Workflows for Continuous Protein Structure Prediction Benchmarking

Haas J[1,2], Gumienny R[1,2], Robin X[1,2], Smolinski A[1,2], Schwede T[1,2]

1 SIB Swiss Institute of Bioinformatics 2 Biozentrum University of Basel 1,2 Klingelbergstr 50/70, 4056 Basel, CH

The critical assessment of structure prediction techniques (CASP[1]) is a pivotal event for the scientific community to establish the state of the art and identify bottlenecks. While it is held every two years, the development of automated prediction methods requires more frequent evaluation cycles. The "Continuous Automated Model EvaluatiOn (CAMEO[2])" platform conducts fully automated blind assessments based on the weekly pre-release of sequences and the subsequent release of PDB Protein Data Bank[3] structures. Each week, CAMEO publishes results for predictions of a set of about 20 targets collected during a four-day prediction window. Benchmarking data is generated consistently for all participants at the same time, ready to be used in publications. CAMEO offers a variety of scores reflecting different aspects of structure modeling, e.g. binding site accuracies or homo-oligomer interface quality. CAMEO is now employing Nextflow[4] workflows to automate each stage of the weekly benchmarking events reducing maintenance work, helping with portable deployment and facilitating future developments.

CAMEO 3D - Protein Structure Prediction

Evaluates monomeric and homo-oligomeric predictions:

- Includes targets structures from X-ray and NMR experiments.
- Scores represent a wide variety of modeling aspects, e.g. overall all-atom quality (lDDT[5], Fig 1), homo-oligomer interface quality (QS-score[6], Fig 2), binding site and model confidence accuracy.

Currently ongoing projects:

- OpenEBench[7] integration (data and operation)
- Target validation (PDB validation reports)
- Hetero oligomer and ligand support

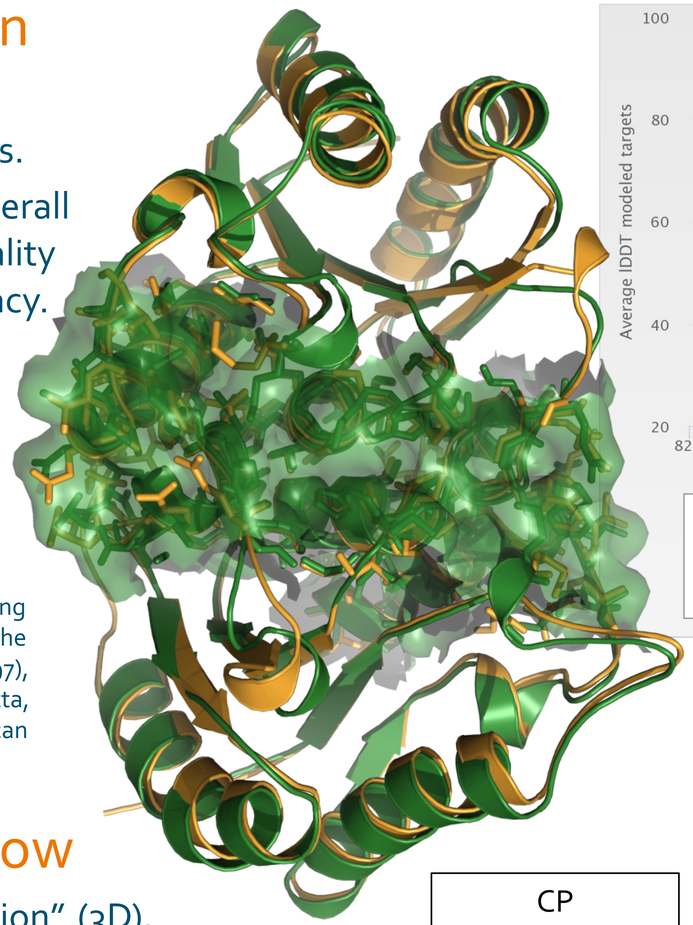


Figure 2: Example of the homodimeric protein–protein interface assessment employing "QS-score" in the case of target 2018-05-12_00000037_1 | 5nna [D] (Isatin hydrolase A). Only the surface depiction of the SWISS-MODEL interface prediction is shown (green, QS-score: 91.97), while the interface residues are shown for both predictions in stick mode (orange: Robetta, QS-score: 85.60). The full dimers are shown in cartoon mode. Only these two servers can routinely predict homo-oligomers at the moment.

