Applications of AlphaFold beyond Protein Structure Prediction

Yuan Zhang[1], Peizhao Li[2], Feng Pan[1], Hongfu Liu[2], Pengyu Hong[2], Xiuwen Liu[3], Jinfeng Zhang[1,*]

[1]Department of Statistics, Florida State University, Tallahassee, FL 32306

[2]Department of Computer Science, Brandeis University, Waltham, MA 02453

[3]Department of Computer Science, Florida State University, Tallahassee, FL 32306

[*]Contact: Jinfeng@stat.fsu.edu

Abstract

Solving the half-century old protein structure prediction problem by DeepMind's AlphaFold is certainly one of the greatest breakthroughs in biology in the twenty first century. This breakthrough paved the way for tackling some previously highly challenging or even infeasible problems in structural biology. In this study, we propose strategies to use AlphaFold to address several fundamental problems: (1) protein engineering by predicting the experimentally measured stability changes using the representations extracted from AlphaFold models; (2) estimating the designability of a given protein structure by combining a protein design method (e.g. ProDCoNN), sequential Monte Carlo, and AlphaFold. The designability of a protein structure is defined as the number of sequences that encode that protein structure.; (3) predicting protein stabilities using natural sequences and designed sequences as training data, and representations extracted from AlphaFold models as input features; and (4) understanding the sequence-structure relationship of proteins by computational mutagenesis and testing the foldability of the mutants by AlphaFold. We found the representations extracted from AlphaFold models can be used to predict the experimentally measured stability changes accurately. For the first time, we have estimated the designability for a few real proteins. For example, the designability of chain A of FLT3 ligand (PDB ID: 1ETE) with 134 residues was estimated as 3.12±2.14E85.

Introduction

Protein structure prediction (PSP) has been one of the most challenging problems in computational biology[1]. It is also a problem whose solution will have a profound impact in many areas of biology and biomedical sciences. Not surprisingly, the problem has attracted researchers from many different disciplines for half a century since it was originally proposed[2]. Critical Assessment of Structure Prediction (CASP) was initiated in 1994 to provide a blind test of methods for PSP[3,4], which has played a key role for advancing PSP methods since then. Except the early years of CASP with some substantial progress, the field had come to a standstill for quite some years, until in 2018 when Google's DeepMind joined the game with their deep learning powered method, AlphaFold, which substantially improved the prediction accuracy in CASP13[5]. And in CASP14 held in 2020, AlphaFold2[6] (we will call it AlphaFold in the rest of the paper) has pushed the numbers so much so that the organizers of CASP declared that the PSP problem has been finally solved[4]. Since then, DeepMind team has applied AlphaFold to predict more than 350,000 protein structures in human and other species, and released the structures for biological community to use freely[7]. These predicted protein structures will help biologists understand the functions of these proteins and the mechanisms of those biological processes or diseases, in which these proteins are involved. In a follow-up study[8], and also by Baker and co-workers who developed RoseTTAFold using a deep learning framework inspired by AlphaFold[9], it has been shown that the protein-protein interaction problem can be considered as a protein structure prediction problem by putting two or more protein chains together to predict their complex structure.

High accuracy protein structure prediction enabled new research strategies to tackle previously highly challenging or infeasible problems. Since the release of AlphaFold, our group has explored some potential applications by using it alone or combined with other methods. In this paper, we report our preliminary results and propose some novel strategies, enabled by AlphaFold, to address some fundamental problems in structural biology.

We first looked at the modeling of point mutations, which was generally considered as not suitable for AlphaFold. Indeed, our first finding was that the output confidence scores of AlphaFold correlates poorly with the experimentally measured stability changes. However, further investigation showed that the representations extracted from AlphaFold models can be used to predict stability changes very accurately. We obtained the state-of-the-art performance using a relatively simple model with a much smaller training dataset compared to existing methods.

We then used AlphaFold to predict the structures for the sequences designed to fold to target structures using a protein design method we developed recently, ProDCoNN[10]. ProDCoNN used a deep neural network architecture to model the local three dimensional environment of individual residues. It achieved the best performance for the inverse protein folding (IPF) problem tested on benchmark datasets. We found that some designed sequences can fold to structures quite close (<4Å RMSD) to the target structures while others cannot (as predicted by AlphaFold). AlphaFold can thus be used to select promising sequences designed for an IPF problem. We then propose a new framework by combining AlphaFold, ProDCoNN, and sequential Monte Carlo (SMC) to estimate the designability of a given protein structure. The designability of a given protein structure is defined as the number of sequences that encode the structure[11,12]. A sequence encodes a structure if it can fold to a structure very close to that structure. Here we can use RMSD to measure the similarity between two structures and select a reasonable cutoff value to define designability. For the first time, we have estimated the designability of a real protein structure, chain A of FLT3 ligand (PDB ID: 1ETE) with 134 residues, as 3.12±2.14E85.

