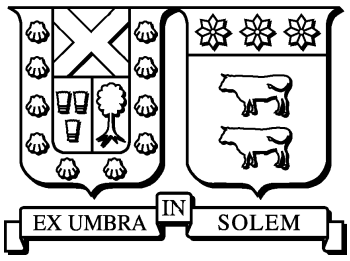
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INFORMÁTICA

SANTIAGO – CHILE



"ALGORITMO PARA EL CÁLCULO DE FRAGMENTOS DE PROTEÍNAS EN LOS ORGANISMOS SECUENCIADOS"

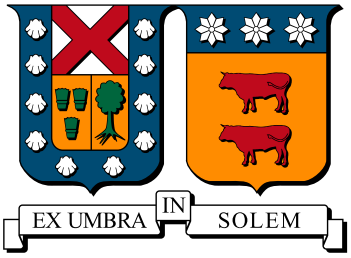
FELIPE NICOLÁS ARAYA BARRERA

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: LIOUBOV DOMBROVSKAIA

OCTUBRE 2017

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE



"ALGORITMO PARA EL CÁLCULO DE FRAGMENTOS DE PROTEÍNAS EN LOS ORGANISMOS SECUENCIADOS"

FELIPE NICOLÁS ARAYA BARRERA

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: LIOUBOV DOMBROVSKAIA
PROFESOR CORREFERENTE: DIEGO ARROYUELO BILLIARDI

OCTUBRE 2017

Agradecimientos

Se escribirán los agradecimientos una vez este documento haya sido terminado.

Dedicatoria

Se escribirá la dedicatoria una vez este documento haya sido terminado.

Resumen

Se completará este capítulo una vez que el cuerpo de este trabajo haya finalizado.

Abstract

Same thing as the previous section.

Índice de Contenidos

Índice de cuadros

Índice de figuras

Glosario

Introducción

Para el actual documento, las razones que motivaron al alumno presente a realizar la investigación de su tesis son la de comenzar a entrar en una rama que en los últimos años ha tomado bastante importancia en la informática, conocido como la *Bioinformática* [4], la cual aplica las tecnologías computacionales contemporáneas a datos biológicos que pueden pertenecer a estructuras como ADN, proteínas, entre otras estructuras biológicas complejas y los cuales están a la mano del ser humano como archivos de cadenas de secuencias (en varios formatos) y que pueden ser usados a voluntad.

En el caso puntual de esta memoria, se trabajarán con proteínas (compuestas de combinaciones de 20 aminoácidos) que están distribuidas en un *dataset* cuyo formato del archivo está en **.fasta**, donde cada cadena de polipéptido está compuesto por un ID o código identificador, su nombre taxonómico y su posterior secuencia de aminoácidos.

Hablando de las proteínas, se han realizado numerosos análisis [5, 6] en bases de datos de secuencias con respecto a la cantidad de aminoácidos que se encuentran en total en este tipo de estructuras, distribuyéndolos según su tamaño o para determinar cuál es el aminoácido que más aparece en este tipo de archivos; también según el año de descubrimiento de las proteínas o el tipo de proteína. Todo esto usando fuentes como UniProt y EROP-Moskow.

Como una derivación directa, existe una variabilidad casi infinita de combinaciones llamadas residuos (fragmentos) de aminoácidos (*amino acid residues* o AAR de manera simplificada en inglés) que se determinan según su tamaño k y por las posibles opciones a obtener, que sigue la regla de combinatoria 20^k ya que son 20 los animoácidos base que existen en la

actualidad (para dipéptidos serían $20^2 = 400$, para tripéptidos serían $20^3 = 8000$ y así sucesivamente).

Lo que se pretende realizar para esta memoria consiste en obtener de un conjunto predeterminado de proteínas el número máximo de fragmentos de péptidos (AAR) que existen asociado a un valor k determinado, donde k se ubicará en el intervalo entre 2 hasta 50, y en base a aquello obtener y determinar para cada k cuáles son los fragmentos de aminóácidos que más se repiten para posteriormente realizar un análisis tanto matemático y biológico de los resultados obtenidos. Para realizar esta tarea se usará como *dataset* la base de datos de proteínas de SwissProt (550100 proteínas) y de TrEMBL (88032926 proteínas).

