그래프 기반 딥러닝을 활용한 항체 서열 디자인

Begench Hangeldiyev, Anar Rzayev, Azamat Armanuly , Luiz Felipe Vecchietti , 차미영, 김호민
한국과학기술원, 기초과학연구원
{begahan, rzayev.anar1, azeke908}@kaist.ac.kr, {lfelipesv}@gmail.com, {mcha, kimhm}@ibs.re.kr

Antibody Sequence Design With Graph-Based Deep Learning Methods

Begench Hangeldiyev, Anar Rzayev, Azamat Armanuly , Luiz Felipe Vecchietti , Meeyoung Cha, Ho Min Kim
KAIST, IBS

요 약

Recently, deep learning methods based on graph representations achieved breakthrough results in protein sequence design. One of these methods, named ProteinMPNN, increased substantially the success rate of designs in wet lab experiments. However, for antibody structures, given the limited number of experimental structures and the diversity of loop-like conformations in variable areas, sequence design remains challenging. In this paper, we conduct a study on the performance of protein sequence design approaches for a collection of antibody structures and compare the findings with a proposed model trained solely on antibody data. When compared to a model trained on general protein datasets, structures predicted for sequences designed by the proposed antibody-specific model achieve a reduced root-mean-square deviation with the native structure.

1 Introduction

Lately, the Covid-19 pandemic highlighted the need for the fast development of therapeutics able to neutralize a target antigen. For that, a protein structure able to interact with the antigen is desired. With this designed structure, a protein sequence should be decoded to maximize the chances of effective expression and binding in wet lab experiments. This process of decoding an amino acid sequence from a protein structure is defined as a protein sequence design problem.

Recent methods that represent a protein as a proximity graph over amino acids achieved breakthrough results in protein sequence design [1-5]. In a graph-based protein representation, each node represents an amino acid (or atom) and each edge of the graph represents the structural neighborhood of an amino acid (or atom). Representing the protein as a graph, the features are associated to nodes, i.e. node features, and to edges, i.e. edge features. Relational reasoning over this graph structure can be performed using deep learning methods such as Graph Neural Networks (GNNs) [6] and Geometric Vector Perceptrons (GVPs) [3], and, from the features learned, the protein sequence is decoded. These methods are highly computationally efficient being able to design sequences in a few seconds. Additionally, a method name ProteinMPNN [5] has been experimentally validated with high developability in wet lab experiments across different tasks.

In this paper, we aim to study the performance of protein sequence design methods for a set of antibodies. Sequence design of antibodies is challenging, given that only a small subset of protein datasets contain antibody structures. Also, the loop-like conformations and vast diversity of possible amino acid sequences in the antibody variable regions, e.g. the complementary-determining regions (CDRs), increases the difficulty of this problem. We propose and train an antibody-specific model, named Ing_Ab, based on the algorithm introduced by Ingraham et al [1]. For comparison, we choose the current state-of-the-art method for protein sequence design, ProteinMPNN [10]. A set of antibodies introduced after the training of our model is collected from an antibody structure database, SabDab [9], for evaluation. Sequences obtained from Ing_Ab and ProteinMPNN then have their structure predicted by an antibody-specific structure prediction network named IgFold [7], and the predicted structures are compared with the native antibody structures. We show that the structures predicted for the sequences designed by the proposed antibody-specific model achieves lower root-mean-square deviation (RMSD) when compared with ProteinMPNN, a model trained using a general protein dataset.

2 Background

2.1 Graph-Based Protein Sequence Design *(Ingraham et al, 2019)*

Ingraham et al [1] presented a relational language model for decoding a protein sequence from a graph representation of a target structure. The architecture is divided into an encoder and a decoder. The encoder is responsible to extracting features from the 3D structures using multi-head self-attention on the graph. Only the k-nearest neighbors of a node are considered. Then a decoder

0) This work was supported by the Institute for Basic Science in South Korea [IBS-R029-C2, IBS-R030-C1].

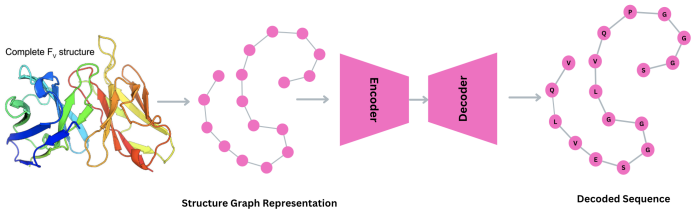

그림 1: Framework overview of graph-based deep learning methods for protein/antibody sequence design.

predicts the protein sequence autoregressively using causal self-attention, i.e. taking into account nodes that were already decoded. In [1] each node is related to an amino acid in the protein sequence. The node features includes the amino acid identity and dihedral angles of the protein backbone. The edge features are composed by the distance and orientations between two residues. Also, as edge features, hydrogen bonds and contact information are included.

