International Conference on Technologies for Sustainability- Engineering, Information Technology, Management and the Environment
DAV Institute of Management, Faridabad (SUSTECH-15)
28 November 2015 www.conferenceworld.in ISBN: 978-81-931039-7-5

IDENTIFYING NEUROTOXINS IN SNAKE VENOMS THROUGH AUTOMATED GENERATION OF PROBABILISTIC CONSENSUS MOTIFS

Akash Nag[1], Dr. Sunil Karforma[2]

[1,2]*Department of Computer Science, The University of Burdwan, Burdwan, WB, (India)*

ABSTRACT

In this paper, we propose an algorithm that automatically generates a consensus motif identifying a protein family given an alignment of a set of protein sequences. The algorithm also assigns probabilistic scores to each amino-acid choice at each position in the sequence that will better aid in matching the motif against a protein database. Using aligned neurotoxin sequences from 91 species of Elapid snakes, the neurotoxin-motif generated by our algorithm successfully matched 175 proteins, including the 91 input proteins, in the UniProt database, most of which were also neurotoxins from other snake venoms.

Keywords: Algorithms, Bioinformatics, Neurotoxins, Protein Motifs, Snake Venoms

I. INTRODUCTION

The PROSITE [1] database is a manually curated protein database created in 1988. It consists of various consensus motifs identifying various protein families. These patterns can help in the identification of short and well-conserved regions in proteins. Regular expressions can then be easily constructed from these patterns and any protein database can be parsed using automatons derived from these regular expressions. The matching proteins can then be said to belong mostly to the protein family identified by the motif. Thus proteins with similar functions and structure can be quickly and easily identified by PROSITE entries. However, construction of these motifs is usually a manual and slow task.

Snake venoms are a complex cocktail of toxins and other proteins. The snake toxins can be categorized into various groups consisting of short and long neurotoxins, cytotoxins and short toxins, and miscellaneous other toxins. These toxins mostly act by binding to the nicotinic acetylcholine receptors in the postsynaptic membrane of the skeletal muscles thus preventing the binding of acetylcholine thereby blocking muscle excitation. These proteins usually vary in size from 60 to 75 amino acids. The PROSITE entry PS00272 describes most of these snake toxins.

In this paper, we focus on the short neurotoxins from 91 species of Elapid snakes, which are the most prevalent and deadly of all snakes responsible for deaths around the world. From analyzing the sequences of these toxins, we have aimed to create a consensus motif which would correctly identify other neurotoxins from other snakes in the same family. To automate the generation of motifs, we propose a simple process to generate patterns from alignments of these sequences. In Section 2, we analyze the tools currently available for automated generation of motifs. In Section 3, we present a simple process to quickly generate consensus motifs for a given set of

sequences; and then how these sequences can be used to estimate the confidence of matched proteins. Finally in Section 4 we analyze our algorithm's performance and present our results when applied to generate motifs for identification of neurotoxins.

II. RELATED WORK

The most important work in the field of automated motif generation was done by Jonassen et al. [2] through the development of the PRATT algorithm. Using wildcards and depth-first-search, PRATT works in 2 phases to identify all conserved regions in a given set of protein sequences. From these, it identifies several patterns with varying degree of fitness. The primary advantage of PRATT is that it does not require the input sequences to be aligned.

Several other algorithms like Roytberg et al. [3], Saki and Sternberg [4], and Lawrence et al. [5] are precursors to the PRATT algorithm in a way and show various strategies of pattern detection in sequences aiding in arriving at a high quality alignment. A combination of these algorithms can be used to arrive at a consensus pattern. Koza and Andre [6] developed an algorithm for motif generation based on genetic programming.

A comprehensive work related to the analysis of Elapid snake venoms – both neurotoxins as well as cytotoxins was done by Dufton and Hider [7].

III. THE PROPOSED ALGORITHM

3.1 Automated Motif Generation

The proposed process for automated motif generation can be enumerated as a sequence of steps as shown in Fig. 1. The algorithm starts off by performing a multiple sequence alignment using any MSA algorithm. For our purposes we used MAFFT (version 7) [8][9] for performing the MSA. The scanning phase starts next, which constructs the groups from the choices at each amino-acid position. The merging phase merges every similar successive group. The fourth step is optional and is used to further reduce the size of the motif by replacing long choice lists by "x" signifying that any amino acid can be present at that position.

After the output from the algorithm is generated, the conserved, conservative and semi-conservative regions can be manually scanned from the output of the alignment performed by the MSA algorithm. This can be then be used to further shorten the motif by replacing non-conserved regions by "x"s when choice-lists are sufficiently large or when it seems that the amino acid at that position does not play any pivotal role in the structure or function of the protein family.

