VNIVERSIDAD NACIONAL AVTNOMA DE MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

"EVOLUCIÓN MOLECULAR MEDIANTE LA EXPLOTACIÓN DE BASES DE DATOS GENÓMICAS CON HERRAMIENTAS DE BÚSQUEDA LOCAL"

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS (COMPUTACIÓN)

P R E S E N T A:

FELIPE DE JESÚS GUTIÉRREZ LÓPEZ

DIRECTOR DE TESIS: DRA. AMPARO LÓPEZ GAONA

México, D.F. **2005.**

Agradecimientos:

"A Dios Por Ser la Razón Del Ser"

INDICE

Pag.

INTRODUCCIÓN 1

CAPITULO I
1. CONCEPTOS BÁSICOS PARA 4
LA ALINEACIÓN LOCAL DE SECUENCIAS 4
1.1 Genoma 4
1.1.1 ADN 5
1.1.1.1 Nucleótidos 7
1.1.1.2 Relaciones 8
1.1.1.3 Alineamiento de secuencias 8
1.1.1.4 Similitud y homología 10
1.1.1.5 Resumen 10

CAPITULO II
2. EVALUACIÓN DE LA EVOLUCIÓN MOLECULAR 12
2.1 Bases de datos genómicas 12
2.1.1 Fasta (formato de almacenamiento de bases de datos genómicas) 14
2.2 Mecanismo para simular la evolución molecular 16
2.2.1 Generación de puntos atractores 18
2.2.1.1 Generación de subcadenas binarias 18
2.2.1.1.1 Subcadena YR 19
2.2.1.1.2 Subcadena WS 19
2.2.1.1.3 Subcadena MK 19
2.2.1.2 Generación de la matriz de cantidades de nucleótidos 20
2.2.1.3 Generación de la matriz de cantidades binarias 20
2.2.1.4 Evaluación de los modelos matemáticos de evolución 21
2.2.1.4.1 Modelo dYR 21
2.2.1.4.2 Modelo dWS 22
2.2.1.4.3 Modelo dMK 22
2.2.1.4.4 Modelo d_4 22
2.2.1.4.5 Modelo d_5 23
2.2.1.4.6 Modelo d_6 23
2.2.1.4.7 Modelo d_7 23
2.2.1.4.8 Modelo d_8 24
2.2.2. Generación de puntos evolutivos 24
2.2.2.1 Generación de cadenas aleatorias de nucleótidos 25
2.2.2.2 Filtros de proporcionalidad 25
2.2.2.3 Obtención de puntos evolutivos 25
2.2.3 Filtro de atracción 26
2.2.4 Alineación local 26
2.2.5 Resumen 26

CAPITULO III

3. INTERBLAST 27
3.1 Blast 27
3.2 Interblast 34
3.3 Arquitectura de Interblast 36
3.3.1 Capa de la interfaz de usuario 36
3.3.1.1 Árbol Mad* 36
3.3.1.2 Tarjetas del árbol Mad* 38
3.3.1.3 Diagrama general de casos de uso 45
3.3.1.4 Detalle de los casos de uso 46
3.3.2 Capa de dominio del problema (diseño de clases) 54
3.3.3 Capa del manejo de datos 57
3.3.3.1 Estructura de la base de datos y diccionario de datos 59
3.4 Interblast en funcionamiento 65
3.4.1 Formatear archivo genómico (pre-procesamiento de datos) 66
3.4.2 Consultar bases de datos genómicas 69
3.4.3 Generar punto atractor 70
3.4.4 Consultar punto atractor 73
3.4.5 Generar punto evolutivo 74
3.4.6 Consultar punto evolutivo 78
3.4.7 Salir 81
3.5 Resumen 82

CAPITULO IV

4. RESULTADOS 83
4.1 Resultados 83
4.2 Minería de datos 85
4.2.1 Weka 85
4.2.1.1 Pre-procesamiento de datos 85
4.2.1.2 Clasificación 88
4.2.1.3 Agrupación 89
4.2.1.6 Visualización 90
4.3 Resumen 91

CONCLUSIONES 92

BIBLIOGRAFÍA 94

GLOSARIO 95

INTRODUCCIÓN

Los seres vivos al igual que el universo estelar guardan dentro de sí reglas de comportamiento muy complejas que desde tiempos antiguos la humanidad ha tratado de entender.

En la década de los cincuentas se produjo una revolución científica en el campo de la biología al descubrirse que dentro de cada una de las células que componen a un ser vivo y que son capaces de reproducirse, se encuentra una macromolécula denominada ADN (Ácido Desoxirribonucléico) que contiene toda la información morfo-funcional de dicho ser vivo.

El ADN mantiene una configuración de una doble hélice enroscada y enlazada por peldaños compuestos por pares de nucleótidos (sólo existen combinaciones entre cuatro nucleótidos distintos), que la ciencia a lo largo de varios años se dedicó a decodificar para beneficio de la humanidad.

Actualmente se encuentra decodificada toda la información del ADN de seres vivos como: *E. coli*, rata, *gallo, perro, Homo Sapiens*, entre otros. Sin embargo, dicha información no ha sido explotada adecuadamente por la misma naturaleza compleja de su estudio y por las restricciones científico-tecnológicas de la actualidad.

En los Estados Unidos de América, el Centro Nacional para la Información Biotecnológica (NCBI por sus siglas en inglés) ha trabajado durante varios años en el desarrollo de software para el estudio de los genomas (todo el material de ADN de un ser vivo). Uno de sus productos más importantes es: Blast (*Basic Local Alignment Search Tool o Herramienta de Búsqueda de Alineación Local*) que mediante técnicas matemáticas complejos permite comparar grandes cadenas de nucleótidos contra algunas bases de datos genómicas, lo que permite a muchos científicos realizar investigaciones de composición y transformación de los seres vivos.

Actualmente el NCBI ofrece un servicio de alineación de secuencias con Blast vía Internet y el servidor utilizado para dicha tarea es una supercomputadora con varios procesadores de alta velocidad y Gigabytes de memoria RAM, sin embargo utilizar el mecanismo remoto resulta ineficiente debido a que el tiempo de respuesta vía web depende principalmente de la carga de la red y del servidor, de la velocidad de transferencia, del tamaño de la cadena que se desea alinear y de la base de datos a utilizar. Por ejemplo, en una prueba de alineación realizada mediante una conexión por módem de 56Kb a las 2:00 hrs., con una cadena de 100 nucleótidos y usando la base de datos del genoma humano, el tiempo de respuesta fue de aproximadamente de 70 segundos.

El objetivo de esta tesis es crear un mecanismo de respuesta rápida que ayude a comprender la evolución molecular de los seres vivos, mediante procesos de alineación local de secuencias de nucleótidos contra bases de datos genómicas.

Para lograr el objetivo, se desarrolló un sistema de software denominado "Interblast" el cual se encargará de preprocesar archivos genómicos, aplicar los modelos matemáticos elaborados por un grupo de trabajo de la Facultad de Ciencias de la UNAM, realizar las alineaciones locales correspondientes con un motor de alineación de Blast, almacenar y recuperar los resultados en una base de datos relacional para que sean analizados bajo interfaces de consulta web y sistemas de minería de datos.

La estructura de este trabajo es el siguiente:

- *Conceptos básicos para la alineación local de secuencias:* en este capítulo se incluyen conceptos relacionados con los seres vivos (a nivel celular) y con la forma en la cual se lleva a cabo la alineación de secuencias, con el propósito de comprender algunos aspectos de la evolución molecular a través de procesos computacionales.

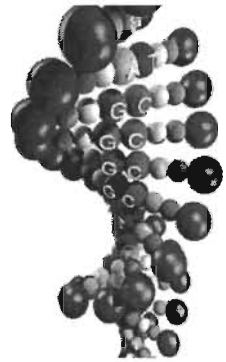
- *Evaluación de la evolución molecular:* en este apartado se muestra la manera en la que la información genética de los seres vivos es almacenada en la computadora, así como también se incluye el mecanismo matemático utilizado para analizar la evolución molecular.

- *Interblast:* contiene información relacionada con el sistema Blast, así como todo el análisis y diseño de la herramienta principal para analizar la evolución molecular denominada "Interblast".

- *Resultados:* muestra los datos obtenidos apartir de la aplicación Interblast y una explicación de los mismos.

Es importante que la ciencia y la tecnología actuales sean aprovechadas al máximo para comprender de una mejor manera la composición, origen y evolución de los seres vivos, ya que de esa manera podremos mejorar día a día nuestra calidad de vida.

CAPÍTULO I



CONCEPTOS BÁSICOS PARA LA ALINEACIÓN LOCAL DE SECUENCIAS

El estudio de la biología ha formado parte de la historia científica del ser humano. Desde sus primeras etapas de razonamiento, el *Homo Sapiens* ha cuestionado el funcionamiento de su entorno y principalmente el de los seres que podemos decir que están vivos (por los procesos de metabolismo que desempeñan).

El ADN desde su descubrimiento en 1869 hasta la fecha ha representado una de las piedras angulares del conocimiento de los seres vivos, ya que dentro de sí se encuentra inscrita gran parte de su información morfológica y funcional.

Gracias a la biología, se han podido erradicar enfermedades mortales y se ha mejorado la calidad de vida de los seres vivos, por eso es importante conocer la composición y evolución de los organismos vivos a lo largo del tiempo.

1.1 Genoma

Al igual que las computadoras requieren de software para que la unidad central de proceso (CPU) realice las tareas que se deseen, así las células de los seres vivos necesitan de determinados códigos para realizar sus funciones [1].

A un conjunto de código que las células necesitan para llevar a cabo ciertas tareas de producción de proteínas (materia prima de todo ser vivo) se le conoce con el nombre de "genoma".

Cada una de las especies que existen en la tierra tiene un genoma específico, el cual ha empezado a secuenciar el ser humano con ayuda de los avances científico-tecnológicos.

El genoma de las especies se determina por su material genético dispuesto en largas cadenas de nucleótidos (ADN), y dentro de dichas cadenas se describen los valores que permiten lograr la diferenciación de las células que forman los diferentes tejidos y órganos de un individuo.

1.1.1 ADN

Los seres vivos se componen de células y dentro de ellas se encuentran macromoléculas de ácido desoxirribonucléico (ADN) con la información necesaria para la producción de cadenas de aminoácidos (polipétidos) que a su vez, dan lugar a las proteínas que son la materia prima de todo ser vivo. Además de lo anterior, en el ADN se encuentra el código de herencia morfológica y funcional que una célula transmite a sus descendientes.

La macromolécula del ADN es un polímero lineal y se compone de dos largas cadenas de monómeros enlazadas a manera de escalera helicoidal. Cada uno de los peldaños de dicha escalera se constituye del enlace de dos nucleótidos con moléculas de fosfato y desoxirribosa (azúcar). Véase la figura 1.1:

Fig. 1.1. Escalera helicoidal del ADN

La cantidad de nucleótidos contenidos dentro del ADN es enorme, pero la diversidad de su alfabeto se reduce sólo a cuatro símbolos que son:

- A (adenina)

- G (guanina)
- T (timina)
- C (citosina)

El hecho de que sólo existan cuatro nucleótidos distintos no significa que la posibilidad de obtener una gran biodiversidad se encuentre reducida, sino que por el contrario, si se mezcla un factor extra llamado *localidad de enlace* dentro de la hélice, el total de combinaciones es enorme.

Los cuatro nucleótidos pueden ser agrupados en dos clases: *Purinas* y *Pirimidinas*.

Purinas: moléculas de dos anillos o también llamadas *grandes*, que incluyen a la adenina y a la guanina.

Pirimidinas: moléculas de un solo anillo o *pequeñas*, dentro de las cuales se encuentra la timina y citosina.

Los enlaces entre nucleótidos son siempre del mismo tipo, de tal manera que si dentro de la escalera se encuentra una molécula de adenina (A) del otro lado siempre existirá una timina (T) y viceversa, y siempre que se halle citosina (C) del lado contrario se verá una guanina (G) y viceversa. Por ello, la información contenida en el ADN se puede deducir apartir de una sola rama de la doble hélice.

En la figura 1.2 se puede observar la regla de asociación de nucleótidos, en la que la adenina sólo puede enlazarse con timina y la guanina con citosina, para formar los pasamanos de la escalera:

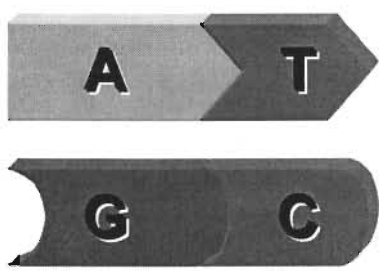


Fig. 1.2. Complementariedad de nucleótidos

La unión de adenina – timina se produce a través de dos puentes de hidrógeno y la de citosina – guanina se hace mediante tres, lo cual hace que el enlace A-T sea más débil que el de G-C. Además al existir una diferencia de tamaño entre purinas y pirimidinas hace que la escalera no sea geométricamente uniforme.

En la figura 1.3 se muestra un esquema detallado de la relación que puede existir entre los cuatro nucleótidos:

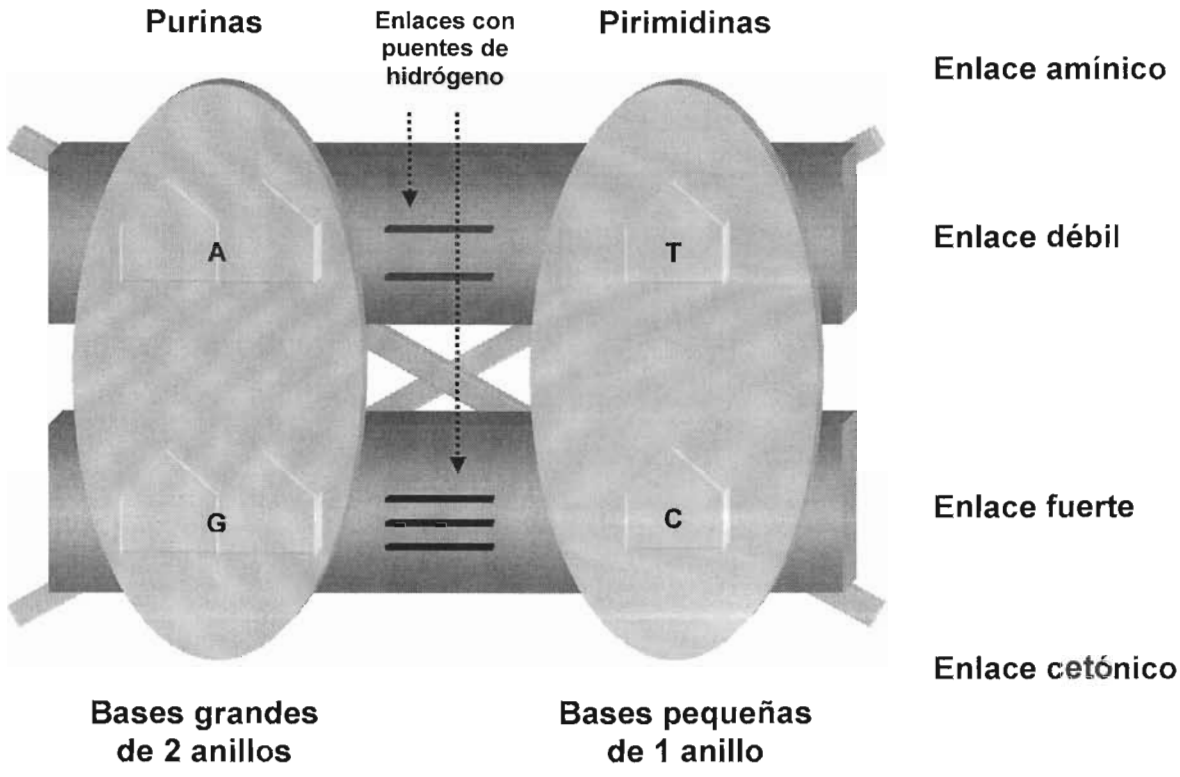


Fig. 1.3. Relación entre nucleótidos

1.1.1.1 Nucleótidos

Los nucleótidos son la unidad de construcción de los peldaños de la escalera del ADN y cada nucleótido se compone de una base nitrogenada (purina o pirimidina), un

azúcar (D-ribosa o 2-deoxi-Ribosa) y un ácido Fosfórico, tal y como se observa en la figura 1.4:

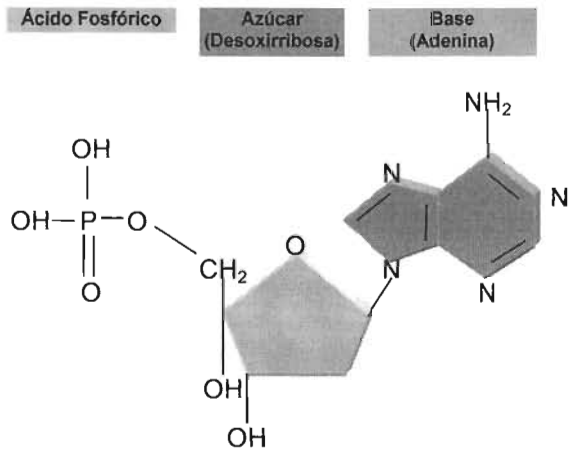


Fig. 1.4. Composición de un ácido nucleico.

1.1.1.2 Relaciones

El genoma de un ser vivo puede ser descrito mediante la formación de una larga cadena de nucleótidos (A, C, G y T), y dicha cadena puede ser transformada en tres cadenas binarias distintas considerando las relaciones siguientes:

1. Grandes – Pequeñas
2. Fuertes – Débiles
3. Amidas - Acetonas

A la primera de las relaciones se le llamará YR, a la segunda WS y a la última MK.

1.1.1.3 Alineamiento de secuencias

El alineamiento de secuencias es un proceso en el que una cadena de nucleótidos A es comparada contra otra cadena de nucleótidos B para encontrar el mayor número posible de coincidencias [8].

Este proceso puede ser clasificado por:

- **Número de secuencias analizadas:**

  - *Alineamiento de un par de secuencias*: este tipo de alineamiento tiene por objetivo encontrar el segmento mejor alineado entre dos secuencias.

  - *Alineamiento múltiple*: como su nombre lo indica, este alineamiento opera sobre varias secuencias al mismo tiempo para obtener una secuencia concenso, esta secuencia concenso tiene en cada posición el nucleótido o el aminoácido (en caso de las proteínas), que más se ha conservado en esa posición en todas la secuencias estudiadas.

- **Nivel de análisis:**

  - *Alineamiento global*: su función es obtener el mejor alineamiento entre dos secuencias.

  - *Alineamiento local*: es el proceso de encontrar el segmento mejor alineado existente entre dos secuencias. Por ejemplo si se tiene una cadena A = ACACGTTGGCATCGATCGACTACG que se desea alinear localmente con una cadena B = CGCGTTTGGCATGACGAACCTACA el resultado es:

Segmento mejor
alineado

1.1.1.4 Similitud y homología

La similitud es el resultado del análisis (observación cuantitativa) de la estructura primaria de dos o más secuencias; la 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.

La homología es una medida cualitativa entre las secuencias, se presenta cuando la similitud que estás 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.

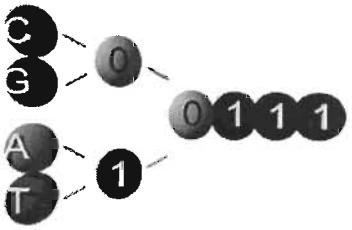
Es válido establecer el porcentaje de similitud de dos o más secuencias, pero esto no es posible para la homología, ya que las secuencias son o no son homólogas.

1.1.1.5 Resumen

Los seres vivos esconden gran parte de su información morfo-funcional dentro del ADN inscrito en cada una de sus células.

El total de información contenida en el ADN de un organismo es enorme, y para explotarla resulta imprescindible utilizar ciertas herramientas y algoritmos computacionales.

CAPÍTULO II



EVALUACIÓN DE LA EVOLUCIÓN MOLECULAR

A mediados del siglo XIX los científicos Alfred Russel Wallace y Charles Darwin descubrieron que los seres vivos modificaban su morfología y funcionamiento para adaptarse a las necesidades del medio ambiente en el que se encontraban, y a este tipo de cambio constante lo denominaron "*evolución por selección natural*".

La ciencia ha postulado que la variabilidad que produce la evolución biológica en sus niveles más bajos (genéticos), se lleva a cabo de manera aleatoria, de tal forma que la mutación azarosa de nucleótidos en el ADN determinará cual ser vivo se adaptará mejor a las condiciones de su medio ambiente.

Uno de los retos de las ciencias genómicas, se encuentra en descubrir si la composición de la biodiversidad a través de la evolución se debe a factores fortuitos o existe una razón que demuestre la causalidad de los seres vivos para adaptarse de una mejor manera al cambio constante de los factores del medio ambiente, sin embargo alcanzar este tipo de conocimiento es muy complicado y resultaría imposible si no se contara con las herramientas computacionales de la actualidad.

2.1 Bases de datos genómicas

En el capítulo I se menciona que el genoma entero de un ser vivo puede ser descrito a través de una larga cadena de nucleótidos (A [Adenina], C [Citosina], G [Guanina] y T [Timina]), por lo tanto es factible plasmar toda esa información en archivos digitales para su procesamiento en computadora.

A todo el conjunto de datos de nucleótidos que describen el genoma de un organismo y que son almacenados de una determinada manera se le denomina "base de datos genómica".

Las formas digitales de almacenar los genomas puede variar dependiendo de las necesidades y/o maneras en que la información será explotada.

Las formas más comunes en las que es posible encontrar bases de datos genómicas se muestran en la tabla 2.1:

Base de Datos	Ventajas	Desventajas
Archivo plano con los nucleótidos consecutivos.	Permite contabilizar el contenido de nucleótidos de una manera fácil y rápida.	El tratamiento de información específica es complejo.
Base de datos relacional con la separación de segmentos en varios campos.	Permite procesar y analizar información en particular ya sea por cromosomas o genes[1].	Imposibilita la alineación local de secuencias.
Archivo plano con bloques de tamaño regular.	Facilita la alineación local de secuencias dentro de Blast.	Complica el procesamiento de información específica, así como la contabilización del contenido de nucleótidos

Tabla 2.1 Diversidad de bases de datos genómicas

Es importante remarcar que en la actualidad los investigadores han descifrado y almacenado los genomas de diversas maneras, y que la mejor configuración es aquella que permita lograr una óptima explotación de la información.

[1] Los genes son subsegmentos de un cromosoma y un cromosoma es un subsegmento de un genoma.

2.1.1 Fasta (formato de almacenamiento de bases de datos genómicas)

Dentro de este trabajo se utiliza un formato de almacenamiento de bases de datos genómicas denominado: "Fasta", el cual guarda la información en un archivo plano, dividido en secciones o bloques de caracteres (nucleótidos) bien determinados. Por ejemplo el genoma completo de la bacteria *E. coli* se estructura en 400 secciones con su respectivo encabezado, y cada una de las secciones contiene a su vez 10,595 caracteres, de tal manera que la base de datos completa se compone de 10,595 x 400=4,238,000 caracteres (bytes).

A continuación se incluye un segmento de una sección correspondiente al archivo Fasta de la bacteria *E. coli*:

```
>gi|1786181|gb|AE000111.1|AE000111 Escherichia coli K-12 MG1655 section 1 of 400 of the complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTG
GTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGAC
AGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGT
AACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGG
TAACGAGGTAACAACCATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCG
ATATTCTGGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTG
GCGATGATTGAAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTT
GACGGGACTCGCCGCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTTGCCCAAATAA
AACATGTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAA
ATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCT
GGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTG
ATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGAC
TACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTG
CGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCG
CTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATACCGGAAATCCT
CAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCCAATCTGAATAACAT
GGCAATGTTCAGCGTTTCTGGTCCGGGGATGAAAGGGATGGTCGGCATGGCGGCGCGCGTCTTTGCAGCGATGTCACGCG
...
```

Diversos centros, institutos y universidades han trabajado de manera conjunta para decodificar los genomas de ciertas especies para su aprovechamiento por la ciencia.

El Centro Nacional para la Información Biotecnológica (NCBI) de los Estados Unidos de Norteamérica [8], comparte al mundo mediante servicios de Internet cierta gama de bases de datos genómicas en formato fasta, y algunas de ellas son:

a) AeropYRum pernix
b) Aquifex aeolicus
c) Arabidopsis thaliana
d) Bacillus subtilis
e) Bos taurus
f) Caenorhabditis elegans
g) Canis familiares
h) Danio rerio
i) Dictyostelium discoideum
j) Drosophila melanogaster
k) Escherichia coli
l) Gallus gallus
m) Homo sapiens
n) Human immunodeficiency virus type 1
o) Methanococcus jannaschii
p) Mus musculus
q) Oryctolagus cuniculus
r) Oryza sativa
s) Ovis aries
t) Plasmodium falciparum
u) Rattus norvecigus
v) Saccharomyces cerevisiae
w) Schizosaccharamyces pombe
x) Xenopus tropicales
y) Simian immunodeficiency virus
z) Sus scrofa
aa) Synechocystis
bb) Takifugu rubripes

cc) Xenus lavéis

dd) Zea mays

2.2 Mecanismo para simular la evolución molecular

Antes de incluir el procedimiento para simular la evolución molecular es preciso definir los dos conceptos siguientes:

1) Punto atractor: es un vector de ocho componentes que definen de manera matemática y única a un organismo.

2) Punto evolutivo, es un vector de ocho componentes (de una cadena aleatoria) que se acerca en un determinado radio al punto atractor de un organismo. Dicho punto también cumple con una cierta proporcionalidad de nucleótidos y su cadena origen al ser alineada localmente contra la base genómica de un organismo, devuelve alineaciones exitosas. Por lo tanto, dicho punto define a la cadena como candidata para ser considerada evolutivamente compatible con el organismo

Para intentar descubrir patrones de evolución molecular en los seres vivos, computacionalmente se deben de llevar a cabo las siguientes tareas:

1) Descargar la base de datos genómica en formato Fasta (desde el portal de NCBI) de algún organismo que se vaya a analizar.

2) Preprocesar (formatear) el archivo Fasta para que en un paso más adelante se puedan llevar a cabo alineaciones locales de secuencias de manera automatizada.

3) Generar un punto atractor de ocho dimensiones, apartir de la aplicación de ciertos modelos matemáticos evaluados sobre el genoma de un organismo (en formato Fasta).

4) Crear puntos evolutivos apartir de cadenas de nucleótidos generadas aleatoriamente.

