SUPPLEMENTARY MATERIAL

Appendix 1

Structure of the Experiment

Participants registered for the experiment in two ways: as human-expert teams, where a combination of computational methods and investigator expertise may be used; or as servers, where methods are only computational and fully automated, so that a target sequence is sent directly to a machine. Investigators could register in both categories and limited additional groups could be registered by the same participants, to allow for testing of different methods. The expert groups were allowed a longer time period (typically three weeks versus 72 hours for servers) between the release of a target and submitting a prediction. There are now very few groups where significant human expertise is brought to bear, and the longer period was primarily utilized in two ways – to make use of initial models produced by the rapid server stage, and to perform longer calculations.

Information about 'soon to be solved' structures was collected from the experimental community and passed on to the modeling community. As is usual, the main CASP prediction season lasted for three months, from May through July; the season for the contact-assisted and refinement experiments lasted till mid-August; and the CASP ROLL experiment for predicting free modeling targets in between CASPs ran from December 2013 until the start of CASP11. For the last two CASP experiments, the PDB has provided an ongoing system for depositors to identify a structure as a CASP target, greatly helping the flow of the process.

Groups were limited to a maximum of five models per target, and were instructed that most emphasis in assessment would be placed on the model they designated as the most accurate (referred to as 'model 1'), particularly for template based modeling. The models were compared with experiment, using numerical evaluation techniques and expert assessment.

All predictions were submitted to the Prediction Center in a machine-readable format. Accepted submissions were issued an accession number, serving as the record that a prediction had been made by a particular group on a particular target. Predictions were numerically evaluated using the same set of core metrics as in CASP10, including measures based on rigid-body model-target

superpositions (RMSD, GDT_TS [1,2], GDT_HA[3]) for overall backbone accuracy; measures based on local similarity of models and targets: CAD[4] (similarity of contact areas) , LDDT[5] (similarity of distance patterns) and Sphere Grinder[6] (similarity of local substructures); and measures evaluating strereochemical correctness of models (Molprobity[7]). . Besides these, some other measures and statistical tests that were employed by previous assessors were also calculated (see the Prediction Center paper, this issue). As always, assessors were encouraged to develop their own additional measures to complement the established CASP ones.

A planning meeting was held in October, at which the assessors presented their findings to each other and to the organizers. All predictions were anonymized until that had been done. After the assessors reported their conclusions, group identities were revealed and the most successful groups as well as those with the promising novel methods were invited to talk at the predictor's conference. The conference to discuss the results was held in Riviera Maya, Mexico in December 2014. The meeting program can be found at http://predictioncenter.org/casp11/doc/CASP11_Meeting_Program.html.

Targets and Submissions

In the main CASP11 experiment 100 protein sequences were released as modeling targets, of which 55 were designated 'all groups' (human and server) targets. Seven targets were cancelled, leaving 93 where the experimental structures were available for evaluation and assessment. Seven out of the 93 assessed targets were designated and evaluated as hetero-multimers and additional 23 targets were assessed as homo-multimers in the CASP/CAPRI joint experiment. Also, between CASP10 and CASP11, 29 challenging targets were released for prediction in the CASP ROLL experiment.

In cases where significant domain movements were observed, or individual domains were classified in different categories (FM, TBM), the targets were divided into separate evaluation units. In all, 136 evaluation units were included. For 37 TBM domains, selected models were released as starting points for the refinement exercise, and for 23 FM or harder TBM domains, sets of contacts and/or experimental sparse data were released for contact-assisted prediction after the initial models had been collected. Contact-assisted prediction was carried out in four

categories: Tp (modeling based on predicted CASP contacts, both correct and incorrect), Ts (modeling based on simulated sparse experimental data obtained by NMR), Tx (modeling based on experimental cross-linking data) and Tc (modeling based on the correctly predicted contacts in CASP contact prediction category).

There were 207 research groups from about 100 labs around the world participating in CASP11 and submitting a total of 58,835 models, of which 36,776 were three dimensional co-ordinate sets. The remaining submissions are for residue–residue contacts (2,332), structural disorder (790, not assessed), and estimation of three-dimensional model quality (6,953). 6,639 3D structures were refinements of initial models and 5,314 were contact assisted models.

Management and Organization

The CASP11 organizers were unchanged from CASP10 and are the authors of this paper. There is an advisory board composed of senior members of the modeling community who advise the organizers on aspects of the CASP experiments and related activities. A participants' meeting during each CASP conference allows for more direct interaction, including votes on issues of CASP policy. The Protein Structure Prediction Center is responsible for all data management aspects of the experiment, including the distribution of target information, collection of predictions, generation of numerical evaluation data, developing tools for data analysis, data security, and maintenance of a web site where all data are available. Two prominent members of the field, Roland Dunbrack and Nick Grishin, were invited to serve as independent assessors and judge about the quality of the models received as well as to interpret the results in terms of progress and bottlenecks.

