Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig–Maximilians–Universität München

Characterization of Protein Interactions by Mass Spectrometry and Bioinformatics

Victor Manuel Solis Mezarino
aus Lima - Peru

2019

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Dr. Franz Herzog betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 26.03.2019 .....

Victor Manuel Solis Mezarino

Dissertation eingereicht am: 08.02.2019

Erstgutachter: Dr. Franz Herzog

Zweitgutachter: Prof. Dr. Axel Imhof

Tag der mündlichen Prüfung: 20.03.2019

Contents

List of Figures

Summary

The characterization of physical and functional interactions between molecules is of vital importance in biology. It is vital because it improves our understanding of biological processes, their regulatory mechanisms and thereby their disease-associated malfunctions, impairments and disruptions. In this thesis, I focused on the study of protein-protein interactions (PPIs). Powerful experimental methods coupled with mass spectrometry (MS) have been developed to study PPIs. However, methods always have limitations, particularly of sensitivity and false discoveries. No method on its own is able to accurately reproduce the whole interactome of the subject under study, and thus bioinformatics tools that overcome such limitations and improve the capabilities of the methods are always in demand.

In this work, I have developed bioinformatics tools and pipelines for the interpretation and integration of MS-based PPI data. Additionally, I have broadened the applicability of chemical cross-linking followed by MS, through the incorporation of quantitative information during data analysis and modeling. As a result, the main contributions of my work have been on resolving protein interaction networks through data integration and on predicting protein binding interfaces and their affinities by chemical cross-linking and quantitative mass spectrometry.

The first chapter of my thesis gives a general introduction to the study of protein interactions and mass spectrometry based methods to discover, measure and characterize PPIs. This chapter provides the theoretical background for a clear understanding of the remaining chapters in this document.
The second chapter describes a bioinformatics tool, called compleXView, which I published during my doctoral work. This chapter shows that the combination of interactomics data obtained with different experimental methods improves the prediction of protein complexes in protein interaction networks, and that the incorporation of information from knowledge databases facilitates these predictions and the interpretation of the data.
The third chapter outlines a bioinformatics pipeline that combines sequence-level properties with the quantification of protein-protein cross-links to infer binding interfaces in macromolecular complexes. Three predictions that were experimentally validated are presented as proofs of concept. This chapter shows that the effective use of quantitative MS information in chemical cross-linking experiments allows the characterization of PPI binding interfaces.
Finally, the fourth chapter describes a bioinformatics method to estimate the dissociation constants of protein interactions through the quantification of protein-protein cross-links.

The applicability of the method is proven in a well-benchmarked trimeric complex and in a multimeric protein association. This chapter shows that the effective use of quantitative MS information in chemical cross-linking experiments allows the measurement of PPI binding affinities.

Overall, my work extends the applications of mass-spectrometry-based methods for the molecular characterization of protein complexes. The tools and concepts that were developed in this endeavor will help the scientific community with the study of protein interactions. As a result, we will improve our understanding of protein complexes and their vital role in biology.

Chapter 1

Theoretical Background

1.1 Proteins and the importance of their interactions

The name 'protein' is derived from the Greek word proteios that means 'the first', 'in the lead' or 'on the top'. The denomination of proteins as 'in the lead' is very appropriate. Proteins are one of the main building blocks that constitute a cell and a whole organism. They catalyze biological processes that take place within and outside cells and carry out the roles of signaling, kinesis, synthesis of molecules and their maturation, translocation and degradation. Moreover, they regulate these processes at different levels.

For executing these roles, physical associations between the same or different proteins are established. These associations are known as protein complexes and are built upon protein-protein interactions (PPI). The overall protein levels within the cell may remain relatively unchanged, even those of individual complex members, and yet, because of the formation or disassembly of certain protein complexes, biological processes can be initiated, modulated and terminated [30]. It is not surprising then that protein complexes constitute the primary targets of drugs of all kinds. The study of proteins and their interactions is fundamental for understanding the molecular mechanisms of diseases and opens doors for discovering new clinical markers and protein therapeutics.

Ever since their discovery, proteins have posed hard challenges to scientists. In part due to the difficulties of isolating them in good quantities, and due to the complex composition of proteins, their large diversity and their dynamism. Proteins were discovered in the 18th century and were first described in 1838, 30 years before nucleic acids. However, discoveries about the properties of the molecules themselves and the elucidation of their roles have not always been in the lead. The constitution of proteins by amino acids was first proposed in 1902, whereas DNA constitution by nucleotides was proven between 1885-1901. The first protein structures to be resolved were those of hemoglobin and myoglobin in 1958, whereas the structure of DNA was published in 1953. And even though Edman's method for protein sequencing was invented before Sanger's method for nucleic acids sequencing, the breakthrough of next-generation DNA sequencing technologies in the 1990s, allowed a faster and thorough study of genomes and transcriptomes, while the study of proteomes lagged behind.

Paradoxically, not being in the lead ended up being fruitful. The in-silico translation of gene sequences resulted in the creation of protein sequence databases for different organisms. These databases would later allow the high-throughput identification of numerous protein samples and whole proteomes by mass spectrometry from the 90s up to now.

Given the relevance of proteins and protein-protein interactions, the main purpose of my doctoral work was to characterize protein complexes through the bioinformatic analysis of mass spectrometry data obtained from protein interaction experiments. In this endeavor, I managed to characterize the composition of protein complexes and the binding interfaces and affinities of their protein members. Thus, the work presented in this thesis extends the applications of mass-spectrometry-based methods for the study of protein complexes. In the remaining parts of this chapter, a theoretical background is provided in order to allow the readers to understand each of the three contributions of my thesis.

1.2 Experimental methods to detect proteins and their interactions

A number of methods have been developed to detect and enrich protein complexes from cells. These methods are classified based on the main technology that supports them. In the proteomics field, liquid chromatography coupled with mass spectrometry (LC-MS) has become the main technology to analyze proteins. Mass spectrometry (MS) is a technology that analyzes molecules based on their mass and charge. Together with chromatography and ionization technologies, MS can separate proteins or peptides, ionize them, sort them based on their masses, quantify them and fragment them into smaller molecules. The masses of these molecules can be compared to the masses of the sequences stored in protein databases, and thereby provide an identification of the proteins in a sample. A typical workflow to identify proteins by MS is depicted in Figure 1.1. Proteins are first extracted and purified from the cell or any other biological sample. They are cut into peptides by proteases with cleavage specificity (e.g., trypsin). Peptides are separated by chromatography based on their hydrophobicities, charged by an ionization source, separated further by the mass spectrometer based on their mass to charge (m/z) ratios, quantified by a detector within the machine (MS1 spectrum). Some of these peptides are selected and then separately fragmented into smaller molecules, whose masses are again analyzed and stored in a spectrum (MS2 spectrum). The fragmentation of a peptide is controlled such that it occurs at the peptide bonds of the amino acid sequence. Thus, the MS2 spectrum of a peptide contains masses of its complete sequence and of its fragments. Peptide candidates from a sequence database can be then selected and be split in silico into fragment sequences. The candidate that shows the best match to the masses of the experimental spectrum is chosen as the identity of the peptide. Moreover, the quantification of peptides is also possible, because the mass spectrometer also records the intensity of the peptide ions before fragmentation (MS1 spectrum), which corresponds to the relative abundances of the peptides in the sample. The identity and intensity of the peptides are used then to infer the identity of the proteins and their quantities in the sample.

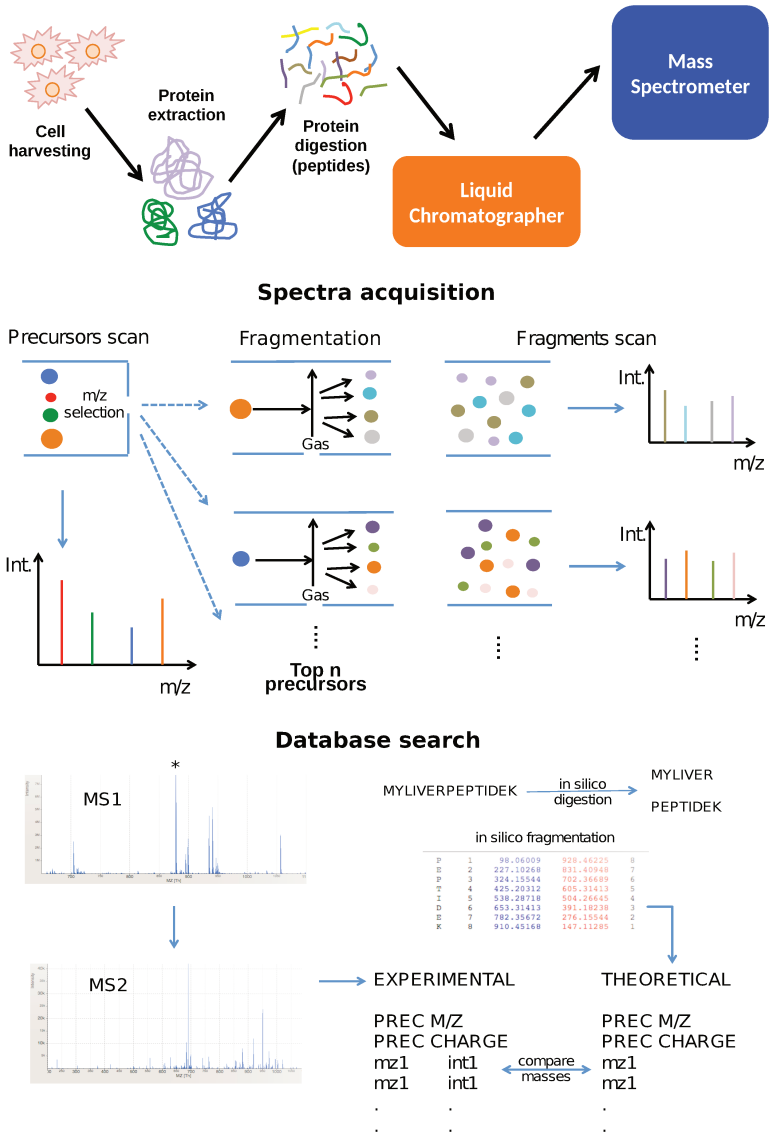


Figure 1.1: Typical experimental workflow to analyze peptides by MS. The workflow consists of 4 steps: Protein extraction and purification, protein digestion, peptide separation and spectra acquisition (LC-MS/MS), and database search (MS2 spectra identification).

The most widespread methods to enrich PPIs are coupled with MS in order to detect and quantify interactions. These methods have been reviewed elsewhere [99] and are depicted in Figure 1.2.

The first category of methods relies on the affinity of a protein to its interactors. An antibody against the protein is used in order to co-purify its interactors from the rest of the complex sample. In affinity-purification coupled to MS (AP-MS) workflows, the protein that is targeted by the antibody is called the bait and the co-purified proteins, the preys (Figure 1.2, left method). The antibody targets the bait itself or an epitope tag that was genetically engineered to the bait sequence. AP-MS methods can be used in

vitro and in vivo. Due to unspecific interactions with the tag or the antibodies, a negative control is always required. The control usually consists of a pull-down where the tag is linked to a decoy protein (e.g., GFP) that is used as bait. An alternative control is a pull-down where the real bait protein has been knocked down in the cell population. To increase specificity, some versions of AP-MS use two epitope tags in tandem, which allows for two consecutive purifications, and thus reduces the number of contaminant proteins [76]. However, only very stable complexes survive this procedure, but dynamic and weak interactions may pass undetected [61]. This would result in a loss of sensitivity that could be compensated for by cross-linking with formaldehyde in order to stabilize complexes previously to cell lysis or purification steps. Otherwise, a protein identified as an interactor in one co-purification experiment can be used as bait in another co-purification experiment. This strategy achieves a broader coverage of interactors, allowing the study of not only protein complexes but also protein interaction networks. AP-MS experiments require high amounts of input material, and thus methods with less starting material and higher sensitivity are needed. The use of nanobodies has to some extent solved the sensitivity and specificity issues because nanobodies have affinities for their epitopes in the sub-nanomolar range and contain a single antigen-binding domain [99]. On the other hand, AP-MS is less suitable for detecting integral-/trans-membrane proteins and their interactors. Membrane complexes are mainly involved in translocation and signaling processes, which means their interactions with other proteins is dynamic and often short-lived. Membrane complexes are relatively low abundant and hydrophobic and thus, their purification requires high sample amounts and harsh extraction conditions. All this leads to destabilization of the interactions, aggregation of proteins due to hydrophobicity and MS-signal suppression due to high lipid contaminations [73].

The second category of methods relies on chromatographic separation based on charge (Ion Exchange Chromatography, IEX) or size (Size Exclusion Chromatography, SEC; Figure 1.2, top method). Protein complexes have higher masses and higher charges than their individual subunits. Thus, different chromatographic fractions are selectively enriched with one or another protein complex or subcomplex. High-throughput fractionation of whole cell extracts is achievable and allows global profiling of PPIs [109, 27]. However, co-elution of non-related complexes may lead to the determination of false interactions. Hence, subcellular fractionation (e.g., into cytoplasmic and nuclear extracts) and other fractionation methods (e.g., sucrose gradient or isoelectric focusing) may be used along with SEC and IEX to decrease the complexity of the overall protein extract. SEC-MS allows the distinction between stable and dynamic interactions as well as the elucidation of the multiple complex memberships of a protein [44]. Stable complexes (or its core components) should have very highly correlated elution profiles across the fractions of an experiment. Stable complexes involved in dynamic, physical associations show multiple apexes in their elution profiles. So do proteins that belong to multiple complexes. Complexes that transiently interact with each other will only show high correlation in local regions of the elution range.

The third category of methods relies on the enzymatic modification of the interactors of a protein to facilitate their purification. These methods are called proximity ligation assays. They fuse a catalytic domain to a protein of interest (i.e., the bait), which then modifies its

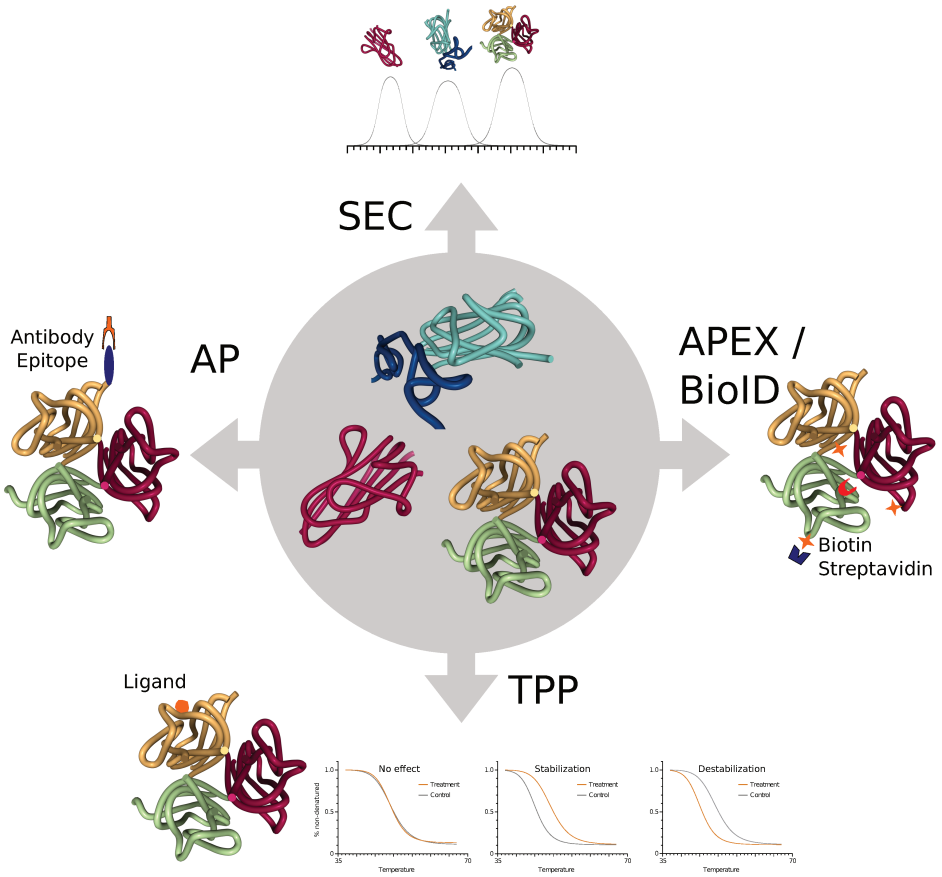


Figure 1.2: Experimental methods for the study of protein-protein interactions coupled to mass spectrometry. AP: Affinity purification. SEC: Size exclusion chromatography. APEX/BioID: ascorbate peroxidase/biotinylation proximity assays. TPP: Thermal proteome profiling.

interactors and other proximal proteins. Modified proteins are subsequently purified using the modification as affinity tag (Figure 1.2, right method). To avoid false identifications, the modification must be relatively uncommon in vivo. Biotinylation has this characteristic, and two methods have been developed to achieve this: APEX [75] and BioID [80]. In APEX the enzyme fused to the protein of interest is ascorbate peroxidase, whereas in BioID the enzyme is the promiscuous biotin ligase BirA*. Both enzymes catalyze the generation of biotin radicals that react with specific amino acids of proteins within a radius of action. The main advantage of APEX is that it is faster than BioID (minutes versus hours) and its radius of action is around 20 nm (twice wider than BioID). None of the methods is, however, able to discriminate between real interactors and proximal proteins. Moreover, the fusion of the catalytic domain to each terminus of the bait may be required in order to achieve good sensitivity. As in AP-MS, specificity is assessed based on the significance of the protein abundances relative to a negative control. The control consists of the catalytic

domain fused to GFP or to a peptide localization signal that puts the mock protein in the same subcellular location as the actual bait in the experiment. BioID is dependent on the availability/proximity of primary amines that are presented by lysine side chains and protein N-termini, whereas APEX is dependent on the presence of aromatic groups like side chains of tyrosine and phenylalanine. Compared to AP-MS methods, proximity ligation assays offer three advantages: i) apart from stable interactions, dynamic and weak associations are also detected, ii) interactors of membrane proteins can be probed as the purification does not depend on a stable bait-prey interaction but on the biotin modification of the prey, and iii) the cellular localization of biotinylated proteins can be observed with confocal microscopy before protein harvesting. A recent modification to the BioID protocol showed that BioID has another advantage over AP-MS. Because of its relatively short radius of action, enrichment of biotinylated peptides can be used to inform about the interfaces of direct protein-protein interactions [55].

The fourth category of methods relies on the correlated behavior of interacting proteins under a perturbation. Thermal proteome profiling (TPP) is the canonical method in this category (Figure 1.2, bottom method). TPP triggers the perturbation through temperature and has been used on ligand-complex stability assays [21]. At each temperature, proteins that are still soluble are quantified and their denaturation curves are plotted with these values. If the ligand (de-)stabilizes the protein, a shift should be observed on the denaturation curve of the protein respect to the control experiment, which consists of a cell culture without the ligand. Accordingly, interactors of the protein should also show a similar shift of their curves. The main limitation of this method is its inability to distinguish between physical and functional interactions [99].

Despite advances in these methods, none of them can distinguish between direct and indirect protein interactions neither can they resolve the topology of protein complexes. Protein cross-linking followed by mass spectrometry (XL-MS) can overcome these limitations. XL-MS uses a chemical cross-linker to covalently link residues spatially close to each other. Hence, the detection of a crosslink is a good indication for a direct protein-protein interaction. Following complex purification and protein digestion, cross-linked peptides are enriched by SEC or strong cation exchange (SCX) chromatography, and analyzed by LC-MS (Figure 1.3). Subsequent to peptide identification the resulting information restricts the plausible protein-protein interfaces to specific regions (e.g., domains, helices, etc.), thereby revealing the topology of the protein complex [50]. Due to the specific length of the cross-linker, only residues separated by a distance below this length are linked. This has made XL-MS an important source of distance information for refining structural models obtained by cryo-electron microscopy [87] and even by computational predictor tools [89]. Moreover, the quantification of cross-links has permitted the study of conformational changes within protein complexes [86, 108]. However, XL-MS may suffer from limitations due to the requirement of specific amino acid residues at the protein-protein interfaces, as well as the relatively low abundance of cross-linked peptides compared to linear ones. Recently, a modified method has been shown to be able to distinguish between intra-protein interfaces and homodimeric interactions [52]. However, proteins sharing a common interactor within different complexes cannot yet be resolved.

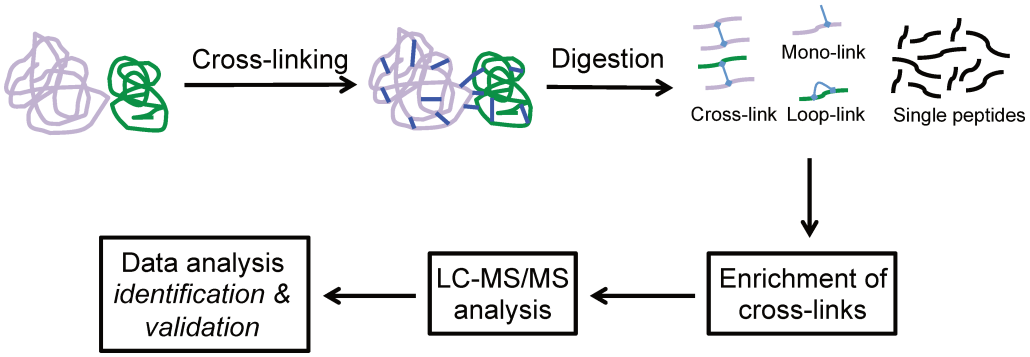


Figure 1.3: XL-MS workflow for the identification of putative protein interactions and their binding interfaces.

1.3 Quantification of PPIs using Mass Spectrometry

Distinguishing real interactors from false positives is not a trivial task. Unspecific binders and contaminants not always co-purify or co-elute in small amounts. Thus, a naive abundance ranking of the proteins identified within the sample will lead to false positive results. Comparing protein abundances in the sample versus the abundances in the negative control has proven to be already a better strategy [83]. The determination of abundances is also important for prioritizing interactors in follow-up experiments. Furthermore, it provides a good measure for estimating the stoichiometry of the components in a complex [98]. Quantification can reveal core components from adjacent ones: core subunits are generally more stable and thus can be detected according to protein abundances [99].

Different MS approaches to measure protein abundances have been developed. Quantification of proteins by MS can be performed by either label-free or labeled methods or by absolute quantification methods. Label-free methods measure peptides in their natural chemical occurrence with no modifications in the isotopic composition of the proteins and no incorporation of isobaric tags in the peptides. Label-free samples are obtained with the standard protease digest workflow (Figure 1.2). Cells are grown in a normal medium and harvested for protein extraction and digestion, which occurs in the absence of chemical tags. Peptides are analyzed in the mass spectrometer with a data-dependent acquisition strategy. As peptides elute from the LC, the most abundant ones at that time point are selected for fragmentation. Abundances are estimated from the peptide intensities in the acquired MS1 spectra.

There are basically two labeling methods used in MS-based interaction studies: SILAC (stable isotope labeling by amino acids in cell culture; [72]) and iTMT (isobaric tandem mass tags; [103, 12]). SILAC experiments (Figure 1.4 A) are performed on cell cultures where lysine and/or arginine sources are composed of either the light or heavy isotopic versions of these amino acids. It is common practice to use light medium for the negative control culture and heavy medium for the experiment. Pull-downs from the control and the experiment are performed separately and then combined in a 1:1 weight ratio. Peptides with the same molecular identity, but different sample origin can be distinguished

based on their mass difference introduced by the isotopically labeled amino acids. Specific interactors of the bait should have a significantly higher intensity in the real experiment relative to the control experiment. SILAC has been performed in AP-MS [90], in SEC-MS [44] and in proximity ligation assays [75].

The second labeling approach, iTMT facilitates the differential labeling of more than three samples in a given experimental design (Figure 1.4 B). This allows the experimenter to assess in the same experiment the effect of mutations in the bait [31] or of different perturbation conditions as in TPP. In iTMT, differential labeling occurs after protein digestion. Isobaric tags modify the amino groups at the N-terminus of the peptide and at the lysine side-chains. Digested labeled extracts are then pooled in a single tube and analyzed by mass spectrometry. As the tags are isobaric, peptides from different sample origins but with the same sequence identity accumulate in the same MS1 peak. This composite peak is selected for fragmentation under high collision-induced dissociation, which produces two classes of ions: peptide fragment ions and reporter ions from the cleavage of the isobaric tags. These reporter ions appear in the low m/z range at specific mass shifts from one another. Their intensities are quantified and used for measuring the relative abundances of the peptide in the different samples. The sequence identity of the peptide is derived from the fragment ions in the MS2 spectrum.

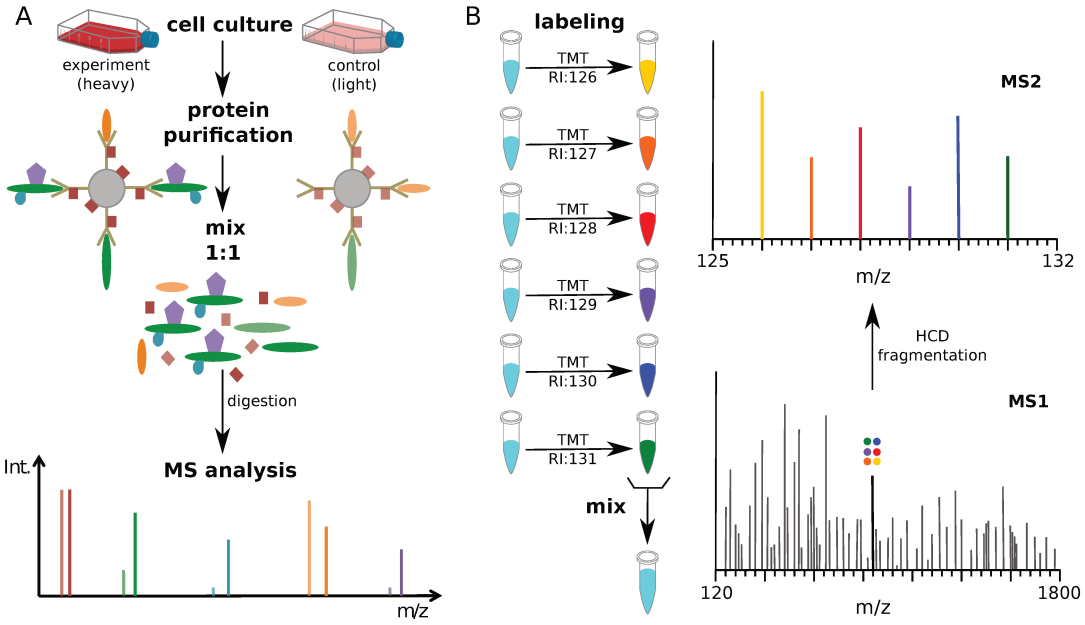


Figure 1.4: Isotopic protein labeling methods to quantify protein interactions by mass spectrometry: SILAC (A) and iTMT (B).

Absolute quantification aims to measure the molar amounts of the interactors with the help of spiked-in synthetic peptides and the use of multiple reaction monitoring (MRM; [85]). MRM is a targeted MS acquisition strategy that selects peptides for fragmentation based on their specific LC elution times and m/z values rather than their intensities. The sequences of the target peptides, as well as their elution times and m/z values, need

to be known prior to MS analysis. Thus, absolute quantification is not used directly for interactome screening, but for stoichiometry determination of discovered or known interactions and complexes. Heavy-labeled synthetic counterparts of these peptides are ordered and spiked into the sample at known concentrations. Their intensities are used for estimating the abundance of the non-labeled native peptides by comparison of peak areas [85]. Alternatively, synthetic peptides with no relation to the targeted peptides can be spiked in the sample at known and distinct concentrations to each other. The intensities of the spiked-in peptides are then used for learning the relation between MS intensity and the real abundance [7]. This relationship is then used to interpolate the intensities of the targeted peptides and inferred their amounts. While absolute quantification and targeted MS methods are more accurate to measure stoichiometries, relative quantification with label-free MS methods can yield satisfactory estimations of the ratios for the core components of a complex [99].

1.4 The inference of PPIs and protein complexes

Comparing protein abundances between the experimental cases and the control, and setting a minimum threshold for the abundance ratios is not enough to determine genuine interactions [69]. Statistical and computational models must be used in order to assess the plausibility and significance of a putative interaction [69]. Apart from the protein abundance, these models take into account the reproducibility of the observations, the specificity of the prey to co-purify with one or only certain baits, and the co-occurrence of preys across different purifications. These models can be classified based on two criteria. The first is the type of input data that they accept, which could be qualitative or quantitative. And the second is the type of interactions that they infer, which can be only bait-prey interactions or both bait-prey and prey-prey interactions.

Independent of those criteria, the input data is always a matrix, where each column represents a purification experiment or a negative control, and each row represents a protein. The content of the matrix cells are either all qualitative (i.e., in a binary format: 0 for absence and 1 for presence) or all quantitative (e.g., spectral counts or intensity). In the case of quantitative information, spectral counts indicate the number of MS2 spectra matched to the protein, whereas intensities indicate the sum of intensities of the corresponding MS1 peaks of the peptides. Both measures are normalized to account for the different number of peptides that the protein generates during digestion. Most modeling methods prefer the use of spectral counts despite intensity-based measures being arguably more accurate and representative of the actual abundance of a protein in the purification [19].

The output of processing the input matrix is another matrix where both, columns and rows, represent proteins. The cells in the matrix contain scores that indicate the plausibility of the protein in the row to interact with the protein in the column. Some methods only output bait-prey interactions and are called 'Spoke' models [4] because the output network graphically resembles spokes attached to the hub of a wheel. Other methods will output interactions between preys in addition to bait-prey interactions and are called 'Matrix' models [4]. Matrix models provide more connectivity, which translates in higher

sensitivity to detect interactions. As they incorporate more levels of inference, their specificity decreases. Spoke models, on the other hand, require a large number of purification experiments to achieve a good level of sensitivity. And these purifications must come from baits that share a relatively good number of preys in common. In principle, integrating data from different MS-based interactomics methods can overcome the limitations of the Spoke and Matrix models.

Even though some Matrix models could also process SEC-MS data, Correlation models and Machine Learning are preferred for this kind of data [44, 27, 42, 109]. As the assumption for SEC-MS is that proteins from the same complex elute in the same SEC fraction (Figure 1.2), they must be identified in the same MS runs. Nonetheless, interactions can be disrupted during SEC fractionation and result in shifts in the elution profiles of the complex components. Cross-correlation models account for this by taking the highest similarity between two elution profiles shifted from one another by a maximum number of SEC fractions. On the other hand, protein complexes that interact transiently with each other have multimodal elution profiles, which may be deconvoluted to detect them and also to find proteins with multiple complex memberships. Correlations on their own could be misleading because co-elution can be due to complexes with similar sizes but not really interacting. Therefore, a subset of known protein complexes can be searched in the data and used as the training set for a machine-learning (ML) model that, together with the elution correlations of the training complexes, will learn to discern true from possibly false interactions in the whole dataset.

Matrices output by 'Spoke', 'Matrix' and 'Correlation' models can be graphically represented as networks. Within these networks, nodes represent proteins, and the edges between them represent interactions. Edges do not have directionality and can contain weights that indicate the plausibility of the interaction. Networks have been used for a long time in scientific applications and plenty of research has been done in the fields of Mathematics, Computer Science, Physics and Biology. Network properties have been thoroughly studied [111, 6] and algorithms have been optimized to find and extract clusters from them [101]. Nevertheless, identifying complexes in a PPI network is not a trivial task. Biological networks have unique properties and protein complexes are a special case of clusters: physical interactions, and not only functional associations, define membership. Algorithms to predict protein complexes in PPI networks are classified based on two criteria: i) those that use only network topology information and ii) those that use network topology and additional biological information [101].

The first category of algorithms searches for highly dense regions of connections within the network. They either take an agglomerative or partitioning approach to discover clusters in the network. While some methods in this class will be strict about the membership of a protein to one cluster or another, others allow overlapping clusters by assigning fuzzy memberships. The Markov clustering (MCL; [17]) algorithm is a member of this category. MCL iteratively performs two matrix operations on the underlying matrix of the network: expansion and inflation. At each iteration, highly connected proteins are revealed clearer and clearer as a group, because their connectivity to other groups in the network decreases while the strength of the connections within the group increases. The algorithm stops when the operations change the underlying matrix no more, which results in the identification

of non-overlapping clusters in the network.

The second category of algorithms incorporates previous knowledge about complexes. This knowledge can be general or particular. A general property of protein complexes is the core-attachment categorization and organization of the components. The core-attachment principle states that proteins are either part of the core of a complex or simply attach themselves to the core to modulate its function or transiently interact with it [24]. Clustering methods that use this principle determine core proteins by the degree of common interactors between core members respect to all their interactors. Once cores are defined, a protein p outside the core is defined as an attachment component if it has interactions with at least half of the core components. Attachment proteins can belong to more than one core, but core proteins have unique membership. And two cores can interact without attachment proteins as intermediaries. Particular properties of protein complexes are related with their functions and cellular localization. Proteins within the same complex share a particular function. Clustering methods that guide themselves through this principle use databases such as Gene Ontology to incorporate functional annotations and cellular localization of proteins to improve the performance of complex prediction.

A relevant aspect of protein complex membership is to determine if the physical interaction between any two members is direct or indirect. For some complexes in the network, this might be already known and can be retrieved from structural databases like the Protein Data Bank (PDB). For others, incorporating data from XL-MS and two-hybrid assays can help to elucidate the answer. Thus, both methods can be used to annotate direct physical interactions within a protein network and infer the topology of protein complexes.

In chapter 2 of my thesis, I show that combining data from quantitative AP-MS or BioID experiments with XL-MS information increases the sensitivity and specificity of protein complex detection and allows the estimation of their stoichiometries.

1.5 Characterization of protein binding interfaces and their affinities

A complete characterization of a protein complex goes beyond determining its members and their stoichiometries. It is highly relevant to elucidate the affinity of the proteins to each other and the binding interfaces that establish the affinity. Studying binding interfaces is important because mutations in these regions can lead to diseases [23, 35]. Knowing the physical interfaces of the interacting proteins in a complex provides helpful information to understand the mechanisms of a disease. Similarly, binding affinities between proteins are very important to explain the formation of complexes and their ontology [39].

Knowledge about binding interfaces shows that these sites have properties that are relatively specific and distinct from other protein regions [117, 105]. Binding sites display high evolutionary conservation and low soluble surface area. Binding sites attract each other by physicochemical complementarity of their amino acids, such as hydrophobicity,

hydrogen bridges and electrostatic interactions. And the interaction is further governed by the shape and molecular flexibility of the binding pockets. Binding interfaces can be directly observed on resolved structures in PDB and have been deposited in databases such as EPPIC, SCOPPI and others.

However, for thousands of binary interactions and protein complexes, their interfaces remain unknown because their structures have not been resolved. In order to uncover their binding interfaces, experimental and computational methods have been proposed. Low-resolution experimental methods to predict binding interfaces include alanine scanning mutagenesis assays [64] and MS-based methods such as hydrogen/deuterium (H/D) exchange [71] and XL-MS [96]. On the other hand, a plethora of predictor software uses either homology-based structures or sequence-level properties to directly infer binding interfaces from protein sequences [114].

In chapter 3 of my thesis, I outline a binding interface predictor that combines sequence-level properties with qXL-MS to infer binding sites in dimeric and multimeric complexes. Three cases in which this strategy was employed are shown as proofs of concept.

A protein-protein interaction is a reversible chemical reaction governed by the concentration of the interacting proteins (law of mass action). The affinity of the binding interfaces attracting each other defines the strength of the interaction [39]. When the reaction reaches the equilibrium, the concentrations of the proteins in the free and bound states do not change any longer with time. At equilibrium, the affinity is inversely related to the dissociation constant (K_d) of the reaction. The K_d indicates the molar ratio of the two free proteins relative to the complex. A low K_d indicates strong affinity whereas a high one weak affinity. Classical technologies to measure K_d values include surface plasmon resonance (SPR), isothermal titration calorimetry (ITC), fluorescence polarization (FP) and fluorescence resonance energy transfer (FRET). Chemical proteomics approaches are less commonly used for this task. However, coupled to MS have been fundamental to perform binding assays between proteins and small molecules [94, 5]. Similarly, thermal proteome profiling (Figure 1.2) has been employed to estimate the affinity of proteins for drugs in a proteome-wide manner [84, 60]. And recently, Makowski et al. [57] presented a new method that uses iTMT-MS in binding assays to estimate the dissociation constants of nuclear proteins for specific DNA sequences and nucleosomes.

Apart from its use in protein interaction studies, XL-MS is mainly employed in structural approaches to reveal the topology and structural features of native proteins and protein complexes [1, 95, 89]. Nonetheless, further applications for XL-MS can be envisioned if quantitative information is taking into account. Quantitative cross-linking mass spectrometry (qXL-MS) has the potential to measure the dynamic cooperation of proteins in biological networks.

In chapter 4 of my thesis, I describe a qXL-MS approach to estimate the affinities of protein interaction assemblies of macromolecular complexes. The method estimates the amount of bound and unbound partners from the intra- and inter-protein cross-link intensities of protein complexes and subsequently calculates the Kd of the interaction. Its applicability is proven in the trimeric complex CNN1-SPC24/25 and in the multimeric complex PRC2 bound to its cofactors AEBP2 and JARID2.

1.6 Aim and Contribution

In summary, numerous experimental methods exist to study protein complexes by MS. The aim of my thesis was to improve these applications through data integration and quantification of protein interactions. The main contributions of my work are summarized by the following achievements:

i Resolving protein interaction networks through the integration of MS-based interactomics data;

ii Predicting binding interfaces through the combination of qXL-MS, sequence conservation and secondary structure prediction; and

iii Measuring protein affinities by qXL-MS.

Chapter 2

Integration of MS-based interactomics data to infer protein complexes in PPI networks

2.1 Introduction

Data integration is understood as the collection of data from different sources, which are then combined, re-analyzed and interpreted in the contextual information provided by each of the data sets [47]. The expectation of integrating data is to obtain novel insights and conclusions, which are shared again with the scientific community as new knowledge. Protein interactomics data is vast and is centralized by efforts such as PRIDE [106] and IntAct [41]. Most of them have been acquired via AP-MS experiments and Y2H assays. In recent years, however, methods like XL-MS and BioID have contributed largely to these repositories. The integration of these data may lead to the discovery of novel interactions and protein complexes.

