단백질 구조를 통한 기능 예측 모델 고안

이민지[1]○ Anar Rzayev[1] 정현규[1,2] Luiz Felipe Vecchietti[2] 차미영[1,2] 김호민[3]
[1]한국과학기술원 전산학부 [2]기초과학연구원 데이터 사이언스 그룹 [3]기초과학연구원 단백질 커뮤니케이션 그룹
{haewon_lee, rzayev.anar1, dino8egg, hm_kim}@kaist.ac.kr, {lfelipesv, mcha}@ibs.re.kr

Structure-based representation for protein functionality prediction using machine learning

Minji Lee[1]○ Anar Rzayev[1], Hyunkyu Jung[1,2] Luiz Felipe Vecchietti[2] Meeyoung Cha[1,2] Homin Kim[3]
[1]School of Computing, KAIST [2]Data Science Group, Institute for Basic Science [3]Protein Communication Group, Institute for Basic Science

요 약

Enhancing the properties of a protein is crucial for its pharmaceutical and industrial usages. However, due to the vast protein space, evaluating the functionality of a protein is not trivial. Using the recent breakthroughs in deep learning-based structure prediction networks, we investigate a structure-based representation to train a protein functionality regressor that can accurately guide protein engineering. The proposed structure-based representation was used to train the functionality regressor of a green fluorescent protein from *Aequorea victoria*, and it outperformed sequence-based representation baselines.

1. Introduction

Proteins are complex molecules responsible for different functions in the human body. Proteins are also an important part of products in the food, pharmaceutical, and chemical industries. Protein engineering, or the capacity to improve a protein's function and properties, is critical for its effective usage in various applications. Protein engineering is accomplished by introducing a series of mutations from a wild-type amino acid sequence, *i.e.,* the protein to be optimized. However, because the sequence space for a target function is usually sparse [1], assessing the quality of mutations is not trivial. For a green fluorescent protein from *Aequorea victoria* (avGFP) wild-type, for example, 50% of the mutants with just four amino acid mutations become completely non-fluorescent [9]. Because the functionality to be optimized is directly guided by protein engineering, it is critical to accurately predict the change in functionality from experimental data.

Proteins are represented by a sequence of characters, each of which represents one of the 20 standard amino acids. A protein sequence can be represented in its simplest form by a one-hot encoding vector given the number of amino acids. However, because proteins have evolutionary related sequences, previous studies [1,3,7] investigated the protein space by attempting to decipher their language: the amino acid sequence. These methods focus on obtaining a good representation of an amino acid sequence. Sequence-based methods typically train a feature extractor to generate continuous, fixed-size embeddings from the discrete, variable-size amino acid sequence, where the feature extractor consists of complex and state-of-the-art natural language processing models such as transformers [10]. The representations obtained by these methods can be used as input features to train machine learning algorithms and predict important functionalities and properties that can guide protein engineering.

Even though sequence-based methods can extract important features of the protein space, they do not take the protein structure into account. The protein structure in 3-dimensional space of its sequence of amino acids is closely related to its function. For example, the fold of avGFP is 11-stranded beta-barrels wrapped around a chromophore which emits fluorescence (Fig. 1). The beta-barrels form a cylinder around the chromophore, which is thought to be responsible for the high yield of fluorescence and stability [8].

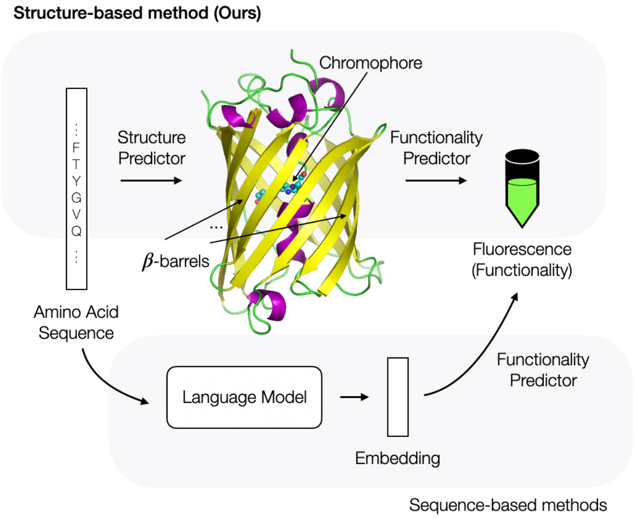

Fig 1. Overview of the functionality prediction methods. Structure-based methods (top), sequence-based methods (below).

