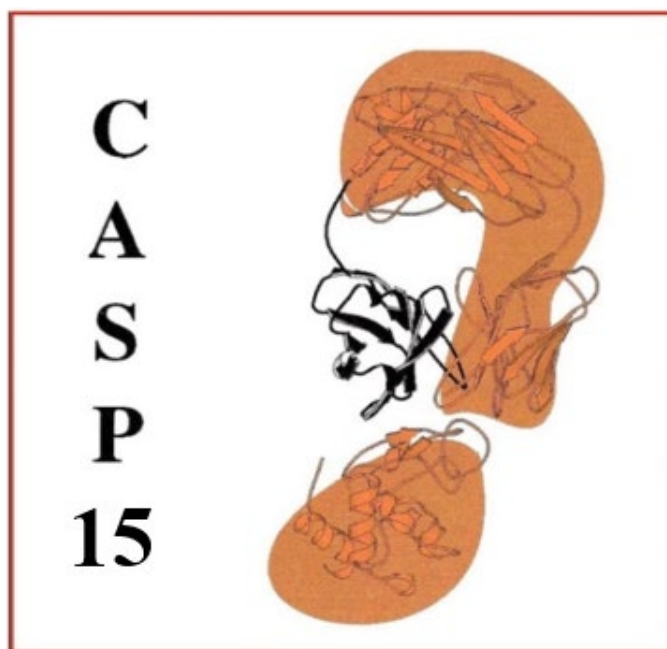


CRITICAL ASSESSMENT OF TECHNIQUES
FOR PROTEIN STRUCTURE PREDICTION



ABSTRACT BOOK

Fifteenth round
May-August 2022

TABLE OF CONTENTS

Agemo	11
Omegafold-based Optimization	11
Agemo_mix	12
AIRFold-Omegafold Ensemble.....	12
Alchemy_LIG	14
Per-atom deviation prediction of protein-ligand binding pose.....	14
with a transformer-based model.....	14
Alchemy_LIG3	16
A hybrid potential energy function for protein-ligand pose scoring and optimization.....	16
Alchemy-RNA	18
Alchemy-RNA2	19
RNA Structure Prediction Using Motif Assembling and Optimization with Statistical Potential	19
APOLLO (QA)	20
Accuracy estimation of individual multimeric protein models using energy-based models	20
BAKER (TS, Assembly, Ligand, RNA)	22
Accurate predictions of protein structures and interactions utilizing RoseTTAFold2.....	22
BAKER-SERVER (TS)	25
Automated protein tertiary structure prediction using RoseTTAFold2	25
BeijingAIProtein	27
OpenComplex: template and docking enhanced protein quaternary structure prediction	27
by geometric deep learning	27
Bench	30
BhageerathH-Pro	31
BhageerathH-Pro: Protein Tertiary Structure Prediction Server.....	31
Bhattacharya	35
Protein modeling and accuracy estimation by Bhattacharya group in CASP15.....	35
CABS-dock	36

Flexible docking of cyclic peptides to proteins using a combination of CABS-dock with FlexPepDock refinement.....	36
Cerebra	38
Cerebra: a convolution-attention-mixed model for protein structure prediction	38
ChaePred (assembly, EMA)	39
Protein Complex Structure Prediction and Scoring Using Machine Learning-based Docking Potential.....	39
ChaePred (TS)	41
Prediction of Protein Tertiary Structure with Accuracy Estimation Using Statistical Potentials and Predicted Distance	41
Chen	43
RNA 3D structure prediction by the hybrid methods	43
ClusPro, Kozakov/Vajda	45
Prediction of protein assemblies and ligand binding modes using a combination of ClusPro and AlphaFold.....	45
CoDock	48
CoDock: Template-based docking and AI-based scoring using in ligand binding prediction..	48
ColabFold	50
Benchmarking ColabFold in CASP15	50
CoMMiT-server	52
Template-based RNA structure prediction guided by deep learning predicted distances	52
Convex-PL; Convex-PL-R; Korp-PL	54
Predicting protein-ligand binding with Convex-PL, Korp-PL, AutoDock Vina, and VinaCPL	54
Coqualia	56
Deep Learning based protein structure prediction model	56
DELCLAB	57
An AI platform oriented to prediction of high-level structures of biomacromolecules.....	57
ddquest	59
Manual trial and error using AlphaFold 2 and conventional ligand docking.....	59
DF_RNA	62
RNA 3D structure prediction by DeepFoldRNA in CASP15	62

DFolding, DFolding-server, DFolding-refine	65
Protein 3D Structure Prediction with DeepFold	65
DLA-Ranker	69
Deep Local Analysis Ranker Server for estimating inter-subunit interfaces accuracy in multimeric complexes	69
DMP	72
DMPfold3 : Minimalist models for end-to-end protein structure prediction	72
Elofsson, NBIS-AF2-standard, NBIS-AF2-multimer	74
Elofsson group using AlphaFold2 and MolPC in CASP15	74
EMBER3D	78
EMBER3D: Fast protein structure prediction for protein mutation movies	78
ESM-single-sequence	80
Atomic resolution protein structure prediction using a language model	80
FALCON2	82
FALCON2: a web server for high-quality prediction of protein tertiary structures	82
FEIGLAB	85
Fernandez-Recio	87
Protein assembly modeling by pyDock: integration of <i>ab initio</i> docking and energy-based scoring of AlphaFold interfaces	87
FoldEver (TS)	89
Protein Structure Prediction based on Adaptive Quality Assessment from Multiple Sequence Alignment.....	89
FoldEver	91
AnglesRefine: refinement of 3D protein structures using Transformer based on torsion angles	91
FoldEver (QA)	93
ZoomScore: residue-level protein complex assessment with machine learning on sequential and 3D structural features	93
FTBiot0119 (assembly)	95
Assembly prediction in CASP15 with <i>ab initio</i> docking and template-based modeling	95
FTBiot0119 (ligand)	96
Protein-ligand complexes docking by using AutoDock Vina and template informations	96

FTBiot0119 (TS)	97
Template-Based Structure Prediction by using dPPAS alignment method	97
GatorsML	98
Predicting conformational diversity in AlphaFold with masked sequence alignments	98
GeneSilico	99
Computational modeling of RNA 3D structures and interactions	99
GinobiFold	100
GinobiFold: MSA-free Superposition Model	100
GinobiFold-SER	102
GinobiFold-SER: Deep Learning based protein structure prediction model	102
Graphen_Medical (ligand)	103
Cross-geometric vector perceptron networks for denoised reconstruction of protein-ligand conjugated conformation.....	103
Graphen_Medical (RNA)	105
Deep reinforcement learning for de novo RNA structure prediction.....	105
Graphen_Medical (TS)	107
Equivariational Graph Attention Learning for Protein Structure Prediction	107
Grudin	109
Protein assembly prediction in CASP15 using a combination of physics-based approaches with AlphaFold2 models	109
GuijunLab-Assembly	110
Multidomain-based protein structure prediction and protein complex structure assembly	110
GuijunLab-DeepDA	112
Protein structures assembly using inter-domain interactions from deep learning	112
GuijunLab-Human	114
Multi-structure refinement using multi-objective optimization and graph neural network-based model quality assessemnt.....	114
GuijunLab-Meta	116
Multiple objective population optimization-based protein structure prediction	116
GuijunLab-RocketX	118
Deep learning-based protein structure prediction and complex model quality assessment....	118

GuijunLab-Threader	120
Protein structure prediction based on enhanced remote homologous template recognition ...	120
GWxraylab	122
SAVARNA: Structure Assembly via Alignment of RNA Secondary Structures	122
HADDOCK	124
HADDOCK scoring of CAPRI Round54 models	124
hFold	126
Protein structure prediction in CASP15	126
through MSA-based HelixFold and MSA-free HelixFold-Single	126
hks1988 (TS)	128
Selection of a good model by single model accuracy estimation method SARTlDdt in CASP15	128
IntFOLD7	130
Automated Prediction of Protein Tertiary Structures with Local Model Quality Scores Using the IntFOLD7 Server.....	130
Kiharalab	134
Integrated structure modeling protocol for human and server prediction for biomolecular structures	134
KORP-PL	138
LAW, MASS (QA)	139
Accuracy estimation of individual multimeric protein models using graph neural network and heterogeneous graph neural network.....	139
LCBio	141
Structure prediction of RNA and RNA complexes using a combination of different modeling methods	141
Manifold, Manifold-E	143
Proteins and Protein Complexes prediction powered by Uni-Fold.....	143
ManiFold-serv	146
An Enhanced protein structure prediction model with integrated domain knowledge and interaction constraints	146
MASS (QA)	148
McGuffin	149

Manual Prediction of Protein Tertiary and Quaternary Structures and Protein-Ligand Interactions.....	149
ModFOLDdock, ModFOLDdockR, ModFOLDdockS	152
Automated Quality Assessment of Protein Quaternary Structure Models using the ModFOLDdock Server.....	152
MUFold, MUFold2	155
Protein Multimer QA with AlphaFold-Multimer and Machine Learning.....	155
MULTICOM_egnn, MULTICOM_deep, MULTICOM_qa (QA).....	157
Multimer Model Quality Assessment Using Gated-Graph Transformer, Steerable Equivariant Graph Neural Networks, and Pairwise Model Similarity	157
MULTICOM_human, MULTICOM, MULTICOM_deep, MULTICOM_qa (Assembly)	160
Improving Assembly Structure Prediction by Sensitive Alignment Sampling, Template Identification, Model Ranking, and Iterative Refinement.....	160
MULTICOM, MULTICOM_human, MULTICOM_egnn, MULTICOM_refine, MULTICOM_deep, MULTICOM_qa (TS).....	163
Improving Tertiary Structure Prediction by Alignment Sampling, Template Identification, Model Ranking, Iterative Refinement, and Protein Interaction-Aware Modeling.....	163
MULTICOM, MULTICOM_qa (LG).....	167
Template-based Modeling for Accurate Prediction of Ligand-Protein Complex Structures in CASP15.....	167
MultiFOLD	170
Automated Prediction, Quality Assessment and Refinement of Tertiary and Quaternary Structure Models using the MultiFOLD Server.....	170
NBIS-AF2-standard, NBIS-AF2-multimer.....	172
Noxelis.....	173
Ligand pose estimation guided by predicted geometrical constraints with templates	173
Oliva.....	176
Machine learning classifiers for protein-protein docking models and the Effect of Training Data Augmentation	176
Oliva.....	179
Introducing an iterative process in a consensus algorithm for the scoring of protein-protein docking models	179

OpenFold	182
OpenFold: A trainable reproduction of AlphaFold2.....	182
Panlab, Pan_Server	185
Improved Template-based Protein Structure Prediction using PBEscore.....	185
PerezLab_Gators	187
Incorporating AF and server derived ambiguous data into MD simulations	187
PEZYFoldings	190
Protein tertiary and quaternary structure prediction using AlphaFold2 with various metagenomic databases.....	190
PICNIC	193
QUIC	193
Refining AlphaFold TS models using 3D residual and convolutional neural networks.....	193
RaptorX	195
RaptorX: protein structure prediction by deep attention network.....	195
rDP	198
<i>Ab initio</i> RNA structure prediction with deep end-to-end potential	198
RNApolis	200
RNAComposer-based modeling of RNA 3D structures in CASP15	200
Rookie	202
Rookie	202
Schug_Lab	203
Selbstaufsicht - Pre-Trained Transformer Models for RNA Contact Prediction.....	203
Seder 2022	205
Seder in CASP15.....	205
server_122 - 126	206
Server122-126: Protein tertiary structure prediction by MEGA-Protein in CASP15.....	206
ShanghaiTech	209
Improved MSA for Better Protein Structure Prediction.....	209
ShanghaiTech-TS-SER	212
ShanghaitechFold: Hybrid MSA Embedder Model	212
SHORTLE	214

Refinement of Protein Models via Fragment Replacement To Improve Local Energies	214
SHT	216
Prediction of tertiary and quaternary structures of biomacromolecules by integrating evolutionary information and energy functions	216
SoutheRNA.....	219
Hierarchically combining multiple methods for 3D RNA structure prediction.....	219
Spider (assembly).....	222
A Novel Statistical Energy Function and Effective Conformational Search Strategy based Protein Complex Structure Prediction.....	222
Spider (TS)	224
A Novel Statistical Energy Function and Effective Conformational Search Strategy based <i>ab initio</i> Protein Structure Prediction.....	224
SUN_Tsinghua	227
Ab Initio Protein Structure Prediction by Conditioned Self-Avoiding Walk and Monte Carlo Tree Search.....	227
Takeda-Shitaka_Lab	229
Structure Prediction for CASP15 Assembly Targets	229
TRFold, TRComplex.....	231
Deep Learning based Protein and Complex Structure Prediction.....	231
UltraFold	234
OpenComplex-RNA predicts RNA 3D structure at the atomic level	234
UM-TBM	237
Integrating multi-MSA, threading templates and deep learning for protein structure prediction	237
UNRES	241
Protein structure prediction in CASP14 with the coarse-grained UNRES model	241
Venclovas, VorolF, VoromQA-2020	243
Modeling and Scoring Protein Assemblies in CASP15	243
Wallner_TS	247
AlphaFold with improved sampling.....	247
WL_team.....	249
Protein Complex Structure Prediction by Multiple Strategies in CASP 15.....	249

Yang, Yang-Server, Yang-Multimer, bench	252
Protein and RNA structure prediction with trRosettaX2, trRosettaRNA and AlphaFold2.....	252
Zheng.....	254
Multi-MSA strategy for protein complex structure modeling	254
Zou_lab.....	257
A template-guiding and docking strategy for protein-ligand binding mode prediction in CASP15.....	257

CASP-RELATED PUBLICATIONS FROM NON-PARTICIPANTS.....259

The ResiRole server provides automated assessments of structure models presented in CAMEO using functional site predictions and has demonstrated applicability to CASP14 SARS-2-CoV Targets.....	260
ProFold – Quantum Computing at Davis.....	262
Comparative Recall (CR) Analysis for Assessment of Protein Structure Models Against Experimental NMR Data: Characterizing Multiple Conformational States	264

Omegafold-based Optimization

Ruihan Guo, Ruidong Wu

Helixon Inc.

guoruihan.sansi@gmail.com

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

We participate in the CASP15 as a human group “Agemo”. Based on Alphafold¹ and Omegafold² released in colabfold, we have tried different tricks like MSA-generation and contact-based MSA filter. In addition, we have built a structure evaluation model for ranking structures from different models.

Methods

In our method, we try to optimize the pipeline of Alphafold and Omegafold. Specifically, we try to augment the co-evolution information from MSA for those sequences that lacking that. We have proposed a profile-based method and a language-model-based method to generate MSAs.

MSA generation and filter: Before folding, we used a pretrained language model for MSA sequence generation. Sequences with 15% masked are sent into PLM to regrow, which will be augmentation for co-evolution. The generated MSA sequences are filtered by length, coverage and abundance of co-evolution information. To obtain more diverse and useful co-evolution information, we filter the natural and generated MSA sequences by contact-based score predicted by Omegafold.

Structure Evaluation: To evaluate the structures predicted by different models, we have trained a scoring function to rank the candidates. The training data is generated by Omegafold. Adding noise to the single and pair representation, we can control the IDDT of the predicted structure. The augmented training data is also satisfied with the basic constraints of protein folding.

Availability

The method is not publicly available. The components Omegafold can be found in <https://colab.research.google.com/github/sokrypton/ColabFold>.

1. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>.
2. Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuo fan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, Jian Peng. bioRxiv 2022.07.21.500999; doi: <https://doi.org/10.1101/2022.07.21.500999>.

AIRFold-Omegafold Ensemble

Ruihan Guo^{1*}, Yuxuan Song^{2*}, Jingjing Gong^{2*}, Xin Hong²,
Jianzhu Ma¹, Jian Peng¹, Yanyan Lan²

1 – Helixon Inc., 2 – AIR, * – equal contribution

guoruihan.sansi@gmail.com

Key: Auto:N; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y

We participate in the CASP15 as a human group ‘Agemo_mix’. In this method, we use an improved version of Alphafold named AIRFold which integrated a novel homology-miner module to better process the co-evolution information. We also show that Omegafold¹ can be a complementary method for Alphafold², especially when facing orphan sequences and loop fragments.

Methods

Homology Miner: To better utilize the co-evol(co-evolution) in the homology sequences, we design and implement a novel homology miner module. The module could in general be divided into two parts, the co-evol information augmentation part and the co-evol information processing part. The augmentation part includes both the homology sequence retrieval methods based on deep representation learning and homology information generation methods based on deep generative model. To process the gained co-evol information from multiple sequence alignments data, the processing part uses information theoretic inspired quantity as the optimization metric which has demonstrated the effectiveness and robustness.

AIRfold-Omegafold Ensemble: Omegafold doesn’t need co-evol information or template to predict, hence it can be a good complementary method for Alphafold. To evaluate the necessity using Omegafold, we’ve designed a scoring function based on pLDDT and abundance of MSA. For fragments that acquire more co-evol information, we will use embeddings and results from Omegafold to reinforce.

Monte-Carlo based Complex Prediction: For complicated complexes such as H1111, H1114, and H1137, Alphafold-Multimer and Omegafold-Complex can hardly predict the correct docking position in one shot. To increase the diversity of docking positions and avoid long time computing of large complex targets, we propose a method to assemble the complex step by step. First of all we separate the complex and propose several possible subsets by contact prediction. Then we predicted the subsets of complex and selected top five models for future prediction. With the structure of randomly selected subsets of each component as template, the component can be predicted with higher quality and better interpretability.

Multiple-conformation Selection: Given the protein structures obtained through different models or strategies, we use strategic process to conduct multiple-conformation selection. The

key components of this part include diversity encouraged conformation cluster, evaluation metric debiasing, and energy evaluation. With the above-mentioned components, it is expected to rank different conformation results in proper order and also get a more diverse candidate set which could cover as many possible stationary states as possible.

Availability

The method is not publicly available. The component Omegafold can be found in <https://colab.research.google.com/github/sokrypton/ColabFold>.

1. Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuo fan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, Jian Peng
bioRxiv 2022.07.21.500999; doi: <https://doi.org/10.1101/2022.07.21.500999>.
2. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>.

Per-atom deviation prediction of protein-ligand binding pose with a transformer-based model

Tao Shen¹, Fuxu Liu¹, Jinyuan Sun², Yifan Bu¹, Zechen Wang³, Weifeng Li³, Peng Xiong¹, Liangzhen Zheng^{1,4} and Sheng Wang^{1,4}

1 - Shanghai Zelixir Biotech Co. Ltd, 2 - Institute of Microbiology, Chinese Academy of Sciences. 3 - School of Physics, Shandong University. 4 - Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

yishan@zelixir.com, wangsheng@zelixir.com

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA:N; Fragm:N; Cont:Y; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

Protein-ligand binding patterns are key information for drug discovery. Although the protein structure could be accurately predicted through AlphaFold2¹, it is reported that the ligand binding pocket residue side chains may not be precise enough to model the ligand binding in some cases. How to make use of the available AlphaFold-DB² of protein structures to accelerate the drug discovery process could be an interesting topic.

Methods

From RCSB PDB, we collected all protein-ligand complexes, and based on which we predicted all the protein structures using fastMSA³ based accelerated torch-version AF2, and docked all the ligands into their corresponding predicted protein structures by AutoDock Vina⁴ to assemble a very large decoy dataset. The docking pose RMSD was determined by firstly superimposing the AF2 predicted structure to the native structure and using the native ligand pose as the ground truth as did in other cross-docked dataset. For this decoy dataset, the docking poses with very large RMSD (>15 angstrom) were discarded. Based on this dataset (with similarity-based train-test split), we designed a transformer-based model to predict per-atom RMSD of the docking poses⁴ with respect to the native conformations calculated. The model directly takes 3D coordinates and atom types as input. Self-attention mechanisms allow the model to capture protein-ligand interactions. The RMSD value is discretized into 20 bins with an interval of 0.5. Cross-entropy loss is employed to calculate if the predicted per-atom RMSD falls in the ground truth bin. A mask atom prediction task is also applied to improve robustness. For CASP15 protein-ligand predictions, the AF2 predicted protein structure was used for ligand docking and template-based ligand modeling, and the poses were scored by the deep learning model and clustered into 5 groups for submission.

1. Jumper, J., Evans, R., Pritzel, A. et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583-589.
2. Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 2022, 50(D1): D439-D444.

3. Hong L, Sun S, Zheng L, et al. fastmsa: Accelerating multiple sequence alignment with dense retrieval on protein language. bioRxiv, 2021.
4. Trott O, Olson A J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 2010, 31(2): 455-461.

A hybrid potential energy function for protein-ligand pose scoring and optimization

Zechen Wang¹, Tao Shen², Fuxu Liu², Yifan Bu², Jinyuan Sun³, Sheng Wang^{2,4}, Yanjie Wei⁴,
Liangzhen Zheng^{2,4} and Weifeng Li¹

1 - School of Physics, Shandong University, 2 - Shanghai Zelixir Biotech Co. Ltd, 3 - Institute of Microbiology, Chinese Academy of Sciences, 4 - Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

yishan@zelixir.com, lwf@sdu.edu.cn

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:Y.8-10; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

Accurate prediction of protein-ligand interaction patterns is one of the key issues in computer-aided drug design. The advent of AlphaFold2¹ made it possible to obtain more protein structures at a lower cost. Nonetheless, predicting the specific location and orientation of ligands in protein binding pockets remains a challenge. This increases the need for more refined protein pocket modeling and for protein-ligand interaction description. Here, we predicted protein structures and binding pockets based on AlphaFold-2 and Pointsite², respectively, and then docked ligands into the binding pockets. We employed a hybrid scoring function based on deep learning combined with traditional scoring functions to rank protein-ligand binding poses, and further optimized the binding poses of ligands based on this scoring function.

Methods

The initial ligand binding poses were generated by AutoDock Vina³ and other tools based on predicted protein structure and pocket. In addition, protein-ligand templates based on combined similarity search against whole-PDB level structures were adopted to generate more reliable ligand poses by structure superposition. Inspired by physics, we designed a deep learning-based scoring function DeepRMSD⁴ to predict the root mean square deviation (RMSD) of the docking pose with respect to the native pose. We extract interaction features based on pseudo-van der Waals and Coulomb potentials, which are fed into a multilayer perceptron to predict RMSD. Combining DeepRMSD with Vina, a new hybrid scoring function called DeepRMSD+Vina is constructed. The training and testing sets were adopted from PDDBind2019 dataset by generating large-scale ligand binding poses. Furthermore, in view of the differentiability of DeepRMSD+Vina for molecular coordinates and its superior performance in docking ability, we designed a ligand conformation optimization framework. The optimization algorithm does not directly change the coordinates of each atom of the ligand, but changes the conformation by translating, rotating and twisting the rotatable bonds inside the molecule, thus ensuring the rationality of the structure of the ligand molecule in the optimization process.

1. Jumper, J., Evans, R., Pritzel, A. et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583-589.
2. Yan, X., Lu, Y., Li, Z., Wei, Q., Gao, X., Wang, S., Wu, S. & Cui, S. (2022) PointSite: A Point Cloud Segmentation Tool for Identification of Protein Ligand Binding Atoms. *J Chem Inf Model*, 62, 2835-2845.
3. Trott O, Olson A J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 2010, 31(2): 455-461.
4. Wang Z, Zheng L, Wang S, et al. A fully differentiable ligand pose optimization framework guided by deep learning and traditional scoring functions. *arXiv preprint arXiv:2206.13345*, 2022.

Alchemy-RNA

Yu Li¹, Tao Shen², Sheng Wang², Jiayang Chen¹, Siqi Sun³, Zhangzhi Peng³, Liang Hong¹

¹ - CUHK; ² - Zelixir Biotech, Shanghai, China; ³ - Fudan University

RNA structure determination and prediction served as a critical task across various biological applications. Currently, none of the existing approaches is accurate enough, and most of them utilize large-scale sampling, which is time-consuming. Here in CASP15, we develop a deep learning approach, RhoFold, to accurately predict RNA three-dimensional structures.

Several strategies are proposed to tackle the computational challenges in this problem.

Firstly, RhoFold is composed of a fully differentiable end-to-end deep learning model, which takes the maximal usage of the existing data with minimal human interference and directly outputs the coordinates of all atoms in a valid RNA 3D structure.

Secondly, our method utilizes multi-aspect information of the RNA sequence, including multiple sequence alignments (MSAs) and RNA foundation model (RNA-FM) embedding, to infer the 3D structures.

Thirdly, by introducing secondary structure information into the loss function, RhoFold is forced to capture the RNA folding process instead of only memorizing the training data and thus avoid overfitting.

Finally, inspired by AlphaFold , we also developed a novel procedure to perform self-distillation and use confidently predicted structures to augment the training dataset.

RNA Structure Prediction Using Motif Assembling and Optimization with Statistical Potential

Peng Xiong[#], Huan Yang, Bin Huang, Ke Chen, Liangzhen Zheng, Tao Shen, Sheng Wang

Zelixir Biotech, Shanghai, China

xiongpeng@zelixir.com

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:N; MD:Y.*

In CASP 15, we tested our RNA tertiary structure pipeline using structure motif assembling and full atom optimization with BRiQ statistical potential¹. This method took the strategy of Monte Carlo sampling to generate structure motifs and final models. The energy function we used is a high-resolution statistical potential, describing all bonding, base pairing, steric clashes and other polar interactions in a high-dimensional statistical manner. Homology motifs were used when available.

Methods

Our structure prediction pipeline consists of the following steps 1) Get the secondary structure of the target sequence according to literature, homology structures, prediction results or manually; 2) Assign structure motifs based on secondary structure; 3) Predict motif structures by ab initial structure sampling or structure refinement from homology template. For target R1107, R1108, R1116, R1117 homology models were used, for target R1126 and R1136 G-quadruplex motif were extracted from native structures; 4) Run Monte Carlo sampling for motif assembling and structure refinement using SWORD-RNA program.

For each target, we generated thousands of models and selected five best models to submit.

1. Xiong, P. et al. (2021). Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving RNA model refinement. *Nature Communications* **12**(2777).

APOLLO (QA)

Accuracy estimation of individual multimeric protein models using energy-based models

Andrew Jordan Siciliano¹, Chenguang Zhao¹, Tong Liu¹, Zheng Wang^{1*}

1- Department of Computer Science, University of Miami

zheng.wang@miami.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:N; MD:N.*

APOLLO used Energy-Based Models (EBMs) to assess the overall fold accuracy (global score), overall interface accuracy (global interface score), and confidence scores for interface residues (local interface scores) in the CASP15 experiment. The intuition behind our use of EBMs stems from the fact that proteins fold to their lowest energy state. The EBMs were trained using the proposed loss for energy-based regression¹. Feature embeddings (local or global, from a pre-trained transformer² model) and respective logit values were taken as input to the EBMs. The supervised transformer models that produced the embeddings (denoted as pre-trained) were trained separately from the EBMs. The embeddings were taken as outputs from earlier layers in their respective pipelines. Given an embedding-target (logit) pair, our EBM outputs a compatibility score (negative energy¹). During inference, the logit that produces the highest EBM output (negative energy) given an input embedding is denoted as the prediction. A sigmoid function is used to convert the predicted logit back to the probability space.

Methods

We took input features (details of input features to these pre-trained models can be found in LAW, MASS (QA) abstract) and fed them into the described above pre-trained models to produce learned feature embeddings. These pre-trained models utilized graph transformer layers and/or convolutional 1D layers and BCE/L1 loss functions (with their respective targets).

The input to an EBM is a logit-embedding pair. Negative (adversarial) samples were sampled from a normal distribution¹, with the modification of adding a sliding standard deviation to account for asymptotes in the logit function. Fixed quartiles for a given sample (hyper-parameter) in the probability space were used to calculate the standard deviation in the logit space. To perform inference given an embedding, we iteratively check a fixed span of logits associated with the probability space. Dropout layers were frequently used throughout our models to help prevent overfitting.

The local-interface-score predictor of APOLLO passed input features into a pre-trained model consisting of transformer layers² and linear layers. These embeddings were then passed into an EBM that contained recurrent linear and standard linear layers. The recurrent linear layers propagated the logit score down the pipeline. Since this predictor was performing a binary classification task, we assigned fixed values in the logit space to be 0 and 1 accordingly.

The global-score predictor of APOLLO passed input features into a pre-trained model consisting of transformer and convolutional 1D layers. These embeddings were given as input to an EBM consisting of recurrent and standard linear layers. The recurrent linear layers propagated the logit score down the pipeline. The inference was performed using a separate model as opposed to brute force checking. This separate inference model took in energy values associated with (across the fixed span of potential logit values) both the global-score and local-interface EBMs. The model also took the predicted global score (final output) from the pre-trained model (used for producing learned embeddings). This inference model consisted of convolutional 1D and linear layers. The supervised inference model was trained separately from and after the EBM. The final output of the inference model is the global-score prediction.

The global-interface-score predictor of APOLLO passed input features into a pre-trained model consisting of transformer and 1D convolutional layers. This pre-trained model produced learned feature embeddings which were then used, alongside potential logit values, as input to the EBM. The global-interface-score EBM consisted of recurrent and standard linear layers. The recurrent linear layers propagated the logit score down the pipeline.

1. Gustafsson, F.K., et al., How to train your energy-based model for regression. arXiv preprint arXiv:2005.01698, 2020.
2. Shi, Y., et al., Masked label prediction: Unified message passing model for semi-supervised classification. arXiv preprint arXiv:2009.03509, 2020.

BAKER (TS, Assembly, Ligand, RNA)

Accurate predictions of protein structures and interactions utilizing RoseTTAFold2

Minkyung Baek^{1,2}, Guangfeng Zhou^{1,2}, Ian Humphreys^{1,2}, Ivan Anishchenko^{1,2}, Qian Cong^{3,4}, Frank DiMaio^{1,2}, and David Baker^{1,2,5}

1 - Department of Biochemistry, University of Washington, Seattle, WA, USA; 2 - Institute for Protein Design, University of Washington, Seattle, WA, USA; 3 - Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA; 4 - Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA; 5 - Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

dabaker@uw.edu

Key: Auto:N; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N

The BAKER group performed various ranges of predictions from protein tertiary structures to protein-small molecule interactions by utilizing the RoseTTAFold2 with human interventions.

Methods

Protein tertiary structure prediction and alternative state sampling

For human group predictions, the multiple sequence alignments (MSAs) generated by our server protocol (see BAKER-SERVER) were manually inspected to fine-tune the e-value and coverage cutoffs and enriched with metagenomic sequences¹. Templates were re-searched if MSAs were updated. We also tried more diverse sampling by running the same model multiple times with five different random seeds, and the number of recycling was also manually adjusted if necessary. The resulting 25 models were ranked by predicted CA-IDDT (pIDDT), and the top five models were submitted after relaxation.

For the alternative state sampling, we gave biases to RoseTTAFold2 prediction by providing a biased set of templates corresponding to a certain state. To make RoseTTAFold2 more sensitive to the given templates, we sub-sampled MSAs to have less than 30 sequences.

Protein assembly prediction using RoseTTAFold2

For the homo-oligomeric proteins, the MSA is copied multiple times for each component with gap padding. For the hetero-oligomeric proteins where subunits are from the same species, we generated paired alignments by the following. All proteomes for bacteria or eukaryotes were downloaded from NCBI² and JGI³ genome databases. We used BLAST⁴ to generate loose orthologous sequences for each query protein by identifying the forward best hit from proteomes in the database. Using these similar sequences, we used *phmmer*⁵ to search against the loose orthologues. We filtered the alignments to create an initial seed alignment which was then converted into a HMM with *hmmbuild*⁵. This seed alignment is used to align the remainder of the orthologous sequences using *hmmsearch*⁵. Sequences from the same proteome were paired, redundancy was removed using *hhfilter*⁶, and additional unpaired sequences were added to these

alignments with gap padding. For the hetero-oligomers having subunits from different species (e.g. antigen-antibody), MSAs for each chain are stacked in a block-diagonal fashion with padding on the off-diagonals. We also utilized templates for subunits and complexes. We ran all five RoseTTAFold2 models five times, resulting in 25 predictions. The final models were ranked by pLDDT, and the top five models were submitted after energy minimization using Rosetta FastRelax with coordinate constraints on given structures.

Protein-ligand docking with receptor models predicted by RoseTTAFold2

Initial ligand conformation was generated from the SMILES string using UCSF Chimera⁷. Ideal bond geometry and partial charges were computed using AmberTools⁸ for small molecules. We used our human prediction models for receptors. The binding site was determined based on the existing template structures with similar small molecules bound to the protein structure. For a few targets that have no ligand-bound template, we manually selected a few putative binding sites based on the shape of the protein surface. Metal ions were manually placed in the receptor structures based on the templates, and they were fixed during small molecule docking.

Docking of small molecule ligands was performed using Rosetta GALigandDock⁹, which uses a physically realistic energy model with genetic algorithm optimization. 20 independent docking runs were performed for each receptor model, and all the generated models were ranked by the estimated binding affinity ($dG = dH + TdS$). The enthalpy change (dH) upon binding was estimated using Rosetta energy: $dH = E(\text{complex}) - E(\text{receptor}) - E(\text{ligand})$, and the entropy change was estimated by a short Monte Carlo simulation of the torsional entropy of the ligand.

Since GALigandDock can only dock one ligand each time, for targets that have multiple ligands, we docked one ligand at a time and used the top five ranked complex structures for the next ligand. For homo-oligomer targets such as T1124, T1170, etc., a ligand was docked to one of the subunits, and the final complex structure was reconstructed by aligning the ligand-bound subunit to the others. For hetero-multimer targets such as H1171 and H1172, we docked ligands to all of the subunits independently and reconstructed whole complex structures by combining top-ranked predictions for each subunit. The final structures were relaxed with coordinate restraints to remove potential clashes before submission.

RNA structure and protein-RNA complex structure prediction using RoseTTAFoldNA

During CASP15, we extended the RoseTTAFold2 to predict structures of nucleic acid and protein-nucleic acid complexes. We developed a single trained network, RoseTTAFoldNA¹⁰, that rapidly produces 3D structures with estimated confidences for protein-DNA and protein-RNA complexes, and for RNA tertiary structures. We trained this end-to-end RoseTTAFoldNA model using a combination of protein monomers, protein complexes, RNA monomers, RNA dimers, protein-RNA complexes, and protein-DNA complexes, with a 60/40 ratio of protein-only and NA-containing structures.

To generate MSAs for RNA, sequences were searched using *blastn*⁴ over three databases (RNACentral¹¹, rfam¹², and nt²) to identify hits, then using *nhmmer*¹³ to rerank hits. Similar to protein MSA generation, we used successive e-value cutoffs (1e-8, 1e-7, 1e-6, 1e-3, 1e-2, 1e-1),

stopping when the MSA contains more than 10,000 unique sequences with >50% coverage. The models were generated using RoseTTAFoldNA with manual intervention on a choice of MSAs, the number of recycling steps, etc. For large RNA molecules, models were predicted in segments and combined. The top five models ranked by estimated confidence were submitted after relaxation.

Availability

The RoseTTAFoldNA is available at <https://github.com/uw-ipd/RoseTTAFold2NA>, and Rosetta GALiagndDock is available through Rosetta modeling package, downloadable from <https://www.rosettacommons.org/software>.

1. Wu, Q., Peng, Z., Anishchenko, I., Cong, Q., Baker, D., Yang, J. (2020). Protein contact prediction using metagenome sequence data and residual neural networks. *Bioinformatics* 36, 41-48.
2. Wheeler, D. L., et al. (2006). Database resources of the national center for biotechnology information. *Nucleic acids research*, 34(suppl_1), D173-D180.
3. Clum, A., et al. (2021). DOE JGI metagenome workflow. *Msystems*, 6(3), e00804-20.
4. Camacho, C., et al. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, 10(1), 1-9.
5. Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS computational biology*, 7(10), e1002195.
6. Steinegger, M., et al. (2019). HH-suite3 for fast remote homology detection and deep protein annotation, *BMC bioinformatics*, 20(1), 1-15.
7. Pettersen, E. F., et al. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13), 1605-1612.
8. Case, D. A., et al. (2021). Amber 2021. University of California, San Francisco.
9. Park, H., Zhou, G., Baek, M., Baker, D., DiMaio, F. (2021). Force field optimization guided by small molecule crystal lattice data enables consistent sub-angstrom protein-ligand docking. *Journal of chemical theory and computation*, 17(3), 2000-2010.
10. Baek, M., McHugh, R., Anishchenko, I., Baker, D., DiMaio, F. (2022). Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA, *biorxiv*.
11. RNAcentral Consortium. (2021). RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic acids research*, 49(D1), D212-D220.
12. Kalvari, I., et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, 49(D1), D192-D200.
13. Wheeler, T. J., & Eddy, S. R. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, 29(19), 2487-2489.

BAKER-SERVER (TS)

Automated protein tertiary structure prediction using RoseTTAFold2

Minkyung Baek^{1,2}, Ivan Anishchenko^{1,2}, Frank DiMaio^{1,2}, and David Baker^{1,2,3}

1 - Department of Biochemistry, University of Washington, WA, USA; 2 - Institute for Protein Design, University of Washington, WA, USA; 3 - Howard Hughes Medical Institute, WA, USA

dabaker@uw.edu

Key: Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N

BAKER-ROSETTASERVER performed fully automated protein tertiary structure predictions for TS targets. The server employed the RoseTTAFold2 method, an improved version of RoseTTAFold¹, which combines several components of AlphaFold² to the RoseTTAFold's three-track architecture.

Methods

Sequence and template searches

Multiple sequence alignments (MSAs) for the target sequences were generated by several rounds of iterative *hhblits*³ search against the Uniclust30 database⁴ (June 2021 version) with gradually relaxed e-value cutoffs (1e-30, 1e-10, 1e-6, and 1e-3). If the resulting MSA contains less than 4,000 sequences after filtering with coverage of 50% and sequence identity of 90%, it performs a final *hhblits* search against the BFD database² with e-value cutoff 1e-3. The generated MSAs were then used to search for putative structural templates in the PDB (Apr 2022 version) by *hhsearch*⁵.

Predicting protein structures using RoseTTAFold2

We developed RoseTTAFold2, which consists of pure three-track architecture and incorporates the FAPE loss² and recycling during training. We took the concept of structurally coherent attention from AlphaFold, but implemented it in a way that scales with $O(L^2)$ rather than $O(L^3)$, enabling much more efficient modeling for large proteins. RoseTTAFold2 is trained on not only experimentally determined protein monomer structures but also model structures predicted by AlphaFold and protein complex structures in the PDB. RoseTTAFold2 takes MSAs and templates as inputs, and provides full-atom models with residue-wise model confidence in terms of CA-IDDT score (pLDDT). Five models were generated by utilizing five different checkpoints from RoseTTAFold2 training. The final models were ranked by pLDDT. All models underwent a final relax with coordinate constraints on all heavy atoms using Rosetta⁶ before submission.

Availability

The original RoseTTAFold is available at <https://github.com/RosettaCommons/RoseTTAFold> (source code) and <https://rosetta.bakerlab.org> (web-server).

1. Baek, M., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557), 871-876.
2. Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-590.
3. Remmert, M., Biegert, A., Hauser, A. (2012) HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment *Nature Methods*, 9(2), 173-175.
4. Mirdita, M., et al. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research*, 45(D1), D170-176.
5. Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21, 951–960.
6. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E., Baker, D. (2014). Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Science*, 23(1), 47-55.

OpenComplex: template and docking enhanced protein quaternary structure prediction by geometric deep learning

He Huang¹, Jingcheng Yu², Daiwen Sun¹, Jiashan Li¹, Tong Wu¹, Yu Shu², Siwei Dong², Wenda Wang¹, Yemin Shi², Yue Cao², Wenhao Huang², Qiwei Ye^{2*} and Xinqi Gong^{1,2*}

1 - Institute for Mathematical Sciences, Renmin University of China, Beijing, China,

2 - Beijing Academy of Artificial Intelligence, Beijing, China

qwye@baai.ac.cn, xinqigong@ruc.edu.cn

Key: Auto:N; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y

In CASP15, we developed OpenComplex, an end-to-end deep learning based model with MSA encoders and structural decoder that predicted protein structures based on the layout of AlphaFold2¹ and AlphaFold-Multimer². Moreover, when it's difficult, this strategy has been enhanced with multimer complex templates and deliberate docking, minimization and scoring.

Methods

OpenComplex: OpenComplex is a deep learning based model with MSA encoders for biomolecules and a structural decoder. Trained on the Protein Data Bank (PDB), RNA families database (RFRAM), and other databases, OpenComplex can produce suitable models for monomers, multimers and big protein complexes. In addition, we screened MSAs using a variety of design options, incorporated multimer templates for protein complexes, and introduced additional randomness to produce diversity.

Template Search: we focus on the construction and selection of protein templates for multimer as follow, (i) we close the duplication condition in the original AlphaFold-Multimer template search before looking for a template of higher grade. (ii) We retain more monomer candidate templates. (iii) If there are templates for each chain from the same complex, we select the multimer protein template based on the template ID that was used to build the mask.

Manual intervention: for certain big or intricate targets, we employ manual involvement, such as assembly and docking. We employ symmetric docking and binary docking tools, respectively, for targets with stoichiometry A_n or A_1B_1 . When the stoichiometry of the target does not fulfill the aforementioned conditions, we are unable to build the complete structure in a single shot and must employ other assembly procedures. For target H1137, the membrane protein structure with high homology is searched through sequence alignment, and the complex is divided into several subunits for docking with reference to the spatial arrangement of the template structure. For other symmetric hetero-multimers, such as H1114, we predicted the structures of BC , AB_2C_2 and other subunits respectively. We applied C_8 symmetry docking to subunit BC , and then docking with chain A , C_2 , D_2 symmetry docking to subunit AB_2C_2 , generating a large number of candidate conformations. In addition, we alter the predicted structure using the Jackal³ package, which provides modeling using multiple anticipated models as templates. We handle excessively lengthy monomer sequences by separating them into two or

more domains, which are independently inferred by OpenComplex and subsequently superimposed by PyMol⁴ based on aligned sections.

Model selection: we reranked the decoys using a hybrid scoring strategy, grading complex created interfaces by integrating energy functions and monomer structure by plddt score, quality of secondary structure, and disordered area. We manually filtered out these decoys for some large complexes in which inaccurate prediction of local domains resulted in long loops and increased structural clashes. In addition, we leverage the literature-reported interface zones, which provide a crucial basis for our selection of decoys.

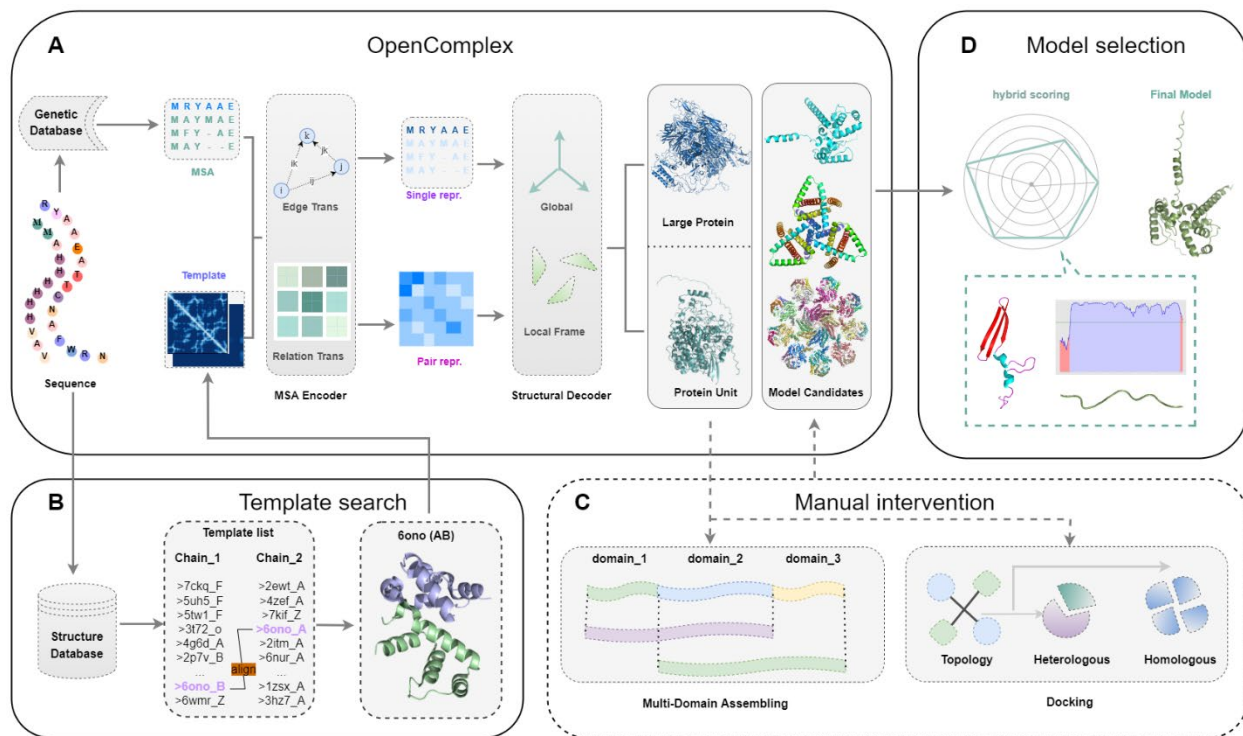


Figure 1: Pipeline overview: A) The architecture of OpenComplex, with MSA encoders and structural decoder, was used for candidate structure generation. B) Construction and selection of protein templates for multimer. C) Manual intervention such as assembling and docking for some large or complicated targets. D) Final models were selected by a hybrid scoring strategy.

Results

At the present date, several TS targets are released on PDB. Our proposed prediction models compared favorably to structures that were made public. We calculated TM-score, RMSD and LDDT of the top five models for two monomer targets (T1120, T1133) and DockQ for two heterodimers (H1106, H1134) (Figure 2).

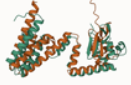
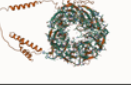
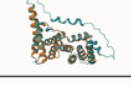
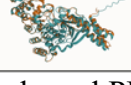
Target ID	Predicted model	PDB ID (chain)	TM score	RMSD	LDDT	DockQ	Structure alignment
T1120	model 5	7qyb (B)	0.342	9.738	#	#	
T1133	model 5	8yds (A)	0.976	1.261	0.887	#	
H1106	model 2	7qih (AB)	#	#	#	0.760	
H1134	model 2	7ubz (AB)	#	#	#	0.424	

Figure 2: Structural comparison of OpenComplex predicted structural model and released PDB structure: our models for target T1133 and H1106 are close to native structures.

Availability

The code and models will be made available to the public shortly.

1. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021.
2. Evans R, O'Neill M, Pritzel A, et al. Protein complex prediction with AlphaFold-Multimer. *BioRxiv*, 2021.
3. Petrey, D., Xiang, Z.X., Tang, C.L., et al. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*, 2003. 53: 430-435.
4. Schrodinger, LLC, The PyMOL Molecular Graphics System, Version 1.8. 2015.

Bench

See: *Yang, Yang-Server, Yang-Multimer*

BhageerathH-Pro: Protein Tertiary Structure Prediction Server

Akshata Hegde¹, Smriti Pranjal¹, Devendra Prajapat¹, Shashank Shekhar¹ and B. Jayaram^{1,2,3}

1 – Supercomputing Facility for Bioinformatics & Computational Biology, 2 – Department of Chemistry, 3 – Kusuma School of Biological Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India - 110016

akshata@scfbio-iitd.res.in, bjayaram@scfbio-iitd.res.in

Key: Auto:Y; CASP_serv:N; MSA:N; Fragm:Y.v; EMA:Y; MD:Y

BhageerathH-Pro protein tertiary structure prediction server is based on an *ab initio*/homology hybrid methodology. It integrates several methodological innovations designed over the years in SCFBio, such as grid sampling, empirical energy-based scoring, physicochemical filters for screening decoys, RM2TS and NCL methodologies for structure generation, ProTSAV for structure selection, together with template-based homology modeling and molecular dynamics refinement.

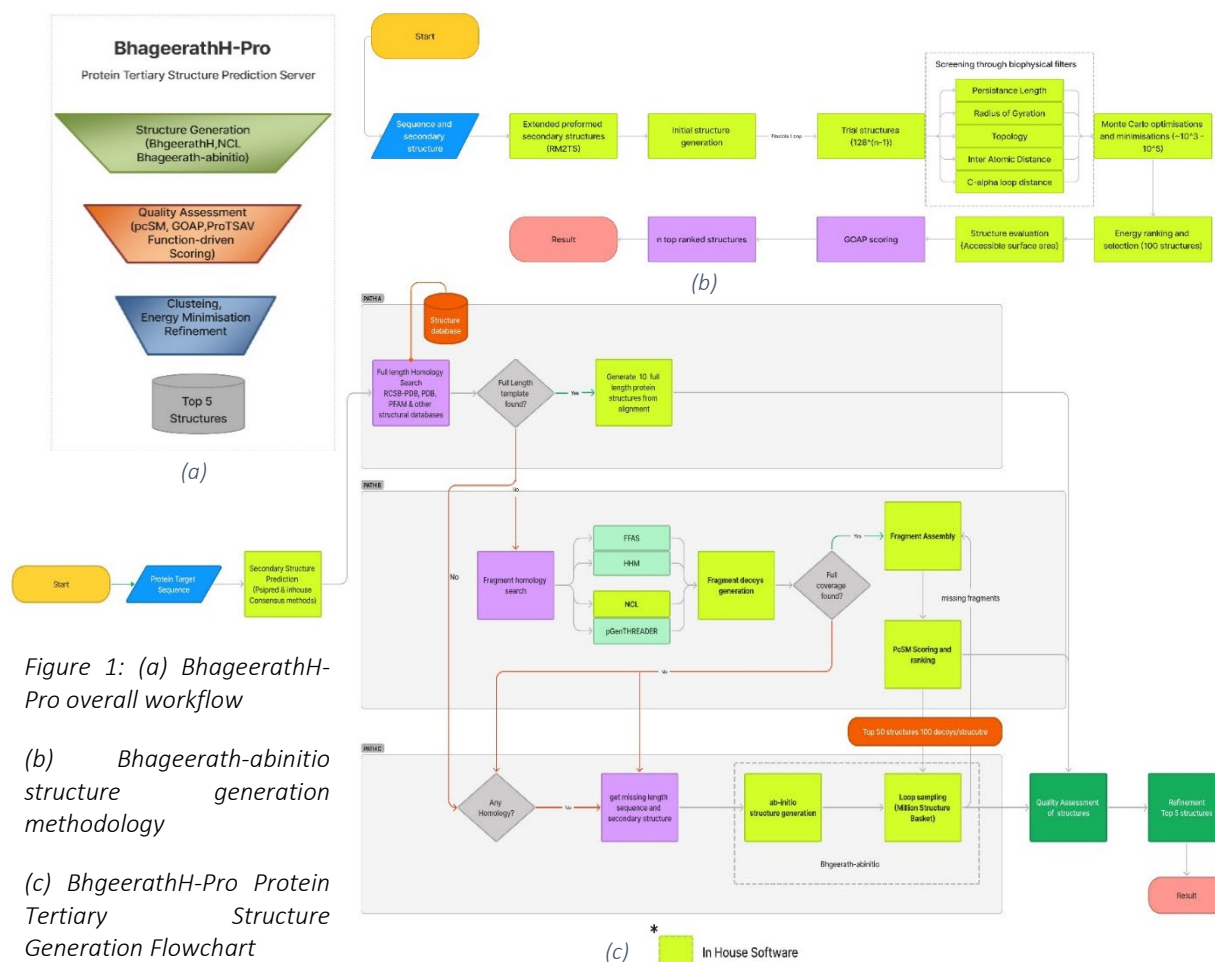


Figure 1: (a) BhageerathH-Pro overall workflow

(b) Bhageerath-abinitio structure generation methodology

(c) BhageerathH-Pro Protein Tertiary Structure Generation Flowchart

Despite the now well-known successes of Alphafold2, we refrained from inserting alphafold2 codes into our methodology and intended to benchmark our server against released experimental structures, and Alphafold predicted structures. The expectation, as we analyze the shortcomings of our server, from CASP to CASP, is that someday soon, physics-based methods can yield reliable tertiary structures.

Methods

BhageerathH-Pro is an advanced version of Bhageerath-H¹ which consists of three major steps of structure generation, quality assessment, and refinement respectively [Figure 1(a)]. The structure generator methodology follows three paths namely, Path A, Path B, and Path C [Figure 1(c)]. Path A caters to the full-length template modeling, while Path B includes the New Chemical Logic of amino acids driven protein alignment and decoys generation² and fragment assembly using the previously developed StrGen³ algorithm. The Bhageerath-*ab initio* module [Figure 1(b)] is implemented as Path C for *ab initio* structure prediction in the absence of homology with an advanced version of RM2TS⁴ and an updated smart Bhageerath⁵, that works on loop sampling. The quality assessment pipeline is designed to pull out the best possible structure from the basket of almost a million conformations generated [Figure 2].

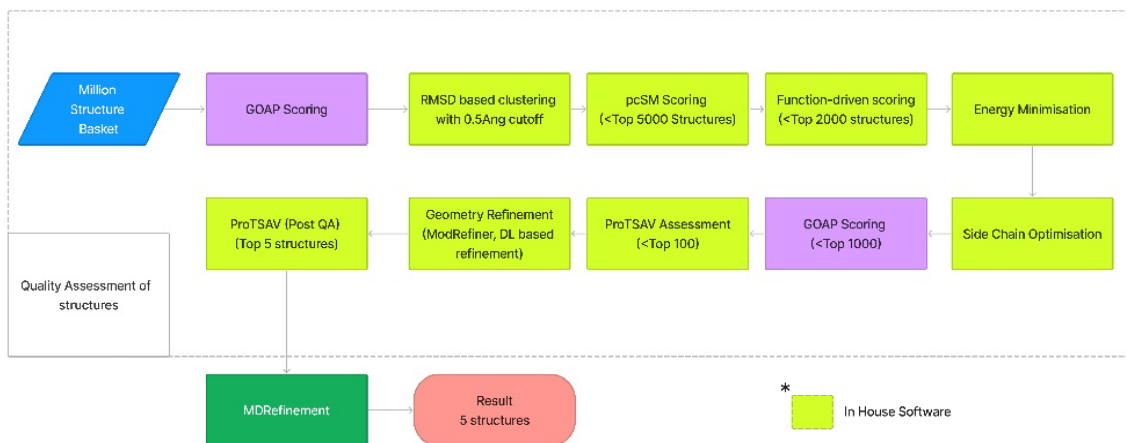


Figure 2: Quality Assessment pipeline for BhageerathH-Pro

The sample conformations are clustered and filtered to retain mutually exclusive topologies, that are scored and ranked using iterative scoring modules like pcSM⁶, Function-driven scoring⁷, and ProTSAV⁸ to extract the top 100 structures. The selected structures are energy minimized using PROSEE⁹, and geometry optimized using the deep encoder-based refiner.

The top 5 models thus generated are subjected to molecular dynamics for final refinement. The models resulting from path A are refined using conventional molecular dynamics (5ns). [Figure:3 Path A], whereas the models obtained from path B or C are refined using Conventional MD, Accelerated MD, and Annealing [Figure:3 Path B or C]. Once the simulations are finished, the lowest energy conformer from each simulation is subjected to

MMPBSA/GBSA calculations to find the most stable conformer, which is further refined using iterative main chain and side chain minimizations.

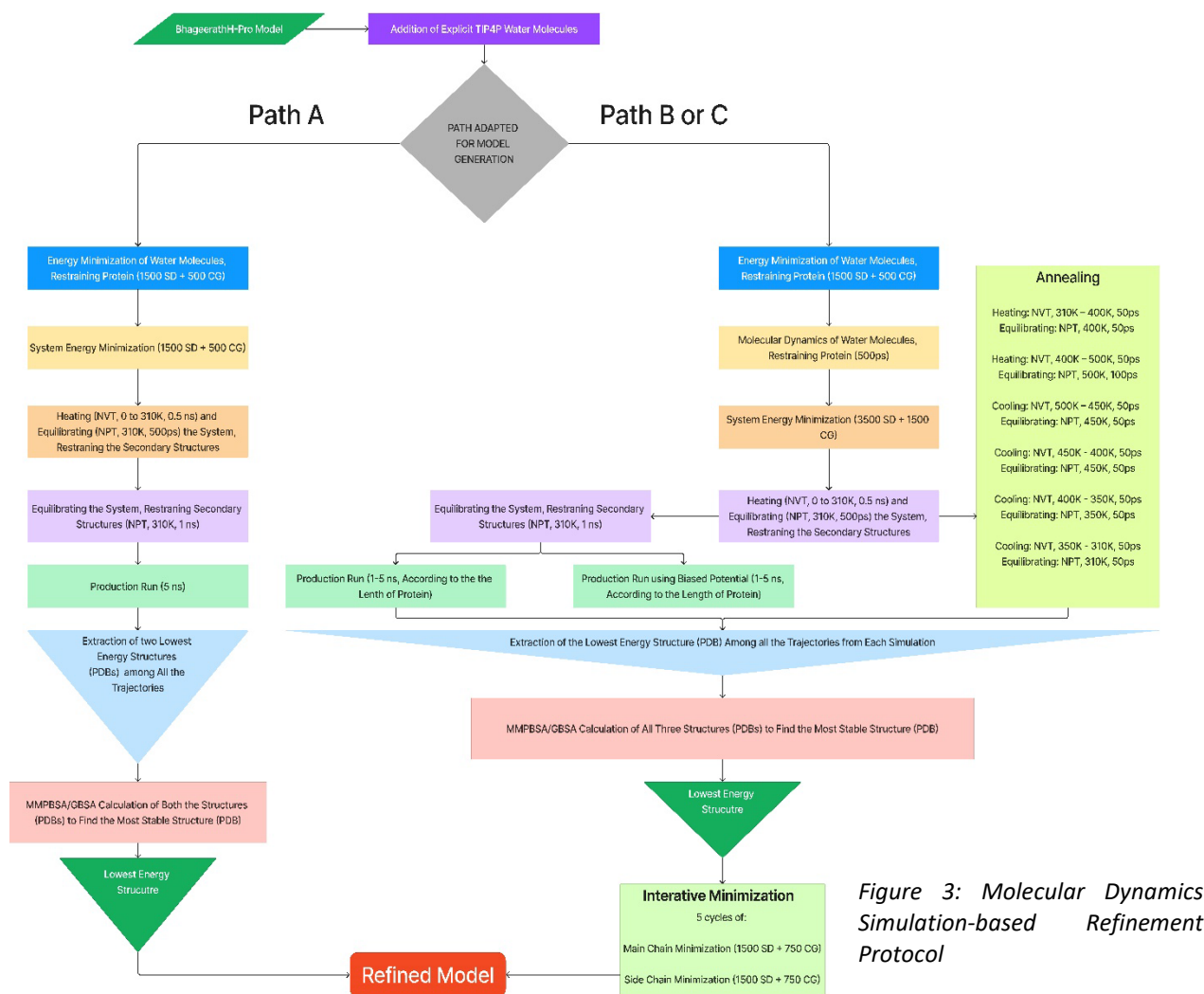


Figure 3: Molecular Dynamics Simulation-based Refinement Protocol

Results

The pipeline outlined above is fully automated and fielded in the recently concluded CASP15 under the TS category as the BhageerathH-Pro server. BhageerathH-Pro has performed reasonably well so far for the targets whose native structures are released in RCSB till now. Out of 7 targets (whose native information is released), BhageerathH-Pro succeeded in predicting 4 of them under low-resolution structures and 3 of them under 5Å of RMSD. Structures submitted from our server are benchmarked with the now famous Alphafold2 structures, and it is observed that for more than 50% of the targets (i.e., 49 out of 94), BhageerathH-Pro structures are within 3Å of RMSD. We have further noticed that mutational effects and functions are well distinguishable in BhageerathH-Pro structures. Most of BhageerathH-Pro modules and tools are freely available in the public domain for the user community.

Availability

BhageerathH-Pro is an open-source web server and is available at the SCFBio website: <http://www.scfbio-iitd.res.in/bhageerathH+/>.

Acknowledgments

This research is supported by the CoE project of the Department of Biotechnology and NSM project administered by Meity and CDAC, Govt. of India to SCFBio. The authors are thankful to Prof. Aditya Mittal for helpful discussions and suggestions.

