Using Chemical Crosslinking and Mass Spectrometry for Protein Model Validation and Fold Recognition

by

Esther W. M. Mak

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Biology

Waterloo, Ontario, Canada, 2006

Author's Declaration for Electronic Submission of a Thesis

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The 3D structures of proteins may provide important clues to their functions and roles in complex biological pathways. Traditional methods such as X-ray crystallography and NMR are not feasible for all proteins, while theoretical models are typically not validated by experimental data. This project investigates the use of chemical crosslinkers as an experimental means of validating these models. Five target proteins were successfully purified from yeast whole cell extract: Transketolase (TKL1), inorganic pyrophosphatase (IPP1), amidotransferase/cyclase HIS7, phosphoglycerate kinase (PGK1) and enolase (ENO1). These TAP-tagged target proteins from yeast *Saccharomyces cerevisiae* allowed the protein to be isolated in two affinity purification steps. Subsequent structural analysis used the homobifunctional chemical crosslinker BS^3 to join pairs of lysine residues on the surface of the purified protein via a flexible spacer arm. Mass spectrometry (MS) analysis of the crosslinked protein generated a set of mass values for crosslinked and non-crosslinked peptides, which was used to identify surface lysine residues in close proximity. The Automatic Spectrum Assignment Program was used to assign sequence information to the crosslinked peptides. This data provided inter-residue distance constraints that can be used to validate or refute theoretical protein structure models generated by structure prediction software such as SWISS-MODEL and RAPTOR. This approach was able to validate the structure models for four of the target proteins, TKL1, IPP1, HIS7 and ENO1. It also successfully selected the correct models for TKL1 and IPP1 from a protein model library and provided weak support for the HIS7, PGK1 and ENO1 models.

Acknowledgements

I would like to thank my supervisor, Dr. Brendan McConkey, for giving me this wonderful opportunity to study with him these past three years. He has given me valuable insights and guidance in my research as well as patience and support in everything. I would also like to Andrea Spires for all her technical input and guidance. She has also been wonderful friend who kept me company through all the good and the bad times in my research as well as my personal life. Ted Quon and Andrew Doxey have been tremendously helpful in getting me through the computational aspects of my project, despite having their own busy work schedule. Furthermore, I would like to thank everyone in my lab; they have made the lab fun to work in and they are the reason why I enjoy coming into school everyday. Last but not least, I would like to thank the members of the Duncker lab for all their assistance in TAP-tagging and their generosity in allowing me to use their lab resources.

I would like to thank my parents for their unconditional love and support. They have never asked anything of me except to be happy. They had made difficult choices and sacrifices in the past to ensure that I will have the best opportunities and the happiest life they feel I deserve and I will be forever grateful for everything they have done for me.

Finally, I would like to thank my fiancé, Ian Washburn, for all his love and support. The past few years have been difficult and he had stuck by me through everything, the good and the bad, the mood swings, the unreasonable demands and then some more. He had never given up hope in my abilities to finish this degree. His constant reassurance has helped me believe that I can do anything I set my mind to. This degree is as much of an accomplishment for him as it is for me.

Dedication

I would like to dedicate this thesis to my grandmother, Luk Mui Mak, who passed away in 1995. She was the smartest, sweetest and the most loving person I have ever known. She taught me that there is no language barrier as long as you keep smiling. She showed me that intelligence is not dependent on the amount of education you have received. Most importantly, she taught me to always see the goodness in people and to treat everyone with kindness and respect. Even though she cannot be here today to share my happiness and success, I know she is watching me from heaven and she is proud of the kind of person I have grown up to be.

Table of Contents

List of Tables

List of Figures

1 Overview

1.1 Introduction

Thousands of proteins participate in biological pathways that intertwine to form complex networks. The functions of individual proteins in these pathways are largely dependent on their three-dimensional structures. Therefore, protein structure analyses may provide important clues to their functions and their roles in biological pathways. To examine the structures, the target proteins must be isolated from thousands of proteins existing in the cell.

According to the statistics release in April 2006, there are a total of 215,741 protein sequence entries in the SWISS-PROT database (Swiss-Prot 2005) but only 36,121 of these sequences have corresponding protein structures in the Protein Data Bank (PDB) (O'Donovan 2002; Westbrook et al. 1997). NMR spectroscopy and X-ray crystallography are the traditional methods for studying protein structures. These techniques contributed to approximately 15% and 85% of the solved structure in the PDB (Westbrook et al. 1997), respectively. However, these techniques are time consuming and are not feasible for all proteins. NMR is limited by the size of the protein molecules while X-ray crystallography is limited by the availability of the protein crystal. These reasons have created a bottleneck in the rate of solving protein structures. Thus there is a need for developing more high-throughput methods for studying protein structures (Bourne and Weissig 2003).

While new experimental techniques for structure analysis are being developed, advances in mathematics and computing technology have improved the efficiency and accuracy of structure prediction algorithms. If the appropriate template structures are available, protein structures can be predicted by comparative modeling. Comparative modeling predicts a protein structure through the use of a template. For this method to be successful, a certain level of sequence identity between the target and template is required. The higher the sequence identity, the better the model will be. Fold recognition is also an important method of structural prediction. In general, the structure of a protein is dictated by the composition of its amino acid sequence and sequences that have high sequence identity to each other will most likely have the same fold. But it is also possible for two unrelated proteins with low sequence identity to have similar folds, which can be predicted by protein threading methods (Bourne and Weissig 2003). The protein sequence is compared to a structural template library and each sequence-structure fit is evaluated by a scoring function which takes into account the sequence homology between the sequence and the template, the secondary

structures as well as factors such as solvent accessibility and residue interactions. The fit with the highest score is considered to be the optimal fold for the sequence (Rost et al. 1997). Finally, ab initio prediction is the most difficult method because it attempts to predict a structure using the amino acid sequence alone. Using energy scoring functions, an ab initio method will attempt to find a state of the protein structure where its free energy is at a global minimum (Bernasconi and Segre 2000). This method can be used when homology modeling fails due to the lack of a good template and it is particularly useful for predicting a protein with a novel fold (Bourne and Weissig 2003). In comparison, fold recognition and ab initio prediction are inferior to homology modeling and they produce less reliable structures (Bourne and Weissig 2003).

1.2 Overview of research methodology

The main purpose of this research project is to investigate experimental methods that can quickly produce data to support protein models generated by computational means. Several research groups have explored the possibility of using chemical crosslinkers in protein structure analysis. In essence, chemical crosslinkers join together specific amino acid residues on the surface of the protein and introduce distance constraints on the structure, thus limiting the number of folds available to the protein. This method has been shown to be successful for structural validation as well as fold recognition (Kruppa et al. 2002; Muller at al. 2001; Pearson et al. 2002; Young et al. 2000).

The following methodology has been proposed for protein structure analysis. First, the protein of interest is purified under native conditions. The purified protein is then incubated with chemical crosslinker to allow the conjugation of specific surface residues. This chemical crosslinking will result in a set of crosslinked peptides, in addition to the normal set of peptides broken down from a non-crosslinked protein. The crosslinked peptides can be identified by mass spectrometry and the positions of the crosslinked residues can be determined. This provides a set of distance constraints between specific residues, which can be compared with the distances between residues in model structures. This approach is validated by selecting proteins with known structures and the experimental crosslinking data is tested against the crosslinked residues predicted from the protein structure. Once the method is validated, the possibility of using crosslinking data for model validation and fold recognition will be investigated.

1.3 Protein purification

The first step in developing experimental procedures for studying protein structures is the extraction of native target proteins from thousands of proteins in the cell. Many methods can be used to separate the protein of interest from crude cell extracts, but affinity chromatography provides an efficient means of isolating proteins using fusion tags. This approach involves the binding of the target protein to a small molecule immobilized to a solid support in a chromatography column. A fusion tag is added to the target protein to facilitate this binding event; Hexahistidine (His), Glutathione S-Transferase (GST) and Tandem Affinity Purification (TAP) fusion tags are only a few examples. Each fusion tag will bind to different small molecules in the column and their binding mechanisms also vary with each tag. A His-tag is a histidine-rich amino acid sequence (Hochuli et al. 1987), which will chelate to nickel or cobalt metals while GST-tags bind to the reduced glutathione in the column (Smith and Johnson 1988). A TAP-tag contains two different binding domains; one is antibody specific and the other is peptide specific (Rigaut et al. 1999; Puig et al. 2001).

Tandem Affinity Purification is a generic method for protein purification developed by Rigaut et al. (1999). It was originally developed as a complementary technique to yeast-two-hybrid for protein-protein interaction studies (Ito et al. 2001; Uetz et al. 2000). However, TAP purified proteins can also be used for structural and functional analysis since the purification method can be carried out under native conditions. Furthermore, TAP-tagged proteins are expressed at physiological level and low abundance proteins can also be purified efficiently (Rigaut et al. 1999).

A TAP-tag is fused to target protein at the N-terminal domain. This tag consists of two IgG binding domains (BD), a TEV cleavage site and a calmodulin binding domain (CaM BD) (Figure 1-1a). This two-step purification protocol is illustrated in Figure 1-1b. The TAP-tagged protein is first separated from the other proteins via the binding of IgG BD in the IgG bead column. IgG BD recognizes the protein A which is conjugated to agarose beads. A protease from the tobacco etch virus (TEV) is used to remove the target protein from the IgG beads by cleaving at the TEV cleavage site. The eluted protein enters the second step of purification which involves the binding of CaM BD with Ca^{2+} which enables the CaM BD to bind to CaM beads (Rigaut et al. 1999; Puig et al. 2001). Ca^{2+} is removed by the addition of a chelating agent and CaM BD dissociates from the CaM beads, thus releasing the protein from the column. The TAP protein will retain the small CaM BD at the end of the purification process.

a)

Calmodulin Binding Domain — TEV Cleavage Site

SMEKRRWKKNFIAVSAANRFKKISSSGALDAYDIPTTASENLYFQGELKTAALAQHDEA

Two IgG Binding Domains

VDNKFNKEQQNAFYEILHLPNLNEEQRNAFIQSLKDDPSQSANLLAEAKKLNDAQAPK

VDNKFNKEQQNAFYEILHLPNLNEEQRNAFIQSLKDDPSQSANLLAEAKKLNGAQAPK

VDANSAGKST

b)

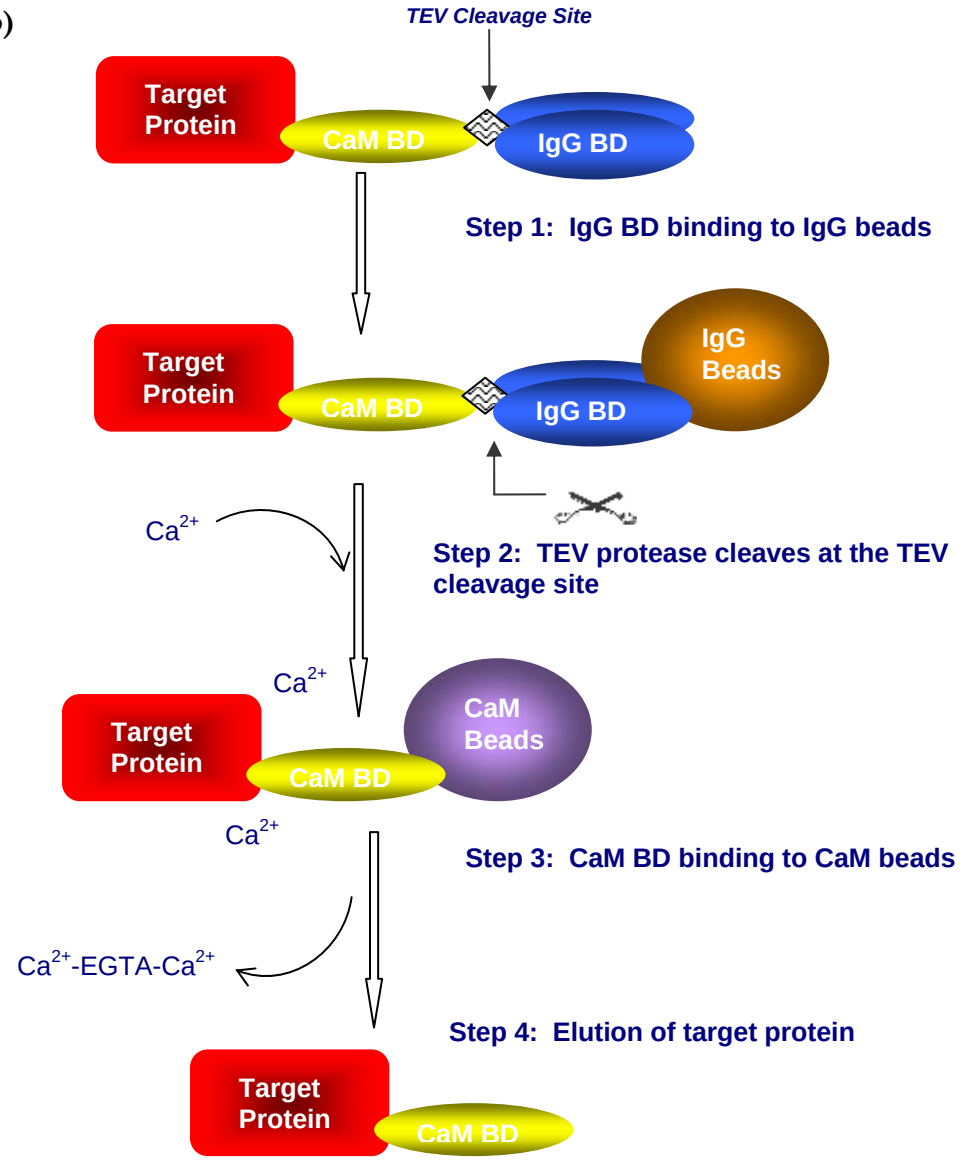


Figure 1-1. Illustration of tandem affinity purification. a) Components of TAP-tag: calmodulin binding domain, TEV cleavage site and two IgG binding domains (modified from Rigaut et al. 1999). b) Details of the two-step purification process.

1.3.1 Protein analysis by mass spectrometry

Isolated native proteins can be analyzed by Mass Spectrometry (MS). MS can measure the mass of peptide fragments from the target protein with great accuracy, speed and sensitivity. Each protein sample must be broken down into smaller pieces for mass analysis because smaller protein fragments generate more accurate measurements. Fragments of 6-20 amino acids are desired for sequence matching and trypsin is the most popular choice for proteomic analysis (Liebler 2002).

Each MS instrument is composed of three parts: the source where ions are produced from the sample, the analyzer where the ions are resolved based on their mass/charge (m/z) ratio, and the detector which detects the resolved ions. The mass data generated are analyzed by computer software (Liebler 2002). For this project, ESI-nanospray Q-TOF was used to analyze the target proteins. For the ESI (electrospray ionization) nanospray process, the ESI samples are introduced into the mass spectrometer in solution. With a controlled pH environment, peptides can exist in ionized form. Nanoelectrospray can reduce the sample flow rate when entering the mass spectrometer. The ESI sample passes through a high-voltage needle and forms mist droplets that contain a mixture of peptide and solvent ions. A desolvation process removes the solvent ions from the mixture and sending only the peptide ions into the mass analyzer (Liebler 2002). Quadrupole Time-of-flight (Q-TOF) is the mass analyzer that measures the time required for the ions to travel through the analyzer and reach the detector. The m/z ratio of the ion is directly proportional to its flight time (Liebler 2002), i.e. low mass ions will reach the detector faster than high mass ions. The resulting m/z ratios are recorded and displayed on a mass spectrum.

1.3.2 Protein identification

There are two methods for protein identification using MS: peptide mass fingerprinting (PMF) and tandem MS (or MS/MS). PMF is the conventional method for obtained information from a single stage MS analysis. This involves taking the measured masses of the protein fragments and matching them to the theoretical peptide masses from protein sequence databases (Liebler 2002). Approximately 50-90% of proteins can be identified successfully from organisms with fully sequenced genomes (Mann et al. 2001). There are several drawbacks to PMF. Firstly, the protein sequence of interest must be in the database. Secondly, data from MS analysis will include a margin of error. Thirdly, it is possible that a match of the sample mass and the database mass is nothing more than a simple coincidence; for example, peptide fragments with the same amino acids but in different order will generate the same mass (Liebler 2002; Mann et al. 2001).

For protein identification, MS/MS is the preferred method. The Q-TOF used in this project consists of a quadrupole MS and a TOF MS, connected together by a collision cell. Doubly charged peptides are preferred because they produce better MS/MS results (Mann et al. 2001). A peptide species of interest can be isolated from the peptide mixture based on its mass in the first mass spectrometer. These peptides enter the collision cell where they collide with a highly pressured gas and fragment into a series of b- and y-ions. The b-ion fragments contain the N-terminus ends of the peptides while the y-ion fragments contain the C-terminus end. Fragments without a charge are not detected by the spectrometer (Liebler 2002). These ions are then separated in a second mass spectrometer stage and analyzed by the TOF mass analyzer (MS-Labor 2005). Given data of sufficient quality, the sequence of the peptide can often be deduced from the resulting MS/MS spectrum.

To confirm the identity of the purified protein using database searching, MS data can be submitted to the web-based search engine, Mascot (Perkins et al. 1999). Both PMF and MS/MS Ion Search are available on Mascot. This tool requires the user to input the MS data (either for PMF or MS/MS), the search database, the taxonomic classification of the study organism, information about the protease used for fragmentation, protein molecular weight and chemical modifications as well as error tolerance. For MS/MS, the proper selection of peptides for MS/MS analysis can improve the quality of Mascot results. High intensity peaks are preferable because MS/MS spectra often have high background noise (Perkins et al. 1999). Experimental mass values for the ion fragments were compared to the calculated values based on the sequence. Mascot identifies the best matches using a scoring system that takes into account the probability of random matches between the experimental and calculated values as well as the size of the search database.

1.4 Chemical crosslinking

Various research groups have used chemical crosslinking reagents to study protein structures and complexes. Used in conjunction with mass spectrometry, these reagents introduce distance constraints by linking proximal residues on the surface of the protein which can yield low resolution structure information. This is particularly useful for model validation or fold recognition since the generated distance constraints can greatly limit the number of folds.

Young et al. (2000) was one of the first groups to use this approach for fold recognition. They succeeded in selecting the correct model (template with IL-1β) for fibroblast growth factor FGF-2 from a set of 20 models, determining that FGF-2 belongs to the β-trefoil fold family. They

have shown that this is a powerful method for fold recognition since the sequence identity between FGF-2 and IL-1β is <13%. Also, without the distance constraints from the chemical crosslinks, FGF-2 would have been placed in the wrong fold family (β-clip fold family) based on the rankings given by the threading algorithm used to generate the models.

In the Young method, crosslinked peptides were isolated from the peptide mixture by size exclusion chromatography (SEC) separation (Young et al. 2000). This presents a clear disadvantage because the end product of the SEC may contain the crosslinked peptides as well as a mixture of unmodified peptides, peptides with dangling crosslinkers or multiple crosslinked peptides (Kruppa et al. 2002). Since then, several other groups have improved this experiment by modify the chemical crosslinkers and detection methods. Isotopically labeled crosslinkers can facilitate the identification of crosslinked peptide peaks from the mass spectrum. Muller et al. (2001) used a deuterium-labeled crosslinker (d_0/d_4 isotopes) to investigate the structures of the microtubule-destabilizing protein Op18/stathmin. The resulting MS spectra contained only singlet and doublet peaks where the doublets represented the crosslinked peptides and allowed the easy detection of the crosslinked peptides from the rest of the peptide mixture. In 2002, Pearson et al. modified the labeling technique by using a larger isotopic mass difference of 8 Da (d_0/d_8). This enabled the detection of doublet peaks for ions up to a 4+ charge. The major advantage of these isotopically labeled crosslinker is that the crosslinked peptides are easily distinguished from the non-crosslinked peptides from the singlet-doublet peak pattern. However, the problem of the dangling crosslinkers and multiply crosslinked peptides remained indistinguishable from singly crosslinked peptides.

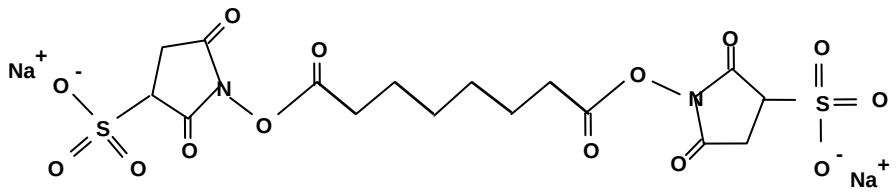
Kruppa et al. (2003) developed a top-down approach in which the crosslinked protein sample can be analyzed directly by Fourier transform mass spectrometry (FT-MS). In this case, the protein is broken up by gas-phase fragmentation, replacing the traditional proteolytic digestion in MS sample preparation. Then, the crosslinked peptides are identified by the gas-phase purification process in the FT-MS analyzer cell based exact mass measurements. This top-down approach greatly enhances the overall sensitivity of the crosslinked peptides. The high resolution and mass accuracy of the FT-MS enables the unambiguous identification of the intramolecular, singly crosslinked peptides.

1.4.1 Properties of chemical crosslinkers

A crosslink reagent typically consists of a flexible spacer arm with a reactive group at each end. There are four types of crosslink designs: homobifunctional, heterobifunctional, zero-length and trifunctional crosslinks. Homobifunctional crosslinks, such as the one used in this project,

contain two identical reactive groups, connected by a carbon chain with various lengths (Figure 1-2), whereas heterobifunctional crosslinks contain two different reactive groups. The major disadvantage of these crosslinks is their susceptibility to form a multitude of poorly defined, aggregated products. Protein aggregation occurs when the crosslink reagent reacts with a protein, forming an intermediate product. This intermediate then reacts with another protein or with a neighbouring functional group within the same protein. To overcome this problem, multiple-step protocols are developed to minimize protein aggregation by eliminating excess crosslink reagents after the initial reaction occurred between a crosslink and the protein (Sinz 2003). The zero-length and trifunctional crosslinks are not as commonly used as the homo- and heterobifunctional crosslinks. Zero-length crosslinks, as the name implies, are compounds that link two amino acid residues together without a flexible spacer arm. Trifunctional crosslinks are a relatively new member in the family of chemical crosslinks. In essence, they are heterobifunctional crosslink reagents with an additional reactive group specific for a third protein. The trifunctional crosslink can be used in affinity purification when the third reactive group is a biotin moiety (Sinz 2003; Trester-Zedlitz et al. 2003).

The most common reactive groups target the amino groups in the proteins via acylation or alkylation reactions, producing stable amide or secondary amine bonds (Sinz 2003). Due to the high abundance of lysines on the surfaces of proteins, N-hydroxysuccinimide (NHS) esters are the most widely used and the acylation reaction occurs at the primary amines of lysine and N-termini of the protein, forming amide bonds. This amine reaction is highly sensitive to its environment and therefore, the selection of the reaction buffer is very important. The pH of the reaction buffer must be close to physiological pH (7.0-7.5); the half-life of BS^3 decreases as the pH of the buffer diverges from the physiological pH (Pierce Biotechnology 2005). Furthermore, the reaction buffer must not contain any primary amines because those primary amines will hydrolyze the NHS-esters on the functional groups, thus reducing the amount of crosslinker available to the proteins. Buffers such as HEPES or phosphate buffers contain only tertiary amines and therefore are suitable for crosslinking reactions with BS^3 (Sinz 2003). Other factors affecting the reaction between crosslink and protein are salt concentration, temperature, hydrophobicity, number of reactive sites on the protein surface as well as the length of the spacer arm (Haniu et al. 1993). For these reasons, the concentration of crosslinkers with respect to proteins must be adapted for each individual application.



BS^3

M.W. 572.43
Spacer Arm 11.4 Å

Figure 1-2. Diagram of the chemical crosslinker BS^3.

1.4.2 Intramolecular crosslinking for protein structure analysis

For structure analysis of individual proteins, intramolecular crosslinks can be used to introduce distance constraints within the three-dimensional structure. Either homo- or heterobifunctional crosslinks can be used. Optimal crosslinking conditions (ie. ratio of protein and crosslink reagents) can be determined using 1D-PAGE and mass spectrometry. Protein concentration should be in the micromolar range to discourage intermolecular crosslinking between proteins, which can result in protein aggregation. Also, excess crosslinking on individual proteins may cause distortion to the tertiary structure (Young et al. 2000) but insufficient crosslinking may not produce enough crosslinked products for MS detection (Sinz 2003).

Crosslinked proteins can be isolated using 1D-PAGE or size exclusion chromatography. The proteins are then subjected to enzymatic digestion in solution. This is favorable over in-gel digestion because it is less time consuming, has more efficient proteolysis and most importantly, higher sample recovery (Back et al. 2002). The choice crosslinker used in this project is bis(sulfosuccinimidyl)-suberate ester (BS^3) which is a homobifunctional crosslinking reagent (Figure 1-2). It has two identical reactive groups that will react specifically with lysine residues, joining them together via a 11.4 Å spacer arm (Pierce Biotechnology 2002; Young et al. 2000). Because trypsin typically cleaves adjacent lysine residues, the addition of BS^3 on the protein will prevent cleavage at that position; thus the crosslinked peptides are composed of multiple peptides (Figure 1-3). The proteolysis of the crosslinked protein produces a mixture of crosslinked and non-crosslinked peptides and this mixture can be analyzed by liquid chromatography/ESI MS or MALDI-TOF MS. The crosslinked peptides generate additional mass values in the peptide mass spectra which can be

identified by comparing the mass spectra with a non-crosslinked control sample. Although the mass of these peptides are 3-4 times higher than the normal peptide fragments, they will still fall within the detection range of the mass spectrometer because these peptides will be multiply charged. Once the crosslinked peptides are identified, the corresponding distance constraints between linked residues can provide structural information at the level of protein fold (Rappsilber et al. 2000; Sinz 2003; Young et al. 2000).

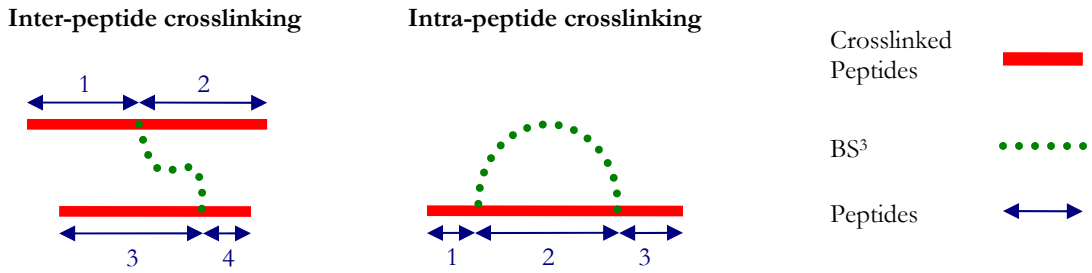


Figure 1-3. Two possible crosslinking scenarios. Inter-peptide crosslinking joined together four peptides while intra-peptide crosslinking joins together three peptides.

1.5 Computational analysis of crosslinking data

The high sensitivity of a MS instrument can allow the detection of low abundance crosslinked peptides, present only in the crosslinked protein spectrum. The Automatic Spectrum Assignment Program (ASAP) was used to obtain useful sequence information from the masses of the crosslinked peptides (Young et al. 2000). ASAP was developed by Young et al. (2000) at the University of California and is available to the public on the World Wide Web (http://roswell.ca.sandia.gov/~mmyoung/asap.html). ASAP constructs a virtual peptide library using the input protein sequence and user-specified parameters. Then, it searches for the experimental masses in the virtual library within a given error limit. Finally, it returns a list of plausible crosslinked peptide assignments and the distance between the crosslinked residues can be calculated using the 3D coordinates from a protein structure or model. Since the chemical crosslinkers introduce distance constraints between pairs of residues, a list of all possible crosslinked residues falling within those distance constraints can be determined.

In the works of Young et al. (2000) and Schilling et al. (2003), it was estimated that the maximum straight C^{α}-C^{α} distance between two residues crosslinked by BS^3 is 24 Å, the length of the spacer arm (11.4 Å) plus the C^{α} to the terminal N^{ζ} distances (2 x 6.2 Å) for each reactive group.

