Proteomics in 3D: Development of new technology and computational tools for structural analysis

Chunxiang Zheng

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

James E. Bruce, Chair

Robert E. Synovec

Matthew F. Bush

Program Authorized to Offer Degree:

Chemistry

University of Washington

Abstract

Proteomics in 3D: Development of new technology and computational tools for structural analysis

Chunxiang Zheng

Chair of the Supervisory Committee:
Professor James E. Bruce
Department of Genome Sciences

Proteins form complexes with specific structures to carry out biological function. Better understanding of protein complex structures will allow scientists to gain deeper insights on biological function, pathway regulation and disease mechanisms. But characterization of protein interactions is challenging. High-throughput methods, such as yeast two-hybrid and co-immunoprecipitation, provide little structural information for the formed complexes. X-ray crystallography and NMR enable high resolution structure measurement, but those methods have very low through-put, and could only be used on purified complexes. Chemical cross-linking coupled with mass spectrometry is a high through-put structure characterization method. It has not been widely applied to complex biological systems due to the technical difficulties of sample preparation and data analysis. Protein Interaction Reporter (PIR) recently developed in Bruce lab

in UW, has overcome some of the major technical difficulties in cross-linking technique. PIR has been proved to be applicable in multiple biological systems, and many novel structure measurements were made. Some of the sample preparation method improvements were covered in this dissertation. In addition to the advancements of sample preparation methods, a public database for cross-linking data, XLink-DB, was developed to accommodate the fast growing amount of data. XLink-DB is a data storage, analysis, and visualization platform. It has a web interface for user to upload and analyze their own results. XLink-DB automatically fetches related information from other public databases such as UniProt and PDB to help user analyze and visualize their results. It also contains a protein modeling and docking pipeline to generate model complex structures with cross-linking distance constraints. As the first data analysis and storage platform for cross-linking results, XLink-DB fills the need of a high through-put pipeline of data processing and visualization, greatly improves the efficiency of data analysis, and provides deeper insight into the data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Thank Dr. Bruce for academic advices and guidance. Thank everyone in Bruce lab and my committee members for their help and encouragement.

DEDICATION

I dedicate this dissertation for the ones who loved and trusted me, as well as the ones who hated and doubted me. I also dedicate this dissertation for the exhausting journey of graduate school, numerous of good or bad decisions I made along the way and the unknown future.

Chapter 1. INTRODUCTION

1.1 AN EXAMPLE

Before we jump into the details of this dissertation, I would like to use an example to illustrate how cross-linking technologies have been used to study the structure of Outer membrane protein A (OmpA) *in vivo*. The detailed discussion of this example is included in Chapter 2. OmpA is one of the most abundant outer membrane proteins in *E. coli*. [1] It was commonly believed to be one of the major structural proteins in the outer membrane. For some strains of *E.coli*, OmpA is also responsible for host pathogen interaction. [1] Most outer membrane proteins of *E.coli* form multimeric complexes, but OmpA was considered to be only forming monomer. [1-3] The structure of OmpA is a mystery because it shows some porin activity at certain conditions, but it is not big enough to form a porin as monomer. [4] A variety of different techniques have been used to examine the structure of OmpA, including X-ray crystallography, NMR, and non-denatured gel electrophoresis. [2, 5, 6] But these techniques cannot provide *in vivo* structure information. First in vivo structural measurement of OmpA was made in 2010 with chemical cross-linking technologies.[7] For the first time, OmpA was shown to be forming multimers in vivo based on crosslinking results. Potentially, the porin activity of OmpA could be explained by the formation of OmpA multimers. The results were quickly adapted in the society. In 2014, intact protein analysis were performed on OmpA, confirming the formation of OmpA multimers.[8] Recent results greatly advanced the understanding of OmpA in vivo, and this all starts from the cross-linking results.

1.2 PROTEIN STRUCTURE

Proteins are comprised of linear amino acid chain, starting from N-terminus of the first amino acid of the sequence to the C-terminus of the last amino acid. The linear sequence of amino acids is also called as the primary structure of a protein. On top of primary structure, proteins have three additional levels of structures including protein secondary, tertiary and quaternary structures. Protein secondary structure refers to highly regular local sub-structures. Alpha helix and beta sheets are the two major types of protein secondary structure.[9] They both have regular structures with saturated hydrogen bonding for donors and acceptors on the peptide back-bone. Based on their unique structural pattern, highly accurate algorithms were developed to predict secondary structures. Protein tertiary structure is the 3D structure of a single protein. It is formed by the secondary structure elements of the protein structure folds in three dimension to form a globular protein. The folding is driven by hydrophobic interactions.[10] For cytosolic proteins, the hydrophilic residues were exposed on the surface of a protein structure, the hydrophobic residues tend to be buried inside. Member proteins, on the contrary, could have hydrophobic residues on the surface because of the hydrophobic membrane environment. The protein tertiary structure is flexible and dynamic, but certain part of the structure could be locked by specific interactions such as salt bridge and di-sulfide bonds. The protein quaternary structure refers to the three-dimension arrangement of more than one protein subunit. When protein chains form quaternary structures, they are also called as protein complex. A protein complex can be formed by protein chains with same or different sequences. The complex formed with same sequences are homo-multimers. The complex with different sequences are hetero-multimers. The protein quaternary structure is driven by non-covalent interactions between protein chains.

1.3 PROTEIN INTERACTIONS

Many proteins, if not all proteins, form complex to carry out biological functions. Therefore, structural analysis of protein complex become the key of understanding biological functions. Traditional ways of characterizing protein interactions includes two major category. The first category includes methods such as affinity purification coupled to mass spectrometry (AP-MS) and yeast two-hybrid.[11-13] Both methods are high through-put methods. In affinity purification coupled to mass spectrometry method, the target protein will be tagged with an affinity purification tag. When the target protein is purified, proteins interacting with the target protein will also be purified. The purified protein complexes will then be digested and analyzed in mass spectrometry. AP-MS method is potentially a quantitative method which provide quantitative measurements of protein interactions. Yeast two-hybrid was first introduced by Field and Song in 1989.[11] This method is first applied in yeast cells, which gives the name of the method. The yeast cells were transfused with two plasmids, prey and bait. Both plasmids carry part of reporter gene and a gene which encodes a target protein. The reporter is only transcribed if the two target proteins interact. Therefore, protein interaction will be detected by observing the presence of reporter protein. AP-MS and yeast two-hybrid are detection methods of protein interactions. They can be used to generate large scale protein interaction networks, but they do not provide structural information of protein interactions.

The second category of protein interaction characterization methods are structural characterization methods, including X-ray crystallography and NMR.[14, 15] In X-ray crystallography method, the electron scattering pattern of protein crystal is used to reconstruct the three dimension structure of protein or protein complexes. The experiment is conducted on purified and crystallized protein complex. When the protein complex is hard to purify, such as

some membrane protein complexes, it is very difficult to perform X-ray crystallography experiment on those complexes.[16] NMR, on the other hand, also measures purified complexes. NMR is also a high resolution technique. Unlike X-ray crystallography, NMR measures protein complexes in liquid phase. Most of the high resolution protein complex structures were measured by these two methods. But both methods are low through-put technology requiring purification of interested protein complexes.

1.4 CHEMICAL CROSS-LINKING

Chemical cross-linking was introduced to be an alternative method to measure protein complex structures.[17] Chemical cross-linking utilize cross-linker to impart covalent bonds in protein complexes. The structure of cross-linker has three major parts. Two reactive groups and a spacer arm. Based on the selection of reactive groups, the cross-linker can react with different functional groups in a protein. The most popular choice is primary amine, given the high abundance of lysine in protein sequence and high reactivity of primary amine. Once two sites on a single protein or a protein complex are cross-linked, the distance between the two sites are known to be within the maximum length of the cross-linker spacer arm. In this way, distance constraints in the three dimension are measured, and will be used for reconstruct the protein structure or protein complex structure. Although the principle of chemical cross-linking is rather straightforward, it used to be limited in practical use due to a few technical challenges. First, it is difficult to identify the cross-linked sites. In a common cross-linking experiment, the protein complex is incubated with cross-linker to allow the covalent bonds to form. After quenching the reaction, the protein complex is then digested. Because the cross-linkers impact covalent bonds, the two cross-linked sites stay connected. The cross-linked peptides are then analyzed with mass spectrometry to identify the

peptide sequences. Because all the combination of tryptic peptides are considered in the search space, the size of search space grows exponentially with the increase of number of peptides. This is also known as the N-square problem. Second, multiple products and high dynamic range of the sample make cross-linked peptides to be very difficult to detect. Because primary amine reactive cross-linkers also react with water, and the experiment is often conducted in aqueous phase to reserve protein complex structure, majority of the cross-linkers hydrolyzes instead of reacting with proteins. The quantity of cross-linked peptides is usually low comparing to non-cross-linked peptides. The high dynamic range of the sample make the cross-linked products to be difficult to detect. Because of these reasons, chemical cross-linking coupled with mass spectrometry is usually only applied to purified protein and protein complexes.

1.5 PROTEIN INTERACTION REPORTER

In order to apply cross-linking technology to complex systems, Protein Interaction Reporter technology, PIR, was developed to address some of the technical difficulties which are mentioned above.[18, 19] The cross-linker used in PIR technology has two mass spectrometry labile bonds which can be selectively cleaved during low energy collision induced dissociation (CID) to release a reporter ion. The sum of the mass of the reporter ion and the two released peptides equal to the mass of the precursor ion. This mass relationship provides two advantages for identifying cross-linked peptides. First, the mass relationship could be used to distinguish cross-linked products with other non-cross-linked peptide. Second, because the two cross-linked peptides are separated in mass spectrometry. They can be identified based on the fragmentation pattern of each single peptide, therefore the search space for peptide identification will linear increase with database size. In addition to the two mass spectrometry cleavable bonds, the PIR cross-linker also has an affinity

purification tag on the reporter. With the affinity purification tag, the cross-linked peptides are enriched from non-cross-linked peptides. The dynamic range of the sample will then be decreased. Some of the technical advancement of PIR technology will be discussed in Chapter 2. Because of the unique advantages, PIR technology has been successfully applied to a variety of different complex systems such as *E.coli*, human, *Pseudomonas Aeruginosa* and yeast.[7, 18, 20-22] Large protein interaction networks were generated from these results.

1.6 FROM CROSS-LINKING TO PROTEIN COMPLEX STRUCTURE

The distance constraints from cross-linking experiments can be used to produce model structure for protein and protein complexes. Protein modeling methods usually generate thousands of candidate models for one modeling experiment. Ranking the models in order to correctly model the protein structure *in vivo* is not an easy task, especially with limited amount of *in vivo* experimental data. Chemical cross-linking methods provide valuable *in vivo* measurement to the protein or protein complex structure, therefore, they can be used to provide rankings for the generated models. Leading modeling algorithms such as I-TASSER, already has included cross-linking data as optional input to guide their modeling experiments.[23] Integrative Modeling Platform (IMP) uses leading docking algorithms PatchDock to generate protein complex models.[24, 25] The complex model candidates will be scored by how the distance measured on each model fits with the distance distribution. The distance distribution is predicted by using minimum and maximum distance constraints of the cross-linker as lower and upper bounds to generate a Gaussian distribution. Although these leading software tools have incorporated cross-linking data in their input, there's no high through-put platform to perform these computations automatically. XLink-DB was designed to handle large scale cross-linking data storage, analysis

and visualization problems.[26] It is the first cross-linking data repository with a variety of software tools which help scientist analyze and visualize their data. The functionalities of XLink-DB will be discussed in detail in Chapter 3 and Chapter 4.

Chapter 2. CROSS-LINKING MEASUREMENTS OF *IN VIVO* PROTEIN COMPLEX TOPOLOGIES.

2.1 SUMMARY

Identification and measurement of *in vivo* protein interactions poses a critical challenge in the goal to understand biological systems. The measurement of structures and topologies of proteins and protein complexes as they exist in cells is particularly challenging, yet critically important to improve understanding of biological function since proteins exert their intended function only through the structures and interactions they exhibit *in vivo*. In the present study, protein interactions in *E. coli* cells were identified in our unbiased cross-linking approach, yielding the first *in vivo* topological data on many interactions and the largest set of identified *in vivo* cross-linked peptides produced to date. These data show excellent agreement with protein and complex crystal structures where available. Furthermore, our unbiased data provides novel *in vivo* topological information that can impact understanding of biological function, even for cases where high resolution structures are not yet available.

2.2 INTRODUCTION

Protein interactions and topologies are key features that enable specificity, function and the evolution of highly integrated, regulated networks in biological systems. Primary challenges associated with the study of biological systems include identification of protein interactions and measurement of topological features of proteins and their interactions *in vivo*. Advancements such as the Yeast Two-Hybrid[11], co-immunoprecipitation[27], Tandem Affinity Purification tags[28]

have greatly increased the ability to identify hundreds or even thousands of interactions from complex biological samples[13, 27, 29, 30]. Despite the many thousands of protein interactions that are now known[31] however, for only a tiny fraction is there any knowledge of their *in vivo* topology. On the other hand, if topologies of interactions were more widely known, this information could improve understanding of underlying fundamental factors that drive interactions, improve development of highly specific modulators of protein interactions, improve interaction prediction capabilities, and improve comprehension on biological systems. Unfortunately, exceedingly few methods exist to allow unbiased measurement of protein-protein interaction topological features in cells.

Chemical cross-linking has great potential for *in vivo* interaction topological studies[32-34]. Cross-linked peptides contain information about interacting protein identities and can uniquely define regions of protein sequences that are near one another when proteins are present within the native cellular environment. Challenges associated with *in vivo* cross-linking analysis that have precluded this achievement include the difficulty in identification of cross-linked peptides and the severe dynamic range constraints resultant from the overwhelming majority of non-cross-linked peptides. Our efforts to overcome these challenges resulted in development of Protein Interaction Reporter (PIR) technology[35] that uses a novel type of cross-linker and mass spectrometry to identify peptides that are close to one another within protein complexes in cells. These efforts resulted in the first reported identification of cross-linked peptides from live cells[33] including the first *in vivo* identification of an interaction among two outer membrane cytochrome *c* proteins, an interaction that appears to be critical to electron transport properties of *Shewanella oneidensis*[36].

Here we present the first application of PIR technology to the study of interactions in *E. coli* cells where 65 cross-linked peptide pairs were unambiguously identified. To date, this constitutes the largest *in vivo* cross-linked peptide dataset ever produced. In this system, we are also able to compare many of our results with known protein and protein complex crystal structures that demonstrate excellent agreement with our *in vivo* data. Importantly, this comparative analysis was also used to define distance constraints that enable refinements of structural prediction of *in vivo* protein complexes never before possible. Furthermore, within our large-scale cross-linking dataset, we identified cross-linked peptides from the periplasmic, C-terminal domain of Outer membrane protein A (OmpA). These results illustrate for the first time that OmpA can exist within cells as a dimer, presenting new possibilities for improved comprehension of the function of OmpA and its possible role in ion transport. Finally, structural predications of this domain, together with our cross-linking data provide the first ever *in vivo* topological picture of periplasmic domain of the OmpA dimer and serve to illustrate the unique capabilities of PIR technology for *in vivo* interaction identification and topology studies.

2.3 MATERIAL AND METHODS

2.3.1 *Material*

PIR cross-linker was synthesized in-house following the protocol discussed elsewhere [33, 35, 37]. Tergitol solution (70% NP40 in water), urea, iodoacetamide (IAA) and ammonium acetate were purchased from Sigma-Aldrich (St. Louis, MO) and used without further purification. TCEP, monomeric avidin ultralink resin and mass spectrometry-grade Trypsin endoproteinase were purchased from Pierce (Rockford, IL). Amicon Ultra -0.5 mL 10K centrifugal filters were from Millipore (Billerica, MA), C18 Sep-Pak Cartridges were from Waters (Milford, MA), Macro Spin

Columns were from the Nest Group, Inc. (Southborough, MA) and phenylmethylsulfonyl fluoride (PMSF) was from GBionsciences (Maryland Heights, MO). Anti-OmpA antibody was a generous gift from Professor Prasadarao Nemani at the Children's Hospital Los Angeles and University of Southern California.

2.3.2 *In Vivo Cross-linking and Cell Lysis*

A fifty milliliter volume of *E. coli* K12 cell suspensions was harvested at O.D. 0.6-0.8 for cross-linking reaction. The cells were pelleted and washed 5 times with 1 mL phosphate buffer saline (PBS) each time before cross-linking. The cells were then suspended in 1 mL PBS solution and PIR cross-linker was added to the suspension with a final concentration of 1mM. The mixture was incubated at 4 °C for 1 hr. The cells were then washed 5 times with 1 mL PBS each time and then suspended in 1 mL PBS solution with 0.1% NP40. The cells were lysed by sonication and concentrated by centrifugation at 4 °C and 15K g for 45 min. The soluble fraction was used directly in the next step. The insoluble fraction was dissolved in 200 µL of 8M urea in 100mM Tris·HCl buffer and then the proteins were precipitated with cold acetone. The protein pellet was then dissolved in 200 µL urea buffer and diluted with 0.1% NP40 in PBS to 1 mL.

