Modelo In Silico de Evolución Dirigida para Genes Cry11 de Bacillus Thuringiensis

Efraín Hernando Pinzón Reyes

Trabajo de grado para optar al título de

Doctor en Ingeniería. Área Electrónica

Director:

Daniel Alfonso Sierra Bueno, Ph.D.

Codirector:

Álvaro Mauricio Florez Escobar, Ph.D.

Universidad Industrial de Santander

Facultad de Ingenierías Físico Mecánicas

Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones

Bucaramanga

2017

DEDICATORIA

A Dios y a mi hermosa familia,

Mi amada esposa *Lina Marie* y a nuestros tesoros,

Mi lluvia temprana *Abdeel David* y mi lluvia tardia *Lina Antonella*.

AGRADECIMIENTOS

A todos aquellos que han hecho parte de este proyecto de vida, mis padres Javier Pinzón y Merida Reyes, demás familiares y amigos.

Al Dr. Daniel Alfonso Sierra Bueno por su apoyo y dirección, gracias por los consejos y orientaciones tan acertadas.

Al Dr. Alvaro Mauricio Florez Escobar por brindarme la oportunidad de iniciarme en la biología computacional, gracias por el apoyo y el camino de vida mostrado.

A los compañeros de los grupos de investigación CEMOS-UIS y BIOMOL-UDES por sus valiosos aportes y disertaciones que hacen parte de lo aprendido.

A los miembros del Centro de Bioinformática y Simulación Molecular de la Universidad de Talca-Chile, en especial al Dr. Jans Alzate y al Dr. Francisco Adasme, por la fraternal acojida y acompañamiento en mi estancia Doctoral.

A las diversas instituciones como Colciencias, UIS, UDES y la Plataforma Alianza del Pacifico por brindar los recursos y los espacios para mi formación Doctoral.

Tabla de Contenido

Pag.

Introducción 15
1. Evolución Dirigida 19
1.1 Modelos in silico de ED 21
1.2 Conclusiones 25
2. Generación de Diversidad 27
2.1 Barajado de ADN 27
2.2 Modelos in silico de Barajado de ADN 29
2.2.1 Selección de los genes parentales. 30
2.2.2 Fragmentación de los genes. 33
2.2.3 Desnaturalización. 34
2.2.4 Hibridación. 34
2.2.5 Extensión. 39
2.2.6 Biblioteca Recombinada. 40
2.3 Puntos de quiebre 42
2.4 Conclusiones 44
3. DEVISING: Una estrategia in silico de ED 47
3.1 SANAFold 48
3.1.1 Uso de métodos predictivos de estructuras secundarias de ADN. 49
3.1.2 Arquitectura software de SANAFold. 53
3.2 SAssembly 63
3.2.1 Módulo gestor de condiciones de barajado 64
3.2.2 Módulo gestor de fragmentos. 64
3.2.3 Módulo gestor de hibridación. 66
3.2.4 Módulo gestor de bibliotecas quiméricas. 68
3.3 GenE-in 72
3.3.1 Módulo Evaluador de Eficiencia. 73
3.3.2 Módulo Evaluador de Diversidad. 73

3.3.3 Módulo Descriptor de Contribución. 74
3.4 Conclusiones 74
4. Caso de Estudio: Experimento de Evolución Dirigida in Silico para Toxinas Cry11 de Bacillus Thuringiensis 77
4.1 Modelo biológico: Toxinas Cry11 de Bacillus thuringiensis 77
4.1.1 Modo de Acción de las Toxinas Cry. 78
4.1.2 Ingeniería de proteínas con Toxinas Cry 80
4.2 Experimentos in vitro de ED con Toxinas Cry11 81
4.2.1 Condiciones experimentales. 82
4.2.2 Resultados. 84
4.3 Uso de la estrategia DEVISING 84
4.3.1 Uso de SANAFold. 85
4.3.2 Uso de SAssembly & GenE-in. 99
4.4 Conclusiones 113
5. Conclusiones Generales 116
6. Divulgaciones y Trabajos Dirigidos 118
Referencias Bibliograficas 121
Apendices 131

Lista de Figuras

Pag.

Figura 1. Evolución Dirigida in vitro 20
Figura 2.Técnica del Barajado de ADN 29
Figura 3. Representación de un fragmento de ADN 35
Figura 4. Selectividad de fragmentos 36
Figura 5. Cambio de estados 37
Figura 6. Proceso de Extensión 39
Figura 7. Motivos estructurales de un fragmento de la región I del gen cry11A 43
Figura 8. Hibridación (Participación de las estructuras secundarias de ADN) 44
Figura 9. Comparación entre la Arquitectura de DEVISING y los pasos de ED 48
Figura 10. Arquitectura software de SANAFold 53
Figura 11. Esquema del proceso algorítmico para parametrización de condiciones in vitro 55
Figura 12. Esquema del proceso algorítmico de la fragmentación del ADN 57
Figura 13. Esquema del proceso algorítmico de mínima energía UNAFold 3.8 58
Figura 14. Esquema del proceso algorítmico de cálculos estadísticos y análisis inferencial 60
Figura 15. Arquitectura software de SAssembly 63
Figura 16. Esquema del proceso algorítmico de la gestión de fragmentos 65
Figura 17. Modelo markoviano de conversión de estados (Cadena sencilla, Cadena doble) 67
Figura 18. Esquema del proceso algorítmico de la gestión de bibliotecas quiméricas 70
Figura 19. Arquitectura software de GenE-in 72
Figura 20. Filogenia de las toxinas Cry y letalidad asociada 78
Figura 21. Estructura de la proteína Cry2Ab con alta homología con Cry11Aa 80
Figura 22. Diversidad y Eficiencia de barajado de ADN in vitro con genes cry11 84
Figura 23. Comportamiento de los estimadores estadísticos en los clústeres cry 89
Figura 24. Comparativo termodinámico (Kcal/mol) entre el escenario de referencia y el estado de transición en condiciones LE-MA de barajado de ADN 90
Figura 25. Asociación de la filogenia de los clústeres cry con su comportamiento termodinámico para la formación de estructuras secundarias de ADN. 91

Figura 26. Diversidad y Eficiencia de las mejores variantes cry11 in silico de barajado de ADN 107
Figura 27. Bloques conservados en la conformación de las mejores variantes cry11 in silico...110
Figura 28. Escenarios de simulación in silico vs. Longitud de las mejores variantes Cry11 in silico 111
Figura 29. Temperatura de alineamiento vs. Longitud de las mejores variantes Cry11 in silico. 112
Figura 30. Longitud inicial de fragmentación vs. Longitud de las mejores variantes Cry11 in silico. 113

Lista de Tablas

Pag.

Tabla 1. Primers para la obtención de la biblioteca quimérica.....83
Tabla 2. Clústeres de genes cry estudiados. 87
Tabla 3. Valores f-ratio con diferencia estadística significativa (regiones génicas de cada gen). 88
Tabla 4. Bibliotecas quiméricas in silico de genes cry11 (36 escenarios de simulación).....101
Tabla 5. Mejores variantes cry11 in silico 103
Tabla 6. Bloques conservados de las mejores variantes cry11 in silico (Barajado de ADN) 108

Lista de Apendices

Pag.

Apendice A. Diagramas de Casos de Uso 131
Apendice B. Análisis de Costos de Experimentos in Vitro e in Silico de Barajado de ADN 137

Resumen

Título: Modelo *In Silico* de Evolución Dirigida para Genes *Cry11* de *Bacillus Thuringiensis*[*]

Autor: Efraín Hernando Pinzón Reyes[**]

Palabras Claves: Modelo *In Silico*, Evolución Dirigida, Barajado de Adn, Estructuras Secundarias de Adn, *Bacillus Thuringiensis.*

Descripción:

El presente trabajo de investigación propone un modelo *in silico* del proceso completo de la técnica de evolución dirigida. El modelo integra la selección de genes parentales, la generación de diversidad mediante el barajado de ADN y la selección de variantes candidatas. Para ello, el modelo aprovecha el método de mínima energía como elemento integrador mediante el cual puede también incorporar los efectos de la formación de estructuras secundarias de ADN en el proceso. El modelo fue usado para estudiar la familia de genes *cry11 de Bacillus thuringiensis* y predecir bibliotecas quiméricas de genes ensamblados. Dentro de los principales hallazgos se encuentran la obtención de bibliotecas quiméricas *in silico* de potenciales variantes *Cry11* y la caracterización de los genes desde sus propiedades intrínsecas para reaccionar ante condiciones experimentales de evolución dirigida, explicada desde una perspectiva evolutiva entre las diferentes familias *Cry*. Estas innovaciones trazan dos nuevas líneas de trabajo: la evolución dirigida *in silico* y la caracterización de genes parentales a partir de sus variaciones termodinámicas para participar de forma eficiente en experimentos de barajado de ADN. Ambas líneas fortalecen desde lo computacional el campo de estudio de la ingeniería de proteínas, superando las limitaciones de los modelos de *mutagénesis asistida por computador*, cuyo punto de partida son las bibliotecas quiméricas *in vitro*, mientras que el modelo aquí reportado permite la predicción de estas bibliotecas quiméricas y su posterior análisis, siendo la primer aproximación de la cual se tiene conocimiento para realizar evolución dirigida *in silico*.

---

[*] **Trabajo de grado**

[**] **Facultad de Ingenierías Físico Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Director: Daniel Alfonso Sierra Bueno, Ph.D. Codirector: Álvaro Mauricio Florez Escobar, Ph.D.**

Abstract

Title: *In Silico* Model of Directed Evolution for *Cry11* Genes of *Bacillus Thuringiensis.*[*]

Author: Efraín Reyes Hernando Pinzón[**]

Keywords: *In Silico* Model, Directed Evolution, Dna Shuffling, Dna Secondary Structures, *Bacillus Thuringiensis*.

Description:

This doctoral thesis proposes an *in silico* model of the entire process of directed evolution technique. This includes the selection of parental genes, the generation of diversity by DNA shuffling and selection of candidate variants. To do so, the model takes advantage on the method of minimum energy as an integrating element by which the effects of the formation of DNA secondary structures in the process can be also incorporated. The model was used to study *cry11* family of genes from *Bacillus thuringiensis* and predict assembled libraries of chimeric genes. Among the main findings are the obtention of chimerical *in silico* libraries of potential Cry11 variants and the characterization of genes from their intrinsic properties to react under experimental conditions on directed evolution, explained from an evolutionary perspective between different families Cry obtained. These innovations draw two new lines of work: *in silico* directed evolution and characterization of parental genes from their thermodynamic variations to participate efficiently in DNA shuffling experiments. Both lines strength, from the computational field, the study of protein engineering, overcoming the limitations of the mutagenesis models assisted by computer whose starting point is the *in vitro* chimeric libraries, while the model reported here allows the prediction of these chimeric libraries and subsequent analysis, being the first approximation which is known for directed evolution *in silico*.

[*] **Bachelor Thesis**

[**] **Facultad de Ingenierías Físico Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Director: Daniel Alfonso Sierra Bueno, Ph.D. Codirector: Álvaro Mauricio Florez Escobar, Ph.D.**

Introducción

La ingeniería de proteínas explora mediante técnicas biomoleculares la obtención de nuevas proteínas funcionalmente mejoradas, para ello, usa el conocimiento funcional de la proteína silvestre y realiza modificaciones estructurales racionales o introduce cambios aleatorios, recreando mediante diseños experimentales, la evolución darwiniana. En ambos casos, se espera la mejora funcional de las nuevas proteínas. La técnica más ampliamente usada en ingeniería de proteínas es la evolución dirigida (ED), que permite una mayor exploración del espacio de posibles variantes.

Los modelos *in silico* han sido grandes aliados para los ingenieros de proteínas, en el diseño de experimentos que optimizan la obtención de bibliotecas quiméricas, implementando criterios de eficiencia y diversidad. Sin embargo, los trabajos de modelado *in silico* de ED que imitan el proceso completo presentan la limitación de no contemplar elementos propios de la composición del ADN y los hallazgos obtenidos por estos, no dejan de ser simples recomendaciones generales sobre algunos parámetros experimentales. Mientras que otras aproximaciones, que si contemplan elementos propios de la composición del ADN, descuidan procesos biológicos fundamentales como la formación de estructuras secundarias de ADN y sus implicaciones en la formación de variantes, a la vez que son trabajos muy específicos, centrados en simular la generación de diversidad de la ED, y sus resultados no pasan de ser recomendaciones acerca de condiciones experimentales óptimas.

Este trabajo es el primero en desarrollar un modelo *in silico* de ED completo, que incorpora elementos propios de la composición del ADN y contempla los efectos de la formación de estructuras secundarias en la generación de variantes con mejoras funcionales. Este trabajo es pionero, según tenemos conocimiento, en la implementación de los métodos de mínima energía como estrategia de modelado integral, en el uso de una arquitectura hardware tipo cluster para simular procesos completos de ED y en lograr la fabricación de variantes mejoradas 100% *in silico*. Como resultado de esta investigación, sentamos las bases para plantear una nueva forma de hacer ED, que pone en un nuevo contexto el uso del modelado computacional, ya no restringido a dar soporte al diseño experimental, sino ahora como un proveedor de insumos en el proceso mismo de expimentos de ED *in vitro*.

Para nuestros experimentos *in vitro* e *in silico* usamos toxinas Cry11 de *Bacillus thuringiensis*, conocidas por su potencial bioinsecticida contra vectores propagadores de enfermedades. Nuestros resultados con toxinas Cry nos permitieron sentar las bases para una nueva línea de trabajo, tecnológicamente facultada para caracterizar de forma sistemática las toxinas Cry como insumos de experimentos futuros de ED. Las primeras caracterizaciones de toxinas Cry11 han sido de utilidad para encontrar variantes con potencial funcional; estos primeros resultados ejemplifican la contribución más representativa de este trabajo, es decir, la fabricación de toxinas recombinadas *in silico*.

Este documento está compuesto por cuatro capítulos principales, conclusiones y divulgaciones. En el capítulo 1, denominado **Evolución Dirigida**, se describe el potencial de la

técnica biomolecular para explorar nuevas variantes y se exponen los modelos computacionales usados para estudiar el proceso completo de ED, con sus respectivas limitaciones.

En el capítulo 2, denominado **Generación de Diversidad**, se plantea cómo las técnicas ADN recombinantes son una mejor alternativa que las técnicas mutagénicas aleatorias para generar diversidad. Se explica el proceso de la técnica ADN recombinante por excelencia, el barajado de ADN, y se presentan los métodos computacionales usados para su simulación, desde una perspectiva crítica de sus principales aportes y limitantes.

En el capítulo 3, denominado **DEvISING: Una estrategia *in silico* de ED**, se presenta nuestra propuesta *in silico*, incluyendo los métodos computacionales implementados y su arquitectura software, basada en tres componentes principales (SANAFold, SAssembly, GenE-in), donde cada componente software simula cada paso experimental de ED, articulando trece módulos constitutivos.

En el capítulo 4, denominado **Caso de estudio, experimento de ED *in silico* para toxinas Cry11 de Bt,** se presenta el modelo biológico de las toxinas Cry, un experimento de ED *in vitro* de toxinas Cry11 llevadas a cabo por nuestro grupo, y se ejemplifica el uso modular de DEvISING. Los resultados *in silico* de esta ejemplificación son interpretados desde un enfoque biológico. El capítulo se cierra con la obtención de variantes Cry11 *in silico*.

Finalmente, en la sección de **conclusiones** se presentan de forma condensada nuestros principales hallazgos y en **divulgaciones** se enuncian los productos de difusión y trabajos dirigidos en el transcurso del desarrollo de la investigación.

1. Evolución Dirigida

La evolución dirigida (ED) es un conjunto de técnicas que imitan la evolución darwiniana a escala de laboratorio (Cobb, Sun, & Zhao, 2013) y permiten, sin un conocimiento profundo de la codificación de una proteína, descubrir nuevas proteínas útiles (Romero & Arnold, 2009). Este conjunto de técnicas se soportan en rondas iterativas de mutación y selección artificial cuyo propósito es mejorar características de las proteínas parentales proporcionadas por la evolución natural como: solubilidad, termo-estabilidad, afinidad con el sustrato y actividad catalítica (Packer & Liu, 2015).

La ED junto con técnicas de diseño racional de proteínas son el arsenal de herramientas con las cuales la ingeniería de proteínas enfrenta el desafío de encontrar proteínas útiles mejoradas (Lane & Seelig, 2014). Sin embargo, debido a la complejidad de los sistemas biológicos, se dificulta el diseño racional que demanda un conocimiento previo de la codificación y funcionalidad de una proteína, haciendo de la ED una herramienta valiosa para la biología sintética (Cobb, Si, & Zhao, 2012).

El proceso *in vitro* de ED se puede describir en tres pasos iterativos (Figura 1): 1) Selección de unos genes parentales; 2) Generación de diversidad genética y 3) Selección de genes mutados (Leemhuis, Kelly, & Dijkhuizen, 2009; Packer & Liu, 2015).

Selección de unos genes parentales (Figura 1): A partir de un grupo de genes de interés biotecnológico se realiza una selección de los genes parentales (genes de referencia) que codifican proteínas con características biotecnológicas que se desea sean mejoradas. Criterios como homología de los genes parentales o la selección de genes que codifiquen la característica deseada en una tasa baja, son por lo general criterios que hacen exitosa la creación de variantes deseables (Leemhuis et al., 2009).

Diversidad Genética (Figura 1): Este paso es fundamental en ED pues es allí donde las mutaciones de la evolución darwiniana son emuladas *in vitro*. Esto se logra a través de técnicas de mutagénesis aleatoria o técnicas ADN recombinantes, a partir de las cuales se obtienen bibliotecas quiméricas, conformadas por genes mutados o genes recombinados, según la técnica implementada (Packer & Liu, 2015).

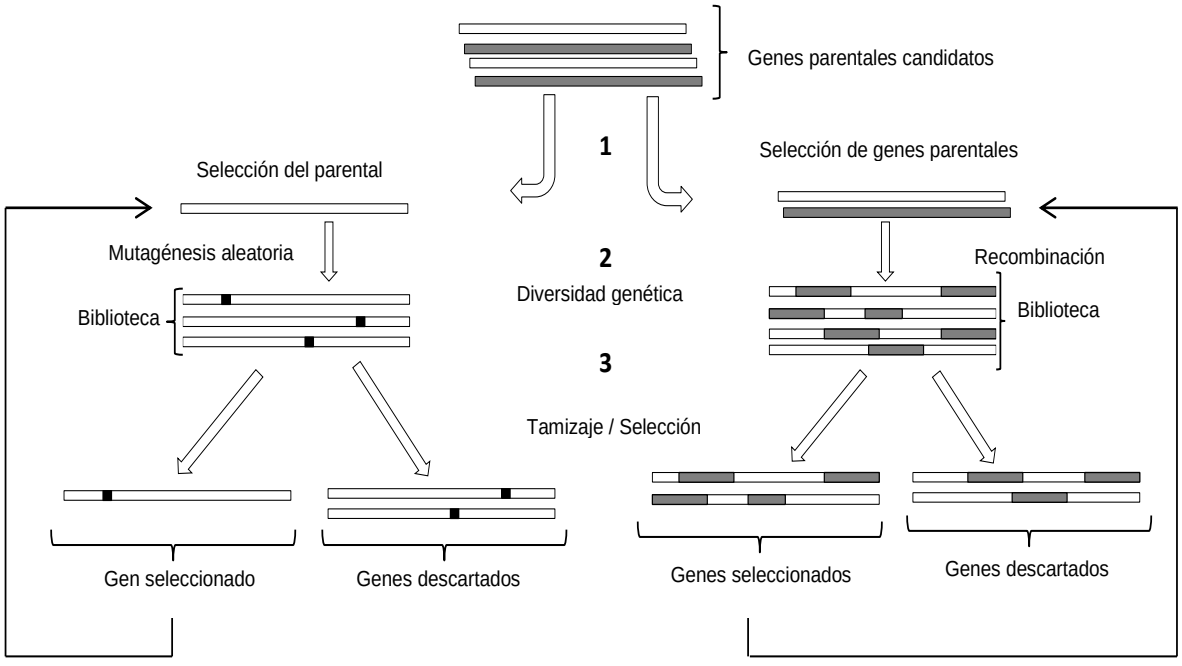


Figura 1. Evolución Dirigida in vitro

Selección de genes mutados/recombinados (Figura 1): Después de creadas las bibliotecas quiméricas, se realiza un proceso de depuración, mediante un diseño de selección artificial (tamizaje), basado en criterios de mejoras potenciales de los genes quiméricos, por ejemplo: la longitud de los genes quiméricos, ya que esto aumenta la probabilidad de codificación de la proteína deseada; o la presencia de mutaciones o material genético de interés evidenciado mediante la observación de marcadores. De esta forma se espera obtener genes quiméricos prometedores (Packer & Liu, 2015).

Estos pasos suelen hacerse con varios ciclos de iteración (Figura 1), sin embargo, la realización de un ciclo de ED con alta presión selectiva y varios ciclos de ED con baja presión selectiva, han mostraron ser estrategias eficientes *in vitro* (Lane & Seelig, 2014).

1.1 Modelos in silico de ED

El espacio de búsqueda de posibles proteínas está dado por un número de 20^{L} posibles configuraciones de aminoácidos, donde L es la longitud de la secuencia de aminoácidos de una proteína de interés. Una pequeña proteína de 100 aminoácidos genera un espacio de búsqueda de 20^{100} ($\approx 10^{130}$) secuencias (Romero & Arnold, 2009). Pese a esta gran complejidad, el algoritmo darwiniano mutación-selección *in vitro* de la ED ha resultado ser un método poderoso de exploración para generar proteínas funcionales (Romero & Arnold, 2009).

Con el fin de refinar la exploración del espacio de búsqueda, la ED ha involucrado estrategias acompañadas de modelos *in silico* (Verma, Schwaneberg, & Roccatano, 2012),

configurando el área de trabajo de la ED de proteínas asistida por computadora, CAPDE (*Computer–aided protein directed evolution*) (Verma et al., 2012).

Una revisión reciente de herramientas CAPDE logró clasificar 53 trabajos computacionales que incluían aplicaciones software, servidores web y bases de datos para atender cuatro aspectos de la ED: i) Caracterización estadística de las bibliotecas generadas; ii) Análisis de bibliotecas focalizadas en la conservación evolutiva; iii) Análisis de bibliotecas focalizadas en la estructura; iv) Efectos de las mutaciones en la proteína (Verma et al., 2012). Sin embargo, en esta revisión no se incluyeron los trabajos computacionales que estudian la formación de las bibliotecas quiméricas, probablemente por considerar que los avances actuales en secuenciación y síntesis del ADN asociado a su costo decreciente facilitan la creación de bibliotecas quiméricas (Cobb et al., 2012; Romero & Arnold, 2009). En contraposición, se focalizó la asistencia computacional a la exploración del nuevo espacio de búsqueda de las bibliotecas ya generadas ($<< 20^L$ secuencias posibles).

Por otra parte, otros modelos *in silico* han sido desarrollados para estudiar los efectos de parámetros como tasa de mutación, recombinación y presión de selección en el funcionamiento de la evolución dirigida (Fox, 2005; Fox et al., 2003; Peng, Levine, Hwa, & Kessler, 2004; Wedge, Rowe, Kell, & Knowles, 2009). Estos estudios usaron algoritmos genéticos que simularon los pasos de la ED, para recorrer modelos de paisajes NK (Fox, 2005; Fox et al., 2003; Wedge et al., 2009), o implementaron modelos evolutivos multi-locus, propios de los estudios de genética de poblaciones para revisar los efectos de la recombinación (Peng et al., 2004).

Los modelos de paisajes NK son útiles para describir bibliotecas de proteínas, donde cada secuencia de aminoácidos de longitud N es representada como un sistema binario con K posiciones interactuando. Cada posición representa un aminoácido que interactúa con K aminoácidos de la misma secuencia, representando el fenómeno biológico de epistasis. De esta forma se logra la representación de paisajes con 2^N posibilidades (Wedge et al., 2009).

Con estos paisajes establecidos se usan algoritmos genéticos (μ, λ), para explorar las soluciones optimas. Los algoritmos genéticos emulan el método de mutación-selección de la ED mediante los siguientes pasos: i) Generar una población λ de secuencias al azar, ii) Evalúar las aptitudes de las secuencias generadas, según las reglas del paisaje NK, iii) Seleccionar las μ mejores secuencias de la población, iv) Generar la descendencia mediante operadores de cruce y mutaciones, v) Agregar población haciendo (iv) hasta que la población sea nuevamente λ. Estos pasos se realizan hasta cumplir el criterio de parada, por ejemplo, haber generado un número de 10 poblaciones (Wedge et al., 2009).

Para que el algoritmo genético pueda explorar soluciones óptimas, se debe establecer una función de aptitud por cada secuencia (Ecuación 1). En estos estudios se obtuvo mediante la suma de las aptitudes parciales de cada posición variable de la proteína W_i, que fueron calculadas a partir de tablas de transición de estado previamente establecidas para las K interacciones (Fox et al., 2003).

$$W = \frac{100}{N} \sum_{i=1}^{N} W_i \qquad (1)$$

Los paisajes NK simulados han permitido entender que para ED: i) Son beneficiosas las tasas de mutación moderadamente altas en presencia de fuerte presión selectiva, ii) La inclusión de métodos de recombinación hacen que la tasa de mutación óptima sea menos relevante, iii) La eficiencia se mide maximizando la creación de diversidad y minimizando el tamaño de las bibliotecas obtenidas (Wedge et al., 2009). Estas son apreciaciones que corresponden a evidencia del experimento ED *in vitro* (Currin, Swainston, Day, & Kell, 2015).

Por su parte los modelos evolutivos multi-locus, permiten recrear técnicas ADN recombinantes y evaluar los efectos de estas técnicas en ED. Para ello se establece un modelo de dos estados que puede representar secuencias de ADN, en donde cada posición (x) de la secuencia génica puede tomar el valor de 1 si la contribución del nucleótido es favorable y (0) si es desfavorable. Con este sistema binario representativo por secuencia de ADN se puede calcular la energía de unión de la secuencia, que es simplemente el número de sitios con nucleótidos favorables (Ecuación 2) (Peng et al., 2004).

$$m = \sum_{i=1}^{N} x_i \qquad (2)$$

El esquema de selección del modelo evolutivo multi-locus obedece a una función de truncamiento dinámico en donde las secuencias que tengan una energía de unión m < m0 (energía de umbral) son secuencias descartadas para participar en iteraciones posteriores. La fuerza de selección de una población está dada por la fracción Φ de secuencias seleccionadas de una población (Ecuación 3), donde P_m es una distribución de la población en términos de la

energía de unión (m=0,1, …, L), L es el número total de sitios activos y α contempla la selección parcial de algunas secuencias en estado m0 (Peng et al., 2004). El modelo incluye como método para simular la recombinación una tasa de mutación μ0 que es usada por cada nucleótido de forma independiente para mutar en una nueva generación. El número de generaciones de simulación es el criterio de parada del algoritmo (Peng et al., 2004).

$$\phi = \alpha P_{m0} + \sum_{m=m0+1}^{L} P_m \quad (3)$$

Sobre el efecto de las técnicas de recombinación en modelos de ED, los modelos evolutivos multi-locus permiten observar importantes conclusiones: i) la recombinación acelera el proceso de evolución, ii) y presenta mejor rendimiento que la sola mutación para los diferentes esquemas de selección, iii) cuanto menos recombinaciones por secuencia menor es el rendimiento evolutivo (Peng et al., 2004). Estas afirmaciones corresponde con estudios de ED *in vitro*, donde se ha observado que las técnicas ADN recombinantes amplían el espectro de búsqueda (Currin et al., 2015) y combinan de forma efectiva mutaciones mientras purgan mutaciones perjudiciales. Estos resultados no se pueden lograr con métodos no recombinantes (Leemhuis et al., 2009).

1.2 Conclusiones

La búsqueda de nuevas proteínas con fines biotecnológicos es un problema de alta complejidad y de interés creciente, que encuentra en el método de mutación-selección darwiniano de la ED una

alternativa plausible para explorar el amplio espectro de posibilidades. Modelos *in silico* han sido desarrollados para representar mediante algoritmos genéticos y modelos evolutivos multi-locus los procesos de ED, y han sido usados con el solo propósito de observar efectos de parámetros propios del método de evolución darwiniana tales como: presión de selección, mutación y recombinación, sobre la eficiencia en la generación de variantes. Estos modelos *in silico* permitieron construir un sustento teórico para mostrar los beneficios de las técnicas ADN recombinantes, en la búsqueda del amplio espectro de nuevas variantes por ED.

Por otra parte, se observó un gran número de trabajos computacionales que han sido integrados de forma conceptual como herramientas CAPDE (*Computer–aided protein directed evolution*). Esta integración es realizada a nivel teórico, para mediante una categorización, dar forma al gran volumen de trabajos desagregados del campo de estudio. Estos trabajos computacionales fueron diseñados para dar apoyo especializado en el proceso experimental de ED y asumen la construcción *in vitro* de las bibliotecas quiméricas, excluyendo de su arsenal los modelos *in silico* de creación de bibliotecas.

En el actual campo de estudio, nosotros proponemos el desarrollo de una estrategia *in silico* integrada de ED, que permita: i) Seleccionar genes parentales, que hagan eficiente la creación de diversidad; ii) Generar diversidad genética, mediante la aplicación de técnicas ADN recombinantes; y iii) Seleccionar variantes génicas que se puedan sintetizar *in vitro*.

2. Generación de Diversidad

La generación de diversidad es el paso que hace potente la técnica de ED dado que, como se expuso en la sección anterior, la mutagénesis (por recombinación o aleatoria) es la opción más prometedora en ausencia de información previa de la proteína blanco (Packer & Liu, 2015). La mutagénesis se implementa en el laboratorio mediante variaciones a la técnica tradicional de PCR (*Polymerase Chain Reaction*) que fue diseñada originalmente para la replicación exacta de ADN (Mullis et al., 1986). Contrario a su interés original, la mutagénesis aleatoria utiliza PCR propensas a error (Leemhuis et al., 2009), mientras que la técnica de recombinación modifica audazmente las PCR al eliminar el uso de cebadores (Stemmer, 1994).

Dado que las técnicas ADN recombinantes han resultado ser más eficientes en la exploración del amplio espacio de búsqueda de diversidad (Currin et al., 2015; Leemhuis et al., 2009; Moore & Maranas, 2002; Peng et al., 2004), esta sección se centrará en la técnica ADN recombinante denominada barajado de ADN (*DNA shuffling*), que ha resultado ser la técnica recombinante más ampliamente aceptada para la generación de diversidad en experimentos de ED (Moore & Maranas, 2002; Packer & Liu, 2015; Stemmer, 1994).