Most integrative approaches that combine and re-analyze different data types use either a supervised or a semi-supervised approach. This means that previous knowledge is required to guide or supervise the discovery of new knowledge. In the context of protein interactions, supervised approaches rely on gold standard lists of protein complexes, and are usually applied on large networks (i.e., with thousands of proteins). A machine-learning (ML) framework can efficiently execute a supervised integration provided that the data set and the standard set share a significant subset of proteins. As nobody knows the most effective way of combining data from different sources, one expects that an ML algorithm would do this efficaciously. For medium-sized networks and relatively unexplored interaction networks, the overlap between the gold standard and the data may be poor. Thus, previous knowledge should be used as guidance and has to be weighted equally or lower than the experimental data. For these cases, a semi-supervised approach may be preferred, and manual validation and curation of interactions is very important.

Here, I aimed to establish a framework and a strategy to analyze and combine MS-based

interactomics data sets and to generate a unified view of their outcomes. As a result, I introduce a semi-supervised tool called compleXView [100], which infers pairwise protein interactions and complexes in small and medium data sets by integrating MS-based quantitative interactomics data with functional annotations. compleXView integrates AP-MS, XL-MS, BioID data and Gene Ontology (GO) functional similarities. The tool was published in Nucleic Acids Research, and a copy of the original article can be found in Appendix B.

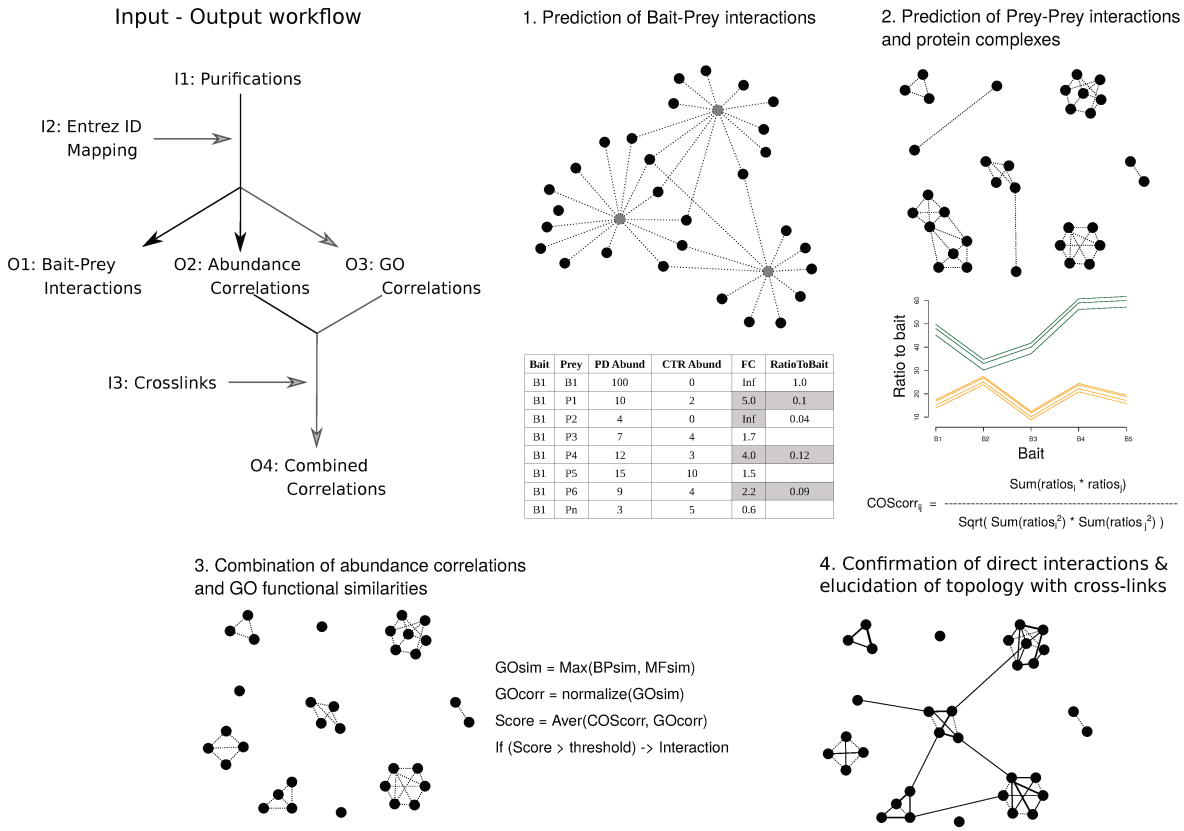
This chapter starts by describing the main idea behind compleXView and the visualization tools that it provides to validate protein complex members, estimate their stoichiometries and infer topologies. It demonstrates the applicability of the tool by analyzing two protein networks: the Protein Phosphatase 2A network and the Mini-chromosomal Maintenance complex and its interactors. The protein interactions and clusters discovered in both networks are discussed in detail.

2.2 Results

2.2.1 Analysis workflow of compleXView

The workflow of compleXView is shown in Figure 2.1. Unprocessed AP-MS or BioID interaction data is usually incomplete and noisy. Indeed, any experimental method produces false positives and negatives. Thus, in order to assess the signal over noise ratio, statistical methods estimate the probability of an interaction based on its abundance, reproducibility and specificity. compleXView measures the specificity as the enrichment of the prey protein in the purification relative to its abundance in the negative control. The reproducibility is accounted for by filtering out preys observed in less than N replicates and/or by penalizing absences. The abundance of the prey relative to the bait is taken as a decision criterion for the acceptance of a putative interaction. The significance of the remaining interactions is assessed either by a mixture probability model similar to SAINT [10] or a t-test. compleXView is built upon the assumption that the user knows better. Thus, it makes few assumptions about the experiment and is very transparent and flexible with its parameters and thresholds, even allowing the user to reduce the number of replicates for cases where a study is in an exploratory/pilot stage.

The first output of compleXView is a network of bait-prey interactions (Figure 2.1, step 1). The enrichment of the preys in the purifications can be assessed using blot plots or heatmap-colored networks (Figure 2.2 A and B). Medium sized bait-prey networks lack the connectivity between preys. Thus inferring higher order structures, such as protein clusters and complexes, is limited. Discovering functionally and physically associated proteins is highly relevant and can be learned from the data. To achieve this, compleXView correlates the abundances of the preys across the purifications (Figure 2.1, step2). This correlation approach is based on the fact that baits not only interact with individual proteins but also with protein complexes. If the direct interaction with one subunit in the complex changes, the abundances of the other members will change accordingly. Thus, one can expect high correlations in the co-variation of the abundances of proteins that belong



Bait	Prey	PD Abund	CTR Abund	FC	RatioToBait
B1	B1	100	0	Inf	1.0
B1	P1	10	2	5.0	0.1
B1	P2	4	0	Inf	0.04
B1	P3	7	4	1.7	
B1	P4	12	3	4.0	0.12
B1	P5	15	10	1.5	
B1	P6	9	4	2.2	0.09
B1	Pn	3	5	0.6	

$$COScorr_{ij} = \frac{Sum(ratios_i * ratios_j)}{Sqrt(\ Sum(ratios_i^2) * Sum(ratios_j^2)\ )}$$

Figure 2.1: Analysis workflow of compleXView for the inference of protein interactions and complexes. Bait-prey interactions are detected in co-purification experiments based on their abundance, reproducibility and specificity with respect to the negative control (1). Prey-prey interactions are predicted based on the correlation of protein abundances across the purifications of the different baits (2). Abundance correlations are combined with GO functional similarity information in order to improve the prediction of complexes in the network (3). Physical proximity of the predicted interactions is validated with cross-linking data (4).

to the same complex. Correlations above a specific threshold are fed to a Markov clustering algorithm (MCL), which uses them as interaction strengths to group proteins based on their local connectivity and interaction strength [17]. After clustering, compleXView outputs a network where clusters represent putative protein complexes or functionally related groups.

Functionally unrelated proteins could exhibit high correlations by simple chance. Thus, GO functional similarities between proteins are incorporated into the analysis workflow. Moreover, functionally related proteins do not necessarily interact, as they can come from different complexes or pathways, or their interaction does not occur in the cellular context or conditions of the experiment. Thus, compleXView combines the abundance correlation and GO similarity of a protein pair into a single score by average, which can be performed

with the same or different weights. The combined scores of all protein pairs are fed to the MCL algorithm, and the resulting clusters are visualized as a network. Clusters in this network represent functionally or physically associated proteins supported by the AP-MS data and previous knowledge.

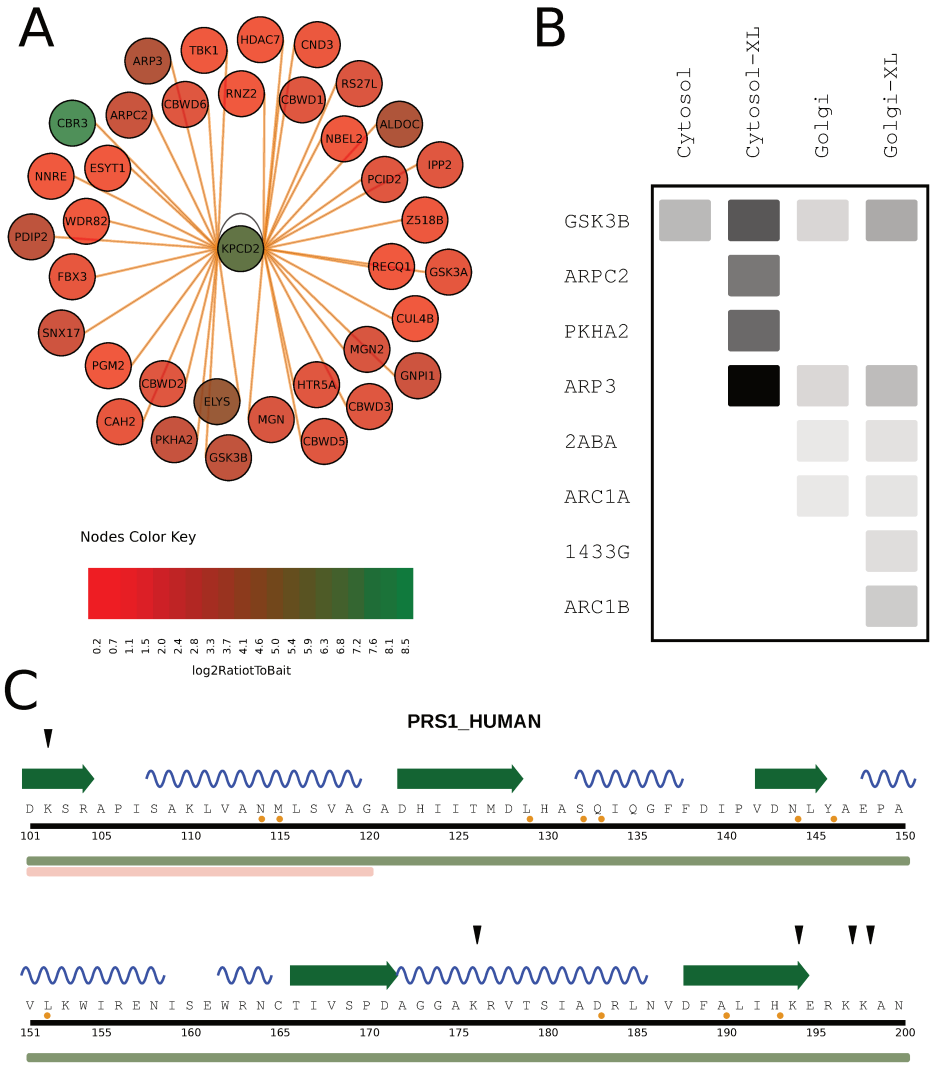


Figure 2.2: compleXView visualizations enable the evaluation of prey abundances within a single bait purification (A) or across different purifications (B). Blot plots like the one showed in B allow for the estimation of the subunit stoichiometry of a complex. Protein cross-links obtained from XL-MS data are depicted in plots annotated with functional information retrieved from UniProt and InterPro (C), allowing the inference of the minimal binding domains of a protein interaction by inspecting the proximity of inter-protein cross-links (black triangles).

To confirm physical connectivity XL-MS and Y2H data may be included in the analysis. The advantage of providing cross-linking data is that the elucidation of the topology and

the binding interfaces of the interactions can be revealed by further in-depth analysis of the distance restraints and the proximal secondary structure domains (Figure 2.2 C and Chapter3).

compleXView provides visualization tools to facilitate the interpretation of the networks generated in each of the data integration steps. The tools allow the discovery of single binary interactions, protein complexes, their topology, stoichiometry, and binding interfaces in a single framework.

In order to validate the applicability of compleXView, I tested the software on two datasets that were obtained by AP-MS, BioID and XL-MS. These datasets comprised the protein interaction networks of the Protein Phosphatase 2A and the Minichromosome Maintenance complex, respectively. The data was acquired by other authors and is publicly available on the PRIDE server. The analysis with compleXView reproduced the findings of the respective publications [29, 15] and was able to provide further insights. In the following sections, the results obtained from the first dataset are presented as an example. The results of the second data set are discussed in Appendix A.

2.2.2 The Protein Phosphatase 2A complex and its regulators

Protein Phosphatase 2A (PP2A) is a large protein complex that dephosphorylates proteins in a multitude of signal transduction pathways. PP2A acts on serine/threonine (S/T) residues and thereby plays the antagonistic role of S/T kinases like cyclin-dependent kinases or polo-like kinases. Together with protein phosphatases of type 1, PP2A is responsible for more than 90% of the S/T phosphatase activity in the cell [74]. Thus, it is frequently associated with a plethora of clinical implications where mutations cause misregulation of its activity. Studying the interactome and substrates of PP2A is therefore of high relevance for the molecular understanding of these diseases.

Class	Members
B	2ABA, 2ABB, 2ABD, 2ABG
B'	2A5A, 2A5B, 2A5D, 2A5G, 2A5E
B"	P2R3A, P2R3B, P2R3C
B"'	STRN1, STRN3, STRN4
Adapters	ANKL2, DAB2P, EST1A, IER5, SMG5/7
Biogenesis	IGBP1, LCMT1, PPME1, PTPA
Inhibitors	AN32A, AN32E, ARP19, CIP2A, ENSA, F122A, IEX1, PA216, PPR1A, PPR17, SET, TIPRL
Retainers	MFHAS1
Activators	NXN

Table 2.1: PP2A canonical regulatory subunits, other activators and inhibitors

Structurally, the PP2A holoenzyme is composed of 3 subunits: a catalytic subunit that encompasses a serine/threonine phosphatase (also called C subunit), a scaffold subunit

(also called A subunit) and a regulatory subunit (also called B subunit). However, about one third of PP2A in the cell is in a dimeric state, consisting of a catalytic protein and a scaffold protein [74]. The C protein exists either in the alpha isoform (PP2AA) or in the beta isoform (PP2AB) within a protein phosphatase complex. Despite their high sequence similarity, the substrates of the C isoforms are apparently not redundant. The A protein is also represented by either one of two isoforms, either 2AAA or 2AAB, which are as well not redundant in their function. The C-A dimer can further associate with B regulatory proteins in a mutually exclusive way, which modulates the subcellular localization, activity and substrate specificity of PP2A [74].

PP2A regulatory proteins are classified into the following families: B, B', B" and B"' (Table 2.1). In addition, many other regulatory proteins have been identified including activators and inhibitors. The way they modulate PP2A, in terms of subcellular location and substrate specificity, is not fully understood.

In this first analysis by compleXView, I integrated MS-based interactomics data from AP-MS and XL-MS experiments [29]. This data is publicly available at `https://xvis.genzentrum.lmu.de/compleXView/docs/PP2A_dataset/RAW`. The results showed a series of regulatory proteins and substrates of the PP2A network and allowed a detailed characterization of their associated complexes.

Bait-prey interaction network of the PP2A data

In order to identify PP2A interactors, a series of AP-MS pull-downs of the catalytic, scaffold and canonical regulatory subunits of PP2A (see Materials and Methods) were analyzed, and the preys were filtered by an FDR of 0.05 and a minimum abundance relative to the bait of 1%.

In agreement with the literature the catalytic subunits PP2AA and PP2AB did not co-purify one another. Nonetheless, they shared a large number of regulators and substrates more than their unique interactors (Figure 2.3). This indicates that their activities may overlap in these cases, but are not fully redundant. Isoform B exhibited more unique interactors than isoform A. Ribosomal proteins and cell-cell adhesion proteins were preferentially pull-downed with PP2AB. Similarly, the prefoldin complex was one of PP2AB main interactors. Prefoldins are chaperone proteins mainly localized in the nucleus and the mitochondrion. RNA polymerase II subunits co-purified preferentially with PP2AA as well as some ribosomal subunits. Protein 2AAB was the most abundant scaffold subunit that associated with the two phosphatases. Isoform A of the scaffold proteins did not pass the thresholds of either fold change to the negative control or minimal abundance relative to the bait. 2ABD was uniquely detected with the PP2AB bait, whereas the remaining regulatory subunits were co-purified with any of the phosphatases. Among the shared interactors were also the TCP cytoplasmic chaperons, liprins and integrator proteins. Overall, the data show the broad implication of PP2A in basic cellular tasks that occur in different subcellular locations, such as transcription, translation, protein folding and cell adhesions.

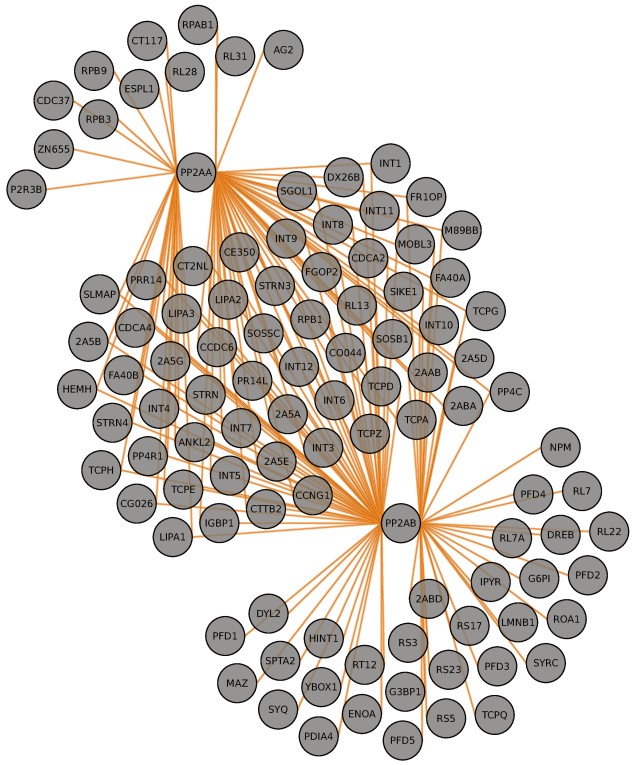


Figure 2.3: Prey interactions of the PP2A catalytic subunits PP2AA and PP2AB. Proteins were considered interactors of the baits if their relative abundance to the bait >2% and >2-fold relative to the negative control with an FDR <5%.

Regulatory subunits of PP2A are supposed to confer substrate specificity and cellular localization to the phosphatase catalytic subunits [74]. Accordingly, the analysis showed that the number of shared interactors between regulatory subunits was smaller than the unique ones (Figure 2.4). The data suggested that the regulatory subunit 2ABG might attribute substrate specificity for TCP chaperones and proteins involved in spindle assembly; 2ABA for a set of transcription-involved proteins and mitochondrial proteins; 2A5E for a set of regulators of autophagy; whereas regulatory subunit 2A5G and D might provide specificity for liprins and importins.

IGBP1, a non-canonical regulatory subunit and PP2A stabilizing factor, was also used as bait in the AP-MS data set. Previous work has proposed a PP2A biogenesis model where IGBP1, TIPRL, TCP proteins and prefoldins cooperate [91]. Accordingly, the data suggested that IGBP1 might associate with TCP proteins during the biogenesis and folding of PP2A and the phosphatases PP4C and PP6 (Figure 2.5). TIPRL, an inhibitor of PP2A, was co-purified in this pull-down, suggesting that the binding domains of these two regulators do not overlap and their roles might be coordinated. Indeed, Smetana & Zanchin [97] report that TIPRL binds to residues 210-309 of PP2A, whereas Jiang et al. [33] report that IGBP1 binds to residues 1-153. Furthermore, Wu et al. [112] have shown the orchestrated operation of TIPRL and IGBP1 in the dynamic assembly-disassembly process of PP2A through the interaction with the C subunits.

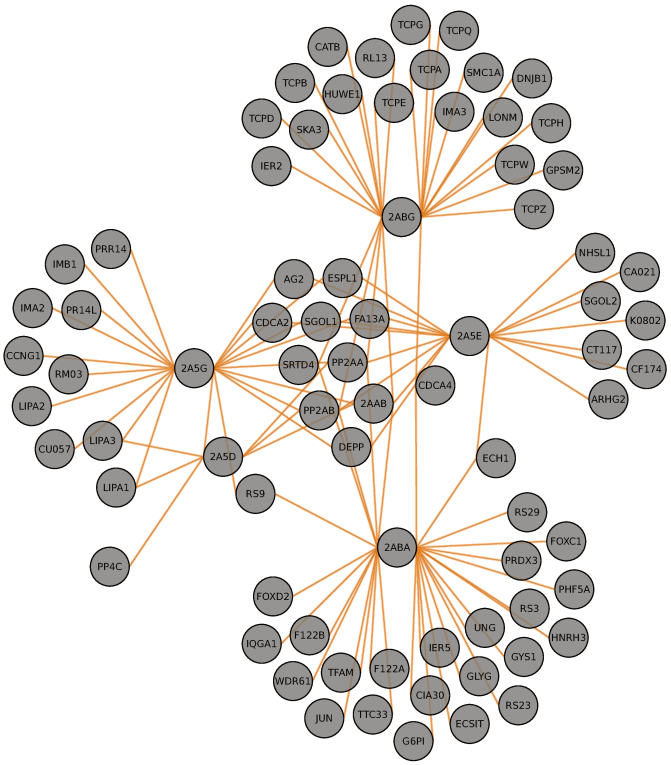


Figure 2.4: Protein interactions of the PP2A canonical regulatory subunits show that they share a relatively low number of interactors.

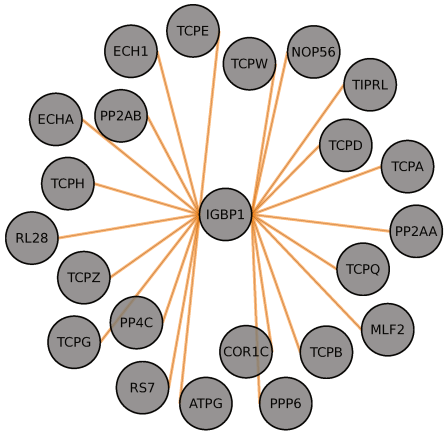


Figure 2.5: Prey interactions of the IGBP1 bait show that TCP proteins are the most numerous interactors of IGBP1 and that they co-purified with PP2A catalytic subunits.

Another bait protein was the catalytic subunit of the protein phosphatase 4 complex, PP4C. The main interactors of this protein were the TCP chaperones, its canonical regulatory proteins and IGBP1 (Figure 2.6). TCP proteins were in the range of 55-60 percent respect to the amount of PP4C. The low abundant interactors may be substrates or adapter proteins. As stated before, AP-MS has limited sensitivity for transient and weak

associations such as enzyme-substrate interactions [43]. Dephosphorylation is relatively transient, and thus it is more plausible that low abundant proteins represent substrates of the phosphatases; in particular, those preys that have known or putative phospho-serine/-threonine sites. In the case of the PP4C pull-down, these include ribosomal proteins and metabolic pathways proteins. On the other hand, those with no putative phosphosites might simply act as adaptors or enzymes that regulate/modify PP4C, e.g. CUL1, a core component of E3 ubiquitin-protein ligase complexes.

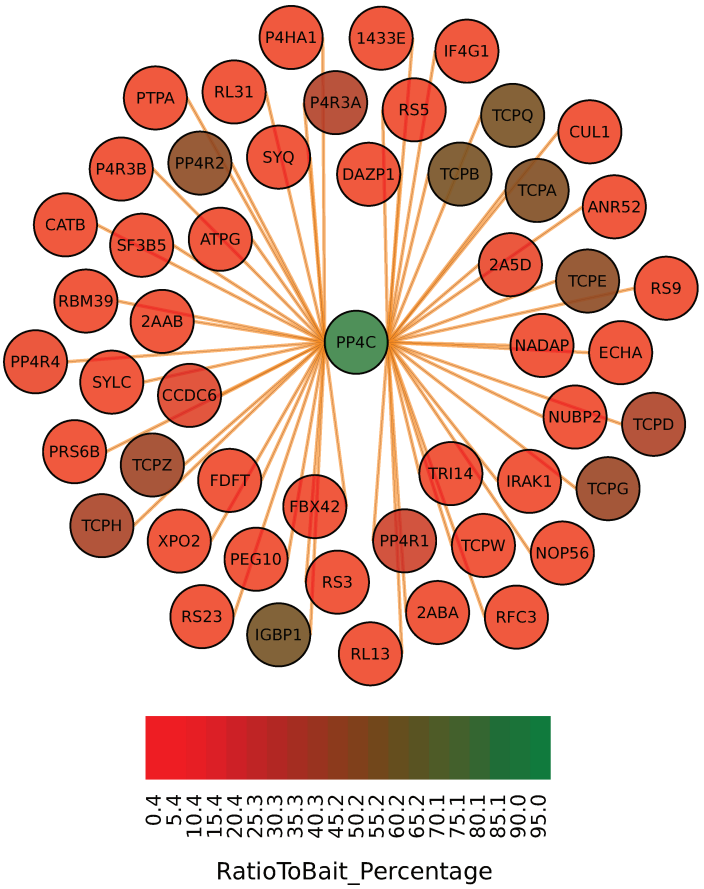


Figure 2.6: TCP proteins and IGBP1 are the most abundant interactors of the PP4 phosphatase catalytic subunit.

Detection of protein complexes in the PP2A network

As shown above, bait-prey interaction maps are helpful to detect interactors of specific bait proteins, as well as to identify interactors shared between baits. In addition, it is important to also detect putative protein complexes in the dataset. One way to achieve this in AP-MS data is by correlating the abundances of the preys across the purifications. This is based on the assumption that preys that are part of the same complex should show similar abundance variations across the pull-downs. Using these correlations, one

can create a denser protein-protein interaction map where the edge between two proteins indicates a high probability of being part of the same complex. A high correlation in the next analyses means a value above 0.8 within the range of 0 to 1. compleXView provides links to the UniProt database as a mean to curate clusters in the networks. In the remaining of this chapter, when no citation is provided, the information was retrieved from UniProt.

Clustering preys based on the correlation of their abundances across purifications identified protein complexes such as the TRiC, Integrator and Striatin (Figure 2.7). A closer look using compleXView blotplots reveals that the members of these protein complexes have highly correlated variations across the PP4C, PP2AA, PP2AB, IGBP1 and 2ABG purifications (Figure 2.8). Moreover, with the blotplots, one can estimate the stoichiometry of the components within the complexes. For example, TRiC subunits exhibited abundance ratios between 0.75 and 1.33. Integrator subunits 9 and 11 were twofold enriched over subunits 4, 5, 6, 8 and 12, and were up to 3 times more abundant than subunits 1, 3 and 10. In the case of striatin proteins, STRN3 was two times more abundant than STRN4, but similar to STRN. These ratios are in good agreement with the literature. The TRiC complex is assembled upon the stack of two hetero-oligomeric rings, each constituted by the eight TCP proteins [36, 51]. In the case of STRN3, mutational studies have revealed that its homo-dimerization is essential for its interaction with PP2A [9]. For the Integrator nothing is known about is stoichiometry. Nevertheless, it is known that in Drosophila, INT9 and INT11 have the nuclease activity required for the processing of snRNAs, whereas INT3 and INT10 are dispensable for this process [18]. Overall, this shows the capability of label-free AP-MS quantification to estimate the stoichiometries within protein complexes.

Previously known interactors of these complexes were also detected. For example, the associations of the Integrator with SOSB1, SOSSC ANKL2 and RNA polymerase II subunits, and the interaction of striatins with CTTB2, CT2NL and a set of kinases. The latter represents the STRIPAK complex [32] without the phosphatase PP2AA, which is also proper of this complex.

Additional complexes, such as prefoldins and liprins, were clustered with some apparently spurious interactors. Prefoldins are chaperone proteins that localize to the nucleus and the mitochondrion (source: UniProt) and associate distinctively with phosphatase PP2AB but not PP2AA (see previous section and [25]). Liprins are important for the disassembly of focal adhesions between the cell and the extracellular matrix (source: UniProt). They regulate the association of tyrosine phosphatases type 2A with their extracellular substrates by localizing these enzymes to the cell membrane [92, 93]. The presence of other proteins in each of these two clusters may be spurious given the lack of functional relation between them with prefoldins and liprins, respectively. Another distinct cluster in the network is the PP4 complex including its catalytic and regulatory subunits.

The remaining clusters and the large groups in the network of Figure 2.7 may represent proteins highly correlated due to chance. They can be partially resolved by applying a second iteration of the MCL algorithm or by grouping them based on their Gene Ontology similarities. compleXView calculates GO similarities in the whole network and detect functional groups present in the dataset. compleXView uses annotations from the

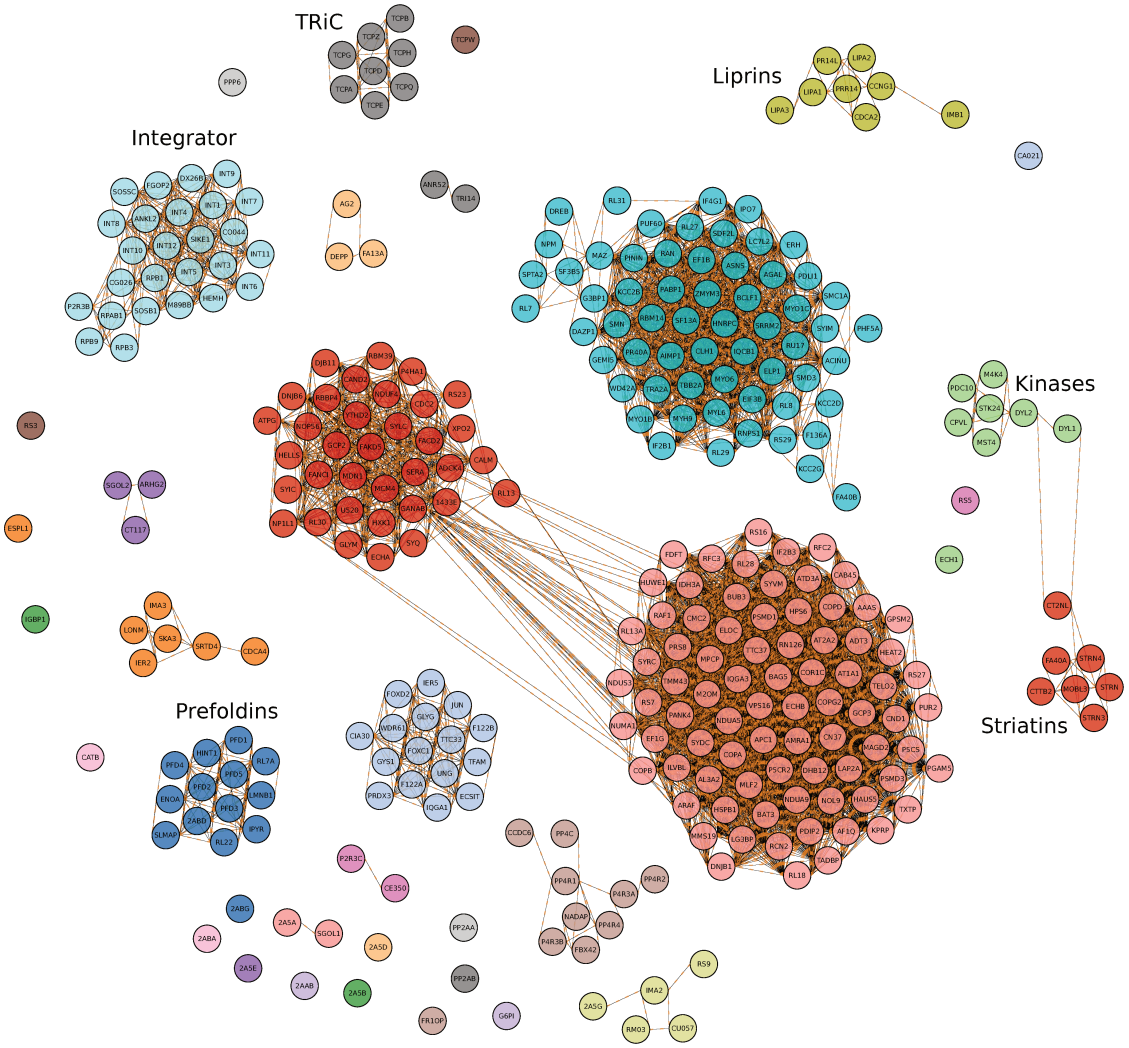


Figure 2.7: Complexes identified in the PP2A data set by correlation of abundances and MCL clustering.

Molecular Function and Biological Process categories of the GO database and calculates a similarity score (see Materials and Methods). Proteins with a similarity above a threshold of 0.6 (in the range from 0 to 1) are kept and used to cluster the proteins.

Like abundance correlations, GO similarities also spotted the TRiC complex (Figure 2.9). Whereas the integrator complex clustered with its interactor SOSB1, the subunits DX26B (INT6L) and M89BB (INT13) did not appear in the same cluster. Striatin proteins formed a group together with CT2NL, which linked them to PP2A. Liprins formed a group of their own, whereas prefoldins appeared in different functional clusters. The PP2A catalytic subunits clustered together with their specific regulators and with some regulatory subunits of the PP4 phosphatase, which clustered separately with its regulatory subunit PP4R2.

GO similarities detected other functional clusters. For example, ribosomal proteins had

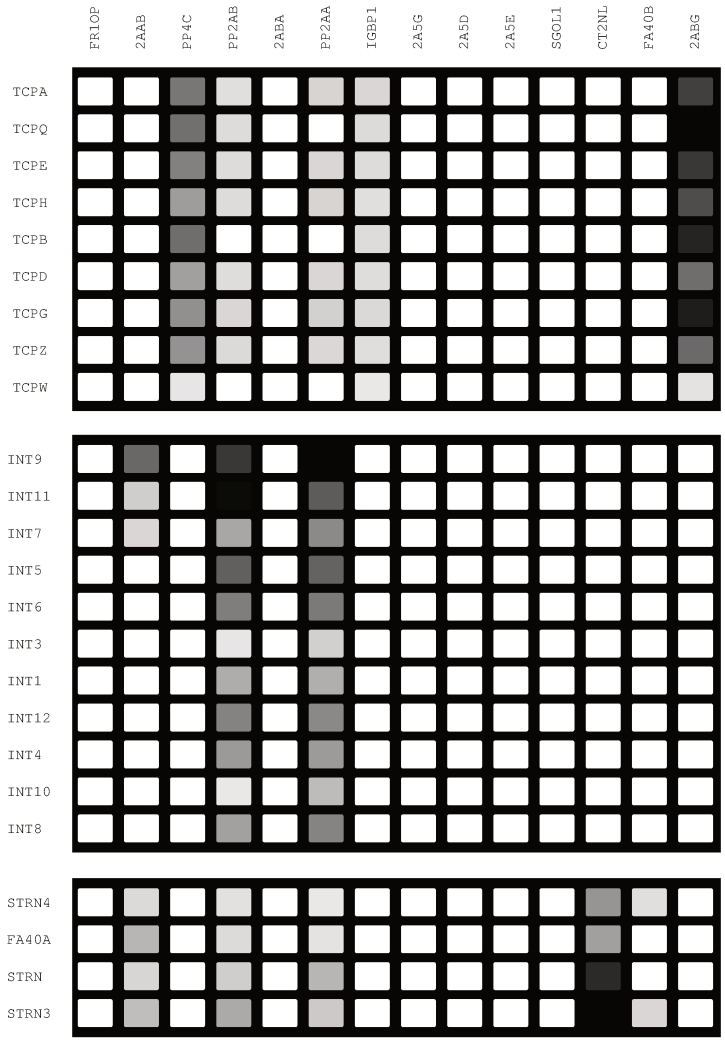


Figure 2.8: Correlation of relative abundances between the members of the TRiC complex, the Integrator and Striatins across purifications. Abundances are computed relative to the bait and increase from light grey to black. In the online implementation, exact abundance values can be observed upon hovering over each cell.

their own group. MDN1 and NOP56 clustered with NOL9 and CU057, which are all involved in the maturation of ribosomal subunits (source: UniProt). Amino acid tRNA ligases also grouped together, so do RNA polymerase II subunits, importins, gamma-tubulin components 3 and 2, replication factors 2 and 3, mitochondrial complex I proteins, serine/threonine-protein kinases (RAF1, ARAF, STK24 and 26), dynein light chains 1 and 2, proteins involved in lipid metabolism (ECH 1, A, B and AL3A2), proteins involved in transcription regulation (RBM39, PININ, PHF5A, and PUF60), proteins involved in mRNA splicing (SRRM2, SF3B5 and TRA2A) and spliceosome synthesis (SMN and SMD3), translation factors (EIF3B, IF4G1 and GEMI5, and EF1G and EF1B), 26S proteasome regulatory subunits (PRS8, PSMD1 and 3), NADH dehydrogenase complex (NDU proteins and CIA30), and the anaphase-promoting complex component APC1 to-

gether with BUB3 (a regulatory protein of this complex).

compleXView takes into consideration the information content of the GO terms during the assignment of GO similarity scores. For example, GO terms with general semantic, like 'protein binding' and 'ATP hydrolysis', are weighted weaker than more specific terms, like 'cytokine binding' and 'DNA helicase'. In spite of this, some proteins will still group due to the generality of their GO terms. A number of clusters in the network of Figure 2.9 show this. For example, the cluster formed by BCLF1 and PDIP2 grouped because they are DNA binding proteins. The cluster formed by ELOC and HUWE1 was due to their involvement in proteasomal degradation (source: UniProt). Proteins FA40B and PR40A clustered due to their involvement in cytoskeleton organization and cell migration. Proteins FOXC1 and IER2 are DNA binding proteins that act as transcription factors, but they are not necessarily involved in the transcription of the same genes (source: UniProt). Proteins ESPL1 (caspase-like protease separin) and CATB (26s proteasome regulatory subunit) appeared in the same cluster due to their involvement in proteolysis, but they are indeed from different pathways. Similarly, SMC1A, ATD3A, ADCK4 and PANK4 were grouped simply due to their ATP binding activity.

