A novel methodology on distributed representations of proteins using their interacting ligands:

A case-study on Sphingolipid Metabolic Pathway

Hakime Öztürk[1,*], **Mehmet Aziz Yirik**[2,*], **Arzucan Özgür**[1], **Kutlu O. Ulgen**[2,3] and **Elif Ozkirimli**[2,3]

Department of { **1** Computer Engineering, **2** Computational Science and Engineering, **3** Chemical Engineering }
Boğazici University, 34342, Istanbul, Turkey

* equal contribution

Abstract

❖ We propose a novel machine learning based method to represent proteins with their ligands.

❖ The proteins are represented using the word-embeddings of the SMILES representation of their ligands and the performance of the protein representation is evaluated on protein clustering task.

❖ The results show that ligand-based representation of proteins perform as well as protein sequence based methods.

Introduction

❖ Representation of proteins is an important task in many bioinformatics problems.

❖ Based on the chemogenomics assumptions, we propose that proteins can be described using the set of ligands that they interact with.

❖SMILES representation of ligands [1] is utilized.

❖ Two different methods are used for comparison:

- A protein sequence-based model that uses word-embeddings [2],
- A ligand-centric protein-protein interaction (PPI) network [3].

METHODS

Data Collection

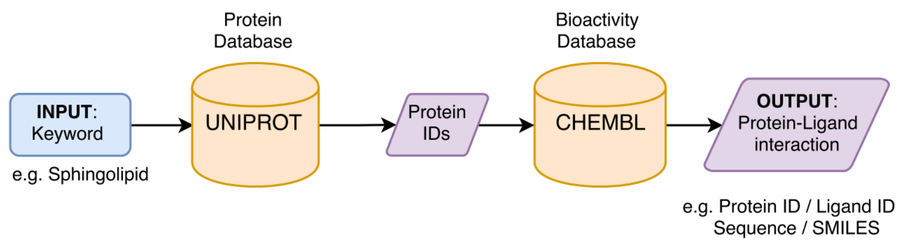


Fig 1: Collecting protein-ligand interaction information

Distributed Representation of Proteins and Ligands

❖ Distributed word representations models (word embeddings) comprise the syntactic and semantic features of the words [4].

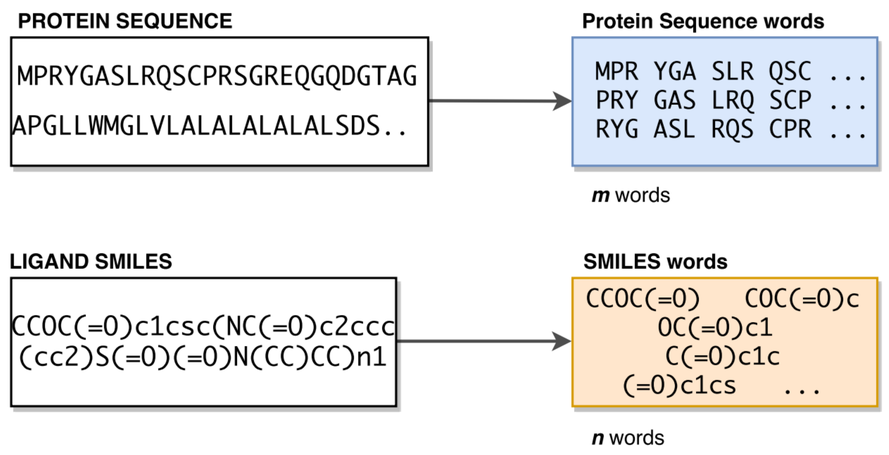


Fig 2: For each protein/ligand word that is extracted from protein sequence/ligand SMILES, a real-valued vector (embedding) is learned from a large training set.

❖Prot2Vec describes a protein as follows [2]:

$$\text{Prot2Vec} = vector(protein) = \frac{\sum_{k=1}^{m} vector(word_k)}{m}$$

$vector(word_k)$ is the word-embedding of k^{th} word of protein sequence (e.g. “MPR”).