2.1 Barajado de ADN

El Barajado de ADN es una técnica propuesta por Stemmer que permite crear una biblioteca de genes recombinados. Para ello, se realizan los pasos de la PCR tradicional con una variación

importante, la exclusión de cebadores. Los cebadores en un PCR tradicional sirven como marcadores para la reconstrucción con alto grado de fidelidad de la cadena original de ADN. Al eliminar los cebadores, esta función es realizada por pequeños fragmentos de las mismas secuencias de ADN que están siendo barajadas, de tal forma que el reensamble de los genes se hace de forma aleatoria promoviendo la recombinación de los diferentes fragmentos (Stemmer, 1994).

Se pueden establecer seis pasos en la técnica de barajado de ADN *in vitro* (Figura 2): 1) Selección de los genes parentales: donde el principal criterio es que exista homología entre los genes con el propósito de generar suficiente diversidad, sin perder la funcionalidad en las bibliotecas resultantes; 2) Fragmentación de los genes parentales: se hace como un preparativo a los ciclos de PCR, donde el ADN es sometido a la acción de una nucleasa (DNasa I) que corta de forma aleatoria las cadenas de ADN. Después de ello se someten estos fragmentos a ciclos iterativos de PCR sin cebadores, en tres pasos: desnaturalización, hibridación y extensión; 3) Desnaturalización: los fragmentos de ADN se someten a temperaturas altas, cercanas a los 94°C, para romper los enlaces de hidrógeno de las hebras dobles de ADN, obteniendo hebras sencillas de ADN; 4) Hibridación: en temperaturas que dependen de las características de los genes parentales, los fragmentos de hebra sencilla de ADN buscan complementariedad para unirse entre ellos en esas regiones específicas; 5) Extensión: en temperaturas cercanas a los 72°C los fragmentos que han hibridado son complementados en sus nucleótidos faltantes para conformar nuevas estructuras de ADN de hebra doble; 6) Después de varios ciclos de iteración, se espera tener una población de genes recombinados cuya longitud sea cercana a la longitud de los genes

parentales, esperando garantizar así la funcionalidad de las proteínas que esos genes puedan codificar.

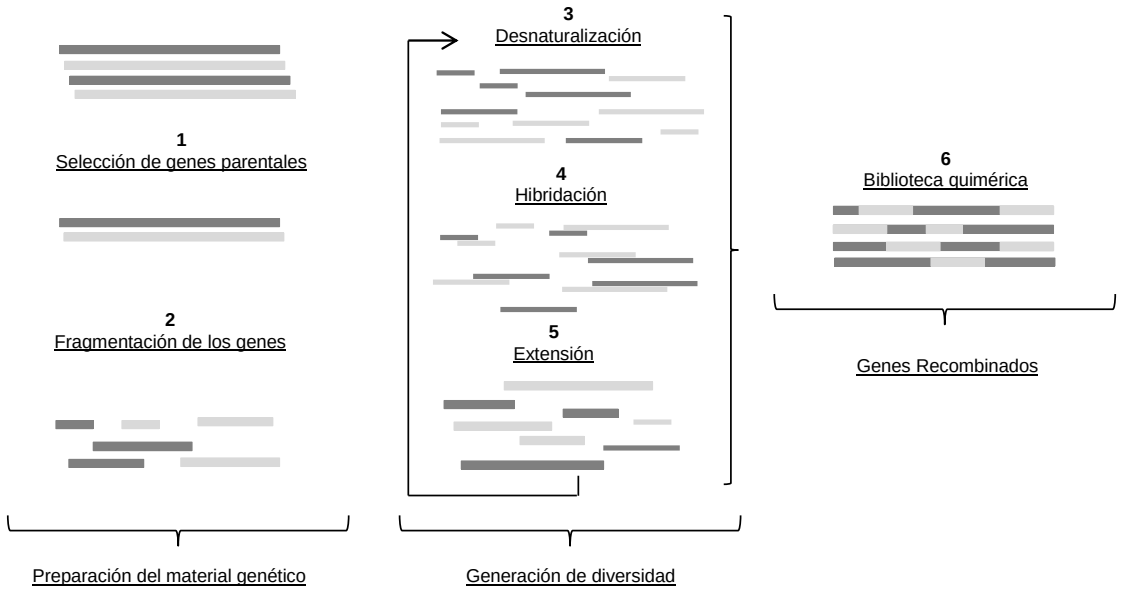


*Figura 2.*Técnica del Barajado de ADN

2.2 Modelos in silico de Barajado de ADN

Los modelos *in silico* de barajado de ADN se pueden categorizar desde su intencionalidad y desde su método. Desde su intencionalidad, los primeros trabajos *in silico* buscaron replicar el proceso de barajado de ADN y capturar la complejidad propia del instante de la hibridación de fragmentos, para dar luces sobre los parámetros experimentales que optimizan la recombinación (Joern, Meinhold, & Arnold, 2002; Maheshri & Schaffer, 2003; Moore & Maranas, 2000; Sun, 1999). Con los hallazgos realizados en estos primeros trabajos de modelado, el nuevo foco de trabajo se trasladó hacia la caracterización de los genes parentales, asumiendo que la selección adecuada de ellos o el diseño previo de los mismos mejorarían los resultados del barajado (He,

Friedman, & Bailey-Kellogg, 2012; Moore & Maranas, 2002). Finalmente, el análisis detallado de métricas de bibliotecas obtenidas en barajado de ADN *in vitro* cobraron relevancia (Patrick & Firth, 2005; Patrick, Firth, & Blackburn, 2003).

Por otra parte, una revisión de los trabajos realizados desde el método permite la categorización de al menos tres métodos a saber: 1) métodos de predicción físico-química del proceso de barajado (Joern et al., 2002; Maheshri & Schaffer, 2003; Moore & Maranas, 2000), 2) métodos estocásticos (Patrick & Firth, 2005; Patrick et al., 2003; Sun, 1999) y 3) métodos de optimización (He et al., 2012; Moore & Maranas, 2002). Cabe notar que pese a esta clasificación tan determinística, la gran mayoría de trabajos usaron métodos híbridos, combinando métodos de predicción físico-química con métodos estocásticos (Joern et al., 2002; Maheshri & Schaffer, 2003; Moore & Maranas, 2000), o con métodos de optimización (He et al., 2012; Moore & Maranas, 2002).

A continuación se realiza una descripción de los trabajos de barajado de ADN desde el método, ordenando los principales aportes de estos trabajos en los seis pasos de esta técnica ADN recombinante (Figura 2).

2.2.1 Selección de los genes parentales. Los modelos de barajado de ADN de selección de genes parentales (He et al., 2012; Moore & Maranas, 2002) fueron trabajos posteriores a los modelos *in silico* que intentaron replicar los procesos del barajado de ADN, y se basan en el supuesto que los genes parentales pueden a nivel experimental ser sintetizados para mejorar las métricas obtenidas en bibliotecas quiméricas.

Los dos trabajos reportados para la selección se basan en métodos de optimización de codones de los genes parentales (He et al., 2012; Moore & Maranas, 2002). Estos métodos asumen que se puede explorar la secuencia de nucleótidos de genes parentales que resulte mejor para la generación de bibliotecas quiméricas. La búsqueda se realiza haciendo una combinación de variaciones de los nucleótidos que codifican para las proteínas parentales, mediante la inclusión de dos reglas de variación: sustituciones silenciosas (variaciones de los nucleótidos de los codones que no cambian el aminoácido) y sustituciones conservadas (variaciones de los nucleótidos de los codones que cambian a otro aminoácido que puede sustituirlo sin perjuicio funcional, por ejemplo una Arginina por una Lisina) (He et al., 2012; Moore & Maranas, 2002). Estas reglas de variación son usadas para evaluar mediante métodos de optimización cuatro funciones objetivo: i) La orientación de los nucleótidos comunes (Ecuación 4), ii) aproximación del vecino más cercano al cambio de la energía libre de recocido (Ecuación 5), iii) corridas de nucleótidos comunes (Ecuaciones 6 y 7), iv) Diversidad de la biblioteca (Ecuación 8).

- Optimización de los nucleótidos comunes:

El objetivo es maximizar el número de nucleótidos idénticos (Ecuación 4) en las posiciones comunes (He et al., 2012; Moore & Maranas, 2002).

$$O_{nt} = \sum_{i=1}^{3n} I\{d_1[i] = d_2[i]\} \quad (4)$$

Donde *I* es una función indicador (1 si es verdadera, 0 si es falsa).

- Optimización del ΔG:

El objetivo es minimizar el cambio de energía libre, donde un enfoque común es aproximar el cálculo de la energía libre con la suma de las contribuciones de los pares de dinucleótidos, denominada la aproximación del vecino más cercano (Ecuación 5) (He et al., 2012; Moore & Maranas, 2002).

$$O_{nn} = \sum_{i=1}^{3n-1} \Delta G°_{nn}\,(d_1[i] \cdot d_1[i+1], d_2[i] \cdot d_2[i+1]) \quad (5)$$

- Optimización de corridas:

El objetivo es optimizar las corridas, maximizando las longitudes de subcadenas apareadas en posiciones alineadas (R) (Ecuación 6) después de cada corrida. Para ello se usa la función f(r) (Ecuación 7), que permite el conteo del número total de nucleótidos por corrida, basado en un umbral θ. Esto supone que los cruces no son posibles para carreras con menos de θ nucleótidos comunes (He et al., 2012).

$$O_{run} = \sum_{R} f(|R|) \quad (6)$$

$$f(r) = \begin{cases} 0 & r < \theta \\ r & r \geq \theta \end{cases}$$

Donde (7)

- Optimización de diversidad:

El objetivo es minimizar la varianza de la diversidad (Ecuación 8) donde λ es el número de fragmentos, m (H_i, H_j) es el nivel de mutación o número de aminoácidos diferentes entre el par de quimeras H_i, H_j y $\bar{m}$ es el promedio de m en la biblioteca (He et al., 2012).

$$O_{div} = \frac{1}{2^{\lambda}(2^{\lambda}-1)} * \sum_{i=1}^{2^{\lambda}-1} \sum_{j=i+1}^{2^{\lambda}} (m\left(H_i, H_j\right) - \bar{m})^2 \quad (8)$$

Estos trabajos presentaron buenas aproximaciones para modificar genes parentales eficientes en estudios *in vitro* donde la identidad de las proteínas eran bajas (15-47%) (He et al., 2012).

2.2.2 Fragmentación de los genes. La representación de la fragmentación de los genes parentales por acción de la digestión de la DNasa I fue uno de los resultados más importantes que arrojó el primer modelo *in silico* de barajado de ADN (Sun, 1999). Este trabajo logró demostrar que el proceso se podía modelar a partir de una distribución de probabilidades tipo Poisson (Ecuación 9), donde λ=1/l, siendo l la longitud promedio de nucleótidos que se espera tengan los fragmentos de ADN digeridos.

$$f_x(x, \lambda) = \lambda e^{-\lambda x} \quad (9)$$

Este resultado fue derivado de la implementación del modelo de secuenciación de ADN (Lander-Waterman) para explicar el proceso de fragmentación y ensamblaje del barajado de ADN (Sun, 1999). Aunque es un resultado puramente estadístico, es ampliamente aceptado como insumo en los modelos *in silico* de barajado posteriores (He et al., 2012; Maheshri & Schaffer, 2003; Moore & Maranas, 2000).

2.2.3 Desnaturalización. El proceso de desnaturalización, que simula la ruptura de los puentes de hidrogeno, que unen la doble cadena de ADN, para generar cadenas de ADN sencillas, es un tema de mera implementación computacional, en donde las secuencias de ADN en dirección 5'- 3' y su complementaria 3'- 5' son almacenadas de forma independiente.

2.2.4 Hibridación. Los trabajos realizados para simular el proceso de hibridación se basan en el supuesto que los fragmentos están compitiendo entre sí para encontrar otro fragmento complementario, cuyo apareamiento le genere estabilidad. El método usado para definir en estos trabajos qué fragmentos hibridan entre sí, es el método de la mínima energía (Maheshri & Schaffer, 2003; Moore & Maranas, 2000).

El método de mínima energía asume que la energía requerida para la formación de una cadena doble de ADN se puede calcular a partir de la suma de las energías libres de los dinucleótidos que conforman la secuencia de ADN. Para ello se aplica el modelo del vecino más cercano, que supone la estabilidad de un par de bases, gobernada por la identidad y orientación de sus pares de bases vecinos (SantaLucia, 1998).

Para calcular la energía libre de una hebra doble de ADN, por ejemplo: CGTTGA·TCAACG (Figura. 3), se debe calcular la energía de los dinucleótidos de forma independiente (Ecuación 10, 11).

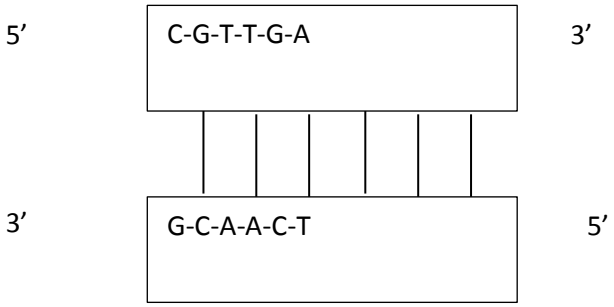


Figura 3. Representación de un fragmento de ADN

$$\Delta G^{\circ}_{37} = \Delta G^{\circ}\,(\mathrm{CG/GC}) + \Delta G^{\circ}\,(\mathrm{GT/CA}) + \Delta G^{\circ}\,(\mathrm{TT/AA}) + \Delta G^{\circ}\,(\mathrm{TG/AC}) + \Delta G^{\circ}\,(\mathrm{GA/CT}) + \Delta G^{\circ}\,(\mathrm{inicio}) \quad (10)$$

$$\Delta G^{\circ}_{37} = -2{,}17 - 1.44 - 1.00 - 1.45 - 1.30 + 0{,}98 + 1.03 = -5.35\ kcal/mol \quad (11)$$

Los cálculos de la energía libre (ΔG) de los dinucleótidos están dados por la temperatura del sistema en el cual estén inmersos y por los valores de entalpia (ΔH) y entropía (ΔS), de cada dinucleótido (Ecuación 12).

$$\Delta G^{\circ}_{T} = \Delta H^{\circ} - T\Delta S^{\circ} \quad (12)$$

Los valores de entalpia y entropía de los diferentes dinucleótidos de ADN fueron encontrados empíricamente y consolidados por SantaLucia (SantaLucia, 1998; SantaLucia & Hicks, 2004).

El método de energía libre es usado para representar el proceso de hibridación al menos desde dos posturas diferentes (Maheshri & Schaffer, 2003; Moore & Maranas, 2000): i) para establecer una fórmula de selectividad de fragmentos que pueden hibridar (Moore & Maranas, 2000); ii) para evaluar si un fragmento cambia de estado al colisionar con otro (Maheshri & Schaffer, 2003).

La primera postura asume que una reacción entre un fragmento F puede alinearse con un fragmento molde A y formar una hebra doble AF (Figura 4), en donde la fracción del molde (X_A), del fragmento (X_F) y de la hebra doble de ADN (X_{AF}) a diferentes temperaturas se puede expresar como una constante de equilibrio K(T) (Ecuación 13) (Moore & Maranas, 2000).

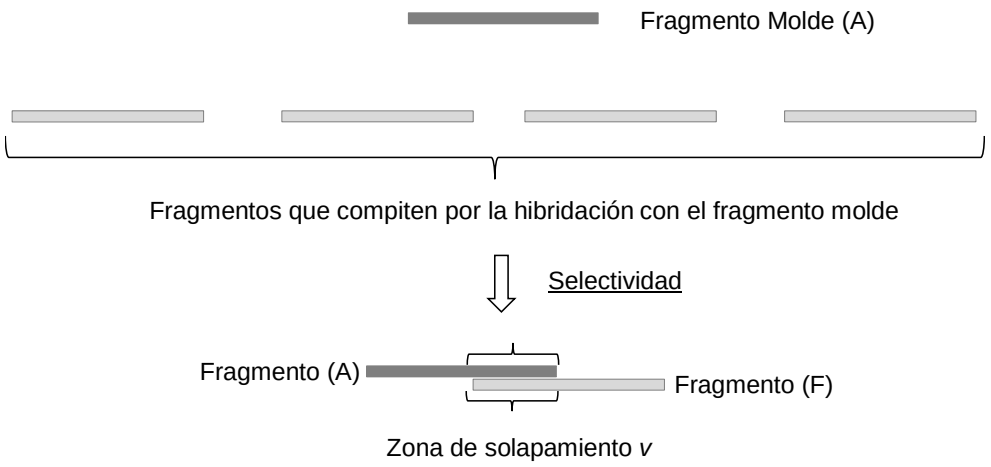


Figura 4. Selectividad de fragmentos

Dado que la competencia es entre varios fragmentos F por hibridar con una zona de solapamiento (v), la fracción para un fragmento particular se puede estimar mediante una ecuación de selectividad (Ecuación 14), donde el fragmento F corresponde al gen parental m, que

se solapa con el fragmento molde A mediante la unión de (v) nucleótidos (Moore & Maranas, 2000).

$$K(T) = \exp\left(-\frac{\Delta G(T)}{RT}\right) = X_{AF}/X_A X_F \tag{13}$$

$$S_{mv}(T) = X_{AFmv} / \left(\sum_{m',v'} X_{Fm'v'}\right) \tag{14}$$

En forma práctica esta postura se traduce en que la X_{AF} correspondiente a la mayor energía libre será el fragmento elegido para hibridar.

Por otra parte una segunda postura propone un modelo de hibridación de cambio de estados, en el cual dos fragmentos seleccionados al azar colisionan para generar un variado número de posibilidades de caminos de hibridación (Figura 5) (Maheshri & Schaffer, 2003).

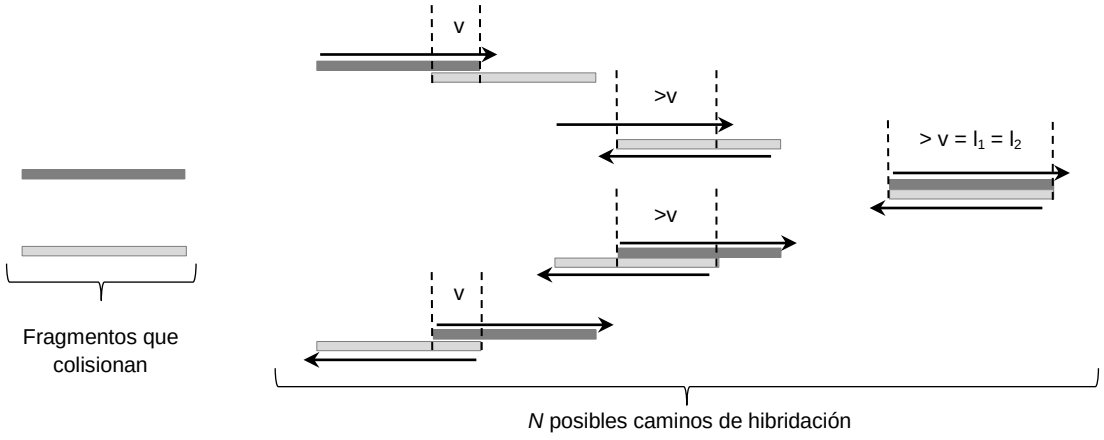


Figura 5. Cambio de estados

La cantidad de posibles caminos de hibridación generados por la colisión de dos fragmentos de ADN al azar, de longitudes l_1 y l_2, se puede evaluar contemplando el recorrido de ambos fragmentos solapados en un mínimo de *v* nucleótidos, desplazándose nucleótido a nucleótido en sentidos opuestos (Figura 5), es decir, una colisión de dos fragmentos con longitudes en su secuencias $l_1 = l_2$ de 50 nucleótidos cada una, con un solapamiento entre ellas mínimo de *v*=7 nucleótidos, genera 50+50-14+1 = 87 posibles caminos de hibridación (Ecuación 15) (Maheshri & Schaffer, 2003).

$$N = l_1 + l_2 - 2v + 1 \quad (15)$$

Para evaluar cuál de esos caminos es el más estable, la constante de equilibrio K(T) (Ecuación 13) se expresa en términos de una variable α que depende de un valor de concentración total de fragmentos de ADN y un término *b* que actúa como una penalización de complementariedad de los caminos, siendo $b = 1$ cuando existe complementariedad entre los nucleótidos solapados y $b = 4$ cuando no existe complementariedad (Ecuación 16) (Maheshri & Schaffer, 2003).

$$\alpha = \frac{1}{KC_T b} + 1 \quad (16)$$

$$X = \alpha - \sqrt{\alpha^2 - 1} \quad (17)$$

La variable α se usa para expresar una conversión de equilibrio X que cuantifica en términos de probabilidad la viabilidad de selección de cada uno de los caminos. El camino con mayor probabilidad será el camino seleccionado. La probabilidad de alineamiento X decrece de 1 hasta 0 en la misma proporción en que α crece desde ≈ 1 hasta ∞. También existe la probabilidad que los fragmentos no cambien de estado y la colisión no resulte en hibridación, esto es ajustado por un modelo markoviano de dos estados (Maheshri & Schaffer, 2003).

2.2.5 Extensión. El proceso de extensión es simulado en los diferentes trabajos como la acción de una polimerasa con el 100% de fidelidad (Maheshri & Schaffer, 2003). La extensión se simula en dirección 5’ – 3’ del fragmento molde (Maheshri & Schaffer, 2003; Moore & Maranas, 2000), es decir, si dos fragmentos con longitudes K y L hibridan, solapándose en v nucleótidos complementarios, el nuevo fragmento de ADN tendrá una nueva longitud de K + (L-v). Se genera en el momento de ensamblado un punto de cruce, después para colisionar en un siguiente ciclo de iteración (Figura 6) (Moore & Maranas, 2000).

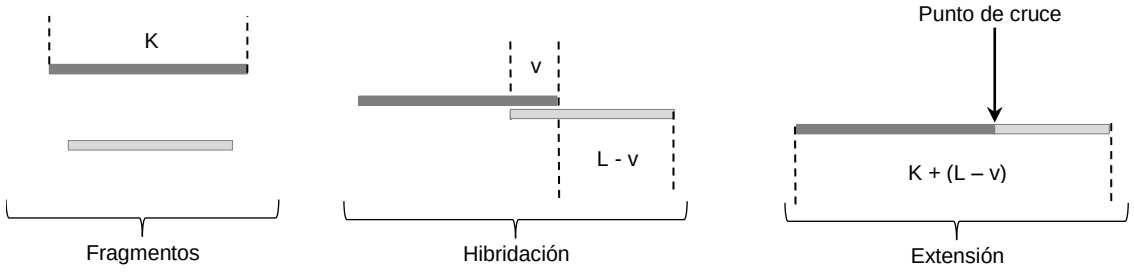


Figura 6. Proceso de Extensión

2.2.6 Biblioteca Recombinada. Hasta el momento, el autor no tiene conocimiento de algún trabajo orientado hacia la generación de bibliotecas quiméricas *in silico*. Todos los trabajos reportados se limitan al cálculo de métricas o generalizar hallazgos computacionales para ser contemplados en las condiciones experimentales *in vitro* del barajado de ADN, de tal forma que se mejore la eficiencia en el ensamblado, y por consiguiente el potencial de las bibliotecas quiméricas generadas (Joern et al., 2002; Maheshri & Schaffer, 2003; Moore & Maranas, 2000).

Los trabajos reportados han logrado establecer relaciones entre los parámetros experimentales del barajado de ADN (temperatura, concentración de ADN y longitud de los fragmentos), con la eficiencia del ensamblado. Dentro de los hallazgos más significativos están que: i) Las condiciones de baja temperatura sesgan los cruces hacia la región de mayor identidad de la secuencia (Joern et al., 2002; Maheshri & Schaffer, 2003); ii) Las condiciones de concentración de ADN bajas permiten acceso a eventos de cruce, mientras que concentraciones altas de ADN promueven la formación de secuencias basura; iii) Los fragmentos de ADN con longitudes pequeñas promueven la formación de heterodúplex (alineamiento de fragmentos de diferentes padres), mientras que fragmentos con longitudes mayores promueven la formación de homoduplex (alineamiento de fragmentos del mismo padre). La formación de heterodúplex genera diversidad, pero de forma desmedida, y puede hacer que las variantes obtenidas pierdan funcionalidad (Moore & Maranas, 2000).

Algunas métricas *in silico* que se pueden obtener de experimentos de barajado de ADN son: i) la fracción de la biblioteca generada que se espera tenga 1, 2,…, 10 puntos de cruce; ii) promedio de entrecruzamiento por secuencia; iii) Probabilidad por cada nucleótido del gen

parental para ser un punto de cruce (Moore & Maranas, 2000). Aunque estas métricas de eficiencia de ensamblado son de interés en la formación de bibliotecas quiméricas, son difíciles de implementar en experimentos *in vitro* de barajado de ADN, por tal razón resultan de interés alternativas de evaluación de una biblioteca desde la diversidad obtenida, (Patrick & Firth, 2005; Patrick et al., 2003).

La medida de diversidad resulta relevante si se piensa que al maximizar la diversidad de una biblioteca se aumenta la probabilidad de encontrar variantes con propiedades mejoradas (Patrick & Firth, 2005; Patrick et al., 2003). La evaluación de la diversidad se puede hallar desconociendo la información completa de las secuencias parentales, basta con conocer la distancia entre las diferencias consecutivas entre las secuencias parentales. Se asume que el número de cruces (X=0,1, 2,…) que puede producirse entre dos mutaciones consecutivas sigue una distribución tipo Poisson (Ecuación 18) (Patrick & Firth, 2005; Patrick et al., 2003; Sun, 1999).

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \tag{18}$$

Donde λ es el número real de cruces en una muestra de las secuencias resultantes, este λ excluye el número de cruces silenciosos, y es una buena aproximación a la diversidad que se puede lograr en bibliotecas quiméricas (Patrick & Firth, 2005; Patrick et al., 2003).

2.3 Puntos de quiebre

En los diferentes métodos hasta ahora expuestos, hemos identificado dos puntos de quiebre. El primero desde la intencionalidad de los métodos y otro desde el grado de representación de los mismos en el paso de hibridación del barajado. En cuanto al punto de quiebre de la intencionalidad, observamos que el alcance de los métodos ha estado en tres tareas muy específicas: i) Entender las relaciones que pueden existir entre parámetros experimentales y eficiencia de los ensamblados, resultados útiles para refinar condiciones experimentales; ii) Diseñar genes parentales que permitan optimizar la eficiencia de los ensamblados, gobernados por los hallazgos del literal anterior; iii) Predecir la diversidad que tendría una biblioteca quimérica a partir de las diferencias de identidad de dos genes parentales. Aunque estas representaciones y sus hallazgos son útiles, pensamos que pueden ser llevados a cabo nuevos desarrollos que generen bibliotecas quiméricas *in silico* y del mismo modo como se pueden sintetizar los genes parentales, las mejores variantes *in silico* pueden ser sintetizadas. Aparte de nuestro trabajo (Capitulo 3) que permite la generación de bibliotecas quiméricas *in silico*, no tenemos conocimiento de un trabajo orientado hacia dicho propósito.

En cuanto al nivel de representación del paso de hibridación en los modelos *in silico*, hemos observado que el fenómeno biológico de las estructuras secundarias de ADN no es tenido en cuenta en ninguno de los modelos *in silico* hasta ahora expuestos. Las estructuras secundarias de ADN tienen un papel importante a nivel biológico, interviniendo en procesos como la replicación, transcripción, recombinación y reparación del ADN (Bikard, Loot, Baharoglu, & Mazel, 2010; Muhire et al., 2014; Sander et al., 2014) y tienen lugar cuando hebras sencillas

ADN se auto-pliegan por complementariedad Watson-Crick (Figura 7) (SantaLucia & Hicks, 2004).

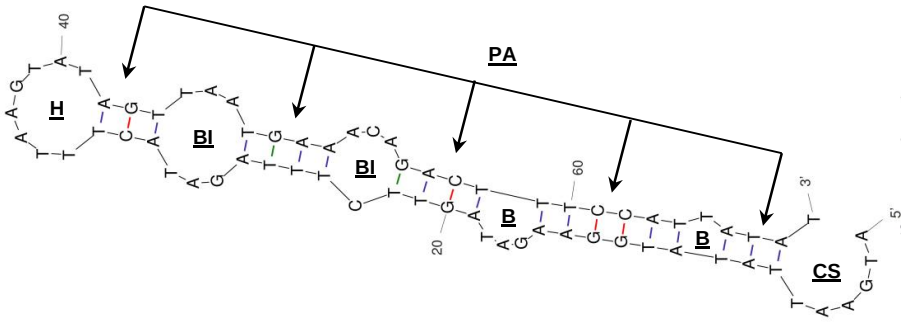


Figura 7. Motivos estructurales de un fragmento de la región I del gen cry11A.a. De izquierda a derecha: Horquilla (H), pares apilados (PA), bucle interno (BI), pares apilados (PA), bucle interno (BI), pares apilados (PA), bulto (B), pares apilados (PA), bulto (B), pares apilados (PA), ADN cadena sencilla (CS).

En técnicas moleculares, la formación de estructuras secundarias de ADN tiene efectos en la hibridación produciendo falsos positivos y reacciones cruzadas (Koehler & Peyret, 2005; Lu, Guo, Marky, Seeman, & Kallenbach, 1992; Nazarenko, Pires, Lowe, Obaidy, & Rashtchian, 2002) y en técnicas como el barajado de ADN (Stemmer, 1994), donde se realizan ciclos sucesivos de amplificación de ADN sin y con cebadores se pueden generar estructuras secundarias que alteran la variabilidad genética durante la recombinación (Figura 8).

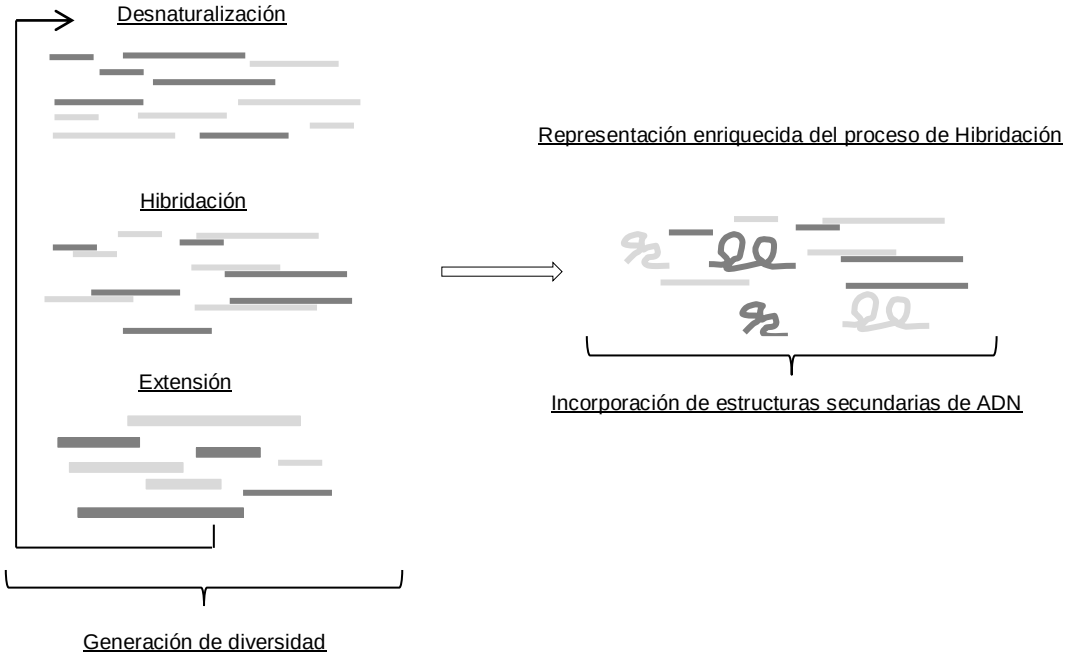


Figura 8. Hibridación (Participación de las estructuras secundarias de ADN)

Las estructuras secundarias de ADN y sus efectos en la generación de diversidad no son considerados en los modelos *in silico* hasta ahora expuestos. En el presente trabajo se maneja la hipótesis que la inclusión de los efectos de la formación de las estructuras secundarias en modelos de barajado de ADN es necesaria para generar bibliotecas quiméricas *in silico* (Capitulo 3).