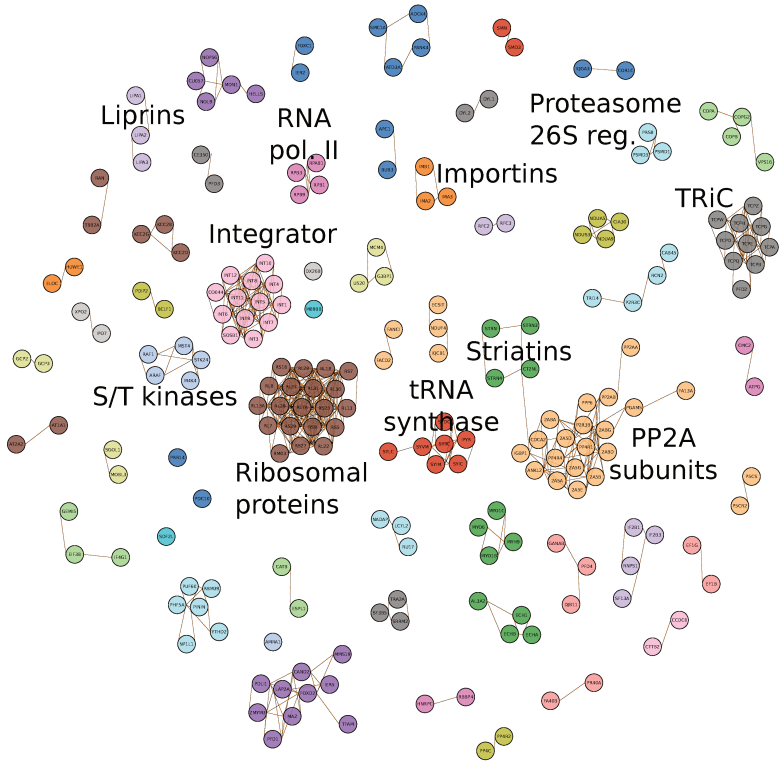


Figure 2.9: Clusters identified in the PP2A dataset by GO functional similarities and MCL clustering.

In order to improve the clusters inferred solely by either the Abundance Correlation score or the GO Similarity score, compleXView combines both indexes into a single score. In the PP2A dataset, protein-protein interactions with a combined score above 0.65 were

used for MCL clustering. The combined score increased the recognition of clusters and improved the membership plausibility of the proteins within each group (Figure 2.10).

SGOL1 clustered now with PP2A regulators. SGOL1 is required for proper chromosome segregation during mitosis and this function requires the interaction with the PP2A complex (source: UniProt). In the network, it was found also associated with the regulatory subunits 2A5A and 2A5B. Prefoldin 3 did not associate with CE350 any more. On the contrary, it grouped again with other prefoldins. RBBP4 did not associate anymore with HNRPC (a heterogenous ribonucleoprotein). RBBP4 is a histone-binding protein important for chromatin assembly and remodeling (source: UniProt). It clustered now with other chromatin/DNA-binding proteins, but also with tRNA ligases. Separase ESPL1 appeared now associated with 26S proteasome subunits and through this to the anaphase-promoting complex subunit APC1 and BUB3. LC7L2 and RU17 did not associate with NADAP any more, but as expected with other RNA binding proteins that are also involved in mRNA processing and splicing.

Additional clusters were also detected in this network constructed with the combined scores. For example, the cluster formed by NUMA1 and GPSM2 is known to interact in order to regulate the recruitment of the dynein-dynactin complex to the mitotic cell cortex regions, which is important for correct spindle orientation (source: UniProt). In the cluster formed by AMRA, BAT3 and BAG5, the latter two proteins are a chaperone and a chaperone-regulator, respectively. Their association with AMRA may be due to their involvement in apoptosis, while AMRA is involved in autophagy. In the cluster formed by AIMP1 and the t-RNA ligase SYIM, AIMP1 is an interactor of tRNAs and a component of the tRNA multisynthase complex. HAUS5 clustered with gamma tubulins, which are required for microtubule nucleation at the centrosome (source: UniProt). HAUS5 is a component of the HAUS augmin-like complex, which contributes to the assembly of the mitotic spindle, centrosome integrity and completion of cytokinesis (source: UniProt). In the cluster formed by TXTP, M2OM, ADT3 and MPCP, all the proteins are involved in the transport of metabolites from the cytoplasm to the mitochondrion. TXTP exchanges citrate/malate, M2OM exchanges oxoglutarate/malate, ADT3 exchanges ADP/ATP, and MPCP imports phosphate groups (source: UniProt). In the cluster formed by WDR61, FOXD2, IER5, TTC33 and F122A, the first three proteins are involved in transcription regulation. TTC33 function is unknown, but its tetratricopeptide repeat motif and its presence in this cluster suggest that it is involved in the control of transcription. F122A function is also unknown, and no domains are annotated for the protein. The cluster formed by FOXC1, JUN and TFAM is due to their role as transcription factors. While FOXC1 and JUN express uniquely in the nucleus, TFAM predominantly exists in the mitochondrion (source: UniProt) but has been found to express in the nucleus too [49]. The green cluster in Figure 2.10 contains a number of mRNA-binding proteins involved in different stages of transcription and splicing. LC7L2 biological function is unknown, its presence in this cluster might indicate a role during transcription or mRNA splicing. Additionally, three proteins involved in mRNA translation were present in this cluster: PABP1, IF2B1 and HNRPC (source: UniProt).

Surprisingly, the PP2A subunits remained separated in this network. This is mainly due to the low abundance correlation between its subunits. Because the composition of the

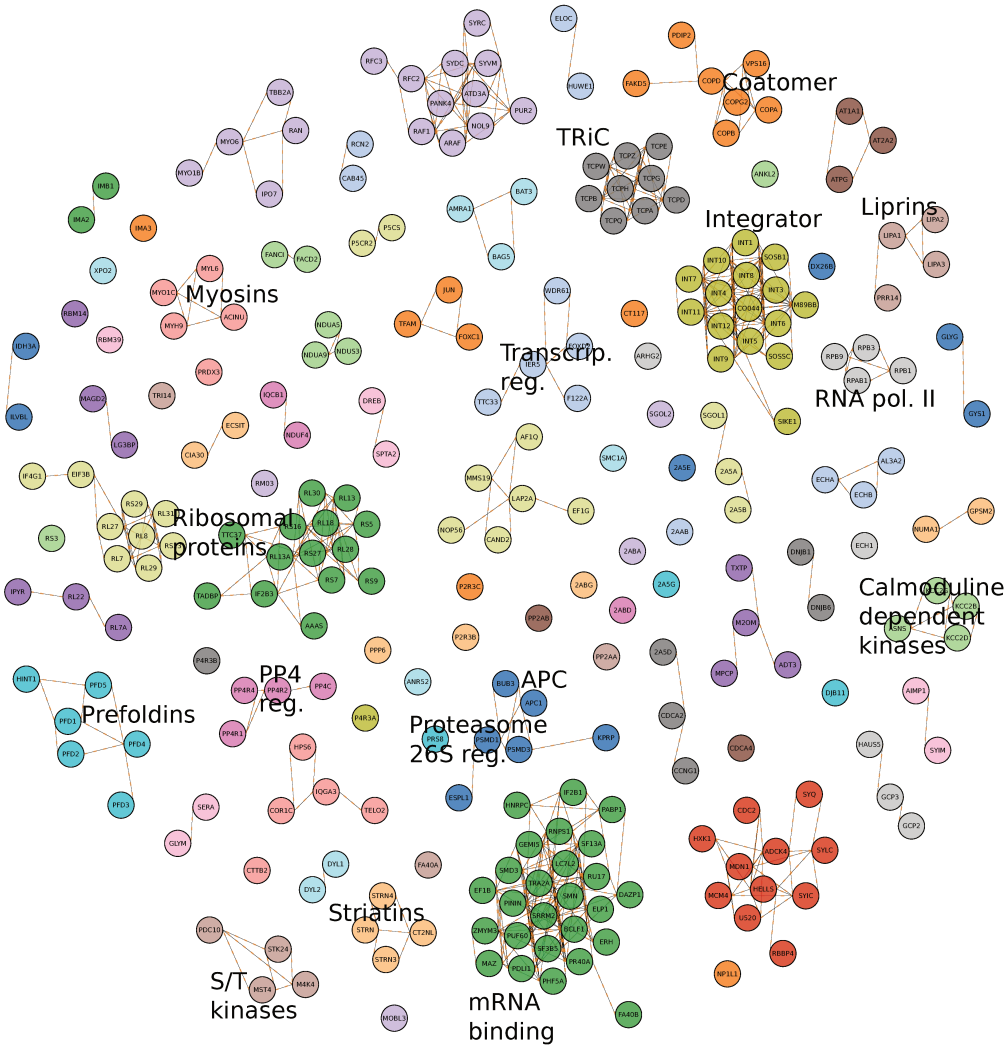


Figure 2.10: Complexes and clusters identified in the PP2A data set by the combination of Abundance Correlation and GO functional similarities followed by MCL clustering.

PP2A complex changes according to their substrate specificity, the abundance correlations between its subunits are expected to be low, and thus the complex is only revealed by GO functional similarities (as shown in the network of Figure 2.9).

Seemingly false clusters were also predicted in the network constructed with combined scores. However, compleXView visualizations allow the user to set interaction strength score thresholds, thereby observing which interactions are putatively less plausible in the network. For example, the association of PDIP2 and FAKD5 with coatomer proteins had a combined score below 0.7, and the cluster may be an artifact of the mitochondrial localization of these proteins. The red cluster contains RNA-/DNA-helicases, chromatin remodelers and tRNA ligases. This cluster seems to originate from the generality of the shared GO functionalities of its members, as all are nucleotide/ATP-binding proteins. Thus, it may simply represent a false cluster that dissociated upon a score threshold

of 0.8. Similarly, the cluster formed by EF1G, LAP2A, AF1Q, MMS19, NOP56 and CAND2 has relatively low functional similarities. Apart from their principal roles, the first four proteins in this cluster are involved directly or indirectly in transcription regulation (source: UniProt), which might explain their association. The cluster started to dissociate upon a score threshold of 0.7. Other clusters that lose association after a cutoff of 0.7 were, for example, the group formed by the transcription factors IF4G1 and EIF3B with ribosomal proteins, and the cluster between IPYR (an inorganic diphosphatase involved in tRNA aminoacylation for translation) and two ribosomal proteins.

Adding cross-linking information to the network

The single analysis of XL-MS data also identified the main protein complexes detected by AP-MS. Thereby, cross-links validated the interactions inferred in previous steps for the TRiC and STRIPAK complexes as well as for the cluster of calmodulin-dependent kinases (KCC proteins). Furthermore, cross-links revealed the interaction between the complexes predicted in previous steps of the analysis workflow. For example, it revealed that the TRiC complex interacts with PP2A and its regulatory subunits 2ABA and 2ABG.

Additionally, XL-MS revealed two new clusters. The first consisted of FR1OP, CE350 and P2R3C. The first two proteins are known to be required for anchoring microtubules to the centrosomes (source: UniProt), while the presence of P2R3C in this cluster is perhaps to give specificity to PP2A for FR1OP and CE350. The second cluster showed the association of the PP2A catalytic subunits, the regulatory subunits and IGBP1. With them, other proteins, like CT117 (SOGA1), appeared to interact directly with PP2A phosphatases. SOGA1 has a putative KEN motif (source: ELM database) and thus might bind to the APC/C complex, which promotes the transition from metaphase to anaphase via degradation of cohesin (source: UniProt). Linked to PP2A was also SGOL1, which prevents the degradation of cohesin by separase ESPL1, which formed a subcluster with the 26S proteasome component and the APC1 subunit. SGOL1 is also required for proper attachment of the spindle microtubule to the kinetochore (source: UniProt). The remaining proteins in this cluster were CCDC6, PR14L and FA13A. The former has been shown to interact with components of the SCF E3 ubiquitin ligase complex; the second is a proline-rich protein whose function is unknown; and the latter is an activator of Rho GTPases. The members of this cluster may not represent a single protein complex, and might be rather due to their different interactions with the PP2A complex.

Finally, the integration of both MS datasets showed that they complement each other. Among the subunits of the Integrator only two physical contacts were detected by XL-MS, and thus AP-MS data was fundamental to keep the cluster integrity and recognition of the complex. Protein abundances of the Integrator subunits were relatively high across the purifications indicating that the lack of detection by XL-MS could not have been due to low sample amounts. Recent studies have suggested that the Integrator might not exist as a single physical entity, but rather that its subunits accomplish the Integrator's function in a sequential manner [16]. The cross-linking data indicated that the nuclease core of the complex, represented by INT9 and 11, are in close proximity to INT4. It has been shown that INT4 knockdown reduces the association of INT11 with snRNAs [16]. Thus, the data

support the close association of INT4 with the INT9/11 catalytic subunit. Among the STRIPAK complex a relatively high number of interactions were observed. The STRIPAK complex appeared as the association of two modules: a kinase module and a larger module containing striatin proteins and microtubule/PP2A associated proteins. It is important to stress that the subcluster of kinases was only detected by AP-MS, showing again the synergy of integrating both datasets.

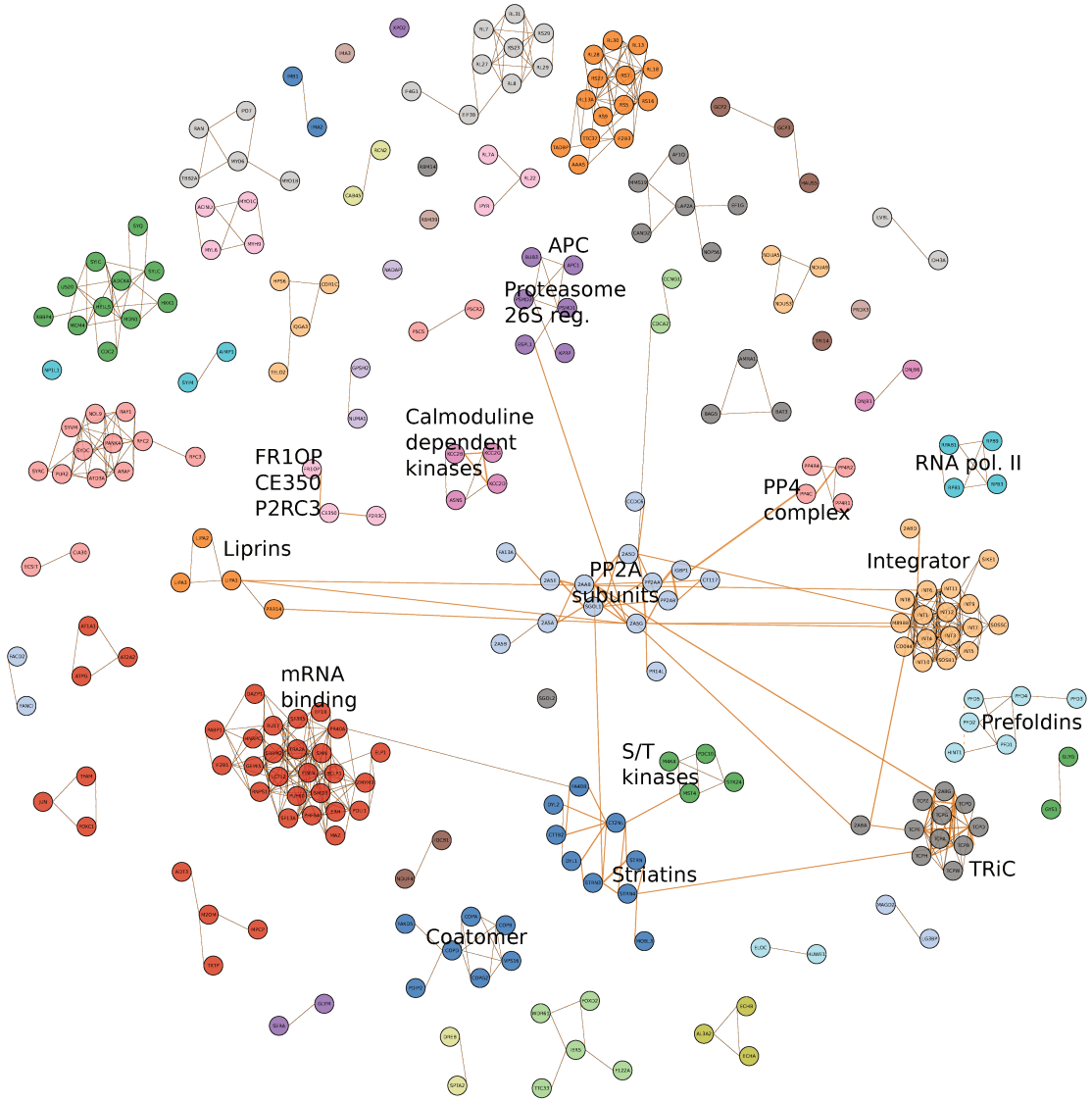


Figure 2.11: Complexes and clusters identified in the PP2A data set by the combination of Abundance Correlation and GO functional similarities followed by MCL clustering and incorporation of cross-links.

2.3 Discussion

Here I described the implementation of a web server tool for the analysis and integration of PPI data. Given the simplicity of its input tables, compleXView can be employed to combine AP-MS, BioID, APEX-labeling, XL-MS and Y2H data. The software does not aim to be a highly precise predictor of protein complexes. False interactions may be wrongly inferred during analysis. Nevertheless, compleXView provides tools for the validation and interpretation of the protein complexes and clusters that it predicts. It is able to resolve known protein complexes and can be the starting point for the characterization of novel interactions and their functional relevance.

An advantage of compleXView is that it is applicable to relatively small datasets. Previous work on inferring complexes used different clustering algorithms on very large networks that included datasets from Gavin and Krogan in 2006. Gavin et al. [24] performed a proteome-wide screen for complexes in budding yeast by systematic tagging of 6466 open reading frames. A concurrent effort was carried out by Krogan et al. [45] to process 4562 tagged yeast proteins by AP-MS. The purification of thousands of baits generated a high level of protein connectivity in these datasets, which served in subsequent years for the training of clustering algorithms and machine learning methods that inferred novel protein complexes in these networks [101]. In contrast, I showed here that compleXView is applicable to medium and small AP-MS datasets, with at least 5 baits, that together with Gene Ontology information can predict protein complexes in such networks.

A frequent discussion in PPI studies is whether a bait-prey interaction network (Spoke model) or a prey-prey interaction network (Matrix model) is better for the inference of protein complexes. In large data sets, Bader et al. [4] showed that a Spoke model of PPIs was more accurate than a Matrix model. This may be true for large data sets but not for small ones where the degree of connectivity is low, and thus, a Matrix model becomes necessary. Strictly speaking, compleXView does not model the data as a full Matrix model because it does not allow every possible protein-protein interaction in the network but only those with high correlation values. Furthermore, these correlations are calibrated with GO similarity scores by average or multiplication. Threshold filters and calibration of correlation values reduced the number of spurious interactions and resulted in better clustering of the data into sub-networks and complexes in the PP2A network and the MCM network (see Appendix A).

Recently, Montano-Gutierrez et al. [63] have shown by simulation studies that protein correlations between true interactors are robust enough to distinguish them as complexes, especially when the quantitative signal of the preys is high across experiments. The PP2A dataset is a good example of this case, and compleXView was able to recognize such correlated signatures within the TriC, Integrator and other complexes in this network. In another dataset, where the overlap of the baits interactomes was small, and therefore the correlation signals were low across purifications, compleXView was able to resolve complexes after the incorporation of GO similarities [100].

Different clustering algorithms have been used to group PPI networks into sub-modules and complexes. compleXView uses the simple and robust MCL algorithm for this task.

MCL does not require prior knowledge of protein domains or functional annotations. It groups proteins solely based on their predicted interaction strengths. Thanks to compleXView, prior information can be incorporated into the clustering process by fetching GO functional similarities between proteins. However, using only GO information was not as reliable as in combination with abundance correlations. Based on the compleXView analysis of the studied networks, the combination of these scores has more power than any of the two alone (see Results and Appendix A).

A closely related strategy to the one implemented in compleXView was carried out by Saha et al. [81], where MS quantitative information was also exploited and combined with GO functional similarities to infer protein complexes in a dataset of 384 human bait proteins. However, the authors did not provide any tool that makes their method publicly accessible. The advantage of compleXView is that it provides an automated web server for the analysis and visualization of PPI data from different sources.

Other methods are more general and have been developed for proteomics expression data obtained under different conditions or perturbations. An example is the pipeline Nano Random Forest [63]. Approaches like this rely again on machine learning algorithms that require positive and negative training sets of known interactions. In comparison, compleXView is straightforward in inferring complexes. Though, it may be less accurate, it has the advantage that the user can curate putative interactions thanks to the quick UniProt links that are provided. On top of that, compleXView integrates cross-linking information as a mean of validating physical interactions. Moreover, the integration of different data types enriches the insights that one obtains from each source alone, as shown here.

In summary, compleXView provides a simple bioinformatics pipeline with a user-friendly interface and highly annotated graphs that make it useful for the exploration, integration and interpretation of MS-based interactomics data.

2.4 Materials and Methods

Datasets

The two datasets analyzed in this chapter and Appendix A includes the quantification of protein abundances and the identification of chemical cross-links by mass spectrometric analyses. The first dataset was previously published by [29] and comprises AP-MS pull-downs of 14 different bait proteins from the PP2A complex and interactors. The baits include: PP2A catalytic subunit alpha (PP2AA), PP2A catalytic subunit beta (PP2AB), PP2A regulatory subunit A beta (2AAB), PP2A regulatory subunit B alpha (2ABA), PP2A regulatory subunit B gamma (2ABG), PP2A regulatory subunit delta (2A5D), PP2A regulatory subunit epsilon (2A5E), PP2A regulatory subunit gamma (2A5G), protein phosphatase 4 catalytic subunit (PP4C), Immunoglobulin-binding protein 1 (IGBP1), Shugoshin-like 1 (SGOL1), CTTNBP2 N-terminal-like protein (CT2NL), Striatin-interacting protein 2 (FA40B or STRP2) and FGFR1 oncogene partner (FR1OP).

The second data set [15] includes 6 bait proteins, each a member of the MCM2-7 subcomplex.

Data analysis

In order to quantify peptide abundances in the PP2A data set, raw files were analyzed with MaxQuant version 1.5 [11] and the results were filtered at 1% FDR. For the second data set, MaxQuant tables were directly retrieved from their respective PRIDE repository PXD004089. The cross-links for this data set were retrieved from PXD002987 (only cross-links between MCM components were considered).

In order to identify and quantify putative interactors of the bait proteins, raw peptide intensities obtained by MaxQuant were analyzed within the statistical environment R. Only unique peptides and proteins with a minimum of 2 identified peptides were considered for quantification. Median normalization between experiments was performed at the peptide level. Normalized peptide intensities were averaged within replicates in order to obtain protein abundance estimations. Protein identifications were required to be present in at least 2 replicates of the respective bait for the PP2A case and 1 for the MCM case (this latter data set did not contain replicates). Protein abundances across the same bait purifications were averaged and the significance of their fold-changes to the negative control was assessed by a Posterior Probability method (see below). Protein identifications were regarded as interactors if their enrichment to the negative control was at least two-fold and significant with an FDR of 0.05. The abundance ratios to the respective bait were calculated and interactors with ratios <2% were not included. As a result we obtained a 'Bait-Prey Interactions Table' listing the putative bait-prey interactions with their respective abundance ratios.

Posterior Probabilities were estimated using a mixture of 3 Gaussian distributions, whose parameters (mean and variance) were estimated using the ratios of the abundances in the positive experiment over the abundances in the negative control. The 3 Normal distributions were found and fitted using the Expectation-Maximization algorithm from the 'mclust' package in R. After parameter estimation, the left distribution was used to estimate the probability of an abundance ratio to be false, whereas the middle distribution was used to estimate the probability of an abundance ratio to be true if the ratio was below a threshold of 10 otherwise the right distribution was used. FDR and posterior error probabilities were estimated using the method of Kaell et al. [37].

The bait-prey interaction tables were used as input to infer prey-prey interactions. Pairwise cosine correlations were calculated using the prey-to-bait abundance ratios across different protein samples. Hence, this mathematical term is referred to as abundance correlation. GO similarities were calculated using the getGeneSim function from the GOSim Bioconductor package [22] with the following parameters: similarity method, 'dot'; normalization method, 'sqrt'; and similarity term, 'relevance'. Only 'Biological Process' (BP) and 'Molecular Function' (MF) categories were used. UniProt accession numbers were mapped to Entrez IDs using the UniProt 'Retrieve/ID mapping' tool. The BP and MF similarity values were summarized by keeping the maximum of the two per protein-protein

pair. Abundance correlations were combined with GO correlations by calculating the average of their values. Minimum thresholds of 0.8, 0.6 and 0.65 were allowed for abundance, GO and combined correlations, respectively in the case of the PP2A data, and 0.9, 0.7 and 0.75 in the case of MCM. Proteins were clustered using the MCL algorithm [17] on the abundance correlations, GO correlations or the combination of the two, respectively. The following parameters were used, expansion: 2, inflation: 3, maximum iterations: 50. Protein interactions were considered as true, if either i) any of the two proteins was a bait and their correlation was above the respective threshold or ii) both proteins were preys in the same MCL cluster with at least one showing a relative ratio to the bait higher than 2%, and their correlation value above the respective threshold, or iii) at least one protein-protein contact was detected by XL-MS. The results are summarized in 3 different tables with interactions based on abundance correlations, GO correlations or the combination of both correlations. These tables are annotated with the respective number of protein-protein contacts detected by XL-MS.

Result tables from the cross-linking experiments were directly retrieved from the PRIDE database. Intra-protein cross-links were filtered from the list whereas inter-protein cross-links were summarized to the number of cross-links per protein-protein pair.

compleXView modules

compleXView offers two different modules, which operate independently of each other. One module is for the analysis of AP-MS data and performs part of the analysis workflow described in Figure 2.1. The main input file for the 'Analysis' module is the 'Purifications Table' containing the protein abundances across all purifications. Its first column must be named Prey and contains the protein IDs of the co-purified proteins. The second and all other columns must contain the abundances of the preys in each of the purification experiments. These columns have to be named according to the following format: BaitID__ReplicateNumber__Condition. The name in the 'BaitID' field must match the format of the entries in the 'Prey' column and the bait itself has to be detected in the respective purification. Negative controls must be named 'NegCtr' in this field. The 'ReplicateNumber' field contains any number or code for the identification of technical or biological replicates (e.g., R1, R2, R3). The 'Condition' field is optional and should be provided in cases where purifications of the same bait under different biological conditions are compared. compleXView requires abundance values like iBAQ or other normalized intensities without log-transformation. Median or quantile normalization between conditions is optional.

The basic output of the 'Analysis' module is the 'Bait-Prey Interactions Table' visualized as a spoke network. Abundance correlations will only be computed if the number of baits or conditions is >4. The output is a protein-protein interaction table that we call the 'Abundance Correlations Table'. In order to compute GO functional similarities between proteins, an optional input table with two columns must be provided. The first column named 'From' contains the Protein IDs in the same format as in the 'Prey' column of the 'Purifications Table'. The second column named 'To' contains the respective UniProt Entrez ID of the protein. The compleXView output is a protein-protein interaction table

called ‘GO Correlations Table’, where each row contains a pair of preys and their corresponding GO similarity values. For the implementation of inter-protein cross-links an input table of at least four columns with the following headings is required: ‘Protein1’, ‘Protein2’, ‘AbsPos1’ and ‘AbsPos2’. The IDs in the first two columns should have the same format as the ‘Prey’ column in the ‘Purifications Table’. The numbers in the ‘AbsPos’ columns indicate the positions of the cross-linked amino acid residues. For more details, see online manual at `https://xvis.genzentrum.lmu.de/compleXView`.

On the other hand, the ‘Visualization’ module displays all bait-prey interaction tables and correlation-based tables generated by the ‘Analysis’ module. Both modules operate independently, which allows the visualization of output tables generated by other programs, such as SAINT, MiST or compPASS. The input table must contain at least 2 columns named ‘Bait’ and ‘Prey’; optional columns are used to represent quantitative information for the node edges. The ‘Visualization’ module generates two types of representations the ‘Network’ and ‘Blot’ plots. The former represents proteins as circular nodes and linear edges indicate their interactions, which are deduced from AP-MS abundances or indicated by XL-MS restraints. The ’Blot’ plot is designed as a western blot diagram displaying protein abundances across the different bait purifications.

Chapter 3

Inferring protein binding interfaces using amino acid sequence-level information and quantitative XL-MS

3.1 Introduction

Protein-protein interactions (PPI) are established upon the non-covalent binding of protein domains. These regions are called binding sites in the free state, and binding interfaces in the bound state. The establishment of binding interfaces is governed by amino acid sequence and structural properties of the interacting proteins. Physicochemical complementarity, shape, solvent accessibility, and evolutionary conservation are the most prominent properties of binding sites [34, 26, 114]. The elucidation of the binding interfaces in a PPI is important because mutations in these regions can lead to impairment of the interaction. In turn, this may lead to a loss of function and the disruption of a biological process. In a clinical context, dysfunctions due to abrogation or excessive formation of PPIs can lead to diseases [23, 35]. Thus, the characterization of protein binding interfaces is relevant for understanding the mechanisms of certain diseases and for the development of therapeutic interventions and drugs [48, 110].

For some protein complexes, binding interfaces can be retrieved from their X-Ray or NMR structures in the Protein Data Bank (PDB). A residue is considered to be part of a binding site if its distance to residues of the interacting protein is below 4-6 Å [114]. Among the set of binding residues, some are more relevant than others as most of the overall interaction energy resides on them. These residues are called hotspots [53]. High-resolution methods are not able to elucidate the majority of protein complexes and PPIs. In order to characterize the binding interfaces of these complexes, low-resolution experimental methods and computational predictors have been developed. Experimental methods are either laborious (e.g., alanine mutagenesis) or have low accuracy (e.g., hydrogen/deuterium exchange and XL-MS). Computational predictors, on the contrary, suffer from low specificity. Combining experimental information and bioinformatics analyses has been shown to be advantageous, as exemplified by homology-based predictors of binding interfaces [115].

Homology-based approaches search for structural models with high sequence similarity to the proteins in the queried PPI. Binding interfaces can be inferred on the queried proteins based on the assumption that high sequence similarity results in similar structural and functional domains. Compared to other computational methods, homology-based predictors are more accurate and reliable [115]. Nevertheless, they require the presence of homologous high-resolution structures in PDB, which, as mentioned, are missing for many PPIs.

In order to address the limitations of the available computational approaches to identify protein binding interfaces, the method described in this chapter aimed to combine quantitative XL-MS (qXL-MS) data and sequence-level properties in order to obtain a more accurate prediction of binding sites. The proposed approach constraints the search space by inferring the most probable regions in a protein sequence that may be binding sites. The method uses chemical cross-linking data of the protein complex, identifies the inter-protein cross-link sites and quantifies them relative to other cross-links in the protein. Subsequently, the regions around these sites are manually ranked as candidates for binding interfaces. The ranking considers the evolutionary conservation, secondary structure and relative accessible surface area of these regions, which are all properties derived from the amino acid sequences of the proteins.

This chapter describes the complete bioinformatics framework where the cross-linking and sequence-level information were combined to predict protein binding interfaces. Finally, it presents three proof-of-concept cases where the predictions were validated by in vitro and in vivo assays testing deletions of motifs or point mutants.

3.2 Results

3.2.1 Properties of binding interfaces

In order to find the sequence properties that discriminate binding from non-binding regions, I used density plots and random forest to evaluate the discriminant powers of a number of amino-acid-based sequence-level indexes. The analysis was performed on a set of 27 non-redundant protein complexes downloaded from PDB. Binding interfaces were defined based on a maximum threshold distance of 4.5 Å between any two residues on different partner proteins. The result of this evaluation showed that the most differentiating sequence properties were amino acid type, evolutionary conservation, relative accessible surface area, disorder, and relative position. The amino acids isoleucine, leucine, arginine and tyrosine were more frequently found at interfaces, whereas alanine, proline and serine were predominantly detected in non-binding regions (Figure 3.1 A). Furthermore, evolutionary conservation was shown to be relatively higher in binding interfaces. Nonetheless, the majority of residues within these regions had conservation values below 60% (Figure 3.1 B). Conservation was particularly distinctive on the following residues: tyrosine, arginine, glutamine, proline, asparagine, isoleucine, glycine, glutamate, aspartate and alanine (Figure 3.2). The predicted relative accessible surface areas also showed that interface residues are less accessible than their counterparts (Figure 3.1 C). Additionally, interfaces

occurred more often on relatively ordered regions (Figure 3.1 D), which could be biased as the analyzed structures come from crystallizable proteins. Interestingly, the relative position of the interface residues along the protein sequence was also distinctive. Binding sites tended to occur right before the middle of the protein and/or at the C-terminus (Figure 3.1 E).

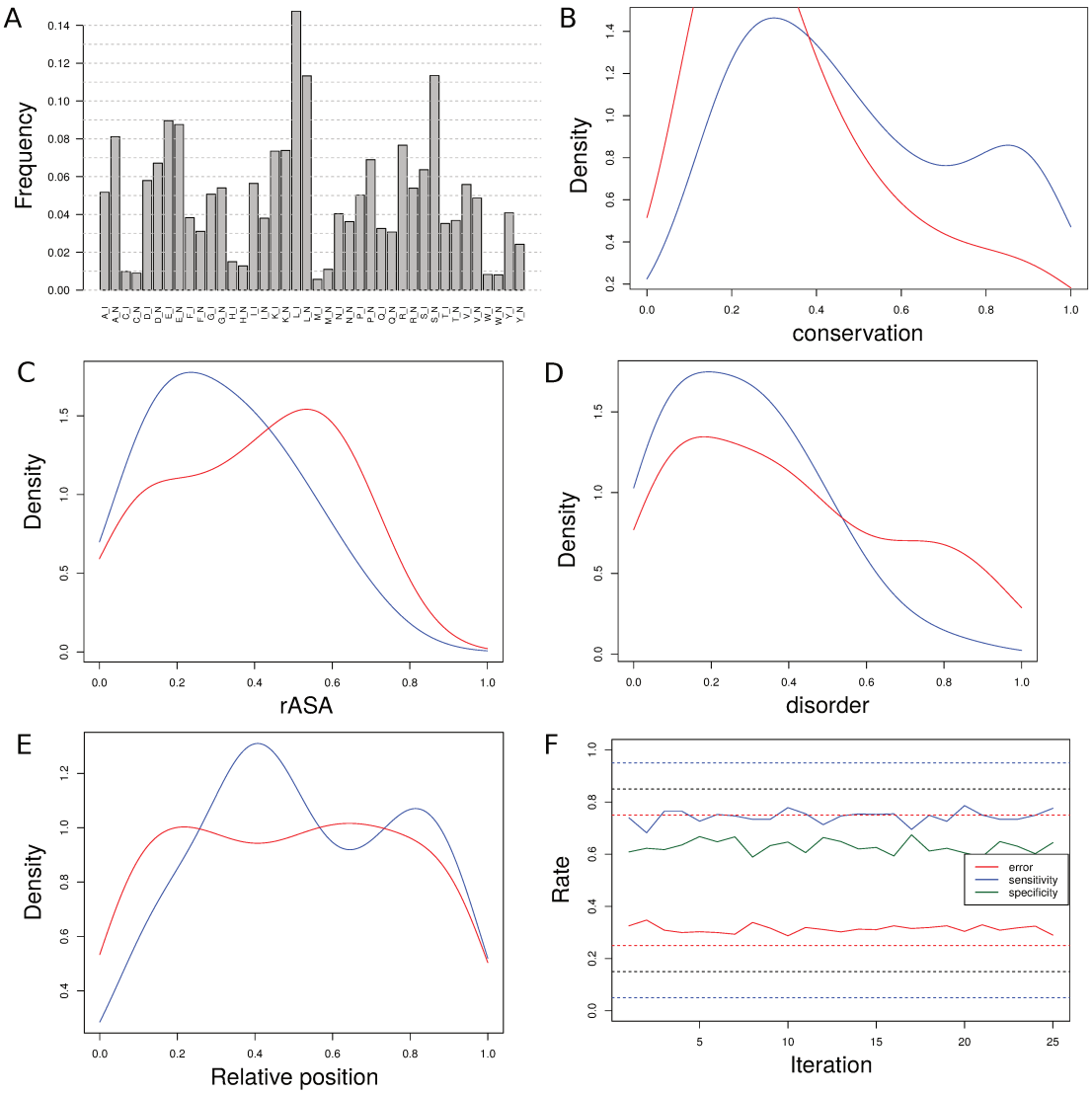


Figure 3.1: Distinctive properties of interface residues have low discriminant power to recognize binding interface residues from non-binding residues. A. Frequency of amino acids in interface (_I) versus non-interface regions (_N). B-E. Conservation, relative accessibility, disorder and relative position of interface regions (blue) versus non-interface regions (red). F. Sensitivity, specificity and error rates of the random forest model on different subsets of the training data set.

3.2.2 Machine learning models for the prediction of binding residues

In addition to the properties mentioned above, a long list of physicochemical variables (from the AAindex database; [40]) was also evaluated within the random forest framework. In order to find the indexes that could best predict interface residues, each variable was mean-smoothed using window sizes from 5 to 25. For each variable, the prediction errors (i.e., the OOB indexes) of the smoothed versions were computed and thereby the best smoothing window was individually chosen. A variable was smoothed as long as its prediction error decreased by more than 3% after smoothing. Next, as features may be highly correlated and thus redundant, they were clustered, and the best of each group was selected according to their discriminant powers (i.e., the mean decrease accuracy). The final random forest predictor was constructed with these features.