Also, the span of the crosslinker is approximately seven times the distance between C^{α} carbons adjacent in sequence. Therefore, it was reasoned that for a pair of crosslinked lysines from the experimental data to be a true positive, the distance must be less than 24 Å and greater than 7 amino acids apart in sequence (Schilling et al. 2003; Young et al. 2000). Other studies have also used the same method for estimating the crosslinking distances with other chemical crosslinkers (Dihazi and Sinz 2003; Kruppa et al. 2003; Pearson et al. 2002). However, Green et al. (2001) reported that the maximum distances are highly improbable. They argued that most crosslinkers are not in a fully extended conformation and the crosslinking distances are variable. In fact, Ye et al. (2004) determined that there is a 5 Å ambiguous region where the crosslinking distance with BS^3 is possible but less likely to occur. They also suggested that reducing ambiguous crosslinked sites would increase the confidence level of the protein model validation. Therefore, a shorter distance of 19 Å was recommended for the straight distance constraint rather than 24 Å to minimize ambiguity.

1.6 Structural model validation

A protein structure is required to evaluate the quality of the crosslinking data. In the event where the protein structure is not available, software tools are available to generate models based on sequence information. One structural modeling technique is homology modeling, also known as comparative modeling, and it generally involves the following steps. It starts by selecting appropriate templates for the initial sequence-template alignment. The success of the model will increase as the sequence similarity between the target and template sequence increases. The next step is to adjust the alignment to find the best fit. Once the optimal alignment is determined, a model can be generated by superimposing the backbone of the target sequence onto the template according to the alignment. Loops and side-chains are then added to the model. When the model has been assembled with all the components, it may be necessarily to optimize the model by adjusting the backbone again to accommodate the loops and side-chains (Bourne and Weissig 2003). The quality of this model is dependent on the sequence similarity between the target protein and the available template; high sequence similarity generally produces good protein models.

One example of homology modeling software is SWISS-MODEL, which is the most widely used protein modeling tool available to the public. The homology modeling server searches for a template in its template library which is extracted from the Protein Data Bank (Berman et al. 2000). Homology modeling by SWISS-MODEL requires at least one template with greater than 25% similarity between the template and the target sequence. SWISS-MODEL can produce highly

accurate models if the sequence identity is greater than 95% between the target and structure template (Schwede et al. 2003).

However, sometimes what the homology modeling software considers to be the best model (ie. one with the highest score) may not necessarily be the correct model (Xu et al. 2003). Therefore, it would be advantageous to obtain experimental data to validate or refute these computational models. Chemical crosslinkers conjugating surface protein residues can generate distance constraints between the residues, thus greatly limiting the number of theoretical folds. Young et al. (2000) had demonstrated the success of this approach. Using the distance information acquired from crosslinking experiments, they were able to select the correct protein model for FGF-2, even though the model was not identified as the best by the modeling software.

1.7 Protein threading

Protein threading is one of the most promising homology modeling techniques for protein structure prediction and fold recognition. This technique attempts to find the best match between the target sequence and a series of structural templates and use this template to predict the sequence's structure. This technique generally involves the following steps. First, a template database is constructed with proteins with low pairwise sequence similarity to minimize computational time. Energy scoring functions are developed to assess the quality of the predicted structures. Then, the target sequence is aligned with each template, optimizing the scoring function. Finally, a structure is predicted by placing the target sequence onto the backbone of the template of the most probable sequence-template alignment. This method can be used for protein structure prediction as well as fold recognition (Xu et al. 2003).

Protein threading has been proven to be a NP-hard problem when the length of variable gaps and pairwise amino acid interactions are considered simultaneously for the scoring function; this means that no polynomial time algorithm exists for finding the optimal solution to the threading problem (Lathrop 1994). Xu and Li (2003) have developed a linear programming approach to formulate this threading problem and find the optimal alignment between the target and template sequence while considering both variable gaps and pairwise interactions. This method was incorporated into a homology modeling software called RApid Protein Threading by Operation Research technique (RAPTOR).

RAPTOR was developed by Xu and colleagues at the University of Waterloo and it is one of the most accurate 3D structure prediction software tools available. It incorporates both secondary

structure and homology information in its prediction. For any given target sequence, RAPTOR uses PSI-BLAST (position-specific iterated BLAST) to construct a sequence profile from a multiple sequence alignment with sequence homologues including the target. This profile is a position specific scoring matrix that summarizes the frequency of each amino acid in each position of the target sequence and high scores are given to highly conserved positions (Altschul et al. 1997; Cates 2005).

For protein threading, RAPTOR aligns this sequence profile to a template and attempts to find the best fit between the sequence profile and the template by optimizing the scoring function. This scoring function is optimized based on sequence homology, types of secondary structures (helix, sheets and loops), solvent accessibility (buried, intermediate and accessible) and pairwise interactions. One of the options for threading methods available within RAPTOR is NP-Core. The templates are treated as a series of cores, where each core represents a conserved region of a secondary structure (α-helix or β-sheet), and the cores are joined together by loops. Only interactions between the core residues were considered during the alignment between the target and template sequences. Loop regions are generally considered insignificant with regards to fold recognition and they often require manual refinement in the predicted models (Xu and Li 2003). Additionally, RAPTOR employs a support vector machine (SVM) to select the best templates for the target sequence. Final z-scores are computed to standardize the scores from SVM and allow unbiased ranking of the threading results (Xu et al. 2003).

The combination of linear programming and SVM allows RAPTOR to perform structural predictions both optimally and efficiently. RAPTOR was ranked first among all non-meta servers (servers that perform structure prediction by combining outputs from other servers) in the Critical Assessment of Fully Automated Structure Prediction (CAFASP3) competition which evaluates the performance of structure prediction servers without expert interventions. RAPTOR had performed very well in both homology modeling and fold recognition (Xu et al. 2003).

1.8 Project summary

The purpose of this research is to investigate the possibility of using chemical crosslinking and MS data to provide experimental validation of proposed structural models from comparative modeling and fold recognition algorithms. Proteins were selected based on the availability of a solved structure. Proteins were purified by TAP and subject to chemical crosslinking experiments with crosslinker, BS^3. MS was used to collect data for the crosslinking peptides and the data was

submitted to ASAP for sequence assignment. This crosslinking data set was then used for protein structure validation for the five target proteins. Finally, the possibility of using this crosslinking data for fold recognition was explored.

2 Protein target selection and purification

2.1 Introduction

One of the first tasks in this project is to select the appropriate protein targets and purified these proteins for method validation and structural analysis. Tandem affinity purification (TAP) enables protein isolation to be performed under mild, non-denaturing conditions. It also allows a normal expression of the protein to ensure the state of the protein is as close to its physiological condition as possible. Other purification methods often require the overexpression of proteins that can lead to protein aggregation (Rigaut et al. 1999). A downside to TAP is that the quantity of the purified protein is dependent on its natural abundance in the cell. Therefore, it was important to investigate the efficiency of each step and to optimize the purification process.

The TAP-tag fused to the target protein allows the proteins to be purified via two simple steps. Purification is first carried out by percolating the protein samples through an IgG bead column, followed by a calmodulin bead column (Rigaut et al. 1999). Only TAP-tagged proteins remain in the sample buffer at the end of the purification process. Finally, tandem MS can be used to confirm the identity of the purified samples.

In 2003, Ghaemmaghami et al. created a TAP fusion library where each ORF in yeast was fused to a TAP-tag. They showed that the TAP-tag did not interfere with normal protein function. They analyzed Clb2 and Sic1, two cell-cycle-regulated proteins, and found that their regulation was unaffected by the presence of the TAP-tag. They also found that the TAP-tag was degraded along with its fusion protein. However, although 83% of the tagged proteins appeared to retain their normal functions, smaller proteins were more difficult to tag successfully, implying that the TAP-tag may interfere with their functions. Moreover, a small group of proteins require the C-termini to be conserved for subcellular localization. Therefore, the TAP-tag, which was fused to the C-terminus of the protein, could disturb this localization process (Ghaemmaghami et al. 2003; Hud et al. 2003).

A TAP fusion library is now commercially available from OpenBioSystems (http://www.openbiosystems.com/) for *Saccharomyces cerevisiae* and thus making this the choice organism for this project. Protein targets for this study were selected based on three criteria. Firstly, they must be high abundance protein in the yeast cells, which will ensure sufficient quantities for the development of experimental protocols. OpenBioSystems provides a spreadsheet that categorizes each TAP-fusion strains by relative abundance level; only strains which contained relatively highly

abundant TAP-tagged proteins were considered. Secondly, the target protein must either be in monomeric or homodimeric form to reduce the complexity of the crosslinking and MS data. Finally, each target protein should have either a known crystal structure or sufficient sequence similarity to another protein with a solved structure so that a reliable model can be built. Each highly abundant protein was investigated to ensure latter two conditions were met.

2.2 Materials and methods

2.2.1 Yeast culture and sample preparation

All yeast strains containing TAP-tagged target protein were purchased from OpenBioSystems. Yeast strains were grown in 1L of liquid YPD media (1% Bacto Yeast Extract, 2% Bacto Peptone, 2% Glucose) at 30ºC to near saturation. Cells were centrifuged at 8000 *g* for 15 min and the cell pellet was washed twice with ultrapure water followed by one wash with NP-40 buffer (15 mM Na_2HPO_4, 10 mM NaH_2PO_4-H_20, 1.0% NP-40, 150 mM NaCl, 2 mM EDTA, 50 mM NaF, 0.1 mM Na_3VO_4). Cell suspension was poured into a 50 mL bead beater chamber along with 200 µL of protease inhibitor mixture (Sigma) and 25 mL of 0.5 mm zirconia/silica beads (BioSpec). The chamber was then filled up to the rim with NP-40 buffer and the bead beating unit was assembled with 2 layers of ice slurry. The cells were lysed using 1 minute ON/ 1 minute OFF cycle, repeated 10 times, with a 5 minute OFF period half-way through. The cell lysate was then transfer to two 50 mL conical tubes and centrifuged at 10,000 *g* for 20 minutes. The supernatants were combined and protein concentration was measured.

Protein concentration was measured based on the Bradford method (Bradford 1976). A 4-point standard curve was constructed using freshly prepared BSA (Sigma), ranged from 0.5 – 10.0 mg/ml. For each standard and sample, 2 µL of protein was mixed with 200 µL of Bio-Rad protein assay reagent and 798 µL of ultrapure water. The blank contained 200 µL of protein assay reagent and 800 µL of ultrapure water. The absorbance was measured at 595 nm and protein concentration of the yeast cell lysate was extrapolated using the BSA standard curve.

2.2.2 Tandem affinity purification

The TAP protocol from the Yeast Resource Center (2000) was followed with minor modifications. To minimize protein degradation, all steps were performed on ice or in a 4ºC walk-in incubator for overnight steps. Because all agarose beads used for protein purification were preserved in buffers containing 20% ethanol, they were washed 3 times using their respective buffers and centrifuged at 4000 *g* (Gingras 2003). All plastic columns, conical and microfuge tubes were washed twice with 70% ethanol and 3 times with ultrapure water to remove residue contaminants that may interfere with MS analysis.

The cell lysate were divided into two 50 mL conical tubes. A Sepharose 6B bead (Sigma) slurry (1:1) was prepared with NP-40 buffer and 500 µL was added to each fraction of the cell lysate.

The lysates were incubated with the Sepharose 6B beads for 1 hour on a rocker and then poured into a Poly-prep chromatography column (Bio-Rad). The eluates were drained into 2 new 50 mL conical tubes.

The concentration of NaCl in the lysate was adjusted to 300 mM and 250 µL of IgG Sepharose 6 Fast Flow (GE Healthcare) bead slurry (1:1) in NP-40 buffer was added to each lysate fraction. The lysates were allowed to incubate with the IgG beads for 2-4 hours on a platform rocker. Then, the lysate/IgG bead mixture was poured into new chromatography columns (Bio-Rad); 2 mL of eluate was collected for analysis while the rest was discarded. The IgG beads were washed twice with 10 mL IPP300 (25 mM Tris-HCl, pH 8.0, 300 mM NaCl, 0.1% NP-40), once with 10 mL IPP150 (25 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.1% NP-40) and once with 10 mL TEV cleavage buffer (CB) (25 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.1% NP-40, 0.5 mM EDTA, 1.0 mM DTT). The bottom of the column was sealed and 1 mL of TEV CB containing 5 µL AcTEV protease (Invitrogen) was added to the IgG beads. The beads were incubated with the TEV protease, rotating overnight at 4ºC.

The eluate from the IgG column was collected in 2 clean 15 mL conical tubes and 30 µL of the eluate was saved for analysis. The IgG beads were washed once with 1 mL of TEV CB and combine with the first IgG eluate. A 300 mL calmodulin (CaM) bead slurry (1:1) was prepared with 0.1% CaM binding buffer (CBB) (25 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1 mM Mg acetate, 1 mM Imidazole, 2mM $CaCl_2$, 10 mM β-mercaptoethanol, 0.1% NP-40) and 150 µL was added to each eluate. Also, 6 mL of 0.1% CBB and 6 µL of 1M $CaCl_2$ were added to the lysate/CaM beads mixture. The protein was incubated for 2-4 hours with the Calmodulin Sepharose 4B (GE Healthcare) beads and then combined into a single chromatography column. The beads were washed once with 1 mL of 0.1% CBB, once with 0.02% CBB (25 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1 mM Mg acetate, 1 mM Imidazole, 2mM $CaCl_2$, 10 mM β-mercaptoethanol, 0.02% NP-40) and finally twice with CBB with no NP-40 (25 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1 mM Mg acetate, 1 mM Imidazole, 2mM $CaCl_2$, 10 mM β-mercaptoethanol). To dissociate the proteins from the beads, the column was sealed and 1 mL of CaM elution buffer (CEB) (25 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1 mM Mg acetate, 1 mM Imidazole, 20 mM EGTA, 10 mM β-mercaptoethanol) was added to the beads. The column was vortexed gently to ensure proper contact between CEB and beads. The eluate from the column was drained into a siliconized microfuge tube. A second mL of CEB was used to wash the calmodulin beads again and drained into another siliconized microfuge

tube. The eluates were then combined and the protein concentration was determined using the Bradford protein assay.

2.2.3 Protein concentration and buffer exchange

Prior to analysis, the purified proteins must be concentrated and resuspended in appropriate buffers. Three methods were explored for concentrating the protein and buffer exchange.

Dialysis of the TAP eluate was done by transferring 2 mL of the final TAP eluate to dialysis tubing (Fisherbrand). The protein sample was dialyzed with 1L of crosslinking buffer diluted 10-fold, and gently stirring overnight at 4°C. The protein sample was removed from the dialysis tubing and placed in a new siliconized microfuge tube. The volume of the sample was reduced to 200 µL in a speed vacuum. Protein concentration was measured using the Bradford protein assay.

The Nanosep devices (Pall Corporation) allowed protein concentration and buffer exchange to be done simultaneously, following the instructions from the user manual. The Nanosep membrane was washed twice with 500 µL of crosslinking buffer (100 mM HEPES, pH 7.0, 1M NaCl, 1mM EDTA, pH 8.0, 20 mM Sodium Phosphate, pH 7.5) to remove any residual contaminants that may interfere with subsequent protein analysis. The crosslinking buffer was centrifuged at 13000 *g* until the buffer had passed through the membrane. 500 µL of the final eluate from TAP was added to the Nanosep and centrifuged at 1500 *g* at 4°C until all buffer had passed through the membrane. This step was repeated until all of the TAP eluate was processed. The membrane was washed twice with 200 µL of crosslinking buffer by gentle vortexing and then centrifuged at 1500 *g* until all the buffer has passed through the membrane. The protein was resuspended in 100 µL of crosslinking buffer by gentle vortexing and repeated pipetting. Finally, the protein was transferred to a new siliconized microfuge tube and the concentration was determined by the Bradford protein assay.

The function of the microcon device (Millipore) was very similar to the Nanosep devices, in which the protein sample was concentrated and a buffer exchange was done simultaneously. Instructions from the user manual were used with minimal variation. To remove any residual contaminants that may interfere with subsequent analysis, the Microcon membrane was washed twice with 400 µL of crosslinking buffer and centrifuged at 10000 *g* until all the buffer had passed through. The volume of the purified protein from TAP was reduced to 500 µL using the speed vacuum. The protein sample was then added to the Microcon and centrifuged at 10000 *g* until all the buffer had passed through the membrane. The membrane was washed once with 400 µL of

crosslinking buffer. Finally, the protein on the membrane was resuspended in 200 µL of crosslinking buffer by gentle vortexing. The membrane was located on the sample reservoir; to remove the protein sample from the membrane, the sample reservoir was inverted, placed into the microcon microfuge tube and centrifuged at 1000 *g* for 30 seconds. Protein concentration was determined using the protein assay.

2.2.4 Protein separation and visualization

To visualize the results of the TAP procedure, protein aliquots taken from various steps were visualized on a 1D SDS-PAGE (Laemmli 1970). Due to the wide range in protein concentration, the amount of protein loaded in each lane was not uniform and was determined by the amount required to be visualized on the gel. The 0.5 to 4 µg of protein samples or 15 µL of beads were added to 2x Laemmli loading buffer (0.16 M Tris-HCl, pH 6.8, 4% SDS, 20% Glycerol) with the addition of 100 mM DTT. The samples were boiled for 5 minutes and loaded onto a 3% stacking gel, pH 6.8, which overlaid a 10% resolving gel. For the molecular weight markers, 1 µL of protein standards (Bio-Rad) with a range of 10 – 250 kDa were added to lane 1 and 10. Cold running buffer (25 mM Tris-HCl, pH 8.3, 192 mM Glycine, 0.1% SDS) was used in both upper and lower chamber. Electrophoresis was run at 50 V, 40 mA for 30 minutes and then 180 V, 40 mA until the dye front just passed the bottom of the gel. The gel was then stained using the PlusOne Silver Staining Kit, Protein (GE Healthcare). For mass spectrometry analysis, gels were washed three times for 5 minutes using ultrapure water, stained for 2 hours using Bio-Safe Coomassie stains (Bio-Rad) and destained with ultrapure water.

2.2.5 Mass spectrometry analysis

All DTT and IAA solutions are freshly prepared before use and only HPLC Grade water was used to prepare the buffers. All microfuge tubes and pipette tips were washed with detergent and 70% ethanol, and rinsed with HPLC Grade water. All steps were performed under the flowhood and all buffers were changed every 2 weeks.

The crosslinking and control samples were evaporated to dryness using a speed vacuum and then resuspended in 6 M urea in 100 mM Tris. For the reduction reaction, 0.5 µL of the reducing agent (200 mM DTT, 100 mM Tris) was added to the samples and incubated at room temperature 1 hour. For the alkylation reaction, 2 µL of the alkylation agent (200 mM IAA, 100 mM Tris) was added to the samples and incubated at room temperature for 1 hour. To eliminate the excess IAA, 2

μL of the reducing agent was added to the samples and incubated at room temperature for 30 minutes. Finally, the concentration of urea was diluted to 0.6 M by adding 75.5 μL of HPLC Grade water.

Proteolytic resistant trypsin (Sigma) was resuspended in 100 mM Tris-HCl, pH 7. Trypsin was added to the protein samples at a 1:10 trypsin:protein ratio. The reaction was incubated in a 37°C water bath for 18-20 hours. The volume of the digest was reduced to 10 μL using the speed vacuum. To stop the tryptic digestion, 1 μL of 1.0% formic acid was added to the samples.

For in-gel digestion, protein samples were first separated by SDS-PAGE and stained using Bio-Safe Coomassie (Bio-Rad). Protein bands were excised from the gel and cut up into pieces approximately 1 mm^3. The gel pieces were washed three times with HPLC Grade water by soaking and vortexing for 5 minutes. To remove the Coomassie dye, the gel pieces were vortexed in 100 μL of 50 mM NH_4HCO_3/50% ACN for 10 minutes; this wash step was repeated two more times. After the final wash, 100 μL of 100% ACN was added to the gel pieces; this step was repeated once to ensure the gel pieces were white and shrunken.

For the reduction reaction, 100 μL of the reducing agent (10 mM DTT, 100 mM NH_4HCO_3) was added to the gel pieces and incubated at 50°C for 30 minutes. To remove the excess moisture, 100 μL of 100% ACN was added to the gel pieces and incubated for 5 minutes; this step was repeated once. For alkylation, 100 μL of alkylating agent (55 mM IAA, 100 mM NH_4HCO_3) was added to the gel pieces and incubated at room temperature for 30 minutes. To remove the excess alkylating agent, 100 μL of the reducing agent was added and the reaction was incubated at room temperature for 30 minutes. The gel pieces were then washed three times by soaking and occasional vortexing with 100 μL of 100 mM NH_4HCO_3 for 15 minutes. Finally, excess moisture was removed by soaking the gel pieces in 100 μL of 100% ACN and this step was repeated once. The gel pieces were air-dried under the flowhood.

The trypsin was prepared as in the in-solution digestion. The gel pieces were rehydrated with trypsin solution at a 1:10 trypsin:protein ratio. After 10 minutes, 50 μL of 100 mM NH_4HCO_3 was added to the gel pieces and the tryptic digest reaction was incubated in a 37°C water bath for 18-20 hours.

To collect the peptides after tryptic digestion, 50 μL of HPLC Grade water was added to the samples and sonicated for 10 minutes. The peptide solution was transferred to a new microfuge tube containing 5% formic acid in 50% ACN. The gel pieces were sonicated in 75 μL of 5% FA/50% ACN for 7 minutes and the peptide solution was added to the microfuge tube; this process

was repeated once. The volume of the peptide solution was reduced to approximately 10 – 15 μL using the speed vacuum and finally, 1 μL of 1% formic acid was added for every 10 μL of the peptides mixture.

To wet the packing column in the C18 ziptip (Millipore), 10 μL of 100% ACN was pipetted into the ziptip and discarded; this was repeated three times. The ziptip was then equilibrated using 10 μL of 0.1 % formic acid; this was repeated three times. To bind the peptides to the ziptip column, the peptide solution was slowly pipetted up and down 10 times. The ziptip was washed three times with 10 μL of 0.1 % formic acid. To elute the peptides from the column, the 5 μL of 50% ACN was slowly pipetted up and down five times. Finally, the eluted peptides were collected into a 0.5 mL microfuge tube and the concentration of formic acid was adjusted to at least 0.2% using 1% formic acid.

2.2.6 Confirmation of protein identity by MS/MS

The identity of the purified proteins were confirmed by tandem MS. Proteins were analyzed by ESI-nanospray Q-TOF within the Mass Spectrometry Facility, University of Waterloo. Selected multiply-charged peaks were subjected to MS/MS analysis. Using raw MS/MS data, the MS analysis program Mascot (Perkins et al 1999) searched the MSDB (Pappin and Perkins 2005) for peptide information with *Saccharomyces cerevisiae* as the target organism and trypsin as the protease, allowing up to 1 missed cleavages. It also accounted for the possible carbamidomethyl (C) and oxidation (M) modifications in its search due to post-translational modifications or oxidation during sample preparation. Mascot calculated a score for each matched peptide and ranked the peptides in descending order by score.

2.3 Results

2.3.1 Protein target selection

Based on the selection criteria, five target proteins were selected for this project (Table 2-1). Transketolase (TKL1) is a homodimeric protein with a molecular weight of 74 kDa and 680 amino acids per subunit. This enzyme is thiamine dependent and is involved in the pentose phosphate pathway (Nikkola et al. 1994). Inorganic pyrophosphatase (IPP1) is also a homodimer with each subunit containing 287 amino acids, with molecular weight of 32 kDa. It is a cytoplasmic phosphoryl-transferase that plays a role in controlling the level of pyrophosphates in the cell, which is a by-product of biosynthesis reactions (Harutyunyan et al. 1996). The amidotransferase/cyclase (HIS7) is a monomer with a molecular weight of 62 kDa and is 552 amino acids in length. The N-terminal domain is an amidotransferase while the C-terminal domain is a cyclase. This enzyme is involved in the histidine biosynthetic pathway (Chaudhuri et al. 2003). Phosphoglycerate kinase (PGK1) is also a monomer with molecular weight of 44 kDa and 416 amino acids in sequence. This enzyme is involved in the glycolytic pathway and it catalyzes an important phosphorylation step (McPhillips et al. 1996). Finally, enolase (ENO1) is a 46 kDa protein and has a sequence length of 437 amino acids. This is a phosphopyruvate hydratase that catalyzes the reaction where 2-phosphoglycerate is converted to phosphoenolpyruvate in glycolysis as well as its reverse reaction in gluconeogenesis (Larsen et al. 1996; Balakrishnan et al. 2005).

The selected targets were high abundance proteins relative to other proteins in the cell. The crystal structure of TKL1 was readily available while structural homologues were available for the other four target proteins. The homologues contained greater than 99% sequence identity to the target sequences in all cases. This provided a correct structure for comparison with predicted structures in the later chapters. A summary of the target proteins is presented in Table 2-1.

Table 2-1. Summary of the target protein selected for TAP.

TAP-tagged Protein	PDB ID	NCBI number	Size (kDa)	Length of sequence	Number of subunits	Sequence identity to homologues
Transketolase	TKL1	NP_015399	73.8	680	homodimer	---
Inorganic pyrophosphatase	IPP1	NP_009565	32.3	287	homodimer	100% (1M38)
Amidotransferase / cyclase HIS7	HIS7	NP_009807	61.1	552	monomer	100% (1OX4A)
Phosphoglycerate kinase	PKG1	NP_009938	44.7	416	monomer	100% (1FW8)
Enolase	ENO1	NP_011770	46.8	437	monomer	99% (1EBGA)

2.3.2 Tandem affinity purification of target proteins

To visualize the protein purification process, small aliquots of samples and agarose beads were taken at various steps and compared by SDS-PAGE gels. For all target proteins in Figure 2-1 and 2-2, gel lanes represented molecular weight markers (MWM), whole cell lysate (WCL), eluate from IgG beads column (WCL-P), IgG beads (IgG.B), eluate after TEV protease incubation (P+TEV), calmodulin beads (CaM.B), eluate from calmodulin beads column (TEV-P) and target protein. HIS7, PKG1 and ENO1 had an additional lane for Sepharose 6B beads (S6B) (Figure 2-2 only). The purpose of running the Sepharose 6B, IgG and CaM beads was to determine how much protein was lost during the purification process due to non-specific protein binding to the beads as well as unsuccessful protein elution from the beads. Note that the amount of protein samples loaded across individual gels were not uniform due to the range of protein concentrations. Therefore, these results are qualitative and indicate the presence or absence of proteins rather than quantitative measurements.

The WCL lane contained the most soluble proteins in the cell. After the incubation with Sepharose 6B beads, the beads were loaded onto the S6B lanes (Figure 2-2). Multiple bands appeared in this lane indicating that a large number of proteins in the cell have a natural affinity to agarose.

The first purification step involved the binding of the target protein to the IgG beads. The WCL-P lane contained the elution from the IgG column and it should contain all the proteins in the WCL except for the target protein. However, the WCL and WCL-P lanes appeared to be very

similar and the missing target protein in the WCL-P lane was not clearly visible. In all cases except for IPP1, the target proteins were faintly visible in both WCL and WCL-P lanes.

The P+TEV lane showed the protein contents from the elution after cleavage by the TEV protease; they included the target proteins, the TEV protease as well as other proteins bound to the IgG beads. The bottom left arrow indicates the approximate location of the TEV protease.

To determine whether the TEV cleavage was complete, samples of IgG beads were taken after the cleavage step and the beads were loaded onto the gel. Multiple bands appeared in the IgG.B lane and the target proteins were clearly visible in all cases. No MS experiments were done to identify the other bands; however, these additional bands were possibly proteins on the IgG beads, proteins in the yeast cells that have a natural affinity to the IgG protein on the beads or non-specific binding to the agarose.

The second purification step was incubating the product of IgG column with CaM beads. Only the target protein containing a CaM binding domain would bind to the CaM beads. The proteins that did not bind to the CaM beads were eluted and were separated in the TEV-P lane. A small amount of target protein was visible in this lane which indicated that not all target protein bound to the CaM beads. The CaM.B lane showed what was left on the CaM beads after the addition of Ca^{2+}, eluting the target protein from the bead. There was a visible amount of target protein remaining in the CaM column, indicating that protein elution was incomplete. However, sufficient protein was purified to perform later crosslinking experiments.

