computers

MDPI

Article

Neural Network Methodology for the Identification and Classification of Lipopeptides Based on SMILES Annotation

Manisha Yadav and Satya Eswari Jujjavarapu *

Department of Biotechnology, National Institute of Technology Raipur, Chhattisgarh 492010, India; manishayadav.nitrr@gmail.com

* Correspondence: satyaeswarij.bt@nitrr.ac.in or eswari_iit@yahoo.co.in

Abstract: Artificial Neural Networks can be applied for the identification and classification of prospective drug candidates such as complex compounds, including lipopeptide, based on their SMILES string representation. The training of neural networks is done with SMILES strings, which are predictive of structural identification; the ANNs are efficient of correctly classifying all compounds, substructures and their analogues distinguishing the drugs based upon atomic organization to obtain lead optimization in drug discovery. The proficiency of the trained ANN models in recognizing and classifying the analogous compounds was tested for analysis of similar compounds, which were not taken previously for training and achieved results with correct classification in the validation set. The best result was achieved with 10 numbers of hidden layers. The R2 value for training is 0.90586; the R2 value for testing is 0.99508; the R2 value after validation is 0.94151; the final value of R2 for total sets is 0.89456. The graphs are plotted between 21 epochs and mean square error (MSE) to report the performance of the model. The value of 798.1735 for the gradient of the curve after 21 iterations and 6 validation checks was obtained. A successful model was developed for the identification and classification of lipopeptides from their SMILES annotation that efficiently classifies similar compounds and supports in decision making for analogue-based drug discovery. This will help in appropriate lead optimization studies for the prediction of potential anticancer and antimicrobial lipopeptide-based therapeutics.

Keywords: neural network; drug discovery; SMILE strings; classification; lipopeptides; surfactin analogs

Citation: Yadav, M.; Jujjavarapu, S.E. Neural Network Methodology for the Identification and Classification of Lipopeptides Based on SMILES Annotation. *Computers* **2021**, *10*, 74. https://doi.org/10.3390/computers10060074

Academic Editors: Antonio Celesti, Ivanoe De Falco, Antonino Galletta and Giovanna Sannino

Received: 31 March 2021
Accepted: 3 June 2021
Published: 10 June 2021

1. Introduction

Currently, along with in vitro and in vivo, in silico analysis such as machine learning for the prediction of chemical properties of compounds has become an efficient way in chemical analysis. One such example is the prediction of protein-ligand interaction, which facilitates the identification of novel compounds through screening of lead compounds in the process of drug discovery [1].

For computational analysis, various file formats are used to define the chemical compounds digitally, which facilitates the reading of compounds through computers. Some of the file formats such as SDF (structure data file), MOL (molfile) (.sdf and .mol are extensions developed by MDL Molecular design limited for saving chemical files in the computer), SMILES (simplified molecular line-entry system) and fingerprints are used widely used in computer-aided drug discovery [2]. MOL format is used to represent a compound in graph connection table form, wherein an atom is represented by each node and the bonds are in the form of edges between atoms. SDF is used to write more than one compound in a single file and is an extension of MOL format, or it can be said its extended version [1].

SMILES, abbreviated as a simplified molecular line-entry system and proposed by Weininger, is broadly perceived and utilized as a standard interpretation system of compounds currently for processing chemical information [3]. A linear notation method is provided by SMILES for the representation of chemical compounds in a distinct way,

which is in the string form over fixed alphabets. SMILES utilize explicit characters and a grammatical way for the description of each atom and the whole structure of a chemical compound. Structural differences can also be distinguished by SMILES, which includes the chirality of compounds. The representation of SMILES in a linear structure is alluded to as SMILES string, which facilitates the straightforward application of artificial neural network for (ANN) for the identification and classification of chemical compounds. The ANN techniques with SMILES annotation are utilized further for virtual screening in lead identification and elucidation of functional substructures (chemical motifs) in modern age drug discovery [1,3].

Analysis of chemical compounds using the artificial neural network is researched and explored actively. TOX 21 Challenge 2014 and Molecular Activity Challenge 2013 motivate such studies, wherein the results achieved from machine learning techniques are found superior to those obtained by utilizing other methods [3–5]. Though, the utilization of deep learning methods to their full capability is yet to be explored.

The artificial neural network is basically inspired by the architecture of the human brain and is an artificial model analogous to the nervous system. As the human brain is well organized and has tremendous decision-making capability because of a complex network of neurons in the human brain. Similar to the human brain, the artificial neural network also learns through the examples on which it has already worked or trained. After sufficient learning or training, the neural network is capable of working on examples by themselves [4]. The machine learning technique for the current work is based on the artificial neural network.

The observations of the above approaches led us to propose a new model by using SMILES annotation, which is a linear representation of a chemical compound to apply ANNs for the classifications and identification of chemical compounds and chemical motifs. A string is a one-dimensional structure (1-D grid) of atoms and symbols similar to molecular sequences such as protein and DNA sequence represented in strings. Neural network techniques have already been applied for the classification of DNA and RNA sequences and for the extraction of conserved motif sequences [4–8]. In such methods, the representation of the one-hot coding region of four nucleotides of DNA, a kernel (filter) with convolution operation of one-dimensional is applied on a sequence for the consideration of position weight matrix to represent a motif. The learning of filters is done by training the artificial neural network on sequences of negative and positive samples, which are obtained through experimental studies of chromatin immunoprecipitation with ChIP-seq high-throughput sequencing [9–12].