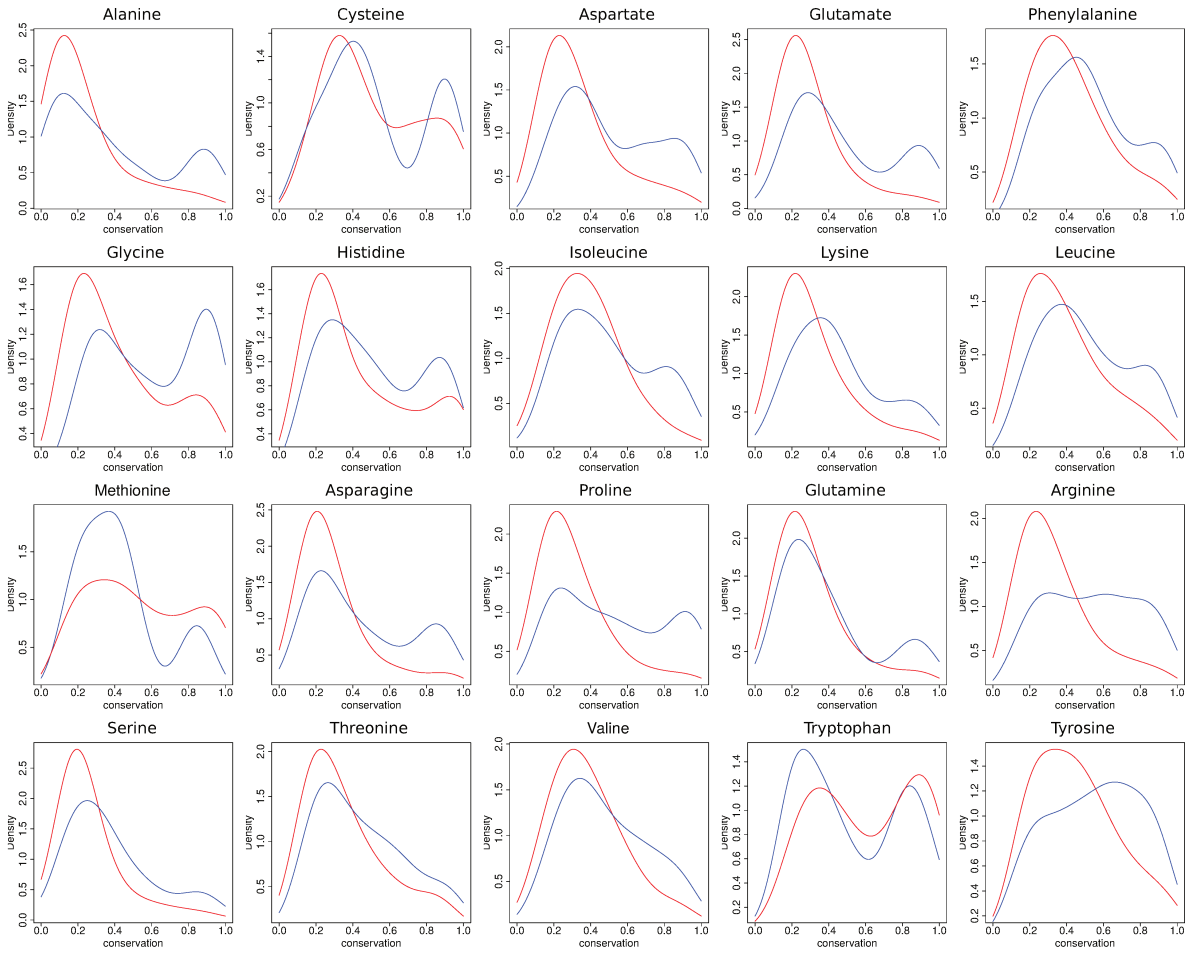


Figure 3.2: Specific types of amino acids display higher conservation on binding (blue) versus non-binding (red) interfaces.

The sensitivity, specificity and error of the model on the training set are shown in Figure 3.1 F. The qualities of these performance indexes were relatively poor. Another machine learning model was trained with the SVM algorithm. The performance of both models

was then assessed on 4 testing data sets. Although the sensitivity of the SVM model was higher, the specificity decreased (Table 3.1). Previous predictors have achieved similar sensitivities and specificities with only sequence-level properties [116, 67]. More limiting is the fact that the predicted interface residues are not partner-aware; i.e., they do not tell to which protein they bind. These two reasons make the applicability of the models very limited. Combining experimental data with the main predictor variables may result in better prediction of binding domains.

Model	Data	Sensitivity	Specificity	Error
RF	Test1	61.3	56.5	41.1
SVM	Test1	75.3	41.1	41.8
RF	Test2	62.8	52.7	42.2
SVM	Test2	79.7	36.1	42.1
RF	Test3	44.0	61.9	47.0
SVM	Test3	57.6	44.0	49.2
RF	Test4	40.6	70.2	44.5
SVM	Test4	56.9	54.5	44.3

Table 3.1: Performance of the RF and SVM models on four test sets

3.2.3 Inter-protein cross-link intensities as indicators of binding interfaces

Inter-protein cross-linked sites in XL-MS experiments occur mostly at the binding interface or close to it. However, some cross-links also occur in regions that are flexible and only transiently come close to the interface or in contact with other parts of the binding partner. Distinguishing these regions is essential to make a precise prediction of the binding interface. By theorizing about the establishment of binding interfaces and the formation of lysine-specific cross-links, I deduced the following pseudo axioms: i) Cross-link intensity depends on the spatial distance of the linked sites (Figure 3.3 A); ii) Binding interfaces and structural regions tend to be ordered and not flexible regions (Figure 3.3 B); iii) During cross-linking of a PPI, intra- and inter-protein cross-links have to compete for the cross-linker amount and for lysine sites at the binding interface or nearby.

From the pseudo axioms above, one can deduce that the inter-protein cross-link intensity of a specific lysine site relative to the total sum of intensities on that site will be higher if the lysine is close to the binding interface. Structural regions that are not part of the interface will have the smallest intensity ratio, whereas flexible regions that transiently come in proximity to the interface (or other parts of the binding partner) will have a middle value. I call this ratio the relative interface propensity index (RIPI) of a cross-linked residue (Equation 3.1).

$$RIPI_{Ki} = \frac{Inter\ XL\ Intensity_{Ki}}{Intensity(Monolink_{Ki} + Loop\ Link_{Ki} + Intra\ XL_{Ki} + Inter\ XL_{Ki})} \quad (3.1)$$

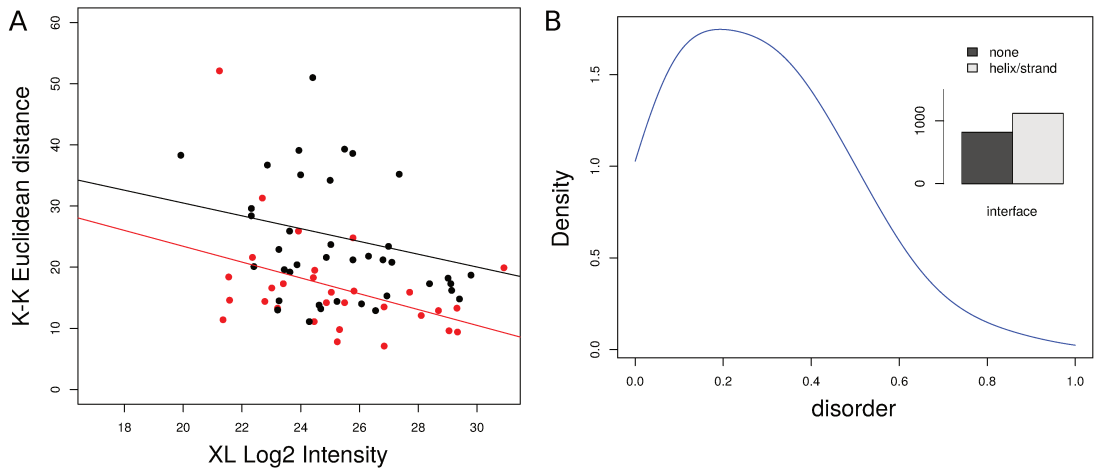


Figure 3.3: The intensity of cross-links depends on the lysine-lysine distance (A), and binding interfaces tend to be structured or display low disorder (B). Red dots in A: intra-protein cross-links; black dots: inter-protein cross-links; lines represent regression lines.

In order to validate the applicability of the RIPI, I analyzed the cross-linking data of 3 complexes for which the binding interfaces are known. For example, in the RNA polymerase II complex, some of the lysines around binding interface residues have a high RIPI indeed (Figure 3.4).

Similarly, the interface lysines of the CENPA-MIF2 dimer (Figure 3.5) and the CNN1-SPC24/25 hetero-trimer (Figure 3.6) have high RIPIs. In the case of the CENPA-MIF2 complex, there is no structural model of their interaction. Nevertheless, previous work [113] has shown that the interacting regions of the dimer are between residues 279-313 of MIF2 and residues 170-229 of CENPA. In the case of the CNN1-SPC24/25 complex, the PDB structure 4GEQ was used to obtain the real interfaces. This structural model only contains information for residues 155-213 of SPC24, residues 61-80 of CNN1 and residues 132-121 of SPC25. Therefore regions outside these ranges can contain binding interfaces that are unknown. The cross-linking data indicates that is the case of the CNN1 N-terminus that seems to be in close proximity with SPC24 (Figure 3.6).

In the three protein complexes shown above, high RIPIs occur next to predicted secondary structures with relative high conservation, low disorder and low rASA. Therefore, I propose the RIPI index as an indicator of binding interfaces, which together with sequence features, can be used to infer binding domains.

3.2.4 Combination of RIPI and sequence features predicts binding interfaces

As the machine-learning models alone have low discriminant power, a heuristic-based framework was followed to infer binding interfaces. The rules of this framework are:

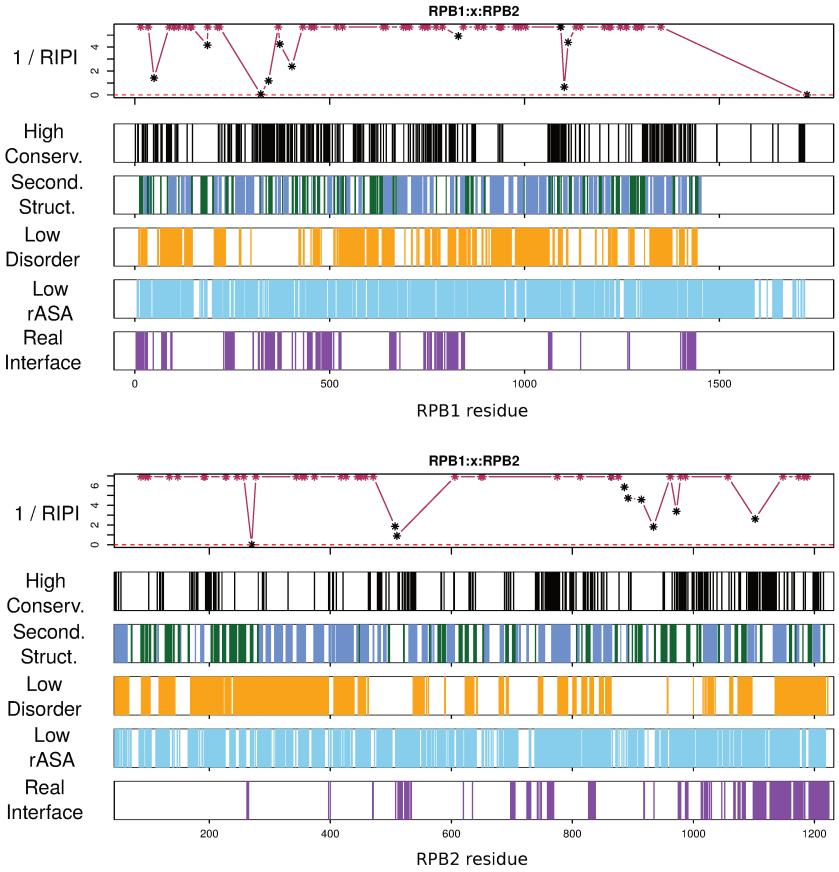


Figure 3.4: High RIPI values coincide with the interface regions between RPB1 and RPB2 in the RNA polymerase II complex. Inter-protein cross-linked lysines are represented as black asterisks. Only the top 20% conserved residues within the protein sequence are depicted. Secondary structure is differentiated between alpha helices (blue) and beta strands (green). Residues were considered to have a low disorder if their IUPred index was below 0.25 in a scale of 0 to 1, and were considered to have low accessibility if their rASA was below 40%. Real interfaces were extracted from the PDB model 5IP and were considered as such if their residue-residue distance was below 4.5 Å.

i Rank cross-linked regions according to their RIPI.

ii Look for the closest (predicted) secondary structure or ordered domain in the vicinity of the RIPI peaks.

iii Assess the degree of conservation, rASA and RFM/SVM prediction for these regions.

iv Taking all into consideration decide which residue range to delete.

v If point mutations are the aim take into consideration the interacting secondary structures and experimental/functional data on their interactions.

As proofs of concept, I show a couple of deletions and point mutations predicted with

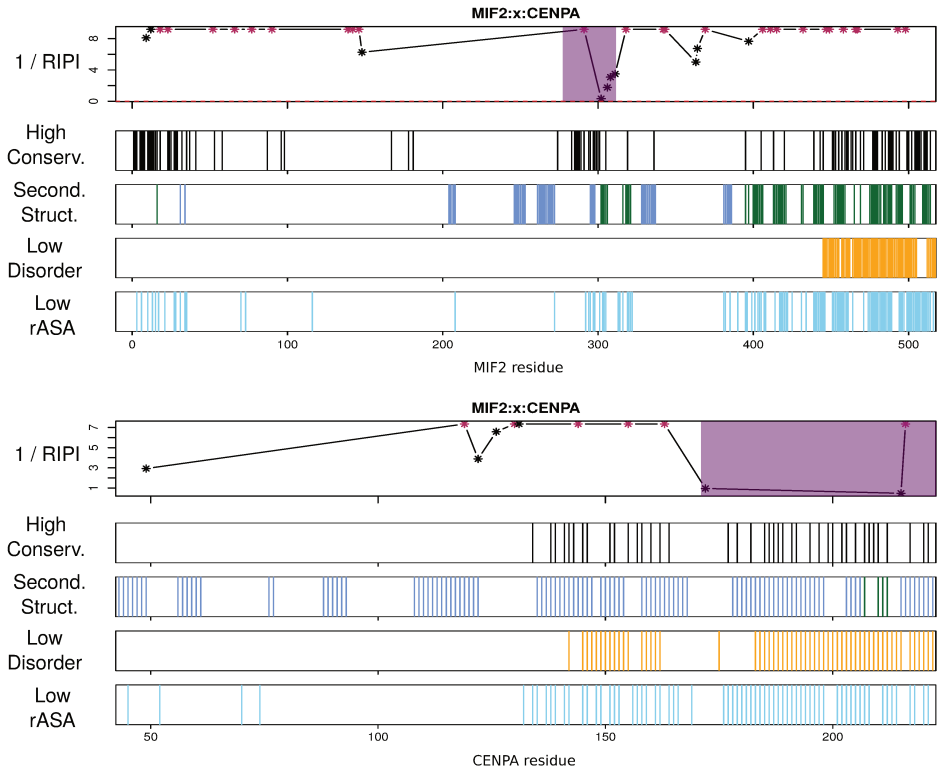


Figure 3.5: High RIPI values coincide with the binding regions of CENPA and MIF2. Binding regions (colored in purple) were retrieved from the literature (see main text). The other indexes were obtained as in Figure 3.4.

the above framework that led to positive experimental results, i.e., to the disruption or reduction of the interaction between the implicated proteins. Binding assays followed by western blot analysis were used to validate the proposed mutants. The gels of the actual experiments are not shown here because these experiments were performed within the context of another yet unpublished study.

As first case, the binding interface between CENPA and OKP1 was inferred (Figure 3.7). The RIPI values along the respective sequences indicate that OKP1-binding domain on CENPA is around residue number 50. The secondary structure prediction suggests that the domain is one of two alpha helices or both, but preferentially the helix downstream of residue 50. Indeed, these predicted helices are not reported in PDB structures of nucleosomes. Thus, it is unclear if they exist or may form upon binding to OKP1. Size exclusion chromatography (SEC) analyses upon deletion of these helices, showed that the downstream helix is required for the binding of CENPA to OKP1 (Figure 3.8). There is a third candidate for a binding region between residues 100 and 150 that contain lower RIPI values. However, this deletion was not tested. In the case of OKP1, the RIPI values indicate that its binding domain to CENPA is localized between residues 130 and 200. Due to the relatively higher degree of conservation, the two helices exactly located within residues 150 and 200 were chosen for deletion. The deletion of both helices individually

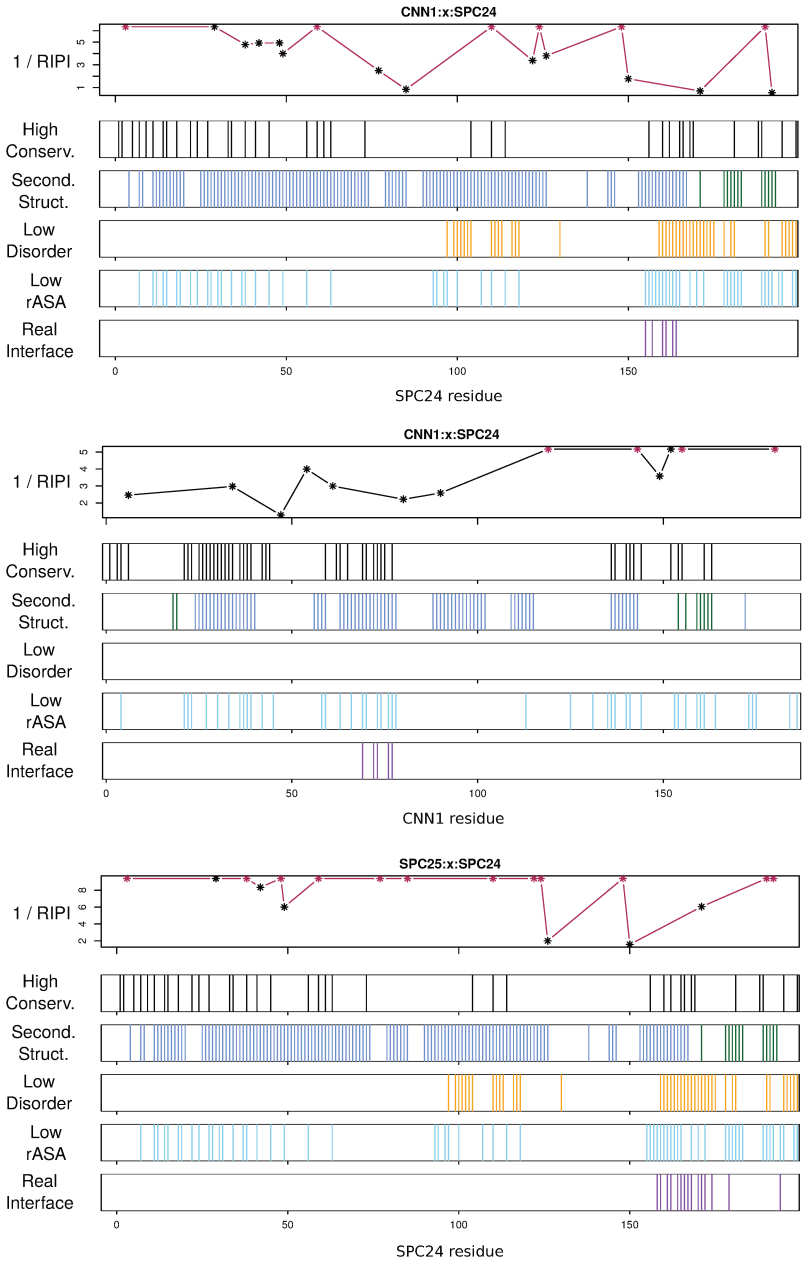


Figure 3.6: High RIPI values coincide with the binding interfaces of the CNN1-SPC24/25 complex. Binding regions were retrieved from the PDB model 4GEQ.

led to a reduction of the interaction with CENPA (data not shown).

As second case, the binding interface between KRE28 and MTW1 was inferred (Figure 3.9). The MTW1 helix between residues 225 and 250 seemed to be the best candidate for a deletion experiment. Similarly, the KRE28 helix between 220 and 265 seemed to be the corresponding binding domain. Deletion of these regions led to a decrease in the interaction (data not shown). Helix-helix interactions are the most predominant kind

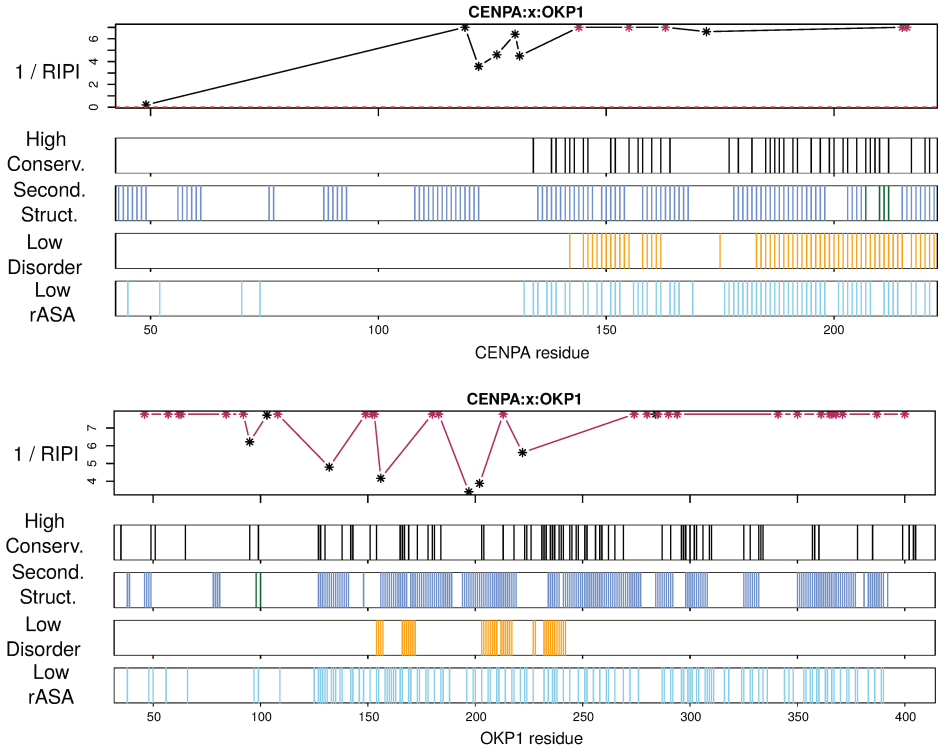


Figure 3.7: RIPI values and sequence properties predict minimum binding regions in the CENPA-OKP1 complex.

of interfaces in protein interactions. The hotspot residues in this kind of interfaces are hydrophobic residues located along the helix [3]. To test the relevance of these residues on the KRE28 protein, pairwise mutations of L231 and L234, L241 and L245 and M248 and V251 were designed. Effectively, each of these pair mutations decreased the interaction with MTW1 (data not shown).

In the third case study, the binding interface between CBF3A and the MTW complex (composed of DSN1, NSL1, NNS1 and MTW1) was inferred (Figure 3.10). The cross-linking data suggested that CBF3 interacts with the tetramer complex through the NLS1/MTW1 dimer. The RIPI index indicates that the helices between residues 115-130 and 330-350 on CBF3A seemed to be the binding domains in this protein. On the other hand, the helix located C-terminally of residue 250 on MTW1 and the helix located N-terminally of residue 40 on NSL1 seemed to be the corresponding binding domains of the interaction. The experimental validation of this prediction was not completed at the time of submission of this thesis.

3.3 Discussion

In this chapter, I have described a heuristic approach that combines information from qXL-MS experiments and sequence-level features to infer binding domains. The approach

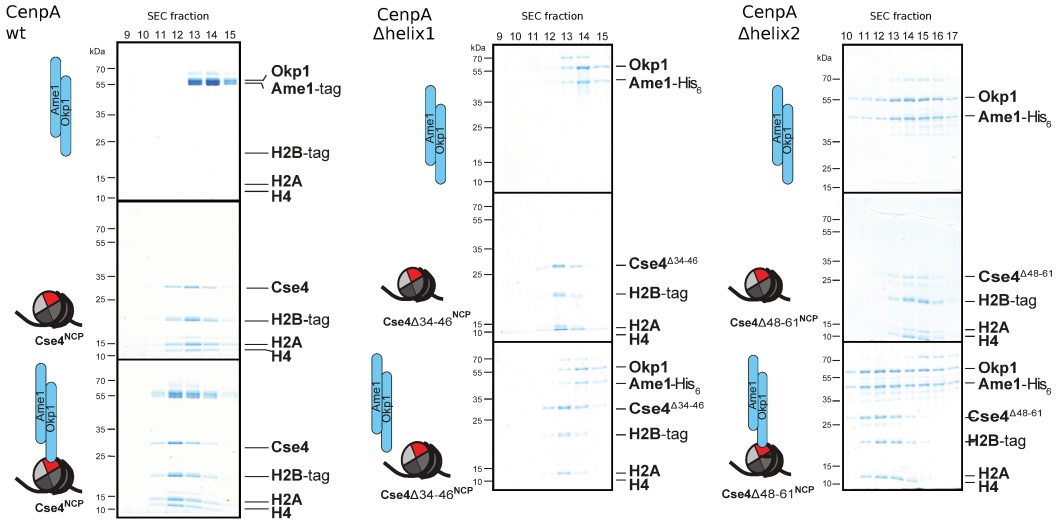


Figure 3.8: Shift assays of the AME1-OKP1 dimer and yeast nucleosomes to heavier SEC fractions upon pre-incubation and interaction of both complexes with each other. Eluted proteins were visualized with SDS-PAGE and Coomassie staining. Smaller SEC fractions indicate heavier macromolecules. Nucleosomes contained wild type CENPA (left gels) or deletion mutants of residues 34-46 (middle gels) and residues 48-61 (right gels). Deletion mutant Δ34-56 did not cause any shift to heavier SEC fractions, which indicates loss of interaction of nucleosomes with the AME1-OKP1 dimer.

facilitates a more educated and efficient design of deletion mutants for binding assays.

Similar observations were reported recently by Liu et al. [54]. Here, I improve on their approach by incorporating quantitative information of inter-protein cross-links and properties such as evolutionary conservation and secondary structure prediction of the involved proteins. Liu et al. observed that some cross-links occur on regions dispensable for interaction, showing that XL-MS without quantitative information is prone to present cross-links around regions that are not part of the direct binding interface. Those dispensable regions are intrinsically disordered or have to be very flexible in order to intermittently come close to the interface. With qXL-MS information and the RIPI index these limitations are overcome. This is best exemplified by the CENPA-MIF2 interaction presented above (Figure 3.5). A number of cross-link sites were observed outside the minimal binding regions of both interactors. Nevertheless, the RIPI index reveals their dispensability and directs the design of deletion mutants to the actual binding regions.

An alternative experimental method to elucidate binding interfaces is hydrogen-deuterium exchange (HDX) MS. Schmitzberger et al. [88] used this method to find the distinct binding sites on OKP1 for its interaction with AME1 and CTF19-MCM21. They found that the sequence ranges 166-211 and 234-264, where protected from deuterium exchange in the presence of AME1, CTF19 and MCM21. Figure 3.11 shows how quantitative cross-link-derived restraints and the RIPI index partially overlap with the observations obtained by HDX-MS. Nevertheless, XL-MS has the advantage of resolving the binding partners that interact with these regions. As demonstrated above, part of the first HDX region on

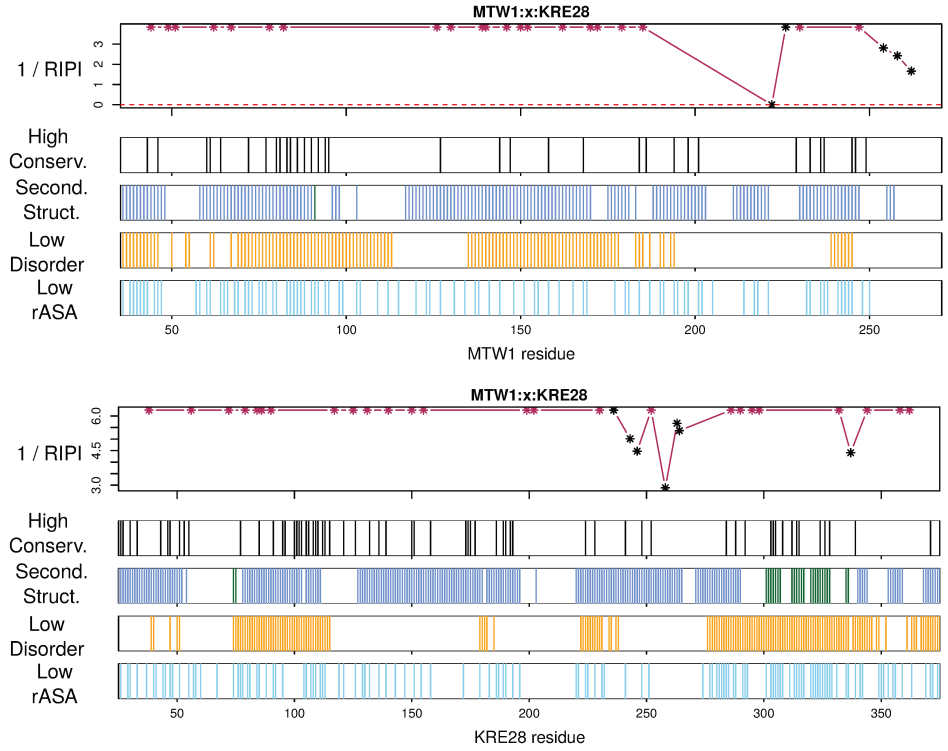


Figure 3.9: RIPI values and sequence properties predict minimum binding regions in the MTW1-KRE28 complex.

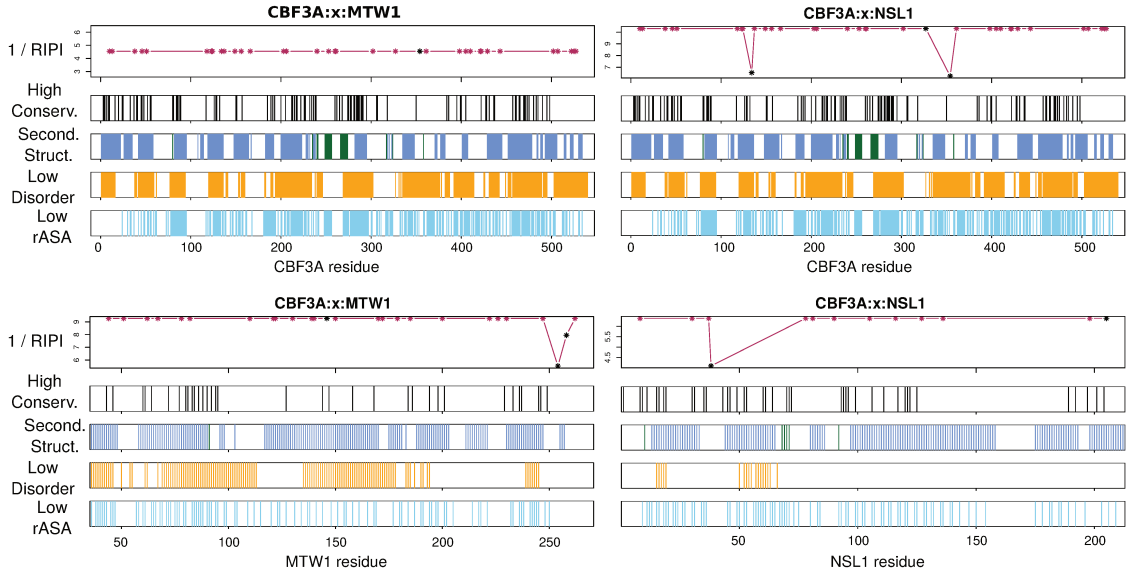


Figure 3.10: RIPI values and sequence properties predict the minimum binding regions of the CBF3A-MTWc complex.

OKP1 together with downstream residues constitute the binding domain of this protein for CENPA, whereas the second HDX region and two other upstream putative helices seemed to be relevant for its interaction with AME1. The corresponding OKP1-binding domain on AME1 appears to be located on the putative C-terminal helices, which is consistent with the HDX region on this protein.

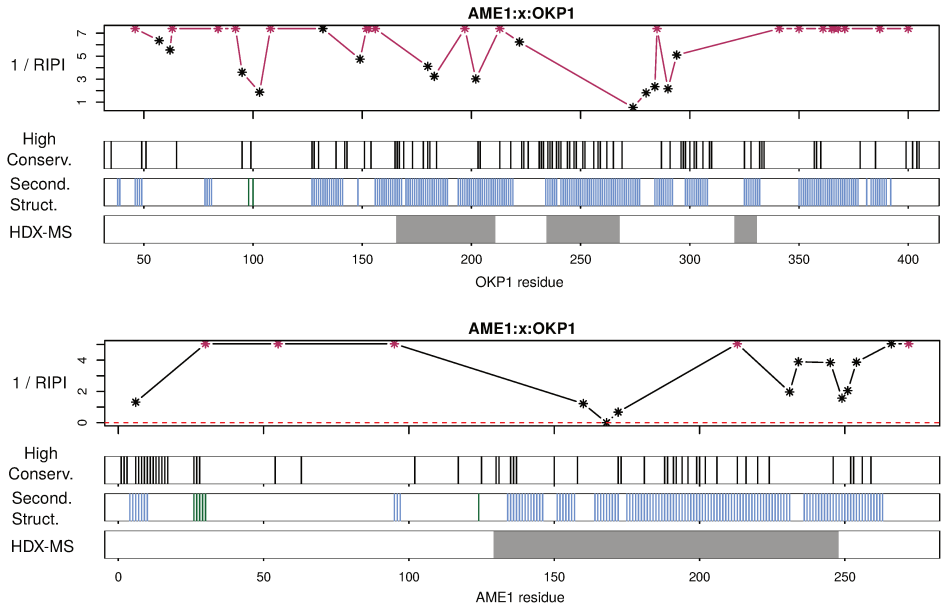


Figure 3.11: High RIPI values from XL-MS experiments occurred at or around deuterium protected regions observed in HDX-MS experiments.

In summary, I demonstrated that qXL-MS data and sequence-level properties can facilitate the characterization of interaction domains. Lysine-targeted qXL-MS information has the ability to precisely predict the binding domains and always depends on the presence of this amino acid at both sides of the interface or in its proximity. Nevertheless, using another cross-linker with different specificity or alternative methods such as HDX-MS can help to elucidate the most probable candidate region in cases where lysines are absent. Taken together the proposed approach highlights the strength of quantitative XL-MS experiments for methods to predict and characterize PPIs.

3.4 Materials and Methods

Datasets

To learn about the properties of binding interfaces, the following PDB structures were downloaded: 1ID3, 2PK9, 3R7W, 4CGY, 4TU3, 5J9Q, 5SVA, 1JEY, 3IAB, 3SJA, 4GEQ, 4XR7, 5L4G, 5WXM, 1SZA, 3MKS, 3ZDM, 4KRD, 5BW8, 5L4G, 6ENY, 2NPI, 3NBH, 3ZS9, 4M75, 5IP7, 5L4K.

From these structures, the following fasta files were retrieved from Human UniProt: 1A03, PSA6, B2MG, PSA7, CALR, PSB1, PSB2, PSB3, PSB4, PSB5, PSB6, PSB7, PDIA3, PSD11, PSD12, PSD13, PRS10, PSDE, PRS4, PSMD1, SEM1, PRS6A, PSMD2, PRS6B, PSMD3, TPSN, PRS7, PSMD4, PRS8, PSMD6, PSA1, PSMD7, XRCC5, XRCC6, PSA2, PSA3, PSA4, PSMD8, RMI1, RMI2, PSA5; and from S. cerevisiae UniProt: LSM7, RPAB2, MDY2, RPAB3, MPP10, RPAB4, CDC4, PAN2, RPAB5, CLP1, PAN3, RPB11, EAF6, PCF11, RPB1, EPL1, PCL10, RPB2, ESA1, RPB3, GET3, PHO80, RPB4, GET4, PHO85, RPB7, GTR1, POP6, RPB9, GTR2, POP7, SAC1, H2A1, H2B2, H3, SKP1, H4, IMP3, LSM1, VPS74, LSM2, LSM3, LSM4, YNG2, LSM5, LSM6, RPAB1.

The data was split into training and testing sets. The following complexes were in the testing set: the yeast RNA polymerase II complex (PDB: 5IP7), the Proteasome 26S complex (PDB: 5L4G, 5L4K), the Peptide loading complex (PDB: 6ENY) and the Ku heterodimer complex (PDB: 1JEY).

The cross-linking data for the RNA polymerase II was acquired as explained below. The cross-linking data for the complexes CNN1-SPC, the CENPA nucleosomes-MIF2-OKP1-AME1, the KRE28-MTWc and the CBF3A-MTWc was acquired similarly, but by other doctoral candidates in the group of Dr. Franz Herzog.

Feature matrices for machine learning

Binding interface residues were labeled as such if their distance to a residue on an interacting protein was less than 4.5 Å. Distance was measured from any heavy atom in one residue to any heavy atom in the other residue.

To measure evolutionary conservation, fasta sequences were given as input to the PSIBlast [2] standalone software. The search was done against the non-redundant and Uref90 databases separately. For the yeast proteins, the genus Saccharomyces was excluded from the search. The following indexes (obtained from the PSSM matrix) were used as indicators of conservation: the conservation of the query residue, the highest conservation on the MSA position, and the information content per position.

To predict the secondary structure and accessible surface area (ASA) of the protein sequences, the PSSM matrices obtained with PSIBlast were also used as input for the SPIDER2 [28] software. The relative ASA was computed from the ASA using a home script. The script calculated these values using previously reported maximal ASA values of each amino acid from the SPIDER2 publication itself as well as from Tien et al. 2013 (empirical and theoretical; [104]), Miller et al. 1987 [62], and Rose et al. 1985 [77].

In order to predict disorder and induced-order-upon-interaction, the IUPRED [13] and ANCHOR [14] software were used.

To obtain the physicochemical properties of the amino acids, the index tables provided by the AAindex database and ExPASY ProtScale were used.

Overall 603 sequence properties were present in the features set.

Learning of the RF and SVM predictors

The R packages randomForest and e1071 were used for the training and prediction of the random forest (RF) and support vector machine (SVM) models. In order to select the best features for interface prediction, the variables in the features matrices were smoothed by the average method using windows of odd sizes from 5 to 25. A feature was replaced by its smooth version if the resulting average decrease in the RF out-of-bag error was greater than 3%. Subsequently, features were clustered using the hierarchical clustering algorithm to avoid redundancy from highly correlated features. For the physicochemical indexes, their tree was divided into 10 clusters. For the conservation, ASA, secondary structure and (dis-)order properties, the tree was divided into 5 clusters. From each of the 15 clusters, the best 2 features were selected based on their 5 cross-validation RF variable importance index with a sequentially reduced number of predictors. The importance index was determined by the mean decrease accuracy after randomization of the feature in question.

The remaining features (41 in total) were fed to the RF and SVM algorithms. Their performances on the training and testing data sets were evaluated using the sensitivity, specificity and overall error indexes.

