EXTENSION DEL ALGORITMO PARA ANÁLISIS FILOGENÉTICO UPGMA (Unweighted Pair Group Method using Arithmetic averages) APLICADO A BASES DE DATOS

JOSÉ GABRIEL RAMÍREZ SUÁREZ

UNIVERSIDAD DEL NORTE
DIVISIÓN DE INGENIERÍAS
PROGRAMA DE INGENIERÍA DE SISTEMAS
BARRANQUILLA
2008

EXTENSION DEL ALGORITMO PARA ANÁLISIS FILOGENÉTICO UPGMA (Unweighted Pair Group Method using Arithmetic averages) APLICADO A BASES DE DATOS

Por

José Gabriel Ramírez Suárez

Tesis propuesta como cumplimiento de los requisitos para optar al título de:

Ingeniero de sistemas

Universidad del Norte

2008

Aprobada por

_____

_____

_____

_____

Programa autorizado para obtener el titulo

_____

Fecha _____

Nota de aceptación:

_____
_____
_____
_____
_____
_____

_____

Ing. Eduardo Enrique Zurek Varela, Ph.D.

Director del proyecto

_____

Ing. José Rafael Capacho Portilla

Coord. Programa de Ing. Sistemas

_____

Corrector

_____

Jurado

_____

Jurado

UNIVERSIDAD DEL NORTE

RESUMEN

EXTENSION DEL ALGORITMO PARA ANÁLISIS FILOGENÉTICO UPGMA (Unweighted Pair Group Method using Arithmetic averages) APLICADO A BASES DE DATOS

Por **José Gabriel Ramírez Suárez**

Director de tesis: Ing. Eduardo Zurek Varela, Ph.D.
Departamento de Ingeniería de Sistemas

Tesis presentada sobre el diseño y la implementación de una interfaz de software que permita el análisis filogenético de un gran número de secuencias genéticas que se encuentran almacenadas en una bases de datos, implementando para ello el algoritmo para análisis filogenético UPGMA (Unweighted Pair Group Method using Arithmetic averages).

Como resultado de este proyecto de grado, se generará una interfaz de software que no solo permite llevar a cabo el análisis filogenético de un gran número de secuencias genéticas, las cuales se encuentran almacenadas en una base datos, utilizando para ello el método para análisis filogenético UPGMA, sino que además incluye una función extra que permite a los investigadores obtener una predicción de asociación a partir de una secuencia de ADN entrante, utilizando para ello un método conocido como GraphDatabases.

Esta investigación hace parte de un proyecto de mayor envergadura que está siendo desarrollado de manera interdisciplinaria por los profesores investigadores

Dr. Eduardo Zurek, Dr. Guillermo Cervantes y Dr. Homero San Juan, y la Joven Investigadora de COLCIENCIAS Ingeniera Sandra Acero; dicho proyecto, busca diseñar un herramienta que permita a los investigadores llevar a cabo el análisis e interpretación de mutaciones del VIH para determinar susceptibilidad o resistencia a drogas.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

AGRADECIMIENTOS

“Seguiré con aguante la carrera que está puesta delante de mí”. Hebreos 12: 1B:

Agradezco a mi madre *YANETH CECILIA SUAREZ CABALLERO*, por haberme dado la vida y brindarme todo el apoyo y cariño necesario para afrontar los momentos más difíciles no solo de mi vida sino de mi carrera profesional. Agradezco además toda su dedicación y sacrificios para así poder brindarme todas las herramientas necesarias para afrontar los futuros retos que se me han de presentar y ser siempre un gran ejemplo de superación a seguir.

A mis abuelos *ODILA* y *TITO*, por todos sus cuidados y por procurar hasta el último día de su vida que me convirtiese en una persona de bien con grandes valores y principios morales.

A mis tías, tíos y demás familiares por estar ahí siempre brindándome su apoyo, amor y cariño de manera incondicional. A mis amigos, en especial a *SARAI REBECA*, por todo su apoyo y por convertirse en los hermanos que nunca tuve.
A mi novia *SUSANA* por convertirse en este último año en parte esencial de mi vida y estar ahí siempre que la necesite brindándome todo su cariño y amor.

Al profesor *EDUARDO ZUREK*, por toda su dedicación a esta tesis, ya que fue pieza fundamental para que los objetivos de esta se cumplieran a cabalidad y se entregara un producto de calidad. A los Doctores GUILLERMO CERVANTES y HOMERO SAN JUAN, por estar tan comprometidos con este proyecto y servirme de guías en la exploración de un área tan fascinante y a la vez tan compleja como lo es la medicina.

INTRODUCCIÓN

A lo largo de la historia de la humanidad se ha visto como esta ha sido afectada por un sin fin de enfermedades que han cobrado la vida de un gran número de personas. En los últimos 30 años el mundo se ha tenido que enfrentar al surgimiento de nuevas enfermedades cada vez más peligrosas como por ejemplo el ébola y la neumonía asiática, pero ninguna había sido tan mortal y había causado tanto impacto como lo ha hecho la aparición del Síndrome de la Inmunodeficiencia Adquirida (SIDA), enfermedad que es producida por el Virus de inmunodeficiencia humana (VIH). La causa de esto es básicamente que una vez el virus se instala en el huésped no puede ser eliminado y a que aún no existe una vacuna eficaz debido a la gran variabilidad viral.

Desde la aparición del VIH, alrededor de los años 80, se ha tratado desesperadamente encontrar una cura eficaz para este gran mal, a través de diversas investigaciones realizadas por la comunidad científica, pero los resultados no han sido del todo satisfactorios, debido a la alta variabilidad genética que presenta el virus.

A través de los estudios e investigaciones epidemiológicas que se han llevado a cabo en múltiples laboratorios de alto reconocimiento, se ha logrado determinar que existen diversos subtipos del virus, dependiendo del área geográfica donde éste se encuentre, lo que ha sido unos de los principales obstáculos para combatir esta enfermedad, debido a que sería necesario crear una vacuna para cada subtipo, dependiendo del área donde éste se esté desarrollando. Por esto es de vital importancia establecer de forma precisa el subtipo predominante en cada región del mundo, lamentablemente esto no ha sido posible en muchos países como por ejemplo Colombia, lo que es un tema de preocupación para la sociedad.

Este hecho ha despertado el interés de diversos grupos de investigación, como lo son el Grupo de Investigación en Biotecnologías y el Grupo de Virología y Patologías de la Universidad del Norte, que están llevando a cabo estudios para determinar el subtipo predominante es nuestro país. Para esto se está implementando una técnica denominada análisis filogenético, que les permita a los investigadores obtener patrones para realizar la clasificación de las muestras que obtengan.

El presente proyecto pretende colaborar con esa investigación llevando a cabo una extensión del algoritmo UPGMA (Unweighted Pair Group Method Using Arithmetic Averages) utilizado para realizar la categorización, implementando una base de datos que permita llevar a cabo el análisis filogenético de un gran número de secuencias genéticas.

CAPÍTULO 1: ESPECIFIACIONES DEL PROYECTO

1.1 PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN

La aparición del virus de inmunodeficiencia humana (VIH) ha significado un gran reto para la comunidad tanto científica como médica, debido a que este virus posee la capacidad de mutar, lo que lo hace difícil de combatir, debido a que de él se desprenden diversos subtipos que varían dependiendo de la región donde esta se encuentre.

En Colombia se pretende determinar de manera precisa el subtipo de VIH predominante. Una manera de llevar a cabo esto es implementando el algoritmo UPGMA (Unweighted Pair Group Method using Arithmetic averages) para construir árboles filogenéticos. Unos de los principales logros que pretende desarrollar este proyecto es el de mejorar significativamente los tiempos de respuesta del sistema cuando se requiere aplicar el algoritmo a un gran número de secuencias genéticas, ayudándose de una base de datos para la reutilización de los árboles filogenéticos que ya hayan sido generados.

1.2 OBJETIVOS

Este proyecto pretende servir de apoyo al proceso de clasificación de los distintos subtipos de VIH existentes en nuestro país. Dicha clasificación forma parte de un estudio realizado en el marco de un macro-proyecto que está siendo desarrollado por el grupo de Virología y Patologías Asociadas de la Universidad del Norte.

1.2.1 Objetivo general

Implementar el algoritmo UPGMA (Unweighted Pair Group Method using Arithmetic averages) para el análisis filogenético de un gran número de secuencias genéticas almacenadas en una base de datos.

1.2.2 Objetivos específicos

- Revisar el algoritmo UPGMA (Unweighted Pair Group Method using Arithmetic averages) implementado para el análisis de un pequeño número de secuencias genéticas, para así diseñar las estrategias computacionales y algorítmicas que se llevaran a cabo a la hora de desarrollar la nueva aplicación.

- Buscar la estructura de bases de datos más adecuada para el almacenamiento y recuperación de secuencias genéticas.

- Implementar una base de datos para el almacenamiento de secuencias genéticas que permita la reutilización de los árboles filogenéticos obtenidos, para así mejorar los tiempos de respuesta cuando se requiera analizar un gran número de cadenas genéticas.

- Diseñar un software basado en UPGMA (Unweighted Pair Group Method using Arithmetic averages) para que interactúe con la base de datos que contiene las secuencias genéticas.

- Validar la base de datos generada e implementar en el software diseñado después de cumplir el objetivo anterior.

1.3 ANTECEDENTES

Este desarrollo se fundamenta en proyectos de grado realizados dentro del Programa de Ingeniería de Sistemas de la Universidad del Norte. Los antecedentes más importantes son: la extensión de otro proyecto ya existente que hace parte de un macro-proyecto que lleva a cabo el grupo de Virología y Patologías Asociadas de la Universidad del Norte, que pretende llevar a cabo la clasificación de los distintos subtipos de VIH existentes en nuestro país, ayudándose para ello de un Software que implementa el algoritmo para análisis filogenético UPGMA (Unweighted Pair Group Method using Arithmetic averages) para la realización de dicha clasificación.

Proyectos de grados anteriores relacionados con esta investigación:

- ALGORITMO PARA ANÁLISIS FILOGENÈTICO: UPGMA (Unweighted Pair Group Method using Arithmetic averages). Autor: Sandra Milena Acero Barraza.
- ANÁLISIS FILOGENÉTICO DEL VIH BASADO EN EL MÉTODO NEIGHBOR – JOINING. Autor: Alejandro Fidel Pedrozo.
- ANÁLISIS FILOGENÉTICO UTILIZANDO EL MÉTODO “PARSIMONY”. Autor: Mat Max Montalvo.
- Análisis Filogenético del Virus de la Inmunodeficiencia Humana VIH a partir de virus aislados en una población del Caribe. Autor: Guillermo Cervantes

Publicaciones relacionadas con los proyectos anteriores:

- PEDROZO, Alejandro y ZUREK, Eduardo. Artículo: “Análisis Filogenético del VIH basado en el Método Neighbor-Joining”. Memorias del “V Encuentro Regional de Electrónica y Sistemas: ERES 2007”. Este artículo recibió el “PREMIO AL MEJOR TRABAJO EN EL ÁREA DE COMPUTACIÓN”

presentado en el evento. Universidad del Norte, Barranquilla, Colombia. Agosto 16, 17 y 18 de 2007.

- PEDROZO, Alejandro y ZUREK, Eduardo. Ponencia: “Análisis Filogenético del VIH basado en el Método Neighbor-Joining”. Presentada durante el “V Encuentro Regional de Electrónica y Sistemas: ERES 2007”. Universidad del Norte, Barranquilla, Colombia. Agosto 16, 17 y 18 de 2007.
- ZUREK, Eduardo, ACERO, Sandra y MONTALVO, Mat. Conferencia: “Secuencias Genéticas y Algoritmos”. Presentada durante el “V Encuentro Regional de Electrónica y Sistemas: ERES 2007”. Universidad del Norte, Barranquilla, Colombia. Agosto 16, 17 y 18 de 2007.

Proyectos paralelos:

- Desarrollo e Implementación de un Software que permita el Análisis e Interpretación de Mutaciones del VIH para Determinar Susceptibilidad o Resistencia a Drogas. Autor: Sandra Acero.
- "Desarrollo de una Aplicación de Unificación de Implementaciones de Algoritmos Filogenéticos". Tesis de pregrado, Departamento de Ingeniería de Sistemas, Universidad del Norte. Autores: Melvin David Faillace Teuta y Jonathan Rudolf Montalvo Alcázar. Director: Dr. Eduardo Zurek.

1.4 ENTIDADES INTERESADAS

- Departamento de Ciencias Básicas Médicas de la Universidad del Norte
- Departamento de Ingeniería de Sistemas de la Universidad del Norte
- Colciencias
- Grupo de investigación en biotecnologías
- Grupo de Virología y patologías asociadas

1.5 HIPOTESIS DE TRABAJO

La hipótesis central de este proyecto afirma que es posible implementar una base de datos que interactuando con una aplicación del algoritmo de análisis filogenético UPGMA (Unweighted Pair Group Method using Arithmetic averages), permita la utilización de árboles filogenéticos ya generados para optimizar el análisis de grandes cantidades de secuencias genéticas.