2.4 Conclusiones

La generación de diversidad es el punto crítico en ED y de las técnicas existentes generadoras de diversidad génica (mutagénesis aleatoria y recombinación). Las técnicas ADN recombinantes han demostrado ser más potentes para explorar el amplio espacio de búsqueda de variantes

funcionales. La técnica ADN recombinante más usada y aceptada en ED es la técnica de barajado de ADN, que permite la obtención de bibliotecas quiméricas a partir de la recombinación de genes homólogos, mediante el uso de ciclos iterativos de PCR sin cebadores.

El desarrollo de modelos *in silico* de barajado de ADN se dio tempranamente, poco después de la aparición de la técnica. Los primeros trabajos en esta materia presentaron hallazgos que relacionaban los parámetros experimentales con la eficiencia de ensamblado, promoviendo esfuerzos posteriores en la caracterización de los genes parentales como un insumo determinante en la optimización de métricas eficientes. La revisión crítica de estos hallazgos perfiló el concepto emergente de diversidad como una nueva métrica para evaluar bibliotecas quiméricas obtenidas. Los modelos *in silico* de barajado de ADN reportados han hecho uso de métodos tan variados como: métodos de mínima energía, estadísticos y de optimización. La gran mayoría de los trabajos usaron estos métodos de forma hibrida, logrando una mayor aproximación al proceso simulado.

Pese a la importancia de los hallazgos *in silico* sobre la técnica de barajado de ADN, se puede apreciar una deficiencia generalizada en los trabajos reportados, lo cual constituye una oportunidad: Un desconocimiento del fenómeno biológico de la formación de estructuras secundarias de ADN y su efecto en la generación de bibliotecas quiméricas. Dado que la formación de estructuras secundarias, o auto-plegamiento del ADN, se presenta en estado de hebra sencilla, resulta un evento altamente probable en experimentos de barajado, pues los pequeños fragmentos desnaturalizados pueden producirlos, limitando la alineación de los fragmentos para hibridar, con un efecto importante sobre la eficiencia de ensamblado y la

diversidad de las bibliotecas quiméricas. Por otra parte se cuenta con una oportunidad de investigación al ampliar el alcance limitado del uso de los métodos, pues estos estaban centrados en la búsqueda de métricas experimentales, cuando es posible incursionar computacionalmente en la obtención de bibliotecas quiméricas *in silico*.

Este trabajo es el primero en desarrollar un modelo integrado de ED *in silico*, que incorpora de forma novedosa los efectos de la formación de estructuras secundarias de ADN en el proceso de ED y que orienta su modelo generador de diversidad a la obtención de bibliotecas quiméricas *in silico*. En el próximo capítulo se detallará el modelo *in silico* de ED implementado (Capitulo 3).

3. DEVISING: Una estrategia in silico de ED

En este capítulo se presentará una estrategia *in silico* de ED novedosa, denominada DEvISING (**D**irected **Ev**olution ***I**n **S**ilico* Model**ing**). La estrategia integra un conjunto de herramientas software (SANAFold, SAssembly, GenE-in) desarrolladas en el lenguaje de programación Python 3.2, bajo el paradigma de programación orientado a objetos (Figura 9).

DEVISING es el primer desarrollo, según tenemos conocimiento, que permite la obtención de variantes génicas *in silico* para sintetizar y transformar *in vivo* y que contempla los efectos de la formación de estructuras secundarias de ADN en las variantes obtenidas. Esta estrategia combina métodos de mínima energía, métodos estadísticos, procesos estocásticos y técnicas tradicionales bioinformáticas como: El cálculo del porcentaje de identidad entre secuencias, el alineamiento múltiple de secuencias y la traducción de secuencias de ORF (Figura 9).

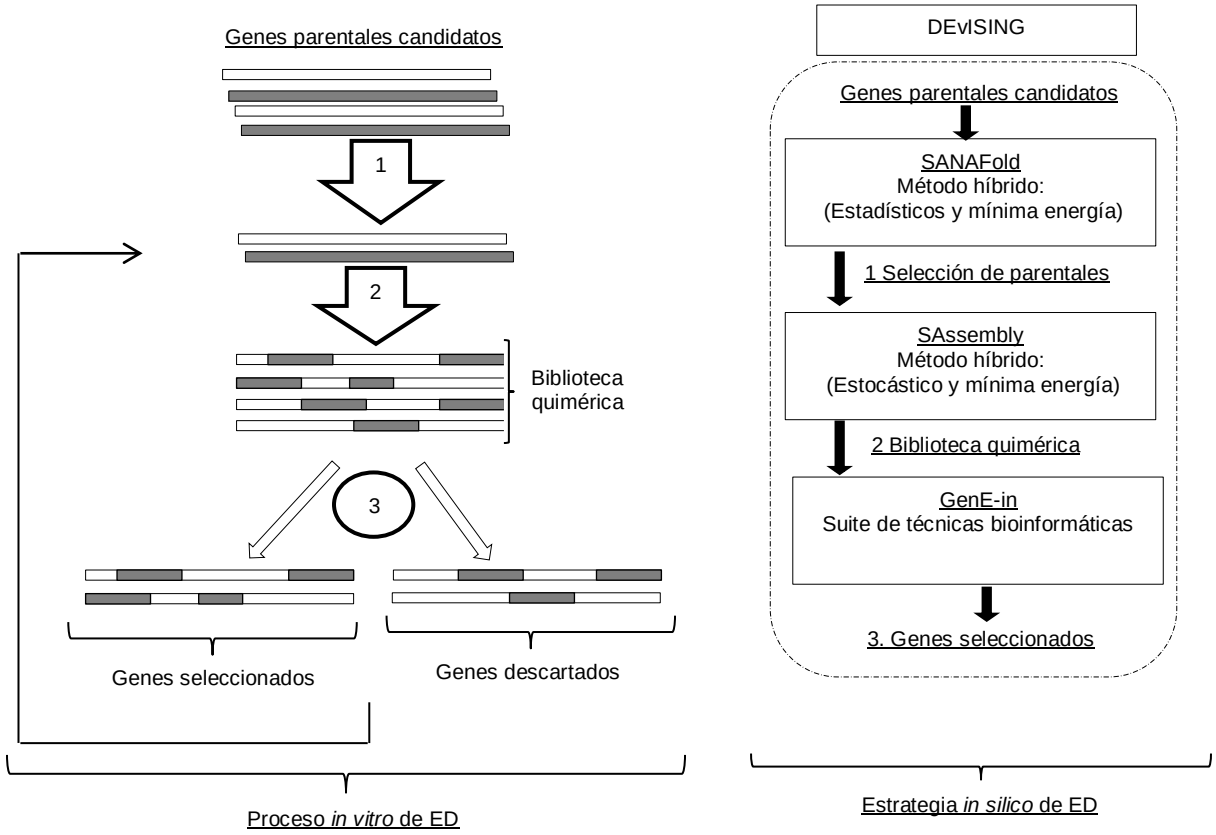


Figura 9. Comparación entre la Arquitectura de DEVISING y los pasos de ED. 1) Selección de genes parentales; 2) Generación de diversidad; 3) Tamizaje/ Selección.

A continuación se realizará una descripción de cada uno de los componentes software (SANAFold, SAssembly, GenE-in) que conforman la estrategia DEVISING.

3.1 SANAFold

SANAFold (***S**tatistical **A**nalysis of **N**ucleic **A**cid **Fold**ing*) es el primer componente software de DEvISING y fue diseñado para simular el paso de selección de genes parentales de ED *(Anexo 1)*. Como se argumentó en el capítulo 2, la caracterización de los genes parentales representa un

paso crítico en ED. Una correcta selección de genes parentales puede mejorar la eficiencia del ensamblado y la diversidad de las bibliotecas quiméricas (He et al., 2012; Moore & Maranas, 2002; Patrick & Firth, 2005; Patrick et al., 2003).

Dado que la caracterización de genes parentales es de alta relevancia para obtener bibliotecas quiméricas mejoradas y es fácilmente reproducible *in vitro*, se propone con SANAFold una forma novedosa de realizar la caracterización de los genes parentales *in silico*, mediante el análisis del comportamiento físico-químico de los genes candidatos, cuando forman estructuras secundarias de ADN en condiciones experimentales de barajado. Esta caracterización permite superar las deficiencias de los modelos *in silico* de generación de diversidad que fueron revisados en el capítulo 2 y traza la ruta para obtener mediante la estrategia integrada DEvISING bibliotecas quiméricas *in silico*.

3.1.1 Uso de métodos predictivos de estructuras secundarias de ADN. Los métodos computacionales de predicción de estructuras secundarias de ADN se pueden agrupar en dos vertientes: métodos de energía libre y métodos comparativos (Spirollari, 2010; Xiong, 2006). Los métodos de energía libre se basan en el supuesto que la estructura secundaria de una hebra sencilla de ADN se determina únicamente a partir de la secuencia, al conseguir una estructura con el número máximo de pares de bases se asume una disminución de la energía requerida y una mayor estabilidad de la estructura (Xiong, 2006). Por su parte, los métodos comparativos se basan en el supuesto que una estructura secundaria se puede predecir de la comparación de las secuencias de genes homólogos, debido a que las estructuras funcionales se han conservado

evolutivamente. En consecuencia, los datos de entrada de estos algoritmos asumen dos o más secuencias de ADN como punto de partida para los cálculos (Xiong, 2006).

Dado que SANAFold está orientada hacia la caracterización físico-química de genes parentales en condiciones de barajado de ADN, el método de mínima energía resulta ser el más adecuado para el proceso de implementación de la predicción de estructuras secundarias de ADN. Este método permite introducir como parámetros para el cálculo termodinámico: concentraciones iónicas de Mg^{++}, la temperatura del medio y la secuencia de nucleótidos (Xiong, 2006); los mismos parámetros que se contemplan en experimentos de barajado de ADN, *in vitro* (Stemmer, 1994) y en modelos *in silico* (He et al., 2012; Joern et al., 2002; Maheshri & Schaffer, 2003; Moore & Maranas, 2000, 2002).

- Método de mínima energía (Estructuras secundarias de ADN)

El método de mínima energía asume que la energía requerida para la formación de una estructura secundaria de ADN se puede calcular a partir de la suma de las energías libres de cada uno de sus motivos estructurales (Tinoco, Uhlenbeck, & Levine, 1971) (Ecuación 19). El método asume como principio que los diversos motivos estructurales tales como: horquillas, pares apilados, bucles internos, bucles multiples y bultos (Figura 7) son facilmente identificables. Esto hace del cálculo de la energía de toda la estructura un ejercicio de descomposición de la misma, en función de sus motivos estructurales, que contribuyen de forma aditiva a la energía total del sistema (Tinoco et al., 1971).

$$\Delta G^{\circ}(s) = \Delta G^{\circ}(s_1) + \Delta G^{\circ}(s_2) + \cdots + \Delta G^{\circ}(s_t) \quad (19)$$

Para calcular la energía libre de cada motivo estructural $\Delta G°(S_x)$, siendo x desde 1 hasta t, se suman las energías libres de los dinucleótidos que conforman un motivo estructural, mediante el modelo del vecino más cercano, que fue explicado en la sección 2.2.4. La principal variación en el cálculo de la energía para un motivo estructural es que cada motivo estructural de ADN está gobernado por una tabla de energía particular (SantaLucia & Hicks, 2004).

Actualmente existen herramientas computacionales que implementan el método de mínima energía para la predicción de estructuras secundarias de ARN-ADN, disponibles en plataformas web-server. Dentro de las ampliamente aceptadas y usadas por la comunidad académica se encuentran: RNAFold (Hofacker, 2003) y UNAFold (Markham & Zuker, 2008; Zuker, 2003), siendo esta última la más renombrada (http://unafold.rna.albany.edu/?q=mfold).

- UNAFold 3.8

UNAFold 3.8 (*Unified Nucleic Acid Folding*) es un paquete integrado de herramientas computacionales escrito en Perl, dispuesto en la web para predecir estructuras secundarias ARN-ADN (Markham & Zuker, 2008). UNAFold 3.8 integra una paquetería y un entorno web con los desarrollos realizados previamente con el software Mfold, que realiza la predicción de estructuras secundarias de ARN y ADN (Zuker, 2003).

UNAFold 3.8 no solo ha sido ampliamente usado para estudios de predicción, sino como un módulo para el desarrollo de software de análisis dada su disponibilidad de descarga y uso, tal

es el caso de NASP (*Nucleic Acid Secondary Structure Predictor*) (Semegni et al., 2011), software de análisis que aprovecha el método de mínima energía de UNAFold para construir un umbral termodinámico de referencia a partir de una matriz de consenso. Esta matriz es modificada entre columnas de forma aleatoria para garantizar rigurosidad estadística en caso de hallar regiones génicas conservadas (Semegni et al., 2011). Estas herramientas han demostrado ser útiles en la predicción de estructuras secundarias y recientemente se han utilizado para el diseño de código de barras de ADN en plantas basados en transcriptos de ARN ITS/ITS2 (Zhang, Yuan, Yang, Huang, & Huang, 2015) y para conocer el papel de estas estructuras y su relevancia biológica en algunos virus (Muhire et al., 2014).

UNAFold 3.8 cuenta con capacidad de predicción de estructuras secundarias a partir del método de mínima energía, lo que permite representar condiciones *in vitro* de barajado de ADN y su característica de reusabilidad; también existe disponibilidad de sus códigos fuentes a la comunidad académica. Dadas las anteriores características, UNAFold 3.8 resultó la mejor alternativa predictiva de estructuras secundarias para estudiar el comportamiento físico-químico de genes parentales en condiciones de barajado de ADN.

3.1.2 Arquitectura software de SANAFold. SANAFold soporta una estrategia estadística diseñada para caracterizar genes parentales candidatos a ser barajados (Figura 10).

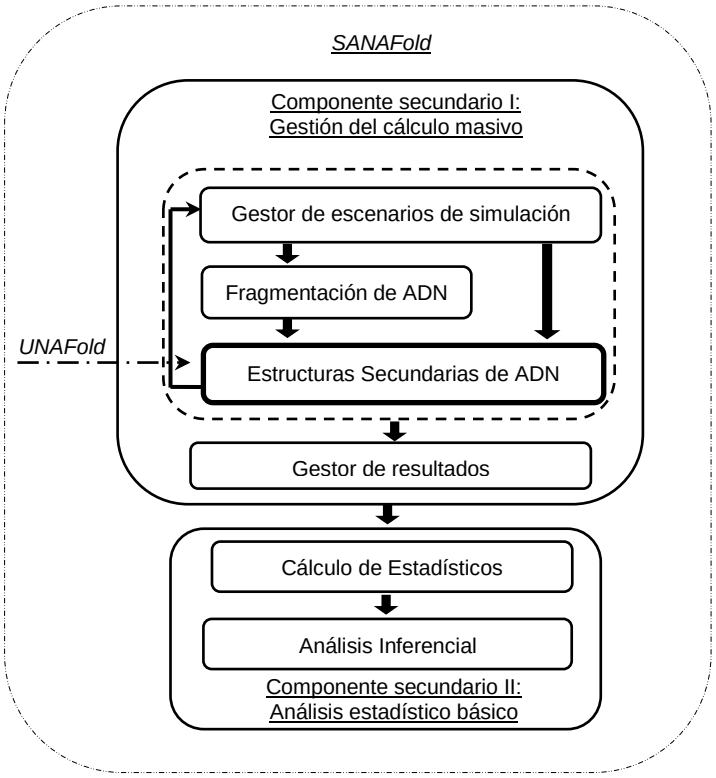


Figura 10. Arquitectura software de SANAFold

Para ello, se han desarrollado dos subcomponentes software: i) El subcomponente de gestión de cálculo masivo: es el encargado de preparar los escenarios de simulación, en condiciones experimentales de barajado de ADN. De esta forma, gestiona los parámetros de ingreso de simulación, a su vez que hace llamados masivos a los módulos de UNAFold. Los resultados de estas simulaciones son devueltos en múltiples archivos por cada llamado, por lo tanto este subcomponente también se encarga de seleccionar y almacenar la información útil para los análisis estadísticos posteriores (Figura 10). ii) El subcomponente de análisis estadístico básico: encargado de la indagación sobre los datos almacenados. Para ello, se realiza una

clasificación de los datos desde su origen génico, a partir de los cuales se indagan comportamientos diferenciados entre las regiones génicas de un gen, mediante la creación de indicadores estadísticos, análisis de varianza (ANOVA) y análisis de distribución de datos. Estos resultados permiten inferir la preferencia de recombinación más probable en las bibliotecas quiméricas construidas con un gen parental ya caracterizado (Figura 10).

- Módulo de gestión de escenarios de simulación

El módulo de gestión de escenarios de simulación almacena los parámetros de simulación y los suministra al módulo de fragmentación y a los módulos de UNAFold para las predicciones de estructuras secundarias de ADN. Los parámetros son ingresados conformando escenarios de simulación basados en tres parámetros de interés en barajado: temperatura, concentración iónica de Mg^{++} y longitud promedio de fragmentación. Estos escenarios se forman combinando rangos de dos parámetros manteniendo el tercer parámetro constante, permitiendo tres escenarios de simulación: Escenario Te-ma, donde la longitud de fragmentación es constante; escenario Le-ma, donde la temperatura es constante y escenario Le-te, donde la concentración iónica de Mg^{++} permanece constante (Figura 11).

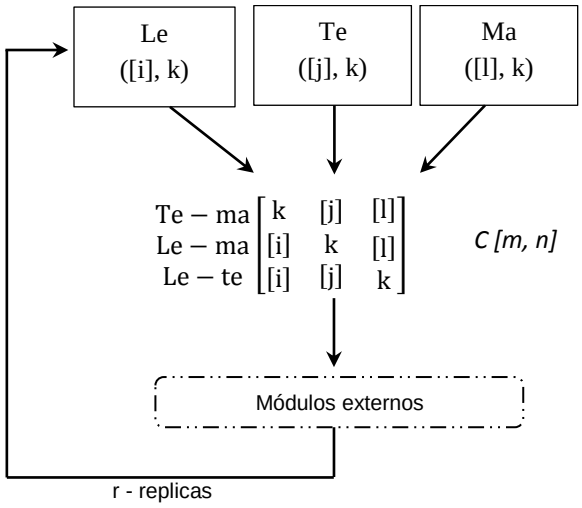


Figura 11. Esquema del proceso algorítmico para parametrización de condiciones in vitro

Por otra parte, el módulo de gestión de escenarios de simulación suministra un cuarto parámetro *r* al subcomponente de gestión del cálculo masivo, que corresponde al número de iteraciones que se van a realizar en el subcomponente. Esto implica que se replican *r* veces las simulaciones usando los mismos escenarios asegurando significancia estadística, dado que en cada replica el módulo de fragmentación se comporta como una distribución Poisson, generando nuevos fragmentos en cada iteración, usando un mismo parámetro de longitud.

Una generalización del algoritmo utilizado por el módulo de gestión de escenarios de simulación es (Figura 11):

- ✓ Leer las condiciones experimentales de Le (longitud de fragmentación), Te (Temperatura), Ma (Concentraciones iónicas de Mg^{++}) y *r* (número de réplicas).
- ✓ Realizar la combinatoria de condiciones experimentales: Te-ma (Le=k, Te[j], Ma[l]), Le-ma (Le[i], Te=k, Ma[l]), Le-te (Le=[i], Te[j], Ma=k), donde *k* implica un valor constante en ese

escenario y donde [i], [j] y [l] son arreglos unidimensionales con un rango de valores de los parámetros.

- ✓ Suministrar los valores de Le al módulo de fragmentación.
- ✓ Suministrar los valores de Te, Ma a los módulos de UNAFold.
- ✓ Repetir *r* veces los pasos anteriores, después de activar el módulo de predicción de estructuras secundarias.

- Módulo de Fragmentación

El módulo de fragmentación se encarga de simular la acción de la digestión de la DNasa I sobre las cadenas de ADN. Para ello, el módulo utiliza el modelo propuesto en los trabajos de Sun, donde los cortes obedecen a una distribución tipo Poisson (Sección 2.2.2.) (Sun, 1999).

Una generalización del algoritmo utilizado por el módulo de fragmentación es (Figura 12):

- ✓ Leer la secuencia de nucleótidos de la región génica de interés en un archivo FASTA
- ✓ Leer la longitud promedio de fragmentación (Dada por el módulo de gestión de condiciones de barajado)
- ✓ Generar las distancias de corte aleatorias a partir de la implementación de una distribución tipo Poisson
- ✓ Almacenar los fragmentos de secuencia en un archivo unidimensional F[i]

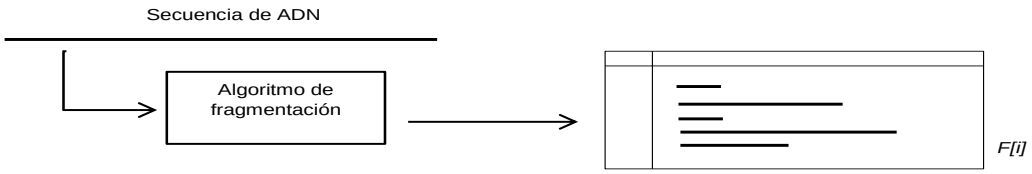


Figura 12. Esquema del proceso algorítmico de la fragmentación del ADN

- Módulo de predicción de estructuras secundarias de ADN

El módulo de predicción de estructuras secundarias de ADN está constituido por una adaptación propia de los sub-módulos de predicción de UNAFold 3.8, disponibles en código libre para uso de la comunidad académica. UNAFold realiza la predicción de estructuras secundarias haciendo uso del método de mínima energía (Sección 3.1.1.1.) y de técnicas computacionales de programación dinámica que permiten encontrar caminos con motivos estructurales termodinámicamente estables (Markham & Zuker, 2008; Zuker, 2003).

Una generalización del algoritmo utilizado por UNAFold 3.8, para realizar la predicción de estructuras secundarias, es (Figura 13):

- ✓ Realizar la construcción de una matriz de pares de bases a partir de la secuencia de nucleótidos.
- ✓ Introducir en la matriz el cálculo unificado de la energía libre del vecino más próximo contemplando las concentraciones iónicas y la temperatura del medio (SantaLucia, 1998; SantaLucia & Hicks, 2004).

- ✓ Buscar la ruta de pares de bases más estable o de mínima energía mediante programación dinámica (Markham & Zuker, 2008; Zuker, 2003). De esta forma UNAFold entrega unos archivos que contienen la información de las estructuras secundarias más probables estructuralmente y más estables termodinámicamente.

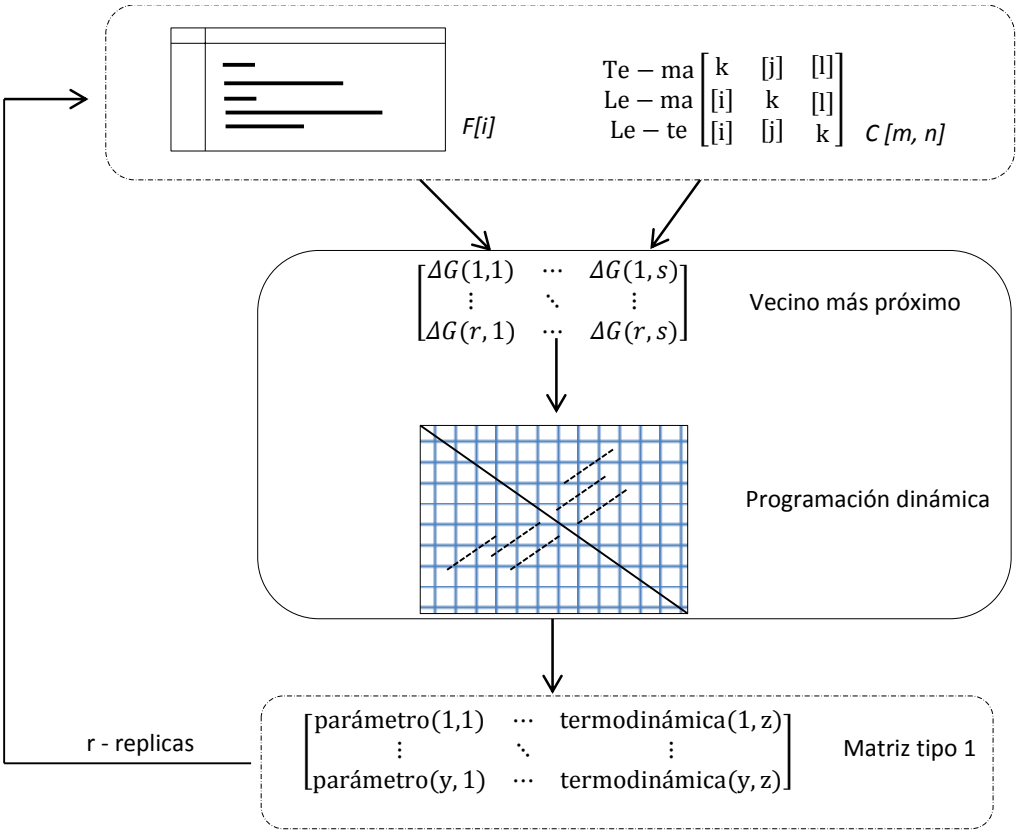


Figura 13. Esquema del proceso algorítmico de mínima energía UNAFold 3.8

- Módulo gestor de resultados

Este módulo es el último del subcomponente de cálculo masivo. Se encarga de ordenar y almacenar, por cada una de las estructuras secundarias de ADN predichas, los parámetros de simulación utilizados y las características termodinámicas de las estructuras, tales como: ΔG

(Energía libre), ΔH (Entalpía), ΔS (Entropía) y T_m (Temperatura de fusión). Este gran volumen de datos es gestionado bajo criterios de almacenamiento como: origen génico del fragmento utilizado y escenario de simulación (Le-ma, Te-ma, Le-te). La forma de almacenamiento es mediante arreglos bidimensionales .csv, denominados matrices tipo 1 (Figura 13), donde cada matriz alberga los datos de *r* – replicas simuladas.

- Módulo cálculo de estadísticos

Los datos de simulación suministrados por el módulo gestor de resultados, en forma de matrices tipo 1, son recorridos por el módulo de cálculo de estadísticos que consolida, a partir de tres variables de interés, seis estimadores estadísticos para cada matriz tipo 1 recorrida. Los estimadores establecidos fueron el promedio $\overline{\Delta G}$ de las estructuras secundarias de ADN con $\Delta G(-)$ y $\Delta G(+)$, el número promedio de estructuras secundarias de ADN formadas a partir de los fragmentos de ADN con $\Delta G(-)$ y $\Delta G(+)$, y finalmente el porcentaje de estructuras secundarias de ADN con $\Delta G(-)$ y $\Delta G(+)$. Estos estimadores son almacenados en arreglos bidimensionales .csv, denominados matrices tipo 2 (Figura 14).

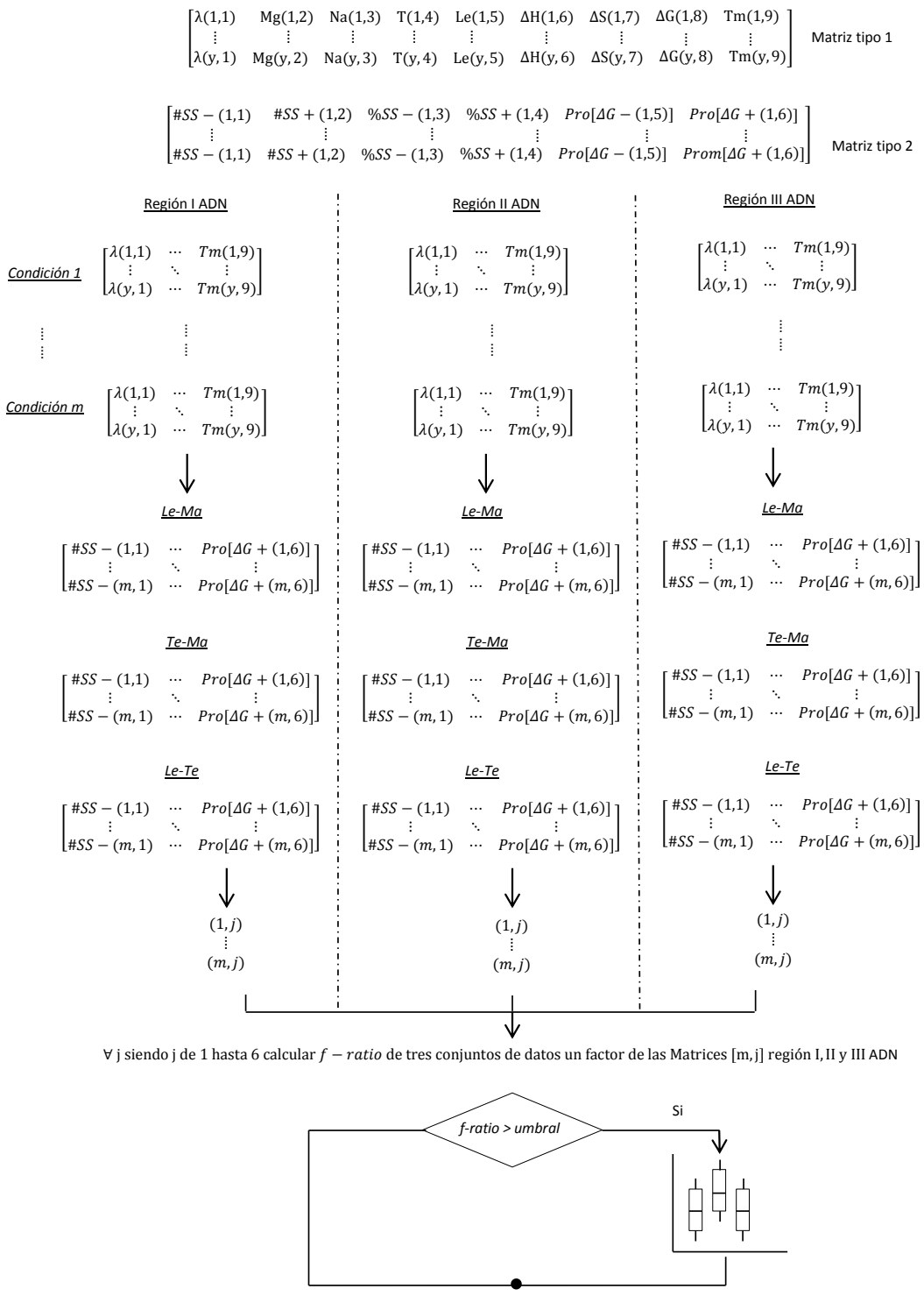


Figura 14. Esquema del proceso algorítmico de cálculos estadísticos y análisis inferencial

- Módulo de Análisis Inferencial

El módulo de análisis inferencial es el encargado de encontrar diferencias estadísticas significativas que permitan inferir el comportamiento diferenciado entre las regiones génicas de un gen parental. Para ello, el módulo implementa un análisis ANOVA de un factor y *n* variantes, donde el factor es el origen génico de los datos y las *n* variantes son el número de regiones génicas conservadas del gen parental. En la Figura 14 se observa el ejemplo algorítmico para un $n = 3$ regiones génicas.

