AEM
Journals.ASM.org

Mining New Crystal Protein Genes from *Bacillus thuringiensis* on the Basis of Mixed Plasmid-Enriched Genome Sequencing and a Computational Pipeline

Weixing Ye,[a] Lei Zhu,[a] Yingying Liu,[a] Neil Crickmore,[b] Donghai Peng,[a] Lifang Ruan,[a] and Ming Sun[a]

State Key Laboratory of Agricultural Microbiology, College of Life Science and Technology, Huazhong Agricultural University, Wuhan, China,[a] and Department of Biochemistry, School of Life Sciences, University of Sussex, Falmer, Brighton, United Kingdom[b]

We have designed a high-throughput system for the identification of novel crystal protein genes (*cry*) from *Bacillus thuringiensis* strains. The system was developed with two goals: (i) to acquire the mixed plasmid-enriched genomic sequence of *B. thuringiensis* using next-generation sequencing biotechnology, and (ii) to identify *cry* genes with a computational pipeline (using BtToxin_scanner). In our pipeline method, we employed three different kinds of well-developed prediction methods, BLAST, hidden Markov model (HMM), and support vector machine (SVM), to predict the presence of Cry toxin genes. The pipeline proved to be fast (average speed, 1.02 Mb/min for proteins and open reading frames [ORFs] and 1.80 Mb/min for nucleotide sequences), sensitive (it detected 40% more protein toxin genes than a keyword extraction method using genomic sequences downloaded from GenBank), and highly specific. Twenty-one strains from our laboratory's collection were selected based on their plasmid pattern and/or crystal morphology. The plasmid-enriched genomic DNA was extracted from these strains and mixed for Illumina sequencing. The sequencing data were *de novo* assembled, and a total of 113 candidate *cry* sequences were identified using the computational pipeline. Twenty-seven candidate sequences were selected on the basis of their low level of sequence identity to known *cry* genes, and eight full-length genes were obtained with PCR. Finally, three new *cry*-type genes (primary ranks) and five *cry* holotypes, which were designated *cry8Ac1*, *cry7Ha1*, *cry21Ca1*, *cry32Fa1*, and *cry21Da1* by the *B. thuringiensis* Toxin Nomenclature Committee, were identified. The system described here is both efficient and cost-effective and can greatly accelerate the discovery of novel *cry* genes.

Bacillus thuringiensis is a ubiquitous Gram-positive, spore-forming bacterium that produces parasporal crystals during the stationary phase of its growth cycle (42). The crystals comprise one or more crystal proteins (encoded by *cry* or *cyt* genes) that show specific toxicity against several orders of insects, including Lepidoptera, Diptera, Coleoptera, Hymenoptera, Homoptera, Orthoptera, and Mallophaga, and also against nematodes, mites, and protozoa (19, 20, 42). A portion of *B. thuringiensis* strains also secrete vegetative insecticidal proteins (VIPs) showing activity against Lepidopteran insect larvae (18). Generally, *cry* genes are located on plasmids, while few of them are reported to reside on the chromosome (31). Since the cloning of the first *cry* gene by Schnepf and Whiteley (43), more than 700 *B. thuringiensis* toxin genes (including 586 *cry* genes in 70 primary ranks, 96 *vip* genes in 4 primary ranks, and 34 *cyt* genes in 3 primary ranks; http://www.lifesci.sussex.ac.uk/home/Neil_Crickmore/Bt/) have been isolated and classified according to the nomenclature system described by Crickmore et al. (14).

The insecticidal activity of Cry proteins has led to the global development of bioinsecticides based upon *B. thuringiensis* for pest control (42). The bacterium is also a key source of genes for transgenic expression to provide pest resistance in plants (6, 26). However, the continuous use of *B. thuringiensis* products leads to resistance being evolved by insects (1, 22, 47, 49). Several strategies, such as the use of multiple toxins (9, 42, 50), spatial or temporal refugia (2, 27), and high or ultrahigh doses (27), have been employed to delay insect resistance to transgenic plants. The search for novel toxins with high toxicity is considered one of the major approaches to counter the potential resistance evolved by insects as well as in developing products against a wider spectrum of insect pests.

A series of approaches have been utilized for isolating novel *cry* genes, such as PCR, initially used by Carozzi et al. (10), for predicting the insecticidal activity of previously uncharacterized *B. thuringiensis* strains, which was followed by variations such as PCR hybridization (29), PCR-RFLP (restriction fragment length polymorphism) (32), E-PCR (exclusive PCR) (28), and PCR-SSCP (PCR-single-stranded conformation polymorphism profiling) (34). The construction of *B. thuringiensis* DNA libraries in *Escherichia coli*, followed by screening by Western blotting (36, 43) or a hybridization-based method (5, 30, 33, 35), or the development of DNA libraries in an acrystaliferous mutant of *B. thuringiensis* followed by microscopic observation and/or SDS-polyacrylamide gel (SDS-PAGE) detection of expressed genes in our laboratory (23) have also been used to detect novel *cry* protein genes.

In all of these methods, PCR-based systems are the most widely used for the identification of novel *cry* genes (7, 28, 37, 48). The design of these systems is based on five conserved blocks originally reported by Hofte and Whiteley (26). There are several limitations to these systems. For instance, all of these systems focus on three-

Received 5 February 2012 Accepted 23 April 2012
Published ahead of print 27 April 2012
Address correspondence to Ming Sun, m98sun@mail.hzau.edu.cn.
Supplemental material for this article may be found at http://aem.asm.org/.

doi:10.1128/AEM.00340-12

domain *cry* genes, and most of them were limited to finding *cry* genes with sufficiently high sequence similarity to the primers used; thus, few of them were reported to be able to identify novel *cry* genes (with less than 45% amino acid sequence identity to the known *cry* genes) (37). Also, none of these systems are able to obtain the full-length *cry* gene sequences. The library-based methods are time-consuming and laborious. A general screening strategy for isolating *cry* genes therefore is required to identify more diverse sequences.

