This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: http://orca.cf.ac.uk/95739/

This is the author’s version of a work that was submitted to / accepted for publication.

Citation for final published version:


Publishers page: http://dx.doi.org/10.1016/j.jip.2016.07.015
<http://dx.doi.org/10.1016/j.jip.2016.07.015>

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher’s version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.
Structural classification of insecticidal proteins – towards an *in silico* characterization of novel toxins.

Colin Berry¹* and Neil Crickmore²

¹Cardiff School of Biosciences, Cardiff University, Park Place, Cardiff CF10 3AT, UK
²School of Life Sciences, University of Sussex, Falmer, Brighton, BN1 9QG, UK

*Corresponding author: Colin Berry, Cardiff School of Biosciences, Cardiff University, Park Place, Cardiff CF10 3AT, UK, Berry@cf.ac.uk

**Abstract:**
The increasing rate of discovery of new toxins with potential for the control of invertebrate pests through next generation sequencing, presents challenges for the identification of the best candidates for further development. A consideration of structural similarities between the different toxins suggest that they may be functionally less diverse than their low sequence similarities might predict. This is encouraging from the prospective of being able to use computational tools to predict toxin targets from their sequences, however more structure/function data are still required to reliably inform such predictions.

**Introduction:**
The insecticidal toxins of *Bacillus thuringiensis*, *Lysinibacillus sphaericus*, *Photobacterium fischeri* and other bacteria represent a rich resource for the control of pest insects. The increasing rate of discovery of new toxins, driven by next-generation sequencing, will expand our arsenal of potential biocontrol agents but this, in itself, presents new challenges. Even with past rates of toxin discovery, toxins have rarely been tested against more than a few species of insects (van Frankenhuyzen, 2009) and, in the future, toxicity testing of large numbers of new toxins
against a wide range of insects will not be feasible. To facilitate the selection of toxins for study, different criteria may be applied, including identification of the toxin in a strain known (from a previous screening) to have interesting biocidal activity or relatedness to known toxins. Here we consider the prospects for a further, selective method through the prediction of activity. We highlight some of the challenges that may be encountered and propose steps that will bring us closer to this goal. Useful predictions would not only assist in the selection of toxins for development but would also have value in support of the regulatory process of biopesticide product registration, where the potential to predict off-target activities would be valuable.

The *B. thuringiensis* nomenclature system (Crickmore et al., 1998) currently contains several hundred individual sequences, divided between 74 classes of Cry toxin, 3 classes of Cyt toxin, 4 classes of Vip toxin and one SIP toxin. *L. sphaericus* strains may produce the BinA/B toxin, Mtx1, Mtx2, Mtx3, Mtx4, sphaericolysin, Cry48 and Cry49 (reviewed in (Berry, 2012)) and *Photorhabdus* strains can produce Tc toxins, PirA/B and Mcf toxins (ffrench-Constant et al., 2007). This represents a great diversity of toxins but some simplification can be achieved by considering these proteins in terms of their structural characteristics (known or predicted). Table 1 shows the toxin classes, colour-coded by sequence homology groups. As can be seen, the 3-domain Cry toxins represent the largest structural family (and also encompass the PirA/B toxin, recently shown to be equivalent to a 3-domain toxin with a dissociated domain III (Lee et al., 2015)). There is also a large group of toxins that is rich in beta-sheets with general structural similarity to aerolysin. This group includes Cry46 and toxins identified by Pfam (Bateman et al., 1999) to be members of either the Etx/Mtx2 family or the Toxin_10 family. Other groups include the Cyt toxins, the ADP-ribosyl transferase toxins Mtx1 and Vip1/2 (along with the Vip1-like Vip4 protein). Cry34 is an aegerolysin like protein and with Cry35 is part of a two-component toxin (Kelker et al., 2014). Cry37, which itself is part
of a two-component toxin with Cry23, which shows structural homology with Cry34 (Rydel et al., 2001). Other toxins, which appear unrelated and have no published structures, are Cry6, Cry22, Cry55, Vip3 and Mcf. Our knowledge of the structure and function of toxins within these groups varies and it will be useful to consider the major groups separately.

**The 3-domain toxins:** These toxins are the best-characterised, with the first structure published in 1991 (Li et al., 1991) and with several decades of studies on the specificity and mode of action of members of this family. The steps leading to toxicity for this family are well-known and involve ingestion by the invertebrate target, solubilisation of toxin crystals in the gut, proteolytic activation by gut enzymes, one or more receptor binding step, followed by membrane insertion (Pardo-Lopez et al., 2012). Insect specificity could be mediated by any of the above steps, for example changes in proteinase activity (Loseva et al., 2002) but the most important determinants of specificity are the binding to and specificity for receptors on the surfaces of target cells.