Finally, all purified proteins were loaded onto their respective gels. In all cases, the proteins migrated to locations consistent with their expected molecular weight. Because the CaM binding domain was not removed from the purified proteins, the experimental molecular weight of the purified proteins appeared approximately 3.3 kDa higher than expected for the untagged proteins. With ENO1 being the only exception, each purified protein sample separated into multiple bands, which suggested the presence of additional polypeptides being co-purified. MS/MS analysis provided sequence information for ENO1 and the various bands for the other four proteins, confirming their respective identities.

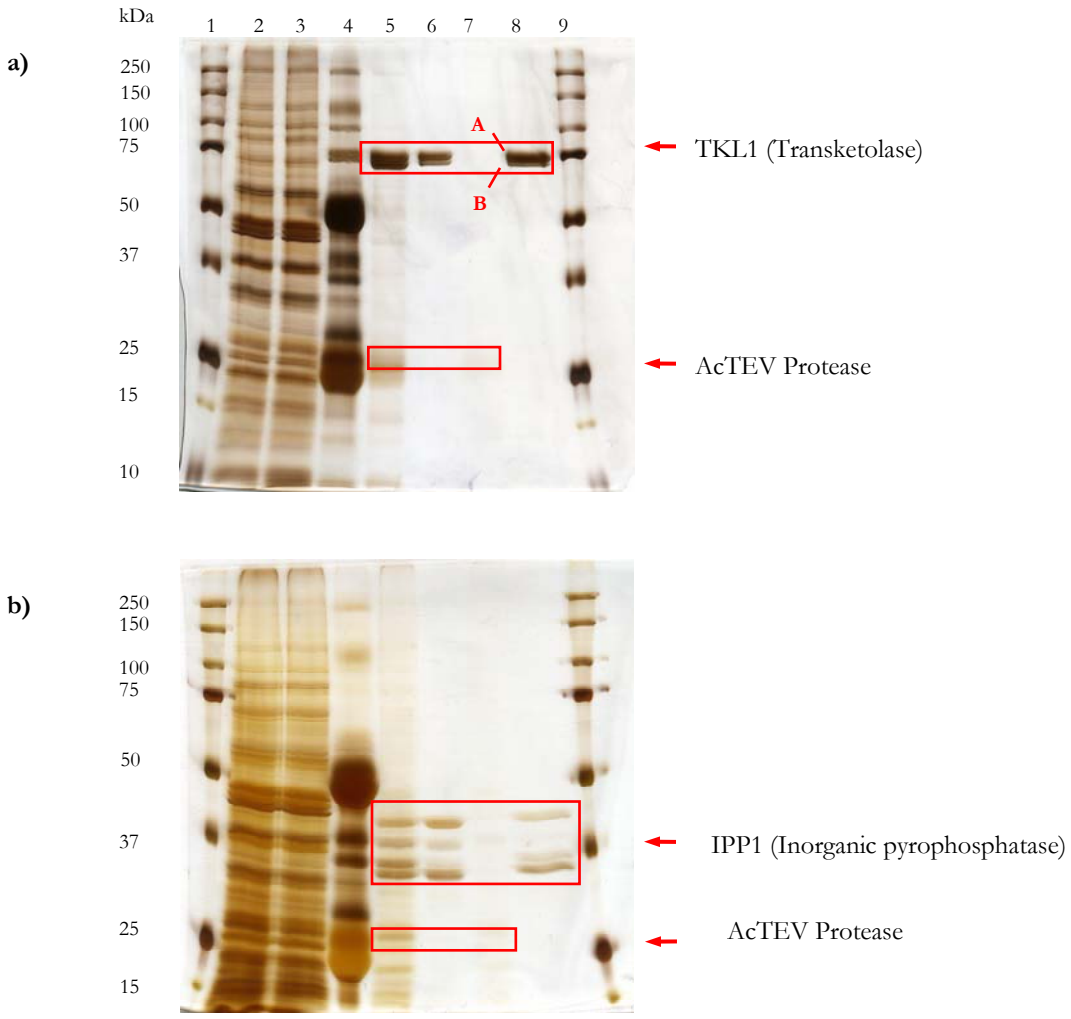


Figure 2-1. Stepwise investigation of tandem affinity purification for TKL1 and IPP1 using SDS-PAGE and silver staining. a) TKL1 and b) IPP1. Lane 1 – MWM; 2 - whole cell lysate (WCL); 3 - elution after incubation with IgG beads (WCL-P); 4 - IgG beads (IgG.B); 5 - elution after incubation with TEV protease (P+TEV); 6 - CaM beads (CaM.B); 7 - elution after incubation with CaM beads (TEV-P); 8 – target protein; 9 - MWM.

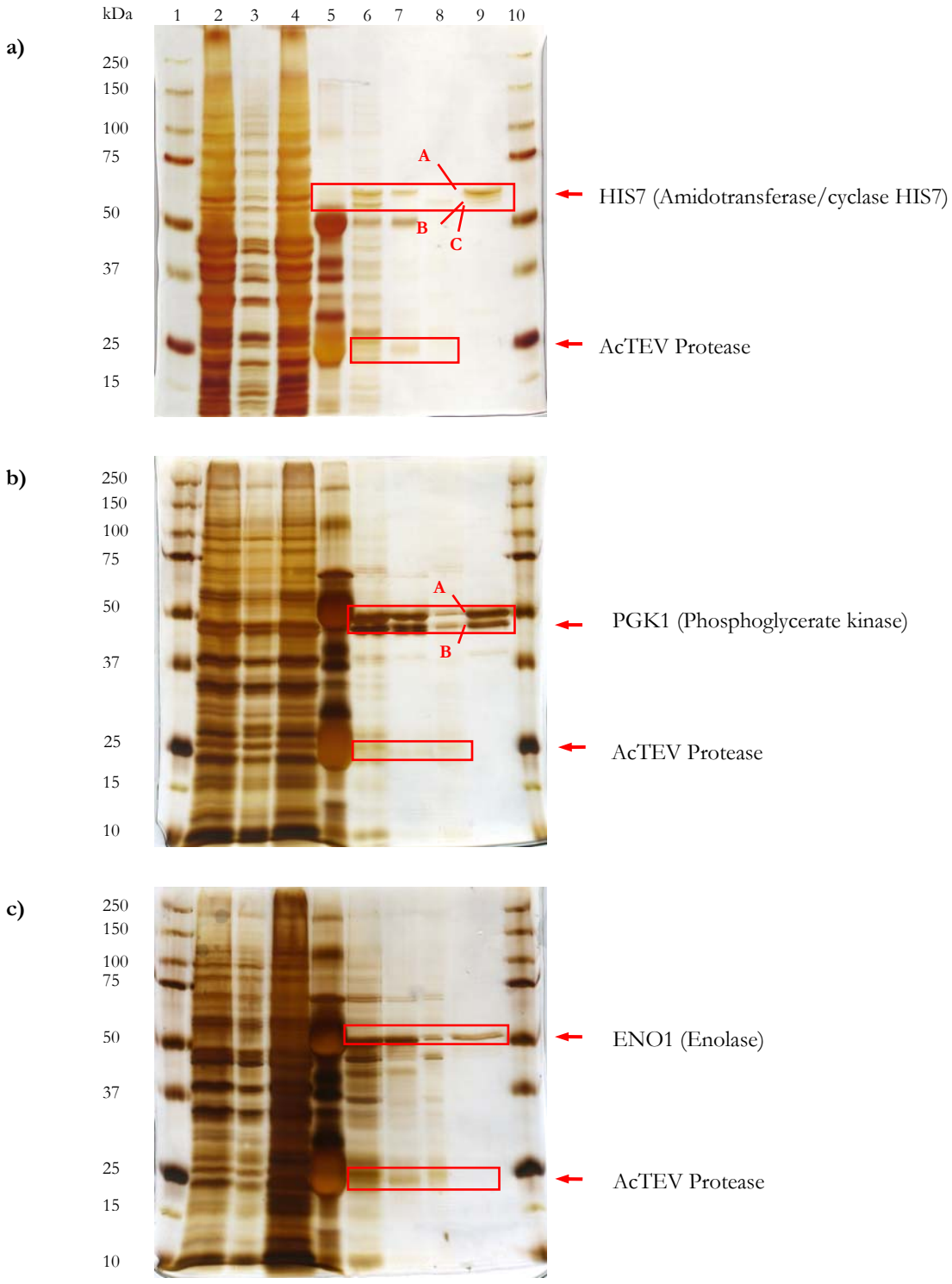


Figure 2-2. Stepwise investigation of tandem affinity purification for HIS7, PGK1 and ENO1 using SDS-PAGE and silver staining. a) HIS7, b) PGK1 and c) ENO1. Lane 1 – MWM; 2 - whole cell lysate (WCL); 3 - Sepharose 6B beads (S6B); 4 - elution after incubation with IgG beads (WCL-P); 5 - IgG beads (IgG.B); 6 - elution after incubation with TEV protease (P+TEV); 7 - CaM beads (CaM.B); 8 - elution after incubation with CaM beads (TEV-P); 9 – target protein; 10 - MWM.

2.3.3 MS identification of TAP proteins

To determine the identity of the purified proteins, TKL1, IPP1, HI7 and PGK1 were separated on 1D SDS-PAGE and excised protein bands were subjected to MS analysis. An exception to this was ENO1, which showed only one band on the 1D gel. ENO1 was therefore prepared directly for MS analysis using an in-solution digestion protocol. Tandem MS was used to obtain sequence information for positive identification of all proteins (Table 2-2). Transketolase was the top query result returned by Mascot for both band A and B. The same results were found for HIS7, PGK1 and ENO1, confirming the identity of each band to be the target protein. Inorganic pyrophosphatase was returned as the top query for IPP1 bands, A and C, but it was only returned as the third on the list for band C. The verification of the target protein identities proved the success of the protein purification.

Table 2-2. Top Mascot results for the protein identification of the TAP proteins using MS/MS.

PDB ID	Protein	Band	Peptide Sequences	Score
TKL1	Transketolase	A	RYEAYGWEVLYVENGNEDLAGIAK LGNLIAIYDDNKITIDGATSISFDEDVAK LSETVLEDVYNQLPELIGGSADLTPSNLTR	143
		B	HTPSIIALSR QNLPQLEGSSIESASK TQFTDIDKLAVSTIR	115
IPP1	Inorganic pyrophosphatase	A	VIAIDINDPLAPK	59
		B	AASDAIPPASPK	11
		C	AASDAIPPASPK VIAIDINDPLAPK ALGIMALLDEGETDWK LEITKEETLNPIIQDTK	224
HIS7	Amidotransferase / cyclase HIS7	A	SGADKVSIGTDAVYAAEK YYFVHSFAAILNSEK.K TTQQGADEVTFLNITSFR	204
		B	GDGTSPIETISK AGLNVIENFLK DCPLKDTPMLEVLK	120
		C	DLGVWELTR STGLNYIDFK AGLNVIENFLK YGSEEFIAAVNK AYGAQAVVISVDPK	127
PGK1	Phosphoglycerate kinase	A	AHSSMVGFDLPQR	27
		B	ASAPGSVILLENLP VLENTEIGDSIFDK TIVWNGPPGVFEFEK	206
ENO1	Enolase		NVNDVIAPAFVK GNPTVEVELTTEK	18 59

2.4 Discussion

2.4.1 Detail analysis of TAP protocol

To investigate the efficiency of each stage, small aliquots of samples were taken at various steps of TAP and the protein contents of these aliquots were visualized by 1D SDS-PAGE. Due to the difference in protein concentration, the amount of protein loaded into each lane was not uniform across the gels and thus, it was a qualitative rather than quantitative measure.

Although the target proteins were relatively high in abundance, it is not surprising that they were not clearly visible in the whole cell lysates. It is likely that the target proteins were hidden by the thousands of proteins present in the lysates. Protein IPP1 was the only exception where the target protein was visible in the whole cell lysate. This protein band contained three possibilities: target protein IPP1 only, some other protein(s) with the same molecular weight as IPP1, or a combination of IPP1 and some other protein(s). No MS analysis was done to identify the protein(s) present in this band.

The first purification step involved TAP-tagged target proteins binding to the IgG beads via the IgG Binding Domain. This binding domain recognized protein A that was conjugated to the agarose beads and any protein without this IgG binding domain should pass through the column. The WCL-P sample was taken after the incubation with IgG beads and therefore, IgG binding proteins, including the target protein, would be trapped in the IgG beads column while the unbound proteins were eluted. Thus, the protein content in this sample should be identical to whole cell lysate except for the absence of the target proteins. Unfortunately, the target proteins were not visible in the whole cell lysates and therefore, the absences of the target proteins in the WCL-P samples were inconclusive. IPP1 was the only exception.

A band, potentially the target protein IPP1, was present in the WCL-P sample. As in the case of the WCL sample, no MS analysis was done and therefore the protein identity of this band cannot be confirmed. However, the presence of this band implied that the band contained the target protein as well as other protein(s) with the same molecular weight. If this band contained IPP1 only, then one would expect to see this band disappear in the WCL-P lane because IPP1 should be bound to the IgG beads in the column. But since the darkness and size of the band was very similar between the WCL and WCL-P sample, this suggested that the bands contained proteins other than IPP1.

To retrieve the target protein from the IgG column, an enhanced derivative of the Tobacco Etch Virus proteases was used. The AcTEV protease recognizes a conserved sequence, Glu-Asn-Leu-Tyr-Phe-Gln-Gly, and it cleaves between Gln and Gly with high specificity (Carrington and Dougherty 1988; Dougherty et al. 1988; Nayak et al. 2003). The TAP-tagged proteins contained this cleavage sequence and allowed the AcTEV proteases to release the target proteins from the beads with high specificity. After incubation with TEV, the eluted sample should contain primarily the target protein and the AcTEV protease. The P+TEV sample was taken after AcTEV incubation and the lanes for all cases clearly showed the presence of many proteins in addition to the targets. AcTEV has a molecular weight of 29 kDa (Nayak et al. 2003) and based on the molecular weight markers, it can be deduced that the faint band appearing slightly above the 25 kDa marker was the AcTEV, indicated by the lower left arrow in Figure 2-1 and 2-2. No MS analysis was done to determine the identities of these bands since this was only an intermediate step in the purification protocol. However, these results did confirm the need for a second purification step. If IgG beads were the only purification step, the additional bands would suggest that the target protein is a member of a complex, leading to false conclusions regarding the target protein.

To determine the efficiency of the AcTEV protease, a small aliquot of IgG beads was loaded onto the gel after the target protein and the protease were eluted from the columns. Multiple bands, including the target protein, appeared in the IgG.B lane. The presence of the target protein suggested that the cleavage by AcTEV protease was not 100% successful. The protease efficiency may possibly improve by increasing the amount of AcTEV or increasing the incubation period. Also, insufficient washing of the IgG beads after the cleavage step could also contribute to incomplete protein elution from the column; additional wash steps should improve recovery. When a comparison was made, more than 10 bands uniformly appeared in the IgG.B lane across all 5 gels and it suggested that these bands did not reflect on the efficiency of protease activity, but rather on an issue intrinsic to the IgG beads. Firstly, IgG beads contained bound protein A from the TAP-tag (Rigaut et al. 1999) and it was expected that protein A or IgG could appear on the gel. Secondly, there may be other unknown proteins conjugated to the beads that may also be visible on the gel. Thirdly, there may be other non-TAP-tagged proteins in the cell lysate that have an affinity to IgG and these will also appear on this lane. Fourthly, some of these protein bands may be a result of unspecific binding to the agarose bead itself and not to protein A. This suggested that incubation with Sepharose 6B beads may not have completely eliminated all agarose binding proteins.

The second step of the purification procedure required the target protein mixture eluted from the IgG column to be incubated with CaM beads. The addition of Ca^{2+} induced a conformational change in the CaM, exposing a region with a high affinity for the CaM binding domain (Voet and Voet 1995). The TAP-tag contained a CaM binding domain that allowed the target protein to bind to the CaM beads. Proteins lacking this TAP-tag should not bind to the CaM beads and therefore would pass through the column. These proteins were separated in the TEV-P lane. An aliquot of sample was taken after incubation with the CaM beads. This sample should only contain the TEV protease as well as the other non-target proteins that appeared in the P+TEV lane since the target protein was bound to the CaM beads. In Figure 2-1 and Figure 2-2a, the TEV-P lanes showed almost no protein bands, indicating that any proteins present were so low in concentration that they could not be effectively visualized by silver stain. However, in Figure 2-2b and c, the target protein bands were clearly visible. There are four possible explanations for this. First of all, there were more target proteins in the buffer than available binding sites on the CaM beads; therefore, the CaM bead column could not capture all of the target proteins and thus the proteins were eluted with the TEV protease. Secondly, the non-covalent binding between the CaM beads and the CaM binding domain was not very strong and some of the proteins dissociated from the beads during the elution with TEV. These two reasons are less likely, as the same results should appear in all of the gels rather than only in a few. Another plausible explanation is the length of the incubation time with the CaM beads. The incubation time was between 2-4 hours; more target proteins may bind to the beads if the incubation period were longer. Finally, the binding affinity of the CaM binding domain to the CaM beads may be altered by the overall structure of the tagged protein. This will also explain why the results were inconsistent between protein samples, since steric effects will be dependent on the individual protein structures.

The final step of TAP was to elute the target protein from the CaM bead column. The addition of the chelating agent, EGTA, releases the Ca^{2+} from the CaM, inducing a conformational change. This allows the CaM binding domain of the target protein to dissociate from the CaM and thus elute from the column. The target proteins were visualized in the second last lane on each gel, labeled with their respective identifier. Protein ENO1 (Figure 2-2c) was the only case where a single band appeared in the lane, indicating that the protein is either monomeric or homomultimeric. In the other four cases, the purification process resulted in multiple bands, leading to two possible explanations. A protein sample separating into multiple bands generally implies that the protein is a complex with several different subunits. This was unlikely the case since one of the criteria for

target selection was that the protein must be a monomer or a homodimer. Furthermore, the bands for TKL1, HIS7 and PKG1 were located very close together, leading to a second possibility that these bands were the result of post-translational modifications (PTM) on the target proteins. A sufficient number of PTMs on a protein could lead to significant changes in molecular weight or charge, altering the protein's mobility and producing a visible separation on a gel. To verify the identity of the purified proteins and their respective bands, MS/MS analysis was done. The results confirmed the identity of all target proteins and thus the success of the TAP method. Furthermore, the identity of each protein band corresponded to the expected target, indicating that the target proteins were pure and the molecular weight differences were likely the result of PTMs.

To determine whether the elution from the CaM beads was efficient, small aliquots of CaM beads were taken after the elution of the target protein for electrophoretic analysis. In all cases, target proteins were present in the CaM.B lane, indicating that the dissociation of the target from the CaM was incomplete. The elution of the target protein was simply a "wash" step with the CaM elution buffer, containing EGTA. It is possible that the EGTA did not have enough time to release all of the Ca^{2+} from the CaM and the target protein could not dissociate from the CaM. Instead of simply passing the elution buffer through the column, it may be beneficial to allow the beads to soak in the buffer for an extended period, giving ample time for EGTA to chelate the Ca^{2+} and release the target protein into the solution.

2.4.2 Verification of TAP proteins

Although the molecular weights of the purified proteins were close to expected results, MS/MS analysis can provide a positive confirmation of the proteins' identities. The raw MS/MS data for the selected peptides were combined and submitted to the Mascot for ion searches. For TKL1, HIS7, PKG1 and ENO1, the expected proteins were returned as the top match from the MSDB database. For IPP1, inorganic pyrophosphatase were ranked highest for band A and C (Table 2-2); however, the middle band was only ranked third. In Figure 2-2b and c, it appeared that band B did not stain as dark as the other two bands and this indicated that the quantity of protein in this band was comparably less. The quality of the MS/MS data is dependent on the quantity of the protein. Therefore, a band with lesser amount of protein was expected to result in poorer MS/MS data, which in turn would affect the outcome of the Mascot search. However, Mascot matched inorganic pyrophosphatase to this band and it was an expected result. Given the positive results for the other proteins, it is likely that IPP1 band B was also an inorganic pyrophosphatase. MS analysis

confirmed the identity of the TAP proteins and therefore proved the success of the TAP procedure for isolating target proteins.

3 Chemical crosslinking of target proteins

3.1 Introduction

To bridge the gap between genomic and proteomic information, it is necessary to develop high throughput methods for studying protein structures. Traditional approaches such as NMR or X-ray crystallography are time consuming and may not be applicable to all proteins, while protein models are not often experimentally validated. In 2000, Young et al. explored the possibility of using chemical crosslinking in combination with mass spectrometry to obtain low resolution structural information. Using a homobifunctional chemical crosslinker, BS^3, they were able to obtain experimental data to select the correct model for the FGF-2, fibroblast growth factor 2.

The same crosslinker, BS^3, was used in this project. BS^3 has two lysine specific reactive groups at each end of a flexible spacer arm. The maximum span of BS^3 from one lysine to the other is less than 24 Å when it is crosslinked to protein surface. This introduces a distance constraint to the structural analysis, where any two peptides crosslinked by BS^3 must be close in proximity. This distance constraint could be used to validate or refute any protein model.

One important aspect of this crosslinking technique is determining the ideal crosslinking conditions which will result in 1-2 BS3 per protein molecule. Excessive intramolecular crosslinking may result in structural distortion (Sinz 2003; Young et al. 2000). It can also decrease solubility and affect the efficiency of proteolysis during MS preparations (Pearson et al. 2002). Moreover, excessive crosslinking would increase undesirable intermolecular crosslinking and the formation of protein aggregates (Sinz 2003). Thus a delicate balance is necessary to avoid excessive crosslinking while ensuring sufficient crosslinking for MS detection. Optimal crosslinking conditions can be determined by gel electrophoresis. Figure 3-1 is a diagram showing three possible crosslinking scenarios. In the ideal situation, scenario C has only 1 to 2 crosslinks present on the protein surface. Since the molecular weight of BS^3 is only 138 Da, the addition of 1 to 2 BS^3 would not dramatically alter the molecular weight of the crosslinked protein. However, the migration rate of the crosslinked protein will partly depend on where the crosslinking occurs. For scenario C in lane $+BS^3$, the crosslinker did not significantly change the shape of the unfolded protein and therefore, those proteins will migrate close to the rate of a non-crosslinked protein (in lane $–BS^3$). For scenario B, the location of the crosslinker increased the bulkiness of the protein and thus decreasing its

mobility. Finally, intermolecular crosslinking, indicated by A, should be minimized since it will not contribute to structural analysis and will increase the complexity of data analysis.

Crosslinked proteins are digested by proteases prior to MS analysis. In the case of trypsin where proteolysis occurs at lysine residues, lysines modified by BS^3 cannot be cleaved and the resulting product will be comprised of four peptides (Figure 3-2a). Proteases with non-lysine specificity, such as Asp-N, will result in crosslinking product composed of two single peptides (Figure 3-2b).

Upon analysis by MS, all crosslinked products will produce an additional set of mass values that is absent from the non-crosslinked sample. To identify crosslinked peptides, crosslinked and non-crosslinked spectra are compared and mass peaks unique to the crosslinked spectrum are considered to be potential crosslinked peptides. The MS data will be analyzed in the next chapter using the Automatic Spectrum Assignment Program (ASAP) and sequence information will be assigned to some of the crosslinked peptides (Young et al. 2000). The experimental data will then be validated against the structures of the target proteins.

In this chapter, the five target proteins were crosslinked with BS^3 which conjugates surface lysine residues. The crosslinked proteins were digested with trypsin and the peptide mixture were analyzed by MS.

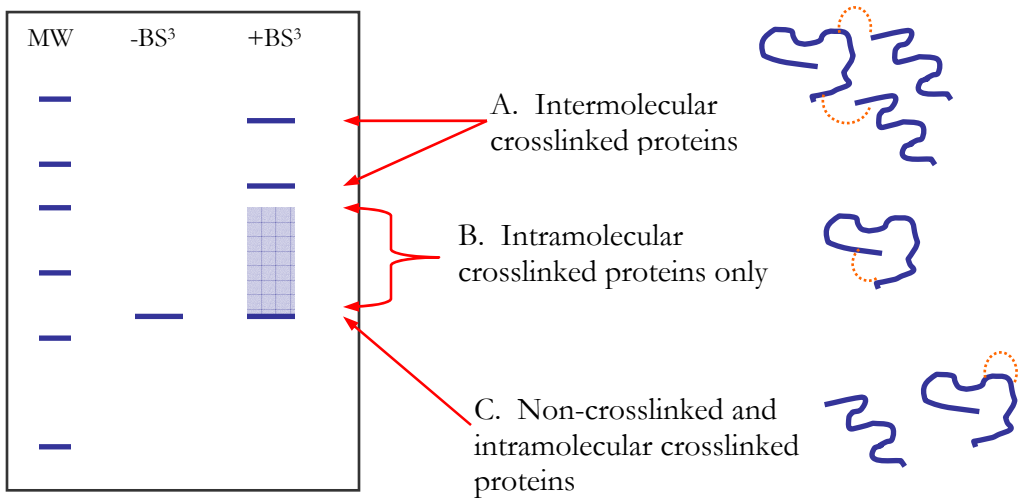


Figure 3-1. Illustration of a 1D SDS-PAGE gel with intermolecular and intramolecular crosslinked proteins.

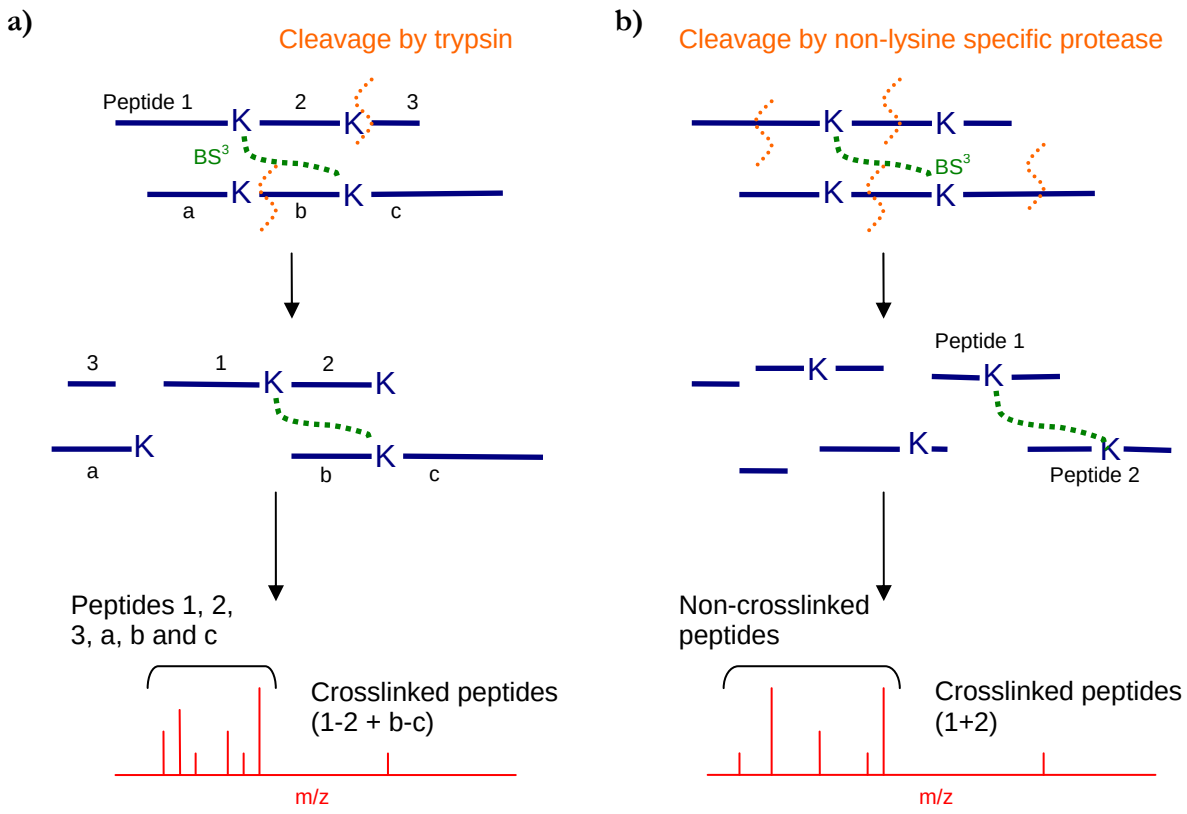


Figure 3-2. Diagram of MS spectra from protein crosslinked with BS^3, followed by proteolysis with trypsin or a non-lysine specific protease. a) Digestion with trypsin results in the crosslinking of 4 peptides. b) Digestion with a non-lysine protease results in the crosslinking of 2 peptides.

