CAMEO - Perspectives on the future of fully automated evaluation of structure prediction methods

Xavier Robin[1], Juergen Haas[1], Rafal Gumienny[1], Gerardo Tauriello[1], and Torsten Schwede[2]

[1]Swiss Institute of Bioinformatics
[2]University of Basel

June 1, 2021

Abstract

The Continuous Automated Model EvaluatiOn (CAMEO) platform complements the biennial CASP experiment by conducting fully automated blind evaluations of 3D protein prediction servers based on the weekly pre-release of sequences of those structures, which are going to be published in the upcoming release of the Protein Data Bank (PDB). While in CASP14 significant success was observed in predicting the structures of individual protein chains with high accuracy, significant challenges remain in correctly predicting the structures of complexes. By implementing fully automated evaluation of predictions for protein-protein complexes, as well as for proteins in complex with ligands, peptides, nucleic acids, or proteins containing non-canonical amino acid residues, CAMEO will assist new developments in those challenging areas of active research.

Introduction

The 2020 CASP14 experiment saw an unprecedented improvement in the performance of 3D protein structure prediction. One method (AlphaFold2) was able to generate highly accurate predictions even for the most challenging *de novo* targets. Beyond the CASP community, this breakthrough has implications for the entire field of structural biology: accurately predicting the structure of a single protein chain has never been closer to being considered a solved problem. But far from being the end of structure prediction, this might instead be the beginning of a new era in the 3D modeling of biomolecular structures. Areas that have been limited so far due to the inability to produce sufficiently accurate *de novo* protein models in the first place, such as the prediction of protein-ligand interactions, large macromolecular complexes and assemblies, or variant effects, might now be within reach of the next generation of structural prediction methods. Independent blind assessment of these techniques will be more than ever required in order to support the development of reliable and reproducible methods. In order to assist the community to tackle those challenges, we are introducing an extension of CAMEO (available at beta.cameo3d.org) with the aim to shift the focus from the prediction of individual protein chains to the prediction of macromolecular complexes as determined experimentally by X-ray crystallography or increasingly cryo-EM techniques and deposited to the PDB (wwPDB consortium et al. 2018).

In this new CAMEO category, participating methods receive the sequences of all unique polymer chains, as well as the InChI codes of non-polymer entities composing the complex as prediction targets. The challenges of the modeling task are to: 1) predict the stoichiometry of the complex; 2) predict the 3D structure of all the components: proteins, peptides, DNA, RNA and ligands, including their orientation and interfaces; and 3) provide per-residue confidence estimates of the model. This CAMEO category is based on an opt-in model: participants only receive the target type(s) their method is able to model. This means that a method that only predicts single protein chains can still participate and will receive the targets composed of only

one protein sequence, which can be either monomers or homo-oligomers, while another method by the same group might be designed to predict e.g. complexes of proteins with drug-like small molecules.

In this manuscript, we describe the different types of prediction targets that CAMEO enables in the new category, and estimate the number of expected validation targets for each category based on PDB statistics observed in 2020. One major challenge will be the scoring of the new type of predictions with regard to the actual experimental structures. Wherever appropriate, we comment on scores that are foreseen to be applied to the various prediction types. We are welcoming feedback from the community regarding complementary scoring approaches.

Material and Methods

Sequence filtering and clustering

The pre-release sequences of polymer entities as well as InChI code of non-polymer ligands were downloaded every Saturday from the PDB (wwPDB consortium et al. 2018) (*http://www.wwpdb.org/files/*). Structures containing sequences with unknown residues, starting with caps, or whose type (protein, DNA or RNA) couldn't be assigned unambiguously were discarded. Within a pre-release week, amino acid sequences of 30 amino acid residues or longer ("protein") were clustered with CD-HIT (Li and Godzik 2006) applying a 99% sequence identity threshold. Amino acid sequences of less than 30 amino acid residues ("peptides"), as well as DNA and RNA sequences were clustered based on exact identity (100%). One representative sequence per cluster was selected as target for structure prediction.

Template searches

Target protein sequences were submitted to two template searches. First, a BLAST+ v. 2.2.31 (Camacho et al. 2009) search against a database of current PDB entries at the time of pre-release was performed. A threshold of 85% sequence identity and at least 70% coverage was used to identify target sequences with very high similarity to a protein with known structure. Next, sequence profiles were built using 1 iteration of HHblits v. 3.2.0 (Steinegger et al. 2019) against Uniclust30 (2018_08) (Mirdita et al. 2017). The profiles were used to search a database of PDB entries available on 2021-03-19, with an HHblits probability threshold of 70% and a coverage threshold of 70% in order to identify target sequences with more remote similarity to a protein with a known structure. Since this was done as a retrospective analysis, hits that were released after the date of the pre-release of the target were filtered out. For peptide sequences of less than 30 amino acid residues and sequences of nucleic acid residues, a lookup was performed against a database of current PDB entries at the time of pre-release with a 100% identity threshold.