As suggested by the name of this family, the structure of the active toxin is composed of 3 distinct structural domains. Domain I is formed from a bundle of alpha helices and is involved in pore formation by the toxin. Domain II has a beta prism structure that appears to be related to carbohydrate binding proteins and Domain III has a beta sandwich fold.

Domains II and III appear to have roles in receptor binding and specificity of the toxins as demonstrated by domain swapping experiments that have altered target specificity (Lee et al., 1995; Pigott and Ellar, 2007). Bioinformatic analysis suggests that the 3 toxin domains evolve at different rates (Bravo, 1997) and this may have implications for target specificity.

Within the 3-domain toxin family, we find toxins with activity against insects in several orders, principally amongst the Lepidoptera with fewer active against the orders Diptera and Coleoptera, and with small numbers active against Hymenoptera and Hemiptera as well as toxins affecting nematodes and gastropods (reviewed recently (Palma et al., 2014a)).
Members of this family active against human cancer cells have also been reported (Ohba et al., 2009), although it is clearly unlikely that they have co-evolved with this host. However, correlation between sequence identity and target range is generally poor even when analysis is carried out at the level of the individual domains (de Maagd et al., 2001). This highlights the need for analysis at a level below that of the domains themselves. Within domain II, several exposed loops (the $\alpha$8 loop, and loops 1, 2 and 3) have been identified as potentially important for receptor binding. The variability of these regions and their dispositions in 3 example toxins is shown in Figure 1. These surface loops are amongst the most variable in sequence and in length between individual toxins and even minor modifications have been shown to change targeting (Bravo et al., 2013). For example, Cry4Ba has no significant activity against Culex mosquitoes but the substitution of Asp454 in domain II loop 3 with the sequence Pro-Ala-Thr results in high toxicity to Culex species without reduction in toxicity towards Aedes aegypti (Abdullah et al., 2003). However, these loops may not be the sole domain II mediators of specificity, for example it has been shown that residues remote from these loops (illustrated in Figure 2) contribute to dipteran/lepidopteran specificity in the Cry2Aa/Cry2Ab toxins (Morse et al., 2001).

An understanding of the receptor-binding interactions of the 3-domain proteins is made complex by the diversity of putative receptor proteins for these toxins. The most commonly identified binding partners include cadherins, aminopeptidases and alkaline phosphatases (Pigott and Ellar, 2007). However, even for single toxins such as Cry1Ac and Cry4Ba, a large number of interacting proteins can be identified through proteomic studies (Bayyareddy et al., 2009; Krishnamoorthy et al., 2007) (Table 2). Although the significance of these binding interactions for toxicity is not known, we cannot discount possible physiological significance for these interactions, for example prohibitin, identified as a potential Cry4Ba binding-protein has been demonstrated to bind toxin in Aedes cells in culture (Kuadkitkan et al., 2012). Glycolipids are also potential receptors for 3-domain toxin binding as shown for
Cry1Ac (Garczynski and Adang, 2000), Cry2Ab (Ma et al., 2012) and may be particularly important for the nematode-active Cry5B (Barrows et al., 2007; Griffitts et al., 2005). The diversity of possible receptors and the probable involvement of more than one receptor in toxicity, adds to the challenge of predicting toxin activity. Several studies have attempted to map regions of the toxins that may interact with receptor proteins. For example, residue Tyr445 in loop 3 of Cry1Aa domain II was identified as being important in binding to the cadherin BtR175 of *Bombyx mori* (Atsumi et al., 2005) while Val582 in domain III was shown to be important for interaction with aminopeptidase N (Atsumi et al., 2008). In parallel, there have been studies to map regions of receptors that may interact with the toxins (Pigott and Ellar, 2007) but even when a common class of receptor (eg a cadherin) is considered, there is little correspondence between regions interacting with different toxins.