3.2 Materials and methods

3.2.1 Chemical crosslinking

Two μg of TAP purified protein was added to a pre-washed siliconized microfuge tube at a concentration of 5 μM. Chemical crosslinking reagent, BS^3 (Pierce Biotechnology) was dissolved in crosslinking buffer at a concentration of 400 mM with 10 mM DTT. In the crosslinking reaction, BS^3 was added to the pure protein at 20 times molar excess. The crosslinking sample was incubated overnight at 4°C. The reaction was quenched with 10 mM Tris-HCl, pH 8. Iodoacetamide was not added to react with the DTT since the samples were immediately denatured for either SDS-PAGE or MS. A control was prepared for each crosslinked reaction; it is identical to the crosslinking sample in every respect except BS^3 was omitted from the crosslinking buffer. Crosslinked and control samples are then prepared for mass spectrometry analysis.

3.2.2 MS spectra analysis of crosslinked proteins

Crosslinked samples were visualized on SDS-PAGE gel to verify that there were relatively few intermolecular crosslinking events. Crosslinked protein bands excised from some of the gels were prepared for MS analysis using the in-gel digestion protocol. Crosslinked proteins that were not prepared by electrophoresis were subjected to in-solution digestion immediately after the completion of the crosslinking reaction. Trypsin was used to reduce all protein samples into peptide fragments. The digestion of the crosslinked protein produced a mixture of crosslinked and non-crosslinked peptides which was analyzed by ESI Q-TOF. To identify the crosslinked peptides, the crosslinked protein spectrum was compared to the control spectrum. Peaks that are unique to the crosslinked samples are recorded as potential crosslinked peptides. Details of SDS-PAGE and MS sample preparation are as described in Chapter 2, Section 2.2.4 and 2.2.5.

3.3 Results

3.3.1 Chemical crosslinking of TAP purified proteins

To ensure interprotein crosslinking was minimal, the control and crosslinked samples were separated on 1D SDS-PAGE and stained with Bio-Safe Coomassie. Figure 3-3 showed sample gel images for TKL1 and PGK1. As expected, the bands in the crosslinked sample were at approximately the same molecular weight as the control in both gels. Given the molecular weight of BS^3 was 138 Da and the objective was achieving 1 to 2 crosslinks per protein, the addition of BS^3 to the protein should not drastically change the molecular weight. Any intermolecular crosslinked proteins would appear at a higher molecular weight (Figure 3-4). TKL1 in Figure 3-3a has an expected molecular weight of 73.8 kDa (control, -BS^3); as expected, the crosslinked sample (+BS^3) also appeared at approximately 75 kDa. Since bands representing higher molecular weight proteins were not visible on these gels, this suggested that the intermolecular proteins were at a much lower concentration than the intramolecular crosslinked proteins. Therefore, the ratio of protein: BS^3 mixture of 1:20 did not result in significant intermolecular crosslinking. Similar results were also obtained for IPP1, HIS7 and ENO1.

Figure 3-4 showed an example of over-crosslinking where the ratio of protein to BS^3 was 1:40 where as the ratio in Figure 3-3 were at 1:20. The presence of the 147.6 kDa band indicated that protein to BS^3 ratio of 1:40 was too high, leading to intermolecular crosslinking for TKL1. Because TKL1 is a homodimer, the upper band represents the sum of two crosslinking possibilities. Crosslinking may have occurred between the two subunits of the same dimer (Figure 3-4 A) or between subunits from different dimers (Figure 3-4 B), both of which are expected in the upper band in Figure 3-4. For monomeric protein, the gel would be expected to look the same as a homodimeric protein sample but the upper band would represent intermolecular crosslinking occurring between two different proteins.

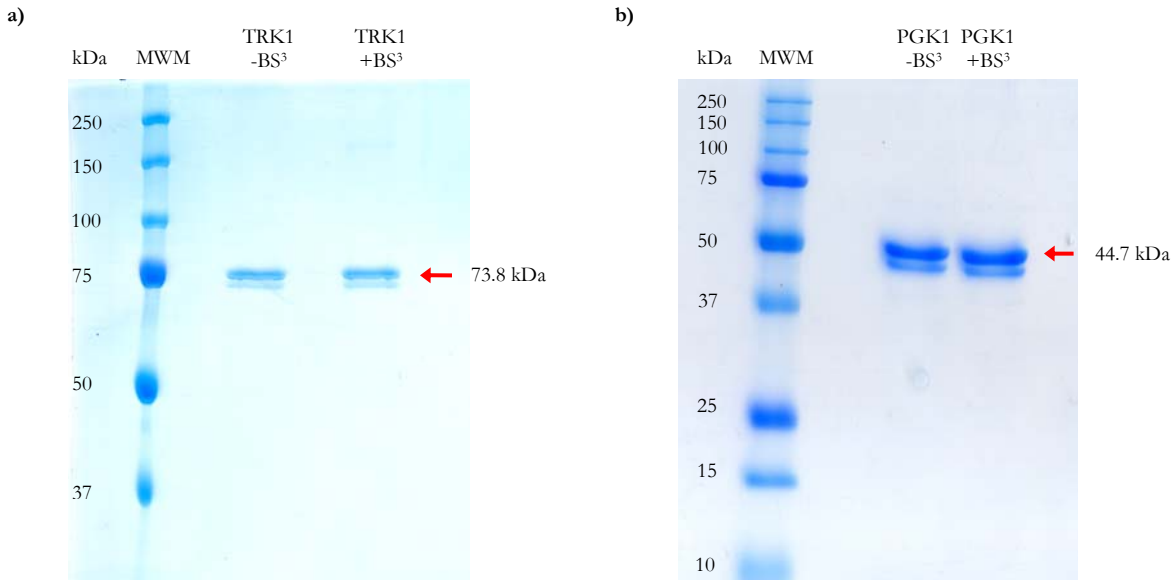


Figure 3-3. Establishing chemical crosslinking conditions using 1D SDS-PAGE and Coomassie staining. The –BS³ lane was the noncrosslinked protein sample (control) and the +BS³ lane was the crosslinked protein sample. The arrows on the right side indicated the expected molecular weight for non-crosslinked protein (–BS³). a) TKL1. b) PKG1.

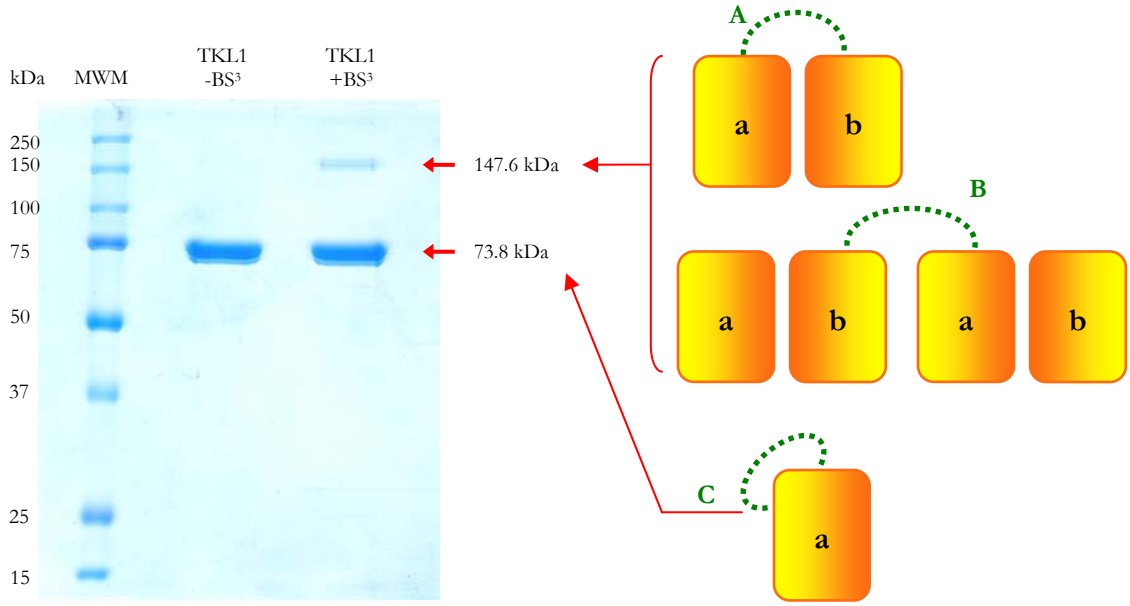


Figure 3-4. Crosslinking conditions resulting in detectable amount of intermolecular crosslinking. Here, a ratio of protein:BS³ of 1:40 was used instead of 1:20. The lower band in the +BS³ lane showed crosslinking within a single subunit of TKL1 (C). The upper band represents the intermolecular crosslinking between two TKL1 subunits, either within the same protein dimer (A) or between subunits from two separate dimers (B).

3.3.2 Identification of potential crosslinking peptides

The protein bands isolated from intramolecular crosslinked proteins, as shown in Figure 3-3, were subjected to in-gel digestion with trypsin and then MS analysis by ESI Q-TOF. Digestion with trypsin was allowed to incubate for 16 to 20 hours. The comparison of spectra from in-gel digestion was difficult because the differences between the control and crosslinkers were not always clear. Examples of the MS spectra for the control and the crosslinked sample of IPP1 from an in-solution digestion are shown in Figure 3-5 and Figure 3-6. The top spectrum was the control sample while the bottom two represented crosslinked samples. No prominent differences could be seen between the crosslinked and control MS spectra (Figure 3-5). Comparing the spectra in the m/z region of 718 to 722 (Figure 3-5a), the 719.0 peak appeared in both crosslinking spectra at an intensity of 1410 and 2100 while the same peak in the control spectrum, with an intensity of approximately 200, was barely distinguishable from the background. In this example, the sequence separation between the two lysines is less than 7 amino acids apart and did not meet the criteria for structurally relevant crosslinked peptides. Next to the 719.0 peak, another crosslinked peptide was visible at a lower intensity. Zooming in on this region (Figure 3-6b), it was evident that this group of peaks represented a potential crosslinked peptide since the intensity of the peaks from the non-crosslinked spectrum (top) was indistinguishable from the background. The relative intensities of the peaks from the crosslinked spectrum were 640 (middle) and 1070 (bottom), respectively, compared to the 17 in the non-crosslinked sample. This low peak intensity indicated that the peaks in the non-crosslinked spectrum may be background noise. Table 3-1 summarizes the number of peaks identified as potential crosslinked peptides for each protein. A complete list of m/z value and charge for each peak can be found in Appendix A.

Sequence assignments were performed using the Automatic Spectrum Assignment Program (ASAP); computational analysis of crosslinking data will be discussed in more detail in the following chapter. Mass peaks unique to the crosslinked protein spectra were found for all of the different target protein samples. Note that special care was taken when selecting singly charged peaks because they can also be produced by contamination present in the sample or the equipment. Therefore, singly charged peaks were only included in the crosslink data set if they were unique to the particular crosslinked protein and not present in other crosslinked samples.

Table 3-1. Number of potential experimental crosslinking sites for each target protein.

Protein	TKL1	IPP1	HIS7	PGK1	ENO1
Number of peaks unique to crosslinked spectrum	216	168	43	43	112

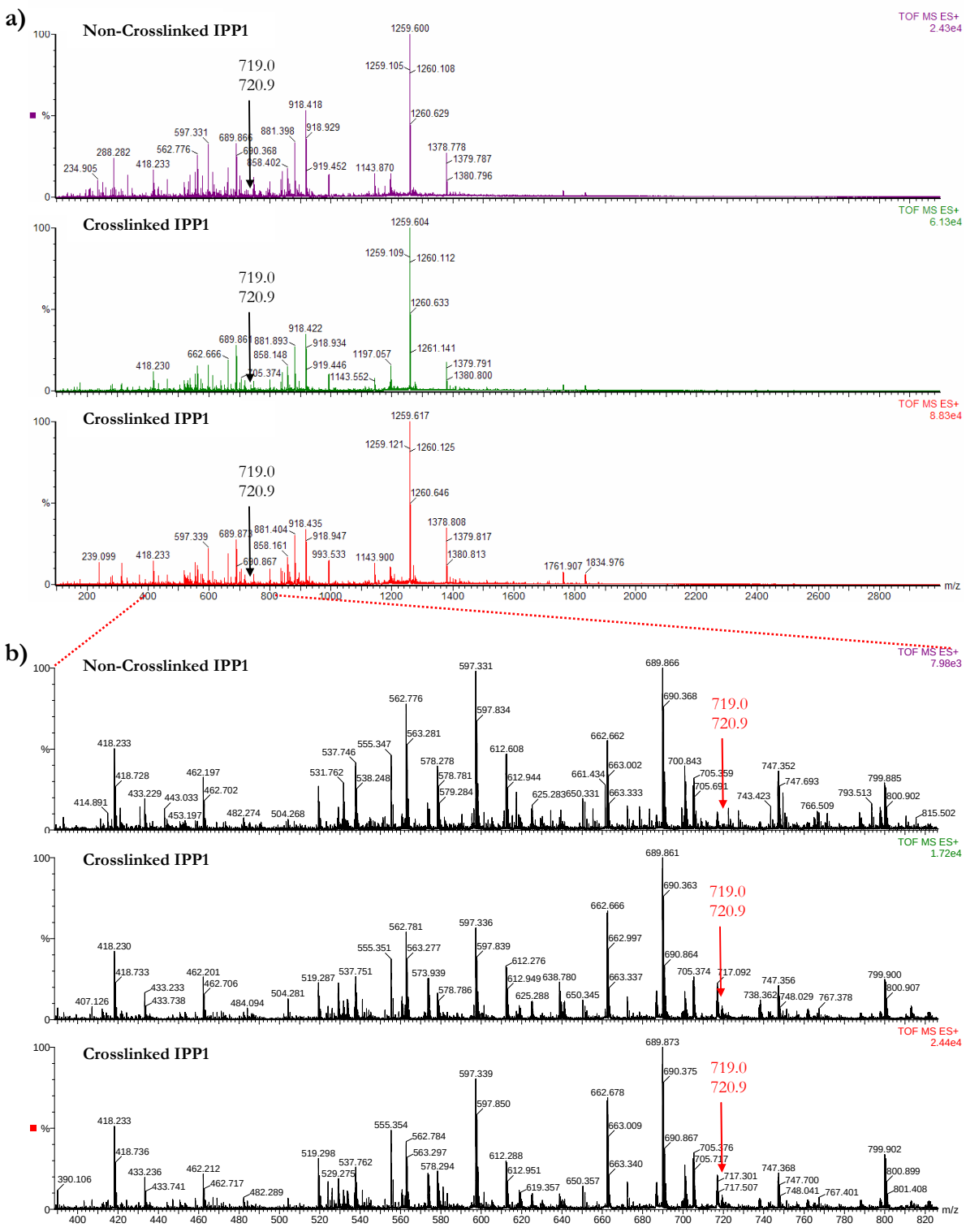


Figure 3-5. MS spectra comparison between non-crosslinked and crosslinked IPP1 using in-solution digestion. a) The peak patterns were very similar between the full non-crosslinked (top) and crosslinked spectra (middle and bottom). b) Small differences can be seen between the non-crosslinked and the crosslinked spectra near 720 m/z area when the 400-800 m/z region is expanded.

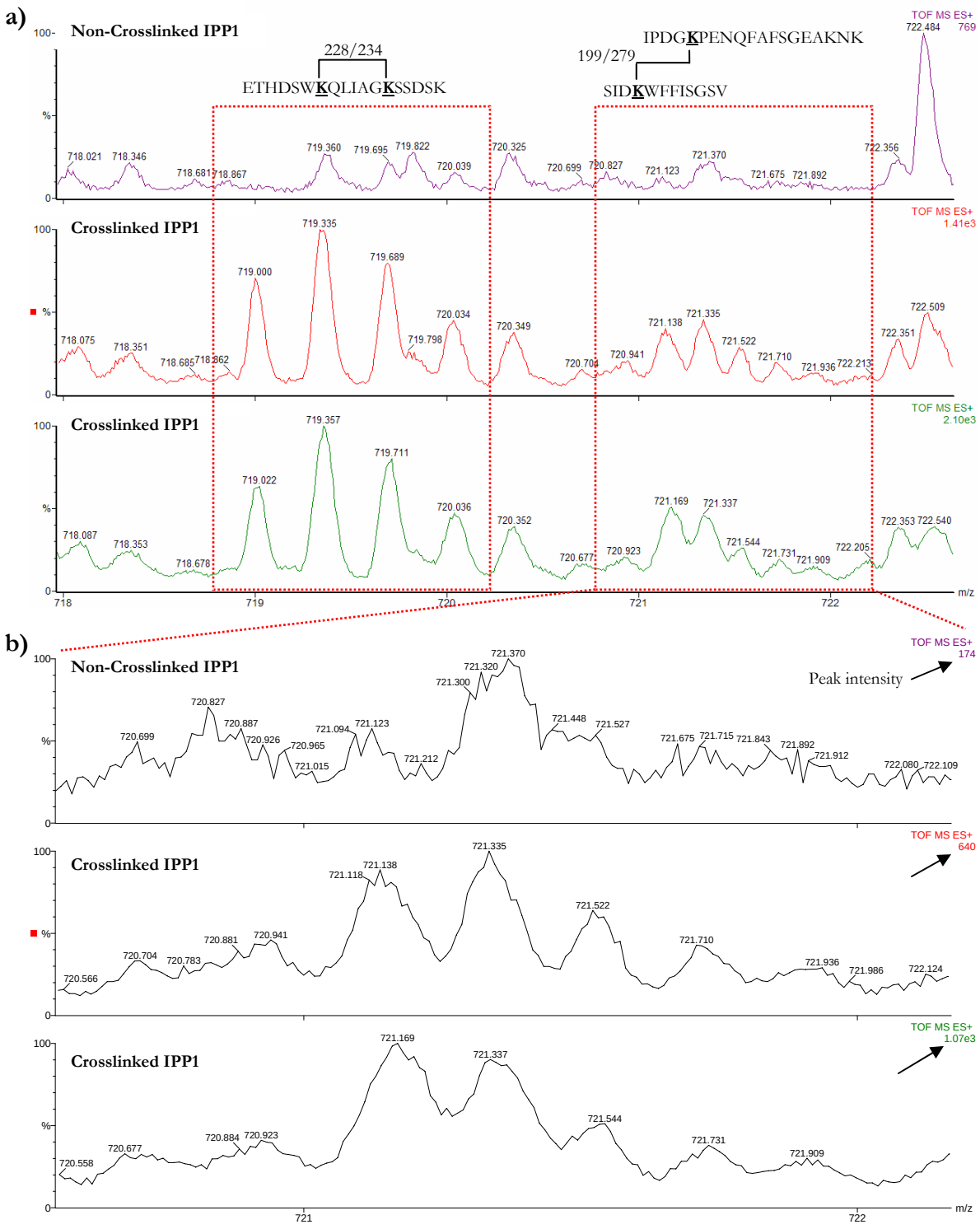


Figure 3-6. Comparison of mass peaks at 719 and 720 (m/z) between non-crosslinked and crosslinked IPP1. a) Peaks are visibly higher in the crosslinked spectra (middle and bottom) than the non-crosslinked spectrum (top). b) The arrow at the top right corner of each spectrum showed the intensity of the highest peak and it showed that the peak intensity is much higher in the crosslinked spectra than in the non-crosslinked.

3.4 Discussion

3.4.1 Establishing optimal crosslinking conditions

Gel electrophoresis enables a simple inspection of crosslinked products and it was useful for establishing crosslinking conditions. This project made the assumption that if the intermolecular crosslinked proteins cannot be visualized by Coomassie staining, then the amount present in the sample is negligible. However, this assumption does not imply the complete absence of intermolecular crosslinked proteins. For a more sensitive detection method, MS can be used to establish crosslinking conditions (Young et al. 2000). Multiple crosslinking conditions can be setup with various protein:crosslinker ratios. These crosslinked proteins are then input into the mass spectrometer without proteolytic digestion. Intact non-crosslinked proteins will result in a single peak on the mass spectrum while the intramolecular crosslinked proteins will produce a second peak to the right of the non-crosslinked protein with the difference in mass corresponding to the number of crosslinker present on the protein. Intermolecular crosslinked proteins will appear further down the spectrum and the mass difference will correspond to the number of copies of the protein as well as the number of crosslinkers present. The optimal crosslinking conditions are a balance between maximizing the intramolecular crosslinking peaks and minimizing the intermolecular crosslinking peaks (Sinz 2003; Young et al. 2000).

3.4.2 Analysis of crosslinked proteins

Crosslinked proteins were subjected to MS analysis and crosslinked peptides were identified by selecting differences between the control and crosslinked spectra. These differences were difficult to distinguish and also these differences were few in number (Figure 3-5 and Figure 3-6). Loss of peptide information is an intrinsic problem to in-gel digestion for MS preparation. On the other hand, in-solution MS preparation was a more favourable method because it is less time consuming, sample loss is minimal and digestion by trypsin is more efficient (data not shown). In comparison, a spectrum from an in-solution digest has higher peak intensity and contains more information than a spectrum from an in-gel digest. It is also reasonable to assume that the non-crosslinked proteins in the band were more abundant than the crosslinked proteins, therefore reducing the relative signal of the crosslinked peptides. Moreover, information on proteins whose mobility was altered due to chemical crosslinking would be lost using a band excised from a gel since they could not be detected by Coomassie (Figure 3-1 B), but these would also provide the most

valuable structure information. Crosslinkers connecting two lysines that were far apart in sequence but close together in space can change the shape of the protein in a denaturing gel, thus altering its mobility. Whereas for the crosslinked proteins in the visible band, the crosslinked lysines are closer together in sequence rather than in 3D space, and therefore no significant change protein mobility is observed (Figure 3-1 C). Since the goal of this project was to determine whether the crosslinking method could identify the correct protein models from a library of folds, crosslinked lysines that are close in spatial proximity but separated in sequence can provide more valuable information than lysines that are close together in sequence. For these reasons, it was decided to forgo the SDS-PAGE and the crosslinked samples were prepared directly for MS analysis, although 1D gel was still used to determine the crosslinking condition to minimize intermolecular crosslinking. Intermolecular crosslinked proteins would produce crosslinked peptides that could be detected by MS. An alternative method to eliminate the intermolecular crosslinked peptides without using 1D gel is to perform a size exclusion (SE) experiment immediately after the initial protein crosslinking reaction. This method would eliminate the intermolecular crosslinked proteins and therefore, only intramolecular crosslinked peptides would be present in the final mixture for MS analysis. However, SE was not required in the experimental design as it was shown that there are relatively few intermolecular crosslinking events.

3.4.3 Alternative crosslinking strategies

Mass spectrometry is a highly sensitive method for detecting the presence of low abundance peptides; however, identifying the crosslinked peptides in pool of peptides is a difficult task. Given protein concentrations at a micromolar range, the crosslinked peptides were even less abundant. It is possible that the signals for non-crosslinked peptides or the background noise overpowered the weaker signals of crosslinked peptides. Also, a single peak can represent a mixture of crosslinked and non-crosslinked peptides. Such issues can render the crosslinked peptides undetectable and valuable information could be lost. To take full advantage of the crosslinking method, other recently introduced chemical crosslinkers can be employed. Different reactive groups in a homo- or heterobifunctional crosslinker allow chemical specificities for different amino acids. Also, the length of the spacer arms can range from 0 to 24 Å (Sinz 2003; Pierce Biotechnology 2002).

Isotopically labeled crosslinkers provide a means to distinguish the crosslinked peptides from non-crosslinked peptides. In 2002, Pearson et al. synthesized an isotopically labeled crosslinker, disuccinimidyladipate (DSA) and a 1:1 mixture of d_0/d_8-DSA was introduced to cytochrome c for

structural analysis. MS results generated pairs of distinct doublets that were 8 Da apart. Muller et al. (2001) also performed a similar experiment using d_0/d_4-labeled crosslinker in a protein interaction study for the Op18/stathmin and tubulin; this resulted in the formation of doublet pairs spacing 4 Da apart. They also showed that by using combinations of labeled crosslinkers with the same functional groups but different arm lengths, the detection of low abundant crosslinked peptides can be enhanced. At the time this research was conducted, isotopic labeled BS^3 (d_0/d_4) was not available but is now commercially available from Pierce Biotechnology for future improvement of the method proposed in this project. Additionally, BS^2 (d_0/d_4) with a shorter arm length of 7.7 Å is also available. The combination of using BS^2- and BS^3-d_0/d_4 can further enhance the detection of the crosslinked peptides.

In the event where the amount of crosslinked peptides is insufficient to produce clear peak signals that can be distinguished from the background, a tri-functional chemical crosslinker can be used. An affinity handle can be incorporated into the crosslinker as third functional group and used to purify the crosslinked peptide from the crude mixture and enrich the sample for MS analysis. Trester-Zedlitz et al. (2003) synthesized tri-functional crosslinkers with a biotin moiety that allowed only crosslinked peptides to bind to the avidin column while the non-crosslinked peptides were removed from the sample. The crosslinkers were successfully tested on the heterodimeric transcription repressor, NC2, in which the complex structure was already solved. Although protein-protein interaction was the focus of their research, these tri-functional crosslinkers can also be applied to protein structure analysis and tri-functional crosslinkers can be synthesized with specific reactive or chemical properties.

4 Computational analysis of crosslinking data

4.1 Introduction

One major goal of this project was to develop methods for generating experimental data to support protein structure models. To assess the experimental data generated by crosslinking experiments, a computational approach was used. The Automatic Spectrum Assignment Program (ASAP) was developed by Young et al. (2000) for their analysis of MS data generated by crosslinked FGF-2. This program requires a list of m/z values and their corresponding charges for each crosslinked peptide peak. It also requires information pertaining to the protein and the crosslinkers, such as the protein sequence and properties of the crosslinker. Based on this information, ASAP creates a virtual crosslinked peptide library and it searches within this library for the experimental masses that fall within the error limit. These experimental peaks are given putative crosslinking assignments which can be used for low resolution model validation.

Solved protein structures or reliable models are required for this method development. TKL1 has a solved crystal structure (Nikkola et al. 1994) while models were generated for IPP1, HIS7, PGK1 and ENO1 using the software program, SWISS-MODEL. SWISS-MODEL is an automated homology-modeling server, which is publicly available on the web (Perkins et al. 1999) Based on the given sequences, SWISS-MODEL automatically selected five protein templates and it can produce highly accurate models if the sequence identity is greater than 95% between the target and structure template (Schwede et al. 2003). Given these protein structures, putative crosslinking sites assigned by ASAP can be verified.

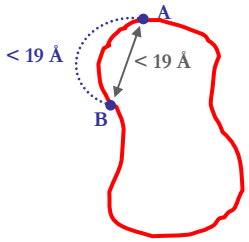
To verify the crosslinking sites, two distance criteria must be met. The first is that the distance between conjugated lysines must be less than the maximum crosslinking distance. The crosslinking distance is calculated by adding the length of the spacer arm to the distances between the lysine's terminal N^{ζ} atom and the C^{α} atom for each reactive group. Young et al. (2000) used the maximum span of 24 Å in their research with crosslinker BS^3. However, Green et al. (2001) argued that it is highly improbable that the crosslinker will be in a fully extended conformation, stretching to its maximum length. Ye et al. (2004) have determined that using a crosslinking distance of 19 Å resulted in a higher confidence level in their protein model validation exercise. In addition to the spatial distance constraint, a second distance constraint is placed on the protein sequence where the crosslinked lysines must be at least 7 amino acids apart. The reason for this is that the crosslinker

BS^3 can span up to 7 amino acids when it is fully extended and thus residue pairs less than 7 amino acids apart will not provide useful information.