Templates found by BLAST, HHblits and lookup on single chains were aggregated into complexes. A structure was considered to be a template if all the chains of the target structure could be uniquely mapped to the chains of the template structure, and the template structure didn't contain any extra polymer chain.

Scores

Single-chain predictions were evaluated against the reference structure with the lDDT score (Mariani et al. 2013) using OpenStructure v. 2.1.0 (Biasini et al. 2013), the global CAD atom-atom (AA) score v. 1646_-63d6b800098c (K. Olechnovič, Kulberkytė, and Venclovas 2013), and the GDT_TS score using LGA v. 05/2009 (Zemla 2003). When the target structure contained more than one copy of the sequence, more than one biological assembly, or for homo-oligomeric predictions, the scores were calculated between all possible combinations of target assembly and target and model chains, and only the most favorable score was kept.

Homo- and hetero-oligomeric predictions were evaluated with the oligo-lDDT and QS-score (Bertoni et al. 2017) using OpenStructure v. 2.1.0 (Biasini et al. 2013), as well as the MM-align-based TM-score v.

20190426 (Mukherjee and Zhang 2009). The oligomeric lDDT score (oligo-lDDT) is an extension of the lDDT score for protein complexes and has also been used in CASP since CASP13 (Guzenko et al. 2019; Kryshtafovych et al. 2019). It relies on the QS-score to identify the mapping of chains and residues between the model and target structure. Once the mapping is identified, the all-atom lDDT score can be applied on the protein complex in the same way as it is applied for single chains with the advantage that it now also considers inter-chain contacts. Extra atoms in the model for mapped chains have no effect on lDDT scores, while extra atoms in the target structure reduce the score. For the oligomeric lDDT score, we penalize extra chains in both reference and model by including them as non-conserved contacts.

Ligand analysis

Functional domain annotation was extracted from CATH (Sillitoe et al. 2021) version 4.3.0. We used the "Structure external" links from DrugBank (Wishart et al. 2006) version 5.1.8 to identify drug-containing targets. The analysis was performed with Python 3.6.6, OpenStructure v. 2.1.0 (Biasini et al. 2013), and pandas v. 1.1.5 (McKinney 2010).

Structure visualization

Figures were generated with the Mol* Viewer (Sehnal et al. 2021).

Results and Discussions

Current CAMEO results

Since 2012, CAMEO has been leveraging the pre-release of structures to be published in the upcoming release of the PDB Protein Data Bank to conduct weekly, blind, fully automated benchmarking experiments. Every Saturday, we download the pre-release data, which contains the sequences of polymer entities, as well as InChI codes of non-polymer entities contained in the PDB structures to be released on the following Wednesday. We selected a set of 20 interesting protein modeling targets which were submitted to registered participants, who have 4 days to predict the 3D structure of those targets. We collect those predictions and, upon release of the structures by the PDB on Wednesdays, compare the predictions with the experimental ground truth.

The CAMEO evaluation provides a wide variety of scores measuring different aspects of protein structure prediction accuracy, and accordingly does not establish a single unique ranking between the methods. However, some of the scores are featured more prominently on the web site, as we consider them more useful estimations of the model quality. The focus of CAMEO has always been on all-atom scores to capture the ability of participants to accurately model proteins including biologically relevant protein side chain conformations. In addition, as CAMEO is a fully automated workflow without human intervention, we have been focusing on superposition-free scores which alleviate the need to manually split proteins into evaluation units (Kinch et al. 2019, 2011) to account for domain movements. Therefore, CAMEO has been showcasing scores like lDDT (Mariani et al. 2013) and CAD-score (K. Olechnovič, Kulberkytė, and Venclovas 2013), both of which are all atom scores and superposition independent. In addition, our server summary page features the lDDT-BS score which measures the accuracy of predictions in the region of ligand binding sites, as well as a measure for model confidence, which evaluates the ability of participants to estimate the accuracy of their own predictions. Additional scores are displayed on the target details page and available in the downloads.

