PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

CASP11 statistics and the prediction center evaluation system

Andriy Kryshtafovych, Bohdan Monastyrskyy, and Krzysztof Fidelis*

Protein Structure Prediction Center, Genome and Biomedical Sciences Facilities, University of California, Davis, California 95616

ABSTRACT

We outline the role of the Protein Structure Prediction Center (predictioncenter.org) in conducting the CASP11 and CASP ROLL experiments, discuss the experiment statistics, and provide an overview of the present CASP infrastructure. The biggest changes compared to the previous CASPs are the implementation of the evaluation system incorporating practically all evaluation measures, statistical tests, and visualization tools historically used by the CASP assessors, the expansion of the infrastructure to incorporate new categories of contact-assisted and multimeric predictions, and the redesign of the assessors' web-workspace enabling assessments based on multiple measures for different group categories and target sets.

Proteins 2016; 84(Suppl 1):15–19.

Key words: CASP; protein structure prediction; protein structure modeling.

INTRODUCTION

Similarly to previous CASPs, in CASP11 the Protein Structure Prediction Center at the University of California, Davis was involved in all aspects of data handling and evaluation. The tasks performed by the Center included: disseminating information about the CASP experiment; registering participants; providing assistance in connecting registered servers to the CASP distribution/acceptance system; soliciting, verifying, selecting, preprocessing, and releasing targets for prediction in the full range of modelling categories; accepting submitted predictions and posting server-generated models at the website for public use; monitoring public release of target structures; preprocessing coordinates for evaluation purposes; searching for available structural templates and preliminary division of targets into evaluation domains; evaluating models, analyzing and presenting the evaluation results in a textual and graphical form to assessors and public; providing assistance to CASP assessors and predictors; and—working with the CASP Organizing Committee—planning the CASP conference and publishing meeting materials. In this article, we report on the data (targets, predictions, and results) processed by the Prediction Center in CASP11, and provide an update on the newly introduced evaluation measures and web resources.

PREDICTION TARGETS

In CASP11, one hundred sequences have been selected as targets (T0759 through T0858) and released for prediction. Majority of these targets were obtained from the Structural Genomics centers, but a significant portion (>40%)—from outside of the PSI. Such a diversification is important to CASP, as number of structures solved by the PSI Centers has winded down and we are exploring new avenues for target supply. Details of some of the most interesting targets are discussed elsewhere in this issue.[1] Targets in CASP11 were divided into two categories: (1) targets for prediction by all groups (all-group targets, or expert/server targets), and (2) server-only targets (Kinch *et al.* CASP11 target classification—this issue). Similar to previous CASPs, the all-group targets were typically selected from among the more challenging targets. Targets were released to predictors through the CASP11 website, May 1 through July 16, 2014. At the time of the web posting, targets were also automatically

Grant sponsor: NIH/NIGMS; Grant numbers: R01GM100482; R13GM109649.
*Correspondence to: Krzysztof Fidelis, Protein Structure Prediction Center, Genome and Biomedical Sciences Facilities, University of California, Davis, CA 95616. E-mail: kfidelis@ucdavis.edu
Received 19 November 2015; Revised 18 January 2016; Accepted 4 February 2016
Published online 9 February 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.25005

forwarded to the participating servers through an automatic distribution system.

After completion of the CASP10 experiment, we continued releasing prediction targets on the rolling basis. A protein sequence was selected as a CASP ROLL target only if it was sufficiently challenging for prediction as indicated by the lack of suitable modeling templates (see paper[2] for the adopted verification protocol). Between CASP10 and CASP11, we have prepared and released 29 CASP ROLL targets (R0019 through R0047). Models submitted on these targets were evaluated by the Prediction Center and assessed by the CASP11 free modeling assessors (Kinch *et al.* Evaluation of free modeling targets in CASP11 and ROLL—this issue).

