Methods 89 (2015) 79–90

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Protein structure prediction guided by crosslinking restraints – A systematic evaluation of the impact of the crosslinking spacer length

Tommy Hofmann [a,1], Axel W. Fischer [b,1], Jens Meiler [b,*], Stefan Kalkhof [a,c,*]

[a] Department of Proteomics, Helmholtz-Centre for Environmental Research – UFZ, Leipzig D-04318, Germany
[b] Department of Chemistry and Center for Structural Biology, Vanderbilt University, Nashville, TN 37232, USA
[c] Department of Bioanalytics, University of Applied Sciences and Arts of Coburg, D-96450 Coburg, Germany

ARTICLE INFO

Article history:
Received 15 February 2015
Received in revised form 21 April 2015
Accepted 12 May 2015
Available online 15 May 2015

Keywords:
Crosslinking
Protein structure prediction
De novo folding
Mass spectrometry
Protein modeling

ABSTRACT

Recent development of high-resolution mass spectrometry (MS) instruments enables chemical crosslinking (XL) to become a high-throughput method for obtaining structural information about proteins. Restraints derived from XL-MS experiments have been used successfully for structure refinement and protein–protein docking. However, one formidable question is under which circumstances XL-MS data might be sufficient to determine a protein's tertiary structure *de novo*? Answering this question will not only include understanding the impact of XL-MS data on sampling and scoring within a *de novo* protein structure prediction algorithm, it must also determine an optimal crosslinker type and length for protein structure determination. While a longer crosslinker will yield more restraints, the value of each restraint for protein structure prediction decreases as the restraint is consistent with a larger conformational space.

In this study, the number of crosslinks and their discriminative power was systematically analyzed *in silico* on a set of 2055 non-redundant protein folds considering Lys–Lys, Lys–Asp, Lys–Glu, Cys–Cys, and Arg–Arg reactive crosslinkers between 1 and 60 Å. Depending on the protein size a heuristic was developed that determines the optimal crosslinker length. Next, simulated restraints of variable length were used to *de novo* predict the tertiary structure of fifteen proteins using the BCL::Fold algorithm. The results demonstrate that a distinct crosslinker length exists for which information content for *de novo* protein structure prediction is maximized. The sampling accuracy improves on average by 1.0 Å and up to 2.2 Å in the most prominent example. XL-MS restraints enable consistently an improved selection of native-like models with an average enrichment of 2.1.

1. Introduction

'Structural Genomics' – the determination of the structure of all human proteins – would have profound impact on biochemical and biomedical research with direct implication to functional annotation, interpretation of mutations, development of small molecule binders, enzyme design or prediction of protein/protein interaction [1]. While significant progress towards this goal has been made through X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR), tertiary structure determination continues to be a challenge for many important human proteins. At present, high resolution structures exist for about 5% of all human proteins in the protein data bank (PDB) [2]. For many uncharacterized human proteins, construction of a comparative model is possible starting from the experimentally determined structure of a related protein. Nevertheless, for about 60% (~7800) of known protein families in the Pfam database [3] not a single structure is deposited [4]. Many of these proteins will continue to evade high-resolution protein structure determination.

Accordingly, researchers strive to develop alternative approaches. The most extreme approach includes computational methods that predict the tertiary structure of proteins from their sequence alone. While computational methods are sometimes successful at the predicting the tertiary structure of small proteins with up to one hundred residues [5], for larger proteins the size of the conformational space to be searched as well as the discrimination of incorrectly folded models hinder structure prediction [6–8].

Abbreviations: LC–MS, liquid chromatography–mass spectrometry; SSE, secondary structure element.

* Corresponding authors at: Helmholtz-Centre for Environmental Research – UFZ, Department of Proteomics, Permoserstr. 15, D-04318 Leipzig, Germany. Fax: +49 (341) 235 1786 (S. Kalkhof). Vanderbilt University, Nashville, TN 37232, USA. Fax: +1 (615) 936 2211 (J. Meiler).

E-mail addresses: jens@meilerlab.org (J. Meiler), stefan.kalkhof@ufz.de (S. Kalkhof).

[1] Contributed equally to this article.

http://dx.doi.org/10.1016/j.ymeth.2015.05.014
1046-2023/

However, recent studies demonstrate that combining *de novo* protein structure prediction with limited experimental data [9–13], i.e. experimental data that alone is insufficient to unambiguously determine the fold of the protein, can yield accurate models for larger proteins. The structural restraints in those studies were acquired using electron paramagnetic resonance (EPR) spectroscopy [9,14], cryo-Electron Microscopy (cryo-EM) [13,15], or NMR spectroscopy [12].

As an alternative technique, chemical crosslinking (XL) in combination with mass spectrometry (MS) can be applied to obtain distance restraints, which can be used to guide protein structure prediction (for review see [16–19]). Using bifunctional reagents with a defined length, functional groups within the protein can be covalently bridged in a native-like environment. Thus, it is possible to determine an upper limit for the distance between those residues after enzymatic proteolysis and identification of crosslinked peptides.

This method allows for a fast analysis of protein structures in a native-like environment at a low concentration and can even be applied to high molecular weight proteins [20], membrane proteins [21], or highly flexible proteins [22]. If combined with affinity purification it becomes possible to study proteins inside the cell [23]. Currently, the XL-MS technology is rapidly gaining importance driven by the liquid chromatography–mass spectrometry (LC–MS) instrument development, the generation of advanced analysis software [24], and the direct integration in protein structure prediction workflows [25–27]. Furthermore, hundreds of different crosslinking reagents with different spacer lengths, reactivities, and features for specific enrichment and improved detectability are now commercially available [28].

However, whereas the potential to combine XL-MS and computational modeling has been frequently demonstrated and many technical problems of XL-MS have been solved, several central questions have not yet been evaluated systematically.

(i) Crosslinking reagents are available with a spacer length ranging from 0 Å to more than 35 Å. Whereas longer reagents are likely to provide more distance restraints, shorter crosslinks have higher information content in *de novo* structure prediction as the conformational search space is more restricted. Thus, the question arises, which crosslinker spacer length supports structure prediction best?
(ii) Crosslinking results are often used to confirm already existing structures. However, what is the average gain in model accuracy and selection of correct models when using crosslinking data in conjunction with *de novo* protein structure prediction?
(iii) Crosslinking reagents vary in reactivity towards different functional groups present in different amino acids. For *de novo* protein structure prediction, which is the gain of using additionally crosslinker with different reactivities?

In this study, we simulated crosslinking experiments on more than 2000 non-redundant protein structures to determine the number of possible and structurally relevant crosslinks depending on the size of the protein as well as on the length and reactivity of the applied crosslinking reagents. We then tested the impact of crosslinking restraints on *de novo* protein structure prediction for fifteen selected proteins.

2. Experimental procedures

2.1. Software and databases

A subset of the PDB containing 2055 non-redundant protein structures was downloaded from the PISCES server (version 08.2012) [29]. This PDB subset was created by filtering all available structures with a resolution of at least 1.6 Å, a maximum sequence identity of 20% and an R-factor cutoff of 0.25. Euclidean distances and shortest solvent accessible surface (SAS) path lengths between Cβ–Cβ, Lys-Nz–Lys-Nz, Lys-Nz–Asp-Cγ, and Lys-Nz–Glu-Cδ, as well as Arg-N_{H2}–Arg-N_{H2} and Cys-S_G–Cys-S_G atom pairs with a maximal intramolecular distance of 60 Å were determined through the command line version of Xwalk [30].

2.2. Generation of sequence dependent distance functions

Tables containing the Euclidian distances and the sequence separation between crosslinking target amino acids ((i) Lys–Lys, (ii) Lys–Asp, (iii) Lys–Glu, (iv) Arg–Arg, and (v) Cys–Cys) were generated. Amino acid pair distances were sorted into 2.5 Å bins. The total number of observed pairs for each sequence and Euclidean distance was counted. Based on the result an approximation of the distance distribution for every sequence distance was created. The median of the distribution was determined. A logarithmic function was calculated as a regression curve in the form $E_{med} = a * \ln(S) + b$ to correlate sequence separation S to the median Euclidean distances E_{med}.

2.3. Calculation of the amino acid side chain length

Based on the structure of calmodulin (PDB entry 2ksz) the average Cβ–Nz, Cβ–Cγ, Cβ–Cδ, Cβ–N_{H2}, and Cβ–S_G distances of the sidechains of lysine, aspartic, glutamic acid, arginine, and cysteine were determined to be 4.5 Å, 2.3 Å, 3.6 Å, 5.1 Å, and 1.8 Å, respectively.

2.4. Distinguishing impossible, possible and structurally valuable crosslinks

Crosslinker spacer lengths between 1 and 60 Å distances were evaluated and classified in either (i) impossible crosslinks, meaning that the distance between the Cβ-atoms of the crosslinked amino acids exceeds the sum of the spacer lengths and the side chain lengths, or (ii) possible crosslinks, meaning that the Cβ–Cβ distance is below the sum of the spacer lengths and side chain lengths. The latter group was subdivided into crosslinks potentially useful for structure determination (valuable XL) and those that are unlikely to contribute much information (non-valuable XL). We defined crosslinks as valuable if the spacer length is shorter than the median distance expected for the given sequence separation by the equations derived in Section 2.2 (Fig. 2B, C and D). For these calculations, all proteins were grouped into 2.5 kDa bins. The calculations were performed for crosslinker lengths from 1 to 60 Å with a step size of 1 Å.