A similar approach is utilized for the classification studies for the lipopeptide-based compounds. Lipopeptides are complex peptide-based compounds of natural origin, which are of interest in various therapeutic applications and diversified biological activity ranging from broad-spectrum antimicrobial, antiviral, antifungal, anticancer and anti-obesity etc. Lipopeptides are natural biomolecules, which are majorly produced by *Bacillus* species. These compounds are amphiphilic in nature and composed of six to ten amino acids comprising of hydrophilic cyclic peptide moiety attached to a C13–15 fatty acid side chain of hydrophobic nature (Figure 1) [13,14]. The classification model is an obvious study for such complex compounds to facilitate the preparation of a combinatorial library for the screening of novel lead compounds. In the current work, interpretation of a neural network operation of "one-dimension" for sequences (canonical SMILES annotation of lipopeptides) is used for scanning of the input sequence in one direction only along the sequence having filter of the same dimension (width) as of the representation of the input. Now, a straightforward approach of applying an artificial neural network to the SMILES strings, which are representing chemical compounds for further identification and classification of the library of lipopeptide-based chemical compounds and the extraction of the conserved structures (chemical motifs) among the compounds [1,3].

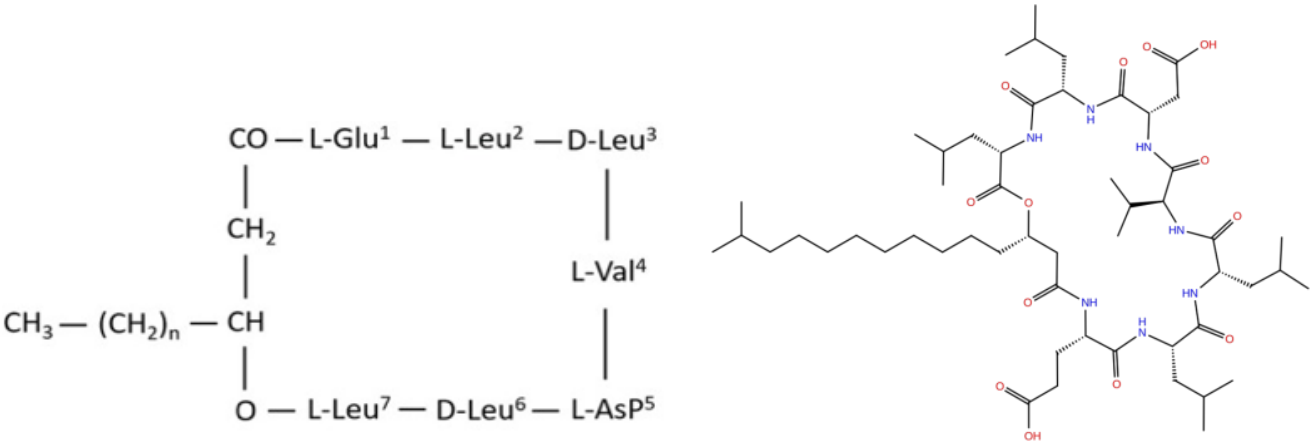

Figure 1. The primary structure of lipopeptide Surfactin; n = 9–11.

2. Materials and Method

2.1. Data Collection

We first tried to obtain the SMILES strings of the lipopeptide compounds. A broad-spectrum antimicrobial and potential anticancer lipopeptide surfactin were retrieved for the preparation of the compound library, which are found as substructures, analogs, isoforms and similar compounds of surfactin lipopeptide using virtual screening [2]. The library preparation for surfactin analogs is done for the study of structural modification in drug compounds and identify the compounds with greater potency for further anticancer studies [15,16]. The canonical smiles of the peptidolipidic compounds are obtained from the PubChem compound database. A total of 22 compounds were obtained, and the SMILES strings of these compounds were retrieved from National Centre for Biotechnology Information (NCBI) database PubChem compound (https://pubchem.ncbi.nlm.nih.gov) (accessed on 20 May 2021) [17]. Ten of the lipopeptides are substructures of surfactin and have a modification in the varied carbon chain length. The rest of the 12 compounds are analogous structures, isoforms and similar compounds.

2.2. Data Preparation

Data encoding has an essential part to improve the performance of the network. ANNs have the capability to process presented data of diverse forms to the network in an appropriate format. Inputs of various classes are distinguished properly by a neural network. Methodology of data encoding is used for data presentation. Lipopeptide-based compounds are biomolecules that are produced as secondary metabolites by various strains of *Bacillus* genus [1,12]. These are composed of a cyclic hexa to deca peptide core attached to a fatty acid side chain. Surfactin is one of the prominent compounds produced by the *Bacillus subtilis* [15,18]. Out of the total 20 amino acids, there are some specific amino acids, which are found predominantly in lipopeptides. A two-dimensional structure of surfactin lipopeptide is depicted in Figure 1 [14,19,20]. A library of similar compounds is prepared through a virtual screening-based method. SMILES string format of each peptidolipidic compound is used to prepare the input file. The annotation of canonical SMILES is represented for Lipopeptide Surfactin A below in Table 1.

Table 1. Lipopeptide Surfactin A and its canonical SMILES annotation.

Compound Name	Canonical SMILES
Surfactin A	CC(C)CCCCCCCCCC1CC(=O)NC(C(=O)NC(C(=O)NC(C(=O)NC(C(=O)NC(C(=O)O1)CC(C)C)CC(C)C)CC(=O)O)C(C)C)CC(C)C)CC(C)C)CCC(=O)O

SMILES strings are basically a way of representation through an elementary arrangement of composed atoms such as C, O, N, S, Ca and Na with certain bonds, which are used to establish the compound's structure. SMILES, abbreviated for Simplified Molecular Line-Entry System, is a specific form of line annotation to describe the structure of a compound through short ASCII strings. SMILES strings are basically imported by molecular editors, which can be back converted to their 2-D drawing format or in 3-D models of the molecule. Usually, various valid SMILES strings can be used to represent a molecule in a 1-D format, such as canonical SMILES and isomeric SMILES. Certain algorithms have been established for generating such SMILES representation of one-dimensional strings. Similar to DNA and protein sequences, SMILES are also unique for each structure. A canonicalized algorithm is used for the generation of the string is called canonical SMILES [2]. The surfactin substructures and similar compounds are retrieved from the PubChem compound database. The canonical smile strings are utilized to conduct classification studies using an artificial neural network. The symbols of SMILES strings are converted into numerical values to generate a tabulated file [1,12,20]. The interpretation of symbols is represented in Table 2.

