Received: 4 July 2019 | Revised: 7 October 2019 | Accepted: 3 November 2019

DOI: 10.1002/prot.25843

RESEARCH ARTICLE

PROTEINS WILEY

Structural complementarity of distance constraints obtained from chemical cross-linking and amino acid coevolution

Ricardo N. dos Santos[1,2] | Guilherme F. Bottino[1,2] | Fábio C. Gozzo[1] | Faruck Morcos[3,4] | Leandro Martínez[1,2]

[1]Institute of Chemistry, University of Campinas, Campinas, São Paulo, Brazil

[2]Center for Computing in Engineering & Sciences, University of Campinas, Campinas, São Paulo, Brazil

[3]Department of Biological Sciences, University of Texas at Dallas, Richardson, Texas

[4]Department of Bioengineering, University of Texas at Dallas, Richardson, Texas

Correspondence
Leandro Martínez, Institute of Chemistry and Center for Computing in Engineering & Science, University of Campinas, Campinas, SP, Brazil.
Email: leandro@iqm.unicamp.br

Funding information
Fundação de Amparo à Pesquisa do Estado de São Paulo, Grant/Award Numbers: 2010/16947-9, 2013/08293-7, 2014/17264-3, 2015/13667-9, 2018/14274-9, 2018/24293-0; School of Natural Sciences and Mathematics at the University of Texas at Dallas

Peer Review
The peer review history for this article is available at https://publons.com/publon/10.1002/prot.25843.

Abstract

The analysis of amino acid coevolution has emerged as a practical method for protein structural modeling by providing structural contact information from alignments of amino acid sequences. In parallel, chemical cross-linking/mass spectrometry (XLMS) has gained attention as a universally applicable method for obtaining low-resolution distance constraints to model the quaternary arrangements of proteins, and more recently even protein tertiary structures. Here, we show that the structural information obtained by XLMS and coevolutionary analysis are effectively complementary: the distance constraints obtained by each method are almost exclusively associated with non-coincident pairs of residues, and modeling results obtained by the combination of both sets are improved relative to considering the same total number of constraints of a single type. The structural rationale behind the complementarity of the distance constraints is discussed and illustrated for a representative set of proteins with different sizes and folds.

KEYWORDS
coevolution, cross-linking, DCA, folding, mass-spectrometry

1 | INTRODUCTION

Protein structural modeling research has successfully explored distance constraints obtained from the analysis of amino acid coevolution.[1-8] For the first time, clearly successful strategies to model proteins without using structural information from analogous templates were developed to the point that structures with near-atomic accuracy can be obtained.[2-4,9,10] At the same time, the structural information that can be derived from amino acid coevolution is limited by the availability of a large set of evolutionarily related protein sequences. Contact information can be obtained only for evolutionarily conserved domains. Also, because of the nature of coevolutionary data, it is intrinsically difficult to obtain structural constraints which are specific for a subfamily of proteins, and even less so for the fold of a protein variant of an individual organism.

Concurrently, chemical cross-linking/mass spectrometry (XLMS) is an experimental strategy that is gaining momentum in the structural biology community.[11-14] It relies on the exposure of protein structures to reactants (the cross-linkers) which react to pairs of residues, forming cross-links. The identification of the amino acid pairs cross-linked is performed with high-resolution mass spectrometry, providing a distance constraint that can be used to aid protein structure modeling.[15] This strategy has been successful to obtain the quaternary arrangement of protein complexes[12,16-18] and attempts to obtain the tertiary structure of proteins have been reported.[15,19,20] Nonetheless, the effectiveness of these methods still needs to be consolidated in a general way.[21-23] XLMS is a promising strategy for large-scale proteomic approaches, as the experimental procedures are relatively simple

Ricardo N. dos Santos and Guilherme F. Bottino contributed equally to this study.

when compared to high-resolution protein structure determination methods like X-ray crystallography, nuclear magnetic resonance, and cryoelectron microscopy.[24,25]

In this work, we explore the structural nature of the contact information obtained by XLMS in comparison with that obtained by amino acid coevolutionary analysis. We quantify the complementarity between the sets of contacts obtained by either method and find out that the contacts obtained by one of the methods are almost exclusively associated with pairs of residues not covered by the other method. We demonstrate that this complementarity is consistent over a variety of protein folds, and provide a structural rationale for the result. Additionally, we show that the combination of the contact information of both sets leads to optimal protein fold predictions relative to the use of one type of contact only, even for contact sets of identical size, demonstrating the practical implications of the information complementarity of these types of contacts.

2 | MATERIALS AND METHODS

2.1 | Target proteins and contact prediction

To compare the contact information derived from amino acid coevolution and cross-linking methods, we analyzed 15 X-ray crystallographic models from distinct families of monomeric proteins with distinct sizes and topological complexity (Table 1). The residue coevolution analysis was performed using the direct-coupling analysis (DCA)[4,5] implementation and multiple sequence alignments for family domains annotated in Pfam[41] (see Supporting Information). DCA contacts are classified according to direct information[4,5] (DI), and we take the N contact predictions with higher DI scores, where N is the number length of the Pfam protein domain (See Table S1). These top N predictions were shown to be typically in contact in the three-dimensional structure.[4] Chemical cross-links were theoretically predicted using TopoLink[42] considering the crystallographic models and the reactivity of three types of state-of-the-art linkers: 1,6-Hexanediamine linker which binds acidic side chains at both ends, the disuccinimidyl suberate (DSS) linker, which binds nonspecifically Lys and Ser residues, and a recently developed zero-length cross-linker, which induces the formation of direct bonds between pairs of acidic/basic side chains[43] (see Supporting Information Section S2 for further details). In all cases, only contacts between atoms belonging to residues distant on the primary sequence by more than four residues were considered in our analyses. Figure 1A portrays an example of a native contact map with highlights to DCA and XL constraints, and Figure 1B allows for visual appreciation of both constraint types superimposed on the crystallographic structure of a protein.

