Diseño, desarrollo e implementación de una aplicación para revelar los aminoácidos ultraconservados que dan forma al núcleo interno hidrofóbico de las proteínas

Jorge Andrés Hernández Pabón

Trabajo de grado para optar el título de Ingeniero de Sistemas

Director:

Fernando Antonio Rojas Morales

Magíster en Ciencias Computacionales

Codirector:

Jorge Hernández Torres

Doctor en Biología Molecular

Universidad Industrial De Santander

Facultad De Fisicomecánicas

Escuela De Ingeniería De Sistemas

Bucaramanga

2020

Agradecimientos

A todos los profesores de la Escuela de Ingeniería de Sistemas e Informática, por todas las enseñanzas que me permitieron emprender y concluir este trabajo de grado.

A todos mis amigos y compañeros de carrera por haberme acompañado en este ciclo con su increíble amabilidad y solidaridad.

A mi director, el profesor Fernando Rojas Morales, quien aparte de ser un excelente docente y orientador, proyecta muy buenas energías con su pasión por la música.

A mi madre por su absoluto respaldo, paciencia y confianza durante el duro trayecto académico, quien, a pesar de las adversidades y desacuerdos, siempre estuvo a mi lado.

Especialmente a mi padre por su incondicional apoyo, extraordinaria dedicación, increíble sabiduría e incomparable amor por su labor de docente, y como cabeza de familia. Le agradezco infinitamente por haberme dado la oportunidad de trabajar a su lado en el desarrollo de este proyecto, construyendo una memorable experiencia como futuro profesional y como hijo, gracias a sus constantes enseñanzas e infatigables esfuerzos.

Mis más sinceros agradecimientos a la estudiante de Maestría en Biología Leidy Rocío Pico Martínez por su ayuda con la dinámica molecular con GROMACS y sus aportes en el campo de la bioinformática.

Contenido

Pág.

Lista de figuras

Pág.

Lista de tablas

Pág.

Resumen

Título: Diseño, desarrollo e implementación de una aplicación bioinformática para revelar los aminoácidos ultraconservados que dan forma al núcleo interno hidrofóbico de las proteínas.[*]

Autores: Jorge Andrés Hernández Pabón[**]

Palabras Clave: Estructura De Proteinas**,** Alineamiento De Secuencias, Aminoácido Hidrofóbico, Nucleo Hidrofóbico, Consenso.

Descripción:
Las proteínas se pliegan espontáneamente para lograr estructuras termodinámicamente estables. Las bases fisicoquímicas del plegamiento aún se investigan. Los pasos iniciales están determinados por la composición específica de aminoácidos y la hidrofobicidad es la principal fuerza motriz. Respecto a las proteínas globulares solubles, existe una alta correlación entre cadenas laterales no hidrofóbicas e hidrofóbicas y su ubicación en la superficie o el núcleo interno (baja accesibilidad al solvente), respectivamente. Existen útiles que permiten alinear múltiples secuencias. El análisis del producto de los alineamientos se centra en la conservación estricta de los aminoácidos, sin tener en cuenta su naturaleza polar/apolar. No obstante, el consenso derivado del alineamiento de una gran cantidad de secuencias homólogas puede revelar posiciones hidrofóbicas fijas, estructuralmente significativas. Se acepta que las posiciones hidrofóbicas ultraconservadas tienen relación directa con la composición del núcleo hidrofóbico interno. No existe una herramienta que construya un consenso hidrofóbico, basado en un alineamiento múltiple y teniendo en cuenta las tablas de hidrofobicidad más comunes. En este trabajo de grado, desarrollamos la aplicación JHydroScaffold, la cual construye un consenso de los residuos hidrofóbicos ultraconservados, en un conjunto de secuencias alineadas con la aplicación ClustalO. Un consenso de esta naturaleza puede ser empleado para predecir la composición del núcleo hidrofóbico interno de una proteína y, por consiguiente, un modelo predictivo de su estructura, tanto como de las propiedades fisicoquímicas y aspectos evolutivos. Para ilustrar la utilidad de la información que arroja la aplicación, se realizó un análisis del núcleo hidrofóbico interno de la enzima beta-lactamasa de *Citrobacter freundii*.

[*] Trabajo de investigación.

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Fernando Rojas Morales, MSc. Codirector: Jorge Hernández Torres, PhD

Abstract

Title: Design, development and implementation of a bioinformatics tool to reveal the ultra-conserved hydrophobic amino acids that are part of the protein hydrophobic core.[*]

Authors: Jorge Andrés Hernández Pabón[**]

Keywords: Protein Structure**,** Sequence Alignment, Hydrophobic Amino Acid, Hydrophobic Core, Consensus.

Description:
Proteins fold spontaneously to achieve thermodynamically stable structures. The physicochemical bases of the correct folding of native polypeptides are still being investigated. The initial steps are determined by the specific composition of amino acids and hydrophobicity is the main driving force. Regarding soluble globular proteins, there is a high correlation between non-hydrophobic and hydrophobic side chains and their location on the surface or the inner core (low solvent accessibility), respectively. A set of tools able to align protein sequences have been developed. The analysis of multiple alignments focuses on the strict conservation of amino acids (identity), irrespective of their polar/non-polar nature. Nevertheless, sequence consensuses derived from the alignment of a high number of homologous sequences can reveal ultraconserved hydrophobic positions of structural significance. It is accepted that the conservation of hydrophobic sites is closely related to the composition of the internal hydrophobic core. Hitherto, a tool able to build a hydrophobic consensus based on a multiple alignment and a selected hydropathy scale is not available. In this work, we developed the app JHydroScaffold, which builds hydrophobic consensuses from a set of sequences aligned with ClustalO. A consensus of this nature can be employed to predict the hydrophobic core composition of a protein and, therefore, a predictive model for its structure, as well as other physicochemical properties and evolutive aspects. Consequentially, we made an analysis of the hydrophobic core of the enzyme 'beta-lactamasa' from *Citrobacter freundii*, to illustrate the utility of the data offered by the app.

[*] Bachelor Thesis

[**] Faculty of Physics-Mechanics Engineering. School of Systems Engineering and Informatics. Advisor: Fernando Rojas Morales, MSc. Co-advisor: Jorge Hernández Torres, PhD

Introducción

El análisis estructural de las proteínas (plegamiento, arquitectura y estabilidad) ha evolucionado, en gran parte, gracias al desarrollo de una inmensa variedad de herramientas bioinformáticas. Existen programas para comparar, jerarquizar y predecir propiedades moleculares, basados únicamente en el ordenamiento primario de los aminoácidos (Carpentier & Chomilier, 2018). Al mismo tiempo, los métodos experimentales para adquirir nuevas secuencias también han evolucionado (i.e., secuenciación de nueva generación) y hoy día se ha acumulado un inmenso caudal de información en las bases de datos de proteínas, a la espera de ser analizados (Benson, y otros, 2017). Inicialmente, los modelos tridimensionales de las proteínas provenían en su mayoría de predicciones basadas en las secuencias de aminoácidos. En la actualidad, la determinación experimental de estructuras 3D es relativamente accesible y toda la información derivada se aloja en bases de datos de acceso público (Berman, y otros, 2000). No obstante, sigue habiendo un vasto universo de secuencias de proteínas de las cuales se desconoce su conformación 3D y la dinámica de su plegamiento. Por consiguiente, lo que se conoce de ellas se mantiene a nivel predictivo basándose únicamente en el estudio de la estructura primaria (secuencia de aminoácidos) (Nadzirin & Firdaus-Raih, 2012). Es en este contexto que proponemos construir un software para predecir cuáles son los aminoácidos hidrofóbicos esenciales para la conformación del núcleo hidrofóbico, con base en alineamientos múltiples de secuencias de la misma proteína, pero de organismos diferentes (proteínas ortólogas).

1. Objetivos

1.1. Objetivo general

Construir una herramienta bioinformática que establezca el consenso de los aminoácidos hidrofóbicos ultraconservados de una proteína, basada en alineamientos múltiples de secuencias.

1.2. Objetivos específicos

1. Desarrollar un algoritmo para la entrada de datos de alineamientos de secuencias de proteínas, realizados por la herramienta de código abierto ClustalO.
2. Desarrollar un algoritmo para la clasificación de los aminoácidos del alineamiento en las categorías hidrofóbicos/no hidrofóbicos, basado en las escalas de hidrofobicidad más comunes.
3. Desarrollar un algoritmo que construya un consenso en tiempo real, de los residuos hidrofóbicos/no hidrofóbicos, según el porcentaje mínimo de a considerar.
4. Desarrollar herramientas para exportación, visualización, búsqueda y análisis computacional de los datos de salida.
5. Desarrollar herramientas para la representación gráfica del consenso y del alineamiento.
6. Aplicar una prueba piloto que demuestre el correcto funcionamiento de las herramientas construidas.

2. Marco de referencia

2.1. Análisis estructural de las proteínas

Las propiedades estructurales de las proteínas se pueden establecer analizando su secuencia (1D), la sucesión de estructuras secundarias (2D) y la conformación de modelos tridimensionales (3D) (McWilliam, y otros, 2013). Comparar la información estructural de la misma proteína, en dos o más grupos taxonómicos, enriquece el conocimiento que se tiene de la misma. Hasta la fecha, los tres métodos de comparación siguen vigentes porque, como se dijo anteriormente, se desconoce la estructura tridimensional de muchísimas proteínas. La comparación entre proteínas a nivel 1D se hace mediante alineamientos de secuencias y la valoración de la conservación estricta de aminoácidos (Notredame, Higgins, & Heringa, 2000); (Sievers & et.al., 2011). La valoración cuantitativa del alineamiento se expresa en porcentaje de identidad. Por ejemplo, una proteína de una bacteria puede contener la secuencia aminoacídica "...ASDFPLKHG..."y la misma proteína, en otra especie bacteriana, podría ser, por ejemplo,"...ATDWPLKKG...". Si alineamos las dos secuencias, tendremos que hay aminoácidos que se han conservado durante el proceso de adaptación (i.e. A, D, P, L, K y G), mientras que otros han ido divergiendo por causa de las mutaciones (i.e., SxT, FxW, HxK). En este ejemplo, las dos proteínas compartirían 6/9 residuos, es decir 66,6% de identidad. Estos cambios no afectan la función, ya que sigue siendo la misma en las dos especies bacterianas. Esto ocurre, porque los aminoácidos claves para la estructura y la función están cuidadosamente conservados (Russell, Saqi, Sayle, Bates, & Sternberg, 1997). A

mayor identidad de secuencia, mayor relación evolutiva entre las especies que las albergan. (Brenner, Chothia, & Hubbard, 1998)

Por otra parte, las proteínas tienden a conservar las estructuras secundarias como α-hélice y hoja β. Si no se conoce la estructura 2D de una proteína, existen abundantes herramientas bioinformáticas para predecirla. Finalmente, las estructuras 3D tienden a conservarse muy bien, a pesar de que la estructura primaria haya cambiado durante la evolución. Esto es posible porque las proteínas toleran cambios entre aminoácidos de la misma naturaleza (e.g., hidrofóbico por hidrofóbico, polar por polar, etc.), sin afectar la estructura fundamental ni la función. Las estructuras 3D se comparan mediante superposición y la calidad de la superposición se realiza calculando la distancia promedio entre los átomos (*root-mean-square deviation* o RMSD), expresada en unidades Armstrong (Å), equivalentes a 10^{-10} m.

2.2. Hidrofobicidad y plegamiento

A pesar de que se ha avanzado en el entendimiento de la dinámica del plegamiento de las proteínas, queda mucho por explicar a nivel fisicoquímico y termodinámico. Lo que sí está bien establecido es el rol de los aminoácidos hidrofóbicos y su interacción con el solvente (agua) en la conformación final (Eisenhaber & Argos, 1994); (Gowder, Chatterjee, Chaudhuri, & Paul, 2014); (Munson, y otros, 1996); (Pace, Scholtz, & Grimsley, 2015). Se sabe que las proteínas pueden prescindir de algunos de sus residuos hidrofóbicos —especialmente de superficie—, sin que se vea afectada su estructura o función, pero sí su solubilidad. En otras palabras, no todos los aminoácidos hidrofóbicos intervienen de la misma manera en la búsqueda y estabilización de la conformación tridimensional. El problema es cómo determinarlos únicamente a partir de las

secuencias. Existen métodos bioinformáticos para analizar la hidrofobicidad local de una única secuencia, pero eventualmente con poca resolución (Kyte, Doolittle, Diego, & Jolla, 1982) o considerados enfoques idealizados que contrastan con la realidad biológica (Banach, Prymula, Konieczny, & Roterman, 2011, págs. 375–377). Es decir, contienen un margen de incertidumbre sobre el rol espacial de residuos claves para alcanzar la estructura plegada, sea desde el núcleo interno de la proteína o en la superficie. El análisis de los agregados hidrofóbicos (HCA) es un análisis de tipo bidimensional (2D), el cual establece la conservación local de grupúsculos hidrofóbicos entre aminoácidos vecinos (Callebaut, y otros, 1997, págs. 621–645). Esta aproximación, aunque de mayor resolución que la simple identidad de secuencia, se restringe a la comparación de agregados hidrofóbicos entre pares de secuencias. Sin embargo, de nuevo, carece de precisión en revelar posiciones hidrofóbicas fundamentales para la estructura proteica como el núcleo interno hidrofóbico o los hidrofóbicos de superficie.

Como dicho anteriormente, el propósito principal de los alineamientos es el de encontrar la mayor cantidad de coincidencias entre secuencias de proteínas homólogas. Aunque un alineamiento 1D arroja información de qué tanto se parecen dos o más secuencias, no toma en cuenta la naturaleza hidrofóbica o no hidrofóbica de las cadenas laterales y, por lo tanto, es poca la información estructural que revela. No obstante, los alineamientos múltiples de secuencia, tal vez por su simplicidad, pueden estar siendo subestimados, ya que no se le da suficiente importancia a la conservación de la naturaleza hidrofóbica o no hidrofóbica de cada posición en el alineamiento.

2.3. Alineamientos múltiples con el programa clustalo, como punto de partida.

La entrada de datos al software aquí desarrollado son los alineamientos de secuencias construidos con el programa ClustalO (Sievers & et.al., 2011); (Sievers & Higgins, 2014), el cual fue integrado a la aplicación. El código fuente de la herramienta ClustalO está disponible gratuitamente al público, amparado bajo una Licencia Pública General Reducida de GNU (GNU *Lesser General Public License*). ClustalO arroja como salida las secuencias alineadas en formato FASTA (Pearson & Lipmant, 1988), las cuales adquieren un nuevo formato en nuestra aplicación, para discriminar gráficamente entre aminoácidos hidrofóbicos y no hidrofóbicos. Se construye un nuevo consenso de posiciones hidrofóbicas/ no hidrofóbicas, según la escala hidrofóbica de Sweet & Eisenberg (1983). En ese estado, una serie de herramientas estarán a disposición del especialista para extraer valiosa información adicional de orden estructural y cuantitativa.

3. Metodología

3.1. Etapa 1

Realizar el análisis de requerimientos de la nueva herramienta, elaborar las labores de entrada y salida, y vincularla con la herramienta externa ClustalO.

1. Analizar la API de la herramienta de código abierto ClustalO, para develar el funcionamiento de entrada y salida de datos.

2. Construir un algoritmo que permita ingresar los datos a la aplicación ClustalO, mediante la invocación de un comando Shell.

3. Implementar un componente gráfico que permita intuitivamente ingresar los datos de entrada de la herramienta ClustalO mediante cadenas de proteínas en formato FASTA y egresar dicha salida al algoritmo de clasificación para su análisis.

3.2. Etapa 2

Diseñar el algoritmo que clasifica los aminoácidos del alineamiento en las categorías hidrofóbicos y no hidrofóbicos, utilizando las convenciones sugeridas por los especialistas.

1. Investigar las escalas de hidrofobicidad estándares (Sweet & Eisenberg, 1983)), (Kyte, Doolittle, Diego, & Jolla, 1982), Janin (1979)).

2. Construir un algoritmo que permita extraer la salida de la herramienta ClustalO, mediante la lectura de archivos de texto plano.

3. Construir un algoritmo que interprete y guarde dicha salida de manera ordenada, de tal manera que se pueda manipular cómodamente.

4. Construir un algoritmo de manipulación y análisis de los datos, que edifique el alineamiento de forma matricial. Al mismo tiempo, se construirá el consenso que clasifica los aminoácidos resultantes del alineamiento en las categorías hidrofóbicos y no hidrofóbicos. Dicho algoritmo estará basado en las escalas de hidrofobicidad estándares estudiadas anteriormente.

5. Implementar un componente gráfico que permita la visualización tanto del alineamiento como del consenso, de manera ordenada. Esta ventana contendrá las diferentes herramientas para la manipulación de contenidos, búsqueda, exportación y diferentes operaciones.

3.3. Etapa 3

Diseñar el algoritmo que modifique el consenso en tiempo real, basado en el porcentaje de secuencias tomadas en conjunto.

1. Construir un modelo que permita la construcción de los consensos según el número de secuencias tomadas en consideración.

2. Construir un algoritmo que modifique el alineamiento y el consenso, en tiempo real, de acuerdo con el porcentaje elegido de secuencias.

3. Construir un algoritmo que modifique el alineamiento y el consenso, tomando en cuenta la relevancia de la existencia de gaps en determinada zona, en tiempo real.

Implementar en dicho algoritmo el uso de las escalas de hidrofobicidad estándares.

4. Construir un componente gráfico que permita modificar en tiempo real el porcentaje de secuencias tomadas en conjunto, e implemente los algoritmos anteriores para su funcionamiento.

3.4. Etapa 4

Construir las herramientas ejecutables que soportan la exportación, visualización, búsqueda y análisis computacional de los datos de salida. Componentes ejecutables:

1. Ventana con la implementación gráfica de la etapa 1.

2. Ventana que muestre información pertinente del alineamiento actual.

3. Botones que copien el consenso al escritorio, al *clipboard*, o a otro destino.

4. Botones que copien el alineamiento al escritorio, al *clipboard*, o a otro destino.

5. Botón que exporte el alineamiento/consenso en formato .pdf al escritorio.

6. Botón que calcule el cociente hidrofóbico para el alineamiento actual.

7. Botón que calcule el consenso hidrofóbico en escalas.

8. Botón que realice la búsqueda de un patrón en el alineamiento/consenso.

9. Botón que cambie de escala hidrofóbica según la preferencia del especialista.

10. Botón que la autoría del software.

Componentes visuales adicionales:

a. Ventana que muestre el nombre de las proteínas alineadas.

b. Señalizador numérico de la región actual del alineamiento.

3.5. Etapa 5

Desarrollar la interfaz gráfica (*frame* general), para el consenso, el alineamiento y sus herramientas.

1. Construir y vincular todos los botones a su parte ejecutable en el *frame* general.

2. Vincular el componente gráfico construido en el punto (3.e).

3. Vincular el componente gráfico especificado en la parte (1.c), al botón (4.a.i.), y posteriormente al *frame* general.

4. Vincular el componente gráfico especificado en la parte (2.e) al *frame* general.

3.6. Etapa 6

Iniciar con una exhaustiva comprobación del correcto funcionamiento del algoritmo de construcción y análisis del alineamiento/consenso. Adicionalmente, añadir otra etapa de prueba, donde se ensayen todas las interfaces gráficas, botones, operaciones estadísticas, motores de búsqueda, importación, exportación, etc.

Iniciar con una exhaustiva comprobación del correcto funcionamiento del algoritmo de construcción y análisis del alineamiento/consenso. Adicionalmente, añadir otra etapa de prueba, donde se ensayen todas las interfaces gráficas, botones, operaciones estadísticas, motores de búsqueda, importación, exportación, etc.

1. Comprobar el funcionamiento de entrada y salida de la herramienta ClustalO, descrita en el punto (1.2).

2. Comprobar el correcto funcionamiento de los algoritmos de extracción, ordenamiento y constructor alineamiento/consenso, descritos en los puntos (2.2, 2.3 y 2.4).

3. Comprobar el correcto cambio en tiempo real del consenso manipulado por los algoritmos de los puntos (3.2 y 3.3).

4. Comprobar las funciones de entrada/salida (exportación en .txt, .pdf, etc.) de los puntos (4.3, 4.4 y 4.5).

5. Comprobar las operaciones cuantitativas de la herramienta, para que la salida sea la correcta, del punto (4.6, y 4.7).

6. Comprobar el correcto funcionamiento de la herramienta de búsqueda del punto (4.8).

7. Comprobar el correcto funcionamiento del visor de nombres de proteínas y señalizador de regiones (4.a y 4.b).

8. Comprobar las funcionalidades gráficas de la herramienta, como la correcta visualización del alineamiento/consenso del punto 5, y la vinculación de los demás componentes gráficos construidos de las etapas 1, 2, 3 y 4, de los puntos 3, 5, 5 y 4, respectivamente.