2.5. Estimation of optimal spacer lengths for a given protein molecular weight

Over all proteins in each molecular weight (MW) bin, the total number of possible distance pairs (#possible cross links) as well as the number of distance pairs useful for structure determination (#valuable cross links) were computed for each crosslinker spacer length. Furthermore, the maximum number of valuable crosslinks observed for all spacer length ($\#valuable_{max}$) was determined. For each MW bin the ratios ($\frac{\#valuable\ cross\ links}{\#possible\ cross\ links}$) and ($\frac{\#valuable\ cross\ links}{\#valuable_{max}}$) were plotted as a function of the crosslinker spacer length. The optimal crosslinker length for each MW bin was approximated as intersection points of the two functions using a local regression (see Fig. 4). The estimated values for the optimal crosslinker spacer length were plotted as a function of the MW and were fitted using a cubic

regression curve. The script used for the calculation is available at http://www.ufz.de/index.php?en=19910.

2.6. Simulation of crosslinking restraints

Seventeen proteins with known tertiary structure determined via X-ray crystallography (resolution of <1.9 Å) were selected from the dataset of structures as test cases to evaluate the influence of crosslinking restraints on *de novo* protein structure prediction. To thoroughly benchmark the algorithm, the benchmark set covers a wide range of protein topologies and structural features. The sequence lengths of the proteins range from 105 to 303 residues, the number of secondary structure elements ranges from 5 to 19 with varying α-helical and β-strand content (Table 1). For these proteins, all solvent accessible surface Cβ–Cβ distances between target amino acids in the structure which were within the range of either homobifunctional Lys-reactive crosslinkers or heterobifunctional Lys–Asp/Glu reactive crosslinkers were determined through Xwalk. For the predicted optimal crosslinker length (read above) and spacer lengths of 2.5, 7.5, 17.5, and 30 Å lists of structurally possible crosslinks were generated.

For the two proteins horse heart cytochrome c (PDB entry 1hrc) and oxymyoglobin (PDB entry 1mbo) restraints were also derived from published crosslinking MS experiments deposited in the XL database [25]. Experimental crosslinking data of FGF2 (PDB entry 1fga) and p11 (PDB entry 4hre) were derived from Young et al. [19] and Schulz et al. [31], respectively.

2.7. Translating crosslinking data into structural restraints

Explicitly rebuilding coordinates for a crosslink is comparable to solving the loop closure problem [32]. During *de novo*, protein structure prediction the crosslink would have to be reconstructed each time the conformation of the protein changes. In a typical Monte Carlo simulation with a maximum of 12000 Monte Carlo steps per model and 5000 models for each protein this would result in a maximum number of 60 million attempts to build the crosslink, which is too resource demanding for usage in *de novo* protein structure prediction. Therefore, we developed a fast approach to estimate the chance that a particular model fulfills a XL-MS restraint. The surface path of a crosslink is approximated by laying a sphere around the protein structure and computing the arc length between the crosslinked residues (Supplementary Fig. S3). The geometrical center of the protein structure is used as the center of the sphere. If takeoff and landing point have different distances to the center of the sphere, the longer distance is used as the radius. During the protein structure prediction process, the side-chains of the residues are not modeled explicitly but represented on a simplified way through a super atom. While this simplification vastly reduces the computational demand of the algorithm, it also adds additional uncertainty due to the unknown side-chain conformations. The agreement of the model with the crosslinking data is quantified by comparing the distance between the crosslinker lengths ($l_{XS} + l_{SS}1 + l_{SS}2$) with the computed arc lengths (d_{arc}), with −1 being the best agreement and 0 being the worst agreement. To account for the uncertainty of side-chain conformations a cosine-transition region of 7 Å was introduced (Supplementary Fig. S3).

2.8. Structure prediction protocol for the benchmark set

The protein structure prediction protocol is based on the BCL::Fold protocol for soluble proteins [33]. In a preparatory step, the secondary structure elements (SSEs) are predicted using the SSE prediction methods PsiPred [34] and Jufo9D [35] (Supplementary method S1) and an SSE pool is created (Supplementary method S2). Subsequently a Monte Carlo Metropolis energy minimization algorithm draws random SSEs from the predicted SSE pool and places them in the three-dimensional space. Random transformations like translation, rotation or shuffling of SSEs are applied. After each Monte Carlo step the energy of the resulting model is evaluated using knowledge-based potentials which, among others, evaluate the packing of SSEs, exposure of residues, radius of gyration, pairwise amino acid interactions, loop closure geometry and amino acid clashes [36] (Supplementary method S3). Based on the energy difference to the previous step and the simulated temperature a Metropolis criterion decides whether to accept or reject the most recent change.

The protein structure prediction protocol is broken into multiple stages, which differ regarding the granularity of the transformations applied, and the emphasis of different scoring terms. The first five stages apply large structural perturbations, which can alter the topology of the protein. Each of the five stages lasts for a maximum of 2000 Monte Carlo steps. If an energetically improved structure has not been generated within the previous 400 Monte Carlo steps, the stage terminates. Over the course of the five assembly stages, the weight of clashing penalties in the total score is ramped up as 0, 125, 250, 375, and 500.

The five protein assembly stages are followed by a stage of structural refinement. This stage lasts for a maximum number of 2000 Monte Carlo steps and terminates if no energetically improved model is sampled for 400 Monte Carlo steps in a row. Unlike the assembly stages, the refinement stage only consists of small structural perturbations, which will not drastically alter the topology of the protein model.

Through multiple prediction runs with different score weights, the optimal contribution of the crosslinking score to the total score was determined to be 40–50%. Consequently, the weight for the scoring term evaluating the agreement of the model with the crosslinking data was set to 300 over all six stages, which ensures that the crosslinking score contributes between 40% and 50% to the total score.

2.9. De novo folding simulations without and with crosslinking restraints

To evaluate the influence of crosslinking restraints on protein structure prediction accuracy, each protein was folded in the absence and in the presence of Lys–Lys, Lys–Glu, and Lys–Asp crosslinking restraints. Independent structure prediction experiments were performed for the predicted optimal as well as two shorter and two longer crosslinker spacer lengths each of the five spacer lengths (Table 2). Additionally, predictions were performed using combination of all spacer lengths as well as using restraints obtained by the optimal spacer length of all three crosslinker reactivities. For the two proteins of which experimentally determined crosslinking data were available, protein structure prediction was additionally performed for the experimentally determined restraints. For each protein and crosslinker length used, 5000 models were sampled in independent Monte Carlo Metropolis trajectories. Due to the randomness of the employed Monte Carlo algorithm, ten sets of 5000 models were sampled for each protein without restraints. Improvements in prediction accuracy can be compared to the standard deviations to identify statistically significant improvements (Table 3).

2.10. Metrics for comparing calculating model accuracy and enrichment

The results were evaluated using the RMSD100 [37] and enrichment [36] metrics. The RMSD100 metric was used to quantify the

Table 1
Chosen Proteins for modeling benchmark test. The fifteen proteins for the benchmark set were selected from high-resolution structures deposited in the Protein Data Bank. The structures were selected to cover a wide range of the structural feastures sequence length (#res), percentage of residues within SSEs (%res_{SSE}), number of SSEs (#SSEs), number of α-helices (#α) and number of β-strands (#β) while having a mutual sequence identity of less than 20%.

Structure	Uniprot	Resolution [Å]	Molecular weight [Da]	Sequence length [aa]	Lys portion [%]	α-helix [%]	β-sheet [%]
1hrc	**P00004**	**1.9**	**12368**	**105**	**18**	**40**	**1**
3iv4	Q7A6S3	1.5	13235	112	6	49	25
1bgf	P42228	1.45	14504	124	5	79	1
1t3y	Q14019	1.15	15835	141	9	35	29
3m1x	C4LXT9	1.2	15882	138	7	25	28
1x91	Q9LNF2	1.5	16419	153	7	76	0
1jl1	P0A7Y4	1.3	17483	155	7	34	30
1mbo	**P02185**	**1.6**	**17980**	**153**	**12**	**77**	**0**
2qnl	Q11XA0	1.5	19218	162	5	70	2
2ap3	Q8NX77	1.6	23190	199	23	81	0
1j77	Q9RGD9	1.5	24226	209	8	62	1
1es9	Q29460	1.3	25876	232	3	41	11
3b5o	D0VWS1	1.35	27506	244	3	71	0
1qx0	P0A2Y6	2.26	32821	293	7	38	20
2ixm	Q15257	1.5	34798	303	7	60	3
fgf2	**P09038**	**1.5**	**17859**	**145**	**10**	**9**	**34**
P11	**P60903**	**2.0**	**11071**	**95**	**13**	**63**	**3**

Table 2
Crosslinks obtained for the benchmark proteins. Simulated and experimentally determined crosslinks were obtained for the fifteen benchmark proteins. For each protein, an optimal spacer length was determined (optimal). Additional crosslinks were simulated for two shorter (short1 and short2) and two longer (long1 and long2) spacer. The number of yielded crosslinks (#rest) is shown for each spacer length.

