Published in final edited form as:

Proteomics. 2012 May ; 12(10): 1565–1575. doi:10.1002/pmic.201100516.

***In vivo* protein complex topologies: Sights through a cross-linking lens**

James E. Bruce

Department of Genome Sciences, University of Washington, Seattle, WA 98109

Abstract

Proteins are a remarkable class of molecules that exhibit wide diversity of shapes or topological features that underpin protein interactions and give rise to biological function. In addition to quantitation of abundance levels of proteins in biological systems under a variety of conditions, the field of proteome research has as a primary mission the assignment of function for proteins and if possible, illumination of factors that enable function. For many years, chemical cross-linking methods have been used to provide structural data on single purified proteins and purified protein complexes. However, these methods also offer the alluring possibility to extend capabilities to complex biological samples such as cell lysates or intact living cells where proteins may exhibit native topological features that do not exist in purified form. Recent efforts are beginning to provide glimpses of protein complexes and topologies in cells that suggest continued development will yield novel capabilities to view functional topological features of many protein and complexes as they exist in cells, tissues or other complex samples. This review will describe rationale, challenges and a few success stories along the path of development of cross-linking technologies for measurement of *in vivo* protein interaction topologies.

1.1 Proteins and their interactions are critical determinants of cellular function

Life on earth has evolved to utilize proteins as primary functional molecules with very few exceptions due to the structures, topological features and interactions that can be generated with amino acid sequences [1, 2]. Diversity in function and physical properties of proteins results from amino acids that yield local physical and chemical properties and overall protein topological features. In terms of physical and functional diversity, proteins are unparalleled by any other single polymer type naturally occurring or otherwise. Cells regulate function at the protein level in inter-related ways including: regulation of protein abundance levels, protein localization, and protein conformation and interactions. Great progress has been made in the previous decade with technology development that enables large-scale protein abundance measurements from cells, yet the ability to understand functional regulation in cells is critically dependent on further advances to allow improved visualization at levels involving interactions and conformational changes in cells

1.2 Protein topological features enable interactions

Proteins interact with one another through their surface residues and topological features to perform highly specific functional roles. Kuriyan and Eisenberg present an inspiring account of the origin of regulated protein interaction networks that includes co-localization of binding partners that increase local concentration and adaptive mutations that give rise to

specific interactions that confer advantage [3]. Based on this, two considerations seem likely. First, homomultimers are anticipated to be the most frequent type of specific interaction since self-co-localization is unavoidable and therefore, local binding partner concentration is expected to be high for homomultimers. And second, evolution of protein coding genes is in some ways, directed by advantageous protein topological features that result. That is, genetic mutations that confer advantage can do so by fine tuning protein topological features in beneficial ways. Thus, proteins, and more specifically, protein topological features can serve as primary determinants in genomic evolution. Similarly, posttranslational modifications can result in changes to protein amino acid sequence, local changes in amino acid size, charge, or hydrophobicity and enable topological variation beyond the confines of a single gene or amino acid sequence. In fact, the remarkable impact of posttranslational modification serves as a reminder of the obvious: proteins actually *only* have topology as their singular attribute – all else originates from this. From these considerations, one key conclusion is that the proteome is, at any given instant, an optimized set of topological features derived through protein abundance levels, modifications, localization and interactions to provide needed functionality.

1.3 Topology is regulated by many factors

Accessible surface areas of proteins are most sensitive to environmental change and provide important recognition sites for protein interacting partners that can serve to regulate function. Not surprisingly, accessible surface residues also serve as important sites for posttranslational modifications since change in charge or physical properties at a particular site can impact topological features and interactions. Phosphorylation and acetylation are two such examples that can contribute strongly to local protein topology and serve to alter sites to produce a strong effect on protein and interaction topologies [4–6]. An interesting observation of the correlation between posttranslational modifications and *in vivo* cross-linking sites is described below. Measurement of topological features of proteins in cells or within a biologically relevant context could dramatically advance understanding of cellular function, how protein interactions change with disease, how cells functionally adapt to a changing environment, and how sensitive and drug resistant cancer cells regulate function differently.

1.4 Disordered protein regions are important for protein function and interactions in cells

Contrary to the traditional view that protein function is explicitly derived from stable three-dimensional structure, growing recognition exists that disordered or “unstructured” regions are often functional, fold on binding to their biological targets or exist as critical, flexible linkers that have a role in the assembly of macromolecular complexes [7]. Thus, disorder is an aspect of protein topology that is also critical to *in vivo* function. As reviewed by Dyson and Wright [8, 9], disordered proteins have only recently been discovered because of the strong bias that exists with classical biochemical methods geared towards the production and characterization of fully folded, active proteins. In general, unstructured regions in proteins serve to complicate and hinder crystallographic studies. Biological systems are not similarly biased against intrinsic disorder and employ disorder in many functional strategies [10–12]. Typically, structural studies can only indicate the presence of disorder through the absence of observed electron density in local regions. NMR, circular dichroism (CD), and other spectroscopic methods as well as traditional biochemical studies of proteolytic susceptibility, can give important information on disordered or unstructured protein regions *in vitro*. However, few existing methods can provide information on protein disorder in cells where binding with native partner proteins takes place and such disordered regions hold significance. Additionally, many of these interactions involving classes of molecules that

require non-physiological solution conditions for their extraction, such as membrane proteins [7, 13], may not survive outside the cellular environment [14]. For these, measurements on protein structures and interactions as they exist in cells may be the only hope for improved understanding and visualization of relevant topological features. These aspects stand as major challenges for proteomics research and the overall goal to improve understanding of protein function. New technology that can enable measurement of interacting protein topologies in cells could significantly advance understanding of protein function in cells or complex systems.