A partir de esto las implementaciones de código para trabajar estos archivos fueron realizados en los lenguajes de programación **C++** y **Python**, con diferentes finalidades que serán explicados a medida que se avance en el documento.

Capítulo 1

Definición del Problema

1.1. Información previa a considerar

Para entrar de lleno en la tema, es necesario conocer de antemano varios aspectos básicos de la biología.

1.1.1. Biomoléculas

Cada vez que se habla de la biología, este concepto se relaciona directamente con la ciencia que estudia a los seres vivos. Ahora bien, las estructuras o compuestos que constituyen una parte esencial de los seres vivos son conocidas como **biomoléculas**. Estas biomoléculas están principalmente constituidas por elementos químicos como el carbono (C), hidrógeno (H), oxígeno (O), nitrógeno (N), fósforo (P) y azufre (S) [7] y se pueden clasificar en biomoléculas inorgánicas, que se encuentran tanto en seres vivos como en los cuerpos inertes, no obstante son imprescindibles para la vida; y las biomoléculas orgánicas, que son sintetizadas por los seres vivos y tienen una estructura con base en carbono. Estas biomoléculas orgánicas se pueden separar en 4 grandes grupos:

1. Glúcidos (hidratos de carbono o carbohidratos): son la fuente de energía primaria que

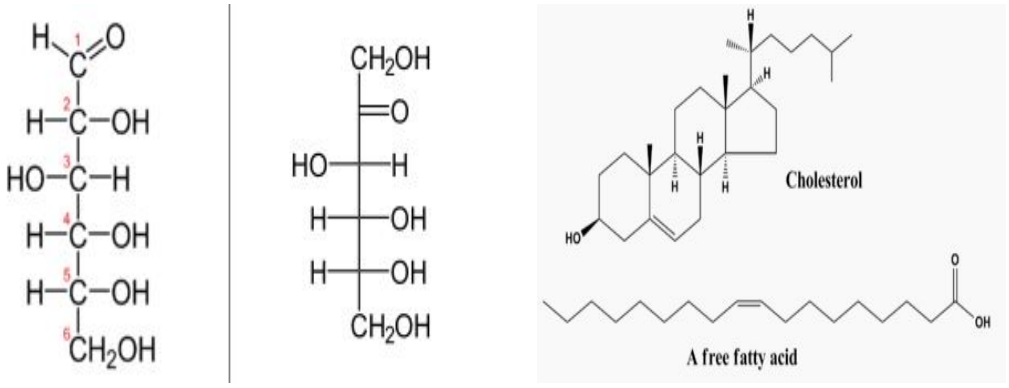
utilizan los seres vivos para realizar sus funciones vitales. Los ejemplos más conocidos son la glucosa, el almidón y el glucógeno.

2. Lípidos: conforman el principal almacén de energía de los animales y desempeñan funciones reguladores de enzimas y hormonas.

3. Ácidos nucleicos: El ácido desoxirribonucleico y el ácido ribonucleico, mayormente conocidos como ADN (DNA) y ARN (RNA y sus derivados) desarrollan posiblemente la función más importante para la vida: contener, de manera codificada, las instrucciones necesarias para el desarrollo y funcionamiento de la célula. El ADN tiene la capacidad de replicarse, transmitiendo así dichas instrucciones a las células hijas que heredarán la información.

4. Proteínas: poseen la mayor diversidad de funciones que realizan en los seres vivos; prácticamente todos los procesos biológicos dependen de su presencia y/o actividad. Son proteínas casi todas las enzimas, catalizadores de reacciones metabólicas, hemoglobina, anticuerpos, entre otros. Su unidad base es el *aminoácido*, por el cual se van formando los péptidos según la cantidad de unidades bases enlazadas.

Dentro de estas biomoléculas, el análisis detallado de los carbohidratos y los lípidos depende en demasía de su estructura química (elementos químicos asociados y tipo de enlaces entre ellos), por lo mismo es una materia más ligada a los químicos (ver Figura 1.1).

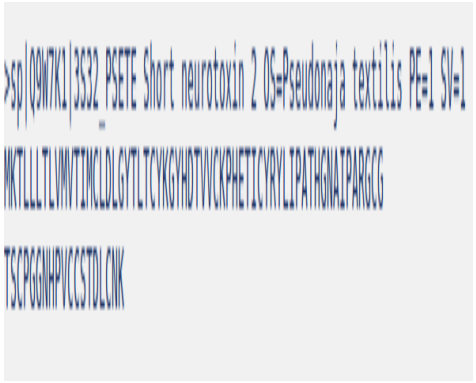


(a) Glúcidos (glucosa y fructosa) (b) Lípidos (colesterol y un ácido graso)

Figura 1.1: Estructura química de los carbohidratos y lípidos.

