Cogentech - Consortium for genomic technologies

Doctoral school in COMPLEX SYSTEMS IN MEDICINE AND LIFE SCIENCES
Ph.D. program in COMPLEXITY IN POST-GENOMIC BIOLOGY
XXIII Cycle

Analysis of protein complexes by cross-linking, mass spectrometry and computational tools

A work submitted by DAVIDE CITTARO

Internal tutor:	PROF. MICHELE CASELLE
External tutors:	DR. ANDREA MUSACCHIO
	DR. ANGELA BACHI
Coordinator:	PROF. RAFFAELE CALOGERO

Academic years: 2008-2010

SSD: BIO/18

Table of Contents

List of abbreviations

AI: Atomic Identification
AP: Affinity Purification
CID: Collision Induced Dissociation
CXMS: Chemical Cross-linking coupled with Mass Spectrometry
ESI: Electrospray Ionization
FP: False Positive
FT-ICR: Fourier-Transform Ion Cyclotron
LC: Liquid Chromatography
MALDI: Matrix-Assisted Laser Desorption/Ionization
MMS: Median Match Score
MS: Mass Spectrometry
MS/MS: Tandem Mass Spectrometry
NPV: Negative Predictive Value
OMSSA: Open Mass Spectrometry Search Algorithm
PMF: Peptide Mass Fingerprinting
PPI: Protein-Protein Interaction
PPV: Positive Predictive Value
R1: Results allowing for best matches only
RAM: Random Access Memory
ROC: Receiver Operating Characteristic
TAP: Tandem Affinity Purification
TEV: Tobacco Etch Virus
TOF: Time of Flight
TP: True Positive
Y2H: Yeast two-Hybrid system

Figure Index

Introduction

Proteins are the main actors within the cell, carrying out most of the duties specified by the information encoded into genes. Proteins, however, rarely act alone. Most of the times they interact in higher-order structures and show complex dynamic connections. Mapping protein-to-protein physical interactions is a critical step towards unwinding the complex molecular relationships in living systems. The complete map of protein interactions is called interactome. Interactome mapping has become one of the main scopes of current biological research (Cusick, *et al.* 2005).

Large-scale technologies, that measure the physical connections of proteins at the proteome scale are essential to achieve a comprehensive portrait of the interactome itself. Advances in development of high-throughput technologies resulted in a tremendous increase of reported protein-protein interactions (PPIs). Collections of PPIs produce the desired “omic” scale views of protein partners and protein memberships in complexes and assemblies (Blow 2009). Also, smaller scale experiments continuously describe a number of interactions. These experimental efforts are being organized in a great variety of digital databases, a glimpse of which can be found at http://www.pathguide.org/.

PPIs can be defined both from a topological and from a functional point of view. The issue of whether two proteins share a “functional contact” is quite distinct from the question of whether the same two proteins interact directly with each other. The evidence of a functional interaction does not imply physical contact nor does the opposite. Therefore, definition of PPI has to consider that the contact should be intentional, *i.e.* the result of specific biomolecular forces/ events; as well as the contact could also be non-generic and it should have evolved for a specific purpose (De Las Rivas and Fontanillo 2010).

Given that the cell’s environment undergoes continuous rearrangement, PPIs are not necessarily static or permanent. Also, not all possible interactions will occur in any cell at any time. Instead, interactions depend on cell type, cell cycle phase and state, developmental stage, environmental conditions, protein modifications, presence of cofactors, and presence of other binding partners.

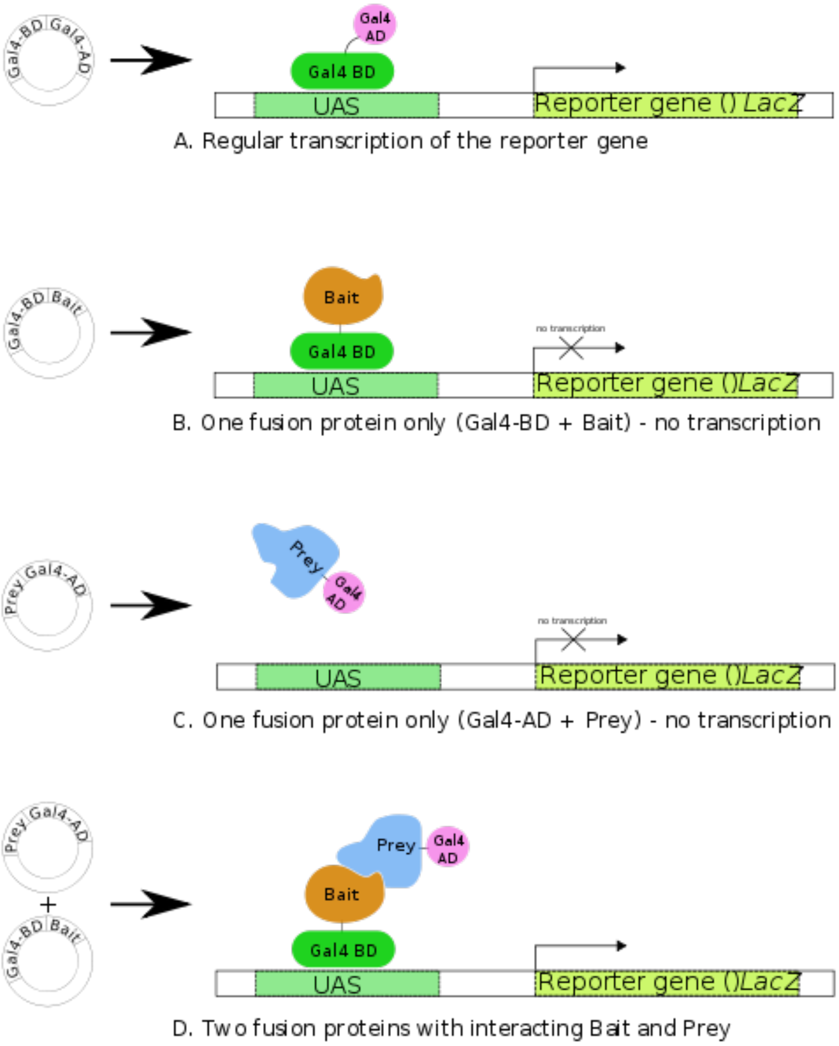
Methodologies to study PPIs

Two main approaches have been introduced to study PPIs: the “binary” approach and the “co-complex” approach (Yu, *et al.* 2008). The binary approach measures the direct interaction between protein pairs; it is mainly represented by the yeast two-hybrid system (Y2H) (Fields and Song 1989). The co-complex approach is more comprehensive and tries to identify physical partners within a whole protein complex. Affinity purification coupled with mass spectrometry (AP-MS) is the most often used methodology with this approach.

The Y2H system relies on the activation of downstream reporter gene(s) by the binding of a transcription factor onto an upstream activating sequence (UAS). The transcription factor is split into two separate fragments: the binding domain (BD), which binds the UAS, is fused with one of the examined proteins (the bait protein); and an activating domain (AD), which is responsible for the activation of transcription, is fused with a single protein (the prey protein) or a library of unknown proteins. If bait and prey proteins interact, they restore the transcription factor functionality, allowing for the expression of the reporter gene (Figure 1).

As only binary interactions can be spotted, it may be difficult to describe a big complex with this method. Also, Y2H system relies on the assumption that two proteins interact in an assay as they interacted in vivo. Unfortunately this is not always the case: false negatives may for instance arise when the interaction is a consequence of a specific post-translational modification (*i.e.* phosphorylation); likewise, false positives may happen since two proteins may interact in Y2H system just by chance, while in the original cellular environment are constitutively separated (in space or in time).

In AP-MS experiments, a protein is fused with a protein tag, or two sequential affinity tags spaced by a cleavage site of tobacco etch virus (TEV) protease (TAP, Tandem affinity purification) (Rigaut, *et al.* 1999). The TEV protease cleaves a sequence (Glu-X-X-Tyr-X-Gln/Ser), which is very uncommon in mammalian proteins. The use of TEV protease therefore minimizes the risk of cleaving inside the bait protein and/or associated proteins. First, TAP-tagged proteins are retained on an affinity column thanks to the first part of the TAP-tag. After rinsing, the TAP is cleaved with the TEV protease, exposing the second affinity tag; this is then exploited to bind a second column (*e.g.*, calmodulin-coated beads) (Figure 2).



A. Regular transcription of the reporter gene

B. One fusion protein only (Gal4-BD + Bait) - no transcription

C. One fusion protein only (Gal4-AD + Prey) - no transcription

D. Two fusion proteins with interacting Bait and Prey

Figure 1 *Overview of two-hybrid assay. The assay is utilized to identify interactions between two proteins, called here Bait and Prey. [A] Gal4 transcription factor gene produces two domain protein (BD and AD) which is essential for transcription of the reporter gene (LacZ). [B,C] Two fusion proteins are prepared: Gal4BD+Bait and Gal4AD+Prey. None of them is usually sufficient to initiate the transcription of the reporter gene alone. [D] When both fusion proteins are produced and Bait part of the first interact with Prey part of the second, transcription of the reporter gene occurs (Wikipedia 2007)*

Although this method allows the identification of more than two proteins involved in a complex, it cannot identify the protein-protein interactions within it; in fact, once the complex is eluted from the second column, any topology information is lost.

Surprisingly enough, the number of interactions confirmed by more than one method is low; a 40-80% estimate of false negatives and 30-60% estimate of false positives have been assigned to Y2H system and affinity purification methods (Aloy and Russell 2004).

In addition, TAP-MS cannot identify transient interactions, as they are usually too weak to be retained during purification. This last issue can be resolved using a chemical cross-linker to freeze PPIs (Guerrero, *et al.* 2006). Cross-linked proteins, however, are not ready to be analyzed by mass spectrometry, as the cross-linker must be usually reversed. In recent years, a number of methodologies have been introduced to identify proteins after chemical fixation of a complex without reversing the chemical cross-link (Schilling, *et al.* 2003; Koning, *et al.* 2006; Gao, *et al.* 2006; Maiolica, *et al.* 2007; Rinner, *et al.* 2008; Panchaud, *et al.* 2010). The possibility to identify cross-linked proteins in mass spectrometry experiments opens a new scenario in which complex topologies can be analyzed in a high- or mid-throughput manner. The methodologies described in this work fall into these last approaches.

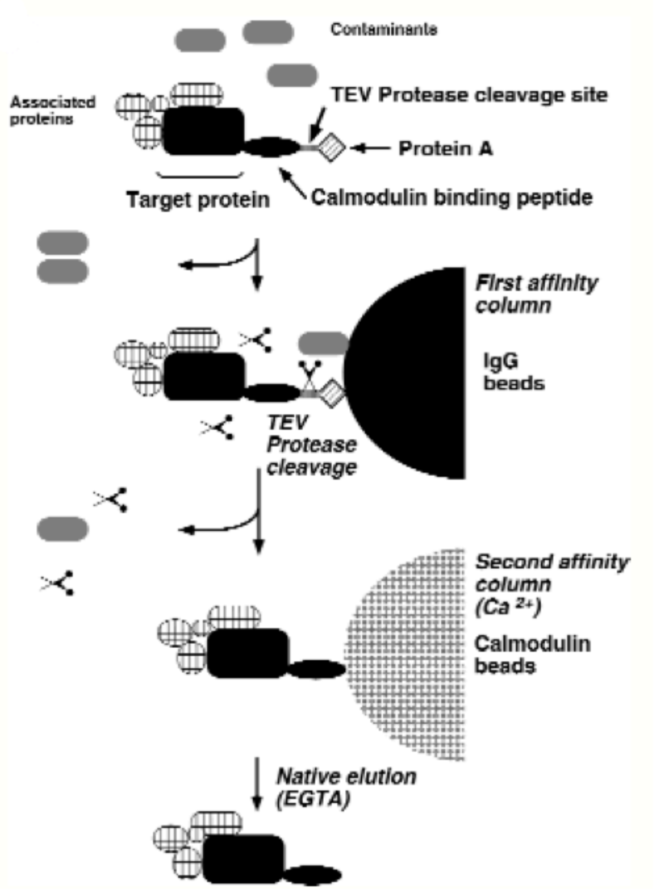


Figure 2 *Overview of the TAP procedure (Rigaut, et al. 1999)*

Mass Spectrometry based proteomics

Mass spectrometric measurements are carried out in the gas phase on ionized analytes. By definition, a mass spectrometer consists of an ion source, a mass analyzer that measures the mass-to-charge ratio (*m/z*) of the ionized analytes, and a detector that registers the number of ions at each *m/z* value (Aebersold and Mann 2003). Electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) are the two techniques most commonly used to volatize and ionize proteins or peptides for mass spectrometric analysis. ESI ionizes the analytes out of a solution and is coupled to liquid-based (*e.g.* chromatographic and electrophoretic) separation tools. MALDI sublimates and ionizes the samples out of a dry, crystalline matrix via laser pulses. MALDI-MS is normally used to analyze relatively simple peptide mixtures, whereas integrated liquid-chromatography ESI-MS systems (LC-MS) are preferred for the analysis of complex samples.

The mass analyzer is central to the technology. There are two broad categories of mass analyzers: the scanning and ion-beam mass spectrometers, such as time-of-flight (TOF) and quadrupole; and the trapping mass spectrometers, such as ion trap (IT), Orbitrap, and Fourier-transform ion cyclotron (FT-ICR) (Yates, *et al.* 1995). The scanning mass analyzers like TOF are usually interfaced with MALDI to perform pulsed analysis, whereas the ion-beam and trapping instruments are frequently coupled to a continuous ESI source.

A typical MS-based proteomic experiment is exemplified in Figure 3. In the first step, the proteins to be analyzed are isolated from cell lysate or tissues by biochemical fractionation or affinity selection. MS of whole proteins is less sensitive than peptide MS and the mass of the intact protein by itself is insufficient for identification. Therefore, proteins are degraded using enzymes, usually trypsin, leading to peptides with C-terminally protonated amino acids, providing an advantage in subsequent peptide sequencing. In the third step, the peptides are separated by high-pressure liquid chromatography (HPLC) in very fine capillaries and eluted into an electrospray ion source where they are nebulized in small, highly charged droplets. After evaporation, multiply protonated peptides enter the mass spectrometer and a mass spectrum of the peptides eluting at this time point is taken (MS^1 spectrum). The computer generates a prioritized list of these peptides for fragmentation and a series of tandem mass spectrometric or 'MS/MS' experiments. These consist of isolation of a given peptide ion, fragmentation by energetic collision with gas, and recording of collision induced dissociation (CID) spectrum (MS/MS spectrum).

Each MS¹ and MS/MS spectrum is typically acquired for about one second or less, depending on the instrument duty cycle, and stored for matching against protein sequence databases. The outcome of the experiment is the identity of the peptides and therefore the proteins making up the purified protein population.

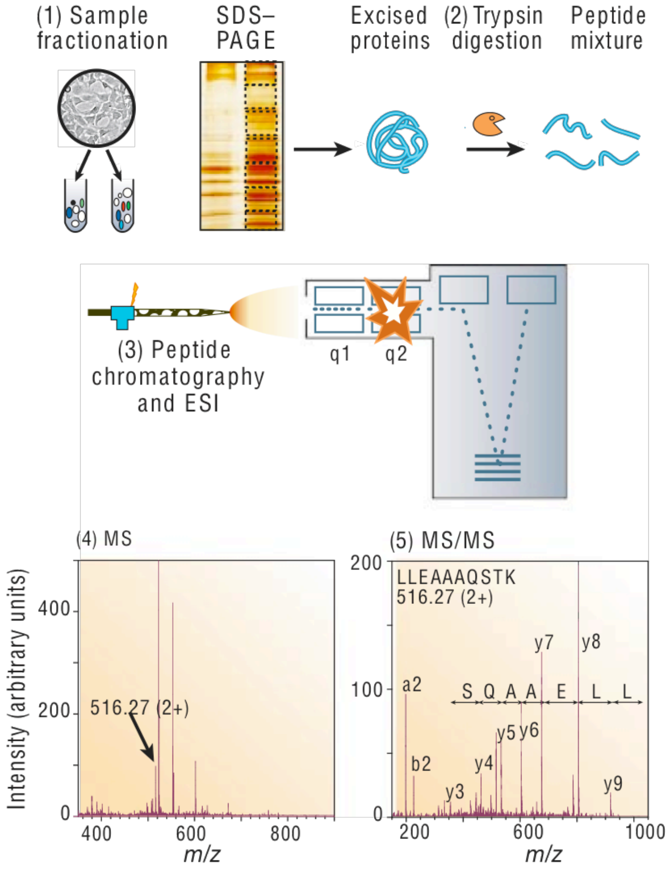


***Figure 3** The typical proteomics experiment consists of five stages. In stage 1, the proteins to be analysed are isolated from cell lysate or tissues by biochemical fractionation or affinity selection. This often includes a final step of one-dimensional gel electrophoresis, and defines the 'sub-proteome' to be analysed. MS of whole proteins is less sensitive than peptide MS and the mass of the intact protein by itself is insufficient for identification. Therefore, proteins are degraded enzymatically to peptides in stage 2, usually by trypsin, leading to peptides with C-terminally protonated amino acids, providing an advantage in subsequent peptide sequencing. In stage 3, the peptides are separated by one or more steps of high-pressure liquid chromatography in very fine capillaries and eluted into an electrospray ion source where they are nebulized in small, highly charged droplets. After evaporation, multiply protonated peptides enter the mass spectrometer and, in stage 4, a mass spectrum of the peptides eluting at this time point is taken (MS1 spectrum, or 'normal mass spectrum'). The computer generates a prioritized list of these peptides for fragmentation and a series of tandem mass spectrometric or 'MS/MS' experiments ensues (stage 5). These consist of isolation of a given peptide ion, fragmentation by energetic collision with gas, and recording of the tandem or MS/MS spectrum. The MS and MS/MS spectra are typically acquired for about one second each and stored for matching against protein sequence databases. The outcome of the experiment is the identity of the peptides and therefore the proteins making up the purified protein population (Aebersold and Mann 2003)*

Although protein identification may be possible using only MS[1] spectra (peptide mass fingerprinting, PMF), the methods described in this work rely on tandem mass spectrometry; hence we will only focus on the analysis of MS/MS spectra.

Tandem mass spectra generated by the fragmentation of peptide ions in the gas phase at low collision energy are dominated by fragment ions resulting from cleavage at the amide bonds (Aebersold and Goodlett 2001). The nomenclature differentiates fragment ions according to where the amide bond breaks and the end of the peptide that retains a charge after fragmentation (Figure 4). If the positive charge associated with the parent peptide ion remains on the amino-terminal side of the fragmented amide bond, then this fragment ion is referred to as a *b* ion. However, the fragment ion is referred to as a *y* ion if the charge remains on the carboxyl-terminal side of the broken amide bond. Since in principle every peptide bond can fragment to generate a *b* or *y* ion, respectively, subscripts are used to designate the specific amide bond that was fragmented to generate the observed fragment ions. *b* ions are designated by a subscript that reflects the number of amino acid residues present on the fragment ion counted from the amino-terminus, whereas the subscript of *y* ions indicates the number of amino acids present, counting from the carboxyl-terminus.

Individual fragment ion *m/z* values can be easily calculated from the amino acid sequence. To calculate the masses of the *b* ion series, 1 u (for 1 H^+) is added to the nominal mass for the first residue. Similarly, to calculate the masses for the *y* ion series, 19 u (for H_3O^+) is added to the nominal residue of the carboxy-terminal amino acid.

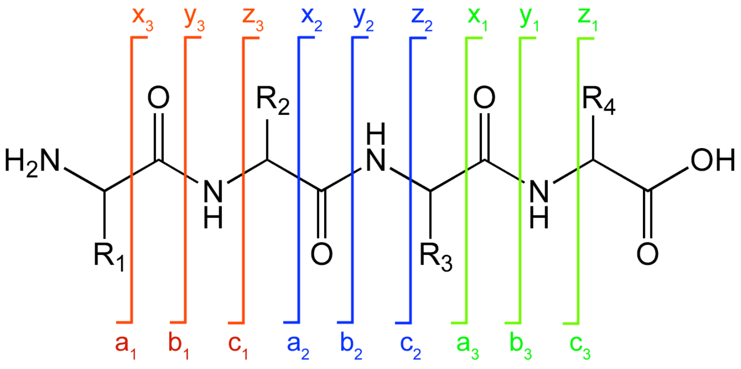


Figure 4 *Peptide ion fragment nomenclature.*

While it is relatively simple to calculate the elements of the *b* and *y* ion series from the peptide sequence, it is much less straightforward to read the amino acid sequence from the CID spectrum of a peptide ion. This is mainly because peptide fragmentation is sequence dependent, and the rules for fragmentation are not completely understood.

Each peptide tandem mass spectrum will contain *b* and *y* ions as well as other fragment ions that can be used to interpret the amino acid sequence. These include ions generated by the neutral loss of specific groups from amino acid side chains: Gln, Lys and Arg may loose ammonia (-17 u), while Ser, Thr, Asp and Gly may loose water (-18 u). In addition, *b* ions typically undergo a neutral loss of carbon monoxide, producing satellite ions (*a* ions), whose signal is 28 u lower than the corresponding *b* ion.

Although spectrum interpretation may be a difficult task and peptide trace is scattered on several ions, only a small subset of the CID spectra may be used to achieve a confident peptide identification (Cittaro, *et al.* 2005).