Cross-linking of the RNA polymerase II complex

The purified protein complex was kindly provided by the group of Dr. Patrick Crammer. A total of 52.8 μg of complex was diluted in cross-linking buffer to a concentration of 0.8 μg/μl. Under the assumption that 1 μg of protein contains 0.5 nmol of lysine, 26.4 nmoles of BS3 cross-linker were added and left to react for 30 min at 10 C. The reaction was stopped by adding 8 μl of AMBIC 1M. Subsequently, proteins were denaturated by adding 132 μl of urea 8M. Reduction and alkylation was done with 20.6 μl of TCEP and 22.7 μl of iodo acetamide. Protein digestion was performed with 1 μg of lysC for 2 h at 35 C, followed by 1 ug of trypsin overnight. Digestion was stopped by medium acidification using TFA to a final percentage of 1% and ACN to a final percentage of 3%. Peptide cleanup was performed using C18 columns using ACN 100% as activator solution, 3% ACN + 0.2% formic acid as washing solution, and 60% ACN as eluting solution, to a final elution volume of 700 μl. The solution was then dry using the speedvac, and re-suspended on size exclusion chromatography (SEC) buffer. After SEC separation, the fractions corresponding to the cross-linked peptides were selected and pooled for mass spectrometric analysis.

Mass spectrometric analysis

Peptides were separated by a Thermofisher nano-HPLC machine and analyzed with an LTQ-Orbitrap Elite instrument. A flow rate of 20 nl/min at incremental gradients of buffer B from 3% to 98% was used. At each MS cycle, the top 10 intense peptides with charges greater than 2 were selected for fragmentation and MS2 scanning, with exclusion

times of 30 s. MS1 spectra were acquired in the orbitrap analyzer at 12K resolution, and MS2 fragment scans at low resolution in the ion trap analyzer.

Identification and quantification of cross-links

Peptide-peptide cross-links were identified using the Xquest/Xprophet software [107]. Mono-links, loop-links and intra-/inter-protein cross-links with a score above 25 were retained and quantified. The quantification was performed using a modified version of the OpenMS FeatureFinderCentroided tool. This tool and the changes that I made to the source code are presented in detail in the next chapter of my thesis.

The RIPI index

Cross-link intensities were summarized to lysine site intensities, by summing up all the intensities were the protein lysine site was involved. This total sum includes mono-links, loop-links and intra-/inter-protein cross-links. Next, the site intensity due to inter-protein cross-links from a specific dimer interaction was divided by the total sum. The resulting value was called the relative interface propensity index (RIPI) of a cross-linked residue. Lysine sites, for which none inter-protein cross-link was observed, were assigned a RIPI value equal to the minimum RIPI in the set. This was done to avoid infinite values for the plotted inversed RIPIs in the figures above.

Other indexes in the RIPI plots

Conservation in the RIPI plots was computed using PSIBlast against the UNIREF90 database. Only residue positions with conservation above the 80% quantile within the protein sequence were plotted. Secondary structure and rASA were predicted using the SPIDER2 software against the UNIREF90 database. Residues were considered to have low disorder if their IUPred index was below 0.25 in a scale of 0 to 1. Residues were considered to have low accessibility if their rASA was below 40%. Real interface residues were extracted from PDB models. The predicted interface residues were obtained with the RF and SVM models trained on the best-selected features.

Chapter 4

Estimation of dissociation constants by quantitative XL-MS

4.1 Introduction

An important step towards the characterization of protein complexes is to determine the affinities of protein-protein interactions (PPI) because they guide and stabilize the assembly of protein complexes and enrich the information provided by biochemical pathways. Protein complex formation is a non-covalent biochemical reaction that reaches equilibrium when the concentrations of the free subunits and the associated subunits are constant over time (Equation 4.1 and 4.2).

$$aA + bB \underset{k_r}{\overset{k_f}{\rightleftharpoons}} C \tag{4.1}$$

$$where\ C = A_aB_b := the\ protein\ complex$$

.

$$At\ equilibrium,$$

$$k_f[A]^a[B]^b = k_r[C] \tag{4.2}$$

$$K_d = \frac{[A]^a[B]^b}{[C]} = \frac{k_r}{k_f} \tag{4.3}$$

The affinity of an interaction is measured in solution by the association rate k_f of the proteins relative to their dissociation rate k_r. This ratio is called the association constant and equals the ratio between bound and unbound partners at equilibrium. Association constants of biochemical reactions range from the mili-molar to the nano-molar scale. They

often play critical roles in biological pathways, where molecular interactions are highly regulated to modulate the rate of biochemical reactions. Thus, the manipulation of association rates provides opportunities to control cellular processes naturally or through drugs. Examples are protein inhibitors, co-activators, the preference of enzymes for different substrates and the use of drugs to ameliorate disease symptoms.

Association constants are inversely related to dissociation constants (K_d; Equation 4.3). Thus, a low K_d indicates high affinity between the subunits of a protein complex, whereas a high K_d indicates the opposite. There are three widely used technologies to measure the K_d of protein complexes: surface plasmon resonance (SPR), isothermal titration calorimetry (ITC) and fluorescence-based methods such as fluorescence polarization (FP) and fluorescence resonance energy transfer (FRET). Advantages and disadvantages exist for all these techniques [78], ranging from cost, required sample amounts, dynamic range, and interferences caused by fluorescence tags and platform immobilization.

Chemical cross-linking in combination with mass spectrometry (XL-MS) has become increasingly popular in hybrid structural biology approaches to reveal the topology and structural features of native proteins and protein complexes. Nevertheless, the development of bioinformatics pipelines for the quantification of cross-links can open new applications for XL-MS that go beyond the sole structural elucidation of proteins. Quantitative XL-MS (qXL-MS) has the potential to measure the dynamic cooperation of proteins in biological networks. It is suitable for calculating the stoichiometry of protein complexes, estimating the relative affinities between their members and measuring complex turnover. qXL-MS can also allow the quantification of post-translational modifications (PTMs) at or next to the contact interface and thereby evaluate the effect of PTMs on the affinity of binary interactions [86]. All these potential applications are important to advance the understanding of the dynamics of protein interactions.

Here, I describe a method that combines chemical cross-linking and mass spectrometry to estimate the K_d of protein complexes (Figure 4.1). The method is based on a series of titrations where the concentration of one of the proteins remains constant whereas the concentration of the other is varied. Proteins are incubated with a deuterated cross-linker, and the cross-linked peptides are enriched by size exclusion chromatography and identified and quantified by mass spectrometry. The next step models the relation of the intra-protein cross-link intensities with the initial concentrations of the subunits through linear regression. Subsequently, it interpolates the intensities of the inter-protein cross-links to estimate the amount of formed complex and the K_d of the interaction. Measuring the intensities of the cross-links thoroughly and accurately is critical for the success of our method. Thus, I developed a new bioinformatics quantification pipeline. The pipeline couples the identification results from popular cross-links search engines with peptide ion quantification. This is achieved sensitively and accurately, yielding a useful tool for the estimation of dissociation constants in protein-protein interactions.

This chapter starts with the description of the bioinformatics pipeline used for the quantification of cross-links. It further describes the K_d estimation of the CNN1-SPC24/25 complex. The estimated value is benchmarked to an ITC measurement as proof of concept. Subsequently, the ability of the method to measure affinities below the μM range

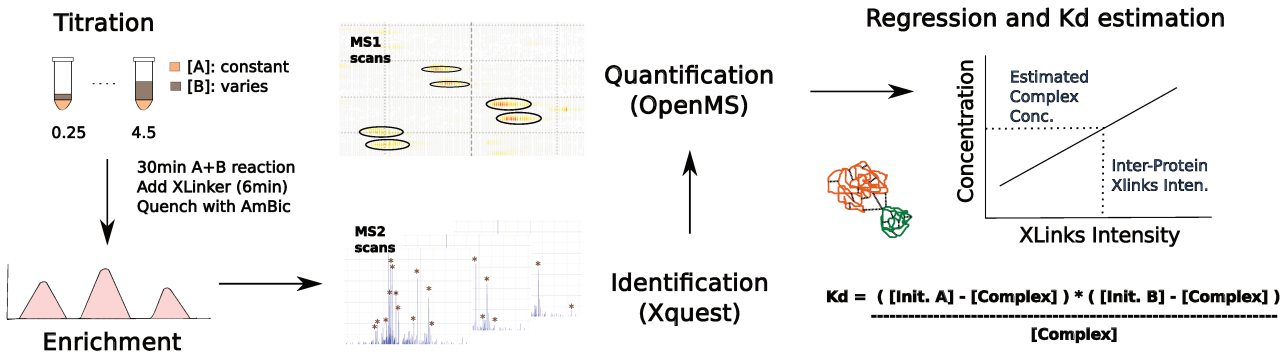


Figure 4.1: Proposed workflow for the estimation of K_d with qXL-MS. Titrations of a protein-protein interaction are performed and treated with a deuterated cross-linker. Cross-linked peptides are enriched by size exclusion chromatography and analyzed by mass spectrometry. Bioinformatics pipelines are used for the identification of cross-links and the extraction of their quantities. Linear regression is employed to infer the amount of formed complex in the samples. The calculated concentrations are finally used to estimate the K_d of the interaction.

is tested. Finally, the method is applied to a multimeric complex to measure relative affinities and changes upon the incorporation of PTMs and an extra protein member.

4.2 Results

4.2.1 Description and validation of the quantification pipeline

In order to quantify dissociation constants, the extraction of MS1 peaks (or features) in the mass over charge (m/z) and retention time (RT) space should be sufficiently sensitive and accurate. To this end, I built a bioinformatics pipeline using tools and data formats from the OpenMS framework [79]. The pipeline starts with the conversion of identification tables from the Xquest search engine [107] to the idXML format. In parallel, raw files are converted to the mzML format and fed to a modified version of the FeatureFinderAlgorithm from OpenMS. Given that the FeatureFinderAlgorithm extracts heavy and light pairs separately its sensitivity is higher than tools developed for SILAC experiments. The FeatureFinderAlgorithm was developed for the quantification of linear peptides. Nonetheless, introducing some modifications to the algorithm and fine-tuning the search parameters, makes it highly suitable for the extraction of features from cross-linking experiments. The modifications that I introduced to the source code mainly consider changes in the final score of the extracted feature, which is now computed as the geometric mean of i) the correlation to the theoretical elution profile model, ii) the fit of the isotopomers to the averagine model and iii) the deviation of the feature's m/z values. Moreover, if identifications are provided as seeds for the extraction, the degree of parameterization is loosened, because identifications at the MS2 level are already proof that the peaks ex-

ist at the MS1 level. After feature extraction, peaks are matched to the identifications using the IDMapper tool. This was also modified from its original version, mainly to allow the matching of an identification to the second isotopomer in the extracted feature. The changes introduced to both tools are based on empirical observations in processing cross-linking data. I observed that the elution profiles of highly abundant cross-linked peptides are mostly asymmetric, whereas low abundant cross-links have symmetric peaks (i.e., Gaussian-shaped). Isotopomers from low-abundant peaks may have a lower fit to the averagine model, and frequently a pseudo monoisotopic peak (i.e., lighter than the actual monoisotopic peak) co-occurs not only for mono-links but also cross-links. This pseudo monoisotopic peak can deceive the feature extraction algorithm in some cases, which ends up in its erroneous incorporation to the feature, thereby avoiding an identification match. The next tool in the pipeline is the FeatureLinkerUnlabeled tool, which matches features between different runs as long as they are replicates or from analog fractions. This resembles the known 'match-between-runs' strategy developed in the quantification software MaxQuant [11]. Finally, the annotated features are summarized to unique peptide-peptide cross-links and unique site-site cross-links. Like other OpenMS tool, the pipeline can be operated in the command line or in a graphical user interface. Running times are faster than the Xtract software [108] and comparable or faster than MaxQuant [20].

To validate the pipeline, I tested its performance on previously published datasets that were acquired for other extraction tools. They include the dataset used for the development of the Xtract algorithm [108] and the datasets used in the extraction pipelines described in [20, 66].

The first dataset comprises 4 dilution experiments from bovine albumin, bovine transferrin, and chicken transferrin [108]. Each protein was cross-linked separately and then pooled before spectra acquisition. In each dilution experiment, the concentration of albumin was kept constant, whereas the concentration of transferrins decreased monotonically in 2:4:8 ratios for the bovine homolog and 4:16:64 for the chicken homolog. As shown in Figure 4.2A, the feature extraction pipeline is able to reproduce the dilution steps of the experimental design. In the 2 most diluted experiments of chicken transferrin, the extracted features were only detectable by the 'match-between-runs' strategy since the protein was not identified in these 2 dilutions. In the most diluted experiment, the features that passed undetected by the 'match-between-runs' strategy were noisy and had too many missing mass traces, thus the failure of detection was justified. Overall, the proposed pipeline was able to quantify a similar number of site-site cross-links as the Xtract software, for which this dataset was created. The overall recall of site-site cross-links was 97.5% (118/121) and the accuracy of the quantification close to the expected values (Figure 4.2A).

The second dataset [20] consists of a SILAC-like experiment, where the protein C3 was cross-linked in its native and cleaved form, C3b, in forward (C3-BS3d0 and B3b-BS3d4) and reverse labeling (C3-BS3d4 and B3b-BS3d0). I used the identifications provided by the authors as feature extraction seeds, without 'match-between-runs' as it was not applicable to this dataset. As benchmark pipeline, the authors used the Pinpoint software with manual curation of the extracted features. I also used this reference as ground truth to compare the performance of their MaxQuant version tailored for cross-linking

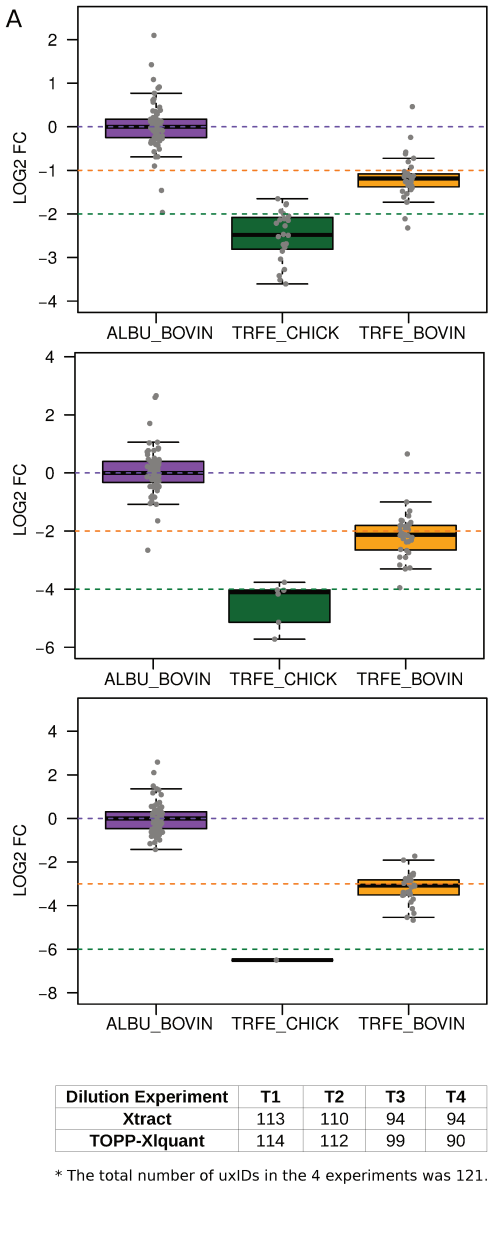
quantification against the here proposed extraction pipeline. The latter showed a 10% increase in sensitivity by recalling features not detected by MaxQuant. Moreover, the pipeline could quantify certain identifications, whose recall was not possible even with the benchmark Pinpoint pipeline (Figure 4.2B). Thus, the reported sensitivity might be even higher than 79%.

A third dataset that was published for the assessment of the reproducibility of cross-link quantification [66] was also used for evaluation. This dataset comprises 2 experiments with 10 replicates each. In the first one, human serum albumin was cross-linked in 10 cross-linking reaction replicates that were analyzed separately by mass spectrometry. In the second, the 10 reaction replicates were pooled and analyzed 10 times by mass spectrometry, so they represent injection replicates. The IDs provided by the authors of the publication were matched to the extracted features from the MS1 maps using a 'match-between-runs' strategy. As a result, unique cross-linked residues were quantified with a recall between 67-76% in the injection replicates and 55-66% in the reaction replicates respect to the overall number of unique cross-links in each of the two experiments (Figure 4.2 C, first row). Pooling all the quantifications across their replicates results in a sensitivity of 86% and 79% for the injection and reaction experiments, respectively. The sensitivities concerning the number of identifications within each replicate are even higher and fluctuate between 84-92% and 82-92% for the injection and reaction experiments, respectively (Figure 4.2 C, second row). Overall, these quantification rates highlight the high sensitivity of the proposed pipeline. Next, I calculated the correlations and coefficients of variation of the abundances at the peptide-peptide cross-link level (Figure 4.2 C, third and fourth rows). The minimum Pearson correlation between replicates was 0.92 in the injection experiment and 0.67 in the reaction experiment. Regarding the coefficients of variation, the median value was 14.6% in the injection experiment and 42.3% in the reaction experiment. Similar coefficients of variation were observed at the unique cross-linked residues level: 15% and 43% respectively. These values are higher relative to the values reported in the original publication of 14% and 32%, respectively. Part of this variation is due to the higher sensitivity achieved with the 'match-between-runs' strategy, as the values decrease to 11% and 38%, respectively, after removing quantifications extracted with this strategy. In summary, this analysis shows that cross-linking reaction replicates show 2-3 times higher variation than technical replicates, which should be taken into consideration when interpreting and benchmarking the estimation of K_d values by mass spectrometry with respect to established methods such as ITC.

The experiments described in the following sections were performed by Goetz Hagemann in the Herzog group. My work involved the data processing in order to identify and quantify cross-links for K_d estimation.

4.2.2 K_d estimation of a CNN1 short peptide and the SPC24/25 dimer

In order to evaluate the applicability of qXL-MS for the estimation of dissociation constants, a series of titrations between a CNN1-peptide (residues 60-85) and the SPC24/25



Dilution Experiment	T1	T2	T3	T4
Xtract	113	110	94	94
TOPP-Xlquant	114	112	99	90

* The total number of uxIDs in the 4 experiments was 121.

B

	MQ for Xlinks	TOPP-Xlquant
Sensitivity (%)	68	79.07
Accuracy (%)	88	89.04
Quantified in one of pair (No.)	11	22
Quantified consistently (No.)	65	65
Conflict quantitation	16	12
Conflict quantitation & Quantified in one of pair (No.)	0	3
Not quantified (No.)	11	1

* For this dataset, sensitivity and accuracy were measured using as reference extraction pipeline "Pinpoint + Manual Inspection". We observed that the reference pipeline could not extract features for some identifications provided by the authors, but our pipeline did. Thus we want to pinpoint that the sensitivity of our pipeline is even higher than the above reported percentage.

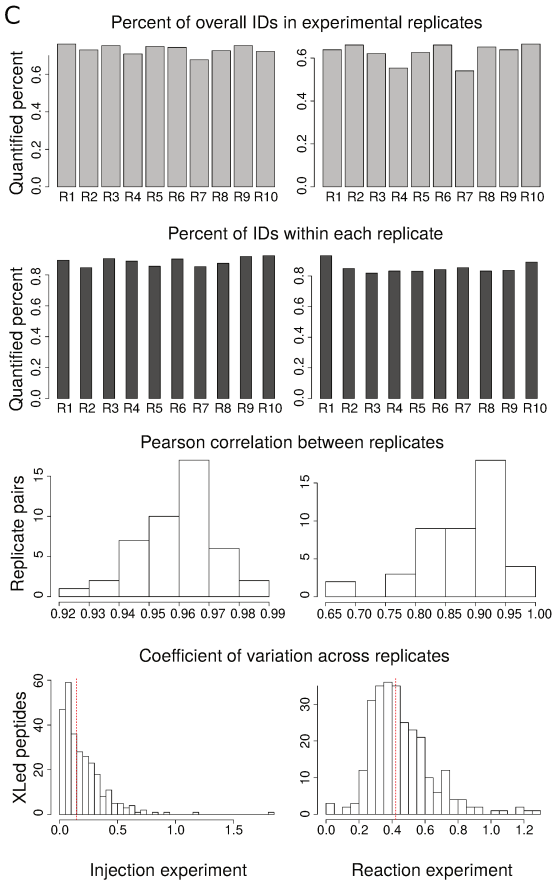


Figure 4.2: Validation of the quantification pipeline sensitivity and accuracy to quantified cross-linking identifications. A. Quantification of fold change ratios in dataset 1. Dashed lines indicate the expected ratios between the dilutions and the stock sample; boxplots' median values indicate the measured fold changes after peak extraction with the quantification pipeline. The bottom table indicates the number of unique cross-links quantified by the presented pipeline and compared to the Xtract pipeline. B. Statistics of the quantification in dataset 2. C. Sensitivity, correlation between replicates and coefficient of variations in dataset 3. Summarizations for injection replicates and reaction replicates are shown on the left and right sides, respectively.

dimer were performed. CNN1 interacts with SPC24/25 with relatively high affinity at a K_d of 2-3 μM as measured by ITC by our group and previous work [59]. This value served as a benchmark for the method proposed here and outlined in Figure 4.1. The CNN1 peptide and SPC24/25 were cross-linked at increasing molar ratios from 0.25 to 4.5, by varying the concentration of the CNN1-peptide while keeping the concentration of SPC24/25 constant. Cross-linked peptides were enriched with size-exclusion chromatography and analyzed by MS. Identification was performed with the Xquest/Xprophet search engine and quantified with the pipeline described in the previous section. The quantification of intra-protein cross-links reproduced the CNN1 and SPC ratios (Figure 4.3, left plot). The titration experiments 9-10,13-16 deviated from the expected ratios presumably as a consequence of their high CNN1 intensity deviations (Figure 4.3, middle plot) as the SPC24/25 concentrations remained as expected relatively constant in these experiments (Figure 4.3, right plot).

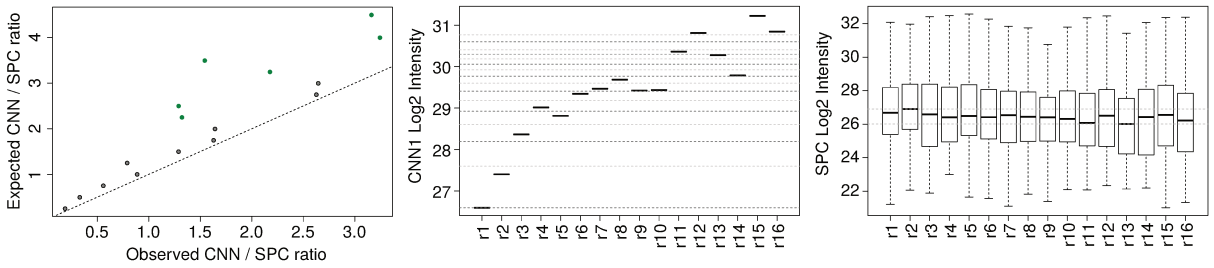


Figure 4.3: Quantification of CNN1 and SPC intra-protein cross-links. Left plot: expected ratios (calculated from the titrated concentrations) versus observed ratios (calculated from the average intensities of the intra-protein cross-links). The dotted line indicates the identity line. Middle plot: CNN1 intensities across the titration experiments. Dotted lines represent the expected intensities based on the intensity in titration r1. Right plot: SPC24/25 average intensities across the titration experiments.

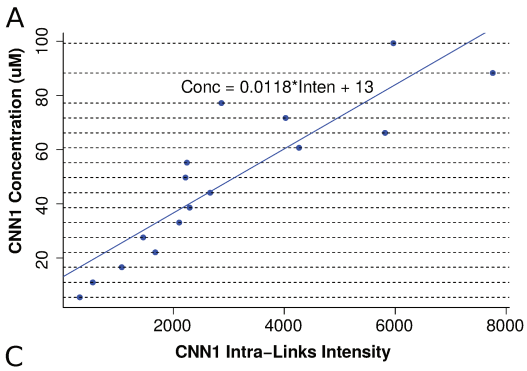
Next, the relation of the intra-protein cross-link intensities to the initial concentrations of CNN1 was modeled through linear regression (Figure 4.4 A). Subsequently, the inter-protein cross-link intensities between CNN1 and SPC were interpolated within the regression model to estimate the amount of formed complex. With these concentrations the K_d of the interaction could be estimated using the kinetic equation:

$$K_d = [FreeCNN1] * [FreeSPC]/[CNN1 : x : SPC]$$

The 'free' concentrations were calculated by subtracting the estimated complex concentration from the fitted initial concentrations of the subunits. The use of the fitted values rather than the original initial concentrations is not arbitrary, but rather to correct for pipetting errors in the experiments and to use the same method to estimate the concentration of the bound and unbound components of the reaction.

The estimated average K_d was relatively close to the benchmark value obtained by ITC and improves after filtering putative outliers (Figure 4.4 B). However, the standard deviation was relatively high. Using an alternative approach, which is more familiar in kinetic

studies, I obtained an even narrower estimation range (Figure 4.4 C and D). If the deviations of experiments 9-10 and 13-16 are indeed a product of technical variability (i.e., introduced during peptide enrichment, spectra acquisition and/or feature extraction), removing them from the analysis is justified. After doing this and recalculating values, the estimated K_d remained unchanged (Figure 4.4 E and F).



B

CNN1:SPC	Estimated Kd
0.25	0.73
0.50	1.50
0.75	2.06
1.00	1.72
1.25	1.89
1.50	2.93
1.75	2.86
2.00	4.52
2.25	3.64
2.50	0.15
2.75	4.43
3.00	4.97
3.25	2.05
3.50	2.51
4.00	8.96
4.50	10.17
Aver. Kd	**3.44**
Std. Kd	**2.74**
Aver. Kd (1:14)	**2.57**
Std. Kd (1:14)	**1.43**

D

	Estimate	Std. Error	Kd= 1/Estimate
(Intercept)	0.81	0.22	1.23
x	-0.85	0.25	1.18
		Aver. Kd	1.21
		Range	0.94 – 1.68

F

	Estimate	Std. Error	Kd= 1/Estimate
(Intercept)	0.71	0.16	1.41
x	-0.79	0.20	1.27
		Aver. Kd	1.34
		Range	1.08 – 1.76

Figure 4.4: Estimation of K_d values for the short peptide CNN1-SPC24/25 complex. A. Linear regression model to estimate the relation between cross-link measured intensities and initial protein concentrations. B. Estimated K_d values for each titration experiment. The average K_d and standard deviations were calculated either using all values or after excluding the 2 most outliers. C and D. Estimation of K_d using a line method. The inverse of the interception with the y-axis and the negative of the inverse of the slope should both equal the K_d of the protein interaction. The standard error of the intercept and slope estimates were used to calculate a -/+ SE range for the K_d. E and F. As in C and D, but after excluding titration experiments 9-10 and 13-16.

In order to assess the reproducibility of the experiments and the above estimation, two replicate experiments were performed. The results and the respective K_d values (Figure 4.5 A and B) showed that indeed the measurement was reproducible. All in all, the proposed method is capable of estimating the affinity of protein interactions with sufficient accuracy and reproducibility.

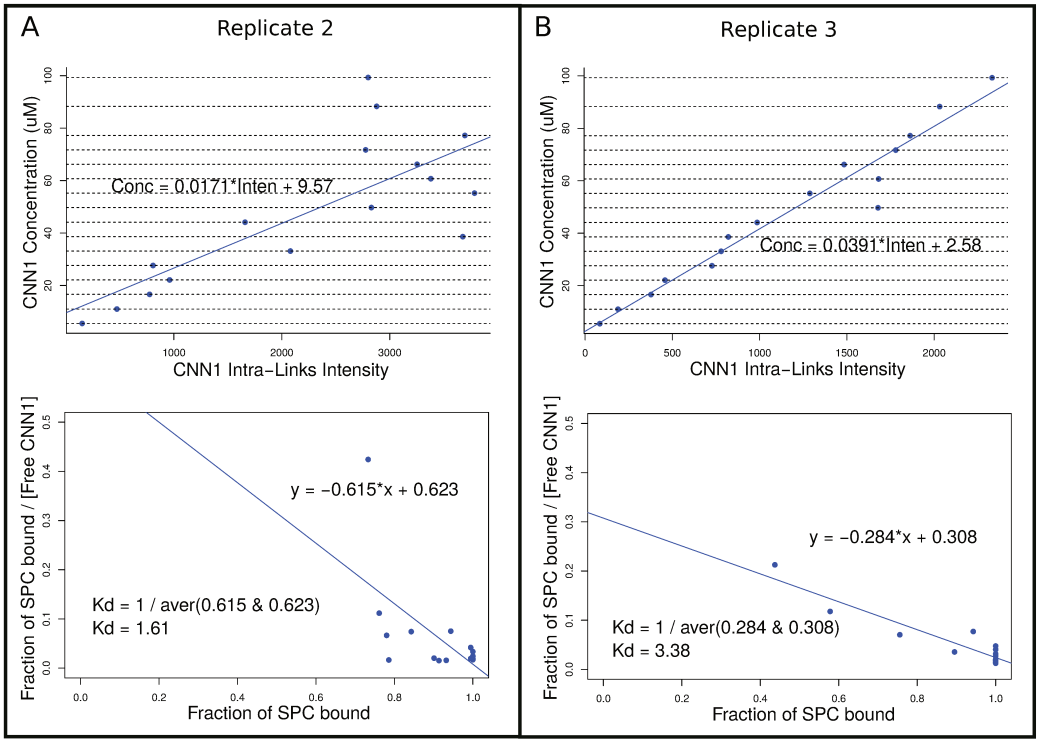


Figure 4.5: Experimental replicates of the short peptide CNN1-SPC24/25 titrations show that the K_d estimation is reproducible.

4.2.3 K_d estimation of a CNN1 long peptide and the SPC24/25 dimer

To assess the dynamic range of the method, and its accuracy and precision for estimating K_d values in the nM range, a series of titrations between a longer CNN1-peptide (residues 1-200) and the SPC24/25 dimer were performed. This peptide has been previously shown to have a higher affinity for the SPC dimer than the short CNN1 peptide, with a K_d value of approximately 16 nM [59]. As for the previous titration experiment, cross-links were quantified using a match-between-runs strategy, and the intra-/inter-protein cross-link intensities were used for estimating the concentration of the free subunits and the complex. A K_d value of 115 nM was eventually estimated (Figure 4.6, top left plot). This value presumably differs from the one in the literature due to the fact that the authors of the recent publication used truncated versions of the SPC24/25 dimer that only included the globular regions of these proteins. In the titrations performed here, the whole SPC24

and SPC25 sequences were used. To assess whether restricting the analysis to cross-linked sites that fell within globular regions of the SPC24/25 could result in a lower K_d, I decided to perform an incremental filtering of cross-links. Indeed the K_d decreased to a value of 49 nM after considering cross-links, for which at least one of the two sites fell in a globular domain (Figure 4.6, top right plot). Restricting the analysis further to cross-links for which both sites fell in regions with reported secondary structures resulted in an even lower K_d of 16.6 nM (Figure 4.6, bottom plot), which is surprisingly close to the literature value.

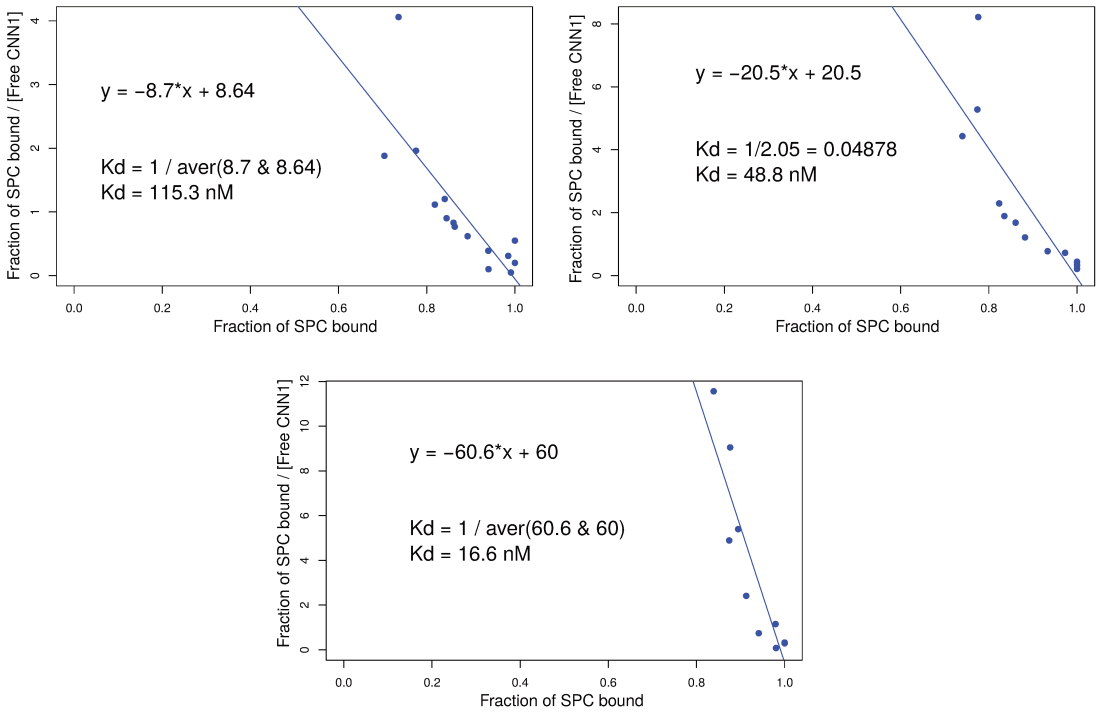


Figure 4.6: K_d estimation of the interaction between a long CNN1 peptide and the SPC24/25 dimer. Top left plot: all cross-links considered. Top right plot: cross-links for which at least one site falls in a SPC globular domain. Bottom plot: cross-links for which both sites fall in globular domains.

4.2.4 Changes in affinity upon the presence of a PTM and a third subunit

In order to assess the applicability of qXL-MS for measuring the relative affinities of two proteins for a complex and to evaluate the change of affinity of one protein in the presence of the other, I quantified the interaction of the Polycomb regulator complex PRC2 to its cofactors AEBP2 and JARID2 from a recently published XL-MS dataset [38]. PRC2 is a histone H3 methyl-transferase complex whose activity is enhanced upon binding of either methylated JARID2 or AEBP2. However, it can also bind to both cofactors simultaneously (Figure 4.7 A). Non-methylated JARID2 binds to PRC2 as a substrate to be modified

at lysine 116, but its allosteric effect on the complex has not been reported. I wanted to determine which cofactor, JARID2 or AEBP2, has a higher affinity for PRC2. To this end, I quantified the cross-links detected within PRC2 co-purified with either non-methylated JARID2 or AEBP2 as baits and found that non-methylated JARID2 bound between 3.5 to 6.25 times more efficiently to PRC2 (Figure 4.7 B).

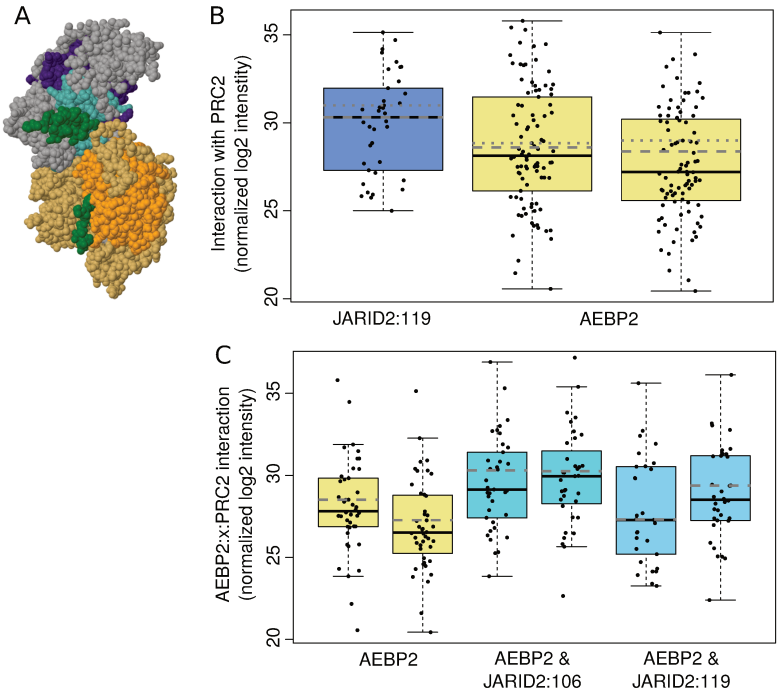


Figure 4.7: Estimation of relative affinities between the PRC2 complex and its cofactors AEBP2 and JARID2. A. Structure of PRC2 in complex with its 2 cofactors AEBP2 and JARID2 (PDB 6C23): JARID2 in dark green, AEBP2 in cyan, SUZ12 in grey, EED in orange, EZH2 in kaki and RBAP48 in violet. B. Comparison between the affinity of non-methylated JARID2 for PRC2 (JARID2:119) and the affinity of AEBP2 for PRC. Boxplots with the same color indicate replicates. Black median: estimation based on cross-links within and to PRC2. Dashed median: estimation based on cross-links within PRC2. Pointed median: estimation based on intra-protein cross-links within PRC2. C. Change in affinity of AEBP2 for PRC2 in the presence of methylatable and non-methylated JARID2. Black median: estimation based on cross-links of AEBP2 to PRC2. Dashed median: estimation based on cross-links of AEBP2 to PRC2 in common between runs.

As both cofactors can simultaneously bind to PRC2, I also wanted to evaluate how AEBP2 affinity for PRC2 changes in the presence of the non-methylated and methylated versions of JARID2. I found that AEBP2 affinity increases by 5.22 fold in the presence of methylated JARID2, but remains almost the same with a 1.36 fold change in the presence of non-methylated JARID2 (Figure 4.7 C).

The relative affinity estimations are in agreement with the literature. The K_d of AEBP2 peptide 379-390 for the PRC2 subunit RbAp48 has been previously measured to be 7.6

μM [102], whereas, the affinity of JARID2 peptide 110-122 K116me3 has been found to be higher with a K_d of about 3 μM [82]. Kasinath et al [38] mentioned that their electron microscopy structures show that JARID2 stabilizes the interaction between AEBP2 and PRC2. The calculations presented here not only corroborate this, they also measure the increment in the stabilization.

Taken together, this case shows that qXL-MS can be employed to compare the affinities of two proteins for a third one and to asses affinity changes in the presence of another subunit.

