SUPPORT VECTOR MACHINE AND GAUSSIAN CLASSIFICATION OF ANTIMICROBIAL PEPTIDES

Kandasamy Nagarajan[1#], Anjali Saini[2#], Siddheshwari Dutt Mishra[3], Ajay Kumar Shrivastava[4], Garima Kapoor[5*], Parul Grover[6]

Abstract

Today's scenario of rapidly increasing antibacterial drug resistance is a major problem. So, to prevent from microbial infection and multi-drug resistance AMP$_S$ (antimicrobial peptides) has been highlighted in recent years. AMP$_S$ are a unique group of shorter to longer chain of molecules that can target and resist the bacterial infection directly. However, identifying AMP$_S$ by lab-experiments is time consuming and costly. Therefore, it is significant to develop computational tool for AMP$_S$ prediction. Though some AMP$_S$ prediction tools have been developed recently, their performances are not well enough to distinguish the AMP$_S$ from anticancer peptides and anti-diabetic peptides. In this systematic study, the selected 180 peptide's predictions are analyzed through the SVM (Support Vector Machine) machine learning method. In addition to SVM, Gaussian classifier is used in this research for optimizations. The linear SVM method shows the best model for the classification of AMP$_S$. The best performance was shown in class of zero violations as compared with class of one, two, three violations. This study anticipated smaller chain dipeptides show high potency against antibacterial drug resistance and prevent bacterial infections.

Key words- Antimicrobial, Dipeptide, Gaussian classifier SVM.

[1,2,5*,6]KIET School of Pharmacy, KIET Group of Institutions, Delhi-NCR, Ghaziabad, Uttar Pradesh, India E-mail addresses: [2]anjalikiet1997@gmail.com, [5]kapoor27garima@gmail.com
[3,4]KIET Group of Institutions, Delhi-NCR, Ghaziabad, Uttar Pradesh, India
# Both have contributed equally so consider both of them as first Author.

***Corresponding Author:** Dr. Garima Kapoor
*Assistant Professor, KIET School of Pharmacy, KIET Group of Institutions, Delhi NCR Ghaziabad-Meerut Road, NH-58, Ghaziabad, UP-201206, India, Phone: +91-964-350-3380,
E-mail addresses: anjalikiet1997@gmail.com, kapoor27garima@gmail.com

DOI: 10.48047/ecb/2023.12.si5a.0374

Introduction

Kernel methods are a type of machine learning algorithm that incorporates a kernel function to model nonlinear relationships. Standard kernel methods assume that a given kernel function is positive-definite, but recent developments in the theory of learning with indefinite kernels have now removed this requirement, allowing a much broader class of functions to be incorporated into kernel methods [1-3]. We study the effectiveness of learning with established sequence-similarity functions for the classification of antimicrobial peptides (AMPs) based on their amino acid sequences. We evaluate the ability of the proposed methodology to predict both general and species-specific antimicrobial activity. AMPs, also known as host defense peptides, are a class of molecules that form an important component of the innate immune system [4-6]. These molecules typically have certain properties, including cationic, 30-50% hydrophobicity, and amphiphilicity. They exhibit good antimicrobial activity against a broad range of bacteria, viruses, fungi, and parasites. In addition, they have an inherently low risk of developing antimicrobial resistance. Such broad-spectrum and rapid antimicrobial activity has prompted researchers to consider AMPs as a potential remedy to the growing problem of AMR [4-8].

Despite some early success in translating AMP-based therapies to clinical use, there are many challenges that researchers have yet to overcome, such as the complex structure-activity relationship (SAR) and the high cost and time required for wet-lab experiments [9] [10]. Many researchers have turned to computational approaches, such as molecular dynamics (MD) simulations and machine learning (ML) algorithms, to speed up the discovery and development of potential AMPs [11] [12]. Several studies have highlighted the potential of ML algorithms in predicting antimicrobial activity, dissecting the SAR, and informing the drug design of AMPs. A wide range of ML algorithms has been used, including random forests, support vector machines (SVMs), and artificial neural networks [13-19]. Many of these algorithms are used in combination with a carefully selected set of peptide features, which can be divided into two categories: compositional and physicochemical. The amino acid composition is a simple example of a compositional feature, which is a vector containing counts of each amino acid in a given peptide [20-24]. There are various extensions, such as the reduced amino acid composition and the pseudo amino acid composition, which take into account composition as well as sequence-order information [27]. The set of physicochemical features includes peptide properties such as the charge, hydrophobicity, and isoelectric point [28] [29].

Classical arrangement calculations, like the Smith Waterman [30] and Needleman- Wunsch [31] calculations, are computationally concentrated and don't scale well to enormous issues. Many papers have upheld the utilization of arrangement-free techniques to decide succession comparability [19], [32-35]. The outcome of these undertakings regardless, of grouping arrangement capabilities, are compelling thoughts of natural succession similitude that can reflect hereditary, primary, or utilitarian comparability and subsequently ought not to be ignored. A few investigations have used grouping arrangement capabilities for AMP expectations. For instance, Wang et al. [36] and Ng et al. [37] used the Impact algorithm [38] in a grouping model. While these methodologies prompted exact models, the Impact calculation is a heuristic technique that tracks down just inexact ideal arrangements. Subsequently, it is fascinating to consider whether similar methodologies utilizing the ideal arrangement score would work on the models. The SVM is a notable ML calculation for grouping and can consolidate a bit of capability to learn non-direct characterization limits. The piece capability enormously impacts the presentation of the subsequent arrangement model. At the point when properly standardized, a portion capability can be viewed as a comparability capability. A valuable bit capability ought to create likenesses that are pertinent to the issue. Numerous expressive thoughts of similitude are not legitimate portion capabilities [39-42], in that they are endless, meaning they can't be utilized with an SVM. Late improvements have now reduced this issue, working with a lot bigger class of similitude capabilities to be utilized related to an SVM. Loosli et al. [1] present a calculation for learning an SVM with endless pieces. Their methodology depends on a strategy for adjustment, significance there is no assurance of worldwide optimality. Then again, the SVM is a calculation of endless portions that are ensured to find an internationally ideal solution. In this work, we used the SVM calculation to evaluate the viability of grouping arrangement capabilities for AMP orders. We played out an observational examination of both the neighborhood arrangement score and the AMP datasets from the writing.

Material and Methods

Materials

In this study, SVM and Gaussian classifier to use for categorize the drugs data set based on their ascent. Initially *In silico* approaches shows different 14 parameters, here we considered 180 types of drugs (trained data) with 14 features, and these features were MI log p, total polar surface area (TPSA), molecular weight, violations, number of rotatable bonds, volume, enzyme bioactivity score (GPCR ligand, ICM, KI, NRL, PI, EI, number of hydrogen bond acceptor, number of hydrogen bond donor, molar refractivity, Log P. These data obtained from SWISS ADME predictions online tool (available in http://www.swissadme.ch/) [43], Chemi-informatics Molinspiration software (available in https://www.molinspiration.com/cgi-bin/properties) [44].

