SISTEMATIZACIÓN DEL ANÁLISIS ESTRUCTURAL DE PROTEÍNAS MEDIANTE EL DISEÑO E IMPLEMENTACIÓN DE UNA APLICACIÓN BASADA EN HCA (*HYDROPHOBIC CLUSTER ANALYSIS*)

FAVIO ALBERTO HERNÁNDEZ AMAYA

JUAN CAMILO MOYA MUÑOZ

UNIVERSIDAD INDUSTRIAL DE SANTANDER

FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS

ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

BUCARAMANGA

2010

SISTEMATIZACIÓN DEL ANÁLISIS ESTRUCTURAL DE PROTEÍNAS MEDIANTE EL DISEÑO E IMPLEMENTACIÓN DE UNA APLICACIÓN BASADA EN HCA (*HYDROPHOBIC CLUSTER ANALYSIS*)

FAVIO ALBERTO HERNÁNDEZ AMAYA

JUAN CAMILO MOYA MUÑOZ

Trabajo de Grado para optar al Título de
Ingeniero de Sistemas

Director
Alfonso Mendoza Castellanos
Bachelor of Science DEA

Codirector
Jorge Hernández Torres
Ph.D. Biología Molecular

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2010

AGRADECIMIENTOS

A los profesores Alfonso Mendoza y Jorge Hernández por el tiempo dedicado y facilitarnos los medios para completar el proyecto.

A nuestros padres, hermanos y familiares por la paciencia y comprensión.

Al profesor Jaques Chomilier por su tiempo, orientación y herramientas proporcionadas.

CONTENIDO

INDICE DE FIGURAS

GLOSARIO

BIOINFORMÁTICA: aplicación de las técnicas informáticas al estudio de la información genética.

AMINOÁCIDO: sustancia química orgánica en cuya composición molecular entran un grupo amino y otro carboxilo. 20 de tales sustancias son los componentes fundamentales de las proteínas.

PROTEÍNA: sustancia constitutiva de las células y de las materias vegetales y animales. Es un biopolímero formado por una o varias cadenas de aminoácidos, fundamental en la constitución y funcionamiento de la materia viva, como las enzimas, las hormonas, los anticuerpos, etc.

PÉPTIDO: molécula formada por la unión covalente de dos o más aminoácidos.

POLIPÉPTIDO: nombre utilizado para designar un péptido de tamaño suficientemente grande, se puede hablar de más de 10 aminoácidos. Cuando el polipéptido es suficientemente grande y, en particular, cuando tiene una estructura tridimensional única y estable, se habla de una proteína.

RESUMEN

TÍTULO: SISTEMATIZACIÓN DEL ANÁLISIS ESTRUCTURAL DE PROTEÍNAS MEDIANTE EL DISEÑO E IMPLEMENTACIÓN DE UNA APLICACIÓN BASADA EN HCA (*HYDROPHOBIC CLUSTER ANALYSIS*)[*]

Autores:
Favio Alberto Hernández Amaya
Juan Camilo Moya Muñoz[**]

Palabras claves: HCA, Windows Presentation Foundation, 1D

Descripción: con el fin de facilitar el análisis de secuencias y ante la ausencia de herramientas gráficas que permitan optimizar el proceso investigativo, se ha creado un software de soporte gráfico y analítico al proceso de secuenciación múltiple utilizando el método HCA.

Dada la importancia de HCA como un método que permite el análisis de secuencias con rangos de identidad de entre 25-30%, intervalo en que los métodos tradicionales no muestran gran nivel de efectividad y teniendo en cuenta la dificultad gráfica del uso del método, debido a la ausencia de herramientas a la medida para desarrollar el trabajo, la creación del Editor HCA representa un avance para el desarrollo del método y permite a los investigadores corroborar el trabajo realizado frente a los métodos comunes 1D.

La herramienta HCA EDITOR fue desarrollada mediante un proceso iterativo, empleando WPF (Windows Presentation Foundation) y utilizando las pautas de desarrollo grafico aportadas por miembros del grupo creador de HCA para la creación de clústeres y herramientas de análisis.

Las pruebas realizadas sobre HCA EDITOR han mostrado su eficacia y confiabilidad en el trabajo realizado por los investigadores, obteniendo la identidad de las secuencias analizadas instantáneamente a partir del trabajo realizado.

[*] Proyecto de Grado

[**] Facultad de Ingeniería Físico Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director Alfonso Mendoza Castellanos Codirector Jorge Hernández Torres

ABSTRACT

TITLE: SYSTEMATIZATION STRUCTURAL ANALISIS OF PROTEINS BY THE DESIGN AND IMPLEMENTATION OF AN APPLICATION BASED ON HCA (Hydrophobic Cluster Analysis). *

Authors:
Favio Alberto Hernández Amaya
Juan Camilo Moya Muñóz**

Key words: HCA, Windows Presentation Foundation, 1D.

In order to facilitate sequence analysis and in the absence of graphical tools to optimize the research process has created a graphic and analytical software to support the process of secuence's alignment using the HCA method.

Given the importance of HCA as a method that allows analysis of sequences with identity range of 25-30%, range in which traditional methods do not show high level of effectiveness and taking into account the difficulty of using graphical method and the lack of tools tailored to develop the work, the creation of HCA Editor represents a breakthrough for the development of the method and allows researchers to verify the work done against the common methods 1D.

EDITOR HCA tool was developed through an iterative process, using WPF (Windows Presentation Foundation) and using graphic development patterns made by members of the creator of HCA.

Tests on HCA have shown EDITOR efficiency and reliability in the work of researchers.

* Project of Grade

** Faculty Physic Mechanical Engineering. School Systems and Computer Engineering. Director Alfonso Mendoza Castellanos. Codirector Jorge Hernandez Torres

INTRODUCCIÓN

La última década ha sido testigo del surgimiento de una nueva disciplina de la biología, *in silico*. La Bioinformática nació con los primeros computadores y se constituyó en una herramienta indispensable para los investigadores. El inmenso caudal de programas ha permitido realizar análisis genómicos y proteómicos imposibles de ejecutar manualmente y en tiempos tan sorprendentemente cortos.

El propósito de la genómica es la caracterización de la estructura, organización y expresión de los genes. Con el advenimiento de la secuenciación de genomas completos, las bases de datos de dominio público han multiplicado exponencialmente la cantidad de secuencias nucleotídicas disponibles a través de la red. Por ejemplo, actualmente Genbank aloja un número aproximado de 108.431.692 millones de secuencias y aumenta diariamente (Benson *et al.*, 2008;Pressman, 2005).

Por otra parte, la proteómica (el análisis de todas las proteínas que expresa un organismo) busca elucidar la composición, estructura, función e interrelación entre las proteínas codificadas por los genes de un organismo.

Uno de los mayores retos actuales de la Bioinformática es la determinación de la estructura y función del inmenso caudal de secuencias polipeptídicas producto de la secuenciación automatizada. Esto no es tarea fácil, pues las bases de datos están abarrotadas de secuencias nuevas y desconocidas y la caracterización individual de cada proteína consume tiempo y esfuerzo. A veces la atribución de la función que podría tener una secuencia no es posible por los métodos actualmente disponibles.

Por ejemplo, la secuenciación del genoma completo y dos plásmidos de *Pseudomonas syringae* Pv. phaseolicola 1448A (cromosoma circular: 5,928,785 bp; p1448A-A: 131,950 bp y p1448A-B: 51,711 bp, respectivamente) (Joardar et al., 2005) arrojó un total de 5,353 secuencias de proteínas potenciales (*open reading frames*, ORFs). En este estudio se pudo asignar una función a 3,626 (68%) ORFs; de las 1,727 proteínas restantes se pudieron anotar 224 (4%) como proteínas hipotéticas y 822 (15%) como proteínas hipotéticas conservadas. Finalmente, 681 proteínas (13%) fueron registradas como de función desconocida. Es decir, no se les pudo reconocer una función con las herramientas bioinformáticas actuales. Es aquí donde pueden intervenir métodos de análisis alternativos como el de HCA.

En efecto, la asignación de funciones biológicas potenciales a secuencias polipeptídicas nuevas se hace con base en alineamientos con proteínas ya caracterizadas. A mayor grado de identidad (aminoácidos iguales) entre dos secuencias proteicas, mayor probabilidad de que las dos proteínas cumplan con la misma función biológica, aunque no siempre ocurre así (Gerlt and Babbitt, 2000).

De manera estimativa, se considera que dos proteínas comparten un ancestro común, es decir provienen evolutivamente del mismo antecesor, cuando comparten el 25% de aminoácidos idénticos sobre más de 100 aminoácidos. Sobre este criterio, se han asignado funciones de manera automatizada a secuencias nuevas (Henikoff and Henikoff, 1994;Pearson, 2001).

Uno de los mejores algoritmos diseñados hasta la fecha para la búsqueda de homólogos a una secuencia específica es Blast (Altschul *et al.*, 1997). Blast, (*Basic Local Alignment Search Tool*), permite comparar una secuencia desconocida con todas las alojadas en una base de datos como Genbank y proponer una lista de las más parecidas con base en la identidad

(aminoácidos iguales/aminoácidos totales). Posteriormente, con base en los resultados, el investigador o un algoritmo computacional decide a cuál proteína es la que más se asemeja y esta información se registra en la base de datos como función potencial (Pearson, 2001).

El problema surge cuando los niveles de identidad están por debajo de 30%, aproximadamente A partir de ese umbral y hasta 15% o menos, las herramientas bioinformáticas actuales pierden su sensibilidad y no producen resultados confiables (Rost, 1999).

Como una alternativa para superar esa dificultad, se creó un método de alineamiento manual de secuencias que se denominó HCA (*Hydrophobic Cluster Analysis*) ó Análisis de los Agregados Hidrofóbicos (Callebaut *et al.*, 1997a). Se trata de un método no convencional que ha mostrado ser particularmente eficiente y sensible para secuencias con bajos niveles de identidad. Esta alta sensibilidad ha hecho posible predecir la función de proteínas con niveles de identidad muy bajos y ha abierto la posibilidad de clasificar una inmensa cantidad de secuencias que esperan su análisis.

El secreto de HCA se basa en alineamientos de representaciones bidimensionales (2D) en lugar de unidimensionales (1D), de las secuencias polipeptídicas. Todas las herramientas bioinformáticas actuales, incluyendo BLAST, acuden a comparaciones lineales de las cadenas de aminoácidos. El resultado puede ser pobre cuando la identidad es inferior al 30% y conducir a errores. HCA en cambio, representa las secuencias aminoacídicas en 2D y es el investigador quien efectúa el alineamiento manualmente de manera que, con base en su experiencia y razonamiento, toma decisiones difícilmente automatizables por un programador. El resultado es un alineamiento con mayor valor de identidad que en 1D y con más información biológica relevante. Con este método se puede llegar a determinar la proteína precursora ancestral y, desde luego, la función bioquímica actual.

El presente trabajo de grado se llevó a cabo con el fin de proveer a los investigadores de una herramienta bioinformática que facilite el análisis gráfico de alineamientos de secuencias de proteínas, utilizando el método HCA. Vale la pena recordar que los alineamientos mediante HCA se realizan manualmente. El programa HCA EDITOR evitará el trabajo manual, el cual se realizará *in silico*. Como resultado, el usuario puede realizar sus alineamientos mucho más rápidamente y directamente en el computador, lo que le permitirá editarlos previamente a su publicación, imprimir copias adicionales e intercambiarlos con otros investigadores. Además, cuenta con un algoritmo que convierte el alineamiento 2D en 1D, lo que hasta la fecha no efectúa ningún otro programa en el mundo bioinformático.

La metodología de desarrollo utilizada para la construcción de la herramienta se basó en el proceso de desarrollo por incrementos, aproximación que permitió la depuración de factores funcionales importantes para facilitar el trabajo y el análisis realizado por el investigador.

