Departamento de
Informática e Ingeniería
de Sistemas
Universidad Zaragoza

Escuela de
Ingeniería y Arquitectura
Universidad Zaragoza

Proyecto Fin de Carrera
Ingeniería Informática
Curso 2013/2014

Diseño y Estudio de herramientas para el Análisis del Índice de Conservación del ADN mitocondrial

Autor:

Francisco Merino Casallo

Bajo la dirección de:

Jorge Álvarez Jarreta
Elvira Mayordomo Cámara

Departamento de Informática e Ingeniería de Sistemas
Área de Lenguajes de Sistemas Informáticos
Escuela de Ingeniería y Arquitectura
Universidad de Zaragoza

Junio de 2014

Diseño y Estudio de herramientas para el Análisis del Índice de Conservación del ADN mitocondrial

RESUMEN

El objetivo principal de este Proyecto Fin de Carrera es el estudio y desarrollo de herramientas software para calcular de manera automatizada el índice de conservación de conjuntos de secuencias biológicas. Este conjunto de herramientas, pese a poder utilizarse de forma independiente, constituyen un potente sistema cuando se utilizan de forma combinada.

El índice de conservación es un estadístico muestral que tiene especial importancia en el estudio de la patogenicidad de mutaciones, el cual consiste en determinar si una mutación puede producir una enfermedad o por el contrario, es inocua al organismo en que se produce. Constituye una de las herramientas de las que disponen los biólogos para realizar los estudios evolutivos.

En la actualidad, los esfuerzos realizados por la comunidad científica para automatizar este proceso han producido contadas herramientas web con algunas limitaciones. Sorprende especialmente el no haber encontrado ninguna que permita utilizar este valor estadístico sobre grandes volúmenes de datos como un valioso instrumento adicional en este tipo de estudios.

Para resolver este problema, se han implementado una serie de algoritmos haciendo uso de técnicas de programación modular para desarrollar el núcleo del sistema; y pequeñas utilidades para automatizar tareas auxiliares como el tratamiento de los conjuntos de secuencias o los análisis estadísticos requeridos para validar los experimentos realizados.

Entre las dificultades encontradas durante el desarrollo de este proyecto cabe destacar la necesidad de formación previa en temas biológicos, ya que no se habían tratado estos temas a lo largo de la carrera. Por otro lado, la necesidad de realizar un tratamiento sobre los conjuntos de secuencias biológicas también implicó importantes esfuerzos en distintas fases del trabajo realizado. Por último, la obtención de información adicional relativa a la división de dichas secuencias para facilitar su análisis supuso un importante obstáculo a resolver en el último tramo del proyecto.

Los resultados del sistema diseñado para los conjuntos de secuencias biológicas han sido muy positivos, especialmente teniendo en cuenta el volumen de datos con el que se ha estado trabajando, muy superior al empleado en estudios anteriores. Gracias a la información extraída de los informes generados por el sistema desarrollado se ha podido constatar la presencia de mutaciones ya conocidas por la comunidad bióloga, lo que demuestra la validez de los resultados obtenidos.

Agradecimientos

Me gustaría agradecérselo a muchas personas, pido perdón por no haber podido incluir a todos los que se lo merecen, ¡no me olvido de vosotros!

En primer lugar quiero agradecer a mis dos directores, Jorge y Elvira, su inestimable ayuda al final de esta aventura que comenzó hace ya casi 7 años. Sin ellos no habría redescubierto mi pasión por la biología y comprendido que nuestra profesión es mucho más divertida y reconfortante cuando la pones al servicio de otras disciplinas, aunque muchos no comprendan por qué tu sueño no es trabajar en Google, Facebook o Twitter.

Desde mis inicios en el antiguo CPS he podido disfrutar de muy valiosas amistades. Javier, que siempre ha querido estar a mi lado aunque dejásemos de compartir asignaturas, intentando enseñarme la importancia de los valores éticos pese a mi supuesto fanboyismo. Rafael, que además de aguantar mis lamentos en más de una práctica comparte conmigo la idea de "Mens sana in corpore sano". Los compañeros de Púlsar e ISC, con los que he aprendido todo eso que no te enseñan dentro de las aulas y alguna otra cosa. Y a Andrés, César, Daniel, David, Jorge y Víctor, gracias por acompañarme de principio a fin, a pesar de todo.

Estaré eternamente agradecido a mis padres, que siempre me animaron a estudiar aquello que me apasionaba, aunque no fuese medicina; y con sacrificio y paciencia han costeado mi formación y me han soportado en los momentos más complicados. A mi hermana Blanca, que me enseñó que en esta vida hay que ser valiente y luchar por lo que te importa, aunque sea en el extranjero. A mi hermana María, que me enseña todos los días que no hay que tirar la toalla, por muchos obstáculos que tenga el camino. A mi hermano Lucas, compañero en muchas penas madridistas y alguna alegría mundialista, que me demuestra a diario que hay otra forma de hacer las cosas. A mi abuelo Antonio, que me enseñó la importancia de cultivar la lectura y el conocimiento, y que si hay algo que te apasiona, no debe terminar el día de tu jubilación. A mi abuela Blanca, que se esforzó mucho por mantener esa "conexión especial" que teníamos cuando era pequeño, y que es la única que me acompaña las tardes de los domingos, por muy mala que sea la película que escoja. Y a mi tía Pilar, que a su manera, siempre intenta que no me aleje demasiado del camino.

Y como dicen que los últimos serán los primeros... A Elena, que me recuerda todos los días lo maravillosa que puede llegar a ser la vida si encuentras a la persona adecuada y que no hay que tener miedo al cambio pues es la clave para continuar avanzando.

Índice general

Índice de figuras

Índice de tablas

1 Introducción

1.1 Contexto del proyecto

Este proyecto de fin de carrera se ha desarrollado en el Departamento de Informática e Ingeniería de Sistemas de la Universidad de Zaragoza, dentro del ámbito de la bioinformática [22], en un grupo de investigación formado por los dos directores de este proyecto y otros miembros del departamento. Pese a que este equipo lleva ya muchos años en activo, los esfuerzos han estado más centrados en otros temas, como por ejemplo, el análisis y construcción de filogenias; por lo que este proyecto extiende el ámbito de la investigación. Respecto al contexto biológico del proyecto, el cálculo del índice de conservación (IC) es uno de los instrumentos utilizados en el estudio de la patogenicidad de las mutaciones producidas en el código genético de los organismos. En el caso concreto de este proyecto, se ha trabajado no solo con el ADN mitocondrial (ADNmt de ahora en adelante); sino que también han sido objeto de estudio las secuencias de aminoácidos resultantes de la traducción de aquellos genes del ADNmt que codifican proteínas. La inclusión de secuencias proteicas en el análisis llevado a cabo en este proyecto se debe a que su estudio permite detectar con mayor precisión enfermedades raras y, muchas de ellas, asociadas a la muerte prematura del individuo [11]. La detección de estas enfermedades se basa en la localización de aquellas posiciones cuyo IC sea muy elevado pero se mantenga por debajo de 1[24].

1.2 Objetivos

Tras establecer el contexto de este proyecto, se procede a describir los objetivos que lo han guiado. La investigación se ha centrado en el desarrollo de herramientas software para calcular el IC, prestando especial atención a aquellos casos en los que se trabaje con grandes cantidades de datos, tomando como unidad de trabajo la secuencia de ADNmt. Pese a que en los últimos años se han ido desarrollando distintas herramientas que hacen uso del IC, cabe destacar que con el trabajo realizado en este proyecto se ofrecen herramientas adicionales que suponen una mejora considerable dotando de mayor profundidad a los estudios centrados en la patogenicidad de las

mutaciones. Esta profundidad se consigue especialmente cuando se combina el uso de las distintas herramientas desarrolladas, formando un sistema (al que se hará referencia de ahora en adelante).

El proyecto consta de dos objetivos fundamentales: introducción y formación, y creación de un sistema de cálculo de IC y generación de los informes correspondientes.

El primero de los objetivos queda inherente a todo proyecto, ya que suele ser imprescindible una pequeña fase de formación para adquirir los conocimientos necesarios para afrontarlo. El problema adicional que se ha encontrado en este proyecto es que no existe formación en estas temáticas durante la carrera, por lo que ha sido necesario un esfuerzo añadido desde el comienzo. Prácticamente este objetivo se ha mantenido durante todo el proyecto, con una ligera variación: al principio fue más guiada por los directores del proyecto y, conforme se ha ido acercando la fase final, se ha dispuesto de mayor grado de libertad para profundizar en aquellos temas que se considerasen convenientes para desempeñar el trabajo.

El segundo objetivo consiste en abordar el cálculo del IC, un estadístico muestral que indica la tasa de variabilidad de una posición concreta, para cada una de las posiciones de un conjunto de secuencias dado. El estudio de dicho valor permite determinar qué posiciones del conjunto de secuencias analizado sufren mutaciones y con qué frecuencia. Esto permite decidir a su vez cuáles son consideradas no neutrales y deben ser estudiadas en mayor profundidad, ya que pueden tener efectos perjudiciales para la supervivencia y reproducción de los organismos. Se recomienda al lector acudir al Capítulo 2 para obtener una definición de las mutaciones génicas y al Apéndice B si se desea conseguir información más detallada.

1.3 Metodología y herramientas

Como ya se ha comentado, la carencia de formación en contextos biológicos ha requerido un esfuerzo más intenso, tanto en dedicación como en tiempo de adquisición de conocimientos. De forma específica, se ha profundizado más en temas de métodos de alineamiento de conjuntos de secuencias, cálculo e interpretación del IC, traducción de secuencias de ADNmt a proteínas y mutaciones génicas.

De forma independiente se ha procedido a desarrollar cada uno de los módulos o herramientas que conforman el sistema. Todos ellos constan de una fase inicial en la que se ha determinado su diseño. A continuación se ha realizado un estudio del marco tecnológico más adecuado que ofreciese las mejores garantías para poder concluir las aplicaciones satisfactoriamente. Por último, una vez terminado el estudio y seleccionados los entornos de trabajo, se ha llevado a cabo la fase de implementación.

Se han realizado pruebas durante y tras la finalización de la implementa-

ción de cada una de las herramientas software desarrolladas con el objetivo principal no solo de detectar posibles errores en los resultados y poder corregirlos lo antes posible, sino también de depurar el formato de los informes generados.

Por último, cabe destacar la importancia de que los resultados finales del trabajo se han obtenido a partir de datos reales pertenecientes al proyecto ZARAMIT [5]. Tanto los alineamientos como los resultados obtenidos a partir del cálculo del IC han sido validados mediante la comprobación de algunos puntos comunes con las conclusiones de estudios biológicos realizados con anterioridad.

1.4 Software

Se ha trabajado con varias aplicaciones software a lo largo del proyecto. Las más utilizadas han sido la librería PhyloDAG (del inglés *Phylogenetic Direct Acyclic Graph*) desarrollada por uno de los directores de este proyecto, Jorge Álvarez, en su versión 1.0; y BioPython en su versión 1.63[10]. De la librería PhyloDAG se han utilizado tres de las herramientas que la componen: la que automatiza el proceso de obtención de conjuntos de secuencias, la que ofrece distintos métodos y configuraciones para llevar a cabo el alineamiento de secuencias biológicas, y la que permite realizar la división por genes.

Los conjuntos de secuencias con los que se ha trabajado provienen de una base de datos creada y depurada a partir de la sincronización básica con GenBank[2], una prestigiosa base de datos con una gran cantidad de información biológica (y un gran número de secuencias). Por otro lado, los alineamientos que se han realizado se han obtenido mediante el uso de dos de las herramientas más conocidas y utilizadas por los bioinformáticos hoy en día, Mafft en su versión 7.130b[16] y ClustalW en su versión 2.1[17], ambas incluidas en PhyloDAG. Por supuesto estas secuencias, una vez terminado el proceso de alineamiento, han pasado varias pruebas de control de calidad basadas en la detección de datos atípicos.

1.5 Entorno tecnológico

Para facilitar el trabajo se ha permitido el acceso a uno de los laboratorios de investigación del Grupo de Ingeniería de Sistemas de Eventos Discretos (GISED), el L1.03b, en el que se ha reservado un espacio físico y el computador Duero durante el periodo que ha durado la realización del proyecto.

1.6 Estructura de la memoria

La memoria se ha dividido en varias secciones y apéndices que se describen brevemente a continuación.

El primer capítulo tras esta introducción se ha titulado Glosario biológico. Como su nombre sugiere, en esta sección se encontrarán todos aquellos conceptos y definiciones, sobre todo de índole biológica, que se han considerado básicos y necesarios para poder comprender los siguientes capítulos en su totalidad. Se recomienda encarecidamente al lector que dedique unos minutos a su lectura.