5) Evaluar la proximidad de los puntos evolutivos con respecto a los puntos atractores.

   Si el punto evolutivo atraviesa ciertos filtros y se aproxima al punto atractor en un radio determinado, la cadena apartir de la que se creó dicho punto evolutivo será alineada de forma local en contra de la base de datos genómica (en formato Fasta).

6) Almacenar los resultados de la alineación local en una base de datos para su futuro análisis.

En la figura 2.1 se puede ver que se generan muchos puntos evolutivos para evaluarlos sobre un punto atractor, que permita decidir si la cadena de origen es digna de ser alineada:

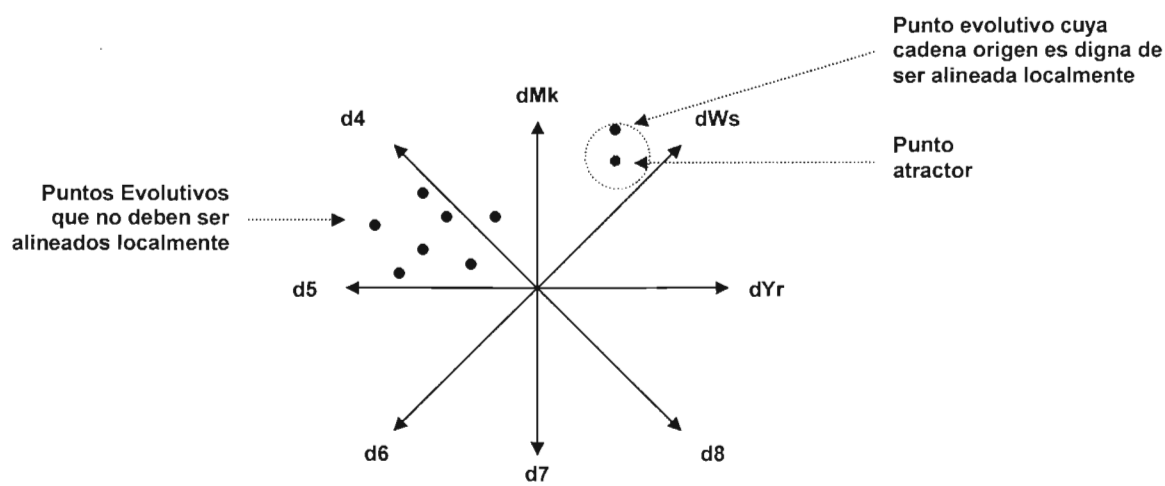


Fig. 2.1 Evaluación de puntos evolutivos sobre un punto atractor

Cabe señalar que los modelos matemáticos que rigen cada una de las dimensiones, fueron desarrollados en la Facultad de Ciencias de la UNAM para buscar patrones de evolución no azarosos.

2.2.1 Generación de puntos atractores

Para generar un punto atractor, es necesario extraer y procesar subcadenas de tamaño M del archivo Fasta de un determinado genoma.

El punto atractor de cada genoma se obtendrá de promediar la aplicación de ocho modelos matemáticos por cada una de las subcadenas.

Para obtener subcadenas de nucleótidos de tamaño M, es necesario depurar el contenido del genoma de un ser vivo apartir de un archivo en formato Fasta, tal y como se muestra en la figura 2.2:

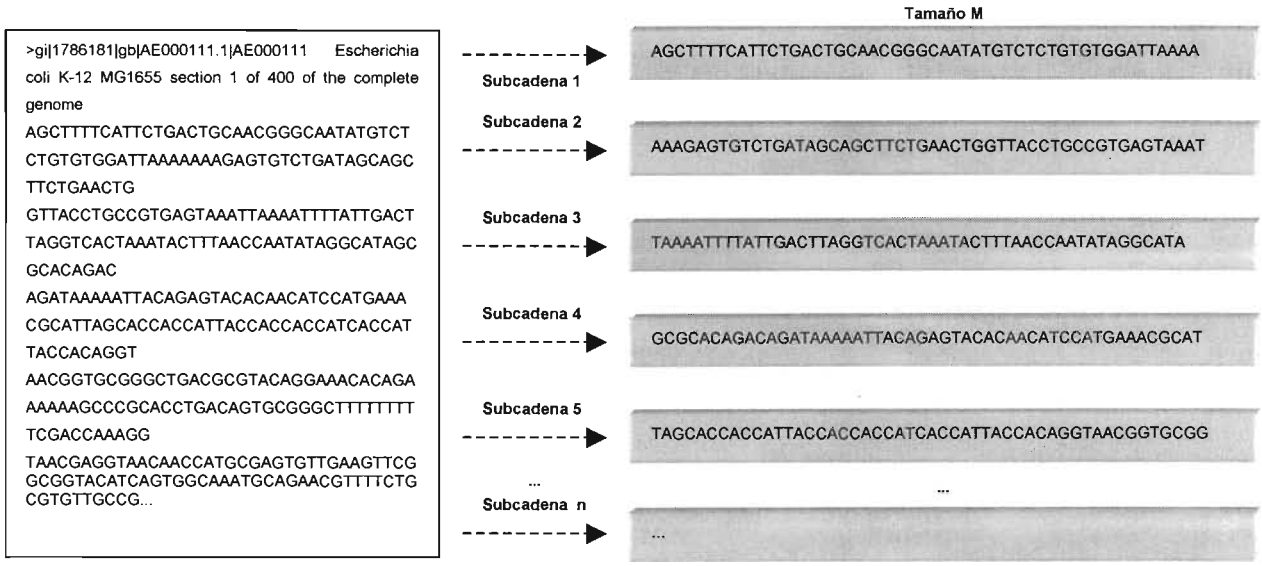


Fig. 2.2 Extracción de subcadenas de un archivo Fasta

2.2.1.1 Generación de subcadenas binarias

Apartir de cada subcadena de nucleótidos extraída del archivo Fasta, es posible obtener tres subcadenas binarias (ceros y unos) distintas, considerando las reglas de asociación siguientes:

2.2.1.1.1 Subcadena YR

Esta subcadena binaria se obtiene de considerar la relación molécular: grandes – pequeñas [10]. En la tabla 2.1 se muestran las equivalencias de cada una de las bases nitrogenadas:

YR		
Grandes	**A**	**0**
	G	
Pequeñas	**T**	**1**
	C	

Tabla 2.2 Conversión a la subcadena YR

2.2.1.1.2 Subcadena WS

Para la subcadena WS, es necesario considerar la naturaleza: fuertes – débiles [10] , y en la tabla 2.3 se incluyen las equivalencias:

WS		
Fuertes	**G**	**1**
	C	
Débiles	**A**	**0**
	T	

Tabla 2.3 Conversión a la subcadena WS

2.2.1.1.3 Subcadena MK

La última cadena tiene que ver con la relación: amina – cetona [10], y su mapeo es el siguiente:

MK		
Aminas	A	1
	C	
Cetonas	T	0
	G	

Tabla 2.3 Conversión a la subcadena MK

2.2.1.2 Generación de la matriz de cantidades de nucleótidos

Apartir de la subcadena de nucleótidos se contabilizan de manera unitaria y por pares a todas las bases, de tal forma que se obtenga una matriz de 5x4 como la que se muestra en la tabla 2.4:

	A	C	G	T	Total
A	N_{AA}	N_{AC}	N_{AG}	N_{AT}	N_A
C	N_{CA}	N_{CC}	N_{CG}	N_{CT}	N_C
G	N_{GA}	N_{GC}	N_{GG}	N_{GT}	N_G
T	N_{TA}	N_{TC}	N_{TG}	N_{TT}	N_T

Tabla 2.5 Matriz de cantidades de nucleótidos

El conteo de la matriz permite que más adelante se apliquen ciertos modelos matemáticos.

2.2.1.3 Generación de la matriz de cantidades binarias

Al igual que con la matriz de nucleótidos, es preciso crear una matriz de cantidades unitarias y por pares de todas las cadenas binarias (YR, WSy MK) tal y como se muestra en la tabla 2.5:

	N_0	N_1	N_{00}	N_{01}	N_{10}	N_{11}
YR	N_0	N_1	N_{00}	N_{01}	N_{10}	N_{11}
WS	N_0	N_1	N_{00}	N_{01}	N_{10}	N_{11}
MK	N_0	N_1	N_{00}	N_{01}	N_{10}	N_{11}

Tabla 2.6 Matriz de cantidades binarias

2.2.1.4 Evaluación de los modelos matemáticos de evolución

La extracción de subcadenas del archivo Fasta, la creación de subcadenas binarias y el conteo de sus contenidos, son parte del proceso de obtención de los datos necesarios para generar cada uno de los ocho valores del punto atractor. Y los modelos matemáticos a utilizar para dicho fin se denominarán: dYR, dWS, dMK, d4, d5, d6, d7 y d8.

2.2.1.4.1 Modelo dYR

Este modelo matemático [10] utiliza los valores obtenidos en la matriz de combinaciones binarias correspondientes a la cadena YR:

$$dYR = \frac{N_{00}N_{11} - N_{01}N_{10}}{N_0 N_1}$$

donde:

N0: total de 0's en la cadena YR.

N1: total de 1's en la cadena YR.

N00: total de 00's en la cadena YR.

N11: total de 11's en la cadena YR.

N01: total de 01's en la cadena YR.

N10: total de 10's en la cadena YR.

2.2.1.4.2 Modelo dWS

El valor de la variable dWS [10] se obtiene apartir de los datos generados con la cadena WS.

$$dWS = \frac{N_{00}N_{11} - N_{01}N_{10}}{N_0 N_1}$$

2.2.1.4.3 Modelo dMK

La materia prima de este modelo se obtiene de la cadena binaria MK [10].

$$dMK = \frac{N_{00}N_{11} - N_{01}N_{10}}{N_0 N_1}$$

2.2.1.4.4 Modelo d_4

De aquí en adelante, todos los modelos utilizan la matriz de cantidades de nucleótidos:

$$d_4 = \%C + G(f_C + f_G) = \frac{C+G}{M}$$

donde:

C+G: suma de C's con G's

M: tamaño de la subcadena de nucleótidos.

f= frecuencia.

2.2.1.4.5 Modelo d_5

$$d_5 = \%CpG(f_{CG}) = \%[CG] = \frac{CG}{M-1}$$

donde:

CG: total de CG's en la cadena

M: tamaño de la subcadena de nucleótidos.

f= frecuencia

2.2.1.4.6 Modelo d_6

$$d_6 = \frac{1}{4}\left[C_{CA} + C_{CG} + C_{TA} + C_{TG}\right]$$

$$d_6 = ¼\ [f_{CA} - f_C f_A + f_{CG} - f_C f_G + f_{TA} - f_T f_A + f_{TG} - f_T f_G]$$

$$d_6 = \frac{1}{4}\left[\left(\frac{N_{CA}}{M-1}\right) - \left(\frac{N_C}{M}\right)\left(\frac{N_A}{M}\right) + \left(\frac{N_{CG}}{M-1}\right) - \left(\frac{N_C}{M}\right)\left(\frac{N_G}{M}\right) + \left(\frac{N_{TA}}{M-1}\right) - \left(\frac{N_T}{M}\right)\left(\frac{N_A}{M}\right) + \left(\frac{N_{TG}}{M-1}\right) - \left(\frac{N_T}{M}\right)\left(\frac{N_G}{M}\right)\right]$$

donde:

NCA: total de CA's en la cadena

NC: total de C's en la cadena

NA: total de A's en la cadena

...

M: tamaño de la subcadena de nucleótidos.

2.2.1.4.7 Modelo d_7

$$d_7 = \frac{1}{3}\left[C_{TC} + C_{GC} + C_{GA}\right]$$

$$d_7 = 1/3\ [\ f_{TC} - f_T f_C + f_{GC} + f_G f_C + f_{GA} - f_G f_A]$$

$$d_7 = \frac{1}{3}\left[\left(\frac{N_{TC}}{M-1}\right)-\left(\frac{N_T}{M}\right)\left(\frac{N_C}{M}\right)+\left(\frac{N_{GC}}{M-1}\right)-\left(\frac{N_G}{M}\right)\left(\frac{N_C}{M}\right)+\left(\frac{N_{GA}}{M-1}\right)-\left(\frac{N_G}{M}\right)\left(\frac{N_A}{M}\right)\right]$$

2.2.1.4.8 Modelo d_8

$$d_8 = \frac{1}{6}\left[C_{AA}+C_{TT}+C_{CC}+C_{CT}+C_{AG}+C_{GG}\right]$$

d8= *1/6* $[f_{AA} - f_A f_A + f_{TT} - f_T f_T + f_{CC} - f_C f_C + f_{CT} - f_C f_T + f_{AG} - f_A f_G + f_{GG} - f_G f_G]$

$$d_8 = \frac{1}{4}\left[\begin{array}{l}\left(\frac{N_{AA}}{M-1}\right)-\left(\frac{N_A}{M}\right)\left(\frac{N_A}{M}\right)+\left(\frac{N_{TT}}{M-1}\right)-\left(\frac{N_T}{M}\right)\left(\frac{N_T}{M}\right)+\left(\frac{N_{CC}}{M-1}\right)-\left(\frac{N_C}{M}\right)\left(\frac{N_C}{M}\right)+\left(\frac{N_{CT}}{M-1}\right)-\left(\frac{N_C}{M}\right)\left(\frac{N_T}{M}\right)\\ +\left(\frac{N_{AG}}{M-1}\right)-\left(\frac{N_A}{M}\right)\left(\frac{N_G}{M}\right)+\left(\frac{N_{GG}}{M-1}\right)-\left(\frac{N_G}{M}\right)\left(\frac{N_G}{M}\right)\end{array}\right]$$

Las ocho variables anteriores definen las propiedades del punto atractor de un genoma.

Para más información sobre los modelos d_6, d_7 y d_8 refiérase al artículo [9].

2.2.2. Generación de puntos evolutivos

Una vez que se cuenta con un punto atractor, es necesario generar muchos puntos evolutivos apartir de cadenas de nucleótidos fabricadas de manera aleatoria. El resultado de la comparación de los puntos evolutivos contra el atractor, definirá si la cadena de la que proviene se deberá alinear de manera local contra la base de datos en formato Fasta.

2.2.2.1 Generación de cadenas aleatorias de nucleótidos

El primer paso consiste en construir una cadena aleatoria de nucleótidos de tamaño M, para después aplicarle dos filtros de proporcionalidad para saber si dicha cadena puede ser evaluada en un siguiente nivel o deberá ser desechada.

2.2.2.2 Filtros de proporcionalidad

De la cadena aleatoria generada, es necesario contabilizar todos los nucleótidos contenidos. Si la proporción de las bases se encuentra dentro del rango siguiente:

$$A+T \rightarrow [40 - 60]\%$$
$$C+G \rightarrow [40 - 60]\%$$

la cadena podrá ser procesada en el siguiente paso, de lo contrario deberá de ser ignorada y será preciso repetir el paso anterior hasta encontrar una cadena que supere dicho filtro.

2.2.2.3 Obtención de puntos evolutivos

Apartir de este paso, el proceso para crear un punto evolutivo es exactamente igual al de la obtención de un punto atractor, es decir, se debe hacer lo siguiente:

- Generar las subcadenas binarias:
  - Subcadena YR.
  - Subcadena WS.
  - Subcadena MK.
- Generar la matriz de cantidades de nucleótidos.
- Generar la matriz de cantidades binarias.
- Aplicar los modelos matemáticos de evolución:
  - dYR.

- dWS.
- dMK.
- d4.
- d5.
- d6.
- d7.
- d8.

2.2.3 Filtro de atracción

Una vez que se cuenta con el punto evolutivo, se compara con el atractor de algún genoma determinado y se verifica que la distancia entre los ocho valores no sea mayor al 10%.

2.2.4 Alineación local

Una vez que se han superado todas las barreras anteriores, la cadena aleatoria se alinea de manera local sobre la base de datos genómica en cuestión. Y los resultados se almacenan en una base de datos relacional para ser analizados de manera manual y con la ayuda de herramientas de minería de datos.

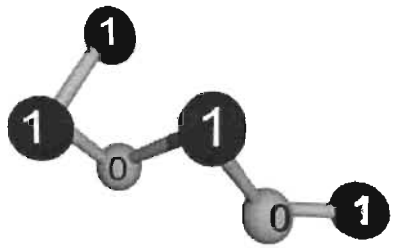
2.2.5 Resumen

De forma matemática es posible definir a un organismo, mediante la obtención de un punto de ocho componentes, que reflejan el contenido y proporción de nucleótidos de su ADN.

Con la proporción genética de una cadena determinada, es posible evaluarla matemáticamente para saber si puede ser considerada evolutiva a un organismo.

En el siguiente capítulo se verá cómo construir una aplicación eficiente (Interblast), que lleve a cabo los procesos de evaluación molecular para intentar encontrar patrones evolutivos en los seres vivos.

CAPÍTULO III



INTERBLAST

Dentro de todo el proceso de deducción de la evolución molecular existe un procedimiento muy relevante denominado: "Alineación local".

Alinear cadenas de nucleótidos sobre una base de datos genómica no es un proceso sencillo y exige el consumo de mucho tiempo y grandes cantidades de recursos computacionales

Durante varios años, el NCBI ha trabajado en la construcción de un software denominado Blast [8], capaz de realizar alineaciones locales sobre bases de datos genómicas en formato Fasta.

A lo largo del presente capítulo se muestra como adaptar y mejorar los servicios que ofrece NCBI para llevar a cabo procesos computacionales no existentes, que beneficien el estudio de la evolución molecular.

3.1 Blast

Es un conjunto de programas creados por NCBI para la alineación local de secuencias contra bases de datos genómicas. Sus siglas significan: *Basic Local Alignment Search Tools o Herramientas de Búsqueda de Alineación Local Básica*.

Los algoritmos implementados en Blast son muy complicados pero sus resultados son altamente confiables y en tiempos muy reducidos.

Actualmente el NCBI ofrece un servicio de búsqueda sobre Blast vía Internet, lo cual permite a muchos científicos realizar sus investigaciones de una manera veraz y

eficiente. Sin embargo, dicho servicio resulta inoperante para las necesidades planteadas en el presente trabajo, debido a que el tiempo de espera es alto y no incluye los mecanismos de aplicación de los modelos matemáticos desarrollados en la Facultad de Ciencias.

Afortunadamente el código fuente y ejecutable de Blast se encuentran disponible para el uso libre de la comunidad científica, lo cual permite que cualquier persona lo adapte a sus necesidades.

Los programas contenidos dentro de Blast son:

blastp: permite comparar una secuencia de aminoácidos contra una base de datos de secuencias de proteínas.

blastn: compara una secuencia de nucleótidos contra una base de datos de secuencias de nucleótidos.

blastx: compara una secuencia de nucleótidos traducida en sus seis posibles marcos de lectura contra una base de datos de secuencias de proteínas.

tblastn: compara una secuencia de aminoácidos contra toda la base de datos de nucleótidos traducida en sus seis posibles marcos de lectura. Si se necesitara realizar este cálculo con Fasta sería necesario realizar las traducciones de las secuencias en los distintos marcos de lectura y ejecutar la búsqueda para cada uno de los seis marcos.

tblastx: compara las seis traducciones en sus marcos de lectura de la secuencia de nucleótidos, contra las seis traducciones en sus marcos de lectura de toda la base de datos de nucleótidos.

En la figura 3.1 se pueden ver algunos de los servicios Blast que NCBI ofrece al mundo vía Internet:

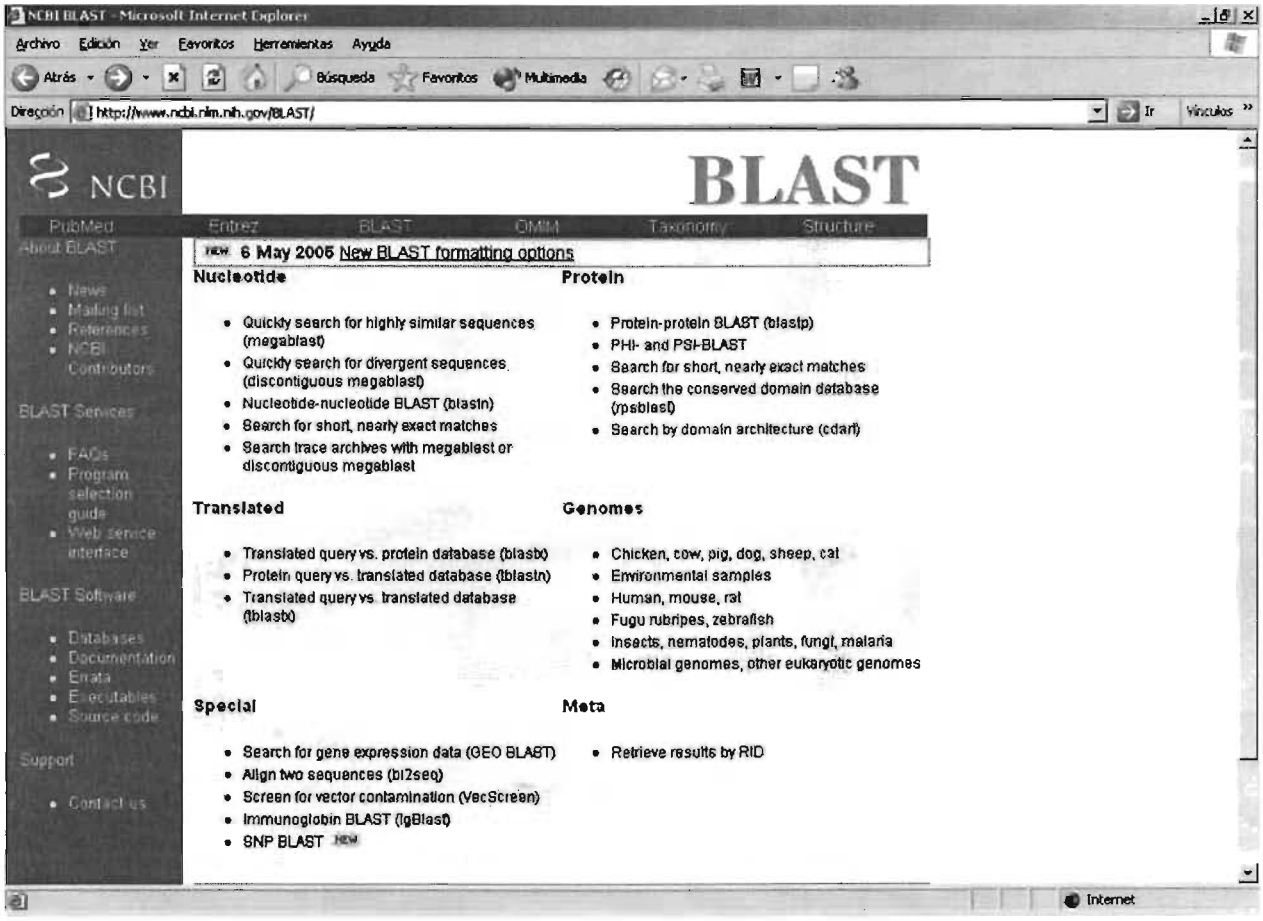


Fig. 3.1. Servicios de Blast

En la figura 3.2 se muestra un ejemplo de uso del sistema Blast vía Internet, en el cual se alinea localmente una cadena de nucleótidos sobre el genoma de una bacteria, en este caso *E. coli*:

Cadena para alinear

Genoma objetivo

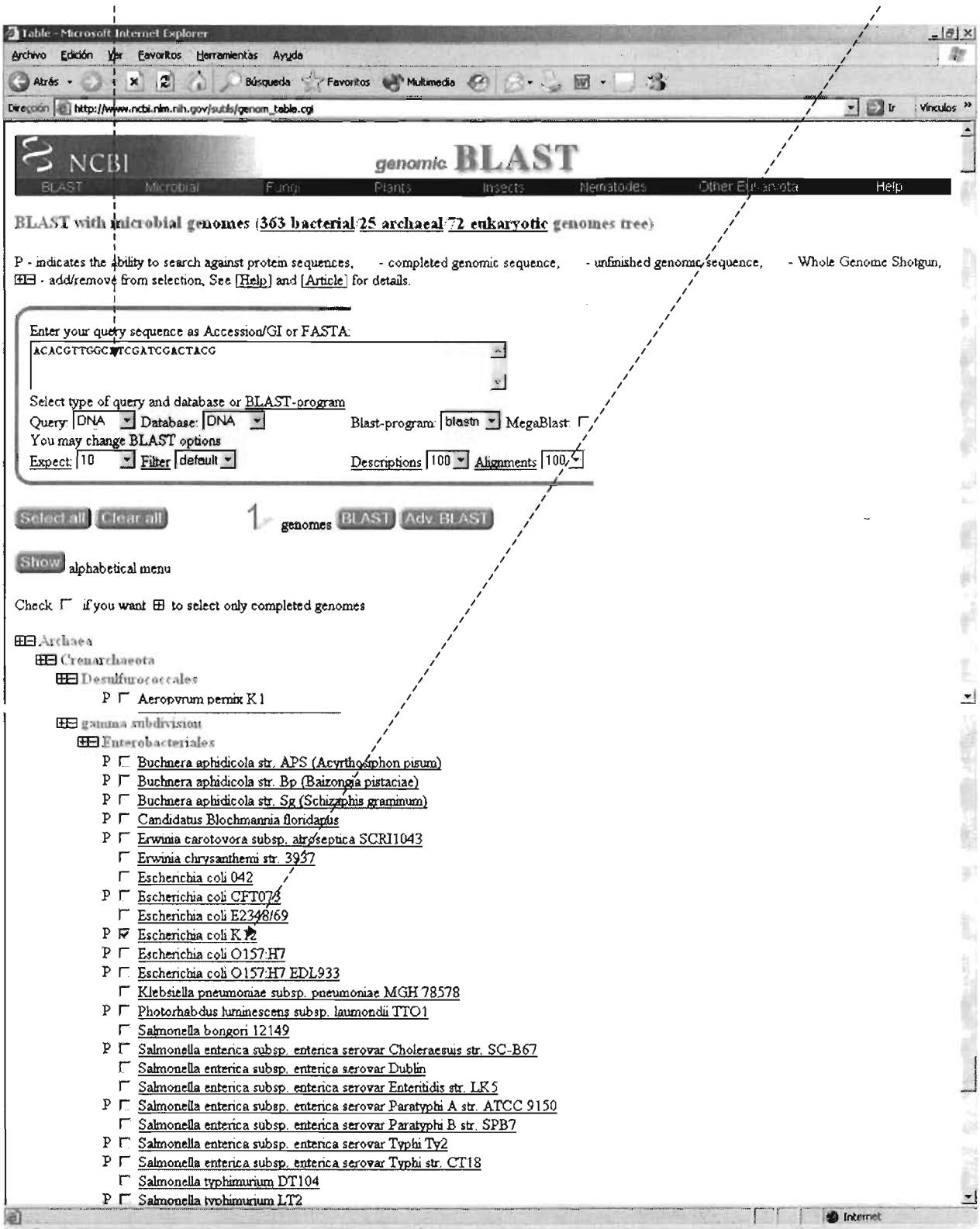


Fig. 3.2. Escritura de la cadena de nucleótidos y selección del genoma sobre el que se llevará a cabo la alineación local

