KNOWLEDGE ENGINEERING: PRINCIPLES AND TECHNIQUES
Proceedings of the International Conference on Knowledge Engineering,
Principles and Techniques, KEPT2007
Cluj-Napoca (Romania), June 6–8, 2007, pp. 130–137

MINING AN ANIMAL TOXIN DATABASE: CHARACTERIZING PROTEIN FOLDS

KENNETH REVETT[(1)], FLORIN GORUNESCU[(2)], AND MARINA GORUNESCU[(3)]

Abstract. The vast majority of animal toxins acts either on sodium or potassium channels. These channels regulate neuronal activity by either inhibiting or activating various neuronal systems. Toxins have been a useful tool for mapping the distribution of various channels types within a variety of organisms. The aim of this paper is to present an automated approach for detecting whether a toxin acts on voltage-sensitive sodium versus potassium channels. In addition, our consensus sequence is also able to reliably determine whether the toxin acts as a gate modifier or pore blocker ($> 93\%$ accuracy). Lastly, we present evidence regarding the existence of two new putative potassium channel gate binding motifs through the use of a learning vector quantization neural network.

1. Introduction

Evolution has endowed animal species with the ability to produce a bewildering array of toxic substances, for both protection and predation. The effects of toxins range from being mildly irritating to lethal on the inflicted victim. Toxins can be broadly classified as either venoms or poisons. Toxins interfere with the normal functioning of specific cell types within the prey organism, although the mechanism(s) of action of several toxins remain to be elucidated [6].

Considering the wide range of toxicities exhibited by animal toxins, it would be very helpful to be able to have a structure function relationship for a given toxin. This task has not yet been accomplished, as there appears to be a small number of structural features (i.e. protein folds) for the vast array of toxins that produce overlapping mechanisms of action (e.g. $\alpha\alpha$ or $\beta\alpha\beta\beta$, [7]). This lack of specificity has made the automated determination of site of action versus structure very difficult. What is still required is a clear consensus sequence that relates structure to site of action. We have attempted to address the issue of generating a consensus sequence, focusing solely on potassium and sodium channels [2]. The challenge has

2000 *Mathematics Subject Classification.* 68P05, 68T05.

Key words and phrases. animal toxins, consensus sequence, learning vector quantization, codebook vector.

been to find a direct association between the structure of the toxin and its specific site of action.

There are several internet-based databases that contain specific information about various toxins ([9], http://www.ncbi.nlm.nih.gov/, http://ca.expasy.org/, http://www.expasy.org/sprot/tox-prot etc.).

In this work, we sought to determine if we could discover one or more consensus sequences for sodium and potassium channels. In addition, we wished to determine if a consensus sequence could be obtained, providing a classification based on whether they acted on the pore or the gate. The results would allow one to determine directly from sequence databases whether the toxin acted on sodium/potassium channels and also whether they would inactivate the channel either by binding to the gate or by blocking the channel pore.

This paper is organized as follows: the next section presents a description of the basic methodology employed, followed by a presentation of the major results, and lastly by a brief conclusion section.

2. Methods

This study entailed the use of several internet based protein structure repositories. The basic outline of our consensus development strategy is as follows:

1. Obtaining the toxins targeting sodium and potassium channels.
2. Refining toxins by extracting the active forms.
3. Separate toxins based on their site of actions.
4. Using various consensus sequence extraction tools sat as PRATT.
5. Comparing the resultant consensus sequences thus obtained by doing a search through the PDB looking to see what the sensitivity and specificity of the resultant hit list was.
6. PRATT consensus builder was employed.
7. Database search using the generated consensus sequences.
8. Repeat from step 2 as necessary.

Below we describe the process in more detail.

2. 1. *Obtaining the toxins targeting Na and K Voltage gate ion channels*: The keywords 'potassium channel inhibitor' were entered into the SRS on the ExPASy server returning a list of peptide toxins targeting Kv channels, returning a list of 155 peptides. Sequences were in FASTA format and output saved into a text document named Master-K.txt. FASTA format was chosen as it is recognized by many types of bioinformatics analysis tools available online [7]. The process was repeated for the Sodium channel toxins using the keywords 'Sodium channel inhibitor', returning 283 toxins.

Refining toxins by extracting active forms: Toxins in ExPASy can contain entire active sequences but sometimes contain Signal peptides and/or Propeptide

sequences. These regions were removed in order to obtain active peptide sequences only, as programs may take those features to generate results. Under the region 'Features' is where details are presented if available, that describes details such as domains and disulphide bonds, e.g. Charybdotoxin b precursor from the scorpion *Leiurus quinquestriatus hebraeus*, SWISSPROT id P59943:

MKILSVLLLA LIICSIVGWS EAQFTDVSCT TSKECWSVCQ RLHNTSIGKC MNKKCRCYS

Key/From/To/Length/Description ⇒ [SIGNAL/1/22/22/"by similarity"]

MKILSVLLLA LIICSIVGWS EA

(leaving the active peptide):

QFTDVSCT TSKECWSVCQ RLHNTSIGKC MNKKCRCYS

These features in ExPASy are not all experimentally derived and have been assigned 3 types of comments [7]: (a) Potential, (b) Probable and (c) By similarity.

