Journal of proteome research

Article

Cite This: J. Proteome Res. 2019, 18, 3077−3085

pubs.acs.org/jpr

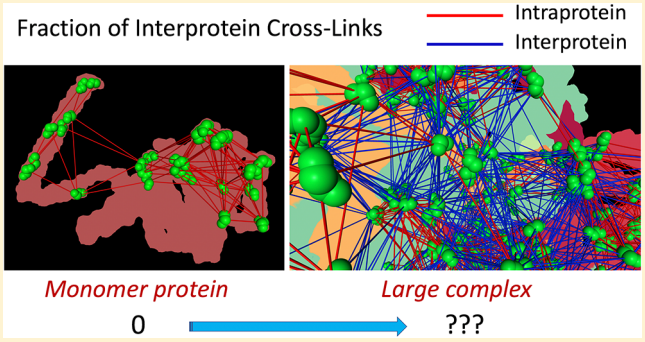
Prediction of an Upper Limit for the Fraction of Interprotein Cross-Links in Large-Scale In Vivo Cross-Linking Studies

Andrew Keller,[†] Juan D. Chavez,[†] Kevin C. Felt,[†] and James E. Bruce*[,†]

[†]Department of Genome Sciences, University of Washington, Seattle, Washington 98195 United States

Supporting Information

ABSTRACT: Chemical cross-linking and mass spectrometry is of growing use for establishment of distance constraints on protein conformations and interactions. Whereas intraprotein cross-links can arise from proteins in isolation, interprotein cross-links reflect proximity of two interacting proteins in the sample. Prediction of expected ratios of the number of interprotein to intraprotein cross-links is hindered by lacking comprehensive knowledge on the interactome network and global occupancy levels for all interacting complex subunits. Here we determine the theoretical number of possible inter- and intraprotein cross-links in available PDB structures of proteins bound in complexes to predict a maximum expected fraction of interprotein cross-links in large scale in vivo cross-linking studies. We show how the maximum fraction can guide interpretation of reported interprotein fractions with respect to the extent of sample protein binding, comparing whole cell and lysate cross-linked samples as an example. We also demonstrate how an observation of interprotein cross-link fractions greater than the maximum value can result from the presence of false positive cross-links which are predominantly interprotein, their number estimable from the observed surplus fraction of interprotein cross-links.



KEYWORDS: *interactomics, cross-linking, maximum interprotein, extent of binding, quality check, FDR lower bound, PIR, DSSO, cross-linker span, mass spectrometry*

INTRODUCTION

Protein interactions and conformations in complex samples, in cells, isolated organelles, tissues, and organisms are primary mediators of normal biological function in healthy states, dysfunction in pathological conditions, and general response to most perturbations.[1−3] The ability to identify and measure protein interactomes in cells has evolved in recent years to include methods based on chemical cross-linking and mass spectrometry technologies.[4,5] Cleavable cross-linkers such as BDP,[6] PIR,[7] DSSO,[8] and others[9] are particularly useful for detecting protein interactions in situ in biological contexts such as tissue culture cells and tissues.

Interprotein and intraprotein cross-linked peptides provide information on protein interactions and protein conformations present during cross-linker application to samples, respectively. Identified interprotein cross-links can originate from two different interacting proteins (e.g., heterodimer), or from interactions between two identical proteins, cross-linked at the same residue position (e.g., homodimer). In contrast, cross-links attached at two different residues of the same protein are assumed to be intraprotein, originating from spatially proximal residues in the same protein molecule, though in rare cases could also be interprotein. Structure models can sometimes help assess the relative likelihoods of those possibilities. While both inter- and intraprotein cross-linked peptides can provide unique biological insight, the fraction of identified cross-links that are interprotein is dependent upon the extent to which the sample proteins are bound together in complexes when cross-linked. However, the relationship between the observed fraction of cross-links that are interprotein and what fractions might be expected with in vivo cross-linking experiments employing any of a variety of chemical cross-linkers, is currently unknown. As such, it is difficult to interpret how the fraction of interprotein cross-links routinely reported by researchers reflects the degree of protein association in their samples.

Prediction of precise interprotein cross-link fractions in cross-linking studies is hindered by lacking comprehensive knowledge on the interactome network and global occupancy levels for all interacting complex subunits. In the absence of this predictive ability, however, expectations on what fraction of interprotein cross-links could be observed are possible by assuming extreme end point conditions. For instance, at one extreme, the fraction of interprotein cross-links will be 0 when all proteins exist in the sample in unbound, noncomplexed forms, as in the case of a cross-linked purified monomer sample protein cross-linked under conditions where no protein

Received: March 22, 2019
Published: July 3, 2019

DOI: 10.1021/acs.jproteome.9b00189
J. Proteome Res. 2019, 18, 3077−3085

oligomers are present. At the other extreme, in the hypothetical scenario where all sample proteins exist stably in fully bound complexes, interprotein as well as intraprotein cross-links are possible. Knowing the maximum possible fraction of interprotein cross-links for fully bound complexes would help to define the expected range of interprotein fractions in cross-linked samples and to guide interpretation of those fractions with respect to the extent of sample protein binding. For example, an observed sample interprotein fraction of 0.2 would indicate a much higher average degree of protein binding if the maximum fraction was found to be 0.25 rather than 1.0. The maximum interprotein fraction can also serve as a quality check, calling attention to unusual cases in which a reported sample fraction is outside of the expected range.

In this study, we investigate this expected maximal possible fraction of interprotein cross-links to be observed in samples by examining 18 676 available PDB structure files of protein complexes ranging in sizes from complexes of 2 proteins to very large complexes such as 5Y6P,[10] a structure of the phycobilisome from the red alga *Griffithsia pacifica* with 862 bound components. For each structure, we calculate its fraction of predicted cross-links that are interprotein. Using this approach, the measurements enable proposal of maximal expected interprotein fractions for large-scale in vivo cross-linking studies employing cross-linkers of specified length. We also demonstrate how a False Discovery Rate (FDR) can be estimated to explain unusual cases of a reported fraction of interprotein cross-links exceeding that maximum, and show how the fraction, normalized to the maximum value, generally reflects the average overall extent to which the sample cross-linked proteins were bound. As cross-linking mass spectrometry continues the current rapid growth in capabilities relevant to interactome and large-scale studies, we feel the concepts and tools developed to investigate interprotein cross-link fraction upper bound estimates can provide general utility for this community.

