View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Digitale Hochschulschriften der LMU

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

Conserved architecture of the core RNA polymerase II transcription initiation complex and an integrative model of Ctk3

Wolfgang Mühlbacher
aus
Bad Reichenhall, Deutschland

2015

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Patrick Cramer betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde selbstständig und ohne unerlaubte Hilfe erarbeitet.

Göttingen, den 04.08.2015

.....

Wolfgang Mühlbacher

Dissertation eingereicht am	29.06.2015
1. Gutachter	Prof. Dr. Patrick Cramer
2. Gutachter	PD Dr. Dietmar Martin
Mündliche Prüfung am	21.07.2015

Summary

Within this work, insights into the structure of the transcription initiation complex of RNA polymerase (Pol) II as well as into the CTDK-I complex that promotes transcription elongation were gained. During transcription initiation at promoters of protein-coding genes, Pol II assembles with TBP, TFIIB, and TFIIF into a conserved core initiation complex that recruits additional factors. The core complex stabilizes open DNA and initiates RNA synthesis, and it is conserved in the Pol I and Pol III transcription systems.

In the first part of this thesis, a protein-protein crosslinking approach was used to identify side-specific distance restraints by using mass spectrometry (MS). With this method, the domain architecture of the yeast core pol II initiation complex during transcription initiation was derived. The yeast complex resembles the human initiation complex and reveals that the TFIIF Tfg2 winged helix domain shows unexpected movement and swings over promoter DNA. An ‘arm’ and a ‘charged helix’ in TFIIF function in transcription start site selection and initial RNA synthesis, respectively, and apparently extend into the active center cleft of Pol II. Our model provides the basis for further structure-function analysis of the entire transcription initiation complex.

The second part of this work focuses on CTDK-I, a yeast kinase complex, that phosphorylates the C-terminal repeat domain (CTD) of RNA polymerase II (Pol II) to promote transcription elongation. CTDK-I consists of the cyclin-dependent kinase Ctk1 (homologous to human CDK12 and to a lower degree CDK9), the cyclin Ctk2 (homologous to human cyclin K), and the yeast-specific subunit Ctk3, which has been shown to be required for CTDK-I stability and activity. Ctk3 consists of a non-canonical CTD-interacting domain (CID) located at the N-terminal end and a predicted three-helix bundle domain at the C-terminal. We determine the X-ray crystal structure of the N-terminal domain of the Ctk3 homologue Lsg1 from the fission yeast *Schizosaccharomyces pombe* at 2.0 Å resolution. The structure reveals eight helices arranged into a right-handed superhelical fold that resembles the CID domain present in the yeast transcription termination factors Pcf11, Nrd1, and Rtt103.

Ctk3 however shows different surface properties and no binding to the CTD, which was determined by fluorescence anisotropy binding assays. Together with the known structure of Ctk1 and Ctk2 homologues, our results lead to a molecular framework for future work to further analyze the structure and function of the CTDK-I complex.

Acknowledgments

First of all, I want to thank Prof. Dr. Patrick Cramer for giving me the opportunity to work on such challenging projects in this outstanding scientific environment.

My great thanks also go to Andreas Mayer, who had constantly provided incredible advice throughout my master thesis and finally supported my transition from a master to a PhD student.

Next, I would like to thank the members of the Cramer lab. They all contributed to a wonderful and highly collaborative working atmosphere, which laid the foundation for the results presented in this thesis. In particular, I thank Clemens Plaschka, Sarah Sainsbury, and Jürgen Niesser for our constant and fruitful discussions and their support in general. In addition, a huge word of gratitude goes to all my collaborators inside and outside the lab including Franz Herzog, Merle Hantsche, Mai Sun, Simon Neyer, Johannes Soeding, Alan Cheung, Matthias Hemann and Michael Remmert. You did a remarkable work! Also, thanks to my assistant Laura Jochem, who completed a great research internship in the Cramer lab.

Moreover, I thank my dear colleagues and friends Sofia Battaglia, Carina Demel, Michael Lidschreiber, Margaux Michel, and Björn Schwalb. All of you were essential in helping me to get integrated in the lab at the beginning and throughout my thesis.

I would also like to thank Kerstin Maier, Claudia Buchen and Stefan Benkert for keeping everything together and gave me all sorts of advices.

I am thankful to my dear friends outside the lab: Tristan Philipp Harzer and Andrej Angelovski. We had an amazing time in Munich!

Very special thanks to my parents Andreas and Hildegard as well as my siblings Eva Maria, Andreas and Mathias for all your patient support!

Last but not least, I want to mention the help of my dear friends Julia Linke and Sara Osman trying desperately to teach me some of the deep mysteries of the English language. Thanks to you, Jule and Sara!

Publications

Part of this work has been published or is in the process of being published.

Wolfgang Mühlbacher*, Sarah Sainsbury*, Matthias Hemann, Merle Hantsche, Franz Herzog, and Patrick Cramer. Conserved architecture of the core RNA polymerase IIinitiation complex. *Nature comm*. 2014;5:4310.

* equally contributed.

Wolfgang Mühlbacher, Andreas Mayer, Mai Sun, Michael Remmert, Alan C.M. Cheung, Jürgen Niesser, Johannes Soeding and Patrick Cramer. The RNA polymerase II CTD kinase complex subunit Ctk3 contains a non-canonical CTD-interacting domain. *Proteins* 2015; Accepted Article.

Jürgen Niesser, Felix Roman Wagner, Dirk Kostrewa, Wolfgang Mühlbacher, Patrick Cramer. Structure of a GPN-loop GTPase chaperone and RNA polymerase II assembly factor. *EMBO J.* 2015*;* under review.

Contents

1 Introduction

1.1 Gene transcription

The transcription of DNA to RNA molecules catalyzed by DNA dependent RNA polymerases (Pol) represents a fundamental biological process found in all living organisms. During gene transcription the generated RNA molecules often serve as templates for protein synthesis[1]. Transcription in eukaryotes is carried out by three different nuclear RNA polymerases: RNA polymerase (Pol) I, Pol II, and Pol III[2]. Pol I synthesizes ribosomal RNAs (rRNAs), whereas Pol II produces all protein-coding messenger RNA (mRNA), small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs). Finally, Pol III synthesizes transfer RNAs (tRNAs), 5S rRNA, and diverse small RNAs[3,4]. In plants, a fourth (Pol IV) and fifth (Pol V) RNA polymerase synthesize RNAs involved in gene silencing. For genome transcription in chloroplasts and mitochondria, dedicated polymerases are present[5,6]. All RNA polymerases exhibit homology within their largest subunits, ranging from bacterial over archaeal to eukaryotes.