1. PRESENTACIÓN DEL PROYECTO

1.1 OBJETIVOS

1.1.1 Objetivo General: Facilitar el análisis estructural de proteínas mediante la construcción de un software que otorgue al investigador herramientas de carácter gráfico y analítico, que le permita realizar dicho análisis en un menor tiempo y con una mayor confiabilidad.

1.1.2 Objetivos Específicos:

- La herramienta debe identificar como parámetros de entrada dos *secuencias de proteínas*[1] que deben tener como formato de entrada el formato FASTA[2] o el formato GC\MSF[3] utilizados ampliamente en aplicaciones de carácter bioinformático.

- Representación gráfica bidimensional de las proteínas de entrada, asignando un conjunto de símbolos y colores determinados con una disposición espacial de los mismos basada en el método HCA.

- Desarrollo de un algoritmo que agrupe gráficamente los aminoácidos hidrofóbicos presentes en una vecindad dentro de la gráfica en un bloque llamado cluster.

---

[1] Secuencia lineal de aminoácidos.

[2] Formato fasta (>nombre de la secuencia y luego la secuencia). Ej.: >rbp ACTAGGACAGCCACTAGACC….ETC.

[3] Formato para dos o más secuencias alineadas, donde elcaracter "." representa un *gap(brecha)*, los nucleótido o aminoácidos son representados en su código de una letra, y la secuencia es escrita en columnas de diez (10) letras cada una. El comienzo de la secuencia es marcado por dos *backslashes*: //.

- Implementación de un formato de impresión que se adecue a las gráficas, ya que por su extensión deben ser fraccionadas.

- Permitir que el usuario guarde de forma digital el trabajo realizado para su posterior utilización.

- Un menú de opciones que permita al usuario manejar de forma simple la inclusión de gaps (brecha espacial en una secuencia de proteína, que se utiliza en el método HCA).

- Desarrollo de un algoritmo de conversión HCA 2D-1D, como medio de comparación con los métodos de alineamiento tradicionales.

1.2 DEFINICIÓN DEL PROBLEMA

La ciencia de la secuenciación comenzó lentamente. Antes de 1945 no existía ningún análisis cuantitativo disponible para ninguna proteína. Sin embargo, posteriores avances hicieron posible que hacia 1960 se hubieran secuenciado unas 20 proteínas; para 1980 el orden se estimaba en 1500. Hoy se encuentran millones de proteínas disponibles en las bases de datos mundiales y su cantidad que aumenta constantemente.

HCA es un método de comparación de proteínas particularmente eficiente y sensible en familias de secuencias con bajos niveles de identidad. Esta extrema sensibilidad ha hecho posible predecir las funciones de genes con secuencias difíciles de alinear por métodos unidimensionales y ofrece una nueva vía para explorar la enorme cantidad de datos generados por el secuenciamiento de genomas completos. HCA provee de originales herramientas para entender aspectos fundamentales de la estructura y

función de las proteínas. Desde la última publicación sobre HCA en 1990, han surgido nuevos e importantes desarrollos computacionales.

La dificultad existente en el uso HCA se encuentra en el extensivo trabajo gráfico necesario para el desarrollo del método y en la falta de herramientas bioinformáticas a la medida, disponibles en el mercado. La importancia de este Trabajo de Grado radica en proporcionar al investigador un entorno gráfico de análisis con opciones ajustadas al método, que reduzcan el tiempo empleado actualmente en la representación y trabajo sobre las secuencias y que el usuario pueda concentrarse en la interpretación de los resultados.

1.3 IMPACTO ESPERADO

Se espera que los estudiantes e investigadores tanto del país como de otras latitudes que trabajan actualmente con el método HCA, puedan utilizar la herramienta en sus investigaciones e incluir el trabajo gráfico realizado directamente en sus publicaciones.

A nivel investigativo la creación de la herramienta constituye un avance hacia posteriores desarrollos que automaticen en mayor grado el trabajo realizado por el investigador.

1.4 VIABILIDAD

Existe la necesidad del apoyo de los investigadores del área de Biología Molecular de la Escuela de Biología, porque son quienes utilizan el método HCA en sus proyectos de investigación. Por lo demás, la Escuela de Ingeniería de Sistemas nos ha aportado los conocimientos necesarios para abordar el problema y contamos con el apoyo permanente de nuestro director y codirector del Trabajo de Grado para resolver problemas específicos.

2. MARCO TEÓRICO

2.1 Proteínas: de la estructura primaria (1D) a la estructura terciaria (3D)

Las técnicas más tempranas de determinación de la secuencia de una proteína se basaban en métodos para la separación de proteínas y péptidos, asociados con métodos para la identificación y cuantificación de aminoácidos (Steen and Mann, 2004). Antes de 1945 no existía ningún análisis cuantitativo disponible para ninguna proteína. Pero los avances en las técnicas de cromatografía y etiquetado a lo largo de la siguiente década desembocaron finalmente en la elucidación de la primera secuencia completa, la de la hormona peptídica insulina (Sanger, 1959). Con todo, aún transcurrieron años hasta que se completó la secuencia de la primera enzima, que fue una ribonucleasa (Kresge *et al.*, 2005). Hacia 1965 se habían secuenciado unas 20 proteínas y para 1980 el número se estimaba en 1500. Hoy existen cientos de miles de secuencias disponibles (Benson *et al.*, 2008).

Es importante tener presente la diferencia de escala en el manejo de información de secuencias y estructuras. A principios del año 2000, en las bases de datos no redundantes disponibles públicamente, se habían depositado más de 300.000 secuencias de proteínas y el número de secuencias parciales en bases de datos se estima en el orden de millones. Por el contrario el número de estructuras 3D únicas en el Protein Data Bank es todavía inferior a 1500 (Berman *et al.*, 2008). Estas cifras subrayan un enorme déficit de información, ya que la adquisición de datos estructurales llega a producir unas 2000 estructuras al año, que es una producción pequeña en comparación con las bases de datos de secuencias, que duplican su tamaño cada año, con la adición de una nueva secuencia cada medio minuto.

La extracción de sentido biológico en la información de secuencias es una ciencia en vías de ser exacta. En esencia, los investigadores se enfrentan al problema de decodificar un lenguaje desconocido y el desafío central de las aplicaciones bioinformáticas es la racionalización de la masa de información de secuencias, con vistas a desarrollar medios más eficaces de análisis. El propósito de esas nuevas herramientas es el de descifrar las pistas estructurales, funcionales y evolutivas codificadas en el lenguaje de las secuencias biológicas.

La pregunta central alrededor de la cual orbitan los métodos de análisis de secuencias es cómo determinar la forma y función a partir de un ordenamiento lineal de aminoácidos. Teóricamente, es posible derivar reglas que determinan el modo en que una proteína se pliega, a partir del análisis de proteínas con estructura conocida. Luego, aplicar tales reglas a una predicción, con base exclusivamente en una secuencia lineal de aminoácidos. A primera vista, considerando las bases de datos disponibles en constante crecimiento, esta no podría parecernos una esperanza irreal. Sin embargo, a pesar de más de tres décadas de investigación las reglas de plegamiento de proteínas no se comprenden en su totalidad y todavía no es posible la predicción de estructuras *ab initio* (Chen and Johnson, 2009).

2.2 Alineamiento de secuencias aminoacídicas

El alineamiento de estructuras primarias (secuencias en 1D) es una de las formas más utilizadas para representar y comparar dos o más proteínas. Los aminoácidos se simbolizan con letras ordenadas paralelamente una a una, con el fin de encontrar regiones de composición idéntica o similar.

Si dos secuencias están relacionadas ancestralmente, esto se reflejará visiblemente en el alineamiento. Una de las aplicaciones de mayor uso por su alta eficiencia es ClustalW (Larkin *et al.*, 2007). En la siguiente figura se

muestra un ejemplo de alineamiento de dos secuencias utilizando esta aplicación disponible en la Web para uso en línea o descargable:

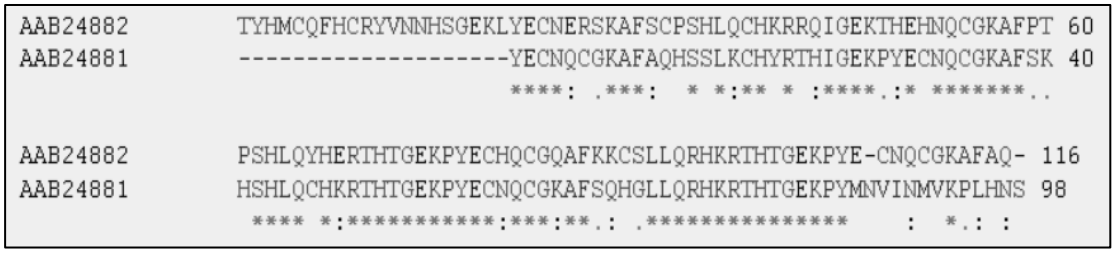


Figura 1. Alineamiento ClustalW.

Fuente: http://www.ebi.ac.uk/Tools/clustalw2/index.html. Agosto 2010.

Como puede observarse, los aminoácidos idénticos se marcan con el símbolo "⁎". Los pares marcados con ":" y "." corresponden a aminoácidos que no son idénticos, pero que conservan propiedades bioquímicas similares. Finalmente, el guión “-” representa ajustes que hace el programa para desfasar una secuencia respecto de la otra y así incrementar el número de aminoácidos idénticos.

Como hemos visto hasta ahora, un alineamiento arroja resultados cualitativos; es decir, el investigador interpreta visualmente el resultado, con especial atención en los bloques de "⁎ ⁎ ⁎ ⁎ ⁎ ⁎ ⁎ ⁎ " que equivalen a un alto parecido entre las dos secuencias.

No obstante, el algoritmo de ClustalW también está diseñado para expresar una valoración cuantitativa del alineamiento. A esto se le llama "identidad de secuencia" o "simplemente identidad" y se refiere al número de residuos idénticos, dividido por el número de aminoácidos totales de la secuencia más larga. Para el ejemplo anterior, la identidad sería de 61/116 = 52%. En el mundo de la Bioinformática, 52% de identidad es un valor elevado y conduce a concluir que esas dos secuencias provienen de una única proteína ancestral y durante la evolución derivó en dos secuencias parcialmente emparentadas en su composición 1D.

Un ejemplo en el cual dos secuencias tienen un bajo nivel de identidad sería:

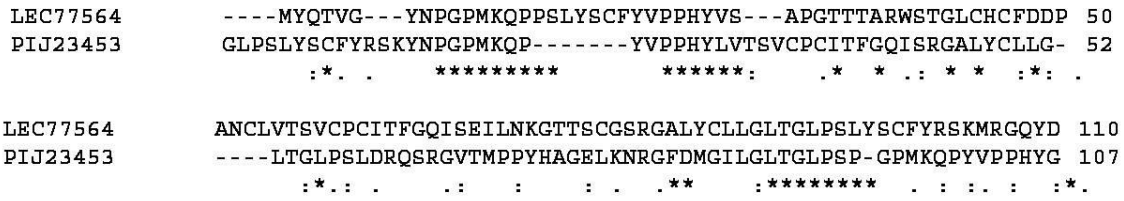


Figura 2.Alineamiento ClustalW con bajo nivel de identidad.

Fuente: Archivo Cinbin.

La poca cantidad de "**x**" indica un bajo nivel de identidad (30%) y si recordamos lo dicho en la introducción, por consenso en la comunidad científica, se considera que dos proteínas provienen evolutivamente del mismo antecesor, cuando comparten alrededor del 25% de aminoácidos idénticos sobre más de 100 aminoácidos (Henikoff and Henikoff, 1994;Pearson, 2001). En este umbral de identidad, llamado *the twilight zone* (Rost, 1999), es muy difícil concluir si existe algún nivel de parentesco entre las dos secuencias. Es aquí donde métodos como el HCA intervienen y revelan lo que aparentemente podría ser o no ser.

