**Protein Sequence Profiler (PsP): An Automated Way of   
Characterizing Protein Sequence Entries in a Database**

**A Special Problem   
Presented to the Faculty of the   
Computer Science Program of the   
Department of Computer Science  
University of the Philippines Cebu**

**In Partial Fulfillment of the   
Requirements for the   
Bachelor of Science in Computer Science**

**Submitted by:**

**Kezah P. Alferez  
Jessica D. Pacilan**

**Bachelor of Science in Computer Science**

**Prof. Demelo M. Lao  
Adviser**

**May 2016**

Approval Page

This undergraduate special problem entitled **“Protein Sequence Profiler (PsP): An Automated Way of Characterizing Protein Sequence Entries in a Database”** prepared and submitted by **KEZAH P. ALFEREZ** and **JESSICA D. PACILAN** in fulfillment of the requirements for the degree **BACHELOR OF SCIENCE IN COMPUTER SCIENCE** has been examined and hereby recommended for approval and acceptance.

**PROF. DEMELO M. LAO**

Adviser  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Accepted and approved in partial fulfillment of the requirements for the degree **Bachelor of Science in Computer Science**.

**DR. ROBERT R. ROXAS**

Department Chairperson

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date

Abstract

The study presents a Protein Sequence Profiler (PsP): An Automated Way of Characterizing Protein Sequence Entries in a Databasewhich aims to provide a coherent system for all protein data taken from different sources. It serves as a unifying facility or a central hub for students and researchers for their source of functional information on proteins. The enormous amount data for proteins mean that having a unified source of protein data would greatly benefit students and researchers or generally, the curious people. This will help improve better understanding of the data for the users. This will also mean better or improved data accuracy for protein as the users have the freedom to import data from different sources and the system has the capacity to update existing data from time to time. Also provides a better usage of the protein information as they can download it from the system.

**Keywords:**

Introduction

# Background

In the drama of life on a molecular scale, proteins are where the action is [1]. Proteins are important to life as they are constituted by amino acids which are the building blocks of life. Although the information necessary for life to go on is encoded by the DNA molecule, the dynamic process of life maintenance, replication, defense and reproduction are carried out by proteins [2]. So, indeed proteins are essential to life.

With that in mind and because of the vast data for proteins, we wanted to build a Protein Sequence Profiler (PsP): An Automated Way of Characterizing Protein Sequence Entries in a Database to better understand what these proteins are and their constitution. There are already a lot of profilers available in the web but these mostly concentrate on signal peptides.

The protein database profiler is a system that contains sequences for both signal and non-signal peptides for different organisms such as archaea, bacteria, eukaryotes and viruses. The profiler basically filters these protein sequences according to what conditions the user wants to show or conditions that are applicable. It also filters the globular and transmembrane proteins, single spanning or multi-spanning for transmembrane and experimental or nonexperimental proteins. These data are stored in a MySQL database and are presented in an aesthetically and non-intimidating way.

The system also has an import and export feature wherein the user can upload and download entries to and from the system via text format. The system also provides table and graphs very useful for comparison and for visual detailing of protein entries that can also be downloaded. Most of all, one unique feature of the system is the internal redundancy checker that reduces the output of its redundant entries.

This study aims to address the issue of unifying the enormous protein that we have from different sources in a way that these data can be easily used for different purposes. And for improving the quality and accuracy of information that each protein holds. The system will not only cater to students, teachers, and researchers but to the general public for feeding their curious minds.

# Statement of the Problem

The question of what a protein does inside a living cell is not a simple one to answer [3]. Ardala Breda, Valadares, de Souza and Garratt believes that the reason for studying proteins and why we should understand how these folds, how they assemble into complexes, how they function is if we wish to answer questions as why we have cancer, why we grow old, why we get sick, how can we find cures for many diseases, why life as we know it has evolved in this way and on this planet and not anywhere else, at least for the moment [4]. The study intends to aid the growing need of researchers, teachers and students to search in a facility where proteins can be found but from different sources. To give them power to contribute to protein data by freely giving them access to import an entry. Also to give them the chance to create a smaller database using the data retrieved from the system containing the user-defined entries for their proteins. To give the user a better chance at understanding these proteins, that has been said over and over again, the one responsible for life of every living creature. At the very least, the system provides a fast but simple search feature for refinement of the data, a redundancy checker for eliminating redundant entries, an import feature for adding entries as well as an export feature for output and generation of reports.