MATERIALS AND METHODS

PDB Structure Files Used in Analysis

All 148,586 PDB structure files in mmCIF format were downloaded from the RCSB Protein Data Bank Web site ftp://ftp.wwpdb.org/pub/pdb/data/structures/divided/pdb/. When multiple PDB files with the same title were encountered, the single one with the highest resolution (lowest Angstrom value) was kept. In addition, if multiple structures contained the same protein constituents and numbers of chains, only a single representative was kept. In each PDB file, all lysine Cα positions and chains were recorded. Each chain in the structure corresponds to a unique bound constituent protein or ligand. All nonidentical lysine pairs within specified maximum Euclidean distance were assessed as either interprotein (originating from different chains in the structure), or intraprotein (originating from the same chain). Protein N-termini were not included as cross-linkable sites since they occur infrequently in cross-linked peptides (comprising less than 0.3% of cross-linked peptides in the XLinkDB database[11]), and are not always included in structure files. Only PDB files having two or more bound constituents, 50 or more lysine pairs, and resolution 5 Å or less, were included in the analysis. The number of bound constituents in each PDB structure was determined as its number of chains with potentially cross-linkable lysine residues.

Calculation of SASD using Jwalk

Jwalk[12] version 1.1 was run locally in a directory with PDB files using the −lys option to specify lysine starting and ending amino acids.

Determination of Empirical Cross-Linker Maximum Spans in the Context of Structure Files

Empirical cross-link maximum span distances were determined using public data sets available on XLinkDB produced using one of the two different cross-linkers: PIR or DSSO. Intraprotein cross-links originating from a protein for which a PDB structure was available were included. The PIR data comprise 5028 nonredundant cross-links from all public Bruce Lab data sets, and the DSSO, 1008 nonredundant cross-links from two publicly available data sets[13,14] and reanalysis by Mango[15] and XLinkProphet[16] of a third.[17] Distances of each cross-link in the structure were computed as the Euclidean distance between the lysine residue Cα positions attached at each end.

Effect of FDR on the Observed Fraction of Interprotein Cross-Links

The following five data sets were used in the analysis displayed in Figure 5:

1. Nuclear DSSO Mango: Data acquired from DSSO treated intact cell nuclei[17] were reanalyzed with Mango, Comet, and XLinkProphet.
2. Histone Protein Mixture: A single run acquired from a purified bovine histone protein mixture cross-linked with PIR and analyzed with Mango, as previously described,[18] was analyzed alone with XLinkProphet.
3. HeLa PTX: Data from HeLa cells treated with Paclitaxel and cross-linked with PIR (submitted) were acquired with ReACT,[19] searched with a full human proteome database with Comet, and validated with XLinkProphet.
4. *Escherichia coli* Mango: Cells cross-linked in vivo with PIR were analyzed with Mango as described.[15]
5. Mouse Heart ReACT: Mouse heart tissue cross-linked with PIR and analyzed with ReACT, as described.[20] MS^3 spectra were searched using Comet with a mouse Uniprot database downloaded on April 2018 with 33 936 protein sequences, including decoys.

In each case, decoys were included in the XLinkProphet output. When multiple identifications of the same cross-linked peptide pair were present, only that with the lowest (best) maximum Comet expect score was kept. Nonredundant cross-links were sorted by maximum Comet expect score as described in the Results.

Intact HeLa Cell and Lysate Cross-Linking

HeLa cells were cultured in RPMI at 37°C and harvested at confluence with 20 mM EDTA solution (two 15 cm culture dishes per condition, ~4×10^7 cells). Cell pellets were washed with PBS containing 1 mM $CaCl_2$, 1 mM $MgCl_2$, and then resuspended in 500 μL 170 mM disodium phosphate buffer to a 1:1 pellet-buffer ratio by volume. Cells used for intact cross-linking were immediately mixed with BDP-NHP to a final concentration of 10 mM, incubated while shaking at room temperature (hereafter RT) for 30 min, and then washed with 0.1 M ammonium bicarbonate until the pellet was no longer yellow. Cell samples used for lysate cross-linking were first snap frozen in liquid nitrogen and then stored at −80°C until they were cryoground at −80°C with a Retsch MM 400 cryomill for 1 min at 30 Hz. The milled lysate was allowed to

DOI: 10.1021/acs.jproteome.9b00189
J. Proteome Res. 2019, 18, 3077−3085

warm to RT and was then cross-linked in the same manner as the intact cells. Samples were mixed with 8 M urea to approximately a 1:1 ratio by volume and sonicated using a GE-130 ultrasonic processor (five pulses at amplitude 40 for five seconds each). Samples were reduced by the addition of TCEP to a final concentration of 5 mM and incubated for 30 min at RT. Reduced thiols were then alkylated by the addition of iodoacetamide to a final concentration of 10 mM and incubated at RT for 30 min. The intact cross-linking condition contained 8.09 mg of protein, and the lysate condition, 8.23 mg, as determined by Bradford assay. Trypsin was added to samples in a 1:200 ratio by protein mass and allowed to digest overnight. Sample digests were centrifuged at 16 000*g* for 15 min. Supernatants were desalted using C18 Sep-Pak cartridges and bound peptides eluted into 50% and 80% acetonitrile (ACN), 0.1% trifluoracetic acid washes (washes were later mixed). Samples were dried using vacuum centrifugation and resuspended in 0.5 mL of strong cation exchange solvent A (5 mM KH_2PO_4, pH 2.6, 30% ACN).

Cross-linked peptides were fractionated from samples using strong cation exchange chromatography (Agilent 1200 series HPLC attached to a Phenomenex Luna SCX column). Fractionation was achieved using a 1.5 mL/min flow rate over a 97.5 min gradient of increasing percentage of strong cation exchange solvent B (5 mM KH_2PO_4, pH 2.6, 30% ACN, 350 mM KCl), resulting in 14 fractions total. Fractions were pooled into six larger fractions: 1−5, 6−7, 8, 9, 10, 11−14. Cross-linked peptides were enriched separately from the 5 pooled fractions by the addition of monomeric avidin resin (Ultralink, Pierce) and incubation for 30 min at RT. Cross-linked peptides were eluted off the avidin resin with 70% ACN, 1% formic acid, and samples were dried using vacuum centrifugation. Samples were resuspended in 0.1% formic acid and injected into an EASY-nLC 1000 coupled to a Q Exactive Plus mass spectrometer. Samples were separated with a 60 cm × 75 μm inner diameter fused silica analytical column packed with ReproSil-Pur C8 (5 μm diameter, 120 Å pore size particles) by applying a linear gradient from 90% solvent A (0.1% formic acid in water), 10% solvent B (0.1% formic acid in acetonitrile) to 60% solvent A, 40% solvent B over 240 min at a flow rate of 300 nL/min. The mass spectrometer was operated using a data dependent analysis (DDA) method performing one high-resolution (70 000 resolving power (RP) at m/z 200) MS^1 scan from 400 to 2000m/z followed by MS^2 (17 500 RP) on the 20 most abundant ions with a charge between 4+ and 8+ inclusive detected in the MS^1. Parameters for MS^2 scans included an automatic gain control target of 50 000 ions, a maximum ion accumulation time of 100 ms, an isolation window of 3.0m/z, and a normalized collision energy of 30. A dynamic exclusion window of 30s was used to reduce redundant selection of the same parent ion. MS^2 spectra were processed using Mango 2017.01 rev. 0 beta 3 with mass tolerance relationship set to 40.00 ppm, and Mango output was searched using Comet with a human Uniprot database downloaded on April 2018 with 40 632 protein sequences, including decoys. Search results of the five fractions were combined separately for the intact (in vivo) and lysate cross-linked samples, and their identified cross-linked peptides validated with XLinkProphet. Nonredundant cross-links were filtered at 1% FDR and uploaded to XLinkDB to assess their interprotein fractions. The mass spectrometry proteomics data (five raw files and Comet search result pepXML files, each, for HeLa samples cross-linked in vivo and as a lysate) have been deposited to the ProteomeXchange Consortium via the PRIDE[21] partner repository with the data set identifier PXD013063.