Using the protein structure as an input feature has been difficult due to the fact that protein structures could not be accurately predicted by in silico methods in the last decades. Recently, however, breakthrough results were obtained via deep neural network architectures for the structure prediction problem [2,4,5]. Structure prediction networks, notably trRosetta [4], are used to infer the structure of a mutant to guide the functionality predictor. The high accuracy of structure prediction networks opened the possibility of using these methods as tools for different applications, including the investigation of structure-based representations for predicting functionalities important for protein engineering. Changes in structure help improve the prediction methods when compared to using only sequence data, noting the difference between mutant sequences and wild type sequences

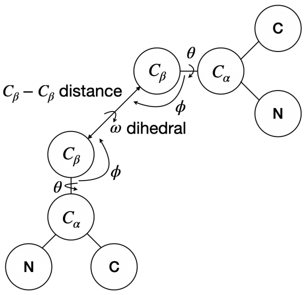

Fig 2. Features of trRosetta

consists only on a few mutations on a long amino acid sequence.

In this paper, we investigate the use of a structure-based representation for the prediction of green fluorescence in a protein. The structure of a mutant is obtained using a deep learning-based structure prediction network [4], and a feature vector representation is proposed by comparing the mutant structure to the wild-type structure. This feature vector is then used to train an explainable machine learning (ML) regressor for the green fluorescent functionality prediction problem using an experimental dataset [9]. The proposed structure-based representation improves the regression accuracy when compared to the one-hot and sequence-based representations.

2. Proposed Method

The proposed method is described in three parts. In the first and second parts, the structure-based representation used to prepare the dataset for training is presented. In the third part, ML methods used for training the functionality regressor are described. An overview of the training framework is presented in Fig. 3.

2.1. trRosetta structure prediction network

trRosetta [4] considers structure prediction as a classification problem of four features: $C_\beta - C_\beta$ distance between amino acids, *i.e.,* $C_\beta - C_\beta$ distance, ω dihedral, *i.e,.* rotation along the virtual axis connecting the C_β atoms of two amino acids, and two angles specifying the direction of the C_β atom of one amino acid in a reference frame centered on another amino acid (Fig. 2). The $C_\beta - C_\beta$ distance is discretized from a range of 2Å to 20Å with each bin size set to 0.5Å, also including one additional bin when the distance exceeds 20Å for a total of 37 discrete classes. The Rosetta software suite [2] refines and reconstructs the final protein 3D structure using the probability distribution predicted by trRosetta for these properties.

2.2. Structure-based representation

The protein structure is inferenced from a sequence using the trRosetta structure prediction network described above. Here, our focus is on the $C_\beta - C_\beta$ distance probability distribution, or *distogram,* of the mutant sequence and wild type sequence. As distances are discretized into discrete bins, the argument max of the distogram is taken to obtain the bins with the maximum predicted probability. Protein distograms are symmetric and usually sparse. In order to reduce the sparsity of the final feature vector, the distograms are filtered to the pixels in which the wild-type distogram is non zero, *i.e.*, the amino acid pair is closer than 20Å. The formalization of this process is as

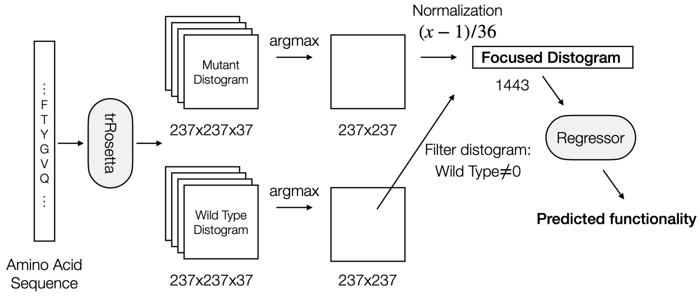

Fig 3. Overview of the framework.

follows. For a mutant distogram $M_{i,j}$ and wild-type distogram $W_{i,j}$ where $i,j \in \{1 \cdots 237\}$, the proposed set of normalized features F is defined as $F = \{(\mathrm{M_{i,j}} - 1)/36 | \mathrm{W_{i,j}} \neq 0\}$.