1.5. Protein topological measurements are predominantly done with purified proteins

Significant effort geared towards protein crystallography, modeling of protein structures and interacting proteins has already significantly advanced understanding of protein function. In this area, protein crystallography currently stands as an irreplaceable pillar that, given the widespread success with this technology, is unlikely to be displaced within the near future. This is true despite the limitations with intrinsically disordered protein regions described above. However, while crystallography and molecular biology techniques like the yeast two-hybrid [15] have enabled measurements of thousands of structures and testing of thousands of interactions among proteins, for exceedingly few interactions does *any* topological data exist to describe how these interactions occur *in vivo*. Finally, many topological features, as described above, are beyond current capabilities for *in vitro* measurements and will require alternative technologies.

1.6 Chemical cross-linking technology has potential to be a proteome-wide topological tool

Cross-linking methods have long been employed for protein interaction [16–18] and protein topological studies [19–24]. Cross-linking can provide useful data on protein topological features and those of interacting partners by imparting new covalent bonds that serve as identifiable features. Because of the finite distance between cross-linker reactive groups, only amino acid residues that co-exist within this immediate radius can react and form a stable linkage. The observed cross-linked sites can be used to produce low resolution structures for proteins and complexes [25–27].

A large fraction of chemical cross-linker work has utilized N-hydroxy succinamide (NHS) esters as reactive groups since these show good selectivity for reaction with primary amine groups. This, coupled with the fact that lysine is an amino acid that occurs in the proteome with fairly high frequency (5–7%), increases chances for cross-link formation. Several recent reviews exist on the evolution of the cross-linking field and its use in probing native protein structure which comprises a majority of current cross-linker work, including reviews by Back [21], Sinz [28], and Leitner et al. [29]. However, as described by Leitner et al., chemical cross-linking has grown from very small segment of the proteomics community, but still accounts for a “niche” technique. Part of this stems from the difficulty in application to increasingly complex samples and the additional challenges that arise. The present review is focused on an area where cross-linking is just beginning to have impact; that of *in vivo* protein topological measurements. If further developed, this area can help fill an unmet need in the demanding challenges of *in vivo* protein topological studies. This review will attempt to describe the challenges unique to *in vivo* topological studies with cross-linking technologies and highlight a few successful glimpses of protein topologies features as they exist in cells.

2.1 Proteome-wide cross-linking greatly increases dynamic range challenges

Cross-linking methods have been used to make contributions to protein and proteome research at many levels from purified proteins and protein complex topological studies [19, 21, 27, 30] to interacting protein identification [31–33]. Most chemical cross-linking efforts have involved analysis on purified proteins and complexes and topological information is derived from sites of cross-linking that are identified. Research geared toward interacting protein partner identification typically does not require cross-linked *site* identification and the cross-linker serves to aid in co-purification of binding partners. Despite the topological information they could provide, cross-linked sites from these complex samples are rarely identified. Reasons for this include difficulty in identification of cross-linked peptides and low yield of cross-linked peptides that create detection dynamic range constraints. Typically proteins are cross-linked with bifunctional reactive molecules, digested with an enzyme, and subjected to mass spectrometry analysis. A few examples exist where top-down measurements have been applied directly to cross-linked intact proteins [34, 35]. Although this approach provides the most definitive method of identification of cross-linked sites, handling of intact cross-linked proteins poses an additional set of challenges. However, identification of cross-linked peptides after digestion is more widely used and can give rise to useful constraints for structural prediction or proteins and protein complexes. With *in vitro* studies, high cross-linker-to-protein ratios are possible and readily enable detection and analysis of cross-linked species with mass spectrometry. Thus, dynamic range issues are less severe in cross-linking studies with purified protein samples as compared with complex samples like cells or cell lysates. With the latter, production of high quantities of cross-linked species for all proteins becomes more difficult since cross-linker molecules have limited solubility. For example, with a single protein of approximate mass of 50,000 Da, one can anticipate approximately 25 different lysine sites could be cross-linker-reactive, whereas with an entire proteome this number could easily reach hundreds of thousands of different lysine sites. In addition, although they do exhibit differing cross-linker accessibilities, the 25 sites within a single protein are present at the same concentration. Over the entire proteome, individual site concentration levels vary widely, as do the protein abundance levels which is further increased by site accessibility. For experiments with cells or tissues, cross-linker penetration also becomes a limitation on reagent availability. It is unlikely that a single cross-linker structure would penetrate all subcellular locations equally well and therefore, the range of cross-linked product concentration levels is further increased. Finally, the much larger fraction of non-cross-linked peptides present in cross-linked proteome samples exacerbates the wide dynamic range demands of proteome applications. *In vivo*, cross-linked peptides comprise a much smaller fraction of the total number of peptides as compared with *in vitro* experiments. Detection of cross-linked peptides from these proteome samples normally requires enrichment by chromatographic and/or affinity purification techniques. Dynamic range challenges for cross-linker application to cells or complex samples are among the leading deterrents to successful application of cross-linker technology to *in vivo* topological studies. Methods used to help overcome this dynamic range issue include biotin-based affinity purification [36], click chemistry [37, 38] and strong cation exchange [39].

