Generation of Cry11 Variants of *Bacillus thuringiensis* by Heuristic Computational Modeling

Evolutionary Bioinformatics
Volume 16: 1–14
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176934320924681
SAGE

Efraín Hernando Pinzón-Reyes[1,2], Daniel Alfonso Sierra-Bueno[3], Miguel Orlando Suarez-Barrera[1], Nohora Juliana Rueda-Forero[1], Sebastián Abaunza-Villamizar[1] and Paola Rondón-Villareal[1]

[1]Universidad de Santander, Faculty of Health Sciences, Laboratory of Molecular Biology and Biotechnology, Bucaramanga, Colombia. [2]Centro de Bioinformática Simulación y Modelado (CBSM), School of Bioinformatic, Universidad de Talca, Talca, Chile. [3]E3T, School of Engineering, Universidad Industrial de Santander (UIS), Bucaramanga, Colombia.

ABSTRACT: Directed evolution methods mimic in vitro Darwinian evolution, inducing random mutations and selective pressure in genes to obtain proteins with enhanced characteristics. These techniques are developed using trial-and-error testing at an experimental level with a high degree of uncertainty. Therefore, in silico modeling of directed evolution is required to support experimental assays. Several in silico approaches have reproduced directed evolution, using statistical, thermodynamic, and kinetic models in an attempt to recreate experimental conditions. Likewise, optimization techniques using heuristic models have been used to understand and find the best scenarios of directed evolution. Our study uses an in silico model named HeurIstics DirecteD EvolutioN, which is based on a genetic algorithm designed to generate chimeric libraries from 2 parental genes, *cry11Aa* and *cry11Ba*, of *Bacillus thuringiensis*. These genes encode crystal-shaped δ-endotoxins with 3 conserved domains. *Cry11* toxins are of biotechnological interest because they have shown to be effective as biopesticides for disease-spreading vectors. With our heuristic model, we considered experimental parameters such as DNA fragmentation length, number of generations or simulation cycles, and mutation rate, to get characteristics of *Cry11* chimeric libraries such as percentage of population identity, truncation of variants obtained from the presence of internal stop codons, percentage of thermodynamic diversity, and stability of variants. Our study allowed us to focus on experimental conditions that may be useful for the design of in vitro and in silico experiments of directed evolution with *Cry* toxins of 3 conserved domains. Furthermore, we obtained in silico libraries of *Cry11* variants, in which structural characteristics of wild *Cry* families were observed in a review of a sample of in silico sequences. We consider that future studies could use our in silico libraries and heuristic computational models, as the one suggested here, to support in vitro experiments of directed evolution.

KEYWORDS: Heuristics, directed molecular evolution, protein engineering, *Bacillus thuringiensis*

RECEIVED: September 9, 2019. **ACCEPTED:** April 7, 2020.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by Conicyt, Fondecyt Postdoctoral Program, No. 3190852, Universidad de Talca, Chile.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Efraín Hernando Pinzón-Reyes, Universidad de Santander (UDES), Faculty of Health Sciences, Laboratory of Molecular Biology and Biotechnology, Street 70 N° 55-210, Building Arhuaco, Office 201, Bucaramanga, Colombia. Email: ehpinzon@udes.edu.co

Introduction

Directed evolution methods mimic evolutionary principles at the laboratory level. For this, strategies for the generation of genic diversity are implemented, either by inducing mutations or by recombining DNA and, after this, selective pressure is carried out.[1] Directed evolution methods, through modified polymerase chain reaction (PCR) cycles, allow obtaining chimeric libraries of recombined genes. In this regard, 2 highly homogeneous and fragmented parental genes are subjected to PCR cycles without primers, until obtaining recombined genes with lengths close to the parental genes.[2] From these libraries, enhanced genes are selected, which are used to obtain a new chimeric library (see Figure 1).

First, in silico approaches of directed evolution were present from the construction of statistical models[3-5]; then models were enriched, including intrinsic thermodynamic information of parental genes[6,7]; and later, kinetic information of reactions was included.[8] These initial models laid the basis for understanding optimal experimental conditions that favored the efficiency and diversity of chimeric protein libraries such as triazine hydrolase, dioxygenases, green fluorescent protein, and beta-lactamases.[4,8,9]

Subsequent studies explored the potential of heuristic techniques to model directed evolution experiments. They assessed the incidence of experimental parameters in the generation of chimeric libraries, recreating the epistasis given in genic sequences through NK landscapes and providing suggestions about favorable experimental conditions in experiments of directed evolution. The experimental parameters assessed were the number of cycles, selective pressure, and mutation rate under high- and low-stringency conditions.[10]

This study presents a heuristic model based on a genetic algorithm, designed to obtain chimeric libraries of *cry11* genes. We have selected this group of genes as our biological model, given their high biotechnological potential.[11,12] These genes are known to be present in a sporulated Gram-positive bacteria named *Bacillus thuringiensis*[13] and encode toxic proteins (δ-endotoxins), which are useful for the biological control of disease-spreading Diptera.[14]

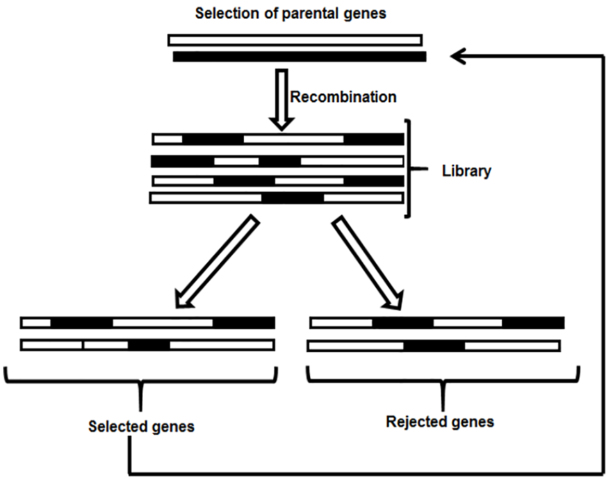

Figure 1. Recombinant DNA techniques of directed evolution.

The biological model of Cry toxins presents at least 2 relevant characteristics for our study. First, Cry toxins have a structure of 3 conserved domains for the 74 or so groups and more than 290 holotypes described so far, and despite their structural conservation, each reported group has a high specificity in its target organism (http://www.lifesci.sussex.ac.uk/home/Neil_Crickmore/Bt/).

Second, experimental models of directed evolution have been reported, where at least 2 *Cry* holotypes have been used, *cry1Ca* and *cry11A12* genes. In the study of Lassner and Bedbrook,[15] an increase in toxicity of the *Cry1Ca* protein has been reported against green doughnut (*Spodoptera exigua*) and fruit worm (*Helicoverpa zea*), while the study developed by Craveiro et al[16] was able to extend the action spectrum of the *Cry11A12* toxin to the giant sugarcane borer species (*Telchin licus licus*), for which the toxin produced by the parental gene was not lethal.[15,16] These studies are an alternative to increase the biopesticide action of native toxins and react to resistant insects.[17-19]

Our study uses a heuristic model, which considers the intrinsic information of *cry11Aa* and *cry11Ba* genes, to generate chimeric libraries and explore the incidence of experimental parameters of directed evolution on the characteristics of chimeric libraries generated in silico, in terms of Diversity, Identity, Delta Energy, and Sequence Truncation.

Materials and Methods

We have implemented a software named HeurIstics DirecteD EvolutioN (HIDDEN), which was written in Python 3 language and simulates a recombining technique of directed evolution by using a genetic algorithm, predicting chimeric libraries from 2 parental genes (http://soft-hidden.com)

HeurIstics DirecteD EvolutioN takes advantage of the common basis of Darwinian evolution used by the evolutionary techniques of artificial intelligence and the recombinant DNA techniques, achieving to reproduce the processes of diversity and selective pressure generation through a genetic algorithm, by which potentially improved genetic variants are obtained. This software is designed to generate libraries from genes with high homology, with preference to genes encoding proteins with 3 conserved domains; it can be used for other gene sequences other than *Cry*, as an example, *Botulinum* parental gene cross-linking is presented (see http://soft-hidden.com/help).

