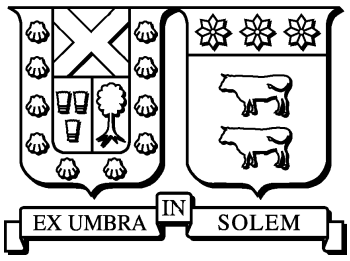
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INFORMÁTICA

SANTIAGO – CHILE



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

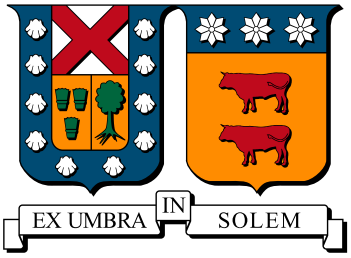
FELIPE NICOLÁS ARAYA BARRERA

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 [2][3] 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

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.