Como resultado de los desarrollos enfocados a la demostración de esta hipótesis, se generará una herramienta de soporte para la clasificación de los distintos subtipos de VIH existentes en la Región Caribe Colombiana. Esta herramienta permitirá mejorar significativamente los tiempos de respuestas del sistema, cuando este requiera analizar grandes secuencias de ADN. Permitiendo así a los investigadores del Grupo de Virología y patologías asociadas de La Universidad del Norte y otras entidades interesadas, obtener información importante de la clasificación hecha a un gran número de secuencias. El posterior análisis e investigación de los datos obtenidos permitirá llevar a cabo avances significativos que ayuden a determinar el subtipo de VIH predominante en nuestra región.

1.6 JUSTIFICACION

El uso del análisis filogenético como herramienta de investigación en la Biología ha sido de vital importancia, ya que a través de ella se han conseguido importantes logros científicos, entre los que tenemos el haber logrado identificar un gran número de subtipos del VIH. A través de este proyecto de grado se pretende colaborar en dichas investigaciones, llevando a cabo el diseño y desarrollo de un software que implemente el algoritmo para análisis filogenético UPGMA (Unweighted Pair Group Method using Arithmetic averages) que permita a los investigadores analizar

grandes cadenas de ADN y reutilizar árboles filogenéticos aplicando técnicas de bases de datos, que les permita la clasificación y estudio de estas distintas variantes del VIH.

1.7 PLAN DE TRABAJO

1. Establecimiento del marco teórico sobre los estudios e investigaciones referentes a la evolución del VIH y al método para análisis filogenético UPGMA (Unweighted Pair Group Method using Arithmetic averages).

2. Estudio del lenguaje de programación Python, en el cual se desarrollara la aplicación.

3. Estudio del lenguaje de consulta estructurado SQL.

4. Estudio de la interacción entre Python y SQL.
5. Búsqueda y escogencia de la estructura de bases de datos que mejor se adapte al sistema y que va a ser utilizada para el almacenamiento y recuperación de los árboles filogenéticos.

6. Diseño del algoritmo basado en el método para análisis filogenético UPGMA (Unweighted Pair Group Method using Arithmetic averages) que permita trabajar con un gran número de secuencias genéticas.

7. Desarrollo de la implementación basada en el método para análisis filogenético UPGMA (Unweighted Pair Group Method using Arithmetic averages) que interactúe con el modelo de bases de datos escogido.

8. Llevar a cabo pruebas en la implementación, utilizando grandes secuencias de genéticas.

9. En paralelo con las actividades anteriores, se elaborará la monografía del proyecto de grado.

1.8 METODOLOGIA

1.8.1 Tipo de estudio

Debido a que en Colombia son escasas las investigaciones que conlleven a determinar el subtipo de VIH predominante, el presente proyecto tiene como fin desarrollar una herramienta que le permita a los investigadores y demás interesados en el tema, llevar a cabo dicha clasificación cuando se cuente con un gran número de cadenas genéticas.
Por lo anterior se adopta un tipo de estudio exploratorio, el cual permitirá el desarrollo de un proyecto que se convierte en un aporte al campo de las investigaciones referentes al VIH.

1.8.2 Método de investigación

El método de investigación a utilizar es del tipo síntesis, debido a que después de haber recopilado la información necesaria, y haber desarrollado el marco teórico, se procederá a interrelacionar todos los datos obtenidos tanto en el área de la Biología y de la Virología, como en el área de la computación, para así diseñar e implementar el sistema de información que se necesita.

1.8.3 Técnicas de recolección de información

La información que será utilizada para el desarrollo de este proyecto, será suministrada primeramente por el grupo de Grupo de Virología y patologías asociadas de la Universidad del Norte y el grupo de investigación en biotecnologías.

Por otra parte también se obtendrá información de artículos y/o boletines publicados por entidades y universidades reconocidas que estén adelantando investigaciones referentes al tema.

1.8.4 Pasos metodológicos

1. Se hará un compendio de información pertinente al tema. La información se obtendrá de: las entidades interesadas, las bibliografías de consulta, fuente y soporte, y la interacción con el director y los asesores de la tesis.

2. Se procederá a realizar un estudio y análisis de toda la información obtenida, para así explotar de manera eficaz la más importante y relevante para el desarrollo del proyecto.

3. Se estudiará el algoritmo para análisis filogenético UPGMA (Unweighted Pair Group Method using Arithmetic averages), para así determinar que estrategias algorítmicas y computacionales se llevaran a cabo con él, para el diseño de la implementación.

4. Se implementará el algoritmo en la plataforma Python para que este interactúe con una base de datos que permitirá la reutilización de los árboles filogenéticos.

5. Se realizarán pruebas para comprobar el correcto funcionamiento de la implementación.

6. Se elaborará el manual necesario para guíe al usuario final en la forma en como este debe interactuar con el sistema para obtener los resultados que busca.

7. Por último se llevará a cabo la elaboración de la monografía, para mostrar tanto la investigación realizada como los resultados obtenidos.

CAPÍTULO 2: BIOINFORMATICA

Esta disciplina se interesa principalmente del análisis, interpretación y procesamiento de datos mediante el uso de herramientas computacionales de información obtenida de procesos biológicos. Según la definición del Centro Nacional para la Información Biotecnológica "National Center for Biotechnology Information" (NCBI por sus siglas en Inglés, 2001): *"Bioinformática es un campo de la ciencia en el cual confluyen varias disciplinas tales como: biología, computación y tecnología de la información. El fin último de este campo es facilitar el descubrimiento de nuevas ideas biológicas así como crear perspectivas globales a partir de las cuales se puedan discernir principios unificadores en biología"*.[1]

La bioinformática se ha consolidado como una nueva disciplina dentro del área de la Biología y esto ha permitido que esta amplíe su rango de estudio, debido a que su utilización ha ido mucho más allá del manejo y análisis de bases de datos biológicas, y se ha centrado en fusionar paralelamente técnicas computacionales con los diversos procesos de análisis e investigación que existen dentro de la Biología. Surge debido a los múltiples interrogantes que se generaron a raíz del estudio del genoma humano y de cómo la puesta en marcha de la investigación genómica puede ayudar dramáticamente a mejorar la condición y calidad de vida humana.[2] Hoy en día es considerada una de las principales herramientas utilizadas por médicos e investigadores, para lograr importantes avances en el campo médico.

[1] NCBI. http://www.ncbi.nlm.nih.p-ov/AbouUprimer/bioinformatics.html. Fecha de consulta: 11 de marzo de 2007.

[2] http://www.solociencia.com/biologia/bioinformatica-concepto.htm. Fecha de consulta: 12 de marzo de 2007.

Esta disciplina comprende múltiples áreas de investigación, entre las principales tenemos:

- Predicción de estructuras proteicas: Esta área se encarga de la predicción de la estructura tridimensional de las proteínas, que es conocida como estructura terciaria, lo cual se hace a partir de la secuencia de aminos ácidos[3].

- Análisis de Secuencias Genéticas (Genomics): Esta área se encarga del análisis de grandes secuencias genéticas que han sido decodificadas y almacenadas en poderosas bases de datos, para su posterior estudio que permita establecer posibles similitudes entre ellas[4].

- Medición de la biodiversidad: La biodiversidad es definida como la totalidad de los genes, las especies y los distintos ecosistemas de una región[5]. Por lo tanto esta área se encarga del estudio y diseño de grandes bases de datos, las cuales son capaces de brindar importante información acerca de determinada especie o población, o de diseñar modelos de alta precisión que permitan determinar la dinámica poblacional de una especie[6].

- Biología Sistémica (System Biology): Esta área se encarga del estudio de subsistemas celulares, mediante la simulación de estos para poder llevar a

[3] http://es.wikipedia.org/wiki/Predicci%C3%B3n_de_estructura_de_prote%C3%ADnas. Fecha de consulta: 21 de marzo de 2008. 12:09 a.m.
[4] http://www.merriam-webster.com/dictionary/genomics. Fecha de Consulta: 20 de marzo de 2008. 1:35 p.m.
[5] Gonzalo Halffter y Exequiel Ezcurra. Que és biodiveridad?. Disponible en: http://web.minambiente.gov.co/biogeo/menu/biodiversidad/bioespa.htm. Fecha de consulta: 21 de marzo de 2008. 10:30 p.m.
[6] CLAUDIA E. MORENO. Métodos para medir la biodiversidad. M&T – Manuales y Tesis SEA, vol. 1. Primera Edición: 2001. Zaragoza, 84 pp.

cabo un profundo análisis de las complejas conexiones que existen entre ellos[7].

La bioinformática hace uso de otras disciplinas para abarcar diversos campos como son la Medicina molecular y la Biotecnología, las cuales buscan principalmente ampliar la investigación genómica.

- MEDICINA MOLECULAR

La medicina molecular es una ciencia que lleva a cabo la utilización de unas técnicas denominadas técnicas de ADN recombinante, las cuales han permitido a los biólogos moleculares la manipulación de los ácidos-nucleicos. Su investigación está centrada más que todo en la identificación y clasificación de genes y mutaciones responsables de diversas enfermedades y en la comprensión de los mecanismos patofisiológicos moleculares[8].

Gracias a los diversos avances que ha proporcionado la Medicina molecular en conjunto con la investigación biomédica, hoy en día contamos con diversas pruebas, las cuales a partir de muestras de ADN y ARN, permiten identificar mutaciones patogénicas y ofrecer un diagnóstico mucho más preciso en cuanto al curso clínico de diversas patologías. Por otra parte los conocimientos que se han obtenido a partir de investigaciones en esta área, han impulsado el planteamiento de diversas hipótesis entre las que se encuentra la hipótesis del origen genético del cáncer[9].

[7] HIROAKI KITANO. Systems Biology: A Brief Overview. 1 de Marzo de 2002. 3 pp.
[8] SIMON KAWA KARASIK. Medicina Molecular. [online]. Disponible en: http://www.medigraphic.com/espanol/e-htms/e-h-gea/e-gg2000/e-gg00-2/em-gg002e.htm. Fecha de consulta: 20 de Marzo de 2008. 12:49 a.m. 2 Páginas.
[9] SIMON KAWA KARASIK. Medicina Molecular. [online]. Disponible en:

- BIOTECNOLOGÍA

La biotecnología es una ciencia que se encarga del uso y la manipulación de organismos vivos y de sus componentes para la fabricación de productos encaminados al beneficio del ser humano. Entre sus principales logros se tiene la producción de la penicilina y de la insulina humana, medicinas de vital importancia para el mejoramiento de la calidad de vida de muchas personas.

Esta ciencia ha contribuido de manera importante a la prevención de enfermedades infecciosas, gracias por ejemplo a la manipulación genética efectuada de algunas bacterias y levaduras que permiten su posterior transformación con los genes que codifican para proteínas de interés farmacológico. Es así como la Biotecnología ha desarrollado poderosas vacunas, que permiten a los individuos desarrollar los anticuerpos necesarios para prevenir una posible infección. Dos ejemplos destacados de esto es la obtención de la vacuna para la hepatitis B y la rabia[10].

2.1 PROYECTO DEL GENOMA HUMANO Y LAS BASES DE DATOS GENÉTICAS

El proyecto del genoma humano es un programa a nivel internacional que busca obtener un conocimiento completo de toda la información genética humana que se encuentra en las células del cuerpo y codificada en nuestro ADN[11].

Los principales objetivos del proyecto son los de lograr identificar los genes que hacen parte del núcleo de la célula humana, además poder determinar su

---

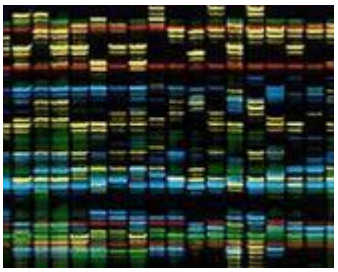
http://www.medigraphic.com/espanol/e-htms/e-h-gea/e-gg2000/e-gg00-2/em-gg002e.htm. Fecha de consulta: 20 de Marzo de 2008. 12:49 a.m. 2 Páginas.

[10] Enciclopedia Encarta 2006. Biotecnología. Fecha de consulta: 21 de Marzo de 2008. 11:49 p.m.

[11] DI MARIO GABRIEL.2005. Proyecto del genoma humano.Ilustrados.com. Fecha de Consulta: 21 de Agosto de 2007.

localización a nivel de los cromosomas y lograr determinar mediante unos procesos conocidos como secuenciación y cartografía, la información genética codificada por el orden de las subunidades químicas de nuestro ADN. El resultado es un mapa que contiene toda la organización tanto de los genes como de los cromosomas, que podría lograr prevenir o tratar diversas enfermedades y conocer cuáles son los factores bioquímicos que las desarrollan[12].

Figura 1. Mapa genoma humano [13]



Este proyecto en sus múltiples faces, ha logrado exitosamente identificar 25.000 genes que se encuentran en el núcleo de la célula humana y ha logrado determinar la localización de ellos en los 23 pares de cromosomas que hay en el núcleo. Por lo general las muestras de sangre o tejido utilizados en la investigación, provienen de personas anónimas, sin que esto afecte los datos, debido a que la diferencia entre muestras de dos individuos es de entre 0.05% y 0.1%[14].