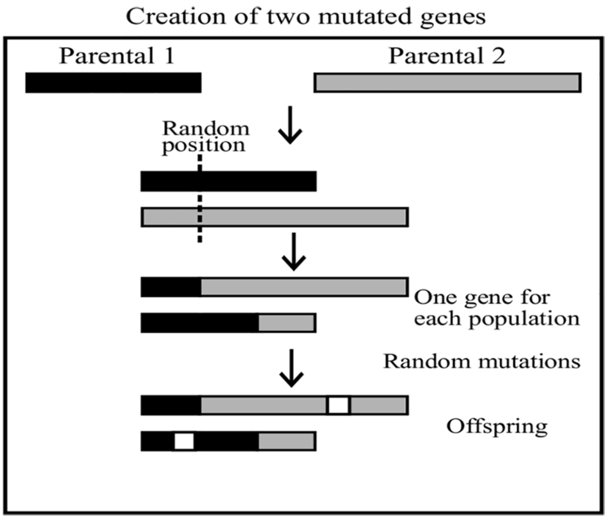

Figure 2. Creation of mutated genes.

Creation of initial populations

The genetic algorithm generates 2 initial populations, 1 for each parental gene, so that mutated genes are created from the *cry11Aa* and *cry11Ba* parental genes until completing the desired number of individuals in the initial populations. The new mutated genes correspond to the genes obtained by crossing the parental genes and performing random mutations. For each cross, 2 parental genes are used, and as a result, a mutated gene is obtained for each of the 2 populations (see Figure 2).

Iterative cycles given by the number of generations

Once the initial populations have been created, the genetic algorithm starts its iterative process given by the desired number of generations. This iterative cycle includes the following actions: assessment of individual fitness, creation of the offspring, and replacement of a percentage of the population.

Assessment of individual fitness. The process of fitness assessment is carried out by evaluating the energy delta for every gene in each population. The energy delta is calculated by dividing the sequence of a gene in its possible 2-mers. Later, the energy contribution of each 2-mer is added, and the final energy delta of the gene is obtained without exclusion of nucleotides in the generated sequence, because the DeltaG allows all the combinations in the genetic code.

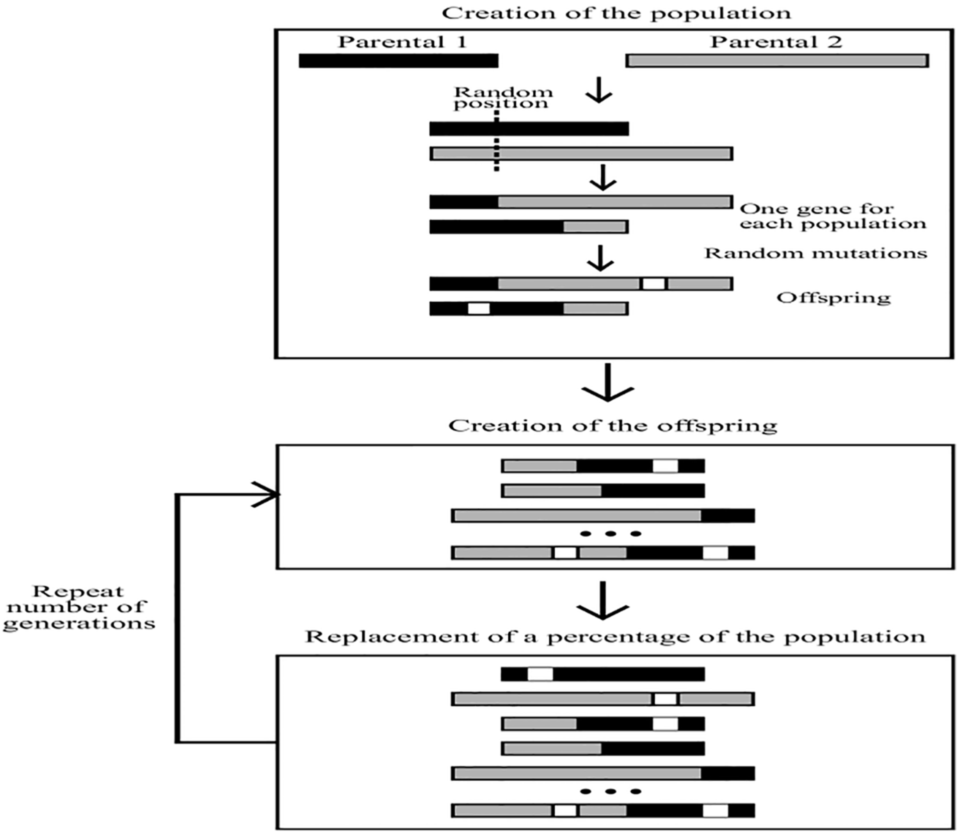
Creation of the population
Parental 1
Parental 2
Random position
One gene for each population
Random mutations
Offspring
Creation of the offspring
Repeat number of generations
Replacement of a percentage of the population

Figure 3. HIDDEN genetic algorithm. HIDDEN indicates HeurIstics DirecteD EvolutioN.

Creation of the offspring. For each individual in the offspring, a parent from each of the populations must be selected. The process of choosing a parent of the population includes selecting 2 candidates by using the roulette method, considering their fitness value. Then both candidates must compete, but only the best one in terms of penalty gets selected as the parent of such population to be used in a cross. The penalization value is calculated as the sum of 3 terms: penalization for mutations (pm), penalization for the size of the open reading frame (ORF) (ps), and penalization for delta value (pd)

The penalization for mutations is equal to 0 if the mutations in domains 1, 2, and 3 of the new gene are not higher than the desired number of mutations for each domain. If the number of mutations is higher than the desired number, then the penalization value is the sum of the additional mutations in each of the 3 domains.

The penalization for the size of the ORF is equal to 10 if its size is smaller than the ORF of the parental gene. Otherwise, the value is equal to 0.

The penalization for delta value is the absolute value of the difference between the delta value of the new gene and the limit of the desired interval. If the value belongs to this interval, then the penalization value is equal to 0.

In this regard, the contest for choosing a parent from each population must be carried out as many times as new individuals are to be created in each population. Once the 2 parents have been selected, the new children will be created by following the methodology explained in Figure 2.

Replacement of a percentage of the population. Then, selective pressure is applied, rejecting a part of the population with deficient fitness values to be replaced by the individuals of the new population. The number of individuals to be replaced is ruled by the parameter of population replacement, which aims to ensure that the most recent generation or population contains the most suitable individuals from the previous generation.

These actions are repeated cyclically over a determined number of generations (see Figure 3).

Diversity generation

HeurIstics DirecteD EvolutioN uses 2 parameters to generate diversity: mutation rate and fragmentation length. The mutation rate is a probability value that rules the allowed rate of DNA changes in the parental genes and can be considered as homogeneous or non-homogeneous throughout the parental gene. In the latter case, HIDDEN takes advantage of intrinsic thermodynamic markers of parental genes, calculated with the SANAFold software.[20] SANAFold allows the characterization of genes from the thermodynamic behavior of their genic regions to favor the formation of secondary DNA structures, under conditions of directed evolution experiments.[21,22] The

Table 1. Information on *cry11* genes used in in silico simulations.

TOXIN	SOURCE OF EXTRACTION	OPEN READING FRAME		
		AA	ST-SP (BP)	GENBANK ACCESS ID
Cry11Aa1	*Bt israelensis*	646	32-1972	M31737-J03510
Cry11Ba1	*Bt jegathesan*	724	64-2238	X86902

Abbreviation: AA, amino acids; BP, base pairs.

mutation rate with non-homogeneous distribution is based on the assumption that the formation of secondary DNA structures does not favor the gene recombination or the mutation appearance.[20]

On the contrary, the fragmentation length is considered as the second parameter that generates diversity, because it regulates the crossing operation of individuals in the genetic algorithm. This parameter corresponds to the number of base pairs (bp) expected in the small pieces of DNA of parental genes, which act as substitutes for primers in a DNA shuffling experiment.[2] This parameter is incorporated into HIDDEN through a Poisson probability model, indicating to the genetic algorithm the location of the crossing point where any 2 genes of the population are recombined.[5]

Simulation scenarios

Simulation scenarios were designed for conditions of low computational performance,[10] in which each simulation produces a library of up to 100 individuals or chimeric sequences. The DNA sequences used as parental genes belong to the group of Cry11 toxins of *Bacillus thuringiensis*, and *cry11Aa* and *cry11Ba* genes with high homology were selected from this group (see Table 1).

Each simulation scenario includes the parameters of diversity (Mutation Rate and Fragmentation Length) and selection (Number of Generations). The ranges established for the variation of these parameters were taken from in vitro DNA shuffling conditions.

