Search for Cry proteins expressed by *Bacillus* spp. genomes, using hidden Markov model profiles

J. Francisco Castillo-Esparza¹ · Ismael Hernández-González¹ · Jorge E. Ibarra¹

Received: 5 November 2018 / Accepted: 14 December 2018 / Published online: 2 January 2019
© King Abdulaziz City for Science and Technology 2019

Abstract

This report focuses on a systematic search for Cry proteins in *Bacillus* spp. other than *B. thuringiensis* by analyzing reported *Bacillus* spp. genomes, using conserved sequences from the C-terminal half of reported Cry proteins in hidden Markov model profiles. A high-throughput model based on the use of HMMER and CD-HIT tools was designed, which identified Cry proteins. This model was used on 857 reported *Bacillus* spp. genomes, where 174 Cry protein sequences were identified, mostly, as expected, in *B. thuringiensis* genomes but, interestingly, 42 were identified on other species. Despite including 89 species of *Bacillus* in the HMMER analysis, Cry protein sequences were found only in genomes from species within the *B. cereus* group. According to the species registered at the NCBI database containing each genome, this group was formed by 18 non-*B. thuringiensis* strains. However, when sequences in those genomes were analyzed by multilocus sequence typing, the number of non-*B. thuringiensis* strains increased to 39, indicating that as many as 119 Cry protein sequences were found in four non-*B. thuringiensis* species. Therefore, dispersion of Cry proteins is much wider and frequent than previously thought, questioning its role in nature.

Keywords Cry proteins · Thuringiensis · Markov profiles · Bacillus · Genomes

Introduction

Bacillus thuringiensis is a Gram-positive, sporogenic bacterium, biotechnologically developed as bioinsecticide due to its potent δ -endotoxins with activity against a number of insect pests important in agriculture and health programs (Palma et al. 2014). The proteinaceous δ -endotoxins form parasporal bodies, typically called “crystals”, simultaneously to the spore, and show insecticidal activity against many lepidopteran, dipteran, and coleopteran pests (Peng et al. 2015). The δ -endotoxins can be classified within two families: Cry and Cyt proteins, being the former the most diverse type, with more than 800 reported sequences to date, most of them of so-called “three-domain” Cry toxins (Crickmore

et al. 2017). Their mode of action has been widely studied, mostly on lepidopteran species. Once activated in the insect’s midgut, Cry proteins are able to form pores in gut cells, which causes osmotic instability and cytolysis, leading to a general septicemia of the insect (De Maagd et al. 2001; Bravo et al. 2007, 2013; Sauka and Benintende 2008; Palma et al. 2014; Melo et al. 2016; Zhao et al. 2017).

From all the reported Cry proteins known to date, a overwhelming majority have been found in *B. thuringiensis* (Crickmore et al. 1998). In fact, the presence of the parasporal body, or crystal within the sporangium, segregates this species from other highly related species within the *B. cereus* group. However, some papers have reported the presence of Cry proteins (or *cry* genes) in other bacilli. Barloy et al. (1996) found a Cry protein (Cry16Aa1) from *Clostridium bifermentans*, showing mosquitocidal activity, and 2 years later they found another new Cry protein (Cry17Aa1) in the same species, with no significant mosquitocidal activity, though (Barloy et al. 1998). Later on, a Cry protein classified as Cry18Aa1, highly similar to Cry2Aa1, was found in the obligate pathogen of white grubs (scarabeid larvae) *Paenibacillus popilliae* (Zhang et al. 1997), followed by the discovery of Cry18Ba1 and

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13205-018-1533-3>) contains supplementary material, which is available to authorized users.

✉ Jorge E. Ibarra
jorge.ibarra@cinvestav.mx

¹ Cinvestav-Irapuato, Apartado postal 692, 36500 Irapuato, Guanajuato, Mexico

Cry18Ca1 proteins in the same bacterial species. Also, in a related species, *P. lentimorbus*, Yokoyama et al. (2004) discovered the Cry43Aa1 and Cry43Ba1 proteins, as well as some other *cry43*-like sequences, close-by in the genome. Cry43Aa1 showed some ingestion inhibition and high mortality on *Anomala cuprea* larvae. More recently, two Cry proteins, Cry48Aa1 and Cry49Aa1, were discovered in *Lysinibacillus sphaericus* (Jones et al. 2007). Interestingly, similar to the interaction between the Bin toxins of the same species, Cry48Aa1 and Cry49Aa1 complement each other to express their insecticidal activity. This was the first time this type of phenomenon was observed among Cry proteins.

In spite of these reports, the distribution of Cry proteins among bacillaceous bacteria is poorly studied, and new findings have been the result of chance or serendipity, rather than of a systematic exploration. For example, universal primers designed to amplify any three-domain Cry protein were used in a systematic search for *cry* genes from a collection of *B. thuringiensis* isolates, discovering three new type Cry proteins: Cry57Aa1, Cry58Aa1, and Cry59Aa1 (Noguera and Ibarra 2010).