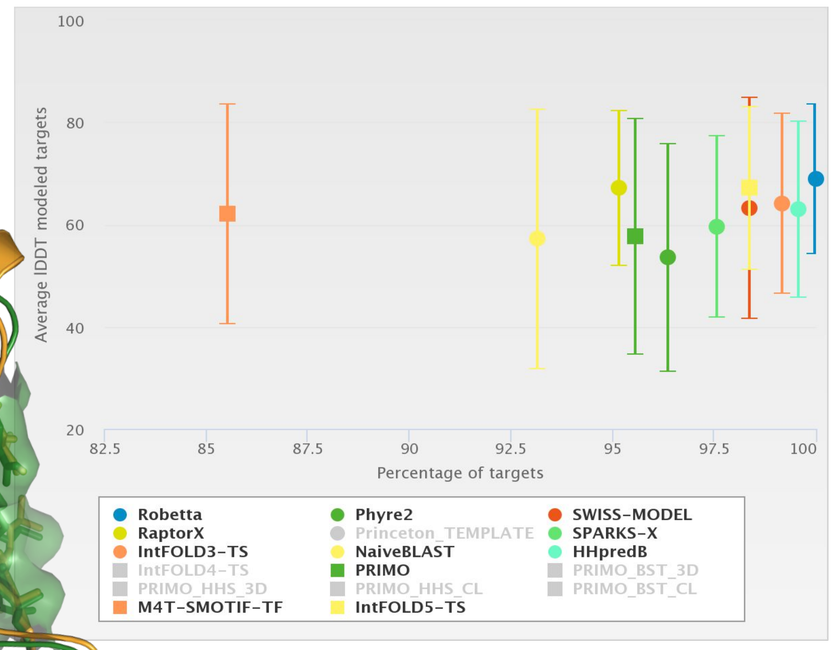


Figure 1 (above): Global all-atom quality (lDDT) by returned percentage of targets for a 3-months time frame (2019-03-01 - 2019-05-25). Apart from one method, all returned more than 90% of the targets. Phyre2 and NaiveBLAST are historic approaches to compare to.