Después de enviar la cadena para que Blast la procese, el sistema muestra una ventana (Fig. 3.3) con el tiempo estimado de respuesta:

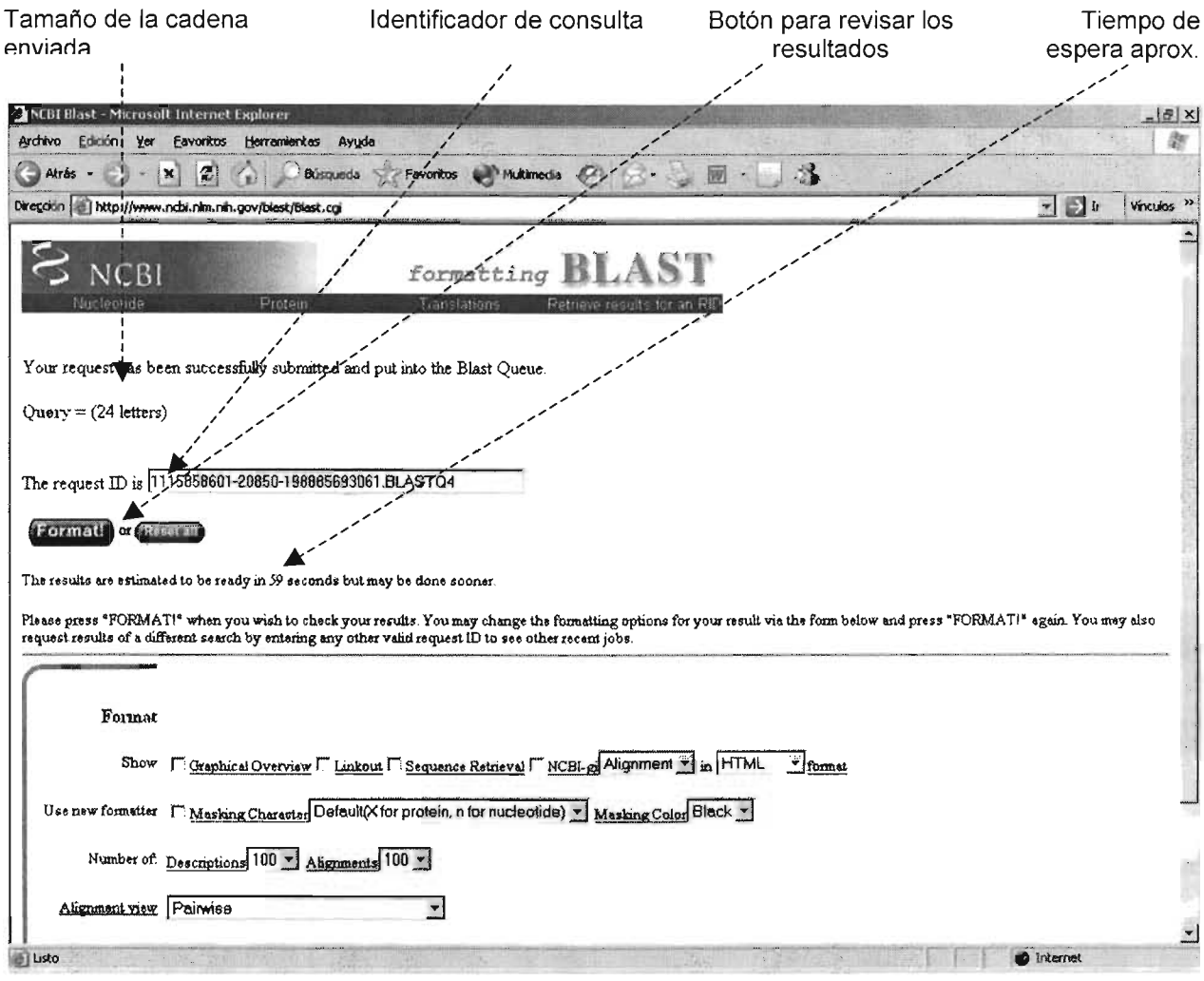
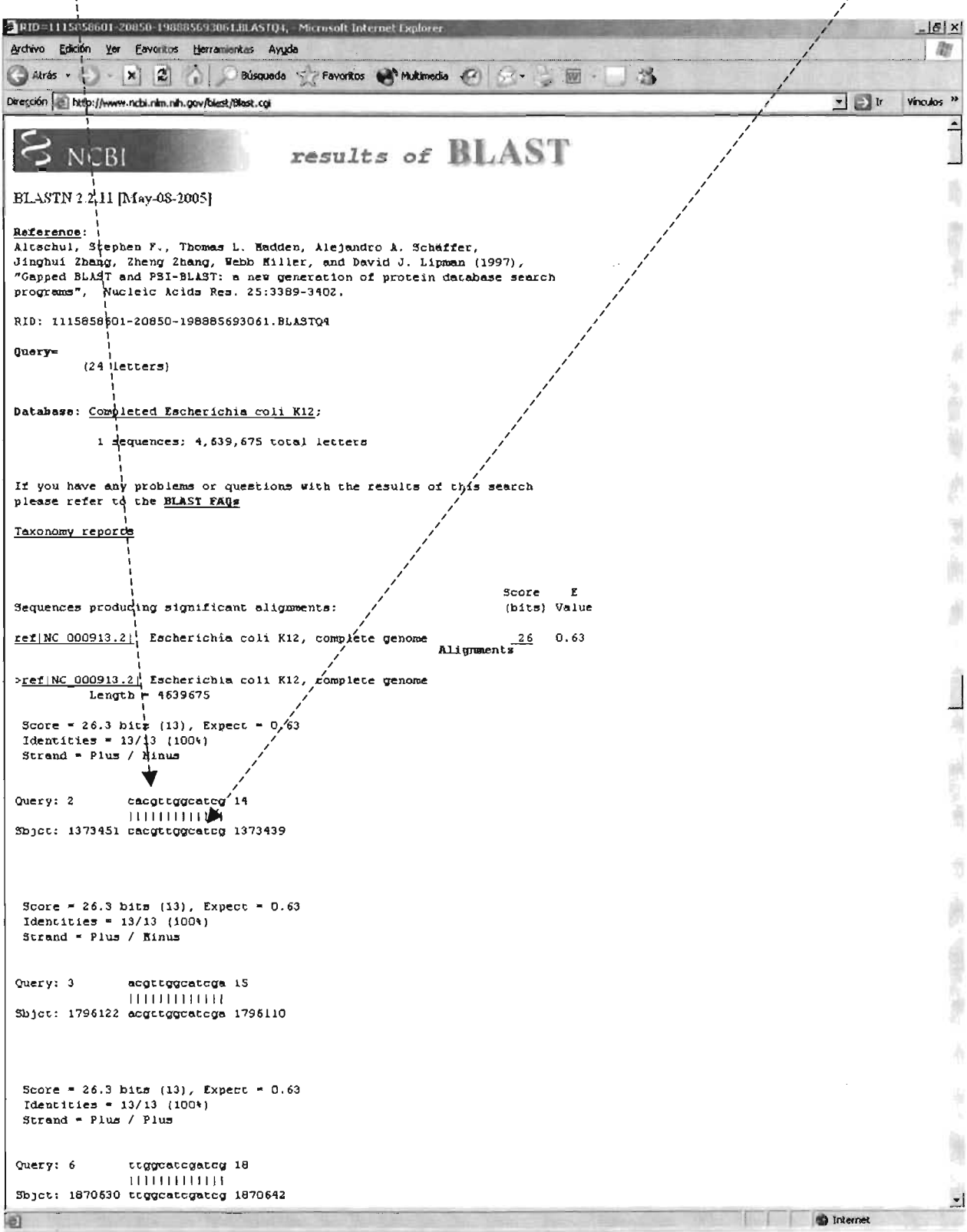


Fig. 3.3. La pequeña cadena de nucleótidos escrita en la caja de texto superior se alineará sobre el genoma del *E. coli* K 12, en aproximadamente 59 segundos

Una vez transcurrido el tiempo de procesamiento, los resultados se muestran en una ventana como la de la figura 3.4:

Subcadena alineada

Subcadena del genoma



RID=1115058601-20850-198885693061.BLASTQ4 - Microsoft Internet Explorer

Archivo Edición Ver Favoritos Herramientas Ayuda

Atrás Búsqueda Favoritos Multimedia

Dirección http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

Vínculos

NCBI results of BLAST

BLASTN 2.2.11 [May-08-2005]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

RID: 1115058601-20850-198885693061.BLASTQ4

Query=
(24 letters)

Database: Completed Escherichia coli K12;
1 sequences; 4,639,675 total letters

If you have any problems or questions with the results of this search
please refer to the BLAST FAQs

Taxonomy reports

Sequences producing significant alignments:	Score (bits)	E Value
ref|NC_000913.2| Escherichia coli K12, complete genome	26	0.63

Alignments

>ref|NC_000913.2| Escherichia coli K12, complete genome
Length = 4639675

Score = 26.3 bits (13), Expect = 0.63
Identities = 13/13 (100%)
Strand = Plus / Minus

Query: 2 cacgttggcatcg 14
||||||||||||
Sbjct: 1373451 cacgttggcatcg 1373439

Score = 26.3 bits (13), Expect = 0.63
Identities = 13/13 (100%)
Strand = Plus / Minus

Query: 3 acgttggcatcga 15
|||||||||||||
Sbjct: 1796122 acgttggcatcga 1796110

Score = 26.3 bits (13), Expect = 0.63
Identities = 13/13 (100%)
Strand = Plus / Plus

Query: 6 ttggcatcgatcg 18
|||||||||||||
Sbjct: 1870630 ttggcatcgatcg 1870642

Internet

Fig. 3.4.a Resultados de la alineación local sobre el genoma *E.coli*

Otras subcadenas alineadas

Otros resultados

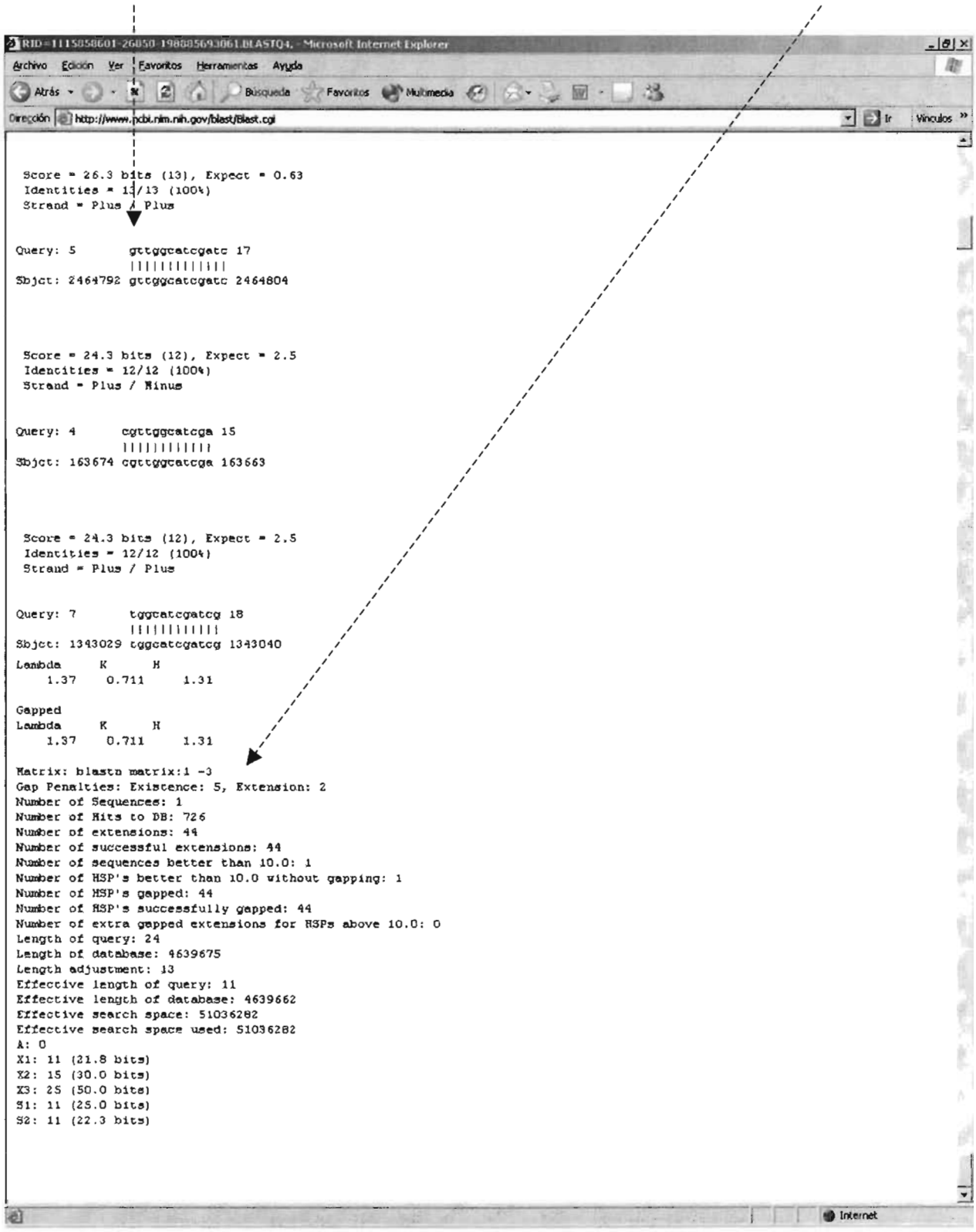


Fig. 3.4.b Continuación de resultados

Desarrollar los algoritmos de Blast para utilizarlos en un sistema local en la Facultad de Ciencias de la UNAM resultaría muy laborioso, por lo cual para dar una mejor solución en el proceso de evolución molecular, para esta tesis se adaptó el código a un sistema propio que cuenta con más funcionalidades.

3.2 Interblast

Para realizar las tareas de pre-procesamiento de datos, generación de puntos atractores, generación de puntos evolutivos, alineación local, almacenamiento y recuperación de resultados en una base de datos relacional, se desarrolló un sistema vía web llamado: Interblast, el cual representa el objeto principal de esta tesis.

Cabe señalar que todo el proceso de alineación local es posible, gracias a que se adaptó el Blast de manera local al software Interblast, con el fin particular de beneficiar el grupo de trabajo de la Facultad de Ciencias que estudia la evolución molecular.

Con la ayuda de Interblast y algunas herramientas de minería de datos, es posible trabajar para encontrar patrones significativos sobre la evolución molecular.

Para asegurar un consumo eficiente de recursos y tiempos de respuesta reducidos, la solución completa se implementó bajo un clúster o agrupación de computadoras con el sistema operativo Linux, de tal manera que se emula la operación de una supercomputadora, tal y como se muestra en la figura 3.5:

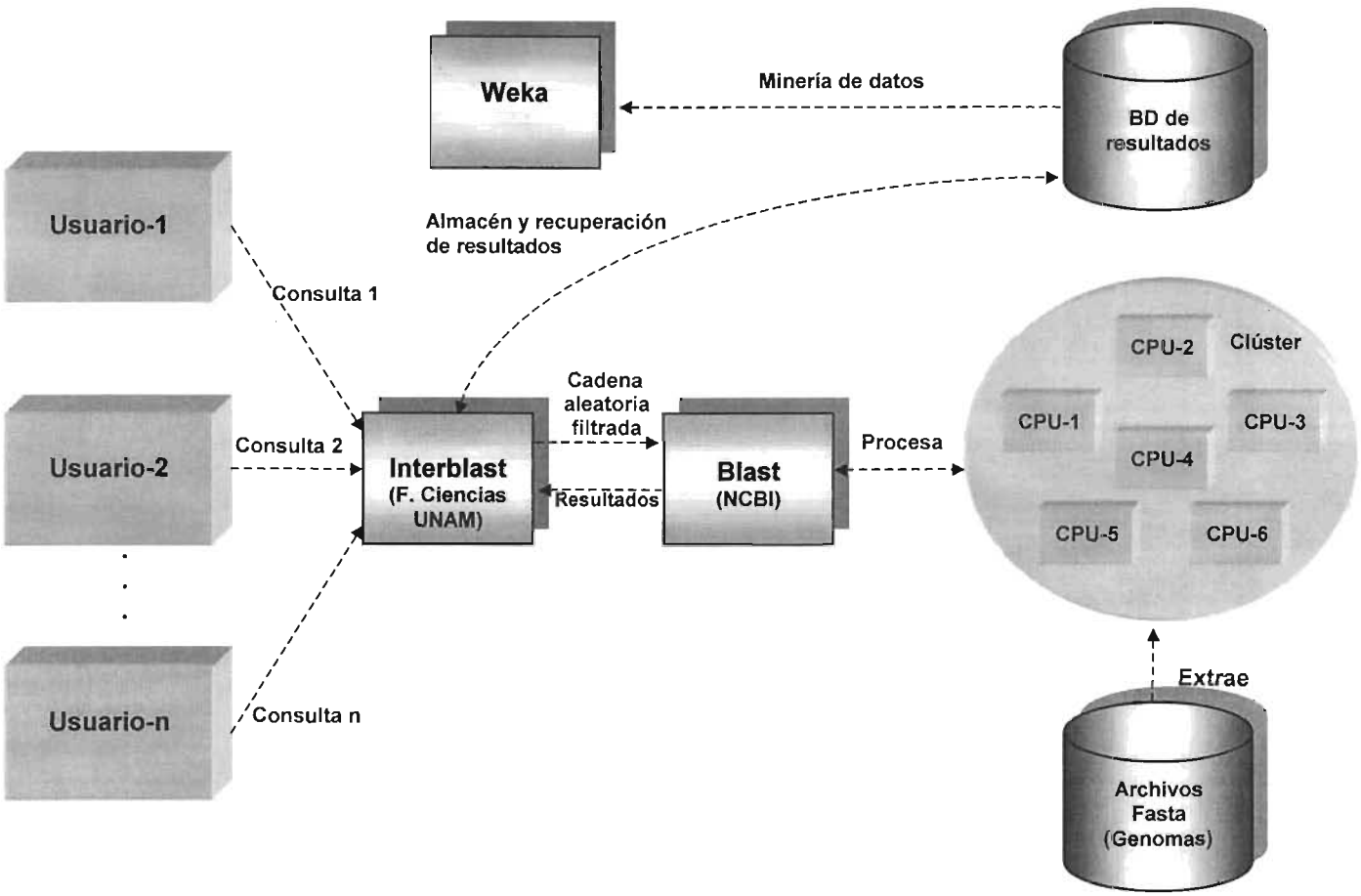


Fig. 3.5 Proceso de interacción Humano - Interblast - Blast

En el esquema anterior se puede observar que Interblast, genera de manera automática secuencias genómicas aleatorias que se filtran para ser enviadas a Blast, el cual a su vez realiza eficientemente las operaciones de alineación correspondientes con la ayuda de un clúster. Los resultados son almacenados por Interblast en una base de datos que puede ser consultada por varios usuarios de manera simultánea, y explotados con herramientas de minería de datos.

3.3 Arquitectura de Interblast

La arquitectura del software del sistema esta definida en tres capas independientes que son:

- Interfaz de usuario
- Dominio del problema
- Manejo de datos

3.3.1 Capa de la interfaz de usuario

La capa de software de Interblast encargada de la interacción humano-máquina se desarrolló para operar vía web de forma funcional.

3.3.1.1 Árbol Mad*

Antes de construir la interfaz de usuario, es necesario comprender la forma en la que el usuario realiza la tarea que se desea automatizar, y para ello se muestra en la figura 3.6 el árbol MAD* correspondiente junto con sus tarjetas:

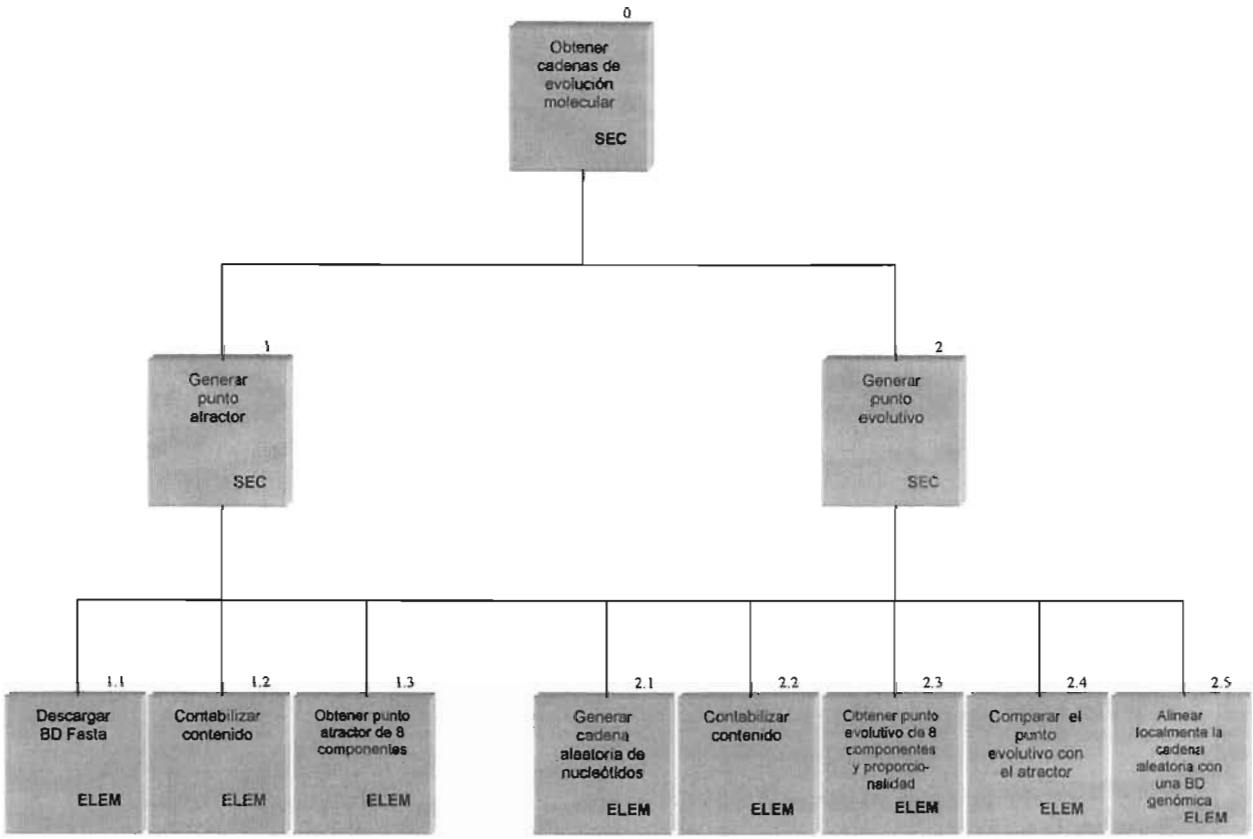


Fig. 3.6 Árbol MAD* que define el proceso de trabajo del usuario

Para que el árbol anterior esté completo, se requieren de tarjetas descriptivas, las cuales se muestran enseguida:

3.3.1.2 Tarjetas del árbol Mad*

Núcleo	
Número	0
Nombre	Obtener cadenas de evolución molecular
Objetivo	Encontrar cadenas de ADN que puedan ser analizadas para descubrir patrones evolutivos
Constructor	Secuencial
Facultativa	Falso
Interruptible	1
Prioridad	1
Operario	Investigador biólogo
Modo	Manual
Pre y Post condiciones	
Precondición de inicio	Hay bases de datos genómicas disponibles
Precondición de arranque	Ninguna
Precondición de paro	Ninguna
Postcondiciones	Cadena evolutiva generada

Núcleo	
Número	1
Nombre	Generar punto atractor
Objetivo	Obtener un punto de 8 componentes que identifique de manera única a un organismo
Constructor	Secuencial
Facultativa	Verdadero
Interruptible	1
Prioridad	1
Operario	Investigador biólogo
Modo	Manual
Pre y Post condiciones	
Precondición de inicio	Hay bases de datos genómicas disponibles
Precondición de arranque	Ninguna
Precondición de paro	Ninguna
Postcondiciones	Punto atractor generado

Núcleo	
Número	1.1
Nombre	Descargar archivo Fasta
Objetivo	Disponer de un archivo genómico en un formato que permita contabilizar el contenido de un organismo
Constructor	Elemental
Facultativa	Verdadero
Interruptible	1
Prioridad	1
Operario	Investigador biólogo
Modo	Manual
Pre y Post condiciones	
Precondición de inicio	Hay bases de datos genómicas disponibles para descargar
Precondición de arranque	Ninguna
Precondición de paro	Ninguna
Postcondiciones	Archivo Fasta almacenado

Núcleo	
Número	1.2
Nombre	Contabilizar contenido
Objetivo	Obtener las cantidades de los datos incluidos en un archivo Fasta
Constructor	Elemental
Facultativa	Verdadero
Interruptible	1
Prioridad	1
Operario	Investigador biólogo
Modo	Manual
Pre y Post condiciones	
Precondición de inicio	Hay por lo menos un archivo Fasta almacenado
Precondición de arranque	Ninguna
Precondición de paro	Ninguna
Postcondiciones	Contenido del archivo Fasta contabilizado

Núcleo	
Número	1.3
Nombre	Obtener punto atractor de 8 componentes
Objetivo	Encontrar 8 valores que identifiquen de manera única a un organismo
Constructor	Elemental
Facultativa	Verdadero
Interruptible	1
Prioridad	1
Operario	Investigador biólogo
Modo	Manual
Pre y Post condiciones	
Precondición de inicio	Contenido de archivo Fasta contabilizado
Precondición de arranque	Ninguna
Precondición de paro	Ninguna
Postcondiciones	Punto atractor de 8 componentes obtenido