2.2 ProteinMPNN

ProteinMPNN [5] uses a message passing neural network (MPNN) with an encoder-decoder architecture, similarly to [1], to predict the protein sequence. In addition to decoding the sequence in from N to C terminus (first-to-last amino acid in sequence) in a sequential manner, an order agnostic autoregressive model in which the decoding order is randomly sampled was proposed. This feature gives additional flexibility to the model, including the possibility to design protein complexes. A main difference between ProteinMPNN and the method proposed by Ingraham et al [1] is observed in the modeling of node features and edge features. ProteinMPNN embed edges but do not include any node features. The edge features includes the distances between atoms of the protein backbone, distances between amino acids in the same chain calculated in the sequence space (relative positional encoding), and an binary feature indicating if residues are from the same or different chains.

3 Methodology

3.1 ProteinMPNN Model

In our experiments, we use the ProteinMPNN model available in [5]. The model architecture consists of 3 encoder and 3 decoder layers with hidden dimensions set to 128. The dataset used for training the model consists of 19700 single-chain protein structures which are chosen and split based on the CATH dataset [10]. It is noted that CATH includes both both general proteins and antibodies. Even ProteinMPNN includes order agnostic decoding, in our experiments we decode the antibody structure from the beginning of the heavy chain to the end of the light chain using the sampling temperature set to 0.1.

3.2 Antibody-specific Model

3.2.1 Antibody Data

For training the proposed antibody-specific model, we need a dataset which contains a diverse set of antibody structures. In this work, the SAbDAb [9] is used. SabDab currently includes approximately 6685 antibody structures that are available in the Protein Data Bank (PDB) [11]. Here, our focus in only in generating the sequence for the variable regions of the antibody. With that in mind, the IMGT server [12] is used to obtain parts of the antibody structure related to its variable region (heavy and light chains). After pre-processing the data, our final dataset size consists of 6299 antibodies.

3.2.2 Ing_Ab Model

The architecture for training our antibody-specific model is similar to the one proposed in Ingraham et al [1]. For training the dataset was split into train (95%), validation (2.5%), and test sets (2.5%). As Ingraham et al can only decode single chain proteins, we concatenated heavy and light chains as a single chain. After performing a preliminary hyperparameter search, we train our model for 100 epochs with the batch size set to 6000. Similarly to the ProteinMPNN model, the sampling temperature is set to 0.1.

4 Results

To evaluate ProteinMPNN and Ing_Ab we prepared a reference dataset consisting of 112 antibody structures, that were added to SabDab after the training of our model. For each model, ProteinMPNN and Ing_Ab, we generated 50 amino acid sequences for each antibody structure in the evaluation set. Each of the generated sequences was added as an input to IgFold for antibody structure prediction. In this way, we have predicted 112x50 structures for each model. The main metric used for comparison is the RMSD value between the predicted structure and the native antibody structure. The structures are aligned and the distances calculated using Cealigner. Additionally, we add the predicted RMSD (pRMSD) that is the output of IgFold that indicates the confidence of the networks in its predictions for further analysis.

The results obtained for sequences generated by ProteinMPNN and Ing_Ab are shown in Table 1. It is shown that the structures

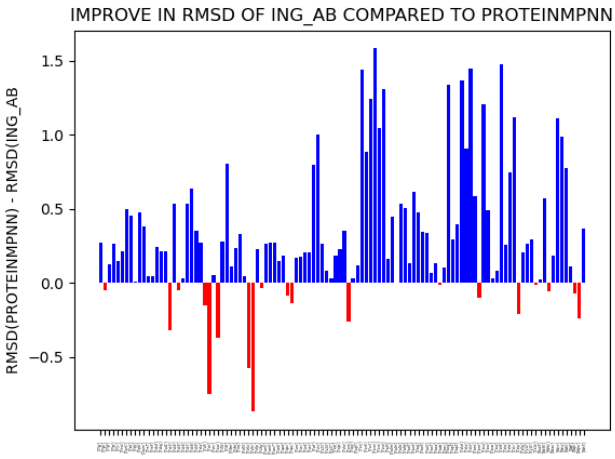

그림 2: Comparison between Ing_Ab and ProteinMPNN for each structure in the evaluation set (x-axis). A positive number (blue) indicate structures in which Ing_Ab outperforms ProteinMPNN.