Algorithm for automated motif generation

1. Input a set of protein sequences and align them using any multiple sequence alignment algorithm.
2. For each column position in the aligned set, do the following:
   a) Set *CHOICES* = the distinct amino acids present in that column
   b) For each amino acid, calculate the probability to be its frequency (in that column) divided by the number of protein sequences.
   c) Set *MIN*=1 and *MAX*=1 for that position. If there is a gap in this position for any sequence, set *MIN*=0.
   d) If *CHOICES* consist of only 1 amino acid, append it to output.

e) If *CHOICES* contains more than 1 amino acid, check if its length is more than half of that of *ALPHABET*, where *ALPHABET* is the list of possible amino acids (usually 20). If it is less, append *CHOICES* to output inside square brackets, else, append *ALPHABET – CHOICES* to output inside curly braces.

f) If *MIN≠MAX*, append their values to output, separated by a comma, inside parenthesis. If *MIN=MAX* and *MAX*>1, append the *MAX* value to output inside parenthesis.

3. After the preliminary output is produced, the merging phase commences. In this step, every successive group is merged if their constituent *CHOICES* are the same. If they are same, their *MIN* and *MAX* values are separately added to arrive at the combined *MIN* and *MAX* values.
4. A user tunable parameter called *CHOICE-THRESHOLD* determines which *CHOICES* are are converted to "x" (i.e. "ANY amino acid"). This can be tuned by the user according to the number and type of sequences being worked on. If the threshold is set to 5, then any *CHOICE* consisting of 5 or more amino acids are converted to "x"s. After all such conversions have been applied, a second merging phase merges all successive "x" groups into a single group.

Fig 1. Algorithm for Automated Motif Generation

3.2 Output Motif Format

The motif format output by the algorithm consists of both a pattern as well as a set of probabilities. A pattern consists of groups separated by hyphens. Each group may either consist of a single upper-case letter, a lower-case "x", or a group of letters enclosed within [ ] or { }. A single upper-case letter denotes a single amino-acid at that position, e.g. A (for Alanine). A lower-case "x" will match any amino-acid at that position. A group of letters inside [ ] signify matching against any one of those amino-acids, while a group of letters inside { } will match any amino-acid NOT specified by the group of letters. Additionally, after a group, one or two comma-separated numbers may be present inside parenthesis. A single number, e.g. A(2) represents 2 successive Alanines, while two numbers, e.g. A(2,4) represents 2, 3 or 4 Alanines. An example motif format may be: C-N-x-[ARV]-G(2,3)-N(0,1). For each column position, the proposed algorithm also outputs a set of probability scores. E.g. for [ARV] the scores may be 0.03, 0.94, 0.03; meaning that the amino acid at that position must be R for 94% of the cases, and A or V each for 3%.

3.3 Estimating Confidence of Matched Proteins

The output motif from the algorithm can then be used to scan for proteins in any protein database like UniProt [10]. The PROSITE web interface can be used to scan for proteins using any input motif for this purpose. The confidence or probability that a matched protein belongs to the target protein family can be computed using the probability scores. This can be done by taking the average of the probability scores for those amino acids which are present in the matched protein.

3.4 Performance

The algorithm has a running-time complexity of $O(N^2.M^2)$, where N is the number of protein sequences, while M is the length of each aligned sequence. The space-complexity of the algorithm is $O(M)$.

IV. MOTIF GENERATION FOR NEUROTOXINS

The proposed algorithm was applied to generate a motif for identification of neurotoxins in Elapid snake venoms. For input, we used short neurotoxin protein sequences from the UniProt database from 91 species for Elapid snakes. These sequences were aligned using the MAFFT algorithm. The minimum sequence length was 58 and the maximum was 86. The average sequence length was 71.

The program was developed using Java 8. The original output motif was 977 characters in length consisting of 97 groups, while the reduced motif (with THRESHOLD setting = 4) was 358 characters in length. The process was completed in 0.016 seconds on a 1.6GHz Intel Celeron M processor. The motifs and probability scores are presented in Fig. 2.

The output motif was then presented to the PROSITE web interface and matched against proteins from the SwissProt database. The motif successfully matched a total of 175 proteins, among them were: 107 short neurotoxins, 3 neurotoxin-like proteins, 12 weak toxins, 1 long neurotoxin homolog, 14 weak neurotoxins, 8 three-finger toxins, as well as 25 other neurotoxins. Thus, it is clear that the generated motif was general enough to match other neurotoxins as well, but not too loose to match unrelated proteins.