1. Jayaram, B., Dhingra, P., Mishra, A. et al. Bhageerath-H: A homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins. *BMC Bioinformatics*, **2014**, *15*(Suppl 16), S7.
2. Rahul Kaushik, Ankita Singh and B. Jayaram. "Where informatics lags, chemistry leads," *Biochemistry*, **2018**, *55*(5), 503-505.
3. Dhingra, P. and Jayaram, B. A homology/ab initio hybrid algorithm for sampling near-native protein conformations. *J. Comput. Chem.*, **2013**, *34*, 1925-36
4. Debarati DasGupta, Rahul Kaushik, and B. Jayaram "From Ramachandran Maps to Tertiary Structures of Proteins," *J. Phys. Chem. B*, **2015**. *119* (34), 11136 - 11145.
5. B. Jayaram, K. Bhushan, et al., "Bhageerath: An energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins" *Nucl. Acids Res.*, **2006**, *34*, 6195-6204.
6. Avinash Mishra *et al*, "Capturing Native/Native like Structures with a Physico-Chemical Metric (pcSM) in Protein Folding" *BBA-Proteins and Proteomics*. **2013**, *1834*, 1520-31.
7. P. Amita, *et al*; "Mask Blast with a new chemical logic of amino acids for improved protein function prediction" *Proteins: Structure, Function, and Bioinformatics*, **2021**, *89*(8), 922-924.
8. Ankita Singh, Rahul Kaushik, Avinash Mishra, Asheesh Shanker, and B. Jayaram "ProTSAV: A Protein Tertiary Structure Analysis and Validation Server," *BBA-Proteins and Proteomics*, **2016**, *1864*(1), 11-19.
9. P. Narang, K. Bhushan, S. Bose, and B. Jayaram, "Protein structure evaluation using an all-atom energy based empirical scoring function", *J. Biomol. Str. Dyn.*, **2006**, *23*, 385-406.

Protein modeling and accuracy estimation by Bhattacharya group in CASP15

Md Hossain Shuvo¹, Mohimenuul Karim¹, and Debswapna Bhattacharya^{1,*}

¹*Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA.*

* dbhattacharya@vt.edu

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

We participated in the CASP15 protein modeling category as "Bhattacharya" group which is the result of a system integration of our monomeric quality estimation and refinement methods for structure prediction; and in the accuracy estimation category with our unpublished ongoing work in accuracy estimates for multimeric complexes and inter-subunit interfaces using graph neural networks.

Methods

Our protein modeling pipeline leveraged the AlphaFold2 and RoseTTAFold predictions released by the CASP Prediction Center to perform model selection using our rapid multi-model structural consensus approach clustQ¹. For each of the top selected models, we independently performed structure refinement using refinedD² method and ranked them using the method's internal scoring scheme following a similar strategy as employed by QDeep³, to submit five top-ranked models.

Our unpublished ongoing work in protein complex accuracy estimation first extracts the interface of interacting residues from the complex model and constructs a graph representation of the interfacial region using sequence- and structure-derived node features and novel geometric edge features of the interface interactions. The accuracy of the interfacial region is then estimated using a graph neural network with graph embeddings and multi-head attention. The estimated interfacial accuracy is then combined with consensus-based accuracy estimates of the interacting monomers¹ for estimating the accuracy of the whole multimeric complex.

Availability

QDeep is freely available at <https://github.com/Bhattacharya-Lab/QDeep/>, clustQ is freely available at <http://watson.cse.eng.auburn.edu/clustQ/>, refinedD is available via the DeepRefiner server available at <http://watson.cse.eng.auburn.edu/DeepRefiner/>.

Flexible docking of cyclic peptides to proteins using a combination of CABS-dock with FlexPepDock refinement

M. Zalewski¹, A. E. Badaczewska-Dawid², M. Kurciński¹, S. Kmiecik¹

1 - Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, 1 Pasteura St., 02-093 Warsaw, Poland, 2 - Department of Chemistry, Iowa State University, Ames, IA 50011, USA.

sekmi@chem.uw.edu.pl

The number of peptide-based therapeutics is growing continuously and it is expected to increase even more in coming years. Recently a wide attention has been drawn to cyclic peptides as potential modulators of biomolecular interactions with improved biological activity. Only a few methods exist that enable molecular docking of cyclic peptides, however with some limitations. Here, we present a protocol for a flexible docking of cyclic peptides. Method is based on the combination of a well-established tool for protein-peptide docking – the CABS-dock¹, and on the Rosetta FlexPepDock² refinement.

Methods

Proposed protocol consists of the following steps:

Step 1: Generating peptide starting conformations: 10 random starting models in C-alpha trace representation are generated.

Step 2: Conformational space sampling: Conformational space sampling is performed using Replica Exchange Monte Carlo sampling scheme. Generated models are saved into trajectory for every starting structure.

Step 3: Reconstruction: All generated models are reconstructed to CABS (C-Alpha, C-Beta, Side chains) coarse-grained representation.

Step 4: Scoring: 10 best structures are selected using a combination of CABS energy-based scoring function and k-medoid structural clustering.

Step 5: Reconstruction and refinement: 10 top-scored models are rebuilt to all-atom representation and refined using PD2³ method and Rosetta FlexPepDock tool.

Results

The proposed protocol was evaluated on a set of 38 cyclic peptide complexes. Provided results show that the combination of CABS-dock with Rosetta refinement may be an effective way for docking not only linear, but also cyclic peptides.

Availability

The CABS-dock repository is available at <https://bitbucket.org/lcbio/cabsdock/src/master/>

Rosetta FlexPepDock tool is available at <http://flexpepdock.furmanlab.cs.huji.ac.il/>

1. Kurcinski, M. et al. CABS-dock standalone: a toolbox for flexible protein–peptide docking. *Bioinformatics* 35, (2019).
2. London, N., Raveh, B., Cohen, E., Fathi, G. & Schueler-Furman, O. Rosetta FlexPepDock web server - High resolution modeling of peptide-protein interactions. *Nucleic Acids Research* 39, (2011).
3. Moore, B. L., Kelley, L. A., Barber, J., Murray, J. W. & MacDonald, J. T. High-quality protein backbone reconstruction from alpha carbons using gaussian mixture models. *Journal of Computational Chemistry* 34, (2013).

Cerebra: a convolution-attention-mixed model for protein structure prediction

Jian Hu, Yunxin Xu, Hanyang Zhou, Tianyu Mi and Haipeng Gong

School of Life Sciences, Tsinghua University

hgong@tsinghua.edu.cn

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:Y; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:Y*

In this CASP, we built a protein structure prediction model called Cerebra, which is a hybrid architecture deep learning neural network model, which used convolutional neural network (CNN) and attention mechanism to obtain local information and global information respectively. Cerebra is an end-to-end model, which only needs to input multiple sequence alignment (MSA) to predict the coordinates of each target residue.

Methods

Training data: Cerebra was trained on subset of CATH S35 v4.2¹, then we searched their homologous sequences to build MSA as model input.

Structure prediction: This model could predict the CA atoms coordinates of each residue. We trained the model by minimizing the difference between the predicted coordinates and the real coordinates.

Generate **CASP15 target sequence MSA:** we searched the following databases during CASP15 to obtain related sequences: UniRef30², UniRef90², BFD³, MGnify⁴, etc.

Refinement: The CA atoms from our models were fixed using Modeller⁵ and PDBFixer⁶, and then relaxed using OpenMM⁶ minimization function with Amber ff14SB and customized restrained gradient descent method.

Availability

Cerebra is still under development. Once completed, we will open the source code and related data.

1. Sillitoe, I. *et al.* (2021). CATH: increased structural coverage of functional space. *Nucleic Acids Res* **49**, D266-D273.
2. Bateman, A. *et al.* (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506-D515.
3. Steinegger, M., Mirdita, M. & Sodin g, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat Methods* **16**, 603-+.
4. Mitchell, A. L. *et al.* (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res* **48**, D570-D578.
5. Fiser, A., Do, R. K. G. & Sali, A. (2000). Modeling of loops in protein structures. *Protein Sci* **9**, 1753-1773.
6. Eastman, P. *et al.* (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *Plos Comput Biol* **13**.

Protein Complex Structure Prediction and Scoring Using Machine Learning-based Docking Potential

Se-Ung Han¹, Yu-Chol Choe¹, Myong-Ho Chae¹

¹Department of Life Science, University of Sciences, Unjong-District, Pyongyang, DPR Korea

cmh1971@star-co.net.kp

Our protein complex structure prediction and scoring is based on an optimized protein docking potential (ODP) derived using a single-layer perceptron and docking decoys.

Methods

Docking Potential for complex structure scoring. ODP potential is an upgraded version of neural network-based distance-dependent atom-pair potential for protein docking as described in ref¹. The potential was trained on a large protein complex dataset from ref². To train the neural network, 4400 decoys of even distance(iRMSD) distribution generated by in-house docking decoy generation program were used for each of the 500 protein complex training set.

Decoys were computed applying to one protein of the native protein complex random rotations (with rotation axes going through the interface centroid) and subsequent translations. We considered these artificially generated protein-pair structures to be valid decoys only, if the fraction of atom-pair contacts (maximum distance between atoms 3.8 Å) between atoms belonging to different proteins (ignoring hydrogen atoms) is above 50% relative to the contacts that are present in the corresponding native complex structure. We also checked for atom clashes occurring in these decoys, which we defined for atom pairs, which are closer than 2.4 Å. The maximum number of allowed atom clashes in a decoy was set to the number of clashes found in the native complex geometry plus four additional clashes.

We use atom-pair distance distributions to evaluate docking geometries. The neural network input information uses type specific atom-pairs (with atoms belong to different proteins of the complex), which are assigned to different distance classes (bins). The complete distance range that we take into account extends from 0.0 (although distances close to zero do not occur) to 10 Å and is divided into 16 distance bins.

As a target function we used F_{nat} which is the fraction of native interfacial contacts preserved in the decoy.

Assembly prediction. For each assembly target, the models of the individual subunits were taken from the af2-multimer, af2-standard and BAKER CASP-hosted servers.

Free docking of these subunit CASP server models was done by ZDOCK3.0 for homodimer or hetero-complexes and SymmDock³ for homomultimers. The predicted quaternary structures were then ranked for submission using the ODP potential. Furthermore, the information from assembly templates (if any) along with visual inspection was used for some targets in order to manually filter the modelled complexes.

Multimer structure model quality assessment. For each predicted quaternary structure model, ODP score was calculated and normalized to estimate the overall fold accuracy and overall interface accuracy.

1. Chae, M.H., Krull, F., Lorenzen, S., Knapp, E.W.(2010) Predicting protein complex geometries with a neural network. *Proteins*, 78, 1026-1039.
2. Ravikant, D.V., Elber, R. (2010). PIE-efficient filters and coarse grained potentials for unbound protein-protein docking. *Proteins*, 78, 400–419
3. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., Wolfson, H.J. (2005) PatchDock and SymmDock: servers for rigid and symmetric docking, *Nucleic Acids Res.*, 33(Web Server issue), W363–W367.

Prediction of Protein Tertiary Structure with Accuracy Estimation Using Statistical Potentials and Predicted Distance

Se-Ung Han¹, Yu-Chol Choe¹, Myong-Ho Chae¹

¹*Department of Life Science, University of Sciences, Unjong-District, Pyongyang, DPR Korea*

cmh1971@star-co.net.kp

We participated in the CASP15 tertiary structure prediction experiment as a human group "ChaePred", which is based on our recently developed deep learning based single-model quality assessment method DeepEMA.

Methods

CASP15 server models were evaluated by single-model quality assessment method DeepEMA, and the top model was selected.

We trained our method on the CASP7-CASP10 datasets. Prior to training, we filtered out models with the same GDT-TS scores for a given targets as well as low quality models (IDDT score¹<0.3). In total, the training dataset includes 33,000 models for 435 target proteins.

From each 3D model, we derived the following residue-specific features: 1) one-hot encoded secondary structure (3-state), 2) relative solvent accessibility (RSA) calculated by DSSP, 3) sine and cosine of backbone torsion angles, 4) residue-residue contact environment (the numbers of 20 residue types contacting with a given residue).

One-dimensional sequence features include one-hot encoded predicted secondary structure and relative solvent accessibility from SSpro² and one-hot encoded amino acid sequence.

For statistical potential terms, we used GOAP³ and two versions of DOOP⁴ potential, DOOP-CB which incorporates main-chain atoms and CB atoms, and DOOP-CBCG which incorporates main-chain atoms, CB, and CG atoms. Potential features include per-residue DOOP-CB/DOOP-CBCG potentials, per-residue GOAP potential (in-house implemented), DOOP-CB and GOAP potential averaged on residues within 8Å sphere of a specific residue.

For each protein sequence, we run PSI-BLAST⁵ against UniRef50 to construct multiple sequence alignment (MSA) (e-value 0.001). If there were still not enough sequences in the MSA, sequences were searched again using e-value cutoff of 100.

Distance distributions were predicted by trRosetta⁶ from the MSA, and three (2-8Å, 8-12Å and 12-16Å) distance distribution probabilities were derived from trRosetta prediction. For each residue, these three distance distribution probabilities for residue-pairs within corresponding distance bin were summed up and used as distance features.

Our deep neural network consists of five 1D convolutional blocks which are composed of a convolutional layer with 8 filters and kernel size of 3 with varying dilation rates(1, 2, 4, 8 and 1), a batch normalization layer, an elu layer, a dropout layer with a dropout rate at 0.4. We trained this network using ADAM optimizer with the learning rate of 0.001 and 10^{-4} penalty for the L2 regularization. The 1D-CNN was trained to predict the local IDDT scores of residues. The loss of our deep neural network model is the MSE(Mean Square Error) between predicted local quality and its ground truth(IDDT score). The global accuracy score of a model is derived by averaging the predicted local IDDT scores of residues.

1. Mariani V, Biasini M, Barbato A, Schwede T.(2013). IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29, 2722-2728.
2. Magnan C.N, Baldi P. (2014). SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, 30, 2592-2597.
3. Zhou H, Skolnick J. (2011). GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.* 101, 2043-2502.
4. Chae M.H, Krull F, Knapp E.W. (2015). Optimized distance-dependent atom-pair-based potential DOOP for protein structure prediction, *Proteins*, 83, 881-890.
5. Altschul S.F, Madden T. L , Schaffer A.A, Zhang J, Zhang Z, Miller W, Lipman D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.*, 1997, 25, 3389-3402.
6. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. (2020). Improved protein structure prediction using predicted inter-residue orientations. *Proc. Natl. Acad. Sci. USA*, 117, 1496-1503.

RNA 3D structure prediction by the hybrid methods

Jun Li¹, Sicheng Zhang¹ and Shi-Jie Chen¹

1 - Department of Physics, Department of Biochemistry, and Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA

chenshi@missouri.edu

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:Y*

We used three different methods, Vfold-Pipeline¹, IsRNA^{2,3,4} and RNAJP⁵, to generate RNA 3D structures from the sequence. If structural templates for some motifs in the RNA targets were found in the current PDB database, they would be constrained in the sampling process in our methods.

Methods

Vfold-Pipeline¹: Vfold-Pipeline is a pipeline connecting 2D and 3D structure predictions from the sequence, where 2D structures are predicted using the Vfold2D⁶ method combined with sequence analysis, and 3D structures are predicted using the Vfold3D and VfoldLA methods. Vfold2D is a physics-based model that predicts 2D structures based on the free energies of mismatched base pairs and various RNA loop motifs, including pseudoknots. Vfold3D and VfoldLA methods are based on the assembly of A-form helices with loop and/or motif templates, extracted from the known RNA 3D structures. Up to five structures are obtained through this pipeline, and an energy-minimization is performed for structure refinement.

IsRNA^{2,3,4}: IsRNA is a coarse-grained model for RNA 3D structure prediction for a given 2D structure and sequence. To efficiently sample the conformational space, it performs replica exchange molecular dynamics simulations with the coarse-grained force field built from an iterative simulated reference state approach to decipher the correlations between different structural parameters. The low-energy structures sampled in the simulations are clustered into five groups and the centroid structures in the clusters are chosen as the predicted structures. The all-atom structures are rebuilt based on the coarse-grained structures and an energy-minimization process is performed for structure refinement.

RNAJP⁵: RNAJP is a nucleotide- and helix-level coarse-grained model for RNA 3D structure prediction with a primary focus on junction structures. Given the RNA 2D structure, it performs global sampling of the 3D arrangements of the helices using molecular dynamics simulations with explicit consideration of non-canonical base pairing and base stacking interactions as well as long-range loop-loop interactions. It also clusters the low-energy structures sampled in the simulations into five groups, and then rebuilds the all-atom structures, and finally performs an energy-minimization structure refinement.

Model ranking: When we had more than five structural candidates by the above three 3D structure prediction methods, we ranked them by calculating the AMBER energies in implicit solvent for the RNA structures after energy minimization.

Availability

Vfold2D is available at <http://rna.physics.missouri.edu/vfold2D/index.html>

Vfold-Pipeline is available at <http://rna.physics.missouri.edu/vfoldPipeline/index.html>

IsRNA is available at <http://rna.physics.missouri.edu/IsRNA/index.html>

1. Li, J., Zhang, S., Zhang, D., & Chen, S. J. (2022) Vfold-Pipeline: a web server for RNA 3D structure prediction from sequences. *Bioinformatics*, **38(16)**: 4042-3.
2. Zhang, D., & Chen, S. J. (2018) IsRNA: An iterative simulated reference state approach to modeling correlated interactions in RNA folding. *J. Chem. Theory Comput.*, **14**: 2230-2239.
3. Zhang, D., Li, J., & Chen, S. J. (2021) IsRNA1: de novo prediction and blind screening of RNA 3D structures. *J. Chem. Theory Comput.*, **17**, 1842-1857.
4. Zhang, D., Chen, S. J., & Zhou, R. (2021). Modeling Noncanonical RNA Base Pairs by a Coarse-Grained IsRNA2 Model. *J. Phys. Chem. B*, **125(43)**, 11907-11915.
5. Li, J., & Chen, S. J., (2022) RNAJP: enhanced RNA 3D structure predictions with noncanonical interactions and global topology sampling. Under review.
6. Cheng, Y., Zhang, S., Xu, X., & Chen, S. J. (2021). Vfold2D-MC: A physics-based hybrid model for predicting RNA secondary structure folding. *J. Phys. Chem. B*, **125(36)**, 10108-10118.

Prediction of protein assemblies and ligand binding modes using a combination of ClusPro and Alphafold

Omeir Khan¹, Sergei Kotelnikov², Usman Ghani¹, Dzmitry Padhorny², Dmitri Beglov¹, Sandor Vajda¹, Dima Kozakov²

1 - Boston University, 2- Stony Brook University

midas@laufercenter.org

Key: Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Y_Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N

In the latest CASP-CAPRI round, our group generated models of protein assemblies using a combination of Alphafold-Multimer (AFM), Alphafold2 (AF2), and docking using the ClusPro webserver.¹⁻³ Here, we will describe the methods used for both generating and ranking ensembles of protein—protein complexes. For ligand modeling we have used ClusPro LigTBM approach.⁴

Methods

Assembly Prediction

For assembly prediction, our group utilized a two-stage methodology in which we first generate an ensemble of initial models, which are subsequently provided to Alphafold-Multimer (AFM) as templates for generating “refined” structures of the target complex. Briefly, the protocols used for initial model generation are described below:

Docking Alphafold2 Models with Cluspro (AF2+ClusPro): The structure of each chain in the assembly is independently predicted using the pTM parameter set of AF2. These single-chain predictions are then ranked by the predicted LDDT (pLDDT). The top ranked model for each chain is selected, and low confidence residues (pLDDT < 0.50) are cut from the termini. The trimmed models are then docked using the ClusPro web server.³ All models generated using the “Electrostatic-favored” coefficient set are downloaded and retained for further processing. For antibody and nanobody targets, ClusPro was run in antibody mode.⁵ For homomeric complexes additional models were generated using ClusPro’s multimer docking mode.

Multimer Prediction with Templates (AFM-Temp): An unmodified version of AFM was used to generate 25 models of the target complex. For template searching, the maximum template release date was set to May 14th, 2022.

Multimer Prediction without Templates (AFM-NoTemp): The MMseqs2 API was used to generate multiple sequence alignments (MSAs) for each subunit of the complex.⁶ The pTM

parameter set was then used to generate 5 models of the target assembly. No templates were used in the generation of these models.

Template-based modeling with ClusPro-TBM (TBM): The ClusPro template-based modeling functionality was used to generate templates for a given target-complex. The sequences and stoichiometry of the assembly are given as inputs. Templates were found using a local installation of HHPred, from which the HHblits and HHsearch commands were used to search the uniclust30 and pdb100 databases.⁷⁻⁹ Search results were for hits with > 20% probability and > 20% query sequence coverage. If the stoichiometry of a template matches that of the target, the template is retained and used for model generation in the next step.

The models generated using each of the aforementioned approaches are then refined with AlphaFold-Multimer. The refinement stage is dual purpose, as it can not only improve the quality of template models, but also produce a confidence score for each model that can be used for ranking. For refinement, MSAs were prepared for each subunit using the AFMMseqs2 API.⁶

Ligand Docking

We applied the template-based small-molecule docking algorithm ClusPro LigTBM to build the model of the ligand. If no global template was found, LigTBM was extended to consider local templates of binding pockets in PDB structures containing fragments of the candidate metabolite. Instead of searching for fully homologous receptor-ligand pairs, our approach identifies ligand substructures and matching binding pockets on the target protein surface.

Availability

ClusPro and ClusPro LigTBM are available as webservers that are free for academic and governmental use.

1. Evans,R., O'Neill,M., Pritzel,A., Antropova,N., Senior,A., Green,T., Zidek,A., Bates,R., Blackwell,S., Yim,J., Ronnenberger,O., Bodenstein,S., Zielinski,M., Bridgland,A., Potapenko,A., Cowie,A., Tunyasuvunakool,K., Jain,R., Clancy,E., Kohli,P., Jumper,J. & Hassabis,D. (2022). Protein complex prediction with AlphaFold-Multimer. Preprint at *Biorxiv*.
2. Jumper, J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronnenberger,O., Tunyasuvunakool,K., Bates,R., Zidek,A., Potapenko,A., Bridgland,A., Meyer,C., Kohl,S.A.A., Ballard,A.J., Cowie,A., Romera-Paredes,B., Nikolov,S., Jain,R., Adler,J., Back,T., Petersen,S., Reiman,D., Clancy,E., Zielinski,M., Steinegger,M., Pacholska,M., Berghammer,T., Bodenstein,S., Silver,D., Vinyals,O., Senior,A.W., Kavukcuoglu,K., Kohli,P. & Hassabis,D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**, 583-589
3. Kozakov,D., Hall,D.R., Xia,B., Porter,K.A., Padhorny,D., Yueh,C., Beglov,D. & Vajda,S (2017). The ClusPro web server for protein-protein docking. *Nat. Protoc.* **12**, 255-278
4. Alekseenko,A., Kotelnikov,S., Ignatov,M., Egbert,M., Kholodov,Y., Vajda,S. & Kozakov,D. (2020). ClusPro LigTBM: Automated Template-based Small Molecule Docking. *J. Mol. Biol.* **432**, 3404-3410

5. Brenke,R., Hall,D.R., Chuang,G-Y., Comeau,S.R., Bohnuud,T., Beglov,D., Schuler-Furman,O., Vajda,S., & Kozakov,D. (2012). Application of asymmetric statistical potentials to antibody-protein docking. *Bioinformatics*. **28**, 2608-2614
6. Steinegger,M. & Soding,J. (2017). MMseqs2 enabled sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. **35**, 1026-1028
7. Mirdita,M.,von den Driesch,L., Galiez,C., Martin,M.J., Soding,J. & Steinegger,M. (2016) Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res*. **45**, D170-D176
8. Hildebrand,A., Remmert,M., Biegert,A. & Soding,J. F (2009). Fast and accurate automatic structure prediction with Hhpred. *Proteins*. **77**, 128-132
9. Padhorny,D., Porter,K.A., Ignatov,M., Alekseenko,A., Beglov,D., Kotelnikov,S., Ashizawa,R., Desta,I., Alam,N., Sun,Z., Brini,E., Dill,K., Schueler-Furman,O., Vajda, S. & Kozakov, D. (2020) Cluspro in rounds 38-45 of CAPRI: Toward combining template-based methods with free docking. *Proteins*. **88**, 1082-1090

CoDock: Template-based docking and AI-based scoring using in ligand binding prediction

Ren Kong¹, Xufeng Lu², Shan Chang^{1*}

¹ Institute of Bioinformatics and Medical Engineering, School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou 213001, Jiangsu, China, ² Primary Biotechnology Inc., No. 3 Meishan Avenue, Changzhou 213125, Jiangsu, China.

E-mail: schang@jsut.edu.cn

In CASP15, we participated the category of ligand binding prediction. The template-based docking method and AI-based scoring function are used to predict the binding mode of small molecules or metal ions. The template-based docking method adopts the align algorithm developed by our group, which can accurately identify templates from the structure database.

Method

Template searching

In previous work, a sequence-based template search strategy was applied for protein-protein docking problem ¹. Here for ligand binding prediction, a structure-based 3D align algorithm is developed by our group and used for both of pocket template searching and ligand alignment. For pocket template searching, the CA atoms in protein pocket are set as nodes. For ligand alignment, all atoms in ligand are set as nodes. A set of fully adjacent nodes is defined as a clique. Then, matching is formulated as a graph theoretical problem, which attempts to find the maximum clique between the request structure and the template structure. We use the exhaustive matching algorithm as DOCK program ² to search the most similar template and generate orientation for alignment.

Binding pose prediction

With all the complex templates found by the 3D align algorithm, the macro-molecule structures of potential templates are extracted and compared with the target structure provided by AF2 or disclosed by CASP organizing committee. In some of the target systems, complex templates with high similarity score are found, such as H1114, R1117, H1135, R1136, T1146, T1127, etc. For those ligands identical to ligands in complex template, structure-based alignment is directly used to get the ligand position in the predicted target. For example, the structure of R1117 and metal ions for H1114 are obtained by this way. For those ligands chemically similar with ligands in complex templates, template guided docking protocol are used to obtain the target-ligand complex structure. For target systems with no appreciate complex templates, such as T1181 and T1187, traditional docking is performed by using glide ³.

Pose ranking by AI score

An AI-based scoring is applied to evaluate the results. We use Convolutional Neural Network (CNN) ⁴ to train the scoring function for protein small molecule complex prediction, and the final docking poses are evaluated and ranked by this scoring function.

Acknowledgement

We would like to thank Lianhua Piao, Ying Gao for their helpful discussion. We also want to thank the CASP15 organizer and the experiment specialist who kindly provide the structures for assessments.

1. Kong R, Liu R R, Xu X M, Zhang D W, Xu X S, Shi H, Chang S. Template-based modeling and *ab-initio* docking using CoDock in CAPRI. *Proteins*. 2020; 88(8): 1100-1109.
2. Ewing T J A, Kuntz I D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J Comput Chem*. 1997; 18: 1175-1189.
3. Friesner R A, Banks J L, Murphy R B, Halgren T A, Klicic J J, Mainz D T, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*. 2004; 47: 1739-49.
4. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes D R. Protein-Ligand Scoring with Convolutional Neural Networks. *J Chem Inf Model*. 2017; 57: 942-957.

Benchmarking ColabFold in CASP15

Sergey Ovchinnikov^{1,2}, Martin Steinegger^{3,4,5} and Milot Mirdita³

1 - JHDSF Program, Harvard University, Cambridge, MA 02138, USA, 2 - FAS Division of Science, Harvard University, Cambridge, MA 02138, USA, 3 - School of Biological Sciences, Seoul National University, Seoul, 08826, South Korea, 4 - Artificial Intelligence Institute, Seoul National University, Seoul, 08826, South Korea, 5 - Institute of Molecular Biology and Genetics, Seoul National University, Seoul, 08826, South Korea

so@fas.harvard.edu, martin.steinegger@snu.ac.kr, mmirdit@snu.ac.kr

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:Y*

ColabFold-AlphaFold2 is an enhanced implementation of the accurate protein structure prediction method AlphaFold2. While its main goal was to build upon AlphaFold2 to make it widely accessible, we also implemented a series of performance enhancements: Besides modifications to reduce model compilation time and runtime, we also incorporate the fast MSA search and generation method MMseqs2. For this report, we limit our description, as the details have previously been published¹.

Methods

For monomer prediction, we used the AlphaFold2-ptm model² and for multimeric protein structure prediction we used the AlphaFold-multimer-v2 model³. The MSA was generated following the protocol described in ¹. The sequence and template databases we used were UniRef100 2103 and 2202 (switched to 2202 from target T1119 on), ColabFoldDB 202108, and PDB70 220313. For the AlphaFold runs we used 12 recycles without early-stopping and 3 ensembles. For extremely large oligomeric proteins, we predicted a maximum number of components that could fit into memory and submitted these for the automated submission. Additional, manual parameter tweaks were required for these large models to fit them into system memory (reduced recycles, ensembles, etc.) For the manual-intervention submission (colabfold_human), we used symmetry operations to generate the entire complex. More specifically, we used the make_symmdef_file.pl script from Rosetta⁵. The AlphaFold predictions were ranked using the predicted confidence metrics. For monomers, this was the predicted LDDT (pLDDT) and for multimers the predicted interface TMscore (ipTM).

Results

Mean pLDDT of the non-cancelled targets is 80.85. (1st Quantile: 80.10, Median: 85.21, 3rd Quantile 91.98). Mean N_{eff} value (see HH-suite Userguide section “How is the MSA diversity

Neff calculated?”) for the generated MSA for all 132 chains is 4.217 (1st Quantile: 2.475, Median: 4.350, 3rd Quantile: 5.925). Mean ipTM score of the 20 hetro-oligomeric complexes is 73.10 (1st Quantile: 63.15, Median: 74.35, 3rd Quantile: 83.35). Final results are not yet known.

Availability

ColabFold is free open source software that can be installed locally from <https://github.com/sokrypton/ColabFold> or used online with a web browser through Google Colab at <https://colabfold.com>. The ColabFold databases can be found at <https://colabfold.mmseqs.com>. Submitted predictions, including MSAs and confidence metrics, were uploaded immediately after prediction to CASP15 to <https://wwwuser.gwdg.de/~mmirdit/casp15>.

1. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6), 679-682.
2. Steinegger, M. & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* 35, 1026–1028.
3. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589.
4. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A. W., Green, T., ... & Hassabis, D. (2021). Protein complex prediction with AlphaFold-Multimer. *bioRxiv*.
5. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D., & André, I. (2011). Modeling symmetric macromolecular structures in Rosetta3. *PloS one*, 6(6), e20450.

Template-based RNA structure prediction guided by deep learning predicted distances

Chengxin Zhang^{1,2}, Anna Marie Pyle^{1,2}

1 - Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06511, USA

2 - Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA

zcx@umich.edu

Key: Auto:Y; CASP_serv:Y; Templ:Y; MSA:N; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:N

CoMMiT (Comparative Modeling of RNAs by Multi-Threading), is a fully automated template-based RNA tertiary structure prediction program that uses deep learning-predicted distances for template-based modeling of RNAs.

Methods

CoMMiT performs sequence and secondary structure alignment between an input RNA sequence and template structures in the PDB database using five threading methods: RATEs, MapAlignG, LocARNA¹, LaRA² and Foldalign³. Among these five methods, RATEs and MapAlignG are threading methods developed in-house. RATEs aligns both sequence and secondary structure while MapAlignG aligns secondary structure only. Structure fragments derived from templates are then assembled into a full-length structure by simulated annealing Monte Carlo (SAMC) simulation. The simulation is guided by a hybrid energy function consisting of template-derived distance restraints, statistical energy terms⁴, secondary structure formation terms defined by CSSR⁵ and nucleotide-nucleotide distance constraints predicted by a deep residual convolutional neural network (ResNet) using features derived from the input sequence. Conformations generated by the simulation are grouped by the SPICKER density-based clustering algorithm⁶ and the final structure model is derived from the centroid of the largest cluster.

1. Will, S., Joshi, T., Hofacker, I. L., Stadler, P. F., & Backofen, R. (2012). LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA*, 18(5), 900-914.
2. Winkler, J., Urgese, G., Ficarra, E., & Reinert, K. (2022). LaRA 2: parallel and vectorized program for sequence–structure alignment of RNA sequences. *BMC bioinform*, 23(1), 1-22.
3. Sundfeld, D., Havgaard, J. H., de Melo, A. C., & Gorodkin, J. (2016). Foldalign 2.5: multithreaded implementation for pairwise structural RNA alignment. *Bioinformatics*, 32(8), 1238-1240.
4. Zhang, T., Hu, G., Yang, Y., Wang, J., & Zhou, Y. (2020). All-atom knowledge-based potential for RNA structure discrimination based on the distance-scaled finite ideal-gas reference state. *Journal Comput Biol*, 27(6), 856-867.
5. Zhang, C., & Pyle, A. M. (2022). CSSR: assignment of secondary structure to coarse-grained RNA tertiary structures. *Acta Crystallogr D*, 78(4).

6. Zhang, Y., & Skolnick, J. (2004). SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem*, 25(6), 865-871.

Predicting protein-ligand binding with Convex-PL, KORP-PL, AutoDock Vina, and VinaCPL

M. Kadukova^{1,2} and S. Grudinin¹

¹Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble France, ²Astex Pharmaceuticals, Cambridge UK

maria.kadukova@astx.com, sergei.grudinin@univ-grenoble-alpes.fr

Key: *Auto:N; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:N; EMA:Y; MD:N*

In CASP15, we assessed our docking and scoring techniques specifically developed for small-molecule docking. These include an in-house docking engine VinaCPL¹ and four types of docking scoring functions Convex-PL², Convex-PL^{R 1}, KORP-PL^w, and KORP-PL³.

Methods

We have submitted predictions for H1114, T1105v1, T1118, T1118v1, T1124, T1146, and T1152 targets.

We obtained protein models from AlphaFold2 and BAKER-SERVER server predictions. Ideal point-group symmetry was set in targets H1114, T1124, and T1152 using the AnAnaS tool⁴⁻⁵. For two targets, T1105 and T1118/T1118v1, we simulated continuous structural heterogeneity of the binding pocket using nonlinear principal component analysis with the NOLB tool⁶.

Depending on the target, binding pockets were predicted with Fpocket⁷, using the UniProt⁸ annotations, or from homology to the known complexes from the PDB. 3D structures of carbohydrate ligands in T1146 and T1152 were obtained from the PDB. In T1124, we have used both RDKit⁹ and PDB complex 7clf as a source of ligand structures. Other ligands were generated with RDKit. For the macrocycle in T1118, we clustered 1,000 of the RDKit's conformations¹⁰. For all targets, we then ran docking of all starting ligand conformations to all available pockets with AutoDock Vina¹¹ and in-house VinaCPL¹ and re-scored the docking poses with Convex-PL² (Convex-PL team), Convex-PL^{R 1} (Convex-PL-R team), KORP-PL^{w 3} (Grudinin team) and KORP-PL³ (KORP-PL team). We then clustered the ligand poses of each model with 1 or 2 Å RMSD cutoff. Finally, we submitted the top-5 models based on the ligand docking score, keeping top-5 ligand poses per model.

Availability

Our methods are available on our website at <https://team.inria.fr/nano-d/software/>.

1. Kadukova,M., Chupin,V., & Grudin,S. (2021). Convex-PLR-Revisiting affinity predictions and virtual screening using physics-informed machine learning. bioRxiv.
2. Kadukova,M., & Grudin,S. (2017). Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization. *J. Comput. Aided Mol. Des.* 31, 943-958.
3. Kadukova,M., Machado,K.D.S., Chacón,P., & Grudin,S. (2021). KORP-PL: a coarse-grained knowledge-based scoring function for protein–ligand interactions. *Bioinformatics* 37, 943-950.
4. Pagès,G., Kinzina,E., & Grudin,S. (2018). Analytical symmetry detection in protein assemblies. I. Cyclic symmetries. *J. Struct. Biol.*, 203(2), 142-148.
5. Pagès,G., & Grudin,S. (2018). Analytical symmetry detection in protein assemblies. II. Dihedral and cubic symmetries. *J. Struct. Biol.*, 203(3), 185-194.
6. Hoffmann,A., & Grudin,S. (2017). NOLB: Nonlinear rigid block normal-mode analysis method. *J. Chem. Theory Comput.*, 13(5), 2123-2134.
7. Le Guilloux,V., Schmidtke,P., & Tuffery,P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC bioinform.* 10, 1-11.
8. UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic acids res.* 47, D506-D515.
9. Landrum,G. <http://www.rdkit.org>, RDKit: Open-source cheminformatics.
10. Wang,S., Witek,J., Landrum,G.A., & Riniker,S. (2020). Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *J. Chem. Inf. Model.* 60, 2044-2058.
11. Trott,O., & Olson,A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455-461.

Deep Learning based protein structure prediction model

Zhigang Sun, Kexin Zhang, Fenglei Li, Anqi Pang and Jingyi Yu

ShanghaiTech University

pangaq@shanghaitech.edu.cn

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

Coqualia is a hybrid multi-model deep learning network. The whole architecture uses metagenomic data to obtain multi-sequence alignments (MSA) with different searching methods. According to different training configurations, a total of 15 sets of parameters were obtained. Finally, we leverage averaged predicted local distance distribution test (pLDDT) value to rank multiple predictions.

Methods

The whole model architecture includes data processing, feature embedder, evoformer, structure module and structure refinement. In data processing part, we used GPU accelerated techniques to build MSA database for NCBI and Mgnify. In feature Embedder part, we search MSA from above metagenomic data and build different features with different methods. In evoformer and structure module part, we used alphafold2¹ architecture. For all self-attention layers, we used dynamic axial parallelism technique to save GPU memory and accelerate forwarding and backpropagation speed. In structure refinement, we used OPENMM² with CUDA platform. We got several hundred models for each target with different MSA features and models and rank them by averaged pLDDT, then we select top 5.

Results

For most of CASP targets, we can get reliable predictions based on pLDDT values.

Availability

Coming soon.

1. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
2. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang LP, Simmonett AC, Harrigan MP, Stern CD, Wiewiara RP, Brooks BR, Pande VS. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol*. 2017 Jul 26;13(7):e1005659.

An AI platform oriented to prediction of high-level structures of biomacromolecules

Carlos A. Del Carpio Muñoz

Choju-Medical Institute, Fukushima Hospital. Noyori-cho, Yamanaka 19-14. Toyohashi-City, Aichi-ken 441-8124, Japan

delcmca@gmail.com

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA:N; MetaG:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:N; MD:Y*

We report on the improvement of our original bioinformatics platform oriented to predict the 3D and 4D structure of proteins and RNA's^{1,2} that has been developed by our group to include machine learning based algorithms. While the hitherto methodologies developed by our group enable the prediction of protein folding patterns as well as local substructures and domains, prediction of long-range interactions among amino acids can be improved by exploiting structural information in protein structural knowledge bases. In past CASP rounds we have combined classical homology methods with our genuine method based on spectral analysis of the sequences of the amino acids represented by their physicochemical properties. While this methodology assists in determining the overall putative folding family of a target sequence, amino acid and atomic long-range interactions are critical in predicting the close to native structure. On the other hand, CASP14 results produced by Alpha-Fold³ have been highly accurate owing its success to a sophisticated treatment of the stereo-chemical features of proteins. Consequently, consideration of this aspect led us to introduce to a limited extent Alpha-Fold predicted structures in the CASP15.

This has been especially the case in the prediction of protein quaternary structure or protein complexes. Thus, when computationally possible, besides the structures predicted by our original procedure, Alpha-Fold predicted structures were also used to predict these higher structures. Energy minimization and molecular dynamics to correct ill placed atoms were used to rank the final structures. The process has been handled using our system for the assessment of complex structures MIAx⁴, the main characteristics of which consist on the prediction of binding sites and a new protocol for the evaluation of the plausibility of contact regions.

Method

The multi-platform automatic system proposed starts with the selection of the best homologs for the sequence in question with orthodox methodologies. When no homologs are found for the target, the process shifts to the spectral analysis of the sequences and homologs from this point of view are output that are analysed in a piece-wise manner with the target sequence. Then the required 3D sequence for the target structure is built by the platform. Loop and structural stability analysis is then carried out with our system for protein stability analysis. Molecular

dynamics and other minimization processes are then applied to the most plausible candidate structures which are then ranked according to the energetic characteristics.

On the other hand, protein assemblies are predicted using the system MIA³ for protein interaction assessment, which consists on protein interaction region prediction and docking of the structures. For hetero multimer structure prediction, prediction of the binding sites was performed based on a new way to assess the order of interaction of the subunits⁴.

1. Del Carpio, C. A. & Yoshimori, A. (2002). Fully automated protein tertiary structure prediction using Fourier transform spectral methods. *Protein Structure Prediction: Bioinformatics*, University of California, International University Line.
2. Del Carpio, C. A. & Carbajal, J. C. (2002). Folding pattern recognition in proteins using spectral analysis methods. *Genome Inform* 13, 163-72.
3. Jumper J. et al (2021). Highly Accurate Protein Structure Prediction with AlphaFold, *Nature* 596(2021)
4. Del Carpio, C.A., Ichiishi E. (2017). Inference of Protein Multimeric Complex Dynamic Order of Formation: An Active Region Recognition Based Approach. *International Journal of Genomics and Data Mining* 2017, 1.
5. Del Carpio, C. A., Ichiishi, E., Yoshimori, A. & Yoshikawa, T. (2002). A new paradigm for modeling biomacromolecular interactions and complex formation in condensed phases. *Proteins: Structure, Function, and Genetics* 48, 696-732.

Manual trial and error using AlphaFold 2 and conventional ligand docking

Kazuki Yamamoto^{1,2}, Yoshitaka Moriwaki¹, Motoyuki Hattori³,

Keisuke Yanagisawa², and Masahito Ohue²

1 – The University of Tokyo, 2 – Tokyo Institute of Technology, 3 – Fudan University

kazuki@ric.u-tokyo.ac.jp

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

To test whether predicted structures generated by AlphaFold2¹ (AF2) would be useful for ligand docking, we worked on the ligand prediction targets. Basically, the predictions were made using conventional methods and target-specific tricks.

Methods

The predictions were made according to the flow chart in Figure 1.

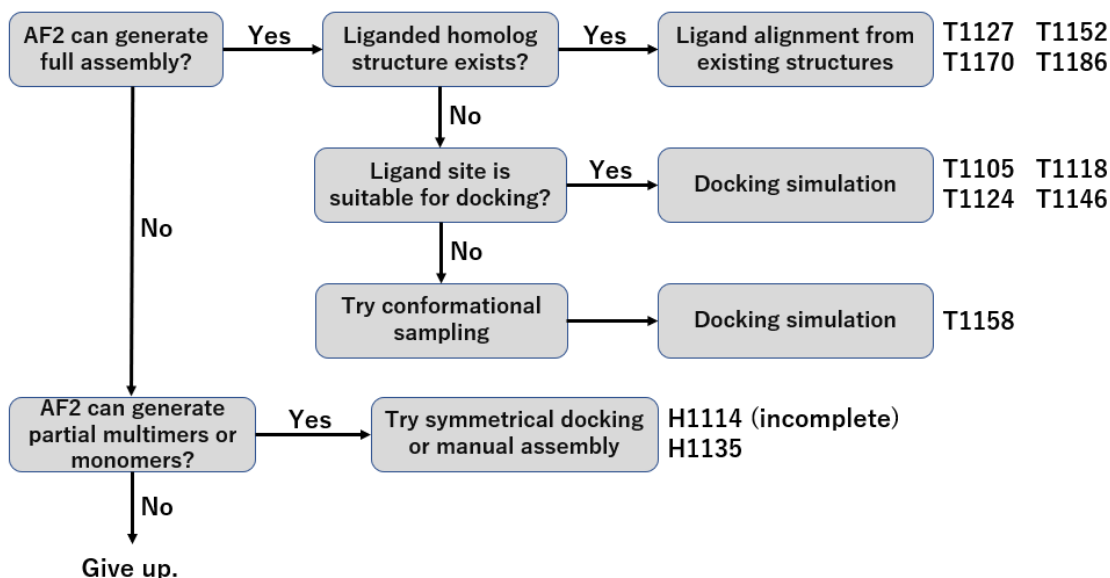


Figure 1. Flow chart of modeling process.

Protein modeling: AF2 was executed via localcolabfold²⁻⁴. When needed, M-ZDOCK⁵ was used for symmetrical assembly.

Ligand docking: Small molecule ligands were docked using conventional docking software (Molegro Virtual Docker 7.0⁶). When needed, manual positioning was performed using PyMOL⁷.

Molecular Dynamics: For H1135, residues 63–77, 111–116, and 156–165 are estimated to be candidate regions to bind potassium ions according to insights of a structural biologist (M. Hattori). A 10-ns production run with AMBER 22⁸ were then conducted for the 12-mer H1135 in complex with 27 K⁺ complex to evaluate their stability. (The ff14SB force field and ion parameters created by Joung and Cheatham were used for the protein and the K⁺ ion, respectively.) After the simulation, stable K⁺ ions bound to the protein were submitted as the answer.

Results

H1114: A4 was modelled reasonably by AF2. B2C2 was modelled reasonably by AF2 and could be symmetrically assembled by M-ZDOCK. A4 + (B2C2)_{x4} could be assembled manually. Ligand sites could not be identified. (incomplete)

H1135: A3B1 was modelled by AF2. Then M-ZDOCK to assemble 12-mer. K⁺ binding site was identified through visual inspection by a structural biologist and optimized by molecular dynamics.

T1105: Removed His-tag. Then AF2. Ligand docked reasonably.

T1118: Iron ions substituted with bridging carbons. Then docked. After docking, iron ions were manually positioned.

T1124: Docked to AF2 model. A docking pose appropriate for enzyme reaction was selected manually.

T1127: AF2, then docking.

T1146: AF2. GlcNAc-MurNAc was docked. Then the ligand was separated as two GlcNAcs manually.

T1152: AF2. The ligand was placed using template docking, aligning 5C8Q.PDB.

T1158: AF2, then docking. The open/closed conformations of the protein were controlled by removing long unstructured interdomain region. The ligand v2 was placed aligning LTC4 in 5UJA.PDB. In v4, AF2 model was morphed to 6S7P.PDB, then the ligands were placed aligning the ligands in 6S7P.pdb.

T1170: AF2. The ligand was aligned from 1J7K.PDB.

T1186: AF2. The ligand was placed aligning 5KMW.PDB.

Acknowledgements

We would like to thank Masakazu Sekijima for providing us with computational resources of his laboratory in Tokyo Institute of Technology.

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589.
2. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, 1-4.
3. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A. W., Green, T., ... & Hassabis, D. (2021). Protein complex prediction with AlphaFold-Multimer. *BioRxiv*.
4. Moriwaki, Y. LocalColabFold. <https://github.com/YoshitakaMo/localcolabfold>
5. Pierce, B., Tong, W., & Weng, Z. (2005). M-ZDOCK: a grid-based approach for C n symmetric multimer docking. *Bioinformatics*, 21(8), 1472-1478.
6. Thomsen, R., & Christensen, M. H. (2006). MolDock: a new technique for high-accuracy molecular docking. *Journal of medicinal chemistry*, 49(11), 3315-3321.
7. Schrödinger, L., & DeLano, W. (2020). PyMOL. <http://www.pymol.org/pymol>
8. D.A. Case, H.M. Aktulga, ... & P.A. Kollman (2022), Amber 2022. University of California, San Francisco.

RNA 3D structure prediction by DeepFoldRNA in CASP15Robin Pearce^{1*}, Gilbert S. Omenn^{1,3}, Peter Freddolino^{1,2}, Yang Zhang^{1,2}

¹Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, United States of America; ²Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan, United States of America; ³Departments of Internal Medicine and Human Genetics and School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America.

*To whom correspondence should be addressed. E-mail: robpearc@umich.edu

Key: Auto:Y; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y

The RNA tertiary structure prediction for the DF_RNA group in CASP15 is based on the DeepFoldRNA pipeline ¹, which uses deep self-attention networks to predict inter-residue distance and orientation maps as well as backbone pseudo-torsion angles for a query sequence, where full-length structure models are generated through L-BFGS optimization of the backbone and base torsion angles using a potential constructed from the predicted restraints. The pipeline consists of 3 consecutive steps. First, starting from the query sequence, a set of multiple sequence alignments (MSAs) are created by rMSA ², which iteratively searches the query through the nt³, RNACentral⁴, and Rfam⁵ sequence databases using blastn⁶, nhmmer⁷, and CMsearch ⁷. Additionally, PETfold⁸ is used to predict the secondary structure information from the MSAs.

In the second step, the generated MSAs and secondary structure predictions are used as the input to the deep self-attention networks to predict an ensemble of restraints, including pairwise distance and orientation maps that consider the interactions between multiple sets of backbone and base atoms as well as the backbone pseudo-torsion angles. Briefly, three embeddings are used by the network: the MSA, pair, and sequence embeddings. The MSA embedding captures the evolutionary information contained in the alignment of homologous sequences, the pair embedding captures the pairwise spatial relationships between each nucleic acid, and the sequence embedding captures the network's representation of the query sequence. The MSA embedding is first initialized from the one-hot encoded MSA and processed by multiple rounds of row-wise and column-wise self-attention, while the pair embedding is initialized from the predicted secondary structure and paired sequence information and is processed using a triangular self-attention scheme⁹. Interaction is encouraged between the two representations by biasing the MSA self-attention using the pair embedding and by updating the pair embedding using the outer product mean of the MSA embedding. Next, the sequence embedding is extracted from the first row in the processed MSA embedding and is similarly refined using multiple self-attention layers. The binned pairwise distance and orientation maps are predicted from a linear projection of the pair embedding, while the binned backbone pseudo-torsion angles are predicted from the sequence embedding. Lastly, the network generates predicted error maps from the final pair representation, which are used to determine the optimal

ensemble weights for the restraint sets generated by the different network variants and parameter files.

In the third step, full-length structure models are created using L-BFGS optimization based on a potential derived from the predicted restraints. The potential is constructed by taking the negative log-likelihood of the binned probability distributions for each restraint and made continuously differentiable by fitting a smooth curve through the histogram distributions using cubic spline interpolation. The optimization is carried out on the backbone pseudo-torsion angles (η and θ) and the base torsion (χ). A pool of models is generated by taking different restraint ensembles, where the model with the lowest predicted error is selected as the first model and clustering is used to introduce conformational diversity for models 2-5 by selecting the largest cluster centers as the representative models. Lastly, the full-atomic structures are refined using SimRNA¹⁰ and QRNAS¹¹ with restraints on the backbone atom positions. The modeling procedure is fully automated.

Availability

<https://zhanggroup.org/DeepFoldRNA/>

1. Pearce R, Omenn GS, Zhang Y. De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning. bioRxiv. 2022:2022.05.15.491755.
2. Zhang C, Zhang Y, Pyle AM. rMSA: database search and multiple sequence alignment generation to improve RNA structure modeling. ISMB. 2022:In press.
3. Coordinators NR. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2014;42:D7-D17.
4. Consortium RN. RNACentral 2021: secondary structure integration, improved sequence search and new member databases. Nucleic acids research. 2021;49:D212-D20.
5. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic Acids Res. 2003;31:439-41.
6. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997;25:3389-402.
7. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics. 2013;29:2487-9.
8. Seemann SE, Menzel P, Backofen R, Gorodkin J. The PETfold and PETcofold web servers for intra- and intermolecular structures of multiple RNA sequences. Nucleic Acids Research. 2011;39:W107-W11.
9. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583-9.
10. Boniecki MJ, Lach G, Dawson WK, Tomala K, Lukasz P, Soltysinski T, et al. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. Nucleic acids research. 2016;44:e63.
11. Stasiewicz J, Mukherjee S, Nithin C, Bujnicki JM. QRNAS: software tool for refinement of nucleic acid structures. BMC Struct Biol. 2019;19:5.

Protein 3D Structure Prediction with DeepFold

Jong-Hyun Won^{1,2*}, Jae-Won Lee^{1,2*}, Seonggwang Jeon^{1,2*}, Yujin Choo^{1,2*}, Yubin Yeon^{1,2*}, Jin-Seon Oh^{1,2*}, Minsoo Kim^{3*}, SeonHwa Kim⁴, InSuk Joung⁵, Cheongjae Jang¹, Sung Jong Lee⁶, Tae Hyun Kim¹, Kyong Hwan Jin⁴, Giltae Song⁷, Eun-Sol Kim¹, Jejoong Yoo³, Eunok Paek¹, Yung-Kyun Noh^{1,8†}, Keehyoung Joo^{2†}

1 - Department of Computer Science, Hanyang University, Korea, 2 - Center for Advanced Computation, Korea Institute for Advanced Study, Korea, 3 - Department of Physics, Sungkyunkwan University, Korea, 4 - Department of Electrical Engineering and Computer Science, DGIST, Korea, 5 - Standigm Inc. Seoul, Korea, 6 - Basic Science Institute, Changwon National University, Korea, 7 - School of Computer Science and Engineering, Pusan National University, Korea, 8 - School of Computational Sciences, Korea Institute for Advanced Study, Korea

* Equal contribution

† Corresponding authors: Keehyoung Joo (newton@kias.re.kr), and Yung-Kyun Noh (nohyung@hanyang.ac.kr)

Key: Auto:N; CASP_serv:Y; Templ:Y; MSA:Y; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:Y

For CASP15 experiment, we have developed a new pipeline, DeepFold based on AlphaFold2 (AF2)¹, focusing on the more precise backbone and the sidechain prediction. We modified loss functions in AF2 and introduced additional loss functions for the sidechain confidence and the secondary structure prediction. We have trained the DeepFold models on a training dataset built from PDB with 40% sequence identity level using sequence clustering. We updated the MSA/template features by searching the sequence databases and re-aligning (by CRFalign²) between the target and the templates. Protein 3D structures were generated with DeepFold network models (DFolding-server), followed by the refinement using the MD simulation protocol (DFolding-refine) and the re-optimization using the conformational space annealing^{3,4} (DFolding).

Loss functions: Torsion angle loss in AF2 was modified, improving the reliability of sidechain prediction by providing sequential conditioning. In addition, we introduced a new sidechain confidence measure to predict the reliability of sidechains. FAPE (Frame Aligned Point Error) loss was modified by giving distance-dependent weights to residue pairs following the predicted distogram information (closer residues are emphasized). Finally, a secondary structure loss was introduced, which measures the cross-entropy loss between the 8-state secondary structure prediction and its ground truth.

Dataset and training: The training dataset containing about 31k (20k for fine-tuning) protein chains was built by CD-HIT sequence clustering using the latest PDB data at 40% sequence identity. For training, we cropped the input sequences to 256 and 386 residue sizes as in AF2. The 256-crop dataset was used for initial training, and the 386-crop was used for fine-tuning. We built a training system using uni-fold⁵ (a trainable modification of AF2) with the following protocols: all models were trained from AF2 parameters. The parameters were

optimized using the new datasets and the modified loss functions. In total, twelve different models were generated. Six were trained in the transfer-learning style, utilizing the fixed AF2 representation by freezing the evoformer block and updating the parameters only in the structure-module. The remaining six were trained without freezing.

MSA and template feature processing: We prepared additional features by searching databases. For MSA features, we used HHpred⁶, Kalign⁷ and HHblits⁸ outputs to build 4 different alignments. Then, we replaced template features by using CRFalign method, which is based on conditional random fields for searching best templates, and the optimal alignments between the target sequence and the templates.

Ranking the prediction outputs: DeepFold pipeline generates about 50 different protein structures for each target sequence. The generated structures were clustered by the hierarchical clustering method based on the TM-score and the five best structures in terms of the plddt score. In the case when plddt is larger than 0.85, the weighted average of the plddt and the sidechain confidence score in each cluster was used for selecting models for submission as DFolding-server.

Multimer prediction: Multimer structures were generated by AF2Complex^{9,10} using non-paired MSA between protein chains for the features generated by the DeepFold pipeline. For large complex targets, we generated sub-complexes by domains and combined them using the Modeller program¹¹. The multimer models were clustered into five based on interface RMSD, and we selected the model with the highest interface score in each cluster.

Refinement: Selected five models were further refined by restrained molecular dynamics (MD) simulations. We applied restraint forces to the backbone to prevent extreme conformational change, which also reduced the simulation time. Furthermore, we used pairwise distance restraints potential between alpha carbons giving the force constant in proportion to the plddt of the structure. The MD trajectory was averaged, and the structure was energy-minimized for submission as DFolding-refine.

DFolding as a human prediction protocol: we have applied the global optimization method of conformational space annealing (CSA)^{3,4} to the full atom force field with distance restraints and sidechain torsion restraints generated by the DeepFold pipeline.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub). We thank Korea Institute for Advanced Study for providing computing resources (KIAS center for Advanced Computation Linux Cluster) for this work. The authors would like to acknowledge the support from AI One Team – GPU Server Infrastructure Build for Large Scale Korean Language Model Development project funded by KT (KT award B210001432).

Availability

A github repository for DeepFold is getting ready and will be opened later.

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., and Potapenko, A. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583-589.
2. Lee, S.J., Joo, K., Sim, S., Lee, J., Lee, I.-H., and Lee, J. (2022). CRFalign: A Sequence-structure alignment of proteins based on a combination of HMM-HMM comparison and conditional random fields. *Molecules* 27, 3711.
3. Joung, I., Kim, J.Y., Gross, S.P., Joo, K., and Lee, J. (2018). Conformational Space Annealing explained: A general optimization algorithm, with diverse applications. *Computer Physics Communications* 223, 28-33.
4. Joo, K., Lee, J., Lee, S., Seo, J.H., Lee, S.J., and Lee, J. (2007). High accuracy template based modeling by global optimization. *Proteins: Structure, Function, and Bioinformatics* 69, 83-89.
5. Li, Z., Liu, X., Chen, W., Shen, F., Bi, H., Ke, G., and Zhang, L. (2022). Uni-Fold: An Open-Source Platform for Developing Protein Folding Models beyond AlphaFold. *bioRxiv*.
6. Söding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research* 33, W244-W248.
7. Lassmann, T., and Sonnhammer, E.L. (2005). Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC bioinformatics* 6, 1-9.
8. Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* 9, 173-175.
9. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A.W., Green, T., Židek, A., Bates, R., Blackwell, S., and Yim, J. (2021). Protein complex prediction with AlphaFold-Multimer. *BioRxiv*.
10. Gao, M., Nakajima An, D., Parks, J.M., and Skolnick, J. (2022). AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nature communications* 13, 1-13.
11. Šali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology* 234, 779-815.

Deep Local Analysis Ranker Server for estimating inter-subunit interfaces accuracy in multimeric complexes

S. Grudinin^{1*}, Y. Mohseni Behbahani^{2*}, E. Laine^{2*} and A. Carbone²

1 - Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France; 2 - Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France

* - Equal contributions

sergei.grudinin@univ-grenoble-alpes.fr, yasser.mohseni_behbahani@sorbonne-universite.fr,
elodie.laine@sorbonne-universite.fr, alessandra.carbone@lip6.fr

Key: Auto:Y; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N

We have recently developed Deep Local Analysis (DLA)- Ranker, a new deep learning framework for discriminating near-native complex conformations from incorrect ones by exploiting local 3D environments around interfacial residues¹. DLA-Ranker applies 3D convolutions to interfaces represented as sets of locally oriented cubic volumetric maps. In CASP15, we have set up a fully automated server based on DLA-Ranker for assessing the quality of inter-subunit interfaces in multimeric complexes.

Methods

Starting from a candidate conformation, we detected its interfacial residues using FreeSASA algorithm² (with a probe radius of 1.4 Å) as those displaying a change in solvent accessibility between the free (isolated) partners and the complex. We associated each interfacial atom with a feature vector of length 6 one-hot encoding its chemical element (O,C,N or S) and the partner to which it belonged (either “receptor” or “ligand”)¹. We computed a density function from the atomic 3D coordinates and feature vectors, and we mapped it to grids of 24x24x24 voxels of side 0.8 Å. To build the maps, we adapted the method proposed in Ornate³. Each map is centered on an interfacial residue and it is oriented by defining a local frame based on the common chemical scaffold of amino acid residues in proteins. Thanks to this local frame definition, the map not only is invariant to the candidate conformation’s initial orientation but also provides information about the atoms and residues relative orientations.

DLA-Ranker takes as input a cubic volumetric map and outputs a score between 0 and 1 reflecting the probability of finding this local 3D environment in a near-native complex conformation¹. The DLA-Ranker architecture comprises three 3D convolutional layers, a max pooling layer and three fully connected layers. To avoid overfitting, we used 40%, 20% and 10%

dropout regularization on the input, first and second layers of the fully connected subnetwork, respectively. The loss function is the binary cross-entropy measuring the difference between the probability distribution of the predicted output and the given label (0 or 1).