Recently, next-generation sequencing technology has been employed for the discovery of new *cry* genes (41). The major advantage of this biotechnology is that it provides a great deal of genomic data efficiently. Two issues remain to be resolved for such a strategy for the identification of new *cry* genes. One is that the average cost for identifying a *cry* gene is much higher than that of a PCR-based strategy, and the other is that no bioinformatics tool is publicly available for predicting *cry* genes from genomic sequences. A large number of protein classification algorithms are available, such as the BLAST method using pairwise local alignments to measure sequence similarity (3), the hidden Markov model (HMM) method based on multiple alignments generated by a statistical profile HMM (16), and the support vector machine (SVM) (15) method, which transforms protein sequences into fixed-length feature vectors. They represent three different kinds of protein prediction algorithms.

In this study, we describe a system for isolating new *cry* genes by combining mixed plasmid-enriched genome sequencing and a computational pipeline. The system was validated by using 21 *B. thuringiensis* strains from our laboratory. Finally, we identified three new *cry*-type genes (primary ranks) and five *cry*-holotype genes (>45% amino acid sequence identity to the known *cry* genes).

MATERIALS AND METHODS

Strain selection. Most *cry* genes are located on large plasmids (31). For the purpose of finding novel *cry* genes efficiently, strains that harbor abundant plasmids and could produce parasporal inclusions were our preferred strains. We screened more than 100 strains, assessing their plasmid profile and/or parasporal inclusion formation. Finally, 21 candidate *B. thuringiensis* strains were selected for further analysis.

Plasmid-enriched genomic DNA preparation and Illumina sequencing. These selected strains were subjected to plasmid-enriched genomic DNA extraction by incubating the cells with lysozyme (20 mg/ml) in TE (50 mM Tris base, 10 mM EDTA, 20% sucrose, pH 7.5) at 37°C, with shaking at 75 rpm for 2.5 h. Samples were then subjected to alkaline lysis (39) and further purified through ultracentrifugation in the presence of cesium chloride and ethidium bromide (40). After ultracentrifugation, the plasmid DNA (the closed circular plasmid DNA) was identified and withdrawn slowly from the tube. Ethidium bromide was removed from the solution of DNA by repeated extraction with organic solvents, and cesium chloride was removed by ethanol precipitation. The DNA precipitates were then washed twice with 70% ethanol and evaporated at room temperature. Finally, the DNA was dissolved in 400 μl of Tris-EDTA (pH 7.5) buffer and analyzed using pulsed-field gel electrophoresis to determine the concentration. The total plasmid-enriched genomic DNA from 21 candidate *B. thuringiensis* strains then were mixed together in equal amounts for sequencing.

A library for Illumina paired-end sequencing was prepared from 5 μg mixed plasmid-enriched DNA using a paired-end DNA sample preparation kit (PE-102-1001; Illumina Inc.). The DNA was fragmented by hydrodynamic shearing to generate <800-bp fragments. For end repair and phosphorylation, sheared DNA was purified using a QIAquick PCR purification kit (28104; Qiagen). The end-repaired DNA was A-tailed, and adaptors were ligated according to the manufacturer's instructions. The products of this ligation reaction were size selected by agarose gel electrophoresis and purified. The 5′ adaptor extension and enrichment of the library were performed using 10 PCR cycles with primers PE1.0 and PE2.0, which were supplied by Illumina. The library was finally purified using a QIAquick PCR purification kit and adjusted to an appropriate concentration. The flow cell was prepared according to the manufacturer's instructions using a paired-end cluster generation kit (PE-103-1001; Illumina Inc.) and a Cluster Station. Sequencing reactions were performed on a 1G GA2 equipped with a paired-end module (Illumina Inc.). *De novo* assembly was done by ABySS (45) using the following parameters: kmer (a parameter of ABySS software), 25; and *n*, 10.

Computational pipeline construction. To facilitate the identification of *cry* gene processes, a computational pipeline named BtToxin_scanner was constructed. The program consists of BioPerl software modules (46) and freely available third-party software, including a stand-alone BLAST application (8), HMMER 3.0 (17), and LIBSVM 2.91 (12). The *cry* database and the background database were integrated into the pipeline. Figure 1 shows the experimental strategy. The details are below.

Preprocessing of input sequences. Based on different types of sequence input, corresponding modules would be called to convert the input sequences into protein sequences. For nucleotide sequences obtained by next-generation sequencing technology, a six-frame translation module was employed to find all possible protein sequences. We did not use a direct open reading frame prediction module, since it might lose dozens of potential *cry* sequences.

Length filter. Short sequences did not provide any information for the cloning of *cry* genes, thus sequences with lengths of less than 115 amino acids were eliminated.

