Snake Toxin Classification using Neural Networks

Akash Nag[#1], Sunil Karforma[*2]

[1,2]Dept. of Computer Science, The University of Burdwan
Rajbati, Burdwan, 713104

Abstract— *In this paper, a method is presented for classifying snake toxins into neurotoxins, cytotoxins or cardiotoxins, using a four-stage neural network. A training phase was used to teach the network to recognize the type of toxin based on the number of amino acid residues between the disulphide bridges, on a sample set of 139 snake toxins. When the trained network was used to classify a set of 239 toxins, the system achieved an accuracy of 74%.*

Keywords — *snake toxins, snake venom, neurotoxin, cytotoxin, cardiotoxin, artificial neural networks, classification, bioinformatics*

I. INTRODUCTION

Snake toxins are a group of peptides found in snake venoms that have one or more adverse physiological effects on humans and/or animals [1] [2]. These toxins can be classified into broadly two groups: neurotoxins and cytotoxins. Neurotoxic proteins work by binding to the nicotinic acetylcholine receptors in the post-synaptic membrane of muscles, thereby blocking acetylcholine binding and in turn preventing the excitation of muscles, often leading to paralysis and death. Cytotoxins work by making cells undergo necrosis as a result of which cell lysis occurs. During this phase, cells often exhibit swelling, and lose membrane integrity, thereby releasing cell contents into its environment. Both of these toxins are hazardous to human health, and snake antivenoms are often the only alternative left for mitigating its effects.

Neurotoxic proteins also exhibit a wide variety, for example long and short neurotoxins, with the short neurotoxins being more similar to the cytotoxins than to the long chain neurotoxins [3]. Cardiotoxins are a type of cytotoxin, which have a direct impact on cardiac muscle tissues. Preliminary work on classification of snake toxins was done by Dufton [3], based on overall chain lengths between the eight cysteine residues which invariably occur in almost all snake toxins and are responsible for forming four disulphide bridges. These cysteines were used as reference points and the inter-cysteine chain length was used for classification. Classification using decision trees was done by Nag et al. [4] based on the presence or absence of each possible tri-mer in the protein sequence.

The basic idea of neural computing was first developed from the computational model of a physiological brain based on threshold logic [5]. Artificial neural networks are a computational tool modelled on the structure and behaviour of neurons and synapses in the human brain, and these networks can be trained to recognize and classify complex patterns [6]. Since its inception, neural networks have been used in many classification tasks such as facial recognition [7], fingerprint recognition [8], speech recognition [9], and handwriting recognition [10]. Artificial neural networks are generally organized into layers, also called stages. Each layer has one or more nodes. A network has atleast two stages: the input layer, and the output layer. Between these two layers, zero or more hidden layers may exist. The neural network used in our work are of the feed-forward type, wherein nodes in one layer are connected (via edges) to nodes in the next layer only. Edges have weights associated with it. Each layer also usually has a bias node. Nodes receive inputs from nodes in the previous layer and output a weighted sum of those inputs, often modified using some threshold logic or activation function. Pattern recognition is achieved by adjusting the weights of the edges to minimize the error at each step, also called epoch, and the network is said to be learning from its experience of repeated classification and misclassification efforts during the training phase.

A major feature of snake toxins is the presence of eight cysteine residues, which form four disulphide bridges among them. This feature indicates that the snake toxin family is homologuous [3], and this signature is also used for deriving a motif or pattern in recognizing this protein family [11]. In this paper, we improve on Dufton's method [3] by using these cysteines as fixed reference points, and counting the frequency of each amino-acid residue in every inter-cysteine region, and feed this as an input to a four-stage artificial neural network. The outputs are decoded to classify each toxin into one of the three categories of neurotoxin, cytotoxin or cardiotoxin variety. The rest of this paper is organized as follows: in Section II we discuss the methods used, the network topology, and the data. In Section III, we discuss the results obtained by us and the accuracy of the system, and finally we conclude with the general implications of our results in Section IV.

II. METHODS

A. Data

The snake toxin data was obtained from SwissProt (Nov. 2015 Release) [12]. The training set used for training the neural network consisted of 122 snake toxins ranging in length from 60 to 108 residues, having 87 neurotoxins (71%), 30 cytotoxins (25%) and 5 cardiotoxins (4%), from 37 species of snakes. The validation set consisted of 239 snake toxins from 53 species, ranging in length from 57 to 147 residues, having 176 neurotoxins (74%), 58 cytotoxins (24%) and 5 cardiotoxins (2%). The training and test datasets are listed fully in Table I and Table II respectively. Both long and short chain neurotoxins as well as elapitoxins, bungarotoxins, cobrotoxins, neurotoxin homologs and weak neurotoxins have been collectively grouped under neurotoxins.