On the other hand, given the cumulative number of genomes available in the NCBI gene bank, an *in silico* approach to find Cry genes from other bacilli is feasible. For this purpose, the hidden Markov models (HMM) have proven to be highly effective to analyze massive amounts of data. These models have been successfully applied on sequence analyses, discovery of genes, and characterization of protein families (Restrepo-Montoya et al. 2011). One HMM profile is statistically and probabilistically intrinsic, which is ideal to quantitatively evaluate if an individual sequence belongs to a given profile (Gong et al. 2012). HMM profiles have been widely used in the protein families Pfam database, which makes possible to search, classify, and characterize protein families. Also, these models have been used to identify transmembrane proteins, signal peptides, identification of carotenoid genes, identification of new genes coding for trypsin type proteases, among other uses (Chen et al. 2003; Liu et al. 2003; Tonhosolo et al. 2009). This approach has been used before to search for Cry proteins (Ye et al. 2012); however, the model detected many false positives.

Search for new Cry proteins is still valid as some have shown new host ranges (Méndez-López et al. 2003), higher toxicity (Reinoso-Pozo et al. 2016), or different mode of action, which may render alternatives in case of resistance development (Tabashnik et al. 1993). This is the reason why this report is focused on a systematic search for Cry proteins, focused on non-*B. thuringiensis* species, as a source. Our approach followed an *in silico* search for Cry proteins among 857 reported genomes of *Bacillus* spp.

Materials and methods

Search for representative clusters

A database was constructed with most of the Cry protein sequences found in the master database described in http://www.lifesci.sussex.ac.uk/home/Neil_Crickmore/Bt/, excluding those out of the “three-domain” group of proteins and those showing minimum differences. Preliminary attempts to build a cluster of sequences from the N-terminal half of the molecules failed when trying to build a hidden Markov model. Therefore, only those so-called “complete” sequences, which include the non-toxic, C-terminal half of the molecule were selected. That is, all the so-called “naturally truncated” proteins were excluded. All selected sequences were obtained from the NCBI database, available by 2017. This database of Cry protein sequences was scrutinized with the cd-hit-v4.6.7. software (<https://github.com/weizhongli/cdhit/releases>), which has been widely used to analyze large groups of sequences. Clusters were integrated with sequences showing between 50 and 90% identity. Then, the suitability of each cluster was analyzed individually with the MEGA7 software (http://www.megasoftware.net/download_form). The basic CD-HIT algorithm classified the sequences, both long and short ones, taking representative sequences from each cluster, which was classified as redundant or representative sequence according to its similarity to other representative sequences. The best advantage of this algorithm is its high speed to analyze the sequences and its use of filters of “short words” which cancel many unsuitable alignments (Huang et al. 2010; Fu et al. 2012; Weizhong et al. 2012). For the development of this model, cluster construction was based on 70% sequence similarity between the Cry proteins, along with some other criteria mentioned below to increase its stringency.

Construction of Hidden Markov Models

The HMMER V3.1b2 software (Eddy 1998) was used to construct a high-throughput HMMs profile to analyze the Cry protein sequences from the constructed database. Once clusters were obtained with the CD-HIT software as explained above, other criteria were added to the analysis, as follows: (a) sequences were eliminated from the database if they were incomplete, or if their δ -endotoxins do not show the typical three domains of the toxic region of the Cry proteins; (b) the Cry proteins called “naturally truncated” lack the half C-terminus, only the toxic fragment remaining; as the model is based on the C-terminal fragment, those “naturally truncated” Cry proteins, with

less than 800 amino acids, were also excluded from the analysis; and (c) only clusters containing more than five aligned sequences that fulfilled the described criteria were included in the analysis.

Sequences with at least 70% identity that fulfilled the criteria described above were read with the *hmmbuild* algorithm within the HMMER software to create an HMM profile. The algorithm generated a *.hmm file (Cry.hmm, available on request) containing a consensus sequence for the “complete” Cry protein family. Then the *hmmsearch* algorithm was used to search for Cry proteins in the *Bacillus* spp. database constructed for this purpose, which contains all the registered genomes within this genus. The model was restricted to show only Cry proteins with an e-value lower than 1×10^{-5} and more than 30% identity, as used earlier (Gong et al. 2012; Muñoz-Medina et al. 2015).

Validation of the HMM profile

To validate the HMM profile constructed as described above, a strain that includes both “complete” and “naturally truncated” Cry proteins was used as well as its acrystalliferous mutant. This strain was *B. thuringiensis* ssp. *israelensis* IPS-82, which contains four Cry proteins, but only two are complete (Cry4A and Cry4B). Its mutant 4Q7 (Jeong et al. 2014), lacks the pBtoxis plasmid (Berry et al. 2002), which contains all the *cry* genes of this strain. Additionally, the validation of the HMM profile verified if the Cry proteins identified by the algorithm were actually annotated as Cry proteins.