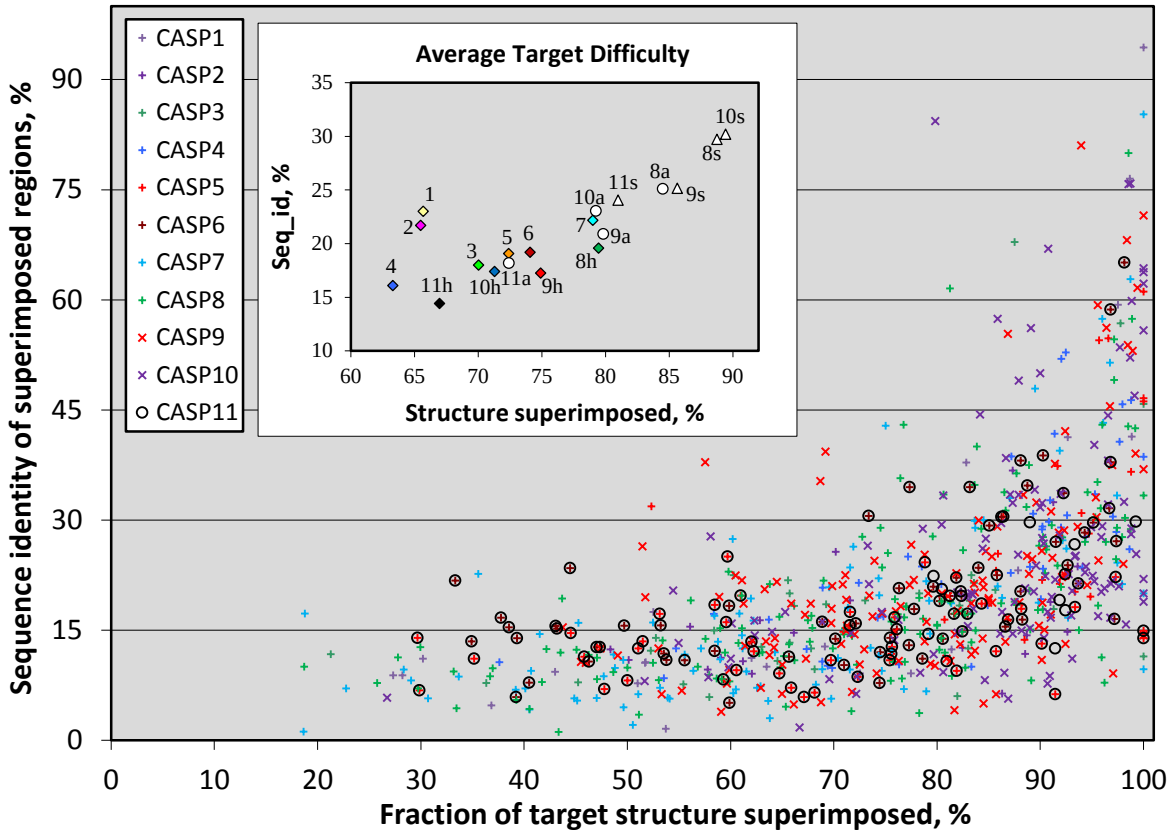
Appendix 2

Target difficulty

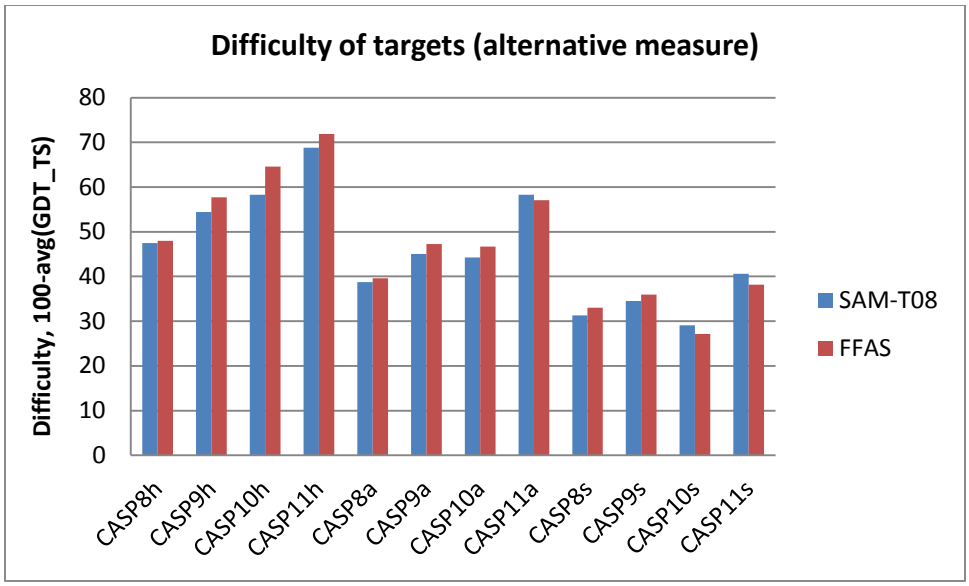
With the large number of experimental structures now determined (currently about 105,000 entries in the protein databank at this writing), most modeling is now based at least partly on homology to known structure. The accuracy of the resulting models varies widely, depending on two primary factors – the level of sequence similarity to proteins with experimental structure, and the extent of the structure covered by such relationships. In CASP, we use a difficulty scale related to these two factors[8]. Supplementary figure S1 shows that by these criteria, the CASP11 'h' targets are among the most difficult of any CASP, and the full set of targets ('11a') are about as difficult as the 'h' targets in CASP10.