2.3.3 *Enrichment of the Cross-linked Products*

Strong cation exchange chromatography: The cell lysate was reduced with 5 mM TCEP and alkylated with 10 mM IAA. Then the sample was digested with trypsin (trypsin to sample ratio 1:300, specifically cleave at C-terminus of arginine and lysine) at 37 °C overnight. After digestion, the sample was purified with a C18 sep-pak column. Desalted peptides were then re-suspended in 0.5% formic acid and 5% acetonitrile and loaded on a HIL-SCX macro spin column. The columns were then washed with 5% acetonitrile and 0.5% formic acid to remove low charged peptides and

eluted stepwise with increasing concentrations of ammonium acetate solutions. Ammonium acetate was removed by speedvac before further analysis.

Avidin affinity purification: As a complementary method to the strong cation exchange method, avidin capture affinity purification can also enrich cross-linked products. The purification generally followed the previous literature[33]. In brief, for a 1 mL *E. coli* cell lysate, 100 µL avidin beads slurry was added and incubated at room temperature for 2 hrs. The beads were washed with 1 mL 100 mM NH_4HCO_3 five times. The beads were then suspended with 100 µL NH_4HCO_3 solution, reduced with 1 µL of 0.5 M TCEP and alkylated with 1 µL of 1 M IAA . The sample was incubated in 1 mL of 10 ng/µL trypsin solution at 37 °C for 2 hrs to digest proteins. Then 10 µL 100mM PMSF was added to the solution to quench digestion, and the mixture was incubated at room temperature for 30 min. The second avidin capture was performed by adding 100 µL avidin slurry to the mixture and incubated for 2 hrs. Then the beads were washed with 0.1% NP40 in NH_4HCO_3 solution 3 times and with NH_4HCO_3 3 times. Finally the beads were eluted with 75% acetonitrile and 0.5% formic acid 4 times, each time with 400 uL eluting buffer.

2.3.4 *Acid cleavage*

Cross-linked products were cleaved in 95% trifluoroacetic acid and 5% water with 2 hours incubation at room temperature. The reaction mixture was then re-suspended in 5-fold volume of cold triethylether and incubated in -80^{o}C freezer overnight. The mixture was then centrifuged at 14,000g for 30min to precipitate the peptides. The peptides were then re-suspended in 0.1% formic acid for mass spectrometry analyses.

2.3.5 *Mass spectrometry analyses*

The enriched cross-linked products were further separated with UPLC (Waters nanoAcquity, Milford, MA) and detected with LTQ-FT MS (Thermo Fisher Scientific Inc., Waltham, MA). A 30-cm long C18 column was made in-house by packing fused silica capillary (360 μm × 75 μm) with MAGIC C18AQ 100A 5U beads (Michrom Bioresources, Inc., Auburn, CA). A 2-cm long trap column was prepared similarly by packing fused silica capillary (360 μm × 100 μm) with MAGIC C18AQ 200A 5U beads. The following LC gradient was employed in both LC runs: 0-60 min 5%-60% buffer B, 60-85 min flushing with 80% buffer B and 85-120 min equilibrating with 5% buffer B (buffer A: 0.1% formic acid in DI water; buffer B: 95% acetonitrile and 0.1% formic acid in DI water). The MS instrument resolution was set at 25K to obtain high mass accuracy data. Two scan types, the first one with in-source CID (ISCID) off and the second one with ISCID on (80V), was applied alternatively. Searches with software tools *X-links*[38] or BLinks[39] were used to identify the PIR relationships. Relationship mass tolerance was set to 10 ppm. The *m/z* values of the identified released peptides and the LC elution times were then used to generate a mass-and-time targeted inclusion list, which was used in the second data-dependent LC-MS/MS run.

In the second LC-MS/MS, the same columns and gradient as the first LC-MS experiment were used to maintain the retention time information. The MS/MS experiments were set up as follows: one MS with 25K resolution was followed by 5 data-dependent MS/MS experiments with the precursor masses and retention time values obtained from the mass-and-time targeted list. Dynamic exclusion repeat and exclusion duration were both 15 sec. The acid cleavage sample was analyzed with the same LC-MS/MS method, except that none of the mass-and-time targeted inclusion list was used. Instead, the MS/MS precursors were selected according to their intensities.

A third LC-MS/MS was performed on the same column with the same gradient as the first and second runs. A mass-and-time targeted list of the identified cross-linked parent ions was used in this third run to trigger three MS/MS events for each MS event. The dynamic exclusion repeat and exclusion duration were both set as 5 sec and the CID energy as 27.

2.3.6 *Data Analysis*

The peak lists were generated with Hardklor (version 1.34)[40]. The LC-MS/MS data were searched against the *E. coli* K-12 database with Mascot (Version: 2.3.01). Three possible missed cleavages were allowed, the precursor error tolerance was 25 ppm and the fragmentation error tolerance was 0.6 Da. The remaining tag from the PIR cross-linker (mass= 99.0320) was treated as a variable modification on lysine residues or protein N-terminus. Carbamidomethylaion on cysteine is considered as fixed modification. Oxidation on methionine, loss of water on N-terminal glutamine and glutamate are considered as variable modifications. All MS/MS data were searched against *E.coli* database (Release date, 02-19-2009). In total, 4178 sequences were included in the database. All peptide IDs were filtered with expect score threshold 0.1. Minimum 2 unique peptides were required for each protein ID. Then the software tools *X-links* and BLinks were used to search for the identification of relationships. *X-links* parameters were: filter *m/z* range 300-2000, filter isotopic fit 0.0-0.2 and mass tolerance 10ppm. BLinks parameters were: PPM tolerance 10ppm, p-value threshold 0.1 and minimum data points 5. Finally manual inspection of the results was required to confirm that all the identified cross-links were correct.

2.3.7 *Protein Docking*

The OmpA dimeric structural model was computed with tools on the Symmdock web server[41]. The starting structural file was model P0A910 from the SWISS-MODEL Repository

containing residues from position 208 to position 337. A symmetry order of 2 was used. The top 100 models from the docking results were manually filtered with cross-linking data. Thirty-five Å was chosen as the distance threshold based on the distance distribution of our cross-linking results. Only 2 very similar models were left after evaluation with distance constraints and accessibility of the labeled sites. These two models are shown in **Figure. 7**.

The heterodimeric complex model of FKBP-type peptidyl-prolyl cis-trans isomerase FkpA (FkpA) and 30S ribosomal protein S6 (RpsF) was computed with Hex 6.1[42]. Structural files 1Q6H and 2QAL from RCSB Protein Data Bank (PDB) were used as starting structures for FkpA and RpsF. RpsF with part of ribosome was used as the receptor and the FkpA dimer was used as the ligand. The starting points for docking experiments were determined with the cross-linking results. Shape and electrostatics were used as the correlation methods. 3D was used as FFT mode. The docking results were refined with MM minimization. The receptor range was set to 180 with step size 7.5. The ligand range was set to 180 with step size 7.5. The twist range was set to 360 with step size 5.5. The distance range was set to 40 with scan step 0.8. We have also computed the dimeric structure model of OmpA periplasmic domain with Hex 6.1. Very similar results were obtained.

2.4 RESULTS/DISCUSSION

2.4.1 *PIR Technology*

The PIR approach has been described in detail previously[33, 35, 37, 43] and successfully applied on *Shewanella oneidensis* bacteria[33]. Here, the principles of PIR technology will be only briefly discussed. The structure and cleavage characteristics of the PIR cross-linker used for these studies are presented in **Figure 1A**. The employed PIR molecule contains two mass

spectrometry-labile bonds, two reactive groups and biotin which serves as an affinity purification tag. The two labile bonds are specifically cleaved during mass spectrometry analysis without fragmentation of other peptide bonds. All cross-linked product types, such as dead-end, intra- and inter-cross-linked peptides are cleaved at these labile bonds, yielding product type-specific mass relationships. For example, inter-cross-linked peptides can be cleaved to yield three components as illustrated in **Fig. 1B** (cleavage reaction). The observed fragment masses sum to match the measured intact cross-linked product mass. **Fig. 1C** illustrates three PIR mass relationships that define each product type. The inter-cross-linked pair M_p is cleaved to yield a reporter ion m_r and two released peptides m_{p1} and m_{p2} (**Fig. 1C**, inter-cross-link). The sum of these three fragment masses matches the inter-cross-linked product mass ($M_p = m_r + m_{p1} + m_{p2}$). Analogous mass relationships also define dead-end and intra-cross-linked PIR products. PIR cleavable features allow differentiation of cross-linked and non-cross-linked peptides and identification of all cross-link product types.

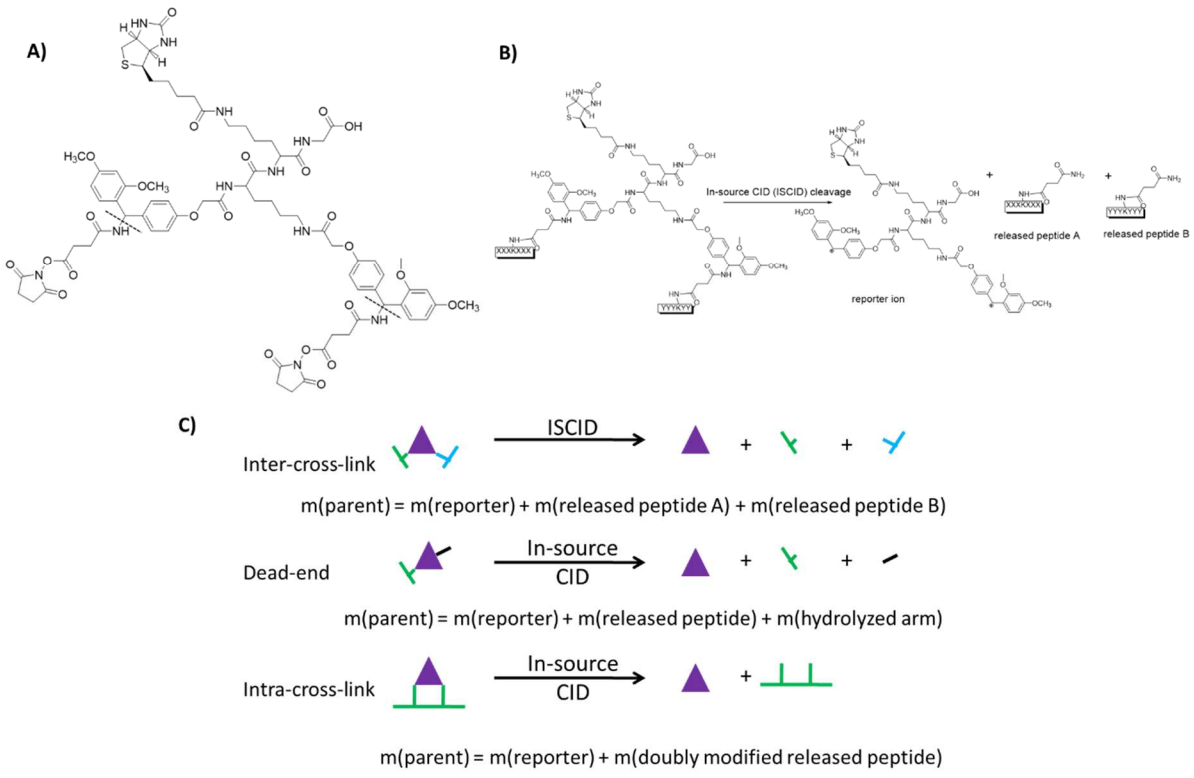

Figure 1 PIR technology

A two-step mass spectrometric identification strategy (**Figure 2)** was developed to enable relationship and cross-linked protein sequence identification. Precursor and ISCID scans are alternated throughout the whole LC-MS run to separately detect cross-linked precursors (M_p) and cleavage products (m_r and m_p). PIR mass relationships are identified with high mass measurement accuracy in neighboring scans[44]. Upon relationship identification, a second LC-MS/MS experiment where in-source CID activation was continuously applied to release all peptides was used to identify peptide sequences. Each of these peptides was selected for MS/MS analysis based on accurate mass and retention time. In the data analysis process, the residual PIR-induced mass modification was included during the database search accomplished with Mascot[45]. Accurate and reliable cross-linked sequence identification can be achieved even with whole genome database searches since each query is based on a single peptide fragmentation pattern. In contrast,

conventional non-cleavable cross-linked peptides can be challenging to identify with whole genome database searches and proteome-wide samples due to the lack of individual peptide mass information and complex fragmentation patterns[33, 44, 46]. Advanced features of the PIR strategy enable application directly on whole cells without purifying targeted proteins/complexes or sample fractionation prior to cross-linking. Finally, the affinity tag on PIR molecules allows enrichment of labeled peptides from complex sample digests.

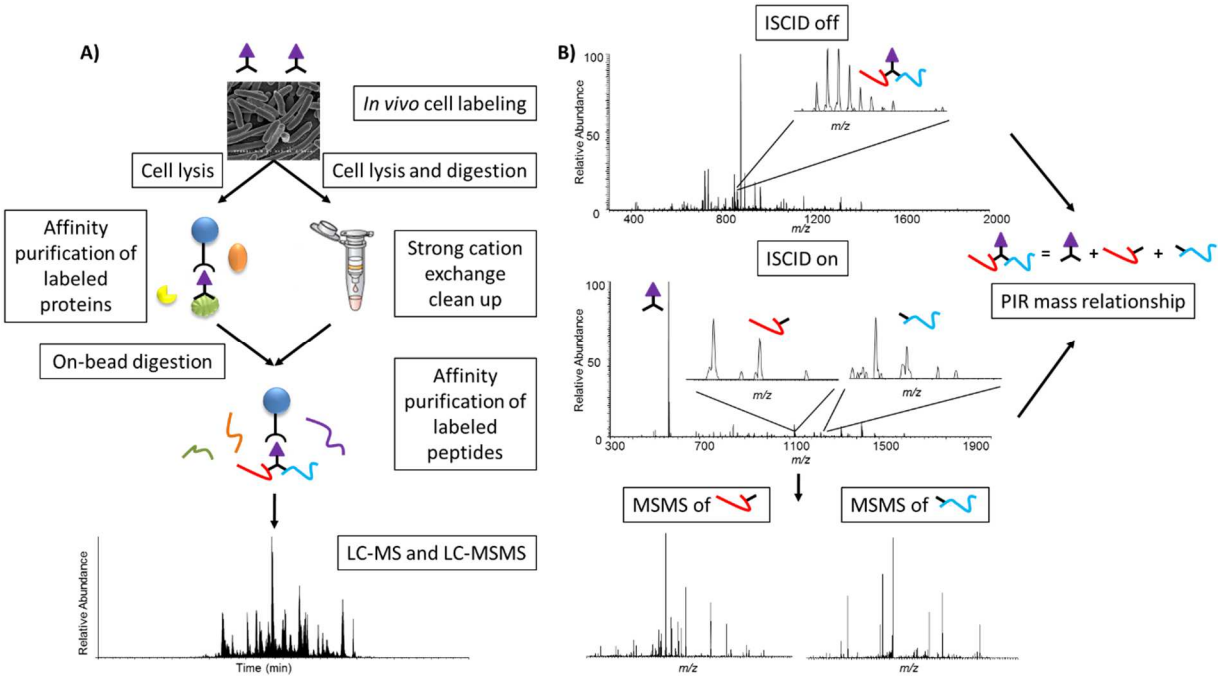

Figure 2 Sample preparation scheme

2.4.2 *Cross-linking on E. coli cells*

The application of PIR technology in *E. coli* cells was achieved through the steps outlined in **Figure 2**. Verification that cross-linking was achieved in cells included anti-biotin western blot analysis, extensive cell washing after labeling prior to lysis, avidin-biotin affinity purification of PIR-labeled proteins and peptides after digestion and mass spectrometry identification of the cross-

linked products. Anti-biotin western blot analysis is a highly sensitive method to detect biotinylated (or PIR-labeled) proteins that are present within cells or in subsequent cell washing steps prior to lysis. **Figure. 3** shows an anti-biotin western blot comparison of *E. coli* cell lysates from cells treated with or without PIR reaction. Both the soluble and insoluble fractions of the lysates from labeled cells have significantly more biotin-containing protein bands than those from the control cell lysates. Cells can undergo unintended lysis as well as normal secretion of proteins during cross-linking reaction. A negative anti-biotin western blot analysis of PIR-labeled cell wash solutions (normally within 5 washes with PBS – **see Figure. 4**) was required prior to cell lysis and sample preparation for mass spectrometry experiments. This requirement allowed the greatest chance for observation of topological features of protein interactions that were present in cells during PIR reaction.