El módulo recorre de forma selectiva las matrices tipo 2, y agrupa conjuntos de datos de las diferentes variantes (regiones génicas del parental) que pertenezcan a un mismo estimador y un mismo escenario de barajado de ADN, luego estos nuevos conjuntos de datos son evaluados mediante análisis de varianza (Figura 14). Para realizar el análisis de significancia de los datos, el módulo hace un test F (test de Fisher), obteniendo un valor *f-ratio* con sus respectivos grados de libertad y una confianza del 95.5%, para cada conjunto de datos.

Los valores *f-ratio* obtenidos son evaluados con respecto a la distribución de Fisher. Si el valor de *f-ratio* está por debajo del umbral establecido, no se rechaza la hipótesis nula (H_0) que asume la igualdad de medias de los datos analizados (Ecuación 20). En este caso, pese a la presencia de estructuras secundarias de ADN en las diferentes regiones génicas, se infiere que el gen se comporta de la misma forma en todas sus regiones, ante condiciones de barajado. Esto garantiza que no se presentarán sesgos o preferencias de recombinación, si el gen es considerado como un parental en experimentos de ED.

$$f - ratio < umbral \Rightarrow H_0: m_1 = m_2 = m_3 \quad (20)$$

$$f - ratio > umbral \Rightarrow H_1: \exists\, i, j; i \neq j\ ;\ m_i \neq m_j \quad (21)$$

Por otra parte, si el valor de *f-ratio* está por encima del umbral establecido, se rechaza la hipótesis nula (H_0) (Ecuación 21). En este caso, se infiere que al menos dos de las *n* regiones génicas del gen estudiado presentan comportamiento con diferencias estadísticas significativas, que ante las demás regiones presentarán sesgos o preferencias de recombinación, y que se verán reflejadas en las bibliotecas quiméricas, si ese gen es usado como un parental en experimentos de ED.

A partir de los datos con diferencias estadísticas significativas, el módulo calcula y presenta datos cuantitativos del estimador, por cada uno de las variantes analizadas, además de diagramas de caja-bigote con la distribución de los datos de cada variante (Figura 14). De esta forma se aprecia cuáles regiones génicas generan las diferencias de medias del comportamiento, ya sea porque favorecen (alta espontaneidad) o desfavorecen (baja espontaneidad) de forma diferenciada la formación de estructuras secundarias de ADN (Figura 14).

El autor propone el criterio de espontaneidad para calificar una región génica en su capacidad termodinámica para formar estructuras secundarias, cuantificable a partir de la energía libre de Gibbs. Si $\overline{\Delta G}$ de las estructuras secundarias de ADN de una región génica es más negativa que otra, se entiende que es más espontánea y tiene una mayor tendencia a favorecer la formación de estructuras secundarias de ADN, siempre y cuando existan diferencias

estadísticamente significativas. De esta forma, se pueden caracterizar las regiones de un gen con calificativos de alta, media y baja espontaneidad.

3.2 SAssembly

Sassembly (***S**tochastic **Assembly***) es el segundo componente software de DEvISING y fue diseñado para simular la generación de diversidad en ED *(Anexo 1)*. La arquitectura software del componente es de tipo modular, donde cuatro módulos integrados permiten su funcionalidad (Figura 15). SAssembly hace un uso híbrido de los métodos de mínima energía, métodos estadísticos y procesos estocásticos, para implementar la técnica recombinante de barajado de ADN.

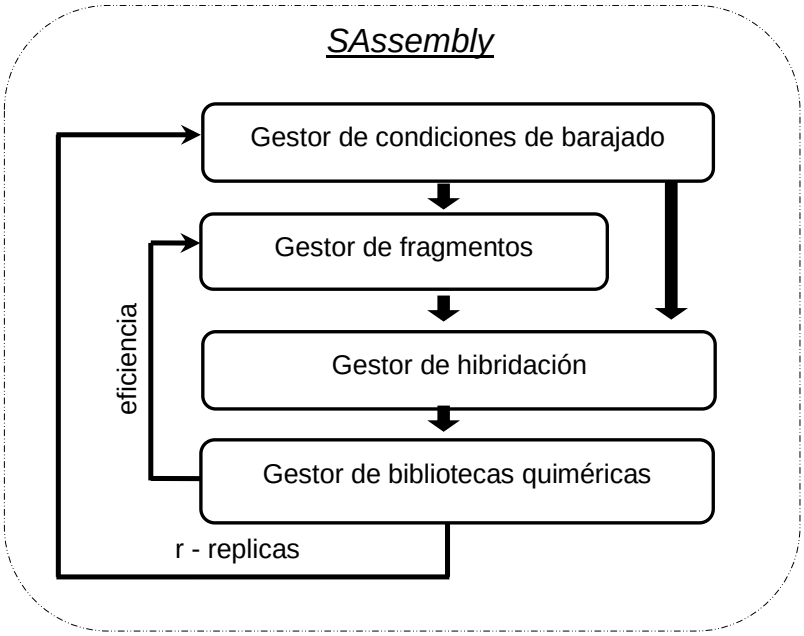


Figura 15. Arquitectura software de SAssembly.

3.2.1 Módulo gestor de condiciones de barajado. El módulo gestor de condiciones de barajado se encarga de leer y suministrar los parámetros de los diferentes escenarios de simulación. Dentro de los parámetros recibidos están: el rango de temperaturas del barajado (Te[i]), La longitud de fragmentación de los genes parentales (Le [j]), el archivo multifasta con las secuencias en dirección 5'-3' de los genes parentales y el número de r - réplicas de las simulaciones.

El archivo multifasta con las secuencias de ADN es entregado al módulo gestor de fragmentos, mientras que los demás parámetros de simulación son entregados al módulo gestor de hibridación (Figura 15). Este procedimiento de gestión de condiciones experimentales se debe replicar *r* veces, lo cual sirve para aumentar la diversidad de las bibliotecas al tomar como punto de partida en cada iteración nuevos cortes de ADN (Figura 15).

3.2.2 Módulo gestor de fragmentos. Este módulo aprovecha el código de fragmentación descrito en la sección 3.1.2.2., que simula la acción digestiva de la DNasa I, realizando cortes aleatorios según el parámetro Le[j] sobre las secuencias que lee de los archivo multifasta. Ambos datos son suministrados por el módulo gestor de condiciones de barajado (Figura 16). Por otra parte, además de la fragmentación de las secuencias, el módulo gestor de fragmentación se encarga de simular la desnaturalización y colisión de fragmentos de ADN.

Para simular la desnaturalización, el módulo implementa un algoritmo de almacenamiento y reproducción, donde las secuencias fragmentadas son almacenadas en un archivo plantilla unidimensional P[i], que asume la orientación de los fragmentos de ADN

almacenados en dirección 5’-3’. Este archivo es duplicado como un archivo F[i] para preservar los datos, mientras que se crea un archivo unidimensional adicional C[i] que almacena las secuencias complementarias de los fragmentos en dirección 3’-5’ (Figura 16).

Por otra parte, para recrear el proceso de colisión basta con usar un generador aleatorio que permite la selección al azar de dos fragmentos cualesquiera de los archivos F[i] o C[i]; estos fragmentos son ingresados al módulo gestor de hibridación (Figura 16).

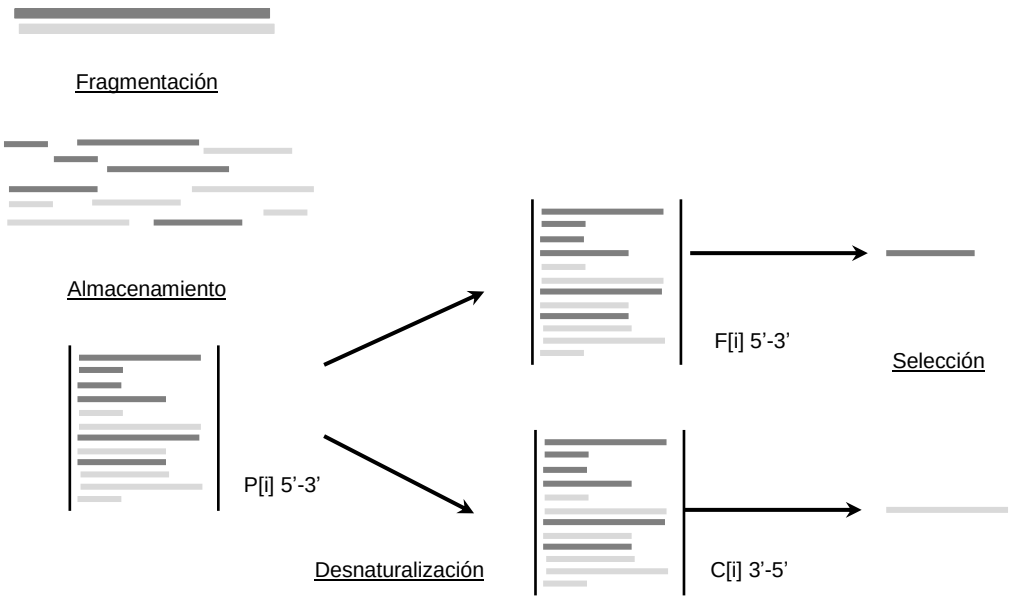


Figura 16. Esquema del proceso algorítmico de la gestión de fragmentos

Una generalización del algoritmo utilizado por el módulo de fragmentación es (Figura 16):

- ✓ Leer las secuencia de nucleótidos de los genes parentales en el archivo multifasta.
- ✓ Leer la longitud promedio de fragmentación (Dada por el módulo de gestión de condiciones de barajado)

- ✓ Generar las distancias de corte aleatorias a partir de la implementación de una distribución tipo Poisson
- ✓ Almacenar los fragmentos de secuencia 5'-3' en un archivo unidimensional P[i]
- ✓ Duplicar la información en un archivo unidimensional F[i]
- ✓ Encontrar los fragmentos complementarios 3'-5'
- ✓ Almacenar los fragmentos complementarios 3'-5' en un archivo unidimensional C[i]
- ✓ Generar dos números aleatorios, mediante una distribución uniforme en el rango desde 1 hasta i, para obtener los índices correspondientes a los archivos F[i] y C[i] que almacen los fragmentos.
- ✓ Seleccionar a partir de los índices correspondientes a los números aleatorios los fragmentos que colisionan.

3.2.3 Módulo gestor de hibridación. El módulo gestor de hibridación utiliza el método de mínima energía expuesto en la sección 2.2.4., en donde se calcula la probabilidad de hibridación de los diferentes N caminos, a partir de la comparación de la energía libre de cada uno de ellos (Ecuación 15). El módulo usa para el cálculo de las energías libres el modelo del vecino más cercano (Ecuación 12). Estos valores energéticos encontrados se transforman en una variable α dependiente de la ecuación de equilibrio de Boltzman (Ecuaciones 13, 16), que expresa en últimas la probabilidad de hibridación (Ecuación 17).

Encontrar esta probabilidad por camino de hibridación es suficiente para predecir cuál es el más favorable, pero dado que no en toda colisión se da hibridación, un modelo estocástico de estados resulta una mejor aproximación al proceso *in vitro* (Maheshri & Schaffer, 2003). Por

ello, el módulo gestor de hibridación implementa un modelo markoviano de dos estados que aprovecha la probabilidad de hibridación (X) hasta ahora expuesta (Figura 17) (Maheshri & Schaffer, 2003).

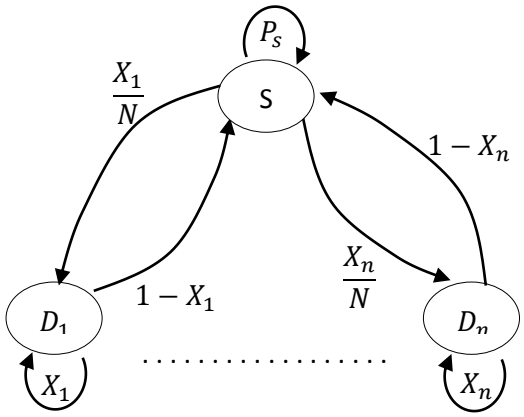


Figura 17. Modelo markoviano de conversión de estados (Cadena sencilla, Cadena doble)

Después de calculada la probabilidad (X) de cada camino, el módulo gestor de hibridación pasa a la construcción de una matriz de probabilidad de dos estados (Ecuación 22). Esta matriz representa el modelo markoviano (Figura 17), de tal forma que los fragmentos colisionados pueden pasar a cualquiera de los D_n estados posibles (Estados de hibridación o de doble hebra de ADN) o pueden permanecer en el estado S (Estado de no hibridación o de hebra sencilla de ADN) (Figura 17). En la matriz de probabilidad P (i x i), la probabilidad de cambiar del estado S a un estado D está dado por X_i/N, la probabilidad de transitar de un estado D al estado S está dado por $1\text{-}X_i$, mientras que permanecer en un estado sin realizar transición está dado por P_s o X_i. La suma de las probabilidades de transición en cada fila es igual a 1 (Ecuación 22).

$$[P] = \begin{bmatrix} P_s & \frac{X_1}{N} & \frac{X_2}{N} & \frac{X_3}{N} & \dots & \frac{X_i}{N} \\ 1 - x_1 & x_1 & 0 & 0 & \dots & 0 \\ 1 - x_2 & 0 & x_2 & 0 & \dots & 0 \\ 1 - x_3 & 0 & 0 & x_3 & \dots & 0 \\ 1 - x_4 & 0 & 0 & 0 & \dots & 0 \\ 1 - x_i & 0 & 0 & 0 & \dots & x_i \end{bmatrix} \quad (22)$$

De esta forma, el módulo de gestión de hibridación implementa una matriz de probabilidades de dos estados, para cada colisión, que sea predicha por el módulo gestor de fragmentos.

3.2.4 Módulo gestor de bibliotecas quiméricas. Este es el último módulo del componente SAssembly, encargado de: i) resolver la matriz de probabilidad para definir si se realiza una transición a un estado D_n o si se mantienen los fragmentos en estado *S*; ii) Realizar el proceso de extensión; iii) Alimentar los archivos unidimensionales con los fragmentos extendidos y con los fragmentos no colisionados (Figura 18). Este módulo administra un criterio de parada para cada ciclo de PCR y otro para el proceso de barajado total. El primer criterio de parada permite la simulación de un ciclo de PCR hasta que menos de *z* fragmentos han sido sometidos a más de *x* colisiones, sin resultar en modificación de la población de fragmentos hibridados. Estos fragmentos no hibridados también son tenidos en cuenta para colisionar en el siguiente ciclo PCR. Por su parte, el segundo criterio, el poblacional, permite el proceso de barajado evaluando la eficiencia de la biblioteca, que equivale a barajar hasta que un porcentaje *w* de secuencias de la población han logrado una longitud de nucleótidos aproximadamente igual a los genes parentales.

La matriz de probabilidad se puede resolver dado que la matriz cumple con las características de matriz ergódica, ya que todos los estados del modelo de Markov aquí propuesto son: recurrentes, aperiódicos y se comunican entre sí (Winston, 2004). Estas características, permiten mediante n productos matriciales sucesivos de P encontrar una matriz de transición T, que garantiza encontrar todas las probabilidades de transición entre los estados (Figura 18). Es decir, siendo T(i x i) la matriz de transición de una cadena ergódica P(i x i), existe un vector $\pi = [\pi_1 \ \pi_2 \ldots \pi_i]$ o distribución de estado estable, que contiene, las probabilidades de todas las posibles transiciones (Ecuación 23) (Winston, 2004).

$$T_{(ixi)} = \lim_{n\to\infty}[P^n] = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_i \\ \pi_1 & \pi_2 & \cdots & \pi_i \\ \pi_1 & \pi_2 & \cdots & \pi_i \end{bmatrix} \quad (23)$$

Sin embargo, el coste computacional de realizar n productos matriciales es alto e incremental, debido que en cada ciclo de PCR, las longitudes de los fragmentos que colisionan son cada vez mayores, acrecentando las dimensiones de las matrices de probabilidad, gobernadas por los N posibles caminos de hibridación (Ecuación 15, sección 2.2.4). Por ello, se implementó una solución algebraica que calcula la distribución de estado estable mediante la resolución de un sistema de ecuaciones (Ecuación 24) En este sistema, la distribución de estado estable es el producto matricial entre el vector π y la matriz de probabilidades P (Winston, 2004).

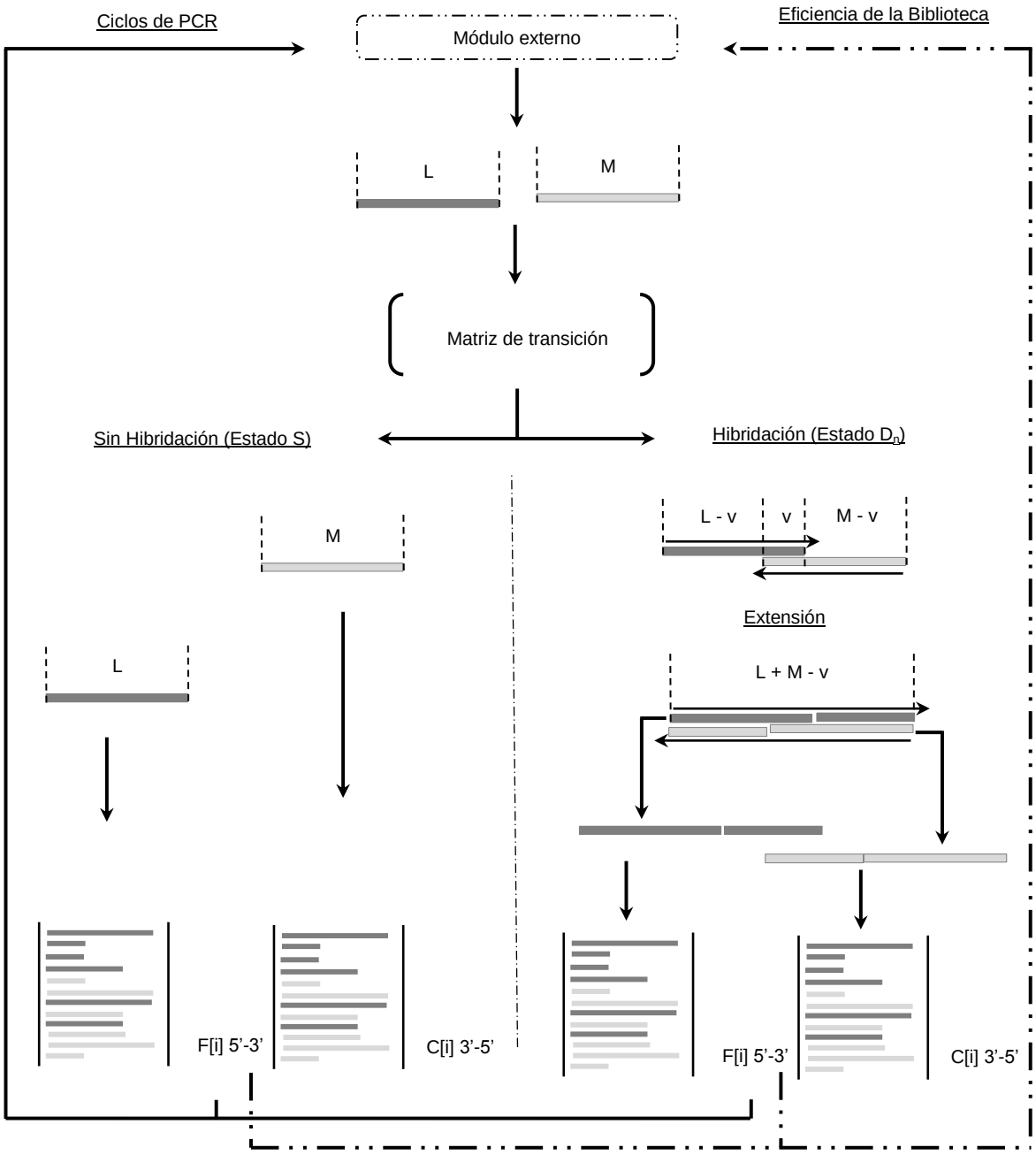


Figura 18. Esquema del proceso algorítmico de la gestión de bibliotecas quiméricas

Al intentar resolver el sistema de ecuaciones (Ecuación 24), la solución es un conjunto de alternativas infinitas. Para que el sistema permita la obtención de un sistema único de soluciones, se introdujo una ecuación de restricción (Ecuación 25), que reemplaza una ecuación cualquiera

del sistema inicial (Winston, 2004). Mediante este artificio el módulo gestor de bibliotecas quiméricas encuentra la distribución de estado estable y decide cuál camino de hibridación es el más estable o si los fragmentos quedan sin hibridar.

$$\pi_j = \sum_{k=1}^{k=i} \pi_k P_{kj} \tag{24}$$

$$\pi_1 + \pi_2 + \cdots \pi_i = 1 \tag{25}$$

Si los fragmentos hibridan, el módulo hace uso de un algoritmo de extensión, que simula la acción de una polimerasa con el 100% de fidelidad. El algoritmo de extensión toma dos fragmentos con longitudes L, K y zona de solapamiento *v*, agregando los nucleótidos complementarios en las zonas incompletas. Es decir, completa en dirección 5'-3' la sección con longitud M-v y en dirección 3'-5' la sección con longitud L-v, obteniendo así un fragmento de doble hebra con una longitud de L + M – v (Figura 18).

Después de realizada la extensión, estos fragmentos al igual que aquellos que no lograron hibridar son almacenados por el módulo en archivos unidimensionales. Los que son extendidos esperando el siguiente ciclo de PCR y los que no hibridan son candidatos para nuevas hibridaciones en el mismo ciclo de PCR si aún existe un número importante de fragmentos por colisionar o en el siguiente ciclo PCR si se cumple el criterio de terminación del ciclo.

De esta forma, SAssembly logra simular el barajado de ADN y obtener bibliotecas quiméricas *in silico*, que corresponden a genes que pueden ser analizados, sintetizados y transformados *in vivo.*

3.3 GenE-in

GenE-in (**Gen**es that **E**volved - **in**), es el tercer y último componente software de DEvISING y fue diseñado para aprovechar técnicas bioinformáticas que permitieron filtrar las bibliotecas quiméricas *in silico*, en búsqueda de las mejores variantes *(Anexo 1)*. Este componente imita el resultado de los pasos de tamizaje y selección de la ED. La arquitectura software del componente está basada en tres módulos integrados que permiten la revisión de las mejores variantes *in silico* (Figura 19).

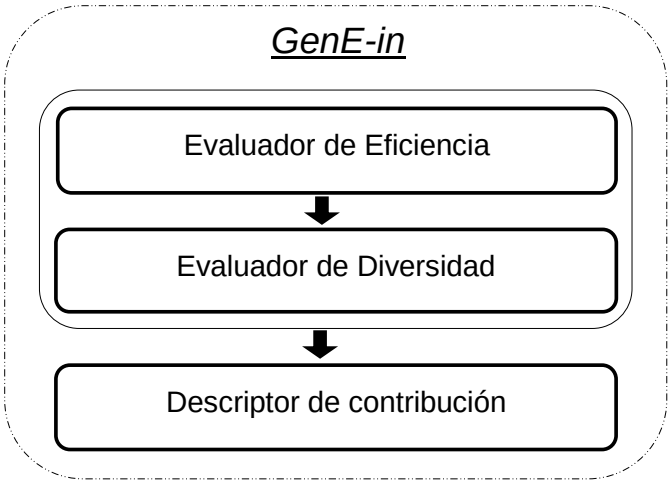


Figura 19. Arquitectura software de GenE-in

GenE-in aprovecha un conjunto de herramientas de análisis bioinformático disponibles para la comunidad académica y las adapta para la evaluación de las bibliotecas quiméricas *in silico* en tres momentos: i) primer momento: evalúa y selecciona los genes más eficientes; ii)

segundo momento: a partir de los genes más eficientes, evalúa y selecciona los genes con mayor identidad con relación a la familia de los genes parentales; iii) tercer momento: describe el grado de contribución de los parentales en la formación de las variantes *in silico* (Figura 19).

3.3.1 Módulo Evaluador de Eficiencia. Este primer módulo de GenE-in hace un filtro a los genes *in silico* barajados. El módulo usa un algoritmo para recorrer una biblioteca quimérica, secuencia a secuencia, identificando marcos abiertos de lectura (secuencias que codifican proteínas), almacenándolas en orden descendente por longitud de nucleótidos. Este primer paso está basado en el concepto de eficiencia en donde, por primera vez, se puede hacer un filtrado a bibliotecas quiméricas *in silico* haciendo posible la reducción del universo de variantes con posibles mejoras funcionales. La eficiencia del ensamblado indica que la longitud de los marcos abiertos de lectura debe ser cercana a la longitud de las proteínas que codifican los genes parentales. Para hacer el cálculo de los marcos abiertos de lectura, GenE-in adapta el código disponible de Oligo 7 http://www.oligo.net/, útil para este propósito (Rychlik, 2007).

3.3.2 Módulo Evaluador de Diversidad. Este segundo módulo de GenE-in evalúa la diversidad proteica de los mejores marcos abiertos de lectura identificados. Para ello, las mejores secuencias almacenadas por el módulo evaluador de eficiencia son comparadas con las bases de datos *online* suministrada por el NCBI (The National Center for Biotechnology Information). Allí se calcula un estimador de identidad para cada secuencia. Este estimador se almacena en orden descendente, siempre y cuando el estimador relacione la variante con la familia a la cual pertenecen los genes parentales. Para realizar este segundo filtro sobre las variantes candidatas,

GenE-in aprovecha el código disponible de BLAST para proteínas el BLASTp https://blast.ncbi.nlm.nih.gov/Blast.cgi (Altschul et al., 1997).

3.3.3 Módulo Descriptor de Contribución. Este último módulo de GenE-in revisa el aporte de los diferentes genes parentales y sus regiones génicas en la formación de las mejores variantes. Estos resultados permiten analizar las implicaciones de la selección de los genes parentales desde el enfoque SANAFold (sección 3.1.).

Para realizar la descripción de contribución, este módulo utiliza la técnica de alineamiento múltiple de secuencias mediante la adaptación del algoritmo ClustalW que provee el software MEGA 4.0.2. http://www.megasoftware.net/mega4/mega.html (Tamura, Dudley, Nei, & Kumar, 2007), y realiza una estimación de valores de identidad entre las mejores variantes y los genes parentales. Para ello, el módulo incorpora el código disponible de MatGAT 2.02., como parte de su arsenal http://ww3.bergen.edu/faculty/jsmalley/matgat.html (Campanella, Bitincka, & Smalley, 2003).

3.4 Conclusiones

Este trabajo ha logrado el desarrollo computacional de DEvISING (Directed Evolution *In Silico* Modeling), mediante la articulación de tres componentes software: SANAFold (*Statistical Analysis of Nucleic Acid Folding*), SAssembly (*Stochastic Assembly*) y GenE-in (Genes that Evolved - in), que simulan: la selección de genes parentales, la generación de diversidad y el

proceso de tamizaje/selección, respectivamente; haciendo de este trabajo la primera aproximación *in silico* que integra paso a paso la técnica de ED.

La estrategia desarrollada logró incorporar, por primera vez, los efectos de las estructuras secundarias de ADN en la generación de bibliotecas quiméricas. El presente trabajo es el primero en proponer una herramienta computacional (SANAFold) para caracterizar y seleccionar genes parentales, a partir del comportamiento termodinámico de sus regiones génicas, en cuanto a su capacidad para favorecer la formación de estructuras secundarias en condiciones experimentales de ED.

Por otra parte, se puede destacar la forma novedosa de implementación de modelos de hibridación y extensión de SAssembly. Ésta facilita el almacenamiento de las secuencias recombinadas, entregando verdaderas bibliotecas de secuencias de ADN, las cuales pueden ser analizadas mediante las técnicas bioinformáticas de GenE-in (proceso *in silico* de tamizaje/selección), logrando, mediante criterios de eficiencia y diversidad, las mejores variantes.

DEVISING propone un nuevo enfoque al modelado *in silico* de ED, presentando con este trabajo una herramienta como proveedora de insumos en los experimentos *in vitro* de ED, en donde las variantes resultantes pueden ser sintetizadas, transformadas *in vivo* y cuya funcionalidad puede ser evaluada *in vitro*.

En el próximo capítulo se presenta un caso de estudio, en donde DEvISING es usado para modelar un experimento de ED, donde se obtienen variantes *in silico* de toxinas Cry11 de *Bacillus thuringiensis*.

4. Caso de Estudio: Experimento de Evolución Dirigida in Silico para Toxinas Cry11 de Bacillus Thuringiensis

En esta sección, se presenta el uso de la estrategia *in silico* DEvISING para encontrar variantes de toxinas Cry11 de *Bacillus thuringiensis* (Bt). Se inicia describiendo las toxinas Cry de Bt, nuestro modelo biológico de interés; posteriormente se hace una presentación de un experimento de *ED in vitro* con toxinas Cry11 de Bt, que es simulado mediante el uso de los tres componentes software de DEvISING: SANAFold, SAssembly y GenE-in. Por medio de éstos se hace la caracterización de genes parentales *cry11*, la generación de bibliotecas quiméricas y la selección de las mejores variantes. Finalmente, los hallazgos *in silico* son presentados y articulados con los hallazgos *in vitro*.