Multilocus Sequence Typing (MLST)

As the main objective of the work was to find Cry protein sequences in non-*Bacillus thuringiensis* strains, and due to the possibility of misidentification of the bacterial species whose sequences are registered in the NCBI database (Liu et al. 2015), further verification of the species identification which each genome was obtained from was carried out. All the ST and allelic profiles from each genome were obtained from PubMLST.org, to be examined by MLST analysis (Larsen et al. 2012). Some incomplete loci were sent to the MLST-1.8 server at the Center for Genomic Epidemiology (CGE). MLST concatamers were analyzed with MEGA7 (Tamura et al. 2011) and iTOL (Interactive Tree of Life) (Letunic and Bork 2016) software to build a dendrogram, both, to correlate the analyzed genomes to a species associated with a reference strain, according to the MLST analysis; and to correlate the presence of a Cry protein to a particular species. The analysis included starting values with 1000 replicates. This information helped to compare the distribution of Cry proteins among the different species of *Bacillus*

described in the NCBI database, with those species identified by our MLST analysis.

Results

Clustering of Cry proteins by CD-HIT and HMM profile construction

By 2017, 791 sequences were reported in the *cry* gene database (http://www.lifesci.sussex.ac.uk/home/Neil_Crickmore/Bt/). Once sequences were excluded, according to the criteria mentioned in M&M, alignments of the remaining sequences showed identities between 50 and 90%. As expected, clusters were fewer if the identity level was set at 50%, but the number of sequences per cluster was higher. Obviously, if identity was set at 90%, the number of clusters increased but with a lower number of sequences per cluster. Preliminary analyses of the sequences within each cluster indicated clearly that the N-terminal half of the molecules was more diverse than the C-terminal half, which was found highly conserved. From these observations, it was decided to design the HMM profiles from the C-terminal halves of the Cry proteins, being conscious that the “naturally truncated” Cry proteins were excluded from the analysis but, instead, the HMM profile precision increased significantly.

Once the criteria to design the HMM profiles were fulfilled (see M&M), six clusters were obtained, which included 111 conserved sequences. These sequences were manually edited by eliminating almost all the N-terminal halves and keeping the C-terminal halves to develop the HMM model. With these sequences, an HMM profile was obtained that was able to find 586 consensus positions (Cry.aln file available on request).

Validation of the HMM model

Once the HMM profile was designed, it was validated with the genomic DNA from *B. thuringiensis* ssp. *israelensis* (wild type) and its acrystalliferous mutant 4Q7. *B. thuringiensis* ssp. *israelensis* harbors the pBtoxis plasmid (missing in the 4Q7 mutant), which expresses four Cry proteins: Cry4Aa, Cry4Ba, Cry11Aa and Cry10Aa. The model accurately identified only the Cry4Aa and the Cry4Ba proteins, as they are the only ones in this strain that present the C-terminal half. Additionally, the model identified three annotations named pBt025, pBt026, and pBt048, which correspond to fragments of a Cry28-like protein (pBt025 and pBt026) and to a C-terminal fragment of a Cry4-like protein (pBt048). The model showed no identification in the mutant strain 4Q7, as expected.

Identification of Cry proteins by *hmmsearch*

Once 857 reported genomes from bacteria registered within the genus *Bacillus*, contained in 89 different species (including *B. thuringiensis*), were used to construct a database, the HMM model was tested and a total of 174 putative Cry proteins were identified (Online resource 1). From the 89 different species of *Bacillus* analyzed, Cry proteins were found only in three species: *B. thuringiensis*, *B. cereus*, *B. subtilis*; and two unknown species: *Bacillus* sp. strain L1B05 and *Bacillus* sp. strain LIB08, according to the species registered at the NCBI database. Once sequences of the 174 proteins were analyzed by *hmmsearch*, 125 were confirmed to be Cry proteins, 44 were identified as “hypothetical proteins”, and five more as “unknown” protein products. However, when these unmatched sequences were analyzed “manually” with the NCBI’s tool Blastp, all these sequences showed significant similarities with Cry proteins. Table 1 shows the 12 Cry subgroups identified in the genomes by the HMM model, as well as the number of proteins in each group and their frequency. As expected, Cry1-type proteins were more frequently found (43.68%), represented by 15 different subgroups, followed by the Cry4-type (19.54%), represented by four subgroups (Table 1).

Most importantly, from the 174 proteins identified by the HMM model, 42 were found in species different to *B. thuringiensis*, according to the species registered at the NCBI database. From these 42 sequences, 36 were found in genomes of *B. cereus*, five in unidentified bacilli (*Bacillus* spp.), and one in a genome registered at the NCBI database as a strain of *B. subtilis*. However, when detailed analyses of several gene sequences from the latter genome were made (data not shown), they indicated that this strain actually belongs to the *B. cereus* group rather than to *B. subtilis*, and, therefore, this report includes this genome as another

Table 2 Cry protein subgroups found in species different to *B. thuringiensis* by the HMM model, according to the species reported in the NCBI database