9. Comprobar el manejo de la memoria que la herramienta consume, dado un grande volumen de datos.

4. Desarrollo

4.1. Vista general

Para el desarrollo de este proyecto se utilizó el lenguaje de programación Java, debido a su popularidad, buena documentación y fundamentos vistos a través de la carrera. Conjuntamente, el manejo de *strings*, base de este proyecto, es una práctica muy común en este lenguaje con múltiples librerías externas de apoyo. Al mismo tiempo, el lenguaje Java facilitará la compatibilidad con la herramienta de código abierto ClustalO (desarrollada en C++), un alineador de múltiples secuencias de ADN o proteínas.

Debido a la gran cantidad de secuencias que dicha herramienta es capaz de alinear, se decidió implementar un modelo que acelerara la velocidad de ciertos algoritmos. Se eligió un modelo

concurrente debido a la necesidad de una interfaz gráfica intuitiva y, argumentalmente, que el volumen de datos no amerita la construcción de un modelo de supercomputación que soporte paralelismo. Finalmente, se efectúa la ejecución de ciertos algoritmos en modo multi-hilo.

4.2. Planificación y especificación

El desarrollo de dicha herramienta fue concebido por bloques (controladores), debido a la complejidad y exactitud que demanda la misma. Al finalizar la construcción de cada bloque, se efectuaron pruebas con el fin de constatar su correcto funcionamiento.

4.3. Diseño del controlador de la herramienta clustalo

4.3.1. API de la herramienta ClustalO. ClustalO es un software de código abierto desarrollado para alinear múltiples secuencias de ADN o de proteínas. La aplicación emplea dos o más secuencias de entrada y genera una representación matricial de cadenas de texto. El propósito es determinar los nucleótidos o los aminoácidos comunes en la vertical de cada posición del alineamiento. Paralelamente, ClustalO compara todas las secuencias por pares y arroja una matriz de porcentaje de parecido entre ellas (llamado identidad de secuencia).

Para revelar las posiciones hidrofóbicas verticalmente ultraconservadas, en la construcción de nuestra herramienta se tomó exclusivamente la matriz de cadenas de texto. Esto significa que la herramienta ClustalO ejecuta las primeras acciones de la aplicación desarrollada en este trabajo. El input inicial, es decir, las secuencias aminoacídicas, están dispuestas una seguida de la otra en un formato denominado *.fa o FASTA (Pearson & Lipmant, 1988). El output de la herramienta

ClustalO es un archivo de texto plano (.aln) que contiene la matriz alineada de secuencias. La aplicación aquí desarrollada retoma el archivo .aln como input y lo representa gráficamente, de forma que se revela nueva información.

ClustalO puede ejecutarse online (Madeira, y otros, 2019) o en el PC de escritorio mediante comandos ejecutables vía intérprete de comando de Windows, en el directorio de origen. Los siguientes comandos permiten ejecutar, analizar y modificar la data circulante (Figura 1):

1. --infile, el archivo que contiene las secuencias en formato FASTA.
2. --wrap, el tamaño máximo de las secuencias en el alineamiento.
3. --outfmt, el formato de salida.
4. -o, el nombre del archivo de salida.
5. –force, forzar sobrescribir el archivo si ya existe.
6. Otros, consultar API de ser necesario.

```
C:\Users\Zed\Desktop\JHydroScaffold> clustalo.exe --infile=InputClustalO.fa --wrap=100000 --outfmt=clu -o clustal.aln -v --force
Using 8 threads
Read 51 sequences (type: Protein) from InputClustalO.fa
not more sequences (51) than cluster-size (100), turn off mBed
Calculating pairwise ktuple-distances...
Ktuple-distance calculation progress done. CPU time: 0.05u 00:00:00.04 Elapsed: 00:00:00
Guide-tree computation done.
Progressive alignment progress done. CPU time: 2.24u 00:00:02.24 Elapsed: 00:00:03
Alignment written to clustal.aln
```

Figura 1. Ejemplo de ejecución de la herramienta ClustalO vía intérprete de comandos.

4.3.2. Algoritmo del controlador. Para la construcción de este algoritmo, se genera una invocación del intérprete de comandos de Windows, con el siguiente comando:

```
Process process = Runtime.getRuntime().exec("cmd /c clustalo.exe --infile=InputClustalO.fa --wrap=100000 --outfmt=clu -o clustal.aln -v --force");
```

Este comando ejecutará el software ClustalO para que alinee un número máximo de 100000 secuencias del archivo InputClustalO.fa, dispuestas en formato FASTA. El output por defecto (secuencias alineadas) se guarda en el archivo por defecto clustal.aln, sobrescribiendo el existente. Clustal.aln puede abrirse como archivo de texto plano y está listo para ser leído por el siguiente algoritmo aquí desarrollado.

4.3.3. GUI del controlador. La aplicación aquí desarrollada se compone de una interfaz gráfica que permite la entrada de datos (serie de secuencias de proteínas) en el mismo formato FASTA que exige ClustalO. Para facilitar el uso del software por expertos angloparlantes, los botones de la aplicación se diseñaron en idioma inglés. Desde la interfaz gráfica, activada al ejecutar el archivo JHydroScaffold.jar, se inicializa la construcción de un alineamiento con ClustalO, desde el botón '*New alignment'* (Figura 2), en la sección *'File'*. Esta acción abrirá el controlador desarrollado en la sección 3.2.

Figura 2. Botón para realizar un nuevo alineamiento.

En consecuencia, se abrirá un cuadro de diálogo sobre el cual se pega el conjunto de secuencias por alinear, en formato FASTA (Figura 3). El botón *'Align'* ejecutará ClustalO en segundo plano, lo cual puede tomar unos segundos, en función del número de secuencias por alinear. El comando shell '*Wait*' (Figura 4) informará el estado del proceso, bien sea en espera ('*Status*: *Waiting*'), en proceso de alineación ('Status: *Processing*') o tarea concluida ('*Status*: *Completed*!'). Completado

el proceso de alineamiento, el botón '*Plot*' representará el output "clustal.aln" de forma matricial en un *textpane* con el mismo nombre. En el caso en que el usuario decida descartar el alineamiento e iniciar uno nuevo, el botón '*Clear'* borrará lo ejecutado hasta el momento y el cuadro de diálogo estará disponible para un nuevo proceso. El botón '*Plot'* estará inhabilitado hasta que se produzca un alineamiento previo.

Figura 3. Ejemplo de la ventana de diálogo abierta mediante la ejecución del controlador para input de secuencias.

Adicionalmente, se muestra el *status* actual en el que se encuentra la ejecución del algoritmo, para mantener al usuario informado del estado del programa.

Figura 4. Ejemplo de finalización de ejecución.

4.4. Diseño del controlador y algoritmo que clasifica los aminoácidos según los valores de hidrofobicidad.

4.4.1. Escalas de hidrofobicidad de referencia. Las escalas hidrofóbicas son valores asignados que determinan la hidrofobicidad relativa de los aminoácidos (Simm, Einloft, Mirus, & Schleiff, 2016). Por convención, los valores positivos significan mayor hidrofobicidad y los negativos mayor afinidad por el agua (Tabla 1). La multiplicidad de escalas hidrofóbicas deriva en el método con el cual se mide la hidrofobicidad. Por ejemplo, algunos autores han construido sus escalas basados en las propiedades fisicoquímicas de las cadenas laterales "R" de los aminoácidos. Otros, por el contrario, se basan en datos estadísticos derivados de la posición que ocupa cada residuo dentro de las estructuras 3D alojadas en las bases de datos.

Tabla 1. *Ejemplos de atribución de hidrofobicidad según los autores con mayor aceptabilidad*

AA	K & D	S & H	Janin
Ala	1.8	0.62	0.3
Arg	-4.5	-2.53	-1.4
Asn	-3.5	-0.78	-0.5
Asp	-3.5	-0.9	-0.6
Cys	2.5	0.29	0.9
Gln	-3.5	-0.85	-0.7
Glu	-3.5	-0.74	-0.7
Gly	-0.4	0.48	0.3
His	-3.2	-0.4	-0.1
Ile	4.5	1.38	0.7

AA	K & D	S & H	Janin
Leu	3.8	1.06	0.5
Lys	-3.9	-1.5	-1.8
Met	1.9	0.64	0.4
Phe	2.8	1.19	0.5
Pro	-1.6	0.12	-0.3
Ser	-0.8	-0.18	-0.1
Thr	-0.7	-0.05	-0.2
Trp	-0.9	0.81	0.3
Tyr	-1.3	0.26	-0.4
Val	4.2	1.08	0.6

Nota. Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., & Bairoch A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press. pp. 571–607

Como puede observarse en la Tabla 1, los valores de hidrofobicidad son considerablemente diferentes entre los autores, lo que hace que los investigadores tengan preferencias específicas entre más de 50 escalas (Simm, Einloft, Mirus, & Schleiff, 2016), basadas en sus propias experiencias. No obstante, la tabla de mayor universalidad y aceptación en la comunidad científica es la de Kyte & Doolittle (1982). Existen algoritmos esencialmente basados en el análisis de segmentos sucesivos que calculan la hidropatía promedio dentro de un tramo de longitud predeterminada, a medida que se avanza a través de una secuencia. De esta manera, una secuencia de aminoácidos puede graficarse de acuerdo con los valores promedio de hidropatía (score) y así predecir las tendencias hidrofóbicas o hidrofílicas locales (Figura 5).

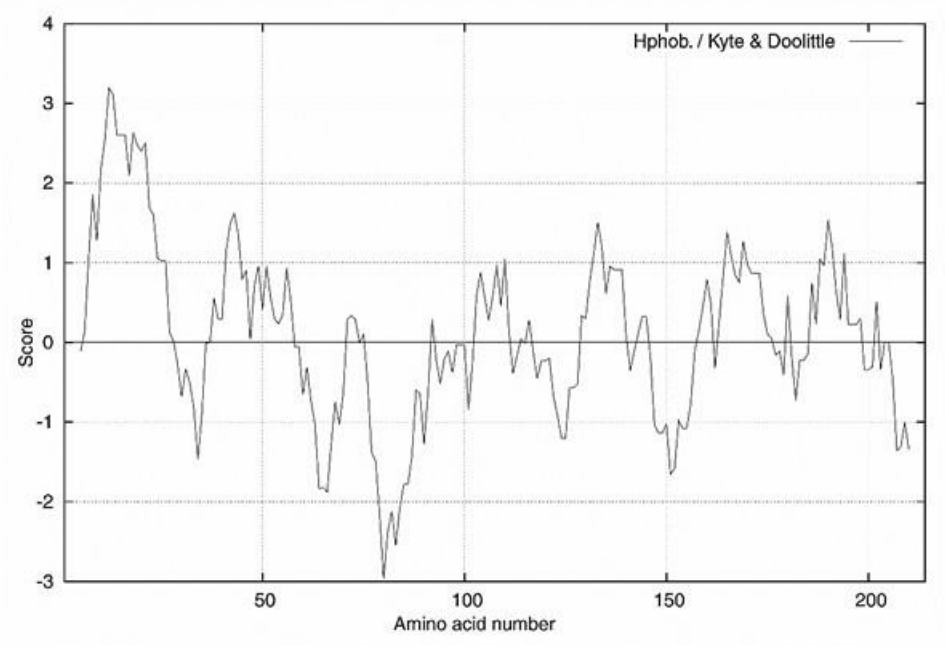


Figura 5. Representación gráfica del perfil de hidrofobicidad de la lipasa B de Bacillus subtilis (código de la Protein Data Bank 2qxu), según la escala de Kyte & Doolittle (1982). Segmentos aminoacídicos con score >0: regiones hidrofóbicas y <0: regiones hidrofílicas. Gráfica construida con la aplicación Protscale. Adaptada de Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., & Bairoch A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press. pp. 571–607

El software aquí diseñado está especialmente dirigido a diferenciar de modo gráfico los aminoácidos hidrofóbicos de los no hidrofóbicos. Aunque existe una multitud de escalas de hidrofobicidad (Simm, Einloft, Mirus, & Schleiff, 2016), se incluyeron tres de las más citadas en la literatura científica para la representación gráfica de los alineamientos (output):

1. Sweet & Heisenberg (1983). Residuos hidrofóbicos: VILFMWY.

2. Kyte & Doolite (1982). Residuos hidrofóbicos: IVLFCMA.

3. Janin (1979). Residuos hidrofóbicos: CIVLFMAGW.

Cada una de estas tablas será un *string* único en el que los residuos hidrofóbicos se representarán de un color verde, a diferencia de los demás que permanecerán en negro. Así mismo, estos *strings* serán fácilmente invocados para vincularlos con el algoritmo de la sección 4.4.3 y la 4.5.1.

4.2.2. Algoritmo de extracción y almacenamiento del output ClustalO. Para la siguiente etapa, se empleó el *output* del controlador anterior como entrada. La salida de ClustalO está conformada por los nombres de las secuencias alineadas, las cadenas de aminoácidos respectivas y un consenso que, para los efectos de nuestra aplicación, es irrelevante y por tanto se ignora (Figura 6).

```
CLUSTAL O(1.2.2) multiple sequence alignment


G_max           MAEEAKAKGNAAFSAGDFAAAVRHFSDAIALSPSNHVLYSNRSAAHASLQNYAEALADAQ
A_arabicum      MADEAKARGNAAFSSGDFNAAVTHFTDAINLDPTNHVLFSNRSAAHASLHQYVEALNDAN
A_lyrata        MADEAKAKGNAAFSSGDFNSAVNHFTDAINLSPTNHVLFSNRSAAHASLHHYDEALSDAK
                **:****:******:*** :** **:*** *.*:****:**********:.* *** **:
```

Figura 6. Ejemplo del contenido del output clustal.aln. (i.e., alineamiento de secuencias proteicas), construido con ClustalO. Las secuencias fueron alineadas buscando las mejores coincidencias estrictas en la vertical de cada posición, representadas por asteriscos (consenso, abajo).

Nuestra herramienta está en capacidad de extraer y almacenar dichos datos en diferentes campos para el análisis: en *strings* para los nombres de las secuencias y en vectores de caracteres para las secuencias aminoacídicas. Como dicho anteriormente, el consenso (asteriscos) no es tenido en cuenta porque un nuevo consenso será construido por la aplicación, en función de la hidrofobicidad de los residuos.

4.4.3. Algoritmo de manipulación y representación gráfica del output de ClustalO. Esta etapa es considerada la más importante del presente proyecto. Con base en las cadenas de caracteres generadas por ClustalO y almacenadas en *strings* y vectores, se construyó un algoritmo que genera un nuevo tipo de consenso. En efecto, el algoritmo discrimina los aminoácidos del alineamiento entre hidrofóbicos y no hidrofóbicos, según las escalas de hidrofobicidad citadas anteriormente, tomando la tabla de Sweet & Heisenberg (1983) por defecto. Trabajos previos han demostrado que la escala de Sweet & Heisenberg refleja una alta precisión tanto a nivel predictivo como en el laboratorio. No obstante, el algoritmo permite que el usuario cambie de escala hidrofóbica a su criterio.

Para lo anterior, se programaron los siguientes pasos en código:

1. Inicializar un vector de caracteres, en donde se almacenará el consenso final (señalado con una flecha en la Figura 7).

2. Inicializar una variable con la primera secuencia (indicado con un cuadro horizontal rojo en la Figura 7).

3. Recorrer la lista de secuencias.

4. Para cada secuencia, fragmentar la variable que contiene la primera secuencia y la secuencia actual, en un vector de caracteres.

5. Realizar las siguientes comparaciones con todos los elementos de los vectores de caracteres del paso 4, en la vertical.

- Si, verticalmente, los caracteres son iguales, se añade el carácter coincidente al vector de caracteres del consenso final, en letras azules para 'no hidrofóbico' y en verdes para 'hidrofóbico' (ejemplo indicado con recuadro rojo ('A', no hidrofóbico) y verde ('L' hidrofóbico, en la Figura 7).
- Si los caracteres no son iguales, pero están dentro de la misma categoría de 'hidrofóbico', se añade el carácter "I" al vector de caracteres del consenso final (ejemplo indicado con un recuadro verde en la Figura 7).
- Si los caracteres no son iguales y están dentro de la categoría de 'no hidrofóbico', estén o no entremezclados con 'hidrofóbicos', se añade el carácter "x" al vector de caracteres del consenso final (ejemplos rodeados con recuadro azul en la Figura 7).

6. Realizar el mismo proceso para cada cadena existente.

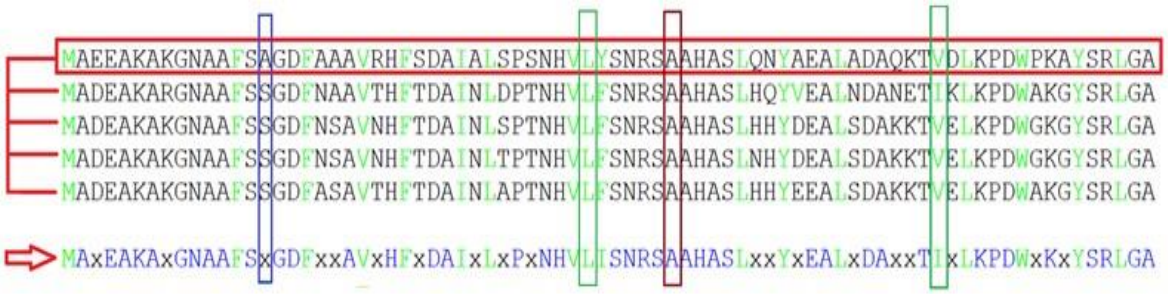


Figura 7. Ejemplo de un consenso (flecha roja) derivado de un alineamiento de 5 secuencias y representado gráficamente con el algoritmo de manipulación (escala hidrofóbica de Sweet & Heisenberg (1983), por defecto). Los recuadros rodean las variantes consensuales en la vertical (ver texto).

Como dicho previamente, al final del alineamiento surgirá un vector único de caracteres. Este recibirá el nombre de ‘consenso hidrofóbico’ o simplemente ‘consenso’ y es el resultado final que contiene la información valiosa para el investigador. En resumen, el consenso hidrofóbico constituirá el output de este controlador y el foco del resto de los scripts desarrollados.

4.4.4. GUI del controlador. Con el fin de presentar la información de manera intuitiva y ordenada se diseñó un visualizador gráfico que mostrara el nombre, la secuencia y el consenso en caracteres coloreados (Figura 8).

Figura 8. Ejemplo de GUI del controlador 4.4.

La idea subyacente de este GUI es la sencillez, de tal forma que toda secuencia sea fácilmente identificable por su respectivo nombre a la izquierda. El consenso hidrofóbico se representa al final, en una zona despejada. Es importante anotar que es posible expandir y contraer la barra entre el nombre y las secuencias, en los casos en que los nombres sean muy largos y se requiera una lectura en extenso. Las barras de desplazamiento están diseñadas para avanzar de manera ágil, en caso de que el alineamiento se extienda más allá del ancho del monitor.

Como dicho anteriormente, se agregaron colores al alineamiento para facilitar la interpretación del resultado:

1. Los aminoácidos hidrofóbicos se grafican de color verde. Cuando toda una columna esté compuesta de residuos hidrofóbicos, sean el mismo o no, el color verde se extiende hasta el consenso. Esta posición se denomina hidrofóbica ultraconservada.

2. Los aminoácidos no hidrofóbicos se grafican de color negro. Cuando una columna esté compuesta del mismo residuo no hidrofóbico, el aminoácido consensual se representa en color azul. Esta posición también se reconoce como ultraconservada (normalmente residuos del sitio activo), pero se mantiene en la categoría de no hidrofóbica.

3. Una posición en la vertical, en la cual no existe ninguna coincidencia relevante, permanece de color negro y estará ocupada por una ‘x’.

4.5. Diseño del controlador y algoritmo que actualizan el consenso en tiempo real, basados en el porcentaje mínimo de secuencias con un hidrofóbico por cada columna.

4.5.1. Algoritmo de modificación del consenso. Como explicado anteriormente, el consenso se construye a partir del análisis de la composición de residuos hidrofóbicos/no hidrofóbicos, en la vertical de cada posición individual del alineamiento, considerando el 100% de las secuencias.