Interpretation of spectra is nowadays a computer-aided process. Whole protein databases are *in silico* cleaved, simulating enzymatic digestion, so that all theoretical peptides are calculated. For each peptide, a theoretical spectrum is inferred and matched to experimental MS/MS spectra in order to find the best matches. After peptide identification, proteins can be identified with a summarization strategy.

There are many dozens of scoring systems described in the literature, but in most cases these consist in attributing a score for each protein in the database and then calculating a measure of confidence that the top-ranking identified protein is not a false positive, such as in the case where the protein being investigated does not exist in the database (McHugh and Arthur 2008).

A recent benchmarking paper compared four publicly available algorithms and showed that of the 608 proteins identified by at least one of the algorithms, 335 were identified by all of the algorithms, with 70 being identified by only a single algorithm (Kapp, *et al.* 2005). The proteins identified by only a single algorithm were then independently manually verified, with most being determined by this expert validation to be correct identifications.

Chemical cross-linking

Cross-linking of protein species results in the formation of a covalent bond between two spatially proximal residues within a single or between two polypeptide chains. Besides the main reaction products, a number of side-reaction species may be observed. We will classify them either as mono-links (having one free end), loop-links (having both ends linked to the same peptide), and cross-links (having ends linked to different peptides) (Leitner, *et al.* 2010).

A large number of chemical cross-linking reagents have been developed. They may be classified in several categories according to their reactivity (*e.g.* amine- or thiol-reactive and homo- and heterobifunctional) or the incorporation of additional functional groups (*e.g.* cleavable sites and affinity tags).

Chemical cross-linking reagents consist of two reactive groups connected through a spacer or linker region, typically an alkyl chain. Usually, the reactive groups of cross-linkers target the primary amino group of lysine (and the proteins' N-termini). Most commonly succinimide-type linkers are used. The most common cross-linkers in this family are disuccinimidyl suberate (DSS; six-carbon linker) and disuccinimidyl glutarate (DSG; three-carbon linker) as well as their sulfo analogs bis(sulfosuccinimidyl) suberate (BS^3) and bis (sulfosuccinimidyl) glutarate (BS^2G), which are more soluble in purely aqueous solutions. DSS and DSG, in contrast, require prior dissolution in small volumes of polar organic solvents such as N,N-dimethylformamide or DMSO before addition to the sample. Structures of these cross-linkers are reported in Figure 5

Lysine cross-linking has several advantages, including the high prevalence of lysine residues (about 6 % in the proteome) and relatively high reaction specificity. Side reactions of N-hydroxysuccinimide esters with other amino acids usually do not occur at relevant levels under carefully controlled reaction conditions (pH, reaction times, and reagent excess). Similar specific cross-linking reactions can be carried out when targeting cysteine residues, *e.g.* by maleimides, but the low abundance of cysteine (2%) makes this approach less attractive. Other cross-linking chemistries are not frequently used either because the reactions cannot be performed under appropriate ("native") conditions or because reaction products are unstable or inhomogeneous.

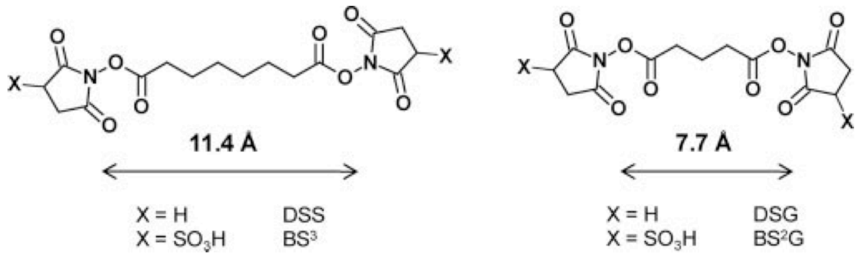


***Figure 5** Structures of most commonly used amine-reactive cross-linking reagents: DSS, BS[3], DSG, and bis(sulfosuccinimidyl) glutarate (BS[2]G) (Leitner, et al. 2010)*

A notable exception to the general linker design is formaldehyde, which only contains a single aldehyde group but is able to connect two amino acid side chains via a two-step reaction. Formaldehyde is a less specific reagent, although lysine and tryptophan residues are primarily targeted. Coupling reagents, for example carbodiimides such as ethyl diisopropyl carbodiimide, are only involved in an intermediate reaction step but do not introduce additional atoms into the molecule. The result is a so-called “zero-length” cross-link in the form of an amide bond between Lys and Asp/Glu residues that, however, requires very close spatial proximity (El-Shafey, *et al.* 2006).

To facilitate the analysis of the products of cross-linking reactions by mass spectrometry, different types of functionalized cross-linking reagents have been proposed. These include linkers carrying stable isotope labels that give characteristic fragmentation patterns in tandem mass spectrometry experiments. Stable isotope-labeled cross-linkers are used in a mixture of a cross-linker containing only light isotopes and a heavy (usually deuterated or C13 labeled) form of the reagent, reaction products carry a unique isotopic signature, a *m/z* shift in MS^1 spectra equal to the different number of neutrons. This feature is used for detecting peptides carrying mono- or cross-links among a large excess of unmodified peptides in enzymatic digests of complex samples but also facilitates the interpretation of MS/MS spectra of cross-linked peptides provided that the heavy and light form of the cross-linked peptides are sequenced. This is possible because only fragment ions containing the cross-link site are shifted in mass between light and heavy forms.

Analysis of mass spectrometry data from cross-linked samples

Chemical cross-linking coupled with mass spectrometry (CXMS) (Singh, Panchaud and Goodlett 2010) has gained popularity in recent years for characterization of inter- and intra-protein interactions in protein complexes; nevertheless it is a challenging undertaking mainly because of the overwhelming numbers of possible combinations that have to be considered by the search engine. In addition, the development of instruments suitable for high throughput LC-MS/MS analysis (*e.g.* FT-ICR and Orbitrap hybrid instruments) now allows the analysis of complex samples. Just like in many other modern research fields, technological improvements raised the demand for novel data analysis software to deal with advanced cross-linking workflows (Leitner, *et al.* 2010).

The algorithms that were developed for CXMS analysis use precursor mass and fragment ion mass information to identify cross-linked peptides and follow a strategy similar to commonly used search engines for peptide identification. The assignments are based on (a) selection of candidate cross-links from the sequence database, (b) matching of theoretical MS/MS spectra against acquired MS/MS spectra, and (c) scoring of possible candidate/spectrum matches to separate true from false positive identifications.

The number of possible peptide pairs is equal to

$$\binom{n+k-1}{k}$$

where n is the number of distinct predicted peptides (*i.e.* they have different sequences and also different modifications) and k=2, as we are only considering binary combination. As a consequence, the search space grows exponentially with increasing numbers of peptides. This "explosion" of the search space is the reason why the identification of cross-linked peptides from large sequence databases is such a challenging task.

Several strategies recently developed use restricted databases to reduce the search space. A common strategy in fact is to reduce the database to a small number of "target" proteins, possibly included in the purified complex under study. The composition of these samples is either known in advance or may be determined by common proteomics search strategies.

The approach previously described by Maiolica *et al.* is based on the generation of a database containing all possible linearized peptide pair permutations (XDB) where a single residue is considered a mono-link site modified with the cross-linker mass. The rationale is that the two peptides present in a cross-link cover the entire set of possible single bond fragments of the cross-link. The advantage of this method is that the MS/MS data may be searched using most of the search algorithms available in standard proteomics workflows, although it has been only published with MASCOT (Pappin, Hojrup and Bleasby 1993). The first implementation of the project described here follows and improves the very same idea and deploys it with an open source MS/MS search engine. The same approach has been later refined and implemented in xComb (Panchaud, *et al.* 2010). xComb generates a database of putative cross-linked peptides to be used with any MS/MS search algorithm. xComb approach has been recently deployed into commercial package Phenyx (GeneBio, Geneve, Switzerland).

The recently developed tools to identify cross-linked peptides using restricted databases are very useful if single proteins or small protein complexes are studied. Nevertheless, these approaches cannot be used if the composition of the sample is largely unknown or the sample is more complex, *e.g.* in the case of whole proteomes or subcellular fractions.

Currently, the most serious unresolved issue in computational approaches to cross-linking is the verification and validation of the results that the different algorithms provide without relying on manual validation. Several approaches have been developed to assess the quality of the match of a candidate cross-link spectrum to the theoretical spectrum. So far, cross-correlation scores, match ratio scores, and probabilistic (based on E-value) scores have been reported to separate true positive from false positive identifications. In this respect, it also has to be considered that the likelihood of false positive identifications does not increase in a linear fashion with the database size but rather quadratically like the number of combinations themselves (Leitner, *et al.* 2010). In addition, evaluation of published cross-linking data is typically very difficult. Frequently, the programs used for the analysis are no longer available or have never been released for public usage. Also, experimental sections often lack details about the databases used and validation criteria. Although this is a general problem of the proteomics community, the enormous diversity of workflows in the cross-linking field makes it particularly problematic.

Aim of the Work

Modern molecular biology has entered the “omics” era; the approach to the biological problem is no more focused on a specific process or molecule but rather on the global functional network. The interest of the researcher is increasingly pointed to the relationships among the parts. This is true at each level of the cellular environment, from genes to metabolic pathways.

The study of protein complexes becomes essential because proteins play primary roles in cellular processes. By contrast, few techniques are available to efficiently study protein-protein interactions and they often require manual investigation of the results. Recent advances in mass spectrometry opened the doors to an increasing number of applications to study protein complexes.

In this thesis, I am presenting two computational approaches for semi-automatic identification of cross-linked peptides using MS/MS spectra. The first approach (**silk**) extends and improves a previously published work (Maiolica, *et al.* 2007) with the addition of a more refined scoring system. The second approach (**nsilk**) is a new implementation that overcomes usage limitation of **silk**, it has been designed to be more flexible and, possibly, more powerful.

These programs have been used for structural analysis of three complexes with known three-dimensional structure (Ndc80 tetramer, GINS tetramer and RNA Polymerase II) and three complexes with unknown structure (RNA Polymerases I and III and the KMN complex). Both the programs have been benchmarked using Ndc80 data, as we already validated those.

Analysis of GINS and RNA Polymerase II was necessary to establish the power of our approach and to tune the running parameters. In addition, the two complexes represent the ideal lower and upper limits for real case analysis: GINS is a small complex composed by homogeneous sequences while RNAP II is a big complex composed by a heterogeneous set of proteins, varying in a wide range of length.

Materials and methods

Protein purification

Ndc80 complex has been purified as described in (Maiolica, *et al.* 2007). KMN network proteins have been purified as described in (Petrovic, *et al.* 2010). GINS complex has been kindly provided by Gulliermo Montoya laboratory (Structural Biology and Biocomputing Programme, Centro Nacional de Investigaciones Oncólogicas, Madrid, Spain). RNA polymerase I and RNA polymerase II have been provided by Patrick Cramer laboratory (Genzentrum, Department Chemie und Biochemie, München, Germany). RNA polymerase III has been provided by Cristoph W. Muller laboratory (Structural and Computational Biology Programme, EMBL, Heidelberg, Germany).

Cross-linking

Protein complexes analyzed were mixed with a 100x excess of isotope-labeled cross-linker bis(sulfosuccinimidyl)glutarate (BS^2G) (Pierce) in a final volume of 150 μl of 10 mM Hepes, pH 7.5, 100 mM NaCl at room temperature. The cross-linker, a 1:1 mixture of light BS^2G-d0 and heavy BS^2G-d4, was freshly prepared as a 10 nmol/μl solution in DMSO. The reaction was stopped after 30 min by adding 5 μl of 1 M ammonium bicarbonate.

Sample preparation for MS analysis

Proteins, after reduction, alkylation and digestion with trypsin, were analyzed by LC-MS/MS using an HPLC system (1100 binary nanopump, Agilent, Palo Alto, CA) coupled on line to an ion trap FTICR hybrid mass spectrometer (LTQ-FT, ThermoElectron, Bremen, Germany). C_{18} material (ReproSil-Pur C18-AQ 3 μm, Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) was packed into a spray emitter (75-μm inner diameter, 8-μm opening, 70-mm length; New Objectives) using an air pressure pump (Proxeon Biosystems, Odense, Denmark) to prepare an analytical column with a self-assembled particle frit. Mobile phase A consisted of water, 5% acetonitrile, and 0.5% acetic acid, and mobile phase B consisted of acetonitrile and 0.5% acetic acid. The samples were loaded from an Agilent 1100 autosampler onto the column at a 700 nl/min flow rate. The gradient had a flow rate of 300 nl/min, and the percentage of buffer B varied linearly from 0 to 20% in the first 77 min and then from 20 to 80% in a further 15 min.

Peaks were picked from the raw data files using DTAsupercharge version 0.94 (http://msquant.sourceforge.net/) with the following settings: precursor mass deviation, *m/z* 0.08; smart picking for MS/MS activated; maximum search level, 8. Raw data files were converted to mzXML format (Pedrioli, *et al.* 2004) using ReAdW.exe (http://ionsource.com/functional_reviews/readw/t2x_update_readw.htm); since ReAdW.exe is only available for MS Windows and uses XDA-api (Xcalibur Development kit, Thermo Inc.), we used wine (http://www.winehq.org) to run in UNIX environments.

MS/MS data analysis

In order to confirm the composition of each protein complex, a preliminary analysis has been performed using Mascot version 2.2 (MatrixScience, London, UK) with the following parameters: monoisotopic mass values; peptide tolerance 0.08 Da; MS/MS tolerance 0.5 Da; instrument ESI-TRAP; fully tryptic specificity; cysteine carbamidomethylation as fixed modification; oxidation on methionine and hydrolyzed cross-linker on protein N-terminus, lysine, serine, and tyrosine as variable modifications; two missed cleavage sites allowed.

Programming tools

Both **silk** and **nsilk** have been developed using python programming language (http://www.python.org); version 2.5 has been used for **silk**, version 2.6 has been used for **nsilk**.

silk depends on the following python libraries:

library	**what provides**
base64, struct, xml.dom	mzXML file import facilities
math	log() and exp() functions
scipy.stats	poisson distribution
zlib	crc32 hashing functions
csv	importing OMSSA result
io, os, sys	OS interaction
getopt	user interaction

nsilk depends on the following python libraries:

library	**what provides**
base64, struct, xml.dom	importing mzXML files
hashlib	md5 hashing function
numpy	numpy arrays and most of the mathematical functions
scipy.stats	distribution definitions
biopython	importing fasta files
io, os, sys, time	OS interaction
getopt, optparse	user interaction

Charts have been plotted using pylab (http://www.scipy.org/PyLab) supported by matplotlib v 1.0.0, scipy v 0.8.0 and numpy v1.5.0.

Results and discussion

Preprocessing MS/MS spectra

To enforce results and filter out false positives, every cross-link experiment has been carried out with a 1:1 mixture of light and heavy forms of BS^{2}G, having a 4 Da mass difference within the 2 forms.

This difference can be used to identify doublets in the MS spectrum and filter out all those fragmentation spectra not belonging to peptides present in both isotopic forms, hence likely to contain the cross-linker. To achieve this, we developed a small script (doubletCheck.py) that reads full raw data in mzXML format (Pedrioli, *et al.* 2004) and outputs two spectra files, one containing fragmentation spectra for light form of cross-linked peptides and one containing fragmentation spectra for the heavy form of the same peptides. An example of MS spectra for isotopic doublet is reported in Figure 6.

Two twin spectra are considered the heavy and the light form if the correlation between them is higher than 0.6. Correlation coefficient is calculated from the arrays of intensities of both spectra.

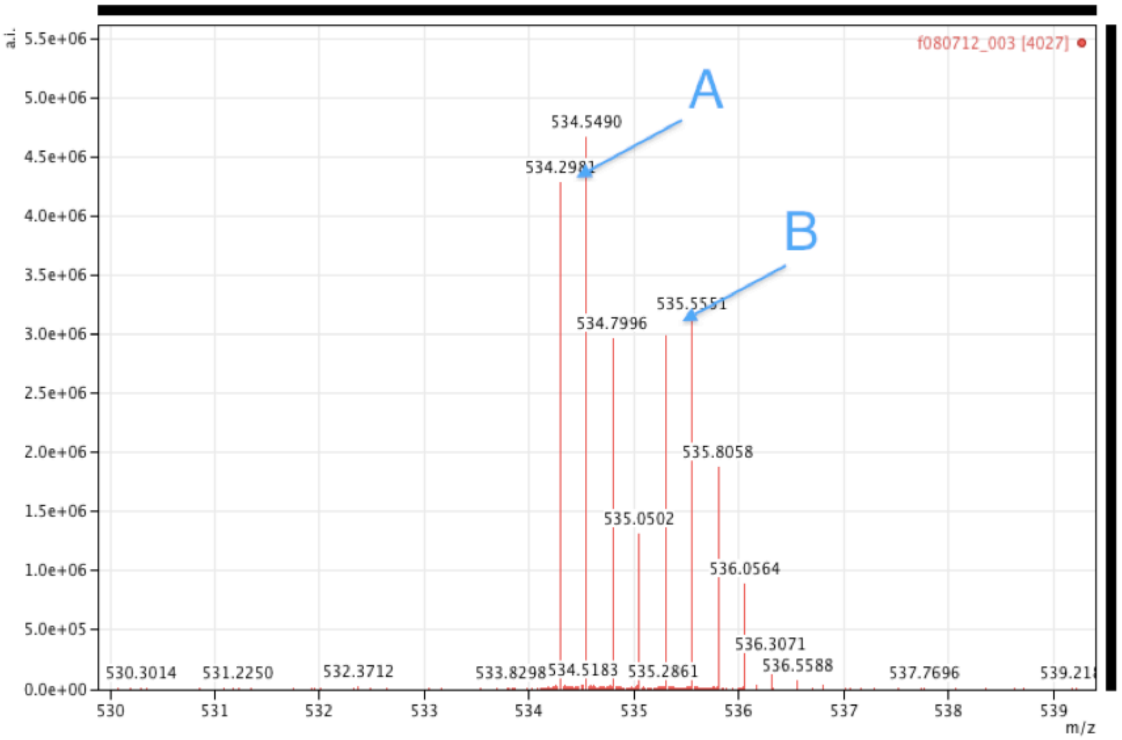


Figure 6 *MS spectrum of a doublet. The trace of a quad-charged light peptide (A) starts at 534.2981 m/z. The twin spectrum from the heavy form (B) starts at 535.5551 m/z. Four intensities of trace A and trace B are used to calculate the correlation coefficient.*

silk

The **silk** algorithm follows the algorithm previously described in XDB methodology (Maiolica, *et al.* 2007). The rationale behind this approach is to exploit common MS/MS search engines to identify cross-linked peptides. The trick, here, is to feed the search engine with fake linear peptides, whose matches identify couples of cross-linked peptides.

Formally we start from a database D made of n sequences S. D is also the set of supposed interacting proteins we want to study:

$$D = \{S_1, S_2, \dots, S_n\}$$

Each sequence S_n can be digested into m peptide sequences p_n; m may vary between different sequences.

$$S_n \rightarrow P_n = \{p_n^1, p_n^2, \dots, p_n^{'}\}$$

A linearized peptide X will be made by juxtaposition of peptides coming from two different proteins j and k:

$$X = p_j^a + p_k^b$$

The database of cross-linked peptides will contain all the possible pair wise combinations of single peptides. To collapse sequences and save space, a convenient representation in the cross-link database will be

$$D_x = p_j^a + p_k^b + p_j^a$$

To keep track of source proteins, we build the cross-links database naming each linearized peptide X with the source peptides and the source protein accession numbers. A typical example will look like

```
>xlh|xpep0x19| MFCEKAMELIR:GINS1 TLVKDMWDTRIAK:GINS2
MFCEKAMELIRTLVKDMWDTRIAKMFCEKAMELIR
```

The accession number of each X peptide should be unique. We chose to assign the combination of two progressive integers, each being the index of one of the source peptides.

Using an appropriate number of missed cleavages as parameter for the search engine ($2n + 1$, where n is the number of missed cleavages usually adopted), a combination of the original peptides p_j and p_k can be found from their *b* and *y* ion series. In the example above, the search engine may be able to find peptide MFCEKAMELIRTLVKDMWDTRIAK if there is a sufficient number of *b* ions belonging to peptide MFCEKAMELIR and a sufficient number of *y* ions belonging to peptide TLVKDMWDTRIAK. *Vice versa* is also true.

The approach described so far is not mass conservative: once we create a linearized peptide X, its molecular mass will be reduced by the introduction of a (fake) peptidic bond. The mass of X has to be

$$M(X) = M(p_j^a) + M(p_k^b) + M(C) + M(H_2O)$$

where $M(C)$ is the mass of the chemical cross-linker. To comply with the conservation of masses, we add the mass of water to the mass of the cross-linker. This species will be considered as a variable modification of X.

The search engine chosen for **silk** is the Open Mass Spectrometry Search Algorithm (OMSSA) (Geer, *et al.* 2004). This software has the major advantage to be open source and portable on different platforms.

Each MS/MS spectra file is searched against the cross-linked peptides database using OMSSA. Search parameters are tuned to be as little specific as possible. OMSSA is designed for reliable peptide identification and we should use it in a sub-optimal manner; the E-value filter is set to 1000 times the number of entries in the database. OMSSA results are collected and merged together. At this point, most of the identifications are not compatible with a cross-link. A filtering step is needed to retain only candidate cross-link.