En el siguiente capítulo se describe en detalle el conjunto de herramientas software desarrolladas en en este proyecto. Dicho capítulo comienza describiendo las motivaciones que han guiado su construcción y a continuación se explica el proceso que se ha seguido para la creación de dichas herramientas. Se termina detallando las pruebas realizadas y los resultados obtenidos.

En el último capítulo antes de los apéndices se recogen todas las conclusiones relacionadas con el proyecto.

El primero de los apéndices se centra en la descripción de la asignación de tiempo dedicada a cada parte del proyecto, incluyendo el diagrama de Gantt.

Debido a las limitaciones de longitud del cuerpo de la memoria, se ha incluido en el segundo apéndice un desarrollo más extenso sobre los fundamentos biológicos en los que se sustenta el proyecto.

El tercer apéndice se centra en describir de forma más detallada las motivaciones que guiaron la elección del lenguaje de programación empleado.

En el penúltimo apéndice se ha incluido el manual de usuario de la implementación realizada de las distintas herramientas construidas.

En el último apéndice se incluyen en un tamaño más legible la Figura 3.15 y la Figura 3.16.

Finalmente se adjunta la bibliografía consultada durante el proyecto.

2
Glosario biológico

En este capítulo se encuentra la definición de aquellos conceptos que se han considerado imprescindibles para la comprensión del proyecto. Para más detalles sobre aspectos biológicos el lector puede acudir al Apéndice B.

ADN: siglas en castellano de ácido desoxirribonucleico. Es una macromolécula formada por una doble cadena de nucleótidos (por lo general siguiendo una estructura de doble hélice) que forma parte de todas las células, y es usada para su desarrollo y funcionamiento. En ella se encuentra toda la información genética y es, por tanto, el componente responsable de la transmisión hereditaria.

ADN mitocondrial (ADNmt): en algunas células existen unos orgánulos denominados mitocondrias. En estos orgánulos se produce la oxidación de las moléculas de glucosa, obteniendo energía para la célula. Poseen ADN propio que gestiona sus funciones internas, siendo independiente del ADN nuclear. Este ADN posee unas características únicas que lo hacen idóneo para el estudio de la patogenicidad de las mutaciones, debido a su alta tasa de mutación y a su gran conservación entre organismos de la misma especie [14, 23, 29].

Gen: segmento de una secuencia de ADN o ARN que contiene toda la información necesaria para codificar un elemento funcional de la célula. Codifican proteínas o secuencias de ARN con objetivos muy específicos. El ADNmt animal contiene, salvo alguna excepción, 37 genes en su secuencia [6].

Proteína: macromolécula formada por cadenas de aminoácidos. Es fundamental para la vida, ya que puede desempeñar una gran cantidad de funciones básicas para el correcto funcionamiento del organismo (estructural, inmunológica, enzimática, etc.). 13 de los genes contenidos en el ADNmt animal codifican proteínas.

Mutación génica: alteraciones producidas en la secuencia de nucleótidos de un gen. Estos cambios pueden provocar a su vez modificaciones en las cadenas de aminoácidos que tengan como resultado efectos patógenos sobre el individuo. Dichas alteraciones se clasifican en dos grupos: neutrales y no neutrales. El primer grupo está formado por aquellas que no afectan ni a la supervivencia del organismo ni a su reproducción mientras que las

mutaciones no neutrales están formadas tanto por aquellas que tienen un efecto beneficioso como las que son perjudiciales.

Índice de conservación: estadístico muestral que permite medir la frecuencia con la que cada nucleótido o aminoácido (dependiendo del tipo de secuencia manejada) aparece en una posición concreta para un conjunto de secuencias dado. Dicho valor estadístico tiene especial importancia en el estudio de la patogenicidad de mutaciones.

Alineamiento: análisis y modificación de un conjunto de secuencias con el fin de conseguir la mayor cantidad posible de nucleótidos iguales en la misma posición. Para ello se realizan inserciones de huecos (habitualmente llamados *gaps*) en algunas de las secuencias.

Secuencia de referencia: secuencia, normalmente de ADN, de un individuo específico de una especie determinada que ha sido ampliamente estudiada, y por lo tanto, es muy poco probable que contenga errores. Suele existir una por cada especie. Además contiene información referente al principio y final de las secciones de cada uno de sus genes.

3

Automatización del Estudio del Índice de Conservación

En este capítulo se van a exponer de forma detallada tanto las motivaciones del proyecto como todo el trabajo realizado para el desarrollo de un sistema formado por distintas herramientas software que permitan el cálculo del índice de conservación (IC) de forma automática.

3.1 Estado del Arte

La motivación principal para el desarrollo de un conjunto de herramientas que permitan la automatización del cálculo del IC era la de ofrecer un potente instrumento adicional para facilitar y dotar de mayor profundidad a los estudios de mutaciones a lo largo de la evolución.

Este estadístico muestral se ha estado utilizando desde hace ya varios años para determinar si una mutación es neutral o no neutral [25, 26, 30]; e incluso se ha conseguido asociar algunas de estas mutaciones no neutrales con distintos tipos de cáncer, como el de mama o endometrio [7] o el gástrico [3]. También se han llevado a cabo otros estudios en los que se utiliza el IC junto con otros instrumentos para detectar una nueva mutación en un gen del ADNmt asociada a la pérdida auditiva hereditaria [20] o enfermedades cardiovasculares como la hipertensión [9] y la miocardiopatía no compactada [18, 28].

Aunque inicialmente el cálculo del IC se realizaba de forma manual, con el paso de los años se han ido desarrollando herramientas que facilitan el cálculo de este estadístico muestral de una forma cada vez más eficiente y automatizada. Ejemplos de estas herramientas son MITOMASTER [8] y MitoTool [12]. A pesar de ello, todavía existen limitaciones en dichas propuestas que sería beneficioso subsanar de cara a mejorar la precisión y eficiencia de los resultados. MITOMASTER se centra en el cálculo del IC interespecífico, es decir, entre individuos de distintas especies, obviando su estudio entre individuos de la misma especie; mientras que MitoTool realiza un cálculo del IC también únicamente entre individuos de distintas especies y además, fija esta comparación a 43 especies de primates, lo cual limita la

profundidad del estudio. El desarrollo de este tipo de sistemas es cada vez más importante debido al crecimiento exponencial que se está produciendo en cuanto al número y la variedad de las secuencias disponibles y que permite sacar conclusiones cada vez con una base más sólida. El objetivo de este proyecto es desarrollar un conjunto de herramientas que permitan realizar cálculos del IC no solo de forma interespecífica como MITOMASTER y MitoTool, sino también entre individuos de una misma especie. Además, el sistema que conforman este conjunto de herramientas permite realizar un análisis más profundo al ofrecer la posibilidad de trabajar con secuencias de aminoácidos. Los resultados de dicho estudio se presentan al personal investigador de una forma clara con el objetivo de facilitar su comprensión además de ofrecer la información necesaria para determinar cuáles deberían ser los objetivos de un estudio en mayor profundidad.

3.2 Diseño

La construcción del sistema se ha llevado a cabo teniendo muy en cuenta el esquema de caja negra y la idea de modularidad. El objetivo no era otro que el facilitar la comprensión por parte del usuario de qué es lo que hace cada una de las herramientas desarrolladas, sin prestar atención a cómo lo hacen. Al desarrollar módulos independientes no solo se consigue agilizar la comprensión global del sistema, sino que además se obtiene mayor robustez y se facilita su mantenimiento.

El sistema construido está compuesto por tres módulos principales:

1. Cálculo del IC y generación de informes.
2. Traducción de conjuntos de secuencias de nucleótidos a aminoácidos.
3. Combinación de informes para aquellos genes que codifican proteínas.

Se considera importante mencionar el hecho de que gracias a la colaboración de investigadores expertos en estos temas como Eduardo Ruiz Pesini y su doctorando Antonio Martín Navarro, se han podido aplicar sus conocimientos en la materia para determinar tanto la información útil que debía incluirse en dichos informes como el formato de los mismos.

A continuación se va a proceder a describir de forma detallada el diseño de cada uno de los módulos.

La Figura 3.1 muestra el diseño del primer módulo del sistema construido. Como ya se ha comentado anteriormente, cada módulo que se expone a lo largo de este capítulo en sus distintas fases puede operar de forma independiente. Durante la fase de diseño se ha visto la posibilidad de combinar todos los módulos para formar un sistema de análisis del IC en secuencias biológicas.

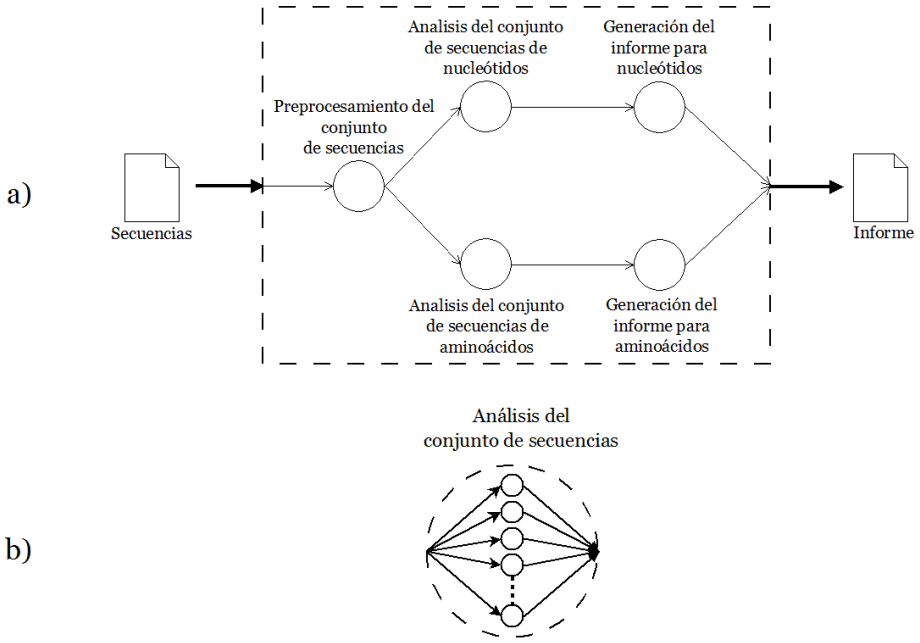

Figura 3.1: Diseño del primer componente del sistema. En el apartado a) se indica el diseño global del módulo: tras el preprocesamiento del conjunto de datos de entrada, realiza un análisis de la variabilidad de dichos datos y genera un informe adaptado a las necesidades indicadas por el usuario. El apartado b) refleja la paralelización de distinta granularidad según los datos que se manejan en la fase de análisis.

A continuación se va a proceder a detallar las operaciones internas que este primer módulo realiza. Como se puede apreciar en el apartado a) de la Figura 3.1, el primer paso consiste en realizar el preprocesamiento del conjunto de secuencias dado. Si se está trabajando con secuencias de nucleótidos, se ofrece la posibilidad de realizar una división por genes (en el caso de secuencias completas de ADNmt) o por secciones (tanto si se trata de secuencias completas de ADNmt como si son las asociadas a un gen concreto). Por otro lado, si se trabaja con secuencias de aminoácidos solo se permite la división por secciones ya que únicamente interesa estudiar las secuencias asociadas a los genes que codifican proteínas. En ambos casos, de forma

adicional se permite realizar un análisis global, es decir, sin realizar ningún tipo de división. La segunda fase del preprocesamiento consiste en realizar el alineamiento de las secuencias. Esta fase puede resultar muy compleja en función de varios aspectos (precisión, longitud de las secuencias, etc.) como se comenta en la Sección 3.4 y en la Sección 3.5.

Posteriormente se realiza el análisis del conjunto de secuencias indicado. Puesto que se ofrece la posibilidad de trabajar con distintos tipos de secuencias (nucleótidos o aminoácidos) es necesaria la inclusión de dos bloques diferenciados, cada uno con operaciones específicas para realizar el análisis del IC sobre un alfabeto distinto como se puede apreciar en la Tabla 3.1 y en la Tabla 3.2. El algoritmo en el que se basa el análisis consiste en contabilizar el número de veces que cada nucleótido o aminoácido aparece en cada una de las posiciones del conjunto alineado de secuencias y dividir el resultado por el número total de secuencias para obtener su valor del IC. En aquellos casos en los que el nucleótido o aminoácido contemple varias posibilidades, como por ejemplo el símbolo R en nucleótidos que equivale a un nucleótido G o a un nucleótido A, se le asigna a cada una de estas equivalencias un peso de forma equitativa (0.5 en este caso). Con este comportamiento se pretende penalizar de algún modo el hecho de no conocer con certeza el nucleótido que aparece en dicha posición. Merece especial atención el caso del *gap*, que tiene un peso de 0 para conseguir un efecto más penalizante sobre el cálculo del IC que el caso anterior, ya que dicho elemento representa la falta de información. Como se puede ver en el apartado b) de la Figura 3.1, dicho análisis se puede realizar de forma paralela ya que cada columna o sección del conjunto de secuencias se puede analizar de forma independiente bajo este estadístico. Pese a ello hay que tener muy en cuenta que solo resulta interesante su paralelización en caso de que la longitud de las secuencias así lo recomiende, especialmente en aquellos alineamientos en los que no se haya realizado la división de las secuencias por genes o secciones. Una vez se ha llevado a cabo el análisis de la variabilidad de cada uno de los genes o secciones en los que se ha dividido el conjunto de secuencias, se generan los informes correspondientes. En caso de que no se haya realizado ninguna división previa se generará un único informe. Además, también se calcula la secuencia más frecuente (SMF), que incluye en cada posición el nucleótido (o aminoácido) que tiene un IC más elevado, es decir, aquel que aparece más veces en dicha posición.