Una columna de solo hidrofóbicos (VILFMYW) se destacará en el consenso con una “I”. Sin embargo, puede ocurrir que, por ejemplo, una sola secuencia tenga un residuo no hidrofóbico y las demás, e.g., 99 secuencias, tengan todas un hidrofóbico. En ese caso, el 99% de las secuencias tendría una posición hidrofóbica y habría una sola no hidrofóbica en esa posición. Los investigadores pueden argumentar, por ejemplo, que un no hidrofóbico puede camuflarse como hidrofóbico, formando enlaces de hidrógeno. La conclusión sería que esa posición en la vertical podría ser considerada globalmente como hidrofóbica y esa decisión se debería ver reflejada en el consenso, remplazando la ‘x’ con una ‘I’ de color verde. Lo mismo aplicaría para valores decrecientes de ‘estringencia’, i.e., 98%, 97%, 96%, etc. de las secuencias. Por esta razón, se incluyó un algoritmo que analice cada posición vertical del alineamiento, y reconfigure el consenso, ignorando secuencias que crean ruido de fondo. Al seleccionar con la ‘barra de estringencia’ (ver sección 5d) el porcentaje mínimo de secuencias que deben tener un aminoácido hidrofóbico en la vertical, en tiempo real se modificará el consenso. La ecuación para definir si

una posición será o no hidrofóbica en el consenso, al mover la 'barra de estringencia' es la siguiente:

$$\text{Estringencia (\%)} = \frac{\#\ \text{mínimo de secuencias con un residuo hidrofóbico en una posición 'x'}}{\#\ \text{total de secuencias}} * 100$$

Para ejemplificar la ecuación anterior, construimos un alineamiento de cinco secuencias y los consensos que arroja nuestra aplicación, con 100 y 80% de estringencia. Como puede apreciarse en la Figura 9, en el consenso con 100% de estringencia (izquierda) surgen solo 4 posiciones hidrofóbicas ultraconservadas (F, Y, V y F). Sin embargo, al disminuir la estringencia al 80% (derecha), el programa recalcula y actualiza el consenso, arrojando la aparición de dos nuevas posiciones hidrofóbicas señaladas con flechas azules y rojas (F, Y, V, F, I e I). La barra de estringencia recalcula el consenso en saltos de una unidad, i.e., 99%, 98%, 97%, 96%, 95%, etc. Queda a criterio del investigador interpretar los nuevos consensos que surjan a medida que la barra de estringencia se aleja del 100%. Para el ejemplo de la Figura 9, el consenso no se modificará hasta tanto la barra de estringencia se posicione sobre el 80%.

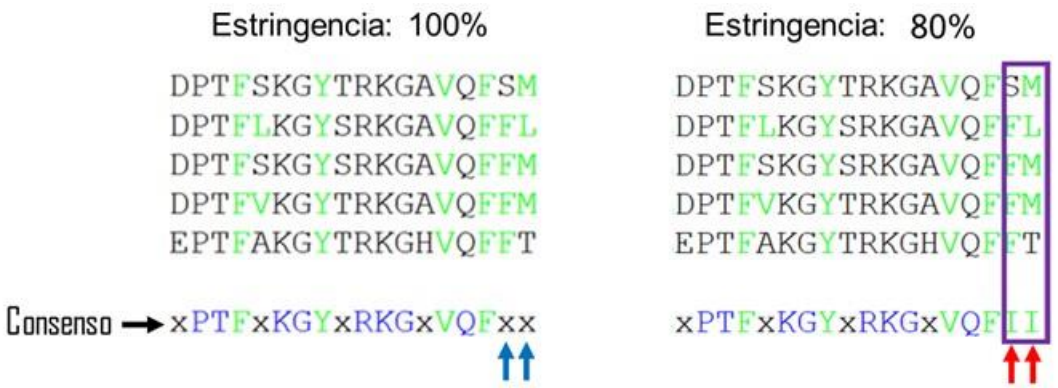


Figura 9. Modificación del consenso en función del número mínimo de secuencias tomadas en conjunto, para determinar si una posición en la vertical es o no hidrofóbica.

Como ejemplo, se muestran los consensos con 100 y 80% de estringencia. Flechas azules: posiciones consensuales no hidrofóbicas (‘x’). Flechas rojas: posiciones consensuales hidrofóbicas tomadas al menos el 80% de las secuencias.

4.5.2. Algoritmo de modificación según la existencia de gaps. Las secuencias proteicas varían en su longitud dependiendo de las especies biológicas. Es decir, una misma proteína puede tener, por ejemplo, 455 aminoácidos en una especie (e.g., fríjol) y 463 en otra (e.g., habichuela). Esto ocurre porque los cambios evolutivos de cada proteína (o gen) siguen rumbos diferentes según la especie y la presión de selección. Por consiguiente, si alineamos dos secuencias de longitudes diferentes, necesariamente quedarán espacios vacíos en el alineamiento, como consecuencia de la existencia de un determinado aminoácido en una proteína, pero no en la otra. Esos espacios vacíos se llaman ‘*gaps’* y ClustalO los rellena con un guion (-). Los gaps permiten ajustar el alineamiento a una misma longitud, porque de lo contrario surgiría un desfase adverso entre las secuencias. En la figura 10 se aprecia la presencia de gaps en un alineamiento y la forma como ClustalO los representa.

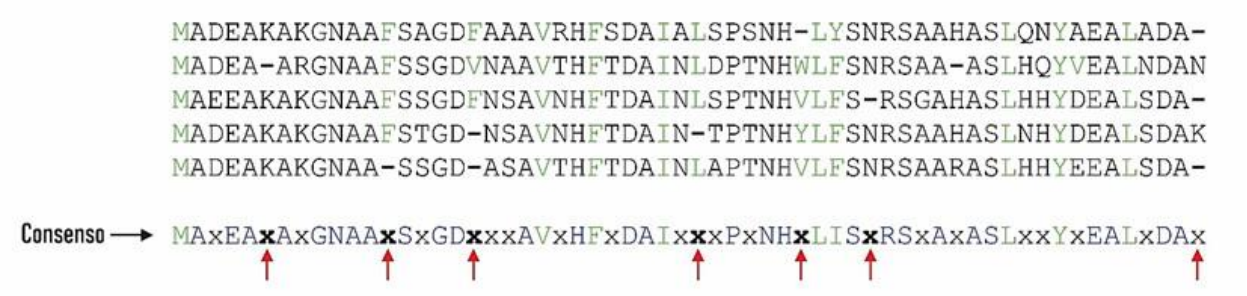


Figura 10. Ejemplo de un alineamiento de cinco secuencias con la presencia de '*gaps'* (guiones). La presencia de gaps afecta el consenso (x en negrilla, flechas rojas), a pesar de que el resto de las secuencias comparta el mismo residuo o aminoácidos hidrofóbicos.

Desde el punto de vista de nuestra aplicación, la inserción de gaps afecta considerablemente la interpretación algorítmica del alineamiento y la representación gráfica del consenso. Por este motivo, se implementó un algoritmo [□ Ignore gaps] que permite desconocer la existencia de los gaps en el alineamiento y reconfigure el consenso de acuerdo con la naturaleza de los aminoácidos de cada columna (Figura 11).

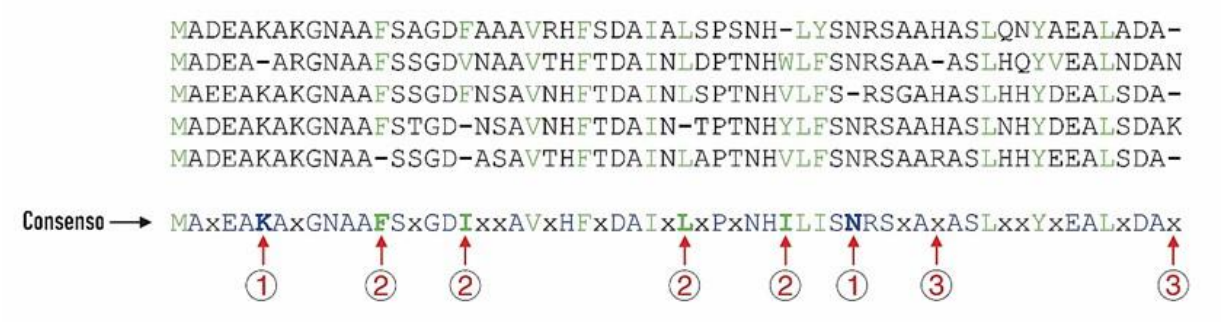


Figura 11. Al activar la herramienta [□ Ignore gaps], el programa actualiza el consenso de acuerdo con la composición aminoacídica de cada columna, i.e., residuo no hidrofóbico ultraconservado (#1 en círculos), posición hidrofóbica ultraconservada (#2) o sin efecto (#3).

Es de anotar que el investigador debe recurrir a esta opción con prudencia, sobre todo cuando hay presencia de muchos gaps en el alineamiento, para evitar conclusiones erróneas desde el punto de vista evolutivo.

4.5.3. Reconfiguración del consenso con otras escalas hidrofóbicas. Como dicho anteriormente, la tabla de hidrofobicidad por defecto para la construcción del consenso es la de Sweet & Eisenberg (1983). Sin embargo, los investigadores tienen preferencias personales por otras escalas, de acuerdo con su experiencia. Por esta razón, nuestra aplicación permite reconfigurar el consenso con base en otras dos tablas de hidrofobicidad de amplio uso en la comunidad científica. Esta opción añade versatilidad y empleabilidad a la aplicación. La pestaña '*Swap Tables'* ofrece la posibilidad de navegar entre las escalas y facilitar la elección. Los aminoácidos considerados hidrofóbicos se listan frente al nombre de la tabla (Figura 12).

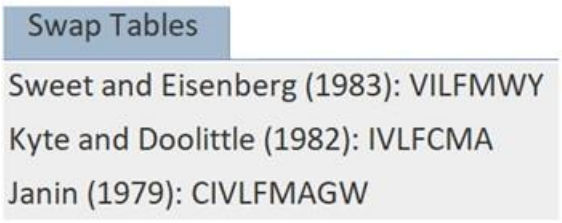


Figura 12. La pestaña '*Swap Tables'* permite alternar entre las tres tablas de hidrofobicidad más comunes.

Al escoger una de las opciones, el consenso se actualiza automáticamente. Obviamente, las posiciones hidrofóbicas ultraconservadas de cada tabla serán coloreadas en verde en el GUI, como vimos en los ejemplos anteriores.

Adicionalmente, los aminoácidos específicos según cada tabla serán recoloreados de verde en el GUI, como vimos en los ejemplos anteriores.

4.5.4. GUI del controlador. El diseño de este controlador hace referencia a una barra deslizante que el investigador desplaza para reconfigurar el consenso en tiempo real, dependiendo del porcentaje mínimo de secuencias que deben tener un aminoácido hidrofóbico en una posición de la vertical. El propósito es evitar que el especialista, tenga que efectuar un conteo manual para construir consensos con porcentajes de identidad variables. De hecho, es una tarea ardua de realizar manualmente y con alta posibilidad de error.

Adicionalmente, se añadió un visor para monitorear el porcentaje de estringencia en tiempo real, un contador de aminoácidos ultraconservados en el consenso según la estringencia y el combo-box para activar la opción de ignorar gaps, vista anteriormente.

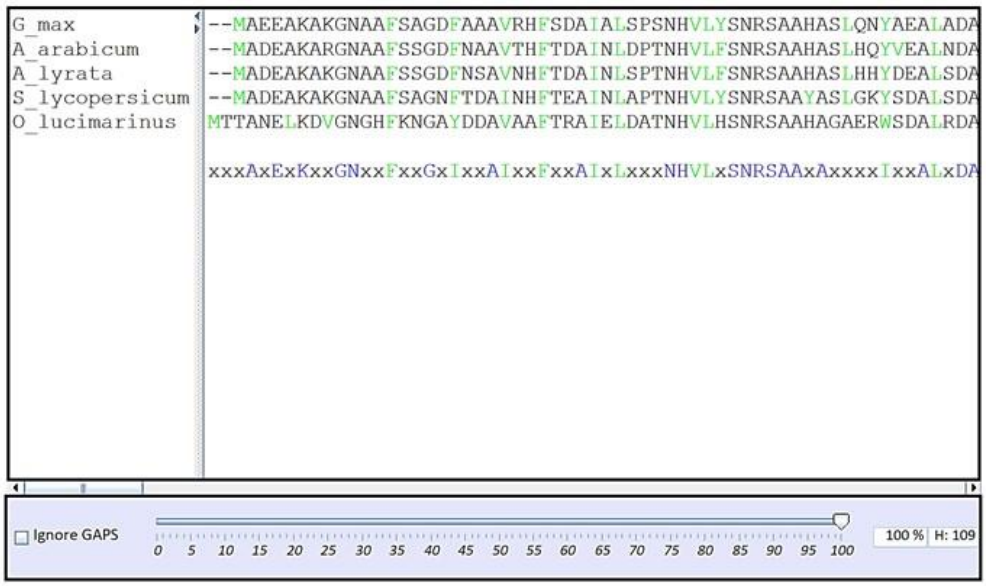


Figura 13. Ejemplo del GUI del controlador.

4.6. Diseño del controlador y construcción de herramientas de exportación, visualización, búsqueda y análisis computacional.

En el diseño de este controlador se crearon categorías para cada tipo de herramienta. Dichas categorías están distribuidas de la siguiente forma:

4.6.1. Menú 'File'. Contiene las herramientas de información del alineamiento, copia y exportación.

4.6.1.1. Información del alineamiento: Además de categorizar los aminoácidos y construir consensos hidrofóbicos, la aplicación aquí diseñada arroja una lista de metadatos que son importantes para el investigador. Se trata de la escala hidrofóbica en uso, el número de secuencias alineadas, la longitud total del alineamiento (incluyendo gaps), el número de posiciones hidrofóbicas y no hidrofóbicas ultraconservadas, el porcentaje de estringencia actual y el estado activado/desactivado de la casilla '[□ *Ignore gaps*]'.

La lista de metadatos fue diseñada para ser extensible, dependiendo de las nuevas herramientas incorporadas. Al activar la pestaña '*Alignment info'*, se obtendrá una tabla de propiedades como se muestra en la Figura 14.

Figura 14. Tabla de propiedades del alineamiento en tiempo real. Se proporcionan datos globales importantes para el investigador, de acuerdo con la sección 6.1.1.

4.6.1.2. Herramientas de copia: Para el funcionamiento de la aplicación, el usuario necesita transferir la data de input (secuencias de proteínas) hacia el visor, desde archivos de texto, .docx, un navegador de internet, etc. En sentido contrario, antes de construir una versión definitiva de un alineamiento, es usual que el investigador quiera exportar información previamente obtenida en nuestra aplicación y pegarla en archivos como Microsoft Word o PowerPoint. Por esta razón, se añadieron instrucciones de copiado y pegado desde y hacia el clipboard (Figura 15, izquierda y centro).

Figura 15. Ilustración de herramientas de copia. Izquierda: herramientas de 'deshacer' (undo), 'cortar', 'copiar' y 'pegar' y 'seleccionar todo', para importar o editar datos dentro del visor de la aplicación.

Las instrucciones se activan presionando el botón derecho del mouse y seleccionando la opción requerida. Centro: herramientas de copia de datos de salida desde la aplicación. Desde el menú 'File' se puede copiar al clipboard, bien sea el consenso, el alineamiento o los dos y pegar los datos hacia archivos txt, docx, etc. Derecha: herramientas de exportación de datos de salida en formatos .txt y .pdf.

1. Undo: Se activa presionando el botón derecho del mouse dentro del visor. Deshace el borrado o inserción de caracteres dentro del visor.

2. Cut: Corta los caracteres seleccionados dentro del visor y conserva una copia en el *clipboard*.

3. Copy: Copia los caracteres seleccionados dentro del visor y los almacena en el *clipboard*.

4. Paste: Pega en el visor la información almacenada en el *clipboard* (típicamente secuencias en el formato Fasta).

5. Select all: Selecciona todo el contenido del visor.

6. Copy consensus to clipboard: Se activa desde el menú '*File'*. Realiza una copia al *clipboard* de la última línea proveniente del visor del alineamiento, es decir, el consenso.

7. Copy alignment to clipboard: Realiza una copia al *clipboard* del alineamiento proveniente del visor, exceptuando el consenso.

8. Copy all to clipboard: Realiza una copia del alineamiento y el consenso al *clipboard*, provenientes del visor.

4.6.1.3. Herramientas de exportación. Una vez perfeccionado un alineamiento que el investigador quiera conservar para futuro análisis o diagramar para una publicación, las herramientas de exportación facilitan esta tarea (Figura 15, derecha). Tanto el consenso, el alineamiento o los dos, pueden ser exportados en formato .txt. Por otra parte, el alineamiento y el consenso pueden ser exportados a formato .pdf el cual es fácilmente editable por aplicaciones como Adobe Illustrator, CorelDraw o un editor de PDF.

1. Consensus to .txt: Exporta el consenso al escritorio del usuario en formato .txt.

2. Alignment to .txt: Exporta el alineamiento, excluyendo el consenso al escritorio del usuario en formato .txt.

3. All to .txt: Copia al *clipboard* todo el contenido del visor.

4. All to .pdf: Exporta el alineamiento con su respectivo consenso al escritorio del usuario, en formato .pdf. Para generar este tipo de archivos, se utilizó una librería externa llamada IText. Se trata de un poderoso toolkit de generación, programación, manejo y manipulación de PDFs. IText es de uso gratuito, mientras el software esté respaldado por una licencia AGPL (https://bit.ly/2qyoMOM).

4.6.1.4. Botón cerrar. Cierra el programa y cualquier instancia generada por éste. Se activa desde el menú '*File'*, '*Exit'*.

4.6.2. Menú *'tools'*. Además de construir consensos, nuestra aplicación está diseñada para exportar archivos con información cuantitativa referente a las secuencias alineadas. El menú '*Tools'* contiene herramientas que demandan bastantes recursos de la máquina. En la versión actual del software, estas funciones solo aplican para la escala hidrofóbica Sweet & Heisenberg (1983).

4.6.2.1. 'Calculate stringency vs hydrophobic ratio (csv file)'. Esta función construye una tabla de valores en formato .csv que contiene un valor aquí denominado 'cociente hidrofóbico' o '*hydrophobic ratio'*, en función de todos los valores de porcentaje de estringencia (1 a 100%). El '*hydrophobic ratio'* se calcula mediante la siguiente ecuación:

$$\text{Cociente hidrofóbico} = \frac{\text{cantidad de aminoácidos ultraconservados (consenso)}}{\text{longitud total del alineamiento}}$$

Para lanzar el proceso, se selecciona la pestaña '*Tools'* y posteriormente *'Calculate stringency vs hydrophobic ratio* (*csv file*)'. El resultado será un archivo .csv editable con una hoja de cálculo como Microsoft Excel. En la figura 16 se muestra un ejemplo del *output*. Se observa que se crean dos columnas, una con los valores de estringencia de 1 a 100% (incrementos de una unidad) y otra con el cociente hidrofóbico calculado con cada valor de estringencia. Más adelante se ilustrará el propósito de estos cálculos.

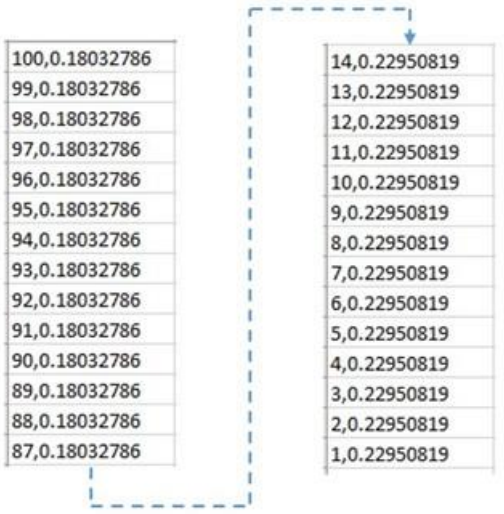


Figura 16. Ejemplo del contenido del archivo '*Stringency vs hydrophobic ratio*.csv'. Se tabulan los valores de estringencia de 1 a 100%, en función del cociente hidrofóbico.

En la figura se muestran los valores decrecientes de 100% de estringencia hacia 1%. La línea punteada representa los datos desde 86% hasta 15%. La tabla de valores fue invertida de manera decreciente (de 100% a 1%) con el programa Microsoft Excel.

4.6.2.2. 'Scaled hydrophobic consensus': Otra herramienta añadida a la pestaña '*Tools'* es el cálculo escalado del consenso hidrofóbico. Se trata de la generación de un archivo .txt con el consenso hidrofóbico en función de los valores decrecientes de estringencia. El objetivo es visualizar los porcentajes de estringencia máximos en los cuales surgen las posiciones hidrofóbicas ultraconservadas. Recordemos que el consenso hidrofóbico estricto del alineamiento se construye con el 100% de estringencia. Sin embargo, recordando el ejemplo de un alineamiento con una posición en la vertical que tiene un residuo no hidrofóbico contra 99 secuencias que sí tienen un hidrofóbico, al final de cuentas esa columna podría considerarse como hidrofóbica. Lo mismo se aplicaría con 99%, 98%, 97%, etc. Con los datos del archivo *'Scaled* hydrophobic

consensus.txt' se puede construir una figura que permite visualizar cada posición en la vertical y a partir de cuál valor máximo de estringencia se convierte en hidrofóbica (Figura 17).

