BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 14 2011, pages 1941–1947
doi:10.1093/bioinformatics/btr292

Structural bioinformatics

Advance Access publication May 17, 2011

Hidden conformations in protein structures

Haim Ashkenazy[1,2,3], Ron Unger[2] and Yossef Kliger[1,*]

[1]Compugen LTD, Tel Aviv 69512, [2]The Mina and Everard Goodman Faculty of Life Sciences, Bar Ilan University, Ramat-Gan, 52900 and [3]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Prediction of interactions between protein residues (contact map prediction) can facilitate various aspects of 3D structure modeling. However, the accuracy of *ab initio* contact prediction is still limited. As structural genomics initiatives move ahead, solved structures of homologous proteins can be used as multiple templates to improve contact prediction of the major conformation of an unsolved target protein. Furthermore, multiple templates may provide a wider view of the protein's conformational space. However, successful usage of multiple structural templates is not straightforward, due to their variable relevance to the target protein, and because of data redundancy issues.

Results: We present here an algorithm that addresses these two limitations in the use of multiple structure templates. First, the algorithm unites contact maps extracted from templates sharing high sequence similarity with each other in a fashion that acknowledges the possibility of multiple conformations. Next, it weights the resulting united maps in inverse proportion to their evolutionary distance from the target protein. Testing this algorithm against CASP8 targets resulted in high precision contact maps. Remarkably, based solely on structural data of remote homologues, our algorithm identified residue–residue interactions that account for all the known conformations of calmodulin, a multifaceted protein. Therefore, employing multiple templates, which improves prediction of contact maps, can also be used to reveal novel conformations. As multiple templates will soon be available for most proteins, our scheme suggests an effective procedure for their optimal consideration.

Availability: A Perl script implementing the WMC algorithm described in this article is freely available for academic use at http://tau.ac.il/~haimash/WMC.

Contact: kliger@compugen.co.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 30, 2010; revised on April 14, 2011; accepted on April 30, 2011

1 INTRODUCTION

Protein structure representations vary in their level of detail from precise atomic resolution models to the coarser contact maps that provide information at the level of the interactions between amino acid residues. Despite its low resolution, contact map representation holds advantages for specific applications (Holm and Sander, 1993; Kliger, 2010; Kliger *et al.*, 2009; Zaki, 2003). For example, in multiple conformation analysis, whereas structural alignment of conformations requires structural superimpositions, contact map representation (which is translation and rotation invariant) allows straightforward comparison of multiple conformations. These advantages motivate the development of algorithms aimed at predicting intramolecular residue–residue interactions. These algorithms often do not use any explicit structural information, but instead rely on statistical methods that quantify the correlation between columns in a multiple sequence alignment (MSA), also known as correlated mutation analysis (Ashkenazy and Kliger, 2010; Dekker *et al.*, 2004; Dunn *et al.*, 2008; Dutheil *et al.*, 2005; Eyal *et al.*, 2007; Fleishman *et al.*, 2004; Gobel *et al.*, 1994; Kass and Horovitz, 2002; Martin *et al.*, 2005; Olmea and Valencia, 1997). Improved performance has been achieved by various machine learning methods that incorporate correlated mutation measures and other features (Cheng and Baldi, 2005, 2007; Fariselli *et al.*, 2001; Miller and Eisenberg, 2008; Punta and Rost, 2005; Shackelford and Karplus, 2007; Tegge *et al.*, 2009). Currently, the best performing '*ab initio*' contact predictor is NNCon (Tegge *et al.*, 2009) as determined by the analysis performed during the Critical Assessment of Techniques for Protein Structure Prediction (CASP8) (Ezkurdia *et al.*, 2009). However, despite the considerable progress, the performance of *ab initio* contact map prediction methods is still insufficient for most potential applications (Grana *et al.*, 2005; Horner *et al.*, 2008; Izarzugaza *et al.*, 2007).

Fortunately, structural information extracted from templates can facilitate contact prediction. Some template-based predictors use various approaches to construct 3D models of the target, then extract a contact map from each model, and finally produce the consensus contact map (Gao *et al.*, 2009; Stehr and Lappe, 2008). The rationale behind this approach is that a consensus of many inaccurate models will have a much higher reliability than predictions that are based on any single model. Other template-based predictors implement machine learning techniques to incorporate contact map information extracted from several templates (Walsh *et al.*, 2009; Wu and Zhang, 2008a, c).

Whereas template-based structure prediction may allow more accurate models, it has an intrinsic drawback: the template captures the protein in a single conformation. However, the existence of an ensemble of conformations is often important for the function, interactability and evolvability of proteins (Boehr *et al.*, 2006; Boehr and Wright, 2008; Lange *et al.*, 2008; Tokuriki and Tawfik, 2009). In this study, we suggest that taking into account multiple templates offers two opportunities: (i) the ability to improve contact map

[*]To whom correspondence should be addressed.

prediction and (ii) the capacity to capture more than a single snapshot of the protein's conformational space. The latter is very appealing when different experimentally derived 3D structures of identical, or highly similar, sequences are available. Such data, which are often considered as 'redundant' when sequence-based redundancy removal procedures are used, may contain information about alternate conformations (Dan *et al.*, 2010; Kosloff and Kolodny, 2008; Zhang *et al.*, 2007). However, taking into account multiple templates raises the challenge of avoiding bias toward sequences that are overrepresented in the databases. To address this challenge, here we tested whether contact prediction that is based on multiple templates can be further improved by considering appropriate weighting of multiple templates. Our results demonstrate that an algorithm that uses multiple templates performs better than any previous contact prediction algorithm. This holds true even when the best template (which can, in reality, be chosen only in retrospect) available in the database is chosen as a single template. Furthermore, this study demonstrates that by using multiple templates for the target protein, we can capture a broader view of the protein conformational space thereby identifying contacts derived from the various conformations that the protein can adopt.

