CASP6 Data Processing and Automatic Evaluation at the Protein Structure Prediction Center

CASP6 Data Processing and Automatic Evaluation at the Protein Structure Prediction Center

Andriy Kryshtafovych, Maciej Milostan,† Lukasz Szajkowski, Pawel Daniluk, and Krzysztof Fidelis*
Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California

ABSTRACT **We present a short overview of the system governing data processing and automatic evaluation of predictions in CASP6, implemented at the Livermore Protein Structure Prediction Center. The system incorporates interrelated facilities for registering participants, collecting prediction targets from crystallographers and NMR spectroscopists and making them available to the CASP6 participants, accepting predictions and providing their preliminary evaluation, and finally, storing and visualizing results. We have automatically evaluated predictions submitted to CASP6 using criteria and methods developed over the successive CASP experiments. Also, we have tested a new evaluation technique based on non-rigid-body type superpositions. Approximately the same number of predictions has been submitted to CASP6 as to all previous CASPs combined, making navigation through and understanding of the data particularly challenging. To facilitate this, we have substantially modernized all data handling procedures, including implementation of a dedicated relational database. An overview of our redesigned website is also presented (http://predictioncenter.org/casp6/). Proteins 2005;Suppl 7:19–23.**

Key words: protein structure prediction; evaluation methods; CASP; local descriptors of protein structure

PROTEIN STRUCTURE PREDICTION CENTER IN CASP6: MAIN TASKS

The Livermore Protein Structure Prediction Center has been providing the infrastructure and automatic evaluation of predictions for the CASP experiments since 1996.[1,2] Though many elements of the CASP6 infrastructure carry over from previous CASPs,[2] the current system was reorganized and in many places essentially rewritten using newer Web technologies. We also added features recommended by the organizers, assessors, and predictors, and eliminated those rarely used.

The primary purpose of this article is to give an overview of the redesigned Prediction Center's system for data management and evaluation, as well as the tools and measures used in the automated assessment of predictions. The main tasks for this system can be outlined as follows:

- Registration of participants for the experiment and later for the meeting.
- Solicitation and selection of targets (soon to be solved protein structures), verification of target sequence data, and release of targets for prediction.
- Acceptance of protein models from both human-expert and server prediction groups; verification of prediction formats and compliance with submission/correction deadlines.
- Monitoring the public release of target coordinates, especially instances compromising the blind prediction premise of the CASP experiment.
- Automatic evaluation of prediction quality.
- Analysis and presentation of prediction results.
- Interaction with assessors, predictors, and observers.

The connections between modules of the data management system are presented in Figure 1, while details of the implementation are discussed in the following paragraphs. We concentrate on modifications and extensions to the system and present a summary for the rest.

REGISTRATION OF PARTICIPANTS

CASP participation is open to all. As in previous CASPs, CASP6 had three different types of registration. First, prospective predictors registered as traditional human-expert groups, where it was possible to use any combination of human knowledge and computations, and where there were 2–9 weeks (depending on the target expected public release date) in which to prepare models. Second,

Abbreviations: 3D, three-dimensional; AL, alignment; DAL, descriptor-based alignment; DP, domain prediction; DR, disordered regions; FN, function; FR, fold recognition; GDT, Global Distance Test; GDT_TS, Global Distance Test Total Score; NF, new fold; OLAP, online analytical processing; OLTP, online transactional processing; ORDBMS, object-relational database management system; PDB, Protein Data Bank; RMSD, root-mean-square distance; RR, residue–residue contacts; TS, tertiary structure.

†Current address: Institute of Computing Science, Poznan University of Technology, Poznan, Poland

Grant sponsor: U.S. Department of Energy; Grant number: W-7405-Eng-48 (to the University of California, Lawrence Livermore National Laboratory). Grant sponsor: National Institutes of Health; Grant number: LM07085-01 (to K. Fidelis).

*Correspondence to: Krzysztof Fidelis, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551. E-mail: fidelis@llnl.gov

Received 13 May 2005; Accepted 30 June 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20718

This article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version.