In many applications, especially protein engineering and design, it is important to know the stability of the designed protein. However, predicting stability for any given sequence is still a question not answered by AlphaFold. We believe this is a major weakness of AlphaFold since it cannot guarantee that the predicted structure can actually fold stably. We propose a strategy that can take natural protein sequences (positive cases) and design sequence (negative cases) as training data, and use the representations extracted from AlphaFold models as input features to build machine learning models to predict whether a sequence has stability comparable to natural proteins.

Finally, we proposed that we can study the sequence-structure relationship of proteins by performing computational mutagenesis and testing the foldability of the mutants using AlphaFold. This may help identify the minimum elements for folding (MEF) for a protein structure, where MEF for a protein structure is defined as the minimum requirement at each residue position for a sequence to fold to the structure. Such studies may shed light on the fundamental principles of protein folding.

Method and Data

AlphaFold and RoseTTAFold programs were obtained from their GitHub release.

Point mutations with experimentally measured stability changes

To study the correlation between AlphaFold confidence scores and the experimentally measured stability changes, we randomly selected 3507 experiments from protein single-point mutants stability database FireProtDB[13], corresponding to 1251 mutants from 86 protein chains. The dataset contains 2557 experiments with Gibbs free energy changes (ΔΔG) upon mutation and 952 experiments with changes in melting temperatures (ΔTm). There are 328 stabilized single-point mutant experiments, 1842 destabilized mutants, and 1337 neutral ones. The stabilization status follows the definition in FireProtDB, such as stabilizing mutations ($\Delta Tm > 1$ or $\Delta\Delta G < 1$ kcal/mol), destabilizing ($\Delta Tm < 1$ or $\Delta\Delta G > 1$ kcal/mol), and neutral ($-1 \leq \Delta Tm \leq 1$ or $-1 \leq \Delta\Delta G \leq 1$ kcal/mol).

To train a model for predicting point mutation stability changes, we randomly select 7777 experiments from FirProtDB with a valid ΔΔG, corresponding to 2854 mutants from 114 protein chains. There are 499 stabilized single-point mutant experiments, 3653 destabilized mutants, and 3625 neutral ones. We calculate the median of ΔΔG if multiple experiments are taken for each mutant sequence, which results in 2854 data points, with 149 stabilized mutant, 1311 destabilized mutant, and 1394 neutral ones. The dataset is separated into 10 folds for 10-fold cross-validation. The separation is residue-based and guaranteed that two mutants at the equivalent site from two homologous proteins were always in the same fold. The homologous proteins are defined as sequence identity higher than 25%, which is calculated by using T-coffee[14].

Designed sequences for inverse protein folding problem

We used a modified ProDCoNN[10] to design sequences for 9 protein structures selected from SCOPe database[15] which belong to 7 major classes defined by SCOPe. Specifically, two structures are from the *all alpha proteins* class (3lynA, 1e2aA), two structures from *all beta proteins* (1g6vK, 1kexA), one structure from each of the *alpha and beta (a/b)* (1h4yA), *alpha and beta(a+b)* (1a2pA), *multi-domain* (2avuF), *membrane and cell surface* (1g4yB), and *coiled coil proteins* (1ujwB) classes. The lengths of the proteins range from 76 to 156 residues. All the selected protein structures from PDB are single chains without missing residues and uncommon amino acids. No binding ligands were included.

The original ProDCoNN[10] was not suitable for sampling protein sequences from the space of all possible sequences following certain probability distributions. To sample a sequence, we used a sequential Monte Carlo approach[16-23] by sampling one residue at a time. A new residue is sampled conditioning on the residues sampled before it. This required us to train a partial protein design model based on ProDCoNN. The partial model takes a backbone structure with the types of some of the residues already sampled and sample one additional residue conditioning on the already sampled residues. The sampling order (which residue should be sampled the first, which residue the second, etc.) is decided based on the calculated entropy along the sequence:

$$\mathrm{Entropy(x)} = -\sum_{\mathrm{i}} \mathrm{p_i} \ln \mathrm{p_i},$$

where $\mathrm{p_i}$ is the probability that residue *x* is predicted as amino acid type *i*. And the sampling probability is normalized by

$$P_x \sim e^{-Entropy(x)/T_1},$$

which includes a parameter temperature $\mathrm{T_1}$. The residues with lower entropy have a higher chance to be sampled first, while different sampling orders could be generated by different sampling iterations.