**The beta sheet toxins:** There are several classes of toxins in Table 1 that appear to be rich in beta sheets. These include Cry37 (de Maagd et al., 2003) and the aegerolysin-like Cry34 (Kelker et al., 2014) which both form part of binary toxins (with Cry23 and Cry35 respectively, discussed below). The presence of extended beta sheets also characterises the sphaericolysin/anthrolysin family of toxins that is highly conserved across isolates from *L. sphaericus, B. thuringiensis, Bacillus cereus* and *Paenibacillus alvei* (Berry, 2012; Bourdeau et al., 2009) and appear to act as cholesterol dependent cytolysins (From et al., 2008; Nishiwaki et al., 2007). At present, the literature lacks sufficient information to allow structure/function predictions for these toxins. However, a number of proteins in Table 1 belonging to the Etx/Mtx2 family, the Toxin_10 family, along with Cry46 are rich in beta sheets and show a general fold similar to aerolysin. This group of toxins is the most numerous after the 3-domain toxins and some analysis of structure and specificity is possible. The general structure of these toxins features a head region and a tail region that features long beta
strands (available structures for these proteins are shown in Figure 3). Cry45 and Cry46 are 
produced by *B. thuringiensis* but have no known activity against invertebrates. However, 
activity against mammalian cancer cells has been demonstrated (Ohba et al., 2009). 
Structurally, Cry46 most closely resembles the Etx/Mtx2 family (although Pfam analysis of its 
primary sequence does not assign it to this family). The Etx/Mtx2 family differs from the 
Toxin_10 family in two clear features. In the Toxin_10 family, the head domains contain beta 
trefoil motifs similar to carbohydrate-binding domains and these heads are formed 
exclusively from the N-terminal end of the proteins. The beta trefoil domains may have a role 
in toxin interactions with glycoproteins or glycolipids to facilitate receptor binding or other 
stages of the mechanisms of action of the toxins. In contrast, in the Ext/Mtx2 family, the head 
lacks the beta trefoil and is composed of residues from the N-terminal region and from a 
further stretch of amino acids much closer to the C-terminus (before the C-terminal sequence 
completes the last long beta strand of the tail domain). In these features, Toxin_10 proteins 
resemble toxins such as the haemolytic lectin from the parasitic mushroom *Laetiporus sulphureus* (Mancheno et al., 2005), while the Etx/Mtx2 family resembles mammalian toxins 
such as aerolysin (Figure 3) and epsilon toxin from *Clostridium perfringens*.

When the ability of these aerolysin-like toxins to target invertebrate or mammalian 
cells is compared to their overall sequence, there is no obvious correlation (Figure 4A). When 
analysis is performed at the level of the phylogeny of individual head and tail domains 
(Figure 4B and C) there is still no clear correlation, indicating once again the need for deeper, 
subdomain analysis to be undertaken to predict activity.

Understanding the receptor binding and specificity of the Toxin_10 family is further 
complicated by the existence of partner proteins for many of these toxins. Cry36 is a clear 
exception since it is reported to act alone to kill *Diabrotica* larvae (Rupar et al., 2000). 
Another protein in the Toxin_10 family (41.9 kDa protein) may be encoded by *B. thuringiensis*
but no toxicity has been discovered to date and no partner protein has been identified (Palma
All of the other Toxin_10 proteins act with their specific partner proteins to form binary toxins as follows: BinA and BinB (both Toxin_10 proteins)(Broadwell et al., 1990; Oei et al., 1990); Cry34 (aegerolysin-like) and Cry35 (Toxin_10)(Kelker et al., 2014; Masson et al., 2004); Cry48 (3-domain) and Cry49 (Jones et al., 2007). The role of each protein in these binary pairs is clearly significant to understanding the specificity of the toxins and potential binding of both components presents further challenges to prediction. The binding of both Cry48 and Cry49 to Culex brush border membrane fractions has been shown, with Cry49 suggested to be the principal binding component (Guo et al., 2016). In the case of the Bin toxin, BinB appears to be the major receptor-binding component in Culex mosquitoes but in Anopheles BinA also appears to be able to bind (Charles et al., 1997; Oei et al., 1992). Binding of BinB to target membranes appears to be mediated by residues at its N-terminal end (Oei et al., 1992; Romao et al., 2011; Singkhamanan et al., 2010), consistent with receptor recognition via the head domain (Srisucharitpanit et al., 2014). The BinA/BinB toxin appears to bind to a single toxin receptor, a GPI anchored α-glycosidase (Silva-Filha et al., 1999), which may simplify the investigation of binding interactions (particularly when compared to the complex receptor binding of 3-domain toxins). Both the receptor, Cqm1, from the Bin-sensitive Culex quinquefasciatus and the ortholog from the insensitive Aedes aegypti are known and a region involved in binding has been identified. This includes a Gly-Gly motif, potentially on a surface loop, which may be required for productive interaction (Ferreira et al., 2014). If binding to single receptors proves to be a general feature of this class of toxin, prediction of activity may be made more straightforward but this may also have implications for the ease with which insects may acquire resistance. A number of the Etx/Mtx2 family of Cry toxins are also reported to require a binary partner for full activity. Cry23Aa (also known as ET33) acts with Cry37 (ET34) and has activity against various coleopteran insects (Donovan et al., 2000; Ekobu et al., 2010). Cry15Aa and Cry33Aa form binary partnerships with two other proteins (40kDa and NT32KD respectively) that, due to their lack of individual activity, have not been
assigned Cry names, and share no obvious sequence similarity with any other characterized protein (Kim et al., 2003; Naimov et al., 2011).

