ORIGINAL RESEARCH
published: 28 June 2021
doi: 10.3389/fmicb.2021.664476

IDOPS, a Profile HMM-Based Tool to Detect Pesticidal Sequences and Compare Their Genetic Context

*Stefani Díaz-Valerio, Anat Lev Hacohen, Raphael Schöppe and Heiko Liesegang**

Genomic and Applied Microbiology & Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August University of Göttingen, Göttingen, Germany

Biopesticide-based crop protection is constantly challenged by insect resistance. Thus, expansion of available biopesticides is crucial for sustainable agriculture. Although *Bacillus thuringiensis* is the major agent for pesticide bioprotection, the number of bacteria species synthesizing proteins with biopesticidal potential is much higher. The Bacterial Pesticidal Protein Resource Center (BPPRC) offers a database of sequences for the control of insect pests, grouped in structural classes. Here we present IDOPS, a tool that detects novel biopesticidal sequences and analyzes them within their genetic environment. The backbone of the IDOPS detection unit is a curated collection of high-quality hidden Markov models that is in accordance with the BPPRC nomenclature. IDOPS was positively benchmarked with BtToxin_Digger and Cry_Processor. In addition, a scan of the UniProtKB database using the IDOPS models returned an abundance of new pesticidal protein candidates distributed across all of the structural groups. Gene expression depends on the genomic environment, therefore, IDOPS provides a comparative genomics module to investigate the genetic regions surrounding pesticidal genes. This feature enables the investigation of accessory elements and evolutionary traits relevant for optimal toxin expression and functional diversification. IDOPS contributes and expands our current arsenal of pesticidal proteins used for crop protection.

Keywords: biopesticide, hidden markov model, insecticidal protein, toxin identification, pesticidal, genetic context, IDOPS, comparative genomics

OPEN ACCESS

Edited by:
*Tereza Castellane,
Sáo Paulo State University, Brazil*

Reviewed by:
*Ricardo Polanczyk,
Sáo Paulo State University, Brazil
Jie Zhang,
Institute of Plant Protection, Chinese Academy of Agricultural Sciences, China*

***Correspondence:**
*Heiko Liesegang
hlieseg@gwdg.de*

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *05 February 2021*
Accepted: *13 May 2021*
Published: *28 June 2021*

Citation:
Díaz-Valerio S, Lev Hacohen A, Schöppe R and Liesegang H (2021) IDOPS, a Profile HMM-Based Tool to Detect Pesticidal Sequences and Compare Their Genetic Context. Front. Microbiol. 12:664476. doi: 10.3389/fmicb.2021.664476

1. INTRODUCTION

Agricultural management strategies regard biopesticides as an environmentally friendly alternative to the chemical formulations used to suppress invertebrate pests (Mnif and Ghribi, 2015; Kachhawa, 2017). Plant-associated, soil, and entomopathogenic bacteria are a natural source of agents with pesticidal potential. Among those, *Bacillus thuringiensis* (Bt) strains and their derived crop protection products are safe for humans, highly specific to the targeted pests, and affordable to manufacture in bulk, making Bt the most successful biopesticide implemented worldwide (George and Crickmore, 2012; Jouzani et al., 2017). Nevertheless, nature fights back and target insects develop resistance mechanisms against Bt, which creates a constant need for novel and improved pest control agents (Vílchez, 2020).

Recently, it became clear that there is a wider variety of pesticidal proteins, synthesized by other bacteria that also interact with insects as part of their lifestyle (Castagnola and Stock, 2014; Ruiu, 2018). These have the potential to replace, supplement, and expand current options for biopest control (Waterfield et al., 2001). Bacteria like *Dickeya* spp., (Loth et al., 2015) *Yersinia* spp., and *Photorhabdus* spp. (Heermann and Fuchs, 2008) are now known sources of such proteins. The varied pesticidal proteins were re-classified in 16 structural groups by the Bacterial Pesticidal Protein Resource Center (BPPRC) (Crickmore et al., 2020a,b).

Biological databases are constantly enriched due to the proliferation of sequencing projects and advancements in sequencing technologies, thus they are a promising reservoir of sequences with uncharacterized pesticidal potential. However, screening such vast amounts of data requires sophisticated computational approaches in order to gain insight and take advantage of less explored resources. The use of profile hidden Markov models (HMMs) to analyze biological data has proven to be robust and sensitive (Eddy, 1998). A profile HMM condenses the information of a multiple alignment of homologous sequences, and therefore, has a higher discriminative power than pairwise similarity-based search tools like BLAST (Söding, 2005). Profile HMMs are broadly applied for protein family assignment, domain analysis, and detection of remote homologies in resources such as InterProScan (Jones et al., 2014; Blum et al., 2020), PFAM (Sonnhammer et al., 1998; Finn et al., 2007), and TIGRFAM (Haft et al., 2012). Examples of dedicated collections of profile HMMs are RVDB-prot (Bigot et al., 2019), a database for detection of viral proteins, and TASmania, a tool for the discovery of toxin-antitoxin systems in bacterial genomes (Akarsu et al., 2019).

Consequently, previous efforts to implement profile HMMs for detection of pesticidal proteins are not surprising. Cry_Processor (Shikov et al., 2020) is a tool for identification of 3 domain Cry sequences based on 4 profile HMMs, one for each domain and one full-length protein, making single domain delimitation possible. BtToxin_Digger (Liu et al., 2020), the successor of BtToxin_scanner, relies on a combination of HMMs, BLAST, and support vector machine (SVM) for prediction of not only 3 domain toxins but also members of the other structural groups.

The focus of the existing toxin prediction tools is to recognize and classify pesticidal protein sequences. Currently, none of them take into consideration the genetic environment of the genes coding for pesticidal proteins. Surrounding elements often include chaperones, crystallization domains, mobile elements, transporters, prophages, and virulence factors (Koni and Ellar, 1993; Shao et al., 2001; Elleuch et al., 2016; Adalat et al., 2017; Fayad et al., 2020; Lechuga et al., 2020). Moreover, the arrangement and distribution of such elements across genomes reveal crucial details about toxin functionality, host adaptation, diversification, and evolution of biopesticides (Khasdan et al., 2007; Peng et al., 2015; Ruffner et al., 2015; Fiedoruk et al., 2017; Zheng et al., 2017; Fayad et al., 2020; Wang et al., 2020). Hence, a more exhaustive approach would not only detect pesticidal sequences but also enable comparative genomics analysis of the candidate toxin in order to characterize the complete expression unit.