Species	Sequences found	Subgroups of Cry proteins found	Sequences per subgroup	Identity to highest hit (%)
<i>Bacillus cereus</i>	36	Cry1Aa	3	100*
		Cry1Ab	2	100*
		Cry1Ac	4	100*
		Cry1Ca	1	99
		Cry1Da	1	100
		Cry4Ba	11	63–100
		Cry5Aa	2	100*
		Cry8Aa	2	88, 90
		Cry9Ea	2	100*
		Cry21Ba	2	34, 46
		Cry21Fa	2	43, 49
		Cry32Aa	1	56
		Cry32Da	2	59, 74
		Cry32Eb	1	73
<i>Bacillus</i> spp	6	Cry14Aa	1	100
		Cry4Aa	3	99–100
		Cry40	2	62*

*All sequences showed the same identity level

Bacillus sp. strain. In general, similar to the total recount, Cry1-type proteins were more frequent in *B. cereus*, among other types. Only Cry4A, Cry14, and Cry40A were found in genomes of unidentified bacilli (Table 2).

Species identification by MLST analysis of genomes with Cry sequences

Because the 857 analyzed genomes belong to 89 different species within the *Bacillus* genus, and 174 putative Cry proteins were found in 60 different genomes, the identification of the bacterial species containing those 60 genomes was

Table 1 Subgroups from the main Cry-type proteins identified in silico by the HMM model in 857 genomes of *Bacillus* spp

Cry types	No. of sequences	Subgroups	Frequency (%)
Cry1	76	Cry1Aa, Cry1Ab, Cry1Ac, Cry1Ae, Cry1Ba, Cry1Bb, Cry1Bc, Cry1Bd, Cry1Ca, Cry1Cb, Cry1Da, Cry1Db, Cry1Fb, Cry1Ga, Cry1Ha	43.68
Cry4	34	Cry4Aa, Cry4Ba, Cry4Cb, Cry4Da	19.54
Cry7	16	Cry7Aa, Cry7Ab, Cry7Ba, Cry7Ca, Cry7Ea	9.20
Cry21	10	Cry21Ba, Cry21Fa, Cry21Ga, Cry21Ha	5.75
Cry39	8	Cry39	4.60
Cry8	6	Cry8Aa, Cry8Ba, Cry8Ea, Cry8X	3.45
Cry9	6	Cry9Ea	3.45
Cry32	6	Cry32Da, Cry32Eb	3.45
Cry5	5	Cry5Aa, Cry5Ba	2.87
Cry40	3	Cry40Aa	1.72
Cry61	3	Cry61Aa	1.72
Cry14	1	Cry14Aa	0.57

verified by MLST analysis to corroborate (or refute) the identification registered at the NCBI database. It is a fact that, with some frequency, a proper identification of bacteria from the registered genomes to the NCBI database might be erroneous (Liu et al. 2017), mostly if those species belong to the *B. cereus* group, where the species discrimination is difficult. Therefore, concatamers were built from the reported genomes that showed the presence of Cry proteins and an MLST analysis was carried out, to verify the species identification. Figure 1 shows the distribution of these genomes in a dendrogram, where 12 reference species were added. As observed, all the analyzed genomes containing Cry protein sequences are distributed among species that belong to

the *B. cereus* group (separated by lines in Fig. 1). Cry proteins detected in each genome are shown in Fig. 1 (see the corresponding genome in Online resource 1). Twenty-five genomes were related to the *B. cereus* clade, 21 genomes to the *B. thuringiensis* clade, five genomes to the *B. anthracis* clade, one genome to the *B. mycoides* clade, and one more to the *B. toyonensis* clade, a species that was recently reported to belong to the *B. cereus* group (Liu et al. 2017). No concatamers could be made from the remaining seven genomes. It is important to notice that 38 species reported in the NCBI database containing these genomes show a discrepancy with the identification based on our MLST analysis (Online resource 2). If this analysis is valid, then the number of Cry

Fig. 1 Distribution of Cry proteins found in different species of *Bacillus*, when a dendrogram was inferred from the concatamers analyzed by MLST to identify the species containing the 60 genomes that showed putative Cry protein sequences, according to the HMM

model, and their relationship with type strains. Numbers in parenthesis correspond to the genome shown on Online resource 1, and the abbreviation of the species name registered at NCBI, followed by their Cry content. Bc: *B. cereus*, Bt: *B. thuringiensis*, Bs: *Bacillus* sp

protein sequences found in non-*B. thuringiensis* increased from 42 to 119, indicating that Cry protein sequences were more frequent in non-*B. thuringiensis* strains than in *B. thuringiensis* strains.

Discussion

The great diversity of Cry proteins, totaling more than 800 different sequences registered to date (http://www.lifesci.sussex.ac.uk/home/Neil_Crickmore/Bt/), within 75 different holotypes, may indicate that this family of proteins may not be restricted only to the species *B. thuringiensis*. In fact, there are previous reports about the presence of Cry proteins in other bacilli, such as *C. bifermentans*, *P. popilliae*, *P. lentimorbus*, and *L. sphaericus* (Barloy et al. 1996, 1998; Zhang et al. 1997; Yokoyama et al. 2004; Jones et al. 2007). Nevertheless, those findings were more the result of chance or serendipity, rather than following a systematic search, either based on an in silico procedure or the use of a large collection of different species of bacilli. The approach followed here may change the concept that associates Cry proteins only with *B. thuringiensis* and may even change the association of Cry proteins only to an insecticidal activity, as suggested earlier (Melo et al. 2016).