2.3 Representación bidimensional (2D) de las secuencias para HCA

El método HCA (*Hydrophobic Cluster Analysis*) o análisis de los agregados hidrofóbicos es una técnica no convencional de análisis de proteínas, en la cual se realizan alineamientos manualmente. Es un método particularmente eficiente para alinear secuencias que comparten bajos niveles de identidad (<30%) (Callebaut *et al.*, 1997a). Esta particularidad hace de HCA una técnica mucho más sensible que las herramienta de análisis 1D como ClustalW, en el rango de la *twilight zone*.

El método se basa en la representación gráfica de las secuencias en 2D. Básicamente, se trata de asumir que cada secuencia conforma una única estructura secundaria en forma de cilindro, o más precisamente, α-hélice (**a**). Al abrir el cilindro, los aminoácidos quedan proyectados en un espacio bidimensional (**b**). Esta imagen es duplicada con el fin de revelar las vecindades existentes entre los aminoácidos (**c**) (Callebaut *et al.*, 1997a):

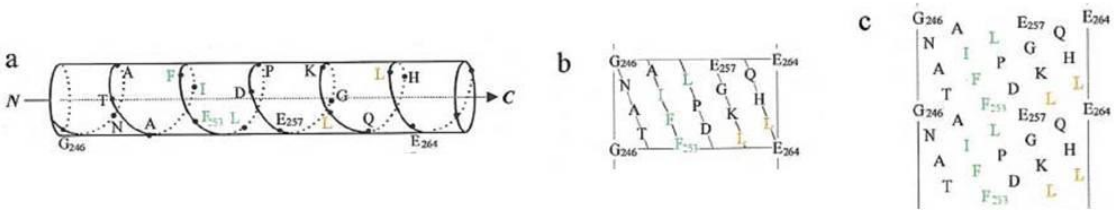


Figura 3. Conformación gráfica HCA

Fuente: (Callebaut *et al.*, 1997a).

Posteriormente, los aminoácidos hidrofóbicos colindantes en el espacio se agrupan en un racimo o *cluster*, teniendo en cuenta que cada residuo (0) queda espacialmente rodeado de sus vecinos que le preceden (-1, -4, -3) y le suceden (+1, +3, +4):

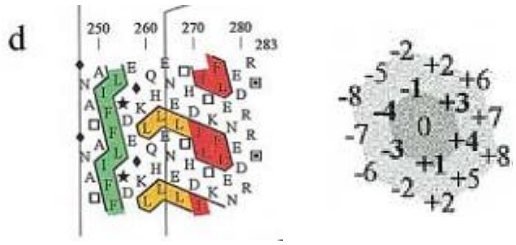


Figura 4. Conformación de clusters HCA

Fuente: (Callebaut *et al.*, 1997a).

El hecho de asumir que toda la proteína es una sola α-hélice es válido, porque se ha demostrado que el centro de los *clusters* coincide con el centro de las estructuras secundarias (Woodcock *et al.*, 1992).

Por otra parte, La forma de los *clusters* provee una indicación del tipo de estructura secundaria adoptada en un segmento de la secuencia. En efecto,

los clusters horizontales se corresponden estadísticamente con hojas β y los verticales con α-hélices (Woodcock *et al.*, 1992). Ejemplo:

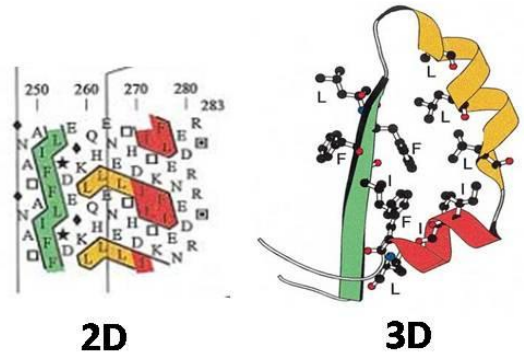


Figura 5. Gráfica HCA – estructura secundaria de proteína

Fuente: (Callebaut *et al.*, 1997a).

2.4 Alineamiento de secuencias mediante el método HCA

Como dicho anteriormente, el alineamiento mediante HCA se hace de manera manual. Es decir, una vez representada cada secuencia en 2D, se procede a descifrar visualmente los aminoácidos comunes, agrupados en clusters:

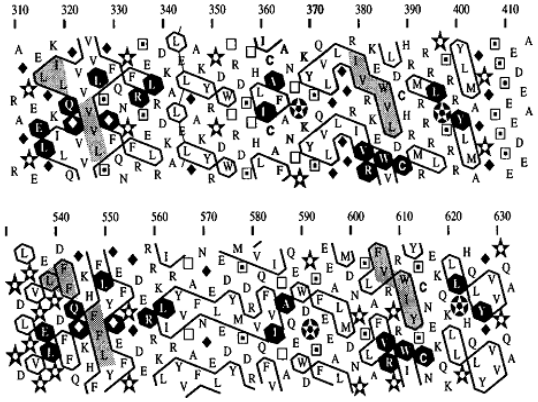


Figura 7. Clusters

Fuente: Archivo Cinbin.

En la figura anterior, se destacan dos tipos de aminoácidos. Los hidrofóbicos, asociados en clusters con la misma forma (áreas sombreadas), indistintamente de su estructura. Basta con que pertenezca a la categoría de hidrofóbicos (VILFMWY). Los no hidrofóbicos comunes (letras blancas dentro de círculos negros), o sea todos los 13 restantes (ARCQTEKHSNDPG). Cuatro residuos aminoacídicos se representan con símbolos, por tener radicales R especiales: Glicina –R = H (♦)– y Prolina –R = ciclo (✯)–. Serina y Treonina son dos residuos hidroxilados (susceptibles de camuflarse con puentes de hidrógeno) representados respectivamente con un cuadrado y un cuadrado con un punto.

Una vez establecidos manualmente los residuos comunes, se procede a relacionar los clusters con líneas verticales para una mejor comprensión del alineamiento:

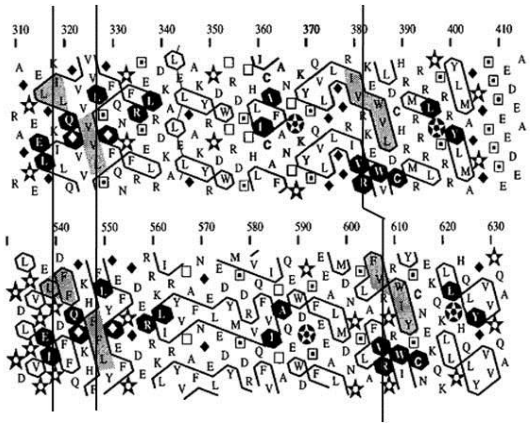


Figura 8. Clusters relacionados

Fuente: archivo Cinbin.

Y finalmente, si se cuenta con estructuras 3D alojadas para una o las dos secuencias en bases de datos como la PDB (Berman *et al.*, 2008), se dibujan debajo de cada secuencia, para revelar el tipo de estructura secundaria que adoptan los clústeres. Las hojas β serán cilindros y las α-hélices espirales o flechas. Además, si es pertinente, algunos residuos importantes se pueden

resaltar en negrilla o en un fondo de otro color o tamaño de fuente y regiones específicas se pueden destacar con cuadros:

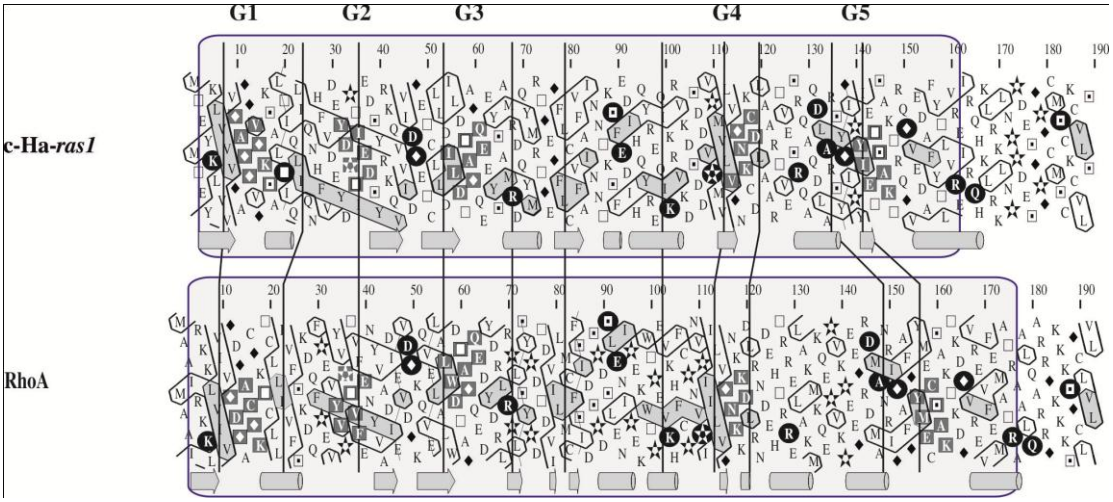


Figura 9. Alineamiento HCA con cuadros destacados.

Fuente: Archivo Cinbin.

Para que el lector pueda interpretar correctamente el alineamiento, es bueno incluir siempre esta figura de referencia:

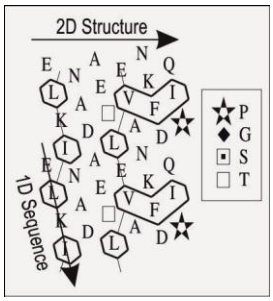


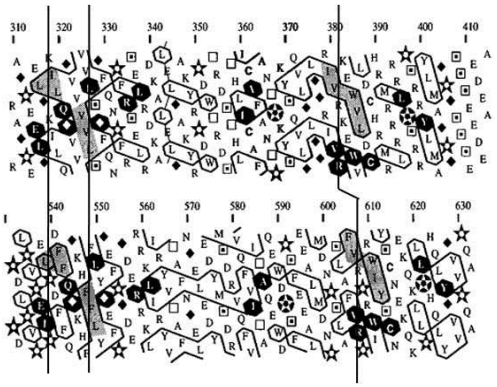
Figura 10. Figura de referencia HCA.

Fuente: Archivo Cinbin.

Como se puede apreciar, la secuencia 1D se lee diagonalmente obviando la duplicación de la secuencia y la estructura 2D se interpretará horizontalmente.

2.5 Cálculo de la identidad de secuencia mediante el método HCA

HCA no es únicamente un método de alineamiento que arroja resultados visuales y cualitativos. También permite calcular la identidad de secuencia (residuos estrictamente idénticos), mediante el conteo de aminoácidos comunes sobre aminoácidos totales. Por ejemplo, el valor de identidad para el siguiente alineamiento se calcularía así:



Hidrofóbicos comunes = 4

No hidrofóbicos comunes = 17

Cálculo: (4 + 17)/ = 114

Total = 15%

Figura 11. Calculo de identidad

Con frecuencia, los valores de identidad calculados mediante HCA son superiores a los obtenidos con programas como ClustalW.

2.6 Trabajando con el método HCA

De acuerdo con los desarrolladores del método (Callebaut *et al.*, 1997b;Woodcock *et al.*, 1992), el desarrollo completo de un alineamiento mediante HCA cumple las siguientes etapas:

- Obtener la secuencia de la proteína de interés, de acuerdo con la pregunta de investigación.
- Analizar individualmente la representación en 2D, teniendo en cuenta la forma de los *clusters* hidrofóbicos (hojas β, α-hélices) y las regiones “coil” (al azar, ni hojas β, ni α-hélices), regiones “hinges” (bisagra), identificar posibles dominios, detectar visualmente eventuales repeticiones internas.

- Buscar proteínas potencialmente homólogas mediante herramientas como BLAST, usando como patrón la secuencia total o parcial de la proteína de interés.
- Representar en 2D la secuencia de las dos proteínas y con la ayuda de marcadores de colores identificar los aminoácidos hidrofóbicos comunes que se agruparán en clusters y los no hidrofóbicos comunes que se marcarán con otro color.
- Sobre la base del análisis de la organización global del alineamiento, de las regiones y residuos conservados y de los niveles de identidad, sacar conclusiones acerca de la relación estructural y funcional entre las dos proteínas.
- Emitir un juicio final sobre la posible relación evolutiva entre las dos proteínas.