Table 2. Data Interpretation: The replacement of atoms and symbol of SMILES annotation in numerical values.

Symbol	Replaced with
1	1
C	2
O	3
N	4
Na	5
Ca	6
S	7
+	8
=	9
(	10
)	11
.	12
-	13
[	14
]	15
Empty space	0

2.3. Experimental Protocol

We retrieved all the required strings from the PubChem compound database for data collection. The SMILES strings are arranged according to the requirement based on varied fatty acids carbon chain length, side chain modification, specific additional functional group modification and other similar compounds. The compounds, which were analyzed in the present work, are listed below in Table 3. The SMILES strings are converted into numerical values and tabulated to generate an input data file. The sequences are completed up to 153 places in the columns of the input sheet. Two files are prepared, one sheet for input data (ipp) and another sheet for output (opp). Further, MATLAB software is used to run the program for maximum output.

Table 3. Lipopeptide-based compounds and numerical values used for the preparation of the input file.

Compound Name	Numerical Value
Surfactin A	10
Surfactin B	20
Surfactin C	30
Surfactin D	40
Surfactin-C13	50

Table 3. *Cont.*

Compound Name	Numerical Value
Surfactin-C14	60
Surfactin-C15	70
Surfactin Peptide	80
Methylated Surfactin	90
Amidated Surfactin	100
Linear Surfactin	110
Sodium Surfactin	120
Disodium Surfactin	130
Calcium Surfactin	140
Aminomethane Sulfonated Surfactin	150
[Leu-7]Surfactin	160
[Ile4]Surfactin	170
[Ile7]Surfactin	180
[Ile4,7]Surfactin	190
[Ile2,4,7]Surfactin	200
3(R)-Aza-Surfactin	220
3(S)-Epi-Aza-Surfactin	220

2.4. Neural Network Methodology

Various methodologies are used to analyze chemical compounds from the perspective of drug discovery. A few of such methods are Artificial Neural Network preliminaries, Gradient Descent Algorithm (GDA), Multi-layered Feed Forward ANN (MLFANN), Leven–Marquardt algorithm (LM), Conjugate Gradient Descent Algorithm etc. Here, the LM algorithm and Scaled Conjugate Gradient (SCG) algorithm are used for the performed work of the article. It is a technique used for solving problems of nonlinear least squares. In the case of nonlinear functions in parameters, it is aroused as nonlinear quadrangles glitches. To decrease the errors of quadrangle sum between sedate data joints and parameters, the nonlinear least square method incorporates iterative progress to the parameter value. Actually, the LM curve-fitting is a combined method of two minimization techniques that is the Gauss–Newton method and grade lineage technique.

The summation of squared errors gets reduced in the incline descent method through updating the parameters in the direction of steepest-descent. The Gauss–Newton method includes the summation of squared errors, which is abridged through an assumption of a locally quadratic function of minimum squares and ultimately finds out the minimum of quadratic. The functionality of the LM method is quite comparable with the gradient-descent method, wherein the parameters are actually distant from the optimal value. It also acts similar to the Gauss–Newton method if structures are nearer to the optimal value [9,12]. In current work, we have taken input file as (ipp), which will be trained, and an output file is received as opp after training. Hidden layers interplay an indispensable role during the process of training. The selection of the number of hidden nodes is a very prominent factor on which the final output is dependent. Meanwhile, training, testing and validation processes take place. The two networks are run parallelly. The structure of the neural network is shown in Figure 2.

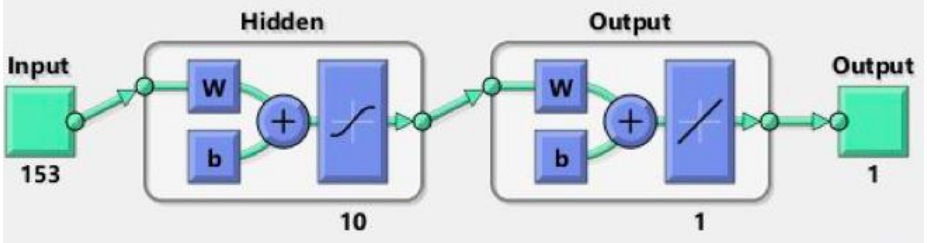

Figure 2. Neural network architecture: input and output with hidden nodes.

3. Results and Discussion

The artificial neural networks are applied due to their efficiency in tackling a huge amount of data with the ease and good convergence rate through which it can be trained, and as a result, discreet modeling of big datasets. This helps in the prediction of identification and classification of lead molecule out of the library of a similar category of compounds, which ultimately supports novel drug discovery. The principal purpose of the current work is to develop a neural network model for accurate identification and classification of lipopeptide-based candidate drugs, which are of various medicinal properties such as anticancer, antiviral, antibacterial and antifungal and deduce a relationship between them [18–20].