SMILES-based Protein Representation

❖ SMILES2Vec describes a ligand as follows:

$$\text{SMILES2Vec} = vector(ligand) = \frac{\sum_{k=1}^{n} vector(word_k)}{n}$$

❖ For a protein interacting with *p* ligands:

$$vector(protein) = \frac{\sum_{k}^{p} vector(SMILES2Vec_k)}{p}$$

Ligand-centric PPI Network

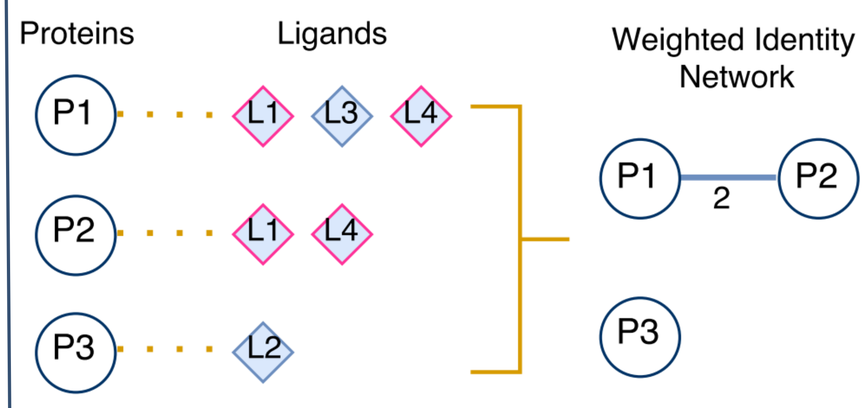


Fig 3: The weighted identity network (WIN).[3]. .

Experiment Setting

For Sphingolipid (SL) metabolism related proteins:

❖ their ligands were collected (51 prot, ~84.5K lig).

❖ vector forms were created using Prot2Vec and SMILES2Vec and K-means was applied for clustering.

❖ WIN was constructed and Markov Clustering (MCL) algorithm was used.