2.7 Limitaciones del método HCA

Sin duda la mayor limitación del método HCA es que se realiza manualmente y hasta ahora no ha sido posible automatizar el proceso. Esto sucede porque como hemos visto, son muchas las tareas que se realizan, en las cuales interviene la perspicacia y la experiencia del investigador.

No obstante, con el presente trabajo de grado se pretende reemplazar el trabajo manual de imprimir, recortar, pegar y colorear trazados 2D mediante una herramienta de análisis gráfico, hasta lograr la impresión final que se actualmente hace con programas no muy adecuados como CorelDraw®.

Para los investigadores que trabajan con HCA sería de gran utilidad una aplicación de este tipo y estamos seguros del impacto dentro de la comunidad científica, la publicación de un editor específico para agilizar y mejorar la calidad de los productos derivados del método.

3 Desarrollo del Modelo Computacional

3.1 Metodología

La metodología sobre la cual basamos el desarrollo del proyecto fue iterativa e incremental. Esta metodología de desarrollo se escogió con el fin de realimentar aspectos de utilidad práctica del investigador a la herramienta a medida que se evolucionaba en el desarrollo y se hacían entregas.

El modelo incremental combina elementos del modelo en cascada aplicado en forma iterativa(Pressman, 2005). Cada una de las iteraciones lleva un enfoque secuencial que consta de cinco fases:

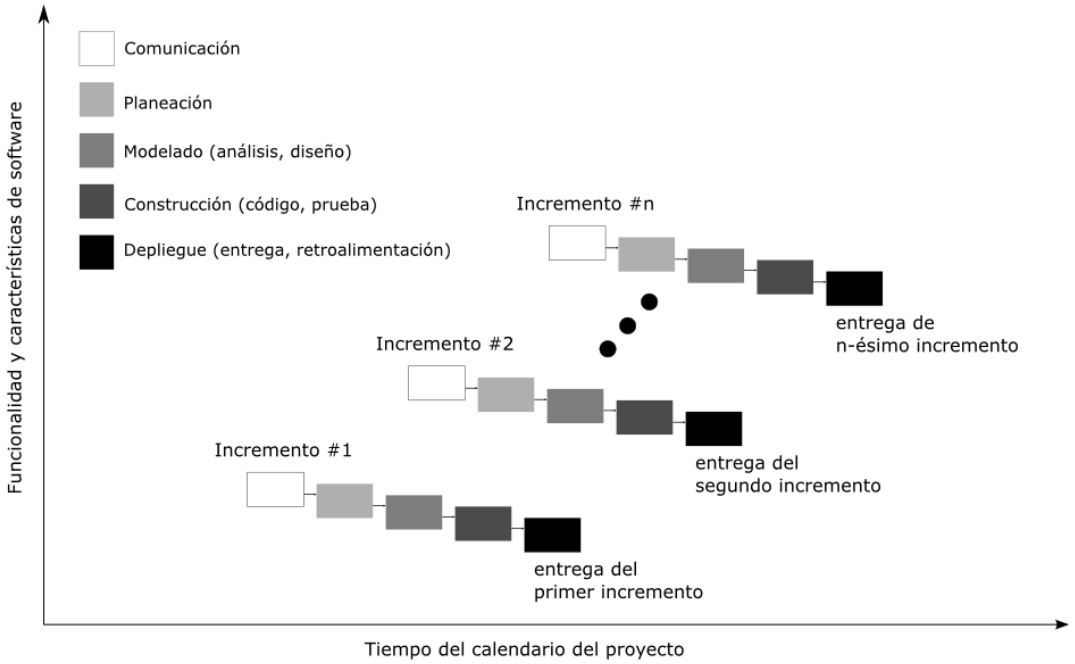


Figura 12. Metodología iterativa e incremental.

Fuente: (Pressman, 2005).

Comunicación: tiene como propósito establecer colaboración y comunicación con el cliente en el análisis y toma de requerimientos.

Planeación: establece un plan de consecución de desarrollo, definiendo las tareas a realizarse, posibles riesgos, recursos requeridos, productos que ha de obtenerse con el trabajo realizado y la elaboración de un plan de trabajo.

Modelado: creación de modelos que permiten al desarrollador y al cliente entender mejor los requisitos del software y el diseño que logra satisfacerlos.

La actividad de elaboración de un modelo implica:

- **Análisis:** conjunto de tareas de trabajo como son investigación, negociación y especificación de tecnologías.

- **Diseño:** abarca tareas de trabajo como definición de interfaz, componentes y arquitectura.

Construcción: generación del código y elaboración de pruebas del desarrollo realizado.

Despliegue: entrega del software parcial o completo, para que el usuario lo evalúe y proporcione información basada en su evaluación.

Cada incremento es una versión incompleta del producto final, que proporciona al usuario alguna funcionalidad y le da la posibilidad de evaluarlo, con el fin de corregir aspectos del desarrollo realizado y replantear nuevos requisitos para la siguiente iteración.

3.2 Diagramas UML

Los diagramas de UML que se desarrollan fueron creados con la intención de encontrar un modelo que describiera el uso y las actividades que se desarrollaría con el programa, a forma muy general se presentan los modelos de caso de uso y de actividades que el usuario (biólogo) desarrolla en el programa.

Digrama Casos de Uso

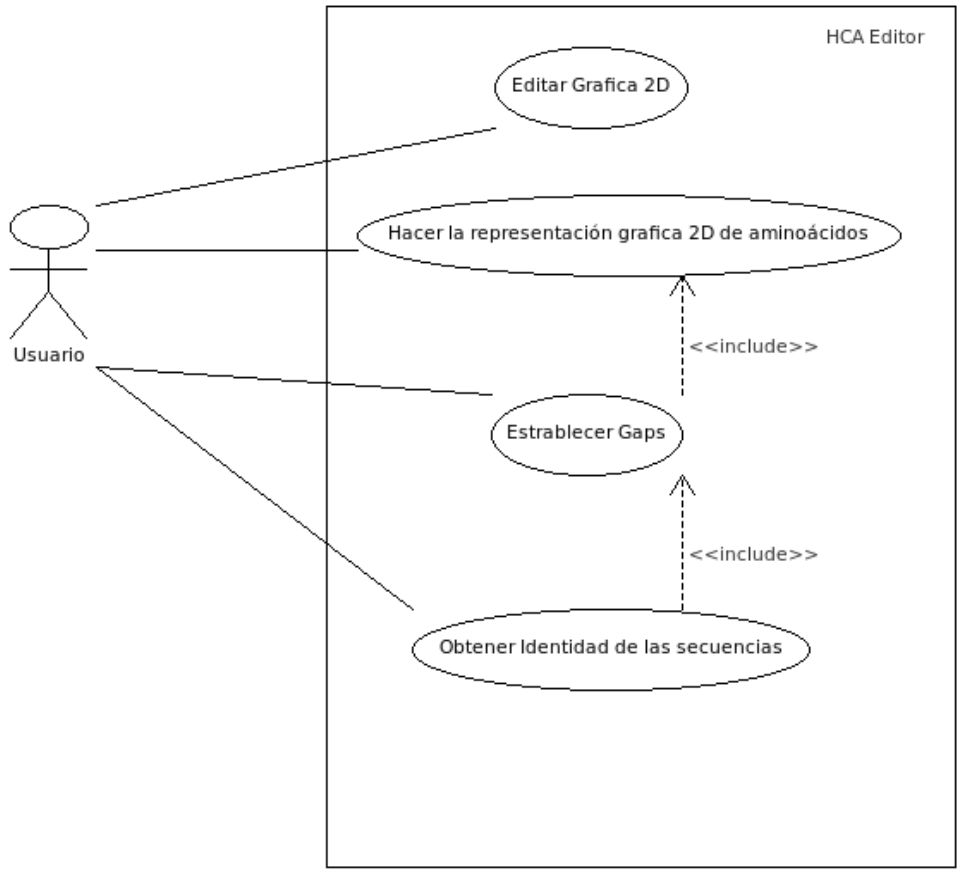


Figura 13. Diagrama de casos de uso.

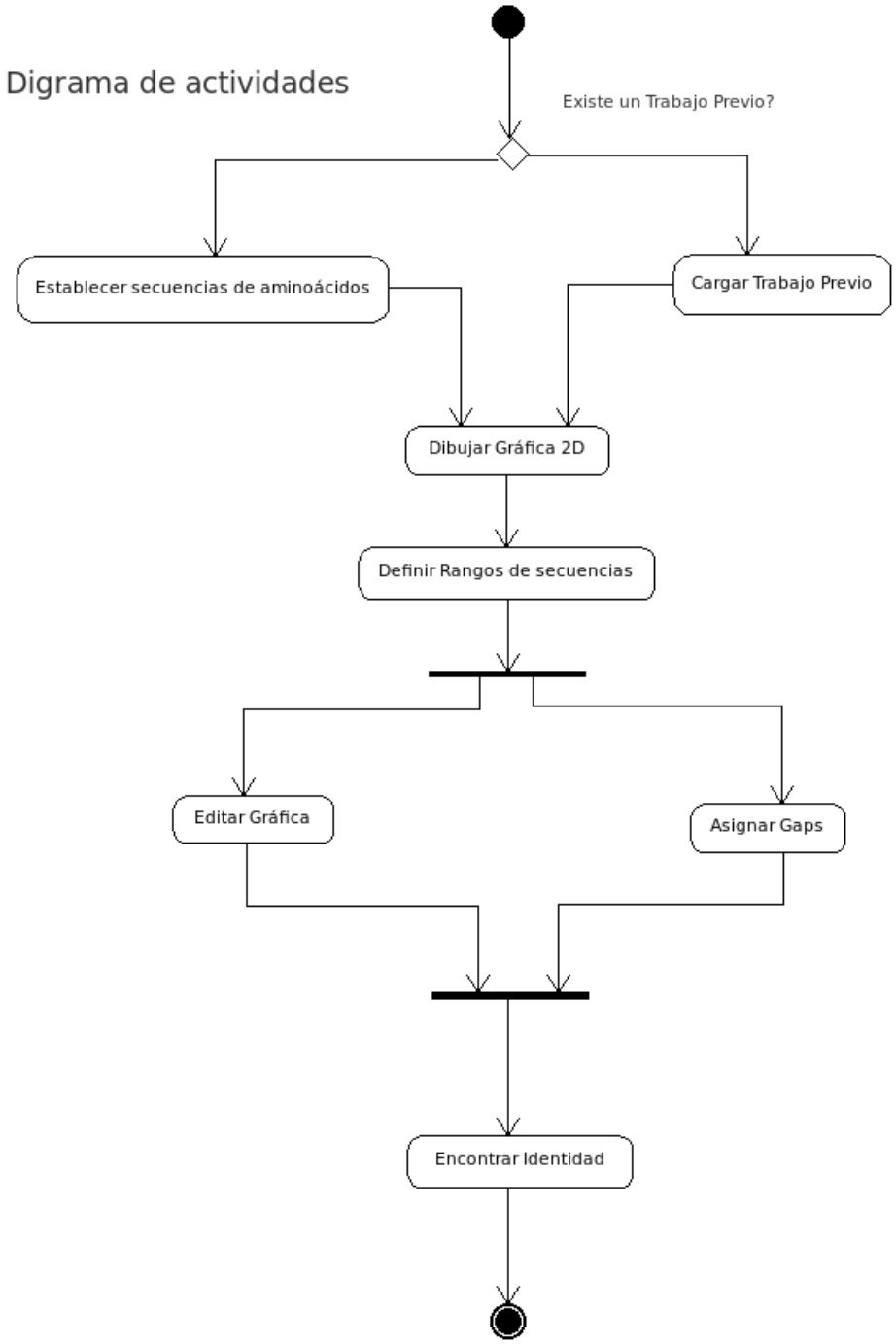


Figura14. Diagrama de actividades.