2 METHODS

Structural data: 3D structures were obtained from PDB (Berman *et al.*, 2000).

Template identification: the templates were collected as described by Godzik and colleagues (Rychlewski *et al.*, 2000) with these modifications—templates were considered if (i) they had a BLAST *e*-value $< 10^{-5}$; (ii) their release date predate the CASP8 start date (in order to fairly compare our performance to that of CASP8 predictors); (iii) they shared $<90\%$ identity with the target (to avoid homologs that are too similar to the target); (iv) the alignment between the target and template covered at least 50% of the target length and at least 50% of the template length [as was found to be the optimal criteria for homologous selection in correlated mutation analysis (Ashkenazy *et al.*, 2009)]; (v) they were determined by X-ray crystallography.

Building MSA: MSA of the target and templates sequences were built with MAFFT (v6.240) (Katoh and Toh, 2008) using the L-INS-I procedure.

Contact maps: were extracted using contacts defined by the CSU algorithm (Sobolev *et al.*, 1999), which is based on a detailed analysis of interatomic contacts and interface complementarity.

Evolutionary distance: the distances between the templates and the target protein were extracted from an evolutionary tree constructed by the SEMPHY algorithm (Ninio *et al.*, 2007). The distance matrix represented by the phylogenetic tree was extracted using the Bio::TreeIO module, which is part of the BioPerl (Stajich *et al.*, 2002) package.

Performance evaluation: the number of interactions is linearly correlated with the protein length, while the number of possible interactions is quadratic; this may result in a bias in performance evaluation in favor of short proteins. To allow a true comparison between proteins with different lengths, it is common to compare the precision [Precision = TP/(TP + FP)] achieved for the top-ranked prediction in proportion to the protein length. Thus, in this study we sorted the predictions according to their scores and evaluated the precision achieved for the Top L (L = target length) (or Top L/5, when indicated) ranked predictions for residue pairs separated by more than six residues (or 24 residues, when indicated) on the primary sequence. When we compared two contact prediction methods, the significance of the advantage of one method was based on the difference in the number of per-protein 'wins' which was compared to the null hypothesis that each of the two methods have an equal probability to outperform the other. This difference was standardized to a *Z*-score, using the total number of proteins, and the corresponding *P*-value was computed based on a standard normal distribution.

Contact map of the best template: evaluation was performed for the 65 target proteins that had at least two templates, and their best template structure (as determined by CASP8 assessors) was based on a crystal structure (Tress *et al.*, 2009).

CASP8 predictions: Predictions submitted to the CASP8 experiment were downloaded from the protein structure prediction center web site.

3 RESULTS

Multiple structural templates are already available in the PDB for about half of the CASP8 targets (Supplementary Figs S1 and S2). Below, we demonstrate that incorporating information extracted from multiple templates substantially boosts correct prediction of protein contact maps compared with all alternative approaches (Figs 1 and 2 and Supplementary Figs S5 and S7). Furthermore, we demonstrate that our algorithm often results in a contact map that captures an ensemble of conformations that the protein can adopt as illustrated for calmodulin (Fig. 3).

3.1 Current availability of potential templates for structure prediction

Historical analysis reveals that as the number of structural domains that were subject to CASP experiments increases, the fraction of domains for which reliable templates cannot be identified [termed FM (free modeling) since CASP7 and NF (new fold) before that] or that have only small resemblance to known structures (termed FM/TBM and FR/NF) sharply decreases (Supplementary Fig. S2). Furthermore, for 65 out of the 112 CASP8 TBM (template-based modeling) targets, more than one template was available (Supplementary Fig. S1). It is noteworthy that many of these templates share high sequence similarity (Supplementary Fig. S1). These findings suggest that using multiple templates for structure prediction is feasible.

3.2 Multiple conformations and evolutionary information benefits contact map prediction

Many proteins are dynamic machines, and their intramolecular interactions may vary between conformations. A comprehensive survey of PDB structures having identical or similar sequences can capture part of the conformational space (Dan *et al.*, 2010; Kosloff and Kolodny, 2008; Zhang *et al.*, 2007). Such a survey uses information that is habitually discarded in computational biology to avoid bias toward sequences overrepresented in databases such as PDB. In this study, we looked for a way to incorporate such 'redundant' information for contact prediction while minimizing the potential bias that might be caused by the unbalanced representation of these structures in PDB. We suggested that weighting the templates according to their evolutionary distances from the target protein should be adequate for this purpose.

The evolutionary distances between each template and the target protein were estimated according to the evolutionary tree that was constructed by SEMPHY (Ninio *et al.*, 2007) based on the protein sequences. Then, templates (t) were grouped according to their evolutionary distances, so all templates with a distance of less than 0.02 to the group's common ancestor were grouped together. Then, for each group (g), a binary contact map (M) was greedily constructed, where all interactions (between residue i and residue j)

found in at least one of the group members were noted.

$$M_{i,j}^{g} = \bigcup_{t \in g} M_{i,j}^{t}$$

Finally, the groups' contact maps are summed up in a weighted fashion to construct the predicted contact map. The prediction score ($S_{i,j}$) for each pair of residues is calculated as follow:

$$\forall i,j \le L \,|\, |i-j| \ge 6 : S_{i,j} = \sum_{g=1}^{G} W^{g} \cdot M_{i,j}^{g}$$

Where L is the length of the target protein, G is the number of groups and W^g is the group's weight. The weight of the contact map of each group was calculated to be equal to d^{-3}, where d is the evolutionary distance and the third power was empirically determined. This method was named 'Weighted Multiple Conformations' (WMC).