Protein	Optimal		Short1		Short2		Long1		Long2	
	Length	#rest	Length	#rest	Length	#rest	Length	#rest	Length	#rest
1hrc	**10.2**	**13**	**2.5**	**0**	**7.5**	**7**	**17.5**	**27**	**30**	**107**
3iv4	10.4	5	2.5	2	7.5	2	17.5	7	30	13
1bgf	10.7	6	2.5	3	7.5	4	17.5	10	30	13
1t3y	10.9	35	2.5	9	7.5	20	17.5	42	30	63
3m1x	10.9	1	2.5	0	7.5	0	17.5	5	30	19
1x91	11	2	2.5	0	7.5	1	17.5	8	30	27
1jl1	11.2	7	2.5	0	7.5	3	17.5	11	30	24
1mbo	**11.3**	**9**	**2.5**	**0**	**7.5**	**3**	**17.5**	**23**	**30**	**77**
2qnl	11.5	6	2.5	4	7.5	4	17.5	8	30	15
2ap3	12.1	53	2.5	0	7.5	19	17.5	136	30	427
1j77	12.2	29	2.5	7	7.5	16	17.5	36	30	70
1es9	12.5	8	7.5	0	17.5	1	37.5	17	45	20
3b5o	12.7	15	7.5	2	17.5	8	37.5	21	45	25
1xq0	13.3	9	7.5	0	17.5	4	37.5	14	45	44
2ixm	13.5	41	7.5	20	17.5	41	37.5	49	45	57

sampling accuracy by computing the normalized root-mean square distance between the backbone atoms of the superimposed model and native structure. The enrichment metric was used to quantify the discrimination power of the scoring function by computing which percentage of the most accurate models can be selected by the scoring function. The enrichment metric is used to assess the influence of the crosslinking restraints to discriminate among the sampled models. First, the models of a given set S are sorted by their RMSD100 relative to the native structure. The 10% of the models in S with the lowest RMSD100 are assigned to subset P (positives) and the remaining 90% of the models are assigned to subset N (negatives). Second, the models in S are sorted by their BCL score. The 10% of the models in S with the best score are assigned to subset FS (favorable score). The intersection $TP = FS \cap P$ contains the most accurate models which the scoring function can select (true positives). The enrichment $= \frac{\#TP}{\#P} \cdot \frac{\#P+\#N}{\#P}$, with # denoting the size of the given sets, measures which ratio of the most accurate models the scoring function can select. In order to reduce the influence of the sampling accuracy on the enrichment values, the positive models are considered the 10% of the models with the lowest RMSD100 and $\frac{\#P+\#N}{\#P}$ is fixed at a value of 10.0. Therefore, the enrichment ranges from 0.0 to 10.0, with a score of 1.0 indicating random selection and a value above 1.0 indicating that the scoring function enriches for native-like models.

3. Results

3.1. Creation of an in silico crosslinking database

We performed *in silico* crosslinking experiments on 2055 non-redundant proteins. Covering a MW range from 1.4 to 139 kDa, 59% of the proteins have a MW below 25 kDa (Supplementary Fig. S1). For each of those proteins all Lys–Lys, Lys–Asp, and Lys–Glu sequence and Euclidean distances as well as the solvent accessible surface (SAS) distance between the Cβ-atoms were determined. Thus, the resulting database contained information on 391,902 Lys–Lys, 395,815 Lys–Glu, and 360,101 Lys–Asp pairs which built the basis for the determination of the number of possible crosslinks, crosslinks useful for structure prediction, and finally for the prediction of the optimal crosslinker length for studying a selected protein (Fig. 1A).

3.2. Estimation of the possible crosslinks per protein

Next we estimated how many and which of the distances could be crosslinked with a crosslinker of a given length and specificity. We considered crosslinks possible if the sum of the spacer length and the length of the two connected sidechains (Cβ–Cβ-, Lys-Nz–Lys-Nz, Lys-Nz–Asp-Cγ or Lys-Nz–Glu-Cδ) is longer than the Cβ–

Table 3
Comparison between structure prediction results with and without crosslinking restraints by using geometrical restraints obtained from crosslinking experiments, the size of the sampling space can be reduced resulting in an improved sampling accuracy. This is shown by significant improvements in the RMSD100 value of the most accurate model (best). Furthermore, crosslinking restraints provide geometrical information, which improves the discrimination power of the scoring function, leading to an improvement in the enrichment (e). Bold entries indicate proteins for which experimental data was available.

Protein	Without restraints		Optimal Lys/Lys		All Lys/Lys lengths		All reactivities	
	Best	*E*	Best	*e*	Best	*e*	Best	*e*
1hrc	**4.5**	**0.8**	**3.8**	**2.0**	**3.8**	**2.0**	**3.7**	**5.9**
3iv4	6.7	1.2	5.7	2.5	5.3	2.5	5.2	1.9
1bgf	6.6	1.0	5.7	2.1	4.9	2.4	6.2	1.6
1t3y	7.0	1.7	6.4	2.9	5.7	3.0	6.2	2.3
3m1x	3.8	0.7	3.8	0.7	3.6	1.5	3.6	1.7
1x91	4.8	2.0	4.8	2.0	2.0	3.2	2.1	3.5
1jl1	6.4	1.2	5.6	2.1	5.3	2.8	5.1	2.7
1mbo	**7.1**	**0.8**	**6.4**	**2.0**	**6.5**	**1.6**	**4.2**	**2.5**
2qnl	7.0	1.0	4.8	1.9	4.1	2.1	6.1	2.1
2ap3	2.5	1.6	2.0	3.0	1.6	3.1	2.2	2.0
1j77	6.8	0.5	5.0	2.0	4.0	2.4	3.8	3.2
1es9	7.3	1.1	5.7	2.1	5.6	2.8	6.3	2.9
3b5o	9.2	1.4	8.6	1.9	9.0	2.6	7.1	1.9
1xq0	9.9	1.1	8.3	1.9	8.5	2.4	7.4	2.1
2ixm	9.4	1.1	7.9	1.7	8.5	1.7	7.0	1.9
Ø	6.6	1.1	5.6	2.1	5.2	2.4	5.1	2.6

Cβ-SAS-distance between the amino acids. As the lengths of the sidechains of Lys (Cβ-Nz), Asp (Cβ-Oz), and Glu (Cβ-Oz) 4.5 Å, 2.4 Å, and 3.6 Å were used which were determined as average values from the crystal structure of Calmodulin (PDB entry 1cll). *In silico* crosslinking experiments were conducted for all of the 2055 proteins using homobifunctional Lys–Lys-reactive, as well as heterobifunctional (Lys–Asp- and Lys–Glu-reactive) crosslinking reagents with lengths from 1 to 60 Å (step size 1 Å).

To draw conclusions from the correlation of this *in silico* crosslinking experiments to the MW of the studied proteins the proteins were grouped into 45 bins with a step size of 2.5 kDa. For example, a protein with a MW in the range of 25–27.5 kDa contains on average 15.1 Lys, 14.4 Asp, and 16.7 Glu. On average 182 Lys–Lys, 173 Lys–Glu, and 144 Lys–Asp crosslinks exist per protein within this specific MW bin. Theoretically, all of those could be crosslinked with a crosslinker of 60 Å. In contrast by utilization of crosslinkers of 13 Å (as e.g. BS3) only about 33% of the crosslinks are formed *in silico*. When going to a crosslinker of length of 1 Å (e.g. close to EDC), only 10% of all possible amino acid pairs are linked.

3.3. Estimation of structurally relevant crosslinks

In protein structure prediction approaches, the enrichment of low RMSD structures among thousands of generated models is crucial. Therefore, we hypothesized that restraints that are valuable for structure prediction will reduce the conformational search space substantially. For the present study, we classify a crosslinking restraint as useful for structure prediction if it discriminate at least 50% of all possible conformations. Thus, in a second step each of the possible crosslinks was evaluated in terms of its potential to discriminate at least 50% of incorrect structures (useful for structure determination) or whether the crosslinked amino acids are so close in sequence that it can be derived from sequence separation that the distance can be bridged by the crosslinker independently of the protein's structure (not useful for structure determination).

In order to develop a stringent measure for usefulness we did not simply assume the maximum distance that can be bridged by an amino acid chain of a certain length. Rather the Euclidian distance distributions for Lys–Lys, Lys–Glu, and Lys–Asp were computed for the sequence separations ranging from 1 to 60 amino acids within our database of protein structures. For example, in the more than 2000 analyzed structures there are 3132 Lys–Lys pairs, which are separated by 10 amino acids. For this sequence distance Euclidian distances bins of 2.5 Å where defined in which the occurrences of residue pairs were counted. The pairs were present in bins ranging from 2.5 to 35.0 Å. As the median distance, we found 15.5 Å. For the same sequence distance the distribution of Lys–Glu (3336 pairs) and Lys–Asp (3010 pairs) are quite similar and the median values were 15.6 Å and 15.3 Å.

Similarly, for sequence separations of 15 amino acids we observed 3024 Lys–Lys pairs, 3200 Lys–Glu pairs, and 2835 Lys–Asp pairs. The median values are 20.8 Å, 20.9 Å, and 20.7 Å, respectively. For sequence separations of 60 amino acids, we observed 2167 Lys–Lys pairs, 2212 Lys–Glu pairs, and 2167 Lys–Asp pairs. The median values are 23.0 Å, 23.0 Å, and 23.0 Å, respectively (Fig. 2A).

