Bioinformatics, 34(13), 2018, 2201–2208
doi: 10.1093/bioinformatics/bty074
Advance Access Publication Date: 12 February 2018
Original Paper

OXFORD

Structural bioinformatics

Enhancing protein fold determination by exploring the complementary information of chemical cross-linking and coevolutionary signals

Ricardo N. dos Santos[1,2], Allan J. R. Ferrari[1], Hugo C. R. de Jesus[1], Fábio C. Gozzo[1], Faruck Morcos[3] and Leandro Martínez[1,2,*]

[1]Institute of Chemistry, [2]Center for Computational Engineering and Sciences, University of Campinas, Campinas, SP 13083-970, Brazil and [3]Department of Biological Sciences, University of Texas at Dallas, Richardson, TX 75080-3021, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on September 19, 2017; revised on January 19, 2018; editorial decision on February 7, 2018; accepted on February 10, 2018

Abstract

Motivation: Elucidation of protein native states from amino acid sequences is a primary computational challenge. Modern computational and experimental methodologies, such as molecular coevolution and chemical cross-linking mass-spectrometry allowed protein structural characterization to previously intangible systems. Despite several independent successful examples, data from these distinct methodologies have not been systematically studied in conjunction. One challenge of structural inference using coevolution is that it is limited to sequence fragments within a conserved and unique domain for which sufficient sequence datasets are available. Therefore, coupling coevolutionary data with complimentary distance constraints from orthogonal sources can provide additional precision to structure prediction methodologies.

Results: In this work, we present a methodology to combine residue interaction data obtained from coevolutionary information and cross-linking/mass spectrometry distance constraints in order to identify functional states of proteins. Using a combination of structure-based models (SBMs) with optimized Gaussian-like potentials, secondary structure estimation and simulated annealing molecular dynamics, we provide an automated methodology to integrate constraint data from diverse sources in order to elucidate the native conformation of full protein systems with distinct complexity and structural topologies. We show that cross-linking mass spectrometry constraints improve the structure predictions obtained from SBMs and coevolution signals, and that the constraints obtained by each method have a useful degree of complementarity that promotes enhanced fold estimates.

Availability and implementation: Scripts and procedures to implement the methodology presented herein are available at https://github.com/mcubeg/DCAXL.

Contact: leandro@iqm.unicamp.br

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

1 Introduction

Elucidation of the three-dimensional functional conformation of proteins is a key step to understand fundamental biochemical processes of living organisms (Alberts, 1998; Alberts *et al.*, 2014; Piccolino, 2000). Although the native state of a protein is directly determined by its amino acid sequence (as stated by Anfinsen's dogma), the very large number of degrees of freedom and consideration of various physico–chemical environments turns the prediction of protein 3-D structures a perplexing problem (Anfinsen, 1973; Dobson *et al.*, 1998; Dobson, 2003; Dill and MacCallum, 2012). During the last two decades, numerous methodologies have been developed to perform *in silico* prediction of native states of proteins (Baker, 2014; Baker and Sali, 2001; Cooper *et al.*, 2010; Honig, 1999; Rohl *et al.*, 2004; Roy *et al.*, 2010; Webster, 2000;). Although several methods have shown substantial accuracy in identifying folding architectures of specific systems, their applicability is usually limited to comparative modeling or requires massive computational power (Bender *et al.*, 2016; Dill and MacCallum, 2012; Freddolino *et al.*, 2010; Kelley *et al.*, 2015; Piana *et al.*, 2014; Roche and McGuffin, 2016; Yang *et al.*, 2015).

In this context, coevolutionary signals have been used with remarkable success in the identification of inter- and intramolecular protein interactions related to a broad range of functional states (Göbel *et al.*, 1994; de Juan *et al.*, 2013; Morcos *et al.*, 2011; 2014; Taylor *et al.*, 2013). It is based on the principle that during the differentiation of a protein family along divergent evolution, mutation events in residues that are critical to protein functionality are compensated by complementarity mutations (Göbel *et al.*, 1994; de Juan *et al.*, 2013; Morcos *et al.*, 2011; Shindyalov *et al.*, 1994). When sufficient sequence data are available, statistical methods can be applied in multiple sequence alignments (MSAs) to estimate the correlation between these pairwise mutations and to identify co-evolving residues that typically are a proxy for spatial proximity in the native state (Morcos *et al.*, 2011). Several molecular modeling techniques have been successfully adapted to include such co-evolutionary couplings as parameters to assist fold recognition and elucidate the organization of oligomeric complexes (Hopf *et al.*, 2012; Marks *et al.*, 2012; Morcos *et al.*, 2011; Ovchinnikov *et al.*, 2014; dos Santos *et al.*, 2015; Sułkowska *et al.*, 2012; Sutto *et al.*, 2015). RNA structure elucidation has also benefit from the use of global methods to extract residue interactions (De Leonardis *et al.*, 2015; Taylor and Hamilton, 2017; Weinreb *et al.*, 2016).

Another distinct and promising state-of-the-art methodology to infer structural information about biomolecular systems is the combination of chemical cross-linking (XL) and mass spectrometry (MS) techniques (XL-MS) (Sinz *et al.*, 2015; Young *et al.*, 2000). Most commonly, cross-links are obtained by the exposure of a target protein to bifunctional chemical linkers able to react with specific protein residue side chains. Typically, proteins are subjected to tryptic digestion followed by MS analysis. Identification of modified peptides provides information from pairwise maximum distance limits that can be used to restrict search through the protein conformational space (Sinz, 2006; Sinz *et al.*, 2015). Recent advances in mass spectrometry instrumentation, the establishment of robust cross-linking protocols and the development of specialized software for cross-linking identification have expanded the applicability of XL-MS to assist protein fold determination and complex predictions (Brodie *et al.*, 2017; Hofmann *et al.*, 2015; Jin Lee, 2008; Liu *et al.*, 2015; Nguyen-Huynh *et al.*, 2015; Paramelle *et al.*, 2013; Petrotchenko *et al.*, 2014; Pereira *et al.*, 2014; Santos *et al.*, 2011; Sinz, 2006).