Pol I, II and III are multi-subunit complexes and share a conserved core of ten subunits and additional subsets of up to seven subunits. Pol II is composed of 12 subunits Rpb1 to Rpb12, of which Rpb1 represents the largest subunit.

Rpb1 has a unique, highly repetitive C-terminal domain (CTD) which plays a key role in the cycle of eukaryotic transcription (see Sections 1.4)[7,8].

1.2 Transcription cycle and the chromatin environment

The Pol II transcription cycle has been divided into five defined phases: Transcription pre-initiation, initiation, elongation, termination and re-initiation[9,10]. All transcription cycle events are precisely coordinated and controlled. Specific subsets of accessory proteins are needed to form complex networks which are required of regulation[11].

In a higher level of complexity, the genome is organized as chromatin. Various proteins are associated with chromatin, including the histones which are needed for the compact packaging of the genome. Moreover, interactions of diverse factors, such as histone modifying enzymes and chromatin remodelers may facilitate and regulate gene expression (for a more detailed overview see Figure 1B)[12].

1.2.1 Initiation

Before initiation occurs, Pol II needs to bind the promoter DNA together with the general transcription factors TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH to form a pre-initiation complex[13-15]. In addition, the pre-initiation complex bind to large coactivator complexes like Mediator or SAGA to facilitate transcription in different gene classes[16]. Coactivators link signals from gene-specific activators which bind upstream to the core promoter region (see Figure 1A)[17]. In the complete pre-initiation complex, the double stranded DNA is melted (open complex) and the nascent RNA molecules can be synthesized (initially transcribing complex)[18,19]. A detailed list of all GTFs and the corresponding functions is depicted in Table 1.

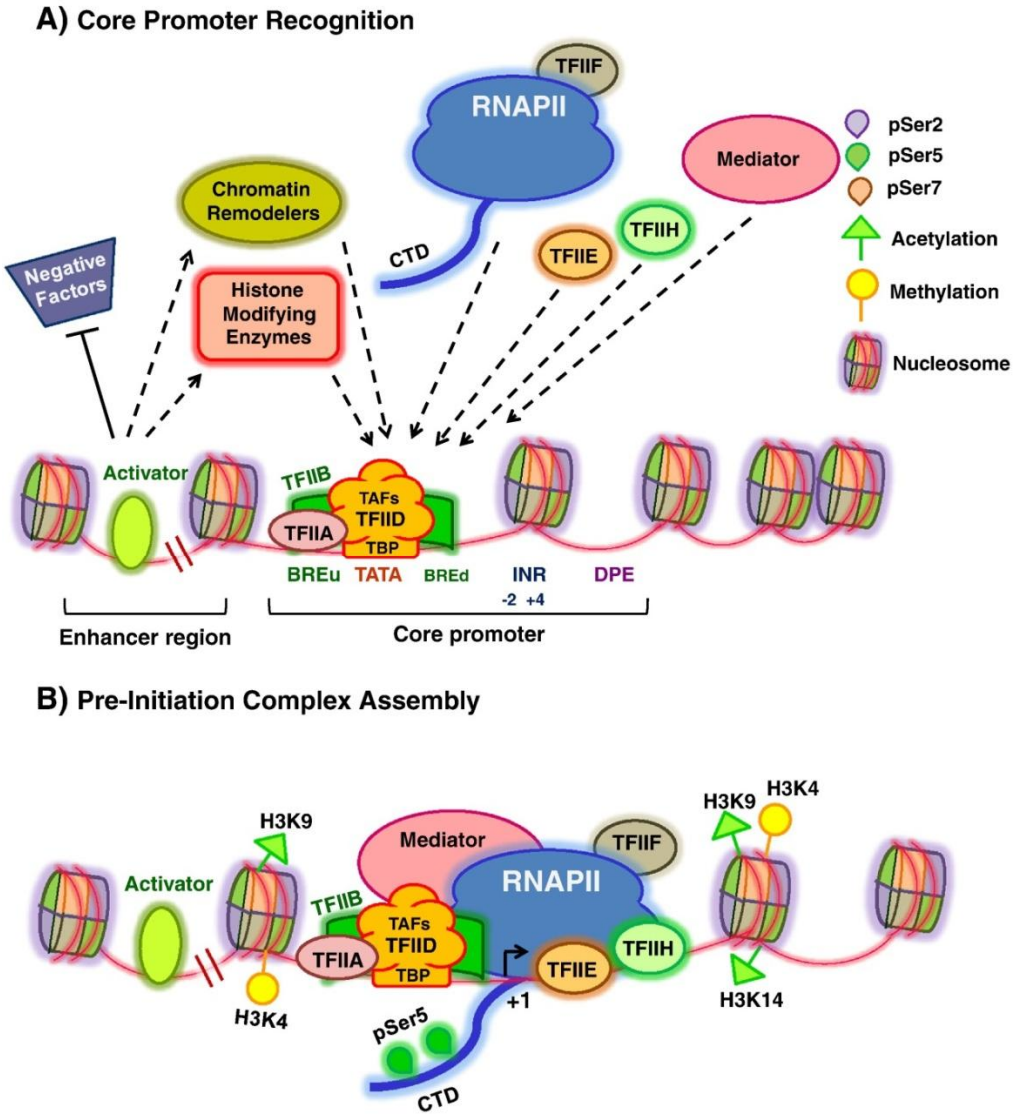


Figure 1 Promoter recognition and assembly of the pre-initiation complex**.**

(A) Binding of the activator to its enhancer sequence leads to the recruitment of the GTFs which bind to the core promoter elements (CPEs): TATA box, upstream and downstream B recognition element (BREu/d), Initiator sequence (Inr) and downstream promoter element (DPE). The TATA box is occupied by the TBP containing TFIID complex and TFIIB binds the BRE elements which play a role in the recruitment of Pol II and TFIIF and further GTFs[20]. Histone modifying enzymes (e.g. acetyltransferases, methyltransferases and nucleosome remodelers) alter the chromatin environment which is required for transcription. **(B)** The assembled PIC consists of the Pol II, GTFs and the mediator[13]. Nucleosomes which are close to the promoter DNA comprise distinct histone modifications like methylation of H3K4 (at the lysine 4 which of histone 3; yellow circles) and acetylation of H3K9/14 (green triangles) for active gene transcription. The repetitive CTD tail of Pol II is hyper-phosphorylated at position serine 5 (green circles). Adapted from[12].