2.2 | Evaluation of contact overlap and contact structural properties

The overlap of the sets of contacts obtained by DCA or XL is reported by a mutual overlap (MO) coefficient, defined as the fraction of contacts of the smallest set (typically the XL set) that is coincident with the largest set. That is, if n_{XL} is the number of XL contacts and $n_{XL\cap DCA}$ is the number of contacts in the intersection of both sets,

$$MO = n_{XL\cap DCA}/n_{XL} \quad (1)$$

The MO between the XL and DCA sets is usually small, as we will show. The random probability that two sets of n_{XL} and n_{DCA} contacts have $n_{XL\cap DCA}$ contacts in common is dependent on the total number

TABLE 1 Structural models studied, contact information obtained by each method, set intersection and mutual correlations. n_{DCA} denotes the number of contacts obtained by Direct Coupling Analysis. n_{XL} is the number of potential cross-links predicted for each model. $n_{XL}\cap n_{DCA}$ is the number of contacts in both sets, and MO is the mutual overlap as defined in Equation (1). Probabilities of obtaining a random overlap smaller or equal the observed $n_{XL}\cap n_{DCA}$ are shown, for a pool of contacts with size based cutoff distance of 8 Å

PDB	Size	n_{DCA} (Pfam length)	n_{XL}	$n_{XL}\cap n_{DCA}$	MO	random $p(n \le n_{XL}\cap n_{DCA})$ δ = 8 Å
1G6X[26]	58	52	2	0	0.00	35.5%
1C75[27]	71	66	20	3	0.15	<0.1%
1D4T[28]	115	80	27	3	0.11	0.7%
1FK5[29]	93	86	14	1	0.07	<0.1%
1C52[30]	131	89	66	1	0.02	<0.1%
1EW4[31]	106	102	33	3	0.09	<0.1%
1B0B[32]	142	106	33	0	0.00	<0.1%
1D06[33]	130	112	17	2	0.12	<0.1%
1AMM[34]	174	162	28	1	0.04	0.1%
1G67[35]	450	184	97	5	0.05	<0.1%
1ARB[36]	263	194	38	0	0.00	<0.1%
1G8A[37]	227	221	104	2	0.02	<0.1%
1ATG[38]	231	226	41	3	0.07	<0.1%
1GCI[39]	269	232	45	3	0.07	<0.1%
1BXO[40]	323	306	63	7	0.11	<0.1%

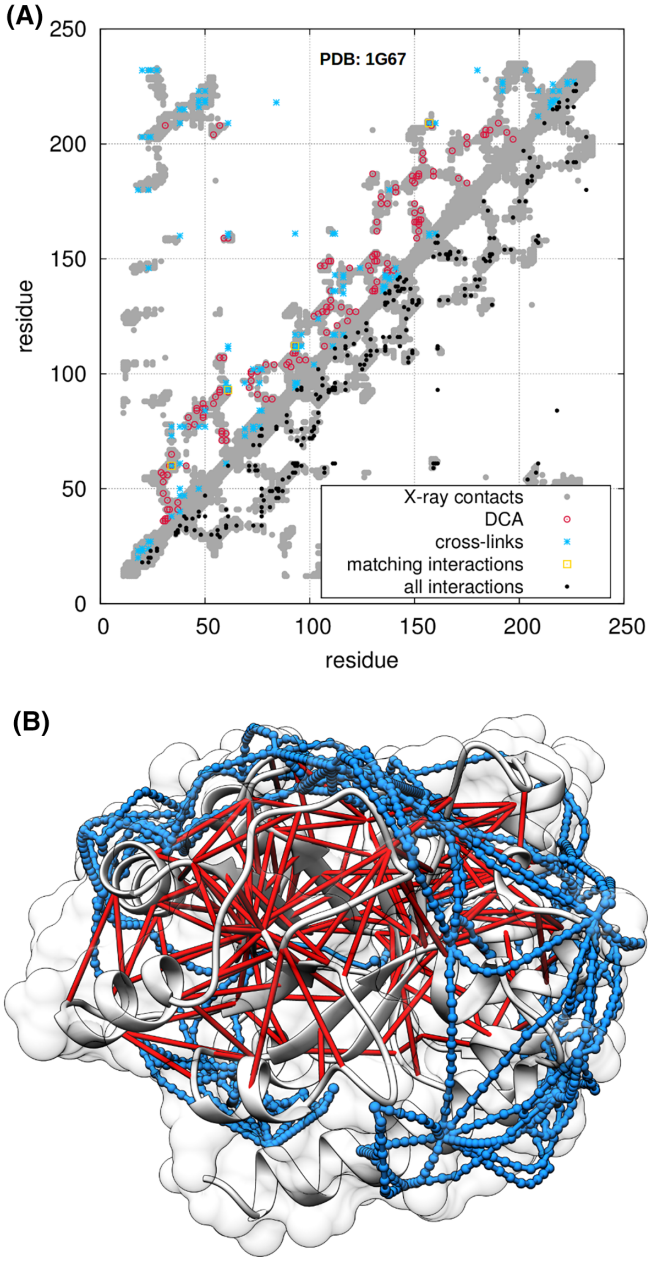

FIGURE 1 A, Contact map of the Thiamin Phosphate Synthase (1G67) compared with the distance constraints obtained by DCA (red) and potential XLs (blue). The few matching interactions are shown by yellow squares. B, Visual representation of XLs constraints (blue) and DCA constraints (red). The structural nature of each set leads to the almost complete orthogonality of the contact information. Similar results for all other proteins are shown in the SI