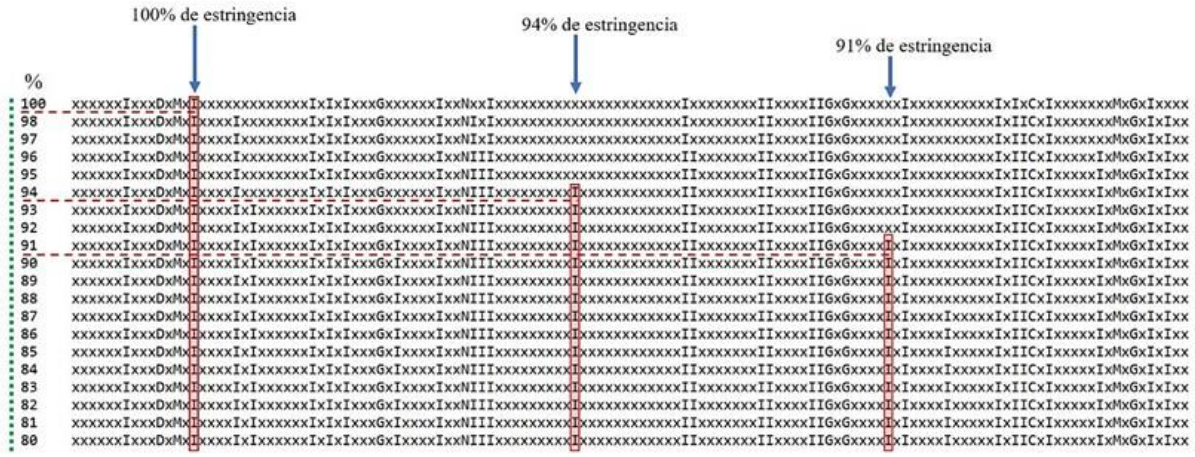
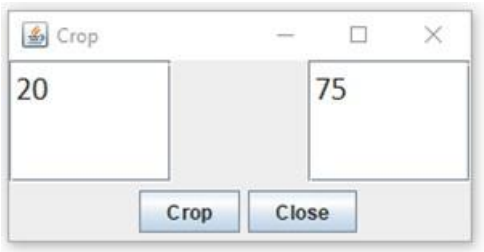


Figura 17. Ejemplo del *output* del consenso hidrofóbico escalado. Solo se muestran los valores decrecientes de 100% a 80%, pero la línea vertical verde (margen izquierda) va hasta 1%.

Las flechas azules señalan tres posiciones con 100%, 94% y 91% de estringencia. Se aprecia que al 100%, toda la columna es hidrofóbica (recuadro rojo). La segunda solo es hidrofóbica desde el 94% hacia abajo, así como la tercera lo es desde el ≤91%.

4.6.2.3. 'Crop from initial alignment' (BETA): Cuando el total del alineamiento resulta demasiado largo, es concebible que el investigador se interese por una porción específica y quiera eliminar el resto, para no crear ruido de fondo. La herramienta '*Crop from initial alignment*' permite recortar porciones a izquierda y derecha del alineamiento (Figura 18). Por ahora, esta función está en versión beta y solo permite recortar valores, siempre desde la longitud inicial del alineamiento. Es decir, si inicialmente se recorta desde, por ejemplo, 50 residuos a la izquierda y

posteriormente se quiere recortar 10 residuos más, en la casilla '*crop'* se escribirá 60 (50+10) y no 10, de la siguiente manera:



4.6.3 Menu *'Find'*. Cuando se alinean múltiples secuencias, como será frecuente con nuestra aplicación (>100)

Puede resultar difícil ubicar posiciones específicas dentro del alineamiento, e.g., el sitio activo, estructuras secundarias, etc. Por esta razón, una herramienta de búsqueda de secuencias es fundamental.

4.6.3.1. 'Find *sequence'*: El menú '*Find'* despliega la opción '*Find sequence'*. Al activarla, se abre un cuadro de diálogo dentro del cual se escribe el orden de aminoácidos que se quiere buscar (Figura 18).

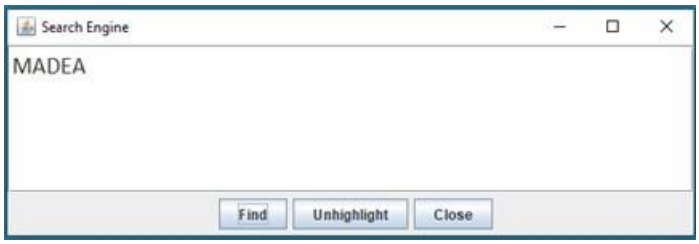


Figura 18. Cuadro de diálogo de la herramienta '*Find sequence'*.

Al presionar el botón '*Find'*, las secuencias estrictamente idénticas serán resaltadas en color rojo dentro del alineamiento y/o consenso. La secuencia buscada debe estar cuidadosamente escrita, sin espacios adicionales ni otros caracteres que no se encuentren dentro del alineamiento (Figura 19). Con frecuencia, las proteínas tienen segmentos de secuencia repetidos, por lo que la herramienta de búsqueda puede resaltar más de una opción, como se muestra en la Figura 19. En caso de querer quitar el resaltado en rojo, por ejemplo, para una nueva búsqueda, basta con presionar el botón '*Unhighlight'*.

Figura 19. Ejemplo de resultado de una búsqueda de un ordenamiento específico de aminoácidos dentro del alineamiento (MADEA, en este caso).

4.6.3.2. 'Unhighlight': En caso de que se haya presionado el botón 'Close' y se quiera remover el resaltado en color rojo dentro del alineamiento, una segunda opción 'Unhighlight' dentro del menú 'Find' permitirá hacerlo.

4.6.4. Menú *'swap tables'*. Como se expuso al principio, la tabla de hidrofobicidad por defecto es la de Sweet & Heisenberg (1983). No obstante, se incluyeron otras escalas de uso frecuente para que el investigador pueda comparar los consensos. El menú '*Swap tables'* contiene los botones que alternan entre las diferentes escalas hidrofóbicas de la Tabla 1. Para que el investigador tenga presente la lista de aminoácidos hidrofóbicos según la escala, éstos se relacionan frente a cada opción del menú *'Swap tables'.*

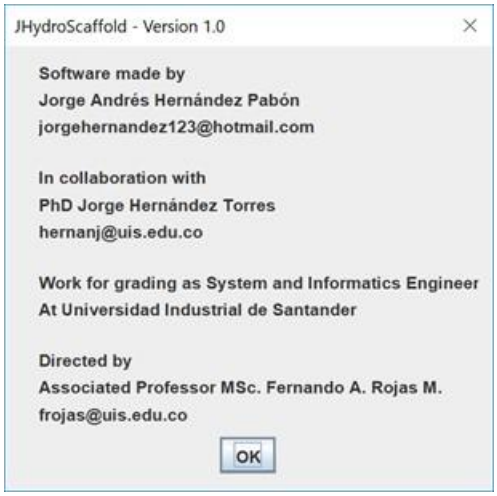
4.6.4.1. Sweet & Heisenberg (1983): VILFMWY: Cambia la apariencia de hidrofóbicos/no hidrofóbicos del alineamiento y el consenso, en conformidad con la escala hidrofóbica de Sweet & Heisenberg (1983). Para esta tabla, 7/20 aminoácidos son considerados hidrofóbicos.

4.6.4.2. Kyte & Doolitle (1982): IVLFCMA: Cambia la apariencia de hidrofóbicos/no hidrofóbicos del alineamiento y el consenso, en conformidad con la escala hidrofóbica de Kyte and Doolittle (1982). Para esta tabla, 7/20 aminoácidos son hidrofóbicos, excluyendo la metionina e incluyendo la cisteína.

4.6.4.3 Janin (1979): CIVLFMAGW: Cambia la apariencia de hidrofóbicos/no hidrofóbicos del alineamiento y el consenso, en conformidad con la escala hidrofóbica de Janin (1979). Según Janin, 9/20 aminoácidos son hidrofóbicos. Incluye la cisteína, la glicina y la alanina.

4.6.5. Menú *'About'*. Este menú despliega un panel de información que contiene la lista de autores y colaboradores de este proyecto.

4.6.5.1. 'Authors'.



4.6.6. Herramienta de visualización de número de campo actual. Además de la herramienta '*Find'* para localizar segmentos aminoacídicos, es también útil que el investigador esté informado de la ubicación de cada residuo en el alineamiento. La herramienta de visualización de número de campo actual muestra, en tiempo real, la posición relativa de un aminoácido en el alineamiento, al señalarlo con el cursor (Figura 20, flecha roja).

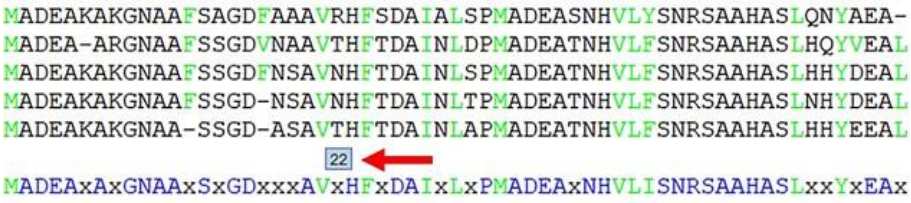


Figura 20. La herramienta de visualización de número de campo actual muestra la posición relativa de un aminoácido en el alineamiento. El cursor se posicionó sobre un residuo de valina (V) y un recuadro informa que ocupa la posición 22 (flecha roja).

4.7. Vinculación de controladores al gui principal

Todos los controladores, fueron articulados para ofrecer una interfaz intuitiva y fácilmente navegable (Figura 21). Las herramientas de esta sección fueron integradas en un toolbar estándar. Es importante resaltar que el usuario (normalmente biólogo molecular) no necesitará tener conocimientos de programación porque la interfaz gráfica es muy clara y estable. En el uso cotidiano no se ha detectado ningún tipo de fallo.

4.8. Multi-threading

Para este software se desarrollaron varias herramientas con alta demanda de recursos de máquina (i.e., lectura/escritura). Debido a la arquitectura lineal del código, la interfaz gráfica se bloqueaba, hasta el cumplimiento de la tarea en proceso. Por este motivo, se implementó un controlador multi-hilo para las herramientas de las secciones 4.3.3, 6.2.1 y 6.2.2. Esta configuración independiza las herramientas para que se ejecuten de manera concurrente.

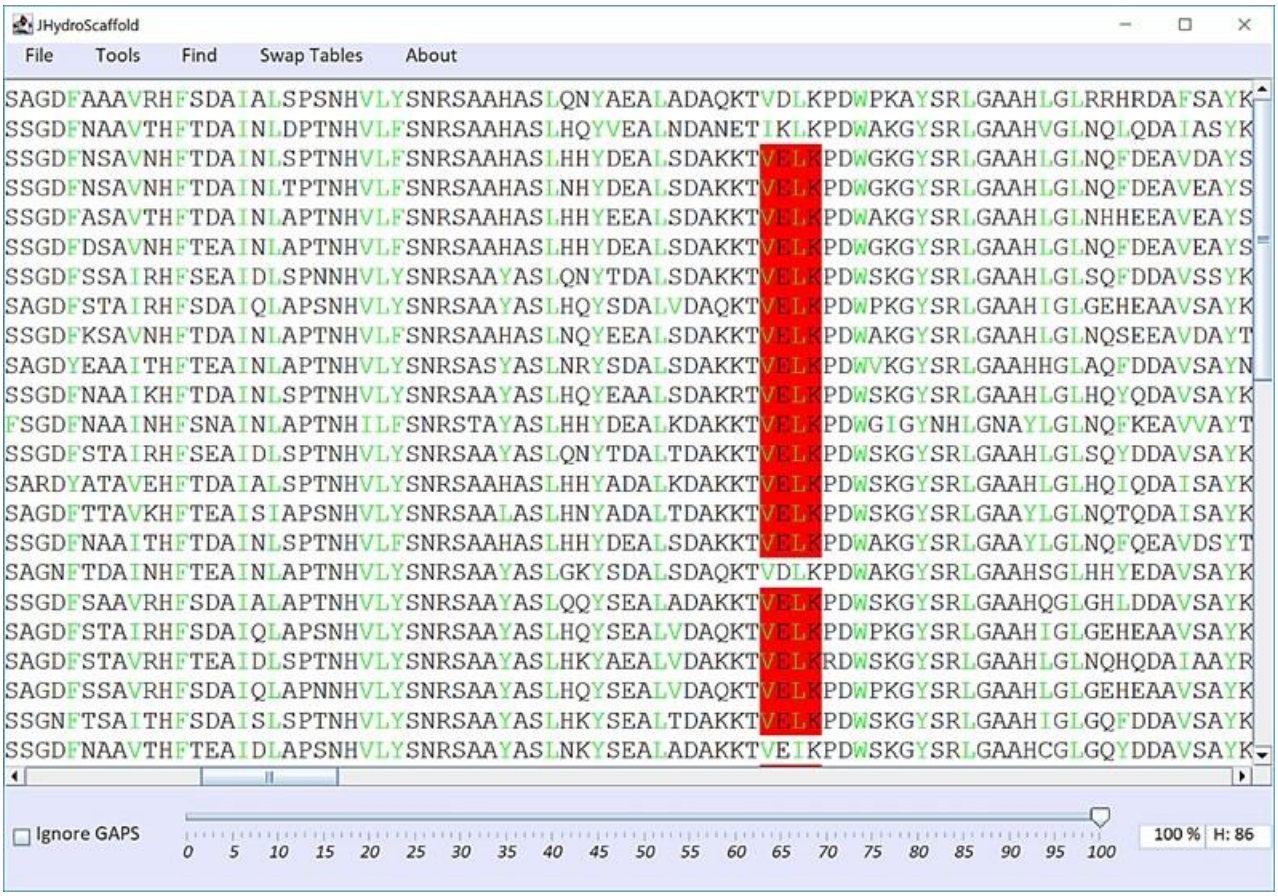


Figura 21. Vista general de la interfaz gráfica de la aplicación.

5. Resultados

En el presente trabajo de grado se construyó un software bioinformático de código abierto, concebido para generar, manipular y exportar una nueva forma de consenso derivada de alineamientos de secuencias de proteínas ortólogas (i.e., misma proteína en especies diferentes). Esta forma innovadora de consenso revela información estructural oculta en las secuencias mismas de las isoformas presentes dentro de una rama taxonómica. Es importante aclarar que

hasta donde alcanza nuestro conocimiento, no existen programas que generen este tipo de consensos y tampoco la manipulación subsiguiente de los mismos. Es decir, sobre la marcha se ha ido construyendo la herramienta, pero también se ha ido comprendiendo el significado molecular y el impacto que puede tener la información generada para la biología estructural.

En lo referente al diseño de software, se intenta mantener un balance entre poder de computación y sencillez gráfica. Esto significa que la aplicación es lo suficientemente poderosa como para abordar una mediana densidad de secuencias, a diferencia de otras alternativas como el complejo uso de soluciones en paralelo que no ofrecen ningún tipo de interfaz gráfica. Al mismo tiempo es fácil e intuitivamente controlable por el investigador, opuesto a otros diseños sobrecargados de herramientas irrelevantes, que solo entorpecen, dificultan y ralentizan los procesos.

Por otra parte, el impacto en el ámbito de la bioquímica, la biofísica y la biología molecular, disciplinas enfocadas en estudiar la estructura tridimensional de las proteínas está aún por definirse. Por analogía con las estrategias de gestión de datos, la información emanada de nuestra aplicación se podría entender como una forma de extracción de metadatos. Con el propósito de ilustrar lo novedoso que resulta analizar las secuencias proteicas desde este nuevo enfoque, a continuación, se presenta un análisis estructural de la enzima bacteriana beta-lactamasa.

5.1. Análisis estructural de la Beta-Lactamasa clase C.

La beta-lactamasa (BRENDA:EC 3.5.2.6, 'penicilinasas') es una enzima bacteriana que provee resistencia a los antibióticos β-lactámicos como las penicilinas, las cefalosporinas, monobactámicos y carbapenémicos. Las beta-lactamasas por lo general son secretadas (es decir,

liberadas al medio circundante) por las enterobacterias y otros bacilos Gram negativos y bacterias Gram positivas. Los antibióticos β-lactámicos son inhibidores de la última etapa de la síntesis de la pared celular bacteriana, como se muestra en la Figura 22. La enzima DD-Transpeptidasa une las cadenas aminoacídicas intermediarias que forman la red del peptidoglicano. La DD-Transpeptidasa pierde su actividad ante la presencia de los antibióticos β-lactámicos y las bacterias no pueden sintetizar la pared celular, lo cual hace a las bacterias sensibles a la ruptura. Para contrarrestar este efecto inhibitorio, en el curso de la evolución surgieron las beta-lactamasas, las cuales hidrolizan el antibiótico y hacen que pierdan su actividad bactericida.

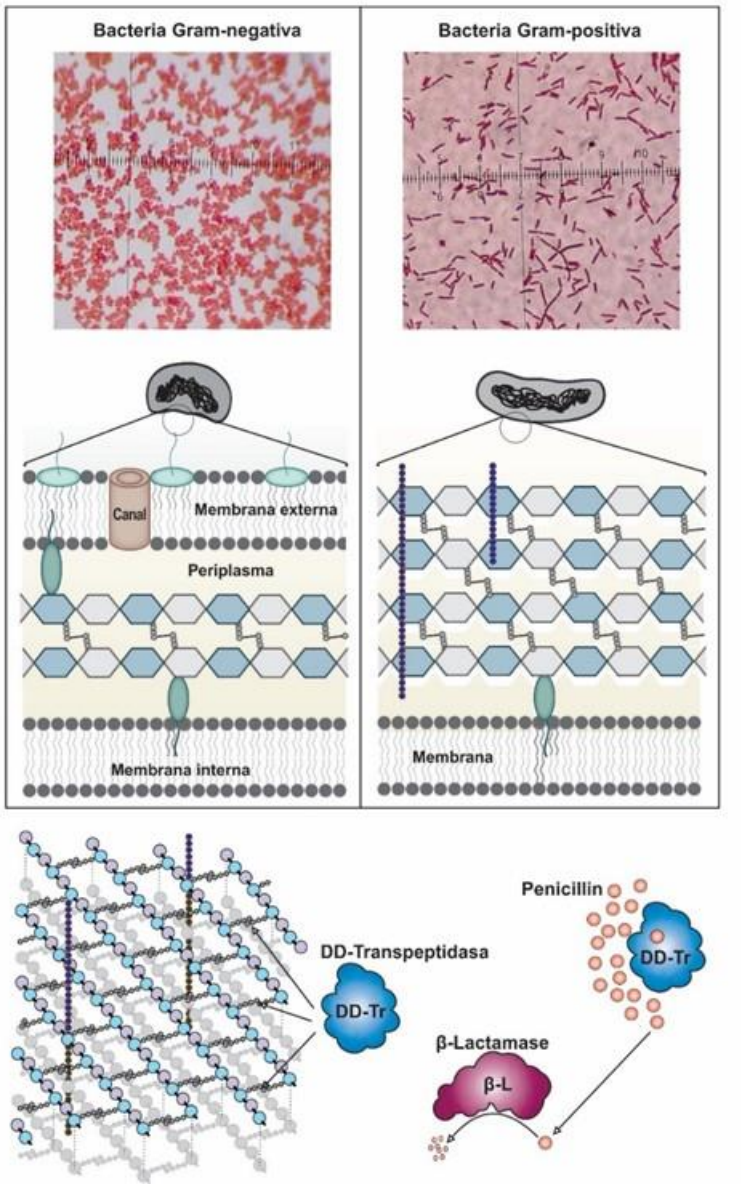


Figura 22. Las bacterias Gram negativas (izquierda) y Gram positivas (derecha) se diferencian en la permeabilidad y retención de un colorante.

Las bacterias Gram negativas contienen dos membranas (interna y externa) que impiden el paso del colorante, mientras que las Gram positivas sí lo permiten. La envoltura de peptidoglicano las hace resistentes a la presión osmótica y al daño mecánico. Se trata de una red entrelazada de

azúcares y aminoácidos. La enzima encargada de fusionar las cadenas laterales de aminoácidos es la DD-Transpeptidasa. Los antibióticos β-lactámicos inhiben la enzima e impiden la reconstitución de la red protectora. En consecuencia, la bacteria sufre ruptura y muerte celular. La enzima beta-lactamasa degrada el antibiótico (penicilina) y protege a la bacteria de la acción bactericida.

Las beta-lactamasas hidrolizan los β-lactámicos y los convierten en un ácido inactivo, como se muestra en la Figura 23.

