Protein Database Profiler

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

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

Kezah P. Alferez
______Jessica D. Pacilan_____
Department of Computer Science

__Prof. Demelo M. Lao__
Adviser

May 2016

Approval Page

This undergraduate special problem entitled **“Protein Database Profiler”** 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**.

ATTY. LIZA D. CORRO
Dean

____
Date

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:

i. 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.

ii. 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:

i. Keyword (KW) section

ii. Features (FT) section,

iii. 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; 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