Targets, for which the Prediction Center was able to obtain structures shortly after the original sequence release, were considered as candidates for refinement or contact-assisted experiments. Selection of targets for this purpose was performed at the Prediction Center. A target (or its constitutive domain) was considered appropriate for refinement if it was relatively short, had no significant gaps in structure, had no apparent crystal contact distortions, and the best submitted server predictions were of relatively high accuracy (usually better than 50 GDT_TS). In CASP11, we released refinement targets for more proteins (37 vs. 28) and in a wider range of difficulty compared to CASP10. The majority of the refinement targets (26 out of 37) were shorter than 200 residues; the longest had 288 residues. Accuracy of the starting models ranged from 46 to 90 GDT_TS units, with the vast majority of targets (31 out of 37) scoring above 60 GDT_TS.

More challenging targets (best server models usually below 50 GDT_TS) were considered as candidates for contact-assisted experiments, in which we probed the extent to which sparse experimental data or contact predictions might improve model accuracy. In CASP11, we conducted contact-assisted experiments in a conceptually different manner compared to CASP10. The general idea was to explore more realistic scenarios where restraints are obtained from typically accessible sources rather than selected from experimental structure. The contact-assisted category included 23 Tp targets (modeling based on predicted contacts), 19 Ts targets (modeling based on simulated sparse experimental data obtained by NMR), and 4 Tx targets (modeling based on experimental crosslinking data). In the Tp category, predicted three dimensional contacts collected in the CASP residue–residue contact prediction category (RR) were released shortly after completing the unassisted prediction. For each target, we released approximately $L/5$ (L, target length) long-range contacts from 10 historically better performing CASP11 contact prediction groups. These predictions included both correct and incorrect contacts. After the collection of structure predictions in the Tp category, for selected targets we released bigger sets of contacts simulating the data available in the initial stages of a typical NMR study (the Ts category). Such constraints are sparse and usually not sufficient to refine the structure using standard NMR packages. As in Tp, the provided sets contained both correct and incorrect contacts. The simulated sparse NMR contacts were generated in Gaetano Montelione's group. In the Tx category, predictors were provided with distance restraints obtained with crosslinking mass spectroscopy studies. The studies were carried out in Juri Rappsilber's group (Technical University of Berlin) on the biological material obtained from crystallographers determining the structure. In addition to the above three categories, we also provided a fourth (Tc) set of contacts. These were generated for 24 targets using the knowledge of structure. In this category, we released $\sim L/5$ *correct* contacts predicted by the same ten groups as in the Tp category. The contacts were usually released after completing the Ts prediction. Finally, some of CASP targets were designated for quaternary structure (multimeric) prediction. In this category, three target tandems (T0787/788, T0797/798, T0840/841) and target T0825 were assessed as heteromultimers, and—additionally—23 targets were assessed as homomultimers predictions. These assessments were performed by the CAPRI team (Lensink *et al.* Prediction of homo- and heteroprotein complexes by ab-initio and template-based docking: a CASP-CAPRI experiment—this issue).

PARTICIPANTS AND PREDICTIONS

Over 200 groups participated in each of the CASP rounds held since 2002. In the latest, 11th round, 123 human-expert groups and 84 automatic servers representing 102 research centers world-wide registered and actively participated. In CASP ROLL, 9 to 24 groups (depending on the target) submitted predictions on targets released between CASP10 and CASP11. The total number of models evaluated for the latest round of CASP and CASP ROLL exceeded 60,000. All predictions were collected, checked for format consistency, and stored in relational databases. In CASP11, we accepted predictions in three different formats: tertiary structure (TS), residue-residue contacts (RR), and estimates of model accuracy, a.k.a. quality assessment (QA) (see http://predictioncenter.org/casp11/index.cgi?page = format for details); in CASP ROLL we accepted tertiary structure and contact predictions.