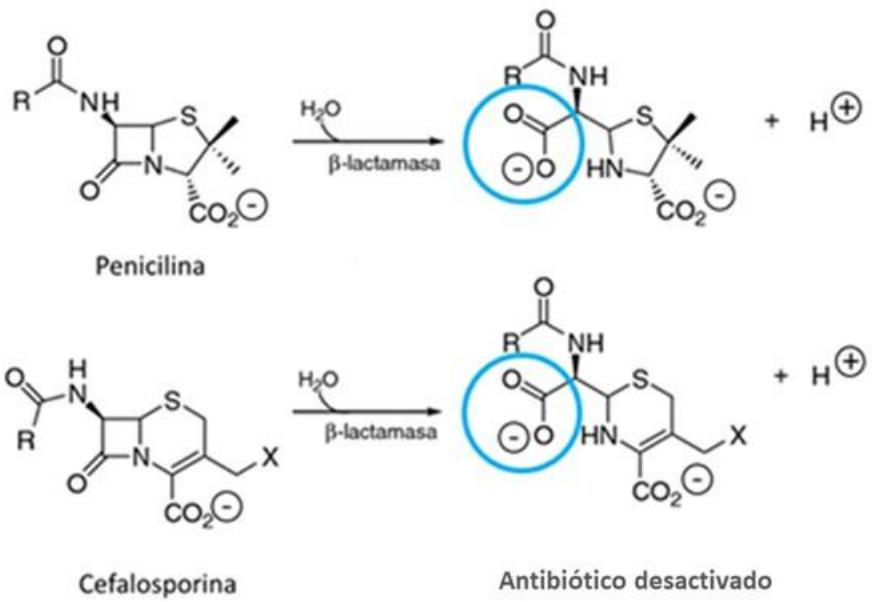


Figura 23. Esquema general de la reacción de hidrólisis de los β-lactámicos por las beta-lactamasas.

El producto es un ácido inactivo que pierde la acción antibiótica. El círculo azul rodea el grupo CO_2 resultante de la ruptura del anillo β-lactámico.

Se han identificado cuatro clases de beta-lactamasas, A, B, C y D, que se diferencian en su secuencia aminoacídica, aunque las clases A y C tienen el mismo origen evolutivo. Las beta-lactamasas de clase C contienen un residuo de serina en el sitio activo, implicado en la acción catalítica, en la formación de una acil—enzima con sustratos β-lactámicos. La isoforma de referencia empleada en este ejemplo es la beta-lactamasa de *Citrobacter freundii*, con código 1fr1 de la Protein Data Bank (Berman, y otros, 2000). La enzima madura consta de un homodímero, cada uno de 361 aminoácidos (Figuras 24B y 24D). El peso molecular se estima en 39.7 kDa y el pI predictivo es de 8.7. En conformidad con la tabla de hidrofobicidad de Sweet & Heisenberg (residuos VILFMWY), el porcentaje de aminoácidos hidrofóbicos es del 33%, es decir, la tasa común entre las proteínas globulares. Se ha determinado la estructura tridimensional refinada de la beta-lactamasa clase C de varias especies bacterianas. En la Figuras 24A y 24C se presenta una vista general de la isoforma de *Enterobacter cloacae* (código 1bls), la cual contiene un sustrato en el sitio activo.

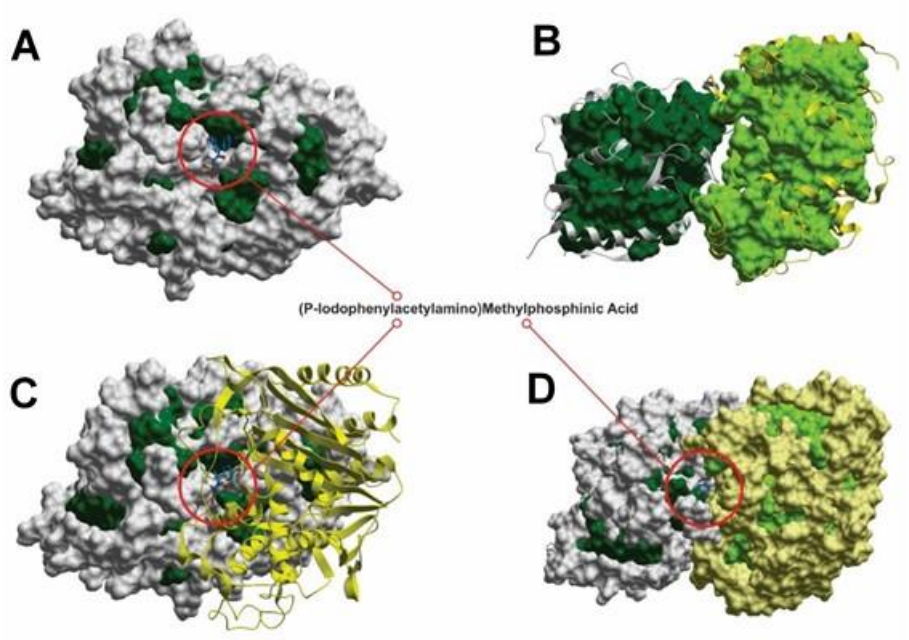


Figura 24. Vista general de la isoforma de *Enterobacter cloacae* (código 1bls), sustrato en el sitio activo

A. Vista superficial de la subunidad A de la beta-lactamasa clase C de Enterobacter cloacae (código 1bls) con el sitio activo ocupado por el antibiótico m-carboxifenil [[N-[(p-iodofenil)acetil]amino]metil]fosfonato (rodeado con un círculo rojo). La superficie de los aminoácidos se ha coloreado de gris, excepto los hidrofóbicos que están de color verde. B. La mayoría de las beta-lactamasas consta de dos subunidades idénticas. Se muestran los aminoácidos que hacen parte del corazón hidrofóbico interno de las dos subunidades de la beta-lactamasa clase C de Citrobacter freundii, con código 1fr1 (verde oscuro y claro para las subunidades A y B, respectivamente). C. Vista de superficie de la subunidad A y en hélices alfa y hojas beta de la subunidad B, de E. cloacae. El sitio activo de cada subunidad queda confinado en medio del dímero. D. Vista de superficie de las dos subunidades de C. freundii, donde se aprecia el estrecho

contacto entre las dos subunidades. La dimerización involucra algunos residuos hidrofóbicos y otros no hidrofóbicos.

De la Figura 24 se puede concluir que la proteína es globular y tiene diversos parches hidrofóbicos de superficie, que facilitan la interacción con las membranas (Bowden & Georgiou, 1990); (Ciofu, Beveridge, Kadurugamuwa, Walther-Rasmussen, & Høiby, 2000). Uno de los predictores de solubilidad más recientes y precisos (Hebditch, Carballo-Amador, Charonis, Curtis, & Warwicker, 2017) estima que la proteína es insoluble (*scaled solubility value* = 0,363/1,0) y esta propiedad se ha observado en la práctica (Bowden & Georgiou, 1990). Sin embargo, la isoforma de Chromohalobacter (bacteria halofílica moderada) se ha podido expresar de manera recombinante en *Escherichia coli*, aunque disuelta en solución salina baja (Tokunaga, Ishibashi, Arakawa, & Tokunaga, 2004).

En la base de datos SCOPe (Structural Classification of Proteins — extended), La beta-lactamasa (e.3.1.1 AMPC beta-Lactamase) se clasifica en la Clase e: Multi-domain proteins (alpha and beta). En la figura 25 se muestra la secuencia de la isoforma de referencia para este trabajo (*C. freundii*) con las respectivas estructuras secundarias. Se destacan en recuadros verdes todos los aminoácidos hidrofóbicos de la proteína, según la tabla de Sweet & Heisenberg (1983).

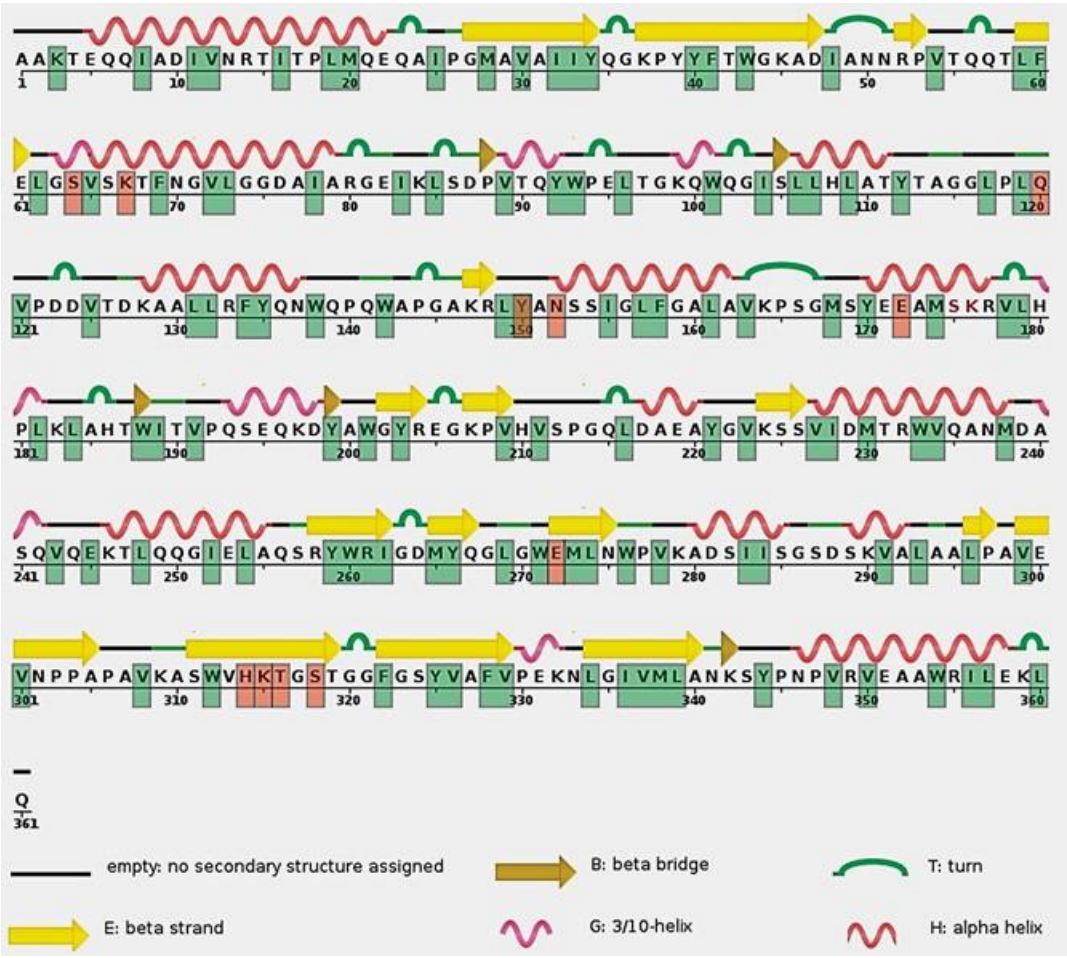


Figura 25. Secuencia aminoacídica de las subunidades A y B de la beta-lactamasa clase C de *Citrobacter freundii* (Código PDB 1fr1).

Todos los aminoácidos hidrofóbicos según la tabla de Sweet & Heisenberg (residuos VILFMWY) se encuentran encuadrados en color verde. Los residuos del sitio activo se encuadran en color anaranjado. Puede apreciarse que la proteína es una combinación de hélices alfa y hojas beta.

Con la secuencia de la beta-lactamasa clase C de *Citrobacter freundii* como referencia, se hizo una búsqueda con la herramienta BlastP del NCBI (*National Center for Biotechnology Information*) contra las bases de datos no redundantes de Genbank y la PDB (Protein Data Bank).

Se obtuvo un total de 38 secuencias ortólogas (i.e., la misma proteína en diferentes especies bacterianas), todas pertenecientes al Phylum Proteobacteria. El valor mínimo de identidad fue de 40% y el promedio de 55%. Con el fin de mostrar evidencia de las distancias entre las secuencias, se construyó un árbol de distancia. Como puede apreciarse en la Figura 26, las secuencias se agrupan en dos clados principales de acuerdo con la taxonomía bacteriana vigente.

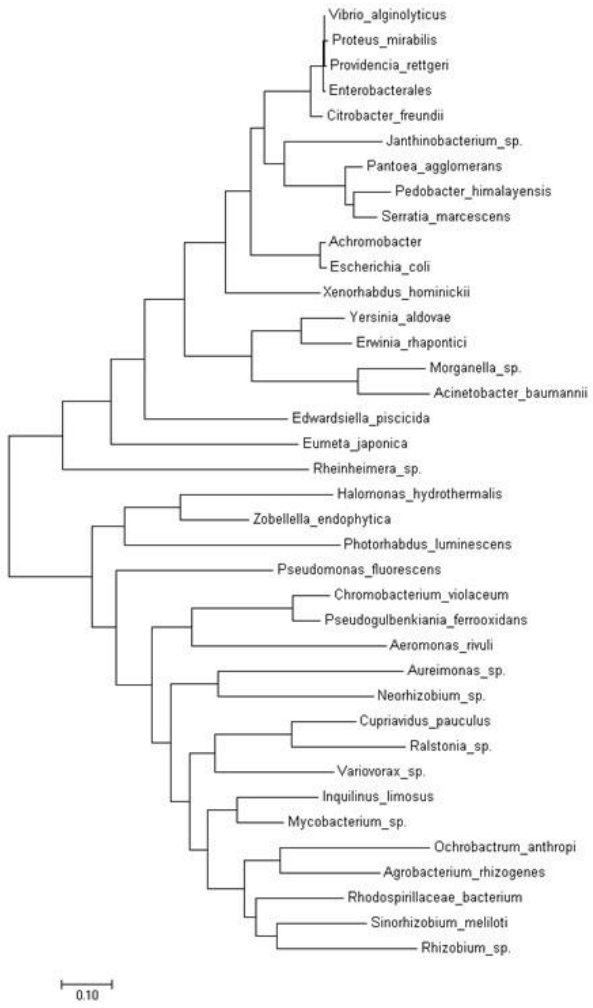


Figura 26. Árbol de distancia construido con las secuencias aminoacídicas de las beta-lactamasas de 38 géneros bacterianos.

El árbol fue construido a escala con la aplicación MEGA7 (Kumar et al., 2016) se basó en la matriz de distancias.

Con las 38 secuencias recopiladas, se construyó un consenso de los aminoácidos hidrofóbicos ultraconservados, con la herramienta desarrollada en este trabajo. El número de hidrofóbicos ultraconservados con una estringencia del 100% fue de 66 sobre 361 totales (i.e., 18%), distribuidos de manera homogénea en toda la longitud del consenso. Por otra parte, El número de hidrofóbicos ultraconservados con una estringencia del 90% fue de 88 (i.e., 24%). No obstante, al disminuir la estringencia a 89% surgen cuatro hidrofóbicos adicionales que pueden integrarse dentro del consenso hidrofóbico principal (Figura 27). De esta manera, el número total de residuos hidrofóbicos ultraconservados es de 92 (i.e., 25,4%).

En la Figura 28 se muestra una representación escalonada de las posiciones hidrofóbicas obtenidas entre 1 y 100% de estringencia. Esta información se obtuvo utilizando la opción '*Scaled hydrophobic consensus*' (menú '*Tools*'). Se puede observar que, a medida que disminuyen los porcentajes mínimos de secuencias con un aminoácido hidrofóbico en cada columna, surgen nuevas posiciones hidrofóbicas en el consenso del alineamiento. Las posiciones totalmente hidrofóbicas están resaltadas en columnas verdes, incluyendo las cuatro que surgen con 89% de estringencia. Este umbral se delimita con una línea horizontal verde. Por debajo de ese límite, solo hay columnas anaranjadas (rango de estringencia entre 1 y 89%), lo que significa la presencia de aminoácidos hidrofóbicos para unas proteínas, pero no para otras en cada vertical, dependiendo de la conservación entre las especies. En total hay 206 posiciones anaranjadas, pero solo el 27 superan el 50% de estringencia.

Por ejemplo, la tercera y cuarta columnas del alineamiento tienen posiciones hidrofóbicas anaranjadas; es decir, el 23% (9/38) y el 5% (2/38) de las secuencias tienen un aminoácido

hidrofóbico en esas posiciones verticales, respectivamente. De esta manera, la Figura 28 representa la relevancia estructural (hidrofóbica/no hidrofóbica) de cada posición en el alineamiento, para la conformación de cualquier beta-lactamasa. La conclusión principal es que existen tres tipos de posiciones hidrofóbicas: las que surgen con 100% de estringencia (i.e., estrictamente hidrofóbicas), las que surgen entre 89-90% de estringencia y las demás que están en el rango de 1 a 89-90%. En comparación con los datos publicados en la literatura científica, las posiciones entre 90% (89% en este caso) y 100% coinciden con el núcleo hidrofóbico de la proteína (*hydrophobic core*), calculado por otros métodos. Las posiciones por debajo de 90% son susceptibles de mutar hacia aminoácidos polares o con carga y no son fundamentales para la conformación compacta hidrofóbica, llamada en este trabajo el 'andamiaje hidrofóbico' o '*hydrophobic scaffold*'.

Figura 27. Captura de pantalla del consenso de aminoácidos hidrofóbicos ultraconservados, con un nivel de estringencia de 89%.

El consenso hidrofóbico, derivado de 38 secuencias de beta-lactamasa Clase C de Proteobacteria, se muestra en la parte inferior. Los guiones representan los gaps (i.e., residuo no existente en esa secuencia).

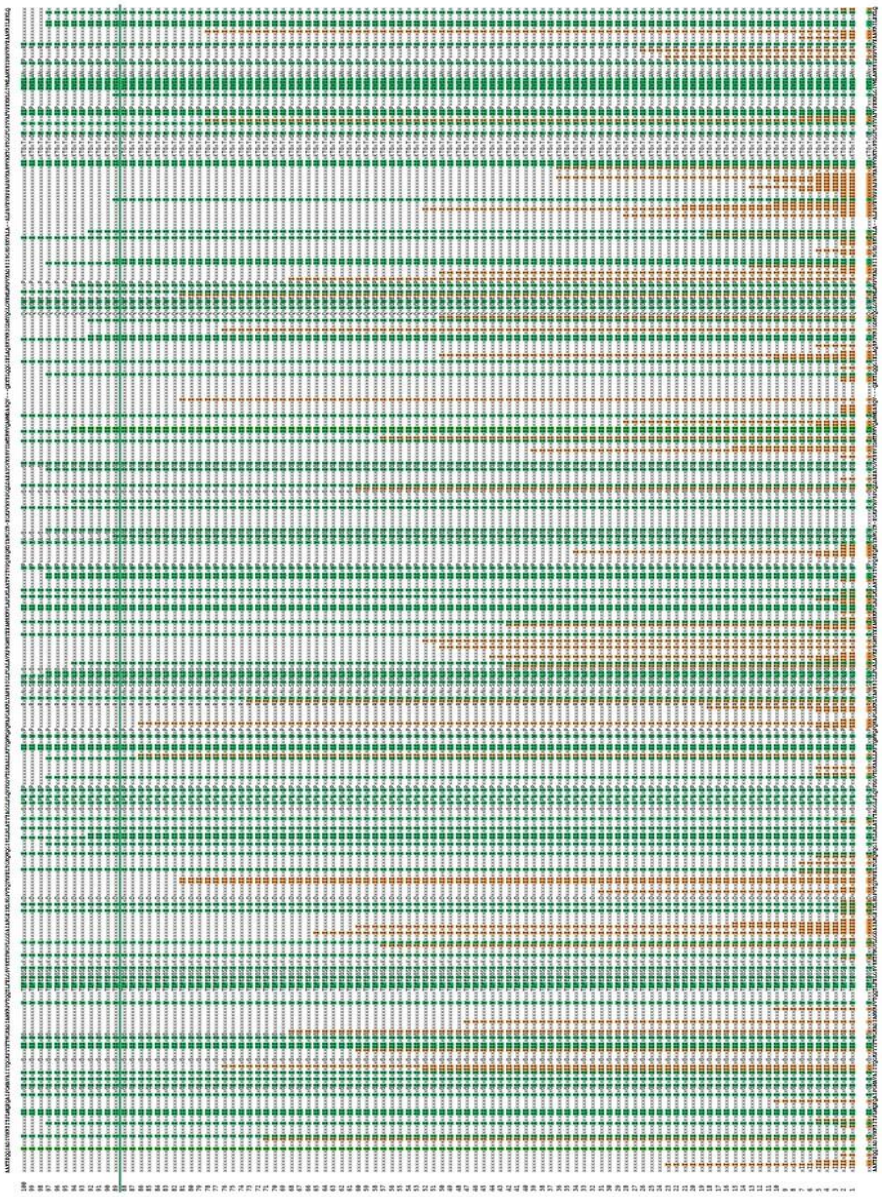


Figura 28. Representación escalonada de las posiciones hidrofóbicas obtenidas desde 1 a 100% de estringencia.

Para profundizar en esta importante conclusión, se recurrió a otra opción de la aplicación que arroja datos cuantitativos para graficar la tasa de residuos hidrofóbicos (aquí llamados 'cociente hidrofóbico'), en relación con los niveles de estringencia. En el menú '*Tools*', la opción '*Calculate stringency vs hydrophobic ratio*' genera un archivo en formato .scv como explicado anteriormente. Los datos obtenidos se graficaron como se representa en la Figura 29.

