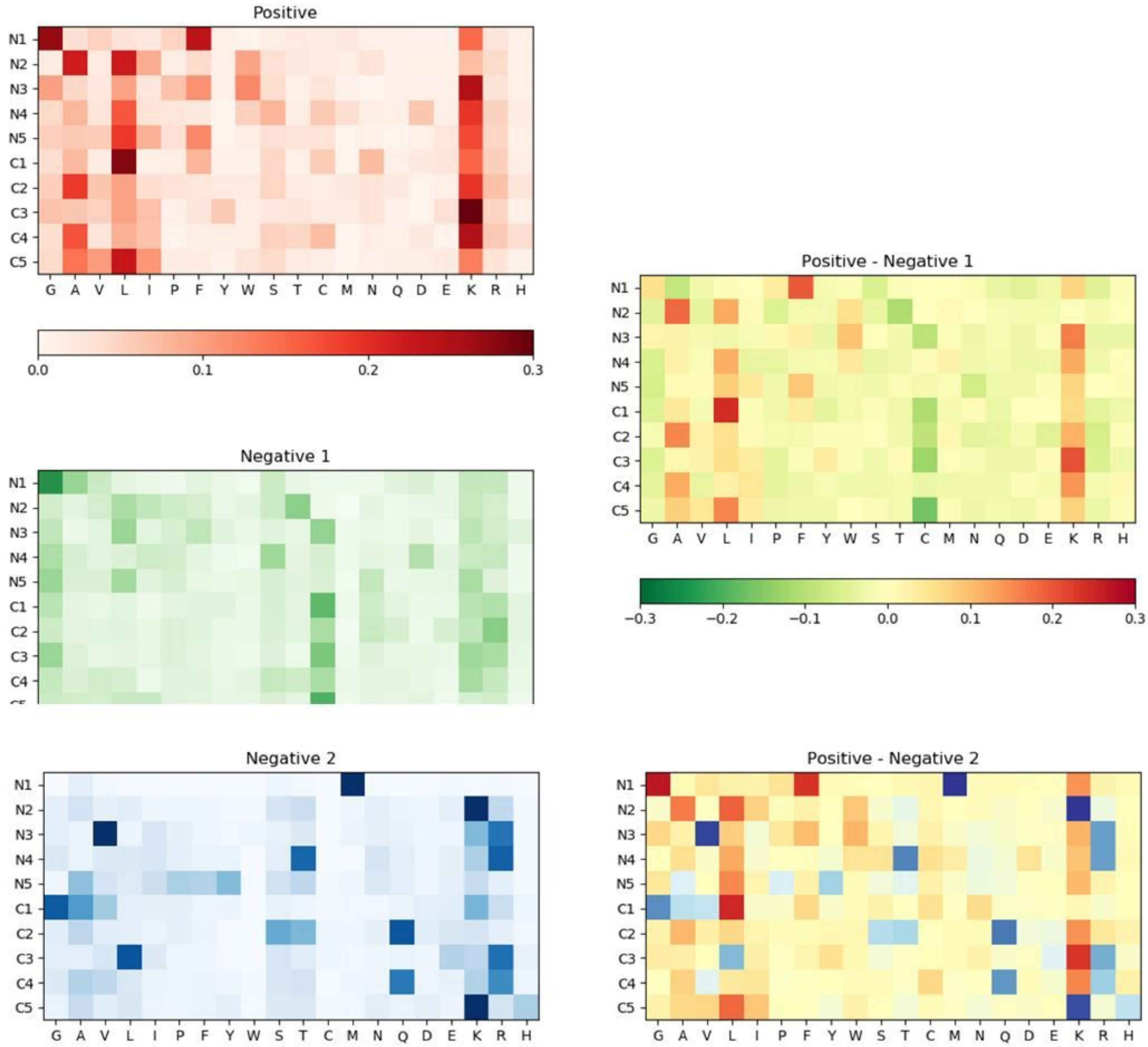


Figure 2

N5C5 analysis results

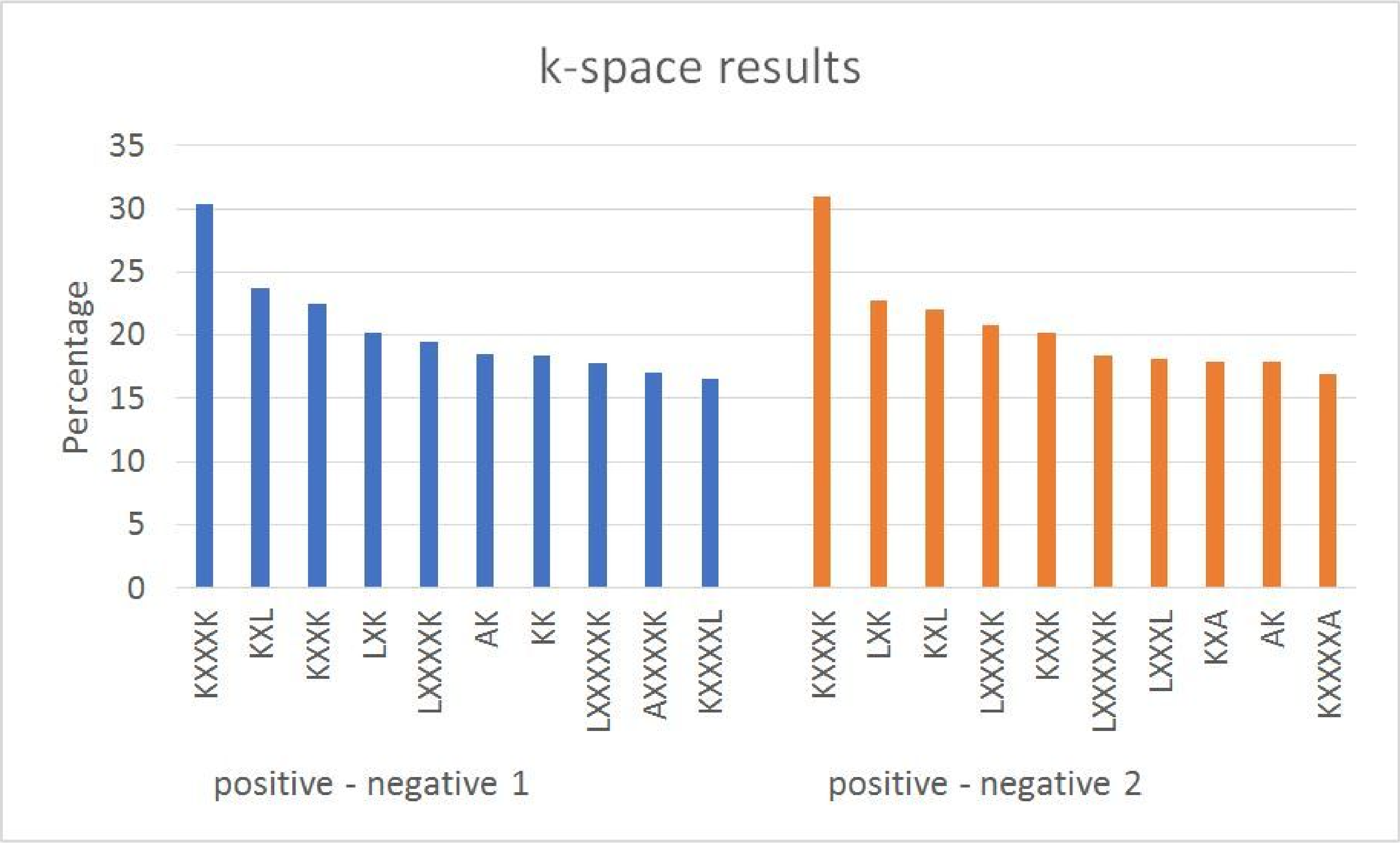