The number of genomes available in the NCBI database has increased geometrically in the past years (Land et al. 2015). That makes it an extraordinary source to search for specific genes; however, the amount of data to analyze is exceptionally large. That is why special tools have been developed for this type of analyses, such as the hidden Markov model profiles (Eddy 1998; Yoon 2009). The development of a specific profile to search for Cry genes in 857 genomes of *Bacillus* spp. clearly indicated that this family of proteins can be found in other species different to *B. thuringiensis*, as 42 sequences related to Cry proteins were found in *B. cereus* and three unidentified bacilli.

The HMM profile developed here proved to be valid to detect Cry proteins containing the C-terminal half as it verified the detection of Cry4A and Cry4B from *B. thuringiensis* ssp. *israelensis*, but excluded Cry11 and Cry10, which lack the C-terminal half of the molecule, during the validation test of the model. Interestingly, the profile was able to detect other sequences related to other Cry genes in the same strain. According to the pBtoxis plasmid sequence, these sequences exist, but lack their N-terminal half. Several reports support this observation as well as natural disconnections of the N-terminal half from its C-terminal half (Ito et al. 2002; Ohgushi et al. 2005; Hernández-Soto et al. 2009; Noguera and Ibarra 2010; Barboza-Corona et al. 2012; Sun et al. 2013). In any case, this validation indicated that the developed HMM profile was able to detect any C-terminal sequence of a Cry protein, independently of the presence

of its N-terminal half. Besides, the same subspecies, but without the pBtoxis plasmid, showed negative results, as expected. However, at the same time, a shortcoming of this HMM profile is evident, as the so-called “naturally truncated” Cry proteins, such as Cry2, Cry3, Cry11, etc., cannot be detected by this profile. Nevertheless, the HMM profile developed here was totally efficient to detect three-domain “complete” Cry proteins, as compared with a similar report (Ye et al. 2012) where the development of an HMM profile to search for Cry proteins also detected many false positives.

The number of Cry proteins found with the HMM profile was higher than expected (174), but it was biased due to the large number of *B. thuringiensis* genomes included in the database. However, the number is still high (Vilas-Bôas et al. 2007) if it is considered that 42 sequences were found in species different to *B. thuringiensis*. As expected, most of these sequences were found in the highly related species *B. cereus* (Rasko et al. 2005) but, still, five more were found from two unidentified bacilli (*Bacillus* spp.), and one from a strain originally reported as *B. subtilis* at the NCBI database; however, our analysis showed that it also belongs to the *B. cereus* group, as well as the other two *Bacillus* sp. strains.

It is important to note that the results from this work showed that all the putative Cry proteins were found in species belonging to the *B. cereus* group, even if the 857 genomes analyzed belonged to 89 *Bacillus* species. This observation may indicate that all cry genes are innate to this group, regardless if they form a parasporal crystal or not. And yet, a work is underway on a database with genomes from other genera of the family Bacillaceae. Although this approach is still too far from clarifying how disperse and diverse is this protein family among bacteria, it is expected to answer many of these questions, as the number of genomes increases as well as of metagenomes.

These results validate the question about the evolutionary advantage and/or physiological role of Cry proteins, whether in *B. thuringiensis* or not, as just a handful of these proteins show high insecticidal toxicity. As shown here, Cry proteins are highly dispersed and frequently found, at least in the *B. cereus* group. Labeling all these as insecticidal proteins would be a mistake, meaning that much work is still to be done to find the real context of these proteins.

Furthermore, it was clear that there is a lack of taxonomic confirmation in the NCBI sequence database. The most evident case was the reported genome from a *B. subtilis* strain which was actually a species within the *B. cereus* group. Most importantly, from the MLST analysis of the 60 strains showing putative Cry proteins, 38 (63%) showed a difference with the species identification reported at the NCBI database (Liu et al. 2015). If the MLST results were accepted as valid in our analysis, the proportion between the Cry proteins found in *B. thuringiensis* strains versus non-*B. thuringiensis* strains changes dramatically. From 132 Cry

proteins originally found in *B. thuringiensis*, it would change to only 55, and from 42 Cry proteins originally found in non-*B. thuringiensis*, it would increase to 119 found in four species, indicating that Cry protein sequences are more frequent in non-*B. thuringiensis* strains than in *B. thuringiensis* strains. This change would affect the conclusion about the distribution of Cry proteins in the *Bacillus* species, although it would not change about their distribution within the *B. cereus* group.

