Bioinformatics, 35(17), 2019, 3005–3012
doi: 10.1093/bioinformatics/btz013
Advance Access Publication Date: 10 January 2019
Original Paper

Structural bioinformatics

Statistical force-field for structural modeling using chemical cross-linking/mass spectrometry distance constraints

Allan J. R. Ferrari[1], Fabio C. Gozzo[1] and Leandro Martínez[1,2,*]

[1]Institute of Chemistry and [2]Center for Computing in Engineering & Sciences, University of Campinas, Campinas, SP, Brazil

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on September 18, 2018; revised on December 3, 2018; editorial decision on December 29, 2018; accepted on January 4, 2019

Abstract

Motivation: Chemical cross-linking/mass spectrometry (XLMS) is an experimental method to obtain distance constraints between amino acid residues which can be applied to structural modeling of tertiary and quaternary biomolecular structures. These constraints provide, in principle, only upper limits to the distance between amino acid residues along the surface of the biomolecule. In practice, attempts to use of XLMS constraints for tertiary protein structure determination have not been widely successful. This indicates the need of specifically designed strategies for the representation of these constraints within modeling algorithms.

Results: A force-field designed to represent XLMS-derived constraints is proposed. The potential energy functions are obtained by computing, in the database of known protein structures, the probability of satisfaction of a topological cross-linking distance as a function of the Euclidean distance between amino acid residues. First, the strategy suggests that XL constraints should be set to shorter distances than usually assumed. Second, the complete statistical force-field improves the models obtained and can be easily incorporated into current modeling methods and software. The force-field was implemented and is distributed to be used within the Rosetta *ab initio* relax protocol.

Availability and implementation: Force-field parameters and usage instructions are freely available online (http://m3g.iqm.unicamp.br/topolink/xlff).

Contact: leandro@iqm.unicamp.br

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

1 Introduction

The number of protein structures determined is much smaller than that of proteins known at the sequence level (Bairoch and Apweiler, 2000; Bateman *et al.*, 2017; Berman *et al.*, 2000; Pundir *et al.*, 2017). This discrepancy is the result of experimental limitations to obtain high-resolution structures and, in parallel, the experimental advances in genome sequencing. *In silico* approaches for protein structure prediction have been applied to fill that gap. If the structure of homologous proteins has already been solved, modeling the target protein is relatively simple (Eswar *et al.*, 2006; Fiser, 2010; Song *et al.*, 2013). Without structural homologous, however, the determination of a protein fold is a major challenge in computational biology.

Many types of data can be used to improve structural modeling of protein structures. For example, sparse NMR data (Bowers *et al.*, 2000; Tang *et al.*, 2015; Thompson *et al.*, 2012), Small Angle X-Ray Scattering (Schneidman-Duhovny *et al.*, 2012), Cryo-electron Microscopy (DiMaio *et al.*, 2015), distance constraints derived from residue coevolution statistics (Ovchinnikov *et al.*, 2016, 2017) and, more recently, from chemical cross-linking/mass spectrometry (XLMS) (Belsom *et al.*, 2016; Brodie *et al.*, 2017; Santos *et al.*, 2018).

XLMS is attractive experimentally because it requires cheap and more accessible instrumentation, simple sample handling and small amounts of sample. Furthermore, the results are tolerant to contaminants and, in principle, XLMS data can be obtained for any protein, as MS is a universally applicable technique.

In some sense, the information XLMS provides is similar to that obtained from NMR, that is, a list of distance constraints between atoms. Nevertheless, important differences exist: (i) the XLMS constraint is a distance along the surface of the protein; (ii) the constraint is in principle associated only to the maximum linker reach, that is, it is only an upper bound to the distance between the residues; (iii) the length of the linker can be of the order of several Angstroms and, thus, geometrically associates residues which are relatively far on the protein structure. Experimentally, XLMS presents its own limitations, which are a field of intense state-of-the-art research: the number XLMS constraints is limited by the diversity of the reactivity of the linkers, and by the exposure of the residues on the protein surface. Also, the interpretation of XLMS spectra is still a complex task (Iacobucci and Sinz, 2017), requiring specific algorithms and software (Götze *et al.*, 2012; Hoopmann *et al.*, 2015; Kosinski *et al.*, 2015; Lima *et al.*, 2015; Sarpe *et al.*, 2016) and possibly manual curation.

Because of the current limitations in experimental and modeling techniques, the use of chemical cross-links has been indisputably successful only for the determination of quaternary arrangements. Their use for tertiary structure modeling remains a challenge. For instance, in the CASP11 and CASP12 assisted competitions, no clear improvements in the quality of the models were observed from the use of experimental cross-linking constraints (Schneider *et al.*, 2016; Tamò *et al.*, 2017). Recently, we were able to model the tertiary structures of a variety of models with the support of XLMS distance constraints, but only in combination with distance constraints derived from amino acid coevolution analysis, which played a determinant role in obtaining models with fold-level accuracy (Santos *et al.*, 2018).

The incorporation of XLMS constraints in structural modeling strategies is indeed a challenge. The distance constraints are along the protein surface; thus, their precise evaluation depends on the model structure which, in principle, is not known. Furthermore, the evaluation of the surface-accessible distance between two residues requires specialized strategies and, of course, is much more computationally demanding than the evaluation of straight Euclidean distances. Therefore, these constraints have been implemented in the modeling process through Euclidean-distance-dependent energy functions that aim to constrain the maximum distance between residues observed to be cross-linked. The maximum distance is usually derived from the maximum cross-linker and side chain extensions, through simple geometrical arguments.