# Objectives of the Study

There are already a lot of existing protein databases in the internet. There are some of general character and some of specific aspects. “*Probably the first question, when working with a protein structure, would be where to find the structure of interest. And another question, which many people need to ask, once they get access to a protein structure file is: What is actually inside that file? What information can be found there apart from the structure as such [5]?”* The output of this study is a protein database profiling system that aims to provide a complete but concise information for protein data that will give the user a better understanding of the complexities of proteins. Furthermore, this study intends to answer the following questions:

1. Did it simplify the complex data that each protein constitutes?
2. Will the system be publicly available and up-to-date?

# Scope and Limitation of the Study

The automatic update on the information of the protein data is at the moment, confined to the changes provided by UniProt (http://www.uniprot.org/).

Review of Related Literature

The study of proteins and their function is central to understanding both cells and organisms [6]. In the early 80’s, because of the technology advances, the paradigm shifted from studying single proteins to whole set of proteins of an organism according to Burley, et al [7]. And by the 90’s there is already an explosive growth with the amount of data. These data are stored in databases that seemingly grow in number as each day passes. According to Galperin, there are 858 databases, 139 more than the previous year, available to the public [8]. The amount of data produced urged the necessity for fast and reliable ways of accessing, retrieving, researching and understanding these data [9, 10]. With the rapid increase of protein data in the databases, there was a need for a data profiler.

“Like it or not, many of the assumptions you have about your data are probably not accurate.  Despite our best efforts, gremlins inevitably find their way into our systems [11].” (Hover, 2013).  That is why the quality of data becomes sacrificed as an end-result of that problem. To give solution to this problem, many researchers found a way for data profiling. Data profiling, also called data archeology, is the statistical analysis and assessment of [data](http://searchdatamanagement.techtarget.com/definition/data)values within a [data set](http://whatis.techtarget.com/definition/data-set) for consistency, uniqueness and logic. Profiling tools evaluate the actual content, [structure](http://searchsqlserver.techtarget.com/definition/data-structure) and [quality](http://searchdatamanagement.techtarget.com/definition/data-quality) of the data by exploring relationships that exist between value collections both within and across data sets [12]. For Ralph Kimball, data profiling is a systematic analysis of the content of a data source. It is “systematic” in the sense that it’s thorough and looks in all the “nooks and crannies” of the data [11].

However, there are already a lot of existing databases for proteins to date. “*Some of them are of general character, but some are dedicated to specific aspects of protein structures or to specific protein families, specific metabolic pathways, etc.*” (Al Karadaghi, 2015). Some of the popular databases that are of general character include UniProt. Uniprot’s mission is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information [13]. On the other hand, some popular databases that has specific protein structures is the SPdb: A Signal Peptide Database. The SPdb is a signal peptide database containing signal sequences of archaea, prokaryotes and eukaryotes [14]. Our system may be similar to that of the UniProt which have a general character of the protein databases but there is probably none that has an internal redundancy checker which our system provides. Pennisi, Wiley and Michaels tells us that still up to today, many researchers are confronted by similar obstacles in accessing up-to-date data, which are withheld from public access by method developers [15, 16]. This study is aiming at providing the needed data for the general public for any use this might be beneficial for them.

Methodology

**A. Data sets**

Partial sets of signal and non-signal peptides are examined in this study. These were obtained from UniProt release 2016\_04 and are downloaded in text format. Entries that belong to the Swiss-Prot section of UniProtKB are those entries that have been manually annotated by experts and reviewed by UniProtKB curators. These entries are tagged with a yellow star of their entry status and are the subject of this study. The compilation of dataset consisted of the following:

1. **Signal Peptide Entries Retrieval**

In order to retrieve signal peptide entries, entries are fed into a keyword filter. A term ‘Signal’ indicated in the Sequence Annotation, Features (FT) section, or in Kewyords (KW) section retrieves a signal peptide entry.

1. **Non-Signal Peptide Entries Retrieval**

In order to retrieve non-signal peptide entries, entries are fed into a keyword filter. A term without ‘Signal’ indicated in the Sequence Annotation, Features (FT) section or in Keywords (KW) section, retrieves a non-signal peptide entry.