4.3 Discussion

Here I described a method to estimate constants of dissociation in protein interactions using qXL-MS. To my knowledge, Maedler et al [56] carried out the only previous attempt in this direction. The authors found that the amount of cross-linked complex correlates with the binding affinity of protein interactions with K_ds between 30 nM and 25 μM. For cases above 25 μM, unspecific cross-linking between not real interactors may as well occur [56]. Affinities in ranges lower than 30 nM were not investigated and K_d values could not be estimated. The method proposed in this chapter corroborates the correlation of protein complex amount and cross-links abundances, and moreover, it shows that the relation of intra-/inter-protein cross-links can be used to estimate K_d values.

Is it correct to use intra- and inter-protein cross-links alone to estimate the amounts of the free subunits and the complex? Would the incorporation of mono-links and loop-links improve the analysis? The short answer to the latter question is no. The estimated K_d values deviated considerably from the benchmarks when these types of links were considered in the analysis. They increased by more than a factor of 10. Nonetheless, I observed that their incorporation improved the fit of experimental intensities to the expected changes in the concentration of the varying subunit. However, the deviation of the ratio to the constant subunit with respect to the expected ratios increased. Moreover, incorporating these type of links may be detrimental to the estimation of the amount of complex, as it is impossible to decide whether a mono-link or loop-link that involves a lysine site that was also observed in inter-protein cross-links occurred in the free-subunit or in the complex species. In other words, it is apparently better to restraint the analysis to the use of intra-/inter-protein cross-links for estimating the K_d and relative affinities.

How applicable is the method presented here? The kinetic equation for the complex model used in the present study assumes that the 2 subunits (CNN1 and SPC24/25) associate at a 1:1 stoichiometry, which is indeed the case as proven by X-ray structures (PDB: 4GEQ). For complexes where the stoichiometry differs from this relation, the method should still be applicable, because such cases should affect the number of identified inter-protein cross-linked sites, but not their average intensity. Other advantages of the method are that the reaction is performed in solution and thus steric effects do not pose limitations and no additional tagging or labeling that could affect the affinity is performed on the protein sequences. The method requires small sample amounts and is not limited by the size of

the subunits. In principle, higher affinities could be measured as it has been shown that mass spectrometers could measure analytes from the μM to the fM range. However, this method cannot measure the forward and reverse kinetic constants, but only the relation of them in the equilibrium state. Moreover, measurements do not occur in real time. Thus, the method is limited by technical variabilities that could be introduced during digestion, peptide clean-up or SEC/MS analysis. The method is also limited by the presence of cross-linkable amino acids at or proximal to the interface. Therefore, experiments using a variety of cross-linkers with different spacer lengths and reactive groups may boost the applicability and performance of the method.

Taken together, I demonstrated the feasibility of determining K_ds on dimeric/trimeric complexes using qXL-MS data. Properly normalized inter-protein cross-link intensities can facilitate the characterization of relative binding affinities or even the estimation of absolute K_d values. Furthermore, the proposed approach provides a unique method for following relative affinities of several binding interfaces in multimeric complexes simultaneously.

4.4 Materials and Methods

Expression and Protein Purification of SPC24/25 and the CNN1 peptide

For the expression of the budding yeast Spc24/25 complex in E. coli the respective genes were amplified from genomic DNA and cloned into the pETDuet-1 vector (Novagen). Expression and purification of the Spc24/25 complex were performed as described previously [46]. In brief, pDuet1-Spc24-6xHis/Spc25 was transformed into E. coli strain BL21 DE3 (EMD Millipore). Bacteria were grown to an OD600 of 0.6 at 37C and protein expression was induced with 0.2 mM IPTG for 18 h at 18C. Cells were lysed in lysis buffer (30 mM HEPES, pH 7.5, 300 mM NaCl, 5% glycerol, 30 mM imidazole, 5% glycerol, Complete EDTA-free protease inhibitors [Roche]) and the cleared lysate was incubated with Ni-NTA agarose beads (Qiagen). The protein complex was eluted with buffer containing 30 mM HEPES (pH 7.5), 150 mM NaCl, 0.01% NP40, 2% glycerol and 250 mM imidazole. The Spc24/25 complex was further purified on a Superdex 200 HiLoad 16/60 column (GE Healthcare) applying 30 mM HEPES (pH 7.5), 150 mM NaCl and 5% glycerol as the mobile phase.

For the CNN1 peptide, the respective nucleotide sequence was cloned into Insect cells. Cells were lysed in buffer containing 30 mM HEPES (pH 7.5), 400 mM NaCl, 10% glycerol and protease inhibitor cocktail (Roche) using a cell disruptor at 18000 psi. The complex was purified on Ni-NTA resin (Qiagen) and eluted in 30 mM HEPES (pH 7.5), 150 mM NaCl, 5% glycerol and 250 mM imidazole. The eluate was further purified on a Superdex 200 HiLoad 16/60 column (GE Healthcare) applying 30 mM HEPES (pH 7.5), 150 mM NaCl and 5% glycerol as the mobile phase.

Complex titration, chemical cross-linking and mass spectrometry

Purified in vitro reconstituted dimers and peptides were titrated in different molar ratios and incubated for 45 min to allow complex formation. The SPC dimer concentration was kept constant, while the CNN1 peptide concentration varied to fit the following molar ratios: 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 3.25, 3.5, 4.0, and 4.5. Subsequently, protein complexes were cross-linked by addition of an equimolar mixture of isotopically light (hydrogen) and heavy (deuterium) labeled bis[sulfosuccinimidyl]suberate (BS3, H12/D12) (Creative Molecules). BS3 was added at a final concentration of 2 fold the total protein concentration and let react at 30C for 6 min. The crosslinking reaction was quenched by adding ammonium bicarbonate to a final concentration of 100mM for 20 min at 30C. Samples were then reduced with 5mM TCEP (Thermo Fisher Scientific) at 35C for 15min and alkylated with 10mM iodoacetamide (Sigma-Aldrich) at room temperature for 30 min in the dark. Proteins were digested with Lys-C (1:50 (w/w), Wako Pure Chemical Industries) at 35C for 2 h, diluted with 50 mM ammonium bicarbonate to 1 M urea, and digested with Trypsin (1:50 (w/w), Promega) overnight. Peptides were acidified with trifluoroacetic acid (TFA) at a final concentration of 1% and purified by reversed phase chromatography using C18 cartridges (Sep-Pak, Waters).

Cross-linked peptides were enriched by size exclusion chromatography on a Superdex Peptide PC 3.2/30 column using water/acetonitrile/TFA (77.4/22.5/0.1, v/v/v) as mobile phase at a flow rate of 50l/min. Fractions typically containing cross-linked peptides were analyzed by liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) using a nano-HPLC and an LTQ-Orbitrap Elite instrument. A flow rate of 20 nl/min at incremental gradients of buffer B from 3% to 98% was used. At each MS cycle, the top 10 intense peptides with charges >2 were selected for fragmentation and MS2 scan, with exclusion times of 30 s. MS1 spectra were acquired in the Orbitrap analyzer at 12K resolution, and MS2 fragment scans at low resolution in the ion trap.

Identification of cross-linked spectra

Raw spectra were converted to mzXML format using MSConvert from the ProteoWizard suit tools and analyzed with xQuest/XProphet for the identification of cross-linked peptides. Peptide spectrum matches were performed against a database containing the subunits of the complex in question (i.e., SPC24, SPC25 and CNN1) and 22 E. coli decoy sequences. A maximum of 2 trypsin missed cleavages was allowed, and peptide lengths between 4 and 45 amino acids. Carbamidomethyl-Cys was set as a fixed modification and a mass shift of 138.068 for intra-/inter-protein cross-link candidates with an additional shift of 12.075321 to account for cross-links with the heavy version of BS3. A precursor mass tolerance of -/+ 10 ppm was used and a tolerance of 0.2 and 0.3 Da for linear and cross-linked fragment ions, respectively. The search was performed in the so-called 'ion-tag' mode. Identifications were filtered at the Xquest score threshold of 25; precursor errors above 5.0 ppm were filtered out; a maximum of 0.95 delta score was allowed and a minimum of 3 ions matches per peptide was imposed. Final identification tables were downloaded as xtract.csv-formatted files from the xQuest/xProphet visualization tool.

Quantification of cross-linked peptides

Quantification was performed with an in-house established workflow implemented in the OpenMS software version 2.0 and described in the following lines. Identifications contained in the xtract.csv files were converted to idXML format using our house script xtractToIdXML.py. Files in the mzXML format were converted to mzML using the FileConverter function with default parameters, except for the filter of MS2 scans and MS1 peaks with intensities <100.0. Peak features in the mzML files and their respective profile chromatograms were extracted with an in-house modified version of the FeatureFinderAlgorithmPicked tool from OpenMS. Parameters fed to this tool are found in the file 'ffc_params.ini'. Detected features were then annotated with their putative peptide identifications in the idXML files using the IDMapper function with an m/z tolerance of 7 ppm and RT tolerance of 10 s. Retention times between runs were aligned using the MapAlignerIdentification function with default parameters. Finally, consensus tables were generated using the FeatureLinkerUnlabeled function with default parameters and converted to CSV format with the TextExporter function. The intensity of the quantified peptide ions was summarized to protein-protein cross-linked sites using an in-house script.

Estimation of K_d

Protein-protein cross-linked sites intensities were loaded and analyzed in the statistical environment R as described in the following lines. Technical replicates were averaged, with non-assigned values being ignored in this step. The intensities of peptides seen in >1 SEC fraction were summed up, and peptide-peptide cross-links were summarized to protein site-site cross-links by addition of their intensities. The intensities of the subunit whose concentration was constant in all titrations were used to normalize the intensities between runs. Finally, a linear model was fitted between the initial concentrations of the varying subunit and the intensities of its intra-protein cross-links. This linear relation was used to predict the concentration of the complex from the median intensity of the inter-protein cross-links. Subsequently, the constants of dissociation were calculated as in the kinetic equation shown in the RESULTS section. The initial concentrations of the protein subunits were recalculated based on the linear relation of Concentration and Intensity. The estimated complex concentration was subtracted from the initial concentrations to obtain the amount of the free subunits. For each titration, a K_d was calculated, and the mean and standard deviation of these values was reported.

We also used a more common method to estimate the K_d, namely plotting the linear relation of 'fraction of SPC bound over concentration of free CNN1' (y-axis) versus 'fraction of SPC bound' (x-axis). In this approach, the K_d should equal the negative inverse of the slope as well as the inverse of the intersection coefficient.

Relative affinities in the AEBP2-JARID2-PRC2 complex

Raw files from the relevant experiments were directly downloaded from the PRIDE repository with entry number PXD008605. Cross-links were identified with xQuest/xProphet

with the same parameters specified in the authors publication. Quantification was performed as explained above. Match between runs was applied whenever replicates were available.

Intensities were summarized to protein-protein cross-linked sites, which were normalized by either the median intensity of the final bait in order to control for the initial abundance of AEBP2 or JARID2:106 or by the sum of both medians (in the case of the double pull-down, where AEBP2 and JARID2:119 have the same flag tag). The affinity for PRC2 was determined using either cross-links within and to PRC2 or cross-links within PRC2 or intra-protein cross-links within PRC2 (Figure 4.7 B) or, when applicable, using the inter-protein crosslinks between AEBP2 and PRC2 (Figure 4.7 C). For this latter case, the median intensity of AEBP2:x:PRC2 inter-protein crosslinks common across all samples was also computed and indicated in Figure 4.7 C. This was done in order to discount for conformational changes when both cofactors are bound to PRC2.

Conclusion and Outlook

In this doctoral work, I have developed bioinformatics tools and concepts for the molecular characterization of protein complexes through mass spectrometry.

In the second chapter of my thesis, I presented compleXView, a tool for the integration and interpretation of MS-based interactomics data. Previous works on the matter have already combined AP-MS data with Gene Ontology information. My work improves on them by automatizing their ideas into a software tool, which additionally incorporates, for the first time, XL-MS and BioID data in its workflow. Not least, I showed in this chapter that the integrated data is more powerful than any of its sources alone, as regards sensitivity and specificity in the discovery of physical and functional protein associations. Useful insights will be obtained with my tool from small and medium protein interaction studies that use mass spectrometry. I anticipate that future bioinformatics tools will improve by incorporating quantitative information of cross-links in network clustering algorithms, and information from other knowledge databases besides Gene Ontology. All together, this will lead to greater insights and accuracy on the elucidation of protein complexes in PPI networks.

In the third chapter of my thesis, I presented a bioinformatics workflow for the prediction of minimal binding domains in protein complexes. Previous work also used XL-MS data for the same purpose. My approach improves on it by incorporating for the first time quantitation of the cross-links in order to rank protein regions as potential candidates for binding domains and to elucidate dispensable from indispensable regions. As proven, my workflow will facilitate a more educated and data-driven design of deletion mutants in protein interaction experiments. I anticipate that future improvements will automatize the workflow into a software tool that employs a better machine-learning algorithm, which this time will successfully combine the protein sequence-level information with the quantitated cross-links. All together, this will lead to finer tools that predict hot spot residues within the binding domains.

In the last chapter of my thesis, I presented a method for the estimation of protein binding affinities through the quantification of inter-protein crosslinks. Previous efforts in this direction could not achieve the calculation of K_ds. The method proposed here did achieve this in a trimeric complex. Future work will have to expand the capability of the method to measure K_ds in multimeric complexes. Moreover, I anticipate that quantitative XL-MS will be highly useful in the characterization of post-translational modifications that affect protein-binding affinities. Thus, future work in this line will allow the elucidation on how protein subunits assemble collaboratively and dynamically into macromolecular

structures.

Overall, the tools and concepts that were developed here will help the scientific community in the molecular characterization of protein interactions. As a result of this and future work, we will improve our understanding of protein complexes and their vital role in biology and human diseases.

Appendices

Appendix A

Supplement to Chapter 2

A.1 Results

A.1.1 The Minichromosome Maintenance complex and interactors

The Minichromosome Maintenance Complex (MCM) is a protein complex that cleaves hydrogen bonds between DNA double strands to allow the replication of the genome. The proposed physiologically active complex consists of six protein subunits, each with DNA-helicase and ATPase activities. The current model that describes the assembly of the pre-replication complex [118] proposes that two MCM hexamers are loaded sequentially onto an origin of replication. In more detail, one copy of each subunit (from MCM2 to MCM7) assembles into a hexamer ring structure that binds an origin of replication, through its interaction with the Origin Recognition Complex (ORC) and the CDC6 protein. Subsequently, a second MCM hexamer is loaded. In the overall process, CTD1 is important to maintain the stability of each MCM ring and to load them. Eventually, the two rings move in opposite directions, separating the DNA strands along their ways. MCM2-7 proteins have high sequence similarity, but are all and each indispensable for growth and DNA replication. Thus, they are associated with a number of cancer types [70]. The observation of other n-mer MCM complexes and the existence of other MCM proteins (i.e., MCM8 to 10), suggest that these proteins may be involved in other cellular processes. Therefore, studying the interactome of MCM proteins is highly relevant.

To that aim I integrated 3 interactomics data sets from BioID, AP-MS and XL-MS experiments. This data is publicly available at PRIDE under the accession numbers PXD004089 and PXD002987, respectively, where MCM proteins were used as baits. The data shows the implication of MCM proteins in DNA repair and putative involvements in ribosome biogenesis and splicing processes. The integrative approach revealed the protein complexes involved in this network. Because the AP-MS abundance measurements were not suitable for quantification, only the protein identifications were used as an inclusion filter for the identifications in the BioID data set. The BioID data has good quantification

measurements, but contains proteins that are clear false positives (even though their high enrichment respect to the negative control and their high abundance respect to the baits). For example, two of the most highly abundant proteins were LIAS (involved in lipoyl modification of proteins) and ABCAD (involved in transmembrane transport of lipids). Their functionality is clearly out of relation with MCM proteins. Using this filtering, false interactors enriched simply due to physical proximity to the BirA*-bait will be discarded. To still account for transient/weak interactions with nuclear proteins, any BioID-labeled protein with nuclear localization was not excluded.

A note aside: In the rest of this section, when no citation is provided then the information was retrieved from the UniProt or DAVID databases.

Interactors of the MCM2-7 proteins

In total 154 MCM2 putative interactors were identified. Proteins involved in mRNA processing, particularly splicing (41 and 35) were the most enriched. The next category was rRNA processing (29), overlapping with 17 proteins involved in ribosome biogenesis. The next group was DNA replication proteins, which included (apart from the 6 MCM components) 7 proteins: replication factor protein RFC1, origin recognition protein ORC2, and the proteins FANCJ, MRE11, NASP, SP16H and SSRP1. There were 10 helicases in the MCM2 interactome apart from the MCM proteins: the DNA-helicases RUVB2, SMCA5 and FANCJ, as well as the RNA-helicases DDX27/42/46, DHX15, DX39B, IF4A3 and U520. Transcription regulators were also enriched (12, that included the negative elongation factors NELF proteins). Finally, there were 9 proteins involved in DNA damage and repair. Leaving the MCM aside, the 5 most abundant proteins were histone variants H2AW/Y, HNRPF (involved in mRNA processing), CBX3 (part of heterochromatin-like complexes) and SSRP1 (component of the FACT complex, which reorganizes nucleosomes).

In total 158 MCM3 putative interactors were identified. Proteins involved in mRNA processing, particularly splicing (34 and 28) were the most enriched. They were followed by proteins involved in DNA replication (24). There were 17 helicases: the 6 MCM proteins, 5 more DNA-helicases (RUVB1/2, FANCJ, SMCA5 and CHD4) and 6 RNA-helicases (DDX10/42/46, DHX15, IF4A3 and U520). There were 14 proteins involved in DNA damage and repair, and 14 in DNA replication. This latter group contained MCM proteins, DNA polymerases DPOD1/2 and DPOLA, replication factors RFC1/3/4/5 and protein FEN1. There were 42 transcription regulatory proteins (18 of them repressors); 7 proteins involved in ribosome biogenesis and 7 in mismatch repair. Leaving the MCM aside, the 5 most abundant proteins were the general transcription co-activator TCP4, DPOLA, RL27A, PCNP (putative cell cycle regulator), and LRWD1/PARP3 (the first a stabilizer of the ORC; and the second poly ADP-ribose polymerase involved in base excision repair).

In total 120 MCM4 putative interactors were identified. The most enriched category was DNA replication with 22 proteins. Besides the 6 MCM helicases, there were 4 DNA helicases (RUVB1/2, FANCJ, and CHD4) and 4 RNA helicases (DDX10/42/46 and MTREX).

The next category was for proteins involved in mRNA processing and splicing (24 and 19), followed by proteins involved in DNA damage and repair (15). There were 32 transcription regulatory proteins (15 of them repressors). Leaving the MCM aside, the 5 most abundant proteins were TADBP (a DNA-/RNA-binding protein that regulates transcription and splicing), ORC3, NASP (DNA replication protein), WDR46 (component of the nucleolar structure) and AQR (component of the spliceosome).

In total 164 MCM5 putative interactors were detected. The most enriched category was for proteins involved in mRNA processing and splicing (36 and 33), followed by DNA replication (28). There were 22 rRNA processing proteins overlapping with 13 proteins involved in ribosome biogenesis. Apart from the MCM, there were 11 helicases: 5 DNA-helicases (RUVB1/2, SMCA5, FANCJ, and CHD4) and 6 RNA-helicases (DDX27/42/46, IF4A3, MTREX and U520). There were 17 proteins involved in DNA damage and repair; 8 proteins involved in mRNA export from the nucleus and 36 involved in transcription regulation. Leaving the MCM aside, the 6 most abundant proteins were the transcription initiation factor TF2AA, followed by PARP3, ORC2/3, LRWD1, TCP4, and T2EA (transcription factor the recruits the initiation complex).

In total 181 MCM6 putative interactors were detected. The most enriched category was mRNA processing and splicing (35 and 31), followed by 30 rRNA processing proteins overlapping with 14 ribosome biogenesis proteins, and then DNA replication with 20 proteins. There were 13 helicases beside the MCM: G3BP1, 3 DNA helicases (RUVB1/2 and CHD4) and 9 RNA helicases. There were 43 transcription regulators (22 repressors) and 12 proteins involved in DNA damage and repair. Leaving the MCM aside, the 6 most abundant proteins were the ribosomal proteins RS28 and RL27A, MCMBP, H2AY, HNRH1 (heterogeneous nuclear ribonucleoprotein involved in splicing) and PPIA (isomerase that accelerates the folding of proteins).

For MCM7 the least number of interactors were detected: 70. The most enriched category was DNA replication with 10 members (MCM proteins, ORC2/3/5 and DPOD3). There were 6 helicases besides the MCM complex: G3BP1, RUVB1 and the RNA helicases DDX27/46/47 and DX39B. There were 7 proteins involved in protein folding; 9 in mRNA processing and 8 in splicing. Leaving the MCM aside, the 5 most abundant proteins were TADBP, HS90B (a chaperone), PPIA, STMN1 (destabilizes microtubules and prevents their assembly) and TCPZ (chaperone).

What is the stoichiometry of the MCM subunits in the data? As explained before, the MCM holocomplex is a hetero-hexamer with equal ratios of the MCM2-7 proteins. However, one can observe 2 MCM subcomplexes in the data, with more stability than the holocomplex (Figure A.1). These subcomplexes are apparently formed by subunits 2, 4, 6 and 7 on the one hand, and 3 and 5 on the other. Why are the MCM protein abundances different? Is the tag on the N or C-terminus of the protein affecting the assembly or stability of the hexamer? Cdt1 acts as a chaperone that stabilizes the hexameric ring by interacting with the N-termini of MCM2, 4 and 6 [118]. However, Cdt1 was not detected in any of the purifications. The interaction of Cdt1 with the hexamer might be prevented if the exogenous constructs have birA* tagged to the N-termini of the MCM proteins. The original publication of this data set does not report any information about the location

of the tags. In any case, the data suggest that the subcomplexes might exist in natural conditions. The question that arises is whether they are physiologically active or not.

The regulatory protein MCMBP promotes the disassembly of the ring. MCMBP was indeed detected in all purifications, but in some cases was not 2-fold higher than in the negative control. It was detected at ratios <10 percent for MCM2-4, while for MCM5-7 at ratios between 10 and 20 percent. Its highest interaction was with MCM6. It is curious that other MCM proteins (MCM8 to 10) did not purify in any of the experiments. To my knowledge, at least MCM10 may interact with subunits MCM2, 4, 6, and 7 [58].

All in all, the data indicates that MCM proteins may be involved in other processes besides replication, such as mRNA processing/splicing, DNA damage/repair and rRNA processing/ribosome biogenesis.

The BioID data set was obtained under normal cellular conditions and after treatment with etoposide. Etoposide is a drug that forms a complex with DNA and topoisomerase II (TOP2A), causing breaks in the DNA strands. One should expect then the treatment to trigger an increase in the MCM abundance, apart from changes in DNA damage and repair proteins. The data shows that neither the amounts of the MCM holocomplex nor the subcomplexes MCM2/4/6/7 and MCM3/5 change after treatment.

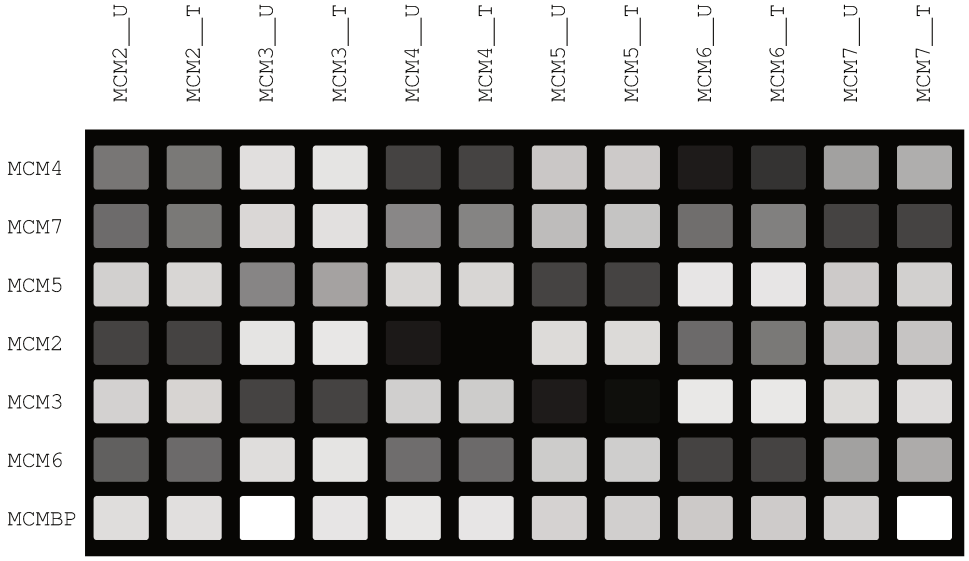


Figure A.1: Blot plot of the abundances of the MCM subunits and MCMBP across purifications of untreated (_U) and etoposide-treated (_T) cells.

Does the interaction to DNA repair proteins change after treatment? Blot plots of this class of proteins showed that indeed 11 of them change by 2 fold or more, but only in the pull-downs of certain MCM proteins (data not shown). In the rest the interaction remains the same. These proteins are SFPQ (down in MCM7), BRD4 (down in MCM5), MSH2 (up in MCM5), CDK9 (down in MCM7), SSRP1 (down in MCM2), SMC1A (down in MCM2; down in MCM5), RIF1 (down in MCM6), UBP7 (down in MCM3), CDC5L (down in MCM2; down in MCM5), SYF1 (down in MCM2 down; up in MCM5), BAZ1B

(down in MCM2), SMC3 (down in MCM2), SP16H (down in MCM2; up in MCM3). SFPQ is a splicing factor; in association with NONO may be involved in DNA non-homologous end joining after dsDNA break repair. BRD4 isoform B is a chromatin insulator of DNA damage response; i.e., it inhibits the response signaling by recruiting condensing-2 complex to DNA regions rich in acetylated histones. MSH2 is a component of the mismatch repair system, which binds to mismatches and initiates DNA repair. CDK9 in association with cyclin-K prevents DNA damage, while isoform 2 in interaction with Ku70/XRCC6 may have a role in DNA repair. SSRP1 is a component of the FACT complex, which reorganizes nucleosomes during DNA replication and repair. SMC1A is involved in chromosome cohesion and is involved in DNA repair via its interaction with BRCA1. RIF1 is a telomere-associated protein that regulates TP53BP1, which is a positive regulator and key for dsDNA repair. UBP7 is recruited to DNA damage sites to promote the deubiquitination of ERCC6, a protein positively involved in the transcription-coupled nucleotide excision repair. CDC5L is a DNA-binding protein that is required for the efficient expression and splicing of genes involved in DNA damage response [65]. SYF1 is involved in the transcription-coupled repair of DNA. BAZ1B is a kinase that phosphorylates H2AX at Y142, which is central for DNA repair. SMC3 is a component of cohesin. Cohesion presumably prevents DNA repair by impairing the use of strands as templates for repair [68]. SP16H is also a component of the FACT complex, which promotes the dissociation of nucleosomes and their subsequent reestablishment. Taken together, it seems that proteins that promote DNA repair interact more with MCM3 and MCM5 after the damage induced by etoposide, but less with MCM7, while MCM2 interacts less with both positive and negative effectors of DNA repair.

During the functional analysis it was observed that proteins associated with telomere, heterochromatin and remodelers were mildly enriched. The blot plots of these proteins showed that the abundance of 9 of them change after treatment: H2AY (down in MCM2; up in MCM3; down in MCM5; up in MCM6), H2AW (down in MCM2; down in MCM5; up in MCM6), SMCA5 (down in MCM2), CBX1 (down in MCM2), CBX3 (down in MCM2; up in MCM3; up in MCM6), DEK (down in MCM3; down in MCM7) and BRD4 (down in MCM5). Histone variants H2AY/W are proper of heterochromatic nucleosomes. SMCA5 is an essential component of the nucleolar-remodeling complex that leads to the formation of heterochromatin. CBX1 is a usual component of heterochromatin. CBX3 seems to be part of heterochromatin-like complexes, and can also recruit NIPBL to DMA damage sites; NIPBL has a role in cohesin loading to these sites. DEK changes the topology of DNA by inserting positive supercoils. Taken into account the DNA damage context after treatment, it seems logical that the chromatin should be in a more open state to allow repair. Strikingly, heterochromatic proteins interact more with MCM3 and MCM6 after treatment, but less with MCM2 and MCM5.

What is the effect of etoposide on the other biological processes that were enriched across the pull-downs? Blot plots showed that mRNA processing proteins tend to interact less with MCM2 after treatment; more with MCM3; no change with MCM4; less with MCM5; more with MCM6; and mildly less with MCM7. Regarding rRNA processing and ribosome biogenesis proteins, they tend to interact less with MCM2 after treatment; no change with MCM3; no change in MCM4; less with MCM5; mildly more with MCM6; and both

ways with MCM7. On the other hand, DNA replication is apparently not affected by the treatment. Regarding transcription regulation proteins, their interaction with MCM2 decreases; with MCM3, 3 interactions change up and 3 down; with MCM4, only 2 interactions change up; with MCM5, 5 increase and 1 decreases; with MCM6, 4 proteins increase interaction; with MCM7, 6 change up and 3 down.

Detection of protein complexes in the MCM network

The use of correlation of abundances combined with GO similarities and XL-MS information, revealed a number of complexes in the MCM network related with DNA replication and repair (Figure A.2). The MCM subcomplex MCM2/4/6/7 is validated by the cross-linking data, and its interaction with MCMBP is revealed (Cluster C1). Moreover, the subcomplex interacts via MCM7 with a group of RNA helicases (Cluster C2) and the DNA-helicases RUVB1 and 2 (Cluster C3). The MCM3/5 subcomplex interacts with ORC3 and 5, and with 2 DNA-polymerases involved in repair: FANCJ and SMCA5 (Cluster C4). RBBP7, which is also in the cluster, is a protein required for the reorganization of chromatin after replication. Replication factors (RFC proteins) cluster with DNA polymerase delta subunits and CTF18 (Cluster C5). CTF18 is known to form a complex with RFC proteins that binds to ssDNA to load the proliferating cell nuclear antigen-sliding clamp, which in turn is important for the recruitment of DNA polymerases during replication and repair. The origin of recognition protein ORC1 clusters with the kinase CDC7 (which regulates DNA replication by phosphorylation) and with the mismatch repair protein MLH1 (Cluster C6). The structural maintenance of chromosomes proteins SMC1A and SMC3 also formed a cluster (C7); both proteins are required for cohesion of sister chromatids. Cohesion is coupled to DNA replication and is involved in DNA repair. In the cluster formed by HNRPK and MEF2D (Cluster C8), HNRPK is both a coactivator and corepressor of p53 response to DNA damage, while MEF2D is a transcriptional activator that decreases etoposide-induced damage [8]. The proteins SP16H and SSRP1, both components of the FACT complex (which reorganizes chromatin to regulate processes such as transcription elongation and DNA replication and repair), cluster with Z280C, whose function is unknown, but might be a transcription factor (Cluster C9).

Complexes whose members are involved in rRNA processing and ribosome biogenesis are also detected. The PeBow complex (PESC and WDR12) associates with SFBP1 (Cluster C10); the three proteins are required for the maturation and processing of rRNA. Another cluster of rRNA processing proteins is formed by IMP3/4, NH2L1, UT14A and WDR3 (Cluster C11). Aside, WDR43 and UTP15, both involved in ribosome biogenesis, also form a group (C12). Similarly, another set of this kind of proteins cluster with PCID2, which is a component of the TREX2 complex, responsible for the export of ribonucleoproteins from the nucleus (Cluster C13). The rRNA processing proteins in this cluster include EBP2, DCA13, RRP9, TBL3 and UTP11; while the ribosome biogenesis proteins include HEAT1, NOL11, NOP2 and WDR74. Another cluster is formed by the ribosome biogenesis proteins BMS1 and NOG1 together with the splicing protein U5S1 (Cluster C14). Two components of the cleavage and poly-adenylation specificity factor (proteins CPSF1 and FIP1) cluster with WDR46, which is involved in rRNA processing (Cluster C15). Another cluster is

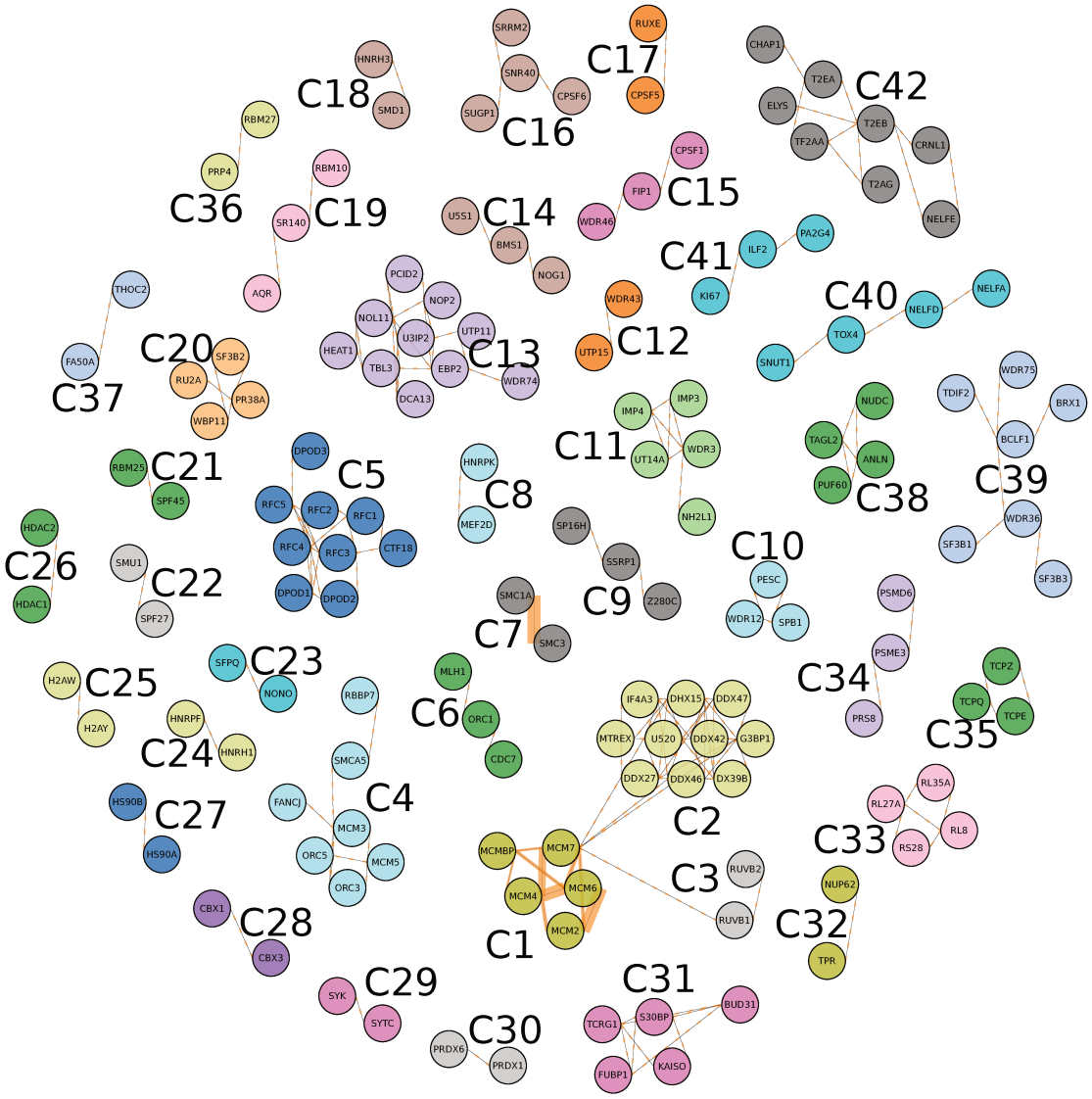


Figure A.2: Complexes and clusters identified in the MCM data set by the combination of Abundance Correlation and GO functional similarities followed by MCL clustering and incorporation of cross-links.

represented by the association of the splicing proteins SRRM2, SUGP1 and SNR40 with the mRNA cleavage and the poly adenylation specificity factor CPSF6 (Cluster C16). Aside a component of the cleavage and polyadenylation specificity factor (protein CPSF5) clusters again with a protein involved in splicing (RUXE), which reflects the interrelation of both biological processes (Cluster C17).

Other RNA splicing proteins also show good correlations. For example, proteins HNRH3 and SMD1 formed their own group (C18), and so do the splicing proteins AQR, RMB10 and SR140 (C19). Similarly do proteins RU2A, SF3B2, WBP11, and PR38A (Cluster 20); RMB25 and SPF45 (Cluster 21); and the proteins SMU1 and SPF27, which are specifically

involved in splicing of mRNAs (Cluster 22). The splicing factors SFPQ and NONO, whose interaction is known to be essential for mRNA splicing formed their own group (C23). And two components of the heterogeneous nuclear ribonucleoprotein (hnRNP) complexes do the same: proteins HNRPF and HNRH1 (Cluster 24). hnRNP complexes are required for the processing and maturation of pre-mRNAs to become translatable; this includes regulation of alternative splicing events.

Other complexes in the network include heterochormatin related complexes. Histones H2AW and H2AY, proper of heterochromatic nucleosomes, formed their own group (C25). Histones deacetylases 1 and 2 do the same (Cluster C26). And also the heat shock proteins 90A and 90B, which constitute a group of chaperones for histone deacetylases and DNA methyltransferases (Cluster C27). The chromobox protein homologs CBX1 and 3 also form their own cluster (Cluster C28). These proteins are involved in transcription silencing via the formation of heterochromatin-like complexes.

Additional clusters include, the amino acid t-RNA ligases SYK and SYTC (Cluster C29), and the peroxiredoxin proteins PRDX1 and 6 (Cluster C30). The group formed by BUD31, FUBP1, KAISO, S30BP and TCRG1 represent a cluster of transcription regulators and factors (Cluster C31). Nucleoporins TPR and NUP62 also group (C32), and so do ribosomal proteins RL8, RL27A, RL35A and RS28 (Cluster C33). The 26S proteasome complex is partially revealed by the cluster formed by PRS8, PSMD6 and PSME3 (Cluster C34). And so is the TRiC complex, revealed here by 3 TCP proteins (Cluster C35).

Other clusters include proteins of unknown function. For example, PRP4 and RBM27 are two RNA binding proteins that form a group (C36). While the former is a component of the splicesome, the function of the latter is unknown. Similarly, THOC2 and FA50A cluster together (C37). While THOC2 is a component of the TREX complex (which is responsible for the translocation of spliced mRNAs to the cytoplasm), FA50A is a putative DNA binding protein of unknown function.