of contacts considered, n_T. We computed n_T for each protein as a function of the distance cutoff, δ, between Cα atoms, and simulated 1000 times the random sampling of n_{XL} and n_{DCA} contacts in the set to obtain the probability that the overlap of the sets is smaller or equal than that the observed overlap. In Table 1, we report this probability for δ = 8 Å, and in Figure 2A, we report the random MO as a function of δ. The number of contacts for δ = 8 Å is ∼2 times the number of residues of the protein, excluding all contacts between residues for which the distance in the sequence is smaller than five residues.

The distances between the pair Cα atoms of each contact, and the distance of each Cα atom to the surface of the protein were also computed to produce the plots in Figure 2C,D. The algorithm to compute the distance to the surface of the protein is described in Supporting Information S6.

2.3 | Protein modeling experiments

Modeling experiments were performed for targets 1C52, 1C75, and 1D06. 1C75 is the smallest protein of the set, except for 1G6X which we did not consider for this analysis because it has only two potential XLs. 1C52 and 1D06 were chosen for having similar sizes and number of DCA contacts, while having very different sets of XLs (66 and 17, respectively—Table 1). Models were generated with the Rosetta ab initio framework[44-49] (see Supporting Information S5), employing for all contacts quadratically bounded potentials centered on the true crystallographic distances between the residues' Cα atoms, with a tolerance of ±1 Å. Five experiments were performed, with (1) No constraints, and constraint sets comprised of (2) all n_{XL} XLs for each target, (3) randomly generated subsets of n_{XL} DCA constraints for each target, (4) randomly generated subsets of $2n_{XL}$ DCA contacts for each target, and (4) the combination of the n_{XL} XLs and randomly selected subsets of n_{XL} DCA contacts. In the case of 1C52, where $n_{DCA} < 2n_{XL}$ the $2n_{XL}$ DCA set was completed with the subsequent contacts predicted by the DCA calculations. Thousand models were generated in each experiment for each target. The quality of the models was evaluated with the TM-score relative to the crystallographic structure.[50] A TM-score greater than 0.5 indicates that the overall fold of the structure is correct.[51] The fraction of models obtained with a TM-score > 0.5 can be considered a measure of the overall success of a modeling experiment.[23]

3 | RESULTS AND DISCUSSION

A comparative example of the contact information obtained by DCA, cross-linking, and the contact matrix of the X-ray model of a protein structure is shown in Figure 1A. To build the figure, the native contact matrix was computed considering a pair of residues in contact if the distance between their Cα atoms was less than 10 Å. The figure shows Thiamin Phosphate Synthase, a 225-amino acid residue protein with a composition of 43% α-helical and 15% β-sheet structures (PDB id. 1G67).[35] DCA obtained 143 contacts compatible within 10 Å with the crystallographic contacts within the top 184 best-ranked predictions (where 184 is the number of residues of the conserved Pfam domain). At the same time, 97 potential XLs were predicted for the crystallographic structure. Visual inspection suggests that DCA contacts and XLs rarely involve the same pairs of residues, as shown in Figure 1B. In this case, only five contacts found by DCA associate pairs of residues that are within cross-linking distance.

The mutual overlap, MO, of the XL and DCA contacts of 1G67, is 0.05, implying that ∼5% of the XL contacts were predicted as from DCA. The probability that this small overlap is fortuitous is dependent