Values established for the Fragmentation Length were 75bp and 150bp, which correspond to experimental length values of fragments obtained through the action of DNAsaI in in vitro DNA shuffling experiments with *cry11* genes.[23]

On the contrary, the values of 15 generations and 100 generations established for the parameter of number of generations correspond to values close to the range reported as favorable for the obtention of chimeric libraries, in in vitro[10] and in silico directed evolution studies.[23] This parameter allows replicating the number of cycles of directed evolution that is used as stop criteria for the execution of the HIDDEN genetic algorithm.

As for the third parameter (Mutation Rate) that has an impact on the simulation scenarios established, 6 values (0.001, 0.003, 0.005, 0.01, 0.02, 0.05) were selected, from which 0.02 and 0.05 have been reported in some in silico studies as optimal simulation values. However, to explore the incidence of mutation rate, lower values reported in other experimental scenarios of directed evolution were introduced.[10] This third parameter is incorporated into the simulation scenarios by using 2 strategies: the first one assumes that the Mutation Rate is homogeneously distributed throughout the gene; the second one assumes that the *cry* gene, made up of 3 genic regions, distributes the mutation rate preferentially in these regions, where the distribution criteria are ruled by the intrinsic thermodynamic information of the region.[20] In this regard, each mutation rate value constitutes itself as a parameter that generates 2 possible scenarios, simulations in mutation conditions with Homogeneous and Non-Homogeneous Distribution. Based on these, 48 simulation scenarios were constructed by combining these parameters: distribution of mutation rate (homogeneous distribution, non-homogeneous distribution), fragmentation length (75bp, 150bp), number of generations (15, 100), and mutation rate (0.05, 0.02, 0.01, 0.005, 0.003, 0.001)

Statistical analysis

Four estimators were established to assess the results of the 48 simulation scenarios: Diversity, Identity, Truncated Proteins, and Fitness (Energy Delta). Each individual obtained through HIDDEN is an amino acid sequence of a *Cry11* variant, and each sequence provides biological information that can be simplified in the constructed estimators. The arithmetic mean of the information provided by a group of 100 individuals that make up a chimeric library generated by HIDDEN constitutes the value of an estimator, which is useful to assess the incidence of the diversity and selection parameters in the different simulation scenarios.

The simulations of the 48 scenarios were performed in triplicate to ensure the statistical variability of the data. In total, 144 simulations were performed and provided that each simulation scenario gives 2 populations, 1 population for each parental gene, 288 values were obtained for each of the 4 population estimators, ie, a total of 1152 values were calculated.

For the statistical analysis, tests were performed to analyze the variance of a single factor and *n* variables, depending on the configuration of the data required. These tests were useful in

Table 2. Significant statistical difference (ANOVA), based on the number of generations parameter.

CRY11 15 GENERATIONS VS 100 GENERATIONS					
15 GENERATIONS	DELTA ENERGY	TRUNCATED ENERGY	IDENTITY	DIVERSITY	100 GENERATIONS
Cry11Aa 15 Generations	22/24	1/24	1/24	8/24	*Cry11Aa* 100 Generations
Cry11Aa 15 Generations	20/24	2/24	2/24	6/24	*Cry11Ba* 100 Generations

Abbreviation: ANOVA, analysis of variance.
Number of ANOVAs of 24 that has a value of $P < .05$ for each estimator.

establishing whether the differences in estimators among simulation scenarios were statistically significant or not.

Results and Discussion

As a result of the 144 simulations, a total of 288 chimeric libraries were obtained, from the 48 simulation scenarios run in triplicate, because each simulation produces 2 chimeric libraries: a chimeric library with 100 sequences of *Cry11Aa* variants and another chimeric library with 100 sequences of *Cry11Ba* variants. A total of 28 800 sequences were obtained, corresponding to 14 400 sequences of *Cry11Aa* variants and 14 400 sequences of *Cry11Ba* variants.

Four population estimators were calculated from these sequences: (a) Library Diversity, ie, the percentage of the obtained population that was not repeated; (b) Population Identity, which corresponds to the average of population identity with the parental gene, obtained from the calculation of the identity matrix using the ClustalW algorithm (https://www.ebi.ac.uk/Tools/msa/clustalo/); (c) Library Functionality, consisting of the arithmetic mean of the penalty factor of individuals concerning the presence of coding fragments for internal stop codons; and (d) Library Fitness, calculated from the arithmetic mean of the Gibbs Energy, constituting the thermodynamic stability of each individual. The 48 scenarios were organized in pairs of comparable scenarios for the first 3 parameters: 24 pairs of comparable scenarios based on the distribution of mutation rate (homogeneous distribution vs non-homogeneous distribution), 24 pairs of comparable scenarios depending on the fragmentation length (75bp vs 150bp), and 24 pairs of comparable scenarios based on the number of generations (15 generations vs 100 generations).

Finally, the 48 simulation scenarios were organized in 8 groups to assess the mutation rate parameter, making 1 group for every 6 simulation scenarios, where only the mutation rates change (0.05, 0.02, 0.01, 0.005, 0.003, 0.001). For these 8 groups, the Tukey test was performed when there was a statistically significant difference among them.

Thus, a total of 72 pairs of comparable scenarios and 8 groups were assessed for a review of 80 evaluation environments. For each one of them, 4 ANOVAs (analyses of variance) were estimated, 1 ANOVA for each population estimator, for a total of 240 ANOVAs evaluated in *Cry11Aa* chimeric libraries and 240 ANOVAs in *Cry11Ba* chimeric libraries. A statistical review was performed based on the 480 ANOVAs calculated.

Number of generations

The Number of Generations parameter showed significant differences in the 4 estimators for the chimeric libraries with *Cry11Aa* and *Cry11Ba* variants. However, the Energy Delta and Diversity estimators showed a greater number of significant differences when varying the number of generations parameter. For *Cry11Aa* libraries, the number of differences was 22 and 8, and for *Cry11Ba* libraries, it was of 20 and 6, corresponding to Delta Energy and Diversity, respectively. On the contrary, the Truncated Proteins and Identity estimators of the variants concerning the parental gene showed differentiated behaviors that were not very representative (see Table 2).

The Number of Generations, as a parameter of the genetic algorithm, showed effects on the Diversity estimator of the chimeric libraries obtained from *cry11Aa* and *cry11Ba* genes. A significant decrease in Diversity was observed in simulation scenarios where the number of generations was equal to 100, compared with the simulation scenarios using a number of generations equal to 15. Diversity values near 0.95 and 0.96 that were obtained with 15 generations decreased to values near 0.84 and 0.86 with 100 generations for *cry11Aa* and *cry11Ba* gene libraries, respectively (see Table 3). This can be attributed to the characteristics of the genetic algorithm developed: by increasing the number of generations, the best individuals are conserved for the next generations. Thus, a selected group tends to remain and duplicate individuals. Therefore, if a greater Diversity is desired, few generations of directed evolution are preferred.[10]

On the contrary, though a high number of generations has an undesirable effect on the Diversity estimator, it favors the Delta Energy estimator (see Table 3). The Delta Energy estimator represents the thermodynamic stability of the variants obtained, which showed an increase in stability in simulation scenarios of 100 generations, demonstrating an average energy increase of 46.53 Kcal/mol for *Cry11Aa* variants and 45.23 Kcal/mol for *Cry11Ba* variants. However, this can be also attributed to the characteristics of the genetic algorithm, given that the

Table 3. Population averages of the estimators in simulation scenarios with variation in the number of generations.

ESTIMATORS	*CRY11AA* VARIANTS		*CRY11BA* VARIANTS	
	15 GENERATIONS	100 GENERATIONS	15 GENERATIONS	100 GENERATIONS
Diversity	0.95 ± 0.03	0.84 ± 0.1	0.96 ± 0.03	0.86 ± 0.09
Identity	0.67 ± 0.12	0.63 ± 0.14	0.62 ± 0.13	0.59 ± 0.12
Truncated proteins	0.92 ± 0.17	0.81 ± 0.34	0.92 ± 0.16	0.80 ± 0.35
Delta energy[a]	–2391.33 ± 31.17	–2437.86 ± 30.02	–2675.83 ± 33.48	–2721.06 ± 32.71

[a]Units of energy Delta are Kcal/mol; other estimators are in proportions.

Table 4. Significant statistical difference (ANOVA), based on the fragmentation length parameter.