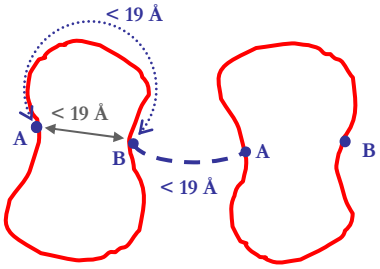
One challenge in this project is to eliminate the false positives from both the data sets by the comparison between the experimental data and predicted crosslinking sites. The predicted crosslinking data set is generated based on the distance constraints using protein structures and models. Following the trend of current crosslinking research, Euclidian distances are calculated for the two crosslinked lysines (Dihazi and Sinz 2003; Kruppa et al. 2003; Pearson et al. 2002; Schilling et al. 2003; Young et al. 2000), where in reality, the crosslinker path is along the surface of the protein. Figure 4-1 shows five possible crosslinking scenarios between residue A and B. The ideal situation (scenario 1) is where the straight distance and the surface distance are both less than 19 Å; this will be classified as a true positive result. Scenario 2 shows an intermolecular crosslinking that will produce the same result as scenario 1 since both straight-through and surface distances are both less than 19 Å. Although this is a false positive result, it will not affect the outcome of the experiment and therefore tolerated in this analysis. Scenario 3 presents crosslinking events where crosslinked peptides can be identified by ASAP but will be eliminated as a false positive by the comparison with the predicted crosslinking data because the straight-line distance between A and B is greater than 19 Å. In scenario 4, any inaccessible lysine pairs present in the predicted list will not be assigned by ASAP and are eliminated as false positives. An instance in which a false positive cannot be filtered is shown in scenario 5. The straight-line distance between two amino acids is less than 19 Å and therefore is recorded in the predicted data set but an intermolecular crosslinking event will produce a crosslinking product that can be detected by ASAP. However, this is a false positive result because of the crosslinking distance on the surface of the protein is greater than 19 Å. Currently, this method is not capable of eliminating this false positive and thus this crosslinking pair will be recorded as a true crosslinking result. Note that the occurrences of scenarios 2, 3 and 5 would be relatively low since the experimental conditions were designed to minimize the number of inter-molecular cross-links.

In this chapter, MS data from the crosslinking experiments with BS3 was analyzed by ASAP. The putative crosslinking assignments were verified against predicted crosslinking data. False positives were eliminated from both the experimental and the predicted data set. The set of valid crosslinking sites for each target proteins will be used in the next chapter for model validation.

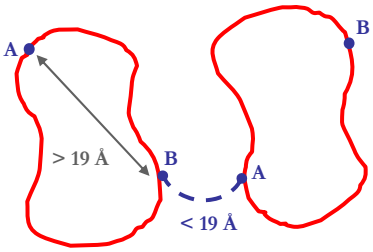
Scenario 1: True positive
Both straight and surface distances are less than 19 Å.



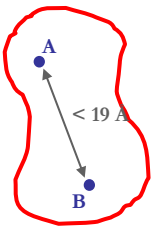
Scenario 2: False positive (tolerated)
This intermolecular crosslinking event is tolerated since it will not affect the outcome of the crosslinking analysis.



Scenario 3: False positive
This intermolecular crosslinking event will be eliminated by the predicted crosslinking data set since the straight distance is greater than 19 Å.



Scenario 4: False positive
Crosslinking between internal residues will not be detected by ASAP.



Scenario 5: False positive
This intermolecular crosslinking event cannot be eliminated by this analysis and may skew the final results.

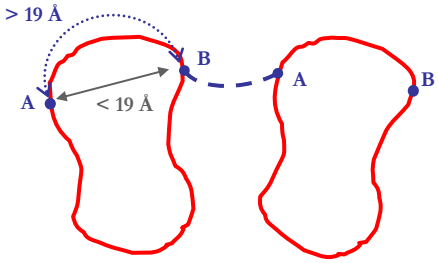


Figure 4-1. Illustration of five potential crosslinking events between residue A and B. The straight arrow (grey) indicates the Euclidian distance between A and B. BS^3 crosslinking is represented by the blue dashed line with dots on the ends. The blue dotted lines with arrowheads on the ends represent the surface path of the crosslinker connecting A and B.

4.2 Materials and methods

4.2.1 Automated spectrum assignment program

To analyze the crosslinking data acquired from the MS spectra, the following information was submitted to ASAP: the sequence of the crosslinked protein, the list of peaks unique to the crosslinked MS spectrum, the molecular weight of BS^3 (138.08373 Da), amino acid specificity of BS^3 (lysine), protease used (trypsin), number of missed cleavages (2) and maximum allowed error (0.1 Da). The ASAP displayed all plausible assignments for the crosslinked peptides. From the output of ASAP, pairs of potential crosslinked lysines can be identified by their position number in the protein sequence.

4.2.2 Structural model generation

Protein models were generated using SWISS-MODEL (Schwede et al. 2003), with the exception of TKL1 which had a corresponding PDB file. The First Approach mode was used to generate models for IPP1, HIS7, PGK1 and ENO1. Each protein sequence was submitted to SWISS-MODEL with no template specified. SWISS-MODEL returned the protein models and their corresponding templates in PDB file format. Protein-protein BLAST was used to determine the sequence identity between the targets and their respective templates (Altschul et al. 1997). Due to high sequence identity between the targets and their respective templates, these models can be assumed to represent accurate structures for these target proteins. In addition, each model was visually inspected against their respective templates using SWISS-PDB Viewer (Guex and Peitsch 1997) to ensure the quality of the model.

4.2.3 Identifying potential crosslinking sites on target proteins

The perl program *LinkLys* was written to predict potential crosslinking sites given a protein structure model and to generate the predicted crosslinking data set. Using the coordinates provided in the model PDB file, this program calculates the C^{α}- C^{α} distance between all pairs of lysine residues in the protein and lists pairs whose distances were less than 19 Å in 3D space but greater than 7 amino acids apart in sequence.

To determine which pairs of lysines are consistent with both the model and the MS data pairs, *Cmp_crosslinks* was written to compare the list of potential crosslinking sites obtained from ASAP with the predicted data set generated by *LinkLys*. All lysine pairs that appeared on both lists were

considered experimentally validated crosslinking sites. The output of this program contained the following information: target PDB ID, CATH code for the target protein and a list of crosslinking sites.

4.2.4 Visualization of crosslinked protein structures

PyMOL and SWISS-PDB Viewer were used for visualization of the protein structure and models (DeLano 2002; Schwede et al. 2003). All figures of protein structures in this thesis were generated by PyMOL.

4.3 Results

4.3.1 Protein structures and models

The crystal structure of TKL1 was solved at 2.0 Å resolution by Nikkola et al. in 1994. The structures for the other target proteins were obtained using the fully automated modeling software, SWISS-MODEL. To ensure the quality of these models, the sequence identity between targets and their respective templates were verified using protein-protein BLAST (Altschul et al. 1997). In each case, the sequence identities between the target and one or more templates used were greater than 99%, indicating that the models and the target structure share the same fold and are highly similar. All template structures were solved by x-ray crystallography. Table 4-1 showed a summary of the templates used for modeling the target structures.

Table 4-1. Sequence identity between target and template sequences.

Target Protein	Protein templates used for model generation (% identity)
IPP1	1M38A,B (100%); 1E6AA, 1E9GA, 1YPPA (99%)
HIS7	1OX4A, 1OX6B (100%); 1OX4B, 1JVNA,B (99%)
PGK1	1FW8A (100%); 1QPG (99%); 3PGK (97%); 1VJDA, 1VJCA (65%)
ENO1	1EBGA,B, 2ONEB, 1ONEA,B (99%)

4.3.2 Structural description of target proteins

TKL1 had an available crystal structure refined to 2.0 Å resolution. TKL1 is a homodimer and the structure of a subunit contains three domains with 42% of residues being α-helices and 12% of residues being β-sheets (Figure 4-2a). The N-terminal domain is the largest domain and it consists of approximately half of the subunit. Five-stranded parallel β-sheets form the core of the domain and are surrounded by clusters of α-helices. The middle domain is the second largest domain with six parallel β-sheets at the center and surrounded by α-helices. The C-terminal domain is the smallest of the three and it is made up of four parallel plus one anti-parallel β-strands. These β-sheets are also surrounded by several α-helices (Nikkola et al. 1994).

SWISS-MODEL selected 1M38 chain A and B as the templates for homology modeling of IPP1. The fold of the 1M38 subunits is conserved among other pyrophosphatases with known structures (Harutyunyan et al. 1996). Both chains have 100% sequence identity to the target and each chain represents a 32 kDa subunit. IPP1 was modeled to a single subunit with a compact globular shape (Figure 4-2b). A 4-stranded β-barrel is located at the core of the protein and is

surrounded by two long and one short helix on the surface of the protein. A second part of the structure is made of two long and two short anti-parallel β-strands, surrounded by 2 short helices at the C-terminus of the protein. The two regions do not form separate domains since they are not packed tightly together.

The model of HIS7 consists of a C-terminal and a N-terminal domain and the components of each domain are packed tightly together (Figure 4-2c). The N-terminal domain contains a series of β-strands at the core, five parallel sheets followed by four anti-parallel strands. The four anti-parallel β-strands loosely formed a structure resembling a barrel. Also in the core, a short α-helix is located approximately between the parallel and anti-parallel sheets. This core region is surrounded by 6 helices on the surface. The center of the C-terminal domain consists of a 7-stranded β-barrel which is surrounded by 9 α-helices. Two short anti-parallel β-strands, between β1 and β2, extend from the barrel and onto the surface. The length of β7 is slightly longer than the other β-strands and it protrudes from the core, forming a single short anti-parallel segment which in turn is connected to an α-helix on the surface of the structure. On the opposite side, three strands of anti-parallel β-sheets and one helix are located on one side of the barrel.

The model of PGK1 shows two distinct globular shaped domains with similar architecture (Figure 4-2d). The N-terminal domain contains six parallel β-strands at the center and surrounded by four helices on the surface. Additionally, two anti-parallel β-strands protrude onto the surface of the protein, towards the C-terminal domain. The C-terminal domain is made up of seven α-helices on the surface, surrounding the five parallel β-sheets at the core. Three anti-parallel β-strands are located on the surface of the protein, away from the N-terminal domain. The region between the two domains is occupied by two α-helices: one α-helix is connecting the two domains together while the other is an α-helix from the N-terminus, traversing across this region from the C- to the N-terminal domain.

The model of ENO1 shows no distinctive domains but it can be described as two regions, joined together by a short α-helix (Figure 4-2e). The smaller region consists of three α-helices and three anti-parallel β-strands. The larger region contains an 8-stranded β-barrel structure which is enclosed by 8 α-helices. Additionally, two short parallel β-strands are located on the surface of the larger region.

4.3.3 MS data analysis using ASAP and *LinkLys*

Two steps were involved in identifying valid crosslinked peptides from the MS results. The first step was to assign sequence information to each of the mass peaks using ASAP and the second step was to compare the ASAP crosslinking sites with those generated from structural models using *LinkLys*. The MS peaks present in the crosslinked peptide spectra but absent from the non-crosslinked spectra were input into ASAP. Each pair of putative crosslinking sites was labeled according to their lysine positions within the target sequence; this represented the experimental data set. A second set of crosslinking sites was generated by *LinkLys* using solved or model structures of the protein targets; this provided a predicted data set for each model. This list included all pairs of lysines that are less than 19 Å in space and greater than 7 amino acids apart in sequence. Since BS^3 can span up to 7 amino acids, anything less than 7 amino acids in length would not provide any useful crosslinking information. For example, crosslinked lysine 228-234 in Figure 3-6a would be excluded from this predicted data set because lysine 228 and 234 are less than 7 amino acids apart.

The experimental and predicted data sets were compared and any matches between the lists were considered as validated assignments for chemically crosslinked lysines, providing experimental evidence to support the model structures. Table 4-2 summarizes the results of crosslinking MS data analysis by ASAP and *LinkLys*. Details of crosslinking assignments by ASAP can be found in Appendix B. From these results, it appeared the number of putative crosslinking site assignments by ASAP increases with the number of peaks input into the program. Also, crosslinking site prediction by *LinkLys* had resulted in a large number of possibilities and only a fraction of these predicted sites were validated by experimental data. In the example of TKL1, 216 peaks were found to be unique to the crosslinking spectra but ASAP only identified 100 of these peaks as putative crosslinking sites. *LinkLys* predicted 292 potential sites from the TKL1 structure but only 8 lysine pairs were experimentally confirmed. Similar results were found for the other four target proteins.

Table 4-3 presented the complete list of crosslinking sites identified by residue number using this method. TKL1 and IPP1 both resulted in the highest number of 8 crosslinking sites identified among the five target proteins. PGK1 contained only 1 crosslinking site, the least number of sites identified. It is expected that one would see the number of crosslinking sites increase with the number of lysine residues in the protein sequence. However, the results showed no clear relationship between these two factors. For example, TKL1 contained 8 crosslinking sites while PGK1 contained only 1 but their lysine content was approximately the same. Eight crosslinks were verified for IPP1 but the sequence contained the least number of lysine residues. This process of

comparing experimental against predicted data had eliminated a large number of false positives from both data sets.

Table 4-2. Summary of MS data analysis by ASAP and *LinkLys*.

Protein Name	Number of peaks submitted to ASAP	Number of putative crosslinking sites assigned by ASAP	Number of possible crosslinking sites predicted by *LinkLys*	Number of validated crosslinking sites
TKL1	216	100	173	8
IPP1	168	64	83	8
HIS7	43	16	109	4
PGK1	43	13	104	1
ENO1	112	46	76	5

4.3.4 Visualizing crosslinker BS^3 on target proteins

The structural models of each target protein were visualized using PyMOL (Figure 4-1). The lysine residues and their residue labels were highlighted for a selected number of crosslinking sites. The dotted lines represented the chemical crosslink formed by BS^3, connecting pairs of crosslinking sites identified by both ASAP and *LinkLys*.

Table 4-3. Crosslinked lysine pairs present in both experimental and predicted data set.

PDB ID	Length of Protein Sequence	Number of Lysines	Crosslinked Lysine Pairs	Crosslinking Distances (Å)
TKL1	680	43	9/276	13.937
			9/278	12.828
			303/311	12.386
			303/314	16.867
			311/321	16.028
			311/322	16.722
			345/392	17.080
			582/671	18.214
IPP1	287	29	11/74	17.072
			17/74	17.084
			57/199	16.007
			57/268	18.662
			77/194	15.885
			168/177	9.828
			177/211	15.552
			199/279	17.603
HIS7	552	43	172/546	15.074
			199/210	16.037
			258/441	18.777
			419/432	13.658
PGK1	416	42	244/258	10.774
ENO1	437	37	5/28	12.216
			5/85	16.341
			54/67	18.504
			56/67	16.073
			178/241	9.916

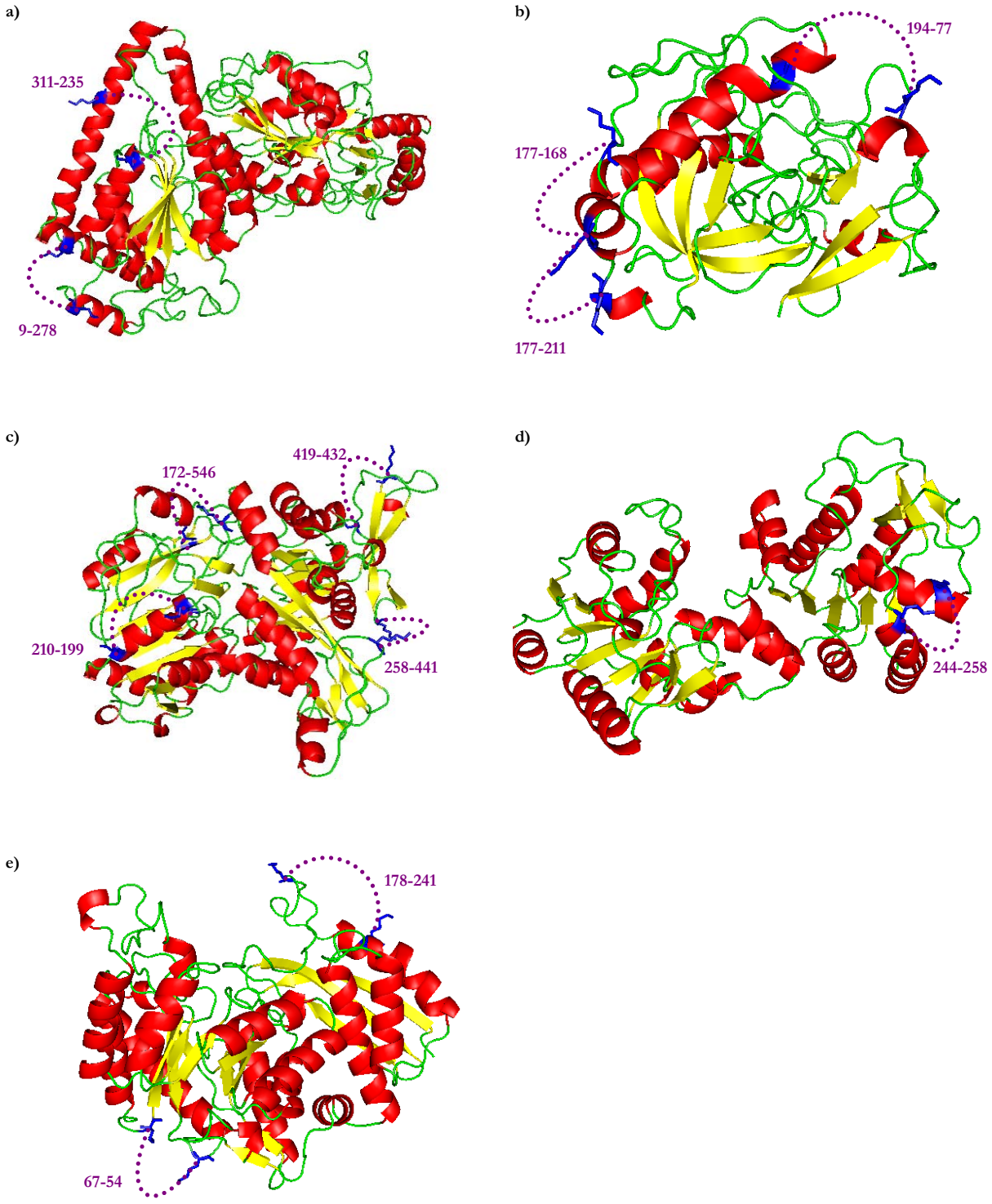


Figure 4-2. Structures of target proteins with selected BS^3 crosslinking sites. a) IPP1, b) HIS7, c) PGK1, d) ENO1 and e) TKL1.

4.4 Discussion

4.4.1 Assessment of the computational approach to data analysis

Both the experimental and computationally derived data sets were expected to contain a high rate of false positives, as observed in the results. This can be attributed to several factors. First of all, the peptide mixture entering the mass spectrometer may contain products from both inter- and intramolecular crosslinking; it was not possible to distinguish the mass peaks for inter- and intramolecular crosslinked peptides by visual comparison of the spectra. Secondly, ASAP assignments were obtained by matching peptide fragments with the mass values; it is possible that more than one crosslinked peptide pair can be assigned to some mass values. Both of these can contribute to the presence of false positive assignment by ASAP. Thirdly, the *LinkLys* predicted list was generated by determining the C^{α}-C^{α} distances between pairs of lysine residues in the protein structure and the program did not discriminate against inaccessible lysine residues. Although this is not expected to have large impact on the data analysis, as very few lysines are located in the core of the protein and they are physically inaccessible to BS^3, a method eliminating these internal lysines should improve the accuracy of the data set. More significantly, *LinkLys* determines Euclidian distance between two residues as a straight line through the protein; ideally, the program would calculate the distance of the actual crosslinking path which travels across the surface of the protein. This surface distance will always be greater than the Euclidian distance between two points and by incorporating this additional piece of distance information into the program, the accuracy of the predicted data set can be further enhanced.

Potluri et al. (2004) developed algorithms for analyzing the crosslinking data for the purpose of protein model discrimination and their work addressed the concerns with internal lysines and surface distances. Their algorithms identified the surface residues by assessing the residue's solvent accessibility (Lee and Richards 1971). Then they find the shortest path between two specific surface residues traveling on the surface of the protein, taking into account the surface geometry of the protein by providing an upper and lower bound on the distances for each crosslinking path. They compared their result with the results published by Young et al. (2000) and found that a distance calculation using C^{β} was better than using C^{α}. Moreover, they also found their method performed better in model discrimination than Young *et al.* While this result in an improvement over the Euclidian distance, the algorithms presented by Potluri et al. (2004) would be more difficult to implement. Furthermore, high quality models are necessary for the success of this approach. The

computation of upper and lower bounds for each crosslinking path requires accurate atom positions, which may not be feasible for constructed protein models.

4.4.2 Improving the quality of crosslinking data

The reduction of false positives in the experimental and predicted crosslinking data sets may enhance the effectiveness of this method. Changes can be made in the experimental protocol to improve the quality of the crosslinking data.

Intuition suggests that the higher number of lysines available on the protein sequence should result in a higher number of crosslinking sites; however, the results showed no significant trends between these two parameters. The IPP1 sequence contained the least number of lysines but had one of the highest number of crosslinking sites while PGK1 showed the opposite trend. These results implied that the abundance of lysines cannot be used as a clear indication of the degree of crosslinking. Recall that all proteins were crosslinked under the same conditions, in terms of protein to BS^3 ratio, concentration of both protein and BS^3 as well as the reaction time. It is possible that the experimental conditions used in crosslinking were not optimal for PGK1 and the results for PGK1 would be improved if a higher concentration of BS^3 or longer reaction time was used. Such alteration to the crosslinking conditions would likely increase the number of crosslinking sites detected by this method.

Chapter 3 (Section 3.4.3) of this thesis described two alternatives for improving the quality of the crosslinking data. Preliminary MS analysis immediately after chemical crosslinking can provide feedback on the extent of crosslinking on the protein. Once the optimal crosslinking condition is established, size exclusion chromatography can be used to eliminate intermolecular crosslinked proteins. The use of isotopically labeled or tri-functional crosslinkers will also improve the detection of crosslinked peptides from mass spectra.

In addition to technical alternatives, careful planning of the crosslinking strategy can also increase the efficiency of the experiment and Ye et al. (2004) proposed a probabilistic analysis for finding the optimal crosslinking condition to produce experimental data for supporting or discriminating against protein models. They developed an algorithm called XlinkPlan that can determine the best experimental strategies to enhance the quality of the crosslinking data prior to executing the experiment. This planning approach would be particularly useful when only small amounts of the proteins or crosslinkers are available. XlinkPlan can help select the optimal crosslinking condition without wasting precious materials.

5 Model validation and fold recognition

5.1 Introduction

One application for chemical crosslinking is structural model validation. Similar to the methodology in Chapter 4, predicted crosslinking sites can be generated from any given protein model and can be compared to the experimental data. Predicted crosslinking sites consistent with the experimentally observed sites can be used to support the model structure. Similar studies have been done using ubiquitin with homobifunctional crosslinkers (Kruppa et al. 2003) and cytochrome c and ribonuclease A using isotopically labeled crosslinkers (Pearson et al. 2002). However, their research focused on improving the crosslinking/MS methodology for structural validation. The aim of this project is to determine whether the experimental data can be used to increase the confidence of the structural models and whether the distance constraints can be used to identify the best model.

A large number of models can be generated by homology modeling, where models are constructed by fitting a target sequence to template structures. The protein modeling tool used for generating protein models in this chapter is RApid Protein Threading by Operation Research technique (RAPTOR). Given a target protein sequence, RAPTOR will construct a sequence profile using multiple sequence alignments with sequence homologues. This sequence profile will then be aligned to each structure in the template library via protein threading. One of the protein threading methods available in RAPTOR is the NP-core (non-pairwise) algorithm. In this method, each template is treated as a series of core structures, represented by conserved regions of α-helices and β-sheets, connected together by loops. NP-core only considers interactions between the core residues for the alignment between the sequence profile and the templates and the best alignment can be obtained by optimizing the scoring function. RAPTOR employs a support vector machine (SVM) to compare each alignment against the rest of the alignment data set. Z-scores are then computed for each alignment and RAPTOR ranks the alignment by their z-score values (Xu and Li 2003; Xu et al. 2003).

Normally, a protein model is built by superimposing the backbone of the target onto the template structure provided by the alignment. Manual refinement is then required to optimize the fit between the sequence and the structure to accommodate the loops and side-chains (Bourne and Weissig 2003). However, RAPTOR will generate 2,983 sequence-structure alignments for each target protein and close to 15,000 models will need to be generated all together for this project. It is

not feasible to generate models for a data set of this magnitude. Therefore, the RAPTOR alignments are treated as protein models in this project and they can be used directly for crosslinking site prediction. These models are the superposition of the alignments onto the template structure and crosslinking information can be extracted by using the C^{α} positions of the crosslinked residues from the template structures.

In addition to model validation and selection, chemical crosslinking may also be used for fold recognition. Chemical crosslinkers connecting surface protein residues can generate distance constraints between the residues and thus greatly limit the number of theoretical folds. This approach of using chemical crosslinking for fold recognition has been shown to be successful. Young et al. (2000) used a sequence threading program to select the top 20 structural models for the protein FGF-2 from a set of 635 templates sharing less than 30% sequence identity. FGF-2 clearly belongs to the β-trefoil family and without the distance information from the crosslinking experiments, the sequence threading algorithm would have placed FGF-2 in the β-clip fold family. In this project, the fold recognition capability will be tested against a fold library containing 2,983 structural models. Models containing the highest number of experimentally observed crosslinking sites will be considered to be the correct model.

5.2 Materials and methods

5.2.1 Model generation by protein threading

RAPTOR (version 3.0) contains over 6000 non-redundant structural files within its template library. From these, structures with a CATH Hierarchical Classification (Orengo et al. 1997) were selected for protein threading. A list of CATH classified PDB codes (v.2.6.0, released April 2005) were first cross-referenced with the RAPTOR structure library. Then, templates with missing structural regions in the PDB files were excluded from the template set. The resulting template library contained 2938 structures. Alignments between the five target sequences (TKL1, IPP1, HIS7, PGK1 and ENO1) and the template database were generated using RAPTOR using the non-pairwise core threading algorithm. The RAPTOR output included all sequence-structure alignments (*.pir) as well as files (*.scoreRank) containing the z-scores calculated for each alignment. The z-score is a measure of the quality of the alignment between the target sequence to the template structure and it provides a means to rank each alignment for a given target protein. Only the top 100 ranking alignments were used for model validation.

The protein models are the superposition of the RAPTOR generated alignments onto the template structures. The prediction of the models' crosslinking sites utilized the C^{α} positions of the crosslinked residues from the template structures. This eliminated the need to construct detailed three-dimensional models for each sequence-structure alignment for crosslinking site prediction.

5.2.2 Crosslinking site prediction algorithm for protein models

The algorithm for the virtual crosslinking of the models and the prediction of the potential crosslinking sites is illustrated in Figure 5-1. A perl program called *ParsePIR* was built to parse the alignment output generated by RAPTOR. This program located each lysine residue in the target sequence and found the corresponding residue in the template sequence (Figure 5-1, Step 1). If the target residue corresponds to a gap in the alignment, then this position was omitted from the output. The output of this program contains the following information (Figure 5-1, Step 2): model identifier (template PDB ID), template sequence, lysine positions in the target sequence, template residue names and numbers corresponding to the lysine positions.

The second program, *LinkModel*, predicted potential crosslinking sites on the protein model using the target-template alignment. Using the output from *ParsePIR* (Figure 5-1, Step 2), this program extracted the template positions and residue names and then found the coordinates for

each residue from the template PDB file (Figure 5-1, Step 3). The method for calculating the distance between each pair of residues was the same as *LinkLys*, described in Chapter 3 of this thesis. Once potential crosslinking sites were identified, each template position was mapped back to the original target lysines in the model (Figure 5-1, Step 4 and 5). The output contained the following information: model identifier, position numbers and distance for each pair of crosslinked lysines. The list of potential crosslinking sites generated by *LinkModel* was equivalent to the output from *LinkLys* used to identify crosslinking sites on target structures.