La Figura 3.2 muestra la estructura del segundo módulo del sistema. La motivación que guió la creación de este módulo consiste en realizar un análisis más profundo de los efectos que la mutación o mutaciones que ha sufrido el ADNmt provocan en los organismos. Dicho análisis permite prestar especial atención a las posiciones de las secuencias con un IC próximo al 100 %

Símbolo	Equivalencia	Significado	Peso
G	G	Guanina	1
A	A	Adenina	1
T	T	Timina	1
C	C	Citosina	1
R	G o A	Purina	0.5 por nucleótido
Y	T o C	Pirimidina	0.5 por nucleótido
M	A o C	Amina	0.5 por nucleótido
K	G o T	Cetona	0.5 por nucleótido
S	G o C	Interacción fuerte (enlaces 3 H)	0.5 por nucleótido
W	A o T	Interacción débil (enlaces 2 H)	0.5 por nucleótido
H	A o C o T	no G	0.33 por nucleótido
B	G o T o C	no A	0.33 por nucleótido
V	G o C o A	no T	0.33 por nucleótido
D	G o A o T	no C	0.33 por nucleótido
N	G o A o T o C	Cualquier nucleótido	0.25 por nucleótido
-	*gap*	Ningún nucleótido	0

Tabla 3.1: Elementos del alfabeto utilizado en secuencias de nucleótidos y el peso asignado.

pero sin alcanzarlo. En estas posiciones existe una mínima probabilidad de mutación, lo que hace que sea más interesante analizar si dicha mutación afecta a la secuencia de aminoácidos, en cuyo caso la probabilidad de que esta sea mortal aumenta considerablemente. Como se ha comentado brevemente en el Capítulo 2 y se describe en mayor detalle en el Apéndice B, existen mutaciones en el ADNmt que no tienen efectos perjudiciales para el organismo, bien por el hecho de que no provocan cambios en la secuencia de aminoácidos correspondiente, o porque dichos cambios no afectan a las funcionalidades de las proteínas.

En GenBank aunque el número de secuencias de ADNmt es muy elevado y sigue creciendo rápidamente, no se encuentran las traducciones a proteínas de todas estas secuencias. Por ello, se consideró necesario el desarrollo de una herramienta capaz de realizar la traducción de secuencias de nucleótidos y permitir un análisis más exhaustivo del IC.

A la hora de diseñar esta herramienta se ha empleado como guía el proceso biológico que realiza la célula. Recordar que solo tiene sentido realizar esta traducción sobre el conjunto de secuencias asociadas a aquellos genes que codifican proteínas (13 de los 37 genes que se encuentran en el ADNmt animal).

Al igual que en el primer módulo, se incluye una fase de preprocesamiento. Sin embargo, existen ciertas diferencias: la división por genes solo

Símbolo	Equivalencia	Significado	Peso
G	G	Glicina	1
A	A	Alanina	1
V	V	Valina	1
L	L	Leucina	1
I	I	Isoleucina	1
P	P	Prolina	1
F	F	Fenilalanina	1
Y	Y	Tirosina	1
C	C	Cisteína	1
M	M	Metionina	1
H	H	Histidina	1
K	K	Lisina	1
R	R	Arginina	1
W	W	Triptófano	1
S	S	Serina	1
T	T	Treonina	1
D	D	Ácido aspártico	1
E	E	Ácido glutámico	1
N	N	Asparagina	1
Q	Q	Glutamina	1
B	D o N	Ácido aspártico o Asparagina	0.5 por aminoácido
Z	E o Q	Ácido glutámico o Glutamina	0.5 por aminoácido
-	-	Terminador	1
X	X	Desconocido	1
=	*gap*	Ningún aminoácido	0

Tabla 3.2: Elementos del alfabeto utilizado en secuencias de aminoácidos y el peso asignado.

selecciona aquellos que codifican proteínas, y no está permitida la división por secciones. Después se lleva a cabo la traducción a proteínas, que consiste en traducir cada triplete o codón (se denomina así a los conjuntos de tres nucleótidos) en su aminoácido correspondiente (los 64 tripletes posibles codifican solo 20 aminoácidos distintos). En la Tabla B.1 se indican los codones que codifican cada uno de los aminoácidos. Como se puede ver en el apartado b) de la Figura 3.2, dicha traducción se puede realizar de forma paralela ya que la traducción de cada secuencia del conjunto se puede llevar a cabo de forma independiente.

A continuación se va a proceder a describir las operaciones internas que realiza el tercer módulo, cuya estructura se muestra en la Figura 3.3. El

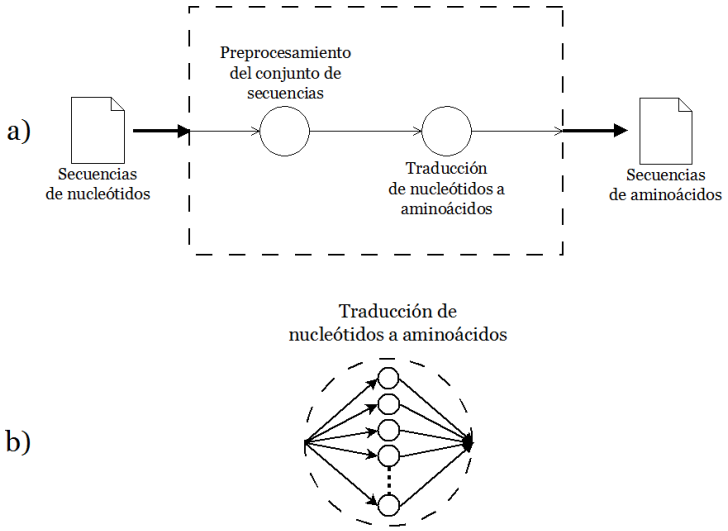

Figura 3.2: Diseño del segundo componente del sistema. En el apartado a) se indica el diseño global del módulo: tras el preprocesamiento del conjunto de secuencias de ADNmt de entrada, realiza la traducción de dichas secuencias a proteínas. El apartado b) refleja la paralelización de distinta granularidad según los datos que se manejan en la fase de traducción.

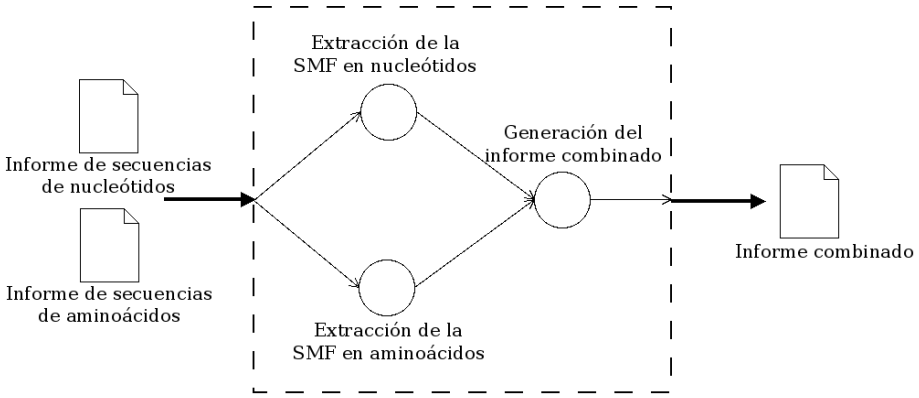

Figura 3.3: Diseño del tercer componente del sistema. Realiza la combinación en un único informe de los generados previamente tanto para las secuencias de nucleótidos como de aminoácidos.

primer paso consiste en extraer la SMF de los informes tanto de nucleótidos como de aminoácidos. Para obtener dichos informes es necesaria la invo-

cación del primer módulo dos veces, una con el conjunto de secuencias de nucleótidos y otra con el conjunto de secuencias de aminoácidos obtenido con el segundo módulo desarrollado. Esta primera etapa se realiza de forma paralela ya que son procesos totalmente independientes. Por último, tras realizar la extracción de dichas secuencias, se procede a la generación del informe combinado, que permite estudiar de una forma más cómoda cuales de las mutaciones que aparecen en las secuencias de nucleótidos producen una modificación en las de aminoácidos.

3.3 Implementación

La implementación del sistema se ha llevado a cabo utilizando principalmente el lenguaje de programación Python. Una de las razones más importantes de su elección ha sido que dispone de librerías muy potentes como BioPython que simplifican el trabajo realizado. Los motivos de la elección del lenguaje de programación se estudian en profundidad en el Apéndice C.

Gracias a la modularidad que Python ofrece, ha sido posible implementar cada uno de los bloques del sistema bajo la misma idea de caja negra que se lleva mencionando a lo largo de este capítulo.

En la Figura 3.4, la Figura 3.7 y la Figura 3.8 se encuentran los esquemas a nivel de implementación de los diseños desarrollados anteriormente.

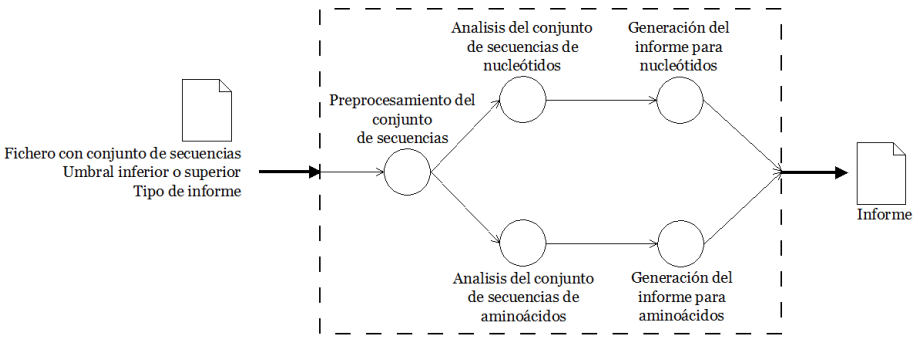

Figura 3.4: Primer componente del sistema. Tras realizar el preprocesamiento del conjunto de secuencias, realiza el análisis del IC de las mismas y genera un informe adaptado a las necesidades del usuario.

Las entradas principales de este primer módulo, como se puede apreciar en la Figura 3.4, son: el fichero en el que se encuentra almacenado el conjunto de secuencias, el umbral que indica el límite inferior o superior del IC y el tipo de informe que se desea obtener. Si el análisis se ha llevado a cabo de forma global, la única salida generada por este primer módulo es el fichero con el informe (básico o detallado). En el caso de haber realizado algún tipo

de división (por genes o secciones), se creará un fichero para cada uno de los informes generados. A continuación se describe de forma más detallada el formato y la información contenida en cada una de las posibilidades.

Como se ha comentado previamente, el hecho de poder conocer las necesidades y preferencias de los biólogos ha permitido establecer un diseño y formato de los informes que facilita enormemente la interpretación de la información disponible en los mismos por parte de los investigadores.

```
10470: MPLIYINIILAFTISLLGILVYRSHLISSLLCLEGIILSLFIIATLITLN
       ·························+····················

10620: THSLLANIVPIAILVFAACEAAVGLALLVSISNTYGLDYVHNLNLLQC-
       ·················································

There is 1 column with a high degree of variation (< 95.00%) in the peptide sequence:

> 10548 ('X':   0.0871%, 'I':  94.7155%, 'L':   0.0131%, 'M':   5.1799%, '=':   0.0044%)
```

Figura 3.5: Informe básico del IC para un conjunto de secuencias de aminoácidos del gen ND4L y un umbral superior del 0.95.