The goal of this study was to develop such tool. Our efforts originated IDOPS (Identification of Pesticidal Sequences), a software based on an extensive collection of high-quality profile HMMs. IDOPS aims to provide (i) a detection unit to aid in the finding of pesticidal proteins, especially novel variants within recently expanding groups, (ii) a basic classification system in accordance with the BPPRC nomenclature system, and (iii) a comparative genomics module to investigate toxin genes and their complete expression unit within their genomic environment.

2. MATERIALS AND METHODS

2.1. Building Profile HMMs

In order to better represent the great diversity of pesticidal proteins, we created a collection of known and putative pesticidal sequences. Our initial collection combined data from the previous *Bacillus thuringiensis* Toxin Nomenclature website (Crickmore et al., 1998), and matches for various pesticidal sequences found at UniProtKB-2020_06 (UniProt Consortium, 2019). Since UniProt-TrEMBL includes fragmented and repeated entries, redundancy was removed by clustering sequences of 100% identity with CD-HIT v.4.8.1 (Li and Godzik, 2006); then the longest member of each cluster was preserved. The representative sequences were used to create an all vs. all matrix with BLASTp v.2.9.0 (Camacho et al., 2009). This matrix was the input for clustering with the Markov Clustering Algorithm implemented by MCL (Enright et al., 2002). Resulting groups containing at least five members were aligned by ClustalO 1.2.4 (Sievers and Higgins, 2018), and the alignment was passed to hmmbuild-HMMER v.3.3 (Eddy, 1998) to create profile hidden Markov models (HMMs). Furthermore, an individual profile HMM was created to represent the C-terminal region of Cry toxins longer than 1,000 amino acids. For such purpose, the toxin core of the Cry sequences was removed from the alignment before the hmmbuild step. The preliminary profile HMM database accurately described pesticidal proteins of *B. thuringiensis* and was useful for detection of novel toxins. Nevertheless, at the time of this study, the BPPRC released the updated classification of pesticidal proteins and their source organisms. Therefore, the initial profile model collection undergone further examination and additional models were created to represent sequences from non Bt organisms.

2.2. Model Refinement and Validation

Several rounds of manual refinement and optimization were necessary to ensure high sensitivity and specificity of the models. This step included evaluation of the sequences used for each profile HMM, their phylogenetic relationship, domain signature predicted by the InterPro consortium (Blum et al., 2020), removal of biases produced by over-represented sequences and performance when databases were scanned using hmmsearch-HMMER v.3.3 (Eddy, 1998).

Protein sequences from the BPPRC database were considered as true positives. For a control dataset of true negatives, a collection of bacterial pore-forming toxins and related proteins from other bacteria was used (Gonzalez et al., 2008; **Supplementary Table 1**). The overall performance of the models was evaluated by scanning the UniProtKB databases. Furthermore, the distribution of the sequences matched by each model was analyzed and served to determine a trusted cutoff value above which true protein class members are found.

Criterion to define a good model are:

- Identification of all the true positive members of the protein class (or subclass) described by the model within the high score range.
- Additional matches within the high score range can be consistently assigned to the protein group by evaluation of domain signatures, source organism, and quality of alignment with true members.
- Distantly related pore-forming toxins from the true negative control dataset are not found at all or found below the trusted cutoff value.
- The distribution of the proteins matched by the model when searching UniProtKB databases indicates a clear separation between true members and other protein matches.

2.3. Pipeline Implementation

This collection of profile HMMs can be used effectively to search for matches within query sequences, genomes, and whole databases. It was implemented in a pipeline named Identification of Pesticidal Sequences (IDOPS). The software applied in IDOPS and the corresponding versions are presented in **Table 1**. To ease distribution and installation, IDOPS is available as a conda package (https://anaconda.org/GAMB-GO/idops) and its source code is found at Github (https://github.com/GAMB-GO/IDOPS), under a GPLv3 license.

Valid input formats for IDOPS are fasta (for single or multiple proteins) and genbank (for complete or draft genomes). Candidate pesticidal sequences are identified by hmmscan-HMMER v.3.3 (Eddy, 1998) and evaluated against the trusted cutoffs of the refined profile models. To facilitate the assessment of the found candidates, IDOPS internally generates profile alignments and reports phylogenetic trees for each match with the 10 nearest sequences of the corresponding protein groups.

Furthermore, in order to characterize the genomic context of identified pesticidal proteins and produce relevant insights regarding evolution and functionality, IDOPS offers an additional feature when genomic data is provided in genbank format. First, it retrieves the 5000 bp upstream and downstream regions of each match. Secondly, it does a Prokka annotation (Seemann, 2014) of such sequences. Finally, it creates an EasyFig (Sullivan et al., 2011) comparison that depicts BLAST identities between conserved regions.

TABLE 1 | Software dependencies and versions implemented in IDOPS.

Software	Version
Python	3.7.6
HMMer (hmmbuild, hmmsearch, hmmscan, hmmalign)	3.3
Biopython	1.76
Easyfig	2.2.5
Clustal Omega	1.2.4
BLAST	2.9.0
Prokka	1.14.6

2.4. Comparative Analysis of Genetic Environments

We used IDOPS to analyze eight *B. thuringiensis* plasmids carrying *cry1* genes (detailed strains and accession ID list in **Supplementary Table 2**). It was previously reported that *cry1* can occur alone or as part of an insecticidal pathogenicity island (Fiedoruk et al., 2017). We used this observation to demonstrate the value of IDOPS to facilitate the discovery of different genetic arrangements.

2.5. Benchmarking

The performance of IDOPS was compared with that of current tools for pesticidal protein detection: Cry_Processor (Shikov et al., 2020) and BtToxin_digger (Liu et al., 2020). Cry_Processor was developed to identify 3 domain Cry sequences. Therefore, the tool was examined taking in consideration only this group of proteins. It provides two search modes, Find Domains (FD) that is based upon a hmmsearch against generalized HMM models, and Domains Only (DO) which searches directly for each of the domains without a filtering step. The two modes were applied. Both tools were tested with sequences from the BPPRC as positive dataset and the true negative collection of distantly related pore-forming toxins.