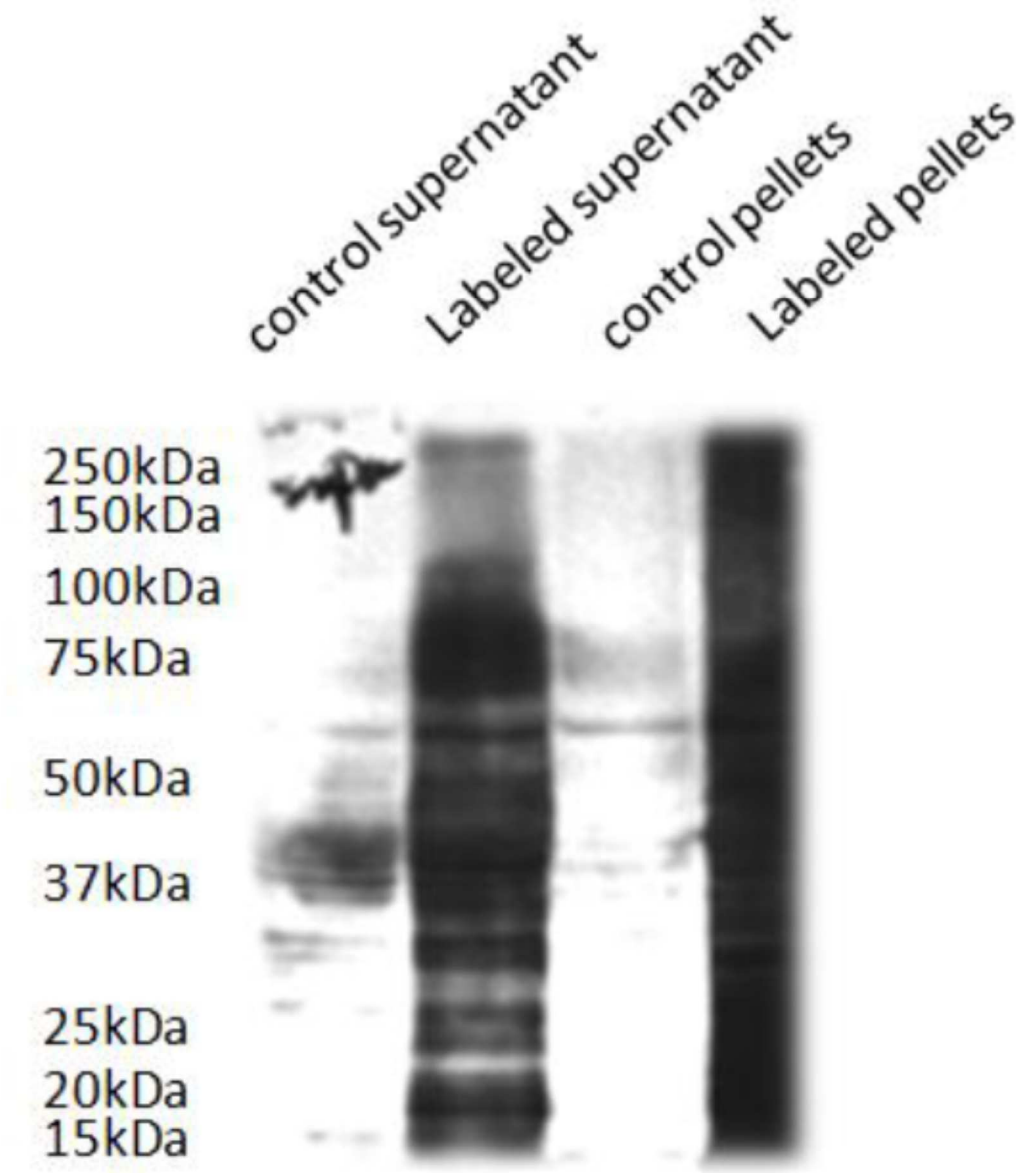

Figure 3 Anti-biotin western blot of control vs. labeled E. coli cells

The un-labeled and labeled E. coli cells were lysed, centrifuged, and the supernatant solutions were separated from the pellets, respectively. The pellets were re-dissolved with 8M urea and diluted for the subsequent purification experiments. On the membrane from left to right: Lane 1, control (un-labeled) E. coli cell lysate, supernatant portion; Lane 2, supernatant portion from the lysate of the labeled E. coli; Lane 3, re-dissolved pellet portion from the control E. coli; and Lane 4, re-dissolved pellets from the labeled E. coli. Equal amounts of total protein were loaded onto each lane.

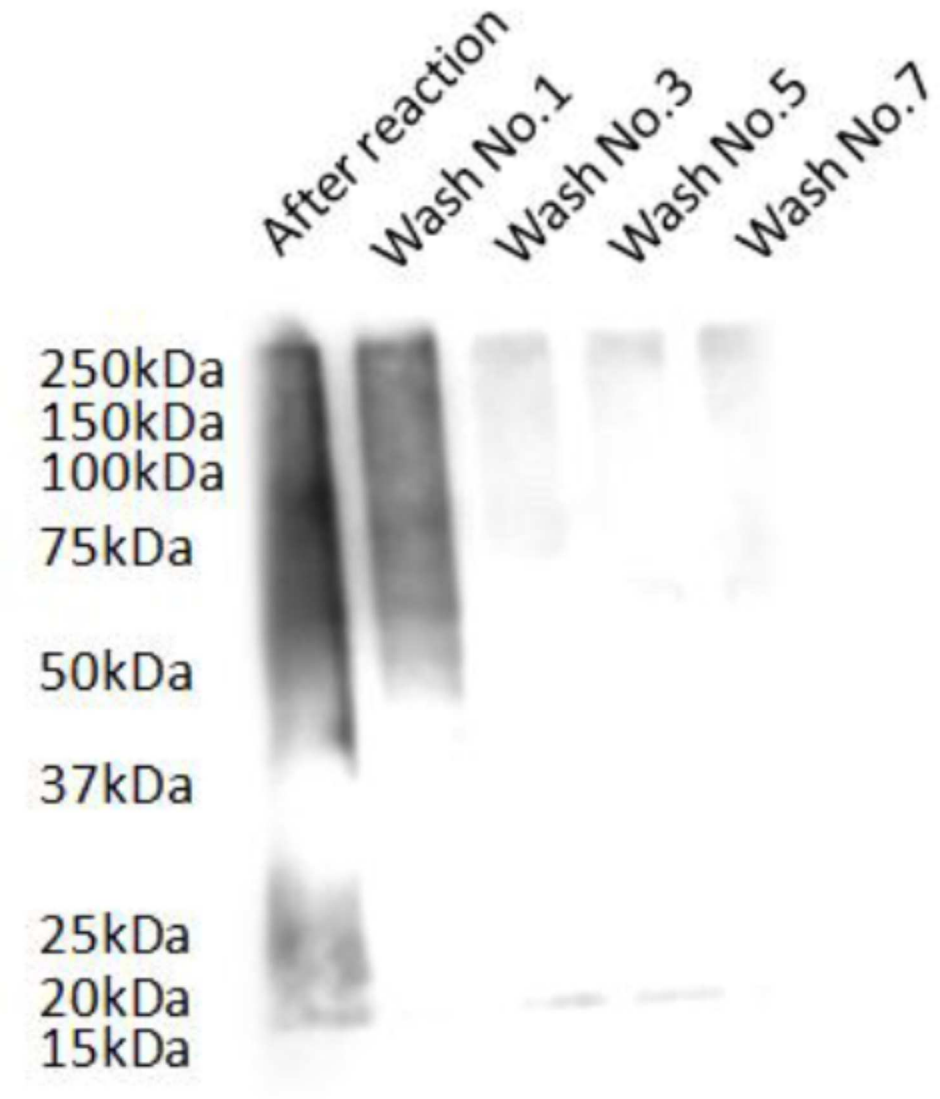

Figure 4 Anti-biotin western blot of cell washing prior to lysis

The labeled E. coli cells were washed with 0.1% NP40 in PBS seven times to get rid of self-lysed products. The supernatant solutions after cross-linking reaction and from the washes were loaded onto an SDS-PAGE gel, separated, transferred to a membrane and visualized by anti-biotin western blot analysis. Equal volumes of supernatant or wash solutions were loaded onto each lane. The amount of biotin-signal present in solutions from the 5th and 7th wash steps is not significant, so 5 wash steps were used in all subsequent experiments.

Shotgun proteomics analysis of acid-cleaved PIR labeled species from *E. coli* cells yielded a set of 1503 peptide sequences from 416PIR-reacted proteins. Analysis of these species to ascertain

their subcellular location was performed with psort2.0 which illustrated that approximately half of the observed PIR-labeled *E. coli* peptides were derived cytoplasmic proteins. Previous efforts with other gram negative bacteria indicated a similar fraction of cytoplasmic proteins are reactive with PIR molecules in cells and PIR reactive proteins were visualized on cell membranes and the cytoplasm with immunogold electron microscopic methods (Tang et al., J. Proteome Res. 2007).

A summary of all the inter-cross-linked peptide pairs from *E. coli* cells is listed in **Table 1**. PIR experiments on *E. coli* cells resulted in identification of 65 cross-linked pairs in total, which is the largest *in vivo* cross-linking dataset produced to date. *E. coli* was chosen as a biological system for application and demonstration of PIR technology since many X-ray crystal, NMR and EM structures of *E. coli* proteins and protein complexes exist[47]. PIR technology complements those efforts by enabling *in vivo* studies of protein complex topologies, supporting structural assignments in cells for cases where crystal structures are known, and by providing novel information for proteins and interactions without prior structural data. The PIR results in **Table 1** include inter-cross-linked peptides from homo- and heterodimeric interactions, as well as many which likely result from cross-linking of intra-molecular interactions. Identified PIR cross-link relationships can be validated by selecting cross-linked products for MS/MS. Fifteen out of 65 cross-linked peptide pairs were subjected to validation experiments. In all cases, identified peptide relationships resultant from precursor and ISCID scan comparisons were correctly assigned. The identified reactive sites are indicated in **Table 1** with an underline on each sequence. The distances between reactive sites were measured from corresponding X-ray crystal or NMR structures if available. A few of the identified protein complexes are discussed in detail in the following paragraphs, including those with and without previously determined crystal structures. The known structures are compared to the *in vivo* cross-linking data which showed excellent agreement.

Furthermore, this dataset provides new information on protein complexes without known structures. These results demonstrate the unique utility of the PIR cross-linking strategy to enable protein complex topological studies in cells.

Table 1 Inter-cross-linked pairs from in vivo PIR application

Index	distance (Å)	sequence	gene	expect score	sequence	gene	expect score
1	5.4	ILLINPTDSDAVGNAVKMANQANIPVITLDR	rbsB	6.00E-11	QATKGEVVSHIASDNVLGGK	rbsB	2.30E-06
2	X	STDISVKTDQK	osmY	0.0063	VKAALVDHDNIK	osmY	3.70E-06
3	X	STDISVKTDQK	osmY	0.0063	GVEGVTSVSDKLHVR	osmY	9.10E-07
4	X	VKAALVDHDNIK	osmY	3.70E-06	DAKEGSVK	osmY	0.021
5	X	VKAALVDHDNIK	osmY	3.70E-06	VDSSMNKVGNFMDDSAITAK	osmY	2.40E-07
6	X	GYAGDTATTSEIKAK	osmY	3.40E-07	SVKNDLKTK	osmY	0.022
7	X	GVEGVTSVSDKLHVR	osmY	9.10E-07	DAKEGSVK	osmY	0.021
8	X	GVEGVTSVSDKLHVR	osmY	9.10E-07	VKAALVDHDNIK	osmY	3.70E-06
9	X	GVEGVTSVSDKLHVR	osmY	9.10E-07	GYAGDTATTSEIKAK	osmY	3.40E-07
10	14.5	GLTFTYEPKVLR	tnaA	0.0017	HFTAKLK	tnaA	0.019
11	22.2	GLTFTYEPKVLR	tnaA	0.0017	YADMLAMSAKK	tnaA	6.00E-06
12	15.3	TGKQLPCPAELLR	tnaA	0.0075	GLTFTYEPKVLR	tnaA	0.0017
13	27	KDAMVPMGGLLCMK	tnaA	5.40E-05	TGKQLPCPAELLR	tnaA	0.0075
14	12.7	GAEQIYIPVLIKK	tnaA	6.80E-05	GLTFTYEPKVLR	tnaA	0.0057
15	20.2	AVEIGSFLLGRDPKTGK	tnaA	0.054	GAEQIYIPVLIKK	tnaA	6.80E-05
16	**X**	**GIKDVVTQPQA**	**ompA**	**0.00029**	**GIKDVVTQPQA**	**ompA**	**0.00029**
17	X	FGQGEAAPVVAPAPAPAPEVQTKHFTLK	ompA	0.00016	GIPADKISAR	ompA	0.0045
18	10.2	GIKDVVTQPQA	ompA	0.00029	GIPADKISAR	ompA	0.0045
19	15	FGQGEAAPVVAPAPAPAPEVQTKHFTLK	ompA	0.00016	GIKDVVTQPQA	ompA	0.00029
20	20	SSIPVFGVDALPEALALVKSGALAGTVLNDANNQAK	mglB	3.20E-07	KAIEQDAK	mglB	0.045
21	5.5	SSIPVFGVDALPEALALVKSGALAGTVLNDANNQAK	mglB	3.20E-07	VPYVGVDKDNLAEFSKK	mglB	3.60E-05
22	**17.2**	**TKLEKDVMAQR**	**skp**	**0.00014**	**TKLEKDVMAQR**	**skp**	**0.00014**
23	23.5	TKLEKDVMAQR	skp	0.00014	QTFAQKAQAFEQDR	skp	0.00011
24	X	NPQKNLYTFK	hdeB	0.0032	VIEYCKK	hdeB	0.06
25	X	MNAPDIKALFSSVR	ynhG	4.30E-05	VGQKIPNPTWTPTAGIR	ynhG	0.024
26	X	MNAPDIKALFSSVR	ynhG	4.30E-05	LVGQNQTYTVQEGDKNLQAIAR	ynhG	0.0027
27	13.1	KVVMTGPSKDNTPMFVK	gapA	0.00014	KHITAGAK	gapA	0.0024

28	16	TVDGPSHKDWR	gapA	0.0092	KHITAGAK	gapA	0.00046
29	X	VLAQKAVR	ybaY	0.09	VLAQKAVR	ybaY	0.09
30	X	TEGKQSPFSFVLSFNPADVQPNAR	ybaY	1.20E-05	VLAQKAVR	ybaY	0.09
31	18.7	YMENSLKEQEKLGIK	fkpA	7.00E-05	VKSSAQAK	fkpA	0.012
32	**28.1**	**VKSSAQAK**	**fkpA**	**0.012**	**VKSSAQAK**	**fkpA**	**0.012**
33	**21.8**	**QSIHSAHAKTLDTQGLR**	**kduI**	**0.00024**	**QSIHSAHAKTLDTQGLR**	**kduI**	**0.00024**
34	14	AAETNVAKSEAEKR	ytfQ	0.015	GITLKIADGQQK	ytfQ	0.006
35	**18.3**	**LDNMATKYR**	**lpp**	**0.026**	**LDNMATKYR**	**lpp**	**0.026**
36	10.5	SKATNLLYTR	dps	0.0051	KATVELLNR	dps	0.0053
37	16.4	EAKDLVESAPPAALK	rplL	0.00043	VAVIKAVR	rplL	7.00E-06
38	17.4	NLTGKAEADAALGR	glyA	0.0082	GGSEELYKK	glyA	0.005
39	8.6	NLTGKAEADAALGR	glyA	0.0082	YAEGYPGKR	glyA	0.00073
40	8.2	ANITVNKNSVPNDPK	glyA	0.0022	YAEGYPGKR	glyA	0.00073
41	17	ANITVNKNSVPNDPKSPFVTSGIR	glyA	0.0058	GGSEELYKK	glyA	0.005
42	X	YRLGETGDAIAKQTR	yqjD	0.063	SKAEQALKQSR	yqjD	0.0014
43	31.2	GRNVVLDKSFGAPTITK	groL	0.023	GIVKVAAVK	groL	0.0059
44	24.6	RVVINKDTTTIIDGVGEEAAIQGR	groL	0.0058	GIVKVAAVK	groL	0.0059
45	13.8	RVVINKDTTTIIDGVGEEAAIQGR	groL	0.0058	GRNVVLDKSFGAPTITK	groL	0.023
46	31.2	NVVLDKSFGAPTITK	groL	3.90E-05	GIVKVAAVK	groL	0.0026
47	24.6	VVINKDTTTIIDGVGEEAAIQGR	groL	0.0098	GIVKVAAVK	groL	0.0026
48	22.8	VVINKDTTTIIDGVGEEAAIQGR	groL	0.0098	NVVLDKSFGAPTITK	groL	3.90E-05
49	22.8	RVVINKDTTTIIDGVGEEAAIQGR	groL	0.0058	NVVLDKSFGAPTITK	groL	3.90E-05
50	X	NGGVAGNTTVNQKGR	flu	0.00047	TTINKNGR	flu	0.0027
51	X	GAKAIGTTGR	purA	0.0004	LKAFDHR	rpsJ	0.0045
52	X	ATLKPEGQAALDQLYSQLSNLDPK	ompA	2.80E-06	KAIEQDAK	mglB	0.045
53	X	TKHAVTEASPMVK	rpsF	0.0055	VKSSAQAK	fkpA	0.012
54	X	AQVAQIAGKPSSEVSMIHAR	osmE	6.20E-07	VLAQKAVR	ybaY	0.09
55	X	YADMLAMSAKK	tnaA	6.00E-06	KLLDEGR	tufB	0.1
56	X	TKPHVNVGTIGHVDHGK	tufB	0.00011	DAKEGSVK	osmY	0.021
57	X	KLLDEGR	tufB	0.1	GVEGVTSVSDKLHVR	osmY	9.10E-07
58	X	STDISVKTDQK	osmY	0.0063	KAGEGAK	rbsB	0.071
59	X	LDNMATKYR	lpp	0.026	DAKEGSVK	osmY	0.021
60	X	YRLGETGDAIAKQTR	yqjD	0.063	GFAGTVKR	rplC	0.0058
61	X	VKSVTFCAR	pta	0.019	GGSEELYKK	glyA	0.005

62	X	SNVPALEACPQKR	rpsL	0.0081	SHALNATKR	rpmB	0.00087
63	X	FGAKAISTIAESK	gadA	0.014	TSAALAKMQER	accD	0.001
64	X	NAIASVKAINAAR	sdaA	0.0075	GIGPAYEDKVAR	purA	0.0064
65	X	SKEHTTEHLR	yqjD	0.033	TAEICEHLKR	glpK	0.028

2.4.3 *PIR results of protein complexes with known structures*

As Kuriyan and Eisenberg discussed in their insightful review[48], regulated protein interactions are the inevitable consequence of colocalization and adaptive mutation. Thus, the most prevalent interactions are likely to involve homomultimeric interactions, since the probability of colocalization is maximal for proteins of identical sequence. However, these can be challenging to identify with cross-linking approaches, because it can be difficult to unambiguously determine whether the two cross-linked peptides come from the same protein subunit or two separate subunits. However in many cases in this study, cross-linked peptides with identical amino acid sequences were observed that allow unambiguous identification of homomultimeric interactions, since both peptides can come only from the same protein sequence and occur only once in each sequence. An exciting example is the periplasmic chaperone Skp which is believed to form a homotrimeric protein complex[49, 50]. Skp is involved in the folding and insertion of proteins in the outer membranes of many gram-negative bacteria, for example such as Outer membrane protein A (OmpA)[50-52]. As such, Skp plays a critical role in prevention of misfolding and aggregation of many outer membrane proteins. Improved knowledge of the *in vivo* structure and topology of Skp can increase understanding of this important process. The crystal structure of Skp illustrates one hairpin-shaped α-helical extension from each of three subunits that form the trimer (PDB entry 1SG2, **Figure 5A**). These extensions form the fingers or arms of a "three-pronged grasping forceps" model and form the cavity that serves in its chaperone function[49, 53].