Debido a la gran variedad de proyectos y estudios que se están llevando a cabo sobre el genoma humano, día a día se obtiene más información acerca de ellos, lo que ha generado grandes volúmenes de datos y que ha conllevado a crear lo que hoy es conocido como bases de datos genéticas, las cuales tienen como función almacenar toda esa información no solo para su conservación, sino también para

[12] DI MARIO GABRIEL.2005. Proyecto del genoma humano.Ilustrados.com. Fecha de Consulta: 21 de Agosto de 2007.
[13] http://www.latercera.cl/medio/articulo/0,0,38035857__147605214__1,00.html. Imagen tomada el 22 de Marzo de 2008.
[14] ALCALÁ O. MARIA A. MARVELIS RONDÓN. 2005. Genoma Humano. Ilustrados.com. Fecha de consulta: 23 de Agosto de 2007.

facilitar su gestión y análisis, que entre otras cosas es una de los objetivos principales de la bioinformática.

Existen diversas bases de datos genéticas, cada una con distintos tipos de información. Algunas son utilizadas para el almacenamiento de secuencias de ADN, mientras que otras almacenan datos de mutaciones genéticas y demás. Todo esto con el fin de que los investigadores cuenten con una poderosa herramienta que les permita relacionar todo este tipo de información, para así tener una perspectiva clara ay concisa sobre alguna enfermedad genética o de un gen en particular.

Una desventaja que poseen estas herramientas de almacenamiento genético, consiste en que es necesario que el investigador o persona que necesite algún tipo de información de ellas, conozca muy bien su estructura funcional, que tipo de datos contiene y que clases de palabras claves puede utilizar para lograr acceder a los datos que le sean relevantes. Entre las principales bases de datos genéticas tenemos, DNA Database of Japan (DDBJ), dbSTS, GOLD (Genomes Online Databases) y Unigene[15].

[15] www.**ddbj**.nig.ac.jp/. Fecha de consulta: 3 e Diciembre de 2007.

CAPÍTULO 3: MÉTODO PARA EL ANÁLISIS FILOGENÉTICO UPGMA

A continuación se hará una descripción del funcionamiento del método para el análisis filogenético UPGMA (Unweighted Pair Group Method using arithmetic Averages).

Primero se procederá a explicar de manera detallada cada uno de los aspectos principales del algoritmo, luego se plantea la estructura del mismo, y por último se procede a mostrar un ejemplo que ilustrará cada uno de los pasos que se deben llevar cabo a la hora de ser implementado.

3.1 DESCRIPCIÓN DEL ALGORITMO

El método para el análisis filogenético UPGMA (Unweighted Pair Group Method using arithmetic Averages), es un algoritmo heurístico que fue definido por los Doctores Peter H. A. Sneath y Robert R. Sokal en el año de 1973[16].

Dicho método consiste principalmente en el cálculo de una matriz de distancia, en la cual se lleva a cabo la búsqueda de la distancia más pequeña que se haya generado para así relacionar los dos grupos de especies más cercanos. Una vez realizada esta operación, se procede a re calcular la ya mencionada matriz de distancias y todo el proceso es repetido hasta que todas las especies se encuentren relacionadas a un único grupo[17].

---

[16] SANDRA ACERO. ALGORITMO PARA ANÁLISIS FILOGENÈTICO: UPGMA (Unweighted Pair Group Method using Arithmetic averages). Tesis de grado. 2007.

[17] DAN E. KRANE, MICHAEL L. RAYMER. Fundamental Concepts of Bioinformatics. Pearson Education. 2003

El UPGMA basa todos sus cálculos en la matriz de distancia que genera a partir de las secuencias genéticas que recibe, dicha matriz, es una *matriz* cuadrada cuyo tamaño en *n,* dónde *n* representa el número de secuencias a clasificar y la *matriz[i][j]* es la distancia genética que hay entra la especie *i* y la especie *j*. Cuando este método lleva a cabo el re calcular la matriz de distancias, emplea para ello la ejecución de una fórmula denominada la Fórmula de Tajima, la cual será explicada posteriormente a profundidad y la cual fue objeto de estudio y de modificación para así lograr que el UPGMA fuese capaz de trabajar con un gran número de secuencias genéticas de gran longitud.

La matriz de distancia obtenida y en la cual se basan todos los cálculos que se realizan, debe cumplir tres requisitos para ser válida:

- La matriz de distancias debe ser métrica. Una matriz es métrica si satisface la siguiente condición:
  - Simetría: $matriz_{ij} = matriz_{ji}$ y $matriz_{ii} = 0$ [18]
  - Desigualdad triangular: $matriz_{ij} + matriz_{jk} \geq matriz_{ik}$

- La matriz de distancias debe ser métrica aditiva. Una matriz es métrica aditiva si satisface la condición de que existe un árbol donde:
  - Cada una de las ramas tiene un peso positivo y cada hoja corresponde a una especie distinta.
  - $\forall i, j, 1 \leq i \leq n, i < j \leq n$, $matriz_{ii}$ es la suma de los pesos de las ramas desde la hoja *i* hasta la hoja *j.*

[18] NING K., SHAN T., XIANG S. L., SHEN W. Phylogenetic Tree Reconstruction: Distance Based. . [online]. Octubre 10 de 2003. Disponible en: <http://www.comp.nus.edu.sg/~bioinfo/phylogenetic%20tree%20reconstruction_2_8.pdf> . Fecha de consulta: 22 de Junio de 2007. 20 páginas.

- La matriz de distancias debe ser ultramétrica. Una matriz es ultrametrica si se cumple la siguiente condición:
    - La raíz del árbol que se forma a partir de la matriz de distancia, y de todos los sub árboles que contiene, es tal que la suma de todos los pesos de las ramas salientes de esta es la misma.

3.2 ALGORITMO UPGMA

Entrada:

n secuencias de ADN o proteína las cuales son denominadas OTUs.

Salida:

El fenograma que representa la relación evolutiva existente entre las n OTUs.

Algoritmo:

Sea $OTUS = \{t1, t2, t3, ..., tn\}$ el conjunto de las n secuencias, donde cada ti representa una especie.
Sea $MatDistancia$ una matriz cuadrada de tamaño n.
Repetir para $i = 1, ..., n$
  Repetir para $j = i, ..., n$
    Calcular $dist(ti, tj)$
    $MatDistancia[i][j] = dist(ti, tj)$
    $MatDistancia[j][i] = dist(ti, tj)$
    $MatDistancia[i][i] = 0$
  Fin repetir
Fin repetir
Repetir n-1 veces
  Buscar ti, tj tal que $dist(ti, tj)$ en $MatDistancia$ sea mínima.
  Hacer $f = i, c = j$.
  Definir un nuevo $tk = ti \cup tj$. En $OTUS$
  Reemplazar ti y tj por tk
  Recalcular la matriz de distancia.

```
  Hacer n = n − 1, tamaño de matDistancia igual a n
  Repetir para i =,...,n
    Repetir para j = i,...n
      Si i ≠ f ∧ i ≠ c ∧ j ≠ f ∧ j ≠ c :
        matDistancia[i] = dist(i, tj)
      Fin Si
      Si j = f :
        matDistancia[i] = (dist(i, tf) + dist(i, tc)) / 2
      Fin si
      Si i = f :
        matDistancia[i] = (dist(c, tj) + dist(f, tj)) / 2
      Fin si
      matDistancia[i] = matDistancia[i]
      matDistancia[i] = 0
    Fin repetir
  Fin repetir
  Hacer distancia(k) = dist(f, tc) / 2
  Crear Subárbol uniendo rama tf con rama tc, con una distancia
  de dist(k)[1]
Fin repetir
```

3.3 GENERACIÓN DEL ÁRBOL FILOGENÉTICO

El árbol producto del análisis filogenético se conoce con el nombre de fenograma, dicho árbol para ser válido debe cumplir la condición de ser ultramétrico. Su función consiste en representar la relación de evolución existente entre un grupo de OTUs basándose en la similitud entre ancestros comunes hasta llegar a uno que sea común para todos.

En cada una de las hojas que conforman el árbol filogenético, se encuentran las partes de las secuencias que diferencian cada OTU de las demás, y habrán tantas hojas como secuencias estemos analizando. Las hojas pueden estar asociadas a otras hojas o a otros subárboles y estos a la vez a otros subárboles, hasta completar el fenograma, donde cada uno estará etiquetado con los ancestros que tengan en común y que han sido determinados a medida que se va ejecutando el método.

Figura 2. Árbol Filogenético

3.4 EJEMPLO

A continuación se desarrollará un ejemplo para describir paso a paso, la ejecución del método y los distintos cálculos que este ejecuta. Para este ejemplo se trabajará solo con 4 secuencias de longitud 108, ya que este solo es un ejemplo de carácter ilustrativo.

A las distintas matrices de distancias que vayan surgiendo se les aplicará la fórmula de Tajima, la cual fue mencionada anteriormente, pero por cuestiones de simplicidad, ese paso no será mostrado aquí, pero más adelante será ilustrado, ya que aquí radica la clave, para que el UPGMA sea capaz de trabajar con grandes secuencias genéticas.

Sean las siguientes secuencias:

1. GGTGGTACTTAAAATAAAGTTAACAATTACATTTTTTTTTTTTTCGTTTTCCAAACGTCTTTAATAAGTAAATAAAGGAGCAATGTAAAATGTTGCAGTTTGCTTGGT
2. CATGGGATGAATTTTAAGATTCAGAAATATCCTTTACTTACATTGTTTTGTTTTTTAAAACTCTCTTCTAGGTCTACTTGAAGATTTTTTTTCTTCGTTAAGGTTCAA
3. AACTATTAAATTTTAATCTTGACAGTTTTTACATATCCATGAGTGTTTTTATTTAATCAAAGTATCCTTTTCCGACATCTTAAAATTATTTTTATGAGTTTATGATC
4. TATCATTTCAGAAAACTGTTGCATCAAATAATATACAACCAGGTATCAGTATGAAAAAGGATCTTTGTTCATCACTATTTCTTACAAATAAAATAACAAATAAATGA

- Primer paso: Crear la matriz de distancia

Se compara la primera secuencia con todas las demás, para determinar que partes de ella no tienen en común y así determinar la primera distancia de la matriz, y así sucesivamente hasta llegar a la última secuencia.

1. **GG**TGG**T**A**CTT**A**AAA**TAA**AG**TT**AA**C**AA**T**TA**CAT**TTT**TT**TT**TTT**T**CG**TTTT**CCAAACGTCTTTAA**T**AAGTAA**ATAAAG**
GAGCAA**TG**T**AAAA**T**GT**TG**C**AG**T**TT**G**C**TT**GGT**
2. **CA**TGG**G**A**TGA**A**TTT**TAA**GA**TT**C**A**G**AA**A**TA**TCC**TTT**AC**TT**ACA**T**TG**TTTT**GTTTTTT**AAAACTC**T**CTTCT**A**GGTCTA**
CTTGAA**GA**T**TTTT**T**TC**T**TC**G**T**T**AA**G**GT**T**CAA**

$$dist(1,2) = \frac{70}{107} = 0.65$$

1. **GGTGG**T**ACTTAAAATAAAG**T**T**AACAA**TT**AC**A**TTTTTTTTT**TTTC**GTTTT**CCAAACG**TC**TTTAATAAG**T**AAATAA**AG**
GAGCAA**TGTAA**AA**TG**TT**GCAGT**TT**GCT**TGGT**
3. **AACTA**T**TAAATTTTAATCT**TGA**CAGT**TTT**TT**A**CA**TA**TCCA**T**GAGT**GTTTTT**TATTTAA**TC**AAAGTATCC**T**TTTCCG**AC
ATCTTA**AAATT**A**T**TTT**ATGAG**TT**TA**T**GATC**

$$dist(1,3) = \frac{81}{107} = 0.75$$

Tabla 1. Matriz de distancia por factor de corrección Tajima

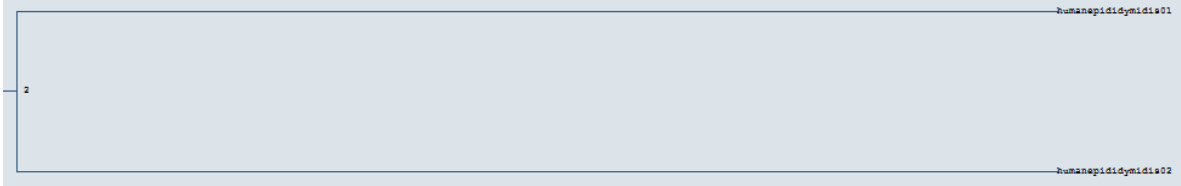
OTU	1	2	3	4
1	0	1.41314284055	2.55495311954	3.03328942723
2	**1.41314284055**	0	1.5385217862	2.21950312243
3	2.55495311954	1.5385217862	0	1.5385217862
4	3.03328942723	2.21950312243	1.5385217862	0

- Segundo paso: Hallar la menor distancia en la matriz
  - Menor distancia: 1.41314284055

- Tercer paso: Unir las OTUs de menor distancia, en este caso la 1 y 2.

$$\text{Distancia} = \frac{dist(1,2)}{2} = \frac{1.4131428455}{2} = 0.7065714027$$

Figura 3. Subárbol resultante de unir OTUS 1 y 2



- Cuarto paso: Re calcular la matriz de distancias del primer paso, una vez unidas las dos primaras OTUs.