'Potential' -there is some logical evidence that given annotation could apply. This non-experimental qualifier is often used to present the results from protein sequence analysis tools if the results make sense in respect to a given protein [1], [2]. 'Probable' -is a stronger indicator than 'Potential' and is based on some experimental evidence, that the given information is expected to be found in the natural environment the protein [3], [4].'By similarity' -facts proven for the protein or part of it, and then transferred to other protein family members within a certain taxonomic range. Sites within conserved domains to each other include active sites and disulfide bonds [5]. The Master files contain active toxin peptides, were 'saved as' Primary files: Prim-Na.txt and Prim-K.txt files. Master files remained untouched since downloading.

Separating toxins by site of action: The toxins in the Primary files were separated by site of action, sites where determined through literature reading from associated links of ExPASy. The resulting data was stored on disc for further analysis (see below).

Prim-K.txt: K-pores.txt & K-gaters.txt
MasterNa.txt: Napores.txt & Na-gaters.txt

K-files: SWISSPROT ID's for each toxin in the Prim-K.txt file was entered into ExPASy. The resulting page has links to related (published) literature that was reviewed to determine the toxin as a pore blocker or a gate modifier. Once determined, the toxin entry was cut and pasted into the respective file.

Na files: The process for Kv toxins was repeated, however due to the structural differences between the channels, Nav was found to have 6 binding sites, 5 of which are associated with the gate (sites 2-6) and site 1 the pore.

Veratridine, Batrachotoxin and Grayanotoxin are toxins acting on site 2, but are not peptides. They were not present on the list of the ExPASy output for Sodium channel inhibitors. Brevetoxin and Ciguatoxin acting on site 5 again are

not peptide toxins that were returned on the output for Sodium channel inhibitors, therefore these toxins will be excluded.

PRATT Consensus builder: PRATT (http://www.ebi.ac.uk/pratt/) is part of EBI (http://www.ebi.ac.uk) and generates consensus sequence motifs from unaligned fasta files. ExPASy has many other databases and tools for analysis of proteins and PROSITE will be initially used to search against the generated consensus. PRATT search input parameters can be modified to the user preferences, i.e. consensus must match at least 50% of inputted sequences. The output of PRATT is in PROSITE format and feed directly into the PROSITE database search [9].

2.2. *Database search*: PROSITE is a database that can be searched when a Motif is entered as a parameter (http://www.ExPASy.org/prosite/). The website is able to understand patterns and motifs by using the accepted one letter abbreviations of the amino acids (e.g. G is Glycine), i.e. the usual format of a typical pattern/motif can be something like:

G-[ASP]-V-X(2)-GLA-SP,

where letters correspond to their aminoacids, '-' separates the next aminoacid, X calls for any aminoacid (the number in the brackets determines how many of the preceding letter), {∗} -aminoacids within curly braces are not to be included in the pattern/Motif search.

Example of interpretation: "Starting with a Glycine, followed by Alanine, Serine or Proline, followed by a Valine, then by any 2 aminoacids, followed the aminoacid order Glycine, Leucine and Alanine and finally ending on any aminoacid barring Serine or Proline".

3. Results

Potassium - GATE: The gate (voltage sensor) of the Kv channel currently has two known folds targeting it. These folds are: $\beta\beta\beta$, $3_{10}\beta\beta$ and 2 non-categorised folds $\beta\beta$ and $\beta\beta\beta\beta$.

The consensus that was applied to the structures was:

C-X(3)-[WMILF]-X(9,10)-C-X(0,3)-[REKH]-X(1,5)-C-X(3,10)-C

This consensus finds 19 out of the 20 toxins that have been classed to target the potassium gate. The only toxin that does not contain this consensus is Kappa-conotoxin BtX from the Cone snail.

3.1. **POTASSIUM-PORE**: Starting with the potassium channel pore consensus,

C-X(9,12)-C-X(2,5)-C,

Folds include: $\beta\beta\beta$, $\alpha\alpha$ (hairpin), $\alpha\alpha$ (helical cross), $3_{10}\alpha\alpha$, $\alpha\beta$, $\beta\alpha\beta\beta$ and $3_{10}\beta\beta\alpha$

$\beta\beta\beta$ PRATT ouput: CXK7A-CONPU∗: 8- 26: Cfqhlddccsrk-CnrfnkC.