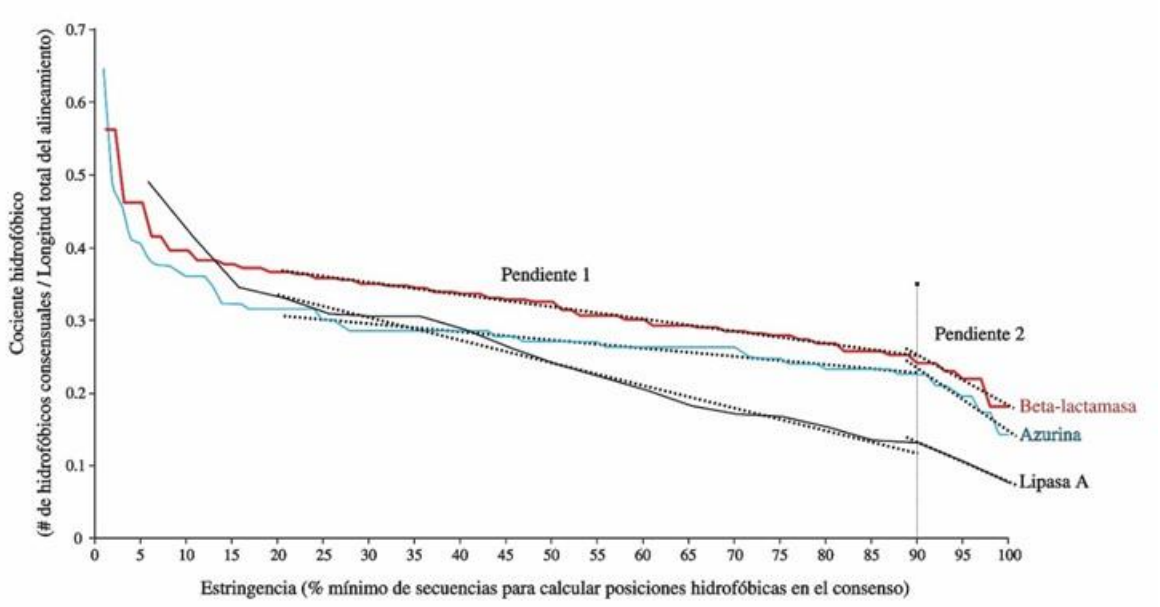


Figura 29. Cociente hidrofóbico (número de aminoácidos hidrofóbicos en el consenso sobre longitud total del alineamiento [367 posiciones, según '*alignment info*']) en función de la estringencia (porcentaje mínimo de secuencias que tienen un residuo hidrofóbico en una posición vertical, para construir el consenso hidrofóbico).

Como se aprecia en la Figura 29, la aparición de posiciones hidrofóbicas a medida que disminuye la estringencia, no es directamente proporcional. Por el contrario, sorprendentemente, surgen dos pendientes separadas por una inflexión justo alrededor del ≥90% de estringencia.

Desde el punto de vista molecular, esto quiere decir que el núcleo hidrofóbico de las proteínas beta-lactamasa, azurina y lipasa A (y muchas otras ya analizadas) está compuesto rigurosamente por ~10% del número de aminoácidos totales. Este es un descubrimiento importante hasta el momento desconocido para la comunidad científica. Surge una pregunta de orden molecular: ¿Por qué las posiciones ≥90% no son totalmente hidrofóbicas en lugar de tener variantes no hidrofóbicas? Nuestra hipótesis es que en las secuencias en las que esas posiciones han cambiado, lo han hecho hacia aminoácidos que forman puentes de hidrógeno y se camuflan con hidrofóbicos. (Baker & Hubbard, 1984, págs. 97–179)

De esta manera, la proteína ganaría cierta flexibilidad interna y además contribuiría a su solubilidad, ya que algunos de estos residuos (≥90%-99) también se expanden hacia la superficie (Tabla 2). Cabe reiterar que, interesantemente, en nuestros análisis comprobamos que la composición del núcleo hidrofóbico interno (*hydrophobic core*), calculado mediante otros métodos *in silico* como el *fuzzy oil drop model* (Banach, Konieczny, & Roterman, 2014), coincide perfectamente con los obtenidos mediante este método (datos no mostrados).

Seguidamente, con base en el consenso arrojado por nuestra aplicación, se procedió a representar gráficamente la estructura del núcleo hidrofóbico interno (exclusivamente 100% de estringencia) y superficial de la beta-lactamasa de *C. freundii* (Figura 30).

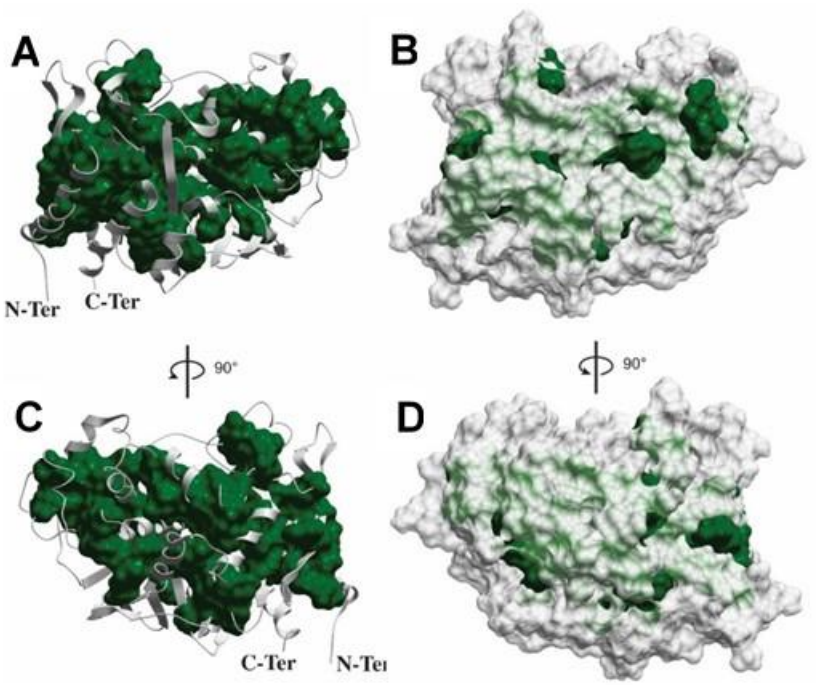


Figura 30. A y C, representación del núcleo hidrofóbico de la beta-lactamasa de *C. freundii* (código PDB 1fr1) conformado por los residuos del consenso con 100% de estringencia. B y D, representación superficial de la enzima.

Como se aprecia en las Figura 30A y 30C, los residuos hidrofóbicos ultraconservados, conforman un núcleo denso y compacto. El método fuzzy oil drop model (Banach, Konieczny, & Roterman, 2014) arroja los mismos resultados, pero con menor resolución, ya que solo reporta el núcleo hidrofóbico interno y excluye los residuos hidrofóbicos ultraconservados expuestos a la superficie, que también tienen función estructural (Figuras 30B y 30D). Posteriormente, a los residuos correspondientes al 100% de estringencia se añadieron los que surgen en el intervalo entre ≥89% y 99%; es decir, todos los que están cobijados por la pendiente 2 de la Figura 29. En la Figura 31, se representan gráficamente ambas categorías y se puede observar que la compactación se mantiene y fortalece, confiriendo la forma y estabilidad de la proteína.

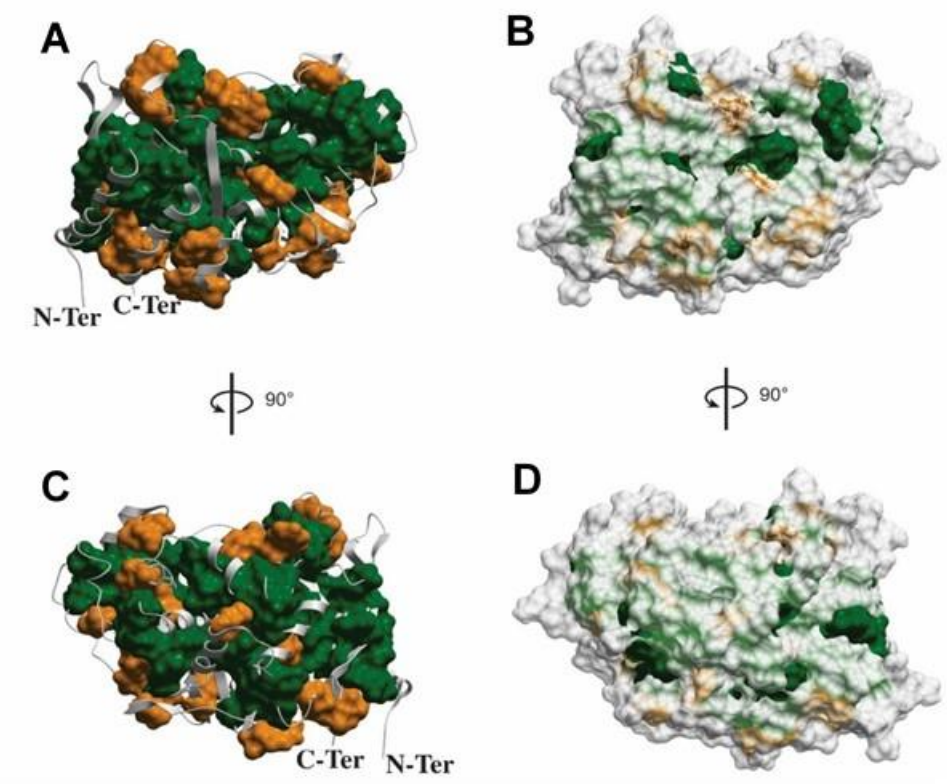


Figura 31. A y C, representación del núcleo hidrofóbico de la beta-lactamasa de *C. freundii* (código PDB 1fr1) conformado por los residuos del consenso con 89%-99% (anaranjados) y 100% (verdes) de estringencia. B y D, representación superficial de la enzima.

Como puede apreciarse en las Figuras 31A y Figura 31C, los aminoácidos de la fracción ≥89% y 99% (anaranjados) se ubican alrededor de los residuos ultraconservados (verdes), sin quebrantar la compactación. La superficie de la proteína (Figuras 31B y 31D) se torna ligeramente anaranjada, lo cual refleja la tendencia de algunos de esos aminoácidos (fracción 89%-99%) a proyectarse hacia la superficie (SASA = 0 a 6%) y siete de ellos alcanzan valores SASA entre 7 y 27% (Tabla 2). Interesantemente, se aprecian varios parches verdes superficiales que evidencian la exposición de residuos ultraconservados (100% de estringencia), que contribuyen a la conformación y la estabilidad de la enzima.

Tabla 2. *Apreciación de parches verdes superficiales que evidencian la exposición de residuos ultraconservados (100% de estringencia), que contribuyen a la conformación y la estabilidad de la enzima*

A

#	P	AA	%
1	8	ILE	0
2	12	VAL	0
3	16	ILE	0
4	19	LEU	2,5
5	20	MET	10
6	25	ILE	3,9
7	28	MET	0,1
8	30	VAL	0
9	32	ILE	0
10	40	TYR	25
11	41	PHE	17
12	43	TRP	18
13	54	VAL	5,9
14	59	LEU	0,8
15	60	PHE	0,1
16	62	LEU	0
17	65	VAL	0
18	69	PHE	0,1
19	73	LEU	0,8
20	83	ILE	2,1
21	85	LEU	12
22	96	LEU	1,8
23	101	TRP	0,8
24	104	ILE	1,8
25	106	LEU	0
26	107	LEU	12
27	109	LEU	0
28	112	TYR	1,5
29	117	LEU	1,5
30	119	LEU	32
31	121	VAL	12
32	125	VAL	5,8
33	131	LEU	5,7
34	134	PHE	7,9
35	135	TYR	0
36	138	TRP	6,6
37	150	TYR	11
38	155	ILE	0,1
39	157	LEU	1,1
40	158	PHE	0
41	161	LEU	1
42	170	TYR	5,2
43	174	MET	0,2
44	178	VAL	0

B

#	P	AA	%
45	179	LEU	5,6
46	182	LEU	9,3
47	184	LEU	1,8
48	188	TRP	24
49	189	ILE	12
50	191	VAL	11
51	199	TYR	11
52	203	TYR	4,8
53	209	VAL	14
54	211	VAL	25
55	216	LEU	7,9
56	221	TYR	17
57	223	VAL	0,5
58	230	MET	0,1
59	233	TRP	0
60	234	VAL	0
61	238	MET	5,6
62	248	LEU	0,4
63	252	ILE	0,8
64	259	TYR	2,2
65	260	TRP	2,1
66	265	MET	17
67	269	LEU	0
68	271	TRP	0
69	274	LEU	4,6
70	276	TRP	11
71	283	ILE	1,7
72	291	VAL	33
73	293	LEU	27
74	312	TRP	0
75	313	VAL	0
76	322	PHE	0,1
77	325	TYR	0
78	328	PHE	0
79	329	VAL	0
80	336	ILE	0
81	337	VAL	0
82	338	MET	0,8
83	339	LEU	0,2
84	344	TYR	0,7
85	350	VAL	0
86	356	ILE	0,1
87	357	LEU	0,2
88	360	LEU	5,9

C

#	P	AA	%
1	11	ILE	4,7
2	33	ILE	0,1
3	34	TYR	6,9
4	39	TYR	14,5
5	48	ILE	45,7
6	72	VAL	0
7	78	ILE	16,4
8	89	VAL	0
9	92	TYR	8,5
10	93	TRP	13,6
11	132	LEU	16,3
12	142	TRP	22,2
13	149	LEU	15,8
14	163	VAL	5,1
15	168	MET	13,2
16	201	TRP	34
17	227	VAL	0
18	228	ILE	17,9
19	243	VAL	4,6
20	254	LEU	23
21	262	ILE	5,1
22	266	TYR	12,8
23	273	MET	0,9
24	278	VAL	12,8
25	293	LEU	26,6
26	299	VAL	47
27	308	VAL	43,9
28	326	VAL	0
29	348	VAL	27,1
30	354	TRP	10,1

A y B, valores de superficie expuesta al solvente (SASA, Accessible Surface Area and Accessibility) de los aminoácidos del consenso, emanado de las fracciones 100% de estringencia (fondo verde) y 89%-100% (fondo anaranjado), respectivamente, para la beta-lactamasa de C. freundii (código PDB 1fr1,

cadena A). Las posiciones hidrofóbicas de los consensos 1%-88% se representan sobre fondo celeste (C). Símbolos: #: número consecutivo de aminoácido (88 verdes y 30 anaranjados, respectivamente). P: posición del residuo dentro de la proteína. AA: aminoácido. %: Porcentaje de exposición del residuo al solvente. Los valores SASA se obtuvieron con el servicio online del Center for Informational Biology de la Universidad de Ochanomizu (Japón). http://cib.cf.ocha.ac.jp/bitool/ASA/

La última categoría de posiciones hidrofóbicas surge entre 1 y 88% de estringencia. Obviamente, estos sitios no son compartidos por la mayoría de las secuencias entre las especies, así que la presencia y el rol estructural de estos residuos debe estudiarse individualmente. Como puede observarse en la Tabla 2, solo 4 posiciones de 30 (Val72, Val89, Val227 y Val326) presentan un valor SASA de 0, es decir, están totalmente ocultos del solvente. Hay que precisar que valina es el aminoácido más hidrofóbico según la escala de Sweet & Heisenberg (1983), por lo cual su valor SASA es correcto. Los demás residuos de esta categoría presentan valores SASA variables que alcanzan el 46%. En la Figura 31 se representan las superficies de los aminoácidos hidrofóbicos que ocupan estas posiciones, para la beta-lactamasa de *C. freundii* (código PDB 1fr1, cadena A).

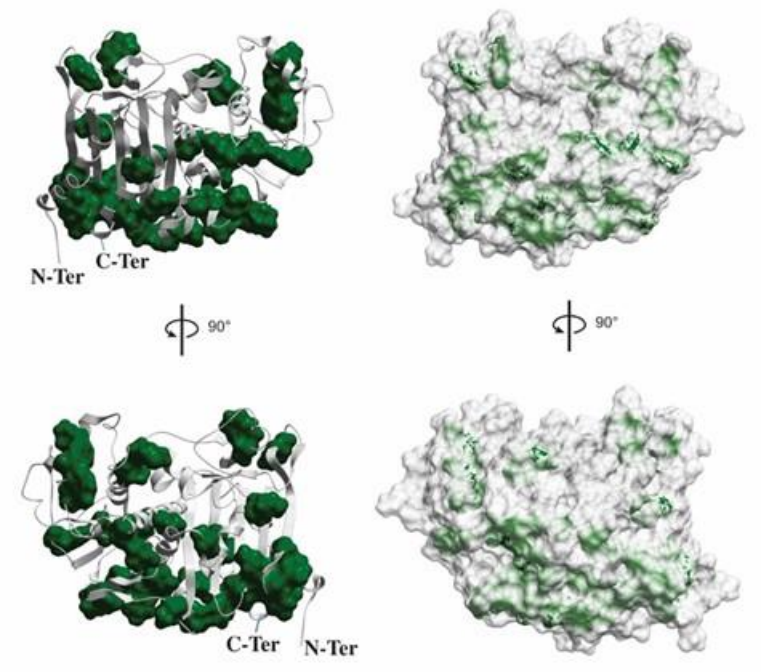


Figura 32. A y C, representación de superficie de los aminoácidos (verdes) que conforman la fracción 1-88% de estringencia de la beta-lactamasa de *C. freundii* (código PDB 1fr1). B y D, representación superficial de la enzima.

Uno de los métodos bioinformáticos más empleados actualmente para simular el comportamiento de las proteínas en solución es la dinámica molecular. Este método calcula *in silico* el movimiento y la deformación de las moléculas en un ambiente predeterminado por el investigador, durante un tiempo que normalmente oscila entre 10 y 50 ns. (Alder & Wainwright, 1959, págs. 459–466)

En lo que concierne el análisis estructural de las proteínas, el método emplea las estructuras 3D (formato .pdb) determinadas en el laboratorio (e.g., por difracción de rayos X de cristales proteicos) y alojadas en las bases de datos. La Protein Data Bank (PDB) es la base de datos pública y más completa a disposición de los científicos. (Berman, y otros, 2000)

Prácticamente, la PDB contiene las estructuras de todas las proteínas determinadas a la fecha. Por ejemplo, a la estructura de la beta-lactamasa de *C. freundii*, se le asignó el código PDB 1fr1. Un archivo con formato .pdb puede abrirse con un editor de texto y lo que se visualizará serán las coordenadas espaciales X, Y y Z de cada átomo de la proteína. La estructura 1fr1.pdb (beta-lactamasa de *C. freundii* tipo salvaje) fue descargada de la PDB y simulada con dinámica molecular (Figura 33).

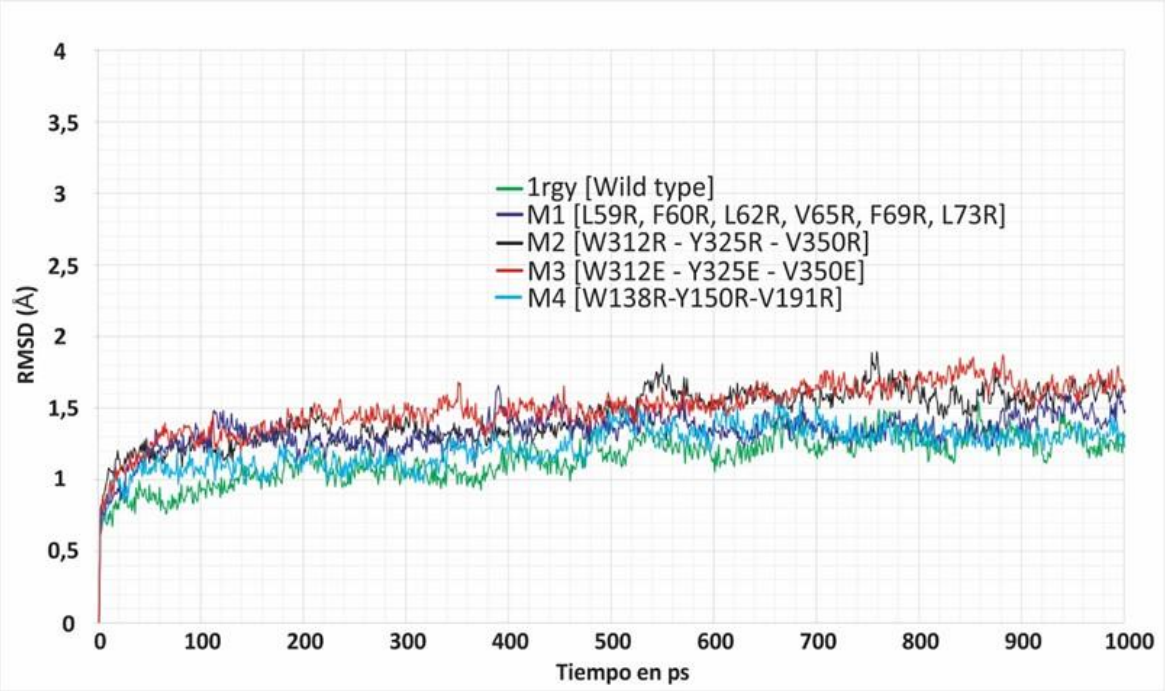


Figura 33. Simulación de dinámica molecular (1 ps) de la beta-lactamasa tipo salvaje de *C. freundii* (código PDB 1rgy, '*wild type*') y cuatro mutantes modelados *in silico*.