***Cry* candidate prediction module.** *cry* sequences were predicted with three modules: BLAST, HMM, and SVM. The BLAST database was constructed with the data set manually maintained by Crickmore et al. (http://www.lifesci.sussex.ac.uk/home/Neil_Crickmore/Bt/) (*B. thuringiensis* toxins are available at http://bcam.hzaubmb.org/BtToxin_scanner/), and the expected cutoff value was set to 1e−25. The HMM model (Cry.hmm, Cyt.hmm, and Vip.hmm; all available at http://bcam.hzaubmb.org/BtToxin_scanner/) of Cry proteins was built with the HMMER 3.0 package, and the expected cutoff value was set to 1e−10. The SVM classifier for Cry proteins was developed with amino acid and dipeptide composition using LIBSVM 2.91 (12). We implemented machine learning with radial basis function (RBF) for the training of various models. The training models were optimized for best performance by adjusting the kernel parameter gamma (G = 100) and the regularization parameter C (C = 100) (the SVM model is available at http://bcam.hzaubmb.org/BtToxin_scanner/). Any sequences that passed either of the prediction modules were considered candidate Cry proteins. Candidate Vip protein sequences and Cyt protein sequences were predicted only with the HMM prediction module, and the expected cutoff value was set to 1e−10.

Background elimination. To reduce the number of false-positive Cry protein sequences that would arise as a result of using all three prediction methods, a further background elimination step was employed by screening candidate Cry proteins against the background database and eliminating those proteins with significant similarity (less than 1e−30) (designated the background data set; available at http://bcam.hzaubmb.org/BtToxin_scanner/). The final results become available as a detailed prediction report along with the corresponding protein sequences.

Gene cloning, expression, and microscopic observation. Three strategies were applied for the cloning of novel *cry* genes from pipeline outputs. For candidate *cry* sequences with an intact promoter and terminator, they were directly amplified by PCR and then were ligated to the *Escherichia coli-B. thuringiensis* shuttle vector pHT304 (4). For those that only contained full-length *cry* genes, the entire genes were amplified, and then they were ligated to vector pBMBL (a plasmid derived from pHT304 and con-

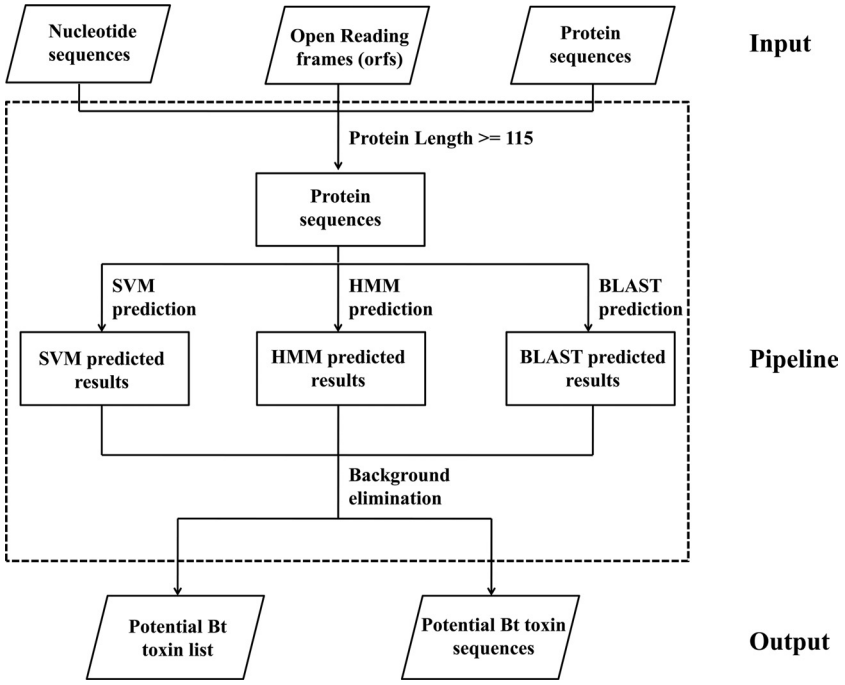

FIG 1 Workflow for computational pipeline construction.

taining a promoter and terminator from *cry1Ac10*; unpublished data). For candidate *cry* sequences with only partial segments, either a manual assembly (the contig was searched against the original reads, and those reads that could extend the contig were then extracted and assembled to the contig) or reverse PCR (Fig. 2) followed by sequencing was performed to obtain the remaining coding sequences. After that, recombinant vectors containing the amplified *cry* genes were introduced by electroporation (39) into an acrystaliferous strain, BMB171 (25), for expression. Parasporal inclusions were observed by phase-contrast microscopy, and the major protein bands were detected by SDS-PAGE analysis.

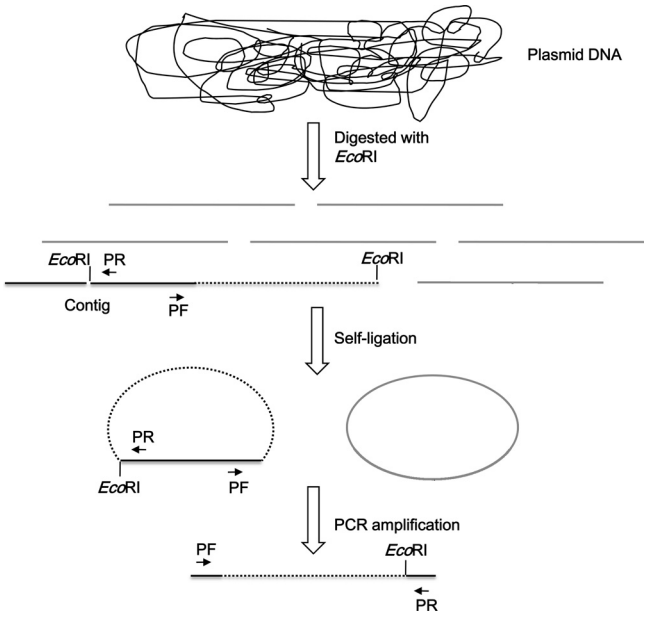

FIG 2 Schematic diagram of primer design for reverse PCR. The solid dark line represents the contig that needs to be extended, and the dotted line represents DNA downstream of the cloned *cry* gene. PF means forward primer, and PR means reverse primer. The plasmid DNA was digested completely by a restriction enzyme (such as EcoRI), and the DNA sample was self ligated overnight. The region downstream of the cloned gene was amplified using the PR and PF primers.