However, the tips of the α-helical extensions are highly charged and flexible regions. As a result, only two of the three tips appear in the crystal structures, with the third tip disordered and not resolved[49]. Interestingly, *in vivo* PIR results demonstrate that at least two tips of the "prongs" were cross-linked in cells by the PIR cross-linker (**Figure 5A** red-red cross-link). The distance between the two cross-linked sites is approximately 17.2 Å based on measurements from the crystal structure. These data are the first such visualization of the Skp chaperone complex in cells and support the forceps model *in vivo*. In addition, this same site (Lys85) was found to be cross-linked to another peptide from the Skp complex (**Figure 5A** red-yellow cross-links). In this case, it is unclear whether this cross-link is within or between subunits, because the second cross-linked site (Lys97) may reside on the same subunit as the first, but may also reside on another subunit. According to the distance measurements from the crystal structure, the intra-subunit cross-link (distance 23.5 Å) is more likely than the inter-subunit cross-link (distances 31.9 Å). Further development and application of PIR technology can help resolve these sites. Nonetheless, the data presented here demonstrate that *in vivo* cross-linking can yield useful topological data in protein regions that, because of charge, flexibility or other factors, present difficulties with crystal structure measurements.

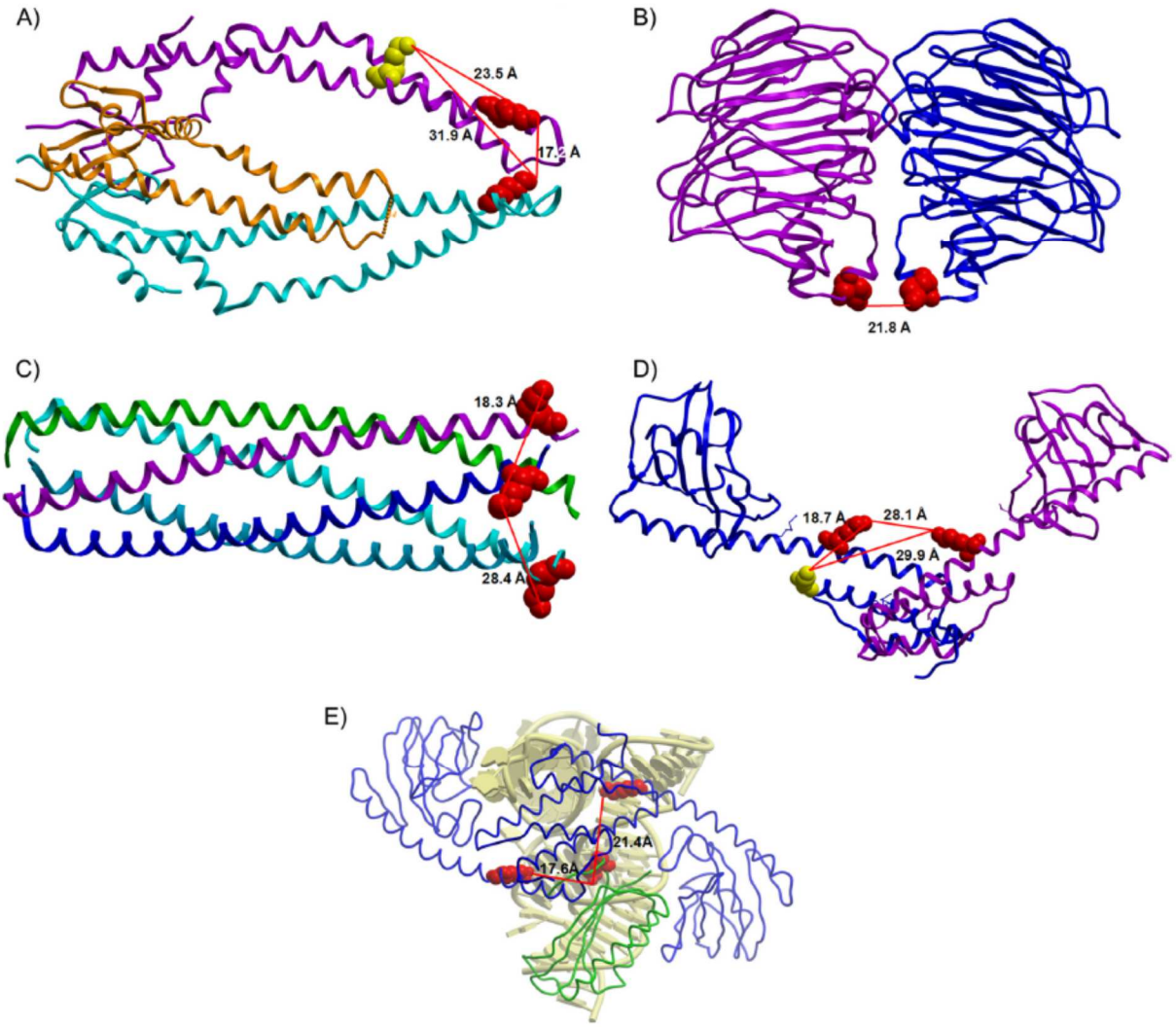

Figure 5 Cross-linked relationships shown on the protein complex crystallography structures A, Skp trimeric complex (PDB entry 1SG2). Red: lysine 85; Yellow: lysine 97. B, KduI dimeric complex (1XRU). Red: lysine 13. C, Lpp multimeric model (2GUV). Red: lysine 75. D, FkpA dimeric complex (1Q6H). Red: lysine 110; Yellow: lysine 85. E, The docking result of FkpA dimer and 30S ribosome (2QAL) based on the PIR cross-linking result. Blue: FkpA dimer; Green: 30S ribosomal protein S6 (RpsF); Yellow: Ribosomal RNA proximal to RpsF. Remainder of ribosome is not shown.

Several other homomultimeric protein complexes were identified in these PIR studies on *E. coli* cells. For example, 5-keto-4-deoxyuronate isomerase (KduI) has been reported to exist as a homohexamer[54] (PDB entry 1XRU), and serves as an important metabolic enzyme that supports the use of pectin as a carbon source in some bacteria. Two identical peptides from KduI were

identified as cross-linked species in the present *in vivo* PIR studies. The crystal structure for two subunits and the observed cross-linked sites are shown in **Figure 5B**. The distance between these two sites is 21.8 Å based on the crystal structure model, which matches very well with the PIR cross-linker arm length range[39]. Another example is murein lipoprotein (Lpp), which is an abundant homomultimer complex[55] (PDB entry 2GUV) in *E. coli* that is important for maintenance of cell envelope integrity. Two cross-linked sites were observed on the C-termini of Lpp chains from *in vivo* PIR experiments. The cross-linking distance between the two adjacent subunits is 18.3 Å, and 28.4 Å between the non-adjacent subunits (**Figure 5C**). For each peptide pair, the cross-linked precursor, released peptide, MS/MS identification of each released peptide and the verification of the relationship are shown in sequence.

2.4.4 *PIR results of protein complexes without known structure*

The unbiased *in vivo* PIR cross-linking strategy is especially advantageous for discovering novel protein interactions along with topological information for the interacting interfaces. The PIR approach can also reveal topological information on protein regions where it may be difficult to acquire with other methods like X-ray crystallography or NMR. Two examples will be presented here. One is a novel interaction between FKBP-type peptidyl-prolyl cis-trans isomerase FkpA and 30S ribosomal protein RpsF. The other is the discovery of OmpA periplasmic region homodimer. The OmpA C-terminal domain has remained resistant to crystallization, despite more than 30 years of effort to study this protein[56]. *In vivo* measurements present new opportunities to identify structural and topological features that can help improve understanding of function.

The first example is the study of FkpA topology and interactions. Proper folding of proteins in cells is essential to cell structures and functions. In this process, chaperones catalyze the correct

polypeptide folding, prevent and remove incorrect or incompletely folded products and assist translocation of proteins. Generally, there are two types of chaperones; one that catalyzes the proper disulfide exchange and another to attain correct cis/trans isomerization of peptidyl-prolyl bonds[57]. Chaperones can also be grouped based on their different subcellular localizations. The major cytoplasmic chaperones in *E. coli* include the well-studied DnaK and GroEL, while the periplasmic chaperones are less well understood. FkpA is a periplasmic bifunctional chaperone and cis/trans peptidyl-prolyl isomerase. The structure and function of this protein have been studied due to its interesting combination of two functional roles and its localization in the periplasmic region of the cells[58-60]. The present study provides new *in vivo* information on the structure and interactions of FkpA. Lys110 from FkpA was observed in several different cross-linked relationships, including a hetero-protein interaction, a homodimer protein interaction and an intra-FkpA cross-link. This kind of multiply cross-linked site appears hyper-reactive possibly due to accessibility, local protein disorder, or charge. Lys110 is labeled with red color in **Figure 5D-E** (PDB entry 1Q6H). FkpA is known to form a homodimeric complex from ultracentrifugation and crystallography measurements[58, 59]. The data presented here are the first to visualize such dimeric FkpA complex structures *in vivo*. The three helices on the N-terminal of each monomer were reported to be essential for maintenance of the dimeric structure[61] and are observed to be the most reactive regions for cross-linking in this study (**Figure 5D** red spot). A pair of peptides from FkpA with identical sequence was cross-linked, indicating that the cross-link was formed between two subunits of this dimeric protein complex (**Figure 5D**, red-red cross-link). The distance between these two sites was measured as 28.1 Å based on data from the crystallography structure. The same site (Lys110) was also cross-linked to a 30S ribosomal protein RpsF which represents a novel finding from these efforts. Using these

identified sites and the known structures of each proteins, the *in vivo* structure of an FkpA-RpsF protein complex was modeled with docking software *Hex* 6.1[42]. The result shown in **Figure 5E** is the first ever prediction of the *in vivo* complex of FkpA dimer and RpsF as it resides within the 30S ribosome. For clarity, only RNA proximate to RpsF is displayed in this figure. In addition to its periplasmic functional role, FkpA is known to interact with ribosomes. Previous studies on the isomerase activity of the FkpA-ribosomal protein complex showed that *in vitro* FkpA demonstrated isomerase activity when associated with 50S or 70S ribosomal protein complexes[62]. Other results from yeast-two-hybrid and high throughput experiments (*E. coli* interaction database, http://ecid.bioinfo.cnio.es/) demonstrated that FkpA can interact with 30S ribosomal protein RpsD. The results presented here support the interaction between FkpA and the 30S ribosomal protein complex, and demonstrate that an additional 30S subunit, RpsF, can interact with FkpA *in vivo*. In addition, another cross-link between this hyper-reactive site and a nearby lysine within the FkpA sequence was also identified (Lys110-Lys85). This is shown as the red-yellow cross-link on **Figure 5D**. However, this pair is most likely an intra-protein cross-link (distance 18.7 Å), but could also be an inter-protein cross-link (29.9 Å). Further cross-linking studies with alternate PIR structures may resolve these two possibilities. However, the results presented here are the first ever to reveal regions within the sequences of the complex components that are close to one another *in vivo*. On FkpA, this region is located on the center of the protein, which is supported by previous protein functional studies[61] where the chaperone and isomerase functional regions of FkpA were reported to localize to the distinct N- and C-terminus separately. Deletion experiments involving each terminal showed that these two domains can function independently, leaving the center of FkpA available for maintenance of the structure[62]. Interestingly, many other lysine residues exist near this hyper-reactive site. However few were

observed to be labeled or cross-linked in these experiments. All of these non-labeled sites are located on the opposite side of the protein complex. Such specific cross-linking and localization of these lysine residues indicate that, in addition to solvent accessibility and the presence of lysine residues, other factors affect the likelihood that reactive sites will be observed as cross-linked or labeled peptides.

A novel homomultimeric Outer membrane protein A (OmpA) interaction was also identified in the *in vivo* PIR studies presented here. OmpA in *E. coli* has been the subject of study for more than 30 years. Highly abundant and conserved among gram negative bacteria, OmpA is thought to play a key role in bacterial outer membrane structural stability, serve as a receptor for bacteriocins and phages, promote adhesion in pathogenic strains that transmit food-born infection[56, 63, 64]. As suggested[65], OmpA is perhaps the most important structural protein in *E. coli*. The first crystal structure of the N-terminal 171 residues of OmpA showed that OmpA can form an 8-stranded anti-parallel β-barrel[66]. The C-terminal domain of OmpA has been resistant to crystallization and its role in OmpA function has remained elusive. A structural model of this domain has been developed and accepted by many researchers, which is based on a highly homologous protein RmpM from *Neisseria meningitides*[67]. Despite these efforts, the oligomeric state of OmpA *in vivo* has remained unclear[55, 65, 68-70]. In general, membrane proteins are the most difficult proteins to study structurally, because they can be strongly influenced by the presence of the lipid bilayer, are difficult to overexpress and fold properly, and frequently contain disordered regions[71].

In vivo PIR experiments with *E. coli* have yielded several cross-linked peptide pairs from OmpA. Interestingly, all the observed cross-linked sites are in the C-terminal domain of OmpA, including lysine residues 213, 294, and 338. These data are supportive of the RmpM-based

structural model and can help answer key questions of *in vivo* OmpA topology. Based on prediction of sequence disorder[72, 73] and protein binding sites[74], the C-terminal domain of OmpA is highly disordered as shown in **Figure 6B** and it also contains several potential binding sites as shown in Figure 6C. This suggests that local protein flexibility of disordered regions and the proximity of binding sites affect which *in vivo* cross-linked sites are likely to be observed. These regions are most difficult to study structurally with most conventional techniques and PIR technology may provide unique insight on protein interactions that involve disorder. Other proteins investigated for disordered content also show good correlation between disorder and PIR reactive sites including, GAPDH (Fig.6A), Skp and GroEL[71].

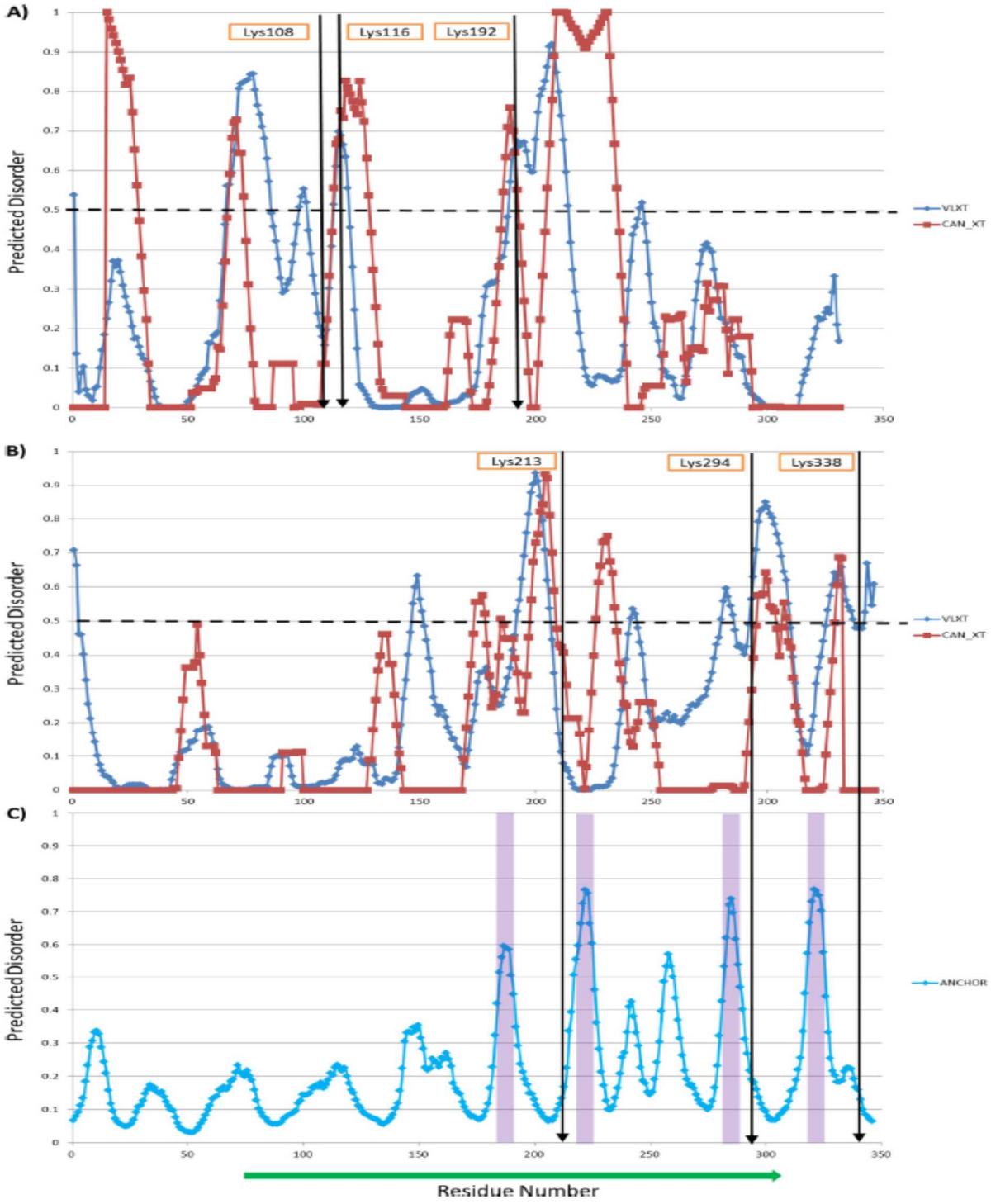

Figure 6 Prediction of disordered regions and binding sites in *E. coli*

A) Prediction of disorder in GapA; B) Prediction of disorder in OmpA; and C) Prediction of binding regions in OmpA.