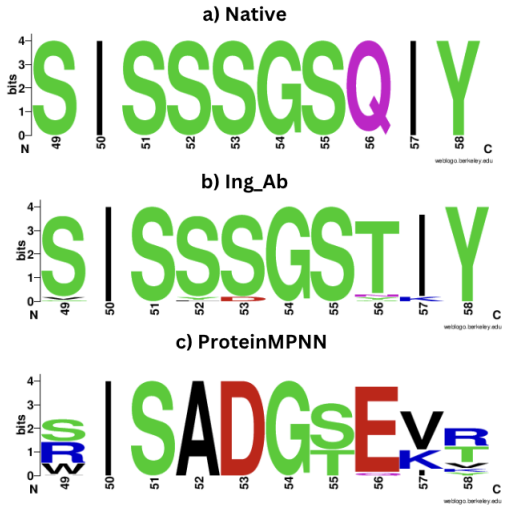

그림 3: Amino acid distribution for the CDR-H2 of one of the antibodies in the evaluation set (PDB=7v24). It is seen that Ing_Ab learns a distribution closer with the native antibody sequence.

predicted for sequences generated by Ing_Ab outperforms ProteinMPNN in all metrics evaluated. In general, the structures predicted by IgFold for sequences generated by Ing_Ab achieves lower mean RMSD and minimum RMSD when compared to the native structure. Additionally, the pRMSD metric, indicating the confidence of IgFold in the antibody structure prediction, is also lower for Ing_Ab when compared to ProteinMPNN.

표 1: Comparison of RMSD and confidence metrics with structures predicted by IgFold for the sequences generated with ProteinMPNN and Ing_Ab.

Model	RMSD_IgF (mean)	RMSD_IgF (min)	pRMSD
ProteinMPNN	1.9156	0.7252	0.8333
Ing_Ab (ours)	1.5866	0.5624	0.6977

Figure 2 show how the mean RMSD is improved using Ing_Ab for each test sequence. The improvement in RMSD, calculated by subtracting the mean RMSD obtained by ProteinMPNN and the mean RMSD obtained by Ing_Ab is shown. It can be observed in Fig. 2 that Ing_Ab achieves lower mean RMSD for the majority of the antibody structures evaluated. In Fig. 3 we also evaluate the amino acid distribution generated by Ing_Ab and ProteinMPNN for one of the antibodies in the evaluation set. It is seen that the sequence generated by Ing_Ab is closer to the native sequence, differing with only one residue if sampling is performed in a greedy fashion. We suggest that training the network with antibody-only data may lead the generation of sequences with closer characteristics to native antibodies when compared to ProteinMPNN that is trained with general protein datasets.

5 Conclusion

In this paper we train an antibody-specific graph-based deep learning model for sequence design. The proposed model is compared with a state-of-the-art protein sequence design model and the results show that training antibody-specific models leads to lower RMSD with the native structure for the majority of test sequences evaluated. As a future work, we plan to investigate different loss functions that emphasize the importance of variable regions of antibody structures.

References

[1] Ingraham, John, et al. "Generative models for graph-based protein design." Advances in neural information processing systems 32 (2019).

[2] Strokach, Alexey, et al. "Fast and flexible protein design using deep graph neural networks." Cell systems 11.4 (2020): 402-411.

[3] Jing, Bowen, et al. "Learning from protein structure with geometric vector perceptrons." arXiv preprint arXiv:2009.01411 (2020).

[4] Hsu, Chloe, et al. "Learning inverse folding from millions of predicted structures." bioRxiv (2022).

[5] Dauparas, Justas, et al. "Robust deep learning–based protein sequence design using ProteinMPNN." Science (2022): eadd2187.

[6] Battaglia, Peter W., et al. "Relational inductive biases, deep learning, and graph networks." arXiv preprint arXiv:1806.01261 (2018).

[7] Ruffolo, Jeffrey A., and Jeffrey J. Gray. "Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies." Biophysical Journal 121.3 (2022): 155a-156a.

[8] Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021): 583-589.

[9] Dunbar, James, et al. "SAbDab: the structural antibody database." Nucleic acids research 42.D1 (2014): D1140-D1146.

[10] Pearl, Frances MG, et al. "The CATH database: an extended protein family resource for structural and functional genomics." Nucleic acids research 31.1 (2003): 452-455.

[11] Bank, Protein Data. "Protein data bank." Nature New Biol 233 (1971): 223.

[12] Ruiz, Manuel, et al. "IMGT, the international ImMunoGeneTics database." Nucleic acids research 28.1 (2000): 219-221.