Table 1: Pol II, its corresponding General Transcription Factors and the coactivators. Text in table was adapted from[14]

complex	Subunits; Functions
Pol II	12; transcription of all mRNAs and a subset of noncoding RNAs including snoRNAs and miRNAs
TFIIA	2–3; counteract negative cofactors; interact with activators and components of the basal initiation machinery
TFIIB	Single subunit; stabilizes TFIID-promoter binding; recruits TFIIF/Pol II to the promoter; start site selection
TFIID	14; nucleates PIC assembly; coactivator activity through direct interaction with gene specific activators
TFIIE	2; helps recruit TFIIH to promoters; stimulates helicase and kinase activities of TFIIH; promoter melting
TFIIF	2–3; associated with Pol II; enhances affinity of Pol II for TBP-TFIIB-promoter complex; recruitment of TFIIE/TFIIH to the PIC; start site selection and promoter escape; enhances elongation efficiency
TFIIH	10; promoter opening and promoter clearance; transcription coupled DNA repair; phosphorylation of Pol II CTD; transition from initiation to elongation
Mediator	At least 24; bridges interaction between activators and basal factors; stimulates both activator dependent and basal transcription; required for transcription from most Pol II dependent promoters
SAGA	20; interacts with activators, histone H3, and TBP; histone acetyltransferase activity; deubiquitinating activity

1.2.2 Elongation

To facilitate the transition from transcription initiation to the processive elongation phase, a transcription elongation complex (TEC) needs to be formed, comprising its own set of distinct proteins, the transcription elongation factors (EFs)[21,22]. The EFs can influence the processivity and rate of Pol II as well as the chromatin environment to guarantee and control the productive synthesis of nascent RNA molecules[23-25]. TFIIS is a key player to stimulate the productive transcription state of Pol II, by inhibition of promoter pausing, backtracking events and cleavage of the mRNA [25,26].

1.2.3 Termination and re-initiation

Transcription termination is coupled with the cleavage of the 3'-end of the synthesized mRNA strand. Thereby, the polyadenylation (pA) site of the nascent mRNA serves as a marker to initiate mRNA cleavage[27]. Downstream of the pA site Pol II dissociates from the DNA template and transcription is terminated[28,29].

To close the transcription cycle, Pol II and the GTFs can again re-initiate another cycle of gene transcription. Some factors remain bound to the promoter DNA such as: TFIID, TFIIE, TFIIF and the Mediator complex to facilitate a subsequent round of transcription[30].

1.3 Architecture of the core RNA polymerase II transcription initiation complex

The following introductory text in Section 1.3 was published.

Wolfgang Mühlbacher*, Sarah Sainsbury*, Matthias Hemann, Merle Hantsche, Franz Herzog, and Patrick Cramer. Conserved architecture of the core RNA polymerase IIinitiation complex. *Nature comm*. 2014;5:4310.
*These authors contributed equally.

1.3.1 From pre-initiation to the initially transcribing complex (ITC)

During assembly of the transcription pre-initiation complex (PIC), the Pol II-TFIIF complex binds to a TFIIB-TBP-DNA promoter assembly, resulting in a core initiation complex[31,32]. The structure and function of the core initiation complex is conserved from yeast to human, and also in the two other eukaryotic transcription systems[33]. Pol I and Pol III both contain a TFIIF-like subcomplex, and they also use TBP and a TFIIB-like factor for initiation. The conserved core initiation complex stabilizes open promoter DNA and directs initial RNA synthesis, resulting in the initially transcribing complex (ITC).

In the Pol II system, the core initiation complex additionally binds TFIIE and TFIIH to form a complete pre-initiation complex (PIC). Architectural models of the yeast Pol II PIC were obtained by site-specific protein cleavage mapping[34-36]. The architecture of the human PIC was obtained by electron microscopy (EM)[37], and generally resembled that of the yeast PIC. Recently, an alternative model of the yeast PIC was derived based on a combination of EM and protein crosslinking coupled to mass spectrometry (XL-MS)[38], raising the question whether the PIC architecture is indeed conserved between eukaryotic species.

We have previously modelled the architecture of the core Pol II initiation complex[39] by structural superposition of our Pol II-TFIIB crystal structures[40,41] with a Pol II-TFIIF complex model obtained by XL-MS[42]. However, the model awaited experimental confirmation because both TFIIF and TFIIB are modular factors with flexible domains that may be repositioned upon complex assembly.

The N-terminal regions of TFIIF subunits Tfg1 and Tfg2 form a dimerization module, whereas their flexibly linked C-terminal regions each include a winged helix (WH) domain. TFIIB consists of a N-terminal zinc ribbon domain followed by the reader and linker regions and two C-terminal cyclin domains.

1.3.2 Protein crosslinking of the ITC

In the first part of the thesis we used protein-protein crosslinking and mass spectrometric identification to derive a model of the core ITC from yeast. The yeast complex resembles the previously published human counterpart[37], indicating that the core initiation complex is conserved between eukaryotic species. The results also reveal a new element, the charged helix in the TFIIF subunit Tfg1, demonstrate that the Tfg2 WH domain can swing over the DNA after it was loaded into the active center cleft, and provide a basis for elucidating the architecture of the entire initiation complex.

1.4 The CTD of Pol II and the CTD kinase I complex (CTDK-I)

Wolfgang Mühlbacher, Andreas Mayer, Mai Sun, Michael Remmert, Alan C.M. Cheung, Jürgen Niesser, Johannes Soeding and Patrick Cramer. The RNA polymerase II CTD kinase complex subunit Ctk3 contains a non-canonical CTD-interacting domain. *Proteins* 2015; Accepted Article.

1.4.1 The Pol II C-terminal repeat domain (CTD)

The CTD of Rpb1 in Pol II consists of 26 (yeast) and 52 (human) heptapeptide repeats with the consensus sequence Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7[8,43]. The CTD serves as a binding platform for various factors during transcription, including pre-mRNA processing factors. During the transcription cycle, the CTD changes its phosphorylation pattern and this alters its binding affinity to factors[7,21,44,45]. Phosphorylation at position Ser5 is associated with pre-mRNA capping in early transcription elongation complexes[46]. Ser2 phosphorylation has been implicated in both elongation and termination events. The CTD residues Tyr1, Thr4, and Ser7 can also be phosphorylated[47-50].

