

Evaluation of residue-residue contact prediction in CASP10.

Bohdan Monastyrskyy, Daniel d'Andrea, Krzysztof Fidelis, Anna Tramontano, Andriy Kryshtafovych

▶ To cite this version:

Bohdan Monastyrskyy, Daniel d'Andrea, Krzysztof Fidelis, Anna Tramontano, Andriy Kryshtafovych. Evaluation of residue-residue contact prediction in CASP10.. Proteins - Structure, Function and Bioinformatics, 2014, 82 Suppl 2, pp.138-53. 10.1002/prot.24340 . pasteur-01202597

HAL Id: pasteur-01202597 https://riip.hal.science/pasteur-01202597

Submitted on 21 Sep 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License



NIH Public Access

Author Manuscript

Proteins. Author manuscript; available in PMC 2015 February 01

Published in final edited form as:

Proteins. 2014 February ; 82(0 2): 138–153. doi:10.1002/prot.24340.

Evaluation of residue-residue contact prediction in CASP10

Bohdan Monastyrskyy^{1,§}, Daniel D'Andrea^{2,§}, Krzysztof Fidelis¹, Anna Tramontano^{2,3}, and Andriy Kryshtafovych^{1,*}

¹Genome Center, University of California, Davis, 415 Health Sciences Dr., Davis, CA 95616, USA

²Department of Physics, Sapienza – University of Rome, 5 P. le Aldo Moro, 00185 Rome, Italy

³Istituto Pasteur - Fondazione Cenci Bolognetti – University of Rome, 5 P. le Aldo Moro, 00185 Rome, Italy

Abstract

We present the results of the assessment of the intra-molecular 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.

Keywords

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 intra-molecular contacts in proteins can serve as an intermediate step towards 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

^{*}To whom the correspondence should be addressed: Andriy Kryshtafovych, Genome Center, University of California, Davis, 415 Health Sciences Dr., Davis, CA, 95616, USA, akryshtafovych@ucdavis.edu, Tel/Fax: (+1) 530-754-8977. [§]Authors contributed equally

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 contacts⁴⁻⁸. A few features related to the tolerance of these methods to data uncertainty and incompleteness were discovered. In particular, in a pioneering work¹, Havel and co-workers 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 quality of the reconstructed structures is the long-range geometrical constraint. Skolnick et al suggested⁷ that knowing contacts for one in every seven residues would be sufficient to recover the structure of short proteins. Later, Vassura et al⁹ claimed that knowing 1 in 4 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 3D 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 paper [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 intra-molecular 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 papers 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 literature¹⁵. Quite recently, the 20-year-old idea received a new twist as several papers 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 algorithms³⁶, random forest models³⁷, and learning classifier systems³⁸. 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 CASP2³⁹ (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 results⁴⁴. 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 intra-molecular 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 6–11, 12–23 and at least 24 residues, respectively. The contacts with a separation of less than 6 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, i.e. the domains classified in the FM, TBM/FM or TBM_hard categories⁴⁵. 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 units⁴⁵.

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 repeats⁴⁶, while the last one is a three-helical spectrin bundle repeated 5 times [EXPERIMENTAL TARGET HIGHLIGHT PAPER – THIS ISSUE]. 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 Fig. S1 of Supplementary Material.

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" sub-section 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.

Pre-processing of predictions—For multi-domain 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 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.

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

Page 5

"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 paper 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 residueresidue 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 papers^{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 paper, while the *Xd*-based results are shown in the Supplementary material.

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" comparisons⁴³.

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:

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

They were already used in previous CASPs, but were shown to be equivalent on the reduced prediction sets⁴¹. 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 paper⁴⁷, 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 any classes. Therefore, in this paper 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 cutoff 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 analysis⁵², but differs in that the parametric curves are plotted in the (*recall, precision*) coordinates. Davis and Goadrich⁵³ 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 Goadrich⁵³ and freely available from their website⁵⁴.

The Jaccard distance for clustering methods

The dissimilarity between two groups for each target is defined in terms of the Jaccard distance⁵⁵:

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

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 regardless of whether there are correct predictions 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 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 4 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 Book⁵⁶ and in the papers listed in Table I.