Neural Network Training Results

The parameters are set in software MATLAB R2016a for conducting an artificial neural network. The canonical SMILES strings of 22 lipopeptides are used to develop neural network model. Out of the total 22, 16 compounds are used for training the network, 3 are used for validation and 3 are used for testing. The performance measures of the network such as R2 and MSE are shown in the figures below. The ANN is performed with 70%, 15% and 15% of the total 22 lipopeptides being used for training, validation and testing, respectively. Similarly, the training, testing and validation are performed using all various possible combinations. Different numbers of the feasible mixture and hidden layers are used to generate the architecture of the network with which the least error is provided. The network is trained with various kinds of hidden nodes with validation. The sets for testing were changed accordingly every time. The output as the best network was obtained after rigorous training, which is mentioned in the figures below. It is depicted from these figures that the optimum performance was obtained with 10 hidden nodes with a 15% testing set and 15% validation set. Out of the total 22 sets, 16 were allocated for training, 3 for testing and 3 for validation. The results for the Levenberg–Marquardt algorithm (LM) are depicting the best validation performance (Figure 3), error histogram (Figure 4), regression plot (Figure 5) and training state (Figure 6). The R2 value is 0.90586 with training, the R2 value is 0.99508 with testing, the R2 value after validation is 0.94151, and the R2 value obtained with the total number of sets is 0.89456 (Figure 5). The results have shown that the experimental results are found closer to the predicted neural network results. Figure 3 is depicting the best validation performance. The figures are drawn against MSE vs. epochs. In total, nine epochs were taken for the modeling. From Figure 4, the error histogram can be seen. In each plot, the dashed line denotes the perfect result-outputs = targets, wherein the solid line indicates the line of best fit linear regression between targets and outputs. The R2 value signifies the relationship between the targets and outputs. The greater value of R2 (close to 1) denotes the greater accuracy in the linear relationship between targets and outputs. The value of MSE and R2 for training, testing, validation and overall data is shown in Figures 3 and 5, respectively. The validation for predicted and actual output is depicted in Figure 6 for the training of the neural network. Similarly, the results for the Scaled Conjugate Gradient algorithm SCGA) are showing the best validation performance (Figure 7), error histogram (Figure 8), regression plot (Figure 9), and training state (Figure 10). In Figure 10, the training state can be seen for the SCG algorithm, wherein decreasing error with each iteration is observed.

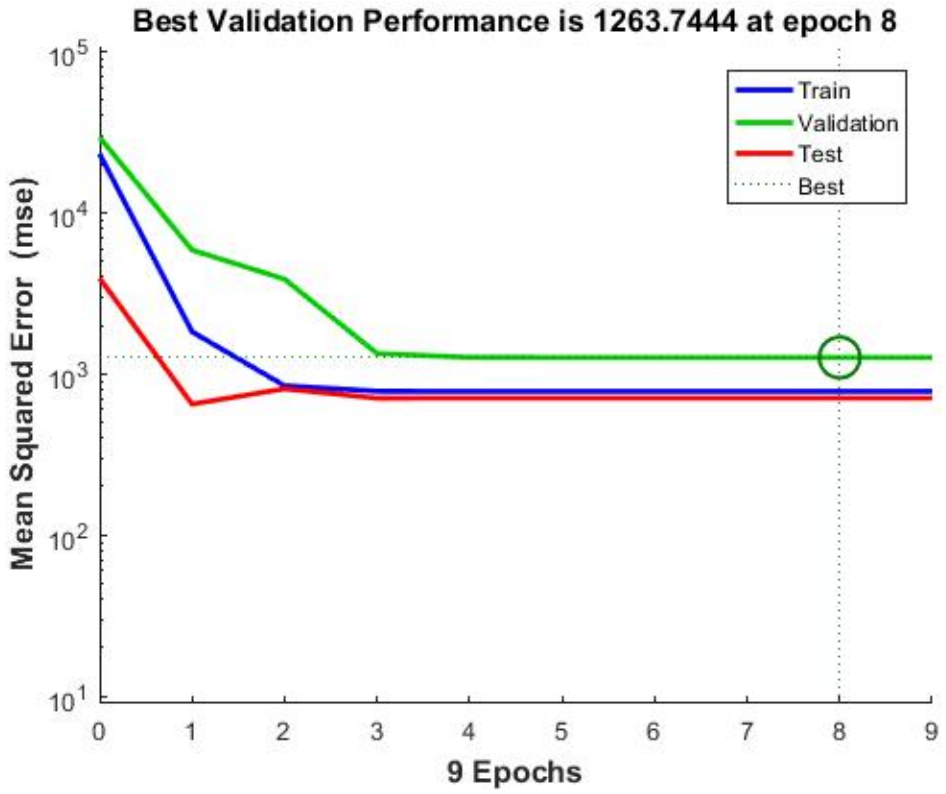

Figure 3. Performance: The best validation performance at epoch 8 (LMA).

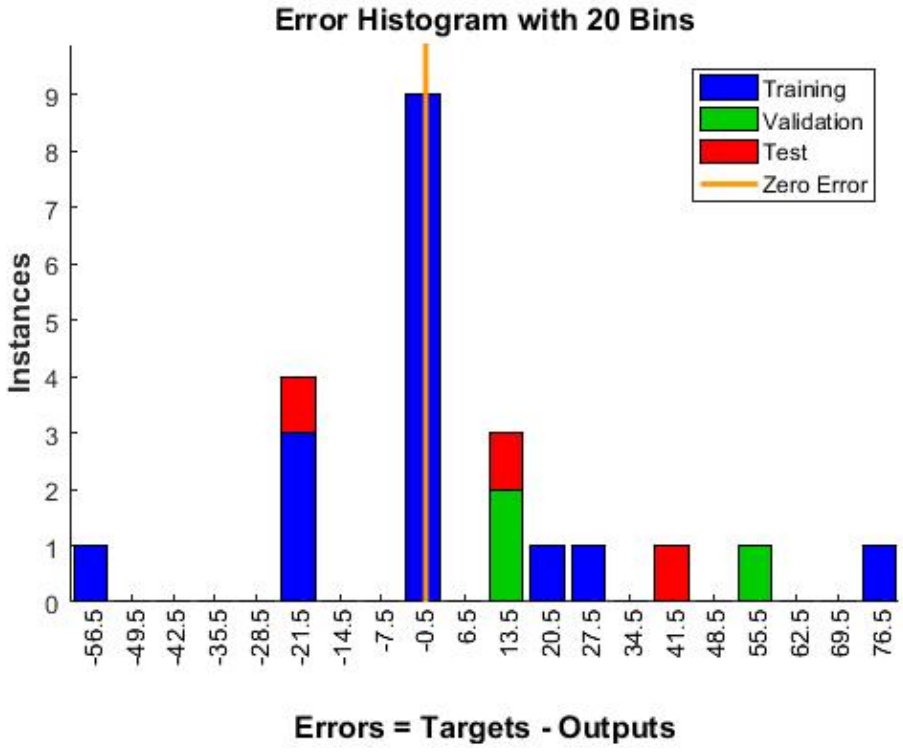

Figure 4. Histogram: Error Histogram (LM Algorithm).