4. Ingreso de secuencias

El primer paso del proceso de análisis implica el ingreso de las secuencias a trabajar. Los formatos disponibles para trabajar con la herramienta son el formato FASTA (US National Center for Biotechnology Information, 2010) y el formato CG/MSF (US National Center for Biotechnology Information, 2010), las secuencias se ingresan una a la vez para un total de dos secuencias.

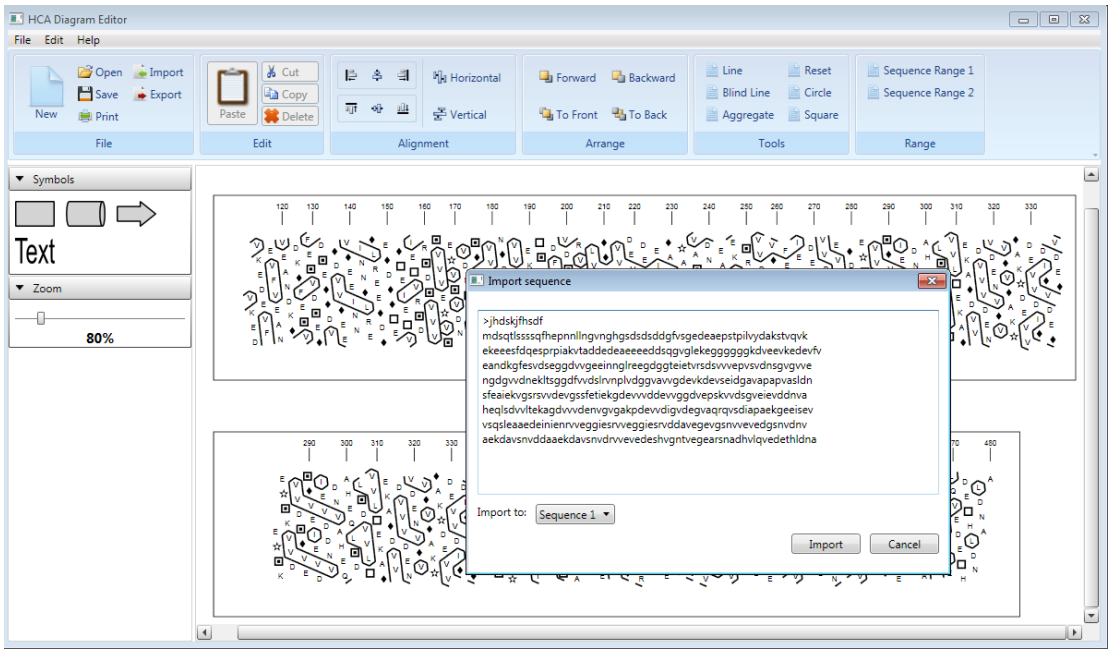


Figura 15. Ingreso de secuencias.

Una vez se han importado las secuencias, se realiza el proceso de graficar en el formato del HCA.

5. Representación gráfica HCA

El proceso de graficar se inicia con la disposición espacial de los aminoácidos presentes en las secuencias según el método HCA. Completado el proceso de graficar se realiza un análisis de los clústeres de aminoácidos hidrofóbicos presentes en cada una de las secuencias de acuerdo al modelo de vecindad HCA. Los clústeres se agrupan gráficamente en zonas que facilitan al investigador la observación de motivos similares dentro del alineamiento. La grafica muestra un alineamiento realizado con el software construido:

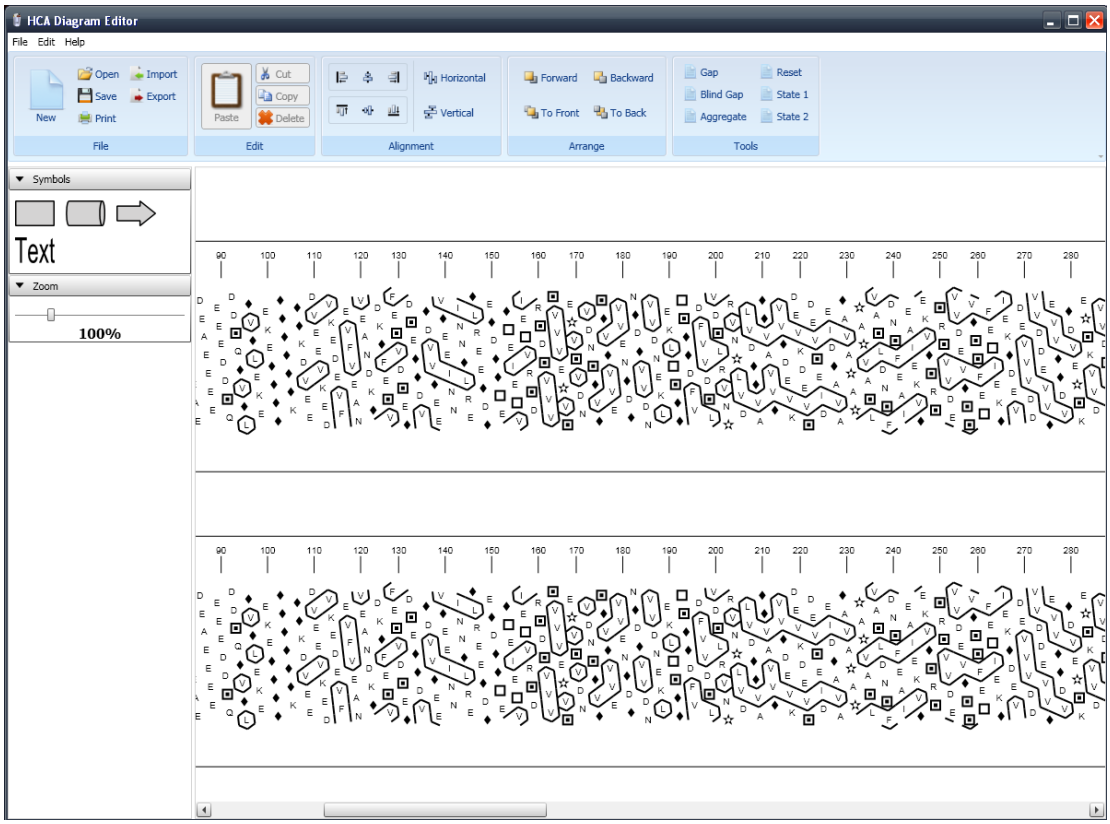


Figura 16. Secuencias graficadas utilizando HCA.

Una vez se ha completado la presentación inicial de las secuencias con sus respectivos clústeres en el formato HCA, el investigador puede iniciar su trabajo sobre las secuencias, con las herramientas proporcionadas por el software.

6. Trabajo gráfico

El objeto final del trabajo investigativo radica en la obtención de la identidad existente entre las secuencias analizadas. El software provee a los investigadores herramientas que le permite la inserción de gaps, creación de agrupamientos, inserción de líneas ciegas e inclusión de texto y símbolos representativos propios del método, útiles para la consecución exitosa del trabajo realizado.

6.1 Inserción de gaps

Un gap constituye un espacio dentro del alineamiento. El método HCA maneja la inserción de gaps a través de líneas trazadas desde una secuencia a otra y entre dos aminoácidos.

La opción gap del software elaborado ejecuta el proceso de trazado de líneas y reordenamiento interno de los aminoácidos dentro del alineamiento de secuencias, esto facilita el análisis grafico observable por el investigador, la obtención del alineamiento 1D final y el cálculo de la identidad de las secuencias analizadas. La grafica muestra creación de un gap con la herramienta desarrollada.

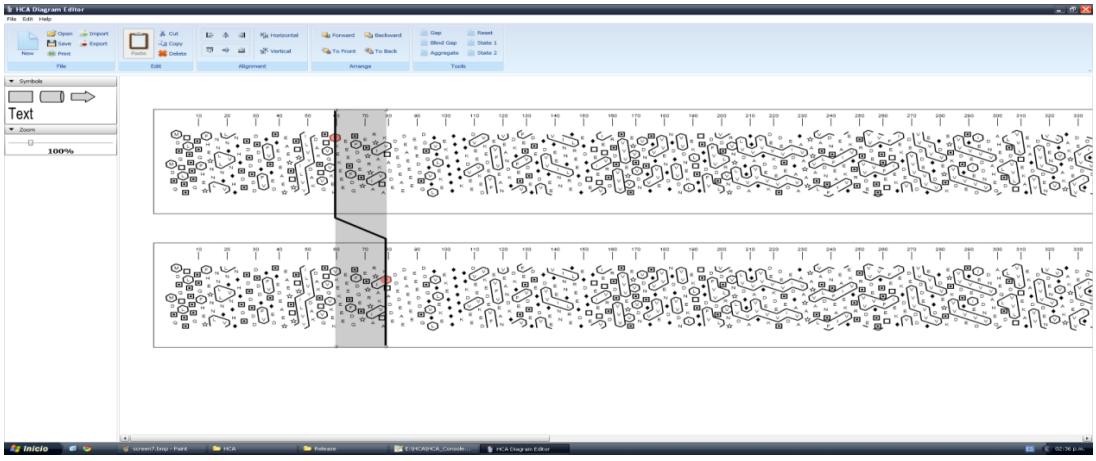


Figura 17. Creación de gap.

6.2 Creación de agrupamientos

Un agrupamiento constituye una guía visual utilizada para agrupar motivos de aminoácidos hidrofóbicos con el fin de presentar a ayuda visual al investigador en el proceso de búsqueda de analogías entre las secuencias analizadas. La herramienta Aggregate que forma parte del software desarrollado permite la representación de agrupamientos dentro de las secuencias analizadas, facilitando el proceso de señalar zonas con grado de similaridad importante dentro de clústeres. La siguiente gráfica muestra el señalamiento de un cluster realizado:

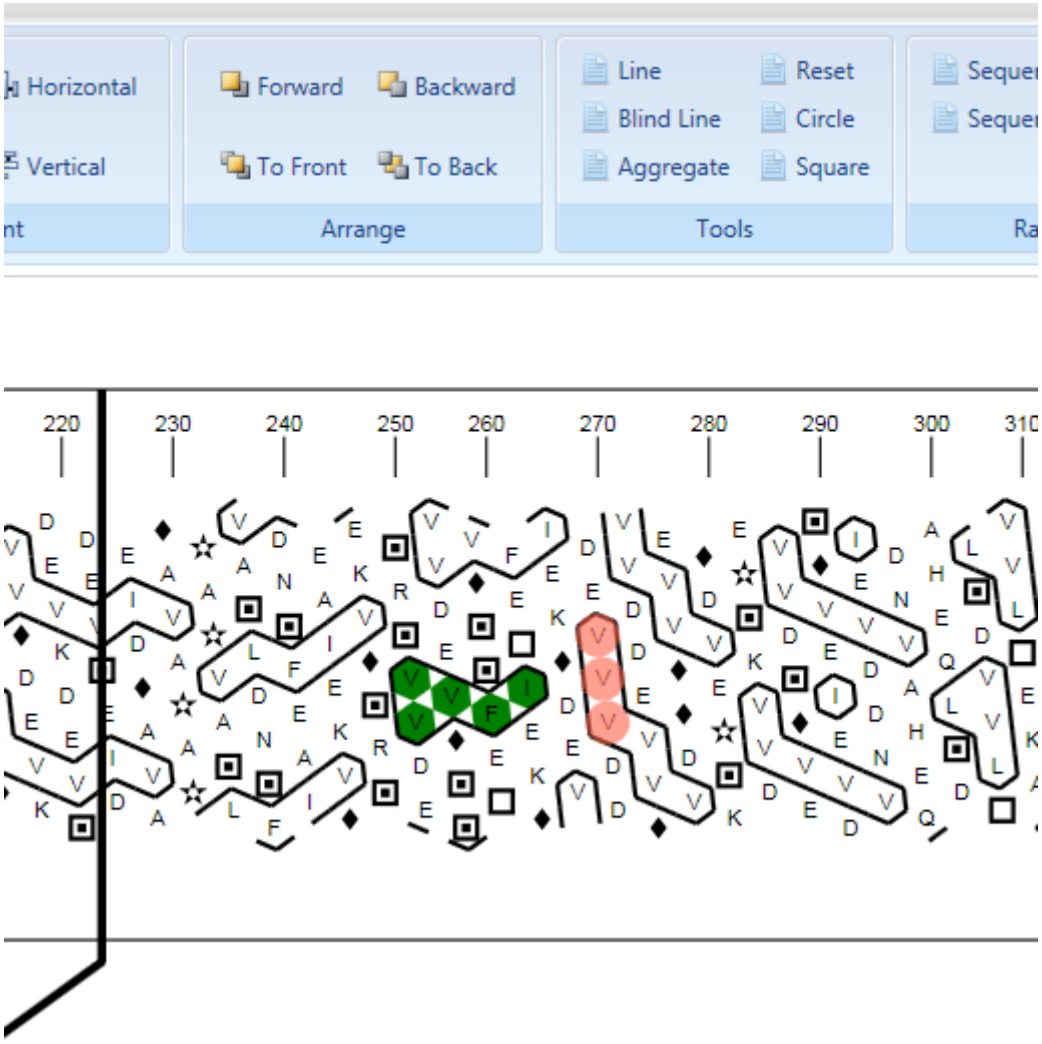


Figura 18. Creación de Agrupamientos.