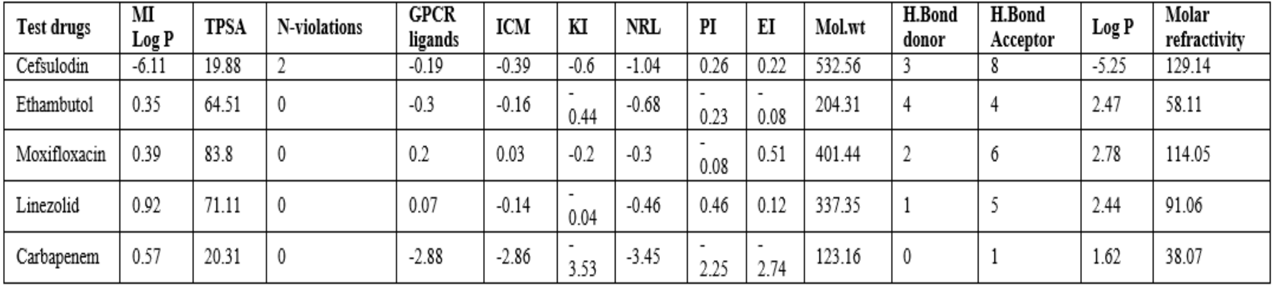
Methods for performance of SVM and Gaussian classifier

In Support Vector Machine, ordering model has been developed using SVM algorithm. Total of 180 peptides sequences classified according to the n-violations which were recorded from MOLINSPIRATION online tool and 5 anti-bacterial drugs were considered. The peptides were converted into format suitable for SVM algorithm. To better classification, we need to do preprocessing of the data set, so to do preprocess in our data set; we need to make then a numerical value. we have some value which is not numeric, so based on their nature we can assign them as a number value, we here just make a change 0 for low value and 1 for high value and at the same time we need to rescale our data set, rescale means we need to make the range of data set in a fixed value because, if our data set is in very large range so, the machine learning classifier is not able to perform well, to overcome this problem, here we make our drug data set in a fixed range value of -10 to 5. Since machine learning, we need to extract the features manually, so with the help of these features we can classify the data to do better classification, we need to split our data set into two parts namely training and testing. So, to do this training and test ratio we used 80:20 ratios. In this study 180 peptides with the following mentioned 14 parameters containing 4 groups Gly-Asp to Asp-Val (33 Trained Data set) were in Class-0 violations as from **Table-2**, Ala-Asn to Gly-Asn-Gly (26 Trained Data set) were in Class-1 violations as from **Table-3**, Gly- Asn-Asp to Gly-Gly-Gly-Gly-Gly-Ser (28 Trained Data set) were in Class-2 violations as from **Table-4**, Gly-Ala-Arg-Asn to Gly-Gly- Gly-Gly-Val (64 Trained Data set) were in Class-3 violations as from **Table-5**. Whereas the five drug samples (Cefsulodin, Ethambutol, moxifloxacin, linezolid and Carbapenem were used name as test data set with same 14 parameters. Validate the data through the Support Vector Machine (SVM) and Gaussian classifier, and check whether the trained data set is near the test data set or not.

Results and Discussion

The result of the standard drug (Cefsulodin, Ethambutol, moxifloxacin, linezolid and Carbapenem) used as test data sets for SVM algorithm and Gaussian classifier were represented as follows:

Table1: Standard drug as Test data set for SVM and Gaussian classifier.



Test drugs	MI Log P	TPSA	N-violations	GPCR ligands	ICM	KI	NRL	PI	EI	Mol.wt	H.Bond donor	H.Bond Acceptor	Log P	Molar refractivity
Cefsulodin	-6.11	19.88	2	-0.19	-0.39	-0.6	-1.04	0.26	0.22	532.56	3	8	-5.25	129.14
Ethambutol	0.35	64.51	0	-0.3	-0.16	-0.44	-0.68	-0.23	-0.08	204.31	4	4	2.47	58.11
Moxifloxacin	0.39	83.8	0	0.2	0.03	-0.2	-0.3	-0.08	0.51	401.44	2	6	2.78	114.05
Linezolid	0.92	71.11	0	0.07	-0.14	-0.04	-0.46	0.46	0.12	337.35	1	5	2.44	91.06
Carbapenem	0.57	20.31	0	-2.88	-2.86	-3.53	-3.45	-2.25	-2.74	123.16	0	1	1.62	38.07

The peptides were arranged according to their n-violations number which was recorded by using chemo-information Molinspiration online tool. In the below mentioned table all the selected peptides shows the class of zero violations.

Table-2: Represent the trained data sets of Class 0 violation peptides.

Trained data (Peptides)	MI Log P	TPSA	N-violations	GPCR ligands	ICM	KI	NRL	PI	EI	Mol.wt	H. Bond donor	H. Bond Acceptor	Log P	Molar refractivity
Gly-Asp	-3.98	129.72	0	0.04	0.17	-0.85	-0.83	0.29	0.24	190.6	4	6	-2.25	40.21
Gly-Ala	-3.16	92.42	0	-0.53	-0.21	-1.29	-1.59	-0.09	-0.21	146.15	3	4	-1.51	33.63
Gly-Glu	-3.71	129.72	0	0.09	0.17	-0.7	-0.64	0.3	0.31	204.18	4	6	-1.93	45.02
Gly-Lue	-1.85	92.42	0	-0.05	0.18	-0.94	-0.76	0.38	0.16	188.23	3	4	-0.58	48.05
Gly-Ser	-4.12	112.65	0	-0.19	0.03	-0.96	-1.29	0.13	0.16	162.15	4	5	-2.26	34.79
Gly-Cys	-3.18	92.42	0	-0.2	-0.08	-1.1	-1.29	0.66	0.46	178.21	3	4	-1.53	41.56
Gly-Ile	-1.87	92.42	0	-0.23	0.04	-0.99	-1.19	0.22	0.08	188.23	3	4	-0.62	48.05

Gly-Met	-2.7	92.42	0	-0.26	-0.12	-1.43	-1.04	0.25	0.29	206.27	3	4	-0.82	50.83
Gly-Val	-2.38	92.42	0	-0.3	-0.04	-1.03	-1.25	0.12	-0.03	174.2	3	4	-0.86	43.24
Gly-phe	1.7	92.42	0	0.25	0.24	-0.41	-0.47	0.43	0.27	222.24	3	4	0.23	58.12
Pro-Gln	-3.36	121.52	0	0.39	0.43	-0.18	-0.26	0.82	0.44	243.26	4	5	-1.62	62.47
Pro-Thr	-2.9	98.65	0	0.2	0.27	-0.57	-0.45	0.55	0.35	216.24	4	5	-1.35	55.91
Pro-Phe	-0.83	78.42	0	0.52	0.45	-0.1	-0.11	0.78	0.39	262.31	3	4	0.46	74.43
Pro-Val	-1.51	78.42	0	0.14	0.36	-0.53	-0.64	0.59	0.23	214.26	3	4	-0.29	59.56
Pro-ile	-1	78.42	0	0.17	0.39	-0.51	-0.63	0.64	0.3	228.29	3	4	0.04	64.37
Pro-Ala	-2.29	78.42	0	-0.02	0.3	-0.7	-0.88	0.44	0.11	186.21	3	4	-0.89	49.95
Pro-Cys	-2.31	78.42	0	0.24	0.37	-0.57	-0.66	1.02	0.62	218.28	3	4	-0.93	57.88
Pro-Met	-1.83	78.42	0	0.15	0.26	-0.86	-0.51	0.67	0.46	246.33	3	4	-0.38	67.15
Pro-Tyr	-1.3	98.65	0	0.57	0.47	-0.01	0.09	0.78	0.43	278.31	4	5	0.02	76.46
Pro-Lys	-2.78	104.45	0	0.57	0.65	-0.07	-0.25	0.95	0.59	243.31	4	5	-0.88	67.07
Ala-Cys	-3.28	92.42	0	-0.08	-0.08	-0.92	-1.01	0.83	0.58	192.24	3	4	-1.22	46.37
Ala-Phe	-1.79	92.42	0	0.32	0.2	-0.3	-0.3	0.6	0.37	236.27	3	4	0.05	62.92
Ala-Ile	-2.78	104.45	0	0.57	0.65	-0.07	-0.25	0.95	0.59	104.45	4	5	-0.88	67.07
Ala-Met	-2.8	92.42	0	-0.14	-0.12	-1.23	-0.81	0.45	0.42	92.42	3	4	-0.61	55.64
Ala-Val	-2.48	92.42	0	-0.17	-0.03	-0.83	-1	0.33	0.15	92.42	3	4	-0.6	48.05
Ala-Tyr	2.27	112.65	0	0.4	0.24	-0.18	-0.06	0.61	0.42	112.65	4	5	-0.54	64.95
Ala-Trp	-1.64	108.21	0	0.65	0.3	0.17	-0.1	0.7	0.5	108.21	4	4	0.06	74.78
Asp-Ile	-2.8	129.72	0	0.24	0.32	-0.45	-0.43	0.71	0.41	246.26	4	6	-0.83	59.44
Asp-Gly	-3.87	129.72	0	0.01	0.19	-0.73	-0.63	0.46	0.25	190.16	4	6	-2.11	40.21
Asp-Phe	-2.62	129.72	0	0.56	0.38	-0.09	0.02	0.83	0.46	280.28	4	6	-0.69	69.5
Asp-Cys	-4.1	129.72	0	0.31	0.31	-0.5	-0.44	1.06	0.73	236.25	4	6	-1.72	52.94
Asp-Met	-3.63	129.72	0	0.22	0.2	-0.79	-0.33	0.73	0.56	264.3	4	6	-1.22	62.22
Asp-Val	-3.3	129.72	0	0.22	0.3	-0.46	-0.44	0.66	0.36	232.24	4	6	-1.39	54.63