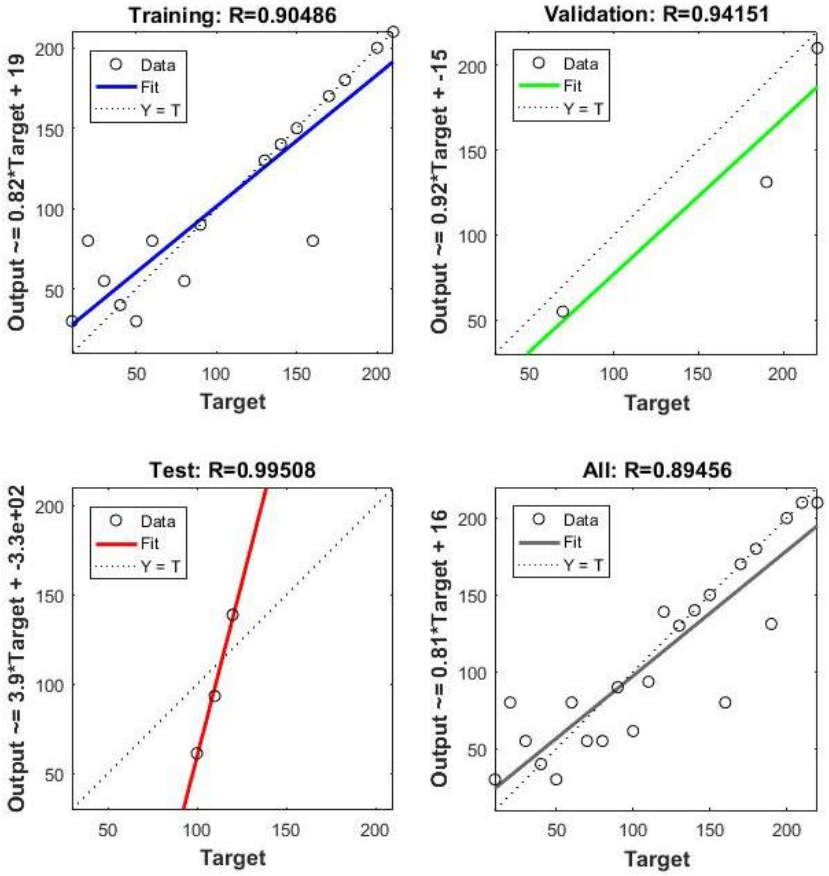

Figure 5. Regression Plot: training, testing and validation (LM Algorithm).

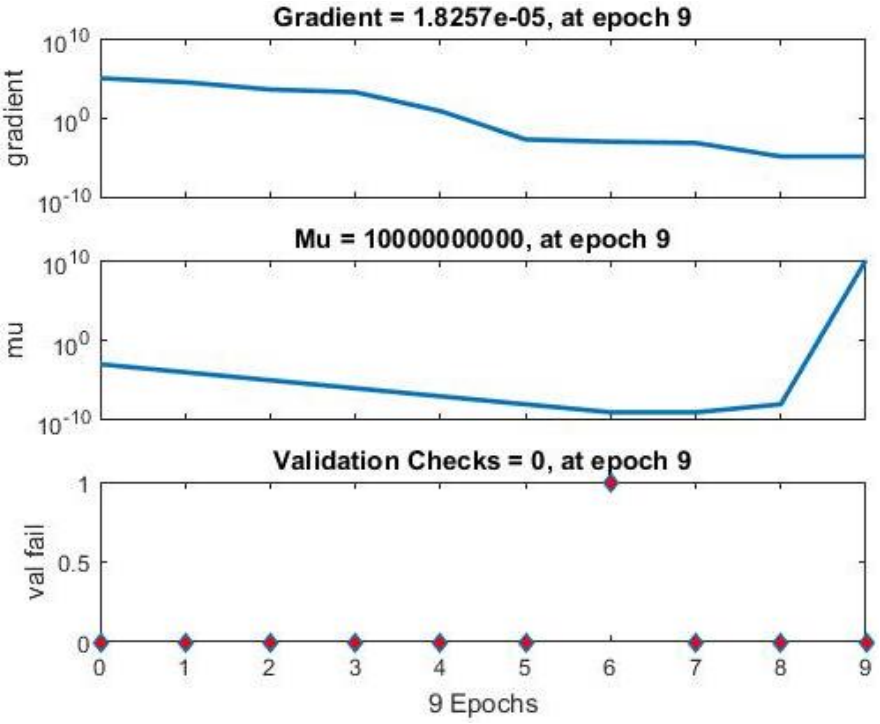

Figure 6. Validation (LMA): gradient at epoch 9.