Another important advantage of PIR technology is the potential to learn about native protein structures and interactions during cross-linker application. This is especially advantageous for membrane proteins since *in vivo* cross-linking with PIR technology takes place prior to cell lysis while the proteins still reside within their native environment or cellular location. Therefore, the potential to obtain useful topological information on membrane proteins is a compelling aspect of this approach. As such, PIR technology enabled the first observation of the OmpA dimer *in vivo* and the visualization of the interacting interface. Among these data, the most interesting result is that a homodimeric cross-linked species with two identical peptides containing lysine 338 was observed (**Figure 7**) and validated. Previous efforts to purify outer membrane protein complexes from *E. coli* resulted in identification of OmpA dimer bands on non-denaturing gels[55]. Folding studies of OmpA have also indicated higher order structures may exist[68]. We also observed OmpA dimer bands with anti-OmpA western blot analysis of samples from cross-linked *E. coli* cells. The present PIR data are first to unambiguously demonstrate that OmpA can form homomultimeric complexes in cells. To further evaluate this result, SymmDock[41] was used to compute the dimeric structure of the OmpA periplasmic domain. The model structure of the C-terminal domain was used as input in Symmdock which was used to compute putative structural orientation of the dimer. The top 100 models from docking results were evaluated against the cross-linking constraints from PIR studies. Using 35 Å as a distance threshold for cross-linked sites, 92 of the model structures were eliminated. Furthermore, only two very similar model structures remained after testing for accessibility of all reactive lysine residues observed in cross-linked pairs (discussed below). PIR data, together with computational analysis allows prediction of *in vivo* dimeric model structures of OmpA periplasmic domain (**Figure 7**). In addition to the homodimer cross-link, three other OmpA cross-links were also identified and validated. Among

these cross-linked species, two were considered to be intra-protein, and the other one was ambiguous and could be either intra- or inter- protein cross-linked (**Figure 7D**).

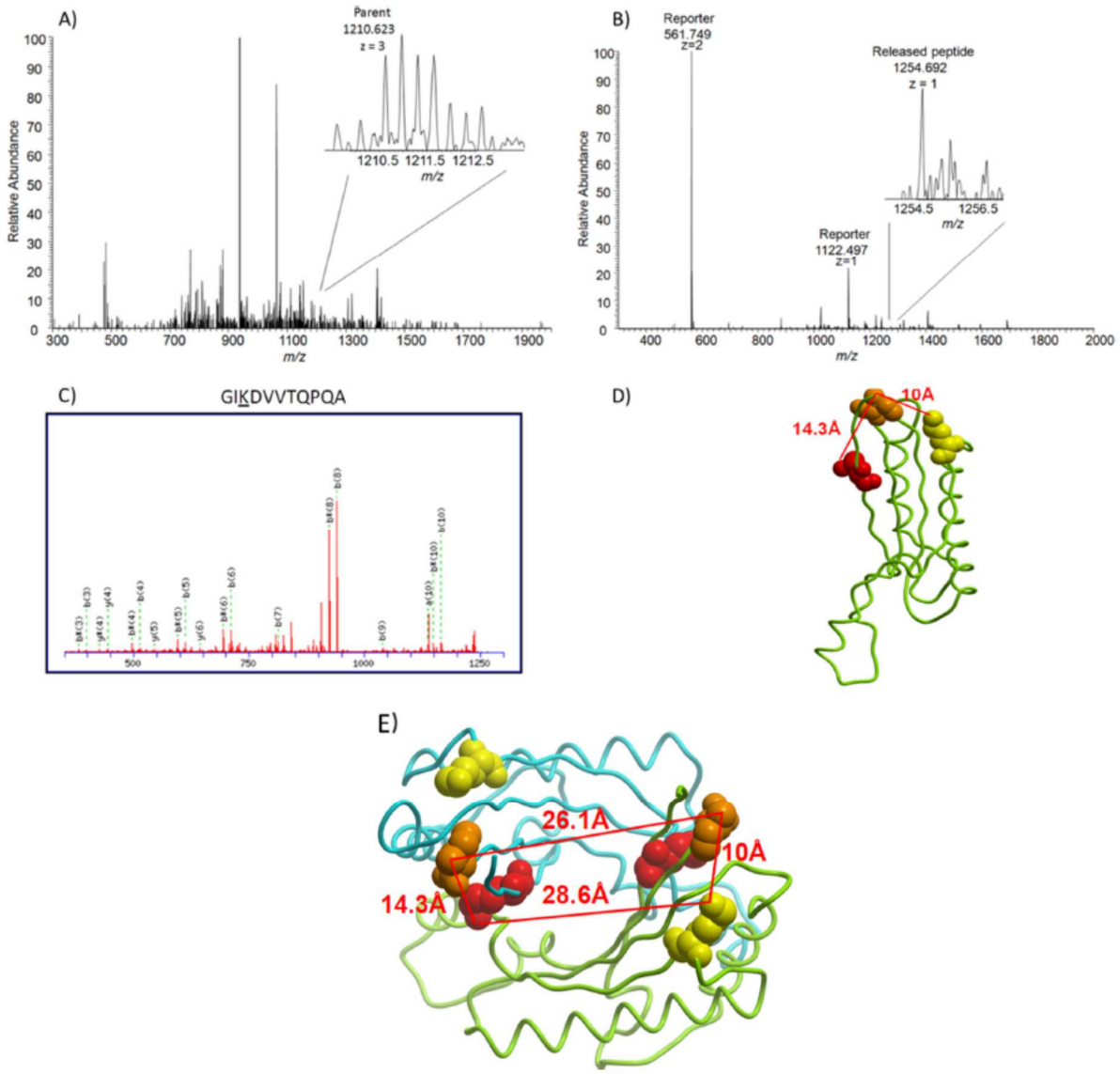

Figure 7 OmpA dimeric model and mass spectra of the homodimer cross-linking relationship

A, Precursor scan showing the homodimer cross-linked complex. Inset: isotopic distribution of the parent ion. B, ISCID scan showing the reporter and the released peptide. C, MS/MS spectrum of the OmpA released peptide with the sequence and assignment of fragments. D, OmpA monomer model. Red: Lysine 213; Yellow:Lysine 294; Orange: Lysine 338. E, A dimeric structural model from docking results (viewing angle: top-down). Red: Lysine 213; Yellow: Lysine 294; Orange: Lysine 338. The distances between each lysine are labeled in red.

It is important to consider this new topological data in the context of previous results on OmpA. First, OmpA has been demonstrated to function as a porin with temperature-dependent pore size[4, 75]. However, the smaller 8-strand beta barrel observed at lower temperature has no free pathway for the ions to get through the outer membrane and the ion pair-gated model derived from this structure cannot be used to explain the temperature effect[6]. Second, single channel conductance measurements[75] suggest full length OmpA can form a 16-strand beta barrel at physiological temperatures while truncated OmpA containing only the N-terminal domain yields only smaller channels. Incorporation 8 additional beta strands from the C-terminal domain has been proposed to be the origin of the larger pore structure[76] although this was not supported by secondary structural measurements[77]. The cross-linking data presented here show that OmpA can exist as a dimer in cells. Thus a dimeric beta barrel structure may be present, where 8 strands from each OmpA monomer combine to yield a larger channel. This has also been observed with other beta barrel containing multimeric proteins like Skp. Interactions within the C-terminal domain may be important for dimer stabilization, and if so, achieved crystal structures to date would fail to show the dimeric structures since full length OmpA crystals have not yet been obtained. Furthermore, this result is consistent with the *in vitro* observations that show larger OmpA channel formation only when the C-terminal domain is present[4].

Chapter 3. XLINK-DB: DATABASE AND SOFTWARE TOOLS FOR STORING AND VISUALIZING PROTEIN INTERACTION TOPOLOGY DATA

3.1 INTRODUCTION:

Protein interactions support most biological function and are directed by shapes or topologies of the interacting proteins. Improved measurements of protein interaction topologies in cells are needed to increase our understanding of how protein interactions carryout their life supporting functions. Chemical cross-linking with mass spectrometry has been used to study protein structures and complex topologies for several years [19, 78-93]. Most prior applications have been limited to either purified proteins or complexes due to the complexity and wide dynamic range presented by complex biological samples. Recent technical advancements of the chemical cross-linking methods achieved in a number of labs have allowed this technique to be extended to complex systems [7, 90, 94, 95]. Successful applications of chemical cross-linking to studies of intact virus particles, cell lysates, and even intact bacterial and human cells suggest that in the future, cross-linking methods may provide a majority of structural and topological data on protein complexes as they exist in cells or other complex samples[7, 20, 95, 96].

As is the case with most large-scale biological data, its usage among investigators in biochemistry, biophysics, cellular and molecular biology, as well as proteomics requires that new tools be developed to visualize, share and compare these results. This is especially true for large-scale cross-linking data since current growth in data quantity exceeds manual data analysis capabilities. Furthermore cross-linking with mass spectrometry datasets are unique in that they

contain multiple tiers of information on protein sequence, interaction, and structural levels for which no single existing data analysis tool can sufficiently support. Often data analysis requires comparison of cross-linking results with existing crystal structure data if available. In addition, cross-linking data are often compared with existing protein interaction data. If previously unknown interactions are discovered, the cross-linked site information can be superimposed by computational docking of interacting structures. These steps can require hours of efforts even with only a few cross-linked peptide pairs in a single experiment and this approach becomes intractable for hundreds of cross-linked peptides.

Here we report development of XLink-DB which was designed to serve both as a storage site and an online data processing and visualization tool to enable analysis of large-scale cross-linked peptide datasets. Importantly, XLink-DB will be useful among biological and proteomics research communities since it provides new analysis capabilities and improved access to complex cross-linking topological data. XLink-DB allows users to upload their cross-linking data and populate a relational database, as well as browse existing datasets. XLink-DB automatically retrieves related protein sequence information from UniProt[97] and high resolution structure information from the Protein Data Bank (PDB)[98]. If relevant structures are available, cross-linked site annotation is automatically performed with XLink-DB and visualized within the incorporated protein structure viewer[99]. The cross-linking data is also visualized in a protein interaction network view with an embedded web-based Cytoscape tool[100]. The data stored in XLink-DB will be compared to existing protein interaction databases such as IntAct[101] and EciD[102]. We anticipate that XLink-DB will be a useful tool and benefit the proteomics research community as well as all researchers interested in protein topologies and interactions.

3.2 OVERVIEW

The XLink-DB website was developed with PHP 5.5 and JavaScript, data analysis tools were programmed with Java 1.6 and data were stored in a MySQL database. The functionality of the website also depends on both Java applets and flash plug-in. As shown in **Figure 8**, the website contains two major modules: 1) Data upload, process and storage and 2) Data visualization. Five different views (interaction network, protein structure, search, site and table views) are available for cross-linked peptide data analysis. Interaction network view shows the protein interaction network generated from the dataset. Protein structure view shows the cross-linking peptide pairs on the existing PDB structure. A key feature of XLink-DB is the ability to map cross-linked sites on protein complexes for which individual proteins crystal structures exist, but no co-crystal have been reported. Site view is designed to display the sites when the co-crystal structure does not exist. Search view is a sub-network of the dataset. The table view is a summary of the dataset in a table. To help users get familiar with the features of the database, we have created a video tutorial which can be found in the help page. In addition, we have also put tooltips on some parameters to guide the users. Details on each module are discussed below.

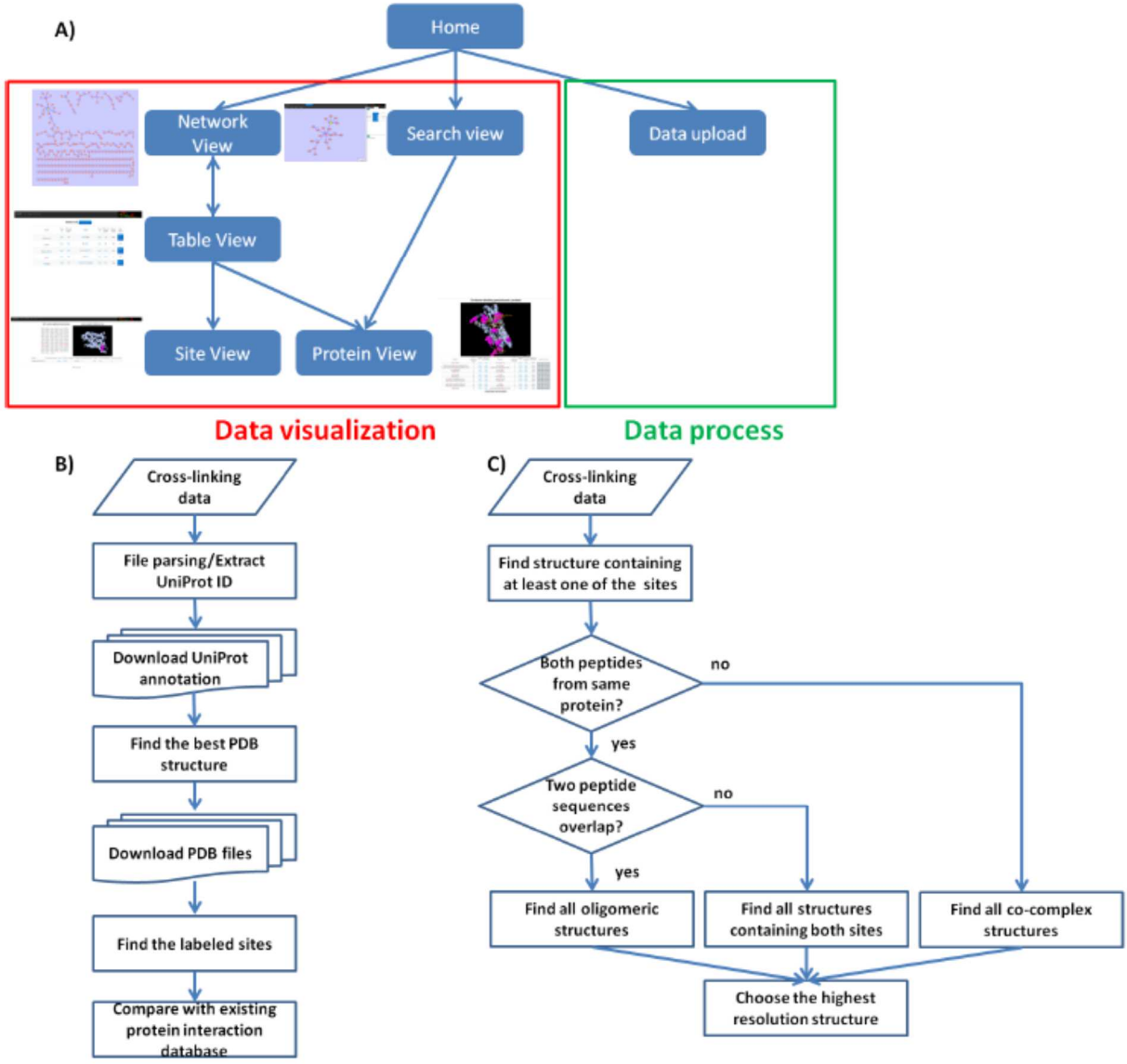

Figure 8 Internal structure and algorithms in XLink-DB

A) Web structure of CrossLink-DB; B) The data process scheme for uploaded data; C) The algorithms of choosing the best PDB structure

3.2.1 *Data upload, process and storage:*

The users can choose if they want their data to be public available. If they choose not to release their data to the public, they will get a table name after the data upload is finished. Their data will not appear in the drop-down list to choose. Instead, the users can use the table name to access

their data. Their data will be stored in the database for 90 days. If the user chooses to make their data public available, the data will be permanently stored in the database. The users can access their published and previously uploaded data from the drop-down list. Data are uploaded in XLink-DB in a tab-delimited file format with column arrangements as indicated on XLink-DB help page (http://brucelab.gs.washington.edu/crosslinkdbv1/help.php). XLink-DB parses the input file to extract the UniProt identifiers for each cross-linked protein contained within the dataset. The UniProt files (.txt files) containing protein annotation is then automatically downloaded from the UniProt database. The sequence information and identifiers for each labeled protein are parsed from the UniProt file and stored within the database in XLink-DB. If available, the PDB code associated with each protein is also retrieved from the UniProt annotation. For cases where more than one PDB code is associated with one protein, XLink-DB will select and retrieve the PDB structure based on the following rules: First XLink-DB will find all the PDB files which contain structural information covering the cross-linked site. If two cross-linked peptides originate from different protein sequences, which identifies a hetero interaction, all the co-crystal structures containing the two labeled proteins will be put in the candidate pool for later selection. Next, if the cross-linked peptide pair contains identical or overlapping peptide sequences that originate within a single protein sequence, all oligomer structure files containing both sites will be put in the candidate pool. If the cross-linked peptide pair does not fall into either of the two categories above, individual structure files containing both sites will be put into the candidate pool. Finally, the software will choose the structure with highest sequence coverage from the candidate pool to use for visualization of the cross-linked peptide pair. The structure with highest sequence coverage is chosen because they allow the best representation of the entire protein and greatest chance to

cover cross-linked sites. If no structural file can be found which contains both labeled sites, the software will choose the best individual structures for each labeled site.

