PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Evaluation of residue–residue contact prediction in CASP10

Bohdan Monastyrskyy,[1] Daniel D'Andrea,[2] Krzysztof Fidelis,[1] Anna Tramontano,[2,3] and Andriy Kryshtafovych[1]*

[1] Genome Center, University of California, Davis, California 95616

[2] Department of Physics, Sapienza—University of Rome, 00185 Rome, Italy

[3] Istituto Pasteur—Fondazione Cenci Bolognetti—University of Rome, 00185 Rome, Italy

ABSTRACT

We present the results of the assessment of the intramolecular residue-residue contact predictions from 26 prediction groups participating in the 10th round of the CASP experiment. The most recently developed direct coupling analysis methods did not take part in the experiment likely because they require a very deep sequence alignment not available for any of the 114 CASP10 targets. The performance of contact prediction methods was evaluated with the measures used in previous CASPs (i.e., prediction accuracy and the difference between the distribution of the predicted contacts and that of all pairs of residues in the target protein), as well as new measures, such as the Matthews correlation coefficient, the area under the precision-recall curve and the ranks of the first correctly and incorrectly predicted contact. We also evaluated the ability to detect interdomain contacts and tested whether the difficulty of predicting contacts depends upon the protein length and the depth of the family sequence alignment. The analyses were carried out on the target domains for which structural homologs did not exist or were difficult to identify. The evaluation was performed for all types of contacts (short, medium, and long-range), with emphasis placed on long-range contacts, i.e. those involving residues separated by at least 24 residues along the sequence. The assessment suggests that the best CASP10 contact prediction methods perform at approximately the same level, and comparably to those participating in CASP9.

Proteins 2014; 82(Suppl 2):138–153.

Key words: CASP; residue-residue contact prediction; RR.

INTRODUCTION

Inter-residue contacts have been shown instrumental in reconstructing protein backbones by means of distance geometry or restrained molecular dynamics.[1–3] This finding suggested that the prediction of intramolecular contacts in proteins can serve as an intermediate step toward accurate prediction of the three-dimensional structure, and triggered extensive research to connect protein sequence and structure with a "two-span bridge": from sequence to contacts and from contacts to structure. To build such a bridge, the researchers focused on predicting contacts with accuracy sufficiently high to be useful for structure modeling on one side, and on building a structure from incomplete/inaccurate contact data, on the other.

As far as the area of structure rebuilding is concerned, a series of papers published in the 1990s demonstrated that protein contact maps can indeed serve as scaffolds for building protein structures even when the maps are sparse or contain just a fraction of correct contacts.[4–8] A few features related to the tolerance of these methods to data uncertainty and incompleteness were discovered. In particular, in a pioneering work,[1] Havel *et al.* speculated that it is better to know many distances imprecisely rather than a few distances accurately. Saitoh *et al.*[5] noticed that the only factor largely influencing the

Additional Supporting Information may be found in the online version of this article.

Abbreviations: FM, free modeling; MCC, the Matthews correlation coefficient; RR, residue-residue (contacts); TBM, template-based modeling.

Grant sponsor: US National Institute of General Medical Sciences (NIGMS/NIH); Grant number: R01GM100482 (to K.F.); Grant sponsor: KAUST Award; Grant number: KUK-I1–012-43 (to A.T.); Grant sponsor: EMBO.

Bohdan Monastyrskyy and Daniel D'Andrea contributed equally to this work

*Correspondence to: Andriy Kryshtafovych, Genome Center, University of California, Davis, 415 Health Sciences Dr., Davis, CA 95616. E-mail: akryshtafovych@ucdavis.edu

Received 18 April 2013; Revised 14 May 2013; Accepted 21 May 2013
Published online 12 June 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24340

quality of the reconstructed structures is the long-range geometrical constraint. Skolnick *et al.* suggested[7] that knowing contacts for one in every seven residues would be sufficient to recover the structure of short proteins. Later, Vassura *et al.*[9] claimed that knowing one in four actual contacts might be enough to facilitate rebuilding tertiary structure with 5 Å accuracy. Although in general it is still unclear what accuracy, coverage, and distribution of contacts along the sequence are needed to be useful in practice, it has become common knowledge that information on just a few correct contacts can be valuable for improving structure prediction. This is especially true for the long-range contacts, which impose strong constraints on the three-dimensional structure and effectively narrow the search space of possible conformations. The usefulness of the contact approach was illustrated in the current edition of CASP, where predictors in the newly introduced contact-assisted structure prediction category (see the contact-assisted assessment article, this issue) were able to build substantially better models using information provided by the organizers on some of the long-range contacts in the target structures. Other studies also report that incorporating contact information into protein folding programs such as Rosetta and I-TASSER leads to improvement of the 3D models.[10,11]

Returning to the first bridge span in the "two-span bridge" analogy, substantial attention was dedicated to the prediction of intramolecular contacts. Much of the research in this area stemmed from the hypothesis of correlated mutations, suggesting that pairs of residues that mutate in a coordinated fashion during evolution are likely to be in contact. In the 1990s, the first articles demonstrating the applicability of this idea to contact prediction were published.[12–14] After these promising results, a series of contact prediction methods developing this concept further appeared in the literature.[15] Quite recently, the 20-year-old idea received a new twist as several articles claimed improved accuracy of contact prediction through disentangling the direct pairwise couplings from the background network of coordinately mutating positions.[15–22] Besides the coordinated mutations approaches, many other contact prediction methods were developed based on different or hybrid methodological concepts. In general, they are based on machine-learning techniques incorporating sequence-related features such as the sequence evolutionary profile of the target, secondary structure, and solvent accessibility—to name just a few. These methods use neural networks,[23–29] support vector machines,[30–32] hidden Markov models,[33–35] genetic algorithms,[36] random forest models,[37] and learning classifier systems.[38] Many of the methods mentioned above were tested in CASP experiments achieving different levels of success.

The prediction of residue-residue contacts has been a part of the CASP experiment since CASP2[39] (1996), however, the prediction format and the assessment procedures have been standardized only in CASP6–CASP9.[40–43] For CASP10, we developed an infrastructure for an automatic evaluation of the RR predictions and visual analysis of the results.[44] Here we analyze the results obtained by groups participating in CASP10 and quantify progress in the area compared with the previous CASPs.

MATERIALS AND METHODS

RR prediction format and definition of a contact

The RR prediction format and definition of intramolecular contacts in CASP10 have not changed since previous rounds of CASP. A pair of residues is defined to be in contact when the distance between their C_β atoms (C_α in case of GLY) is less than 8.0 Å. Depending on the separation along the sequence, short-, medium- and long-range contacts are between residues separated by 6 to 11, 12 to 23, and at least 24 residues, respectively. The contacts with a separation of less than six residues are not considered as they typically correspond to contacts within secondary structure elements. The participating groups were asked to submit a list of pairs of residues predicted to be in contact. Each reported contact had to be annotated with a probability score in the [0;1] range, reflecting the predictor confidence in assigning the contact. Unlike the previous rounds of CASP, only one set of contact predictions per target was allowed in CASP10 for each participating group.

Sets of domains evaluated

The evaluation of predictions was carried out on a per-domain basis. The domains with detectable homology to proteins of known structures were not included in the evaluation as in these cases contacts could easily be derived from the template structures. Thus, we used only the domains for which structural templates did not exist or were very difficult to identify, that is, the domains classified in the FM, TBM/FM, or TBM_hard categories.[45] The complete list of CASP10 domains with their classifications is available at http:/predictioncenter.org/casp10/domains_summary.cgi.

We assessed the performance of contact prediction methods on two sets of domains.

Set 1 (denoted as "FM") comprises 15 FM and 1 FM/TBM domains. For these domains templates did not exist or could not be reliably identified based on the target sequence. Set 1 is our main evaluation set and is consistent with the sets used in previous rounds of CASP.

Set 2 (hereinafter referred to as "FM + TBM_hard") is an extension of the previous set obtained by adding the domains from the TBM_hard category (13 entries). These are the hardest TBM targets, for which

templates exist but are hard to identify or to properly align with the target. As a consequence, the scores of all submitted three-dimensional models for these targets were rather poor, not exceeding 50 GDT_TS units.[45]

We also performed the assessment on two sets of targets generated from the original two sets by eliminating non-globular proteins consisting of repeated structural blocks: Set 1R = Set 1 – {T0653-D1, T0695-D1}, and Set 2R = Set 2 – {T0653-D1, T0671-D2, T0690-D1, T0695-D1}.

The first three targets removed from Set 2 are the well-known leucine-rich repeats,[46] while the last one is a three-helical spectrin bundle repeated five times.[62] All four structures are built with repeated structural blocks for which good templates exist. Since the majority of contacts for these domains could be derived from the templates, their inclusion could introduce a bias in the evaluation. In practice, differences in the results on the original and the reduced sets were minor for the majority of analyses, and therefore we present here the results only for the original datasets, except for the domain-length dependence analysis, where using the reduced sets is more appropriate.

An estimate of the difficulty of individual domains for contact prediction is shown in Supporting Information Figure S1.