En la parte superior del informe básico que se muestra en la Figura 3.5 se puede apreciar la secuencia de aminoácidos más frecuente para un conjunto de secuencias de proteínas. El número que acompaña cada sección de 50 aminoácidos por línea indica la posición absoluta del primer aminoácido de dicha sección respecto de la secuencia completa de ADNmt humano. La línea inmediatamente inferior a cada una de estas secciones ofrece información sobre el valor del IC de cada posición en relación con el umbral establecido. Se indica con un '·' aquellas posiciones en las que el aminoácido más frecuente tiene un porcentaje de aparición superior (o inferior, si se hubiese marcado un umbral superior) al umbral establecido (0.95 en este caso). Por otro lado, se marca con un '+' aquellas posiciones de la sección en las que dicho aminoácido tiene un porcentaje de aparición inferior (o superior, si se hubiese marcado un umbral inferior) a dicho umbral. En los informes básicos asociados a conjuntos de secuencias de nucleótidos el formato y la información ofrecida es la misma.

La parte inferior de la Figura 3.5 indica aquellas posiciones en las que el aminoácido más frecuente tiene un porcentaje de aparición inferior (o superior en caso de indicarse un umbral inferior). Además se ofrece un informe más detallado de su IC en el que se incluyen todas aquellos aminoácidos que han aparecido por lo menos una vez en dicha posición así como sus porcentajes de aparición. Los *gaps* introducidos durante el proceso de alineamiento vienen indicados por el carácter '=' ('-' en el caso de que el análisis se centre en conjuntos de secuencias de nucleótidos).

Por otro lado, la Figura 3.6 muestra un fragmento de la parte superior de un informe detallado. En él se ofrece de forma pormenorizada, para cada posición, el IC de cada uno de los nucleótidos. En la parte inferior del informe

12207:	A	0.0022%	C	0.0022%	G	99.9891%	T	0.0022%
12208:	A	99.9924%	C	0.0011%	G	0.0011%	T	0.0011%
12209:	A	0.0022%	C	0.0022%	G	99.9891%	T	0.0022%
12210:	A	99.9848%	C	0.0022%	G	0.0065%	T	0.0022%
12211:	A	99.9858%	C	0.0033%	G	0.0033%	T	0.0033%
12212:	A	99.9891%	C	0.0022%	G	0.0022%	T	0.0022%
12213:	A	0.0033%	C	0.0033%	G	99.9858%	T	0.0033%
12214:	A	0.0022%	C	99.9891%	G	0.0022%	T	0.0022%
12215:	A	0.0033%	C	0.0447%	G	0.0033%	T	99.9445%
12216:	A	0.0022%	C	99.9848%	G	0.0022%	T	0.0065%
12217:	A	99.9379%	C	0.0120%	G	0.0425%	T	0.0033%

Figura 3.6: Fragmento de un informe detallado del IC para un conjunto de secuencias de ADNmt humano del gen Ser2 y un umbral del 0.5.

detallado no se añade ni se modifica información a la que ya se incluye en el informe básico.

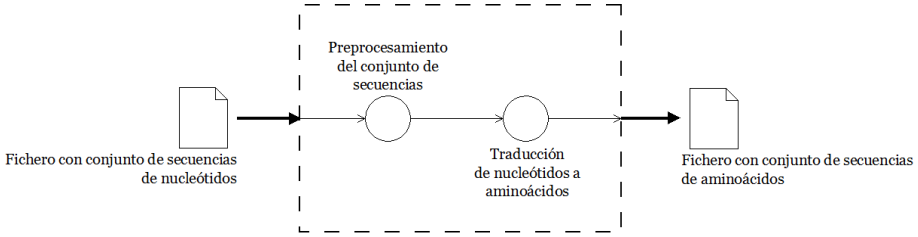

Figura 3.7: Segundo componente del sistema. Realiza la traducción de un conjunto de secuencias de ADNmt a proteínas.

Antes de decidir realizar una implementación propia de la traducción de secuencias de nucleótidos a secuencias de aminoácidos se han revisado algunas de las aplicaciones existentes. La única que se ha valorado ha sido la incluida en BioPython por su fácil disponibilidad. Tras su evaluación se ha detectado que no soporta secuencias que incluyan *gaps*. Sin embargo, en este tipo de estudios es necesario realizar un preprocesamiento de los conjuntos de secuencias, lo cual produce de forma muy frecuente la inserción de *gaps* en ellas. Por este motivo se ha optado por el desarrollo de una nueva herramienta.

El segundo módulo, como se indica en la Figura 3.7, tiene como entrada principal el fichero que almacena el conjunto de secuencias de nucleótidos que se desean traducir. La única salida que genera consiste en un fichero que almacena las secuencias de aminoácidos resultantes. Para poder llevar a cabo esta traducción es necesaria la comprobación de la longitud de las secuencias de nucleótidos, ya que deben ser múltiplos de tres. Como sucede en el proceso biológico, en el caso de que no lo sean, se realiza la inserción en el extremo de terminación de uno o dos nucleótidos 'A' según corresponda, con el objetivo de generar un codón de terminación que como su nombre

indica, establece el final de la secuencia a traducir.

Por último, en la Figura 3.8 se puede apreciar que el tercer módulo del sistema tiene como únicas entradas los ficheros en los que se almacenan los informes para las secuencias de nucleótidos y sus traducciones a aminoácidos. Se recuerda de nuevo al lector que solo interesa utilizar este módulo para combinar los informes de las secciones de las secuencias asociadas a aquellos genes que codifican proteínas (13 de 37 en el ADNmt). La única salida que genera este módulo es un fichero que almacena el informe combinado.

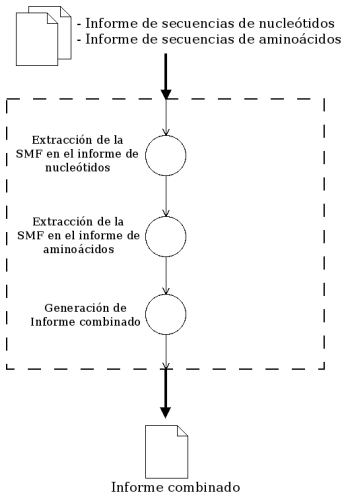

Figura 3.8: Tercer componente del sistema. Genera un informe en el que se incluye la información obtenida en el análisis del IC realizado por el primer módulo del sistema para secuencias de ADNmt y su traducción a aminoácidos.

En la Figura 3.9 se puede apreciar un fragmento de la parte superior de un informe combinado (salida del último módulo). En él se incluye la misma información que en los correspondientes informes para conjuntos de secuencias de nucleótidos y aminoácidos de un gen del ADNmt humano que codifique una proteína. La ventaja de este tipo de informes consiste en la posibilidad de ver en un único fichero tanto la secuencia de nucleótidos como su traducción a proteínas, facilitando así el estudio en profundidad de la repercusión de las mutaciones en las secuencias de ADNmt.

Finalmente se procede a describir el montaje del sistema completo mediante dos ejemplos. En el primero de ellos se dispone inicialmente de un conjunto de secuencias de ADNmt y mediante el primer módulo se realiza un análisis del IC tras alinear las secuencias con división por genes, y se obtienen 37 informes distintos (primera fase). A continuación, utilizando el segundo módulo se obtiene la traducción de las secciones asociadas a los

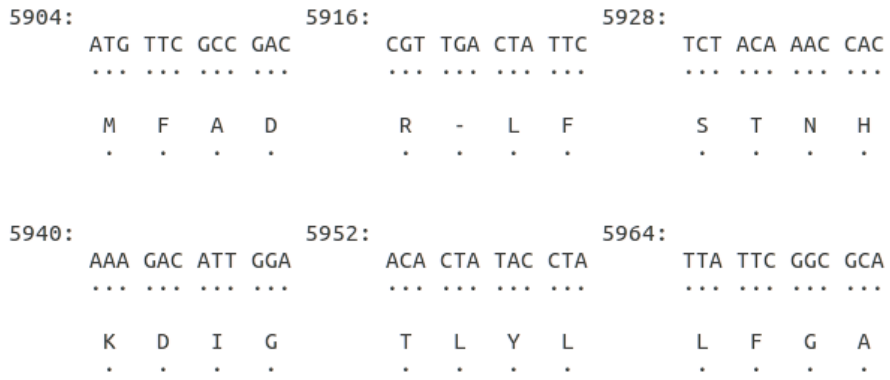

Figura 3.9: Fragmento de un informe combinado del IC para un conjunto de secuencias de ADNmt humano y proteínas del gen CO1, para un umbral superior del 0.75.

13 genes que codifican proteínas a secuencias de aminoácidos (segunda fase). Estas secuencias de aminoácidos son analizadas por el primer módulo para estudiar su IC y se generan 13 nuevos informes (tercera fase). Por último, seleccionando únicamente aquellos informes asociados a los 13 genes que codifican proteínas, se combinan los informes obtenidos en la primera y tercera fase obteniendo 13 nuevos informes. En el segundo ejemplo se dispone inicialmente de un conjunto de secuencias asociado a uno de los 13 genes que codifican proteínas. En este caso basta con realizar un análisis global mediante el primer módulo realizando previamente un alineamiento sin divisiones.

El Apéndice D contiene el manual de usuario y en él se describen de una forma más detallada las entradas y salidas de cada uno de los módulos desarrollados así como algunos consejos de uso de las herramientas.

3.4 Pruebas

Los datos que se han utilizado durante la fase de pruebas se dividen en dos grandes grupos: 22954 secuencias de ADNmt humano y 442 secuencias de ADNmt primate, con las correspondientes traducciones a proteínas.

Las herramientas utilizadas para realizar el preprocesamiento de las secuencias han sido Mafft y ClustalW, por los motivos ya comentados en la Sección 1.4.

Hay que tener muy presente que los diferentes métodos de alineamiento de secuencias con sus distintas configuraciones permiten obtener resultados más o menos precisos, lo cual tiene una gran repercusión en el tiempo de ejecución de los mismos. En el caso de la herramienta Mafft, se han seleccionado

las siguientes configuraciones por ser las más comunes en los estudios bioinformáticos. La configuración *parttree*, que tiene como objetivo la rapidez en la ejecución. La configuración *linsi*, que tiene como objetivo la precisión de los resultados obtenidos. Por último, la configuración *auto* selecciona en función de las características de los datos de entrada la opción más adecuada de entre 3 posibles: la primera opción se centra en la precisión de los resultados obtenidos, la segunda tiene como objetivo minimizar el coste temporal y la tercera establece un equilibrio entre las dos anteriores. Hay que aclarar que, obviamente, ninguna de las opciones disponibles en Mafft - *auto* equivale a las 2 configuraciones previamente mencionadas, sino que se encuentran en un punto intermedio. Para poder realizar todas las pruebas deseadas ha sido necesaria la adaptación de la herramienta de división por genes de la librería PhyloDAG, que en su versión actual utiliza Mafft - *auto* para generar los alineamientos.

El motivo por el que en la comparativa con secuencias humanas solo aparecen los métodos Mafft - *auto* y ClustalW es que el principal objetivo de dicha prueba ha sido el de verificar si el tiempo adicional que requiere ClustalW permite obtener resultados más precisos. Por otro lado, en la comparativa con secuencias de primates, debido a limitaciones temporales, se ha realizado la equiparación únicamente entre las distintas configuraciones de la herramienta Mafft.

Como se puede apreciar en la Figura 3.10 y la Figura 3.11, el alineamiento con división de genes utilizando Mafft en su configuración *auto* tiene un tiempo de ejecución dos órdenes de magnitud inferior al mismo alineamiento utilizando ClustalW para las 22954 secuencias de ADNmt humano. Por otro lado, pese a utilizar un mismo método de alineamiento (Mafft en este caso particular), el emplear configuraciones distintas genera importantes variaciones en lo que respecta a los tiempos de ejecución y por lo general, en la precisión de los alineamientos obtenidos. Otra de las cuestiones que hay que tener muy en cuenta a la hora de trabajar con secuencias de distintas especies (como ha sucedido en el caso de los primates) es la alta variabilidad que existe entre ellas, que tiene efectos importantes en las propiedades ya mencionadas de los alineamientos.

Por lo tanto, resulta evidente que antes de realizar un alineamiento hay que considerar el número de secuencias que se pretende alinear y su longitud, así como la precisión que se espera obtener y las limitaciones temporales que puedan existir.

Uno de los obstáculos que ha habido que superar se encontraba en la fase de obtención de los conjuntos de secuencias de GenBank, de la que ya se ha hablado en la Sección 1.4. Ha sido necesario llevar a cabo un estudio en profundidad para construir las consultas necesarias para obtener los conjuntos de secuencias de ADNmt que nos interesan; filtrando tanto los fragmentos de secuencias, como aquellas secuencias que carecen de toda la información necesaria para el correcto funcionamiento de las herramientas.