Not all methods are conceptually different as oftentimes they rely on similar prediction techniques using similar mathematical apparatus and predictive features. 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 Figure S2 in Supplementary Material), and is ranked first according to the cumulative z-score (Figure 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 Supplementary Material for *Xd*). This statement is supported by the results of the "head-to-head" comparison (Table III and Table S2 in Supplementary Material) 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 (Figure 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 Supplementary Materials) and consistently better than other methods in head-to-head comparisons (Table III and Table S2 in Supplementary Materials). 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 3D 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 (Supplementary Materials) 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 4A shows the *precision* of the best ten performing groups as a function of domain length. The CASP10 FM dataset covers a wide range of domain length spanning from 58 to 535 residues. Two domains are short (under 60 residues), two rather long (over 390 residues) and the remaining twelve 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-BLAST⁵⁷ running 5 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 4B 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 group does not predict contacts directly from the sequence but relies on the submitted 3D 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 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 references^{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 6A shows, for each group, the percentage of times in which the first correct prediction is found in a given position; figure 6B shows the rank of the first incorrectly predicted contact. 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.

Inter-domain contact predictions

The prediction of contacts between different domains can be extremely useful in cases where multi-domain 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 inter-domain 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 (Supplementary Material) 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 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 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 [see PROGRESS PAPER – THIS ISSUE].

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 3D 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank John Moult and David Jones for useful suggestions. This work was partially supported by the US National Institute of General Medical Sciences (NIGMS/NIH) – grant R01GM100482 to KF, by KAUST Award KUK-I1-012-43 to AT and by EMBO.

Abbreviations

FM	free modeling
TBM	template-based modeling
RR	residue-residue (contacts)
MCC	the Matthews correlation coefficient

LITERATURE

- Havel TF, Crippen GM, Kuntz ID. Effects of Distance Constraints on Macromolecular Conformation .2. Simulation of Experimental Results and Theoretical Predictions Biopolymers. 1979; 18(1):73–81.
- 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 U S A. 1986; 83(11):3801–3805. [PubMed: 3459158]
- 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(2):269–277. [PubMed: 3030815]
- 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(3):861–869. [PubMed: 7685827]
- Saitoh S, Nakai T, Nishikawa K. A geometrical constraint approach for reproducing the native backbone conformation of a protein. Proteins. 1993; 15(2):191–204. [PubMed: 8441754]
- 6. Taylor WR. Protein fold refinement: building models from idealized folds using motif constraints and multiple sequence data. Protein Eng. 1993; 6(6):593–604. [PubMed: 8234230]
- 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(2):217–241. [PubMed: 9020984]
- Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. Fold Des. 1997; 2(5):295–306. [PubMed: 9377713]
- 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(10): 1313–1315. [PubMed: 18381401]
- 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 U S A. 2006; 103(14):5361–5366. [PubMed: 16567638]
- Wu S, Szilagyi A, Zhang Y. Improving protein structure prediction using multiple sequence-based contact predictions. Structure. 2011; 19(8):1182–1191. [PubMed: 21827953]

- Shindyalov IN, Kolchanov NA, Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng. 1994; 7(3):349–358. [PubMed: 8177884]
- Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins. 1994; 18(4):309–317. [PubMed: 8208723]
- 14. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des. 1997; 2(3):S25–32. [PubMed: 9218963]
- de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat Rev Genet. 2013; 14(4):249–261. [PubMed: 23458856]
- Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol. 2012; 30(11):1072–1080. [PubMed: 23138306]
- 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(2):184–190. [PubMed: 22101153]
- Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. Proc Natl Acad Sci U S A. 2012; 109(26):10340–10345. [PubMed: 22691493]
- 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 U S A. 2011; 108(49):E1293–1301. [PubMed: 22106262]
- 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(12):e28766. [PubMed: 22163331]
- Burger L, van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol. 2010; 6(1):e1000633. [PubMed: 20052271]
- 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 U S A. 2009; 106(1):67–72. [PubMed: 19116270]
- Fariselli P, Casadio R. A neural network based predictor of residue contacts in proteins. Protein Eng. 1999; 12(1):15–21. [PubMed: 10065706]
- Vullo A, Walsh I, Pollastri G. A two-stage approach for improved prediction of residue contact maps. BMC Bioinformatics. 2006; 7:180. [PubMed: 16573808]
- 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(2):123–136. [PubMed: 18490258]
- Shackelford G, Karplus K. Contact prediction using mutual information and neural nets. Proteins. 2007; 69 (Suppl 8):159–164. [PubMed: 17932918]
- 27. Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. Bioinformatics. 2012; 28(19):2449–2457. [PubMed: 22847931]
- Xue B, Faraggi E, Zhou Y. Predicting residue-residue contact maps by a two-layer, integrated neural-network method. Proteins. 2009; 76(1):176–183. [PubMed: 19137600]
- 29. Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng. 2001; 14(11):835–843. [PubMed: 11742102]
- Wu S, Zhang Y. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics. 2008; 24(7):924–931. [PubMed: 18296462]
- Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics. 2007; 8:113. [PubMed: 17407573]
- 32. Chen P, Han K, Li X, Huang DS. Predicting key long-range interaction sites by B-factors. Protein Pept Lett. 2008; 15(5):478–483. [PubMed: 18537737]
- Bjorkholm P, Daniluk P, Kryshtafovych A, Fidelis K, Andersson R, Hvidsten TR. Using multidata hidden Markov models trained on local neighborhoods of protein structure to predict residueresidue contacts. Bioinformatics. 2009; 25(10):1264–1270. [PubMed: 19289446]