The peptides were arranged according to their n-violations number which was recorded by using chemo-information Molinspiration online tool. In the below mentioned table all the selected peptides shows the class of one violations.

Table-2: Represent the trained data sets of Class 1 violation peptides.

Trained data (Peptides)	**MI Log P**	**TPSA**	**N-violations**	**GPCR ligands**	**ICM**	**KI**	**NRL**	**PI**	**EI**	**Mol.wt**	**H.Bond donor**	**H.Bond Acceptor**	**Log P**	**Molar refractivity**
Ala-Asn	-3.37	135.51	1	-0.01	-0.04	-0.64	-0.64	0.47	0.27	203.2	4	5	-2.14	46.15
Ala-Gln	-4.28	135.51	1	0.11	0.07	-0.51	-0.46	0.52	0.35	217.22	4	5	-1.9	50.96
Ala-Lys	-3.75	118.44	1	0.33	0.34	-0.37	-0.46	0.69	0.52	217.27	4	5	-0.96	55.57
Asp-Glu	-4.49	167.02	1	0.42	0.36	-0.25	-0.1	0.72	0.51	262.22	5	8	-2.12	55.4
Asp-Thr	-4.53	149.95	1	0.22	0.34	-0.55	-0.24	0.62	0.37	234.21	5	7	-2.15	50.98
Asp-Tyr	-3.1	149.95	1	0.6	0.39	-0.01	0.19	0.82	0.49	296.28	5	7	-1.18	71.53
Asp-Lys	-4.45	155.74	1	0.6	0.55	-0.09	-0.07	0.93	0.63	261.28	5	7	-2.02	62.14
Gly-Ala-Arg	-4.58	185.93	1	0.66	0.41	-0.11	-0.32	1.02	0.5	302.33	6	6	-2.25	74.57
Gly-Cys-Glu	-4.26	158.82	1	0.51	0.25	-0.04	-0.14	1.08	0.76	307.33	5	7	-1.98	70.37
Gly-Gln-Gly	-4.46	164.61	1	0.28	0.1	-0.2	-0.21	0.68	0.34	260.25	5	6	-2.5	58.77
Gly-His-Ile	-2.75	150.2	1	0.65	0.39	0.16	-0.69	0.91	0.61	325.37	5	6	-0.89	82.11
Gly-Leu-Lys	-2.9	147.54	1	0.56	0.34	-0.06	0.06	0.93	0.45	316.4	5	6	-0.58	82.6
Gly-Thr-Trp	-2.72	157.54	1	0.65	0.21	0.12	0.05	0.76	0.42	362.39	6	6	-0.94	93.37
Gly-Tyr-Val	-1.96	141.75	1	0.46	0.25	0.01	0.01	0.74	0.33	337.38	5	6	-0.34	87.18
Gly-Ala-Asp	-4.44	158.82	1	0.38	0.24	-0.35	-0.42	0.65	0.38	261.23	5	7	-2.31	57.63
Gly-Gln-Ile	-3.51	164.61	1	0.36	0.14	-0.12	-0.2	0.78	0.34	316.36	3	6	-1.47	77.99
Gly-His-Lys	-4.42	176.22	1	0.8	0.49	0.3	-0.44	0.97	0.71	340.38	6	7	-1.74	84.81
Gly-Met-Ser	-4.22	141.75	1	0.31	0.1	-0.49	-0.36	0.74	0.55	293.35	5	6	-1.75	69.42
Gly-Gln-Lys	-4.83	190.63	1	0.53	0.25	0.05	0.06	0.83	0.43	331.37	6	7	-2.81	80.7
Gly-His-Phe	-2.57	150.2	1	0.72	0.41	0.26	-0.45	0.86	0.58	359.38	5	6	-0.77	92.17
Gly-Leu-Ser	-3.38	141.75	1	0.45	0.31	-0.16	-0.16	0.84	0.47	275.3	5	6	-1.58	66.64
Gly-Met-Trp	-1.66	137.31	1	0.55	0.18	-0.08	-0.1	0.74	0.45	392.47	5	5	0.29	104.6
Gly-Thr-Val	-3.55	141.75	1	0.31	0.07	-0.34	-0.24	0.7	0.35	275.31	5	6	-1.77	66.64
Gly-Tyr-Asp	-3.56	179.04	1	0.58	0.32	0.02	0.14	0.79	0.42	353.33	6	8	-1.58	84.14
Gly-Ala-Glu	-4.24	158.82	1	0.4	0.2	-0.26	-0.3	0.65	0.41	275.26	5	7	-1.87	62.44
Gly-Asn-Gly	-3.62	164.61	1	0.19	0.03	-0.28	-0.35	0.65	0.28	246.22	5	6	-2.96	53.96

The peptides were arranged according to their n-violations number which was recorded by using chemo-information Molinspiration online tool. In the below mentioned table all the selected peptides shows the class of two violations.

Table-2: Represent the trained data sets of Class 2 violation peptides.

Trained data (Peptides)	**MI Log P**	**TPSA**	**N-violations**	**GPCR ligands**	**ICM**	**KI**	**NRL**	**PI**	**EI**	**Mol.wt**	**H.Bond donor**	**H.Bond Acceptor**	**Log P**	**Molar refractivity**
Gly-Asn-Asp	-4.51	201.91	2	0.49	0.17	-0.07	-0.06	0.82	0.41	304.26	6	8	-3.42	65.35
Gly-Asn-Glu	-4.33	201.91	2	0.48	0.14	-0.03	0.01	0.79	0.42	318.29	6	8	-3.11	70.15
Gly-Gly-Gly-Gly-Gly	-2.97	232.78	2	0.02	-0.1	-0.11	-0.13	0.09	-0.04	456.41	6	28	-2.24	66.67
Gly-Gly-Gly-Gly-Ala	-4.29	179.71	2	0.25	0.03	-0.14	-0.26	0.5	0.18	317.3	6	30	-1.85	71.48
Gly-Gly-Gly-Gly-Asp	-4.6	220.09	2	0.26	-0.02	-0.24	-0.05	0.43	0.16	403.34	7	35	-2.53	87.96
Gly-Gly-Gly-Gly-His	-4.5	208.4	2	0.56	0.25	0.14	-0.54	0.67	0.47	383.36	7	33	-1.91	88.11