Approximating the proteins structures as spheres, we applied a logarithmical model to fit the relationship between the sequence separation (S) and the median Euclidean distance (E_{med}) (Fig. 2B–D). We find (i) $E_{Lys-Lys} = 5.46^{*}\ln(S_{Lys-Lys}) + 2.2$, (ii) $E_{Lys-Glu} = 5.37^{*}\ln(S_{Lys-Glu}) + 2.36$, and (iii) $E_{Lys-Asp} = 5.19^{*}\ln(S_{Lys-Asp}) + 2.36$ for Lys–Lys, Lys–Glu, and Lys–Asp distances, respectively.

Secondly, using our derived functions constituting the S/E relationships, we considered every crosslink as of reasonable discriminative power, i.e. which fulfills the criterion that the sum of the crosslinker spacer length and the average length of both contributing side chains is shorter than the median of the sequence/Euclidean-distance distribution. If we examine the 25 kDa MW bins of Lys–Lys targets with a 1 Å spacer crosslink 1167 of the possible 22,398 target pairs fulfilled this criterion and were considered as of sufficient discriminative power (Fig. 3A). These crosslinks, which represent 4% of all Lys–Lys distances we defined therefore as useful for protein structure prediction. Application of a 13 Å spacer length results in 2935 valuable target pairs (12% of all Lys–Lys distances) (Fig. 3B). In contrast, a crosslinker with a spacer length of 60 Å would allow to crosslink all distances. However, none of the crosslinks would have discriminative power for native-like models (Fig. 3C). For the proteins of the 25 kDa MW bins the number of valuable crosslinks as a function of the crosslinker length has a log-normal character never exceeding a roughly 25 Å spacer. The intermediate length of 13 Å resulted in an almost equal contribution of valuable and structurally invaluable crosslinking pairs. Whereas 29% of all possible reactive amino acid pairs are linked, 12% are considered valuable according for structure prediction (Fig. 3).

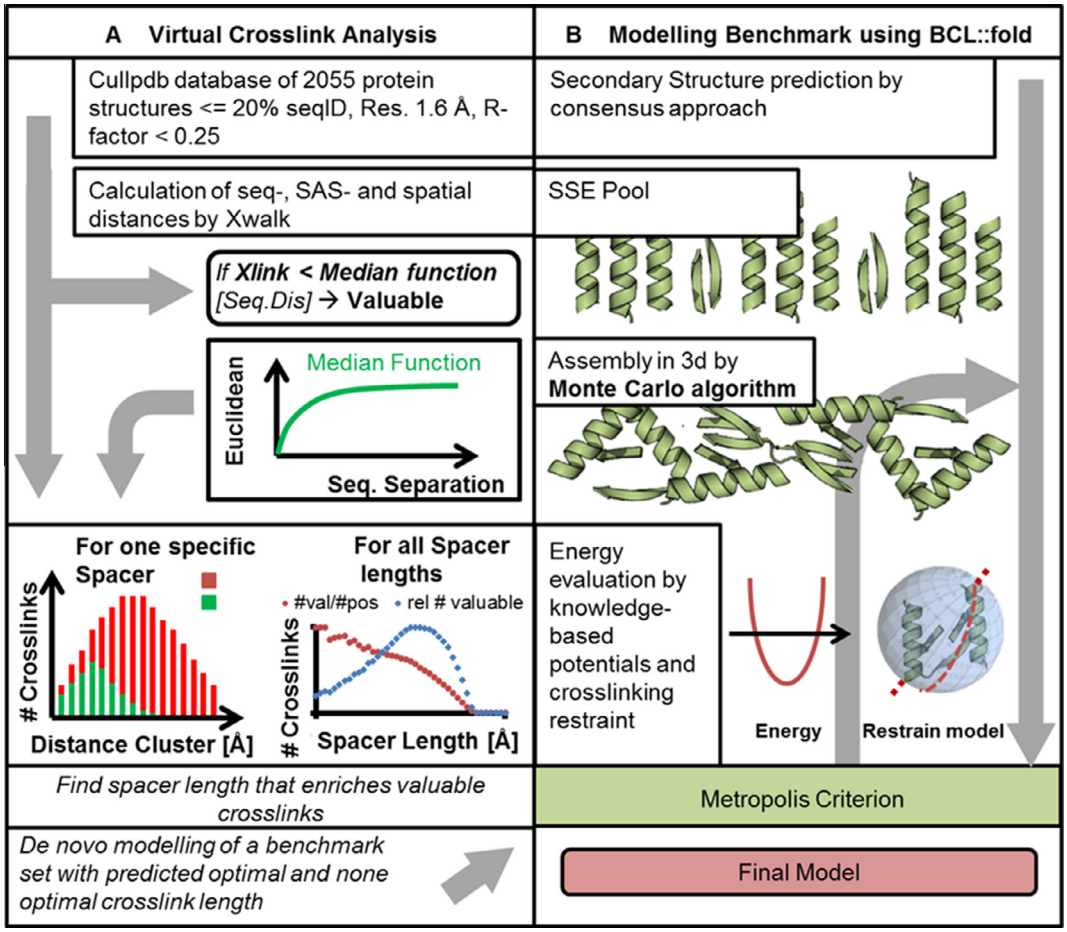

Fig. 1. Workflow for (A) the prediction of optimal crosslinker spacer length and (B) for *de novo* protein structure prediction using BCL::Fold. (A) Workflow for the prediction of the optimal spacer length depending on the MW of the protein of interest. (B) Workflow for *de novo* protein structure prediction using BCL::Fold. Secondary structure elements (SSEs) are predicted using PsiPred and Jufo9D. A Monte Carlo Metropolis algorithm subsequently searches the conformational space for the structure with most favorable score.

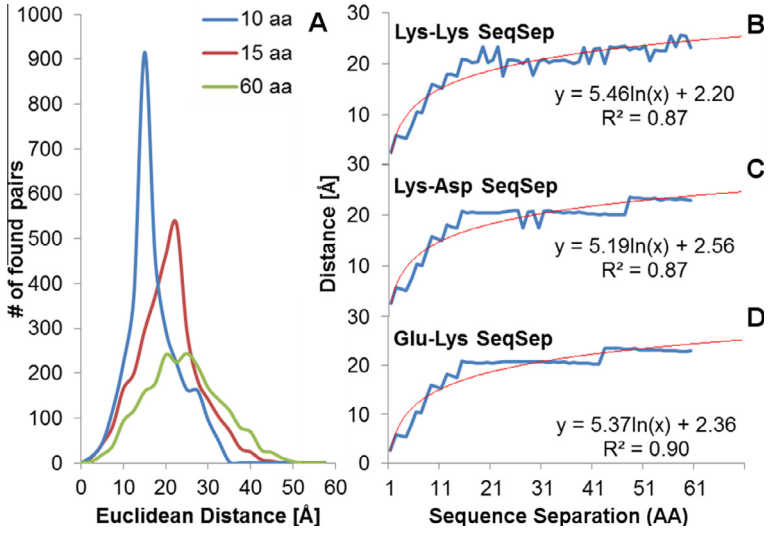

Fig. 2. (A) Distribution of the number of Lys–Lys pairs in respect to their Euclidean distance and (B–D) functions representing the relationship between sequence and spatial distance approximated by method of least squares to a logarithmic equation for (B) Lys–Lys, (C) Lys–Glu, and (D) Lys–Asp.

3.4. Prediction of MW dependent optimal crosslinker spacer lengths

Whereas usage of a short crosslinker will result in only a few but mostly structurally valuable restraints, a longer crosslinker will yield more restraints but a lower ratio of valuable restraints. Furthermore, the ratio of valuable restraints as well as the number of possible restraints depends on the size of the protein. In agreement with prior studies regarding structural modeling driven by

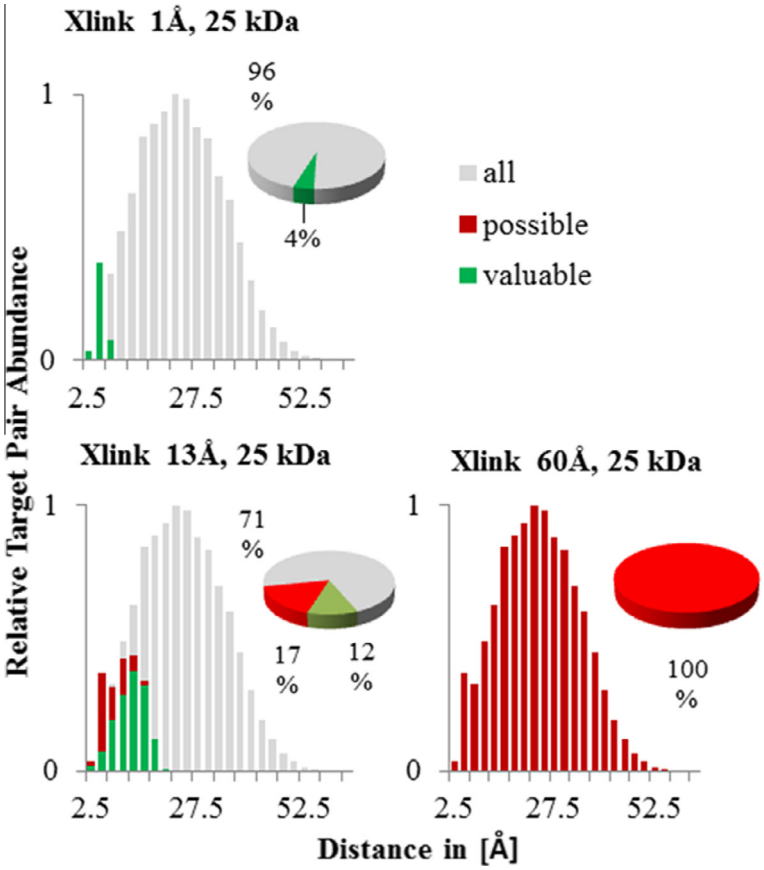

Fig. 3. Distribution of all possible and valuable Lys–Lys pairs for a 25–27.5 kDa weight bin. Gray bars show all theoretical pairs in their specific distance cluster of ±2.5 Å. Red bars show pairs that could be connected in respect to their surface distance by a specific crosslink (here 1 Å, 13 Å and 60 Å) always including the side chain contribution to the overall length. Green bars show pairs that are considered valuable by our proposed scoring function. Pie charts show the accumulated number of crosslinks for every spacer length.