RESULTS AND DISCUSSION

We seek to predict the maximal fractions of interprotein cross-links possible for large-scale samples treated with chemical cross-linkers. Taking advantage of an abundance of high resolution structures of proteins bound in complexes, we compute for each its fraction of cross-linkable lysine pairs that are interprotein. Using these calculated fractions for structures of highly bound constituents, we predict expected interprotein fractions not for average large-scale cross-linking studies, but for those involving maximally bound proteins cross-linked in crowded cellular environments.

Structures of bound protein complexes were downloaded as PDB files in mmCIF format from the RCSB Protein Data Bank[22,23] (see Materials and Methods). For each PDB structure, a value of its fraction of cross-links expected to be interprotein, ξ, was computed as its fraction of lysine pairs within a 35 Å Euclidean distance, a maximum cutoff used in many large-scale cross-linking studies,[24−26] that were found to be interchain rather than intrachain. The fraction of intra-protein cross-links, by definition, is always equal to $1 - \xi$. All lysine pairs with Cα carbons within 35 Å Euclidean distance were assumed to be accessible to a cross-linker, though this is strictly untrue due to possible occlusion by protein structural volumes. However, as illustrated further below, estimates made with surface accessible distance calculations[12] illustrated little effect on calculated interprotein fractions. Of course, in actual cross-linking applications to complex samples, lysine residues in certain proteins or certain protein regions may exhibit greater apparent reactivity with cross-linkers which is unaccounted for in this estimation. In addition, some lysine pairs may be more or less detectable depending on the digestive enzyme, mass spectrometer, and database search parameters used to identify cross-linked peptides, and whether cross-link data is reported at the nonredundant peptide pair level or at the residue pair level. Since there is no reason a priori to expect cross-linked lysine pairs detected in stably bound protein complexes to bias toward interprotein or intraprotein, calculated interprotein fractions using all lysine pairs in the PDB structures, within specified maximum distance, should apply robustly to a wide variety of cross-link studies.

Computed interprotein fractions of the protein complex structures were pooled together into a frequency distribution. Each PDB file with a collection of predicted cross-links contributes a single interprotein fraction to the distribution, weighted by its total number of cross-links. The weighted distribution thereby ensures equal contributions of cross-links from each complex, taking into account the greater numbers of predicted cross-links in large versus small complexes. The resulting frequency distribution of computed fractions of interprotein cross-links, ξ, is shown below in Figure 1, plotted separately for subsets of PDB files with an increasing minimum number of bound constituents. Distributions are normalized to the total number of contributing predicted cross-links from the included PDB structures, ranging from 23 236 286 cross-links in 18 676 structure files with two or more bound constituents, to the subset of 5 406 977 cross-links in 562 structures having 25 or more. The distribution means and standard deviations are shown in Table 1. The distribution means indicate the

DOI: 10.1021/acs.jproteome.9b00189
J. Proteome Res. 2019, 18, 3077−3085

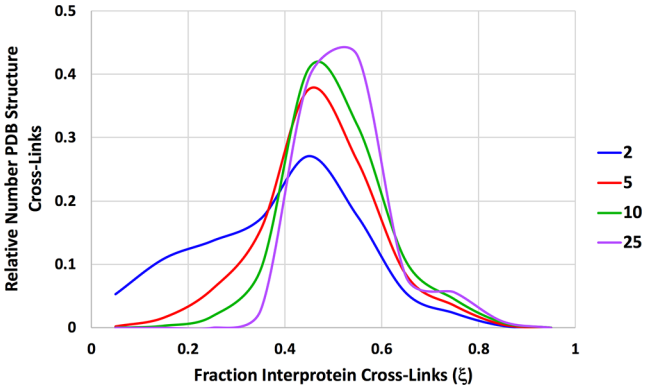


Figure 1. Frequency distribution of computed fractions of predicted cross-links that are interprotein, plotted separately for PDB structures with indicated minimum number of bound constituents ranging from 2 to 25, reflecting increasingly large protein complexes. Interprotein fractions of contributing PDB structures are weighted by their total numbers of cross-links.

Table 1. Fractions of Predicted Cross-Links That Are Inter-Protein, Calculated Among PDB Structures with Indicated Minimum Number of Bound Constituents. Also Shown Are the Numbers of Contributing PDB Structures and Cross-Links

	calculated fraction interprotein cross-links (ξ)			
minimun number PDB bound constituents	mean	std. dev.	total number PDB structures	total number cross-links
2	0.431	0.1653	18 676	23 235 286
5	0.513	0.1211	5 973	14 850 864
10	0.545	0.1048	2 176	10 025 147
15	0.555	0.0984	1 133	7 850 261
20	0.557	0.0951	770	6 885 062
25	0.568	0.0901	562	5 406 977
30	0.569	0.0895	463	4 984 794
35	0.571	0.0888	388	4 584 411
40	0.572	0.0889	350	4 327 806
45	0.572	0.0884	300	3 907 944
50	0.571	0.0868	266	3 647 974

average overall fraction of interprotein cross-links among all included cross-linked complexes. As the number of bound constituents in the PDB files increases, reflecting larger protein complexes, the mean fraction of interprotein cross-links increases, eventually leveling off to a mean ξ value for 25 or more bound constituents of 0.568 (Supporting Information (SI) Figure S1). This suggests that PDB structures with 25 or more constituents are maximally bound, the addition of more components not increasing their interprotein fraction, and thus serve as good models for protein interactions in very crowded in vivo environments.[27,28] The average resolution among the 562 PDB structures with 25 or more bound constituents was 3.4 Å.