2.2 Proteome-wide application of cross-linkers creates new challenges for identification

For cross-linker application to complex samples such as cells, identification of cross-linked peptides can be challenging in the absence of molecular weight measurement for each linked peptide chain. Even in conventional proteomics measurements where peptides are isolated

fragmented and identified by database search, peptide identification attempts can fail when two or more peptides are co-isolated and fragmented. For purified samples, accurate mass measurement of cross-linked products may yield sufficient specificity to allow putative identification of cross-linked peptides. These putative peptide sequences can then be supported by complex fragmentation spectra. In extrapolation to proteome-wide applications however, accurate mass measurements of cross-linked peptide complexes from these samples are seldom unique. Fragmentation analysis of intact cross-linked peptide complexes can yield convoluted mass spectra that, without peptide mass information, are prone to failure with conventional proteomics search tools. Rinner *et al.* present a sophisticated search algorithm called xQuest which shows remarkable promise with cross-linked *E. coli* cell lysate samples [39]. Du *et al.*, [40] also present a platform that hold potential for proteome-wide cross-linker identification. Thus, advanced informatics can help improve upon this limitation, as can advances in cross-linker design.

2.3 Random cross-linking of proteins

An additional concern for cross-linker application to proteome-wide samples is that of “random passerby” cross-linking, where proteins that happen to be close to one another are cross-linked. This is a direct consequence of nonspecific nature of the reactive groups like NHS esters commonly used for cross-linking that possess no inherent preference for reaction with specific interacting proteins, i.e., true binding partners, over the random passerby proteins. Although definitive data has yet to emerge, random passerby cross-linking almost certainly takes place in complex samples. At first consideration, this type of product might be expected to confound proteome-wide applications. However, *in vivo* topological data acquired thus far suggest detection of random cross-linking in cells is not yet possible with current capabilities. In fact, random cross-links are likely to be lower in abundance than those between *bona fide* interacting proteins for several reasons. First, truly interacting proteins are likely resultant from random interactions that give rise to some specific biological advantage [3]. This advantage must originate from specific partner orientation and specific topological features, otherwise any random protein, with random orientation could yield similar advantage. This specificity differentiates true interactions from random passerby, nonspecific interactions where the relative orientation of binding partners is not restricted. Second, observed cross-linked peptides in mass spectrometry experiments normally on the order of femtomoles or more to yield detectable signals. Thus, any cross-linked peptide pair observed in the mass spectrometer must have formed hundreds of millions to billions of times or more during cross-linker reaction with cells. This fact really means that during the cross-linker reaction period of 10 to 30 minutes, the two protein species that were cross-linked and identified were close to one another with billions of examples. Furthermore, in all cases these interactions occurred with the same orientation so as to allow the same two lysine sidechains to be cross-linked. Third, the high frequency of these proteins being in the same proximity, with the same orientation is a hallmark of a specific protein interaction and means that cross-linking on cells and mass spectrometry detection is biased towards detection of those proteins that exhibit these features. The fact that many cross-linked peptides are observed where both sequences originate within the same protein, some with exactly the same sequence which define an unambiguous homomeric interactions, strongly suggests cross-linking experiments can yield useful topological information on non-random, specific interactions in cells [41–43]. Finally, the fact that many homodimeric species appear most abundant among cross-linked peptides identified from cells is supportive of the notion about prevalence of homomeric interactions derived from Kuriyan’s and Eisenberg’s insight mentioned above.

2.4 Advanced chemistry for *in vivo* topological studies

Cross-linker application to cells or complex biological samples requires advanced technology and informatics. Our lab reported development of cleavable cross-linker molecules we call Protein Interaction Reporter (PIR) technology several years ago [44]. PIR technology was recently reviewed [45] so will only briefly be described here. PIR molecules are cross-linkers that can be cleaved in the mass spectrometer to allow complex measurement and subsequent release of cross-linked peptides. PIR molecules also include biotin or other affinity groups to allow enrichment of cross-linked peptides from proteome samples. The PIR concept was pursued with the thought that, if peptides could be released from the cross-linker in the mass spectrometer, MS and MS/MS measurements on the released peptides could enable identification of the peptide sequences in much the same way as is performed in conventional proteomics experiments with LC/MS/MS on peptide mixtures. We have since reported development of several different PIR molecules using a variety of activation methods to enable peptide release [46, 47], development of PIR informatics strategies [48, 49] and PIR applications to live cells [41–43, 50]. In addition, several labs have reported development of chemical cross-linkers that are cleavable within the mass spectrometer [51–55], suggesting this approach may hold promise for proteome applications.

3.0 *In vivo* topological information from cross-linking studies

Despite the promise of cross-linking technology, at this point the reports of *in vivo* topology acquired by these methods are few. Nonetheless, the examples that do exist are beginning to provide a better picture of the challenges and the unique information that can be derived from cross-linking experiments with cells. Most efforts have thus far focused on the gram negative bacteria *Shewanella oneidensis* MR-1 and *Escherichia coli*. *S. oneidensis* is a bacterium that has generated significant interest due to its ability to reduce metal ions in solution. Cross-linking efforts yielded the first glimpse of interacting protein topologies involving outer membrane cytochrome proteins that are required for metal reduction. Two outer membrane decaheme cytochrome proteins, *omcA* and *MtrC* that are known to be required for electron transport processes that facilitate metal ion reduction, were identified as part of an interacting network by cross-linking on cells followed by lysis and immunoprecipitation of either *OmcA* or *MtrC*, digestion and LC/MS/MS identification of cross linked proteins [32]. In addition, these two proteins were shown to form a complex *in vitro* that can facilitate iron ion reduction in solution [56]. In subsequent untargeted work with *in vivo* cross-linking, cross-linked sites on *OmcA* and *MtrC* were identified yielding the first topological data on these and other interacting proteins as they exist in cells [43].