Filtering is made using some “structural” rules: (1) the identified sequence should contain exactly the source peptide sequences, regardless of their order and (2) the identified sequence must contain exactly one site modified with the cross-linker.

In the example above, MFCEKAMELIRTLVK or KAMELIRTLVKDMWDTR are not valid cross-links because they do not contain exactly the source peptides (MFCEKAMELIR and TLVKDMWDTRIAK).

The aminoacid that is recognized as modified with the cross-linker by the search engine will be one of the cross-link sites on one peptide. The other cross-linked aminoacid is inferred by scanning the sequence of the second peptide for potential cross-linking sites (typically Lys). For each i position in the m possible cross-linking sites, the sum of *b* and *y* ions up to the position i matching to ions in the fragmentation spectrum is calculated. Maximizing this sum will identify the cross-link site:

$$\arg\max_{i=0}^{m} b(i) + y(i)$$

A score based on Poisson distribution is assigned to every identified cross-link:

$$p = \frac{\lambda^n e^{-\lambda}}{n!}$$

where λ is the number of expected ions according to the sequence, n is the number of matched ions. This scoring method is directly inherited from OMSSA peptide matching score.

Every cross-link matched to a spectrum is called “atomic identification” (AI). Many AIs may describe the same cross-link event. For this reason, multiple AIs that map the same cross-link on whole proteins are grouped into a single event. Every event is then scored using a mixed scheme. A pseudo-likelihood score G is calculated:

$$G = 2\sum_{i=1}^{n} k_i \ln(\frac{p_i}{p_i^0})$$

where r is the number of AIs, k is the number of istances of the i-th AI, p_i is the score of the i-th AI and p_i^0 is the null probability of the i-th AI. The G score is inherited from OMSBrowser scoring framework (Xu, *et al.* 2006).

In addition, a structural score is given according to the event properties. This score, called “rank”, is defined as a 5-bit mask (Table 1).

bit	**description**
0x0001	There are multiple spectra assigned to the event
0x0002	Both sequence have cross-linker as modification
0x0003	Different AI have different peptide sequences
0x0004	There are different cross-linkers (typically heavy and light form)
0x0010	Forward and reverse conformation have been identified

***Table 1** Bit meaning for rank score*

The condition described by bit 0x0010 implies that *b* and *y* ions for both peptides have been found in the fragmentation spectra. As showed in Figure 7, a single spectrum may contain clear traces of both peptides, and therefore OMSSA is able to identify both the “forward” and “reverse” conformations.

Among the command line options, a bitwise mask can be specified to filter rank score according to desired properties (Table 2). It should be pointed that OMSSA uses much more ion information than the *b* and *y* ion series alone, therefore the number of matching ions may seem lower in **silk**.

```
[ Event 1 ]
Coordinates: HEC1_SPC25:59 HEC1_SPC25:156
G-Score: -659.756195175  [ 2.17 ]
Rank: 23
N. of atoms: 19
N. of spectra: 6
Xlink atomic identifications
…
[ 6 ]
        Left Peptide: VSLFGKR [   ]
        Right Peptide: IFKDLGYPFALSK [ x-BS2GD4 (K):3 ]
        p-value: 0.000335462627903
        Spectrum Number: 1220 [ FinneganScanNumber: 12440 ]
        Charge: 4       m/z: 601.845325032      Error (ppm): 0.0
VSLFGKR                                      IFKDLGYPFALSK
b1 V 100.076 []                              b1 I 114.092 []
b2 S 187.108 []                              b2 F 261.160 []
b3 L 300.192 [300.097, 416.30]               y10 D 1110.58296072 [1110.933, 1056.60]
b4 F 447.261 []                              y9 L 995.556 [995.84, 395.90]
b5 G 504.282 []                              y8 G 882.472 [882.92, 6608.10]
y1 R 175.119 []                              y7 Y 825.450 [825.47, 1482.40]
                                             y6 P 662.3872 [662.87, 9049.20]
                                             y5 F 565.334 [565.01, 1170.40]
                                             y4 A 418.266 []
                                             y3 L 347.229 []
                                             y2 S 234.145 [234.20, 355.60]
                                             y1 K 147.113 []
[ 7 ]
        Left Peptide: IFKDLGYPFALSK [   ]
        Right Peptide: VSLFGKR [ x-BS2GD4 (K):6 ]
        p-value: 0.000335462627903
        Spectrum Number: 1220 [ FinneganScanNumber: 12440 ]
        Charge: 4       m/z: 601.845325032      Error (ppm): 0.0
IFKDLGYPFALSK                                VSLFGKR
b1 I 114.092[]                               b1 V 100.076 []
b2 F 261.160 []                              b2 S 187.108 []
y10 D 1110.583 [1110.93, 1056.60]            b3 L 300.192 [300.10, 416.30]
y9 L 995.556 [995.83, 395.90]                b4 F 447.261 []
y8 G 882.472 [882.92, 6608.10]               b5 G 504.282 []
y7 Y 825.450 [825.47, 1482.40]               y1 R 175.119 []
y6 P 662.387 [662.87, 9049.20]
y5 F 565.334 [565.01, 1170.40]
y4 A 418.266 []
y3 L 347.229 []
y2 S 234.145 [234.20, 355.60]
y1 K 147.113 []
…
```

Figure 7 *Example of forward/reverse cross-link identification*

option	description	default value
-d STRING	database with protein sequences	nr
-i STRING	csv formatted OMSSA results	none
-f STRING	mgf formatted spectra	none
-t FLOAT	MS/MS tolerance	0.5
-u [ppm I Da]	MS/MS toleance unit	Da
-e FLOAT	MS tolerance (ppm)	10
-l INT	minimum peptide length	3
-g FLOAT	*G* score filter	0.0
-p FLOAT	*p*-value filter	0.05
-q FLOAT	*q*-value (FDR) filter	None
-m [bitmask I INT]	bitmask or integer mask for rank filtering	None
-x	include complex ions when scoring	False
-c	csv output	FALSE
-n INT	min. number of cross-links/interaction	1
-r FLOAT	noise reduce ratio	0.03
-D	disable any filter (debug)	False
-F STRING	Fixed Modifications, comma separated	carbamidomethyl C

***Table 2 silk** command line parameters*

nsilk

The **silk** approach is not straightforward; the rationale behind is elegant but not easy to understand and to spread; it also needs two steps to identify a cross-link. In addition, the OMSSA comes with a high number of options and it may be hard to tune it properly.

To overcome all these difficulties, we implemented a new approach, called **nsilk** (new **silk**). This new approach is founded on more straightforward principles but it is easy to tune and may become distributable.

nsilk takes as input a fasta file containing the sequences of candidate proteins (that may have been previously identified with standard procedures) and the MS/MS spectra file(s).

Proteins are *in silico* digested and peptides that can be involved in a cross-link (according to the cross-linker definition) are retained in a list. A loop cycles over this list to create all the peptide couples. For each couple, the molecular mass is calculated and spectra having a matching precursor mass are assigned to the new cross-link. Each cross-link/spectrum association is given a score S based on the negative binomial distribution:

$$S = \sum_{i=0}^{r} NB(1, k-r, p)v_i$$

where k is the length of the peptide, r is the number of matching ions (*b* and *y* ions only), p is the maximum likelihood estimator (MLE) of the probability of matching a single ion and v_i is the percentile of the intensity of the i-th ion in its spectrum. For negative binomial distribution

$$p = \frac{r}{r+k}$$

In other words, we calculate the probability of matching one ion given $k - r$ non-matching ions; we weight this probability on the ion intensity.

Logarithms of individual scores are used to model a normal distribution, so that a p-value can be assigned to each cross-link identification (Figure 8).

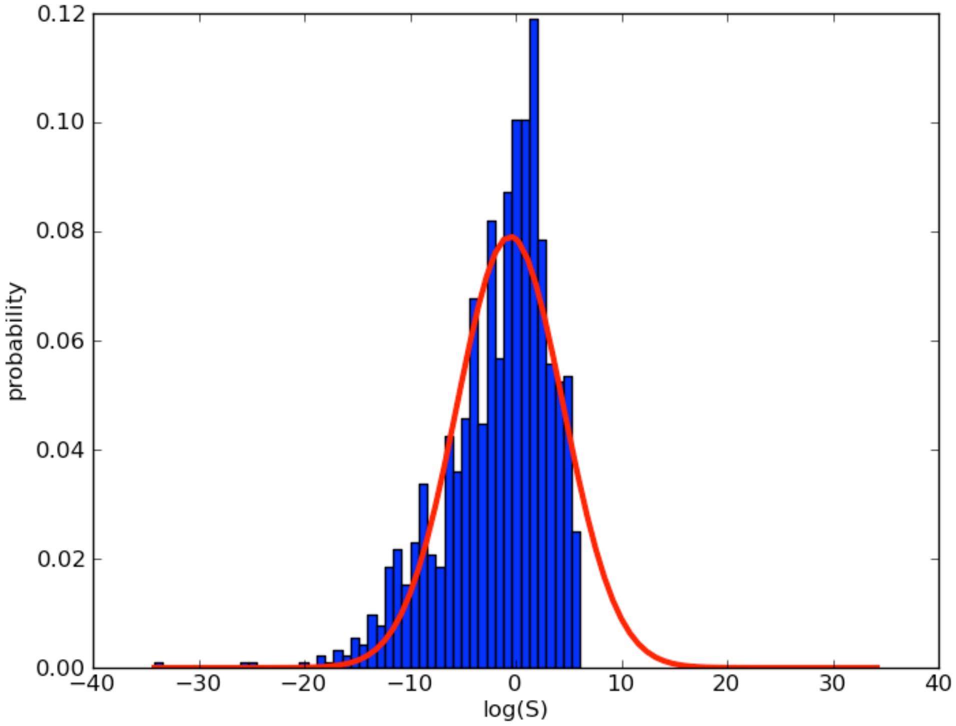


Figure 8 *Distribution of log(S). Red line shows the normal distribution modeled after score values*

As previously described in the **silk** approach, multiple identifications of the same physical event are collapsed into an “interaction”. Each interaction is scored using either the median of individual scores S and a likelihood G score, similar to the one previously defined and conceptually backward compatible:

$$G = 2\sum_{i=1}^{n} k_i \ln(\frac{p_i}{p_0})$$

where n is the number of cross-link identifications in the event, k_i is the number of spectra matching the i-th cross-link, p_i is the p-value for the cross-link and p_0 is the inverse of the number of cross-links that can match the spectrum (*i.e.* the possible alternative matches). It should be noted that multiple spectra could be assigned to single cross-link identifications; also, many cross-link identifications, possibly differing in sequence or modifications, can explain a single spectra. With this “many-to-many” relationship in mind, any interaction gets higher scores if a cross-link explains many spectra and if each spectrum is assigned to a small number of cross-links.

nsilk command line options are explained in Appendix A.

Experimental application

To test the performance of the **nsilk** software, we benchmarked it using data from a previously published experiment, the manually validated analysis of the Ndc80 Complex (Maiolica, *et al.* 2007).

We also tested the software with two protein complexes having publicly available three-dimensional structures (GINS and Polymerase II).

Finally, we used **nsilk** to evaluate the topology of a protein complex with unknown structure (KMN).

Ndc80

Ndc80 is a tetrameric complex that plays a key role at the kinetochore-microtubule interface (Ciferri, Musacchio and Petrovic 2007). It is conserved in higher eukaryotes and it is composed of four subunits, known as Hec1, Nuf2, Spc24 and Spc25.

The Ndc80 complex localizes at centrosomes during the interphase; from late G2 phase it relocates to the kinetochore outer plate, where it remains bound at nearly constant levels until late anaphase. Analyses in different organisms have demonstrated that interference with the Ndc80 complex affects microtubule–kinetochore attachment, chromosome congression and chromosome segregation (Kline-Smith, Sandall and Desai 2005).

The molecular weights of the subunits of human Ndc80 are 73.9 kDa, 54.3 kDa, 26.1 kDa and 22.4 kDa, respectively. All four subunits contain long coiled coils that form a rod structure with globular domains at both ends. One globular end of Ndc80, made by Hec1 and Nuf2, binds the microtubules, while the other, made by Spc24 and Spc25, is important for kinetochore localization.

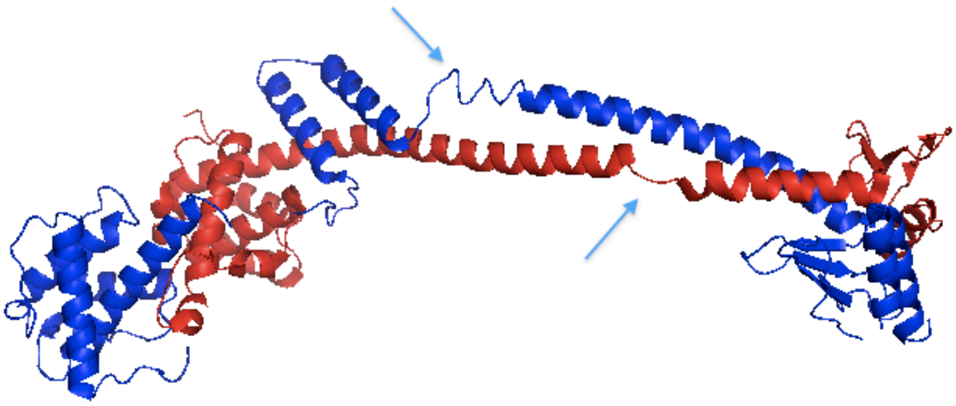


Figure 9 *3D structure of Ndc80 complex, bonsai construct. Blue chain is Hec1-Spc25 fusion protein; red chain is Nuf2-Spc24 fusion protein. Blue arrows indicate where proteins fuse.*

The Ndc80 complex was included in our analysis because we have a list of 26 manually verified cross-links. Among the 26 validated cross-links, 6 have not been considered in our analysis, because they include peptides that are only 3 residues long; such small peptides can easily generate false positives in a standard analysis when manual validation is not possible; in practice, it is hard to correctly map a tripeptide in a set of protein sequences, as it may be not uniquely represented. The three-dimensional structure of an engineered version of the Ndc80 complex has been recently reported (Ciferri, *et al.* 2008) (Figure 9). This implies that some regions, especially those that map in the middle of the central shaft, which contains coiled-coil regions, cannot be verified, as they have not been characterized by X-ray crystallization.

When running with standard parameters, **silk** produces a list of 127 events. 13 events are included in the validated list (Figure 10). The G score for these identifications varies in the range between -11.06 and -1536.83. The rank for the same events varies in a range from 1 to 31. The last observation implies that an event should be always described by more than one spectrum. Hence, we can constrain the rank to be always an odd value.

A total of 33 identifications show a G score lower than -11.06 and an odd rank. As stated above, 13 of these have been previously validated and 11 involve a tripeptide with multiple positions on different proteins (*e.g.* Nuf2 contains two occurrences of peptide KEK). Among the remaining 9 events, 3 are compatible with Ndc80 structure (Hec1:381-Nuf2:271, Nuf2:255-Hec1:287, Nuf2:271-Nuf2:354), two arise from a software bug and 5 are false positives (Nuf2:248-Nuf2:1, Hec1:122-Spc24:1, Spc24:117-Spc25:10, Spc25:152-Nuf2:26, Spc25:215-Nuf2:356).

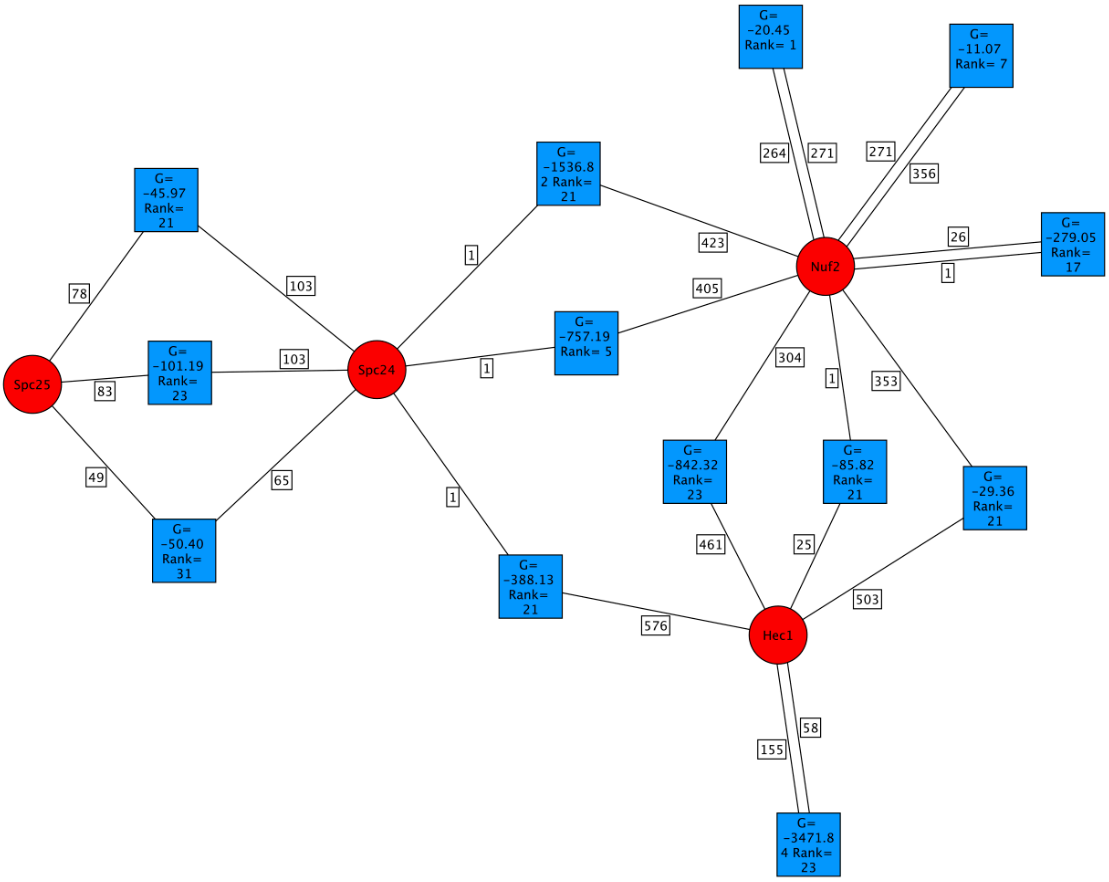


Figure 10 *Topology map of validated interactions in Ndc80 complex identified by* ***silk****. Red edges are proteins; blue edges are interaction scores; each vertex is labeled with the aminoacid involved in the cross-link*

nsilk performance is quite similar. The **nsilk** run (standard parameters except for MS tolerance set to 0.5 Da) gives 197 interactions, 13 of which are among the verified interactions (Figure 11). *G* scores vary in a range between -0.41 and -227.03. Median Match Scores (*MMS*) vary in a range between 13.43 and 256.34. A total of 30 interactions fall in the same *G* and *MMS* ranges. Interestingly, **nsilk** is able to spot at least one interaction between Spc25 and Hec1, allowing for a slightly more precise description of the complex topology. Also, default **nsilk** parameters filter peptides shorter than five aminoacids. In the results of **nsilk**, *G* values for positive identifications cover approximately half of the whole range (-227.03 – 44.82). The *MMS* values for the same identifications all lie in the 4-th quartile. If we exclude the worst *MMS* value, the remaining twelve lie in the 90-th percentile (P_{90}), which includes 20 elements. It is worth noting that among the 8 not validated interactions in the 90-th *MMS* percentile (P_{90}*(MMS)*, *i.e.* *MMS* ≥ 43.02), two involve Lys residues 2-4 aa afar (Hec1:420-Hec1:422 and Hec1:632-Hec1:628); one is a mono-link (Hec1:628-Hec1:628); two are compatible with the Ndc80 model (Hec1:503-Spc24:1, Nuf2:463-Spc25:72): globular domains are mapped near the N-terminus of Hec1 and Nuf2 and near the C-terminus of Spc24 and Spc25. The remaining three interactions (Hec1:122-Spc24:1, Hec1:36-Hec1:632, Spc24:1-Spc25:147) can be considered false positives.

Using the P_{90}*(MMS)* rule for **nsilk** and the *G* range plus odd rank rule for **silk**, excluding dubious identifications (*i.e.* the ones that cannot be validated on Ndc80 structure) from our analysis, we can summarize the results as follows:

silk			Sensitivity=0.65
	T	**F**	Specificity=0.85
T	*13*	*16*	PPV=0.45
F	*7*	*94*	NPV=0.93

nsilk			Sensitivity=0.6
	T	**F**	Specificity=0.98
T	*12*	*3*	PPV=0.8
F	*8*	*177*	NPV=0.96

Although **nsilk** shows a slightly lower sensitivity, it outperforms **silk** on specificity rate. These observations and the easiness of use make **nsilk** the preferred approach for any further analysis.