When the sampling order is decided, the types of amino acids at each position will be sampled based on the predicted probabilities of the twenty amino acids by the trained partial model. To adjust the relative probabilities of the types with non-zero probabilities, we used the following normalized probability:

$$P_i \sim e^{p_i/T_2} - 1.$$

In our sampling, the parameter T_1 was set to 1 and T_2 was set to 0.05. The goals of tuning T_1 and T_2 are to have enough diversification in sampled sequences while also make sure a significant of the sequences are foldable. The similarities between the sampled sequences and the wild type sequences range from 15.8% to 36%.

Estimating the designability of a protein structure

To estimate the designability of a protein structure, we first used the SMC strategy described above to sample a number of sequences for the protein structure. We then used AlphaFold to predict the structures of these sequences. Those sequences with predicted structures with RMSD smaller than 4 Å to the target structure are considered as foldable and used for estimating the designability. The other sequences are discarded. Each sampled sequence using SMC has a weight, which is updated recursively as follows: $w_t = w_{t-1}/p_t$, where w_t is the weight at step t, w_{t-1} is the weight at step t-1 and p_t is the probability the actual amino acid type at step t is sampled. The designability can then be estimated using the equation: $D = \frac{1}{n}\sum_i^n w_i$, where D is designability, w_i is the weight of sequence i, and n is the total number of foldable sequences.

Deep learning model for predicting point mutation stability changes

We implemented a multilayer perceptron regression model for ΔΔG prediction using features extracted from the representations of AlphaFold models as input.

We first used AlphaFold to predict the structures of both wild type and mutant sequences. We then extracted the feature vectors from the position of the mutated residue from the “single representation” of the AlphaFold models for both wild type and mutant sequences as model input. We additionally include the subtraction of the two vectors from wild type and mutant as the input feature. As the dimension for each residue for the single representation is 384, the dimension of the final input feature vector is 3x384 = 1,152.

We built the model with four linear projections with input and output feature dimensions (1,152, 1,152), (1,152, 512), (512, 512), and (512, 1), and used the output in the last layer as the ΔΔG value. Non-linear activation function ReLU was inserted between the linear projections. We used Adam optimizer and set the batch size as 1,024, learning rate as 1e-4, and training epochs as 1,000. The model was trained with Smooth L1 Loss.

Definition of foldable sequences

In this study, we define a foldable sequence to a given structure as one that can fold to a structure with a RMSD smaller than 4Å to the given structure. Since we use AlphaFold to predict the structures of designed sequences, the foldability is predicted, not experimentally measured. In fact, the designed sequences, even predicted as foldable, may not have stabilities similar to real proteins. Our assumption here is that even a designed sequence is not actually foldable, there is a sequence in the neighborhood of it, which is actually foldable. Here the neighborhood of a sequence is defined as sequences with small number of mutations (i.e. smaller than 5) from the given sequence.

Results

Predicting stability changes of point mutations

We first tested AlphaFold on the point mutation dataset to see whether there is any relationship between the confidence scores it outputs and the experimentally measured stability changes. For a protein, *p*, we used AlphaFold to predict structures for both its wild type sequence and single-point mutant to generate two structures $S_{p,w}$ and $S_{p,m}$, respectively.

We examined the correlation between the stability changes and confidence scores output by AlphaFold. There are two different measures of stability changes for the mutants from FireProtDB database, ΔΔG and ΔT_m. There are also two different confidence scores (CSs) for the predicted structures, predicted TM-score, pTM, and predicted LDDT value for the mutated residue, pLDDT. It is more meaningful to look at the changes in pTM and pLDDT between $S_{p,w}$ and $S_{p,m}$. We plotted the scatter plots for all four possible pairs between stability changes and CS changes upon mutation (Fig 1). None of the pairs showed any significant correlations.