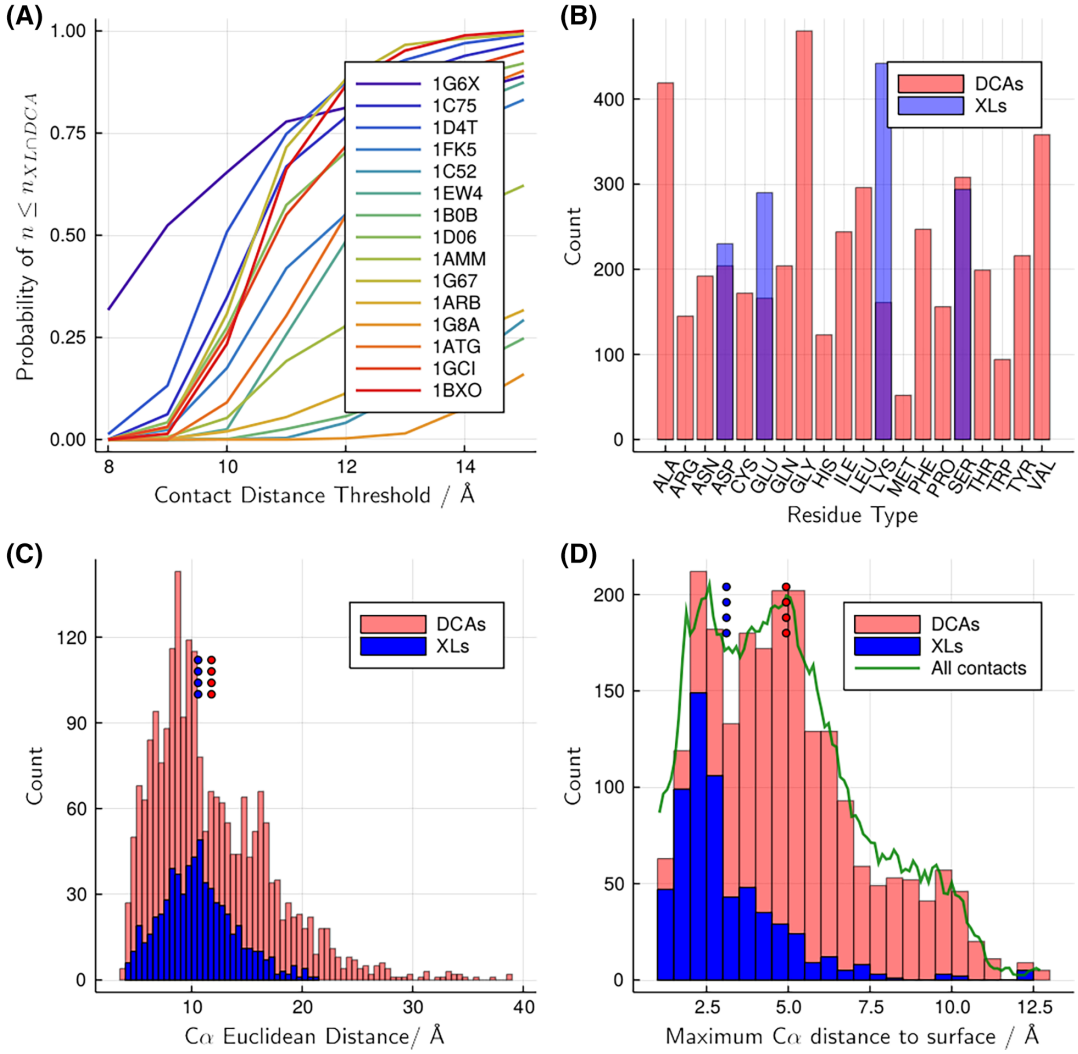

FIGURE 2 A, Probability of observing a random mutual overlap, MO, smaller or equal to the observed MO, as a function of the contact distance threshold that defines the number of contacts of the pool from which the random contacts are selected. B, Residue types involved in DCA and XL contacts. C, Distance distribution of Cα atoms of DCA and XL contacts. The distributions are different, but with similar averages. D, Distribution of the *maximum* distance to the protein surface of the Cα atoms of each pair of DCA, XL and all protein pairs of residues. At least one of the DCA residues is significantly buried in the protein, with a distribution very similar to that of all protein contacts. XLs, as expected, are concentrated on the protein surface [Color figure can be viewed at wileyonlinelibrary.com]

on the number of possible contacts that could be probed by the methods randomly. For a protein of *N* residues, this number is at most $(N^2–N)/2$, and this pool of contacts is so large that the probability of overlap of two sets of contacts of the order of *N* contacts is very small. Since this pool of contacts seems to be unrealistically large, we computed the number of contacts that each protein has assuming different distance cutoffs between the Cα atoms. Usually, two residues are said to be in contact if the distances between their Cα atoms is less than ~8 Å.[52] The number of contacts of a globular protein considering a cutoff distance of 8 Å, disconsidering residues vicinal in the primary structure by less than five positions, is about two times the number of residues. If a pool of 2 *N* contacts is considered for a random selection of n_{XL} and n_{DCA} contacts, the probability of an overlap equal or smaller than the observed ones is, now, very small (<0.1% in most cases), as shown in Table 1. This probability increases with the increase in the number of possible contacts, thus with the cutoff used to count them from the protein structure. For the 1G67 protein, in particular, this probability of a fortuitous lack of overlap increases to about 30% if the pool of possible contacts contains about six times the number of residues (for a cutoff of ~10 Å). For each of the proteins evaluated, the probability of fortuitous MO smaller than the observed overlap is shown in Figure 2A, as a function of the threshold defining the total number of contacts of the protein (see Supporting Information S4). For small cutoffs the probability of a fortuitous result is small, but for larger cutoff distances it becomes non-depreciable for each protein (the systematic observation of an event of 50% probability in 15 proteins is very small, nevertheless). Therefore, despite the fact that the overlaps of the sets are very small, it is not possible from simple combinatorial arguments to exclude completely the possibility that the contacts result from random sampling within the possible protein residue pairs. A

rationale behind the lack of overlap and the demonstration that it has practical implications is necessary. The visual appeal of the contacts in Figure 1B suggests that the complementarity of the sets has structural origins. Similar structural representation of DCA and XL contacts for all proteins studied are shown in section S7 of the Supporting Information.

The first obvious reason for the lack of overlap of the two sets of contacts is that the chemical nature of the residues involved in XL contacts is restricted by the linkers considered. Most commonly, linkers react with charged and polar residues. The linkers considered in this study can react with Lys, Ser, Glu, and Asp residues, and this is reflected in the distribution of residue types shown in Figure 2B. DCA contacts can, in principle, span any type of residue. Indeed, Figure 2B shows that the DCA contacts computed here are distributed among all residue types, perhaps with a particular preference towards amino acid residues with short side-chains as Ala, Val, and Gly. Therefore, the lack of overlap of the contact sets is in part trivially explained by the nature of the residues involved, yet this would not have any impact on the use of these contacts for modeling purposes.