Output from the proposed algorithm:

Original Motif:

M(0,1)-K(0,1)-[NT](0,1)-L(0,3)-[IST](0,1)-[FLP](0,1)-[LV](0,1)-[MV](0,1)-[LMV](0,1)-T(0,1)-[IMTV](0,1)-[AIMV](0,1)-[CF](0,1)-L(0,1)-D(0,1)-L(0,1)-V(0,1)-[ACG](0,1)-[CHY](0,1)-[ST](0,1)-[IKLMR]-[EIKNQRT]-C-[CFHLNRVY]-[GIKMNQRT]-[DGHIKLNQRSV](0,1)-[DFLMPQRY](0,1)-[ALQST](0,1)-[AILQSTY](0,1)-[AEFNQT](0,1)-[AFGHINPRT](0,1)-E(0,1)-D(0,1)-[PQ](0,1)-[AEFHIKNPQRSV]-[DGIKST]-[EFILNPST]-[EGKMQTV]-[ADGHILNPQRSTV]-C-[AEGKPRSY]-[DEGHLPQRS](0,1)-[ADGHIQSW](0,1)-[DEGHKMQV](0,1)-[DGHKLNSTY]-[DFILMNQSV]-C-[EFLQY]-[AEIKNQRS]-[EFKLMNQRTY]-[ADFILQRSTVWY]-[EHITVWY]-[HIKMNPQRSTWY](0,1)-[ADEFGPTV](0,1)-[FHMNSTY](0,1)-[FGHINPQ](0,1)-[GNRSY]-[GHNRSV]-[AEIKMPSTVWY]-[DEIKLMRV]-[HILPSTVY]-[ADEHKLSTY]-[KMRW]-G-C-[AGSTV]-[ALSTY](0,1)-[KNST](0,1)-C-[HP](0,1)-[DHKNPQRST](0,1)-[AEIMPV](0,1)-[KR](0,1)-Y(0,1)-[GHKNPQRS](0,1)-[GHNPR](0,1)-[DEIKLNPV](0,1)-[DEGHIKLNQRTY]-[ADILPRSV]-[AEHIKMNQRSTVY]-C(2)-[ADEGHKQRST]-[KRST]-[DEN]-[DEKLNR]-[CS]-N-[AEGKLNRY](0,1)-P(0,1)-S(0,1)-T(0,1)-P(0,1)-S(0,1)-T(0,1)

Reduced Motif:

M(0,1)-K(0,1)-[NT](0,1)-L(0,3)-[IST](0,1)-[FLP](0,1)-[LV](0,1)-[MV](0,1)-[LMV](0,1)-T(0,1)-x(0,2)-[CF](0,1)-L(0,1)-D(0,1)-L(0,1)-V(0,1)-[ACG](0,1)-[CHY](0,1)-[ST](0,1)-x(2)-C-x(2,8)-E(0,1)-D(0,1)-[PQ](0,1)-x(5)-C-x(3,6)-C-x(12,16)-G-C-x(1,3)-C-[HP](0,1)-x(0,2)-[KR](0,1)-Y(0,1)-x(3,6)-C(2)-x(2)-[DEN]-x-[CS]-N-x(0,1)-P(0,1)-S(0,1)-T(0,1)-P(0,1)-S(0,1)-T(0,1)

0	M(0,1)	0.56			
1	K(0,1)	0.56			
2	[NT](0,1)	0.04	0.52		
3	L(0,3)	0.56			
6	[IST](0,1)	0.01	0.02	0.53	
7	[FLP](0,1)	0.09	0.46	0.01	
8	[LV](0,1)	0.09	0.47		
9	[MV](0,1)	0.08	0.48		
10	[LMV](0,1)	0.02	0.01	0.53	
11	T(0,1)	0.56			
12	[IMTV](0,1)	0.51	0.03	0.01	0.01
13	[AIMV](0,1)	0.01	0.01	0.08	0.46
14	[CF](0,1)	0.55	0.01		
15	L(0,1)	0.56			
16	D(0,1)	0.56			

