Volume : 5 | Issue : 8 | August 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

Original Research Paper

Classifying Snake Toxins Using Decision Trees

Computer Science

KEYWORDS : snake toxins; classification; decision trees; protein sequence analysis; bioinformatics

Akash Nag — Dept. of Computer Science, The University of Burdwan, Burdwan, WB, India 713104

Sunil Karforma — Dept. of Computer Science, The University of Burdwan, Burdwan, WB, India 713104

ABSTRACT

Snake toxins are a group of peptides found in snake venoms, which have some form of adverse physiological effects on humans and/or animals. Primarily these consist of neurotoxins, cytotoxins and cardiotoxins, which directly affect the nervous system, skin and muscle tissues, and the heart muscle tissues respectively. In this short paper, we show how, given a protein sequence, we can classify it either as a neurotoxic, a cytotoxic or a cardiotoxic peptide using decision trees. Such classification can be helpful when the phsiological effects of a new venom peptide need to be determined given its sequence information only.

1. INTRODUCTION

Snake toxins are a group of peptides found in snake venoms [1] [2] [3]. These consist of various types of toxins, namely long and short neurotoxins, cytotoxins and various other short and weak toxins. Cardiotoxins are a special type of cytotoxins that directly affect the heart.

A distinct feature of all snake toxins is the presence of cysteines [1], which are involved in disulphide bridges thought to be essential for its molecular structure and hence function. Classification of snake toxins was attempted by Dufton [1] based on inter-cysteine chain length. The approach is based on the evolutionary view of the homology of all snake toxins. The eight cysteins have been used as fixed reference points in the phylogeny and the chain length (i.e. number of amino acid residues between two cysteines) have been used as a measure of the number of insertion and deletion events during the course of evolution.

In this paper, we present a decision-tree based approach to performing the classification wherein we detect the presence or absence of all possible k-mers in the whole sequence.

2. METHODS

For the purposes of classification, 403 different snake toxins were selected from the broad groups of neurotoxins and cytotoxins. The neurotoxic group consisted of short and long neurotoxins, elapitoxins, alpha neurotoxins, weak neurotoxins, cobrotoxins and bungarotoxins. The cytotoxic group consisted of cytotoxins and cardiotoxins. All proteins were from snakes in the Elapidae family. A class, one of three: neurotoxins, cytotoxins or cardiotoxins, was assigned to each of these sequences. In total there were 293 neurotoxins, 99 cytotoxins and 11 cardiotoxins in the dataset.

A representative sample, about 34.5% the size of the original dataset, was selected for training purposes. The sample consisted of 85 neurotoxins, 43 cytotoxins and 11 cardiotoxins.

For every protein sequence, the absence or presence of all possible 3-mers were computed. This was done by breaking the sequence into overlapping and consecutive 3-mers, and matching against all of the possible 20 × 20 × 20 = 8,000 k-mers.

This count-data from the sample was then presented to the popular decision tree algorithm C4.5 [4] to generate a decision-tree. The input to the decision tree is an $N \times 8000$ matrix M of boolean values (indicating absence or presence of a tri-mer), where N is the number of sequences. For the whole dataset, N=403, for the training-set, N=139. $M[i,j]$ indicates whether the tri-mer represented by j is absent or present in sequence i, where $1 \le j \le 8000$. The C4.5 algorithm successively partitions the sample, in this case into two-subsamples, at each iteration based on a specific attribute. The partitioning attribute is chosen based on the one which has the maximum gain ratio. For our dataset, the classes C = { neurotoxin, cytotoxin, cardiotoxin }.

A sample S is partitioned as follows:

When all records in S belong to the same class, it is assigned to be a leaf of the tree.

When S contains no records, it is assigned to be a leaf of the tree.

When S contains records belonging to more than one class, S must be partitioned or refined into subsamples. A node for S is assigned to the tree, and children nodes are created under it which will hold the subsamples.

There are many ways for testing which attribute should be chosen for partitioning the sample, but the most common test is the test of entropy. The entropy of a sample S is given by:

$$Info(S) = \sum_{i=1}^{k} \left[\left(\frac{freq(C_i, S)}{|S|} \right) . \log_2 \left(\frac{freq(C_i, S)}{|S|} \right) \right] \quad (1)$$

Where, k is the number of classes; in our case k=3,

$|S|$ represents the number of records in sample S,

freq(C_i,S) represents the number of records in S belonging to class C_i.

After S has been partitioned based on the n possible outcomes (values) for each attribute X, we compute the following:

$$Info_X(S) = \sum_{i=1}^{n} \left[\frac{|S_i|}{|S|} . Info(S_i) \right] \quad (2)$$

Volume : 5 | Issue : 8 | August 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

$$Gain(X) = Info(S) - Info_X(S) \quad (3)$$

In our case, n=2 (as absent or present are the only possible outcomes for each tri-mer), so S_1 represents the records having value false, and S_2 represents the records having value true for the tri-mer identified by X.