Since 2016, CAMEO (Jürgen Haas et al. 2018) has been evaluating the ability of modeling servers to correctly predict the oligomeric state of a target protein and model the correct assembly, based solely on the amino acid sequence. As targets are submitted as a single protein sequence, participants need to predict whether the protein is likely to assemble into a homo-oligomer and, if that's the case, to predict the exact stoichiometry as well as the correct interfaces. The complex models are evaluated with the oligo-lDDT score (Juergen

Haas et al. 2019), which is a modified version of lDDT that looks at the whole complex and accounts for missing or extra chains; the MM-align-based (Mukherjee and Zhang 2009) TM-score and RMSD, which are superposition-dependent; and the QS-score (Bertoni et al. 2017), which looks specifically at the conservation of interface residues.

In 2020, we performed 52 prediction rounds and provided targets to 15 public modeling servers (from 9 groups) and 25 development servers (from a total of 18 groups). After filtering problematic targets of low or uncertain quality, or targets causing technical issues to scoring tools for formatting reasons, we evaluated and scored 812 targets, 453 of which were oligomeric. Compared with 84 3D modeling targets of CASP14, CAMEO enables participants to accurately assess the accuracy of their prediction servers on a wide variety of targets in much shorter time intervals.

Protein Complexes

With the new version of CAMEO we are extending the scope of the assessment to structures and complexes. Instead of considering every protein sequence separately, a prediction target is now defined as a complete experimental structure with all the chemical entities it contains. In the case of monomeric and homo-oligomeric protein entries, this would be identical to the current CAMEO-3D targets and contain only one unique protein sequence. However, for hetero-oligomeric targets, evaluation is only performed in the context of the whole complex, and no longer as individual iosolated protein chains taken out of context. Methods registered to receive hetero-oligomeric complexes as targets thus receive all sequences of the proteins that form a complex, and are expected to predict the oligomeric structure of the complex. All participating methods receive the sequences of monomeric or homo-oligomeric targets. This allows establishing a common baseline where all participating servers can be compared with each other on a subset of common targets.

In order to select interesting targets for this category, we search for the presence of homologous complexes (Figure 1). Closely related homologs are first identified with BLAST for every protein sequence with 30 or more amino acid residues separately. Complexes containing DNA, RNA, or peptide sequences shorter than 30 amino acids are excluded at this stage, and handled separately (see following sections). For every target, we consider the complete set of proteins that compose it, and search for a homologous template that covers all the protein entities. We ignore templates that only cover some of the target sequences, or that contain extra polymer entities (proteins, peptides, DNA or RNA). We consider targets to be interesting if such a closely related homologous complex cannot be found. This includes cases of novel complexes (where all the proteins can be modeled separately easily, but where the complex has never been observed experimentally in its entirety, and therefore the interface(s) is unknown) or if at least one of the protein sequences in the complex is a non-trivial modeling target on its own.

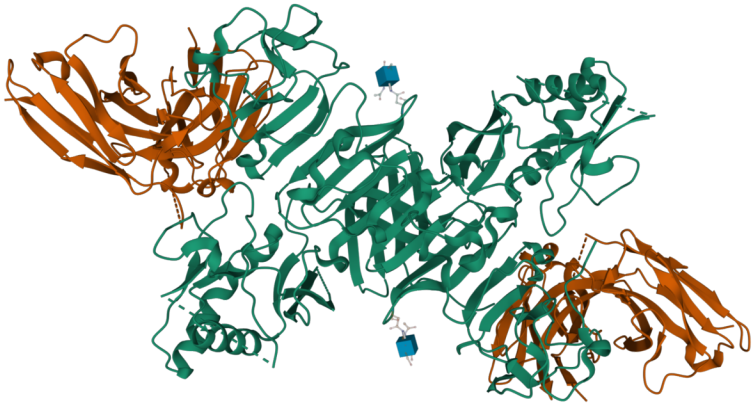

Figure 1: Target 2020-12-19_00000231 (PDB ID 7K93) is a hetero-2-2mer protein complex of a Dengue virus non-structural protein (NS1) (green) in complex with a mouse neutralizing single chain Fab variable region (orange) (Biering et al. 2021). While templates can be easily identified with HHblits for both entities, there is no overlap between the template lists, meaning the two proteins have never been observed in a homologous complex. Specifically, no homologs of this Dengue virus protein have been observed in complex with an antibody. This therefore constitutes an interesting target for the modeling of heteromeric protein complexes.

Looking at the data we collected in the 52 pre-release weeks of 2020, 3158 interesting protein structures where no closely related homolog could be found with BLAST were released by the PDB. Among those, 1017 were monomers, 1011 homo-oligomeric complexes (which can't be distinguished from monomers from the sequence-only pre-release data) and 1130 were hetero-oligomers.