As a first approximation, the frequency distribution with PDB files having 25 or more bound constituents is a good indicator of the relative likelihood of observing the highest possible fractions of interprotein cross-links from samples complexed to a maximal extent in a crowded cellular environment. It suggests that if only a single cross-link originating from a random complex with 25 or more bound constituents is observed, it has a 57% chance of being interprotein, and similarly if 100 cross-links are independently obtained from random complexes in this distribution, on average 57 are expected to be interprotein and 43, intraprotein. In a typical in vivo cross-linking study, ordinarily between 1000 and 5000 nonredundant cross-links are identified involving a wide variety of proteins.[17,29−31] If we assume they originate from a multitude of random large complexes represented in the above distribution of interprotein fractions based on PDB structures with 25 or more bound constituents, then the likelihood that each cross-link is interprotein can be independently derived from that distribution. A predicted maximum sample interprotein fraction was thus estimated using Monte Carlo simulation of sample fractions by selection with replacement of 1000 random PDB files with 25 or more bound constituents from the above distribution, each contributing a single cross-link. Whether a cross-link was interprotein was determined according to the calculated interprotein fraction of the PDB complex from which it originated, specifically if a random number between 0 and 1 was less than that fraction. The total number of resulting interprotein cross-links divided by 1000 was recorded as that sample's overall interprotein fraction. This was repeated 10 000 times to generate a distribution of sample interprotein fractions, found to have a mean value of 0.563, close to the PDB-based distribution mean of 0.568. A maximum inter-protein fraction, ξ_{max}, was then estimated with p-value 0.01 as the value greater than 99% of sample fractions, equal to 0.6. This closely matches the maximum value predicted according to the Central Limit Theorem (see SI). Thus, 99% of data sets with 1000 or more cross-links are expected to have a fraction of interprotein cross-links in the range between 0 (cross-linked sample proteins completely unbound) and 0.6 (cross-linked sample proteins completely in large complexes). The maximum is conservative since it is unlikely that an in vivo cross-linked sample would be sufficiently bound in large complexes to have an interprotein fraction equal to, let alone greater than, its value. It is important to note, however, that this maximum is predicted for large-scale studies in which identified cross-links arise from a multitude of different complexes. A sample with all cross-links obtained from a single protein complex could have a higher interprotein fraction, as indicated by the PDB-based distribution.

Fractions of interprotein cross-links in the PDB files can also be estimated based on the numbers of interchain and intrachain lysine pairs within 35 Å solvent accessible surface distance (SASD), as calculated by Jwalk.[12] This measure predicts which lysine pairs in a structure are accessible to a cross-linker better than Euclidean distance, taking into account whether other parts of proteins will obstruct the path. Calculating SASDs for lysine pairs of all PDB files is very time-consuming, so they were calculated for a random subset of 1000 PDB files analyzed above in mmCIF format, and used to compute fractions of interprotein cross-links based on the numbers of interchain and intrachain lysine pairs with SASD distances within the maximum specified bound. Figure 2 shows a comparison of fractions of interprotein cross-links for the PDB files computed based on a maximum 35 Å SASD versus those based on a maximum 35 Å Euclidean distance. One can see very good agreement over a wide range, suggesting that values based on Euclidean distance are a valid measure of the fraction of interprotein cross-links predicted in PDB structure files. The major difference between using the two distance

DOI: 10.1021/acs.jproteome.9b00189
J. Proteome Res. 2019, 18, 3077−3085

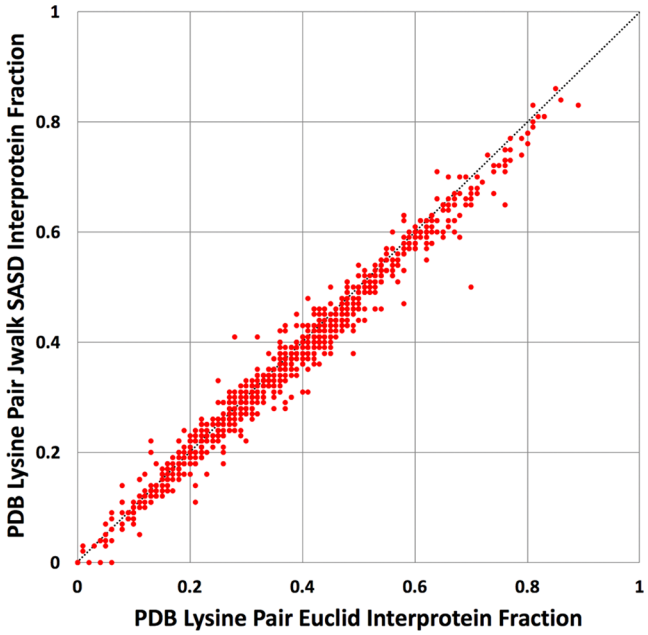


Figure 2. Fractions of interprotein cross-links computed among PDB structures based on lysine pairs within 35 Å SASD or Euclidean distance. Dashed line indicates perfect correlation.

criteria is the overall lower number of SASD accessible lysine pairs, not the proportion of interprotein cross-links. SI Table S1 shows the correlation coefficients comparing the computed fractions of interprotein cross-links among the 1000 PDB structures using SASD versus Euclidean distance, for a range of maximum interlysine distances including 35 Å.

Different chemical cross-linkers can vary in the maximum distance spanned,[32] not always equal to 35 Å. For this reason, the predicted maximum fraction of interprotein cross-links based on PDB structure files with a minimum number of 25 bound constituents was recalculated over a range of maximum cross-link lysine pair Cα Euclidean distances. For each maximum distance, the frequency distribution of lysine pair fractions within range that were interchain was compiled and used to estimate a maximum interprotein sample fraction by Monte Carlo simulation, as described above. As the allowed distance reached by a cross-linker increases, the predicted maximum fraction of interprotein cross-links increases, as shown in Figure 3 below. This observation occurs because more neighboring proteins and neighboring protein residues become accessible to a cross-linker of greater length, enabling more interprotein cross-links. Yet beyond some distance approaching a protein's diameter, no additional intraprotein cross-links are possible, all being already within reach. As a consequence of this trend, the predicted maximum fraction of interprotein cross-links should take into consideration the maximum distance spanned by the applied cross-linker in the context of structure files. One generally expects lower fractions of interprotein cross-links using shorter-spanning chemical cross-linkers.