Each of the two steps used in WMC, i.e. template weighting and redundancy consideration, were evaluated separately and found to improve prediction performance. First, the importance of template weighting was demonstrated by comparing the performance of two simple contact map predictors: 'NR Weighted' (where each template weight equals d^{-3}, where d is the evolutionary distance) and 'NR Average' (where each template weight equals 1). These methods, 'NR Weighted' and 'NR Average', consider only a single representative structure out of the templates that share 100% sequence identity (the structure with the highest resolution and the lowest R factor was selected as the representative template); thus, both methods are designated as non-redundant (NR) methods. The results of this comparison reveal that weighting templates according to their evolutionary distance (i.e. 'NR Weighted') achieves significantly higher prediction performance than using a simple average ('NR Average') (Supplementary Fig. S3). The second comparison demonstrates that the way used by our 'WMC' scheme to incorporate 'redundant' information, i.e. taking into account all, rather than only one, PDB structures per sequence, does not reduce the performance as could be expected from the unequivocal representation of proteins in the PDB even when a large number of templates are used (Supplementary Fig. S4). As will be demonstrated below, such incorporation of 'redundant' information enables to capture a wider view of the protein conformational space.

3.3 Improvement over 'best template'

Predictors that use sequence and secondary structure prediction to determine structural templates often do not identify the most appropriate (best) template. Of course, the 'best template' can be only determined with hindsight knowledge by using structural alignment between the experimentally solved structure of the target protein against the PDB. Thus, at the end of the prediction period, CASP assessors use LGA (Zemla, 2003) and Mammoth (Ortiz *et al.*, 2002) to search the PDB for the 'best template' for each target (Tress *et al.*, 2009). Remarkably, most homology modeling prediction methods have difficulties in generating 3D models that are more accurate than the structure simply projected from the target's 'best template' (Kopp *et al.*, 2007). This difficulty found in 3D homology modeling prediction occurs also in contact map prediction. Thus, to demonstrate the strength of our multiple template approach, we compared our predicted contact maps with those derived from the 'best template' that were determined by CASP8 assessors by after-the-fact analysis (Tress *et al.*, 2009). The results revealed that the WMC method performs better than the contact map projected from the best template (P-value = 5.2×10^{-6}; Fig. 1). The same finding was obtained when the CASP evaluation procedure was used—calculating the precision of the top L/5 (L = protein length) predictions while taking into account only residues which are separated by > 24 residues on the primary sequence (P-value = 2.3×10^{-4}; Supplementary Fig. S5). Some of the improvement of the WMC prediction over the 'best template' can be explained by the number of available templates [Pearson's correlation coefficient (PCC) of 0.236], and by the correlation between the improvement of WMC and the root-mean-square deviation (RMSD) between the 'best-template' and the structure of the target protein (PCC = 0.353).

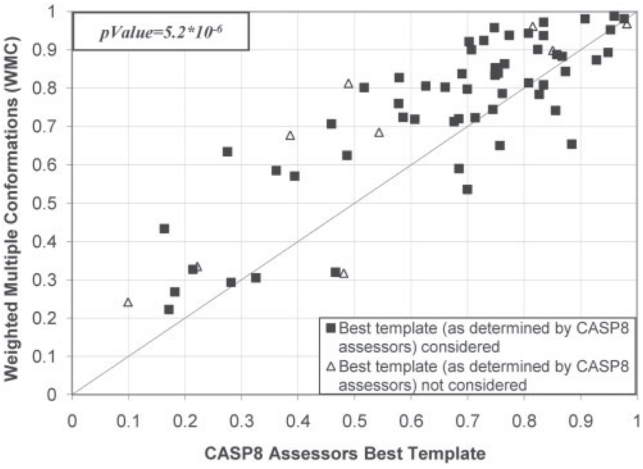

Fig. 1. Improvement over the contact map derived from the best template. The contact map projected from the best template [as determined by CASP8 assessors by structural alignment of the target structure (determined experimentally) (Ortiz *et al.*, 2002; Zemla, 2003) after the end of CASP8] is compared to our weighted multiple conformation (WMC) prediction. The precision achieved when considering the Top L (L = protein length) ranked predictions is shown. The figure reveals that the WMC predictor is more accurate than the contact map projected from the best template.

These results suggest that it is beneficial to consider templates other than the 'best' one even in the hypothetical optimal case where the 'best template' is known in advance. This conclusion is further supported by looking at the nine targets for which the CASP 'best template' was not considered by our algorithm (open triangles on Fig. 1). In seven out of these nine cases, the precision of our predictor was higher than that of the contact map projected directly from the best template.

3.4 WMC is the best contact map predictor when templates are available

CASP experiments aim to objectively compare performance of structure prediction methods on proteins whose structure is not yet available, and thus to determine the current 'state-of-the-art' method (Moult *et al.*, 2009). In CASP, contact prediction evaluations are made only for the '*free modeling*' (FM) targets, i.e. when all predictors failed to find a template for that target.