In order to retrospecitvely analyse the complexity of the hetero-oligmeric target set, we repeated the template search with HHblits to identify more remotely related homologous complexes. We could identify a homologous hetero-oligomeric complex with HHblits for 565 of these 1130 targets, where all entities of the target could be uniquely mapped to the template, and reciprocally. In 240 hetero-oligomeric complexes, templates for individual entities could be identified with HHblits, but not in the same complex (or the template contained extra entities); and 113 complexes could similarly be identified with BLAST. These 353 "novel complex" targets are of particular interest, as an accurate prediction would have to successfully predict the assembly mode of the complex, and accurately model the (unknown) interfaces, therefore going beyond the classical reach of homology modeling. Finally for the remaining 212 complexes, no template could be identified by HHblits for at least one of the target entities.

HHblits was able to identify homologs in the vast majority (1734) of the 2028 monomeric (1017) or homo-oligomeric (1011) interesting protein structures contained in the CAMEO target set. We note, however, that 43 of the targets could only be mapped to templates in complex with a different partner. The interfaces are likely to differ from the templates, and therefore we consider these targets as interesting modeling targets for CAMEO. Finally HHblits was unable to identify a template for 294 of these targets.

In order to evaluate the predictions, we are using the same scores as for the homo-oligomers: oligo-lDDT, QS-score and TM-score. In addition, other single-chain scores can be generalized to evaluate heteromers in the same fashion as the oligo-lDDT score is a generalization of the lDDT score to oligomers. Finally we are also looking at the applicability of the scores used by the CAPRI community for automated evaluation.

It should be noted that the selection of interesting protein target structures is performed regardless of ligand contents, but non-polymer ligands are submitted nonetheless to participating servers that support it. 76% of the structures released by the PDB in 2020, and 65% of the interesting protein structures selected in this category, contain at least one ligand. In addition, we are considering specifically selecting interesting ligand modeling targets, which we describe in the following section.

Non-polymer Ligands

Small chemical compounds which are not part of a polymer chain are provided as InChI codes and PDB chemical components in the pre-release of the PDB. They are included in the target definition together with the polymer entities for participating servers that support predicting small chemical compounds in complex with proteins. Consequently, in addition to predicting the correct protein structure, predictors are challenged to include the ligands in their models at the correct binding site in an accurate conformation.

However predicting the exact pose of a ligand within a theoretical model remains a challenge which is out of reach for most current protein prediction servers. To specifically facilitate the development of such methods, these should be evaluated separately to the prediction of protein complexes. Therefore we are proposing a specialized CAMEO category, where easy protein modeling targets (as per the opposite of the definition in the previous section) are selected if they contain novel ligands that haven't been seen in a template.

We analyzed the feasibility of this approach on the current data in the PDB. In 2020, we observed 4870 protein targets that would be trivial to solve with comparative modeling but included a combination of non-polymer ligands never seen before in a template for those structures. Furthermore, 4486 of them contained only homo-oligomeric or monomeric targets, which would enable many current protein structure prediction servers to participate without having to implement new modeling approaches for protein complexes.

Interestingly, 3491 of these 4870 structures contained a known drug from DrugBank (Wishart et al. 2006). Figure 2 shows a typical example of such a target, the SARS-CoV-2 main protease in complex with Boceprevir, an FDA-approved drug for the treatment of the hepatitis C virus (Fu et al. 2020). Drug repurposing studies are common in the PDB, and the CAMEO target set is therefore representative of current areas of active research and can help developers to assess the performance of their methods on relevant datasets. For instance 149 DrugBank drug-containing ligand modeling targets were identified by CATH as containing the 3CL-PRO main protease domain 3 (CATH ID 1.10.1840.10), and an additional 70 targets had ligands not known to DrugBank.

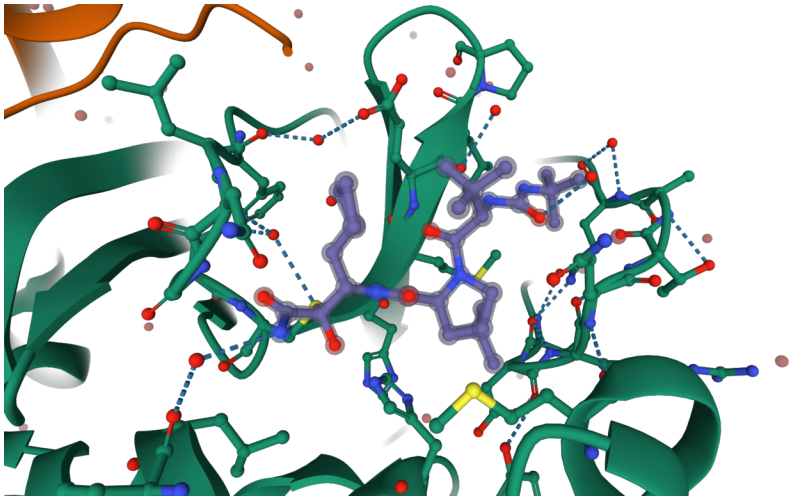

Figure 2: Target 2020-05-09_00000305 (PDB ID 7BRP) is a structure of the SARS-CoV-2 main protease in complex with Boceprevir (Fu et al. 2020). At the time of pre-release, the structure of the protease had already been solved, and was therefore a trivial modeling target on its own. However it had not been observed in complex with Boceprevir, and therefore the complex was deemed interesting for ligand modeling.