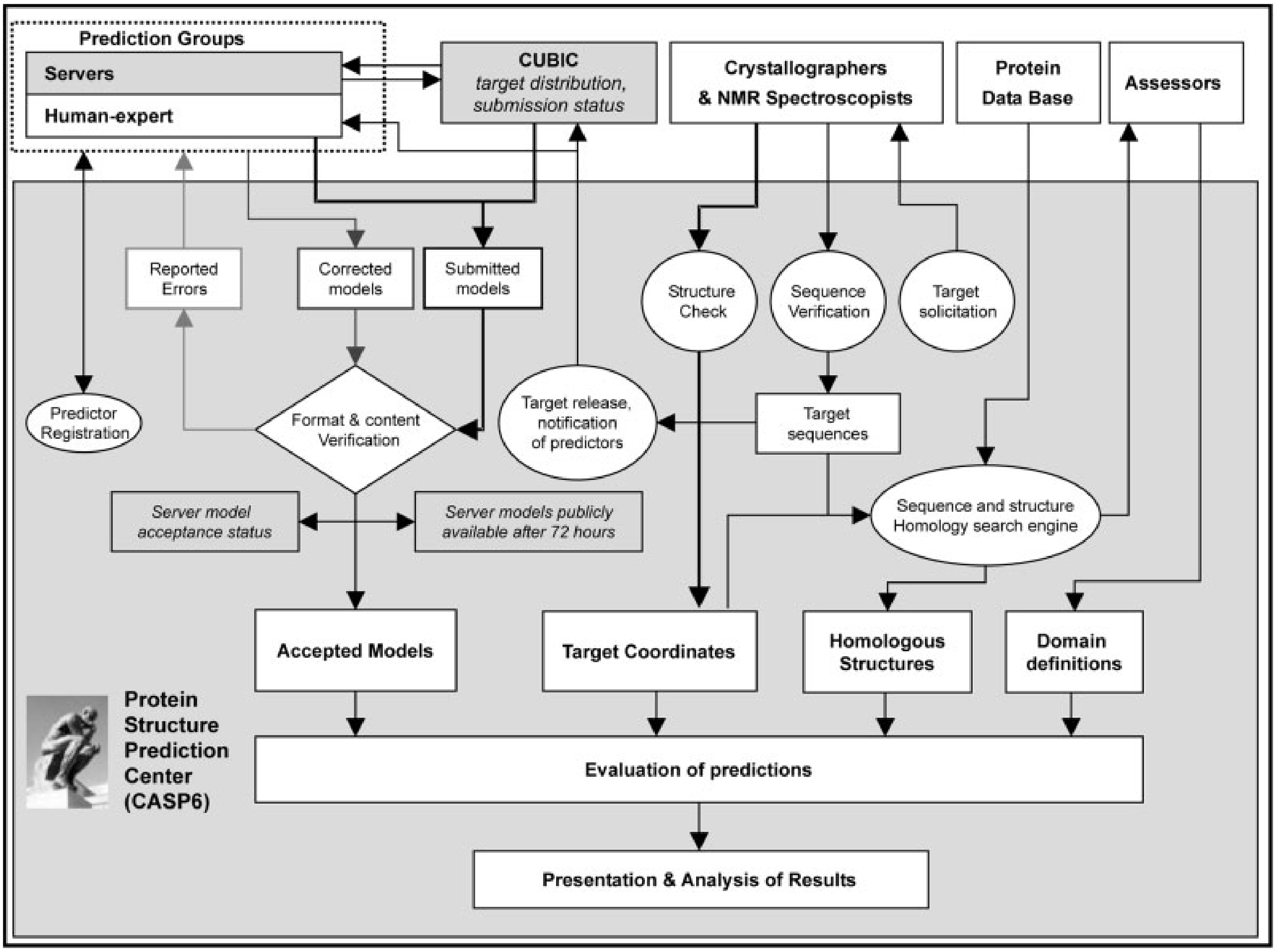
Prediction Groups
Servers
Human-expert
CUBIC
target distribution, submission status
Crystallographers & NMR Spectroscopists
Protein Data Base
Assessors
Reported Errors
Corrected models
Submitted models
Structure Check
Sequence Verification
Target solicitation
Predictor Registration
Format & content Verification
Target release, notification of predictors
Target sequences
Server model acceptance status
Server models publicly available after 72 hours
Sequence and structure Homology search engine
Accepted Models
Target Coordinates
Homologous Structures
Domain definitions
Protein Structure Prediction Center (CASP6)
Evaluation of predictions
Presentation & Analysis of Results

Fig. 1. Schematic of the CASP6 data management system.

predictors could register as server groups, which implied that target sequences were automatically distributed to the registered server machines, with models to be returned within 48 hours without any human intervention. Predictors planning to provide both *human-expert* and *server* models had to register separately for each. Third, those not intending to submit predictions but interested in the process and in receiving mailings concerning progress of the experiment had an opportunity to register as observers.