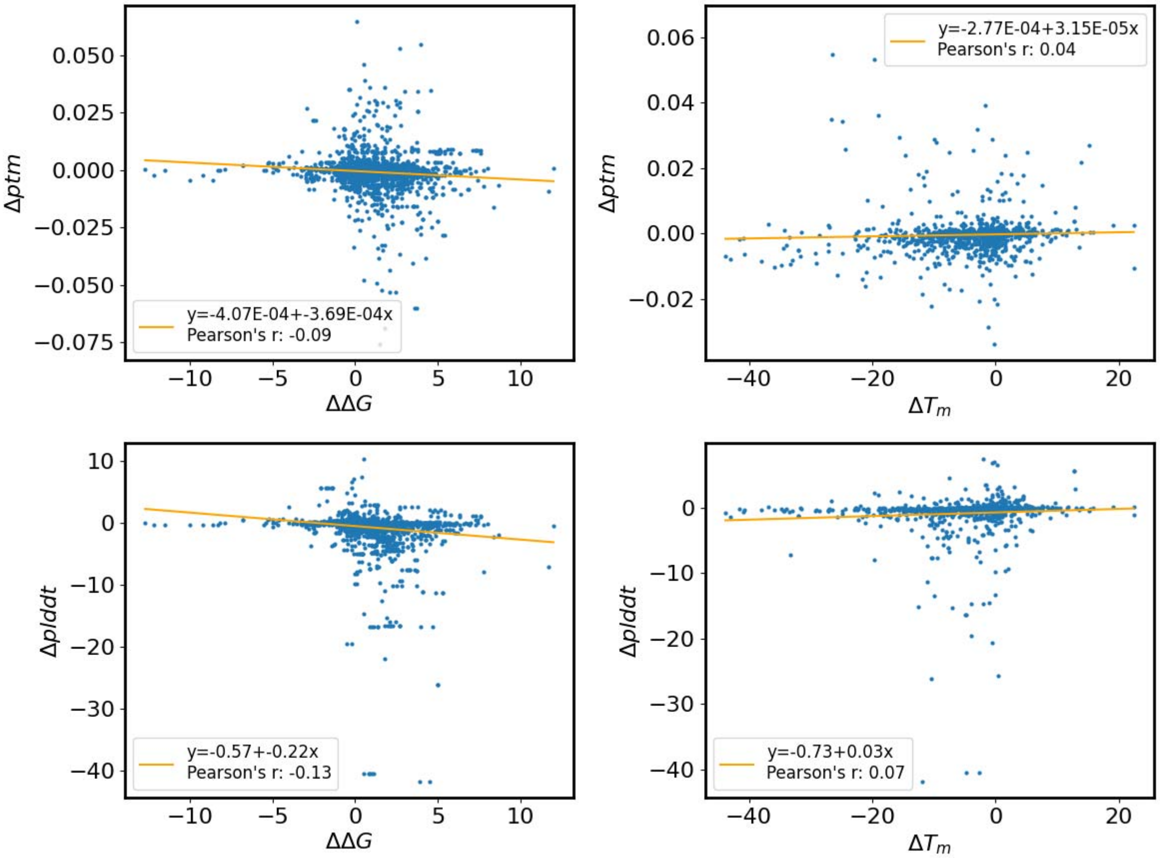

Fig 1: Difference of confidence score from AlphaFold (top row: whole sequence score ΔpTM; bottom row: residue wise score ΔpLDDT) vs. Gibbs free energy changes upon mutation ΔΔG (left column) and changes in melting temperatures ΔT_m (right column)

We then extracted some features from the "single representations" of AlphaFold models and built a simple neural network model to predict ΔΔG. We used 10-fold cross validation and the Pearson's correlation coefficient is 0.58, which is significantly higher than those observed between the confidence scores and stability changes. It is also slightly higher than the state-of-the-art performance achieved by a recent deep learning method[24]. Fig 2 shows the scatter plot between ΔΔG and the predicted ΔΔG from a 10-fold cross-validation.