1.4.2 Phosphorylation of the CTD

CTD phosphorylation is accomplished by four different cyclin-dependent kinases (CDKs) in yeast, namely Kin28, Srb10, Bur1, and Ctk1[7]. The Kin28 kinase and its human counterpart CDK7 are subunits of the initiation factor TFIIH and phosphorylate the CTD at position Ser5[51-53]. The Srb10 kinase associates with cyclin Srb11 and resides within the Mediator coactivator complex. The Srb10-Srb11 pair phosphorylates both Ser2 and Ser5 residues and is related to the mammalian pair CDK8-cyclin C[54,55]. The CTD kinases Bur1 and Ctk1 are Ser2 kinases and both share homology with mammalian CDK9, a subunit of positive transcription elongation factor b (P-TEFb), which induces productive elongation[56-58]. Ctk1 is the main Ser2 kinase in yeast, whereas Bur1 phosphorylates both Ser2 and the elongation factor Spt4/5[59],[60,61]. Bur1 also play a role in histone modification[62]. The transition from transcription initiation to elongation requires, in addition to Ser2 phosphorylation, dephosphorylation of Ser5 residues by Rtr1 and Ssu72[63,64]. Bur1 activity is controlled by cyclin Bur2[58,65]. Yeast Ctk1 and Bur1 kinases appear to be orthologues to metazoan Cdk12 and Cdk9, respectively[66].

1.4.3 The elongation promoting complex CTDK-I

Ctk1 (also known as Lsk1 in *S. pombe*) associates with its cyclin partner Ctk2 (*S. pombe* Lsc1) and a third subunit, Ctk3 (*S. pombe* Lsg1), to form the CTD kinase I (CTDK-I) complex[67-71]. This trimeric structure is unique amongst CDK complexes[68,72]. *S. cerevisiae* Ctk3 and *S. pombe* Lsg1 share 24% amino acid sequence identity, and associate with Ctk1/Ctk2 and *S. pombe* Lsk1/Lsc1, respectively[69-71,73,74]. Throughout this work, we refer to *S. pombe* Lsg1 as Ctk3. In *S. cerevisiae*, the activity of Ctk1 and Ctk2 are strongly dependent on the binding to Ctk3[67,75]. The Ctk3 C-terminal region is involved in the stabilization of the Ctk2-Ctk3 heterodimer and CTDK-I function[67]. Recruitment of Ctk1 *in vivo* relies to some extent on the completion of the pre-mRNA 5'-cap structure[76], but there are additional, unknown mechanisms of CTDK-I recruitment. Ctk3 may function in CTDK-I recruitment, although Ctk3 does not have counterparts in metazoa[68,72].

2 Materials and Methods

2.1 Materials

2.1.1 Bacterial and Yeast strains

Table 2: Bacterial and Yeast strains.

Bacterial Strain	Description	Company
XL1- blue	recA1 endA1 gyrA96 thi-1 hsdR17 supE4 relA1 lac(F' proAB lacIqZDM15Tn10(Tetr))	Stratagene
BL21-(DE3)RIL	B F- ompT hsdS(rB- mB-) dcm+ Tetr gal l (DE3) endA The (argU ileY leuW Camr), extra copies of argU ileY and leuW tRNA genes on a COlE1-compatible plasmid with chloramphenicol resistance marker, protease deficiency, chromosomal T7-polymerase gene	Stratagene
Rosetta B834 (DE3)	F- ompT hsdSB(rB- mB-) dcm+ metB, methionine auxotroph	Novagen
Yeast Strain	**Description**	**Company**
BJ5464 His-Bio	BJ5464 His-Bio tag introduced at 5'end of Rpb3 gene, use of URA3 selection marker	(Kireeva et al., 2000b)

2.1.2 Plasmids and oligonucleotides

Table 3: Plasmids used in this study.

Vector	Species	Insert	Type	Resistance.	Source
WM01	*Sc*	TBP (61-240), N-6xHis	pET28b	Amp	MB
WM02	*Sc*	TFIIB	pOPINE	Amp	SS
WM03	*Sc/Sm*	TFIIF	pETduet	Amp	KK
WM04	*Sp*	Ctk3 full-length	pET28b	Amp	This work
WM05	*Sp*	Ctk3 (1-140)	pET28b	Kan	This work
WM06	*Ca*	Ctk3 (1-146)	pET28b	Kan	This work
WM07	*Sc*	Ctk3 (1-149)	pET28b	Kan	This work

Sc, *Saccharomyces cerevisiae*; *Sm, Saccharomyces mikatae; Sp, Schizozaccharomyces pombe*; *Ca, Candida albicans;* Kan, Kanamycin; Amp, Ampicillin; cloned by Kerstin Kinkelin, KK; Michela Bertero, MB; Sarah Sainsbury, SS. A detailed list of DNA oligonucleotides used for cloning can be obtained from the Cramer group.

Table 4: Oligonucleotides used for Protein crosslinking.

Type	Sequence (5’ - 3’)
Template DNA	CGA GAA CAG TAG CAC GCT GTG TAT ATA ATA GTG TGT TGT ACA TAG CGG AGG TCG GTG GGG CAC AAC TGC GCT
Non-template DNA	AGC GCA GTT GTG CTA TGA TAT TTT TAT GTA TGT ACA ACA CAC TAT TAT ATA CAC AGC GTG CTA CTG TTC TCG
RNA	AUA UCA

2.1.3 Growth media and additives

Table 5: Growth media for E. coli and S. cerevisiae cultures.

Media	Organism	Description
LB	*E. coli*	1 % (w/v) tryptone, 0.5 % (w/v) yeast extract. 0.5 % (w/v) NaCl
SeMet	*E. coli*	22 g/L SeMet base, 5g/L nutrient mix, 40 µg/mL SeMet (Molecular Dimensions)
YPD	*S. cerevisiae*	2 % (w/v) peptone, 2 % (w/v) glucose, 1 % (w/v) yeast extract

Table 6: Additives for E. coli cultures.

Additive	Application	Stock solution	Applied concentration
Ampicillin	Antibiotic	100 mg/mL in H2O	0.1 mg/L
Chloramphenicol	Antibiotic	30 mg/mL in Ethanol	0.03 mg/L
Kanamycin	Antibiotic	20 mg/mL in H2O	0.02 mg/L
IPTG	Protein expression	1 M in H2O	0.5 mM

IPTG = Isopropyl-β-D-1- thiogalactopyranoside.