3.1 *In vivo* cross-linked sites are fewer than those observed *in vitro*

Application of PIR technology to bacterial cells resulted in the identification of reactive sites on approximately 1500 peptides from 400 proteins. Thus, on average 3 peptides were found labeled per protein. Of those sites in *E. coli*, 65 cross-linked peptide pairs were identified from 20–30 proteins. Thus, 2–3 cross-linked peptides were identified per protein from *in vivo* cross-linking experiments. This is fewer reactive sites per protein than is typically observed with *in vitro* cross-linking experiments on purified proteins and complexes. Similar numbers of cross-linked peptides per protein were observed by Rinner et al. in cross-linking experiments with *E. coli* cell lysates [39]. This is likely resultant from the large number of lysine sites presented by whole cell lysates or intact cells, compared with the finite concentration of the cross-linker. Under these conditions, the total number of reactive lysine sites is much greater than the number of cross-linker molecules. For *in vitro* experiments with a small number of proteins, this is likely not the case. For example, the

final concentration of cross-linker in cell-based experiments is typically 1 mM which is at or near the maximum achievable concentration. For *in vitro* cross-linking experiments with protein concentration in the range of 1 to 10 μ M and 1 mM cross-linker concentration, the ratio of cross-linker molecules to available lysine sites can greatly exceed 10:1, assuming an average protein contains 25 lysine amino acids per protein molecule. Thus one could conceivably react cross-linker molecules with every lysine in simple protein mixture. However, each additional protein included in the cross-linking reaction mixture rapidly increases the available lysine groups. A single proteome-wide application of cross-linker molecules is unlikely to yield a large number of cross-links for any single protein. Thus, in the absence of significantly increased cross-linker solubility, the anticipation that many cross-linked sites will be identified on large numbers of proteins is unlikely and perhaps, an unrealistic expectation. Nonetheless, improvements in cross-linker solubility and the use of multiple cross-linker additions will improve this area.

3.2 *In vivo* cross-linking experiments can yield data on disordered protein regions

Due to the reduced cross-linker-to-lysine ratio experienced with cross-linking experiments with cells, selectivity is observed in the sites that are cross-linked [41–43]. From the sites observed on proteins like OmpA, Skp, GAPDH and others, *in vivo* cross-linking reactivity appears to preferential target disordered protein regions. This could be due to the effect of local protein disorder and increased solvent (and cross-linker) accessibility that may result. Under reaction conditions with limited cross-linker, those sites with higher disorder and accessibility should appear more reactive. This was exemplified with the outer membrane protein OmpA, an outer membrane protein that appears to function as an ion channel in artificial membranes [57, 58]. While the N-terminal domain of this protein forms a highly ordered beta barrel structure upon crystal formation, the C-terminal domain is predicted to be disordered and in practice, resistant to attempts at crystal structure analysis [59]. The structure and function of the C-terminal domain have remained unknown, although hypotheses exist that suggest it may be involved in formation of a larger, 16 strand beta barrel [60]. Three *in vivo* cross-linked sites were reported on OmpA, all which occur within the C-terminal domain [41]. No cross-links were observed in the ordered, N-terminal domain of this protein. One OmpA cross-linked peptide pair observed within the disordered domain allowed unambiguous identification of an OmpA multimer that exists in *E. coli*, providing novel topological information on this protein and how it may function in cells.

Disordered regions are important in interactions among many proteins [7, 10, 11, 61, 62], and thus, *in vivo* cross-linking may in general provide unique opportunity to help improve understanding of protein function in cells for this class of interactions. To highlight another example, Figure 1 illustrates the cross-linked sites reported on the *E. coli* periplasmic chaperone protein, Skp [42]. This protein is thought to exist as a trimeric species in cells and serves as a chaperone for proteins destined for the outer membrane. Walton et al. reported *in vitro* measurement of the protein OmpA, a substrate for Skp bound within the cavity of Skp formed by the three finger-like projections of the “forceps” model [63]. The ends of these arms are highly disordered and as such, only two of the three Skp subunits are resolved in the trimer structure [64] (PDB entry 1SG2). PIR experiments with *E. coli* cells resulted in the identification of two cross-linked peptide pairs, including one between lysine 77 and lysine 89 in the sequence. This pair likely originated within a single subunit as indicated in Fig. 1A and are separate by 21.9 Å. The other peptide pair consisted of two identical tryptic peptide sequences: TLE**K**DVMAQR both of which included lysine 77 (in bold) as the site of PIR modification. This pair defined an unambiguous homodimeric interaction between two Skp monomers and presents the first data to supports the existence of Skp multimeric

topology *in vivo*. As shown in Fig 1C however, these two lysine residues occur on opposite sides of the two forceps fingers which requires that either the cross-linker wrap around the ends of the fingertips or that the fingers exhibit orientation different from that shown in the crystal structure, as might be expected given the high levels of disorder in this region. Future *in vivo* cross-linking studies with alternate PIR structures will shed further light on this interesting topology. Finally, given the preferential cross-linker reactivity for disordered regions, *in vivo* cross-linking can provide highly complementary topological data to that acquired from highly detailed crystallography and other conventional structural measurements.