Sin embargo, el ADN y los polipéptidos poseen unidades base que pueden ser codificadas como letras, por consiguiente pueden ser secuenciados como *cadenas de strings* y en donde los avances computacionales y la evolución informática toman una importante relevancia (ver Figura 1.2).

(a) Cadena de ADN aleatoria



(b) Cadenas α y β de hemoglobina bovina

Figura 1.2: Biomoléculas de ADN y péptidos llevadas a cadenas de strings.

Con respecto a estas 2 últimas estructuras, la diferencia visual más notoria radica en la cantidad de diferentes letras (strings) que las componen, para el ADN son 4 [8] y son denominadas **bases nitrogenadas** que son las siguientes:

1. Adenina
2. Timina
3. Citosina
4. Guanina

Para las proteínas, su elemento básico, como ya se mencionó anteriormente es el **aminoácido**, pero ahora se adentrará en más detalle sobre esta molécula.

1.1.2. Aminoácidos

Los aminoácidos tienen diferentes funciones en el organismo [8] pero ante todo sirven como **las unidades básicas de los péptidos y de las proteínas.** A nivel orgánico el aminoácido es una molécula compuesta con un grupo amino (-NH2) y un grupo carboxilo (-COOH) y que pueden tener distintas distribuciones. Para el caso de los que componen las proteínas se consideran como alfa-aminoácidos:

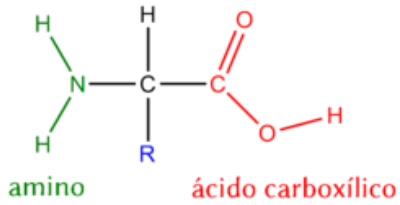


Figura 1.3: Estructura general de un alfa-aminoácido

En la imagen anterior se puede identificar el carbono central (alfa) unido al grupo carboxilo (rojo), grupo amino (verde), un hidrógeno (imagen superior color negro) y el grupo radical (azul) o R. Este grupo radical es el que determina la identidad y las propiedades de cada uno de los diferentes aminoácidos.

El primer aminoácido fue descubierto a principios del siglo XIX, y a partir de ese entonces hasta la actualidad son miles los aminoácidos que han sido descubiertos, pero solo 20 se consideran como los componentes esenciales para las proteínas (y los que se considerarán como parte de esta memoria) que se presentarán a continuación en conjunto con su respectiva abreviación utilizada en las cadenas de proteínas de los archivos FASTA:

1. Alanina - A
2. Cisteína - C
3. Ácido aspártico - D
4. Ácido glutámico - E
5. Fenilalalina - F

6. Glicina - G
7. Histidina - H
8. Isoleucina - I
9. Lisina - K
10. Leucina - L
11. Metionina - M
12. Aspargarina - N
13. Prolina - P
14. Glutamina - Q
15. Arginina - R
16. Serina - S
17. Treonina - T
18. Valina - V
19. Triptófano - W
20. Tirosina - Y

Existen otras abreviaturas en las cadenas como B, X o J, pero para el alcance de esta memoria no serán considerados como objeto de estudio y análisis posterior.

A partir de este pequeño elemento se forman las macromoléculas que se identifican según la cantidad de aminoácidos (a partir de ahora se mencionarán como aa.) que lo compongan:

Tamaño	Tipo de estructura
2 aa.	Dipéptido
3 aa.	Tripéptido
Entre 2 y 8 aa.	Oligopéptido
Menos de 100 aa.	Péptido
Mayor o igual de 100 aa.	Proteína o polipéptido