4.1 Modelo biológico: Toxinas Cry11 de Bacillus thuringiensis

Las Toxinas Cry son una familia de proteínas de tres dominios, compuesta por 74 grupos diferentes, con 290 holotipos registrados (http://www.lifesci.sussex.ac.uk/home/Neil_Crickmore/Bt/). Es un grupo ampliamente estudiado y usado con fines biotecnológicos como insecticida microbiano, para el control de plagas agrícolas y vectores propagadores de enfermedades (Abdullah, 2012; J. Y. Wu et al., 2007). La familia de tres dominios de toxinas Cry es la mejor caracterizada evolutivamente, se han realizado varios estudios de cristalografía que han establecido que poseen estructura y modo de acción similar. A pesar de ello, los estudios de letalidad de estas proteínas muestran especificidad

de cada familia de toxinas hacia insectos blanco diferentes: Lepidópteros, coleópteros, dípteros, etc., (Figura 20) (Bravo et al., 2013; George, 2012). Además, estas proteínas resultan inocuas para el resto de especies, incluyendo al ser humano (Soberón, 2007), propiedad que ha permitido su uso en la industria como un producto bio-pesticida por excelencia (Mahadeva, 2012).

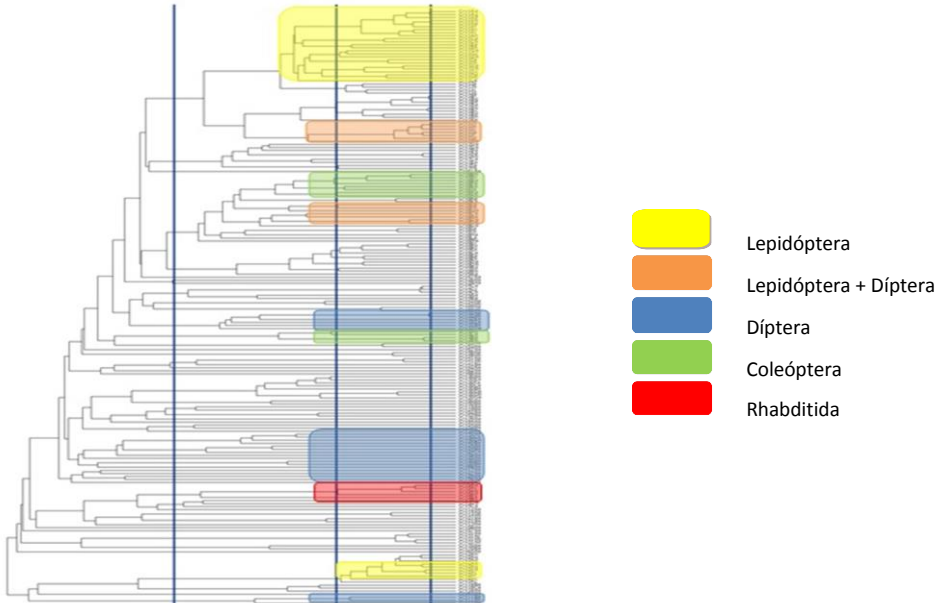


Figura 20. Filogenia de las toxinas Cry y letalidad asociada. Este Dendograma de toxinas Cry de tres dominios fue adaptado de http://www.lifesci.sussex.ac.uk/home/Neil_Crickmore/Bt/ y coloreado para indicar la especificidad insecticida de las toxinas, según la información registrada en las bases de datos del Natural Resources Canada, http://www.glfc.forestry.ca/bacillus/.

4.1.1 Modo de Acción de las Toxinas Cry. Cada uno de los tres dominios (Figura 21) que conforman esta familia de proteínas tiene un papel en la acción tóxica. Cuando un organismo blanco (insecto) ingiere la toxina Cry se desencadenan una serie de acciones bioquímicas letales (George, 2012). La estructura de δ-endotoxinas de B. *thuringiensis* incluye las regiones

responsables de encontrar el objetivo de membrana en el intestino del insecto blanco y crear los poros de iones. Esta función se le atribuye particularmente al dominio I de la δ-endotoxina, que está formado por siete α-hélices (Figura 21) (Schnepf et al., 1998). Por su parte, el dominio II, que consta de tres hojas β plegadas (Figura 21), actúa anclando la toxina a cada uno de los bucles expuestos en la superficie del intestino del insecto, adhiriéndola de forma irreversible. Finalmente, el dominio III de la toxina Cry formada por una estructura β-sandwich (Figura 21), interactúa con otros receptores desencadenando una acción conjunta de desequilibrio osmótico que finaliza con la muerte del insecto (George, 2012; J. Y. Wu et al., 2007).

Sin embargo, la resistencia de los insectos a las toxinas ha sido demostrada y empieza a evidenciarse como un problema de aplicación (George, 2012; Storer et al., 2010; Tabashnik, Gassmann, Crowder, & Carriere, 2008), que impulsa el desarrollo de investigaciones hacia la búsqueda de nuevas variantes más potentes o que controlen una mayor variedad de organismos blanco (Bravo et al., 2013). Es en esta búsqueda de nuevas variantes donde la ingeniería de proteínas toma un lugar protagónico en las investigaciones biotecnológicas con *B. thuringiensis*. Así mismo, las técnicas de ED con mutagénesis ADN recombinante como el barajado de ADN son ahora las alternativas tecnológicas para mejorar la toxicidad y diversificar la actividad a otras especies blanco de interés (Flórez, 2012).

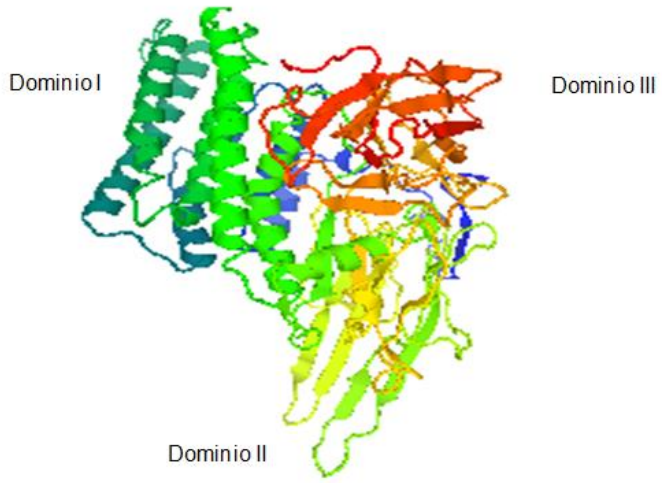


Figura 21. Estructura de la proteína Cry2Ab con alta homología con Cry11Aa. Adapatado de el visualizador http://www.proteinmodelportal.org/ (Arnold et al., 2009).

4.1.2 Ingeniería de proteínas con Toxinas Cry. Algunos esfuerzos se han hecho desde la mutagénesis dirigida para encontrar variantes Cry. Un estudio reciente plantea algunos patrones génicos observados en las toxinas Cry silvestres que mejoran la toxicidad y promueven la diversidad (Bravo et al., 2013). Por ejemplo, se ha propuesto que el intercambio *in vitro* del segmento del gen que codifica para el dominio III de la proteína entre familias Cry es una buena estrategia para crear toxinas híbridas contra plagas que no muestran la susceptibilidad a las toxinas Cry originales (Bravo et al., 2013), mientras que mutaciones puntuales en el segmento del gen que codifica para los dominios I y II aumentan la toxicidad de la proteína (Alzate, Osorio, Florez, & Dean, 2010; Hussain, 2010, 2011; S. F. Wu, A; Homoelle, B; Dean, D; Alzate, O., 2012). Estos hallazgos resultan alentadores, pues muestran que algunas variaciones estructurales tienen efectos significativos en la mejora funcional de las toxinas silvestres (Bravo et al., 2013).

Por otra parte, la ED como técnica de la ingeniería de proteínas también ha demostrado casos exitosos de variantes mejoradas: los genes *cry*1Ca y *cry*11A12 han sido usados como parentales en técnicas de barajado y han demostrado mejoras funcionales. En el trabajo de Lassner *et al*., se reportó el aumento de la toxicidad del gen *cry*1Ca contra larvas de Rosquilla verde (*Spodoptera exigua*) y gusano del fruto (*Helicoverpa zea*), mientras que el trabajo desarrollado por Craverio *et al*. logró ampliar el espectro de acción de la toxina *cry*11A12 a la especie barrenados gigantes de la caña de azúcar (*Telchin licus licus),* para la cual la toxina producida por el gen parental no era letal (Craveiro et al., 2010; Lassner & Bedbrook, 2001). Estos dos trabajos demuestran el potencial de la ED *in vitro* para incrementar la actividad letal contra el blanco o ampliar el espectro de acción de una toxina.

El trabajo en ingeniería de proteínas y en especial en ED con Toxinas Cry es prometedor. Es muy probable que desde ED *in silico* se puedan encontrar mejores variantes Cry, con mayor actividad tóxica y mayor espectro de acción, hacia nuevos organismos blanco.

4.2 Experimentos in vitro de ED con Toxinas Cry11

La importancia biotecnológica de las toxinas Cry11 (Cry11Aa, Cry11Ba, Cry11Bb) radica en que se ha identificado que esta familia está conformada por proteínas con actividad contra insectos transmisores de enfermedades del orden díptera tales como: *Anopheles albimanus*, *Aedes aegypti* y *Culex quinquefasciatus* (de Maagd, Bravo, & Crickmore, 2001; Schnepf et al., 1998). Por tanto, la obtención de variantes de Cry11 es un tema de interés sanitario.

4.2.1 Condiciones experimentales. A continuación se presentan las condiciones experimentales y resultados obtenidos por nuestro grupo al realizar ED *in vitro* (barajado de ADN) para obtener variantes de toxinas Cry11.

- Selección del parental

Se aislaron los genes *cry11Aa*, *11Ba* y *11Bb* mediante PCR a partir de los constructos que los contenían (pJEG80.1/cry11Ba, p/cry11Aa, pBTM5/cry11Bb). Cada uno de los fragmentos fue clonado en pCR®4-TOPO (Invitrogen) y se verificó el aislamiento mediante análisis de secuencia de los fragmentos obtenidos (Morales, 2011).

- Generación de diversidad (Barajado de ADN)

Se realizaron ensayos de digestión con DNAsaI®. Para ello se utilizaron 3 µg de ADN en una solución de 0.0006 U/µl de DNAsa I® (Invitrogen) en presencia de Mn^{2+} y se incubaron a temperatura ambiente a diferentes tiempos hasta obtener fragmentos entre 25 y 250 bp. Las digestiones fueron visualizadas en geles de agarosa al 2.5% y purificadas (Morales, 2011).

Con estos fragmentos se procedió a la generación de diversidad: desnaturalización, hibridación y extensión. Para ello se usaron 30 ng/µl del producto purificado, se adicionó a una mezcla sin cebadores que contenía Mix 1X de buffer *Pfx* y 2.5U de *Pfx*® DNA polimerasa (Invitrogen) utilizando 45 ciclos a 94°C por 30 seg, 48°C por 3 min y 68°C por 1min + 12 s. por ciclo y una extensión final de 7 minutos a 72ºC (Morales, 2011).

Para la obtención de la biblioteca quimérica, un microlitro de esta reacción fue usado como molde en una reacción de PCR de 25 µl que contenía: 0.4 µmol de cada primer (Tabla 1), buffer Taq 1X, 0.4mM de cada dNTP y 2.5U Taq® polimerasa (Promega) y sometida a 94ºC por 4 minutos, y 25 ciclos de 94ºC por 45seg, 55ºC por 1 min, 68ºC por 4 minutos + 20 s./ciclo y finalmente a 72ºC por 10 minutos (Morales, 2011).

- Tamizaje/Selección

Los productos de reensamblaje fueron purificados y clonados en el sistema de clonación TOPO® (Invitrogen). La homología entre las secuencias de los clones con las secuencias de los genes *cry11* fue determinada mediante el análisis con la herramienta bioinformática Blast (Altschul et al., 1997; Morales, 2011).

Tabla 1.

Primers para la obtención de la biblioteca quimérica

Nombre Del Primer	Secuencia Del Primer
PCR4F	GAT AAC AAT TTC ACA CAG GA
PCR4R	TTG TAA AAC GAC GGC CAG TG
PGE7F	GAT GTG CTG CAA GGC GAT T
PGE7R	TTA CGC CAA GCT ATT TAG GTG

- Prueba de Toxicidad

Finalmente los productos fueron sometidos a ensayos de letalidad, se realizaron 2 replicas por proteína, cada replica constaba de 30 larvas de *A. aegypti*, se tomaron en cuenta 7 dosificaciones,

utilizando un total de 420 larvas por variante. Se realizo un recuento de las larvas vivas que se tradujo a porcentaje de larvas muertas (Suarez, 2016).

4.2.2 Resultados. Los experimentos de ED con genes *cry11* realizados por nuestro grupo, permitieron la obtención de 94 variantes (clones), cuya distribución fue: 34 variantes [< 1 Kb], 14 variantes [1-2 kb], 22 variantes [> 2.1 Kb], 14 variantes sin homología con *cry11* y 10 clones restantes sin insertos (Florez et al., 2016). De las 70 variantes con insertos se logró una eficiencia cercana al 31% (Figura 22) con longitudes de las variantes [>2.1 Kb]. De estas 70 variantes nuestro grupo ha realizado un primer análisis de 5 de ellas, de las cuales ya se identificaron 2 con mejoras en su actividad toxica frente a los parentales (Suarez, 2016).

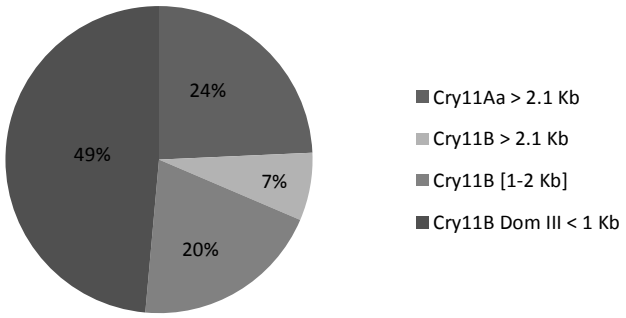


Figura 22. Diversidad y Eficiencia de barajado de ADN in vitro con genes cry11

4.3 Uso de la estrategia DEVISING

A continuación se presenta el uso de DEvISING en el modelado *in silico* de ED, para encontrar variantes de toxinas Cry11 de *Bacillus thuringiensis*. DEvISING fue desarrollado en el lenguaje

de programación Python 3.2 y ejecutado mediante en una arquitectura hardware con cluster, conformado por cinco nodos integrados con el sistema operativo Linux (distribución Fedora 20). Estas características de implementación son innovadoras en el campo de estudio de los modelos *in silico* de ED y hacen de DEvISING la primera estrategia computacional potente, capaz de enfrentar una demanda masiva de simulaciones que recrean las condiciones físico-químicas a nivel biomolecular de los experimentos ADN recombinantes. DEvISING es la primera estrategia computacional que integra las técnicas de ED y que es capaz de predecir bibliotecas quiméricas *in silico*.

En las próximas secciones se muestra el uso de los componentes de DEvISING: SANAFold, SAssembly y GenE-in, que funcionan articulados con el propósito de encontrar variantes *in silico* de Toxinas Cry11.

4.3.1 Uso de SANAFold. SANAFold y su arquitectura software fueron expuestos con detalle en la sección 3.1. Este componente permite la caracterización de genes parentales en experimentos de ED mediante el análisis del comportamiento termodinámico de las regiones génicas, asociadas a la formación de estructuras secundarias de ADN en condiciones de barajado. A continuación se presentan los escenarios de simulación, los resultados y una comparación de los resultados con los datos experimentales encontrados en la literatura científica, después de usar SANAFold.

- Escenarios de simulación

Cada escenario de simulación sometió las secuencias de genes *cry* a tres condiciones experimentales de barajado de ADN: temperatura (TE), concentración de Mg^{++} (MA) y longitud promedio de los fragmentos (LE). Los experimentos *in silico* se diseñaron combinando las tres condiciones donde para cada combinación se mantuvo una condición constante y las otras dos usaron rangos de datos para sus variaciones. Los valores de los datos fueron contemplados a partir de condiciones experimentales de barajado *in vitro* usados por nuestro grupo (sección 4.2.1.). Los rangos establecidos son: TE = [48°C-68°C], MA= [0.02mM-1mM] y LE= [50pb-250pb]. Los valores constantes utilizados como condición no variable fueron seleccionados a partir de resultados obtenidos en escenarios de prueba, donde dichos valores mostraron mayor formación de estructuras secundarias de ADN: TE= 48°C, MA = 0.5mM y LE=50pb. De esta forma se establecieron los siguientes escenarios de simulación: i) LE-TE: se realizaron combinaciones de valores de longitud promedio de los fragmentos y de temperatura, quedando constante MA=0.5mM; ii) LE-MA: se realizaron combinaciones de valores de longitud promedio de los fragmentos de ADN y de concentración iónica de magnesio, quedando constante TE= 48°C, iii) TE-MA: se realizaron combinaciones de valores de temperatura y de concentración iónica de magnesio, quedando constante LE= 150pb.

Por otra parte, se estableció un escenario de referencia en donde todos los genes fueron evaluados. El escenario establecido como referencia fue un LE-MA, con valores de LE= [50pb-250pb], TE=37°C, MA=[0.02mM-1mM]. Los resultados obtenidos con este escenario fueron usados para revisar las variaciones en los comportamientos de los genes *cry* ante condiciones de

barajado y se seleccionaron para la caracterización de los genes parentales *cry11* (clúster I) y familias de genes *cry*, según su distancia filogenética con respecto a *cry11*: familias cercanas (clúster II), medianamente lejanas (clúster IV) y lejanas (clúster III) (Tabla 2).

Tabla 1.

Clústeres de genes cry estudiados.

Clúster	Toxina	Referencia	Fuente de extracción	Marco Abierto de Lectura		
				AA	St – Sp (pb)	# Acceso GenBank
I	Cry11Aa1	Donovan *et al*, 1988	*Bt israelensis*	646	32-1972	M31737 J03510
	Cry11Ba1	Delecluse *et al*, 1995	*Bt jegathesan*	724	64-2238	X86902
	Cry11Bb1	Orduz *et al*, 1998	*Bt medellin*	786	1-2346	AF017416
II	Cry2Aa1	Donovan *et al*, 1989	*Bt kurstaki*	633	156-2057	AF273218
	Cry18Aa1	Zhang *et al*, 1997	*Paenibacillus popilliae**	712	725-2863	X99049
III	Cry1Aa1	Schnepf *et al*, 1985	Bt kurstaki HD1	1176	527-4057	EU357805
	Cry1Ab1	Wabiko *et al*, 1986	*Bt berliner 1715*	1155	1-1695	EU357806
IV	Cry30Aa1	Delecluse *et al*, 2000	*Bt medellin*	662	60-2045	AB125059
	Cry30Ca1	Ohgushi *et al*, 2004	*Bt sotto*	688	1-2064	GQ368655

Nota: *Bacteria de la especie *Firmicutes*, clase *Bacilli*, orden *Bacillaes*, familia *Paenibacillaceae*, género *Paenibacillus*.

- Resultados

Todas las simulaciones fueron replicadas cinco veces para garantizar significancia estadística. Un poco más de $18.2 * 10^6$ datos termodinámicos fueron almacenados. Del análisis de estos datos se obtuvieron un total de 162 valores *f-ratio* que resumen el comportamiento termodinámico por clústeres de genes *cry*. De estos, 41 valores *f-ratio*, un 25,3% de los análisis

realizados, que representan a un poco más de $4.6 * 10^6$ de lo datos termodinámicos calculados, presentaron diferencias estadísticas significativas (Tabla 3).

Tabla 2.

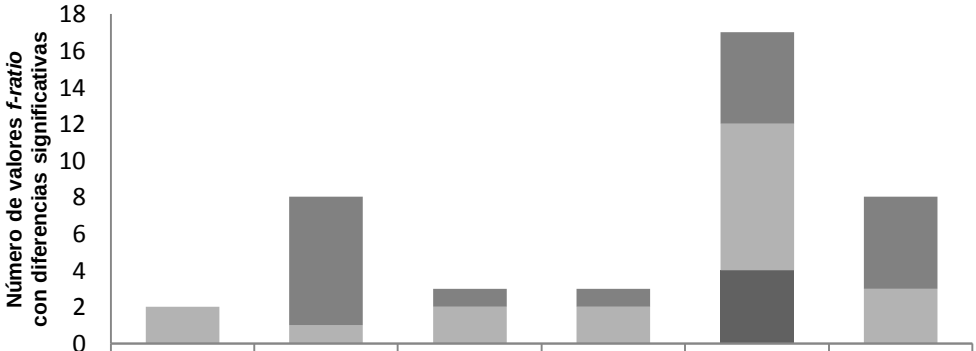
Valores *f-ratio* con diferencia estadística significativa (regiones génicas de cada gen).

Clúster	ADN	*F-ratio*						Condiciones Barajado de ADN	Gl (n:d)	Medida *F-ratio*
		# SSΔG		% SSΔG		$\overline{\Delta G}$				
		-	+	-	+	-	+			
								LE-TE	2:24	3,40
	cry11Aa1					6,60		LE-MA	2:33	3,28
			5,97			8,25	3,30	TE-MA	2:33	3,28
								LE-TE	2:24	3,40
I	*cry11Ba1*	6,45		4,41	4,41	9,90	3,30	LE-MA	2:33	3,28
			8,69				3,30	TE-MA	2:33	3,28
								LE-TE	2:24	3,40
	cry11Bb1						4,12	LE-MA	2:33	3,28
			8,81			4,12		TE-MA	2:33	3,28
						9,00		LE-TE	2:24	3,40
	cry2Aa1					28,00		LE-MA	2:33	3,28
II						6,19	4,12	TE-MA	2:33	3,28
								LE-TE	2:24	3,40
	cry18Aa1					5,50		LE-MA	2:33	3,28
								TE-MA	2:33	3,28
						4,26		LE-TE	2:24	3,40
	cry1Aa1					21,35		LE-MA	2:33	3,28
III			6,85	5,44	5,73	7,50	4,71	TE-MA	2:33	3,28
						3,60		LE-TE	2:24	3,40
	cry1Ab1					16,50	3,30	LE-MA	2:33	3,28
			5,77					TE-MA	2:33	3,28
						6,00		LE-TE	2:24	3,40
	cry30Aa1	4,69				8,25		LE-MA	2:33	3,28
IV			6,67					TE-MA	2:33	3,28
								LE-TE	2:24	3,40
	cry30Ca1		5,22	4,58	4,67	7,62		LE-MA	2:33	3,28
			8,21			4,12	7,07	TE-MA	2:33	3,28

Nota: *# SSΔG (Número de Estructuras secundarias de ADN con energía libre postiva o negativa con f-ratio significativo), % SSΔG (Porcentaje de Estructuras secundarias de ADN con energía libre positiva o negativa con f-ratio significativo), $\overline{\Delta G}$ (Promedio de energía libre positiva o negativa con f-ratio significativo), Gl (n:d) (Grados de libertad de los datos, n: numerador y d: denominador).

Estas diferencias significativas fueron el punto de partida para el análisis de comportamiento. Cada estimador estadístico diferenciado permitió observar las tendencias en la prevalencia de regiones que favorecen o desfavorecen la formación de estructuras secundarias de ADN. De acuerdo con lo anterior, el estimador $\overline{\Delta G(-)}$ fue el más representativo de los seis estimadores utilizados en este estudio, con una presencia del 41.4%. Por tal razón, este estimador fue usado para el análisis del comportamiento de los cuatro clústeres de genes *cry*.

Por otra parte, haciendo una revisión de la frecuencia de variación a partir de los estimadores estadísticos, se observó que las regiones génicas *cry* son variables ante condiciones con variaciones de Mg^{++}. Éstas fueron evidenciadas en 18 variaciones de valores *f-ratio* en condiciones LE-MA y 19 variaciones de valores *f-ratio* en condiciones TE-MA. Cuando se tienen condiciones LE-TE, es decir, ante la ausencia de variaciones de Mg^{++}, disminuye la variación de las regiones génicas a 4 valores *f-ratio* con diferencias estadísticas significativas (Figura 23).



	#SSΔG(-)	#SSΔG(+)	%SSΔG(-)	%SSΔG(+)	Prom_ΔG(-)	Prom_ΔG(+)
Te-Ma	0	7	1	1	5	5
Le-Ma	2	1	2	2	8	3
Le-Te	0	0	0	0	4	0

Figura 23. Comportamiento de los estimadores estadísticos en los clústeres cry.

- Escenario de Referencia

En este escenario de simulación los clústeres de genes *cry* presentaron un comportamiento, favorable y espontáneo, para la formación de estructuras secundarias de ADN. Los rangos energéticos a partir del estimador $\overline{\Delta G(-)}$ que establecen la energía promedio de las estructuras secundarias de ADN formadas por los fragmentos de los genes *cry* variaron entre -1.0 y -2.2 Kcal/mol. Los genes *cry1Aa*, *cry11Aa*, *cry30Ca* presentaron mayor espontaneidad (Figura 24).

Los clústeres *cry* presentaron de forma general una baja en su capacidad termodinámica para formar espontáneamente estructuras secundarias de ADN en condiciones experimentales LE-MA de barajado de ADN, al compararse con el escenario de referencia. Esto fue evidenciado mediante la observación en los valores más negativos del estimador $\overline{\Delta G(-)}$ en condiciones de referencia (Figura 24).

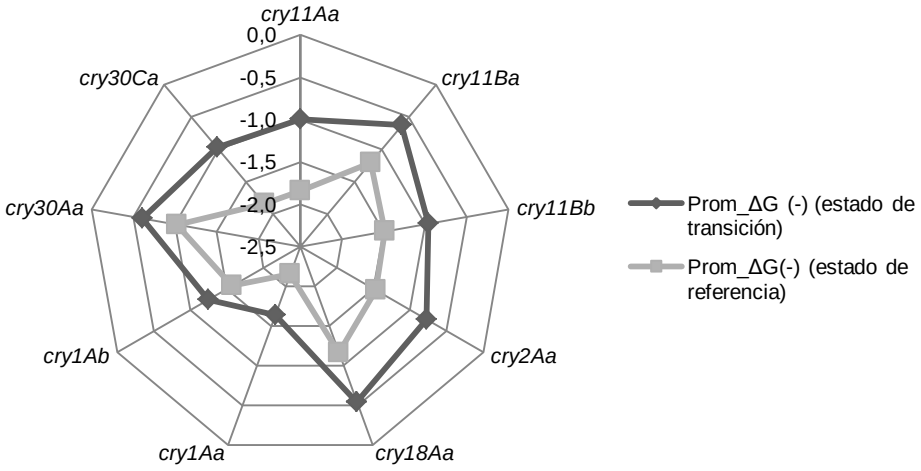


Figura 24. Comparativo termodinámico (Kcal/mol) entre el escenario de referencia y el estado de transición en condiciones LE-MA de barajado de ADN.

- Primer clúster de análisis

El clúster I correspondió al análisis de los genes *cry11Aa1, cry11Ba1 y cry11Bb1*. Para este clúster se evaluaron 54 valores *f-ratio* de los cuales 14, equivalente al 25.9% de los análisis, presentaron diferencias estadísticas significativas (Tabla 3). Al evaluar la dispersión de los datos entre las regiones que codifican para dominios I, II y III del gen *cry11Aa1* se observaron 4 valores *f-ratio* con diferencias significativas en condiciones de simulación LE-MA y TE-MA. Para todos los casos, el análisis del conjunto de datos de cada *f-ratio* mostró el mismo comportamiento termodinámico por regiones dando como resultado una alta espontaneidad en la región del gen que codifica para el dominio I, seguida con media espontaneidad de la región que codifica para el dominio III y una baja espontaneidad de la región del gen que codifica para el dominio II (Figura 25).

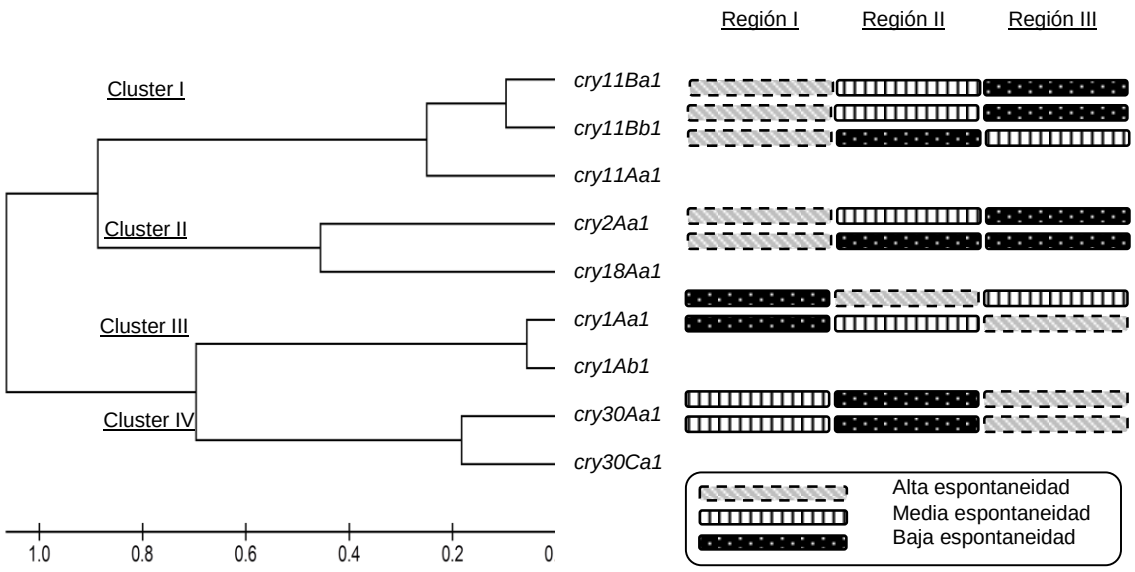


Figura 25. Asociación de la filogenia de los clústeres cry con su comportamiento termodinámico para la formación de estructuras secundarias de ADN.

Al evaluar la dispersión de los datos entre las regiones del gen *cry11Ba1* que codifican para los dominios I, II y III se observaron 7 valores *f-ratio* con diferencias estadísticas significativas en condiciones de simulación LE-MA y TE-MA (Tabla 3). Para todos los casos el análisis del conjunto de datos de cada *f-ratio* mostró el mismo comportamiento termodinámico por regiones dando como resultado una alta espontaneidad en la región del gen que codifica para el dominio I, seguida con media espontaneidad de la región del gen que codifica para el dominio II y con una baja espontaneidad de la región del gen que codifica para el dominio III (Figura 25). Finalmente al evaluar las regiones del gen *cry11Bb1* que codifican para los dominios I, II y III, se observaron 3 valores *f-ratio* con diferencias significativas en condiciones de simulación LE-MA y TE-MA (Tabla 3). Así mismo, el comportamiento termodinámico del conjunto de datos de cada *f-ratio* por regiones se mantuvo igual dando el mismo resultado en términos de espontaneidad, a los obtenidos para el gen *cry11Ba1* (Figura 25). No se observaron diferencias significativas en condiciones de LE-TE para este clúster.