Some clusters have subgroups with apparently unrelated (or not closely related) functions. For example, the cluster formed by ANLN, NUDC, PUF60 and TAGL2 (Cluster C38) is a group of cadherin binding proteins, but each with different biological roles. Splicing factors SF3B1 and SF3B3 group with ribosome biogenesis proteins WDR36/75 and BRX1, as well as with the transcription regulators BCLF1 and TDIF2 (Cluster 39). The cluster formed by the negative factors of transcription elongation (NELF proteins) with TOX4 and SNUT1 seems also to be false (Cluster 40). TOX4 is a component of a phosphatase complex that regulates chromatin structure and cell cycle progression, whereas SNUT1 is involved in mRNA splicing. Similarly, the cluster formed by PA2G4, ILF2 and MKI67 seems to be a false positive as their members have not similar functional relations (Cluster 41). PA2G4 is a co-repressor of transcription, particularly of E2F1-regulated genes. It also associates with rRNA and is thought to be involved in rRNA processing. ILF2, together with ILF3 (not in the network), is a transcription regulator of the gene interleukin 2, which is required for T-cell proliferation. KI67, however, is a protein that acts as a chromosome surfactant preventing their agglomeration into a single chromatin mass. Finally, cluster C42 formed by the negative transcription factor NELFE and the positive transcription factors T2AG, T2EB, TF2AA and T2EA, contains the functionally unrelated proteins CRNL1 (involved

in splicing), ELYS (required for nucleopore assembly) and CHAP1 (required for proper chromosome alignment during metaphase and to maintain the attachment of the mitotic spindle to the kinetochore).

The unclustered proteins in the network comprised 24 proteins involved in transcription regulation, 23 proteins involved in mRNA processing (18 of them involved in mRNA splicing), followed by 13 proteins involved in DNA damage and repair, 7 in DNA replication and 5 in rRNA processing and ribosome biogenesis.

Taken together, the combination of abundance correlations with GO similarities and cross-linking data provides context and validation to the complexes observed in the network and facilitates the biological interpretation of the data.

Appendix B

Original publication of compleXView

W276–W284 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* *Published online 12 May 2017*
doi: 10.1093/nar/gkx411

compleXView: a server for the interpretation of protein abundance and connectivity information to identify protein complexes

Victor Solis-Mezarino and Franz Herzog*

Gene Center and Department of Biochemistry, Ludwig-Maximilians-Universität München, 81377 Munich, Germany

Received March 03, 2017; Revised April 19, 2017; Editorial Decision April 28, 2017; Accepted May 08, 2017

ABSTRACT

The molecular understanding of cellular processes requires the identification and characterization of the involved protein complexes. Affinity-purification and mass spectrometric analysis (AP–MS) are performed on a routine basis to detect proteins assembled in complexes. In particular, protein abundances obtained by quantitative mass spectrometry and direct protein contacts detected by crosslinking and mass spectrometry (XL–MS) provide complementary datasets for revealing the composition, topology and interactions of modules in a protein network. Here, we aim to combine quantitative and connectivity information by a webserver tool in order to infer protein complexes. In a first step, modeling protein abundances and functional annotations from Gene Ontology (GO) results in a network which, in a second step, is integrated with connectivity data from XL–MS analysis in order to complement and validate the protein complexes in the network. The output of our integrative approach is a quantitative protein interaction map which is supplemented with topological information of the detected protein complexes. *compleXView* is built up by two independent modules which are dedicated to the analysis of label-free AP–MS data and to the visualization of the detected complexes in a network together with crosslink-derived distance restraints. *compleXView* is available to all users without login requirements at http://xvis.genzentrum.lmu.de/compleXView.

INTRODUCTION

Proteins interact and build up complexes in order to execute their function rather than acting as individual proteins. The assembly of complexes is a dynamic and highly regulated process which ensures that the protein function is exerted at the proper cellular localization and time. Thus, elucidating the molecular mechanisms of cellular processes requires the biochemical analysis of the involved proteins and their interactions in a signaling pathway.

Affinity purification coupled to mass spectrometry (AP–MS) is a widely used technique to detect protein interactions in biological samples. The identified interactors of a certain bait protein are called preys and their abundances are obtained from the respective peptide intensities by mass spectrometry. In addition, recent efforts have combined chemical crosslinking and mass spectrometry (XL–MS) for the identification of proteins which directly contact each other or are in close proximity within a complex and thus, crosslinks provide topological information. In most cases, XL–MS studies apply amine reactive crosslinking agents to covalently link lysine residues and dedicated software to identify the crosslinked lysines from fragment ion spectra (1, 2).

Affinity-purifications of protein complexes are usually contaminated with unspecific proteins depending on the purification protocol, affinity-tag or cell line. To separate contaminants from interacting proteins is crucial for determining the protein complex composition. Negative control samples are used together with statistical methods to filter out spurious interactions. A frequently used method is SAINT (significance analysis of interactome) (3), which models the abundances of protein identifications in the negative and positive samples into a mixture probability distribution that calculates the odds of an interaction being true rather than false. Additional software programs like MiST (mass spectrometry interaction statistics) (4) and compPASS (comparative proteomic analysis software suite) (5), measure the abundance, reproducibility and specificity of the identification, and combine those into a probability score of interaction. In all three methods, scores above certain thresholds indicate the prey as an interactor of the bait and represent the bait–prey interactions in a table depicting the abundance values of the preys.

There are two different approaches for modeling network topology in the population of interactions: the Spoke model and the Matrix model (6). The Spoke model displays a network as a wheel-like arrangement of baits connected to

*To whom correspondence should be addressed. Tel: +49 89218076937; Email: herzog@genzentrum.lmu.de

multiple preys through spokes lacking connectivity between proteins. Thus, no higher-order structures and very few protein clusters are observed in this kind of network. In contrast, in the Matrix model the input data is first transformed in order to infer interactions between preys, which results in a network with higher-order structures and protein clusters. However, the number of false interactions is proportionally amplified to the size of the dataset.

Approaches for inferring prey–prey interactions include profile correlation, socio-affinity index (7) and hypergeometric probabilities (8). The profile correlation method assumes that protein complexes are regulated and perturbed as a single entity where changes in subunit abundances will change others accordingly. Thus, high correlation in the co-variation of abundances across the different purifications is expected. In the second method, the socio-affinity index measures the number of times two proteins appear in the same purification relative to their frequency in the whole dataset. Other methods rely on machine learning algorithms and require large datasets, bona-fide complexes for training, and the derivation of loose explanatory variables based on measures of abundance, co-purification, and reproducibility (9).

The majority of protein interaction studies includes less than a few tens of baits turning abundance profile correlations into the most appropriate method for the identification of prey–prey interactions as other approaches are tailored to cope with hundreds of baits (7–9).

Subsequent to calculating a measure of interaction strength, proteins are displayed in a network and clustered by different algorithms in order to infer protein complexes and submodules. Clustering algorithms either use properties inherent to the network or introduce prior knowledge into their models. Algorithms such as force-layout, Markov Clustering (MCL) (10) and Molecular Complex Detection (MCODE) (11) belong to the first category and apply the calculated interaction strengths and local connectivity within the network to group proteins into clusters. Algorithms such as CORE (12) and WCOACH (13) belong to the second category, which either adhere to the protein-complex-organization model (7) or use Gene Ontology (GO) functional annotations to weight the membership of a protein in a cluster.

Here, we introduce *compleXView* a webserver that calculates measures of abundance, reproducibility and specificity derived from AP–MS experiments to discriminate true from false bait–prey interactions. Prey–prey interactions are predicted and quantified based on the profile correlation method and these values together with GO functional similarities are supplied to an MCL algorithm. The webserver integrates crosslink data to complement and validate the predicted interactions and to provide connectivity information within and between complexes in a network. *compleXView* is an extension of the previously described *xVis* webserver (14) and facilitates the generation of protein interaction tables at every step and visualizes the network of protein complexes as interactive maps.

MATERIALS AND METHODS

Datasets

Two datasets from previous studies were analyzed, each include label-free quantification of protein abundances and the identification of chemical crosslinks by mass spectrometric analyses.

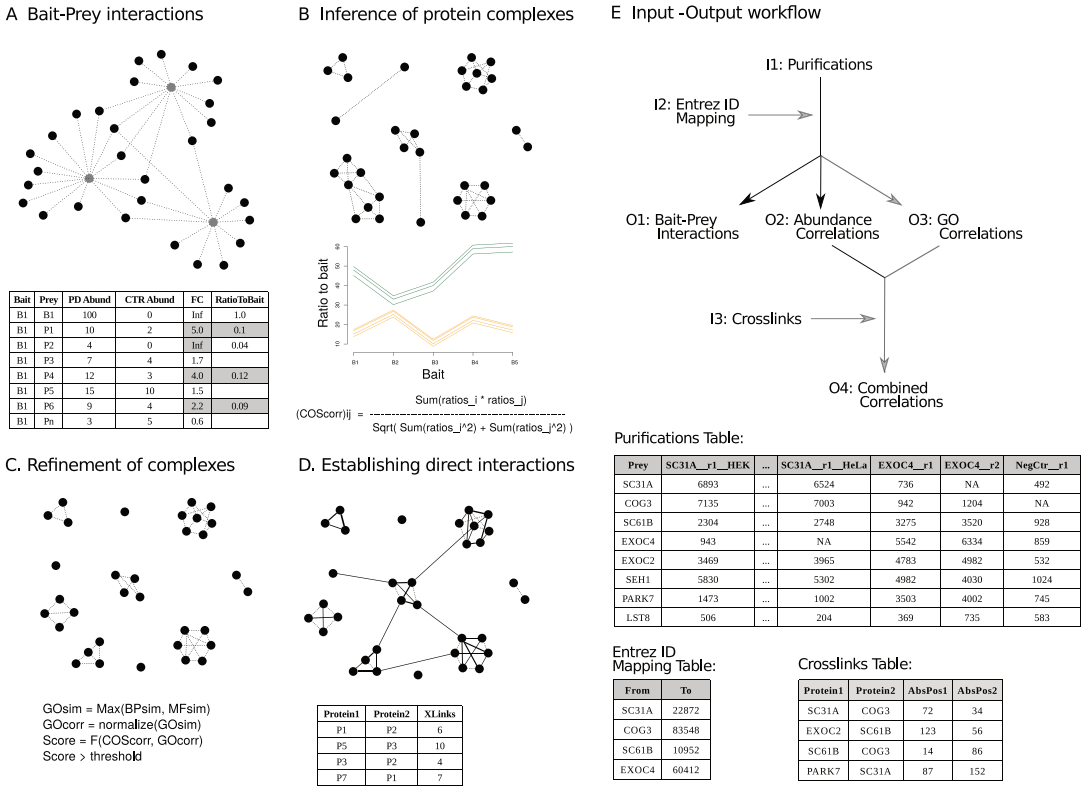
The first dataset (15) comprises affinity-purifications of 14 different bait proteins of the protein phosphatase 2A (PP2A) network, including: PP2A catalytic subunit alpha (PP2AA), PP2A catalytic subunit beta (PP2AB), PP2A regulatory subunit A beta (2AAB), PP2A regulatory subunit B alpha (2ABA), PP2A regulatory subunit B gamma (2ABG), PP2A regulatory subunit delta (2A5D), PP2A regulatory subunit epsilon (2A5E), PP2A regulatory subunit gamma (2A5G), protein phosphatase 4 catalytic subunit (PP4C), Immunoglobulin-binding protein 1 (IGBP1), Shugoshin-like 1 (SGOL1), CTTNBP2 N-terminal-like protein (CT2NL), Striatin-interacting protein 2 (FA40B or STRP2) and FGFR1 oncogene partner (FR1OP).

The second dataset (16) includes five bait proteins of distinct complexes which are associated with DNA including: ribose-phosphate pyrophosphokinase 1 (PRPS1); DNA replication licensing factor MCM6; structural maintenance of chromosomes protein 1A (SMC1A); structural maintenance of chromosomes protein 3 (SMC3); and X-ray repair cross-complementing protein 6 (XRCC6).

Data analysis

In order to quantify peptide abundances in the PP2A dataset raw files were analyzed with MaxQuant version 1.5 (17) at 1% FDR. For the second dataset (16) MaxQuant tables were directly retrieved from their respective PRIDE repository locations (PXD002987).

In order to identify and quantify putative interactors of the bait proteins, raw peptide intensities obtained by MaxQuant were analyzed within the statistical environment R (18). Only unique peptides and proteins with a minimum of two identified peptides were considered for quantification. Median normalization between experiments was performed at the peptide level. Normalized peptide intensities were averaged within replicates in order to obtain protein abundances. Protein identifications were required to be present in at least two replicates of the respective bait. For the PP2A dataset, a plausible set of contaminants was downloaded from the CRAPome database version 1.1 (19), applying the following filters: cell/tissue type, HEK293; epitope tag, Strep-HA; subcellular fractionation, total cell lysate; affinity approach, streptactin; fractionation, 1D LC–MS; and instrument, LTQ-Orbitrap. Proteins observed in six or more CRAPome datasets were considered as contaminants. Protein identifications present in this list were filtered out as well as ribosomal proteins. Protein abundances across the same bait purifications were averaged and the significance of their fold-changes to the negative control was assessed by the Student's *t*-test. Protein identifications were regarded as interactors if their enrichment to the negative control was at least twofold and significant with a Benjamini–Hochberg adjusted *P*-value of 0.05. The abundance ratios to the respective bait were calculated and in-



Bait	Prey	PD Abund	CTR Abund	FC	RatioToBait
B1	B1	100	0	Inf	1.0
B1	P1	10	2	5.0	0.1
B1	P2	4	0	Inf	0.04
B1	P3	7	4	1.7	
B1	P4	12	3	4.0	0.12
B1	P5	15	10	1.5	
B1	P6	9	4	2.2	0.09
B1	Pn	3	5	0.6	

Protein1	Protein2	XLinks
P1	P2	6
P5	P3	10
P3	P2	4
P7	P1	7

Prey	SC31A__r1__HEK	...	SC31A__r1__HeLa	EXOC4__r1	EXOC4__r2	NegCtr__r1
SC31A	6893	...	6524	736	NA	492
COG3	7135	...	7003	942	1204	NA
SC61B	2304	...	2748	3275	3520	928
EXOC4	943	...	NA	5542	6334	859
EXOC2	3469	...	3965	4783	4982	532
SEH1	5830	...	5302	4982	4030	1024
PARK7	1473	...	1002	3503	4002	745
LST8	506	...	204	369	735	583

From	To
SC31A	22872
COG3	83548
SC61B	10952
EXOC4	60412

Protein1	Protein2	AbsPos1	AbsPos2
SC31A	COG3	72	34
EXOC2	SC61B	123	56
SC61B	COG3	14	86
PARK7	SC31A	87	152

Figure 1. Workflow of the *compleXView* 'Analysis' module. (**A**) bait–prey interactions are determined upon enrichment over the negative control and their relative abundance to the bait (PD, pull-down; CTR, control; FC, fold change). (**B**) Pairwise cosine correlations of prey abundance ratio profiles are used to infer interactions between preys. Subunits of a complex are expected to exhibit similar relative abundances to the bait across different bait purifications. Abundance correlations above a certain threshold value are selected for clustering the proteins into modules. (**C**) To eliminate spurious high correlations between two proteins, GO functional similarities between preys are used to refine the protein–protein interactions identified in the previous step. Highly correlated proteins with notably different molecular functions are scored lower. The combined score improves the resolution of the protein complexes in the network. (**D**) Protein interactions are inferred from quantitative AP–MS data. The final analysis step integrates direct protein interactions detected by XL–MS into the network and thereby, validates protein complexes and reveals inter-complex contacts. (**E**) Input (I1–I3) and output (O1–O4) tables required and generated by the compleXView 'Analysis' module (top panel) and example layouts of the input files. Grey arrows indicate optional files.

teractors with ratios <2% were not included. As a result we obtained a 'Bait–Prey Interactions Table' listing the putative bait–prey interactions with their respective abundance ratios.

The bait–prey interaction tables were used as input to infer prey–prey interactions. Pairwise cosine correlations were calculated using the prey-to-bait abundance ratios across different bait purifications. Hence, this mathematical term is referred to as abundance correlation. GO similarities were calculated using the getGeneSim function from the GOSim Bioconductor package (20) with the following parameters: similarity method, 'dot'; normalization method, 'sqrt'; and similarity term, 'relevance'. UniProt accession numbers were mapped to Entrez IDs using the UniProt 'Retrieve/ID mapping' tool (21) and only 'Biological Process' and 'Molecular Function' categories were used. Their values were summarized by keeping the maximum of the two per protein–protein pair. Abundance correlations were combined with GO correlations by calculating the average of their values. Minimum thresholds of 0.8, 0.6 and 0.65 were allowed for abundance, GO and combined correlations, respectively. Proteins were clustered using the MCL algorithm (8) on either the abundance correlations, GO correlations or the combination of the two. Protein interactions were considered as true, if (i) any of the proteins was a bait and their correlation was above the respective threshold or (ii) both proteins were preys in the same MCL cluster with at least one showing a relative ratio to the bait >2%, and their correlation value above the respective threshold or (iii) at least one protein–protein contact was detected by XL–MS. The results are summarized in three different tables with interactions based on either abundance correlations, GO correlations or the combination of both correlations. These tables are annotated with the respective number of protein–protein contacts detected by XL–MS.

Result tables from the crosslink experiments were directly retrieved from the PRIDE database. Intra-protein crosslinks were filtered from the list whereas inter-protein crosslinks were summarized to number of crosslinks per protein–protein pair.

compleXView Analysis Module

compleXView offers two different modules, which operate independent of each other. One module is for the analysis of AP–MS data and performs part of the analy-

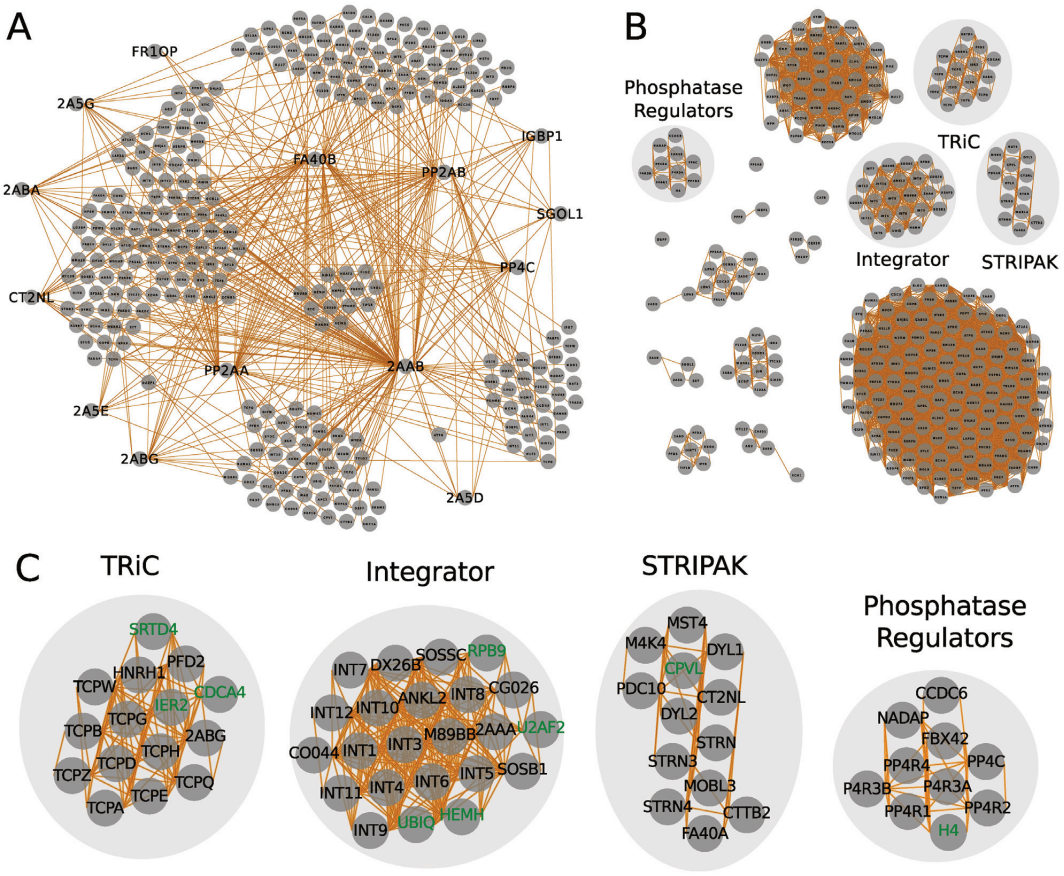


Figure 2. PP2A complexes inferred from bait–prey interactions and abundance correlations. (**A**) bait–prey interactions of the PP2A network. Minimum relative abundance to the bait is 0.02 and the minimum enrichment over the negative control is 2.0. Proteins were grouped by a force-layout algorithm using relative abundances as measure for interaction strength and their inverse values as node-node initial distances. (**B**) PP2A complexes detected based on abundance correlations between preys. Correlation values >0.8 were considered as interactions. Proteins were clustered using the MCL algorithm, arranged by a force-layout algorithm using correlation values as interaction strength and the inverse values for node-node initial distances. (**C**) Zoom-in on complexes indicated in (B). Core subunits and interactors are depicted in black. Putative spurious interactions are shown in green.

sis workflow described in the previous section using protein abundances (Figure 1). Thus, the main input file for the 'Analysis' module is the 'Purifications Table' containing the protein abundances across all purifications. Its first column must be named 'Prey' and contains the protein IDs of the co-purified proteins. The second and all other columns must contain the abundances of the preys in each of the purification experiments. These columns have to be named according to the following format: BaitID_ReplicateNumber_Condition. The name in the 'BaitID' field must match the format of the entries in the 'Prey' column and the bait itself has to be detected in the respective purification. Negative controls must be named 'NegCtr' in this field. The 'ReplicateNumber' field contains any number or code for the identification of technical or biological replicates (e.g. R1, R2, R3). The 'Condition' field is optional and should be provided in cases where purifications of the same bait under different biological conditions are compared.

compleXView requires abundance values like iBAQ or other normalized intensities without log-transformation. Median or quantile normalization between conditions is optional. The basic output of the 'Analysis' module is the 'Bait–Prey Interactions Table' visualized as a spoke network. Abundance correlations will only be computed if the number of baits or conditions is >4. The output is a protein–protein interaction table that we call the 'Abundance Correlations Table'.

In order to compute GO functional similarities between proteins an optional input table with two columns must be provided. The first column named 'From' contains the Protein IDs in the same format as in the 'Prey' column of the 'Purifications Table'. The second column named 'To' contains the respective UniProt Entrez ID of the protein. The *compleXView* output is a protein–protein interaction table called 'GO Correlations Table', where each row contains a pair of preys and their corresponding GO similarity values.

For the implementation of inter-protein crosslinks an input table of at least four columns with the following headings is required: 'Protein1', 'Protein2', 'AbsPos1' and 'AbsPos2'. The IDs in the first two columns should have the same format as the 'Prey' column in the 'Purifications Table'. The numbers in the 'AbsPos' columns indicate the positions of the crosslinked amino acid residues.

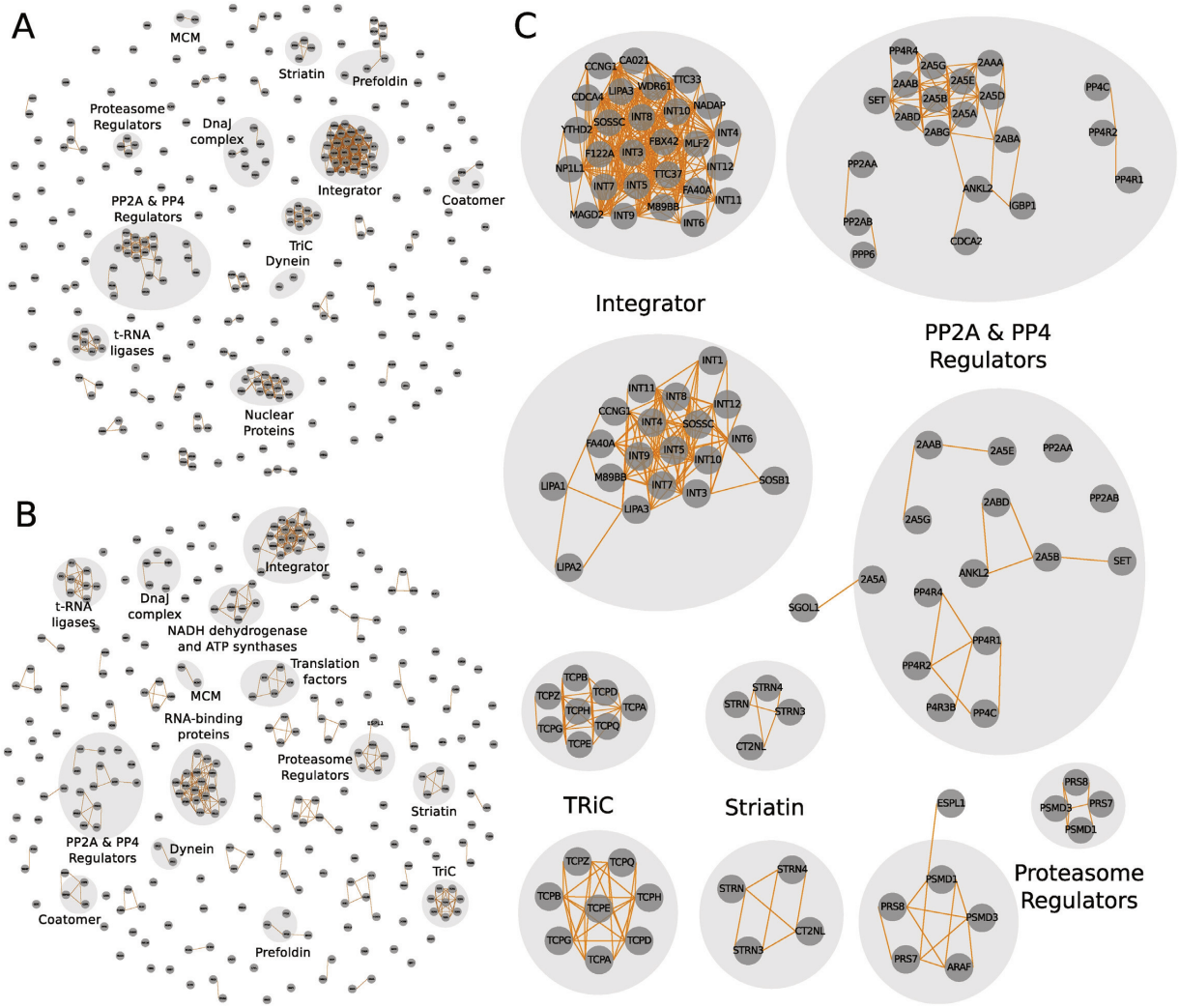


Figure 3. PP2A complexes predicted based on GO functional similarities alone and in combination with abundance correlations. (**A**) PP2A complexes inferred from GO similarities. Similarity values >0.6 were considered as interactions. Proteins were cluster using the MCL algorithm and arranged by a force-layout algorithm as described in (2A). (**B**) PP2A network analysis by applying abundance correlations combined with GO functional similarities between preys. Combined values >0.65 were considered as interactions. Proteins were clustered using the MCL algorithm and arranged by force-layout algorithm using combined values as interaction strength and the inverse values for node-node initial distances. (**C**) Zoom-in on complexes detected in (A) and (B).

The interactions in the output tables can be filtered according to different parameters like fold-change and p-value thresholds (see online Manual).

compleXView Visualization Module

The 'Visualization' module displays all bait–prey interaction tables and correlation-based tables generated by the 'Analysis' module (Figure 1E). Both modules operate independently which facilitates visualization of output tables generated by other programs, such as SAINT (3), MiST (4) or compPASS (5). The input table must contain two columns named 'Bait' and 'Prey' and optional columns to represent quantitative information.

The 'Visualization' module generates two types of representations the 'Network' and 'Blot' plots. The former represents proteins as circular nodes and linear edges indicate their interactions which are deduced from AP–MS abundances or indicated by XL–MS restraints. The 'Blot' plot is designed as western blot diagram displaying protein abundances across different bait purifications. 'Blot' plots are generated by selecting the respective nodes in the network and their quantitative interaction values determine the band intensities.

RESULTS AND DISCUSSION

Workflow

compleXView comprises two independent modules: an 'Analysis' module and a 'Visualization' module. The workflow of the 'Analysis' module is schematically represented in Figure 1. *compleXView* exploits the quantitative informa-

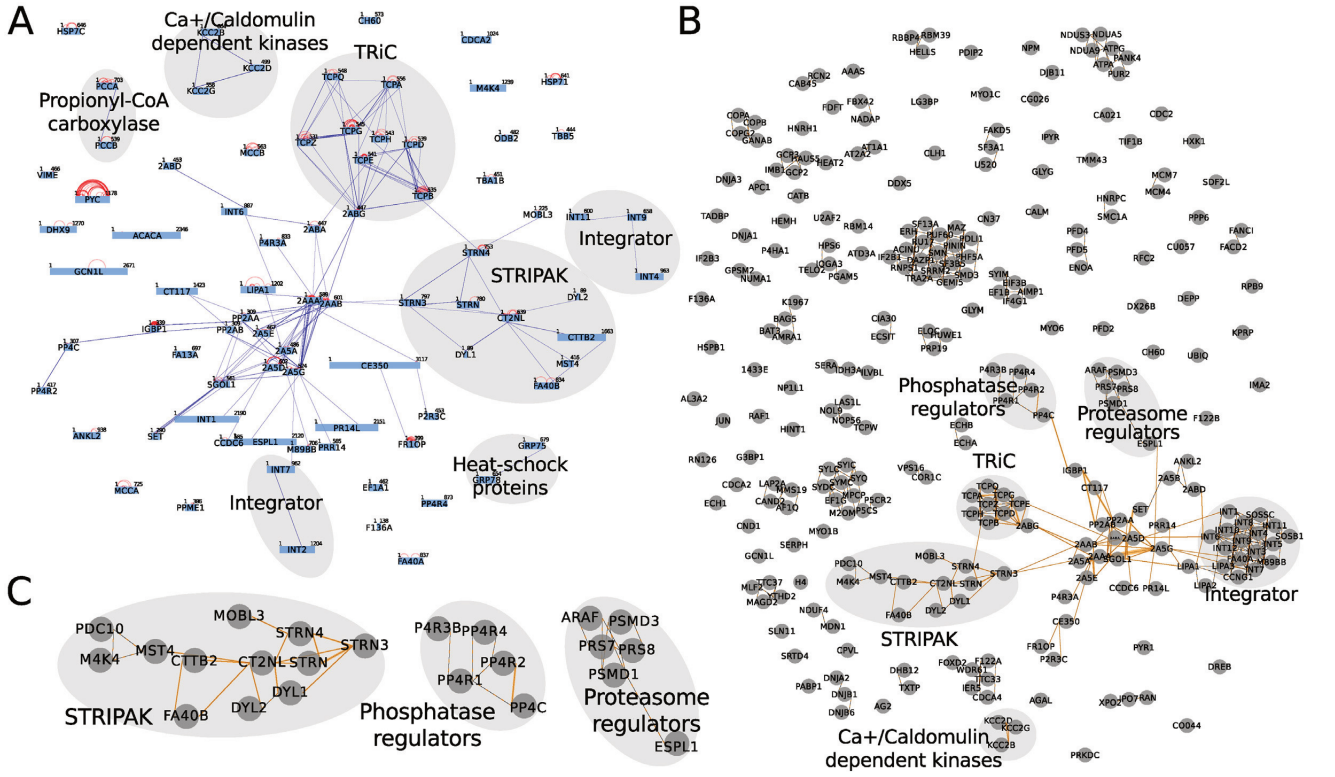


Figure 4. *compleXView* analysis and visualization of the PP2A network based on crosslink-derived protein connectivity in combination with abundance and GO correlations. (**A**) Protein complexes in a PP2A network identified by inter-protein crosslinks. (**B**) Network of PP2A complexes based on the combination of abundance correlations, GO functional similarities and crosslinks. Crosslink-derived restraints validate interactions within predicted complexes, reveal inter-complex contacts and provide insights into the complex topology. Heat shock proteins and propionyl-CoA carboxylases detected in (A) did not pass the threshold values applied in (B). (**C**) Zoom-in on predicted clusters in (B). Interactions predicted by abundance correlations are indicated as dotted lines and interactions identified by crosslinks are depicted as solid lines.

tion of multiple AP–MS experiments as well as GO functional annotations to infer protein complexes in protein interaction studies. Furthermore, *compleXView* implements XL–MS data to establish direct connectivity within or between the predicted complexes. The input data introduced as 'Purifications Table' is used by the 'Analysis' module to determine whether a detected protein is significantly enriched over the negative control and thus, considered as true interactor. Furthermore, only interactors whose relative abundances to the bait are greater than a specified threshold are considered (Figure 1A). The output is a table which serves as input file for the 'Visualization' module.

The 'Bait–Prey Interactions Map' derived from the quantitative AP–MS analysis of a limited number of baits do not provide enough protein connectivity to infer complexes in the network. *compleXView* overcomes this limitation by inferring the relation between preys based on calculating the correlation of their abundances profiles across different bait preparations. Accordingly, *compleXView* moves from a Spoke model of bait–prey interactions to a Matrix model of prey–prey interactions where correlations of abundances between all proteins are calculated. Abundance correlations are computed using the cosine similarity formula as schematically shown in Figure 1B. Although, abundance correlations may be capable of clustering the whole network into submodules and protein complexes, interactions between unrelated proteins may remain. To eliminate these incidents, *compleXView* retrieves GO functional terms and computes the similarity of the GO trees for every pair of proteins (Figure 1C). GO similarities are combined with the abundance correlations in order to obtain a network with higher resolution in terms of protein complex identification. Putative false interactions due to coincidentally occurring high correlations are resolved by accounting GO functional similarities. Low similarity values penalize correlations and only highly correlated or highly functionally similar protein–protein pairs remain.

The integration of direct protein connectivity information from XL–MS experiments with correlated protein abundances advances the approach, aids in inferring protein complex composition and provides additional topological information (Figure 1D). To integrate inter-protein crosslinks into correlation-based protein networks, the user has to provide a table listing the crosslinked amino acid positions between protein pairs. As demonstrated for the test datasets, XL–MS data confirms interactions within complexes and indicates contacts between them (Figures 1D and 4C).

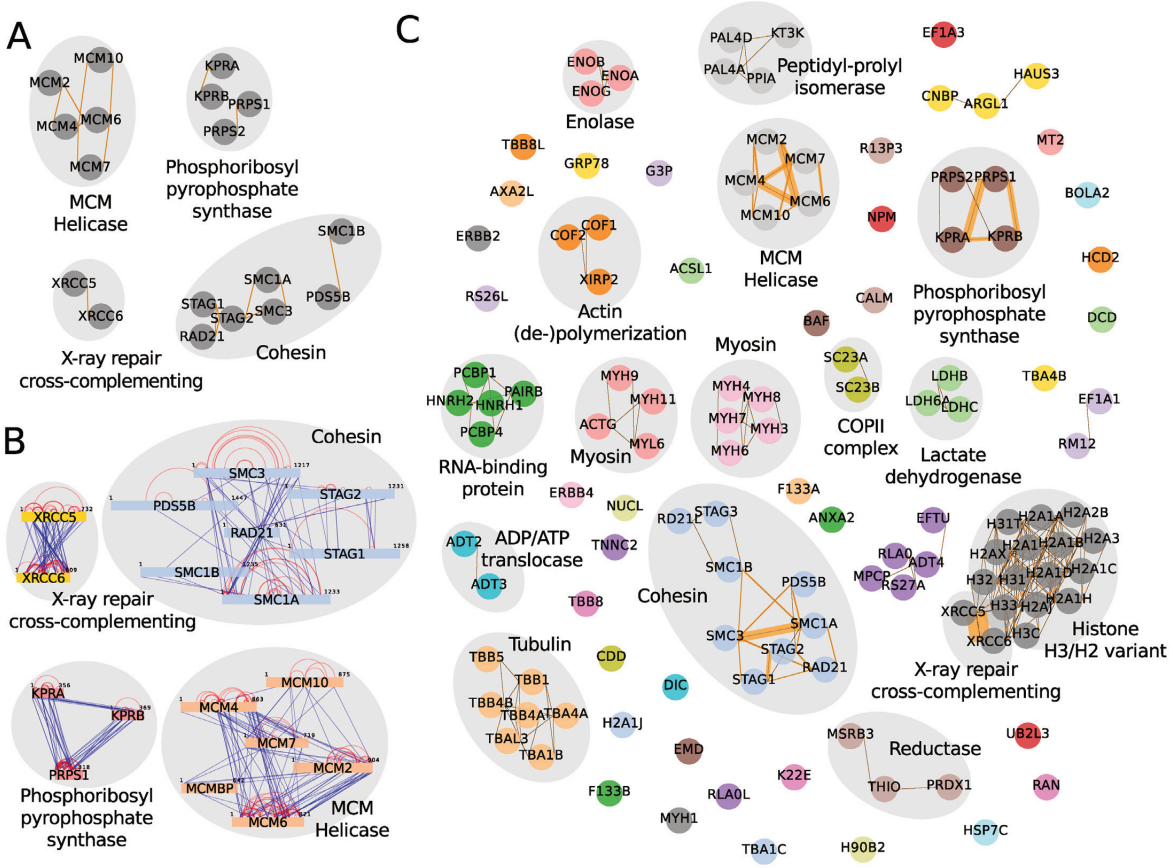


Figure 5. *compleXView* analysis and visualization of chromatin-associated complexes (16) applying abundance correlations combined with GO functional similarities and inter-protein crosslinks. (**A**) Zoom-in on the network solely based on GO similarities depicting only bait complexes. Co-purifying complexes are shown in (C). (**B**) Inter-protein crosslink network. (**C**) Network of protein complexes detected by the combination of abundance correlations, GO functional similarities and inter-protein crosslinks.

Analysis of AP–MS / XL–MS Datasets

We tested *compleXView* on two different datasets which comprise AP–MS analyses and their respective XL–MS experiments (see Materials and Methods).

The first dataset of a PP2A network was obtained from purifications of PP2A core subunits, adapter and substrate proteins (Figure 2). The 'Bait–Prey Interactions Map' derived from data of the 'Purifications Table' depicts the co-purifying proteins of 14 different baits (Figure 2A). To reveal protein complexes in the network, computing abundance correlations between preys resulted in a network with a higher degree of connectivity. Clustering the proteins by a force-layout algorithm which applies the correlation values as measures of interaction strength is able to infer submodules and protein complexes in the network (Figure 2B). In particular, TRiC (TCP-1 ring complex), the Integrator and the STRIPAK complexes are discerned (Figure 2C) from co-purifying proteins. Remaining proteins are associated in large groups due to high random co-variation. Further clustering of proteins based on their GO functional similarities results in higher resolution of the indicated protein complexes in the network (Figure 3A) and reveals additional clusters and interactions (Figure 3B, C). Furthermore, *compleXView* facilitates the interactive manual inspection of putative interactions and protein clusters by providing links to the UniProt database.