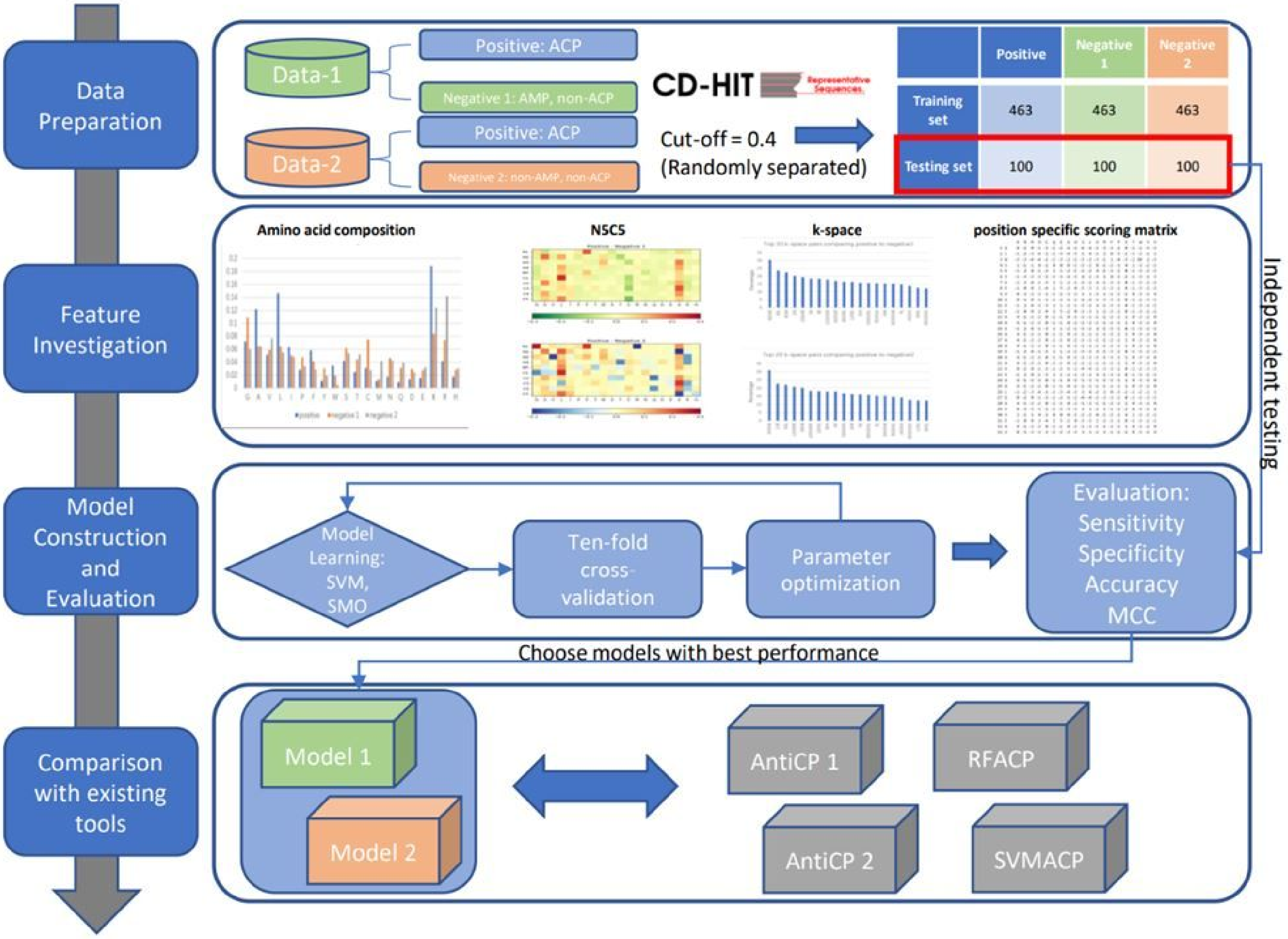


Figure 3

10 most different k-space pairs comparing positive to negative 1, positive to negative 2



	Positive	Negative 1	Negative 2
Training set	463	463	463
Testing set	100	100	100

Figure 4

Flowchart of this work