To score these predictions, we will follow the procedure developed by the CELPP community, and evaluate ligand poses with a symmetry-corrected RMSD (Wagner et al. 2019) .

Peptides

Accurately predicting the structures of short proteins or peptides has always been challenging for comparative modeling. As a consequence, many protein prediction servers have limits on the minimal length of protein sequences that they attempt to predict. CAMEO has so far taken a conservative approach and submitted targets containing at least 30 amino acids to the participants. In the future, participants will be able opt-in to also receive peptides with less than 30 residues as targets. These targets are relevant in areas of research such as for instance host-pathogen interactions.

In order to identify interesting novel targets, we considered a conservative cut-off of 100% sequence identity to a template. In 2020, the PDB released 536 novel structures containing at least one amino acid sequence of less than 30 residues in 2020. In 453 structures, such peptides were in complex with a protein or DNA/RNA, making those structures suitable for instance for peptide-protein docking methods. In 83 structures, the

peptides were observed in monomeric or homo-oligomeric forms, mainly with NMR. Advances in AI and *de novo* modeling technologies may very well make it feasible to predict the structure of those peptides.

The interface (QS-score) and complex (oligo-lDDT) scores can be used to score protein-peptide complexes. However additional scores like those used in the CAPRI experiment (Lensink et al. 2020), and others geared towards protein-peptide docking, will also be considered.

DNA and RNA

Predicting the 3D structure of nucleic acids remains a challenge. To the best of our knowledge, no fully automated prediction server is publicly available, although several standalone approaches have been published. (Wirecki et al. 2020; Orengo et al. 2020; Miao et al. 2020)

Considering a conservative cut-off of 100% sequence identity with previously known structures to identify interesting novel targets, 323 new structures containing RNA were released by the PDB in 2020, and 390 containing DNA. Most of them were in complex with proteins, and only 42, respectively 57 targets contained only nucleic acids. This low number of modeling targets might prove a challenge for blind benchmarking of nucleic acid structure prediction methods.

The CAD-score was reported to be an appropriate score to evaluate DNA and RNA predictions (Kliment Olechnovič and Venclovas 2014). Other all-atom scores are also being considered.

Mixed Complexes

Finally, CAMEO can submit targets containing a mixture of all of the above: complexes with proteins, peptides, nucleic acids and ligands (Figure 3). While this prediction task is to date extremely challenging for most methods, we believe it should be the ultimate goal in 3D structure prediction: the ability to predict any biologically relevant macromolecular structure, regardless of its composition.