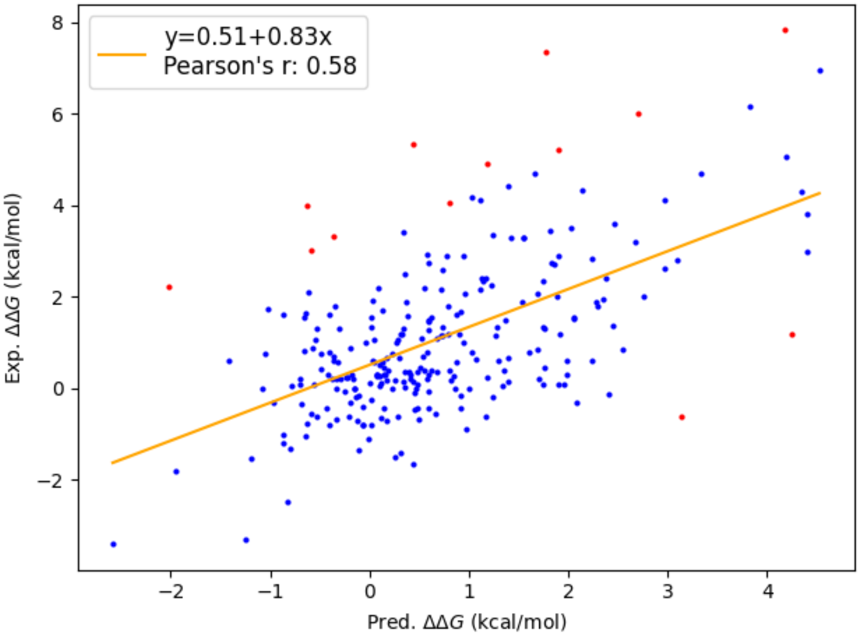

Fig 2. The scatter plot for the experimentally measured stability changes, ΔΔG, vs predicted ΔΔG for point mutations. The red dots are outliers.

Predicting structures for sequences designed for inverse protein folding (IPF)

It is understandable that the predicted structures for point mutations by AlphaFold are always very close to the wild type structures since that is the only pattern it can learn from PDB. What about new sequences that significantly differ from any of the natural sequences? We next used AlphaFold to predict structures for sequences designed to fold to certain target protein structures. Designing sequences that fold to a given protein structure is also called the inverse protein folding (IPF) problem. We selected several proteins from different fold classes from SCOPe database. The sequences were designed using a modified ProDCoNN[10] (see Method and Data for details). The similarities between designed sequences and the corresponding wild type sequences range between 16-36%. Fig 3 shows the predictions from AlphaFold for 4 different proteins. We can see that a significant number of the designed sequences are predicted to fold to relatively small RMSDs compared to the target structures. Our result indicates that AlphaFold can be used to select promising foldable sequences and filter out sequences that are not foldable.

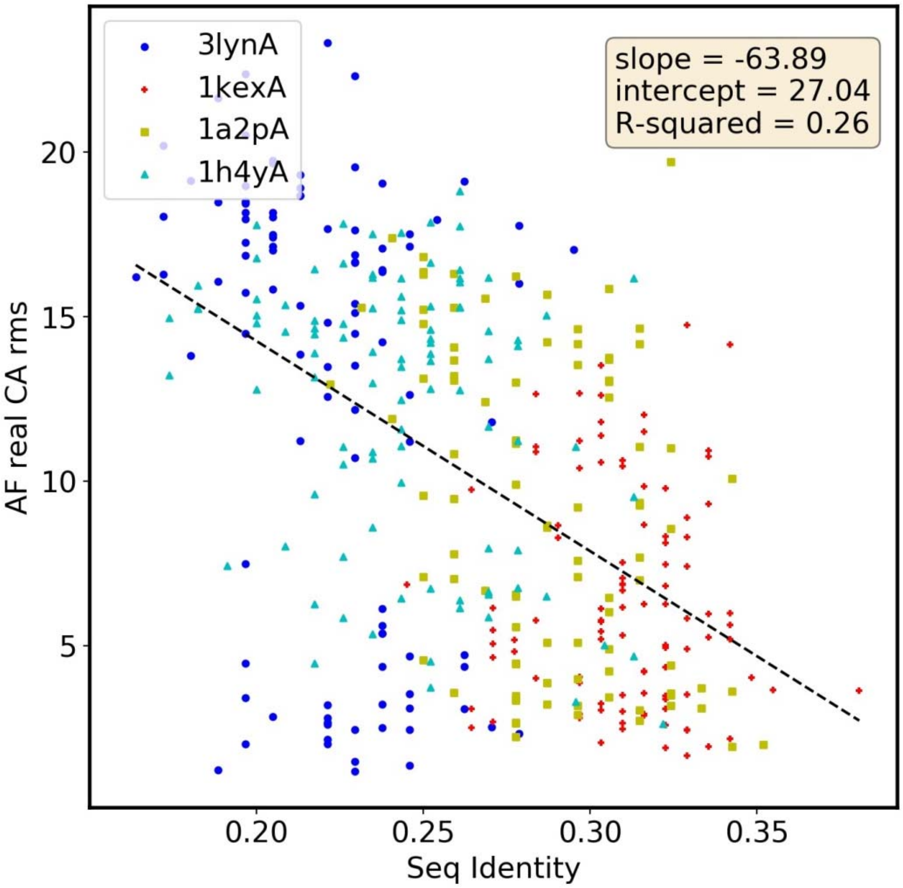

Fig 3. The RMSD of AlphaFold predicted structures to the target structures vs. the sequence identities of the designed sequences for 4 proteins.