Cuadro 1.1: Identificación de macromoléculas según cantidad de aminoácidos

1.2. Secuencias de proteínas

Desde el momento en que se descubrieron los elementos componentes del ADN y las proteínas, se han investigado sobre las posibles combinaciones que se pueden encontrar entre las bases que los conforman y en que cantidad se encuentran. Para el ADN y sus 4 elementos básicos existen millones de seres vivos, parásitos, virus, protozoos y entre otros que se definen por su código genético, por lo cual encontrar los diversos residuos de bases según un determinado tamaño. En el caso puntual para la finalidad de este escrito y según lo mencionado por [9], es posible estudiar de manera teórica y con fórmulas matemáticas la cantidad máxima de fragmentos que puede formar una proteína. Considerando como base que el número posible de estructuras peptídicas naturales P están compuestas de diferentes residuos de aminoácidos (incluyendo repeticiones en cadenas de aminoácidos) sigue la siguiente fórmula:

$$P = A^n \tag{1.1}$$

Donde A es el número de diferentes aminoácidos existentes, y n es la cantidad de aminácidos correspondientes a la estructura estudiada. Por lo mismo y siguiendo esta fórmula (considerando $A = 20$) la cantidad de diferentes combinaciones péptidos de tamaño k que se pueden obtener se aprecian en la siguiente tabla:

Tamaño péptido (k)	Combinaciones posibles (A^k
2 aa.	400
3 aa.	8000
4 aa.	160000
5 aa.	3200000
10 aa.	1.024×10^{13}
20 aa.	1.049×10^{26}
50 aa.	1.126×10^{65}

Cuadro 1.2: Combinaciones posibles a obtener según el tamaño del péptido

Según lo observado en esta tabla se identificar que a medida que el valor de k va en aumento, las posibles combinaciones que se pueden obtener de fragmentos de proteínas pueden llegar a tener valores inimaginables para el ser humano corriente; no obstante, no todas estas estructuras existen o son capaces de ser encontradas en la naturaleza [6], aun así la diversidad de la búsqueda de estos residuos sigue siendo gigantesca, y por ende difícil de solucionar, y este será uno de los problemas que se intentará solucionar en esta memoria.

1.3. Definición

En [5] se menciona que buscar secuencias de proteínas es una tarea muy compleja, ya que se formaría un escenario similar al buscar una *aguja en un pajar* y recorrer millones de secuencias en cada búsqueda no sería lo más convieniente, por consiguiente una herramienta recomendable sería preprocesar la base de datos de proteínas con alguna herramienta o técnica conocida.

Capítulo 2

Estado del Arte

Capítulo 3

Propuesta

Capítulo 4

Implementación

Conclusiones

Bibliografía

[1] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1993.

[2] Albert Einstein. *Zur Elektrodynamik bewegter Körper*. (German) [*On the electrodynamics of moving bodies*]. Annalen der Physik, 322(10):891–921, 1905.

[3] Knuth: Computers and Typesetting,
`http://www-cs-faculty.stanford.edu/~uno/abcde.html`

[4] Rafael Lahoz-Beltrá. *BIOINFORMÁTICA, simulación, vida artificial e inteligencia artificial*. Ediciones Díaz de Santos S.A., Madrid, España, 2004.

[5] Bruce C. Orcutt, Winona C. Barker. *Searching the protein sequence database*. National Biomedical Research Foundation, Georgetown University Medical Center, Washington, D.C., *Bulletin of Mathematical Biology*, 46(4):545-552, 1984.

[6] Alexander A. Zamyatnin. *The Features of an Array of Natural Oligopeptides*. Bakh Institute of Biochemistry, Russian Academy of Sciences, Moskow, Russia; Departamento de Informática, Universidad Técnica Federico Santa María, Valparaíso, Chile, *Neurochemical Journal*, 10(4):249-257, 2016.

[7] Wikipedia, Definición de biomolécula,
`https://es.wikipedia.org/wiki/Biomolécula`.

[8] Jan Koolman, Klaus-Heinrich Röhm *Bioquímica: Texto y Atlas*. Editorial Médica Panamericana, Madrid, España; Georg Thieme Verlag, Stuttgart, Germany, 2004.

[9] A. A. Zamyatnin, *Fragmentomics of Natural Peptide Structures*. Bach Institute of Biochemistry, Russian Academy of Sciences; Universidad Técnica Federico Santa María, Departamento de Informática, Valparaíso, 2009, pp. 405-428.

[10] A. A. Zamyatnin, *Fragmentomics of Oligopeptides and Proteins*. Bach Institute of Biochemistry, Russian Academy of Sciences; Universidad Técnica Federico Santa María, Departamento de Informática, Valparaíso, 2nd Asia-Pacific International Peptide Symposium, 2007.