

**IMPLEMENTACIÓN DEL ALGORITMO SMITH Y WATERMAN PARA EL  
ANÁLISIS DE ALINEAMIENTOS DE SECUENCIAS PROTEÓMICAS**

**DANIEL GALAVÍS IBÁÑEZ  
EDGAR JIMÉNEZ SUÁREZ**

**DIRECTOR  
ING. JUAN CONTRERAS**

**CORPORACIÓN UNIVERSITARIA RAFAEL NÚÑEZ  
FACULTAD DE INGENIERÍA  
PROGRAMA DE SISTEMAS  
CARTAGENA DE INDIAS, OCTUBRE 10 DE 2005**

**IMPLEMENTACIÓN DEL ALGORITMO SMITH Y WATERMAN PARA EL  
ANÁLISIS DE ALINEAMIENTOS DE SECUENCIAS PROTEÓMICAS**

**DANIEL GALAVÍS IBÁÑEZ  
EDGAR JIMÉNEZ SUÁREZ**

**Proyecto de Grado presentado como requisito  
Para optar al título de Ingeniero de Sistemas**

**DIRECTOR  
ING. JUAN CONTRERAS**

**CORPORACIÓN UNIVERSITARIA RAFAEL NÚÑEZ  
FACULTAD DE INGENIERÍA  
PROGRAMA DE SISTEMAS  
CARTAGENA DE INDIAS, OCTUBRE 10 DE 2005**

Nota de aceptación

---

---

---

Presidente del jurado

---

Jurado

---

Jurado

## DEDICATORIA

A DIOS, por guiarnos y darnos la fortaleza necesaria para la culminación de este proyecto.

A nuestros Padres, por impulsarnos a luchar por este sueño, por su abnegación, su constante apoyo y motivación.

## AGRADECIMIENTOS

A Juan Contreras, nuestro director y maestro por toda su colaboración para la realización de este trabajo de grado.

A María Claudia Bonfante, Justo Sarabia, Lisbeth Urueta y demás profesores de la facultad por sus valiosas enseñanzas.

A nuestras familias, por su motivación, paciencia, fe y colaboración para la culminación de este proceso.

## RESUMEN

El creciente volumen de datos derivado de miles proyectos de investigación científica, entre ellos el desarrollo de técnicas de separación y análisis de proteínas, han hecho ver la necesidad de desarrollar e implementar herramientas computacionales para el análisis y manipulación de esta información.

Teniendo en cuenta lo anterior en este trabajo se realiza la implementación del algoritmo Smith y Waterman el cual es una herramienta diseñada para efectuar alineamientos en secuencias de proteínas, con la finalidad de medir la similitud o identidad entre ellas.

## TABLA DE CONTENIDO

	Pág.
INTRODUCCIÓN	1
PROBLEMA DE INVESTIGACIÓN	1
PLANTEAMIENTO DEL PROBLEMA	2
JUSTIFICACIÓN	2
OBJETIVOS	3
OBJETIVO GENERAL	3
OBJETIVOS ESPECÍFICOS	4
MARCO REFERENCIAL	4
MARCO TEÓRICO	5
MARCO CONCEPTUAL	6
1. PROTEÍNAS	8
1.1 GENERALIDADES SOBRE LAS PROTEÍNAS	8
1.1.1 Estructura de las proteínas	9
1.1.2 Propiedades de las proteínas	12
1.1.3 Clasificación de las proteínas	13
1.2 PROTEÓMICA	14
2. ALINEAMIENTO DE SECUENCIAS	17
2.1 SIMILITUD	18
2.2 HOMOLOGÍA	19
2.3 MATRICES DE SUSTITUCIÓN	19
2.3.1 Matrices de sustitución para proteínas	20
2.3.2 Pam	21
2.3.3 Blosum	22
2.4 HERRAMIENTAS DE BÚSQUEDA DE SIMILITUD	23
2.4.1 Blast	26
2.4.2 Fasta	26
3. ALGORITMOS DE ALINEAMIENTO DE SECUENCIAS	27
3.1 Algoritmo de Smith Y Waterman	27
4. ANÁLISIS DEL ALINEADOR	31
4.1 FUNCIONES DEL SISTEMA	31
4.2 DIAGRAMA DE CASOS DE USO DEL ALINEADOR	32
4.2.1 Descripción de Casos de Usos del ALINEADOR	33
5. DISEÑO DEL ALINEADOR	37
5.1 MODELO DE CLASES	37
5.2 DESCRIPCIÓN DIAGRAMA DE CLASES	38
5.3 DIAGRAMA DE SECUENCIAS DEL ALINEADOR	43
5.4 DISEÑO DE LA INTERFAZ GRÁFICA	43
6. DICCIONARIO DE DATOS DEL ALINEADOR	48
6.1 DESCRIPCIONES DE LOS PROCESOS DE NIVEL CONTEXTUAL.	48
6.2 DESCRIPCIONES DE LOS FLUJOS DE DATOS EN EL NIVEL CONTEXTUAL.	49

7. DESARROLLO, IMPLEMENTACIÓN Y PRUEBAS DEL ALINEADOR	50
BIBLIOGRAFÍA	53



## LISTA DE FIGURAS

	Pág.
Figura 1.1 (a) Aminoácido, (b) Enlace Peptídico	8
Figura 1.2 Estructura Primaria de la Proteína	10
Figura 1.3 Estructura Secundaria $\alpha$ (alfa)-hélice	11
Figura 1.4 Estructura Secundaria conformación beta	11
Figura 1.5 Estructura Terciaria	11
Figura 1.6 Estructura Cuaternaria	12
Figura 2.1 Ecuaciones para la construcción de la Matriz PAM	22
Figura 3.1 Calculo del valor de $H_{ij}$ (Matriz de Resultado)	28
Figura 4.1 Diagrama de Casos de Uso del ALINEADOR	32
Figura 4.2 Flujo de Datos General para el ALINEADOR	35
Figura 4.3 Flujo de Datos Conceptual	35
Figura 4.4 Flujo de Datos Nivel 1	35
Figura 4.5 Flujo de Datos nivel 2	36
Figura 4.6 Flujo de Datos nivel 3	36
Figura 5.1 Diagrama de Clase del ALINEADOR	37
Figura 5.2 Diagrama de Secuencias del ALINEADOR	43
Figura 5.3 Diseño de la pantalla	44
Figura 5.4 Interfaz de Usuario	45
Figura 5.5 Menú Archivo	45
Figura 5.6 Menú Edición	46
Figura 5.7 Menú Ayuda	46
Figura 5.8 Barra de Herramientas	47
Figura 5.9 Zona Ingreso de Datos	47
Figura 5.10 Zona de Resultados	47

## LISTA DE TABLAS

	Pág.	
Tabla 1.1	Holoproteínas	13
Tabla 1.2	Heteroproteínas	14
Tabla 2.1	Identificación de Aminoácidos	18
Tabla 2.2	Matriz de Sustitución sencilla	20
Tabla 2.3	Matriz de Sustitución de Ocurrencias	20
Tabla 2.4	Comparación de los programas utilizados en la búsqueda de secuencia en las bases de datos	25
Tabla 3.1	Matriz H (Matriz de Resultado)	29
Tabla 3.2	Máximo valor en Matriz de Resultado y Residuo	41
Tabla 3.3	Inicio de Traceback	41
Tabla 3.4	Recorrido completo en Matriz de Resultado	42
Tabla 4.1	Funciones Básicas	31
Tabla 4.2	Descripción Caso de Uso Ingresar Valores	33
Tabla 4.3	Descripción Caso de Uso Ejecutar ALINEADOR	34
Tabla 5.1	Descripción de la Clase Principal	38
Tabla 5.2	Descripción de la Clase DatosAlineamiento	39
Tabla 5.3	Descripción de la Clase Secuencia	41
Tabla 5.4	Descripción de la Clase Matriz	41
Tabla 7.1	Pruebas al ALINEADOR	50

## LISTA DE ANEXOS

Anexo A	MANUAL DE USUARIO	56
Anexo B	ARTÍCULO CIENTÍFICO. IMPLEMENTACIÓN DEL ALGORITMO SMITH Y WATERMAN PARA EL ANÁLISIS DE ALINEAMIENTOS DE SECUENCIAS PROTEÓMICAS	67
Anexo C	MATRICES DE SUSTITUCIÓN	71
Anexo D	LECTURA DE INFORMACIÓN EN PROSITE	84

## **INTRODUCCIÓN**

### **PROBLEMA DE INVESTIGACIÓN**

La gran cantidad de información contenida actualmente en las bases de datos y los proyectos de estudios masivos de interacción entre proteínas traen consigo algunas consecuencias importantes: como primera medida permiten plantear estudios que hasta ahora eran inabordables; pero por otra parte exigen un cambio en la mentalidad y en las herramientas informáticas de tratamientos de datos.

Se está pasando de una época en la que se disponía de un número relativamente bajo de datos, a otra en la cual se dispone de grandes cantidades de ellos, de los que se sabe mucho menos en detalle pero de los que se pueden comprender sus propiedades globales como sistema.

Esta nueva época constituye un verdadero cambio de paradigma en bioquímica y también en el campo de la Bioinformática, debido a que el tipo de herramientas informáticas y las estrategias para procesar datos deben ser reorientadas.

Hablar de Bioinformática se refiere a un amplio campo de actividades que utilizan la tecnología de la información para organizar, analizar y distribuir información biológica con la finalidad de responder preguntas complejas en biología, incluyendo, entre otras, el análisis y la interpretación de varios tipos de datos, como secuencias de nucleótidos y aminoácidos, dominios de proteínas y estructura de proteínas.

La Bioinformática es una disciplina sin la cual no se puede dar el paso requerido por la aplicación de nuevas metodologías de muestreo masivo de datos, además ella permite generar conocimiento nuevo a partir de la integración de la información y los datos disponibles.

Por otro lado, la Proteómica es una disciplina científica que, al igual que la genómica, es informacional. Los estudios proteómicos no sólo incluyen la identificación y la cuantificación de proteínas, sino también la determinación de su localización, modificaciones, interacciones, actividades y, en última instancia, la determinación de su función. Una de las maneras mas frecuentes de obtener información sobre una o un grupo de secuencias de proteínas incógnitas, es mediante la búsqueda comparativa con otras proteínas existentes. Uno de los métodos comparativos más comunes es el alineamiento de pares de secuencias, que consiste en introducirles espacios para destacar su parecido.

## **PLANTEAMIENTO DEL PROBLEMA**

Debido a los avances en la investigación en ciencias biomédicas, se ha producido un enorme crecimiento en el volumen y en la complejidad de la información biológica generada por la comunidad científica. Anteriormente, los resultados de los experimentos podían interpretarse sobre el cuaderno de laboratorio, actualmente no es conveniente comparar las secuencias de varios nucleótidos o aminoácidos de manera manual, es necesario el uso de técnicas y herramientas las cuales tienen menos posibilidades de arrojar errores que un acercamiento manual.

Las herramientas que se han desarrollado buscan garantizar el obtener más información sobre las secuencias, aprovechando el conocimiento previo almacenado en bases de datos.

## **JUSTIFICACIÓN**

La gran cantidad de información generada por el estudio de la biología molecular así como el progreso en los diferentes proyectos de secuenciación de genomas, trajo consigo la necesidad de organizar y almacenar toda esta información en

bases de datos clasificadas (secuencias, estructuras, expresión de genes, rutas metabólicas, etc.) para su análisis.

Debido a esto se hace necesario la implementación y el desarrollo de herramientas computacionales útiles, como algoritmos, los cuales buscan patrones significativos y relaciones entre las diferentes secuencias de proteínas, complementándose con el uso de procedimientos matemáticos y estadísticos que permiten relacionar los diferentes tipos de datos disponibles.

Por ejemplo, métodos para la localización de genes en secuencias de nucleótidos, predecir características estructurales y funcionales de las proteínas, a partir de su secuencia aminoácida o agrupar las secuencias de aminoácidos en familias de proteínas relacionadas.

El mayor desarrollo de estas herramientas hará posible el progreso de la proteómica, al facilitar el estudio de las interacciones entre proteínas.

Técnicas como el Algoritmo de Smith y Waterman, permiten alinear secuencias de proteínas y cuyos resultados facilitan encontrar patrones de conservación, descubrir homólogos, construir taxonomías e Inferir los eventos del proceso evolutivo. La información obtenida, a su vez, tiene miles de aplicaciones, por lo que los alineamientos son la herramienta base de toda la bioinformática.

## **OBJETIVOS**

### **OBJETIVO GENERAL:**

Obtener la información teórico-práctica necesaria para implementar el algoritmo de Smith y Waterman en el análisis de alineamientos de secuencias proteómicas.

## **OBJETIVOS ESPECÍFICOS:**

- Recopilar y analizar información bibliográfica sobre proteínas, secuencias y alineamientos.
- Analizar información sobre los algoritmos empleados para el análisis de secuencias.
- Indagar sobre las herramientas computacionales existentes que realicen alineamientos de secuencias utilizando el algoritmo de Smith y Waterman.
- Realizar la implementación del algoritmo de Smith y Waterman para el análisis de secuencias proteómicas.

## **MARCO REFERENCIAL**

El término “proteoma” fue usado por vez primera en 1995 para describir el conjunto de PROTEÍNAS de un Genoma, una célula o un tejido. De forma imperceptible, la palabra proteoma dio lugar a una nueva disciplina, la “proteómica”.

El término "proteoma" fue acuñado por Mark Wilkins, para designar la totalidad de proteínas codificadas por un genoma, aunque el término ha ido perdiendo el contexto general y se ha restringido a designar al conjunto de proteínas que expresa un determinado tipo celular en un determinado momento. De este modo el término "proteoma" ha pasado de ser un concepto difuso y abstracto a ser un término funcional asociado a una realidad material. Así pues, la proteómica es el estudio a gran escala de las proteínas, habitualmente por medio de métodos bioquímicos.

## MARCO TEÓRICO

Los estudios proteómicos no sólo incluyen la identificación y la cuantificación de proteínas, sino también la determinación de su localización, modificaciones, interacciones, actividades y, en última instancia, la determinación de su función.

Los ácidos nucleicos y las proteínas pueden ser similares en cuanto a su función, su estructura o su secuencia. El objetivo principal de la comparación entre las secuencias de dos o más biomoléculas es el establecimiento de inferencias en cuanto a su estructura y función, a partir de la similitud en las secuencias. Esto es posible muchas veces, pero otras no, se conocen muchos casos de proteínas con una similitud en secuencia muy baja que, sin embargo, adoptan estructuras tridimensionales similares y comparten la misma función.

Las secuencias suelen compararse alineándolas. Un alineamiento de secuencias constituye una representación cualitativa de la similitud entre dos o más secuencias. Existen dos variantes fundamentales de alineamiento: (1) el alineamiento de una secuencia contra otra u otras (de tipo uno-a-uno ó uno-a-muchos) y (2) el alineamiento simultáneo de varias secuencias entre sí (de tipo muchos-a-muchos), conocido como alineamiento múltiple. Ambas variantes se basan en los mismos principios generales establecidos para comparar una secuencia con otra, introducidos por Smith y Waterman y modificados luego por Needleman y Wunsch. Sin embargo, los algoritmos usados en uno y otro caso son diferentes, optimizados de acuerdo al objetivo final.

El principal problema en la comparación de secuencias consiste en encontrar todas las zonas de similaridad significativas entre dos o mas secuencias y determinar que es significativo cuando se habla de secuencias biológicas. Para cubrir la mayoría de las necesidades hay varios enfoques y distintos programas que se pueden utilizar.



La sensibilidad de los distintos programas a la hora de realizar el análisis de secuencias dependen del algoritmo empleado. Estos algoritmos generan matrices estadísticas que permiten alinear la secuencia problema con otras parecidas, almacenadas en las bases de datos. Las matrices más simples sacrifican cierto grado de significación en la comparación a cambio de una mayor velocidad. Por tanto la velocidad y sensibilidad de la búsqueda va a depender de la complejidad del algoritmo empleado y de la matriz que genere.