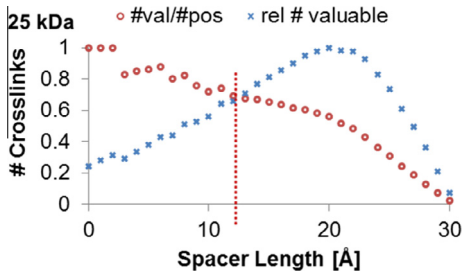

Fig. 4. Behavior of valuable and possible crosslinks in the MW bin 25 kDa and localization of the optimal spacer length. Shown is the number of valuable crosslinks for every tested spacer length in red. These values are normalized to a dimension spanning 1. Blue points show the share of valuable crosslinks among the physical possible ones. The dotted line meets the intersection of both curves and represents the optimal spacer length where the best ratio between valuable and possible crosslinks is attained and the number of valuable crosslinks is maximized in respect to this ratio.

sparse distance restraints [38], we hypothesize that a compromise between maximizing the portion of valuable crosslinks compared to all crosslinks which can be formed with a given crosslinker length $\left(\frac{\#\text{valuable cross links}}{\#\text{possible cross links}}\right)$ and maximizing the relative number of valuable crosslinks compared to the maximal number of achievable valuable crosslinks with any spacer length $\left(\frac{\#\text{valuable cross links}}{\#\text{valuable}_{\max}}\right)$ might yield the best results.

Following our hypothesis, for each MW bin we derived the optimal spacer length as the intersection point of the two functions as it is shown exemplarily for MW 25 kDa in Fig. 4.

The derived optimal spacer lengths for Lys–Lys, Lys–Asp, and Lys–Glu were plotted as function of the MW (Fig. 5A–C). The relationship was fitted using a cube root function. For our observable MW sample space for Lys–Lys crosslinks, all spacer lengths reached dimensions between 5.0 Å and 18.6 Å. No optimal spacer length was further than 2.5 Å separated from the regression curve. The average distance from the modeled spacer lengths was 0.7 Å. The MW term as well as the side chain term has been modeled as an exponential fraction in respect to the relation between volume and distances in spherical objects.

Additionally, the optimal spacer lengths were also predicted for homobifunctional arginine and for homobifunctional cysteine crosslinking reagents analogously to the procedure being descript for the homo- and heterobifunctional lysine-containing crosslinks. (Supplementary Fig. S2). Consistently, the optimal spacer lengths depend on the molecular weight MW as well as the lengths of the crosslinked sidechains SS1 and SS2 and could be calculated by $l_{\text{opt.}}[\text{Å}] = k * \sqrt[3]{\text{MW}} + \sqrt[3]{\text{SS1} + \text{SS2}}$. k was determined to be 0.32, 0.31, 0.34, 0.34, and 0.35 for Lys–Lys, Lys–Asp, Lys–Glu, Arg–Arg, and Cys–Cys, respectively.

3.5. Generation of in silico and experimental crosslinking data for testing the effect of different spacer length for de novo modeling

To evaluate the effect of crosslinking data derived from experiments with different spacer length we folded 17 proteins *de novo* with BCL::fold (Fig. 1B). Thirteen proteins were part of our dataset while for four proteins experimental crosslinking data were available (3mbo, 1hrc, 1fga, and 4hre) (Table 1). All Proteins have a MW between 13 to 27 kDa. Most structures were mainly α-helical with

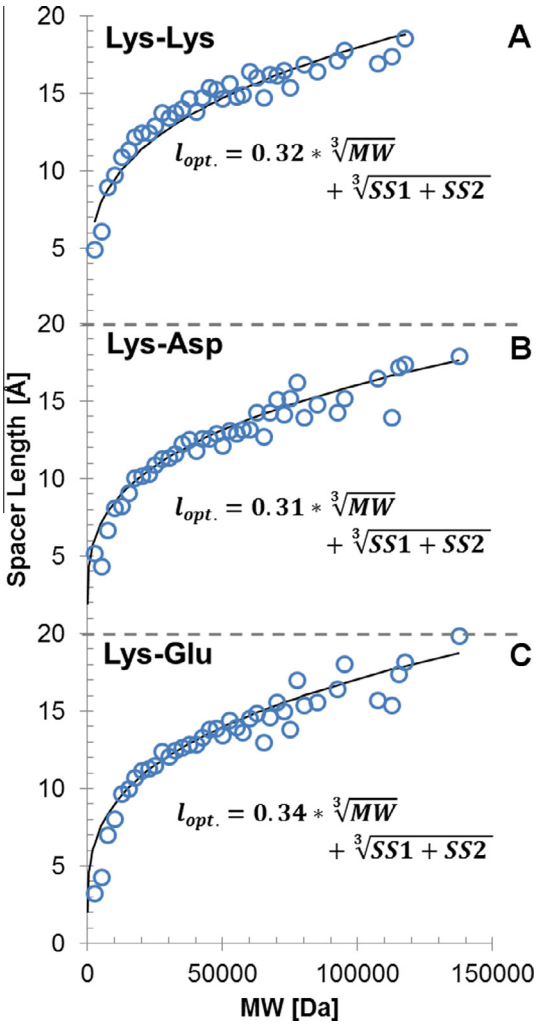

Fig. 5. Functions representing the relationship between sequence (S) and spatial distance (E). The equations approximated by method of least squares to a logarithmic equation for (A) Lys–Lys, (B) Lys–Glu, and (C) Lys–Asp.

fewer β-strand secondary structure elements. The β-sheet content ranged from 0% to 51%. The α-helical content ranges from 2% to 81%. The highest β-sheet content showed 1lmi with also the fewest α-helices. The portion of lysines was between 3% and 23%, which resulted in minimal 4 and maximal 46 lysine residues per structure. For the two structures 3mbo and 1hrc, which were studied experimentally, we used the published experimental data, which were obtained using DSG, DSS/BS3, and DEST [25]. For 3mbo, there were 8 crosslinks in total with the 11.4 Å reagent BS3 four of them confirmed with the 7.7 Å reagent DSG. For 1hrc, 48 crosslinks were reported. 9 DSS, 31 BS3, 6 DSG and 9 with DEST (11 Å). Six crosslinks had been identified with different crosslinking reagents. 18 BS3 crosslinks were published for 1fga [19] whereas 3 intramolecular BS3 crosslinks were available for 1hre [31]. For the thirteen proteins as well as for 3mbo and 1hrc, we predicted all crosslinks, which are possible with the predicted optimal crosslinker length as well as with two shorter and two longer crosslinking reagents (Table 2) and used these data as restraints during modeling (Supplementary Fig. S3).

3.6. Crosslinking restraints improve the sampling accuracy of de novo protein structure prediction

XL-MS data provides structural restraints that reduce the sampling space in *de novo* structure determination. Thereby a fraction of incorrect conformations is excluded and the sampling density in all other areas of the conformational space is increased. To evaluate the power of crosslinking restraints to guide *de novo* protein structure determination we computed the RMSD100 [37] values of the most accurate models (best) for each protein for structure prediction with and without crosslinking restraints. Using crosslinking restraints not only increases the frequency with which accurate models are sampled, but the best models achieve an accuracy not sampled in the absence of crosslinking data (Table 3). Across all benchmark proteins, the accuracy of the best models was, on average, 6.6 Å when no crosslinking data was used. By using crosslinking, data for the spacer length deemed optimal the average RMSD100 value was improved to 5.6 Å, which corresponds to two standard deviations. By using restraints obtained for all five spacer lengths, the average accuracy of the best model improved to 5.2 Å. For the proteins 1xq0, 2ixm, and 3b50 even with crosslinking data, it was not possible to sample a native-like conformation. We attribute this to limitations in the sampling algorithm resulting in the native conformation not being part of the sampling space. For other proteins, significant improvements could be observed. While the accuracy of the best models for 1es9 and 1j77 was 7.3 Å and 6.8 Å, crosslinking restraints improved the accuracy to 5.7 Å and 4.5 Å, respectively. For 1mbo, the accuracy could be increased from 7.1 Å to 4.2 Å by using a combination of Lys–Glu/Asp reactive crosslinkers (Fig. 6).