RESULTS

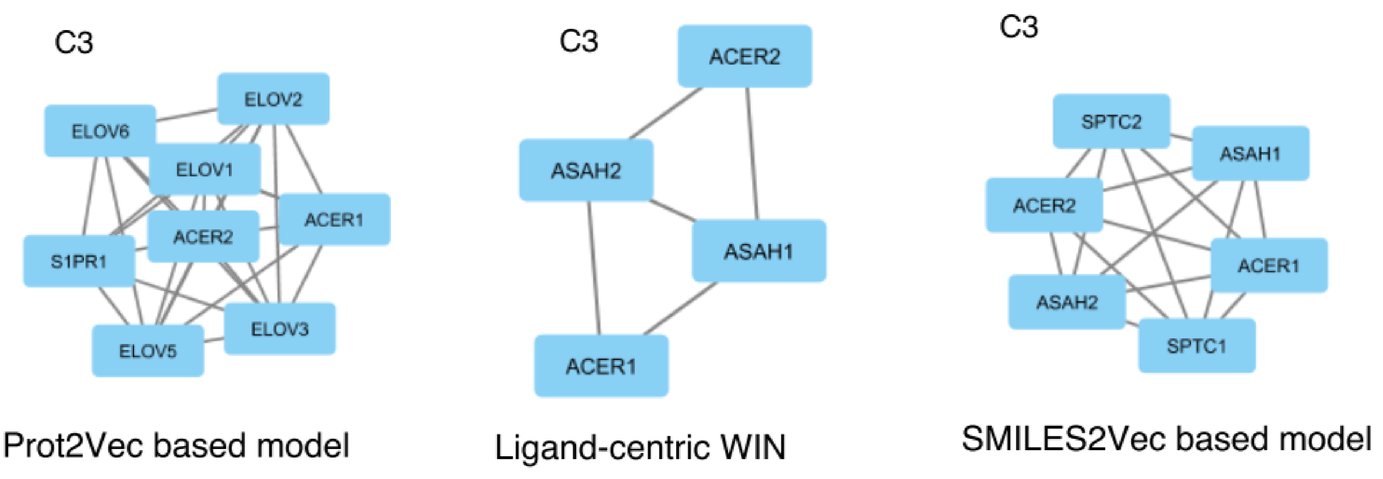
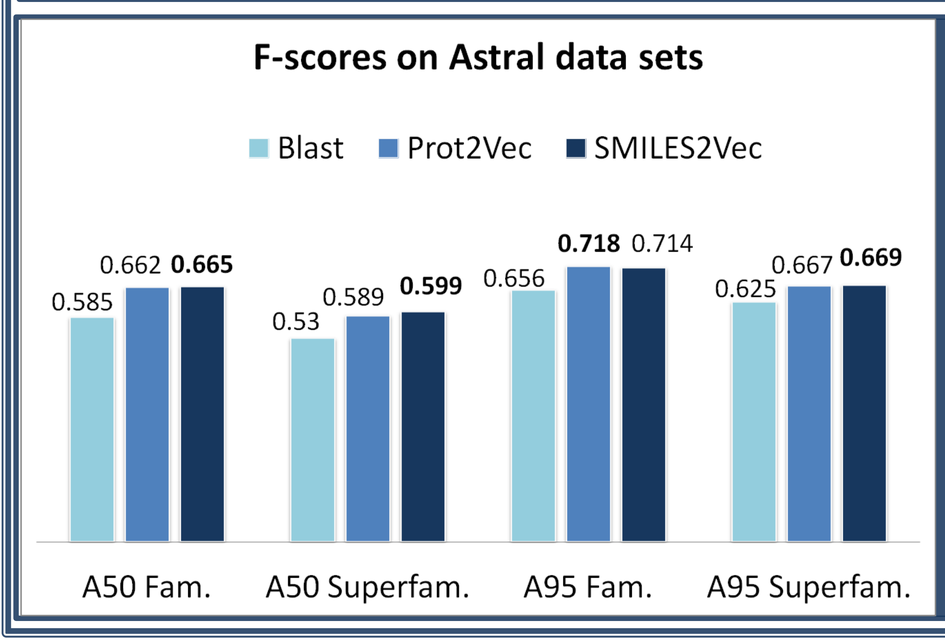


Fig 4: WIN clusters have proteins from same families due to their high number of common ligands and SMILES2Vec brings out the similarity aspect of ligands. The clusters of the Prot2Vec-based proteins have high sequence similarity.



❖ The performances of the protein representation methods were evaluated on Astral50 & Astral95 [5] protein family /superfamily subsets.

❖ These subsets are filtered based on their ligand binding information. (1607p, 2604p)

❖ MCL was used for identifying clusters.

Conclusion

❖ Using SMILES2Vec, we were able to define proteins based on their interacting ligands even in the absence of sequence or structure information.

❖ SMILES2Vec-based protein representation performed as well as protein sequence based methods in protein clustering problem.

References

1. Öztürk, H et al. "A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction." *BMC bioinformatics* , 2016
2. Asgari, E, & M Mofrad. "Continuous distributed representation of biological sequences for deep proteomics and genomics“ *POne*, 2015
3. Öztürk, H,.et al. "Classification of Beta-lactamases and penicillin binding proteins using ligand-centric network models." *POne,* 2015
4. Mikolov, T, et al. "Distributed representations of words and phrases and their compositionality." *Advances in NIPS*. 2013
5. Fox, N.K. *et al. “*Scope: Structural classification of proteins—extended, integrating scop and astral data and classification of new structures.”, *NAR, 2013*

hakime.ozturk; aziz.yirik; arzucan.ozgur; ulgenk; elif.ozkirimli @boun.edu.tr

Acknowledgment

• This research is supported by Bogazici University Research Fund Grant Number 12304.
• TÜBİTAK BİDEB 2211 Scholarship Programme is gratefully acknowledged.

Travel funding to ISMB/ECCB 2017 was generously provided by ISCB.