## **MARCO CONCEPTUAL**

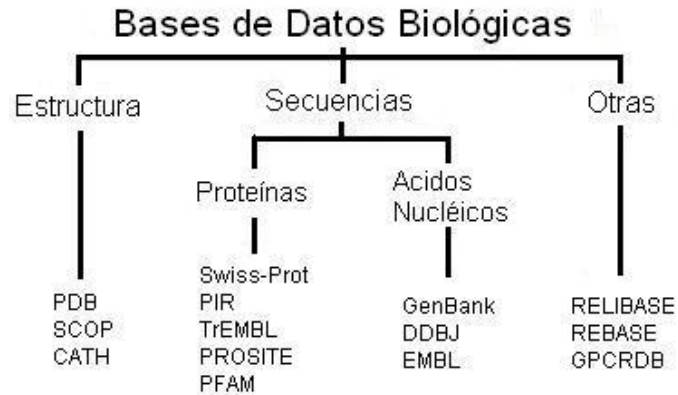
**Proteína.** Una macromolécula compuesta por una o más cadenas polipeptídicas, cada una con una secuencia característica de aminoácidos enlazados por enlace peptídico. Las proteínas son las principales macromoléculas que forman parte de los organismos vivos y son cruciales en prácticamente todas las funciones celulares.

**Secuencia.** Son una serie de elementos encadenados unos detrás de otros, por eso hablamos de secuencias de nucleótidos y de secuencias de aminoácidos. A través de letras podemos identificar los distintos monómeros que conforman la macromolécula (ejemplo: A: alanina; T: treonina; C: cisteína; y G: glicina; D: aspártico; E: glutámico; etcétera).

**Alineamiento de secuencias.** Arreglo mutuo de dos o más secuencias que muestra donde estas son similares y donde difieren. Un alineamiento óptimo es aquel que muestra la mayor cantidad de correspondencias y la menor cantidad de diferencias.

**Aminoácido.** Compuestos cuya estructura incluye un grupo amino, un grupo carboxilo y una cadena lateral de composición variable. Veinte aminoácidos conocidos como "estándar" son los principales monómeros de las cadenas polipeptídicas que forman las proteínas.

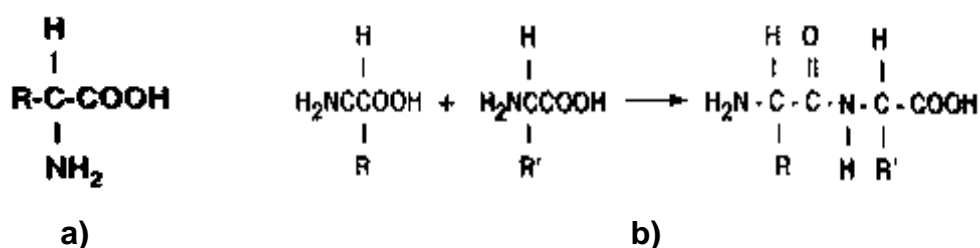
**Bases de datos de secuencias.** Las bases de datos de secuencias son repositorios primarios de datos que aceptan secuencias de ácidos nucleicos y proteínas procedentes de la comunidad científica internacional y los hacen disponibles en forma pública.



# 1. PROTEÍNAS

## 1.1 GENERALIDADES SOBRE LAS PROTEÍNAS

Las proteínas son compuestos orgánicos constituidos por moléculas de aminoácidos<sup>1</sup>, estos aminoácidos están formados por un grupo amino (-NH<sub>2</sub>) y un grupo carboxilo (-COOH) unidos mediante enlaces peptídicos, los cuales son enlaces que se establecen entre el grupo carboxilo y el grupo amino.



**Figura 1.1 (a) Aminoácido (b) Enlace Peptídico**

Las proteínas se componen básicamente por carbono, hidrógeno, oxígeno y nitrógeno. Pueden además contener azufre y en algunos tipos de proteínas, fósforo, hierro, magnesio y cobre entre otros elementos. Ellas intervienen en diversas funciones vitales esenciales, como el metabolismo, la contracción muscular o la respuesta inmunológica. Se descubrieron en 1838 y hoy se sabe que son los componentes principales de las células y que suponen más del 50% del peso seco de los animales.

Las moléculas proteicas van desde las largas fibras insolubles que forman el tejido conectivo y el pelo, hasta los glóbulos compactos solubles, capaces de atravesar la membrana celular y desencadenar reacciones metabólicas. Tienen un peso molecular elevado y son específicas de cada especie y de cada uno de sus órganos. Se estima que el ser humano tiene unas 30.000 proteínas distintas.

---

<sup>1</sup> Los péptidos son moléculas intermedias entre las de los aminoácidos, que son sencillas, y las moléculas complejas de las proteínas.

Las proteínas sirven sobre todo para construir y mantener las células, aunque su descomposición química también proporciona energía, con un rendimiento de 4 kilocalorías por gramo, similar al de los hidratos de carbono.

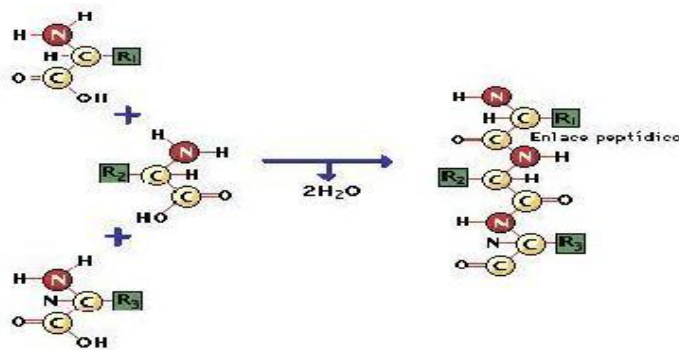
El número casi infinito de combinaciones en que se unen los aminoácidos y las formas helicoidales y globulares en que se arrollan las hileras o cadenas polipeptídicas, permiten explicar la gran diversidad de funciones que estos compuestos desempeñan en los seres vivos.

Los tipos y funciones de las proteínas son muy variados, existen proteínas que actúan como:

- Catalizadores en las reacciones químicas que tienen lugar en el ser vivo, denominándose las enzimas a este tipo, como ejemplo está la peptina.
- Proteínas contráctiles que forman parte de los músculos, como es el caso de la actina y miosina, son proteínas que se encargan del transporte de materiales como la hemoglobina de la sangre que transporta  $O_2$  y  $CO_2$ .
- Proteínas que intervienen en el proceso de coagulación de la sangre, otras proteínas actúan como anticuerpos, por ejemplo la inmunoglobulina.
- Proteínas con función de reserva, como la ovoalbúmina (clara del huevo) o la caseína. Se encuentran también proteínas estructurales como el colágeno, la tubulina, la elastina o la queratina.
- Proteínas reguladoras, las hormonas como la insulina.

1.1.1 Estructura de las Proteínas. La organización de una proteína viene definida por cuatro niveles estructurales denominados: estructura primaria, estructura secundaria, estructura terciaria y estructura cuaternaria. Cada una de estas estructuras informa de la disposición de la anterior en el espacio. Las diferentes secuencias de aminoácidos a lo largo de la cadena afectan de distintas formas a la estructura de la molécula de proteína.

- **Estructura Primaria.** La estructura primaria de una proteína es su secuencia de aminoácidos, se forma al unirse el grupo carboxilo —un átomo de carbono (C), dos átomos de oxígeno (O) y un átomo de hidrógeno H)— de un aminoácido con el grupo amino ( $\text{NH}_2$ ) del siguiente aminoácido, mediante un enlace peptídico. Cada proteína es una cadena larga constituida por muchos aminoácidos, y por cada enlace peptídico que se forma se libera una molécula de agua.



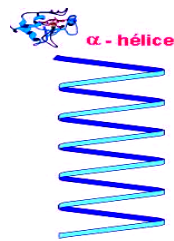
**Figura 1.2 Estructura Primaria de la Proteína**

Fuente: Enciclopedia Microsoft® ENCARTA®

- **Estructura Secundaria.** La estructura secundaria es la disposición de la secuencia de aminoácidos en el espacio. Fuerzas como los enlaces de hidrógeno, la atracción entre cargas positivas y negativas, y los enlaces hidrófobos (repelentes del agua) e hidrófilos (afines al agua) hacen que la molécula se enrolle o pliegue y adopte una estructura secundaria.

Existen dos tipos de estructura secundaria:

**La  $\alpha$ (alfa)-hélice:** Esta estructura se forma al enrollarse helicoidalmente sobre sí misma la estructura primaria. Se debe a la formación de enlaces de hidrógeno entre el  $-\text{C}=\text{O}$  de un aminoácido y el  $-\text{NH}-$  del cuarto aminoácido que le sigue.



**Figura 1. 3 Estructura Secundaria a(alfa)-hélice**

Fuente: Enciclopedia Microsoft® ENCARTA®

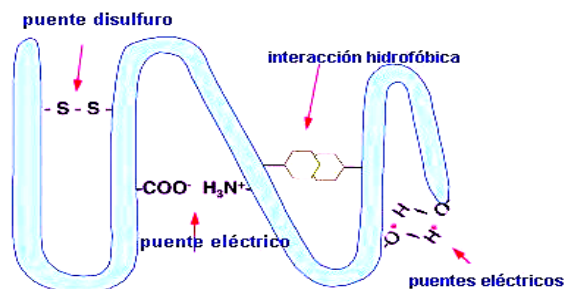
La conformación beta: En esta disposición no forman una hélice sino una cadena en forma de zigzag, denominada disposición en lámina plegada.



**Figura 1.4 Estructura Secundaria conformación beta**

Fuente: Enciclopedia Microsoft® ENCARTA®

- Estructura Terciaria. La estructura terciaria informa sobre la disposición de la estructura secundaria de un polipéptido al plegarse sobre sí misma originando una conformación globular. Esta conformación globular facilita la solubilidad en agua y así realizar funciones de transporte, enzimáticas, hormonales, etc.



**Figura 1.5 Estructura Terciaria**

Fuente: Enciclopedia Microsoft® ENCARTA®

- Estructura Cuaternaria. Esta estructura informa de la unión, mediante enlaces débiles (no covalentes) de varias cadenas polipeptídicas con estructura terciaria, para formar un complejo proteico.



**Figura 1.6 Estructura Cuaternaria**  
Fuente: Enciclopedia Microsoft® ENCARTA®

Determinados factores mecánicos (agitación), físicos (aumento de temperatura) o químicos (presencia en el medio de alcohol, acetona, urea, detergentes o valores extremos de pH) provocan la desnaturalización de la proteína, es decir, la pérdida de su estructura tridimensional; las proteínas se despliegan y pierden su actividad biológica.

1.1.2 Propiedades de las Proteínas. Las dos propiedades principales son:

- Especificidad. La especificidad se refiere a su función; cada una lleva a cabo una determinada función y lo realiza porque posee una determinada estructura primaria y una conformación espacial propia; por lo que un cambio en la estructura de la proteína puede significar una pérdida de la función.

Además, no todas las proteínas son iguales en todos los organismos, cada individuo posee proteínas específicas suyas que se ponen de manifiesto en los procesos de rechazo de órganos transplantados. La semejanza entre proteínas es un grado de parentesco entre individuos, por lo que sirve para la construcción de "árboles filogenéticos".

- **Desnaturalización.** Consiste en la pérdida de la estructura terciaria, por romperse los puentes que forman dicha estructura. Todas las proteínas desnaturalizadas tienen la misma conformación, muy abierta y con una interacción máxima con el disolvente, por lo que una proteína soluble en agua cuando se desnaturaliza se hace insoluble en agua y precipita.

La desnaturalización se puede producir por cambios de temperatura, (huevo cocido o frito), variaciones del pH. En algunos casos, si las condiciones se restablecen, una proteína desnaturalizada puede volver a su anterior plegamiento o conformación, proceso que se denomina renaturalización.

#### 1.1.3 Clasificación de las proteínas.

- **Holoproteínas.** Son aquellas formadas solamente por aminoácidos.

**Tabla 1.1 Holoproteínas**

<b>Globulares</b>	<p><b>Prolaminas:</b> Zeína (maíz), Gliadina (trigo), Hordeína (cebada)</p> <p><b>Gluteninas:</b> Glutenina (trigo), Orizanina (arroz)</p> <p><b>Albúminas:</b> Seroalbúmina (sangre), Ovoalbúmina (huevo), Lactoalbúmina</p> <p><b>Hormonas:</b> Insulina, hormona del crecimiento, prolactina, tirotropina</p> <p><b>Enzimas:</b> Hidrolasas, Oxidasas, Ligasas, Liasas, Transferasas, etc.</p>
<b>Fibrosas</b>	<p><b>Colágenos:</b> en tejidos conjuntivos, cartilaginosos</p> <p><b>Queratinas:</b> En formaciones epidérmicas: pelos, uñas, plumas, cuernos.</p> <p><b>Elastinas:</b> En tendones y vasos sanguíneos</p> <p><b>Fibroínas:</b> En hilos de seda, (arañas, insectos)</p>



- Heteroproteínas. Son aquellas formadas por una fracción proteínica y por un grupo no proteínico.

**Tabla 1.2 Heteroproteínas**

Glucoproteínas	Ribonucleasa Mucoproteínas Anticuerpos Hormona luteinizante
Lipoproteínas	De alta, baja y muy baja densidad, que transportan lípidos en la sangre.
Nucleoproteínas	Nucleosomas de la cromatina Ribosomas
Cromoproteínas	Hemoglobina, hemocianina, mioglobina, que transportan oxígeno Citocromos, que transportan electrones

## 1.2 PROTEÓMICA

La Proteómica es un área de la Biología cuyo objetivo es el estudio de los proteomas, el cual puede describirse como el conjunto de proteínas expresadas por un genoma<sup>2</sup>.

El término “proteoma” fue usado por vez primera en 1995 para describir el conjunto de proteínas de un genoma, una célula o un tejido. La palabra proteoma dio lugar a una nueva disciplina, la “proteómica”. Esencialmente la proteómica es el estudio a gran escala de los productos génicos de un genoma mediante métodos bioquímicos, con el fin de obtener una visión global e integrada de los procesos celulares.

---

<sup>2</sup> Un genoma es el conjunto de los genes que caracterizan a una especie

De la proteómica se puede decir que comenzó en los años setenta cuando se empezaron a construir bases de datos de proteínas utilizando la electroforesis bidimensional<sup>3</sup>, hasta el uso de la espectrometría de masas<sup>4</sup>. Este desarrollo, junto con la disponibilidad de los genomas secuenciados marca el comienzo de una nueva era.

Hubo dos factores decisivos para el desarrollo de la proteómica:

- 1) Por un lado la secuenciación de los genomas a gran escala (se conoce la secuencia de los genes pero no su función).
- 2) Por otro lado, el desarrollo de técnicas de separación y análisis de proteínas (Electroforesis Bidimensional y Espectrometría de Masas).

Se puede hablar de dos tipos de proteómica: proteómica de expresión y proteómica del mapa celular.

**La proteómica de expresión:** es el estudio cuantitativo de la expresión de proteínas entre muestras que difieren en alguna variable. En esta estrategia se compara la expresión del proteoma total o de subproteomas entre diferentes muestras. La información obtenida puede permitir la identificación de nuevas proteínas implicadas en la identificación de proteínas específicas de una enfermedad y proteínas de interés en microbiología médica.

**La proteómica del mapa celular o estructural:** es el estudio de la localización subcelular de las proteínas y de las interacciones proteína-proteína mediante la purificación de orgánulos o complejos y la posterior identificación de sus componentes mediante espectrometría de masas.

---

<sup>3</sup> La electroforesis bidimensional es una técnica de alta resolución cuyo objetivo es la separación de mezclas de proteínas altamente complejas.