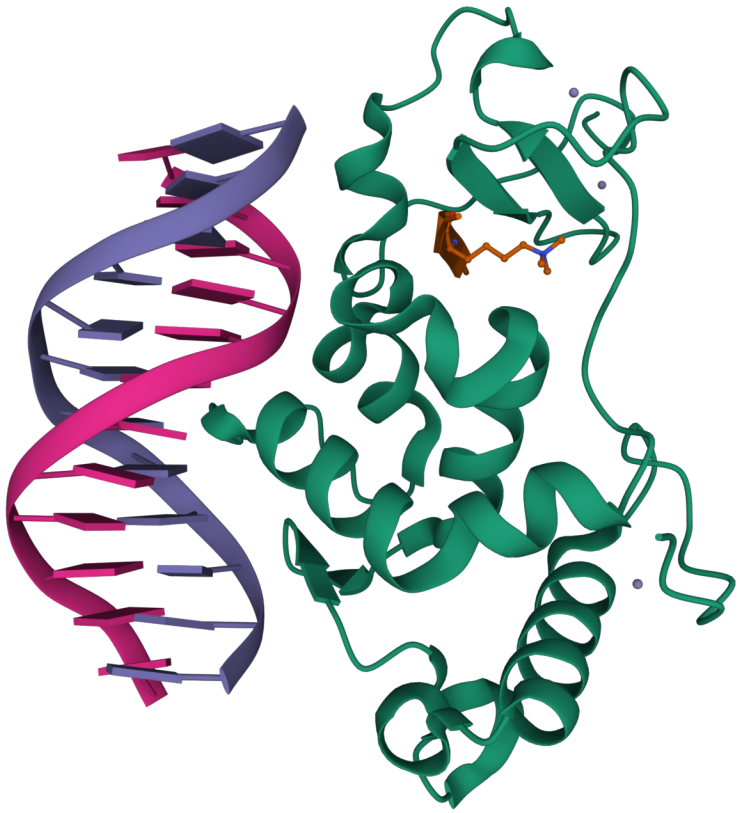

Figure 3: Target 2020-05-30_00000276 (PDB ID 6LQF) is an ARID-PHD protein cassette in complex with a peptide, DNA and zinc ions (Tan et al. 2020). The protein only has remote similarity ($<$ 30% sequence identity) to known structures, and none of them are in complex with DNA or the H3K4me3 peptide, making it an extremely challenging target. We are not aware of any methods that would currently be able to model this type of complex with acceptable accuracy. It should be noted that the peptide contains a non-canonical residue (N-Trimethyllysine, derived from Lysine).

In 2020, following the criteria outlined in the previous sections, we observed 983 structures containing more than one type of polymer entities. All of them were proteins in complex with peptides (421), DNA (279), RNA (199), DNA and RNA (52) or both peptides and nucleic acids (32).

With appropriate extensions, we believe that some of the scores selected for the individual target types such as the oligo-lDDT and CAD-score will be applicable to evaluate all these targets in a consistent manner.

Non-canonical amino acids and bases

Macromolecular structures frequently contain amino (or nucleic) acid residues which are not part of the 20 (respectively 8) standard residues. Traditionally for modeling purposes, the target sequences are canonicalized, that is modified residues are represented by their "parent" or closest canonical amino acid residue. However this may result in suboptimal models which wouldn't accurately represent the region containing

the modification. Post-translational modifications such as phosphorylations can result in significant conformational changes of the protein structure, which would be impossible to correctly model without knowledge of the modification.

As this information is available at the time of pre-release, CAMEO can provide sequences containing non-canonical residues on an opt-in basis (Figure 3). In this case, sequences will contain the PDB component identifier (typically 3 letters) enclosed in round brackets, in place of the parent amino acids. Models correctly representing those residues are expected to obtain higher scores for the all-atom measures such as the lDDT or the CAD-score.

In 2020, 444 of the 4323 protein, DNA, RNA and mixed structures and complexes we observed contained non-canonical residues. We observed these non-canonical residues in proteins (286), peptides (112), DNA (35) and RNA (27). 16 of them were observed in mixed complexes.

Current implementation status of CAMEO

At the time of writing, the CAMEO "Structures & Complexes" functionality is available as a beta version at*https://beta.cameo3d.org/* and is open for registrations. It has been providing targets containing proteins, DNA and RNA to registered servers on a weekly basis since October 2020. Participants can currently choose to receive the non-polymer ligands contained in these targets as InChI codes or PDB component IDs, as well as non-canonicalized sequences including modified residues. Predictions can be returned in PDB or mmCIF format, and are assessed with a fully automated pipeline including the oligo-lDDT and QS-scores. A weekly download of models, reference structures and assessment results is made available for offline analysis.

Our next steps will be to refine the target selection process, especially with respect to selecting relevant ligand targets as described in the previous sections. We are exploring ways to increase the diversity of the target selection, while ensuring that as many participants as possible receive a common subset of targets in order to make comparisons between servers possible for some aspects of the evaluation. We aim to improve the scoring by providing more diverse scores as described in the previous sections. Most groups developing novel methods have implemented their own scoring workflows locally. We therefore consider at this point the raw data downloads of the prediction results as a crucial service to the community developing specialized prediction methods as it allows including independent blind prediction data in publications describing the new method.

Conclusion

With the extension of CAMEO to the fully automated assessment of prediction of complexes (including protein-protein, DNA, RNA, peptides, small molecules), we aim to encourage and facilitate the development of automated structure prediction servers going beyond the modeling of single chains of amino acids. In this manuscript, we identified several challenging aspects of modeling which we believe will become more active areas of research in the future, and that are suitable for benchmarking with CAMEO. By assessing prediction targets with the same complexity as experimental structures using an "opt in" mechanism for the diverse modelling tasks, CAMEO will assist the development of new methods tackling these specific modeling challenges. As demonstrated by analysing the PDB releases of the last year, CAMEO will be able to provide a diverse set of challenging blind prediction targets to enable the community to tackle next generation modeling challenges.

We welcome feedback from the community on which of these aspects should be prioritized and how various predictions should be numerically evaluated in CAMEO. We encourage methods developers to register to the beta CAMEO server to help testing and evolving these new features according to the needs of the prediction community.