Gly-Gly-Gly-Gly-Val	-4.29	237.91	2	0.2	0.04	-0.12	-0.18	0.43	0.14	459.5	8	44	-2.2	106.32
Gly-Gly-Gly-Gly-Trp	-0.69	175.28	2	0.42	0.13	0.2	0	0.75	0.32	416.4	6	34	-0.03	106.25
Gly-Gly-Gly-Gly-Ile	-3.05	179.71	2	0.25	0.06	-0.2	-0.29	0.57	0.21	359.38	6	36	-0.83	85.9
Gly-Gly-Gly-Gly-Ser	-4.84	199.94	2	0.37	0.12	-0.05	-0.19	0.59	0.32	333.3	7	31	-2.88	72.64
Gly-Gly-Gly-Gly-Gln	-4.88	222.81	2	0.33	0.07	-0.09	-0.08	0.58	0.25	374.35	7	35	-1.91	84
Gly-Gly-Gly-Gly-Asn	-4.37	222.81	2	0.31	0.03	-0.1	-0.12	0.6	0.23	360.32	7	33	-2.3	79.19
Gly-Gly-Gly-Gly-Cys	-3.1	217.68	2	0.21	-0.08	0.03	-0.18	0.6	0.25	493.57	9	43	-2.17	115.8
Gly-Gly-Gly-Gly-Thr	-4.67	199.94	2	0.3	-0.01	-0.18	-0.12	0.53	0.26	347.32	7	33	-2.49	77.45
Gly-Gly-Gly-Ala-Gly	-3.94	217.68	2	0.15	-0.11	-0.1	-0.3	0.36	0.02	415.4	7	39	-2.38	95.14
Gly-Gly-Gly-Asn-Gly	-4.34	222.81	2	0.33	0	-0.03	-0.07	0.72	0.24	360.32	7	33	-2.3	79.19
Gly-Gly-Gly-Leu-Gly	-3.77	237.91	2	0.28	0.08	-0.09	-0.01	0.59	0.2	473.5	8	46	-1.81	114
Gly-Gly-Gly-Lys-Gly	-4.94	263.93	2	0.32	0.11	0	-0.02	0.56	0.24	488.5	9	48	-2.03	113.84
Gly-Gly-Gly-Met-Gly	-4.49	237.91	2	0.2	-0.04	-0.28	-0.12	0.53	0.24	491.5	8	45	-2.11	113.91
Gly-Gly-Gly-Pro-Gly	-4.13	208.89	2	0.26	-0.03	-0.08	-0.14	0.51	0.07	414.4	6	41	-2.34	106.55
Gly-Gly-Gly-Ser-Gly	-4.82	199.94	2	0.41	0.12	-0.05	-0.09	0.73	0.36	333.3	7	31	-2.88	72.64
Gly-Gly-Ala-Gly-Gly	-3.94	217.68	2	0.15	-0.11	-0.1	-0.3	0.36	0.02	415.4	7	39	-2.38	95.14
Gly-Gly-Gly-Gly-Gly-Gly	-4.41	208.81	2	0.2	0.01	-0.07	-0.1	0.42	0.17	360.32	7	33	-2.73	79.29
Gly-Gly-Gly-Gly-Gly-Ala	-4.33	243.71	2	0.17	-0.05	-0.16	-0.19	0.46	0.11	472.5	8	47	-1.92	112.27
Gly-Gly-Gly-Gly-Gly-Asn	-4.61	251.91	2	0.26	0.02	-0.08	-0.1	0.52	0.2	417.38	8	38	-2.79	91.81
Gly-Gly-Gly-Gly-Gly-Asp	-4.94	246.11	2	0.34	0.12	-0.11	-0.09	0.53	0.25	418.36	8	37	-2.89	90.76
Gly-Gly-Gly-Gly-Gly-gln	-5.03	251.91	2	0.28	0.06	-0.07	-0.07	0.51	0.22	431.34	5	40	-2.4	96.61
Gly-Gly-Gly-Gly-Gly-Ser	-4.84	199.94	2	0.37	0.12	-0.05	-0.19	0.59	0.32	333.3	7	8	-3.67	72.64

The peptides were arranged according to their n-violations number which was recorded by using chemo-information Molinspiration online tool. In the below mentioned table all the selected peptides shows the class of three violations.

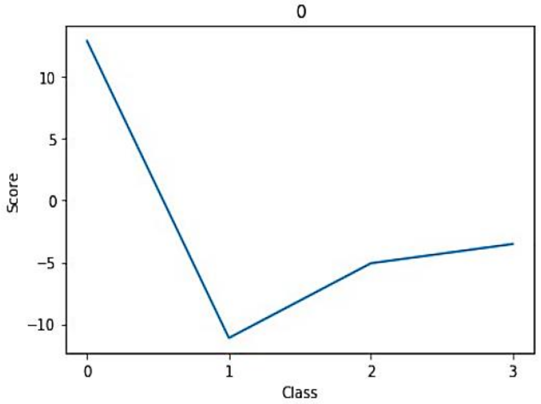
Table-2: Represent the trained data sets of Class 3 violation peptides.

Trained data (Peptides)	**MI Log P**	**TPSA**	**N-violations**	**GPCR ligands**	**ICM**	**KI**	**NRL**	**PI**	**EI**	**Mol.wt**	**H.Bond donor**	**H.Bond Acceptor**	**Log P**	**Molar refractivity**
Gly-Ala-Arg-Asn	-4.86	368.72	3	-0.65	-1.83	-1.59	-1.73	0.03	-1.17	755.83	12	13	-3.39	189.29
Gly-Asp-Cys-Gln	-5.96	573.37	3	-3.82	-3.9	-3.92	-3.92	-3.71	-3.83	1266.47	20	123	-1.01	308.21
Gly-Ile-Leu-Lys	-1.21	365.22	3	-3.63	-3.76	-3.77	-3.79	-3.35	-3.68	1030.32	12	109	5.61	277.02
Gly-Ser-Thr-Trp	-6.05	708.73	3	-4	-4.04	-4.05	-4.07	-3.96	-4.01	1477.91	26	150	-1.44	409.2
Gly-Tyr-Val-His	-4.53	357.54	3	-2.31	-3.32	-3.22	-3.55	-1.44	-2.8	869.98	12	81	0.54	226.05
Gly-Arg-Ala-Asn	-6.48	956.02	3	-4.07	-4.1	-4.11	-4.14	-4.04	-4.08	2009.13	29	204	6.35	552.28
Gly-Gly-Glu-His	-5.37	431.64	3	-3.31	-3.69	-3.67	-3.75	-2.82	-3.53	975.03	15	98	-2.18	254.4
Gly-Leu-Ile-Lys	-6.05	788.72	3	-4.06	-4.09	-4.11	-4.11	-4.02	-4.06	1924.37	34	224	3.25	596.98
Gly-Phe-Met-Pro	-3.71	298.68	3	-1.82	-3.14	-3.02	-2.9	-0.98	-2.38	842.97	10	89	8.31	309.17
Gly-Thr-Ser-Trp	-4.34	432.75	3	-3.6	-3.78	-3.77	-3.79	-3.32	-3.68	1038.13	15	92	0.83	261.88
Gly-Asn-Arg-Ala	-5.39	452.35	3	-3.44	-3.71	-3.74	-3.73	-2.86	-3.58	1008.13	14	93	0.42	253.43
Gly-Gln-Cys-Asp	-6.83	1113.57	3	-4.13	-4.16	-4.16	-4.18	-4.11	-4.13	2473.65	42	216	-6.75	587.63
Gly-His-Glu-Gly	-5.51	367.21	3	-0.21	-1.22	-0.91	-1.31	0.17	-0.59	712.72	12	65	-1.23	167.66
Gly-His-Val-Tyr	-4.49	376.68	3	-3.31	-3.67	-3.67	-3.75	-3.72	-3.52	972.11	20	127	1.45	344.79
Gly-Ala-Asn-Arg	-5.22	414.53	3	-0.83	-2.04	-1.77	-1.91	-0.12	-1.34	771.88	18	124	-1.71	328.48