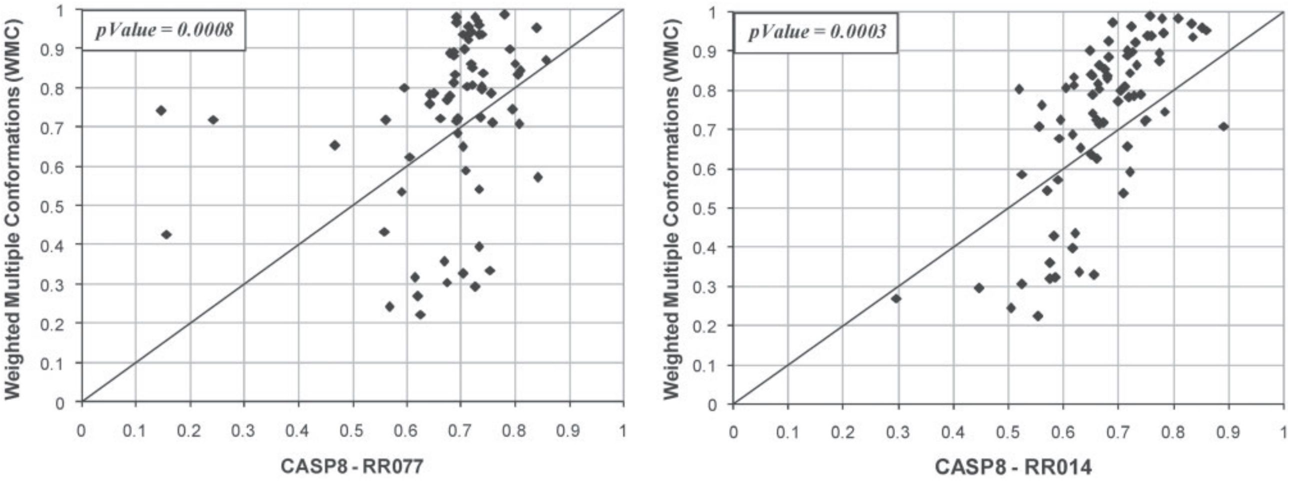

Fig. 2. Comparison of the contact maps predicted by the weighted multiple conformations (WMC) method and state-of-the-art methods. The performance (top L precision) of contact map prediction achieved by the two best groups participated in CASP8 contact-prediction category compared to our WMC method. Notably, the WMC method significantly outperforms these state-of-the-art contact prediction methods.

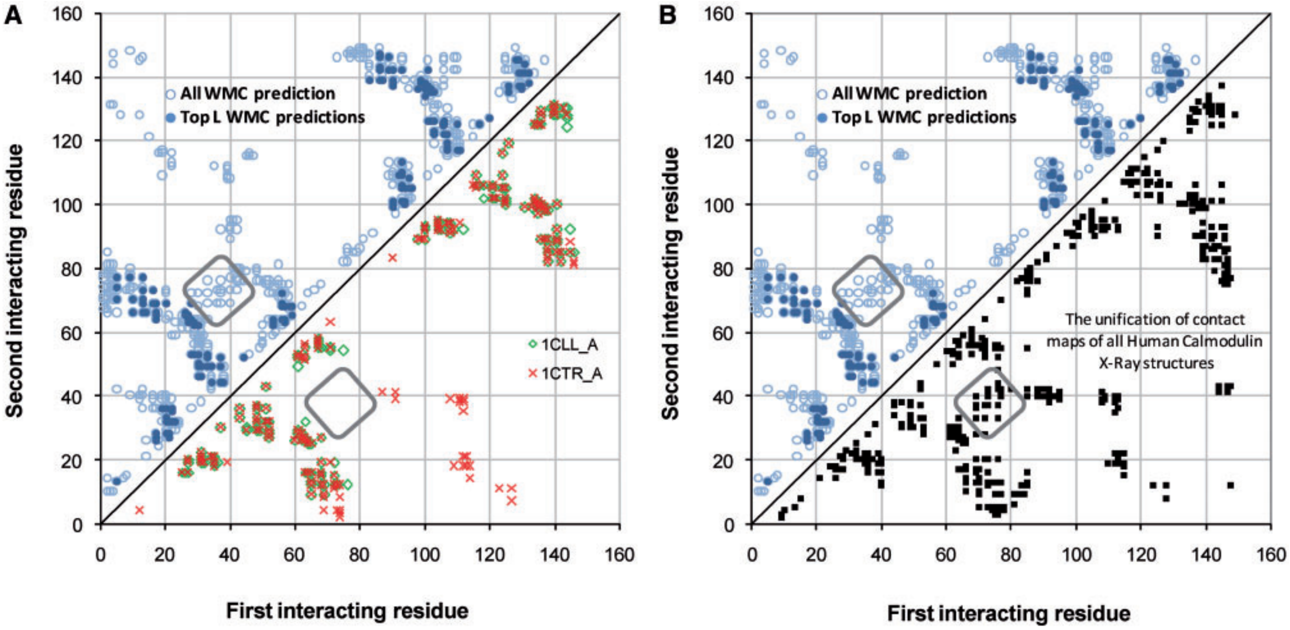

Fig. 3. Calmodulin multiple conformation contact map. (**A**) The full multiple conformation contact map as predicted by the WMC method (circles; upper left segment) is compared with the unification of contact maps of calmodulin's two conformations that appear in MolMovDB (lower right segment): PDB IDs 1CLL and 1CTR (diamonds and 'X', respectively). A helix–helix interaction predicted by the WMC method, but not identified by MolMovDB, is marked by a rectangle. (**B**) The same WMC predicted contact map (circles; upper left triangle) is compared with the 'greedy' contact map composed of the unification of contact maps of all X-ray structures of calmodulin proteins 100% identical in sequence to human calmodulin (squares; lower right triangle). Note that there is at least one PDB structure that confirms the helix–helix interaction predicted by the WMC method.