Nucleotide sequence accession numbers. The nucleotide sequences published in this paper have been submitted to GenBank and assigned accession numbers JF521572 (*cry8Ac1*), JF521575 (*cry7Ha1*), JF521577 (*cry21Ca1*), JF521578 (*cry21Da1*), JF521580 (*cry32Fa1*), JF521582 (Toxin6), JF521583 (Toxin7), and JF521584 (Toxin8).

RESULTS

Construction and validation of the computational pipeline for predicting *cry* genes. To identify *cry* genes from genomic sequences efficiently, a computational pipeline named BtToxin_scanner was constructed as described in Materials and Methods (Fig. 1). The performance of BtToxin_scanner was evaluated using two data sets downloaded from GenBank; one includes genomic sequences of *B. thuringiensis*, which contains a large number of *cry* genes, and the other includes genomic sequences of *Bacillus cereus*, which is supposed to be free of *cry* genes.

BtToxin_scanner shows high sensitivity using data set A with genomic sequences from *B. thuringiensis*. We tested whether BtToxin_scanner was able to identify *B. thuringiensis* toxins from genomic sequences by using data set A, which contained 3 completed genome sequences, 15 whole-genome shotgun (WGS) sequences, and 24 complete plasmid sequences from *B. thuringiensis* (see Table S1 in the supplemental material). We identified 85 candidates from this data set with an average process speed of 1.03 Mb/min. These sequences were downloaded from GenBank, thus they contained annotation information. We extracted protein sequences whose annotation related to *B. thuringiensis* toxins and compared it to the result obtained with BtToxin_scanner. It was found that BtToxin_scanner correctly identified all of the *B. thuringiensis*

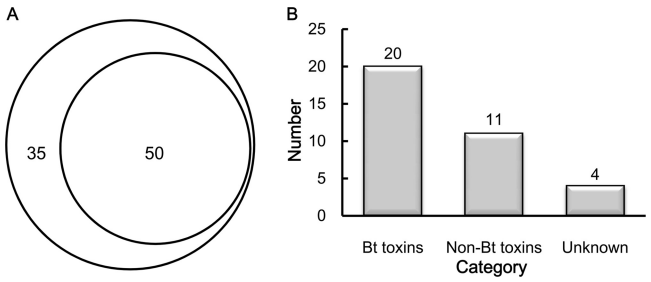

FIG 3 Comparison of BtToxin_scanner result and keyword extraction results. (A) Venn diagram of comparison of BtToxin_scanner result and keyword extraction result. Inner circle, protein number obtained with BtToxin_scanner; outer circle, protein number obtained with keyword extraction. (B) Categories for the 35 proteins identified only by BtToxin_scanner. Bt toxins indicates *B. thuringiensis* toxin proteins correctly identified by BtToxin_scanner but missing from the keyword extraction result. Non-Bt toxins indicates non-*B. thuringiensis* toxin proteins identified by BtToxin_scanner. Unknown indicates proteins identified by BtToxin_scanner with unknown function.

toxins in the keyword extraction list. It also identified an additional 35 candidates (Fig. 3A), among which 20 were considered true positives, 4 of them were marked as unknown due to a lack of homology to any proteins from GenBank, and the remaining 11 were identified as false positives (see Table S2 in the supplemental material).

The WGS sequences in data set A were obtained with a run of 454 sequencing technology, with the average *cry* gene number per run computed as 2.6. We did not include 3 complete genome sequences of *B. thuringiensis* strain BMB171 (25), *B. thuringiensis* strain Al Hakam (11), and *B. thuringiensis* serovar *konkukian* strain 97-27 (24), since these strains did not produce a parasporal crystal.

BtToxin_scanner shows high specificity using data set B with genomic sequences from *B. cereus*. Since BtToxin_scanner integrated the three prediction methods, it was important to assess how many false positives were produced. Thus, we evaluated the performance of BtToxin_scanner on data set B (see Table S3 in the supplemental material), which includes 190,062 protein sequences from *B. cereus* WGS sequences. We identified 11 candidates from data set B with an average process speed of 1.02 Mb/min (protein sequences). These sequences were analyzed by searching against GenBank, and it was found that they contained 6 false positives, 1 unknown protein, 1 Cry protein, and 3 Vip proteins (see Table S4 in the supplemental material).

Mining *cry* genes from genome sequencing of mixed plasmid-enriched total DNAs from 21 *B. thuringiensis* strains. As described above, BtToxin_scanner was able to identify *B. thuringiensis* toxins from genomic sequence efficiently. Thus, we focused on acquiring large numbers of genomic sequences with abundant *B. thuringiensis* toxin sequences in an efficient way.

Selection of strains. The selection of the 21 working strains was based on a combination of the plasmid pattern, the formation of parasporal inclusion, and the major protein profile detected by SDS-PAGE.

Acquisition of genomic sequences. Plasmid-enriched genomic DNA was extracted and mixed as described in Materials and Methods and then loaded onto the Illumina GA2 sequencer. The raw sequence data were *de novo* assembled, thus generating the mixed plasmid-enriched genomic sequence data.