B. Network Topology

The artificial neural network (ANN) used in our work is a feed-forward network with 4 layers: the input layer, two hidden layers and an output layer. The input layer consists of a total of 701 nodes (700 input nodes plus one bias node), while the output layer consists of 3 nodes. The two hidden layers have 101 (100 nodes and 1 bias node), and 11 (10 nodes and 1 bias node) nodes respectively. The input is binary encoded, and the outputs are also received in binary form, which is decoded to obtain the results. The weights on the edges range between −1.0 and +1.0. A sigmoid function (see Eqn. (1)) was used for the producing node outputs. The topology is illustrated in Fig. 1.

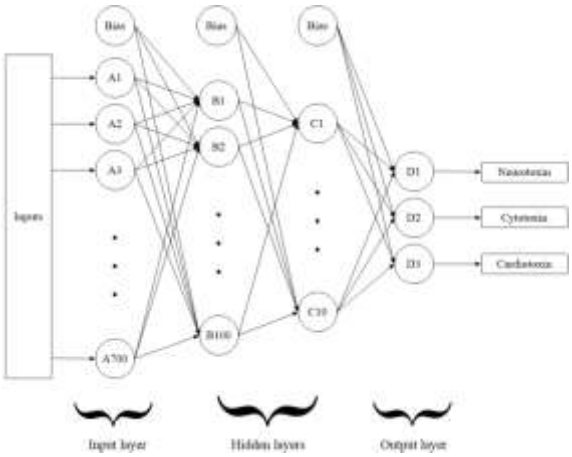

Fig. 1. The neural network topology

C. Input encoding and output decoding

The inputs to the neural network must be provided in binary form, however the snake toxin data consists of a set of protein sequences. That data must be encoded into binary before being fed into the network. This is achieved by first determining the positions of the eight signature cysteine residues in each sequence. Therefore, there are seven inter-cysteine regions (region between two consecutive cysteines) in each sequence, in which the frequency (F) of each amino acid residue is determined. As a result, we obtain 7 × 20 frequency counts (20 amino acids for each of the 7 regions). The logarithmic value of the percentage frequency (L) is then obtained using Eqn. (2) for each of those frequency values.

$$S(x) = \frac{1}{1+e^{-x}} \qquad (1)$$

$$L_b(a) = 10 \times \log_{10}\left(\frac{F_b(a)}{|R_b|} \times 100\right) \qquad (2)$$

Where, $F_b(a)$ is the frequency of amino-acid a in the inter-cysteine region b, and $|R_b|$ is the length of that region.

This logarithmic value ranges between 0 and 20, which is then converted into binary resulting in a 5-bit binary number. Therefore, the whole data can then be quantified in 7 × 20 × 5 = 700 bits. These 700 bits are fed as inputs to the network.

The network has 3 output nodes, which output a binary value, and is decoded as in Table III. Combinations of output values not shown in the table are treated as incorrect classifications.

D. Training and classifying

After the toxin data was encoded into binary, the training-set was fed into the neural network and it was trained till the output error for each sequence dropped below 0.05. Initial weights were assigned randomly. The learning rate was set to a constant 0.9. After the system was trained, the validation-set was fed into the network and the outputs decoded and recorded. The process is illustrated in Fig. 2.

III. RESULTS AND DISCUSSION

The neural network was implemented in Java 8, on a standard Intel Celeron 1 GHz processor with 2GB memory. The experiment was repeated several times, and the system achieved a peak accuracy of 73.64% after it was trained for 75 epochs in 89.9 seconds. The classification statistics for each type of toxin are tabulated in Table IV. The summed mean squared errors (MSEs) seen during the epochs are presented in Fig. 3. The decreasing MSEs signify that the network is being gradually trained to recognize the pattern, as it makes less and less misclassifications.

From Table IV, we can see that there were no false positives, i.e. no toxin was incorrectly classified as being of some other category. Failed

classifications refer to instances where the outputs could not be uniquely decoded, i.e. those outputs which are not present in Table III. It can be seen that the system routinely fails to detect cardiotoxins, however it can be attributed to the lack of sufficient data in the cardiotoxin category.