Concerning the conformational search intrinsic to any structural prediction method, the use of simplified representations of the protein structure allows an efficient search of the conformational space. For example, coarse-grained models using only C_α atoms are proven to be practical in the context of structure prediction from coevolutionary constraints and also for the analysis of folding energy landscapes (Bryngelson *et al.*, 1995; Onuchic *et al.*, 1997; Onuchic and Wolynes, 2004; Wolynes *et al.*, 1995). These simplified models, which originally were conceived using distance constraints obtained from the crystallographic models are called structure-based models (SBMs) and proved to properly represent not only native states but also the multi-dimensional energy funnel that allows the observation of ensembles of intermediate states (Bryngelson *et al.*, 1995; Dill *et al.*, 2008; Onuchic *et al.*, 1997; Wolynes *et al.*, 1995). With an SBM, it is possible to efficiently explore the energy landscape of folding implied by structural data, which is incorporated as interaction potentials in biased molecular dynamic simulations (Clementi *et al.*, 2000; Noel *et al.*, 2010; Onuchic and Wolynes, 2004; Whitford *et al.*, 2009). Lately, the integration of SBM and coevolutionary signals has shown to be an efficient framework to study the protein conformational changes (Morcos *et al.*, 2013; Sfriso *et al.*, 2016), complex formation (dos Santos *et al.*, 2015) and the functional conformation of small globular systems (Sułkowska *et al.*, 2012). Recent coarse grained models like Associative Memory, Water Mediated, Structure and Energy Model (AWSEM, Chen *et al.*, 2016; Davtyan *et al.*, 2012) that include memory terms for fragments and optimized potential have also integrated evolutionary restrains successfully to make estimates of protein folds (Sirovetz *et al.*, 2017).

In this study, we show that SBMs can be used to obtain structural models of the tertiary structure of proteins by incorporating distance constraints obtained from coevolutionary information with those obtained by chemical cross-linking mass-spectrometry, in an efficient and complementary fashion that leads to more robust and accurate structural predictions.

2 Materials and methods

2.1 Estimation of coevolutionary couplings

The comparison of protein sequences within a specific domain can provide information about correlated mutations and aid the inference of physical contacts among residues (Göbel *et al.*, 1994; Shindyalov *et al.*, 1994). A very effective method to compute direct couplings in a MSA, that are typically predictors of physically interacting residue pairs, is direct coupling analysis (DCA; Morcos *et al.*, 2011; Weigt *et al.*, 2009). A detailed description is provided in Supplementary Section S1.

We have used coevolutionary information from DCA to predict protein structures in the past for several systems (Sułkowska *et al.*, 2012), an approach also used successfully by several others (Hayat *et al.*, 2015; Hopf *et al.*, 2015, 2012; Michel *et al.*, 2017; Ovchinnikov *et al.*, 2017). In this work, we are concerned about the effects of integrating experimental data with such coevolutionary signals to improve the process of structural estimation. For this purpose, we selected a set of five protein systems (Table 1) with different degrees of evolutionary coupling accuracies (Cherfils *et al.*, 1997; Luhavaya *et al.*, 2015; Ohren *et al.*, 2007; Stenkamp, 2008; Trajtenberg *et al.*, 2014; Zhang *et al.*, 2010). For all these systems, MSAs were obtained from the Pfam protein family database. Top DCA pairs with highest correlation were selected in equal number L to the length of domain in MSAs. A number of couplings close to

Table 1. Systems selected for folding prediction

System	PDB	Pfam ID	sequences	I ± SD (%)	DL	FL
SalBIII	5CXO	PF12680	8806	0.17 ± 0.05	104	134
DesR	4LE1	PF00072	31596	0.25 ± 0.07	111	132
RAP2A	1KAO	PF00071	16898	0.3 ± 0.1	160	167
Rhodopsin	3C9L	PF00001	27067	0.20 ± 0.07	252	326
Abl Kinase	3K5V	PF07714	16405	0.30 ± 0.08	250	286
Creatine Kinase	1U6R	PF00217	1182	0.3 ± 0.2	214	380
		PF02807	778	0.5 ± 0.2	71	

Note. DL: pfam domain length; FL: full protein length; I: mean identity of multiple alignment.

the length of the protein have been proposed to be sufficient for efficient structure determination (Kamisetty *et al.*, 2013).

2.2 Cross-linking/mass spectrometry constraints

We combined DCA constraints, i.e. couplings with high Direct Information (DI) values, with chemical cross-linking/mass-spectrometry (XL) constraints. The constraints were obtained either experimentally or by modeling the expected constraints from the crystallographic model using the Topolink software (Martinez *et al.*, 2017), which computes the solvent-accessible paths and distances for a linker connecting potentially reactive residues. Effective maximum distances for each type of cross-linking considered are listed in Supplementary Table S2.

The experimental dataset for SalBIII (eXL) was comprised by 38 constraints, which resulted from the use of commonly commercial DSS cross-linker and a novel chemistry, named Xplex, which is able to use 1, 6-hexanediamine as a linker as well as to produce simultaneously the formation of zero-length species (unpublished data). No evidence for quaternary cross-links was obtained, suggesting that SalBIII was monomeric in solution. A scheme of the possible linked residues and the experimental constraints distribution over SalBIII sequence is presented in Supplementary Figure S1.