**Convergent evolution of the beta sheet toxins?:**

Figure 3 indicates that three of the homology groups highlighted in Table 1 (Toxin_10, Cry46 and Etx/Mtx2) share significant sequence similarity and indeed are very similar to toxins such as aerolysin. In the aerolysin family of proteins, the conserved beta sheet dominated “aerolysin fold” is believed to adopt a barrel conformation within a membrane with the associated domains primarily having a binding role (Szczesny et al., 2011). It is conceivable that despite the diversity in primary sequence, many of the non 3-domain Cry toxins could share significant structural/functional homology. This would take the form of the aerolysin tail fold associated with a head domain involved in targeting the protein to a particular receptor. This head domain could be part of the same protein as the aerolysin fold or come from an associated binary partner. A number of Cry and Cyt toxins have acquired a ricin-like beta-trefoil carbohydrate-binding domain and while it is tempting to speculate that this could have given these proteins novel binding activities, there is currently no evidence that this is the case (reviewed by (Adang et al., 2014)).

**The future for in silico analyses:**

The ultimate objective of such work is to be able to predict the likely specificity of a toxin from primary sequence data. While the computational power to be able to achieve such a goal is available, the underlying data required to inform such analyses are still lacking. Significant progress has been made in recent years in the elucidation of new toxin structures and the potential to derive reliable structures from modelling approaches is discussed elsewhere in this issue (Berry and Board, 2016). Similarly, while there are a lot of data available on the target specificity of individual toxins (Palma et al., 2014a; van Frankenhuyzen, 2009), many of
these are contradictory, thus compounding efforts to derive meaningful associations. The area where least information is known concerns the interaction of the toxin with the target cell (Vachon et al., 2012). As discussed above, even when putative receptors are identified, determining which interactions are crucial for toxicity is far from straightforward. There is a need to elucidate more structures for confirmed receptors, and ideally toxin-receptor complexes, which can then lead to in silico predictions of toxin-receptor interactions. A number of studies have used docking analysis to indicate the likely interaction between a Cry toxin and its putative receptor (Ahmad et al., 2015; Tajne et al., 2012; Zhao et al., 2012). Without a detailed understanding of which interactions are crucial for toxicity, such studies are likely to throw up many false positives, indeed one report predicts that Cry1Ac could have activity against cattle (Ebenezer et al., 2013), a prediction that is not supported by experimental observations.

In summary, whilst our understanding of toxin structure is rapidly progressing, and might suggest that the large family of toxins is less diverse than was thought, we are still a long way from the goal of being able to match toxins and hosts based on primary sequence data generated from genome sequencing.
Table 1: Toxins and their homology groups

Toxins of *B. thuringiensis*, *L. sphaericus* and *Photorhabdus* spp. are shown with colouring to indicate homology groups: light blue = 3-domain toxins; peach = Etx/Mtx2 toxins; pink = Toxin_10 family proteins; violet = Cyt toxins; khaki = aegerolysin toxins; grey = ADP ribosyl transferase-related proteins; toxins not falling into these groups are coloured differently.

<table>
<thead>
<tr>
<th>Cry1</th>
<th>Cry21</th>
<th>Cry41</th>
<th>Cry61</th>
<th>Vip3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cry2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry5</td>
<td>Cry25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry6</td>
<td>Cry26</td>
<td>Cry46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry9</td>
<td>Cry29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cry20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cry10 family proteins: Sphaericolysin, PirA, PirB, Mcf, Cyt1, Cyt2, Cyt3, Sip, Vip1, Vip2.
Table 2: Potential receptors for 3-domain toxins.
Proteins discovered through proteomics as potential binding partners for two 3-domain toxins, Cry1Ac and Cry4Ba are shown.