Here, we formulate a Euclidean-distance-dependent structure-based statistical force-field for cross-linking/mass spectrometry constraints, named XLFF. In summary, we compute from a database of nonredundant protein structures the probability of observing two residues at a surface-accessible cross-linking distance as a function of the Euclidean distance between their Cβ atoms. This probability curve is converted into a potential energy function assuming it obeys a Boltzmann distribution. The potential is dependent on the cross-linker length and on the nature of the residues involved, thus defining a residue and linker-dependent force-field for structural modeling with XLMS distance constraints. We implemented the force-field in the Rosetta *ab initio* relax protocol (Bonneau *et al.*, 2001, 4, 2002; Bradley *et al.*, 2005; Raman *et al.*, 2009, 8; Simons *et al.*, 1999) and demonstrate that this statistical force-field increases significantly the probability of obtaining native-like tertiary structures compared to current approaches to represent the constraints. Although here we focus on the more challenging problem of tertiary protein structure determination, the principles here described find application in other structural modeling goals, including the determination of general protein assemblies.

2 Approach

Chemical cross-linking is an experimental method to obtain structural information from a chemical modification of the protein with a reagent called cross-linker, or simply linker. If a residue is found to attach to the linker, it means that the residue is accessible to the solvent in some significantly populated protein conformation in solution. If, additionally, the linker is found to be attached to a pair of residues, A and B, it follows that the reactive atoms A_x and B_y are closer to each other up to the length of the cross-linker spacer arm, L_{XL}. This linker works as a molecular ruler over the protein surface. Thus, when measuring the distance between A_x and B_y, $d(A_x, B_y)$, one should consider the physical path between them, $d_{top}(A_x, B_y)$, where the subscript *top* stands for 'topological' distance, which we define here as the shortest path physically accessible to the linker connecting the reactive atoms (see Fig. 1).

Backbone atoms provide more stable reference positions for the introduction of constraining potentials and Cα atoms have been used to define the distance limits. Here, we define a statistical force-field based on the probability of the reactive atoms of the side chain being at a cross-linkable topological distance given the Euclidean

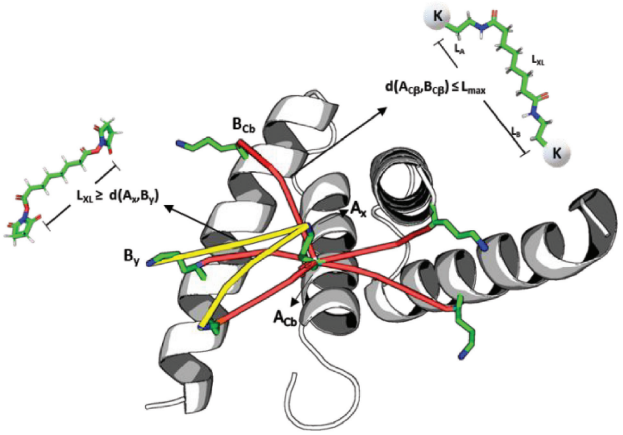

Fig. 1. Cross-linking information. The identification of a cross-link between two residues (yellow lines) implies that the distance between reactive atoms, $d(A_x, B_y)$ is smaller than the extended linker length, L_{XL}. However, a fixed backbone structure cannot represent all cross-linkable side chains configurations. As indicated by the red lines connecting Cβ atoms, alternative configurations of side chains for a single fixed backbone can potentially validate other three cross-links. Therefore, at least the variability of side chain orientations should be considered to define the effective maximum distance, L_{max}, between residues that can be cross-linked. Here, we develop a potential energy function to represent the interactions between pairs of residues that can be cross-linked in monomeric proteins, which will be specific for each type of residue and linker. For a given linker and residue pair, the linker reactivity is associated with a side-chain chemical group. For example, the amine group of Lysine residues, or carboxylate group of acidic side chains. Therefore, in principle, the observation of a cross-link is associated with these side-chain atoms being within the linker length, $d_{top}(A_x, B_y) \leq L_{XL}$. Since structural models are static and side chains exposed to solvent are frequently mobile, the distance between side-chain atoms in a model reflects poorly the possibility of a cross-link

distance between corresponding Cβ atoms. Side chains' atoms fluctuate anchored on Cβ atoms which are static for a given backbone conformation. Therefore, Cβ atoms positions provide an additional precision compared to Cα atoms. In addition, Cβ distances retain information on the directionality of the side chain. When computing the topological distances, that option avoids the validation of topological paths that are nonphysical because would require the linker to assume a conformation that cannot be sampled by the side-chain. Finally, this statistical force-field considers implicitly the flexibility of the side chains and, by being based on Euclidean distances, is practical to use.

The maximum topological distance between Cβ atoms consistent with the formation of a cross-link is $L_{max}=L_{XL}+L_A+L_B$, where L_{XL} is the maximum linker length and L_A and L_B are the lengths of the side chains of the residues involved. If the topological distance between Cβ atoms, $d_{top}(A_{C\beta}/B_{C\beta})$, is smaller than L_{max}, residues *A* and *B* may form a cross-link, since the side-chains can potentially fluctuate to assume conformations compatible with the topological path along the surface associated with $d_{top}(A_{C\beta}, B_{C\beta})$. To a first approximation, using $d_{top}(A_{C\beta}, B_{C\beta})$ and L_{max} to constrain residue distances can be considered a strategy to incorporate the side-chain flexibility into the modeling procedure.

However, L_{max}, when implemented as a Euclidean distance constraint, represents an unlikely scenario in which the linker and both side chains are in their fully extended conformations. Intuitively, constraining Cβ atoms distances to something smaller than L_{max} should be a good strategy in most cases. In this work, we first propose that an effective L_{max} can be assessed by statistical analysis of known protein structures. We compute the frequency distribution of $d_{top}(A_{C\beta}, Bc_{\beta})$ in a protein database, under the condition that $d_{top}(A_x, B_y) \leq L_{XL}$. A significant reduction of the distance-cutoff can be obtained by eliminating a small fraction (i.e. 1%) of unlikely scenarios (see Supplementary Section S1.1). Thus, we defined a more constrained distance-cutoff which encompasses 99% of the possible crosslinks, $L_{max}(0.99)$.