Depending on the location of the residues at the interface, their geometrical and physico-chemical environments are expected to be very different. For instance, the map computed for a residue deeply buried in the interface will be much more dense than that computed for a partially solvent-exposed residue at the rim. This motivated us to explicitly give some information to the network about the location of the input residue at the interface. To do so, we exploited the support-core-rim (S-C-R) classification proposed by Levy⁴. We encoded the input residue's structural class (either S, C, or R) as a one-hot vector which we concatenated to the embedding derived from the convolutional layers¹.

Results

For the CASP15 challenge, we trained DLA-Ranker on 449,158 candidate complex conformations generated by HADDOCK⁵ for 142 dimers from the Docking Benchmark version 5 (BM5)⁶. On average, each target complex has ~230 near-native conformations and ~2,932 incorrect ones, according to CAPRI criteria⁷. In case of multimeric complexes comprising more than two subunits, we defined an interface for each subunit. More precisely, in the cubic volumetric map centered on residue i from subunit j , we labeled the atoms belonging to subunit j as “receptor” and the atoms from any other subunit as “ligand”. We did not consider interfaces with less than 5 residues. For each candidate complex conformation, we submitted to CASP the scores computed by DLA-Ranker for the interfacial residues we detected in the complex, and also a global interface quality score obtained by averaging over all per-residue scores.

Availability

The software and model parameters are available at: <http://gitlab.lcqb.upmc.fr/dla-ranker/DLA-Ranker/tree/CASP15>. DLA-Ranker can be run on Linux, MacOS, and Windows. We recommend running it on GPUs.

1. Mohseni Behbahani, Y., Crouzet, S., Laine, E., Carbone, A (2022). Deep Local Analysis evaluates protein docking conformations with locally oriented cubes, *Bioinformatics*, btac551.
2. Mitternacht S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Res*. 2016 Feb 18;5:189. doi: 10.12688/f1000research.7931.1. PMID: 26973785; PMCID: PMC4776673.
3. Pagès G. et al. (2019) Protein model quality assessment using 3D oriented convolutional neural networks. *Bioinformatics*, 35, 3313–3319.

4. Levy E.D. (2010) A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J. Mol. Biol.*, 403, 660–670.
5. Dominguez C. et al. (2003) HADDOCK: a protein protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, 125, 1731–1737.
6. Vreven T. et al. (2015) Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *J. Mol. Biol.*, 427, 3031–3041.
7. Lensink M.F. et al. (2017) Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins*, 85, 359–377.

DMPfold3 : Minimalist models for end-to-end protein structure prediction

D.T. Jones

Department of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom

d.t.jones@ucl.ac.uk

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; MetaG:Y; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:N; MD:N*

In CASP15, a new version of the DMPfold method^{1,2} for tertiary structure prediction was used. DMPfold3 is aimed at being as simple as possible and using only standard machine learning algorithms to achieve a competitive level of protein structure prediction accuracy. The idea being to try to explore end-to-end prediction methods that do not employ extensive domain-specific engineering, hopefully whilst still producing good enough results. This will offer a much lower bar of entry to exploring these kinds of models (DMPfold³ is only a few hundred lines of code), and also allows generic technological improvements in language models, such as new transformer models, to be quickly tested in the protein structure domain.

Methods

Multiple sequence alignments (MSAs) were built using searches against the latest UniRef30 databases available at the time of target release. Where this retrieved fewer than 1000 hits, deeper MSAs were built using searches of EBI MGnify, NCBI Transcriptome shotgun assembly (TSA), MetaEuk and IMG sequence databases, each time building a list of putative hits and using these as a custom database for a further search.

The MSA is used as input to a stack of axial attention blocks with the coordinates generated simply by projection from the embedding dimension (384-D) down to 3-D (alpha carbon coordinates). No final structure module or explicit pairwise (“distogram”) representation is used in the model. As DMPfold is aimed solely at tertiary structure prediction, for multimeric CASP targets, the initial single chain C-alpha trace coordinates from DMPfold3 were used as input templates for AlphaFold2-Multimer3 modelling.

Results

A total of 111 models were submitted for 95 targets.

Availability

DMPfold3 will be made available on the PSIPRED GitHub page (<https://www.github.com/psipred>) under a permissive license.

1. Greener, J.G. et al. (2019). Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat. Commun.* 10, 3977.

2. Kandathil, S.M. et al. (2022). Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterised proteins. PNAS 119(4) e2113348119.
3. Evans, R. et al. (2021). Protein complex prediction with AlphaFold-Multimer. bioRxiv 2021.10.04.463034.

Elofsson group using AlphaFold2 and MolPC in CASP15

Arne Elofsson¹, Patrick Bryant¹, Petras Kundrotas¹, Aditi Shenoy¹, Wensi Zhu¹,
Gabrielle Pozzatti¹, Claudio Mirabello²

1 - Dep of Biochemistry and Biophysics and Science for Life Laboratory, Stockholm University;

2 - Dept of Physics, Chemistry and Biology, NBIS (National Bioinformatics Infrastructure Sweden), Science for Life Laboratory, Linköping University

arne@bioinfo.se

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y. MetaG:Y Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:N*

We (group Elofsson #320) have based our predictions on the default AlphaFold pipelines (monomer and multimer) submitted by our server as groups NBIS-AF2-standard (#270) and NBIS-AF2-multimer (#390). The predictions were manually examined. We deemed the server prediction accurate and submitted those for 62/81 targets. For four targets, we used metagenomics sequences to provide better alignments (we tried in a few more cases, but there it did not seem to help). We ran AlphaFold2 with extra recycles for five targets, and for five targets (one overlapping with the recycles), we performed a manual ranking of the models. Finally, for 6 large (multichain) targets, we used a manual version of MolPC¹ to create large complexes.

Methods

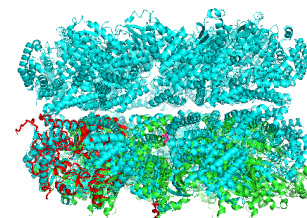
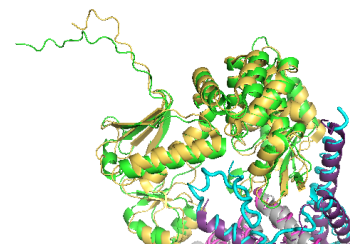
AlphaFold2 is run as the server using the default pipelines.

Metagenomic sequences were searched using jackhmmer² against the MGnify³ database for targets T1123, T1130, T1131, T1154, and T1179 (no hits). For targets T1173, T1174, T1180, T1181, and T1184, 25 recycles were used, while for targets H1171, H1172, T1174, T1179, and T1192, the predicted models were reranked manually. Finally, large targets H1111, H1114, H1115, H1135, H1137, and T1169 were manually built using superpositions of smaller subcomponents, following the methodology of MolPC¹.

Results

Models where the structure is known.

At the time of writing, the only target where we submitted a non-standard AF2 prediction, and the PDB file is available is H1111; these predictions were wrong. Here we predicted the structure of two or three chains and then built the 27-mer from these subcomplexes. The 27 chains are present in the final structure, forming a nice 9-fold symmetric circle. But, a significant part of the 27 chains was not seen.



Modelling only the parts seen in the final model predicts the repeating unit accurately, and using only a 9-mer of chain A makes a nice symmetric structure that can be somewhat superposed to the correct structure, i.e. this indicates that a MolPC-like approach could have worked if a careful selection of what was modelled was made, but no such method exists yet.

Models where the structure is not yet known.

Below, we describe a few manual examples, although we do not know if our predictions are correct. Left side original model, right side submitted submission. Colouring by pLDDT.

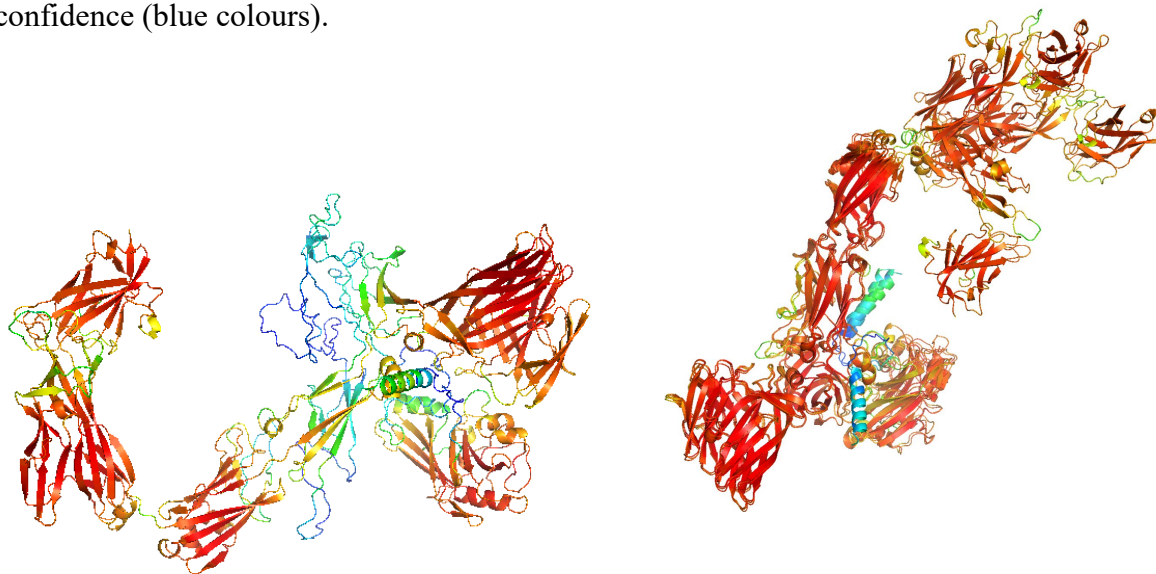
H1135 (A9B3)

Running H1135 through AlphaFold multimer results in a model where all the B chains are not in contact with the A chains. Therefore, we made a model of A3B1 and then superposed this on an A9 model.

H1137 (A1B1C1D1E1F1G2H1I1)

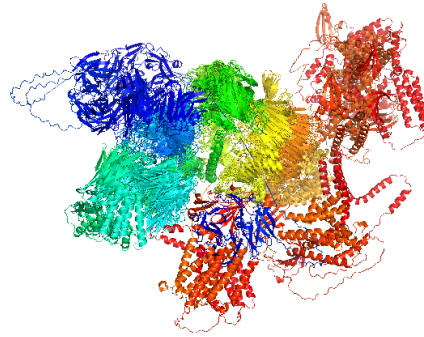
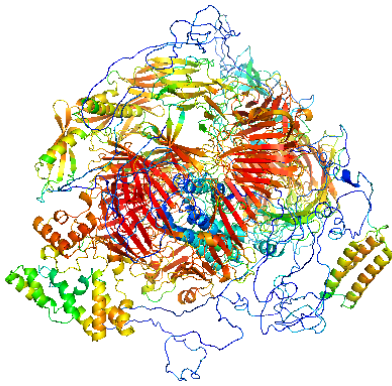
The AlphaFold-multimer model contained large disorder regions. These disappeared when modelled as two parts (6+3 chains).

T1154 (A1) Adding extra hits from metagenomes (right) makes the predictions more consistent. The model also contains fewer regions with low confidence (blue colours).



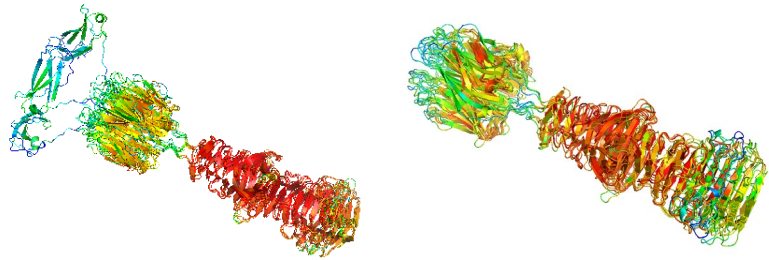
T1169 (A1)

Original AlphaFold model contained large low-confident regions. The protein was split into three parts (domains), which were modelled separately and manually docked with each other.



T1174 (A3)

Extra cycles made the model look better.

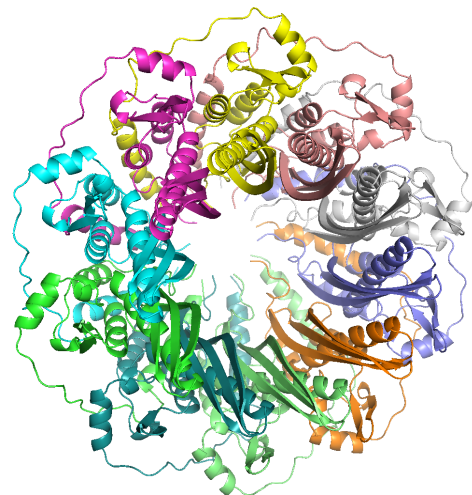
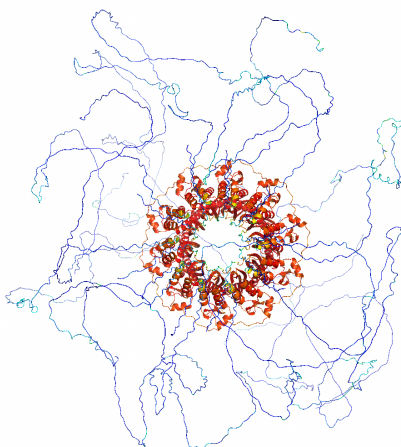


T1192 (A10)

AlphaFold multimer could only generate one model within the 72 hours server limit (on a node with 8 GPUs). We, therefore, modelled the other submissions, excluding the disordered regions.

**Avail
abilit
y**

Alph
aFold
is
freely
avail
able



from DeepMind.

MolPC is available freely at <https://gitlab.com/patrickbryant1/molpc>

References

1. Bryant, P. et al. Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. bioRxiv 2022.03.12.484089 (2022) doi:10.1101/2022.03.12.484089.
2. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11, 431 (2010).
3. Mitchell, A. L. et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. 48, D570–D578 (2020).

EMBER3D

EMBER3D: Fast protein structure prediction for protein mutation movies

Konstantin Weißenow, Michael Heinzinger & Burkhard Rost

Technical University Munich

k.weissenow@tum.de

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

High-quality state-of-the-art protein structure prediction systems such as Alphafold2¹ and RoseTTAFold² rely on evolutionary information captured by multiple-sequence alignments (MSAs), making structures of proteins with few evolutionary relatives tough to predict. Additionally, creating high-quality MSAs is not trivial: the parameters for the alignment process need to be chosen on an individual basis in order to add enough, yet diverse sequences. This is done with the goal of obtaining a rich set of sequences that model structural constraints, whilst avoiding the inclusion of sequences with diverging structure. While inference speeds of trained machine learning systems are fast, the database searches necessary to build MSAs significantly add to the runtime of state-of-the-art structure predictors.

Methods

We present EMBER3D (EMBedding-based inter-residue distance predictor), a novel end-to-end deep learning method used to predict 2D and 3D structure from sequence alone at high speeds.

EMBER3D consists of two major parts: first, we use our pre-trained protein language model ProtT5³ to generate rich, contextual representations of the amino-acid sequence of a query protein by extracting both the last-layer representation (1D) as well as the pairwise attention matrices (2D) of the underlying transformer architecture. Next, a structure module trained on a large set of experimentally determined structures jointly processes these representations to compute inter-residue distance probability distributions, anglegrams and 3D backbone coordinates (C, C-alpha, N and O).

The structure module is closely following the RoseTTAFold architecture with its 2-track and 3-track layout, however with some modifications: we significantly reduced the amount of both 2-track and 3-track blocks as well as the size of intermediate representations to optimize for computation speed. We further increased throughput by replacing the SE(3) transformer implementation of the original RoseTTAFold architecture with a functionally equivalent but much faster version from NVIDIA.

EMBER3D is primarily intended as a high-speed tool for the rendering of structure mutation movies and to assess the structural impact of mutations on large datasets, trading quality for speed. For the purpose of the CASP15 experiment, we therefore included an additional relaxation step with pyRosetta, using the predicted distograms and anglegrams as constraints for folding. We selected the best five models for each target using DeepAccNet⁴.

Availability

Preliminary code for EMBER3D is available at <https://github.com/kWeissenow/EMBER3D>. A preprint of the manuscript will be published soon.

1. Senior,A.W., Evans,R., Jumper,J. et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710.
2. Baek,M. et al. & Baker,D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network, *Science* Vol 373 Issue 6557.
3. Elnaggar,A., Heinzinger,M., Dallago,C. et al & Rost,B. (2021). ProtTrans IEEE TPAMI.
4. Hiranuma,N., Park,H., Baek,M. et al. (2021). Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat Commun* 12, 1340.

ESM-single-sequence

Atomic resolution protein structure prediction using a language model

Zeming Lin^{1,2,*}, Halil Akin^{1,*}, Roshan Rao^{1,*}, Brian Hie^{1,3,*}, Zhongkai Zhu¹, Wenting Lu¹, Allan dos Santos Costa⁴, Maryam Fazel-Zarandi¹, Tom Sercu¹, Salvatore Candido¹, Alexander Rives¹

1 - Meta AI, FAIR Team, 2 - New York University, 3 - Stanford University, 4 - Massachusetts Institute of Technology

* Equal Contribution

arives@fb.com

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

We study a development version of ESMFold in the CASP15 competition. Our objective is to investigate the ability of large scale language models to enable atomic-resolution structure prediction from a single protein sequence. ESMFold uses ESM-2, a new family of language models trained on diverse protein sequences across evolution, building on the findings of our previous generation ESM-1b to further scale up and produce a more performant language model. We train ESMFold to output a structure prediction in the form of atomic coordinates directly from the language model representation of a protein's sequence. This permits predictions to be made end-to-end from the sequence, an order of magnitude faster than current state-of-the-art approaches.

Methods

ESMFold uses a 3 billion parameter ESM-2 transformer language model¹, pretrained with the masked language modeling loss² on the UniRef50³ database. Language model weights are frozen and features are given to a structure prediction network which outputs three-dimensional coordinates. The structure prediction network is trained on a set of experimentally determined structures from PDB⁴ augmented with a set of computationally predicted structures generated with AlphaFold2⁵.

Our approach does not take either MSAs or templates as input. Nevertheless, we find that it produces accurate predictions of many protein structures. Replacing the MSA and templates with the language model enables predictions to be made at least an order of magnitude faster than current state of the art approaches.

At prediction time, we generate samples from the model by masking 1000 different subsets of amino acids from the sequence. The prediction with highest model confidence (pLDDT) is refined via Amber⁶ relaxation. For multimers, we set the residue index between chains to a random large gap, as well as inserting a phantom linker of 25 glycines between chains. For the latter half of the competition, we realized that masking but only accepting the

result if the predicted IDDT was greater than 0.1 compared to the original gave better overall performance, so we switched to this scheme. For longer proteins (>length 1600), to reduce computational cost we simply gave the single best prediction.

Availability

ESM-2 language models are available at github.com/facebookresearch/esm. The ESMFold structure prediction model will also be made available at the same location in the future.

1. Vaswani,A., Shazeer,N., Parmar,N., Uszkoreit,J., Jones,L., Gomez,A.N., Kaiser,L., and Polosukhin,I. (2017) Attention Is All You Need. In, *Advances in Neural Information Processing Systems.*, pp. 5998–6008.
2. Devlin,J., Chang,M.-W., Lee,K., and Toutanova,K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In, *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.
3. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B., Wu,C.H., and Consortium,U. (2014) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31, 926–932.
4. Burley,S.K., Berman,H.M., Bhikadiya,C., Bi,C., Chen,L., Di Costanzo,L., Christie,C., Dalenberg,K., Duarte,J.M., Dutta,S., Feng,Z., Ghosh,S., Goodsell,D.S., Green,R.K., Guranović,V., Guzenko,D., Hudson,B.P., Kalro,T., Liang,Y., Lowe,R., Namkoong,H., Peisach,E., Periskova,I., Prli,A., Randle,C., Rose,A., Rose,P., Sala,R., Sekharan,M., Shao,C., Tan,L., Tao,Y.-P., Valasatava,Y., Voigt,M., Westbrook,J., Woo,J., Yang,H., Young,J., Zhuravleva,M., and Zardecki,C. (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*, 47.
5. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Židek,A., Potapenko,A., Bridgland,A., Meyer,C., Kohl,S.A.A., Ballard,A.J., Cowie,A., Romera-Paredes,B., Nikolov,S., Jain,R., Adler,J., Back,T., Petersen,S., Reiman,D., Clancy,E., Zielinski,M., Steinegger,M., Pacholska,M., Berghammer,T., Bodenstein,S., Silver,D., Vinyals,O., Senior,A.W., Kavukcuoglu,K., Kohli,P., and Hassabis,D. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583–589.
6. Eastman,P., Swails,J., Chodera,J.D., McGibbon,R.T., Zhao,Y., Beauchamp,K.A., Wang,L.-P., Simmonett,A.C., Harrigan,M.P., Stern,C.D., Wiewiora,R.P., Brooks,B.R., and Pande,V.S. (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol*, 13, e1005659.

FALCON2

FALCON2: a web server for high-quality prediction of protein tertiary structures

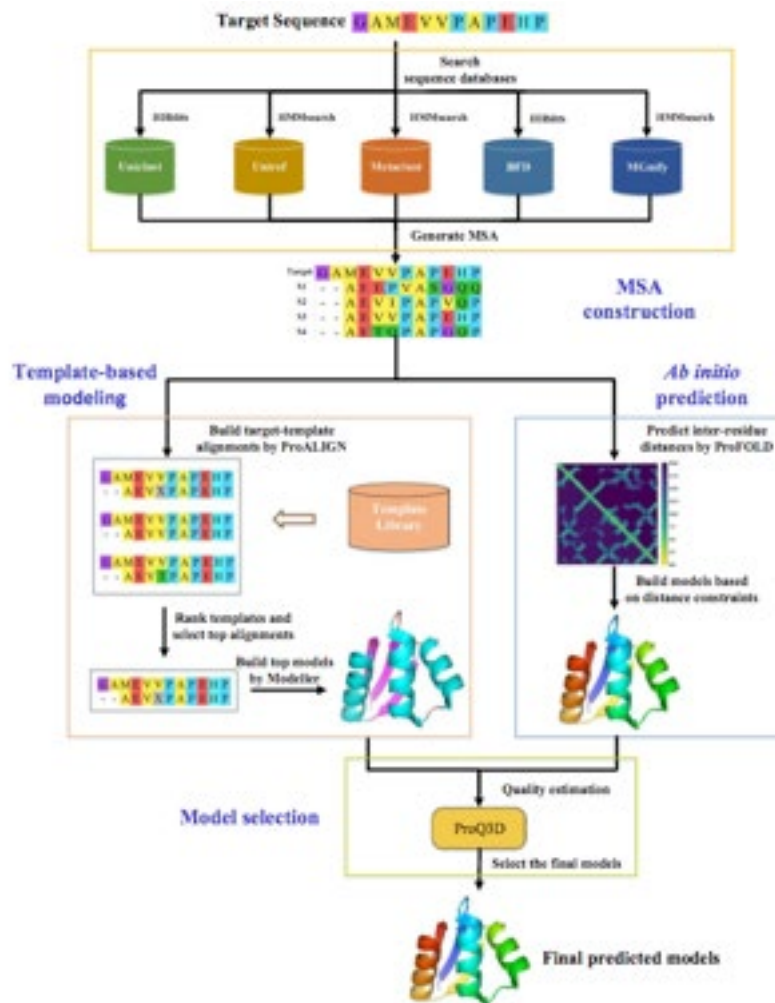
Fangxiong Xiao¹, Chungong Yu¹, HaiCang Zhang¹, and Dongbo Bu¹

1 - Institute of Computing Technology, Chinese Academy of Sciences

dbu@ict.ac.cn

Key: Auto:N; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N

ProALIGN¹ and ProFOLD² have achieved excellent performance in protein structure prediction. However, in-depth examination suggests that when high-quality templates are available, ProALIGN is superior to ProFOLD and in other cases, ProFOLD shows better performance. Therefore, we design a web server to take advantage of ProALIGN and ProFOLD.



Methods

For the candidate structures predicted by ProALIGN and ProFOLD, FALCON2 estimates structure quality by running ProQ3D. Briefly speaking, ProQ3D assesses the quality of a structure by considering a variety of features, including residue contacts, residue conservation, and the agreement with the predicted secondary structure and solvent accessibility area. By running ProQ3D on all candidate structures, FALCON2 obtains the predicted quality value IDDT and normalizes them into Z-score. FALCON2 finally selects the candidate structure with the highest IDDT as the final prediction result.

ProALIGN uses a Deep neural network to learn the patterns of context-specific alignment motifs. These patterns enable ProALIGN to model the dependence among residue pairs and thereafter accurately construct target-template alignments for structure building. Specifically, it consists of the following four main steps: (i) Feature calculation: The features to be used include sequence profile, secondary structure, solvent accessibility, and inter-residue distances. (ii) Alignment likelihood inference: The input features are fed into a pre-trained deep convolutional neural network, which predicts alignment likelihood for each residue pair (one query residue and one template residue). In our approach, alignment likelihood is represented as a matrix form. One entry in the matrix contains the match likelihood value of a residue pair. (iii) Alignment generation: Based on the alignment likelihood, we construct the optimal alignment with maximum likelihood. (iv) Model building: Build a 3D structure model by running MODELLER³ on the generated alignment.

ProFOLD employs an end-to-end framework called CopulaNet to estimate inter-residue distances directly from multiple sequence alignment (MSA) of the target protein. It consists of three key modules: MSA encoder, co-evolution aggregator, and distance estimator. MSA encoder embeds residue mutations using a 1D convolutional residual network. Co-evolution aggregator measures the co-mutations between two residues. Before presenting the design of co-evolution aggregator module. Distance estimator aims to estimate inter-residue distances according to the obtained residue co-evolution.

Results

We first evaluated the performance of FALCON2 over 104 CASP13 official-defined domain targets and 91 CASP14 official-defined domain targets, and the average TM-score is 0.755 and 0.712, respectively. Simultaneously, the FALCON2 server outperforms the two individual approaches.

Availability

Project name: FALCON2 server

Project home page: <http://falcon.ictbda.cn:89/serve/>

1. Ju, Fusong, et al. "CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction." *Nature communications* 12.1 (2021): 1-9.
2. Kong, Lupeng, et al. "ProALIGN: Directly learning alignments for protein structure prediction via exploiting context-specific alignment motifs." *Journal of Computational Biology* 29.2 (2022): 92-105.
3. Eswar, N., Webb, B., Marti-Renom, M.A., et al. 2006. Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics*. 15, 5—6.

Lim Heo and Michael Feig

1 - Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA

mfeiglab@gmail.com

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

Kinases catalyze phosphorylation reactions of various substrates. The phosphorylation state of substrates controls their biological activities such as regulation of metabolic pathways and cell signaling. Thus, the kinase family is one of the most important protein families and draws attention as therapeutic targets among the human proteome. Kinases can adopt various conformations especially at their ATP/substrate binding pocket and activation loop depending on their activation or substrate binding states. More than 8,000 human kinase structures have been experimentally determined. Even though structures for around half of the human kinases were captured by those experimental structures, they are still far from covering diverse conformations of the kinases. Alternatively, computational protein structure prediction methods can predict protein models at near experimental accuracy thanks to recent advances in machine learning-based methods. However, these methods have strong conformational state biases in modeling proteins that can have multiple conformational states. For example, AlphaFold¹ is likely to predict GPCRs in the inactive state² and kinases in the active state³, at which experimentally determined structures are dominant. Recently, we devised a multi-state modeling protocol for GPCRs using AlphaFold with activation state-annotated template databases². The protocol utilized experimental structure templates in either active or inactive states to predict models in the corresponding states. In addition, to make prediction effectively guided by templates in either state, sequences in the input MSA were masked at sequence positions at which templates were aligned. Using this protocol, it was possible to predict models for both active and inactive states of the GPCR at atomic-level accuracy.

Methods and Results

During CASP15, we extended the multi-state modeling protocol towards kinases to model targets T1195-97. To construct conformational state-annotated template databases for kinases, state annotations for experimentally determined kinase structures were taken from KinCoRe³. The annotation is based on the spatial orientation of the DFGmotif (DFG_{in}, inter, and out) and the dihedral angles of the Phe in the motif, and there are 12 conformational states in total. For each target, we attempted to model in every state using the multi-state modeling protocol. According to our benchmark test, different from multi-state modeling of GPCRs, multi-state modeling of kinases often resulted in a state different from that of used templates. For example, modeling in the DFG_{in} state was always successful. On the other hand, modeling in the DFG_{out} state succeeded only for 44% of tested kinases and resulted in DFG_{in} and DFG_{inter} states for

44% and 12% of the targets because AlphaFold has a strong bias for the DFGin state (*c.f.* models from the original AlphaFold adopted DFGin, inter, and out states for 87%, 6%, and 7% of the same set of kinases, respectively). Although our multi-state modeling protocol for kinases cannot model in every state, it is still beneficial as it can expand the conformational state coverage beyond what the original AlphaFold can do, while all the conformational states may not be accessible for every kinase.

For the model selection, according to the benchmark test, we found that a model for a state from our multi-state modeling protocol was more accurate than a model from the original AlphaFold if (1) the sequence identity of templates for the state was high (> 40%) and (2) the predicted model was in the same state of the used templates. Consequently, we selected five models for submission using aforementioned criteria, and selected models were ranked according to their pLDDT scores.

Availability

The multi-state modeling protocol for kinases and GPCRs is available at https://colab.research.google.com/github/huhlim/alphafold-multistate/blob/main/AlphaFold_multistate.ipynb.

1. Jumper, J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589.
2. Heo, L. & Feig, M. (2022) Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins*, doi:10.1002/prot.26382.
3. Modi, V. & Dunbrack, R. L. (2022) Kincore: a web resource for structural classification of protein kinases and their inhibitors. *Nucleic Acids Res.* **50**, D654-D664.

Protein assembly modeling by pyDock: integration of *ab initio* docking and energy-based scoring of AlphaFold interfaces

L.A. Rodríguez-Lumbreras^{1,2} and J. Fernández-Recio^{1,2}

1 - Instituto de Ciencias de la Vid y del Vino (ICVV), CSIC/UR/Gobierno de La Rioja, Spain, 2 - Barcelona Supercomputing Center (BSC), Spain

juan.fernandezrecio@icvv.es

Key: Auto:Y; CASP_serv:Y; Templ:Y; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:Y

We explored here new strategies for modeling of protein assemblies by integrating deep learning approaches like AlphaFold with the docking and energy-based scoring function of pyDock¹, which previously showed successful results on template-based and *ab initio* docking models². For that, we participated in the CASP15 Assembly category, as part of the 5th common CASP-CAPRI Assembly Prediction challenge (CAPRI Round 54), consisting in 39 targets: nine homo-dimers (A2), 13 hetero-dimers (A1B1 or E1I1), five homo-trimers (A3), three hetero-trimers (A2B1), and nine higher-degree homo- and heter-oligomers (ranging from a hetero-pentamer to a homo-16mer). As human predictors, we participated in all of the proposed targets except in H1137/T204. As scorers, we participated in all 38 proposed targets (target H1106/T191 was not included in the scoring experiment).

For each assembly, the coordinates of the individual subunits were taken from the AlphaFold2 models available at CASP site by Elofsson group (except for some subunits in targets H1135/T203 and H1151/T210, which had available structure). The AF2 models were further processed to keep only reliable residues, based on pLDDT values, initially using a cutoff value of pLDDT > 60, and then pLDDT >70 (but for some cases we tolerated smaller pLDDT values in order not to remove >30% of the protein sequence, to be able to pass the CAPRI server verification).

As predictors, we applied our pyDock¹ docking pipeline to the individual subunits, in order to build the binary interactions in each assembly. For that, we used FTDock (electrostatics on; 0.7 Å grid resolution) to generate 10,000 rigid-body docking poses. In homo-oligomers, we assume symmetry and removed docking poses not satisfying the expected symmetry (e.g. cyclic C₂ symmetry for homo-dimers; C₃ for homo-trimers) within a given tolerance³. In target T1132/T201, we selected docking models that fitted into a trimer of dimers symmetry, as expected from an available template (PDB code 2GFF). In target H1135/T203, hetero-tetrameric interfaces (A3B1) were built based on an available template (PDB code 6WME), and these models were used as input for *ab initio* docking to build the final assembly as a trimer of hetero-tetramers (A9B3).

All docking models were scored and ranked by pyDock energy-based function. After removing redundant models (within 4 Å ligand RMSD), we selected automatically the 10 best-ranked models (per rules of CAPRI) from pyDock scoring as our submission set. Interestingly, we checked that in around half of the cases, this initial submission set contained docking poses that were similar (within 10 Å ligand RMSD) to the AF-Multimer assembly models available at CASP site by Elofsson group. In these cases, this consistency between the energy-based pyDock and the AF-Multimer predictions reassured our confidence in the submission set based only on pyDock scoring. In the rest of the cases, to build a more reliable submission set, we combined (in alternative order) the top 5 docking models from pyDock scoring and the 5 docking models that were most similar to the AF-Multimer predictions, independently on their energy-based scoring. Before submission, all models were minimized with AMBER to remove clashes and improve geometries.

In the CAPRI scorers experiment, we first removed models with more than 25 clashes (i.e., intermolecular pairs of atoms closer than 3 Å) per interface. Then, we applied pyDock scoring and used the same criteria to rank the docking models as in predictors (i.e. filter by symmetry, check available templates, compare to AF-Multimer, and minimize clashes).

Availability

The pyDock 3.0 program is available for academic use as a GNU/Linux binary and as a web server (<https://life.bsc.es/pid/pydock/>).

1. Cheng, T.M.-K., Blundell, T.L. & Fernandez-Recio, J. (2007) pyDock: electrostatics and desolvation for effective rigid-body protein-protein docking. *Proteins*. 68, 503-515.
2. Lensink, M.F., Brysbaert, G., Nadzirin, N., et al. (2019). Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins*. 87, 1200-1221.
3. Rosell, M., Rodríguez-Lumbreras, L.A., Romero-Durana, M., et al. (2020) Integrative modeling of protein-protein interactions with pyDock for the new docking challenges. *Proteins*. 88, 999-1008.

Protein Structure Prediction based on Adaptive Quality Assessment from Multiple Sequence Alignment

Andrés Lince¹, Mahdi Rahbar², Gabriel Schwarz³, Luotong Kang¹, Dong Si¹, Renzhi Cao⁴,
Jie Hou^{2*}

1 - Division of Computing and Software Systems, University of Washington Bothell, Bothell, Washington, USA; 2 - Department of Computer Science, Saint Louis University, Saint Louis, Missouri, USA; 3 - Highland High School, Highland, IL, 4 - Department of Computer Science, Pacific Lutheran University, Tacoma, WA 98447

*Corresponding author: jhou4@slu.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

Protein structure prediction has made a breakthrough in recent years under the community-wide efforts in biological research and the advancement of artificial intelligence techniques¹. The ever-growing number of large-scale metagenomic sequences significantly boosts the precision of computational algorithms in predicting the tertiary & quaternary protein structures from the sequences, specifically advancing the capabilities of deep learning (DL) models to extract co-evolutionary structural information from the sequence families^{2,3}. The multiple sequence alignments (MSA) from homologous sequences have become the essential knowledge base from most recent state-of-the-art DL approaches to guide the direct folding of protein sequences. Substantial studies have shown that the quality of 3D protein structure modeling, including multi-metric proteins, largely relies on the resolution of co-evolutionary signals in the sequence alignments, such as the number of effective homologs and diversity of organism species in MSA. In this work, we proposed an adaptive scheme to select the optimal sequence alignments for generating higher-quality 3D structure predictions given the input protein sequences. The quality of all the predicted structures were examined using our latest deep learning-based protein quality assessment algorithms using graph neural networks to determine the top ranked 3D models for the submission to CASP15.

Methods

The proposed 3D protein structure prediction pipeline starts with the generation of MSAs by searching the query protein sequence against the representative biological sequence databases, including Uniref90, Uniprot, MGnify, BFD, uniclust30, PDB70 as used in work ². The algorithm will collect the diverse set of MSAs by executing several well-known sequence alignment techniques (i.e., HHblits, HHsearch, Jackhammer) with different parameters, including the e-value thresholds (i.e., 1e-4, 1e-3, 1e-2, 1e-1, 10), sequence overages (i.e., 30%, 40%, 50%, 70%) and iteration runs (i.e., 1, 2, 3). The MSAs from another alignment tool, DeepMSA⁴, will also be included in the pool of diverse alignments. The template database (PDB70) is used to determine

the existence of domain boundaries in multi-domain proteins. If the multi-metric protein is identified, the MSAs for individual domains will be generated and concatenated as full-length MSAs to be included in the alignment pool. For single-chain tertiary structure prediction, each of the MSAs in the alignment pool is fed into the structure module in Alphafold framework² to generate five 3D structure predictions with Amber constrained relaxation, leading to at least ~70 predicted structures on average from the available MSAs. The predicted structures are finally re-ranked according to the averaged pLDDT scores from our in-house graph-neural network based quality assessment approach⁵ and the local predicted scores by Alphafold. The top 5 ranked models are then submitted as predictions. For quaternary structure prediction, the structures of the protein complex were derived from Alphafold-Multimer⁶.

Acknowledgements

We are grateful to Dr. Cao's lab at Pacific Lutheran University and the DAIS lab at the University of Washington Bothell for additional computing resources and group efforts in helping the structure prediction of large-size proteins in CASP15. Andrés Lince is also sponsored by SENACYT under the scholarship program named "Becas de Pregrado de Excelencia 2019". The work was also partially supported by the PRF Award at Saint Louis University to JH in 2021.

1. Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN. High-accuracy protein structure prediction in CASP14. *Proteins*. 2021 Dec;89(12):1687–99.
2. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021 Aug;596(7873):583–9.
3. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. 2021 Aug 20;373(6557):871–6.
4. Zhang C, Zheng W, Mortuza SM, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics*. 2020 Apr 1;36(7):2105–12.
5. Rahbar M, Chauhan RK, Shah PN, Cao R, Si D, Hou J. Deep graph learning to estimate protein model quality using structural constraints from multiple sequence alignments. *In Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics 2022 Aug 7 (pp. 1-10)*.
6. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun*. 2022 Mar 10;13(1):1265.

AnglesRefine: refinement of 3D protein structures using Transformer based on torsion angles

Junyong Zhu¹, Jie Hou², Dong Si³, Renzhi Cao⁴, Sheng Wang¹ and Lei Zhang¹

1 - Department of Computer Science, AnHui University, Hefei, China, 2 - Department of Computer Science, Saint Louis University, Saint Louis, Missouri, USA, 3 - Division of Computing and Software Systems, University of Washington Bothell, Bothell, Washington, USA, 4 - Department of Computer Science, Pacific Lutheran University, Tacoma, WA 98447

e21201084@stu.ahu.edu.cn

Key: *Auto:Y; CASP_serv:Y; Templ:Y; MSA:Y; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:N*

The protein structure refinement method aims to improve the accuracy of the predicted protein models, especially the local structure at the residue level. Most of the existing refinement methods are based on physics, while molecular simulation methods are computationally expensive and time-consuming. In this work, we utilize deep learning techniques to extract the structural constraints from residues in protein structure to guide protein structure refinement. Different from the existing methods, we proposed a new method AnglesRefine for structure refinement. AnglesRefine is based on the secondary structure of protein and uses a transformer¹ model to refine the various angles of protein structure (psi, phi, omega, CA_C_N_angle, C_N_CA_angle, N_CA_C_angle), and finally generates a higher quality protein model from the refined angles.

Methods

First of all, we use PSIPRED² tool to predict the secondary structure of the protein model from the protein sequence as the target secondary structure, and we compare it with the secondary structure extracted from the predicted protein model, and the unmatched fragments or the incorrect local structures will be refined. (In this CASP, we used the model predicted by AlphaFold³ as the initial protein model, and applied our new method to refine it for the protein structure prediction); Then, we extract the angles of those incorrect local structures, and use six transformer models to optimize 6 angles (psi, phi, omega, CA_C_N_angle, C_N_CA_angle, N_CA_C_angle) of each incorrect local structure to obtain the optimized angles; Finally, our in-house tool is used to convert the optimized angles into refined local structures, and these refined local structures are used to replace the original local structures to obtain the final refined protein structure.

Results

AnglesRefine was trained and tested on casp11-14 dataset, the average accuracy of α -Helix structure's transformer models trained by our method is about 0.7, and the results showed that our method effectively trained angles' transformer models of α -Helix. In addition, the models can transform local structures (β -sheet or random coil, but expected to be α -Helix) into α -Helix

with 100% accuracy in our test dataset. In addition, our method can refine a protein model in about 30 seconds, while other physics-based refinements take several minutes.

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
2. Buchan, D. W., & Jones, D. T. (2019). The PSIPRED protein analysis workbench: 20 years on. *Nucleic acids research*, 47(W1), W402-W407.
3. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589

ZoomScore: residue-level protein complex assessment with machine learning on sequential and 3D structural features

Sheng Wang¹, Jie Hou², Dong Si³, Renzhi Cao⁴, Junyong Zhu¹, Lei Zhang¹

1 -Department of Computer Science, AnHui University, Hefei, China, 2 -Department of Computer Science, Saint Louis University, Saint Louis, Missouri, USA, 3 - Division of Computing and Software Systems University of Washington Bothell, Bothell, Washington, USA; 4 - Department of Computer Science, Pacific Lutheran University, Tacoma, WA 98447

e21201006@stu.ahu.edu.cn

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

The Estimation of Model Accuracy plays an important role in protein structure prediction, especially the local quality of residues that contact with other residues from the interacting proteins. However, traditional QA methods (like DeepQA¹, GraphQA² etc.) usually predict a global quality score of the predicted single protein structure, but cannot be used to evaluate the interfaces between multimeric complexes and subunits. In our previous work, we developed a local model quality assessment method for evaluating single protein models, ZoomQA³, which can assess the accuracy of a tertiary protein structure/complex prediction at the residue level. It has shown the potential to identify problematic regions of the SARS-CoV-2 protein complex. Based on that, we proposed the new model quality assessment method ZoomScore to evaluate the quality of interface residues in protein complexes based on sequence and 3D structural and chemical features and benchmark it on CASP15.

Methods

ZoomScore uses a novel representation of amino acids in the protein structure and addresses the residue level protein quality assessment problem with the help of machine learning techniques.

ZoomScore utilizes 10 properties regarding the chemical and physical properties of the target amino acid and its environment. We will use the term ‘fragment’ to describe a region of a protein that can be generated by including all amino acids within a radius of consideration r (Å) of a target amino acid where r represents a distance measured in angstroms. Two datasets were generated: one where r was set to a minimum of 5 angstroms and a maximum of 25 angstroms, and another where r was set to a minimum of 5 angstroms and a maximum of 55 angstroms. For each dataset generated, the step for considering a new ‘fragment’ was 1 angstrom. According to our analysis, the average proportion of protein residues included at the radius of 25 angstroms was 0.5265 and the average proportion of protein residues included at the radius of 55 angstroms was 0.9393.

The new method ZoomScore fused the following features: (1) the average amino acid density of a fragment; (2). the average hydrophobicity of the fragment. (3). the average monoisotopic mass of the fragment. (4). the average solvent accessibility of the fragment. (5). the

isoelectric point of the fragment. (6). the stability score of the target amino acid's torsion angles. (7-10). We also included the properties of the center amino acid, into the feature set that comprises the amino acids' monoisotopic mass, hydrophobicity, solvent accessibility, isoelectric point, and torsion angles (two values). We include the amino acid letter code and the secondary structure extracted from the protein structure as one-hot encoded vectors, and all features are normalized between 0 and 1.

Interface residues between chains are defined as the amino acids having contacts with at least one residue from different chains with a distance of CB-CB atoms $\leq 8 \text{ \AA}$ (CA atom in the case of Glycine). In the training stage, ZoomScore identified all interface residues and generated a total of 2397 features for each amino acid. We selected the top 100 performing features based on their Pearson correlation with the known quality scores. The final model was trained on 60 000 vectors of the top 100 features for a set of samples of data generated from a maximum radius of consideration of 55 angstroms. The Support Vector Machine is trained as the final model with the RBF kernel, a C value of 1.0, an epsilon value of 0.1, and a gamma value of 1.0. In the prediction stage, we collect features for all interface residues given the input complex structures and generate the predicted quality scores for all residues, including the interface residues between chains.

1. Cao, R., Bhattacharya, D., Hou, J., & Cheng, J. (2016). DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC bioinformatics*, 17(1), 1-9.
2. Baldassarre, F., Menéndez Hurtado, D., Elofsson, A., & Azizpour, H. (2021). GraphQA: protein model quality assessment using graph convolutional networks. *Bioinformatics*, 37(3), 360-366.

Assembly prediction in CASP15 with ab initio docking and template-based modeling

Gwang So¹, Gyong-Jin Pang¹, Chung-Hyok Kang¹, Ryong-Chol Kim¹, Chol Ryom¹, Chang-Hyok Ri¹ and Gwang-Hyok Kim¹

¹- Department of Bioinformatics, Branch of Biotechnology, State Academy of Sciences, Daedonggang-District, Pyongyang, DPR Korea

sg1992@star-co.net.kp

In CASP15, we used ab initio docking(zdock³), template-based modeling (MODELLER⁵), symmetry constraints.

Methods

For each assembly, the models of the individual subunits were taken from server models of the AlphaFold and Baker servers.

First, we searched template of target sequence against the PDB database² by using PSI-blast¹. If the target-like template is identified in PDB, we simulated the quaternary structure of the target using the multi-chain method of MODELLER. Here we use homo-types.

And for targets with a small number of subunits (2 ~ 3), initial models were obtained using zdock³, and excellent models were selected from them.

In addition, for targets in which subunits are in symmetrical conformations, symmdock⁴ was used to assemble the structures of subunits presented in the target, and excellent models were selected from the results.

In the final result model selection stage, the three-dimensional structure of the templates from the PSI-blast search results for the PDB database² was analyzed and final models were selected from the various docking results.

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.
2. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res* 28, 235-242 (2000).
3. Chen,R., Li,L., Weng,Z. (2003) ZDOCK: An initial-stage protein-docking algorithm. *Proteins.* 52, 80–87.
4. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. PatchDock and SymmDock: servers for rigid and symmetric docking. *NAR*, 33: W363-W367, 2005.
5. Webb,B. and Sali,A. (2016) Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinforma.*, 2016, 5.6.1-5.6.37.

Protein-ligand complexes docking by using AutoDock Vina and template informations

Gwang So¹, Gyong-Jin Pang¹, Chung-Hyok Kang¹, Ryong-Chol Kim¹, Chol Ryom¹, Chang-Hyok Ri¹ and Gwang-Hyok Kim¹

¹- *Department of Bioinformatics, Branch of Biotechnology, State Academy of Sciences, Daedonggang-District, Pyongyang, DPR Korea*

sg1992@star-co.net.kp

In the protein-ligand complex modeling class of CASP15, we generated a protein-ligand docking model using AutoDock Vina², a small-molecule docking software. The binding position of the ligand and the final model selection were carried out through repeated calculation experiments by searching for template data in the PDB database¹ and using the information.

Methods

For the initial protein conformation to obtain the protein-ligand complex model, the AlphaFold and Baker sever models or the best model among our predictive models was used.

If a template structure similar to the target is identified from the PDB database, the type of ligand and binding position information are obtained from it.

Then, using AutoDock Vina², the binding sites of the ligand are found by the fixed docking method and the flexible docking method.

Finally, the final result model is selected using the protein-ligand binding energy and template information.

1. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res* 28, 235-242 (2000).
2. Goodsell, D. S., & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins Structure Function and Bioinformatics*, 8(3), 195-202.

Template-Based Structure Prediction by using dPPAS alignment method

Gwang So¹, Gyong-Jin Pang¹, Chung-Hyok Kang¹, Ryong-Chol Kim¹, Chol Ryom¹, Chang-Hyok Ri¹ and Gwang-Hyok Kim¹

¹- *Department of Bioinformatics, Branch of Biotechnology, State Academy of Sciences, Daedonggang-District, Pyongyang, DPR Korea*

sg1992@star-co.net.kp

In CASP15, we used the dPPAS alignment method based on PSI-blast¹ profile information and some structural information. And the database that consist of log-odds profile, relative solvent accessibility, and three-state secondary structure information of templates were constructed, and alignment between the target and the template was performed based on the dPPAS alignment method. Then, the quaternary structure of the target was modeled by MODELLER².

Methods

To obtain the log-odds profile of a target sequence, we used PSI-blast against the latest Uniref50 database.

To predict the three-state secondary structure information and relative solvent accessibility value of a target sequence, we used FTBiot method that is based a deep learning model.

Then, we used the profile, the three-state secondary structure information, relative solvent accessibility value of target and template sequence with the dPPAS alignment method, to obtain pair alignment between a target and template sequence. The weight coefficients of this dPPAS alignment method were optimized through RNN models.

In our 3D-model construction step, we used MODELLER² with the result of the pair alignment between a target and template sequence and PDB entry file of templates.

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.
2. Webb,B. and Sali,A. (2016) Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinforma.*, 2016, 5.6.1-5.6.37.

Predicting conformational diversity in AlphaFold with masked sequence alignments

Liwei Chang^{1,2}, Alberto Perez^{1,2,*}

1 Department of Chemistry, University of Florida, Gainesville, Florida, United States, 2 Quantum Theory Project, University of Florida, Gainesville, Florida, United States;

* perez@chem.ufl.edu

Key: Auto:N; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N

We participated in the Protein conformational ensembles category of CASP15 experiment as group “GatorsML”. The structure prediction was done by using AlphaFold (AF)¹ with masked co-evolution information in the multiple sequence alignments (MSA).

Methods

The alternative conformation predictions were performed as the following steps:

Step 1: initial structure prediction from AlphaFold: we first run the structure prediction using trained AF parameters with the local install ColabFold interface, where the MSA is generated by MMseqs2 searching UniRef100 database². No template information is used in the prediction.

Step 2: alternative conformation prediction from AlphaFold with masked MSA: previous studies have shown that reducing the depth of input sequence alignments is able to generate alternative conformations of transporters and receptors³. Given our assumption that the regions with large structure fluctuation correspond to residues with low confidence prediction scores (under the condition that MSA information at this region is not too sparse, e.g. >100 alignment hits), we first remove the sequence alignments for those residues and then use subsampled MSA sets as input for AF structure prediction. To select final models from predicted structures, we carried out principal component analysis (PCA), which helps view the structure sets in the dimensions of large conformation variance. The top-ranked structure without masked sequence alignments was submitted as model_1 and the other four were selected based on (1) the population of similar structures in the prediction sets under the top two principal components and (2) their RMSD differences against model_1.

1. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. Nature 1–11 (2021).
2. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. Nat Methods 19, 679–682 (2022).
3. Alamo, D. del, Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. Elife 11, (2022).

Computational modeling of RNA 3D structures and interactions

S.Mukherjee, S.N.Moafinejad, F.Jaryani, M.A.Farsani, N.G.Badepally, E.Baulin, T.Wirecki,
F.Stefaniak, and J.M.Bujnicki*

International Institute of Molecular and Cell Biology in Warsaw

*janusz@iimcb.gov.pl

Ribonucleic acid (RNA) molecules are the master regulators of cells. They are involved in many molecular processes: They can transmit genetic information, sense cellular signals, relay responses, and catalyze chemical reactions. The function of RNA, particularly its ability to interact with other molecules, is encoded in the sequence. Understanding how these molecules perform their biological tasks requires detailed knowledge of RNA structure and dynamics, which determine how RNA folds and interacts in the cellular environment.

Methods

Our workflow for computational modeling of RNA 3D structures and their interactions with other molecules uses a suite of methods developed in our laboratory, including PARNASSUS for the remote homology detection, MeSSPredRNA for the prediction of canonical and non-canonical base-pairs, and the SimRNA-family of programs for the modeling of RNA 3D structure and its complexes with other molecules.

Results

We applied our methods to predict RNA 3D structures in the CASP and RNA-Puzzles experiments.

Availability

SimRNA was previously published and is available as a standalone tool at <http://genesilico.pl/software/stand-alone/simrna>

Other elements of our computational workflow are experimental and are not yet available.

GinobiFold: MSA-free Superposition Model

Anqi Pang, Kexin Zhang, Zhigang Sun, Jiale Sun and Jingyi Yu

ShanghaiTech University

pangaq@shanghaitech.edu.cn

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

GinobiFold is a hybrid multi-model deep learning network. The whole pipeline includes Feature Embedding, Evoformer, Structure Module and Structure Refinement. Before the Ginobifold pipeline, we need to train a multi-sequence alignments (MSA) generative model and pLM model from metagenomic datasets to generate MSA features. According to different training configurations, a total of 10 sets of parameters were obtained. Finally, we leverage predicted local distance distribution test(pLDDT) value to generate 5 new superposition models.

Methods

The whole architecture of Ginobifold includes Feature Embedding, Evoformer, Structure Module and Structure Refinement.

Before the Ginobifold pipeline, we need to build metagenomic database that is used to train a MSA generative model and pLM model. we downloaded hundreds of TB sequence data from NCBI and MGnify and used plass to assemble them. Then, we built our customized datasets by a similar pipeline of BFD. Because of the limitation of memory and speed, we split assembled proteins into many small parts and used GPU-accelerated techniques to build large MSA databases that can be searched by hhblits. These databases are about 20x larger than BFD database so that we can get enough homologs for each query sequence.

In Feature Embedding part, we construct representations from Residue Embedder and MSA Embedder. For MSA Embedder, we combined two different Embedders to get the final MSA features including Pseudo MSA Embedder and protein Language Model (pLM) Embedder. In Pseudo MSA Embedder, we first generate pseudo MSA from pre-trained MSA generative model. In pLM Embedder, we output MSA features directly. By different combinations of the above two MSA Embedders, we got 10 kinds of MSA features. Finally, we combine Residue Embedder features and MSA features to build MSA representations and pair representations.

In Evoformer and Structure Module part, we used the same architecture as alphafold2¹. For all self-attention layers, we used dynamic axial parallelism technique to save GPU memory and accelerate forwarding and back-propagation speed.

In structure refinement, we used *OPENMM*² with CUDA platform to firstly optimize structures. We got several hundred predictions for each target with different generated pseudo MSA and different models. Based on the assumption that, the residue with higher pLDDT value among all predictions will have more reliable local regions, we used 5 different strategy to optimize structures that start from the top 1 structure (ranked by mean pLDDT). Thus, we can get 5 so called superposition models. The first strategy is that residues with higher pLDDT value have better global coordinates. The second is that better residues have better local coordinates. The third is that better residues have better distance vector of CA (only center CA atom)-AA pair within certain angstrom shell. The fourth is that the better residues have better distance matrix for all CA-CA atom pairs within certain angstroms. The strategy of superposition5 is that better residues will have better distance matrix of CA-AA atom pairs within certain angstroms.

1. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
2. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang LP, Simmonett AC, Harrigan MP, Stern CD, Wiewiora RP, Brooks BR, Pande VS. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol*. 2017 Jul 26;13(7):e1005659.

GinobiFold-SER: Deep Learning based protein structure prediction model

Zhigang Sun, Kexin Zhang, Fenglei Li, Anqi Pang and Jingyi Yu

ShanghaiTech University

pangaq@shanghaitech.edu.cn

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

GinobiFold-SER is a hybrid multi-model deep learning network. The whole architecture uses metagenomic data to obtain multi-sequence alignments (MSA) with different searching methods. According to different training configurations, a total of 15 sets of parameters were obtained. Finally, we leverage averaged predicted local distance distribution test (pLDDT) value to rank multiple predictions.

Methods

The whole model architecture includes data processing, feature embedder, evoformer, structure module and structure refinement. In data processing part, we used GPU accelerated techniques to build MSA database for NCBI and Mgnify. In feature Embedder part, we search MSA from above metagenomic data and build different features with different methods. In evoformer and structure module part, we used alphafold² architecture. For all self-attention layers, we used dynamic axial parallelism technique to save GPU memory and accelerate forwarding and backpropagation speed. In structure refinement, we used OPENMM² with CUDA platform. We got several hundred models for each target with different MSA features and models and rank them by averaged pLDDT, then we select top 5.

Results

For most of CASP targets, we can get reliable predictions based on pLDDT values.

1. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
2. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang LP, Simmonett AC, Harrigan MP, Stern CD, Wiewiora RP, Brooks BR, Pande VS. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol.* 2017 Jul 26;13(7):e1005659.

Cross-geometric vector perceptron networks for denoised reconstruction of protein-ligand conjugated conformation

Xiang-Yu Lin, Yi-Ting Chen, Sing-Han Huang, Cheng-Tang Wang, Ching-Yung Lin

Graphen Inc., New York, NY 10110, USA

molalin@graphen.ai, timchen@graphen.ai, johnhuang@graphen.ai, tomwang@graphen.ai, cylin@graphen.ai

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:Y.*

Specificity and efficacy are the keys to develop a practical therapeutic-available drug against aimed-diseases. To establish such therapeutic capability of certain molecules, protein-ligand conjugated structures are crucial references to fully understand the mechanism of how a molecule conducts its target inhibition ability. We present a novel protein-ligand orientation denoised model, constructed with evolutionary cross-geometric vector perceptron (cGCP) modules, to fully leverage the space orientation information carried by large (protein) and small (ligand) molecules. Benefited from the convenience of a topological presentation, we decompose a protein description traditionally in the amino acid level further into the atomic level, detailing the interaction correlation lied between protein and ligand. Considering the spectral orientations of separated molecules are the noisy results given a protein sequence and a small molecule, our model reconstructs the binding structure as the de-noised orientation of the conformation by minimizing the overall Gibbs energy.

Methods

Our mode operates on diverse scales of protein graphs along with the atomic-scale graph of ligand. To represent a protein, three scales to describe nodes are considered: (1) protein backbone in the atomic scale; (2) whole protein in the amino acid scale; and (3) whole protein in the atomic scale. They are dedicated to determining the global protein structure, orienting the substructure of each motif, and optimizing the conformation of the protein-ligand conjugating complex, respectively.

As the input features of our model, the nodes in protein and ligand graphs are continuous-positioned embedded, and the edge information is composed with atomic bonds including the inner-protein, inner-ligand, and cross-protein-ligand distance features. Embedded features of protein and ligand are used as the inputs of our cross geometric vector perceptron network to generate the conjugated structure output. The model generates three predictions: (1) distance matrix in the atomic resolution of within and cross protein and ligand, (2) torsions, angles, and coordinates of the atoms, and (3) Gibbs energy of the current protein-ligand conformation.

For each predicted structure, energy relaxing is conducted using restrained gradient descent on CHARMM¹ force field and SIMTK-OPENMM² package to further retrieve the refined protein-ligand complex structure.

Availability

The source codes and models are not publicly available at the moment.

1. MacKerell AD Jr, et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B*. 102:3586–3616.
2. Eastman, P. et al. (2017). OpenMM 7: Rapid development of high-performance algorithms for molecular dynamics. *PLoS Computational Biology*, 13(7), e1005659.12.
3. Altschul, S. F. et al. (1997), Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389-3402.

Deep reinforcement learning for de novo RNA structure prediction

Sing-Han Huang, Yi-Ting Chen, Xiang-Yu Lin, Cheng-Tang Wang, Ching-Yung Lin

Graphen Inc., New York, NY 10110, USA

johnhuang@graphen.ai, timchen@graphen.ai, molalin@graphen.ai, tomwang@graphen.ai, cylin@graphen.ai

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:Y.*

RNAs play many important roles for cellular processes. They need to fold into secondary and tertiary structures to perform their functions and interact with other biomolecules. Experimental determination of RNA three-dimensional (3D) structure is time consuming and challenging, so that the structures of the majority of known RNAs are still not well understood. Here, we present a Deep Reinforcement Learning algorithm to predict the RNA 3D structure. We use the Reinforcement Learning combined with Deep Neural Network to predict the RNA 3D structure of a given RNA sequence.

Methods

The process starts with the input sequence of the target RNA, which is initially an open strand without any base pairing. Then, according to the policy function learned by the deep neural network, the most likely base-pairing is selected from the action space. This will be iterated until folding to the final state; here, we use the value function to score the current state to determine whether to stop. The error between the tertiary structure is returned as a reward by comparing the final folding state with the native state. After training the first round, the process performs the next round of the folding simulation on the new model, and repeatedly loops the above process until 1000 episodes. Different from traditional deep learning models that just need training data to train the model, the training of reinforcement learning generally requires the agent-environment interactions. Here, we used RNAWorld3D from OpenAI Gym¹ to set up the reinforcement learning simulation environment for RNA 3D structure prediction. The predicted RNA 3D structure from reinforcement learning was further refined by the energy function of openMM².