<sup>4</sup> Un espectrómetro de masas es un equipo de análisis muy sofisticado que permite obtener la masa molecular de casi cualquier compuesto, por complicado que sea.

También se utiliza el término de proteómica funcional para referirse a diversas aproximaciones proteómicas que permiten el estudio y caracterización de un grupo de proteínas determinado proporcionando información importante sobre señalización, mecanismos de la enfermedad o interacciones proteína-fármaco.

La proteómica proporciona un conjunto de herramientas muy poderosas para el estudio a gran escala de la función de los genes a nivel de proteína. La aplicación de la proteómica tiene un enorme potencial en el área de la biomedicina para el desarrollo de fármacos (anticancerígenos, para el sistema nervioso, aparato cardiovascular, antimicrobianos, etc.), métodos de diagnóstico, desarrollo de vacunas, etc.

## 2. ALINEAMIENTO DE SECUENCIAS

En primer lugar una secuencia se puede definir como la serie de elementos encadenados uno detrás de otros, debido a eso se habla de secuencias de nucleótidos y de secuencias de aminoácidos. A través de letras podemos identificar los distintos aminoácidos que conforman la macromolécula (ejemplo: A: alanina; T: treonina; C: cisteína; y G: glicina; D: aspártico; E: glutámico; etcétera).

El alineamiento de secuencias biológicas consiste en establecer un segmento entre ellas donde el número de coincidencias (una coincidencia se presenta cuando el nucleótido de la secuencia A sea igual al nucleótido en la secuencia B) sea máximo. El alineamiento de secuencias es una forma de hacer arqueología, de descubrir qué partes de las secuencias son más importantes (están más conservadas), descubrir qué proteínas tienen un origen común (existen modelos estadísticos que ayudan a distinguir parecidos al azar de parecidos que reflejan un mismo origen evolutivo).

El alineamiento de secuencias también puede servir para predecir la estructura de las proteínas (las proteínas homólogas tienen una misma arquitectura tridimensional), o también puede ayudar a predecir la función de las proteínas (aunque en este aspecto hay que ser cautelosos ya que a lo largo de la evolución proteínas con un origen común pueden terminar desarrollando distintas funciones).

La descripción del resultado del alineamiento trae aparejado el uso (o mal uso) de los términos: **identidad**, **similitud** y **homología**. La **identidad** significa que existe exactamente el mismo nucleótido o aminoácido en la misma posición de las secuencias alineadas. **Similitud** expresa una medida observable que considera por un lado las identidades y por otro lado, le da valor a las sustituciones favoreciendo aquellas que sean conservativas respecto de las que no lo son. Finalmente, **homología** significa que las secuencias no sólo se parecen mucho entre sí, sino que también comparten una historia evolutiva.

**Tabla 2.1 Identificación de Aminoácidos**

<b>Símbolo</b>	<b>Significado</b>	<b>Símbolo</b>	<b>Significado</b>
A	Alanina	P	Prolina
B	Asparagina	Q	Glutamina
C	Cisteína	R	Arginina
D	Ácido Aspártico	S	Serina
E	Ácido Glutámico	T	Treonina
F	Fenilalanina	U	Selenocysteína
G	Glicina	V	Valina
H	Histidina	W	Triptófano
I	Isoleucina	Y	Tirosina
K	Lisina	Z	Ácido Glutámico
L	Leucina	M	Metionina
N	Asparagina	*	Fin de la traducción
-	gap de longitud indeterminada		

El alineamiento se puede clasificar por:

1. Número de secuencias analizadas:

- Alineamiento de un par de secuencias: Este método recibe dos secuencias y encuentra el segmento mejor alineado entre ellas.
- Alineamiento múltiple: Este método trabaja sobre muchas secuencias y el resultado que obtiene es una secuencia consenso, que tiene en cada posición el nucleótido o el aminoácido (en caso de las proteínas), que más se ha conservado en esa posición en todas las secuencias estudiadas.

2. Nivel de análisis:

- Alineamiento global: Obtiene el mejor alineamiento de dos secuencias.
- Alineamiento local: Encuentra el mejor segmento alineado existente entre dos secuencias.

## **2.1 SIMILITUD**

Es el resultado del análisis (observación cuantitativa) de la estructura primaria de dos o más secuencias; las secuencias pueden ser ácidos nucleicos o proteínas.

Puesto que la similitud es obtenida de observar las secuencias no puede ser tomada como un indicador para establecer la relación biológica (descendencia) entre las secuencias, ya que el grado de similitud puede deberse a cambios aleatorios acumulados en las secuencias a través del tiempo.

## **2.2 HOMOLOGÍA**

La homología es una medida cualitativa entre las secuencias, se presenta cuando la similitud que estas tienen es atribuible a razones evolutivas y no al azar, es decir, la homología establece regiones entre las secuencias que se han conservado con el tiempo.

La similitud es el resultado de una medida cuantitativa, la homología es una hipótesis postulada por el investigador basándose en la similitud de las secuencias y en otros datos biológicos que previamente conozca sobre el origen de dichas secuencias. Es permitido establecer el porcentaje de similitud de dos o más secuencias, pero esto no es posible para la homología, ya que las secuencias son o no son homólogas.

## **2.3 MATRICES DE SUSTITUCIÓN**

Una matriz de sustitución es una tabla de valores que describe la probabilidad de que un residuo de la secuencia  $m$  en la posición  $i$  tenga ocurrencia en la secuencia  $n$  en la posición  $j$ . Existen por lo tanto matrices de sustitución para ADN y proteínas, sin embargo, las más utilizadas son las matrices para sustitución de aminoácidos (proteínas) ya que la gran mayoría de análisis se realizan con este tipo de secuencias, incluso cuando lo que se requiere es encontrar un nivel de similaridad entre dos secuencias de ADN.

Características:

1. Una matriz de sustitución se elabora bajo una teoría de evolución.

2. El resultado de la comparación de dos o más secuencias depende fuertemente de la matriz de sustitución que se haya seleccionado.
3. Las matrices de sustitución son utilizadas en los análisis comparativos de secuencias.
4. Los algoritmos de alineamiento (comparación) funcionan igual con una matriz de distancias o con una matriz de sustitución (aunque se pueden obtener diferentes resultados).
5. Una matriz de distancias es muy útil en la reconstrucción de un árbol filogenético, mientras que una matriz de sustitución es utilizada para realizar búsqueda en bases de datos.

2.3.1 Matrices de sustitución para Proteínas. La matriz de sustitución más sencilla es aquella que evalúa de manera binaria si un residuo en la posición  $m_i$  es o no idéntico al residuo en la posición  $n_j$ :

**Tabla 2.2 Matriz de Sustitución sencilla**

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

Otra manera de evaluar la similitud es mediante una matriz que evalúe las ocurrencias dado que se cambie una purina por una pirimidina o viceversa (transiciones/transversiones):

**Tabla 2.3 Matriz de Sustitución de Ocurrencias**

	A	C	G	T
A	0	5	5	1
C	5	0	1	5
G	5	1	0	5
T	1	5	5	0

Este tipo de matrices son utilizadas para la evaluación de alineamientos de residuos nucleotídicos. Las matrices de sustitución para proteínas son un poco más complejas, y los valores en una matriz de este tipo son logaritmos del ratio de dos probabilidades:

1. La probabilidad de una ocurrencia aleatoria de un aminoácido en el alineamiento (este valor es el producto de las frecuencias de ocurrencia independientes de cada aminoácido).
2. La probabilidad de una ocurrencia significativa de un par de residuos en un alineamiento. Estas probabilidades se derivan de alineamientos ya conocidos y de los cuales se sabe que son significativos.

2.3.2 Pam. Las matrices PAM (Percent Accepted Mutation) fueron descritas por Daykoff (1978). Están basadas en el alineamiento global de secuencias de proteínas estrechamente relacionadas y asumen que una modificación en algún sitio depende solamente del aminoácido presente en ese sitio. La PAM1 por ejemplo es la matriz calculada a partir de comparaciones de secuencias con no más del 1% de diferencias.

Este tipo de matriz esta indicada especialmente para alineamientos de secuencias con ancestros homólogos. Cuanto mayor es el número que acompaña al nombre de la matriz (PAM40, PAM120), mayor se espera que sea la distancia evolutiva. Si se desconoce la distancia evolutiva es necesario ejecutar al menos tres búsquedas utilizando las matrices PAM40, PAM120 y PAM 250. Generalmente se utilizan para establecer la evolución de una secuencia o para identificar secuencias conservadas.

Procedimiento para construir una matriz PAM:

1. Alinear un conjunto de secuencias que tengan una identidad no inferior al 85%.



2. Reconstruir un árbol filogenético<sup>5</sup> de las secuencias alineadas con el fin de inferir ancestros.
3. Calcular  $A_{ij}$ :  $A_{ij}$  representa el número de veces en las cuales el aminoácido  $j$  fue reemplazado por el aminoácido  $i$  en todas las comparaciones (las secuencias se comparan por pares y los cambios encontrados se presumen que se han presentado por selección natural).
4. Calcular la mutabilidad del aminoácido  $j$  ( $m_j$ ): Es la propensión que dado un aminoácido  $j$  sea reemplazado por cualquier otro aminoácido.
5. Combinar los resultados obtenidos de los numerales 3 y 4 para generar una matriz de probabilidad de mutación para un PAM de distancias evolutivas, la matriz se calcula con la siguiente formula:

$$M_{ij} = \frac{m_j \times A_{ij}}{1 - m_j}$$

$$M_{jj} = 1 - m_j$$

### **Figura 2.1 Ecuaciones para la construcción de la Matriz PAM**

Algunas propiedades de una matriz de probabilidad de mutación:

- La probabilidad que un aminoácido sea sustituido es del orden del 1%.
- La suma de  $m_j$  para toda  $j$  es igual a un (1).
- La matriz M1 establece una unidad de cambio evolutivo. (La PAM 1 acepta una mutación cada 100 aminoácidos).
- Aplicaciones sucesivas de una matriz M1 a una secuencia produce matrices M2, M3, ..., Mn.
- Los elementos de la matriz PAM 0 son 1 para  $M_{ii}$  y 0 para  $M_{ij}$ .

2.3.3 Blosum. La matriz "BLOcks SUBstitution" fue propuesta por Steven Henikoff and Jorja G. Henikoff en el año de 1992, fue creada a partir de un estudio sobre

---

<sup>5</sup> Un árbol filogenético agrupa las proteínas que comparten caracteres derivados de un antecesor común pero que no necesariamente incluye a todos los descendientes.

bloques conservados. Este tipo de matriz está indicada cuando se trata de identificar una secuencia desconocida.

Procedimiento para la construcción de una matriz Blosun:

- I. Se inicia con segmentos (bloques) conservados de secuencias.
- II. Alinear las secuencias sin permitir la presencia de huecos.
- III. Establecer el número de aminoácidos alineados por pares ( $f_{ij}$ ).
- IV. La frecuencia observada de cada par de aminoácidos ( $q_{ij}$ ) es el cociente entre el valor de  $f_{ij}$  y el número total de pares aminoácidos (esto incluye los par  $i=j$ , es decir los casos en los que no se presenta sustitución).

La frecuencia esperada de un par de aminoácidos es el producto de las frecuencias de cada aminoácido en el conjunto de datos.

Las secuencias de un bloque son agrupadas según unos umbrales de similaridad, por ejemplo si se utiliza un 80% de similaridad se construye una matriz Blosun80.

## **2.4 HERRAMIENTAS DE BÚSQUEDA DE SIMILITUD**

Muchos programas implementan diferentes estrategias de búsqueda que permiten encontrar similitudes entre la secuencia query (secuencia a consultar) y las que se encuentran reportadas en la base de datos. El principio básico de cada uno de estos algoritmos es el mismo: la secuencia de interés es comparada con las secuencias en la base de datos para establecer una lista de aquellas con las que se encuentra mayor similaridad.

El método más tradicional para seleccionar un resultado en particular, entre todos los posibles resultados obtenidos tras el alineamiento, puede ser la escogencia de aquel que posea el puntaje o *score* más alto. Existen diferentes formas para

determinar el *score*, pero la más común y simple de ellas, consiste en la suma de los puntajes individuales determinados para cada pareja de residuos comparados.

En estas comparaciones se asignan valores positivos para uniones correctas y negativos para aquellas que no lo son. El *score* asignado a cada pareja depende del tipo de secuencia que se está comparando, por ejemplo, para secuencias de aminoácidos se han construido matrices de sustitución en las cuales se ha asignado un valor diferente para las posibles parejas de aminoácidos comparados.

En el *score* total también se considera la aparición de un carácter adicional, denominado gap (hueco), el cual se incluye en las secuencias comparadas para maximizar su similitud. En un contexto biológico, la inclusión de un gap en una de las secuencias, podría ser equivalente por ejemplo, a la aparición de una inserción en la otra. La inclusión de este nuevo carácter tiene dos componentes. Por una parte se penaliza la aparición como tal del mismo, e independientemente se penaliza la extensión de un gap ya abierto a más de un residuo. No existe una teoría aceptada para la elección de los valores de penalización, y su escogencia depende principalmente de pruebas de ensayo y error.

El siguiente paso consiste en encontrar el alineamiento óptimo de la secuencia. Algoritmos como Smith-Waterman para alineamientos locales llevan a cabo esta tarea, siendo el tiempo consumido en ella proporcional al producto de la longitud de las secuencias comparadas.

Los programas de búsqueda en bases de datos difieren en el núcleo del algoritmo que usan. Los algoritmos de alta velocidad usan principios simplificados para establecer la similitud de las secuencias. Es importante considerar al escoger uno de los posibles programas para búsqueda de similitud entre secuencias, que el tiempo que ésta tarda en llevarse a cabo depende de la sensibilidad del algoritmo, estando fuertemente influenciado por la longitud de la secuencia y el tamaño de la

base de datos. A continuación, se resumen algunos de los programas utilizados para búsqueda de secuencias en bases de datos.

**Tabla 2.4 Comparación de los programas utilizados en la búsqueda de secuencia en las bases de datos**

<b>Programa</b>	<b>Sensibilidad</b>	<b>Velocidad</b>	<b>Tipo de secuencia</b>
BLAST	Medianamente sensitivo	Muy rápido	DNA, Proteína
FASTA	Sensitivo	Rápido	DNA, Proteína
Blitz	Muy sensitivo	Medianamente rápido	DNA, Proteína
SSEARCH	Muy sensitivo	Lento	DNA, Proteína
PSI-BLAST	Extremadamente sensitivo	Lento	Proteína

FASTA y BLAST fueron desarrollados como algoritmos de alta velocidad y baja sensibilidad, en comparación con Smith-Waterman, ya que se basan en estrategias heurísticas que concentran sus esfuerzos en las regiones de la secuencia más probablemente relacionadas. Procedimientos rápidos de unión-exacta identifican inicialmente las regiones promisorias, y solo hasta este momento se acude al algoritmo de Smith-Waterman, lo que permite que estos programas sean 10 a 100 veces más rápidos. Es posible ajustar algunos parámetros cuando se está utilizando FASTA y BLAST, los cuales hacen referencia al procedimiento heurístico y que al ser modificados, permiten establecer una relación velocidad/sensibilidad adecuada.

Una vez obtenido el alineamiento de dos secuencias surge un concepto importante que hace referencia a la relevancia biológica del resultado obtenido. El establecimiento de la similitud de dos secuencias con relación a su origen evolutivo, es decir, con la derivación a partir de un ancestro común (homología), puede ser una de las inquietudes de mayor relevancia biológica en el análisis de secuencias. Extraer esta información simplemente de un valor de similaridad

obtenido resulta difícil, por lo cual se han establecido algunos parámetros estadísticos que permiten estimar la relevancia del resultado.