3.2. **SODIUM-PORE**: Site 1 is the pore of the sodium channel. The folds that recognize the Nav Pore are: $\beta\beta$, $\beta\beta\beta$ and $\beta\alpha\beta\beta$. The Pore consensus will be applied to a structure that represents each fold.

C-x(3,6)-C-x(4,6)-C-[ACGN].

Matches 15 out of 16 known Nav pore blocking toxins.
Example displayed: $\beta\beta$-Conotoxin GS. CXGS-CONGE: 9-20: grgsr Cppqc–Cmglr–CGrgnpq

3.3. **SODIUM-GATE**: Site 3 best consensus for C-X(6,9)-C-X(0,6)-C
This consensus matches some toxin sequences six times, Neurotoxin Tx2-6 from the Brazilian armed spider. As a result alignment B will be chosen as it matches some toxin sequences only 3 times.

C-x(5,7)-C-x(10,13)-C

All folds except $\beta\beta\alpha\beta\beta\alpha$ target site 3 of the sodium channel. Each fold type will be shown $\beta\beta$ Robustoxin (funnel web spider)

TXDT1-ATRRO: 1-20: Cakkrnw-Cgknedcccpmk–C iyawy
TXDT1-ATRRO:14-31: gkned Cccpmk–Ciyawynqqgs—Cqttit

Finally, a consensus was determined for all potassium and sodium channel toxins. This produced an interesting output, in that the individual consensus sequences seem to be the reverse of each other.
Potassium toxins C-X(9,12)-C-X(2,6)-C,
Sodium toxins C-X(3,6)-C-X(5,9)-C.
The potassium toxin consensus has a long spacer region between the first and second Cystiene and the sodium has a longer spacer region between the second and third Cystiene. The predominant type of channel toxin for the potassium channel seems to be Pore specific and for the sodium channel the gate conversely. The relationships between these observations are not clear but may help to understand the differences between the two types of ion channel toxins.

In our previous work [2], we have noticed that there were 2 unreported potassium channel gate modifiers -with the $\beta\beta$ and $\beta\beta\beta\beta$ folds. It is well known that the $\beta\beta$ fold is found in conotoxins, but these act on sodium channels as pore blockers, not gate modifiers. Our initial study of these toxins suggested that these 2 motifs are also able to bind to potassium channels as well. Table 1 summarizes statistics of the overall consensus sequences for each of the categories.

Table 1. The summary statistic

Target	Number	Consensus Sequence	% Found
K Pore	127	C-X(9,12)-C-X(2,5)-C	93.7%
K Gate	20	C-X(3)-[WMILF]-X(9,10)-C-X(0,3)-[REKH]-X(1,5)-C-X(3,10)-C	95.0%
Na Pore	16	C-X(3,6)-C-X(4,6)-C-[ACGN]	93.75%
Na Gate	247	C-X(3,6)-C-X(9,11)-C	93.9%
All K	145	C-X(11,14)-C-X(2,5)-C	91.72%
All Na	262	C-X(3,6)-C-X(10,13)-C	93.13%

This result was based on motif matching from our selected database extracted from various toxin repositories. We sought to strengthen the case for these two new putative potassium gate modifiers in this study.

To provide additional support for this hypothesis, we generated a classifier based on a self-organized type of neural network -the *learning vector quantization* (LVQ2.1). LVQ is a supervised version of vector quantization, similar to Self-Organizing-Maps (SOM). LVQ can be understood as a special case of an artificial neural network, applying a winner-take-all Hebbian learning-based approach. It can be applied to pattern recognition, multi-class classification and data compression tasks. LVQ algorithms directly define class boundaries based on prototypes, a nearest-neighbour rule and a winner-takes-it-all paradigm. The main idea is to cover the input space of samples with "codebook vectors" (CV), each representing a region labelled with a class. A CV can be seen as a prototype of a class member, localized in the centre of a class or decision region ("Voronoi cell") in the input space. As a result, the space is partitioned by a "Voronoi net" of hyperplanes perpendicular to the linking line of two CVs (mid-planes of the lines forming the "Delaunay net"). A class can be represented by an arbitrarily number of CVs, but one CV represents one class only. In terms of neural networks a LVQ is a feedforward net with one hidden layer with Kohonen neurons, adjustable weights between input and hidden layer and a winner takes it all mechanism. The basic LVQ algorithm -LVQ1- tends to push CVs away from Bayes decision surfaces. In order to get better approximations of the Bayes rule by pairwise adjustments of two CVs belonging to adjacent classes, in LVQ2 (LVQ2.1 too) adaptation only occurs in regions with cases of misclassification in order to get finer and better class boundaries. To conclude, a main advantage of using LVQ is that it creates prototypes that are easy to interpret for experts in the field.