En la Figura 33 se aprecia el comportamiento de la beta-lactamasa tipo salvaje en solución acuosa y a temperatura ambiente (trayectoria verde). Se simuló la deformación o estabilidad de la enzima durante 1 ps y, dada la ausencia de fluctuaciones (señal estable del orden de 1,25 Å), se

concluyó que la proteína no sufre cambios estructurales significativos durante 1 ps. La magnitud de estas deformaciones se mide mediante con el RMSD (*Root Mean Square Deviation* o desviación de la media cuadrática). Esto es, qué tanto se alejan los átomos de las posiciones originales, en condiciones ideales (variación medida en nm o Å).

Interesantemente, con herramientas bioinformáticas se pueden alterar los archivos .pdb para construir mutantes de posiciones específicas. Para ello, existen servidores de modelamiento en línea que ofrecen gratuitamente este servicio (Waterhouse & et.al., 2018); (Yang, y otros, 2014). Un tipo de mutación sería, por ejemplo, un modelo .pdb de la beta-lactamasa que no contuviera valina en posición 72 sino glicina (Val72Gly). Posteriormente, con los programas de simulación de dinámica molecular se pueden evaluar, *in silico,* las consecuencias de las mutaciones en la estabilidad estructural (Raval, Piana, Eastwood, Dror, & Shaw, 2012). Los efectos son variables, desde una pequeña deformación durante el tiempo hasta una pérdida irreversible de la estructura.

Con la aplicación Swiss-Model (Waterhouse & et.al., 2018) se construyeron tres mutantes de la beta-lactamasa de *C. freundii* en estado monomérico (i.e., una sola unidad, código PDB 1rgy), direccionados a cambiar aminoácidos hidrofóbicos ultraconservados elegidos al azar, por los siguientes residuos:

Mutante 1: Leu59Arg, Phe60Arg, Leu62Arg, Val65Arg, Phe69Arg, Leu73Arg. En el tipo salvaje, estos aminoácidos se agrupan internamente (SASA=0) y hacen parte del núcleo hidrofóbico o '*hydrophobic core*' (i.e., surgen con 100% de estringencia). arginina ('Arg' o 'R') es un aminoácido de carga positiva.

Mutante 2: Trp312Arg, Tyr325Arg, Val350Arg. En el tipo salvaje, estos aminoácidos hidrofóbicos ultraconservados también se contactan internamente y, a la par, conforman el núcleo hidrofóbico (SASA=0). Las mutaciones también se hicieron hacia el aminoácido arginina.

Mutante 3: Trp312Glu, Tyr325Glu, Val350Glu. Se mutaron las mismas posiciones hidrofóbicas ultraconservadas del mutante 2, por el aminoácido con carga negativa Glutamato (‘Glu’ o ‘E’).

Mutante 4: Trp138Arg, Tyr150Arg, Val191Arg. En el tipo salvaje, estos aminoácidos, aunque también ultraconservados, afloran ligeramente a la superficie (Valores SASA de 6,6%, 11% y 11%, respectivamente).

En la Figura 33 se puede apreciar el efecto en la estabilidad estructural de la beta-lactamasa (dinámica molecular), debido a la sustitución de los aminoácidos hidrofóbicos por residuos cargados. Es necesario precisar que la predicción de las consecuencias moleculares derivadas de cada uno de los mutantes va mucho más allá del alcance de este trabajo. De por medio está la consideración de una plétora de variables fisicoquímicas, termodinámicas, conformacionales, funcionales, etc. y, finalmente, pruebas de laboratorio. No obstante, la sola fluctuación es un primer indicador de la viabilidad estructural y, en última instancia, funcional de la enzima (Maiorov & Crippen, 1994); (Wei, Huang, & Altman, 1999); (Huang, Kalyanaraman, Bernacki, & Jacobson, 2006).

Durante 1 ps, la trayectoria de los mutantes fluctúa en un rango de 0.5 Å. El mutante de mayor fluctuación es el 3 (línea roja), el cual contiene residuos con valor SASA=0 y mutados hacia

glutamato. Le siguen los mutantes 2 y 1 (líneas negra y azul, respectivamente) y finalmente el mutante 4, que contiene residuos que afloran ligeramente a la superficie y fueron mutados hacia arginina. Grosso modo, la tendencia de las trayectorias es hacia el ascenso gradual, por lo que nuevas simulaciones con mayores tiempos podrían revelar los valores RMSD de estabilización. No obstante, en todos los casos, se observaron fluctuaciones mayores que el tipo salvaje. El efecto deletéreo de una o más mutaciones depende de cada proteína. Para citar solo un ejemplo, el mutante P722S (prolina por serina, posición 722), con 1.21 Å de fluctuación, es suficiente para desactivar la proteína FGFR1 (Doss, y otros, 2012, págs. 37–43). No obstante, valores más bajos también pueden tener el mismo efecto.

La plataforma CABSflex2 realiza simulaciones de dinámica molecular, pero, a diferencia de Gromacs que calcula las fluctuaciones de cada átomo de la proteína en relación con el tiempo, CABSflex2 lo hace con respecto a un punto espacial predeterminado (Kuriata, y otros, 2018). En la Figura 33 se correlacionan cuatro parámetros de la beta-lactamasa: el perfil de hidropatía con base en la escala de Sweet & Heisenberg (1983), el área de superficie accesible al solvente (SASA), la dinámica molecular de los aminoácidos del tipo salvaje y la dinámica molecular del mutante 1. Resulta oportuno recordar que un análisis exhaustivo de resultados como los de la Figura 33 puede dar para una tesis aparte. No obstante, nos referiremos al impacto general de las mutaciones. Puede observarse que las mutaciones de posiciones hidrofóbicas ultraconservadas (recuadro vertical gris) recaen en residuos de un segmento hidrofóbico (WT S&E) con valor SASA=0 (i.e., cero exposición al solvente) (WT SASA) y con escasa fluctuación (WT RMSF). Aunque algunos residuos del mismo segmento sufren mínimas fluctuaciones en el mutante, otros residuos de la proteína (flechas negras) fluctúan del simple al doble. Este importante resultado puede interpretarse con una visión holística del polipéptido, en el sentido de que una alteración de

uno de los componentes afecta el todo, aún en los sitios más inesperados. Estas fluctuaciones, de acuerdo con la dinámica molecular arrojada por Gromacs, podrían desestabilizar toda la proteína e impedir su correcta conformación y finalmente su actividad biológica. La posición interna de los aminoácidos mutados se representa gráficamente en la Figura 34.

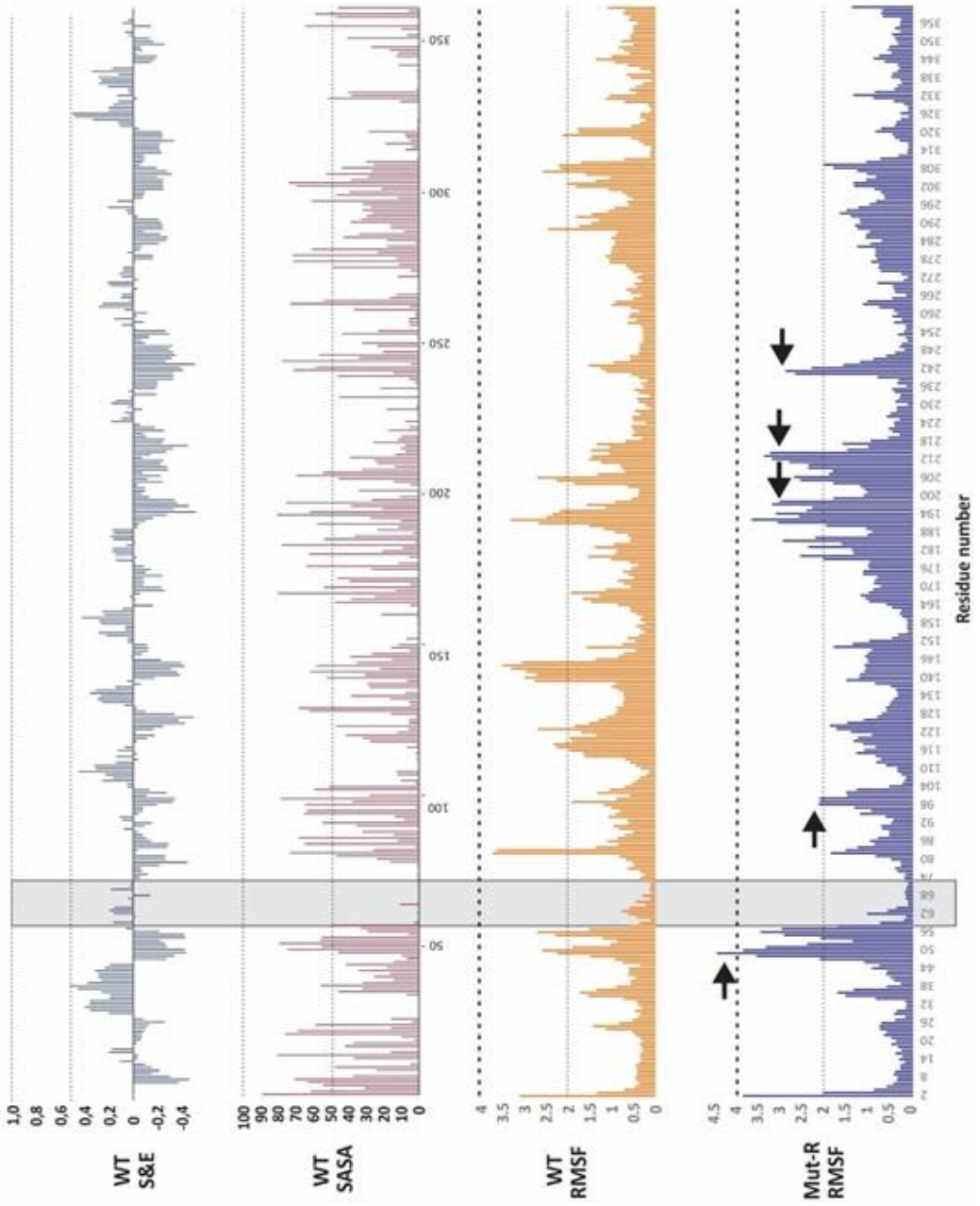


Figura 34. Perfil de hidropatía con base en la escala de Sweet & Heisenberg (1983) (WT S&E); SASA (*solvent-accessible surface área*) de los aminoácidos del tipo salvaje (WT SASA); dinámica molecular del tipo salvaje (WT RMSF) y del mutante L59R, F60R, L62R, V65R, F69R y L73R (Mut-R RMSF).

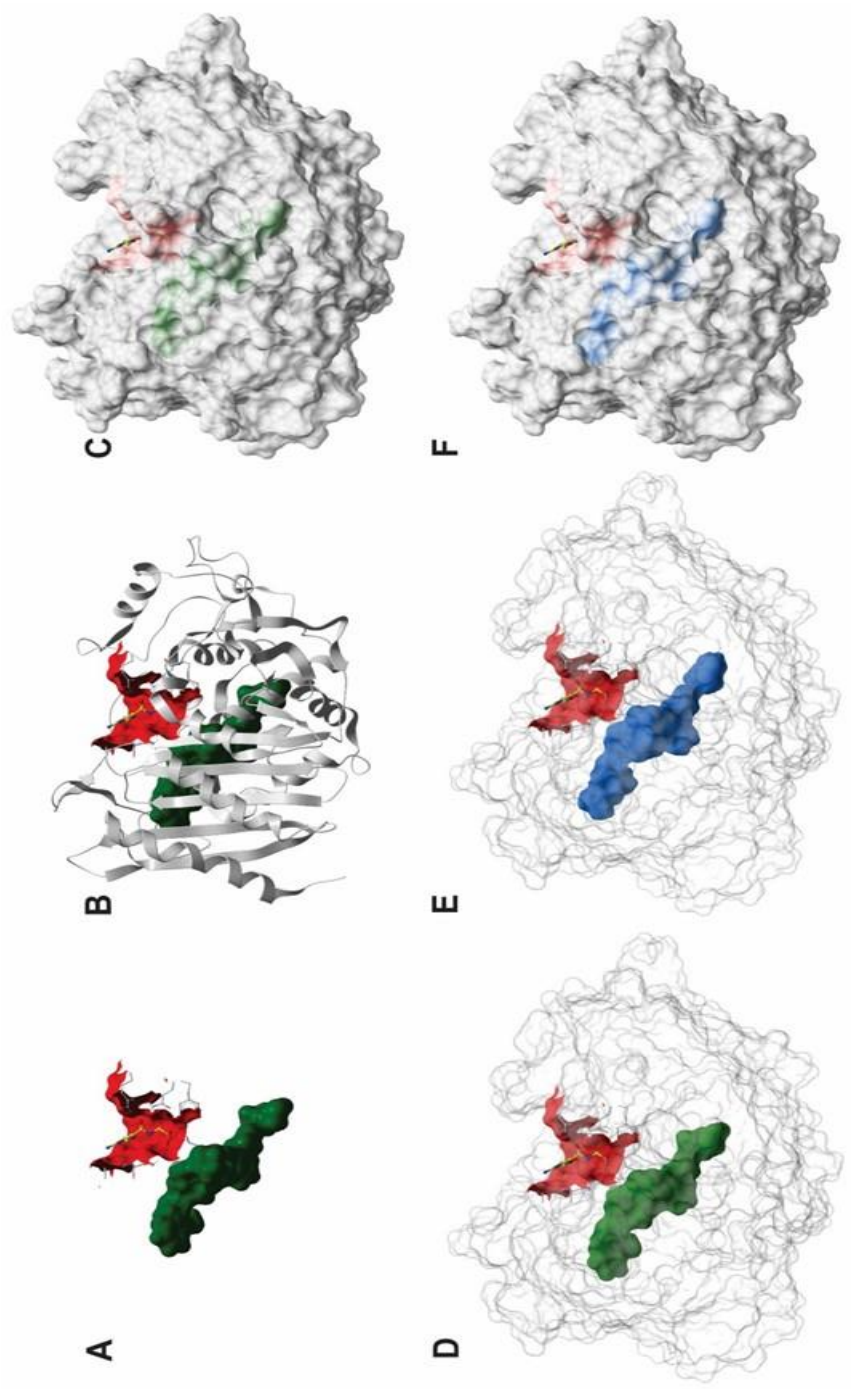


Figura 35. Representación gráfica de la ubicación interna de los aminoácidos L59R, F60R, L62R, V65R, F69R y L73R mutados por arginina. En color verde se representa la superficie de los residuos hidrofóbicos y, en azul, las mismas posiciones ocupadas por argininas.

Por otra parte, los mutantes 2 y 3 fluctúan un poco menos que mutante 1 y, además, se observa una estabilización de la señal alrededor de 0.3 nm de RMSD (Figura 32). Este resultado es interesante porque las mutaciones de los residuos Trp312, Tyr325 y Val350, aun siendo ultraconservados, parecen no afectar la estabilidad de la proteína. Sin embargo, como se puede apreciar en la Figura 35, si bien es poca la fluctuación alrededor de los sitios de las mutaciones, la trayectoria de otros aminoácidos se ve notablemente alterada (flechas negras). Solo una evaluación exhaustiva bioinformática y molecular podría dilucidar el efecto sobre la estabilidad y sobre la estructura de la beta-lactamasa. En cualquier caso, la trayectoria no es la misma del tipo salvaje y por lo tanto sí hay un impacto estructural por el hecho de mutar dichas posiciones.

Finalmente, el mutante 4 resulta de gran interés porque las predicciones de dinámica molecular sugieren un efecto estabilizador de las mutaciones. Recordaremos que los aminoácidos Trp138, Tyr150 y Val191, aunque también ultraconservados, afloran ligeramente a la superficie. En ese sentido, los resultados son muy satisfactorios porque mutaciones de estas posiciones hacia residuos con carga deberían tener un efecto solubilizante en contacto con el solvente. En la Figura 36 se observa una menor fluctuación de los aminoácidos situados entre los residuos Trp138 y Tyr150, lo que es coherente con las predicciones de dinámica molecular de Gromacs. Por el contrario, otras posiciones, señaladas con flechas, tienden a fluctuar más que el tipo salvaje, siendo impredecibles por ahora los efectos sobre la función de la beta-lactamasa.

Tomando en conjunto los resultados obtenidos para la beta-lactamasa, queda claro que la dinámica conformacional de las proteínas es un asunto complejo, pero que su entendimiento pasa por la discriminación de los residuos hidrofóbicos que contribuyen a la geometría de la enzima. El método de análisis basado en la información que arroja la aplicación diseñada en este trabajo ha permitido dilucidar el núcleo hidrofóbico de una vasta cantidad de proteínas en nuestro

laboratorio, incluyendo la beta-lactamasa (Figura 37). Sin embargo, no solo nos hemos restringido a lo predictivo. Por el contrario, hemos construido mutantes basados en estas conclusiones y la coherencia entre predicciones *in silico* y resultados de laboratorio ha sido muy satisfactoria. En vista de que la aplicación aquí diseñada y construida contribuye a desvelar el andamiaje hidrofóbico de las proteínas, le hemos nombrado el nombre de JHydroScaffold. El siguiente paso en la sería integrarla a un programa de mayor robustez, que traslade los consensos directamente sobre archivos PDB para visualizar el núcleo hidrofóbico en tiempo real, después de un alineamiento de rutina.

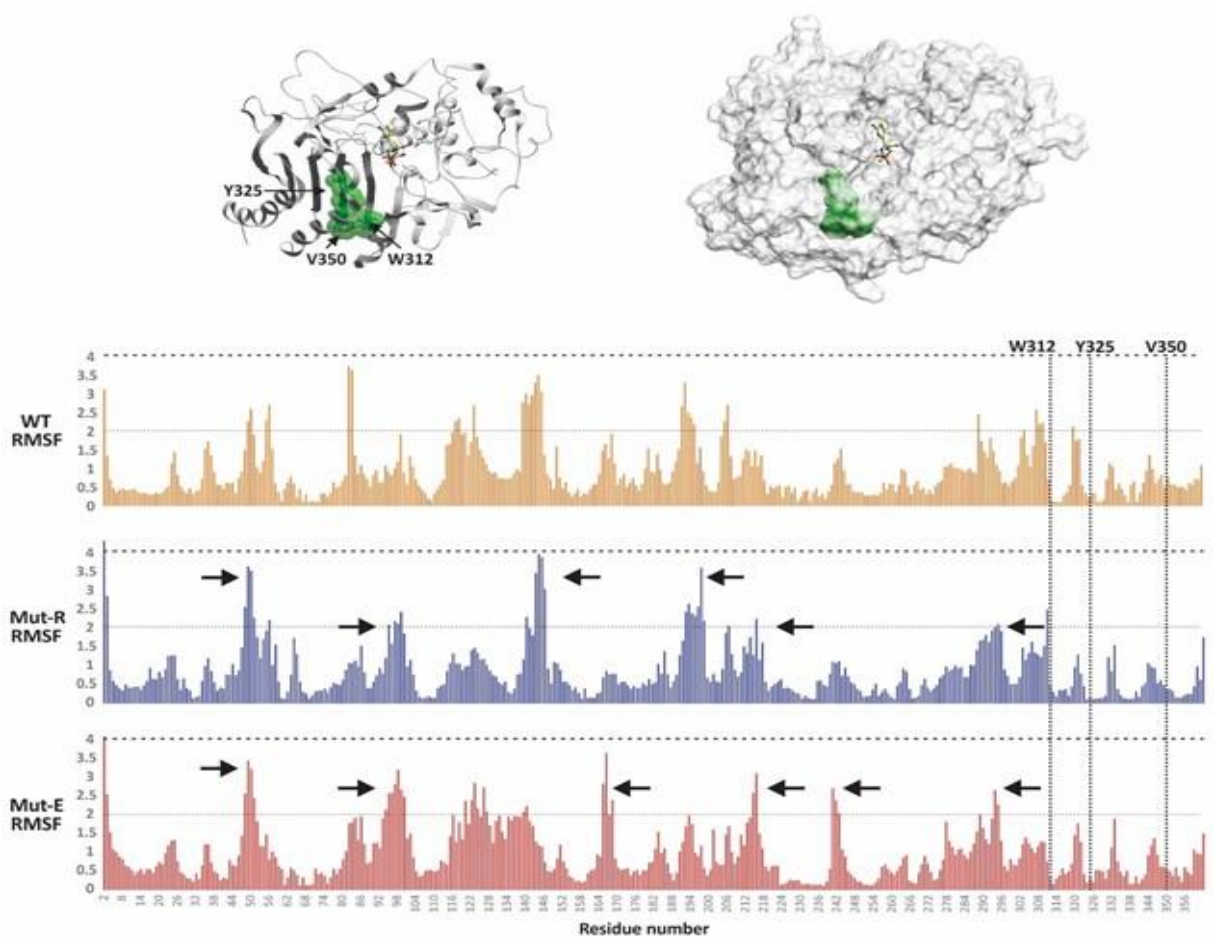


Figura 36. Dinámica molecular de la beta-lactamasa (1rgy) del tipo salvaje (WT RMSF), mutante 2 (Mut-R RMSF [Trp312Arg, Tyr325Arg, Val350Arg]) y mutante 3 (Mut-E RMSF [Trp312Glu, Tyr325Glu, Val350Glu]).