2.4.1 Blast. Es un conjunto de programas de búsqueda de similitud diseñados para explorar todas las bases de datos de secuencias. BLAST es el acrónimo de *Basic Local Alignment Search Tool*. Fue desarrollado por Altschul en 1990. La principal característica del BLAST es su velocidad, pudiendo tomar pocos minutos cualquier búsqueda en la totalidad de la base de datos. De hecho, los resultados se presentan en pantalla inmediatamente después de calculados. El BLAST puede hacer búsquedas en una base de datos no redundante (nr) la cual tiene los registros no redundantes entre las dos bases de datos principales a nivel mundial: GenBank en USA y EMBL (European Molecular Biology Laboratories) en Europa.

2.4.2 Fasta. Fue el primer algoritmo ampliamente utilizado para búsqueda de similitud en una base de datos. FASTA busca alineamientos locales óptimos buscando coincidencias de pequeñas subsecuencias denominadas palabras ("words o k-tuplas"), el score del primer segmento en el que se aparean varias palabras se denomina "init1" y la suma de todos los score de los segmentos se denomina "initn". La sensibilidad y velocidad del algoritmo es inversamente proporcional a la longitud de la palabra utilizada en la búsqueda.

Desarrollado por David Lipman y William Pearson en el año de 1985. es empleado principalmente por el EMBL - EBI (European Molecular Biology Laboratories - European Bioinformatics Institute), si se compara su velocidad con BLAST se notará que es mucho más lento, incluso llega a emplear varias horas para obtener los resultados, es por esta razón que el EMBL envía los cálculos al usuario por correo electrónico.

FASTA compara una secuencia de DNA o de proteínas contra todas las secuencias de una base de datos y devuelve los mejores segmentos alineados.

### 3. ALGORITMOS DE ALINEAMIENTO DE SECUENCIAS

Los algoritmos de comparación de secuencias miden similitud o identidad entre secuencias, pero no miden homología. La forma que estos algoritmos emplean para darle significado a cada uno de los alineamientos posibles, es asignándole un valor (*score*) a cada uno de ellos: el *score* más alto corresponde al mejor alineamiento.

La manera más común de asignar este valor es, a través de una suma simple de *scores* especificados para cada alineamiento de pares de letras (que representan igualdades o sustituciones), y de letras con caracteres nulos (que representan delecciones o inserciones). El conjunto de estos *scores* mencionados representa una matriz de *scores*, las más populares son las matrices de bloques de sustitución (BLOSUM) y las matrices de mutaciones puntuales aceptables (PAM).

#### 3.1 ALGORITMO DE SMITH Y WATERMAN.

El algoritmo de Smith y Waterman (1981) encuentra el segmento mejor alineado entre un par de secuencias.

Características:

- A las no coincidencias se le asigna un puntaje (peso) negativo.
- El mínimo puntaje guardado en la matriz debe ser cero.
- El inicio y final de un camino óptimo puede encontrarse en cualquier sitio de la matriz, no sólo en la última fila o columna.
- El puntaje se incrementa en una región de alta similaridad y disminuye fuera de tales regiones.
- Si hay dos segmentos de alta similaridad, entonces deben estar lo suficientemente cercanos para que puedan encadenarse por un hueco o quedarán como segmentos independientes de similaridad local. La notación de un hueco se hace por medio de un guión bajo (\_).

- Cada segmento de similaridad debe iniciar con un puntaje de cero.
- Se debe buscar en toda la matriz las regiones de alta similaridad local.

Descripción:

Dadas dos secuencias A y B,  $A=a_1a_2...a_n$  y  $B=b_1b_2...b_m$ . Se define:

- Una función de similitud (coincidencias o Matriz de Sustitución),  $S(a_i,b_j)$ , entre los elementos  $a_i$  y  $b_j$  de las secuencias a alinear, como las secuencias son proteínas se puede utilizar una matriz de sustitución como la PAM250 .
- Los in/dels (inserciones o deleciones) de longitud  $k$  se penalizan con un peso  $W_k$ . Estos valores hacen referencia a los valores de los gaps (huecos), que son valores aleatorios.
- Se construye una matriz H de  $n+1$  filas y  $m+1$  columnas. La secuencia B se ubica en las filas y la secuencia A en las columnas.

Procedimiento:

1. Se inicializa la matriz H o Matriz de Resultado con ceros.
2. La posición  $H_{ij}$  es la máxima similitud de dos segmentos que terminan en  $a_i$  y  $b_j$  respectivamente. Se obtiene de la siguiente expresión:

$$H(i, j) = \max \begin{cases} 0 \\ H(i-1, j-1) + S(a_i + b_j) \\ H(i-1, j) + W_0 \\ H(i, j-1) + W_e \end{cases}$$

**Figura 3.1 Cálculo del valor de  $H_{ij}$  (Matriz de Resultado)**

3. El segmento mejor alineado se obtiene de la matriz H ubicando en esta matriz, la posición con el valor más alto.
4. Por medio de un procedimiento traceback, se recuperan los residuos que conforman el segmento.
5. El traceback se detiene cuando encuentra un cero en la diagonal de la matriz.

Por ejemplo, si se tienen las siguientes secuencias:

Secuencia A: CAGCCNCGCENNAG,  $m=13$

Secuencia B: AANGCCANNGACGG,  $n=14$

Después de realizar los cálculos para llenar las posiciones, se llega a la siguiente Matriz de Resultado:

**Tabla 3.1 Matriz H (Matriz de Resultado)**

.	<i>j</i>	1	2	3	4	5	6	7	8	9	10	11	12	13
<i>i</i>	.	C	A	G	C	C	N	C	G	C	N	N	A	G
1	A	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2	A	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
3	N	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
4	G	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
5	C	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
6	C	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0
7	A	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0
8	N	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0	1.0
9	N	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0
10	G	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.3	2.0	1.7	1.3	2.3	2.7
11	A	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0
12	C	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0
13	G	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0
14	G	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0

El siguiente paso es encontrar el máximo score de la matriz H (Tabla 1). El máximo valor es 3.3 y se encuentra en la posición (10,8). A partir de esta posición se empieza a recorrer la matriz hacia atrás, guardando el residuo del máximo valor que se encuentra entre las posiciones arriba, abajo o derecha. Los residuos se guardan teniendo en cuenta los recorridos en las dos secuencias.



**Tabla 3.2 Máximo valor en Matriz de Resultado y Residuo**

i	j	1	2	3	4	5	6	7	8	9	10	11	12	13	G G
i		C	A	G	C	C	N	C	G	C	N	N	A	G	
1	A	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
2	A	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7	
3	N	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7	
4	G	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0	
5	C	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3	
6	C	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0	
7	A	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0	
8	N	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0	1.0	
9	N	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0	
10	G	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.3	2.0	1.7	1.3	2.3	2.7	
11	A	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0	
12	C	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0	
13	G	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0	
14	G	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0	

**Tabla 3.3 Inicio de Traceback**

i	j	1	2	3	4	5	6	7	8	9	10	11	12	13	NG CG
i		C	A	G	C	C	N	C	G	C	N	N	A	G	
1	A	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
2	A	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7	
3	N	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7	
4	G	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0	
5	C	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3	
6	C	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0	
7	A	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0	
8	N	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0	1.0	
9	N	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0	
10	G	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.3	2.0	1.7	1.3	2.3	2.7	
11	A	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0	
12	C	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0	
13	G	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0	
14	G	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0	

El procedimiento finaliza encuentra un cero en la diagonal (posición (3,2)).

**Tabla 3.4 Recorrido completo en Matriz de Resultado**

.	j	1	2	3	4	5	6	7	8	9	10	11	12	13
i	.	C	A	G	C	C	N	C	G	C	N	N	A	G
1	A	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2	A	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.7
3	N	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.7
4	G	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0	0.0	0.0	0.7	0.7	1.0
5	C	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3	2.0	0.7	0.3	0.3	0.3
6	C	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0	1.3	1.7	0.3	0.0	0.0
7	A	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0	0.7	1.0	1.3	1.3	0.0
8	N	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0	0.7	1.7	2.0	1.0	1.0
9	N	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0	0.7	1.7	2.7	1.7	1.0
10	G	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.3	2.0	1.7	1.3	2.3	2.7
11	A	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0	3.0	1.7	1.3	2.3	2.0
12	C	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7	3.0	2.7	1.3	1.0	2.0
13	G	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7	1.7	2.7	2.3	1.0	2.0
14	G	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3	2.3	1.3	2.3	2.0	2.0

El segmento alineado es:

**GCCANNG**

**GCC\_NCG**

## 4. ANÁLISIS DEL ALINEADOR

El análisis, y el diseño posterior del ALINEADOR, está basado en la metodología propuesta por Craig Larman, a través del uso del lenguaje UML (Uniform Modelling Language), que especifica un método uniforme para describir y especificar las distintas etapas de análisis y diseño de un sistema de software.

Las etapas del proceso comprendidas en este capítulo son:

- Detalle de funciones.
- Casos de uso.

### 4.1 FUNCIONES DEL SISTEMA

**Tabla 4.1 Funciones Básicas del ALINEADOR**

Ref.	Función	Categoría
R.1.1	Validación de los valores de las secuencias y los valores de los gaps.	Evidente
R.1.2	Carga de la Matriz de Sustitución.	Oculto
R.1.3	Creación de la Matriz de Resultado y búsqueda del valor mayor.	Oculto
R.1.4	Procedimiento Traceback.	Oculto
R.1.5	Visualización de la Salida.	Evidente

## 4.2 DIAGRAMA DE CASOS DE USO DEL ALINEADOR

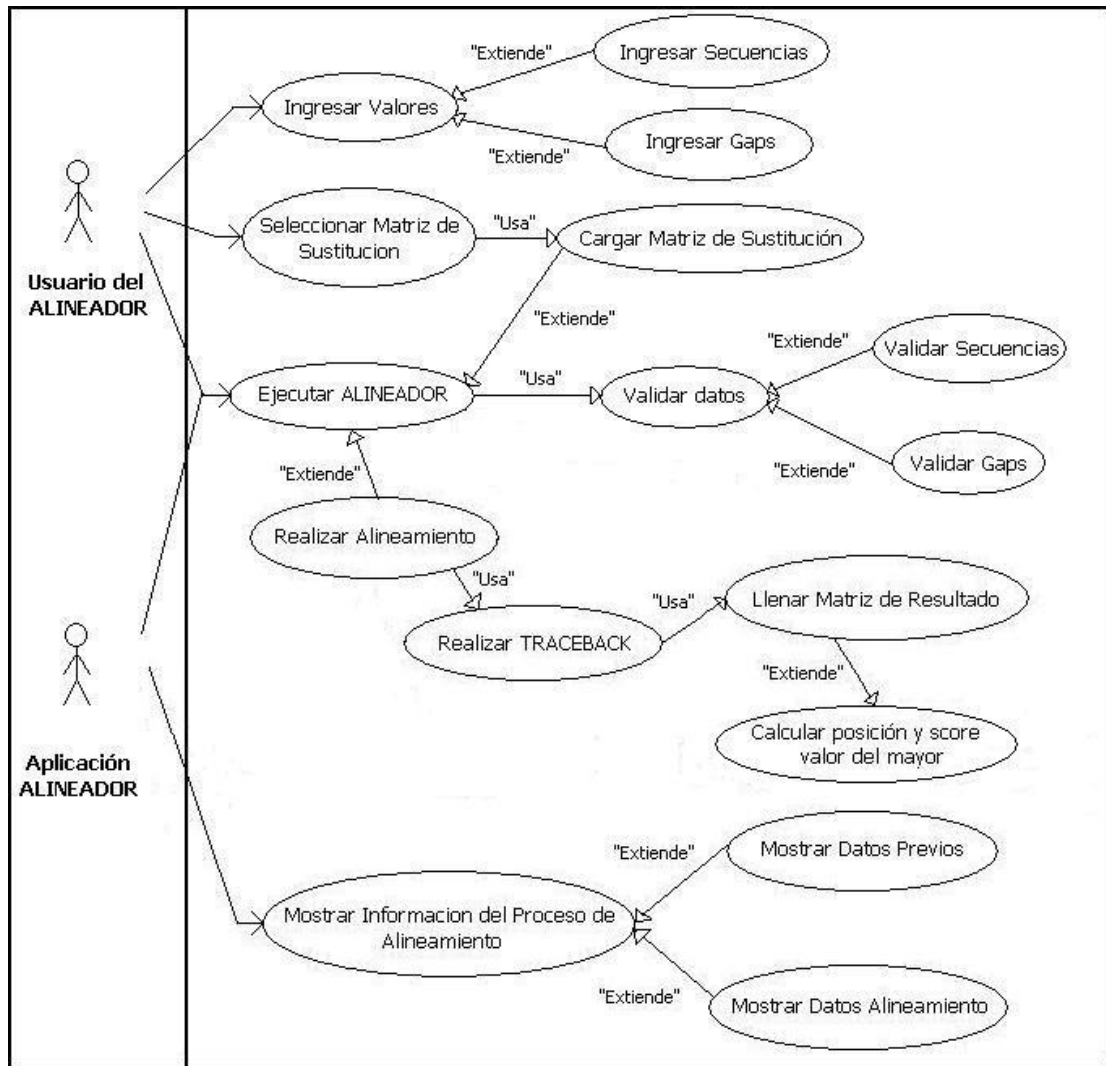


Figura 4.1 Diagrama de Casos de Uso del ALINEADOR

#### 4.2.1 Descripción de Casos de Usos del ALINEADOR

**Tabla 4.2 Descripción Caso de Uso Ingresar Valores**

CASO DE USO: Ingresar valores			Activación: Usuario del ALINEADOR			
PROPÓSITO: Permitir que el usuario ingrese los valores correspondientes a las secuencias y los valores de los Gaps.						
n	Flujo Principal de Eventos			Variaciones		Excepciones
	ACTOR	SISTEMA				
1	Ingresa las secuencias, teniendo en cuenta que cada letra de las cadenas corresponde a un aminoácido.	El Software captura los valores digitados por el usuario o cargados desde archivo. Realiza la validación teniendo en cuenta que no existan caracteres no válidos en las secuencias.				El software bloquea las teclas que corresponden a caracteres especiales, números y caracteres que no correspondan a aminoácidos.
2	Ingresa los valores de los Gaps: Open y Extended.	El software captura los datos y realiza la validación teniendo en cuenta que estos deben ser valores de tipo numérico.				El software bloquea las teclas que no correspondan a números, como los caracteres especiales y letras.