Empirical distances spanned by PIR and DSSO cleavable cross-linkers were investigated by observing distances of intraprotein cross-links on our online cross-linked peptide database and analysis tool suite, XLinkDB,[11] in the context of PDB structure files. Figure 4 illustrates the cumulative distance distributions for PIR and DSSO cross-links. The PIR cross-links include 5028 nonredundant cross-links from public Bruce

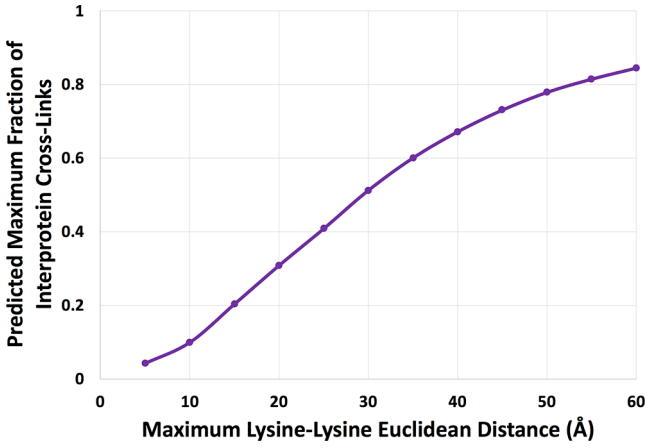


Figure 3. Predicted maximum fraction of interprotein cross-links based on lysine pairs within indicated maximum Cα−Cα distances in PDB structure files with 25 or more bound constituents.

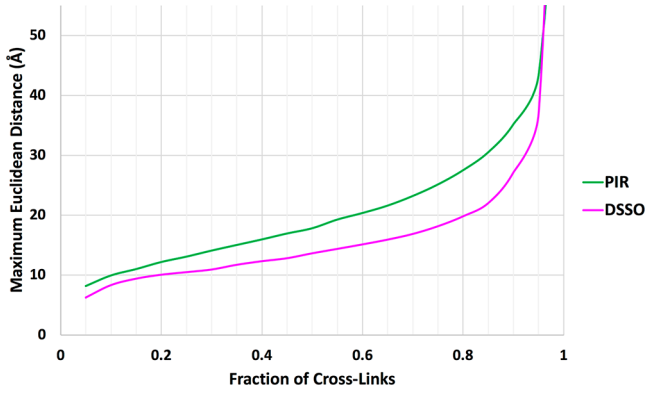


Figure 4. Fraction of cross-links within indicated maximum Euclidean distance in context of PDB structure files, plotted separately for samples treated with PIR and DSSO cross-linkers. Included are intraprotein cross-links originating from proteins with available PDB structures.

Lab data sets on XLinkDB, and the DSSO, 1008 nonredundant cross-links from three public data sets (see Materials and Methods). It is evident that greater than 90% of PIR cross-links are within 35 Å Euclidean distance, and 90% of DSSO cross-links, within 27 Å. These distances are the empirically derived effective maximum spans of the cross-linkers in the context of PDB structure files. The former value matches our initial estimated cross-link span with a predicted maximum fraction of interprotein cross-links for PIR equal to 0.6, whereas the latter value, a maximum span of 27 Å, corresponds, according to Figure 3, to a predicted maximum fraction of interprotein cross-links for DSSO of 0.45. Since some proteins have multiple alternative conformations that differ from the structure file to varying degrees, a fraction of intraprotein cross-links is routinely found to have distances in structure files exceeding the expected range. This is evidenced in Figure 4 by the large increase in maximum distance observed as the fraction of cross-links approaches 100%. For this reason, a lower fraction of 90% was used to infer the maximum distance reachable by the cross-linkers. Because of the imperfect resolution of structure files and possible existence in samples of flexible and alternative protein conformations without available structures, one can only roughly estimate the effective maximum spans of chemical cross-linkers. As has been

DOI: 10.1021/acs.jproteome.9b00189
J. Proteome Res. 2019, 18, 3077−3085

previously noted,[25,33−35] these empirical measurements for a variety of reasons do not always coincide precisely with predictions based on the cross-linker molecular structures.

The most likely reason for observing a fraction of interprotein cross-links greater than its predicted maximum value in any large-scale cross-linked sample is the presence of false positive results, the great majority of which are interprotein. That is because two random peptides assigned to a cross-link most often correspond to different proteins. Only when small databases are used to assign PSMs to the cross-linked peptides, particularly those containing some very large proteins, are a significant fraction of false positives intraprotein cross-links due to the greater chance that two random peptides correspond to the same protein. In most data sets analyzed with whole proteome databases, we observe the great majority of extremely low scoring cross-links to be interprotein.[20] Theoretically, assuming all false positives are interprotein cross-links, then the FDR is proportional to the observed surplus of the fraction of interprotein cross-links, ξ, above the actual value among only true positives in the data set, ξ_0, according to the following equation (see SI):

$$\mathrm{FDR} = \frac{\xi - \xi_0}{1 - \xi_0} \text{ where } \xi > \xi_0 \tag{1}$$

The FDR will be even greater, of course, if some of the false positives are intraprotein, as may occur when small search databases with few protein sequences are used.

To investigate the effect of false positives on the fraction of interprotein cross-links observed, several data sets of non-redundant cross-links from samples treated with PIR or DSSO were sorted and filtered by increasing maximum Comet[36] expect score, the worse of the scores assigned to each cross-link's two released PSMs.[37] The resulting FDRs were estimated from the numbers of decoy and decoy−decoy cross-links in the filtered data sets,[38,39] and the fraction of interprotein target cross-links recorded. Figure 5 below plots the surplus fraction of interprotein cross-links, $\xi - \xi_0$, normalized to the theoretical proportionality constant, $1 - \xi_0$, versus FDR. For each data set, ξ_0 was estimated as its fraction of interprotein cross-links at decoy-estimated 0% FDR. Also shown is the predicted line with slope 1, according to eq 1. It is evident that for all data sets the fraction of interprotein cross-links, ξ, increases in proportion to FDR, with overall good agreement between the observed and predicted relationships. These data sets, analyzed by ReACT[19] or Mango[15] and XLinkProphet, span a range of protein complexity and search database size, including the Histone sample, a purified protein mixture that was cross-linked and acquired in a single run for Mango analysis, the identified cross-links of which were over 40% interprotein (ξ_0 = 0.4). SI Table S2 shows several features of the data sets, such as their numbers of correct cross-link identifications at 1% FDR, and the fraction of interprotein cross-links among their decoy results, results of low search score, and results at decoy-estimated 0% FDR (i.e., ξ_0). As expected, 96% or more of target cross-links with very poor search scores (35 or greater maximum Comet expect score), and all of the decoy−decoy cross-links, were interprotein. These results thus illustrate how the FDR of cross-linking studies can be inferred from the surplus of their observed fraction of interprotein cross-links, ξ, relative to their actual value, ξ_0, according to eq 1.

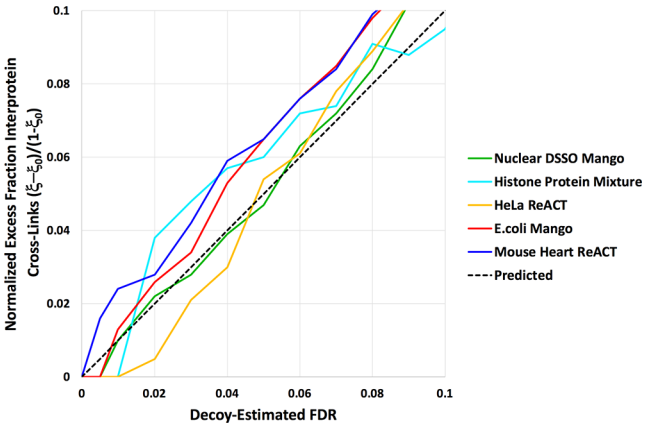


Figure 5. Observed excess fraction of interprotein cross-links in data filtered for increasing decoy-estimated FDR. The true interprotein fraction, ξ_0, was computed among results filtered at decoy-estimated 0% FDR.