3.3 *In vivo* cross-linking sites appear to also be sites of lysine acetylation

Zhang et al. reported the interesting finding in 2009 that many proteins in *E. coli* appear to be modified by acetylation on lysine side chains [65]. Their report focused on metabolic proteins and showed that lysine acetylation in *E. coli* is regulated in response to stress stimuli. Interestingly, several of the same metabolic proteins that were identified as lysine acetylation targets, such as serine hydroxymethyltransferase (GlyA), tryptophanase (tnaA), Glyceraldehyde 3 phosphate dehydrogenase (GAPDH) were also observed in our *in vivo* cross-linking studies. Furthermore, in all three examples, at least one of the reported lysine acetylation sites for each protein was found cross-linked *in vivo* and two (GlyA and tnaA) sites were reportedly found cross-linked in *E. coli* cell lysates [39]. Figure 2 illustrates crystal structure data on the metabolic protein tryptophanase (tnaA) [66] (PDB entry 2C44). This figure illustrates cross-linked sites identified in cells that include lysine 115, lysine 459, and lysine 409. Rinner et al also reported cross-linking with lysine 115 and lysine 459 in cell lysate cross-linking experiments [39]. Their report suggested this cross-link was intermolecular between two subunits and this seems most plausible, although cross-linking within a single subunit is also feasible, particularly for larger cross-linker structures. The distances between intra- and inter-linked sites based on the crystal structure are shown. Admittedly, linkages with lysine 115 could be within a single subunit in this case, although local protein topology precludes direct line-of-site linkage in both cases. Thus, the actual distances between lysine 115 and lysine 459, and 409 are likely much greater than those shown. In either case however, *in vivo* cross-linking results suggest high reactivity on lysine 115. Zhang et al., identified tnaA peptides containing lysine 115 in their acetylation study suggesting that this site is important for functional regulation of tnaA [65]. Given that some sites appear common to both datasets, the combined acetylation and cross-linking studies may prove useful for mapping topological features and dynamic effects inherent in response to stress and other stimuli. For example, if lysine acetylation were used as a modulator of metabolic activity in *E. coli* as suggested by Zhang et al., it would seem reasonable to anticipate that some such modifications may occur near regions that interface with other proteins. Future studies with *in vivo* cross-linking will likely include both lysine acetylation profiling measurements as well as altered metabolic stimuli to investigate these relationships.

Figure 3 illustrates the crystal structure of the protein GAPDH [67] (PDB entry 1DC6), reported acetylated lysine sites and cross-linked sites observed in cells. The lysine at position 191 was reported acetylated by Zhang et al. [65] and was found cross-linked in cells with lysine at position 107 (numbering based on PDB file). As with other examples, lysine 191 appears in a flexible loop region of the molecule that appears at the entrance to the NAD + cofactor binding pocket. All three observed cross-linked sites in GAPDH are associated with high predicted disorder as indicated by disorder prediction algorithms VLXT, CAN_XT, and VL3, that use various sequence lengths and datasets for training as described on website: www.pondr.com. Cross-linked peptides linking lysine 107 and that at position 122a (position 124 in the Uniprot sequence for *E. coli* GAPDH, accession number: P0A9B2)

were also reported. Interestingly, cross-linked species containing lysine 191 and 122a were not observed, despite the fact that these two appear very close in the crystal structure. Acetylation of site 191 may alter structure and binding of NAD⁺ and thus this site and its reactivity may be important to function. Finally, in PIR experiments with human cells we have also identified cross-linked sites on human GAPDH that strikingly, were found to be conserved with the link between lysine 107 and lysine 191 shown in Fig. 3 (2011 ASMS Proceedings, Denver, CO). In human GAPDH this link includes lysine 107 and lysine 194 and was the only GAPDH cross-link yet found in human cells; human GAPDH does not contain lysine in the analogous vicinity of lysine 122a of *E. coli*. The observation of conserved cross-linked sites from *in vivo* cross-linking with cells from highly divergent species is striking, but at this point will require additional examples and analyses to comprehend the significance. At the very least however, this correlation strongly suggests that results from *in vivo* cross-linking are non-random in nature.

4. Future directions of *in vivo* cross-linking for protein topological measurements

If it were possible, technologies that could penetrate into cells and tissues and provide records of protein topologies and topological features of their interactions could drastically change the current understanding of biological function. To achieve this, such technologies need the ability to provide protein-specific information; one needs to know what proteins are being measured. Since cells contain thousands of proteins, this requires the specificity of proteomic technologies. This technology also needs to provide location-specific information; one needs to know which other proteins are within the vicinity of each measured protein and if possible, what parts of the proteins are close to one another. Finally, this technology needs to provide these records with high sensitivity and wide dynamic range to be broadly applicable. Although far from mature, *in vivo* cross-linking technologies do provide many of the elements of this technology: protein specificity from the peptide sequences and location-specificity from the reactive sites and cross-linker finite length. Despite the relatively small number of proteins and cross-linked sites identified from *in vivo* studies thus far, these examples illustrate the unique type of information that can be obtained from these methods. The challenges for future directions relate to the required sensitivity and dynamic range of these measurements and to the difficulty in maintaining all desirable attributes in the cross-linker structure, like cleavability and affinity tags while minimizing undesirable ones like poor solubility and size. Future directions must include smaller cross-linkers which will increase demands on this balance. However, decreased cross-linker size is important for the overall molecular length, but even more but more so for decreased compound MW, increased solubility and increased cellular penetration. Current experiments are carried out under conditions of limited cross-linker availability. Increased solubility and cell penetration will improve both the number and types of proteins that can be studied *in vivo*.

Finally, advances in informatics technologies that can make use of cross-linking constraints are critically needed. Mapping cross-linked sites back to the known crystal structures is a fundamental part of developing *in vivo* topological capabilities. However, as was discovered with OmpA, proteins with highly disordered regions may not have relevant crystal structures. However, these proteins are ones where *in vivo* cross-linking technologies can have the greatest impact. In these cases, protein structure prediction capabilities could strongly synergize with *in vivo* cross-linking capabilities. For example, the use of only three cross-link constraints drastically reduced the large number of possible model structural models that result from docking predictions for OmpA dimers to two very similar models. Finally, even with the limited capabilities for *in vivo* cross-linking measurements that currently exist, making use of all the peptide pairs information is challenging. As *in vivo*

topological data become widely available, computational tools that can better utilize these cross-linking constraints are needed.