The search for new Cry proteins is important not only for the reasons explained above, but new Cry proteins in other species different to *B. thuringiensis* may also explain the origin of such a diverse family and, perhaps, new roles, other than their insecticidal activity, mostly on the ecology of these bacteria, rather than on its metabolism. These results question the idea about the role of *B. thuringiensis* in nature which is still uncertain, mostly because, as mentioned, the overwhelming majority of natural isolates show very low or no insecticidal activity, and new roles in nature have been observed (Lopez-Meza and Ibarra 1996; Zhang et al. 1997; Juárez-Pérez et al. 2003; Yokoyama et al. 2004; Ito et al. 2006; Peng et al. 2015; Melo et al. 2016; Jouzani et al. 2017). This report shows that it is feasible to find Cry proteins beyond *B. thuringiensis*, and how important the systematic search for Cry proteins is, to solve many question on the real role of these proteins (and *B. thuringiensis*) in nature.

Acknowledgements Authors are in debt for the excellent technical support of Regina Basurto-Ríos, Javier Luévano-Borroel, Africa Islas-Robles, and Leandro Gabriel Ordóñez-Acevedo. JFCE and IHG received PhD fellowships from Consejo Nacional de Ciencia y Tecnología (CONACYT, Mexico).

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

- Barboza-Corona JE, Park HW, Bideshi DK, Federici BA (2012) The 60-kilodalton protein encoded by orf2 in the cry19A operon of *Bacillus thuringiensis* ssp. jegathesan functions like a C-terminal crystallization domain. *Appl Environ Microbiol* 78:2005–2012. <https://doi.org/10.1128/AEM.06750-11>
- Barloy F, Delécluse A, Nicolas L, Lecadet MM (1996) Cloning and expression of the first anaerobic toxin gene from *Clostridium bifermentans* subsp. malaysia, encoding a new mosquitocidal protein with homologies to *Bacillus thuringiensis* delta-endotoxins. *J Bacteriol* 178:3099–3105. <https://doi.org/10.1128/jb.178.11.3099-3105.1996>
- Barloy F, Lecadet MM, Delécluse A (1998) Cloning and sequencing of three new putative toxin genes from *Clostridium bifermentans* CH18. *Gene* 211:293–299. [https://doi.org/10.1016/S0378-1119\(98\)00122-X](https://doi.org/10.1016/S0378-1119(98)00122-X)
- Berry C, O'Neil S, Ben-dov E, Jones AF, Murphy L, Quail MA, Holden MTG, Harris D, Zaritsky A, Parkhill J (2002) Complete sequence and organization of pBtoxis, the toxin-coding plasmid of *Bacillus thuringiensis* subsp. israelensis. *Appl Environ Microbiol* 68:5082–5095. <https://doi.org/10.1128/AEM.68.10.5082-5095.2002>
- Bravo A, Gill SS, Soberon M (2007) Mode of action of *Bacillus thuringiensis* Cry and Cyt toxins and their potential for insect control. *Toxicon* 49:423–435. <https://doi.org/10.1016/j.toxicol.2006.11.022>
- Bravo A, Gómez I, Porta H, García-Gómez BI, Rodríguez-Almazan C, Pardo L, Soberón M (2013) Evolution of *Bacillus thuringiensis* Cry toxins insecticidal activity. *Microb Biotechnol* 6:17–26. <https://doi.org/10.1111/j.1751-7915.2012.00342.x>
- Chen Y, Yu P, Luo J, Jiang Y (2003) Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. *Mamm Genome* 14:859–865. <https://doi.org/10.1007/s00335-003-2296-6>
- Crickmore N, Zeigler DR, Feitelson J, Schnepf E, Van Rie J, Lereclus D, Baum J, Dean DH (1998) Revision of the nomenclature for the *Bacillus thuringiensis* pesticidal crystal proteins. *Microbiol Mol Biol Rev* 62:807–813
- Crickmore N, Baum J, Bravo A et al (2017) *Bacillus thuringiensis* toxin nomenclature. <http://www.btnomenclature.info/>. Accessed 30 May 2018
- De Maagd RA, Bravo A, Crickmore N (2001) How *Bacillus thuringiensis* has evolved specific toxins to colonize the insect world. *Trends Genet* 17:193–199. [https://doi.org/10.1016/S0168-9525\(01\)02237-5](https://doi.org/10.1016/S0168-9525(01)02237-5)
- Eddy S (1998) Profile hidden Markov models. *Bioinformatics* 14:755–763. <https://doi.org/10.1093/bioinformatics/14.9.755>
- Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>
- Gong YN, Chen GW, Shih SR (2012) Characterization of subtypes of the influenza A hemagglutinin (HA) gene using profile hidden Markov models. *J Microbiol Immunol Infect* 45:404–410. <https://doi.org/10.1016/j.jmii.2011.12.018>
- Hernández-Soto A, Del Rincón-Castro MC, Espinoza AM, Ibarra JE (2009) Parasporal body formation via overexpression of the Cry10Aa toxin of *Bacillus thuringiensis* subsp. israelensis, and Cry10Aa-Cyt1Aa synergism. *Appl Environ Microbiol* 75:4661–4667. <https://doi.org/10.1128/AEM.00409-09>
- Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26:680–682. <https://doi.org/10.1093/bioinformatics/btq003>
- Ito T, Sahara K, Bando H, Asano S (2002) Cloning and expression of novel crystal protein genes cry39A and cry39ORF2 from *Bacillus thuringiensis* ssp. aizawai Bun1-14 encoding mosquitocidal proteins. *J Insect Biotechnol Sericol* 128:123–128. <https://doi.org/10.11416/JIBS2001.71.123>
- Ito T, Ikeya T, Sahara K, Bando H, Asano SI (2006) Cloning and expression of two crystal protein genes, cry30Ba1 and cry44Aa1, obtained from a highly mosquitocidal strain, *Bacillus thuringiensis* subsp. entomocidus INA288. *Appl Environ Microbiol* 72:5673–5676. <https://doi.org/10.1128/AEM.01894-05>
- Jeong H, Park SH, Choi SK (2014) Genome sequence of the Acrytalliferous *Bacillus thuringiensis* serovar israelensis strain 4Q7, widely used as a recombination host. *Genome Announc* 2:e00231–e00214. <https://doi.org/10.1128/genomeA.00231-14>
- Jones GW, Nielsen-Leroux C, Yang Y, Yuan Z, Dumas VF, Gomes Monnerat R, Berry C (2007) A new Cry toxin with a unique two-component dependency from *Bacillus sphaericus*. *FASEB J* 21:4112–4120. <https://doi.org/10.1096/fj.07-8913com>
- Jouzani GS, Valijaniani E, Sharafi R (2017) *Bacillus thuringiensis*: a successful insecticide with new environmental features and