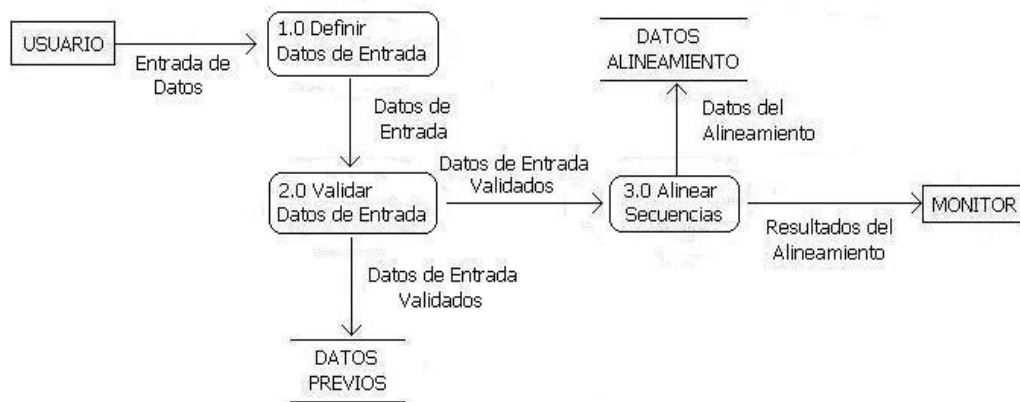
**Tabla 4.3 Descripción Caso de Uso Ejecutar ALINEADOR**

CASO DE USO: Ejecutar ALINEADOR				Activación: Usuario del ALINEADOR		
PROPÓSITO: Iniciar el proceso de Alineamiento de Secuencias						
n	Flujo Principal de Eventos			Variaciones		Excepciones
	ACTOR	SISTEMA				
1	Inicia el proceso.	Valida los valores o Datos de Entrada	a	Validar Secuencias		Mensaje de Error informando, si las cadenas están vacías, o si existen caracteres no validos en las cadenas.
			a	Validar Gaps		Mensaje de Error informando si los valores de los gaps, no tienen formato numérico o si no existen datos.
2		Cargar Matriz de Sustitución				
3		Realizar Alineamiento	a	Realizar TRACEBACK		

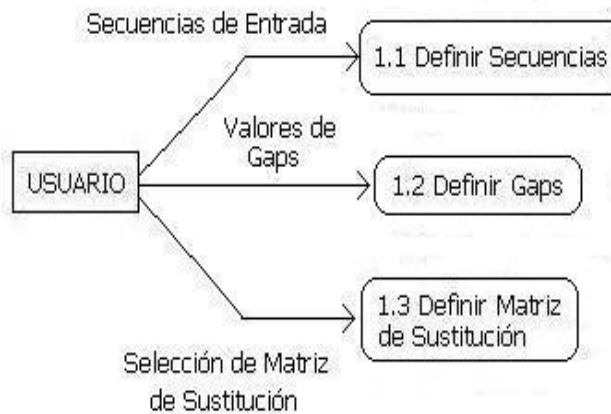
### 4.3 DIAGRAMA DE FLUJO DE DATOS PARA EL ALINEADOR



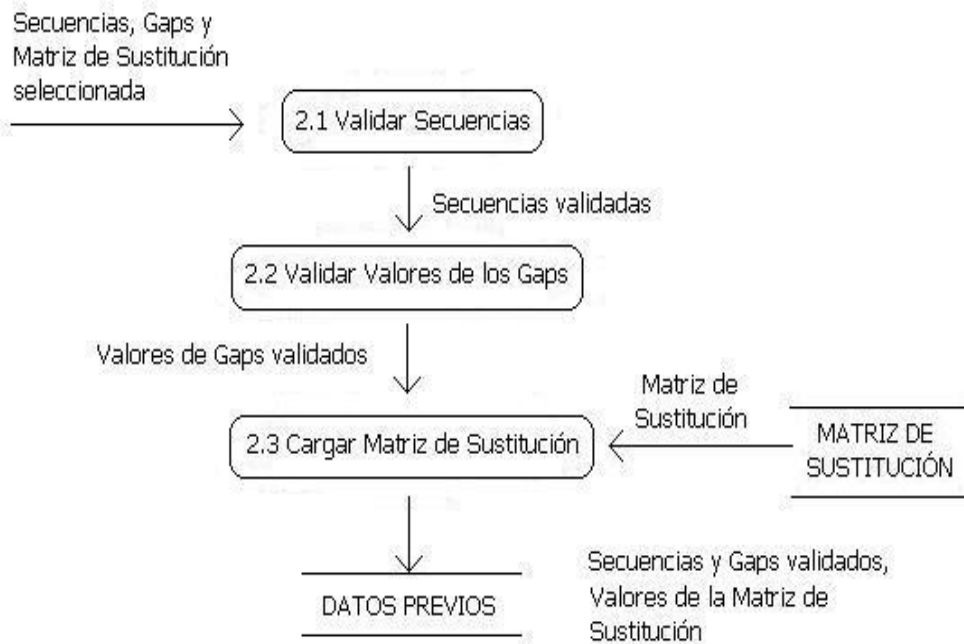
**Figura 4.2 Flujo de Datos General para el ALINEADOR**



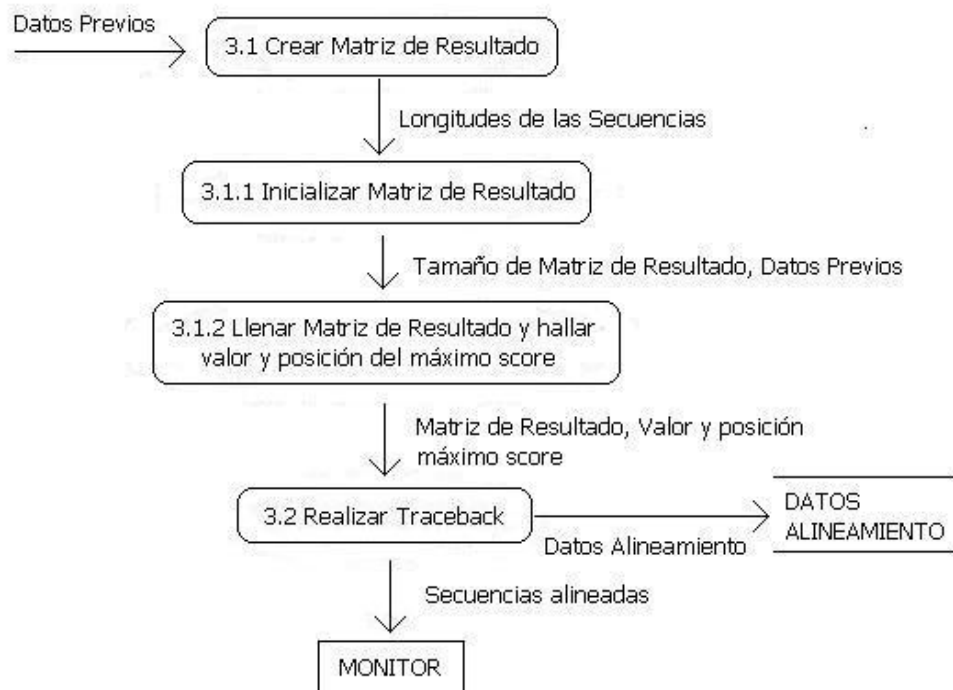
**Figura 4.3 Flujo de Datos Conceptual**



**Figura 4.4 Flujo de Datos Nivel 1**



**Figura 4.5 Flujo de Datos nivel 2**



**Figura 4.6 Flujo de Datos nivel 3**



## 5. DISEÑO DEL ALINEADOR

### 5.1 MODELO DE CLASES

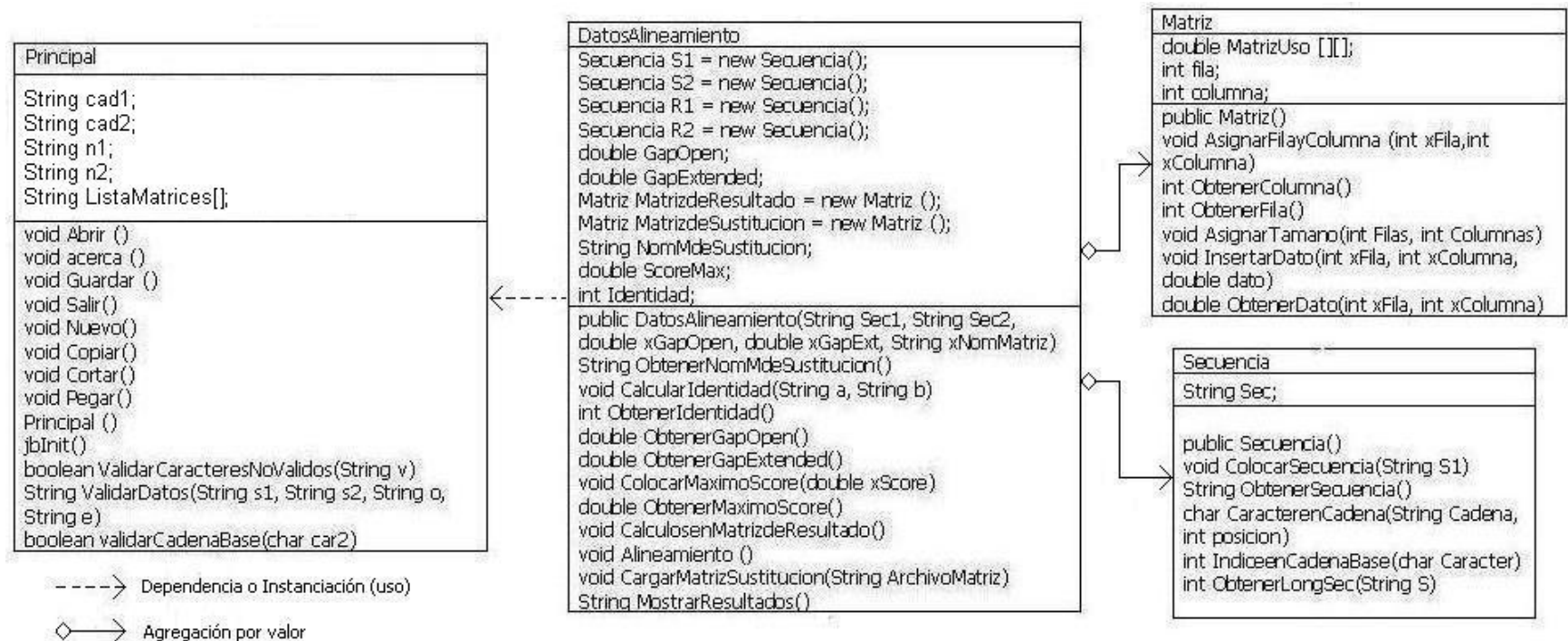


Figura 5.1 Diagrama de Clase del ALINEADOR

## 5.2 DESCRIPCIÓN DIAGRAMA DE CLASES

Tabla 5.1 Descripción de la Clase Principal

Nombre de la Clase: Principal	
<b>Descripción:</b> Clase que contiene los elementos de la Interfaz gráfica. Es la encargada de validar todos los datos que ingresa el usuario, antes de continuar con el proceso de alineamiento.	
Atributos	Descripción
cad1	Elementos de tipo String, que almacena el contenido del cuadro de texto de la secuencia 1.
cad2	Elementos de tipo String, que almacena el contenido del cuadro de texto de la secuencia 2.
n1	Elementos de tipo String, que almacena el contenido del cuadro de texto del Gap Open.
n2	Elementos de tipo String, que almacena el contenido del cuadro de texto del Gap Extended.
ListaMatrices[];	Es un vector tipo String, que carga los nombres de las Matrices de Sustitución desde el archivo.
Métodos	Descripción
Principal()	Constructor de la clase
Abrir()	Permite cargar al ALINEADOR, una cadena de caracteres, contenida en un archivo
Nuevo()	Reinicia la pantalla del ALINEADOR. Por defecto asigna 1.0 y 2.0 a los valores de los Gap Open y Gap Extended respectivamente, y borra las secuencias.
Salir()	Cierra el ALINEADOR.
Guardar()	Guarda la información del proceso de Alineamiento, incluyendo los datos previos y los resultados, en un archivo, permitiendo que el usuario escoja la ubicación de éste en el computador.

Cortar(), Copiar() y Pegar()	Métodos que permiten cortar, copiar y pegar las secuencias.
acerca()	Muestra información sobre los autores del ALINEADOR.
ValidarCaracteresNoValidos(String v)	Cada aminoácido está representado por una letra. Esta función recibe la cadena contenida en una caja de texto y valida que no existan en ésta, caracteres que no correspondan a aminoácidos. Retorna Verdadero, en el caso de que los encuentre, y Falso en el caso contrario.
ValidarDatos (String S1, String S2, double o, double e)	Método que realiza secuencialmente la validación de los datos ingresados por el usuario.
ValidarCadenaBase(char car2)	Valida que un carácter se encuentre dentro de la cadena base (ARNDCQEGHILKMFPSTWYVBZ) que corresponde a la forma como se encuentran creadas las matrices de sustitución. Ésta validación se hace en tiempo real a medida que el usuario digita una letra.

**Tabla 5.2 Descripción de la Clase DatosAlineamiento**

<b>Nombre de la Clase: DatosAlineamiento</b>	
<b>Descripción:</b> Clase principal del ALINEADOR, contiene todos los datos necesarios para la ejecución del programa.	
<b>Atributos</b>	<b>Descripción</b>
Secuencia S1 = new Secuencia (); Secuencia S2 = new Secuencia ();	Tipos de dato Secuencia, que permite almacenar las dos secuencias ingresadas por el usuario (S1 y S2).
Secuencia R1 = new Secuencia (); Secuencia R2 = new Secuencia ();	Tipos de dato Secuencia, que permite almacenar las dos secuencias generadas por el alineamiento.
GapOpen;	Variable de tipo double. Almacena el valor del Gap Open.

GapExtended;	Variable de tipo double. Almacena el valor del Gap Extended.
MatrizdeResultado = new Matriz ();	Es un atributo tipo Matriz, que se encarga de almacenar los valores que se crean en el proceso del alineamiento de las secuencias.
MatrizdeSustitucion = new Matriz ();	Es un atributo tipo Matriz, usado para cargar los valores de la Matriz de Sustitución seleccionada por el usuario.
NombreMatrizdeSustitucion;	Variable de tipo String. Almacena el nombre asignado a la Matriz de Sustitución.
ScoreMax;	Variable de tipo double. Almacena el máximo valor (score) obtenido en la Matriz de Resultado.
Identidad;	Variable de tipo int. Este atributo guarda el valor de la identidad (número de caracteres iguales), entre las secuencias.
<b>Métodos</b>	<b>Descripción</b>
DatosAlineamiento(String Sec1, String Sec2, double xGapOpen, double xGapExt, String xNomMatriz)	Es el constructor de la clase. Inicializa los atributos S1, S2, GapOpen, GapExtended y carga MatrizdeSustitucion.
ObtenerGapOpen();	Permite visualizar el contenido del atributo GapOpen.
ObtenerGapExtended();	Permite visualizar el contenido del atributo GapExtended.
CalculosenMatriz();	Permite que se inserten los datos a la matriz de resultado.
ObtenerNombreMatrizdeSustitucion();	Obtiene el valor del atributo NombreMatrizdeSustitucion.
CargarMatrizSustitucion(String ArchivoMatriz)	Carga la Matriz de Sustitución requerida por el usuario.
ColocarMaximoScore(double xScore)	Asigna el valor al atributo máximo score a ScoreMax.
Alineamiento();	Realiza el alineamiento implementando el algoritmo de Smith y Waterman.
CalcularIdentidad(String a, String b);	Realiza el cálculo de la Identidad, recorriendo cada carácter de las cadenas y contando el número de caracteres iguales en cada posición. Este contador inicializa el atributo Identidad.

ObtenerIdentidad();	Obtiene el valor del atributo Identidad.
MostrarResultados();	Muestra toda la información obtenida en el proceso de alineamiento: las secuencias originales, las secuencias alineadas y el nombre de la matriz de sustitución utilizada. Incluye el valor de la identidad como porcentaje y el máximo score en la matriz de resultado.