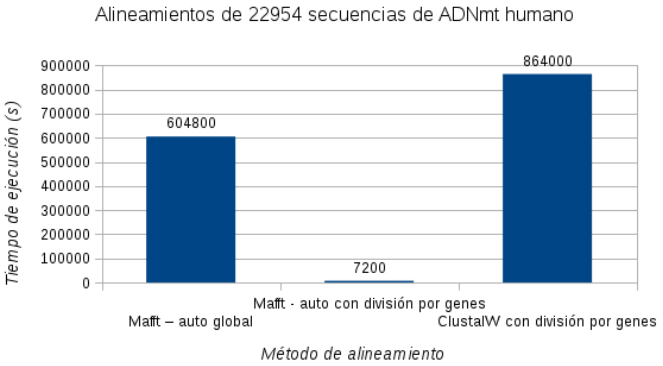

Figura 3.10: Tiempo de ejecución en segundos de los diferentes alineamientos de 22954 secuencias de ADNmt humano.

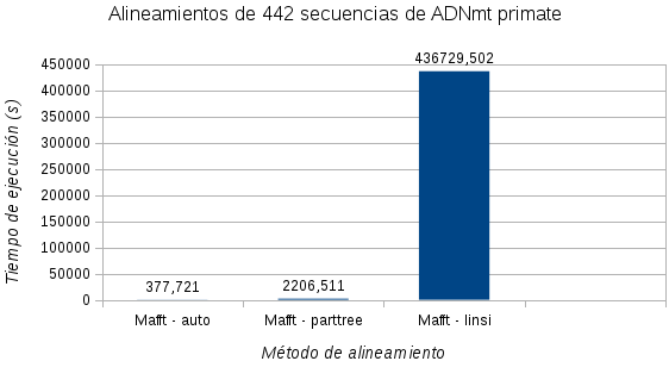

Figura 3.11: Tiempo de ejecución en segundos de los diferentes alineamientos de 442 secuencias de ADNmt primate.

El principal motivo por el que se han omitido los fragmentos en esta versión del sistema es porque su alineamiento supone un problema muy complejo y su tratamiento supera los objetivos de este proyecto.

Cabe destacar que durante el proceso de construcción de las consultas, han sido detectados varios errores de notación en algunas de las secuencias disponibles en GenBank que han sido notificados para su futura corrección. La existencia de este tipo de secuencias inválidas en los primeros conjuntos de ADNmt primate descargados hizo que el alineamiento global no tuviese ninguna validez práctica y fuese necesaria una división por familias taxonó-

micas como etapa intermedia. En estas primeras pruebas de alineamiento global se generaba la inserción de más de 300000 *gaps*, lo cual supone un incremento de 20 veces la longitud de la secuencia más larga en el conjunto de secuencias original. Tras resolver este problema, se han conseguido alinear las secuencias de ADNmt primate de forma global, sin que fuese necesaria la división por familias taxonómicas.

Durante el proceso de obtención de conjuntos de secuencias de ADNmt primate se detectó que para algunas especies existen varias secuencias de referencia. Esto supone un gran problema de cara a la validez de los resultados ya que no existe un protocolo biológico para seleccionar una secuencia de referencia en el caso de que existan varias para la misma especie. Debido al coste que supone resolver este problema (formación y pruebas) ha quedado pendiente el desarrollo de mecanismos que permitan llevar a cabo dicho proceso.

Por último, también se ha llevado a cabo la actualización de la herramienta de PhyloDAG encargada de descargar los conjuntos de secuencias de GenBank, para obtener información adicional referente a la localización de cada gen en las secuencias de ADNmt. Esto ha permitido simplificar el proceso de división por genes y alineamiento de las secuencias. Como se puede apreciar en la Figura 3.12 y la Figura 3.13, a pesar de que se sufre una penalización en cuanto al tiempo de descarga de las secuencias, el espacio en disco que ocupan los ficheros que las almacenaban y el consiguiente tiempo de lectura de dichos ficheros, el ahorro en costes posterior resulta muy superior y beneficioso, especialmente a la hora de resolver el problema que se ha comentado previamente sobre las múltiples secuencias de referencia asociadas a una misma especie.

3.5 Resultados

Uno de los objetivos principales en la construcción de las herramientas desarrolladas a lo largo de este proyecto era el poder realizar análisis del IC con grandes cantidades de datos. El coste de realizar dicho análisis de forma manual resultaba inadmisible desde el punto de vista de los investigadores.

Antes de que comenzase este proyecto, el alineamiento más grande realizado por el grupo de bioinformática de la Universidad de Zaragoza era el llevado a cabo en el proyecto ZARAMIT (centrado en la construcción de árboles filogenéticos), que contaba con 7390 secuencias de ADNmt humano. Además, dicho alineamiento se realizó mediante una reconstrucción *bottom-up* con la ayuda de un árbol filogenético, lo que simplifica el problema pero exige que se disponga de un árbol filogenético que contenga todas las secuencias que se vayan a alinear. En este proyecto se ha estado trabajando con alineamientos de 22954 secuencias de ADNmt humano para las cuales no se ha construido todavía ningún árbol filogenético. Además en el tramo

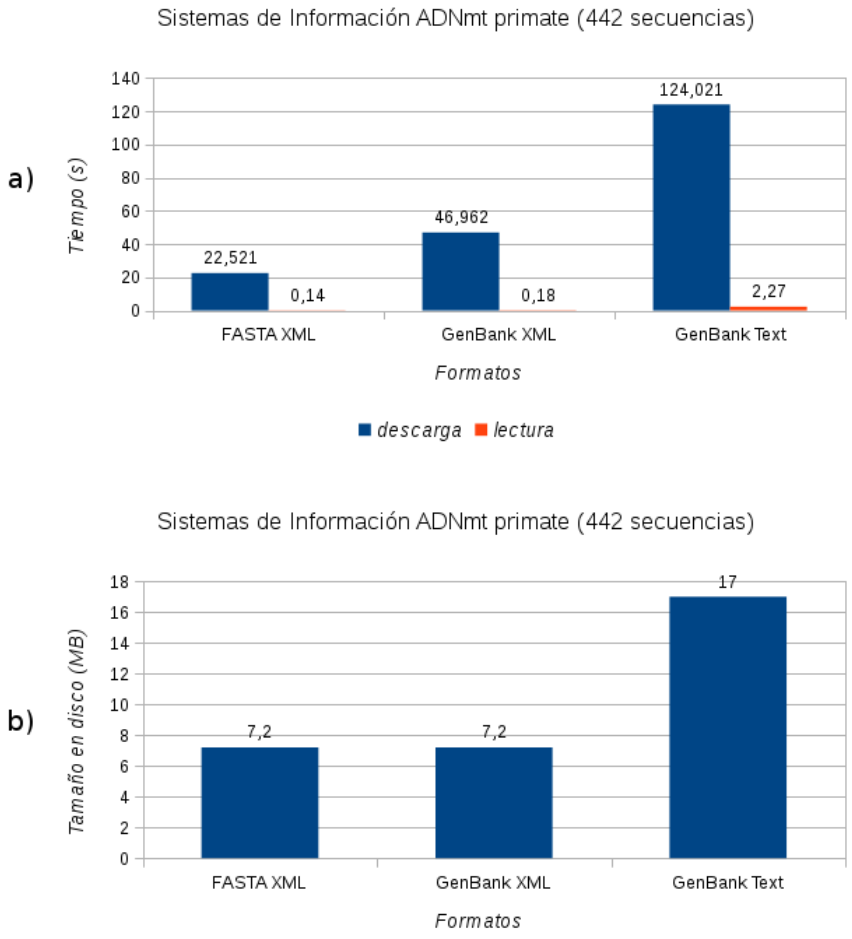

Figura 3.12: Análisis de descarga y almacenamiento de 442 secuencias de ADNmt primate: a) Tiempos de descarga y lectura en segundos; b) Tamaño del fichero en MB.

final se ha creado otro grupo de estudio en el que se incluyen 442 secuencias de ADNmt de distintas especies de primates, en las que se han excluido a los seres humanos. Los últimos esfuerzos se han enfocado en dotar al sistema de los mecanismos necesarios para realizar un tratamiento gen a gen no solo en humanos, sino también en el resto de las especies animales.

Como se ha comentado en la Sección 3.4, la decisión de utilizar una herramienta de alineamiento u otra afecta, entre otras cosas, a la precisión del resultado obtenido, como queda reflejado en la Figura 3.14.

En el caso de las secuencias de ADNmt humano, existen dos secuencias

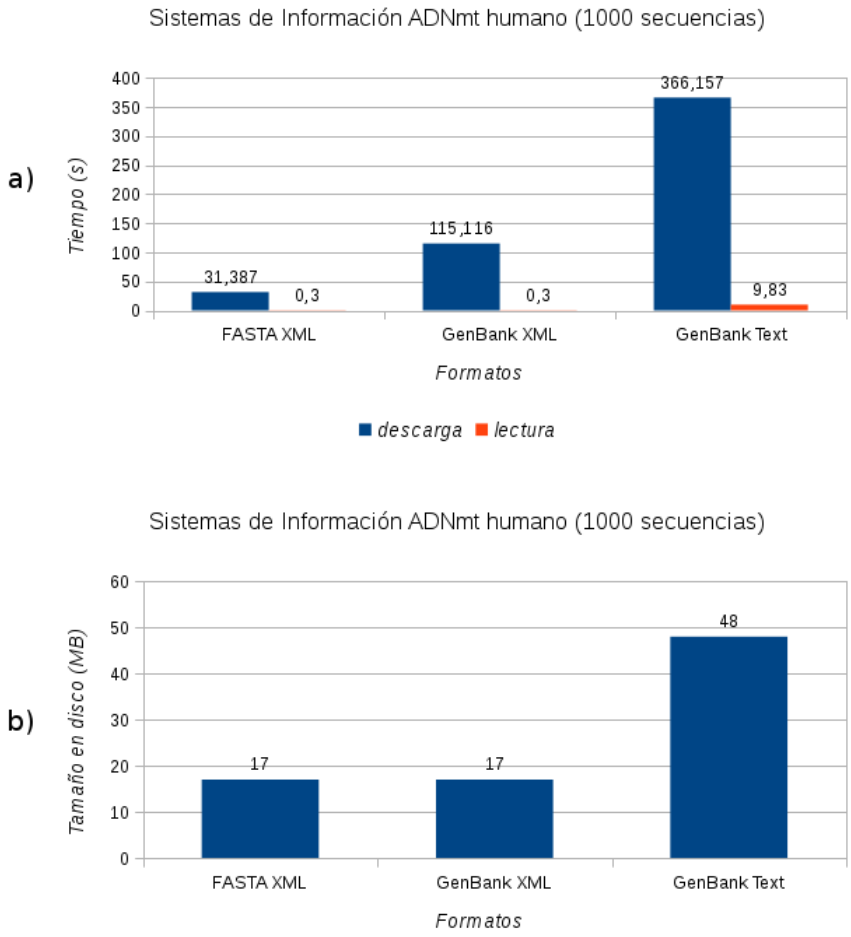

Figura 3.13: Análisis de descarga y almacenamiento de 1000 secuencias de ADNmt humano: a) Tiempos de descarga y lectura en segundos; b) Tamaño del fichero en MB.

de referencia (rCRS y RSRS), lo que genera una disyuntiva biológica. La rCRS es una revisión de la primera secuencia de referencia publicada para este tipo de secuencias y la segunda es una secuencia sintética creada en el año 2012 que se sitúa en la raíz del árbol filogenético del ADNmt humano. Puesto que la disyuntiva todavía no se ha resuelto se han duplicado las pruebas realizadas con el objetivo de comprobar si existe algún tipo de diferencia entre los resultados obtenidos con cada una de ellas. A la vista de los resultados, se ha decidido descartar el obtener más resultados con la RSRS ya que no se han detectado diferencias que justifiquen el coste del

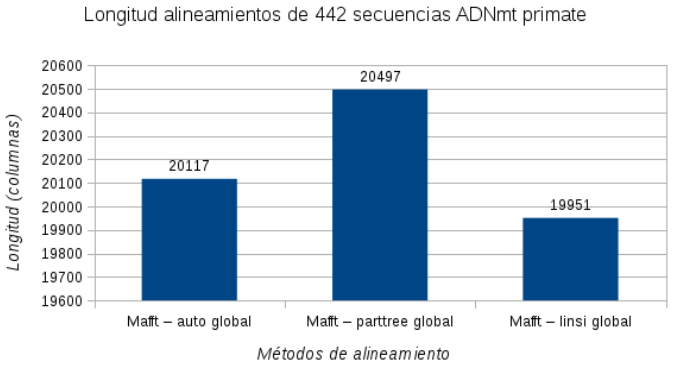

Figura 3.14: Longitud de los alineamientos de 442 secuencias de ADNmt primate.

doble análisis, como se puede apreciar en la Figura 3.15 y la Figura 3.16.

Como se ha comentado al comienzo de la Sección 3.3, se ha realizado un estudio estadístico del IC en cada gen del ADNmt humano con la ayuda de una de las herramientas incluidas en la librería PhyloDAG. En la Figura 3.15 y la Figura 3.16 se muestran los resultados obtenidos normalizados dividiendo los valores absolutos de las secciones asociadas a cada gen entre su longitud.