Continuous Evaluations Based on Nextflow

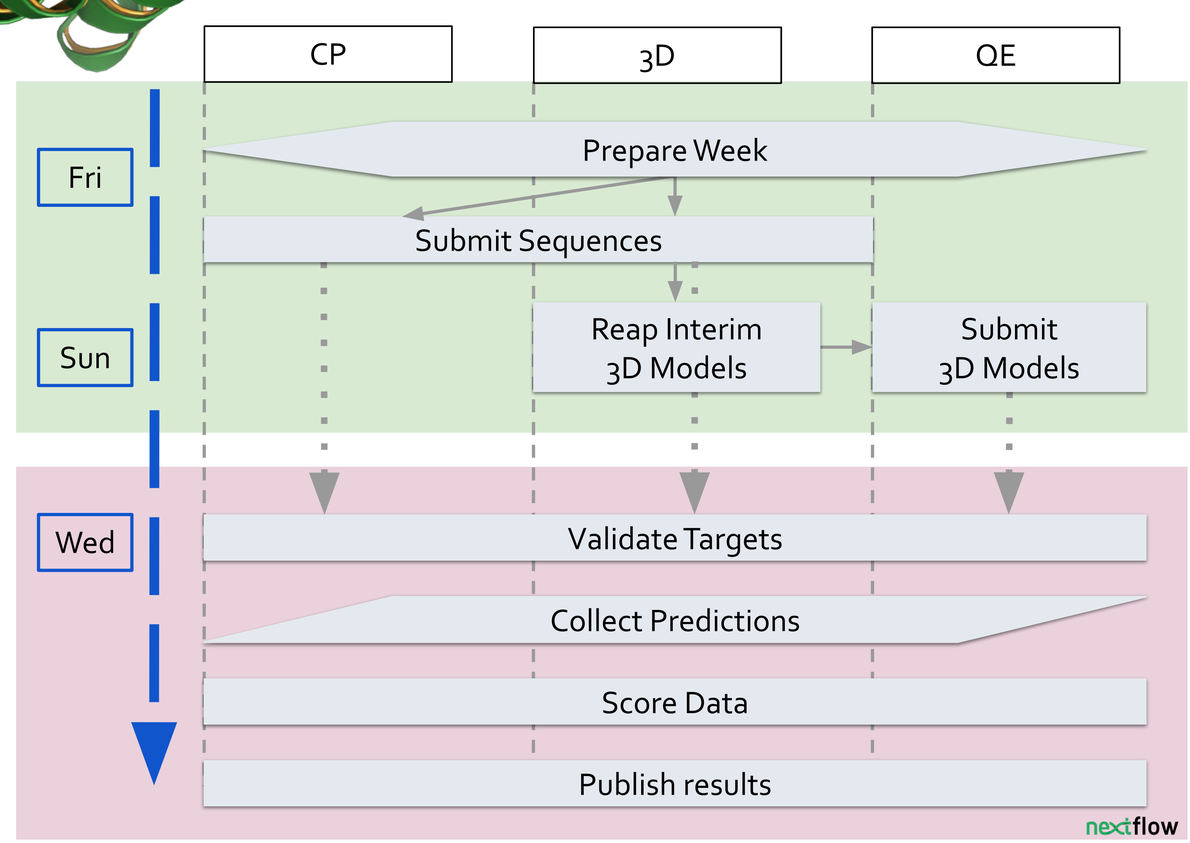
CAMEO supports three categories: "protein structure prediction" (3D), "quality estimation" (QE) and "contact prediction" (CP). Each evaluation cycle in CAMEO consists of six steps: prepare a new week, submit targets, validate targets, collect predictions, score and publish to cameo3d.org. The 30 processes are split into two Nextflow scripts handling the steps execution and restart:

1. prepare-and-submit

Retrieves the PDB pre-release data on Friday night and sets up the new cycle executing some sanity checks. Next for each category this Nextflow script selects the targets and submits them to the participants.

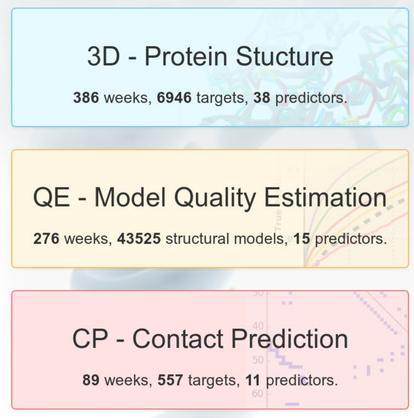
2. publish-results

Upon PDB release on Wednesday at midnight, CAMEO closes the prediction window, collects the predictions sent via email, extracts them and checks the format. Target structures are validated and predictions are scored against the experimental evidence deposited with PDB. In a final stage all data is published to the website.



Conclusions

- Publication-Ready Data
- Weekly Cycle
- Variety of Scores - Different Quality Aspects



References

1. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A, *Proteins* (2018), 86 Suppl 1:7–15.
2. Haas J, Barbato A, Behringer D, et al. *Proteins* (2018), 86 Suppl 1:387–398.
3. Berman HM, Henrick K, Nakamura H, *Nature Structural Biology* (2003), 10: 980.
4. Di Tommaso P, Chatzou M, Floden EW, et al. *Nat Biotechnol* (2017) 11;35(4):316-319.
5. Biasini M, Mariani V, Barbato A, Schwede T, *Bioinformatics* (2013), 29(21):2722-8.
6. Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T, *Sci Rep.* 2017 Sep 5;7(1):10480.
7. Capella S, De la Inglesia D, Haas J, et al. *BioRxiv* (2017), DOI: 10.1101/181677.

Contact:

Torsten Schwede
help-cameo3d@unibas.ch
SIB Swiss Institute of Bioinformatics
Klingelbergstr 50/70, 4056 Basel, CH

https://www.cameo3d.org

ELIXIR-EXCELERATE is funded by the European Commission within the Research Infrastructures programme of Horizon 2020, grant agreement number 676559.