Most importantly, the residues involved in DCA or XL contacts are distributed differently in the protein structures. First, as shown in

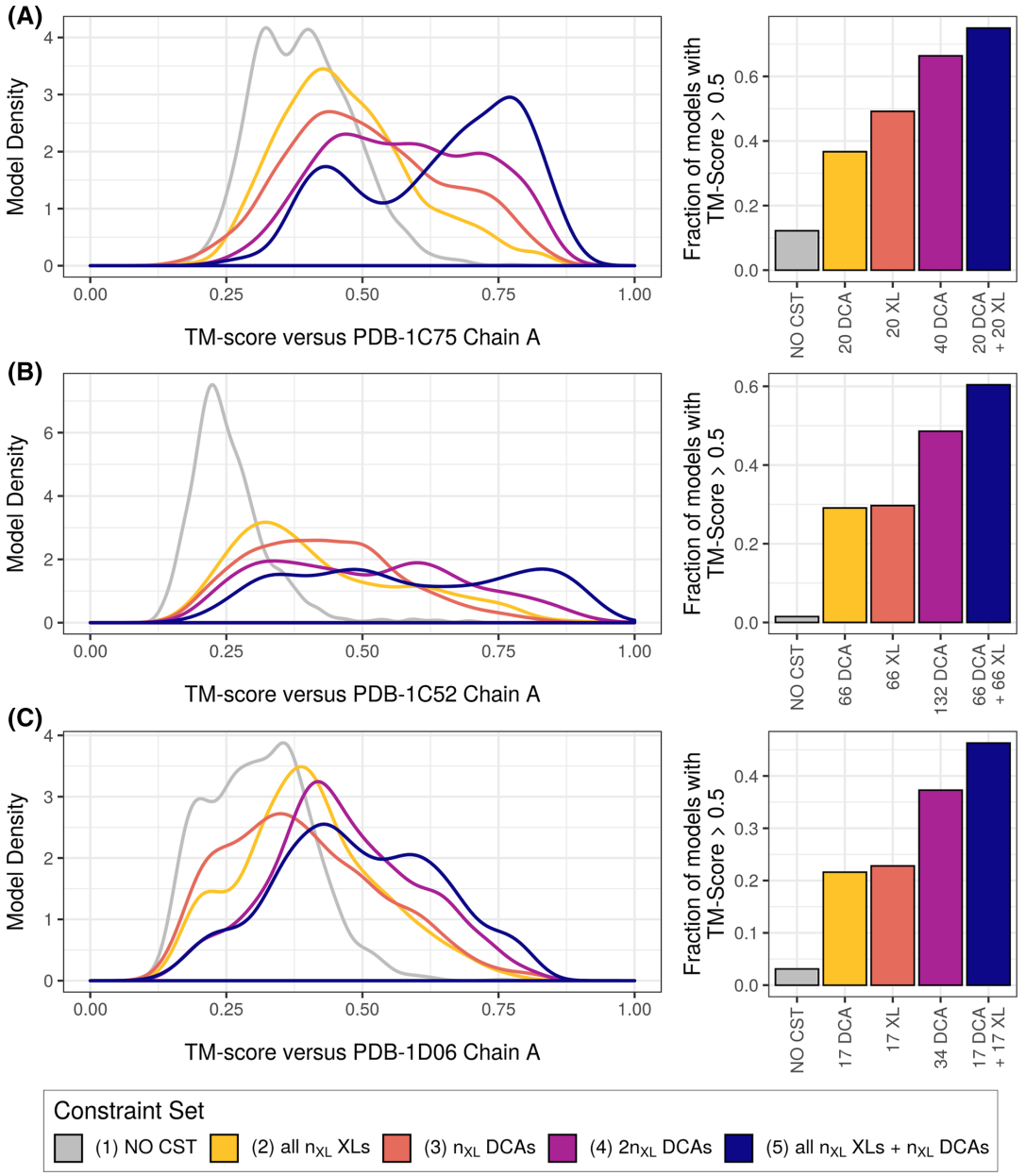

FIGURE 3 Model quality assessment for the modeling experiments performed. Distribution of output model quality determined by the TM-Score of the alignment of each model against the crystallographic structure, and fraction of successful (TM-Score > 0.5) models for targets 1C75 (A), 1C52 (B), and 1D06 (C). Quality distributions illustrate successive elongations on the model quality peak since the unconstrained modeling case all the way to the combination of XLs and DCA constraints, with successive increases of the population over 0.5 TM-Score. Successful model fraction plots clearly show that the bar related to the combination of XLs and DCAs is roughly double the size of the DCA-only case, whereas the *2 N* DCA performance is slightly more modest in all cases [Color figure can be viewed at wileyonlinelibrary.com]

Figure 2C, the distance distribution between Cα atoms of the contacts is different. DCA contacts involve pairs of residues for which the distances appear to follow a binomial distribution, with distinct peaks at ∼9 Å and a shoulder at ∼15 Å. The distance distribution of XL contacts displays a single peak at ∼12 Å. The distribution averages are shown by dots in Figure 2C and are surprisingly similar, being 11.8 Å for DCA contacts and 10.6 Å for XL contacts. Nevertheless, the sole facts that the distributions are different imply an additional factor for the non-overlapping of the sets.