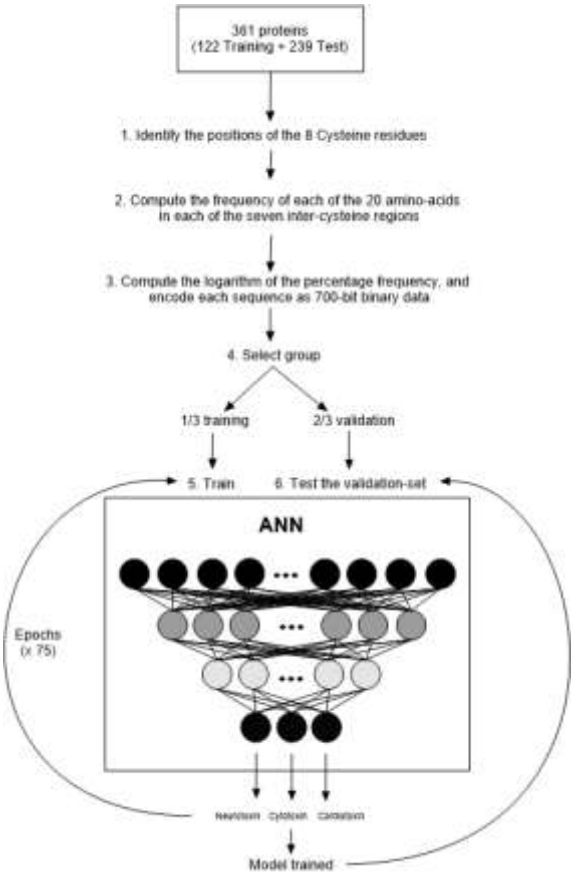

Fig. 2. The training and classification process

IV. CONCLUSIONS

Neural networks have been widely used in bioinformatics for various classification purposes, such as prediction of secondary structure [13]. However, one of the main difficulties in using neural networks is deciding on the optimum network topology. The universal approximation theorem [14] states that feed-forward networks with a single hidden layer are universal approximators in $C(R^m)$. Networks with two hidden layers can approximate any arbitrary function.

Increasing the number of hidden layers further increases training time exponentially, without any significant gain in accuracy. Therefore, we have decided on using two hidden layers in our neural network. The accuracy ranged from 55% to 74% during our experiments with an average accuracy of 64%. The training time ranged from 55 seconds to 5 minutes. Since the number of input nodes in our network is huge, increasing the nodes in the hidden layers was not feasible, and during our trials, even increasing the number of hidden neurons did not have any appreciable gain in performance.

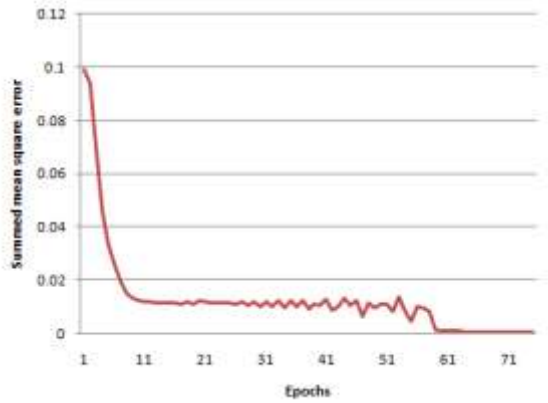

Fig. 3. Plot of MSE as the network is trained through the epochs

It is clear that the entire snake toxin family is homologous, and particularly the long and short chain neurotoxins are quite similar to each other. However, there are two other points to consider: firstly, the short neurotoxins are more similar to the cytotoxins than the long neurotoxins, and secondly, the cytotoxic group is quite varied implying that it has undergone extensive functional evolution, and do not have any close affinity with one another [3]. This is further strengthened when we look at the accuracy of the neural network in each toxin category, where we find (see Table IV) that accuracy rates for cytotoxins and cardiotoxins (which are also cytotoxins) are quite lower than those of neurotoxins.

From Table IV, we can see that there were no false positives, i.e. no toxin was incorrectly classified as being of some other category. Failed classifications refer to instances where the outputs could not be uniquely decoded, i.e. those outputs which are not present in Table III. It can be seen that the system routinely fails to detect cardiotoxins, however it can be attributed to the lack of sufficient data in the cardiotoxin category.