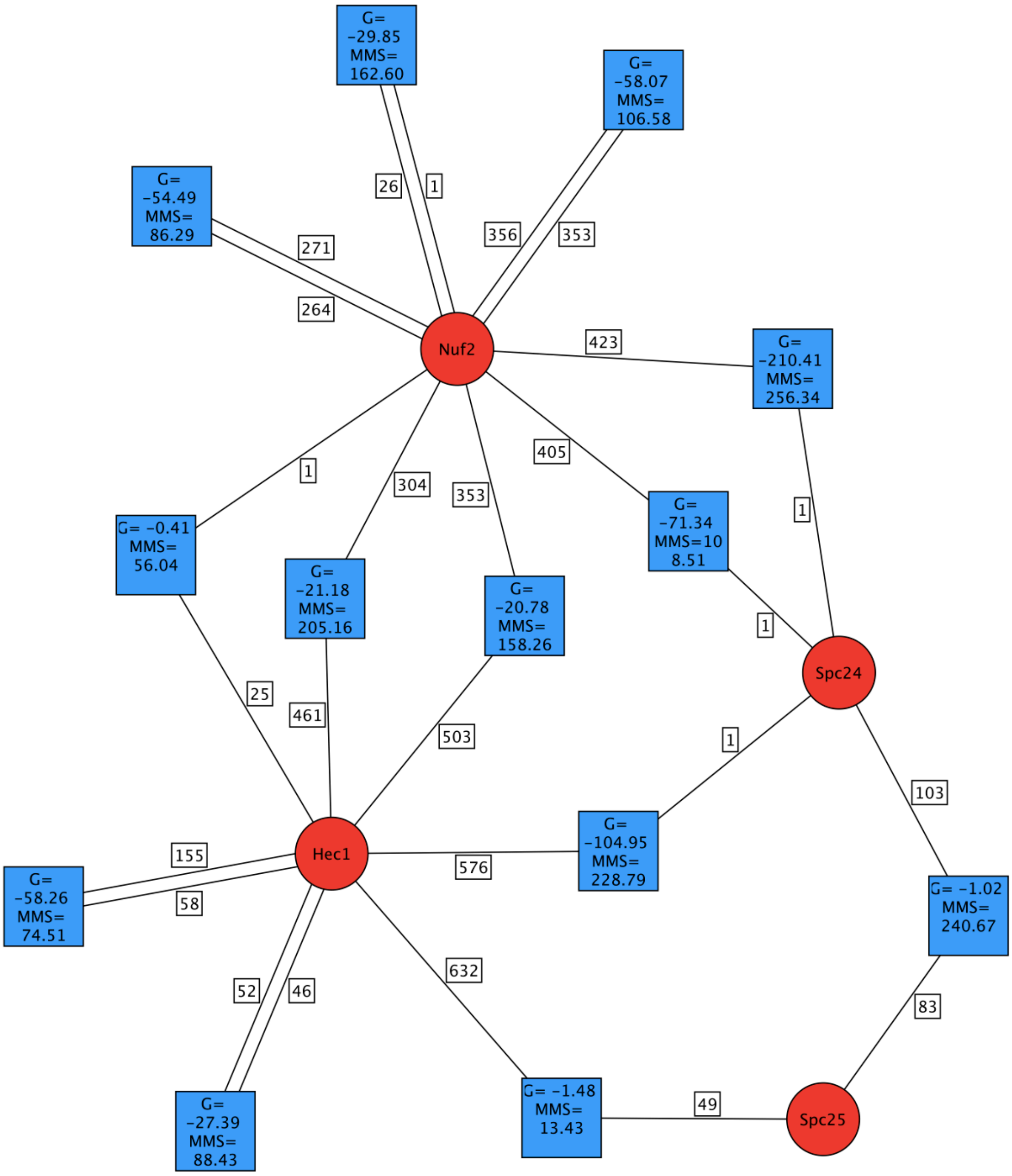


***Figure 11** Topology map of validated interactions in Ndc80 complex mapped by **nsilk**.*

GINS

The GINS (四, 一, 二, 三, Go, Itchi, Ni, San) complex was first described in yeast as a result of genetic analyses aiming to discover proteins that interact with DNA polymerase B possible subunit 11 (Dpb11) (Takayama, *et al.* 2003). It is a heterotetrameric complex consisting of Sld5 (synthetic lethality with Dpb11), Psf1 (Partner of Sld5-1), Psf2 and Psf3; each subunit is relatively small (~200 aa) and highly conserved in all eukaryotes. GINS has been shown to interact directly with the DNA primase and is essential for initiation of DNA replication and normal progression of the replication fork (De Falco, *et al.* 2007).

The three-dimensional structure of GINS complex has been determined; the four subunits assemble around a central hole predicted to host DNA.

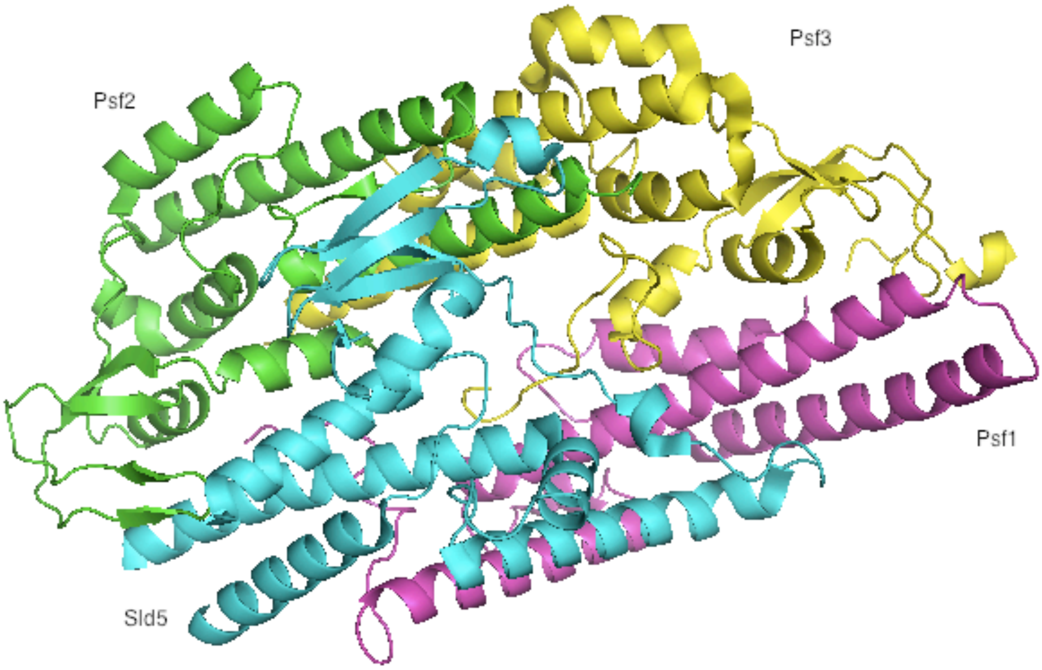


Figure 12 *3D structure of GINS complex (PDB ID:2Q9Q)*

The 3D structure (PDB ID: 2Q9Q) was used to validate cross-link experiments on GINS (Figure 12). It is worth warning that some protein regions are missing from the crystal structure, as they are part of unstructured regions that cannot be resolved. 12 MS datasets were available for this analysis. From a total of 22267 fragmentation spectra, 5764 have been retained as potentially containing the isotopic doublet. A cross-link has been considered valid if the C-α distance of the cross-linked aminoacids is less than 25 Å. This value comes from the sum of the length of the cross-linker arm (7.7 Å), the length of Lys residues side chain (6.3 Å) and a ~20% tolerance introduced to deal with chain flexibility.

Using standard run parameters, 94 out of 11989 peptide combinations are compatible with cross-linked peptides masses. 29 out of 64 interactions are retained after filtering. 21 interactions involve a peptide that has been resolved in the crystal structure. Assuming the $P_{90}(MMS)$ rule adopted above is valid ($MMS \geq 140.64$), we get the following confusion matrix:

$\boldsymbol{MMS \geq 140.64}$			Sensitivity=0.5
	T	**F**	Specificity=0.67
T	*1*	*6*	PPV=0.14
F	*1*	*12*	NPV=0.92

Loosing the filter to $P_{70}(MMS)$ ($MMS \geq 9.06$), we slightly improve the performance in terms of mapped interactions, with a minor impact on precision:

$\boldsymbol{MMS \geq 9.06}$			Sensitivity=0.4
	T	**F**	Specificity=0.67
T	*2*	*5*	PPV=0.28
F	*3*	*10*	NPV=0.77

The G scores distribute in different ways between positive and negative matches.

G score distribution for false identifications is centered near the zero (μ=-2.62, σ=15.74) and shows a small variance; distribution for true identifications is shifted to negative values (μ=-168.67, σ=450.92), although it shows huge variance. A plot of both distributions (Figure 13) suggests that very negative values for G may help in discriminating false positives (FP) from true positives (TP). Adding a second filter based on G score allows for a loose MMS filter keeping the performances acceptable. To define a reasonable value for G, a Receiver Operating Characteristic (ROC) plot of some scores was plotted (Figure 14).

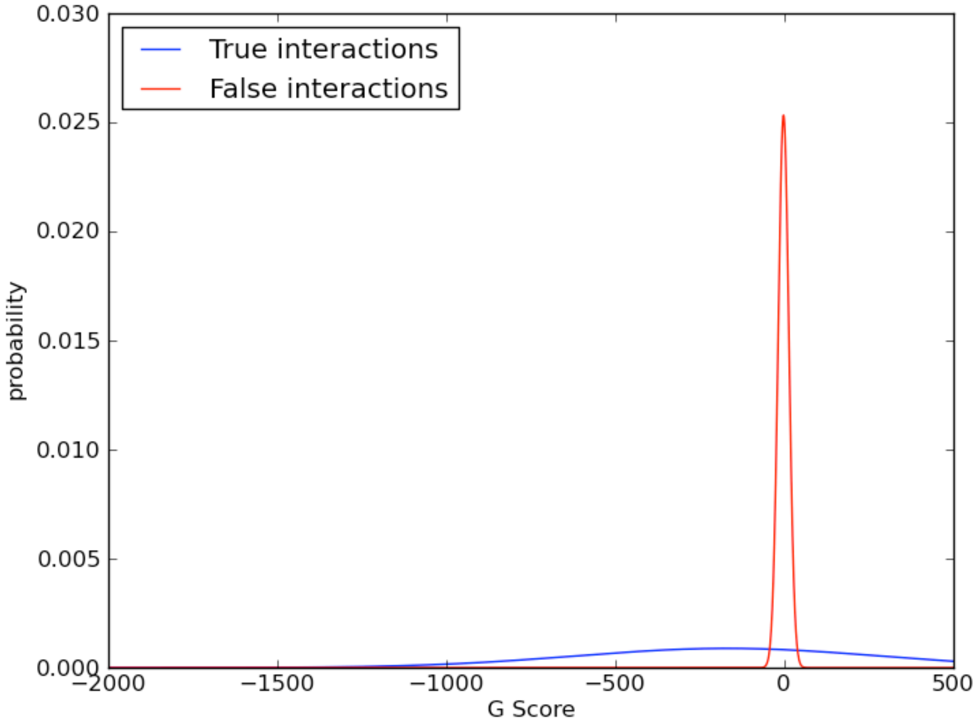


Figure 13 *Distribution of G scores for true and false identifications in GINS dataset.*

Setting a threshold on *G* to -30 gives better performances in terms of classification; unfortunately, the number of positive interactions is low:

MMS ≥ 9.06 & G ≤ -30			Sensitivity=0.5
	T	**F**	Specificity=0.67
T	*1*	*5*	PPV=0.17
F	*1*	*10*	NPV=0.91

One **nsilk** parameter may be relevant for refining the cross-link identification. The maximum number of cross-links that can match to a single spectrum is tunable; a default search will run without a limit.

If we apply a limit of one spectrum per cross-link (*i.e.* we retain best matches only), the number of interactions will be lower (21 interactions, 14 can be mapped on 3D structure), a threshold equal to $P_{80}(MMS)$ performs much better in terms of sensitivity:

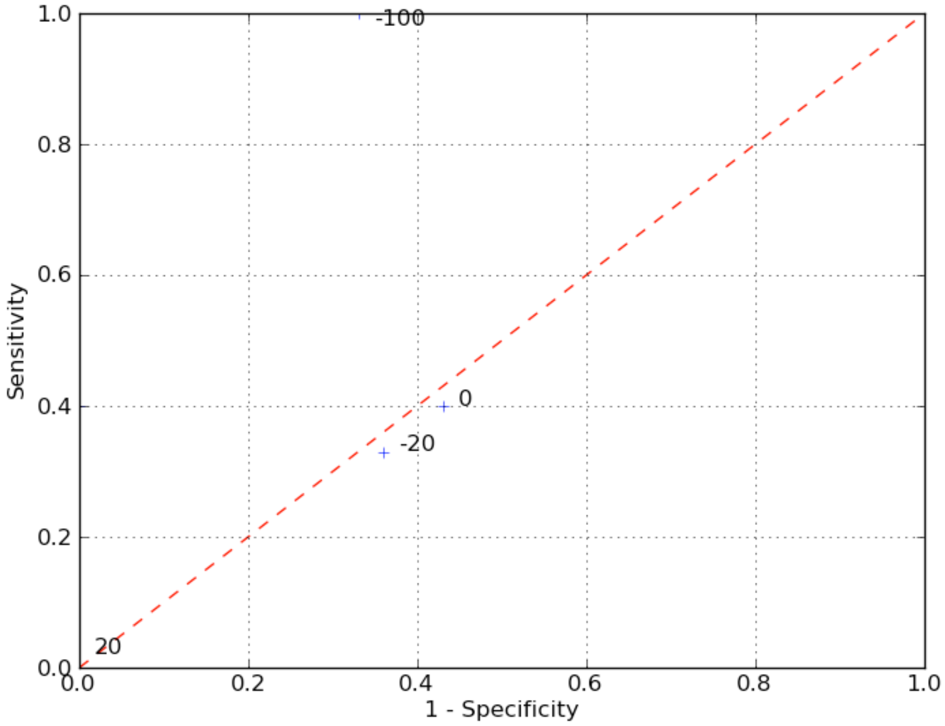


Figure 14 *Receiver operating characteristic plot for some G values at fixed MMS threshold. The dashed line represents random assignments.*

Assuming that the rule $P_{80}(MMS)$ is valid when the number of retained spectra is limited to 1 (R1), the threshold to filter 21 interactions is 147.48; five interactions are above this threshold, one of which was reputed false based on the 3D structure analysis. Five interactions are not sufficient to build the topology map of GINS complex (Figure 15).

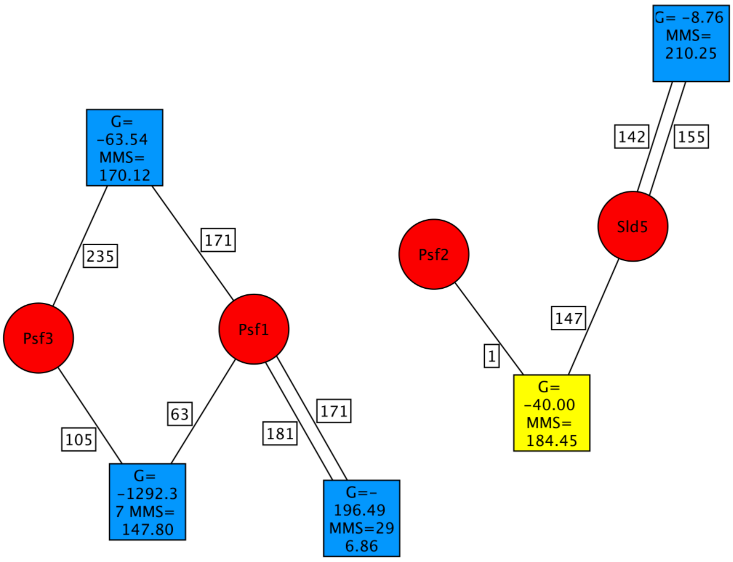


Figure 15 *Topology map of interactions in GINS assuming R=1 and MMS ≥ 147.48. Yellow nodes mark false positive interactions according to the 3D strutcture.*

RNA Polymerase II

The DNA-directed RNA Polymerase II (RNAP II) is an enzyme that catalyzes the transcription of DNA to synthesize precursors of mRNA and most snRNA and miRNA (Carty, *et al.* 2000; Cho, *et al.* 1998). The mass of the entire complex, composed of 12 subunits, is approximately 550 kDa. The high-resolution structure of RNAP II has provided detailed insight at the atomic level into how an active RNAP II is structured (Cramer, Bushnell and Kornberg 2001; Woychik and Hampsey 2002; Kettenberger, Armache and Cramer 2003; Brueckner and Cramer 2008) (Figure 16).

RPB1 is the largest subunit; it forms the DNA binding domain in combination with RPB9. It also interacts with RPB8. The second largest subunit, RPB2, forms a structure that maintains contact in the active site of the enzyme between the DNA template and the synthesized RNA. RPB3 is part of a core subassembly with RPB11 and is involved in RNAP II assembly; it interacts with RPB1-5, RPB7, RPB10-12. RPB5 is a dimer within the mature structure of RNAP II and interacts with RPB1, RPB3 and RPB6. RPB4 and RBP7 are present at substoichiometric levels and form a subcomplex, which has recently been described by X-ray diffraction along with the whole RNAP II (Woychik and Hampsey 2002). RPB5, together with RPB3, binds all other subunits of the complex, except for RPB9.

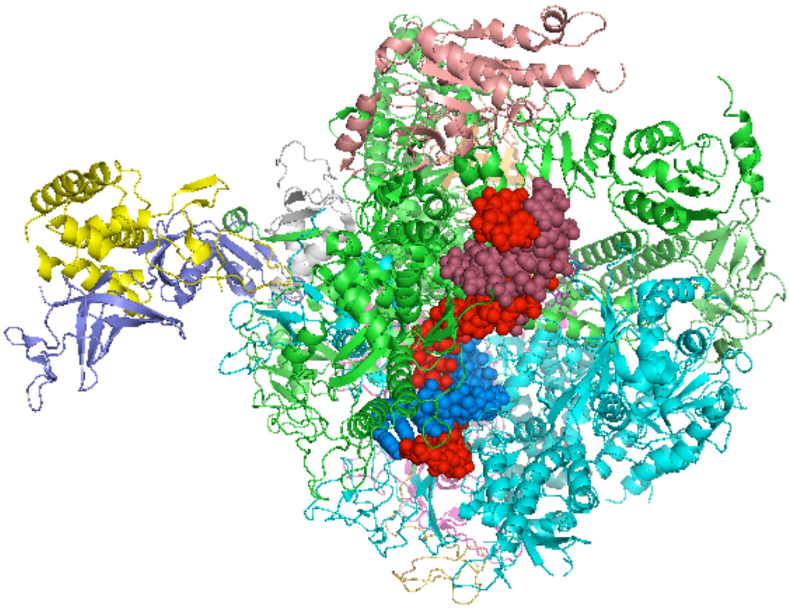


Figure 16 *3D structure of RNA polymerase II (PDB ID: 2VUM). Protein chains are rendered as ribbons, nucleic acids as spheres. The emergin RNA filament is colored in blue*

A total of 15 MS/MS acquisitions were used for **nsilk** analysis. 18959 fragmentation spectra out of 74158 have been identified as potentially derived from cross-linked peptides by searching the isotopic doublet. Standard **nsilk** parameters match 10451 out of 256541 possible peptide combinations with MS data according to the molecular mass; these combinations have been collapsed into 3134 valid interactions (out of 5933).

In order to validate the interactions, the 12-subunit structure has been used as a reference (PDB ID: 2VUM). 2359 interactions can be mapped on the resolved structure. If all possible matches are assigned to a spectrum and scored (R0), both the *MMS* and *G* are useless in discriminating true and false identifications. The plot of distributions for both scores (Figure 17) shows completely overlapping curves, although curves for true interactions show a slightly better pattern (*i.e.* higher *MMS* and lower *G*).

As expected, the performance in identification is extremely low:

$MMS \geq 29.89$ (P_{90})			Sensitivity=0.16
	T	**F**	Specificity=0.95
T	*38*	*104*	PPV=0.27
F	*192*	*2024*	NPV=0.91

The reason of this behavior may be the high number of possible cross-links and spectra. Spectra that match more than twenty species are quite common. The same effect was not observed in previous experiments, presumably due to the smaller number of fragmentation spectra and peptides.

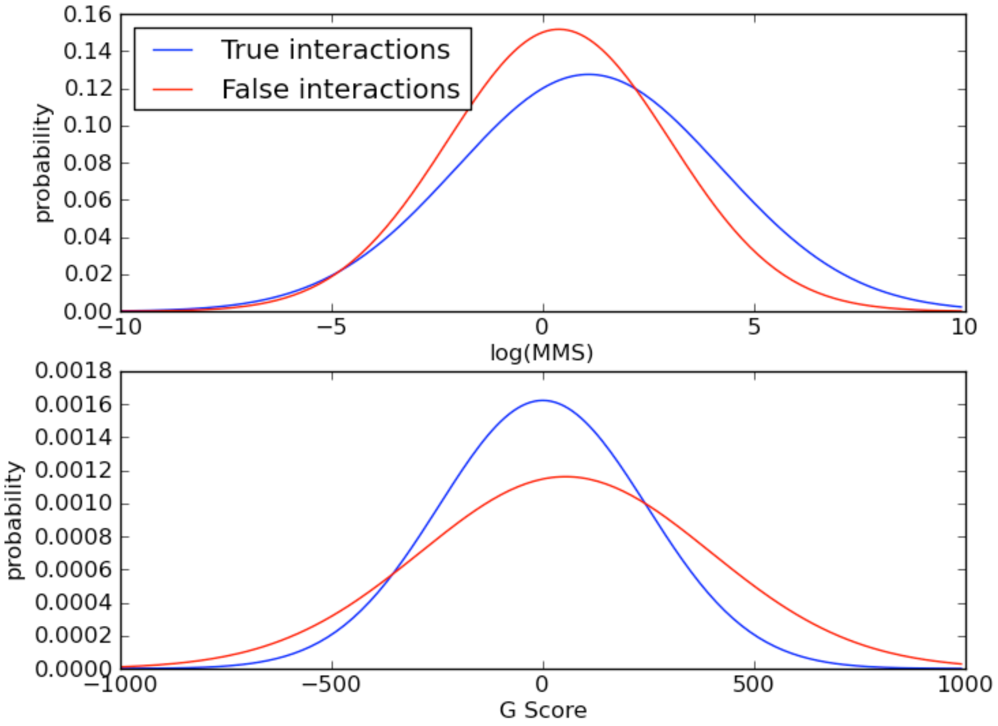


Figure 17 *Distribution of **nsilk** MMS scores (top) and G scores (bottom) for PolII experiment when all spectra are retained.*

For this reason, it is mandatory to filter results allowing a maximum number of matches per spectrum. If we allow for best matches only (R1), 759 valid interactions are retained, 510 of which can be mapped on the 3D structure.