3.7. Crosslinking restraints improve the discriminative power of the scoring function

The ability of the scoring function to identify the most accurate models among the sampled ones was quantified using the enrichment metric (see Methods). Enrichments were computed for proteins predicted without crosslinking data, for each spacer length and for all spacer lengths combined. For protein structure prediction without crosslinking restraints an average enrichment of 1.1 was measured, which is barely better than random selection. The scoring function has almost no discriminative power. Using crosslinking restraints yielded by the optimal spacer length improved the enrichment to 2.1 (Table 3), which corresponds to three standard deviations. Using all five spacer lengths to obtain additional restraints further improves the enrichment to 2.4. The most significant improvement could be observed for 1j77, for which the enrichment could be improved from 0.5 to 2.4.

3.8. The crosslinker length determines improvements in sampling accuracy and discrimination power

The length of the crosslinker determines the number of obtainable restraints as well as their information content [9]. While a longer crosslinker is able to form more crosslinks and therefore yields a larger number of restraints, the longer crosslinker length can be fulfilled by a larger number of conformations, reducing the discriminative power of the restraint. In order to assess the influence of the crosslinker length, and therefore the number of restraints and restraint distances, on the sampling accuracy and discrimination power, the protein structure prediction protocol was conducted with restraints derived from different crosslinker lengths.

The crosslinker lengths were separated into five groups: *optimal*, which is the predicted optimal crosslinker length, *short1* and *short2*, which are shorter crosslinker lengths, and *long1* and *long2*, which are longer crosslinker lengths. The crosslinker length predicted to be optimal yielded the most useful restraints for protein structure prediction in terms of sampling accuracy and discriminative power. Across all proteins the average RMSD100 values of the most accurate models for the optimal crosslinker length were 5.6 Å

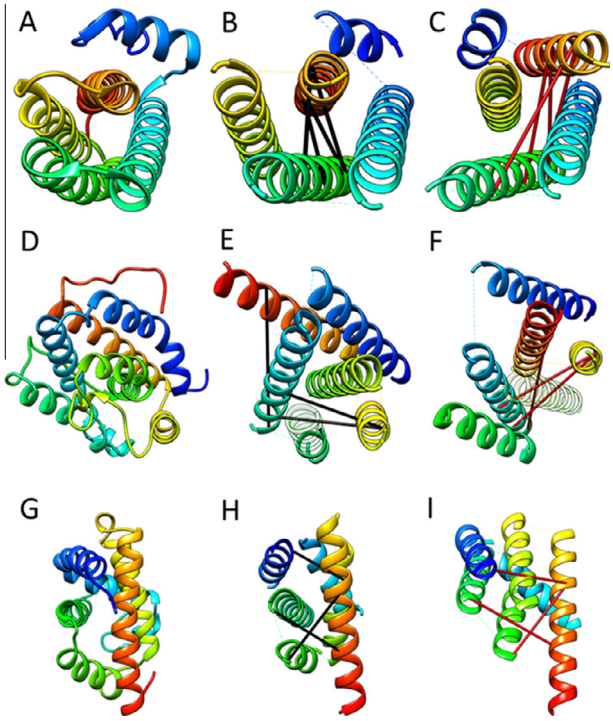

Fig. 6. Most accurate models sampled with and without using crosslinking restraints. The RMSD100 values of the most accurate models sampled for 1x91, 1j77, and 1mbo were 4.8 Å, 6.8 Å, and 7.1 Å. By using restraints yielded by Lys–Lys/Asp/Glu reactive crosslinkers, the accuracy could be improved to 2.7 Å, 5.0 Å, and 4.2 Å. Shown are the native structures of 1x91, 1j77, and 1mbo (A, D, G), the most accurate models sampled without crosslinking restraints (B, E, H), and the most accurate models sampled with crosslinking restraints (C, F, I). Selected restraints are shown that are not fulfilled in the model predicted without crosslinking data (red bars), but that are fulfilled in the model predicted with crosslinking data (black bars).

– an improvement by 15% – while they were 6.3 Å, 6.2 Å, 5.9 Å and 6.1 Å – improvements by 5%, 6%, 11% and 8% – for the shorter and longer crosslinker lengths, respectively (Fig. 7A, Supplementary Table S1). The longest crosslinkers have a less significant impact on the sampling accuracy due to their ambiguity, while the shortest crosslinkers failed to yield a sufficient number of distance restraints to impact prediction accuracy. The discriminative power, quantified through the enrichment metric, for the optimal crosslinker length was 2.1, while it was 1.4, 1.5, 1.9, and 1.7 for the shorter and longer crosslinkers, respectively (Fig. 7B). For the proteins 1x91 and 3m1x the optimal crosslinker length did not yield any crosslinks with a sequence separation of at least ten and therefore did not provide relevant structural information. In those cases protein structure prediction with longer crosslinker lengths provided better results. By combining restraints obtained for all five crosslinker lengths, the average enrichment value could be improved to 2.4.

3.9. Combination of crosslinkers with different reactivities results in improvements larger than seen when varying the spacer lengths

In order to obtain valuable restraints for *de novo* protein structure prediction a maximum number of SSE pairs needs to be crosslinked. The availability of Lys–Asp/Glu reactive crosslinkers allows for a better sequence coverage and therefore a wider coverage of SSE pairs. Crosslinks with different spacer lengths were simulated for the proteins in the benchmark set using Xwalk (Supplementary Table S2). To assess the influence of Lys–Asp/Glu reactive crosslinkers on protein structure prediction the same protocol was applied as for the Lys–Lys reactive crosslinkers. For the Lys–Glu reactive crosslinkers a prediction accuracy of 5.7 Å and enrichment of 2.2 on average could be achieved, which is comparable to the results for the Lys–Lys reactive crosslinkers (Supplementary Table S3).

While Lys–Asp reactive crosslinkers also achieve improvements in prediction accuracy and enrichment when compared to protein structure prediction without restraints, the results are slightly worse than for Lys–Lys reactive crosslinkers with an average prediction accuracy of 6.0 Å versus 5.6 Å and an average enrichment of 1.7 versus 2.1 (Supplementary Table S3). To a large part, the difference in the overall results is caused by the results for the proteins 1es9, 1t3y, and 3m1x for which Lys–Asp reactive crosslinkers failed to yield restraints between SSE pairs and therefore failed to reduce the conformational space significantly. Besides deviations regarding the average improvements over all proteins, the spacer length deemed optimal also provides the best results for Lys–Asp/Glu reactive crosslinking (Supplementary Tables S4 and S5). Combining the restraints yielded for the optimal spacer lengths with Lys–Lys/Asp/Glu reactive crosslinks improves the sampling average sampling accuracy to 5.1 Å and the average enrichment to 2.6. Combining the restraints yielded by all spacer lengths and crosslinker reactivities failed to further improve prediction results.

4. Discussion

4.1. Prediction of the optimal crosslinker spacer length

It has been shown frequently that chemical crosslinking data can be used to guide *de novo* structure prediction and selection of native-like models. Surely, the sensitivity, the broad applicability to almost all proteins, the nearly physiological experimental condition during the chemical crosslinking reaction, and the potential of automation are the main advantages for using XL-MS to generate such restraints. However, the small number and high uncertainty of restraints from chemical crosslinks limit impact on *de novo* proteins structure prediction, in particular when compared to more data rich methods such as NMR spectroscopy [12].

One major limitation is the fact that distances between the functional groups in long and flexible amino acid sidechains are measured. Therefore, a significant uncertainty is added to the crosslinker length when converting XL-MS data into Cβ–Cβ restraints. A second challenge of chemical crosslinks is that only the maximum distance is restricted, but no information is obtained on the minimum distance or the favored distance distribution. In result, even a "zero length" crosslinker restricts is the Cβ–Cβ distance to the sum of the lengths of the two connected sidechains (e.g. for Lys–Lys crosslinks 9.1 Å).

In most of the crosslinking experiments, lysine residues are targeted. Lysines are excellent targets because of their overrepresentation on protein surfaces and the clean chemistry of amine modification. Nevertheless, their frequency is on average only about 7%. As a consequence the number of crosslinks which can be formed e.g. in a 25 kDa protein with a standard homobifunctional Lys–Lys-reactive crosslinking reagents with a spacer length of 6.4 Å (length of DST) are in the range of about 20. Only a small fraction of these restraints will substantially limit the conformational space for the protein. This number is usually too small to restrict the conformational space to an unambiguous single protein fold. To increase the number of restrains it is possible to use crosslinkers with longer spacer length or target amino acids such as Asp, Glu, Tyr, Ser, Thr, Arg, or Cys.