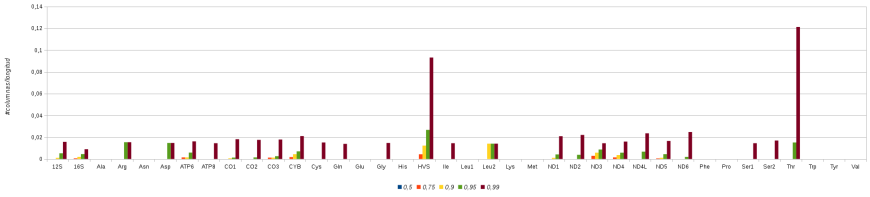

Figura 3.15: Gráfica con los resultados del cálculo del IC del alineamiento normalizado de 22954 secuencias de ADNmt humano con división por genes utilizando Mafft - *auto* y la secuencia de referencia rCRS.

Como se puede apreciar en la Figura 3.15 y la Figura 3.16 el gen de la Treonina (Thr) consta de un número de posiciones con una alta tasa de variabilidad muy superior al resto. Tras contactar con el personal investigador biólogo se confirmó que en dicho gen existen una serie de mutaciones génicas conocidas. Solo en la región de control (también conocida como región hipervariante) se pueden encontrar valores similares. La hipótesis más aceptada actualmente es que esta región no contiene información genética por lo que

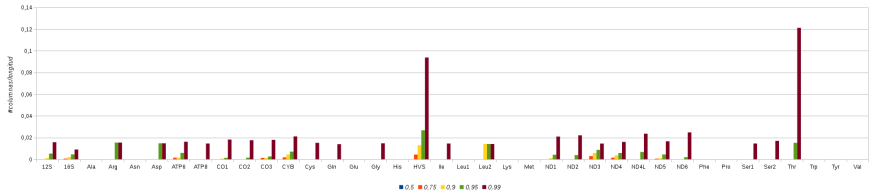

Figura 3.16: Gráfica con los resultados del cálculo del IC del alineamiento normalizado de 22954 secuencias de ADNmt humano con división por genes utilizando Mafft - *auto* y la secuencia de referencia RSRS.

el encontrar posiciones altamente conservadas puede resultar interesante.

Por otro lado, hay que tener en cuenta que la elección de un método de alineamiento concreto no solo afecta al tiempo de ejecución como se ha visto en la Figura 3.10 y Figura 3.11, sino que como se ha comentado anteriormente, la precisión del alineamiento también varía y por lo tanto, los IC son objeto de modificaciones debido a este problema.

4
Conclusiones

4.1 Trabajo realizado

El trabajo realizado supone una nueva aportación al conjunto de herramientas software para estudios de patogenicidad de mutaciones, extendiendo su ámbito a las secuencias de aminoácidos y a especies distintas del ser humano con el objetivo de profundizar en el estudio de los efectos que las mutaciones tienen en los organismos. Se han cumplido satisfactoriamente no solo aquellos objetivos que se plantearon inicialmente, sino también aquellos que han ido surgiendo a lo largo del desarrollo de este proyecto.

4.2 Con vistas al futuro

Aunque el sistema desarrollado se considera por si solo un potente instrumento adicional para el estudio de la patogenicidad de mutaciones, existen varios aspectos en los que se va a continuar trabajando. Se considera especialmente importante desarrollar los mecanismos necesarios para establecer una correlación entre las secuencias de las distintas especies y poder realizar estudios del IC interespecífico. Otro de los aspectos más interesantes es el de ampliar los estudios realizados a otras especies, no solo primates como se ha hecho en la segunda mitad de este proyecto, sino a otros grandes grupos como pueden ser los mamíferos. El hecho de manejar conjuntos de secuencias con mayor variabilidad entre ellas hace que las mutaciones en aquellas posiciones en las que el IC esté cercano al 100 % sin llegar a alcanzarlo, tengan mayores probabilidades de ser mortales. Como se ha comentado con anterioridad, el hecho de utilizar GenBank como fuente de los conjuntos de secuencias biológicas provoca que la información que obtenemos no sea todo lo homogénea que nos gustaría, ya que se trata de un proyecto colaborativo. Una de las consecuencias de esto consiste en que no todas las secuencias cuentan con la sección hipervariante ya que entre la comunidad científica estuvo extendida durante mucho tiempo la creencia de que dicha sección no contenía información biológica relevante, por lo que no resultaba beneficioso su secuenciación tanto desde el punto de vista académico como desde el punto de vista aplicado. A este respecto, se considera interesante

añadir la capacidad de detectar aquellas secuencias que no cuenten con esta sección ya que la ausencia de información por este motivo no debería tener el mismo efecto penalizante en el cálculo del IC que aquellos generados en el proceso de alineamiento. En los estudios llevados a cabo hasta la fecha el número de secuencias sin región hipervariante ha sido ínfimo por lo que no se ha considerado necesario, pero con vistas a un futuro próximo en el que los estudios se amplíen a otras especies o se incluyan los fragmentos de secuencias disponibles, resultará conveniente. Con el objetivo de extender y facilitar el uso de dichas herramientas, también se ha pensado en ofrecer la capacidad de trabajar con conjuntos de secuencias almacenadas en formatos distintos a FASTA, que pese a ser el más extendido, no es el único manejado por la comunidad bióloga. Por último, pensando en adaptar el sistema para que pueda trabajar con secuencias de ADN nuclear, se considera también importante implementar aquellos mecanismos de paralelización que durante la etapa de diseño fueron valorados. Quedaron descartados en la implementación por resultar innecesarios al trabajar con secuencias de ADNmt, con una longitud en el caso peor inferior a los 22000 nucleótidos (tras la fase de preprocesamiento). Sin embargo, al plantear la aplicación del sistema a secuencias de ADN nuclear, que en el caso de los seres humanos tiene una longitud total aproximada de 3200 millones de nucleótidos, estos mecanismos resultan muy necesarios. Para que el lector se haga una idea más clara de la magnitud del problema que se está describiendo, se ofrece el siguiente ejemplo: en el peor caso antes mencionado, los tiempos de ejecución rondaban los 25 minutos para el análisis del IC. Puesto que el algoritmo tiene un coste temporal lineal, esto supone que en el caso de disponer del mismo número de secuencias de ADN nuclear humano, este mismo análisis tendría un tiempo de ejecución superior a 7 años.

4.3 De lo profesional a lo personal

Desde el punto de vista profesional debo realizar una valoración muy positiva del trabajo que se ha llevado a cabo a lo largo de este proyecto ya que se ha conseguido desarrollar un sistema completo y robusto que pueda ser utilizado como herramienta de trabajo por los investigadores, a pesar de los distintos obstáculos que se han ido encontrando a lo largo del camino. Lo más valioso ha sido la posibilidad de demostrar la capacidad de trabajo, esfuerzo y aprendizaje, no solo en cuestiones relacionadas con la informática, sino aquellas que tienen un corte más biológico. Todo esto ha desembocado en la posibilidad de disfrutar de un contrato de 3 meses de duración en el mismo grupo donde se ha desarrollado este trabajo dentro del proyecto de investigación BASMATI y queda todavía pendiente la participación en la redacción de dos artículos de investigación de futura publicación, lo que ha hecho que se considere cursar un máster para continuar la formación en estos

temas o incluso la realización de una tesis doctoral. Pese a que al empezar en la carrera de Ingeniería Informática mi interés por la biología quedó enterrado, debo agradecer a mis directores que me ofreciesen la posibilidad de realizar este proyecto, que me ha hecho recuperar mi pasión por la biología y descubrir el poder de la informática cuando la pones al servicio de otras especialidades. Por otro lado, desde el punto de vista personal, me he sentido muy cómodo, valorado y acogido en el grupo de bioinformática, permitiéndome colaborar de forma activa en otros proyectos de investigación.

A

Diagrama de Gantt

En la Figura A.1 se puede ver el diagrama de Gantt correspondiente a este proyecto.

A continuación se va a comentar un poco las distintas tareas y el por qué de su orden y duración.

Conforme se iban cumpliendo objetivos y encontrando obstáculos durante el desarrollo del proyecto se han realizado reuniones con los directores del mismo. Estas han sido frecuentes y muy valiosas, puesto que además de permitir conocer la situación en la que se encontraba el proyecto en todo momento, se trataba de buscar soluciones para los problemas encontrados y determinar las próximas tareas a realizar. También se asistió a una reunión con el personal investigador biólogo que colabora con el grupo de bioinformática al que pertenecen los dos directores de este proyecto.

Como ya se ha comentado varias veces a lo largo de la memoria, la fase de documentación ha sido larga y continuada prácticamente a lo largo de todo el proyecto, de ahí su extensión.

Puesto que el desarrollo de los distintos módulos se ha realizado de forma independiente, las fases de diseño, implementación y pruebas se han llevado a cabo de forma iterativa, siguiendo las pautas del modelo ágil de desarrollo de software en la medida de lo posible.

Aunque desde el principio se hizo un esfuerzo por documentar el trabajo que se iba desarrollando, los esfuerzos más intensos en este apartado se han llevado en la fase final del proyecto.

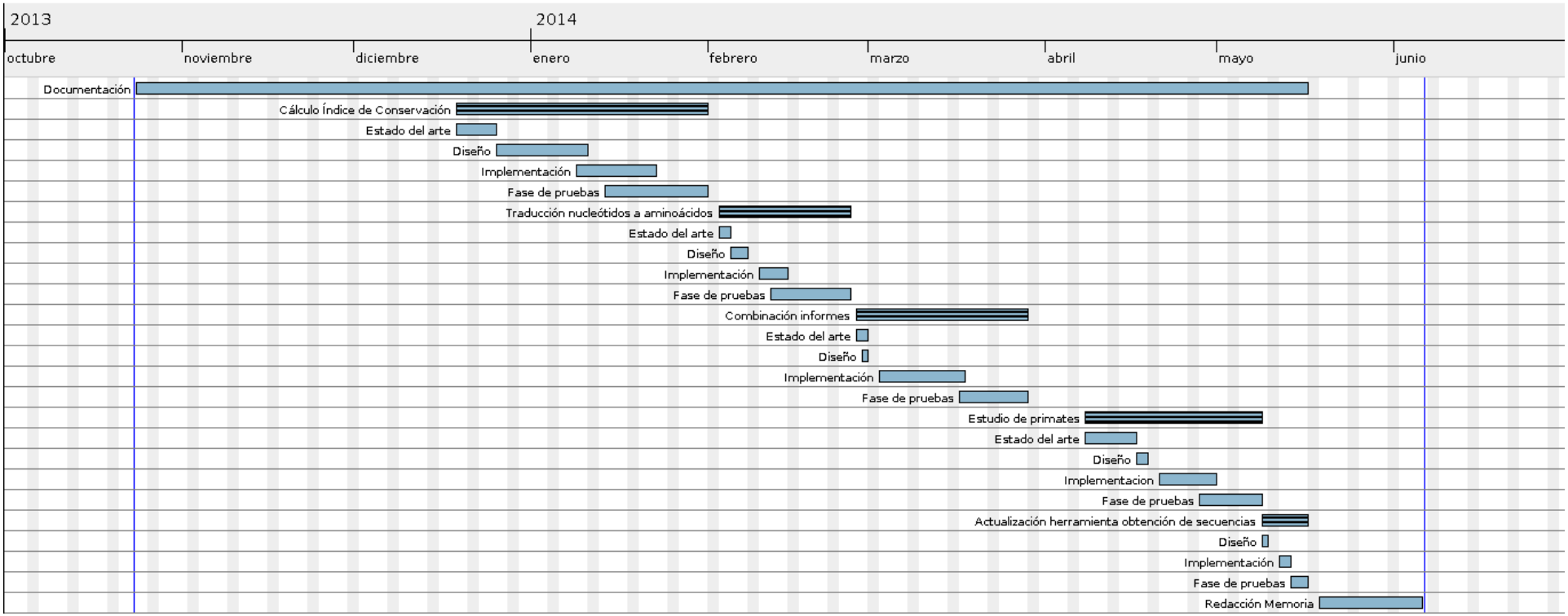

Figura A.1: Diagrama de Gantt.

B
Fundamentos biológicos

A lo largo de este apéndice se van a explicar todos aquellos términos y conceptos que se han considerado necesarios para la completa comprensión del trabajo realizado en este proyecto. Antes de continuar me gustaría agradecer a uno de mis directores, Jorge Álvarez, su consentimiento para usar la memoria de su proyecto de fin de carrera [1] como inspiración y base para la realización de este capítulo.

B.1 Base biológica

Se ha considerado oportuno hacer una breve pausa en la mente informática del lector para mostrar el proyecto desde el punto de vista de un biólogo, permitiendo así apreciarlo en su conjunto y comprender mejor el contexto del trabajo.