In the case of the SalBIII system, an ideal cross-linking experiment would provide 74 constraints, as predicted by Topolink. Experimentally, 38 constraints were recovered (51%). This relation was used to estimate the limitations in XL experimental determination. Therefore, for the other examples, for which experimental restrictions are not available, XL constraints were obtained by randomly selecting 50% of the crosslinks predicted by Topolink, analogous to the observed SalBIII results.

2.3 Estimation of pairwise equilibrium distances for DCA constraints

Previous conformational studies using DCA signals as interaction potentials for SBM systematically employ a unique equilibrium distance for C_α pairs to represent predicted interaction restrictions independently of residue types (dos Santos *et al.*, 2015; Schug *et al.*, 2009; Sfriso *et al.*, 2016). In an effort to provide a better description of residue–residue interaction distances, we performed a statistical analysis of a large dataset of protein conformations from protein data bank (PDB). We analyzed 43 606 deposited crystallographic structures within 2 Å resolution and computed the C_α–C_α distances for all physically interacting pairs in unique chains. An statistical estimator corresponding to the peak of the distance distribution was designated as equilibrium distance for each specific pair of residue types (Supplementary Figure S2). This estimation resulted in a general improvement in prediction accuracy.

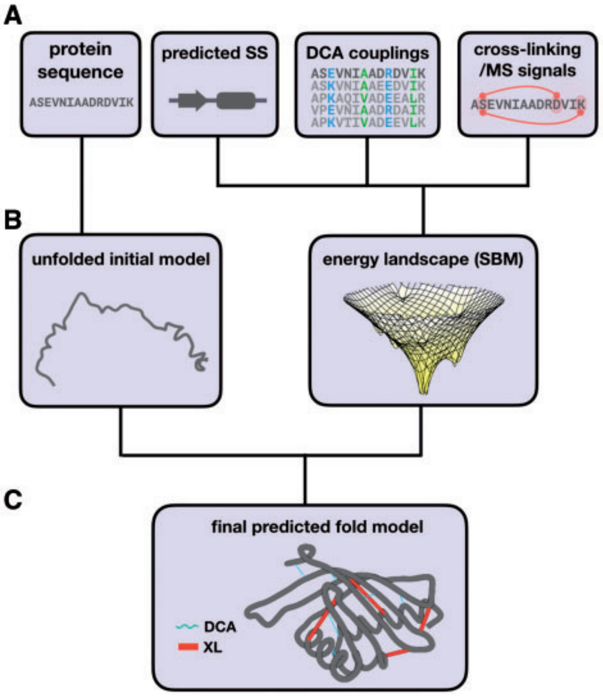

Fig. 1. Schematic representation of the steps required for generating protein fold predictions. (**A**) Primary sequence of a target protein is used to predict the type of secondary structure. An MSA of a protein family is used to estimate coevolving pairs. Interaction signals obtained by chemical cross-linking coupled to MS are also obtained. (**B**) These datasets are merged to generate an initial unfolded model and a energy landscape (SBM, i.e a customized force-field) for conformational search. (**C**) Short molecular dynamics simulations using temperature annealing are carried out to identify conformations with optimal restriction agreement

2.4 SBMs

In order to explore the folding of a diverse set of proteins driven by coevolution and XL/MS signals, initial unfolded models for each system were generated using coarse-grained SBMs with residues represented only by C_α atoms (Clementi *et al.*, 2000; Matysiak and Clementi, 2004). These unfolded models are composed by a linear arrangement of C_α beads with null parameters for all dihedrals. Parameters for bound interactions were generated using an in-house algorithm (available at: https://github.com/mcubeg/DCAXL). Bonding angles and dihedrals for each region of protein sequence were estimated by computing the secondary structure with Jpred (Drozdetskiy *et al.*, 2015) and setting optimal parameters based on ideal α-helix and β-sheet structures (Supplementary Table S1). Furthermore, this simple strategy allows the application of secondary structure predictions from diverse sources.

Based on the top ranked coevolving pairs from DCA and observed cross-linking, Gaussian-like potentials were generated to represent each pairwise interaction as described in the Supplementary Material (Lammert *et al.*, 2009; Noel and Onuchic, 2012; dos Santos *et al.*, 2015). For cross-linking interactions, a maximum effective cross-linking distance (Supplementary Table S2) was used to approximate a flat harmonic potential by summation of Gaussian functions with distinct equilibrium positions (Supplementary Section S2).

A schematic of the entire process for merging coevolution and cross-linking signals as structure based models is depicted in Figure 1.

2.5 Folding simulations

Simulations of protein folding were performed using a modified version of Gromacs package with support for Gaussian-like potentials (Noel *et al.*, 2016, 2010; Lammert *et al.*, 2009). Each simulation was developed using an annealing protocol where the system temperature was reduced from 200 to 1 in steps of 100 ps. With this protocol, each folding simulations takes about 1 h of computing time using 4 CPUs (Intel Xeon E5-2670 v2 of 2.50 GHz) for proteins of medium range sizes (250 aa). We performed 1000 independent folding runs for each system to obtain statistically meaningful data on the accuracy of the folding protocol (Fig. 1). Also, the ensemble of models obtained allows for the use of clustering methods for final model evaluation (see Section 2.5). The final template modeling (TM)-score (Xu and Zhang, 2010; Zhang and Skolnick, 2004) and root-mean square deviation (RMSD) values were computed using LovoAlign (Martínez *et al.*, 2007) considering the last frame of each simulation and the reference crystallographic models. All-atom models of folded conformations were constructed with REconstruct atomic MOdel from reduced representation (REMO, Li and Zhang, 2009) from the final C_α coordinates obtained from Molecular Dynamics (MD) simulations with no further refinement.