Registration was made available through the Prediction Center website. Each participant registering for a specific group was automatically assigned a number encoding his/her CASP status (*human-expert*, *server*, *observer*) and role/privileges in the group (*leader* or *member*). Group leaders were asked to specify the name for the group and only they subsequently had access to submit/correct predictions. The registration information was stored in a database linked to the validation module of the prediction acceptance software, where predictor registration codes were used to verify identity of a submission and the correspondence with the deadlines (set differently for human-expert and server groups).

Registration for CASP6 included 289 predictor groups representing 25 countries; of these, 228 have actively participated in the experiment (165 human-expert and 63 server). Assessment included 208 groups; 19 were disqualified for submitting predictions that were not original, and one group withdrew its results. Please note that because the same people could have participated in more than one human-expert and/or server group, these numbers overestimate participation of different scientific centers. By combining groups with at least one predictor in common, we obtain 137 clusters, which can be interpreted as the lower bound for the number of research groups actively participating in CASP6.

PREDICTION TARGETS

X-ray crystallographers and NMR spectroscopists around the world actively participated by submitting the amino acid sequences of potential CASP6 targets (i.e., structures either about to be solved or not yet publicly released). Structural genomics projects contributed significantly, providing the majority of CASP6 targets. Accepted sequences were prescreened for homology with known structures, so that a daily package of released targets would be difficulty balanced. No more than three targets, containing at most 700 residues total, were released on any single day. During the CASP6 prediction season (June 7 through September 2), a total of 87 targets was released.

We have set up an automatic monitoring of the structure databases to identify any release of pertinent information happening prior to target deadline. In this regard we have received substantial help from many of the CASP participants, who monitored conference programs, publications, and structural genomics websites (especially warm thanks to Alexey Murzin). Unfortunately, we did have to cancel 11 targets because of structural information leaks during the prediction season. To reduce the number of these clearly undesired occurrences, substantially shorter prediction windows are considered for CASP7 and future experiments.

SUBMISSION OF PREDICTIONS

All CASP6 predictions were collected at the Livermore Prediction Center. The main difference between the CASP5 and CASP6 submission processes lied in the reorganized software system and in accepting server type predictions, which is outlined in more detail in a separate article in this issue.

For CASP6, the prediction acceptance system was integrated into a multifunctional data management system, designed to handle in one package (1) acceptance of predictions, (2) evaluation, and (3) presentation of results. Improvements were dictated by the demands of the ever-increasing number of submissions. As previously, predictors could submit their models either through an HTML form or by sending e-mail to our verification server. A maximum of five models per group per target were allowed. The acceptance module of the system processed submissions in the following order. First, the software read in the required information headers and checked whether a submission arrived through the proper communication channel (different gateways were assigned to human-expert, original server, and corrected server submissions), and verified the author registration data and the compliance with submission deadlines. Each submission was assigned a PIN number (if a server submission contained more than one model, up to five models for one target were accepted and assigned a PIN number separately). The parser then verified compliance with one of the six CASP6 formats[3]:

1. TS—atomic coordinates in the Protein Data Bank (PDB) format[a]
2. AL—sequence alignments to structures available in the PDB
3. RR—distances between residues close in three-dimensional (3D) space ($|C^i_\beta - C^j_\beta| < 8$ Å)
4. DR—residues in the protein sequence corresponding to disordered regions
5. DP—number of domains and domain residue assignments
6. FN—prediction of the target protein function.

Submissions were verified for their consistency with the sequence of the target. In addition, all references to the PDB, especially for the AL-type submissions, were verified. Properly formatted predictions were then assigned an accession number composed of the prediction target ID, format category designator, predictor group number, and model index (1 through 5 as assigned by predictors to rank their models starting from the best). Acceptance notifications or error messages with suggestions on how to amend format inconsistencies were mailed back to predictors. Subsequent successful submissions of a duplicate model (same target, group, model index) replaced the previously accepted model if received prior to the particular target prediction deadline.

Finally, using the system described above, more than 41,000 predictions in all six categories were accepted for evaluation. This number includes almost 34,000 3D structure predictions, which approximately equals the number of 3D predictions submitted to CASP1–5 combined.

The increased number of submissions to CASP6 prompted our concerns regarding the reliability of data management, as well as the speed and ease of the subsequent prediction assessment. To address this, several improvements described in the following sections have been implemented.