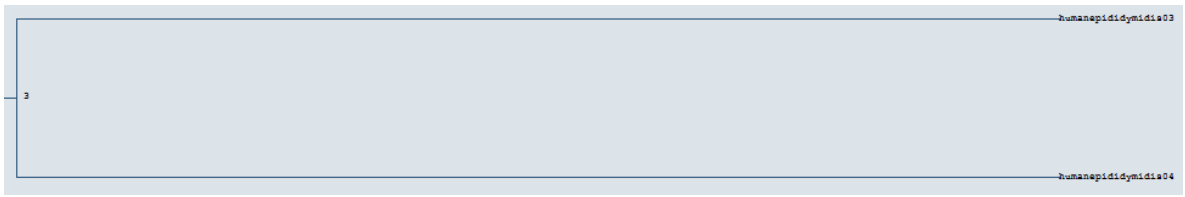
Tabla 2. Matriz de distancia por factor de corrección Tajima- Cuarto paso

OTU	12	3	4
12	0	2.04673745287	2.62639627483
3	2.04673745287	0	1.5385217862
4	2.62639627483	**1.5385217862**	0

- Quinto paso: Repetir el paso 2, hasta que todas las OTUs estén relacionadas con el ancestro en común.

  - Menor distancia: 1.5385217862
  - OTUS a unir: 3 y 4

Figura 4. Subárbol resultante de unir OTUS 1 y 2



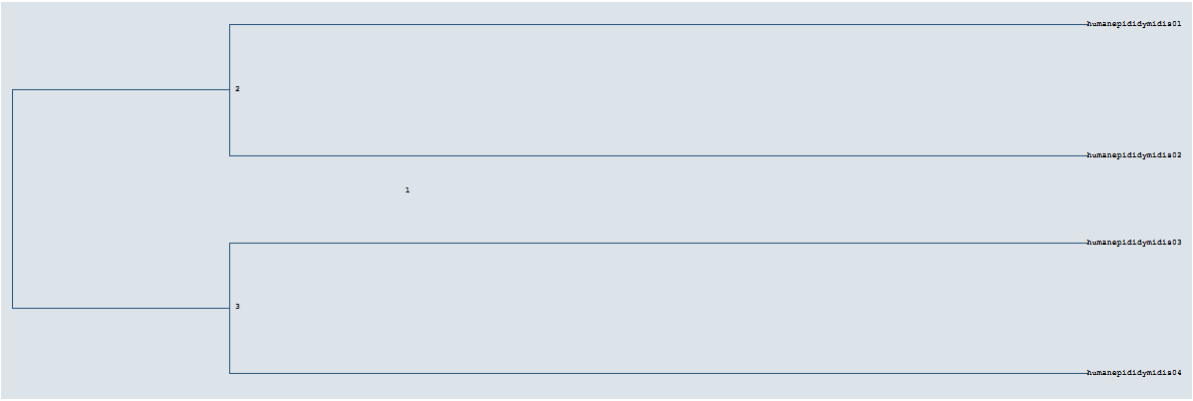
- Nuevamente se re calcula la matriz de distancias:

Tabla 3. Matriz de distancia por factor de corrección Tajima

OTUs	12	34
12	0	**2.33656686385**
34	2.33656686385	0

- Menor distancia: 2.33656686385
- OTUs a unir: 12 y 34

Figura 5. Subárbol final



CAPÍTULO 4: IMPLEMENTACIÓN DEL ALGORITMO

En el capítulo anterior se dio una descripción del funcionamiento básico del algoritmo UPGMA, así como un ejemplo ilustrativo del mismo, para brindar mayor claridad y entendimiento sobre la manera cómo este opera. A continuación se presentan las distintas estrategias tanto algorítmicas como computacionales que se llevaron a cabo para lograr que la aplicación fuese capaz de procesar un gran número de secuencias genéticas.

4.1 ETAPAS DE DESARROLLO

El desarrollo del algoritmo se llevo a cabo en 6 etapas hasta obtener el resultado deseado:

- Primera etapa: Creación y montaje de la bases de datos que contiene las secuencias genéticas que serán analizadas.
- Segunda etapa: Creación del módulo de conexión, que permitirá la interacción entre la aplicación y la base de datos.
- Tercera etapa: Diseño e implementación de una fórmula que fuese la equivalente a la planteada por Tajima, para lograr que el algoritmo soporte el análisis de un gran número de secuencias genéticas.
- Cuarta etapa: Creación del módulo para la formación del GraphDatabases, el cual será explicado con detalle en el capítulo 6.
- Quinta etapa: Creación de un área de dibujo óptimo para el despliegue del árbol filogenético generado.
- Sexta etapa: Diseño de la nueva interfaz gráfica con todos sus elementos.

4.1.1 Creación de la bases de datos

Debido a que es necesario el manejo de un gran volumen de información, y mejorar la disponibilidad de los datos, fue necesaria la creación de una base de datos que permite el almacenamiento sistemático de las secuencias genéticas para su posterior utilización.

Dicha base de datos fue diseñada en MySQL, que es un sistema de gestión de base de datos relacional, multihilo y multiusuario con más de seis millones de instalaciones[19], que opera bajo licencia dual GNU y GPL y que es de fácil comprensión y manejo para usuarios inexpertos. La principal razón por la que se decidió trabajar con este sistema de gestión de datos, fue porque este es capaz de interactuar con el lenguaje de programación Python, plataforma sobre la cual fue diseñada la aplicación.

La base de datos se encuentra estructurada en dos campos:

- Campo de descripción: El cual contiene el nombre de la secuencia genética.
- Campo de la secuencia: El cual contiene la secuencia en sí.

Ejemplo:

Tabla 4. Tabla en MySQL con su respectiva secuencia genética

Descripción	Secuencia
human epididymidis	**GGTGG**T**ACTTAAAATAAAG**TT**AACAA**TT**AC**A**TTTTTTTTTTC**GTTTT**CCAAACG**TC**TTTAATAAG** T**AAATAA**A**GGAGCA**A**TGTAA**AA**TG**TT**GCAGT**TT**GC**T**TGGT**

[19] <http://dev.mysql.com/tech-resources/articles/dispelling-the-myths.html>. Fecha de consulta: 11 de marzo de 2008.

4.1.2 Creación del módulo de conexión

Una vez creada la base de datos se procedió a desarrollar el módulo de conexión, que permitiría la captura de los datos que serían procesados en el algoritmo UPGMA.

Para llevar a cabo este paso fue necesaria la implementación de una librería conocida como MySQL-python-1.2.2.win32-py2.5[20], de uso gratuito distribuida por SourceForge.net, y que tiene la responsabilidad de permitir la interacción entre el módulo de conexión creado en Python y la bases de datos de MySQL.

A continuación se muestra la implementación de la conexión con la bases de datos en Python.

```
def conectar(self):
    """Función para conectarse a la bases de datos"""
    #Crea la lista líneas que contiene los registros leídos de la bases de datos
     self.registro=[]
     self.OTUS=[]
     db=MySQLdb.connect(host='localhost',user='root',passwd='',db='upgma')
     cursor=db.cursor()
     sql="SELECT secuencia FROM informacionsecuencias"
     cursor.execute(sql)
     resultado=cursor.fetchall()

     cursor.close ()
     db.close()


        for self.registro in resultado:
        self.OTUS.append(self.registro[0])

     return self.OTUS
```

[20] http://alexandria.wiki.sourceforge.net/What+is+SourceForge.net%3F

4.1.3 Diseño e implementación de una aproximación numérica a la fórmula a la de Tajima

Como se mencionó en el capítulo anterior, para que se lleve a cabo la ejecución del método UPGMA, es necesario que este reciba un conjunto de secuencias genéticas conocidas como OTUs (Unidad Taxonómica Operativa). Cada una de estas secuencias está conformada por una serie de caracteres llamados nucleótidos, los cuales están representados cada uno por una letra distinta, A (Adeline), C (Cytosine), G (Guanine) y T (Thymine).

Una vez se cargan las secuencias genéticas, en este caso de la base de datos, se procede a calcular la matriz de distancias que es el cálculo de la distancia genética que existe entra cada una de las OTUs que se estén analizando. Comúnmente para llevar a cabo el cálculo de esta matriz de distancia, es necesario emplear una técnica conocida como método de Hamming, que consiste básicamente en contar el número de cambios que es necesario realizar para que una secuencia sea igual a otra y dividir entre el total de nucleótidos. A continuación se muestra un ejemplo del método de Hamming

Ejemplo:

1. **GG**TGG**T**A**CTT**A**AAA**TAA**AG**TT**AA**C**AA**T**TA**CAT**TTT**TT**TT**TTT**T**CG**TTTT**CCAAACGTCTTTAA**T**AAGTAA**ATAAAG **GAGC**AA**TG**T**AAAA**T**GT**TG**C**AG**T**TT**G**C**TT**GGT**
2. **CA**TGG**G**A**TGA**A**TTT**TAA**GA**TT**C**A**G**AA**A**TA**TCC**TTT**AC**TT**ACA**T**TG**TTTT**GTTTTTTAAAACTCT**C**TTCT**A**GGTCTA **CTTG**AA**GA**T**TTTTT**TC**TT**C**GT**T**AA**GG**TT**CAA**

$$dist(1,2) = \frac{70}{107} = 0.654205607476355$$

Como se puede apreciar, para que la secuencia 1 sea igual a la secuencia 2, es necesario que en esta se lleve a cabo el cambio de 70 nucleótidos y luego dividirlo

entre el total de estos, en este caso 107, así que la distancia de Hamming entre la secuencia 1 y 2 es de 0.6542056074766355.

Pero existe un inconveniente, y es que este método en algunos casos falla, ya que cuando se está llevando a cabo el análisis de las secuencias se puede presentar el hecho de que la adquisición de un mismo carácter en dos nucleótidos no es causa de una descendencia común y esto conlleva a que se den mutaciones múltiples en el nucleótido o conjunto de nucleótidos. A este fenómeno se le conoce con el nombre de Homoplasia[21].

Para superar este inconveniente se implementó una fórmula resuelta por Jukes y Cantor[22], la cual era un factor de corrección:

$$k(A,B) = -\frac{3}{4}\log_e\left[1-\frac{4}{3}\overbrace{dist(A,B)}\right]$$

Donde $k(A,B)$ es la distancia corregida y $\overbrace{dist(A,B)}$ es la distancia de Hamming.

Esta fórmula es válida para $\overbrace{dist(A,B)} < 0.75$, ya que de no ser así no se puede llevar a cabo el cálculo del logaritmo. Para superar este impase Tajima y Nei plantearon una nueva fórmula:

$$D(A,B) = \sum_{i=1}^{k} \frac{k^{(i)}}{i\left(\frac{3}{4}\right)^{i-1} m^{(i)}}$$ [23]

Donde:

[21] http://attila.inbio.ac.cr:7777/pls/portal30/INBIO_BIODICTIONARY.DYN_WORD_DETAIL.show?p_arg_names=_show_header&p_arg_values=YES&p_arg_names=pTermino&p_arg_values=Homoplasia. Fecha de consulta: 12 de marzo de 2008.

[22] JUKES, T.H., y C.R. CANTOR. 1969. Evolution of Proteins Molecules. Pp 21-132 en H:N: MUNRO, ed. Mammalian protein metabolism. Academis Press, New York.

[23] TAJIMA, Fumio. MSATOCHI, Nei. Unbiased Estimation of Evolutionary Distance between Nucleotide Sequences. Department of population Genetics, Nacional Institute of genetics. University of Chicago. Mol. Biol. Evol. 10(3). Pp. 677-688. Octubre 6 de 1993.

- k es el número total de pares bases que son diferentes en las dos secuencias y
- m es el total de pares de las secuencias
- $k = m - dist(a,b)$, por lo tanto: $k^{(i)} = \frac{k!}{(k-i)!}$ y $n^{(i)} = \frac{n!}{(n-i)!}$

A continuación se muestra una implementación en Python de dicha fórmula[24]:

```
self.s = 0.0
self.d = (int)(self.dist/m)
for l in range(1,self.d+1,1):
    self.num = (float)(self.factorial(k))/(self.factorial(k-l))
    self.den = (float)(l*((3.0/4.0)**(l-1.0)))*((self.factorial(m))/(self.factorial(m-l)))
    self.s += (self.num/self.den)
if self.d == 1.0:
   self.num = (float)(self.factorial(k))/(self.factorial(k-1))
   self.den = (float)(1*((3.0/4.0)**(0)))*((self.factorial(m))/(self.factorial(m-1)))
   self.s = (self.num/self.den)

self.dist = self.s
```

Pero con esta fórmula de Tajima y Neil, surge un nuevo inconveniente, y es que cuando se requiere analizar un gran número de secuencias genéticas la fórmula falla, ya que esta necesita del múltiple cálculo de un factorial, como se puede apreciar en la fórmula implementada en Pyhon, y cuando el número que llega a este factorial es muy grande la aplicación arroja un error. Para superar esta limitante fue necesario replantear esta fórmula y diseñar una equivalente, que fuese capaz de particionar los datos que se van generando a medida que se hace la comparación nucleótido a nucleótido de las dos secuencias, para que así los números que lleguen al factorial no sean muy grandes y permitan la ejecución normal del programa.