The FDR for a sample with a fraction of interprotein cross-links exceeding the predicted maximum can in a similar manner be estimated using eq 1. Though the actual fraction among true positives, ξ_0, is not generally known, it must be less than or equal to the predicted maximum, ξ_{max}. Inserting the inequality $\xi_{max} \geq \xi_0$ into eq 1 yields an expression for a lower bound on the FDR responsible for a reported interprotein fraction exceeding the maximum value (see SI):

$$\mathrm{FDR} \geq \frac{\xi - \xi_{max}}{1 - \xi_{max}} \text{ where } \xi > \xi_{max} \geq \xi_0 \tag{2}$$

Plugging in the predicted maximum fraction of interprotein cross-links for samples treated with PIR and DSSO, one obtains the following equations for the FDR of a cross-linked data set due to an observed value of ξ above the maximum:

$$\mathrm{FDR}_{\mathrm{PIR}} \geq \frac{\xi - 0.6}{0.4} \text{ where } \xi > 0.6 \geq \xi_0 \tag{2a}$$

$$\mathrm{FDR}_{\mathrm{DSSO}} \geq \frac{\xi - 0.45}{0.55} \text{ where } \xi > 0.45 \geq \xi_0 \tag{2b}$$

These equations are not generally applicable to large-scale cross-linking studies, but only to those unusual cases with a reported fraction of interprotein cross-links above the predicted maximum value, 0.6 for PIR and 0.45 for DSSO, and are minimum estimates. The FDR will be greater as the actual fraction among true positives, ξ_0, decreases relative to the predicted maximum value when not all cross-linked proteins are bound in large complexes, which will be the case for any study that is not of purified interacting proteins. It will also be greater if some of the false positives are intraprotein.

Observed fractions of interprotein cross-links reflect the degree to which cross-linked proteins are associated in complexes. For large-scale studies, assuming no false positives among the identified cross-links, the fraction of interprotein cross-links, ξ, can range from 0, when cross-linked proteins are all unbound, to its predicted maximum value ξ_{max}, when cross-linked proteins are completely associated in large complexes. One can therefore estimate the average extent to which the cross-linked sample proteins were bound in complexes, D_{bind}, varying from 0 to 1, as

$$D_{\mathrm{bind}} = \frac{\xi}{\xi_{max}} \tag{3}$$

DOI: 10.1021/acs.jproteome.9b00189
J. Proteome Res. 2019, 18, 3077−3085

Normalizing observed interprotein fractions to the maximum expected value indicates the average extent of sample protein binding relative to the maximum extent possible in crowded in vivo fully bound complexes. It enables comparison across data sets employing chemical cross-linkers with different maximum spans, and hence different ξ_{max} values. For example, a data set using PIR with an interprotein fraction of 0.3 and a one using DSSO with an interprotein fraction of 0.22 would both reflect the same degree of binding, at half that of maximum. Using eq 3 with the ξ_{max} value of 0.6 for PIR cross-linkers spanning 35 Å, extents of binding of public data sets on XLinkDB cross-linked in vivo with PIR were computed. All were found to have fractions of interprotein cross-links in the range of 0.21−0.44. Because some intraprotein cross-links may also originate from distinct copies of the same protein bound as a homodimer, the actual fraction of interprotein cross-links may be a bit higher (see SI). Nevertheless, they correspond roughly to extents of binding between 0.35 and 0.73, reflecting their average extent of protein association during the cross-linking reaction. The Bioplex interactome[2] now includes 23 744 interactions, both direct and indirect, among 7668 human proteins, suggesting that a significant proportion of cellular proteins may be found in complexes. Interestingly of human proteins in the interactome, 14% have only a single interactor, and 43%, 5 or fewer. Though the interactome is still incomplete, this suggests that many proteins are bound only in small complexes, consistent with a lower than maximum fraction of interprotein cross-links.

We analyzed HeLa cell samples cross-linked in vivo and as a lysate (see Materials and Methods), and compared their fractions of interprotein cross-links. As expected, we found the in vivo cross-linked sample to have a higher fraction of interprotein cross-links than the lysate, 0.26 versus 0.13. These values, according to eq 3, correspond to degrees of binding of 0.43 and 0.22 for the two samples, respectively. This indicates that proteins in the in vivo cross-linked sample are bound in complexes to significantly greater extents than in the lysate sample, on average at 43% maximally complexed levels. The degree of binding in the lysate sample is the lowest observed in our lab, where all other samples were cross-linked in vivo. It is likely that significant dissociation of protein interactions occurs during processing of lysate samples prior to cross-linking. Interestingly, the nuclear DSSO Mango data set in Figure 5 has an interprotein fraction of 0.25, less than that of the in vivo HeLa sample, yet due to the lower DSSO ξ_{max} value, corresponds to a higher average degree of binding, at 56% of maximally complexed levels.

The overlap between cross-links identified in the two HeLa samples is shown in Figure 6. Also indicated are the fraction of cross-links that are interprotein in the three overlap segments. Cross-links identified only in the in vivo cross-linked sample are enriched for interprotein associations. Interestingly, intraprotein cross-links are over-represented among the cross-links in common to the two samples. Intralinks reflect protein conformations, and are thus more likely to be consistent across all samples in which the proteins are observed. In contrast, interprotein cross-links reflect protein interactions that are more likely to vary from sample to sample. The small number of interprotein cross-links identified only in the lysate may arise from interactions that are more accessible to cross-linkers in the lysate than in vivo, from interactions that occur only during sample processing, or more generally, due to incomplete sampling by mass spectrometry.

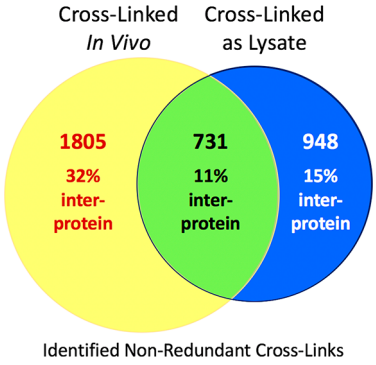


Figure 6. Comparison of cross-links identified in HeLa samples cross-linked in vivo versus as a purified lysate, filtered for 1% FDR. Indicated are the numbers of cross-links and their interprotein fractions.