FRAGMENTATION LENGTH: *CRY11* 75BP VS 150BP					
75BP	DELTA ENERGY	TRUNCATED PROTEINS	IDENTITY	DIVERSITY	150BP
75bp fragmentation length for *Cry11Aa*	4/24	1/24	3/24	0/24	150bp fragmentation length for *Cry11Aa*
75bp fragmentation length for *Cry11Ba*	2/24	2/24	3/24	2/24	150bp fragmentation length for *Cry11Ba*

Abbreviation: ANOVA, analysis of variance; BP, base pairs.
Number of ANOVAs of 24 that has a $P < .05$ value for each estimator.

greater the number of generations, the better the values of the fitness function are likely to be obtained.

In addition, a high number of generations is supported by a good average in the truncated protein estimator (see Table 3), where there is a slight improvement of the indicator, to interpret this estimator, values that are very close to 1 imply a high penalty of population individuals due to the presence of internal stop codons. Thus, in comparison with arithmetic means, an improvement near 0.11 is observed for *Cry11Aa* and 0.12 for *Cry11Ba*. Then, for the benefit of the thermodynamic stability and with average favorable truncations, scenarios with a high number of generations are preferred.[10] This is because the genetic algorithm (HIDDEN) simulates a high selective pressure that includes energy delta penalties and internal stop codons.[10]

Furthermore, average population values of identity around 0.6 were observed. Although it is not a high population identity value, it is maintained despite the number of generations implemented. The above seems to indicate that without another modified experimental condition, 3-domain *Cry11* variants are highly conserved.[24]

To conclude, a high number of generations compared with a low number of generations favor the efficiency (Delta Energy, Truncated Proteins) of the *Cry11* chimeric libraries and do not favor the generation of diversity (Diversity). These data coincide with the reported one in previous studies where a strong selective pressure has been proved to be beneficial in directed evolution experiments.[10,25]

Fragmentation length

The results of the ANOVA to evaluate the behavior of the population estimators from a 75bp-to-150bp fragmentation length showed some significant differences in 3 out of the 4 estimators: Delta Energy, Truncated Proteins, and Identity for the *Cry11Aa* variant populations as well as in all 4 estimators for the *Cry11Ba* variant populations (see Table 4). In both cases, the number of scenarios with significant differences was very low [0%-16%] (see Table 4).

When checking the incidence of fragmentation length on the population arithmetic mean of the estimators, an improvement of about 0.09 was observed for *Cry11Aa* and 0.11 for *Cry11Ba* on the truncated protein estimator, when using 75bp fragmentation lengths (see Table 5). This situation may be because 75bp fragmentation lengths do not favor the formation of heteroduplex in the process of genetic algorithm recombination, reducing the possibility of including restriction codons within the sequences,[6] then favoring the efficiency of the library.

Distribution of mutation rates

The Mutation Rate parameter was applied using a Homogeneous and a Non-Homogeneous Distribution,[20] presenting significant differences in 3 out of the 4 estimators: Delta Energy, Truncated Proteins, and Identity for the variant populations of *Cry11Aa* as well as in all 4 estimators for the variant populations of *Cry11Ba* (see Table 6). In both cases, the

Table 5. Population averages of the estimators in simulation scenarios with variation in FL.

ESTIMATORS	*CRY11AA* VARIANTS		*CRY11BA* VARIANTS	
	FL 75BP	FL 150BP	FL 75BP	FL 150BP
Diversity	0.88 ± 0.09	0.88 ± 0.07	0.92 ± 0.07	0.91 ± 0.06
Identity	0.67 ± 0.97	0.63 ± 0.13	0.62 ± 0.10	0.59 ± 0.12
Truncated proteins	0.82 ± 0.23	0.91 ± 0.10	0.80 ± 0.20	0.91 ± 0.10
Delta energy[a]	−2411.33 ± 37.28	−2417.80 ± 36.65	−2694.32 ± 38.37	−2702.48 ± 38.15

Abbreviation: FL, fragmentation length; BP, base pairs.
[a]Delta energy units are in Kcal/mol; other estimators are in proportions.

Table 6. Significant statistical difference (ANOVA), based on the parameter of mutation rate distribution.

MUTATION RATE DISTRIBUTION: *CRY11* HOMOGENEOUS VS NON-HOMOGENEOUS					
HOMOGENEOUS	DELTA ENERGY	TRUNCATED PROTEINS	IDENTITY	DIVERSITY	NON-HOMOGENEOUS
Homogeneous distribution for *Cry11Aa*	3/24	2/24	1/24	0/24	Non-homogeneous distribution for *Cry11Aa*
Homogeneous distribution for *Cry11Ba*	3/24	2/24	2/24	1/24	Non-homogeneous distribution for *Cry11Ba*

Abbreviation: ANOVA, analysis of variance.
Number of ANOVAs of 24 that has a value of $P < .05$ for each estimator.

Table 7. Population averages of estimators in simulation scenarios with variation in the number of generations.

ESTIMATORS	*CRY11AA* VARIANTS		*CRY11BA* VARIANTS	
	H-MR	NH-MR	H-MR	NH-MR
Diversity	0.89 ± 0.08	0.88 ± 0.11	0.92 ± 0.07	0.90 ± 0.10
Identity	0.65 ± 0.13	0.65 ± 0.13	0.62 ± 0.13	0.59 ± 0.12
Truncated proteins	0.92 ± 0.17	0.81 ± 0.34	0.92 ± 0.16	0.80 ± 0.35
Delta energy[a]	−2415.35 ± 40.25	−2413.85 ± 36.52	−2698.99 ± 42.29	−2697.89 ± 37.74

Abbreviations: H-MR, homogeneous mutation rate; NH-MR, non-homogeneous mutation rate.
[a]Delta energy units are expressed in Kcal/mol; other estimators are in proportions.

number of scenarios with significant differences was very low [0%-12%].

When reviewing the incidence of mutation rate distribution on the population arithmetic mean of the estimators, an improvement of about 0.11 for *Cry11Aa* and 0.12 for *Cry11Ba* was observed on the truncated protein estimator, when using non-homogeneous mutation rates, ie, distributed according to thermodynamic criteria differentiated in the 3 conserved *Cry* domains (see Table 7).

The above suggests that the intrinsic thermodynamic information in the *cry11Aa* and *cry11Ba* gene sequences affects the recombination process, favoring the efficiency of the recombined sequences. These differences may be associated with the evolutive characteristics of the parental genes.[20,24]

Mutation rates

Delta energy. Significant differences are observed in the "Delta Energy" values, among simulation scenarios with medium-low mutation rates [0.001-0.001] and high mutation rates [0.02, 0.05]. This occurs for both *Cry11Aa* (see Table 8) and *Cry11Ba* (see Table 9).

The percentage of scenarios with high mutation rates and significant differences when comparing the 8 groups was in the range of [12%-100%] for *Cry11Aa* (see Table 8) and in the range of [0%-75%] for *Cry11Ba* (see Table 9).

These significant differences in the Delta Energy value lead to the conclusion that high mutation rates [0.02, 0.05] favor the thermodynamic stability of the obtained libraries (see

Table 8. Significant statistical difference of the energy Delta estimator for *Cry11Aa* based on the variation in mutation rates.

DELTA ENERGY FOR CRY11AA						
	0.001	0.003	0.005	0.01	0.02	0.05
0.001		0	0	1/8	8/8	8/8
0.003	0		0	0	8/8	8/8
0.005	0	0		0	5/8	7/8
0.01	1/8	0	0		1/8	6/8
0.02	8/8	8/8	5/8	1/8		6/8
0.05	8/8	8/8	7/8	6/8	6/8	

Number of $P < .05$ values obtained with the Tukey test among simulation scenarios for 8 compared groups.

Table 9. Significant statistical difference of the Delta energy estimator for *Cry11Ba* based on the variation in mutation rates.

DELTA ENERGY FOR CRY11BA						
	0.001	0.003	0.005	0.01	0.02	0.05
0.001		0	0	0	5/8	6/8
0.003	0		0	0	4/8	6/8
0.005	0	0		0	2/8	6/8
0.01	0	0	0		0	4/8
0.02	5/8	4/8	2/8	0		4/8
0.05	6/8	6/8	6/8	4/8	4/8	

Number of $P < .05$ values obtained with the Tukey test among simulation scenarios for 8 compared groups.