El ADN mitocondrial humano, elemento tratado a lo largo de este proyecto, (ADNmt de ahora en adelante) es un elemento biológico muy interesante desde muchos puntos de vista. Al estar dentro de las mitocondrias y ser independiente del ADN celular es centro de muchas teorías y controversias respecto al origen o inclusión de estos corpúsculos en la célula (y gracias a los cuales hoy estamos nosotros aquí). Pero ese no es el tema que atañe a este proyecto. El ADNmt posee una tasa de cambio muy elevada, lo que lo hace idóneo para el estudio de individuos y su estrecha relación, como miembros de una misma especie. Además el ADNmt está muy ligado a ciertas enfermedades que provocan una muerte prematura en los individuos que las padecen, lo cual lo convierte en idóneo como objeto de análisis en aquellos estudios que se centren en la patogenicidad de mutaciones [15].

El ADNmt ha sido largamente estudiado, y actualmente se sabe que lo forman entre 16557 y 16576 nucleótidos, conteniendo 37 genes distintos, de los cuales 13 tienen como objetivo la creación de proteínas, 22 codifican ARNt (ARN transferente o de transferencia) y, cómo no, las dos unidades que conforman el ribosoma (ARNr) [11]. Hay que indicar que el ADNmt es circular, a diferencia del ADN nuclear, del que todo el mundo conoce su estructura de doble hélice. Por tanto, el ADNmt posee una región de control, también conocida como bucle D, que no tiene como objetivo la codifica-

ción sino la unión para conformar dicha estructura. Esta zona contiene dos regiones hipervariantes: HVR_1 y HVR_2.

Las enfermedades mitocondriales son desórdenes resultantes de la deficiencia de una o más proteínas localizadas en las mitocondrias e involucradas en el metabolismo [27, 11]. Dichas enfermedades pueden estar causadas por mutaciones en el ADN mitocondrial o bien por mutaciones en genes nucleares que codifican proteínas implicadas en el correcto funcionamiento de la mitocondria [13].

B.1.1. Expresión Génica

La expresión génica es el proceso por medio del cual todos los organismos procariotas y células eucariotas transforman la información codificada por los ácidos nucleicos en las proteínas necesarias para su desarrollo y funcionamiento[21]. Consta de dos fases:

1. Transcripción: proceso mediante el cual se transfiere la información contenida en la secuencia del ADN hacia la secuencia de proteína utilizando diversos ARN como intermediarios.

2. Traducción: proceso en el cual el ARN mensajero se decodifica para producir un polipéptido específico de acuerdo con las reglas especificadas por el código genético (Tabla B.1.

B.1.2. Mutaciones Génicas

Las mutaciones génicas son alteraciones en la secuencia de nucleótidos de un gen y pueden producirse por dos causas:

1. Por sustitución de base: Suponen el 20 % de las mutaciones espontáneas. En ocasiones, el nuevo triplete codifica el mismo aminoácido o uno distinto pero que no hace que se altere la función de la proteína, con lo que la mutación no tiene consecuencias perjudiciales. En otras ocasiones, la mutación hace que cambie un aminoácido del centro activo de una enzima o afecta a un triplete de finalización. En estos casos dicha mutación puede resultar perjudicial. Se dividen en dos grupos:

   a) Transiciones

   b) Transversiones

2. Por pérdida o inserción de nucleótidos: Constituyen el 80 % de las mutaciones espontáneas. En este tipo de mutaciones génicas resulta afectado el proceso de síntesis de proteínas. La consecuencia de estas mutaciones es un corrimiento en el orden de lectura de los tripletes

Aminoácido	Codones	Con alfabeto extendido
A	GCU, GCC, GCA, GCG	GCN
R	CGU, CGC, CGA, CGG, AGA, AGG	CGN, MGR
N	AAU, AAC	AAY
D	GAU, GAC	GAY
C	UGU, UGC	UGY
Q	CAA, CAG	CAR
E	GAA, GAG	GAR
G	GGU, GGC, GGA, GGG	GGN
H	CAU, CAC	CAY
I	AUU, AUC, AUA	AUH
L	UUA, UUG, CUU, CUC, CUA, CUG	YUR, CUN
K	AAA, AAG	AAR
M	AUG	
F	UUU, UUC	UUY
P	CCU, CCC, CCA, CCG	CCN
S	UCU, UCC, UCA, UCG, AGU, AGC	UCN, AGY
T	ACU, ACC, ACA, ACG	ACN
W	UGG	
Y	UAU, UAC	UAY
V	GUU, GUC, GUA, GUG	GUN
- (Terminación)	UAA, UGA, UAG	UAR, URA

Tabla B.1: Reglas de traducción del código genético.

a partir del punto en el que ocurre la mutación y por lo tanto, alteran todos los tripletes siguientes, teniendo por lo general efectos muy graves. Se dividen en dos tipos:

a) Deleciones

b) Inserciones

B.2 Introducción a la bioinformática

La bioinformática puede considerarse una rama de reciente aparición en la informática y tiene como objetivo la aplicación de la informática a la gestión y análisis de datos biológicos. Su necesidad se ha visto acentuada en los últimos años debido a la magnitud y complejidad que están adquiriendo ciertas investigaciones referentes a la biología, que hacen intratable su resolución de forma manual [22, 19].

En concreto en este proyecto se han tratado temas del área de análisis de secuencias, más en concreto, sobre el cálculo del índice de conservación.

Para poder llevar a cabo dicho estudio estadístico es indispensable disponer de secuencias previamente alineadas. El alineamiento de secuencias no ha sido objetivo de este proyecto, por lo que se ha trabajado con las técnicas que se estudiaron y desarrollaron en anteriores proyectos de fin de carrera [4]. Para dar una idea al lector de en qué consisten estas técnicas, el alineamiento de secuencias es una operación que establece equivalencias entre los distintos caracteres de las secuencias introducidas. Haciendo la suposición de que entre dichas secuencias existe un parentesco, pretende detectar aquellos hechos que las separan.

B.2.1. Alineamientos de secuencias

Para poder realizar un análisis de la información contenida en los conjuntos de secuencias se requiere su tratamiento previo, lo que se conoce como alineamiento. Este proceso es necesario debido a que no todas las secuencias tienen exactamente la misma longitud ya que pueden producir inserciones y deleciones en distintas secciones de la secuencia, especialmente en aquellas que no codifican información relevante, como la región hipervariante. Dicho tratamiento consiste, como ya se ha comentado en el Capítulo 2, en el análisis y modificación de estos conjuntos de secuencias con el fin de conseguir la mayor cantidad posible de nucleótidos iguales en la misma posición. Para ello se realizan inserciones de huecos (habitualmente llamados *gaps*) en algunas de las secuencias. La precisión del alineamiento se mide contabilizando el número de *gaps* insertados en las secuencias que componen el conjunto así como la longitud, en cada secuencia, de estos fragmentos de *gaps*.

B.2.2. Índice de Conservación

El Índice de Conservación (IC) es un estadístico muestral que indica la frecuencia con la que los elementos de un alfabeto (nucleótidos, aminoácidos, etc.) aparecen en cada una de las posiciones de un conjunto de secuencias alineadas. Su cálculo en secuencias biológicas es un importante instrumento adicional en los estudios de la patogenicidad de las mutaciones. Su análisis permite centrar las investigaciones en aquellas secciones de las secuencias en las que es más probable que las mutaciones sean mortales. Como se ha comentado en la Sección 3.1, la utilización de dicho estadístico muestral ha permitido asociar distintos tipos de cáncer, como el de mama o endometrio [7] o el gástrico [3] a mutaciones en secciones concretas del ADNmt humano. También se han llevado a cabo otros estudios en los que se utiliza el IC junto con otros instrumentos para detectar una nueva mutación en un gen del ADNmt asociada a la pérdida auditiva hereditaria [20] o enfermedades cardiovasculares como la hipertensión [9] y la miocardiopatía no compactada [18, 28].

Selección del Lenguaje de Programación

Tras realizar un primer análisis del problema al que se buscaba dar solución se decidió escoger como paradigma de programación la programación por procedimientos, que se deriva de la programación estructurada. La elección de dicho paradigma permitió explotar la técnica de programación modular, cuyas ventajas más destacables son la facilidad de depurar, actualizar y modificar el código.

Entre los múltiples lenguajes de programación que se consideraron inicialmente se encontraban lenguajes de bajo (C) y alto nivel (C++, Java, Python).

Pese a que lenguajes como C permiten una gestión de los recursos computacionales más eficiente, puesto que en este proyecto no se trabajaba con sistemas que tuviesen grandes limitaciones en este aspecto, se decidió elegir un lenguaje de alto nivel y aprovechar sus múltiples ventajas. Además, debido a la creciente complejidad de las arquitecturas de los microprocesadores modernos, los compiladores para lenguajes de alto nivel cada vez generan código más eficiente.

Los lenguajes candidatos fueron C++, Java y Python por múltiples razones (experiencia previa, documentación, gestión de excepciones, robustez, etc.). Pese a que la experiencia previa utilizando C++ y Java era muy superior, varios factores fueron determinantes a la hora de seleccionar a Python. Dicha elección como lenguaje principal de este proyecto se basó en: legibilidad, librería estándar muy extensa y potente; existencia de la librería BioPython, que incluye herramientas para la bioinformática; y la utilización de dicho lenguaje en las herramientas desarrolladas por Jorge Álvarez, que permitieron simplificar la complejidad del problema a resolver. Por último, merece la pena comentar que pese a que Python se un lenguaje interpretado y esto penaliza el tiempo de ejecución de los algoritmos, debido a la alta interacción con herramientas externas como Mafft, dicha penalización se puede considerar despreciable.

D

Manual de usuario: Conjunto de herramientas desarrolladas para el análisis del IC

D.1 Manual de usuario: *Herramienta de cálculo del IC*

El programa de cálculo del IC dispone de los siguientes parámetros de entrada:

- *fich_secs* [obligatorio]: ruta de acceso al fichero que contiene el conjunto de secuencias de ADNmt o proteínas. El fichero tiene que estar en formato FASTA.

- *dir_informe* [opcional]: ruta del directorio donde se guardará el fichero con el informe generado. Por defecto, el directorio actual. De no existir el directorio se creará.

- *umbral* [obligatorio]: indica el límite inferior o superior del IC. En el informe se indican aquellas posiciones del conjunto de secuencias que tienen un IC superior o inferior al indicado por este umbral.

- *tipo_umbral* [opcional]: limite inferior o superior (por defecto el límite es superior). Permite realizar búsquedas de las posiciones con un IC tanto por encima como por debajo del umbral establecido.

- *rango* [opcional]: rango de la sección a analizar (por defecto la sección comprende la longitud total del de las secuencias, si se introducen alineadas, o de la longitud del alineamiento una vez efectuado). Permite realizar secciones no asociadas a los distintos genes del ADNmt.

- *gen* [opcional]: gen al que pertenece el conjunto de secuencias a analizar. Necesario para que en el informe generado se indique correctamente la posición absoluta de cada nucleótido. Por defecto el primer nucleótido de la secuencia comienza en 1, y en caso de indicarse un gen,

la primera posición corresponde a la posición inicial del gen en la secuencia de ADNmt. En la versión actual el parámetro solo es aplicable en ADNmt humano.

- *informe_detallado* [opcional]: bandera (más conocida por el término anglosajón *flag*) que indica el deseo de obtener como salida un informe detallado del análisis del IC realizado (por defecto no se incluye indicando que se desea obtener un informe básico). En la Sección 3.5 se explican las principales diferencias entre estos dos tipos de informes.

- *verbose* [opcional]: bandera (o *flag*) que indica el deseo de obtener información por pantalla relativa al estado en el que se encuentra la ejecución del módulo (por defecto no se incluye ningún tipo de información del estado del proceso).

En el caso de que las secuencias no estén alineadas, se solicitará por pantalla al usuario que indique la herramienta de alineamiento que se desea utilizar y su configuración.

Dicho programa genera la siguiente salida:

- Fichero de texto que contiene el informe con los resultados del estudio del IC para cada posición del conjunto de secuencias indicado en el parámetro de entrada.

D.2 Manual de usuario: *Herramienta de traducción de nucleótidos a aminoácidos*

El programa de traducción de nucleótidos a aminoácidos dispone de los siguientes parámetros de entrada:

- *fich_secs* [obligatorio]: ruta de acceso al fichero que contiene el conjunto de secuencias de ADNmt. El fichero tiene que estar en formato FASTA.

En el caso de que las secuencias no estén alineadas, se solicitará por pantalla al usuario que indique la herramienta de alineamiento que se desea utilizar y su configuración.

Dicho programa genera la siguiente salida:

- Alineamiento del conjunto de proteínas resultante de aplicar el método desarrollado en este módulo que simula el proceso de traducción a proteínas que ocurre en las células.