Restraints obtained with longer crosslinking reagents are less restrictive to the conformational space. To evaluate the value of

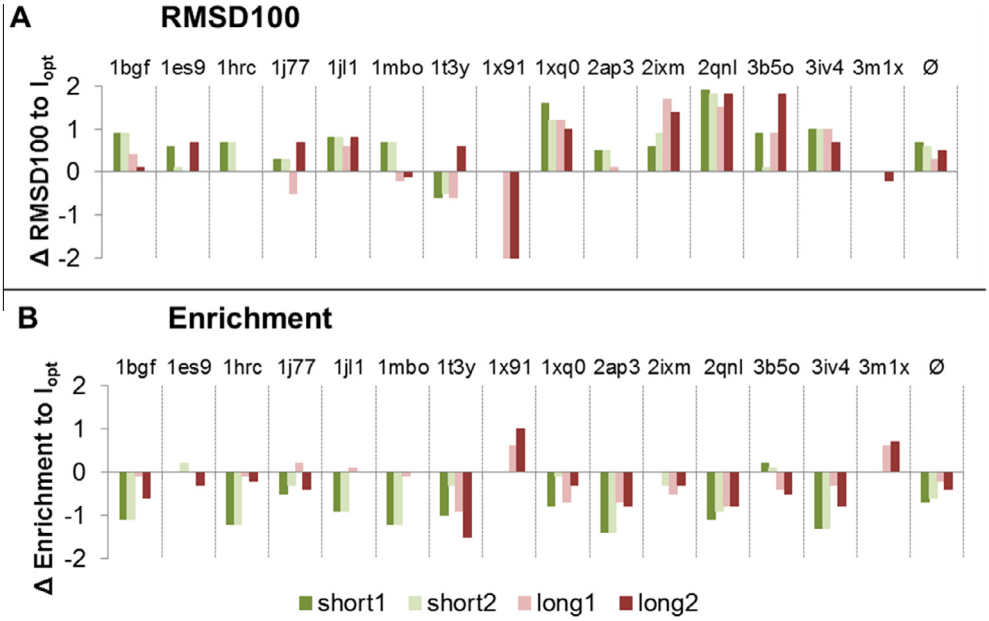

Fig. 7. Crosslinking data improve prediction accuracy and discrimination power. Using geometrical restraints derived from crosslinking experiments reduces the size of the conformational space, which needs to be searched for the conformation with the lowest free energy. This results in a higher likelihood of sampling accurate models and an improved discrimination power of the scoring function. Panel A compares the RMSD100 values of the most accurate model for structure prediction from different spacer lengths to the results for the optimal spacer length (horizontal line). Panel B compares the enrichments for different spacer lengths likewise.

crosslinks for protein structure prediction we determined for each sequence distance (0–60 amino acids) how long a crosslinker has to be to link the target amino acids. For example two lysines, which are separated by eight amino acids in sequence were found to be linkable in all 3488 cases by a homobifunctional crosslinker with a length of >30 Å (as it is in BS(PEG)9). In our study, we stated the hypothesis that it would be desirable if two target amino acids can only be linked in 50% of all models created meaning that 50% of all structures could be discarded based on a single crosslink. E.g. for two lysines separated by 10 amino acids this would be the case for crosslinker lengths of 14.8 Å (for further amino acids distances see Fig. 2B). Crosslinks, which could only be formed in less than 50% for the corresponding sequence distance, were considered as being valuable. Based on this definition for all 2055 structures of the applied non-redundant protein structure database the optimal spacer length was calculated. With this optimal spacer length, the number of structural valuable crosslinks has been maximized taking into account that in general for modeling approaches few distance restraints of highly discriminative character are less favorable than a higher number with a smaller discriminative power [27,38].

Since the optimal crosslinker length should depend on the protein size in a cubic root fashion to convert volume into distance, it is not unexpected that this was also observed for the dependency on the MW (Fig. 5). However, one has to keep in mind that the formula might not be applicable to non-globular proteins and multi-domain proteins. However, in case of multidomain proteins this formula should be applicable to the separated domains. Remarkably, based on our simulation for proteins with a MWs of 10, 25, 50, and 100 kDa the recommended spacer length are 9.0, 11.5, 13.9 and 17.0 Å, respectively, which is quite close to the homobifunctional amine-reactive commercially available crosslinkers DSG (7.7 Å), BS3 (11.4 Å) and EGS (16.1 Å) which are currently the preferred choice to study small (<20 kDa), medium (20–50 kDa) and large proteins (>50 kDa), respectively.

Addressing different functional groups is a second approach to increase the total number of distance restraints. The consequence is that the crosslinking reaction is either less effective or specific (Asp, Glu, Tyr, Ser, Thr) creating challenges in data interpretation or the target amino acids are less frequent (Arg and Cys) limiting the number of restraints observed. However, using the same theoretical approach revealed that optimal spacer length for heterobifunctional Lys–Asp and Lys–Glu crosslinker (Fig. 5) as well as homobifunctional Cys–Cys and Arg–Arg crosslinker (Supplementary Fig. S2) can be predicted with the same equation: $l_{\text{opt.}}[\text{Å}] = k * \sqrt[3]{\text{MW}} + \sqrt[3]{\text{SS1} + \text{SS2}}$ with $k \approx \frac{1}{3}$ in which SS1 and SS2 are the average lengths of the crosslinked sidechains.

Comparing the two approaches to increase the number of valuable crosslinks, it should be pointed out that using several crosslinking reagents with different reactivities results in significantly higher improvement of the model quality than using only lysine reactive crosslinking reagent but with different spacer length.

4.2. Challenges in using crosslinking data to guide de novo modeling

To test whether the crosslinker with the predicted optimal spacer length indeed perform best in modeling we have chosen a *de novo* structure prediction approach BCL::Fold for testing. Even though comparative modeling using known protein structures as a template usually performs better then *de novo* modeling, our goal was to maximize impact of XL-MS restraints.

A major limiting factor for *de novo* protein structure prediction is the vast size of the conformational space. Crosslinking restraints can aid the computational prediction of a protein's tertiary structure by drastically reducing the size of the sampling space. Crosslinking experiments yield a maximum Euclidean distance between the crosslinked residues, which increases the sampling density in the relevant part of the conformational space.

A major limitation of using crosslinking restraints to guide protein structure prediction when compared to restraints obtained from EPR and NMR spectroscopy is that the crosslinker length cannot be directly translated into a Euclidean distance between the crosslinked residues. While crosslink prediction and evaluation methods like Xwalk [22] are successful at predicting if a certain crosslink can be formed in a given structure, explicit modeling approaches are computationally too expensive for usage in a rapid

scoring function required for protein structure prediction. Approximations, such as the great circle on a sphere presented here, are fast to compute but associated with increased uncertainties. Most of the crosslinkers used can cover a long Euclidean distance and therefore the yielded restraints can be fulfilled by a wide variety of conformations. In spite of this, crosslinking restraints display some potential in limiting the size of the sampling space, resulting in a higher density of accurate models. Further, the geometrical restraints derived from XL-MS allow for the discrimination of a significant fraction of models representing incorrect topologies and therefore improve the discriminative power of the scoring function.

4.3. Abilities and limitations of protein structure prediction from limited experimental data

We showed that incorporation of crosslinking data into a *de novo* protein structure prediction method improves the accuracy of the structure prediction. The two major challenges of *de novo* predictions are the sampling of structures as well as the discrimination of inaccurate structures. In this study reduction of the conformational space was achieved through the assembly of predicted SSEs with limited flexibility and the incorporation of geometrical restraints derived from crosslinking data. The discrimination of inaccurate models is performed through a scoring function which approximates the free energy. Assuming that the native structure is in the global energy minimum, complete sampling and an accurate methods to measure free energy would lead to the correct identification of the native conformation. However, the conformational space is too large to be extensively sampled and the free energy needs to be approximated, which results in ambiguity regarding the model which is most similar to the native structure. Incorporating crosslinking data provides geometrical restraints which can be used as additional criteria to discriminate inaccurate models. While an average sampling accuracy of 5.1 Å, when using restraints yielded XL-MS, is a significant improvement over the 6.6 Å, when not using crosslinking data at all, only for four proteins it was possible to sample models with an RMSD100 of less than 4 Å when compared to the crystal structure. Crosslinking data yields an upper boundary for the Euclidean distance of the crosslinked residues, which allows for the placement of the second residue within a sphere of volume $4/3\pi r^3$ around the first residue. Depending on the crosslink distribution, topologically different models can fulfill the same restraint set. Discrimination among those models is impossible with XL-MS restraints.

4.4. Comparison of experimental and in silico crosslinks

In order to draw general conclusion based on the analysis of hundreds of different structures this study relies mainly on virtual crosslinking experiments. Unfortunately, although extensive XL-MS datasets have been published for several proteins, it proved difficult to obtain suitable experimental datasets for the present benchmark due to additional requirements: (i) the protein must be monomeric and small enough for *de novo* protein folding with BCL::Fold (ii) an experimental atomic detail structure for comparison and (iii) a large dataset of intramolecular crosslinks must be available. Results for the four cases p11, FGF2, cytochrome c, and oxymyoglobin that came closest are reported to demonstrate our efforts to work not only with simulated data. However, for p11 and FGF2 using experimentally determined restraints did not improve the prediction results in a statistically significant way. For p11, only three restraints were available with a maximum sequence separation of nine residues. Because of the small sequence separation, these restraints contain very limited structural information and no improvement in *de novo* folding can be expected. The tertiary structure of FGF2 contains twelve β-strands with several β-strands that are strongly bent. This protein is too large for *de novo* structure determination with BCL::Fold. As BCL::Fold is unable to sample the conformation of the protein in the first place, no significant improvement was expected or observed when XL-MS data were added. Nevertheless, the value of the predicted crosslinks in comparison to experimental crosslinks could be validated with the two proteins cytochrome c and oxymyoglobin for which experimental crosslinks had been published in the XL database [16]. For cytochrome c (PDB entry 1hrc), we indeed found that the crosslinker with predicted optimal spacer length of 10.2 Å performed best. However, for oxymyoglobin (PDB entry 1mbo) the longer spacers improved the accuracy slightly more than the crosslinker with the optimal spacer length. Interestingly, on the one hand for both proteins several crosslinks, which should be possible, could not be detected, which might be due to experimental or analytical reasons. On the other hand, also several crosslinks, which were experimentally, identified which were not predicted. An examination of these data revealed that most of these crosslinks are not present in the virtual data set because their Cβ–Cβ distances exceed the expected maximum length. This finding is in agreement with Merkley et al. [39] who evaluated protein structures by molecular dynamics and reported that usually a high number of experimental approved crosslinks exceed the theoretical maximal spatial distance due to structure flexibility. It was concluded for the investigation of Lys–Lys distances using a BS3/DSS crosslinking reagent an upper bound of 26–30 Å for Cα atoms [39].