However, our method is relevant only for targets for which a template can be assigned (TBM targets). Thus, for the purpose of comparison between our method, and the current 'state-of-the-art' methods we evaluated the predicted contact maps submitted by CASP8 participants for the TBM targets (Supplementary Fig. S6). Whereas some of the predictors are based on templates, for example SVMSEQ/LOMETS (Wu and Zhang, 2007, 2008a, c) (group RR413), SMEG-CCP (Stehr and Lappe, 2008) (group RR014) and Pairings (group RR077), others do not directly use templates, for example: Hamilton-Torda-Huber (group number RR424), MULTICOM-CMFR (Cheng and Baldi, 2005; Tegge *et al.*, 2009) (group RR069) MULTICOM-RANK (Cheng and Baldi, 2007) (group RR131) and SAM-T06 (Shackelford and Karplus, 2007) (group number RR477). Notably, the methods used by groups RR069, RR131 and RR477 were ranked among the best *ab initio* methods (i.e. on FM targets), during CASP7 (Izarzugaza *et al.*,

2007) and CASP8 (Ezkurdia *et al.*, 2009); however, as can be expected, the groups that were using the information derived from templates (RR014, RR077, RR413) perform much better than the others (RR477, RR424, RR131, RR069) on the TBM targets (Supplementary Fig. S6).

A comparison between the contact maps predicted by the WMC method and those submitted by the two TBM best groups, RR077 and RR014, reveals that, for most proteins, the WMC method resulted in significantly higher precision (Fig. 2). These results are statistically significant having *P*-value of 0.0003 when compared with RR014, and *P*-value of 0.0008 when compared with RR077. This conclusion is also supported when evaluating only the top L/5 (L = protein length) predictions achieved for residues that are separated by >24 residues on the primary sequence (*P*-value of 3.47×10^{-7} when compared with RR014, and *P*-value of 0.0006 when our WMC method is compared with RR077 (Supplementary Fig. S7).

3.5 Multiple conformation prediction—calmodulin as a test case

Calmodulin is a Ca^{2+} binding protein that mediates many cellular processes in response to changes in calcium concentrations. Calmodulin interacts with many cellular partners and is known to change conformation accordingly (Akke and Chazin, 2001; Chou *et al.*, 2001). Here, human calmodulin was used to demonstrate the ability of the WMC method to predict a contact map that represents an ensemble of physiologically relevant conformations rather than a single conformation (Fig. 3).

MolMovDB (Flores *et al.*, 2006; Krebs and Gerstein, 2000) defines two representative structures for calmodulin: PDB|1CLL_A, which represents an open conformation, and PDB|1CTR_A, which represents a closed one (see lower triangle of Fig. 3A). We extracted contact maps from each of these structures, and used them to evaluate the performance of WMC. The WMC prediction was based on 114 templates that share 21–89% sequence identity with human calmodulin (we discarded all templates that share more than 90% sequence identity with human calmodulin). The results reveal that both the open and closed conformations are predicted by the WMC method (Fig. 3A). For the top L predictions (L = protein length = 149 residues), WMC achieved a precision of 0.92 and recall of 0.73 for the closed conformation, and precision of 0.93 with recall of 0.86 when tested against the open conformation.

Note that WMC also suggests another conformation for calmodulin that involves an interaction between two helices (encircled by a rectangle in Fig. 3). Interestingly, this helix–helix interaction, which is absent in PDB|1CLL_A and PDB|1CTR_A structures, does appear in the 3D structures of calmodulin complexed with the adenylyl cyclase domain of anthrax edema factor (PDB|1SK6_D and PDB|1S26_D) (Guo *et al.*, 2004; Shen *et al.*, 2004) (Fig. 3B; lower right triangle). As mentioned above, the WMC prediction was based on 114 templates. The prediction of the helix–helix interaction was supported by 32 of these templates, which share 33–53% sequence identity with calmodulin. Among these templates, there are proteins that, like calmodulin, have four calcium-binding EF hands: chicken, turkey and rabbit Troponin-C and human Centrin-2. In addition, there are templates that have three calcium-binding EF hands (myosin regulatory light chain, myosin light chain 6B, myosin light polypeptide 6 and myosin light chain 1) and one template that has only two such motifs (myosin essential light chain).

These findings demonstrate the ability of WMC to accurately predict an ensemble of contact maps that represents multiple conformations.

4 DISCUSSION

In structural biology, the ability to represent a protein as a contact map is instrumental. For example, contact map representation is useful for the identification of structural motifs (Zaki, 2003) and for quantification of structural alignment (Holm and Sander, 1993). Recently, prediction of intramolecular helix–helix interactions, which was based on predicted contact maps, enabled the design of biologically active peptides that hold therapeutic promise in cancer and inflammation (Kliger, 2010; Kliger *et al.*, 2009). In addition, several groups suggested approaches for reconstructing 3D models from contact maps (Baú *et al.*, 2006; Di Lena *et al.*, 2009; Pollastri *et al.*, 2007; Porto *et al.*, 2004; Sathyapriya *et al.*, 2009; Vassura *et al.*, 2007, 2008a, b; Vendruscolo and Domany, 2000; Vendruscolo *et al.*, 1997). Other studies use predicted contact maps as constraints in 3D modeling to enhance convergence to the best model (Latek and Kolinski, 2008; Michino and Brooks, 2009), to refine the created models (Misura *et al.*, 2006) or to eliminate incorrect models after the modeling process is completed (Miller and Eisenberg, 2008; Paluszewski and Karplus, 2009; Tress and Valencia, 2010).