2.6 Blind selection of correct folds

The modeling performed here generated sets of 1000 models for each target. Therefore, we can explore the properties of the ensemble to classify models, using consensus methods (Kryshtafovych *et al.*, 2014). Here, we opt to use a blind selection of folding conformations from decoy ensembles consisting of evaluating the average similarity of each model to all other models of the ensemble. This classificator is known as the 'Davis-QA-consensus' method (Kryshtafovych *et al.*, 2014). All-on-all structural alignments of the models were performed within each ensemble using LovoAlign (Martínez *et al.*, 2007) and the average TM-score was computed for each model.

3 Results

3.1 Coevolutionary and XL/MS signals contribute synergistically to folding prediction

We investigate the contribution of coevolutionary and cross-linking signals on prediction accuracy of the native state of full proteins using SBMs. As an initial case of study, we performed *ab initio* structure predictions of SalBIII protein from *Streptomyces albus* and compared with its respective full X-ray structure (PDB ID: 5CXO - chain B), for which experimental cross-linking constraints were obtained recently by Gozzo and co-workers (unpublished data). Coevolutionary constraints were inferred for SalBIII using DCA and the MSA for its family (SnoaL-like domain; Table 1). Top *L* pairs used for simulations are shown as blue dots in Figure 2. Comparison with monomeric contacts from an X-ray crystal shows substantial agreement of DCA within SalBIII assigned domain. Experimental cross-linking constraints (black circles, Fig. 2) also agree very well with the monomeric X-ray map and are mainly found outside the regions covered by DCA, providing distinct and complementary contact information from that obtained by coevolution. A possible reason to this complementarity is the fact that highly coevolved couplings are originated from interactions that are crucial to preserve minimal function and are usually located in the deep core of macromolecular structure. On the other hand, chemical cross-linking reactions are limited to exposed amino acids and can only account for contacts within surface vicinity. Therefore, important structural information from both sources can be used to get a refined description of interaction patterns.

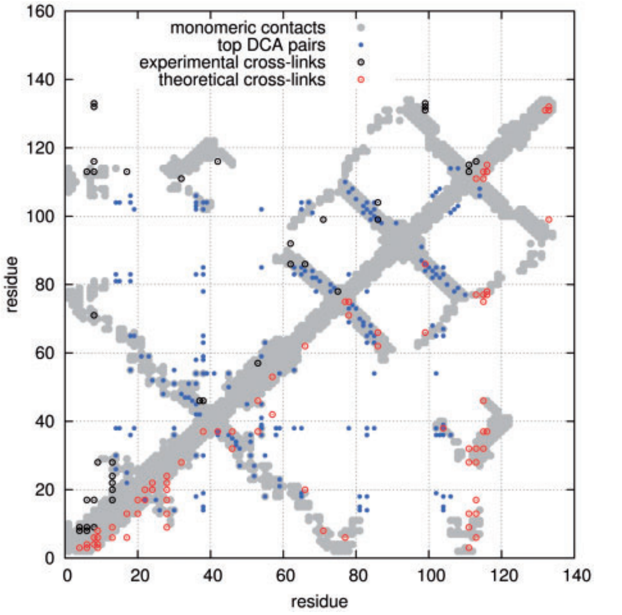

Fig. 2. Pairwise interactions used for prediction of SalBIII structure. Comparison of distinct residue–residue distance restrictions obtained from coevolution analysis (blue dots) and experimental and theoretical cross-linking/MS signals (black and red circles, respectively) with monomeric contacts of SalBIII X-ray structure. Physical contacts were computed considering C_α pairs within a distance of 10 Å

In order to evaluate the extent to which cross-linking data can contribute to increase folding accuracy of SalBIII, we also considered an ideal cross-linking experiment corresponding to the set of all possible cross-links that can be expected from the set of linkers used and crystallographic models (red dots, Fig. 2, see Section 2 for details). Consideration of this idealized experiment confirmed the low overlap with evolutionary couplings, evidencing that both techniques can provide unique structural information, which can be utilized to increase prediction accuracy of any computational protocol.

Figure 3 shows the distributions of TM-scores for 1000 simulations in ensembles considering each set of distance restrictions. As expected, simulations using only secondary structure information and no distance restrictions resulted in TM-score values below 0.2 (distributions in grey, Fig. 3 and Table 2), corresponding to random conformations.

When considering only predicted coevolutionary contacts (Fig. 3, blue distribution) simulations were able to reach folded models with TM-scores relative to the crystallographic model greater than 0.5, meaning that the overall correct fold was obtained (Zhang and Skolnick, 2004), however with low frequency (Table 2). When using exclusively experimental cross-linking signals (38 pairs, see Section 2), an improvement in the TM-score distribution is obtained relative to unrestrained simulations (Fig. 3A, green distributions). Therefore, although the restrictions are quite broad in terms of equilibrium distances, they contribute with meaningful information to structure prediction. Nevertheless, no models with proper folds were obtained.

When considering all possible cross-linking signals (74 pairs), cross-linking constraints provided and additional shift of the ensemble towards higher TM-scores but were still insufficient to achieve fold-level predictions for this system (Fig. 3B, green distributions

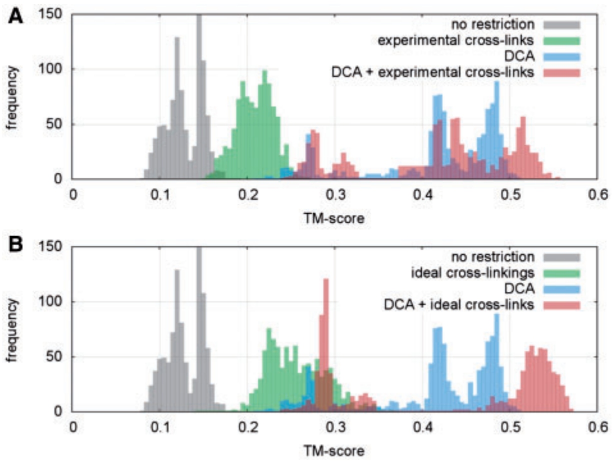

Fig. 3. Contributions of distinct distance constraints to folding prediction. Comparison of TM-score distributions for folding of SalBIII protein using coevolution signals and (**A**) experimental cross-links or (**B**) theoretical cross-links as interaction data. SalBIII X-ray structure was used as reference model (PDB: 5CXO)