Sets of evaluated contacts

To compare the performance of contact prediction methods we used two different approaches. In the first approach, we trimmed the predicted lists of contacts to the same number of contacts per target (see the Reduced contact lists subsection below); in the second, we "padded" the lists by assigning a probability value of 0 to all non-listed contacts. The both procedures ensure that the participating groups are compared on the same number of contacts.

Preprocessing of predictions

For multidomain targets, we extracted the lists of inter-residue contacts for each individual domain. This step was necessary as predictions were submitted for the entire targets, but evaluated on a per-domain basis (see above). We also considered contacts between residues from different domains as their correct prediction can be useful in predicting the orientation of the interacting domains.

For each prediction, we separated short-, medium-, and long-range contacts and assessed them independently. The medium and long-range contacts were also assessed together.

Reduced contact lists

For every domain, the lists described above were trimmed to the $L/5$ and $L/10$ contacts predicted with higher probability (L is the length of the domain). The number $L/5$ (or $L/10$) is rounded to the closest integer, and if there are multiple entries corresponding to the same probability they are considered in the order provided by the predictor. To be included in the evaluation, the filtered list of contacts had to comprise at least $L/5$ or $L/10$ contacts. In order to assess also the groups that submitted only very small numbers of contacts, we also evaluated predictions on the five contacts with the highest assigned probability values, regardless of the domain length.

Thus, for every group we generated 12 reduced lists of contacts per predicted domain, whenever possible. The results for all lists of contacts and all contact range categories are available at http:/predictioncenter.org/casp10/rr_results.cgi. In this paper we focus on the results for the $L/5$ lists of long-range contacts. The numbers of domains predicted on these datasets for each of the participating groups are summarized in Figure 1. Two groups (G334 and G077) submitted just a few predictions for the evaluated domains and one (G246) did none, so we excluded them from the analysis and present the results on the reduced lists for the remaining 23 groups. For every group, the final scores on the reduced datasets are averages of the per-domain scores.

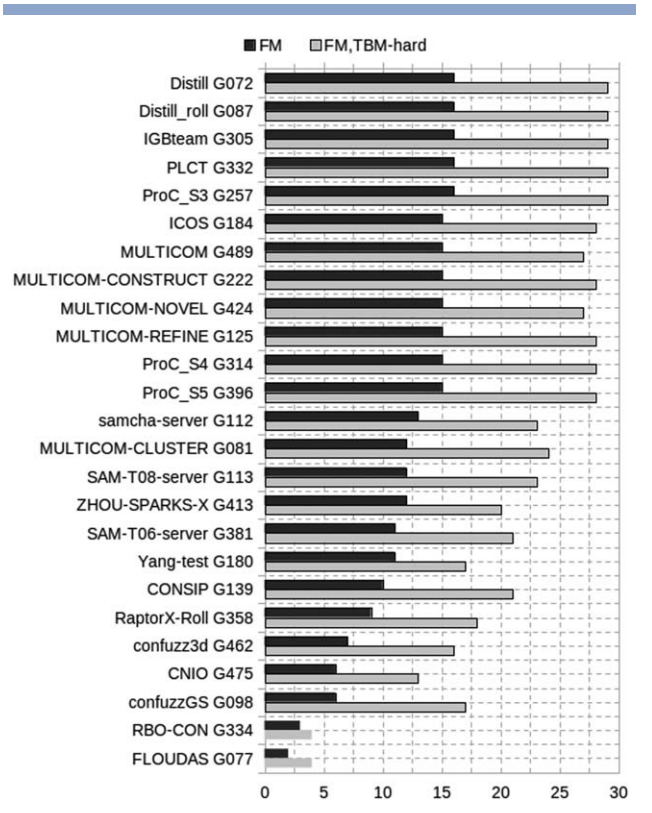

Figure 1

Number of domains per group for which the L/5 list of long-range contacts were evaluated. Two groups RBO-CON (G334) and FLOUDAS (G077) submitted too few predictions and are not included in the subsequent analyses.

Padded contact maps

As contact probability maps generated from submitted predictions are sparse, they are usually unsuitable for many analyses that require complete predictions (i.e. we need each pair of residues to be predicted either in contact or not). We remediate the "sparseness" problem here by setting the values of the empty cells of contact probability maps to zero ("padded" lists). In other words, pairs of residues that are missing in predictions are considered as non-contacts. Under such assumption, each prediction list classifies every pair of residues within the selected range to one of the four cases: TP, correctly predicted contact; FP, non-contact predicted as contact; TN, correctly "predicted" non-contact (i.e., the non-contact not included in the predicted contact list); and FN, contact "predicted" as non-contact (i.e., the contact missing in the submitted list).

We only assessed the groups that submitted predictions for at least 10 domains on the "padded" datasets—these are the same 23 groups as above, plus group G334. As in the case of the reduced contact lists, in this article we concentrate on the analysis of the performance of the participating groups for the long-range contacts only. Differently from the assessment on the reduced contact lists, the final group scores on the padded datasets are calculated from the data on all domains pooled together.

Evaluation procedure

In CASP10 we have substantially expanded the set of evaluation tools to assess residue-residue contact predictions. Besides the methods used in the previous CASPs, we introduced several new evaluations providing an alternative point of view on methods' performance. While in previous CASPs the assessors analyzed the results exclusively on the "reduced" datasets, implicitly concentrating on two aspects of contact prediction: (1) how good are methods in identifying the most reliable predicted contacts and (2) how accurate are the methods in predicting contacts with the highest reliability, in this CASP we complemented the assessment with analyses on the full sets of contacts addressing the issue of how accurate are all submitted contact predictions, including those predicted with lower reliability. Below, we briefly outline all evaluation procedures, focusing in more detail on the new evaluation measures.

Basic scoring functions and group performance on the reduced datasets

Since CASP6, predictions in the RR category have been evaluated on the reduced contact lists using two main scores: precision = TP/(TP + FP), and Xd. The detailed description of these scores can be found in the previous CASP contact assessment articles.[40–43] Note, that in those papers the measure defined by the formula TP/(TP + FP) was called "accuracy" (Acc); here we have changed its name to "precision" to be consistent with the classic descriptive statistics definition. The precision-based results are discussed in the main text of this article, while the Xd-based results are shown in the Supporting Information.

Based on these two scores, the performance of groups was further compared with two strategies: cumulative *z*-score ranking (sum of precision-based and Xd-based *z*-scores) and "head-to-head" comparisons.[43]

Evaluation measures for the padded datasets

Matthews' correlation coefficient and other binary descriptive statistics measures

For the assessment of the effectiveness of the predictive methods as binary classifiers we used four evaluation measures.

The first two are precision and recall, a.k.a. sensitivity:

$$precision = \frac{TP}{TP+FP}, \qquad recall = sensistivity = \frac{TP}{TP+FN}.$$

They were already used in previous CASPs, but were shown to be equivalent on the reduced prediction sets.[41] On the complete datasets, precision and recall are not inter-dependent any more as the number of predicted contacts is different for different predictions. Based on the formulae, one can notice that each of these measures takes into account only two out of the four parameters of prediction quality (TP, FP, TN, and FN) and therefore focuses on the specific aspects of predicting contacts only (ignoring non-contacts).

The *F*-score is a more comprehensive measure as it combines precision and recall

$$F_1 = 2\frac{precision * recall}{precision + recall}$$

and inherits useful features typical to both measures. However, the *F*-measure still does not take the true negative rate into account.

Even though employing measures that take all parameters of contact prediction into account may seem beneficial, it should be approached with caution, as in our case two binary classes of prediction (contacts and non-contacts) are disproportionally distributed in the structure (contacts constitute just a small fraction of all pairs of residues). As it was discussed in the CASP9 disorder assessment article,[47] the Matthews correlation coefficient (MCC)

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

is a well-suited measure for handling cases with imbalanced class frequencies. The MCC was shown to provide a more appropriate account of the skewed data than many other methods, and not to favor over-prediction of

Table I
The Publicly Available Contact Prediction Servers Participating in CASP10

Server name and URL address	CASP10 group	Brief description of the method
CMAPpro[a]. Available at: http:/scratch.proteomics.ics.uci.edu/	G305	Deep neural networks architecture allowing progressive refinement of contact prediction.
Distill, Distill-roll. Available at: http://distill.ucd.ie/distill/	G072 and G087	Two-dimensional-recursive neutral networks.
ICOS. Available at: http:/icos.cs.nott.ac.uk/servers/psp.html	G184	Inhouse machine-learning technique taking into account nine-residue window profiles, secondary structure, and other features.
MULTICOM-CLUSTER. Available at: http:/casp.rnet.missouri.edu/svmcon.html	G081	An SVM tool. The input data include secondary structure, solvent accessibility, and sequence profile.
MULTICOM-CONSTRUCT[a]. Available at: http:/iris.rnet.missouri.edu/dncon/	G222	Ensembles of deep networks.
MULTICOM-NOVEL,MULTICOM-REFINE. Available at: http:/casp.rnet.missouri.edu/nncon.html	G424 and G125	Recursive neural networks. MULTICOM-REFINE has a separate module to predict contacts in beta-sheets.
PROC_S3. Available at: http:/www.abl.ku.edu/proc/proc_s3.html	G257	Random Forest models incorporating more than 1000 sequence-related features.
SAM-T06, SAM-T08. Available at: http:/compbio.soe.ucsc.edu/SAM06/ and http:/compbio.soe.ucsc.edu/SAM08/	G381 and G113	Recursive neural networks using the correlated mutations in MSA.
Samcha-server[a]. Available at: http:/binfolab12.kaist.ac.kr/conti/	G112	SVM incorporating more than 800 sequential features.