TABLE I. THE TRAINING-SET USED FOR TRAINING THE NEURAL NETWORK

Species	Min. length	Max. length	No. of proteins	Neurotoxins	Cytotoxins	Cardiotoxins
Acanthophis antarcticus	74	79	3	3	-	-
Aspidelaps scutatus	68	68	1	1	-	-
Austrelaps labialis	96	107	2	2	-	-
Austrelaps superbus	90	92	2	2	-	-
Bungarus caeruleus	76	76	1	1	-	-
Bungarus candidus	73	87	5	5	-	-
Bungarus multicinctus	87	103	8	6	1	1
Demansia vestigiata	88	88	2	2	-	-
Dendroaspis jamesoni kaimosae	72	72	1	1	-	-
Dendroaspis polylepis polylepis	72	72	4	4	-	-
Dendroaspis viridis	72	73	2	2	-	-
Drysdalia coronoides	88	108	7	7	-	-
Hydrophis hardwickii	92	93	3	3	-	-
Hydrophis stokesii	70	72	2	2	-	-
Laticauda colubrina	69	69	1	1	-	-
Laticauda laticaudata	87	87	1	1	-	-
Laticauda semifasciata	87	87	1	1	-	-
Naja anchietae	72	72	1	1	-	-
Naja annulata annulata	71	71	1	1	-	-
Naja annulifera	60	60	2	-	2	-
Naja atra	81	81	10	-	10	-
Naja haje haje	71	71	1	1	-	-
Naja kaouthia	60	81	3	1	2	-
Naja melanoleuca	60	71	3	2	1	-
Naja mossambica	60	60	2	-	2	-
Naja naja	60	71	7	5	2	-
Naja nivea	60	71	3	1	2	-
Naja oxiana	60	73	3	1	2	-
Naja pallida	60	60	1	-	1	-
Naja sputatrix	81	90	6	1	5	-
Notechis scutatus scutatus	94	94	1	1	-	-
Ophiophagus hannah	72	94	24	20	-	4
Oxyuranus microlepidotus	92	92	4	4	-	-
Oxyuranus scutellatus scutellatus	92	92	1	1	-	-
Pseudechis australis	89	89	1	1	-	-
Pseudonaja textilis	89	89	1	1	-	-
Tropidechis carinatus	93	93	1	1	-	-
37 species	**60**	**108**	**122**	**87**	**30**	**5**

TABLE II. THE VALIDATION-SET USED FOR TESTING THE NEURAL NETWORK

Species	Min. length	Max. length	No. of proteins	Neurotoxins	Cytotoxins	Cardiotoxins
Acanthophis antarcticus	62	62	1	1	-	-
Aipysurus laevis	60	81	3	3	-	-
Aspidelaps scutatus	63	64	2	-	2	-
Austrelaps superbus	81	81	1	1	-	-
Bungarus caeruleus	147	147	2	2	-	-
Bungarus candidus	84	147	9	9	-	-
Bungarus fasciatus	63	86	7	7	-	-
Bungarus flaviceps flaviceps	146	146	2	2	-	-
Bungarus multicinctus	83	147	29	28	-	1
Cryptophis nigrescens	81	81	2	2	-	-
Demansia vestigiata	84	84	1	1	-	-
Dendroaspis angusticeps	66	66	1	1	-	-
Dendroaspis jamesoni kaimosae	60	60	1	1	-	-
Dendroaspis polylepis polylepis	60	65	3	3	-	-
Dendroaspis viridis	60	60	1	1	-	-
Drysdalia coronoides	78	81	2	2	-	-
Hemachatus haemachatus	61	63	6	2	4	-
Hoplocephalus stephensii	81	81	1	1	-	-
Hydrophis cyanocinctus	60	79	2	2	-	-
Hydrophis hardwickii	81	81	3	3	-	-
Hydrophis lapemoides	60	60	1	1	-	-
Hydrophis ornatus	60	60	1	1	-	-
Hydrophis peronii	81	81	2	2	-	-
Hydrophis schistosus	60	60	1	1	-	-
Laticauda colubrina	83	83	5	5	-	-
Laticauda crockerii	62	62	3	3	-	-
Laticauda laticaudata	62	83	7	7	-	-
Laticauda semifasciata	83	83	1	1	-	-
Micrurus corallinus	78	86	3	3	-	-
Micrurus pyrrhocryptus	60	60	1	1	-	-
Micrurus surinamensis	58	64	4	4	-	-
Naja annulata annulata	61	61	1	1	-	-
Naja annulifera	60	62	13	4	9	-
Naja atra	60	86	31	9	20	2
Naja christyi	62	62	1	1	-	-
Naja haje haje	60	61	2	1	1	-
Naja kaouthia	60	86	10	5	5	-
Naja melanoleuca	61	61	3	1	2	-
Naja mossambica	60	62	5	2	3	-
Naja naja	60	83	8	5	3	-
Naja nivea	60	61	2	1	1	-
Naja oxiana	60	61	2	1	1	-
Naja pallida	61	61	1	1	-	-
Naja philippinensis	61	61	1	1	-	-
Naja sagittifera	60	60	1	-	1	-
Naja samarensis	61	61	1	1	-	-
Naja sputatrix	60	86	16	10	6	-
Notechis scutatus scutatus	81	81	1	1	-	-
Ophiophagus hannah	57	86	15	13	-	2
Oxyuranus microlepidotus	83	83	2	2	-	-
Oxyuranus scutellatus scutellatus	62	83	3	3	-	-
Pseudechis australis	83	83	1	1	-	-
Pseudechis porphyriacus	83	83	1	1	-	-
Pseudonaja textilis	79	79	7	7	-	-
Tropidechis carinatus	81	81	2	2	-	-
53 species	**57**	**147**	**239**	**176**	**58**	**5**