6.3 Inserción de líneas ciegas

Las líneas ciegas se han incluido en el software con el fin de proveer una guía en el proceso de selección de gaps. Se describen como líneas que actúan como probables gaps, pero no tenidos en cuenta dentro del alineamiento final; se representan como una línea punteada que une dos aminoácidos para formar un gap provisional, solamente indicativo en el proceso de alineamiento. La grafica muestra una línea ciega creada entre dos aminoácidos.

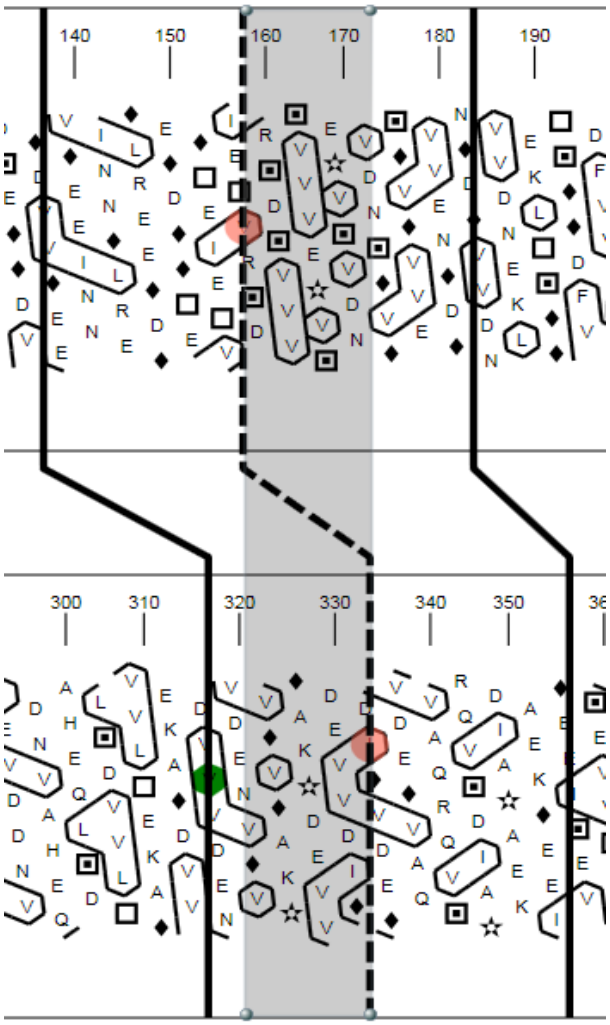


Figura 19. Líneas ciegas.

6.4 Inserción de texto

Es útil para el investigador crear anotaciones sobre el alineamiento realizado con el fin de señalar aspectos de estructura probable de las secuencias analizadas a partir del trabajo realizado. La herramienta provee un control drag and drop llamado Text que permite la inclusión de texto definido por el usuario y la ubicación del mismo dentro del alineamiento en el sitio definido por el usuario. La siguiente grafica muestra texto agregado al alineamiento:

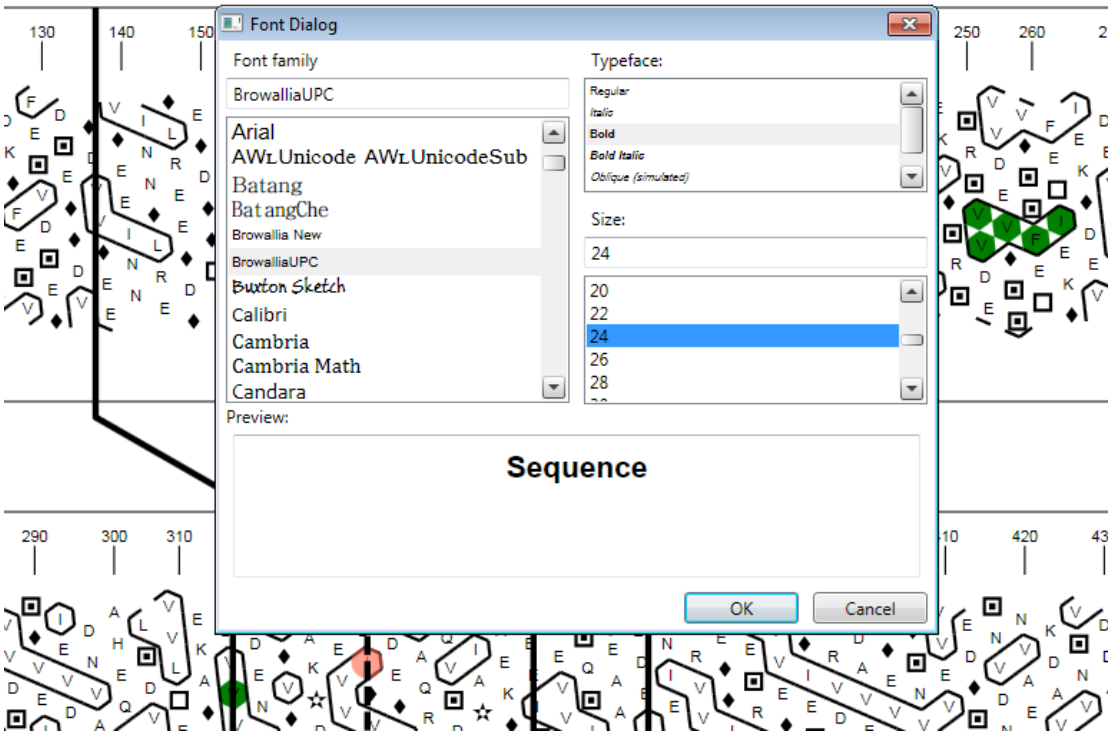


Figura 20. Inserción de texto.

6.5 Inserción de símbolos

Algunas figuras son utilizadas con el fin de representar el comportamiento estructural de las secuencias solamente observable en un entorno 3D, en las secuencias analizadas utilizando el método HCA, para esto se incluyen figuras como cilindros y flechas que describen la forma que podría tener la secuencia a partir del trabajo realizado, como hélices u hojas. El software provee controles gráficos drag and drop que le permiten al usuario ubicar y

dimensionar estos símbolos en el área de trabajo a su gusto, como podemos observar en la gráfica:

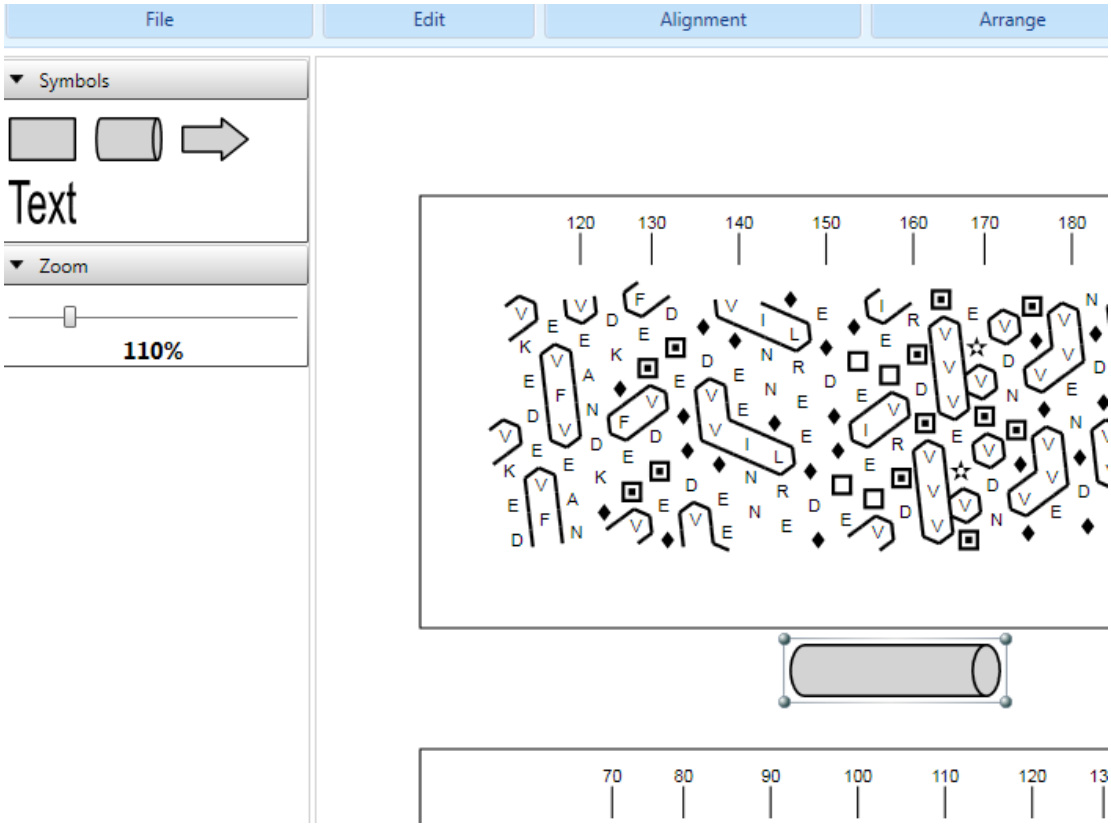


Figura 21. Inserción de símbolos.

7. Calculo de Identidad del alineamiento

La conclusión del trabajo realizado por el investigador se constituye en el cálculo de identidad de las secuencias analizadas. Este resultado permite analizar si el alineamiento realizado es consistente.

Para el cálculo de identidad se desarrolló un algoritmo que le permite al investigador obtener un alineamiento final 1D a partir del trabajo realizado con el método HCA, esto hecho con el fin de comparar el resultado obtenido con métodos 1D como ClustalW, y obtener el cálculo del porcentaje de identidad entre las secuencias analizadas.