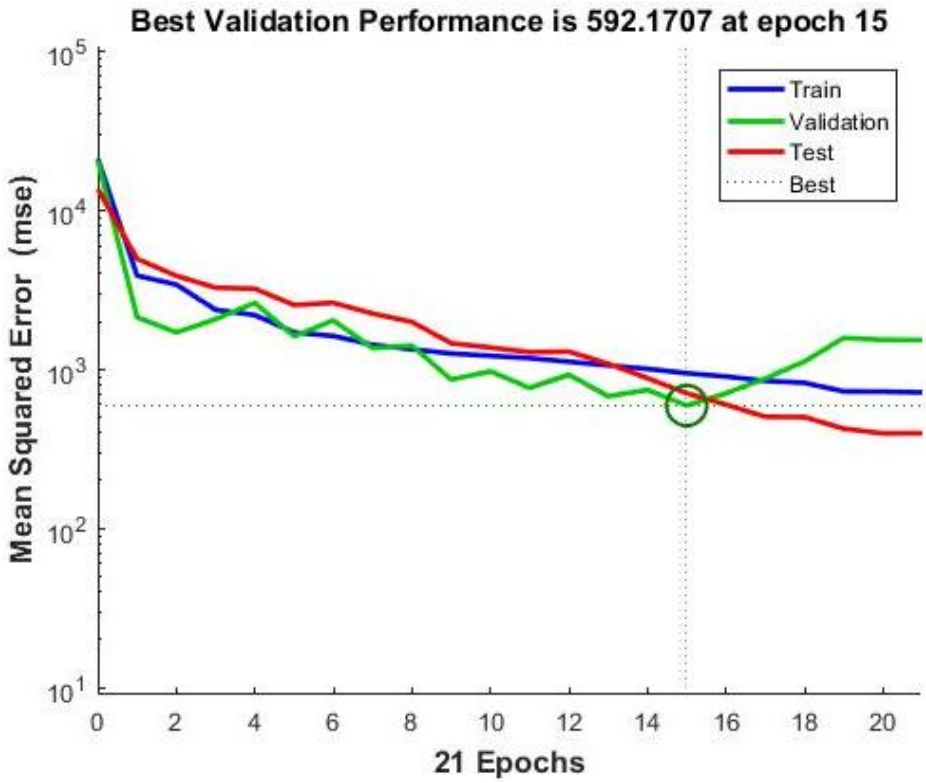

Figure 7. Performance Model for the SCG Algorithm: the best validation performance at epoch 15.

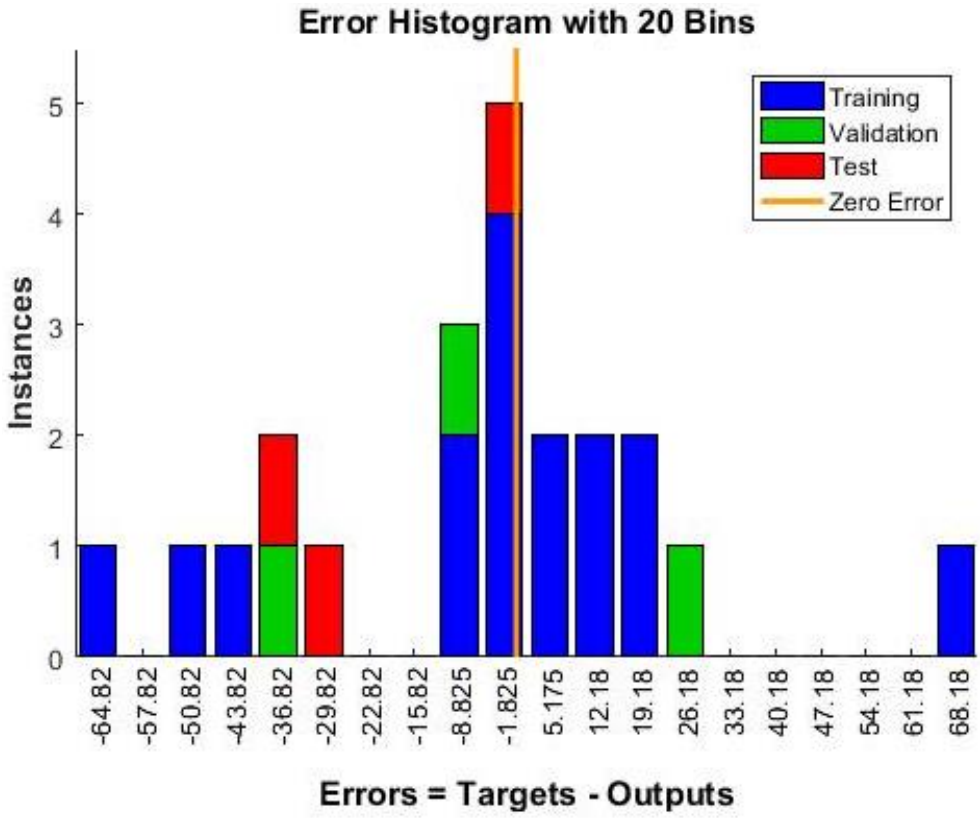

Figure 8. Histogram (SCGA): the error histogram with 20 bins.