**B. Experimental Findings**

All entries are subjected to experimental findings. Entries whose protein existence are not clear will fall as non-experimental entries. These entries have not been strictly proven, have probable evidence, have unsure evidence or without protein existence. These types of non-experimental evidence are termed as follows:

***‘Evidence at transcript level’*** indicates that the existence of a protein has not been strictly proven but that expression data (such as existence of cDNA(s), RT-PCR or Northern blots) indicate the existence of a transcript.

***‘Inferred by homology’*** indicates that the existence of a protein is probable because clear orthologs exist in closely related species.

***‘Predicted’* indicates that an entry is** without evidence at protein, transcript, or homology levels.

***‘Uncertain’*** indicates that the existence of the protein is unsure.

These terms are indicated in the Protein Attributes, Protein Existence (PE) section, of an entry. A clear evidence set are entries with the term ‘Evidence at protein level’ in the PE section. Stated below is the full description of the term.

***‘Evidence at protein level’*** indicates that there is a clear experimental evidence for the existence of the protein. The criteria include partial or complete Edman sequencing, clear identification by mass spectrometry, X-ray or NMR structure, good quality protein-protein interaction or detection of the protein by antibodies as stated in the user manual of UniProt.

**C. Classification Based on Taxonomy**

All entries are classified to the following Taxonomy (Superkingdom): Archaea, Bacteria, Eukaryota and Viruses. These are indicated in the Taxonomy (OC) section, a subsection of the ‘Names and Taxonomy’. This contains the taxonomic hierarchal classification lineage of the source organism. Only the first listed classification, the Superkingdom, in the hierarchy is being used as the keyword.

**D. Classification Based on Subcellular Location**

An entry can either be classified as a Transmembrane or a Non-Transmembrane. To check where an entry belongs, check on these three sections:

1. Keyword (KW) section
2. Features (FT) section,
3. Subcellular Location (CC) section

Any of the three sections signifies a membrane-spanning or a non-membrane-spanning region of the protein. If a ‘Transmembrane’ or ‘Transmem’ keyword exists, a protein is classified as a Transmembrane protein. If not, a protein is classified as a Non-transmembrane protein.

For Transmembrane proteins, each of the sections have different uses as a keyword filter. In KW section, this would signify a Transmembrane entry. In FT section, the count of ‘TRANSMEM‘ keyword signifies a Single-pass or a Multi-pass membrane. In CC section, this signifies a more detailed classification of a Single-pass membrane protein.

The definition of a Single-pass membrane protein is a protein spanning the membrane once. These have terms on the CC section such as ‘Single-span’, ‘Singlespan’, ‘Single-pass’ or ‘Singlepass’. The definition of a Multi-pass membrane is a protein spanning the membrane more than once. It is based on its N-terminus and C-terminus. These have terms on the CC section such as ‘Multi span’, ‘Multispan’, ‘Multipass’, ‘Multi pass’, ‘Multipass’ or ‘Polytopic membrane protein’.

We already defined a single-pass membrane protein as a protein spanning the membrane once. Further breakdown of the classification of a single-pass membrane protein is based on its N-terminus and transmembrane domain location. These are classified to four types namely: ‘Single-pass Type I’, ‘Single-pass Type II’, ‘Single-pass Type III’, and ‘Single-pass Type IV’.

There are Multi-spanning entries having ‘Transmembrane’ keywords but don’t have FT TRANSMEM line in which contradicts our method of classifying a transmembrane protein. All Multi-spanning membrane proteins have transmembrane regions. It just happen, for beta-stranded transmembrane regions, these are not annotated. As explained by a curator of UniProt, Beta strand transmembranes are found in the outer membranes of bacteria (both Gram negative and acid fats Gram positive) and etc. Such transmembrane domains are however not predicted by prediction programs such as TMHMM or ESKM. As a consequence, such entries frequently have no FT TRANSMEM line, although they contain the ‘Transmembrane’ keywords in KW section.

For Non-transmembrane protein, also termed as Globular proteins, have two classifications. One is the secretory protein and the second one is the non-secretory protein. These are not further classified by the system but it provides the location of where the protein exits or resides.

**E. Redundancy Checker**

After entries are filtered, we may apply data reduction or data redundancy check procedure. According to Sikic [17], the inclusion of similar sequences in certain analyses will introduce undesirable biases. Therefore, removing data redundancy procedure is important for it removes protein sequences that overreach certain similarity thresholds.