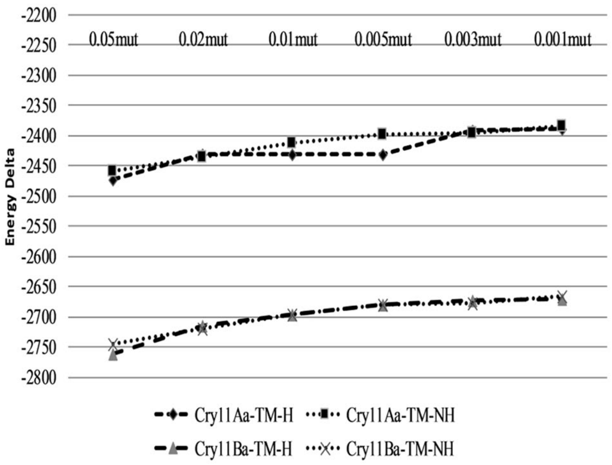

Figure 4. Behavior of the population "Delta energy" estimator, when varying the mutation rate and the H-MR (homogeneous mutation rate) and NH-MR (non-homogeneous mutation rate) distribution.

Table 10. Significant statistical difference of the identity estimator for *Cry11Aa* based on the variation in mutation rates.

CRY11AA IDENTITY						
	0.001	0.003	0.005	0.01	0.02	0.05
0.001		0	0	0	0	7/8
0.003	0		0	0	0	6/8
0.005	0	0		0	1/8	2/8
0.01	0	0	0		0	4/8
0.02	0	0	1/8	0		2/8
0.05	7/8	6/8	2/8	4/8	2/8	

Number of $P < .05$ values obtained with the Tukey test among simulation scenarios for 8 compared groups.

Figure 4). From the algorithmic point of view, we may consider that higher values in the mutation rates allow greater exploration of the spectrum and favor the optimization of the fitness value of the obtained sequences.

Identity. Significant differences are observed in the "Identity" values among simulation scenarios with medium-low mutation rates [0.001-0.002] and the highest mutation rate [0.05], with respect to *Cry11Aa* (see Table 10) and between medium-low mutation rates [0.001-0.001] and high mutation rates [0.02, 0.05] for *Cry11Ba* (see Table 11).

The percentage of scenarios with high mutation rates and significant differences when comparing the 8 groups was in the range of [25%-87%] for *Cry11Aa* (see Table 10) and in the range of [0%-25%] for *Cry11Ba* with respect to a mutation rate of 0.02 and [0%-62%] for *Cry11Ba* with respect to a mutation rate of 0.05 (see Table 11).

These significant differences in the Identity value lead to the conclusion that high mutation rates [0.02, 0.05] do not favor the identity of the obtained libraries (see Figure 5). This conclusion was expected as mutation rates allow the algorithm to explore sequence changes that favor the fitness function (Gibbs Free Energy of the sequence).

Truncated proteins. Significant differences are observed in the "truncated proteins" values among simulation scenarios with medium-low mutation rates [0.005-0.05] and lower mutation rates [0.001, 0.003], for both *Cry11Aa* (see Table 12) and *Cry11Ba* (see Table 13).

The percentage of scenarios with low mutation rates and significant differences when comparing the 8 groups was in the range of [0%-12%] for *Cry11Aa* with a mutation rate of 0.003 and of [12%-25%] for *Cry11Aa* with a mutation rate of 0.001 (see Table 12). Similar behavior was observed for *Cry11Ba* (see Table 13).

These significant differences in the "Truncated Proteins" value lead to the conclusion that low mutation rates [0.001] favor the formation of sequences without internal stop codons (see Figure 6). This conclusion was to be expected as the high mutation rates increase the inclusion of variations in the sequence, increasing the probability of incorporating stop codons.

Table 11. Significant statistical difference of the identity estimator for *Cry11Ba* based on the variation of mutation rates.

CRY11BA IDENTITY						
	0.001	0.003	0.005	0.01	0.02	0.05
0.001		0	0	0	2/8	5/8
0.003	0		0	1/8	2/8	5/8
0.005	0	0		0	1/8	4/8
0.01	0	1/8	0		0	4/8
0.02	2/8	2/8	1/8	0		0
0.05	5/8	5/8	4/8	4/8	4/8	

Number of $P < .05$ values obtained with the Tukey test among simulation scenarios for 8 compared groups.

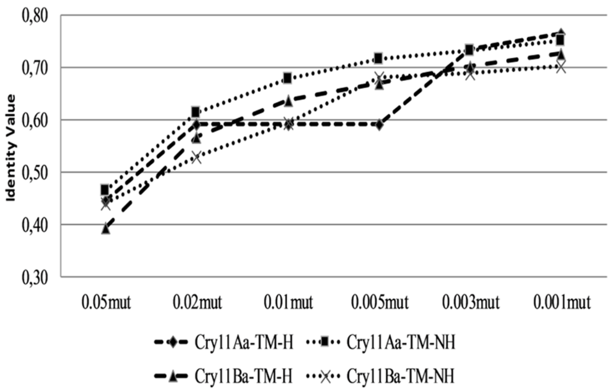

Figure 5. Behavior of the population "identity" estimator when varying mutation rates and H-MR (homogeneous mutation rate) and NH-MR (non-homogeneous mutation rate) distribution.

Table 12. Significant statistical difference of the truncate proteins estimator for *Cry11Aa* based on the variation in mutation rates.

CRY11AA TRUNCATED PROTEINS						
	0.001	0.003	0.005	0.01	0.02	0.05
0.001		1/8	1/8	1/8	2/8	2/8
0.003	1/8		1/8	1/8	1/8	0
0.005	1/8	1/8		0	0	0
0.01	1/8	1/8	0		0	0
0.02	2/8	1/8	0	0		0
0.05	2/8	0	0	0	0	

Number of $P < .05$ values obtained with the Tukey test among simulation scenarios for 8 compared groups.

Diversity. Significant differences were observed in the "Diversity" values among simulation scenarios with low mutation rates [0.001-0.005] and the highest mutation rate [0.05] for *Cry11Aa* (see Table 14), while there were no significant differences for *Cry11Ba*.

Table 13. Significant statistical difference of the truncate proteins estimator for *Cry11Ba* based on the variation in mutation rates.

CRY11BA TRUNCATED PROTEINS						
	0.001	0.003	0.005	0.01	0.02	0.05
0.001		1/8	1/8	2/8	2/8	2/8
0.003	1/8		0	1/8	1/8	0
0.005	1/8	0		0	0	0
0.01	2/8	1/8	0		0	0
0.02	2/8	1/8	0	0		0
0.05	2/8	0	0	0	0	

Number of $P < .05$ values obtained with the Tukey test among simulation scenarios for 8 compared groups.

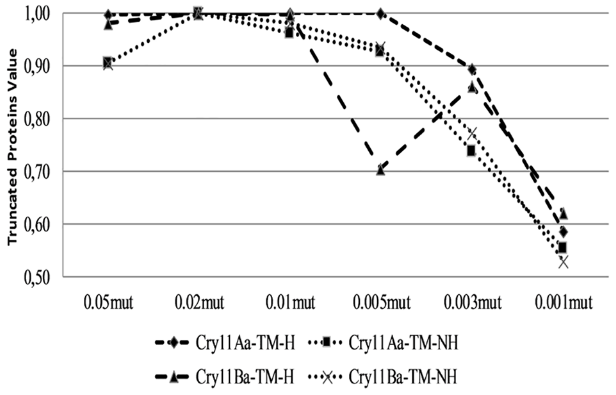

Figure 6. Behavior of population "truncated proteins" estimator when varying the mutation rate and H-MR (homogeneous mutation rate) and NH-MR (non-homogeneous mutation rate) distribution.

These significant differences in the "Diversity" value lead to the conclusion that low mutation rates [0.001-0.005] favor the formation of conserved sequences (see Figure 7). This conclusion was expected as the low mutation rates make the exploration of new global optimizations through the function of the algorithm optimization difficult, so it tends to replicate the best solutions in high numbers.

Then, the analysis of the incidence of mutation rates on the estimators resulted as expected. High mutation rates [0.02-0.05] favor the Energy Delta estimator but do not favor the identity estimator. Meanwhile, low mutation rates [0.001-0.003] favor the Truncated Protein estimator but do not favor the Diversity estimator.

These analyses of population estimators provide some guidance on how directed evolution parameters may affect results in libraries generated from parental genes that code for proteins from 3 conserved domains.