[24] SANDRA ACERO. ALGORITMO PARA ANÁLISIS FILOGENÈTICO: UPGMA (Unweighted Pair Group Method using Arithmetic averages). Tesis de grado. 2007.

A continuación se plantea la nueva implementación en Pyhton de la fórmula de Tajima y Neil:

```
self.matDistancia = matrices.create((n),(n))

    for i in range(0,n-1,1):
        for j in range(i+1,n,1):
            self.dist = 0.0
            self.otu1 = OTUS[i]
            self.otu2 = OTUS[j]

            for k1 in range(0,m,1):
                if self.otu1[k1] != self.otu2[k1]:
                    self.dist = self.dist + 1

            # EQUIVALENCIA DE LA FÓRMULA DE  TAJIMA

            self.s = 0.0
            self.k = (int)(self.dist)

            for self.i1 in range(1,self.k+1,1):

                self.num = self.factorial(self.k)/self.factorial(self.k-self.i1)
                self.den = self.i1*((3.0/4.0)**(self.i1-1.0))*(self.factorial(m))/(self.factorial(m-self.i1))
                self.s += (self.num/self.den)
            self.dist = self.s

            self.d_ab = 0
            self.d_ab2 = 0
            self.i1 = 1.0
            self.sw = 0
            while self.i1 <= self.k and self.sw == 0:
                self.t1 = 1.0
                self.t2 = (fact(self.k)/fact(self.k-self.i1))/(fact(m)/fact(m-self.i1)*pow(0.75,self.i1-1)*self.i1)
                self.j1 = 1

                while ((self.j1 <= m) and (self.t1 > 1e-20)):
                    self.m1 = m - self.j1 + 1.0
                    self.k1 = self.k - self.j1 + 1.0
                    if self.k1 < 1:
                        self.k1 = 1.0
                    self.ki1 = self.k1 - self.i1
```

```
                if self.ki1 < 1:
                    self.ki1 = 1.0
                self.mi1 = self.m1 - self.i1
                if self.mi1 < 1:
                    self.mi1 = 1.0
                self.t1 = self.t1 * self.k1*self.mi1/self.ki1/self.m1
                if self.j1 > 1 and self.j1 < self.i1+1:
                    self.t1 = self.t1 * 4/3
                if self.t1 < 1e-20:
                    self.sw = 1
                self.j1 = self.j1+1
            self.t1 = self.t1 / self.i1
            self.d_ab = self.d_ab + self.t1
            self.d_ab2 = self.d_ab2 + self.t2
            self.i1 = self.i1 + 1.0

        self.matDistancia[j][i] = self.d_ab
        self.matDistancia[i][j] = self.d_ab

    return self.matDistancia
```

4.1.4 Creación del módulo para la formación del GraphDatabases

Los GraphDatabases son representaciones gráficas que se pueden hacer con cualquier tipo de dato, en este caso, con las secuencias genéticas, para obtener información de ellas. Esta definición y otros pasos posteriores serán desarrollados en el capítulo 6.

Para el desarrollo de esta etapa fue necesario el diseño de un módulo para la formación de GraphDatabases que permitiera poder almacenarlos en una lista para posteriormente trabajar con ellos. Este módulo lo que hace es que a partir del árbol filogenético que se generó cuando se ejecutó el método UPGMA, crea una lista que contiene las cadenas comunes del grupo de OTUs que se analizó para

posteriormente ser comparados con una nueva OTU para establecer probabilidades de asociación.

4.1.5 Creación de un área de dibujo óptimo para el despliegue del árbol filogenético generado

Como se mostrará más adelante, el software cuenta con tres áreas de despliegue de la información. La primera donde muestra las OTUs cargadas de la base de datos, la segunda donde muestra la matriz de distancia calculada y la tercera donde despliega el árbol filogenético que fue generado. Un inconveniente que surgió en el despliegue del árbol por pantalla, fue que cuando se analiza un gran número de secuencias genéticas el árbol que se dibuja es demasiado grande, ya que el tamaño promedio del área que se necesita para que este sea dibujado es de 31257*31257, lo que equivaldría a 31 monitores de 900*900. El cual es un área demasiado grande que no solo era imposible de mostrar por pantalla, sino que además provocaba una demora significativa en el tiempo de ejecución de la aplicación y en muchos casos el desplome del programa por la cantidad de memoria que consumía en el proceso. Para solucionar este impase fue necesario hacer un gran número de prueba y error, hasta encontrar un área con el tamaño indicado, para el despliegue del árbol sin importar su tamaño y hacer uso de dos funciones que son *zoomin* y *zoomout*, que son explicadas en el manual del usuario, y que permiten ampliar o disminuir la vista del árbol. El tamaño ideal del área es de 700*300, lo que despliega un árbol de tamaño pequeño, pero que haciendo uso de estas dos funciones permite verlo en sus dimensiones reales.

4.1.6 Diseño de la nueva interfaz gráfica con todos sus elementos

La interfaz gráfica que fue diseñada para la aplicación es una interfaz amigable con el usuario con distintos elementos que son fácilmente reconocidos y manejables. Una versión preliminar de esta interfaz se puede encontrar en los trabajos de Sandra Acero[25], Alejandro Pedrozo[26] y Mat Montalvo[27]. Para el diseño de esta fue necesario el uso de un conjunto de módulos incorporados en Python para el diseño de este tipo de interfaces gráficas, conocido como Tkinter, que son de fácil utilización. Además también fue necesaria la utilización de un módulo llamado Pmw para el diseño de elementos gráficos un poco más complejos y especializados[28].

La interfaz cuenta con diversas áreas que son explicadas a continuación:

1. Área de la barra de menú: Esta se encuentra en la parte superior de la ventana. Está conformada por: Bases de datos, Analizar, GraphDatabases, Ver y Ayuda. Donde cada una está despliega un submenú con funciones específicas.

2. Área de la barra de herramientas: Esta se encuentra debajo de la barra de menú, son una serie de iconos que permiten al usuario acceder de manera directa a las distintas aplicaciones del programa, sin necesidad de recurrir a la barra de menú. Los iconos empleados se encuentran disponibles de manera libre para su utilización en el sitio:

[25] SANDRA ACERO. ALGORITMO PARA ANÁLISIS FILOGENÈTICO: UPGMA (Unweighted Pair Group Method using Arithmetic averages). Tesis de grado. 2007.

[26] ALEJANDRO PEDROZO. ANÁLISIS FILOGENÉTICO DEL VIH BASADO EN EL MÉTODO NEIGHBOR – JOINING. Tesis de grado. 2007.

[27] MAT MONTALVO. ANÁLISIS FILOGENÉTICO UTILIZANDO EL MÉTODO "PARSIMONY". Tesis de grado. 2007.

[28] Paquete: Pmw . Release: 1.2 . Date: agosto 4 2003. URL: http://sourceforge.net/projects/pmw/. Fecha de descarga: Marzo 7 de 2007 9:46 a.m. Administrador del proyecto: Greg McFarlane. Sistema Operativo: OS Independen. Licencia: MIT License.

http://www.neatui.com/. Esta página por razones de mantenimiento se encuentra deshabilitada desde el 29 de febrero de 2008 hasta la fecha.

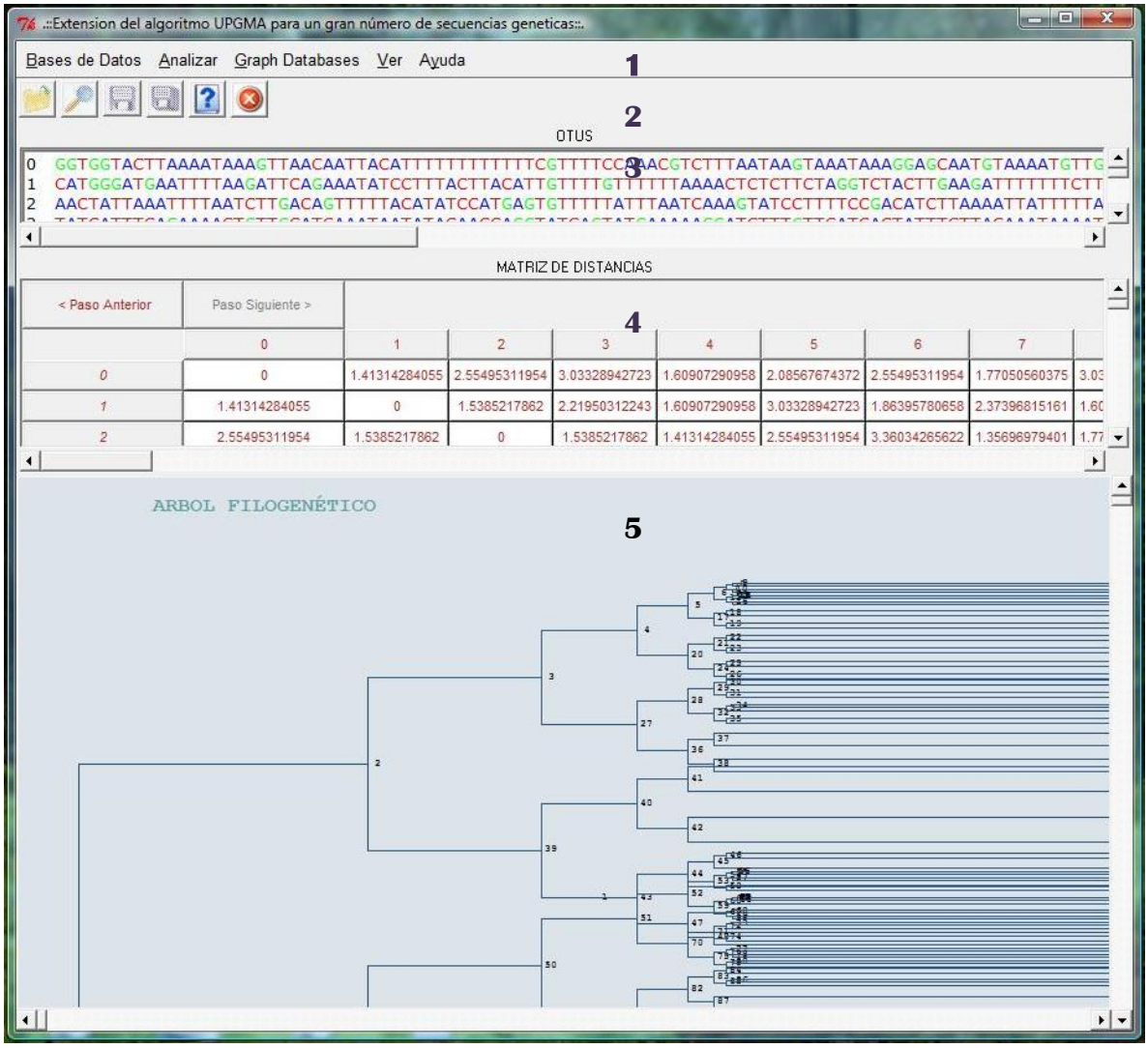
3. Área de visualización de las secuencias: Es el área donde son mostradas las secuencias genéticas que fueron obtenidas de la bases de datos de MySQL.

4. Área de visualización de la matriz de distancia: Es el área donde se despliega la matriz de distancia que fue obtenida luego de realizar los respectivos cálculos en las OTUs.

5. Área de visualización del árbol: Es el área donde es desplegado el árbol filogenético obtenido a partir de la matriz de distancia. Como se explico anteriormente este es mostrado en un tamaño reducido pero tiene la opción de ser mostrado en tamaño real, y puede ser guardado en formato jpg gracias a la utilización de una librería llamada Python Imaging Library (PIL), la cual le brinda la posibilidad a Python de procesar imágenes gif y las ya mencionadas jpg[29].

[29] Versión: PIL 1.1.6. Disponible en: http://www.pythonware.com/products/pil/index.htm. Author: Secret Labs AB (PythonWare). Description: Python Imaging Library. Name: PIL. Url: http://www.pythonware.com/products/pil.

Figura 6. Interfaz gráfica de la aplicación.



CAPÍTULO 5: GRAPHDATABASES

En capítulos anteriores se menciona el término GraphDatabases, a continuación se dará la definición formal del término y se explicará en detalle la forma cómo estas Bases de Datos fueron integradas con el método UPGMA.

5.1 DEFINICIÓN

Los GraphDatabases son representaciones gráficas que se pueden llevar a cabo con cualquier tipo de dato para facilitar el acceso a la información que estos contienen y su posterior extracción. Con su implementación lo que se pretende es mostrar la relación existente entre los datos, para facilitar la reestructuración de los mismo y que se puedan llevar a cabo actualizaciones en la información contenida en ellos. Debido a esto son muy utilizados para el almacenamiento de información genética[30].