Table 2. Comparison of SalBIII folding prediction using distinct distance constraints

Restriction	None	DCA	eXL	iXL	DCA + eXL	DCA + iXL
Best TM-score	0.19	0.51	0.27	0.37	0.56	0.57
%TM-score > 0.5	0	1.3	0	0	26.0	51.9

Note. DCA: coevolution signals obtained from direct-coupling analysis; eXL: experimental cross-linking/MS signals; iXL: ideal theoretical cross-linking/MS signals based on available X-ray models.

and Table 2). Therefore, in the context of structure prediction using SBMs, the cross-linking distance restraints appear not to be precise enough to obtain correctly folded structures. This is expected given that neither the SBMs (which in this case do not carry any a priori information on the folded structure) nor the cross-links provide precise residue distance information.

Finally, when we integrate both interactions signals from coevolution and cross-linking, we observe an overall and significant improvement of folding prediction (Table 2). The joint use of experimental cross-links and DCA constraints promoted an increase of 10% in the TM-score of the best predicted model and, most importantly, a 20-fold enhancement of population of models displaying the overall correct fold (Table 2). When considering every possible cross-linking pair along with DCA predictions, we observed an increase of 12% in the TM-score of the best prediction, with an improvement of approximately 40 times on the frequency of simulations reaching the correct fold.

These results motivate the notion that information obtained from coevolution and cross-linking are complementary and can be synergistically applied to increase accuracy and the rate of success in current structure prediction methods.

3.2 Fold of proteins with diverse topologies

We applied the proposed methodology (Fig. 1) to a set of systems with diverse topologies (Table 1). In these cases, cross-linking constraints were determined computationally by using the Topolink package and a random subset of theoretical restrictions (tXL) was utilized to represent the average number of links obtained in XL/MS experiments (see Section 2). Predicted DCA couplings and cross-linking/MS signals for each system are shown in Supplementary Figure S3. A combination of constraints obtained from distinct methodologies (DCA or theoretical cross-linking) promoted substantial improvement in prediction accuracy in all systems selected in this study, when compared with predictions based solely in unique sources of pairwise distance restrictions (only DCA or cross-linking restrictions, Table 3).

Table 3. Combination of coevolutionary and cross-linking distances restrictions to predict folding of diverse protein systems

System		Distance restriction			
		None	tXL	DCA	DCA + tXL
DesR	Best TM-score	0.19	0.38	0.56	0.60
	%TM-score > 0.5	0	0	69.7	83.5
RAP2A	Best TM-score	0.16	0.50	0.68	0.72
	%TM-score > 0.5	0	0	80.5	75.9
	%TM-score > 0.6	0	0	53.4	75.6
Rhodopsin	Best TM-score	0.20	0.35	0.60	0.62
	%TM-score > 0.5	0	0	79.5	81.3
	%TM-score > 0.6	0	0	0	12.6
Abl kinase	Best TM-score	0.15	0.47	0.58	0.64
	%TM-score > 0.5	0	0	58.2	58.7
	%TM-score > 0.6	0	0	0	40.8
Creatine kinase	Best TM-score	0.14	0.50	0.36	0.59
	%TM-score > 0.5	0	0	0	12.1

Note. DCA: coevolutionary signals obtained from direct-coupling analysis. tXL: theoretical cross-linking/MS signals based on available X-ray models.

Coevolutionary pairs obtained from DCA integrated in SBM potentials showed fold-level accuracy in all systems selected containing single families, with a considerable higher statistics (~80% for RAP2A and Rhodopsin, Table 3 and Supplementary Fig. S4). Despite cross-link signals only provide upper-limit distances for residue pair interactions, in some cases, the constraints predicted (tXL) were sufficient to drive the simulations towards correct folds (RAP2A, Abl Kinase and Creatine Kinase, Table 3 and Supplementary Fig. S4). These results provide evidence that cross-linking data can improve conformational search and folding predictions and validate this approach as an efficient methodology to assist protein structural characterization.

An special case considered in this study is creatine kinase. This protein contains two distinct conserved domains (Pfam families: PF00217 and PF02807) with limited sequence data, hindering the application of coevolution for structural characterization. As shown in Table 3 and Supplementary Figure S4, using only distance constraints from DCA was insufficient to recover protein native conformations. This same limitation was observed when using only cross-linking restrictions, although a small fraction of predicted conformations achieved near fold-level accuracy (Supplementary Fig. S4). For this case, the combination of DCA and cross-linking signals showed to be crucial to improve the prediction into fold-level accuracy (TM-score > 0.5). This is a representative case of the types of systems where the proposed methodology would be more beneficial.

3.3 Blind selection of native folds

Discrimination of protein native state from folding decoys is a difficult problem (Park *et al.*, 1997). This problem is even more challenging when *ab initio* predictions provide large decoy ensembles with a plethora of possible folding architectures (Brodie *et al.*, 2017; Cooper *et al.*, 2010; Kosciolek and Jones, 2014; Rohl *et al.*, 2004). Even though there has been a significant refinement improvement in theoretical models describing physical interactions, successful identification of correct folded states based solely in energy functions is rare (Deng *et al.*, 2016; Mishra *et al.*, 2016; Mirny and Shakhnovich, 1996; Sankar *et al.*, 2017; Uziela *et al.*, 2017; Zhou *et al.*, 2014a, 2014b). Recent progress has been achieved using alternative approaches such as entropy estimation and machine learning methods (Sankar *et al.*, 2017; Uziela *et al.*, 2017).