- 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 (Suppl 1):S62–70. [PubMed: 12169532]
- 35. Karplus K. SAM-T08, HMM-based protein structure prediction. Nucleic Acids Res. 2009; 37(Web Server issue):W492–497. [PubMed: 19483096]
- Wang Z, Eickholt J, Cheng J. MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. Bioinformatics. 2010; 26(7):882–888. [PubMed: 20150411]
- 37. Li Y, Fang Y, Fang J. Predicting residue-residue contacts using random forest models. Bioinformatics. 2011; 27(24):3379–3384. [PubMed: 22016406]
- Stout M, Bacardit J, Dirst JD, Smith RE, Krasnogor N. Prediction of topological contacts in proteins using learning classifier systems. Soft Computing. 2009; 13:245–258.
- Lesk AM. CASP2: report on ab initio predictions. Proteins. 1997; (Suppl 1):151–166. [PubMed: 9485507]
- 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 (Suppl 7):214–224. [PubMed: 16187364]
- Izarzugaza JM, Grana O, Tress ML, Valencia A, Clarke ND. Assessment of intramolecular contact predictions for CASP7. Proteins. 2007; 69 (Suppl 8):152–158. [PubMed: 17671976]
- Ezkurdia I, Grana O, Izarzugaza JM, Tress ML. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. Proteins. 2009; 77 (Suppl 9):196–209. [PubMed: 19714769]
- 43. Monastyrskyy B, Fidelis K, Tramontano A, Kryshtafovych A. Evaluation of residue-residue contact predictions in CASP9. Proteins. 2011; 79 (Suppl 10):119–125. [PubMed: 21928322]
- 44. Kryshtafovych A, Monastyrskyy B, Fidelis K. Prediction Center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins. 2013 (THIS ISSUE).
- 45. Taylor T, Tai C-H, Bai H, Kryshtafovych A, Montelione G, Lee B. Definition and classification of evaluation units for CASP10. Proteins. 2013 THIS ISSUE.
- 46. Enkhbayar P, Kamiya M, Osaki M, Matsumoto T, Matsushima N. Structural principles of leucinerich repeat (LRR) proteins. Proteins. 2004; 54(3):394–403. [PubMed: 14747988]
- Monastyrskyy B, Fidelis K, Moult J, Tramontano A, Kryshtafovych A. Evaluation of disorder predictions in CASP9. Proteins. 2011; 79 (Suppl 10):107–118. [PubMed: 21928402]
- 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. Journal of Artificial Intelligence in Medicine. 2004:139–155.
- 49. Kok, S.; Domingos, P. Learning the structure of Markov Logic Networks. ACm Press; 2005. p. 441-448.
- He HB, Garcia EA. Learning from Imbalanced Data. Ieee T Knowl Data En. 2009; 21(9):1263– 1284.
- Goadrich, M.; Oliphant, L.; Shavlik, J. Learning ensembles of first-order clauses for recallprecision curves: A case study in biomedical information extraction. 2004. p. 421-456.
- 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(1):33– 38.
- 53. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC Curves. 2006. p. 233-240.
- 54. http://www.cs.wisc.edu/~richm/programs/AUC/.
- 55. Levandowsky M, Winter D. Distance between Sets. Nature. 1971; 234(5323):34.
- 56. http://predictioncenter.org/casp10/doc/CASP10_Abstracts.pdf.
- 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(17):3389–3402. [PubMed: 9254694]