Gly-Asp-Gln-Cys	-5.56	324.52	3	0.53	0.06	-0.03	-0.07	0.98	0.44	577.6	11	54	-0.94	136.45
Gly-Glu-His-Gly	-4.62	288.79	3	0.59	0.21	0.2	-0.31	0.71	0.46	512.5	9	44	-1.73	116.82
Gly-Ile-Lys-Leu	-4.67	487.51	3	-3.84	-3.93	-3.94	-3.98	-3.77	-3.87	1354.7	0	0	0	0
Gly-Met-Pro-Phe	-1.86	382.04	3	-3.73	-3.84	-3.87	-3.9	-3.58	3.76	1141.4	0	0	0	0
Gly-Ser-Trp-Thr	-3.67	359.36	3	-3.64	-3.8	-3.8	-3.81	-3.45	-3.7	1086.2	0	0	0	0
Gly-Thr-Ser-Pro	-5.15	267.1	3	-0.39	-1.52	-1.26	-1.29	0.18	-0.86	722.8	8	69	-2.53	191.33
Gly-Phe-Met-Lys	-2.77	255.06	3	-0.18	-1.14	-0.98	-0.88	0.27	-0.52	701.8	18	125	4.43	370.49
Gly-Leu-Ile-His	-3.86	300.82	3	0.17	-0.69	-0.45	-0.8	0.62	-0.08	653.7	10	64	0.12	162.48
Gly-Gly-Glu-Gln	-5.68	410.72	3	-1.51	-2.81	-2.56	-2.56	-0.77	-1.96	834.9	0	0	0	0
Gly-Cys-Asp-Asn	-5.97	847.93	3	-4.09	-4.12	-4.13	-4.14	-4.06	-4.09	2095.1	0	0	0	0
Gly-Arg-Ala-Val	-3.6	244.12	3	0.5	0.05	-0.07	-0.19	0.8	0.26	548.6	8	54	1.61	143.52
Gly-Tyr-Trp-Thr	-3.87	415.68	3	-3.72	-3.84	-3.84	-3.84	-3.6	-3.75	1136.3	17	95	2.61	293.72
Gly-Ser-Pro-Phe	-4.88	368.88	3	-2.27	-3.29	-3.31	-3.43	-1.35	-2.81	873	12	88	1.79	233.13
Gly-Ile-Gly-Glu	-3.48	414.9	3	-3.44	-3.73	-3.73	-3.73	-2.9	-3.57	986.1	0	0	0	0
Gly-Gln-Cys-Tyr	-5.37	434.19	3	-3.62	-3.78	-3.78	-3.79	-3.34	-3.68	1090.2	0	0	0	0
Gly-Asn-Arg-Gly	-5.8	474.23	3	-1.07	-2.32	-2.01	-2.17	-0.39	-1.59	785.9	18	100	-0.1	272.01
Gly-Gly-Gly-Gly-Arg	-5.16	302.32	3	0.39	0.18	-0.06	-0.13	0.61	0.28	516.5	10	50	-2.4	120.23
Gly-Gly-Gly-Gly-Tyr	-2.32	244.83	3	0.34	-0.32	-0.11	-0.18	0.44	0.02	595.6	9	47	0.68	151.75
Gly-Gly-Gly-Gly-Phe	-1.9	208.81	3	0.32	-0.06	-0.07	-0.04	0.47	0.15	540.6	7	45	0.49	137.87
Gly-Gly-Gly-Gly-Pro	-5.73	348.08	3	-0.44	-1.49	-1.23	-1.29	0.07	-0.88	728.7	11	67	-5.14	174.32
Gly-Gly-Gly-Gly-Glu	-5.2	312.51	3	0.28	-0.06	-0.09	-0.02	0.47	0.2	561.5	10	50	-2.76	124.29
Gly-Gly-Gly-Gly-Lys	-4.78	289.96	3	0.4	0.06	0.2	-0.04	0.58	0.29	517.54	10	52	-2.57	120.94
Gly-Gly-Gly-Gly-Met	-3.92	208.81	3	0.2	-0.02	-0.37	-0.18	0.5	0.27	434.5	7	40	-1.61	101.3
Gly-Gly-Gly-Arg-Gly	-5.15	302.32	3	0.14	0.16	-0.01	-0.09	0.69	0.29	516.5	10	50	-2.4	120.23
Gly-Gly-Gly-Asp-Gly	-5.56	368.72	3	0.18	-0.61	-0.48	-0.63	0.69	-0.06	677.7	13	62	-2.76	156.97
Gly-Gly-Gly-Cys-Gly	-4.83	237.91	3	0.27	0.05	0.01	-0.1	0.72	0.37	509.6	10	44	-2.63	117.37
Gly-Gly-Gly-Gln-Gly	-5.85	468.02	3	-1.26	-2.54	-2.21	-2.34	-0.34	-1.78	800.8	15	79	-2.23	191.12
Gly-Gly-Gly-Glu-Gly	-4.53	324.45	3	0.08	-0.6	-0.4	-0.24	0.54	-0.1	630.6	10	55	-3.66	140.43
Gly-Gly-Gly-His-Gly	-5.02	295.69	3	0.39	0.02	0.11	-0.36	0.54	0.31	554.5	10	48	-3.39	125.96
Gly-Gly-Gly-Ile-Gly	-5.39	424.93	3	-1.12	-2.34	-2.06	-2.22	-0.42	-1.59	785.9	14	80	-1.15	111.13
Gly-Gly-Gly-Phe-Gly	-4.34	273.33	3	0.46	0.08	-0.01	-0.05	0.76	0.28	549.6	9	51	-0.3	136.91
Gly-Gly-Gly-Thr-Gly	-5.76	377.02	3	-0.37	-1.28	-1	-0.94	0.06	-0.68	707.7	14	68	-5.22	160.24
Gly-Gly-Gly-Trp-Gly	-3.68	297.49	3	-0.62	-0.18	-1.39	-1.19	-0.17	-1.15	737.8	11	61	-0.48	188.23
Gly-Gly-Gly-Tyr-Gly	-0.08	194.82	3	0.18	-0.14	-0.05	-0.02	0.27	0.04	538.6	7	43	-0.4	159.8
Gly-Gly-Gly-Val-Gly	0.68	204.05	3	0.18	-0.14	-0.11	-0.1	0.53	0.11	550.61	7	52	1.16	140.32