The statistical analysis above can be further refined for the establishment of a distance-dependent force field for XLMS constraints. Imagine that a pair of Cβ atoms from residues *A* and *B* are found at a Euclidean distance $d_{euc}(A_{C\beta}, B_{C\beta})$. This distance is associated with a topological distance, $d_{top}(A_{C\beta}, B_{C\beta})$, as defined above. Given a database of known protein structures, we ask what is the probability that the topological distance $d_{top}(A_{C\beta}, B_{C\beta})$ is smaller than $L_{max}(0.99)$ given the Euclidean distance, $d_{euc}(A_{C\beta}, B_{C\beta})$, that is, $p[(d_{top}(A_{C\beta}, B_{C\beta})<L_{max}(0.99))|d_{euc}(A_{C\beta}, B_{c\beta})]$. The potential energy that would imply this probability distribution, assuming Boltzmann sampling is

$$V(d_{euc}) = -\mathrm{RT}\ln\ p[\left(d_{top}(A_{C\beta}, B_{C\beta}) < L_{max}(0.99)\right)|d_{euc}(A_{C\beta}, B_{c\beta})], \quad (1)$$

where at room temperature, RT = 0.569 kcal mol^{-1}. This potential can be directly incorporated into most modeling procedures, as it is dependent on the Euclidean distances between Cβ atoms and on the $L_{max}(0.99)$ from the structural database. Section 3.1 and Supplementary Section S1 describes the details of the parameterization and implementation of this potential energy function.

The statistical force-field was implemented in Rosetta *ab initio* relax protocol and proved to be superior in terms of modeling quality to current state-of-art approaches for XLMS constraint representation, as we will show. Modeling details are available in Supplementary Section S2. All modeling results, including input and output raw files, are available at http://m3g.iqm.unicamp.br/topolink/xlff. Each modeling round consisted of generating 1000 models with Rosetta (Supplementary Protocol S1 for more details). We evaluated the quality of the models by the distribution of the structural similarity of the models to the crystallographic structure, as given by the TM-score metric computed with LovoAlign (Andreani *et al.*, 2009). Structures with TM-scores greater than 0.5 relative to the crystallographic structure are considered to have roughly the correct fold (Xu and Zhang, 2010). Structures with TM-scores greater than 0.6 are likely winner candidates at the CASP modeling competitions.

3 Results

3.1 Parametrization of the statistical force field

Figures 2 and 3 exemplify the construction of the statistical force-field for a pair of reactive residues. In Figure 2A, we display the frequency of observation of topological distances between Lys Nζ atoms in the CATH database of nonredundant domains. The subset of pairs of Lys residues for which the Nζ are within the length of the linker molecule (11.5 Å) was shortlisted, and the distribution of Cβ distances for these pairs is obtained, as shown in Figure 2B. The most frequent Cβ–Cβ topological distance associated with Nζ–Nζ topological distances smaller than 11.5 Å is about 7.5 Å, corresponding to Cβ–Cβ topological distances of vicinal residues in α-helices and 99% of the Cβ atoms are closer than 17.8 Å. Therefore, we consider 17.8 Å the maximum effective distance between Cβ atoms for this linker. This maximum effective distance will be named the *Statistical Limit* of the linker.

Then, we compute the probability of finding the Cβ atoms of the Lys residues closer than 17.8 Å as a function of their Euclidean distance, as shown in Figure 3A. This probability shows, for example, that if the Euclidean distance between Cβ atoms of Lys residues is greater than ~14 Å, there is only 50% probability that a topological path connecting these residues exists within the reactive distance.

This probability distribution is translated, according to Equation (1), into a statistical potential, which is represented in Figure 3B. For instance, this potential introduces an increasing energy for the Euclidean distance between Cβ atoms at all distances, but which is

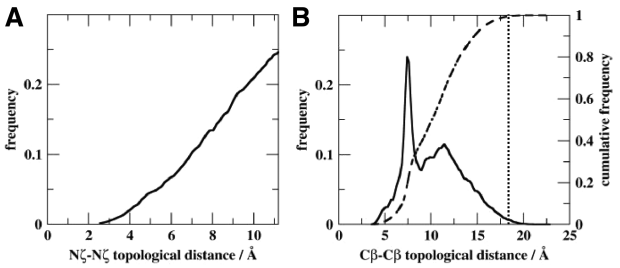

Fig. 2. Statistical-based definition of L_{max}. (**A**) After computing the topological distance between Nζ atom pairs from structures on CATH S40 database, we selected the subset of pairs with distances shorter than the linker length, 11.5 Å. (**B**) Next, we selected the subset of topological distances between Cβ atoms pairs contained in the previous subset. The topological distance distribution reveals that distances corresponding to side chains and linker in extended conformations (~22 Å) are never observed. We define a cross-linkable distance for Lysine pairs and DSS/BS3 cross-link after removing unlikely scenarios (1%) as $L_{max}(0.99)$ = 17.8 Å (vertical dashed line), increasing the restrictive role of the constraint by more than 4 Å (see Supplementary Section S1.1). Similar profiles for other residue pairs are shown in Supplementary Figure S1