DATA MANAGEMENT SYSTEM DESIGN

The PostgreSQL–ORDBMS was used to store predictions and results of evaluation. It was chosen because of stability, modern features, open source, and liberal license. The database schema was constructed with the OLTP paradigm in mind. It means that the database was prepared for fast processing of large number, concurrent, and nonanalytical transactions. Such a schema is ideal for data gathering process, but not for more complex data analysis. During CASP6, the database evolved toward the OLAP paradigm due to the necessity of generating a wide range of statistics and status pages. The database incorporated indexes, relations, views, and materialized views refreshed in a timely manner. Certain tables were clustered, and the database parameters were tuned to provide a faster access to the relevant information.

The evaluation process was managed by the scheduler connected to the database of collected predictions and target structures. The scheduler distributed tasks to a set of 24 processors, uploading necessary structures and running Perl scripts controlling evaluation. This allowed us to optimize the evaluation process for a set of targets or to obtain data just for a selected group, target, or type of evaluation. As in CASP4 and CASP5, the LGA[4] program was used to find best structural model–target superposition. We also tested our new software to estimate the level of structural similarity between model and target in a non-rigid-body regimen.

Dynamic management and presentation of evaluation results, realized through a set of Perl scripts connected

[a] A specially formatted TS template was prepared for target T0206, enabling predictors to submit quaternary structure models for this trimer (information provided by the crystallographer).

with the database, enabled expanding existing reports without regenerating all the content. A user-friendly organization of the system allows for easy handling of statistics and visualization. For example, results can be multisorted by different columns according to user preference, and each dynamic table can be easily transformed into a chart or a plain text file to be parsed by an external script. However, the advanced functionality, together with the large size of the Web page content (500 rows per page), necessitated productivity optimization. To speed up the process of loading data, a special caching mechanism was employed and the Web server was reconfigured to compress data before sending it to a client. In addition, all static links were replaced by the hashed dynamic equivalents. These improvements, together with the database and query optimization, substantially reduced the amount of transmitted information (for a single Web page the size was reduced from 800 kB to 30 kB), thus considerably shortening the response time.

Outline of the Similarity Measures Used in CASP6 Automatic Evaluation

Techniques for comparing models with targets remained generally intact compared to previous rounds of CASP.[2] The structural alignment and analysis packages ACE[2] and LGA[4] (2004 version) were used for these calculations. Here we provide a short overview of the main evaluation measures (and their graphical representations) incorporated in or based on results of the aforementioned packages, referring the reader to the previously published accounts[2,4] for a more detailed description:

- *RMSD.* The root-mean-square-deviation between model and target calculated for all atomic positions, their subsets (C_α's, main-chain, and side-chain atoms) and over dihedral angles (separately for ϕ/ψ, and χ angles).
- *NP_P* reports percent of residues predicted in the model.
- *GDT, GDT_TS (basic measure), and GDT summary graphs.* A GDT value reports the largest number of residues in a prediction that can be fitted to the target structure in the sequence-dependent mode under a specified distance cutoff D0. GDT_TS score is an average of 4 GDT values calculated at D0 = 1, 2, 4, and 8 Å and normalized by the number of residues in target structure. GDT summary graphs comprise all predictions submitted on a given target, with quality of each prediction represented by a separate line with the percent of the superimposed residues plotted along the x axis and the corresponding distance cutoffs along the y axis. In general, the smaller the area under the curve, the better the prediction. Each line also serves as a link to the corresponding 3D superposition of model and target.
- *AL0 and alignment accuracy charts.* The alignment accuracy measure AL0 reports the number of correctly aligned residues in the LGA 5Å sequence-independent superposition of the modeled and experimental structures of a target. A model residue is considered to be correctly aligned if its C_α atom falls within 3.8 Å of the corresponding atom in the experimental structure, and there is no other experimental structure C_α atom nearer. The strip charts (cumulative and sequence-based views) show correspondingly the number and location of the correctly aligned residues, those aligned within a ±4 residue window, and those aligned with an error of more than four residues. Each strip on the graph provides a link to the corresponding 3D superposition of model and target.