On the other hand, spacer conformations usually adapt lengths that are somehow distributed between their minimal and maximal lengths. In line it was also reported that many spacers in commercially available crosslink agents preferable adopt conformations which are significantly below the cited maximal spacer length [40]. Thus, ideally crosslinking results should be evaluated based on experimentally derived or simulated ensembles of in-solution structures instead of using X-ray structures as reference. However, to address all degrees of flexibility during the *de novo* structure prediction is currently too resource intensive. Furthermore, there are many additional practical challenges, which may prevent the formation or identification of crosslinks, and thus may result in more meaningful results using a crosslinker with a non-optimal length. Nevertheless, for both structures the sampling accuracies could also be improved by 0.7 Å based on the experimental restraints, which is only slightly worse than the improvement of 1.0 Å observed based on *in silico* crosslinks.

5. Conclusion

Recent development of high-resolution MS instruments enables the analysis of proteins not accessible to NMR spectroscopy and X-ray crystallography. Data obtained from those experiments can be translated into structural restraints to guide protein structure prediction. The information content of a geometrical restraint obtained from XL-MS experiments is directly dependent on the used spacer length. Thus, the choice of the spacer length is an important step.

Firstly, for amino acids pairs close in sequence only minimum structural information is obtained if the spacer is too long. Here we determine the optimal spacer length to gain structural information on lysines with a sequence separation of X, we estimated a length as $E = 5.5 * \ln(S) + 2.2$. Secondly, we demonstrate that for *de novo* protein structure prediction the optimal spacer length depends on the MW of the protein of interest and the length of the crosslinked sidechains (ss) and can be predicted as $l_{\text{opt.}}[\text{Å}] = k * \sqrt[3]{\text{MW}} + \sqrt[3]{\text{SS1} + \text{SS2}}$ with $k \approx \frac{1}{3}$.

We also demonstrate that restraints obtained from crosslinking experiments contribute moderately to solving the major challenges of *de novo* protein structure prediction – the vast size of the conformational space and discrimination of inaccurate models. Using restraints from crosslinking experiments significantly increases the sampling density of native-like models and contribute to the discrimination of incorrect models. By combining crosslinking restraints with knowledge-based scoring functions, the average accuracy of the sampled models could be improved by up to 2.2 Å and the average enrichment of accurate models could be improved from 11% to 24%.

Conclusively we believe this study can help in the planing of XL-MS experiments as well as to evaluate how much information can be gained by XL-MS experiments and the ambiguity that remains.

Acknowledgments

This study was supported by grants from Deutsche Forschungsgemeinschaft Transregio 67 (subproject Z4) and ESF Investigator group GPCR 2. Work in the Meiler laboratory is supported through NIH (R01 GM080403, R01 GM099842, R01 DK097376) and NSF (CHE 1305874). This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ymeth.2015.05.014.

References

[1] D. Baker, A. Sali, Science 294 (2001) 93–96.
[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Res. 28 (2000) 235–242.
[3] M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn, Nucleic Acids Res. 40 (2012) D290–D301.
[4] K. Khafizov, C. Madrid-Aliste, S.C. Almo, A. Fiser, Proc. Natl. Acad. Sci. U.S.A. 111 (2014) 3733–3738.
[5] J. Moult, Curr. Opin. Struct. Biol. 15 (2005) 285–289.
[6] R. Bonneau, I. Ruczinski, J. Tsai, D. Baker, Protein Sci. 11 (2002) 1937–1944.
[7] R. Bonneau, C.E.M. Strauss, C.A. Rohl, D. Chivian, P. Bradley, L. Malmstrom, T. Robertson, D. Baker, J. Mol. Biol. 322 (2002) 65–78.
[8] R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C.E.M. Strauss, D. Baker, Proteins (2001) 119–126.
[9] N. Alexander, M. Bortolus, A. Al-Mestarihi, H. McHaourab, J. Meiler, Structure (London, England: 1993), 16 (2008) 181–195.
[10] P.M. Bowers, C.E.M. Strauss, D. Baker, J. Biomol. NMR 18 (2000) 311–318.
[11] S.J. Hirst, N. Alexander, H.S. McHaourab, J. Meiler, J. Struct. Biol. 173 (2011) 506–514.
[12] B.E. Weiner, N. Alexander, L.R. Akin, N. Woetzel, M. Karakas, J. Meiler, Proteins Struct. Funct. Bioinform. 82 (2014) 587–595.
[13] S. Lindert, N. Alexander, N. Wötzel, M. Karakaş, P.L. Stewart, J. Meiler, Structure(London, England:1993), 20 (2012) 464–478.
[14] A.W. Fischer, N.S. Alexander, N. Woetzel, M. Karakas, B.E. Weiner, J. Meiler, Proteins Struct. Funct. and Bioinform., (2015) n/a-n/a.
[15] S. Lindert, R. Staritzbichler, N. Wötzel, M. Karakaş, P.L. Stewart, J. Meiler, Structure (London, England: 1993), 17 (2009) 990–1003.
[16] E.V. Petrotchenko, C.H. Borchers, Mass Spectrom. Rev. 29 (2010) 862–876.
[17] A. Sinz, Mass Spectrom. Rev. 25 (2006) 663–682.
[18] J. Rappsilber, J. Struct. Biol. 173 (2011) 530–540.
[19] M.M. Young, N. Tang, J.C. Hempel, C.M. Oshiro, E.W. Taylor, I.D. Kuntz, B.W. Gibson, G. Dollinger, Proc. Natl. Acad. Sci. 97 (2000) 5802–5806.
[20] K. Lasker, F. Förster, S. Bohn, T. Walzthoeni, E. Villa, P. Unverdorben, F. Beck, R. Aebersold, A. Sali, W. Baumeister, Proc. Natl. Acad. Sci. U.S.A. 109 (2012) 1380–1387.
[21] R.B. Jacobsen, K.L. Sale, M.J. Ayson, P. Novak, J. Hong, P. Lane, N.L. Wood, G.H. Kruppa, M.M. Young, J.S. Schoeniger, Protein Sci. 15 (2006) 1303–1317.
[22] S. Kalkhof, C. Ihling, K. Mechtler, A. Sinz, Anal. Chem. 77 (2004) 495–503.
[23] A. Sinz, Anal. Bioanal. Chem. 397 (2010) 3433–3440.
[24] V. Tinnefeld, A. Sickmann, R. Ahrends, Eur. J. Mass. Spectrom. (Chichester, Eng). 20 (2014) 99–116.
[25] A. Kahraman, F. Herzog, A. Leitner, G. Rosenberger, R. Aebersold, L. Malmström, PLoS ONE 8 (2013) e73411.
[26] S. Kalkhof, S. Haehn, M. Paulsson, N. Smyth, J. Meiler, A. Sinz, Funct. Bioinform. 78 (2010) 3409–3427.
[27] A. Leitner, T. Walzthoeni, A. Kahraman, F. Herzog, O. Rinner, M. Beck, R. Aebersold, Mol. Cell. Proteomics 9 (2010) 1634–1649.
[28] B.L. Zybailov, G.V. Glazko, M. Jaiswal, K.D. Raney, J. Proteom. Bioinform. 6 (2013) 001.
[29] G. Wang, R.L. Dunbrack, Nucleic Acids Res. 33 (2005) W94–W98.
[30] A. Kahraman, L. Malmström, R. Aebersold, Bioinformatics 27 (2011) 2163–2164.
[31] D.M. Schulz, S. Kalkhof, A. Schmidt, C. Ihling, C. Stingl, K. Mechtler, O. Zschörnig, A. Sinz, Funct. Bioinform. 69 (2007) 254–269.
[32] A.A. Canutescu, R.L. Dunbrack, Protein Sci. 12 (2003) 963–972.
[33] M. Karakas, N. Woetzel, R. Staritzbichler, N. Alexander, B.E. Weiner, J. Meiler, PLoS ONE 7 (2012).
[34] D.T. Jones, J. Mol. Biol. 292 (1999) 195–202.
[35] J.K. Leman, R. Mueller, M. Karakas, N. Woetzel, J. Meiler, Prot. Struct. Funct. Bioinform. 81 (2013) 1127–1140.
[36] N. Woetzel, M. Karakas, R. Staritzbichler, R. Muller, B.E. Weiner, J. Meiler, PLoS ONE 7 (2012).
[37] O. Carugo, S. Pongor, Protein Sci. 10 (2001) 1470–1473.
[38] T.F. Havel, G.M. Crippen, I.D. Kuntz, Biopolymers 18 (1979) 73–81.
[39] E.D. Merkley, S. Rysavy, A. Kahraman, R.P. Hafen, V. Daggett, J.N. Adkins, Protein Sci. 23 (2014) 747–759.
[40] N.S. Green, E. Reisler, K.N. Houk, Protein Sci. 10 (2001) 1293–1304.