Núcleo	
Número	2
Nombre	Generar punto evolutivo
Objetivo	Obtener una cadena aleatoria que pueda ser alineada localmente sobre una base de datos genómica, después de haberse obtenido un punto evolutivo de 8 componentes compatible con el punto atractor de un organismo determinado.
Constructor	Secuencial
Facultativa	Verdadero
Interruptible	1
Prioridad	1
Operario	Investigador biólogo
Modo	Manual
Pre y Post condiciones	
Precondición de inicio	Se cuenta con el punto atractor de un organismo
Precondición de arranque	Ninguna
Precondición de paro	Ninguna
Postcondiciones	Punto evolutivo generado

Núcleo	
Número	2.1
Nombre	Generar cadena aleatoria de nucleótidos
Objetivo	Obtener una cadena aleatoria de nucléotidos que pueda ser procesada para saber si puede ser alineada localmente con una BD genómica.
Constructor	Elemental
Facultativa	Verdadero
Interruptible	1
Prioridad	1
Operario	Investigador biólogo
Modo	Manual
Pre y Post condiciones	
Precondición de inicio	Se cuenta con el punto atractor de un organismo
Precondición de arranque	Ninguna
Precondición de paro	Ninguna
Postcondiciones	Cadena aleatoria generada

Núcleo	
Número	2.2
Nombre	Contabilizar contenido
Objetivo	Extraer las cifras del contenido de la cadena aleatoria
Constructor	Elemental
Facultativa	Verdadero
Interruptible	1
Prioridad	1
Operario	Investigador biólogo
Modo	Manual
Pre y Post condiciones	
Precondición de inicio	Hay cadena aleatoria
Precondición de arranque	Ninguna
Precondición de paro	Ninguna
Postcondiciones	Contenido contabilizado

Núcleo	
Número	2.3
Nombre	Obtener punto evolutivo de 8 componentes y proporcionalidad
Objetivo	Obtener 8 valores que definan una cadena aleatoria de manera única y que cumplan con cierta proporcionalidad, para decidir si deberá ser alineada con una BD genómica.
Constructor	Elemental
Facultativa	Verdadero
Interruptible	1
Prioridad	1
Operario	Investigador biólogo
Modo	Manual
Pre y Post condiciones	
Precondición de inicio	El contenido de la cadena aleatoria se ha contabilizado
Precondición de arranque	Ninguna
Precondición de paro	Ninguna
Postcondiciones	Punto evolutivo de 8 componentes y proporcionalidad obtenidos

Núcleo	
Número	2.4
Nombre	Comparar el punto evolutivo con el atractor
Objetivo	Saber si la cadena aleatoria cumple con los requisitos para ser alineada localmente con una base de datos genómica.
Constructor	Elemental
Facultativa	Verdadero
Interruptible	1
Prioridad	1
Operario	Investigador biólogo
Modo	Manual
Pre y Post condiciones	
Precondición de inicio	Hay punto evolutivo y atractor
Precondición de arranque	Ninguna
Precondición de paro	Ninguna
Postcondiciones	Punto evolutivo y atractor comparados

Núcleo	
Número	2.5
Nombre	Alinear localmente la cadena aleatoria con un aBD genómica
Objetivo	Obtener subcadenas de la cadena aleatoria que se encuentren dentro de la BD genómica de un organismo
Constructor	Elemental
Facultativa	Verdadero
Interruptible	1
Prioridad	1
Operario	Investigador biólogo
Modo	Automático
Pre y Post condiciones	
Precondición de inicio	Hay cadena aleatoria, punto evolutivo y atractor.
Precondición de arranque	Ninguna
Precondición de paro	Ninguna
Postcondiciones	Cadena aleatoria alineada localmente con una BD genómica

3.3.1.3 Diagrama general de casos de uso

Una vez que contamos con la lógica de trabajo del usuario, es necesario crear un diagrama que demuestre la forma en la que ese mismo trabajo se realizará pero a través de la interfaz de usuario de Interblast, para ello en la figura 3.7 se incluye el diagrama general de casos de uso:

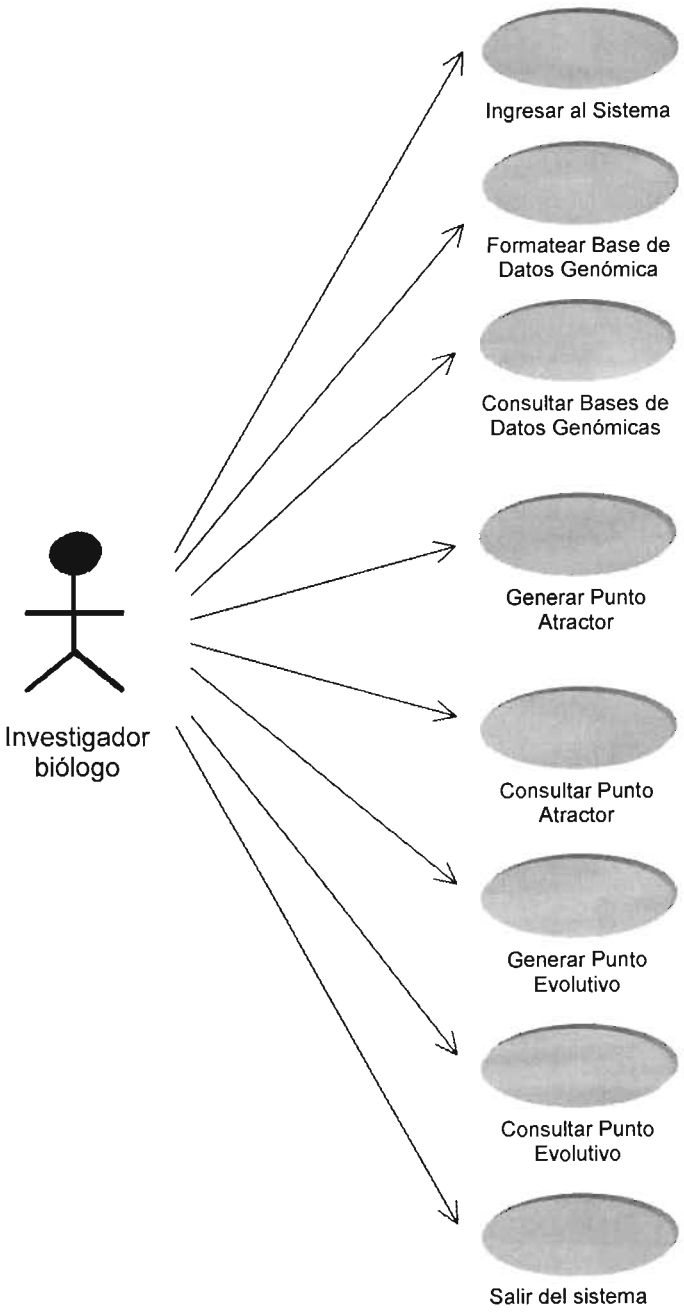
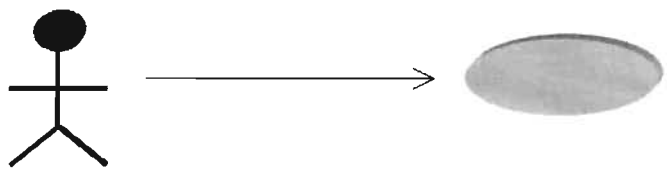


Fig. 3.7 Diagrama general de casos de uso

3.3.1.4 Detalle de los casos de uso

Caso de uso: ingresar al sistema

Actor: Investigador biólogo.



Descripción: el Investigador biólogo ingresa al sistema sin necesidad de pedirle contraseña.

Flujo:

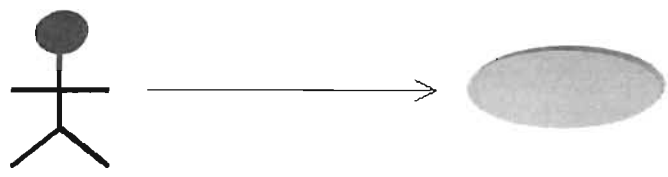
Actor		Sistema		
Paso	**Acción**	**Paso**	**Acción**	**Excepción**
1	Ingresar la dirección del sitio que aloja el sistema	2	Mostrar la página de inicio y dar la bienvenida	

Excepciones:

Excepción	Nombre	Acción

Caso de uso: formatear archivo genómico

Actor: Investigador biólogo.



Descripción: el Investigador biólogo selecciona un archivo genómico válido para que se genere un conjunto de archivos útiles para el sistema.

Flujo:

Actor		Sistema		
Paso	**Acción**	**Paso**	**Acción**	**Excepción**
1	Seleccionar un archivo genómico válido y lo envía.	2	Procesar el archivo Fasta, generar los archivos índices y dar de alta el nombre del genoma dentro de la Base de Datos.	E_1 y E_2
		3	Informar que el archivo se ha formateado.	

Excepciones:

Excepción	Nombre	Acción
E_1	No seleccionar un archivo	Avisar al usuario que para procesar un archivo primero se debe seleccionar éste.

E2	Seleccionar un archivo no válido	Indicar que el archivo seleccionado no es válido, por lo que es necesario utilizar alguno que si lo sea.

Caso de uso: consultar bases de datos genómicas

Actor: Investigador biólogo.

Investigador biólogo

Consultar Bases de Datos Genómicas

Descripción: el usuario consulta las bases de datos genómicas disponibles.

Flujo:

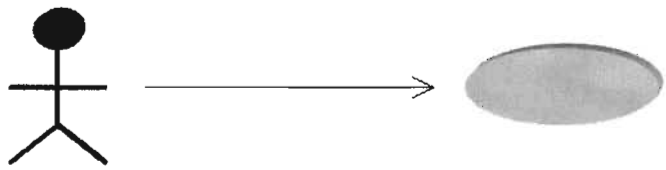
Actor		Sistema		
Paso	Acción	Paso	Acción	Excepción
1	Consultar las bases de datos genómicas que puede utilizar.	2	Mostrar una lista de bases de datos genómicas disponibles para su uso.	

Excepciones:

Excepción	Nombre	Acción

Caso de uso: generar punto atractor.

Actor: Investigador biólogo.



Investigador biólogo

Generar Punto Atractor

Descripción: el usuario selecciona una base de datos genómica para que el sistema genere el punto atractor correspondiente.

Flujo:

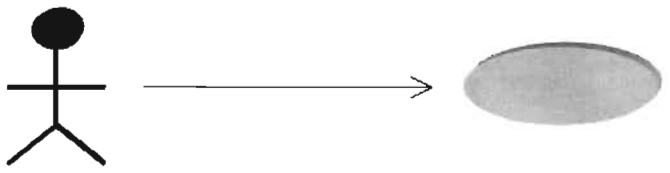
Actor		Sistema		
Paso	**Acción**	**Paso**	**Acción**	**Excepción**
1	Seleccionar la Base de Datos genómica de un organismo.	2	Generar el punto atractor del organismo seleccionado.	
		3	Mostrar una pantalla en la que señala que el punto atractor se ha generado	

Excepciones:

Excepción	Nombre	Acción

Caso de uso: consultar punto atractor

Actor: Investigador biólogo.



Investigador biólogo

Consultar Punto Atractor

Descripción: el investigador consulta los puntos atractores apartir de los cuales podrá generar puntos evolutivos.

Flujo:

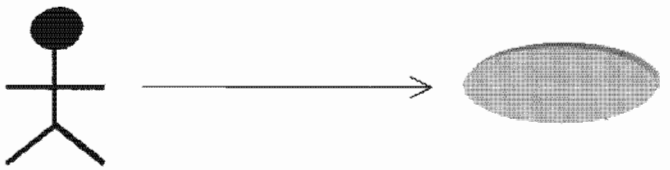
Actor		Sistema		
Paso	**Acción**	**Paso**	**Acción**	**Excepción**
1	Consultar los puntos atractores generados.	2	Mostrar todos los puntos atractores que sirven para generar puntos evolutivos.	

Excepciones:

Excepción	Nombre	Acción

Caso de uso: generar punto evolutivo

Actor: Investigador biólogo.



Descripción: el investigador genera puntos evolutivos sobre un determinado organismo que ya cuenta con un punto atractor,

Flujo:

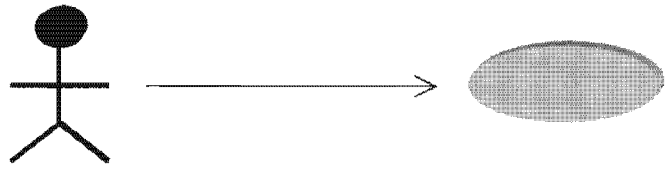
Actor		Sistema		
Paso	Acción	Paso	Acción	Excepción
1	Seleccionar una base de datos genómica a utilizar y el total de puntos evolutivos a generar.	2	Generar todos los puntos evolutivos solicitados.	E_1, E_2 y E_3

Excepciones:

Excepción	Nombre	Acción
E_1	Generar puntos sin ninguna base de datos.	Indicar que antes de generar puntos evolutivos, es necesario contar con puntos atractores.
E_2	No se indicó el total de puntos a generar	Señalar que se debe de especificar la cantidad de puntos a generar.
E_3	El dato ingresado no es numérico.	Solicitar cantidades válidas.

Caso de uso: consultar punto evolutivo.

Actor: Investigador biólogo.



Descripción: el investigador consulta los puntos evolutivos que se han generado sobre un determinado organismo.

Flujo:

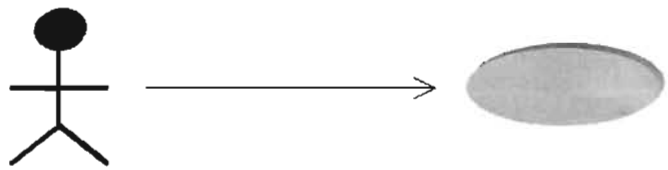
Actor		Sistema		
Paso	Acción	Paso	Acción	Excepción
1	Consultar una cierta cantidad de puntos evolutivos de un cierto organismo.	2	Mostrar el total de puntos evolutivos solicitados.	E_1 y E_2

Excepciones:

Excepción	Nombre	Acción
E_1	No se indicó el total de puntos a consultar	Señalar que se debe de especificar la cantidad de puntos a consultar.
E_2	El dato ingresado no es numérico.	Solicitar cantidades válidas.

Caso de uso: salir del sistema

Actor: Investigador biólogo.



Descripción: el usuario finaliza su sesión con el sistema.

Flujo:

Actor		Sistema		
Paso	**Acción**	**Paso**	**Acción**	**Excepción**
1	Cerrar el sistema.	2	Cerrar las conexiones y muestra un mensaje de finalización de sesión.	E_1

Excepciones:

Excepción	Nombre	Acción
E_1	Abandonar el sistema sin cerrarlo	El sistema finaliza por si sólo la sesión después de un determinado tiempo sin uso.

Apartir de los casos de uso es posible desarrolla la interfaz de usuario y el sistema.

3.3.2 Capa de dominio del problema (diseño de clases)

La capa de dominio del problema o de codificación del software del sistema, se sustenta bajo el lenguaje de programación Java [5], para asegurar la portabilidad entre plataformas.

En esta tesis no se incluye código fuente, solamente los diagramas de clases correspondientes.

El diagrama de clases que define la operación del sistema Interblast es el siguiente:

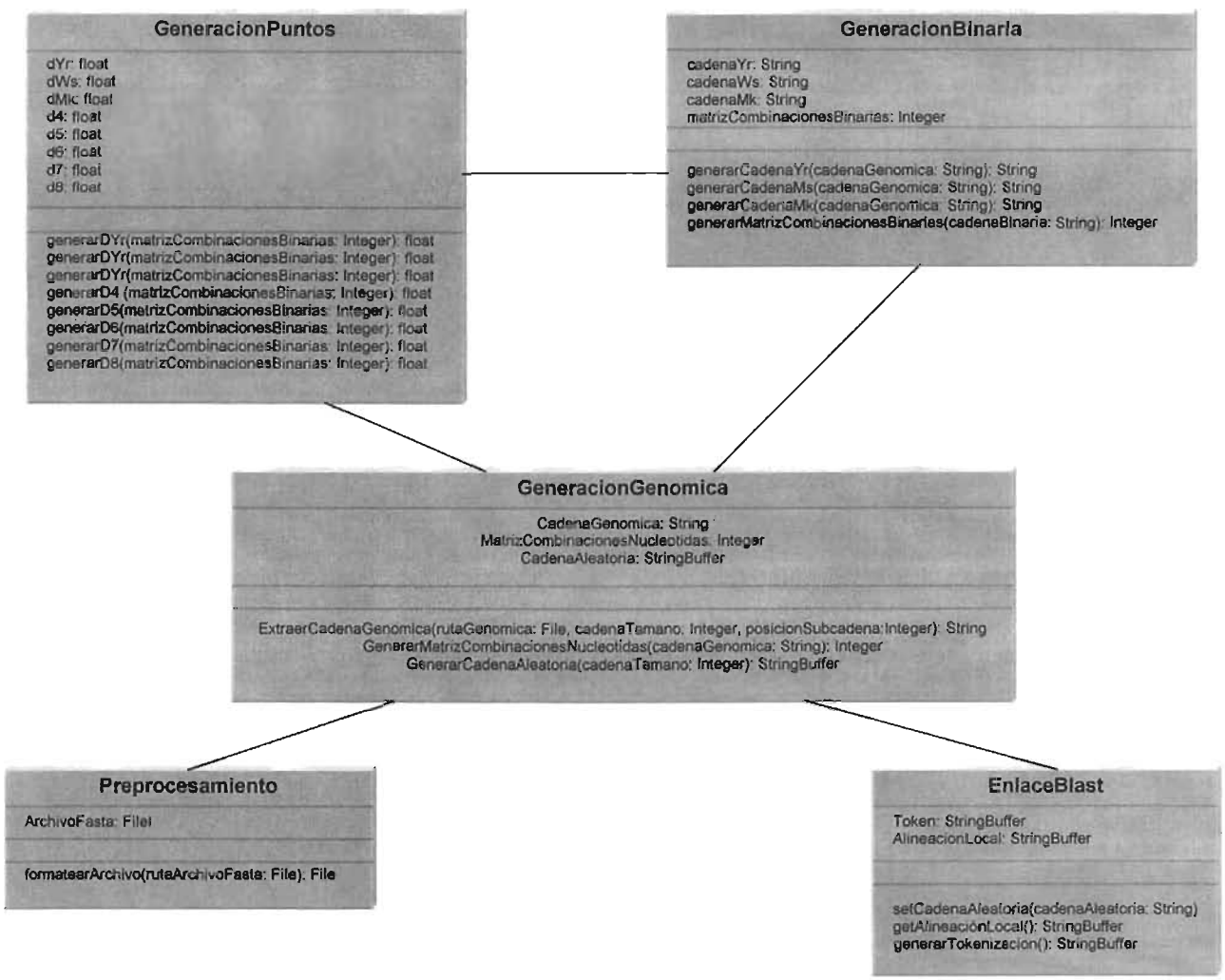


Fig. 3.8 Diagrama de clases para el funcionamiento de Interblast.

La explicación a las clases mostradas es:

- La función de la clase "Preprocesamiento" es la formatear un archivo Fasta, es decir se encarga de generar ciertos archivos índice que necesita el Blast adaptado localmente para realizar las alineaciones.

- La función de la clase "Generación Genómica" es la de extraer subcadenas de los archivos Fasta, generar cadenas aleatorias y obtener las matrices de cantidades de nucleótidos correspondientes.

- La clase "Generación Binaria" será la responsable de transformar las cadenas genómicas en sus respectivas cadenas binarias (YR, WS y MK), además de producir sus matrices de cantidades.

- La clase "Generación de Puntos" tiene como objetivo el aplicar los ocho modelos matemáticos necesarios para la construcción de puntos atractores y evolutivos.

- La clase "Enlace Blast" es la responsable de enviar las cadenas aleatorias hacia Blast para que sean alineadas con las bases de datos y genómicas, y también se encarga de recuperar los resultados.

Además de todas estas clases, se requiere de otras clases que permitan almacenar y recuperar los datos de la base de datos. Véase la figura 3.8:

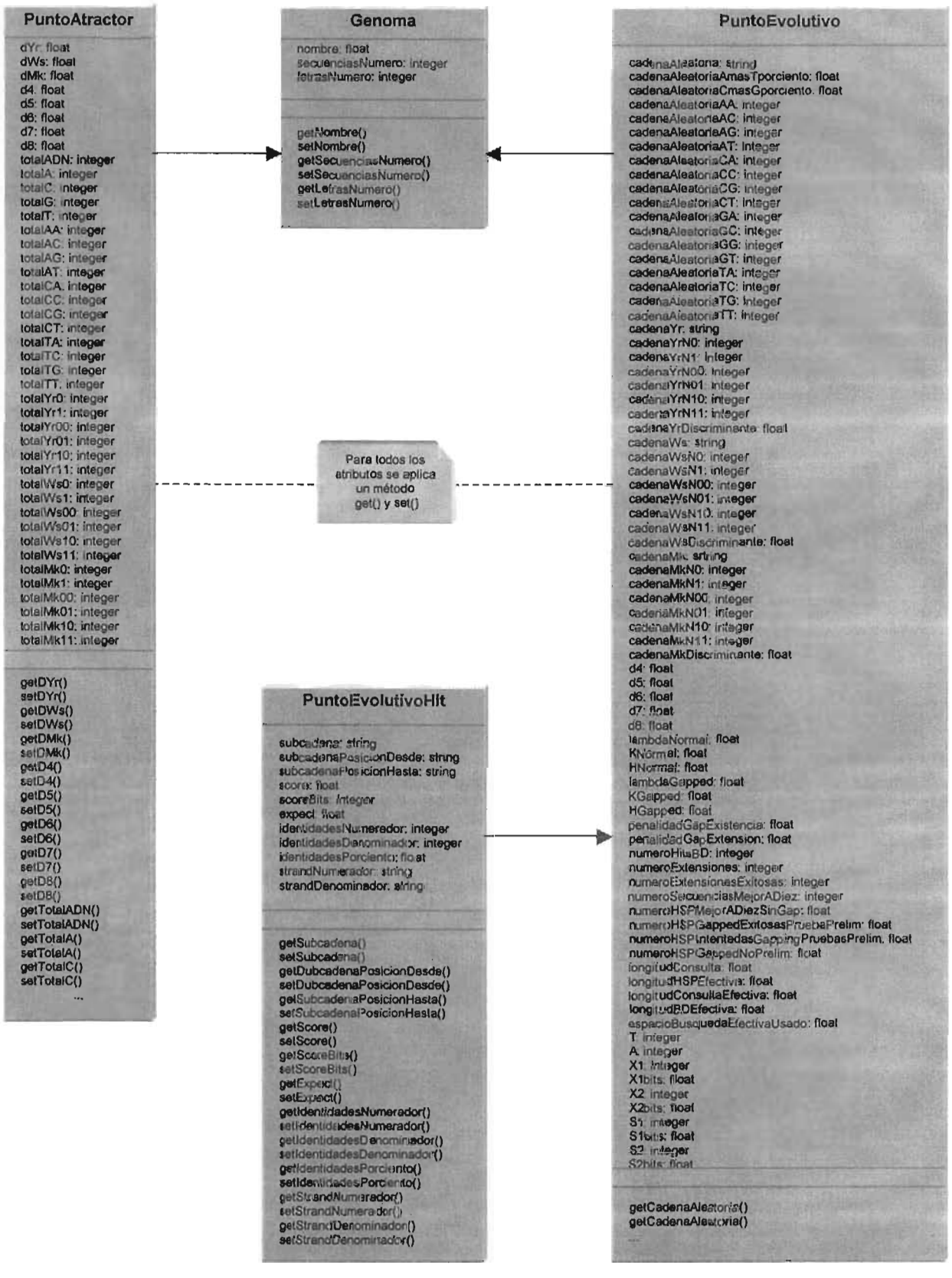


Fig. 3.9 Diagrama de clases para la inserción y recuperación de datos en la BD.

Las clases de la figura anterior son las responsables de almacenar y recuperar todos los datos que son parte de los procesos de alineación.

3.3.3 Capa del manejo de datos

Para almacenar y consultar los resultados de los puntos atractores, puntos evolutivos y alineaciones locales, se creó una base de datos relacional con el nombre de Interblast, cuyo diseño general se muestra en la figura 3.10.