The most clear physical rationale for the complementarity of DCA and XL contacts is, however, the fact that XL contacts involve residues that must lay on the protein surface. Figure 1B and the corresponding figures for all the other proteins (Supporting Information Section S7) are clearly suggestive that DCA contacts, on their turn, involve residues at the protein core. We computed the distance of all the Cα atoms to the surface of the protein (Figure S1). Each contact involves two residues, and we are interested in the greatest distance of the Cα atoms of these two residues to the surface, indicating the penetration of the contact in the protein core. The distributions of the maximum distances to the surface are shown in Figure 2D. XL contacts almost invariably involve residues for which the corresponding Cα atoms are closer than 3 Å from the protein surface. DCA contacts display again a bimodal distribution, with a peak at ∼2.5 Å and a second peak at ∼5 Å. The distribution of maximum distances to the surface between the residues of all protein contacts is shown in green and, interestingly, it is very similar to that of DCA contacts. This means that the DCA is sampling residues with a mostly random distribution of positions in the protein in what concerns the distance of the residues to the surface. The distributions of XL and DCA maximum Cα distances to the surfaces are, clearly, very different, and this is the most relevant structural rationale behind the small overlap between these contact sets.

If the structural nature is different for DCA and XL contacts, the information provided by each set in a modeling experiment may be complementary. To test this hypothesis, we modeled the structures of three of the proteins studied here with different constraints sets. For these experiments, all contacts were considered to be correct, and the harmonic constraining potentials were introduced centered at the distances associated to each pair of Cα atoms on the crystallographic structures. Thus, we simulate an ideal scenario in which the true distances are known for all contacts, a no errors associated to information uncertainty are involved. The goal of these experiments was to show that the addition of XL constraints to a DCA contact set provided more structure information than the addition of the same number of new DCA constraints to the set. Figure 3 displays the results of these modeling experiments for all three proteins. The quality of the models is evaluated by computing their TM-score relative to the crystallographic structure, and a TM-score greater than 0.5 usually indicates that the overall fold of the structure is correct.

For example, for 1C75, about 12% of the models were obtained with TM-scores greater than 0.5 in the modeling without constraints (NO CST—Figure 3A). With 20 DCA contacts, ∼37% of the models were obtained with the correct fold (20 DCA). With 20 XL constraints, this fraction was ∼49% (20 XL). Thus, in this case the 20 XL contacts appear to be more informative than the sets of 20 DCA contacts. Using 40 DCA constraints the fraction of correct folds increased to ∼66% (40 DCA), and using 20 DCA + 20 XL constraints about 75% of the models were obtained with the correct fold (20 DCA + 20 XL). Therefore, a better modeling is obtained if XL and DCA contacts are combined relative to doubling the number of DCA contacts. In this case, it is not possible to discern if this increased quality is a result of the combined information of the sets or simply by the fact that the XL contacts are more informative.

On the other side, for the other two proteins, 1C52 and 1D06, the use of the same number of DCA or XL constraints provided very similar modeling outputs when employed separately. For 1C52, using 66 DCA contacts, ∼29% of the models were obtained with the correct fold, while ∼30% of correct models were obtained using 66 XL contacts. Thus, the DCA and XL sets are almost equally informative. By doubling the DCA set to 132 constraints, ∼48% of the models are obtained with TM-scores greater than 0.5. Finally, using 66 DCA constraints and 66 XL constraints, ∼60% of the models were obtained with the correct fold. Therefore, the combined use of the constraints resulted in a greater fraction of correct folds than the modeling with the same number of DCA-only contacts, even though the types of constraints were, individually, equally informative. The same result was obtained for 1D06: using 34 DCA constraints resulted in ∼37% of models with TM-score greater than 0.5, while ∼46% of the models with the correct fold were obtained using 17 DCA + 17 XL constraints.

Thus, these modeling experiments endorse the thesis that constraint sets containing a combination of XL and DCA information have inferior cross-redundancy when compared to pure DCA sets. This is, of course, a combination of different factors: addition of extra DCA contacts means sampling more from the same residue and distance distribution, increasing the probability of redundant or correlated constraints, whereas introducing XLs enables sampling contacts that, for the most part, are structurally unprecedented; with a small amount of DCA contacts solving the protein core and another, uncorrelated set of XLs solving the protein surface, constraint synergy is maximized in a way that the duplication in the constraint set size translated to a duplication in modeling performance when DCA and XLs were combined, in comparison to the singular addition of extra DCAs, which resulted in more humble increases, in the order of 60% to 70%.

4 | CONCLUSION

In summary, we quantify the degree of complementarity of the distance constraints obtained by coevolutionary analysis and those obtained by chemical cross-linking. The contact sets obtained involve residue pairs in regions of the structure that rarely coincide, because of the nature of the structural information probed by each method. The combined use of DCA and XL contacts in modeling experiments provides better results than using a single contact type, even if considering the sets of contacts of similar number. The present results explain the potential of the combination of these constraints to

improve structural modeling protocols.[53] Since both methods are developed with computational and experimental simplicity in mind, and have widespread applicability, these results encourage the specific development of models and algorithms to combine DCA and XL sources of interaction data.