Acknowledgements

We are thankful for the invaluable feedback that we obtained from observers and participants alike, in particular the Schwede group members for testing new CAMEO releases, open discussions, and critical feedback. We would like to thank the community for their support, their new scores, and prediction methods. We are grateful to the sciCORE team for providing support and computational resources. We would like to thank the RCSB PDB for publishing the pre-release data openly and their invaluable input on experimental structure matters. We are grateful for funding: from the SIB Swiss Institute of Bioinformatics toward the development of CAMEO and OpenStructure and the use of sciCORE computing infrastructure; from NIH and National Institute of General Medical Sciences (U01 GM093324-01) partially to CAMEO; and from ELIXIR EXCELERATE to CAMEO.

Bibliography

Bertoni, Martino, Florian Kiefer, Marco Biasini, Lorenza Bordoli, and Torsten Schwede. 2017. "Modeling Protein Quaternary Structure of Homo- and Hetero-Oligomers beyond Binary Interactions by Homology." *Scientific Reports* 7 (1): 1–15.

Biasini, M., T. Schmidt, S. Bienert, V. Mariani, G. Studer, J. Haas, N. Johner, A. D. Schenk, A. Philippsen, and T. Schwede. 2013. "OpenStructure: An Integrated Software Framework for Computational Structural Biology." *Acta Crystallographica. Section D, Biological Crystallography* 69 (Pt 5): 701–9.

Biering, Scott B., David L. Akey, Marcus P. Wong, W. Clay Brown, Nicholas T. N. Lo, Henry Puerta-Guardo, Francielle Tramontini Gomes de Sousa, et al. 2021. "Structural Basis for Antibody Inhibition of Flavivirus NS1-Triggered Endothelial Dysfunction." *Science* 371 (6525): 194–200.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (December). https://doi.org/10.1186/1471-2105-10-421.

Fu, Lifeng, Fei Ye, Yong Feng, Feng Yu, Qisheng Wang, Yan Wu, Cheng Zhao, et al. 2020. "Both Boceprevir and GC376 Efficaciously Inhibit SARS-CoV-2 by Targeting Its Main Protease." *Nature Communications* 11 (1): 4417.

Guzenko, Dmytro, Aleix Lafita, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Jose M. Duarte. 2019. "Assessment of Protein Assembly Prediction in CASP13." *Proteins*87 (12): 1190–99.

Haas, Juergen, Rafal Gumienny, Alessandro Barbato, Flavio Ackermann, Gerardo Tauriello, Martino Bertoni, Gabriel Studer, Anna Smolinski, and Torsten Schwede. 2019. "Introducing 'Best Single Template' Models as Reference Baseline for the Continuous Automated Model Evaluation (CAMEO)." *Proteins* 87 (12): 1378–87.

Haas, Jürgen, Alessandro Barbato, Dario Behringer, Gabriel Studer, Steven Roth, Martino Bertoni, Khaled Mostaguir, Rafal Gumienny, and Torsten Schwede. 2018. "Continuous Automated Model EvaluatiOn (CAMEO) Complementing the Critical Assessment of Structure Prediction in CASP12." *Proteins*86 Suppl 1 (March): 387–98.

Kinch, Lisa N., Andriy Kryshtafovych, Bohdan Monastyrskyy, and Nick V. Grishin. 2019. "CASP13 Target Classification into Tertiary Structure Prediction Categories."*Proteins* 87 (12): 1021–36.

Kinch, Lisa N., Shuoyong Shi, Hua Cheng, Qian Cong, Jimin Pei, Valerio Mariani, Torsten Schwede, and Nick V. Grishin. 2011. "CASP9 Target Classification." *Proteins*79 Suppl 10 (October): 21–36.

Kryshtafovych, Andriy, Sony Malhotra, Bohdan Monastyrskyy, Tristan Cragnolini, Agnel-praveen Joseph, Wah Chiu, and Maya Topf. 2019. "Cryo-electron Microscopy Targets in CASP13: Overview and Evaluation of Results." *Proteins: Structure, Function, and Bioinformatics*. https://doi.org/10.1002/prot.25817.

Lensink, Marc F., Nurul Nadzirin, Sameer Velankar, and Shoshana J. Wodak. 2020. "Modeling Protein-protein, Protein-peptide, and Protein-oligosaccharide Complexes: CAPRI 7th Edition." *Proteins: Structure, Function, and Bioinformatics*. https://doi.org/10.1002/prot.25870.

Li, W., and A. Godzik. 2006. "Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences." *Bioinformatics* 22 (13). https://doi.org/10.1093/bioinformatics/btl158.