- 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(25):17763–17770. [PubMed: 10364219]
- 59. Domanski TL, Halpert JR. Analysis of mammalian cytochrome P450 structure and function by site-directed mutagenesis. Curr Drug Metab. 2001; 2(2):117–137. [PubMed: 11469721]
- 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(3):1450–1460. [PubMed: 18076053]
- 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. [PubMed: 21083931]



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.



Figure 2.

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



Figure 3.

Precision (panel A) and cumulative z-score (panel 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.



Figure 4.

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



Figure 5. PR-curves for all predicted long-range contacts on FM domains.



Figure 6.

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



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 3rd quartile and above the 3rd 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.





Table I

The publicly available contact prediction servers participating in CASP10.

Server name and URL address	CASP10 group	Brief description of the method
CMAPpro* http:/scratch.proteomics.ics.uci.edu/	G305	Deep neural networks architecture allowing progressive refinement of contact prediction.
Distill, Distill-roll http://distill.ucd.ie/distill/	G072, G087	2D-recursive neutral networks.
ICOS http:/icos.cs.nott.ac.uk/servers/psp.html	G184	In-house machine-learning technique taking into account 9-residue window profiles, secondary structure, and other features.
MULTICOM-CLUSTER http://casp.rnet.missouri.edu/svmcon.html	G081	An SVM tool. The input data include secondary structure, solvent accessibility, and sequence profile.
MULTICOM-CONSTRUCT* http://iris.rnet.missouri.edu/dncon/	G222	Ensembles of deep networks.
MULTICOM-NOVEL,MULTICOM-REFINE http://casp.rnet.missouri.edu/nncon.html	G424 G125	Recursive neural networks. MULTICOM- REFINE has a separate module to predict contacts in beta-sheets.
PROC_S3 http://www.abl.ku.edu/proc/proc_s3.html	G257	Random Forest models incorporating more than 1000 sequence-related features.
SAM-T06, SAM-T08 http:/compbio.soe.ucsc.edu/SAM06/ http:/compbio.soe.ucsc.edu/SAM08/	G381 G113	Recursive neural networks using the correlated mutations in MSA.
Samcha-server [*] http://binfolab12.kaist.ac.kr/conti/	G112	SVM incorporating more than 800 sequential features.

new methods according to the CASP10 Abstract Book

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 zscore ranking.

Monastyrskyy et al.

																						e common set of domains (the numbers above the diagonal). Shaded cells indicate
G314	15	6	15	11	14	14	12	15	14	x	G125	26	28	27	19	17	28	28	22	27	х	each pair of the groups on t
G489	14	8	15	11	14	14	11	15	x	0.48	G424	25	27	26	18	17	27	27	21	х	0.39	ormed for
G087	15	6	16	12	15	15	12	x	0.49	0.48	G113	21	23	23	14	18	23	23	х	0	0	ests perfc
G113	12	6	12	8	11	11	x	0.44	0.31	0.37	G305	27	29	28	20	18	29	х	0.14	0.17	0.19	tudent t-ti 0.05
G125	14	×	15	11	15	х	0.08	0.32	0.32	0.34	G072	27	29	28	20	18	x	0.21	0.24	0.11	0.11) of the S
G424	14	8	15	11	х	0.26	0.08	0.36	0.34	0.21	G358	16	18	18	10	×	0.11	0.25	0.47	< 0.01	< 0.01	diagonal
G413	11	9	12	х	0.13	0.13	0.24	0.46	0.4	0.09	G413	18	20	19	х	0.12	0.43	0.18	0.38	0.36	0.33	elow the
G305	15	6	х	0.19	0.44	0.39	0.39	0.4	0.37	0.41	G222	26	28	x	0.08	0.26	0.36	0.01	0.18	< 0.01	< 0.01	ss (cells b a rasults
G358	6	x	0.3	0.19	0.09	0.1	0.34	0.46	0.1	0.21	G087	27	х	0.34	0.34	0.03	0.08	0.07	0.12	0.03	0.03	le p-value
G222	×	0.49	0.18	0.11	0.07	0.16	0.5	0.26	0.33	0.19	G489	х	0.03	0.02	0.05	< 0.01	< 0.01	0.01	< 0.01	0.01	0.01	s show th
A	G222	G358	G305	G413	G424	G125	G113	G087	G489	G314	В	G489	G087	G222	G413	G358	G072	G305	G113	G424	G125	The table.