Distributions of both G and MMS score can be discriminated in a way similar to the GINS example (Figure 18).

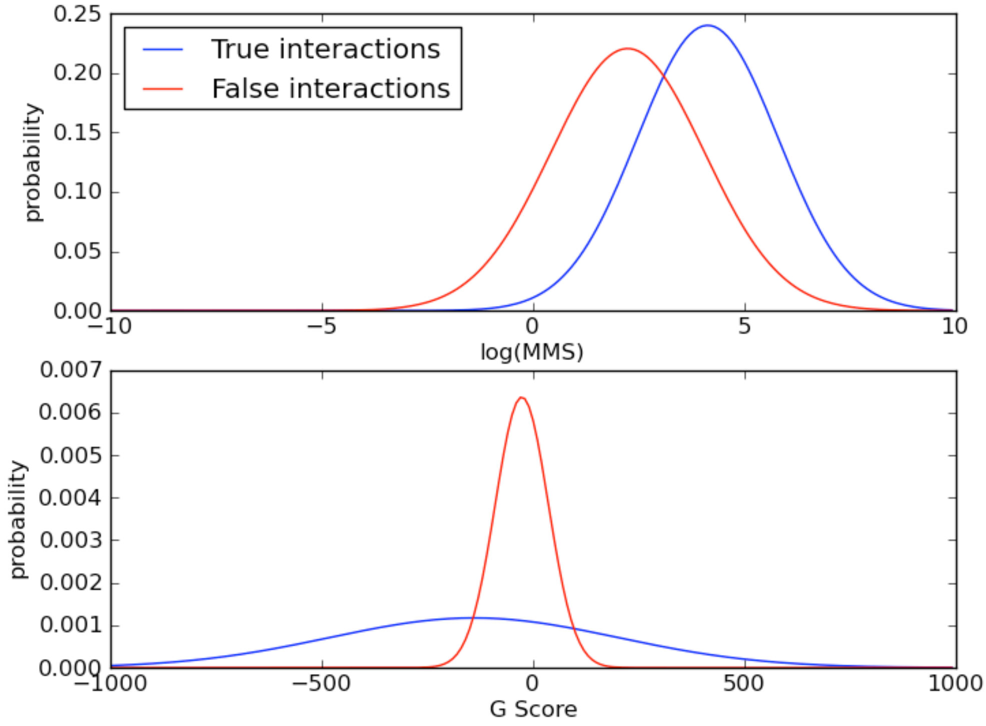


***Figure 18** Distribution of **nsilk** MMS score (top) and G score (bottom) when only best matching cross-links are retained.*

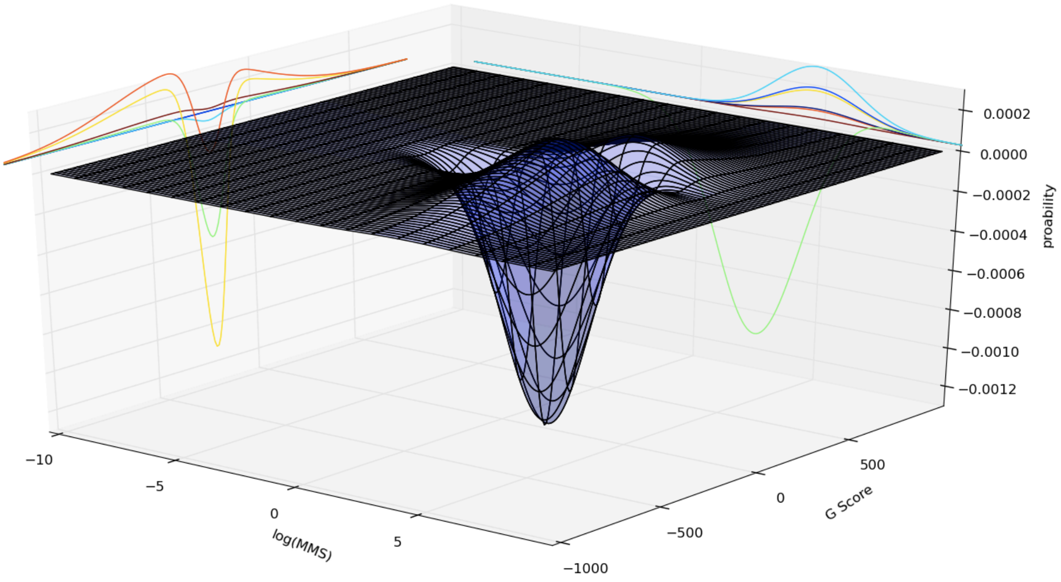


***Figure 19** Difference of distributions for true and false identifications. Ranges in which the surface is different from zero identify the possibility to differentiate true from false identifications*

Also, the difference of the distributions shows there are ranges in which true identifications emerge (Figure 19).

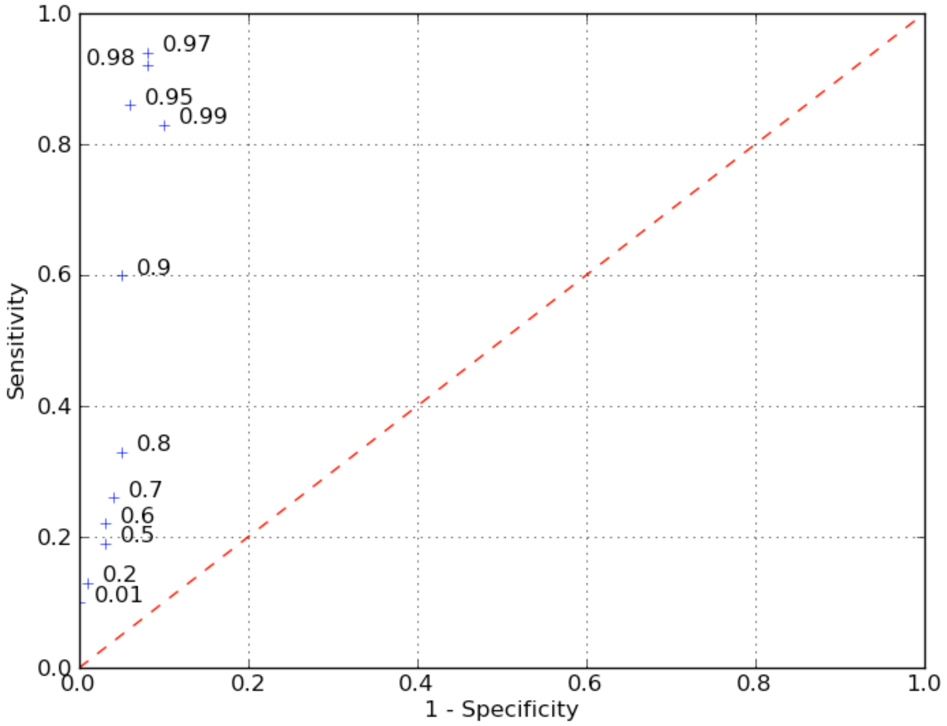


Figure 20 *ROC analysis of MMS score for PolII data. Numbers next to points indicate the threshold percentile*

The high amount of data allows for a more detailed choice of MMS and G thresholds. ROC analysis shows that above P_{90} we achieve a sensitivity higher than 0.5 (*i.e.* we can identify more true positives than false positives) (Figure 20); above P_{97} there is a degradation of sensitivity. If we set the threshold to P_{95} (MMS) we have a good balance between performance and number of identifications:

$MMS \geq 131.24$ (P_{95})			Sensitivity=0.86
	T	**F**	Specificity=0.94
T	*25*	*30*	PPV=0.45
F	*4*	*451*	NPV=0.99

In order to find a good trade-off between MMS and G threshold, a ROC curve analysis has been performed for different percentile levels (Figure 21).

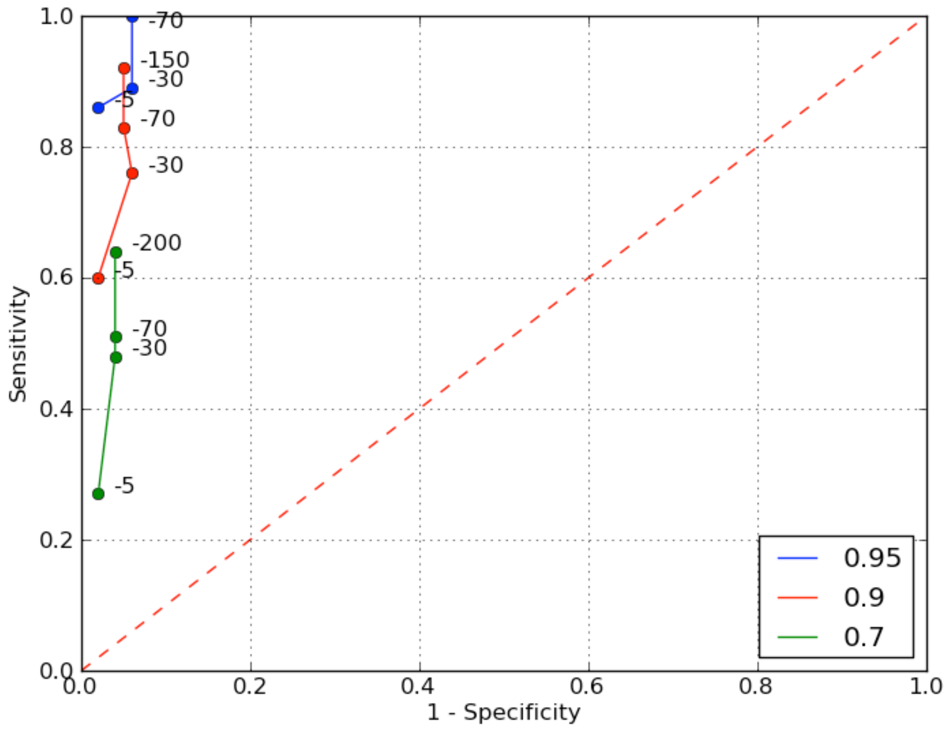


***Figure 21** ROC analysis of different G scores at varying MMS threshold. The MMS threshold are expressed as percentile*

The analysis suggests that setting the MMS threshold at 131.24 (P_{95}) outperforms any other choice, implying that a G threshold may be useful for lower MMS cutoffs. Setting MMS threshold to 80.48 (P_{90}) and G threshold at -70, we get the following confusion matrix:

$MMS \geq 80.48$ & $G \leq -70$			Sensitivity=0.83
	T	**F**	Specificity=0.95
T	*19*	*23*	PPV=0.45
F	*4*	*464*	NPV=0.99

A further filter can be applied on **nsilk** output. Individual cross-links can be filtered using the p-value threshold. In order to evaluate the effect of this filter, we applied three different filters on the “best match” dataset. The number of resulting interactions decreases with the p-value threshold (Table 3)

p-value threshold	number of interactions	max. number of cross-links/ interaction
1	759	27
0.5	585	27
0.2	355	27
0.1	203	26

Table 3 *Number of interactions that can be spotted at different p-value filters.*

We expect G scores to be affected by the p-value threshold, as they are calculated using individual p-values; by examining the distribution curves for different p-values, we notice that the influence is barely evident (Figure 22).

G scores for true identifications are quite conserved, while the curve for false identifications increases in skewness but retains the same central value.

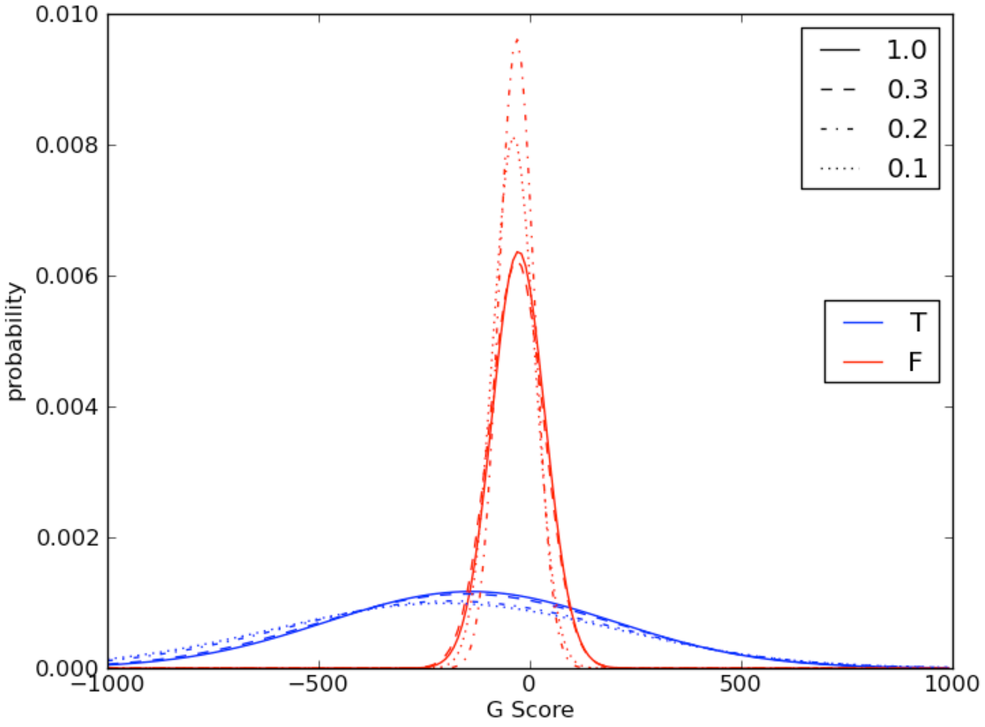


Figure 22 *Distribution of G score at different p-value thresholds*

MMS scores are also affected by the p-value threshold, essentially because "top scoring" identifications are retained. Although a total separation of distributions cannot be observed, the skewness increases for both curves as the as p-value decreases (Figure 23); the direct effect of this is that the area where the curves overlap decreases as well. This is a positive outcome of the analysis, because the overlap is correlated to the amount of false positives and false negatives

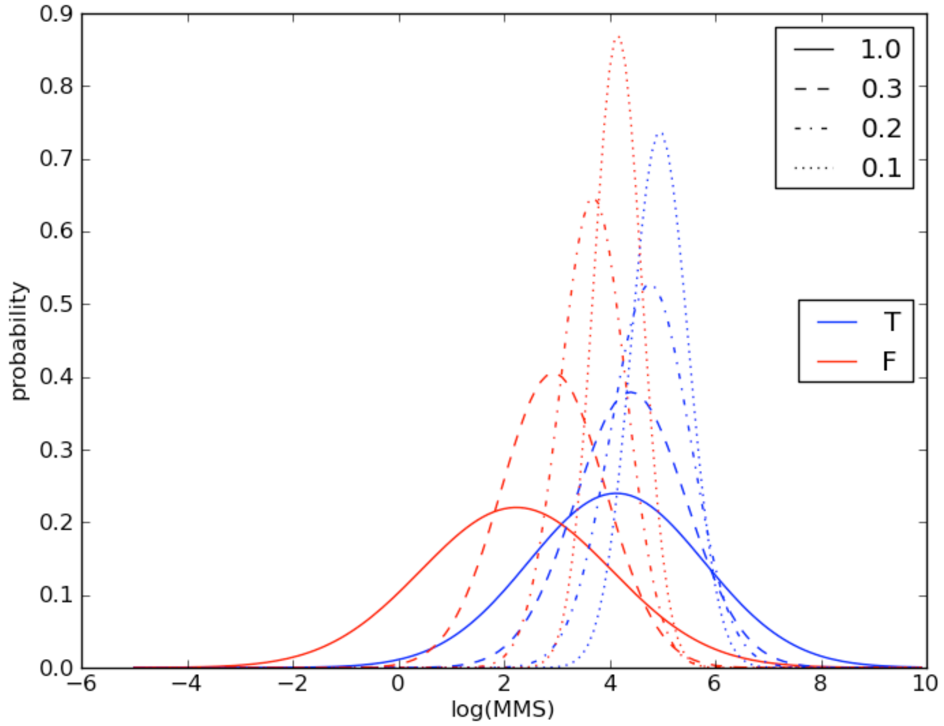


Figure 23 *Distribution of MMS scores at different p-value thresholds*

In order to choose a further combination of G and MMS thresholds when p-value filter is engaged ($p \leq 0.1$), we performed another ROC analysis (Figure 24).

Cutting MMS at the P_{90} or P_{95} does not make a remarkable difference, especially when G is lower than -30; it is worth noting that enabling p-value filter allows for a sensitivity higher than 0.5 also when lower MMS or G thresholds are engaged. Setting MMS threshold to $P_{90}(MMS)$ and G threshold to -30, the sensitivity is above 0.9, at the cost of a low number of interactions:

$MMS \geq 178.25$ & $G \leq -30$			Sensitivity=0.92
	T	**F**	Specificity=0.94
T	*13*	*7*	PPV=0.65
F	*1*	*109*	NPV=0.99

Using the same thresholds on whole results we are able to draw a map of interactions for RNAP II (Figure 25). No valid interactions could be found for seven subunits using these parameters (RPB3, RPB7-12). One cross-link involving RPB5 and RPB2 is a false positive according to the 3D structure.

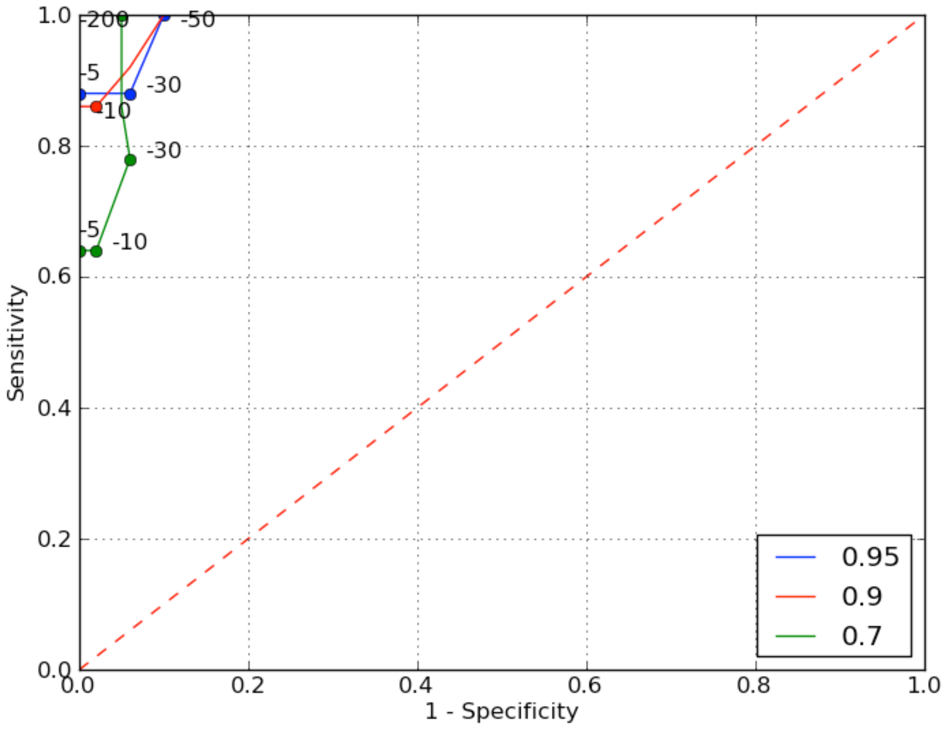


Figure 24 *ROC analysis for different G scores (labeled dots) at varying MMS threshold (colored lines) when p-value filtering is engaged and best matching spectra are retained*

Most of valid interactions are intra-chain cross-links mapped on RPB1 and RPB2; this finding suggests that the number of cross-links that can be mapped to a protein is likely to be proportional with its mass.

With another set of thresholds (no p-value filter, $MMS \geq 80.45$ and $G \leq -70$) we are able to build a better topology map (Figure 26).

Many interaction spotted in the last condition were sub-optimal identifications in previous selections, especially interactions involving RPB5 and RPB6. This suggests that a manual inspection of results may be recommended for borderline identifications.