PREPROCESSING OF TARGET STRUCTURES, DOMAINS, AND TEMPLATES

For evaluation purposes, the Prediction Center preprocessed coordinate files obtained from crystallographers

and NMR spectroscopists bringing the coordinates to agree with the residue naming and numbering of the released CASP targets. The most typical chains (X-ray) or models (NMR) were selected as representatives in case of X-ray homo-multimers or NMR ensembles. For hetero-multimers, reference structures were prepared for all possible structurally different combinations of chains to allow evaluation of all submitted models. Only well-defined regions of targets were included in the reference structures. In many cases, we could obtain additional information on protein function, binding sites, ligands, resolution, oligomerization, and at times even preliminary coordinates already at the time of target selection or soon afterwards. This enabled us to designate more targets for refinement and sparse data-assisted experiments. Specifically, the number of refinement targets grew from 28 in CASP10 to 37 in CASP11, and the number of contact-assisted targets from 15 to 24, respectively. Target coordinates and the associated information were posted in a secure web workspace for analysis by the assessors. This additional information available early in the prediction process is potentially useful in formulating challenging target-specific questions before the modeling process ends.

For parsing targets into evaluation domains we used the DomainParser2[3] and DDomain[4] packages. Results of the automatic parsing are used for preliminary evaluation of models at the domain level and subsequent checks of whether dividing into domains is needed for final evaluation. The checks are based on the Grishin plots (Kinch *et al.* CASP11 target classification—this issue) illustrating differences between the whole-target evaluation scores and weighted domain-based scores and indicating the necessity for a domain split in evaluation. All the data obtained in these analyses are provided to the assessors for the purpose of defining boundaries of the final evaluation units.

We have also searched for appropriate modeling templates for both domains and whole targets. Information on structural homologues is needed to identify the level of target difficulty, and to define the type of questions that may be addressed by structure prediction. Identification of homologous structures is also needed in a more detailed evaluation of submitted models, where comparisons with template structures are necessary. It is also important to keep a record of all the homology-related structures available for any given target by the target prediction deadline. This information is useful in future benchmarking experiments allowing for comparisons with the original CASP predictions, and for estimating progress in the field.

The lists of related structures were compiled, together with the corresponding levels of structure similarity to target proteins, using two strategies. First, the Protein Data Bank was searched for homologues with sequence-based methods PSI-BLAST[5] and HHblits[6] to estimate the difficulty of targets and their constituent domains. Second, once the target closed for prediction and the structure become available, a direct structure similarity search versus the whole PDB was performed with MAMMOTH[7] and LGA.[8] Scientific literature and databases were searched for any structural information available on targets and their homologues. Results were carefully analyzed and any relevant structural information found was made available to the assessors.

EVALUATION AT THE PREDICTION CENTER

In CASP11, for the first time all the basic evaluation scores and statistical tests needed for the assessors' analyses were calculated at the Prediction Center. This lessened the burden on the assessors and allowed them to concentrate on the analysis of the evaluation results rather than on their generation.

Already in CASP10, the following measures were calculated for the tertiary structure evaluation (regular, refinement, and contact-assisted categories): GDT-like measures of global model accuracy (GDT_TS, GDT_HA, GDC_SC, GDC_ALL)[8–10] (Definitions of the measures used in CASP are also available via the Prediction Center website, e.g.: http://predictioncenter.org/casp11/doc/help.html); alignment accuracy measures (AL0, AL4); results of the sequence-independent model-target comparisons (LGA_S, Mammoth,[7] DALI[11]); RMSD (root mean square deviation); strereochemical correctness measure (Molprobity[12]); as well as measures based on local correctness of models (CAD-score,[13] LDDT,[14] SphereGrinder[2,15] and RPF[16]). In addition to these measures, in CASP11 we calculated: QCS and TenS scores[17]; CoDM, DFM, and Handedness,[18] TM-score,[19] FlexE,[20] SOV,[21] and QSE measures. All these evaluation measures (with the exception of ASE, see below) are comprehensively described in the referenced papers and on the Prediction Center website. In CASP11, the SphereGrinder score was calculated and then averaged for two different RMSD cutoffs—2 Å and 4 Å (cf. a single 2 Å cutoff in CASP10)—to allow a more relaxed fit between model and target. The new ASE (Accuracy of Self-Estimates) measure was developed by the authors of this article for CASP11 to evaluate the accuracy of submitted per-residue error estimates. The score evaluates how far away are the submitted error estimates from the actual errors (distances between the corresponding residues in the LGA model-target superposition). For each residue, the distance d is normalized to the [0;1] range using the S-function

$$S(d) = \frac{1}{1 + \left(\frac{d}{d_0}\right)^2}$$

and then averaged for the whole model and rescaled to the [0;100] range using the following formula

$$\mathrm{ASE} = 100 \times \left(1 - \frac{1}{N}\sum_{i=1}^{N} |S(e_i) - S(d_i)|\right)$$

where e_i is the estimated distance as submitted by predictors, d_i is the actual distance in the LGA superposition, d_0 is a scaling factor set here to 5. The higher the score, the more accurate the prediction of the distance errors in a model. If error estimates for some residues are not included in the prediction, they are set to a high value so the contribution of that specific error to the total score is negligible.