- Segundo clúster de análisis

El clúster II correspondió al análisis de los genes *cry2Aa1* y *cry18Aa1*. Para este clúster se evaluaron 36 valores *f-ratio* de los cuales 5 valores, 13.8% de los análisis, presentaron diferencias estadísticas significativas (Tabla 3). Al evaluar la dispersión de los datos entre las regiones del gen *cry2Aa1* que codifica para los dominios I, II y III, se observaron 4 valores *f-ratio* con diferencias significativas en condiciones de simulación LE-MA, LE-TE y TE-MA (Tabla 3). Para estos casos, la tendencia del comportamiento termodinámico por regiones mostró el mismo resultado en términos de espontaneidad a los obtenidos con los genes *cry11Ba* y *Bb*

(Figura 25). Finalmente, para este clúster al evaluar las regiones del gen *cry18Aa1* que codifica para los dominios I, II y III se observó 1 valor *f-ratio* con diferencia significativa en condiciones LE-MA (Tabla 3). La tendencia del comportamiento termodinámico mostró los mismos resultados a los obtenidos con los genes *cry11Ba* y *Bb* para las regiones I y III (Figura 25).

- Tercer clúster de análisis

El clúster III correspondió al análisis de los genes *cry1Aa1* y *cry1Ab1*. Para este clúster se evaluaron 36 valores *f-ratio* de los cuales 11 valores, 30.5% de los análisis, presentaron diferencias estadísticas significativas (Tabla 3). Al revisar la dispersión de los datos entre las regiones del gen *cry1Aa1* que codifican para los dominios I, II y III, se observaron 7 valores *f-ratio* con diferencias significativas en todas las condiciones de barajado con una mayor variación de los estimadores cuando las secuencias génicas fueron evaluadas en condiciones de simulación TE-MA (Tabla 3). Para todos los casos, el análisis del conjunto de datos de cada *f-ratio* mostró el mismo comportamiento termodinámico por regiones dando como resultado una alta espontaneidad en la región del gen que codifica para el dominio II, seguida con media espontaneidad de la región del gen que codifica para el dominio III y con una baja espontaneidad de la región I (Figura 25).

Finalmente, en las regiones del gen *cry1Ab1* que codifican para los dominios I, II y III se observaron 4 valores *f-ratio* con diferencias significativas en LE-TE, LE-MA y TE-MA. El análisis del conjunto de datos de cada *f-ratio* fue diferente a *cry1Aa1* indicando una alta espontaneidad en la región que codifica para el dominio III, seguida con media espontaneidad de

la región del gen que codifica para el dominio II y con una baja espontaneidad de la región del gen que codifica para el dominio I (Figura 25).

- Cuarto clúster de análisis

El clúster IV correspondió al análisis de los genes *cry30Aa1, cry30Ca*. Para este clúster se evaluaron 36 valores *f-ratio* de los cuales 11 valores, equivalente al 30.5% de los análisis, presentaron diferencias estadísticas significativas (Tabla 3). La dispersión de los datos entre las regiones del gen *cry30Aa1* que codifica para los dominios I, II y III, presentó 4 valores *f-ratio* con diferencias significativas en condiciones LE-TE, LE-MA y TE-MA (Tabla 3). El análisis del conjunto de datos de cada *f-ratio* mostró el mismo comportamiento termodinámico dando como resultado una alta espontaneidad en la región que codifica para el dominio III, mientras que presentó variaciones entre las regiones que codifican para los dominio I y II, entre la media y baja espontaneidad; prevaleciendo el comportamiento para el estimador $\overline{\Delta G(-)}$ con la región que codifica para el dominio I con espontaneidad media y la región que codifica para el dominio II, con espontaneidad baja (Figura 25). Por otra parte, la dispersión de los datos entre las regiones del gen *cry30Ca1* que codifica para los dominios I, II y III mostró 7 valores *f-ratio* con diferencias significativas con LE-MA y TE-MA. Para este gen la tendencia del comportamiento termodinámico por regiones se mantuvo dando como resultado una alta espontaneidad en la región que codifica para el dominio III, seguida con media espontaneidad de la región que codifica para el dominio I y con una baja espontaneidad de la región correspondiente al dominio II (Figura 25).

- Validación de hallazgos con datos reportados en la literatura científica

La importancia biológica de las estructuras secundarias de ADN ha sido descrita en distintos eventos celulares; en células eucariotas, procariotas y virus (Muhire et al., 2014) (Sander et al., 2014) (Bikard et al., 2010). Así mismo, se ha descrito su importancia en las aplicaciones en biología molecular y la biotecnología asociadas a las técnicas donde ocurre la desnaturalización de las cadenas de ADN ocasionando inhibición en la hibridación o reacciones cruzadas (SantaLucia & Hicks, 2004). En técnicas como el barajado de ADN, la recombinación es la base para realizar ensamblados entre genes parentales y promover la variabilidad genética; por lo tanto, la formación de estructuras secundarias de ADN tiene un papel importante durante la recombinación. En este estudio, se analizaron las variaciones termodinámicas asociadas a la formación de estructuras secundarias de ADN en condiciones de barajado de ADN para un grupo de genes pertenecientes a la familia *cry* de *Bt* organizados en cuatro clústeres según la relación filogenética y los tres dominios altamente conservados relacionados con la función de las toxinas.

El comportamiento termodinámico de los genes *cry* fue similar en todos los clústeres durante las simulaciones en el escenario de referencia. Aunque no presentaron variaciones termodinámicas significativas por regiones, los genes presentaron una mayor espontaneidad $[< \Delta G(-)]$ en comparación con las condiciones de barajado de ADN. En los escenarios de simulación LE-MA, TE-MA y LE-TE el comportamiento termodinámico fue similar entre genes *cry* pertenecientes a un mismo clúster indicando que las variables de longitud de los fragmentos, temperatura y concentración de magnesio son determinantes en las condiciones de barajado y

que, de acuerdo a los análisis basados en la espontaneidad de las regiones génicas, permiten inferir un patrón termodinámico para cada uno.

El primer patrón termodinámico en el clúster I (*cry11Aa1, cry11Ba1, cry11Bb1*) mostró, en la región que codifica para el dominio I, la mayor tendencia a formar estructuras secundarias debido a su alta espontaneidad y por ende valores $< \Delta G(-)$. Sin embargo, en lo que respecta a las regiones que codifican para los dominios II y III, el patrón termodinámico en *cry11Aa* fue distinto. Teniendo en cuenta el comportamiento termodinámico del clúster I, las regiones, y en especial la III para *cry11Ba* y cry11*Bb*, presentaron una menor tendencia a formar estructuras secundarias con respecto a la región II seguido de la región I, siendo esta última la más alta en términos de espontaneidad. Lo interesante de este comportamiento termodinámico es que se mantiene entre las toxinas que están filogenéticamente más relacionadas como ocurre con las toxinas Cry11Ba y Cy11Bb (Figura 25) y a su vez, con el porcentaje de identidad que presentan a nivel de sus secuencias de ADN. Mientras que *cry11Ba1 y cry11Bb1* tienen el 83%, *cry11Aa1* tiene el 62% y 60 % respectivamente (Orduz, Realpe, Arango, Murillo, & Delecluse, 1998). Esto podría explicar las diferencias termodinámicas encontradas entre los mismos genes del mismo holotipo Cry11. La conformación de estos patrones termodinámicos en condiciones de barajado de ADN podrían sugerir la disposición que tienen ciertas regiones para recombinarse entre sí, lo cual favorecería la variabilidad genética. Esto podría tener una relación directa con los resultados obtenidos en experimentos de barajado de ADN realizados por nuestro grupo empleando los tres genes *cry11*. Se ha observado que de los fragmentos ensamblados, el 48.5% son preferiblemente genes incompletos que contienen el dominio III con homología a *cry11B* y que el 31.4% que logra ensamblarse corresponde a genes completos que expresan proteínas con actividad tóxica

pero de Cry11Aa y que presentan una mayor variabilidad genética en deleciones, inserciones y sustituciones en el dominio III en relación con las regiones que codifican para los demás dominios (Florez et al., 2016). Las variaciones termodinámicas obtenidas a nivel del ADN son similares entre aquellos genes que están más relacionados filogenéticamente. Para saber si este mismo comportamiento se presentaba en otros genes relacionados al holotipo Cry11, se hicieron los mismos ensayos en secuencias génicas de *cry2Aa* y *cry18Aa* pertenecientes al clúster II. Estos dos genes codifican para dos toxinas que se encuentran filogenéticamente más cercanas al clúster I (Crickmore et al., 2016), tienen un 51,3% de identidad y la divergencia de las estructuras de los dominios II es la que otorga la especificidad de las toxinas, de tal forma que la toxina Cry2Aa interactúa con receptores de dípteros y hemípteros mientras que la toxina Cry18Aa, interactúa con receptores de coleópteros (van Frankenhuyzen, 2009). A pesar de estas diferencias, el clúster II también mostró un comportamiento termodinámico conservado entre sus genes. Las regiones que codifican para los dominios I y III se comportaron igual a *cry11Ba* y *Bb* pero todas coinciden en mostrar el mismo comportamiento termodinámico en la región del dominio I como la región que mayor tendencia tiene a formar estructuras secundarias (Figura 25). El comportamiento termodinámico encontrado en los clústeres I y II tendría una relación con la conformación de los bloques conservados y a las estructuras proteicas de los dominios II y III de las toxinas Cry, cuya diferenciación evolutiva está determinada por la selección positiva para interactuar con distintos receptores localizados en el intestino medio de los insectos (Bravo et al., 2013). Del mismo modo, lo observado termodinámicamente en la región que codifica para el dominio I, podría estar relacionado con preservar la funcionalidad del dominio encargado de la formación de poro y la oligomerización (Bravo et al., 2013). Esto concuerda con los ensayos experimentales de barajado con genes *cry11* en donde de los productos re ensamblados de

cry11Aa se presentaron únicamente deleciones de 3 a 90 amino ácidos en la región amino terminal de la toxina. Lo interesante es que ninguna de las variantes ensambladas compromete las hélices α4 y α5 involucradas en la formación del poro y todas exhiben actividad tóxica (Florez et al., 2016).

Para conocer si este mismo comportamiento se mantenía en otras toxinas, se seleccionaron genes alejados filogenéticamente de los clúster I y II, como fueron los genes *cry1Aa* y *cry1Ab*, pertenecientes al clúster III, y *cry30Aa* y *Cry30Ca* pertenecientes al clúster IV. Estos grupos se sometieron a los mismos análisis. El clúster III presentó un comportamiento distinto a los clústeres I, II y IV, y entre los genes de este clúster se observa una coincidencia con la región en el dominio I. Esta región presentó menor tendencia, a nivel termodinámico, de formar estructuras secundarias y se evidencia un patrón invertido entre ambos genes con respecto a las regiones del dominio II y III. Sin embargo, en el holotipo de Cry30, el patrón se mantiene entre los genes pero es distinto a los que se obtuvieron en los clúster I, II y III en condiciones de barajado (Figura 25). Los resultados obtenidos para Cry1 contrastan con los obtenidos para Cry11, siendo la diferencia más importante la baja espontaneidad de la región I. esto podría indicar que esta región, que codifica para el primer dominio de las proteínas Cry1Aa y Cry1Ab, es la más estable termodinámicamente y en principio menos propensa a la formación de estructuras secundarias, que en condiciones simuladas de barajado de ADN utilizadas en este estudio, podría favorecer la recombinación y por ende una mayor variabilidad genética. Sin embargo, de los estudios experimentales en toxinas Cry1 que emplean el barajado de ADN, así como metodologías combinadas y que muestran preferencia por modificaciones en el dominio III asociadas con un incremento de la actividad tóxica (Lucena et al., 2014), solo uno hace

referencia a la fragmentación previa del ADN, obteniendo pocos clones con actividad y ninguno con actividad aumentada (Knight, Broadwell, Grant, & Shoemaker, 2004). Según los autores esto se produce debido a que las toxinas no son tolerantes al intercambio de dominios ni a las mutaciones de los dominios conservados en las cuales es probable que se produzca intercambio de dominios (Knight et al., 2004).

El comportamiento conservado entre las regiones para los genes *cry30Aa y cry30Ca* (Figura 7), se relaciona con la identidad que tienen entre ellos: 78,1%. Estos genes conservan un patrón termodinámico en todos los dominios y de acuerdo a los análisis estructurales realizados en Cry30Ca2 comparten una topología estructural con Cry4Ba con mayores diferencias en el dominio II (Zhao, Zhou, & Xia, 2012). De todos modos, aún faltan estudios estructurales y de letalidad de estas toxinas que permitan contrastar los resultados encontrados. Siendo este el primer estudio en toxinas Cry, no hay datos puntuales que permitan contrastar los resultados termodinámicos encontrados en nuestro estudio con otros en términos de variabilidad genética en regiones que codifican para sus dominios durante eventos de recombinación. Sin embargo, estos hallazgos demuestran la complejidad del barajado de ADN a nivel experimental en toxinas Cry y la necesidad de diseñar modelos *in silico* que permitan estudiar variables termodinámicas asociadas a mejorar la eficiencia de ensamblados que codifiquen para proteínas funcionales.

4.3.2 Uso de SAssembly & GenE-in. Sassembly y su arquitectura software fueron expuestos con detalle en la sección 3.2. Este componente permite la creación de bibliotecas quiméricas de genes cry, generando diversidad mediante la simulación de la técnica recombinante, barajado de ADN. Por otro lado, GenE-in y su arquitectura software fueron

expuestos con detalle en la sección 3.3., y éste modulo es el encargado de seleccionar las variantes quiméricas más prometedoras en términos de eficiencia y diversidad, simulando mediante técnicas bioinformáticas los resultados del paso de tamizaje/selección de ED.

A continuación se presentan los escenarios de simulación usados con SAssembly y los criterios de selección tenidos en cuenta por GenE-in, para filtrar las mejores variantes *in silico*. Luego se presenta un resultado consolidado de las bibliotecas quiméricas obtenidas y el resultado de la selección de las mejores variantes, finalmente estos resultados en conjunto son comparados con los resultados experimentales *in vitro* llevados a cabo en nuestro grupo.

- Escenarios de simulación y resultados preliminares de SAssembly

Cada escenario de simulación se diseñó para someter a genes parentales *cry11* a condiciones experimentales de barajado de ADN. Estas condiciones incluyen: la temperatura del barajado (TE), la longitud promedio en pares de bases de los fragmentos iniciales de ADN a barajar (LE), con un número mínimo de *v* nucleótidos de solapamiento para la hibridación (v=7) (Maheshri & Schaffer, 2003); y los parámetros termodinámicos del vecino más próximo (SantaLucia, Allawi, & Seneviratne, 1996). Los diseños experimentales *in silico* variaron las condiciones de LE y TE en rangos de valores, para los cuales LE = [50pb-250pb] y TE=[48 °C-68 °C]. Se implementaron 180 escenarios de simulación de donde se obtuvieron el mismo número de bibliotecas quiméricas.

Para la obtención de bibliotecas quiméricas con SAssembly se usaron variadas asociaciones de genes parentales *cry11*, formando cuatro grupos. El primer grupo se conformó con la selección de tres genes parentales, todos los miembros de la familia *cry11* (*cry11Aa, cry11Ba y cry11Bb*), mientras que los tres grupos restantes son la asociación de dos genes de la familia *cry11* (*cry11Aa y cry11Ba; cry11Aa y cry11Bb; y cry11Ba y cry11Bb*) (Tabla 4). Cada grupo de genes fue fragmentado simulando el efecto de la DNasa I que se emplea en condiciones experimentales de barajado de ADN (Sun, 1999) y usando el módulo de fragmentación que es compartido por SANAFold y SAssembly. Se fragmentaron cerca de 405 veces los genes completos *cry11* para llevar a cabo el número de experimentos in silico de barajado aquí reportados.

Tabla 3.

Bibliotecas quiméricas in silico de genes cry11 (36 escenarios de simulación)

Condiciones de barajado			Métricas de las bibliotecas obtenidas					
Genes parentales	T °C	Fragmentación (pb)	Longitud promedio de los genes ensamblados (pb)	DE	Proporción de la Población con mayor o igual longitud que los parentales (Eficiencia)	# Genes	Longitud promedio de los mejores ORF	Identidad Promedio de los mejores ORF (Blast)
Aa_Ba_Bb	68	250	3750	567	0,97	244	278	0,93
Aa_Ba_Bb	58	250	3784	562	0,96	242	346	0,93
Aa_Ba_Bb	48	250	3785	556	0,98	242	336	0,98
Aa_Ba_Bb	68	150	2242	246	0,85	414	203	0,79
Aa_Ba_Bb	58	150	2223	239	0,81	418	223	0,90
Aa_Ba_Bb	48	150	2209	246	0,80	420	189	0,79
Aa_Ba_Bb	68	50	2687	169	0,98	1276	162	0,51
Aa_Ba_Bb	58	50	2681	167	0,98	1278	153	0,64
Aa_Ba_Bb	48	50	2673	173	0,98	1282	161	0,44
Aa_Ba	68	250	3718	663	0,96	150	265	0,98
Aa_Ba	58	250	3681	704	0,96	152	291	0,92
Aa_Ba	48	250	3637	716	0,95	156	243	0,87
Aa_Ba	68	150	2639	371	0,83	264	219	0,61
Aa_Ba	58	150	2991	388	0,88	270	184	0,90
Aa_Ba	48	150	2603	357	0,83	268	202	0,79
Aa_Ba	68	50	2638	206	0,97	828	138	0,41
Aa_Ba	58	50	2654	207	0,98	824	138	0,57
Aa_Ba	48	50	2644	208	0,96	826	136	0,34
Aa_Bb	68	250	3606	698	0,96	160	291	0,82

Aa_Bb	58	250	3622	651	0,95	158	266	0,93
Aa_Bb	48	250	3547	652	0,94	162	310	0,98
Aa_Bb	68	150	2553	348	0,82	278	203	0,90
Aa_Bb	58	150	3400	409	0,92	274	195	0,87
Aa_Bb	48	150	3429	456	0,90	274	205	0,91
Aa_Bb	68	50	2647	204	0,96	842	157	0,56
Aa_Bb	58	50	2649	208	0,97	842	157	0,45
Aa_Bb	48	50	2639	203	0,97	844	136	0,46
Ba_Bb	68	250	3712	739	0,92	166	265	0,92
Ba_Bb	58	250	3662	704	0,93	168	294	0,93
Ba_Bb	48	250	3629	604	0,94	168	280	0,91
Ba_Bb	68	150	4217	550	0,99	294	236	0,93
Ba_Bb	58	150	3852	514	0,93	289	192	0,78
Ba_Bb	48	150	4307	562	0,98	288	214	0,76
Ba_Bb	68	50	2657	194	0,97	888	134	0,15
Ba_Bb	58	50	2660	212	0,96	888	149	0,47
Ba_Bb	58	50	2668	209	0,97	886	155	0,70

Dado que el proceso de colisión de fragmentos se hace de forma cíclica (simulando ciclos iterativos de una PCR), los resultados parciales de las colisiones, después de hibridar, extenderse y desnaturalizarse, deben ser almacenados temporalmente en arreglos unidimensionales $N[i]$ para ser usados en los siguientes ciclos de colisión. En las simulaciones realizadas, donde el criterio de parada de SAssembly fue una alta eficiencia, es decir, que al menos el 80% de la población de los genes almacenados debían cumplir con una longitud igual o cercana al promedio de la longitud de los genes parentales, fueron necesarios cerca de $3.1 * 10^4$ arreglos temporales de almacenamiento $N[i]$. En éstos se usaron rangos entre [4 - 6] ciclos computacionales (ciclos de PCR simulados) para obtener 180 bibliotecas quiméricas, cinco bibliotecas por cada escenario de simulación (Tabla 4), correspondientes a 17423 genes ensamblados, donde el 93 % de ellos cumplieron con el criterio de eficiencia (Tabla 4).

- Consolidado de las mejores variantes *in silico* mediante el uso de GenE-in

GenE-in identificó que los cerca de 17423 genes ensamblados codificaron para un poco más de $4.1 * 10^5$ marcos abiertos de lectura (ORF por sus siglas en inglés), de los cuales solo una proporción de $5.25 * 10^{-5}$ cumplieron con la longitud mínima de una toxina Cry11. Este grupo de proteínas candidatas contó con una identidad promedio de 0.77 ± 0.11 respecto a toxinas Cry11. Las características de las 22 mejores variantes *in silico* de toxinas Cry11, según los criterios de diversidad y eficiencia de GenE-in, han sido descritas y reafirman la importancia del parámetro de longitud de fragmentación en los resultados de ED (Tabla 5) (Moore & Maranas, 2000; Sun, 1999).

Tabla 4.

Mejores variantes cry11 in silico

Gen	Condiciones experimentales		Parentales *cry11* Barajados	Longitud del ORF candidato (aa)	Identidad Blast
	Temperatura de alineamiento °C	Longitud de Fragmentación (pb)			
01	58	250	Aa_Ba	466	0,78
02	58	250	Aa_Ba	459	0,80
03	58	250	Aa_Ba	358	0,88
04	58	250	Aa_Ba	342	0,60
05	58	250	Aa_Ba	342	0,60
06	68	250	Aa_Ba	321	0,87
07	68	250	Aa_Ba	307	0,90
08	68	250	Aa_Ba	394	0,89
09	68	250	Aa_Ba_Bb	489	0,64
10	58	250	Aa_Ba_Bb	471	0,66
11	58	250	Aa_Ba_Bb	471	0,64
12	48	250	Aa_Ba_Bb	456	0,76
13	48	250	Aa_Ba_Bb	434	0,69
14	58	250	Aa_Ba_Bb	434	0,68
15	48	250	Aa_Ba_Bb	374	0,86
16	48	250	Aa_Ba_Bb	374	0,87
17	48	250	Aa_Bb	570	0,81
18	68	250	Aa_Bb	423	0,66
19	58	250	Aa_Bb	319	0,76

20	58	250	Ba_Bb	371	0,97
21	48	250	Ba_Bb	356	0,78
22	68	250	Ba_Bb	306	0,90

- Descripción de resultados de SAssembly & GenE-in

  - Primer grupo de análisis

El primer grupo de análisis incluyó la recombinación de los genes *cry11Aa, cry11Ba y cry11Bb*, generando 45 librerías quiméricas, para un total de 5816 genes ensamblados, con una longitud promedio por gen de 2893 pb (Tabla 4). Este primer grupo generó el 36.3 % de los mejores ORF con una longitud promedio de 374 aa (Tabla 5) y una identidad del 79%. Se presentó la mayor eficiencia para el ensamblado en condiciones experimentales con temperaturas de 48°C, 58 °C y fragmentos de ADN con longitudes iniciales cercanas a los 250 pb.

  - Segundo grupo de análisis

El segundo grupo de análisis incluyó la recombinación de los genes *cry11Aa*, *cry11Ba*, generando 45 librerías quiméricas, para un total de 3738 genes ensamblados, con una longitud promedio por gen de 3023 pb (Tabla 4). Este segundo grupo generó un 36.3 % de los mejores ORF con una longitud promedio de 438 aa (Tabla 5) y una identidad del 72.5%. Se presentó la mayor eficiencia para el ensamblado en condiciones experimentales con temperaturas de 58 °C, 68 °C y fragmentos de ADN con longitudes iniciales cercanas a los 250 pb.

- Tercer grupo de análisis

El tercer grupo de análisis incluyó la recombinación de los genes *cry11Aa, cry11Bb*, generando 45 librerías quiméricas, para un total de 3834 genes ensamblados, con una longitud promedio por gen de 3121 pb (Tabla 4). Este tercer grupo generó un 13.6 % de los mejores ORF con una longitud promedio de 437 aa (Tabla 5) y una identidad del 74.3%. Se presentó la mayor eficiencia para el ensamblado en condiciones experimentales con temperaturas de 48°C, 58 °C, 68 °C y fragmentos de ADN con longitudes iniciales cercanas a los 250 pb.

- Cuarto grupo de análisis

El cuarto grupo de análisis incluyó la recombinación de los genes *cry11Ba, cry11Bb*, generando 45 librerías quiméricas, para un total de 4035 genes ensamblados, con una longitud promedio por gen de 3485 pb (Tabla 4). Este cuarto grupo generó un 13.6 % de los mejores ORF con una longitud promedio de 344 aa (Tabla 5) y una identidad del 88.3%. Se presentó la mayor eficiencia para el ensamblado en condiciones experimentales con temperaturas de 48°C, 58 °C, 68 °C y fragmentos de ADN con longitudes iniciales cercanas a los 250 pb.

- Validación de hallazgos con experimentos *in vitro* de barajado de genes *cry11*

Los experimentos de evolución dirigida pueden ser medidos a partir de los indicadores de eficiencia y diversidad de las librerías obtenidas (Maheshri & Schaffer, 2003; Moore & Maranas, 2000; Volkov & Arnold, 2000; Volkov, Shao, & Arnold, 2000). La eficiencia de un barajado de

ADN es la proporción de la población de genes de la librería con una longitud cercana a la longitud promedio de los genes parentales usados (Volkov & Arnold, 2000; Volkov et al., 2000). La diversidad es producida por la presencia de entrecruzamientos y de mutaciones puntuales o inserciones (Maheshri & Schaffer, 2003; Moore & Maranas, 2000), que corresponde a la proporción de mutaciones de los genes ensamblados.

Los experimentos de barajado de ADN con genes *cry11* realizados por nuestro grupo permitieron la obtención de 94 variantes de las cuales: 34 variantes [< 1 Kb], 14 variantes [1-2 kb], 22 variantes [> 2.1 Kb], 14 variantes no presentaron homología con *cry11* y las 10 variantes restantes sin insertos (Florez et al., 2016). De las 74 variantes con insertos se logró una eficiencia cercana al 31% (Figura 22) con longitudes de las variantes [>2.1 Kb] y los resultados logrados con SAssembly presentaron 22 variantes candidatas con una eficiencia del 25 % con variantes [1.38 Kb – 1.71 Kb] (Tabla 5, Figura 26). La población inicial que presentó SAssembly mostro una eficiencia en el ensamblado de los genes del 93.36% (Tabla 4), sin embargo, la gran utilidad de la implementación de SAssembly es que permite obtener las secuencias de genes ensamblados de donde se infiere un eficiencia del 25% a partir de variantes codificantes para toxinas Cry11, un valor que corresponde más con los resultados *in vitro* (Figuras 22, 26).

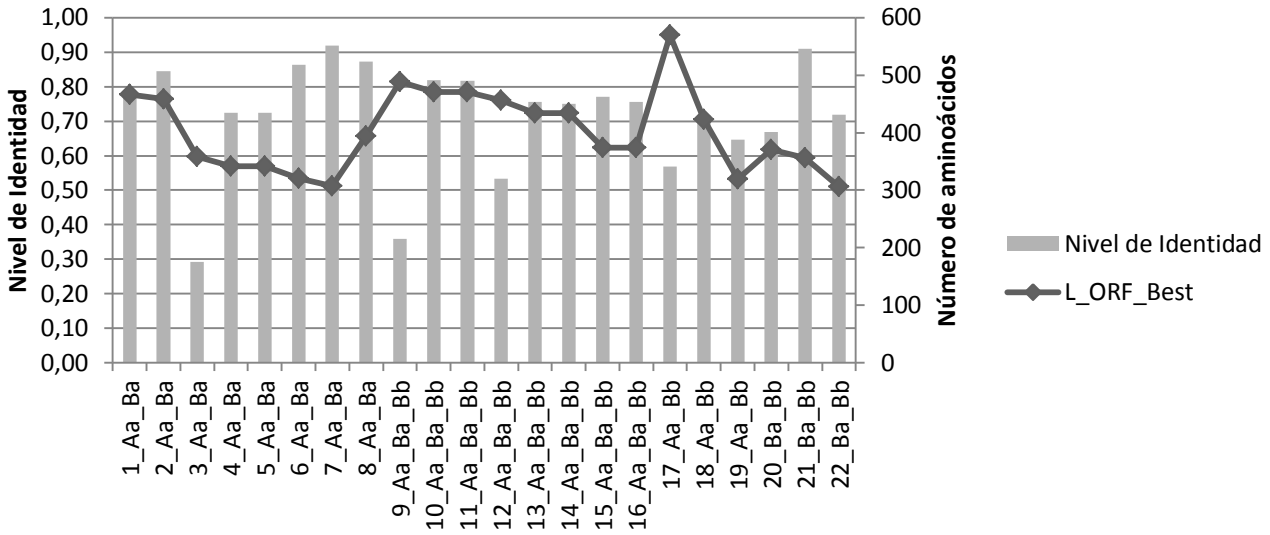


Figura 26. Diversidad y Eficiencia de las mejores variantes cry11 in silico de barajado de ADN

En cuanto a diversidad, las 70 variantes con insertos obtenidas *in vitro* no mostraron entrecruzamientos sino deleciones, inserciones y mutaciones puntuales (Morales, 2011), sin embargo, posibles entrecruzamientos silenciosos no se descartan en el proceso de barajado de ADN (Morales, 2011). Estudios de toxicidad asociada a cinco variantes: 1, 8, 23, 79 y 81; fueron realizados por nuestro grupo y se observó que la tasa de mutación de las variantes fue de 15, 13, 1, 20 y 8% respectivamente. Las tasas de mutación bajas 1 y 8% de las variantes 23 y 81, reportaron la mayor pérdida de actividad toxica (6.8 y 11.6 veces) con respecto al parental Cry11Aa; las variantes 1 y 79 con mayor tasa de mutación, 15 y 20 % respectivamente, presentaron toxicidad similar a la parental, mientras que la variante 8 con una tasa de mutación de 13% mostró un incremento de hasta 6.09 veces en su toxicidad con referencia al parental (Suarez, 2016). Esto parece indicar que tasas de mutación moderadas [13% - 20%] en variantes de Cry11 son candidatas para mantener o aumentar la toxicidad con respecto a los parentales. Las 22 variantes *in silico* presentan una proporción de identidad 0.77 ± 0.11 (Tabla 5, Figura 26), equivalente a una tasa de mutación de 22.7 ± 10.9%, que concuerda con la diversidad de las

variantes *in vitro* hasta ahora estudiadas. Estos hallazgos sugieren que SAssembly, en términos de eficiencia y diversidad, es una buena aproximación para simular experimentos de barajado de ADN de genes *cry11*.

En cuanto a la conformación de las secuencias de las variantes obtenidas *in vitro*, se pudo observar que cerca del 49 % se conformaban por inserciones del dominio III de genes *cry11B* y el 31% de las variantes completas [>2.1 Kb] presentaron mayor variabilidad en la región III que codifica para el dominio III de las toxinas Cry11 (Figura 22). Estos hallazgos son correspondientes con la conformación de las variantes *in silico* que presenta mayor homología en su dominio III, equivalente a un 44% con respecto a los demás dominios (I y II), proveniente en un 65% de *cry11B* (Tabla 6, Figura 27).

Tabla 5.