In the current study, we suggest an approach for combining structural data from several templates to enhance contact map prediction of novel proteins. The basic idea is that tools that capture a wider view of the protein conformational space can provide useful insights on the contact map level, and serve as a better starting point for downstream analysis involving contact maps (e.g. 3D modeling). Indeed, as structural genomics initiatives move ahead (Bonanno *et al.*, 2005; Burley *et al.*, 2008; Chandonia and Brenner, 2006; Levitt, 2007; Nair *et al.*, 2009), it has become more likely to find at least one structural template for most proteins (Tress *et al.*, 2009). Furthermore, as revealed in Supplementary Figure S1, many proteins already have at least two potential templates.

The contact map predicted by the WMC algorithm succeeded in two different tests: WMC exhibits better performance than any other contact prediction method when evaluated against a single conformation (as done in CASP8; Fig. 2 and Supplementary Figs S5 and S7), and it also enables the prediction of a united contact map that considers the multiple conformations that a protein can adopt (as demonstrated for human calmodulin; Fig. 3).

This study suggests that the performance of template-based contact prediction methods can be improved significantly by using information extracted from multiple templates. A similar conclusion was suggested for the case of 3D modeling (Cheng, 2008; Larsson *et al.*, 2008).

Several methods such as Modeller (Eswar *et al.*, 2006; Sali and Blundell, 1993) and SwissModel (Bordoli *et al.*, 2009) utilize structural constraints derived from a template, or multiple templates, to assist 3D modeling. The dynamic nature of proteins, as exemplified by the variation of intramolecular contacts observed in different conformations (Bywater, 2010) pose the challenge of predicting contacts in each of the different conformations. Our study suggests a scheme for template-based contact prediction that captures a wider view of the protein conformational space. The scheme described here is based on three concepts that—to the best of our knowledge—were not described previously: (i) weighting multiple templates according to their evolutionary proximity to the target protein; (ii) using multiple templates in the prediction process in a fashion that avoids bias toward 'redundant' templates; and (iii) predicting a contact map that may represent more than one snapshot of the protein's conformational space. The first two ideas described above can be further utilized for 3D modeling.

Several major directions can be considered when trying to improve the framework suggested in this study. One major direction is the use of structural data for searching and aligning the templates. Optimal selection of templates, including remote homologs, may be achieved by algorithms that identify evolutionarily remote templates sharing similar structures with the target, even in the absence of sequence similarity (Cheng and Baldi, 2006; Debe *et al.*, 2006; Ginalski *et al.*, 2003; Jaroszewski *et al.*, 2005; Karplus *et al.*, 1998; Soding, 2005; Wu and Zhang, 2008b).

WMC uses MAFFT (Katoh and Toh, 2008; Katoh *et al.*, 2002, 2005), which is a method for MSA based only on sequence information. A more accurate MSA, which may be offered by methods that incorporate structural information through structural alignment (Armougom *et al.*, 2006; O'Sullivan *et al.*, 2004), may further improve performance. In addition, considering the evolutionary forces for each pair of residues, rather than for the entire protein, might also dramatically improve performance.

To further benefit from the information about possible multiple conformations suggested by the WMC method, it is reasonable to try to separate different conformations. Whereas the 'united contact map' may include some mutually exclusive contacts that cannot coexist in a single state of the protein—each reflects a different conformation—appropriate mathematical manipulations can make it useful for various structural analyses, even without such separation. For example, Fourier transform can enhance the signal of a specific conformation and enable reliable predictions of helix–helix interactions (Kliger *et al.*, 2009). In addition, using the scores assigned to each of the predicted residue–residue interactions can give some information about the abundance of each predicted conformation (at least as reflected by the current databases).

In summary, this study describes an effective approach for automatic integration of multiple templates in a weighted fashion for contact map prediction in proteins. The best choice of weighting functions to utilize multiple templates and to consider multiple conformations will require further research. Yet, the results presented in this study clearly demonstrate that using multiple templates and multiple conformations in a weighted fashion can dramatically improve protein contact map prediction. Our method provides the most accurate prediction when multiple templates are available, and we believe that a similar approach will prove useful to developers of 3D modeling methods.

ACKNOWLEDGEMENTS

The authors are grateful to A. K. Dunker, E. Schreiber, I. Borukhov, T. Menasherov, A. Dan, I. Shub, G. Almogy, O. Cohen, D. Burstein, M. Peeri, R. Khosravi and A. Apatoff for technical help, useful comments and helpful discussions.

Conflict of Interest: none declared.

REFERENCES

Akke,M. and Chazin,W.J. (2001) An open and shut case. *Nat. Struct. Biol.*, **8**, 910–912.

Armougom,F. *et al.* (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.*, **34**, W604–W608.

Ashkenazy,H. and Kliger,Y. (2010) Reducing phylogenetic bias in correlated mutation analysis. *Protein Eng. Des. Sel.*, **23**, 321–326.

Ashkenazy,H. *et al.* (2009) Optimal data collection for correlated mutation analysis. *Proteins*, **74**, 545–555.