Cuando se trabaja con este tipo de estructura de datos se cuenta con plena libertad para decidir qué datos se encuentran conectados entres sí para posteriormente construir un árbol que contenga las distintas relaciones. Una vez completado este paso para poder llevar a cabo la extracción de la información se puede proceder de diversas maneras, una de las más utilizadas es a través de sentencias SQL que son recibidas por medio de un intérprete diseñado para ese propósito. La principal desventaja que existe al utilizar sentencias SQL, es que en muchos casos estas pueden llegar a ser muy complejas y esto conlleva a que a usuarios inexpertos se les dificulte obtener la información que necesitan.

[30] ADRIAN SILVESCU, DOINA CARAGEA, ANNA ATRAMENTOV. 2002. GraphDatabases. Iowa State University. Pp, 1-14. Fecha de consulta: 5 de Junio de 2007.

5.2 IMPLEMENTACION DE LOS GRAPHDATABASES CON EL MÉTODO UPGMA

Como se mencionó anteriormente cuando se diseña un GraphDatabases por lo general se utilizan sentencias SQL para obtener la información, pero esto en algunos casos resulta inconveniente ya que se requiere que la persona que esté haciendo uso de ellos tenga conocimientos en dicho lenguaje. Por tal motivo, para su aplicación en el método UPGMA fue necesario darles otra perspectiva que facilitara la extracción de la información.

Primeramente se creó un módulo en Python llamado graDibujarArbolNew, este módulo fue diseñado para que a partir del árbol filogenético que fue generado, se cree una lista que contenga las cadenas comunes del grupo de OTUs que se analizó. Dichas cadenas, son subgrupos de nucleótidos que las distintas OTUs tienen en común y que fueron formados durante el proceso de ejecución del método UPGMA. Una vez hecho esto se procede a cargar un archivo en formato FASTA que contiene una nueva secuencia genética, dicha secuencia es comparada nucleótido a nucleótido con cada uno de los elementos del GraphDatabases para establecer probabilidades de asociación. Estas probabilidades van a permitirle a los investigadores determinar que tan relacionada está una secuencia genética recibida con los distintos grupos genéticos que conforman un árbol filogenético dado

Una vez hecho todo el proceso se despliega por pantalla la información y automáticamente la aplicación genera un archivo en el disco C:/ donde se encuentra todo el proceso.

5.2.1 Captura de la secuencia a comparar

La secuencia genética que va a ser recibida en un archivo se encuentra en formato FASTA, el cual es muy usado para representar secuencias de ADN. Dicho formato está conformado por un área para el encabezado y un área para las secuencias. El encabezado comienza con el símbolo > seguido de un código que identifique a la secuencia, de una breve descripción, y por último la longitud de la misma, la cual está precedida del símbolo | y la palabra len seguida de =. Y el área de la secuencia que contiene a la secuencia en sí, en la cual luego de escribir la secuencia se debe dar <Enter>, para que el formato quede completo.

El formato es el siguiente:

>CódigoIdentificador Descripción |len=(entero)longitud[Enter]
Secuencia[Enter]

A continuación se muestra un ejemplo con el procedimiento completo para formar el GraphDatabases y el cálculo de las probabilidades para dar mayor claridad.

EJEMPLO:

Sea el GraphDatabases formado al ejecutar el método UPGMA, que está conformado por cada uno de los grupos de OTUs que se encuentran en el árbol filogenético.

['AGTAG---', 'AGTAGT--', 'AGTAGTT-', 'AGTAGGG-']

Y sea la secuencia AGTAGTTC, que es reciba en un archivo en formato FASTA.

Lo que se procede hacer en comparar nucleótido a nucleótido para ver qué tanta similitud guardan entre sí, para finalmente establecer probabilidades de asociación.

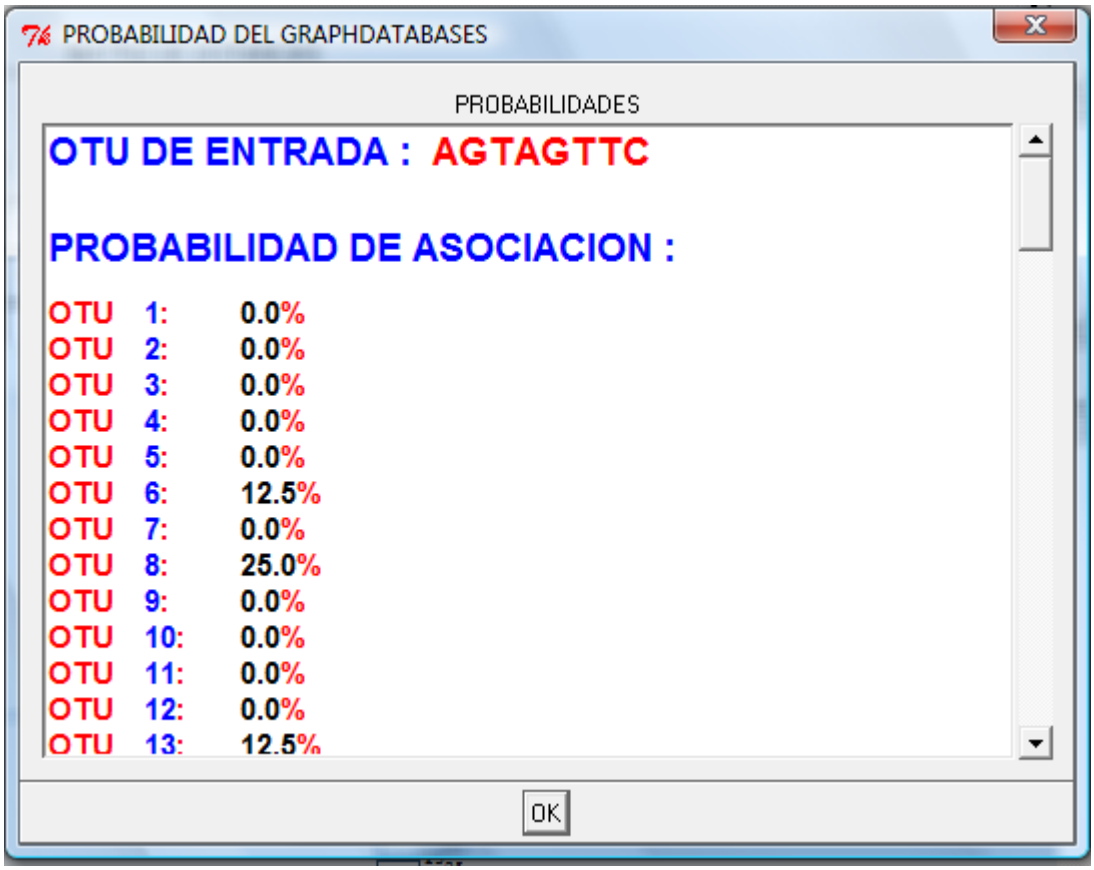
1. **AGTAG**TTC
2. **AGTAG**---

$$proAsoc(1,2) = \frac{5}{8} = 0.625$$
$$proAsoc(1,2) = 0.625*100 = 62.5\%$$

5.2.2 Interfaz gráfica del GraphDatabases

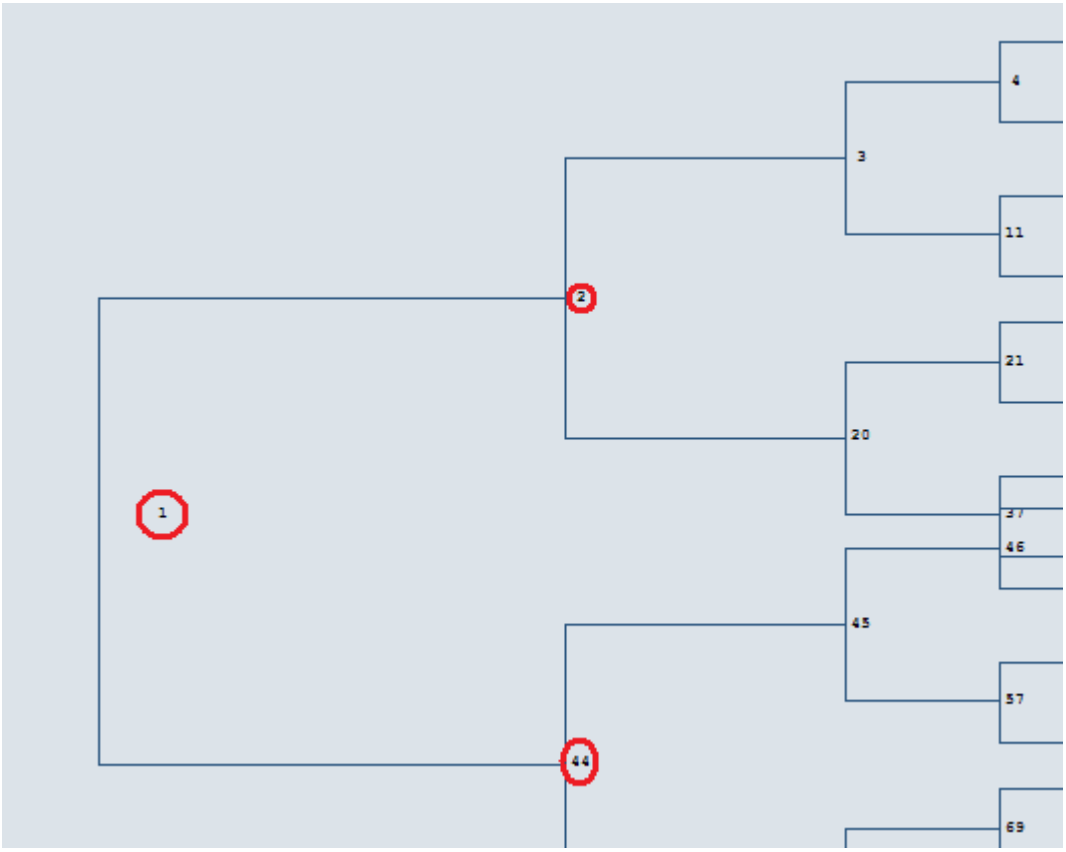
Una vez el usuario ejecuta el GraphDatabases y carga el respectivo archivo con la secuencia a comparar se desplegará la ventana de la figura 7, en la cual se encuentra la OTU de entrada, seguida de las distintas OTUs con las que se armó el GraphDatabases, las cuales están acompañadas con sus respectivas probabilidades.

Figura 7. Interfaz gráfica del GraphDatabases.



Las OTUs que conforman el GraphDatabases van enumeradas, esto para que el usuario pueda identificar más fácilmente a cual parte del árbol filogenético quedaría asociada la OTU de entrada, si esta fuese incluida en la bases de datos.

Figura 8. Numeración para ser usada en el GraphDatabases.



CAPÍTULO 6: PRUEBAS Y RESULTADOS OBTENIDOS

En capítulos anteriores se hizo una descripción completa sobre los distintos procedimientos que se llevaron a cabo para que la aplicación no solo fuese capaz de interactuar con una base de datos previamente diseñada, sino que además fuese capaz de analizar un gran número de secuencias genéticas. En este capítulo se mostrarán los resultados obtenidos al analizar un gran número de secuencias de longitud variable.

6.1 PRUEBAS CON UN GRAN NÚMERO DE SECUENCIAS GENÉTICAS

En la aplicación fue necesario llevar a cabo pruebas con un gran número de secuencias genéticas para comprobar que la aplicación era capaz de soportar los distintos cálculos que son llevados a cabo. Aunque no es uno de los objetivos explícitos del programa, también se llevaron a cabo pruebas con secuencias de gran longitud para así verificar la eficacia de la implementación con métodos numéricos de la fórmula Tajima, la cual es el pilar de la aplicación, ya que es la que permite llevar a cabo el análisis filogenético de un gran número de secuencias genéticas de cualquier longitud. En los resultados que se muestran a continuación se incluye también el tiempo de ejecución de la aplicación.

Las pruebas que se llevaron a cabo se hicieron en un equipo que presenta las siguientes características:

- Equipo: HP Pavilion Slimline Pc s3041la
  - Sistema operativo: Microsoft Windows Vista Home Premium

- Características: Procesador AMD Athlon 64 X2 Dual Core. 2.0 GB de RAM. 400 GB de disco duro.

Las secuencias que fueron cargadas en la bases de datos de MySQL, son secuencias reales obtenidas de la siguiente página: **ftp://ftp.ncbi.nih.gov/blast/db**, las cuales son de uso libre.

- Pruebas con secuencias de longitud constante

Tabla 5. Tabla de tiempo de ejecución con cantidad de secuencias constante.

Número de secuencias	Longitud de las secuencias	Tiempo de ejecución (Milisegundos)
50	**8**	**8481**
60	**8**	**8373**
70	**8**	**11330**
80	**8**	**20722**
90	**8**	**38283**
100	**8**	**41893**
105	**8**	**43522**

- Pruebas con secuencias de longitud variable

Tabla 6. Tabla de tiempo de ejecución con cantidad de secuencias constante.