Structural analysis of the best sequences obtained using HIDDEN

The HIDDEN algorithm creates libraries of 3-domain *Cry* variants. The generated sequences have associated scores:

Table 14. Significant statistical difference of the diversity estimator for *Cry11Aa* based on the variation in mutation rates.

CRY11AA DIVERSITY						
	0.001	0.003	0.005	0.01	0.02	0.05
0.001		0	0	0	0	1/8
0.003	0		0	0	0	1/8
0.005	0	0		1/8	1/8	1/8
0.01	0	0	1/8		0	0
0.02	0	0	1/8	0		0
0.05	1/8	1/8	1/8	0	0	

Number of $P < .05$ values obtained with the Tukey test among simulation scenarios for 8 compared groups.

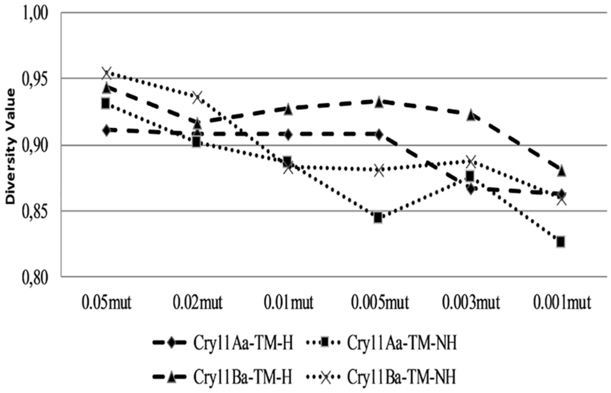

Figure 7. Behavior of the population "diversity" estimator when varying the mutation rate and H-MR (homogeneous mutation rate) and NH-MR (non-homogeneous mutation rate) distribution.

penalties due to energy delta, penalties due to protein truncation, penalties due to the number of mutations in the DNA sequence, and in the amino acid sequence. Six libraries generated by HIDDEN were randomly selected, and from these libraries, the best sequence was selected by filtering the truncation penalty equal to 0. A structural review was performed on these randomly obtained sequences.

As a first approximation, the structural analysis was focused on establishing the percentages of identity and similarity of each of the mutant proteins obtained by using the in silico tools previously described, compared with the *Cry11Aa* and *Cry11Ba* parental ones (see Table 15).

These results show that the variant with the lowest identity and similarity in relation to the *Cry11Aa* parental protein was H15pop1-25 with values of 75.5% and 84.8%, respectively. On the contrary, the variant with the highest identity and similarity values was S15pop1-1, with percentages of 85.1% and 89.9%. Regarding the *Cry11Ba* parental protein, the mutants with the highest and lowest percentages of identity and similarity were H15pop1-25 (67.1%/76.1%) and S100pop1-1b (59.4%/70.4%), respectively. As the highest percentages of the identity of the entire set of mutant proteins were given with *Cry11Aa*, the subsequent analyses were performed considering its amino acid sequence.

Mutation rates of the mutant protein group ranged between 14.9% and 24.5%, showing that the lowest mutation rate protein was S15pop1-1 and the highest one was H15pop1-25. On the contrary, the H15pop1-6, H15pop1-15, S100pop1-1a, and S100pop1-1b mutants showed values of 19.1%, 17.7%, 16.9%, and 19.9%, respectively.

Later, an analysis of substitutions, deletions, and insertions domain-wise was performed. This information is vital to determine how thermodynamic parameters are involved in the mutant recombination. It is possible to observe that domain III is the one that presented the least number of variations, both in variants of homogeneous mutation rate and in mutants of restricted variation by domains, according to the parameters of DNA thermodynamic spontaneity with SANAFold.[20] The mutants with the most variations in the domain I were S100pop1-1b, H15pop1-6, and S100pop1-1a, with values of 10.73, 8.53, and 7.10, respectively. As additional information, it was observed that the S15pop1-1 mutant did not present any mutation in domain III, confirming at the same time that it is the least changing with respect to *Cry11Aa* (see Table 16).

There are positions in domain I that are described as important factors in the toxicity, all this in relation to the formation of lithic pore, regions such as the α4 helix and α5 helix, to which the formation of the oligomer is attributed.[26] On the contrary, regions located in domain II have been identified, specifically, the positions corresponding to the loops 3, α8, β4, and these regions are reported as residues involved in receptor bindings.[27] Finally, domain III, which has been reported to play an important role in the stability of receptor bindings, mainly with aminopeptidases (aminopeptidase N (APN)) and alkaline phosphatase (ALP), also plays an important role in the conservation of protein integrity with the solvent. It also presents important regions such as the coding zone to the β16 leaf, which is mainly involved in the stability of the toxin-receptor bindings.[28]

In this regard, it is important to verify the changes that the mutants maintain concerning these zones. As for the coding region of the α4 helix, the variants H15pop1-6, H15pop1-25, H15pop1-15, S15pop1-1 show changes, being H15pop1-6 the one with the highest number of changes with 6 substitutions (LF-TS) in positions 138-139 and (LSGA-QLIS) in positions 141-144. Studies conducted by Girard et al[29] show the importance of this region with specific substitutions. This study reports a reduction in the toxic activity of *Cry11Aa* compared with *Manduca sexta* due to changes in the positions I132C, S139C, and V150C.

With respect to the α5 helix, it was one of the important regions with a great variation by the mutants, only the mutant S100pop1-1 maintained this region with respect to the

Table 15. Table of percentage of identity (upper triangular matrix) and similarity (bottom singular matrix—region in gray) of the variants for the parental proteins.

	CRY11AA	CRY11BA	H15POP1-6	H15POP1-25	H15POP1-15	S15POP1-1	S100POP1-1A	S100POP1-1B
Cry11Aa		53.7	80.9	75.5	82.3	85.1	83.1	81.1
Cry11Ba	67.7		61.0	67.1	65.2	64.1	62.4	59.4
H15pop1-6	87.8	71.4		69.8	79.5	78.9	79.3	73.9
H15pop1-25	84.8	76.1	81.4		81.8	83.1	77.7	67.7
H15pop1-15	89.3	75.0	86.5	89.8		91.4	87.1	75.6
S15pop1-1	89.9	74.7	86.8	89.9	95.0		84.0	76.8
S100pop1-1a	89.8	73.1	86.2	87.0	91.3	89.9		75.2
S100pop1-1b	88.5	70.4	82.6	80.6	86.0	85.7	84.8	

The values show the identity and similarity of the variants for *Cry11Aa* and *Cry11Ba*.

Table 16. Changes in the mutant domains.

VARIANT	DOMAIN I	TOTAL[a]	DOMAIN II	TOTAL[a]	DOMAIN III	TOTAL[a]	TOTAL
H15pop1-6	6.59 SUS	8.53	5.68 SUS	6.73	2.64 SUS	2.64	6.40
	0.78 INS		0.75 INS		0.0 INS		
	1.16 DEL		0.30 DEL		0.0 DEL		
H15pop1-25	4.52 SUS	5.3	11.21 SUS	11.96	8.92 SUS	9.12	8.57
	0.39 INS		0.45 INS		0.0 INS		
	0.39 DEL		0.30 DEL		0.20 DEL		
H15pop1-15	4.01 SUS	5.18	9.87 SUS	10.32	1.01 SUS	1.01	5.89
	0.39 INS		0.45 INS		0.0 INS		
	0.78 DEL		0.0 DEL		0.0 DEL		
S15pop1-1	3.49 SUS	4.27	9.12 SUS	9.72	0.0 SUS	0.0	5.06
	0.39 INS		0.30 INS		0.0 INS		
	0.39 DEL		0.30 DEL		0.0 DEL		
S100pop1-1a	5.04 SUS	7.10	7.32 SUS	7.77	1.42 SUS	1.82	5.99
	0.90 INS		0.45 INS		0.20 INS		
	1.16 DEL		0.0 DEL		0.20 DEL		
S100pop1-1b	7.63 SUS	10.73	4.48 SUS	5.08	2.43 SUS	3.25	6.92
	1.55 INS		0.30 INS		0.41 INS		
	1.55 DEL		0.30 DEL		0.41 DEL		

Abbreviations: DEL, deletion; INS, insertion; SUS, substitution; Total, percentage of changes throughout the sequence.
The analyses were performed comparing the amino acid sequence of the *Cry11Aa* parental protein. The sizes for Domain I, Domain II, and Domain III are 774, 669, and 493, respectively. The size of the parental protein is 1936.
[a]Total: percentage of changes in each domain.