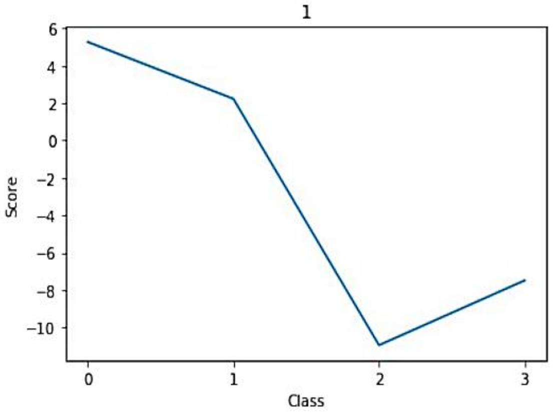
Gly-Gly-Gly-Gly-Gly-Arg	-5.16	302.32	3	0.39	0.18	-0.06	-0.13	0.61	0.28	516.52	10	50	-2.4	120.23
Gly-Gly-Gly-Gly-Gly-Cys	-3.38	275.88	3	0.16	-0.29	-0.04	-0.21	0.48	0.13	607.67	11	53	-3.15	141.03
Gly-Gly-Gly-Gly-Gly-Glu	-5.2	312.51	3	0.28	-0.06	-0.09	-0.02	0.47	0.2	561.5	10	50	-2.76	124.29
Gly-Gly-Gly-Gly-Gly-His	-4.71	353.89	3	-1.07	-2.34	-1.91	-2.4	-0.42	-1.55	780.8	12	74	-1.54	189.64
Gly-Gly-Gly-Gly-Gly-Ile	-1.97	319.41	3	-1.21	-2.6	-2.25	-2.27	-0.6	-1.7	799.9	11	75	0.43	202.72
Gly-Gly-Gly-Gly-Gly-Leu	-3.49	294.12	3	0.22	-0.28	-0.17	-0.22	0.81	0.18	571.6	10	59	-0.98	140.69
Gly-Gly-Gly-Gly-Gly-Lys	-4.78	289.96	3	0.4	0.06	0.2	-0.04	0.58	0.29	517.5	10	52	-2.57	120.94
Gly-Gly-Gly-Gly-Gly-Met	-3.92	208.81	3	0.22	0.02	-0.32	-0.09	0.55	0.28	508.62	7	47	-0.49	123.31
Gly-Gly-Gly-Gly-Gly-Phe	-1.9	208.81	3	0.32	-0.06	-0.07	-0.04	0.47	0.15	540.58	7	45	0.49	137.87
Gly-Gly-Gly-Gly-Gly-Pro	-5.73	348.08	3	-0.44	-1.49	-1.23	-1.29	0.07	-0.88	728.72	11	67	-5.95	174.32
Gly-Gly-Gly-Gly-Gly-Thr	-5.28	315.29	3	0.37	-0.25	-0.11	-0.46	0.54	0.18	598.57	11	12	-4.83	136.73
Gly-Gly-Gly-Gly-Gly-Trp	-5.3	345.76	3	-1.75	-3.19	-2.66	-3.06	-1.1	-2.26	822.87	11	76	-0.45	200.79
Gly-Gly-Gly-Gly-Gly-Tyr	-2.23	244.83	3	0.34	-0.32	-0.11	-0.18	0.44	0.02	595.61	9	47	0.68	151.75
Gly-Gly-Gly-Gly-Gly-Val	-3.22	325.2	3	-1.33	-2.63	-2.25	-2.42	-0.6	-1.86	790.88	11	75	-0.27	197.5

Graph-1: Representation of SVM classification of anti-microbial peptides as compared with standard drug Cefsulodin



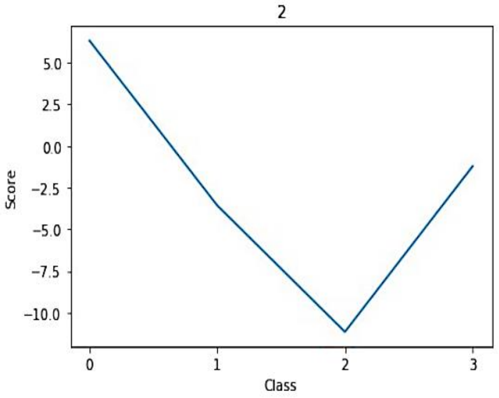
As from **graph-1**, shows the SVM algorithm result of standard drug Cefsulodin in which the similarity index above 90% with the class of zero violations (Gly-Asp to Asp-Val) for antimicrobial peptides as from **Table-2**. Through the above mention of graph, we may predict the best performance activity of class zero violations peptides for prevention of bacterial infection as well as the antimicrobial drug resistance. Although, other peptides of classes one, two, and three violation showed the least activity against the microbial infection and multi-drug resistance.

Graph-2: Representation of SVM classification of anti-microbial peptides as compared with standard drug Ethambutol



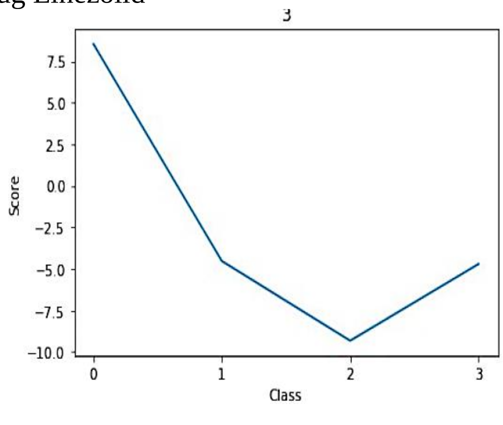
In **graph-2**, shows the SVM algorithm result of standard drug Ethambutol in which the similarity index above 90% with the class of zero violations (Gly-Asp to Asp-Val) for antimicrobial peptides as from **Table-2**. Through the above mention graph, we may predict the best performance activity of class zero violations peptides for prevention of bacterial infection as well as the antimicrobial drug resistance. Although, others peptides of class one, two, three violations shows the least activity against the microbial infection.

Graph-3: Represents the graphically, SVM classification of anti-microbial peptides as compared with standard drug Moxifloxacin



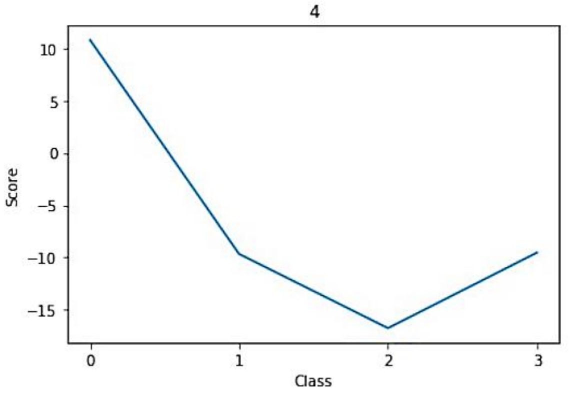
As per the **graph-3**, shows the SVM algorithm result of standard drug Moxifloxacin in which the similarity index above 90% with the class of zero violations (Gly-Asp to Asp-Val) for antimicrobial peptides as from **Table-2**. Through the above mention graph we may predict the best performance activity of class zero violations peptides for prevention of bacterial infection as well as the antimicrobial drug resistance. Although, others peptides of class one, two, three violations shows the least activity against the microbial infection.

Graph-4: Representation of SVM classification of anti-microbial peptides as compared with standard drug Linezolid



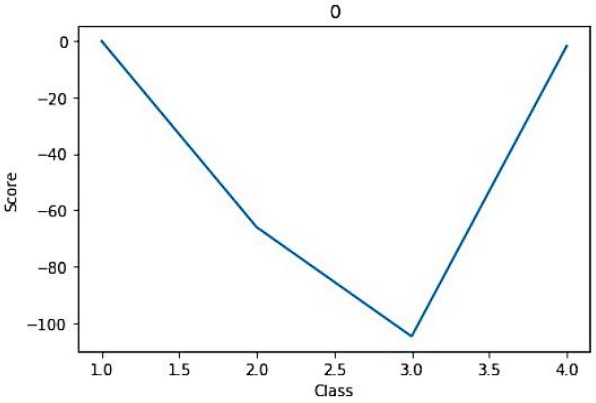
In **graph-4**, shows the SVM algorithm result of standard drug Linezolid in which the similarity index above 90% with the class of zero violations (Gly-Asp to Asp-Val) for antimicrobial peptides as from **Table-2**. Through the above mention graph, we may predict the best performance activity of class zero violations peptides for prevention of bacterial infection as well as the antimicrobial drug resistance. Although, others peptides of class one, two, three violations shows the least activity against the microbial infection.