The TRiC complex is revealed subsequent to clustering the proteins based on their abundance correlations (Figure 2C). Correlation values >0.9 are calculated between core components of the complex: TCPA, TCPB, TCPD, TCPE, TCPG, TCPH, TCPQ TCPW and TCPZ. In addition, known interactors of the TRiC core complex are identified: the heterogeneous nuclear ribonucleoprotein H (HNRH1), prefoldin subunit 2 (PFD2) and the PP2A regulatory subunit 2ABG. These interactions are annotated in the BioGRID and Intact databases. The associated proteins, SRTD4, IER2 and CDCA4, are putative interactors with high correlations to the TRiC complex. Clustering the network solely based on GO similarities only maintains the core subunits of the TRiC complex in the same group (Figure 3C). The functional similarities of HNRH1, PFD2 and 2ABG to TRiC subunits are low and their low correlation values are insufficient to keep them in the combined network.

Similarly, the Integrator complex is delimited upon clustering the proteins based on their abundance correlations. Integrator core subunits form a group with other known

interactors, such as the ankyrin repeat and LEM domain-containing protein 2 (ANKL2), the PP2A regulatory subunit 2AAA, the integrator subunit 6-like (DX26B), the uncharacterized protein CG026, von Willebrand factor A domain-containing protein 9 (CO044), SOSS complex subunits C and B1 and the cell cycle regulator Mat89Bb homolog (Figure 2C). For the cluster members, RPB9, U2AF, UBIQ and HEMH, no previous evidence for their association with the Integrator complex has been reported. Interestingly, the Integrator complex was found to regulate RNA polymerase II activity (22) indicating that RPB9 may be directly associated with the Integrator complex and thus, these interactions have to be further evaluated. Clustering the network based on GO similarities maintains the Integrator core subunits in a group. However, many of the known Integrator interactors mentioned above are eliminated from the cluster. On the other hand, proteins, exclusively implicated by GO similarities in binding the Integrator complex, are possibly false interactors as their high GO similarity scores result from very general 'Molecular Process' terms (Figure 3C). Moreover, they lack previous evidence of interaction with the Integrator in the BioGRID and Intact databases and are removed from the cluster upon combining abundance correlations with GO similarities. The presence of LIPA1 and 2 in the cluster is due to its high correlation with LIPA3 which is based on a general Molecular Function similarity to Integrator subunits (Figure 3C).

Clustering based on abundance correlations also distinguishes the STRIPAK complex comprising kinases and kinase-associated proteins such as MAP4K4, MST4 and PDCD10 and proteins which interact with striatin like dynein light chains (DYL1 and 2), the Mps one binder-like protein (MOBL3) and the Cortactin-binding protein 2 (CTTB2) (Figure 2C). However, applying GO similarities alone or in combination with abundance correlations results in loss of STRIPAK interacting proteins (Figure 3C) Thus, correct clustering based on weaker abundance correlations may be abrogated once combined with GO functional similarities.

Regulatory subunits of protein phosphatase 4 (PP4) are clustered by applying abundance correlations. However, regulatory subunits of PP2A are dispersed in different groups in the network (Figure 2C). Clustering solely based on GO similarities groups all PP2A regulatory subunits into a cluster and leaves some PP4 regulators outside (Figure 3C). In this case, the clustering with abundance correlations and functional similarities splits the PP2A regulators into subgroups and unifies PP4 regulators with its original cluster.

Bait–prey and prey–prey interactions which are abundant in the affinity-purifications are usually sufficiently covered by the XL–MS analysis detecting at least one crosslink per interaction. Hence, the composition and topology of the PP2A core complexes, TRiC and the STRIPAK complex were revealed solely based on inter-protein crosslinks (Figure 4A). Protein interactions below the detection limit of XL–MS were inferred from AP–MS data revealing clusters of phosphatase and proteasome regulators and interactions of MAP4K4 and PDC10 with the STRIPAK complex (Figure 4B). Thus, the integration and visualization of AP–MS and XL–MS data through *compleXView* analysis complements the protein interactions of complexes indicated by crosslink-derived restraints and validates interactions inferred from abundance correlations (Figure 4C).

The second dataset analyzed by *compleXView* is comprised of five bait proteins with four of them assembled in chromatin-associated complexes and one enzyme involved in the nucleotide metabolic pathway (16). Clustering solely based on abundance correlations did not resolve these protein complexes (data not shown). Indeed, clustering by GO similarities alone was sufficient to group many subunits into the respective complexes (Figure 5A). Importantly, only the combination of both, abundance correlations and GO functionalities, associated STAG3 and RD21L to the cohesin complex and PRPS2 to the phosphoribosyl pyrophosphate synthase complex (Figure 5C). Several other complexes with relative abundances $<10\%$ of the bait, which were not detected by XL–MS, were distinguished (Figure 5B and C).

compleXView offers interactive graphical features for the manipulation and interpretation of the interaction maps. In single-bait experiments users can color preys based on their relative abundances and multiple purifications can be directly compared in a 'Blot' plot representation (see online Manual for detailed description).

compleXView aims to provide an analysis tool for biologists to identify and interpret protein complexes in their pull-down studies. In particular, the combination and visualization of quantitative and connectivity data obtained by mass spectrometry complements the standard maps of co-purifying proteins with structural restraints between subunits and modules in the network.

FUNDING

LMU*excellent* Initiative, the Bavarian Research Center of Molecular Biosystems (to F.H.); German Excellence Initiative (Graduate School QBM); German Research Foundation [GRK1721], the European Research Council [MolStruKT StG no. 638218]; Human Frontier Science Program Research Grant [RGP0008/2015]. Funding for open access charge: Bavarian Research Center of Molecular Biosystems.
Conflict of interest statement. None declared.

REFERENCES

1. Leitner,A., Walzthoeni,T., Kahraman,A., Herzog,F., Rinner,O., Beck,M. and Aebersold,R. (2010) Probing native protein structures by chemical cross-linking, mass spectrometry and bioinformatics. *Mol. Cell. Proteomics*, **9**, 1634–1649.
2. Liu,F. and Heck,A.J. (2015) Interrogating the architecture of protein assemblies and protein interaction networks by cross-linking mass spectrometry. *Curr. Opin. Struct. Biol.*, **35**, 100–108.
3. Choi,H., Larsen,B., Lin,Z.-Y., Breitkreutz,A., Mellacheruvu,D., Fermin,D., Qin,Z.S., Tyers,M., Gingras,A.-C. and Nesvizhskii,A.I. (2011) SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat. Methods*, **8**, 70–73.
4. Jäger,S., Cimermancic,P., Gulbahce,N., Johnson,J.R., McGovern,K.E., Clarke,S.C., Shales,M., Mercenne,G., Pache,L., Li,K. *et al.* (2012) Global landscape of HIV-human protein complexes. *Nature*, **481**, 365–370.
5. Sowa,M.E., Bennett,E.J., Gygi,S.P. and Harper,J.W. (2009) Defining the human deubiquitinating enzyme interaction landscape. *Cell*, **138**, 389–403.

6. Bader,G.D. and Hogue,C.W. V (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.*, **20**, 991–997.
7. Gavin,A.-C., Aloy,P., Grandi,P., Krause,R., Boesche,M., Marzioch,M., Rau,C., Jensen,L.J., Bastuck,S., Dümpelfeld,B. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.
8. Hart,G.T., Lee,I. and Marcotte,E.R. (2007) A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, **8**, 236.
9. Krogan,N.J., Cagney,G., Yu,H., Zhong,G., Guo,X., Ignatchenko,A., Li,J., Pu,S., Datta,N., Tikuisis,A.P. *et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643.
10. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
11. Bader,G.D. and Hogue,C.W. V (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.
12. Leung,H.C.M., Xiang,Q., Yiu,S.M. and Chin,F.Y.L. (2009) Predicting protein complexes from PPI data: a core-attachment approach. *J. Comput. Biol.*, **16**, 133–144.
13. Kouhsar,M., Zare-mirakabad,F. and Jamali,Y. (2015) WCOACH: protein complex prediction in weighted PPI networks. *Genes Genet. Syst.*, **90**, 317–324.
14. Grimm,M., Zimniak,T., Kahraman,A. and Herzog,F. (2015) XVis: a web server for the schematic visualization and interpretation of crosslink-derived spatial restraints. *Nucleic Acids Res.*, **43**, W362–W369.
15. Herzog,F., Kahraman,A., Boehringer,D., Mak,R., Bracher,A., Walzthoeni,T., Leitner,A., Beck,M., Hartl,F.-U., Ban,N. *et al.* (2012) Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science*, **337**, 1348–1352.
16. Makowski,M.M., Willems,E., Jansen,P.W.T.C. and Vermeulen,M. (2016) Cross-linking immunoprecipitation-MS (xIP-MS): Topological analysis of chromatin-associated protein complexes using single affinity purification. *Mol. Cell. Proteomics*, **15**, 854–865.
17. Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.
18. R Core Team. (20016) R: a language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna.
19. Mellacheruvu,D., Wright,Z., Couzens,A.L., Lambert,J.-P., St-Denis,N.A., Li,T., Miteva,Y. V, Hauri,S., Sardiu,M.E., Low,T.Y. *et al.* (2013) The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods*, **10**, 730–736.
20. Fröhlich,H., Speer,N., Poustka,A. and Beissbarth,T. (2007) GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics*, **8**, 166.
21. The UniProt Consortium (2014) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
22. Stadelmayer,B., Micas,G., Gamot,A., Martin,P., Malirat,N., Koval,S., Raffel,R., Sobhian,B., Severac,D., Rialle,S. *et al.* (2014) Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. *Nat. Commun.*, **5**, 5531.

Bibliography

[1] Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Rout, M. P., and Sali, A. Determining the architectures of macromolecular assemblies. *Nature 450*, November (2007), 683–694.

[2] Altschul, S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research 25*, 17 (1997), 3389–3402.

[3] Azzarito, V., Long, K., Murphy, N. S., and Wilson, A. J. Inhibition of α-helix-mediated protein-protein interactions using designed molecules. *Nature Chemistry 5*, 3 (2013), 161–173.

[4] Bader, G. D., and Hogue, C. W. V. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol 20*, 10 (2002), 991–997.

[5] Bantscheff, M., Hopf, C., Savitski, M. M., Dittmann, A., Grandi, P., Michon, A. M., Schlegl, J., Abraham, Y., Becher, I., Bergamini, G., Boesche, M., Delling, M., Dümpelfeld, B., Eberhard, D., Huthmacher, C., Mathieson, T., Poeckel, D., Reader, V., Strunk, K., Sweetman, G., Kruse, U., Neubauer, G., Ramsden, N. G., and Drewes, G. Chemoproteomics profiling of HDAC inhibitors reveals selective targeting of HDAC complexes. *Nature Biotechnology 29*, 3 (2011), 255–268.

[6] Barabási, A.-L., and Albert, R. Emergence of Scaling in Random Networks. *Science 286* (1999), 509–513.

[7] Barnidge, D. R., Dratz, E. A., Martin, T., Bonilla, L. E., Moran, L. B., and Lindall, A. Absolute quantification of the G protein-coupled receptor rhodopsin by LC/MS/MS using proteolysis product peptides and synthetic peptide standards. *Analytical Chemistry 75*, 3 (2003), 445–451.

[8] Chan, S. F., Sances, S., Brill, L. M., Okamoto, S.-i., Zaidi, R., McKercher, S. R., Akhtar, M. W., Nakanishi, N., and Lipton, S. A. ATM-Dependent Phosphorylation of MEF2D Promotes Neuronal Survival after DNA Damage. *Journal of Neuroscience 34*, 13 (2014), 4640–4653.

[9] Chen, C., Shi, Z., Zhang, W., Chen, M., He, F., Zhang, Z., Wang, Y., Feng, M., Wang, W., Zhao, Y., Brown, J. H., Jiao, S., and Zhou, Z.

Striatins Contain a Noncanonical Coiled Coil That Binds Protein Phosphatase 2A A Subunit to Form a 2:2 Heterotetrameric Core of Striatin-interacting Phosphatase and Kinase (STRIPAK) Complex. *Journal of Biological Chemistry 289*, 14 (2014), 9651–9661.

[10] Choi, H., Larsen, B., Lin, Z.-Y., Breitkreutz, A., Mellacheruvu, D., Fermin, D., Qin, Z. S., Tyers, M., Gingras, A.-C., and Nesvizhskii, A. I. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nature methods 8*, 1 (2011), 70–3.

[11] Cox, J., and Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology 26*, 12 (2008), 1367–72.

[12] Dayon, L., Hainard, A., Licker, V., Turck, N., Kuhn, K., Hochstrasser, D. F., Burkhard, P. R., and Sanchez, J. C. Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags. *Analytical Chemistry 80*, 8 (2008), 2921–2931.

[13] Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics 21*, 16 (2005), 3433–3434.

[14] Dosztanyi, Z., Meszaros, B., and Simon, I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics 25*, 20 (2009), 2745–2746.

[15] Dubois, M. L., Bastin, C., Lévesque, D., and Boisvert, F. M. Comprehensive Characterization of Minichromosome Maintenance Complex (MCM) Protein Interactions Using Affinity and Proximity Purifications Coupled to Mass Spectrometry. *Journal of Proteome Research 15*, 9 (2016), 2924–2934.

[16] Egloff, S., Zaborowska, J., Laitem, C., Kiss, T., and Murphy, S. Ser7 phosphorylation of the CTD recruits the RPAP2 ser5 phosphatase to snRNA genes. *Molecular Cell 45*, 1 (2012), 111–122.

[17] Enright, A. J., Van Dongen, S., and Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research 30*, 7 (2002), 1575–1584.

[18] Ezzeddine, N., Chen, J., Waltenspiel, B., Burch, B., Albrecht, T., Zhuo, M., Warren, W. D., Marzluff, W. F., and Wagner, E. J. A Subset of Drosophila Integrator Proteins Is Essential for Efficient U7 snRNA and Spliceosomal snRNA 3'-End Formation. *Molecular and Cellular Biology 31*, 2 (2011), 328–341.

[19] Fabre, B., Lambour, T., Bouyssié, D., Menneteau, T., Monsarrat, B., Burlet-Schiltz, O., and Bousquet-Dubouch, M. P. Comparison of label-free quantification methods for the determination of protein complexes subunits stoichiometry. *EuPA Open Proteomics 4* (2014), 82–86.

[20] Fischer, L., Chen, Z. A., and Rappsilber, J. Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. *Journal of Proteomics 88* (2013), 120–128.

[21] Franken, H., Mathieson, T., Childs, D., Sweetman, G. M., Werner, T., Tögel, I., Doce, C., Gade, S., Bantscheff, M., Drewes, G., Reinhard, F. B., Huber, W., and Savitski, M. M. Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nature Protocols 10*, 10 (2015), 1567–1593.

[22] Fröhlich, H., Speer, N., Poustka, A., and Beissbarth, T. GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC bioinformatics 8*, 1 (2007), 166.

[23] Gao, M., Zhou, H., and Skolnick, J. Insights into disease-associated mutations in the human proteome through protein structural analysis. *Structure 23*, 7 (2015), 1362–1369.

[24] Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. Proteome survey reveals modularity of the yeast cell machinery. *Nature 440* (2006), 631–636.

[25] Glatter, T., Wepf, A., Aebersold, R., and Gstaiger, M. An integrated workflow for charting the human interaction proteome: Insights into the PP2A system. *Molecular Systems Biology 5*, 237 (2009).

[26] Guharoy, M., and Chakrabarti, P. Conserved residue clusters at protein-protein interfaces and their use in binding site identification. *BMC Bioinformatics 11*, 1 (2010), 286.

[27] Havugimana, P. C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, A. L., Li, Z., Wang, P. I., Boutz, D. R., Fong, V., Phanse, S., Babu, M., Craig, S. A., Hu, P., Wan, C., Vlasblom, J., Dar, V. U. N., Bezginov, A., Clark, G. W., Wu, G. C., Wodak, S. J., Tillier, E. R. M., Paccanaro, A., Marcotte, E. M., and Emili, A. A census of human soluble protein complexes. *Cell 150*, 5 (2012), 1068–1081.

[28] Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., and Zhou, Y. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports 5*, 1 (2015), 11476.

[29] Herzog, F., Kahraman, a., Boehringer, D., Mak, R., Bracher, a., Walzthoeni, T., Leitner, a., Beck, M., Hartl, F.-U., Ban, N., Malmstrom, L., and Aebersold, R. Structural Probing of a Protein Phosphatase 2A Network by Chemical Cross-Linking and Mass Spectrometry. *Science 337*, 6100 (2012), 1348–1352.

[30] Hewick, R. M., Zhijian, L., and Wang, J. H. Proteomics in Drug Discovery. In *Proteome Characterization and Proteomics*, R. D. Smith and T. D. Veenstra, Eds., vol. 65 of *Advances in Protein Chemistry*. Academic Press, 2003, pp. 309–342.

[31] Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., Dong, R., Guarani, V., Vaites, L. P., Ordureau, A., Rad, R., Erickson, B. K., Wühr, M., Chick, J., Zhai, B., Kolippakkam, D., Mintseris, J., Obar, R. A., Harris, T., Artavanis-Tsakonas, S., Sowa, M. E., De Camilli, P., Paulo, J. A., Harper, J. W., and Gygi, S. P. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell 162*, 2 (2015), 425–440.

[32] Hwang, J., and Pallas, D. C. STRIPAK complexes: Structure, biological function, and involvement in human diseases. *International Journal of Biochemistry and Cell Biology 47*, 1 (2014), 118–148.

[33] Jiang, L., Stanevich, V., Satyshur, K. A., Kong, M., Watkins, G. R., Wadzinski, B. E., Sengupta, R., and Xing, Y. Structural basis of protein phosphatase 2A stable latency. *Nature Communications 4* (2013), 1611–1699.

[34] Jones, S., and Thornton, J. M. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences 93*, 1 (1996), 13–20.

[35] Jubb, H. C., Pandurangan, A. P., Turner, M. A., Ochoa-Montaño, B., Blundell, T. L., and Ascher, D. B. Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Progress in Biophysics and Molecular Biology 128* (2017), 3–13.

[36] Kalisman, N., Adams, C. M., and Levitt, M. Subunit order of eukaryotic TRiC/CCT chaperonin by cross-linking, mass spectrometry, and combinatorial homology modeling. *Proceedings of the National Academy of Sciences 109*, 8 (2012), 2884–2889.

[37] Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. Posterior error probabilities and false discovery rates: Two sides of the same coin. *Journal of Proteome Research 7*, 1 (2008), 40–44.

[38] Kasinath, V., Faini, M., Poepsel, S., Reif, D., Feng, X. A., Stjepanovic, G., Aebersold, R., and Nogales, E. Structures of human PRC2 with its cofactors AEBP2 and JARID2. *Science 944*, February (2018), 1–10.

[39] Kastritis, P. L., and Bonvin, A. M. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface 10* (2013), 20120835.

[40] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research 36* (2007), D202–D205.

[41] Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., and Hermjakob, H. The IntAct molecular interaction database in 2012. *Nucleic Acids Research 40*, D1 (2012), D841–D846.

[42] Kirkwood, K. J., Ahmad, Y., Larance, M., and Lamond, A. I. Characterization of Native Protein Complexes and Protein Isoform Variation Using Size-fractionation-based Quantitative Proteomics. *Molecular & Cellular Proteomics 12*, 12 (2013), 3851–3873.

[43] Koh, G. C., Porras, P., Aranda, B., Hermjakob, H., and Orchard, S. E. Analyzing protein-protein interaction networks. *Journal of Proteome Research 11*, 4 (2012), 2014–2031.

[44] Kristensen, A. R., Gsponer, J., and Foster, L. J. A high-throughput approach for measuring temporal changes in the interactome. *Nature Methods 9*, 9 (2012), 907–909.

[45] Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H. Y., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature 440*, 7084 (2006), 637–643.

[46] Lampert, F., Hornung, P., and Westermann, S. The Dam1 complex confers microtubule plus endtracking activity to the Ndc80 kinetochore complex. *The Journal of Cell Biology 189*, 4 (2010), 641–649.

[47] Lapatas, V., Stefanidakis, M., Jimenez, R. C., Via, A., and Schneider, M. V. Data integration in biological research: an overview. *Journal of Biological Research-Thessaloniki 22*, 1 (2015), 9.

[48] Leal, A. S., Williams, C. R., Royce, D. B., Pioli, P. A., Sporn, M. B., and Liby, K. T. Bromodomain inhibitors, JQ1 and I-BET 762, as potential therapies for pancreatic cancer. *Cancer Letters 394* (2017), 76–87.

[49] Lee, E. J., Kang, Y. C., Park, W.-H., Jeong, J. H., and Pak, Y. K. Negative transcriptional regulation of mitochondrial transcription factor A (TFAM) by nuclear TFAM. *Biochemical and Biophysical Research Communications 450*, 1 (2014), 166–171.

[50] Leitner, A., Faini, M., Stengel, F., and Aebersold, R. Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends in Biochemical Sciences 41*, 1 (2016), 20–32.

[51] Leitner, A., Joachimiak, L. A., Bracher, A., Mönkemeyer, L., Walzthoeni, T., Chen, B., Pechmann, S., Holmes, S., Cong, Y., Ma, B., Ludtke, S., Chiu, W., Hartl, F. U., Aebersold, R., and Frydman, J. The Molecular Architecture of the Eukaryotic Chaperonin TRiC/CCT. *Structure 20*, 5 (2012), 814–825.

[52] Lima, D. B., Melchior, J. T., Morris, J., Barbosa, V. C., Chamot-Rooke, J., Fioramonte, M., Souza, T. A., Fischer, J. S. G., Gozzo, F. C., Carvalho, P. C., and Davidson, W. S. Characterization of homodimer interfaces with cross-linking mass spectrometry and isotopically labeled proteins. *Nature Protocols 13*, 3 (2018), 431–458.

[53] Liu, Q., and Li, J. Protein binding hot spots and the residue-residue pairing preference: a water exclusion perspective. *BMC Bioinformatics 11*, 1 (2010), 244.

[54] Liu, Q., Remmelzwaal, S., Heck, A. J., Akhmanova, A., and Liu, F. Facilitating identification of minimal protein binding domains by cross-linking mass spectrometry. *Scientific Reports 7*, 1 (2017), 1–11.

[55] Mackmull, M., Klaus, B., Heinze, I., Chokkalingam, M., Beyer, A., Russell, R. B., Ori, A., and Beck, M. Landscape of nuclear transport receptor cargospecificity. *Molecular Systems Biology 13*, 12 (2017), 962.

[56] Mädler, S., Seitz, M., Robinson, J., and Zenobi, R. Does Chemical Cross-Linking with NHS Esters Reflect the Chemical Equilibrium of Protein-Protein Noncovalent Interactions in Solution? *Journal of the American Society for Mass Spectrometry 21*, 10 (2010), 1775–1783.

[57] Makowski, M. M., Gräwe, C., Foster, B. M., Nguyen, N. V., Bartke, T., and Vermeulen, M. Global profiling of proteinDNA and proteinnucleosome

binding affinities using quantitative mass spectrometry. *Nature Communications 9*, 1 (2018), 1653.

[58] Makowski, M. M., Willems, E., Jansen, P. W., and Vermeulen, M. Cross-linking immunoprecipitation-MS (xIP-MS): Topological Analysis of Chromatin-associated Protein Complexes Using Single Affinity Purification. *Molecular & Cellular Proteomics 15*, 3 (2016), 854–865.

[59] Malvezzi, F., Litos, G., Schleiffer, A., Heuck, A., Mechtler, K., Clausen, T., and Westermann, S. A structural basis for kinetochore recruitment of the Ndc80 complex via two distinct centromere receptors. *EMBO Journal 32*, 3 (2013), 409–423.

[60] Mateus, A., Määttä, T. A., and Savitski, M. M. Thermal proteome profiling: Unbiased assessment of protein state through heat-induced stability changes. *Proteome Science 15*, 1 (2017), 1–7.

[61] Meyer, K., and Selbach, M. Quantitative affinity purification mass spectrometry: a versatile technology to study proteinprotein interactions. *Frontiers in Genetics 6* (2015), 237.

[62] Miller, S., Janin, J., Lesk, A. M., and Chothia, C. Interior and surface of monomeric proteins. *Journal of Molecular Biology 196*, 3 (1987), 641–656.

[63] MontañoGutierrez, L. F., Ohta, S., Kustatscher, G., Earnshaw, W. C., and Rappsilber, J. Nano Random Forests to mine protein complexes and their relationships in quantitative proteomics data. *Molecular biology of the cell 28*, 5 (2017), 673–680.

[64] Morrison, K. L., and Weiss, G. A. Combinatorial alanine-scanning. *Current Opinion in Chemical Biology 5*, 3 (2001), 302–307.

[65] Mu, R., Wang, Y. B., Wu, M., Yang, Y., Song, W., Li, T., Zhang, W. N., Tan, B., Li, A. L., Wang, N., Xia, Q., Gong, W. L., Wang, C. G., Zhou, T., Guo, N., Sang, Z. H., and Li, H. Y. Depletion of pre-mRNA splicing factor Cdc5L inhibits mitotic progression and triggers mitotic catastrophe. *Cell Death and Disease 5*, 3 (2014), 1–12.

[66] Müller, F., Fischer, L., Chen, Z. A., Auchynnikava, T., and Rappsilber, J. On the Reproducibility of Label-Free Quantitative Cross-Linking/Mass Spectrometry. *Journal of the American Society for Mass Spectrometry 29*, 2 (2018), 405–412.

[67] Murakami, Y., and Mizuguchi, K. Applying the Na??ve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics 26*, 15 (2010), 1841–1848.

[68] Murayama, Y., Samora, C. P., Kurokawa, Y., Iwasaki, H., and Uhlmann, F. Establishment of DNA-DNA Interactions by the Cohesin Ring. *Cell 172*, 3 (2018), 465–477.e15.

[69] Nesvizhskii, A. I. Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments. *Proteomics 12*, 10 (2012), 1639–1655.

[70] Neves, H., and Kwok, H. F. In sickness and in health: The many roles of the minichromosome maintenance proteins. *Biochimica et Biophysica Acta - Reviews on Cancer 1868*, 1 (2017), 295–308.

[71] Oganesyan, I., Lento, C., and Wilson, D. J. Contemporary hydrogen deuterium exchange mass spectrometry. *Methods 144*, April (2018), 27–42.

[72] Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Molecular & Cellular Proteomics 1*, 5 (2002), 376–386.

[73] Pankow, S., Bamberger, C., Calzolari, D., Bamberger, A., and Yates, J. R. Deep interactome profiling of membrane proteins by co-interacting protein identification technology. *Nature Protocols 11*, 12 (2016), 2515–2528.

[74] Reynhout, S., and Janssens, V. Physiologic functions of PP2A: Lessons from genetically modified mice. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 1866*, 1 (2019), 31–50.

[75] Rhee, H.-w., Zou, P., Udeshi, N. D., Martell, J. D., Mootha, V. K., Carr, S. A., and Ting, A. Y. Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science 339*, March (2013), 1328.

[76] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. A generic protein purification method for protein complex characterization and proteome exploration Guillaume. *Nature Biotechnology 17* (1999), 1030–1032.

[77] Rose, G., Geselowitz, A., Lesser, G., Lee, R., and Zehfus, M. Hydrophobicity of amino acid residues in globular proteins. *Science 229*, 4716 (1985), 834–838.

[78] Rossi, A. M., and Taylor, C. W. Analysis of protein-ligand interactions by fluorescence polarization. *Nature Protocols 6*, 3 (2011), 365–387.

[79] Röst, H. L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H. C., Gutenbrunner, P., Kenar, E., Liang, X., Nahnsen, S., Nilse, L., Pfeuffer, J., Rosenberger, G., Rurik, M., Schmitt, U., Veit, J., Walzer, M., Wojnar, D., Wolski, W. E., Schilling, O., Choudhary, J. S., Malmström, L., Aebersold, R.,

Reinert, K., and Kohlbacher, O. OpenMS: A flexible open-source software platform for mass spectrometry data analysis. *Nature Methods 13*, 9 (2016), 741–748.

[80] Roux, K. J., Kim, D. I., Raida, M., and Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *Journal of Cell Biology 196*, 6 (2012), 801–810.

[81] Saha, S., Dazard, J.-e., Xu, H., and Ewing, R. M. Computational Framework for Analysis of PreyPrey Associations in Interaction Proteomics Identifies Novel Human Protein-Protein Interactions and Networks. *Journal of proteome research 11*, 9 (2012), 4476–4487.

[82] Sanulli, S., Justin, N., Teissandier, A., Ancelin, K., Portoso, M., Caron, M., Michaud, A., Lombard, B., da Rocha, S. T., Offer, J., Loew, D., Servant, N., Wassef, M., Burlina, F., Gamblin, S. J., Heard, E., and Margueron, R. Jarid2 Methylation via the PRC2 Complex Regulates H3K27me3 Deposition during Cell Differentiation. *Molecular Cell 57*, 5 (2015), 769–783.

[83] Sardiu, M. E., Cai, Y., Jin, J., Swanson, S. K., Conaway, R. C., Conaway, J. W., Florens, L., and Washburn, M. P. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proceedings of the National Academy of Sciences of the United States of America 105*, 5 (2008), 1454–1459.

[84] Savitski, M. M., Reinhard, F. B., Franken, H., Werner, T., Savitski, M. F., Eberhard, D., Molina, D. M., Jafari, R., Dovega, R. B., Klaeger, S., Kuster, B., Nordlund, P., Bantscheff, M., and Drewes, G. Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science 346*, 6205 (2014).

[85] Schmidt, C., Lenz, C., Grote, M., Lührmann, R., and Urlaub, H. Determination of protein stoichiometry within protein complexes using absolute quantification and multiple reaction monitoring. *Analytical Chemistry 82*, 7 (2010), 2784–2796.

[86] Schmidt, C., and Robinson, C. V. A comparative cross-linking strategy to probe conformational changes in protein complexes. *Nature Protocols 9*, 9 (2014), 2224–2236.

[87] Schmidt, C., and Urlaub, H. Combining cryo-electron microscopy (cryo-EM) and cross-linking mass spectrometry (CX-MS) for structural elucidation of large protein assemblies. *Current Opinion in Structural Biology 46* (2017), 157–168.

[88] Schmitzberger, F., Richter, M. M., Gordiyenko, Y., Robinson, C. V., Dadlez, M., and Westermann, S. Molecular basis for inner kinetochore configuration through RWD domainpeptide interactions. *The EMBO Journal 36*, 23 (2017), e201796636.

[89] Schneider, M., Belsom, A., and Rappsilber, J. Protein Tertiary Structure by Crosslinking/Mass Spectrometry. *Trends in Biochemical Sciences 43*, 3 (2018), 157–169.

[90] Selbach, M., and Mann, M. Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). *Nature Methods 3*, 12 (2006), 981–983.

[91] Sents, W., Ivanova, E., Lambrecht, C., Haesen, D., and Janssens, V. The biogenesis of active protein phosphatase 2A holoenzymes: A tightly regulated process creating phosphatase specificity. *FEBS Journal 280*, 2 (2013), 644–661.

[92] Serra-Pagès, C., Kedersha, N. L., Fazikas, L., Medley, Q., Debant, A., and Streuli, M. The LAR transmembrane protein tyrosine phosphatase and a coiled-coil LAR-interacting protein co-localize at focal adhesions. *The EMBO journal 14*, 12 (1995), 2827–38.

[93] Serra-Pagès, C., Medley, Q. G., Tang, M., Hart, A. C., and Streuli, M. Liprins, a family of LAR transmembrane protein-tyrosine phosphatase-interacting proteins. *Journal of Biological Chemistry 273*, 25 (1998), 15611–15620.

[94] Sharma, K., Weber, C., Bairlein, M., Greff, Z., Kéri, G., Cox, J., Olsen, J. V., and Daub, H. Proteomics strategy for quantitative protein interaction profiling in cell extracts. *Nature Methods 6*, 10 (2009), 741–744.

[95] Shi, Y., Pellarin, R., Fridy, P. C., Fernandez-Martinez, J., Thompson, M. K., Li, Y., Wang, Q. J., Sali, A., Rout, M. P., and Chait, B. T. A strategy for dissecting the architectures of native macromolecular assemblies. *Nature methods 12*, 12 (2015), 1135–1138.

[96] Sinz, A. Cross-Linking/Mass Spectrometry for Studying Protein Structures and ProteinProtein Interactions: Where Are We Now and Where Should We Go from Here? *Angewandte Chemie - International Edition 57*, 22 (2018), 6390–6396.

[97] Smetana, J. H. C., and Zanchin, N. I. T. Interaction analysis of the heterotrimer formed by the phosphatase 2A catalytic subunit, ??4 and the mammalian ortholog of yeast Tip41 (TIPRL). *FEBS Journal 274*, 22 (2007), 5891–5904.

[98] Smits, A. H., Jansen, P. W. T. C., Poser, I., Hyman, A. A., and Vermeulen, M. Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Research 41*, 1 (2013), 1–8.

[99] Smits, A. H., and Vermeulen, M. Characterizing Protein-Protein Interactions Using Mass Spectrometry: Challenges and Opportunities. *Trends in Biotechnology 34*, 10 (2016), 825–834.

[100] Solis-Mezarino, V., and Herzog, F. CompleXView: A server for the interpretation of protein abundance and connectivity information to identify protein complexes. *Nucleic Acids Research 45*, W1 (2017), W276–W284.

[101] Srihari, S., Yong, C., Patil, A., and Wong, L. Methods for protein complex prediction and their contributions towards understanding the organization, function and dynamics of complexes. *FEBS letters 589*, 19 (2015), 2590–2602.

[102] Sun, A., Li, F., Liu, Z., Jiang, Y., Zhang, J., Wu, J., and Shi, Y. Structural and biochemical insights into human zinc finger protein AEBP2 reveals interactions with RBBP4. *Protein & Cell 9*, 8 (2018), 738–742.

[103] Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., and Hamon, C. Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry 75*, 8 (2003), 1895–1904.

[104] Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., and Wilke, C. O. Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS ONE 8*, 11 (2013), e80635.

[105] Tuncbag, N., Kar, G., Keskin, O., Gursoy, A., and Nussinov, R. A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Briefings in Bioinformatics 10*, 3 (2009), 217–232.

[106] Vizcaíno, J. A., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q.-W., Wang, R., and Hermjakob, H. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research 44*, D1 (2016), D447–D456.

[107] Walzthoeni, T., Claassen, M., Leitner, A., Herzog, F., Bohn, S., Förster, F., Beck, M., and Aebersold, R. False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nature Methods 9*, 9 (2012), 901–903.

[108] Walzthoeni, T., Joachimiak, L. A., Rosenberger, G., Röst, H. L., Malmström, L., Leitner, A., Frydman, J., and Aebersold, R. XTract: Software for characterizing conformational changes of protein complexes by quantitative cross-linking mass spectrometry. *Nature Methods 12*, 12 (2015), 1185–1190.

[109] Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., Xiong, X., Kagan, O., Kwan, J., Bezginov, A., Chessman, K., Pal, S., Cromar, G., Papoulas, O., Ni, Z., Boutz, D. R., Stoilova, S., Havugimana, P. C., Guo, X., Malty, R. H., Sarov, M., Greenblatt, J., Babu, M., Derry, W. B., Tillier, E. R., Wallingford, J. B., Parkinson, J., Marcotte, E. M., and Emili, A. Panorama of ancient metazoan macromolecular complexes. *Nature 525* (2015), 339–344.

[110] Watson, G. M., Lucas, W. A. H., Gunzburg, M. J., and Wilce, J. A. Insight into the Selectivity of the G7-18NATE Inhibitor Peptide for the Grb7-SH2 Domain Target. *Frontiers in Molecular Biosciences 4*, September (2017), 1–8.

[111] Watts, D. J., and Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature 393* (1998), 440–442.

[112] Wu, C. G., Zheng, A., Jiang, L., Rowse, M., Stanevich, V., Chen, H., Li, Y., Satyshur, K. A., Johnson, B., Gu, T. J., Liu, Z., and Xing, Y. Methylation-regulated decommissioning of multimeric PP2A complexes. *Nature Communications 8*, 1 (2017), 1–13.

[113] Xiao, H., Wang, F., Wisniewski, J., Shaytan, A. K., Ghirlando, R., Fitzgerald, P. C., Huang, Y., Wei, D., Li, S., Landsman, D., Panchenko, A. R., and Wu, C. Molecular basis of CENP-C association with the CENP-A nucleosome at yeast centromeres. *Genes and Development 31*, 19 (2017), 1958–1972.

[114] Xue, L. C., Dobbs, D., Bonvin, A. M. J. J., and Honavar, V. Computational prediction of protein interfaces: A review of data driven methods. *FEBS Letters 589*, 23 (2015), 3516–3526.

[115] Xue, L. C., Dobbs, D., and Honavar, V. HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC Bioinformatics 12*, 1 (2011), 244.

[116] Yan, C., Dobbs, D., and Honavar, V. A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics 20*, SUPPL. 1 (2004), 371–378.

[117] Yan, C., Wu, F., Jernigan, R. L., Dobbs, D., and Honavar, V. Characterization of protein-protein interfaces. *Protein Journal 27*, 1 (2008), 59–70.

[118] Zhai, Y., Li, N., Jiang, H., Huang, X., Gao, N., and Tye, B. K. Unique Roles of the Non-identical MCM Subunits in DNA Replication Licensing. *Molecular Cell 67*, 2 (2017), 168–179.

Acknowledgements

Along this doctoral work, a number of people supported me with ideas and the provision of their data. I want to thank in particular to four scientists in my group. First of all, to Franz Herzog for his mentoring and supervision in this endeavor. Second, to Goetz Hagemann for his huge effort on the acquisition of the data used in the last chapter of my thesis. And to Mia Potocnjak and Josef Fischboeck, whose cross-linking data and their own projects motivated the idea presented in the third chapter of my thesis.

Finally, I want to deeply thank my family and close friends for their emotional support along these eight years ever since I moved to Europe for my Master's and Doctoral studies. I was blessed with two parents that always put high relevance in education. And thus, I also want to thank you, Julia Mezarino and Hugo Solis, for all the effort you put on supporting my career decisions in each and every single aspect that involved it.

Thank you very much, everyone!