A total of 6,070,863 paired-end reads of 100 bp were generated, with an average length of inserts of paired-end reads at 174 bp. All short paired-end reads generated by the Illumina genome analyzer were inputted into the *de novo* assembly software ABySS (45). A total of 9,846 contigs were produced with a predefined cutoff rate (N50, 1,365 bp; maximum contig size, 58,460 bp). The total size of the produced contigs was 9,884,426 bp, with 2,365 contigs having a length greater than 1 kbp (Table 1) (mixed plasmid-enriched genomic sequences are available at http://bcam.hzaubmb.org/BtToxin_scanner/).

Prediction of candidate *cry* gene sequences. Sequence data generated from the mixed plasmid-enriched genome sequencing project was analyzed with BtToxin_scanner, and 143 candidates were identified with an average process speed of 1.80 Mb/min (nucleotide sequences). (The detail report and the sequence file of mixed plasmid-enriched genomic sequence prediction results are available at http://bcam.hzaubmb.org/BtToxin_scanner/.) Among those 143 candidates, there were 113 candidate *cry* sequences, 23 candidate *vip* sequences, and 7 candidate *cyt* sequences. Among the 113 candidate *cry* sequences, 48 of them showed highest-hit identities of less than 45% to the known *cry* genes, and 36 hits showed identities of between 45 and 78%. The BLAST module identified 89 *cry* hits, and the level of identity of those hits to the known *cry* genes ranged from 27 to 100%. *cry32*-type hits were the most frequently found among the 89 sequences analyzed, with a frequency of 19%, followed by the *cry1*-type hits, with a frequency of 18%. These results indicated the existence of multiple potential novel *cry* genes.

Sequence analysis and characterization of full-length *cry* genes. Based on the sequence identity to known *cry* genes and sequence length, we chose 27 sequences from the BtToxin_scanner prediction results. We used different strategies to clone these novel *cry* genes as described in Materials and Methods. Finally, 8 of them were obtained with full-length sequences. The recombinant plasmids were resequenced and, as expected, all of them shared 100% amino acid sequence identity with the original sequencing data. Their sequences were submitted to the *Bacillus thuringiensis* delta-endotoxin nomenclature committee. Five of them have received official names (Cry8Ac1, Cry7Ha1, Cry21Ca1, Cry32Fa1, and Cry21Da1), while the remaining 3 were considered novel, since they shared less than 45% amino acid sequence iden-

TABLE 1 Summary of mixed plasmid-enriched genomic sequence reads and contigs

Property	Value
Total no. of paired-end reads	6,070,863
Total length (Mb)	1,214
Quality score of >20 (%)	78
Total no. of contigs	9,846
N_{50}[a] (bp)	1,365
N_{90}[b] (bp)	356
Total length (bp)	9,884,426
Maximum length (bp)	58,460
Minimum length (bp)	300
No. of contigs longer than 1 kbp	2,365
GC content (%)	33.54

[a] N_{50}, the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50% of the assembly.
[b] N_{90}, the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 90% of the assembly.

TABLE 2 Description of novel crystal proteins from BtToxin_scanner prediction

Toxin ID	Size (aa)	Name[a]	Best hit	
			Name[b]	Identity (%)
Toxin1	1,225	Cry8Ac1	Cry8Ab1	83
Toxin2	1,160	Cry7Ha1	Cry7Ba1	66
Toxin3	1,301	Cry21Ca1	Cry21Ba1	53
Toxin4	1,267	Cry32Fa1	Cry32Aa1	45
Toxin5	1,289	Cry21Da1	Cry21Ba2	48
Toxin6	802	Novel	Cry41Aa1	43
Toxin7	1,262	Novel	Cry21Ba1	40
Toxin8	802	Novel	Cry31Aa6	31

[a] Official name received from the *Bacillus thuringiensis* Toxin Gene Nomenclature Committee. Novel indicates protein sequences with less than 45% sequence identity to known Cry proteins.
[b] Name from the *Bacillus thuringiensis* toxin nomenclature website (http://www.lifesci.sussex.ac.uk/home/Neil_Crickmore/Bt/).

tity to the known *cry* proteins. Basic information on these selected *cry* genes is listed in Table 2, and the complete sequences of these proteins are provided as Data set S1 in the supplemental material.

Cry8Ac1, Cry7Ha1, Cry21Ca1, Cry32Fa1, Cry21Da1, and Toxin7 sequences were considered to constitute complete toxins. Toxin6 protein and Toxin8 protein sequences were considered to constitute naturally truncated proteins, as both of them contained only 802 amino acids. Interestingly, the C-terminal sequence of Toxin6 is similar to that of *Clostridium botulinum* hemagglutinin HA-17 (21). A BLAST search against GenBank revealed that the Toxin8 protein shares 67% sequence identity to the 83-kDa crystal protein from *B. cereus* strain AH603, which is not included in the *B. thuringiensis* Toxin Nomenclature Committee system.

All of these toxin genes were ligated into pHT304 and used to transform the acrystaliferous strain BMB171 for expression. One of them (*cry8Ac1*) could produce bipyramidal parasporal inclusion (Fig. 4A), and SDS-PAGE analysis revealed the correct size with a predicted molecular mass of 138 kDa (Fig. 4C). The *cry8Ac1* gene was amplified from *B. thuringiensis* strain Sbt006. Sbt006 forms a spherical parasporal inclusion and a >140-kDa protein band as detected by SDS-PAGE (Fig. 4B and C). The difference in protein size and parasporal inclusion shape between the recombinant strain and Sbt006 implies that *cry8Ac1* is cryptic in Sbt006. The expression of the other seven crystal genes did not succeed in BMB171.