**Tabla 5.3 Descripción de la Clase Secuencia**

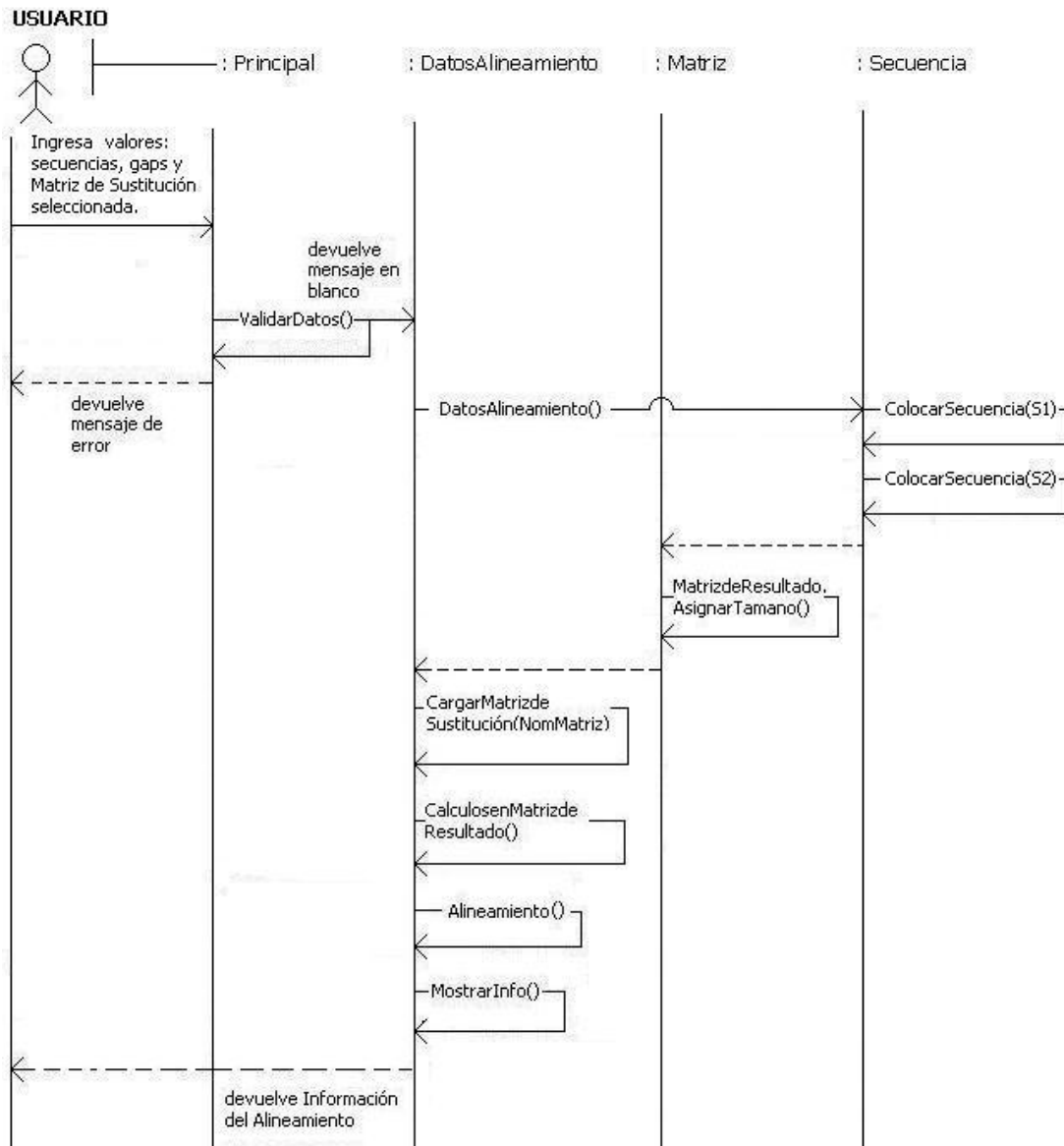
<b>Nombre de la Clase: Secuencia</b>	
<b>Descripción:</b> Contiene todos los atributos y operaciones con las secuencias.	
<b>Atributos</b>	<b>Descripción</b>
Secuencia1	Almacena la cadena de caracteres que representa la secuencia.
<b>Métodos</b>	<b>Descripción</b>
Secuencia();	Constructor de la clase Secuencia.
ColocarSecuencia(String S1)	Asignar la cadena de caracteres que recibe al atributo Secuencia,
ObtenerSecuencia();	Obtiene el valor contenido del atributo Secuencia.
CaracterenCadena(String Cadena, int posicion);	Recibe una cadena y una posición, y devuelve el carácter en esa posición.
IndiceenCadenaBase(char Caracter);	Recibe un carácter y devuelve la posición de ese carácter en la Cadena Base. La Cadena Base es una variable tipo cadena que guarda la forma como están creadas las matrices de sustitución, en donde cada índice de ésta, corresponde a un aminoácido.
ObtenerLongSec(String S);	Obtiene la longitud de una cadena.

**Tabla 5.4 Descripción de la Clase Matriz**

<b>Nombre de la Clase: Matriz</b>
<b>Descripción:</b> La clase Matriz, permite realizar operaciones básicas con matrices, que van desde la asignación del tamaño, hasta la visualización de los datos contenidas en ellas. El ALINEADOR utiliza dos matrices que son: La Matriz de Resultado y la Matriz de Sustitución.

<b>Atributos</b>	<b>Descripción</b>
MatrizUso [][];	Es el tipo de dato que almacena la información.
Fila;	Almacena la fila del mayor elemento de la Matriz de Resultado.
Columna;	Almacena la columna del mayor elemento de la Matriz de Resultado.
<b>Métodos</b>	<b>Descripción</b>
Matriz();	Constructor de la clase Matriz.
AsignarFilayColumna(int xFila, int xColumna);	Guarda una fila y una columna de la matriz.
ObtenerFila();	Obtiene el valor de Fila.
ObtenerColumna();	Obtiene el valor de Columna.
InsertarDato(int xFila, int xColumna, double dato);	Inserta un dato en la Matriz en la posición requerida.
ObtenerDato(int xFila, int xColumna);	Obtiene un dato en la Matriz en la posición requerida.
CargarMatrizSustitucion(int nummatriz);	Las Matrices de Sustitución son archivos “.dat”, y son cargadas dependiendo de los requerimientos de los usuarios.
AsignarTamano(int Filas, int Columnas);	Asigna el tamaño de la MatrizUso. Recibe dos valores, que representan el tamaño de filas y columnas. Para la Matriz de resultado, las filas serán del tamaño de la segunda secuencia y las columnas del tamaño de la primera. Para la Matriz de Sustitución el tamaño siempre será de 23, tanto para las filas como para las columnas.

### 5.3 DIAGRAMA DE SECUENCIAS DEL ALINEADOR



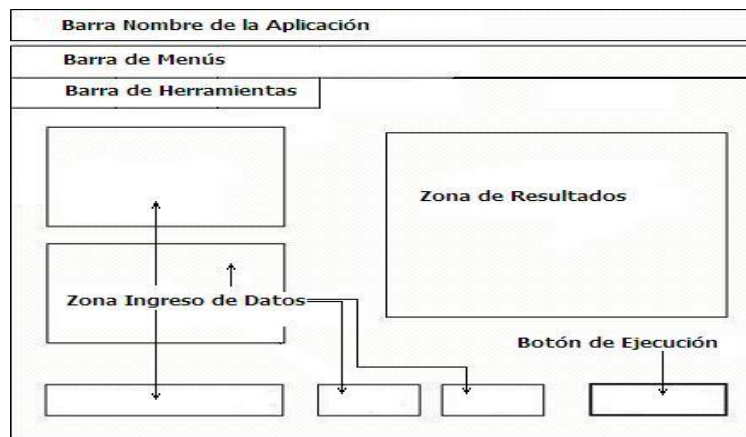
**Figura 5.2 Diagrama de Secuencias del ALINEADOR**

### 5.4 DISEÑO DE LA INTERFAZ GRÁFICA

El buen diseño de una interfaz gráfica es una característica indispensable para lograr una mejor comunicación entre el usuario y el sistema con el cual interactúa.

Mientras mejor es la interfaz desarrollada, mejor será la aceptación del producto diseñado. La interfaz implementada en el software ALINEADOR presenta unas características importantes como son: sencillez y facilidad de manejo, permitiendo con esto que el usuario en poco tiempo interaccione fácilmente con ella.

La interfaz de la aplicación fue diseñada de la siguiente forma:

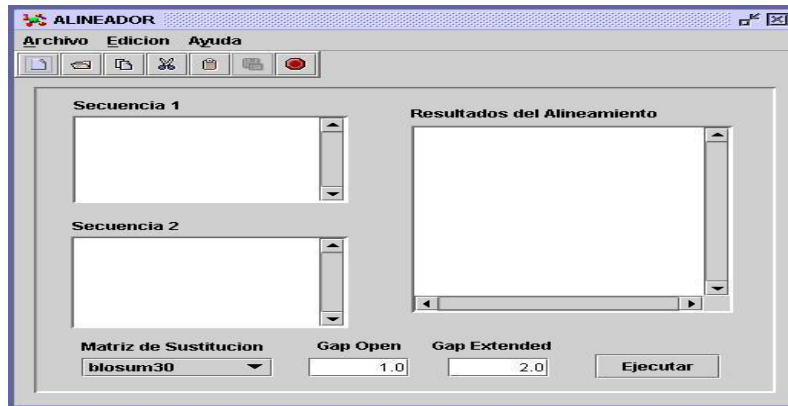


**Figura 5.3 Diseño de la pantalla**

Entre los objetos que contiene encontramos tres barras, una con el nombre de la aplicación, la otra es una barra de menús y la última es una barra de herramientas. Además consta de una zona de ingreso de datos, una zona de resultados y un botón de ejecución

El resultado final de la interfaz grafica es el siguiente:





**Figura 5.4 Interfaz de Usuario**

**La Barra de Menús:** agrupa de manera específica las funciones que se realizan durante la ejecución del ALINEADOR. Consta de las siguientes opciones:

- **Archivo:** Por medio de este menú accedemos a las opciones Nuevo, Abrir, Guardar y Salir.



**Figura 5.5 Menú Archivo**

**Nuevo:** permite reiniciar el programa para iniciar un nuevo alineamiento.

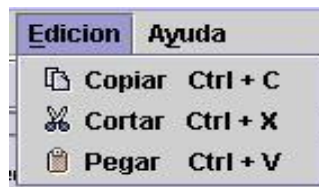
**Abrir:** Carga desde un archivo (de preferencia .txt, o .dat), que contenga una de las secuencias que se necesiten alinear.

**Guardar:** Permite almacenar los resultados obtenidos del alineamiento, incluyendo: Las secuencias originales, las secuencias alineadas, las longitudes de todas las secuencias, la Matriz de Sustitución utilizada, el

valor de la identidad y el máximo valor en la Matriz de Resultado. Todos estos resultados se guardan con formato “\*.doc”

**Salir:** Permite finalizar la ejecución del ALINEADOR.

- **Edición:** por medio de este se accede a las opciones Copiar, Cortar y Pegar.



**Figura 5.6 Menú Edición**

**Copiar:** permite el copiado de las secuencias entradas y las secuencias alineadas.

**Cortar:** permite cortar las secuencias de entrada.

**Pegar:** permite pegar las secuencias copiadas.

- **Ayuda:** por medio de este menú se accede a las opciones Conceptos y acerca de.



**Figura 5.7 Menú Ayuda**

**Conceptos:** muestra algunos conceptos importantes relacionados con el alineamiento de secuencias de proteínas.

**Acerca de:** presenta una ventana con información de los autores.

**La Barra de Herramientas:** contiene accesos directos a las funciones contenidas en la Barra de Menús.



**Figura 5.8 Barra de Herramientas**

**La Zona de Ingreso de Datos:** esta zona está compuesta por los cuadros de texto para ingreso de secuencias, por los cuadros de texto para ingreso de los valores de los Gaps y por un combo para la selección de la matriz de sustitución a utilizar.

Matriz de Sustitucion	Gap Open	Gap Extended
blosum30	1.0	2.0

**Figura 5.9 Zona Ingreso de Datos**

**La Zona de Resultados:** esta zona es un cuadro de texto para mostrar los resultados del Alineamiento.

**Figura 5.10 Zona de Resultados**

## 6. DICCIONARIO DE DATOS DEL ALINEADOR

### 6.1 DESCRIPCIONES DE LOS PROCESOS DE NIVEL CONTEXTUAL.

NOMBRE DEL PROCESO	1.0 Definir Datos de Entrada
DESCRIPCIÓN	Se definen los valores de las Secuencias, los Gaps y la Matriz de Sustitución que va a ser utilizada.
FLUJO DE DATOS INTERNO	Datos ingresados por el usuario.
FLUJO DE DATOS EXTERNO	Secuencia1, secuencia2, Gap Open, Gap Extended, nombre de la Matriz de Sustitución

NOMBRE DEL PROCESO	2.0 Validar Datos de Entrada
DESCRIPCIÓN	Se validan las secuencias, para que éstas no contengan caracteres no válidos (como caracteres especiales, números y letras que no representan aminoácidos). Se valida que los valores de los gaps sean de formato numérico.
FLUJO DE DATOS INTERNO	Secuencia1, secuencia2, Gap Open, Gap Extended.
FLUJO DE DATOS EXTERNO	Valores validados de las variables secuencia1, secuencia2, Gap Open, Gap Extended

NOMBRE DEL PROCESO	3.0 Alinear secuencias
DESCRIPCIÓN	Calcula el mejor alineamiento entre dos secuencias.
FLUJO DE DATOS INTERNO	Valores validados de las variables secuencia1, secuencia2, Gap Open, Gap Extended y nombre de la Matriz de Sustitución.
FLUJO DE DATOS EXTERNO	Datos del Alineamiento, que incluyen los segmentos de resultado, el valor de la identidad y el máximo score en la Matriz de Resultado

## 6.2 DESCRIPCIONES DE LOS FLUJOS DE DATOS EN EL NIVEL CONTEXTUAL.

Nombre del flujo de datos	Entrada de Datos
Descripción	Incluye todos los datos que el usuario ingresa en el ALINEADOR
Desde los procesos	---
Hasta los procesos	1.0 Definición de datos de Entrada.

Nombre del flujo de datos	Datos de Entrada
Descripción	Todos los datos definidos por el usuario: las secuencias, el valor de los gaps y la matriz de sustitución seleccionada.
Desde los procesos	1.0 Definición de Datos de Entrada.
Hasta los procesos	2.0 Validación de Datos de Entrada.

Nombre del flujo de datos	Datos de Entrada Validados
Descripción	Los datos son validados para que permitan el buen funcionamiento del ALINEADOR.
Desde los procesos	2.0 Validación de Datos de Entrada.
Hasta los procesos	3.0 Alineamiento de Secuencias.

Nombre del flujo de datos	Datos del Alineamiento
Descripción	Son los datos arrojados por el ALINEADOR, producto de la aplicación del algoritmo.
Desde los procesos	3.0 Alineamiento de Secuencias.
Hasta los procesos	---

Nombre del flujo de datos	Resultados del Alineamiento
Descripción	Es la suma de los datos previos y los datos obtenidos durante el alineamiento de las secuencias
Desde los procesos	3.0 Alineamiento de Secuencias.
Hasta los procesos	---

## **7. DESARROLLO, IMPLEMENTACIÓN Y PRUEBAS DEL ALINEADOR**

El ALINEADOR fue desarrollado bajo los siguientes criterios:

### **Reusabilidad**

Capacidad de los elementos de software ALINEADOR, de servir para la construcción de muchas aplicaciones diferentes. Los sistemas software a menudo siguen patrones similares. Esta capacidad evita reinventar soluciones a problemas que ya han sido encontrados, y ofrece un nuevo punto de partida para las nuevas aplicaciones.

### **Mantenibilidad y Reparabilidad**

Los pequeños cambios en las especificaciones del problema, no se transformen en altos costos en términos de tiempos de desarrollo. Así mismo debe ser reparable, es decir que si se detecta algún error en el mismo, sea fácil repararlo sin que esto redunde en un cambio total del diseño ni demasiado tiempo de desarrollo.

### **Amigabilidad**

La amigabilidad se relaciona con la facilidad con la cual otras personas con diferentes formaciones pueden aprender a usar el ALINEADOR.

### **Portabilidad**

Hace referencia a la facilidad de utilizar el producto en varios ambientes: un producto de software es más portable cuando más fácil sea lograr que funcione

en diferentes entornos de hardware y software. El requerimiento de portabilidad consiste en que el ALINEADOR funcione en cualquier versión de sistema operativo utilizado por los usuarios potenciales.

El ALINEADOR fue desarrollado en el lenguaje Java, para aprovechar las ventajas que éste lenguaje de programación ofrece y que concuerdan con los criterios establecidos para la elaboración del software.

Las pruebas realizadas al ALINEADOR, se centraron en la búsqueda de posibles escenarios de ejecución que crearan mal funcionamiento del software. A partir de las pruebas realizadas al código (Pruebas de Caja Blanca), se mejoraron las bases necesarias para desarrollar una interfaz gráfica que evitará la aparición de dichos escenarios y su comprobación a través de las pruebas de Caja Negra.