1. B. Greg, C. Vicki, P. Ludwig, S. Jonas, S. John, T. Jie, and Z. Wojciech. "OpenAI Gym." arXiv:1606.01540. (2016)
2. P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande. "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics." PLOS Comp. Biol. 13(7): e1005659. (2017)

Equivariational Graph Attention Learning for Protein Structure Prediction

Yi-Ting Chen, Sing-Han Huang, Xiang-Yu Lin, Cheng-Tang Wang, Ching-Yung Lin

Graphen Inc., New York, NY 10110, USA

timchen@graphen.ai, johnhuang@graphen.ai, molalin@graphen.ai, tomwang@graphen.ai, cylin@graphen.ai

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:Y.*

Single structure prediction challenge of protein structure has been mostly solved at CASP14 as shown by AlphaFold2. However, a single protein structure prediction can only solve a limited number of problems in the process of drug development. We found the characteristics of proteins at the atomic scale, using the equivariational graph attention learning model to understand the interaction between branched-chain atoms, can further deduce the direction of the backbone. In order to enable the model to simulate the structure with a small number of coevolutionary features, we have added atomic-level physicochemical features to the feature set. By this way, we can improve the model to become a better structural simulator with a small number of reference sequences. In CASP15, we apply a new way to solve multiple protein interaction problems by using Equivariational Graph Attention Learning to model structures.

Methods

For Co-Evaluational features, we use JackHMMER¹, HHblits² and HH-suite³ tools to search related sequences and templates in UniRef90⁴, BFD^{5,6}, PDB⁷ and MGnify⁸ clusters databases to obtain related sequences.

For physicochemical force features, we use pepdata⁹ and prody¹⁰ tools to find residue-residue force features. Not only using basic peptide features, but we also apply non-covalent bonding forces to present the relation of local side chains by our formular from PDB.

The input features and initial coordinates are used as model inputs for the equivariational graph attentional learning method that produces a variety of predictions including atom-based distances, torsions, atom coordinates, and estimates the per-residue value of the C α -IDDT¹¹. The predicted structures are selected according to the predicted value of model reward. All models were trained using publicly-available structures in the PDB.

Each structure prediction is relaxed using restrained gradient descent on the Amber ff99SB force field¹² using SIMTK-OPENMM¹³.

Availability

The source codes and models are not publicly available at the moment.

1. Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11(1), 431.
2. Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2), 173-175.
3. Steinegger, M. et al. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, 20(1), 1-15.
4. UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1), D506-D515.
5. Steinegger, M., Mirdita, M., & Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods*, 16(7), 603-606.
6. Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1), 1-8.
7. wwPDB consortium. (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research* 47: D520-D528.
8. Mitchell, A. L. et al. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*, 48(D1), D570-D578.
9. Alex Rubinsteyn, Tavi Nathanson, Arun Ahuja. *Openvax/pepdata*, (2018), GitHub repository.
10. Zhang S, Krieger JM, Zhang Y, Kaya C, Kaynak B, Mikulska-Ruminska K, Doruker P, Li H, Bahar I ProDy 2.0: Increased scale and scope after 10 years of protein dynamics modelling with Python 2021 *Bioinformatics*, btab187.
11. Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21), 2722-2728.
12. Hornak, V. et al. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3), 712-725.
13. Eastman, P. et al. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology*, 13(7), e1005659.
12. Altschul, S. F. et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389-3402.

Grudinin

Protein assembly prediction in CASP15 using a combination of physics-based approaches with AlphaFold2 models

S. Grudinin

Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

sergei.grudinin@univ-grenoble-alpes.fr

Key: *Auto:N; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:N; MD:N*

In this CASP round, we assessed bioinformatics tools developed in our lab combined with initial structural predictions of protein assemblies by AlphaFold2.

Methods

We used initial AlphaFold2 models of the monomers and assemblies provided by the Elofsson group. We used NOLB normal modes¹ to sample conformations in T1113, T1115, H1135, T1184, T1185, and T1187 targets. We also used NOLB nonlinear PCA¹ to sample conformations in the H1151 target. SAM ab-initio symmetry-constrained docking method² was applied to targets T1187, H1114, and T1176. Ideal point-group symmetry was applied to the assemblies, according to the provided stoichiometry, with the AnAnaS tool³⁻⁴. Final docking solutions were clustered with 1-5 Å RMSD cutoff. The models were ranked by the initial deviation from the ideal symmetry (as reported by AnAnaS), by the SAM scores, and by pDockQ scores provided by the Elofsson group⁵.

Availability

NOLB and AnAnaS are available on our website at <https://team.inria.fr/nano-d/software/>. SAM is available at <http://sam.loria.fr>.

1. Hoffmann,A., & Grudinin,S. (2017). NOLB: Nonlinear rigid block normal-mode analysis method. *J. Chem. Theory Comput.*, 13(5), 2123-2134.
2. Ritchie,D.W. & Grudinin,S. (2016). Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry. *J. Appl. Cryst.* 49, 158-167.
3. Pagès,G., Kinzina,E., & Grudinin,S. (2018). Analytical symmetry detection in protein assemblies. I. Cyclic symmetries. *J. Struct. Biol.*, 203(2), 142-148.
4. Pagès,G., & Grudinin,S. (2018). Analytical symmetry detection in protein assemblies. II. Dihedral and cubic symmetries. *J. Struct. Biol.*, 203(3), 185-194.
5. Bryant,P., Pozzati,G. & Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.*, 13(1), 1-11.

Multidomain-based protein structure prediction and protein complex structure assembly

Chunxiang Peng, Qianliang Wang, Yuhao Xia, Jun Liu, Kailong Zhao, Minghua Hou, Dong Liu,
Xiaogen Zhou and Guijun Zhang*

1 - College of Information Engineering, Zhejiang University of Technology, Hangzhou, China

zgj@zjut.edu.cn

Key: *Auto:Y; CASP_serv:Y; Templ:Y; MSA:N; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:N; EMA:Y; MD:N*

In this CASP, we used docking-based monomer assembly simulations to generate oligomer structures based on predicted monomer structures. For the monomer structure prediction, we first predict the domain boundaries of the target sequence using in-house domain boundary prediction method DomBPred, and then the structural domains are individually modelled using our template-enhanced AlphaFold2 (Pathreader). Finally, the domain structure models are assembled into full-length structures using proposed SADA protocol. Based on the predicted monomeric structure, we detected structural analogues for generating distance profiles between monomeric structures. Guided by general geometric constraints, distance profile, and oligomer interactions, a two-stage docking-based approach is proposed to generate oligomer structures. The final models are ranked by in-house DeepUMQA.

Methods

For the monomer structure prediction, we combined the domain boundaries prediction results of DomBpred¹ to determine the domain boundary of the target sequence. For single-domain proteins, we used Pathreader to directly predict the full-length structures. For multi-domain proteins, we used Pathreader to predict the single-domain structures, and then SADA² was used to assemble the domain structure models into full-length structures.

Based on the predicted monomeric structures, we used a two-stage docking-based approach to generate oligomer structures guided by general geometric constraints, distance profile, and oligomer interactions.

For the oligomer target, we extracted oligomer interactions from AlphaFold-Multimer, which can be used as a global restraint between monomers to guide overall conformation space sampling⁴. Based on the predicted monomer structures by Pathreader, the initial oligomer structure model is generated under the guidance of the oligomer interactions.

We detected suitable multi-domain protein templates from the constructed MPDB³ and then extracted the distance profiles of residues between different monomers based on the

template information⁴. The distance profiles of residues between different monomers can be regarded as a local restraint, which may be complementary to the oligomer interactions.

The assembly engine for monomer assembly was carried out through the simultaneous rotation and translation of each monomer. Under the guidance of the general geometric constraints (e.g. atom clash), distance profile, and oligomer interactions, a two-stage differential evolution algorithm was proposed to determine the optimal solution. The 5 models that best fit these constraints are selected and ranked by DeepUMQA⁵. The temperature factor values of these models are predicted by DeepUMQA⁵.

Availability

SADA is available at <http://zhanglab-bioinf.com/SADA>

DomBpred is available at <http://zhanglab-bioinf.com/DomBpred>

PAthreader is available at <http://zhanglab-bioinf.com/PAthreader>

DeepUMQA is available at <http://zhanglab-bioinf.com/DeepUMQA>

1. Yu, Z., Peng, C., Liu, J., Zhang, B., Zhou, X., & Zhang, G. (2022). DomBpred: protein domain boundary prediction based on domain-residue clustering using inter-residue distance. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, DOI: 10.1109/TCBB.2022.3175905.
2. Peng, C., Zhou, X., Xia, Y., Liu, J., Hou, M., & Zhang, G. (2022). Structural analogue-based protein structure domain assembly assisted by deep learning. *Bioinformatics*, <https://doi.org/10.1093/bioinformatics/btac553>.
3. Peng, C., Zhou, X., Xia, Y., Zhang, Y., & Zhang, G. (2021). MPDB: a unified multi-domain protein structure database integrating structural analogue detection. *bioRxiv*, 2021.2010.2027.466092.
4. Peng, C., Zhou, X., & Zhang, G. (2022). De novo Protein Structure Prediction by Coupling Contact with Distance Profile. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 19, 395-406.
5. Guo, S., Liu, J., Zhou, X., & Zhang, G. (2022). DeepUMQA: ultrafast shape recognition-based protein model quality assessment using deep learning. *Bioinformatics*, 38, 1895-1903.

Protein structures assembly using inter-domain interactions from deep learning

Yuhao Xia, Kailong Zhao, Chunxiang Peng, Jun Liu, Minghua Hou, Dong Liu,
Haitao Zhu, Xiaogen Zhou and Guijun Zhang*

College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

zgj@zjut.edu.cn

Key: *Auto:Y; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

In CASP15, we developed a pipeline for the assembly of protein monomeric structures as well as oligomeric structures by predicting inter-domain interactions from deep learning, with protein domains as assembly units. We first used the domain segmentation methods to identify the domain boundaries of the target sequence, then predicted each domain individually based on a template-enhanced AlphaFold2 (PATHreader), and then specifically inferred inter-domain interactions through a pre-trained network model. Finally, the predicted interactions were used as the driving force to assemble the protein monomer or oligomer structures.

Methods

The key components of our method include: 1) a deep learning network that predicted inter-domain interactions (including affine transformations of inter-domain residues); 2) a population-based optimization method that assembled the domain structure models into full-length structures by predicted inter-domain interactions.

Inter-domain interaction prediction. We pre-trained a deep learning network model over multi-domain proteins from the Multi-domain Protein Structure Database (MPDB¹⁻²) to specifically predict inter-domain interactions. Given a target sequence, we searched Uniclust30 and BFD databases to generate the multiple sequence alignments (MSAs), and used our three-track alignment-based remote homology template detection method, PATHreader, to search for templates from the PDB and AlphaFold DB³. In addition, we extracted the inter-domain features according to the domain segmentation information of the target sequence, and fed them into an attention-based convolutional neural network together with MSAs and templates, and finally inferred the interactions between domains.

Protein structures assembly. We first used the domain segmentation tool (DomBpred⁴) to partition the target sequence (for oligomer target, the sequence of each monomer was divided into different domains), then used the template-enhanced AlphaFold2⁵ (templates were identified by PATHreader) to predict each domain structure separately, and constructed the atomic coordinate deviation potential according to the predicted inter-domain interactions. Finally, we

assembled the domain structures by a population-based optimization algorithm⁶ to obtain the monomer or oligomer structure. The generated models were evaluated by our model quality assessment method, DeepUMQA⁷, and the top five models were selected.

Availability

MPDB is available at <http://zhanglab-bioinf.com/SADA>

DomBpred is available at <http://zhanglab-bioinf.com/DomBpred>

PAthreader is available at <http://zhanglab-bioinf.com/PAthreader>

DeepUMQA is available at <http://zhanglab-bioinf.com/DeepUMQA>

1. Peng, C., Zhou, X., Xia, Y., Zhang, Y., & Zhang, G. (2021). MPDB: a unified multi-domain protein structure database integrating structural analogue detection. *bioRxiv*, 2021.2010.2027.466092.
2. Peng, C., Zhou, X., Xia, Y., Liu, J., Hou, M., & Zhang, G. (2022). Structural analogue-based protein structure domain assembly assisted by deep learning. *Bioinformatics*, <https://doi.org/10.1093/bioinformatics/btac553>.
3. Varadi, M. et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*, **50**, D439-D444.
4. Yu, Z., Peng, C., Liu, J., Zhang, B., Zhou, X., & Zhang, G. (2022). DomBpred: protein domain boundary prediction based on domain-residue clustering using inter-residue distance. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, DOI: 10.1109/TCBB.2022.3175905.
5. Jumper, J. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583-589.
6. Xia, Y., Peng, C., Zhou, X., & Zhang, G. (2021). A sequential niche multimodal conformational sampling algorithm for protein structure prediction. *Bioinformatics*, **37**, 4357-4365.
7. Guo, S., Liu, J., Zhou, X., & Zhang, G. (2022). DeepUMQA: ultrafast shape recognition-based protein model quality assessment using deep learning. *Bioinformatics*, **38**, 1895-1903.

Multi-structure refinement using multi-objective optimization and graph neural network-based model quality assessemnt

Dong Liu, Minghua Hou, Chunxiang Peng, Yuhao Xia, Jun Liu, Kailong Zhao, Xiaogen Zhou, Biao Zhang and Guijun Zhang *

College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

zgj@zjut.edu.cn

Key: Auto:N; CASP_serv:N; Templ:Y; MSA:N; Fragm:N. Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:Y;

In this CASP, we used three different strategies to refine protein model structures and selected the best structure through the model quality assessment method. For each target, we generated the initial five structure models by our GuijunLab-Threader, GuijunLab-Assembly, GuijunLab-DeepDA, and GuijunLab-RocketX. Then, we further refined the local structure using deep learning geometric constraint adjustment and multi-objective structure optimization developed by our group and the molecular dynamics refinement¹ from Feig Lab. Finally, we utilized a protein model evaluation method, which are based on graph coupled networks, to submit the top five structures.

Methods

Deep learning geometric constraint adjustment. We trained a deep-learning graph neural network (DN1) to predict the geometrically constrained profile, where the sequence and structural features extracted from model were employed. The predicted geometric constraints were then combined with the structural template to guide the refinement of the local structure. Meanwhile, a model quality assessment network (DN2) was used to improve the quality of the local structure and select the best structure.

Multi-objective structure optimization. A two-stage multi-objective population optimization module was constructed to further optimize the screened conformations, mainly driven by the deviation between the residue geometry information derived from the input conformation and those extracted from MSAs and remote homologous structure recognition. Among them, the model was refined by the full-chain three-segment insertion and the secondary structure spatial position adjustment of the loop region in two stages respectively.

Model quality assessment method based on graph coupled networks. To train a coupled network of our method, we screened structures from the Protein data bank² (PDB) and AlphaFold Protein Structure Database³ (AlphaFoldDB), used different methods to generate protein models (i.e. decoys). The method used the following part important features: (1) the geometrical triangle localization feature based on the model structure (2) the predicted

geometrical constraints from GuijunLab-Threader, a template-based enhanced deep learning structure prediction method. The protein structure features were used to evaluate model quality through a coupled neural network, which consists of a graph module that encodes sequence-structure relationships and a transform-based convolution module that decodes structure-quality connection.

Results

We postpone the assessment of the approach until the official release of CASP15 results.

1. Heo, L., & Feig, M. (2018). Improvement of the global structure of template-based models via MD and structural averaging. *Bioinformatics*, **34**, 1063-1065.
2. Berman, H. M., et al. (2000). The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.
3. Varadi, M. et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*, **50**, D439-D444.

Multiple objective population optimization-based protein structure prediction

Minghua Hou, Sirong Jin, Dong Liu, Jun Liu, Chunxiang Peng, Kailong Zhao, Xuhao Xia
and Guijun Zhang*

College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

zgj@zjut.edu.cn

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:Y.MetaG; Fragm:Y; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

For this round of CASP, we developed a multi-objective population optimization modeling pipeline based on the template-enhanced AlphaFold2 (PAthreader), where the main driving force of model building was residue geometric constraints extracted from different coevolutionary data and deep learning models. The resulting conformations were further refined by a loop-specific dihedral angle optimization strategy based on residue distance bias and ranked by a model quality assessment model based on a deep graph coupling network and a protein language model (GraphCPLMQA).

Methods

Multiple sequence alignments (MSAs) of the target sequence were generated by searching Unilust30 (version January 2020) and BFD, which will be used to predict residue constraint information and construct a variable-length fragment library¹ of the target. Among them, the predicted distance information will be used for the three-track alignment-based remote homology template detection method to obtain template information to enhance AlphaFold2² (PAthreader). Further, different residue constraint information was extracted using PAthreader and RoseTTAFold³, which will be combined with the original predicted distance information and used for modeling in a multi-objective population optimization method⁴. This multi-objective strategy can make up for the deficiencies between different methods to improve prediction accuracy, and make it possible to predict different conformations of multiple states. Based on residue distance deviation, conformations were further refined by adjusting the spatial position relationship between the secondary structures using the dihedral angle rotation model of loop region with partial inter-residue distance constraints⁵. Finally, the refined populations were clustered by using the DMscore⁶, the conformations in each class will be ranked by our in-house model quality assessment model (GraphCPLMQA), which is developed based on a deep graph coupling network and a protein language model, and the rank 1 model of each class was outputted.

Availability

GraphCPLMQA is available at <http://zhanglab-bioinf.com/Panda>

PAthreader is available at <http://zhanglab-bioinf.com/PAthreader>

MultiDFold is available at <http://zhanglab-bioinf.com/MultiDFold>

1. Feng, Q., Hou, M., Liu, J., Zhao, K. & Zhang, G. (2022). Construct a variable-length fragment library for de novo protein structure prediction. *Briefings in Bioinformatics*, **23**, bbac086.
2. Jumper, J. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583-589.
3. Baek, M. et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871-876.
4. Hou, M., Peng, C., Zhou, X., Zhang, B. & Zhang, G. (2022). Multi contact-based folding method for de novo protein structure prediction. *Briefings in Bioinformatics*, **23**, bbab463.
5. Liu, J., Zhao, K., He, G., Wang, L., Zhou, X., & Zhang, G. (2021). A de novo protein structure prediction by iterative partition sampling, topology adjustment and residue-level distance deviation optimization. *Bioinformatics*, **38**, 99-107.
6. Zhao, K., Liu, J., Zhou, X., Su, J., Zhang, Y. & Zhang, G. (2021). MMpred: a distance-assisted multimodal conformation sampling for de novo protein structure prediction. *Bioinformatics*, **37**, 4350-4356.

Deep learning-based protein structure prediction and complex model quality assessment

Jun Liu, Kailong Zhao, Guangxing He, Chunxiang Peng, Xuhao Xia, Minghua Hou,
Dong Liu and Guijun Zhang*

College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

zgj@zjut.edu.cn

Key: *Auto:Y; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

Deep learning (DL) has been successfully applied to protein structure prediction and model quality assessment in recent CASP. In this CASP, we used deep learning to predict the inter-residue geometric constraints, model's inter-residue distance deviation and per-residue accuracy, and tried to improve the structure accuracy through iterative geometries prediction, structural folding simulation, and model quality assessment. We also extended the model quality assessment of single chains to complexes.

Methods

Based on the framework of RocketX¹, we designed a protein structure prediction method based on iterative geometries prediction, structural folding simulation and model quality assessment. For a target sequence, MSA and template were searched using HHblits² and our in-house template recognition method PAtreader, respectively. The MSA and template are embedded and fed into the geometries prediction network based on the triangle update and self-attention mechanism to predict the inter-residue geometric constraints. The inter-residue geometric constraints are construct as continuous energy function, and then used to predict the structural models by IPTDFold³, a structure prediction algorithm based on iterative partition sampling, topology adjustment, and residue-level distance deviation optimization. The improved model quality assessment method, DeepUMQA2⁴, is incorporated to assess the quality of the predicted model, i.e. predicting inter-residue distance deviation and per-residue IDDT⁵ of the model. The assessment results will be fed back into a new round of geometric constraint prediction and structural folding simulations. On the one hand, it is used as the dynamic features of the geometries prediction network to re-predict geometric constraints. On the other hand, it is used as an additional constraint of IPTDFold, and combined with the updated geometric constraints to correct the predicted structural model. The final models are generated by five iterations and ranked by model quality assessment.

DeepUQMA⁶ was used to evaluate the structural model of the complex. Firstly, the residue-level USR features and other 1D, 2D and 3D features are extracted from the complex model, and then the IDDT of each residue was predicted by deep residual neural network. The

overall fold accuracy is calculated as the average value of IDDT for all residues, and the overall interface accuracy is calculated as the average value of IDDT of residues being in the interface. We also provide estimates of residues being in the interface.

Availability

DeepUMQA is available at <http://zhanglab-bioinf.com/DeepUMQA>

DeepUMQA2 is available at <http://zhanglab-bioinf.com/DeepUMQA2>

PAthreader is available at <http://zhanglab-bioinf.com/PAthreader>

IPDFold is available at <https://github.com/iobio-zjut/IPDFold>

1. Liu, J., He, G., Zhao, K., & Zhang, G. (2022). De novo protein structure prediction by incremental inter-residue geometries prediction and model quality assessment using deep learning. bioRxiv, doi: <https://doi.org/10.1101/2022.01.11.475831>
2. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., & Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics*, **20**, 1-15.
3. Liu, J., Zhao, K., He, G., Wang, L., Zhou, X., & Zhang, G. (2021). A de novo protein structure prediction by iterative partition sampling, topology adjustment and residue-level distance deviation optimization. *Bioinformatics*, **38**, 99-107.
4. Liu, J., Zhao, K., & Zhang, G. (2022). Improved model quality assessment using sequence and structural information by enhanced deep neural networks. bioRxiv, doi: <https://doi.org/10.1101/2022.08.12.503819>
5. Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, **29**, 2722-2728.
6. Guo, S., Liu, J., Zhou, X., & Zhang, G. (2022). DeepUMQA: ultrafast shape recognition-based protein model quality assessment using deep learning. *Bioinformatics*, **38**, 1895-1903.

Protein structure prediction based on enhanced remote homologous template recognition

Kailong Zhao, Yuhao Xia, Fujin Zhang, Jun Liu, Guangxing He, Zhaohong Huang, Chunxiang Peng, Minghua Hou, Dong Liu, Guijun Zhang*

College of Information Engineering, Zhejiang University of Technology, HangZhou 310023, China

* zgj@zjut.edu.cn

Key: Auto:Y; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:Y.8-10; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N

Recognition of remote homologous template is a necessary module for deep learning-based modeling methods. In CASP15, we used the proposed PAtreader to recognize remote templates and predict 3D structures. First, we constructed the template library based on PDB and AlphaFold DB with a threshold of 80% structural similarity. The structure profile extracted from the PAcluster80 and the distance profile predicted by our in-house program. Then, the structure profile and distance profile are aligned by a three-track alignment and get the maximum alignment score (alignScore). As an effective supplement to alignScore, the physical and geometric features of the alignment structure are extracted and fed into the convolutional network with self-attention to predict DMScore (pDMScore)¹, which is linearly weighted with the alignScore for ranking templates. Finally, we used the identified remote templates to enhance AlphaFold2 for protein structure prediction, and ranked the predicted models through our in-house model quality assessment (GPSEM).

Methods

Construction of template library. We constructed a template library based on PDB and AlphaFold DB². We removed structures with 100% sequence identity from the PDB, since identical sequences often correspond to very similar structures. Then, we calculated the structural similarity of the retained 106,275 proteins using TM-align and classified them into 34,701 structural classes by an 80% similarity threshold. We used a greedy incremental clustering approach similar to CD-HIT³, which avoids many pairwise structure alignments. On this basis, we further extended the structure of AlphaFold DB. We selected 100,912 structures with pLDDT ≥ 90 as available templates and clustered with PDB according to 80% structure similarity, resulting in 55.7% of the structures could be classified into 34,701 PDB clusters and the remaining structures could form 22,105 new clusters.

Three-track alignment. To take full advantage of the deposited structure information to identify templates, we developed a three-track alignment (residue-residue alignment, residue pair-residue pair alignment, and distance profile-structure profile alignment) with two stages. In the first stage, the optimal N_{clu} structural clusters are identified by three-track alignment between the query sequence and the representative structures of clusters. In the second stage, the optimal templates are identified from the structures within the clusters determined in the first stage by

repeating the three-track alignment. The purpose of the three-track alignment is to find an optimal alignment between query sequence and template sequence by maximizing the alignScore. The first track is to calculate the protein-specific score matrix and find the optimal sequence alignment by dynamic programming. The protein-specific score matrix is obtained from second track by a second dynamic programming to find the optimal residue pair alignment that only considered the inter-residue distance. The residue pair alignment is performed based on the construction of residue pair-residue pair score matrix, where the values are calculated from the third track by maximizing the probability product and minimizing the distance difference.

Modeling and folding. In structure modeling, we enhanced AlphaFold2⁴ with the template recognition, which provides accurate local atomic information for single-domain proteins and accurate domain orientation for multi-domain proteins. Furthermore, the templates are aligned by TM-align and the frequency distribution of residues is calculated based on different distance deviation thresholds and secondary structures, which are used to identify folding intermediates for exploring folding pathway.

Model quality assessment. We used PDB and AlphaFold DB to construct a non-redundant dataset with a 35% sequence similarity, 80% of which was divided into training set and 20% as validation set. For each protein in this dataset, we generated 100 decoys structures using the dihedral angle perturbation method. Using this dataset, we retrained a neural network model for global protein model quality assessment based on equivariant graph neural network⁵. The input features of the model include geometric information and physicochemical energy information of protein structure. We used the neural network model to score the structure predicted in the previous step, and select the best model according to the score.

Availability

PAthreader is available at <http://zhanglab-bioinf.com/PAthreader>

1. Zhao, K., Liu, J., Zhou, X., Su, J., Zhang, Y. & Zhang, G. (2021). MMpred: a distance-assisted multimodal conformation sampling for de novo protein structure prediction. *Bioinformatics*, **37**, 4350-4356.
2. Varadi, M. et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*, **50**, D439-D444.
3. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150-3152.
4. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583-589 (2021).
5. Satorras, VcG., Hoogeboom, E. & Welling, M. (2021). E (n) equivariant graph neural networks. In: International Conference on Machine Learning. *PMLR*. p. 9323-9332.

SAVARNA: Structure Assembly via Alignment of RNA Secondary Structures

Xiangyun Qiu

Department of Physics, George Washington University, Washington DC 20052

xqiu@gwu.edu

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:Y*

RNA folding is a hierarchical process in which secondary structural motifs are the first glimpses of an orderly organization, followed by dynamic optimizations of tertiary contacts. As such, RNA secondary structural elements are the basic building blocks for constructing tertiary structures and the 2D layouts of secondary structures serve as the blueprints for 3D folding. In recognition of their close connections and in the light of recent advances in RNA secondary structure prediction, we have sought to investigate the feasibility of using RNA secondary structure similarity as the key metric for searching for candidate tertiary structures. The structure candidates provide the backbones for assembling target atomic structures to be relaxed with physics-based energy minimization.

Methods

We first build a PDB library of known RNA 3D structures by downloading all RNA PDB structures and removing redundant sequences with sequence identity over 80% (via CD-HIT-EST1). Then, for a given target RNA sequence, our pipeline goes through the following steps to predict 3D structure models.

- 1) The secondary structure of the target sequence is predicted with a home-developed deep learning (DL) model, SeqFold2D2;
- 2) The predicted secondary structure is aligned against the PDB library with three secondary structure pairwise alignment programs (Gardenia3, RNAforester4, and RNAdistance4);
- 3) PDB Structure candidates are selected from the PDB library with best structure alignment scores;
- 4) The backbones of the structure candidates are used to assemble 3D structure models by filling in full atomic details and adding missing residues;
- 5) The structure models are relaxed with the molecular dynamics (MD) program NAMD5;
- 6) Confidence scores are estimated as the absolute displacement of each residue before and after energy minimization after structure alignment.

Results

While our methodology strongly depends on the existence of similar RNA secondary structures in our PDB library with just ~600 structures, we found over the period of the CASP15 that reasonable structure candidates can be found for almost all target sequences and energy minimization can proceed without major issues. This observation may suggest that the diversity of RNA secondary and tertiary structures is rather limited, especially compared with that of proteins, such that a rather small library provides decent coverage of the structure space. It is also expected that additional refinements of thus obtained structures can further improve the quality of prediction. Nonetheless, this method is expected to perform poorly for out-of-distribution sequences. Work is in progress to make better use of molecular dynamics simulations to refine obtained structure models, as well as to incorporate DL-based methods to improve the generalizability of the pipeline.

Availability

The DL model for RNA secondary structure prediction is available at <https://github.com/qiuresearch/SeqFold2D>. The complete pipeline is still under active development and will be made available at the earliest.

1. Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28, 3150-3152.
2. Qiu, X. (2022) Decisive Roles of Sequence Distributions in the Generalizability of de novo Deep Learning Models for RNA Secondary Structure Prediction. *bioRxiv*, 2022.2006.2029.498185.
3. Blin, G., Denise, A., Dulucq, S., Herrbach, C. and Touzet, H. (2010) Alignments of RNA Structures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7, 309-322.
4. Lorenz, R., Bernhart, S.H., Honer Zu Siederdisen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol*, 6, 26.
5. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L. and Schulten, K. (2005) Scalable molecular dynamics with NAMD. *J Comput Chem*, 26, 1781-1802.

HADDOCK scoring of CAPRI Round54 models.

Marco Giulini, Rodrigo V. Honorato, Charlotte van Noort, Joao M.C. Teixeira
and A.M.J.J. Bonvin

Computational Structural Biology, Faculty of Science - Chemistry, Utrecht University, Utrecht, the Netherlands

a.m.j.j.bonvin@uu.nl

Key: Auto:N; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:Y

The HADDOCK team participated as a scorer group in the CAPRI 54 round. The scoring protocol was entirely based on the energetics of the complex, with no information coming from other sources such as deep learning methods or bioinformatic predictions. The scoring module of the new modular version of HADDOCK, HADDOCK3¹ (<https://github.com/haddocking/haddock3/>) was used throughout the whole round (in its beta version).

Methods

The HADDOCK scoring pipeline consists of few different building blocks, namely preprocessing including topology generation, energy minimisation and scoring, and Fraction of Common Contacts (FCC) clustering.

In the *preprocessing* stage the input ensemble is parsed and potentially problematic models are identified and corrected. Topologies are created for each model and missing atoms are added. The protonation state of histidine residues and the presence of disulfide bonds are automatically handled.

The *emscoring* step follows, which consists of a short energy minimisation (50 steps) carried out with the OPLS² force field. The minimized models are then ranked based on their HADDOCK score³:

$$HS = 1.0 E_{vdw} + 0.2 E_{ele} + 1.0 E_{desolv}$$

where E_{vdw} and E_{elec} correspond to the intermolecular van der Waals and electrostatic energies, respectively, and E_{desolv} is a solvent accessible surface area-dependent empirical desolvation energy term⁴.

During the last step of the scoring process, the structures are clustered using the FCC clustering algorithm⁵. This procedure lumps together models that share a consistent part of their interfacial contacts. A minimum of four models is required to form a cluster, and the input structures that fail to satisfy this criterion are labeled as “unclustered”.

The model selection is based on the HADDOCK score of each cluster, calculated as the average score of the best four models. A short visual inspection is carried out to exclude biologically implausible complexes, whose excellent scores are mainly due to unphysically optimal energetics. For submission, we typically select one model for each of the top five clusters as top 1-5 predictions. Positions 6-10 are usually filled with additional models coming either from the top clusters or from other clusters, depending on the case and number of clusters. Occasionally, unclustered structures with a particularly good HADDOCK score are also considered.

Results

The results for the CAPRI 54 round are yet to be published and will be summarized once available.

Availability

HADDOCK3 is freely available from <https://www.bonvinlab.org/software/haddock3/>

1. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: A Protein–Protein Docking Approach Based on Biochemical or Biophysical Information. *J Am Chem Soc.* 2003;125(7):1731-1737.
2. Jorgensen WL, Tirado-Rives J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc.* 1988;110(6):1657-1666
3. Vangone A, Rodrigues JPGLM, Xue LC, et al. Sense and simplicity in HADDOCK scoring: Lessons from CASP-CAPRI round 1. *Proteins Struct Funct Bioinforma.* 2017;85(3):417-423
4. Fernández-Recio J, Totrov M, Abagyan R. Identification of protein-protein interaction sites from docking energy landscapes. *J Mol Biol.* 2004;335(3):843-865
5. Rodrigues J.P.G.L.M., Trellet M., Schmitz C., Kastiris P.L., Karaca E., Melquiond A.S.J. and Bonvin A.M.J.J. Clustering biomolecular complexes by residue contacts similarity. *Proteins: Struc. Funct. & Bioinformatic.* 2021;80:1810-1817

Protein structure prediction in CASP15 through MSA-based HelixFold and MSA-free HelixFold-Single

Yingfei Xiang, Lihang Liu, Yang Xue, Dayong Lin, Xiaomin Fang and Fan Wang

Baidu Inc., Shenzhen, Guangdong, China

xiangyingfei01@baidu.com, fangxiaomin01@baidu.com, wangfan04@baidu.com

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; MetaG; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

We developed HelixFold¹ and HelixFold-Single², and combined them for protein structure prediction in CASP15. HelixFold and HelixFold-Single are mainly based on the pipeline of AlphaFold2³, where HelixFold relies on the MSA to extract the co-evolution knowledge, while HelixFold-Single is an MSA-free method that takes advantage of a large-scale protein language model. Large-scale unlabeled data and more distillation data are collected for model training. Besides, for each target, many variant versions of HelixFold and HelixFold-Single are utilized to produce the candidate structures, and the best are selected and submitted to CASP15.

Methods

Abundant Data: We collected about 200 million targets from UniRef30⁴ (2021-03) and hundreds of thousands of protein structures from RCSB PDB^{5,6}. We also produced and collected about one million self-distillation protein structure data from UniClust30⁷ and AlphaFold Protein Structure Database⁸. We applied HHblits^{9,10} and JackHMMER¹¹ to search against BFD^{12,13}, UniRef90⁴ and MGnify¹⁴ clusters for generating MSAs separately. In addition, we used HHsearch^{9,10} tool to find potential templates from PDB70.

HelixFold and HelixFold-Single: HelixFold is an MSA-based protein structure prediction pipeline based on the architecture of AlphaFold2. Compared with the original AlphaFold2, we utilized much more distillation data for training HelixFold to improve the prediction accuracy. HelixFold-Single is an MSA-free model, combining the protein language model and the geometric modeling for protein structure prediction. The protein language model is served as an alternative to MSA to extract the co-evolution. First, the protein language model is trained with 200 million unlabeled targets. Second, the protein language model and the geometric modeling are jointly trained with the structure data to provide efficient and accurate protein structure predictions.

Ensemble of Multiple Variants: With the increase in the number of recycling and the number of layers in the structure module, we found that the accuracy of some targets can achieve further improvement. Consequently, for each target, we predict the candidate structures by

multiple variant versions of HelixFold and HelixFold-Single with the different numbers of recycling and layers in the structure module. The top five predictions are selected according to pLDDT and pTM³ scores.

Availability

The data and tools for data processing used in our system are publicly available. We provide online service for HelixFold at <https://paddlehelix.baidu.com/app/drug/protein/forecast> and HelixFold-Single at <https://paddlehelix.baidu.com/app/drug/protein-single/forecast>.

1. Wang, G., Fang, X., Wu, Z. et al. (2022). Helixfold: An efficient implementation of alphafold2 using paddlepaddle. arXiv preprint arXiv:2207.05477.
2. Fang, X., Wang, F., Liu, L. et al. (2022). HelixFold-Single: MSA-free Protein Structure Prediction by Using Protein Language Model as an Alternative. arXiv preprint arXiv:2207.13921.
3. Jumper, J., Evans, R., Pritzel, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583-589.
4. UniProt Consortium. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31(6), 926-932.
5. Berman, H. M., Westbrook, J., Feng, Z. et al. (2000). The protein data bank. *Nucleic Acids Res.* 28(1), 235-242.
6. Burley, S. K., Bhikadiya, C., Bi, C. et al. (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* 49(D1), D437-D451.
7. Mirdita, M., Von Den Driesch, L., Galiez, C. et al. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* 45(D1), D170-D176.
8. Varadi, M., Anyango, S., Deshpande, M. et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50(D1), D439-D444.
9. Remmert, M., Biegert, A., Hauser, A. et al. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* 9(2), 173-175.
10. Steinegger, M., Meier, M., Mirdita, M. et al. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 20, 1-15.
11. Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* 11(1), 1-8.
12. Steinegger, M., Mirdita, M., & Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods* 16(7), 603-606.
13. Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications* 9(1), 1-8.
14. Mitchell, A. L., Almeida, A., Beracochea, M. et al. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* 48(D1), D570-D578.

hks1988 (TS)

Selection of a good model by single model accuracy estimation method SARTliddt in CASP15

Kun-Sop Han, Chol-Song Kim and Myong-Chol Ma

University of Science, Unjong-District, Pyongyang, DPR Korea

hks1981@star-co.net.kp

We participated in TS category of CASP15 with group "hks1988" and number 354. We submitted a model which is selected from the released TS models (BAKER-SERVER and AlphaFold models, about 15 models) by using our protein model accuracy estimation method SARTliddt. The local accuracy of the submitted model is the same as the one of the selected model and the global accuracy is SARTliddt_G.

A new single model-based local quality score SARTliddt_L is based on linear combination of 4 components extracted from a sphere centered on the residue of interest. For local score, linear regression analysis is performed between 4 components and LDDT of the interested residue. The global score SARTliddt_G is the average of SARTliddt_L. 31369 CASP (CASP9, CASP11 and CASP13) models are used as training set.

Methods

Calculation of four terms in a sphere (step1):

In the first step, we compute four terms in a sphere (radius 12 Å) centered on the residue of interest. The first component is $SS_8 = SS_{8_N} / N$. N is the total number of residues within the sphere. SS_{8_N} is 8 states-agreement number between the predicted (by SSpro8 of SCRATCH1.3^[1]) and the calculated secondary structure (by DSSP^[2]) of amino acid residues within the sphere.

The second component is $SS_{8BIN} = SS_{8BIN_N} / N$. SS_{8BIN_N} is the number of residues on which the predicted secondary structure from primary sequence equals to the calculated one from protein model and the predicted binary solvent accessibility (by ACCpro of SCRATCH1.3) equals to the transformed one. To transform the decimal solvent accessibility calculated by DSSP into binary state, decimal solvent accessibility is divided by maximum one of corresponding amino acid residue. Then, if the divided solvent accessibility is bigger than 0.25, it is in the exposed state. If not, the buried state.

The third is $TOR_{GHA} = TOR_{GHA_N} / N$. TOR_{GHA_N} indicates how well the predicted torsion angles are reproduced in model. That is, it is the number of residues in the sphere on which the difference of the predicted and the calculated ψ is less than 40. The torsion angle is predicted by our torsion angle predictor PredTOR based on deep residual neural network. The mean absolute error (MAE^[3]) between the predicted and the real value of ϕ is 16.58 and that of ψ is 18.63 for the testing set consisting of 716 protein chains.

The last is $ACC20_{GHA} = (P_5 + P_{10} + P_{20} + P_{40})/4$. To compute $ACC20_{GHA}$, we first calculate differences of the predicted and the calculated solvent accessibilities of residues in the sphere. Then, we compute P_5 , P_{10} , P_{20} and P_{40} . Here, P_i is percentage of residues of having the difference smaller than a threshold i in the sphere. The final $ACC20_{GHA}$ is an arithmetic mean of four percentages obtained using 4 thresholds 5, 10, 20 and 40, the same ones applied to calculate GDT-HA score^[4].

Smoothing by the second window (step2):

After calculating four terms in the sphere of residue of interest (step1), we conduct smoothing by a window (step2). The so-called “smoothing by the window” means averaging the terms obtained in step1 in linear sliding window (window size=5) centered on the residue of interest.

Calculation of SARTliddt (step3):

The SARTliddt_L is based on linear combination of 4 components described above (step1 and step2). Weights of 4 components and constant term are obtained by linear regression analysis between 4 components and LDDT score calculated from 6887535 residues of 31369 CASP (CASP9, CASP11 and CASP13) models.

SARTliddt_G is the average of SARTliddt_L.

Selection of a good model (step4):

We selected a model with the highest SARTliddt_G score from the released TS models (BAKER-SERVER and AlphaFold models, about 15 models). The coordinates and local accuracy of the submitted model is the same as the ones of the selected model and the global accuracy is SARTliddt_G.

1. Cheng,J., et al. (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.* 33, W72–W76.
2. Kabsch,W., Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577-2637.
3. Wu S, Zhang Y (2008) ANGLOR: A Composite Machine-Learning Algorithm for Protein Backbone Torsion Angle Prediction. *PLoS ONE* 3(10): e3400. doi:10.1371/journal.pone.0003400
4. Battey, JN, et al. Automated server predictions in CASP7. *Proteins*. 2007;69(Suppl. 8):68–82.
5. Jianlin Cheng, et al. Estimation of model accuracy in CASP13. *Proteins*. 2019; 87:1361–1377
6. Liu,T., Wang,Y., Eickholt,J. & Wang,Z. (2016) Benchmarking Deep Networks for Predicting Residue-Specific Quality of Individual Protein Models in CASP11. *Scientific Report*, 6, 1930.

Automated Prediction of Protein Tertiary Structures with Local Model Quality Scores Using the IntFOLD7 Server

L.J. McGuffin, S.M.A. Alharbi, B.R. Salehe, R. Adiyaman

School of Biological Sciences, University of Reading, Reading, UK

l.j.mcguffin@reading.ac.uk

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y; Fragm:Y.v; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

The IntFOLD server¹ integrates our latest methods for tertiary structure (TS) prediction, domain boundary prediction, prediction of intrinsically disordered regions, prediction of protein-ligand interactions and quality assessment (QA) of predicted 3D models of proteins. Following the successes of our previous IntFOLD servers^{2,3}, which used ModFOLD variants⁴ to rank models, our initial focus for the IntFOLD7 server at CASP15 was the further improvement of global model ranking and local model quality scoring, using our newly improved ModFOLD9 method. In addition, we integrated two new deep learning methods for tertiary structure modelling: LocalColabFold⁵ (1.0.0) and trRosetta2⁶.

Methods

For CASP15, a bespoke version of the IntFOLD7 server was developed in order to return appropriately formatted results for the tertiary structure (TS) prediction category. Additionally, the local quality assessment predictions (QMODE3) using ModFOLD9 were returned as scores in the B-factor column of each TS model file. Our TS method was developed with the aim of fixing local errors, identified in an initial pool of single template models, through iterative multi-template modelling. The method attempts to exploit our previous CASP successes in accurately predicting local errors in our models⁷ by taking the global and local per-residue errors into consideration during the multiple template selection stage⁸. For the initial fold recognition stage, 14 different methods were installed and run in-house to generate up to 10 sequence-to-structure alignments each, resulting in up to 140 alternative single-template-based models being generated for each CASP target. The following fold recognition methods were used: SP3⁹, SPARKS2⁹, HHsearch¹⁰, COMA¹¹, SPARKSX¹², CNFsearch¹³ and the 8 alternative threading methods that are integrated with the LOMETS package¹⁴ (PPA, dPPA, dPPA2, sPPA, MUSTER, wPPA, wdPPA and wMUSTER).

In the next stage, all single-template models were assessed using ModFOLDclust2¹⁵ in order to assign global and local model quality scores. Using the single template model quality scores, and other criteria involving template coverage, the corresponding alignments were then selected from each fold recognition method and used to build multiple-template models, with the

aim of minimizing local errors in the final models. The alternative multi-template modelling alignment selection methods resulted in the generation of a new population of up to 124 multi-template models for each target. Additionally, I-TASSER *light*¹⁶ (for targets <500 residues; run in “light mode” with wall-time restricted to 5h), HHpred¹⁷, DMPfold¹⁸, trRosetta²⁶ and LocalColabFold⁵ version 1.0.0 were used to generate up to 5 models each, which were then added to the final pool of alternative multi-template models for ranking. In the final stage of the IntFOLD7 method, the models in the final reference set were then evaluated using ModFOLD9 (described below) and the top 5 ranked models were submitted.

ModFOLD9 is the latest update to our server for evaluating the quality of tertiary structure models. The ModFOLD9 protocol builds on that of ModFOLD8⁴ by including 6 new integrated scoring methods: 3 new Contact Distance Agreement (CDA) scores, and the 3 variants of the DeepAccNet¹⁹ methods (DeepAccNet, DeepAccNet-Bert and DeepAccNet-MSA). Our CDA scores measure the agreement between the residue contacts predicted from the target sequence and the measured Euclidean distance (in Å) between residues in the predicted 3D model³. The contact predictions from trRosetta²⁶, DeepDist²⁰, and TripletRes²¹, were used for the three new CDA scores, CDA_trR2 and CDA_DD and CDA_TR respectively. As in previous versions of ModFOLD, neural networks were then used to combine the component per-residue/local quality scores from each of the scoring methods, resulting in a final consensus of per-residue quality scores for each model. For each TS model, the model rankings and predicted per-residue quality scores (pLDDT*100) from ModFOLD9 were added to the B-factor column for each set of atom records.

For tertiary structure targets >1200 residues, due to lack of resources, we were unable to complete predictions from all of the IntFOLD7 component methods within 72h. Therefore, for targets >1200 residues we returned our TS models that were already completed using our MultiFOLD protocol (see our MultiFOLD abstract for details).

Results

IntFOLD7 and ModFOLD9 are continuously benchmarked using the CAMEO resource²². According to the 3D and QE benchmark results, both of our new servers show improved performance over our previous versions of those methods and they are competitive with the other public servers in their respective categories.

Availability

The IntFOLD7 server is available at:

https://www.reading.ac.uk/bioinf/IntFOLD/IntFOLD7_form.html

1. McGuffin,L.J., Adiyaman,R., Maghrabi,A.H.A., Shuid,A.N., Brackenridge,D.A., Nealon,J.O. & Philomina,L.S. (2019) IntFOLD: an integrated web resource for high performance protein structure and function prediction. *Nucleic Acids Res.* 47, W408-W413, doi: 10.1093/nar/gkz322
2. Kryshchak,A., Monastyrskyy,B., Fidelis,K., Moutl,J., Schwede,T., Tramontano,A. (2018) Evaluation of the template-based modeling in CASP12. *Proteins.* 86 S1, 321-334. doi: 10.1002/prot.25425.
3. McGuffin, L.J., Shuid, A.N., Kempster, R., Maghrabi, A.H.A., Nealon J.O., Salehe, B.R., Atkins, J.D. & Roche, D.B. (2018) Accurate Template Based Modelling in CASP12 using the IntFOLD4-TS, ModFOLD6 and ReFOLD methods. *Proteins.* 86 S1, 335-344. doi: 10.1002/prot.25360.
4. McGuffin,L.J., Aldowsari, F.M.F., Alharbi,S.M.A., & Adiyaman,R. (2021) ModFOLD8: accurate global and local quality estimates for 3D protein models. *Nucleic Acids Research.* 49(W1), W425–W430. doi: 10.1093/nar/gkab321
5. Mirdita,M., Schütze,K., Moriwaki,Y., Heo,L., Ovchinnikov,S., & Steinegger,M. (2022) ColabFold: making protein folding accessible to all. *Nature Methods.* 19(6), 679–682. doi: 10.1038/s41592-022-01488-1
6. Anishchenko,I., Baek,M., Park,H., Hiranuma,N., Kim,D.E., Dauparas,J., Mansoor,S., Humphreys,I.R., & Baker,D. (2021) Protein tertiary structure prediction and refinement using deep learning and Rosetta in CASP14. *Proteins.* 89(12), 1722–1733. doi: 10.1002/prot.26194
7. McGuffin,L.J., Roche,D.B. (2011) Automated tertiary structure prediction with accurate local model quality assessment using the IntFOLD-TS method. *Proteins.* 79 S10, 137-46. doi: 10.1002/prot.23120
8. Buenavista,M.T., Roche,D.B., McGuffin,L. J. (2012) Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. *Bioinformatics.* 28, 1851-1857. doi: 10.1093/bioinformatics/bts292
9. Zhou,H., Zhou,Y. (2005) SPARKS2 and SP3 servers in CASP6. *Proteins.* 61 S7, 152-156. doi: 10.1002/prot.20732
10. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics.* 21, 951-96. doi: 10.1093/bioinformatics/bti125
11. Margelevičius,M., Venclovas,Č. (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparisons. *BMC Bioinformatics.* 11, 89. doi: 10.1186/1471-2105-11-89
12. Yang,Y., Faraggi,E., Zhao,H., Zhou,Y. (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics.* 27, 2076-2082. doi: 10.1093/bioinformatics/btr350
13. Ma,J., Wang,S., Zhao,F., Xu,J. (2013) Protein threading using context-specific alignment potential. *Bioinformatics.* 29, i257-65. doi: 10.1093/bioinformatics/btt210
14. Wu,S. and Zhang,Y. (2007) LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research.* 35, 3375-3382. doi: 10.1093/nar/gkm251
15. McGuffin,L.J. & Roche,D.B. (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics.* 26, 182-188. doi: 10.1093/bioinformatics/btp629
16. Roy,A., Kucukural,A., Zhang,Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols.* 5, 725-738. doi: 10.1038/nprot.2010.5

17. Meier,A., Söding,J. (2015) Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling. *PLoS Comput Biol.* 11, e1004343. doi: 10.1371/journal.pcbi.1004343
18. Greener,J.G., Kandathil,S.M. & Jones,D.T. (2019) Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat Commun.* 10, 3977. doi: 10.1038/s41467-019-11994-0
19. Hiranuma, N., Park, H., Baek, M., Anishchenko, I., Dauparas, J., & Baker, D. (2021). Improved protein structure refinement guided by deep learning based accuracy estimation. *Nature Communications.* 12(1), 1340. doi: 10.1038/s41467-021-21511-x
20. Wu,T., Guo,Z., Hou,J., & Cheng,J. (2021) DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. *BMC Bioinformatics.* 22(1), 30. doi: 10.1186/s12859-021-03960-9
21. Li,Y., Zhang,C., Bell,E.W., Zheng,W., Zhou,X., Yu,D.-J., & Zhang,Y. (2021). Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *PLoS Computational Biology.* 17(3), e1008865. doi: 10.1371/journal.pcbi.1008865
22. Haas, J., Barbato, A., Behringer, D., Studer, G., Roth, S., Bertoni, M., Mostaguir, K., Gumienny, R., Schwede, T. (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins.* 86 S1, 387-398. doi: 10.1002/prot.25431

Integrated structure modeling protocol for human and server prediction for biomolecular structures

C.W. Christoffer¹, A.J. Jain², Y. Kagaya², H. Kannan^{2,3}, T. Nakamura², G. Terashi², J.C. Verburg², Y. Zhang¹, Z. Zhang¹, Hayato Fujita⁴, Masakazu Sekijima⁴, and D. Kihara^{1,2}

1- Department of Computer Science, Purdue University, West Lafayette, IN, USA, 2 - Department of Biological Sciences, Purdue University, West Lafayette, IN, USA, 3 - Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, India, 4 Department of Computer Science, Tokyo Institute of Technology, Yokohama, 226-8501, Japan

Authors from Purdue are listed in alphabetical order. Correspondence: dkihara@purdue.edu

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

Our group participated in the TS (Tertiary Structure), H (heteromeric assembly), and LG (Ligand) prediction categories for both protein and RNA targets.

Methods

Single-chain protein structure prediction: We generated MSAs for the deep learning network inputs using a following two-step strategy. For a query sequence, we first performed 3 iterations of search on UniRef30 (2021_03) and UniRef_90¹ (2022_01) with hhblits² and jackhmmer³, respectively, to obtain two MSAs. These two MSAs were combined and then used as input profiles for the second step. In the second step, the MGnify⁴ (2019_05), Metaclust_nr⁵ (2018_06), and BFD (Latest) databases were used. The search was performed with only a single iteration with jackhmmer and hhblits. In addition, the MSAs used in ColabFold⁶ were computed independently using mmseqs2⁷ and their corresponding database. All resulting MSAs from both first and second step were simply concatenated and then used as input for the network.

In addition to predictions with the AlphaFold2⁸ that were made available by DeepMind, we trained a separate structure module for single-chain prediction. The input for this structure module was the pair and single representations generated from the regular AlphaFold inference pipeline. The training dataset contained around 30k sequences with 25% sequence similarity. Single chain targets with 25% sequence similarity to the training set were used as a validation set. We used all losses along with its weight settings as claimed in the AlphaFold paper, except for MSA and distogram losses, as they are not applicable to the structure module. Using both original AlphaFold2 and the model we trained by ourselves, in total, we had 10 predictions for each target. Predicted structures were ranked based on the mean pLDDT values. Literature information were used for model selection and modifying models when available. When necessary, manual modification of input MSA as well as the resulting structure models was also performed.

Protein complex assembly: For protein complexes, MSAs generated as above were used to generate AlphaFold models. Models were generated using Alphafold-Multimer and with other existing implementation on servers. Models from different sources were ranked by calibrated model quality estimates. For large prediction targets where reasonable models could not be generated in a single inference, models were broken down, either by reducing stoichiometry or separating plausible domains, and combined using minimization by phenix.geometry_minimization⁹. For homomeric targets where AlphaFold-Multimer still failed to generate reasonable models, we used SAM¹⁰ for symmetrical docking. We first ran AlphaFold-Multimer on repeated target sequences and obtained subunit structures, which were fed to SAM to produce complexes with various symmetries. We essentially ranked the structures using the score output by SAM. Manual modifications of models and input MSAs were performed when necessary.

RNA structure prediction: We mainly used Rosetta FARFAR2¹¹ for modeling RNA targets with secondary structure constraints generated from IPknot++¹². For some targets we manually edited the secondary structure constraints on the basis of literature information as well as particular observations, such as complementarity of bases in the target sequences. Based on literature as well as sequence search in RNACentral and BLAST nt, if suitable 3D structure templates were found, rna_thread in Rosetta was used for a sequence alignment with the target and to generate a (partial) template for the modeling of the aligned region. We used rna_score in Rosetta and ARES¹³ as well as Ranksum of the two scores for model selection as the sum of ranks of independent scores performed well in our experience in protein docking¹⁴. For the RNA-protein complex targets, T1189/R1189 and T1190/R1190, we modeled protein and RNA separately then placed them by superimposing onto template structures, after which we performed phenix.geometry_minimization.

Protein- and RNA-ligand complex assembly: GLIDE XP¹⁵ and Induced Fit Docking (IFD)¹⁶ tools from the Schrödinger software suite were used for RNA and Protein-ligand complex assembly when the ligand was a small molecule. We compared rigid-body docking (GLIDE) with flexible side-chain docking (IFD) and selected the protein-ligand complexes with the best docking score. AutoDock^{17,18} was used when the ligand was a metal ion. In cases where structural templates that had the same ligand were found for the given target sequences on PDB, we superimposed the ligand from the template onto our modeled target using PyMOL¹⁹ followed by restrained Molecular Dynamics simulations or geometry optimization with the Merck Molecular Force Field (MMFF94) implemented in PyMOL to clear steric clashes.

Automatic server: We used an automatic protein structure prediction pipeline to submit both monomer and complex TS targets as Kiharalab_Server²⁰. For monomeric protein modeling, the deep learning method described above was used alongside AlphaFold. pLDDT predictions were calibrated so that model quality estimates from both models could be directly compared. The top 5 models by pLDDT were then submitted. For multimeric protein modeling, AlphaFold-Multimer was run using MSAs generated as described above. In cases where GPU memory was exhausted before models were output, the size or stoichiometry of the input was reduced until

models were generated. For RNA modeling, 10 replicas of SimRNA²¹ were run for 1e6 iterations each and ranked by their output scores.

Acknowledgements

Xiao Wang helped in modeling some RNA targets. We are grateful for ITaP Research Computing at Purdue University for providing us additional computational resources for this project. This work is partially supported by the National Institutes of Health (R01GM133840, R01GM123055), the National Science Foundation (CMMI1825941, MCB1925643, DBI2146026, IIS2211598, and DBI2003635). C.C. and J.V. were supported by a National Institute of General Medical Sciences-funded predoctoral fellowship to C.C. and J.V. (T32 GM132024).

1. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., and Wu, C. H. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31, 926–932. doi: 10.1093/bioinformatics/btu739.
2. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* 41, e121–e121. doi: 10.1093/NAR/GKT263.
3. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 20, 1–15. doi: 10.1186/s12859-019-3019-7.
4. Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., et al. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res* 48, D570–D578. doi: 10.1093/NAR/GKZ1035.
5. Steinegger, M., and Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications* 2018 9:1 9, 1–8. doi: 10.1038/s41467-018-04964-5.
6. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods* 2022 19:6 19, 679–682. doi: 10.1038/s41592-022-01488-1.
7. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* 2017 35:11 35, 1026–1028. doi: 10.1038/nbt.3988.
8. Jumper, J. et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583–589.
9. Liebschner, D. et al. (2019) Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Cryst Sect D Struct Biol*, 75, 861–877.
10. D.W. Ritchie and S. Grudin (2016). Spherical Polar Fourier Assembly of Protein Complexes with Arbitrary Point Group Symmetry. *Journal of Applied Crystallography*, 49(1), 158–167.
11. Watkins, A.M. et al. (2020) FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds. *Structure*, 28, 963–976.e6.
12. Sato, K. and Kato, Y. (2021) Prediction of RNA secondary structure including pseudoknots for long sequences. *Brief. Bioinform.*, 23, bbab395.

13. Townshend, R. J. L. et al. (2021) Geometric deep learning of RNA structure. *Science*, 373, 1047–1051.
14. Christoffer, C. et al. (2019) Performance and enhancement of the LZerD protein assembly pipeline in CAPRI 38-46. *Proteins*, 88, 948–961.
15. Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., ... & Mainz, D. T. (2006). Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein– ligand complexes. *Journal of medicinal chemistry*, 49(21), 6177-6196.
16. Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A., & Farid, R. (2006). Novel procedure for modeling ligand/receptor induced fit effects. *Journal of medicinal chemistry*, 49(2), 534-553.
17. Huey, R., Morris, G. M., Olson, A. J., & Goodsell, D. S. (2007). A semiempirical free energy force field with charge-based desolvation. *Journal of computational chemistry*, 28(6), 1145-1152.
18. Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., & Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of computational chemistry*, 19(14), 1639-1662.
19. DeLano WL (2002) The PyMOL molecular graphics system. <http://www.pymol.org>
20. Christoffer, C. et al. (2021) LZerD Protein-Protein Docking Webserver Enhanced With de novo Structure Prediction. *Front. Mol. Biosci.*, 8.
21. Boniecki, M. J. et al. (2015) SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res*, 44, e63–e63.

KORP-PL

See: *CONVEX-PL*

LAW, MASS (QA)

Accuracy estimation of individual multimeric protein models using graph neural network and heterogeneous graph neural network

Chenguang Zhao, Tong Liu, Carter Karl Falkenberg, Zheng Wang*

Department of Computer Science, University of Miami

zheng.wang@miami.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; Fragn:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:N; MD:N.*

We present two deep learning methods, named LAW and MASS, for protein single-model estimation of multimeric model accuracy or quality assessment (EMA or QA). LAW was implemented with graph and heterogeneous graph neural networks, and MASS was built as graph neural networks. Both methods were trained to predict an overall fold accuracy (global score), an overall interface accuracy (global interface score), and confidence scores for interface residues (local interface score).

Methods

We used the 22 homomers and 10 heteromers of CASP13-14 for training and four homomers and four heteromers of CASP13-14 for validation. TM score¹, QSbests², and the accuracy of the predicted interface were used as the target values for global score, global interface score, and local interface score, respectively. For each residue of a multimeric model, we generated 83 node features, which can be classified into six categories: (1) one hot coding of amino acid sequence; (2) position-specific scoring matrix (PSSM) created using PSI-BLAST from the multiple sequence alignment (MSA); (3) normalized Rosetta energies; (4) SOV_refine scores³ for sequence-based and model-based secondary structure (SS) and solvent accessibility; (5) MASS⁴ protein statistical potentials, including pseudo-bond angle potential (PAP), accessible surface potential at the atomic level (ASPA), sequence separation-dependent potential (SSDP), contact-dependent potential (CDP), relative solvent accessibility potential (RSAP), and volume-dependent potential (VDP); and (6) sinusoidal positional encoding. Additionally, we also used the mean, median, standard deviation, and variance of ESM⁵ features in predicting the global score of LAW and all scores of MASS. Edges were created for any residue pairs that have a CB-CB distance ≤ 8 Å (CA in the case of Glycine). Ten features were generated for each edge: the distance of residues, the angle between two residues, torsional angles (omega, theta, and phi), and extended contact-dependent potential (CDP) & relative solvent accessibility potential (RSAP) (if the residue pairs are in protein interfaces). We generated the global SOV_refine consistency scores for SS and solvent accessibility as global features.