This conclusion is also supported when target difficulty is based on modeling results from reference servers (see section Backbone and Alignment Accuracy). Supplementary figure S2 shows difficulty of targets in four recent CASPs in three target subcategories 'h', 's', 'a'. According to this difficulty definition, CASP11h target set stands out as the most difficult one, consistently with the Figure S1 results.

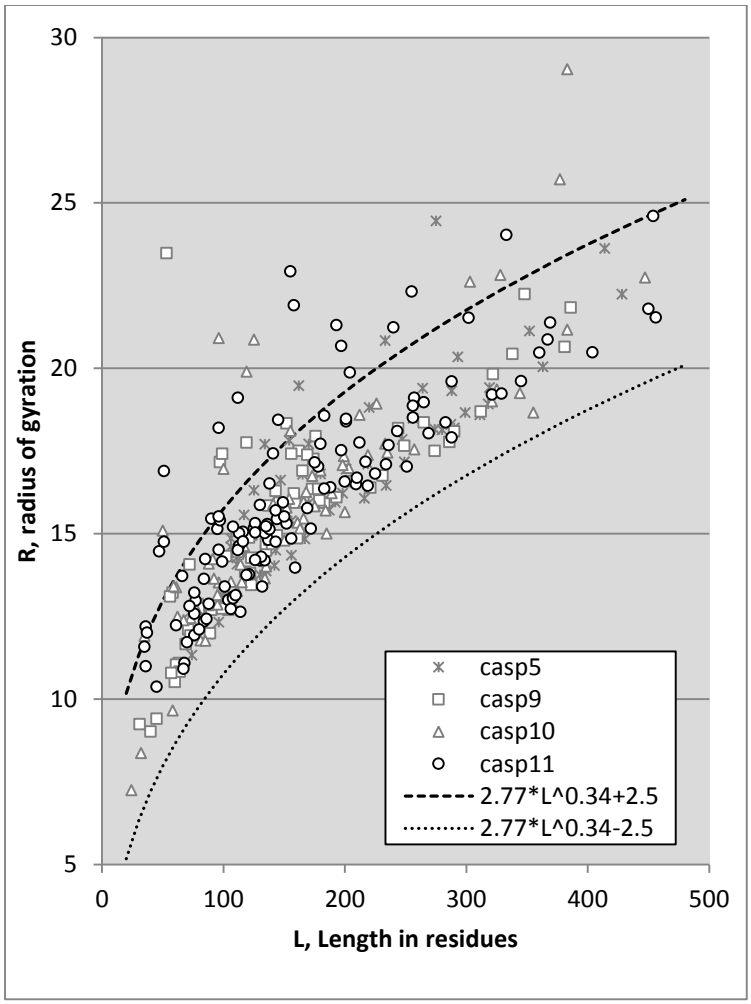
There are a number of additional factors contributing to the difficulty of a target from a modeling perspective. One is that the more non-globular the target, the harder it tends to be. Supplementary figure S3 shows the distribution of target radius of gyration as a function of target length for the three most recent CASPs and CASP5. There are 16 unusually high radius targets in CASP11, which is as many as in CASP 9 and 10 combined.



Supplementary Figure S1. Relative modeling difficulty of CASP targets, as a function of the fraction of each target that can be superimposed on a known structure (horizontal axis) and the sequence identity between target and template for the superimposed region (vertical axis). Each point represents one target. Inset shows the average values for each CASP. For CASPs 8-11, averages are shown separately for server only targets (marked with an "s" suffix), 'all groups' targets (a.k.a. human targets, "h"), and complete set of all targets ("a"). The closer the point to the left lower corner of the graph – the harder the target set. By this criterion, CASP11 human target set (11h) is one of the most difficult sets in all of the CASP. All targets in CASP11 (11a) are of approximately the same difficulty as human targets in CASP10 (10h).