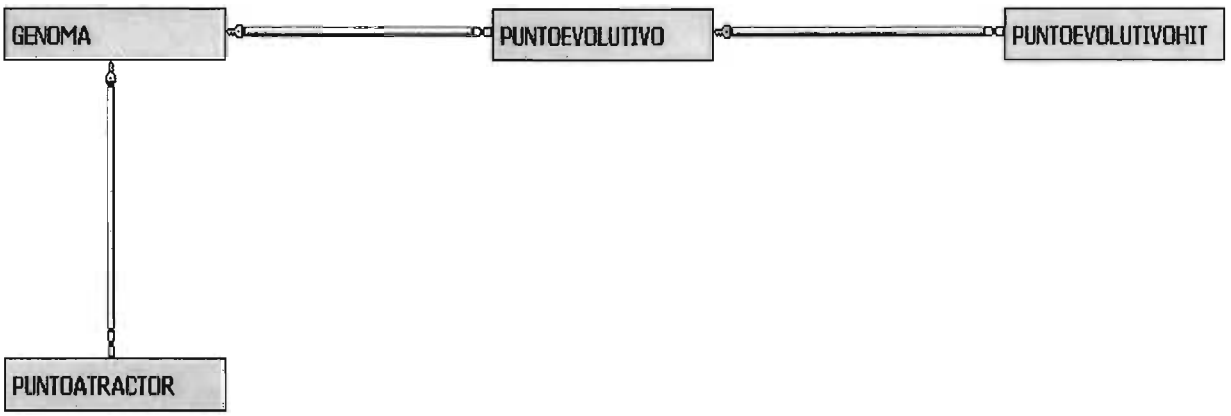


Fig. 3.10 Diseño general de la BD Interblast

En la página siguiente se incluye el diagrama entidad-relación detallado de la BD Interblast:

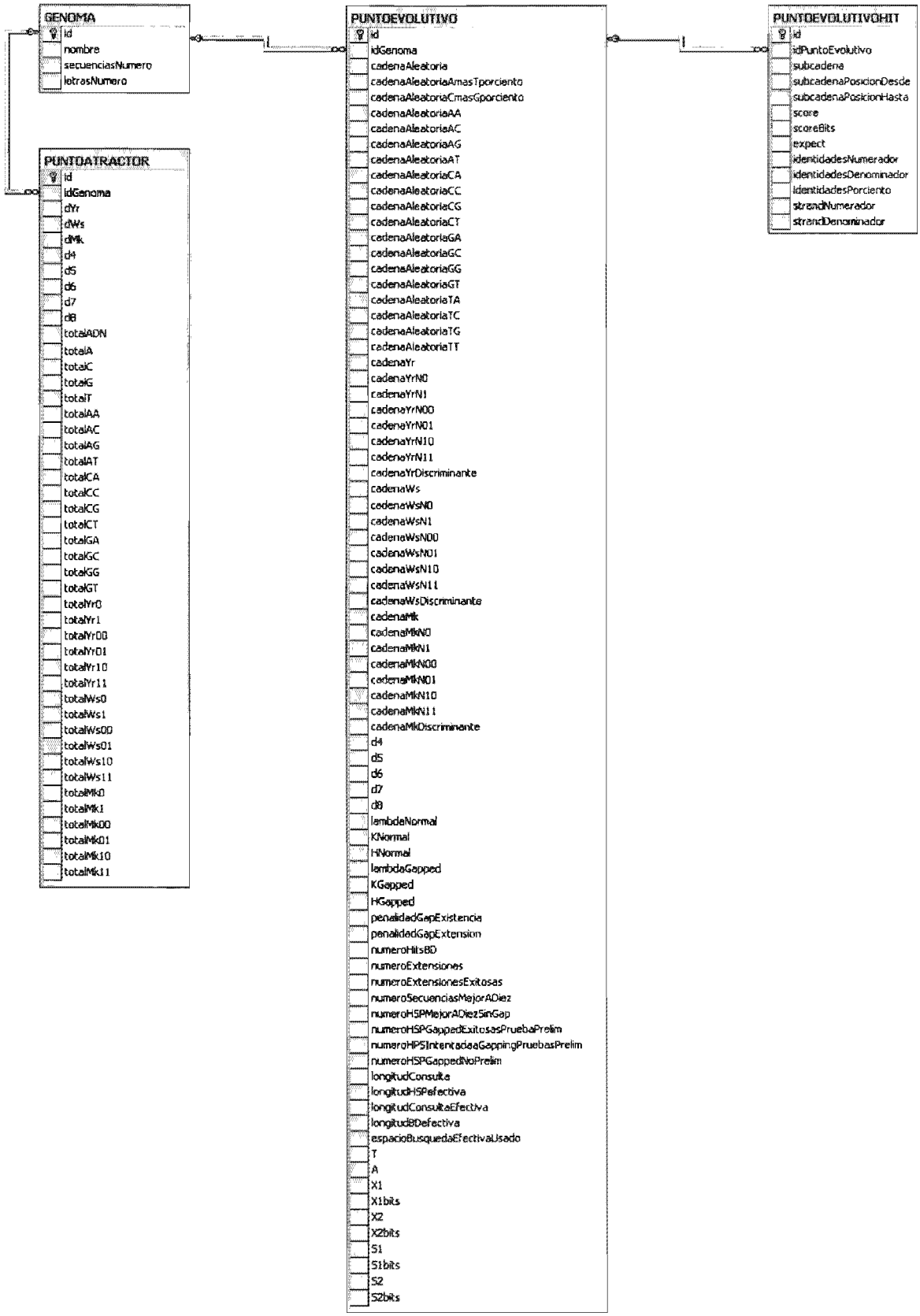
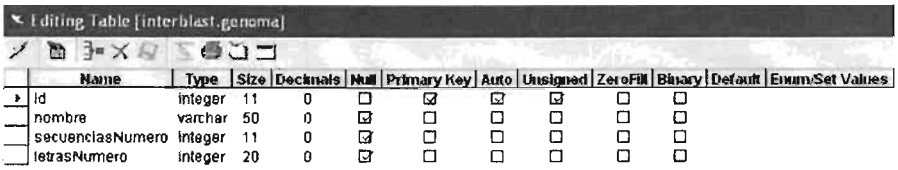


Fig. 3.11 Diagrama E-R de Interblast

3.3.3.1 Estructura de la base de datos y diccionario de datos

En primera instancia es necesaria una tabla para almacenar datos básicos de los genomas, como la siguiente


Editing Table [interblast.genoma]

Name	Type	Size	Decimals	Null	Primary Key	Auto	Unsigned	ZeroFill	Binary	Default	Enum/Set Values
id	integer	11	0	☐	☑	☑	☑	☐	☐		
nombre	varchar	50	0	☑	☐	☐	☐	☐	☐		
secuenciasNumero	integer	11	0	☑	☐	☐	☐	☐	☐		
letrasNumero	integer	20	0	☑	☐	☐	☐	☐	☐		

Tabla 3.1 La tabla "Genoma"

En la tabla genoma se guardan datos de los organismos como:

- id: es un identificador numérico para cada organismo.
- nombre: es el nombre del organismo.
- secuenciasNúmero: es el total de bloques que contiene la base de datos genómica.
- letrasNúmero: es el total de nucleótidos contenidos en el ADN del organismo.

En la tabla 3.2 que nombrada "puntoAtractor", se almacenan los puntos atractores de cada una de las bases de datos genómicas:

Editing Table [interblast.puntoatractor]

Name	Type	Size	Decimals	Null	Primary Key	Auto	Unsigned	ZeroFill	Binary	Default	Enum/Set Values
Id	integer	11	0	☐	☑	☑	☐	☐	☐		
idGenoma	integer	11	0	☑	☐	☐	☐	☐	☐		
dYr	float	8	5	☑	☐	☐	☐	☐	☐		
dWs	float	8	5	☑	☐	☐	☐	☐	☐		
dMk	float	8	5	☑	☐	☐	☐	☐	☐		
d4	float	8	5	☑	☐	☐	☐	☐	☐		
d5	float	8	5	☑	☐	☐	☐	☐	☐		
d6	float	8	5	☑	☐	☐	☐	☐	☐		
d7	float	8	5	☑	☐	☐	☐	☐	☐		
d8	float	8	5	☑	☐	☐	☐	☐	☐		
totalADN	integer	9	0	☑	☐	☐	☐	☐	☐		
totalA	integer	9	0	☑	☐	☐	☐	☐	☐		
totalC	integer	9	0	☑	☐	☐	☐	☐	☐		
totalG	integer	9	0	☑	☐	☐	☐	☐	☐		
totalT	integer	9	0	☑	☐	☐	☐	☐	☐		
totalAA	integer	9	0	☑	☐	☐	☐	☐	☐		
totalAC	integer	9	0	☑	☐	☐	☐	☐	☐		
totalAG	integer	9	0	☑	☐	☐	☐	☐	☐		
totalAT	integer	9	0	☑	☐	☐	☐	☐	☐		
totalCA	integer	9	0	☑	☐	☐	☐	☐	☐		
totalCC	integer	9	0	☑	☐	☐	☐	☐	☐		
totalCG	integer	9	0	☑	☐	☐	☐	☐	☐		
totalCT	integer	9	0	☑	☐	☐	☐	☐	☐		
totalGA	integer	9	0	☑	☐	☐	☐	☐	☐		
totalGC	integer	9	0	☑	☐	☐	☐	☐	☐		
totalGG	integer	9	0	☑	☐	☐	☐	☐	☐		
totalGT	integer	9	0	☑	☐	☐	☐	☐	☐		
totalTA	integer	9	0	☑	☐	☐	☐	☐	☐		
totalTC	integer	9	0	☑	☐	☐	☐	☐	☐		
totalTG	integer	9	0	☑	☐	☐	☐	☐	☐		
totalTT	integer	9	0	☑	☐	☐	☐	☐	☐		
totalYr0	integer	9	0	☑	☐	☐	☐	☐	☐		
totalYr1	integer	9	0	☑	☐	☐	☐	☐	☐		
totalYr00	integer	9	0	☑	☐	☐	☐	☐	☐		
totalYr01	integer	9	0	☑	☐	☐	☐	☐	☐		
totalYr10	integer	9	0	☑	☐	☐	☐	☐	☐		
totalYr11	integer	9	0	☑	☐	☐	☐	☐	☐		
totalWs0	integer	9	0	☑	☐	☐	☐	☐	☐		
totalWs1	integer	9	0	☑	☐	☐	☐	☐	☐		
totalWs00	integer	9	0	☑	☐	☐	☐	☐	☐		
totalWs01	integer	9	0	☑	☐	☐	☐	☐	☐		
totalWs10	integer	9	0	☑	☐	☐	☐	☐	☐		
totalWs11	integer	9	0	☑	☐	☐	☐	☐	☐		
totalMk0	integer	9	0	☑	☐	☐	☐	☐	☐		
totalMk1	integer	9	0	☑	☐	☐	☐	☐	☐		
totalMk00	integer	9	0	☑	☐	☐	☐	☐	☐		
totalMk01	integer	9	0	☑	☐	☐	☐	☐	☐		
totalMk10	integer	9	0	☑	☐	☐	☐	☐	☐		
totalMk11	integer	9	0	☑	☐	☐	☐	☐	☐		

Tabla 3.2 Espacio para almacenar puntos atractores

En la tabla de puntos atractores es posible almacenar los siguientes datos:

- id: identificador auto-numérico del punto atractor.

- idGenoma: valor que relaciona esta tabla con la tabla genoma.
- dYR, WS, dMK, d4, d5, d6, d7 y d8: guardan los valores obtenidos al aplicar los ocho modelos matemáticos.
- totalA, totalC, totalAA, totalAC,...,: alojan los valores de la matriz de cantidades de nucleótidos.
- totalYR0, totalYR1, totalYR00,...,: incluyen los valores de la matriz de cantidades binarias.

La siguiente tabla (puntoEvolutivo) incluye los puntos evolutivos que han atravesado todos los filtros y que cumplen con las reglas de cercanía al punto atractor correspondiente, junto con ciertos resultados de alineación:

Editing Table [interblast.puntoevolutivo]

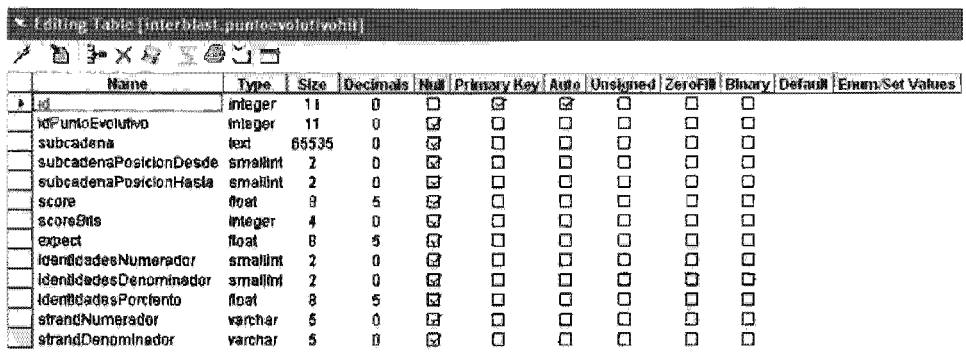
Name	Type	Size	Decimals	Null	Primary Key	Auto	Unsigned	ZeroFill	Binary	Default	Enum/Set Values
id	integer	11	0	☐	☑	☑	☐	☐	☐		
idGenoma	integer	11	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoria	text	65535	0	☑	☐	☐	☐	☐	☐		
cadenaAlestoriaAmasTporciento	float	8	5	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaCmasGporciento	float	8	5	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaAA	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaAC	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaAG	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaAT	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaCA	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaCC	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaCG	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaCT	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaGA	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaGC	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaGG	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaGT	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaTA	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaTC	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaTG	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaAleatoriaTT	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaYr	text	65535	0	☑	☐	☐	☐	☐	☐		
cadenaYrN0	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaYrN1	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaYrN00	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaYrN01	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaYrN10	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaYrN11	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaYrDiscriminante	float	8	5	☑	☐	☐	☐	☐	☐		
cadenaWs	text	65535	0	☑	☐	☐	☐	☐	☐		
cadenaWsN0	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaWsN1	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaWsN00	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaWsN01	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaWsN10	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaWsN11	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaWsDiscriminante	float	8	5	☑	☐	☐	☐	☐	☐		
cadenaMk	text	65535	0	☑	☐	☐	☐	☐	☐		
cadenaMkN0	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaMkN1	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaMkN00	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaMkN01	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaMkN10	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaMkN11	integer	4	0	☑	☐	☐	☐	☐	☐		
cadenaMkDiscriminante	float	8	5	☑	☐	☐	☐	☐	☐		
d4	float	8	5	☑	☐	☐	☐	☐	☐		
d5	float	8	5	☑	☐	☐	☐	☐	☐		
d6	float	8	5	☑	☐	☐	☐	☐	☐		
d7	float	8	5	☑	☐	☐	☐	☐	☐		
d8	float	8	5	☑	☐	☐	☐	☐	☐		
lambdaNormal	float	8	5	☑	☐	☐	☐	☐	☐		
KNormal	float	8	5	☑	☐	☐	☐	☐	☐		
HNormal	float	8	5	☑	☐	☐	☐	☐	☐		
lambdaGapped	float	8	5	☑	☐	☐	☐	☐	☐		
KGapped	float	8	5	☑	☐	☐	☐	☐	☐		
HGapped	float	8	5	☑	☐	☐	☐	☐	☐		
penalidadGapExistencia	smallin	2	0	☑	☐	☐	☐	☐	☐		
penalidadGapExtension	smallin	2	0	☑	☐	☐	☐	☐	☐		
numeroHitsBD	integer	4	0	☑	☐	☐	☐	☐	☐		
numeroExtensiones	integer	4	0	☑	☐	☐	☐	☐	☐		
numeroExtensionesExitosas	integer	4	0	☑	☐	☐	☐	☐	☐		
numeroSecuenciasMejorADiez	integer	4	0	☑	☐	☐	☐	☐	☐		
numeroHSPMejorADiezSinGap	integer	4	0	☑	☐	☐	☐	☐	☐		
numeroHSPGappedExitosasPruebaPrelim	integer	4	0	☑	☐	☐	☐	☐	☐		
numeroHSPIntentadasGappingPruebasPrelim	integer	4	0	☑	☐	☐	☐	☐	☐		
numeroHSPGeppedNoPrelim	integer	4	0	☑	☐	☐	☐	☐	☐		
longitudConsulta	integer	4	0	☑	☐	☐	☐	☐	☐		
longitudHSPEfectiva	integer	4	0	☑	☐	☐	☐	☐	☐		
longitudConsultaEfectiva	integer	4	0	☑	☐	☐	☐	☐	☐		
longitudBDEfectiva	integer	9	0	☑	☐	☐	☐	☐	☐		
espacioBusquedaEfectivaUsado	bigint	20	0	☑	☐	☐	☐	☐	☐		
T	smallin	6	0	☑	☐	☐	☐	☐	☐		
A	smallin	6	0	☑	☐	☐	☐	☐	☐		
X1	smallin	6	0	☑	☐	☐	☐	☐	☐		
X1bits	float	8	5	☑	☐	☐	☐	☐	☐		
X2	smallin	6	0	☑	☐	☐	☐	☐	☐		
X2bits	float	8	5	☑	☐	☐	☐	☐	☐		
S1	smallin	6	0	☑	☐	☐	☐	☐	☐		
S1bits	float	8	5	☑	☐	☐	☐	☐	☐		
S2	smallin	6	0	☑	☐	☐	☐	☐	☐		
S2bits	float	8	5	☑	☐	☐	☐	☐	☐		

Tabla 3.3 En esta tabla se alojan valores de los puntos evolutivos y de alineación

Los valores que puede almacenar la tabla 3.3 son:

- id: es un identificador auto-numérico de cada punto evolutivo.
- idGenoma: campo que relaciona el punto evolutivo con el genoma sobre el cual se evaluó.
- cadenaAleatoria: guarda la cadena generada de manera aleatoria.
- cadenaAleatoriaAmasTporciento: guarda la proporcionalidad de A+T nucleótidos en la cadena aleatoria,
- cadenaAleatoriaCmasGporciento: guarda la proporcionalidad de C+G nucleótidos en la cadena aleatoria.
- cadenaAleatoriaAA, cadenaAleatoriaAC,...: almacena la matriz de cantidades de nucleótidos contenidos en la cadena aleatoria.
- cadenaYR, cadenaWSy cadenaWS: guardan las cadenas binarias obtenidas de las aleatorias.
- cadenaYRN0, cadenaYRN1,...,: incluyen los valores de la matriz de cantidades binarias.
- cadenaYRDiscriminante,cadenaWSDiscriminante, cadenaMKDiscriminante, d4, d5, d6, d7 y d8: guardan los valores obtenidos al aplicar los ocho modelos matemáticos.
- numeroExtensionesExitosas: total de subcandenas de la cadena aleatoria, que alinean en contra de la base de datos del organismo en cuestión.
- lambdaNormal, KNormal, ..., S2bits: valores de interpretación que escapan al entendimiento de esta tesis, pero que sirven para futuros trabajos.

Finalmente la tabla puntoEvolutivoHit, guarda las subcadenas de los puntos evolutivos que tienen alineación local significativa:


Editing Table [interblast.puntoevolutivohit]

Name	Type	Size	Decimals	Null	Primary Key	Auto	Unsigned	ZeroFill	Binary	Default	Enum/Set Values
id	integer	11	0	☐	☑	☑	☐	☐	☐		
idPuntoEvolutivo	integer	11	0	☑	☐	☐	☐	☐	☐		
subcadena	text	65535	0	☑	☐	☐	☐	☐	☐		
subcadenaPosicionDesde	smallint	2	0	☑	☐	☐	☐	☐	☐		
subcadenaPosicionHasta	smallint	2	0	☑	☐	☐	☐	☐	☐		
score	float	8	5	☑	☐	☐	☐	☐	☐		
scoreBits	integer	4	0	☑	☐	☐	☐	☐	☐		
expect	float	8	5	☑	☐	☐	☐	☐	☐		
identidadesNumerador	smallint	2	0	☑	☐	☐	☐	☐	☐		
identidadesDenominador	smallint	2	0	☑	☐	☐	☐	☐	☐		
identidadesPorciento	float	8	5	☑	☐	☐	☐	☐	☐		
strandNumerador	varchar	5	0	☑	☐	☐	☐	☐	☐		
strandDenominador	varchar	5	0	☑	☐	☐	☐	☐	☐		

Tabla 3.4 Incluye espacio para guardar alineaciones de subcadenas

Los campos de la tabla anterior, guardan los siguientes valores:

- id: es un identificador auto-numérico de cada subcadena alineada.
- idPuntoEvolutivo: campo que relaciona la subcadena con su cadena aleatoria.
- subcadena: almacena la subcadena alineada exitosamente.
- subcadenaPosicionHasta: guardan la posición de inicio y término de alineación de la subcadena.
- subcadenaPosicionDesde:
- score, scoreBits, ..., strandDenominador: valores de interpretación que escapan al entendimiento de esta tesis, pero que sirven para futuros trabajos.

3.4 Interblast en funcionamiento

El sistema final se visualiza a través de un navegador web como: Internet Explorer, Netscape Navigator, Mozilla, entre otros.

La figura 3.12 se muestra la primer ventana de Interblast:

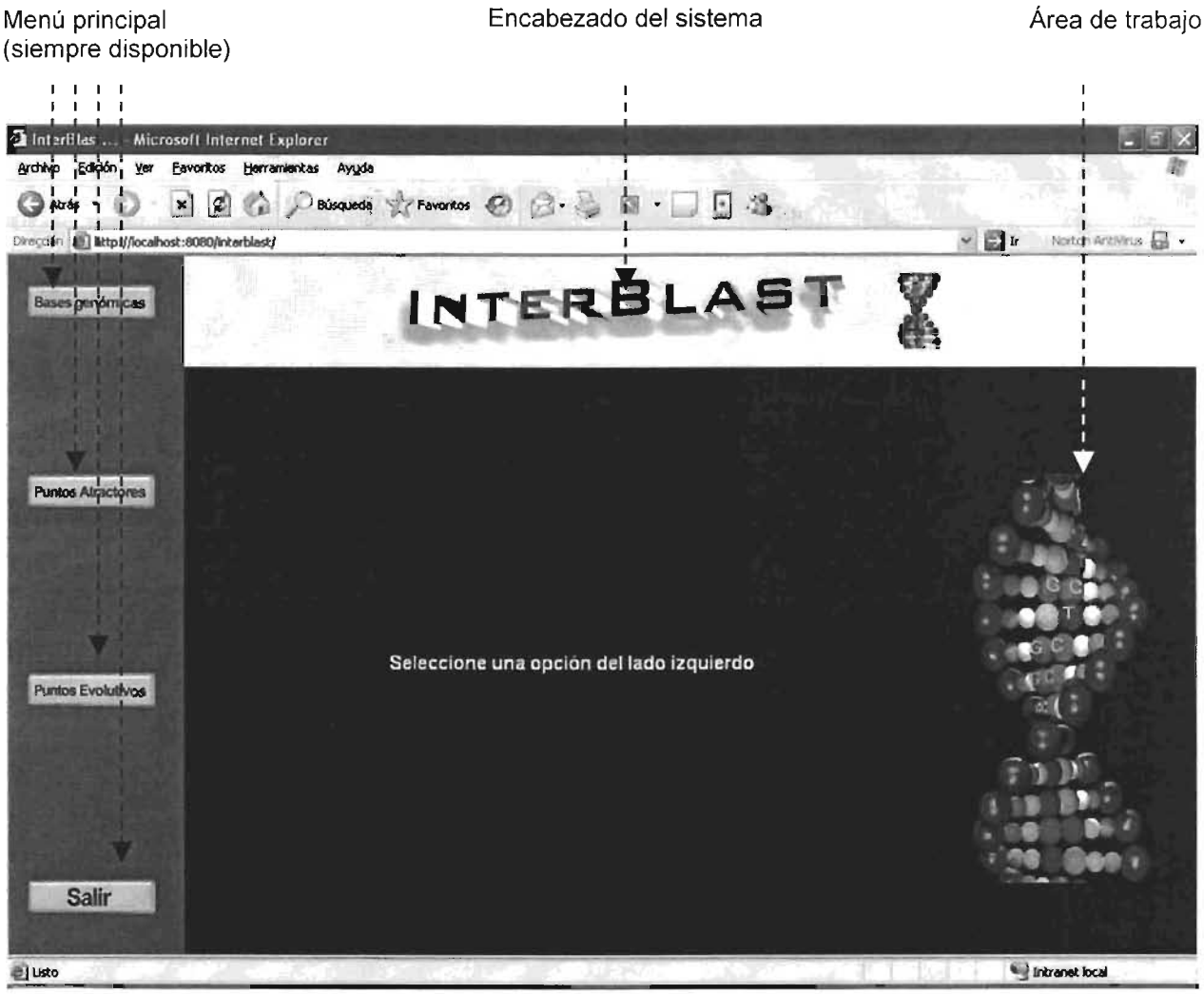


Fig. 3.12 Pantalla de ingreso al sistema

En la ventana de inicio, Interblast ofrece la bienvenida al sistema. Y para obtener una buena funcionalidad contiene un menú siempre visible, con sub-opciones que sólo aparecen al momento de pasar el mouse por encima de un botón.

3.4.1 Formatear archivo genómico (pre-procesamiento de datos)

Antes de generar un punto atractor, es necesario contar con una base de datos genómica (en formato Fasta) que haya sido formateada previamente por Interblast. Por eso dentro de las opciones el menú izquierdo al seleccionar "Bases genómicas", aparece una opción para llevar a cabo dicho proceso (Véase Figura 3.13)

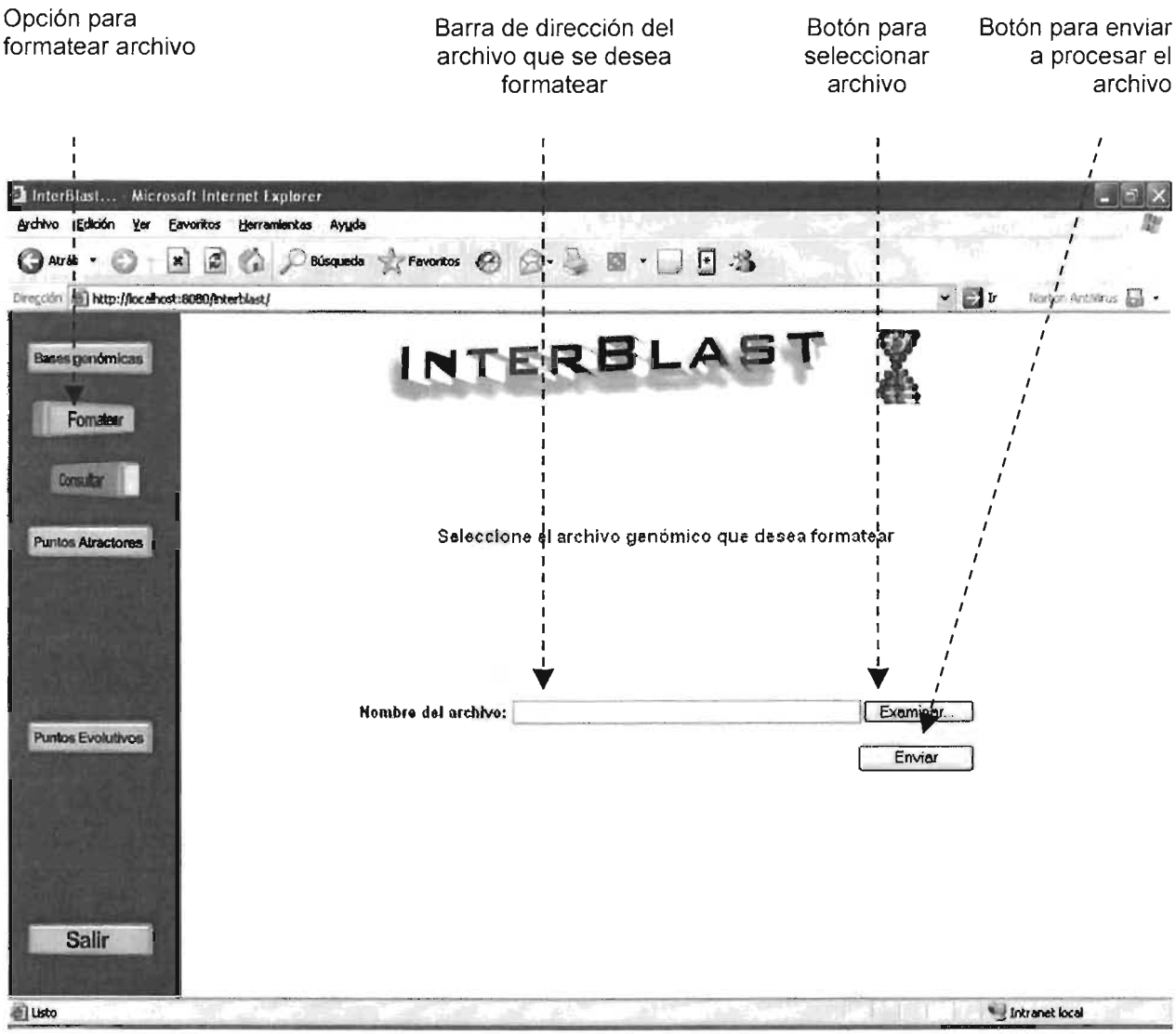


Fig. 3.13 Formateo de archivos genómicos

Después de que el usuario ha elegido y aceptado un archivo válido, Interblast empieza a procesarlo al mismo tiempo que muestra una ventana de espera.