TABLE III. DECODING THE OUTPUT FROM THE NEURAL NETWORK

D1	D2	D3	Classification output
1	0	0	Neurotoxin
0	1	0	Cytotoxin
0	0	1	Cardiotoxin

TABLE IV. CLASSIFICATION RESULTS

Toxin type	No. of proteins	True positives	False positives	False negatives	Failed classifications
Neurotoxin	176	134 (76.1%)	0	42 (23.8%)	42
Cytotoxin	58	40 (68.9%)	0	18 (31%)	18
Cardiotoxin	5	2 (40%)	0	3 (60%)	3
Total	**239**	**176**	**0**	**63**	**63**
Percentages	**-**	**73.64%**	**0%**	**26.35%**	**26.35%**

REFERENCES

[1] Endo, T., and N. Tamiya. "Snake toxins", Harvey A.L., Ed., pp165-222, Pergamon Press, New-York, (1991).

[2] Mebs D., Claus I. "Snake toxins", Harvey A.L., Ed., pp425-447, Pergamon Press, New-York, (1991).

[3] Dufton, M. J. "Classification of elapid snake neurotoxins and cytotoxins according to chain length: evolutionary implications." Journal of molecular evolution 20.2 (1984): 128-134.

[4] Nag, Akash, and Sunil Karforma. "Classifying Snake Toxins using Decision Trees"., IJSR 5.8 (2016): 262-263.

[5] McCulloch, W. & Pitts, W. (1943) Bull. Math. Biophys. 5, pp115-123.

[6] Bishop, Christopher M. Neural networks for pattern recognition. Oxford university press, 1995.

[7] Latha, P., L. Ganesan, and S. Annadurai. "Face recognition using neural networks." Signal Processing: An International Journal (SPIJ) 3.5 (2009): 153-160.

[8] Leung, W. F., et al. "Fingerprint recognition using neural network." Neural Networks for Signal Processing [1991]., Proceedings of the 1991 IEEE Workshop. IEEE, 1991.

[9] Tebelskis, Joe. "Speech recognition using neural networks". Diss. Siemens AG, 1995.

[10] Gentric, Philippe. "Handwritten character recognition using neural networks." ICOHD 93 (1993): 13-15.

[11] Jonassen, Inge, John F. Collins, and Desmond G. Higgins. "Finding flexible patterns in unaligned protein sequences." Protein science 4.8 (1995): 1587-1595.

[12] Bairoch, Amos, and Rolf Apweiler. "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000." Nucleic acids research 28.1 (2000): 45-48.

[13] Holley, L. Howard, and Martin Karplus. "Protein secondary structure prediction with a neural network." Proceedings of the National Academy of Sciences 86.1 (1989): 152-156.

[14] Csáji, Balázs Csanád. "Approximation with artificial neural networks." Faculty of Sciences, Etvs Lornd University, Hungary 24 (2001): 48.