<table>
<thead>
<tr>
<th>Cry4Ba (Bayyareddy et al., 2009)</th>
<th>Cry1Ac (Krishnamoorthy et al., 2007)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cadherin</td>
<td>Cadherin</td>
</tr>
<tr>
<td>Alkaline phosphatases (3)</td>
<td>Alkaline phosphatases</td>
</tr>
<tr>
<td>ATPase</td>
<td>ATPases</td>
</tr>
<tr>
<td>Actin</td>
<td>Actin</td>
</tr>
<tr>
<td>Serine and metallo peptidases</td>
<td>Aminopeptidases</td>
</tr>
<tr>
<td>Prohibitin</td>
<td>Desmocollin-like protein</td>
</tr>
<tr>
<td>Mitoporin</td>
<td></td>
</tr>
<tr>
<td>Flotillin-1</td>
<td></td>
</tr>
<tr>
<td>ATP synthase</td>
<td></td>
</tr>
</tbody>
</table>

Figure Legends

Figure 1: Domain II loops
The structures of Cry1Aa, Cry2Aa and Cry3Aa (PDB codes 1CIY, 1I5P and 1DLC respectively) are shown with domains I and III shown as yellow and brown ribbons while domain III is shown as a surface representation with the exposed regions of the $\alpha 8$ loop, and loops 1, 2 and 3 coloured cyan, orange, green and magenta, respectively.
Figure 2: Residues implicated in Cry2A specificity

The Cry2Aa structure (PDB 1I5P) is shown with domain II in grey. Residues shown to contribute to lepidopteran specificity are coloured blue while those involved in dipteran specificity are shown in magenta.

Figure 3: Structures of beta sheet toxins

The structures of the insecticidal toxins BinB (PDB 3WA1 (Srisucharitpanit et al., 2014)), Cry35 (PDB 4JP0 (Kelker et al., 2014)), Cry23 (PDB 4RHZ), Cry45 (PDB 2D42 (Akiba et al., 2006)), Cry46 (PDB 2ZTB (Akiba et al., 2009)) and Cry51 (PDB 4PKM (Xu et al., 2015)) are shown along with the structures of the haemolytic toxin from L. sulphureus (Lsulph: PDB 1W3A (Mancheno et al., 2005) and proaerolysin (PDB 1PRE (Parker et al., 1994)). Head regions are coloured cyan, tails blue and the extra domain in proaerolysin in green.

Figure 4: Phylogenetic relationship of insecticidal and mammalian-active beta sheet toxins

The amino acid sequences of the toxins were compared using MEGA6 and phylogenetic trees were built using the maximum likelihood algorithm. For those toxins where the head and tail domains are discontinuous, the separate regions encoding each domain were combined and analysed as a contiguous sequence. Those toxins with known activity against mammals are highlighted in bold red.

References:


Ebenezer, K. S., Nachimuthu, R., Thiagarajan, P., Velu, R. K., 2013. in silico identification of cross affinity towards Cry1Ac pesticidal protein with receptor enzyme in *Bos taurus* and sequence, structure analysis of crystal proteins for stability. Bioinformation. 9, 792-795.


Lee, M. K., Young, B. A., Dean, D. H., 1995. Domain III exchanges of *Bacillus thuringiensis* Cry1A toxins affect binding to different gypsy moth midgut receptors. Biochem Biophys Res Commun. 216, 306-312.


Ma, G., Rahman, M. M., Grant, W., Schmidt, O., Asgari, S., 2012. Insect tolerance to the crystal toxins Cry1Ac and Cry2Ab is mediated by the binding of monomeric toxin to lipophorin glycolipids causing oligomerization and sequestration reactions. Developmental and Comparative Immunology. 37, 184-192.


Oei, C., Hindley, J., Berry, C., 1990. An analysis of the genes encoding the 51.4- and 41.9-kDa toxins of Bacillus sphaericus 2297 by deletion mutagenesis: the construction of fusion proteins. FEMS Microbiol. Lett. 72, 265-274.


Palma, L., Muñoz, D., Berry, C., Murillo, J., Caballero, P., 2014b. Draft genome sequences of two Bacillus thuringiensis strains and characterization of a putative 41.9-kDa insecticidal toxin. Toxins. 6, 1490-1504.


Figure 3
Figure 4

A  Overall

Cry15Aa1
ETX
Cry33Aa1
Aerolysin
Cry64Aa1
Cry51Aa1
Cry60Aa1
Cry23Aa1
Cry45Aa1
Cry74Aa1
Cry46Aa1
Hydralysin
Cry38Aa1

B  Head

Cry51Aa1
Cry60Aa1
Cry23Aa1
Cry64Aa1
Cry33Aa1
Cry15Aa1
Cry38Aa1
Cry45Aa1

C  Tail

Cry45Aa1
Cry49Aa1
Cry36Aa1
Cry38Aa1
Cry74Aa1
Cry46Aa1

Toxin_10 proteins

0.5

C. perfringens enterotoxin
L. sulphureus lectin
Cry35Ab1
BinB
Cry36Aa1
Cry49Aa1

0.5