3. RESULTS

3.1. A Collection of High-Quality Profile Hidden Markov Models

We developed highly specific profile HMMs to accurately represent each of the 16 structural groups defined by the Bacterial Pesticidal Protein Resource Center. The iterative refinement and manual curation process led to the creation of subgroups within some of the protein classes; particularly for highly populated and diverse groups, like Cry and Cyt. Our final collection consists of 31 profile hidden Markov models (detailed list in **Supplementary Table 3**).

The Xpp category for unclassified homology groups contains some proteins that could not be modeled based on multiple sequence alignments due to insufficient available sequences at the databases, namely Xpp37, Xpp76, and Xpp77. For those proteins single sequence models were created and incorporated in IDOPS. Furthermore, an individual profile HMM was dedicated to the C-terminal region of the 3 domain pesticidal proteins.

Every profile HMM satisfies the criteria for a good model. The final subsets of sequences selected to construct the models are enough to represent the whole diversity of each group while retaining specificity. Consequently, there is no significant overlap between the matches identified by each profile HMM, as shown at the **Supplementary Table 4**. None of the pore-forming proteins from the true negative dataset was identified by IDOPS models.

The scanning of UniProtKB databases was beneficial to elucidate the selective power of each model and to determine the gathering cutoff values. A density estimate of the hit distribution against UniProt-TrEMBL and UniProt-SwissProt depicts how

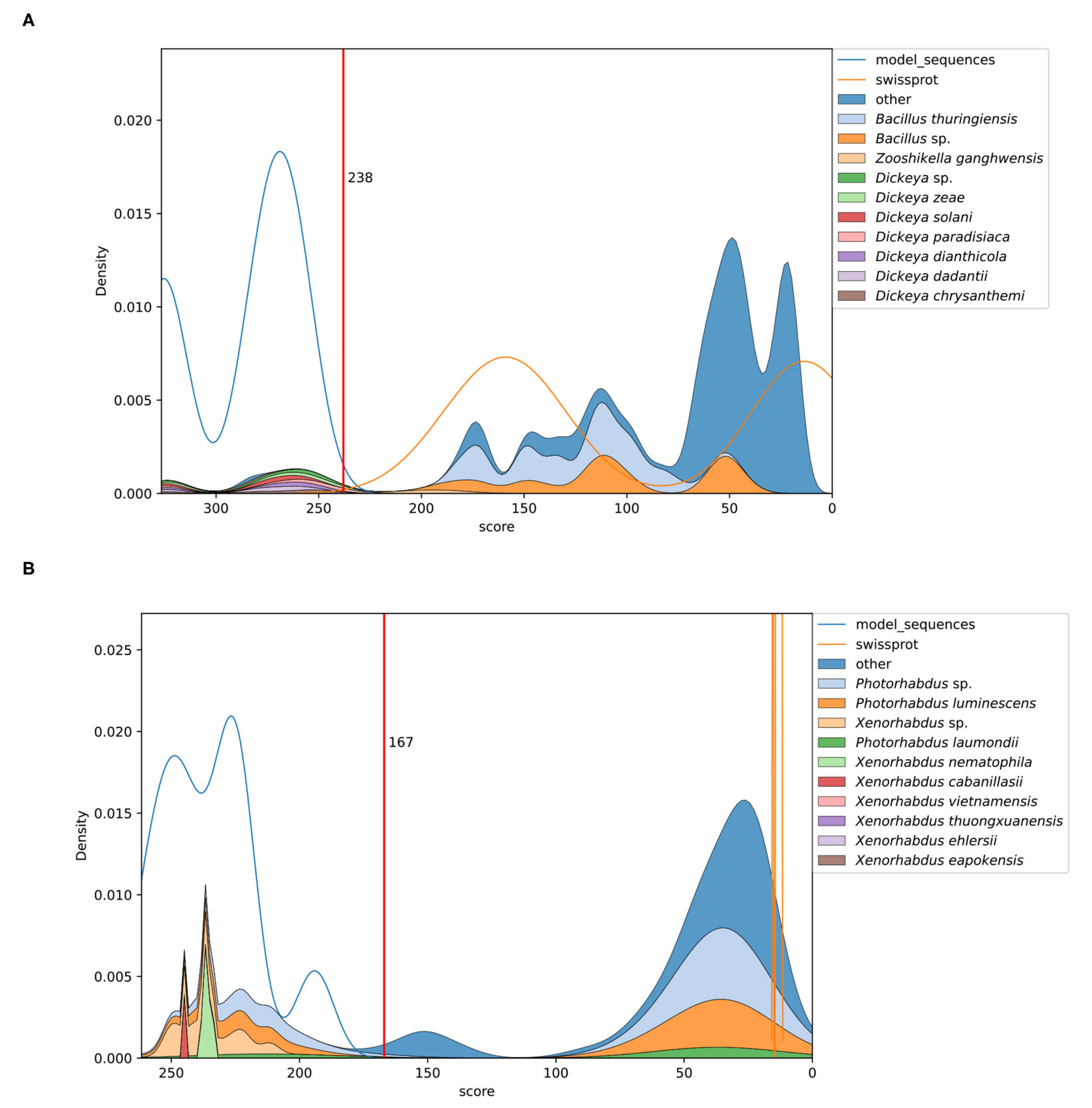

FIGURE 1 | Kernel density plots depicting distribution of model matches in UniProtKb. **(A)** The profile HMM identifies Cyt 4/5/6/7 proteins from *Dickeya* spp. within the high score range. The red line marks the curated gathering cutoff. Matches scoring below 238 correspond to related Cyt proteins from other source organisms. **(B)** A model targeting Pra sequences produced by *Photorhabdus* spp. and *Xenorhabdus* spp. Some *Yersinia* spp. proteins score directly next to the gathering cutoff.

accurately our models discriminate true members of each protein class from the whole database. In **Figure 1**, two examples are shown.

Figure 1A depicts the sequences identified by a model targeting Cyt 4/5/6/7 from *Dickeya* spp. Only proteins from *Dickeya* spp. score above the gathering cutoff and those below it correspond to related Cyt proteins produced by other organisms like *Bacillus thuringiensis*. The example in **Figure 1B** shows the matches of a model built to identify Pra proteins. Here, the high scoring sequences are clearly separated from most of the non-relevant hits. Interestingly, some entries, annotated as "uncharacterized proteins," score

TABLE 2 | Protein sequences identified by IDOPS profile HMMs in UniProtKb-TrEMBL compared with the amount of sequences at the Bacterial Pesticidal Protein Resource Center database.