Acknowledgments

This laboratory is grateful for support from the National Institutes of Health through grant numbers: 5R01RR023334, 1R01GM097112, 5R01GM086688, and 1R01HL110879. JEB also thanks Juan Chavez and Xiaoting Tang for careful reading and helpful suggestions in the preparation of this manuscript.

References cited

1. Goldman AD, Samudrala R, Baross JA. The evolution and functional repertoire of translation proteins following the origin of life. *Biology direct*. 2010; 5:15. [PubMed: 20377891]
2. Caetano-Anolles G, Caetano-Anolles D. Universal sharing patterns in proteomes and evolution of protein fold architecture and life. *Journal of molecular evolution*. 2005; 60:484–498. [PubMed: 15883883]
3. Kuriyan J, Eisenberg D. The origin of protein interactions and allostery in colocalization. *Nature*. 2007; 450:983–990. [PubMed: 18075577]
4. Hughes RM, Waters ML. Effects of lysine acetylation in a beta-hairpin peptide: comparison of an amide-pi and a cation-pi interaction. *Journal of the American Chemical Society*. 2006; 128:13586–13591. [PubMed: 17031973]
5. Inoue M, Yasuda K, Uemura H, Yasaka N, et al. Phosphorylation-dependent protein interaction with *Trypanosoma brucei* 14-3-3 proteins that display atypical target recognition. *PloS one*. 2010; 5:e15566. [PubMed: 21203569]
6. Lange S, Sylvester M, Schumann M, Freund C, Krause E. Identification of phosphorylation-dependent interaction partners of the adapter protein ADAP using quantitative mass spectrometry: SILAC vs (18)O-labeling. *Journal of proteome research*. 2010; 9:4113–4122. [PubMed: 20568816]
7. Dunker AK, Oldfield CJ, Meng J, Romero P, et al. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC genomics*. 2008; 9 Suppl 2:S1.
8. Dyson HJ, Wright PE. Elucidation of the protein folding landscape by NMR. *Methods in enzymology*. 2005; 394:299–321. [PubMed: 15808225]
9. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nature reviews. Molecular cell biology*. 2005; 6:197–208.
10. Meszaros B, Simon I, Dosztanyi Z. The expanding view of protein-protein interactions: complexes involving intrinsically disordered proteins. *Physical biology*. 2011; 8:035003. [PubMed: 21572179]
11. Dagkessamanskaia A, Durand F, Uversky VN, Binda M, et al. Functional dissection of an intrinsically disordered protein: understanding the roles of different domains of Knr4 protein in protein-protein interactions. *Protein science : a publication of the Protein Society*. 2010; 19:1376–1385. [PubMed: 20506404]
12. Mark WY, Liao JC, Lu Y, Ayed A, et al. Characterization of segments from the central region of BRCA1: an intrinsically disordered scaffold for multiple protein-protein and protein-DNA interactions? *Journal of molecular biology*. 2005; 345:275–287. [PubMed: 15571721]
13. Dunker AK, Uversky VN. Signal transduction via unstructured protein conduits. *Nature chemical biology*. 2008; 4:229–230.
14. Bartesaghi A, Subramaniam S. Membrane protein structure determination using cryoelectron tomography and 3D image averaging. *Current opinion in structural biology*. 2009; 19:402–407. [PubMed: 19646859]
15. Fields S, Song O. A novel genetic system to detect protein-protein interactions. *Nature*. 1989; 340:245–246. [PubMed: 2547163]
16. Fancy DA. Elucidation of protein-protein interactions using chemical cross-linking or label transfer techniques. *Current opinion in chemical biology*. 2000; 4:28–33. [PubMed: 10679368]
17. Kluger R, Alagic A. Chemical cross-linking and protein-protein interactions-a review with illustrative protocols. *Bioorganic chemistry*. 2004; 32:451–472. [PubMed: 15530987]