Fig. 3.14 Ventana de espera mientras se formatea un archivo

La ventana de espera permanecerá visible hasta terminar el pre-procesamiento del archivo Fasta.

Una vez concluido el pre-procesamiento de los datos, aparece una ventana anunciándolo.

Fig. 3.15 Ventana formateo exitoso

Cuando ya se ha formateado una base de datos genómica exitosamente, es posible consultarla mediante la opción continua.

3.4.2 Consultar bases de datos genómicas

Para ver las bases de datos genómicas disponibles (previamente formateadas), existe la opción de consulta que devuelve los nombres de los organismos involucrados (Fig. 3.16).

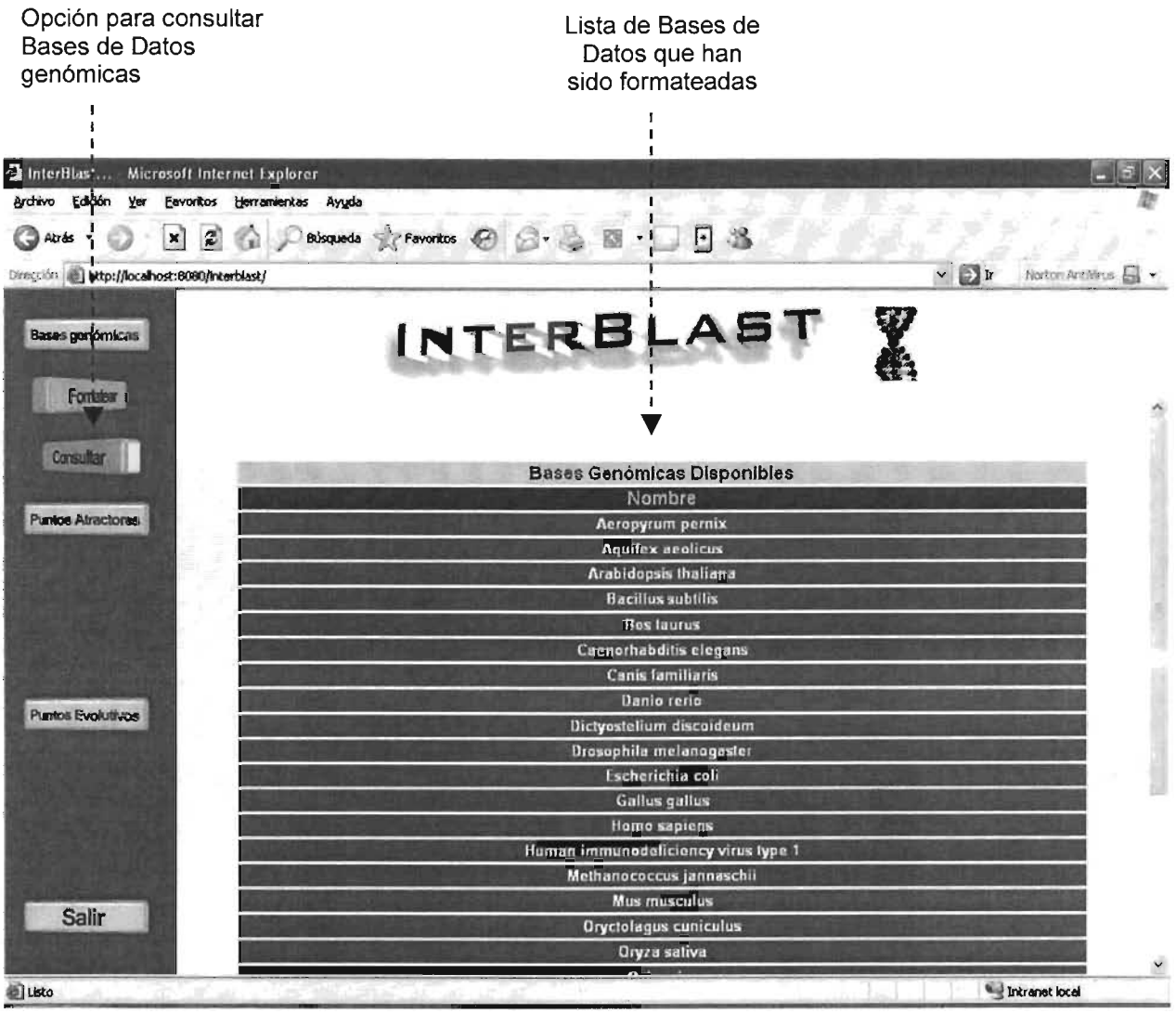


Fig. 3.16 Tabla con listado de todos los organismos cuyos archivos han sido formateados satisfactoriamente

Es importante aclarar que sin una base de datos genómica preprocesada no se pueden obtener puntos atractores y mucho menos evolutivos.

3.4.3 Generar punto atractor

El siguiente paso al pre-procesamiento de un archivo genómico en formato Fasta, es la generación del punto atractor del organismo respectivo. Y para ello existe una opción dentro del menú principal que permite lograrlo.

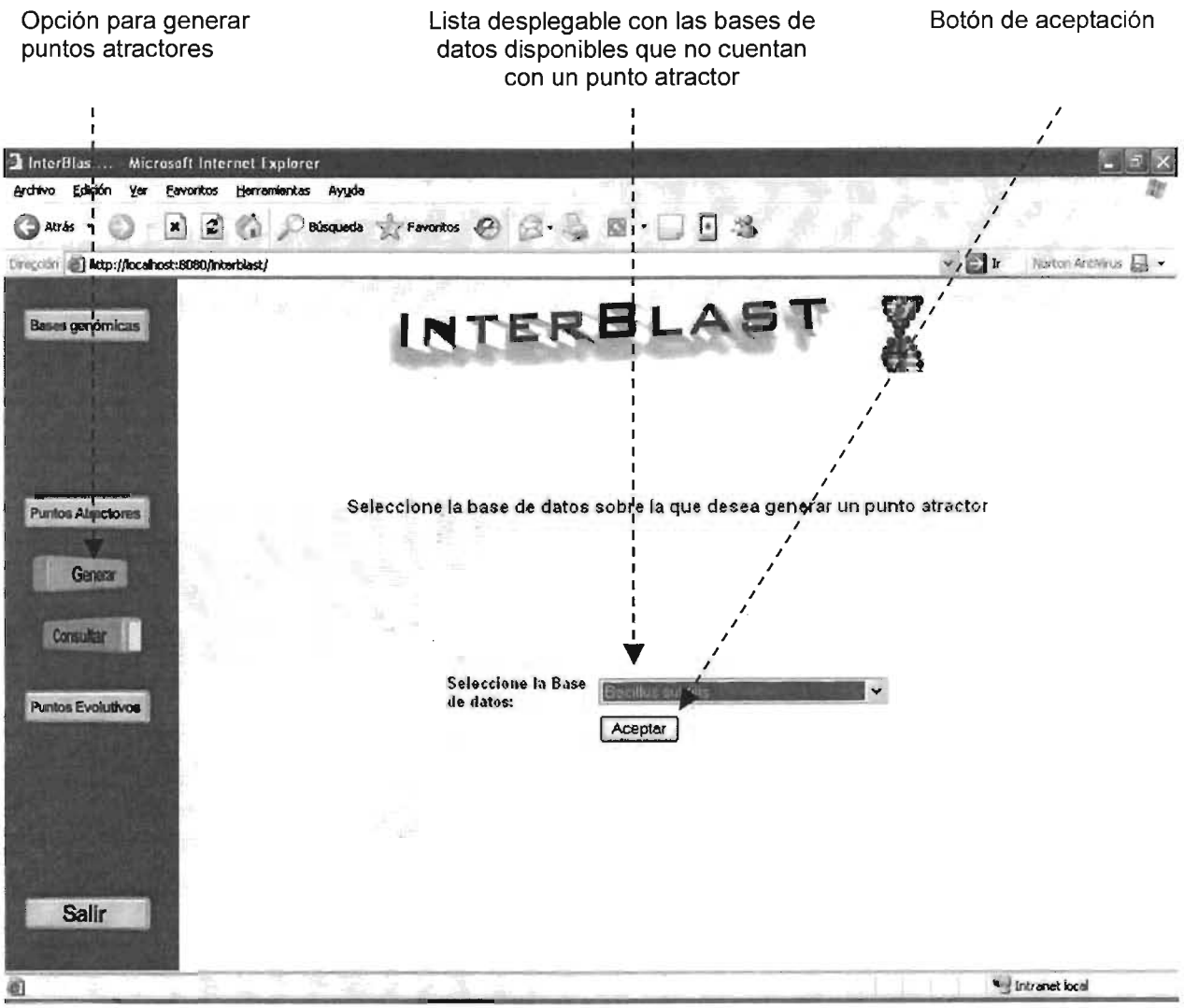


Fig. 3.17 Pantalla de selección de bases de datos pre-procesadas, a las cuales no se les han generado puntos atractores

Para generar el punto atractor se elige el nombre de un organismo que se desea trabajar y se presiona el botón de aceptar.

Mientras se genera el punto atractor del organismo seleccionado, aparece una ventana de espera como la de la figura 3.18.

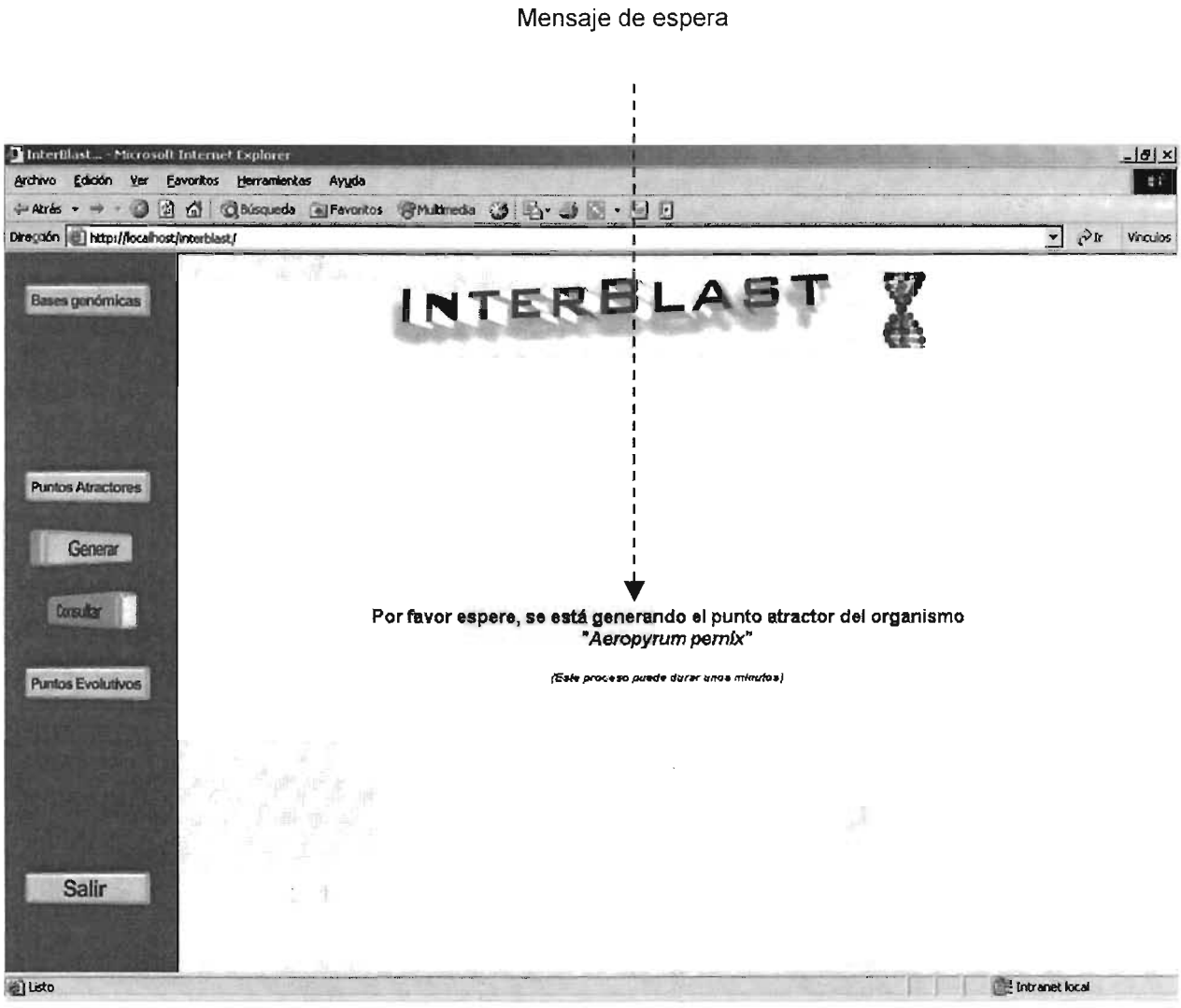


Fig. 3.18 Pantalla de espera mientras se crea el punto atractor

La ventana de espera estará vigente hasta que se termine todo el proceso de generación

Después de generarse el punto atractor de manera satisfactoria, se despliega una pantalla que señala que ahora es posible generar puntos evolutivos para dicho punto atractor.

Fig. 3.19 Mensaje exitoso de generación de un punto atractor

Si se desean consultar los puntos atractores disponibles, sólo es necesario elegir la opción correspondiente.

3.4.4 Consultar punto atractor

Para la consulta de puntos atractores, Interblast muestra una tabla en donde cada registro o renglón corresponde a un organismo (Fig. 3.20).

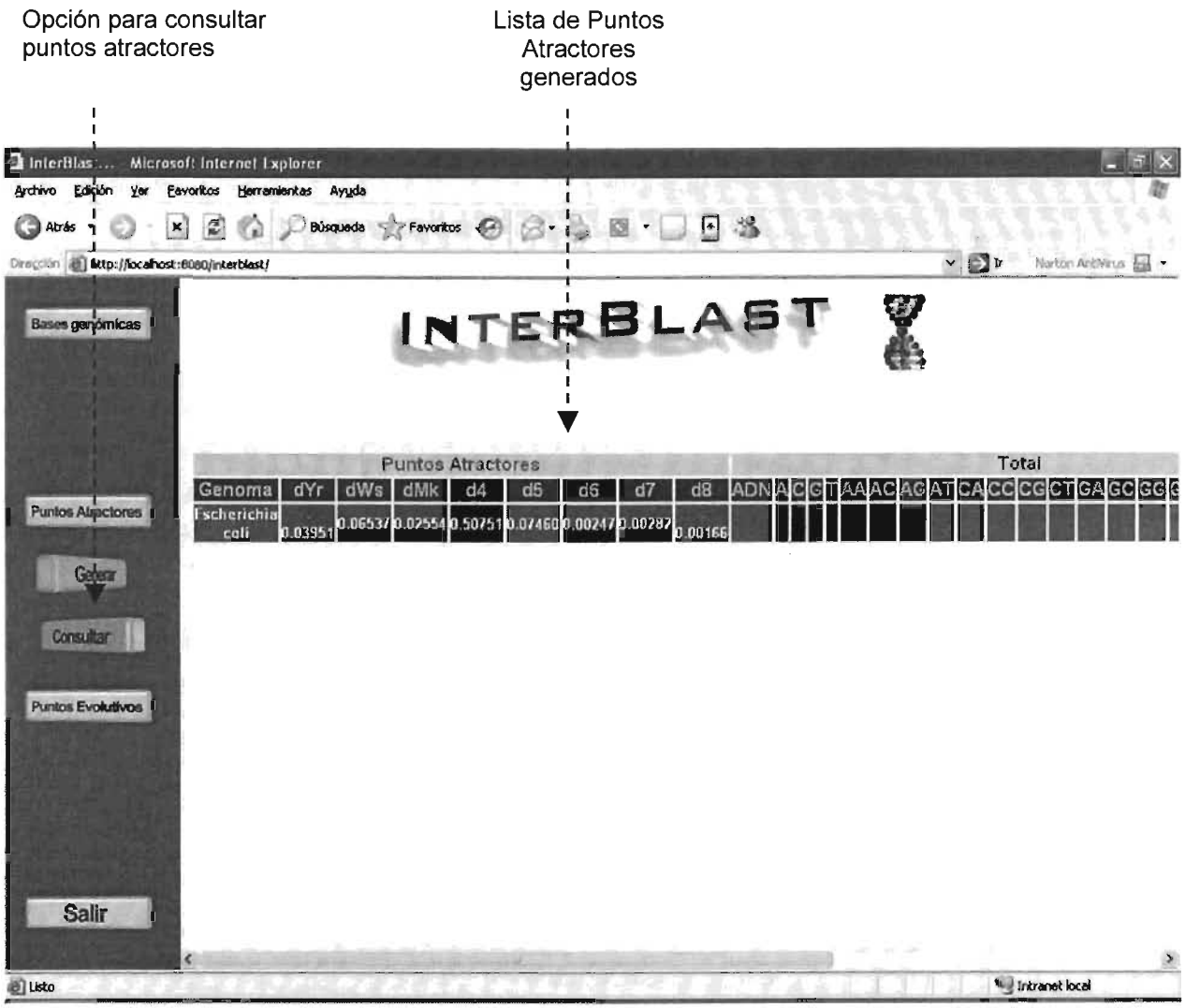


Fig. 3.20 Tabla que describe el punto atractor de cada organismo

3.4.5 Generar punto evolutivo

Para la generación de puntos evolutivos, es necesario seleccionar el punto atractor de un organismo y especificar el total de puntos a construir (véase figura 3.21).

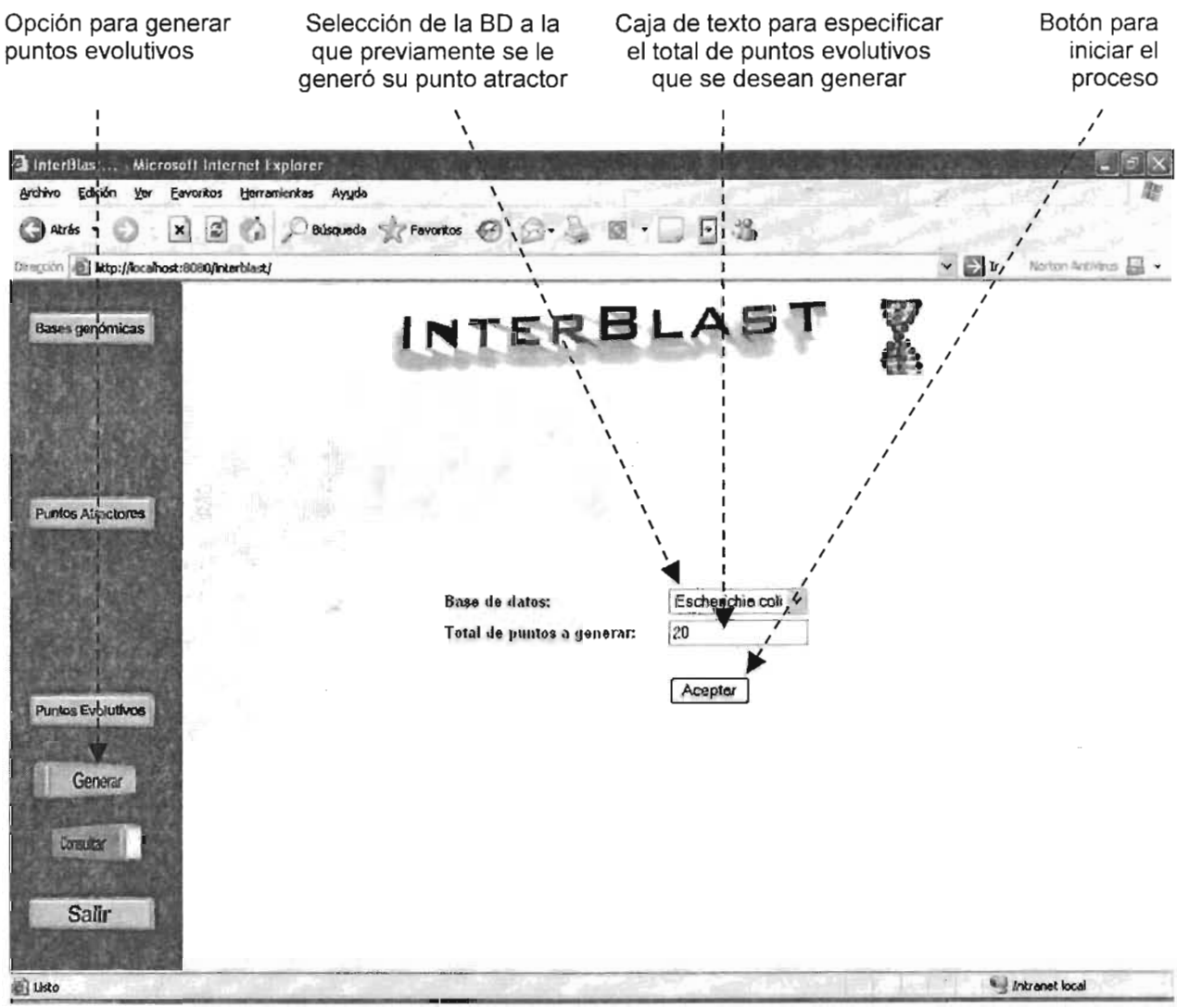


Fig. 3.21 Bases de datos que cuentan con su punto atractor y sobre las cuales es posible generar puntos evolutivos

Al igual que en otras operaciones, al momento en que Inteblast se encuentra procesando la información, una ventana de espera se aparece.

Mensaje de espera

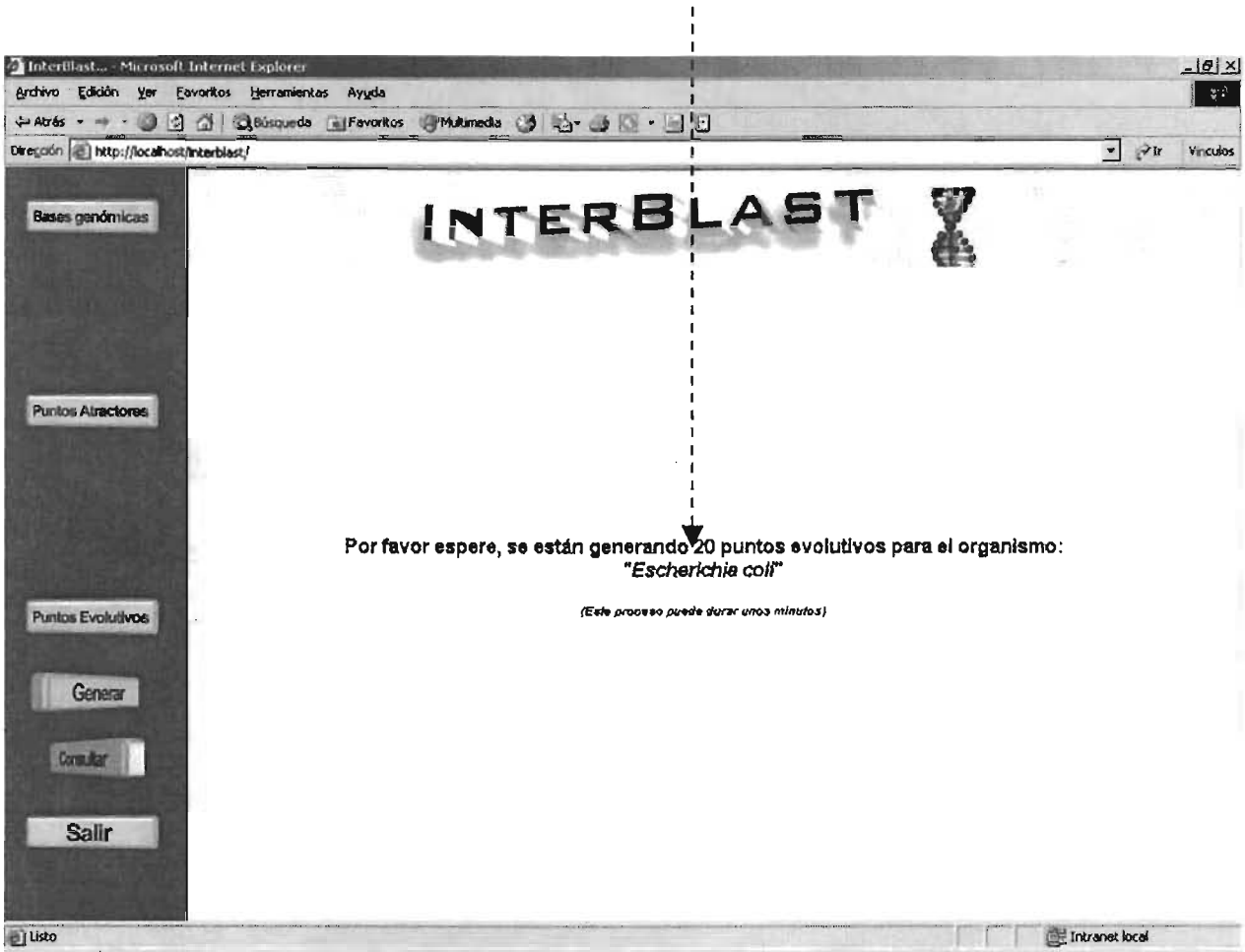


Fig. 3.22 Mensaje de espera mientras se generan los puntos evolutivos

En este proceso se llevan a cabo las tareas más complicados del sistema, tales como: generación de cadenas aleatorias, generación de puntos evolutivos, aplicación de los modelos matemáticos, filtrado de resultados y alineación local de secuencias.