Biological data are vastly increasing and it need tools to eliminate redundancies and able to make full use of the functions of those data that are greatly needed by sciences and those working on it on laboratories such as scientists, doctors, and alike.

**Choosing of a Redundancy Program**

There are different programs available to check data redundancy such as the following ‘Pisces’ [18], ‘BlastClust’ [19], ‘Decrease redundancy’ [20], ‘cd-hit’ [21], ‘SkipRedundant’[22], and etc. The non-redundant datasets resulted from the five programs mentioned are moderately similar to each other where same program is fed and with the same percentage of identity threshold [17]. All of their outputs are more than acceptable in terms of residual similarity between the entries that are grouped in the outputs [17].

For this system, ‘Pisces’ (<http://dunbrack.fccc.edu/Guoli/PISCES_InputD.php>), is used as the data redundancy removal program. The advantage of this program from the other four programs is that it is an open source software and the sequence percentage identity or similarity can range from 0 to 1.

References

[1] Lesk, A. M. Introduction to Protein Architecture. OUP, Oxford, 2001.

[2]Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach [Internet] *Why Is It Important to Study Proteins* [Online] [Accessed May 16, 2016] **<**http://www.ncbi.nlm.nih.gov/books/NBK6824/#A176**>**

[3] Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th Edition. *Analyzing Protein Structure and Function.* New York: [Garland Science](http://www.garlandscience.com/textbooks/0815341059.asp); 2002. [Online] [Accessed May 16, 2016]

< http://www.ncbi.nlm.nih.gov/books/NBK21054/>

[4] Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach *Chapter A06 Protein Structure, Modelling and Applications* [Online] [Accessed May 18, 2016]

< http://www.ncbi.nlm.nih.gov/books/NBK6824/>

[5] Basics of Protein Structure Protein Databases: Short Overview

[Online][Accessed May 18, 2016]

< http://www.proteinstructures.com/Structure/Structure/proteinstructure-databases.html>

[6] Studying Proteins and Protein Purification *Why study proteins*

[Online][Accessed May 18, 2016]

http://www-users.med.cornell.edu/~jawagne/proteins\_%26\_purification.html

[7] Burley S. K., Almo S. C., Bonanno J. B., Capel M., Chance M. R., Gaasterland T., Lin D., Sali A., Studier F. W., Swaminathan S. Structural genomics: beyond the Human Genome Project. Nat. Genet. 1999;23:151–157

[8] Galperin M. Y. The Molecular Biology Database Collection: 2006 update. Nucl. Acids Res.2006;34:D3–D5.

[9] Luscombe N.M., Greenbaum D., Gerstein M. What is Bioinformatics? A proposed definition and overview of the field. Method Inform. Med. 2001;40:346–358.

[10] Ouzonis C. A., Valencia A. Early bioinformatics: the birth of a discipline – a personal view. Bioinformatics. 2003;19:2176–2190.

[11] Jason Hover Data Profiling: What, Why and How? *Intro to Data Quality*

[Online] [Accessed May 18, 2016]

< https://datasourceconsulting.com/data-profiling/>

[12] Search Data Management *Data Profiling* [Online] [Accessed May 18, 2016]

<http://searchdatamanagement.techtarget.com/definition/data-profiling>

[13] http://www.uniprot.org/

[14] Choo KH, Tan TW, Ranganathan S. 2005. SPdb – a signal peptide database. BMC Bioinformatic 6:249

[15] Keeping genome databases clean and up to date.

*Pennisi E* *Science. 1999 Oct 15; 286(5439):447-50.*

[16] Should software hold data hostage? *Wiley HS, Michaels GS* *Nat Biotechnol. 2004 Aug; 22(8):1037-8.*

[17] Kresimir Sikic, Oliviero Carugo. Protein sequence redundancy reduction: comparison of various method. 2010. [Online] [Accessed July, 2012]

< http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3055704/>

[18] http://dunbrack.fccc.edu/Guoli/PISCES\_InputD.php

[19] http://toolkit.tuebingen.mpg.de/blastclust

[20] http://web.expasy.org/decrease\_redundancy/

[21] http://weizhongli-lab.org/cdhit\_suite/cgi-bin/index.cgi?cmd=cd-hit

[22] http://www.bioinformatics.nl/cgi-bin/emboss/skipredundant