The global-score predictor of LAW used three graph network⁶ blocks followed by a fully connected layer and a sigmoid function. The global-interface-score predictor of LAW applied three RGATConv⁷ layers on node features, and then the predictor

concatenated the results of scattering mean function on the node features and scattering max and mean functions on the edge features. A fully connected layer and a sigmoid function were applied to the concatenated results. The local-interface-score predictor of LAW used three graph network⁶ blocks followed by a sigmoid function to update edge and node features.

The global-score predictor of MASS used three principal neighborhood aggregation (PNA)⁸ layers followed by a global mean pooling layer. The last layer contains a fully connected layer and a sigmoid function. The networks of the global-interface-score predictor and local-interface-score predictor of MASS have almost the same structure as the global-score predictor of MASS, except that the global-interface-score predictor used two PNA layers, and the local-interface-score predictor used six PNA layers.

1. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality*. Proteins: Structure, Function, and Bioinformatics, 2004. **57**(4): p. 702-710.
2. Bertoni, M., et al., *Modeling protein quaternary structure of homo-and hetero-oligomers beyond binary interactions by homology*. Scientific reports, 2017. **7**(1): p. 1-15.
3. Liu, T., Z.J.S.C.f.B. Wang, and Medicine, *SOV_refine: A further refined definition of segment overlap score and its significance for protein structure similarity*. 2018. **13**(1): p. 1-10.
4. Liu, T. and Z.J.B.b. Wang, *MASS: predict the global qualities of individual protein models using random forests and novel statistical potentials*. 2020. **21**(4): p. 1-10.
5. Rives, A., et al., *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences*. Proceedings of the National Academy of Sciences, 2021. **118**(15): p. e2016239118.
6. Battaglia, P.W., et al., *Relational inductive biases, deep learning, and graph networks*. arXiv preprint arXiv:1806.01261, 2018.
7. Busbridge, D., et al., *Relational graph attention networks*. arXiv preprint arXiv:1904.05811, 2019.
8. Corso, G., et al., *Principal neighbourhood aggregation for graph nets*. Advances in Neural Information Processing Systems, 2020. **33**: p. 13260-13271.

Structure prediction of RNA and RNA complexes using a combination of different modeling methods

C. Nithin¹, M. Lenart^{1,2}, D. Sztuczka^{1,2}, M. Zalewski¹, M. Kurcinski¹, and S. Kmiecik¹

1 - Laboratory of Computational Biology, Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw, Zwirki i Wigury 101, 02-089 Warsaw, Poland, 2 - Faculty of Physics, University of Warsaw, Ludwika Pasteura 5, 02-093 Warsaw, Poland

sekmi@chem.uw.edu.pl

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:Y.N; Fragm:Y.v; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

The three-dimensional structure (3D) prediction of RNA molecules is in infancy. Various methods were developed for 3D prediction. Here, we use a combination of different methods to perform the predictions for the CASP15 targets.

Methods

Firstly, we gathered information about the RNA sequence from various databases and literature. If available, information on the family and multiple sequence alignments was retrieved from the Rfam and the RNACentral databases¹⁻². We used ViennaRNA, RNAstructure, CentroidFold, ContraFold, Ufold, and IPknot to predict the secondary structure of the RNA³⁻⁴. In addition, whenever RNACentral returned a secondary structure for the sequence, we added it to the set of predicted secondary structures. In addition, we enriched the secondary structures based on the information from the literature. The consensus secondary structure from the multiple predictions served as a guide to prepare restraints for the 3D predictions. We used four different methods, SimRNA, DeepFoldRNA, FARFAR, and Vfold, to predict the 3D structure⁵⁻⁸. The simulations were performed with and without restraints to sample the conformational landscape of the RNA molecule. The top-scored models from the different methods were selected for further analysis and checked for convergence to a similar architecture and topology. The selected models were subjected to high-resolution refinements using QRNAS to minimize the errors introduced by the coarse-grain modeling methods⁹.

For the RNA-ligand complexes, the putative pockets were identified based on the 3D coordinates of structures containing the same ligand with an RNA and the conservation of secondary structure elements. Superposition of the pocket regions in the models to the pockets from known structures served as the guidance to adapt the ligand coordinates for preparing the initial poses of the complex. To optimize the position of the ligand in the complex, we performed short runs of energy minimization.

In the case of RNA-protein complexes, the RNA structure was prepared by Frankenstein modeling, where part of the structure was built by homology modeling and combined with the model of the remaining RNA structure. The protein structures were homology-modeled, and the complex was prepared based on the known complexes.

We optimized the 3D structures further with Molecular dynamics simulations (whenever computationally possible) and performed clustering of the trajectories to pick the representative structures.

Results

The RNA secondary structures were predicted using different methods and the consensus secondary structure from these predictions served as a guide to prepare restraints for the 3D predictions. The information from literature, when available was used to enrich the restraints. The top models from 3D structure predictions were checked for convergence to a similar architecture and topology, and was subjected to further refinement. The complex structures were modeled for RNA-protein and RNA-ligand targets. The 3D structures were further optimized using MD and the representative structures were submitted to CASP.

Availability

The various methods used in this pipeline are available publicly.

1. Kalvari,I., Nawrocki,E.P., Ontiveros-Palacios,N., Argasinska,J., Lamkiewicz,K., Marz, M., Griffiths-Jones,S., Toffano-Nioche,C., Gautheret,D., Weinberg,Z. and Rivas,E., Eddy,S.R., Finn,R.D., Bateman,A., & Petrov,A.I. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, 49(D1), D192-D200.
2. The RNAcentral Consortium (2015) RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Research*. **43(D1)**, D123-D129.
3. Churkin,A., Weinbrand,L., & Barash,D. (2015). Free energy minimization to predict RNA secondary structures and computational RNA design. *In RNA Bioinformatics, Humana Press, New York*, 3-16.
4. Hamada,M. (2015). RNA secondary structure prediction from multi-aligned sequences. *In RNA Bioinformatics, Humana Press, New York*, 17-38.
5. Wirecki,T.K., Nithin,C., Mukherjee,S., Bujnicki,J.M., & Boniecki,M.J. (2020). Modeling of three-dimensional RNA structures using SimRNA. *In Protein Structure Prediction, Humana, New York*, 103-125.
6. Pearce,R., Omenn,G.S., & Zhang,Y. (2022). De Novo RNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning. *BioRxiv*.
7. Watkins,A.M., Rangan,R., & Das,R. (2020). FARFAR2: improved de novo rosetta prediction of complex global RNA folds. *Structure*, **28(8)**, 963-976.
8. Li,J., Zhang,S., Zhang,D., & Chen,S.J. (2022). Vfold-Pipeline: a web server for RNA 3D structure prediction from sequences. *Bioinformatics*, **38(16)**, 4042-4043.
9. Stasiewicz,J., Mukherjee,S., Nithin,C., & Bujnicki,J.M. (2019). QRNAS: software tool for refinement of nucleic acid structures. *BMC structural biology*, **19(1)**, 1-11.

Proteins and Protein Complexes prediction powered by Uni-Fold

Xuyang Liu^{*+}, Weijie Chen^{*}, Fan Shen^{*}, Han Wen^{*}, Qinrui Wang^{*}, Maohua Yang^{*}, Yu Guo^{*},
Xinyan Wang^{*}, Yanze Wang^{*}, Ziyao Li^{*}, Junhan Chang^{*}, Zhenfeng Deng^{*}, Shaowei Li^{*},
Dongdong Wang^{*}, Hang Zheng, Xi Wang, Xinyu Li, Guolin Ke^{*+}, Linfeng Zhang^{*+}

DP technology, Beijing, China.

* Equal contribution; + Corresponding authors: liuxy@dp.tech, kegl@dp.tech, zhanglf@dp.tech

In CASP15, we developed several automated workflows integrating both data-driven and physics-driven methods to model all kinds of systems. For proteins and protein-complexes, we reimplemented AlphaFold¹ and AlphaFold-Multimer² in the PyTorch framework, improved the inference efficiency and reproduced training processes³. Then, we fine-tuned these models with various settings and tried to sample the input features for diversity of predictions. Finally, 5 models were selected for **Manifold-E** submission according to structure confidence score. For predictions of other kinds of systems such as RNA or protein-ligand complexes, we also developed automated workflows combing the predicted structures with physics-driven methods. For the human group **Manifold**, we made some manual interventions based on the predicted performance for hard cases.

Methods

Sequence and template searches: MSAs (multiple sequence alignment) were generated by hhblits⁴ against bfd¹ and UniRef30_2021_03 with e-value 0.001. And JackHMMER⁵ was used to search against mgnify⁶, UniRef90_2022_05 and Uniprot_2022_05⁷ with e-value 0.0001. Templates were searched by hmmsearch⁸ against pdb_seqres.txt, downloaded from ftp://ftp.wwpdb.org/pub/pdb/derived_data/pdb_seqres.txt on 2022_05_01⁹.

Protein modeling pipeline: We trained a series of models with different train configs, such as number of sequences, sequence crop size, number of templates, etc. Then, we tried to predict diverse structures by sampling MSA, templates and number of recycles. Next, all predictions were clustered according to structure's RMSD before relaxation. The cluster results were sorted according to the best conformation's plddt (for monomer) or ptm (for multimer) before relaxation, which was done with OpenMM¹⁰ and AMBER99 force field¹¹. Finally, the top-five structures were selected as the results of the server group.

Protein-ligand prediction: We had five steps to obtain the complex conformations. Firstly, five different conformations of the protein were generated from the Uni-fold³. Second, we performed the mixed-solvent molecular dynamics for identifying binding hotspots¹². Third, Fpocket¹³ was used to detect the binding pockets and these pockets were scored by Uni-mol¹⁴. Fourth, the ligands were docked into the top 20 pockets with Uni-IFD and generated many complex conformations. Finally, we used the GBSA scoring method to select the most likely complex conformations for submitting.

RNA prediction: Our RNA folding process was performed as 3D structure folding with secondary structure information as constraints. For secondary structure prediction, we used RNAStructure¹⁵ and ViennaRNA¹⁶, selecting base pairs that overlap in both predictions as constraints and then calling our own parameterized coarse grained force field for conformational search using parallel tempering.

After the conformational search, we clustered the structures, which were later equilibrated using OpenMM¹⁰ and the AMBER14 force field¹¹ under implicit solvent. A batch of conformations was then filtered based on energy ranking. Finally, we used the statistical force field DFIRE-RNA¹⁷ to fine-tune the conformations and selected the top 5 conformations for submission based on the scores given by the statistical force field.

Manual intervention: For some orphan or designed proteins such as T1119 and T1130, we further increased the diversity by changing the e-value cut-off on MSA searching.

For some very long proteins such as T1169. The server predictions were very messy. In these cases, the structures were predicted in segments and then assembled as the final results.

For some proteins with symmetry such as T1115, in addition to making predictions by Uni-Fold symmetry, we manually built structures based on the proteins' asymmetric unit as extra results.

For some additional protein complexes, there were several binding interfaces. In these cases, the literature's information was used to determine the binding position of protein complexes.

Availability

The code is available through github at <https://github.com/dptech-corp/Uni-Fold>

Uni-Fold paper: <https://www.biorxiv.org/content/10.1101/2022.08.04.502811v3>

Uni-Fold symmetry paper: <https://www.biorxiv.org/content/10.1101/2022.08.30.505833v1>.

1. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583-589.
2. Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; Ronneberger, O.; Bodenstein, S.; Zielinski, M.; Bridgland, A.; Potapenko, A.; Cowie, A.; Tunyasuvunakool, K.; Jain, R.; Clancy, E.;

- Kohli, P.; Jumper, J.; Hassabis, D., Protein complex prediction with AlphaFold-Multimer. **2022**, 2021.10.04.463034.
3. Li, Z.; Liu, X.; Chen, W.; Shen, F.; Bi, H.; Ke, G.; Zhang, L., Uni-Fold: An Open-Source Platform for Developing Protein Folding Models beyond AlphaFold. **2022**, 2022.08.04.502811.
 4. Remmert, M.; Biegert, A.; Hauser, A.; Soding, J., HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **2011**, *9* (2), 173-5.
 5. Johnson, L. S.; Eddy, S. R.; Portugaly, E., Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **2010**, *11*, 431.
 6. Mitchell, A. L.; Almeida, A.; Beracochea, M.; Boland, M.; Burgin, J.; Cochrane, G.; Crusoe, M. R.; Kale, V.; Potter, S. C.; Richardson, L. J.; Sakharova, E.; Scheremetjew, M.; Korobeynikov, A.; Shlemov, A.; Kunyavskaya, O.; Lapidus, A.; Finn, R. D., MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res* **2020**, *48* (D1), D570-D578.
 7. UniProt, C., UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **2021**, *49* (D1), D480-D489.
 8. Potter, S. C.; Luciani, A.; Eddy, S. R.; Park, Y.; Lopez, R.; Finn, R. D., HMMER web server: 2018 update. *Nucleic Acids Res* **2018**, *46* (W1), W200-W204.
 9. Burley, S. K.; Berman, H. M.; Kleywegt, G. J.; Markley, J. L.; Nakamura, H.; Velankar, S., Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol Biol* **2017**, *1607*, 627-641.
 10. Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L. P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S., OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **2017**, *13* (7), e1005659.
 11. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65* (3), 712-25.
 12. Graham, S. E.; Leja, N.; Carlson, H. A., MixMD Probeview: Robust Binding Site Prediction from Cosolvent Simulations. *J Chem Inf Model* **2018**, *58* (7), 1426-1433.
 13. Le Guilloux, V.; Schmidtke, P.; Tuffery, P., Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **2009**, *10*, 168.
 14. Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G., Uni-Mol: A Universal 3D Molecular Representation Learning Framework. *ChemRxiv* **2022**.
 15. DiChiacchio, L.; Mathews, D. H., Predicting RNA-RNA Interactions Using RNAstructure. *Methods Mol Biol* **2016**, *1490*, 51-62.
 16. Lorenz, R.; Bernhart, S. H.; Honer Zu Siederdisen, C.; Tafer, H.; Flamm, C.; Stadler, P. F.; Hofacker, I. L., ViennaRNA Package 2.0. *Algorithms Mol Biol* **2011**, *6*, 26.
 17. Zhang, T.; Hu, G.; Yang, Y.; Wang, J.; Zhou, Y., All-Atom Knowledge-Based Potential for RNA Structure Discrimination Based on the Distance-Scaled Finite Ideal-Gas Reference State. *J Comput Biol* **2020**, *27* (6), 856-867.

**An Enhanced protein structure prediction model
with integrated domain knowledge and interaction constraints**

Jiangbin Zheng, Yufei Huang, Ge Wang, Siqi Ma, Bozhen Hu and Stan Z. Li*

AI Division, School of Engineering, Westlake University, Hangzhou, 310030

* Corresponding author: Stan Z. Li (Stan.ZQ.Li@westlake.edu.cn)

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

In the CASP15 experiment, we developed ManiFold, a protein structure prediction system with integrated domain knowledge and constraints on residues. Specifically, ManiFold follows an encoder-decoder architecture. A single sequence-based protein language model is pre-trained and fused with the Evoformer model of AlphaFold2¹. The decoder part is an improved version of the AlphaFold2 structure module, into which a new invariant point attention (IPA) submodule and an iterative sidechain refinement submodule are incorporated, resulting in enhanced invariant/equivariant constraints.

Methods

An enhanced protein sequence representation: We pre-trained a single-sequence protein language model on the UniRef database² in a self-supervised manner. We also migrated the MSA-based Evoformer model of AlphaFold2¹ into the system. The two models are integrated into a final representation of the input protein sequence. This alleviates the need for MSA information.

Iterative optimization with sidechain information: Sidechain information is crucial for structure decoder but is often overlooked in previous protein structure prediction methods. To this end, we trained an equivariant neural network to refine both sidechain and backbone simultaneously. Through an iterative interaction process, sidechain features and the backbone features are updated alternatively one other.

Improved protein structure decoding module: In AlphaFold2, the IPA module predicts the coordinates of amino acids in a Euclidean space through linear projectors, followed by operations invariant to rotation and translation. In ManiFold, we additionally introduced SE(3) Transformer³ to incorporate higher-order geometric constraints into the IPA module while enhancing rotation and translation equivariance.

1. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, 596(7873): 583-589.
2. Suzek B E, Huang H, McGarvey P, et al. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 2007, 23(10): 1282-1288.

3. Fuchs F, Worrall D, Fischer V, et al. SE(3)-transformers: 3D roto-translation equivariant attention networks[J]. Advances in Neural Information Processing Systems, 2020, 33: 1970-1981.

MASS (QA)

See: *LAW*

Manual Prediction of Protein Tertiary and Quaternary Structures and Protein-Ligand Interactions

L.J. McGuffin, N.S. Edmunds, A.G. Genc, S.M.A. Alharbi, R. Adiyaman

School of Biological Sciences, University of Reading, Reading, UK

l.j.mcguffin@reading.ac.uk

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y; Fragm:Y; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

For our manual predictions, we used several components from our latest IntFOLD¹ and ModFOLD² servers, as well as our newly developed quaternary structure modelling and scoring servers (see our IntFOLD7, MultiFOLD and ModFOLDdock abstracts). For our tertiary and quaternary structure predictions (TS format), we made use of the CASP-hosted 3D server models, which we ranked using our ModFOLD9 and/or ModFOLDdock methods and then refined with our new refinement approach, AF2MRefine. Finally, the likely ligand binding sites (LG format) were predicted with our latest version of FunFOLD³.

Methods

Tertiary structure predictions: The server models were ranked according to the ModFOLD9 global quality scores (see our IntFOLD7 abstract for more details on ModFOLD9). The top 10 ModFOLD9_rank models were then selected and submitted as templates to our AF2MRefine pipeline (see our MultiFOLD abstract), which used the LocalColabFold⁴ 1.3.0 method with the “--custom-template-path” option. We used 12 recycles and AMBER relaxation for targets <1000 residues, and we used 3 recycles without AMBER for larger targets.

Quaternary structure predictions: For multimeric targets, the tertiary structure server models for each subunit and the quaternary structure server models were ranked using the ModFOLD9_rank and ModFOLDdockR methods respectively (see our ModFOLDdock abstract). The top 5 tertiary structure models for each subunit and the top 5 quaternary structure models were then selected and used as input templates for the AF2MRefine pipeline, as above.

For each TS format prediction, the final model rankings and the predicted per-residue quality scores (pLDDT*100) from LocalColabFold were added to the B-factor column for each set of atom records. The overall pLDDT and/or pTM scores were then compared with those of the top-ranked models from our MultiFOLD method. If the scores were improved upon, then the refined server models were submitted. If not, then our MultiFOLD models were submitted instead. We also used some manual inspection of multimer models to check if the final models had correctly interacting subunits. For some of the very large complexes (>2500 residues), due to our limited GPU resources, we had to divide sequences up into overlapping fragments for submission to MultiFOLD. The resulting modelled fragments were then manually assembled,

using structural superposition in PyMOL (<https://www.pymol.org>), to form larger, more complete models.

Ligand binding predictions (FunFOLD4): The FunFOLD server³ was designed to find the biologically relevant binding sites and ligands in 3D models by utilising the relevant similar templates from the Protein Data Bank⁵ (PDB) identified by IntFOLD¹. In CASP15, we upgraded our FunFOLD pipeline to predict the potential binding sites and generate poses for the target ligand. Our top manually selected TS models in the human prediction category were used to initially locate binding sites in the individual subunits. For each top model, relevant template lists were generated using a combination of IntFOLD⁷, LocalColabFold⁴ and Foldseek⁶ (a threshold TM-score of 0.4 was applied for the templates found by Foldseek⁶). FunFOLD4 was run to find the biologically relevant binding sites and ligands based on the combined template list and the top selected TS model. If the chemical properties of the ligands predicted by FunFOLD4 matched those of the CASP target ligand, then the ligand was re-docked using Gnina⁷ to fit into the biologically relevant binding site predicted by FunFOLD4. If the chemical properties of the target ligands were not similar to the ligands predicted by FunFOLD4, then whole protein docking was performed using Gnina to find potential binding sites, which were ranked according to the CNN score generated by Gnina. These potential binding sites were also compared with the FunFOLD4 predictions, and then the most common binding sites were selected to re-dock the target ligands. For each target ligand, five poses were generated using Gnina for the top TS model (the corresponding protein receptor in the same frame of reference), which were then submitted in LG format.

Availability

Server methods are available via: <https://www.reading.ac.uk/bioinf/>

Software is free to download via: <https://www.reading.ac.uk/bioinf/downloads/>

1. McGuffin,L.J., Adiyaman,R., Maghrabi,A.H.A., Shuid,A.N., Brackenridge,D.A., Nealon,J.O. & Philomina,L.S. (2019) IntFOLD: an integrated web resource for high performance protein structure and function prediction. *Nucleic Acids Res.* 47, W408-W413. doi: 10.1093/nar/gkz322
2. McGuffin,L.J., Aldowsari, F.M.F., Alharbi,S.M.A., & Adiyaman,R. (2021) ModFOLD8: accurate global and local quality estimates for 3D protein models. *Nucleic Acids Research.* 49(W1), W425–W430. doi: 10.1093/nar/gkab321
3. Roche, D. B., Tetchner, S. J., & McGuffin, L. J. (2011). FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC bioinformatics.* 12(1), 1-20. doi: 10.1093/nar/gkt498
4. Mirdita,M., Schütze,K., Moriwaki,Y., Heo,L., Ovchinnikov,S., & Steinegger,M. (2022) ColabFold: making protein folding accessible to all. *Nature Methods.* 19(6), 679–682. doi: 10.1038/s41592-022-01488-1

5. Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., & Abola, E. E. (1998). Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallographica Section D: Biological Crystallography*, 54(6), 1078-1084. doi: 10.1107/s0907444998009378
6. Van Kempen, M., Kim, S., Tumescheit, C., Mirdita, M., Söding, J., & Steinegger, M. (2022). Foldseek: fast and accurate protein structure search. *bioRxiv*. doi: 10.1101/2022.02.07.479398
7. McNutt, A.T., Francoeur, P., Aggarwal, R. et al. (2021) GNINA 1.0: molecular docking with deep learning. *J Cheminform.* 13, 43. doi: 10.1186/s13321-021-00522-2

Automated Quality Assessment of Protein Quaternary Structure Models using the ModFOLDdock Server

L.J. McGuffin, S.M.A. Alharbi, A.G. Genc, R. Adiyaman and N.S. Edmunds

School of Biological Sciences, University of Reading, Reading, UK

l.j.mcguffin@reading.ac.uk

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

The ModFOLDdock server is our new web resource for the Quality Assessment (QA) of protein quaternary structure models. Three variants of ModFOLDdock were tested at CASP15, which were optimised for the different facets of the quality estimation problem.

Methods

The ModFOLDdock server uses a hybrid consensus approach for producing both global and local (interface residue) quality scores for predicted quaternary structures. The ModFOLDdock variants use various combinations of scores (detailed in the sections below), which are calculated using the output from 7 individual scoring methods: DockQJury, QSscoreJury, QSscoreOfficialJury, IDDTOfficialJury, voronota-js-voromqa, CDA, and ModFOLDIA.

For the DockQJury, QSscoreJury, QSscoreOfficialJury and IDDTOfficialJury scoring methods, pairwise comparisons were made between each quaternary structure model and every other model and then the mean QS¹, IDDT² or DockQ³ scores were calculated. The difference between the QSscoreJury and QSscoreOfficialJury approaches was that in the former, in-house code was used to calculate the fraction of correctly modelled interface contacts in the complex normalised by the max of either the observed or predicted contacts, while in the latter, the OpenStructure⁴ package was used to obtain QS scores (using the “ost compare-structures” action). The voronota-js-voromqa⁵ method was used off-the-shelf with the “--inter-chain” and “--output-dark-scores” options.

The CDA score was based on the original Contact Distance Agreement (CDA) score^{6,7}, which relates to the agreement between the residue contacts predicted from the sequence and the measured Euclidean distance (in Å) between residues in the model. In this case, we used the contact prediction profiles that resulted from the generation of LocalColabFold⁸ version 1.0.0 multimer models.

The ModFOLDIA method was also used to carry out structure-based comparisons of alternative oligomer models and it produced both global and local/per-residue interface scores. The first stage of the ModFOLDIA method was to identify the interface residues in the model to be scored (defined as $\leq 8\text{Å}$ between C β atoms, or C α for GLY) and then obtain the minimum contact distance (D_{min}) for each contacting residue. The second stage was to locate the equivalent

residues in all other models and then obtain the mean minimum distances of those residues in all other models ($MeanD_{min}$). The final Interface Accuracy (IA) score for each of the interface residues in the model was the absolute difference in the S_i from the mean S_i : $IA = 1 - |S_i - MeanS_i|$, where $S_i = 1 / (1 + (D_{min}/20)^2)$ and $MeanS_i = 1 / (1 + (MeanD_{min}/20)^2)$. The global ModFOLDIA score for a model was then taken as the total interface score (sum of residue scores) normalised by the maximum of either the number of residues in the interface or the mean number of interface residues across all models for the same target.

ModFOLDdock: This variant produced predicted scores optimised for positive linear correlations with the observed scores, i.e., the predicted overall quality scores correlated well with the observed overall quality scores, according to the assessors' formulae for CASP14 multimer models⁹. The overall fold accuracy (column 2 in the QA file) was calculated from the mean of the DockQJury and the IDDTOfficialJury scores. The overall interface accuracy (column 3) was calculated from the mean of the DockQJury and the QScoreOfficialJury scores. Additionally, confidence scores (IA scores) for all of the interface residues in each model were calculated using the ModFOLDIA method (as shown above).

ModFOLDdockR: This variant produced predicted scores optimised for ranking, i.e. the top-ranked models (top 1) should have higher observed overall accuracy, but the relationship between predicted and observed scores may not be linear. The overall fold accuracy (column 2) was calculated from the mean of the QScoreJury, IDDTOfficialJury and voronota-js-voromqa scores. The overall interface accuracy (column 3) was calculated from the mean of the DockQJury, QScoreOfficialJury and voronota-js-voromqa scores. The confidence scores for all of the interface residues in each model were calculated from the mean of the IA score and the per-residue score from voronota-js-voromqa.

For the very large complexes (>1500 residues total length of all subunits), due to CPU and RAM limitations, we could not carry out all-against-all pairwise structural comparisons using all approaches within the 48h time window. So in these cases, for both the ModFOLDdock and ModFOLDdockR methods, we initially scored all models using voronota-js-voromqa and then selected just the top 40 models to act as the reference set for all model comparisons.

ModFOLDdockS: This variant used a quasi-single model approach to score models. Sets of reference multimer models were firstly generated from the input sequences using our MultiFOLD method (see our MultiFOLD abstract for details) and then each model was compared individually against the reference set using the 7 individual scoring methods described above. The overall fold accuracy (column 2) was calculated from the mean of the DockQJury and the IDDTOfficialJury scores. The overall interface accuracy (column 3) was calculated from the mean of the DockQJury and the QScoreOfficialJury scores. The confidence scores for all of

the interface residues in each model were calculated from the mean of the IA, the voronota-js-voromqa and the CDA scores.

Availability

The ModFOLDdock server is available at:

https://www.reading.ac.uk/bioinf/ModFOLDdock/ModFOLDdock_form.html

1. Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., Schwede, T. (2017) Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci Rep.* **7**, 10480. doi: 10.1038/s41598-017-09654-8.
2. Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013) IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics.* **29**(21), 2722–2728. doi: 10.1093/bioinformatics/btt473
3. Basu, S., Wallner, B. (2016) DockQ: A Quality Measure for Protein-Protein Docking Models. *PLoS One.* **11**, e0161879. doi: 10.1371/journal.pone.0161879.
4. Biasini, M., Schmidt, T., Bienert, S., Mariani, V., Studer, G., Haas, J., Johnner, N., Schenk, A.D., Philippsen, A., & Schwede, T. (2013) OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography.* **69**(Pt 5), 701–709. doi: 10.1107/S0907444913007051
5. Olechnovič, K., & Venclovas, C. (2014) Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. *Journal of Computational Chemistry.* **35**(8), 672–681. doi: 10.1002/jcc.23538
6. Maghrabi, A.H.A. & McGuffin, L.J. (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D models of proteins. *Nucleic Acids Res.*, **45**, W416–W421. doi: 10.1093/nar/gkx332.
7. McGuffin, L.J., Aldowsari, F.M.F., Alharbi, S.M.A., & Adiyaman, R. (2021) ModFOLD8: accurate global and local quality estimates for 3D protein models. *Nucleic Acids Research.* **49**(W1), W425–W430. doi: 10.1093/nar/gkab321
8. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022) ColabFold: making protein folding accessible to all. *Nature Methods.* **19**(6), 679–682. doi: 10.1038/s41592-022-01488-1
9. Ozden, B., Kryshtafovych, A., & Karaca, E. (2021) Assessment of the CASP14 assembly predictions. *Proteins.* **89**(12), 1787–1799. doi: 10.1002/prot.26199

Protein Multimer QA with AlphaFold-Multimer and Machine Learning

Wenbo Wang and Yi Shang

Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, USA

wwr34@mail.missouri.edu

Key: Auto:Y; CASP_serv:N; Templ:Y; MSA:Y; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:N

MUFold and MUFold2 are two new protein complex QA methods designed on top of AlphaFold-Multimer. MUFold uses a single-stage machine learning method based on AlphaFold-Multimer results, while MUFold2 features a 2-stage machine learning method. In MUFold2, a model is first trained to make initial predictions using the output of AlphaFold-Multimer. Then, a second pre-trained predictive model is used to generate more accurate predictions.

Methods

TS/QS: The monomer and assembly 3D structure prediction are done using the publicly available AlphaFold-Multimer1 with postprocessing methods we developed for some difficult targets. The output is later used as input to our QA algorithms, which will be the focus of this abstract.

QA: The input of the algorithm is the target protein sequence S and a predicted protein model M .

Step 1. Run AlphaFold-Multimer to get 25 unrelaxed predictions U_iP_j , where $i=1..5$, $j=0..4$, (unrelaxed_model_[1-5]_multimer_v2_pred_[0-4].pdb) and 25 relaxed ones R_iP_j , where $i=1..5$, $j=0..4$, (relaxed_model_[1-5]_multimer_v2_pred_[0-4].pdb). During CASP15, we simply used our results from our 3D structure prediction.

Step 2. Calculate TMscore between model M and $U[1-5]P[0-4]$ using MM-align2 to get $TMS_U[1-5]P[0-4]$. Do the same with $R[1-5]P[0-4]$ to get $TMS_R[1-5]P[0-4]$.

Step 3. Divide unrelaxed and relaxed AlphaFold-Multimer predictions respectively into five groups in a way that each group has the same P index, i.e., $U[1-5]P_0$, $U[1-5]P_1$, $U[1-5]P_2$, $U[1-5]P_3$, $U[1-5]P_4$, and $R[1-5]P_0$, $R[1-5]P_1$, $R[1-5]P_2$, $R[1-5]P_3$, $R[1-5]P_4$.

Step 4. For each group in Step 3, use a machine learning method based on the models in this group and pre-generated features to train a predictive model. Use this machine learning model to make an initial QA prediction on protein model M .

Step 5. Feed the QA predictions on protein model M from different groups and additional protein features like sequence information to a pretrained machine learning model to make the final QA prediction on M. The difference between MUFold and MUFold2 is that MUFold does not consider the output from step 4, whereas MUFold2 considers the outputs from both step 2 and 4.

1. Evans, R., et al. (2021). Protein complex prediction with AlphaFold-Multimer. *BioRxiv* 2021.10.04.463034; doi: <https://doi.org/10.1101/2021.10.04.463034>
2. Mukherjee, S., & Zhang, Y. (2009). MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic acids research*, 37(11), e83.

MULTICOM_egnn, MULTICOM_deep, MULTICOM_qa (QA)

Multimer Model Quality Assessment Using Gated-Graph Transformer, Steerable Equivariant Graph Neural Networks, and Pairwise Model Similarity

Xiao Chen[†], Alex Morehead[†], Raj S. Roy[†], Zhiye Guo, Jian Liu, Nabin Giri, Tianqi Wu, Chen Chen, Jianlin Cheng^{*}

University of Missouri, Columbia, MO 65211, USA

[†] Joint first author, ^{*}Corresponding author: chengji@missouri.edu

Key: Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:Y; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N;

In the CASP15 experiment, we deployed three different quality assessment methods for estimating the accuracy of predicted multimer models. MULTICOM_egnn is a *single-model* deep learning method using the gated-graph transformer architecture to predict the global quality of a multimer model. MULTICOM_deep is a *single-model* deep learning method that uses an equivariant graph neural network to predict both the per-residue local distance difference test (IDDT) scores and the global quality score of a multimer model. MULTICOM_qa is a *multi-model* method that combines the pairwise similarity between multimer models and the inter-chain contact probabilities predicted by deep learning methods to estimate their global quality.

Methods

MULTICOM_egnn

MULTICOM_egnn is built on top of our in-house *single-model* quality assessment method – DProQ [1] that takes a multimer model as input and represents it as a 3D graph to predict the DockQ [2] score of the model. The node features include one-hot amino acid encoding, secondary structure type, relative accessible surface area, phi angle, psi angle, and graph Laplacian positional encoding of residues. The edge features include Ca-Ca distance, Cb-Cb distance, N-O distance, inter-chain contact encoding, permutation-invariant chain encoding, and edgewise positional encoding. It uses a gated-graph transformer architecture to update node and edge embeddings during graph message passing to predict the DockQ score of the multimer model.

MULTICOM_deep

MULTICOM_deep is based on our in-house *single-model* quality assessment method – DeepRefine [3]. It represents a multimer model as a 3D graph, where the nodes of the graph correspond to residues in the model and the edges are defined according to each residue’s 20 nearest neighbors in 3D space. For this 3D graph, DeepRefine generates geometric features for each of its nodes and edges such as cosinusoidal and sinusoidal encodings of the residue’s

backbone dihedral angles and distance, direction, and orientation between residues. With such 3D graphs instantiated with these features, it then applies a steerable equivariant graph neural network [4] to predict the input structure's quality (i.e., nativeness). Specifically, DeepRefine predicts the IDDT score corresponding to the nativeness of the 3D position of each node (residue). The average of the per-residue IDDT scores is used as the predicted global quality of the multimer model.

MULTICOM_qa

MULTICOM_qa is a *multi-model* quality assessment method based on our in-house method MultimerEva. It takes a pool of multimer models of a target as input to predict their global quality score. The multimer models are compared with each other using MMalign [5]. The average TM-score between a model and all the other models in the pool is calculated as one measure of the quality of the model (denoted as *avg_pairwise_score*). Moreover, for each multimer model, the probabilities of interchain residue-residue contacts in the model are predicted by our deep learning tools for predicting interchain residue-residue contacts and/or distances [6-8], which are averaged as another measure of the global quality of the model (denoted as *avg_interface_score*). Finally, the weighted sum of *avg_pairwise_score* and *avg_interface_score* is used as the final predicted quality score of each multimer model in the pool.

Availability

DProQ (MULTICOM_egnn) is available at: <https://github.com/BioinfoMachineLearning/DProQ>; DeepRefine (MULTICOM_deep) is available at: <https://github.com/BioinfoMachineLearning/DeepRefine>.

1. Chen, X., Morehead, A., Liu, J., & Cheng, J. (2022). DProQ: A Gated-Graph Transformer for Protein Complex Structure Assessment. *bioRxiv*.
2. Basu, S., & Wallner, B. (2016). DockQ: a quality measure for protein-protein docking models. *PLoS one*, 11(8), e0161879.
3. Morehead, A., Chen, X., Wu, T., Liu, J., & Cheng, J. (2022). EGR: Equivariant Graph Refinement and Assessment of 3D Protein Complex Structures. *arXiv preprint arXiv:2205.10390*
4. Brandstetter, Johannes, et al. "Geometric and physical quantities improve E(3) equivariant message passing." arXiv preprint arXiv:2110.02905 (2021).
5. Mukherjee, S., & Zhang, Y. (2009). MM-align: A quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Research*, 37(11), e83. <https://doi.org/10.1093/nar/gkp318>
6. Quadir, F., Roy, R. S., Soltanikazemi, E., & Cheng, J. (2021). DeepComplex: A Web Server of Predicting Protein Complex Structures by Deep Learning Inter-chain Contact Prediction and Distance-Based Modelling. *Frontiers in Molecular Biosciences*, 8. <https://www.frontiersin.org/article/10.3389/fmolb.2021.716973>
7. Roy, R. S., Quadir, F., Soltanikazemi, E., & Cheng, J. (2022). A deep dilated convolutional residual network for predicting interchain contacts of protein homodimers. *Bioinformatics*, btac063. <https://doi.org/10.1093/bioinformatics/btac063>

8. Guo, Z., Liu, J., Skolnick, J., & Cheng, J. (2022). *Prediction of inter-chain distance maps of protein complexes with 2D attention-based deep neural networks* (p. 2022.06.19.496734). bioRxiv. <https://doi.org/10.1101/2022.06.19.496734>

Improving Assembly Structure Prediction by Sensitive Alignment Sampling, Template Identification, Model Ranking, and Iterative Refinement

Jian Liu, Zhiye Guo, Tianqi Wu, Raj S. Roy, Farhan Quadir, Chen Chen, Jianlin Cheng*

University of Missouri, Columbia, MO 65211, USA

*Corresponding author: chengji@missouri.edu

Key: Auto:N; CASP_serv:N; Templ:Y; MSA:Y; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N;

For the CASP15 experiment, we developed a protein assembly prediction system on top of AlphaFold2-Multimer¹'s superior capability of generating structural models. While using the deep learning models of AlphaFold2-Multimer as the underlying engine to generate structural models from alignments and templates, our system focuses on improving the input fed to AlphaFold2-Multimer and evaluating and refining the outputs generated by AlphaFold2-Multimer to enhance multimer structure prediction. Specifically, we developed several different algorithms to sample multiple sequence alignments, identify structural templates, rank structural models generated by AlphaFold2-Multimer, and refine the structural models via a novel structure-alignment based refinement method. Particularly, we leverage monomer structure prediction and sensitive structure alignment to generate deep multiple sequence alignments and identify remote templates to improve and refine assembly (multimer) prediction. These different methods were tested as two server predictors (MULTICOM_deep and MULTICOM_qa) and two human predictors (MULTICOM and MULTICOM_human) in the assembly structure prediction in CASP15.

Methods

1. The multimer (assembly) structure prediction pipeline

The protein structure prediction pipeline of our multimer prediction system consists of six sequential steps: **(1) monomer (tertiary) structure prediction for each unit of a multimer**, **(2) multiple sequence alignment sampling**, **(3) template identification**, **(4) model generation**, **(5) model ranking**, and **(6) model refinement**. Except that the model generation is handled by the deep learning models of AlphaFold2 and AlphaFold2-Multimer, all the other steps are largely based on our customized algorithms, which are described in detail below.

Monomer (tertiary) structure prediction for each unit of multimer. Our in-house tertiary structure prediction system of optimizing the multiple sequence alignments and templates fed to AlphaFold2² and ranking and refining structural models generated by AlphaFold2 is used to generate multiple sequence alignments (MSAs) and predict tertiary structures for each unit of a multimer target (see our server tertiary structure prediction abstract entitled “*Improving Tertiary Structure Prediction by Alignment Sampling, Template Identification, Model Ranking, Iterative Refinement, and Protein Interaction-Aware Modeling*” for technical details).

Multimer MSA sampling. The MSAs of the subunits of the multimer target are concatenated using potential protein-protein interactions extracted from multiple sources such as the species information, UniProt accession IDs, the protein-protein interactions in the STRING³ database, and the protein complexes in the Protein Data Bank⁴ (PDB), resulting in a series of MSAs for the multimer. Moreover, the predicted tertiary structures of the units of the multimer target are also searched against an inhouse complex template database built from PDB and against the single-chain models in the AlphaFoldDB (the version released before March 2022) by a structure alignment tool - FoldSeek⁵ to identify similar structural units in a template complex or similar non-overlapped domains of an AlphaFoldDB model, whose sequences are concatenated to generate MSAs for the multimer. *This structure alignment-based method can generate deeper MSAs for some hard targets than traditional sequence alignment methods, leading to better structure prediction.*

Multimer template identification. The sequence of the multimer is searched against PDB70 and an inhouse complex template database built from PDB by HHsearch⁶ to identify the structural templates. The templates for each subunit are concatenated together if they share the same PDB code. Moreover, the predicted tertiary structural model of each unit of the multimer is searched against the inhouse structure template database by FoldSeek to identify more templates, which are concatenated as multimer templates. *This structure alignment-based method can identify some remote structural templates for multimers that cannot be found by traditional sequence alignment methods.*

Multimer structural model generation. Each combination of the concatenated MSAs and templates is fed for the customized AlphaFold2-Multimer to generate multimer structures. Usually, more than 100 models are generated for a target.

Multimer model ranking. MultimerEva, an inhouse tool of evaluating the quality of multimer models based on the average pairwise structural similarity score between models of a target, is used to rank the generated models. The pairwise structural similarity score is calculated by MM-align⁷. The confidence score generated by AlphaFold-Multimer is also used to rank the models. Finally, the average of the two is applied to rank the models as well.

Multimer Iterative model refinement. We developed a novel iterative model refinement method based on structure search and alignments. An initial target structural model is used as input for FoldSeek⁵ to search for similar structures in the template databases curated from the PDB and the AlphaFoldDB (the version released before March 2022). The output of the FoldSeek includes the e-value of the similar structural hits as well as the structural alignments between the target model and the hits, which are converted into the sequence alignments between them. The MSAs and templates of the subunits generated from FoldSeek search are concatenated if they are from the same PDB complex structure or the non-overlapped regions of the same single-chain AlphaFoldDB model. The sequence alignments are added into the original MSA to generate a *deeper* MSA. The new MSA and the top-ranked structural hits found by FoldSeek are used as MSA and template inputs for the customized AlphaFold2-Multimer to generate the refined models. If the highest confidence score of the newly refined models is higher than that of the input model, the refinement process is repeated with the refined model as input until the number of the refinement iterations reaches 5. *This iterative, structure alignment-based refinement method can improve the quality of the final prediction for some hard targets.*

2. Implementation of the CASP15 assembly (multimer) structure predictors

Both the monomer and multimer prediction methods in our prediction system above were executed to generate the models for multimer targets. Our two CASP15 multimer server predictors (MULTICOM_qa and MULTICOM_deep) used the AlphaFold2-Multimer confidence score and the average of the confidence score and MultimerEva score to rank multimer models, respectively. Due to the three-day time constraint, the refinement was only applied to some smaller multimer targets in the server prediction.

The two human multimer predictors (MULTICOM and MULTICOM_human) generated more multimer models from more diverse MSAs thanks to a much longer timeline. The FoldSeek-based iterative model refinement was applied to most targets. MULTICOM_human used the average of the confidence score and the MultimerEva score to rank and select models for final submission, while MULTICOM applied the MultimerEva score to rank models. The ranking may be manually adjusted according to human inspection.

1. Evans R, O'Neill M, Pritzel A, et al. Protein complex prediction with AlphaFold-Multimer. *BioRxiv* 2021.
2. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;596:583-589.
3. Mering Cv, Huynen M, Jaeggi D, et al. STRING: a database of predicted functional associations between proteins. *Nucleic acids research* 2003;31:258-261.
4. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic acids research* 2000;28:235-242.
5. van Kempen M, Kim S, Tumescheit C, et al. Foldseek: fast and accurate protein structure search. *bioRxiv* 2022.
6. Steinegger M, Meier M, Mirdita M, et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics* 2019;20:1-15.
7. Mukherjee, S., & Zhang, Y. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic acids research* 2009; 37(11), e83-e83

MULTICOM, MULTICOM_human, MULTICOM_egnn, MULTICOM_refine, MULTICOM_deep, MULTICOM_qa (TS)

Improving Tertiary Structure Prediction by Alignment Sampling, Template Identification, Model Ranking, Iterative Refinement, and Protein Interaction-Aware Modeling

Jian Liu, Zhiye Guo, Tianqi Wu, Raj S. Roy, Farhan Quadir, Chen Chen, Jianlin Cheng*

University of Missouri, Columbia, MO 65211, USA

*Corresponding author: chengji@missouri.edu

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA:Y; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N;*

Since CASP14, AlphaFold2¹ and its extension AlphaFold2-Multimer² has become the standard method for predicting protein structures. For the CASP15 experiment, we developed a protein tertiary structure prediction system on top of AlphaFold2's superior capability of generating structural models. While using the deep learning models of AlphaFold2 as the underlying engine to generate structural models from alignments and templates, our system focuses on improving the input fed to AlphaFold2 and evaluating and refining the outputs generated by AlphaFold2 to enhance tertiary structure prediction. Specifically, we developed several different algorithms to sample multiple sequence alignments, identify structural templates, rank structural models generated by AlphaFold2, and refine the structural models via a novel structure-alignment based refinement method. Moreover, for monomer targets that are the units of multimer targets, we integrate monomer and multimer prediction to account for the structural change on tertiary structures induced by protein-protein interaction. These different methods were tested as four server predictors (MULTICOM_egnn, MULTICOM_refine, MULTICOM_deep, and MULTICOM_qa) and two human predictors (MULTICOM and MULTICOM_human) in the CASP15 tertiary structure prediction.

Methods

The general tertiary structure prediction pipeline of the system consists of five sequential steps: **(1) multiple sequence alignment sampling**, **(2) template identification**, **(3) model generation**, **(4) model ranking**, and **(5) model refinement**. Except for Step 3 (model generation) that is handled by the deep learning models of AlphaFold2, all the other steps are largely based on our customized algorithms, which are described in detail below.

The tertiary (monomer) structure prediction pipeline

Monomer multiple sequence alignment (MSA) sampling. When the system receives a monomer target, a monomer alignment generation pipeline is applied to generate various kinds of MSAs by using HHblits^{3,4}, JackHMMER⁵ and MMseqs2⁶ to search the sequence databases, including UniRef30, UniRef90⁷, BFD^{8,9}, MGnify clusters¹⁰, UniProt⁷, and the ColabFold DB⁶. Moreover, a DeepMSA-like alignment tool is executed in the background to search the huge Integrated Microbial Genomes (IMG) database to generate alternative alignments for hard targets having few homologous sequences.

Template identification. In addition to using the templates identified by the default AlphaFold2, the MSA generated from UniRef90 is used to search our inhouse template database curated from Protein Data Bank¹¹ (PDB) to identify alternative templates.

Monomer structural model generation. A customized version of AlphaFold2 is used to generate models using the MSAs and templates generated from the previous steps. Each combination of a MSA and a set of templates is used to generate five models. Multiple combinations of MSAs and templates lead to about 50 models generated for each target. If the depth of the MSA generated by the default AlphaFold2 is less than 200, the MSA generated from the IMG database is also used to generate more models.

Model ranking. The APOLLO¹² model ranking score (the average pairwise structural similarity between models) and the global pLDDT score generated by AlphaFold2 are used to rank the structural models. The average of the two is also used to rank them. Moreover, a deep learning method - DeepRank¹³ is used in model ranking. EnQA¹⁴ – a 3D-equivariant deep learning model is applied to rank the structural models when appropriate.

Iterative model refinement. We developed a novel iterative model refinement method based on FoldSeek¹⁵ structure search and alignments. An initial target structural model is used as input for FoldSeek to search for similar structures in the template database curated from the PDB and the AlphaFoldDB (the version released before March 2022). The output of the FoldSeek includes the e-value of the similar structural hits as well as the structural alignments between the target model and the hits, which are converted into the sequence alignments between them. The sequence alignments are added into the original MSA to generate a deeper MSA. The redundant sequences in the new MSA are removed by HHfliter⁴ according to 90% sequence identity threshold. The filtered MSA and the top-ranked structural hits found by FoldSeek are used as MSA and template inputs for the customized AlphaFold2 to generate the refined models. If the highest pLDDT score of the newly refined models is higher than that of the input model, the refinement process is repeated with the refined model as input until the number of the refinement iterations reaches 5.

Implementation of CASP15 predictors & integration of monomer and multimer prediction to account for structural changes induced by protein-protein interaction

If a monomer target was a single-chain target, only the monomer prediction method above was used to generate structural models. MULTICOM_egnn server predictor used the average of the pLDDT score and the APOLLO pairwise similarity score to rank models; MULTICOM_refine refined the top five models selected by the average ranking and selected the final five models with the highest pLDDT scores from the 10 unrefined and refined models (5 unrefined + 5 refined); MULTICOM_deep used pLDDT score to rank and select models; and MULTICOM_qa refined the top 5 models generated by the default AlphaFold2.

The two human predictors (MULTICOM and MULTICOM_human) selected monomer models from a larger model pool generated in a longer period and from more diverse MSAs than the server predictors. The refined models were also added into the pool for ranking. DeepRank¹³ was used to rank models for MULTICOM_human, while the average ranking of the pairwise

similarity score and pLDDT score generated from AlphaFold2 was used to rank models for MULTICOM. The ranking may be manually adjusted according to human inspection.

If a monomer target was a chain of a multimer target, the monomer models extracted from the multimer models predicted for the multimer target were added into the model pool for ranking if available (see our assembly structure prediction abstract entitled “*Improving Assembly Structure Prediction by Sensitive Alignment Sampling, Template Identification, Model Ranking, and Iterative Refinement*” for more details). Because the structure of a chain may change when interacting with other chains in a multimer, the monomer models extracted from the multimer models accommodating protein-protein interaction were preferred to the monomer models generated by the monomer structure prediction without considering the interaction between chains. Therefore, generally, the top ranked models extracted from multimer models were used as top 3-4 models submitted to CASP15, while the remaining models submitted could be the top ranked monomer models generated by the monomer modeling. *This protein interaction-aware prediction of tertiary structure integrating monomer and multimer prediction appeared to improve prediction accuracy for many monomer targets that are a unit of multimer targets in CASP15.*

1. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;596:583-589.
2. Evans R, O'Neill M, Pritzel A, et al. Protein complex prediction with AlphaFold-Multimer. *BioRxiv* 2021.
3. Remmert M, Biegert A, Hauser A, et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* 2012;9:173-175.
4. Steinegger M, Meier M, Mirdita M, et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics* 2019;20:1-15.
5. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics* 2010;11:1-8.
6. Mirdita M, Schütze K, Moriwaki Y, et al. ColabFold: making protein folding accessible to all. *Nature Methods* 2022:1-4.
7. Consortium U. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* 2019;47:D506-D515.
8. Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manifold. *Nature methods* 2019;16:603-606.
9. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nature communications* 2018;9:1-8.
10. Mitchell AL, Almeida A, Beracochea M, et al. MGnify: the microbiome analysis resource in 2020. *Nucleic acids research* 2020;48:D570-D578.
11. Sussman JL, Lin D, Jiang J, et al. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallographica Section D: Biological Crystallography* 1998;54:1078-1084.
12. Wang Z, Eickholt J, Cheng J. APOLLO: a quality assessment service for single and multiple protein models. *Bioinformatics* 2011;27:1715-1716.

13. Hou J, Wu T, Cao R, et al. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins: Structure, Function, and Bioinformatics* 2019;87:1165-1178.
14. Chen C, Chen X, Morehead A, et al. 3D-equivariant graph neural networks for protein model quality assessment. *bioRxiv* 2022.
15. van Kempen M, Kim S, Tumescheit C, et al. Foldseek: fast and accurate protein structure search. *bioRxiv* 2022.

Template-based Modeling for Accurate Prediction of Ligand-Protein Complex Structures in CASP15

Nabin Giri, Ashwin Dhakal, Jian Liu, Jianlin Cheng*

University of Missouri, Columbia, MO 65211, USA

*Corresponding author: chengji@missouri.edu

Key: Auto:N; CASP_serv:N; Templ:Y; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N

We developed a template-based modeling pipeline for accurate ligand-protein interaction prediction (TULIP) for the CASP15 experiment, which participated in the protein-ligand complex structure prediction as a server predictor (MULTICOM_qa) and a human predictor (MULTICOM). TULIP primarily made use of a novel template-based approach to predict the structure of protein-ligand complexes, which was supplemented by an optional deep learning tool for fitting ligands into protein structures.

Methods

Input Data Preparation. Given a target ligand's SMILES¹ string, TULIP generated its 3D structure using the cheminformatics tools (RDKit² and Open Babel³), which could be the global minimum energy conformer⁴. The structure of the receptor protein was predicted by our CASP15 3D protein structure predictors (i.e., MULTICOM_qa for the server prediction and MULTICOM for the human prediction). The predicted protein structures were searched against a database of known experimental protein structures curated from the Protein Data Bank (PDB) using Foldseek⁵ to identify similar protein structure templates containing ligands for the protein-ligand complex structure prediction⁶

Template Based Protein-Ligand Prediction. The target ligand's 3D initial conformer structure, predicted receptor protein structure, and identified template structures were used as inputs for TULIP to extract the ligands from template structures. Like the approach employed in DeepProLigand⁷, TULIP first aligned the template structure containing ligands into the same geometrical space of the predicted receptor structure using UCSF Chimera's⁸ *matchmaker* in the non-interactive mode and saved the superimposed template structures and their ligands relative to the predicted receptor structure in a PDB file. This output PDB file was then processed by PyRosetta's⁹ *is_ligand* function, which identified the template ligands by checking each residue into the chemical component dictionary. The extracted unique ligands from each template and the target ligand were converted into molecular fingerprints to compute the molecular similarity between them. Morgan Fingerprint¹⁰ was used to convert the ligand molecules into fingerprints, and Dice and Tanimoto similarity metrics were used to measure the similarity between the

fingerprints. The similarity was measured as scores between 0 and 1, where scores closer to 0 indicated no structural similarity and scores closer to 1 high structural similarity between the two molecules. This step provided the initial binding location of the target ligand with respect to the receptor protein structure.

Furthermore, to adjust a target ligand's binding pose and orientation by rotation and translation, TULIP used LS-align¹¹ to align the target ligand with the template ligands of higher similarity by both flexible and rigid body alignments. Between the flexible and rigid body alignment's outputs, it selected the alignment that had lowest $RMSD_{LS}$ between the template and target ligands to obtain the predicted coordinates of the target ligand. The target ligand's coordinates were then submitted to CASP15 in the MDL file format.

Optional Deep Learning Based Protein-Ligand Prediction. For several targets released in the early stage of the CASP15 experiment before TULIP was fully developed, we also applied a deep learning tool, EquiBind¹² to make protein-ligand predictions. EquiBind used graph matching networks¹³ and E (3)-equivariant graph neural networks¹⁴ (E(3)-GNN) to perform a direct prediction of protein-ligand complex structure from the input structure of a target ligand and a predicted receptor structure.

1. Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J Chem Inf Comput Sci* **28**, 31–36 (1988).
2. Landrum, G. *et al.* rdkit/rdkit: 2022_03_5 (Q1 2022) Release. (2022) doi:10.5281/ZENODO.6961488.
3. Hutchison Chris Morley Craig James Chris Swain Hans De Winter Tim Vandermeersch Noel M O, G. R. *Open Babel Documentation*. (2021).
4. Yoshikawa, N. & Hutchison, G. R. Fast, efficient fragment-based coordinate generation for Open Babel. *J Cheminform* **11**, (2019).
5. van Kempen, M. *et al.* Foldseek: fast and accurate protein structure search. doi:10.1101/2022.02.07.479398.
6. Dhakal, A., McKay, C., Tanner, J. J. & Cheng, J. Artificial intelligence in the prediction of protein-ligand interactions: recent advances and future directions. *Briefings in Bioinformatics* vol. 23 Preprint at <https://doi.org/10.1093/bib/bbab476> (2022).
7. Giri, N. & Cheng, J. A Deep Learning Bioinformatics Approach to Modeling Protein-Ligand Interaction with cryo-EM Data in 2021 Ligand Model Challenge. *bioRxiv* (2022).
8. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612 (2004).
9. Chaudhury, S., Lyсков, S. & Gray, J. J. PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* vol. 26 689–691 Preprint at <https://doi.org/10.1093/bioinformatics/btq007> (2010).
10. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J Chem Inf Model* **50**, 742–754 (2010).
11. Hu, J., Liu, Z., Yu, D. J. & Zhang, Y. LS-align: An atom-level, flexible ligand structural alignment algorithm for high-throughput virtual screening. in *Bioinformatics* vol. 34 2209–2218 (Oxford University Press, 2018).

12. Stärk, H., Ganea, O.-E., Pattanaik, L., Barzilay, R. & Jaakkola, T. EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. (2022).
13. Li, Y., Gu, C., Dullien, T., Vinyals, O. & Kohli, P. *Graph Matching Networks for Learning the Similarity of Graph Structured Objects*.
14. Satorras, V. G., Hoogeboom, E. & Welling, M. E(n) Equivariant Graph Neural Networks. (2021).

MultiFOLD

Automated Prediction, Quality Assessment and Refinement of Tertiary and Quaternary Structure Models using the MultiFOLD Server

L.J. McGuffin, A.G. Genc, S.M.A. Alharbi, R. Adiyaman, B.R. Salehe and N.S. Edmunds

School of Biological Sciences, University of Reading, Reading, UK

l.j.mcguffin@reading.ac.uk

Key: *Auto:Y; CASP_serv:Y; Templ:Y; MSA:Y; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

The MultiFOLD server is our new integrated pipeline for producing tertiary and quaternary structure models of proteins via modelling, quality scoring and refinement.

Methods

The MultiFOLD protocol has 3 main stages: modelling, scoring, and refinement. In the first stage, 3D models of tertiary and quaternary structures were built using two different versions of LocalColabFold¹ (<https://github.com/YoshitakaMo/localcolabfold>). Firstly, LocalColabFold version 1.0.0, which is based on the ColabFold/AlphaFold2_advanced notebook integrating the AlphaFold2² weights (AlphaFold2-ptm) and modified to additionally produce models for multimers. Secondly, LocalColabFold version 1.3.0, which is based on the ColabFold/AlphaFold2_mmseqs2 notebook integrating the official AlphaFold2-Multimer³ weights (AlphaFold-multimer-v2) and specifically tuned for multimer prediction.

For version 1.0.0, we used the following options for all targets: “--homooligomer”, “--use_ptm”, “--use_turbo” “--max_recycle 3”, and “ --num_relax Top5”. However, for 1.0.0 we were GPU resource-limited to 1800 residues for the total length of all subunits. For 1.3.0, the GPU usage was more efficient and we were able to model complexes up to 2500 residues. We used templates and AMBER relaxation if the total length was <1000 with the following options: “--templates”, “--amber”, “--num-recycle 3”, and “--model-type auto”. For targets with lengths >1000 and <2500 residues, the “--templates” and “--amber” options were not used. The first stage resulted in the generation of up to 20 3D models (5 relaxed and 5 unrelaxed from each of the two methods).

In the second stage of the process, the first stage models were scored and ranked using ModFOLDdockR, which is a hybrid consensus approach for producing both global and local (interface residue) quality scores for predicted quaternary structures (see our ModFOLDdock server abstract for further details).

In the final stage, the top 5 ModFOLDdockR selected models were reformatted to mmCIF files using MAXIT⁴ and then used as input templates for our AF2MRefine protocol. The

AF2MRefine approach used the LocalColabFold 1.3.0 method with the “--custom-template-path” option, with 12 recycles and AMBER relaxation for targets <1000 residues or 3 recycles without AMBER for larger targets. For each model, the model rankings and predicted per-residue quality scores (pLDDT*100) from LocalColabFold were added to the B-factor column for each set of atom records.

For the very large complexes (>2500 residues), due to our limited GPU resources, we had to divide sequences up into overlapping fragments for submission to MultiFOLD. The resulting modelled fragments were then manually assembled, using structural superposition in PyMOL (<https://www.pymol.org>), to form larger, more complete models.