ACKNOWLEDGMENTS

This work was supported by São Paulo Research Foundation (FAPESP, grants: 2015/13667-9, 2010/16947-9, 2013/08293-7; 2016/13195-2; 2014/17264-3; 2018/24293-0; 2018/14274-9) and funds from the School of Natural Sciences and Mathematics at the University of Texas at Dallas.

ORCID

Ricardo N. dos Santos https://orcid.org/0000-0003-3315-747X
Guilherme F. Bottino https://orcid.org/0000-0003-1953-1576
Fábio C. Gozzo https://orcid.org/0000-0002-5270-4427
Faruck Morcos https://orcid.org/0000-0001-6208-1561
Leandro Martínez https://orcid.org/0000-0002-6857-1884

REFERENCES

1. Ovchinnikov S, Kinch L, Park H, et al. Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife*. 2015;4:e09248.
2. Huang P-S, Boyken SE, Baker D. The coming of age of de novo protein design. *Nature*. 2016;537(7620):320-327.
3. Ovchinnikov S, Park H, Varghese N, et al. Protein structure determination using metagenome sequence data. *Science*. 2017;355(6322):294-298.
4. Morcos F, Pagnani A, Lunt B, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA*. 2011;108(49):E1293-E1301.
5. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci USA*. 2009;106(1):67-72.
6. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet*. 2013;14(4):249-261.
7. Taylor WR, Jones DT, Sadowski MI. Protein topology from predicted residue contacts. *Protein Sci*. 2012;21(2):299-305.
8. Kosciolek T, Jones DT. Accurate contact predictions using covariation techniques and machine learning. *Proteins*. 2016;84(suppl 1):145-151.
9. Cheng RR, Raghunathan M, Noel JK, Onuchic JN. Constructing sequence-dependent protein models using coevolutionary information. *Protein Sci*. 2016;25(1):111-122.
10. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. *Proc Natl Acad Sci USA*. 2012;109(26):10340-10345.
11. Dorn G, Leitner A, Boudet J, et al. Structural modeling of protein-RNA complexes using crosslinking of segmentally isotope-labeled RNA and MS/MS. *Nat Methods*. 2017;14(5):487-490.
12. Ferber M, Kosinski J, Ori A, et al. Automated structure modeling of large protein assemblies using crosslinks as distance restraints. *Nat Methods*. 2016;13(6):515-520.
13. Scorsato V, Lima TB, Righetto GL, et al. Crystal structure of the human Tip41 orthologue, TIPRL, reveals a novel fold and a binding site for the PP2Ac C-terminus. *Sci Rep*. 2016;6:30813.
14. Lima DB, de Lima TB, Balbuena TS, et al. SIM-XL: a powerful and user-friendly tool for peptide cross-linking analysis. *J Proteomics*. 2015;129:51-55.
15. Leitner A, Walzthoeni T, Kahraman A, et al. Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol Cell Proteomics*. 2010;9(8):1634-1649.
16. Liu F, Heck AJR. Interrogating the architecture of protein assemblies and protein interaction networks by cross-linking mass spectrometry. *Curr Opin Struct Biol*. 2015;35:100-108.
17. Sharon M, Sinz A. Studying protein-protein interactions by combining native mass spectrometry and chemical cross-linking. *Analyzing Biomolecular Interactions by Mass Spectrometry*. 2015;55-79. https://doi.org/10.1002/9783527673391.ch2
18. Leitner A. Cross-linking and other structural proteomics techniques: how chemistry is enabling mass spectrometry applications in structural biology. *Chem Sci*. 2016;7(8):4792-4803.
19. Brodie NI, Popov KI, Petrotchenko EV, Dokholyan NV, Borchers CH. Solving protein structures using short-distance cross-linking constraints as a guide for discrete molecular dynamics simulations. *Sci Adv*. 2017;3(7):e1700479.
20. Sinz A, Arlt C, Chorev D, Sharon M. Chemical cross-linking and native mass spectrometry: a fruitful combination for structural biology. *Protein Sci*. 2015;24(8):1193-1209.
21. Hofmann T, Fischer AW, Meiler J, Kalkhof S. Protein structure prediction guided by crosslinking restraints – a systematic evaluation of the impact of the crosslinking spacer length. *Methods*. 2015;89:79-90.
22. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round XII. *Proteins*. 2018;86(S1):7-15.
23. Ferrari AJR, Gozzo FC, Martínez L. Statistical force-field for structural modeling using chemical cross-linking/mass spectrometry distance constraints. *Bioinformatics*. 2019;35(17):3005-3012.
24. Lössl P, van de Waterbeemd M, Heck AJ. The diverse and expanding role of mass spectrometry in structural and molecular biology. *EMBO J*. 2016;35(24):2634-2657.
25. Leitner A, Faini M, Stengel F, Aebersold R. Crosslinking and mass spectrometry: an integrated technology to understand the structure and function of molecular machines. *Trends Biochem Sci*. 2016;41(1):20-32.
26. Addlagatta A, Krzywda S, Czapinska H, Otlewski J, Jaskolski M. Ultra-high-resolution structure of a BPTI mutant. *Acta Crystallogr D Biol Crystallogr*. 2001;57(Pt 5):649-663.
27. Benini, S., Rypniewski, W.R., Wilson, K.S., Ciurli, S., and Mangani, S. The complex of Bacillus pasteurii urease with beta-mercaptoethanol from X-ray data at 1.65-A resolution. *J Biol Inorg Chem*. 1998;3(3):268-273.
28. Poy F, Yaffe MB, Sayos J, et al. Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition. *Mol Cell*. 1999;4(4):555-561.
29. Han GW, Lee JY, Song HK, et al. Structural basis of non-specific lipid binding in maize lipid-transfer protein complexes revealed by high-resolution X-ray crystallography. *J Mol Biol*. 2001;308(2):263-278.
30. Than ME, Hof P, Huber R, et al. Thermus thermophilus cytochrome-c552: a new highly thermostable cytochrome-c structure obtained by MAD phasing. *J Mol Biol*. 1997;271(4):629-644.
31. Cho SJ, Lee MG, Yang JK, Lee JY, Song HK, Suh SW. Crystal structure of *Escherichia coli* CyaY protein reveals a previously unidentified fold for the evolutionarily conserved frataxin family. *Proc Natl Acad Sci USA*. 2000;97(16):8932-8937.
32. Bolognesi M, Rosano C, Losso R, et al. Cyanide binding to *Lucina pectinata* hemoglobin I and to sperm whale myoglobin: an x-ray crystallographic study. *Biophys J*. 1999;77(2):1093-1099.
33. Miyatake H, Mukai M, Park SY, et al. Sensory mechanism of oxygen sensor FixL from rhizobium meliloti: crystallographic, mutagenesis and resonance Raman spectroscopic studies. *J Mol Biol*. 2000;301(2):415-431.