After the PDB codes are assigned to each protein, the PDB files for these proteins are automatically downloaded. XLink-DB then computes atom numbers for all cross-linked peptide sites with the following steps: First, the peptide sequence is mapped to the protein sequence in the PDB file. Next, the atom numbers and coordinates of every copy of the cross-linked peptide in the PDB file are identified. The chosen atoms are the alpha carbon of the cross-linked lysine residues. The shortest distance between the two cross-linked sites contained in each cross-linked peptide pair is then calculated from the atomic coordinates. Finally, the associated atom numbers of the cross-linked sites are stored within the database embedded in XLink-DB.

The final data processing step is to compare the uploaded data with an existing protein interaction database. For this case we used the databases IntAct[101] and EciD[102]. We chose these two databases based on the coverage of protein interaction data. IntAct is used for human data. For *E.coli* data, EciD is used instead because it has a better coverage on the *E.coli* protein interaction data. The computed distances between two cross-linked proteins serve as measurement from the reference protein interaction network composed from existing protein interaction database information. For example, if two cross-linked proteins were previously known to interact, the computed distance within the reference protein interaction network is 0, otherwise the computed distance is the smallest number of nodes or proteins that exist in the reference network linking the two cross-linked proteins. If the cross-linked proteins cannot be connected in the reference network, “N/A” will be returned for this computed distance.

3.2.2 *Data visualization*

3.2.2.1 Network view:

In Network View, a protein interaction network of the cross-linked peptide dataset will be displayed on the left side of the page. Each node represents a protein and each edge represents all the cross-linked peptide pairs linking the two proteins. The users can open files, save files and change the layout and style options from the menu on the top. The toolbox at the right bottom corner of the network graph enables panning and zooming in the graph. Every node and edge in the graph can be selected, dragged and edited. The right-hand side of the page contains three tabs: Visual Style, Filter and Properties. The Visual Style tab allows users to change the color of the nodes, edges and background. The Filter tab allows users to filter the nodes based on the value of attributes. The properties tab is automatically activated when nodes or edges are selected. When one or more nodes are selected, the interacting partners of the selected nodes will be listed in a table. The name of each interacting partner is converted into a button which will lead to the Protein View of this protein complex. When one or more edges are selected, the interactions which are represented by the selected edges will be listed in a table. Each interaction is converted to a button which will lead to the Protein View of the pair. In addition, the protein interaction network developed with cross-linking data is compared with previous known protein structural and interaction information. For instance, the size of the node indicates a crystal structure for the protein exists in PDB. The thickness of the edges is related to the number of cross-linked peptide pairs that have been identified in the dataset, with thicker lines indicative of 2 or more cross-links. The color of the edge indicates the distance of connection of the two proteins in reference protein interaction database. Red edges indicate direct interactions between linked proteins are found in IntAct or EciD. Green edges indicate linked proteins have been found to share a common interactor

in the reference database and are therefore one node away. Black edges indicate linked proteins are more than one node away or were not found in the reference databases. It should be noted that, for intra-linkages that contain two peptides from the same protein, green edge color indicates that these proteins were not found to form homomultimers in the reference protein interaction databases.

3.2.2.2 Protein View:

Protein View page contains a structure viewer on the top if the structure is available, and a result table on the bottom. The user can change basic display options with right-click menu in the Jmol layer. Two buttons are available to change the display of cross-linked peptide pairs. "Display all" button illustrates all cross-linked sites associated with the two proteins displayed in the Jmol layer. "Reset complex" button will remove all the cross-linking pairs labeled on the structure. The bottom part of the page contains a result table with all the pairs associated with the two proteins. This table contains peptide sequence, gene name, PDB code, number of cross-linked pairs that involve the peptide and display option button. The number of cross-linking pairs involving the peptide is a measurement of reactivity and spatial proximity of the labeled site. A larger number indicates the labeled site is close to many other sites and the labeled site is highly reactive. The "display single pair" button will display the selected pair on the structure. The users can also use their own favorite structure if they do not appreciate the pre-assigned structures. They need to input the PDB code and the chain IDs for the respective proteins.

3.2.2.3 Table View:

The Table View page can be accessed from the Network View by clicking on the "Generate table view" button. The result table page contains two parts; the top part shows the link to the network view and the title. The bottom part is the result table with peptide sequence, protein

accession, PDB code, distance of connection and links to protein view. This table can be sorted by entries within each column by clicking on the column heading. Each entry in Peptide A/B columns is hyperlinked to the Site view page which will be described later. Protein names shown in columns Protein A/B within the table are hyperlinked to relevant UniProt pages for each protein to facilitate further investigation. Similarly, "PDB code for peptide A/B" names are hyperlinked to the relevant PDB page for additional structure information if needed. The "Show structure" button produces a protein-level view of the cross-linked pair.

3.2.2.4 Site view:

As mentioned above, the Site View shows the two labeled sites in two parallel windows. This enables users to visualize the location of the labeled peptide in the protein. When the crystal structure is available for the either protein but not the complex, the site will be highlighted magenta on the structure; otherwise the entire cross-linked peptide will be highlighted red in the protein sequence.

3.2.2.5 Search view:

Search view can be accessed from the home page. The user can choose UniProt ID, UniProt accession or gene name to search for any protein of interest. The user can either search one protein or give a list of protein IDs to search. The search will be performed against all the datasets for the selected organism.

3.3 RESULTS:

Two datasets are used to demonstrate the features of XLink-DB. One is a large scale cross-linking experiment performed in our laboratory on intact *E. coli* cells (See companion manuscript by Weisbrod et al.) "Weisbrod *et al.*" dataset is used here to denote this data from *E. coli* cells.

The other dataset was extracted from a recent publication by Yang *et al*. in which the researchers performed cross-linking on *E. coli* cell lysate.[95] "Yang *et. al.*" is used here to refer to this dataset. Both datasets comprise the largest reported cross-linking datasets and contain several hundred unique cross-linked sites. There are a few differences in the two experiments. Weisbrod *et al.* used customized cross-linker which is mass spectrometry cleavable and has biotin affinity tag for purification. Yang et. al. used commercially available DSS which is non-cleavable. Both dataset used strong cation exchange to enrich high charge peptides. Weisbrod *et.al.* performed avidin capture to enrich biotin-tagged peptides prior to mass spectrometry analysis. Using XLink-DB to analyze these datasets provides unique insight into datasets which would have been difficult and time consuming to get manually. **Figure 9** illustrates the distribution of cross-linked distances mapped by XLink-DB. These distances are extracted from XLink-DB and plotted in Excel. Both datasets show broad distributions of observed cross-linked distances. Disuccinimidyl suberate (DSS) a cross-linker with a relatively short spacer arm length (11.4Å) was applied in the "Yang *et. al.*" dataset. The cross-linker used in the "Weisbrod *et al.*" dataset has a spacer arm longer than 30Å, the fact that both datasets show similar cross-linked distance distributions suggests that cross-linker size is less important than protein flexibility in determination of which protein sites are cross-linked in complex mixtures.

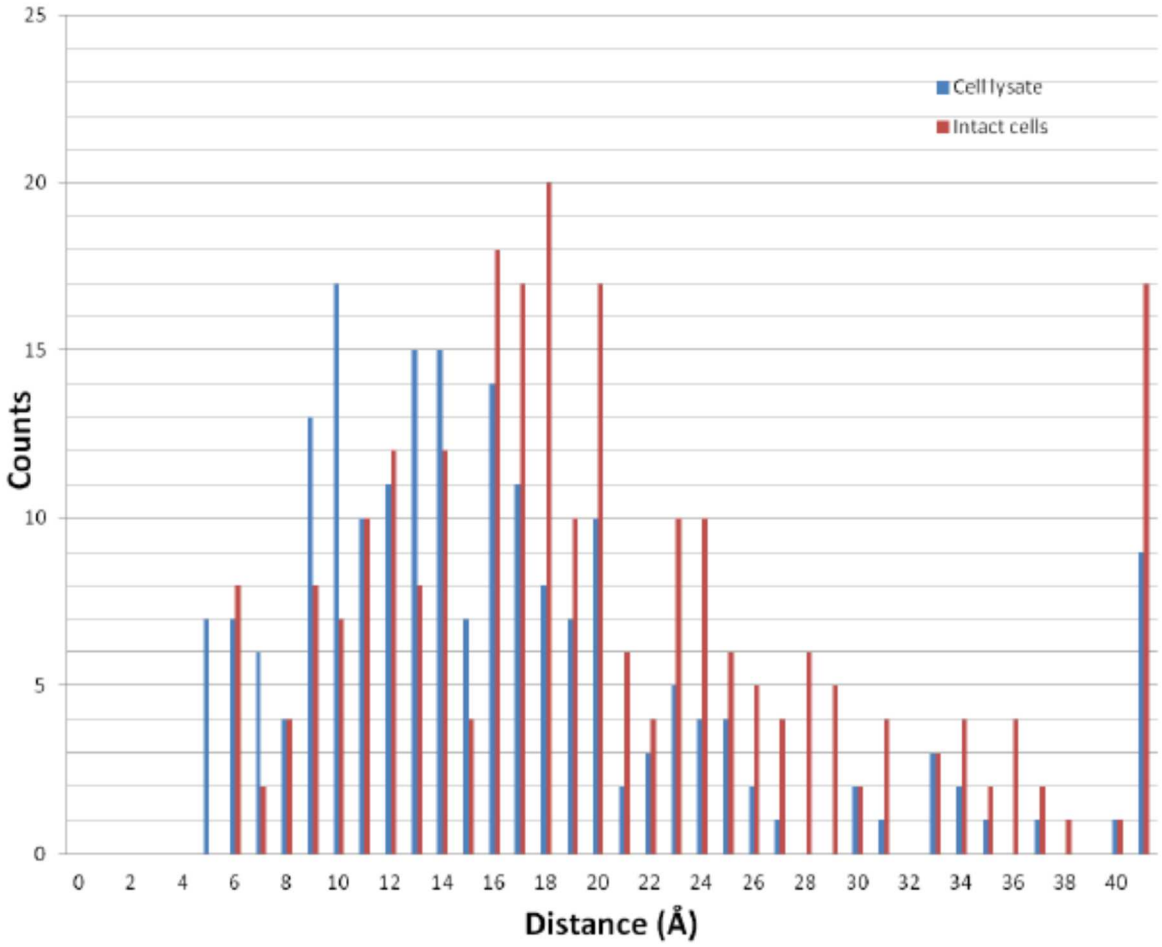

Figure 9 Distribution of interlinked distances of large-scale cross-linked peptide datasets from cells and cell lysates.

Distances are computed in XLink-DB from all cross-linked sites that appear within crystal structures available from the PDB. Cell lysate data (Yang et al, 2012) is shown in blue and intact cell data shown in red.

Using XLink-DB both datasets were compared to the *E.coli* protein interaction database (EciD, only considering interactions from experimentally derived data). Figure 3 shows the distribution of the node distances of both datasets and a Monte-Carlo simulation of the expected distance for randomly selecting two proteins. Both cross-linking datasets consist of approximately 130 inter-protein interactions. For the Monte-Carlo simulation, 130 randomly selected protein pairs were chosen to represent the sample size of the cross-linking experiment. The experiment was repeated

100 times and the average percentage of each distance is plotted in **Figure 10**. Based on the Monte-Carlo simulation, the most probable expected distance of two randomly chosen proteins is 2 nodes. The majority of the distances for the two cross-linking datasets is below or equal to one node, suggesting that both "Weisbrod *et al."* dataset and "Yang *et. al."* cross-linking experiments show good correlation with other experimental techniques. Furthermore, the "Weisbrod *et al.*" dataset contains the highest percentage (25%) of known direct interactors (0 nodes), whereas random simulation predicts about 4%. This suggests that data from either the "Weisbrod *et al."* or "Yang *et. al."* cross-linking experiments is significantly different from random data based on existing known interactions from EciD.

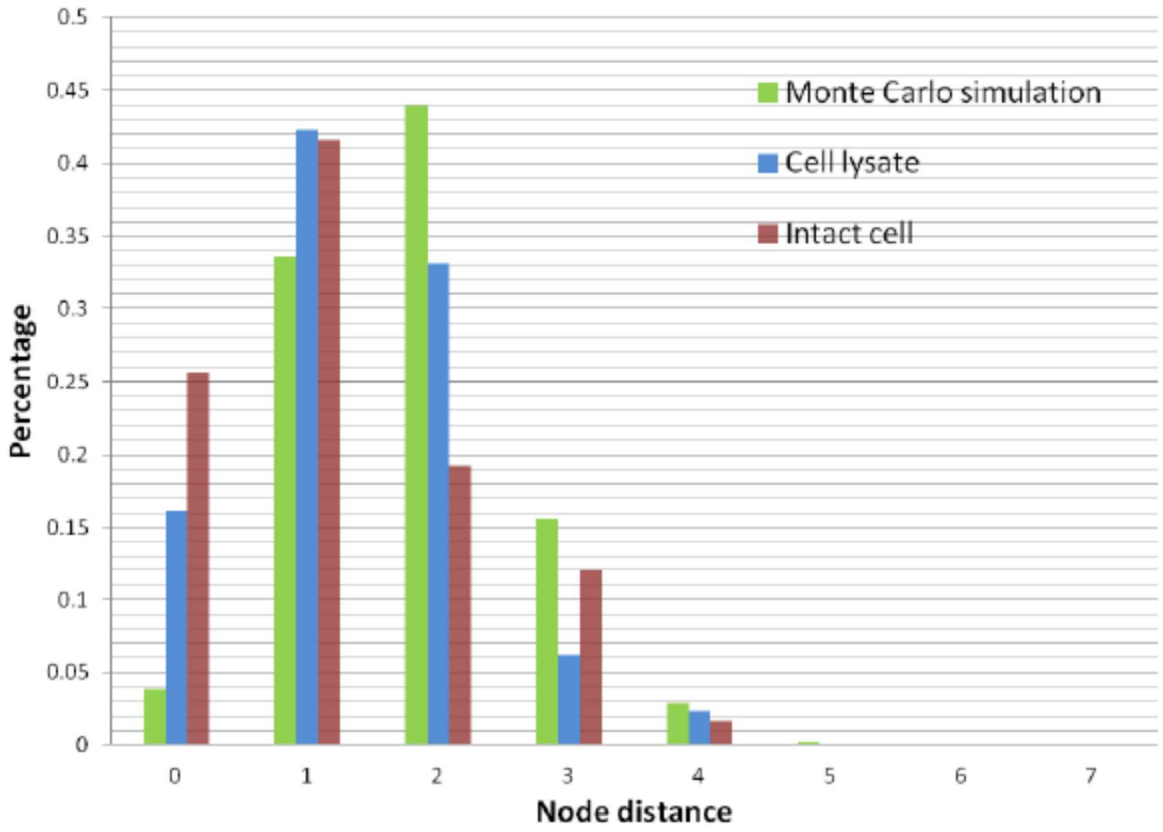

Figure 10 Distribution of the node distances

Distribution of the node distances observed in cross-linked peptide datasets from cell lysates (Yang et al., 2012) shown in blue and intact cells shown in red as determined from the E. coli protein

interaction database EciD. Also shown in green is the expected nodal distance distribution for random selection of 2 proteins shown in green.

3.4 DISCUSSION:

Several protein interaction databases have been established and embraced by the scientific community, such as PDB, EciD and IntAct. But none of them serve the role which XLink-DB will play. While PDB represents a significant resource in terms of available protein crystal structures and databases like EcID and IntAct contain significant wealth of data on protein interactions, there currently is a void of databases that contain protein interaction topological data. This likely stems from the lack of technological capabilities to produce data of this kind, but new technologies and advancements are rapidly changing the situation[7, 19, 90, 96, 103]. XLink-DB was developed to help fill this void in database availability and maximize the access and utility of protein interaction topological data that is now available and will come from these technological advancements.