The normalization is performed to transform the discrete bins representing distances from 2Å to 20Å to values from 0 to 1, and distances larger than 20Å to a negative value. For the rest of this paper, the feature vector F is called *focused distogram*, and the unfiltered feature vector is called *full distogram.*

2.3 ML-based fluorescence regressor

For the experiments, three ML regressors are used: Ridge [11], ElasticNet [12], and gradient boosting [6]. Ridge is used to compare different representations due to its fast training and inference time. All three methods are used for ablation studies with the proposed structure-based representation.

3. Experimental Results

3.1. Experimental settings

The proposed method is trained and tested on the dataset proposed by Sarkisyan et al [9] which is comprised of 54045 sequences of avGFP mutants paired with their fluorescence value. The dataset was randomly split into train (90%) and test (10%). Smaller splits (10000 and 500 samples) were also sampled from the train set for experiments. Distograms for the wild-type and all mutants were collected using trRosetta. Five protein representations, two structure-based and three sequence-based, were used for comparison: focused distogram, full distogram, one-hot encoding, UniRep [1], and UniRep64 [1]. UniRep and Unirep64 differ on the size of the generated feature vector. Using each representation as input, a ridge regressor with $\alpha = 1.0$ is trained to predict fluorescence.

3.2. Results

Table 1. Comparison of mean squared error between representations. All representations are tested on same test set except †,‡. For †, 5-fold cross-validation results are shown due to long inference time of UniRep. For ‡, regression method is linear regression and we compared with the result from [14] which used the same dataset but with a different split. Best results on each dataset are **highlighted.**

Representation	No. Features	Train data size		
		Full	10000	500
Focused distogram	1443	0.4722	0.5488	**0.8049**
Full distogram	56169	**0.4556**	**0.5295**	1.0022
One-hot	4740	1.1171	1.1172	1.1190
UniRep [1]	1900	1.32‡	0.6858†	0.8673†
UniRep64 [1]	64	1.0826	1.0889	1.1267

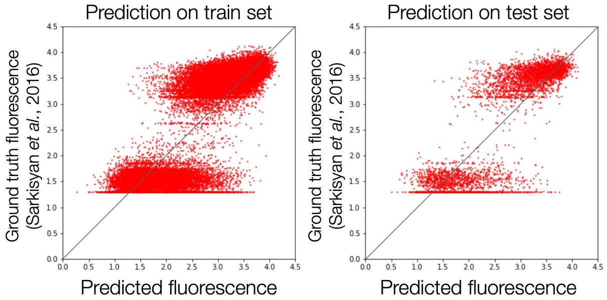

Fig 4. Fluorescence prediction on train and test set. LightGBM is trained on full train set of focused distogram representation.

Table 2. Comparison of mean squared error between different regression methods. All representations are tested on same test set.

Regression Methods	Data size		
	Full	10000	500
LightGBM [6]	**0.3017**	**0.3511**	**0.6467**
Ridge [11]	0.4722	0.5488	0.8049
ElasticNet [12]	1.1171	1.1171	1.1173

As shown in Table 1, the distogram representations achieved the smallest mean-squared error (MSE) for all datasets. For two larger train sets, full distogram achieved the best MSE and for the smallest train set, focused distogram achieved the best MSE. However, focused distogram showed competitive results when compared to full distogram with only 2.6% of the number of features. By filtering amino acid pairs that interact in the wild-type protein, the regressor was able to focus on important interactions.