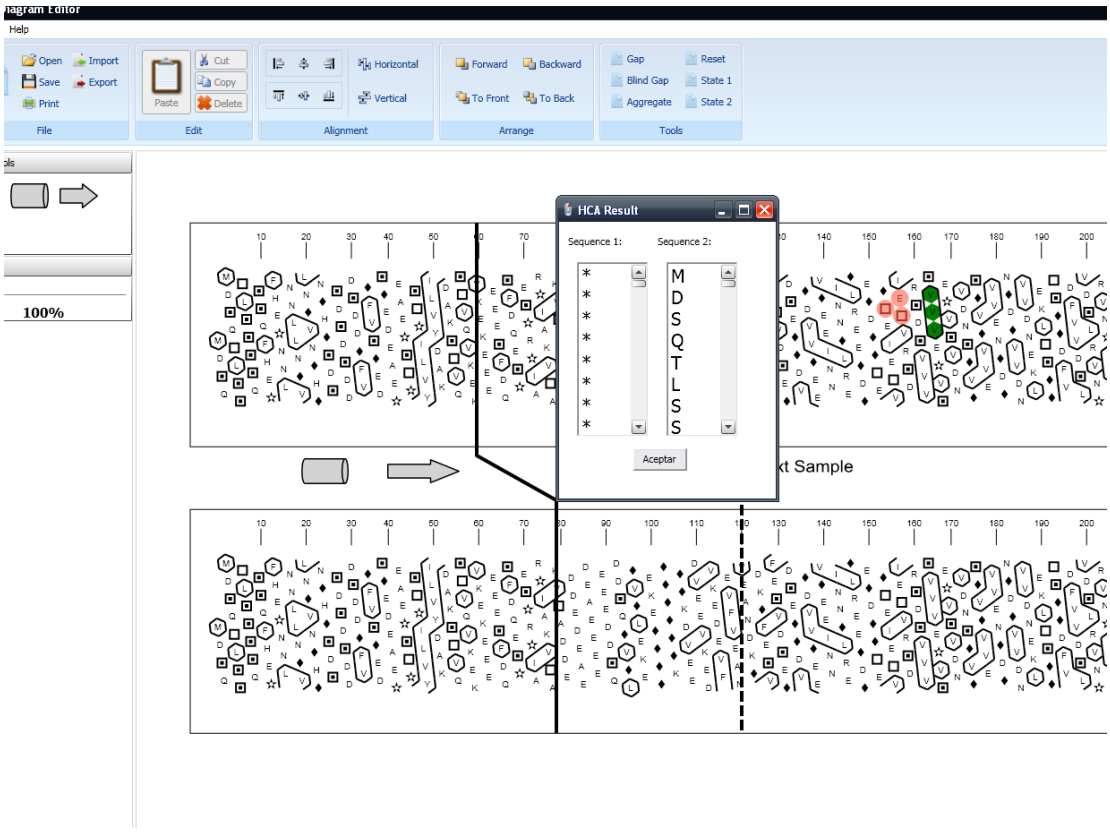


Figura 22. Conversión 1D.

8. Impresión de resultados

Teniendo en cuenta que los investigadores del área constantemente publican artículos utilizando como metodología de análisis al método HCA, se agregó la funcionalidad de impresión del trabajo realizado por el investigador. A continuación se aprecia en la gráfica el proceso de impresión de los resultados de un alineamiento.

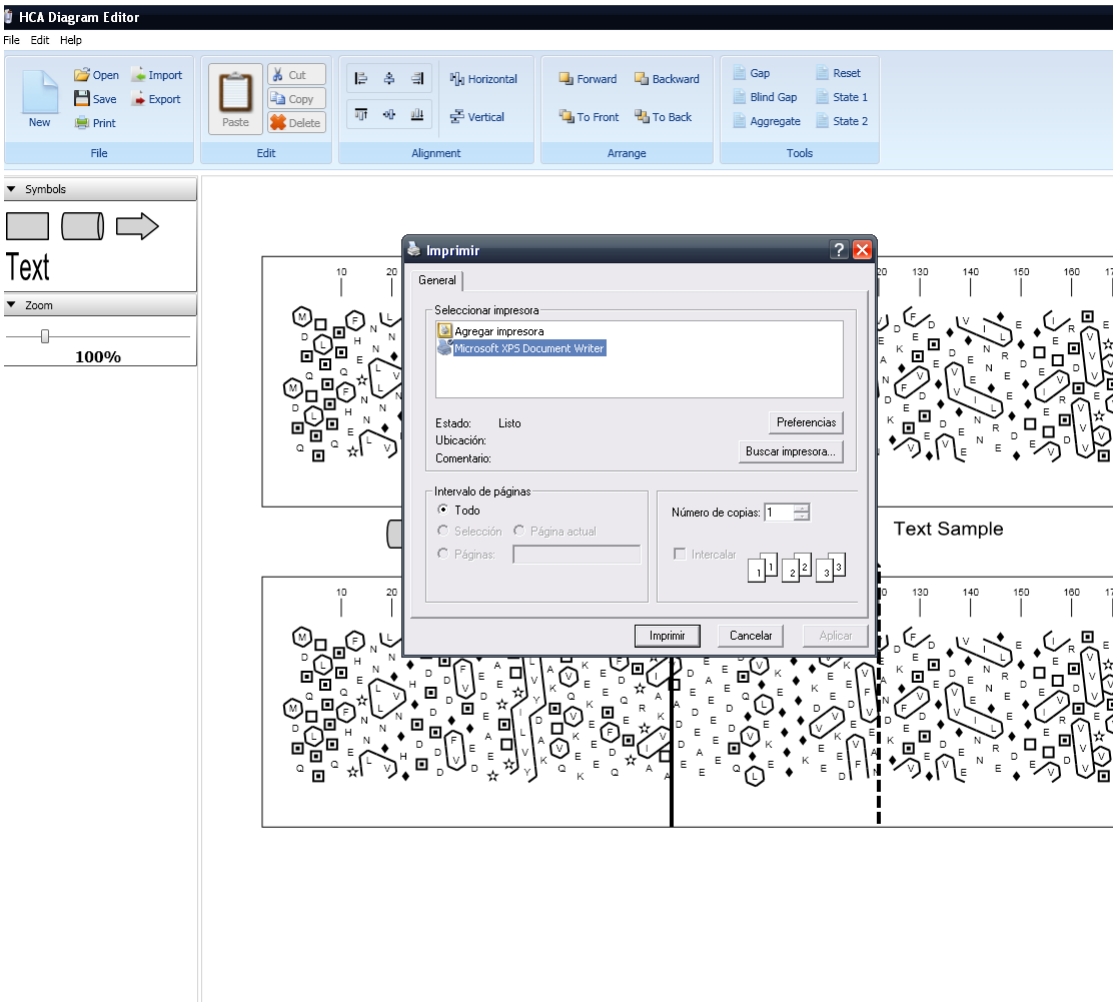


Figura 23. Impresión.

9. Guardar la información

El desarrollo de un alineamiento es un trabajo extensivo, por lo cual fue necesario proveer el guardado del trabajo realizado, para permitirle al usuario retomar el análisis parcial y continuarlo. Se creó un formato con extensión (.hca) que distingue los proyectos realizados con la herramienta y da consistencia a la información presentada.

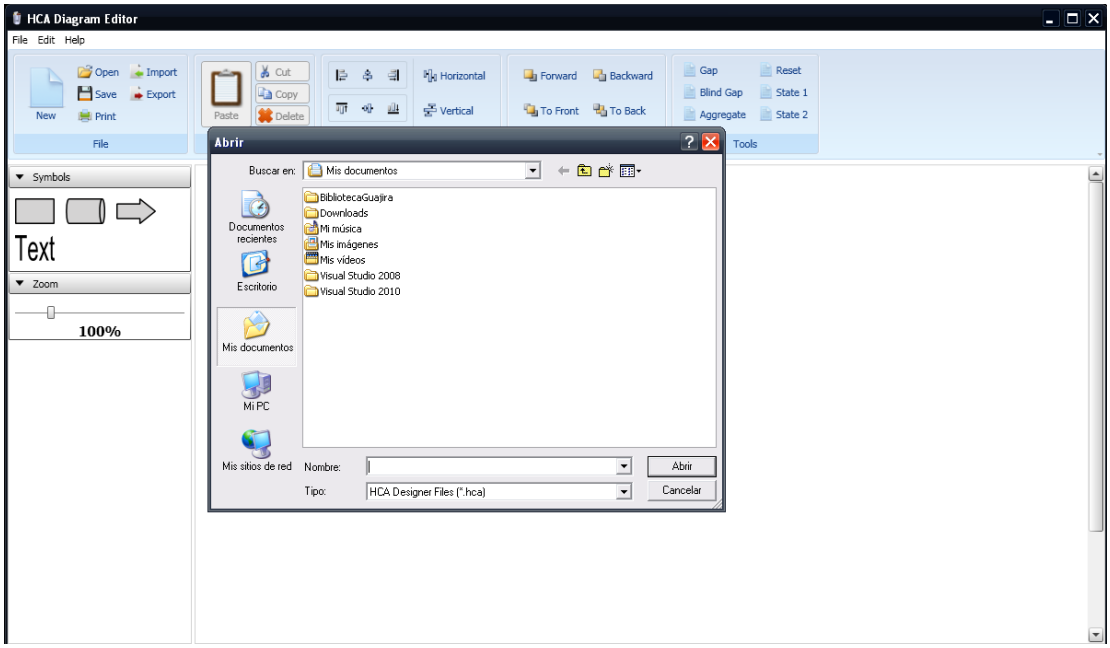


Figura 24. Guardar información.

10. Algoritmo de conversión HCA a 1D

Para explicar el proceso de transformación del trabajo gráfico HCA a un formato de una dimensión, empezaremos por definir algunos objetos utilizados en el desarrollo.

La representación gráfica 2D de los elementos de desarrolla dentro de un control tipo CANVAS en el cual se distinguen los siguientes elementos gráficos:

- Contenedores de secuencias: contenedor en el que se encuentra presente una secuencia de aminoácidos.

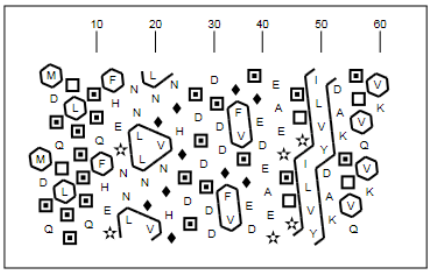


Figura 25 secuencia HCA.

- Contenedor de aminoácidos: contiene los aminoácidos pertenecientes a la secuencia y sus propiedades graficas asociadas.

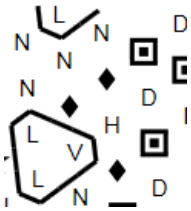


Figura 26. Detalle aminoácidos.

- Líneas: indican el trazo que tiene un gap en dos o más secuencias. Estas líneas están asociadas a un aminoácido de cada secuencia presente.

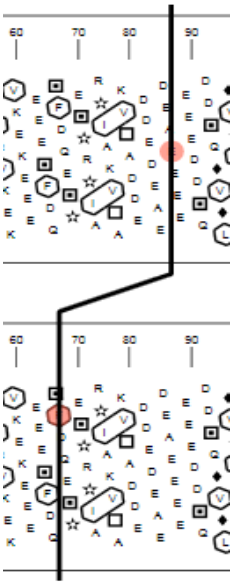


Figura 27. Líneas HCA.

Partiendo de esta información planteamos el desarrollo de un algoritmo que permite el paso de esta representación 2D a un plano 1D

Asumiendo los siguientes datos para el desarrollo de esta solución:

- Se posee un listado con los “Contenedores de secuencias “que ahora vamos a definir como **Secuencias.**

- Hay otra lista presente, “Líneas” llamada **Lineas** que a su vez posee una lista con unos identificadores únicos que indican a que aminoácido está asociado en cada contenedor de secuencias.