Baú,D. *et al.* (2006) Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. *BMC Bioinformatics*, **7**, 402.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Boehr,D.D. *et al.* (2006) An NMR perspective on enzyme dynamics. *Chem. Rev.*, **106**, 3055–3079.

Boehr,D.D. and Wright,P.E. (2008) Biochemistry. How do proteins interact? *Science*, **320**, 1429–1430.

Bonanno,J.B. *et al.* (2005) New York-Structural GenomiX Research Consortium (NYSGXRC): a large scale center for the protein structure initiative. *J. Struct. Funct. Genomics*, **6**, 225–232.

Bordoli,L. *et al.* (2009) Protein structure homology modeling using SWISS-MODEL workspace. *Nat. Protoc.*, **4**, 1–13.

Burley,S.K. *et al.* (2008) Contributions to the NIH-NIGMS Protein Structure Initiative from the PSI Production Centers. *Structure*, **16**, 5–11.

Bywater,R. (2010) Solving the protein folding problems. *Available from Nature Precedings*. Availabe at http://hdl.handle.net/10101/npre.2010.4730.1.

Chandonia,J.M. and Brenner,S.E. (2006) The impact of structural genomics: expectations and outcomes. *Science*, **311**, 347–351.

Cheng,J. (2008) A multi-template combination algorithm for protein comparative modeling. *BMC Struct. Biol.*, **8**, 18.

Cheng,J. and Baldi,P. (2005) Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, **21** (Suppl. 1), i75–i84.

Cheng,J. and Baldi,P. (2006) A machine learning information retrieval approach to protein fold recognition. *Bioinformatics*, **22**, 1456–1463.

Cheng,J. and Baldi,P. (2007) Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, **8**, 113.

Chou,J.J. *et al.* (2001) Solution structure of Ca(2+)-calmodulin reveals flexible hand-like properties of its domains. *Nat. Struct. Biol.*, **8**, 990–997.

Dan,A. *et al.* (2010) Large-scale analysis of secondary structure changes in proteins suggests a role for disorder-to-order transitions in nucleotide binding proteins. *Proteins*, **78**, 236–248.

Debe,D.A. *et al.* (2006) STRUCTFAST: protein sequence remote homology detection and alignment using novel dynamic programming and profile-profile scoring. *Proteins*, **64**, 960–967.

Dekker,J.P. *et al.* (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, **20**, 1565–1572.

Di Lena,P. *et al.* (2009) On the reconstruction of three-dimensional protein structures from contact maps. *Algorithms*, **2**, 76.

Dunn,S.D. *et al.* (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.

Dutheil,J. *et al.* (2005) A model-based approach for detecting coevolving positions in a molecule. *Mol. Biol. Evol.*, **22**, 1919–1928.

Eswar,N. *et al.* (2006) Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics*, **Chapter 5**, Unit 5 6.

Eyal,E. *et al.* (2007) A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction. *Proteins*, **67**, 142–153.

Ezkurdia,I. *et al.* (2009) Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins*, **77** (Suppl. 9), 196–209.

Fariselli,P. *et al.* (2001) Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins*, **45** (Suppl. 5), 157–162.

Fleishman,S.J. *et al.* (2004) An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. *J. Mol. Biol.*, **340**, 307–318.

Flores,S. *et al.* (2006) The Database of Macromolecular Motions: new features added at the decade mark. *Nucleic Acids Res.*, **34**, D296–D301.

Gao,X. *et al.* (2009) Improving consensus contact prediction via server correlation reduction. *BMC Struct. Biol.*, **9**, 28.

Ginalski,K. *et al.* (2003) ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Res.*, **31**, 3804–3807.

Gobel,U. *et al.* (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.

Grana,O. *et al.* (2005) CASP6 assessment of contact prediction. *Proteins*, **61** (Suppl. 7), 214–224.

Guo,Q. *et al.* (2004) Structural and kinetic analyses of the interaction of anthrax adenylyl cyclase toxin with reaction products cAMP and pyrophosphate. *J. Biol. Chem.*, **279**, 29427–29435.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Horner,D.S. *et al.* (2008) Correlated substitution analysis and the prediction of amino acid structural contacts. *Brief Bioinform.*, **9**, 46–56.

Izarzugaza,J.M. *et al.* (2007) Assessment of intramolecular contact predictions for CASP7. *Proteins*, **69**, 152–158.

Jaroszewski,L. *et al.* (2005) FFAS03: a server for profile–profile sequence alignments. *Nucleic Acids Res.*, **33**, W284–W288.

Karplus,K. *et al.* (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Kass,I. and Horovitz,A. (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins*, **48**, 611–617.

Katoh,K. *et al.* (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.

Katoh,K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.*, **9**, 286–298.

Kliger,Y. (2010) Computational approaches to therapeutic peptide discovery. *Biopolymers*, **94**, 701–710.

Kliger,Y. *et al.* (2009) Peptides modulating conformational changes in secreted chaperones: from in silico design to preclinical proof of concept. *Proc. Natl Acad. Sci. USA*, **106**, 13797–13801.

Kopp,J. *et al.* (2007) Assessment of CASP7 predictions for template-based modeling targets. *Proteins*, **69** (Suppl. 8), 38–56.

Kosloff,M. and Kolodny,R. (2008) Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins*, **71**, 891–902.

Krebs,W.G. and Gerstein,M. (2000) The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res.*, **28**, 1665–1675.

Lange,O.F. *et al.* (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, **320**, 1471–1475.