Since we generate an ensemble of 1000 models for each system, we chose to use a consensus method to classify the models (Kryshtafovych *et al.*, 2014). We employed the Davis-QA-consensus classification method for each ensemble of models predicted using DCA and cross-linking/MS data (Table 3). The models with greatest average similarity to all other models in the ensemble were selected. This evaluation allowed to successfully identify the models in the upper limit of TM-score predictions (Fig. 4). Therefore, clustering by similarity resulted to be an effective method for quality assessment of models generated using the current protocol.

4 Discussion

In this work, we provide an effective, computationally inexpensive and robust methodology to predict protein folds using residue–residue couplings from coevolution and cross-linking/MS data integrated with structure based models. We performed a systematic study of the role of each signal component in structure prediction performance for a diverse set of protein topologies. We observe a synergism between coevolutionary and cross-linking restrictions, each contributing with distinct and unique structural information that led to an increase of folding prediction accuracy. While coevolution couplings are usually prevalent in the core of protein structures by key intermolecular contacts that promotes packing, cross-linking reactions are restricted to protein surface due to physical accessibility. Therefore, both components contribute with important information to solve tertiary structure. This is particularly true for the challenging coevolution cases where sequence availability is limited or the domain coverage is insufficient.

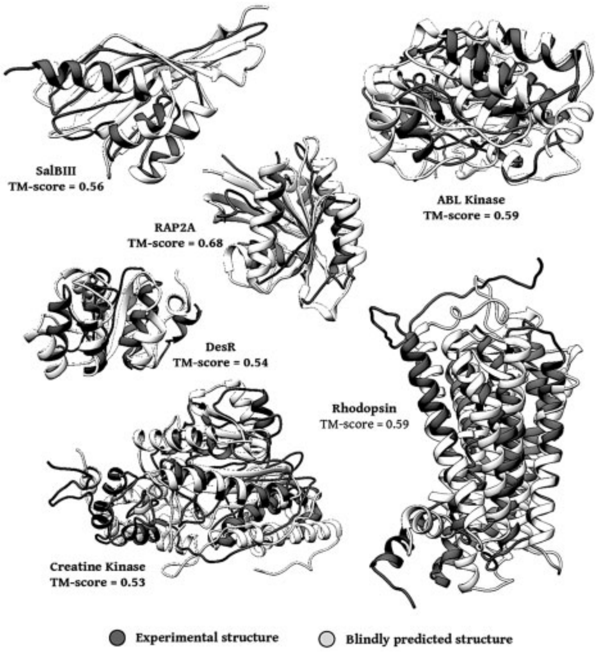

Fig. 4. Blind selection of predicted folding using structural similarity analysis. Folded models with TM-scores ≥ 0.5 for all systems considered in this study were identified. Dark gray structures depict blindly selected predictions compared to the experimentally determined structures in light gray. TM-score$_C$: predicted model selected by consensus score analysis. TM-score$_B$: best model generated, with highest similarity to crystallographic model

Molecular coevolution has recently been established as a significant technique to infer protein interactions and assist structural elucidation. On the other side, the use of experimental cross-linking/MS with long-range linkers as a unique source of interactions is usually insufficient for fine molecular description such as needed for folding prediction (Tamò *et al.*, 2017). Interestingly, although cross-linking signals constitute non-precise distance restrictions, they can provide enough information to allow folding elucidation when substantial data are available (as shown for DesR, RAP2A and Abl and creatine kinases). This observation also suggests that improvement in equipment sensibility for cross-linking/MS signals should boost structural elucidation over the next years.

Finally, we demonstrate how folding ensembles can be used to identify plausible functional conformations by applying a self-consistent similarity analysis. The described methodology can be easily applied to practical problems in structural biology using the protocol and scripts developed in this work and available for others to use. We expect that this approach that integrates and maximizes computational and experimental methodologies will be useful to elucidate important challenges in structural bioinformatics.

Funding

This work was supported by São Paulo Research Foundation [FAPESP, grants: 2015/13667-9, 2010/16947-9, 2013/05475-7, 2013/08293-7; 2016/13195-2; 2014/17264-3] and funds from the School of Natural Sciences and Mathematics at the University of Texas at Dallas.

Conflict of Interest: none declared.

References

Alberts,B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**, 291–294.

Alberts,B. *et al.* (2014) *Molecular Biology of the Cell*. 2nd edn. Garland Science, New York, NY.

Anfinsen,C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

Baker,D. (2014) Centenary award and Sir Frederick gowland hopkins memorial lecture. Protein folding, structure prediction and design. *Biochem. Soc. Trans.*, **42**, 225–229.

Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.

Bender,B.J. *et al.* (2016) Protocols for molecular modeling with Rosetta3 and RosettaScripts. *Biochemistry*, **55**, 4748–4763.

Brodie,N.I. *et al.* (2017) Solving protein structures using short-distance cross--linking constraints as a guide for discrete molecular dynamics simulations. *Sci. Adv.*, **3**, e1700479.

Bryngelson,J.D. *et al.* (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, **21**, 167–195.

Chen,M. *et al.* (2016) Protein folding and structure prediction from the ground up: the atomistic associative memory, water mediated, structure and energy model. *J. Phys. Chem. B*, **120**, 8557–8565.

Cherfils,J. *et al.* (1997) Crystal structures of the small G protein Rap2A in complex with its substrate GTP, with GDP and with GTPgammaS. *EMBO J.*, **16**, 5582–5591.