Next, the performance of the proposed focused distogram with different regression methods is compared. Due to CPU limitations, we chose to conduct the ablation study only with the focused distogram. As shown in Table 2 and visualized in Fig. 4, the LightGBM regressor achieved the smallest MSE for all dataset splits. It is to be observed that the best value of approximately 0.30 is better when compared to other sequence-based benchmark results [14].

4. Discussion

Existing works on protein representation learning [1,3,7] mostly focused on sequence-based methods. However, our comparison between representations shows that the structure of the protein encodes more valuable information than the embeddings extracted from the amino acid sequence. This result is quite reasonable since the structure has explicit relation with function, where amino acid sequence does not. Also, the sequence-based methods are trained on whole protein database, without the information about the target function. So the representation may not include the property of the protein which is crucial when predicting certain functionality. This is also the reason why sequence-based method show increased accuracy when fine-tuned to specific target proteins [13].

It was also interesting to observe that the proposed representation is able to achieve accurate results even with a dataset with only 500 samples. Because collecting large amounts of functionality data experimentally is expensive, the ability to predict functionality successfully with limited data will be critical in guiding protein engineering and protein design in the years ahead.

However, the limitation of our work is that the distogram representation does not include the information of side chains. Side chains of amino acids determine the interaction of the protein with other molecules, and the properties of the protein. Therefore, we aim to design the representation that both includes the structure and side chains of each amino acid. Also, we plan to train deep learning-based regressor with full distogram representation on full train set, since we can overcome CPU limitation with batch training.

5. Conclusion

This paper studied a structure-based representation to train a protein functionality regressor that can be used for protein engineering. The structure-based representation is created by processing the structure inferenced by a deep learning-based structure prediction network. The proposed structure-based representation was used to train a green fluorescent protein functionality regressor and was able to achieve an improvement in the MSE over sequence-based representation baselines. Proposed structure-based representations also surpassed sequence-based representations by using only 5% of the original train set. Such results show the potential of using structure-based regressors as objective functions of various protein engineering and design tasks.

7. References

[1] Alley, Ethan C., et al. "Unified rational protein engineering with sequence-based deep representation learning." Nature methods 16.12 (2019): 1315-1322.

[2] Baek et al. "Accurate prediction of protein structures and interactions using a three-track neural network." Science 373.6557 (2021): 871-876.

[3] Brandes, Nadav, et al. "ProteinBERT: A universal deep-learning model of protein sequence and function." Bioinformatics 38.8 (2022): 2102-2110.

[4] Zongyang Du, Hong Su, Wenkai Wang, Lisha Ye, Hong Wei, Zhenling Peng, Ivan Anishchenko, David Baker, and Jianyi Yang. The trrosetta server for fast and accurate protein structure prediction. Nature protocols, 16(12):5634–5651, 2021.

[5] Du et al. "The trRosetta server for fast and accurate protein structure prediction." Nature protocols 16.12 (2021): 5634-5651.

[6] Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017).

[7] Lu, Tianyu, Alex X. Lu, and Alan M. Moses. "Random Embeddings and Linear Regression can Predict Protein Function." arXiv preprint arXiv:2104.14661 (2021).

[8] Ormö et al. "Crystal structure of the Aequorea victoria green fluorescent protein." Science 273.5280 (1996): 1392-1395.

[9] Sarkisyan et al. "Local fitness landscape of the green fluorescent protein." Nature 533.7603 (2016): 397-401.

[10] Vaswani et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[11] Hoerl, Arthur E., and Robert W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems." Technometrics 12.1 (1970): 55-67.

[12] Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." Journal of the royal statistical society: series B (statistical methodology) 67.2 (2005): 301-320.

[13] Biswas, Surojit, et al. "Low-N protein engineering with data-efficient deep learning." Nature methods 18.4 (2021): 389-396.

[14] Lu et al. "Self-supervised contrastive learning of protein representations by mutual information maximization." BioRxiv (2020).

이 연구는 기초과학연구원[IBS-R029-C2, IBS-R030-C1]과 한국연구재단[NRF-2017R1E1A1A01076400]의 지원을 받았음.