In addition to calculating the many raw scores for structural models, we also performed a series of statistical tests designed to compare the results obtained by participating groups. These tests include *t*-tests, head-to-head comparisons, bootstrapping tests, and *z*-scores. The *z*-scores were computed for each group on all measures so that the assessors could combine them with the desired weights for a final group ranking. A separate web-based infrastructure was developed to simplify the assessors' analysis.

For the assessment of estimates of model accuracy (EMA), in CASP11 we substantially extended the arsenal of evaluation measures. In addition to the comparison of predicted global accuracy scores with the GDT_TS values, we also compared them with the LDDT, CAD, and Sphere Grinder scores that implicitly reward methods recognizing models with accurate local geometry. Adding these measures to the evaluation package extends the scope of the assessment by providing the account of model features not readily identified by the GDT_TS alone. The main emphasis of the EMA assessment was placed on the ability of methods to identify the best models in a decoy set. Bivariate descriptive statistics and ROC analysis were used to additionally assess the correlation between the predicted and observed accuracy of models, the accuracy in distinguishing between good and bad models, the ability to discriminate between reliable and unreliable regions in models, and the accuracy of the self-estimates of coordinate errors. A detailed description of the EMA measures can be found in our assessment article elsewhere in this issue.[22]

The residue–residue contact predictions in CASP11 showed exciting developments and were evaluated with a number of measures, including precision, recall, Xd-scores, the Matthews correlation coefficients, as well as precision-recall curves. Since promising results were obtained by methods using the new co-variation techniques, we carried out the analysis of the dependency of these results on the depth of the corresponding sequence alignments. Our contact assessment paper[23] carries detailed description of these measures and the results of the residue–residue contact evaluation. The comprehensive analysis on the various target sets and contact sets is provided on our webpage (http://predictioncenter.org/casp11/rr_results.cgi).

RELEASE OF RESULTS, VISUALIZATION TOOLS, AND SUMMARY TABLES

During the CASP prediction season and thereafter, the evaluation results discussed above were made available to the independent assessors through a password-protected gateway on a continuing target by target basis as soon as the calculations were completed. The results were provided as plain text files, interactive tables, and as graphical presentations. A week before the CASP11 meeting, final evaluation data for CASP11 and CASP ROLL were publicly released through the Prediction Center website http://predictioncenter.org/{casp11|casprol}/results.cgi.

The skeleton of the infrastructure for displaying CASP results was outlined in our previous papers.[2,24–26] For CASP11, we extended the infrastructure to include additional evaluation measures, additional prediction categories, and interactive cumulative score tables for regular targets, refinement targets, as well as contact-assisted and residue–residue contact targets. The new web interface was also developed to show evaluation results for multimeric targets. For such targets, we presented results for different arrangements of molecules in the asymmetric unit and then, for each model, we selected the highest scores for release in the final evaluation table.

The tables showing summary group performance scores can be generated by the user for a choice of first models or best models; for all groups on "expert/server" targets or for server groups on all targets; for different target difficulty categories separately or combined; and for different evaluation measures—GDT_TS alone or the combined score used by the assessors' in their final rankings.

ACKNOWLEDGMENTS

Authors acknowledge the crystallographers and NMR spectroscopists taking part in CASP11, especially the researchers from the JCSG center, who provided 32 out of the 100 prediction targets (see http://predictioncenter.org/casp11/numbers.cgi). Special thanks are extended to the staff of the Protein Data Bank for providing targets to the experiment through the CASP hold structure submission option.