Construction of *cry* gene recognition web server. To make the pipeline method accessible for the broader biological community, we implemented it as a user-friendly web server accessible at http://bcam.hzaubmb.org/BtToxin_scanner. The common gateway interface (CGI) script for BtToxin_scanner was written using PERL version 5.10. The server was installed on an Apache server (2.2.16) in a Linux (Ubuntu 10.10 server) environment. The server accepts three kinds of sequence (proteins, open reading frames [ORFs], and nucleotide sequences). Users can submit sequences for prediction using file uploading. It might take a few minutes (depending on the sequence file size, the average processing speed was 1.02 Mb/min for protein sequences and ORFs and 1.8 Mb/min for nucleotide sequences) to complete the whole job. The output for each run is displayed in a user-friendly table format as well as two downloadable links, one for the detailed information about the prediction and the other for the protein sequences.

The underlying *cry* database and the background database are updated on a regular basis. The profile HMMs and the SVM model are also updated manually as the new Cry protein members identified by experimental studies are reported.

DISCUSSION

In this study, we combined three different kinds of well-developed algorithms, BLAST, HMM, and SVM, to increase sensitivity, and we used background elimination to increase specificity. We then used two data sets to evaluate the performance of BtToxin_scanner, a computational pipeline, and we proved that it is able to distinguish true positives from negatives efficiently. A small, manually curated data set was used in our pipeline method, which reduced the error level that exists with public databases, such as GenBank (44), and also reduced the time required for the whole prediction process.

Theoretically, next-generation sequencing technology produces large numbers of sequences indiscriminately and quickly, and the computational pipeline could identify all kinds of *B. thuringiensis* toxin proteins, including Cry proteins, Vip proteins, and Cyt proteins. Thus, it is expected that our system is more efficient in identifying all kinds of *B. thuringiensis* toxins than PCR-based systems. In fact, we successfully identified 143 candidate *B. thuringiensis* toxins from the mixed plasmid-enriched genomic sequences. Eight of them were obtained with full-length gene sequences.

A major issue that remains to be resolved is how many strains can simultaneously be analyzed by this method. Previous reports showed that *B. thuringiensis* strain YBT-1520 contains a total plasmid genetic content of 988 kb (51). To acquire high-quality sequencing data, we mixed 21 plasmid-enriched genomic DNA sequences from *B. thuringiensis* strains for Illumina sequencing (>50× coverage), and theoretically the average cost for the identification of a *cry* gene would be about 5% of that of a whole-genome shotgun sequencing strategy. We successfully identified

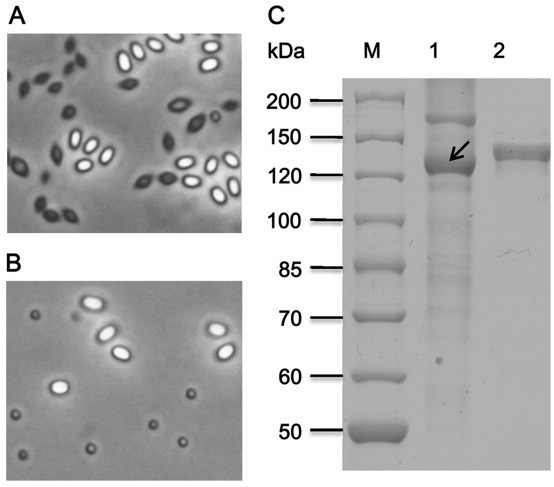

FIG 4 Phase-contrast micrographs (magnification, ×1,000) and SDS-PAGE analysis of crystal proteins from parent and recombinant strains of *B. thuringiensis*. (A) Parent strain Sbt006; (B) strain BMB0659, harboring the 138-kDa protein gene *cry8Ac1* from Sbt006. (C) Lane M, molecular mass standard; lane 1, BMB0659; lane 2, parent strain Sbt006. The arrow points to the position of crystal protein Cry8Ac1.

113 candidate Cry sequences, thus the average cost for the identification of a *cry* gene was reduced significantly.

We noticed that more than 80% of the sequences from the mixed plasmid-enriched sequencing project were shorter than 1 kbp. Previously, we accomplished the genome sequencing of several strains with the same sequencing strategy. We found that the sequence quality is poor (with more contigs and shorter contig length) when the strain harbors more plasmids (unpublished data). We suppose that the poor sequence quality of the mixed plasmid-enriched sequence data is due to both the abundant plasmids and the sequencing technology itself. The complexity of the plasmid sequence, such as repeat elements, and genes with high sequence identity increases the difficulty of the assembly process. We also used an Illumina sequencing platform, which produces short sequence reads (100 bp). In practice, the assembly of shorter sequence reads yields poorer quality assemblies than those with longer capillary reads (38). The poor sequence quality inflates the number of hits being identified and also increases the time and workload needed to clone the complete coding sequence. The only way to solve the poor sequence quality would be the improvement of the sequencing technology itself.

B. thuringiensis is distinguished from *B. cereus* because it can produce parasporal crystals (42). However, we found that *B. cereus* AH603 contained a full-length gene coding for a Cry protein. This strain was originally isolated from a dairy in Norway. The formation of a parasporal crystal phenotype in this strain either has not been established or has not been reported. Thus, whether this strain should be identified as *Bacillus cereus* or *Bacillus thuringiensis* is uncertain based on available information.

In conclusion, we have established a system combining mixed plasmid-enriched genome sequencing and a computational pipeline to mine *cry* genes from *B. thuringiensis*. The system was able to evaluate 21 *B. thuringiensis* strains in a fast and efficient way. A total of 113 candidate Cry sequences were extracted from the 21 strains, and 8 of them were identified. Among them, 3 potentially represent novel *cry* gene types (primary ranks) and 5 of them became *cry* holotypes. These results proved the efficiency of this system to mine *cry* genes. Indeed, the mining of novel sequences must be related to the previous strain selection. Still, it is important to note that the selection of the sequencing strategy affects the final prediction results, thus a choice has to be made between cost and efficiency. Additionally, we have developed a computational pipeline, BtToxin_scanner, which is publicly available at http://bcam.hzaubmb.org/BtToxin_scanner.