Availability

The MultiFOLD server is available at:

https://www.reading.ac.uk/bioinf/MultiFOLD/MultiFOLD_form.html

MultiFOLD is also available as a docker image here: <https://hub.docker.com/r/mcguffin/multifold>

1. Mirdita,M., Schütze,K., Moriwaki,Y., Heo,L., Ovchinnikov,S., & Steinegger,M. (2022) ColabFold: making protein folding accessible to all. *Nature Methods*. **19**(6), 679–682. doi: 10.1038/s41592-022-01488-1
2. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., Bridgland,A., Meyer,C., Kohl,S.A.A., Ballard,A.J., Cowie,A., Romera-Paredes,B., Nikolov,S., Jain,R., Adler,J., ... Hassabis,D. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**(7873), 583–589. doi: 10.1038/s41586-021-03819-2
3. Evans,R., O’Neill,M., Pritzel,A., Antropova,N., Senior,A., Green,T., Žídek,A., Bates,R., Blackwell,S., Yim,J., Ronneberger,O., Bodenstein,S., Zielinski,M., Bridgland,A., Potapenko,A., Cowie,A., Tunyasuvunakool,K., Jain,R., Clancy,E., ... Hassabis,D. (2021) Protein complex prediction with AlphaFold-Multimer. *In bioRxiv*. doi: 10.1101/2021.10.04.463034
4. Adams,P.D., Afonine,P.V., Baskaran,K., Berman,H.M., Berrisford,J., Bricogne,G., Brown,D.G., Burley,S.K., Chen,M., Feng,Z., Flensburg,C., Gutmanas,A., Hoch,J.C., Ikegawa,Y., Kengaku,Y., Krissinel,E., Kurisu,G., Liang,Y., Liebschner,D., ... Young,J.Y. (2019) Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB). *Acta Crystallographica Section D: Structural Biology*. **75**(Pt 4), 451–454. doi: 10.1107/S205979831900045

NBIS-AF2-standard, NBIS-AF2-multimer

See: *Elofsson*

Ligand pose estimation guided by predicted geometrical constraints with templates

Tingzhong Tian¹, Ziheng Zou², Shuya Li¹ and Jianyang Zeng^{1,*}

1 - Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China, 2 - Silexon AI Technology Co., Ltd., Nanjing, Jiangsu Province, China.

*Corresponding author: zengjy321@tsinghua.edu.cn

Key: Auto:Y; CASP_serv:Y; Templ:Y; MSA:N; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:N

Predicting the ligand binding structure is important for the drug design and was often realized by docking and molecular dynamics methods. The classical docking methods generally sample the possible poses of ligands and rank them based on physics-based, knowledge-based, or half-experience-based scoring functions. While those classical methods showed limited docking accuracy, our approach can improve docking performance by combining the classical scoring function with predicted intermolecular distance distributions, which was learned by a graph neural network-based model called Collie. The Collie model predicts ligand poses based on a given protein structure with a specified pocket center and uses AutoDock Vina and GNINA docking poses as input templates. In CASP 2015 ligand pose prediction tasks, we adopted the AlphaFold predicted protein structure provided by the CASP-hosted server and utilized a protein pocket detection model developed in our previous work to get pocket centers. Collie model showed significant improvement compared with the classical docking methods on benchmark datasets.

Methods

Protein preparation. Collie followed the same protein preparation workflow as MoG. In CASP15 ligand tasks, we utilized the AlphaFold¹ protein structures provided by the TS prediction server, which were transformed to pdb files by Biopython² package. Then the pdb files were modified by pdbfixer³. The protein pdb files were then used to build the anchor features and to prepare pdbqt files for the docking step with openbabel⁴.

Binding site prediction. Collie used same predicted binding sites as MoG. We utilized our pocket detection model to predict the pocket centers⁵. The pocket detection model was trained on PDBbind v2020 dataset and could recommend several possible pockets on a given protein. Then, using one pocket or multiple pockets was manually decided according to the protein size.

Ligand – protein anchor distance prediction. Collie model is designed to predict the pairwise distance distributions between protein anchors and ligand atoms with the templates from docking methods. Anchors were sampled points in protein pockets for representing the

subpocket-level protein features⁵. Compared with MoG, Collie accepted the reference anchor-atom distances as extra inputs, which would be transformed into the original distance features. Here we adopted the top 1 docking poses from AutoDock Vina⁶ and GNINA⁷ and calculated the features together.

Collie was trained to maximize the likelihood of the distribution parameter values given the real distances. We built the training set with Vina and GNINA docking poses in the general set of PDBbind v2016 dataset, and we used CASF2016 and newly updated data in 2020 as validation and test sets.

Ligand pose sampling and ranking. Collie and MoG shared the same ligand preparation and docking programs. For a ligand with predicted distance distributions, we built a statistical potential and combined the potential with Autodock Vina scoring function with a certain weight⁶. And we used the same Monto Carlo search algorithm and BFGS optimization algorithm as Vina.

In the ligand pose prediction task, we used rdkit⁸ to generate 3D structures from SMILES and transformed the structures to initial pdbqt files with openbabel for the pose sampling. The docking centers were the pocket centers provided in "Binding site prediction". Similar to Autodock Vina, our method could recommend multiple binding poses and rank them according to the mixed scoring function. We submitted the top 5 poses for the ligands in the task with a single pocket and the top 1 pose for each pocket in those tasks with multiple pockets.

Results

The binding pocket prediction model demonstrated SOTA accuracy in the test datasets, and Collie showed improvement compared with MoG on benchmark datasets.

Availability

The methods will be released when published.

1. J. Jumper, R. Evans, A. Pritzel, T. Green, et al. (2021). Highly accurate protein structure prediction with AlphaFold, *Nature*. 596, 583-589.
2. P.J.A. Cock, T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 25(11), 1422-3.
3. GitHub - openmm/pdbfixer: PDBFixer fixes problems in PDB files
4. O'Boyle, N.M., Banck, M., James, C.A. et al. (2011). Open Babel: An open chemical toolbox. *J Cheminform*. 3, 33.
5. S. Li, T. Tian, Z. Zhang, Z. Zou, D. Zhao, J. Zeng. (2022). PocketAnchor: Learning Structure-based Pocket Representations for Protein-Ligand Interaction Prediction. PREPRINT (Version 1) available at Research Square <https://doi.org/10.21203/rs.3.rs-1583468/v1>.

6. J. Eberhardt, D. Santos-Martins, A.F. Tillack, S. Forli. (2021). AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings, *J. Chem. Inf. Model.* 61, 3891–3898.
7. A.T. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri, D.R. Koes. (2021). GNINA 1.0: molecular docking with deep learning, *J Cheminform.* 13, 43.
8. RDKit: Open-source cheminformatics; <http://www.rdkit.org>.

Machine learning classifiers for protein-protein docking models and the Effect of Training Data Augmentation

D. Barradas-Bautista¹, A. Almajed², A. Vangone³, Z. Cao⁴, Luigi Cavallo⁴, P. Kalnis², R. Oliva⁵

1 Kaust Visualization Core lab, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. 2- InfoCloud Research Group, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. 3- Large Molecule Research, pharma Research & Development, Roche GmbH, Penzberg, Germany. 4- Kaust Catalysis Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. 5- Department of Sciences and Technologies, University "Parthenope" of Naples, Napoli, Italy.

romina.oliva@uniparthenope.it

Key: *Scoring; Machine learning; Docking models benchmark; Data augmentation*

Scoring is a critical step in docking and represents, in fact, a separate challenge of the CAPRI (Critical Assessment of PRedicted Interactions) experiment since 2006.¹ Traditionally, scoring functions for protein-protein docking models (DMs) are either energy-based or knowledge-based. However, over the years, a wide variety of algorithms have been developed, some of them combining the above potentials into a hybrid approach or integrating them with evolutionary information, others based on alternative approaches, such as the consensus of the inter-residue contacts at the interface of the complex.²⁻⁴ Nowadays, over 100 scoring functions are available from the CCharPPI web server,⁵ while more potentials can be obtained from other public sources. These are all descriptors of the protein-protein complexes, which can be in principle combined to gain an improved performance in assessing the quality of predicted 3D models.

Herein, we present the results of a machine learning (ML) approach we developed to exploit all the scoring functions we could collect from public sources. To this aim, we generated a set of $\approx 7 \times 10^6$ DMs with three different docking programs for the 230 complexes in the protein-protein interaction benchmark 5 (BM5).⁶

Furthermore, we explored the effect of training data augmentation on the above models.

Methods

Three different ML approaches, Random Forest (RF), Support Vector Machines (SVMs), and single-layer Perceptron (PRC), were used to train classifiers with 157 different scoring functions (features), including 93 potentials from CCharPPI, and 32 features calculated by our tools CONSRANK and COCOMAPS,⁷ within the scikit learn python library and within pyplot/seaborn for visualization. For each of the 230 protein-protein complexes (targets) in BM5, we generated a total of 30,000 DMs with FTDock,⁸ ZDock,⁹ and HADDOCK.¹⁰ The quality of the generated DMs was assessed following the CAPRI protocol. Balanced and unbalanced "core" datasets were built from the DMs above for the training and test processes.

To augment our training dataset, we started from the 1,392 protein pairs from the high confidence human interactome in the I3D database,¹¹ to create a "silver set" that we then added to the above balanced "core" set to obtain our "augmented set". The augmented dataset consists

of 1,553 protein pairs, with a total of 4,224,740 labeled DMs (being 76-fold larger than the original training set). Labeling of the “silver set” was performed with a Snorkel statistical modeling approach.

Results

We trained three different machine learning approaches on a balanced dataset of $\approx 7 \times 10^4$ DMs, Bal-BM5, and validated them on an unbalanced dataset of $\approx 7 \times 10^5$ DMs, 3K-BM5. To our knowledge, these are the largest datasets used to develop a ML classifier in this field. We made them open access, labeled with their respective quality assignment (incorrect, acceptable, medium- and high-quality, according to the CAPRI criteria) and complete with the values of calculated features, to be used as reference benchmarks both for developing and comparing different scoring methods using classic empirical potentials, and for the training of ML-based methods.

Since the RF approach showed the best performance, we trained a final RF-based classifier by optimizing its hyperparameters. The final RF classifier was named CoDES (COnservation Driven Expert System), as, within the 16 selected features optimizing its performance, the one having by far the highest importance is the CONSRANK score, which represents the average conservation (frequency) of the inter-residue contacts featured by a given DM, relatively to the set of models it belongs to. Testing of CoDES on the CAPRI Score Set showed it to outperform any single scorer in the corresponding CAPRI Rounds and to be able to top rank not just correct but medium- and high-quality DMs.¹² Overall testing on independent datasets resulted in CoDES equaling or exceeding the performance of the few state-of-the-art machine learning methods available in the literature.

Classifiers based on the above machine learning approaches were also trained on an “augmented set” we generated by a weak supervision approach, which proved to significantly improve the performance, at least for the RF-based models.

Availability

Generated DMs in the “core” and “augmented” datasets are available at <https://doi.org/10.5281/zenodo.4012018> and <https://repository.kaust.edu.sa/handle/10754/666961>.

ML algorithms are available at <https://github.com/D-Barradas/CoDES>, https://github.com/D-Barradas/hAIkal_TF2_exploratory_test, and https://colab.research.google.com/drive/1vbVrJcQSf6_C3jOAmZzgQbTpuJ5zC1RP?usp=sharing.

1. Lensink, M. F. et al. (2007) Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins* 69, 704–18.
2. Chermak, E., Petta, A., Serra, L., Vangone, A., Scarano, V., Cavallo, L., Oliva, R. (2015) CONSRANK: a server for the analysis, comparison and ranking of docking models based on inter-residue contacts, *Bioinformatics* 31, 1481–3.

3. Oliva, R., Vangone A., Cavallo L. (2013) Ranking multiple docking solutions based on the conservation of inter-residue contacts. *Proteins* 81, 1571–84.
4. Chermak, E., De Donato, R., Lensink M.F., Petta, A., Serra, L., Scarano, V., Cavallo L., Oliva R. (2016) Introducing a Clustering Step in a Consensus Approach for the Scoring of Protein-Protein Docking Models, *PLoS ONE* 11, e0166460.
5. Moal, I. H. et al. (2015) CCharPPI web server: computational characterization of protein-protein interactions from structure. *Bioinformatics* 31, 123–25.
6. Vreven, T. et al. (2015) Updates to the Integrated ProteinProtein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J Mol Biol* 427, 3031–41
7. Vangone, A., Spinelli R., Scarano V., Cavallo L., Oliva R. (2011) COCOMAPS: a web application to analyze and visualize contacts at the interface of biomolecular complexes. *Bioinformatics* 27, 2915–16.
8. Gabb, H. A. et al. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* 272, 106–20.
9. Chen, R. et al. (2003) ZDOCK: an initial-stage protein-docking algorithm. *Proteins* 52, 80–87.
10. de Vries, S. J. et al. (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins* 69, 726–733.
11. Mosca R, Ceol A, Aloy P. (2013) Interactome3D: adding structural details to protein networks. *Nat Methods* 10, 47–53.
12. Barradas-Bautista D., Cao Z., Vangone A., Oliva R., Cavallo L. (2021) A random forest classifier for protein–protein docking models. *Bioinform Adv* 2, vbab042.

Introducing an iterative process in a consensus algorithm for the scoring of protein-protein docking models

T. Ricciardelli¹, D. Barradas-Bautista², Z. Cao¹, L. Cavallo¹, R. Oliva³

1 Kaust Catalysis Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. 2 Kaust Visualization Core Lab, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. 3 Department of Sciences and Technologies, University "Parthenope" of Naples, Napoli.

romina.oliva@uniparthenope.it

Key: Scoring; Consensus; Iteration; Enrichment; CONSRANK

Correctly scoring protein-protein docking models is an open challenge, continuously monitored in the CAPRI Rounds since 2006.¹ To the traditional scenario of energy-based and knowledge-based scoring potentials, we added CONSRANK (CONSENSUS RANKing), an alternative approach which uses the conservation rate (or frequency within the ensemble of decoys) of inter-residue contacts featured at the interface by a given model, as a measure of its reliability.²⁻⁴ CONSRANK has been blindly tested in CAPRI since 2014, where it proved to provide state-of-the-art predictions, especially in terms of top-1 success rate, i.e. the ability to locate a correct solution at the 1st ranked position.⁵⁻⁸

To further explore the CONSRANK potential, we present here an algorithm development, IterCONSRANK, where CONSRANK is applied iteratively to a set of models to be scored, taking to the next iteration only those ranked in the top $N\%$ positions. Thus, IterCONSRANK decreases the number of models analyzed at each step, discarding the bottom $(100-N)\%$ scored ones. The method performance was tested both in terms of percentage of correct solutions for the ensemble at each step, and of success rate, i.e. the ability of locating correct solutions at the top positions of the ranking

Methods

To test the performance of IterCONSRANK, we used a total of 175,985 models, for 65 targets from two public benchmarks, 3K-BM5up⁹⁻¹⁰ and Score_set.¹¹ Based on the featured percentage of correct solutions (%corr), we classified the 52 3K-BM5up targets as "very difficult" cases (19 targets), "difficult" cases (12 targets), "medium" cases (8 targets), and "easy" cases (13 targets), for scoring. This dataset was used to select the optimal iteration parameters and perform detailed analyses of the performance by different metrics. The second dataset, Score_set, includes 14 interfaces to be scored, of which 3 represent "very difficult", 3 "difficult", 5 "medium" and 3 "easy" cases for scoring. This dataset was used to test again the method performance on an independent CAPRI dataset. Different values have been tested for the iteration threshold ($N\%$): 80%, 85%, 90%, 95%. To fully explore the method potential, the maximum number of steps has been set to 30. At the 30th step, with a cut-off of 85%, the number of models is reduced to $\approx 1\%$ of the original set.

Results

We tested IterCONSRANK on the 3K-BM5up and Score_set benchmarks, by varying the number of iteration steps and ranking thresholds for iteration (N%). Optimizing these parameters allowed us to significantly enrich the docking decoys in correct solutions (i.e. to increase the %corr value) for all the represented ranges of target difficulty.

To get a deeper understanding of the obtained results, we performed detailed analyses of targets for which a strong improvement of the scoring prediction was achieved. Such analyses outlined how the background noise in the recorded inter-residue contacts is cleared over the iteration process while the native consensus clearly emerges. This allows increasing the %corr of the ensembles and the scoring success rate itself. As the %corr values directly correlate with the rate of difficulty of the scoring process, IterCONSRANK proposes itself not only as an efficient scoring algorithm but also as a preprocessing step allowing to increase the percentage of correct solutions in the examined ensemble of decoys, for the subsequent application of other scoring functions.

Finally, the iterative approach was shown to further improve the CONSRANK performance in terms of top-1 success rate, clearly outperforming all the 100+ publicly available scoring potentials¹² we comparatively tested on the above datasets.

Availability

Code available upon request from the authors.

1. Lensink, M. F. et al. (2007) Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins* 69, 704–18.
2. Oliva, R., Vangone A., Cavallo L. (2013) Ranking multiple docking solutions based on the conservation of inter-residue contacts. *Proteins* 81, 1571–84.
3. Chermak, E., Petta, A., Serra, L., Vangone, A., Scarano, V., Cavallo, L., Oliva, R. (2015) CONSRANK: a server for the analysis, comparison and ranking of docking models based on inter-residue contacts, *Bioinformatics* 31, 1481–3.
4. Chermak, E., De Donato, R., Lensink M.F., Petta, A., Serra, L., Scarano, V., Cavallo L., Oliva R. (2016) Introducing a Clustering Step in a Consensus Approach for the Scoring of Protein-Protein Docking Models, *PLoS ONE* 11, e0166460.
5. Lensink M.F. et al. (2016) Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins* 84, 323-348.
6. Barradas-Bautista, D., Cao, Z., Cavallo, L., Oliva, R. (2020) The CASP13-CAPRI targets as case studies to illustrate a novel scoring pipeline integrating CONSRANK with clustering and interface analyses. *BMC bioinformatics* 21, 1-18.
7. Lensink M.F. et al. (2019) Blind prediction of homo-and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins* 87, 1200-21.
8. Lensink M.F. et al. (2021) Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins* 89, 1800-23.

9. Vreven, T. et al. (2015) Updates to the Integrated ProteinProtein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J Mol Biol* 427, 3031–41.
10. Barradas-Bautista D., Cao Z., Vangone A., Oliva R., Cavallo L. (2021) A random forest classifier for protein–protein docking models. *Bioinform Adv* 2, vbab042.5.
11. Lensink M.F., Wodak S.J. (2014) Score_set: a CAPRI benchmark for scoring protein complexes. *Proteins* 82, 3163-3169.
12. Moal, I. H. et al. (2015) CCharPPI web server: computational characterization of protein-protein interactions from structure. *Bioinformatics* 31, 123–25.

OpenFold: A trainable reproduction of AlphaFold2

Gustaf Ahdriz¹, Nazim Bouatta², Sachin Kadyan¹, Mohammed AlQuraishi¹

¹ – Department of Systems Biology, Columbia University, ² – Harvard Medical School

ma4129@cumc.columbia.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:N; Dist:N; Tors:Y; DeepL:Y; EMA:Y; MD:Y.*

At CASP14, AlphaFold2¹ revolutionized structural biology by predicting protein structures with experimental accuracy. The following year, DeepMind released the paper, model weights, and accompanying code for the model’s inference pipeline. This official implementation, however, (i) lacks the code and data required to train models for new tasks, such as predicting protein-ligand complexes or antibody structures, (ii) is unoptimized for commercially available computing hardware, making large-scale prediction campaigns impractical, and (iii) remains poorly understood with respect to how training data and regimen influence accuracy.

Here, we present OpenFold, a fully open-source, trainable, and optimized reproduction of AlphaFold2 using PyTorch. We trained it from scratch and used our model with its original weights at CASP15.

Methods

OpenFold hews as closely as possible to the training and inference procedures described in the original supplement to the AlphaFold2 paper, which we summarize briefly here. The model’s hyperparameter settings are identical to those used by AlphaFold2. Training data consisted of 130k structures from PDB² and a filtered self-distillation set of 270k Uniclust30³ chains with structures predicted by AlphaFold2, all with accompanying MSAs. PDB MSAs are computed from MGnify⁴ and Uniref90⁵ with JackHMMER⁶ and Uniclust30 + BFD with HHblitsv3⁷. Distillation MSAs are computed using an all-against-all HHblits search against Uniclust30 itself. Structural templates are obtained by searching PDB70⁸ using the UniRef90 MSA. We trained the model using the abbreviated training schedule outlined in the AlphaFold2 supplement and all of the same losses and optimizer settings used to train AlphaFold2. At a high level, the model is trained in several phases: one “initial training” phase that uses relatively short MSA inputs and shorter crops of training proteins (256 residues), one “finetuning” phase that uses deeper MSAs, longer crops (384 residues), and a structural violation loss, and finally an additional finetuning phase with the predicted-TM score module activated. After these three phases were finished, we trained an additional branch of the model with templates disabled starting from the end of the initial training phase. For additional details, we direct readers to the aforementioned supplement.

During inference, the model accepts both MSAs and structural templates as inputs. MSAs are computed exactly as they were during training. Output structures are relaxed with AMBER⁹.

Five predictions were generated for each target using as many sets of model parameters, each corresponding to a peak in the validation IDDT-C α during training. To maximize the diversity of outputs, parameters were sampled as sparsely as possible while maintaining output quality; they range from the beginning of the finetuning stage to the very end of the pTM phase. Models were run with their own config settings differing primarily in the number of templates (0 – 6), recycling iterations (4 – 20), and extra MSA depths (1024 – 5120). Outputs were ranked in order of descending mean pLDDT, AlphaFold2’s built-in confidence measure.

To predict complexes, we used the inference-time hack known as AlphaFold-Gap¹⁰. This involves concatenating chains in the complex, marking chain boundaries with constant-sized gaps in the residue indices used to compute positional embeddings. MSAs and templates are combined in similar fashion. The models are run on the resulting “chain” as usual, without any further training. Predictions are then split back into their component parts.

All targets were run on a single 40GB A100 GPU.

Manual interventions were fairly minimal. For a handful of chains with low average confidence, we re-ran predictions with tweaked config settings, mainly adding additional recycling iterations and templates. For extremely long chains and complexes, we reduced the number of recycling iterations and templates to save compute and enabled OpenFold’s memory-saving features, including low-memory attention implementations, tensor offloading, and in-place operations.

It should be noted that the models were not fully trained by the beginning of the competition; additional training and finetuning were performed until July 12, after which the model weights were frozen. It should also be noted that a bug in our inference pipeline initially caused the model to run with shallower MSAs than intended (containing 128 sequences instead of 512). This issue was resolved by late June.

Availability

OpenFold code, weights, model config presets, and, soon, its preprint are publicly available at <https://github.com/aqlaboratory/openfold>. The model’s complete training data is made available via the Registry of Open Data on AWS (RODA) at <https://registry.opendata.aws/openfold/>.

1. Jumper, J. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **577** (7792), 583–589. 10.1038/s41586-021-03819-2.
2. wwPDB consortium (Oct. 2018). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research* **47** (D1), D520–D528. 10.1093/nar/gky949.
3. Mirdita, M., Driesch, L. von den, Galiez, C., Martin, M. J., Söding, J., and Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research* **45** (D1), D170–D176. 10.1093/nar/gkw1081.
4. Mitchell, A. L. et al. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research* **48** (D1), D570–D578. 10.1093/nar/gkz1035.

5. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Uniprot Consortium (2013). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31** (6), 926–932. 10.1093/bioinformatics/btt473.
6. Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11** (1), 431. 10.1186/1471-2105-11-431.
7. Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9** (2), 173–175. 10.1038/nmeth.1818.
8. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20** (1), 473. 10.1186/s12859-019-3019-7.
9. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65** (3), 712–725. 10.1002/prot.21123.
10. Baek, M. (2021). Twitter post: Adding a big enough number for “residue_index” feature is enough to model hetero-complex using AlphaFold (green&cyan: crystal structure / magenta: predicted model w/ residue_index modification). <https://twitter.com/minkbaek/status/1417538291709071362>.

Improved Template-based Protein Structure Prediction using PBEscore

Wei Cao, Lu-Yun Wu, Zhi-Xin Wang and Xian-Ming Pan[†]

1 - Key Laboratory of Ministry of Education for Protein Science, School of Life Sciences, Tsinghua University, Beijing 100084, China.

[†] Corresponding author: Dr. Xian-Ming Pan (pan-xm@mail.tsinghua.edu.cn)

Key: *Auto:Y; CASP_serv:Y; Templ:Y; MSA:Y; Fragm:Y.3,9; Cont:Y; Dist:Y; Tors:N; DeepL:N; EMA:Y; MD:Y*

In CASP15, we have tested our new pipeline based on our workflow in CASP14. Panlab and Pan_Server followed the same pipeline but we have performed some manual intervention in expert group(Panlab). Briefly, we construct 3D-models based on our remoter homologous protein recognition method with alignments generated, RosettaCM¹ to model 3D structures, 3DRobot² to generate decoys which are near top-ranked structures, then, we applied PBEscore, a novel knowledge-based-energy scoring function, to guide protein conformation search and refinement.

Methods

Given a target sequence, our pipeline has 4 steps as fellow.

1. Remote protein homology detection by alignment-based PairThreading. There are many alignment methods for remote protein homology detection, but these methods are based on the assumption that the types of residues at different positions are independent of each other. We break this assumption and propose a method, PairThreading, based on residue pair substitution information. PairThreading obtains position-specific residue pair substitution information indirectly from the position-specific score matrices (PSSMs) rather than directly from the multiple sequence alignments (MSAs) to avoid statistical non-convergence problem. Thus, PairThreading can detect more remote homologous proteins and can generate more accurate alignments. For targets whose sequence length larger than 1000, our method have a dynamic strategy to detect domain region and generate fragment for further 3D modeling and template hybridize.

2. Constructing 3D-models by RosettaCM¹. We use RosettaCM¹ to construct 3D-models based on the single or multiple templates which are selected by PairThreading along with corresponding alignments.

3. Generating decoys using 3DRobot². After selecting top-ranked models by PBEscore, we use 3DRobot to create protein structure decoys which have enhanced hydrogen-bonding and compactness interactions.

4. Ranking 3D-models by PBEscore. PBEscore is a novel knowledge-based-energy scoring function, simply considering the interactions of peptide bonds, rather than, as conventionally, the residues or atoms as the most important energy contribution. This energy function is trained on in-house dataset and has outstanding performance on several independent benchmark datasets. We applied PBEscore in every ranking steps involved in our pipeline.

5. Refinement 3D-models by OpenMM³. We ran the molecular dynamics to refine the Top 5 ranked models. The program PDBFixer was used to add hydrogen atoms, N- and C-terminal patches to selected models. All simulations were run using OpenMM³ under AMBER14⁴ force field.

1. Song, Y., DiMaio, F., Wang, R. Y., Kim, D., Miles, C., Brunette, T., Thompson, J., & Baker, D. (2013). High-resolution comparative modeling with RosettaCM. *Structure*. 21(10), 1735–1742.
2. Deng, H., Jia, Y., & Zhang, Y. (2016). 3DRobot: automated generation of diverse and well-packed protein structure decoys. *Bioinformatics*. 32(3), 378–387.
3. Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L. P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., & Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol*. 13(7), e1005659.
4. Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., & Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput*. 11(8), 3696–3713.

Incorporating AF and server derived ambiguous data into MD simulations

A. Mondal, L. Chang, R. Esmaceli, A. Perez

University of Florida

perez@chem.ufl.edu

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; DeepL:Y; EMA:N; MD:Y*

MELD has been successful in predicting protein and its complex structures. MELD was used in the last couple of CASPs with good successes for predicting protein structure with data-assisted category, mainly NMR data assisted. In this CASP, our goal was to model more RNA structures and conformational ensembles of the kinases.

In CASP 15 event, we have submitted 7 predictions in the- 4 RNA model and 3- kinases conformational ensembles. All simulations were performed with our local resources on HiPerGator Supercomputer at the University of Florida.

Methods

MELD (Modeling Employing Limited Data) is a Bayesian inference-based enhanced sampling tool which accelerates molecular dynamics in presence of data which works a plugin to OpenMM(1). Data from different sources with different levels of quality can be used in MELD. We map data as spatial (e.g. distance or dihedral) information between atoms in the systems to limit the conformation search space. The way these restraints work is, there is no energy penalty when the data is satisfied and increases quadratically (and then linearly after a cutoff) otherwise. The data has the peculiarity that some of the data might not be correct, we generally do not know what percent to trust as well as which data to trust. The Bayesian aspect of the method comes from inferring which subset of the data is most compatible with the system given a prior (given by the molecular dynamics force field). MELD uses H-T,REMD(2) protocol where at higher replicas, temperature is high and distance restraints are weak, and at lower replicas, temperature is low and restraints are strong, which facilitate multiple folding-unfolding events through exploration-exploitation of the energy landscape. All simulations were run for at least a microsecond using 30 replicas with temperature ranging from 300K to 500K, the GBneck2 implicit solvent model(3) and the ff14SB force field for side chains with ff99SB for the backbone for proteins(4) and with OL3 forcefield for RNAs(5).

Data used in simulations:

We used two different approaches for protein and RNA modelling. Details are given in the following paragraph. For RNA, T1107 and T1108 were supposed to be homodimers. We used

RNAfold WebServer(6) (from the ViennaRNA package) to predict their secondary structures and then we used modelling to assemble them into multimers based on known similar systems. From the server predicted structures, we deduced a set of base pairs which are taking part in the helix pairing in the monomer. And from the known complexes we deduced base pair contacts that might be present in the complex. We put distance restraints (as mentioned above) on those base-pairs with 80% and 50% accuracy for the monomer restraints and complex restraints respectively. Target T1116 and T1128 are monomer. For these two we only had distance restraints for monomer base pairs, and we enforced those with 80% accuracy.

We predicted conformational ensemble of 3 kinase systems- T1195, T1196, and T1197. At the first step we predicted their structures with AlphaFold (AF) colab with MSA searched from Uniref100.(7, 8) Then, the top model from AF was selected as the starting conformation for MELD simulation. We also extracted the distograms from AF and created restraints based on their confidence estimation from AF. Then, we apply cartesian restraints on the CA atoms of the residues that are part of high confidence distograms (>90% in 2Å range) to model them as flatbottom harmonic restraints as mentioned above. For the low confidence distograms, we further separated them in two groups: local distograms (for $i-j < 4$, where i and j are two residues in a distogram) and global distograms (for $i-j > 4$, where i and j are two residues in a distogram). We designed two protocols for each, both with the cartesian restraints and one with global distogram restraints and another with local distogram restraints. We used 50% and 80% of the starting accuracy for the global and local distograms respectively, where the accuracy parameter would self-optimize as the simulation proceeds.

In the end of the simulations, we performed hierarchical clustering using cpptraj(9) on five lowest temperature replicas, then we selected the top 5 models based on the population of the clusters. For RNA targets, we submitted those structures as our predictions. For protein conformation ensemble targets we further refined those selected models with AF using them as templates and we submitted AF output as the final predictions.

Results

MELD traditionally has been successful for NMR assisted protein structure prediction and small globular protein structure prediction(10–12). These types of targets were not part of the CASP in the current edition. With the goal of expanding the use of MELD, we attempted different modeling challenges this time. However, none of our predicted target has experimental structures released by CASP yet. We are hopeful that our sampled ensemble has some close to native conformations, but we might be missing them while picking top5. A better scoring method for RNA conformations would help in future.

Availability

OpenMM, AmberTools, MELD and the MELD-OpenMM plugin are all available and free to use. Our MELD frontend can be accessed at: [git@github.com:maccallumlab/meld.git](https://github.com/maccallumlab/meld.git), and the MELD-openMM plugin can be accessed at: [git@github.com:maccallumlab/meld-openmm-plugin.git](https://github.com/maccallumlab/meld-openmm-plugin.git). (font different, and these git links is not valid)

1. JL MacCallum; A Perez; KA Dill. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proceedings of the National Academy of Sciences of the United States of America* 112, 6985–6990 (2015)
2. Y Sugita; Y Okamoto. Replica exchange molecular dynamics method for protein folding simulation. *Chemical physics letter* 314, 141–151 (1999)
3. H Nguyen; DR Roe; C Simmerling. Improved Generalized Born Solvent Model Parameters for Protein Simulations. *Journal of Chemical Theory and Computation* 9, 2020–2034 (2013)
4. JA Maier; C Martinez; K Kasavajhala; L Wickstrom; KE Hauser; C Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* 11, 3696–3713 (2015)
5. M Zgarbová; M Otyepka; J Šponer; A Mládek; P Banáš; TE Cheatham; P Jurečka. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J Chem Theory Comput* 7, 2886–2902 (2011)
6. AR Gruber; R Lorenz; SH Bernhart; R Neuböck; IL Hofacker. The Vienna RNA Websuite. *Nucleic Acids Res* 36, W70–W74 (2008)
7. J Jumper; R Evans; A Pritzel; T Green; M Figurnov; O Ronneberger; K Tunyasuvunakool; R Bates; A Židek; A Potapenko; A Bridgland; C Meyer; SAA Kohl; AJ Ballard; A Cowie; B Romera-Paredes; S Nikolov; R Jain; J Adler; T Back; S Petersen; D Reiman; E Clancy; M Zielinski; M Steinegger; M Pacholska; T Berghammer; S Bodenstein; D Silver; O Vinyals; AW Senior; K Kavukcuoglu; P Kohli; D Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature* (2021)
8. M Mirdita; K Schütze; Y Moriwaki; L Heo; S Ovchinnikov; M Steinegger. ColabFold: making protein folding accessible to all. *Nat Methods* 19, 679–682 (2022)
9. DR Roe; TE Cheatham. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation* 9, 3084–3095 (2013)
10. JC Robertson; R Nassar; C Liu; E Brini; KA Dill; A Perez. NMR-assisted protein structure prediction with MELDxMD. *Proteins: Structure, Function and Bioinformatics* 87, 1333–1340 (2019)
11. A Mondal; A Perez. Simultaneous Assignment and Structure Determination of Proteins From Sparsely Labeled NMR Datasets. *Frontiers Mol Biosci* 8, 774394 (2021)
12. A Perez; JA Morrone; E Brini; JL MacCallum; KA Dill. Blind protein structure prediction using accelerated free-energy simulations. *Science Advances* 2, e1601274 (2016)

Protein tertiary and quaternary structure prediction using AlphaFold2 with various metagenomic databases

Toshiyuki Oda

Infinite Curation Inc.

t_oda@i-curation.co.jp

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:Y*

At CASP14, AlphaFold2, developed by DeepMind, demonstrated outstanding accuracy of monomeric structure prediction¹⁻³. After the competition, its derivative, AlphaFold-Multimer (hereafter, both are referred to as AF2), was presented⁴. It also showed excellent performance in predicting multimeric structures. Since their inference code and weights are publicly available under the generous license, their predictions will be the baseline for CASP15.

Therefore, I set the following challenges for CASP15: (1) Collect a sufficient number of evolutionary related sequences for inputs of AF2. (2) Improve the structures generated by AF2.

Methods

Collecting evolutionary related sequences: I used various tools and DB(database)s to collect evolutionary related sequences (Table 1). The MSA(Multiple Sequence Alignment)s created by B, C, D, E, and F were merged, and the resulting MSAs were clustered by identity threshold (90 %, 95 %), filtered by identity with the query sequence (60 %, 80 %), or sent to the next step without any arrangement.

Model building: The input MSAs are fed into a customized AF2 pipeline using the official AF2 weights. This pipeline is essentially the same as the original pipeline, although it is more flexible (e.g. more internal variables that can be changed with arguments) than the original pipeline. For multimeric targets, the monomer mode of AF2 was also used with adding gaps (as extra-residue index^{5,6}) between different subunits and all final unrelaxed models were processed with refinement script (described below) to produce the self-reliability metrics for multimer.

For multimer targets, template base modelling with TM-align⁷ was applied when I found multimeric templates which matched the stoichiometry information provided by the CASP organizer. When the target was too large to fit in GPU's RAM, two approaches were used to build the model: (1) building the entire structures on CPU, and (2) building the partial models using MSAs which were divided into multiple pieces. And the partial models were concatenated after prediction.

Model selection: The predicted models were sorted according to the self-reliability metrics (plddt or (i)ptm) produced by AF2. The top one model was always selected. The remaining 4 models were selected considering the metrics and TM-score⁸ with the other selected models to take variety of the conformations. Human intervention was introduced when there

were issues in the models (e.g., the refined models sometimes have many atom clashes, the concatenated models did not yield metrics) or there were the specific notes provided by the CASP organizer.

Refinement: The selected 5 models were refined by an in-house fine-tuned AF2, which takes the protein 3D structures as input and outputs the refined structures. The details of this customized AF2 will be described elsewhere.

If structures with good self-reliability metrics could not be obtained, additional runs were performed by selecting other models, or randomly changing the position of chains or atoms in the input structures for refinement.

Relax: The relaxation step defined in official AF2 scripts was run before submission. When the program failed or there was no time to run the step, the models before the relaxation were submitted.

Target specific process: Since the method pipeline was completed during the season and there were many different types of targets, the details of the applied procedure varied from target to target. As an extreme example, T1109 had a mutated residue at position 183, therefore all residues at that position in the MSAs were changed to alanine.

	tool	query	DB	description
A	PZLAST ⁹	target sequence	public metagenomic amino acid sequences	The hit sequences were aligned with jackhmmer ¹⁰ and assembled with a simple python script.
B	PSI-BLAST _{exB} ^{11,12}	MSA made at A	nr+in-house metagenome database	The in-house metagenome database was derived from amino acid sequences or nucleotide sequences in Assembly database ¹³ . The nucleotide sequences were translated to amino acids by prodigal ¹⁴ . The hit sequences were aligned with jackhmmer. And when the number of hits was small, MSAs made at A were merged.
C	hhblits ¹⁵	MSA made at B	Uniclust30 ¹⁶	
D	hhblits	MSA made at B	BFD https://bfd.mmseqs.com/	If the number of hit sequences was large, the sequences were filtered by hhfilter ¹⁵ with the option “-id 100 -cov 30 -diff 10000”.
E	jackhmmer	target sequence	Uniprot/TrEMBL ¹⁷	
F	jackhmmer	target sequence	MGnify ¹⁸	The same procedure was used as in D.

Table 1. The tools and DBs used for sequence similarity search.

1. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).
2. Pereira, J. *et al.* High-accuracy protein structure prediction in CASP14. *Proteins* **89**, 1687-1699, doi:10.1002/prot.26171 (2021).
3. Jumper, J. *et al.* Applying and improving AlphaFold at CASP14. *Proteins* **89**, 1711-1721, doi:10.1002/prot.26257 (2021).
4. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.2010.2004.463034, doi:10.1101/2021.10.04.463034 (2022).
5. Humphreys, I.R. *et al.* Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805, doi:10.1126/science.abm4805 (2021).
6. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun* **13**, 1265, doi:10.1038/s41467-022-28865-w (2022).
7. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-2309, doi:10.1093/nar/gki524 (2005).
8. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-710, doi:10.1002/prot.20264 (2004).
9. Mori, H. *et al.* PZLAST: an ultra-fast amino acid sequence similarity search server against public metagenomes. *Bioinformatics*, doi:10.1093/bioinformatics/btab492 (2021).
10. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).
11. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
12. Oda, T., Lim, K. & Tomii, K. Simple adjustment of the sequence weight algorithm remarkably enhances PSI-BLAST performance. *BMC Bioinformatics* **18**, 288, doi:10.1186/s12859-017-1686-9 (2017).
13. Kitts, P.A. *et al.* Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res* **44**, D73-80, doi:10.1093/nar/gkv1226 (2016).
14. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).
15. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473, doi:10.1186/s12859-019-3019-7 (2019).
16. Mirdita, M. *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res* **45**, D170-D176, doi:10.1093/nar/gkw1081 (2017).
17. The UniProt Consortium UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480-D489, doi:10.1093/nar/gkaa1100 (2021).
18. Mitchell, A.L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res* **48**, D570-D578, doi:10.1093/nar/gkz1035 (2020).

PICNIC

QUIC

Refining AlphaFold TS models using 3D residual and convolutional neural networks

Chenguang Zhao, Tong Liu, Ross Campbell Stewart, Zheng Wang*

Department of Computer Science, University of Miami

zheng.wang@miami.edu

Key: Auto:Y; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:N; MD:N.

We present two computational predictors in CASP15, named QUIC and PICNIC, for protein tertiary structure (TS) prediction. QUIC and PICNIC were trained to refine the predictions of af2-standard₁ provided by the CASP15. QUIC was implemented as a three-dimensional (3D) residual neural network (Resnet), while PICNIC was built as a 3D convolutional neural network. Both servers were trained with 120 AlphaFold2₁ predicted models during CASP14.

Methods

Our refinement is performed at the atom level. For N, CA, C, and O atoms of each amino acid, we created an 81x81x81 mesh cube with each small cell in the cube having a side length of 0.1 Å. We put the atom in the AlphaFold model that our algorithm will refine at the center cell of the mesh cube, and we also put the other neighboring atoms that are within the boundaries of the cube in the corresponding cells. During training, we superpositioned each AlphaFold2 predicted model with the corresponding native structure, and then we used the position of the corresponding atom in the native structure as where the target value 1 was located. All other cells in the cube were assigned target values of 0s.

We generated 10 features for each cell if at least an atom exists in it. If no atom exists in a cell, the features of that cell were assigned 0s, and if there are ≥ 2 atoms existing in a cell, we used the averaged scores of the features. Those features are atom existence, an amino acid token, an atom type token, averaged ESM₂, MASS2-CASP14 and LAW-CASP14 predicted local qualities, and sinusoidal positional encoding. We trained the following two methods using the same features to refine the AlphaFold2 predicted structures.

QUIC contains eight Conv3D-BatchNorm3D-LeakyReLU-Dropout residual blocks followed by one Conv3D-BatchNorm3D layer. QUIC allows each atom to move up to two cubes on an axis.

PICNIC contains 10 Conv3D-BatchNorm3D-LeakyReLU-Dropout layers followed by a Conv3D-BatchNorm3D layer. PICNIC only predicts atoms belonging to the coiled coils.

For both predicted protein structures of QUIC and PICNIC, we added hydrogen and other missing atoms using PDBFixer and relaxed the predicted protein structure using OpenMM₃ with the same parameters used by AlphaFold2. After that, we used MASS2-CASP14 to evaluate the

final predicted structure, the local quality scores of which were submitted to CASP as quality scores.

1. Jumper, J., et al., Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021. 596(7873): p. 583-589.
2. Rives, A., et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 2021. 118(15): p. e2016239118.
3. Eastman, P., et al., OpenMM 7: Rapid development of high-performance algorithms for molecular dynamics. *PLoS computational biology*, 2017. **13**(7): p. e1005659.

RaptorX: protein structure prediction by deep attention networkXiaoyang Jing¹, Fandi Wu^{1,2}, Xiao Luo¹, Lupeng Kong³, Jinbo Xu¹

¹ Toyota Technological Institute at Chicago, Chicago, IL 60637, USA; ² Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 626011, China; ³ Beijing Changping Laboratory, Beijing, China

jinboxu@gmail.com

Significant improvement has been achieved in computational protein structure prediction in recent CASPs, especially the deep ResNet- and Transformer-based methods¹⁻⁵. In CASP15, we developed an automated pipeline for protein structure prediction by employing following strategies: 1) in addition to the uniref90, mgnify, bfd and uniclust30 databases, we collected an in-house metagenome sequence database; 2) in addition to the templates from PDB, we used decoy models predicted by AlphaFold2 as extra templates; 3) we used our in-house attention-based methods and AlphaFold2 to predict structures simultaneously and selected top models based on the predicted confidence (i.e. pLDDT); 4) we used the predicted structure as query structure to search templates from template databases.

Methods*MSA search*

The main pipeline to generate the multiple sequence alignment (MSA) is similar to the AlphaFold2 MSA generation process. We use the Jackhmmer to search the uniref90 and mgnify, and use the HHBlits to search the bfd and uniclust30. It should be noted that all databases (except for the bfd) have been updated at the end of April 2022. If the MSA is shallow, we will use the HHBlits to search our in-house metagenome sequence database to enrich the MSA. Our in-house metagenome sequence databases are built from SMAG⁶, MetaEuk⁷, TOPAZ⁸, MG⁹, GPD¹⁰ and IMG/M (from June 2018 to January 2022)¹¹ with a maximum sequence identity of 90%.

Template search

We built three template databases for CASP15: PDB70 and PDB100, and the DistillPDB. The templates in PDB70 and PDB100 were collected from PDB (released before April 2022) and were filtered based on the sequence identity of 100% and 70%, respectively. The templates in DistillPDB were the predicted decoy structures by AlphaFold2 with pLDDT no less than 90.

We employ two strategies to search templates: 1) using hhsearch to search templates based on MSA; 2) using the predicted model of highest pLDDT as the query structure to search templates by TAlign and generating sequence alignment between the query sequence and templates by DeepAlign.

In-house attention-based methods

The overall architecture of our in-house attention-based methods is a modified version of AlphaFold2 architecture. One difference is that we use a linear layer to integrate the scalar, point, and pair attention values in the IPA model while AlphaFold2 uses only addition. We modified the AlphaFold2 feature module and trained four methods with different feature combinations: 1) MSA, 2) MSA & Template, 3) MSA & MSATransformer & Template and 4) MSA & MSATransformer & Template & AlphaFold2 predicted model. In addition to the MSA based methods, we trained a single sequence method which only uses the query sequence and the sequence representation from the protein language model (ESM-1b) as input.

Repeatedly search new templates using predicted model

We use the predicted model of highest pLDDT as the query structure to search for new templates by TMalign and generate sequence alignment between the query sequence and templates by DeepAlign, then run the template based predictions using the new templates. This procedure may repeat several rounds until we could not improve the pLDDT of predicted models anymore.

Model selection

The predicted models are sorted based on the pLDDT and the top 5 models were submitted. To improve the diversity of models, we cluster all predicted models and select the model at cluster center (based on TMscore) as one of the top 5 models.

1. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology* **13**, e1005324 (2017).
2. Xu, J. Distance-based protein folding powered by deep learning. *PNAS* **116**, 16856–16865 (2019).
3. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences* **117**, 1496–1503 (2020).
4. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
5. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
6. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* **2**, 100123 (2022).
7. Levy Karin, E., Mirdita, M. & Söding, J. MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**, 48 (2020).
8. Alexander, H. *et al.* Eukaryotic genomes from a global metagenomic dataset illuminate trophic modes and biogeography of ocean plankton. Preprint at bioRxiv <https://doi.org/10.1101/2021.07.25.453713> (2021).
9. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
10. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098–1109 (2021).
11. Chen, I-Min A., *et al.* "IMG/M: integrated genome and metagenome comparative data analysis system." *Nucleic acids research* (2016): gkw929.

***Ab initio* RNA structure prediction with deep end-to-end potential**Yang Li^{1,2}, Chengxin Zhang^{1,3}, Chenjie Feng¹, Peter L. Freddolino^{1,2}, and Yang Zhang^{1,2}*1 - Department of Computational Medicine and Bioinformatics, University of Michigan Medical School;**2 - Department of Biological Chemistry, University of Michigan Medical School;**3 - Department of Molecular, Cellular, and Developmental Biology, Yale University*

liyangum@umich.edu

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:N*

Non-coding RNAs are fundamental in living cells, performing specific functions determined by their tertiary architectures. Currently, RNA structures are mostly understood and computationally approached using RNA secondary structures, which are often predicted with high accuracy. On the contrary, obtaining high-resolution RNA tertiary structure predictions from nucleotide sequences is still a challenging task. We describe a novel method, rDP, for prediction of RNA tertiary structure using single sequence information.

Methods

The input of rDP is the RNA sequence which is represented by a 5-D one-hot encoding, including 4 types of nucleotides and an unknown state. Based on the sequence, the secondary structure is predicted, in the form of a binary matrix where each bit is set to 1 if the corresponding residue pair forms a base pair. The query sequence and predicted secondary structure are fed into an embedding layer which outputs the sequence and the pair representations. Next, the embedded representations go through 48 RNA transformer blocks, following the successful design of Evoformer in AlphaFold2¹. The transformer blocks output nucleotide-wise rotation matrices and translation vectors applied on a coarse-grained representation of nucleotides represented by three atom types (P, C4', and the glycosidic N atom of the nucleobase). Considering the higher flexibility of RNA structures compared to that of proteins, we construct reference frames with SVD orthogonalization², instead of Gram-Schmidt orthogonalization as used in Alphafold2. The full model is trained end-to-end. Two types of loss functions, i.e., the main Frame Aligned Point Error (FAPE) loss and the inter-N atom distance loss, are used when training the end-to-end models. Additionally, the pair representations are also used for RNA inter-nucleotide distance and orientation prediction with the supervision of negative log likelihood loss function.

The predicted frames and geometries are then integrated as potentials for RNA structure optimization. The conformations predicted by end-to-end models were also used as initial structures of the optimization system and separately optimized by the same hybrid potential function. The gradient of parameters with respect to the hybrid potential function can be

calculated by the automatic differentiation package in PyTorch. With the energy value and the gradient, we can use the L-BFGS algorithm to iteratively update the parameters of the system, i.e., nucleotide-wise rotation matrices and translation vectors. The conformation with lowest energy is considered as the final predicted structure, among the 6 models from different initial conditions.

Availability

The web server is available at <https://zhanggroup.org/DRfold/>

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583-589. 10.1038/s41586-021-03819-2.
2. Aiken, J.G., Erdos, J.A., and Goldstein, J.A. (1980). On Löwdin orthogonalization. *International Journal of Quantum Chemistry* 18, 1101-1108.

RNAComposer-based modeling of RNA 3D structures in CASP15

M. Popenda², J. Sarzynska², M. Antczak^{1,2}, M. Szachniuk^{1,2}

1- Institute of Computing Science, Poznan University of Technology, 2- Institute of Bioorganic Chemistry, Polish Academy of Sciences

mszachniuk@cs.put.poznan.pl

In CASP15, our group participated in RNA tertiary structure modeling category.

Methods

The general workflow of the 3D structure prediction pipeline consists of three major stages: (1) RNA secondary structure prediction, (2) building of an ensemble of RNA 3D structures, and (3) selecting the final submissions.

RNA secondary structure prediction. We predicted RNA secondary structures using computational tools incorporated into the computational engine of the RNAComposer system^{1,2} or based on the consensus structures obtained from the literature or Rfam database³. In the case of synthetic targets such as R1128 and R1138, we manually modeled pseudoknots following the basic principles of RNA origami.

RNA 3D structure construction. RNA 3D models were generated using the RNAComposer system in either automated or expert modeling manner. In the latter case, we introduced structural elements^{1,2} selected from the RNA FRABASE 2.0 repository⁴. We filtered out identified 3D structure elements of low sequence homology, clustered the remaining ones, and selected best candidates for modeling.

In the case of R1126, we manually constructed a structural element for a branched Kissing Loop (bKL) based on the principles published in the literature⁵.

For R1126 and R1136, we have identified structural elements binding the ligands, including quadruplexes. These elements served as the user-provided blocks in expert modeling scenarios available in RNAComposer^{6,7}.

In the case of R1117, we manually constructed the model using structural elements extracted from the experimentally determined RNA 3D structures for class I preQ1 riboswitch, type 1, and 2⁸. Finally, we applied RNAComposer to refine the prototype 3D models.

Model selection. Obtained RNA 3D models were sorted by the total energy coefficient computed by XPLOR⁹. For the subsequent analysis, we selected promising models in which the total energy is below the threshold, i.e., -20kcal/mol per residue^{1,2}.

Next, the promising models were processed by RNAspider¹⁰ to ensure that they did not include incorrect entanglements of structural elements¹¹. Entangled models were rejected. In some prediction challenges, we performed additional refinement using RNAComposer. When selecting predictions to submit, we verified the post-refinement total energy coefficient and the

presence of non-canonical interactions. The latter was done using RNAPdbec^{12,13,14}. Finally, we performed the RMSD-wise clustering with the OC program¹⁵ and selected centroids of the groups as our submissions that show consistency with the found literature data.

Availability:

The methods developed in our laboratory are available at <https://rnapolis.pl/>.

1. Popena, Mariusz, et al. "Automated 3D structure composition for large RNAs" *Nucleic Acids Research* 40.14 (2012): e112-e112.
2. Antczak, Maciej, et al. "New functionality of RNAComposer: an application to shape the axis of miR160 precursor structure" *Acta Biochimica Polonica* 63.4 (2016): 737-744.
3. Kalvari, Ioanna, et al. "Rfam 14: expanded coverage of metagenomic, viral and microRNA families" *Nucleic Acids Research* 49.D1 (2021): D192-D200.
4. Popena, Mariusz, et al. "RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures" *BMC Bioinformatics* 11.1 (2010): 1-12.
5. Liu, Di, et al. "Branched kissing loops for the construction of diverse RNA homooligomeric nanostructures" *Nature Chemistry* 12.3 (2020): 249-259.
6. Jeng, Sunny CY, et al. "Fluorogenic aptamers resolve the flexibility of RNA junctions using orientation-dependent FRET" *RNA* 27.4 (2021): 433-444.
7. Huang, Kaiyi, et al. "Structure-based investigation of fluorogenic Pepper aptamer" *Nature Chemical Biology* 17.12 (2021): 1289-1295.
8. McCown, Phillip J., et al. "Structural, functional, and taxonomic diversity of three preQ1 riboswitch classes" *Chemistry & Biology* 21.7 (2014): 880-889.
9. Schwieters, Charles D., et al. "The Xplor-NIH NMR molecular structure determination package" *Journal of Magnetic Resonance* 160.1 (2003): 65-73.
10. Luwanski, Kamil, et al. "RNAspider: a webserver to analyze entanglements in RNA 3D structures" *Nucleic Acids Research* 50.W1 (2022): W663-W669.
11. Popena, Mariusz, et al. "Entanglements of structure elements revealed in RNA 3D models" *Nucleic Acids Research* 49.17 (2021): 9625-9632.
12. Antczak, Maciej, et al. "New algorithms to represent complex pseudoknotted RNA structures in dot-bracket notation" *Bioinformatics* 34.8 (2018): 1304-1312.
13. Zok, Tomasz, et al. "RNAPdbec 2.0: multifunctional tool for RNA structure annotation" *Nucleic Acids Research* 46.W1 (2018): W30-W35.
14. Antczak, Maciej, et al. "RNAPdbec - a webserver to derive secondary structures from pdb files of knotted and unknotted RNAs" *Nucleic Acids Research* 42.W1 (2014): W368-W372.
15. Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise" *KDD Proceedings* 96.34 (1996): 226-231.

Rookie

Sen Wei

wes_study@163.com

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:N; Fragm:Y; Cont:Y; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

During the CASP, I use RNAstructure¹, MXFold2², etc, to predict the secondary structure of target RNA. Then I use FARFAR2³ and 3dRNA⁴ to predict the tertiary structure. Finally, I use self-trained ARES⁵ to ranking the prediction, and then choose 5 different predictions from the best 20-50 predictions.

1. <https://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict1/Predict1.html>
2. <http://www.dna.bio.keio.ac.jp/mxfold2/>
3. Wang J, Wang J, Huang Y, Xiao Y. 3dRNA v2.0: An Updated Web Server for RNA 3D Structure Prediction. *Int J Mol Sci.* 2019;20(17):4116. Published 2019 Aug 23. doi:10.3390/ijms20174116
4. Watkins, A. M.; Rangan, R.; Das, R. "FARFAR2: Improved de novo Rosetta prediction of complex global RNA folds." *Structure*, 2020, 28: 963-976.
5. Townshend, R. J. L., et al. (2021). "Geometric deep learning of RNA structure." *Science* 373(6558): 1047-1051.

Selbstaufsicht - Pre-Trained Transformer Models for RNA Contact Prediction

O. Taubert¹, C. Faber², A. Bazarova², F. v.d.Lehr³, P. Knechtges³, S. Kesselheim²,
A. Basermann³, M. Götz¹, A. Streit¹, and A. Schug²

*1 – Karlsruhe Institute of Technology, 2 – Forschungszentrum Jülich, 3 – Deutsches Zentrum für Luft und
Raumfahrt*

schug@kit.edu

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:Y.MetaG Y.; Fragm:N; Cont:Y; Dist:N; Tors:N;
DeepL:Y; EMA:N; MD:Y*

For our CASP15 contribution to RNA structure prediction we use a four step workflow: building an MSA, unsupervised deep learning from MSAs, contact prediction with traditional ML, and finally: structure prediction using Monte Carlo simulations.

We applied deep transformer models combined with gradient-boosted decision trees and replica exchange Monte Carlo simulation to RNA structure prediction. Starting with a multiple sequence alignment we produce attention maps with a pre-trained deep transformer model. These serve as input for gradient boosted decision trees that predict a binary contact map. The contact map is then used to bias a SimRNA simulation.

Methods

Selbstaufsicht is based on the MSA transformer models¹ used for unsupervised protein contact prediction. We pre-train the transformer on MSAs from the Rfam² database using a hidden language modeling or inpainting task. In a second training stage we fit a supervised logistic regression contact predictor using the latent attention maps from the transformer as input. The labels for this stage are the same as described in CoCoNet³. Unlike the MSA transformer we do not freeze the transformer parameters during this process. Instead we use aggressive early stopping to manage overfitting. We do not use this regression model for the final contact predictions. Instead we feed the latent attention maps of this fine-tuned transformer into an XGBoost⁴ model trained on the same labels.

For the CASP15 submissions we first built MSAs. For that we submitted queries to RNA Central⁵ and BLAST⁶ databases. Further we used the ClustalW⁷ and Infernal⁸ analysis tools for alignment. For sequences longer than 300 bp we heuristically split the target sequence into several overlapping fragments according to their secondary structure and created MSAs for each one of those. To bias the structure prediction we implement the L most confidently predicted contacts into the coarse-grained simulation SimRNA⁹, where L is the sequence length. The contact biases add one-well and two-slope potentials to the energy function. We clustered the

lowest energy configurations from ten replicas and submitted a representative of the largest cluster.

Availability

Selbstaufsicht is being prepared for publication. Upon publication a repository of the necessary resources for training and inference including source code and datasets will be made available.

1. Rao, R.M., Liu, J., Verkuil, R., Meier, J., Canny, J.F., Abbeel, P., Sercu, T. & Rives, A. (2021). MSA Transformer. *International Conference on Machine Learning*. PMLR, 8844-8856.
2. Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., Rivas, E., Eddy, S.R., Finn, R.D., Bateman, A. & Petrov, A.I. (2020) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*. 49(D1), D192-D200.
3. Zerihun, M.B., Pucci, F. & Schug, A. (2021). CoCoNet – boosting RNA contact prediction by convolutional neural networks. *Nucleic Acids Research* 49(22), 12661-12672.
4. Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
5. RNAcentral Consortium (2021) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D212–D220
6. Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T.L., Agarwala, R., Schäffer, A.A. (2008) Database indexing for production MegaBLAST searches. *Bioinformatics*. 15;24(16):1757-64. Erratum in: *Bioinformatics*. 2008; 24(24):2942. PMID: 18567917; PMCID: PMC2696921.
7. Thompson, J.D., Higgins, D.G., Gibson T.J. CLUSTAL W (1994): improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*. 22(22):4673-80. PMID: 7984417; PMCID: PMC308517.
8. Nawrocki, E. P. and Eddy, S. R. (2013) Infernal 1.1: 100-fold faster RNA homology searches, *Bioinformatics* 29:2933-2935.
9. Boniecki, M.J., Lach, G., Dawson, W.K., Tomala, K., Lukasz, P., Soltysinski, T., Rother, K.M. & Bujnicki, J.M. (2015). SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Research* 44(7). e63

Seder in CASP15

Eshel Faraggi¹, Robert L. Jernigan², Andrzej Kloczkowski³

*Indiana University Purdue University Indianapolis; 2 - Iowa State University; 3- The Steve and Cindy Rasmussen
Institute for Genomic Medicine*

We used the fifth iteration of the Seder server to participate in the CASP15 experiment. The Seder server attempts to predict the TM score of protein models based on the sequence of the protein and the structure of its model using an iterative deep neural network. This cycle of Seder is unique in that it uses a balanced set of hybrid back-propagation/Levenberg-Marquardt and standard back-propagation neural networks. We have found evidence that this approach is more useful for this problem than either one on its own. The hybrid networks are useful in cases where a crucial part of the network is of a limited extent, while other parts may be very large. This is only partially the case here, since there is not enough data to support larger networks. Both networks use associative memory. The hard/easy flavors of the server are distinguished by the training set used for them. For the hard set only proteins with low identity to templates, less than 40% sequence identity as judged by three iterations of PSIPRED.