[a]New methods according to the CASP10 Abstract Book.

any classes. Therefore, in this article we consider this measure as the main estimator of binary classifiers on the expanded datasets.

Precision-recall curve analysis

In previous rounds of CASP, the probability score assigned to every predicted contact was used in assessment only to select the most reliable contacts (according to the predictors' estimates) for the reduced evaluation datasets. However one can argue that the probability score holds valuable information that can be used both in modeling of the structure and in assessment. For example, it can be used to test the ability of predictors to correctly rank the predicted contacts and select the proper cut-off separating contacts (positive cases) from non-contacts (negative cases).

To address these issues we carried out the analysis based on the precision-recall (PR) curves, which are widely used in statistical evaluations of disproportional datasets.[48–51] The PR-curve analysis is conceptually similar to the well-known ROC-curve analysis,[52] but differs in that the parametric curves are plotted in the (recall, precision) coordinates. Davis and Goadrich[53] proved that the dominant curve in ROC space corresponds to the dominant curve in PR space and vice versa, and showed that the curves in PR space may be more informative for skewed data, as ROC curves tend to provide overly optimistic results in such cases.

In essence, a PR-curve illustrates the relationship between the precision and recall of a predictor for a set of probability thresholds. For each threshold, a record (pair of residues in our case) is considered as a positive example (contact) if its predicted probability is equal to or greater than the threshold value. The area under the PR-curve, AUC_PR, is indicative of the classifier's accuracy, with a value of 1 corresponding to a perfect predictor. The AUC_PR values were calculated using the software developed by Davis and Goadrich[53] and freely available from their website.[54]

The Jaccard distance for clustering methods

The dissimilarity between two groups for each target is defined in terms of the Jaccard distance:[55]

$$J_{ij} = (M_{01} + M_{10})/(M_{11} + M_{01} + M_{10}),$$

where M_{11} is the number of common contacts predicted by groups i and j, M_{10} and M_{01} are the contacts only predicted by group i and j, respectively. The *J-score* has values in the range of [0;1], with the value of 0 corresponding to identical predictors and 1 - to completely dissimilar ones.

The tie-breaking procedure for defining the first correct/incorrect contact

If prediction contains several contacts with the same probability value, the position of the first correct/incorrect prediction is assigned regardless of whether there are incorrect/correct predictions with the same probability. In other words, if the correct prediction with the highest probability has the same probability, and therefore the same rank R, as one or more incorrect predictions, the correct prediction is assigned rank R. Analogously, the position of the first incorrect prediction is assigned regardless of whether there are correct predictions with the same probability, i.e. if the first incorrect prediction

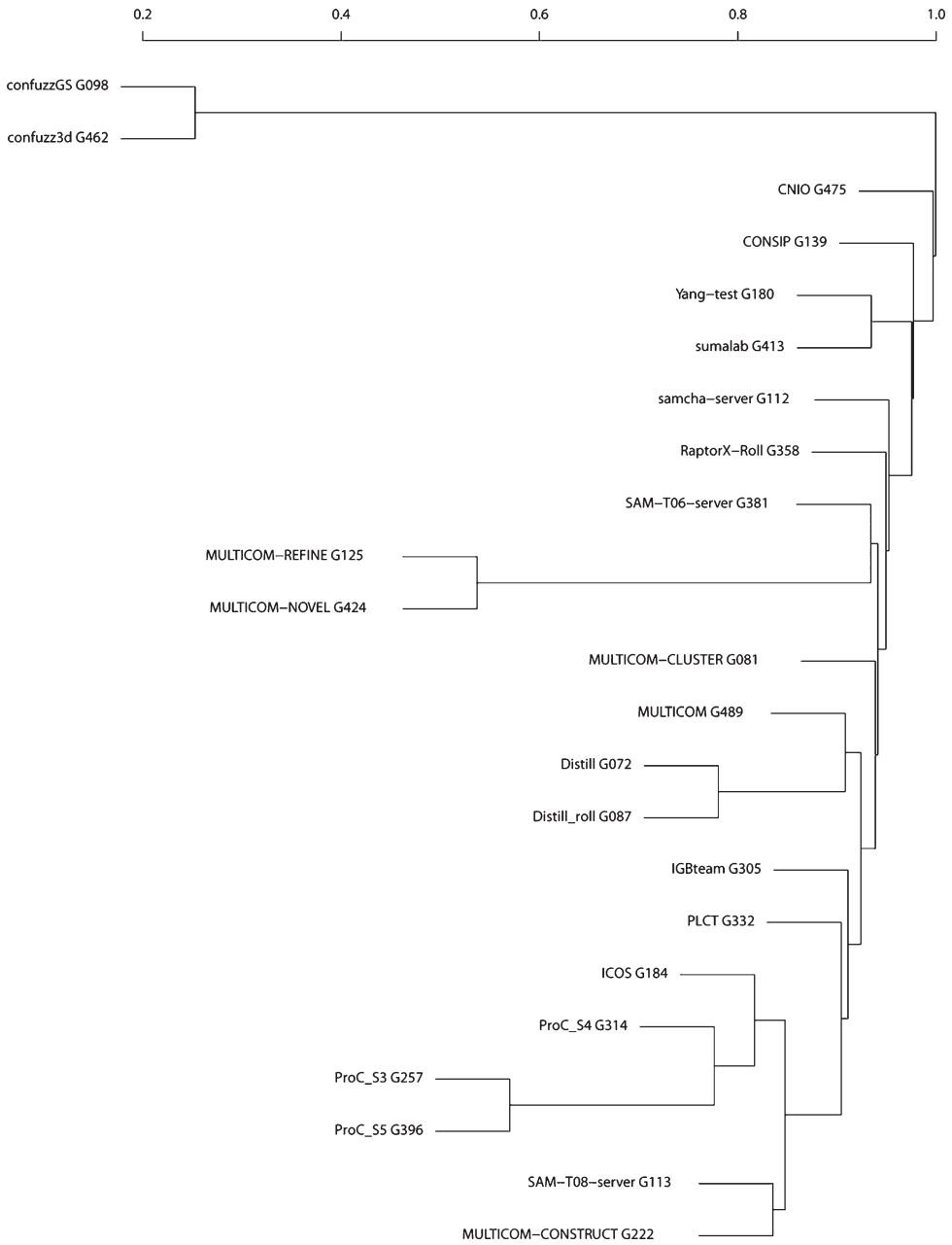

Figure 2
Dendrogram illustrating the similarity among different methods as judged by the number of common predictions for all targets.

has the same rank R as a correct prediction, the first incorrect prediction is assigned rank R.

RESULTS

Participating methods: Brief description and similarity

In CASP10 26 groups submitted predictions of intra-molecular contacts, including 22 automated servers and four expert groups. Three groups used new methods, while others used modified techniques developed earlier and tested in previous rounds of CASP. Table I presents a short description of the participating publicly available contact prediction servers. A more detailed overview of all the methods participating in CASP10 can be found in the CASP10 Abstract Book.[56]

Not all methods are conceptually different as oftentimes they rely on similar prediction techniques using similar mathematical apparatus and predictive features.