- tidings. Appl Microbiol Biotechnol 101:2691–2711. <https://doi.org/10.1007/s00253-017-8175-y>
- Juárez-Pérez V, Porcar M, Orduz S, Delécluse A (2003) Cry29A and Cry30A: Two Novel δ -endotoxins Isolated from *Bacillus thuringiensis* serovar *medellin*. Syst Appl Microbiol 26:502–504. <https://doi.org/10.1078/072320203770865783>
- Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW (2015) Insights from 20 years of bacterial genome sequencing. Funct Integr Genom 15:141–161. <https://doi.org/10.1007/s10142-015-0433-4>
- Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM, Lund OJ (2012) Multilocus sequence typing of total genome sequenced bacteria. Clin Microbiol 50:1355–1361. <https://doi.org/10.1128/JCM.06094-11>
- Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucl Acids Res 44:W242–W245. <https://doi.org/10.1093/nar/gkw290>
- Liu Q, Zhu Y, Wang B, Li Y (2003) A HMM-based method to predict the transmembrane regions of b-barrel membrane proteins. Comput Biol Chem 27:69–76. [https://doi.org/10.1016/S0097-8485\(02\)00051-7](https://doi.org/10.1016/S0097-8485(02)00051-7)
- Liu Y, Lai Q, Göker M, Meier-Kolthoff JP, Wang M, Sun Y et al (2015) Genomic insights into the taxonomic status of the *Bacillus cereus* group. Sci Rep 5:14082. <https://doi.org/10.1038/srep14082>
- Liu Y, Du J, Lai Q, Zeng R, Ye D, Xu J, Shao Z (2017) Proposal of nine novel species of the *Bacillus cereus* group. Int J Syst Evol Microbiol 67:2499–2508. <https://doi.org/10.1099/ijsem.0.001821>
- Lopez-Meza JE, Ibarra JE (1996) Characterization of a novel strain of *Bacillus thuringiensis*. Appl Environ Microbiol 62:1306–1310
- Melo ALDA, Soccol VT, Soccol CR (2016) *Bacillus thuringiensis*: Mechanism of action, resistance, and new applications: a review. Crit Rev Biotechnol 36:317–326. <https://doi.org/10.3109/07388551.2014.960793>
- Méndez-López I, Basurto-Ríos R, Ibarra JE (2003) *Bacillus thuringiensis* serovar israelensis is highly toxic to the coffee berry borer, *Hypothenemus hampei* Ferr. (Coleoptera: Scolytidae). FEMS Microbiol Lett 226:73–77. [https://doi.org/10.1016/S0378-1097\(03\)00557-3](https://doi.org/10.1016/S0378-1097(03)00557-3)
- Muñoz-Medina JE, Sánchez-Vallejo CJ, Méndez-Tenorio A et al (2015) In silico identification of highly conserved epitopes of influenza A H1N1, H2N2, H3N2, and H5N1 with diagnostic and vaccination potential. Biomed Res Int 2015:813047. <https://doi.org/10.1155/2015/813047>
- Noguera PA, Ibarra JE (2010) Detection of new cry genes of *Bacillus thuringiensis* by use of a Novel PCR primer system. Appl Environ Microbiol 76:6150–6155. <https://doi.org/10.1128/AEM.00797-10>
- Ohgushi A, Saitoh H, Wasano N, Uemori A, Ohba M (2005) Cloning and characterization of two novel genes, cry24B and sIorf2, from a mosquitocidal strain of *Bacillus thuringiensis* serovar sotto. Curr Microbiol 51:131–136. <https://doi.org/10.1007/s00284-005-7529-3>
- Palma L, Muñoz D, Berry C, Murillo J, Caballero P (2014) *Bacillus thuringiensis* toxins: an overview of their biocidal activity. Toxins 6:3296–3325. <https://doi.org/10.3390/toxins6123296>
- Peng DH, Pang CY, Wu H, Huang Q, Zheng JS, Sun M (2015) The expression and crystallization of Cry65Aa require two C-termini, revealing a novel evolutionary strategy of *Bacillus thuringiensis* Cry proteins. Sci Rep 5:19–21. <https://doi.org/10.1038/srep08291>