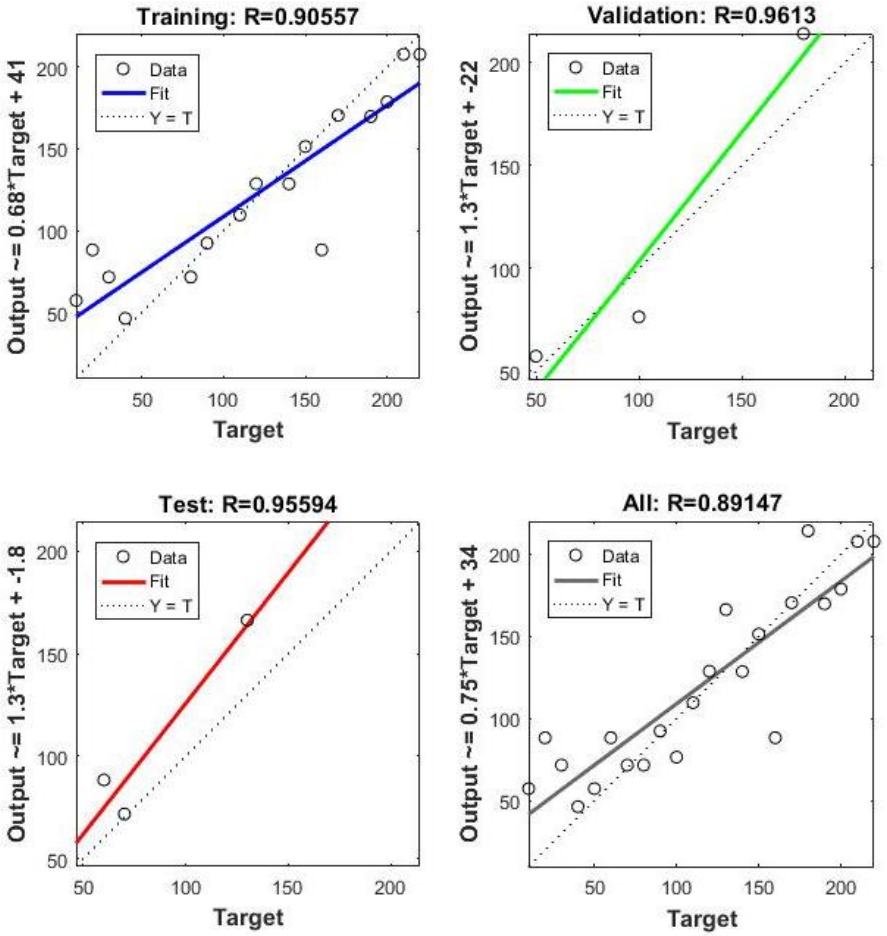

Figure 9. Regression Plot (SCGA): training, testing and validation model.

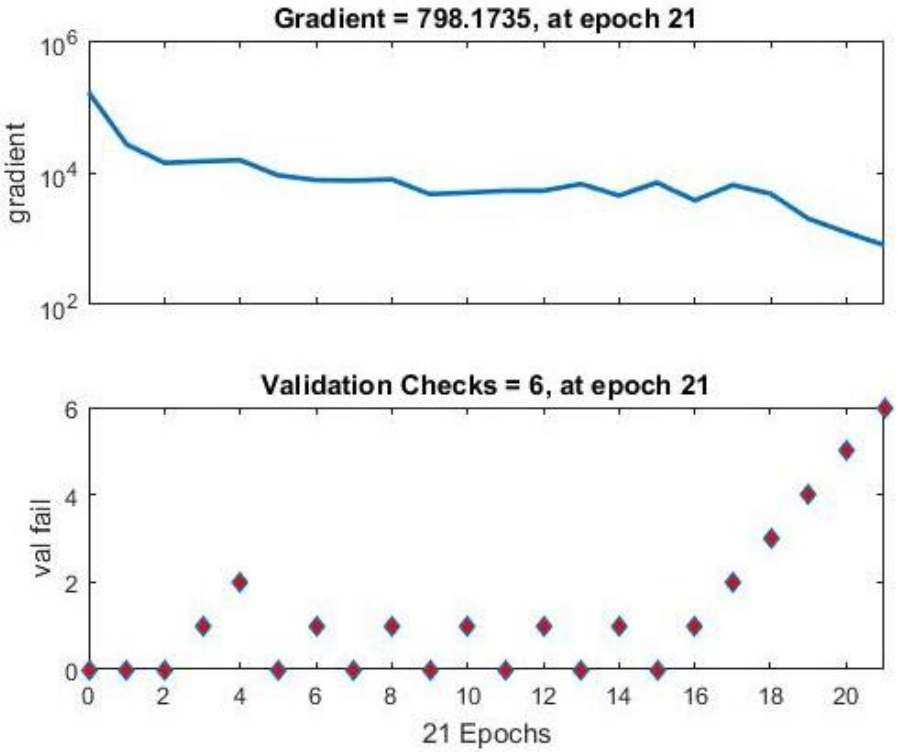

Figure 10. The training state with validation checks for the SCG Algorithm: gradient at epoch 21.

The ANN model results are quite comparable with the RSM modeling (Response surface methodology). Although, in current work, advanced neural network methodology is represented for the classification studies for the library of compounds in drug discovery. The epoch along 0–9 hidden nodes and varied iteration were undergone training. Though, in the course of 23 epochs and 10 hidden layers have shown optimum results and good output. The performance is mentioned as, random data division (dividerand) algorithm was used, scaled conjugated gradient (tracing) was used for training and graphs are plotted between 21 epochs and mean squared error (MSE) to report the performance. Default was used as derivative (defaultderiv). The progress is mentioned as Time: 0.00.05 s epoch with nine iterations (maximum stated–1000). Performance: Gradient: 798 with six validation checks. The value of R2 quantizes the correlation among targets and output in the regression curve. The R2 value close to one depicts the close relationship, and a zero value shows a random relationship between the test compounds. Mean squared error (MSE) is a regular squared difference among targets and productions. According to the errors, adjustment of the network is done. The application of parameters is done for measuring the generalization of the network and to stop the training accordingly when generalization stops to improve. The training is not affected by this; hence it generates an independent measure of the network performance in the duration and after the training of sample output. The training sample output was taken 70% of the total 22 samples. Three samples were used for validation, which gave 15% output. Similarly, three samples gave 15% output for testing. The correlation among targets and output was measured in terms of regression R2 and MSE of the measured performance. The zero value of R2 denotes a random relationship, while one denotes a close relationship. With the training of the model multiple times, it generates different results because of different conditions and initial sampling. Once the generalization stops further improving, the training stops automatically, which is an indication of increased mean square error for validation samples. In parallel network 1 and network 2, 10 hidden nodes were used, and both consist of different numbers of hidden layers. The optimal results were achieved with 10 hidden nodes, and further training stops when generalization also stopped improving. Training does not get affected by the nodes of validation, and it provides independent results for network performance or training.