In conclusion, XLink-DB presents a new way to organize and demonstrate protein interaction data with topological information. Conventional databases either lack the interaction information or lack the topological information for the protein complexes. With the advancement of new cross-linking technologies, large scale protein interaction studies are now becoming reality. XLink-DB is the first database to allow compilation and analysis of large-scale cross-linking data. It will not only help the cross-linking community to store, share and process their data, but also share the data with other scientists with interests in protein interactions and topologies.

Chapter 4. AUTOMATIC MODELING AND DOCKING PIPELINE FOR XLINK-DB

4.1 INTRODUCTION

With the fast growing amount of cross-linking data, an automatic modeling and docking pipeline which is specifically designed for cross-linking data is ever needed. Leading modeling/docking platforms such as Modeller, I-TASSER, Rossetta, PatchDock, ZDock either do not have a web interface for users or only accept singular input each time.[23, 104-106] Phyre2 web interface could accept hundreds of sequence as a single input file for modeling, but it requires special permission from the administrator.[107] For a single cross-linking experiment, thousands of cross-linking pairs could be observed. The number of modeling or docking experiments could be hundreds. Using these existing platforms manually could be cumbersome and prone to errors. There are many disadvantages for manually perform these computations. First, the parameters of the computation experiment could be hidden from the users once the input files are submitted. Therefore, there's no tracking record or limited access to the tracking records of those experiments. Manual input every time is easy to introduce operator error. Second, handling hundreds of input files manually is easy to keep track of finished jobs and ongoing jobs. Third, in addition to how many jobs users can submit to those platforms, they also usually have a limit on how many jobs a single user can keep in their job queue. Fourth, in order to use those existing platforms, the cross-linking data need to be pre-processed to match their required input file format. It is usually a complex job which requires automatic pipelines to handle, but these functions are not included in those platforms. Therefore, a new pipeline which will include automatic modeling, docking and pre-process for cross-linking data is needed for cross-linking community.

Luckily, XLink-DB has already included some of the pre-processing features which are required for modeling/docking experiments.[26] Features such as extracting protein sequence, existing PDB file, automatic aggregating cross-linking pairs for each protein complex or single protein are essential preparation steps prior to modeling/docking experiments. Therefore, XLink-DB is the perfect hosting platform for automatic modeling/docking pipelines. By including existing modeling/docking pipeline into XLink-DB. The data pre-processing, parameter control and job handling could be automatic handled by XLink-DB.

4.2 DOCKING EXPERIMENT WITH CROSS-LINKING DATA ON BENCHMARK DATASET

4.2.1 *Generation of benchmark dataset*

It is important to test the idea of using cross-linking data to improve docking experiment with a benchmark dataset before we apply this idea to large scale dataset. Since we need to test the docking results against something we know the answer, we decided to use protein complex structures which are deposited in PDB. Among the data in current XLink-DB, we have 122 binary protein interactions which already have PDB structure. These 122 binary protein interactions are from 4 different organisms, including 48 binary interactions from *E.coli*, 43 binary interaction from human, 29 binary interactions from yeast and 2 binary interactions from mouse. We first filtered the 122 PDB structures to generate a benchmark dataset for our test. We have two criteria for the filtering process. First, the two interacting proteins need to have a large enough interface. Because the rigid body docking experiment optimizes the contacting area or the interacting interface between the two proteins, a sufficient large interface is required for testing docking experiments. In the presented study, we have defined 20 amino acid at the interface to be the minimum area. An amino acid is considered to be on the interface when it is 10 Å or closer to

another amino acid in the other protein. Direct distance between the two alpha carbon atoms is used in this case. Second, the distances for each pair of cross-linked sites need to be within 35 Å. The distance between two cross-linked sites is used to score each model. Since 95% of the cross-linking distances we observed are within 35 Å, longer distances will mislead the scoring algorithms to yield a low score for the benchmark structure. After the filtering steps, the number of benchmark structures reduced to 84.

4.2.2 *Docking experiment for benchmark dataset*

Each binary protein complex in the benchmark dataset will be separated into two subunits, each set of coordinates is written into a separate file. The two proteins are then used in docking experiment with and without cross-linking data. PatchDock is used for the docking algorithms. [25] The docking results are clustered with maximum RMSD 4Å. The top ten models are used for testing similarity with the original complex structure. For the docking experiments with cross-linking data, the top 100 models are re-ranked with cross-linking distances. The top ten models after re-ranking are used for testing similarity with the original complex structure.

Figure 11 illustrates the success rate of the docking experiment. A successful docking experiment is defined as following, if one of the top 10 models is similar enough (below certain RMSD threshold) to the original structure, the docking experiment is considered as a successful one. Figure 11 shows the success rate of docking experiment by applying different RMSD thresholds. With 14Å RMSD threshold, 80% docking experiments are successful even without using cross-linking data. This is likely resulted from the fact that the docking algorithms we used was trained with PDB structures. Furthermore, since the two proteins are separated from one complex, they already have good complementary shapes, which will benefit the docking experiment. With cross-linking data, the success rate of docking experiments were further

improved to 90% with 14Å RMSD threshold. The result shows rigid body docking achieves high accuracy docking results and cross-linking data improve docking accuracy. Therefore, this approach could be used for large scale docking experiments.

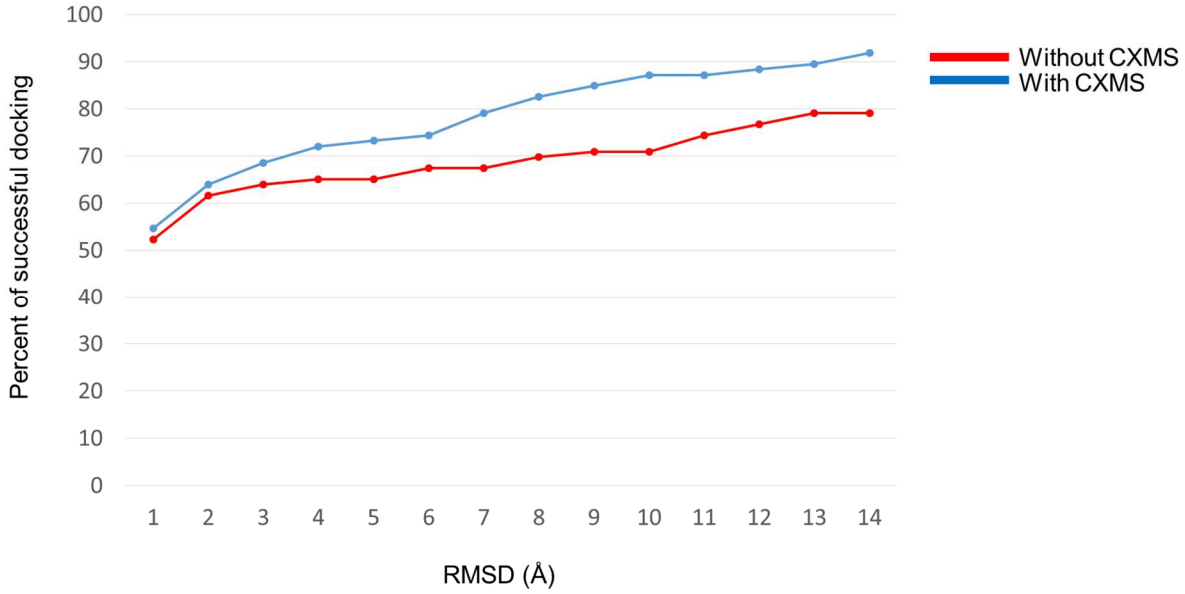

Figure 11 Success rate of docking experiments

4.3 LARGE SCALE MODELING AND DOCKING EXPERIMENT WITH EXISTING DATA IN XLINK-DB

Currently, the large scale cross-linking dataset in XLink-DB contains 6719 distance constraints measured from 6 different organism, including *E.coli*, *Pseudomonas Aeruginosa*, *Acinetobacter Baumannii*, yeast, mouse and human[7, 21, 22, 26, 108]. These 6719 distance constraints are mapped onto 1199 binary protein interactions. Figure 12 shows the breakdown of the entire dataset in organisms. Human dataset is the largest dataset, which contributes 33% of the data. Yeast dataset contributes 20% of the entire dataset, becoming the 2nd largest dataset. *E.coli* data shares

a similar size with yeast dataset with 18%. Mouse, *Acinetobacter* and *Pseudomonas* datasets are 11%, 14% and 4% respectively.

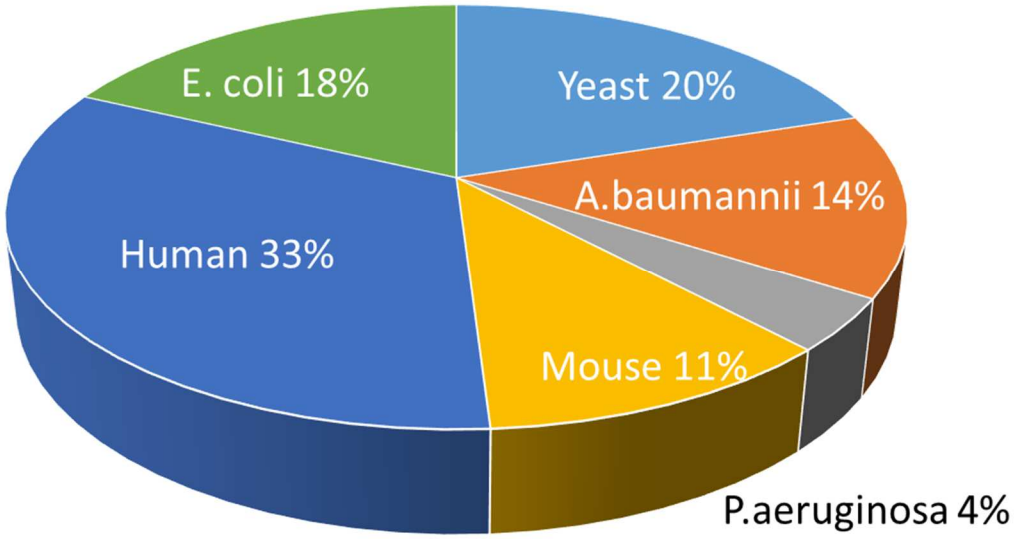

Figure 12. Large scale cross-linking dataset in XLink-DB

Many proteins that are included in the large scale dataset do not have existing PDB structure. We used Phyre2 online server to generate homology models for the proteins missing monomer structures[107]. The best model for each protein was chosen to use in the docking experiments. In total, we generated 740 homology models. Integrative Modeling Platform (IMP) was used for the docking experiments[24]. Since the results of docking experiments are determined by the initial coordinates of the protein, the two proteins are docked with both forward and reverse orders. The top 100 models of forward and reverse docking for each complex are mixed and re-ranked with cross-linking distances measured on the model structure. 1199 docking experiments were performed to predict the complex structures, the top models were deposited in XLink-DB (http://brucelab.gs.washington.edu/xlinkdb/dataRetriever.php?tablename=master_docking_phyre&dataset=&privateFlag=1).

Although no existing complex protein structures are available in PDB for the 1199 protein interactions we used for large scale docking experiment, some protein have highly homologous protein which form similar complexes which has PDB structure. Human mitochondria ATPase alpha-beta interaction is a good example. Human mitochondria ATPase complex, also called as complex V in mitochondria electron transport chain, has two functional domain F_0 and F_1. Alpha subunit and beta subunit both locate at the F_1 domain, each has three copies per complex. Our docking experiment simulated the interaction between one alpha and one beta subunit. Although the structure of this interaction is unknown. The bovine mitochondria ATPase alpha-beta interaction has known structure in PDB database (2WSS). The sequence of bovine mitochondria ATPase alpha subunit and beta subunit both share over 90% sequence identity with human counterpart. As shown in Figure 13, we compared our docking structure with the PDB structure for bovine ATPase alpha-beta interaction. Although the starting structure of human ATPase subunits are different with bovine ATPase subunits, the complex structure we acquired from docking experiment is almost identical to the PDB structure for bovine mitochondria ATPase.

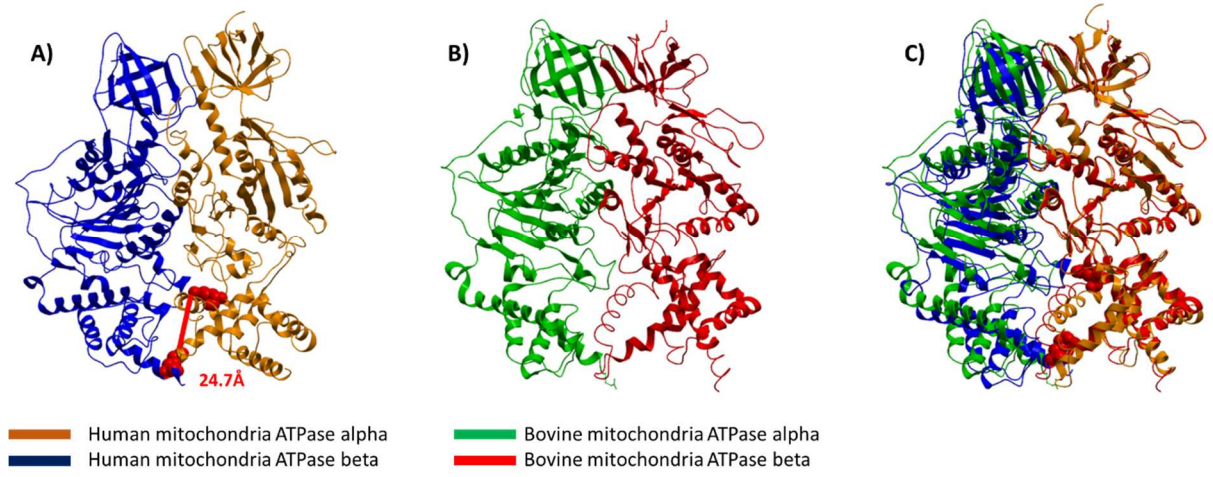

Figure 13 Comparison of the docking structure of human mitochondria ATPase alpha-beta subunits interaction with the crystal structure of bovine mitochondria ATPase alpha-beta subunits interaction

4.4 IMPLEMENTATION OF AUTOMATIC MODELING/DOCKING PIPELINE

This section describes the efforts of integrating the modeling/docking into XLink-DB to enable automatic computation and database maintenance. Integrative modeling platform (IMP) developed by Sali's group is chosen for modeling and docking.[24] IMP uses Modeller for protein homology modeling, and PatchDock for protein docking. IMP presents a few advantages. First, IMP is already an integrated modeling, docking pipeline. Compatibility issues between modeling software and docking software have already been resolved. Second, IMP is invoked via python script, providing flexibility for future addition of new features. Third, both Modeller and PatchDock are leading algorithms.[25] Fourth, the running time of IMP is relative low, therefore it is more compatible with large scale data.

For every protein which does not have existing PDB structure, XLink-DB will generate a structure model. The modeling job working cycle is like following: the sequence of the target protein is submitted to align with PDB sequence database. PDB sequences are ranked based on their homology with the target protein sequence. The best ranked PDB structure is then chosen for homology modeling, and the structure file is automatically downloaded from PDB database. A set of homology models are then created based on the template. The best scored one is then chosen to be stored in XLink-DB. When both proteins of a protein complex have either model or PDB structure, a docking job will be started by XLink-DB automatically. Integrative docking module in IMP is used for docking experiments. XLink-DB fetches all the cross-linking distance constraints related to this complex. The docking model candidates are scored with the build-in CXMS scoring algorithms in IMP. The top one model is accepted and stored in XLink-DB.

The modeling and docking experiments for each dataset could take hours to do. They should be treated differently than cross-linking data uploading job in XLink-DB, which usually only takes

a few minutes. The modeling and docking pipeline in XLink-DB is separated from normal data upload pipeline. XLink-DB uses a Jenkins server to schedule routine jobs such as checking for new data, starting modeling or docking jobs, check modeling job status, and update database to store structure models. By using Jenkins server, the modeling and docking pipeline is entirely separated from the normal cross-linking data upload pipeline. Therefore, the modeling and docking pipeline does not require users to wait for the modeling jobs to finish. Also, the modeling and docking pipeline can also be separately maintained, which will provide system flexibility for future development. All modeling and docking job status are stored with a relational database for querying and updating. In order to efficiently maintain all docking and modeling jobs, a centralized job queueing system is required. XLink-DB uses Sun Grid Engine (SGE) to manage job queues. SGE is commonly used for job schedule system on Linux clusters, but it is also an excellent job management system for single machine.

4.5 CONCLUSION

The automatic modeling/docking pipeline in XLink-DB is the first automatic pipeline which will take cross-linking results as input to produce high quality models. It provides an integrative data analysis platform for large scale cross-linking data, and enables researchers to visualize their results in model structures. Furthermore, the created model structures can be used as starting point for further model structure refining.

BIBLIOGRAPHY

1. Wang, Y., The function of OmpA in Escherichia coli. Biochem Biophys Res Commun, 2002. 292(2): p. 396-401.

2. Stenberg, F., et al., Protein complexes of the Escherichia coli cell envelope. J Biol Chem, 2005. 280(41): p. 34409-19.

3. Smith, S.G., et al., A molecular Swiss army knife: OmpA structure, function and expression. FEMS Microbiol Lett, 2007. 273(1): p. 1-11.

4. Zakharian, E. and R.N. Reusch, Kinetics of folding of Escherichia coli OmpA from narrow to large pore conformation in a planar bilayer. Biochemistry, 2005. 44(17): p. 6701-7.