ACKNOWLEDGMENTS

This work was supported by the National High Technology Research and Development Program (863) of China (2011AA10A203 and 2006AA02Z174), the National Basic Research Program (973) of China (2009CB118902), the National Natural Science Foundation of China (31170047, 30870066, and 31000020), the Genetically Modified Organisms Breeding Major Projects of China (2009ZX08009-032B), and China 948 Program of the Ministry of Agriculture (2011-G25).

We thank the BBSRC for facilitating a closer collaboration with Neil Crickmore through their China Partnering Award scheme. We thank Zheng Jinshui for his help in website construction.

REFERENCES

1. **Akhurst RJ, James W, Bird LJ, Beard C.** 2003. Resistance to the Cry1Ac delta-endotoxin of *Bacillus thuringiensis* in the cotton bollworm, *Helicoverpa armigera* (Lepidoptera: Noctuidae). J. Econ. Entomol. **96**: 1290–1299.
2. **Alstad DN, Andow DA.** 1995. Managing the evolution of insect resistance to transgenic plants. Science **268**:1894–1896.
3. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.
4. **Arantes O, Lereclus D.** 1991. Construction of cloning vectors for *Bacillus thuringiensis*. Gene **108**:115–119.
5. **Balasubramanian P, et al.** 2002. Cloning and characterization of the crystal protein-encoding gene of *Bacillus thuringiensis* subsp. *yunnanensis*. Appl. Environ. Microbiol. **68**:408–411.
6. **Barton KA, Whiteley HR, Yang NS.** 1987. *Bacillus thuringiensis* section sign–endotoxin expressed in transgenic Nicotiana tabacum provides resistance to lepidopteran insects. Plant Physiol. **85**:1103–1109.
7. **Beron CM, Curatti L, Salerno GL.** 2005. New strategy for identification of novel *cry*-type genes from *Bacillus thuringiensis* strains. Appl. Environ. Microbiol. **71**:761–765.
8. **Camacho C, et al.** 2009. BLAST+: architecture and applications. BMC Bioinformatics **10**:421.
9. **Cao J, Zhao JZ, Tang D, Shelton M, Earle D.** 2002. Broccoli plants with pyramided *cry1Ac* and *cry1C Bt* genes control diamondback moths resistant to Cry1A and Cry1C proteins. Theor. Appl. Genet. **105**:258–264.
10. **Carozzi NB, Kramer VC, Warren GW, Evola S, Koziel MG.** 1991. Prediction of insecticidal activity of *Bacillus thuringiensis* strains by polymerase chain reaction product profiles. Appl. Environ. Microbiol. **57**: 3057–3061.
11. **Challacombe JF, et al.** 2007. The complete genome sequence of *Bacillus thuringiensis* Al Hakam. J. Bacteriol. **189**:3680–3681.
12. **Chang C-C, Lin C-J.** 2011. LIBSVM: a library for support vector machines. ACM Transact. Intell. Syst. Technol. **2**:27:21.
13. Reference deleted.
14. **Crickmore N, et al.** 1998. Revision of the nomenclature for the *Bacillus thuringiensis* pesticidal crystal proteins. Microbiol. Mol. Biol. Rev. **62**:807–813.
15. **Cristianini N, Shawe-Taylor J.** 2000. An introduction to support vector machines, p 1–15. Cambridge University Press, New York, NY.
16. **Durbin R, Eddy S, Krogh A, Mitchison G.** 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, United Kingdom.
17. **Eddy SR.** 1998. Profile hidden Markov models. Bioinformatics **14**:755–763.
18. **Estruch JJ, et al.** 1996. Vip3A, a novel *Bacillus thuringiensis* vegetative insecticidal protein with a wide spectrum of activities against lepidopteran insects. Proc. Natl. Acad. Sci. U. S. A. **93**:5389–5394.
19. **Feitelson JS.** 1993. The *Bacillus thuringiensis* family tree, p 63–71. *In* Kim L (ed), Advanced engineered pesticides. Marcel Dekker, Inc., New York, NY.
20. **Feitelson JS, Payne J, Kim L.** 1992. *Bacillus thuringiensis*: insects and beyond. Biotechnology **10**:271–275.
21. **Fujinaga Y, et al.** 1994. Molecular construction of *Clostridium botulinum* type C progenitor toxin and its gene organization. Biochem. Biophys. Res. Commun. **205**:1291–1298.
22. **Goldman IF, Arnold J, Carlton BC.** 1986. Selection for resistance to *Bacillus thuringiensis* subspecies *israelensis* in field and laboratory populations of the mosquito *Aedes aegypti*. J. Invertebr. Pathol. **47**:317–324.
23. **Guo S, et al.** 2008. New strategy for isolating novel nematicidal crystal protein genes from *Bacillus thuringiensis* strain YBT-1518. Appl. Environ. Microbiol. **74**:6997–7001.
24. **Han CS, et al.** 2006. Pathogenomic sequence analysis of *Bacillus cereus* and *Bacillus thuringiensis* isolates closely related to *Bacillus anthracis*. J. Bacteriol. **188**:3382–3390.
25. **He J, et al.** 2010. Complete genome sequence of *Bacillus thuringiensis* mutant strain BMB171. J. Bacteriol. **192**:4074–4075.
26. **Hofte H, Whiteley HR.** 1989. Insecticidal crystal proteins of *Bacillus thuringiensis*. Microbiol. Rev. **53**:242–255.
27. **Huang F, Buschman LL, Higgins RA, McGaughey WH.** 1999. Inheritance of resistance to *Bacillus thuringiensis* toxin (Dipel ES) in the European corn borer. Science **284**:965–967.
28. **Juarez-Perez VM, Ferrandis MD, Frutos R.** 1997. PCR-based approach for detection of novel *Bacillus thuringiensis cry* genes. Appl. Environ. Microbiol. **63**:2997–3002.
29. **Kalman S, Kiehne KL, Libs JL, Yamamoto T.** 1993. Cloning of a novel *cryIC*-type gene from a strain of *Bacillus thuringiensis* subsp. *galleriae*. Appl. Environ. Microbiol. **59**:1131–1137.
30. **Kongsuwan K, Gough J, Kemp D, McDevitt A, Akhurst R.** 2005. Characterization of a new *Bacillus thuringiensis* endotoxin, Cry47Aa, from