Clementi,C. *et al.* (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.*, **298**, 937–953.

Cooper,S. *et al.* (2010) Predicting protein structures with a multiplayer online game. *Nature*, **466**, 756–760.

Davtyan,A. *et al.* (2012) AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B*, **116**, 8494–8503.

De Leonardis,E. *et al.* (2015) Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res.*, **43**, 10444–10455.

Deng,H. *et al.* (2016) 3DRobot: automated generation of diverse and well-packed protein structure decoys. *Bioinformatics*, **32**, 378–387.

Dill,K.A. *et al.* (2008) The protein folding problem. *Annu. Rev. Biophys.*, **37**, 289–316.

Dill,K.A. and MacCallum,J.L. (2012) The protein-folding problem, 50 years on. *Science*, **338**, 1042–1046.

Dobson,C.M. (2003) Protein folding and misfolding. *Nature*, **426**, 884–890.

Dobson,C.M. *et al.* (1998) Protein folding: a perspective from theory and experiment. *Angew. Chem. Int. Ed.*, **37**, 868–893.

Drozdetskiy,A. *et al.* (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.*, **43**, W389–W394.

Freddolino,P.L. *et al.* (2010) Challenges in protein-folding simulations. *Nat. Phys.*, **6**, 751–758.

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

Hayat,S. *et al.* (2015) All-atom 3D structure prediction of transmembrane *β*-barrel proteins from sequences. *Proc. Natl. Acad. Sci. U.S.A*, **112**, 5413–5418.

Hofmann,T. *et al.* (2015) Protein structure prediction guided by crosslinking restraints—a systematic evaluation of the impact of the crosslinking spacer length. *Methods*, **89**, 79–90.

Honig,B. (1999) Protein folding: from the levinthal paradox to structure prediction. *J. Mol. Biol.*, **293**, 283–293.

Hopf,T.A. *et al.* (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**, 1607–1621.

Hopf,T.A. *et al.* (2015) Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nat. Commun.*, **6**, 6077.

Jin Lee,Y. (2008) Mass spectrometric analysis of cross-linking sites for the structure of proteins and protein complexes. *Mol. Biosyst.*, **4**, 816–823.

de Juan,D. *et al.* (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–261.

Kamisetty,H. *et al.* (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U.S.A*, **110**, 15674–15679.

Kelley,L.A. *et al.* (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.*, **10**, 845–858.

Kosciolek,T. and Jones,D.T. (2014) De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS One*, **9**, e92197.

Kryshtafovych,A. *et al.* (2014) Assessment of the assessment: evaluation of the model quality estimates in CASP10. *Proteins*, **82(Suppl 2)**, 112–126.

Lammert,H. *et al.* (2009) Robustness and generalization of structure-based models for protein folding and function. *Proteins*, **77**, 881–891.

Liu,F. *et al.* (2015) Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat. Methods*, **12**, 1179–1184.

Li,Y. and Zhang,Y. (2009) REMO: a new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins: Struct. Funct. Bioinf.*, **76**, 665–676.

Luhavaya,H. *et al.* (2015) Enzymology of pyran ring A formation in salinomycin biosynthesis. *Angew. Chem. Int. Ed Engl.*, **127**, 13826–13829.

Marks,D.S. *et al.* (2012) Protein structure prediction from sequence variation. *Nat. Biotechnol.*, **30**, 1072–1080.

Martínez,L. *et al.* (2007) Convergent algorithms for protein structural alignment. *BMC Bioinformatics*, **8**, 306.

Martinez,L. *et al.* (2017) TopoLink: a software to validate structural models using chemical crosslinking constraints. *Protoc. Exchange*, DOI: 10.1038/protex.2017.035.

Matysiak,S. and Clementi,C. (2004) Optimal combination of theory and experiment for the characterization of the protein folding landscape of S6: how far can a minimalist model go? *J. Mol. Biol.*, **343**, 235–248.

Michel,M. *et al.* (2017) Predicting accurate contacts in thousands of Pfam domain families using PconsC3. *Bioinformatics*, **33**, 2859–2866.

Mirny,L.A. and Shakhnovich,E.I. (1996) How to derive a protein folding potential? a new approach to an old problem. *J. Mol. Biol.*, **264**, 1164–1179.

Mishra,A. *et al.* (2016) Discriminate protein decoys from native by using a scoring function based on ubiquitous Phi and Psi angles computed for all atom. *J. Theor. Biol.*, **398**, 112–121.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A*, **108**, E1293–E1301.

Morcos,F. *et al.* (2013) Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci. U.S A*, **110**, 20533–20538.

Morcos,F. *et al.* (2014) Direct coupling analysis for protein contact prediction. *Methods Mol. Biol.*, **1137**, 55–70.

Nguyen-Huynh,N.-T. *et al.* (2015) Chemical cross-linking and mass spectrometry to determine the subunit interaction network in a recombinant human SAGA HAT subcomplex. *Protein Sci.*, **24**, 1232–1246.

Noel,J.K. *et al.* (2010) SMOG@ctbp: simplified deployment of structure-based models in GROMACS. *Nucleic Acids Res.*, **38**, W657–W661.

Noel,J.K. *et al.* (2016) SMOG 2: a versatile software package for generating structure-based models. *PLoS Comput. Biol.*, **12**, e1004794.

Noel,J.K. and Onuchic,J.N. (2012) The many faces of structure-based potentials: from protein folding landscapes to structural characterization of complex biomolecules. In: Dokholyan,N. (ed.) *Biological and Medical Physics, Biomedical Engineering*. Springer, Boston, MA, pp. 31–54.

Ohren,J.F. *et al.* (2007) Structural asymmetry and intersubunit communication in muscle creatine kinase. *Acta Crystallogr. D Biol. Crystallogr*, **63**, 381–389.