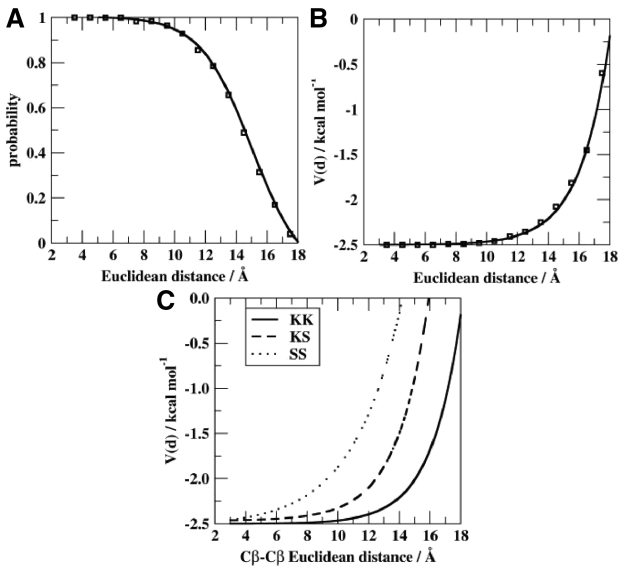

Fig. 3. Statistical force-field determination. (**A**) Probability that a topological distance is below $L_{max}(0.99) = 17.8$ Å as a function of the Euclidean distance between Cβ Lys/Lys pairs. As the Euclidean distance reaches L_{max}, the probability of satisfying the topological length decreases because fewer possible physical paths connecting the residues are possible. (**B**) A potential energy curve can be derived from (A) assuming that this probability distribution is a Boltzmann distribution. (**C**) For each pair of residue types a different energy function is derived. Here, we show the energy functions for KK, KS and SS pairs assuming a DSS/BS3 cross-linker. Potential energy profiles for other residue pairs and linkers are shown in Supplementary Figure S3

particularly noticeable above ∼12 Å. Therefore, effectively, the force field penalizes distances which are greater than ∼12 Å.

In Figure 3C, the potential energy profiles for the pairs KK, KS and SS, obtained with the same protocol, are shown. As expected, for shorter side-chains, the potential energy increases at shorter Cβ–Cβ distances. This reflects the fact that, for example, linkers of the same length may bind Lys residues at larger Cβ–Cβ distances than Ser residues, as result of the difference between side chain lengths. The exact profile of the potential is dependent not only on the lengths of the side chains, but also on the nature of their interactions with the surface of the proteins, and these are implicitly taken into account in the present approach because the profiles are obtained from the structural database. The profiles of the potential energies of all other reactive pairs of residues are shown in Supplementary Figure S3 and are available for download.

3.2 Modeling performance

3.2.1 Overview of previously attempted cross-linking representation strategies

Different interaction potentials have been proposed for the use of XLMS-derived constraints in protein modeling protocols. Kahraman *et al.* (2013) have proposed using a FlatHarmonic potential to integrate cross-linking constraints data to *de novo* and comparative protein prediction. The FlatHarmonic potential penalizes models having Euclidean distance between two atoms farther than an upper distance limit, UL (see Supplementary Fig. S4). In that work, the UL was chosen as 30 Å for all constraints associated with the same linker (DSS/BS3). Belsom *et al.* (2016) have proposed a Lorentz-like potential. As shown in Supplementary Figure S4, instead of penalizing models having Euclidean distance above a certain UL, the Lorentz potential rewards models for which Euclidean distances are below a threshold, that is, in which it is believed that the cross-link should be satisfied. Above this limit, there is a progressive decrease in the energy bonus to zero. This potential is argued to be more tolerant to the presence of incorrect constraints. The Serum Albumin domains were modeled using this function, but in combination with contact prediction constraints from evolutionary information. Therefore, the specific role of the XLMS constraints in model quality was not addressed. Merkley *et al.* (2014) have proposed a justification for this UL = 30 Å for cross-links between Lysine pairs: the correlation between Cα Euclidean distances in a set of crystal structures and molecular dynamics simulation, and a set of experimental cross-linking data for cytochrome C, showed that the experimental information of cross-linking data is often not represented by a single conformational state. It is then proposed that using larger constraint distances would account for the conformational flexibility of the structure. This concept of adding some threshold to the extended conformation of the linker to account for structural variability has guided much of the XL modeling and validation strategies (Chavez *et al.*, 2016, 2018; Fritzsche *et al.*, 2012; Herzog *et al.*, 2012; Kahraman *et al.*, 2013; Kalisman *et al.*, 2012). Alternatively, the structural variability can be addressed by the proposal of multiple models, without sacrificing the precision of the XLMS ruler (Degiacomi *et al.*, 2017).

All these proposals were evaluated here in the context of modeling with Rosetta's *ab initio* relax protocol (Bonneau *et al.*, 2001, 2002; Bradley *et al.*, 2005; Raman *et al.*, 2009, 8; Simons *et al.*, 1997, 1999). We collected experimental data for four different targets: SalBIII (Luhavaya *et al.*, 2015), a 15.6 kDa protein with a low sequence similarity to other proteins in the Protein Data Bank; and the three domains of Albumin (Sugio *et al.*, 1999) (ALB-D1, ALB-D2 and ALB-D3), that have been standard examples in cross-linking experiments (Belsom *et al.*, 2016; Fischer *et al.*, 2013; Huang *et al.*, 2004).

In this section, we will describe the modeling performed with ideal cross-linking datasets, computed from the crystallographic models with Topolink (Ferrari *et al.*, 2019) (http://m3g.iqm.unicamp.br/topolink). The SalBIII set contains 62 constraints compatible with the crystallographic structure. For ALB-D1, ALB-D2 and ALB-D3, the sets contained 125, 153 and 92 constraints, respectively. Section 3.3 describes the modeling results obtained with more limited experimental sets of constraints. In addition, we benchmarked our strategy by randomly choosing 15 proteins of three different categories (mainly alpha, mainly beta and alpha–beta proteins) and sizes between 70 and 250 residues from the CATH S40 database, for which we also modeled the cross-linking results. Due to space limitations, these additional results can be found in Supplementary (Section S3).

Modeling was performed without constraints and applying three cross-linking constraints representations: the FlatHarmonic potential, the Lorentz potential and the statistical potential. The upper limit distances considered for the FlatHarmonic and Lorentz potential were (i) UL = 25 Å between Cβ atoms, assuming that the UL incorporates the conformational diversity of the structure; and (ii) UL = $L_{max}(0.99)$, the *Statistical Limit*, which is computed for each pair of residue-types independently as shown in Table 1.