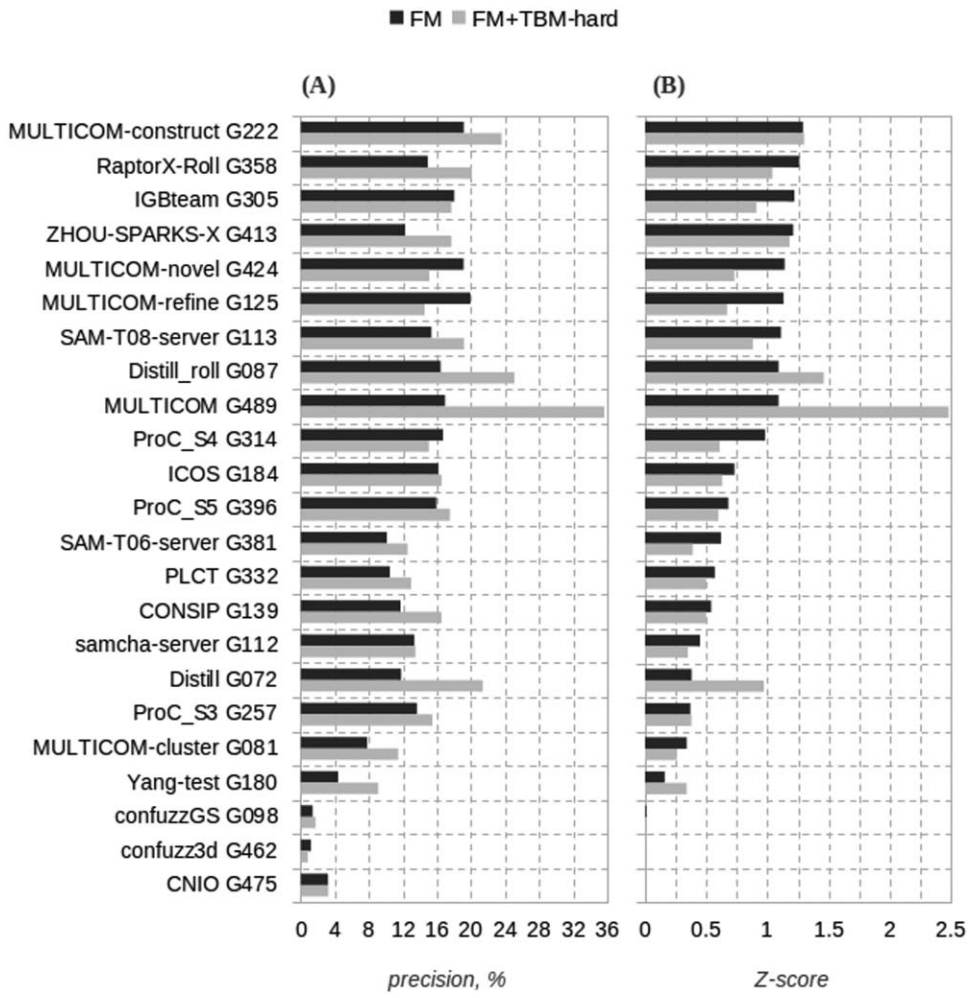

Figure 3

Precision (**A**) and cumulative *z*-score (**B**) for the participating groups on the two sets of the evaluated domains (FM and FM + TBM_hard). The data are shown for the top *L*/5 long-range contacts. Groups in both panels are ordered according to their cumulative *z*-score on FM targets.

To illustrate this, we clustered the methods participating in CASP10 based on the pair-wise Jaccard distance (see Materials). Figure 2 shows the results of the method clustering. As one can notice, four lowest level clusters encompass two prediction groups each from the same research centers, i.e. two Proc-S, Distill, Multicom, and confuzz methods. It is apparent that the clustered groups use similar methodologies with slight modifications in the implementation of the method.

Group performance on the reduced datasets: Precision and Xd

The results of the analysis of the group performance for long-range contacts in the *L*/5 contact lists are presented in Figure 3. For each group we show the values of precision and cumulative *z*-score (sum of precision-based and Xd-based *z*-scores) averaged over all predicted domains from the "FM" and "FM + TBM_hard" datasets (see Materials for a detailed description of the datasets and evaluation measures).

Panel A of Figure 3 demonstrates that the precision of the current prediction methods on FM targets does not exceed 20%. The three best performing groups on the FM targets (G125, G222, and G424) attain precision of 19% and belong to the same family of methods (Multicom, group leader J. Cheng, University of Missouri). Multicom-construct method (G222) was also shown to reach the highest score according to the Xd measure (see Fig. S2 in Supporting Information), and is ranked first according to the cumulative *z*-score (Fig. 3, panel B). It should be mentioned, though, that the difference in performance of this method and the others is marginal, as Student's *t*-tests did not reveal statistically significant difference in the performance of the top ten methods (see Table II for precision and Table S1 in Supporting Information for Xd). This statement is supported by the results of the "head-to-head"

Table II
Results of the Paired Student's *t*-Test on the Precision Score for (A) FM and (B) FM, TBM-Hard Domains for Top 10 Groups According to the Cumulative *z*-Score Ranking

A	G222	G358	G305	G413	G424	G125	G113	G087	G489	G314
G222	x	9	15	11	14	14	12	15	14	15
G358	0.49	x	9	6	8	8	9	9	8	9
G305	0.18	0.3	x	12	15	15	12	16	15	15
G413	0.11	0.19	0.19	x	11	11	8	12	11	11
G424	0.07	0.09	0.44	0.13	X	15	11	15	14	14
G125	0.16	0.1	0.39	0.13	0.26	x	11	15	14	14
G113	0.5	0.34	0.39	0.24	0.08	0.08	x	12	11	12
G087	0.26	0.46	0.4	0.46	0.36	0.32	0.44	x	15	15
G489	0.33	0.1	0.37	0.4	0.34	0.32	0.31	0.49	x	14
G314	0.19	0.21	0.41	0.09	0.21	0.34	0.37	0.48	0.48	x

B	G489	G087	G222	G413	G358	G072	G305	G113	G424	G125
G489	x	27	26	18	16	27	27	21	25	26
G087	0.03	x	28	20	18	29	29	23	27	28
G222	0.02	0.34	x	19	18	28	28	23	26	27
G413	0.05	0.34	0.08	x	10	20	20	14	18	19
G358	<0.01	0.03	0.26	0.12	x	18	18	18	17	17
G072	<0.01	0.08	0.36	0.43	0.11	x	29	23	27	28
G305	0.01	0.07	0.01	0.18	0.25	0.21	x	23	27	28
G113	<0.01	0.12	0.18	0.38	0.47	0.24	0.14	x	21	22
G424	0.01	0.03	<0.01	0.36	<0.01	0.11	0.17	0	x	27
G125	0.01	0.03	<0.01	0.33	<0.01	0.11	0.19	0	0.39	x

The tables show the *P* values (cells below the diagonal) of the Student's *t*-tests performed for each pair of the groups on the common set of domains (the numbers above the diagonal). Shaded cells indicate statistically indistinguishable results at the significance level of 0.05.

Table III
The "Head-to-Head" Comparison of the Performance of the Groups Based on the *precision* Score for (A) FM and (B) FM, TBM-Hard Domains for the Top 10 Groups According to the Cumulative *z*-Score Ranking

		Group 2									
A		G222	G358	G305	G413	G424	G125	G113	G087	G489	G314
Group 1	G222	X	44.4%	46.7%	63.6%	50.0%	50.0%	41.7%	66.7%	64.3%	46.7%
	G358	44.4%	x	44.4%	16.7%	75.0%	75.0%	33.3%	33.3%	37.5%	55.6%
	G305	33.3%	33.3%	x	58.3%	40.0%	40.0%	33.3%	50.0%	60.0%	40.0%
	G413	27.3%	66.7%	25.0%	x	36.4%	36.4%	25.0%	58.3%	45.5%	27.3%
	G424	21.4%	12.5%	53.3%	45.5%	x	13.3%	18.2%	46.7%	35.7%	28.6%
	G125	28.6%	12.5%	46.7%	45.5%	13.3%	x	18.2%	53.3%	35.7%	21.4%
	G113	33.3%	44.4%	50.0%	50.0%	63.6%	63.6%	x	66.7%	45.5%	50.0%
	G087	26.7%	55.6%	37.5%	41.7%	33.3%	26.7%	33.3%	x	26.7%	40.0%
	G489	28.6%	50.0%	40.0%	45.5%	50.0%	50.0%	45.5%	60.0%	x	35.7%
	G314	33.3%	33.3%	46.7%	54.5%	64.3%	57.1%	33.3%	60.0%	57.1%	x

		Group 2									
B		G489	G087	G222	G413	G358	G072	G305	G113	G424	G125
Group 1	G489	x	63.0%	53.8%	55.6%	75.0%	77.8%	66.7%	71.4%	72.0%	73.1%
	G087	25.9%	x	35.7%	45.0%	61.1%	55.2%	51.7%	43.5%	55.6%	50.0%
	G222	42.3%	60.7%	x	63.2%	55.6%	57.1%	60.7%	56.5%	69.2%	74.1%
	G413	38.9%	55.0%	31.6%	x	70.0%	50.0%	50.0%	35.7%	61.1%	63.2%
	G358	18.8%	27.8%	38.9%	20.0%	x	50.0%	50.0%	33.3%	76.5%	70.6%
	G072	14.8%	24.1%	35.7%	45.0%	44.4%	x	44.8%	39.1%	59.3%	57.1%
	G305	33.3%	31.0%	25.0%	40.0%	33.3%	41.4%	x	30.4%	51.9%	50.0%
	G113	23.8%	47.8%	26.1%	35.7%	44.4%	52.2%	47.8%	x	71.4%	72.7%
	G424	20.0%	29.6%	15.4%	27.8%	17.6%	33.3%	40.7%	14.3%	x	29.6%
	G125	19.2%	35.7%	14.8%	26.3%	17.6%	35.7%	39.3%	13.6%	14.8%	x

The rows show the fraction of common domains for which the precision score of the group in the row is higher than that of the group in the column. Cases of equal scores are not counted.