Graph-5: Represenation of SVM classification of anti-microbial peptides as compared with standard drug Carbapenem



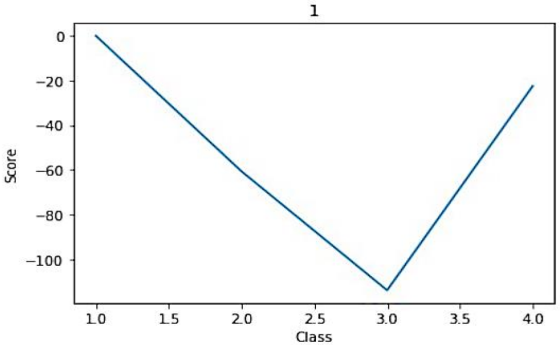
In the **graph-5**, shows the SVM algorithm result of standard drug Carbapenem in which the similarity index above 90% with the class of zero violations (Gly-Asp to Asp-Val) for antimicrobial peptides as from **Table-2**. Through the above mention graph we may predict the best performance activity of class zero violations peptides for prevention of bacterial infection as well as the antimicrobial drug resistance. Although, others peptides of class one, two, three violations shows the least activity against the microbial infection.

Graph-6: Representation of Gaussian classification of anti-microbial peptides as compared with standard drug Cefsulodin



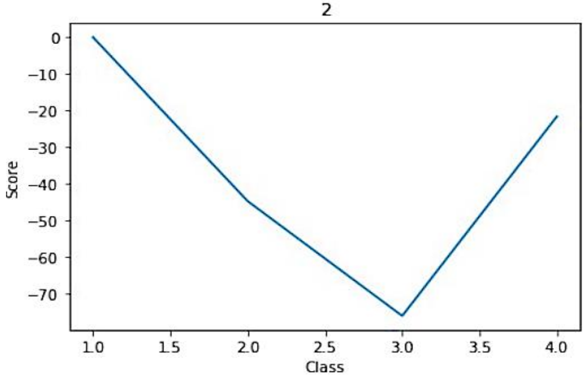
The **graph-6**, shows the Gaussian classifier result of standard drug Cefsulodin in which the similarity index above 90% with the class of zero violations (Gly-Asp to Asp-Val) for antimicrobial peptides as from **Table-2**. Whereas, the class of three violations (Gly-Ala-Arg-Asn to Gly-Gly- Gly-Gly-Val) as from **Table-5** also showed the similarity score above 90% as compared with Cefsulodin. Through the above mention graph we may predict the best performance activity of class zero violations peptides for prevention of bacterial infection as well as the antimicrobial drug resistance. Although, others peptides of class one and two violations shows the least activity against the microbial infection.

Graph-7: Representation of Gaussian classification of anti-microbial peptides as compared with standard drug Ethambutol



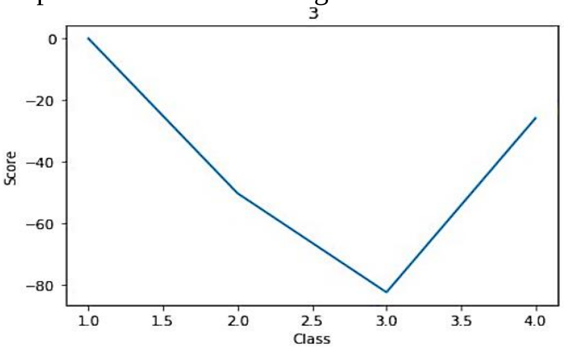
In the **graph-7**, shows the Gaussian classifier result of standard drug Ethambutol in which the similarity index above 90% with the class of zero violations (Gly-Asp to Asp-Val) for antimicrobial peptides as from **Table-2**. Whereas, the class of three violations (Gly-Ala-Arg-Asn to Gly-Gly-Gly-Gly-Val) as from **Table-5** also showed the similarity score above 80% as compared with Ethambutol. Through the above mention graph we may predict the best performance activity of class zero violations peptides for prevention of bacterial infection as well as the antimicrobial drug resistance. Although, others peptides of class one and two violations shows the least activity against the microbial infection.

Graph-8: Representation of Gaussian classification of anti-microbial peptides as compared with standard drug Moxifloxacin



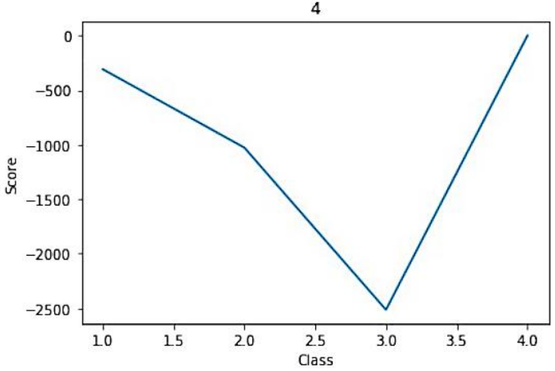
In the **graph-8**, shows the Gaussian classifier result of standard drug Moxifloxacin in which the similarity index above 90% with the class of zero violations (Gly-Asp to Asp-Val) for antimicrobial peptides as from **Table-2**. Through the above mention graph we may predict the best performance activity of class zero violations peptides for prevention of bacterial infection as well as the antimicrobial drug resistance. Although, others peptides of class one, two and three violations shows the least activity against the microbial infection.

Graph-9: Representation of Gaussian classification of anti-microbial peptides as compared with standard drug Linezolid



The **graph-9**, shows the Gaussian classifier result of standard drug Linezolid in which the similarity index above 90% with the class of zero violations (Gly-Asp to Asp-Val) for antimicrobial peptides as from **Table-2**. Through the above mention graph we may predict the best performance activity of class zero violations peptides for prevention of bacterial infection as well as the antimicrobial drug resistance. Although, others peptides of class one, two and three violations shows the least activity against the microbial infection.

Graph-10: Representation of Gaussian classification of anti-microbial peptides as compared with standard drug Carbapenem



In the **graph-10**, shows the Gaussian classifier result of standard drug Carbapenem in which the similarity index above 90% with the class of three violations (Gly-Ala-Arg-Asn to Gly-Gly- Gly-Gly-Val) for antimicrobial peptides. As from **Table-5** Through the above mention graph we may predict the best performance activity of class zero violations peptides for prevention of bacterial infection as well as the antimicrobial drug resistance. Although, others peptides of class zero, one and two violations shows the least activity against the microbial infection.

In the both SVM and Gaussian classifier, we found that counteracting of class 0 drugs (33 trained data sets of peptides) were found to be the most promising lead for treatment of infection caused by bacterial, fungal including mycobacterial as well as for multi-drug resistance problem, which was clearly evidenced with all graphical representation the composition of all classes were depicted in both graphs.