1. Faraggi, Eshel, and Andrzej Kloczkowski. "A global machine learning based scoring function for protein structure prediction." *Proteins: Structure, Function, and Bioinformatics* 82, no. 5 (2014): 752-759.
2. Faraggi, Eshel, Robert L. Jernigan, and Andrzej Kloczkowski. "A Hybrid Levenberg–Marquardt Algorithm on a Recursive Neural Network for Scoring Protein Models." In *Artificial Neural Networks*, pp. 307-316. Humana, New York, NY, 2021.

Server122-126: Protein tertiary structure prediction by MEGA-Protein in CASP15

Sirui Liu,¹ Jun Zhang,¹ Haotian Chu,⁴ Min Wang,⁴ Mengyun Chen,⁴ Ningxi Ni⁴, Jialiang Yu,⁴ Boxin Xue,^{1,2} Chengwei Zhang,⁵ Dechin Chen,³ Yi Isaac Yang³, Hao Chai,² Yuhao Xie,¹ Zidong Wang,⁴ Lijiang Yang,^{1,2,5} Fan Yu,⁴ Yi Qin Gao^{1,2,3,5}

¹Changping Laboratory; ²Beijing National Laboratory for Molecular Science,; College of Chemistry and Molecular Engineering, Peking University; ³Institute of Systems and Physical Biology, Shenzhen Bay Laboratory; ⁴Huawei Technologies Co., Ltd; ⁵Biomedical Pioneering Innovation Center (BIOPIC), Peking University

liusirui@cpl.ac.cn, gaoyq@pku.edu.cn

Key: Auto:N; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:Y.

Deep learning (DL) methods like AlphaFold2¹ has been successfully applied in the last CASP to predict highly accurate protein structure in most cases, but these methods still rely heavily on co-evolution information(MSA) or template information. In CASP15, we sought to further improve performance of DL methods and extend their application. We participated the protein tertiary structure track and developed a more accurate and efficient end-to-end protein structure prediction toolkit called MEGA-Protein. This toolkit mainly consists of three parts: protein structure prediction tool MEGA-Fold², MSA generation tool MEGA-EvoGen³ and protein structure assessment tool MEGA-Assessment. We used five different settings to predict tertiary structure and submitted results to 5 servers separately(named by server 122/123/124/125/126 respectively) .

Methods

Given a query sequence, we prepared templates and four different MSAs as input. Then we adopted MEGA-Protein to predict and selected the best models according to different criteria. Models selected were refined by OpenMM⁴ relaxation.

Input data

Given a query sequence, we prepared co-evolution information(MSA) in 4 different ways, 3 with traditional database query and 4th with our MSA Generation tool MEGA-EvoGen: MSA 1 was searched using MMseqs2 (with default settings from ColabFold⁵). MSA 2 was generated by filtering MSA 1 by HHfilter⁶ (with 95 percent identity and 50 percent coverage filters). For MSA 3, we searched MSA by Jackhmmmer and HHblits following standard AF2 protocol, whose result was then combined with MSA 1 and as MSA 3. For MSA 4, the MEGA-EvoGen was applied to improve the quality of MSA 3, and 8 groups of MSA were generated using different hyper-parameters. All the 4 MSAs were fed into HHSearch individually to generate templates.

Protein Structure Prediction

Server 122 For this server, we used MEGA-Fold to infer structures with checkpoint trained on our own dataset. We used the MSA 1/2/3 + template as input. For each query sequence, multiple models together with their confidence were predicted. We submitted the model with highest confidence as MODEL 1 of server 122. We then selected four most diversified models from the remaining models(confidence higher than 0.7) according to their mutual IDDT and TM-score and submitted as MODEL 2/3/4/5.

Server 123 For this server, we adopted MSA generation with MEGA-EvoGen as input and still used MEGA-Fold for inference. The selection method was the same as server 122 except that results of server122 and server123 were merged together as the candidate pool.

Server 124 For this server, first, we replaced one of the templates in MEGA-Fold with the structures predicted by server 122, while keeping other inputs the same as server 122, called server 122'. we then selected four best models from server 122 and the server 122' by AF2Rank⁷, These models were submitted as MODEL 1/2/3/4 respectively. Then we selected the best model of server 122' according to pLDDT confidence as MODEL 5.

Server 125 The protocol of this server was the same as server 122, except that homo-oligomer was processed in multimer-like protocol: the MSA for single query sequence was repeated and placed block diagonal-wisely to simulate multimeric MSA. the query sequence was also repeated corresponding times in the predicted structure. We predicted confidence for each copy separately and saved the copy with highest confidence as the resulting model. We selected the most confident model from all resulting models predicted and submitted it as MODEL 5 of server 125, while keeping MODEL 1/2/3/4 the same as server 122.

Server 126 The model selection protocol was the same as server 122, but it took all unique models from server 122 and 124 (also unique models in server 123 when MSA depth <128) into consideration.

Manual interventions

Most query sequences were processed with an automatic pipeline, except for 3000+ length targets T1165 and T1169, due to the memory limitation of hardware (Ascend910-32GB). For T1165, we predicted all models with our own checkpoint from MEGA-fold with jax on GPU A100-80GB. For T1169, we predicted models with our own checkpoint with jax on GPU A100-80GB for server 122/123/124/126, and standard AlphaFold2 for server 125. We are working on improving memory efficiency on Ascend.

Availability

The proposed method MEGA-Protein is developed based on [MindSPONGE](#) computational biology/chemistry package and [MindSpore](#) AI framework. The MEGA-Protein package is available at our [gitee](#) or [github](#) page.

1. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596(7873): 583-589.

2. Liu S, Zhang J, Chu H, et al. PSP: million-level protein sequence dataset for protein structure prediction. arXiv preprint arXiv:2206.12240, 2022.
3. Zhang J, Liu S, Chen M, et al. Few-shot learning of accurate folding landscape for protein structure prediction. arXiv preprint arXiv:2208.09652, 2022.
4. Eastman P, Swails J, Chodera JD, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comp. Biol.*, 13(7):1–17, 2017.
5. Mirdita, M., Schütze, K., Moriwaki, Y. et al. ColabFold: making protein folding accessible to all. *Nat Methods*, 19: 679–682, 2022.
6. Remmert M, Biegert A, Hauser A, et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, 9(2): 173-175, 2012.
7. Roney J P, Ovchinnikov S. State-of-the-Art estimation of protein model accuracy using AlphaFold. *BioRxiv*: 2022.03.11.484043, 2022.

Improved MSA for Better Protein Structure Prediction

Huiyu Chen¹², Wentong Wang¹², Liting Zeng¹², Shuyi Yang¹², Kexin Yu¹², Wei Han¹²,
Suwen Zhao¹²

1 - iHuman Institute, ShanghaiTech University, Shanghai 201210, China., 2 - School of Life Science and Technology,
ShanghaiTech University, Shanghai 201210, China.

zhaosw@shanghaitech.edu.cn

Key: Auto:N; CASP_serv:Y; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y;
DeepL:Y; EMA:Y; MD:N

The approach to build up MSA in Alphafold2 has room to be improved. In our method, we attempt to predict better structures by improving the quality of multiple sequence alignment (MSA). For each target, we used several best available distant homolog searching methods such as Hhblits, jackhmmer and MMseqs2 to search a wide range of sequence database at the same time to form multiple MSAs. Then we filtered MSAs according to coverage and identity. When the depth of MSAs are not enough, we tried to merged some/all generated MSAs. We further filtered the MSAs through MSA plots, only MSAs passed the previous filters were used as the input of Alphafold2 framework. Finally, pLDDT was used to rank all models.

Methods

For each target, we used the same prediction method. Our prediction method of protein structures is based on Alphafold2¹. Everything is the same as the Alphafold2 except for the MSA generation section. PDB70 was used as the template database, and AMBER² was used for refinement after model generation. The deep learning part is exactly the same as Alphafold2, without any changes. The pLDDT was used to rank all models. Instead of targeting specific sets of proteins, our approach works on all proteins.

For MSA search on BFD, Uniclust30 and MetaShanghaiTech (a homemade metagenome database), we used HHblits from hh-suite v.3.0-beta.3 release 14/07/2017³. For MSA search on Uniref90, MGnify, IMG/VR3⁴ (a virus metagenome database), PSDB (a homemade virus metagenome database), we used jackhmmer from HMMER v.3.3⁵. For Uniref30 and ColabFoldDB, we used MMseqs2^{6,7}. We will generate a plot for each MSA, which includes the number of homologs, as well as the identity and coverage compared to the query. Based on the plot, we then decided which MSAs to filter on identity and coverage, and/or to combine with other good MSAs after filtering and realign to get some new MSAs. For realignment, we used Kalign2⁸. The identity and coverage when filtering, and the gap open penalty and gap extension penalty when realigning, were decided by us based on the actual situation of each target. All the good MSAs would be sent to Alphafold2 framework for model prediction.

Results

Based on the pLDDT, more than half of the results are better than the CASP-Hosted Server Predictions. By observing the results of MSA plot and pLDDT, we found that using MMseqs2 to search Uniref30 or ColabFoldDB database with reference to Colabfold often had better results.

Availability

The parameters of several methods for searching homologs are detailed in their respective software. The homologs databases: Uniref90, https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2020_01/uniref/; MGnify, http://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2019_05/; Uniclust30, <https://uniclust.mmseqs.com/>; BFD, <https://bfd.mmseqs.com/>; Uniref30, <https://uniclust.mmseqs.com/>; colabfoldDB, <https://colabfold.mmseqs.com/>; IMGVR3, https://genome.jgi.doe.gov/portal/IMG_VR/IMG_VR.home.html; The homemade database PSDB, MetaShanghaiTech and steps for filtering and merging MSAs are not available yet. The template database: PDB70, https://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/. The MSA plot code and pLDDT code refers to the corresponding section of ColabFold python code, <https://github.com/sokrypton/ColabFold/blob/main/beta/colabfold.py>. Except for the MSA generation section, please refer to the Alphafold2 code for other parts (<https://github.com/deepmind/alphafold>).

3. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Židek,A., Potapenko,A., Bridgland,A., Meyer,C., Kohl,S.A.A., Ballard,A.J., Cowie,A., Romera-Paredes,B., Nikolov,S., Jain,R., Adler,J., Back,T., Petersen,S., Reiman,D., Clancy,E., Zielinski,M., Steinegger,M., Pacholska,M., Berghammer,T., Bodenstein,S., Silver,D., Vinyals,O., Senior,A.W., Kavukcuoglu,K., Kohli,P., Hassabis,D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*. **596(7873)**, 583-589.
4. Case,D.A., Cheatham,T.E.3rd, Darden,T., Gohlke,H., Luo,R., Merz,K.M.Jr., Onufriev,A., Simmerling,C., Wang,B., Woods,R.J. (2005). The Amber biomolecular simulation programs. *J Comput Chem*. **26(16)**, 1668-88.
5. Remmert,M., Biegert,A., Hauser,A., Söding,J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods*. **9(2)**, 173-5.
6. Roux,S., Pérez-Espino,D., Chen,I.A., Palaniappan,K., Ratner,A., Chu,K., Reddy,T.B.K., Nayfach,S., Schulz,F., Call,L., Neches,R.Y., Woyke,T., Ivanova,N.N., Elie-Fadroshe,E.A., Kyrpides,N.C.(2021). IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res*. **49(D1)**, D764-D775.
7. Johnson,L.S., Eddy,S.R., Portugaly,E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*. **11**, 431.
8. Mirdita,M., Schütze,K., Moriwaki,Y., Heo,L., Ovchinnikov,S., Steinegger,M. (2022). ColabFold: making protein folding accessible to all. *Nat Methods*. **19(6)**, 679-682.
9. Steinegger,M., Söding,J.. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. **35(11)**, 1026-1028.

10. Lassmann,T., Sonnhammer,E.L. (2005). Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*. **6**, 298.

ShanghaitechFold: Hybrid MSA Embedder Model

Anqi Pang, Kexin Zhang, Zhigang Sun and Jingyi Yu

ShanghaiTech University

pangaq@shanghaitech.edu.cn

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

ShanghaiTechFold is a hybrid multi-model deep learning network. The whole architecture includes Data Processing, Feature Embedding, Evoformer, Structure Module and Structure Refinement. To obtain diverse multi-sequence alignments (MSA) and get MSA features, we build a large metagenomic database. According to different training configurations, a total of 15 sets of models were obtained and we get hundreds of predictions with different combinations of models and MSA searching/generation methods. Finally, we leverage averaged predicted local distance distribution test (pLDDT) value to rank multiple predictions.

Methods

The whole architecture of ShanghaiTechFold includes Data Processing, Feature Embedding, Evoformer, Structure Module and Structure Refinement.

In Data Processing part, we downloaded hundreds of TB sequence data from NCBI and Mgnify and used plasm to assemble them. Then, we built our customized datasets by a similar pipeline of BFD. Because of the limitation of memory and speed, we split assembled proteins into many small parts and used GPU-accelerated techniques to build large MSA databases that can be searched by hhblits. These databases are about 20x larger than the BFD database so we can get enough homologs for each query sequence. Also, we tried different operations to search MSA, such as filtering by coverage, combination and realignment, and random mutations. For each target, we can get many kinds of MSA that are used in the Naive MSA Embedder of Feature Embedding part.

In Feature Embedding part, we construct representations from Residue Embedder and MSA Embedder. For MSA Embedder, we combined three different Embedders to get the final MSA features including Naive MSA Embedder, Pseudo MSA Embedder, and protein Language Model (pLM) Embedder. In Naive MSA Embedder, we search MSA from metagenomic data. In Pseudo MSA Embedder, we first generate pseudo MSA from pre-trained MSA generative model, then we use the same pipeline as Naive MSA Embedder. In pLM Embedder, we output MSA features directly. By different combinations of the above MSA Embedders, we got 15 kinds of MSA features. Finally, we combine Residue Embedder features and MSA features to build MSA representations and pair representations.

In Evoformer and Structure Module part, we used the same architecture as alphafold2¹. For all self-attention layers, we used dynamic axial parallelism technique to save GPU memory and accelerate forwarding and backpropagation speed. In structure refinement, we used *OPENMM*² with CUDA platform. We got several hundred models for each target with different MSA features and models and rank them by averaged pLDDT value, then we select top 5.

Results

For most of CASP targets, we can get reliable predictions based on pLDDT values.

1. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
2. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang LP, Simmonett AC, Harrigan MP, Stern CD, Wiewiora RP, Brooks BR, Pande VS. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol*. 2017 Jul 26;13(7):e1005659.

SHORTLE

Refinement of Protein Models via Fragment Replacement To Improve Local Energies

D.R. Shortle

Thr Johns Hopkins University School of Medicine

dshortl1@jhmi.edu

Key: *Auto:N; CASP_serv:Y; Templ:N; MSA:N.MetaG; Fragm:Y.3-5; ContN; Dist:N; Tors:Y; DeepL:N; EMA:Y; MD:N*

Although models submitted by BAKER_SERVER and Alpha2 have very little atom-atom overlap between residues separated by more than 5 amino acids, the energy of interactions for nearby residues frequently appears to be higher than that expected from high resolution x-ray structures. The sum of overlap between atoms I to i+1 and atoms I to i+2-5 is often significantly larger and the statistics of phi, psi, and chi1 angles in monomeric, dimeric and trimeric segments are often outside the ranges found in real proteins. Disregard of these local energy terms may lead, all things being equal, to small errors in chain direction, potentially allowing conformations which are energetically forbidden to the real protein. In other words, very low global energy without optimization of these local terms is not necessarily an indication of successful refinement of a model.

Method

Our group has spent the past 20 years working on fragment replacement methods using a genetic algorithm and Monte Carlo search strategies that focus on improving these local energies. Fragments with roughly the same secondary structure are selected from the PDB based on scoring with multiple Ramachandran statistical potentials and absence of excessive atom overlap after side-chain replacement. At this stage, all other energy terms are ignored. After randomly joining these fragments of length 3 to 5 residues to generate several thousand crude, full length models or “decoys”, one round of 4 generations using the genetic algorithm is applied to sets of 25 of these decoys, with the full set serving as a library of varying length segments. Heavy selective pressure for reduced local atom overlap and improved Ramachandran energies are the major terms emphasized, with only modest pressure applied to reduce backbone overlap between residues separated by 10 and 30 percent of full length.

In the first round of model refinement, a combined pool of all AF2 and all BAKER_SERVER models serves as a starting population and large segments of these models are forcibly replaced by segments from refined decoys. Subsequently, extensive fragment replacement proceeds using fragments randomly drawn from all the refined decoys, selecting primarily for local energy terms. No direct structural information (CA-CA distance matrix error) from the guide model was needed to generate models converging toward the guide model.

In the subsequent 2 to 5 genetic refinement rounds (4 generations each), scoring for improved global energies was slowly added to the survival function, along with pressure from the CA-CA distance matrix calculated over increasing residue separations. To reduce the loss of structural diversity, which is a major problem for genetic algorithms in general, the principle tactics employed were: (1) selection with replacement rather than children competing with parent (i.e., no expansion of the population); (2) one function selects for successful fragment swaps plus a different survival function that picks one structure generated during the second half of the trajectory to replace the starting structure; (3) Alternating the selection function and survival functions between the weighted sum of 5-12 composite pseudo-energy terms versus the weighted sum of the z-scores of these same terms.

As refinement progresses, more emphasis is given to the standard global energy terms, such as atom-atom statistical potentials, solvation, and hydrogen bonds and relatively less to local energetics. In parallel, pressures for native-like packing density and atom-atom separation statistics are also applied. Our experience is that conventional statistical potentials for atom-atom interactions do not reproduce the statistics of atom-atom distances observed in high resolution x-ray models.

Results

In summary, the refinement method described above was used for all attempts at model refinement. Inspection indicated that the best model seldom deviated by more than 1 Å in CA RMSD from the BAKER_SERVER_TS1 model. This may or may not indicate serious limitations in the conformational search, which possibly only accesses a subset of new structures very similar to the starting models.

Availability

The author will provide information or code from all reasonable requests.

Prediction of tertiary and quaternary structures of biomacromolecules by integrating evolutionary information and energy functions

Lin Wang¹, Fenglei Li^{1,2}, Shihang Wang¹, Yongqi Zhou¹, Shiwei Li¹, Siyuan Tian¹,
Xinyue Ma^{1,2}, Yihao Zhi², Qiaoyu Hu¹, Xianglei Zhang¹, Shenghua Gao², Fang Bai^{1,2}

1 - Shanghai Institute for Advanced Immunochemical Studies, ShanghaiTech University, 393 Middle Huaxia Road, Shanghai, 201210, China, 2 - Information Science and Technology, ShanghaiTech University, 393 Middle Huaxia Road, Shanghai, 201210, China

wanglin3@shanghaitech.edu.cn

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA: Y.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:Y*

AlphaFold2 is changing the paradigm of computational biology research by mining protein sequences from genetic databases and using the evolutionary information between sequences to model protein tertiary structures¹. It provides a great initial structure for many classical computational methods that based on force fields or energy functions, such as molecular dynamics (MD) simulations, Monte Carlo simulations, and molecular docking. In CASP15, we integrate the above simulation methods with AlphaFold2 to explore their potential in accurately predicting the tertiary and quaternary structures of biomacromolecules.

Methods

The AlphaFold methods^{1,2} were used to predict the protein structures for all targets, including those from ColabFold, local-deployed AlphaFold, and structures downloaded from the Elofsson lab. The prediction models were ranked by pLDDT, model confidence, and structural diversity and the top models were selected for further optimization.

For regular targets, the conformation of side chains was predicted by sampling orientations and global energy minimization was conducted by applying Prime protein structure refinement of Schrödinger³. Combining the per-residue pLDDT scores provided by AlphaFold, the protein reliability report generated by Maestro GUI of Schrödinger packages, and the expert experiences, the quality of the structures was evaluated and the rank of models was adjusted. The protein reliability report contained the rationalization of backbone and side chain dihedral angles, polar encapsulation and solvent exposure. If needed, DeepRefiner⁴ or MD simulation were used for further optimization, especially for some flexible fragments of proteins.

For the ligand binding prediction targets, we first predicted 3D structure of target sequence using AlphaFold^{1,2} and then searched for homologous templates in the Protein Data Bank (PDB) to determine the binding pockets of ligands, metal ions and co-factors. After that, the ligands were docked to the ligand-binding pocket on the target protein by Glide⁵ program of Schrödinger packages.

For assembly targets, protein complex structures were predicted using AlphaFold-Multimer² to generate different conformations, then we utilized a Monte Carlo simulation of Prime module³ of Schrödinger packages or a relax⁶ program of Rosetta packages to optimize the backbone structure and refine the side chain conformations. For each conformation, the interface Molecular Mechanics/Generalized Born Surface Area (MMGBSA) binding free energy was calculated by the prime module³ of Schrödinger, and a biological dimer reliability score was estimated by ClusPro-DC⁷. Taking both MMGBSA binding free energy and ClusPro-DC reliability score into consideration, we ranked the conformations of protein-protein complexes.

In particular, for antigen-antibody targets, we applied two methods to build the antibody structures: DeepAb⁸, a deep learning-based method; and Antibody Structure Prediction⁹ of Schrödinger packages, a method derived from homology modeling and loop modeling. Because of the high identity of antigens from CASP15 with current solved proteins, the structures of antigens were built by Advanced Homology Modeling of Schrödinger packages. Subsequently, the global docking using ClusPro^{10,11} in antibody mode or HADDOCK¹², was performed to generate antigen-antibody binding modes. As with other assembled targets, antigen-antibody complex models were ranked using MMGBSA, ClusPro-DC reliability score, and expert experiences.

RNA targets modeling involves four steps: secondary structure prediction, 3D structure modeling, ARES scoring, and energy minimization. Specifically, a rough secondary structure was predicted by RNAfold^{13,14} and MXfold¹⁵. All the stem-loop structures in the predicted secondary structure were then used for the template searching in the nucleic acid database¹⁶. The obtained template structures, together with the predicted second structures, were utilized for RNA 3D structure modeling using FARFAR2¹⁷. All the predicted structures were subsequently scored by ARES¹⁸. The top 20 structures were refined by Prime module^{3,6} of Schrödinger packages. In addition, for the RNA-Ligand system, the template structure with the same ligand was used for RNA modeling. Glide docking¹⁹ embedded in Schrödinger was then engaged in adjusting the ligand poses. And we adopted energy minimization embedded in Schrödinger 2021-1 to refine the final structure. For the RNA-Protein complex, the protein dimer or trimer structures were modeled by AlphaFold¹, and the RNA structure was constructed as previously described. The relative positions between RNA and protein were determined by homologous templates.

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583-589.
2. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Židek, A., Bates, R., Blackwell, S., Yim, J. *et al.* (2022) Protein complex prediction with AlphaFold-Multimer. *BioRxiv*.
3. Jacobson, M.P., Pincus, D.L., Rapp, C.S., Day, T.J.F., Honig, B., Shaw, D.E. and Friesner, R.A. (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins-Structure Function and Bioinformatics*, **55**, 351-367.

4. Shuvo, M.H., Gulfam, M. and Bhattacharya, D. (2021) DeepRefiner: high-accuracy protein structure refinement by deep network calibration. *Nucleic Acids Research*, **49**, W147-W152.
5. Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelley, M., Perry, J.K. *et al.* (2004) Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, **47**, 1739-1749.
6. Conway, P., Tyka, M.D., DiMaio, F., Kondering, D.E. and Baker, D. (2014) Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci*, **23**, 47-55.
7. Yueh, C., Hall, D.R., Xia, B., Padhorny, D., Kozakov, D. and Vajda, S. (2017) ClusPro-DC: Dimer Classification by the Cluspro Server for Protein-Protein Docking. *J Mol Biol*, **429**, 372-381.
8. Ruffolo, J.A., Sulam, J. and Gray, J.J. (2022) Antibody structure prediction using interpretable deep learning. *Patterns*, **3**, 100406.
9. Zhu, K., Day, T., Warshaviak, D., Murrett, C., Friesner, R. and Pearlman, D. (2014) Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins: Structure, Function, and Bioinformatics*, **82**, 1646-1655.
10. Kozakov, D., Hall, D.R., Xia, B., Porter, K.A., Padhorny, D., Yueh, C., Beglov, D. and Vajda, S. (2017) The ClusPro web server for protein–protein docking. *Nature protocols*, **12**, 255-278.
11. Brenke, R., Hall, D.R., Chuang, G.-Y., Comeau, S.R., Bohnuud, T., Beglov, D., Schueler-Furman, O., Vajda, S. and Kozakov, D. (2012) Application of asymmetric statistical potentials to antibody–protein docking. *Bioinformatics*, **28**, 2608-2614.
12. Van Zundert, G., Rodrigues, J., Trellet, M., Schmitz, C., Kastiris, P., Karaca, E., Melquiond, A., van Dijk, M., De Vries, S. and Bonvin, A. (2016) The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes. *J Mol Biol*, **428**, 720-725.
13. Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R. and Hofacker, I.L. (2008) The Vienna RNA Websuite. *Nucleic Acids Research*, **36**, W70-W74.
14. Lorenz, R., Bernhart, S.H., Höner Zu Siederdisen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol*, **6**, 26.
15. Akiyama, M., Sato, K. and Sakakibara, Y. (2018) A max-margin training of RNA secondary structure prediction integrated with the thermodynamic model. *J Bioinform Comput Biol*, **16**, 1840025.
16. Berman, H.M., Lawson, C.L. and Schneider, B. (2022) Developing Community Resources for Nucleic Acid Structures. *Life (Basel)*, **12**, 540.
17. Watkins, A.M., Rangan, R. and Das, R. (2020) FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds. *Structure*, **28**, 963-976.
18. Townshend, R.J.L., Eismann, S., Watkins, A.M., Rangan, R., Karelina, M., Das, R. and Dror, R.O. (2021) Geometric deep learning of RNA structure. *Science*, **373**, 1047-1051.
19. Borrelli, K.W., Cossins, B. and Guallar, V. (2010) Exploring hierarchical refinement techniques for induced fit docking with protein and ligand flexibility. *J Comput Chem*, **31**, 1224-1235.

Hierarchically combining multiple methods for 3D RNA structure prediction

Simon Poblete^{1,2}, Horacio V. Guzman^{3,4} and Fabrizio Pucci^{5,6,*}

1 - Instituto de Ciencias Físicas y Matemáticas, Universidad Austral de Chile, Valdivia 5091000, Chile; 2 - Computational Biology Lab, Fundación Ciencia & Vida, Santiago 7780272, Chile; 3 - Department of Theoretical Physics, Jozef Stefan Institute, Ljubljana, Slovenia; 4- Departamento de Física Teórica de la Materia Condensada, Universidad Autónoma de Madrid, E-28049 Madrid, Spain; 5 - Computational Biology and Bioinformatics, Université Libre de Bruxelles, 1050 Brussels, Belgium; 6 - Interuniversity Institute of Bioinformatics in Brussels, 1050 Brussels, Belgium

*Fabrizio.Pucci@ulb.be

Key: Auto:N; CASP_serv:N; Templ:N; MSA:Y; Fragm:Y; Cont:Y; Dist:N; Tors:N; DeepL:N; EMA:N; MD:Y

The accurate prediction of the 3-dimensional RNA structure from the nucleotide sequences is critically important for understanding the key functions of RNA in cellular processes. Here we present the approach we used in the CASP15 experiment, that consists in the integration of different and complementary methodologies, to predict the structure of RNA molecules. We begin by considering the information that is available for the RNA target such as the structure of homology templates, the secondary structure prediction and/or the multiple sequence alignment of homologous RNA sequences. From all this data, we extract spatial constraints which are then enforced in the 3D modeling step. It consists in building three-dimensional coarse-grained structures at nucleotide-level resolution and then refining them towards an all-atom representation. In the last step of our approach, we screen all generated structures and manually select the best five candidates based on energy, symmetry and geometry considerations.

Methods

(A) Collecting RNA data for structural modeling

We consider different sources of information as input of our pipeline. The type of information used clearly depends on the RNA target input:

RNA 2D structure. Given the RNA sequence we predict its secondary structure using the well known tool RNAfold¹. This has been done for all RNA targets in CASP15.

Homology template. In the case in which there is an homologous RNA sequence with a known 3D structure deposited in the PDB², we use the template for extracting the nucleotide-nucleotide distance constraints that are used in the 3D modeling. In this competition, we used PDB 7LYF and 3Q50 for the modeling of the R1116 and R1117 target, respectively, and fragments from PDB 5OB3 and 7EOP for R1136.

Multiple Sequence Alignment. In the case in which the target sequence matches one of the RFAM families³ with a good score, we re-align it using ClustalW⁴ and employ coevolutionary approaches from pydca⁵ to identify nucleotide pairs that coevolve and thus are

likely to be in spatial proximity in the 3D RNA structure. In CASP15, we used the MSA constructed from the CPEB3 ribozyme family (RF00622) as input for targets 1107 and 1108.

Base-pair complementarity. We manually search for kissing loop structures that result from base-pairing between hairpin or internal loops. We use this approach for different synthetic targets in CASP15 (R1126, R1128, R1136, R1138). Note that we also complemented the manual identification of kissing loops by using the IPKnot⁶ package.

All information retrieved in this step was used in the form of spatial constraints to guide the 3D structure modeling.

(B) 3D RNA structure modeling

The tertiary structure modeling was performed using two methods : SimRNA⁷ and SPQR⁸.

SimRNA is a nucleotide-level coarse-grained model of RNA, developed in the group of J. M. Bujnicki which uses Monte Carlo method for sampling the conformational space and a statistical potential to identify the lowest energy conformations. It was employed for exploring the conformation space, clustering the low-energy structures and obtaining a final set of a few hundreds all-atom structures.

SPQR is a nucleotide-level coarse-grained model of RNA for building models by fragment assembly as done in larger structures⁹, and locally exploring the conformation space and minimizing the energy, with structural restraints for secondary structure and tertiary contacts imposed as a harmonic ERMSD restraint⁹. The all-atom structures were generated by placing an atomistic template on each nucleotide, which in some cases was minimized through a short MD simulation, using the parameters of previous works¹⁰.

(C) Structure selection and refinement

In the previous steps we generated a pool of structures from which we selected manually the best five candidates. Note that small variations of the procedure were applied on each puzzle. Here we list the different criteria to select the candidates:

Energy evaluation. From the structures generated via SimRNA, we collect the 1% with the lowest energy and cluster them. Additionally, we employ SPQR for rescoring the SimRNA structures and select either those with the lowest SPQR energy or the most populated SimRNA cluster representative. We use this approach for the large majority of CASP15 targets. For the structures generated with SPQR, we only used the SPQR energy.

Radius of gyration (R_g). We evaluate R_g and use it as an additional criteria for the choice of the final predicted structure. More compact structures are preferred with respect to extended ones.

Symmetric properties. We did a visual inspection of the 3D RNA structure to verify its symmetric properties and use it as another selection criteria.

Finally, the proposed all-atom models chosen are refined using QRNAS¹¹ or MD relaxation¹⁰.

Availability

SPQR code is freely available at <https://github.com/srnas/spqr>.

1. Lorenz et al. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **6**, 26.

2. Berman HM et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**(1):235-42.
3. Kalvari I et al. (2020). Rfam 14: expanded coverage of metagenomic, viral and microRNA families, *Nucleic Acids Res.* **49**(D1), D192–D200.
4. Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**(22):4673-80.
5. Zerihun MB, Pucci F, Peter EK, Schug A (2020). pydca v1.0: a comprehensive software for direct coupling analysis of RNA and protein sequences. *Bioinformatics.* **36**(7): 2264–2265
6. Sato, K., Kato, Y., Hamada, M., Akutsu, T., Asai, K. (2011) IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, **27**(13):i85-i93.
7. Boniecki MJ et al., SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.* **44**(7), e63.
8. Poblete, S., Bottaro, S. & Bussi, G. (2018). A nucleobase-centered coarse-grained representation for structure prediction of RNA motifs. *Nucleic Acids Res.* **46**, 1674-1683.
9. Poblete, S. & Guzman, H. V. (2021). Structural 3d domain reconstruction of the RNA genome from viruses with secondary structure models. *Viruses* **13**, 1555.
10. Poblete, S., Bottaro, S. & Bussi, G. (2018), Effects and limitations of a nucleobase-driven backmapping procedure for nucleic acids using steered molecular dynamics, *Biochem. Biophys. Res. Comm.* **498** (2), 352-358
11. Stasiewicz, J., Mukherjee, S., Nithin, C. et al. (2019). QRNAS: software tool for refinement of nucleic acid structures. *BMC Struct Biol* **19**, 5 (2019).

Spider (assembly)

A Novel Statistical Energy Function and Effective Conformational Search Strategy based Protein Complex Structure Prediction

Md Wasi Ul Kabir¹, Avdesh Mishra², Md Tamjidul Hoque^{1,*}

1 - Computer Science, University of New Orleans, 2000 Lakeshore Drive, New Orleans, LA 70148, USA;

2 - Electrical Engineering and Computer Science, Texas A&M University-Kingsville, Kingsville, TX, USA

* thoque@uno.edu

Key: Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:Y; DeepL:N; EMA:Y; MD:Y

We used our developed protein structure prediction (PSP) method, called 3DIGARS-PSP, for the prediction of protein complex structure (or assembly prediction) in CASP15. 3DIGARS-PSP uses an effective statistical energy function called 3DIGARS and an advanced search algorithm called KGA. We refer to our assembly prediction method as 3DIGARS-PSP-ASSEMBLY. For the conformational sampling of the protein folding process, the 3DIGARS-PSP method uses a memory-assisted genetic algorithm (GA) which is an extension of the KGA. We design GA using two important operators: memory-assisted crossover and mutation. These operators perform the important function of angle rotation and segment translation to support careful sampling. In order to aid in the search process, we also utilize the propensities of secondary structure and torsion angle. The memory-assisted GA-based sampling generates a large-scale ensemble of decoys to minimize the statistical energy function. Finally, we select the top five models for each CASP15 assembly target by clustering the ensemble of decoys, and consequently, these models are submitted to CASP15.

Methods

The assembly targets have multiple subunits in CASP15, each of which has a corresponding fasta sequence. To create a single fasta sequence, we first combine the fasta sequences of the subunits by inserting 20 Glycine (GLY or G) amino acids in between each fasta sequence. Glycine amino acid is used to combine the fasta sequences of the subunits because of its smallest side chain size among 20 standard amino acids. Then we use the AlphaFold2¹ tool to obtain the predicted models using the combined fasta sequence. The prediction of the 3D structure of the assembly target starts by initializing some of the chromosomes of the GA population with the cartesian coordinates of the backbone atoms of the models obtained from AlphaFold2¹. The rest of the chromosomes are filled by single point torsion angle changes (rotation). To make a guided change of the torsion angles (Φ or Ψ), the occurrence frequency of 20 standard amino acids with different Φ - Ψ angle pairs are constructed from the 4,332 high-resolution experimental structures extracted in our previous work². The Φ and Ψ angle range is divided into 120 bins with an interval of 3 degrees, and the frequencies of the bins are updated based on the value of the Φ and Ψ angles of every amino acid in the protein to obtain the frequency of distribution of 20 standard amino acids. By examining the cluster of frequency values, we further classify the frequency distributions into zones. Then, the most probable torsion angle (namely, $p\Phi$ or $p\Psi$) of the zone is

extracted using the roulette wheel selection method, and a random angle around this angle is selected as a new torsion angle.

Moreover, the propensities of secondary structure (SS) types of amino acids are also extracted from the same experimental structures used above by running the DSSP program to guide the torsion angle rotation. The SS types given by DSSP are broadly categorized into four different SS types (H, G, and I = H; E and B = E; T and S = T; and U). The torsion angle pair and SS types of the amino acids in protein are used to obtain the SS distribution. Later, this distribution of SS is used such that the SS type, which has the largest frequency count, is assigned to the given amino acid having a certain Φ - Ψ angle. Furthermore, the Φ - Ψ angle pairs corresponding to the H and E types are grouped into helix and beta groups and are consequently used to update the Φ or Ψ angle that results in a clash within the structure.

The chromosomes (models) for the next generation of GA are obtained by two different types of structural change operators: *i*) angle rotation and *ii*) segment translation. The mutation in GA involves torsion angle rotation, and crossover involves segment translation followed by torsion angle rotation at the crossover point. The torsion angle rotation technique is based on the principle of rotation about an arbitrary axis. On the other hand, crossover in GA performs segment translation where all the amino acid indexes that are not SS type E or B are considered as possible crossover points. This is done to avoid random changes in the beta-sheet region and make more appropriately guided changes during the mutation operation. The children's structures in the crossover process are generated from two parent structures and a structure with the best fitness saved in the memory ³.

The decoys generated by the conformational change through memory-assisted GA guided by the statistical energy function are then converted into the all-atom level by using Oscar-star software ⁴. The large-scale pool of decoys are clustered into five different cluster groups, at least 5Å apart from each other based on the average root-mean-square deviation (RMSD). Then, we select the top five models in different clusters based on the 3DIGARS energy score ranking. The subunits of the top five models are further refined using the ModRefiner ⁵ software. Then, we use the ResQ ⁶ method to add B-factors to the subunits of the top five models. Finally, the models of the subunits are combined together in the CASP15 assembly format before submission.

Availability

Source code, manual, and example data of 3DIGARS-PSP for Linux are freely available, for non-commercial use, at http://cs.uno.edu/~tamjid/Software/ab_initio/v2/PSP.zip.

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583-9.
2. Mishra A, Hoque MT. Three-Dimensional Ideal Gas Reference Sstate Based Energy Function. *Current Bioinformatics*. 2017;12:171-80.
3. Hoque MT, Iqbal S. Genetic algorithm-based improved sampling for protein structure prediction. *International Journal of Bio-Inspired Computation*. 2017;9:129-41.
4. Liang S, Zheng D, Zhang C, Standley DM. Fast and accurate prediction of protein side-chain conformations. *Bioinformatics*. 2011;27:2913-4.
5. Xu D, Zhang Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophysical Journal*. 2011;101:2525-34.

6. Yang J, Wang Y, Zhang Y. ResQ: An approach to unified estimation of B-factor and residue-specific error in protein structure prediction. *Journal of Molecular Biology*. 2016;428:693-701.

Spider (TS)

A Novel Statistical Energy Function and Effective Conformational Search Strategy based *ab initio* Protein Structure Prediction

Md Wasi Ul Kabir¹, Avdesh Mishra², Md Tamjidul Hoque^{1,*}

1 - Computer Science, University of New Orleans, 2000 Lakeshore Drive, New Orleans, LA 70148, USA;

2 - Electrical Engineering and Computer Science, Texas A&M University-Kingsville, Kingsville, TX, USA

* thoque@uno.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:Y; DeepL:N; EMA:Y; MD:Y*

In CASP15, we evaluate our proposed novel *ab initio* protein structure prediction (PSP) method, called 3DIGARS-PSP. 3DIGARS-PSP method utilizes an advanced search algorithm called KGA and an effective statistical energy function called 3DIGARS. It uses a memory-assisted genetic algorithm (GA) derived from KGA to overcome the critical search process. GA employs two effective operators: memory-assisted crossover and mutation, which are decorated with angle rotation and segment translation features. Likewise, dihedral angle distribution and secondary structure propensities are utilized to guide the conformational search. The GA-based sampling generates a large decoy pool to minimize the statistical energy function. Finally, we cluster the ensemble of decoys to identify the top five models for each CASP15 target and then submit these models to CASP15.

Methods

Protein structure is primarily represented by backbone atoms N, C α , C, and O in 3DIGARS-PSP. For each CASP15 target, we first obtain the predicted models from AlphaFold2 [1]. Then, the method starts by initializing some of the chromosomes of the GA population with the Cartesian coordinates of the backbone atoms of the models from AlphaFold2 ¹. Next, the remaining chromosomes are initialized by single-point torsion angle changes (rotation). We utilize the frequency of occurrence of 20 different amino acids with different Φ - Ψ angle pairs for an informed change of the torsion angles (Φ or Ψ). The frequency of occurrence is calculated from the 4,332 high-resolution experimental structures extracted in our previous work ². The range of both Φ and Ψ angles for every amino acids are divided into 120 bins with an interval of 3 degrees, and the frequencies of the bins are updated based on the value of the Φ and Ψ angles. By examining the cluster of the frequency values, it is possible to further categorize the frequency distribution for each amino acid into zones. Then, to select the most probable torsion angles (namely, $p\Phi$ or $p\Psi$) belonging to the zone, the roulette wheel selection approach is applied. Next, a random Φ or Ψ (say, $r\Phi$ or $r\Psi$) between $p\Phi-3$ and $p\Phi$ or $p\Psi$ and $p\Psi+3$ is

selected, and rotation of the current torsion angle is performed to achieve a new torsion angle, $r\Phi$ or $r\Psi$.

In addition, the change of the torsion angles is further guided by the propensities of secondary structure (SS) types of the amino acids extracted from the 4,332 high-resolution experimental structures by running the DSSP program. The eight different SS types (E, B, H, G, I, T, S, and U) given by DSSP are broadly categorized into four different SS types (H, G, and I = H; E and B = E; T and S = T; and U). The Φ - Ψ angle pair and SS types are used to obtain the index in the SS frequency table and increase the frequency count of the cell in the table by one. Next, the SS type, which has the largest frequency count, is assigned to the given amino acid having a certain Φ - Ψ angle. Furthermore, we collect the Φ - Ψ angle pairs belonging to the H and E types and group them into helix and beta groups. We utilize the Φ - Ψ angle pairs belonging to the helix or sheet group to update the Φ or Ψ angle that results in the clash within the structure.

We apply two types of conformational change operators *i*) angle rotation; and *ii*) segment translation, to generate new chromosomes (structural samples) for the next generation of GA. The mutation operation involves phi or psi angle rotation, and the crossover operation involves segment translation followed by phi or psi angle rotation at the crossover point. The rotation of phi and psi angles is based on the idea of rotation about an arbitrary axis. For segment translation, a set of possible crossover points are selected based on the secondary structure information. All amino acid indexes except the amino acids belonging to the beta-sheet secondary structure type (either E or B) are considered as possible crossover points. This is done to preserve beta-sheet regions in the structure from random changes during the crossover operation and perform more controlled changes to this region while performing mutation operations. We generate four child structures from two parent structures using the crossover process and a structure with the best fitness saved in the memory ³.

Decoys are generated by minimizing the potential energy using associated memory GA discussed above using the statistical energy function. Each decoy generated by 3DIGARS-PSP is then converted into the all-atom level by using Oscar-star software ⁴. Then a single-model based model quality assessment program Qprob ⁵, which predicts a model's quality by estimating the error of structural, physiochemical, and energy-based features using probability density distributions, is used to rank the decoys. Next, the MUFOLD-CL ⁶ method is used to cluster the decoys. Then, we select the top five models in different clusters based on their Qprob rankings. The top five models are further refined using ModRefiner ⁷ software. Then, we use the ResQ ⁸ method to add B-factors to the top five models before submission.

Availability

Source code, manual, and example data of 3DIGARS-PSP for Linux are freely available for non-commercial use at http://cs.uno.edu/~tamjid/Software/ab_initio/v2/PSP.zip.

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583-9.

2. Mishra A, Hoque MT. Three-Dimensional Ideal Gas Reference Sstate Based Energy Function. *Current Bioinformatics*. 2017;12:171-80.
3. Hoque MT, Iqbal S. Genetic algorithm-based improved sampling for protein structure prediction. *International Journal of Bio-Inspired Computation*. 2017;9:129-41.
4. Liang S, Zheng D, Zhang C, Standley DM. Fast and accurate prediction of protein side-chain conformations. *Bioinformatics*. 2011;27:2913-4.
5. Cao R, Cheng J. Protein single-model quality assessment by feature-based probability density functions. *Scientific Reports*. 2016;6:23990.
6. Zhang J, Xu D. Fast algorithm for population-based protein structural model analysis. *Proteomics*. 2013;13:221-9.
7. Xu D, Zhang Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophysical Journal*. 2011;101:2525-34.
8. Yang J, Wang Y, Zhang Y. ResQ: An approach to unified estimation of B-factor and residue-specific error in protein structure prediction. *Journal of Molecular Biology*. 2016;428:693-701.

Ab Initio Protein Structure Prediction by Conditioned Self-Avoiding Walk and Monte Carlo Tree Search

Weitao Sun^{1,2}, Diyao Wang¹

1 – School of Aerospace Engineering, Tsinghua University, Beijing, 100084, China; 2 – Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing, 100084, China

sunwt@tsinghua.edu.cn

Key: Langevin equation; conditioned self-avoiding walk; Monte-Carlo method; rigid cranks; Monte-Carlo tree search

All-Atom Conditioned Self-Avoiding Walk (AA-CSAW) has been used in predicting protein structures since CASP9 (Group name: sun@tsinghua). It is an ab initio protein folding simulation model based on Langevin equation and Monte-Carlo (MC) method^{1,2}. There are no other modeling methods used as integral part of AA-CSAW method. AA-CSAW is the same method for all predictions. We did not perform any manual intervention and no templates, MSA or CASP-hosted servers have been used. The dihedral angle distribution for three-residue fragment has been used in choosing torsion angles. Monte-Carlo tree searching is tested in the prediction, but no advanced (deep learning) machine-learning methods were used. We did not use any methods for ranking predictions. We only used AA-CSAW energy function for the refinement, is constructed by considering hydrophobic effect, desolvation effect and hydrogen bonding interaction³.

Methods

The protein chain is regarded as a series of rigid cranks connected by covalent bonds. Bond lengths and bond angles are obtained from chemical structures of 20 amino acids. Protein structure is determined by backbone and sidechain dihedral angles. An unfolded protein structure is generated as initial structure. The residue in protein chain is pivoted by setting dihedral angles to certain values. In the pivot algorithm, the backbone dihedral angles for each residue are chosen in Ramachandran plot according to a probability distribution derived from 3-residue fragment set. The effective energy of protein structure is constructed by considering hydrophobic effect, desolvation effect and hydrogen bonding interaction. An appropriate three dimensional structure is accepted with a probability according to Metropolis scheme. A ratio of secondary structure content to radius of gyration is proposed to evaluate the predicted structures as a supplement to the energy evaluation criteria.

Based on atom locations in each residue crank and the dihedral angles between every two cranks, the main chain and sidechain atom coordinates are determined. The sidechain rotamer distribution is used to provide sidechain structures with low energy. The hydrophobic energy is estimated based on two factors: the solvent accessible surface area (SASA) and residue types. A residue with more neighbors is buried in protein and has less SASA. In addition, if the

surrounding residues are all hydrophobic residues, the residue has high hydrophobic energy. A pair of residues are considered in contact based on Atom Distance criteria (ADC) model^{4,5}. We introduce a scheme to decrease the hydrophobic energy when the aggregation of hydrophobic residue grows to large size. This method provides more chances to open the hydrophobic core, which is essential for misfolded intermediate structures. The DSSP⁶ method is used as HB criterion. The total number of hydrogen bonds is a measurement of HB energy. An optimal HB strength parameter is used to account for the stability of hydrogen bond at different locations. In order to prevent the formation of tight hydrophobic core without hydrogen bonding, we introduce a penalty to buried NH, CO groups without hydrogen bonds.

Monte Carlo Tree Search (MCTS) is tested in predicting protein structures. In GO game, subsequent possible child nodes are searched through MCTS. In AA-CSAW, the current state of protein structure is determined by the amino acid sequence and the corresponding dihedral angles. The subsequent state towards folded structure is searched by MCTS. Note that there is no way to strictly define the "win or loss" like in GO game. So an upper limit number of pivot iteration and the energy convergence (energy drop between every two pivots is less than one percent of the original energy) is used as the termination of the MCTS searching.

The AA-CSAW is now a parallel code and can produce a bunch of candidate structures at the same time.

Results

All results, intermediate data files, and performance analysis documents are available on the web at <https://www.researchgate.net/profile/Weitao-Sun-4>.

Availability

The AA-CSAW software is written in C++ and have been compiled and tested on both Windows and LINUX systems. The software is available by sending email to sunwt@tsinghua.edu

1. Huang, K. (2007) CONDITIONED SELF-AVOIDING WALK (CSAW): STOCHASTIC APPROACH TO PROTEIN FOLDING. *Biophysical Reviews and Letters* **2**, 139-154.
2. Weitao, S. (2007), *2007 IEEE International Conference on Bioinformatics and Biomedicine Workshops*, pp. 45-52.
3. Sun, W. (2007) *Protein folding simulation by all-atom CSAW method*.
4. Sun, W. and He, J. (2010) Understanding on the Residue Contact Network Using the Log-Normal Cluster Model and the Multilevel Wheel Diagram. *Biopolymers*, **93**, 904-916.
5. Sun, W.T. and He, J. (2011) From Isotropic to Anisotropic Side Chain Representations: Comparison of Three Models for Residue Contact Estimation. *Plos One*, **6**.
6. Kabsch, W. and Sander, C. (1983) Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, **22**, 2577-2637.

Structure Prediction for CASP15 Assembly Targets

Shinpei Kobayashi, Yuta Miyakawa, Yasuomi Kiyota and Mayuko Takeda-Shitaka

School of Pharmacy, Kitasato University, Tokyo, Japan

shitakam@pharm.kitasato-u.ac.jp

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:N*

We participated in the assembly category of CASP15. We predicted both homo- and hetero-oligomeric protein structures according to the oligomeric state in the CASP15 target list.

Methods

Structure Preparation: We generated protein structure models (25 models) using AlphaFold-Multimer¹ for assembly target. We also used AlphaFold-Multimer models (25 models) by Arne Elofsson's group (<http://duffman.it.liu.se/casp15/>).

Template Search: To obtain monomeric and oligomeric templates, we carried out two-step template search². Firstly, templates were searched by HHblits³ against UniRef30 and PDB70 database. Secondly, to search templates more widely, we ran PSI-BLAST⁴ on PDBaa using HHblits hits as inputs.

Additional model construction: We visually inspected structure models, and compared them with templates obtained by two-step template search. When models were unfolded, we constructed structure models additionally using AlphaFold-Multimer based on the information of domains, monomeric and oligomeric templates with human intervention.

Assessment of model quality: For dimeric complex, we used our original quality assessment method based on learning to rank approach⁵. Our original ranking method was developed for predicting the model ranking based on DockQ⁶ score. In this method, SOAP-PP⁷ score, VoromQA⁸ score and Rosetta⁹ score were used as input features into learning to rank with LightGBM¹⁰. For other multimeric complex, we used the model confidence score from AlphaFold-Multimer for ranking. Clash information at the interface was considered manually, if necessary.

1. Evans R, et al. (2021). Protein complex prediction with alphafold-multimer. BioRxiv.
2. Kiyota Y, Kobayashi S, Harada Y, Takeda-Shitaka Y. CASP14 abstract book. (2020). p266-227.
3. Steinegger M, et al. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *Bioinformatics*. 20, 473.
4. Altschul SF, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.
5. Harada Y. Kitasato University, thesis, (2022).

6. Basu S, Wallner B. (2016). DockQ: A Quality Measure for Protein-Protein Docking Models. PLoS ONE 11(8): e0161879.
7. Dong GQ, et al. (2013). Optimized atomic statistical potentials: assessment of protein interfaces and loops. Bioinformatics. 29, 3158- 3166.
8. Olechnovič K, Venclovas Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins. 85, 1131-1145.
9. Alford RF, et al. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput. 2017 Jun 13;13(6):3031-3048.
10. Guolin Ke, et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 31st conference Neural Information Processing Systems NIPS. 2017. 3149-3157.

Deep Learning based Protein and Complex Structure Prediction

Y.Tang, M.Jin, Z.Dong and H.Miao

XLab, Shanghai Tianrang Intelligence

hj.miao@tianrang-inc.com

Key: Auto:N; CASP_serv:N; Templ:Y/N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y

In CASP15 we applied TRFold and TRComplex for the modeling of single-chain protein and assembly targets respectively. For single-chain targets known to be part of an assembly their structures were directly parsed from the complex rather than being modeled separately as monomers. TRFold-single which uses only the target's sequence as input was developed during CASP experiments and subsequently applied.

Methods

Input features: TRFold and TRComplex utilize both multiple sequence alignment (MSA) and template information for structure prediction. Latest UniRef30, UniRef90, and metagenomic database colabfold_envdb are searched and sequences found are concatenated to construct target's MSA. The component from colabfold_envdb is further filtered to limit its size. Templates are found by HHsearch through latest PDB70 downloaded in April. TRFold-single does not use MSA but the target's sequence is first embedded through a transformer-based language model pre-trained on UniRef sequences and used as input.

Single-chain protein modeling: TRFold is a transformer-based neural network that takes a target's sequence, MSA and templates as input and outputs predicted distance matrix, torsion angles, coordinates and associated TM-score. The network was first trained on a non-redundant set of high-resolution single-chain proteins from the Protein Data Bank and then finetuned on both native and predicted structures of a non-redundant subset of UniRef30. The predicted set was fed progressively and several parameter settings were tried during finetuning. For every monomeric target the five highest scoring decoys were submitted. As we noticed, very early on in this CASP experiment, significant conformational difference between assembly component chain predictions and its corresponding monomeric decoys, all single-chain targets that are known to be part of an assembly were parsed directly from the complex structure. Coordinate-constrained relaxation was done using Amber force fields to further reduce clash.

Assembly modeling: instead of training a multimer model from scratch, we applied spatial cropping on single-chain proteins at the finetuning stage to simulate inter-chain interactions. MSAs for assembly targets were constructed by concatenating MSAs of individual chains without pairing. A total of 50 decoys were generated for each target and the five highest

scoring ones were submitted. For large assemblies, the largest substructure that could fit onto our GPU was predicted and submitted through TRFold group whereas whole structure prediction was tried on CPUs and submitted through trComplex group if completed before the deadline.

Manual intervention: templates were manually grouped and selected for several targets including T1109, T1110, T118, T1158, T1162, T1162, T1195, T1196 and T1197. MSAs were also manually filtered to produce alternative conformations for targets T1195 to T1197. Higher number of recycles were applied for orphan targets such as T1130. Aside from predicted scores, visual inspection and selection based on literature information was carried out on antigen-antibody complex targets. Unrelaxed predictions were submitted for several large assemblies as relaxation caused substantial backbone conformation changes.

Development during CASP: TRFold-single, a target sequence-only model was trained seeing the release of single-point mutants T1109 and T1110, and subsequently one of the five submissions was from TRFold-single regardless of its predicted TM-score. By introducing memory saving techniques, the total foldable length increased from ~1400 amino acids to ~3500 on our in-house 24G GPU after target H1137. A model with higher weights on clash and bond length violation loss was trained and preferred for large assemblies.

Results

Earlier version of TRFold trained on pre-CASP14 PDB data was tested on all CASP14 targets by back-rolling sequence and template databases to that of April 2020 and achieved an average TM-score of 0.902. TRFold-single was tested on a set of de novo designed proteins recently deposited in the PDB and achieved an average TM-score of 0.862, ~4% higher than that of AlphaFold2.

Availability

All methods and algorithms applied in CASP15 are available on our AI-directed protein design workbench at <https://xlab.tianrang.com>, where structure prediction, protein design, property analysis and optimization algorithms are hosted in an all-in-one platform.

1. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10), 1282-1288.
2. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, 1-4.
3. Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics*, 21(7), 951-960.
4. Berman, H., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, 10(12), 980-980.

5. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589.
6. Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research*, 33(7), 2302-2309.
7. Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11), 1026-1028.
8. Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham III, T. E., Laughton, C. A., & Orozco, M. (2007). Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophysical journal*, 92(11), 3817-3829.

OpenComplex-RNA predicts RNA 3D structure at the atomic level

Wenjun Lin¹, Zhaoming Chen¹, Zhaoqun Li¹, Jingcheng Yu¹, Wenhao Huang¹,
Yaqing Zhang^{1,2*} and Qiwei Ye^{1*}

¹ Beijing Academy of Artificial Intelligence, Beijing, China;

² Institute of Pharmacy and Molecular Biotechnology, Heidelberg University, Germany

* yaqing.zhang@uni-heidelberg.de, qwye@baai.ac.cn

Key: Auto:Y; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y;
DeepL:Y; EMA:Y; MD:Y

Overcoming the limitation of extremely little trainable data and highly flexible geometry for deciphering RNA 3D structure, we present OpenComplex-RNA, a deep-learning model that jointly incorporates multiple sequence alignment (MSA) encoder and atomic structural decoder with optimized functional scores in CASP15. Its successive feature enables capturing detailed coordinates of RNA base not limited to the complex sequence length, the sophisticated multimer interaction, and the distinct ligands. In short, OpenComplex-RNA predicts accurate RNA structures in diverse application scenarios.

Methods

Feature extraction: By given query sequence, their potential MSAs, structural templates, and secondary structures were fully explored together with relevant published experimental data (Figure 1B). Briefly, (i) MSA was generated using rMSA¹ by querying genetic databases including Rfam, RNACentral, and nt databases. (ii) Local structural templates and RNA-ligand complex templates were collected from the PDB database after searching RNA sequence data and SMILES of ligands. (iii) From the Rfam and RNACentral databases, secondary structural information was gleaned. For specific cases, secondary RNA structures were indirectly retrieved from the published word and then modified manually. For targets that lack experimental secondary structure, we jointly combined three RNA secondary structure prediction techniques (RNAfold,² MXfold,³ and E2Efold⁴) as their consensus predictions.

RNA tertiary structures generation: Using the retrieved information, we created two parallel pipelines to generate RNA tertiary structure (Figure 1A). (i) For each target, the secondary structure and starting pose created by RNAComposer⁵ were fed into FARFAR2⁶ and simRNA⁷ to build tens of thousands of tertiary structures. As an alternate input for some targets, local structure templates were employed. (ii) The secondary structure and MSA were utilized as inputs for the MSA encoder in OpenComplex-RNA to build both single and pair representations, which direct the structure module to generate tertiary structures.

Model selection: To comprehensively examine the quality of generated structures, we ranked all the predicted tertiary structures according to their ARES⁸ scores and physical-based Rosetta⁹ ratings with specified local and global structure qualities (Figure 1D). By using the

Rosetta suite,¹⁰ we additionally estimated the secondary pairings of candidate models as alternative quality measures. In light of those targets that have accessible homologous structures, candidate structures were further identified by their structural similarity compared to the templates.

Docking: We performed RNA-RNA docking using the HNADOCK server,¹¹ which specifically designed an intrinsic scoring system for evaluating nucleic acid interaction (Figure 1C). While RNA-ligand docking was conducted along with the information of RNA-ligand templates that were extracted from the PDB database. Besides, RNA-protein docking was calculated by the HDOCK.¹² The human-optimized RNA-protein docking was viewed as the potential benchmark for downstream application.

Availability

The code and models will be made available to the public shortly.

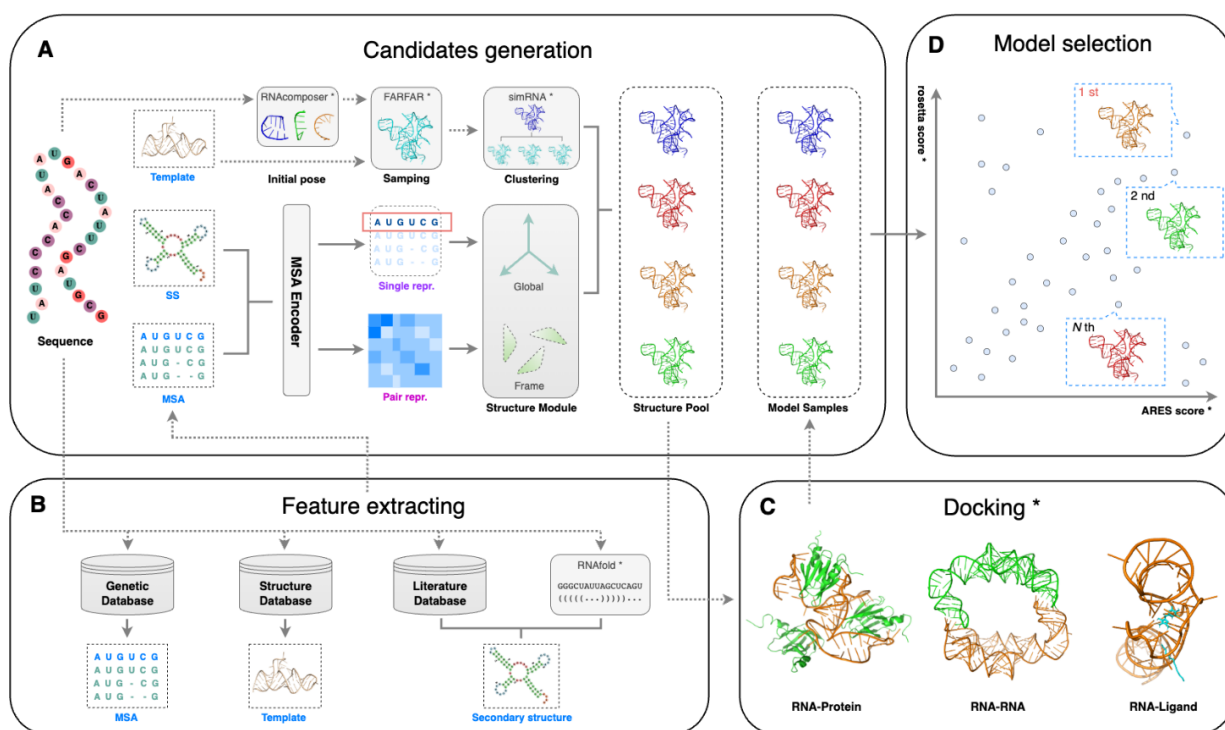


Figure 1: The Pipeline Overview.