	Number of protein Sequences above gathering cutoff	
Protein class	**BPPRC database**	**UniProtKB-TrEMBL**
app	10	121
cry	720	1,123
cyt	40	87
gpp	11	6
mcf	5	82
mpf	5	14
mpp	40	130
mtx	1	3
pra	3	59
prb	3	46
spp	2	482
tpp	30	52
vip	108	120
vpa	20	40
vpb	20	79
xpp	14	16

just below the gathering threshold. These are produced by members of the *Yersinia* genus. A similar distribution is observed when analyzing the matches of the toxin partner component, represented by the Prb HMM (Plots for each model are found as **Supplementary Figures 1–31**).

In addition, it is of notice the abundance of potential new pesticidal proteins identified by our approach. When the UniProtKb-TrEMBL database was scanned, a total of 2,460 protein sequences were found within the trusted score range of IDOPS profile HMMs, many of them annotated as uncharacterized or hypothetical proteins (**Table 2**).

Notably, models representing groups with few members in the BPPRC collection; like App and Spp, with 10 and 2 entries, respectively, detected plenty of potentially novel pesticidal sequences; 121 for App and 482 in the case of Spp. In other cases, no new entry is matched by our profile HMMs, this is true for Xpp76 and Xpp77 proteins, members of the Xpp group.

3.2. Comparative Analysis of Genetic Environments

In order to demonstrate the potential of IDOPS to investigate the genetic context of pesticidal proteins, we tested our tool against known reported *cry1* cassettes in plasmids of *B. thuringiensis* (**Figure 2**).

Using the sequences of eight Bt plasmids (retrieved in genbank format), IDOPS automatically generated a color coded and uniformly annotated alignment of genome regions encoding the *cry1* cassettes. In the red box, the figure shows the *cry1* cassette components (Cry1 -N-acetylmuramoyl-L-alanine amidase - K(+)/H(+) antiporter) and their genetic environment. IDOPS's output already allows the identification and grouping of the *cry1* cassette variants according to their particular elements, such as transposases from different families. Moreover, the figure is in accordance with the manually generated figure created by Fiedoruk et al. (2017), showing how IDOPS identifies, aligns, and displays the genomic surrounding of a toxin of interest.

3.3. Benchmarking

To evaluate the specificity and the sensitivity in comparison with currently implemented toxin identification tools, BtToxin_Digger and Cry_Processor, we applied benchmark tests (**Table 3**). The proteins from the BPPRC database were used as true positives and distantly related pore forming toxins sequences as a true negative dataset.

All three programs performed convincingly throughout the specificity test. None of the sequences from the true negative dataset were missannotated as a pesticidal protein by either Cry_Processor, BtToxin_Digger or IDOPS. The generalized profile HMMs of Cry_Processor returned some single domain matches. Nevertheless, those did not meet the criteria to be reported as pesticidal proteins.

Concerning the sensitivity Cry_Processor showed some discrepancy between the FD (Full Domain) and DO (Domain only) modes, as 696 and 715 out of 720 sequences were identified as 3 domain toxins, respectively (**Table 3**). In a similar way, BtToxin_Digger successfully recognized 994 of the 1,033 input sequences from the BPPRC. It failed to detect any of Spp and Vpb sequences, while identification of Cry, Vip, Mpp, Pra, and Tpp groups was incomplete (**Table 3**). IDOPS had the highest retrieval rate regarding this dataset. It recognized all but three Cry sequences above its gathering cutoff values. In the case of the missed toxins, Cry1Ca10, Cry3Bb3, and Cry11Aa2, a closer look revealed that these three sequences represent truncated toxins. The proteins sequences have been, nevertheless, recognized as hits that scored below trusted cut off. Exclusively IDOPS recovered all complete protein sequences from the tested true positive data set.

4. DISCUSSION

Here we present IDOPS, a tool to detect bacterial pesticidal protein sequences and compare their genetic environment. The power of IDOPS comes from a collection of high-quality profile hidden Markov models; each one carefully designed to represent a structural group as defined by the Bacterial Pesticidal Protein Resource Center (BPPRC) (Crickmore et al., 2020a,b). To this date, the tool comprises the most exhaustive and complete collection of models describing pesticidal proteins.

We compared IDOPS with other tools implementing profile HMMs, Cry_Processor (Shikov et al., 2020) and BtToxin_Digger (Liu et al., 2020). Neither of those recognized all of the complete sequences from the positive dataset, making the search against genomes or full databases potentially incomplete. Moreover, Cry_Processor's database is not up to date with the current BPPRC nomenclature. On the other hand, BtToxin_Digger has a greedy approach behind its profile HMMs. While this could work well for closely related and not so diverse structural groups, it is not the best option when dealing with varying sequences,