Bloques conservados de las mejores variantes cry11 in silico (Barajado de ADN)

Genes	**Gen ensamblado**	**Longitud de bloques parentales conservados (aa) por dominio**			**Longitud de bloques parentales conservados (aa) por parental**		
		DI	DII	DIII	Cry11Aa	Cry11Ba	Cry11Bb
cry11Aa; cry11Ba; cry11Bb	9_Aa_Ba_Bb	20 aa, Cry11Bb	70 aa, Cry11Bb				90 aa
	10_Aa_Ba_Bb	90 aa, Cry11Bb		130 aa, Cry11Bb			220 aa
	11_Aa_Ba_Bb	90 aa, Cry11Bb		130 aa, Cry11Bb			220 aa
	12_Aa_Ba_Bb	90 aa, Cry11Aa 80 aa, Cry11Bb		90 aa, Cry11Bb 40 aa, Cry11Bb	90 aa		210 aa
	13_Aa_Ba_Bb		80 aa, Cry11Bb	20 aa, Cry11Bb			100 aa
	14_Aa_Ba_Bb		80 aa, Cry11Bb	20 aa, Cry11Bb			100 aa
	15_Aa_Ba_Bb	90 aa_Cry11Aa	90 aa_Cry11Ba	130 aa_Cry11Aa	220 aa	90 aa	
	16_Aa_Ba_Bb	90 aa_Cry11Aa	90 aa_Cry11Ba	130 aa_Cry11Aa	220 aa	90 aa	

cry11Aa; cry11Ba;	1_Aa_Ba	70 aa_Cry11Ba	40 aa_Cry11Ba 10 aa, Cry11Aa	160 aa_Cry11Aa	170 aa	110 aa	
	2_Aa_Ba	100 aa_Cry11Ba		160 aa_Cry11Aa	160 aa	100 aa	
	3_Aa_Ba	80 aa, Cry11Ba	90 aa, Cry11Aa		90 aa	80 aa	
	4_Aa_Ba	20 aa, Cry11Ba	10 aa, Cry11Ba 80 aa, Cry11Aa 80 aa, Cry11Ba	20 aa, Cry11Ba	80 aa	130 aa	
	5_Aa_Ba	20 aa, Cry11Ba	10 aa, Cry11Ba 80 aa, Cry11Aa 80 aa, Cry11Ba	20 aa, Cry11Ba	80 aa	130 aa	
	6_Aa_Ba	30 aa, Cry11Aa	50 aa, Cry11Aa 80 aa, Cry11Ba		80 aa	80 aa	
	7_Aa_Ba	80 aa, Cry11Aa		90 aa, Cry11Ba	80 aa	90 aa	
	8_Aa_Ba	90 aa_Cry11Ba		150 aa_Cry11Aa	150 aa	90 aa	
cry11Aa; cry11Bb	17_Aa_Bb	90 aa, Cry11Bb		160 aa, Cry11Aa	160 aa		90 aa
	18_Aa_Bb		90 aa, Cry11Bb	160 aa, Cry11Aa 110 aa, Cry11Bb	160 aa		200 aa
	19_Aa_Bb	30 aa, Cry11Aa	60 aa, Cry11Aa	160 aa, Cry11Bb	90 aa		160 aa
cry11Ba; cry11Bb	20_Ba_Bb	20 aa, Cry11Bb	60 aa, Cry11Bb				80 aa
	21_Ba_Bb	80 aa, Cry11Bb		90 aa, Cry11Ba		90 aa	80 aa
	22_Ba_Bb	20 aa, Cry11Ba	60 aa, Cry11Ba 80 aa, Cry11Ba	80 aa, Cry11Bb		160 aa	80 aa

Esto puede explicarse desde la baja espontaneidad termodinámica para la formación de estructuras secundarias de ADN de la región III de los genes *cry11B*, de tal forma que esta región favorece las nucleaciones de sus fragmentos en experimentos de barajado, prevaleciendo su participación en la conformación de variantes *cry11* (Figura 22).

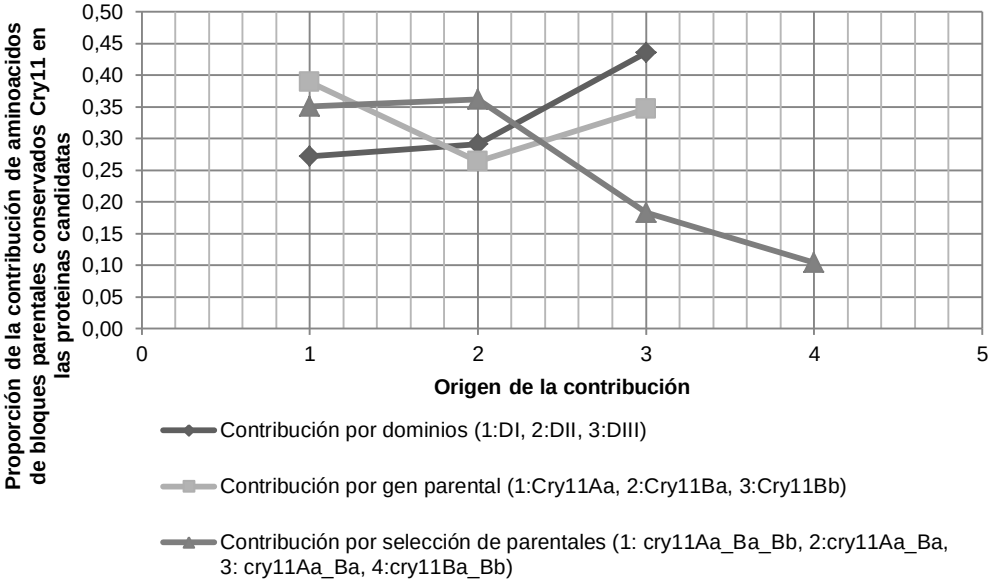


Figura 27. Bloques conservados en la conformación de las mejores variantes cry11 in silico

Por otra parte, las variantes *in vitro* completas [> 2.1 Kb], que fueron cercanas al 31%, presentaron alta homología con el parental *cry11Aa*, de igual forma las variantes *in silico* obtenidas por SAssembly mostraron una homología cercana al 39 % con respecto a Cry11Aa y un aumento de 2 y 3.6 veces de formación de variantes cuando *cry11Aa* es barajado con cry11Ba que con respecto el barajado de *cry11Aa con cry11Bb* y *cry11Ba con cry11Bb* respectivamente (Tabla 6, Figura 27). Por tanto, existe una correspondencia en los hallazgos *in silico* e *in vitro* en donde las variantes completas presentan una prevalencia de la homología de los ensamblados con el parental *cry11Aa*. Esto puede deberse a que Cry11Aa tiene un comportamiento termodinámico para la formación de estructuras secundarias de ADN en condiciones de barajado de ADN diferenciado de las *cry11B* que le hacen de alguna forma prevalecer en el ensamblado (Figura 22).

Finalmente la revisión de las 36 condiciones experimentales a las cuales fueron sometidos los genes parentales *in silico* mediante SAssembly, permiten observar un patrón en la formación de las mejores variantes según los genes parentales *cry11* barajados (Figura 28). Este patrón está asociado a las condiciones experimentales: temperatura de alineamiento y longitud de los fragmentos iniciales de los genes parentales. Una estrategia sugerida para potenciar la formación de bibliotecas obtenidas por experimentos de barajado de ADN consiste en disminuir el tamaño de los fragmentos iniciales de ADN y la temperatura de alineamiento (Moore & Maranas, 2000).

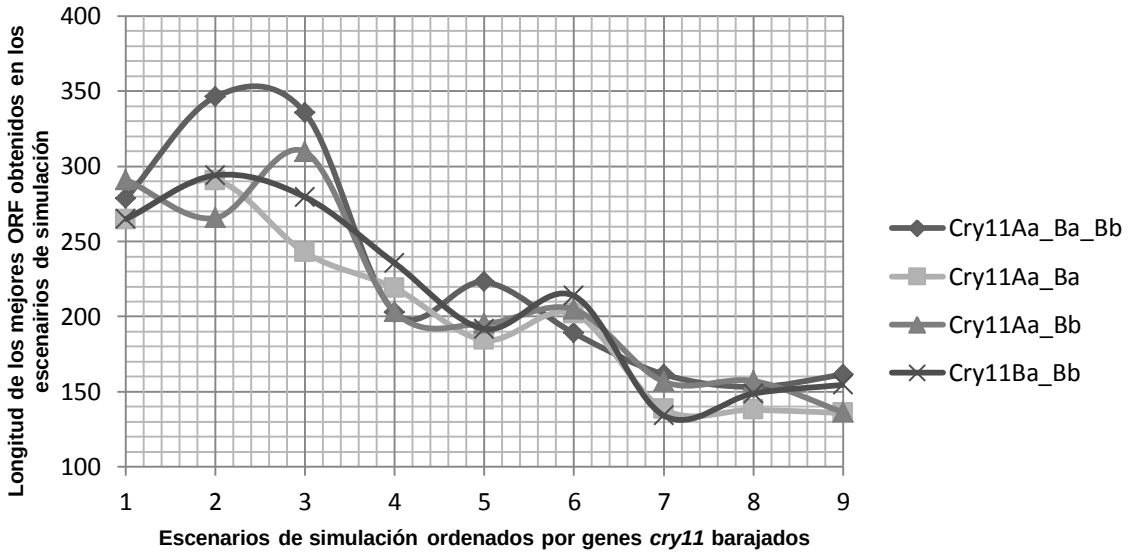


Figura 28. Escenarios de simulación in silico vs. Longitud de las mejores variantes Cry11 in silico.

Revisando la relación que existe en los resultados *in silico* entre la temperatura de alineamiento y la longitud de las mejores variantes en los 36 escenarios de simulación, se encontró un patrón de

decrecimiento de la longitud de los ORF inversamente proporcional a la longitud de los fragmentos sin relación a la temepratura asociada (Figura 29). Este patrón se confirma, al realizar el mismo análisis dependiente con la longitud de los fragmentos iniciales de ADN, se encontró una correlación moderada r = 0.8694 (Figura 30), esto indica que en la medida que se aumenta la longitud de los fragmentos iniciales de ADN [50 pb, 150 pb, 250 pb] aumenta la longitud de los ORF de las mejores variantes (Figura 30).

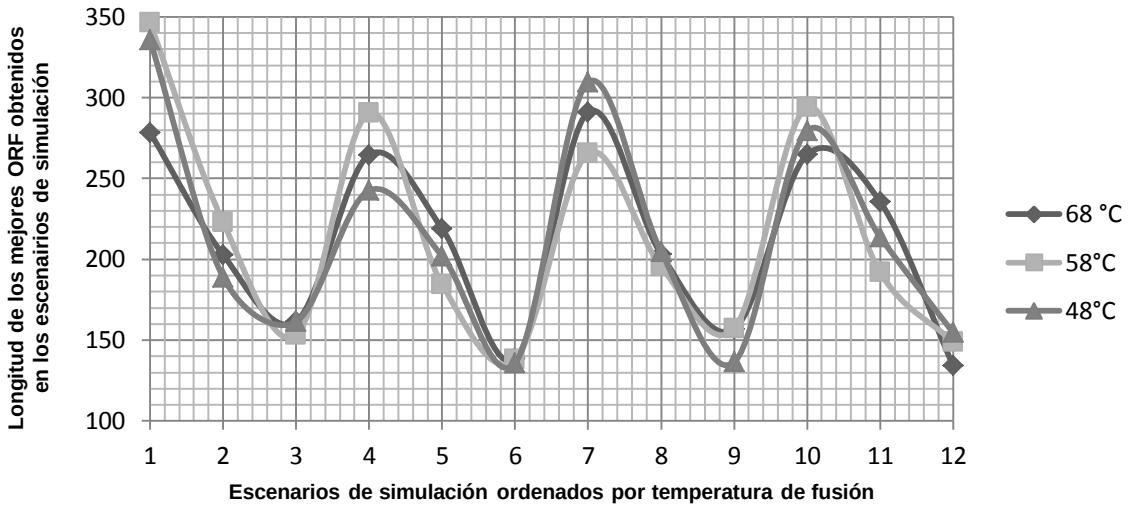


Figura 29. Temperatura de alineamiento vs. Longitud de las mejores variantes Cry11 in silico.

Esto podría indicar que es de gran importancia en los experimentos de barajado la longitud inicial de los fragmentos, como un factor determinante en la formación de variantes. Esta diferencia podría estar asociada con que a mayor tamaño del fragmento inicial hay una mayor posibilidad de ocurrir un efecto negativo en el ensamblado de las estructuras secundarias de ADN, porque es menos favorable energéticamente la formación de estructuras secundarias de ADN (SantaLucia & Hicks, 2004) que iniciar los procesos de nucleación que se dan con al

menos 7 nucleótidos complementados (Maheshri & Schaffer, 2003), favoreciendo desde el inicio el ensamblado (SantaLucia, 1998; SantaLucia et al., 1996; SantaLucia & Hicks, 2004). Estas condiciones experimentales permitieron la obtención de variantes *in vitro*, con un promedio de 6 entrecruzamientos (Morales, 2011) y variantes *in silico* con un promedio de 4.8 ciclos computacionales de ensamblado, mostrando una cercanía en la diversidad de las variantes obtenidas por medio de ambas estrategias (*in vitro e in silico*).

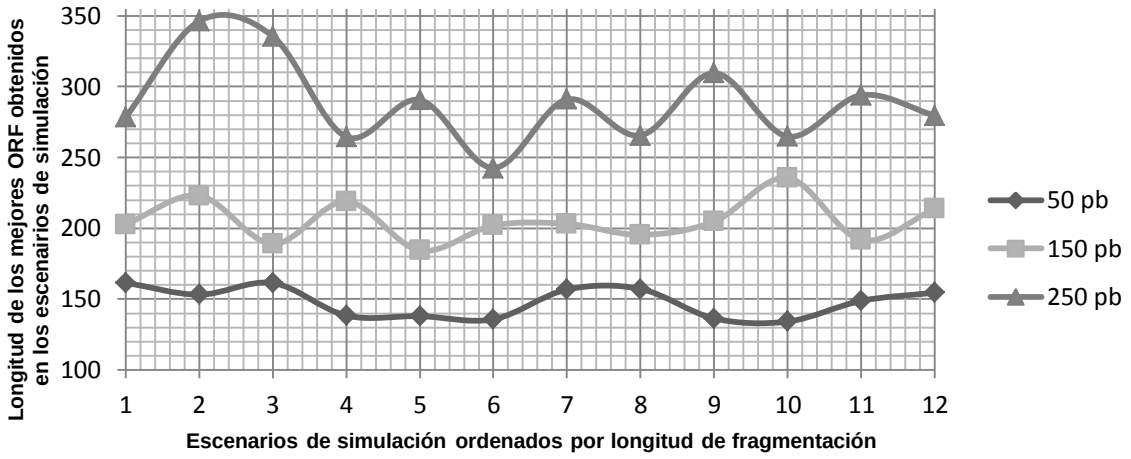


Figura 30. Longitud inicial de fragmentación vs. Longitud de las mejores variantes Cry11 in silico.

4.4 Conclusiones

DEvISING mostró ser una estrategia *in silico* útil para reproducir las condiciones experimentales de ED. En el presente capítulo se expuso: Selección de los genes parentales más favorables para barajar, generación de bibliotecas quiméricas *in silico* mediante la simulación del barajado de ADN y filtrado de las mejores variantes a partir de criterios de eficiencia y diversidad.

Consideramos que DEvISING es un trabajo novedoso, siendo el pionero en integrar métodos estadísticos, de energía libre, técnicas bioinformáticas y procesos estocásticos en una arquitectura computacional cluster, que permite la obtención de variantes *in silico,* en este caso para el estudio de toxinas Cry11.

Estos componentes software desarrollados brindaron otros hallazgos, relativos a la comprensión de fenómenos de interés biotecnológico en el campo de las toxinas Cry. Mediante la ejecución del componente SANAFold se lograron identificar patrones de comportamiento conservado en los diferentes clústeres analizados. Estos patrones se pudieron asociar a la cercanía filogenética de los genes, lo que sugiere que los patrones encontrados obedecen a condiciones intrínsecas de la secuencia de los genes, que se conservan evolutivamente y que se comportan diferenciadamente cuando están bajo condiciones de barajado de ADN. Este hallazgo, además de ser útil para certificar la coherencia de nuestra propuesta de selección parental, presenta nuevas oportunidades de exploración de los genes *cry* con múltiples propósitos.

Con los componentes de DEVISING se lograron articular modelos de selección de parentales y de generación de diversidad, usando el método de mínima energía, siendo los primeros en incorporar en un modelo de ED los efectos de la formación de estructuras secundarias de ADN en la generación de bibliotecas quiméricas con técnicas ADN recombinantes.

Ningún trabajo predecesor había logrado reproducir *in silico* el proceso completo de ED. Hemos optimizado la técnica de obtención de variantes de barajado de ADN mediante el

modelado in silico y con el uso de estos desarrollo hemos disminuido considerablemente los costos experimentales (Anexo 2).

Finalmente, debemos resaltar que todos los experimentos *in silico* que aquí se presentaron son reproducibles, y sus productos susceptibles de verificación mediante la secuenciación y trasformación *in vivo* de las mejores variantes obtenidas mediante nuestras simulaciones, como material biológico que podrá ser evaluado *in vitro* en cuanto a funcionalidad.

5. Conclusiones Generales

Esta tesis doctoral partió del supuesto que era posible el desarrollo de una estrategia *in silico* que contemplara los elementos derivados de la composición del ADN y permitierá el diseño de ED con genes *cry*. Al realizar la revisión de la literatura científica en el campo de estudio, identificamos que las aproximaciones *in silico* de ED tenían dos orientaciones: la primera, limitada a la búsqueda de métricas muy generales que probablemente mejorarían la eficiencia de las bibliotecas, pero sin contemplar los elementos derivados de la composición del ADN; y la segunda, enfocada al desarrollo de herramientas CAPDE (*Computer–aided protein directed evolution*) que solo son útiles cuando ya se cuenta con la biblioteca quimérica resultante, desatendiendo la generación de diversidad *in silico*.

En la búsqueda por incorporar elementos de la composición del ADN en un modelo *in silico* de ED, estudiamos las aproximaciones de generación de diversidad, basados en modelos ADN recombinantes, que mediante métodos de mínima energía lograron estudiar los efectos de las condiciones experimentales en términos de eficiencia y diversidad.

Se ha integrado mediante componentes software una estrategia in *silico* novedosa que hemos denominado DEvISING. Nuestro trabajo logró incluir elementos derivados de la composición del ADN, expresados en el comportamiento termodinámico de los genes y que hemos usado como criterio para: seleccionar los genes parentales, generar mecanismos de selectividad de hibridación, e inferir resultados de eficiencia y diversidad de genes ensamblados.

A partir de nuestros avances desde el campo de estudio de la ED *in silico*, nuestro software posibilita la exploración de dos nuevas líneas de trabajo. En primer lugar, permitirá la fabricación de variantes mejoradas *in silico*; y en segundo lugar, permitirá la identificación de patrones de comportamiento termodinámico diferenciado, asociado a la formación de estructuras secundarias de ADN que podrán ser útiles en trabajos futuros para establecer relaciones evolutivas en familias de interés biotecnológico.

En cuanto a nuestro aporte al campo de estudio de las toxinas Cry (nuestro modelo biológico), logramos identificar patrones termodinámicos en familias de toxinas Cry, útiles para el diseño de experimentos de ED. Actualmente, SANAFold es usado para la caracterización de toxinas Cry con 3 dominios conservados, a la que pertenecen más de 74 familias de toxinas Cry hasta ahora reportadas. Estos resultados serán útiles para quienes trabajan en el desarrollo de nuevos productos biotecnológicos (bio-pesticidas) a partir de estas toxinas.

Podemos concluir que los retos asumidos por esta tesis doctoral fueron logrados, obteniendo un desarrollo computacional que apalancada por las actuales tecnologías de síntesis de ADN, constituye una alternativa prometedora para hacer ED.

6. Divulgaciones y Trabajos Dirigidos

Conferencia: "Algorithmic analysis of a model *in silico* DNA Shuffling for implementation under GPU Architecture". VI Conferencia Latinoamericana de Computación de Alto Rendimiento, CLCAR. San José, Costa Rica, 28-30 de Agosto 2013. Autores: **Efraín Hernando Pinzón Reyes**[1], [2]; Daniel Alfonso Sierra Bueno[1]; Raúl Ramos Pollan[1]; Álvaro Mauricio Flórez Escobar [2]. Afiliación: [1]Department of physical-mechanical science, Electrical electronics and telecommunications engineering, Universidad Industrial de Santander, Bucaramanga, Santander, Colombia. [2] Department of Health, Medicine, Laboratory of Molecular Biology and Biotechnology, Universidad de Santander, Bucaramanga, Santander, Colombia.

Poster: "Formation of secondary structures of *cry11* genes from *Bacillus thuringiensis* in DNA shuffling conditions: an experimental and computational approach". VI International Conference on Environmental, Industrial and Applied Microbiology - BioMicroWorld2015, Barcelona (Spain), 28-30 October 2015. Autores: **Efraín Hernando Pinzón Reyes**[1], [2]; Miguel Orlando Suarez Barrera[2], [3]; Karen Lizeth Rivera Rivera [2]; Daniel Alfonso Sierra Bueno[1]; Sergio Orduz Peralta [4]; Álvaro Mauricio Flórez Escobar [2]* (corresponding author). Afiliación: [1] Department of physical-mechanical science, Electrical electronics and telecommunications engineering, Universidad Industrial de Santander, Bucaramanga, Santander, Colombia. [2] Department of Health, Medicine, Laboratory of Molecular Biology and Biotechnology, Universidad de Santander, Bucaramanga, Santander, Colombia. [3] Department of Health, Bacteriology Faculty, Universidad Industrial de Santander, Bucaramanga, Santander, Colombia.

[4] Department of Ecology and Systematic of Insects; Universidad Nacional de Colombia Sede Medellin.* amflorez@udes.edu.co

Artículo aprobado para publicación (BMC Biophysics): "DNA secondary structure formation by DNA shuffling of the conserved domains of the Cry protein of *Bacillus thuringiensis.* Autores: **Efraín Hernando Pinzón Reyes**[1], [2]; Daniel Alfonso Sierra Bueno[1]; Miguel Orlando Suarez Barrera[2], Sergio Orduz Peralta [3]; Álvaro Mauricio Flórez Escobar [2]* (corresponding author). Afiliación: [1]Department of physical-mechanical science, Electrical electronics and telecommunications engineering, Universidad Industrial de Santander, Bucaramanga, Santander, Colombia. [2] Department of Health, Medicine, Laboratory of Molecular Biology and Biotechnology, Universidad de Santander, Bucaramanga, Santander, Colombia. [3]Department of Ecology and Systematic of Insects; Universidad Nacional de Colombia Sede Medellin.* amflorez@udes.edu.co. 22 de Mayo de 2017. doi: 10.1186/s13628-017-0036-7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5441083/

Trabajo de grado: "Modelo in silico del barajado genético basado en leyes cinéticas y termodinámicos biomoleculares". Autores: Andrés Felipe Guerrero Franco, Cristhian Fabián Ariza Jiménez. Director: **Efraín Hernando Pinzón Reyes**, Facultad de Ingeniería, Escuela de Ingeniería de Sistemas. Universidad de Santander- UDES. 2014.

Trabajo de grado: "Modelado y simulación de la formación de estructuras secundarias genómicas a partir de sus leyes termodinámicas". Autores: María Julieth Acevedo Suarez,

Dimelsa Andrea Cáceres Cáceres. Director: **Efraín Hernando Pinzón Reyes**, Facultad de Ingeniería, Escuela de Ingeniería de Sistemas. Universidad de Santander-UDES. 2014.

Referencias Bibliograficas

Abdullah, M. (2012). Use and Efficacy of Bt Compared to Less Environmentally Safe Alternatives *Bacillus thuringiensis Biothecnology* (pp. 87-92): Springer.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25*(17), 3389-3402. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9254694

Alzate, O., Osorio, C., Florez, A. M., & Dean, D. H. (2010). Participation of valine 171 in alpha-Helix 5 of Bacillus thuringiensis Cry1Ab delta-endotoxin in translocation of toxin into Lymantria dispar midgut membranes. *Appl Environ Microbiol, 76*(23), 7878-7880. doi:10.1128/AEM.01428-10

Arnold, K., Kiefer, F., Kopp, J., Battey, J. N., Podvinec, M., Westbrook, J. D., . . . Schwede, T. (2009). The Protein Model Portal. *J Struct Funct Genomics, 10*(1), 1-8. doi:10.1007/s10969-008-9048-5

Bikard, D., Loot, C., Baharoglu, Z., & Mazel, D. (2010). Folded DNA in action: hairpin formation and biological functions in prokaryotes. *Microbiol Mol Biol Rev, 74*(4), 570-588. doi:10.1128/MMBR.00026-10

Bravo, A., Gomez, I., Porta, H., Garcia-Gomez, B. I., Rodriguez-Almazan, C., Pardo, L., & Soberon, M. (2013). Evolution of Bacillus thuringiensis Cry toxins insecticidal activity. *Microb Biotechnol, 6*(1), 17-26. doi:10.1111/j.1751-7915.2012.00342.x

Campanella, J. J., Bitincka, L., & Smalley, J. (2003). MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences. *BMC Bioinformatics, 4*, 29. doi:10.1186/1471-2105-4-29

Cobb, R. E., Si, T., & Zhao, H. (2012). Directed evolution: an evolving and enabling synthetic biology tool. *Curr Opin Chem Biol, 16*(3-4), 285-291. doi:10.1016/j.cbpa.2012.05.186

Cobb, R. E., Sun, N., & Zhao, H. (2013). Directed evolution as a powerful synthetic biology tool. *Methods, 60*(1), 81-90. doi:10.1016/j.ymeth.2012.03.009

Craveiro, K. I., Gomes Junior, J. E., Silva, M. C., Macedo, L. L., Lucena, W. A., Silva, M. S., . . . Grossi-de-Sa, M. F. (2010). Variant Cry1Ia toxins generated by DNA shuffling are active against sugarcane giant borer. *J Biotechnol, 145*(3), 215-221. doi:10.1016/j.jbiotec.2009.11.011

Crickmore, N. , Baum, J., Bravo, A., Lereclus, D., Narva, K., Sampson, K., . . . Zeigler, D.R. . (2016). Bacillus thuringiensis toxin nomenclature.

Currin, A., Swainston, N., Day, P. J., & Kell, D. B. (2015). Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem Soc Rev, 44*(5), 1172-1239. doi:10.1039/c4cs00351a

de Maagd, R. A., Bravo, A., & Crickmore, N. (2001). How Bacillus thuringiensis has evolved specific toxins to colonize the insect world. *Trends Genet, 17*(4), 193-199. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11275324

Flórez, A. M.; Osorio, C.; Alzate, O. (2012). Protein engineering of Bacillus thuringiensis δ-endotoxins. *Bacillus thuringiensis Biotechnology* (pp. 19-39): Springer.

Florez, A.M., Suarez-Barrera, M.O., Morales, G.M., Rivera, K.V., Orduz, S., Ochoa, R., . . . Muskus, C. . (2016). *Toxic activity, molecular modeling and docking simulations for Cry11variants from*

Bacillus thuringiensis obtained from DNA shuffling.

Fox, R. (2005). Directed molecular evolution by machine learning and the influence of nonlinear interactions. *J Theor Biol, 234*(2), 187-199. doi:10.1016/j.jtbi.2004.11.031

Fox, R., Roy, A., Govindarajan, S., Minshull, J., Gustafsson, C., Jones, J. T., & Emig, R. (2003). Optimizing the search algorithm for protein engineering by directed evolution. *Protein Eng, 16*(8), 589-597. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12968076

George, Z; Crickmore, N (2012). Bacillus thuringiensis Applications in Agriculture *Bacillus thuringiensis Biothecnology* (pp. 19-39): Springer.

He, L., Friedman, A. M., & Bailey-Kellogg, C. (2012). Algorithms for optimizing cross-overs in DNA shuffling. *BMC Bioinformatics, 13 Suppl 3*, S3. doi:10.1186/1471-2105-13-S3-S3

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res, 31*(13), 3429-3431. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12824340

Hussain, S; Flórez, A; Dean, D; Alzate, O. (2010). Preferentially Protection of Domains II and III of Bacillus thuringiensis Cry1Aa Toxin by Brush Border Membrane Vesicles. *Revista Colombiana Biotecnología, 12*, 14-26.

Hussain, S; Flórez, A; Dean, D; Alzate, O. (2011). Characterization of a Mutant Bacillus thuringiensis δ-endotoxin With Enhanced Stability and Toxicity. *Revista Colombiana Biotecnología, 13*, 144-154.

Joern, J. M., Meinhold, P., & Arnold, F. H. (2002). Analysis of shuffled gene libraries. *J Mol Biol, 316*(3), 643-656. doi:10.1006/jmbi.2001.5349

Knight, J. S., Broadwell, A. H., Grant, W. N., & Shoemaker, C. B. (2004). A strategy for shuffling numerous Bacillus thuringiensis crystal protein domains. *J Econ Entomol, 97*(6), 1805-1813. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15666731

Koehler, R. T., & Peyret, N. (2005). Effects of DNA secondary structure on oligonucleotide probe binding efficiency. *Comput Biol Chem, 29*(6), 393-397. doi:10.1016/j.compbiolchem.2005.09.002

Lane, M. D., & Seelig, B. (2014). Advances in the directed evolution of proteins. *Curr Opin Chem Biol, 22*, 129-136. doi:10.1016/j.cbpa.2014.09.013

Lassner, M., & Bedbrook, J. (2001). Directed molecular evolution in plant improvement. *Curr Opin Plant Biol, 4*(2), 152-156. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11228439

Leemhuis, H., Kelly, R. M., & Dijkhuizen, L. (2009). Directed evolution of enzymes: Library screening strategies. *IUBMB Life, 61*(3), 222-228. doi:10.1002/iub.165

Lu, M., Guo, Q., Marky, L. A., Seeman, N. C., & Kallenbach, N. R. (1992). Thermodynamics of DNA branching. *J Mol Biol, 223*(3), 781-789. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1542118

Lucena, W. A., Pelegrini, P. B., Martins-de-Sa, D., Fonseca, F. C., Jr, J. E., de Macedo, L. L., . . . Grossi-de-Sa, M. F. (2014). Molecular Approaches to Improve the Insecticidal Activity of Bacillus thuringiensis Cry Toxins. *Toxins (Basel), 6*(8), 2393-2423. doi:10.3390/toxins6082393

Mahadeva, H.M.; Asokan, R.;Rajasekaran, P.E.;Mahmood, R.; Nagesha, S.N.; Arora, D.K. (2012). Analysis of Opportunities and Challenges in Patenting of Bacillus thuringiensis

Insecticidal Crystal Protein Genes. *Recent Patents on DNA & Gene Sequences.* (Vol. 6, pp. 64-71).