A) Model architecture of candidate structure generation. Arrows indicate the information flow that connects the consecutive computational module. The rectangle with the dashed line represents the core data feature, while the one with the solid line represents the computational module. The reported algorithm was denoted with an asterisk. B) Relevant features were extracted from various databases for the downstream model prediction unit. C) Hyper-interaction related to multi-ligand was conducted along with RNA-RNA docking, RNA-protein,

and RNA-ligand docking accordingly. D) Final models were ranked according to the specified rules of ARES and the rosetta score.

1. “GitHub - pylelab/rMSA: RNA Multiple Sequence Alignment — github.com,” <https://github.com/pylelab/rMSA>, [Accessed 15-Sep-2022].
2. R. Lorenz, S. H. Bernhart, C. Höner zu Siederdisen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, “Viennarna package 2.0,” *Algorithms for molecular biology*, vol. 6, no. 1, pp. 1–14, 2011.
3. K. Sato, M. Akiyama, and Y. Sakakibara, “Rna secondary structure prediction using deep learning with thermodynamic integration,” *Nature communications*, vol. 12, no. 1, pp. 1–9, 2021.
4. X. Chen, Y. Li, R. Umarov, X. Gao, and L. Song, “Rna secondary structure prediction by learning unrolled algorithms,” *International Conference on Learning Representations*, 2020.
5. M. Biesiada, K. J. Purzycka, M. Szachniuk, J. Blazewicz, and R. W. Adamiak, “Automated rna 3d structure prediction with rnacomposer,” in *RNA Structure Determination*. Springer, 2016, pp. 199–215.
6. A. M. Watkins, R. Rangan, and R. Das, “Farfar2: improved de novo rosetta prediction of complex global rna folds,” *Structure*, vol. 28, no. 8, pp. 963–976, 2020.
7. M. J. Boniecki, G. Lach, W. K. Dawson, K. Tomala, P. Lukasz, T. Soltysinski, K. M. Rother, and J. M. Bujnicki, “Simrna: a coarse-grained method for rna folding simulations and 3d structure prediction,” *Nucleic acids research*, vol. 44, no. 7, pp. e63–e63, 2016.
8. R. Townshend, S. Eismann, A. M. Watkins, R. Rangan, M. Karelina, R. Das, and R. O. Dror, “Geometric deep learning of rna structure,” *Science (New York, N.Y.)*, vol. 373, no. 6558, pp. 1047–1051.
9. R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O’Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel *et al.*, “The rosetta all-atom energy function for macromolecular modeling and design,” *Journal of chemical theory and computation*, vol. 13, no. 6, pp. 3031–3048, 2017.
10. S. Chaudhury, S. Lyskov, and J. J. Gray, “Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta,” *Bioinformatics*, vol. 26, no. 5, pp. 689–691, 2010.
11. J. He, J. Wang, H. Tao, Y. Xiao, and S.-Y. Huang, “Hnadock: a nucleic acid docking server for modeling rna/dna–rna/dna 3d complex structures,” *Nucleic Acids Research*, vol. 47, no. W1, pp. W35–W42, 2019.
12. Yan Y, Tao H, He J, et al. The HDock server for integrated protein–protein docking[J]. *Nature protocols*, 2020, 15(5):1829-1852.

Integrating multi-MSA, threading templates and deep learning for protein structure prediction

Wei Zheng^{1,2}, Qiqige Wuyun³ and Peter L Freddolino^{1,2}

1 - Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA,

2 - Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA

3 - Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

zhengwei@umich.edu, wuyunqiq@msu.edu, and petefred@umich.edu

Key: Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y

The UM-TBM server used in CASP15 is designed for modeling regular protein targets (monomer protein) based on a fully automated combination of the extension pipeline of DeepMSA1, LOMETS2, AlphaFold23, and I-TASSER4; 5. The pipeline contains five steps: (i) MSA construction by DeepMSA2, (ii) multi-domain template detection by LOMETS32, (iii) initial models and spatial geometric restraints prediction by AlphaFold23 and other deep learning predictors, (iv) I-TASSER Replica-Exchange Monte Carlo (REMC) simulation for model construction, and (v) atomic-level model refinement by MD simulation.

Methods

We generate the multiple sequence alignments needed by later stages of our pipeline using DeepMSA2, which contains two sub-methods (dMSA and qMSA) to generate seven candidate multiple sequence alignments (MSAs). Here, dMSA is our previous MSA construction program (DeepMSA1) developed during CASP13, where HHblits6, Jackhmmer7 and HMMsearch7 are used to search the query sequence against the Uniclust308, UniRef909 and Metaclust10 databases in three stages (labeled stage 1 – stage 3 in the order listed above). qMSA is an extended version of dMSA with a new search added between stage 2 and stage 3 of dMSA, where HHblits is used to search the BFD11 metagenomic database. In addition, a new iteration stage (stage 4) is added in qMSA to search the query through the Mgnify12 metagenomic database. Thus, five different MSAs are generated by stages 1-3 of dMSA and by stages 3-4 of qMSA. Furthermore, the MSA from qMSA stage3 (obtained from the BFD database) is used as the starting point for HMMsearch to search through the IMG/M13, Tara14 and MetaSource15 metagenome databases, which contain more sequences than the Metaclust, BFD, and Mgnify databases. The resulting sequence hits are converted into a sequence database. This sequence database is then used as the target database for dMSA stage 3 and qMSA stage 4 to generate two additional MSAs. As an additional filtering step, the seven MSAs from DeepMSA2 are used as inputs for separate AlphaFold2 (1-embedding) runs to predict seven sets of models and the associated spatial geometric restraints, and the MSA associated with the highest pLDDT score from the AlphaFold2 models is selected as the final output of DeepMSA2.

The final MSA of DeepMSA2 is used for AlphaFold2 (8-embedding), AttentionPotential, and DeepPotential16 for the predictions of residue-residue contact maps, distance distributions, inter-residue torsion angles, and hydrogen-bond networks. Those deep learning-predicted restraints are utilized to guide the REMC folding simulation with the same set of restraints calculated from templates detected by LOMETS3. AttentionPotential is an extended pipeline from DeepPotential, which utilizes an MSA transformer¹⁷. The full sets of predicted restraints from AttentionPotential and DeepPotential are later fed into DeepFold, an L-BFGS folding system, to get ten full-length models. Those ten models, the five models generated by AlphaFold2, and the LOMETS3 full-chain level threading templates, are used as initial conformations in the REMC folding simulation.

The MSA generated from the DeepMSA2 is also used to produce sequence profiles or profile Hidden Markov Models (HMM) for six profile-based threading methods used by LOMETS3. The contact maps and distance distributions predicted from AlphaFold2, AttentionPotential and DeepPotential are used by five contact and distance-based threading methods in LOMETS3. Different from previous versions of LOMETS, LOMETS3 can automatically handle the multi-domain protein threading problem by adding domain partition and domain assembly modules. FUpred¹⁸ and ThreaDom¹⁹ are used as domain boundary prediction methods in LOMETS3, and DEMO²⁰ is used for assembling domain-level templates to full-chain level templates guided by distance restraints from AlphaFold2, AttentionPotential and DeepPotential. Finally, 110 (10 templates from each component threading method) full-chain level templates are collected by LOMETS3, and then used as initial conformations associated with the models constructed by AlphaFold2 and DeepFold in REMC simulation.

For target proteins with lengths of less than 300 residues, an I-TASSER-based REMC simulation is utilized for generating 10,000 decoy conformations. The REMC simulation is guided by residue-residue contact maps, distance distributions, inter-residue torsion angles, and hydrogen-bond networks that are predicted by deep learning predictors and calculated from LOMETS3 threading templates. The decoys are later clustered by SPICKER²¹ to obtain five clusters for final model selection. For the targets with lengths of greater than 300 residues, the top five ranked AlphaFold2 models are directly used in the next MD refinement.

The five cluster centroids (for target with length<300AA) or the five top ranked AlphaFold2 models (for target with length≥300AA) are further refined by FG-MD²² to remove steric clashes and refine the local structure packing, resulting in the final models.

1. Zhang, C., Zheng, W., Mortuza, S. M., Li, Y. & Zhang, Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* 36, 2105-2112.
2. Zheng, W., Wuyun, Q., Zhou, X., Li, Y., Freddolino, P. L. & Zhang, Y. (2022). LOMETS3: integrating deep learning and profile alignment for advanced protein template recognition and function annotation. *Nucleic Acids Research* 50, W454-W464.
3. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.

- A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583-589.
4. Zheng, W., Zhang, C., Li, Y., Pearce, R., Bell, E. W. & Zhang, Y. (2021). Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Reports Methods* 1, 100014.
 5. Zheng, W., Li, Y., Zhang, C., Zhou, X., Pearce, R., Bell, E. W., Huang, X. & Zhang, Y. (2021). Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. *Proteins: Structure, Function, and Bioinformatics* 89, 1734-1751.
 6. Remmert, M., Biegert, A., Hauser, A. & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* 9, 173-175.
 7. Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R. & Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research* 46, W200-W204.
 8. Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J. & Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research* 45, D170-D176.
 9. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H. & UniProt, C. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)* 31, 926-932.
 10. Steinegger, M. & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications* 9, 2542.
 11. Steinegger, M., Mirdita, M. & Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods* 16, 603-606.
 12. Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A. & Finn, R. D. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research* 48, D570-D578.
 13. Chen, I. M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Huntemann, M., Varghese, N., White, J. R., Seshadri, R., Smirnova, T., Kirton, E., Jungbluth, S. P., Woyke, T., Eloe-Fadrosh, E. A., Ivanova, N. N. & Kyrpides, N. C. (2019). IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research* 47, D666-D677.
 14. Wang, Y., Shi, Q., Yang, P., Zhang, C., Mortuza, S. M., Xue, Z., Ning, K. & Zhang, Y. (2019). Fueling ab initio folding with marine metagenomics enables structure and function predictions of new protein families. *Genome Biology* 20, 229.
 15. Yang, P., Zheng, W., Ning, K. & Zhang, Y. (2021). Decoding the link of microbiome niches with homologous sequences enables accurately targeted protein structure prediction. *Proceedings of the National Academy of Sciences* 118, e2110828118.
 16. Li, Y., Zhang, C., Zheng, W., Zhou, X., Bell, E. W., Yu, D.-J. & Zhang, Y. (2021). Protein inter-residue contact and distance prediction by coupling complementary coevolution features with deep residual networks in CASP14. *Proteins: Structure, Function, and Bioinformatics* 89, 1911-1921.
 17. Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J. F., Abbeel, P., Sercu, T. & Rives, A. (2021). MSA Transformer. *bioRxiv*, 2021.02.12.430858.

18. Zheng, W., Zhou, X., Wuyun, Q., Pearce, R., Li, Y. & Zhang, Y. (2020). FUpred: detecting protein domains through deep-learning-based contact map prediction. *Bioinformatics* 36, 3749-3757.
19. Xue, Z., Xu, D., Wang, Y. & Zhang, Y. (2013). ThreaDom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics* 29, i247-i256.
20. Zhou, X., Hu, J., Zhang, C., Zhang, G. & Zhang, Y. (2019). Assembling multidomain protein structures through analogous global structural alignments. *Proceedings of the National Academy of Sciences* 116, 15930.
21. Zhang, Y. & Skolnick, J. (2004). SPICKER: a clustering approach to identify near-native protein folds. *Journal of computational chemistry* 25, 865-871.
22. Zhang, J., Liang, Y. & Zhang, Y. (2011). Atomic-Level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling. *Structure* 19, 1784-1795.

Protein structure prediction in CASP14 with the coarse-grained UNRES model

M. Maszota-Zieleniak¹, K.K. Bojarski², E.A. Lubecka², M. Marcisz¹, A. Danielson¹,
Ł. Dziadek¹, M. Gaardløs¹, A. Gieldon¹, A. Liwo¹, S.A. Samsonov¹, R. Slusarz¹, K. Zieba¹,
A.K. Sieradzan^{1,*}, C. Czaplewski¹

1- University of Gdańsk, Wita Stwosza 63, Gdańsk; 2- Technical University of Gdańsk, ul. G. Narutowicza 11/12,
Gdańsk

adam.sieradzan@ug.edu.pl

Key: Auto:N; CASP_serv:N; Templ:Y; MSA:N;Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N;
EMA:N; MD:Y

We tested, with the CASP15 targets, our methodology for protein-structure prediction, which is based on the UNRES heavily coarse-grained physics-based model.¹ In UNRES the only interaction sites are united backbone peptide groups and united side chains, the alpha-carbon atoms serving to define backbone geometry.

Methods

Both monomeric and oligomeric targets were treated and the models of oligomeric targets were submitted both to CASP and to CAPRI. The UNRES group used the consensus-fragment restraints derived from the server models and the models generated by the in-house installation of iTASSER and AlphaFold2. For oligomeric targets, the HHpred server was used to obtain hints as to the possible structures of oligomers. The prediction procedure consisted of the following stages² (i) running restrained multiplexed replica-exchange (MREMD) simulations of the targets with UNRES to explore the conformational space, (ii) determining the analysis temperature (before the unfolding-transition temperature), and determining the probabilities of the conformations by using weighted histogram analysis method (WHAM), (iii) dissecting the simulated conformations into 5 (CASP) or 10 (CAPRI) families by minimum-variance clustering and selecting the conformations closest to cluster means for further processing, ranking following cluster free energy (iv) conversion of the coarse-grained structures to all-atom structures to obtain the candidate predictions which were submitted to CASP/CAPRI. MREMD simulations were started from the server models. The recent extension of UNRES enabled us to run simulations for very large targets and DNA/protein targets.

Results

As the official CASP15 results had not been published at the time the poster was created, only the results of comparison of the predictions of the UNRES group with the corresponding experimental structures that had been released in the PDB at that time are presented.

Availability

The software is available at unres.pl and git distribution.

1. Liwo A, Sieradzan AK, Lipska AG, Czaplewski C, Joung I, Żmudzińska W, Hałabis A, Ołdziej S. A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. III. Determination of scale-consistent backbone-local and correlation potentials in the UNRES force field and force-field calibration and validation. *J Chem Phys.* 2019 Apr 21;150(15):155104. doi: 10.1063/1.5093015. PMID: 31005069.
2. Antoniak A, Biskupek I, Bojarski KK, Czaplewski C, Giełdoń A, Kogut M, Kogut MM, Krupa P, Lipska AG, Liwo A, Lubecka EA, Marcisz M, Maszota-Zieleniak M, Samsonov SA, Sieradzan AK, Ślusarz MJ, Ślusarz R, Wesołowski PA, Zięba K. Modeling protein structures with the coarse-grained UNRES force field in the CASP14 experiment. *J Mol Graph Model.* 2021 Nov;108:108008. doi: 10.1016/j.jmgm.2021.108008. Epub 2021 Aug 17. PMID: 34419932.

Modeling and Scoring Protein Assemblies in CASP15

K. Olechnovič, J. Dapkūnas, L. Valančauskas and Č. Venclovas

Institute of Biotechnology, Life Sciences Center, Vilnius University

ceslovas.venclovas@bti.vu.lt

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:Y*

In CASP15 our group participated in predicting 3D structures of protein assemblies, protein-RNA complexes and large multidomain single-chain proteins as well as in estimating accuracy of assemblies and intersubunit interfaces.

Methods

The general workflow for 3D structure prediction consisted of two major stages: (1) construction of an ensemble of multiple diverse structural models and (2) selection of the best models using a newly developed accuracy estimation protocol. Accuracy estimation procedures were based on both new and previously developed contact-area based scoring functions.

Construction of 3D models. Initial ensembles of protein complexes were constructed using both AlphaFold-PTM¹ and AlphaFold-Multimer-v2² versions available either as the ColabFold³ or the original DeepMind's implementation. The DeepMind's AlphaFold modeling pipeline included full databases for multiple sequence alignment generation and PDB templates. The ColabFold-based AlphaFold modeling pipeline employed a variety of different parameters and conditions so as to achieve extensive structure sampling. These variations included the choice of sequence databases, the construction and pairing of multiple sequence alignments as well as the number of AlphaFold recycles. If AlphaFold failed to generate the complex (assembly was too large to handle or subunits did not form a complex), structural models were obtained using docking. Docking models were also added if the resulting AlphaFold models had poor self-estimated accuracy (pLDDT, pTM, ipTM) or did not show structural consensus. Docking was performed using established rigid-body docking tools. FTDOCK⁴ and HEX⁵ were used for generating heterodimers, whereas SAM⁶ was used for generating symmetric homomers. In several cases, when closely related templates were available as identified by PPI3D⁷, Modeller-based homology models were constructed as well⁸. Large monomeric multidomain proteins were modeled using AlphaFold and ColabFold using similar pipeline as for protein complexes. Protein-RNA complexes were constructed as follows. Models for protein subunits were obtained using AlphaFold, whereas CASP server models were used to represent the RNA moiety. The ensembles of protein-RNA complexes were obtained using exhaustive docking. Once sets of 3D structures for protein-protein, protein-RNA or multidomain monomers were obtained, the top five models were selected as described below.

Model ranking and selection. Model ranking and selection was done using a newly developed VoroIF-jury (Voronoi-based InterFace jury) procedure. VoroIF-jury resembles EMA-jury, developed previously for assessing the models in a recent CASP-commons experiment focused on SARS-CoV-2⁹. Given a set of models, VoroIF-jury (a) computes multiple rankings

using different interface-focused scores (most of them based on the VoroMQA interface energy¹⁰; (b) pools the top 1, top 2, ..., top N models selected by each EMA ranking into N corresponding supersets; (c) for every model in each superset calculates the VoroIF-jury interface consensus score, that is an average of the interface CAD-score¹¹ values derived by comparing a given model with other models in the superset; (d) ranks models by the best achieved VoroIF-jury score; (e) removes redundant models from the final ranking using the interface CAD-score-based clustering. For docking models VoroIF-jury was applied in two stages: (1) selecting top 300 from all the docking models, often exceeding 100 000; (2) after relaxing those 300 models using OpenMM¹² to remove clashes, selecting the final top five models.

VoroIF-jury included two newly developed interface scoring methods. The first one, a generic interatomic contact area-based energy potential, applicable for scoring of not only protein-protein, but also protein-nucleic acids interfaces, was derived from the protein-protein VoroMQA potential¹⁰. The second one, VoroIF-GNN, a graph neural network-based method, was developed for predicting the residue-level interface accuracy in models of protein-protein complexes. VoroIF-GNN is based on a graph attention network (GAT) that accepts a Voronoi tessellation-derived graph of inter-chain interface contacts. The network was trained using heterodimeric models produced by rigid body redocking of complexes from PDB. The ground truth interface quality scores were calculated by comparing the models with the corresponding experimental structures using interface CAD-score¹¹.

Analysis of the available structural data. In addition to automatically generated models, available information on the target proteins was also considered, starting with the UniProt database. Available structural data for homologous proteins were queried using HHpred¹³, COMER¹⁴ and DALI¹⁵ servers. Multimeric templates were identified using the PPI3D⁷ server, and the oligomeric states of structurally resolved homologs were additionally checked using the RCSB PDB Advanced search¹⁶. If multimeric structural templates were available, homology models were generated using the PPI3D web server. Disordered regions were predicted by DISOPRED3¹⁷. All the available information was used in manual model selection and re-ranking for harder targets.

Accuracy estimates for multimeric complexes and inter-subunit interfaces. The VoroIF-jury method was also used by the “Venclovas” group to derive accuracy estimates in the EMA category. Among two other EMA groups, “VoroIF” used the newly developed VoroIF-GNN method. “VoroMQA-select-2020” used the same tournament-based scoring procedure that was used by the “Venclovas” group in CASP14⁸.

Availability

The methods developed in our laboratory (PPI3D, VoroMQA, CAD-score, COMER) are available at <https://bioinformatics.lt/software/>.

1. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Židek,A., Potapenko,A., Bridgland,A., Meyer,C., Kohl,S.A.A., Ballard,A.J., Cowie,A., Romera-Paredes,B., Nikolov,S., Jain,R., Adler,J., Back,T., Petersen,S., Reiman,D., Clancy,E., Zielinski,M., Steinegger,M., Pacholska,M., Berghammer,T., Bodenstein,S., Silver,D., Vinyals,O., Senior,A.W., Kavukcuoglu,K., Kohli,P. & Hassabis,D. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589.

2. Evans,R., O'Neill,M., Pritzel,A., Antropova,N., Senior,A., Green,T., Žídek,A., Bates,R., Blackwell,S., Yim,J., Ronneberger,O., Bodenstern,S., Zielinski,M., Bridgland,A., Potapenko,A., Cowie,A., Tunyasuvunakool,K., Jain,R., Clancy,E., Kohli,P., Jumper,J. & Hassabis,D. (2022) Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034 doi:10.1101/2021.10.04.463034.
3. Mirdita,M., Schütze,K., Moriwaki,Y., Heo,L., Ovchinnikov,S. & Steinegger,M. (2022) ColabFold: making protein folding accessible to all. *Nat Methods* 19, 679–682.
4. Gabb,H.A., Jackson,R.M. & Sternberg,M.J. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* 272, 106–120.
5. Ritchie,D.W. & Kemp,G.J. (2000) Protein docking using spherical polar Fourier correlations. *Proteins* 39, 178–194.
6. Ritchie,D.W. & Grudin,S. (2016) Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry. *J Appl Cryst* 49, 158–167.
7. Dapkūnas,J., Timinskas,A., Olechnovič,K., Margelevičius,M., Dičiūnas,R. & Venclovas,Č. (2017) The PPI3D web server for searching, analyzing and modeling protein-protein interactions in the context of 3D structures. *Bioinformatics* 33, 935–937.
8. Dapkūnas,J., Olechnovič,K. & Venclovas,Č. (2021) Modeling of protein complexes in CASP14 with emphasis on the interaction interface prediction. *Proteins* 89, 1834–1843.
9. Kryshtafovych,A., Moutl,J., Billings,W.M., Della Corte,D., Fidelis,K., Kwon,S., Olechnovič,K., Seok,C., Venclovas,Č., Won,J., & CASP-COVID participants. (2021) Modeling SARS-CoV-2 proteins in the CASP-commons experiment. *Proteins* 89, 1987–1996.
10. Olechnovič,K. & Venclovas,Č. (2017) VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins* 85, 1131–1145.
11. Olechnovič,K. & Venclovas,Č. (2020) Contact Area-Based Structural Analysis of Proteins and Their Complexes Using CAD-Score. *Methods Mol Biol* 2112, 75–90.
12. Eastman,P., Swails,J., Chodera,J.D., McGibbon,R.T., Zhao,Y., Beauchamp,K.A., Wang,L.-P., Simonett,A.C., Harrigan,M.P., Stern,C.D., Wiewiora,R.P., Brooks,B.R. & Pande,V.S. (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* 13, e1005659.
13. Zimmermann,L., Stephens,A., Nam,S.-Z., Rau,D., Kübler,J., Lozajic,M., Gabler,F., Söding,J., Lupas,A.N. & Alva,V. (2018) A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.* 430, 2237–2243.
14. Dapkūnas,J. & Margelevičius,M. (2022) The COMER web server for protein analysis by homology. submitted.
15. Holm,L. (2022) Dali server: structural unification of protein families. *Nucleic Acids Res* 50, W210–W215.
16. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G.V., Christie,C.H., Dalenberg,K., Di Costanzo,L., Duarte,J.M., Dutta,S., Feng,Z., Ganesan,S., Goodsell,D.S., Ghosh,S., Green,R.K., Guranović,V., Guzenko,D., Hudson,B.P., Lawson,C.L., Liang,Y., Lowe,R., Namkoong,H., Peisach,E., Persikova,I., Randle,C., Rose,A., Rose,Y., Sali,A., Segura,J., Sekharan,M., Shao,C., Tao,Y.-P., Voigt,M., Westbrook,J.D., Young,J.Y., Zardecki,C. & Zhuravleva,M. (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* 49, D437–D451.

17. Jones,D.T. & Cozzetto,D. (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* 31, 857–863.

AlphaFold with improved sampling

Björn Wallner

IFM Bioinformatics, Linköping University

bjorn.wallner@liu.se

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

In this CASP we utilized a modified version of AlphaFold¹ with improved sampling capability that has demonstrated good performance for peptide-protein structure modeling². The method is completely automated, but was run as a manual server to allow for more computational time.

Methods

Improved sampling using AlphaFold was achieved by activating the dropout layers at inference and generating at least 200 models with a different random seed per neural network weight set. For the monomer case both the original five network weights, *model_[1,2,3,4,5]*, as well as the updated ptm weights, *model_[1,2,3,4,5]_ptm*, were used. For the multimer case, version 1 weight set, *model_[1,2,3,4,5]_multimer*, and version 2 weights, *model_[1,2,3,4,5]_multimer_v2*, were both used.

Additional models were also generated without template information and without template information and also an increased number of recycles.. Thus, for each target the goal was to generate 6,000 models (5x2x200x3), for some large targets this was too time consuming, but for most targets it was not a problem.

In addition, if the best score for any of the 6,000 models was <0.70, more models were generated. For some targets up to 30,000 models were made. The model with the highest score was always submitted as TS1, but to avoid submitting five identical models, a filter was applied to submit the model with the highest score but at least 2Å RMSD or lower than 0.8 MAlign³ from a previously submitted model for monomer, and multimers, respectively.

Multiple sequence alignments provided at <http://duffman.it.liu.se/casp15/> for the CASP community were used to facilitate straight comparison to the baseline AlphaFold versions.

1. Jumper, J. *et al.* (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 1–11 doi:10.1038/s41586-021-03819-2.
2. Johansson-Åkhe, I. & Wallner, B. (2022). Improving Peptide-Protein Docking with AlphaFold-Multimer using Forced Sampling. *Front. Bioinform.* doi: 10.3389/fbinf.2022.959160

3. Mukherjee, S. & Zhang, Y. (2009). MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res* **37**, e83–e83.

Protein Complex Structure Prediction by Multiple Strategies in CASP 15

Wenyi Zhang^{1,2}, Tengyu Xie^{1,2}, Runtong Qian^{1,2}, Zongyang Qiu^{1,2} and Jing Huang^{1,2}

1. Westlake AI Therapeutics Lab, Westlake Laboratory of Life Sciences and Biomedicine, 18 Shilongshan Road, Hangzhou, Zhejiang 310024, China; 2. Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, 18 Shilongshan Road, Hangzhou, Zhejiang 310024, China.

zhangwenyi@westlake.edu.cn

Key: Auto:N; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:N; MD:Y.

The emergence of AlphaFold¹ has elevated the accuracy and robustness of data-driven predictions for protein structures, yet the predictions for multimers and protein-ligand complexes have been severely limited by the lack of well-curated structure datasets. In CASP15, we applied multiple strategies towards different modeling categories to predict and refine structures. Starting from the initial structures obtained by AlphaFold or RoseTTAFold², we analyzed the binding sites and interactions by a synergy of template-based, machine learning-based, and *a priori* knowledge-based approaches. Specifically, we captured the ambiguous binding poses of complexes inspired by information from literature, such as homologous structures, biochemical experiments and bioinformatic analysis. Finally, all-atom explicit solvent molecular dynamics (MD) simulations were employed to refine whole structures.

Methods

Tertiary structure prediction. We obtained the initial conformations from AlphaFold (v2.2.0) in monomer mode for single proteins or in multimer mode (AlphaFold-Multimer³) for the assembly targets whose sequence length was less than 1200. The AlphaFold structures are refined by classical MD simulations with the CHARMM36m⁴ force field. System setup was performed with CHARMM⁵. MD simulations were carried out using OpenMM⁶ with force-switch scheme for van der Waals (vdW) interaction and particle-mesh Ewald (PME) method for long-range electrostatic interaction. For each target, 50-100 ns MD simulations were performed and 5 conformations from approximate converged trajectories with RMSD fluctuation < 4 Å were selected as the final submitted results. If a simulation was hard to converge, we would adjust the initial structure states based on previous studies. One example was WhiB6 in subunit 2 of target H1151. AlphaFold predicted the structure of the holo state binding with Fe₄S₄, but experiments suggested two more disulfide bonds on apo WhiB6 compared with the holo state, which might significantly affect the secondary structures⁷. This observation was consistent with our MD simulations after adding these two disulfide bonds. We also considered the effect of post-translational modifications (PTMs) on protein fold and stability. For instance, target T1154 was a highly glycosylated protein and MD simulation quickly converged if we added oligosaccharide chains on the glycosylated sites identified by experiments⁸ and sequence model inference⁹, while it kept divergent without glycosylation.

Protein-ligand complex. Our prediction pipeline consisted of four steps. Firstly, 100 ns MD simulation was performed to refine the apo structures predicted by AlphaFold. Then, templates were searched by hmmsearch and hmmbuild from HMMER suite (v3.3.2) on PDB dataset and the conserved binding residues were selected for docking based on literature. In case no binding site can be identified with template searching, EquiBind¹⁰ was utilized for blind docking and the centroid coordinate of ligand pose was assigned as the docking center. The 3D conformation of the ligand was generated by energy minimization operation in Molecular Operating Environment (MOE). The third step was to predict the poses of protein-ligand by the two approaches, MOE and EDock¹¹. The induced fit docking of MOE was performed based on physical energy. To obtain protein-ligand structure by REMC simulation, we also run a modified version of EDock. In this modified EDock, the energy term for ligand-receptor atomic distance profile was constructed by searching the pocket-ligand complex templates from BioLiP¹² dataset using PPS-align¹³ and LS-align¹⁴. Finally, 100 ns MD simulation was performed to refine the docking poses predicted by MOE and EDock. We considered the binding poses were stable and submitted in CASP 15 if the RMSD fluctuation of ligand structures was less than 4 Å. If not, the steps 2-4 were repeated.

Assembly. For homologous multimer assembly, such as H1115, the single chain structure was predicted by AlphaFold and was split into the individual domains. We predicted the assembly of an individual domain by AlphaFold-Multimer¹⁵, which usually resulted in the assembly framework of the target. Then, the single chain structure was repeated and aligned to the single-domain assembly framework by TM-align¹⁶ to construct the complete structure. For heteromers, using the interaction information predicted by Pesto¹⁷, we truncated the targets as the interacting domains within 1200 residues, such that they can be predicted and constructed by AlphaFold-Multimer. Then, the single chain structure for each subunit was aligned to the interacting domains to obtain the full-length structure. Finally, 100 ns MD simulations were performed to refine the final structures.

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583-589.
2. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., ... & Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871-876.
3. Feinberg, E. N., Joshi, E., Pande, V. S., & Cheng, A. C. (2020). Improvement in ADMET prediction with multitask deep featurization. *Journal of medicinal chemistry*, **63**, 8835-8848.
4. Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., De Groot, B. L., ... & MacKerell, A. D. (2017). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature methods*, **14**, 71-73.
5. Brooks, B. R., Brooks III, C. L., Mackerell Jr, A. D., Nilsson, L., Petrella, R. J., Roux, B., ... & Karplus, M. (2009). CHARMM: the biomolecular simulation program. *Journal of computational chemistry*, **30**, 1545-1614.

6. Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., ... & Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, **13**, e1005659.
7. Alam, M. S., Garg, S. K., & Agrawal, P. (2009). Studies on structural and functional divergence among seven WhiB proteins of *Mycobacterium tuberculosis* H37Rv. *The FEBS journal*, **276**, 76-93.
8. Peyfoon, E., Meyer, B., Hitchen, P. G., Panico, M., Morris, H. R., Haslam, S. M., ... & Dell, A. (2010). The S-layer glycoprotein of the crenarchaeote *Sulfolobus acidocaldarius* is glycosylated at multiple sites with chitobiose-linked N-glycans. *Archaea*, **2010**.
9. Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., ... & Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in bioinformatics*, **3**, 265-274.
10. Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R., & Jaakkola, T. (2022, June). Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning* (pp. 20503-20521). PMLR.
11. Zhang, W., Bell, E. W., Yin, M., & Zhang, Y. (2020). EDock: blind protein–ligand docking by replica-exchange monte carlo simulation. *Journal of cheminformatics*, **12**, 1-17.
12. Yang, J., Roy, A., & Zhang, Y. (2012). BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, **41**, D1096-D1103.
13. Hu J, Zhang Y. PPS-align. <https://zhanglab.dcmf.med.umich.edu/PPS-align/>.
14. Hu, J., Liu, Z., Yu, D. J., & Zhang, Y. (2018). LS-align: an atom-level, flexible ligand structural alignment algorithm for high-throughput virtual screening. *Bioinformatics*, **34**, 2209-2218.
15. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A. W., Green, T., ... & Hassabis, D. (2021). Protein complex prediction with AlphaFold-Multimer. *BioRxiv*.
16. Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research*, **33**, 2302-2309.
17. Krapp, L. F., Abriata, L. A., Rodriguez, F. C., & Dal Peraro, M. (2022). PeSTo: parameter-free geometric deep learning for accurate prediction of protein interacting interfaces. *bioRxiv*.

Protein and RNA structure prediction with trRosettaX2, trRosettaRNA and AlphaFold2

Wenkai Wang^{1,#}, Hong Wei^{1,#}, Chenjie Feng^{2,#}, Zongyang Du¹, Zhenling Peng² and
Jianyi Yang^{2,*}

1 - School of Mathematical Sciences, Nankai University, Tianjin 300071, China, 2 - Ministry of Education Frontiers Science Center for Nonlinear Expectations, Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao 266237, China.

#co-first authors, *corresponding author: yangjy@sdu.edu.cn

Key: *Templ:Y; MSA:Y;MetaG; Dist:Y; Tors:Y; DeepL:Y;MD:N*

In CASP15, we submitted predictions for protein and RNA structures based on trRosettaX2¹ and trRosettaRNA², respectively. The predicted structures were fed into COACH-D³ to predict protein-ligand complex structures.

Methods

Monomer structure (Yang-Server, Yang) was predicted by trRosettaX2, an improved version of trRosettaX⁴ and trRosetta^{5,6}. The AlphaFold2⁷ prediction was considered in case the trRosettaX2 prediction was not satisfactory. In trRosettaX2, we adopt the attention-based network (i.e., Evoformer) from AlphaFold2 to improve the prediction of inter-residue distance and orientations. The second step of structure realization by energy minimization is the same as in trRosetta. Multiple MSAs were generated based on HHblits and MMseqs2 against the UniRef30 and the Mgnify metagenome databases. For targets (especially virus targets) with limited homologous sequences, additional sequences from manual searching against the NCBI sequences and other virus databases are included in the MSA.

Multimer structure (Yang-Multimer, Yang) was predicted based on a revised version of AlphaFold-Multimer⁸. Two major changes: the template searching was replaced by HHsearch; MSA pairing was disabled.

Protein-ligand complex structure was predicted by the template-based approach COACH-D. Starting from the predicted receptor structures, homologous templates from the BioLiP database⁹ are obtained by structure and sequence alignment. The binding information from the template is transferred to the query receptor structure and molecular docking was applied to dock the query ligand against the receptor structure.

RNA structure was predicted by trRosettaRNA, which is the development of trRosetta for RNA structure prediction. In trRosettaRNA, structural geometries are predicted by a network (RNAformer) inspired by AlphaFold2's Evoformer. The predicted structural geometries are used to fold the 3D structure by energy minimization, similar to trRosetta.

Results

Benchmark tests show that trRosettaX2 is comparable to AlphaFold2 and outperforms RoseTTAFold¹⁰ on the CASP14 datasets. trRosettaRNA outperforms other methods on the RNA-Puzzles targets.

Availability:

<https://yanglab.nankai.edu.cn/trRosetta/>

<https://yanglab.nankai.edu.cn/trRosettaRNA>

<https://yanglab.nankai.edu.cn/COACH-D>

1. Wenkai Wang, Z.P., Jianyi Yang Approaching the AlphaFold2 accuracy with an improved 2D geometry prediction in trRosettaX2. *in preparation* (2022).
2. Feng et al Accurate de novo prediction of RNA 3D structure with transformer networks. *in preparation* (2022).
3. Wu, Q., Peng, Z., Zhang, Y. & Yang, J. COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Res* **46**, W438-W442 (2018).
4. Su, H. *et al.* Improved Protein Structure Prediction Using a New Multi-Scale Network and Homologous Templates. *Adv Sci (Weinh)* **8**, e2102592 (2021).
5. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences* **117**, 1496 (2020).
6. Du, Z. *et al.* The trRosetta server for fast and accurate protein structure prediction. *Nat Protoc* **16**, 5634-5651 (2021).
7. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
8. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *BioRxiv*, 2021.2010.2004.463034 (2022).
9. Yang, J., Roy, A. & Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res* **41**, D1096-1103 (2013).
10. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021).

Multi-MSA strategy for protein complex structure modeling

Wei Zheng^{1,2}, Qiqige Wuyun³ and Peter L Freddolino^{1,2}

1 - Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA,

2 - Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA

3 - Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

zhengwei@umich.edu, wuyunqiq@msu.edu, and petefred@umich.edu

Key: Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N

The protein multimer structure prediction of the Zheng human group in CASP15 is based on a pipeline combining DeepMSA2 with AlphaFold2-Multimer. The procedure is fully automated but the running time for some large protein complexes required more than three days. Thus, Zheng only participated in the human group.

Methods

The full pipeline contains three steps: (i) multiple sequence alignment (MSA) generation for the individual constituent proteins of the complexes by DeepMSA2 (manuscript in preparation), (ii) MSA selection for each constituent, and (iii) complex model construction and ranking by AlphaFold2-multimer¹ pipeline with DeepMSA2 constituent MSAs as input.

For the constituent proteins in the complexes, DeepMSA2 provides two sub-methods (dMSA and qMSA) that were utilized to generate seven multiple sequence alignments (MSAs). Here, dMSA is our previous MSA construction program (DeepMSA²) developed during CASP13, where HHblits³, Jackhmmer⁴ and HMMsearch⁴ are used to search the query sequence against the Uniclust30⁵, UniRef90⁶ and Metaclust⁷ databases in three stages (labeled stages 1, 2, and 3, in the order noted above). qMSA is an extended version of dMSA with a new search added between the second and third stages of dMSA, where HHblits is used to search the BFD⁸ metagenomics database. In addition, a new iteration stage (stage 4) is added in qMSA to search the query through the Mgnify⁹ metagenomics database. Thus, five different MSAs are generated by stages 1-3 of dMSA and stages 3-4 of qMSA. Furthermore, the MSA from qMSA stage 3 (the MSA from the BFD database) is used as the starting point for HMMsearch to search through the IMG/M¹⁰, Tara¹¹ and MetaSource¹² metagenome databases that contain more sequences than the Metaclust, BFD, and Mgnify databases. The resulting sequence hits are converted into a sequence database. This sequence database is then used as the target database for dMSA stage 3 and qMSA stage 4 to generate two additional MSAs. At the end of this process, seven MSAs are generated by the DeepMSA2 method for each constituent protein.

The seven MSAs thus obtained for each constituent protein of the complex modeling target are fed into AlphaFold2 monomer modeling pipeline to get seven models with associated pLDDT scores. The seven MSAs are then ranked by the associated pLDDT scores. For homo-

oligomer complexes, all seven MSAs are utilized for generating paired MSAs using a modified AlphaFold2-multimer pipeline. However, for heteromeric complexes, an additional selection procedure was used to generate an optimal set of paired MSAs based on combinations of the individual constituent MSAs. The top N ranked MSAs for each constituent protein were selected for generating potential paired MSAs. Each selected MSA for one constituent protein can be paired with the MSA of another constituent. Thus, for a heteromeric complex containing M different constituent proteins, N^M distinct paired MSAs are generated and evaluated. To guarantee that AlphaFold2-multimer modeling with N^M set of paired MSAs could be completed within three weeks, N is selected as the maximal value to satisfy $N^M \leq 100$. For example, if a complex contains three different protein components, then N would be set to 4.

In the final step of complex model generation, the selected N^M sets of MSAs are used as input to a modified AlphaFold2-multimer pipeline. For each set of MSAs, 25 models are generated. Finally, the resulting $25N^M$ complex models are ranked by the predicted TM-scores¹³, and the top five complex models are selected as the final set of models.

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589.
2. Zhang, C., Zheng, W., Mortuza, S. M., Li, Y. & Zhang, Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **36**, 2105-2112.
3. Remmert, M., Biegert, A., Hauser, A. & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9**, 173-175.
4. Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R. & Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research* **46**, W200-W204.
5. Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J. & Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research* **45**, D170-D176.
6. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H. & UniProt, C. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)* **31**, 926-932.
7. Steinegger, M. & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications* **9**, 2542.
8. Steinegger, M., Mirdita, M. & Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods* **16**, 603-606.
9. Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A. & Finn, R. D. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research* **48**, D570-D578.

10. Chen, I. M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Huntemann, M., Varghese, N., White, J. R., Seshadri, R., Smirnova, T., Kirton, E., Jungbluth, S. P., Woyke, T., Elie-Fadrosh, E. A., Ivanova, N. N. & Kyrpides, N. C. (2019). IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research* **47**, D666-D677.
11. Wang, Y., Shi, Q., Yang, P., Zhang, C., Mortuza, S. M., Xue, Z., Ning, K. & Zhang, Y. (2019). Fueling ab initio folding with marine metagenomics enables structure and function predictions of new protein families. *Genome Biology* **20**, 229.
12. Yang, P., Zheng, W., Ning, K. & Zhang, Y. (2021). Decoding the link of microbiome niches with homologous sequences enables accurately targeted protein structure prediction. *Proceedings of the National Academy of Sciences* **118**, e2110828118.
13. Zhang, Y. & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **57**, 702-710.

A template-guiding and docking strategy for protein-ligand binding mode prediction in CASP15

Xianjin Xu[†], Rui Duan[†], Xiaoqin Zou^{*}

Dalton Cardiovascular Research Center, Department of Physics and Astronomy, Department of Biochemistry, Institute for Data Science and Informatics, University of Missouri, Columbia, Missouri 65211, United States

[†]equal contribution; ^{*}corresponding author: zoux@missouri.edu

A novel template-guiding strategy^{1,2} was employed for protein-ligand complex structure predictions in CASP15, which allows for the use of both similar and dissimilar ligands as templates via a newly developed intercomparison method, followed by local optimization and ranking with a hybrid scoring function, in addition to the standard docking protocol. Specifically, for each query target, the protein structure was built with AlphaFold v2.2.0³ and an ensemble of ligand 3D conformers were generated from the SMILES string using the OMEGA2 program (Version 3.0.1.2, OpenEye Scientific Software, Santa Fe, NM, USA. <http://www.eyesopen.com>)^{4,5}. Then, the Protein Data Bank⁶ was searched for template structures containing the target protein or its homologies. If a template structure contains one or more co-bound ligands, the target protein structure was superimposed onto the protein structure in the template with the MatchMaker tool in UCSF Chimera⁷ and the conformers of the query ligand were matched to the co-bound ligands in the template using a 3D molecular similarity measurement program, SHAFTS⁸. Then, the superimposed protein structure and the matched ligand conformers were combined, and local minimization was performed by AutoDock Vina⁹ (with option “local_only”). After that, the predicted complex structures were ranked by a hybrid scoring function¹, which combines a protein-ligand binding score (the AutoDock Vina score) and a 3D similarity score (SHAFTS score, characterizing the 3D similarity between the query ligand and the co-bound ligand in the template). Finally, the top 10 predicted complex structures were manually inspected for electrostatic, polar, and nonpolar interactions, and 5 models were selected for further optimization (as described at the end).

For the cases in which either no templates or only low-quality templates (i.e., the co-bound ligands shared low similarity with the query ligands below the cutoff threshold) were found, a standard molecular docking strategy was employed. First, the binding site information about the query ligand was searched from the homologous proteins in the Protein Data Bank. Local dockings were performed with AutoDock Vina [9] (using the default parameters) for the cases with known binding sites. For the cases without any information about the binding location, global docking was performed using AutoDock Vina by increasing the exhaustiveness value (e.g., from the default 8 to 100 independent runs). For both local and global dockings, the protein structure was treated as a rigid body, and the ligand structure was treated to be fully flexible. Finally, the top 10 binding modes were generated and manually examined, and 5 models were selected for further optimization.

For both the template-guiding method and the molecular docking method, the final 5 selected models were further optimized with a simple force-field minimization of the energy in Maestro, Version 12.9.137 (Schrödinger, LLC)¹⁰, and then submitted to CASP15.

1. Xu, X.; Zou, X. Dissimilar Ligands Bind in a Similar Fashion: A Guide to Ligand Binding-Mode Prediction with Application to CELPP Studies. *Int. J. Mol. Sci.* 2021, 22: 12320.
2. Xu, X.; Ma, Z.; Duan, R.; Zou, X. Predicting protein–ligand binding modes for CELPP and GC3: Workflows and insight. *J. Comput.-Aided Mol. Des.* 2019, 33: 367-374.
3. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021, 596(7873):583-589.
4. Hawkins, P.C.; Skillman, A.G.; Warren, G.L.; Ellingson, B.A.; Stahl, M.T. Conformer generation with omega: Algorithm and validation using high quality structures from the protein databank and Cambridge structural database. *J. Chem. Inf. Model.* 2010, 50: 572-584.
5. Hawkins, P.C.; Nicholls, A. Conformer generation with OMEGA: Learning from the data set and the analysis of failures. *J. Chem. Inf. Model.* 2012, 52: 2919-2936.
6. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* 2000, 28: 235-242.
7. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* 2004, 25: 1605-1612.
8. Liu, X.; Jiang, H.; Li, H. SHAFTS: A hybrid approach for 3D molecular similarity calculation. 1. Method and assessment of virtual screening. *J. Chem. Inf. Model.* 2011, 51: 2372-2385.
9. Trott, O.; Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 2010, 31: 455-461.
10. Schrödinger Release 2021-3: Maestro, Schrödinger, LLC, New York, NY, 2021.

CASP-RELATED PUBLICATIONS FROM NON-PARTICIPANTS

The ResiRole server provides automated assessments of structure models presented in CAMEO using functional site predictions and has demonstrated applicability to CASP14 SARS-2-CoV Targets

William A. McLaughlin¹, Thomas K. Parry¹, Geoffrey Huang¹, Joshua M. Toth¹
and Jürgen Haas²

1 - Department of Medical Education, Geisinger Commonwealth School of Medicine; 525 Pine Street, Scranton PA 18509, 2 - Biozentrum, University of Basel and SIB Swiss Institute of Bioinformatics, Klingelbergstrasse 50/70, CH - 4056 Basel

wmclaughlin@som.geisinger.edu

Key: *Auto:Y; EMA:Y*

The Continuous Automated Model Evaluation (CAMEO¹) platform presents results of protein structure predictions from hosted structure prediction servers for prelease sequences in the Protein Data Bank (PDB)². We have developed a method to evaluate the quality of structural models available through CAMEO based on their abilities to have SeqFEATURE functional site predictions like those at corresponding sites in the reference structures³.

Methods

The ResiRole algorithm calculates the average difference scores per structure prediction technique and per structure model⁴. Each difference score is defined as the absolute difference in cumulative probability of functional site prediction in the reference structure versus that at the corresponding site in the structure model. Difference scores are averages across the results obtained using different functional site prediction models.

Results

Results are accessible according to defined intervals in which models and reference structures have been made available in CAMEO and the PDB, respectively. Results are further delineated based on target difficulty according to IDDT score ranges. To expand the utilities of the ResiRole server, we are developing automated routine updates to evaluate models in CAMEO as they become available on a weekly basis. We have further applied the ResiRole algorithm to SARS-CoV-2 protein targets in CASP14⁵ and we found that the ResiRole method has the capability to detect differences in quality estimations for the first and second attempts by the same structure prediction group. Further, the average quality estimates for structure predictions made by the different structure prediction groups provides a means to further estimate average accuracy of the structure prediction methods. Average IDDT scores and difference scores for the different structure prediction methods were found to correlate, and possible outliers help to demonstrate the utility of the difference score measure.

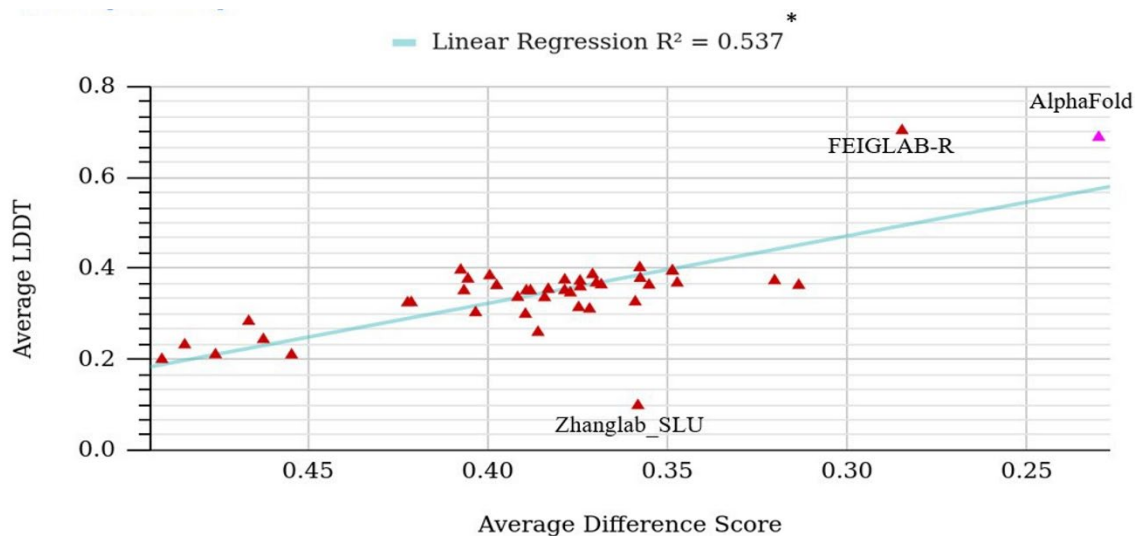


Figure 1. The figure depicts a scatter plot of the average difference scores compared to the average LDDT scores across all SeqFEATURE models for all protein model submission attempts. Each point corresponds to a unique protein prediction group (not all group names shown). Average difference score (ADS) is calculated as $|\text{Probmodel} - \text{Probtarget}|$. Note the descending values of the x-axis, with a lower ADS corresponding to a better model. Three relative outlier group names are shown.

Availability

The ResiRole server is available at the URL <https://protein.som.geisinger.edu/ResiRole/>.

1. Robin, X., *et al.* Continuous Automated Model EvaluatiOn (CAMEO)—Perspectives on the future of fully automated evaluation of structure prediction methods. *Proteins: Structure, Function, and Bioinformatics* 2021;89(12):1977-1986.
2. Berman, H.M., *et al.* The Protein Data Bank. *Nucleic Acids Res* 2000;28(1):235-242.
3. Wu, S., Liang, M.P. and Altman, R.B. The SeqFEATURE library of 3D functional site models: comparison to existing methods and applications to protein function annotation. *Genome biology* 2008;9(1):1.
4. Toth, J.M., *et al.* ResiRole: residue-level functional site predictions to gauge the accuracies of protein structure prediction techniques. *Bioinformatics* 2021;37(3):351-359.
5. Kryshchuk, A., *et al.* Critical assessment of methods of protein structure prediction (CASP) - Round XIV. *Proteins: Structure, Function, and Bioinformatics* 2021;89(12): 1607-1617.

ProFold – Quantum Computing at Davis

Sakib Sazzad, Anurag Kulkarni, Rishab Ghosh, SphoorthyNadimpalli, John Long,
Shuhul Mujoo, Kirk McGregor, Samarth Sandeep

Quantum Computing at Davis

Finding effective folds is most effectively done for a variety of solvents using molecular dynamics, and specifically combined Quantum Mechanics / Molecular Mechanics (QM/MM) methods that utilize both classical force fields and quantum electronic basis sets. However, scaling these solutions for effective, real-time folding determination can be time consuming, with CP2K requiring 14ps/day to complete dynamics for a 1536 atomic box¹. Quantum devices could potentially improve these values by being able to complete large quantum stage operations in a smaller time frame². Here, we develop a framework for protein structure determination utilizing quantum devices combined with STO-3G basis sets and GROMACS CHARMM27 classical modeling; this has allowed us to develop hybrid quantum compute runtimes for molecular dynamics on commercially-available GPUs.

Methods

The method used here is a separated QM/MM process, wherein all MM precede QM. Monomers are created from FASTA using OpenBabel 2.4.1's FASTA to PDB conversion. Then, classical molecular dynamics is completed with GROMACS running CHARMM27 with TIP5P. Next, the user decides which of two processing pathways to implement this quantum mechanics modeling that is to be completed on the GROMACS output: it is either transformed using Givens rotations to directly map STO-nG basis sets onto qubits³, or it is transformed using phase changes from a Quantum Fourier Transform (QFT)⁴ running on qubits to approximate STO-nG basis sets. All code is written in PennyLane with Tensorflow as the backend in order to promote easy testing on NVIDIA GPUs, Xanadu quantum photonic devices, and other superconducting quantum devices. Finally, Fast Fourier Transform (FFT) in Tensorflow transforms the coordinates of each protein, for which the energy from the outputs of VQE/QFT are input to this FFT-transformed structure (i.e., energy gradient?), and inverse Fast Fourier Transform (IFFT) transforms these energy changes into coordinates and places them into PDBs using Biopython. Final PDBs can compare a distance-based metric TM Score using the Zhang Group server (<https://zhanggroup.org/TM-score/>) and a topological metric (Betti numbers) that can be evaluated using Ripser++ (<https://github.com/simonzhang00/riper-plusplus>).

Results

Results for the entire pipeline are still forthcoming. Current updates can be found at <https://github.com/QC-at-Davis/ProFold/>.

1. Sedova A, Davidson R, Taillefumier M, Elwasif W. HPC Molecular Simulation Tries Out a New GPU: Experiences on Early AMD Test Systems for the Frontier Supercomputer. Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States); 2022 Jun 1.
2. Madsen LS, Laudenbach F, Askarani MF, Rortais F, Vincent T, Bulmer JF, Miatto FM, Neuhaus L, Helt LG, Collins MJ, Lita AE. Quantum computational advantage with a programmable photonic processor. *Nature*. 2022 Jun;606(7912):75-81.
3. Streif M, Malone F, Parrish R, Welden A, Fox T, Degroote M, Kyoseva E, Moll N, Santagati R. Towards the Simulation of Large Scale Protein-Ligand Interactions on NISQ-era Quantum Computers. *Bulletin of the American Physical Society*. 2022 Mar 14.
4. Aspuru-Guzik A, Dutoi AD, Love PJ, Head-Gordon M. Simulated quantum computation of molecular energies. *Science*. 2005 Sep 9;309(5741):1704-7.

Comparative Recall (CR) Analysis for Assessment of Protein Structure Models Against Experimental NMR Data: Characterizing Multiple Conformational States

Yuanpeng Janet Huang, Theresa A. Ramelot, Laura Spaman, Roberto Tejero, and Gaetano T. Montelione

*Center for Biotechnology and Interdisciplinary Sciences, and Department of Chemistry and Chemical Biology
Rensselaer Polytechnic Institute, Troy, NY 12180 USA.*

Correspondence may be addressed to: monteg3@rpi.edu

Recent advances in protein structure prediction provide models with accuracies that rival models generated directly from experimental data¹⁻⁸. We have previously described a metric based on information retrieval statistics, the RPF-DP score^{9,10}, for assessing protein structure models against experimental NMR NOESY and chemical shift data. Models generated without sample-specific experimental data using advanced deep learning methods, such as AlphaFold (AF), often exhibit RPF-DP scores, residual dipolar coupling (RDC) quality scores, and other structure quality scores similar to, and sometimes even better than, models reported by experimental NMR groups using standard structure generation methods^{1,4,8}. This is observed even without the use of homologous protein structure templates, and for proteins not available in the machine learning training data.

Methods

We have developed structure analysis metric related to the RPF-DP score, which we call “comparative recall” (CR) analysis. The CR metric assesses a pair of protein structure models against experimental NOESY peak list and chemical shift data, and identifies NOESY peaks that, considering any possible NOESY peak assignment consistent with the chemical shift data, can be explained by the model. In this analysis, “recall violations” for a given model are the experimental NOESY data that are inconsistent with the model, as previously described^{9,10}. Comparing a pair of structures (e.g. an AF model and an experimental NMR model), the CR analysis allows identification of the experimental NOESY data supporting both models, and the specific data supporting only one or the other model. The CR analysis identifies the locations in the models where the data better fit one model rather than the other, and can provide evidence for cases where both models are represented by the solution NMR data; e.g. conformations in dynamic conformational equilibria.

Results

Here we demonstrate the application of CR analysis in multiple scenarios. In the first case, CR analysis reveals that the NMR data equally-well fit NMR, X-ray crystal, and AF models of a target protein structure. This is by far the most common scenario encountered in our studies. In the second case, the CR analysis reveals that the NMR data equally-well support the AF and experimental NMR models, but is partially violated by the corresponding X-ray crystal structure. Detailed structural analysis suggests that the underlying structural differences may be attributed to the differences in pH used in the NMR and X-ray crystallography studies. In a third case, the

NMR data better support an AF model of a target structure rather than the corresponding experimental NMR structure. In a fourth case, the experimental data are not fully consistent with either the AF or experimentally-reported NMR model, but rather suggest a dynamic conformational exchange between these two conformations in solution. The “comparative recall” analysis provides an important tool in our ongoing efforts to use protein structure prediction to guide analysis of experimental NMR data in terms of protein structure and dynamics.

Availability: <https://github.rpi.edu/RPIBioinformatics/ComparativeRecall>

1. Sala, D., Huang, Y.J., Cole, C.A., Snyder, D.A., Liu, G., Ishida, Y., et al. (2019). Protein structure prediction assisted with sparse NMR data in CASP13. *Proteins* 87(12), 1315-1332. doi:<https://doi.org/10.1002/prot.25837>.
2. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. doi:<https://doi.org/10.1126/science.abj8754>.
3. Fowler, N.J., and Williamson, M.P. (2022). The accuracy of protein structures in solution determined by AlphaFold and NMR. *Structure* 30(7), 925-933 e922. doi:<https://doi.org/10.1016/j.str.2022.04.005>.
4. Huang, Y.J., Zhang, N., Bersch, B., Fidelis, K., Inouye, M., Ishida, Y., et al. (2021). Assessment of prediction methods for protein structures determined by NMR in CASP14: Impact of AlphaFold2. *Proteins* 89(12), 1959-1976. doi:<https://doi.org/10.1002/prot.26246>.
5. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*. doi:<https://doi.org/10.1038/s41586-021-03819-2>.
6. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moulton, J. (2021). Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* 89(12), 1607-1617. doi:<https://doi.org/10.1002/prot.26237>.
7. Pereira, J., Simpkin, A.J., Hartmann, M.D., Rigden, D.J., Keegan, R.M., and Lupas, A.N. (2021). High-accuracy protein structure prediction in CASP14. *Proteins*. doi:<https://doi.org/10.1002/prot.26171>.
8. Tejero, R., Huang, Y.J., Ramelot, T.A., and Montelione, G.T. (2022). AlphaFold models of small proteins rival the accuracy of solution NMR structures. *Front Mol Biosci* 9, 877000. doi:<https://doi.org/10.3389/fmolb.2022.877000>.
9. Huang, Y.J., Rosato, A., Singh, G., and Montelione, G.T. (2012). RPF: a quality assessment tool for protein NMR structures. *Nucleic Acids Research* 40(Web Server issue), W542-546. doi:<https://doi.org/10.1093/nar/gks373>.
10. Huang, Y.J., Powers, R., and Montelione, G.T. (2005). Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *Journal of the American Chemical Society* 127(6), 1665-1674. doi:<https://doi.org/10.1021/ja047109h>.