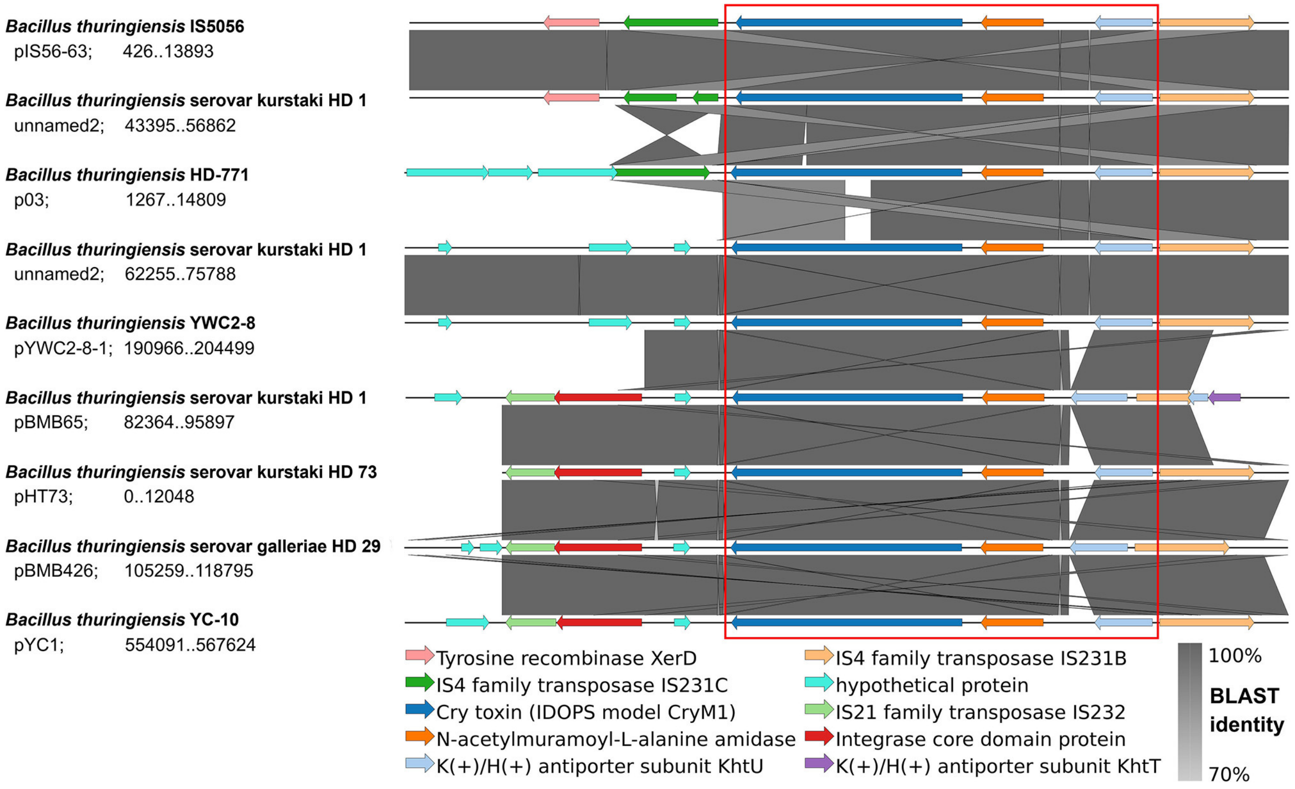
Bacillus thuringiensis IS5056
pIS56-63; 426..13893
Bacillus thuringiensis serovar kurstaki HD 1
unnamed2; 43395..56862
Bacillus thuringiensis HD-771
p03; 1267..14809
Bacillus thuringiensis serovar kurstaki HD 1
unnamed2; 62255..75788
Bacillus thuringiensis YWC2-8
pYWC2-8-1; 190966..204499
Bacillus thuringiensis serovar kurstaki HD 1
pBMB65; 82364..95897
Bacillus thuringiensis serovar kurstaki HD 73
pHT73; 0..12048
Bacillus thuringiensis serovar galleriae HD 29
pBMB426; 105259..118795
Bacillus thuringiensis YC-10
pYC1; 554091..567624
Tyrosine recombinase XerD
IS4 family transposase IS231C
Cry toxin (IDOPS model CryM1)
N-acetylmuramoyl-L-alanine amidase
K(+)/H(+) antiporter subunit KhtU
IS4 family transposase IS231B
hypothetical protein
IS21 family transposase IS232
Integrase core domain protein
K(+)/H(+) antiporter subunit KhtT
100%
BLAST identity
70%

FIGURE 2 | IDOPS genetic environment comparison. IDOPS retrieved, annotated, and aligned the genetic region surrounding *cry1* genes reported as a cassette within Bt plasmids by Fiedoruk et al. (2017). The visualization is generated by EasyFig as implemented internally in IDOPS. The reported *cry1* cassette is marked in red. Strain names, plasmid, and coordinates appear next to each row.

such as members of the Cry group or even proteins of the Xpp class, which lack in shared homology between them. IDOPS overcomes such limitations with its comprehensive collection of profile HMMs. Additionally, it provides a unique genetic context comparison feature.

Gathering cutoff values for IDOPS' models are optimized to recognize full-length proteins, thus shorter or incomplete sequences will score below it and won't be reported. In our tests, IDOPS recognized 717 from the 720 Cry sequences at the BPPRC database. A detailed look revealed that none of the three missed sequences are full-length 3 domain proteins. Cry3Bb3 presents only the InterPro signatures of the central (IPR001178) and the C-terminal domains (IPR005638). Moreover, Cry1Ca10 and Cry11Aa2 are both annotated as partial proteins with sequence lengths of 181aa and 78aa, respectively. Cry1Cb3 is a particular case, since it is not a full-length toxin but rather just the C-terminal portion of a long Cry protein, with the C-terminal (IPR005638) and domain V (IPR041587) regions. Since we developed a dedicated model for the C-terminal region of a 3 domain protein, and the model recognized the sequence as such, it was reported, but under this considerations. However, in case a researcher using IDOPS needs to retrieve the low-scoring proteins, we set up an option to disable the gathering cutoff, so even incomplete matches will be reported for further manual evaluation.

Bias within biological databases affects the creation and refinement of profile HMMs. Cry is one of the most studied pesticidal protein groups, as reflected by the amount of available sequences and their diversity, which are valuable to build rich and diverse profile HMMs. Nevertheless, there is a composition bias within the available Cry sequences. The BPPRC database contains 720 Cry entries and 276 of those correspond to Cry1 sequences. Such skewed composition may be carried over to further studies and databases. To ensure that the profile HMMs do not suffer from this bias, several rounds of model training and refinement were done to find the adequate sequences to represent each pesticidal class. Conversely, other groups, such as Mcf, Mtx, and Spp have significantly fewer representatives, sometimes single entries; this in turn makes model building a challenging task.

IDOPS' models take into consideration common properties of distinct subgroups within each pesticidal protein group. For instance, Cyt proteins that are synthesized by the *Dickeya* spp. cluster separately from the well resolved Cyt proteins of the *Bacillus* clade. They have a shorter N-terminal region and lack hemolytic activity when compared with Cyts from the *Bacillus* spp. (Soberón et al., 2013; Loth et al., 2015). Consequently, it becomes reasonable to have a distinct model to better represent this subgroup. In a similar way, the Pra and Prb proteins of *Vibrio* spp. were modeled separately from their counterparts found in *Xenorhabdus* spp. and *Photorhabdus* spp.

TABLE 3 | Benchmark of tools identifying pesticidal sequences from the Bacterial Pesticidal Protein Resource Center (BPPRC).