3.2.2 The statistical upper limit improves significantly the quality of the models

The distributions of model quality obtained using the different representation of the constraints and upper limits are shown in Figure 4

Table 1. Extended and statistical (L_{max}) distances for cross-linked residue pairs and linkers

Cross-link ID	Extended distance/Å	$L_{max}(0.99)^a$/Å
BS3/DSS		
KK	21.8	17.8
KS	18.0	15.8
SS	14.1	13.4
1,6-Hexanediamine		
DD	14.1	13.5
DE	15.4	14.3
EE	16.7	15.1
Zero-length		
KE	—	10.5
KD	—	9.7
SE	—	7.7
SD	—	7.0

Note: The effective maximum distances that account for 99% of the possible cross-links are significantly more restrictive than the maximum linker lengths. Extended conformations are not frequently observed.

[a]Statistically derived topological Cβ–Cβ distance/Å.

for the SalBIII and Albumin domains. Distributions obtained without constraints are repeated in each graph as a reference. The left-side graphs display the quality distributions of all 1000 models, while the right panels show the fraction of models with fold-level accuracy that are obtained for subsets of the models as classified by percentiles of their total scores (Rosetta + constraints energy). The results are summarized in Table 2.

None of the models obtained without constraints have TM-score greater than 0.5, thus confirming that experimental constraints are essential for the modeling of this protein. The FlatHarmonic function with UL = 25 Å performs slightly better than the modeling without constraints for Albumin domains but has the same performance in modeling SalBIII. Setting the UL to the statistical distances, however, improves dramatically the overall quality of models. Using the statistical limits derived here and the FlatHarmonic energy function, it is possible to obtain 22, 52, 46 and 18% of native-like structures for the Albumin D1, D2 and D3 domains and SalBIII, respectively. Notably, by selecting the 10% best-scored models, the native-like populations increase to 62, 90, 96 and 72%. Therefore, using a more constrained potential increases the quality of the models obtained significantly. Here, this choice is justified by the statistical analysis of cross-linkable pairs in known protein structures. For example, we know that L_{max}=17.8 Å is too restrictive for only 1% of the cross-links between Lysine residues.

The use of the Lorentz potential did not result in any improvement of the models relative to modeling without constraints, independently of the ULs used. This is likely a consequence of the Lorentz potential having a null gradient at almost every distance (see Supplementary Fig. S4). Therefore, in gradient-dependent modeling strategies, like the ones used by Rosetta, this potential could only affect the selection of the models by their final energy. In other words, this potential might be useful for modeling using Monte-Carlo sampling methods alone, but any advantage that might be gained by using gradient information is lost.

3.2.3 The statistical force-field optimally weights constraint penalties

Finally, we used the complete statistical force-field (XLFF) to model protein domains. The statistical weights introduced improved expressively the modeling results as shown in the green curves of

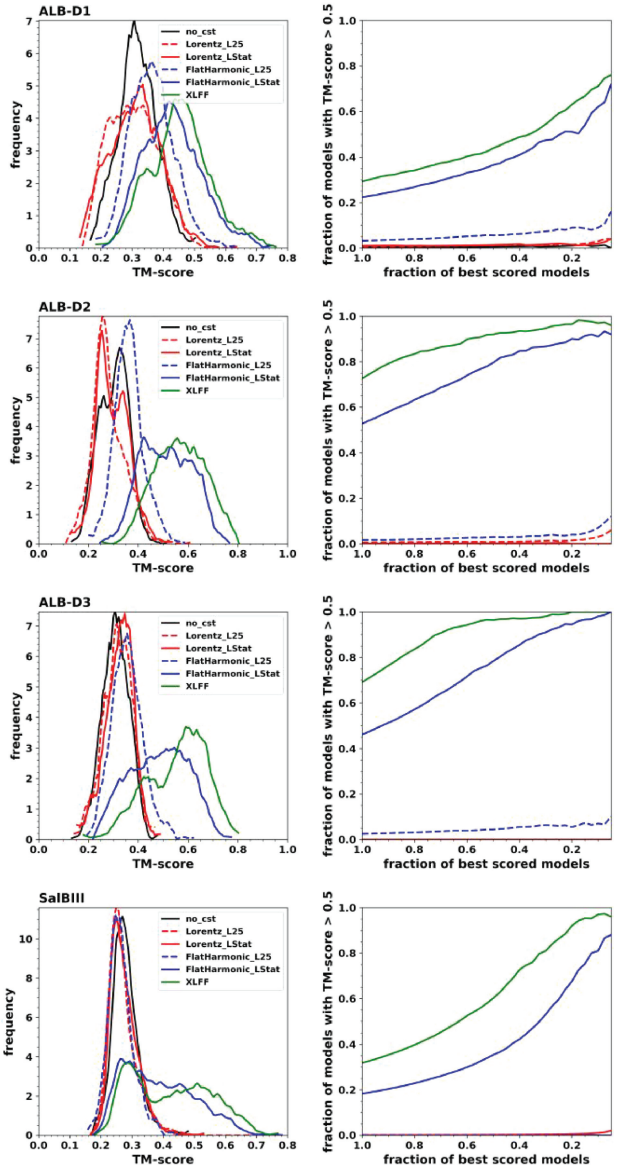

Fig. 4. Performance of cross-linking energy functions in the modeling of Albumin domains and SalBIII with Rosetta *ab initio* relax protocol. (Left panels) Lorentz energy function produces the same distributions as the modeling without constraints, our negative control, regardless the upper limit applied. There is a significant improvement in the quality of the models obtained when using the statistical upper distance limit proposed here with FlatHarmonic potential. In all cases, XLFF outperforms previous cross-linking constraints representations. (Right panels) The fraction of models with TM-score increases by selecting models with better total scores (Rosetta + XL energy) in the case of the FlatHarmonic potential with the statistical limit or by using XLFF. The other cases (no constraints, Lorentz with both limits and FlatHarmonic with loose limits) the effect of the selection is negligible

Figure 4. Using the statistical potential, 29, 73 and 69 of all models of Albumin D1, D2 and D3, respectively, are native-like structures, which correspond to an increase in up to 40% of the fraction of native-like structures obtained with FlatHarmonic energy function when applied the statistical limit. Furthermore, 65, 98 100% of the 10% best Rosetta-scored models are native-like structures. For SalBIII, the fraction of native-like models increases to 32% of all models and 97% out of the 10% best scored models.