Characterizing foldable sequences

For designed sequences, some are predicted to be foldable, while others are not. Studying the foldable sequences may reveal the key residues important for the folding and stability of a protein structure. Fig 4 showed two sequence logos, one plotted using foldable and one using homologous sequences obtained from multiple sequence alignment (MSA) for protein 1a2p. Clearly, despite significant similarity in many positions, the foldable sequences showed marked differences from homologous sequences at certain positions. This indicates that those conserved residues among the homologous sequences may be important for functions instead of stability. The comparison may shed light on potential point mutations that may increase the stability of the protein. It is worth noting that the MSA of 1a2p has much stronger conservations than the foldable sequences. The average entropy for MSA and foldable designed sequences are 0.813 and 1.367, respectively. The numbers of residues with entropy smaller than 1 for MSA and foldable sequences are 65 and 29, respectively. That indicates that the natural sequences may have only explored part of the foldable sequence space for this protein structure.

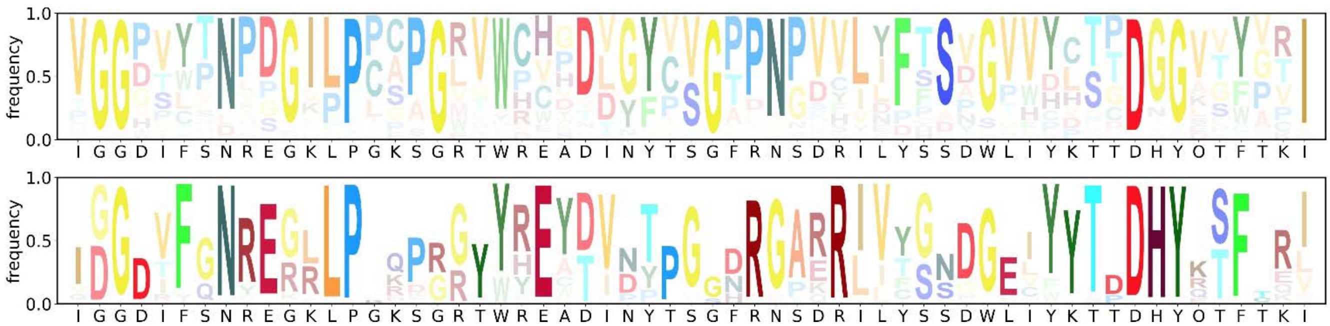

Fig 4. The logo plots for foldable designed sequences and multiple sequence alignment (MSA) of chain A of protein 1a2p. The part of the sequence with good alignment is shown (residues 49-107 out of total 108 residues). The figure with all the residues is provided in Supplementary material. The MSA of 1a2p has much stronger conservations than the foldable sequences. The average entropy for MSA and foldable designed sequences are 0.813 and 1.367, respectively. The numbers of residues with entropy smaller than 1 for MSA and foldable sequences are 65 and 29, respectively. Top panel: The logo plot for the foldable sequences whose predicted structures are within 3Å to the target structure. Bottom panel: The logo plot from multiple sequence alignment of 1a2p.

Conclusion and Discussion

In this study, we used engineered and designed sequences to investigate the applicability of AlphaFold to problems other than structure prediction for naturally occurring sequences. We used mainly two types of sequences: point mutations with experimentally measured stability changes[13]; and sequences designed to fold to target protein structures using a modified algorithm based on ProDCoNN[10]. We found that the representations extracted from AlphaFold models can accurately predict the stability changes of point mutations. We also found that AlphaFold predicted the ProDCoNN designed sequences with a wide range of RMSDs to the target structures, indicating that some are more foldable than others. Finally, comparing the foldable sequences for a target protein with its homologous sequences from a multiple sequence alignment showed significant differences between the two profiles. Studying such differences may shed light on the role of the conserved residues in the two profiles. Based on the findings in this study, we propose four fundamental questions that can be immediately addressed with the help of AlphaFold when combined with other previously developed methods.

Protein engineering

Although the stability changes of point mutations cannot be directly inferred from the confidence scores of AlphaFold predictions, we found that the representations AlphaFold learned during the prediction process can be used to predict the stability change accurately. Several improvements may significantly increase the prediction accuracy. Firstly, the dataset we used can be substantially increased by using the data in a recent study[24], which used more than 5000 point mutations. With more than doubled training data, we expect the model to have substantially improved performance; second, the “pair representation” generated during AlphaFold prediction should also be very useful for predicting stability changes. Of course, the dimensionality of the input also increased significantly, which may need to be regularized; third, in this study, we only took the information of the mutated residue from the single representation. Information of other residues will be very helpful to further improve the prediction performance. For example, we can take a fix number of residues from the sequence neighbors or spatial neighbors of the mutated residue. These improvements are being investigated by us currently.