Structural group	Sequences at the BPPRC	IDOPS	BtToxin_digger	Cry_processor	
				full_domain	domain_only
Cry	720	716	706	696	715
App	10	10	10	N\A	
Cyt	40	40	40	N\A	
Gpp	11	11	11	N\A	
Mcf	5	5	5	N\A	
Mpf	5	5	5	N\A	
Mpp	40	40	39	N\A	
Mtx	1	1	1	N\A	
Pra	3	3	2	N\A	
Prb	3	3	3	N\A	
Spp	2	2	0	N\A	
Tpp	30	30	30	N\A	
Vip	108	108	107	N\A	
Vpa	20	20	20	N\A	
Vpb	20	20	0	N\A	
Xpp	14	14	14	N\A	

The current BPPRC database contains three entries for the Pra category, which corresponds to "*Photorhabdus* Insect-Related toxin A component" produced by *Photorhabdus luminescens* subsp. *luminescens*, *Xenorhabdus nematophila*, and *Vibrio parahaemolyticus* M0605 (Crickmore et al., 2020a). Nevertheless, it is intriguing to find sequences from *Yersinia* spp. scoring very close to the BPPRC proteins. *Yersinia* spp. shares insecticidal potential with *P. luminescens*, as homologous proteins with similar genetic arrangements have been found, perhaps as product of horizontal gene transfer (Heermann and Fuchs, 2008; Ahantarig et al., 2009; Castagnola and Stock, 2014; Hurst et al., 2016). Therefore, the sequences encoded by *Yersinia* spp. and matched by the Pra model might represent related variants of the Pra toxin. Supporting this idea, the match distribution of the Prb model shows *Yersinia* spp. proteins in a similar position, meaning this bacteria might potentially produce both the PirA and the PirB toxin components.

Besides source organisms, structural variations within each pesticidal protein group were contemplated for IDOPS' models. Cry2, Cry11, and Cry18, despite being part of the 3 domain category, lack some of the conserved blocks described for this group (Schnepf et al., 1998; Palma et al., 2014). Accordingly, a distinct model was created for the proteins of this subgroup. In a similar way, another subgroup of Cry proteins present variant and alternate versions of such blocks (Schnepf et al., 1998; de Maagd et al., 2001); for example, Cry5, Cry12, and Cry21, thus they were grouped and modeled apart.

IDOPS provides a profile HMM dedicated to the C-terminal extension of the Cry proteins. The rationale for its creation is the evidence of such sequences encoded in proximity to the short variants of *cry* genes. These proteins have homology to the C-terminal region of the long Cry toxins (de Maagd et al., 2003). A role in crystal formation, packing, and stabilization has been reported for the C-terminal extension (Naimov et al., 2006; Peng et al., 2015). Moreover, chimeric toxins made by artificial recombination of N-terminal and C-terminal regions of Cry proteins have shown increased crystal stability and toxicity (Naimov et al., 2006; Zghal et al., 2016). Therefore, with the C-terminal model, IDOPS contributes to identify independent instances of this C-terminal extension. These instances may be useful to investigate crystallization properties and toxic activity of pesticidal proteins.

IDOPS' profile HMMss were meticulously tested against a) the sequence collection at BPPRC, b) a true negative dataset of pore-forming toxins, and c) the whole UniProtKB database. IDOPS recognized all complete sequences of the true positive dataset, none of the false positives, and 2,460 further sequences with pesticidal potential from the UniProtKB database. The abundance of sequences identified at UniProtKB exposes the unexplored potential of the less investigated pesticidal groups (**Table 1**). Having such dedicated models allows to infer some aspects regarding the candidate pesticidal protein identified by IDOPS. For example, whether it is: a Cry with the conserved blocks, a member of Xpp of a specific subtype, or to which kind of Cyt it belongs. Altogether, the final models are sensitive, specific for each structural group and constitute a promising aid in the search for novel pesticidal protein sequences.

Further developments of IDOPS will include expansion of the search toward other sequence collections, such as metagenomic data. By extended scanning, novel members of the less populated groups could be detected and used to improve the current single sequences models such as Xpp37, Xpp76, and Xpp77. Moreover, as other virulence factors have been shown to support the toxic activity of pesticidal proteins, especially in *B. thuringiensis* (George and Crickmore, 2012; Malovichko et al., 2019), it may be worth to target some of these features by dedicated profile HMMs. Another course of action for IDOPS will be the implementation of the tool as a website service to facilitate its access to the scientific community without the need of local installation.

IDOPS facilitates the comparison of the genetic environment of pesticidal sequences in a systematic and reliable way. The analysis of plasmids carrying *cry1* genes automatically generated an output consistent with Fiedoruk et al. (2017) results. Comparative genomics have proven relevant to understand genetic dynamics of pesticidal proteins. For example, Lechuga et al. (2020) proposed a plasmidial origin to a chromosomally-located Cry1Ba4 only after examination and comparison of the genetic context with that of plasmid-encoded toxins. In a further extensive analysis, we used IDOPS to detect a previously unreported chromosomal *cry* cassette. By comparative genomics we discovered *cry5*, *cry10*, and *cry13* variants in a rather highly conserved genetic environment. Moreover, we consistently found a *Siphoviridae*-like prophage region in the vicinity of the cassette (Lev Hacohen et al. unpublished data). IDOPS was of great help to detect such arrangement in several Bt strains, opening intriguing evolutionary questions.

We achieved a sophisticated and comprehensive tool that provides not only detection and structural classification of pesticidal proteins, but also a feature that identifies, retrieves, and aligns the genomic context of the pesticidal sequences. IDOPS was designed in accordance to the BPPRC nomenclature system. The benchmark confirmed it has the highest sensitivity available among other toxin search tools. This combination of a highly sensitive toxin detection engine with a solid genome comparison module is, to our knowledge, unique. All considered, we created IDOPS as a tool that addresses comparative genomics of pesticidal proteins to gain novel insights on biopesticides.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

SD-V build the initial profile HMMs, performed research, analyzed data, and wrote the manuscript. ALH performed research, refined profile HMMs, and wrote the manuscript. RS performed research, coded the software tool, and wrote the manuscript. HL designed the study, performed research, evaluated results, and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

We acknowledge support by the Open Access Publication Funds of the Göttingen University.