strains that are toxic to the Australian sheep blowfly, *Lucilia cuprina*. FEMS Microbiol. Lett. **252**:127–136.
31. **Kronstad JW, Schnepf HE, Whiteley HR.** 1983. Diversity of locations for *Bacillus thuringiensis* crystal protein genes. J. Bacteriol. **154**:419–428.
32. **Kuo WS, Chak KF.** 1996. Identification of novel *cry*-type genes from *Bacillus thuringiensis* strains on the basis of restriction fragment length polymorphism of the PCR-amplified DNA. Appl. Environ. Microbiol. **62**:1369–1377.
33. **Lee HK, Gill SS.** 1997. Molecular cloning and characterization of a novel mosquitocidal protein gene from *Bacillus thuringiensis* subsp. *fukuokaensis*. Appl. Environ. Microbiol. **63**:4664–4670.
34. **Lin Y, Fang G, Peng K.** 2007. Characterization of the highly variable *cry* gene regions of *Bacillus thuringiensis* strain ly4a3 by PCR-SSCP profiling and sequencing. Biotechnol. Lett. **29**:247–251.
35. **Masson L, Moar WJ, van Frankenhuyzen K, Bosse M, Brousseau R.** 1992. Insecticidal properties of a crystal protein gene product isolated from *Bacillus thuringiensis* subsp. *kenyae*. Appl. Environ. Microbiol. **58**: 642–646.
36. **McLinden JH, et al.** 1985. Cloning and expression of an insecticidal k-73 type crystal protein gene from *Bacillus thuringiensis* var. *kurstaki* into *Escherichia coli*. Appl. Environ. Microbiol. **50**:623–628.
37. **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.
38. **Paszkiewicz K, Studholme DJ.** 2010. De novo assembly of short sequence reads. Brief. Bioinformatics **11**:457–472.
39. **Peng D, et al.** 2009. Elaboration of an electroporation protocol for large plasmids and wild-type strains of *Bacillus thuringiensis*. J. Appl. Microbiol. **106**:1849–1858.
40. **Sambrook J, Russell DW.** 2001. Molecular cloning: a laboratory manual, 3rd ed. Cold Spring Harbor Laboratory Press, New York, NY.
41. **Sampson K, Dunn E, Zeigler J, Tomso D.** 2009. Discovery of novel pesticidal protein genes in *Bacillus thuringiensis* using *de novo* sequencing, abstr B-8, p 41. Abstr. 42nd Annu. Meet. Soc. Invertebrate Pathol.
42. **Schnepf E, et al.** 1998. *Bacillus thuringiensis* and its pesticidal crystal proteins. Microbiol. Mol. Biol. Rev. **62**:775–806.
43. **Schnepf HE, Whiteley HR.** 1981. Cloning and expression of the *Bacillus thuringiensis* crystal protein gene in *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. **78**:2893–2897.
44. **Schnoes AM, Brown SD, Dodevski I, Babbitt PC.** 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput. Biol. **5**:e1000605. doi:10.1371/journal.pcbi.1000605.
45. **Simpson JT, et al.** 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. **19**:1117–1123.
46. **Stajich JE.** 2007. An introduction to BioPerl. Methods Mol. Biol. **406**:535–548.
47. **Tabashnik BE, Van Rensburg JB, Carriere Y.** 2009. Field-evolved insect resistance to *Bt* crops: definition, theory, and data. J. Econ. Entomol. **102**: 2011–2025.
48. **Tan F, et al.** 2009. Cloning and characterization of two novel crystal protein genes, cry54Aa1 and *cry30Fa1*, from *Bacillus thuringiensis* strain BtMC28. Curr. Microbiol. **58**:654–659.
49. **Wright DJ, Iqbal M, Granero F, Ferre J.** 1997. A change in a single midgut receptor in the diamondback moth (*Plutella xylostella*) is only in part responsible for field resistance to *Bacillus thuringiensis* subsp. *kurstaki* and *B. thuringiensis* subsp. *aizawai*. Appl. Environ. Microbiol. **63**:1814–1819.
50. **Yang Z, Chen H, Tang W, Hua H, Lin Y.** 2011. Development and characterisation of transgenic rice expressing two *Bacillus thuringiensis* genes. Pest Manag. Sci. **67**:414–422.
51. **Zhong C, et al.** 2011. Determination of plasmid copy number reveals the total plasmid DNA amount is greater than the chromosomal DNA amount in *Bacillus thuringiensis* YBT-1520. PLoS One **6**:e16025. doi: 10.1371/journal.pone.0016025.