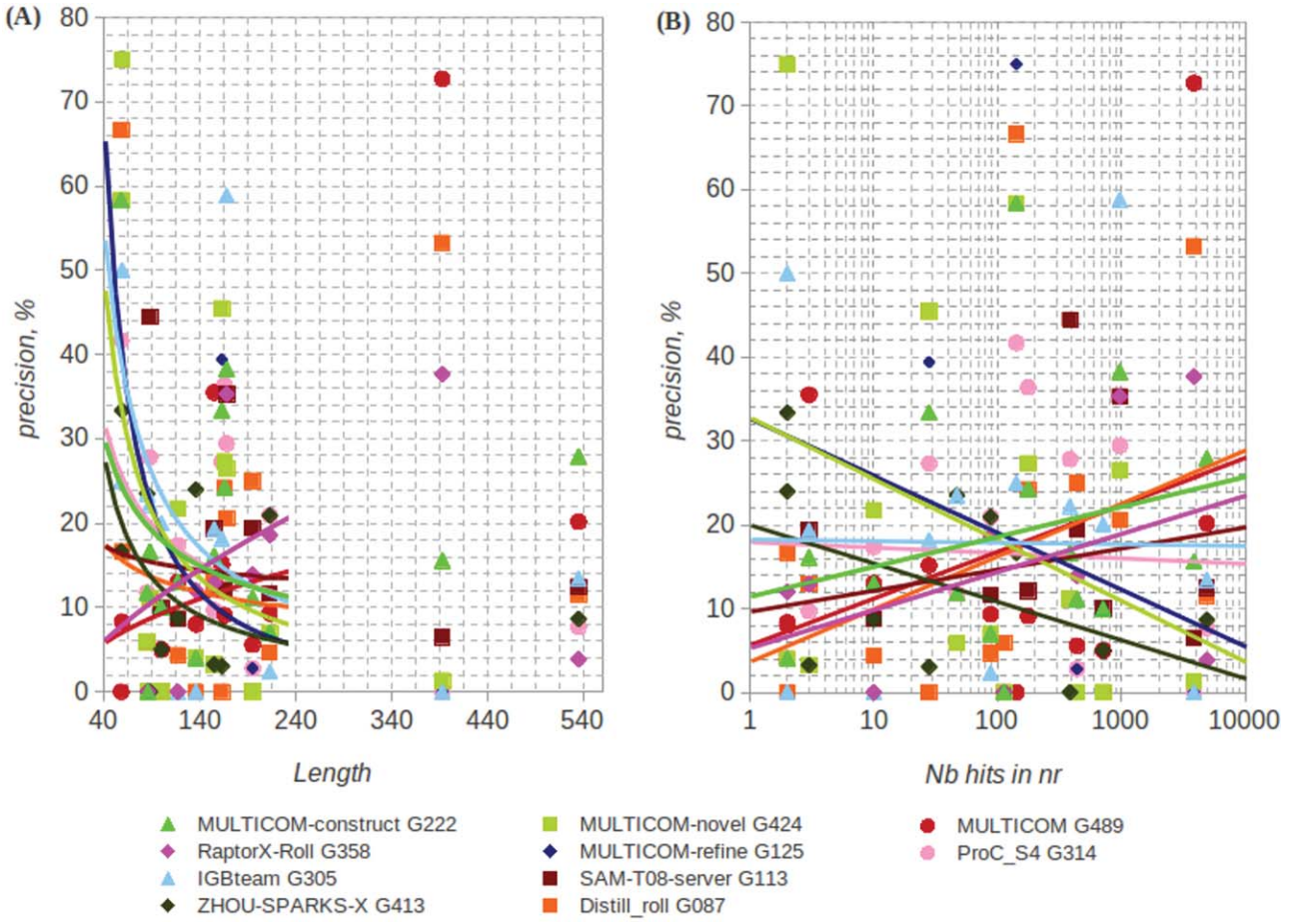

Figure 4

Precision of the prediction methods as a function of domain length (**A**) and depth of the alignment (**B**). The data are shown for the top *L*/5 long-range contacts.

comparison (Table III and Table S2 in Supporting Information) where no method was shown to consistently over-score any other method on more than half of the domains.

For the set of FM and TBM_hard domains, there is a group clearly outperforming the others, Multicom (G489), the results of which (Fig. 3) definitely look better than those of other groups (precision over 35% with the next best value of 24% for the Distill_roll group). The Multicom group is shown to be statistically better than all other predictors on the FM + TBM_hard set of targets (see Table II in the main text and Table S1 in Supporting Information) and consistently better than other methods in head-to-head comparisons (Table III and Table S2 in Supporting Information). However, it should be mentioned that the method used by group G489 is not conceptually an ab initio contact prediction method, as it relies on the three-dimensional models submitted by CASP10 servers. The better performance of this group on the FM + TBM_hard dataset can be explained by the method's consensus strategy, which works well on the TBM targets that constitute a substantial fraction of the FM + TBM_hard dataset.

Dependence of group performance on the domain length and the depth of alignment

Figure S1 (Supporting Information) shows that the contacts are harder to predict for some domains. The predictive difficulty of a domain is not always directly connected with the availability of templates, and from Figure S1 it can be seen that in CASP10 the third easiest target (T0739-D2) is in fact an FM domain, while the second hardest (T0668-D1) is a template-based target. This raises the question of which other features, besides template availability, may influence the accuracy of contact prediction. In particular, we investigated the influence of domain length and depth of alignment.

Figure 4(A) shows the precision of the best 10 performing groups as a function of domain length. The CASP10 FM dataset covers a wide range of domain

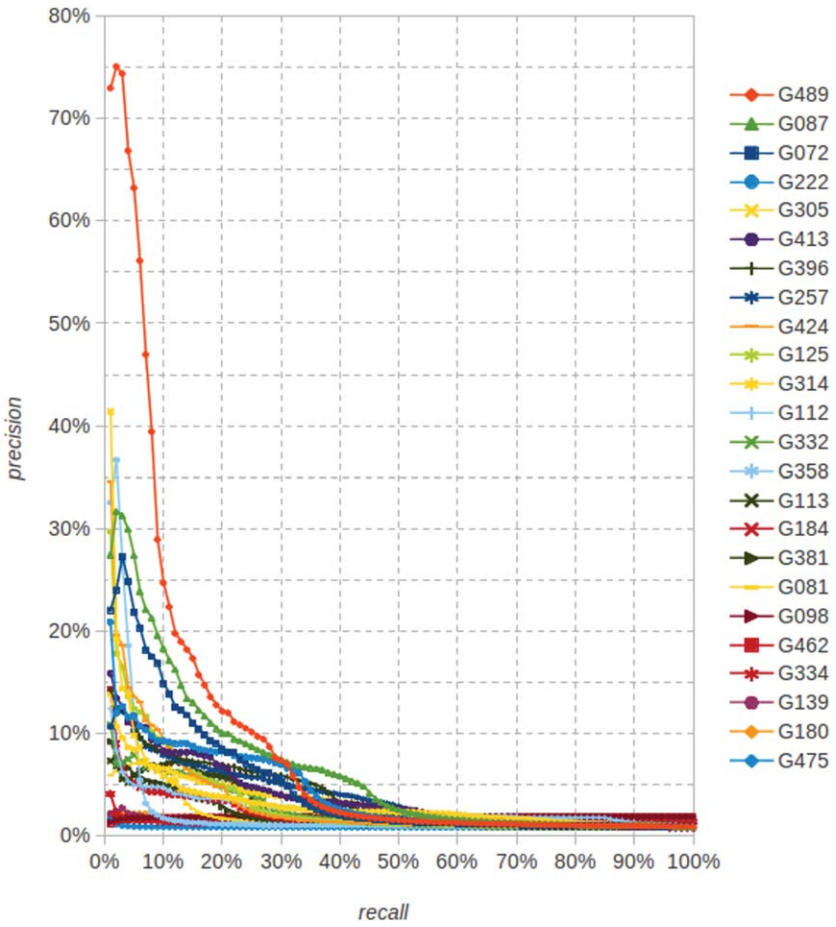

Figure 5

PR-curves for all predicted long-range contacts on FM domains.

length spanning from 58 to 535 residues. Two domains are short (under 60 residues), two rather long (over 390 residues) and the remaining 12 are of medium length (80–220 residues). On four of the domains (the shortest two and one from each of the medium and long sub-ranges), the best groups reach a very high precision (over 50%). It should be noticed, though, that the two longest domains in this graph (T0653-D1 and T0695-D1) represent non-globular targets with a repeated topology (see the description of Set 2R in Materials), and this may introduce bias in the analysis. Therefore, we analyzed per-group trends in the results excluding these two domains. Inspection of the graph reveals that the vast majority of groups reach better precision on shorter targets.

To analyze the dependence of group performance on the depth of the target alignments, we searched for sequence homologs for each target with PSI-BLAST[57] running five iterations against the non-redundant database with parameters "-h 0.05 -v 1000 -b 1000." The number of hits covering at least 75% of target's sequence was used as a measure of the alignment depth. The depth of the alignment for CASP10 FM targets varied from just a few hits (for T0726-D3, T0741-D1, T0740-D1) to more than a thousand for two repeat-topology domains (T0653-D1 and T0695-D1). Figure 4(B) shows that CASP10 methods are in general insensitive to the alignment depth, as no trend in the data can be detected. As precision of group performance depends on target length, we also tested a hypothesis that length can be a contributing factor in how precision depends on depth of alignment. Our additional analysis showed that this is not the case.

Group performance on the untrimmed contact lists: PR-curve and MCC analyses

Figure 5 and Table IV present a different perspective on the methods' performance based on the PR-curve analysis, *MCC* and other descriptive statistics measures (see Materials).