Mariani, Valerio, Marco Biasini, Alessandro Barbato, and Torsten Schwede. 2013. "lDDT: A Local Superposition-Free Score for Comparing Protein Structures and Models Using Distance Difference Tests." *Bioinformatics* 29 (21): 2722.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." *Proceedings of the 9th Python in Science Conference*. https://doi.org/10.25080/majora-92bf1922-00a.

Miao, Zhichao, Ryszard W. Adamiak, Maciej Antczak, Michał J. Boniecki, Janusz Bujnicki, Shi-Jie Chen, Clarence Yu Cheng, et al. 2020. "RNA-Puzzles Round IV: 3D Structure Predictions of Four Ribozymes and Two Aptamers." *RNA*26 (8): 982–95.

Mirdita, Milot, Lars von den Driesch, Clovis Galiez, Maria J. Martin, Johannes Söding, and Martin Steinegger. 2017. "Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments." *Nucleic Acids Research* 45 (D1): D170–76.

Mukherjee, Srayanta, and Yang Zhang. 2009. "MM-Align: A Quick Algorithm for Aligning Multiple-Chain Protein Complex Structures Using Iterative Dynamic Programming." *Nucleic Acids Research* 37 (11): e83.

Olechnovič, K., E. Kulberkytė, and C. Venclovas. 2013. "CAD-Score: A New Contact Area Difference-Based Function for Evaluation of Protein Structural Models." *Proteins*81 (1). https://doi.org/10.1002/prot.24172.

Olechnovič, Kliment, and Ceslovas Venclovas. 2014. "The CAD-Score Web Server: Contact Area-Based Comparison of Structures and Interfaces of Proteins, Nucleic Acids and Their Complexes." *Nucleic Acids Research* 42 (Web Server issue): W259–63.

Orengo, Christine, Sameer Velankar, Shoshana Wodak, Vincent Zoete, Alexandre M. J. J. Bonvin, Arne Elofsson, K. Anton Feenstra, et al. 2020. "A Community Proposal to Integrate Structural Bioinformatics Activities in ELIXIR (3D-Bioinfo Community)." *F1000Research* 9 (April). https://doi.org/10.12688/f1000research.20559.1.

Sehnal, David, Sebastian Bittrich, Mandar Deshpande, Radka Svobodová, Karel Berka, Václav Bazgier, Sameer Velankar, Stephen K. Burley, Jaroslav Koča, and Alexander S. Rose. 2021. "Mol* Viewer: Modern Web App for 3D Visualization and Analysis of Large Biomolecular Structures." *Nucleic Acids Research*, May. https://doi.org/10.1093/nar/gkab314.

Sillitoe, I., N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, C. S. M. Pang, et al. 2021. "CATH: Increased Structural Coverage of Functional Space." *Nucleic Acids Research* 49 (D1). https://doi.org/10.1093/nar/gkaa1079.

Steinegger, Martin, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J. Haunsberger, and Johannes Söding. 2019. "HH-suite3 for Fast Remote Homology Detection and Deep Protein Annotation." *BMC Bioinformatics* 20 (1): 473.

Tan, Lian-Mei, Rui Liu, Bo-Wen Gu, Cui-Jun Zhang, Jinyan Luo, Jing Guo, Yuhua Wang, et al. 2020. "Dual Recognition of H3K4me3 and DNA by the ISWI Component ARID5 Regulates the Floral Transition in Arabidopsis." *The Plant Cell* 32 (7): 2178–95.

Wagner, J. R., C. P. Churas, S. Liu, R. V. Swift, M. Chiu, C. Shao, V. A. Feher, S. K. Burley, M. K. Gilson, and R. E. Amaro. 2019. "Continuous Evaluation of Ligand Protein Predictions: A Weekly Community Challenge for Drug Docking." *Structure* 27 (8): 1326–35.e4.

Wirecki, Tomasz K., Chandran Nithin, Sunandan Mukherjee, Janusz M. Bujnicki, and Michał J. Boniecki. 2020. "Modeling of Three-Dimensional RNA Structures Using SimRNA." *Methods in Molecular Biology* 2165: 103–25.

Wishart, David S., Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. 2006. "DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration." *Nucleic Acids Research* 34 (Database issue): D668–72.

wwPDB consortium, Stephen K. Burley, Helen M. Berman, Charmi Bhikadiya, Chunxiao Bi, Li Chen, Luigi Di Costanzo, et al. 2018. "Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data." *Nucleic Acids Research* 47 (D1): D520–28.

Zemla, Adam. 2003. "LGA: A Method for Finding 3D Similarities in Protein Structures." *Nucleic Acids Research* 31 (13): 3370–74.