CONCLUSIONS

We describe a means to predict the maximum value for the fraction of interprotein cross-links identified in large-scale in vivo cross-linking studies employing cross-linkers of specified length. It is a first approximation based on calculated fractions of interprotein cross-links predicted for large protein complexes with known structures. Proposed maximum fractions are 0.6 for data sets employing the PIR cross-linker and 0.45 for those using the shorter-spanning DSSO cross-linker. One generally expects lower fractions of interprotein cross-links using chemical cross-linkers with shorter linker lengths. These maxima are estimates since the maximum spans of cross-linkers in the context of structure files are not precisely defined. Findings of this study help explain why all of our data with PIR cross-linkers have fractions of interprotein cross-links lower than the 0.6 maximum, and lower for a cross-linked lysate sample than for those cross-linked in vivo, reflecting the average degree of binding of their cross-linked protein constituents. Normalizing observed interprotein fractions to the maximum expected value for fully bound complexes indicates the extent of sample protein binding on a scale of 0 to 1, facilitating comparison across data sets employing chemical cross-linkers with different maximum spans. Since the predicted maxima are calculated based on the assumption that all sample proteins are bound in large complexes in crowded cellular environments similar to those for which structure PDB files with 25 or more bound constituents are available, it is unlikely for any large-scale cross-linked sample to have an observed fraction of interprotein cross-links greater than the predicted maximum value. A violation may indicate the presence of false positive cross-links which are predominantly interprotein, their number estimable from the observed surplus fraction of interprotein cross-links.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.9b00189.

Figure S1: Effect of the PDB structure complex size on interprotein fractions. Estimation of maximum sample interprotein fraction by central limit theorem. Derivation of eq 1. Derivation of eq 2. Table S1: Comparison of PDB interprotein cross-link fractions calculated using

DOI: 10.1021/acs.jproteome.9b00189
J. Proteome Res. 2019, 18, 3077−3085

SASD versus Euclidean maximum distance. Table S2: Effect of FDR on the fraction of interprotein cross-links. Estimated undercount of interprotein cross-links (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: jimbruce@uw.edu.

ORCID

James E. Bruce: 0000-0001-6441-6089

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We give special thanks to Bill Noble for valuable discussions and suggestions. We also thank members of the Bruce Lab for many helpful discussions. This work was supported by grants R01HL110349, R01HL142628, U19AI107775, and R01GM086688.

REFERENCES

(1) Havugimana, P. C.; Hart, G. T.; Nepusz, T.; Yang, H.; Turinsky, A. L.; Li, Z.; Wang, P. I.; Boutz, D. R.; Fong, V.; Phanse, S.; Babu, M.; Craig, S. A.; Hu, P.; Wan, C.; Vlasblom, J.; Dar, V. U.; Bezginov, A.; Clark, G. W.; Wu, G. C.; Wodak, S. J.; Tillier, E. R.; Paccanaro, A.; Marcotte, E. M.; Emili, A. A census of human soluble protein complexes. *Cell* **2012**, *150* (5), 1068−81.

(2) Huttlin, E. L.; Bruckner, R. J.; Paulo, J. A.; Cannon, J. R.; Ting, L.; Baltier, K.; Colby, G.; Gebreab, F.; Gygi, M. P.; Parzen, H.; Szpyt, J.; Tam, S.; Zarraga, G.; Pontano-Vaites, L.; Swarup, S.; White, A. E.; Schweppe, D. K.; Rad, R.; Erickson, B. K.; Obar, R. A.; Guruharsha, K. G.; Li, K.; Artavanis-Tsakonas, S.; Gygi, S. P.; Harper, J. W. Architecture of the human interactome defines protein communities and disease networks. *Nature* **2017**, *545* (7655), 505−509.

(3) Nooren, I. M.; Thornton, J. M. Diversity of protein-protein interactions. *Embo j* **2003**, *22* (14), 3486−92.

(4) Holding, A. N. XL-MS: Protein cross-linking coupled with mass spectrometry. *Methods* **2015**, *89*, 54−63.

(5) Leitner, A.; Faini, M.; Stengel, F.; Aebersold, R. Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends Biochem. Sci.* **2016**, *41* (1), 20−32.

(6) Zhang, H.; Tang, X.; Munske, G. R.; Tolic, N.; Anderson, G. A.; Bruce, J. E. Identification of protein-protein interactions and topologies in living cells with chemical cross-linking and mass spectrometry. *Mol. Cell. Proteomics* **2009**, *8* (3), 409−20.

(7) Tang, X.; Bruce, J. E. A new cross-linking strategy: protein interaction reporter (PIR) technology for protein-protein interaction studies. *Mol. BioSyst.* **2010**, *6* (6), 939−47.

(8) Kao, A.; Chiu, C. L.; Vellucci, D.; Yang, Y.; Patel, V. R.; Guan, S.; Randall, A.; Baldi, P.; Rychnovsky, S. D.; Huang, L. Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Mol. Cell. Proteomics* **2011**, *10* (1), M110−002212.

(9) Muller, M. Q.; Dreiocker, F.; Ihling, C. H.; Schafer, M.; Sinz, A. Cleavable cross-linker for protein structure analysis: reliable identification of cross-linking products by tandem MS. *Anal. Chem.* **2010**, *82* (16), 6958−68.

(10) Zhang, J.; Ma, J.; Liu, D.; Qin, S.; Sun, S.; Zhao, J.; Sui, S. F. Structure of phycobilisome from the red alga Griffithsia pacifica. *Nature* **2017**, *551* (7678), 57−63.

(11) Zheng, C.; Weisbrod, C. R.; Chavez, J. D.; Eng, J. K.; Sharma, V.; Wu, X.; Bruce, J. E. XLink-DB: database and software tools for storing and visualizing protein interaction topology data. *J. Proteome Res.* **2013**, *12* (4), 1989−95.

(12) Matthew Allen Bullock, J.; Schwab, J.; Thalassinos, K.; Topf, M. The Importance of Non-accessible Crosslinks and Solvent Accessible Surface Distance in Modeling Proteins with Restraints From Crosslinking Mass Spectrometry. *Mol. Cell. Proteomics* **2016**, *15* (7), 2491−500.

(13) Liu, F.; Rijkers, D. T.; Post, H.; Heck, A. J. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat. Methods* **2015**, *12* (12), 1179−84.

(14) Liu, F.; Lossl, P.; Rabbitts, B. M.; Balaban, R. S.; Heck, A. J. R. The interactome of intact mitochondria by cross-linking mass spectrometry provides evidence for coexisting respiratory supercomplexes. *Mol. Cell. Proteomics* **2018**, *17* (2), 216−232.