5. Pautsch, A. and G.E. Schulz, Structure of the outer membrane protein A transmembrane domain. Nat Struct Biol, 1998. 5(11): p. 1013-7.

6. Pautsch, A., et al., Strategy for membrane protein crystallization exemplified with OmpA and OmpX. Proteins, 1999. 34(2): p. 167-72.

7. Zheng, C., et al., Cross-linking measurements of in vivo protein complex topologies. Mol Cell Proteomics, 2011. 10(10): p. M110 006841.

8. Marcoux, J., et al., Mass spectrometry defines the C-terminal dimerization domain and enables modeling of the structure of full-length OmpA. Structure, 2014. 22(5): p. 781-90.

9. Pauling, L., R.B. Corey, and H.R. Branson, The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A, 1951. 37(4): p. 205-11.

10. Dill, K.A. and J.L. MacCallum, The protein-folding problem, 50 years on. Science, 2012. 338(6110): p. 1042-6.

11. Fields, S. and O. Song, A novel genetic system to detect protein-protein interactions. Nature, 1989. 340(6230): p. 245-6.

12. Gingras, A.C., et al., Analysis of protein complexes using mass spectrometry. Nat Rev Mol Cell Biol, 2007. 8(8): p. 645-54.

13. Ho, Y., et al., Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature, 2002. 415(6868): p. 180-3.

14. Kay, L.E., NMR studies of protein structure and dynamics. J Magn Reson, 2005. 173(2): p. 193-207.

15. Smyth, M.S. and J.H. Martin, x ray crystallography. Mol Pathol, 2000. 53(1): p. 8-14.

16. Carpenter, E.P., et al., Overcoming the challenges of membrane protein crystallography. Curr Opin Struct Biol, 2008. 18(5): p. 581-6.

17. Sinz, A., Chemical cross-linking and FTICR mass spectrometry for protein structure characterization. Anal Bioanal Chem, 2005. 381(1): p. 44-7.

18. Tang, X. and J.E. Bruce, A new cross-linking strategy: protein interaction reporter (PIR) technology for protein-protein interaction studies. Mol Biosyst, 2010. 6(6): p. 939-47.

19. Tang, X., et al., Mass spectrometry identifiable cross-linking strategy for studying protein-protein interactions. Anal Chem, 2005. 77(1): p. 311-8.

20. Chavez, J.D., et al., Cross-linking measurements of the Potato leafroll virus reveal protein interaction topologies required for virion stability, aphid transmission, and virus-plant interactions. J Proteome Res, 2012. 11(5): p. 2968-81.

21. Chavez, J.D., et al., Quantitative proteomic and interaction network analysis of cisplatin resistance in HeLa cells. PLoS One, 2011. 6(5): p. e19892.

22. Navare, A.T., et al., Probing the protein interaction network of Pseudomonas aeruginosa cells by chemical cross-linking mass spectrometry. Structure, 2015. 23(4): p. 762-73.

23. Yang, J., et al., The I-TASSER Suite: protein structure and function prediction. Nat Methods, 2015. 12(1): p. 7-8.

24. Russel, D., et al., Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol, 2012. 10(1): p. e1001244.

25. Schneidman-Duhovny, D., et al., Geometry-based flexible and symmetric protein docking. Proteins, 2005. 60(2): p. 224-31.

26. Zheng, C., et al., XLink-DB: database and software tools for storing and visualizing protein interaction topology data. J Proteome Res, 2013. 12(4): p. 1989-95.

27. Anderson, N.G., Co-immunoprecipitation. Identification of interacting proteins. Methods in Molecular Biology, 1998. 88: p. 35-45.

28. Rigaut, G., et al., A generic protein purification method for protein complex characterization and proteome exploration. Nature Biotechnology, 1999. 17(10): p. 1030-2.

29. Walhout, A.J. and M. Vidal, High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. Methods, 2001. 24(3): p. 297-306.

30. Gould, K.L., et al., Tandem affinity purification and identification of protein complex components. Methods, 2004. 33(3): p. 239-44.

31. Su, C., et al., Bacteriome.org--an integrated protein interaction database for E. coli. Nucleic Acids Research, 2008. 36(Database issue): p. D632-6.

32. Sinz, A., Chemical cross-linking and FTICR mass spectrometry for protein structure characterization. Analytical and bioanalytical chemistry, 2005. 381(1): p. 44-7.

33. Zhang, H., et al., Identification of protein-protein interactions and topologies in living cells with chemical cross-linking and mass spectrometry. Molecular and Cellular Proteomics, 2009. 8(3): p. 409-20.

34. Leitner, A., et al., Probing Native Protein Structures by Chemical Cross-linking, Mass Spectrometry, and Bioinformatics. Molecular and Cellular Proteomics, 2010. 9(8): p. 1634-49.

35. Tang, X., et al., Mass spectrometry identifiable cross-linking strategy for studying protein-protein interactions. Analytical chemistry, 2005. 77(1): p. 311-8.

36. Shi, L., et al., Direct involvement of type II secretion system in extracellular translocation of Shewanella oneidensis outer membrane cytochromes MtrC and OmcA. Journal of Bacteriology, 2008. 190(15): p. 5512-6.

37. Yang, L., et al., A photocleavable and mass spectrometry identifiable cross-linker for protein interaction studies. Analytical chemistry, 2010. 82(9): p. 3556-66.

38. Anderson, G.A., et al., Informatics strategies for large-scale novel cross-linking analysis. J Proteome Res, 2007. 6(9): p. 3412-21.

39. Hoopmann, M.R., C.R. Weisbrod, and J.E. Bruce, Improved Strategies for Rapid Identification of Chemically Cross-linked Peptides. Journal of Proteome Research, 2010. Accepted.

40. Hoopmann, M.R., G.L. Finney, and M.J. MacCoss, High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. Anal Chem, 2007. 79(15): p. 5620-32.

41. Schneidman-Duhovny, D., et al., PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic acids research, 2005. 33(Web Server issue): p. W363-7.

42. Ritchie, D.W., D. Kozakov, and S. Vajda, Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational FFT generating functions. Bioinformatics, 2008. 24(17): p. 1865-73.

43. Tang, X. and J.E. Bruce, A new cross-linking strategy: protein interaction reporter (PIR) technology for protein-protein interaction studies. Molecular bioSystems, 2010. 6(6): p. 939-47.

44. Anderson, G.A., et al., Informatics strategies for large-scale novel cross-linking analysis. Journal of Proteome Research, 2007. 6(9): p. 3412-21.

45. Perkins, D.N., et al., Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 1999. 20(18): p. 3551-67.

46. Rinner, O., et al., Identification of cross-linked peptides from large sequence databases. Nature methods, 2008. 5(4): p. 315-8.

47. Berman, H.M., et al., The Protein Data Bank. Acta Crystallographica, Section D: Biological Crystallography, 2002. 58(Pt 6 No 1): p. 899-907.

48. Kuriyan, J. and D. Eisenberg, The origin of protein interactions and allostery in colocalization. Nature, 2007. 450(7172): p. 983-90.

49. Korndorfer, I.P., M.K. Dommel, and A. Skerra, Structure of the periplasmic chaperone Skp suggests functional similarity with cytosolic chaperones despite differing architecture. Nature structural & molecular biology, 2004. 11(10): p. 1015-20.

50. Walton, T.A., et al., 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(6): p. 1772-7.

51. De Cock, H., et al., Affinity of the periplasmic chaperone Skp of Escherichia coli for phospholipids, lipopolysaccharides and non-native outer membrane proteins. Role of Skp in the biogenesis of outer membrane protein. European journal of biochemistry 1999. 259(1-2): p. 96-103.

52. Kleinschmidt, J.H., Membrane protein folding on the example of outer membrane protein A of Escherichia coli. Cellular and Molecular Life Sciences, 2003. 60(8): p. 1547-58.

53. Walton, T.A. and M.C. Sousa, Crystal structure of Skp, a prefoldin-like chaperone that protects soluble and membrane proteins from aggregation. Molecular cell, 2004. 15(3): p. 367-74.

54. Dunten, P., H. Jaffe, and R.R. Aksamit, Crystallization of 5-keto-4-deoxyuronate isomerase from Escherichia coli. Acta Crystallographica, Section D: Biological Crystallography, 1998. 54(Pt 4): p. 678-80.

55. Stenberg, F., et al., Protein complexes of the Escherichia coli cell envelope. Journal of Biological chemistry, 2005. 280(41): p. 34409-19.

56. Smith, S.G., et al., A molecular Swiss army knife: OmpA structure, function and expression. FEMS microbiology letters, 2007. 273(1): p. 1-11.

57. Schmid, F.X., et al., Prolyl isomerases: role in protein folding. Advances in protein chemistry, 1993. 44: p. 25-66.

58. Arie, J.P., N. Sassoon, and J.M. Betton, Chaperone function of FkpA, a heat shock prolyl isomerase, in the periplasm of Escherichia coli. Molecular microbiology, 2001. 39(1): p. 199-210.

59. Saul, F.A., et al., Structural and functional studies of FkpA from Escherichia coli, a cis/trans peptidyl-prolyl isomerase with chaperone activity. Journal of Molecular Biology, 2004. 335(2): p. 595-608.

60. Stoller, G., et al., A ribosome-associated peptidyl-prolyl cis/trans isomerase identified as the trigger factor. EMBO journal, 1995. 14(20): p. 4939-48.

61. Saul, F.A., et al., Structural and functional studies of FkpA from Escherichia coli, a cis/trans peptidyl-prolyl isomerase with chaperone activity. J Mol Biol, 2004. 335(2): p. 595-608.

62. Stoller, G., et al., A ribosome-associated peptidyl-prolyl cis/trans isomerase identified as the trigger factor. EMBO J, 1995. 14(20): p. 4939-48.

63. Ma, Q. and T.K. Wood, OmpA influences Escherichia coli biofilm formation by repressing cellulose production through the CpxRA two-component system. Environ Microbiol, 2009. 11(10): p. 2735-46.

64. Morona, R., C. Kramer, and U. Henning, Bacteriophage receptor area of outer membrane protein OmpA of Escherichia coli K-12. Journal of bacteriology, 1985. 164(2): p. 539-43.

65. Wang, Y., The function of OmpA in Escherichia coli. Biochemical and biophysical research communications, 2002. 292(2): p. 396-401.

66. Pautsch, A. and G.E. Schulz, Structure of the outer membrane protein A transmembrane domain. Nature structural biology, 1998. 5(11): p. 1013-7.

67. Grizot, S. and S.K. Buchanan, Structure of the OmpA-like domain of RmpM from Neisseria meningitidis. Molecular microbiology, 2004. 51(4): p. 1027-37.

68. Debnath, D.K. and D.E. Otzen, Cell-free synthesis and folding of transmembrane OmpA reveals higher order structures and premature truncations. Biophysical chemistry, 2010.

69. Stathopoulos, C., An alternative topological model for Escherichia coli OmpA. Protein science, 1996. 5(1): p. 170-3.

70. Sugawara, E., et al., Secondary structure of the outer membrane proteins OmpA of Escherichia coli and OprF of Pseudomonas aeruginosa. Journal of bacteriology, 1996. 178(20): p. 6067-9.

71. Paliy, O., et al., Protein disorder is positively correlated with gene expression in Escherichia coli. Journal of proteome research, 2008. 7(6): p. 2234-45.

72. Li, X., et al., Predicting Protein Disorder for N-, C-, and Internal Regions. Genome informatics. Workshop on Genome Informatics, 1999. 10: p. 30-40.

73. Romero, P., et al., Sequence complexity of disordered protein. Proteins, 2001. 42(1): p. 38-48.

74. Meszaros, B., I. Simon, and Z. Dosztanyi, Prediction of protein binding regions in disordered proteins. PLoS Computational Biology, 2009. 5(5): p. e1000376.

75. Arora, A., et al., Refolded outer membrane protein A of Escherichia coli forms ion channels with two conductance states in planar lipid bilayers. Journal of Biological chemistry, 2000. 275(3): p. 1594-600.

76. Stathopoulos, C., An alternative topological model for Escherichia coli OmpA. Protein Sci, 1996. 5(1): p. 170-3.

77. Sugawara, E., et al., Secondary structure of the outer membrane proteins OmpA of Escherichia coli and OprF of Pseudomonas aeruginosa. J Bacteriol, 1996. 178(20): p. 6067-9.

78. Gingras, A.C., et al., Analysis of protein complexes using mass spectrometry. Nature reviews. Molecular cell biology, 2007. 8(8): p. 645-54.

79. Muller, D.R., et al., Isotope-tagged cross-linking reagents. A new tool in mass spectrometric protein interaction analysis. Analytical chemistry, 2001. 73(9): p. 1927-34.

80. Rappsilber, J., et al., A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. Analytical chemistry, 2000. 72(2): p. 267-75.

81. Huang, B.X., H.Y. Kim, and C. Dass, 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(8): p. 1237-47.

82. Back, J.W., et al., Chemical cross-linking and mass spectrometry for protein structural modeling. Journal of molecular biology, 2003. 331(2): p. 303-13.

83. Young, M.M., et al., High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America, 2000. 97(11): p. 5802-6.

84. Chen, T., J.D. Jaffe, and G.M. Church, Algorithms for identifying protein cross-links via tandem mass spectrometry. J Comput Biol, 2001. 8(6): p. 571-83.

85. Chu, F., et al., Isotope-coded and affinity-tagged cross-linking (ICATXL): an efficient strategy to probe protein interaction surfaces. J Am Chem Soc, 2006. 128(32): p. 10362-3.

86. Gomes, A.F. and F.C. Gozzo, Chemical cross-linking with a diazirine photoactivatable cross-linker investigated by MALDI- and ESI-MS/MS. J Mass Spectrom, 2010. 45(8): p. 892-9.

87. Kalkhof, S., et al., Chemical cross-linking and high-performance Fourier transform ion cyclotron resonance mass spectrometry for protein interaction analysis: application to a calmodulin/target peptide complex. Anal Chem, 2005. 77(2): p. 495-503.

88. Muller, M.Q., et al., A universal matrix-assisted laser desorption/ionization cleavable cross-linker for protein structure analysis. Rapid Commun Mass Spectrom, 2011. 25(1): p. 155-61.

89. Petrotchenko, E.V., et al., BiPS, a photocleavable, isotopically coded, fluorescent cross-linker for structural proteomics. Mol Cell Proteomics, 2009. 8(2): p. 273-86.

90. Rinner, O., et al., Identification of cross-linked peptides from large sequence databases. Nat Methods, 2008. 5(4): p. 315-8.

91. Silva, R.A., et al., A three-dimensional molecular model of lipid-free apolipoprotein A-I determined by cross-linking/mass spectrometry and sequence threading. Biochemistry, 2005. 44(8): p. 2759-69.

92. Sinz, A. and K. Wang, Mapping spatial proximities of sulfhydryl groups in proteins using a fluorogenic cross-linker and mass spectrometry. Anal Biochem, 2004. 331(1): p. 27-32.

93. Yang, T., et al., Mapping cross-linking sites in modified proteins with mass spectrometry: an application to cross-linked hemoglobins. Anal Biochem, 1996. 242(1): p. 55-63.

94. Walzthoeni, T., et al., False discovery rate estimation for cross-linked peptides identified by mass spectrometry. Nat Methods, 2012.

95. Yang, L., et al., A photocleavable and mass spectrometry identifiable cross-linker for protein interaction studies. Anal Chem, 2010. 82(9): p. 3556-66.

96. Zhang, H., et al., Identification of protein-protein interactions and topologies in living cells with chemical cross-linking and mass spectrometry. Mol Cell Proteomics, 2009. 8(3): p. 409-20.

97. Apweiler, R., et al., UniProt: the Universal Protein knowledgebase. Nucleic Acids Res, 2004. 32(Database issue): p. D115-9.

98. Bernstein, F.C., et al., The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol, 1977. 112(3): p. 535-42.

99. Willinghagen, E.L., Processing CML Conventions in Java. Internet Journal of Chemistry, 2001. 4.

100. Lopes, C.T., et al., Cytoscape Web: an interactive web-based network browser. Bioinformatics, 2010. 26(18): p. 2347-8.

101. Kerrien, S., et al., The IntAct molecular interaction database in 2012. Nucleic acids research. 40(Database issue): p. D841-6.

102. Andres Leon, E., et al., EcID. A database for the inference of functional interactions in E. coli. Nucleic Acids Res, 2009. 37(Database issue): p. D629-35.

103. Yang, B., et al., Identification of cross-linked peptides from complex samples. Nat Methods, 2012.

104. Eswar, N., et al., Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics, 2006. Chapter 5: p. Unit 5 6.

105. Pierce, B.G., Y. Hourai, and Z. Weng, Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS One, 2011. 6(9): p. e24657.

106. Simons, K.T., et al., Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins, 1999. Suppl 3: p. 171-6.

107. Kelley, L.A., et al., The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc, 2015. 10(6): p. 845-58.

108. Weisbrod, C.R., et al., In vivo protein interaction network identified with a novel real-time cross-linked peptide identification strategy. J Proteome Res, 2013. 12(4): p. 1569-79.

VITA

Chunxiang Zheng was born in Tianjin, China. He completed his bachelor degree in Fudan University in Shanghai, China. In 2007, he came to US, starting graduate school in Washington State University, Pullman, WA. He got a Master of Science degree in WSU, and transferred to University of Washington, Seattle, WA in 2010.