The PR-curve analysis clearly identifies the top performing group, G489 (Multicom), which reaches an *AUC_PR* score of 9.5%. Again, we remind here that this

Table IV
Descriptive Statistics Scores Calculated for the Predictions Treated in the Context of the Complete Contact Maps for Long-Range Contacts for FM Domains

Group	No dom	TP	FP	TN	FN	MCC	Precision (%)	Recall (%)	F_1	AUC_PR
G489	15	841	11,331	27,5798	1838	0.131	6.9	31.4	0.113	0.095
G087	16	1175	22,200	266,876	1510	0.127	5.0	43.8	0.090	0.065
G222	16	905	15,025	274,051	1780	0.120	5.7	33.7	0.097	0.043
G072	16	814	13,674	275,402	1871	0.112	5.6	30.3	0.095	0.049
G396	16	1006	21,195	267,881	1679	0.109	4.5	37.5	0.081	0.038
G257	16	1331	43,378	245,698	1354	0.092	3.0	49.6	0.056	0.038
G113	13	559	11,025	266,176	1785	0.091	4.8	23.8	0.080	0.025
G314	16	1255	42,987	246,089	1430	0.085	2.8	46.7	0.053	0.032
G125	16	597	12,454	276,622	2088	0.083	4.6	22.2	0.076	0.034
G413	12	841	34,071	152,543	742	0.082	2.4	53.1	0.046	0.038
G424	16	540	10,788	278,288	2145	0.081	4.8	20.1	0.077	0.035
G081	16	312	4287	284,789	2373	0.078	6.8	11.6	0.086	0.018
G112	13	1542	89,198	107,628	269	0.076	1.7	85.1	0.033	0.027
G184	16	570	15,934	273,142	2115	0.065	3.5	21.2	0.059	0.021
G139	10	1294	82,600	167,456	625	0.063	1.5	67.4	0.030	0.015
G332	16	738	25,302	263,774	1947	0.063	2.8	27.5	0.051	0.026
G381	13	448	13,232	263,969	1896	0.061	3.3	19.1	0.056	0.019
G305	16	1952	123,658	165,418	733	0.058	1.6	72.7	0.030	0.038
G358	16	154	1969	287,107	2531	0.057	7.3	5.7	0.064	0.026
G180	12	490	30,968	155,646	1093	0.035	1.6	31.0	0.030	0.012
G334	12	18	241	154,081	2166	0.019	6.9	0.8	0.014	0.015
G475	12	19	857	261,349	2169	0.009	2.2	0.9	0.012	0.009
G098	12	13	982	70,457	1345	−0.005	1.3	1.0	0.011	0.018
G462	12	11	981	70,458	1347	−0.007	1.1	0.8	0.009	0.018

The results are sorted according to the MCC score.

group does not predict contacts directly from the sequence but relies on the submitted three-dimensional models. The two other groups that stand out in the PR-curve analysis are G087 and G072, both from the Distill family of methods (group leader G. Pollastri, University College Dublin).

The results of the PR-analysis (AUC_PR scores) are shown to be well correlated with the MCC and F_1 scores

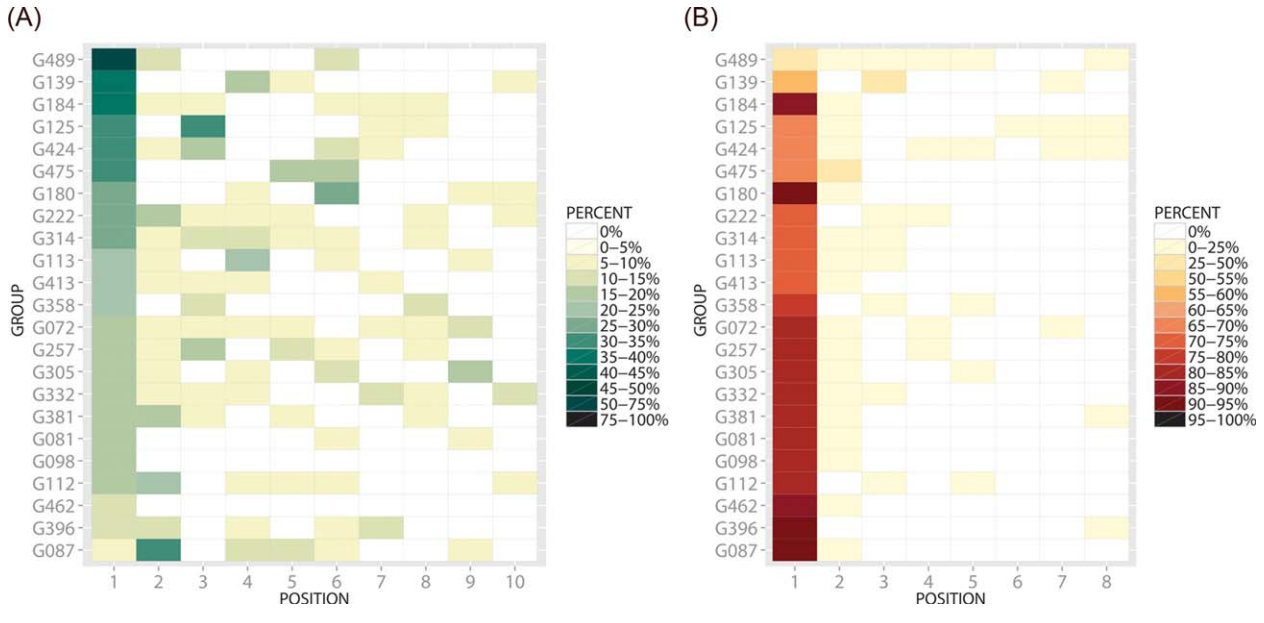

Figure 6
Percent of cases where the first correct (**A**) and first incorrect (**B**) prediction is in the reported position for each group. Rows are ordered according to the percentage in the first column of A. The data are shown for the top $L/5$ long-range contacts in FM domains.

Table V
Results of the Prediction of Long-Range Contacts in Which the Contacting Residues Belong to Two Different Domains

Group	FP	TP	Precision (%)
G489	265	18	6.4
G087	259	7	2.6
G072	261	7	2.6
G475	84	2	2.3
G112	246	5	2.0
G381	213	3	1.4
G334	74	1	1.3
G081	217	2	0.9
G332	261	2	0.8
G139	182	1	0.5
G180	217	1	0.5
G424	231	1	0.4
G077	35	0	0.0
G098	221	0	0.0
G113	231	0	0.0
G125	259	0	0.0
G184	232	0	0.0
G222	249	0	0.0
G257	305	0	0.0
G305	305	0	0.0
G314	305	0	0.0
G358	212	0	0.0
G396	305	0	0.0
G413	218	0	0.0
G462	215	0	0.0

The data are for the $L/5$ contacts with higher predicted probability.

presented in Table IV. The Pearson correlation coefficients for these two pairs of scores are 0.76 and 0.71, respectively. Also there is a high correlation (0.90) between the MCC and F_1 scores. At the same time, the correlation between other measures presented in Table IV is substantially lower (except for the F_1 – precision correlation) confirming that these (low-correlated) measures highlight different aspects of contact prediction.

Position of the first correct and incorrect contact

The prediction of contacts in protein structures can be used as input for computational methods aimed at structure prediction and, in this case, the correct ranking of the contacts in terms of their probability might not be necessarily relevant. On the other hand, prediction of specific contacts in a protein might shed light on its functional or structural properties and in this case, their correctness should be experimentally tested before drawing conclusions. This is usually done by designing appropriate mutations of the residues predicted to be in contact, expressing the mutated protein(s) and testing their function (see for example Refs. [58–61]). Clearly, one would like to perform as few experiments as possible. Since contact predictions are provided together with estimates of their reliability, it is reasonable to expect that the contacts would be tested in the order they appear in the list of predictions. This raises the question of how much down the ordered list of contacts is the first correct prediction for a given method.

We computed the position of the first correct prediction as well as the position of the first error for each target and each group considering short, medium, and long-range contacts. The results of this analysis are available from the CASP10 web site (http://predictioncenter.org/casp10/rr_additional.cgi). As in other sections, here we concentrate on the results for long-range contacts on FM targets.

Figure 6(A) shows, for each group, the percentage of times in which the first correct prediction is found in a given position; Figure 6(B) shows the percentage of times in which the first incorrect prediction is found in a given position. Group G489 that performs better than the other groups has a correct prediction in the first position on the $L/5$ contact lists 56% of the times and in 13% of the cases the first correct prediction is in position 2. Other groups also often have the first correct prediction ranking high in the list. It is instructive to compare the two parts of the figure. For example, group G184 has a correct prediction in one of the top positions about 40% of the time, but also often it has an incorrect prediction in the first positions. This is due to the fact that this group often assigns the same probability values to a set of contacts, some correct and some incorrect.

Interdomain contact predictions

The prediction of contacts between different domains can be extremely useful in cases where multidomain proteins are modeled using different templates for the different domains, since the step of packing together the partial models can, and often does, introduce errors.