Cry11Aa parental protein. The mutant with the highest number of changes in this region was H15pop1-25 with 3 changes, 2 of them (LH-FN) in positions 174-175 and the last one in position T177G. A study in *M. sexta* conducted by Alzate et al showed that a change in position L157 from the α5 helix region provoked a decrease in the toxic activity of the *Cry1Ab* protein. However, the same modification produced an increase of up to 4 times compared with *Lymantria dispar*.[24]

The α8 loop region showed multiple changes. However, the mutants S15pop1-1 and S100pop1-1b did not show any changes in this region. The most variable mutant was H15pop1-15 with 5 substitutions (PVNY286-289NISP) (E291D), an insertion in position 285N. These changes may suggest a variation in mutant toxicity, as in the study conducted by Fuji et al that highlights the importance of the α8 loop of the *Cry1Ab* toxin in the interactions of the toxin with the cadherin BT-R1 receptor.[30,31] This suggests that mutants may be potentiated concerning the binding to the receptor or, in another context, an improvement in additional interactions with target insects that have not been reported yet.

The mutants H15pop1-25 and S100pop1-1a showed changes in the β4 loop region. The first mutant showed a specific change in position N358T, and the second mutant had an insertion in position 362HN. Studies conducted by Fernández et al[32] highlighted the importance of β4 loop in *Cry11Aa* bindings to the corresponding receptors of *Aedes aegypti*. Similarly studies conducted by Li et al, where mutants with variations were obtained in the regions associated to the β4 loop, and 2 potentiated mutants of *Cry2Ah1* (*Cry2Ah1-vp* and *Cry2Ah1-sp*) were produced. The mutant *Cry2Ah1-vp* had an insertion of a Proline in position 354 (V354VP), and Cry2Ah1-sp showed a change from a Valine to a Serine and Proline (V354SP); these mutants had increased toxicity of 1.5 and 5.3 times, respectively, compared with the parental activity. The attribution of the increase according to the authors of the study may be correlated with the toxin-receptor interaction of *Helicoverpa armígera*.[33]

On the contrary, the mutants H15pop1-15 and S100pop1-1a showed changes in the region corresponding to loop 3, the former containing the highest number of changes (4) in positions RI483-484KL, S488G, Q492E. The changes in this position may represent the importance of the acquisition of increased toxicity. Several studies state that loop 3 region of *Cry* toxins is directly related to the significant binding to cadherin BT-R1 receptors with *Cry1Aa* and *Cry1Ab* binding to the BtR receptors of *Heliothis virescens*.[30] Studies on mutagenesis in coding regions to loop 3 conducted by Pacheco et al[34] demonstrate the importance of this amino acid region, in which variations of toxicity could be correlated to changes made in *Cry1Ab* loop 3, ascribing that the lack of binding of toxins to the BBMV of *M. sexta* result in changes in oligomerization and affect the toxicity of the protein

Almost all mutants show changes in domain III, except for S15pop1-1. Studies conducted by Burton et al reported reductions in the toxicity of *Cry1Ac* proteins when making changes in domain III through mutagenesis. The obtained variants showed decreases in the toxic activity compared with *M. sexta*, strongly suggesting a loss in the binding affinity to the receptor.[35,36] In another study conducted by Liu et al, they revealed that the position W544 from β18 to β19 regions plays a fundamental role in maintaining the toxin integrity and made changes in this position. Although the mutant W544F showed no change in its toxicity, which remained the same, it showed greater stability compared with the ultraviolet (UV) radiation exposure than its parental protein.[37] Lucena et al also reported in 2014 the importance of β16-β17 regions. In this study, they made mutations in *Cry1Ac*, increasing the toxic activity in 1.4 times compared with *Spodoptera exigua*, where they conclude the implication of domain III in the aspects mentioned above.[38]

In addition, the mutant S100pop1-b showed no changes in most of the regions mentioned above. However, it exhibited multiple changes throughout the amino acid sequence in regions of domain I other than α4 and α5 helixes, as it is the zone corresponding to the α1 helix with 3 substitutions (NYT-GRL) in positions 25-27. This may suggest a change in the protein toxicity because this helix has also been described in the formation of the oligomer and, additionally, in the binding and stability of the helix to the GPI (glycosylphosphatidylinositol) receptors.[39]

This review exposes that the algorithm generates sequences that could be biologically viable candidates to synthesize in future studies and perform toxicity experiments on them. For example, the sequence S100pop1-1a presents variations in domain II, loop 3, α8 helix, and β4 loop, as well as variations of domain III in β16 region, conserving the regions necessary for pore formation in domain I. These characteristics make it a great candidate as an improved *Cry* variant to control *A. aegypti* or any other biological target.

The number of variations in the sequences reviewed in this section (see Table 16) is explained by the selection process, where the filter used was the truncation penalty score but not the mutation penalty score. We used this filter with the idea of doing future in vitro toxicity evaluation studies. However, despite the high percentage of similarity with respect to *Cry11Aa* protein in wild type, it can be assumed that these in vitro experiments can be carried out.

Conclusions

The genetic algorithm coupled to experimental conditions of DNA shuffling for 3-domain *cry11* genes proved to be a useful computational approach to analyze the incidence of experimental parameters of diversity and selection generation in the obtention of chimeric libraries of *Cry* variants.

Data analysis led to the conclusion that the chimeric libraries of Cry11 variants, ie, *Cry11Aa* and *Cry11Ba* variants, are favored in Diversity with few simulation generations (15 generations) and high mutation rates [0.02, 0.05]; in Energy Delta with many simulation generations (100 generations) and high mutation rates [0.02, 0.05]; in Identity with respect to the parental genes, with low mutation rates [0.001-0.005]; in the decrease of the number of truncated proteins, with low fragmentation lengths (75bp), low mutation rates, and non-homogeneous distribution in the 3 conserved domains, according to the thermodynamics associated to the SANAFold calculations.

Any other way, it was possible to demonstrate from the structural review of a 6-variant sample that the obtained sequences have structural characteristics from the *Cry* protein families of 3 conserved domains. On the other side, a high percentage of identity and similarity was obtained, in relation to the *Cry11Aa* and *Cry11Ba* parental sequences, with variations in structural regions suggesting feasibility to perform in vitro toxicity evaluation studies.

We consider that these findings will be useful for all those working in in silico and in vitro directed evolution methods involving *Cry* proteins of 3 conserved domains.

Author Contributions

EHP-R and PR-V implemented the in silico model, including the design of the genetic algorithm, implementation and the analysis of the obtained results. DAS-B did the analysis of the genetic algorithm and the obtained results. SA-V performed mutagenesis studies in the laboratory. MOS-B and NJR-F performed mutagenesis studies in the laboratory and the analysis of these experimental results. All the authors contributed in the writing and reviewing of the manuscript.

ORCID iDs

Efraín Hernando Pinzón-Reyes https://orcid.org/0000-0002-8131-2772
Daniel Alfonso Sierra-Bueno https://orcid.org/0000-0002-9566-5974
Nohora Juliana Rueda-Forero https://orcid.org/0000-0003-1189-0603