Conclusion

Support vector machine and Gaussian classifier are the informatics technology that helps in the determination of antimicrobial peptides (AMPs). The SVM-based method with the linear kernel function was found to be the best model for the classification of AMP$_S$. Prediction accuracy varied in all classes of peptides. The best performance was obtained from the class of zero violations peptides as trained data set, which indicated the best prediction performance in the classifier compared with the standard drugs and all other classes of peptides. The best potent dipeptide lead obtained from the artificial intelligence (AI) approach through the SVM classifier (Gly-Asp to Asp-Val) of 33 peptides belongs to the Class-0 violations which of trained data sets. The same results have been coinciding with by the using of Gaussian classifier machine learning tool of (Gly-Asp to Asp-Val) belongs to the Class-0 violations which of trained data sets. The results are completely different for both classifiers with regards to standard Carbapenem through AI approach with class-3 peptide drugs as second most potent drugs as from Table-2. This developed model was used for the identification of the most promising lead as a peptide. Meanwhile, computational prediction is a very complementary method but does not replace laboratory experiments. In this study, identified peptides of the class-0 violations peptides category were used to confer and minimized bacterial infection or prevention antibacterial drug resistance as per the SVM classifier and Gaussian classifier, which will be an eye opener for any researcher to proceed the right treatment practical investigation for the tuberculosis and broad spectrum microbial infections in near future.

Acknowledgement

Special thanks to Mr. Siddheshwari Dutt Mishra of MCA department in KIET Group of Institutions. Also, the authors acknowledge and are grateful to Dr. (Col.) A. Garg, Director, and Dr. Manoj Goel who is the Joint Director of KIET Group of Institutions in overall support for carrying out the Work.

Conflict of Interest

Authors, declare they have not any competing interest.

References

1. G. Loosli, S. Canu and C. S. Ong, IEEE Trans. Pattern Anal. Mach. Intell., 2016, 38, 1204–1216.
2. D. Oglic and T. Gärtner, Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 3859–3867.
3. D. Oglic and T. Gärtner, International Conference on Machine Learning, 2019, pp. 4912–4921.
4. N. Mookherjee, M. A. Anderson, H. P. Haagsman and D. J. Davidson, Nat. Rev. Drug Discov., 2020, 19, 311–332.
5. R. E. Hancock, E. F. Haney and E. E. Gill, Nat. Rev. Immunol., 2016, 16, 321–334.
6. D. S. J. Ting, I. Mohammed, L. Rajamani, R. W. Beuerman and H. S. Dua, Front. Med., 1667.
7. D. S. J. Ting, E. T. L. Goh, V. Mayandi, J. M. Busoy, T. T. Aung, M. H. Periayah, M. Nubile, L. Mastropasqua, D. G. Said, H. M. Htoon et al., Sci. Rep., 2021, 11, 1–14.
8. V. Mayandi, Q. Xi, E. T. Leng Goh, S. K. Koh, T. Y. Jie Toh, V. A. Barathi, M. H. Urf Turabe Fazil, M. L. Somaraju Chalasani, J. Varadarajan, D. S. J. Ting et al., J. Med. Chem., 2020, 63, 3522–3537.
9. C. J. Murray, K. S. Ikuta, F. Sharara, L. Swetschinski, G. R. Aguilar, A. Gray, C. Han, C. Bisignano, P. Rao, E. Wool et al., Lancet, 2022.
10. W. Ali, A. Elsahn, D. S. Ting, H. S. Dua and I. Mohammed, Antibiotics, 2022, 11, 475.
11. C. D. Fjell, J. A. Hiss, R. E. Hancock and G. Schneider, Nat. Rev. Drug Discov., 2012, 11, 37–51.
12. D. S. J. Ting, R. W. Beuerman, H. S. Dua, R. Lakshminarayanan and I. Mohammed, Front. Immunol., 2020, 11, 983.
13. P. Das, T. Sercu, K. Wadhawan, I. Padhi, S. Gehrmann, F. Cipcigan, V. Chenthamarakshan, H. Strobelt, C. Dos Santos, P.-Y. Chen et al., Nat. Biomed. Eng., 2021, 5, 613–623.
14. N. Y. Yount, D. C. Weaver, E. Y. Lee, M. W. Lee, H. Wang, L. C. Chan, G. C. Wong and M. R. Yeaman, Proc. Natl. Acad. Sci. U.S.A., 2019, 116, 6944–6953.
15. D. S. J. Ting, J. Li, C. S. Verma, E. T. Goh, M. Nubile, L. Mastropasqua, D. G. Said, R. W. Beuerman, R. Lakshminarayanan, I. Mohammed et al., Front. Pharmacol., 2021, 2793.

A. Capecchi, X. Cai, H. Personne, T. Köhler, C. van Delden and J.-L. Reymond, Chem. Sci., 2021, 12, 9221–9232.

B. Li, D. Sutherland, S. A. Hammond, C. Yang, F. Taho, L. Bergman, S. Houston, R. L. Warren, T. Wong, L. Hoang et al., BMC Genom., 2022, 23, 1–15.

16. P. G. Aronica, L. M. Reid, N. Desai, J. Li, S. J. Fox, S. Yadahalli, J. W. Essex and C. S. Verma, J. Chem. Inf. Model., 2021, 61, 3172–3196.

17. S. A. Pinacho-Castellanos, C. R. García-Jacas, M. K. Gilson and C. A. Brizuela, J. Chem. Inf. Model., 2021, 61, 3141–3157.

18. S. Thomas, S. Karnik, R. S. Barai, V. K. Jayaraman and S. Idicula-Thomas, Nucleic Acids Res., 2010, 38, D774–D780.

19. S. Lata, N. K. Mishra and G. P. Raghava, BMC Bioinform., 2010, 11, 1–7.

20. M. Torrent, D. Andreu, V. M. Nogués and E. Boix, PloS One, 2011, 6, e16968.

21. E. Y. Lee, B. M. Fulan, G. C. Wong and A. L. Ferguson, Proc. Natl. Acad. Sci. U.S.A., 2016, 113, 13588–13593.

22. P. K. Meher, T. K. Sahu, V. Saini and A. R. Rao, Sci. Rep., 2017, 7, 1–12.

23. J. Yan, P. Bhadra, A. Li, P. Sethiya, L. Qin, H. K. Tai, K. H. Wong and S. W. Siu, Mol. Ther. Nucleic Acids, 2020, 20, 882– 894.

24. D. Veltri, U. Kamath and A. Shehu, Bioinformatics, 2018, 34, 2740–2747.

25. P.-M. Feng, W. Chen, H. Lin and K.-C. Chou, Anal. Biochem., 2013, 442, 118–125.

26. K.-C. Chou, Proteins: Struct. Funct. Genet., 2001, 43, 246– 255.

27. G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee and M. H. Zehfus, Science, 1985, 229, 834–838.

28. T. F. Smith and M. S. Waterman, J. Mol. Biol., 1981, 147, 195–197.

29. S. B. Needleman and C. D. Wunsch, J. Mol. Biol., 1970, 48, 443–453.

A. Zielezinski, H. Z. Girgis, G. Bernard, C.-A. Leimeister, K. Tang, T. Dencker, A. K. Lau, S. Röhling, J. J. Choi, M. S. Waterman et al., Genome Biol., 2019, 20, 1–18.

30. M. R. Kantorovitz, G. E. Robinson and S. Sinha, Bioinformatics, 2007, 23, i249–i255.

A. Zielezinski, S. Vinga, J. Almeida and W. M. Karlowski, Genome Biol., 2017, 18, 1–17.

31. P. Kuksa and V. Pavlovic, BMC Bioinform., 2009, 10, 1–18.

32. P. Wang, L. Hu, G. Liu, N. Jiang, X. Chen, J. Xu, W. Zheng, L. Li, M. Tan, Z. Chen et al., PloS One, 2011, 6, e18476. 1–10 | 9

33. X. Y. Ng, B. A. Rosdi and S. Shahrudin, BioMed Res. Int., 2015, 212715.

34. S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, J. Mol. Biol., 1990, 215, 403–410.

35. X. Gao, B. Xiao, D. Tao and X. Li, Pattern Anal. Appl., 2010, 13, 113–129.

36. S. Axelsson, Digit. Investig., 2010, 7, S24–S31.

37. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* (2017) 7:42717.