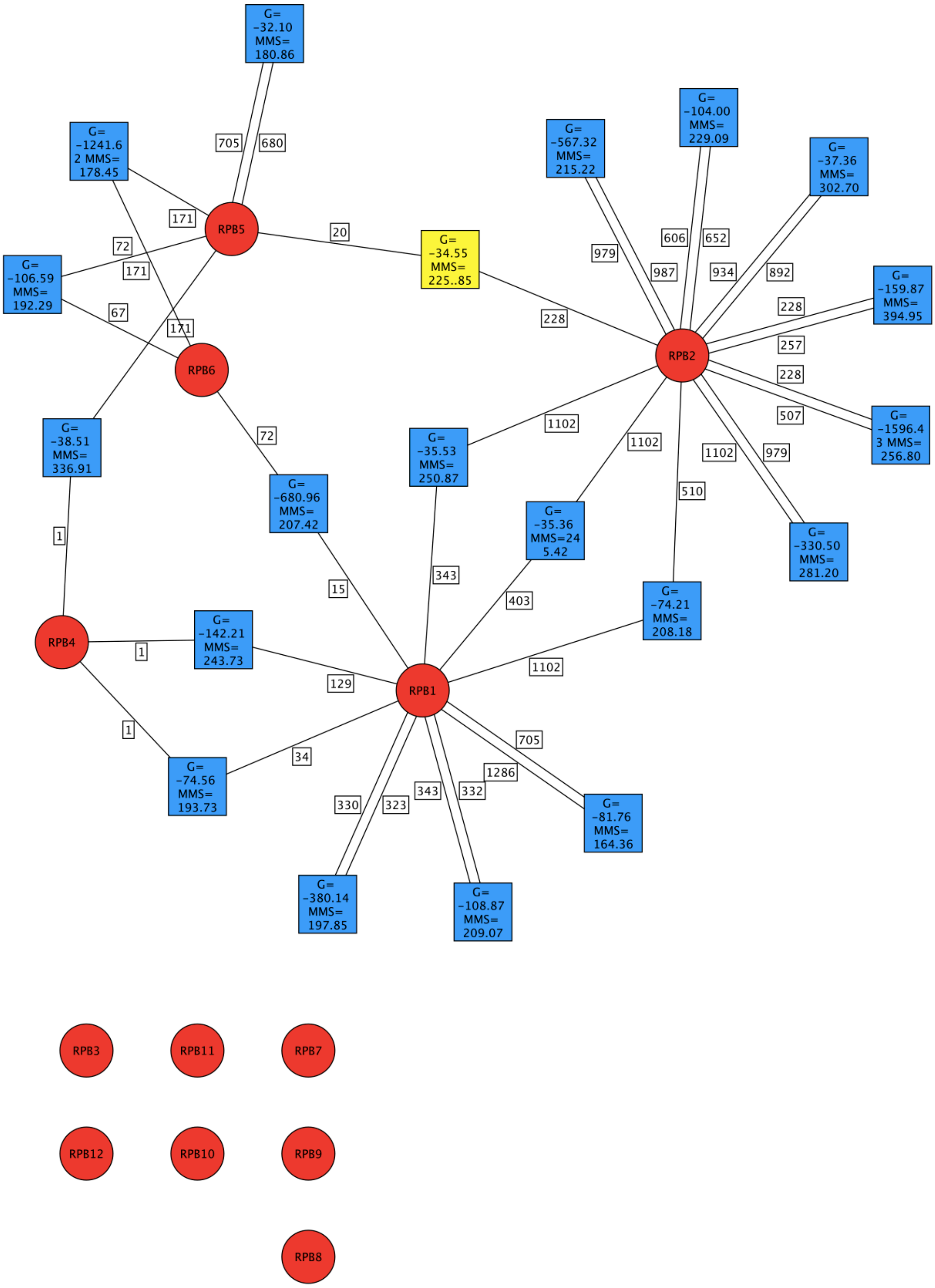


Figure 25 *Topology map of RNAP II interactions using p ≤ 0.1, MMS ≥ 178.25 and G ≤ -30 thresholds. Cross-links that have been found to be false positives have been colored in yellow.*

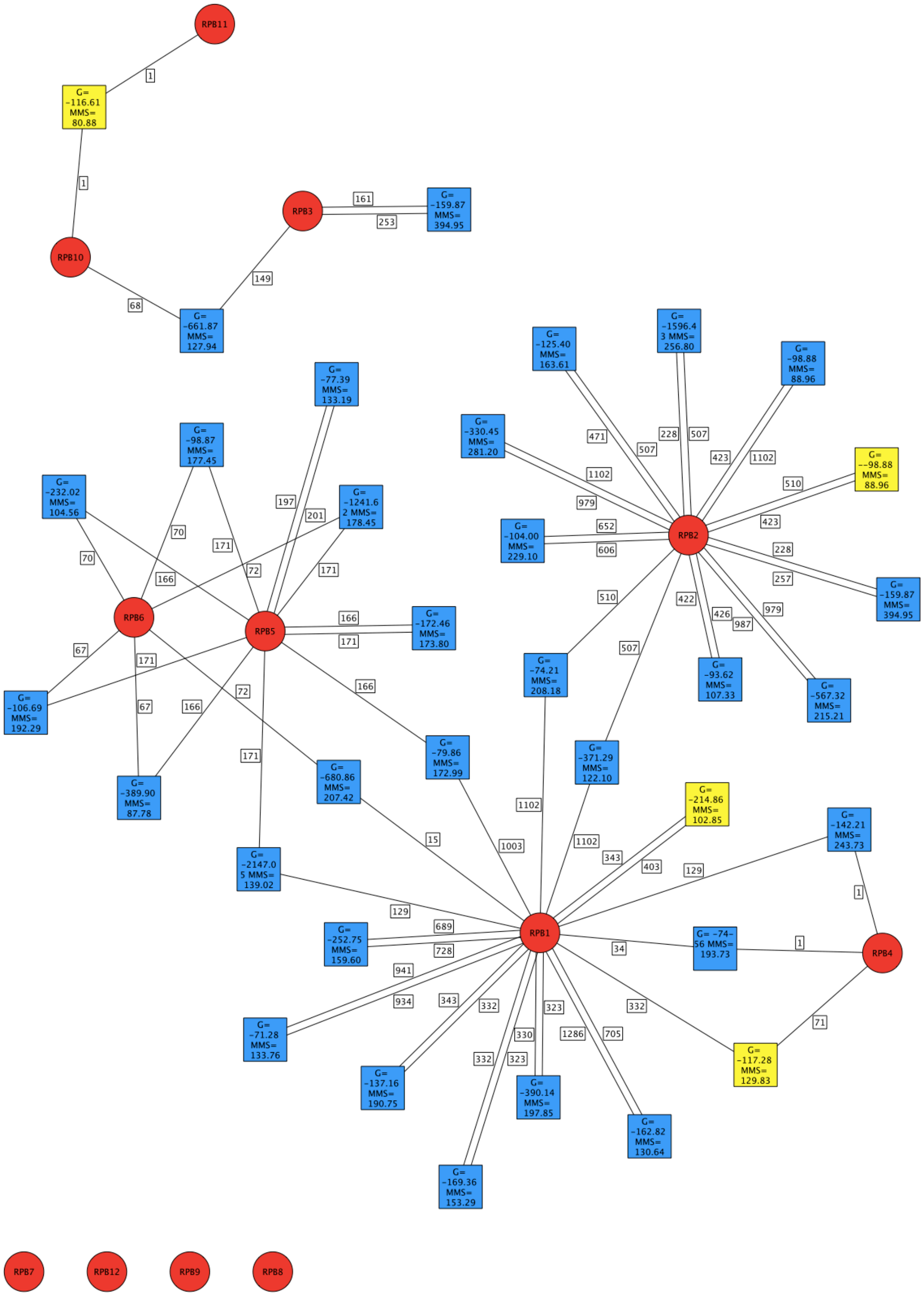


Figure 26 *Topology map of RNAP II using MMS ≥ 80.45 and G ≤ -70.*

RNA Polymerase I and RNA Polymerase III

An initial attempt to apply **nsilk** on unknown complexes has been tested on RNA polymerase I and RNA polymerase III. RNA polymerases I, II and III have related subunit compositions, with twelve homologous or identical core subunits shared by all three enzymes (Werner, Thuriaux and Soutourina 2009). Although the crystal structures of RNAP I and RNAP III have not yet been determined, the information on RNAP II can be used to indirectly validate cross-link analysis on these enzymes (Table 4).

RNAP I	RNAP II	RNAP III
RPA1	RPB1	RPC1
RPA2	RPB2	RPC2
RPAC1	RPB3	RPAC1
RPA14	RPB4	RPC9
RPAB1	RPB5	RPAB1
RPAB2	RPB6	RPAB2
RPA43	RPB7	RPC8
RPAB3	RPB8	RPAB3
RPA12	RPB9	RPC10
RPAB5	RPB10	RPAB5
RPAC2	RPB11	RPAC2
RPAB4	RPB12	RPAB4
RPA34		RPC7
		RPC3
		RPC4
		RPC5
RPA49		RPC6

Table 4 *Summary of subunits correspondence between RNA polymerases I, II and III*

RNAP I is composed of 14 subunits; its molecular weight is approximately 590 kDa. RNAP III is a bit more complex as it is composed by 17 subunits and its molecular weight is approximately 690 kDa.

RNAP I analysis has been performed using a different cross-linker (BS^3). Unfortunately the number of datasets was much lower than for the RNAP II experiment; the direct consequence is a much lower final number interactions (Table 5).

Previous experiments suggest that a *MMS* filter can be chosen above the P_{90} *(MMS)*: $MMS \geq 113$ for RNAP I and $MMS \geq 84$ for RNAP III. *G* score threshold has been set to -30 for both experiments. Unfortunately, given the low number of valid interactions, the topology description is very poor.

	RNAP I	**RNAP III**
N. of datasets	2	1
N. of spectra	10476	1867
N. of cross-links (valid/total)	1/83	5/546
N. of interactions (total)	3921	5111
N. of valid interactions (default param.)	1040	1638
N. of valid interactions (R1)	127	95
N. of valid interactions (R1, p ≤ 0.1)	46	40

Table 5 *Summary counts of datasets and results for RNAP I and RNAP III experiments*

Analysis of RNAP I complex fails to spot interactions for 9 subunits (Figure 27). Also, a limited number of interactions can be spotted for the remaining 5 elements.

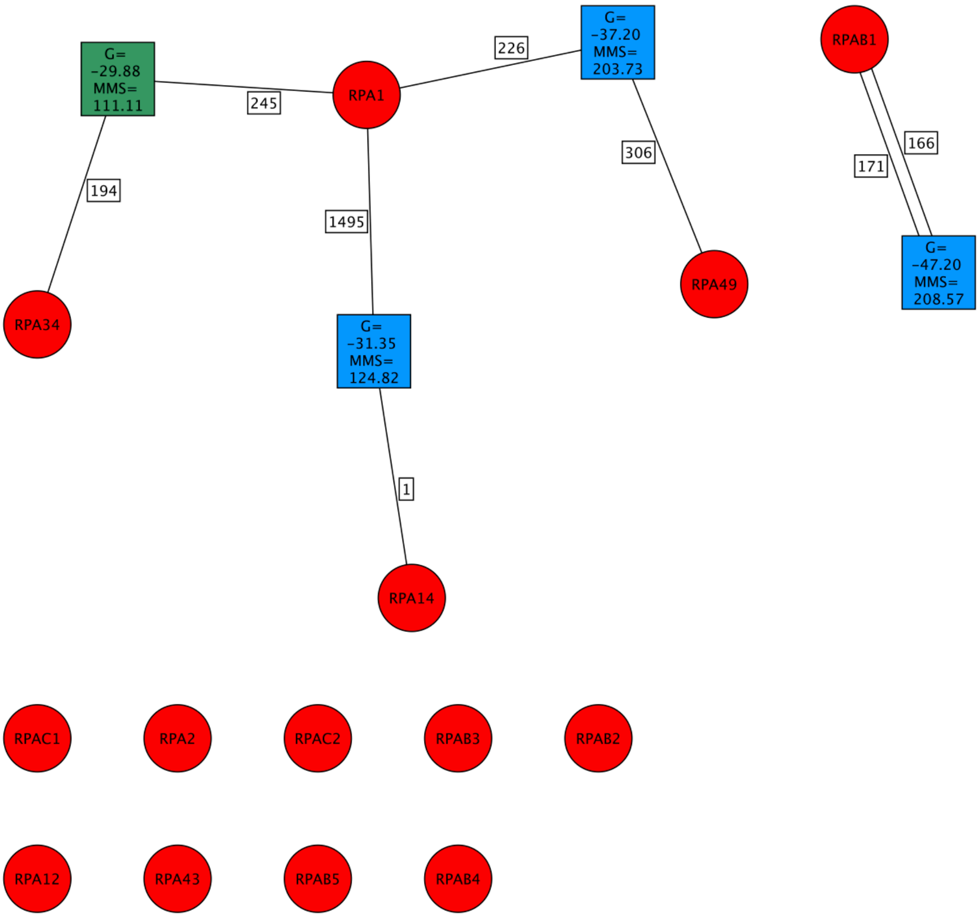


Figure 27 *Topology map of RNAP I complex. Green nodes represent suboptimal interactions found after visual inspection*

According to the subunit similarities, RPA14:1-RPA1:1495 is consistent with the fact that homologous RPB4 and RPB1 interact, although none can be said about the aminoacids involved; the remaining interactions cannot be validated as they involve RPA34 and RPA49: these proteins do not have homologs in the RNAP II complex.

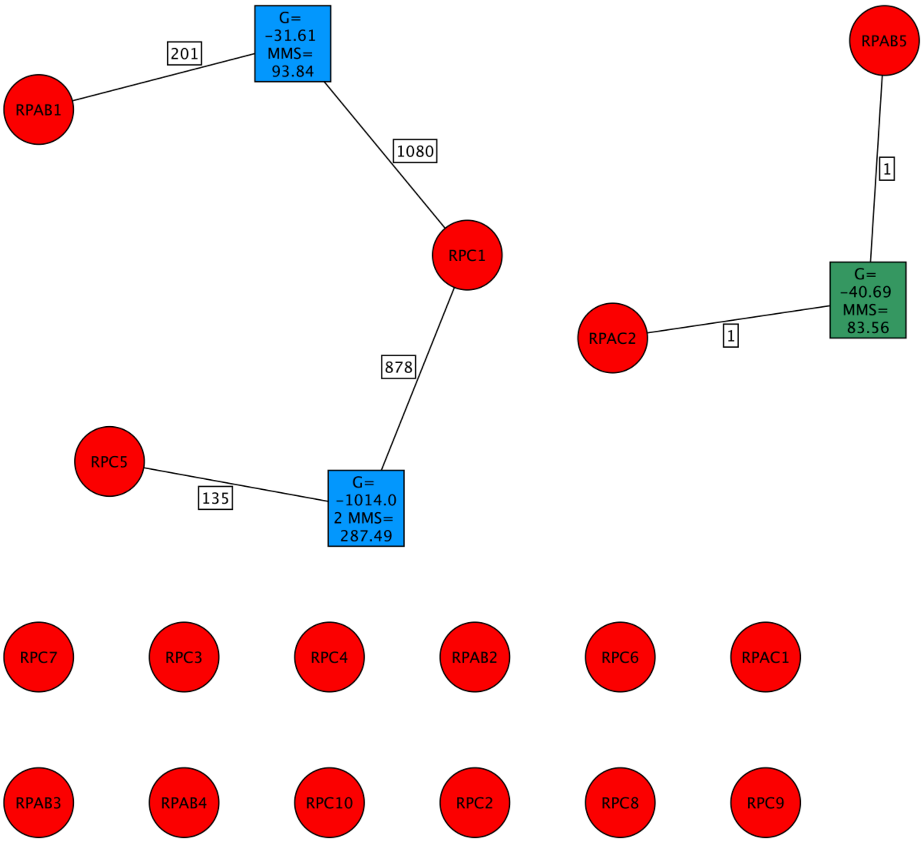


Figure 28 *Topology map of RNAP III*

The number of valid interactions for RNAP III is even lower, 12 subunits out of 17 could not be mapped (Figure 28). RPC1:1080-RPAB1:201 is consistent with RPB1:1003-RPB5:166 on RNAP II. Interestingly, the sub-optimal interaction RPAC2:1-RPAB5:1 is homologous to RPB11:1-RPB10:1 found on RNAP II topology map; unfortunately the latter is known to be a false positive.

The small number of fragmentation spectra available is the most probable cause of these poor results. For previous experiments, the proportion between the number of fragmentation spectra and the possible peptide combinations was clearly higher (1:10 for RNAP II and 1:2 for GINS); the ratio in RNAP I dataset is about 1:50, while it is about 1:550 in RNAP III dataset. We therefore conclude that a larger dataset will be required to evaluate the optimal number of fragmentation spectra.

KMN Network

Kinetochores are the protein structures on chromosomes where the spindle fibers attach during cell division to pull the chromosomes apart (Albertson and Thomson 1993). The kinetochore contains two regions: an inner kinetochore, tightly associated with the centromeric DNA, and an outer kinetochore, which contains components required for microtubule attachment.

The KMN network complex has emerged as a crucial component of the machinery specialized in microtubule binding (Santaguida and Musacchio 2009). The KMN network is a 10-subunit assembly gathering three distinct subcomplexes, known as Knl1, Mis12 and Ndc80. The Ndc80 complex has been described in the first part of this section. Human Knl1 is a 2342 aminoacids protein, its molecular weight is approximately 265 kDa; it is required for chromosome segregation and cell viability (Cheeseman and Desai 2008). Mis12 complex is made of 4 subunits (Mis12, Nsl1, Pmf1, Dsn1), its molecular weight is approximately 120 kDa; recent works show that it is organized in a chain-like structure with Nsl1 at one end; Nsl1 interacts with the globular domain of Spc24/Spc25 (part of the Ndc80 complex) and the N-terminal domain of Knl1 (Petrovic, *et al.* 2010).

Two sets of analyses have been made available for the KMN network (Table 6), addressing two different recombinant sub-complexes. These have been analyzed separately. The final interactions have been produced using best match and $p \leq 0.1$ filters as **nsilk** parameters. Interactions were filtered with the same approach used for RNAP I and RNAP III.

	KM	**KMN**
Proteins included	Dsn1, Nsl1, Mis12, Pmf1, Knl1^{C-212}	Dsn1, Nsl1, Mis12, Pmf1, Knl1^{C-212}, Ndc80^{bonsai}
N. of datasets	*7*	*6*
N. of spectra	*21175*	*14953*
N. of cross-links (valid/total)	*21/566*	*3/140*
N. of interactions (total)	*442*	*724*
N. of valid interactions	*211*	*298*
N. of valid interactions (R1)	*159*	*173*
N. of valid interactions (R1, p ≤ 0.1)	*24*	*27*

Table 6 *Summary counts of datasets and results for three different experiments on the KMN network. Knl1^{C-212} is the C-terminal domain of Knl1, Ndc80^{bonsai} is the engineered version of the Ndc80 complex discussed above.*

As both analyses involved a common core of five subunits (Table 6), it was interesting to evaluate the overlap between the results (Figure 29). The Venn diagram has been elaborated without any filter on **nsilk** results for MMS or G scores. 44 interactions are common between the two experiments if p-value filter is not engaged, but only 10 (about 40% of both datasets) can be identified if cross-links are filtered for $p \leq 0.1$. As there might have been different performances of subsequent step in sample preparation (*e.g.* cross-link reaction or MS acquisition), we want to use the mutual information to infer reasonable filters. Of course, we expect false positives may be common among the two experiments. The biochemical analysis of the KMN network (Petrovic, *et al.* 2010) could be used to exclude potential false positives.

The correlation coefficient for 44 common MMS scores (ϱ_{MMS}) is 0.94, while the correlation coefficient for G scores (ϱ_G) is 0.73; if we consider 10 common interactions when p-value filter is applied, both the coefficients decrease (ϱ_{MMS} = 0.93, ϱ_G = 0.28).

The previously validated experiment suggests that the p-value filter may be useful when we have to deal with larger proteins, when the number of possible peptide combinations is high (i.e in the order of 105 combinations). For smaller protein complexes, we may rely on results after the “best match” rule has been applied.

We start by filtering results using the usual $P_{90}(MMS)$ rule. RNAP II data and GINS data suggest that the specificity should be high enough (close to 70%)

using this thresholding rule. *MMS* threshold for KM dataset is 37.6, and 17 interactions are above this value. *MMS* threshold for KMN dataset is 53.6, and 19 interactions are above this value; 6 interactions out of the common set of 44 are above the thresholds in both the experiments. *G* scores of these 6 interactions are lower than -30 in both experiments, except for Nsl1:262-Nls1:276 which has *G=-24.36* in KM dataset and *G=-19.91* in KMN dataset. All 6 interactions are consistent with known KMN model: 3 interactions involve aminoacids that are closely spaced on the same chain (Pmf1:116-Pmf1:119, Nsl1:142-Nsl1-149 and Nsl1:262-Nsl1-276), 2 interactions involve the C-terminal domains of Dsn1 and Nsl1 (Dsn1:322-Nsl1:259 and Dsn1:322-Nsl-262), one interaction involves the C-terminal domains of Dsn1 and Mis12 (Dsn1:248-Mis12:129); hence, we may consider this small set to assign an upper bound to *G* scores.

