Hidden conformations in protein structures
Haim Ashkenazy¹,²,³, Ron Unger² and Yossef Kliger*¹
¹ Compugen LTD, Tel Aviv, 69512, Israel
² The Mina and Everard Goodman Faculty of Life Sciences, Bar Ilan University, Ramat-Gan, 52900, Israel
³ The Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, 69978, Israel

Associate Editor: Prof. 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 unified 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 multi-faceted 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

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 straight-forward comparison of multiple conformations. These advantages motivate the development of algorithms aimed at predicting intra-molecular 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, 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; Cheng and Baldi, 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, 2008; Wu and Zhang, 2008).

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 con-
tact map 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 towards 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 identify 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 & colleagues (Rychlewski, et al., 2000) with these modifications: Templates were considered if (a) they had a BLAST e-value<10^-5; (b) their release date predate the CASP8 start date (in order to fairly compare our performance to that of CASP8 predictors); (c) they shared less than 90% identity with the target (to avoid homologs that are too similar to the target); (d) 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)); (e) they were determined by X-ray crystallography.

Building MSA. Multiple Sequence Alignment 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 CSU (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=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.

3 RESULTS

Multiple structural templates are already available in the PDB for about half of the CASP8 targets (Figures S1 and S2). Below, we demonstrate that incorporating information extracted from multiple templates substantially boosts correct prediction of protein contact maps compared to all alternative approaches (Figures 1, 2, 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 (Figure 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 (Figure S2). Furthermore, for 65 out of the 112 CASP8 TBM (template based modeling) targets, more than one template was available (Figure S1). It is noteworthy that many of these templates share high sequence simi-
Weighted Multiple Conformations (WMC)

S-score (method to construct the predicted contact map. The prediction found in at least one of the group members were noted: constructed, where all interactions (between residue a b in each group (a binary contact map) was greedily constructed, where residue i and residue j found in at least one of the group members were noted:

\[ M_{i,j}^{\gamma} = \bigcup_{t \in g} M_{i,j}^{t} \]

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:

\[ \forall i, j \leq L, |i-j| \geq 6 : S_{i,j} = \sum_{g=1}^{G} W_{\gamma} \cdot M_{i,j}^{\gamma} \]

Where \( L \) is the length of the target protein, \( G \) is the number of groups, and \( W_{\gamma} \) 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’) (Figure 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 (Figure 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,
least one PDB structure that confirms the helix-helix interaction predicted by the WMC method.

Figure 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 comprised of 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.

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 (pValues=5.2*10^-7, Figure 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 more than 24 residues on the primary sequence (pValues=2.3*10^-4; Figure S5). Some of the improvement of the WMC prediction over the ‘best template’ can be explained by the number of available templates (Pearson Correlation Coefficient (PCC) of 0.236), and by the correlation between the improvement of WMC and the RMSD between the ‘best-template’ and the structure of the target protein (PCC=0.353).

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 Figure 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. 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 (Figure S6). Whereas some of the predictors are based on templates, for example SVMSEQ/LOMETS (Wu and Zhang, 2007; Wu and Zhang, 2008; Wu and Zhang, 2008) (group RR413), SMEG-CCP (Steinh 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; Tege, 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 amongst 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 (Figure 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 (Figure 2). These results are statistically significant having $p$Value of 0.0003 when compared to RR014, and $p$Value of 0.0008 when compared to RR077. This conclusion is also supported when evaluating only the top $L/5$ ($L$=protein length) predictions achieved for residues that are separated by more than 24 residues on the primary sequence ($p$Value of 3.47*10^{-7} when compared to RR014, and $p$Value of 0.0006 when our WMC method is compared to RR077; Figure 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 (Figure 3).

MolMovDB (Flores, et al., 2006; Krebs and Gerstein, 2000) defines two representative structures for Calmodulin: PDB1CLL_A, which represents an open conformation, and PDB1CTR_A, which represents a closed one (see lower triangle of Figure 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 (Figure 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 PDB1CLL_A and PDB1CTR_A structures, does appear in the 3D structures of Calmodulin complexed with the adenyllyl cyclase domain of anthrax edema factor (PDB1SK6_D and PDB1S26_D) (Guo, et al., 2004; Shen, et al., 2004) (Figure 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 contact map 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 (Baü, et al., 2006; Di Lena, et al., 2009; Pollarasti, et al., 2007; Porto, et al., 2004; Sathyapriya, et al., 2009; Vassura, et al., 2007; Vassura, et al., 2008; Vassura, et al., 2008; 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 complete (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 Brenmer, 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 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; Figs 2, 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; Figure 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 intra-molecular 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: (a) weighting multiple templates according to their evolutionary proximity to the target protein; (b) using multiple templates in the prediction process in a fashion that avoids bias towards ‘redundant’ templates; and (c) 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 (Altschul, et al., 1997; Cheng and Baldi, 2006; Debe, et al., 2006; Ginalska, et al., 2003; Jaroszewski, et al., 2005; Karplus, et al., 1998; Rychlewski, et al., 2000; Soding, 2005; Wu and Zhang, 2008).

WMC uses MAFFT (Katoh, et al., 2005; Katoh, et al., 2002; Katoh and Toh, 2008), which is a method for multiple sequence alignment (MSA) based only on sequence information. A more accurate MSAs, which may be offered by methods that incorporate structural information through structural alignment (Arumugom, 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 co-exist 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 (Kluger, 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.

REFERENCES