The code vectors were selected randomly from the set of sequences that were found during our motif search (see Table 1 for details). After seeding the code book vectors by randomly selecting members from of each of the classes (pore blockers - sodium/potassium and gate modifiers - sodium/potassium), the rest of the candidates from the consensus search were presented to the network. If the putative consensus were indeed from a different class then the others, they would tend to aggregate in the vicinity of their respective code-book vectors. A critical

feature in this type of neural network is a distance metric. We encoded the motifs according to aminoacid letter and number of residues. We therefore presented the sequences according to the motifs presented in Table 1. An X(3,6) for instance is matched with the letter 'X' and the numbers 3-6 in the form of a set of strings XXX, XXXX, XXXXX, XXXXX. Then we used the ordinal position within the alphabet as the distance metric to calculate which code book vector a particular consensus sequence was closest too. We performed a 10-fold cross validation, partitioning the dataset into disjoint sets of 10 randomly selected elements until all of the consensus sequences were used as code book vectors. The results indicate that the two new sequences did map to the potassium gate modifier, indicating that they were closest to this code book vector than to any other code book vector.

4. Conclusions

In this study we developed a set of consensus sequences that could differentiate potassium from sodium channel acting toxins. Our methodology was able to generate consensus sequences for potassium gate and pore blockers (2 and 3 respectively) as well as sodium channel gate and pore blockers (4 and 5 respectively). In addition, we were able to come up with a consensus sequence that was able to generically differentiate potassium from sodium channel toxins (6 and 7 respectively). No literature report has provided a consensus sequence that could distinguish pore from gate blockers for either sodium or potassium channels and this is the novel result from this study. When the consensus sequences we have generated are entered into a standard protein sequence database such as PDB, we find that the accuracy, in terms of number of hits per known number of toxins acting on that particular site, is quite high -on the order of 93% or higher in this study. It should be noted however that consensus generation has been optimal when sequence numbers are low or toxins are from a related family. The toxins that act on the channels in this project are from a wide range of organisms.

This approach provides additional evidence that there are two more gate modifier motifs: $\beta\beta$ and $\beta\beta\beta\beta$. The $\beta\beta$ motif is similar to a subset of toxins that are sodium channel pore blockers, yet there is no evidence reported in the literature that indicates they bind the gate (voltage sensor) of potassium channels to our knowledge. This is still a preliminary result and further work must be performed to strengthen the case for this argument. If it holds true, then there will be a total of 10 potassium channel motifs, although one may cross-react with one or more sodium channel pore(s) and potassium gate(s).

References

[1] Boeckmann B., Bairoch A., Apweiler R., Blatter M.C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S. and Schneider

M., 2003. The Swiss-Prot protein knowledgebase and its supplement TrEMBL, Nucleic Acids Res. 31, pp. 365-370.

[2] Bhogal S., Revett, K., 2005, Animal Toxins: What Features Differentiate Pore Blockers From Gate Modifiers, CIMA 2005 Conference, Istanbul Turkey.

[3] Bucher P., Bairoch A., 1994, A generalized profile syntax for biomolecular sequences motifs and its function in automatic sequence interpretation, Proceedings 2nd International Conference on Intelligent Systems for Molecular Biology.

[4] Bucher P., 2002, PROSITE: a documented database using patterns and profiles as motif descriptors, Brief Bioinform, 3, pp. 265-274.

[5] Elsworth J., Deadly by Nature, Web Project, http://www.chm.bris.ac.uk/webprojects 2002/elsworth/deadly-by-nature.htm

[6] Jungo F., Bairoch A., 2994, Tox-Prot, the toxin protein annotation program of the Swiss-Prot protein Knowledgebase, Toxicon., 45, pp. 293-301.

[7] Miller C., Naranjo D., 1996, A strongly interacting pair of residues on the contact surface Charybdotoxin and a shaker channel, Neuron, 16, pp. 123-130.

[8] Possani L.D., Merino E., Corona M., Bolivar F. and Becerril B., 2000, Peptides and genes coding for scorpion toxins that effect ion-channels, Biochemie., 82, pp. 861-868.

[9] Schonbach C., Kowalski-Saunders and Brusic V., 2000, Data warehousing in molecular biology, Brief Bioinformatics., 1, pp. 190-198.

(3) University of Westminster, Harrow School of Computer Science, London, UK
E-mail address: revettk@westminster.ac.uk

(1) Universitatea de Medicina si Farmacie din Craiova, str. Petru Rares, Nr. 2-4
E-mail address: fgorun@rdslink.ro

(2) Universitatea din Craiova, Str. A.I. Cuza, Nr. 13
E-mail address: mgorun@inf.ucv.ro