REFERENCES

1. Kryshtafovych A, Moult J, Baslé A, Burgin A, Craig TK, Edwards RA, Fass D, Hartmann MD, Korycinski M, Lewis RJ, Lorimer D, Lupas AN, Newman J, Peat TS, Piepenbrink KH, Prahlad J, van Raaij MJ, Rohwer F, Segall AM, Seguritan V, Sundberg EJ, Singh AK, Wilson MA, Schwede T. Some of the most interesting CASP11 targets through the eyes of their authors. Proteins 2015 Oct 16. doi: 10.1002/prot.24942 [Epub ahead of print].

2. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins 2014;82:7–13.
3. Guo JT, Xu D, Kim D, Xu Y. Improving the performance of DomainParser for structural domain partition using neural network. Nucleic Acids Res 2003;31:944–952.
4. Zhou H, Xue B, Zhou Y. DDOMAIN: dividing structures into domains using a normalized domain-domain interaction profile. Protein Sci 2007;16:947–955.
5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
6. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 2012;9:173–175.
7. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606–2621.
8. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
9. Zemla A, Venclovas Moult J, Fidelis K. Processing and evaluation of predictions in CASP4. Proteins 2001;(Suppl 5):13–21.
10. Keedy D, Williams CJ, Arendall WB III, Chen VB, Kapral GJ, Gillespie RA, Zemla A, Richardson DC, Richardson JS. The other 90% of the protein: assessment beyond Calphas for CASP8 template-based models. Proteins 2009;77:29–49.
11. Holm L, Kaariainen S, Rosenstrom P, Schenkel A. Searching protein structure databases with DaliLite v.3. Bioinformatics 2008;24:2780–2781.
12. Chen VB, Arendall WB,3, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 2010;66:12–21.
13. Olechnovic K, Kulberkyte E, Venclovas C. CAD-score: a new contact area difference-based function for evaluation of protein structural models. Proteins 2013;81:149–162.
14. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 2013;29:2722–2728.
15. Lukasiak P, Wojciechowski M, Ratajczak T, Hasinski K, Monastyrskyy B, Kryshtafovych A, Fidelis K. SphereGrinder—estimating similarity of structures on a local scale. In: Proceedings of CASP10 conference, Gaeta, Italy; 2012. pp 274–275.
16. Huang YJ, Mao B, Aramini JM, Montelione GT. Assessment of template-based protein structure predictions in CASP10. Proteins 2014;82:43–56.
17. Kinch L, Yong Shi S, Cong Q, Cheng H, Liao Y, Grishin NV. CASP9 assessment of free modeling target predictions. Proteins 2011; 79(Suppl 10):59–73.
18. Tai CH, Bai H, Taylor TJ, Lee B. Assessment of template-free modeling in CASP10 and ROLL. Proteins 2014;82:57–83.
19. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
20. Perez A, Yang Z, Bahar I, Dill KA, MacCallum JL. FlexE: using elastic network models to compare models of protein structure. J Chem Theory Comput 2012;8:3985–3991.
21. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 1999;34:220–223.
22. Kryshtafovych A, Barbato A, Monastyrskyy B, Fidelis K, Schwede T, Tramontano A. Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. Proteins 2015 Sep 7. doi: 10.1002/prot.24919 [Epub ahead of print].
23. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. New encouraging developments in contact prediction: Assessment of the CASP11 results. Proteins 2015 Oct 16. doi: 10.1002/prot.24943 [Epub ahead of print].
24. Kryshtafovych A, Milostan M, Szajkowski L, Daniluk P, Fidelis K. CASP6 data processing and automatic evaluation at the protein structure prediction center. Proteins 2005;61:19–23.
25. Kryshtafovych A, Prlic A, Dmytriv Z, Daniluk P, Milostan M, Eyrich V, Hubbard T, Fidelis K. New tools and expanded data analysis capabilities at the Protein Structure Prediction Center. Proteins 2007;69:19–26.
26. Kryshtafovych A, Krysko O, Daniluk P, Dmytriv Z, Fidelis K. Protein structure prediction center in CASP8. Proteins 2009;77:5–9.