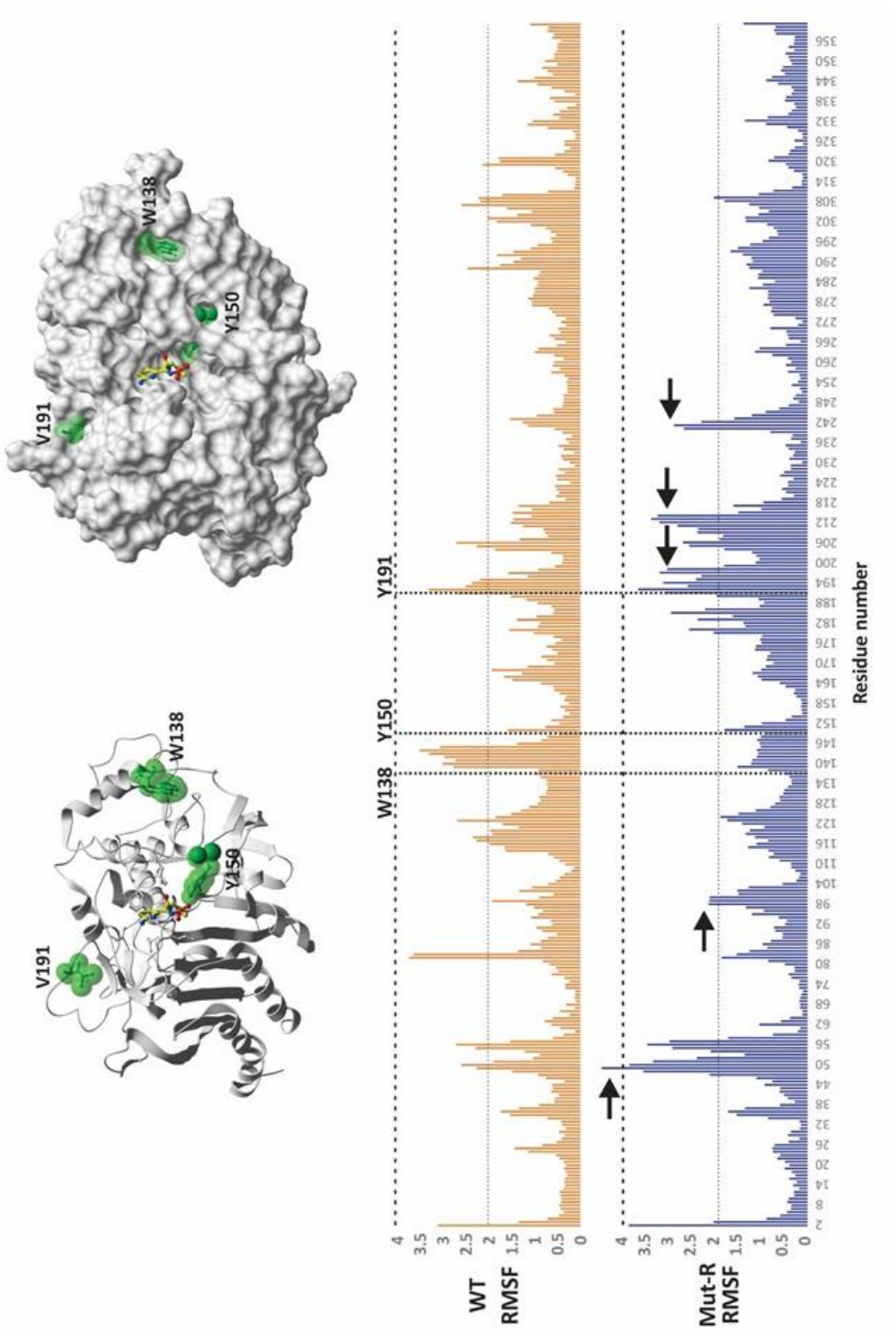


Figura 37. Dinámica molecular de la beta-lactamasa (1rgy) del tipo salvaje (WT RMSF) y del mutante 4 (Mut-R RMSF [W138R, Y150R, V191R]

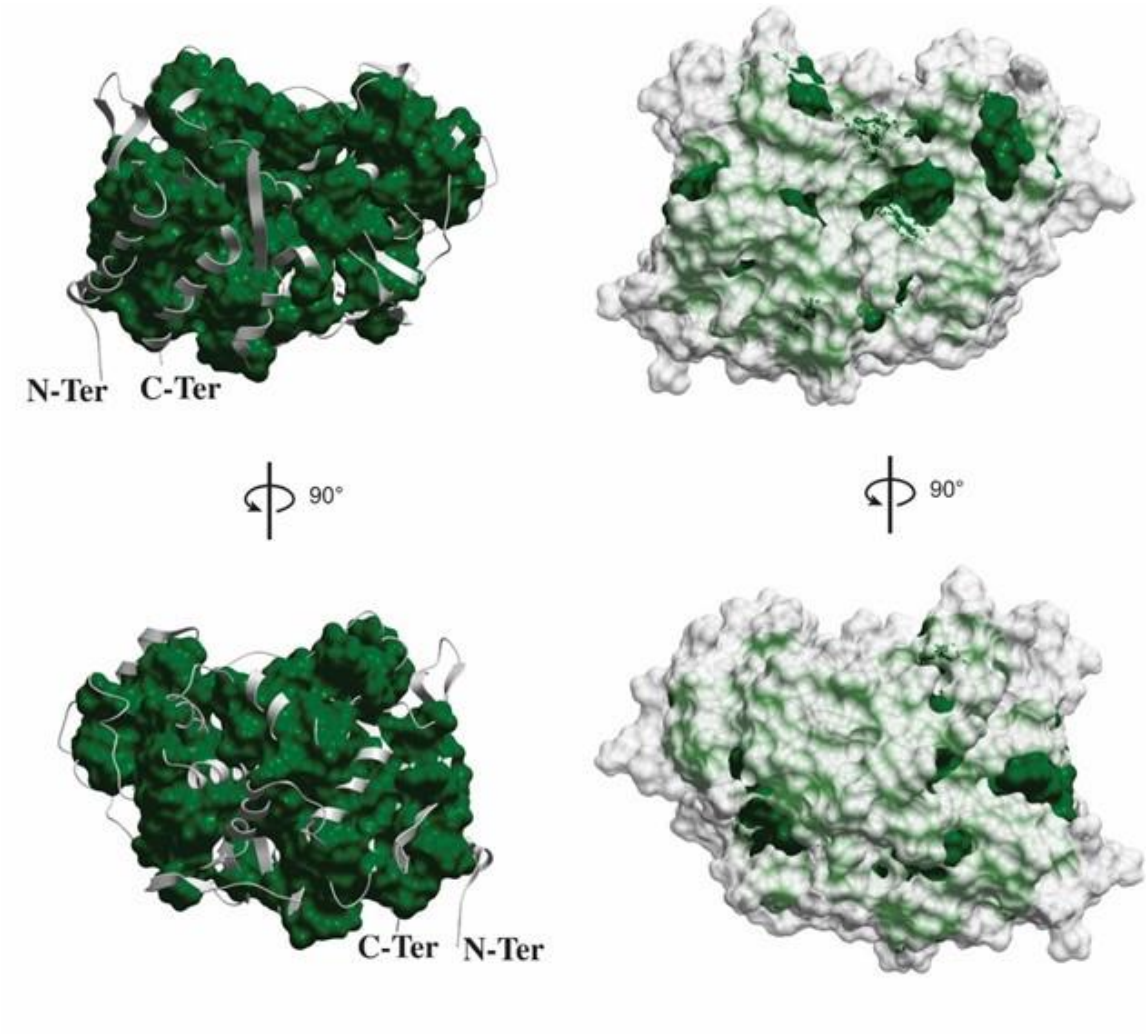


Figura 38. Estructura y representación de superficie del núcleo hidrofóbico de la beta-lactamasa tipo salvaje de *C. freundii* (código PDB 1rgy), basada en los resultados obtenidos con la aplicación desarrollada en este trabajo de grado.

6. Conclusiones

- JHydroScaffold es una solución innovadora que permite establecer el consenso de los aminoácidos hidrofóbicos ultraconservados de una proteína y ofrece herramientas hasta ahora inexistentes para el análisis, manipulación, exportación y búsqueda de consensos derivados de alineamientos múltiples de secuencias.

- El algoritmo de entrada de datos provenientes de la herramienta ClustalO cuenta con un controlador autónomo e independiente, ejecutándose rápidamente en el *background* de tal forma que no confunde o distrae al investigador. La interfaz gráfica tiene un diseño intuitivo y sus botones traen implementada la función *Tooltip*, que muestra una breve descripción de cada botón.

- El algoritmo para la clasificación de los aminoácidos del alineamiento, en las categorías hidrofóbicos y no hidrofóbicos, fue optimizado al máximo reduciendo su complejidad algorítmica e iteraciones innecesarias. Al mismo tiempo, el algoritmo que modifica el consenso en tiempo real fue construido utilizando un modelo que redujera los tiempos de espera entre cada porcentaje de identidad asignado.

- Las herramientas para exportación, visualización, búsqueda y análisis computacional fueron diseñadas para soportar *multi-threading*, de tal forma que se ejecuten independientemente de los procesos que se llevan a cabo en la interfaz principal. Esto le da una increíble sensación de velocidad a la herramienta.

- La interfaz principal cuenta con una unión de controladores independientes, lo que los hace rápidos, accesibles y escalables. La representación gráfica se diseñó de la manera más simple posible, de tal forma que fuera intuitiva y fácil de usar para el investigador.

- Se diseñaron varias rutinas de prueba para todas las herramientas de la aplicación, pero no está comprobado su correcto funcionamiento si están en ejecución al mismo tiempo (Ej: cambio en el *slider*, cambio de tabla, gaps activos, etc.). Esto se debe a que el número de combinaciones de los casos de usos de todos los elementos es demasiado grande para probarlos manualmente.

- La información que provee la aplicación JHydroScaffold será de utilidad para el estudio de las proteínas desde un nuevo enfoque. JHydroScaffold ostenta un alto potencial de expansión y desarrollo en el futuro de la bioinformática.

7. Recomendaciones

- Se recomienda depurar el listado de secuencias (e.g., fijando un umbral máximo de identidad), para alinear la menor cantidad posible. De no ser el caso, es una nueva oportunidad de desarrollo de una herramienta de computación en paralelo o de supercomputación, que sea capaz de soportar un volumen de entrada de datos más grande.

- Esta herramienta fue desarrollada en el *framework* "BlueJ versión 4.2.1" y ejecutada con la versión del "*Java SE Development Kit 12*". Se recomienda obtener la misma versión o una superior de ambas herramientas para el correcto desarrollo y ejecución de la aplicación.

- Para abrir la aplicación, se ejecuta el archivo llamado "JHydroScaffold.jar". De tener la versión del *Kit* mencionada anteriormente, se lanzará automáticamente como si fuera un archivo ejecutable.

8. Limitaciones y problemas

8.1. Herramientas de la sección 4.6.2.

Los *tools* de la sección 4.6.2 sólo están disponibles, por ahora, para la escala hidrofóbica de Sweet & Heisenberg (1983). Esto se debe a que cada herramienta trabaja conjuntamente con una escala en particular y toma gran complejidad y tiempo desarrollarlas.

Fue imposible construir un modelo que permitiera estandarizar dichas herramientas para que funcionaran con todas las escalas al mismo tiempo.

8.2. Depuración y cantidad máxima de secuencias alineables

Actualmente existe una limitación por parte del hardware y la capacidad de procesamiento. Se desconoce la cantidad máxima de secuencias que puede alinear este software. Es posible que cantidades muy grandes de secuencias tomen mucho tiempo en alinear.

8.3. Librería Itext y uso comercial

La generación de archivos .pdf fue construida con base en una librería hecha por terceros. Para poder utilizarla gratuitamente, los desarrolladores especifican que la aplicación debe ser de código abierto, bajo una licencia AGPL. El tiempo que tomaría tramitar una negociación y cotización con los dueños de la librería hubiese comprometido el desarrollo de la herramienta. En consecuencia, de querer lanzar la aplicación al mercado, sería necesario deshabilitar esta funcionalidad temporalmente.

Referencias bibliográficas

Alder, B. J., & Wainwright, T. E. (1959). Studies in molecular dynamics. I. General method. *The Journal of Chemical Physics*, 31(2): 459–466. Disponible en: https://doi.org/10.1063/1.1730376

Baker, E. N,& Hubbard, R.E. (1984). Hydrogen bonding in globular proteins. *Progress in Biophysics and Molecular Biology*, 44(2): 97–179. Disponible en: https://doi.org/10.1016/0079-6107(84)90007-5

Banach, M., Konieczny, L., & Roterman, I. (2014). The fuzzy oil drop model, based on hydrophobicity density distribution, generalizes the influence of water environment on protein structure and function. *Journal of Theoretical Biology*, 359: 6–17. Disponible en: https://doi.org/10.1016/j.jtbi.2014.05.007

Banach, M., Prymula, K., Konieczny, L., & Roterman, I. (2011). "Fuzzy oil drop" model verified positively. *Bioinformation,* 5(9): 375–377. Disponible en: https://doi.org/10.1007/s00894-011-1033-4.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2017). GenBank. *Nucleic Acids Research*, 45(D1): D37–D42. Disponible en: https://doi.org/10.1093/nar/gkw1070

Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N. & Bourne P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1): 235–242. Disponible en: https://doi.org/10.1093/nar/28.1.235

Bowden, G. A., & Georgiou, G. (1990). Folding and aggregation of β-lactamase in the

periplasmic space of *Escherichia coli*. *Journal of Biological Chemistry*, 265(28): 16760–16766. Disponible en: http://www.jbc.org/content/265/28/16760.long

Brenner, S. E., Chothia, C., & Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11): 6073–6078. Disponible en: https://doi.org/10.1073/pnas.95.11.6073

Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., & Henrissat, B. (1997). Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cellular and Molecular Life Sciences CMLS*, 53: 621–645. Disponible en: https://doi.org/10.1007/s000180050082

Carpentier, M., & Chomilier, J. (2019). Protein Multiple Alignments: Sequence-based vs Structure-based Programs. *Bioinformatics*, 35(20): 3970–3980. Disponible en: https://doi.org/10.1093/bioinformatics/btz236

Ciofu, O., Beveridge T.J., Kadurugamuwa J., Walther-Rasmussen J. & Høiby N. (2000). Chromosomal beta-lactamase is packaged into membrane vesicles and secreted from *Pseudomonas aeruginosa*. *Journal of Antimicrobial Chemotherapy*, 45(1): 9–13. Disponible en: https://doi.org/10.1093/jac/45.1.9

Doss, C.G., Rajith, B., Garwasis, N., Mathew, P.R., Raju, A.S., Apoorva, K., William, D., Sadhana, N.R., Himani, T. & Dike, I.P. (2012). Screening of mutations affecting protein stability and dynamics of FGFR1-A simulation analysis. *Applied & translational genomics*, 3(1):37–43.

Eisenhaber, F., & Argos, P. (1994). Hydrophobic regions on protein surfaces : definition based on hydration shell structure and a quick method for their computation. *Protein engineering*, 9(12): 1121–1133.

Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., & Bairoch A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press. pp. 571–607.

Gowder, M.S., Chatterjee, J., Chaudhuri, T., & Paul, K. (2014). Prediction and analysis of surface hydrophobic residues in tertiary structure of proteins. *The Scientific World Journal,* Article ID 971258, 1–7. Disponible en: https://doi.org/10.1155/2014/971258

Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R., & Warwicker, J. (2017). Protein-Sol: A web tool for predicting protein solubility from sequence. *Bioinformatics*, 33(19): 3098–3100. Disponible en: https://doi.org/10.1093/bioinformatics/btx345

Huang, N., Kalyanaraman, C., Bernacki, K., & Jacobson, M. P. (2006). Molecular mechanics methods for predicting protein-ligand binding. *Physical Chemistry Chemical Physics*, 8(44): 5166–5177. Disponible en: https://doi.org/10.1039/b608269f

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, 33(7): 1870–1874. Disponible en: https://doi.org/10.1093/molbev/msw054

Kuriata, A., Gierut, A. M., Oleniecki, T., Ciemny, M. P., Kolinski, A., Kurcinski, M., & Kmiecik, S. (2018). CABS-flex 2.0: A web server for fast simulations of flexibility of protein structures. *Nucleic Acids Research*, 46(W1): W338–W343. Disponible en: https://doi.org/10.1093/nar/gky356

Kyte, J., Doolittle, R. F., Diego, S., & Jolla, L. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology*, 57(1): 105–132. Disponible en: https://doi.org/10.1016/0022-2836(82)90515-0

Madeira F., Park Y.M., Lee J., Buso N., Gur T., Madhusoodanan N., Basutkar P., Tivey A.R.N.,

Potter S.C., Finn R.D. & Lopez R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, 2;47(W1): W636–W641. Disponible en: https://doi.org/10.1093/nar/gkz268

Maiorov, V. N., & Crippen, G. M. (1994). Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *Journal of Molecular Biology* 235: 625–634. Disponible en: https://doi.org/10.1006/jmbi.1994.1017

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cowley A.P., Lopez, R. (2013). Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Research*, 41: 597–600. Disponible en: https://doi.org/10.1093/nar/gkt376

Munson, M., Balasubramanian, S., Fleming, K. G., Nagi, A. D., O'Brien, R., Sturtevant, J. M., & Regan, L. (1996). What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties. *Protein Science*, 5(8): 1584–1593. Disponible en: https://doi.org/10.1002/pro.5560050813

Myers, J. K., & Pace, C. N. (1996). Hydrogen bonding stabilizes globular proteins. *Biophysical Journal*, 71(4): 2033–2039. Disponible en: https://doi.org/10.1016/S0006-3495(96)79401-8

Nadzirin, N., & Firdaus-Raih, M. (2012). Proteins of unknown function in the protein data bank (PDB): An inventory of true uncharacterized proteins and computational tools for their analysis. *International Journal of Molecular Sciences*, 13(10): 12761–12772. Disponible en: https://doi.org/10.3390/ijms131012761

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1): 205–217. Disponible en: https://doi.org/10.1006/jmbi.2000.4042

Pace, C. N., Scholtz, J. M., & Grimsley, G. R. (2015). Forces Stabilizing Proteins. *FEBS Letters*,

588(14): 2177–2184. Disponible en: https://doi.org/10.1016/j.febslet.2014.05.006.Forces

Pearson, W. R., & Lipmant, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85: 2444–2448. Disponible en: https://doi.org/10.1073/pnas.85.8.2444.

Raval, A., Piana, S., Eastwood, M. P., Dror, R. O., & Shaw, D. E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Structure, Function and Bioinformatics*, 80(8): 2071–2079. Disponible en: https://doi.org/10.1002/prot.24098

Russell, R. B., Saqi, M. A. S., Sayle, R. A., Bates, P. A., & Sternberg, M. J. E. (1997). Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation. *Journal of Molecular Biology*, 269(3): 423–439. Disponible en: https://doi.org/10.1006/jmbi.1997.1019

Sievers, F., & Higgins, D. G. (2014). Numbers of Sequences. In: *Multiple Sequence Alignment Methods,* 1079: 105–116. Disponible en: https://doi.org/10.1007/978-1-62703-646-7

Sievers, F. et al., (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(539). Disponible en: https://doi.org/10.1038/msb.2011.75

Simm, S., Einloft, J., Mirus, O., & Schleiff, E. (2016). 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification. *Biological Research*, 49(1): 31. Disponible en: https://doi.org/10.1186/s40659-016-0092-5

Sweet, R. M., & Eisenberg, D. (1983). Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *Journal of Molecular Biology*, 171(4): 479–488. Disponible en: https://doi.org/10.1016/0022-2836(83)90041-4

Tokunaga, H., Ishibashi, M., Arakawa, T., & Tokunaga, M. (2004). Highly efficient renaturation of β-lactamase isolated from moderately halophilic bacteria. *FEBS Letters*, *558*(1–3): 7–12. Disponible en: https://doi.org/10.1016/S0014-5793(03)01508-4

Waterhouse, A. et al., (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1): W296–W303. Disponible en: https://doi.org/10.1093/nar/gky427

Wei, L., Huang, E. S., & Altman, R. B. (1999). Are predicted structures good enough to preserve functional sites? *Structure*, *7*(6): 643–650. Disponible en: https://doi.org/10.1016/S0969-2126(99)80085-9

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2014). The I-TASSER suite: Protein structure and function prediction. *Nature Methods*, *12*(1): 7–8. Disponible en: https://doi.org/10.1038/nmeth.3213