Table 2. Evaluation of population of models with fold-level accuracy of Albumin domains and SalBIII generated using different available energy functions to represent XLMS constraints

Fraction of models with TM-score > 0.5

Target ID	Constraint set	Energy function	UL	All models	50% models with best total score	10% models with best total score
ALB-D1	No constraints	–	–	0.00	0.00	0.01
	Ideal	FlatHarmonic	25 Å	0.03	0.06	0.09
			Statistical limit	0.22	0.37	0.62
		Lorentz	25 Å	0.00	0.01	0.03
			Statistical limit	0.01	0.01	0.02
		XLFF	Statistical limit	0.29	0.44	0.71
	Experimental	FlatHarmonic	25 Å	0.01	0.01	0.02
			Statistical limit	0.03	0.04	0.09
		XLFF	Statistical limit	0.07	0.01	0.18
ALB-D2	No constraints	–	–	0.00	0.00	0.00
	Ideal	FlatHarmonic	25 Å	0.02	0.03	0.07
			Statistical limit	0.53	0.80	0.91
		Lorentz	25 Å	0.00	0.00	0.03
			Statistical limit	0.00	0.00	0.00
		XLFF	Statistical limit	0.73	0.92	0.97
	Experimental	FlatHarmonic	25 Å	0.01	0.02	0.04
			Statistical limit	0.16	0.24	0.34
		XLFF	Statistical limit	0.25	0.43	0.62
ALB-D3	No constraints	–	–	0.00	0.00	0.00
	Ideal	FlatHarmonic	25 Å	0.03	0.04	0.07
			Statistical limit	0.46	0.78	0.98
		Lorentz	25 Å	0.00	0.00	0.00
			Statistical limit	0.00	0.00	0.00
		XLFF	Statistical limit	0.69	0.96	1.00
	Experimental	FlatHarmonic	25 Å	0.00	0.00	0.00
			Statistical limit	0.01	0.01	0.04
		XLFF	Statistical limit	0.02	0.02	0.03
SalBIII	No constraints	–	–	0.00	0.00	0.00
	Ideal	FlatHarmonic	25 Å	0.00	0.00	0.00
			Statistical limit	0.18	0.35	0.81
		Lorentz	25 Å	0.00	0.00	0.01
			Statistical limit	0.00	0.00	0.01
		XLFF	Statistical limit	0.32	0.60	0.97
	Experimental	FlatHarmonic	25 Å	0.00	0.00	0.02
			Statistical limit	0.05	0.09	0.24
		XLFF	Statistical limit	0.12	0.24	0.54

Note: The use of statistical upper limits improves significantly the quality of the models obtained with FlatHarmonic potential, but no difference is observed with Lorentz function. XLFF, the proper statistical representation of XLMS constraints, improves even further the quality of the models obtained. The modeling obtained with XLFF the experimental constraint set can outperforms previous energy functions using the ideal constraint set.

In summary: (i) the extended linker lengths are unnecessarily large, leading to information loss, independently of the potential energy function used; (ii) restricting UL to statistically relevant distances, $L_{max}(0.99)$, improves significantly the probability of generating native-like structures and (iii) the statistical force-field, $V(L_{max}, d_{euc})$, is the best strategy to implement XLMS-derived constraints in modeling.

3.3 Modeling with experimental XLMS constraints

In the previous sections, we evaluated protein tertiary structures modeling performance from an ideal perspective of having identified all potential cross-links consistent with the crystallographic structure.

We discuss now the modeling results obtained using experimental data. In this work, we consider the use of XLMS constraints only. Evidently, the present approach can be, and should be, combined with other source of data if available. Specifically, solvent accessibility information can be obtained from dead-ends in XLMS experiments or protein footprinting/mass spectrometry and can be used for model classification (Aprahamian *et al.*, 2018) and, potentially, model generation.

We performed XLMS experiments using DSS and the Xplex chemistry (Fioramonte *et al.*, 2018) on SalBIII and Serum Albumin domains (Supplementary Section S2) and analyzed the resulting MS raw files with SIM-XL (Borges *et al.*, 2015; Lima *et al.*, 2015) to obtain a list of cross-link candidates. The set of cross-links obtained experimentally which is consistent with the crystallographic structure contains from 25 to 40% of the ideal cross-link set (Supplementary Section S2 and Table S1). This reduced set of cross-links was used for structural modeling with FlatHarmonic and XLFF potentials.

Figure 5 shows, as expected, that reducing the number of constraints worsens the TM-score distribution of the models obtained by Rosetta relative to the ideal set. Nevertheless, there is an improvement relative to modeling without constraints or FlatHarmonic with loose limits. Modeling Albumin D3 with the XLFF provides 1% of native-like models, a fraction that increases to 4% if the 10% best-scored models are considered. This result is