Maheshri, N., & Schaffer, D. V. (2003). Computational and experimental analysis of DNA shuffling. *Proc Natl Acad Sci U S A, 100*(6), 3071-3076. doi:10.1073/pnas.0537968100

Markham, N. R., & Zuker, M. (2008). UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol, 453*, 3-31. doi:10.1007/978-1-60327-429-6_1

Moore, G. L., & Maranas, C. D. (2000). Modeling DNA mutation and recombination for directed evolution experiments. *J Theor Biol, 205*(3), 483-503. doi:10.1006/jtbi.2000.2082

Moore, G. L., & Maranas, C. D. (2002). eCodonOpt: a systematic computational framework for optimizing codon usage in directed evolution experiments. *Nucleic Acids Res, 30*(11), 2407-2416. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12034828

Morales, M. (2011). *Análisis experimental y computacional de genes cry11 de Bacillus thuringiensis mediante evolución dirigida.* (Maestria.), Universidad de Antioquia.

Muhire, B. M., Golden, M., Murrell, B., Lefeuvre, P., Lett, J. M., Gray, A., . . . Martin, D. P. (2014). Evidence of pervasive biologically functional secondary structures within the genomes of eukaryotic single-stranded DNA viruses. *J Virol, 88*(4), 1972-1989. doi:10.1128/JVI.03031-13

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol, 51 Pt 1*, 263-273. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3472723

Nazarenko, I., Pires, R., Lowe, B., Obaidy, M., & Rashtchian, A. (2002). Effect of primary and secondary structure of oligodeoxyribonucleotides on the fluorescent properties of

conjugated dyes. *Nucleic Acids Res, 30*(9), 2089-2195. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11972350

Orduz, S., Realpe, M., Arango, R., Murillo, L. A., & Delecluse, A. (1998). Sequence of the cry11Bb11 gene from Bacillus thuringiensis subsp. medellin and toxicity analysis of its encoded protein. *Biochim Biophys Acta, 1388*(1), 267-272. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9774743

Packer, M. S., & Liu, D. R. (2015). Methods for the directed evolution of proteins. *Nat Rev Genet, 16*(7), 379-394. doi:10.1038/nrg3927

Patrick, W. M., & Firth, A. E. (2005). Strategies and computational tools for improving randomized protein libraries. *Biomol Eng, 22*(4), 105-112. doi:10.1016/j.bioeng.2005.06.001

Patrick, W. M., Firth, A. E., & Blackburn, J. M. (2003). User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Eng, 16*(6), 451-457. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12874379

Peng, W., Levine, H., Hwa, T., & Kessler, D. A. (2004). Analytical study of the effect of recombination on evolution via DNA shuffling. *Phys Rev E Stat Nonlin Soft Matter Phys, 69*(5 Pt 1), 051911. doi:10.1103/PhysRevE.69.051911

Romero, P. A., & Arnold, F. H. (2009). Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol, 10*(12), 866-876. doi:10.1038/nrm2805

Rychlik, W. (2007). OLIGO 7 primer analysis software. *Methods Mol Biol, 402*, 35-60. doi:10.1007/978-1-59745-528-2_2

Sander, A. F., Lavstsen, T., Rask, T. S., Lisby, M., Salanti, A., Fordyce, S. L., . . . Arnot, D. E. (2014). DNA secondary structures are associated with recombination in major

Plasmodium falciparum variable surface antigen gene families. *Nucleic Acids Res, 42*(4), 2270-2281. doi:10.1093/nar/gkt1174

SantaLucia, J., Jr. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A, 95*(4), 1460-1465. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9465037

SantaLucia, J., Jr., Allawi, H. T., & Seneviratne, P. A. (1996). Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry, 35*(11), 3555-3562. doi:10.1021/bi951907q

SantaLucia, J., Jr., & Hicks, D. (2004). The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct, 33*, 415-440. doi:10.1146/annurev.biophys.32.110601.141800

Schnepf, E., Crickmore, N., Van Rie, J., Lereclus, D., Baum, J., Feitelson, J., . . . Dean, D. H. (1998). Bacillus thuringiensis and its pesticidal crystal proteins. *Microbiol Mol Biol Rev, 62*(3), 775-806. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9729609

Semegni, J. Y., Wamalwa, M., Gaujoux, R., Harkins, G. W., Gray, A., & Martin, D. P. (2011). NASP: a parallel program for identifying evolutionarily conserved nucleic acid secondary structures from nucleotide sequence alignments. *Bioinformatics, 27*(17), 2443-2445. doi:10.1093/bioinformatics/btr417

Soberón, M; Bravo, A. (2007). Las toxinas Cry de Bacillus thuringiensis: modo de acción y consecuencias de su aplicación. *Biotecnología, 14*, 303-314.

Spirollari, J; Wang, J. (2010). Consensus Structure Prediction for RNA Alignments. In Stefano Lonardi Jake Chen (Ed.), *Biological Data Mining* (pp. 714). NY: Chapman & Hall/CRC.

Stemmer, W. P. (1994). DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc Natl Acad Sci U S A, 91*(22), 10747-10751. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7938023

Storer, N. P., Babcock, J. M., Schlenz, M., Meade, T., Thompson, G. D., Bing, J. W., & Huckaba, R. M. (2010). Discovery and characterization of field resistance to Bt maize: Spodoptera frugiperda (Lepidoptera: Noctuidae) in Puerto Rico. *J Econ Entomol, 103*(4), 1031-1038. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20857709

Suarez, Miguel O. (2016). *Análisis molecular y determinación de la actividad tóxica de variantes cry11 de Bacillus thuringiensis.* (M.Sc), Universidad Industrial de Santander, Bucaramanga, Colombia.

Sun, F. (1999). Modeling DNA shuffling. *J Comput Biol, 6*(1), 77-90. doi:10.1089/cmb.1999.6.77

Tabashnik, B. E., Gassmann, A. J., Crowder, D. W., & Carriere, Y. (2008). Insect resistance to Bt crops: evidence versus theory. *Nat Biotechnol, 26*(2), 199-202. doi:10.1038/nbt1382

Tamura, K., Dudley, J., Nei, M., & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol, 24*(8), 1596-1599. doi:10.1093/molbev/msm092

Tinoco, I., Jr., Uhlenbeck, O. C., & Levine, M. D. (1971). Estimation of secondary structure in ribonucleic acids. *Nature, 230*(5293), 362-367. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4927725

van Frankenhuyzen, K. (2009). Insecticidal activity of Bacillus thuringiensis crystal proteins. *J Invertebr Pathol, 101*(1), 1-16. doi:10.1016/j.jip.2009.02.009

Verma, R., Schwaneberg, U., & Roccatano, D. (2012). Computer-Aided Protein Directed Evolution: a Review of Web Servers, Databases and other Computational Tools for Protein Engineering. *Comput Struct Biotechnol J, 2*, e201209008. doi:10.5936/csbj.201209008

Volkov, A. A., & Arnold, F. H. (2000). Methods for in vitro DNA recombination and random chimeragenesis. *Methods Enzymol, 328*, 447-456. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11075359

Volkov, A. A., Shao, Z., & Arnold, F. H. (2000). Random chimeragenesis by heteroduplex recombination. *Methods Enzymol, 328*, 456-463. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11075360

Wedge, D. C., Rowe, W., Kell, D. B., & Knowles, J. (2009). In silico modelling of directed evolution: Implications for experimental design and stepwise evolution. *J Theor Biol, 257*(1), 131-141. doi:10.1016/j.jtbi.2008.11.005

Winston, W. L. (2004). *Investigación de operaciones* (4 ed.): Thomson International.

Wu, J. Y., Zhao, F. Q., Bai, J., Deng, G., Qin, S., & Bao, Q. Y. (2007). Adaptive evolution of cry genes in Bacillus thuringiensis: implications for their specificity determination. *Genomics Proteomics Bioinformatics, 5*(2), 102-110. doi:10.1016/S1672-0229(07)60020-5

Wu, S; Floréz, A; Homoelle, B; Dean, D; Alzate, O. (2012). Two disulfide mutants in Domain I of Bacillus thuringiensis Cry3Aa dendotoxin increase stability with no effect on toxicity. *Advance Biological Chemistry, 2*, 123-131.

Xiong, Jin. (2006). *Essential Bioinformatics*: Cambridge University Press.

Zhang, W., Yuan, Y., Yang, S., Huang, J., & Huang, L. (2015). ITS2 Secondary Structure Improves Discrimination between Medicinal "Mu Tong" Species when Using DNA Barcoding. *PLoS One, 10*(7), e0131185. doi:10.1371/journal.pone.0131185

Zhao, X. M., Zhou, P. D., & Xia, L. Q. (2012). Homology modeling of mosquitocidal Cry30Ca2 of Bacillus thuringiensis and its molecular docking with N-acetylgalactosamine. *Biomed Environ Sci, 25*(5), 590-596. doi:10.3967/0895-3988.2012.05.014

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res, 31*(13), 3406-3415. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12824337

Apendices

Apendice A. Diagramas de Casos de Uso

Subsistema SANAFold:

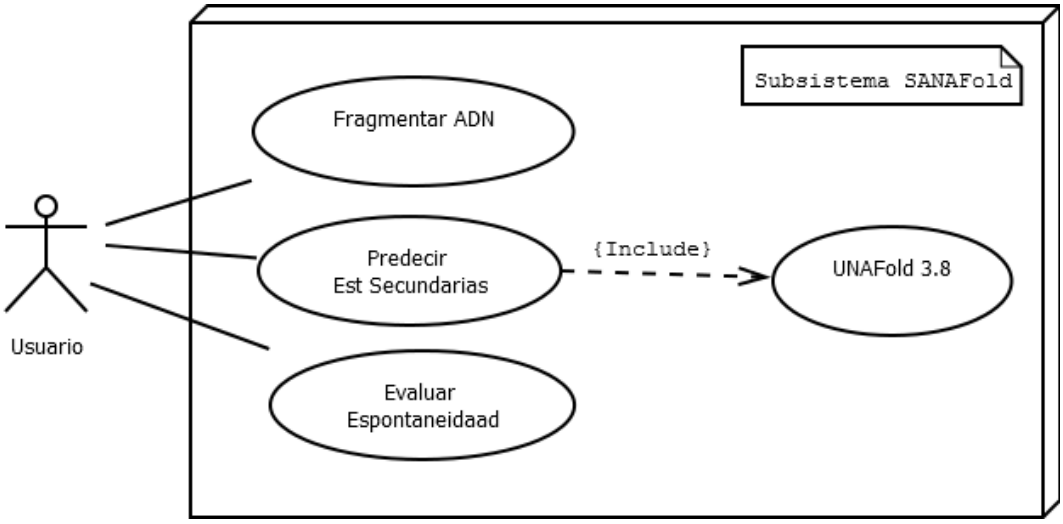


Diagrama casos de uso subsistema SANAFold

Descripción casos de uso SANAFold

Fragmentar ADN:

Entrada:

- Archivo fasta pre-procesado que contiene una secuencia de ADN que codifica para un dominio Cry
- Longitud de fragmentación

Proceso:

- Leer el archivo fasta con la secuencia de ADN
- Leer la longitud promedio de fragmentación
- Generar un valor aleatorio [0-1]
- Calcular la longitud del fragmento de ADN a cortar mediante la implementación de la función de distribución acumulada de Poisson
- Cortar el fragmento de ADN
- Almacenar el fragmento de ADN

Salida:
- Archivo unidimensional con los fragmentos ADN

Predecir Estructuras Secundarias:

Entrada:
- Archivo unidimensional con los fragmentos de ADN
- Condiciones experimentales como: concentraciones iónicas de Mg^{++} y temperatura de fusión.

Proceso:
- Leer el archivo unidimensional de fragmentos de ADN
- Leer las condiciones experimentales
- Llamar el caso de uso (UNAFold 3.8)
- Capturar el archivo de información termodinámica para cada estructura secundaria de ADN simulada
- Generar un archivo bidimensional .csv con la información termodinámica de todas las estructuras secundarias simuladas

Salida:
- Archivo bidimensional .csv con información termodinámica de estructuras secundarias.

Evaluar espontaneidad:

Entrada:
- Archivo bidimensional .csv pre-procesado a partir de estimadores hallados mediante la media de los datos y agrupados para cada gen cry.

Proceso:
- Leer el archivo bidimensional .csv
- Calcular valores f-ratio de análisis de varianza
- Diagramar la distribución de los datos

Salida:
- Valores f-ratio de análisis de varianza y graficas caja-bigote de distribución de los datos.

UNAFold 3.8:

Entrada:
- Fragmento de ADN
- Condiciones experimentales: Concentración iónica de Mg^{++} y temperatura de fusión.

Proceso:
- Leer el fragmento de ADN
- Leer las condiciones experimentales
- Construir una matriz de pares de bases a partir de la secuencia de nucleótidos del fragmento de ADN.
- Calcular la energía libre de las posibles estructuras secundarias de ADN
- Seleccionar las estructuras secundarias de ADN más estables

Salida:
- Esquema bidimensional de la estructura secundaria
- Información termodinámica de la estructura secundaria de ADN

Subsistema SAssembly

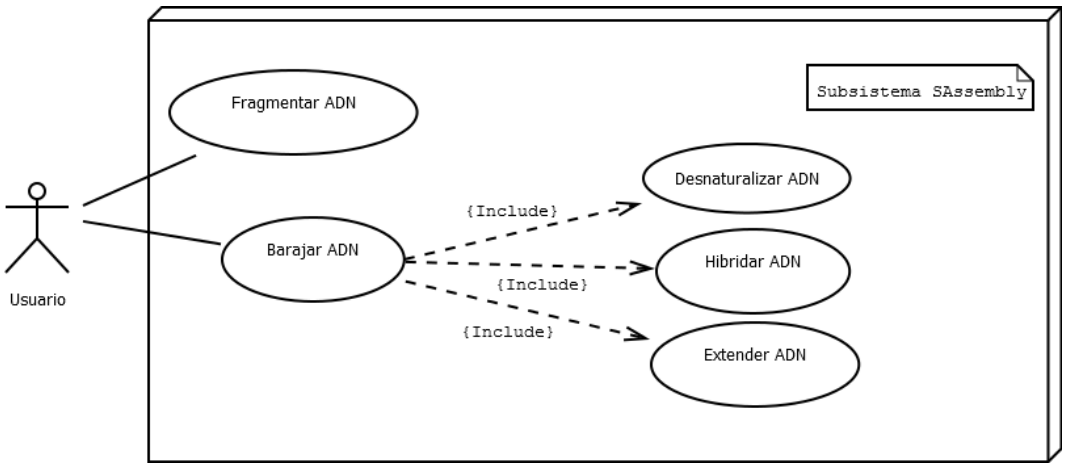


Diagrama casos de uso subsistema SAssembly

Descripción casos de uso SAssembly

Fragmentar ADN:

Entrada: • Archivo multi-fasta pre-procesado que contiene las secuencia de ADN de los genes cry • Longitud de fragmentación
Proceso: • Leer el archivo fasta con la secuencia de ADN • Leer la longitud promedio de fragmentación • Generar un valor aleatorio [0-1] • Calcular la longitud del fragmento de ADN a cortar mediante la implementación de la función de distribución acumulada de Poisson • Cortar el fragmento de ADN • Almacenar el fragmento de ADN
Salida: • Archivo unidimensional con los fragmentos ADN

Barajar ADN:

Entrada: • Archivo unidimensional con los fragmentos de ADN • Condiciones experimentales como: temperatura de fusión y longitud de solapamiento • Criterio de parada

Proceso:
- Leer el archivo unidimensional con los fragmentos de ADN
- Leer las condiciones experimentales
- Leer el criterio de parada
- Generar números aleatorios correspondientes a índices de los fragmentos de ADN
- Seleccionar fragmentos del archivo unidimensional
- Llamar el caso de uso (Hibridar)
- Llamar el caso de uso (Extender)
- Llamar el caso de uso (Desnaturalizar)
- Almacenar los nuevos fragmentos en archivos unidimensionales
- Considerar el criterio de parada en cada ciclo

Salida:
- Archivo multifasta con los genes ensamblados

Desnaturalizar ADN:

Entrada:
- Secuencias de ADN en dirección 5'-3', 3'-5'

Proceso:
- Leer las secuencias de ADN en dirección 5'-3', 3'-5'
- Almacenar en archivos unidimensionales separados las secuencias de ADN 5'-3', 3'-5'

Salida:
- Archivos unidimensionales con fragmentos ensamblados de ADN

Hibridar ADN:

Entrada:
- Archivos unidimensionales con fragmentos ensamblados de ADN
- Condiciones experimentales como: temperatura de fusión y longitud de solapamiento

Proceso:
- Leer los archivos unidimensionales con fragmentos ensamblados de ADN
- Generar números aleatorios
- Seleccionar dos fragmentos ensamblados de ADN
- Calcular la probabilidad de los diferentes caminos de hibridación entre los fragmentos seleccionados
- Almacenar en un archivo bidimensional las probabilidades de hibridación

Salida:
- Archivo bidimensional con las probabilidades de hibridación

Extender ADN:

Entrada:
- Archivo bidimensional con las probabilidades de hibridación

Proceso:
- Leer el archivo bidimensional con las probabilidades de hibridación
- Seleccionar la hibridación más probable
- Completar las secuencias de ADN en dirección 5'-3' y 3'-5'

Salida:
- Secuencias de ADN en dirección 5'-3', 3'-5'

Subsistema GenE-in

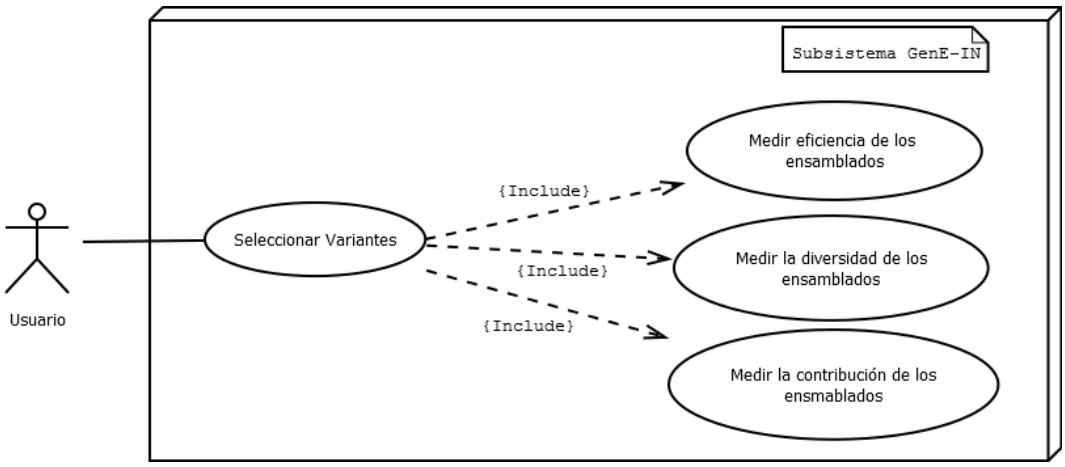


Diagrama casos de uso subsistema GenE-in

Descripción casos de uso GenE-in

Seleccionar variantes:

Entrada: • Archivo multifasta con los genes ensamblados • Criterios de selección: longitud de los parentales, valores de identidad, % conservación de dominios.
Proceso: • Leer el archivo multifasta con los genes ensamblados • Hacer la traducción de los genes ensamblados a secuencias de proteínas • Buscar los ORF a partir de las secuencias de proteínas • Almacenar en un archivo multifasta los ORF obtenidos • Llamar el caso de uso (Medir eficiencia de los ensamblados) • Llamar el caso de uso (Medir diversidad de los ensamblados) • Llamar el caso de uso (Medir contribución de los ensamblados)
Salida: • Archivo multifasta con los mejores ORF en términos de eficiencia, diversidad • Valores de contribución mediante % de conservación de dominios

Medir eficiencia de los ensamblados:

Entrada:
- Archivo multifasta con los ORF obtenidos de los genes ensamblados
- Longitud de los ORF de genes parentales

Proceso:
- Leer el archivo multifasta con los ORF obtenidos de los genes ensamblados
- Leer las longitudes de los ORF de genes parentales
- Comparar la longitud de los ORF de los genes ensamblados y los ORF de los genes parentales
- Almacenar en un archivo multifasta los ORF de longitud cercana a los parentales

Salida:
- Archivo multifasta con los ORF de longitud cercana a la longitud parental (eficiencia)

Medir diversidad de los ensamblados:

Entrada:
- Archivo multifasta con los ORF de longitud cercana a la longitud parental
- Archivo multifasta con los ORF de los genes parentales

Proceso:
- Leer el archivo multifasta con los ORF obtenidos de los genes ensamblados
- Leer el archivo multifasta con los ORF de los genes parentales
- Realizar un alineamiento múltiple de secuencias
- Identificación de la identidad de las secuencias
- Almacenar en un archivo multifasta los ORF de mayor a menor según el criterio de identidad

Salida:
- Archivo multifasta con los ORF con valores de identidad respeto a los parentales (diversidad)

Medir la contribución de los ensamblados:

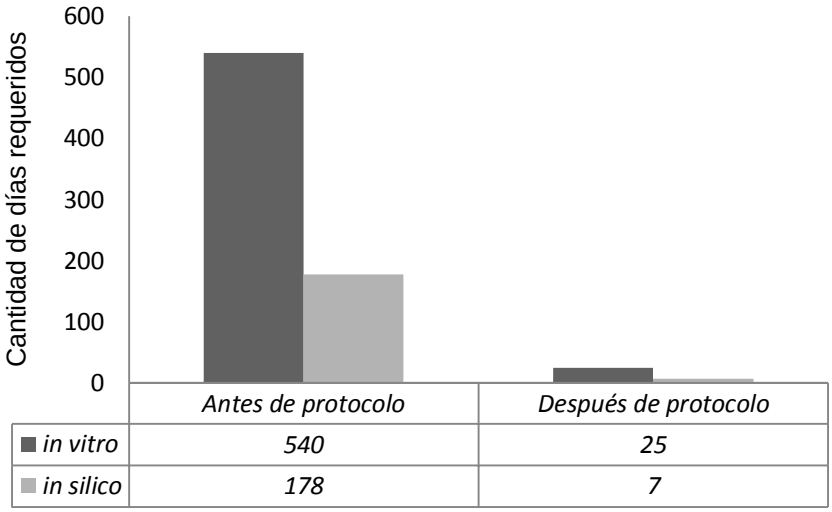
Entrada:
- Archivo multifasta con los ORF con valores de identidad respeto a los parentales (diversidad)
- Archivo multifasta con los ORF de los genes parentales
- Niveles de espontaneidad termodinámica de los dominios estructurales Cry (SANAFold)

Proceso:
- Leer el archivo multifasta con los ORF con valores de identidad respeto a los parentales
- Leer el archivo multifasta con los ORF de los genes parentales
- Leer los nivel de espontaneidad termodinámica de los dominios estructurales Cry
- Realizar un alineamiento múltiple de secuencias
- Comparar dominios conservados entre los mejores ORF

Salida:
- Valores de contribución mediante % de conservación de dominios

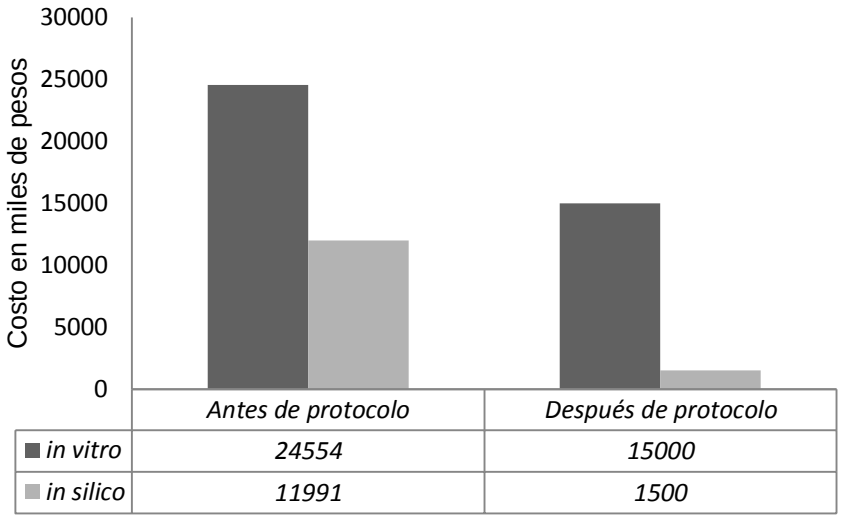
Apendice B. Análisis de Costos de Experimentos in Vitro e in Silico de Barajado de ADN

Análisis de costo (tiempo de realización de una biblioteca quimérica con barajado de ADN)



	Antes de protocolo	*Después de protocolo*
in vitro	*540*	*25*
in silico	*178*	*7*

Análisis de costo (valor de los insumos requeridos para obtener variantes con barajado de ADN)



	Antes de protocolo	*Después de protocolo*
in vitro	*24554*	*15000*
in silico	*11991*	*1500*

Soporte del análisis de costo (tiempo de realización de una biblioteca quimérica con barajado de ADN)

Costos de tiempo experimental para barajado de ADN *in vitro*		
Actividad	**Tiempo antes del protocolo**	**Tiempo después de protocolo**
Cultivos		2 días
Extracción del ADN plasmidico		1 día
PCR		1 día
Digestiones		2 días
Purificación de fragmentos		1 día
Ensamblado PCR	18 meses	1 día
Solubilización		2 días
Selección de transformantes		1 día
Checking		2 días
Secuenciación		7 días
Librerías de bacterias		5 días
18 meses		25 días

Los datos *in silico* fueron medidos a partir del uso de los parámetros por defecto de los software utilizados, la secuencia insumo para realizar las pruebas fue el fragmento del gen cry11Aa que codifica para el dominio I (Para las medidas de SANAFold y análisis de los resultado posteriores) y los genes cry11Aa y cry11Bb (Para las medidas de SAssembly y análisis de los resultado posteriores).

Costos de tiempo de procesamiento para barajado de ADN *in silico*						
		Tiempo después de protocolo computacional				
Actividad	**Tiempo antes del protocolo computacional**	**Frontend**	**Nodos**	**Pre-procesamiento pos-procesamiento**	**Mayor tiempo requerido**	**Tiempo para un barajado *in silico***
SANAFold	6 meses	1m44.304s	2m49.8375s	NA	2m49.8375s	7h6m45.093s*
Consolidación de datos		0m0.298s	0m04.4905s	10m	10m04.4905s	
Preparación ANOVA		0m0.186s	0m04.4102s	10m	10m04.4102s	
ANOVA		0m0.491s	0m04.317s	10m	10m04.317s	
Distribución de datos		0m0.285s	0m04.4603s	10m	10m04.4603s	
SAssembly		18m42.820s	45m54.788s	NA	45m54.788s	

Obtención de ORF		NA	NA	20m	20m	3h7m54.788s
Análisis de diversidad		NA	NA	30m	5m	
Análisis de contribución		NA	NA	2h	2h	
	6 meses	Total			4h19m02.3035s	10h14m39.881s**

Nota. *1h11m07.5155s x 6 dominios ≈ 7h6m45.093s (6 dominios para evaluar el comportamiento de dos genes cry parentales necesarios como mínimo para el proceso de Barajado de ADN. **El barajado de ADN *in silico* tiene un costo de tiempo adicional para tener secuencias biológicas y es el tiempo de síntesis de las secuencias aproximadamente 6 días.

Soporte del análisis de costo (tiempo de realización de una biblioteca quimérica con

Costos de material experimental para barajado de ADN *in vitro*		
Insumo	**Costo**	**Uso del insumo en el proceso *in vitro***
TAQ DNA/Marcador molecular/Tubos para PCR	$ 2.184.583	Taq: Enzima requerida para las PCR tanto de ensamblado de genes como para copiar los genes de interés/Marcador Molécular: Patrón de ADN de pesos conocidos (pares de bases) permite comparar el tamaño de los fragmentos obtenidos por PCR en un gel, como un tamizaje inicial/Tubos para PCR: tubos especiales para montar las PCR
Tetraciclina/Bromuro de Etidio/Eritromicina	$ 545.000	Tetraciclina y Eritromicina: Antibióticos necesarios para tamizar las bacterias transformadas con los plásmidos de interés/Bromuro de Etidio: reactivo necesario para ver el ADN en geles de agarosa
Reactivo LB Agar/botella Virio/Cajas de Petri	$ 4.983.244	Lb: medio de cultivo para bacterias/ Cajas de Petri: material de vidrio para servir el agar Lb y sembrar los microorganismos de interés.
Primers	$ 1.922.600	Cebadores requeridos para las PCR
TAQ Polimerasa/DNASA/X4 Vialesx250UI	$ 3.670.630	Taq: igual que arriba/ DNAsa: Enzima requerida para cortar los genes de interés.
Magnesio Cloruro tetrahidratado Fco x 250 gr	$ 464.000	Requerido como cofactor de la Taq polimerasa. Para las PCR
Enzimas/Wizard plus Minipreps/IPTG	$ 3.827.170	Enzimas de restricción: para verificar el patrón de bandas en los geles de agarosa/ Minipreps: extracción de ADN plasmidico de las bacterias de interés
Sistema de Clonaje TA topo overhangs y blunt	$ 2.126.000	Sistema de clonaje, para tomar los fragmentos obtenidos en el shuffling, ligarlos a un vector (plásmido)
T4 DNA Ligasa	$ 733.250	Para que los fragmentos queden bien adheridos a los vectores (clonados) está enzima es requerida para cerrar los extremos disponibles entre los sitios OH entre el inserto y el plásmido
Wizard gel/pcr clean	$ 293.300	Para purificar las muestras tanto de PCR como de geles y facilitar el proceso de secuenciación
Set DNTP	$ 952.000	Dideoxi nucleótidos, dGTP, dATP, dTTP, dCTP, bases nitrogenadas sintéticas, necesarias para la PCR
Lifagast Rapid DNA ligation System x 30 reacciones Cat M8221	$ 275.000	Sistema más refinado de ligación (igual que la ligada T4)
Células competentes	$ 427.380	Células comerciales, especiales para transformar Los plásmidos de interés con los genes deseados
Enzima ECORI cat 6011	$ 234.640	Enzima de restricción
Lambda DNA	$ 163.410	Lambda, es una polimerasa,
Bacterial Strain JM109, glycerol stock x 500ml cat P9751 Promega	$ 226.260	JM109 bacterias competentes (igual que arriba). Glicerol requerido para la crio preservación de las bacterias mutantes de interés (librería)
Enzima HindIII x 5000U cat R-6041	$ 90.085	Enzima de restricción
Secuenciación 16 reacciones M3 y T7 primers	$ 960.000	Secuenciación, requerida para determinar la secuencia de los genes obtenidos por DNA shuffling clonados en los vectores de interés y transformados en las bacterias competentes
Secuenciación PBTM5 BbseF/R	$ 136.000	
Secuenciación PCR PUCF/R	$ 204.000	

Secuenciación PCR4F/R	$ 136.000	
Total	$ 24.554.552	

Nota. * El barajado de ADN in silico tiene un costo adicional, dado que se requiere el proceso de síntesis de la secuencia obtenida, actualmente cada síntesis tiene un valor de $ 750.000