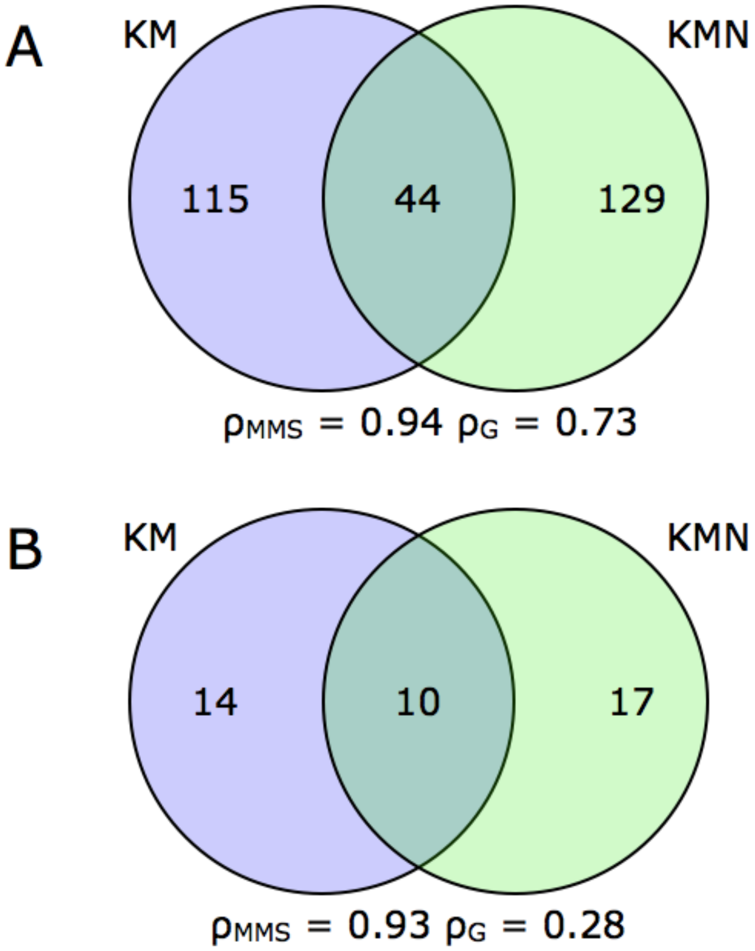

A
KM
KMN
115
44
129
ρ_{MMS} = 0.94 ρ_G = 0.73
B
KM
KMN
14
10
17
ρ_{MMS} = 0.93 ρ_G = 0.28

Figure 29 *Venn diagram of the results for two different analyses on the KMN network. When no p-value filter is applied on cross-links (A), the correlation coefficient for MMS and G scores of common interactions are remarkably high; the number of interactions and the correlation coefficient values decrease when cross-links are filtered for p ≤ 0.1 (B).*

Applying G thresholds on KM and KMN data, we retain 11 and 16 interactions respectively. These interactions can be used to build topology maps (Figure 30, Figure 31, Figure 32). Both in the KM and KMN experiments, no valid interaction between Pmf1 and other subunits could be spotted. In both experiments, the majority of cross-links that do not map on the same chain can be detected between C-terminal domains of Dsn1 and Nsl1. Mis12 C-terminal domain seems to interact with this region, too.

A single Knl1 interaction with the rest of the complex has been identified in the KM experiment, and it is consistent with the biochemical data.

The KMN experiment contains an internal control to evaluate the goodness of the results. Interactions between the artificial Hec1_Spc25 and Nuf2_Spc24 chains of the Ndc80^{bonsai} construct have been described before. The number of cross-links involving these chains is lower than in previous experiments. Nevertheless the interactions found are consistent with known ones, with the big exception of Nuf2_Spc24:1-Hec1_Spc25:302 that is one of the top-scoring cross-links.

Notably, a relevant interaction between the globular domain of Spc25 and the C-terminal domain of Nsl1 (Hec1_Spc25:372-Nsl1:276) has been found in KMN experiment.

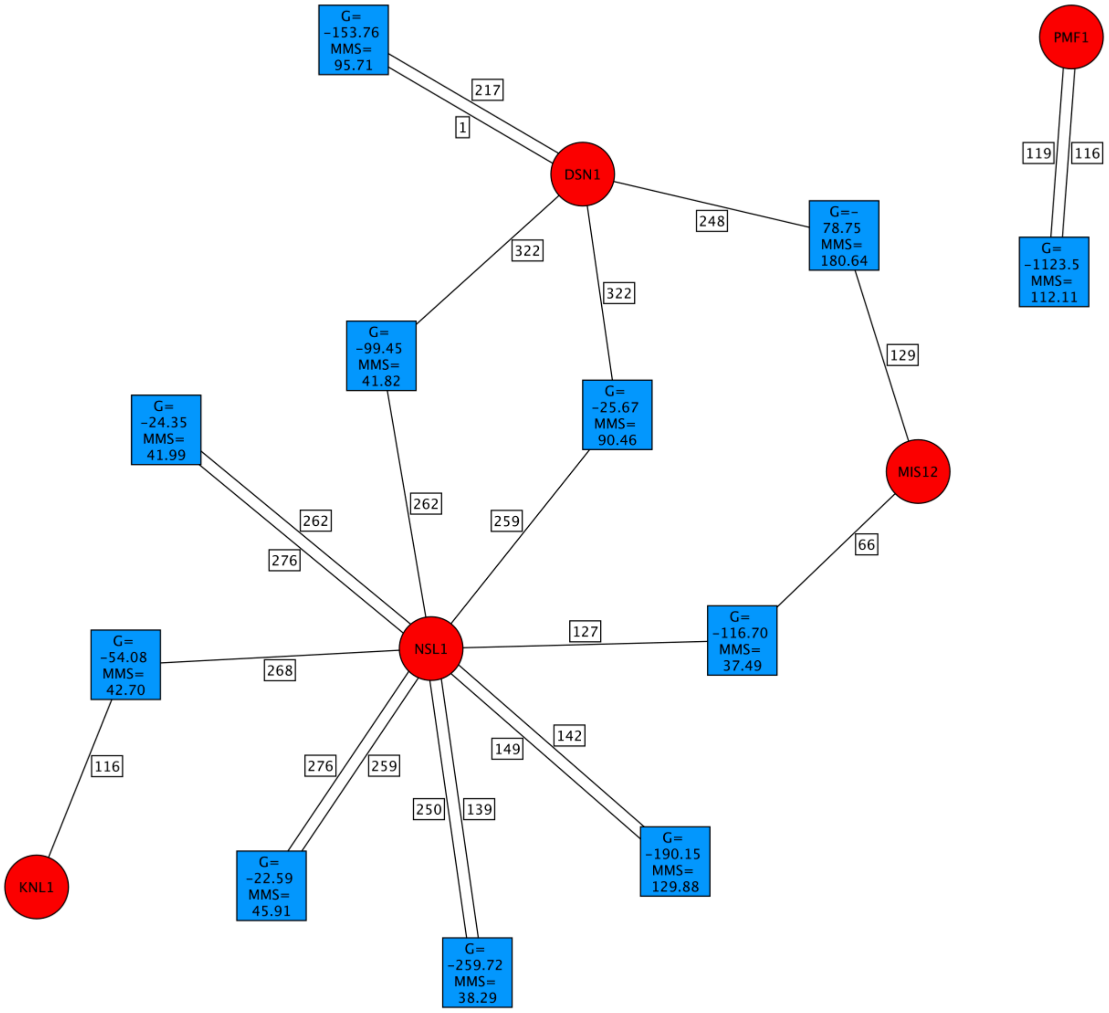


Figure 30 *Topology map of KM experiment*

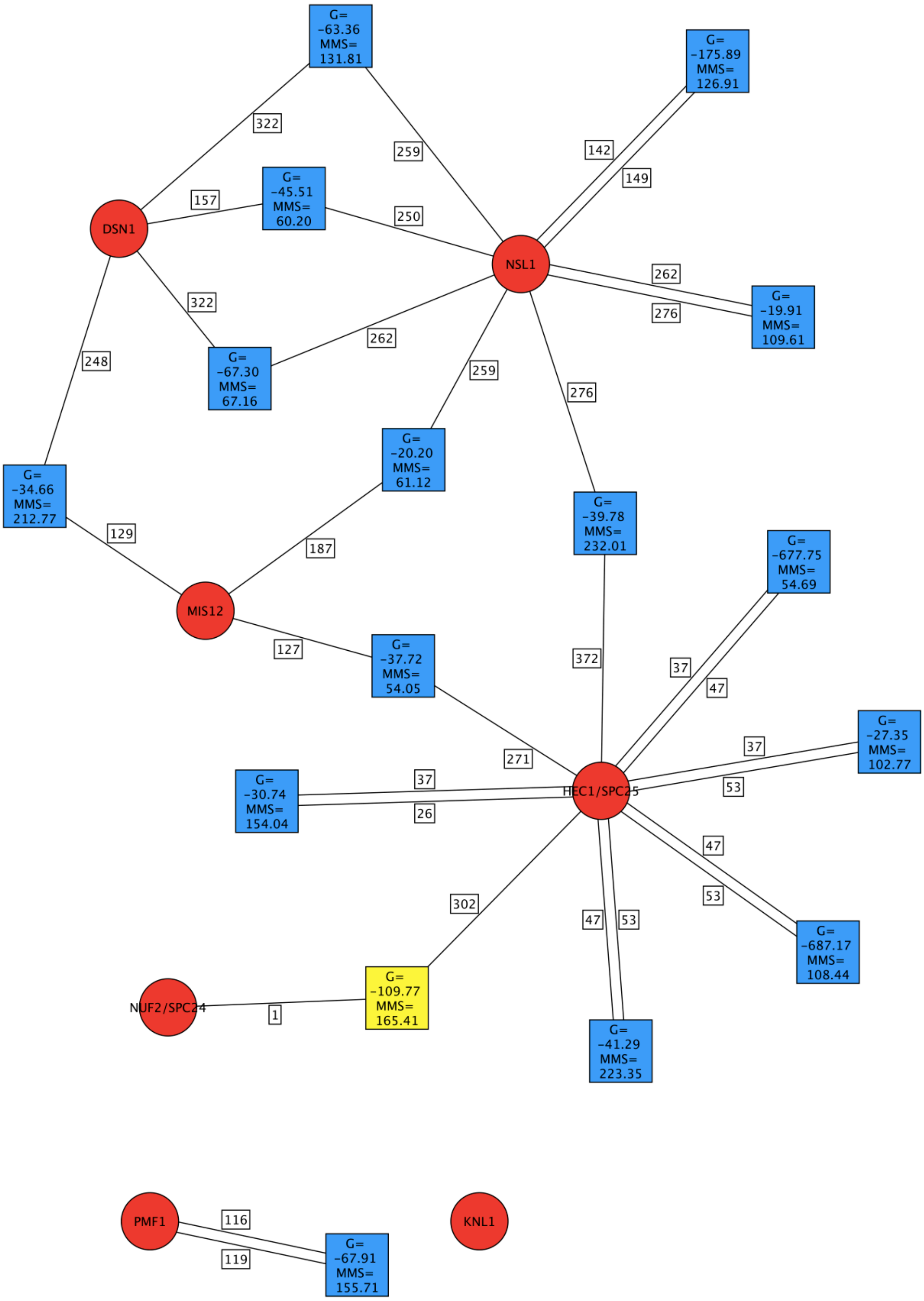


Figure 31 *Topology map of KMN experiment*

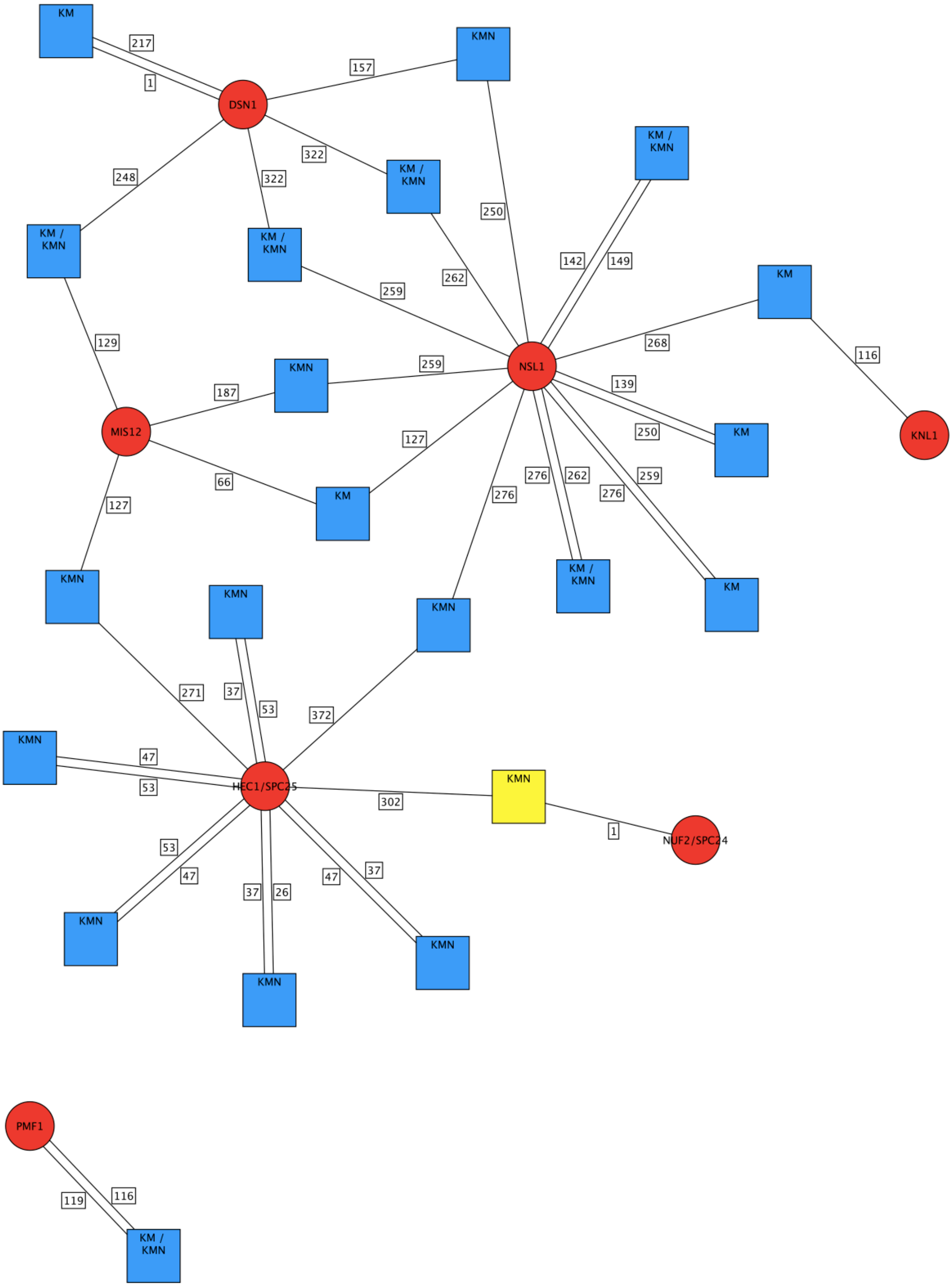


Figure 32 *Topology map obtained by the union of KM and KMN experiments. Each blue node contains the name of the experiment in which the interaction could be found.*

Performance considerations

nsilk scales *O(n · log(n))* with the number of peptides that will be used to build and evaluate a cross-link. The bottleneck of the entire process is the generation of peptide combinations and their match against fragmentation spectra. A major impact on the number of peptides is the tolerance for peptide mass match. For this reason, usage of **nsilk** is indicated for very precise mass spectrometers (*i.e.* FT-ICR, Orbitrap…).

Running times vary in a wide range: 30 seconds for the GINS dataset to 24 minutes for the RNAP III dataset.

Memory usage does not depend much on the number of cross-links allowed (Figure 33): doubling the number of valid cross-links does not affect global memory usage but the running time instead. Most of the memory will be allocated at the beginning of the process, in order to store information about spectra.

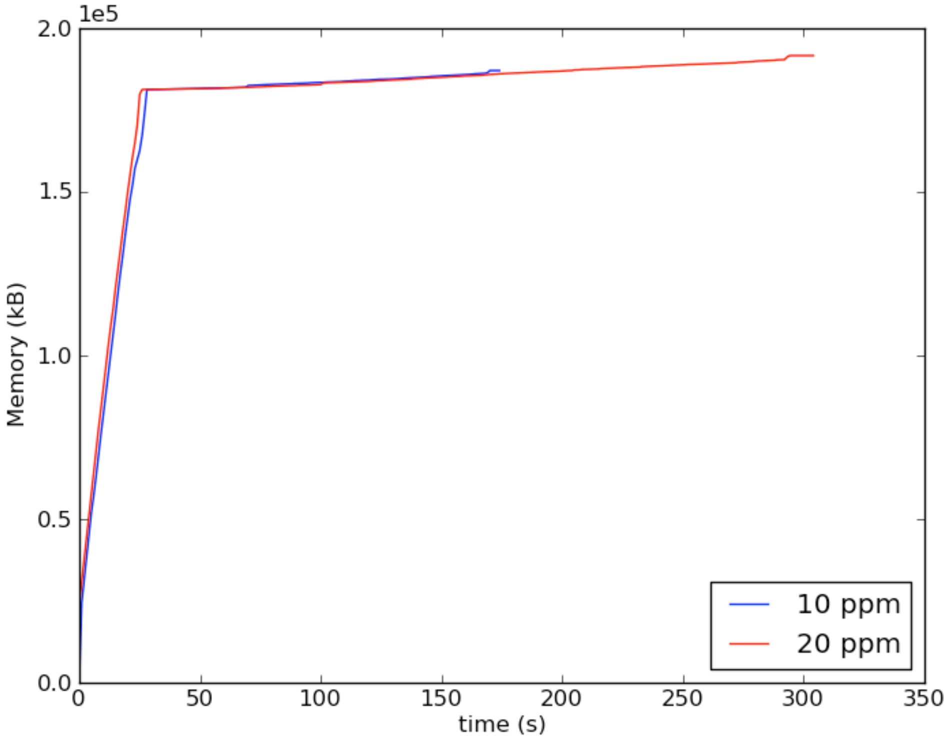


Figure 33 *Memory usage for KMN experiment at two different MS tolerances.*

A mid-sized experiment, such as the RNAP II one, will need approximately 400 Mb RAM for a 20-minute run.

The most time-consuming step of the analysis is the evaluation of single cross-links. This process could be in principle parallelized, as each cross-link is independent from the others. Unfortunately, the python implementation doesn’t allow for an efficient threading environment (http://docs.python.org/c-api/init.html - thread-state-and-the-global-interpreter-lock).

Conclusions and perspectives

Identification of cross-linked peptides is everything but a simple task. The number of possible cross-links grows exponentially with the number and the length of protein sequences. We have developed a lightweight application that identifies cross-links from a known set of proteins in reasonable times. Also, most of the analysis could be performed on a normal desktop computer.

We analyzed complexes in a wide range of molecular size and number of subunits. **nsilk** showed a good specificity in spotting protein-protein interactions, although many true cross-links were poorly scored and tagged as false negatives. We speculate this is due to the fact that we use only b and y ion series to match fragmentation spectra. Regretfully, we do not have a model that includes other frequent ion series, such as a ions, yet.

It is essential for our approach that the biochemical step of the experiment performs well either in terms of cross-link yield or number of spectra collected. Assuming the cross-linking and protein purification efficiency of all our experiments were the same, we had fewer identifications when the number of spectra was low (RNAP I and RNAP III). On the basis of our experience, we estimate the number of fragmentation spectra should be in the range of 1:2-1:20 to the number of all possible cross-links. Also, if the number of spectra is considerably higher than the number of cross-links (*i.e.* 2:1 ratio or higher), the G score statistic may fail as the null probabilities that appear as denominators tend toward zero.

As pointed in the introduction, **nsilk** is not the only software available for CXMS analysis. Unfortunately it is hard to perform an exhaustive comparison benchmark, also because many tools have been published but are no more available. Compared to the published approaches, **nsilk** implements a unique feature: it groups single cross-link identifications into interactions. **nsilk** can take advantage of multiple identifications to increase the confidence on each interaction; ideally, this approach is not different from common MS/MS search algorithms that aggregate peptides increasing likelihood of protein identification. Recent suites like xQuest (Rinner, *et al.* 2008) (http://www.xquest.org) and xComb (Panchaud, *et al.* 2010) (http://phenyx.proteomics.washington.edu/

CXDB/index.cgi) are able to efficiently identify hundreds of cross-links, giving no global picture of the interacting interfaces.

Another strong point of **nsilk** is its ability to process multiple spectra files in a single analysis. Obviously, the same files can be concatenated into a single one, but our approach gives higher modularity and simplifies the user experience.

This said, **nsilk** is not ready for proteome-wide analysis. In order to scale up and analyze an whole proteome, two issues should be solved. First, we need a smart way to compile a cross-link list to evaluate; **nsilk** uses a brute force approach that runs in reasonable times as a single evaluation takes very little time. Second, **nsilk** could take advantage of an implementation in a compiled programming language such as C or C++, although recent python implementations introduced some I/O and threading optimizations; besides the higher performance of such languages, an efficient multi-threaded implementation would be possible in that way.

nsilk source is available for download at http://code.google.com/p/nsilk. Latest **nsilk** versions are available under BSD license from the subversion repository https://nsilk.googlecode.com/svn/trunk, as well as old **silk** versions.