2.1.4 General buffers, markers and solutions

Table 7: List of general buffers and solutions.

Name	Description	Application
Electrophoresis buffer	10x MOPS NuPAGE buffer (Life Technologies)	SDS-PAGE
5x SDS sample buffer	250 mM Tris-HCl (pH 7.0 at 25°C); 50% (v/v) glycerol; 0.5% (w/v) bromophenol blue; 7.5% (w/v) SDS; 500 mM DTT	SDS-PAGE
20 × MES SDS running buffer	50 mM MES; 50 mM Tris Base; 0.1% SDS; 1 mM EDTA; pH 7.3 at 25°C	SDS-PAGE
20 × MOPS SDS running buffer	50 mM MOPS; 50 mM Tris Base; 0.1% SDS; 1 mM EDTA; pH 7.7 at 25°C	SDS-PAGE
Broad range MW marker	Bio-Rad	SDS-PAGE
SDS-PAGE stain	Instantblue (Expedion)	SDS-PAGE
Coomassie gel staining solution	50% (v/v) ethanol; 7% (v/v) acetic acid; 0.125% (w/v) Coomassie Brilliant Blue R-250	SDS-PAGE
100x PI	0.028 mg/mL Leupeptin, 0.137 mg/mL Pepstatin A, 0.017 mg/mL PMSF, 0.33 mg/mL Benzamidine in Ethanol	Protease Inhibitor
10x TAE	50 mM EDTA pH 8, 2.5 M Tris-acetate	Agarose gel electrophoresis
SYBR Safe (10,000 × in DMSO)	Invitrogen	Agarose gel electrophoresis
Gene Ruler 1 kb DNA ladder (0.1 µg/µL)	Fermentas	Agarose gel electrophoresis
1 × TE	10 mM Tris-HCl (pH 8.0 at 25°C); 1 mM EDTA	Oligonucleotides

2.2 Common Methods

2.2.1 Molecular cloning

Polymerase Chain Reaction (PCR)

Primers were designed by using an overhang of nucleotides at the 5' end (5' -AGGAGGAGG- 3'), followed by a restriction side and 20 or more nucleotides complementary to the gene sequence of interest. PCR reactions were carried out with Phusion High-Fidelity DNA Polymerase (Finnzymes), in a 50 µL reaction volume. 50 ng Synthesized oligonucleotide plasmids were used as template DNA. 0.5 pmol/µL PCR primers were used in each reaction. Reactions took place in Biometra T3000 Thermocycler with 30 cycles. Primer annealing temperature and synthesis time varied according to the length of DNA template and primer. PCR products were visualized by using 1% agarose gel electrophoresis and Sybr-Safe staining. Purification of the DNA was carried out with QIAquick gel extraction kit (Quiagen).

Enzymatic restriction cleavage

DNA was digested using restriction endonucleases from Fermentas and New England Biolabs (NEB) as recommended in the producers guidelines. Cleaved PCR products and plasmids were purified using the QIAquick-PCR purification and –gel extraction kits (both Quiagen), respectively.

Ligation

Digested DNA was ligated into linearized vectors at room temperature for 1 hour in a volume of 20 µl using T4 DNA ligase and its corresponding buffer (Fermentas). A 5-fold excess of insert, relative to the linearized vector was used.

Transformation and sequencing

Chemically competent *E. coli* XL-1 blue cells (Table 2) were mixed with 50 µL DNA plasmids and transformed by heat shocking on 42 ^{0}C for 45 sec. 700 µL LB Medium was added and incubated at 37 ^{0}C for 1 h. After sedimentation (30 sec, 14000 rpm), the cells were re-suspended in 200 µL LB-Medium and transferred to LB-Agar plates, containing the corresponding antibiotics for selection. The plates were incubatet at 37 ^{0}C over night. 5 mL LB media was inoculated by a single bacteria colony, representing a single clone, and further incubated at 37 ^{0}C over night and used for the preparation of plasmid DNA using the QIAquick Miniprep Kit (Qiagen). Isolated plasmids were verified by DNA sequencing (Company: GATC).

2.2.2 Protein expression in *E. coli* and complex formation

Protein expression and purification

E. coli cultures were expressed and purified following the respective sections in 2.3 and 2.4. In general, cultures with volumes from 1-8 L LB were inoculated with 50 mL pre-culture, which was incubated over-night. All cultures contained the antibiotics corresponding to the resistance of the transformed vector. Cell were grown to a target OD_{600} of 0.6-0.9 and induced by the addition of 0.5 mM IPTG to start protein expression. Proteins were expressed at 18°C overnight. Recombinant proteins were purified using affinity purification, ion exchange and subsequent size exclusion chromatography.

Protein concentrating, and storage

Proteins were concentrated by AMICON Ultra spin concentrators (Millipore) with defined molecular weight cutoffs, at least three-fold smaller than the target protein. Protein concentration was determined by using the NanoDrop spectrophotometer (absorption at 280 nm). Absorption coefficients were calculated by the ProtParam tool (http://expasy.org/tools/protparam.html). Purified protein samples were frozen in liquid nitrogen and stored at -80 °C.

SDS-PAGE analysis and protein identification

Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was used to determine the stoichiometry and the approximate concentration of protein samples. 20 µL protein solution and 5 µl 5x sample buffer were mixed together and boiled at 95 °C for 3 min. 15 µL protein samples were loaded into the gel-pockets. Gel electrophoresis took place at 100 mA for 30 min. Gel staining was accomplished using InstantBlue (Expedeon) for 30 min. Protein samples were identified by mass-spectrometry analysis from the protein core facility of the Adolf-Butenandt-Institute, LMU.

DNA-RNA Scaffold preparation

DNA and RNA oligonucleotides were separately dissolved in in 1x TE buffer at a concentration of 400 µM. Dissolved oligonucleotides were mixed to reach an equimolar concentration of 100 µM. Annealing took place in a T3000 Thermocycler (Biometra) due to cooling from 95 °C to 10 °C in 1 °C steps occurring every 30 seconds. DNA-RNA scaffold was either directly used in complex formation or stored at -20°C.

