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

Hakime Öztürk[1] , Elif Ozkirimli[2] and Arzucan Özgür[1]

Departments of { **1** Computer Engineering, **2** Chemical Engineering }, Boğaziçi University, İstanbul, Turkey

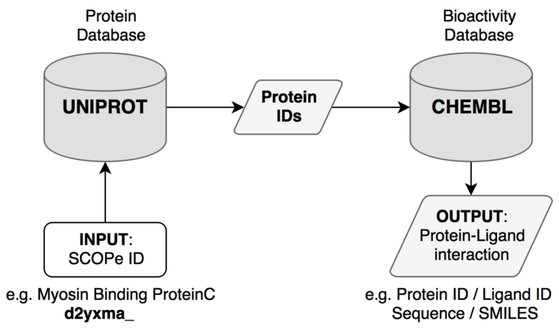
1. ABSTRACT

The effective representation of proteins is a crucial task. Related proteins usually bind to similar ligands. Chemical characteristics of ligands are known to capture the functional and mechanistic properties of proteins suggesting that a ligand based approach can be utilized in protein representation.

We propose a novel method to compute similarity of proteins by describing them based on their ligands' SMILES.

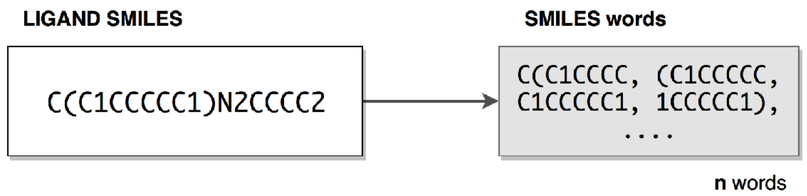
2. METHODS

COLLECTING INTERACTING LIGANDS OF PROTEINS

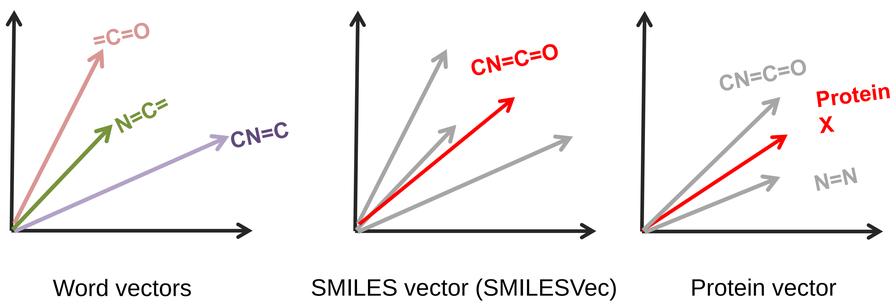


We used ASTRAL A-50 dataset of SCOP database as benchmark **[1]**.

LIGAND-BASED PROTEIN REPRESENTATION



For each chemical word that is extracted from ligand SMILES, a real-valued vector (embedding) is learned from a large training set **[2]**



Word vectors | SMILES vector (SMILESVec) | Protein vector

❖ SMILESVec describes a ligand as follows:

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

❖ For a protein interacting with *nl* ligands:

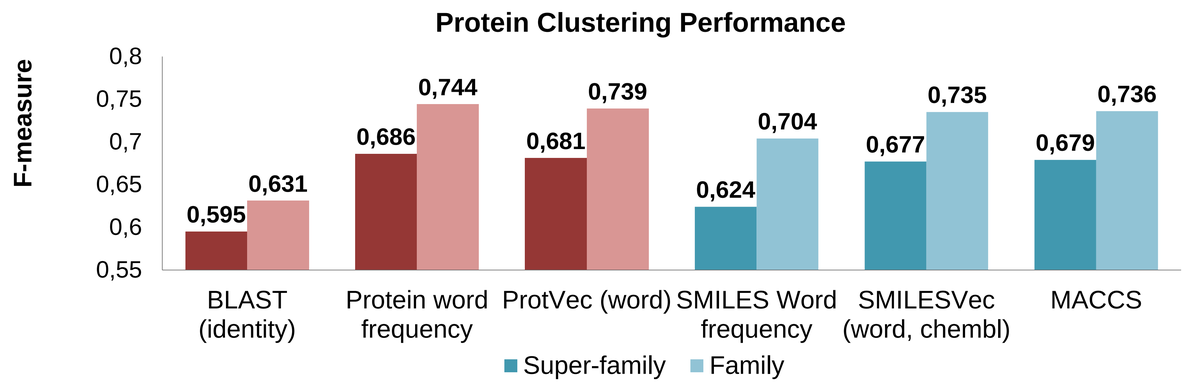
$$vector(protein) = \frac{\sum_{k=1}^{nl} vector(SMILESVec_k)}{nl}$$

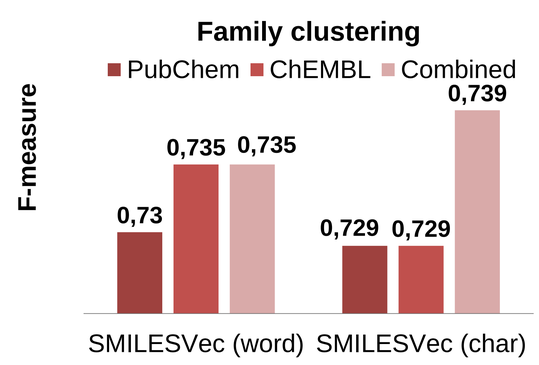
3. RESULTS

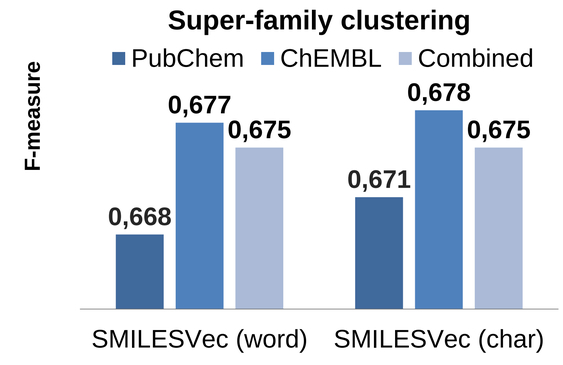
Table 1. Distribution of families and super-families in A-50 dataset before and after ligand-interaction based filtering

Dataset	Num. Seq.	Super-families	Families
Before	10816	1080	2109
After	1639	425	652

- Total 10816 proteins, only (1639) 15% of them interacts at least a ligand.
- 64% of all interacting proteins have ligands fewer than 200 and 0.6% of all proteins are with single ligands.
- The mean number of the interacting ligands is 1791.







The clustering is completed with TransClust **[3]** algorithm following similar pipeline to **[4]**.

4. CONCLUSION

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

❖ We showed that ligand-based protein representation, which uses only SMILES strings of the ligands that proteins bind to, performs as well as protein-sequence based representation methods in protein clustering.

❖ Ligand-based protein description can be applied to different bioinformatics problems such as prediction of new protein-ligand interactions and protein function annotation.

5. References

1. Fox, N. K. *et al.* 2013
2. Mikolov et al., 2013
3. Wittkop *et al.*, 2010
4. Bernardes *et al.*, 2015

TUBITAK-BIDEB 2211-E and Bogaziçi University BAP are gratefully acknowledged.