Appendix A: **nsilk** command line options

nsilk comes with a number of options to refine search and filter results. Current **nsilk** version is 1.0.5.6.

```
Usage: nsilk [options] <spectra file(s)>
```

The following section will describe **nsilk** options.

```
--version
```

Shows program's version number and exit

```
-h
--help
```

Shows the list of available options and exit

```
-f STRING
--fasta=STRING
```

Specifies the name of the FASTA formatted file containing protein sequences under investigation. This parameter is mandatory. Currently there’s no limit to the maximum number of proteins, except for computational power of the underlying hardware.

```
-e ENZYME
--enzyme=ENZYME
```

Specifies the enzyme used for in silico digestion. **nsilk** comes with a configuration file containing enzyme defitions (enzymes.ini).

```
-c INTEGER
--missed=INTEGER
```

Specifies the number of missed cleavages for in silico digestion. This parameter will affect the number of final cross-links as well as the time of processing.

```
-l INTEGER
--minlen=INTEGER
```

Specifies the minimum length for peptides. By default this parameter is set to 5, as shorter peptides may be involved in false identifications.

`-L INTEGER`

`--maxlen=INTEGER`

Specifies the maximum length for peptides. There’s no limit to this value, although longer peptides have less chances to be captured by the mass spectrometer.

`-1 FLOAT`

`--MS=FLOAT`

Specifies the tolerance for precursor mass match. The higher this value, the less specific will be the search

`-2 FLOAT`

`--MSMS=FLOAT`

Specifies the tolerance for ion match. This value will affect the number of b and y ions that will be matched when calcuating the match score.

`-U STRING, --MSu=STRING`

`-u STRING, --MSMSu=STRING`

These options can be used to select between dalton (Da) or part per million (ppm) as tolerance unit.

`-F STRING`

`--fixmod=STRING`

Specifies the list of fixed modification. More modifications can be passed separating them with commas. Single modifications should be included between single or double quotes to escape spaces. Default value is 'carbamidomethyl C’. **nsilk** comes with a configuration file (modifications.ini) containing a number of modifications

`-V STRING`

`--varmod=STRING`

Specifies the list of variable modifications. Again, the list is comma-separated and single values should be quoted. Default value is 'oxidation of M'.

`-M INTEGER`

`--maxmod=INTEGER`

nsilk doesn't calculate all the possible modification. Every peptide will contain at most this value of variable modifications. This heuristic will reduce the total running time. Default value is 2.

`-X STRING`

`--xlink=STRING`

As for modifications, cross-linkers should be specified as quoted and comma separated list. A configuration file (xlinkers.ini) contains the list of available cross-linkers. Default value is 'BS2GD0','BS2GD4'

`-R INTEGER`

`--best=INTEGER`

A single spectrum can be matched against multiple cross-links. **nsilk** allows the control of this behavior by setting the maximum number of cross-links to retain per spectrum. Each match is ranked according to its score. Set this value to 0 to keep all matching spectra. Set this to less than 3 when dealing with big complexes

`-P FLOAT`

`--pvalue=FLOAT`

Specify a p-value filter before cross-links are aggregated to events. Set this value to 1 to keep all matches. Set this to low values when dealing with huge complexes.

`-S FLOAT`

`--mscore=FLOAT`

Specify a minum score for matching cross-links. The effect of this filter has never been evaluated.

`-n INTEGER`

`--intlen=INTEGER`

Each interaction is composed by at least one cross-link. In our experience, such interactions are likely to be false positive. We suggest to set this value to at least 2.

`-T INTEGER, --topions=INTEGER`

`-w FLOAT, --topwin=FLOAT`

As **nsilk** does ion matching, it only consider a subset of ion within each spectrum These options control the behavior of ion selection. A spectrum is windowed according to the “topwin” value, in each window a “topions” number of most intense ions is retained. This behavior is similar to OMSSA. Default is to select 12 most intense ions every 30 Da.

`-m INTEGER`

`--mmatch=INTEGER`

Specify the minimum number of ions each cross-link must match with a spectrum.

`--save=STRING`

`--load=STRING`

Use these options to save and load **nsilk** temporary objects. These options are recommended when testing multiple filters. By specifying a filename prefix, the list of spectra, the list of proteins and the list of all peptide pairs will be saved into specific files. These can be then loaded for a second analysis. Note that changing cleavage conditions require a new saving.

Appendix B: nsilk output

nsilk outputs a results as text stream file `stdout` and logs the processes on `stderr`. The output is composed by a list of run parameters and a list of interactions found, ordered by their *MMS* in descending order (i.e. most relevant interactions first). Each interaction contains a list of the single cross-links forming it.

The run parameters are listed at the beginning, each line starting with a hash ("#"):

```
# Spectra files:
heavy_1_f090703_002.msm,heavy_1_f090703_003.msm,heavy_1_f090703_007.msm,heavy_1_f090703_
008.msm,heavy_1_f090703_009.msm,heavy
_1_f090703_010.msm,heavy_1_f090703_011.msm,light_1_f090703_002.msm,light_1_f090703_003.m
sm,light_1_f090703_007.msm,light_1_f090703_008.msm,lig
ht_1_f090703_009.msm,light_1_f090703_010.msm,light_1_f090703_011.msm
# Sequence file: KM.fasta
# Enzyme: Trypsin
# Missed cleavages: 2
# Min. peptide length: 5
# Max. peptide length: 20
# MS Tolerance: 10.0 ppm
# MS/MS Tolerance: 0.5 Da
# Fixed modifications: carbamidomethyl C
# Variable modifications: oxidation of M
# Max. number of variable modifications: 2
# X-linkers: BS2GD0,BS2GD4
# Number of best spectra to retain: 1
# p-value filter: 1.0
# Match score filter: 0.0
# Min. interaction length: 2
# Number of top matching ions: 12 every 30.0 Da
# Min. number of matching ions per peptide: 2
# Save numpy matrices: None
# Load numpy matrices: LessTolerant
```

Interactions are numbered with a progressive number; aminoacid positions on the proteins involved in the cross-link are reported as "Coordinates". Three parameters are then reported: the Median Match Score (*MMS*), the *G* score and the number of single matches found:

```
[ Interaction 1 ]
Coordinates: DSN1:248 MIS12:129
Median Match Score: 180.6381
G-Score: -78.7511
Number of matches: 4
```

Single matches (“xlink”) are numbered with progressive numbers, starting from 1 for each interaction. Peptides forming the match are reported in the form `PEPTIDE [VARIABLE MODIFICATIONS]`. A number of parameters are then reported: the cross-linker name; the *p*-value, the match score, the match rank (i.e. the rank among the number of matches that can be assigned to the same spectrum); the spectrum description; the precursor charge (read from the spectrum); the *m/z* value and the error on precursor mass:

```
[ xlink 1 ]
        Peptide 1: GSTEAKITEVK [ ]
        Peptide 2: YKTELCTK [ ]
        X-linker: BS2GD0
        p-value: 0.01806
        Match Score: 181.65367
        Match Rank: 1 of 1
        Elution from: 55.97 to 57.27 period: f090703_009.raw experiment: 1 cycles: 1
precIntensity: 21921952.8 FinneganScanNumber: 5302
        Charge: 3       m/z: 767.395    Error (ppm): 1.48
```

nsilk calculates and outputs a list of *b* and *y* ions that could have been matched in the spectrum. This list does include ions up to the supposed cross-linked aminoacid. If any ion is found in the real spectrum, the observed mass and intensity are reported between square brackets:

```
GSTEAKITEVK                         YKTELCTK
b1 G 58.029 [ ]                     b1 Y 164.071 [ ]
b2 S 145.061 [ ]                    y6 T 751.364 [ ]
b3 T 246.109 [ 246.215, 984.470 ]   y5 E 650.316 [ 650.412, 1506.930 ]
b4 E 375.152 [ 375.313, 776.140 ]   y4 L 521.274 [ 521.320, 4535.940 ]
b5 A 446.189 [ 446.335, 49.830 ]    y3 C 408.190 [ 408.204, 1889.620 ]
y5 I 589.356 [ 589.536, 2767.800 ]  y2 T 248.160 [ 248.227, 316.230 ]
y4 T 476.271 [ 476.424, 2573.720 ]  y1 K 147.113 [ ]
y3 E 375.224 [ 375.313, 776.140 ]
y2 V 246.181 [ 246.215, 984.470 ]
y1 K 147.113 [ ]
```

Appendix C: configuration files

nsilk comes with 5 configuration files. These are installed into a specific directory and can be accessed by the software setting the environmental variable `$NSILK_CONFIG` to the correct path.

The syntax of configuration file is the “INI format” (http://www.cloanto.com/specs/ini/).

aminoacids.ini

Contains information about aminoacids. Each field is defined like the following example:

```
[A]                    # Aminoacid symbol
abbr=Ala               # Three-letter code
formula=C3H5NO         # Brute chemical formula
name=Alanine      # Full name
amass=71.0788          # Average mass
mmass=71.03711         # Monoisotopic mass
gravy=1.800       # Grand average of hydropathicity
```

elements.ini

Contains information abouth chemical elements.

```
[H]                    # Symbol
amass=1.00794          # Average mass
mmass=1.0078250321     # Monoisotopic mass
```

enzymes.ini

Contains information about enzymes.

```
[Trypsin]              # Enzyme name
c-term=OH              # C-terminal group after cleavage
expr=[KR][^P]          # Cleavage definition
n-term=H               # N-terminal group after cleavage
```

Cleavage definition must be specified as a POSIX basic regular expression

modifications.ini

Defines chemical modifications for aminoacids, peptide termini or proteins

```
[carbamidomethyl C]    # Modification name
mmass=57.02        # Monoisotopic mass
amass=0            # Average mass
residues=C         # Affected residues
modtype=aa         # Modification type
```

Modification type can be one of “aa”, “np” or “cp”. “aa” modification affect single aminoacids, “np” or “cp” modifications affect peptide N-terminal or C-terminal residue respectively.

xlinkers.ini

Defines cross-linkers properties.

```
[BS2GD0]               # Cross-linker name
mmass=96.02114         # Monoisotopic mass
amass=96.08562         # Average mass
links=K,np         # Cross-linked residues
```

Cross-linked residues must be specified as comma separated values in the “links” section. “np” denotes the N-terminal aminoacid of the intact protein, “cp” can be specified for C-terminal residue.

Bibliography

Aebersold R and Goodlett DR. “Mass spectrometry in proteomics.” Chem. Rev (ACS Publications) 101, no. 2 (2001): 269-296.

Aebersold R and Mann M. “Mass spectrometry-based proteomics.” Nature 422, no. 6928 (Mar 2003): 198-207.

Albertson DG and Thomson JN. “Segregation of holocentric chromosomes at meiosis in the nematode Caenorhabditis elegans.” Chromosome Research 1 (1993): 15-26.

Aloy P and Russell RB. “Ten thousand interactions for the molecular biologist.” Nature biotechnology 22, no. 10 (Oct 2004): 1317-21.

Blow N. “Systems biology: Untangling the protein web.” Nature 460, no. 7253 (Jul 2009): 415-8.

Brueckner F and Cramer P. “Structural basis of transcription inhibition by alpha-amanitin and implications for RNA polymerase II translocation.” Nature structural & molecular biology 15, no. 8 (Aug 2008): 811-8.

Carty SM, Goldstrohm AC, Suñé C, Garcia-Blanco MA and Greenleaf AL. “Protein-interaction modules that organize nuclear function: FF domains of CA150 bind the phosphoCTD of RNA polymerase II.” Proceedings of the National Academy of Sciences of the United States of America 97, no. 16 (Aug 2000): 9015-20.

Cheeseman IM and Desai A. “Molecular architecture of the kinetochore-microtubule interface.” Nature reviews Molecular cell biology 9, no. 1 (Jan 2008): 33-46.

Cho H, Orphanides G, Sun X, Yang XJ, Ogryzko V, Lees E, Nakatani Y and Reinberg D. “A human RNA polymerase II complex containing factors that modify chromatin structure.” Molecular and cellular biology 18, no. 9 (Sep 1998): 5355-63.

Ciferri C, Musacchio A and Petrovic A. “The Ndc80 complex: hub of kinetochore activity.” FEBS Letters (Federation of European Biochemical Societies) 581, no. 15 (2007): 2862-2869.

Ciferri C, Pasqualato S, Screpanti E, Varetti G, Santaguida S, Dos Reis G, Maiolica A, Polka J, De Luca JG, De Wulf P, Salek M, Rappsilber J, Moores CA, Salmon ED and Musacchio A. “Implications for kinetochore-microtubule attachment from the structure of an engineered Ndc80 complex." Cell 133, no. 3 (May 2008): 427-39.

Cittaro D, Borsotti D, Maiolica A, Argenzio E and Rappsilber J. "Peptide identification using vectors of small fragment ions." Journal of proteome research 4, no. 3 (Jan 2005): 1006-11.

Cramer P, Bushnell D and Kornberg R. "Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution." Science (New York, NY) 292, no. 5523 (2001): 1863.

Cusick ME, Klitgord N, Vidal M and Hill DE. "Interactome: gateway into systems biology." Human molecular genetics 14 Spec No. 2 (Oct 2005): R171-81.

De Falco M, Ferrari E, De Felice M, Rossi M, Hübscher U and Pisani F. "The human GINS complex binds to and specifically stimulates human DNA polymerase α-primase." EMBO reports 8, no. 1 (2007): 99.

De Las Rivas J and Fontanillo C. "Protein-protein interactions essentials: key concepts to building and analyzing interactome networks." PLoS computational biology 6, no. 6 (Jan 2010): e1000807.

El-Shafey A, Tolic N, Young M, Sale K, Smith R and Kery V. ""Zero-length" cross-linking in solid state as an approach for analysis of protein-protein interactions." Protein Science (John Wiley & Sons) 15, no. 3 (2006): 429-440.

Fields S and Song O. "A novel genetic system to detect protein-protein interactions." Nature 340, no. 6230 (Jul 1989): 245-6.

Gao Q, Xue S, Doneanu C, Shaffer S, Goodlett D and Nelson S. "Pro-CrossLink. Software tool for protein cross-linking and mass spectrometry." Anal. Chem (ACS Publications) 78, no. 7 (2006): 2145-2149.

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W and Bryant SH. "Open mass spectrometry search algorithm." Journal of proteome research 3, no. 5 (Jan 2004): 958-64.

Guerrero C, Tagwerker C, Kaiser P and Huang L. "An integrated mass spectrometry-based proteomic approach: quantitative analysis of tandem affinity-purified in vivo cross-linked protein complexes (QTAX) to decipher the 26 S proteasome-interacting network." Molecular & cellular proteomics : MCP 5, no. 2 (Feb 2006): 366-78.

Kapp EA, Schütz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS and Simpson RJ. "An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis.” Proteomics 5, no. 13 (Aug 2005): 3475-90.

Keller A, Nesvizhskii AI, Kolker E and Aebersold R. “Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.” Analytical chemistry 74, no. 20 (Oct 2002): 5383-92.

Kettenberger H, Armache KJ and Cramer P. “Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage.” Cell 114, no. 3 (Aug 2003): 347-57.

Kline-Smith S, Sandall S and Desai A. “Kinetochore-spindle microtubule interactions during mitosis.” Current opinion in cell biology (Elsevier) 17, no. 1 (2005): 35-46.

Koning L, Kasper P, Back J, Nessen M, Vanrobaeys F, Beeumen J, Gherardi E, Koster C and Jong L. “Computer-assisted mass spectrometric analysis of naturally occurring and artificially introduced cross-links in proteins and protein complexes.” FEBS Journal (Wiley Online Library) 273, no. 2 (2006): 281-291.

Leitner A, Walzthoeni T, Kahraman A, Herzog F, Rinner O, Beck M and Aebersold R. “Probing native protein structures by chemical cross-linking, mass spectrometry and bioinformatics.” Molecular & cellular proteomics : MCP, Mar 2010: 1634-1649.

Maiolica A, Cittaro D, Borsotti D, Sennels L, Ciferri C, Tarricone C, Musacchio A and Rappsilber J. “Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching.” Molecular & cellular proteomics : MCP 6, no. 12 (Dec 2007): 2200-11.

McHugh L and Arthur J. “Computational methods for protein identification from mass spectrometry data.” PLoS computational biology 4, no. 2 (2008).

Musacchio A and Salmon ED. “The spindle-assembly checkpoint in space and time.” Nature reviews Molecular cell biology 8, no. 5 (May 2007): 379-93.

Panchaud A, Singh P, Shaffer SA and Goodlett DR. "xComb: A Cross-Linked Peptide Database Approach to Protein– Protein Interaction Analysis." Journal of Proteome Research (ACS Publications) 9, no. 5 (2010): 2508-2515.

Pappin DJ, Hojrup P and Bleasby AJ. “Rapid identification of proteins by peptide-mass fingerprinting.” Current biology : CB 3, no. 6 (Jun 1993): 327-32.

Pedrioli PGA, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK, Kapp E, Mccomb ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu W and Aebersold R. “A common open representation of mass spectrometry data and its application to proteomics research.” Nature biotechnology 22, no. 11 (Nov 2004): 1459.

Petrovic A, Pasqualato S, Dube P, Krenn V, Santaguida S, Cittaro D, Monzani S, Massimiliano L, Keller J, Tarricone A, Maiolica A, Stark H and Musacchio A. “The MIS12 complex is a protein interaction hub for outer kinetochore assembly.” The Journal of Cell Biology 190, no. 5 (Sep 2010): 835-52.

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.” Nature biotechnology 17, no. 10 (Oct 1999): 1030-2.

Rinner O, Seebacher J, Walzthoeni T, Mueller LN, Beck M, Schmidt A, Mueller M and Aebersold R. “Identification of cross-linked peptides from large sequence databases.” Nature methods 5, no. 4 (Apr 2008): 315-8.

Santaguida S and Musacchio A. “The life and miracles of kinetochores.” The EMBO Journal 28, no. 17 (Sep 2009): 2511-31.

Schilling B, Row R, Gibson B, Guo X and Young M. “MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides.” Journal of the American Society for Mass Spectrometry 14, no. 8 (2003): 834-850.

Seebacher J, Mallick P, Zhang N, Eddes J, Aebersold R and Gelb M. "Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing." J. Proteome Res (ACS Publications) 5, no. 9 (2006): 2270-2282.

Singh P, Panchaud A and Goodlett DR. “Chemical cross-linking and mass spectrometry as a low-resolution protein structure determination technique.” Analytical chemistry 82, no. 7 (Apr 2010): 2636-42.

Sinz A. “Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes.” Journal of mass spectrometry : JMS 38, no. 12 (Dec 2003): 1225-37.

Takayama Y, Kamimura Y, Okawa M, Muramatsu S, Sugino A and Araki H. “GINS, a novel multiprotein complex required for chromosomal DNA replication in budding yeast.” Genes & development (Cold Spring Harbor Lab) 17, no. 9 (2003): 1153.

Werner M, Thuriaux P and Soutourina J. “Structure-function analysis of RNA polymerases I and III.” Current opinion in structural biology 19, no. 6 (Dec 2009): 740-5.

Wikipedia. “Two-hybrid screening.” Wikipedia, the free encyclopedia, 2007.

Woychik NA and Hampsey M. “The RNA polymerase II machinery: structure illuminates function.” Cell 108, no. 4 (Feb 2002): 453-63.

Xu M, Geer LY, Bryant SH, Kowalak JA and Markey SP. “Protein Scoring for Bottom-up Proteomics Data for OMSSA.” ASMS Poster Session, May 2006: 1-1.

Yates JR, Eng JK, McCormack AL and Schieltz D. “Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database.” Analytical chemistry 67, no. 8 (Apr 1995): 1426-36.

Young M, Tang N, Hempel J, Oshiro C, Taylor E, Kuntz I, Gibson B and Dollinger G. "High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry." Proceedings of the National Academy of Sciences of the United States of America 97, no. 11 (2000): 5802.

Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual JF, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C, de Smet AS, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, Barabási AL, Tavernier J, Hill DE and Vidal M. “High-quality binary protein interaction map of the yeast interactome network.” Science (New York, NY) 322, no. 5898 (Oct 2008): 104-10.

Acknowledgments

There are many people I would like to thank. As soon as I'll start writing I will probably forget some, sorry for that. Also, the order is not related to the importance.

I would like to thank Alessio Maiolica at first: he pointed all my mistakes, inspired a major part of this work and performed all the MS experiments.

Thanks to Andrea Musacchio, Angela Bachi and Michele Caselle for their kind and valuable tutorship.

Thanks to Juri Rappsilber, who introduced me to mass spectrometry and encouraged my very first ideas about cross-link analysis.

Thanks to Sebastiano Pasqualato and Arsen Petrovic for providing critics and, most important, valuable material to analyze.

Thanks to Gulliermo Montoya, Patrick Cramer and Cristoph Muller for providing me protein complexes without asking for results back.

A special thank goes to Barbara Felice and Matteo Cesaroni who persuaded me to attend this PhD course.

Thanks to every PhD classmate who silently attended my presentations over these years and shared some spare time in Turin. Thanks also for the ideas they inspired me, although not related with this work.

Last, but not least, a huge thank goes to my wife, Laura, who patiently waited for me, who pushed me to go on, who literally stared at my shoulders while I was writing pieces of this project.