2.2.3 Crystallization

Initial crystallization

To determine initial crystallization conditions, protein samples were forwarded to the MPI crystallization facility (Max Planck Institute of Biochemistry in Martinsried). Protein samples were applied to diverse sparse matrix screens (96-well plates with sitting drop vapor diffusion technique). All Screens were performed at both, 4 °C and room temperature. Total size of the hanging drop was 200 nL (100 nL protein and reservoir solution, respectively). Following screens from QIAGEN were used: Classics; Classics Lite; $AmSO_4$; Pegs; pH Clear 1, pH Clear 2. Further screens originated from in house source: Crystal platform Magic 1; Crystal platform Magic 2 and from Hampton Research: Index screen.

Optimization of crystallization

The optimization of initial crystallization was carried out manually in 15-well hanging drop crystallization plates. In general, 1 µL pure protein was mixed with 1 µL reservoir buffer and incubated over a 600 µL reservoir solution at either 20°C or 4°C. The ratio of protein to reservoir solution was 1:1, 1:2, or 2:1.

2.3 Specific methods for section 3.1 with focus on protein crosslinking

The following text in Section 2.3 was published.

Wolfgang Mühlbacher*, Sarah Sainsbury*, Matthias Hemann, Merle Hantsche, Franz Herzog, and Patrick Cramer. Conserved architecture of the core RNA polymerase IIinitiation complex. *Nature comm*. 2014;5:4310.
*These authors contributed equally.

2.3.1 Preparation of the yeast core Pol II ITC

Endogenous *S. cerevisiae* 12-subunit Pol II was prepared as described[80]. Full-length TFIIB[41], TFIIF (*S. mikatae* Tfg1, *S. cerevisiae* Tfg2)[34] and TBP[81] (residues 61-240) were prepared as described. Pol II (0.77 mg, 3.5 mg ml^{-1}) was incubated with a four-fold molar excess of TFIIF, TFIIB, and TBP, and a two-fold molar excess of DNA-RNA scaffold (Figure 2a) for 30 min at 298 K and for 5 min at 293K, 288K, and 283K. Size-exclusion chromatography in 250 mM KCl, 20 mM HEPES pH 7.5, 5% glycerol, and 2 mM DTT resulted in a stoichiometric ITC (Figure 2b).

2.3.2 Crosslinking and mass spectrometry

0.9 mg purified ITC (1.2 mg ml^{-1}) was incubated with an eight-fold molar excess of DNA-RNA scaffold and crosslinked with 0.6 mM isotope-labeled disuccimidyl suberate (DSS-d0/d12, Creative Molecules Inc.) as described[82]. Crosslinked protein was digested, and the crosslinked peptides were enriched, analyzed by liquid chromatography coupled to tandem mass spectrometer (Orbitrap Elite^{TM}), and spectra were searched by the xQuest software[83,84]. The resulting cross-link identifications were manually validated and the local false discovery rates for each individual cross-link were estimated as described. The term ‘crosslink’ describes a peptide-peptide pair linked through two specific lysines. A single peptide-peptide combination of peptides containing more than one lysine each can be identified by distinct cross-links which represent a single unique distance restraint.

Finally, we also detected 11 crosslinks, which linked to serine, threonine and tyrosine with maximum Cα distances of 21.7 Å (Supplementary Table 3). Nevertheless, these crosslinks presented no additional structural information.

2.3.3 Structural modeling

All modeling was done manually. To generate the ITC model, we used PyMOL and crosslinking restraints to place homology models of the yeast TFIIF dimerization module and WH domains and the TFIIB C-terminal cyclin domain onto the Pol II-TFIIB (PDB: 4BBR) open promoter complex model[40]. Models for the yeast TFIIF dimerization module and WH domains and TFIIB C-terminal cyclin domain were generated from known crystal structures (dimerization module, chains A and F in PDB 1F3U; Tfg1 WH domain, PDB 1I27; Tfg2 WH domain, PDB 1BBY; TFIIB C-terminal cyclin domain, chain A in PDB 1VOL) using MODELLER[85]. Residues 92-153 and 324-417 in *S. cerevisiae* Tfg1 correspond to residues 5-62 and 73-168 in human Rap74. Residues 54-138 and 208-227 of *S. cerevisiae* Tfg2 align to residues 2-119 of human Rap30, respectively. Residues 678 – 736 of *S. cerevisiae* Tfg1 WH domain align with residues 454 – 517 in human Rap74 and residues 292 – 350 in *S. cerevisiae* Tfg2 WH domain align to residues 176 – 243 in human Rap30. Residues 125-345 of *S. cerevisiae* TFIIB C-terminal cyclin domain align to residues 113-316 in the human counterpart. The Tfg1 sequence of *S. mikatae* was substituted with the one of *S. cerevisiae* since they only differ in three amino acids in the dimerization module model and in six in its winged helix model.

2.4 Specific methods for section 3.2 with focus on crystallography

Wolfgang Mühlbacher, Andreas Mayer, Mai Sun, Michael Remmert, Alan C.M. Cheung, Jürgen Niesser, Johannes Soeding and Patrick Cramer. The RNA polymerase II CTD kinase complex subunit Ctk3 contains a non-canonical CTD-interacting domain. *Proteins* 2015; Accepted Article.

2.4.1 Sample preparation

DNA constructs of *S. pombe* full-length Ctk3 (residues 1-218) and the Ctk3 N-terminal domain (residues 1-140) were synthesized (Mr. Gene GmbH) and cloned into pET28b+ expression vector (Novagen) resulting in C-terminal hexahistidine tags. Both, Ctk3 (1-218) and Ctk3 (1-140) protein variants were expressed overnight at 18ºC in *Escherichia coli* (*E. coli*) BL21 (DE3) RIL cells (Stratagene). *E. coli* strain B834 (DE3) pLsyS (Novagen) was used for selenomethionine (SeMet) labeling. Cells were harvested and resuspended in 50 mM Tris-HCl pH 8.0, 300 mM NaCl and 2 mM DTT, followed by sonication. The resulting slurry was cleared by centrifugation. The cleared lysate was subjected to affinity chromatography on a Ni-NTA column (Qiagen) before dialyzing against 50 mM Tris-HCl pH 8.0, 50 mM NaCl and 1 mM DTT to reduce the high imidazole concentration. The hexahistidine tag was removed by thrombin cleavage at 4ºC overnight. The protein variants were further purified by anion exchange chromatography (MonoQ column, GE healthcare life science). After size exclusion chromatography (Superpose-12 column, GE healthcare life science) in gelfiltration buffer (50 mM HEPES pH 8.0, 50 mM NaCl, 1 mM DTT), the pure protein was concentrated to 14.5 mg ml^{-1}.