Los errores encontrados, no cambiaron la definición de las clases del ALINEADOR. Estos mostraron las fallas que el software tenía en la validación de los datos de entrada y en la presentación de la información final. Los errores fueron corregidos en su totalidad.

La siguiente tabla muestra las diferentes pruebas realizadas al ALINEADOR, indicando el nombre de la prueba, el tipo de prueba, el resultado y la solución al problema

**Tabla 7.1 Pruebas al ALINEADOR**

PRUEBA	Tipo de Prueba	RESULTADO	SOLUCIÓN
Ejecución con cuadros de entrada de datos vacíos.	Caja Negra	No se inicia el alineamiento	Validar que los cuadros de texto de entrada de datos, no se encuentren en blanco, y en el caso de que esto ocurra, mostrar mensaje de error especificando el dato faltante.

Secuencia con espacios en Blanco.	Caja Blanca	No se encuentra el índice del carácter no valido en la Cadena Base.	Eliminar espacios en blanco de las secuencias con la función trim() de Java, después de validar que la cadena no esté vacía y antes de iniciar el alineamiento.
Formato de Gaps. Valor de gap con más de un punto. (Ejemplo: 4..., 4...2.).	Caja Blanca	Excepción de formato numérico.	Crear mensaje de Error que muestre el valor de gap que no tiene el formato numérico correcto.
Secuencias con caracteres no válidos cargados desde archivo.	Caja Negra	No se encuentra el índice del carácter no válido en la Matriz de Sustitución seleccionada.	Las validaciones de las secuencias se realizarán antes del alineamiento y no durante éste.
Matrices de Sustitución no encontradas en el Directorio.	Caja Negra	No ejecuta el alineamiento.	Realizar la búsqueda de los archivos de las secuencias en el directorio y en caso de que no se encuentren, muestra un mensaje de error, solicitando que se compruebe la ubicación de las matrices y se cierre el ALINEADOR.
Guardar información antes de que el proceso de alineamiento se efectúe.	Caja Negra	Crea archivo en blanco	Activar botón de Guardar, solo después de que el alineamiento se realice.
Prueba al formato de muestra de resultados	Caja Negra	Los espacios de las letras son diferentes. No se muestra con claridad la correspondencia entre las secuencias alineadas.	Cambiar el tipo de letra al cuadro de texto que muestra la información del alineamiento.
Prueba al cálculo de la identidad	Caja Blanca	La Identidad, aunque bien calculada, mostraba muchos decimales.	Corrección de la forma como se muestra la Identidad, de la siguiente forma: multiplicando por 100, redondeando y dividiendo por 100.



## BIBLIOGRAFÍA

- ABASCAL, Federico. Alineamiento de secuencias. Búsqueda de parecidos. Alineamientos múltiples. Parte teórica.  
[http://darwin.uvigo.es/people/fabascal/Teaching/Alineamiento\\_secuencias/teoria.html](http://darwin.uvigo.es/people/fabascal/Teaching/Alineamiento_secuencias/teoria.html)
- ABASCAL, Federico. Familias de proteínas.  
[http://darwin.uvigo.es/people/fabascal/Teaching/Familias\\_proteinas/teoria.html](http://darwin.uvigo.es/people/fabascal/Teaching/Familias_proteinas/teoria.html)
- ABASCAL, Federico. Familias de proteínas. Análisis de secuencias: motivos y perfiles. Parte teórica  
[http://www.pdg.cnb.uam.es/fabascal/COMPLU\\_VERANO\\_03/DIA-2/teoria.html](http://www.pdg.cnb.uam.es/fabascal/COMPLU_VERANO_03/DIA-2/teoria.html)
- CORREDOR, Pilar - GUTIÉRREZ, Eugenia Gina. Algoritmos de comparación de secuencias. Centro de Bioinformática del Instituto de Biotecnología. Universidad Nacional de Colombia.  
<http://bioinf.ibun.unal.edu.co/documentos/algoritmos/algor.php>
- ESTACIÓN EXPERIMENTAL DE ZAIDIN. Análisis de secuencias.  
<http://www.eez.csic.es/cursos/bioinf00/cap2.htm>.
- GIL GARCÍA, Concha. La metodología proteómica, una herramienta para la búsqueda de función. Departamento de Microbiología II, Facultad de Farmacia. Universidad Complutense de Madrid
- PINZÓN VELASCO, Andrés M. Alineamiento de secuencias biológicas: algoritmos básicos. [http://www.andrespinzon.com/tutos/comp\\_cient.pdf](http://www.andrespinzon.com/tutos/comp_cient.pdf)

- PINZÓN VELASCO, Andrés Mauricio. Búsqueda e identificación de nuevos candidatos a vacuna contra la malaria producida por plasmodium vivax. Bogotá D.C., 18 de Diciembre de 2003. Informe de pasantía presentado como requisito parcial para optar al título de Biólogo. Universidad Nacional de Colombia. Facultad de Ciencias. Departamento de Biología
- PRESSMAN, Roger. Ingeniería del software. Quinta Edición. Editorial MC Graw-Hill.
- Ruiz, A. M. Bioquímica estructural: conceptos fundamentales y 383 tests con respuesta razonada. México: Alfa Omega Grupo Editor,
- SOMMERVILLE, Ian. Ingeniería de Software. Sexta Edición. México: Pearson Educación, 2002.
- UNIVERSIDAD COMPLUTENSE DE MADRID. Centro de Genómica y Proteómica. Unidad de Proteómica.  
<http://www.ucm.es/info/gyp/proteomica/presentacion.htm>
- UNIVERSIDAD DE LA HABANA.  
Grupo de Informática. Facultad de Biología.  
<http://fbio.uh.cu/bioinfo/educ.html>
- UNIVERSIDAD MICHOACANA DE SAN NICOLÁS HIDALGO.  
Escuela de Químico Farmacología.  
<http://dieumsnh.qfb.umich.mx/bioquimica/cap1d.htm>

- UNIVERSIDAD NACIONAL DE COLOMBIA.

[http://www.virtual.unal.edu.co/cursos/ingenieria/2001832/docs\\_curso/contenido.html](http://www.virtual.unal.edu.co/cursos/ingenieria/2001832/docs_curso/contenido.html)

<http://www.virtual.unal.edu.co/cursos/ingenieria/2001832/lecciones/alineamiento.html>

[http://www.virtual.unal.edu.co/cursos/ingenieria/2001832/lecciones/matrices\\_sustitucion.html](http://www.virtual.unal.edu.co/cursos/ingenieria/2001832/lecciones/matrices_sustitucion.html)

<http://www.virtual.unal.edu.co/cursos/ingenieria/2001832/lecciones/needleman.html>

<http://www.virtual.unal.edu.co/cursos/ingenieria/2001832/lecciones/smith.html>

- ZAMUDIO, Teodora. regulación jurídica de las biotecnologías. medios y finalidades de la proteómica. <http://www.biotech.bioetica.org/docta30.htm>

## ANEXO A

### MANUAL DE USUARIO

#### REQUERIMIENTOS MÍNIMOS

El ALINEADOR, es una herramienta que le permitirá realizar alineamientos entre secuencias de proteínas, de una forma rápida y sencilla. Para su correcto funcionamiento su computador debe cumplir con los siguientes requisitos:

1. Procesador Intel / AMD de 133 Mhz o superior.
2. Al menos 32 Mb de RAM.
3. Windows 98, XP o versiones posteriores.
4. Resolución de pantalla igual o menor de 800 x 600.
5. Tener instalado el Kit de Desarrollo de Java (JDK), necesario para la ejecución del ALINEADOR. Si no lo tiene instalado búsquelo dentro de este CD de instalación.

#### A.1 Instalación del software ALINEADOR

1. Introduzca el CD en la Unidad. Haga doble click en el archivo Instalar.exe.
2. Enseguida aparecerá la ventana de Bienvenida al Instalador del ALINEADOR. Haga click en “Siguiente” para continuar.



3. Aparecerá una ventana con la información de los autores del ALINEADOR. Haga click en “Siguiente” para continuar.



4. Se muestra ventana donde aparece la ubicación de la carpeta donde serán guardados los archivos del ALINEADOR.



Si esta carpeta no existe, el instalador del ALINEADOR, le preguntará si desea crearla. Haga click en “Si” para continuar.



5. La siguiente ventana muestra la ubicación donde se instalarán los archivos del ALINEADOR. Si desea cambiar la ubicación haga click en “<Atrás”. Si no haga click en “Empezar”.



6. La siguiente ventana se muestra el proceso de instalación del ALINEADOR.



7. El ALINEADOR ha sido instalado en su equipo. Haga Click en “Siguiente” para continuar.



La ultima ventana, es de la aplicación que permitió realizar este instalador. Haga click en “Salir” y ha finalizado el proceso de instalación de ALINEADOR.



## A.2 Acceso al programa ALINEADOR.

Dos de las formas más usuales de acceder al ALINEADOR son:

1. En el proceso de instalación del ALINEADOR, se crea un icono en el escritorio. Haga doble click en el icono e ingresará al programa.



2. Haga click en Inicio, ubíquese en “Todos los Programas” y busque una carpeta de nombre ALINEADOR.



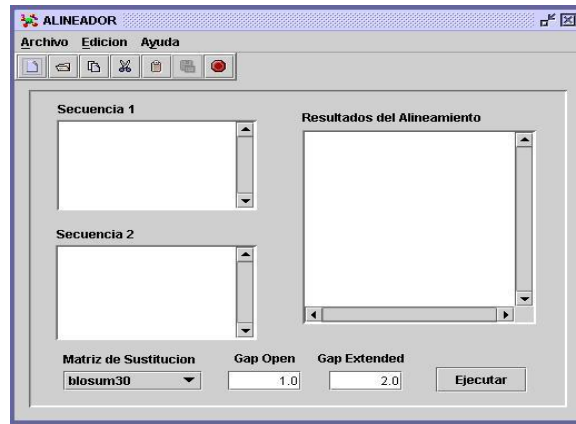
En esta se encuentra el otro acceso que se crea durante la instalación. Haga Click en Alineador y tendrá acceso al programa.



### A.3 Elementos de la aplicación ALINEADOR.

La pantalla de la aplicación ALINEADOR (Algoritmo Smith y Waterman) esta estructurada de la siguiente forma:





Esta ventana presenta una barra de menú con las siguientes opciones:

- **Archivo:** Por medio de este menú accedemos a las opciones Nuevo, Abrir, Guardar y Salir.



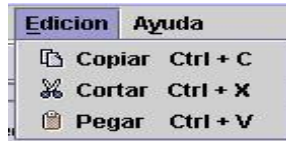
**Nuevo:** permite reiniciar el programa para iniciar un nuevo alineamiento.

**Abrir:** Carga desde un archivo (de preferencia .txt, o .dat), que contenga una de las secuencias que se necesiten alinear.

**Guardar:** Permite almacenar los resultados obtenidos del alineamiento, incluyendo: Las secuencias originales, las secuencias alineadas, las longitudes de todas las secuencias, la Matriz de Sustitución utilizada, el valor de la identidad y el máximo valor en la Matriz de Resultado.

**Salir:** Facilita la salida de la aplicación.

- **Edición:** por medio de este se accede a las opciones Copiar, Cortar y Pegar.



**Copiar:** permite el copiado de las secuencias entradas y las secuencias alineadas.

**Cortar:** permite cortar las secuencias de entrada.

**Pegar:** permite pegar las secuencias copiadas.

- **Ayuda:** por medio de este menú se accede a las opciones Conceptos y acerca de.



**Conceptos:** muestra algunos conceptos importantes relacionados con el alineamiento de secuencias proteómicas.

**Acerca de:** presenta una ventana con información de los autores.

#### A.4 Modo de operación con la interfaz

Para utilizar la aplicación Alineador de Secuencias de Proteínas (Algoritmo Smith y Waterman) se deben seguir los siguientes pasos:

- Carga de Secuencias a alinear.
- Selección de Matriz de Sustitución a utilizar.
- Carga de los valores de los Gap.

- Ejecutar.
- Guardar Resultados.

## 1. Carga de Secuencias:

El usuario ingresa las secuencias que desea alinear, esto lo puede hacer de 2 formas:

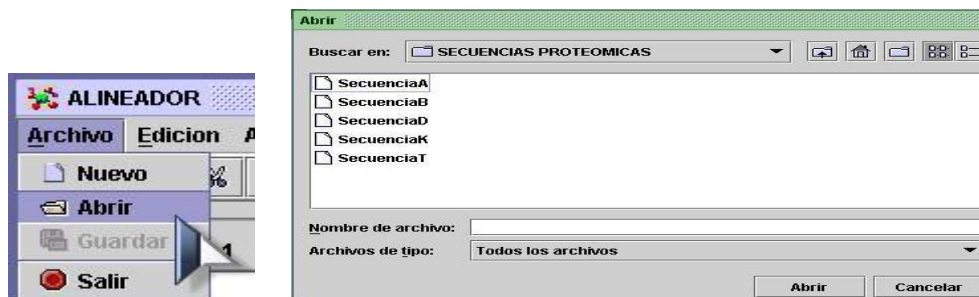
- Ingresándolas mediante el teclado en los espacios asignados.



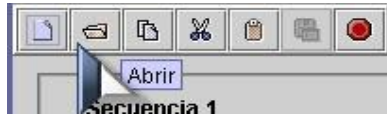
- Ingresándolas por medio de archivos guardados en el disco.

La forma de llegar a estos archivos se puede hacer de 2 maneras:

- Por medio de la barra de Menús.

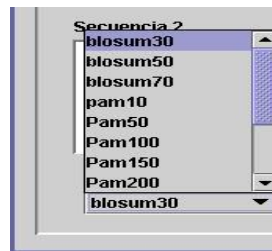


- Por medio de la barra de Herramientas.



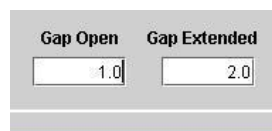
## 2. Selección de Matriz de Sustitución.

El usuario debe escoger la matriz de sustitución que desea utilizar en el alineamiento.



## 3. Carga de los valores Gap.

Se deben ingresar los valores deseados tanto el Gap Open como el Gap Extended.



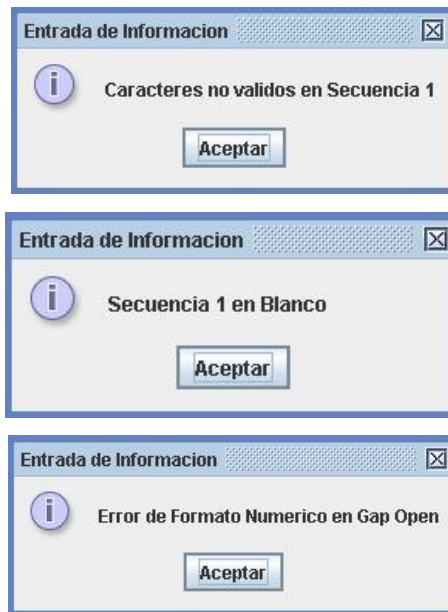
## 4. Ejecutar.



Inicia el proceso de Alineamiento de las secuencias. También valida los datos en los siguientes casos: que en las secuencias no existan caracteres no válidos,

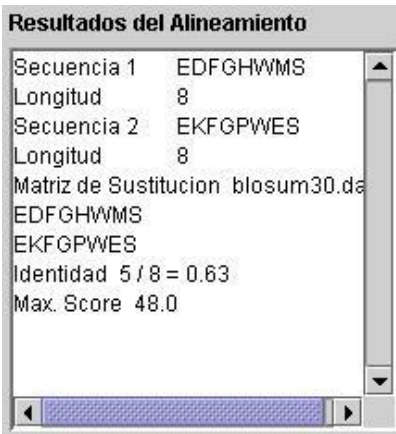
cuando las secuencias hayan sido cargadas desde un archivo; que los cuadros de secuencias y gaps, no estén vacíos; y que el formato de los gaps sea numérico.