We analyzed the number of cases in which different participating groups correctly predicted contacts between residues belonging to two different domains. The results for interdomain long-range contacts in FM targets are summarized in Table V, and the example for target T0658 is shown in Figure 7 Table V shows that in this analysis the best results are achieved by group G489, followed by groups G112 and G072

Also in this case, one can ask the question of how often the contacts predicted with the highest probabilities are correct. The results, shown in Figure S3 (Supporting Information) again highlight that group G489 is particularly effective in ranking the predicted contacts.

Comparison of CASP10 with previous experiments

Establishing progress in contact prediction is not a trivial task as targets, methods, and databases change in time. Unfortunately, no methods are available to adequately take all these relevant factors into account. We report here a

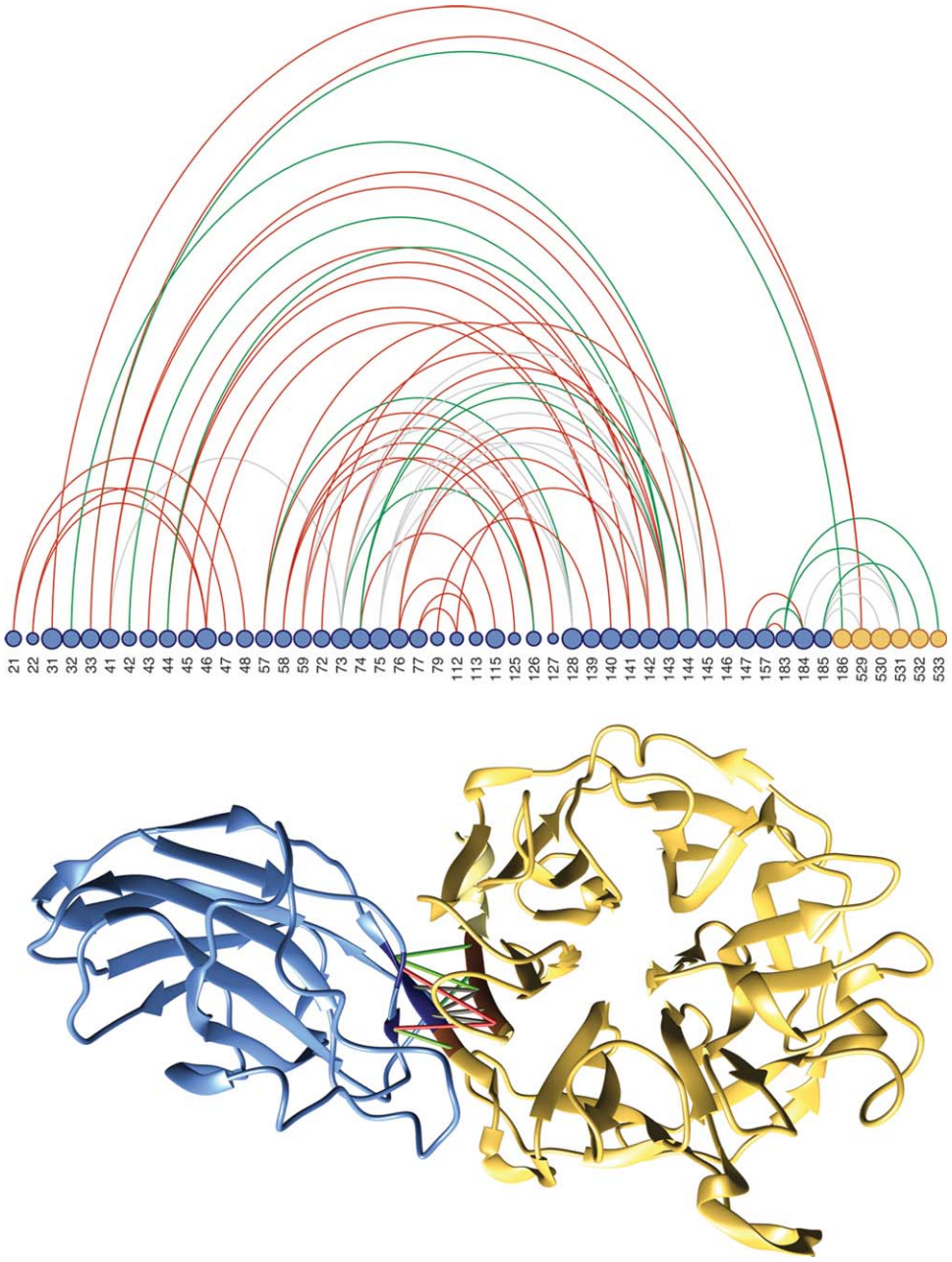

Figure 7

Example of the prediction of inter-domain contacts for target T0658. This is a two domain protein with the first domain (residues 20–185) being an FM target and the second (residues 186–540)—a template based target. The top panel shows *L*/5 contacts correctly predicted by at least one group as arcs connecting the corresponding residues indicated by circles. We show all the residues involved in correctly predicted contacts in the first (FM) domain, both intra- and inter-domain, and only the residues involved in correctly predicted inter-domain contacts for the second (TBM) domain. The size of the circle is proportional to the number of contacts the residue makes in the experimental structure. Blue and yellow circles are residues belonging to the first and second domain, respectively. The color of the connecting arcs indicates the frequency with which the corresponding contact was predicted by the groups. Red, green, and gray lines indicate contacts predicted with a frequency below the median, between the median and the third quartile and above the third quartile, respectively. The bottom figure shows the three-dimensional structure of the protein with the first domain in blue and the second in yellow. The correctly predicted contacts are indicated by sticks with the same color scheme as the corresponding arcs in the top panel.

comparison of the results without attempting to make any claim about the presence of real and measurable progress.

Figure 8 shows the results of the top 10 groups in the latest three CASPs on FM domains for the *L*/5 lists of long-range contacts (CASP10 results for the FM + TBM _hard domains are also included for comparison). On average, the CASP8 predictions (12 domains) have the highest precision—24.6%, followed by CASP9 (29

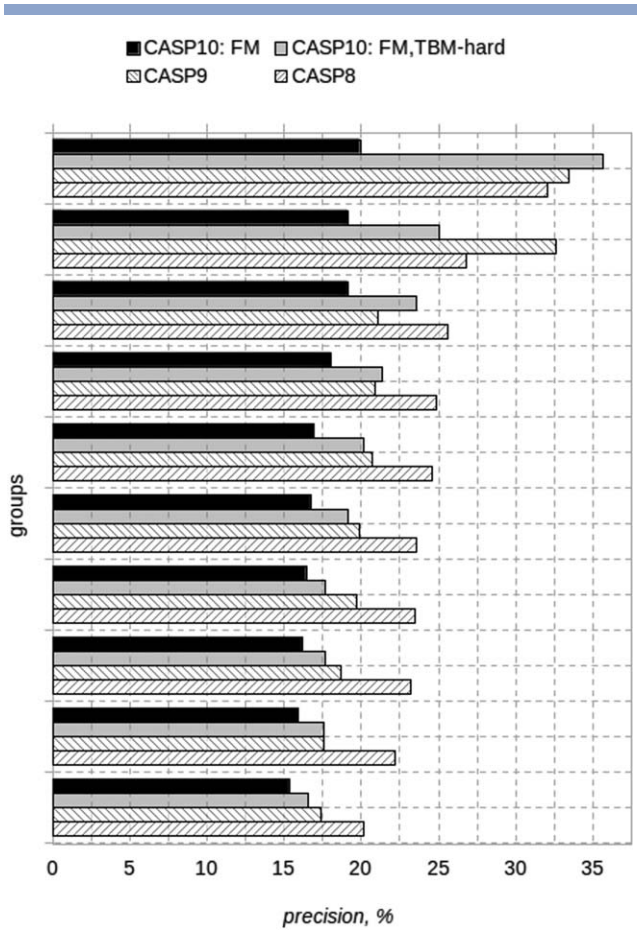

Figure 8

Precision of prediction for the top 10 groups in latest three CASPs.

domains)—21.4%, CASP10 (FM + TBM_hard, 28 domains)—21.4%, and CASP10: (FM, 16 domains)—17.4%. These results may indicate lack of substantial progress or, alternatively, be a consequence of the growing difficulty of targets in subsequent CASPs.[63]

CONCLUSIONS

The assessment of the state-of-the-art in contact prediction shows that the current precision of the best contact prediction methods on long-range contacts averages around 20%—the same limit observed in several previous CASPs. We look forward to seeing the results of the new methods that have recently appeared. Their published results in tests other than CASP have certainly stirred a lot of attention and it is therefore likely that we will see a renewed interest in the development of novel methods in contact prediction that will lead to improved results. We believe that progress in the field is objectively offset by the increased difficulty of the targets in CASP10 and that the depth of the alignments available for these targets made them less attractive for these new methods. At the same time, it should be mentioned that the list of CASP targets does mirror the proteins that the biological community considers interesting and worth an effort.

The predictions submitted by the best performing groups are statistically indistinguishable on the set of free-modeling domains. When hard template-based targets are added to the dataset, the results of the Multicom group, which uses consensus strategy to extract the contacts from predicted three-dimensional structures, are better than the others. Among the remaining groups, two implementations of the Distill method and ab initio predictors from the Multicom series of methods quite consistently perform better.