2.4.2 Crystal structure determination

Crystals for the Ctk3 N-terminal domain variant (residues 1-140) were grown at 4ºC using hanging-drop vapour diffusion. The reservoir solution contained 26% PEG 6000, 100 mM citric acid pH 4.0, 0.8 M lithium chloride and 5 mM Tris(2-carboxyethyl)phosphin (TCEP). Grown crystals were transferred to reservoir buffer containing 10% PEG400. Diffraction data were collected at the Swiss Light Source (SLS) in Villigen, Switzerland. Data were processed by XDS and scaled using XSCALE[86]. The crystal structure was solved by multiwavelength anomalous diffraction (MAD) from SeMet-labeled crystals using SOLVE[87]. Density modification was carried out with RESOLVE[87]. An initial model was automatically built with ARP/wARP[88]. Manual model building was carried out in Coot[89]. The model was refined by PHENIX[90] using individual isotropic B-factors and bulk solvent correction to a free R-factor of 25.4% at 2.0 Å resolution.

2.4.3 Peptide interaction analysis

We measured the protein-peptide interactions by fluorescence anisotropy. The synthetic CTD peptides were labelled by N-terminal aminocaproic-linked fluorescein. Changes in fluorescence anisotropy of the peptide solution were measured by titration of Ctk3 N-terminal domain or full-length Ctk3 (FluoroMaxP, HORIBA). All peptides were dissolved in 20 mM HEPES pH 8.0, 10 mM NaCl and 5 mM DDT, and adjusted to a concentration of 0.4 mM. Ctk3 N-terminal domain and full-length Ctk3 proteins were dissolved in gelfiltration buffer. The FluoroMaxP analyzer was calibrated at 20ºC with gelfiltration buffer and 1 µL peptide in a quartz cuvette (0.4 mM). Before analysis, solutions were mixed by magnetic stirring for 1 min and incubated for an additional minute. The protein solution was titrated to the peptide solution in steps of 20 µL, and three measurements were recorded to enable calculation of an arithmetic average. The excitation wavelength was set to 495 nm (slit width = 2) and the emission wavelength to 520 nm (slit width = 1).

3 Results and Discussion

3.1 Conserved architecture of the core RNA polymerase II initiation complex

The following text in Section 3.1 was published.

Wolfgang Mühlbacher*, Sarah Sainsbury*, Matthias Hemann, Merle Hantsche, Franz Herzog, and Patrick Cramer. Conserved architecture of the core RNA polymerase IIinitiation complex. *Nature comm*. 2014;5:4310.
*These authors contributed equally.

3.1.1 Preparation and crosslinking analysis of the core ITC

To experimentally define the domain architecture of the core initiation complex, we reconstituted a defined yeast core ITC. We previously noted that a stable, defined ITC can be formed by including a 6 nt RNA product[41]. We therefore incubated purified Pol II, TFIIF, TFIIB, and TBP with a DNA-RNA scaffold (Figure 2a) and obtained a stable core ITC after size exclusion chromatography (Figure 2b) (see Online Methods). We then analysed this complex by XL-MS. The power and reliability of XL-MS was recently exemplified by a high agreement between Pol I models derived from XL-MS[82] and subsequently from X-ray analysis[91].

We obtained a total of 472 high-confidence lysine-lysine protein crosslinks (332 distance restraints) within the core ITC (Supplementary Table 1 and 2), of which 241 were inter-subunit and 231 were intra-subunit crosslinks (Table 8). A total of 194 crosslinks within Pol II were readily explained with the Pol II crystal structure[92]. Another 33 crosslinks were observed between TFIIF subunits Tfg1 and Tfg2, and could be explained with the TFIIF dimerization module structure[93]. Only 18 crosslinks showed Cα distances above the maximum expected distance of 27±3 Å[42] (Figure 2c).

Most of these could be explained by the known structural flexibility and higher crystallographic B-factors of the involved lysine residues, leaving only three crosslinks unexplained. Within TFIIB and TBP, 23 and six intramolecular crosslinks were observed, respectively, and could be explained with crystal structures[40,41,94,95]. These multiple internal controls demonstrate the high reliability of the observed crosslinking data.

3.1.2 Positions of TFIIB and TFIIF

We also observed 111 intermolecular crosslinks between transcription factors and Pol II (Table 8) that allowed us to model the core yeast ITC (Figure 3a). Of these crosslinks, 21 were observed between the TFIIF dimerization module and the Pol II lobe and protrusion domains, showing that the module remains at its location observed in the binary Pol II-TFIIF complex[35,42]. Further, 16 crosslinks were obtained between the TFIIB linker and the Pol II domains clamp core, protrusion, and wall. Another six crosslinks were detected between the N-terminal TFIIB cyclin domain and the clamp core, protrusion and wall. The C-terminal cyclin domain did not crosslink to Pol II, consistent with its mobility[40]. All TFIIB-Pol II crosslinks were explained with our crystal structures of the Pol II-TFIIB complex[40,41], demonstrating that TFIIB binds Pol II as observed in the binary structure. These results were corroborated by crosslinks between the TFIIF subunit Tfg2 to TBP (one crosslink) and to the N-terminal cyclin domain of TFIIB (six crosslinks).

3.1.3 The Tfg2 WH domain swings over DNA in the cleft

The crosslinking data also revealed that in the reconstituted core ITC the WH domain in Tfg2 can reside at a position near upstream DNA on the outside of Pol II[42], but also at a position above the DNA in the active center cleft (Figure 3a, d). Thus, in the core ITC, the WH domain remains flexible and adopts both alternative positions. The original WH position near upstream DNA[42] gives rise to 13 crosslinks to the Pol II subunits Rpb2, Rpb3, and Rpb10. The new position above the Pol II cleft was defined by four crosslinks of the WH domain to the clamp, TBP, and the TFIIB N-terminal cyclin domain. These restrains can be satisfied when one assumes a position of the WH domain with respect to promoter DNA that resembles that in a known X-ray structure of a WH domain bound to DNA[96]. This position is also consistent with a recent mapping of the DNA-binding face of the Tfg2 WH domain[97].