Step 1: Identify all template residues that correspond to the lysine positions in the target sequence, IPP1.

```
>P1;template
MGLISDA---------DKKVIKEEFFSKM-----------VNP-VKLIVFVRKDHCQYCDQ---LKQLVQELSELTDK
LSYEIVDF---------DTPEGKELAKRYRI---------------DRAP-ATTITQ----DG-------KDFGV
...

>P1;target IPP1
MTYTTRQIGAKNTLEYKVYIEKDGKPVSAFHDIPLYADKENNIFNMVVEIPRWTNAKLEITKEETLNPIIQDTKKGK
LRFVRNCFPHHGYIHN----------YGAFPQTWEDPNVSHPETKAVGDNDPIDVLEIGETIAYTGQVKQVKALGIM
...
```

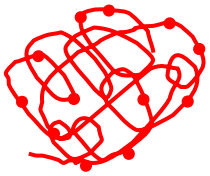
Step 2: Generate *ParsePIR* output.

```
Template Structure: 1a8l
Template Sequence: MGLISDADKKVIKEE...FLEKLLSALS
Template Residue Positions:
   10  15  18  39  41  53  54  56  89  107 ...
Template Residues:
   K   E   S   D   L   D   K   S   K   F   ...
Target K Positions:
   17  22  25  57  62  74  75  77  139 168 ...
```

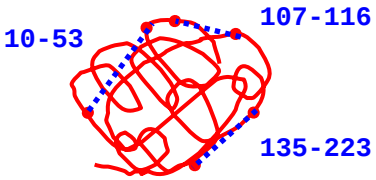
Step 3: Determine the x, y, z coordinates of the template residues from the template PDB file.

```
Template Positions:
   10  15  18  39  41  53  54  56  89  107 ...
Template Name:
   K   E   S   D   L   D   K   S   K   F   ...
```

+



Step 4: Locate all potential crosslinking sites on the template structure.



Step 5: Map the template positions for each crosslinked pairs to the original target lysine positions on the structural model.

Crosslinked residues on template (alignment positions):		Crosslinked lysines on target:
Lys10 – Asp53	⟹	Lys17 – Lys74
Phe107 – Arg116		Lys168 – Lys177
Asp135 – Ser223		Lys199 – Lys279

Figure 5-1. Workflow of the *ParsePIR* and *LinkModel* algorithms for predicting crosslinking sites on protein models. The IPP1-1A8L sequence-template alignment is used for the sample illustration.

5.2.3 Protein model validation using crosslinking data

The predicted data set output by *LinkModel* was compared to the experimental data obtained from chemical crosslinking experiments, as described in the Methods section of Chapter 4. The *Cmp_crosslinks* program, also from Chapter 4, was used to find the predicted crosslinking data from the RAPTOR models to the experimental crosslinking data. The output of *Cmp_crosslinks* includes the following information: model identifier, CATH code for the template used for the modeling, the RAPTOR z-score for the alignment and the list of crosslinking sites. All experimental crosslinking sites matched to the predicted crosslinking sites can provide evidence to support that particular structural model.

5.3 Results

5.3.1 Protein threading by RAPTOR

RAPTOR successfully generated 2983 structural models (sequence-template alignments) for each target protein. Each model was contained in its own individual file. One *.scoreRank* file was generated for all alignments and its contents were sorted by the z-score.

5.3.2 Identifying crosslinking sites for each alignment

The first step of model validation involved the prediction of crosslinking sites on the structural model. *LinkModel* found less than 30% of the models contained predicted crosslinking sites consistent with experimental data (Table 5-1). There appears to be no consistent relation between the length of the protein sequence, the number of lysines present in the sequence and the percentage of models with validated crosslinking sites.

Table 5-1. Summary of model validation for each target protein.

PDB ID	Length of protein sequence	Number of lysines	Total number of models with experimentally observed crosslinking sites	Percentage of models with experimentally observed crosslinking sites
TKL1	680	43	494	16.56%
IPP1	287	29	888	29.77%
HIS7	552	43	581	19.48%
PGK1	416	42	443	14.85%
ENO1	437	37	544	18.24%

In Table 5-2, the total number of validated models was divided into subcategories where each subcategory indicating the number of models with *n* number of experimentally observed crosslinking sites. It was expected to see that the higher number of experimental crosslinking sites would yield a higher number of models with corresponding crosslinking sites by increasing the opportunities of matching with experimental data. For example, IPP1 has a high number of experimental crosslinking sites at 8 and 888 models were consistent with one or more crosslinking sites. Furthermore, HIS7 has 4 experimental crosslinking sites and there are only 581 models with validated crosslinking sites. TKL1 however does not share this pattern; it has the highest number of experimental crosslinking sites at 8 but only 494 models have experimentally observed crosslinking

sites. PGK1 has only one experimental crosslinking site, the lowest of all protein targets, but the number of consistent models is 443. These results suggest that the number of experimentally observed crosslinking sites is not a good indicator of the number of models expected to be consistent with the experimental data.

Table 5-2. Number of protein models supported by crosslinking data. The total number of models with experimentally observed crosslinking information was counted for each target protein. Each column (n = 0 to 8) listed the number of models containing *n* observed crosslinking sites.

PDB ID	Number of experimentally observed crosslinking sites	Models consistent with ≥ 1 experimental crosslinking sites	Number of models matching *n* experimentally observed crosslinking sites								
			8	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**
TKL1	8	494	1	0	1	0	110	40	188	154	2217
IPP1	8	888	1	0	0	0	5	34	215	633	2095
HIS7	4	581	0	0	0	0	2	1	64	514	2402
PGK1	1	443	0	0	0	0	0	0	0	433	2540
ENO1	5	544	0	0	0	0	2	16	128	398	2439

5.3.3 Experimental support for protein structural models

RAPTOR assigns a z-score to each model to indicate the quality of the fit between the target sequence and the template structure. The models were ranked by the z-score, with the best fit represented by the highest value (Xu and Li 2003; Xu et al. 2003). Ideally, if a model with the correct fold is identified, the CATH code for the top scoring model should be the same as the target protein. Appendix C shows a complete list of the top 100 RAPTOR models based on z-score ranking. Only the top 100 models were considered for model validation because they were considered to be the best sequence-template fit by RAPTOR out of the 2983 models available. The top ranking model for each target protein is shown in Table 5-3. RAPTOR successfully identified the correct protein fold for IPP1, HIS7 and ENO1, based on their CATH codes assigned to the template structure. These models were supported by 8, 1 and 2 experimentally observed crosslinking sites, respectively. RAPTOR had also generated a model for TKL1 with the correct fold for one of its two domains, but no experimental data supported this model. Finally, the fold of the top PGK1 model was incorrect as indicated by the CATH code; further examination of the data

showed that no models in the final data set had the correct CATH code 3.40.50.1260. As expected, no crosslinking data was found to support this model.

Only a small group of the top 100 RAPTOR models had supporting experimental evidence. Table 5-4 shows the top ranking models that are supported by experimentally observed crosslinking sites for each target protein. In all cases, these models had the same fold as the target proteins, as indicated by the CATH codes, and each was ranked within the top 5 of the top 100 models. Models for TKL1 and IPP1 were both strongly supported by 8 validated crosslinking sites. Experimental evidence for the HIS7 and ENO1 models were comparably weaker with only 1 and 2 validated crosslinking sites, respectively. The model for PGK1 had the correct fold, down to the level of Topology (3.40.50.x) and it was supported by only one validated crosslinking site.

Table 5-3. Top RAPTOR models for each target protein. All models have the correct fold based on the CATH code assignment.

PDB ID	CATH code	Model ID	Model CATH code	Template Sequence Length (target)	% Identity	RAPTOR z-score	Number of crosslinks
TKL1	3.40.50.920 3.40.50.970	1IK6	3.40.50.920 3.40.50.970	284 (680)	22.9%	197.23	0
IPP1	3.90.80.10	1M38	3.90.80.10	287 (287)	100%	298.14	8
HIS7	3.20.20.70 3.40.50.880	1KA9	3.20.20.70 3.40.50.880	201 (552)	35.8%	204.65	1
PGK1	3.40.50.1260	1BRL	3.20.20.30	252 (416)	22.2%	78.91	0
ENO1	3.30.390.10 3.20.20.120	1JPM	3.30.390.10 3.20.20.120	359 (437)	19.2%	218.65	2

Table 5-4. Top RAPTOR models supported by one or more validated crosslinking sites.

PDB ID	Model ID	Model CATH code	Template sequence length (target)	% Identity	RAPTOR z-score	Ranking by RAPTOR	Number of crosslinks
TKL1	1AY0	3.40.50.920 3.40.50.970	678 (680)	99.9%	163.84	2	8
IPP1	1M38	3.90.80.10	287 (287)	100%	298.14	1	8
HIS7	1KA9	3.20.20.70 3.40.50.880	201 (552)	35.8%	204.64	1	1
PGK1	1BIF	3.40.50.300 3.40.50.1240	432 (416)	19.5%	76.27	5	1
ENO1	1JPM	3.30.390.10 3.20.20.120	359 (437)	19.2%	218.65	1	2

5.3.4 Case-by-case analysis of protein model validation

It was expected that models with the same fold would have a similar number of consistent crosslinking sites, while models with different folds would have fewer consistent crosslinking sites. However, the outcome from this project suggests that other factors, such as the length of the target and template sequences and their sequence identity, may influence the results of model validation by chemical crosslinking experiments. The following sections will discuss the results for each target protein in more detail.

5.3.4.1 Validation of TKL1 models

TKL1 has two domains, differing at the Homologous superfamily level of the CATH hierarchy. RAPTOR had ranked a model based on only one of the TKL1 domains (3.40.50.920) as the top result. 1IK6 was RAPTOR's best model for TKL1 but it was not supported by the experimental data, while 1AY0 was RAPTOR's 2nd best model and it was supported by 8 experimentally observed crosslinking sites. One possible cause for this is the relative sequence length between TKL1 and the templates. Template IIK6 has a sequence identity of 22.9% with TKL1 and was 248 amino acids long while 1AY0 is 678 amino acids long and has a sequence identity with TKL1 of 99.9%. Thus the model based on 1AY0 produced the desired result of a good protein model strongly supported by experimental crosslinking data, increasing the confidence of this model.

5.3.4.2 Validation of IPP1 models

Of the five target proteins, IPP1 produced the highest quality results. The top three RAPTOR models for IPP1 all had the correct structural fold (3.90.80.10). The 1M38 model was ranked 1st and was supported by 8 experimental crosslinking sites. This structural template is 282 amino acids long and has a 100% sequence identity with the target sequence. 1UDE and 2PRD were the 2nd and 3rd ranking RAPTOR models. Their sequence lengths are 168 and 174 and had sequence identity with the target of 26.1% and 25.0%, respectively. Although these models were highly ranked by RAPTOR, they were only supported by 2 and 3 experimentally observed crosslinking sites.

5.3.4.3 Validation of HIS7 models

The top RAPTOR model for HIS7 was 1KA9. The template 1KA9 was only 201 amino acids long and has a sequence identity of 35.8% with the target. Since this model had only 1 consistent crosslinking site, there is not enough experimental data to support or refute the protein models with confidence for both domains. More experimentally observed crosslinking sites may increase the number of crosslinking sites per model and improve the overall confidence of the models.

5.3.4.4 Validation of PGK1 models

There were no models in the data set that exactly matched the fold of the target structure, therefore, models matching to the Topology level were considered in the analysis (CATH code: 3.40.50.x). There are 30 models under this CATH code category in the top 100 results and only 11 have supporting evidence with one experimentally observed crosslinking site. Model 1BIF was ranked 5th by RAPTOR and this model is slightly supported by one validated crosslinking site. PGK1 has only one predicted crosslinking site consistent with the experimental data, which is likely insufficient to provide support for any given model.

5.3.4.5 Validation of ENO1 models

1JPM was the top ranking RAPTOR model for ENO1. The template 1JPM is 359 amino acids long and a sequence identity of 19.2% with the target. This model had 3 predicted crosslinking sites that are consistent with the experimental data. The 2nd and 3rd ranking RAPTOR models also have the correct fold but each are weakly supported by one experimentally observed crosslinking site. However, there are insufficient data to support or refute the protein models.

5.3.5 Enhancement of model selection using crosslinking data

In the case of TKL1 and PGK1, the results for model selection were improved by incorporating the crosslinking data into the analysis. RAPTOR rankings had placed the models 1IK6 and 1AY0 in 1st and 2nd place for TKL1. Both models have the correct fold but 1IK6 has no experimental support while 1AY0 is supported by 8 experimentally observed crosslinking sites. However, 1AY0 is a much better model because the length of the template 1AY0 (678 amino acids) is closer to the target (680 amino acids) compared to 1IK6 (284 amino acids) and the sequence identity between the template and the target is significantly higher with 1AY0 at 99.9% and 1IK6 at 22.9%.

Similarly, RAPTOR had selected 1BRL as the top ranking model for PGK1 and the fold of the model only agreed with the target to the level of Architecture (3.x.x.x). With the aid of crosslinking data, the model 1BIF was selected as the best model instead and it matched the target fold down to the level of Topology (3.40.50.x). The length of the template 1BIF (432 amino acids) is slightly longer than the target (416 amino acids) while the length of template 1BRL (252 amino acids) is approximately 2/3 of the target sequence length. Although the lengths of the 1BIF and PGK1 are very close, their sequence identity is only 19.5%.

These results suggest that sequence coverage may be a larger determinant in the quality of the model than sequence identity. Furthermore, TKL1 and PGK1 have shown that crosslinking data can be used to refine the computational data by providing supporting or discriminating evidence for the models.

5.3.6 Fold recognition using crosslinking data

For each target protein, the templates with the most experimentally observed crosslinking sites were used to examine the fold recognition capability of the crosslinking data. Table 5-5 presents the models with the highest number of crosslinking sites consistent with the experimental data for each target. Crosslinking data had successfully selected models with the correct folds for TKL1, IPP1 and PGK1. However, if the best model is selected based on the maximum number of crosslinking sites, the model 1D0C selected for HIS7 and the model 1TJU for ENO1 would have the incorrect folds, each model supported by 4 crosslinking sites. This indicates that crosslinking data cannot be used by itself for selecting the most appropriate protein model.

Table 5-5. RAPTOR models with the highest number of validated crosslinking sites.

PDB ID	CATH code	Model ID	Model CATH code	Template sequence length (target)	% Identity	RAPTOR ranking (z-score)	Number of crosslinks
TKL1	3.40.50.920 3.40.50.970	1AY0	3.40.50.920 3.40.50.970	678 (680)	99.9%	2 (163.84)	8
IPP1	3.90.80.10	1M38	3.90.80.10	287 (287)	100%	1 (298.14)	8
HIS7	3.20.20.70 3.40.50.880	1GVO	3.90.340.10 3.90.440.10 3.90.1230.10	416 (552)	24.0%	13 (31.6)	4
PGK1	3.40.50.1260	1BIF	3.40.50.300 3.40.50.1240	432 (416)	19.5%	5 (76.27)	1
ENO1	3.30.390.10 3.20.20.120	1TJU	1.10.40.30 1.10.275.10 1.20.200.10	448 (437)	20.9%	32 (50.72)	4

5.4 Discussion

5.4.1 Evaluation of model validation with crosslinking data

One of the major goals of this project was to generate experimental data to provide complementary data for validating computational protein models. Results for TKL1 and IPP1 showed that this method of model validation can be successful. In each case, one single model, with the correct CATH code, was validated by a high number of crosslinking sites while several other models with the same CATH code were weakly supported by lesser number of crosslinking sites. Multiple model validation at the level of fold has increased the confidence in the accuracy of these top models. Similar results were found for the HIS7, PGK1 and ENO1 models where multiple models can be experimentally validated. Unfortunately, the evidence, at 3 or less crosslinking sites, provided weak support for the models and is not strong enough to distinguish these models above the other folds. Therefore, it is not possible to conclusively validate or refute these models with these data.

TKL1 and IPP1 have a high number of experimentally observed crosslinking sites available for successful model validation while the results for HIS7 and PGK1 were inconclusive due to insufficient data. These results implied that the success of model validation can be predicted by the quantity of experimental data available. However, ENO1 has a large number of experimentally observed crosslinking sites but only a few of them are consistent with predicted sites on the protein models. This suggests that the number of experimentally observed crosslinking sites is a poor indicator of the success of model validation.

In addition, TKL1 and PGK1 have shown that model selection can be improved by increasing sequence coverage between the template and the target and by validating the models against experimental data. However, the reverse is not true; high sequence coverage does not guarantee better structural models and not all experimental data will contribute to model validation. ENO1 was a good example where it has seven experimentally observed crosslinking sites but the top RAPTOR models can only be validated by three crosslinking sites. Furthermore, there is a 100% sequence coverage between template 1TJU and ENO1 but 1TJU has the wrong fold and was ranked 32nd by RAPTOR. These observations suggest that no factor should be used by itself for model selection. Instead, the crosslinking data, sequence coverage and RAPTOR rankings should be considered simultaneously in the selection of the best protein model.

5.4.2 Evaluation of fold recognition with crosslinking data

Ideally, it would be possible to use the chemical crosslinking data to select the correct protein model from a given data set. It would be hoped that the best model will have the highest number of validated crosslinking sites. However, results have shown that this is not always the case and sometimes the crosslinking data may even select the incorrect fold. One possible reason is that although all crosslinking data is valuable to model validation, it was evident that some crosslinking sites provided more valuable information on the fold of the protein than others. For example, in Figure 5-2, residues 311-321 and 314-322 (••••••••••) both crosslinked lysine residues on the same long α-helices, providing only local information on the fold of the protein. Another example was 235-311 and 235-314 (−−−−−). These crosslinking sites provided more useful information on the overall protein fold since Lys-235 was located on a different α-helix than Lys-311 and 314; however, Lys-311 and 314 were on the same α-helix and thus, 235-311 and 235-314 provided similar information on the protein fold. Moreover, crosslinking sites on loop regions are typically not useful for fold recognition since the fold of a protein is largely determined by the arrangements of helices and beta-strands. Presently, the *LinkModel* algorithm does not discriminate against crosslinking sites that are located on the same secondary structures or on loop regions. It is possible that the success of fold recognition will increase if these two factors are incorporated into the algorithm.

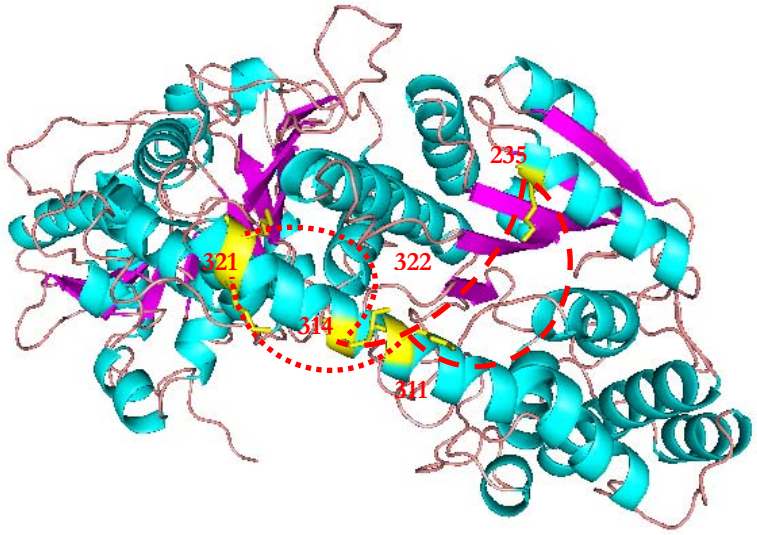


Figure 5-2. Comparison of crosslinking sites contributing to the secondary structures versus the fold of the protein.

6 Conclusion

The main objective of this project was to investigate chemical crosslinking as a method to generate experimental evidence in support of computationally predicted protein structures. A second objective was to determine whether this approach can be used to select correctly folded protein from a library of structures.

To attain these goals, it was necessary to use a purification method that allows the protein to be isolated under non-denaturing conditions and with high product yield. Tandem affinity purification was a simple and efficient technique for purifying target proteins. The purity of the final TAP elution was confirmed by 1D gel electrophoresis and MS analyses. The purified proteins can be used for structural studies similar to those described in this project, as well as protein complex analysis (Gavin et al. 2002; Rigaut et al. 1999). A protein complex isolated by TAP can be separated on a 1D gel and each component can be identified by MS. Furthermore, the chemical crosslinking approach can also reveal spatial organization of the protein subunits within the complex (Back et al. 2001; Rappsilber et al. 2000). Additionally, this generic protein purification approach can be applied to other organisms including mammalian cell lines, as demonstrated by Li et al. (2004).

This chemical crosslinking approach can be applied to any purified protein. However, the results from this research indicated that the crosslinking condition must be optimized for individual proteins. This presents a practical limitation to this TAP-crosslinking approach, especially when working with a low abundance target protein and precious samples are wasted while optimizing crosslinking conditions by trial and error. Therefore, careful planning of the experiments becomes very important in ensuring the success of this method. Computational analysis such as the XlinkPlan algorithm (Ye et al. 2004) can facilitate the experimental design by allowing researchers to explore various combinations of target proteins and crosslinking reagents without wasting precious protein samples. The experimental designs derived from a computational analysis should aid the researcher in selecting the optimal crosslinking reagents and reaction conditions that will generate the high quality crosslinking data.

Potentially all crosslinking data obtained from experiments can contribute to the validation of protein structure models. Any crosslinking sites identified provide experimental evidence for supporting the model; even crosslinking sites that appeared on the same secondary structures or loop regions are useful to model validation, although they do not make a large contribution to fold recognition. Thus, the confidence of the model will increase with respect to the amount of

crosslinking data obtained from the experiments. This was apparent when crosslinking data was used to support the models generated by RAPTOR from a protein library. Several models for target proteins were strongly validated with a high number of experimentally determined crosslinking sites and these results concurred with the scores given by RAPTOR for accessing the quality of the model. In addition to model validation, crosslinking data can also be used in protein model selection. Sometimes the top RAPTOR models are not always the best models due to lower sequence coverage or sequence identity. However, with the aid of crosslinking data, higher quality models can be identified from the data set.

Several future directions have resulted from the outcome of this research. The foremost task is to optimize the chemical crosslinking experiments to generate as many crosslinking sites as possible for model validation. This could be accomplished by using different species of chemical crosslinkers and variations of experimental conditions. Tandem MS on crosslinked peptides can help to definitively identify the peptides involved in crosslinking and minimize the possibility of false positive results. Also, a guideline can be determined to ensure clear distinction between good and bad models. For example, each model must contain a minimum number of experimentally observed crosslinking sites for proper model validation. Furthermore, the success of fold recognition may be improved by excluding crosslinking sites located on the same secondary structures or loop regions from the data set and using only the crosslinking data that are pertinent to the fold of the structure. The combination of these modifications will greatly improve the overall quality of this experimental/computational approach for studying protein structures and enhance its capability in model validation and fold recognition. It is hoped that these modifications will improve the capabilities of this approach in fold recognition and enable the use of this method for fold recognition on novel protein targets.

References

Altschul, S. F. et al. (1997) **Gapped BLAST and PSI-BLAST : a new generation of protein database search programs.** *Nucleic Acids Research* 25: 3389-3402.

Back, J. W. et al. (2001) **A new crosslinker for mass spectrometric analysis of the quaternary structure of protein complexes.** *Journal of American Society for Mass Spectrometry* 12: 222-227.

Back, J. W. et al. (2003) **Chemical cross-linking and mass spectrometry for protein structure modeling.** *Journal of Molecular Biology* 331: 303-313.

Balakrishnan, R. et al. (2005) **Fungal BLAST and model organism BLASTP best hits: new comparison resources at the Saccharomyces Genome Database (SGD).** *Nucleic Acids Research* 33: D374-377. Accessed on Nov. 28, 2005 <URL: http://www.yeastgenome.org/>.

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

Bernasconi, A. and Segre, A. M. (2000) **Ab initio methods for protein structure prediction: A new technique based on Ramachandran Plots.** *European Research Consortium for Informatics and Mathematics News* No. 43. Accessed on Mar. 21 2006 <URL: http://www.ercim.org/publication/Ercim_News/enw43/bernasconi.html>.

Bhat, T. N. et al. (2001) **The PDB data uniformity project.** *Nucleic Acids Research* 29: 214-218.

Bourne, P. E. and Weissig, H., eds. (2003) **Structural Bioinformatics.** New Jersey: Wiley-Liss.

Bradford, M. M. (1976) **A rapid sensitive method for the quantification of microgram quantities of protein utilizing the principle of protein-dye binding.** *Analytical Biochemistry* 72: 248-254.

Carrington, J. C. and Dougherty, W. G. (1988) **A viral cleavage site cassette: identification of amino acid sequences required for tobacco etch virus polyprotein processing.** *Proceedings of National Academy of Science* 85: 3391-3395.

Cates, S. (2005) **PSI-BLAST.** *Connexions* module m11040, version 2.9. Accessed on December 13, 2005 <URL: http://cnx.rice.edu/content/m11040/latest/>.

Chaudhuri, B. N. et al. (2003) **Toward understanding the mechanism of the complex cyclization reaction catalyzed by imidazole glycerolphosphate synthase: crystal structures of a ternary complex and free enzyme.** *Biochemistry* 42: 7003-7012.

DeLano, W. L. (2002) **The PyMOL molecular graphics system.** *DeLano Scientific* San Carlos, CA, USA.

Dihazi, G. H. and Sinz, A. (2003) **Mapping low-resolution three-dimensional protein structures using chemical cross-linking and Fourier transform ion-cyclotron resonance mass spectrometry.** *Rapid Communications in Mass Spectrometry* 17: 2005-2014.

Dougherty, W. G. et al. (1988) **Biochemical and mutational analysis of a plant virus polyprotein cleavage site.** *EMBO J.* 7: 1281-1287.

Gavin, A. C. et al. (2002) **Functional organization of the yeast proteome by systematic analysis of protein complex.** *Nature* 415: 141-147.

Ghaemmaghami, S. et al. (2003) **Global analysis of protein expression in yeast.** *Nature* 425: 737-741.

Gingras, A. C. (2003) **Mammalian TAP-tagging technique – General protocols collection.** *Institute for Systems Biology.* Accessed on September 26, 2005 <URL: www.proteomecenter.org/protocols/Mammalian%20TAP%20protocols.pdf>.

Green, N. et al. (2001) **Quantitative evaluation of the lengths of homobifunctional protein cross-linking reagents used as molecular rulers.** *Protein Science* 10: 1293-1304.

Guex, N. and Peitsch, M. C. (1997) **SWISS-MODEL and the Swiss-PDB Viewer: and environment for comparative protein modeling.** *Electrophoresis* 15: 2714-2723.

Haniu, M. et al. (1993) **Recombinant human erythropoietin (rHuEPO): Cross-linking with disuccinimidyl esters and identification of the interfacing domains in EPO.** *Protein Science* 2: 1441-1451.

Harutyunyan, E. H. et al. (1996) **X-ray structure of yeast inorganic pyrophosphatase complexed with manganese and phosphate.** *European Journal of Biochemistry* 239: 220-228.

Hochuli, E. et al. (1987) **New metal chelate absorbent selective for proteins and peptides containing neighbouring histidine residue.** *Journal of Chromatography* 411: 177-184.

Hud, W. K. et al. (2003) **Global analysis of protein localization in budding yeast.** *Nature* 425: 686-691.

Ito, T. et al. (2001) **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proceedings of the National Academy of Sciencesi* 98: 4569-4574.

Kruppa, G. H. et al (2003) **A top down approach to protein structure studies using chemical cross-linking and Fourier transform mass spectrometry.** *Rapid Communications in Mass Spectrometry* 17: 155-162.

Kwaw, I. et al. (2000) **Thiol cross-linking of cytoplasmic loops in lactose permease of *Escherichia coli*.** *Biochemistry* 39: 3134-3140.

Laemmli, U.K. (1970) **Cleavage of Structural Proteins during the assembly of the head of bacteriophage T4.** *Nature* 227: 680-685.