Número de secuencias	Longitud de las secuencias	Tiempo de ejecución (Milisegundos)
95	107	303747
95	212	1068023

Figura 9. Gráfico longitud de las secuencias VS tiempo de ejecución

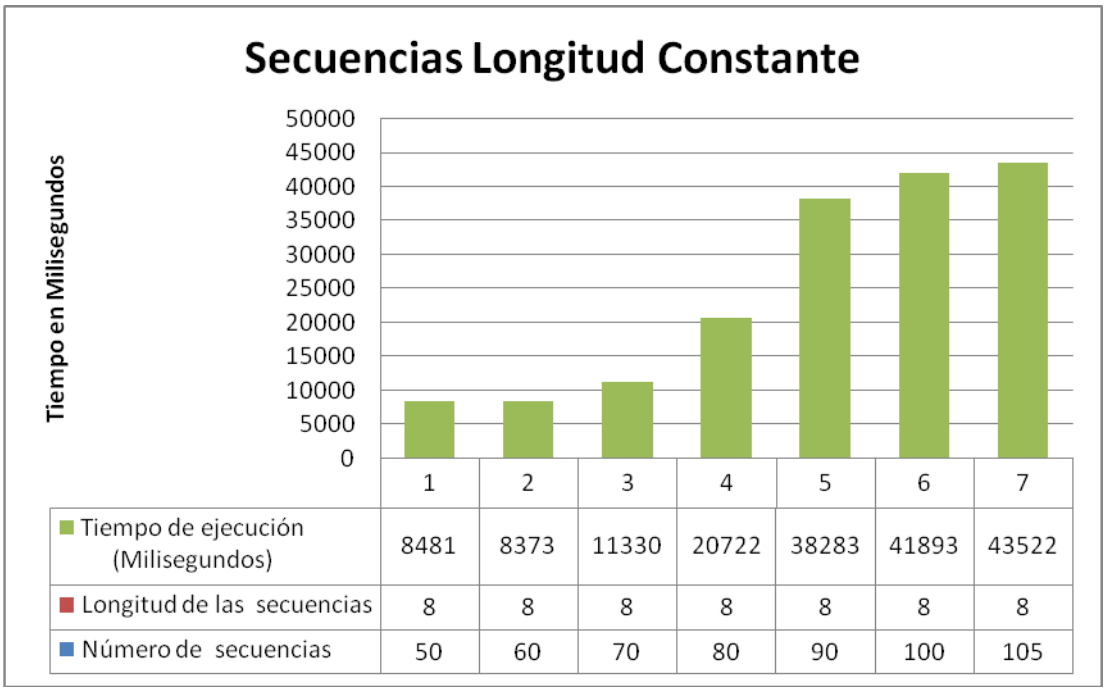
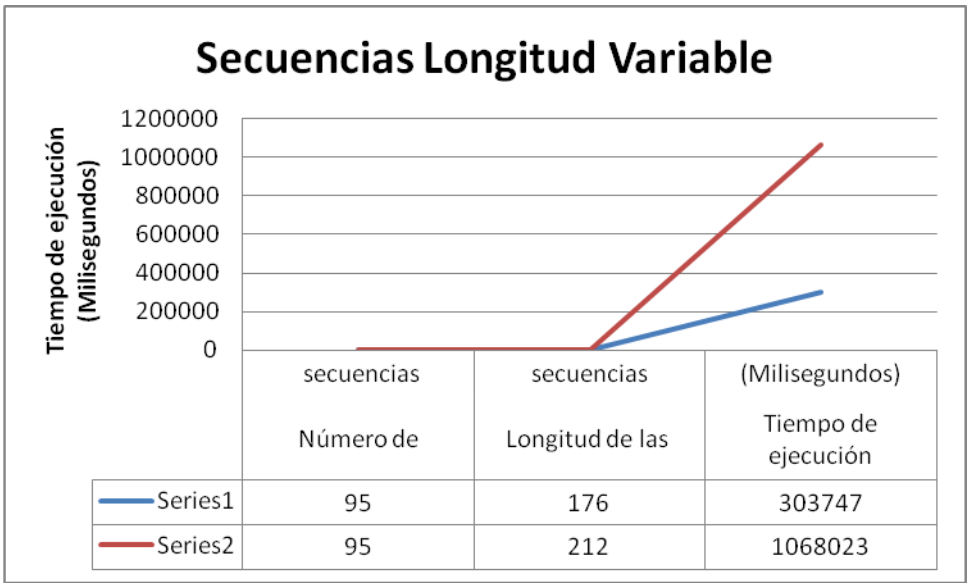


Figura 10. Gráfico longitud de las secuencias VS tiempo de ejecución



6.2 ANÁLISIS DE LOS RESULTADOS OBTENIDOS

La figura 9 se obtuvo a partir de los datos de la tabla 5. Estos datos fueron incorporados de manera independiente en la bases de datos, la primera tanda de datos consistió de 43 secuencias de ADN, la segunda tanda fue 90, la tercera de 100 y la quinta de 105, cada una de una longitud de 8 nucleótidos. Anteriormente el algoritmo estaba limitado a procesara un número de 10 secuencias

Las pruebas realizadas demuestran que el software es capaz de procesar un gran número de secuencias genéticas y además también muestran que el tiempo de ejecución depende de la cantidad de secuencias que se están analizando y de la

manera como Python/Windows (Lenguaje/Sistema Operativo) manejan la memoria y hacen los procesos de paginación, ya que se puede ver claramente que el tiempo crece sigmoidalmente dependiendo del número de secuencias.

La figura 10 se obtuvo a partir de los datos de la tabla 6. Estos datos revelan que el software es capaz no solo de procesar un gran número de secuencias genéticas, sino que además estas pueden ser de gran longitud, ya que anteriormente el algoritmo solo podía procesar secuencias cuya longitud no fuese mayor a 170 nucleótidos. Para este caso las secuencias usadas eran 95 y sus longitudes eran 176 y 212 respectivamente. Para observar el desempeño del sistema se realizaron pruebas con 4 secuencias y el número máximo de nucleótidos que se pudo procesar fue 497. Los resultados obtenidos también demuestran que el tiempo de ejecución también depende de la longitud de las secuencias, debido a que este creció exponencialmente a medida que estas aumentaron su longitud.

6.3 LIMITANTES DE LA APLICACIÓN

Gracias a los resultados obtenidos se pudo ver claramente que la aplicación cumple con el objetivo de que esta fuese capaz de procesar un gran número de secuencias genéticas y añadió el ingrediente de que estas pudiesen ser de gran longitud. Pero el software tiene una limitante que ya no corresponde al área de programación, y es que a pesar de que esta procesa grandes números de secuencias genéticas, cuando estas son mayores a 105, este lleva a cabo todo el proceso pero no es capaz de desplegar la información por pantalla, ya que por la cantidad de cálculos que debe desarrollar durante la ejecución del UPGMA y a que debe mostrarlos gráficamente, muchas veces la memoria del computador es insuficiente para llevar todo esto a cabo.

Otra solución que se implementó para superar este impase, es que el software cuenta con la opción de generar un archivo en formato .txt, que contiene todos los

cálculos que fueron generados y además es capaz de guardar una imagen que contiene el árbol filogenético.

CAPÍTULO 7: CONCLUSIONES

En el capítulo anterior se llevo a cabo la descripción de las distintas pruebas a las que fue sometida la aplicación para corroborar su eficacia. Se llevaron a cabo pruebas con un gran número de secuencias genéticas de gran y corta longitud para analizar el comportamiento del software y de cómo se ve afectado tanto el tiempo de ejecución como la aplicación en sí, cuando se aumentan el numero de secuencias en la bases de datos. En este capítulo se mostrarán las distintas conclusiones a las que se llegaron una vez se concluyó el proyecto.

Los objetivos establecidos para el desarrollo y ejecución del proyecto fueron cumplidos a cabalidad. El principal de los objetivos que era diseñar una aplicación que fuese capaz de procesar un gran número de secuencias genéticas almacenadas en una bases de datos aplicando para ello el método para análisis filogenético UPGMA, fue cumplido con éxito, ya que al concluir este proyecto, se cuenta con una poderosa herramienta que es capaz de procesar grandes números de secuencias genéticas de gran longitud, superando así la limitante con la que se contaba al principio que solo permitía el análisis de máximo 10 secuencias (OTUs) de una longitud no mayor a 170 nucleótidos. Además se incluyó en la aplicación una nueva funcionalidad conocida como GraphDatabases, que permite la predicción de asociación con una nueva cadena de ADN entrante, la cual podría servir como base para un futuro proyecto, a través del cual se implemente algún método para análisis filogenético sin tener que re-calcular la matriz de distancia para incluir una nueva secuencia dentro del árbol filogenético que ya ha sido generado.

Para poder llevar a cabo la ejecución del proyecto, fue necesario estudiar y analizar aspectos correspondientes tanto al área de la genética como al área de la biología para así contar con los conocimientos suficientes y necesarios que permitieran

avances significativos en la ejecución del proyecto. Además fue necesario llevar a cabo un estudio profundo del método UPGMA que ya había sido implementado, para así determinar que partes del algoritmo eran necesarios analizar y mejorar para que se cumplieran los objetivos propuestos y superar las limitantes con las que este contaba.

BIBLIOGRAFÍA

T Leitner , D Escanilla , C Franzen , M Uhlen , J Albert. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. 1996 Oct 1 93:10864-9. Disponible en Internet: www.pubmed.com

JUKES, T.H., y C.R. CANTOR. 1969. Evolution of Proteins Molecules. Pp 21-132 en H:N: MUNRO, ed. Mammalian protein metabolism. Academis Press, New York.

GRIBSKOV, M. Sequence Analysis Primer. New York: Oxford University Press, Incorporated. 1991. 296 p. ISBN 0-19-509874-9.

DAWSON, Michael. Python Programming for the Absolute Beginner. Boston: Course PTR. 2003. 480 p. ISBN 1-59200-073-8.

ALLEN G, Rodrigo; GERALD H, Learn. Computational and Evolutionary Analysis of HIV Molecular Sequences. New York : Kluwer Academic Publishers, 2000. 309 p. ISBN 0-7923-7994-2.

COFFIN, John M.; HUGBERS, Stephen H.; VARMUS, Harold E. Retroviruses. New York: Cold Spring Harbor Laboratory Press. 1999. 843 p. ISBN: 0-87-969571-4.

WANG, Jason T. L.; WU, Cathy H.; WANG, Paul P. Computational Biology and Genome Informatics. Londres: Scientific Publishing Company, Incorporated. 2003. 266 p. ISBN 9-81-238257-7.

CECCHI, MARÍA CLAUDIA, GUERRERO-BOSAGNA, CARLOS y MPODOZIS, JORGE. El ¿delito? de Aristóteles. Rev. chil. hist. nat. [Online]. set. 2001, Vol.74, no.3, Fecha de consulta: 22 de Marzo de 2007, p.507-514. Disponible en Internet: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0716-078X2001000300001&lng=es&nrm=iso. ISSN 0716-078X.

BYSTROFF Chris. Bioinformatics. Fecha de consulta: Marzo 22 de 2007. Depts of Biology & Computer Science Rensselaer Polytechnic Institute. Disponible en: www.bioinfo.rpi.edu/~bystrc/courses/biol4540.html

SORENSEN, Daniel; GIANOLA, Daniel. Likelihood, Bayesian and MCMC Methods in Genetics. New York: Springer-Verlag New York, Incorporated. 2002. 757 p. ISBN 0-387-22764-4.

Cornell University Library. The Phylogeny Inference Package. [Online]. Fecha de consulta: 2 abril 2007. Disponible en Internet: http://vivo.library.cornell.edu/entity?home=1&id=5111

CRISP, Michael. Introductory glossary of cladistic terms. [Online]. Fuente: Invited Contributions of the Society of Australian Systematic Biologists. Fecha de consulta: 2 abril 2007. Disponible en Internet: http://www.science.uts.edu.au/sasb/glossary.html

TAJIMA, Fumio. MSATOCHI, Nei. Unbiased Estimation of Evolutionary Distance between Nucleotide Sequences. Department of population Genetics, National Institute of genetics. University of Chicago. Mol. Biol. Evol. 10(3). Pp. 677-688. Octubre 6 de 1992.

THOMAS Dave, NMSR. Example calculation of phylogenies: The UPGMA method. [Online]. Fecha de consulta: 11 de abril de 2007. Rosswell, New Mexico. Última modificación: 31 de octubre de 2002. Disponible en: http://www.nmsr.org/upgma.htm

MacLEOD, Norman; FOREY, Peter L. Morphology, Shape & Phylogeny. London: Taylor & Francis, Incorporated. 2002. 318 p. ISBN 0-203-16517-9.