El desarrollo del algoritmo en forma de pseudocódigo es el siguiente:

Proceso a_1D

Leer *Secuencias*;

//Es una Lista que contiene las secuencias en el proceso HCA Ordenadas en el mismo orden en que fueron dibujadas en pantalla

Si El número de elementos de *Secuencias* es mayor que cero **Entonces**

Leer *Resultados*;

//Una Lista que contiene cadenas de texto cuyo tamaño es igual a lista *Secuencias*, aquí es donde se almacenan las secuencias aplicando el proceso de 2D a 1D

Leer *Lineas*;

//Es una lista que contiene las *lineas* que representan el alineamiento en el método HCA ordenadas de izquierda a derecha

Si El número de elementos de *Lineas* es 0 **Entonces**

FinProceso

FinSi

Para *i*=0 Menor Nro. de elementos de *Lineas* **Con Paso** 1 **Hacer**

Distancia=0;

Para j=0 Menor Nro. de elementos *Secuencias* **Con Paso** 1 **Hacer**

Temp=0;

Si *i* es igual a 0 **Entonces**

temp = La Posición del aminoácido que pasa por la *Lineas*[i] y que pertenece a la *Secuencias*[j] - La Posición inicial de la *Secuencias*[j];

//La Posición de un aminoácido se determina como el lugar que ocupa en la secuencia

//La Posición inicial de la *Secuencias*[j] se define como la posición del primer aminoácido que se dibuja en la grafica de HCA

Sino

temp = La Posición del aminoácido que pasa por la *Lineas*[i] y que pertenece a la *Secuencias*[j] - La Posición del aminoácido que pasa por la *Lineas*[i-1] y que pertenece a la *Secuencias*[j];

FinSi

Si *temp* es mayor que *distancia* **Entonces**

distancia = *temp*;

FinSi

FinPara

Para *j*=0 Menor Nro. de elementos *Secuencias* **Con Paso** 1 **Hacer**

temp=0;

Si *i* es igual a 0 **Entonces**

temp = La Posición del aminoácido que pasa por la *Lineas*[i] y que pertenece a la *Secuencias*[j] - La Posición inicial de la *Secuencias*[j]

//La Posición de un aminoácido se determina como el lugar que ocupa en la secuencia

//La Posición inicial de la *Secuencias*[j] se define como la posición del primer aminoácido que se dibuja en la grafica de HCA

Sino

temp = La Posición del aminoácido que pasa por la *Lineas*[i] y que pertenece a la *Secuencias*[j] - La Posición del aminoácido que pasa por la *Lineas*[i-1] y que pertenece a la *Secuencias*[j];

FinSi

Para *k*<-0 Menor que *distancia* - *temp* **Con Paso** 1 **Hacer**

Resultados[j]= *Resultados*[j] + '-';

//Agregamos gaps en las secuencias que los necesiten para lograr el alineamiento

FinPara

Si *i* es igual a 0 **Entonces**

Resultados[j]=*Resultados*[j] + Los aminoácidos de la *Secuencias*[j] desde la Posición inicial hasta la posición *temp* + 1;

Sino

Resultados[j]=*Resultados*[j] + Los aminoácidos de la *Secuencias*[j] desde (La posición del aminoácido que pertenece a *Secuencias*[j] que pasa por la *Lineas* [i-1])+ 1 hasta *temp;*

FinSi

FinPara

FinPara

Para *i*<-0 Hasta El número de elementos de *Secuencias* **Con Paso** 1 **Hacer**

Temp=La secuencia completa de aminoácidos de *Secuencias*[i] ;

Pos = La posición del aminoácido que pertenece a *Secuencia*[i] y que pasa por la *Lineas*[Nro de elementos de *Lineas* - 1];

Resultados[j]=*Resultados*[j]+ La subcadena de *temp* desde la posición pos + 1 hasta (La Posición final de la *Secuencias*[j] - pos)

//La Posición final de la Secuencias[j] se define como la posición del ultimo aminoácido que se dibuja en la grafica de HCA

FinPara

maxima_longitud<-La secuencia más larga que este almacenada en la lista *Resultados*;

Para *z*<-0 Hasta El número de elementos de *Resultados* **Con Paso** 1 **Hacer**

Si La longitud de *Resultados*[*z*] es menor que *maxima_longitud* **Entonces**

Resultados[z]<- *Resultados*[z] + '-' Repetido (*maxima_longitud* - la longitud de *Resultados*[z]) veces;

FinSi

FinPara

FinSi

FinProceso

11. Resultados

La herramienta HCA editor fue utilizada en la elaboración de alineamientos por parte de los investigadores de la escuela de biología. La siguiente gráfica muestra un análisis realizado:

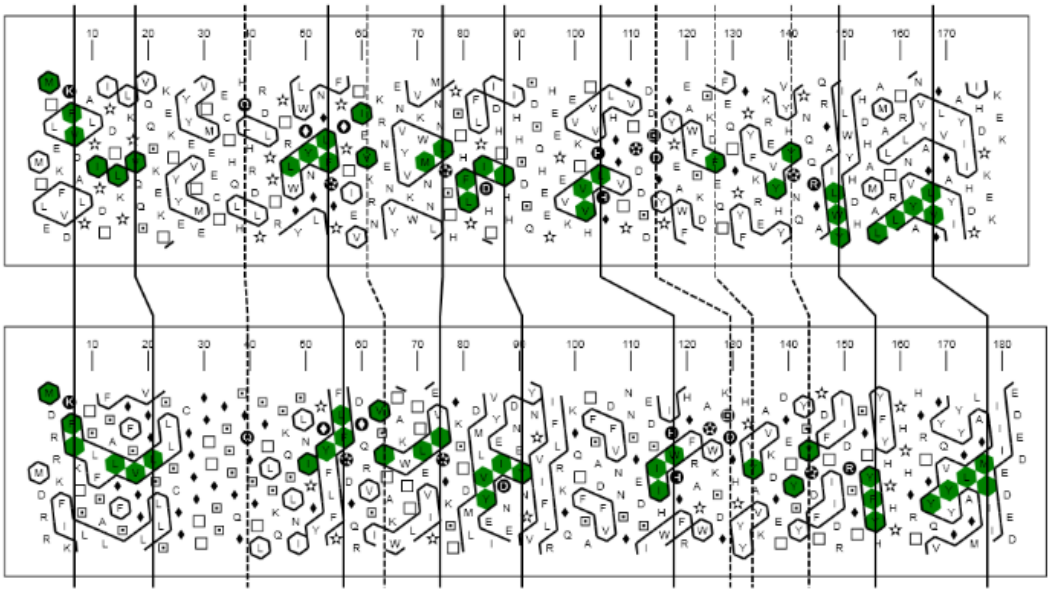


Figura 28. Alineamiento completo.

Se comprobó la facilidad con que se pueden trabajar los alineamientos realizados en HCA, ya que HCA EDITOR suministra herramientas a la medida como la creación de gaps, agregados, líneas ciegas, entre otras, con las que los investigadores pueden realizar el trabajo gráfico en un menor tiempo y con mayor confiabilidad.

Además se comprobó la eficacia del análisis HCA a 1D realizado por HCA EDITOR, que entrega resultados confiables al instante, proceso que comúnmente requiere que los investigadores lo realicen a mano y varias veces para comprobar si se realizaron correctamente los cálculos. A continuación veremos el resultado 1D del alineamiento realizado:

```
MTLEKFV---DALPIPDTLKPVQQSKEKTYYEVTMEECTHQ-LHRDLPPTRLWGYNGLF-PGPTIEVKRNENVYVKWMNNL----PSTHFLPIDHTI
MDRRKFIKTSLFSALGFSVGGL--SLLSCGGGGTTGSSSGQGSGTLSKQSLNIPGYFLFPDGQRVSI----TAKWTTLEVIPGKSTDMLVYEIDNEY
*   **                 *       *     *              **  *                            **


----------HHSDSQHEEPEVKTVVHLHGGVTPDDSDGYPEAWFSKDFEQTGPYFKREVYHY---PNQQRGAILWY---HDHAMALTRLNVYAGLVGAYIIH
NPVIFLRKGQTFSADFVNNSGEDSIIHWHGFRAPWKSD------GHPYY----AVKDGETYSYPDFTIIDRSGTYFYHPHPHGRTGYQVYYGLAGMIIIE---
           *            * **   *  **                  * * *      *    *              **
```

Identity = (25/180)= 13,88%

12. Conclusiones

- Se desarrolló un software prototipo llamado HCA EDITOR como herramienta de apoyo al proceso de análisis de secuencias y obtención de resultados en la utilización del método HCA.

- Se desarrollaron herramientas a la medida que le permiten al investigador la creación de gaps, líneas ciegas, agregados y la inclusión de texto y símbolos referentes al análisis realizado.

- Se implementó un algoritmo de conversión entre el trabajo realizado utilizando el método HCA y el formato utilizado por los programas tradicionales 1D como ClustalW, para comprobar la efectividad del trabajo realizado en la herramienta frente a otros métodos.

- Se implementó un algoritmo de creación de clústeres en la gráfica HCA, basado en el código suministrado por los creadores del método.

- Se desarrolló la opción de impresión que permite al usuario exportar el trabajo realizado en formatos como (.bmp,.jpg,.gif).

- Se creó la opción de guardado del trabajo realizado con extensión de archivo (.hca) que permite al investigador realizar trabajos parciales de alineamientos sin perder el proceso de análisis de los datos.

- Se agregó la funcionalidad de cálculo de identidad de las secuencias alineadas.

- Se incluyó un módulo de cálculo del HCA Score.

- La herramienta HCA EDITOR cuenta con una interfaz gráfica que facilita el trabajo del investigador, ya que provee controles drag and drop con características de posicionamiento y redimensionamiento intuitivas y fáciles de trabajar.

Bibliografía

1. Altschul,S., Madden,T., Schaffer,A., Zhang,J., Zhang,Z., Miller,W., and Lipman,D. (1997) . Gapped-BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. *25*, 3389-3402.

2. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., and Wheeler,D.L. (2008) . GenBank. Nucleic Acids Res *36*, D25-D30.

3. Berman,H.M., Westbrook,Z., Gilliland,F.G., Bhat,T.N., Weissig,H., Shindyalov,I.N., and Bourne,P.E. (2008) . The Protein Data Bank. Nucleic Acids Res. *28*, 235-242.

4. Callebaut,I., Labesse,G., Durand,P., Poupon,A., Canard,L., Chomilier,J., Henrissat,B., and Mornon,J.P. (1997a) . Deciphering protein sequence information through hydrophobic cluster analysis (HCA) : current status and perspectives. Cell. Mol. Life. Sci. *53*, 621-645.

5. Callebaut,I., Labesse,G., Durand,P., Poupon,A., Canard,L., Chomilier,J., Henrissat,B., and Mornon,J.P. (1997b) . Deciphering protein sequence information through hydrophobic cluster analysis (HCA) : current status and perspectives. Cell. Mol. Life Sci. *53*, 621-645.

6. Chen,B. and Johnson,M. (2009) . Protein local 3D structure prediction by Super Granule Support Vector Machines (Super GSVM) . BMC Bioinformatics *10*, S15.

7. Gerlt,J.A. and Babbitt,P.C. (2000) . Can sequence determine function? Genome Biol. *1*, 1-10.

8. Henikoff,S. and Henikoff,J.G. (1994) . Protein family classification based on searching a database of blocks. Genomics *19*, 97-107.

9. Joardar,V. *et al.* (2005) . Whole-genome sequence analysis of *Pseudomonas syringae* pv. phaseolicola 1448A reveals divergence among pathovars in genes involved in virulence and transposition. J Bacteriol *187*, 6488-6498.

10. Kresge,N., Simoni,R.D., and Hill,R.L. (2005) . The Elucidation of the Structure of Ribonuclease by Stanford Moore and William H. Stein. J. Biol. Chem. *280*, e47.

11. Larkin,M.A. *et al.* (2007) . Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

12. Pearson,W.R. (2001) . Protein sequence comparison and protein evolution. Department of Biochemistry and Molecular Genetics,Jordan Hall, Box 800733University of Virginia, Charlottesville, VA 22908, USA.

13. Pressman,S.R. Ingenieria del Software un enfoque practico. 6, 52-53. 2005.

14. Rost,B. (1999) . Twilight zone of protein sequence alignments. Protein Eng. *12*, 85-94.

15. Sanger,F. (1959) . Chemistry of Insulin: Determination of the structure of insulin opens the way to greater understanding of life processes. Science *129*, 1340-1344.

16. Steen,H. and Mann,M. (2004) . The abc's (and xyz's) of peptide sequencing. Nat. Rev. Mol. Cell Biol. *5*, 699-711.

17. US National Center for Biotechnology Information. FASTA format description. 2010.

18. Woodcock,S., Mornon,J.-P., and Henrissat B. (1992) . Detection of secondary structure elements in proteins by Hydrophobic Cluster Analysis. Protein Eng. *5*, 629-635.

19. Real Academia Española. Diccionario de la lengua española. 2010.