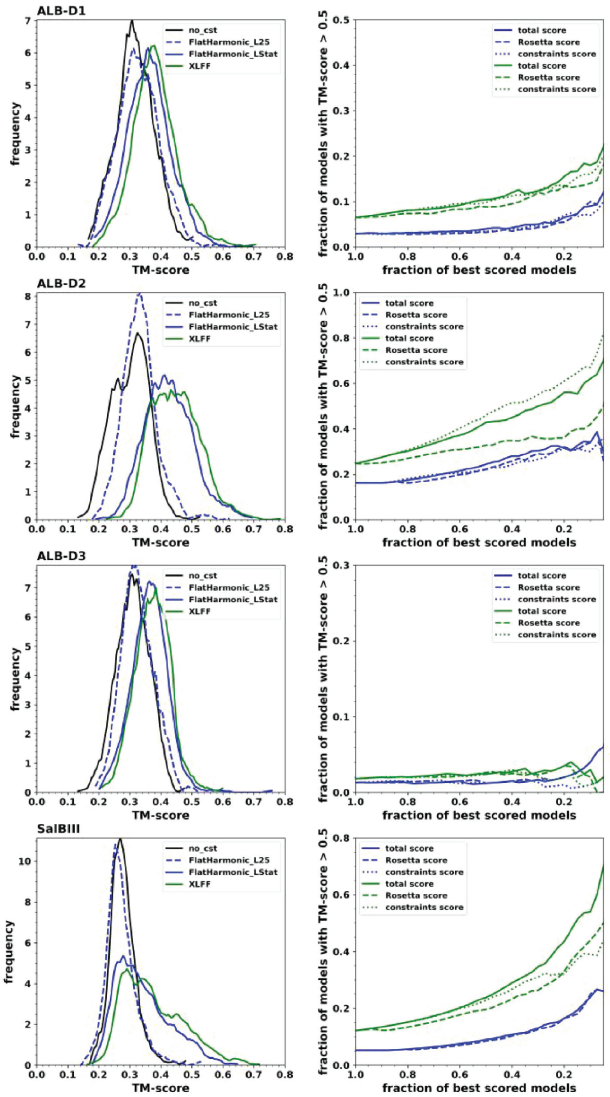

Fig. 5. Modeling Albumin domains and SalBIII with XLMS force field and experimental constraints. (Left panel) As expected, the number of models with TM-score greater than 0.5 decreases compared to the set of *in silico*-derived constraints as result of experimental limitations. However, the use of Statistical limits and, especially, XLFF, is still able to produce an expressive number of native-like models. (Right panels) We rescored the same set of models produced with XLFF (green curves) and FlatHarmonic with the statistical limits (blue curves) with Rosetta energy only, XLFF score only and the composed Rosetta plus constraints energy (named total score). For XLFF, in most cases, the composed score reveals more adequate to select native-like models. For all cases, Rosetta energy only produces worse selection. For FlatHarmonic potential, no significant difference is observed regarding the scoring method applied. Additional information on the number of experimental constraints and how it compares with the theoretical set of constraints if provided on Supplementary Table S1

comparable to that obtained for Albumin D3 using the ideal constraint-set with the FlatHarmonic potential with the extended linker. This domain is particularly challenging because it offers less cross-link possibilities proportionally to its size (Supplementary Table S1). For all other examples, using the XLFF outperform previous constraints representations even if the results are compared to the ones obtained with the ideal set of constraints.

The exact form of the potential energy function depends on the depth of the constraint potential, which was varied by changing the Rosetta constraint weight parameter. Increasing depth of the potential energy appears to improve slightly the fraction of native models obtained, but the difference is not significant for the best scored models (Supplementary Fig. S5). The optimal weight of the constraints is probably dependent on the quality of the dataset (fraction of correct versus incorrect constraints) and can be adjusted by the user. Also, increasing the amount of sampling for each structure in Rosetta improves the quality of the models (Supplementary Fig. S6), as expected. In this case, the user would probably choose to sample as much as possible, although an alternative is to generate more models.

Finally, the XL force-field is also useful to qualify the final models obtained. As shown in the left panels of Figure 5, the complete model potential energy including the XLFF potential outperforms the Rosetta score in its correlation with the quality of the models. For Albumin D2, in particular, the constraint score outperforms even the total model score.

4 Conclusions

A force-field designed for modeling XLMS constraints was developed. The potential energy functions representing the constraints are obtained from the statistics of physically accessible distances between residues in the database of known protein structures. The potential energy function is dependent on the Euclidean distance between residues and on the structural properties of the linker and associates more favorable interactions to pairs of residues which are more likely associated with a valid cross-linking path. The force-field was implemented in the Rosetta modeling suite, and expressively improve the quality of the models obtained. These results bring to reality the possibility of modeling from XLMS constraints the tertiary structures of proteins for which other structural data is not available or is insufficient for characterizing the protein fold. In this work, we only addressed the problem of structure modeling in the presence of distance constraints which are correct from the point of view of a single target model, so additional strategies to deal with multiple conformations and noisy data must be developed.

Funding

We thank the financial support of the funding agency FAPESP [Grants 2010/16947-9, 2013/05475-7, 2013/08293-7, 2013/23814-3, 2014/17264-3, 2018/14274-9 and 2016/13195-2].

Conflict of Interest: none declared.

References

Andreani,R. *et al.* (2009) Low order-value optimization and applications. *J. Glob. Optim.*, **43**, 1–22.

Aprahamian,M.L. *et al.* (2018) Rosetta protein structure prediction from hydroxyl radical protein footprinting mass spectrometry data. *Anal. Chem.*, **90**, 7721–7729.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Bateman,A. *et al.* (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–169.

Belsom,A. *et al.* (2016) Serum albumin domain structures in human blood serum by mass spectrometry and computational biology. *Mol. Cell. Proteomics*, **15**, 1105–1116.

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

Bonneau,R. *et al*. (2002) *De novo* prediction of three-dimensional structures for major protein families. *J. Mol. Biol*., **322**, 65–78.

Bonneau,R. *et al*. (2001) Rosetta in CASP4: progress in *ab initio* protein structure prediction. *Proteins Struct. Funct. Bioinform*., **45**, 119–126.

Borges,D. *et al*. (2015) Using SIM-XL to identify and annotate cross-linked peptides analyzed by mass spectrometry. *Protoc. Exch*.

Bowers,P.M. *et al*. (2000) *De novo* protein structure determination using sparse NMR data. *J. Biomol. NMR*, **18**, 311–318.