34. Kumaraswamy VS, Lindley PF, Slingsby C, Glover ID. An eye lens protein-water structure: 1.2 a resolution structure of gammaB-crystallin at 150 K. *Acta Crystallogr D Biol Crystallogr*. 1996;52(Pt 4):611-622.
35. Peapus DH, Chiu HJ, Campobasso N, Reddick JJ, Begley TP, Ealick SE. Structural characterization of the enzyme-substrate, enzyme-intermediate, and enzyme-product complexes of thiamin phosphate synthase. *Biochemistry*. 2001;40(34):10103-10114.
36. Tsunasawa S, Masaki T, Hirose M, Soejima M, Sakiyama F. The primary structure and structural characteristics of Achromobacter lyticus protease I, a lysine-specific serine protease. *J Biol Chem*. 1989;264(7):3832-3839.
37. Wang H, Boisvert D, Kim KK, Kim R, Kim SH. Crystal structure of a fibrillarin homologue from *Methanococcus jannaschii*, a hyperthermophile, at 1.6A resolution. *EMBO J*. 2000;19(3):317-323.
38. Lawson DM, Williams CE, Mitchenall LA, Pau RN. Ligand size is a major determinant of specificity in periplasmic oxyanion-binding proteins: the 1.2 A resolution crystal structure of *Azotobacter vinelandii* ModA. *Structure*. 1998;6(12):1529-1539.
39. Kuhn P, Knapp M, Soltis SM, Ganshaw G, Thoene M, Bott R. The 0.78 A structure of a serine protease: Bacillus lentus subtilisin. *Biochemistry*. 1998;37(39):13446-13452.
40. Khan AR, Parrish JC, Fraser ME, Smith WW, Bartlett PA, James MN. Lowering the entropic barrier for binding conformationally flexible inhibitors to enzymes. *Biochemistry*. 1998;37(48):16839-16845.
41. Finn RD, Bateman A, Clements J, et al. Pfam: The protein families database. *Nucleic Acids Res*. 2014;42(Database issue):D222-D230.
42. Ferrari AJR, Clasen MA, Kurt L, Carvalho PC, Gozzo FC, Martínez L. TopoLink: evaluation of structural models using chemical crosslinking distance constraints. *Bioinformatics*. 2019;35(17):3169-3170.
43. Fioramonte M, de Jesus HCR, Ferrari AJR, et al. XPlex: an effective, multiplex cross-linking chemistry for acidic residues. *Anal Chem*. 2018;90(10):6043-6050.
44. Bonneau R, Tsai J, Ruczinski I, et al. Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins*. 2001;45(suppl 5):119-126.
45. Bonneau R, Strauss CEM, Rohl CA, et al. De novo prediction of three-dimensional structures for major protein families. *J Mol Biol*. 2002;322(1):65-78.
46. Bradley P, Misura KMS, Baker D. Toward high-resolution de novo structure prediction for small proteins. *Science*. 2005;309(5742):1868-1871.
47. Raman S, Vernon R, Thompson J, et al. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins*. 2009;77(suppl 9):89-99.
48. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*. 1997;268(1):209-225.
49. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*. 1999;34(1):82-95.
50. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*. 2007;68(4):1020-1020.
51. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*. 2010;26(7):889-895.
52. Censoni L, Martínez L. Prediction of kinetics of protein folding with non-redundant contact information. *Bioinformatics*. 2018;34(23):4034-4038.
53. Dos Santos RN, Ferrari AJR, de Jesus HCR, Gozzo FC, Morcos F, Martínez L. Enhancing protein fold determination by exploring the complementary information of chemical cross-linking and coevolutionary signals. *Bioinformatics*. 2018;34(13):2201-2208.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: dos Santos RN, Bottino GF, Gozzo FC, Morcos F, Martínez L. Structural complementarity of distance constraints obtained from chemical cross-linking and amino acid coevolution. *Proteins*. 2020;88:625–632. https://doi.org/10.1002/prot.25843