GONZALEZ, G. Los Coccinellidae de Chile. [Online]. 1996. Fecha de consulta: 24 abril 2007. Disponible en Internet: http://www.coccinellidae.cl

DEPARTMENT OF BIOLOGY. SOUTHERN OREGON UNIVERSITY. Compleat Cladist: A primer of phylogenetic procedures. [Online]. Fecha de consulta: 24 abril 2007. Disponible en Internet: http://www.sou.edu/biology/Courses/Bi332/CompleatCladist.pdf

J. Emilio, CABALLERO López, PÉREZ SUÁREZ Gonzalo. Métodos de análisis en la reconstrucción filogenética. Departamento de Biología Animal. Universidad de Alcalá. Alcalá de Henares. Madrid. España. nº 26, Fecha de consulta: 26 de Septiembre de 2007 1999: 45-56. Disponible en Internet: http://entomologia.rediris.es/sea/bol/vol26/s1/articulo/index.htm

LOS ALAMOS NATIONAL LABORATORY. HIV Sequence Database. [Online]. Fecha de consulta: 22 de marzo de 2008. Disponible en Internet: http://hiv-web.lanl.gov/

FELSENSTEIN, Joe. Phylip. [Online]. Fecha de consulta: 22 de marzo de 2008. Disponible en Internet: http://evolution.genetics.washington.edu/phylip.html

M. CARR Steven, Cluster Analysis: an example. [Online]. Fecha de consulta: 22 de marzo de 2008. Memorial University of Newfoundland. Genetics, Evolution, and

Molecular Systematics Laboratory. Department of Biology. St. John's NF A1B 3X9, Canada. Disponible en: http://www.mun.ca/biology/scarr/Bio4900_UPGMA.html

FACULTAD DE CIENCIAS UNIVERSIDAD AUTÒNOMA DE MEJICO. Index of /Bioinformática. [ONLINE]. Fecha de consulta: 22 de marzo de 2008. Disponible en internet: http://bacteria.fciencias.unam.mx/Bioinformatica

FELSENSTEIN, Joe. Phylogeny programs. [Online]. Fecha de consulta: 3 de mayo de 2007. Disponible en Internet: http://evolution.genetics.washington.edu/phylip/software.html

BECHLY Günter (Böblingen, Alemanha). Glossary of Phylogenetic Systematics. [Online]. Fecha de consulta: 2 octubre 2007. Disponible en Internet: http://www.bernstein.naturkundemuseum-bw.de/odonata/glossary.htm

KUHNER MK; FELSENSTEIN J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. [Online]. Volumen 12, Número 3. Mayo de 1995. Disponible en Internet: http://mbe.oxfordjournals.org/archive/

LOS ALAMOS NATIONAL LABORATORY. HIV AND SIV NOMENCLATURE. [Online]. Fecha de consulta: 5 de junio 2007. Última modificación: Mon Apr 24 19:23 2006. Disponible en: http://hiv-web.lanl.gov/content/hiv-db/HelpDocs/subtypes-more.html.

MIYAMOTO, Michael M. ; CRACRAFT, Joel. Phylogenetic Analysis of DNA Sequences. New York: Oxford University Press, Incorporated. 1991. 369 p. ISBN 0-19-506698-7.

MOTOO Kimura, TOMOKO Ohta. On the stochastic model for estimation of mutational distance between homologous proteins. Journal of Molecular Evolution. [Online]. Volume 2. Issue 1. Monday, May 16, 2005. 25 de Julio de 2007. Disponible en http://www.springerlink.com/content/u244547185w7306w. ISSN 1432-1432 , Pages 87-90.

MOUNT, David W. Bioinformatics: Sequence and Genome Analysis. New York: Cold Spring Harbor Laboratory Press. 2001. 577 p. ISBN 0-87969-597-8.

MURPHY, Robert F. Introduction to Computacional Molecular Biology. [Online]. 2006. Departments of Biological Sciences and Biomedical Engineering Carnegie Mellon University. Última modificación enero 6 2006. Disponible en Internet: http://www.cmu.edu/bio/education/courses/03311/

NING K., SHAN T., XIANG S. L., SHEN W. Phylogenetic Tree Reconstruction: Distance Based. [Online]. Octubre 10 de 2003. Disponible en: http://www.comp.nus.edu.sg/~bioinfo/phylogenetic%20tree%20reconstruction_2_8.pdf . Fecha de consulta: 1 de julio de 2007. 20 páginas.

OMS/ONUSIDA. Los asociados mundiales en pro de una vacuna contra el VIH fortalecen su colaboración para acelerar los avances. [Online]. 7 de febrero de 2005. Fecha de consulta: 2 de julio 2007. Disponible en Internet: http://www.who.int/entity/mediacentre/news/notes/2005/np04/es/index.html

OOH SING Hua, OOI HONG Sain, WONG CHEE Hong, WONG SUM Thai. Phylogenetics Trees Reconstruction. [Online]. Septiembre 26 de 2003. Disponible en:
http://www.comp.nus.edu.sg/~bioinfo/phylogenetic%20tree%20reconstruction_1_7.pdf . Fecha de consulta: 24 de agosto de 2007. 26 páginas.

PYTHON SOFTWARE FOUNDATION. About Python. [Online]. Fecha de consulta: 27 agosto 2007. Disponible en internet: http://www.python.org/about/

PYTHON SOFTWARE FOUNDATION. FAQ General de Python. [Online]. 16 de diciembre del 2005. Fecha de consulta: 6 de marzo de 2007. Disponible en Internet: http://www.python.org/doc/faq/es/general/#faq-general-de-python
RILEY, Sean. Game Programming with Python . Hingham, Massachusetts: Charles River Media. 2003. 484 p. ISBN 1-58450-258-4.

SAKSENA NK, WANG B, Ge YC, XIANG SH, DWYER, DE , CUNNINGHAM AL. Coinfection and genetic Recombination between HIV-1 strains: possible biological implications in Australia and South East Asia. 26:121-7. Disponible en Internet: www.pubmed.com

SIDDALL, Mark E. Phylogentics: Just Methods. [Online]. Fecha de consulta: 30 de agosto 2007. American Museum of Natural History. Disponible en internet: http://research.amnh.org/~siddall/methods/

UNIVERSIDAD DE CALIFORNIA, MUSEUM OF PALEONTOLOGY. Journey into the phylogenetics Systematics. [Online]. Berkeley, EUA. Fecha de consulta: 9 de diciembre de 2007. Disponible en Internet: http://www.ucmp.berkeley.edu/clad/clad4.html

UNIVERSIDAD DE CALIFORNIA, MUSEUM OF PALEONTOLOGY. The Phylogeny of life. [Online]. Berkeley, EUA. Fecha de consulta: 2 de febrero de 2008. Disponible en Internet: http://www.ucmp.berkeley.edu/alllife/threedomains.html

UNIVERSIDAD DE LA REPUBLICA. Sistemática Biológica: Métodos Y Principios. Fecha de Consulta: 2 de febrero de 2008. Facultad de ciencias. Laboratorio de evolución.Montevideo, Uruguay. Disponible en: http://evolucion.fcien.edu.uy/sistematica/intro-cladistica.pdf

ANEXOS

ANEXO 1

Ejemplo del archivo que es cargado en la bases de datos de MySQL

humanepididymidis01	GGCTCACT
humanepididymidis02	GCAACCAC
humanepididymidis03	CGCCTCTC
humanepididymidis04	GGTTTAGT
humanepididymidis05	GGGCCTCC
humanepididymidis06	CTGGGTCG
humanepididymidis07	CCAGCTTC
humanepididymidis08	CAGACCGT
humanepididymidis09	GGGGGCTG
humanepididymidis010	TTTTATGC
humanepididymidis011	GTCATCAT
humanepididymidis012	CCCGCGTA
humanepididymidis013	CGCTCTTC
humanepididymidis014	GGCACCTT
humanepididymidis015	TTCCTGTC
humanepididymidis016	TCAGGAAA
humanepididymidis017	AAAAGGAA
humanepididymidis018	GAAGAGTC
humanepididymidis019	CACCTTGC
humanepididymidis020	GACCGCAG
humanepididymidis021	GGGGGCGG
humanepididymidis022	AGCCCTGC
humanepididymidis023	TCTCGCAA
humanepididymidis024	CTCAATCC

ANEXO 2

Ejemplo del archivo que es generado por la aplicación

Pasos Intermedios
Ejecutando UPGMA
FECHA: 2008-3-22
HORA: 22:1:59
Archivo creado con UPGMA v1.0

SECUENCIAS
AGTAGTTC
AGTAGTTA
AGTAGTAA
AGTAGGGG
AGTAGGGC

MATRIZ DE DISTANCIAS
0 0.125 0.27380952381 0.457010582011 0.27380952381
0.125 0 0.125 0.457010582011 0.457010582011

0.27380952381 0.125 0 0.457010582011 0.457010582011
0.457010582011 0.457010582011 0.457010582011 0 0.125
0.27380952381 0.457010582011 0.457010582011 0.125 0

Inicio

PASO1
Menor distancia 0.125

NUEVA MATRIZ

0 0.199404761905 0.457010582011 0.36541005291

0.199404761905 0 0.457010582011 0.457010582011

0.457010582011 0.457010582011 0 0.125

0.36541005291 0.457010582011 0.125 0

ARBOL[[0, 1, 0.0625, 'AGTAGTT-', 0.0625, 0.0625], 2, 3, 4]

PASO2

Menor distancia 0.125

NUEVA MATRIZ

0 0.199404761905 0.41121031746

0.199404761905 0 0.457010582011

0.41121031746 0.457010582011 0

ARBOL[[0, 1, 0.0625, 'AGTAGTT-', 0.0625, 0.0625], 2, [3, 4, 0.0625, 'AGTAGGG-', 0.0625, 0.0625]]

PASO3

Menor distancia 0.199404761905

NUEVA MATRIZ

0 0.434110449735

0.434110449735 0

ARBOL[[[0, 1, 0.0625, 'AGTAGTT-', 0.0625, 0.0625], 2, 0.099702380952380959, 'AGTAGT--',
0.037202380952380959, 0.099702380952380959], [3, 4, 0.0625, 'AGTAGGG-', 0.0625, 0.0625]]

PASO4

Menor distancia 0.434110449735

NUEVA MATRIZ

0

ARBOL[[[0, 1, 0.0625, 'AGTAGTT-', 0.0625, 0.0625], 2, 0.099702380952380959, 'AGTAGT--', 0.037202380952380959, 0.099702380952380959], [3, 4, 0.0625, 'AGTAGGG-', 0.0625, 0.0625], 0.21705522486772486, 'AGTAG---', 0.1173528439153439, 0.15455522486772486]

ANEXO 3

Ejemplo del archivo .jpg que es generado por la aplicación con el árbol filogenético

ANEXO 4

Ejemplo del archivo de entrada para el GraphDatabases

>001 secuencia de prueba no real|len=8
AGTAGTTC

ANEXO 5

Ejemplo del archivo de salida del GraphDatabases

GRAPH DATABASES

FECHA: 2008-3-22

HORA: 22:4:22

Archivo creado con UPGMA v2.0

GRAPH DATABASE

['--------', '--------', '--------', 'G--T----', 'GG-T-A-T', 'GCTT--C-', '--------', '-------C', '-C-ACC-C', 'GGGC---C', 'GGGC--CC', 'G-------', 'G-------', 'GG--GC--', 'GGGGGC-G', 'GGATGCTA', 'G-A--G--', 'GAA--GT-', 'G-CCG--G', '--------', '--------', 'C-G-G-C-', '-----A--', 'TT-G-A--', 'TT-GGAG-', '--GT-ACA', 'C--G----', 'C--GC-T-', 'CC-GC-T-', 'CCAGT-C-', '--------', '--------', '--C-----', '--C-----', '--C----C', 'CGC---TC', '--CC-TGC', '--CC----', '--CC-A--', '--CCCA-G', 'AGCCCATG', '-GC-----', '-GCA----', '-GCAC-TT', '-T------', '-T---T--', '-TTT-T--', '-TTT-TTT', 'TTTTTTTT', '-T-AAT-C', '-TC-T---', 'GTCAT-A-', 'TTCCT--C', '--------', '--------', '--------', 'CA------', 'CAGA--G-', 'CA---CCT', 'A-GA-C--', 'AGGA-CCA', '-------T', '---G-A-T', '-CTG-A-T', 'ATTA---T', '--------', '--------', '-------A', '-------A', '----G--A', '--A-G-AA', '-CA-GAAA', '-A--A--A', '-A-CA-AA', '-AACA-AA', 'T-T----A', 'T-TT-A-A', '--------', '------A-', '----ACAG', '---TACAG', 'T--TACAG', '-ACT-AA-', 'AAT-T-T-', '--A-----', '--A--T-T', '-AA--TGT', '-CA-T-G-', '-CA-TTGC']

OTU DE ENTRADA:

AGTAGTTC

PROBABILIDADES

OTU1	0.0%
OTU2	0.0%
OTU3	0.0%
OTU4	0.0%
OTU5	12.5%

OTU6	12.5%
OTU7	0.0%
OTU8	12.5%
OTU9	25.0%
OTU10	25.0%
OTU11	25.0%
OTU12	0.0%
OTU13	0.0%
OTU14	25.0%
OTU15	25.0%
OTU16	37.5%
OTU17	0.0%
OTU18	12.5%
OTU19	12.5%
OTU20	0.0%
OTU21	0.0%
OTU22	12.5%
OTU23	0.0%
OTU24	0.0%
OTU25	12.5%
OTU26	0.0%
OTU27	0.0%
OTU28	12.5%
OTU29	12.5%
OTU30	0.0%
OTU31	0.0%
OTU32	0.0%
OTU33	0.0%
OTU34	0.0%
OTU35	12.5%
OTU36	37.5%
OTU37	25.0%
OTU38	0.0%
OTU39	0.0%