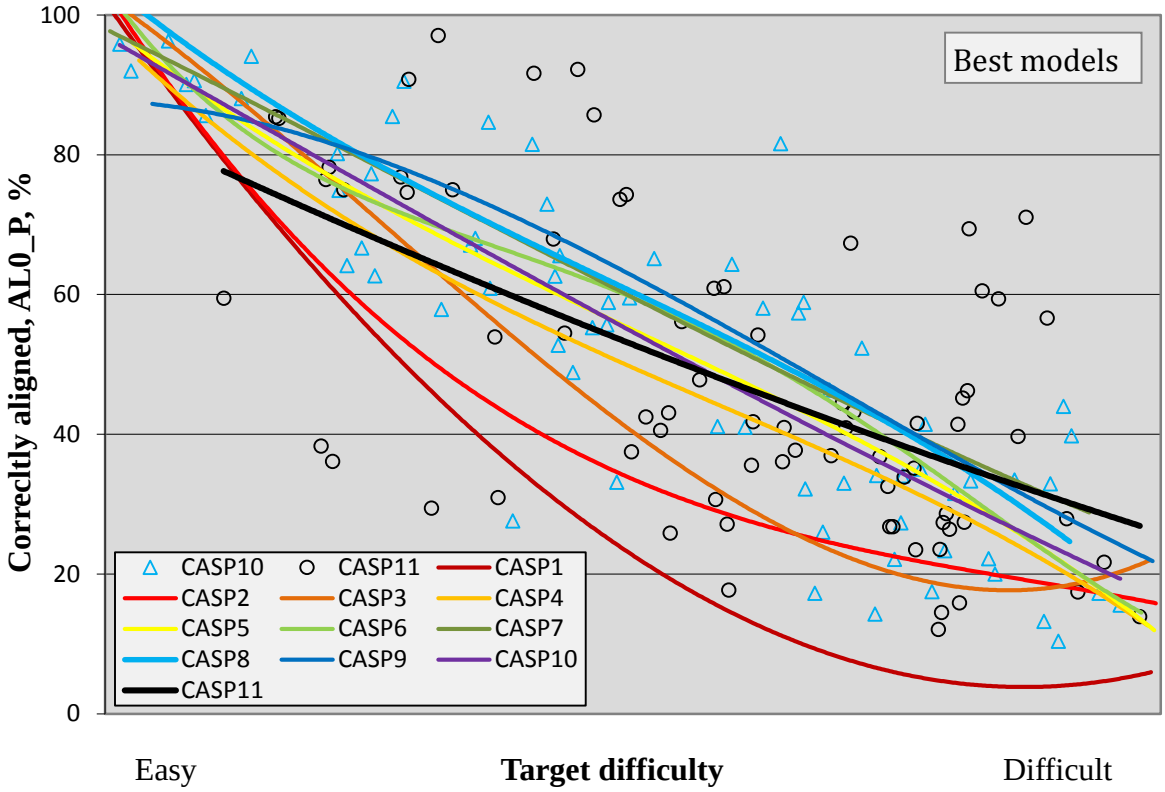
Supplementary Figure S2. Relative difficulty of targets in CASP8-11 based on the evaluation results of frozen in time methods. Difficulty for each target subset ('h', 's' and 'a' – see Fig. S1 caption) is calculated as 100-average(GDT_TS) score of the reference server (either SAM-T08 or FFAS03). The higher the bar – the more difficult the target set.



Supplementary Figure S3. Radius of gyration of CASP targets, *R* as a function of target length *L*. Dashed lines mark the boundaries ±2.5Å on either side of a line $2.77*L^{0.34}$ (not shown) derived from fitting to high resolution crystal structures. CASP11 has 16 targets with a gyration radius higher than defined by the boundaries. This is as many gyration outlier targets as in the previous two CASPs combined.

Appendix 3.

Alignment accuracy



Supplementary Figure S4. Percentage of residues correctly aligned for the best model of each target domain in all CASPs. Trend lines are similar to those in the equivalent GDT_TS plot (Fig. 5), indicating that for many targets, alignment accuracy, together with the fraction of residues that can be aligned to a single template, dominate model quality.

1. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31(13):3370-3374.
2. Zemla A, Venclovas, Moult J, Fidelis K. Processing and evaluation of predictions in CASP4. Proteins 2001;Suppl 5:13-21.
3. 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 Cas for CASP8 template-based models. Proteins 2009; 77 Suppl 9:29-49.
4. Olechnovic K, Kulberkyte E, Venclovas C. CAD-score: a new contact area difference-based function for evaluation of protein structural models. Proteins 2013;81(1):149-162.
5. 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(21):2722-2728.
6. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins 2014;82 Suppl 2:7-13.
7. Chen VB, Arendall WB, 3rd, 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(Pt 1):12-21.
8. Kryshtafovych A, Fidelis K, Moult J. CASP10 results compared to those of previous CASP experiments. Proteins 2014;82 Suppl 2:164-174.