REFERENCES

1. Cobb R, Sun N, Zhao H. Directed evolution as a powerful synthetic biology tool. *Methods*. 2013;60:81-90. doi:10.1016/j.ymeth.2012.03.009.
2. Stemmer WP. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc Natl Acad Sci U S A*. 1994;91:10747-10751. doi:10.1073/pnas.91.22.10747.
3. Firth AE, Patrick WM. Statistics of protein library construction. *Bioinformatics*. 2005;21:3314-3315. doi:10.1093/bioinformatics/bti516.
4. Patrick WM, Firth AE, Blackburn JM. User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Eng*. 2003;16:451-457. doi:10.1093/protein/gzg057.
5. Sun F. Modeling DNA shuffling. *J Comput Biol*. 1999;6:77-90. doi:10.1089/cmb.1999.6.77.
6. Moore GL, Maranas CD. Predicting out-of-sequence reassembly in DNA shuffling. *J Theor Biol*. 2002;219:9-17. doi:10.1016/S0022-5193(02)93102-4.
7. Moore GL, Maranas CD. eCodonOpt: a systematic computational framework for optimizing codon usage in directed evolution experiments. *Nucleic Acids Res*. 2002;30:2407-2416.
8. Maheshri N, Schaffer D. Computational and experimental analysis of DNA shuffling. *Proc Natl Acad Sci USA*. 2003;100:3071-3076. doi:10.1073/pnas.0537968100.
9. He L, Friedman AM, Bailey-Kellogg C. Algorithms for optimizing cross-overs in DNA shuffling. *BMC Bioinformatics*. 2012;13:S3. doi:10.1186/1471-2105-13-S3-S3.
10. Wedge DC, Rowe W, Kell DB, Knowles J. In silico modelling of directed evolution: implications for experimental design and stepwise evolution. *J Theor Biol*. 2009;257:131-141. doi:10.1016/j.jtbi.2008.11.005.
11. Bravo A, Gómez I, Porta H, et al. Evolution of *Bacillus thuringiensis* Cry toxins insecticidal activity. *Microb Biotechnol*. 2013;6:17-26. doi:10.1111/j.1751-7915.2012.00342.x.
12. Bravo A, Likitvivatanavong S, Gill SS, Soberon M. *Bacillus thuringiensis*: a story of a successful bioinsecticide. *Insect Biochem Mol Biol*. 2011;41:423-431. doi:10.1016/j.ibmb.2011.02.006.
13. Swamy HMM, Asokan R, Rajasekaran PE, Mahmood R, Nagesha SN, Arora DK. Analysis of opportunities and challenges in patenting of *Bacillus thuringiensis* insecticidal crystal protein genes. *Recent Pat DNA Gene Seq*. 2012;6:64-71. doi:10.2174/187221512799303181.
14. Gutierrez P, Alzate O, Orduz S. A theoretical model of the tridimensional structure of *Bacillus thuringiensis* subsp. medellin Cry 11Bb toxin deduced by homology modelling. *Mem Inst Oswaldo Cruz*. 2001;96:357-364. doi:10.1590/S0074-02762001000300013.
15. Lassner M, Bedbrook J. Directed molecular evolution in plant improvement. *Curr Opin Plant Biol*. 2001;4:152-156.
16. Craveiro KIC, Gomes Júnior JE, Silva MC, et al. Variant Cry1Ia toxins generated by DNA shuffling are active against sugarcane giant borer. *J Biotechnol*. 2010;145:215-221. doi:10.1016/j.jbiotec.2009.11.011.
17. Siemens G, Baker RSJD. Learning analytics and educational data mining: towards communication and collaboration. Paper presented at: LAK '12: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge; April 29, 2012:252-254; Vancouver, BC, Canada. doi:10.1145/2330601.2330661.
18. Storer NP, Babcock JM, Schlenz M, et al. Discovery and characterization of field resistance to Bt maize: *Spodoptera frugiperda* (Lepidoptera: Noctuidae) in Puerto Rico. *J Econ Entomol*. 2010;103:1031-1038. doi:10.1603/ec10040.
19. Tabashnik BE, Gassmann AJ, Crowder DW, Carriére Y. Insect resistance to Bt crops: evidence versus theory. *Nat Biotechnol*. 2008;26:199-202. doi:10.1038/nbt1382.
20. Pinzón EH, Sierra DA, Suarez MO, Orduz S, Florez AM. DNA secondary structure formation by DNA shuffling of the conserved domains of the Cry protein of *Bacillus thuringiensis*. *BMC Biophys*. 2017;10:4-10. doi:10.1186/s13628-017-0036-7.
21. Zuker M, Jaeger JA, Turner DH. A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucleic Acids Res*. 1991;19:2707-2714. doi:10.1093/nar/19.10.2707.
22. SantaLucia J Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA*. 1998;95:1460-1465. doi:10.1073/pnas.95.4.1460.
23. Florez AM, Suarez-Barrera MO, Morales GM, et al. Toxic activity, molecular modeling and docking simulations of *Bacillus thuringiensis* Cry11 toxin variants obtained via DNA shuffling. *Front Microbiol*. 2018;9:2461. doi:10.3389/fmicb.2018.02461.
24. Alzate O, Osorio C, Florez AM, Dean DH. Participation of valine 171 in α-helix 5 of *Bacillus thuringiensis* Cry1Ab δ-endotoxin in translocation of toxin into *Lymantria dispar* midgut membranes. *Appl Environ Microbiol*. 2010;76:7878-7880. doi:10.1128/AEM.01428-10.
25. Currin A, Swainston N, Day PJ, Kell DB. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem Soc Rev*. 2015;44:1172-1239.
26. Schnepf E, Crickmore N, Van Rie J, et al. *Bacillus thuringiensis* and its pesticidal crystal proteins. *Microbiol Mol Biol Rev*. 1998;62:775-806. doi:10.1128/MMBR.62.3.775-806.1998.
27. Yamagiwa M, Sakagawa K, Sakai H. Functional analysis of two processed fragments of *Bacillus thuringiensis* Cry11A toxin. *Biosci Biotechnol Biochem*. 2004;68:523-528. doi:10.1271/bbb.68.523.
28. Arenas I, Bravo A, Soberon M, Gomez I. Role of alkaline phosphatase from *Manduca sexta* in the mechanism of action of *Bacillus thuringiensis* Cry1Ab toxin. *J Biol Chem*. 2010;285:12497-12503. doi:10.1074/jbc.M109.085266.
29. Girard F, Vachon V, Préfontaine G, et al. Helix 4 of the *Bacillus thuringiensis* Cry1Aa toxin plays a critical role in the postbinding steps of pore formation. *Appl Environ Microbiol*. 2009;75:359-365. doi:10.1128/AEM.01930-08.
30. Fujii Y, Tanaka S, Otsuki M, et al. Cry1Aa binding to the cadherin receptor does not require conserved amino acid sequences in the domain II loops. *Biosci Rep*. 2013;33:103-112. doi:10.1042/BSR20120113.
31. Gómez I, Dean DH, Bravo A, Sobero M. Molecular basis for *Bacillus thuringiensis* Cry1Ab toxin specificity: two structural determinants in the *Manduca sexta* Bt-R1 receptor interact with loops α-8 and 2 in domain II of Cy1Ab toxin. *Biochemistry*. 2003;42:10482-10489.
32. Fernández LE, Pérez C, Segovia L, et al. Cry11Aa toxin from *Bacillus thuringiensis* binds its receptor in *Aedes aegypti* mosquito larvae through loop α-8 of domain II. *FEBS Lett*. 2005;579:3508-3514. doi:10.1016/j.febslet.2005.05.032.

33. Li S, Wang Z, Zhou Y, et al. Expression of cry2Ah1 and two domain II mutants in transgenic tobacco confers high resistance to susceptible and Cry1Ac-resistant cotton bollworm. *Sci Rep*. 2018;8:508-511. doi:10.1038/s41598-017-19064-5.

34. Pacheco S, Gómez I, Arenas I, et al. Domain II loop 3 of *Bacillus thuringiensis* Cry1Ab toxin is involved in a "ping pong" binding mechanism with *Manduca sexta* aminopeptidase-N and cadherin receptors. *J Biol Chem*. 2009;284:32750-32757. doi:10.1074/jbc.M109.024968.

35. Burton SL, Ellar DJ, Li J, Derbyshire DJ. N-acetylgalactosamine on the putative insect receptor aminopeptidase N is recognised by a site on the domain III lectin-like fold of a *Bacillus thuringiensis* insecticidal toxin. *J Mol Biol*. 1999;287:1011-1022.

36. Xu C, Wang BC, Yu Z, Sun M. Structural insights into *Bacillus thuringiensis* Cry, Cyt and parasporin toxins. *Toxins*. 2014;6:2732-2770. doi:10.3390/toxins6092732.

37. Liu YL, Wang QY, Wang FX, Ding XZ, Xia LQ. Residue 544 in domain III of the *Bacillus thuringiensis* Cry1Ac toxin is involved in protein structure stability. *Protein J*. 2010;29:440-444. doi:10.1007/s10930-010-9271-3.

38. Lucena WA, Pelegrini PB, Martins-de-Sa D, et al. Molecular approaches to improve the insecticidal activity of *Bacillus thuringiensis* Cry toxins. *Toxins*. 2014;6:2393-2423. doi:10.3390/toxins6082393.

39. Yamaguchi T, Bando H, Asano S. Identification of a *Bacillus thuringiensis* Cry8Da toxin-binding glucosidase from the adult Japanese beetle, *Popillia japonica*. *J Invertebr Pathol*. 2013;113:123-128. doi:10.1016/j.jip.2013.03.006.