17	L(0,1)	0.56									
18	V(0,1)	0.01									
19	[ACG](0,1)	0.01	0.01	0.54							
20	[CHY](0,1)	0.01	0.01	0.54							
21	[ST](0,1)	0.01	0.55								
22	[IKLMR]	0.01	0.02	0.38	0.35	0.23					
23	[EIKNQRT]	0.14	0.23	0.03	0.01	0.01	0.16	0.41			
24	C	1.00									
25	[CFHLNRVY]	0.25	0.15	0.27	0.02	0.01	0.01	0.01	0.26		
26	[GIKMNQRT]	0.01	0.01	0.12	0.01	0.75	0.05	0.03	0.01		
27	[DGHIKLNQRSV](0,1)	0.01	0.05	0.09	0.01	0.01	0.02	0.01	0.68	0.03	0.01
		0.05									
28	DFLMPQRY](0,1)	0.02	0.01	0.01	0.01	0.08	0.67	0.01	0.09		
29	[ALQST](0,1)	0.01	0.02	0.01	0.75	0.02					
30	[AILQSTY](0,1)	0.01	0.02	0.02	0.02	0.68	0.04	0.01			
31	[AEFNQT](0,1)	0.01	0.08	0.02	0.01	0.65	0.04				
32	[AFGHINPRT](0,1)	0.07	0.02	0.01	0.01	0.01	0.01	0.64	0.02	0.02	
33	E(0,1)	0.01									
34	D(0,1)	0.01									
35	[PQ](0,1)	0.01	0.01								
36	[AEFHIKNPQRSV]	0.03	0.01	0.01	0.14	0.01	0.36	0.01	0.24	0.13	0.01
		0.02	0.01								
37	[DGIKST]	0.07	0.11	0.03	0.01	0.01	0.77				
38	[EFILNPST]	0.01	0.01	0.04	0.09	0.12	0.01	0.01	0.70		
39	[EGKMQTV]	0.02	0.01	0.37	0.02	0.10	0.40	0.08			
40	[ADGHILNPQRSTV]	0.01	0.04	0.01	0.01	0.02	0.01	0.12	0.02	0.02	0.02
		0.21	0.37	0.12							
41	C	1.00									
42	[AEGKPRSY]	0.26	0.11	0.02	0.10	0.35	0.01	0.13	0.01		
43	[DEGHLPQRS](0,1)	0.02	0.02	0.11	0.01	0.02	0.31	0.01	0.01	0.01	
44	[ADGHIQSW](0,1)	0.07	0.02	0.51	0.08	0.01	0.01	0.01	0.02		
45	[DEGHKMQV](0,1)	0.05	0.68	0.04	0.01	0.01	0.01	0.12	0.05		
46	[DGHKLNSTY]	0.01	0.01	0.01	0.13	0.01	0.30	0.22	0.30	0.01	
47	[DFILMNQSV]	0.01	0.11	0.09	0.03	0.01	0.22	0.01	0.51	0.01	
48	C	1.00									
49	[EFLQY]	0.01	0.07	0.01	0.05	0.86					
50	[AEIKNQRS]	0.01	0.04	0.01	0.63	0.15	0.02	0.12	0.01		
51	[EFKLMNQRTY]	0.01	0.01	0.78	0.01	0.01	0.01	0.01	0.07	0.01	0.08
52	[ADFILQRSTVWY]	0.01	0.01	0.03	0.02	0.03	0.23	0.11	0.03	0.44	0.01
		0.05	0.01								
53	[EHITVWY]	0.01	0.01	0.07	0.08	0.05	0.77	0.01			
54	[HIKMNPQRSTWY](0,1)	0.01	0.02	0.05	0.10	0.01	0.08	0.01	0.38	0.29	0.01
		0.01	0.01								
55	[ADEFGPTV](0,1)	0.09	0.71	0.01	0.07	0.03	0.04	0.02	0.01		
56	[FHMNSTY](0,1)	0.09	0.01	0.01	0.01	0.01	0.09	0.01			
57	[FGHINPQ](0,1)	0.01	0.03	0.77	0.02	0.01	0.13	0.01			
58	[GNRSY]	0.10	0.11	0.75	0.03	0.01					
59	[GHNRSV]	0.79	0.09	0.08	0.02	0.01	0.01				
60	[AEIKMPSTVWY]	0.07	0.01	0.01	0.01	0.01	0.09	0.13	0.59	0.02	0.01
		0.04									
61	[DEIKLMRV]	0.01	0.01	0.64	0.01	0.04	0.01	0.21	0.07		
62	[HILPSTVY]	0.01	0.55	0.09	0.02	0.02	0.26	0.02	0.02		
63	[ADEHKLSTY]	0.04	0.02	0.76	0.01	0.01	0.11	0.01	0.02	0.01	
64	[KMRW]	0.01	0.09	0.89	0.01						
65	G	1.00									
66	C	1.00									
67	[AGSTV]	0.03	0.79	0.04	0.12	0.01					