En caso de que ocurra alguno de estos casos, el ALINEADOR muestra un mensaje con la situación, para que el usuario lo corrija.



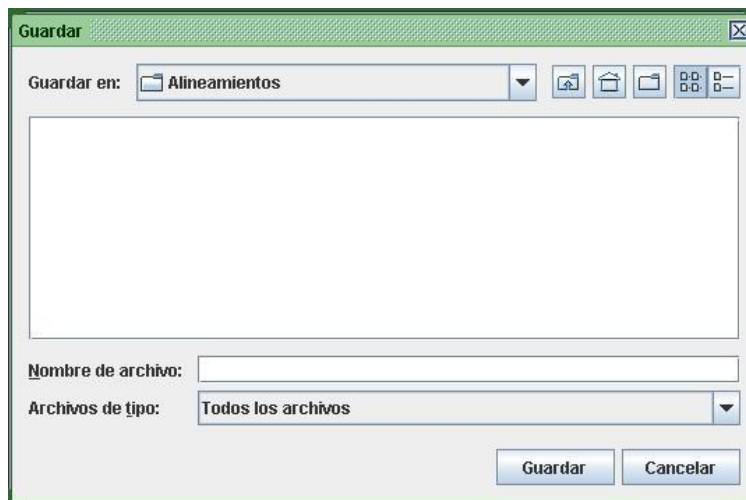
## 5. Resultados del Alineamiento

Este cuadro de texto nos muestra el resultado obtenido de las dos secuencias alineadas. Se aprecia además cuales fueron las dos secuencias originales introducidas, la longitud de cada una de ellas, la matriz de sustitución utilizada, el resultado de la identidad y el máximo score.



## 6. Guardar Resultados

El ALINEADOR, permite guardar los resultados obtenidos durante el proceso de Alineamiento. Haga click en Archivo y luego en Guardar, y busque la ubicación donde desea guardar los resultados. También lo puede hacer haciendo click en el botón Guardar de la Barra de Herramientas.



**ANEXO B**  
**ARTÍCULO CIENTÍFICO**

**IMPLEMENTACIÓN DEL ALGORITMO SMITH Y WATERMAN PARA EL  
ANÁLISIS DE ALINEAMIENTOS DE SECUENCIAS PROTEÓMICAS**

**Daniel Galavís Ibáñez**

**Edgar Jiménez Suárez**

Corporación Universitaria Rafael Núñez  
Facultad de Ingeniería de Sistemas  
Cartagena, Colombia  
dgalavis@hotmail.com  
edjims@hotmail.com

**Resumen:** este artículo describe el desarrollo de la aplicación ALINEADOR, el cual utiliza las técnicas del algoritmo Smith y Waterman, para realizar el alineamiento de secuencias de proteínas, cuyos resultados permitirán medir el grado de similitud o identidad entre el par de secuencias proteómicas alineadas. Se puede decir que el alineamiento de secuencias es una forma de hacer arqueología, ya es indispensable para descubrir las proteínas que tienen un origen común, las partes de las secuencias más conservadas, entre otras muchas aplicaciones. La información obtenida tiene miles de usos, por lo que los alineamientos son una de las herramientas base de toda la Bioinformática.

**Palabras Claves:** alineamiento, secuencias, proteínas, proteómica, bioinformática.

## 1. INTRODUCCIÓN

La proteómica es una disciplina científica que al igual que la genómica, es informacional. Los estudios proteómicos no sólo incluyen la identificación y la cuantificación de proteínas, sino también la determinación de su localización, modificaciones, interacciones, actividades, en última instancia, la determinación de su función.

Una de las maneras más frecuentes de obtener información sobre una o un grupo de secuencias de proteínas incógnitas, es mediante la búsqueda comparativa con otras proteínas existentes. Uno de los métodos comparativos más comunes es el alineamiento de pares de secuencias, que consiste en introducirles espacios para destacar su parecido.

## 2. ALGORITMOS ALINEADORES DE SECUENCIAS

Los algoritmos de comparación de secuencias miden similitud o identidad entre secuencias. La forma que estos algoritmos emplean para

darle significado a cada uno de los alineamientos posibles, es asignándole un valor (*score*) a cada uno de ellos: El score más alto corresponde al mejor alineamiento.

La manera más común de asignar este valor, es a través de una suma simple de scores especificados para cada alineamiento de pares de letras (que representan igualdades o sustituciones), y de letras con caracteres nulos (que representan delecciones o inserciones). El conjunto de estos scores mencionados representa una matriz de scores. Entre las más populares son las **matrices de bloques de sustitución (BLOSUM)** y las **matrices de mutaciones puntuales aceptables (PAM)**.

### 2.1 Algoritmo Smith y Waterman

Este algoritmo encuentra el segmento mejor alineado entre un par de secuencias proteómicas.

Dadas dos secuencias:

$A=a_1a_2...a_n$  y  $B=b_1b_2...b_m$

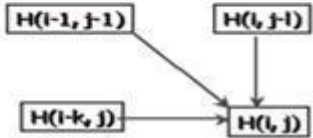


Se define:

- Una función de similitud (coincidencias o Matriz de Sustitución),  $S(a_i, b_j)$ , entre los elementos  $a_i$  y  $b_j$  de las secuencias a alinear. Como las secuencias son proteínas se puede utilizar una matriz de sustitución como la PAM250.
- Los in/dels (inserciones o deleciones) de longitud  $k$  se penalizan con un peso  $W_k$ .
- Se construye una matriz  $H$  de  $n+1$  filas y  $m+1$  columnas. La secuencia  $B$  se ubica en las filas y la secuencia  $A$  en las columnas.

El procedimiento aplicado es el siguiente:

- Se inicializa la matriz  $H$  con ceros.
- La posición  $H_{ij}$  es la máxima similitud de dos segmentos que terminan en  $a_i$  y  $b_j$  respectivamente. Y se obtiene de la siguiente expresión:

$$H(i, j) = \max \begin{cases} 0 \\ H(i-1, j-1) + s(a_i, b_j) \\ H(i-1, j) - w_k \\ H(i, j-1) - w_l \end{cases}$$


El segmento mejor alineado se obtiene de la matriz  $H$  (Matriz de Resultado) ubicando en esta matriz la posición con el valor más alto y por medio de un procedimiento traceback se recuperan los residuos que conforman el segmento. El traceback se detiene cuando encuentra un cero en la diagonal de la matriz. Este procedimiento se puede aplicar de nuevo a la matriz con el fin de recuperar un nuevo segmento alineado.

### 3. MODULO ALINEADOR

El modulo ALINEADOR está desarrollado en un ambiente gráfico fácil de utilizar.

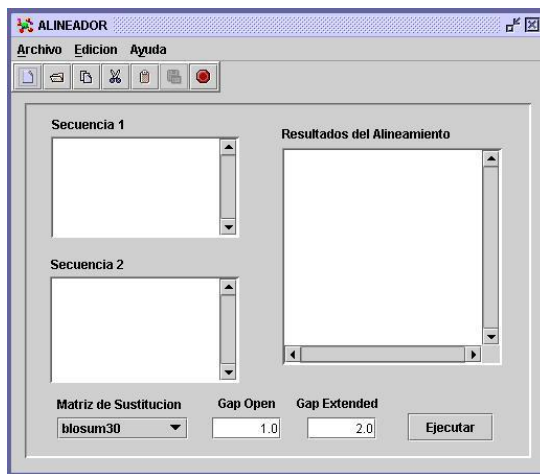


Figura 1. Modulo del ALINEADOR

El usuario ingresa las dos secuencias proteómicas que desea alinear, ingresa los valores del Gap Open y Gap Extended y escoge la matriz de sustitución con la cual desea trabajar.

El sistema, una vez efectuado el alineamiento correspondiente de las dos secuencias, muestra un resumen informativo de los datos de entradas utilizados, el resultado del alineamiento, el máximo score, y el porcentaje de identidad entre las dos secuencias. (Ver Figura 2).

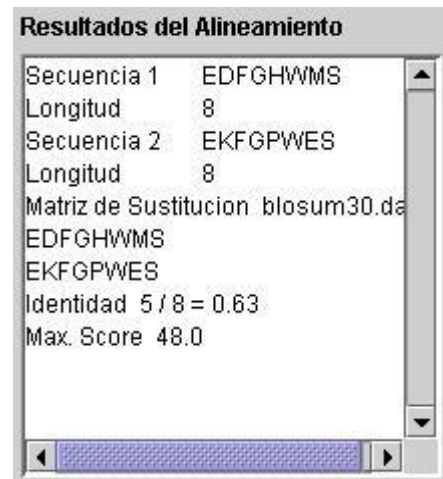


Figura 2. Resultado del Alineamiento

#### 4. CONCLUSIONES

El software ALINEADOR es una sencilla aplicación Bioinformática, se puede decir que es un primer acercamiento al área de la Bioinformática. Esta aplicación se realizó con el propósito de que futuros estudiantes inicien proyectos de investigación en esta área, que se encuentra poco desarrollada en nuestra región.

#### REFERENCIA

- ABASCAL, Federico. ALINEAMIENTO DE SECUENCIAS. BÚSQUEDA DE PARECIDOS. ALINEAMIENTOS MÚLTIPLES. Parte teórica.

[http://darwin.uvigo.es/people/fabascal/Teaching/Alineamiento\\_secuencias/teoria.html](http://darwin.uvigo.es/people/fabascal/Teaching/Alineamiento_secuencias/teoria.html)

- GIL GARCÍA, Concha. LA METODOLOGÍA PROTEÓMICA, UNA HERRAMIENTA PARA LA BÚSQUEDA DE FUNCIÓN.

Departamento de Microbiología II,  
Facultad de Farmacia,  
Universidad Complutense de  
Madrid

- UNIVERSIDAD NACIONAL DE COLOMBIA.  
<http://www.virtual.unal.edu.co/cursos/ingenieria/2001832/lecciones/alineamiento.html>













PAM 100

[illegible]

[illegible]

## PAM 200

[illegible]

[illegible]

[illegible]

[illegible]



## PAM 450

[illegible]



[illegible]

## ANEXO D

### LECTURA DE INFORMACIÓN EN PROSITE

#### -SECUENCIAS CONSENSO Y EXPRESIONES REGULARES-

[http://www.pdg.cnb.uam.es/fabascal/COMPLU\\_VERANO\\_03/DIA-2/teoria.html](http://www.pdg.cnb.uam.es/fabascal/COMPLU_VERANO_03/DIA-2/teoria.html)

Cuando se consulta la información referente a estos aspectos de las proteínas, existe una debilidad y, a veces también mal uso, respecto a algunos términos como son "motivo" o "dominio".

Motivo: si se observa un alineamiento múltiple de proteínas homólogas, algunas columnas varían bastante, mientras que otras están más conservadas. Cuando se observan ciertas columnas cercanas con una alta conservación, o se encuentran ciertos trocitos de las secuencias que se conservan más que otros, y que podrían caracterizar funcionalmente a las proteínas, entonces se habla de MOTIVOS.

Los alineamientos múltiples son la fuente principal para determinar qué partes de la secuencia son más importantes para su función o estructura, y existen diversas aproximaciones para utilizar esta información.

#### **Secuencias consenso:**

La aproximación más sencilla y básica para utilizar la ingente información que contiene un alineamiento múltiple es derivar a partir de éste una *secuencia consenso*, que viene a indicar qué aminoácido es más frecuente en cada posición del alineamiento.

Ejemplo:

```
AGTVATVSC
AGTSATHAC
IGRCARGSC
```

IGEMARLAC  
IGDYARWSC

.....

IGTVARVSC <- Ejemplo de secuencia consenso

### **Patrones - Expresiones regulares:**

Las expresiones regulares se utilizan en muchos ámbitos de la informática. Por ejemplo, cuando se buscan archivos "\*.txt", se buscan todos aquellos que terminen en ".txt". Eso es una expresión regular simple.

Estas expresiones o patrones se pueden utilizar para caracterizar motivos, indicando qué posiciones son más importantes y cuáles pueden variar y qué variaciones pueden sufrir.

¿Cómo expresarse "regularmente"? (el código usado en PROSITE, una base de datos de motivos)

- ♦ Para identificar los aminoácidos se utiliza el código IUPAC: C para cisteína, A para alanina, G para glicina, etcétera.
- ♦ Para identificar posiciones en las que puede haber cualquier aminoácido se usa la "X".
- ♦ Cuando puede haber ambigüedades, es decir, cuando pueden aparecer dos o más residuos en una posición, los ponemos entre corchetes: "[ ]". Por ejemplo: [ATG] significa que puede haber alanina, treonina o glicina.
- ♦ Si queremos indicar qué aminoácidos NO son aceptados, usamos "{ }". Ejemplo: {LIVW} indica que no puede haber ni leucina, ni isoleucina, ni valina, ni triptófano (todos ellos hidrofóbicos).
- ♦ Se usa el separador "-" para separar los diferentes elementos del patrón.

- ♦ (número) se emplea para decir cuántos elementos. Por ejemplo, X(3) se corresponde con x-x-x; A(2,4) con a-a, a-a-a o a-a-a-a.
- ♦ < ó > se utiliza para indicar que la expresión regular es N-terminal o C-terminal.

Ejemplos:

- [AC]-x-V-x(4)-{ED}

Este patrón significa: [Ala o Cys]-cualquiera-Val-cualquiera-cualquiera-cualquiera-cualquiera-{cualquier aa excepto Glu y Asp}

- < A-x-[ST](2)-x(0,1)-V

Este patrón debe encontrarse en posición N-terminal ('<') y significa: Ala-cualquiera-[Ser o Thr]-[Ser o Thr]-(un o ningún aminoácido de cualquier tipo)-Val.

- <{C}\*>

Este patrón lo cumplen todas aquellas proteínas que no contienen cisteínas. El \* significa 'ceros o más elementos'.

Así, a partir del alineamiento anterior:

Ejemplo:

```
AGTVATVSC
AGTSATHAC
IGRCARGSC
IGEMARLAC
IGDYARWSC
```

.....

IGTVARVSC <= Ejemplo de secuencia consenso

Se genera el siguiente patrón:

[AI]-G-X-X-A-[RT]-[SA]-C

Cuando se construyan expresiones regulares se deben tener en cuenta determinados aspectos. Una expresión regular ideal permitirá encontrar a todos los homólogos sin incluir a proteínas no relacionadas. Desafortunadamente, esto no siempre es posible.

Se puede intuir fácilmente que la construcción de un patrón no tiene reglas claras: por ejemplo, en la posición 3 en lugar de X podríamos haber puesto [VSCMY]. Consideramos que es mejor poner X en lugar de [VSCMY] porque ésa parece una posición muy variable y posiblemente cuando conozcamos nuevas proteínas de esa familia alguna tendrá allí algún aminoácido distinto de [VSCMY].

Las expresiones regulares han de ser lo más cortas posibles para evitar ese tipo de situaciones, pero han de ser suficientemente largas para que no aparezcan demasiado frecuentemente por azar, es decir, para que sean específicos de la familia. En cuanto a la penúltima posición observamos que hay serinas y alaninas, pero quizás allí también deberíamos poner una X: siempre es una elección complicada y muchas veces hay que seguir el método de *ensayo y error*, es decir, ir probando y corrigiendo.

Afortunadamente, existen bases de datos como PROSITE donde expertos construyen patrones para los distintos motivos conocidos. Esto lo hacen consultando la bibliografía y analizando alineamientos múltiples. Luego ensayan los patrones sobre Swiss-Prot para estudiar su *sensibilidad y especificidad*. No está de más saber construirlos, especialmente en aquellos casos en que el patrón que nos interesa no está descrito en PROSITE.

La limitación básica de los patrones es la dificultad de definirlos y que son muy estrictos (aunque existen sistemas que pueden buscar con patrones tolerando errores). Básicamente existen dos estados: posiciones importantes y posiciones no importantes (por ejemplo las marcadas como "x"), pero en el mundo real existe una mayor graduación.