Descriptor-Based Analysis of Structural Similarity

In CASP6 we tested a new method capable of identifying multiple independent similarity regions in comparing structures. The method of Local Descriptors of Protein Structure is based on the non-rigid-body structure comparison philosophy and, from the CASP perspective, seems to be especially useful in case of weak structural similarity, significant structural shifts in models or multidomain targets. It focuses on the similarity of local structure in the sense of noncontiguous 3D context. In CASP6, the new method was applied to all FR and NF targets and its results (DAL scores) are accessible through the Prediction Center's CASP6 Web pages. More details on the method and interesting examples can be found at http://predictioncenter.org/local/DALExamples/. A separate article describing the approach will appear shortly.

C_α Clash and Model Similarity Checks

We have also checked distances between C_α atoms and reported (1) geometric irregularities ($0.1\ \text{Å} < |C^i_\alpha - C^j_\alpha| < 3.6\ \text{Å}$ or $|C^i_\alpha - C^{i-1}_\alpha| > 4.0\ \text{Å}$), (2) severe collisions ($0.1\ \text{Å} < |C^i_\alpha - C^j_\alpha| < 1.9\ \text{Å}$), and (3) instances of atoms occupying essentially the same space ($|C^i_\alpha - C^j_\alpha| < 0.1\ \text{Å}$). These checks identify isolated clashes but also more general tendencies to compress/expand models by any particular approach. We have also performed clustering of predictions on each of CASP targets to identify identical or very similar predictions.

ORGANIZATION OF THE WEBSITE

The Protein Structure Prediction Center website provides general information about the prediction experiment and access to prediction targets, original predictions, evaluation results and their visualization. Data for all six CASP experiments are available. Here we provide guidelines for navigation through the results of the automatic evaluation of 3D predictions from the latest experiment. The CASP6 website allows users to look at the evaluation data from three different perspectives:

1. *Targets first perspective* is the default viewing mode and allows selecting a target (domain) and checking the performance of all prediction groups submitting on that target. For a selected target it is possible to access:
   a. A sortable and expandable table containing results for all groups according to the sequence-dependent, sequence-independent, and local descriptor analysis (the latter for FR and NF targets only)

b. Results text files and RASMOL[5] visualizations of the LGA model-target structural superpositions (links embedded into table (a) under section *Charts*)
c. Alignment maps and non-rigid-body structural superpositions (descriptor-based approach, links to results in table (a), section *Descriptor alignment*)
d. GDT plot (LGA sequence-dependent analysis)
e. Alignment strip charts (LGA sequence-independent analysis)
f. Domain boundary prediction plots.

The main *Results* Web page is designed with miniature GDT and alignment plots assembled in one place for easy comparison. An information cell for each target contains GDT and alignment plot pictographs and a header (area above the pictographs). The pictographs are linked to the full size plots (d) and (e). A GDT plot with curves for a selected group highlighted is also accessible from table (a), link *G*. The headers start with a target (domain) ID hyperlink leading to tables of type (a). Headers for one-domain targets (transparent background), and separate domains from multidomain targets (white background) specify the abbreviated target category. Headers for separate domains also specify the range(s) of corresponding residues. Headers for the whole multidomain targets (gray background) also contain a *DP* link pointing to domain boundary plots (f), *Domains* link pointing to the “all domains at once” results table, and the total number of domains identified for that target (in parentheses).

2. In the *groups first perspective*, the data are organized so that the information is group anchored. Groups in the list are sorted alphabetically by the group name. For a selected group it is possible to access:
   a. A dynamic table containing evaluation results for this group only
   b. A rank table displaying relative place of the group for each attempted target according to the GDT_TS and AL0 scores
   c. A page containing pictographs of GDT and alignment plots for all targets, with the results for the selected group highlighted (this way the relative performance of the group on all targets can be assessed from a single page)
   d. A page enabling visual comparison of up to four groups.
3. *Comparative analysis perspective* allows users to generate tables containing results for a selected subset of groups, targets, and evaluation measures. The tables also provide links to graphical representation of the data. It is possible to choose only server predictions for this type of analysis.

Note that it is possible to switch between different viewing modes not only from the main CASP6 Web page but also using the menus at any of the results pages.

REFERENCES

1. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;Suppl 3:22–29.
2. Zemla A, Venclovas, Moult J, Fidelis K. Processing and evaluation of predictions in CASP4. Proteins 2001;Suppl 5:13–21.
3. Available online at http://predictioncenter.org/casp6/doc/casp6-format.html
4. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
5. Sayle RA, Milner-White EJ. RASMOL: biomolecular graphics for all. Trends Biochem Sci 1995;20:370–374.