Larsen, T. M. et al. (1996) **A carboxylate oxygen of the substrate bridges the magnesium ions at the active site of enolase: structure of the yeast enzyme complexed with the equilibrium mixture of 2-phosphoglycerate and phosphoenolpyruvate at 1.8 Å resolution.** *Biochemistry* 35: 4349-4359.

Lathrop, R. H. (1994) **The protein threading problem with sequence amino acid interaction preferences is NP-complete.** *Protein Engineering* 7: 1059-1068.

Lee, B. and Richards, F. M. (1971) **The interpretation of protein structures: estimation of static accessilbility.** *Journal of Molecular Biology* 55: 379-400.

Li, Q. et al. (2004) **A modified mammalian tandem affinity purification procedure to prepare funcational polycystin-2 channel.** *FEBS Letters* 576: 231-236.

Liebler, D. C. **Introduction to proteomics: tools for the new biology.** New Jersey: Humana Press, 2002.

Mann, M. et al. (2001) **Analysis of proteins and proteomes by mass spectrometry.** *Annual Review of Biochemistry* 70: 437-473.

Marchler-Bauer, A. et al. (2005) **CDD : a Conserved Domain Database for protein classification.** *Nucleic Acids Research* 33: D192-6.

McPhillips, T. M. et al. (1996) **Structure of the R65Q mutant of yeast 3-phosphoglycerate kinase complexed with Mg-AMP-PNP and 3-phospho-D-glycerate.** *Biochemistry* 35: 4118-4127.

MS-Labor (2005) **Little encyclopedia of mass spectrometry.** Mass Spectrometry Facility, Intitute of Organic Chemistry, University of Heidelberg. Accessed on December 13, 2005 <URL: http://www.rzuser.uni-heidelberg.de/~bl5/encyclopedia.html>.

Muller, D. R. et al. (2001) **Isotope-tagged cross-linking reagents. A new tool in mass spectrometric protein interaction analysis.** *Analytical Chemistry* 73: 1927-1934.

Muzrin, A. G. et al. (1995) **SCOP : A structural classification of proteins database for the investigation of sequences and structures.** *Journal of Molecular Biology* 247: 536-540.

Nayak, S. et al. (2003) **Enhanced TEV Protease extends enzyme stability for long-term activity.** *Focus* 25: 12-14.

Nikkola, M. et al. (1994) **Refined structure of transketolase from Saccharomyces cerevisiae at 2.0 Å resolution.** *Journal of Molecular Biology* 238: 387-404.

O'Donovan, C. et al. (2002) **High-quality protein knowledge resources: SWISS-PROT and TrEMBL.** *Briefings in Bioinformatics* 3: 275-284.

Orengo, C. A. et al. (1997) **CATH – A hierarchic classification of protein domain structures.** *Structure* 5 : 1093-1108.

Pappin, D. J. C. and D. N. Perkins (2005) **MSDB: Mass spectrometry protein sequence database.** Proteomics Group, Imperial College, London. Accessed on October 11, 2005 <URL: http://csc-fserve.hh.med.ic.ac.uk/msdb.html>.

Pearl, F. M. G. et al. (2005) **CATH v. 2.6.0 (rel. April 2005)** *CATH: Protein structure classification.* Accessed on October 11, 2005 <http://cathwww.biochem.ucl.ac.uk/latest/lists/index.html/>.

Pearson, K. M. et al. (2002) **Intramolecular cross-linking experiments on cytochrome c and ribonuclease A using an isotope multiplet method.** *Rapid Communications in Mass Spectrometry* 16: 149-159.

Perkins, D. N. et al. (1999) **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 20: 3551-3567.

Pierce Biotechnology. (2002) **BS^3 (Bis[sulfosuccinimidyl] suberate) Technical Manual.** Accessed on December 13, 2005 <URL: http://www.piercenet.com/Products/Browse.cfm?fldID=02030212>

Pierce Biotechnology. (2005) **Cross-linking reagents technical handbook.** Pierce Biotechnology, Inc., USA.

Potluri, S. et al. (2004) **Geometric analysis of cross-linkability for protein fold discrimination.** *Pacific Symposium Biocomputing* 447-458.

Puig, O. et al. (2001) **The tandem affinity purification (TAP) method: a general procedure of protein complex purification.** *Methods* 24: 218-229.

Rappsilber, J. et al. (2000) **A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry.** *Analytical Chemistry* 72: 267-275.

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

Rost, B. et al. (1997) **Protein fold recognition by prediction-based threading.** *Journal of Molecular Biology* 270: 471-480.

Schilling, B. et al. (2003) **MS2Assign, Automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides.** *Journal of American Society of Mass Spectrometry* 14: 834-850.

Schwede, T. et al. (2003) **SWISS-MODEL: an automated protein homology-modeling server.** *Nucleic Acids Research* 31: 3381-3385.

Sinz, A. (2003) **Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes.** *Journal of Mass Spectrometry* 38: 1225-1237.

Smith, D. and Johnson, K. S. (1988) **Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione *S*-transferase.** *Gene* 67: 31-40.

Trester-Zedlitz, M. et al. (2003) **A modular cross-linking approach for exploring protein interactions.** *Journal of American Chemistry Society* 125: 2416-2425.

Uetz, P. et al. (2000) **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 403: 601-603.

Voet, D. and Voet, J. G. **Biochemistry.** 2nd ed. New York: John Wiley & Songs, Inc., 1995.

Westbrook, J. et al. (1997) **The Protein Data Bank: unifying the archive.** *Nucleic Acids Research* 30: 245-248.

Xu, J. et al. (2003) **RAPTOR: Optimal protein threading by linear programming.** *Journal of Bioinformatics and Computational Biology* 1: 95-117.

Xu, J. and Li, M. (2003) **Assessment of RAPTOR's linear programming approach in CAFASP2.** *Proteins: Structure, Function and Genetics* 53: 579-584.

Ye, X. et al. (2004) **Probabilistic cross-link analysis and experiment planning for high-throughput elucidation of protein structure.** *Protein Science* 13:3298-3313.

Yeast Resource Center (2002) **Tandem Affinity Purification Protocol.** YRC Mass Spectrometry, University of Washington. Accessed on September 23, 2005 <URL: http://depts.washington.edu/~yeastrc/ms_tap1.htm>.

Young, M. M. et al. (2000) **High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry.** *Proceedings of National Academy of Science* 97: 5802-5806.

Appendix A

Table A-1. M/z values and charges for peaks unique to the crosslinked TKL1 MS spectra.

m/z	charge	m/z	charge	m/z	charge	m/z	charge	m/z	charge
328.1947	2	665.3022	2	856.4230	2	1045.4838	3	1268.5935	4
335.1765	2	667.8047	6	857.6426	3	1050.4742	3	1272.0530	2
360.8760	3	669.3126	2	861.3710	3	1060.4773	2	1274.2792	3
375.2370	3	677.3126	5	865.3974	2	1067.1591	3	1275.5751	2
378.1899	3	678.7923	2	869.8660	2	1067.4745	3	1277.5863	2
380.2310	2	683.8343	2	869.8713	2	1073.8306	3	1284.9137	3
385.7139	2	684.8264	2	873.3937	4	1082.1591	3	1288.0739	2
397.2201	2	686.3107	2	880.8821	2	1089.4828	3	1292.2334	3
398.2286	3	689.3481	2	882.9167	2	1091.9475	2	1297.5613	3
403.2167	2	693.3156	2	884.4011	2	1096.4397	3	1301.5886	2
405.8745	3	697.1544	2	886.8837	6	1101.0048	2	1311.5682	3
410.8853	3	698.7704	5	891.3884	2	1101.0085	2	1312.5890	2
416.2013	2	700.3022	2	895.3865	2	1105.5021	2	1315.6183	3
429.2395	2	701.8354	2	896.8462	2	1107.4690	3	1322.9504	3
435.7701	2	703.3163	2	898.8679	2	1109.7910	3	1349.5900	4
455.9277	4	705.3185	3	902.3995	2	1112.4999	5	1350.6415	3
461.2462	3	707.8267	2	911.4755	2	1118.4757	2	1379.6138	2
476.2739	2	713.0988	4	913.4149	4	1124.8575	3	1390.1615	4
502.2512	3	726.8054	4	915.4197	3	1129.1945	3	1425.6616	2
505.7577	2	728.2946	3	917.4004	2	1132.1873	3	1437.6356	2
506.7579	2	734.3271	3	920.3920	3	1139.1687	3	1481.6812	2
507.2446	2	743.3505	2	933.0496	3	1146.4862	3	1508.2344	2
547.7780	2	772.3481	3	934.8746	2	1158.0205	2	1519.2299	2
547.7949	2	778.3505	2	944.8390	2	1177.5287	2	1519.2543	2
550.7481	2	782.3377	5	950.7524	3	1184.9996	2	1556.2511	2
552.2806	2	785.0092	3	955.4685	4	1197.0049	2	1562.4127	3
557.2850	2	786.3581	3	959.6583	4	1204.2417	3	1580.4930	5
581.2690	4	789.7515	5	963.9130	4	1212.5475	3	1582.6825	2
584.2914	3	796.0052	5	968.4333	3	1215.5032	2	1591.6997	5
585.2874	3	798.3540	5	969.4033	4	1220.5560	2	1610.2527	2
588.9535	3	803.3744	2	974.4073	4	1225.9332	2	1622.7633	2
595.8121	2	805.6233	4	986.9408	4	1228.5490	2	1633.7561	2
597.3188	2	808.3640	3	995.4276	3	1229.5485	2	1644.7362	2
602.7706	4	814.8286	2	996.4521	4	1236.5955	2	1646.2417	2
615.7993	2	817.3684	4	997.7012	4	1239.5649	2	1670.7388	3
615.8065	2	819.3517	3	1005.4371	3	1241.5472	5	1686.7965	2
616.3226	2	824.3556	3	1009.4239	2	1243.5453	2	1697.2980	2
629.5342	4	825.7026	3	1015.0811	5	1247.5618	2	1708.7758	2
634.2909	2	839.0400	3	1018.4572	3	1252.5934	2	1805.8378	2
637.0258	3	844.3692	2	1020.1367	3	1257.0575	2	1816.7723	2
646.3544	3	848.6883	3	1020.4282	2	1263.5580	2		
653.7783	2	851.3940	3	1024.1265	3	1265.0950	2		
655.3016	2	854.4125	2	1030.1650	3	1265.1099	2		
661.2972	3	856.4164	2	1037.8018	3	1267.5914	2		

Table A-2. M/z values and charges for peaks unique to the crosslinked IPP1 MS spectra.

m/z	charge	m/z	charge	m/z	charge	m/z	charge
331.0562	1	724.3449	2	1004.0316	2	1465.7231	3
331.0593	1	724.3570	2	1037.0059	2	1465.7247	3
372.0882	1	725.3256	5	1050.4891	2	1486.7399	2
372.0914	1	726.3134	5	1050.4896	2	1486.7523	2
390.1012	1	728.3715	3	1058.5590	2	1501.7384	3
390.1056	1	729.3625	2	1058.5717	2	1501.7653	3
424.2646	2	729.3845	2	1068.5764	2	1509.7009	2
424.2676	2	737.4077	4	1068.5887	2	1509.7135	2
428.7525	2	742.0271	3	1070.0201	4	1531.7316	2
428.7556	2	742.3593	3	1078.0197	4	1531.7441	2
430.2436	4	743.4098	4	1088.5676	2	1538.7219	2
430.2466	4	772.3691	3	1088.8214	3	1538.7344	2
467.0675	1	772.3813	3	1096.5145	2	1544.7758	2
467.0704	1	810.3904	2	1096.5271	2	1544.7883	2
484.0938	1	810.5558	2	1140.5745	3	1561.3107	2
484.0968	1	812.7625	3	1140.5873	3	1561.8018	2
486.7709	2	812.7742	3	1175.1735	3	1567.7552	2
486.7738	2	831.4123	4	1175.1863	3	1567.7574	2
526.0959	1	831.4352	4	1212.5763	2	1588.7699	2
526.1015	1	843.6823	2	1212.6021	2	1588.7721	2
533.2826	3	843.6840	2	1218.6559	2	1594.7621	3
533.2939	3	859.9238	2	1218.6687	2	1594.7892	3
571.7773	2	859.9362	2	1223.5945	3	1608.7473	2
571.7887	2	876.4180	3	1223.6072	3	1608.7607	2
587.9442	3	876.4304	3	1232.6251	3	1621.7018	3
587.9468	3	898.9033	2	1232.6378	3	1632.7781	2
640.9806	3	898.9048	2	1242.9701	3	1692.8157	3
640.9925	2	905.4084	2	1245.5853	2	1692.8430	3
651.3568	2	905.4319	2	1245.5856	2	1709.9104	2
656.0024	3	912.4248	3	1247.6074	3	1709.9377	2
656.3434	3	912.4372	3	1247.6202	3	1736.8682	2
665.3243	3	968.4566	3	1250.6299	2	1736.9111	2
709.3669	3	970.4911	2	1250.6302	2	1743.8789	1
709.3692	3	970.4922	2	1329.6521	4	1743.8820	1
713.6948	5	973.5124	2	1329.6663	4	1777.8474	1
713.7068	5	973.5135	2	1353.0863	2	1777.8904	1
719.0002	3	991.1613	3	1362.6635	2	1804.8967	1
719.0122	3	991.1738	3	1409.6890	3	1804.9089	1
720.9230	5	997.4696	2	1409.6903	3	1813.3713	2
720.9307	5	997.4822	2	1424.9828	3	1848.4612	2
722.3506	2	1000.4862	2	1451.3831	3	1877.9688	1
722.3627	2	1000.4988	2	1451.3955	3	1877.9966	1

Table A-3. M/z values and charges for peaks unique to the crosslinked HIS7 MS spectra.

m/z	charge	m/z	charge	m/z	charge	m/z	charge
149.5323333	3	299.91295	2	426.078	3	730.64245	2
149.5333333	3	299.91425	2	426.0866	3	738.6305	2
173.0997	3	316.6555	2	432.3174667	3	744.87575	2
173.1006333	3	316.65675	2	438.20205	2	744.87575	2
194.7614667	3	330.7224	3	438.2082	2	744.8821	2
194.7623667	3	331.0597	1	439.22545	2	754.87215	2
207.11685	2	331.0632	1	449.0805333	3	754.8729	2
207.1184	2	339.4913	3	449.0808	3	764.8856	2
211.1041333	3	345.69255	2	540.281	1	782.89315	2
211.1046333	3	345.69365	2	677.3134	2	885.9125	2
230.6228	2	406.20475	2	677.3207	2		

Table A-4. M/z values and charges for peaks unique to the crosslinked PGK1 MS spectra.

m/z	charge	m/z	charge	m/z	charge	m/z	charge
348.0945	1	1026.4794	3	1360.2300	2	1619.3135	2
348.0957	1	1139.5002	2	1374.6461	2	1619.3225	2
366.1079	1	1139.5084	2	1374.6957	2	1695.3147	2
366.1092	1	1215.5354	2	1379.6864	2	1695.3239	2
657.1781	1	1215.5566	2	1379.6952	2	1837.8966	2
657.1804	1	1283.1289	2	1494.2936	2	1837.9215	2
677.1837	1	1306.2120	2	1494.2988	2	1895.9463	2
677.1862	1	1306.2168	2	1545.8000	2	1895.9557	2
695.1090	1	1337.0520	2	1545.8057	2	1971.9547	2
695.1162	1	1337.0568	2	1562.2552	1	1971.9617	2
1026.4640	2	1360.2251	2	1562.2971	1		

Table A-5. M/z values and charges for peaks unique to the crosslinked ENO1 MS spectra.

m/z	charge	m/z	charge	m/z	charge	m/z	charge
421.1898	1	733.1121	1	1302.7292	1	1656.7557	2
421.1913	1	761.7194	3	1302.7338	1	1661.7650	2
434.1325	1	761.7222	3	1327.6547	2	1707.8253	1
434.1340	1	796.2570	1	1327.6633	2	1719.8704	1
453.1748	1	800.4029	2	1329.7104	1	1719.9071	1
453.1764	1	800.4161	2	1329.7554	2	1748.8768	2
465.0513	1	835.1894	4	1353.1173	2	1776.8353	1
465.0529	1	911.4539	2	1353.1355	2	1776.8444	1
471.1906	1	911.4572	2	1383.7292	1	1789.9500	1
477.2033	1	943.4360	3	1390.7245	4	1794.8611	1
510.1972	1	964.4557	2	1390.7295	4	1821.9607	1
510.1990	1	968.4611	3	1407.7239	1	1821.9855	1
528.2066	1	995.4695	3	1407.7424	1	1853.9863	3
528.2132	1	1076.5182	2	1414.6876	2	1916.9297	1
541.1146	1	1076.5339	2	1414.6926	2	1963.0249	1
541.1165	1	1086.5256	3	1439.8444	1	2035.9844	2
546.2242	1	1088.5472	3	1439.8495	1	2035.9939	2
546.2348	1	1142.5881	2	1452.2170	2	2063.9897	1
550.1884	1	1142.5922	2	1452.2261	2	2063.9993	1
550.1950	1	1203.5856	1	1459.7529	3	2117.0256	1
602.3146	2	1203.6664	1	1519.7393	2	2117.0522	1
623.1857	1	1207.5879	2	1520.7501	3	2192.0115	1
623.1926	1	1207.5922	2	1555.7404	2	2192.1069	1
706.1228	2	1225.6199	2	1629.3418	2	2264.0823	1
711.1228	1	1233.1622	2	1629.3475	2	2264.1248	1
711.1253	1	1257.1123	2	1632.7872	2	2392.1990	1
715.3113	1	1257.1207	2	1632.8109	2	2392.2253	1
733.1095	1	1283.6194	2	1640.8069	2	2432.1226	2

Appendix B

Table B-1. Results for MS data analysis of crosslinked TKL1 using ASAP.

Exp Mass	Thr Mass	Chrg	Sequence	Lys Num
581.2690	581.117	4	TLAAKNIKAR - GDKLISPLKK	582-671
667.8047	667.691	6	TILKPGVEANNKWNK - PGVEANNKWNKLFSEYQK	303-311, 303-314
796.0052	795.837	5	TILKPGVEANNKWNK - LFSEYQKKFPELGAELAR	311-321, 311-322
1124.8575	1124.612	3	FGASGKAPEVFK - NIKARVVSLPDFFTFDK	671-582
1129.1945	1128.940	3	QLKSKFGFNPDK - MTQFTDIDKLAVSTIR	276-9, 278-9
1390.1615	1389.955	4	WKEALDFQPPSSGSGNYSGRYIR - LSGQLPANWESKLPTYTAKDSAVATR	392-345
1508.2344	1508.287	2	SKFGFNPDK - MTQFTDIDKLAVSTIR	9-278

Table B-2. Results for MS data analysis of crosslinked IPP1 using ASAP.

Exp Mass	Thr Mass	Chrg	Sequence	Lys Num
713.6948	713.395	5	EETLNPIIQDTKK - QIGAKNTLEYKVYIEK	11/74, 17/74
720.9307	720.771	5	SIDKWFFISGSV - IPDGKPENQFAFSGEAKNK	199/279
973.5124	973.591	2	GKLRFVR – IYKIPDGK	77/194
1140.5745	1140.615 1140.294 1140.288	3	SIDKWFFISGSV – QIGAKNTLEYKVYIEK VIAIDINDPLAPKLNDIEDVEKYFPGLLR KGK - QLIAGKSSDSKGIDLTNVTLPDTPTYSK	168/177
1588.7721	1588.356 1588.328	2	WTNAKLEITK - AASDAIPPASPKADAPIDK WTNAKLEITK - IPDGKPENQFAFSGEAK	57/268 57/199
1813.3713	1812.923	2	PENQFAFSGEAKNK - LNDIEDVEKYFPGLLR	177/211

Table B-3. Results for MS data analysis of crosslinked HIS7 using ASAP.

Exp Mass	Thr Mass	Chrg	Sequence	Lys Num
1020.4739	1020.200	3	SGKAGLNVIENFLKQQSPPIPNYSAEEK	199/210
1298.9524	1298.958	3	YCWYQCTIKGGR - TNDQGDLVVTKGDQYDVREK	258/441
1298.9524	1298.958	3	YCWYQCTIKGGR - TNDQGDLVVTKGDQYDVREK	419/432
1510.7443	1510.817	2	EYLLEHGLKVR - AKYGSEEFIAAVNK	172/546

Table B-4. Results for MS data analysis of crosslinked PGK1 using ASAP.

Exp Mass	Thr Mass	Chrg	Sequence	Lys Num
1306.2168	1306.1768	2	KVLENTEIGDSIFDKAGAEIVPK	244/258

Table B-5. Results for MS data analysis of crosslinked ENO1 using ASAP.

Exp Mass	Thr Mass	Chrg	Sequence	Lys Num
1390.7245	1390.746	4	AAGHDGKIK - TSPYVLPVPFLNVLNGGSHAGGALALQE FMIAPTGAKTFAEALR	178/241
1519.7393	1519.787	2	SKWMGK - DQKAVDDFLISLDGTANKSK MAVSKVYAR - GNPTVEVELTTEKGVFR	5/28
1640.8069	1640.390	2	DGDKSK – WMGKGVLHAVKNVNDVIAPAFVK	54/67
			DGDKSKWMGK - GVLHAVKNVNDVIAPAFVK	56/67
1748.8768	1748.959	2	SKLGANAILGVSLAASR - MAVSKVYARSVYDSR	5/105
1853.9863	1853.992	3	AAGHDGKIK - TSPYVLPVPFLNVLNGGSHAGGALALQE FMIAPTGAKTFAEALR	178/241
2192.1069	2192.191	1	MAVSKVYAR - ANIDVKDQK	5/85

Appendix C

Table C-1-a. RAPTOR models for TKL1, ranking from 1 to 30.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking[1])	z-Score Ranking
1ik6a	284	197.23	3.40.50.920	0 (7)	1
1ay0a	678	163.84	3.40.50.920	8 (1)	2
1deoa	233	85.05	3.40.50.1110	0 (7)	3
1nr0a	610	83.96	2.130.10.10	0 (7)	4
1btma	251	82.38	3.20.20.70	0 (7)	5
1n0ua	819	80.18	2.40.30.10	2 (5)	6
2gbp	309	79.66	3.40.50.2300	0 (7)	7
1rpxa	230	78.75	3.20.20.70	0 (7)	8
1m1ba	291	78.72	3.20.20.60	0 (7)	9
1igs	247	78.45	3.20.20.70	0 (7)	10
1m5wa	242	78.27	3.20.20.70	0 (7)	11
1b5ta	275	76.42	3.20.20.220	0 (7)	12
1h1ya	219	76.08	3.20.20.70	0 (7)	13
1b54	230	74.38	3.20.20.10	0 (7)	14
1bfd	523	73.88	3.40.50.970	2 (5)	15
1vc4a	254	73.76	3.20.20.70	0 (7)	16
1nvma	340	73.61	1.10.8.60	0 (7)	17
1i4na	251	73.53	3.20.20.70	0 (7)	18
1f12a	293	73.12	1.10.1040.10	0 (7)	19
1h70a	255	73.02	3.75.10.10	0 (7)	20
1eepa	314	72.74	3.20.20.70	0 (7)	21
1dd9a	310	72.52	1.20.50.20	1 (6)	22
1zpda	565	72.31	3.40.50.970	1 (6)	23
1qgoa	257	72.11	3.40.50.1400	0 (7)	24
1o5qa	271	72.09	3.20.20.60	0 (7)	25
1byka	255	72.02	3.40.50.2300	0 (7)	26
1euca	306	71.67	3.30.470.20	0 (7)	27
1xyza	320	71.65	3.20.20.80	0 (7)	28
1ovma	535	71.36	3.40.50.970	2 (5)	29
1l6wa	220	71.33	3.20.20.70	0 (7)	30

[1] Crosslink rankings were assigned to each model based on the number of crosslinking sites consistent with experimental data. Models with the highest number of crosslinking sites were ranked as first and all the models with the same number of crosslinking sites were given the same rank.

Table C-1-b. RAPTOR models for TKL1, ranking from 31 to 66.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1v93a	292	71.17	3.20.20.220	0 (7)	31
1kgza	328	71.08	1.20.970.10	0 (7)	32
1smla	266	70.88	3.60.15.10	0 (7)	33
1dkra	298	70.25	3.40.50.2020	0 (7)	34
1dbta	237	70.24	3.20.20.70	0 (7)	35
1bhs	284	69.81	3.40.50.720	0 (7)	36
1de0a	289	69.79	3.40.50.300	1 (6)	37
1p74a	267	69.36	3.40.50.720	0 (7)	38
1gvoa	362	69.32	3.20.20.70	0 (7)	39
1lst	239	69.04	3.40.190.10	0 (7)	40
1dqwa	267	68.79	3.20.20.70	0 (7)	41
1ckea	212	68.63	3.40.50.300	0 (7)	42
1diaa	285	68.59	3.40.50.720	0 (7)	43
1f8ra	483	68.53	1.10.405.10	0 (7)	44
1f6ya	258	68.39	3.20.20.20	0 (7)	45
1oy0a	248	67.47	3.20.20.60	0 (7)	46
1p49a	548	67.14	3.30.1120.10	3 (4)	47
1a8y	338	66.97	3.40.30.10	0 (7)	48
1mo0a	257	66.86	3.20.20.70	0 (7)	49
1bdha	338	66.67	1.10.260.40	0 (7)	50
1jpwa	502	66.61	1.25.10.10	2 (5)	51
1gdha	320	66.54	3.40.50.720	0 (7)	52
1uuma	350	66.41	3.20.20.70	0 (7)	53
1oi2a	336	66.28	5.1.1476.10	0 (7)	54
1dora	311	66.17	2.30.26.10	0 (7)	55
1oata	404	66.12	3.40.640.10	2 (5)	56
1d2fa	361	65.94	3.40.640.10	0 (7)	57
1k2yx	459	65.94	3.30.310.50	0 (7)	58
1ak5	329	65.77	3.20.20.70	0 (7)	59
1vh7a	250	65.69	3.20.20.70	0 (7)	60
1i1ka	298	65.66	3.20.10.10	0 (7)	61
1j93a	343	65.63	3.20.20.210	0 (7)	62
1c3va	245	65.55	3.30.360.10	0 (7)	63
1g61a	225	65.52	3.75.10.10	0 (7)	64
1eixa	231	65.39	3.20.20.70	0 (7)	65
1ak1	308	65.38	3.40.50.1400	0 (7)	66

Table C-1-c. RAPTOR models for TKL1, ranking from 67 to 100.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1h1na	304	65.32	3.20.20.80	0 (7)	67
1puja	261	65.13	1.10.1580.10	0 (7)	68
1k6ia	318	64.87	3.40.50.720	0 (7)	69
1hoza	316	64.73	3.90.245.10	0 (7)	70
1lf1a	296	64.73	3.20.20.80	0 (7)	71
2lbp	346	64.71	3.40.50.2300	0 (7)	72
1ojxa	250	64.66	3.20.20.70	0 (7)	73
1kvka	378	64.18	3.30.230.10	2 (5)	74
1bf6a	291	64.02	3.20.20.140	0 (7)	75
1j1ia	258	64.00	3.40.50.1820	0 (7)	76
1jpha	357	63.85	3.20.20.210	4 (3)	77
1bd3d	224	63.75	3.40.50.2020	0 (7)	78
1jud	220	63.73	1.10.164.10	0 (7)	79
1k4ka	213	63.56	3.40.50.620	0 (7)	80
1sto	208	63.56	3.40.50.2020	0 (7)	81
1a82	224	63.48	3.40.50.300	0 (7)	82
1p44a	268	63.30	3.40.50.720	0 (7)	83
1f0ka	351	63.27	3.40.50.2000	0 (7)	84
1qwka	312	63.27	3.20.20.100	0 (7)	85
1qpba	555	62.87	3.40.50.970	2 (5)	86
1bif	432	62.77	3.40.50.300	1 (6)	87
1mxsa	216	62.58	3.20.20.70	0 (7)	88
1a0ga	280	62.36	3.20.10.10	0 (7)	89
1gega	255	62.26	3.40.50.720	0 (7)	90
1huva	349	62.10	3.20.20.70	0 (7)	91
1jt1a	262	62.00	3.60.15.10	0 (7)	92
1f2da	341	61.99	3.40.50.1100	1 (6)	93
1qh3a	260	61.99	3.60.15.10	0 (7)	94
1uqya	347	61.88	3.20.20.80	0 (7)	95
1dn1a	556	61.85	1.20.1050.30	4 (3)	96
1g2oa	262	61.84	3.40.50.1580	0 (7)	97
1l8xa	355	61.80	3.40.50.1400	0 (7)	98
1gqqa	428	61.79	3.40.50.720	4 (3)	99
1gqna	252	61.71	3.20.20.70	0 (7)	100