68	[ALSTY](0,1)	0.03	0.01	0.03	0.08	0.09					
69	[KNST](0,1)	0.01	0.05	0.14	0.03						
70	C	1.00									
71	[HP](0,1)	0.01	0.98								
72	[DHKNPQRST](0,1)	0.01	0.02	0.14	0.05	0.01	0.20	0.02	0.15	0.26	
73	[AEIMPV](0,1)	0.01	0.09	0.01	0.02	0.01	0.74				
74	[KR](0,1)	0.82	0.05								
75	Y(0,1)	0.01									
76	[GHKNPQRS](0,1)	0.08	0.01	0.21	0.03	0.44	0.01	0.02	0.10		
77	[GHNPR](0,1)	0.87	0.01	0.01	0.01	0.01					
78	[DEIKLNPV](0,1)	0.01	0.01	0.64	0.02	0.01	0.04	0.01	0.16		
79	[DEGHIKLNQRTY]	0.02	0.12	0.05	0.04	0.01	0.43	0.01	0.08	0.05	0.09
		0.01	0.08								
80	[ADILPRSV]	0.03	0.01	0.16	0.56	0.07	0.08	0.07	0.02		
81	[AEHIKMNQRSTVY]	0.01	0.18	0.07	0.07	0.04	0.03	0.15	0.01	0.12	0.01
		0.08	0.19	0.04							
82	C(2)	1.00									
84	[ADEGHKQRST]	0.03	0.01	0.09	0.02	0.11	0.14	0.13	0.09	0.19	0.19
85	[KRST]	0.01	0.01	0.22	0.76						
86	[DEN]	0.68	0.13	0.19							
87	[DEKLNR]	0.16	0.25	0.36	0.11	0.02	0.09				
88	[CS]	0.99	0.01								
89	N	1.00									
90	[AEGKLNRY](0,1)	0.01	0.01	0.02	0.16	0.01	0.68	0.07	0.02		
91	P(0,1)	0.01									
92	S(0,1)	0.01									
93	T(0,1)	0.01									
94	P(0,1)	0.01									
95	S(0,1)	0.01									
96	T(0,1)	0.01									

Fig 2. Output from the proposed algorithm: motif for Elapid neurotoxins and corresponding probability scores; obtained from an input of 91 short neurotoxins from Elapid snakes

V. CONCLUSION

The PROSITE database and format is a great tool for bioinformaticians in this regard but the process of manually generating motifs is time consuming. This is where automated algorithms like ours find their place. Although no automatically generated motif can be sufficient for all purposes and a small amount of manual curation may be necessary, our proposed algorithm largely reduces the workload for such manual modifications. The primary advantage our algorithm has over other contemporary automated motif generation processes is the availability of probability scores to rank the protein matches. However, the quality of motifs generated by our algorithm largely depends on the corresponding quality of the sequence alignment tool used in the first place. A manual curation step may be essential even after motifs are generated by algorithms like ours to identify conserved regions and non-conserved regions and tweak the motifs accordingly. All available algorithms for identification of conserved regions have one or other drawbacks, and this is where the manual step comes into play. However, for all other purposes, our algorithm when coupled with MAFFT for sequence alignment, produces high quality motifs that can be used for quick and accurate identification of protein families.

REFERENCES

[1] Bairoch, Amos. "PROSITE: a dictionary of sites and patterns in proteins." Nucleic Acids Research 19.Suppl (1991): 2241.

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

[3] Roytberg, Mikhail A. "A search for common patterns in many sequences." Computer applications in the biosciences: CABIOS 8.1 (1992): 57-64.

[4] Saqi, Mansoor AS, and Michael JE Sternberg. "Identification of sequence motifs from a set of porteins with related function." Protein Engineering 7.2 (1994): 165-171.

[5] Lawrence, Charles E., et al. "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment." science 262.5131 (1993): 208-214.

[6] Koza, John R., and David Andre. "Automatic discovery of protein motifs using genetic programming." Yao, Xin (editor) (1996).

[7] Dufton, M. J., and R. C. Hider. "Conformational properties of the neurotoxins and cytotoxins isolated from elapid snake venoms." CRC critical reviews in biochemistry 14.2 (1983): 113-171.

[8] Katoh, Kazutaka, et al. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." Nucleic acids research 30.14 (2002): 3059-3066.

[9] Katoh, Kazutaka, and Daron M. Standley. "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." Molecular biology and evolution 30.4 (2013): 772-780.

[10] UniProt Consortium. "The universal protein resource (UniProt)." Nucleic acids research 36.suppl 1 (2008): D190-D195.