ACKNOWLEDGMENTS

We thank Drs. Cynthia Maria Chibani and Bernd Wemheuer for intense scientific discussions. We thank Tarek Morsi for expert IT support with the G2L servers. Last but not least, we are grateful to Prof. Rolf Daniel for hosting and supporting us and giving us unlimited access to the G2L-IT infrastructure.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.664476/full#supplementary-material

REFERENCES

Adalat, R., Saleem, F., Crickmore, N., Naz, S., and Shakoori, A. R. (2017). *In vivo* crystallization of three-domain cry toxins. *Toxins* 9, 80. doi: 10.3390/toxins9030080

Ahantarig, A., Chantawat, N., Waterfield, N. R., Ffrench-Constant, R., and Kittayapong, P. (2009). Pirab toxin from *Photorhabdus asymbiotica* as a larvicide against dengue vectors. *Appl. Environ. Microbiol.* 75, 4627–4629. doi: 10.1128/AEM.00221-09

Akarsu, H., Bordes, P., Mansour, M., Bigot, D.-J., Genevaux, P., and Falquet, L. (2019). Tasmania: a bacterial toxin-antitoxin systems database. *PLoS Comput. Biol.* 15:e1006946. doi: 10.1371/journal.pcbi.1006946

Bigot, T., Temmam, S., Pérot, P., and Eloit, M. (2019). Rvdb-prot, a reference viral protein database and its hmm profiles. *F1000Research* 8, 530. doi: 10.12688/f1000research.18776.1

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al. (2020). The interpro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344–D354. doi: 10.1093/nar/gkaa977

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). Blast+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Castagnola, A., and Stock, S. P. (2014). Common virulence factors and tissue targets of entomopathogenic bacteria for biological control of lepidopteran pests. *Insects* 5, 139–166. doi: 10.3390/insects5010139

Crickmore, N., Berry, C., Panneerselvam, S., Mishra, R., Connor, T. R., and Bonning, B. (2020a). *Bacterial Pesticidal Protein Resource Center*. Available online at: https://www.bpprc.org/

Crickmore, N., Berry, C., Panneerselvam, S., Mishra, R., Connor, T. R., and Bonning, B. C. (2020b). A structure-based nomenclature for *Bacillus thuringiensis* and other bacteria-derived pesticidal proteins. *J. Invertebr. Pathol.* 107438. doi: 10.1016/j.jip.2020.107438

Crickmore, N., Zeigler, D. R., Feitelson, J., Schnepf, E., Van Rie, J., Lereclus, D., et al. (1998). Revision of the nomenclature for the bacillus thuringiensis pesticidal crystal proteins. *Microbiol. Mol. Biol. Rev.* 62, 807–813.

de Maagd, R. A., Bravo, A., Berry, C., Crickmore, N., and Schnepf, H. E. (2003). Structure, diversity, and evolution of protein toxins from spore-forming entomopathogenic bacteria. *Ann. Rev. Genet.* 37, 409–433. doi: 10.1146/annurev.genet.37.110801.143042

de Maagd, R. A., Bravo, A., and Crickmore, N. (2001). How *Bacillus thuringiensis* has evolved specific toxins to colonize the insect world. *Trends in Genet.* 17, 193–199. doi: 10.1016/S0168-9525(01)02237-5

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics* 14, 755–763. doi: 10.1093/bioinformatics/14.9.755

Elleuch, J., Jaoua, S., Ginibre, C., Chandre, F., Tounsi, S., and Zghal, R. Z. (2016). Toxin stability improvement and toxicity increase against dipteran and lepidopteran larvae of *Bacillus thuringiensis* crystal protein cry2aa. *Pest Manag. Sci.* 72, 2240–2246. doi: 10.1002/ps.4261

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Fayad, N., Kambris, Z., El Chamy, L., Mahillon, J., and Awad, M. K. (2020). A novel anti-dipteran *bacillus thuringiensis* strain: Unusual cry toxin genes in a highly dynamic plasmid environment. *Appl. Environ. Microbiol.* 87:e02294-20. doi: 10.1128/AEM.02294-20

Fiedoruk, K., Daniluk, T., Mahillon, J., Leszczynska, K., and Swiecicka, I. (2017). Genetic environment of cry1 genes indicates their common origin. *Genome Biol. Evol.* 9, 2265–2275. doi: 10.1093/gbe/evx165

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., et al. (2007). The pfam protein families database. *Nucleic Acids Res.* 36(Suppl. 1):D281–D288. doi: 10.1093/nar/gkm960

George, Z., and Crickmore, N. (2012). "*Bacillus thuringiensis* applications in agriculture," in *Bacillus thuringiensis Biotechnology* (Dordrecht: Springer), 19–39. doi: 10.1007/978-94-007-3021-2_2

Gonzalez, M., Bischofberger, M., Pernot, L., Van Der Goot, F., and Freche, B. (2008). Bacterial pore-forming toxins: the (w) hole story? *Cell. Mol. Life Sci.* 65, 493–507. doi: 10.1007/s00018-007-7434-y

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., and Beck, E. (2012). Tigrfams and genome properties in 2013. *Nucleic Acids Res.* 41, D387–D395. doi: 10.1093/nar/gks1234

Heermann, R., and Fuchs, T. M. (2008). Comparative analysis of the *Photorhabdus luminescens* and the *Yersinia enterocolitica* genomes: uncovering candidate genes involved in insect pathogenicity. *BMC Genomics* 9:40. doi: 10.1186/1471-2164-9-40

Hurst, M. R., Beattie, A., Altermann, E., Moraga, R. M., Harper, L. A., Calder, J., et al. (2016). The draft genome sequence of the *Yersinia entomophaga* entomopathogenic type strain mh96t. *Toxins* 8, 143. doi: 10.3390/toxins8050143

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). Interproscan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Jouzani, G. S., Valijanian, E., and Sharafi, R. (2017). *Bacillus thuringiensis*: a successful insecticide with new environmental features and tidings. *Appl Microbiol. Biotechnol.* 101, 2691–2711. doi: 10.1007/s00253-017-8175-y

Kachhawa, D. (2017). Microorganisms as a biopesticides. *J. Entomol. Zool. Stud.* 5, 468–473.