Table C-2-a. RAPTOR models for IPP1, ranking from 1 to 36.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1m38a	282	298.14	3.90.80.10	8 (1)	1
1udea	168	204.04	3.90.80.10	2 (4)	2
2prd	174	198.52	3.90.80.10	3 (3)	3
1lst	239	46.79	3.40.190.10	2 (4)	4
1fy7a	273	46.23	1.10.10.10	1 (5)	5
1fdr	244	45.78	2.40.30.10	1 (5)	6
1fj2a	229	44.86	3.40.50.1820	0 (6)	7
1eh6a	168	43.29	1.10.10.10	0 (6)	8
1efza	371	43.15	3.20.20.105	3 (3)	9
1sfe	165	43.14	1.10.10.10	0 (6)	10
1mgta	169	42.30	1.10.10.10	2 (4)	11
1jt1a	262	41.54	3.60.15.10	0 (6)	12
1h1ya	219	41.12	3.20.20.70	1 (5)	13
1ii5a	221	40.87	3.40.190.10	2 (4)	14
1qh3a	260	40.66	3.60.15.10	0 (6)	15
1ggga	220	40.21	3.40.190.10	2 (4)	16
1a2za	220	39.99	3.40.630.20	1 (5)	17
1coy	501	39.89	3.30.410.10	1 (5)	18
1auga	210	39.83	3.40.630.20	2 (4)	19
1aqt	135	39.79	1.20.5.440	0 (6)	20
2bdpa	580	39.56	1.10.150.20	1 (5)	21
1rjpa	474	39.33	3.20.20.140	2 (4)	22
2pth	193	39.22	3.40.50.1470	0 (6)	23
2fhi	124	39.21	3.30.428.10	1 (5)	24
1ac5	483	39.03	3.40.50.1820	1 (5)	25
1ba2a	271	38.77	3.40.50.2300	2 (4)	26
1hjra	158	38.68	3.30.420.10	0 (6)	27
1lap	481	38.36	3.40.220.10	1 (5)	28
1p90a	123	38.17	3.30.420.130	1 (5)	29
1gxsa	267	37.69	3.40.50.1820	2 (4)	30
1ujna	338	37.62	1.20.1090.10	2 (4)	31
1jswa	459	37.53	1.10.40.30	1 (5)	32
1qfja	226	37.48	2.40.30.10	2 (4)	33
1neda	180	37.18	3.60.20.10	0 (6)	34
1d1qa	159	37.08	3.40.50.270	0 (6)	35
1ueka	268	37.07	3.30.230.10	1 (5)	36

Table C-2-b. RAPTOR models for IPP1, ranking from 37 to 72.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1gvoa	362	36.82	3.20.20.70	0 (6)	37
1gmua	138	36.73	3.30.70.790	0 (6)	38
1eiza	180	36.72	3.40.50.150	1 (5)	39
1nhua	558	36.66	1.10.1430.10	1 (5)	40
1oe7a	204	36.56	1.20.1050.10	3 (3)	41
1rpxa	230	36.54	3.20.20.70	0 (6)	42
1gqna	252	36.11	3.20.20.70	1 (5)	43
1d02a	197	35.81	3.40.580.10	0 (6)	44
1jmkc	222	35.76	3.40.50.1820	0 (6)	45
1duba	260	35.66	1.10.12.10	1 (5)	46
1nar	289	35.65	3.20.20.80	1 (5)	47
1o5la	129	35.43	2.60.120.10	0 (6)	48
1esc	302	35.40	3.40.50.1110	3 (3)	49
1hska	303	35.09	3.30.43.10	2 (4)	50
1sgfa	203	35.08	2.10.90.10	1 (5)	51
1e9na	274	34.98	3.60.10.10	2 (4)	52
1bc2a	216	34.91	3.60.15.10	3 (3)	53
1mo0a	257	34.90	3.20.20.70	2 (4)	54
1guba	288	34.88	3.40.50.2300	1 (5)	55
1p49a	548	34.87	3.30.1120.10	1 (5)	56
1hx3a	176	34.78	3.90.79.10	1 (5)	57
1jfma	174	34.66	3.30.500.10	2 (4)	58
1f3va	158	34.56	2.60.210.10	3 (3)	59
1owxa	113	34.52	3.30.70.330	0 (6)	60
1o6ea	225	34.45	3.20.16.10	1 (5)	61
1i36a	258	34.40	1.10.1040.10	2 (4)	62
1psza	286	34.32	3.40.50.1980	2 (4)	63
1j4xa	178	34.32	3.90.190.10	1 (5)	64
1auoa	218	34.30	3.40.50.1820	1 (5)	65
1ia9a	276	34.14	3.30.200.20	2 (4)	66
1m2ta	248	34.14	2.80.10.50	2 (4)	67
1omia	224	34.13	1.10.10.10	4 (2)	68
1h3za	108	34.11	2.30.30.160	0 (6)	69
1cbf	240	34.10	3.30.950.10	3 (3)	70
1nf3a	194	34.10	2.30.42.10	3 (3)	71
1hv2a	99	34.10	3.30.710.10	0 (6)	72

Table C-2-c. RAPTOR models for IPP1, ranking from 73 to 100.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1mejb	201	34.09	3.40.50.170	2 (4)	73
1k9aa	439	34.01	1.10.510.10	2 (4)	74
1atg	231	33.99	3.40.190.10	1 (5)	75
1gyta	503	33.88	3.40.630.10	1 (5)	76
1at3a	217	33.83	3.20.16.10	1 (5)	77
1m3ga	145	33.81	3.90.190.10	1 (5)	78
1e3sa	241	33.77	3.40.50.720	2 (4)	79
1a04a	205	33.72	1.10.10.10	2 (4)	80
1isea	184	33.67	1.10.132.20	0 (6)	81
1e6ya	568	33.66	1.20.840.10	1 (5)	82
1o4va	169	33.59	3.40.50.7700	2 (4)	83
1jdia	223	33.57	3.40.225.10	2 (4)	84
1yaca	204	33.54	3.40.50.850	2 (4)	85
1nrza	163	33.52	3.40.35.10	1 (5)	86
1qhla	203	33.46	3.40.1140.10	0 (6)	87
1dxea	253	33.37	3.20.20.60	1 (5)	88
1c2ta	209	33.21	3.40.50.170	0 (6)	89
1hq0a	295	33.11	3.60.100.10	1 (5)	90
1e5ka	188	33.01	3.90.550.10	0 (6)	91
1mk1a	187	32.98	3.90.79.10	0 (6)	92
1mjfa	271	32.96	2.30.140.10	1 (5)	93
1l1ta	259	32.88	2.60.270.10	1 (5)	94
1in0a	162	32.86	3.30.70.860	0 (6)	95
1mjha	143	32.85	3.40.50.620	1 (5)	96
1iw2a	163	32.84	2.40.128.20	1 (5)	97
1e9ra	420	32.83	1.10.8.80	3 (3)	98
1ej2a	167	32.75	3.40.50.620	1 (5)	99
1f4ja	479	32.68	2.40.180.10	1 (5)	100

Table C-3-a. RAPTOR models for HIS7, ranking from 1 to 36.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1ka9h	195	204.65	3.20.20.70	1 (4)	1
1vh7a	250	190.96	3.20.20.70	0 (5)	2
1igs	247	120.64	3.20.20.70	1 (4)	3
1h1ya	219	105.89	3.20.20.70	0 (5)	4
1i4na	251	98.23	3.20.20.70	0 (5)	5
1vc4a	254	97.22	3.20.20.70	1 (4)	6
1g69a	226	93.38	3.20.20.70	0 (5)	7
1rpxa	230	93.25	3.20.20.70	0 (5)	8
1ep1a	309	83.81	2.10.240.10	1 (4)	9
1huva	349	80.52	3.20.20.70	1 (4)	10
1a50a	260	80.12	3.20.20.70	0 (5)	11
1ojxa	250	77.28	3.20.20.70	0 (5)	12
1gvoa	362	75.02	3.20.20.70	1 (4)	13
1uuma	350	74.64	3.20.20.70	0 (5)	14
1bf6a	291	74.27	3.20.20.140	1 (4)	15
1dora	311	73.92	2.30.26.10	1 (4)	16
1dora	311	73.92	2.30.26.10	1 (4)	17
1m1ba	291	73.90	3.20.20.60	0 (5)	18
1psca	329	72.95	3.20.20.140	1 (4)	19
1oya	399	72.92	3.20.20.70	1 (4)	20
1lbma	194	72.43	3.20.20.70	0 (5)	21
1ex1a	602	72.39	3.20.20.300	1 (4)	22
1q45a	365	71.90	3.20.20.70	1 (4)	23
1ct9a	497	71.87	3.40.50.620	0 (5)	24
1ct9a	497	71.87	3.40.50.620	0 (5)	25
1aj0	282	71.40	3.20.20.20	0 (5)	26
1b4ka	326	70.66	3.20.20.70	0 (5)	27
1o5qa	271	69.27	3.20.20.60	0 (5)	28
1nvma	340	69.21	1.10.8.60	1 (4)	29
1gqna	252	68.98	3.20.20.70	0 (5)	30
1ujpa	243	68.11	3.20.20.70	1 (4)	31
1m5wa	242	67.94	3.20.20.70	0 (5)	32
1a2n	418	67.74	3.65.10.10	0 (5)	33
1o1za	226	67.55	3.20.20.190	1 (4)	34
1dbta	237	66.86	3.20.20.70	0 (5)	35
1dbta	237	66.86	3.20.20.70	0 (5)	36

Table C-3-b. RAPTOR models for HIS7, ranking from 37 to 72.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1fcba	494	66.76	3.10.120.10	1 (4)	37
1geqa	241	66.71	3.20.20.70	0 (5)	38
1o0ya	251	65.82	3.20.20.70	1 (4)	39
1jnda	400	65.76	3.10.50.10	1 (4)	40
1cbf	240	65.53	3.30.950.10	0 (5)	41
1ak1	308	65.10	3.40.50.1400	0 (5)	42
1gqqa	428	65.01	3.40.50.720	1 (4)	43
1jcja	252	64.92	3.20.20.70	0 (5)	44
1euca	306	64.40	3.30.470.20	1 (4)	45
1twsa	273	64.34	3.20.20.20	0 (5)	46
1dxea	253	64.22	3.20.20.60	1 (4)	47
1dxea	253	64.22	3.20.20.60	1 (4)	48
1o4ua	265	63.45	3.20.20.70	1 (4)	49
1b5ta	275	63.09	3.20.20.220	0 (5)	50
1a8y	338	62.94	3.40.30.10	1 (4)	51
1fy2a	220	62.94	3.40.50.880	0 (5)	52
1qgoa	257	62.73	3.40.50.1400	1 (4)	53
1mo0a	257	62.10	3.20.20.70	2 (3)	54
1f6ya	258	61.98	3.20.20.20	0 (5)	55
1c3va	245	61.48	3.30.360.10	0 (5)	56
1v93a	292	61.31	3.20.20.220	0 (5)	57
1b3oa	304	61.04	3.20.20.70	0 (5)	58
1dlia	402	60.55	1.10.1040.10	1 (4)	59
1dlia	402	60.55	1.10.1040.10	1 (4)	60
1hp4a	499	60.49	3.20.20.80	0 (5)	61
1qapa	289	60.35	3.20.20.70	0 (5)	62
1ik6a	284	60.24	3.40.50.920	1 (4)	63
1qnva	328	60.12	3.20.20.70	0 (5)	64
1f8ra	483	59.83	1.10.405.10	2 (3)	65
1dkra	298	59.27	3.40.50.2020	1 (4)	66
1dkra	298	59.27	3.40.50.2020	1 (4)	67
1de0a	289	59.21	3.40.50.300	1 (4)	68
1de0a	289	59.21	3.40.50.300	1 (4)	69
1gqta	305	59.10	2.20.26.10	0 (5)	70
1c2ta	209	59.02	3.40.50.170	0 (5)	71
1cjca	455	59.00	3.40.50.720	1 (4)	72

Table C-3-c. RAPTOR models for HIS7, ranking from 73 to 100.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1n7ka	234	58.65	3.20.20.70	0 (5)	73
1miob	457	58.55	1.20.89.10	0 (5)	74
1jr2a	260	58.38	3.40.50.10090	0 (5)	75
1ecea	358	58.08	3.20.20.80	2 (3)	76
1pjba	361	57.94	3.40.50.720	1 (4)	77
1a9o	289	57.89	3.40.50.1580	1 (4)	78
1qwga	251	57.57	3.20.20.70	0 (5)	79
1dqwa	267	57.41	3.20.20.70	1 (4)	80
1dqwa	267	57.41	3.20.20.70	1 (4)	81
1eixa	231	57.35	3.20.20.70	0 (5)	82
1qnoa	344	57.12	3.20.20.80	2 (3)	83
1psza	286	56.98	3.40.50.1980	0 (5)	84
1oy0a	248	56.73	3.20.20.60	0 (5)	85
1p0ka	306	56.69	3.20.20.70	0 (5)	86
1gs5a	258	56.56	3.40.1160.10	0 (5)	87
1k6ia	318	56.53	3.40.50.720	0 (5)	88
1p74a	267	56.22	3.40.50.720	2 (3)	89
1izca	299	56.22	3.20.20.60	1 (4)	90
1lf1a	296	55.92	3.20.20.80	1 (4)	91
1diaa	285	55.76	3.40.50.720	0 (5)	92
1diaa	285	55.76	3.40.50.720	0 (5)	93
1btma	251	55.43	3.20.20.70	2 (3)	94
1qpba	555	55.39	3.40.50.970	1 (4)	95
1b37a	459	55.36	3.50.50.60	2 (3)	96
1igwa	396	55.27	3.20.20.60	2 (3)	97
1jpma	359	55.25	3.20.20.120	1 (4)	98
1f89a	271	55.07	3.60.110.10	0 (5)	99
1jpdx	318	54.85	3.20.20.120	0 (5)	100

Table C-4-a. RAPTOR models for PGK1, ranking from 1 to 36.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1brla	340	78.91	3.20.20.30	0 (2)	1
1aj0	282	78.66	3.20.20.20	0 (2)	2
1b5ta	275	78.66	3.20.20.220	0 (2)	3
1f07a	321	77.89	3.20.20.30	0 (2)	4
1bif	432	76.27	3.40.50.300	1 (1)	5
1twsa	273	75.77	3.20.20.20	1 (1)	6
1psca	329	75.48	3.20.20.140	1 (1)	7
1vh7a	250	75.36	3.20.20.70	0 (2)	8
1k2yx	459	75.09	3.30.310.50	1 (1)	9
1h1ya	219	74.56	3.20.20.70	0 (2)	10
1pjba	361	74.06	3.40.50.720	1 (1)	11
1mla	305	73.86	3.30.70.250	1 (1)	12
1c4xa	281	73.77	3.40.50.1820	0 (2)	13
1e19a	313	73.60	3.40.1160.10	1 (1)	14
1fdya	292	73.24	3.20.20.70	1 (1)	15
1m5wa	242	73.15	3.20.20.70	0 (2)	16
1qpba	555	73.07	3.40.50.970	0 (2)	17
1qgoa	257	72.95	3.40.50.1400	0 (2)	18
1m9na	589	72.85	1.10.287.440	0 (2)	19
1brlb	319	72.43	3.20.20.30	0 (2)	20
1bf6a	291	72.09	3.20.20.140	0 (2)	21
1btma	251	71.92	3.20.20.70	0 (2)	22
1f0ka	351	71.73	3.40.50.2000	0 (2)	23
1b54	230	71.27	3.20.20.10	1 (1)	24
2lbp	346	71.27	3.40.50.2300	0 (2)	25
1ovma	535	71.21	3.40.50.970	0 (2)	26
1gqna	252	71.07	3.20.20.70	0 (2)	27
1ecea	358	71.06	3.20.20.80	0 (2)	28
1f6ya	258	70.09	3.20.20.20	0 (2)	29
1igs	247	69.49	3.20.20.70	0 (2)	30
1ak1	308	69.47	3.40.50.1400	1 (1)	31
1qtwa	285	68.50	3.20.20.150	0 (2)	32
1eucb	394	67.91	3.30.470.20	0 (2)	33
1sbp	309	67.89	3.40.190.10	0 (2)	34
1fp2a	345	67.67	1.10.10.10	0 (2)	35
1ba3	540	67.66	2.30.38.10	1 (1)	36

Table C-4-b. RAPTOR models for PGK1, ranking from 37 to 72.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1j1ia	258	67.25	3.40.50.1820	0 (2)	37
1jfla	228	67.24	3.40.50.1860	1 (1)	38
1gdha	320	67.18	3.40.50.720	0 (2)	39
1xyza	320	66.86	3.20.20.80	0 (2)	40
1a2oa	347	66.76	3.40.50.180	0 (2)	41
1toaa	277	66.70	3.40.50.1980	0 (2)	42
1vc4a	254	66.37	3.20.20.70	0 (2)	43
1ush	515	66.36	3.60.21.20	0 (2)	44
1abe	305	65.99	3.40.50.2300	0 (2)	45
1psza	286	65.90	3.40.50.1980	1 (1)	46
1b73a	252	65.83	3.40.50.1860	0 (2)	47
1jnda	400	65.82	3.10.50.10	0 (2)	48
1sfra	288	65.66	3.40.50.1820	1 (1)	49
1dlia	402	65.57	1.10.1040.10	0 (2)	50
1cnv	283	65.50	3.20.20.80	0 (2)	51
1yaca	204	65.46	3.40.50.850	0 (2)	52
1eepa	314	64.79	3.20.20.70	0 (2)	53
1puja	261	64.53	1.10.1580.10	0 (2)	54
1mxsa	216	64.39	3.20.20.70	0 (2)	55
1m65a	234	64.37	3.20.20.140	0 (2)	56
1qwka	312	64.13	3.20.20.100	0 (2)	57
1jpma	359	64.05	3.20.20.120	0 (2)	58
1o1za	226	64.03	3.20.20.190	0 (2)	59
1v93a	292	64.01	3.20.20.220	0 (2)	60
1iy8a	258	63.96	3.40.50.720	1 (1)	61
1qwga	251	63.71	3.20.20.70	0 (2)	62
2gbp	309	63.62	3.40.50.2300	0 (2)	63
1jcja	252	63.51	3.20.20.70	0 (2)	64
1geqa	241	63.42	3.20.20.70	0 (2)	65
1cp7a	274	62.83	3.40.630.10	0 (2)	66
1m1ba	291	62.83	3.20.20.60	0 (2)	67
1d2fa	361	62.76	3.40.640.10	0 (2)	68
1g3ua	208	62.58	3.40.50.300	1 (1)	69
1a50a	260	62.47	3.20.20.70	0 (2)	70
1nqka	345	62.24	3.20.20.30	0 (2)	71
1h1na	304	62.20	3.20.20.80	0 (2)	72

Table C-4-c. RAPTOR models for PGK1, ranking from 73 to 100.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1f12a	293	62.05	1.10.1040.10	0 (2)	73
1byka	255	62.04	3.40.50.2300	0 (2)	74
1dqwa	267	61.74	3.20.20.70	1 (1)	75
1bdb	267	61.73	3.40.50.720	1 (1)	76
1gd9a	388	61.70	3.40.640.10	0 (2)	77
1a8y	338	61.68	3.40.30.10	0 (2)	78
1l6wa	220	61.66	3.20.20.70	1 (1)	79
1dbta	237	61.19	3.20.20.70	0 (2)	80
1ojxa	250	60.81	3.20.20.70	0 (2)	81
1bdha	338	60.70	1.10.260.40	0 (2)	82
1k6ia	318	60.70	3.40.50.720	0 (2)	83
1qhwa	300	60.61	3.60.21.10	0 (2)	84
1fmta	308	60.52	3.10.25.10	1 (1)	85
1l8xa	355	60.45	3.40.50.1400	0 (2)	86
1rpxa	230	60.31	3.20.20.70	0 (2)	87
1euca	306	60.27	3.30.470.20	0 (2)	88
1n7ka	234	60.16	3.20.20.70	0 (2)	89
1nvma	340	60.05	1.10.8.60	1 (1)	90
1bhs	284	60.02	3.40.50.720	0 (2)	91
1bd0a	381	60.00	2.40.37.10	1 (1)	92
1b1ya	500	59.93	3.20.20.80	1 (1)	93
1o5qa	271	59.85	3.20.20.60	0 (2)	94
1ak5	329	59.72	3.20.20.70	0 (2)	95
1a04a	205	59.48	1.10.10.10	0 (2)	96
1dxha	335	59.46	3.40.50.1370	1 (1)	97
1ba2a	271	59.45	3.40.50.2300	0 (2)	98
1gv0a	301	59.20	3.40.50.720	1 (1)	99
1ujpa	243	59.15	3.20.20.70	0 (2)	100

Table C-5-a. RAPTOR models for ENO1, ranking from 1 to 36.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1jpma	359	218.65	3.20.20.120	2 (3)	1
1muca	360	217.43	3.20.20.120	1 (4)	2
1bqg	399	199.37	3.20.20.120	1 (4)	3
1mla	305	73.33	3.30.70.250	0 (5)	4
1nm2a	305	67.47	3.30.70.250	0 (5)	5
1a2n	418	62.64	3.65.10.10	2 (3)	6
2lbp	346	62.10	3.40.50.2300	1 (4)	7
1b57a	346	61.56	3.20.20.70	0 (5)	8
1psca	329	59.67	3.20.20.140	0 (5)	9
1f6ya	258	59.21	3.20.20.20	0 (5)	10
1btma	251	59.13	3.20.20.70	1 (4)	11
1v93a	292	58.66	3.20.20.220	0 (5)	12
1uuma	350	58.53	3.20.20.70	0 (5)	13
1o5qa	271	58.50	3.20.20.60	0 (5)	14
1gd9a	388	58.48	3.40.640.10	2 (3)	15
1b54	230	58.38	3.20.20.10	0 (5)	16
1fdya	292	58.37	3.20.20.70	0 (5)	17
1igwa	396	57.74	3.20.20.60	1 (4)	18
1bf6a	291	57.68	3.20.20.140	0 (5)	19
1f12a	293	57.22	1.10.1040.10	0 (5)	20
1gxba	339	56.47	1.20.970.10	0 (5)	21
1bpd	324	56.33	1.10.150.20	1 (4)	22
1azya	440	56.10	1.20.970.10	0 (5)	23
1qtwa	285	54.64	3.20.20.150	0 (5)	24
1jcja	252	54.39	3.20.20.70	0 (5)	25
1vh7a	250	54.28	3.20.20.70	2 (3)	26
1hqta	324	54.13	3.20.20.100	0 (5)	27
1f07a	321	54.05	3.20.20.30	2 (3)	28
1dora	311	54.04	2.30.26.10	0 (5)	29
1fp4a	467	53.94	1.20.89.10	2 (3)	30
1nvma	340	53.89	1.10.8.60	2 (3)	31
1igs	247	53.87	3.20.20.70	0 (5)	32
1jcqa	313	53.78	1.25.40.120	0 (5)	33
1d2fa	361	53.62	3.40.640.10	0 (5)	34
1ak1	308	53.61	3.40.50.1400	0 (5)	35
1jdna	407	53.54	3.40.50.2600	1 (4)	36

Table C-5-b. RAPTOR models for ENO1, ranking from 37 to 72.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1izca	299	53.32	3.20.20.60	0 (5)	37
1aj0	282	53.16	3.20.20.20	2 (3)	38
1dkra	298	53.06	3.40.50.2020	0 (5)	39
1e19a	313	52.75	3.40.1160.10	1 (4)	40
1qwka	312	52.48	3.20.20.100	0 (5)	41
1eu8a	407	52.39	3.40.190.10	1 (4)	42
1o4sa	375	52.29	3.40.640.10	0 (5)	43
1um9a	331	51.96	3.40.50.970	0 (5)	44
1pea	368	51.94	3.40.50.2300	0 (5)	45
1mo0a	257	51.82	3.20.20.70	0 (5)	46
1lf1a	296	51.68	3.20.20.80	0 (5)	47
1fcha	302	51.43	1.25.40.10	2 (3)	48
1a40	321	51.19	3.40.190.10	1 (4)	49
1rpxa	230	50.85	3.20.20.70	0 (5)	50
1tjua	448	50.72	1.10.40.30	4 (1)	51
1rdfa	263	50.53	1.10.164.10	0 (5)	52
1pfka	320	50.21	3.40.50.450	0 (5)	53
1m1ba	291	50.09	3.20.20.60	1 (4)	54
1aq0a	306	49.89	3.20.20.80	0 (5)	55
1umya	374	49.79	3.20.20.330	0 (5)	56
1ecxa	364	49.51	3.40.640.10	2 (3)	57
1o4ua	265	49.39	3.20.20.70	0 (5)	58
1jp4a	302	49.37	3.30.540.10	1 (4)	59
1qora	326	49.21	3.40.50.720	2 (3)	60
1qqea	281	49.16	1.25.40.10	0 (5)	61
1a50a	260	48.97	3.20.20.70	0 (5)	62
1s1pa	315	48.87	3.20.20.100	1 (4)	63
1ba2a	271	48.85	3.40.50.2300	0 (5)	64
2gbp	309	48.75	3.40.50.2300	0 (5)	65
1osna	325	48.74	3.40.50.300	0 (5)	66
1ehia	360	48.72	3.30.470.20	1 (4)	67
1dysa	345	48.69	3.20.20.40	0 (5)	68
1j93a	343	48.67	3.20.20.210	0 (5)	69
1vc4a	254	48.67	3.20.20.70	0 (5)	70
1a59	377	48.59	1.10.230.10	2 (3)	71
1h1ya	219	48.57	3.20.20.70	0 (5)	72

Table C-5-c. RAPTOR models for ENO1, ranking from 73 to 100.

Template	Template Sequence Length	z-Score	CATH Code	Number of Crosslinks (ranking)	z-Score Ranking
1hyea	307	48.42	3.40.50.720	0 (5)	73
1brla	340	48.26	3.20.20.30	0 (5)	74
1f0ka	351	48.24	3.40.50.2000	2 (3)	75
1c7na	394	48.23	3.40.640.10	0 (5)	76
1l8xa	355	48.19	3.40.50.1400	2 (3)	77
1m5wa	242	48.14	3.20.20.70	2 (3)	78
1huva	349	47.98	3.20.20.70	0 (5)	79
1dbta	237	47.96	3.20.20.70	0 (5)	80
1h4pa	408	47.94	3.20.20.80	1 (4)	81
1dn1a	556	47.57	1.20.1050.30	1 (4)	82
1cg2a	389	47.33	3.30.70.360	2 (3)	83
1eyya	504	47.32	3.40.309.10	2 (3)	84
1m65a	234	47.25	3.20.20.140	0 (5)	85
1c3ua	423	47.22	1.10.40.30	1 (4)	86
1m5ya	388	47.22	3.10.50.40	1 (4)	87
1qapa	289	46.95	3.20.20.70	1 (4)	88
1qgoa	257	46.79	3.40.50.1400	0 (5)	89
1o1za	226	46.75	3.20.20.190	1 (4)	90
1fsz	334	46.75	3.30.1330.20	0 (5)	91
1dida	393	46.69	3.20.20.150	2 (3)	92
1jeza	300	46.64	3.20.20.100	1 (4)	93
1brlb	319	46.61	3.20.20.30	0 (5)	94
1cbf	240	46.50	3.30.950.10	0 (5)	95
1eixa	231	46.30	3.20.20.70	0 (5)	96
1ohca	338	46.24	3.90.190.10	0 (5)	97
1rlza	344	46.19	3.40.910.10	2 (3)	98
1dqwa	267	46.12	3.20.20.70	1 (4)	99
1bxza	352	46.11	3.40.50.720	1 (4)	100