Estimating the designability of protein structures

Natural proteins have the capability to withstand a wide range of environmental stress and mutations. As seen from the point mutation data, a large number of mutated sequences of a protein can also fold to its native structure. The larger number of mutations a protein can tolerate, the more robust the protein structure and function is. The "designability" of a protein structure, defined as the number of sequences that encode that structure, has been proposed as an important property that contributes to the functional robustness of proteins[11]. Since protein structures can be organized as a hierarchy[15,25-27] with four levels from folds to super families, to families, and to sequences, the designability of a protein fold is similarly defined as the number of families that take the fold as their native structures. A study has found that many disease-related proteins have folds with relatively few families, and a number of these proteins are associated with diseases occurring at high frequency[12]. This indicates that there is indeed a correlation between designability and functional robustness of proteins.

However, simply looking at the number of families under each fold is not a reliable measure for the fold's designability and its functional robustness, because it has been found that the age of a fold correlates with its "usage" among natural proteins. For instance, eukaryotic folds found only in human, mouse, and yeast contain approximately 2.5 families, on average, compared to an average of 13.8 families per fold for all human proteins[12]. This observation has two implications: firstly, since the number of sequences exists in nature that can fold to a particular protein structure is not necessary a good indicator of its designability, we need to estimate the designability of a protein structure to have a better understanding of its functional robustness; second, since new folds have been much less "explored" by nature, there must exist new families, not related to any families found previously, that can fold to one of the newer folds. These new protein families may be hosts for some interesting, new functions. It is now possible to design new sequences using a program such as ProDCoNN to specifically target on uncharted regions in the foldable sequence spaces and test the design with AlphaFold.

Recently, we have formulated a framework for estimating the designability of a protein structure by combining a protein design algorithm, sequential Monte Carlo (SMC), and AlphaFold[28]. To estimate designability, we need to sample foldable sequences using SMC. SMC is a special type of Monte Carlo method that allows one to estimate the partition function of a system[23], which is usually very challenging to estimate. We have applied SMC in the past to estimate the entropy of lattice polymers[20], the side chain entropy of proteins[19], and other ensemble properties[18,21,29]. The total number of foldable sequences of a given protein structure can be expressed as a partition function, which can be estimated by SMC. In a preliminary study, for the first time, we have estimated the designability of a real protein structure, chain A of FLT3 ligand (PDB ID: 1ETE) with 134 residues, as 3.12±2.14E85[28].

Predicting protein stability using natural and design sequences

As shown earlier, the representations learned by AlphaFold can be used to predict stability changes of point mutations. However, for a designed sequence by a program such as ProDCoNN, we still don't know whether the structure predicted by AlphaFold is stable enough. Stability prediction for any given sequence is one of the key questions in protein folding unanswered by AlphaFold. Although predicting the exact stability can be quite challenging, it may be feasible to predict binary outcomes, such as whether a sequence has the stability as that of natural proteins. To address this with machine learning methods, we need to have stable sequences and unstable sequences. The protein sequences in PDB structures can serve as stable sequences. To obtain unstable sequences, one can randomly sample sequences, but these sequences are not challenging enough to train quality models to distinguish between foldable and unfoldable designed sequences. One reasonable option is to use all the designed sequences as negative sequences or use the designed sequences that are predicted to have RMSD to the target structure greater than certain threshold. We hypothesize that it may work fine if we use all the designed sequences as negative sequences because it is unlikely a designed sequence will have stability comparable to natural proteins. Even some designed sequence do have stability comparable to nature sequences, the number of such sequences should be quite small, which will not impose serious issues for training high quality models. With the sequence decoys and natural protein sequences, we can then train a model to perform a binary prediction: whether a sequence has stability comparable to natural proteins. When constructing the predictive models, we can again use the representations extracted from AlphaFold models as the input.

In protein design practice, we can select the sequences with small RMSDs to the target structure and optimize them by computational mutagenesis using a model for predicting stability changes for point mutations (e.g. the model we developed in this study). The stability of the optimized sequences can then be predicted by the binary model to check whether their stabilities are good enough. This provides a practical pipeline for designing sequences to meet the requirement of real applications.

Using AlphaFold to perform computational mutagenesis to understand the sequence-structure relationship of proteins.