Based on the CASP10 data, we show that shorter domains are in general easier targets for contact prediction, and that the difficulty of predicting contacts in domains is not correlated with the depth of target sequence alignment.

ACKNOWLEDGMENTS

The authors thank John Moult and David Jones for useful suggestions.

REFERENCES

1. Havel TF, Crippen GM, Kuntz ID. Effects of distance constraints on macromolecular conformation. 2. Simulation of experimental results and theoretical predictions. Biopolymers 1979;18:73–81.
2. Brunger AT, Clore GM, Gronenborn AM, Karplus M. Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. Proc Natl Acad Sci USA 1986;83:3801–3805.
3. Clore GM, Nilges M, Brunger AT, Karplus M, Gronenborn AM. A comparison of the restrained molecular dynamics and distance geometry methods for determining three-dimensional structures of proteins on the basis of interproton distances. FEBS Lett 1987;213:269–277.
4. Bohr J, Bohr H, Brunak S, Cotterill RM, Fredholm H, Lautrup B, Petersen SB. Protein structures from distance inequalities. J Mol Biol 1993;231:861–869.
5. Saitoh S, Nakai T, Nishikawa K. A geometrical constraint approach for reproducing the native backbone conformation of a protein. Proteins 1993;15:191–204.
6. Taylor WR. Protein fold refinement: building models from idealized folds using motif constraints and multiple sequence data. Protein Eng 1993;6:593–604.
7. Skolnick J, Kolinski A, Ortiz AR. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–241.
8. Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. Fold Des 1997;2:295–306.
9. Vassura M, Margara L, Di Lena P, Medri F, Fariselli P, Casadio R. FT-COMAR: fault tolerant three-dimensional structure reconstruction from protein contact maps. Bioinformatics 2008;24:1313–1315.
10. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci USA 2006;103:5361–5366.

11. Wu S, Szilagyi A, Zhang Y. Improving protein structure prediction using multiple sequence-based contact predictions. Structure 2011; 19:1182–1191.
12. Shindyalov IN, Kolchanov NA, Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 1994;7:349–358.
13. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
14. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 1997;2:S25–S32.
15. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat Rev Genet 2013;14:249–261.
16. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol 2012;30:1072–1080.
17. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2012;28: 184–190.
18. Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. Proc Natl Acad Sci USA 2012;109: 10340–10345.
19. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA 2011;108:E1293–E1301.
20. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D structure computed from evolutionary sequence variation. PLoS One 2011;6:e28766.
21. Burger L, van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol 2010;6:e1000633.
22. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci USA 2009;106:67–72.
23. Fariselli P, Casadio R. A neural network based predictor of residue contacts in proteins. Protein Eng 1999;12:15–21.
24. Vullo A, Walsh I, Pollastri G. A two-stage approach for improved prediction of residue contact maps. BMC Bioinformatics 2006;7: 180.
25. Chen P, Huang DS, Zhao XM, Li X. Predicting contact map using radial basis function neural network with conformational energy function. Int J Bioinform Res Appl 2008;4:123–136.
26. Shackelford G, Karplus K. Contact prediction using mutual information and neural nets. Proteins 2007;69:159–164.
27. Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. Bioinformatics 2012;28:2449–2457.
28. Xue B, Faraggi E, Zhou Y. Predicting residue-residue contact maps by a two-layer, integrated neural-network method. Proteins 2009;76: 176–183.
29. Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng 2001;14:835–843.
30. Wu S, Zhang Y. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics 2008;24:924–931.
31. Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics 2007;8:113.
32. Chen P, Han K, Li X, Huang DS. Predicting key long-range interaction sites by B-factors. Protein Pept Lett 2008;15:478–483.
33. Bjorkholm P, Daniluk P, Kryshtafovych A, Fidelis K, Andersson R, Hvidsten TR. Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue-residue contacts. Bioinformatics 2009;25:1264–1270.
34. Pollastri G, Baldi P. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. Bioinformatics 2002;18:S62–S70.
35. Karplus K. SAM-T08, HMM-based protein structure prediction. Nucleic Acids Res 2009;37(Web Server issue):W492–W497.
36. Wang Z, Eickholt J, Cheng J. MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. Bioinformatics 2010;26:882–888.
37. Li Y, Fang Y, Fang J. Predicting residue-residue contacts using random forest models. Bioinformatics 2011;27:3379–3384.
38. Stout M, Bacardit J, Dirst JD, Smith RE, Krasnogor N. Prediction of topological contacts in proteins using learning classifier systems. Soft Comput 2009;13:245–258.
39. Lesk AM. CASP2: report on ab initio predictions. Proteins 1997; Suppl 1:151–166.
40. Grana O, Baker D, MacCallum RM, Meiler J, Punta M, Rost B, Tress ML, Valencia A. CASP6 assessment of contact prediction. Proteins 2005;61:214–224.
41. Izarzugaza JM, Grana O, Tress ML, Valencia A, Clarke ND. Assessment of intramolecular contact predictions for CASP7. Proteins 2007;69:152–158.
42. Ezkurdia I, Grana O, Izarzugaza JM, Tress ML. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. Proteins 2009;77:196–209.
43. Monastyrskyy B, Fidelis K, Tramontano A, Kryshtafovych A. Evaluation of residue-residue contact predictions in CASP9. Proteins 2011; 79:119–125.
44. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP Prediction Center infrastructure and evaluation measures in CASP10 and CASP ROLL Proteins, 2013; this issue.
45. Taylor T, Tai C-H, Huang YJ, Block J, Bai H, Kryshtafovych A, Montelione GT, Lee BK. Definition and classification of assessment units for CASP10. Proteins 2014;82(2):14–25.
46. Enkhbayar P, Kamiya M, Osaki M, Matsumoto T, Matsushima N. Structural principles of leucine-rich repeat (LRR) proteins. Proteins 2004;54:394–403.
47. Monastyrskyy B, Fidelis K, Moult J, Tramontano A, Kryshtafovych A. Evaluation of disorder predictions in CASP9. Proteins 2011;79: 107–118.
48. Bunescu R, Ge, R., Kate, R., Marcotte, E., Mooney, R., Ramani, A., Wong, Y. Comparative experiments on learning information extractors for proteins and their interactions. J Artif Intell Med 2004:139–155.
49. Kok S, Domingos, P. Learning the structure of Markov Logic Networks. Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany. 2005. ACM Press, New York, NY, USA.
50. He HB, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21:1263–1284.
51. Goadrich M, Oliphant L, Shavlik J. Learning ensembles of first-order clauses for recall-precision curves: a case study in biomedical information extraction. Proceedings of the 14th International Conference on Inductive Logic Programming, Porto, Portugal, pp. 98–115. Springer-Verlag, Berlin, Heidelberg, 2004.
52. Fawcett T, Flach PA. A response to Webb and Ting's on the application of ROC analysis to predict classification performance under varying class distributions. Mach Learn 2005;58:33–38.
53. Davis J, Goadrich, M. The relationship between precision-recall and ROC curves. Proceedings of the 23rd international conference on Machine learning. Pittsburgh, PA, 2006. ACM Press, New York, NY USA.
54. http:/www.cs.wisc.edu/~richm/programs/AUC/.
55. Levandowsky M, Winter D. Distance between sets. Nature 1971; 234:34–35.
56. http://predictioncenter.org/casp10/doc/CASP10_Abstracts.pdf.
57. 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.
58. Bauer B, Mirey G, Vetter IR, Garcia-Ranea JA, Valencia A, Wittinghofer A, Camonis JH, Cool RH. Effector recognition by the small GTP-binding proteins Ras and Ral. J Biol Chem 1999;274: 17763–17770.

59. Domanski TL, Halpert JR. Analysis of mammalian cytochrome P450 structure and function by site-directed mutagenesis. Curr Drug Metab 2001;2:117–137.
60. Thompson D, Lazennec C, Plateau P, Simonson T. Probing electrostatic interactions and ligand binding in aspartyl-tRNA synthetase through site-directed mutagenesis and computer simulations. Proteins 2008;71:1450–1460.
61. Fremgen SA, Burke NS, Hartzell PL. Effects of site-directed mutagenesis of mglA on motility and swarming of Myxococcus xanthus. BMC Microbiol 2010;10:295.
62. Kryshtafovych A, Moult J, Bales P, Bazan JF, Burgin A, Chen C, Cochran FV, Craig TK, Das R, Fass D, Garcia-Doval C, Herzberg O, Lorimer D, Luecke H, Ma X, Nelson D, van Raaij MJ, Rohwer F, Segall A, Seguritan V, Zeth K, Schwede T. Challenging the state-of-the-art in protein structure prediction: Highlights of experimental target structures for the 10th Critical Assessment of Techniques for Protein Structure Prediction Experiment CASP10. Proteins 2014;82(Suppl 2):26–42.
63. Kryshtafovych A, Fidelis K., Moult J. CASP10 results compared to those of previous CASP experiments, Proteins 2014;82(Suppl 2): 164–174.