D.3 Manual de usuario: *Herramienta de combinación de informes*

El programa de combinación de informes dispone de los siguientes parámetros de entrada:

- *ruta_nucleotidos* [obligatorio]: ruta de acceso al fichero que contiene el informe del IC de las secuencias de nucleótidos.
- *ruta_aminoacidos* [obligatorio]: ruta de acceso al fichero que contiene el informe del IC de las secuencias de aminoácidos.

Ambos informes deben pertenecer al mismo conjunto de secuencias de ADNmt.

Dicho programa genera la siguiente salida:

- Fichero de texto que contiene el informe combinado en el que se muestra la información recopilada de los informes indicados en los parámetros de entrada.

D.4 Consejos de uso

Puesto que el número de ficheros que componen el sistema puede crecer de forma exponencial, tanto al añadir nuevos métodos de alineamiento como al ampliar el número de conjuntos de secuencias que se manejan (por ejemplo, porque se decide ampliar el estudio a otras especies), se propone establecer una estructura lo más sencilla posible en el sistema de ficheros. De este modo, se recomienda al usuario mantener en directorios separados los scripts que componen el núcleo del sistema, los ficheros que almacenan los distintos conjuntos de secuencias, y los informes que dichos scripts generan. Se muestra una propuesta de la estructura de ficheros en la Figura D.1.

Otro de los detalles que se han tenido en cuenta es el de generar ficheros cuyos nombres sean lo más autoexplicativos posible, no solo en el caso de los scripts desarrollados, sino también en aquellos que almacenan los alineamientos y los informes generados. Para facilitar la comprensión del lector a este respecto, se incluyen a continuación tres ejemplos:

- `hmtDNA.fasta`: Conjunto de secuencias de ADNmt humano.
- `hmtDNA_rCRS_aligned_mafft_auto_ATP6_peptide_[all]_0.5.txt`: Informe básico del cálculo del IC con un umbral superior a 0.5, correspondiente al conjunto de secuencias de aminoácidos asociadas al gen ATP6 de ADNmt humano alineadas con Mafft - *auto* y la secuencia de referencia rCRS.

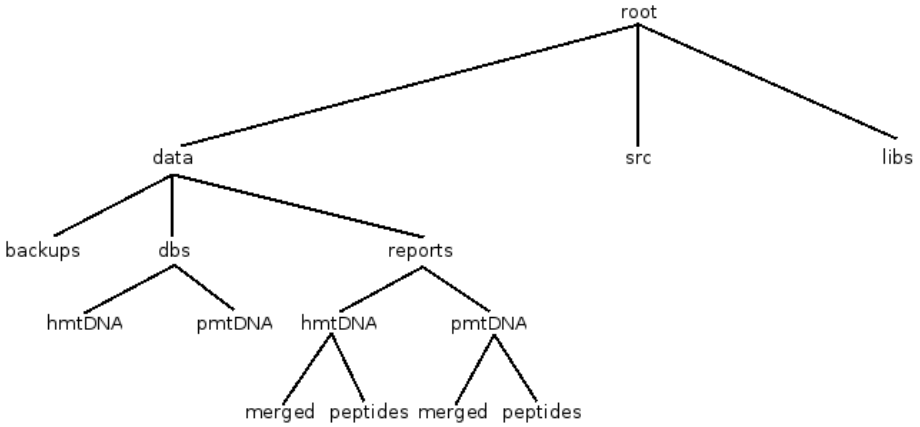

Figura D.1: Propuesta de la estructura del sistema de ficheros.

- `pmtDNA_aligned_mafft_auto_[all]_0.99_details.txt`: Informe detallado del cálculo del IC con un umbral superior a 0.99, correspondiente al conjunto de secuencias de ADNmt primate alineadas con Mafft - *auto*.

E

Gráficas de resultados

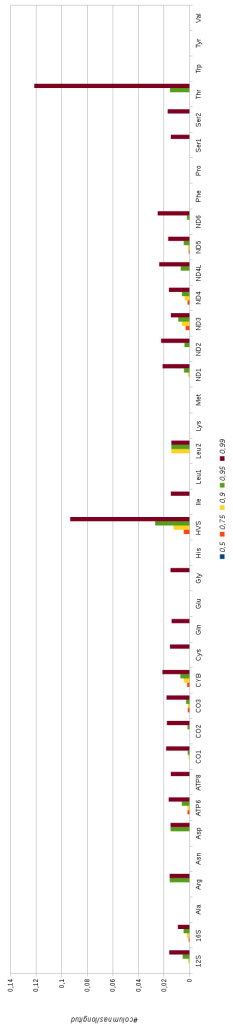

Figura E.1: Figura 3.15 ampliada.

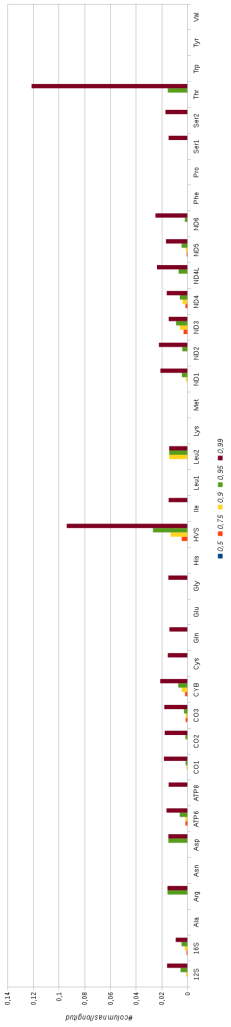

Figura E.2: Figura 3.16 ampliada.

Bibliografía

[1] J. Alvarez-Jarreta. Análisis teórico-práctico de métodos de inferencia filogenética basados en selección de modelos y métodos de superárboles. Master's thesis, Centro Politécnico Superior, Universidad de Zaragoza, 2010.

[2] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler. Genbank. *Nucleic Acids Research*, 36(suppl 1):D25–D30, 2008.

[3] Rui Bi, Wen-Liang Li, Ming-Qing Chen, Zhu Zhu, and Yong-Gang Yao. Rapid identification of mtdna somatic mutations in gastric cancer tissues based on the mtdna phylogeny. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 709–710(0):15 – 20, 2011.

[4] Roberto Blanco. Definición y prototipo de herramienta de análisis filogenético para el adn mitocondrial humano. Master's thesis, Centro Politécnico Superior, Universidad de Zaragoza, 2008.

[5] Roberto Blanco and Elvira Mayordomo. Zaramit: A system for the evolutionary study of human mitochondrial dna. In Sigeru Omatu, MiguelP. Rocha, José Bravo, Florentino Fernández, Emilio Corchado, Andrés Bustillo, and JuanM. Corchado, editors, *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*, volume 5518 of *Lecture Notes in Computer Science*, pages 1139–1142. Springer Berlin Heidelberg, 2009.

[6] Jeffrey L. Boore. Animal mitochondrial genomes. *Nucleic Acids Research*, 27(8):1767–1780, 1999.

[7] M Brandon, P Baldi, and D C Wallace. Mitochondrial mutations in cancer. *Oncogene*, 25(34):4647–4662, 2006.

[8] Marty C. Brandon, Eduardo Ruiz-Pesini, Dan Mishmar, Vincent Procaccio, Marie T. Lott, Kevin Cuong Nguyen, Syawal Spolim, Upen Patil, Pierre Baldi, and Douglas C. Wallace. Mitomaster: a bioinformatics tool for the analysis of mitochondrial dna sequences. *Human Mutation*, 30(1):1–6, 2009.

[9] Hong Chen, Jing Zheng, Ling Xue, Yanzi Meng, Yan Wang, Bingjiao Zheng, Fang Fang, Suxue Shi, Quiaomeng Qiu, Pingping Jiang, Zhongqiu Lu, Jun Qin Mo, Jianxin Lu, and Min-Xin Guan. The 12s rrna a1555g mutation in the mitochondrial haplogroup d5a is responsible for maternally inherited hypertension and hearing loss in two chinese pedigrees. *European Journal of Human Genetics*, 20(6):607–612, 2012.

[10] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[11] Salvatore DiMauro and Guido Davidzon. Mitochondrial dna and disease. *Annals of Medicine*, 37(3):222–232, 2005.

[12] Long Fan and Yong-Gang Yao. Mitotool: A web server for the analysis and retrieval of human mitochondrial {DNA} sequence variations. *Mitochondrion*, 11(2):351 – 356, 2011.

[13] Laura C Greaves, Amy K Reeve, Robert W Taylor, and Doug M Turnbull. Mitochondrial dna and disease. *The Journal of Pathology*, 226(2):274–286, 2012.

[14] C. Herrnstadt, J. L. Elson, E. Fahy, G. Preston, D. M. Turnbull, C. Anderson, S. S. Ghosh, J. M. Olefsky, M. F. Beal, R. E. Davis, and N. Howell. Reduced-median-network analysis of complete mitochondrial dna coding-region sequences for the major african, asian, and european haplogroups. *The American Journal of Human Genetics*, 70(5):1152–1171, 2002.

[15] Donald R. Johns. Mitochondrial dna and disease. *New England Journal of Medicine*, 333(10):638–644, 1995.

[16] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 2002.

[17] Kuo-Bin Li. Clustalw-mpi: Clustalw analysis using distributed and parallel computing. *Bioinformatics*, 19(12):1585–1586, 2003.

[18] Shenghua Liu, Yuanyuan Bai, Jie Huang, Hong Zhao, Xiaoling Zhang, Shengshou Hu, and Yingjie Wei. Do mitochondria contribute to left ventricular non-compaction cardiomyopathy? new findings from myocardium of patients with left ventricular non-compaction cardiomyopathy. *Molecular Genetics and Metabolism*, 109(1):100 – 106, 2013.

[19] N. M. Luscombe, D. Greenbaum, and M. Gerstein. What is bioinformatics? a proposed definition and overview of the field. *Methods of Information in Medicine*, 40(4):346 – 358, 2001.

[20] Emna Mkaouar-Rebai, Nourhene Fendri-Kriaa, Nacim Louhichi, Abdelaziz Tlili, Chahnez Triki, Abdelmoneem Ghorbel, Saber Masmoudi,

and Faiza Fakhfakh. Whole mitochondrial genome screening in two families with hearing loss: detection of a novel mutation in the 12s rrna gene. *Bioscience Reports*, 30(6):405–411, 2010.

[21] Tokumasa Nakamoto. Evolution and the universality of the mechanism of initiation of protein synthesis. *Gene*, 432(1–2):1 – 6, 2009.

[22] Andrzej Polanski and Marek Kimmel. *Bioinformatics*. Springer, 2007.

[23] Martin B. Richards, Vincent A. Macaulay, Hans-Jürgen Bandelt, and Bryan C. Sykes. Phylogeography of mitochondrial dna in western europe. *Annals of Human Genetics*, 62(3):241–260, 1998.

[24] Eduardo Ruiz-Pesini, Dan Mishmar, Martin Brandon, Vincent Procaccio, and Douglas C. Wallace. Effects of purifying and adaptive selection on regional variation in human mtdna. *Science*, 303(5655):223–226, 2004.

[25] Cristina Santos, Rafael Montiel, Adriana Arruda, Luis Alvarez, Maria Aluja, and Manuela Lima. Mutation patterns of mtdna: Empirical inferences for the coding region. *BMC Evolutionary Biology*, 8(1):167, 2008.

[26] Emmanuelle Sarzi, Michael D. Brown, Sophie Lebon, Dominique Chretien, Arnold Munnich, Agnès Rotig, and Vincent Procaccio. A novel recurrent mitochondrial dna mutation in nd3 gene is associated with isolated complex i deficiency causing leigh syndrome and dystonia. *American Journal of Medical Genetics Part A*, 143A(1):33–41, 2007.

[27] Anu Suomalainen. Mitochondrial dna and disease. *Annals of Medicine*, 29(3):235–246, 1997.

[28] Sha Tang, Anjan Batra, Yu Zhang, Eric S. Ebenroth, and Taosheng Huang. Left ventricular noncompaction is associated with mutations in the mitochondrial genome. *Mitochondrion*, 10(4):350 – 357, 2010.

[29] Antonio Torroni, Alessandro Achilli, Vincent Macaulay, Martin Richards, and Hans-Jürgen Bandelt. Harvesting the fruit of the human mtdna tree. *Trends in Genetics*, 22(6):339 – 345, 2006.

[30] Michael V. Zaragoza, Martin C. Brandon, Marta Diegoli, Eloisa Arbustini, and Douglas C. Wallace. Mitochondrial cardiomyopathies: how to identify candidate pathogenic mutations by mitochondrial dna sequencing, mitomaster and phylogeny. *European Journal of Human Genetics*, 19(2):200 – 207, 2011.