(15) Mohr, J. P.; Perumalla, P.; Chavez, J. D.; Eng, J. K.; Bruce, J. E. Mango: A General Tool for Collision Induced Dissociation-Cleavable Cross-Linked Peptide Identification. *Anal. Chem.* **2018**, *90* (10), 6028−6034.

(16) Keller, A.; Chavez, J. D.; Bruce, J. E. Increased sensitivity with automated validation of XL-MS cleavable peptide crosslinks. *Bioinformatics* **2019**, *35* (5), 895−897.

(17) Fasci, D.; van Ingen, H.; Scheltema, R. A.; Heck, A. J. R. Histone Interaction Landscapes Visualized by Crosslinking Mass Spectrometry in Intact Cell Nuclei. *Mol. Cell. Proteomics* **2018**, *17* (10), 2018−2033.

(18) Keller, A.; Chavez, J. D.; Bruce, J. E., Increased Sensitivity with Automated Validation of XL-MS Cleavable Peptide Crosslinks. *Bioinformatics* **2019**, (bty720).35895

(19) Weisbrod, C. R.; Chavez, J. D.; Eng, J. K.; Yang, L.; Zheng, C.; Bruce, J. E. In vivo protein interaction network identified with a novel real-time cross-linked peptide identification strategy. *J. Proteome Res.* **2013**, *12* (4), 1569−79.

(20) Chavez, J. D.; Lee, C. F.; Caudal, A.; Keller, A.; Tian, R.; Bruce, J. E. Chemical Crosslinking Mass Spectrometry Analysis of Protein Conformations and Supercomplexes in Heart Tissue. *Cell Syst* **2018**, *6* (1), 136−141.e5.

(21) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Perez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaino, J. A. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **2019**, *47* (D1), D442−d450.

(22) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−42.

(23) Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Di Costanzo, L.; Christie, C.; Dalenberg, K.; Duarte, J. M.; Dutta, S.; Feng, Z.; Ghosh, S.; Goodsell, D. S.; Green, R. K.; Guranovic, V.; Guzenko, D.; Hudson, B. P.; Kalro, T.; Liang, Y.; Lowe, R.; Namkoong, H.; Peisach, E.; Periskova, I.; Prlic, A.; Randle, C.; Rose, A.; Rose, P.; Sala, R.; Sekharan, M.; Shao, C.; Tan, L.; Tao, Y. P.; Valasatava, Y.; Voigt, M.; Westbrook, J.; Woo, J.; Yang, H.; Young, J.; Zhuravleva, M.; Zardecki, C. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **2019**, *47* (D1), D464−d474.

(24) Navare, A. T.; Chavez, J. D.; Zheng, C.; Weisbrod, C. R.; Eng, J. K.; Siehnel, R.; Singh, P. K.; Manoil, C.; Bruce, J. E. Probing the protein interaction network of Pseudomonas aeruginosa cells by chemical cross-linking mass spectrometry. *Structure* **2015**, *23* (4), 762−73.

(25) Merkley, E. D.; Rysavy, S.; Kahraman, A.; Hafen, R. P.; Daggett, V.; Adkins, J. N. Distance restraints from crosslinking mass spectrometry: mining a molecular dynamics simulation database to evaluate lysine-lysine distances. *Protein Sci.* **2014**, *23* (6), 747−59.

(26) Schneider, M.; Belsom, A.; Rappsilber, J. Protein Tertiary Structure by Crosslinking/Mass Spectrometry. *Trends Biochem. Sci.* **2018**, *43* (3), 157−169.

(27) Rivas, G.; Minton, A. P. Macromolecular Crowding In Vitro, In Vivo, and In Between. *Trends Biochem. Sci.* **2016**, *41* (11), 970−981.

DOI: 10.1021/acs.jproteome.9b00189
J. Proteome Res. 2019, 18, 3077−3085

(28) Ellis, R. J. Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Curr. Opin. Struct. Biol.* **2001**, *11* (1), 114−9.
(29) Schweppe, D. K.; Chavez, J. D.; Lee, C. F.; Caudal, A.; Kruse, S. E.; Stuppard, R.; Marcinek, D. J.; Shadel, G. S.; Tian, R.; Bruce, J. E. Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (7), 1732−1737.
(30) Chavez, J. D.; Lee, C. F.; Caudal, A.; Keller, A.; Tian, R.; Bruce, J. E., Chemical Crosslinking Mass Spectrometry Analysis of Protein Conformations and Supercomplexes in Heart Tissue. *Cell Syst* **2018**.6136
(31) Tan, D.; Li, Q.; Zhang, M. J.; Liu, C.; Ma, C.; Zhang, P.; Ding, Y. H.; Fan, S. B.; Tao, L.; Yang, B.; Li, X.; Ma, S.; Liu, J.; Feng, B.; Liu, X.; Wang, H. W.; He, S. M.; Gao, N.; Ye, K.; Dong, M. Q.; Lei, X., Trifunctional cross-linker for mapping protein-protein interaction networks and comparing protein conformational states. *eLife* **2016**, *5*. DOI: 10.7554/eLife.12509
(32) Paramelle, D.; Miralles, G.; Subra, G.; Martinez, J. Chemical cross-linkers for protein structure studies by mass spectrometry. *Proteomics* **2013**, *13* (3−4), 438−56.
(33) Leitner, A.; Walzthoeni, T.; Kahraman, A.; Herzog, F.; Rinner, O.; Beck, M.; Aebersold, R. Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol. Cell. Proteomics* **2010**, *9* (8), 1634−49.
(34) Ding, Y. H.; Gong, Z.; Dong, X.; Liu, K.; Liu, Z.; Liu, C.; He, S. M.; Dong, M. Q.; Tang, C. Modeling Protein Excited-state Structures from "Over-length" Chemical Cross-links. *J. Biol. Chem.* **2017**, *292* (4), 1187−1196.
(35) Green, N. S.; Reisler, E.; Houk, K. N. Quantitative evaluation of the lengths of homobifunctional protein cross-linking reagents used as molecular rulers. *Protein Sci.* **2001**, *10* (7), 1293−304.
(36) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13* (1), 22−4.
(37) Trnka, M. J.; Baker, P. R.; Robinson, P. J.; Burlingame, A. L.; Chalkley, R. J. Matching cross-linked peptide spectra: only as good as the worse identification. *Mol. Cell. Proteomics* **2014**, *13* (2), 420−34.
(38) Walzthoeni, T.; Claassen, M.; Leitner, A.; Herzog, F.; Bohn, S.; Förster, F.; Beck, M.; Aebersold, R. False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nat. Methods* **2012**, *9*, 901.
(39) Fischer, L.; Rappsilber, J. Quirks of Error Estimation in Cross-Linking/Mass Spectrometry. *Anal. Chem.* **2017**, *89* (7), 3829−3833.

DOI: 10.1021/acs.jproteome.9b00189
J. Proteome Res. 2019, 18, 3077−3085