Bradley,P. *et al*. (2005) Toward high-resolution *de novo* structure prediction for small proteins. *Science*, **309**, 1868–1871.

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.

Chavez,J.D. *et al*. (2018) Chemical cross-linking mass spectrometry analysis of protein conformations and supercomplexes in heart tissue. *Cell Syst*., **6**, 136–141.e5.

Chavez,J.D. *et al*. (2016) *In vivo* conformational dynamics of Hsp90 and its interactors. *Cell Chem. Biol*., **23**, 716–726.

Degiacomi,M.T. *et al*. (2017) Accommodating protein dynamics in the modeling of chemical crosslinks. *Structure*, **25**, 1751–1757.e5.

DiMaio,F. *et al*. (2015) Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat. Methods*, **12**, 361–365.

Eswar,N. *et al*. (2006) Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinform*., **15**, 5.6.1–5.6.30.

Ferrari,A.J.R. *et al*. (2019) TopoLink: evaluation of structural models using chemical crosslinking distance constraints. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btz014

Fioramonte,M. *et al*. (2018) XPlex: an effective, multiplex cross-linking chemistry for acidic residues. *Anal. Chem*., **90**, 6043–6050.

Fischer,L. *et al*. (2013) Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. *J. Proteomics*, **88**, 120–128.

Fiser,A. (2010) Template-based protein structure modeling. *Methods Mol. Biol. Clifton N.J*., **673**, 73–94.

Fritzsche,R. *et al*. (2012) Optimizing the enrichment of cross-linked products for mass spectrometric protein analysis. *Rapid Commun. Mass Spectrom*., **26**, 653–658.

Götze,M. *et al*. (2012) StavroX—a software for analyzing crosslinked products in protein interaction studies. *J. Am. Soc. Mass Spectrom*., **23**, 76–87.

Herzog,F. *et al*. (2012) Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science*, **337**, 1348–1352.

Hoopmann,M.R. *et al*. (2015) Kojak: efficient analysis of chemically cross-linked protein complexes. *J. Proteome Res*., **14**, 2190–2198.

Huang,B.X. *et al*. (2004) Probing three-dimensional structure of bovine serum albumin by chemical cross-linking and mass spectrometry. *J. Am. Soc. Mass Spectrom*., **15**, 1237–1247.

Iacobucci,C. and Sinz,A. (2017) To be or not to be? Five guidelines to avoid misassignments in cross-linking/mass spectrometry. *Anal. Chem*., **89**, 7832–7835.

Kahraman,A. *et al*. (2013) Cross-link guided molecular modeling with ROSETTA. *PLoS One*, **8**, e73411.

Kalisman,N. *et al*. (2012) Subunit order of eukaryotic TRiC/CCT chaperonin by cross-linking, mass spectrometry, and combinatorial homology modeling. *Proc. Natl. Acad. Sci*., **109**, 2884–2889.

Kosinski,J. *et al*. (2015) Xlink analyzer: software for analysis and visualization of cross-linking data in the context of three-dimensional structures. *J. Struct. Biol*., **189**, 177–183.

Lima,D.B. *et al*. (2015) SIM-XL: a powerful and user-friendly tool for peptide cross-linking analysis. *J. Proteomics*, **129**, 51–55.

Luhavaya,H. *et al*. (2015) Enzymology of Pyran Ring A Formation in Salinomycin Biosynthesis. *Angew. Chem. Int. Ed*., **54**, 13622–13625.

Merkley,E.D. *et al*. (2014) Distance restraints from cross-linking mass spectrometry: mining a molecular dynamics simulation database to evaluate lysine–lysine distances. *Protein Sci. Publ. Protein Soc*., **23**, 747–759.

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

Ovchinnikov,S. *et al*. (2016) Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins Struct. Funct. Bioinform*., **84**, 67–75.

Pundir,S. *et al*. (2017) UniProt protein knowledgebase. In: *Protein Bioinformatics, Methods in Molecular Biology*. Humana Press, New York, NY, pp. 41–55.

Raman,S. *et al*. (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins*, **77**, 89–99.

Santos,D. *et al*. (2018) Enhancing protein fold determination by exploring the complementary information of chemical cross-linking and coevolutionary signals. *Bioinformatics*, **34**, 2201–2208.

Sarpe,V. *et al*. (2016) High sensitivity crosslink detection coupled with integrative structure modeling in the Mass Spec Studio. *Mol. Cell. Proteomics*, **15**, 3071–3080.

Schneider,M. *et al*. (2016) Blind testing of cross-linking/mass spectrometry hybrid methods in CASP11. *Proteins*, **84**, 152–163.

Schneidman-Duhovny,D. *et al*. (2012) Integrative structural modeling with small angle X-Ray Scattering profiles. *BMC Struct. Biol*., **12**, 17.

Sergey,O. *et al*. (2016) Improved *de novo* structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins Struct. Funct. Bioinform*., **84**, 67–75.

Simons,K.T. *et al*. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol*., **268**, 209–225.

Simons,K.T. *et al*. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins Struct. Funct. Bioinform*., **34**, 82–95.

Song,Y. *et al*. (2013) High-resolution comparative modeling with RosettaCM. *Structure*, **21**, 1735–1742.

Sugio,S. *et al*. (1999) Crystal structure of human serum albumin at 2.5 Å resolution. *Protein Eng. Des. Sel*., **12**, 439–446.

Tamò,G.E. *et al*. (2017) Assessment of data-assisted prediction by inclusion of cross-linking/mass-spectrometry and small angle X-ray scattering data in the 12th Critical Assessment of Protein Structure Prediction Experiment. *Proteins*, **86**, 215–227.

Tang,Y. *et al*. (2015) Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat. Methods*, **12**, 751–754.

Thompson,J.M. *et al*. (2012) Accurate protein structure modeling using sparse NMR data and homologous structure information. *Proc. Natl. Acad. Sci*., **109**, 9875–9880.

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