Onuchic,J.N. *et al.* (1997) Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.*, **48**, 545–600.

Onuchic,J.N. and Wolynes,P.G. (2004) Theory of protein folding. *Curr. Opin. Struct. Biol.*, **14**, 70–75.

Ovchinnikov,S. *et al.* (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife*, **3**, e02030.

Ovchinnikov,S. *et al.* (2017) Protein structure determination using metagenome sequence data. *Science*, **355**, 294–298.

Paramelle,D. *et al.* (2013) Chemical cross-linkers for protein structure studies by mass spectrometry. *Proteomics*, **13**, 438–456.

Park,B.H. *et al.* (1997) Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.*, **266**, 831–846.

Pereira,M.B.M. *et al.* (2014) αB-crystallin interacts with and prevents stress-activated proteolysis of focal adhesion kinase by calpain in cardiomyocytes. *Nat. Commun.*, **5**, 5159.

Petrotchenko,E.V. *et al.* (2014) Analysis of protein structure by cross-linking combined with mass spectrometry. *Methods Mol. Biol.*, **1156**, 447–463.

Piana,S. *et al.* (2014) Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.*, **24**, 98–105.

Piccolino,M. (2000) Biological machines: from mills to molecules. *Nat. Rev. Mol. Cell Biol.*, **1**, 149–153.

Roche,D.B. and McGuffin,L.J. (2016) Toolbox for protein structure prediction. *Methods Mol. Biol.*, **1369**, 363–377.

Rohl,C.A. *et al.* (2004) Protein structure prediction using Rosetta. *Methods Enzymol.*, **383**, 66–93.

Roy,A. *et al.* (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.*, **5**, 725–738.

Sankar,K. *et al.* (2017) Knowledge-based entropies improve the identification of native protein structures. *Proc. Natl. Acad. Sci. U.S.A*, **114**, 2928–2933.

Santos,A.M. *et al.* (2011) FERM domain interaction with myosin negatively regulates FAK in cardiomyocyte hypertrophy. *Nat. Chem. Biol.*, **8**, 102–110.

dos Santos,R.N. *et al.* (2015) Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci. Rep.*, **5**, 13652.

Schug,A. *et al*. (2009) High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc. Natl. Acad. Sci. U.S.A*, **106**, 22124–22129.

Sfriso,P. *et al*. (2016) Residues coevolution guides the systematic identification of alternative functional conformations in proteins. *Structure*, **24**, 116–126.

Shindyalov,I.N. *et al*. (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng*., **7**, 349–358.

Sinz,A. (2006) Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrom. Rev*., **25**, 663–682.

Sinz,A. *et al*. (2015) Chemical cross-linking and native mass spectrometry: a fruitful combination for structural biology. *Protein Sci*., **24**, 1193–1209.

Sirovetz,B.J. *et al*. (2017) Protein structure prediction: making AWSEM AWSEM-ER by adding evolutionary restraints. *Proteins*, **85**, 2127–2142.

Stenkamp,R.E. (2008) Alternative models for two crystal structures of bovine rhodopsin. *Acta Crystallogr. D Biol. Crystallogr*., **64**, 902–904.

Sułkowska,J.I. *et al*. (2012) Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. U.S.A*, **109**, 10340–10345.

Sutto,L. *et al*. (2015) From residue coevolution to protein conformational ensembles and functional dynamics. *Proc. Natl. Acad. Sci. U.S.A*, **112**, 13567–13572.

Tamò,G.E. *et al*. (2017) Assessment of data-assisted prediction by inclusion of crosslinking/mass-spectrometry and small angle X-ray scattering data in the 12th Critical Assessment of protein Structure Prediction experiment. *Proteins: Struct. Funct. Bioinf*., **86(Suppl 1)**, 215–227.

Taylor,W.R. *et al*. (2013) Prediction of contacts from correlated sequence substitutions. *Curr. Opin. Struct. Biol*, **23**, 473–479.

Taylor,W.R. and Hamilton,R.S. (2017) Exploring RNA conformational space under sparse distance restraints. *Sci. Rep*., **7**, 44074.

Trajtenberg,F. *et al*. (2014) Allosteric activation of bacterial response regulators: the role of the cognate histidine kinase beyond phosphorylation. *mBio*, **5**, e02105-14.

Uziela,K. *et al*. (2017) ProQ3D: improved model quality assessments using deep learning. *Bioinformatics*, **33**, 1578–1580.

Webster,D.M. (2000) *Protein Structure Prediction: Methods and Protocols*. Humana Press, Totowa, NJ.

Weigt,M. *et al*. (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A*, **106**, 67–72.

Weinreb,C. *et al*. (2016) 3D RNA and functional interactions from evolutionary couplings. *Cell*, **165**, 963–975.

Whitford,P.C. *et al*. (2009) An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins*, **75**, 430–441.

Wolynes,P.G. *et al*. (1995) Navigating the folding routes. *Science*, **267**, 1619–1620.

Xu,J. and Zhang,Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, **26**, 889–895.

Yang,J. *et al*. (2015) The I-TASSER Suite: protein structure and function prediction. *Nat. Methods*, **12**, 7–8.

Young,M.M. *et al*. (2000) High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl. Acad. Sci*., **97**, 5802–5806.

Zhang,J. *et al*. (2010) Targeting Bcr-Abl by combining allosteric with ATP-binding-site inhibitors. *Nature*, **463**, 501–506.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zhou,J. *et al*. (2014a) Amino acid network for the discrimination of native protein structures from decoys. *Curr. Protein Pept. Sci*., **15**, 522–528.

Zhou,J. *et al*. (2014b) SVR_CAF: an integrated score function for detecting native protein structures among decoys. *Proteins*, **82**, 556–564.