Khasdan, V., Sapojnik, M., Zaritsky, A., Horowitz, A. R., Boussiba, S., Rippa, M., et al. (2007). Larvicidal activities against agricultural pests of transgenic *Escherichia coli* expressing combinations of four genes from *Bacillus thuringiensis*. *Arch. Microbiol.* 188, 643–653. doi: 10.1007/s00203-007-0285-y

Koni, P., and Ellar, D. (1993). Cloning and characterization of a novel *Bacillus thuringiensis* cytolytic delta-endotoxin. *J. Mol. Biol.* 229, 319–327. doi: 10.1006/jmbi.1993.1037

Lechuga, A., Lood, C., Salas, M., van Noort, V., Lavigne, R., and Redrejo-Rodríguez, M. (2020). Completed genomic sequence of *Bacillus thuringiensis* her1410 reveals a cry-containing chromosome, two megaplasmids, and an integrative plasmidial prophage. *G3* 10, 2927–2939. doi: 10.1534/g3.120.401361

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Liu, H., Zheng, J., Yu, Y., Ye, W., Peng, D., and Sun, M. (2020). Bttoxin_digger: a comprehensive and high-throughput pipeline for mining toxin protein genes from *Bacillus thuringiensis*. *bioRxiv*. doi: 10.1101/2020.05.26.114520

Loth, K., Costechareyre, D., Effantin, G., Rahbé Y., Condemine, G., Landon, C., and Da Silva, P. (2015). New cyt-like δ-endotoxins from *Dickeya dadantii*: structure and aphicidal activity. *Sci. Rep.* 5, 1–10. doi: 10.1038/srep08791

Malovichko, Y. V., Nizhnikov, A. A., and Antonets, K. S. (2019). Repertoire of the bacillus thuringiensis virulence factors unrelated to major classes of protein toxins and its role in specificity of host-pathogen interactions. *Toxins* 11, 347. doi: 10.3390/toxins11060347

Mnif, I., and Ghribi, D. (2015). Potential of bacterial derived biopesticides in pest management. *Crop Protect.* 77, 52–64. doi: 10.1016/j.cropro.2015.07.017

Naimov, S., Martens-Uzunova, E., Weemen-Hendriks, M., Dukiandjiev, S., Minkov, I., and de Maagd, R. A. (2006). Carboxy-terminal extension effects on crystal formation and insecticidal properties of colorado potato beetle-active bacillus thuringiensis δ-endotoxins. *Mol. Biotechnol.* 32, 185–196. doi: 10.1385/MB:32:3:185

Palma, L., Muñoz, D., Berry, C., Murillo, J., and Caballero, P. (2014). *Bacillus thuringiensis* toxins: An overview of their biocidal activity. *Toxins* 6, 3296–3325. doi: 10.3390/toxins6123296

Peng, D.-h., Pang, C.-y., Wu, H., Huang, Q., Zheng, J.-S., and 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.* 5:8291. doi: 10.1038/srep08291

Ruffner, B., Péchy-Tarr, M., Höfte, M., Bloemberg, G., Grunder, J., Keel, C., and Maurhofer, M. (2015). Evolutionary patchwork of an insecticidal toxin shared between plant-associated pseudomonads and the insect pathogens *Photorhabdus* and *Xenorhabdus*. *BMC Genomics* 16:609. doi: 10.1186/s12864-015-1763-2

Ruiu, L. (2018). Microbial biopesticides in agroecosystems. *Agronomy* 8, 235. doi: 10.3390/agronomy8110235

Schnepf, E., Crickmore, N., Van Rie, J., Lereclus, D., Baum, J., Feitelson, J., et al. (1998). *Bacillus thuringiensis* and its pesticidal crystal proteins. *Microbiol. Mol. Biol. Rev.* 62, 775–806. doi: 10.1128/MMBR.62.3.775-806.1998

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Shao, Z., Liu, Z., and Yu, Z. (2001). Effects of the 20-kilodalton helper protein on cry1ac production and spore formation in *Bacillus thuringiensis*. *Appl. Environ. Microbiol.* 67, 5362–5369. doi: 10.1128/AEM.67.12.5362-5369.2001

Shikov, A. E., Malovichko, Y. V., Skitchenko, R. K., Nizhnikov, A. A., and Antonets, K. S. (2020). No more tears: mining sequencing data for novel bt cry toxins with cryprocessor. *Toxins* 12, 204. doi: 10.3390/toxins12030204

Sievers, F., and Higgins, D. G. (2018). Clustal omega for making accurate alignments of many protein sequences. *Protein Sci.* 27, 135–145. doi: 10.1002/pro.3290

Soberón, M., López-Díaz, J. A., and Bravo, A. (2013). Cyt toxins produced by *Bacillus thuringiensis*: a protein fold conserved in several pathogenic microorganisms. *Peptides* 41:87–93. doi: 10.1016/j.peptides.2012.05.023

Söding, J. (2005). Protein homology detection by hmm–hmm comparison. *Bioinformatics* 21, 951–960. doi: 10.1093/bioinformatics/bti125

Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998). Pfam: multiple sequence alignments and hmm-profiles of protein domains. *Nucleic Acids Res.* 26, 320–322. doi: 10.1093/nar/26.1.320

Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039

UniProt Consortium (2019). Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Vílchez, S. (2020). Making 3d-cry toxin mutants: Much more than a tool of understanding toxins mechanism of action. *Toxins* 12, 600. doi: 10.3390/toxins12090600

Wang, Z., Wang, K., Bravo, A., Soberón, M., Cai, J., Shu, C., et al. (2020). Coexistence of cry9 with the vip3a gene in an identical plasmid of *Bacillus thuringiensis* indicates their synergistic insecticidal toxicity. *J. Agric. Food Chem.* 68, 14081–14090. doi: 10.1021/acs.jafc.0c05304

Waterfield, N. R., Bowen, D. J., Fetherston, J. D., Perry, R. D., et al. (2001). The tc genes of *Photorhabdus*: a growing family. *Trends Microbiol.* 9, 185–191. doi: 10.1016/S0966-842X(01)01978-3

Zghal, R. Z., Elleuch, J., Ali, M. B., Darriet, F., Rebaï, A., Chandre, F., et al. (2016). Towards novel cry toxins with enhanced toxicity/broader: a new chimeric cry4ba / cry1ac toxin. *Appl. Microbiol. Biotechnol.* 101, 113–122. doi: 10.1007/s00253-016-7766-3

Zheng, J., Gao, Q., Liu, L., Liu, H., Wang, Y., Peng, D., et al. (2017). Comparative genomics of *Bacillus thuringiensis* reveals a path to specialized exploitation of multiple invertebrate hosts. *MBio* 8:e00822-17. doi: 10.1128/mBio.00822-17

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.