NIH-PA Author Manuscript

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.

A						Gro	up 2				
		G222	G358	G305	G413	G424	G125	G113	G087	G489	G314
Group 1	G222	Х	44.4%	46.7%	63.6%	50.0%	50.0%	41.7%	66.7%	64.3%	46.7%
	G358	44.4%	х	44.4%	16.7%	75.0%	75.0%	33.3%	33.3%	37.5%	55.6%
	G305	33.3%	33.3%	х	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%	х	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%	х	35.7%
	G314	33.3%	33.3%	46.7%	54.5%	64.3%	57.1%	33.3%	60.0%	57.1%	х
в						Gro	up 2				
		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%	х	35.7%	45.0%	61.1%	55.2%	51.7%	43.5%	55.6%	50.0%
	G222	42.3%	60.7%	х	63.2%	55.6%	57.1%	60.7%	56.5%	69.2%	74.1%
	G413	38.9%	55.0%	31.6%	х	70.0%	50.0%	50.0%	35.7%	61.1%	63.2%
	G358	18.8%	27.8%	38.9%	20.0%	х	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%	х	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

Proteins. Author manuscript; available in PMC 2015 February 01.

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.

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. The results are sorted according to the MCC score.

Monastyrskyy et al.

Group	No dom	ΤP	FP	ΠN	FN	MCC	precision, %	recall, %	F_1	AUC_PR
G489	15	841	11331	275798	1838	0.131	6.9	31.4	0.113	0.095
G087	16	1175	22200	266876	1510	0.127	5.0	43.8	060.0	0.065
G222	16	905	15025	274051	1780	0.120	5.7	33.7	0.097	0.043
G072	16	814	13674	275402	1871	0.112	5.6	30.3	0.095	0.049
G396	16	1006	21195	267881	1679	0.109	4.5	37.5	0.081	0.038
G257	16	1331	43378	245698	1354	0.092	3.0	49.6	0.056	0.038
G113	13	559	11025	266176	1785	0.091	4.8	23.8	0.080	0.025
G314	16	1255	42987	246089	1430	0.085	2.8	46.7	0.053	0.032
G125	16	597	12454	276622	2088	0.083	4.6	22.2	0.076	0.034
G413	12	841	34071	152543	742	0.082	2.4	53.1	0.046	0.038
G424	16	540	10788	278288	2145	0.081	4.8	20.1	0.077	0.035
G081	16	312	4287	284789	2373	0.078	6.8	11.6	0.086	0.018
G112	13	1542	89198	107628	269	0.076	1.7	85.1	0.033	0.027
G184	16	570	15934	273142	2115	0.065	3.5	21.2	0.059	0.021
G139	10	1294	82600	167456	625	0.063	1.5	67.4	0.030	0.015
G332	16	738	25302	263774	1947	0.063	2.8	27.5	0.051	0.026
G381	13	448	13232	263969	1896	0.061	3.3	19.1	0.056	0.019
G305	16	1952	123658	165418	733	0.058	1.6	72.7	0.030	0.038
G358	16	154	1969	287107	2531	0.057	7.3	5.7	0.064	0.026
G180	12	490	30968	155646	1093	0.035	1.6	31.0	0.030	0.012
G334	12	18	241	154081	2166	0.019	6.9	0.8	0.014	0.015
G475	12	19	857	261349	2169	0.009	2.2	0.9	0.012	0.009
G098	12	13	982	70457	1345	-0.005	1.3	1.0	0.011	0.018
G462	12	11	981	70458	1347	-0.007	1.1	0.8	0.00	0.018

Table V

Results of the prediction of long-range contacts in which the contacting residues belong to two different domains. The data are for the L/5 contacts with higher predicted probability.

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

NIH-PA Author Manuscript