By studying the foldable sequences, we may identify residues that are important for the folding and stability of a protein structure and gain a deeper understanding of the sequence-structure relationship of proteins. This is a fundamental problem in structure biology. To help the discussion, we call the set of key residues for the folding of a protein as the minimum elements for folding (MEF). Using multiple sequence alignment (MSA) may not be ideal for detecting MEF because some residues may be conserved for binding or dynamics. In addition, the available sequences obtained from a MSA may be a limited or biased subset of all foldable sequences for a protein structure as shown in this study that natural proteins may not have explored the whole foldable sequence space. Starting from foldable sequences, one can perform computational mutagenesis to search for MEF for a protein structure. Each new sequence can be tested by AlphaFold for its foldability to keep updating the sets of foldable sequences. The search may eventually converge to certain sequence pattern satisfying the requirement of a MEF.

Acknowledgement

JZ is supported partially by a grant from National Institute of General Medical Science of National Institutes of Health, grant # R01GM126558. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1 Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol* **20**, 681-697 (2019).

2 Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223-230 (1973).

3 Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* **15**, 285-289 (2005).

4 Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* (2021).

5 Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-710 (2020).

6 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

7 Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590-596 (2021).

8 Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.2010.2004.463034 (2021).

9 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021).

10 Zhang, Y. *et al.* ProDCoNN: Protein design using a convolutional neural network. *Proteins* (2019).

11 Li, H., Helling, R., Tang, C. & Wingreen, N. Emergence of preferred structures in a simple model of protein folding. *Science* **273**, 666-669 (1996).

12 Wong, P. & Frishman, D. Fold designability, distribution, and disease. *PLoS Comput Biol* **2**, e40 (2006).

13 Stourac, J. *et al.* FireProtDB: database of manually curated protein stability data. *Nucleic Acids Res* **49**, D319-D324 (2021).

14 Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-217 (2000).

15 Chandonia, J. M., Fox, N. K. & Brenner, S. E. SCOPe: classification of large macromolecular structures in the structural classification of proteins-extended database. *Nucleic Acids Res* **47**, D475-D481 (2019).

16 Tang, K., Wong, S. W., Liu, J. S., Zhang, J. & Liang, J. Conformational sampling and structure prediction of multiple interacting loops in soluble and beta-barrel membrane proteins using multi-loop distance-guided chain-growth Monte Carlo method. *Bioinformatics* **31**, 2646-2652 (2015).

17 Tang, K., Zhang, J. & Liang, J. Fast protein loop sampling and structure prediction using distance-guided sequential chain-growth Monte Carlo method. *PLoS Comput Biol* **10**, e1003539 (2014).

18 Zhang, J., Lin, M., Chen, R., Liang, J. & Liu, J. S. Monte Carlo sampling of near-native structures of proteins with applications. *Proteins* **66**, 61-68 (2007).

19 Zhang, J. & Liu, J. S. On side-chain conformational entropy of proteins. *PLoS Comput Biol* **2**, e168 (2006).

20 Zhang, J., Chen, Y., Chen, R. & Liang, J. Importance of chirality and reduced flexibility of protein side chains: a study with square and tetrahedral lattice models. *J Chem Phys* **121**, 592-603 (2004).

21 Zhang, J., Chen, R., Tang, C. & Liang, J. Origin of scaling behavior of protein packing density: A sequential Monte Carlo study of compact long chain polymers. *J Chem Phys* **118**, 6102-6109 (2003).

22 Liang, J., Zhang, J. & Chen, R. Statistical geometry of packing defects of lattice chain polymer from enumeration and sequential Monte Carlo method. *J. Chem. Phys.* **117**, 3511-3521 (2002).

23 Liu, J. S. & Chen, R. Sequential Monte Carlo methods for dynamic systems. *J. Am. Stat. Assoc.* **93**, 1032-1044 (1998).

24 Cao, H., Wang, J., He, L., Qi, Y. & Zhang, J. Z. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. *J Chem Inf Model* **59**, 1508-1514 (2019).

25 Lo Conte, L. *et al.* SCOP: a structural classification of proteins database. *Nucleic Acids Res* **28**, 257-259 (2000).

26 Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536-540 (1995).

27 Barton, G. J. Scop - Structural Classification of Proteins. *Trends Biochem.Sci.* **19**, 554-555 (1994).

28 Pan, F., Zhang, Y. & Zhang, J. Estimating the designability of protein structures. *Manuscript in preparation* (2021).

29 Liang, J., Zhang, J. & Chen, R. Statistical geometry of packing defects of lattice chain polymer from enumeration and sequential Monte Carlo method. *J Chem Phys* **117**, 3511-3521 (2002).