18. Vasilescu J, Guo X, Kast J. Identification of protein-protein interactions using in vivo cross-linking and mass spectrometry. *Proteomics*. 2004; 4:3845–3854. [PubMed: 15540166]
19. Singh P, Panchaud A, Goodlett DR. Chemical cross-linking and mass spectrometry as a low-resolution protein structure determination technique. *Analytical chemistry*. 2010; 82:2636–2642. [PubMed: 20210330]
20. Sinz A. Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *Journal of mass spectrometry : JMS*. 2003; 38:1225–1237. [PubMed: 14696200]
21. Back JW, de Jong L, Muijsers AO, de Koster CG. Chemical cross-linking and mass spectrometry for protein structural modeling. *Journal of molecular biology*. 2003; 331:303–313. [PubMed: 12888339]
22. Scherer PE, Krieg UC. Cross-linking reagents as tools for identifying components of the yeast mitochondrial protein import machinery. *Methods in cell biology*. 1991; 34:419–426. [PubMed: 1943816]
23. Pillai S. Chemical cross-linking of a dimeric protein on a modified lectin matrix. A general probe for the chemical topology of oligomeric glycoproteins. *The Biochemical journal*. 1981; 193:825–828. [PubMed: 6796053]
24. Rappsilber J. The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *Journal of structural biology*. 2011; 173:530–540. [PubMed: 21029779]
25. Schulz DM, Ihling C, Clore GM, Sinz A. Mapping the topology and determination of a low-resolution three-dimensional structure of the calmodulin-melittin complex by chemical cross-linking and high-resolution FTICRMS: direct demonstration of multiple binding modes. *Biochemistry*. 2004; 43:4703–4715. [PubMed: 15096039]
26. Huang BX, Kim HY, Dass C. Probing three-dimensional structure of bovine serum albumin by chemical cross-linking and mass spectrometry. *Journal of the American Society for Mass Spectrometry*. 2004; 15:1237–1247. [PubMed: 15276171]
27. Dihazi GH, Sinz A. Mapping low-resolution three-dimensional protein structures using chemical cross-linking and Fourier transform ion-cyclotron resonance mass spectrometry. *Rapid communications in mass spectrometry : RCM*. 2003; 17:2005–2014. [PubMed: 12913864]
28. Sinz A. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass spectrometry reviews*. 2006; 25:663–682. [PubMed: 16477643]
29. Leitner A, Walzthoeni T, Kahraman A, Herzog F, et al. Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Molecular & cellular proteomics : MCP*. 2010; 9:1634–1649.
30. Chang Z, Kuchar J, Hausinger RP. Chemical cross-linking and mass spectrometric identification of sites of interaction for UreD, UreF, and urease. *The Journal of biological chemistry*. 2004; 279:15305–15313. [PubMed: 14749331]
31. Nittis T, Guittat L, LeDuc RD, Dao B, et al. Revealing novel telomere proteins using in vivo cross-linking, tandem affinity purification, and label-free quantitative LC-FTICR-MS. *Molecular & cellular proteomics : MCP*. 2010; 9:1144–1156.
32. Zhang H, Tang X, Munske GR, Zakharova N, et al. In vivo identification of the outer membrane protein OmcA-MtrC interaction network in *Shewanella oneidensis* MR-1 cells using novel hydrophobic chemical cross-linkers. *Journal of proteome research*. 2008; 7:1712–1720. [PubMed: 18303833]
33. Kuo MH, Allis CD. In vivo cross-linking and immunoprecipitation for studying dynamic Protein: DNA associations in a chromatin environment. *Methods*. 1999; 19:425–433. [PubMed: 10579938]
34. Kruppa GH, Schoeniger J, Young MM. A top down approach to protein structural studies using chemical cross-linking and Fourier transform mass spectrometry. *Rapid communications in mass spectrometry : RCM*. 2003; 17:155–162. [PubMed: 12512095]
35. Novak P, Giannakopoulos AE. Chemical cross-linking and mass spectrometry as structure determination tools. *Eur J Mass Spectrom (Chichester, Eng)*. 2007; 13:105–113.

36. Hatanaka Y, Hashimoto M, Kanaoka Y. A novel biotinylated heterobifunctional cross-linking reagent bearing an aromatic diazirine. *Bioorganic & medicinal chemistry*. 1994; 2:1367–1373. [PubMed: 7788299]
37. Chowdhury SM, Du X, Tolic N, Wu S, et al. Identification of cross-linked peptides after click-based enrichment using sequential collision-induced dissociation and electron transfer dissociation tandem mass spectrometry. *Analytical chemistry*. 2009; 81:5524–5532. [PubMed: 19496583]
38. Vellucci D, Kao A, Kaake RM, Rychnovsky SD, Huang L. Selective enrichment and identification of azide-tagged cross-linked peptides using chemical ligation and mass spectrometry. *Journal of the American Society for Mass Spectrometry*. 2010; 21:1432–1445. [PubMed: 20472459]
39. Rinner O, Seebacher J, Walzthoeni T, Mueller LN, et al. Identification of cross-linked peptides from large sequence databases. *Nature methods*. 2008; 5:315–318. [PubMed: 18327264]
40. Du X, Chowdhury SM, Manes NP, Wu S, et al. Xlink-identifier: an automated data analysis platform for confident identifications of chemically cross-linked peptides using tandem mass spectrometry. *Journal of proteome research*. 2011; 10:923–931. [PubMed: 21175198]
41. Zheng C, Yang L, Hoopmann MR, Eng JK, et al. Cross-linking measurements of in vivo protein complex topologies. *Molecular & cellular proteomics : MCP*. 2011
42. Yang L, Zheng C, Weisbrod CR, Tang X, et al. In vivo Application of Photocleavable Protein Interaction Reporter Technology. *Journal of proteome research*. 2011 *submitted*.
43. Zhang H, Tang X, Munske GR, Tolic N, et al. Identification of protein-protein interactions and topologies in living cells with chemical cross-linking and mass spectrometry. *Molecular & cellular proteomics : MCP*. 2009; 8:409–420.
44. Tang X, Munske GR, Siems WF, Bruce JE. Mass spectrometry identifiable cross-linking strategy for studying protein-protein interactions. *Analytical chemistry*. 2005; 77:311–318. [PubMed: 15623310]
45. Tang X, Bruce JE. A new cross-linking strategy: protein interaction reporter (PIR) technology for protein-protein interaction studies. *Molecular bioSystems*. 2010; 6:939–947. [PubMed: 20485738]
46. Yang L, Tang X, Weisbrod CR, Munske GR, et al. A photocleavable and mass spectrometry identifiable cross-linker for protein interaction studies. *Analytical chemistry*. 2010; 82:3556–3566. [PubMed: 20373789]
47. Chowdhury SM, Munske GR, Tang X, Bruce JE. Collisionally activated dissociation and electron capture dissociation of several mass spectrometry-identifiable chemical cross-linkers. *Analytical chemistry*. 2006; 78:8183–8193. [PubMed: 17165806]
48. Hoopmann MR, Weisbrod CR, Bruce JE. Improved strategies for rapid identification of chemically cross-linked peptides using protein interaction reporter technology. *Journal of proteome research*. 2010; 9:6323–6333. [PubMed: 20886857]
49. Anderson GA, Tolic N, Tang X, Zheng C, Bruce JE. Informatics strategies for large-scale novel cross-linking analysis. *Journal of proteome research*. 2007; 6:3412–3421. [PubMed: 17676784]
50. Tang X, Yi W, Munske GR, Adhikari DP, et al. Profiling the membrane proteome of *Shewanella oneidensis* MR-1 with new affinity labeling probes. *Journal of proteome research*. 2007; 6:724–734. [PubMed: 17269728]
51. Soderblom EJ, Goshe MB. Collision-induced dissociative chemical cross-linking reagents and methodology: Applications to protein structural characterization using tandem mass spectrometry analysis. *Analytical chemistry*. 2006; 78:8059–8068. [PubMed: 17134140]
52. Muller MQ, Dreiocker F, Ihling CH, Schafer M, Sinz A. Cleavable cross-linker for protein structure analysis: reliable identification of cross-linking products by tandem MS. *Analytical chemistry*. 2010; 82:6958–6968. [PubMed: 20704385]
53. Liu F, Goshe MB. Combinatorial electrostatic collision-induced dissociative chemical cross-linking reagents for probing protein surface topology. *Analytical chemistry*. 2010; 82:6215–6223. [PubMed: 20560670]
54. Kao A, Chiu CL, Vellucci D, Yang Y, et al. Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Molecular & cellular proteomics : MCP*. 2011; 10 M110 002212.
55. Petrotchenko EV, Serpa JJ, Borchers CH. An isotopically coded CID-cleavable biotinylated cross-linker for structural proteomics. *Molecular & cellular proteomics : MCP*. 2011; 10 M110 001420.