- Rasko DA, Altherr MR, Han CS, Ravel J (2005) Genomics of the *Bacillus cereus* group of organisms. FEMS Microbiol Rev 29:303–329. <https://doi.org/10.1016/j.femsre.2004.12.005>
- Reinoso-Pozo Y, Del Rincón-Castro MC, Ibarra JE (2016) Characterization of a highly toxic strain of *Bacillus thuringiensis* serovar kurstaki very similar to the HD-73 strain. FEMS Microbiol Lett 363:1–6. <https://doi.org/10.1093/femsle/fnw188>
- Restrepo-Montoya D, Becerra D, Carvajal-Patiño J, Mongui A, Niño L, Patarroyo M, Patarroyo M (2011) Identification of plasmodium vivax proteins with potential role in invasion using sequence redundancy reduction and profile hidden Markov models. PLoS One 6:e25189. <https://doi.org/10.1371/journal.pone.0025189>
- Sauka DH, Benintende GB (2008) *Bacillus thuringiensis*: generalidades. Un acercamiento a su empleo en el biocontrol de insectos lepidópteros que son plagas agrícolas. Rev Argent Microbiol 40:124–140
- Sun Y, Zhao Q, Xia L, Ding X, Hu Q, Federici BA, Park HW (2013) Identification and characterization of three previously undescribed crystal proteins from *Bacillus thuringiensis* subsp. *jegathesan*. Appl Environ Microbiol 79:3364–3370. <https://doi.org/10.1128/AEM.00078-13>
- Tabashnik BE, Finson N, Johnson MW, Moar WJ (1993) Resistance to Toxins from *Bacillus thuringiensis* subsp. *kurstaki* Causes Minimal Cross-Resistance to *B. thuringiensis* subsp. *aizawai* in the Diamondback Moth (Lepidoptera: Plutellidae). Appl Environ Microbiol 59:1332–1335
- Tamura K, Peterson D, Peterson N, Stecher G, Ne M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739. <https://doi.org/10.1093/molbev/msr121>
- Tonhosolo R, D’Alexandri FL, de Rosso VV et al (2009) Carotenoid biosynthesis in intraerythrocytic stages of *Plasmodium falciparum*. J Biol Chem 284:9974–9985. <https://doi.org/10.1074/jbc.M807464200>
- Vilas-Boas GT, Peruca APS, Arantes OMN (2007) Biology and taxonomy of *Bacillus cereus*, *Bacillus anthracis*, and *Bacillus thuringiensis*. Can J Microbiol 53:673–687. <https://doi.org/10.1139/W07-029>
- Weizhong L, Limin F, Beifang N, Sitao W, John W (2012) Ultrafast clustering algorithms for metagenomic sequence analysis. Brief Bioinform 13:656–668. <https://doi.org/10.1093/bib/bbs035>
- Ye W, Zhu L, Liu Y, Crickmore N, Peng D, Ruan L, Sun M (2012) Mining new crystal protein genes from *Bacillus thuringiensis* based on mixed plasmid-enriched genome sequencing and a computational pipeline. Appl Environ Microbiol 78:4795–4801. <https://doi.org/10.1128/AEM.00340-12>
- Yokoyama T, Tanaka M, Hasegawa M (2004) Novel cry gene from *Paenibacillus lentimorbus* strain Semadara inhibits ingestion and promotes insecticidal activity in *Anomala cuprea* larvae. J Invertebr Pathol 85:25–32. <https://doi.org/10.1016/j.jip.2003.12.009>
- Yoon BJ (2009) Hidden Markov models and their applications in biological sequence analysis. Curr Genom 10:402–415. <https://doi.org/10.2174/138920209789177575>
- Zhang J, Hodgman TC, Krieger L, Schnetter W, Schairer HU (1997) Cloning and analysis of the first cry gene from *Bacillus popilliae*. J Bacteriol 179:4336–4341. <https://doi.org/10.1128/jb.179.13.4336-4341.1997>
- Zhao M, Yuan X, Wei J, Zhang W, Wang B, Myint Khaing M, Liang G (2017) Functional roles of cadherin, aminopeptidase-N and alkaline phosphatase from *Helicoverpa armigera* (Hübner) in the action mechanism of *Bacillus thuringiensis* Cry2Aa. Sci Rep 7:1–9. <https://doi.org/10.1038/srep46555>