Larsson,P. *et al.* (2008) Using multiple templates to improve quality of homology models in automated homology modeling. *Protein Sci.*, **17**, 990–1002.

Latek,D. and Kolinski,A. (2008) Contact prediction in protein modeling: scoring, folding and refinement of coarse-grained models. *BMC Struct. Biol.*, **8**, 36.

Levitt,M. (2007) Growth of novel protein structural data. *Proc. Natl Acad. Sci. USA*, **104**, 3183–3188.

Martin,L.C. *et al.* (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, **21**, 4116–4124.

Michino,M. and Brooks,C.L. III (2009) Predicting structurally conserved contacts for homologous proteins using sequence conservation filters. *Proteins*, **77**, 448–453.

Miller,C.S. and Eisenberg,D. (2008) Using inferred residue contacts to distinguish between correct and incorrect protein models. *Bioinformatics*, **24**, 1575–1582.

Misura,K.M. *et al.* (2006) Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proc. Natl Acad. Sci. USA*, **103**, 5361–5366.

Moult,J. *et al.* (2009) Critical assessment of methods of protein structure prediction - round VIII. *Proteins*, **77** (Suppl. 9), 1–4.

Nair,R. *et al.* (2009) Structural genomics is the largest contributor of novel structural leverage. *J. Struct. Funct. Genomics*, **10**, 181–191.

Ninio,M. *et al.* (2007) Phylogeny reconstruction: increasing the accuracy of pairwise distance estimation using Bayesian inference of evolutionary rates. *Bioinformatics*, **23**, e136–e141.

O'Sullivan,O. *et al.* (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.*, **340**, 385–395.

Olmea,O. and Valencia,A. (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des.*, **2**, S25–S32.

Ortiz,A.R. *et al.* (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

Paluszewski,M. and Karplus,K. (2009) Model quality assessment using distance constraints from alignments. *Proteins*, **75**, 540–549.

Pollastri,G. *et al.* (2007) Distill: a machine learning approach to ab initio protein structure prediction. In Bandyopadhyay,S. *et al.* (eds), World Scientific, pp. 153–183.

Porto,M. *et al.* (2004) Reconstruction of protein structures from a vectorial representation. *Phys. Rev. Lett.*, **92**, 218101.

Punta,M. and Rost,B. (2005) PROFcon: novel prediction of long-range contacts. *Bioinformatics*, **21**, 2960–2968.

Rychlewski,L. *et al.* (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Sathyapriya,R. *et al.* (2009) Defining an essence of structure determining residue contacts in proteins. *PLoS Comput. Biol.*, **5**, e1000584.

Shackelford,G. and Karplus,K. (2007) Contact prediction using mutual information and neural nets. *Proteins*, **69** (Suppl. 8), 159–164.

Shen,Y. *et al.* (2004) Structure of anthrax edema factor-calmodulin-adenosine 5′-(alpha,beta-methylene)-triphosphate complex reveals an alternative mode of ATP binding to the catalytic site. *Biochem. Biophys. Res. Commun.*, **317**, 309–314.

Sobolev,V. *et al.* (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.

Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Stajich,J.E. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.

Stehr,B.H. and Lappe,M. (2008) Prediction of native contacts, 3D structures and model quality using consensus contacts. In *8th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, Abstracts*, pp. 108–109.

Tegge,A.N. *et al.* (2009) NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res.*, **37**, W515–W518.

Tokuriki,N. and Tawfik,D.S. (2009) Protein dynamism and evolvability. *Science*, **324**, 203–207.

Tress,M.L. *et al.* (2009) Target domain definition and classification in CASP8. *Proteins*, **77** (Suppl. 9), 10–17.

Tress,M.L. and Valencia,A. (2010) Predicted residue-residue contacts can help the scoring of 3D models. *Proteins*, **78**, 1980–1991.

Vassura,M. *et al.* (2007) Fault tolerance for large scale protein 3D reconstruction from contact maps. *Lect. Notes Comput. Sci.*, **4645**, 25.

Vassura,M. *et al.* (2008a) FT-COMAR: fault tolerant three-dimensional structure reconstruction from protein contact maps. *Bioinformatics*, **24**, 1313.

Vassura,M. *et al.* (2008b) Reconstruction of 3D structures from protein contact maps. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **5**, 357–367.

Vendruscolo,M. and Domany,E. (2000) Protein folding using contact maps. *Vitam. Horm.*, **58**, 171–212.

Vendruscolo,M. *et al.* (1997) Recovery of protein structure from contact maps. *Fold Des.*, **2**, 295–306.

Walsh,I. *et al.* (2009) Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. *BMC Struct. Biol.*, **9**, 5.

Wu,S. and Zhang,Y. (2007) LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.*, **35**, 3375–3382.

Wu,S. and Zhang,Y. (2008a) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics*, **24**, 924–931.

Wu,S. and Zhang,Y. (2008b) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*, **72**, 547–556.

Wu,S. and Zhang,Y. (2008c) Protein residue contact prediction by SVMSEQ and LOMETS servers. In *8th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, Abstracts*, pp. 114–115.

Zaki,M. (2003) Mining data in Bioinformatics. In *Handbook of Data Mining*. Lawrence Earlbaum Associates, Mahwah, NJ, pp. 573–597.

Zemla,A. (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374.

Zhang,Y. *et al.* (2007) Between order and disorder in protein structures: analysis of 'dual personality' fragments in proteins. *Structure*, **15**, 1141–1147.