Al concluir el proceso de manera exitosa, el sistema muestra una ventana (Fig. 3.23) de notificación para que el usuario sepa que ya puede consultar sus resultados.

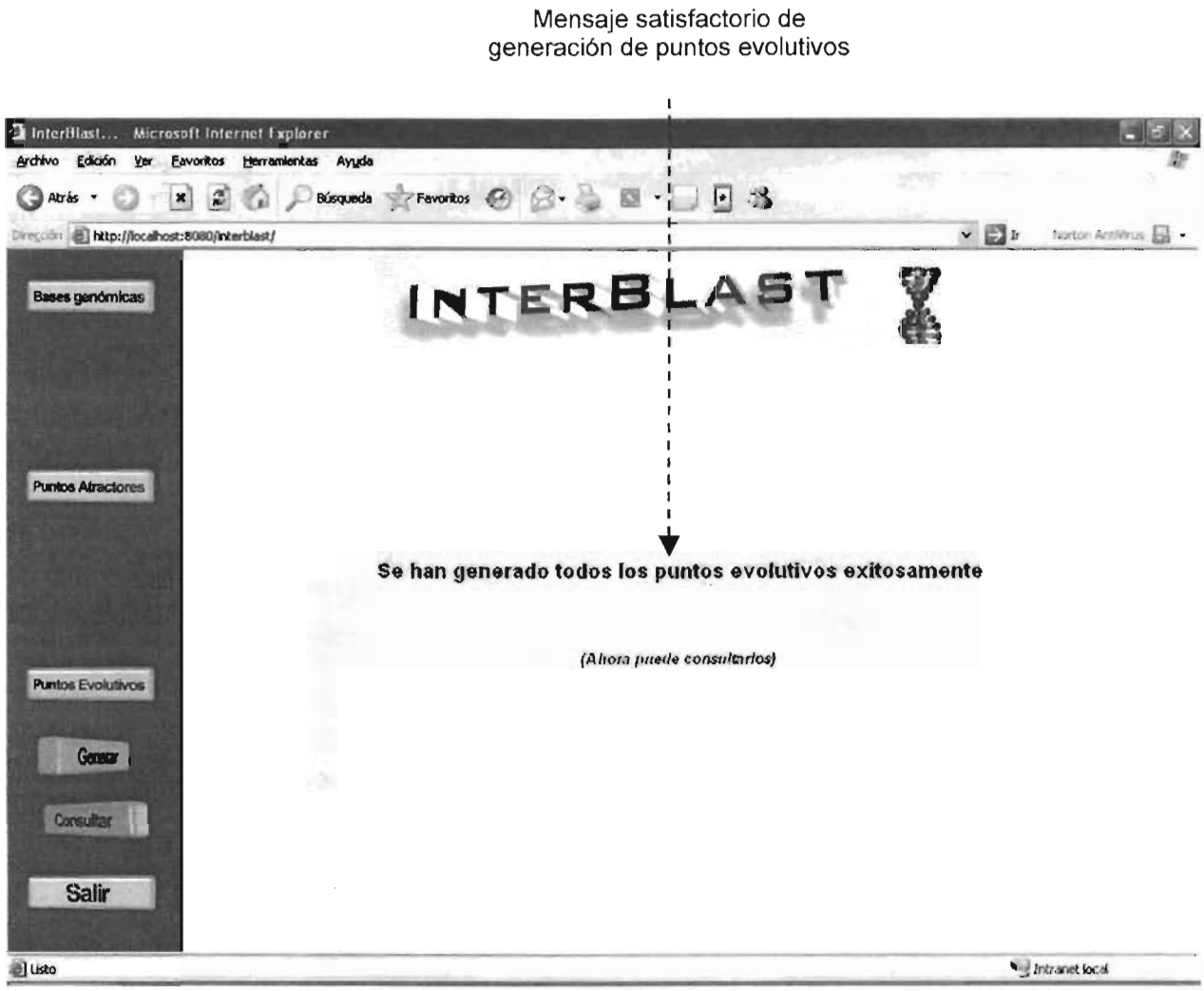


Fig. 3.23 Ventana de generación exitosa de los puntos evolutivos requeridos

En caso de generarse un error, el usuario será notificado mediante una ventana de alerta.

3.4.6 Consultar punto evolutivo

La parte más importante de todo el sistema, la representa la consulta de resultados, ya que apartir de ellos es posible estudiar los mecanismos de evolución molecular de los organismos.

Para consultar los puntos evolutivos obtenidos, es preciso seleccionar un organismo, el total de registros (para evitar sobresaturar el sistema) y el orden de almacenamiento de los mismos, tal y como se muestra en la figura 3.24:

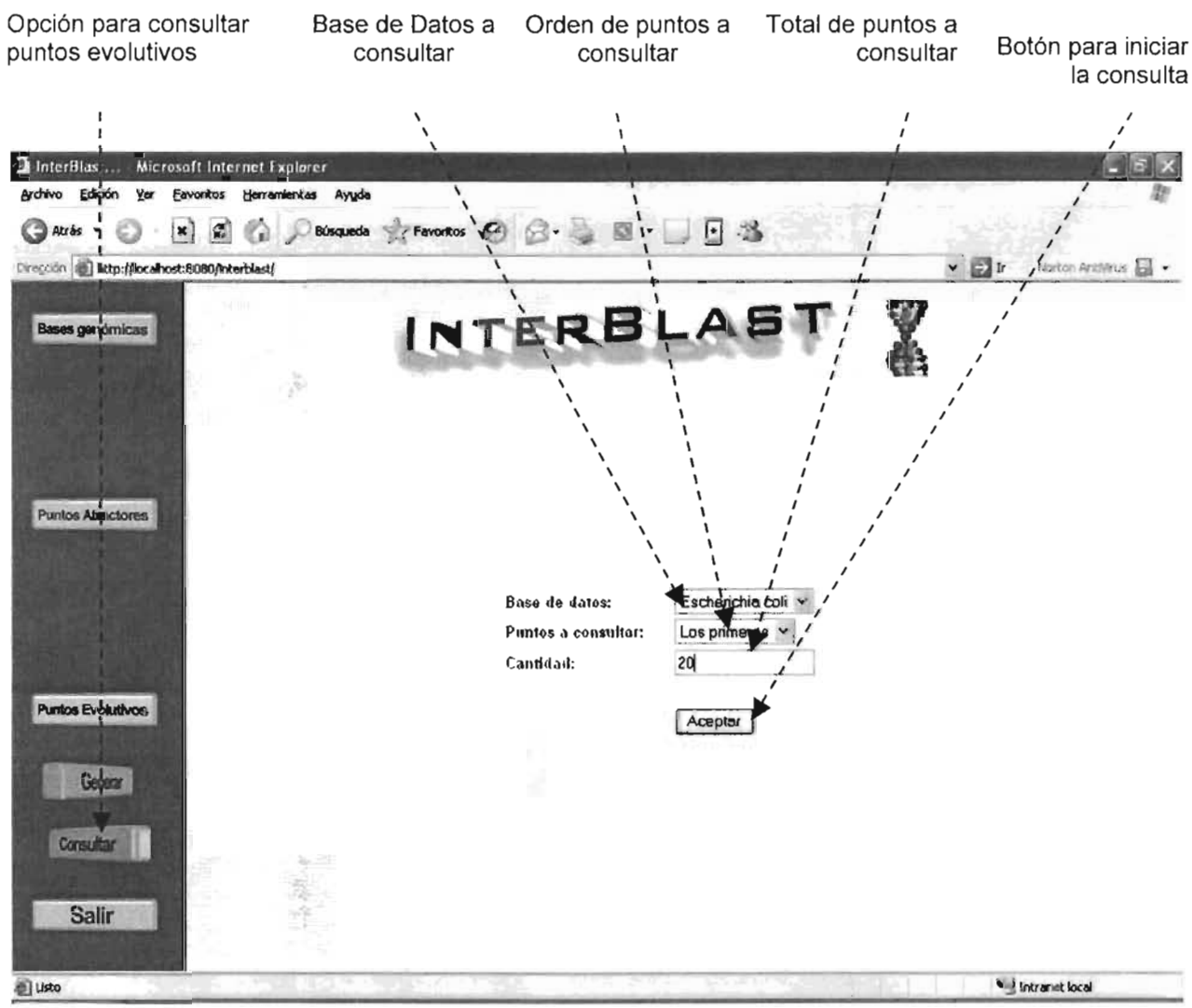


Fig. 3.24 Pantalla de consulta de puntos evolutivos

Los puntos evolutivos se despliegan en una larga tabla que aloja todos los datos descritos en las tablas: puntoEvolutivo y puntoEvolutivoHit de la base de datos Interblast.

En la figura 3.25 se incluye el primer segmento de la tabla resultante de un determinado organismo.

Tabla que muestra los puntos evolutivos solicitados

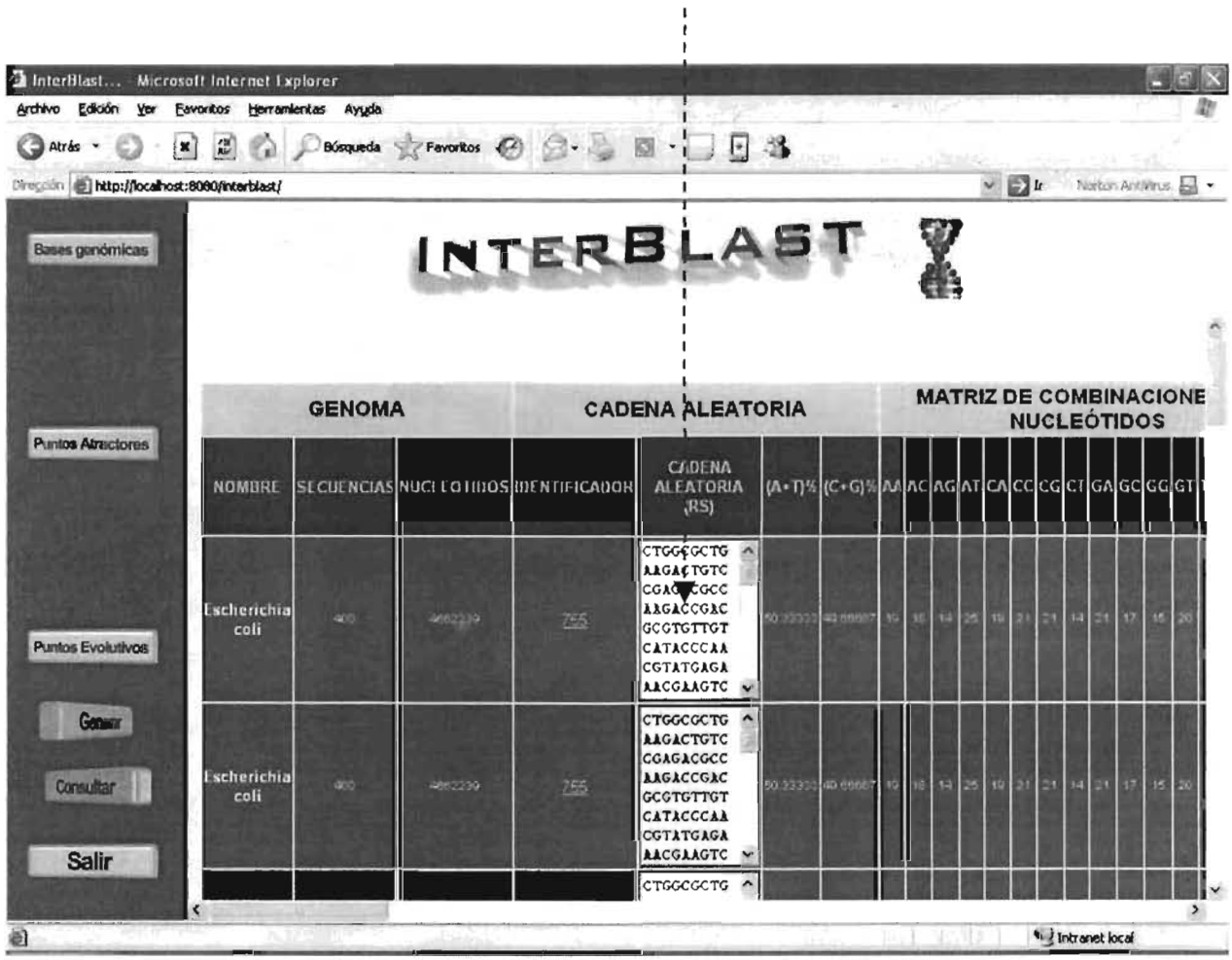


Fig. 3.25 Puntos evolutivos del organismo *E. coli*

Los puntos evolutivos tienen un número identificador a manera de hipervínculo, que si es presionado despliega una ventana con la información individual de dicho punto (Fig. 3.26).

Despliegue individual y mejor
organizado de un punto evolutivo
en particular

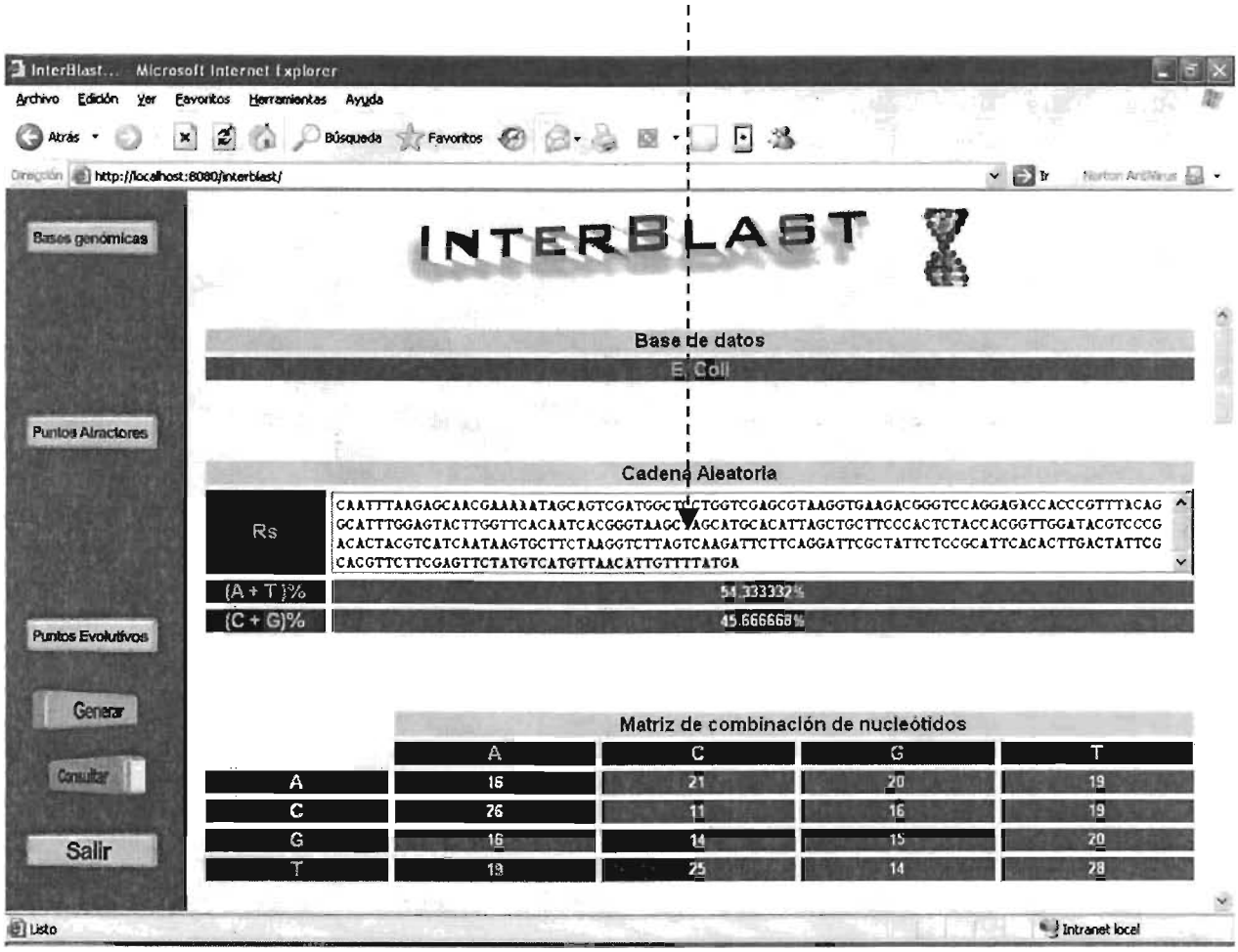


Fig. 3.26 Para analizar un punto evolutivo de mejor manera es preciso presionar sobre la liga de la pantalla de consulta principal

En la ventana descriptiva los datos aparecen mejor distribuidos, e inclusive las matrices aparecen a manera de tablas.

3.4.7 Salir

La última opción es la de salida del sistema, cuya finalidad es evitar que el sistema tenga muchas conexiones inútiles.

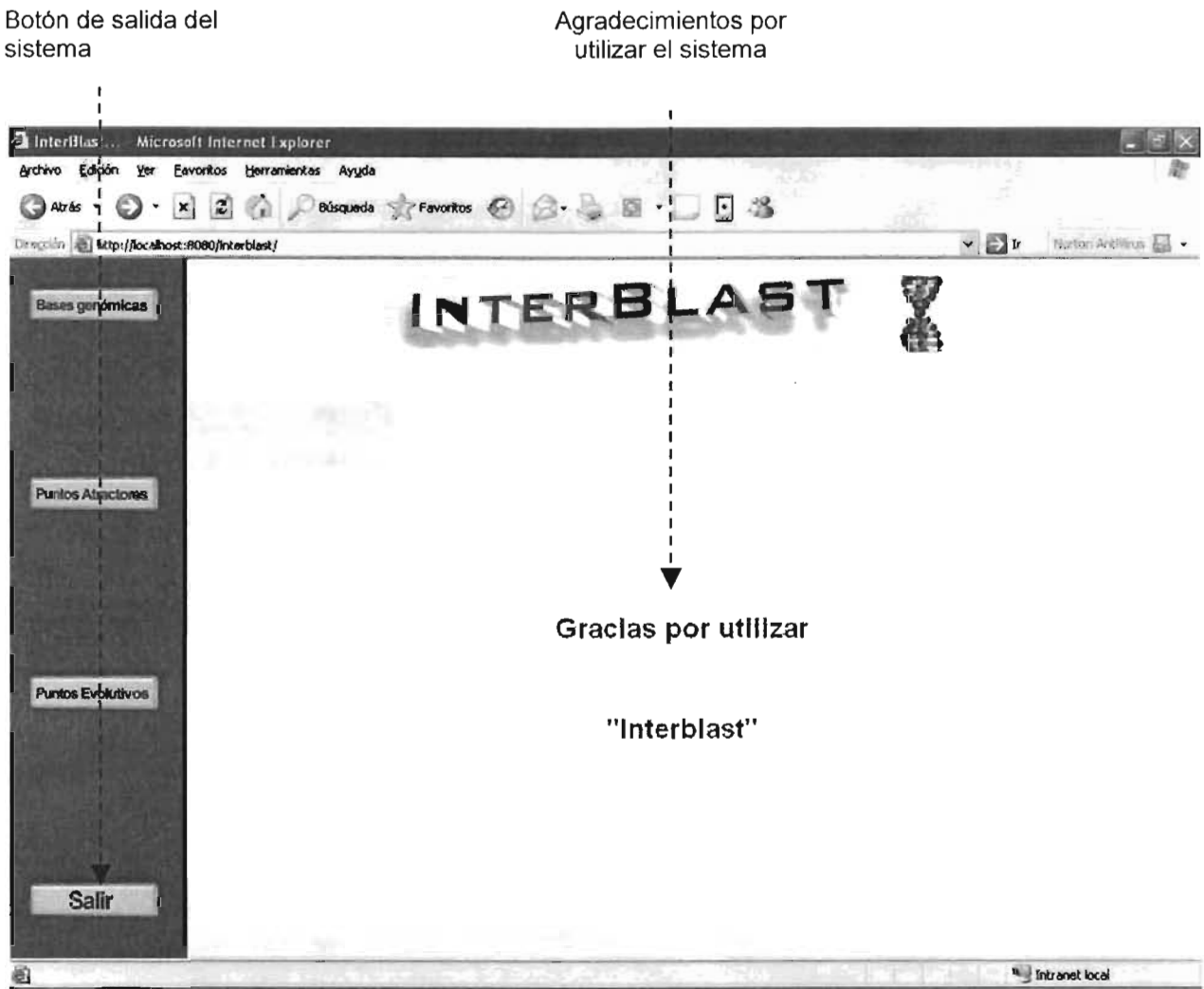


Fig. 3.27 La pantalla de salida cierra el sistema y finaliza la sesión

Es importante señalar que los resultados obtenidos con la ayuda de Interblast son cien por ciento compatibles con las alineaciones que se pueden realizar con el servicio Blast de NCBI vía Internet [8], debido a que Interblast cuenta con el mismo motor de alineación de Blast.

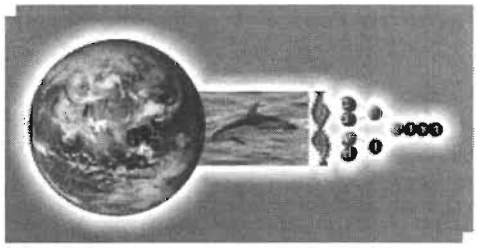
3.5 Resumen

El servicio del sistema Blast que ofrece el NCBI a través de internet, se limita a realizar alineaciones locales de secuencias a velocidades reducidas, por lo cual es necesario un sistema más robusto que además de ser rápido permita generar puntos atractores y evolutivos mediante la aplicación de diversos modelos matemáticos. Es por eso que Interblast además de cumplir con todos los requerimientos señalados, es el único que permite encontrar de manera exitosa cadenas que ofrezcan información valiosa para deducir patrones de evolución molecular, de la manera planteada en el capítulo II.

En el capítulo final se analizan algunos resultados que demuestran parte del potencial de investigación que ofrece Interblast.

CAPÍTULO IV

RESULTADOS



Interblast permite obtener una gran cantidad de resultados, cuya interpretación requiere de un análisis profundo por parte de biólogos especialistas en evolución molecular.

A lo largo del presente capítulo sólo analizaremos ciertos resultados obtenidos para el genoma de la bacteria *E. coli*.

4.1 Resultados

De toda la gama de resultados que ofrece Interblast, sólo se analiza la relación existente entre los valores de los puntos evolutivos, el atractor respectivo, la proporcionalidad de nucleótidos (A+T y C+G), y el total de subcadenas que se alinean localmente sobre la base de datos del organismo que se analice, en este caso *E. coli*.

En la tabla 4.1 se agrega una muestra de ocho puntos evolutivos para conocer cuantas de sus subcadenas tienen alineaciones exitosas:

Organismo	dYRAtractor	dYREvolutivo	dWSAtractor	dWSEvolutivo	dMKAtractor	dMKEvolutivo
Escherichia coli	-0.03951	-0.01151	0.06537	0.02238	0.02554	0.00810
Escherichia coli	-0.03951	-0.04564	0.06537	0.02641	0.02554	0.00943
Escherichia coli	-0.03951	-0.03708	0.06537	0.00293	0.02554	0.00329
Escherichia coli	-0.03951	-0.07640	0.06537	0.00987	0.02554	0.01658
Escherichia coli	-0.03951	-0.03001	0.06537	0.04282	0.02554	0.00293
Escherichia coli	-0.03951	-0.01662	0.06537	0.01307	0.02554	0.00316
Escherichia coli	-0.03951	-0.02488	0.06537	0.02467	0.02554	0.00598
Escherichia coli	-0.03951	-0.03083	0.06537	0.00943	0.02554	0.00943

d4Atractor	d4Evolutivo	d5Atractor	d5Evolutivo	d6Atractor	d6Evolutivo	d7Atractor	d7Evolutivo
0.50751	0.48333	0.07460	0.05351	0.00247	0.00068	0.00282	0.00003
0.50751	0.53000	0.07460	0.08696	0.00247	0.00243	0.00282	0.00077
0.50751	0.49000	0.07460	0.08361	0.00247	0.00228	0.00282	0.00178
0.50751	0.49667	0.07460	0.07023	0.00247	0.00523	0.00282	0.00021
0.50751	0.51000	0.07460	0.06020	0.00247	0.00189	0.00282	0.00135
0.50751	0.47000	0.07460	0.06355	0.00247	0.00105	0.00282	0.00028
0.50751	0.53667	0.07460	0.08027	0.00247	0.00115	0.00282	0.00154
0.50751	0.49000	0.07460	0.06355	0.00247	0.00195	0.00282	0.00083

d8Atractor	d8Evolutivo	AmasTEvolutivo	CmasGEvolutivo	ExtensionesExitosasEvolutivo
-0.00166	-0.00004	51.66667	48.33333	23
-0.00166	-0.00404	47.00000	53.00000	22
-0.00166	-0.00382	51.00000	49.00000	22
-0.00166	-0.00732	50.33333	49.66667	21
-0.00166	-0.00233	49.00000	51.00000	20
-0.00166	-0.00238	53.00000	47.00000	18
-0.00166	-0.00314	46.33333	53.66667	17
-0.00166	-0.00295	51.00000	49.00000	17

Tabla 4.1 Muestra de puntos evolutivos

En la tabla anterior se puede observar que con tan sólo una muestra de ocho puntos evolutivos, el total de subcadenas alineadas por cadena aleatoria es superior a 15 (Extensiones exitosas), de tal manera que los filtros aplicados son acertados y orientan la búqueda a cadenas candidatas a ser consideradas evolutivas al organismo en cuestión.

4.2 Minería de datos

Debido a que el análisis de muchas variables para encontrar patrones resulta en ocasiones complicado, es necesario hacernos valer de algunos algoritmos computacionales capaces de procesar lo que a simple vista no se ve. Es por eso que también se usan mecanismos de minería de datos.

4.2.1 Weka

Es una colección de algoritmos de aprendizaje de máquina, capaces de realizar tareas de minería de datos. Incluye herramientas para el pre-procesamiento de datos, la clasificación, regresión, agrupamiento, reglas de asociación y visualización. Permite también el desarrollo de nuevos esquemas de aprendizaje máquina [3].

Es un software de código abierto (open source) bajo la licencia GNU (Licencia Pública General) [6].

4.2.1.1 Pre-procesamiento de datos

Antes de utilizar la información con esta herramienta de minería de datos, es necesario que los datos mostrados en la tabla 4.1 sean almacenados en un archivo propio de Weka con extensión arff.