The attribute (tri-mer) X, having the highest Gain value is selected as the partitioning attribute. Since all tri-mers can have only one of two choices, each node is partitioned into exactly two child nodes.

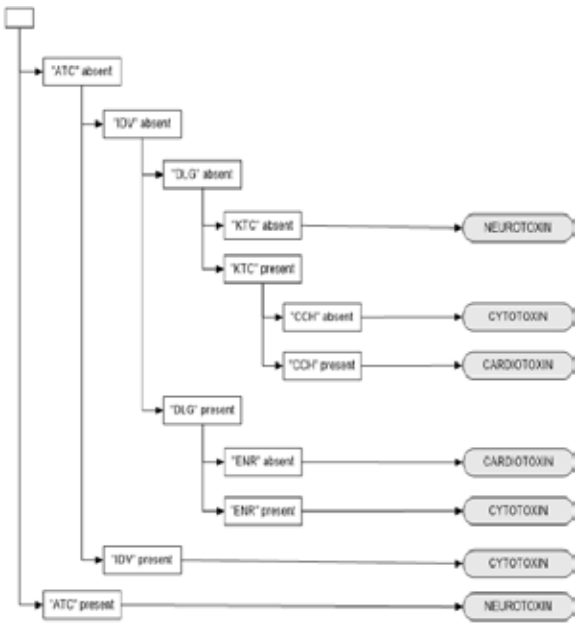

Fig. 1. The sample-trained decision-tree for snake toxin classification

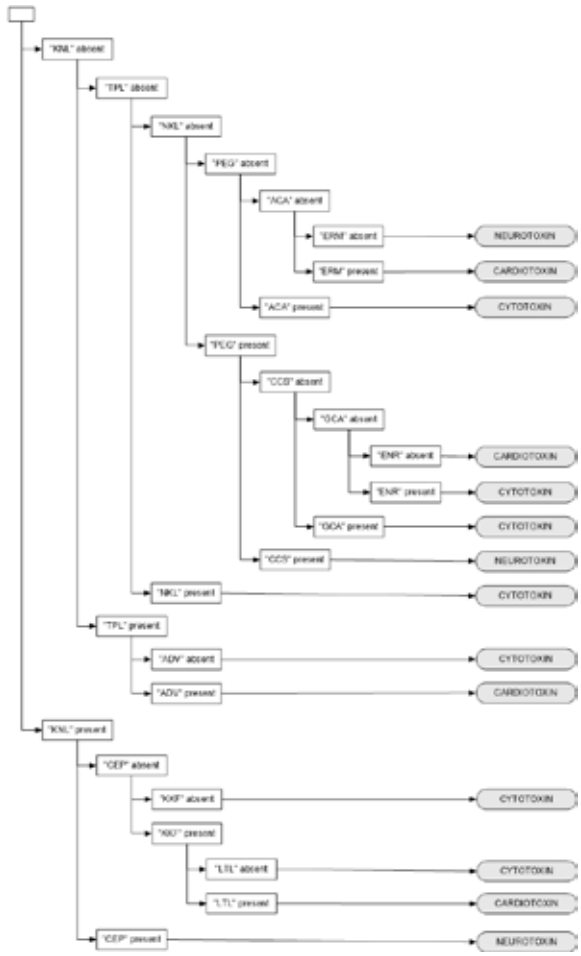

Fig. 2. The decision-tree obtained from the entire dataset

3. RESULTS

The sample trained decision tree is presented in Fig. 1, and has an accuracy of 74.68%. The decision tree obtained from the entire dataset is presented in Fig. 2, which ofcourse exactly fits the data. The snake toxin sequences were obtained from the SwissProt [5] database using the hits of the PROSITE [6] SNAKE_TOXIN motif (PS00272) [7]. The algorithm was implemented using Java 8 on a standard Intel Pentium Celeron 1GHz processor. Both trees were generated in under a second, on the system.

4. CONCLUSION

The approach used by us in this work, is purely from a statistical consideration of the data, and as such does not have any evolutionary implications, as does Dufton's [1] work, which is based on the indel, substitution and mutation events that have occurred through the snake toxin family. However, using the decision tree generated by our method, one can simply and quickly classify a snake toxin into one of the three groups. Our method only requires a linear scan of any given sequence once to identify which tri-mers are present, and then feed that data to our tree (Fig. 2) which has a maximum depth of 6. On the other hand, in Dufton's method we have to first locate the cysteines, and then count the chain lengths in the inter-cysteine regions and then too the classification is complex as we have to determine under which bounds does each region fall, to identify the group.

REFERENCES

[1] 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.

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

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

[4] Quinlan, J. Ross. C4. 5: programs for machine learning. Elsevier, 2014.

[5] 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.

[6] Falquet, Laurent, et al. "The PROSITE database, its status in 2002." Nucleic acids research 30.1 (2002): 235-238.

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