4. Conclusions

In this work, we have developed a neural network model based upon the SMILES annotation of lipopeptide-based compound surfactin and its library of similar compounds. The model is capable of adequately discriminate between the compounds [2,19]. Majorly, computer-aided drug discovery is based upon the three-dimensional visualization of chemical space and its interaction with the surrounding atoms of the target protein. The current approach of utilizing linear notation will further help in a combinatorial chemistry-based chemical library preparation for the identification of novel molecules generated through certain modifications in the given parent compound. The identification and classification model developed using LM and SCG algorithm with the appropriate value of R2 close to one measure the closely related compounds, and R2 value close to zero depicts the random compounds. Such categorization will help in the screening of compounds of interest with slight structural variation to obtain the better affinity in the vicinity of the binding pocket of target proteins of diseases; hence, it will give an insight for analogue-based drug discovery for lead generation and lead optimization. Such model-based classification studies will give a boost to the computational drug discovery process for complex large molecules such as peptide therapeutics. The aim of the work is to utilize the chemical information of complex compounds for the efficient categorization of substructures and analogous compounds of the parent compound. It is concluded from the current study that the performance of predicted results can be increased with a varied number of neurons of hidden layers, even with the increased complexity of the algorithm with equal computational time. Evidently,

prediction performance also increases with enhanced iterations; therefore, higher training of artificial neural network is correspondence to better accuracy with certain limitations.

Author Contributions: Data curation, M.Y.; formal analysis, M.Y.; methodology, M.Y.; resources, S.E.J.; software, S.E.J.; supervision, S.E.J.; validation, S.E.J.; writing–original draft, M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We are thankful to the National Institute of Technology Raipur and the Chhattisgarh Council of Science and Technology (CCOST) (Project number 2487/CCOST/MRP/2016, Raipur dated 25 January 2016), India, for providing the necessary facilities to prepare the manuscript and permission to publish it.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hirohara, M.; Saito, Y.; Koda, Y.; Sato, K.; Sakakibara, Y. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinform.* **2018**, *19*, 83–94. [CrossRef] [PubMed]
2. Ballester, P.J.; Richards, W.G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **2007**, *28*, 1711–1723. [CrossRef] [PubMed]
3. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36. [CrossRef]
4. Helma, C. *Predictive Toxicology*; Taylor and Francis: Washington, DC, USA, 2005.
5. Ma, J.; Sheridan, R.P.; Liaw, A.; Dahl, G.E.; Svetnik, V. Deep neural nets as a method for quantitative structure—Activity relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263–274. [CrossRef] [PubMed]
6. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity prediction using deep learning. *Front. Environ. Sci.* **2016**, *3*. [CrossRef]
7. Alipanahi, B.; Delong, A.; Weirauch, M.T.; Frey, B.J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **2015**, *33*, 831–838. [CrossRef] [PubMed]
8. Zhou, J.; Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **2015**, *12*, 931–934. [CrossRef] [PubMed]
9. Kelley, D.R.; Snoek, J.; Rinn, J.L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **2016**, *26*, 990–999. [CrossRef] [PubMed]
10. Lanchantin, J.; Singh, R.; Wang, B.; Qi, Y. Deep motif dashboard: Visualizing and understanding genomic sequences using deep neural networks. *Pac. Symp. Biocomput.* **2017**, *22*, 254–265. [CrossRef] [PubMed]
11. Zeng, H.; Edwards, M.D.; Liu, G.; Gifford, D.K. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics* **2016**, *32*, i121–i127. [CrossRef]
12. Jujjavarapu, S.E.; Deshmukh, S. Artificial neural network as a classifier for the identification of hepatocellular carcinoma through prognosticgene signatures. *Curr. Genom.* **2018**, *19*, 483–490. [CrossRef] [PubMed]
13. Meena, K.R.; Kanwar, S.S. Lipopeptides as the antifungal and antibacterial agents: Applications in food safety and therapeutics. *BioMed Res. Int.* **2015**, *2015*, 1–9. [CrossRef] [PubMed]
14. Jujjavarapu, S.E.; Dhagat, S.; Yadav, M. *Computer-Aided Design of Antimicrobial Lipopeptides as Prospective Drug Candidates*; CRC Press: Boca Raton, FL, USA, 2019.
15. Kracht, M.; Rokos, H.; Özel, M.; Kowall, M.; Pauli, G.; Vater, J. Antiviral and hemolytic activities of surfactin isoforms and their methyl ester derivatives. *J. Antibiot.* **1999**, *52*, 613–619. [CrossRef]
16. Shoombuatong, W.; Schaduangrat, N.; Nantasenamat, C. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. *EXCLI J.* **2018**, *17*, 734–752.
17. Surfactin | C53H93N7O13. Available online: https://pubchem.ncbi.nlm.nih.gov/compound/443592 (accessed on 20 May 2021).
18. Liu, J.-F.; Mbadinga, S.M.; Yang, S.-Z.; Gu, J.-D.; Mu, B.-Z. Chemical Structure, property and potential applications of biosurfactants produced by Bacillus subtilis in petroleum recovery and spill mitigation. *Int. J. Mol. Sci.* **2015**, *16*, 4814–4837. [CrossRef] [PubMed]
19. Cochrane, S.A.; Vederas, J.C. Lipopeptides from Bacillus and Paenibacillus spp.: A gold mine of antibiotic candidates. *Med. Res. Rev.* **2016**, *36*, 4–31. [CrossRef] [PubMed]
20. Poroikov, V.V. Computer-aided drug design: From discovery of novel pharmaceutical agents to systems pharmacology. *Biochemistry* **2020**, *66*, 30–41. [CrossRef]