La manera de convertir los datos a formato arff, consiste en separar el nombre de la relación a trabajar, los atributos y los datos, tal y como se observa en la tabla 4.2

```
@relation evolucion

@attribute Organismo {Escherichia} ,
@attribute dYRAtractor real,
@attribute dYREvolutivo real,
@attribute dWSAtractor real,
@attribute dWSEvolutivo real,
@attribute dMKAtractor real,
@attribute dMKEvolutivo real,
@attribute d4Atractor real,
@attribute d4Evolutivo real,
@attribute d5Atractor real,
@attribute d5Evolutivo real,
@attribute d6Atractor real,
@attribute d6Evolutivo real,
@attribute d7Atractor real,
@attribute d7Evolutivo real,
@attribute d8Atractor real,
@attribute d8Evolutivo real,
@attribute AmasTEvolutivo real,
@attribute CmasGEvolutivo real,
@attribute ExtensionesExitosasEvolutivo real

@data
Escherichia,-0.03951,-0.01151,0.06537,0.02238,0.02554,0.00810,0.50751,0.48333,0.07460,0.05351,0.00247,0.00068,0.00282,0.00003,-0.00166,-0.00004,51.66,48.33,23
Escherichia,-0.03951,-0.04564,0.06537,0.02641,0.02554,0.00943,0.50751,0.53000,0.07460,0.08696,0.00247,0.00243,0.00282,0.00077,-0.00166,-0.00404,47.00,53.00,22
Escherichia,-0.03951,-0.03708,0.06537,0.00293,0.02554,0.00329,0.50751,0.49000,0.07460,0.08361,0.00247,0.00228,0.00282,0.00178,-0.00166,-0.00382,51.00,49.00,22
Escherichia,-0.03951,-0.07640,0.06537,0.00987,0.02554,0.01658,0.50751,0.49667,0.07460,0.07023,0.00247,0.00523,0.00282,0.00021,-0.00166,-0.00732,50.33,49.66,21
Escherichia,-0.03951,-0.03001,0.06537,0.04282,0.02554,0.00293,0.50751,0.51000,0.07460,0.06020,0.00247,0.00189,0.00282,0.00135,-0.00166,-0.00233,49.00,51.00,20
Escherichia,-0.03951,-0.01662,0.06537,0.01307,0.02554,0.00316,0.50751,0.47000,0.07460,0.06355,0.00247,0.00105,0.00282,0.00028,-0.00166,-0.00238,53.00,47.00,18
Escherichia,-0.03951,-0.02488,0.06537,0.02467,0.02554,0.00598,0.50751,0.53667,0.07460,0.08027,0.00247,0.00115,0.00282,0.00154,-0.00166,-0.00314,46.33,53.66,17
Escherichia,-0.03951,-0.03083,0.06537,0.00943,0.02554,0.00943,0.50751,0.49000,0.07460,0.06355,0.00247,0.00195,0.00282,0.00083,-0.00166,-0.00295,51.00,49.000,17
```

Tabla 4.2 Datos en formato arff

Una vez que se tiene el archivo arff, es necesario cargarlo al programa y seleccionar los atributos que se van a utilizar (Fig. 4.1)

Lista de atributos a utilizar

Nombre y cantidad de las agrupaciones posibles

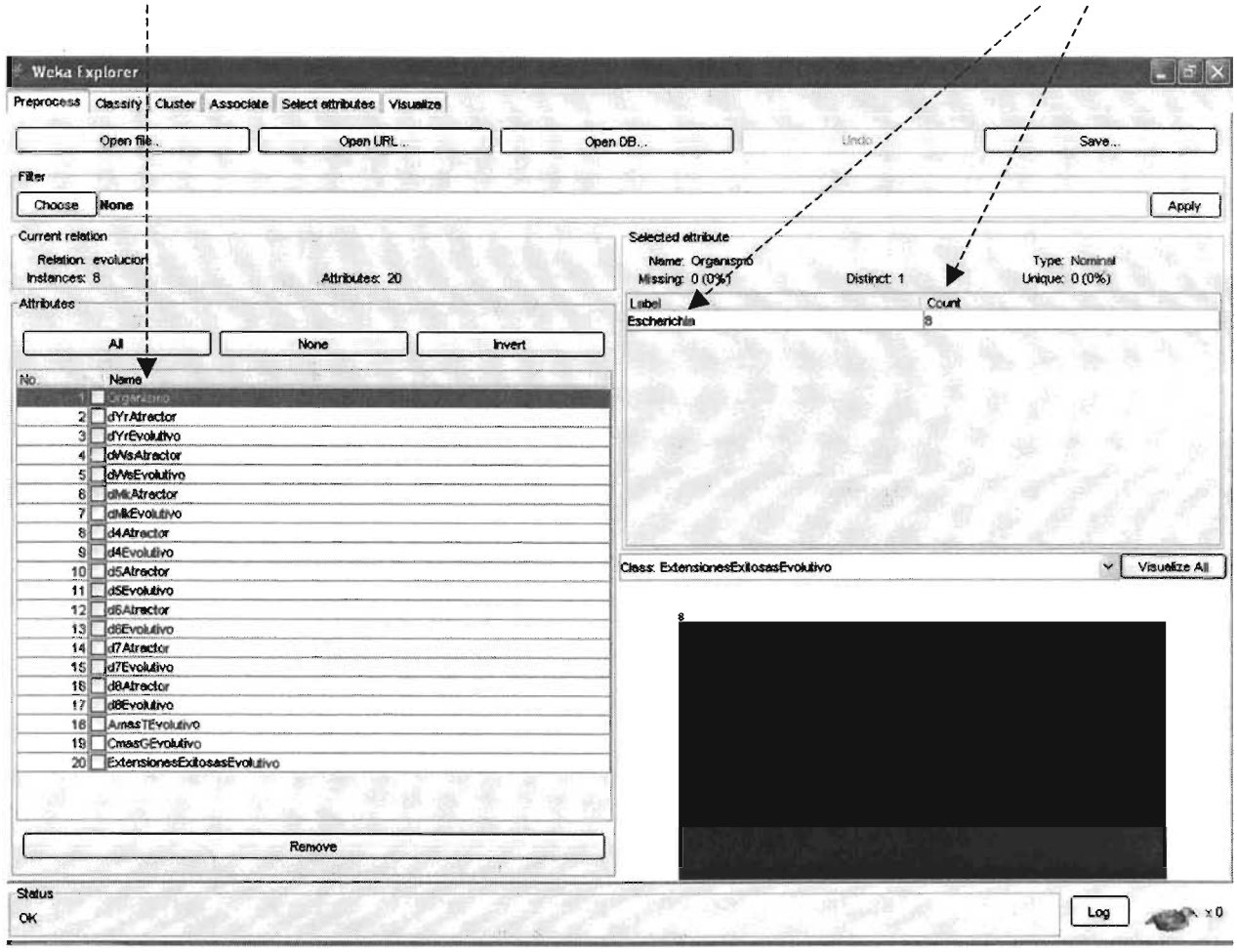


Fig. 4.1 Carga y selección de atributos a utilizar

Se puede notar que la única agrupación posible es la del nombre del organismo (*Escherichia*), ya que a diferencia de los demás atributos, este es el único no numérico.

El desglose de distribución de cada uno de los atributos se muestra en la Figura 4.2, y como es de esperarse los valores atractores generan una gráfica uniforme y las demás no, puesto que para cada valor de un punto evolutivo se permite un cierto radio de aproximación a un único valor atractor.

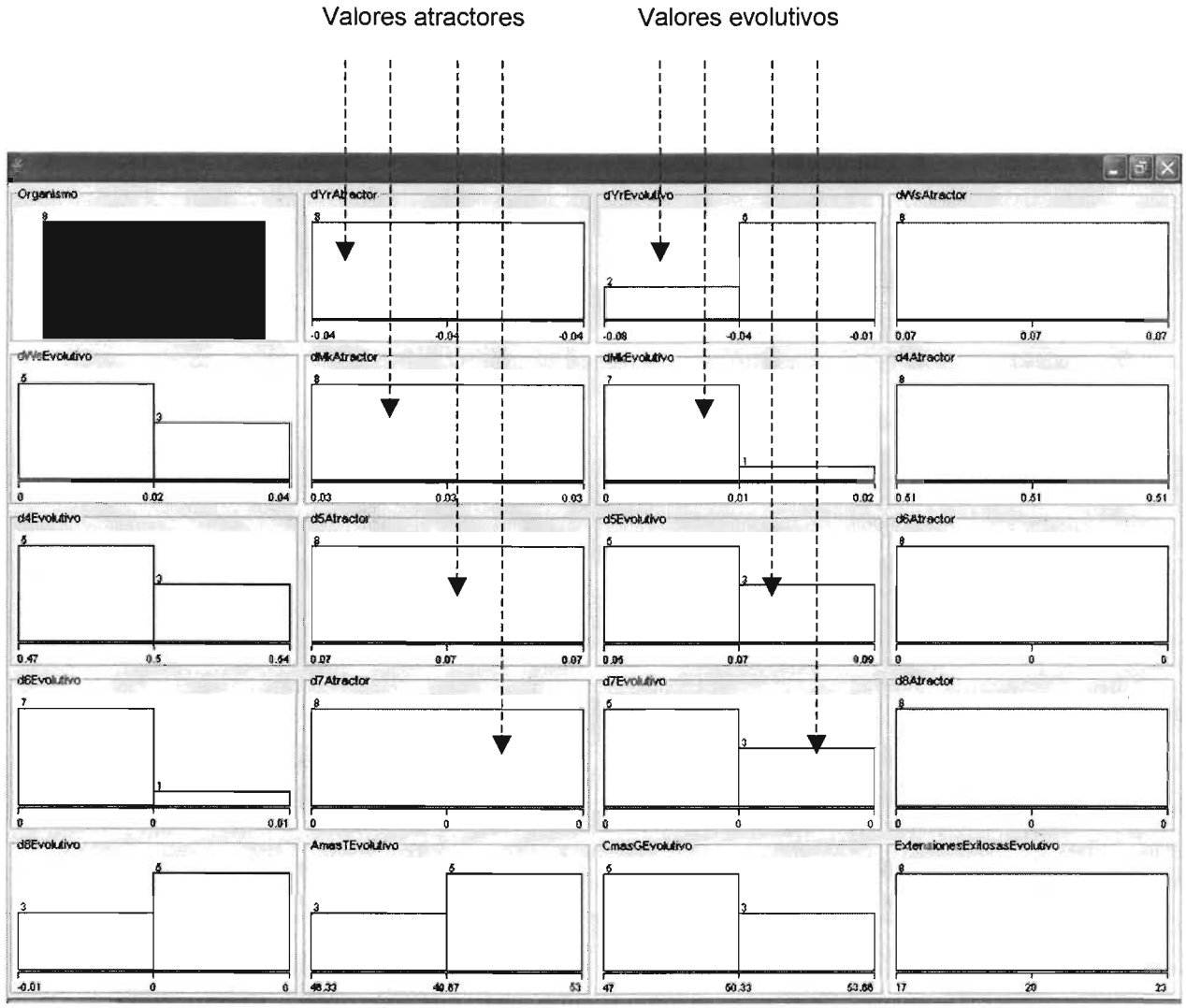


Fig. 4.2 La distribución de los valores de los puntos evolutivos oscila dentro del rango de acercamiento permitido al punto atractor

Las gráficas anteriores hacen pensar que dependiendo el organismo que se utilice será la tendencia de distribución de sus valores evolutivos.

4.2.1.2 Clasificación

Debido a que los atributos seleccionados de toda la tabla de resultados de Interblast, son prácticamente no numéricos, Weka no puede llevar a cabo una clasificación de los datos,

4.2.1.3 Agrupación

La agrupación o clustering de los datos arroja medias de distribución normal junto con las desviaciones estándar de cada uno de los atributos, lo que permite observar qué tanto están alejados los valores de los componentes evolutivos de los atractores.

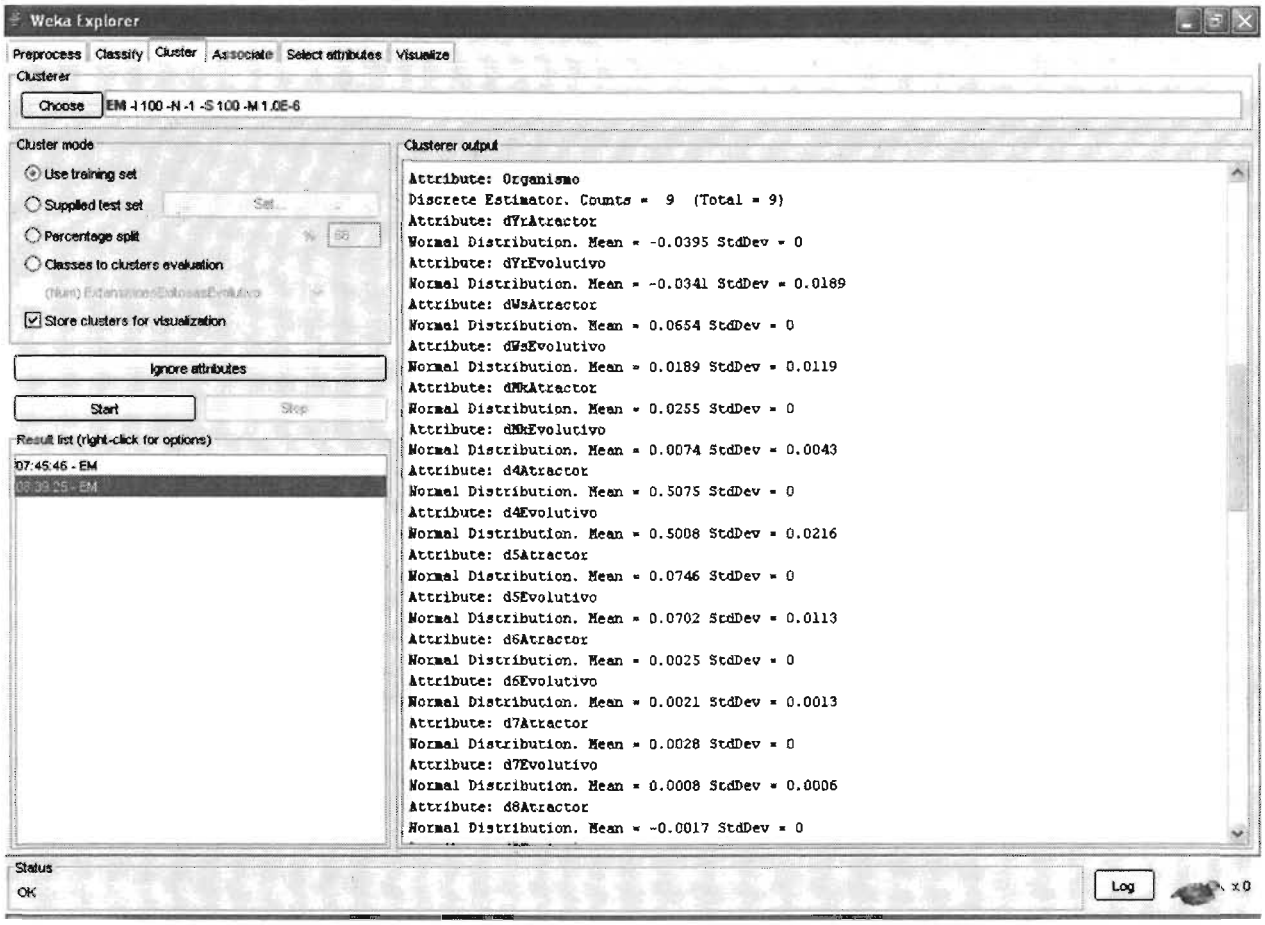


Fig. 4.3 La desviación entre valores evolutivos y atractores es muy pequeña

El hecho de que la cercanía entre los valores del punto atractor y el punto evolutivo sea mínima, abre mayores posibilidades de encontrar mejores cadenas que sirvan de estudio en la evolución molecular.

4.2.1.6 Visualización

El total de gráficas bidimensionales que se visualizan para 20 atributos es de 20x20 = 400, por lo que únicamente se incluye un mapa general cuyo estudio requiere de un análisis minucioso que escapa a los objetivos de esta tesis.

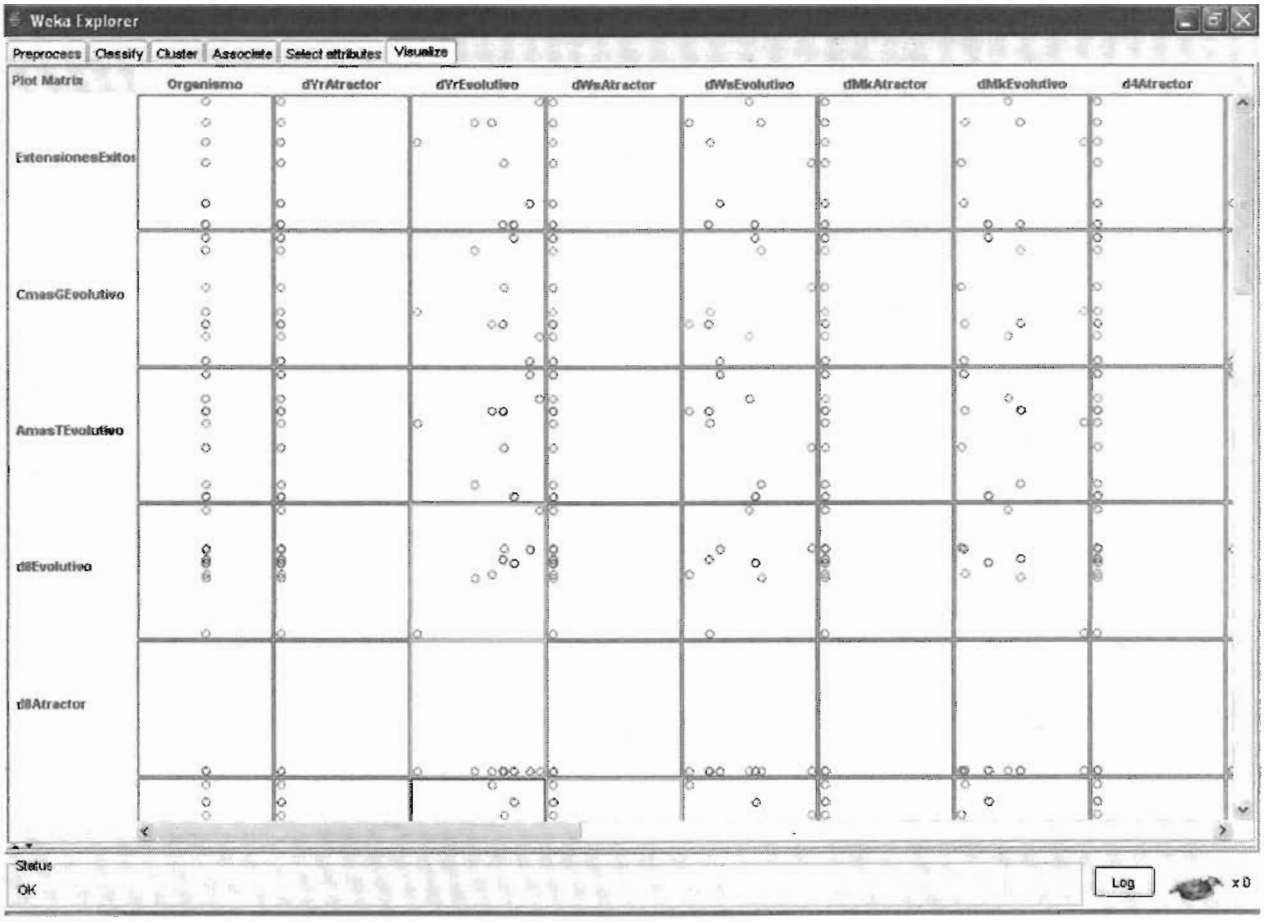


Fig. 4.4 La visualización no es buena por la cantidad de registros utilizados

Debido a que solamente se usaron 8 registros y los valores de las extensiones exitosas se distribuyeron uniformemente, este tipo de visualización requiere de estudios extensos.

4.3 Resumen

Cuando se genera una cadena aleatoria cuyos valores atraviesan de manera satisfactoria los filtros de proporcionalidad y de acercamiento al punto atractor de un organismo, se puede decir que dicha cadena es una fuerte candidata para considerarse evolutivamente compatible con dicho organismo, ya que se si se alinea localmente la cadena en contra de la base de datos genómica del organismo, es posible observar que el total de subcadenas que alinean con el organismo, son diversas, es decir, comparten varios segmentos de ADN.

Comprobar lo anterior es sencillo con la ayuda de Interblast, sólo es necesario consultar los puntos evolutivos obtenidos y revisar cuántos registros se pueden agrupar por el número identificador de la cadena aleatoria, esto quiere decir que cada uno de esos registros contiene una subcadena de esa cadena aleatoria.

CONCLUSIONES

Explotar la información contenida dentro del ADN de un organismo resultaría una tarea imposible si no se contara con los algoritmos y herramientas computacionales de la actualidad.

Llevar a cabo todo el proceso de determinación de evolución molecular, con el servicio que ofrece NCBI a través de Blast en línea no es posible, porque existe una infinidad de cálculos que se deben realizar previamente a la alineación local de cadenas y porque los cálculos demandan mucho tiempo de respuesta.

La elaboración de un software (Interblast) capaz de realizar las tareas de preprocesamiento de datos, aplicación de modelos matemáticos, alineaciones locales (con la adaptación de Blast), y almacenamiento y recuperación de los resultados en una base de datos relacional, es la mejor solución para intentar descubrir patrones de evolución molecular de la manera planteada.

Con Interblast los tiempos de alineación local de secuencias se disminuyen drásticamente en relación al servicio que ofrece NCBI vía web, ya que Interblast también cuenta con el motor de alineación Blast pero adaptado localmente.

Los modelos matemáticos contenidos dentro de Interblast y mostrados en el capítulo II, permiten identificar de manera única a un organismo y también sirven de filtros para decidir si una determinada cadena de nucleótidos pertenece a un organismo.

Cuando se genera una cadena aleatoria cuyos valores del punto evolutivo atraviesan de manera satisfactoria los filtros de proporcionalidad y de acercamiento al punto atractor de un organismo, se puede decir que dicha cadena es una fuerte candidata para considerarse evolutivamente compatible con dicho organismo, ya que se si se alinea localmente la cadena en contra de la base de datos genómica del organismo, es posible observar que son diversas las subcadenas que alinean con el organismo, es decir, comparten varios segmentos de ADN.

Para versiones futuras de Interblast es preciso no sólo alinear el punto evolutivo sobre un solo organismo a la vez, sino hacerlo sobre todas las bases de datos genómicas, ya que si la cadena alinea con varios organismos es posible pensar que dicha cadena puede ser la huella digital de un gen distribuido ampliamente en toda la escala filogenética, pero también es posible detectar un evento horizontal de transferencia de genes.

Finalmente cabe señalar que el mejor aprovechamiento de Interblast queda en manos de los especialistas en evolución molecular, puesto que el estudio de los resultados requiere de un análisis mucho más profundos.

BIBLIOGRAFÍA

1. Attwood, T., Parry, D.: Introduction to bioinformatics. Prentice Hall, Inglaterra, (1999).
2. H., Naya, H., Romero, A., Zavala, B., Alvarez, H.: Musto. Aerobiosis Increases the Genomic Guanine Plus Cytosine Content (CG%) in Prokaryotes. J. Mol. Evol., Vol. 55, 260-264 (2002).
3. H., Witten, Frank, Eibe: Data Mining. MK, E.U.A, (2000)
4. Han, Jiawei: Data Mining: Concepts and Techniques. MK, E.U.A, (2001)
5. Hanna, Phil: Manual de referencia JSP. McGraw-Hill, España (2002).
6. http://www.cs.waikato.ac.nz, (15/Abr/2005).
7. http://www.liebertpub.com, (11/Ene/2005).
8. http://www.ncbi.nlm.nih.gov, (01/Mar/2005).
9. J. R. Quintana, K. Grzeskowiak, K. Yanagi, and R.E. Dickerson.: Structure of a B-DNA Decamer with a Central T-A Step: C-G-A-T-T-A-A-T-C-G, J.Mol. Biol., Vol. 255, 379-395, (1992).
10. P., Miramontes, L., Medrano, C., Cerpa, R., Cedergren, G., FerbeYRe, G., Cocho.: Structural and Thermodynamic Properties of DNA Uncover Different Evolutionary Histories, J. Mol. Evol., Vol. 40, 698-704, (1995).

GLOSARIO

Alineación

Proceso de alineación o emparejamiento de dos o más secuencias para alcanzar los mayores niveles de identidad, con el propósito de valorar el grado de similaridad y la posibilidad de homología.

Alineación global

Es la alineación de toda la longitud de dos secuencias de proteínas o ácidos nucléicos.

Alineación Local

Es la alineación de alguna porción de dos secuencias de ácidos nucléicos o proteínas.

Blast

Es un conjunto de programas con algoritmos de búsqueda de similitud diseñados para explorar todos las bases de datos de secuencias. Sus siglas significan *Basic Local Alignment Search Tools o Herramientas de Búsqueda de Alineación Local Básica*.

Fasta

Es el primer algoritmo de búsqueda de similaridad en bases de datos. El programa busca las alineaciones locales óptimas al escanear la secuencia para emparejamientos pequeños llamadas "palabras".

Filtrado

Conjunto de propiedades que restringen la propagación de resultados no deseados hacia determinados procesos.

Homología

Similaridad atribuida al descendiente de un ancestro común

HSP

High-Scoring Segment Pair o Segmento Par de Mayor Puntuación. Alineación local sin huecos que logra el puntaje más alto dentro de una búsqueda.

Interblast

Software de aplicación de modelos de evolución molecular intercomunicado con Blast y bases de datos genómicas.

Punto Atractor

Es un vector de ocho componentes que definen de manera matemática y única a un organismo.

Punto Evolutivo

Es un vector de ocho componentes (de una cadena aleatoria) que se acerca en un determinado radio al punto atractor de un organismo. Dicho punto también cumple con una cierta proporcionalidad de nucléotidos y su cadena origen al ser alineada localmente contra la base genómica de un organismo, devuelve alineaciones exitosas. Por lo tanto, dicho punto define a la cadena como candidata para ser considerada evolutivamente compatible con el organismo.

Similaridad

Es el grado de relación entre secuencias de nucleótidos o proteínas.