56. Shi L, Chen B, Wang Z, Elias DA, et al. Isolation of a high-affinity functional protein complex between OmcA and MtrC: Two outer membrane decaheme c-type cytochromes of *Shewanella oneidensis* MR-1. *Journal of bacteriology*. 2006; 188:4705–4714. [PubMed: 16788180]
57. Saint N, El Hamel C, De E, Molle G. Ion channel formation by N-terminal domain: a common feature of OprFs of *Pseudomonas* and OmpA of *Escherichia coli*. *FEMS microbiology letters*. 2000; 190:261–265. [PubMed: 11034289]
58. Arora A, Rinehart D, Szabo G, Tamm LK. Refolded outer membrane protein A of *Escherichia coli* forms ion channels with two conductance states in planar lipid bilayers. *The Journal of biological chemistry*. 2000; 275:1594–1600. [PubMed: 10636850]
59. Pautsch A, Schulz GE. High-resolution structure of the OmpA membrane domain. *Journal of molecular biology*. 2000; 298:273–282. [PubMed: 10764596]
60. Stathopoulos C. An alternative topological model for *Escherichia coli* OmpA. *Protein science : a publication of the Protein Society*. 1996; 5:170–173. [PubMed: 8771211]
61. Ren S, Uversky VN, Chen Z, Dunker AK, Obradovic Z. Short Linear Motifs recognized by SH2, SH3 and Ser/Thr Kinase domains are conserved in disordered protein regions. *BMC genomics*. 2008; 9(Suppl 2):S26. [PubMed: 18831792]
62. Simon SM, Sousa FJ, Mohana-Borges R, Walker GC. Regulation of *Escherichia coli* SOS mutagenesis by dimeric intrinsically disordered umuD gene products. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105:1152–1157. [PubMed: 18216271]
63. Walton TA, Sandoval CM, Fowler CA, Pardi A, Sousa MC. The cavity-chaperone Skp protects its substrate from aggregation but allows independent folding of substrate domains. *Proceedings of the National Academy of Sciences of the United States of America*. 2009; 106:1772–1777. [PubMed: 19181847]
64. Korndorfer IP, Dommel MK, Skerra A. Structure of the periplasmic chaperone Skp suggests functional similarity with cytosolic chaperones despite differing architecture. *Nature structural & molecular biology*. 2004; 11:1015–1020.
65. Zhang J, Sprung R, Pei J, Tan X, et al. Lysine acetylation is a highly abundant and evolutionarily conserved modification in *Escherichia coli*. *Molecular & cellular proteomics: MCP*. 2009; 8:215–225.
66. Ku SY, Yip P, Howell PL. Structure of *Escherichia coli* tryptophanase. *Acta crystallographica. Section D, Biological crystallography*. 2006; 62:814–823.
67. Yun M, Park CG, Kim JY, Park HW. Structural analysis of glyceraldehyde 3-phosphate dehydrogenase from *Escherichia coli*: direct evidence of substrate binding and cofactor-induced conformational changes. *Biochemistry*. 2000; 39:10702–10710. [PubMed: 10978154]

A

B

C

FIGURE 1.

A) Side-view of annotated crystal structure image of the *E. coli* periplasmic chaperone, Skp illustrating the observed intramolecular *in vivo* cross-link identified between lysine 77 and lysine 89. B) Alternate side view of crystal structure illustrating position of lysine 77 on the top subunit in this structure. The third subunit, here shown in green, was disordered in crystal structure analysis near the “forceps” tip and lysine 77 was not resolved in this subunit. C) Zoomed view of region encompassing both lysine 77 residues in two resolved subunits. These two lysine amino acids were observed from *in vivo* cross-linking indicating these sites were close to one another in cells.

A

FIGURE 2.

A) Annotate crystal structure image of the *E. coli* metabolic protein, Tryptophanase, tnaA, annotated to illustrate several cross-link sites observed in cells. B) Close up view of cross-linked sites between lysine 115 and lysine 406 and 459.

FIGURE 3.

Annotated crystal structure image of *E. coli* metabolic protein, Glyceraldehyde 3- phosphate dehydrogenase (GAPDH) illustrating *in vivo* cross-linked sites. Even though GAPDH contains more than 25 lysine residues, all observed cross-linked sites appear near the NAD⁺ cofactor binding pocket and are associated with regions of high predicted disorder.