These results indicate that the Tfg2 WH domain can swing over promoter DNA after its loading into the Pol II cleft, and indicate a role of this domain in DNA melting and/or stabilization of the open complex and the ITC. Indeed, this domain binds DNA and is required for initiation[98], and TFIIF suppresses abortive initial transcription[99]. The proximity of the Tfg2 WH domain to TFIIB indicates how TFIIF could stabilize TFIIB on Pol II during initial transcription[100]. The position of the Tfg2 WH domain above the cleft apparently represents its position in a complete ITC. This position is near TFIIE and TFIIA in the PIC[37,38], and is likely stabilized upon TFIIE and/or TFIIA binding. The other WH domain in TFIIF subunit Tfg1 only gave rise to a single crosslink at the Pol II jaw, and does not adopt a defined location[42].

3.1.4 Model of the yeast core ITC

Based on the large number of protein crosslinks we built a reliable three-dimensional model of the yeast Pol II core ITC. First, we derived a homology model of the yeast TFIIF dimerization module based on the human crystal structure[93]. Second, we positioned the resulting yeast TFIIF dimerization module model onto the Pol II-TFIIB-DNA-RNA crystal structure[41] assuming the location of the human module detected by EM[37]. Third, we extended DNA both upstream and downstream using standard B-form duplexes.

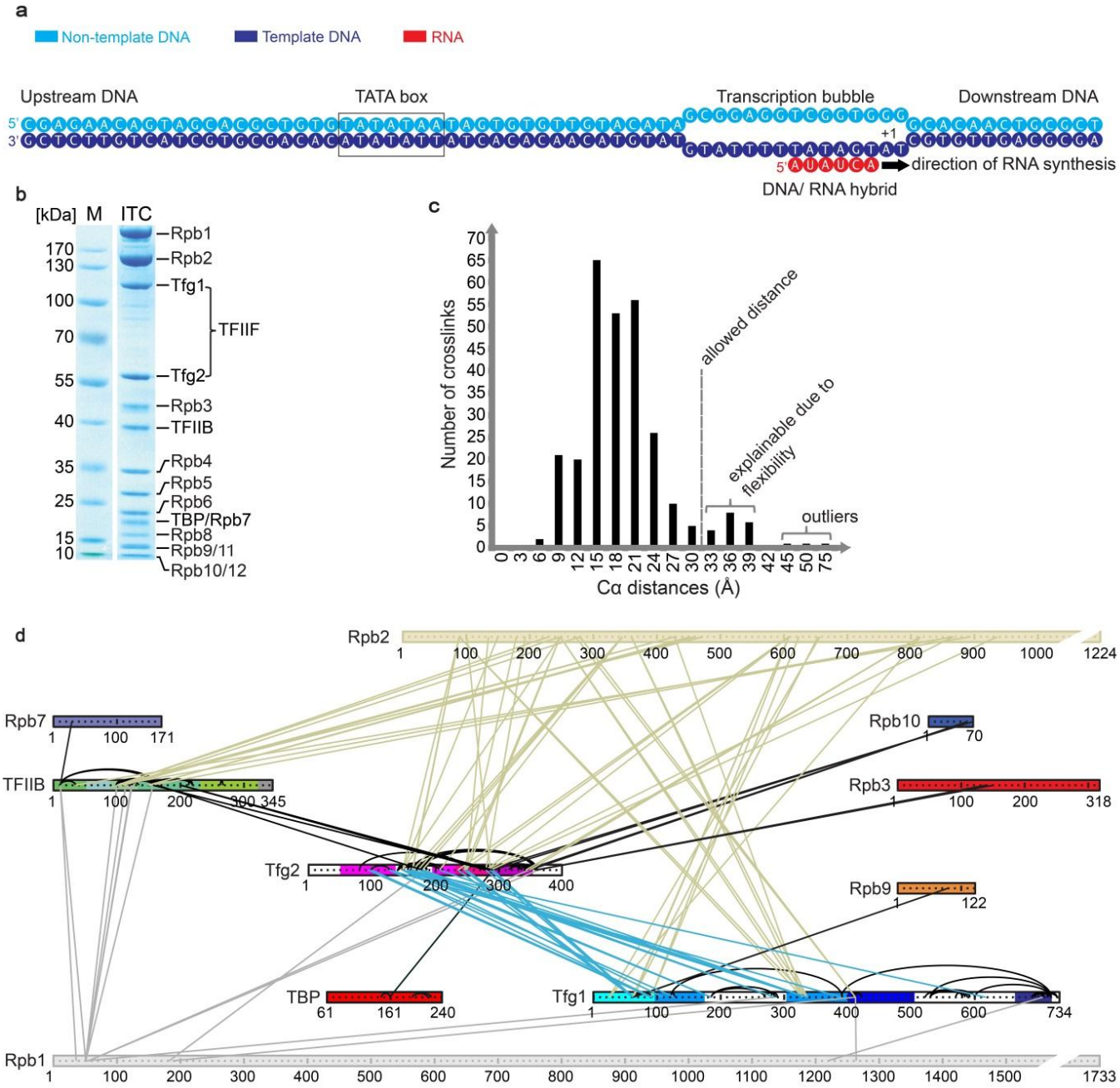

a
Non-template DNA
Template DNA
RNA
Upstream DNA
TATA box
Transcription bubble
Downstream DNA
+1
direction of RNA synthesis
DNA/ RNA hybrid
b
[kDa] M ITC
Rpb1
Rpb2
Tfg1
TFIIF
Tfg2
Rpb3
TFIIB
Rpb4
Rpb5
Rpb6
TBP/Rpb7
Rpb8
Rpb9/11
Rpb10/12
c
Number of crosslinks
allowed distance
explainable due to flexibility
outliers
Cα distances (Å)
d
Rpb2
Rpb7
Rpb10
TFIIB
Rpb3
Tfg2
Rpb9
TBP
Tfg1
Rpb1

Figure 2: Preparation and XL-MS analysis of the yeast core ITC.
(A) DNA-RNA scaffold based on a *HIS4* DNA promoter with a mismatched bubble region containing a 6 nt RNA transcript formi ng a hybrid duplex with the DNA template strand[41]. **(B)** SDS-PAGE analysis of the purified Pol II ITC revealing its 16 polypeptide subunits. **(C)** Cα distance distribution for observed lysine-lysine crosslink pairs (unique distance restraints). Crosslinks with distances of 30-39 Å are explainable due to protein mobility (four crosslinks) or because of lysine location in mobile protein loops with high crystallographic B-factors (14 crosslinks). Only three crosslinks cannot be explained and are classified as outliers. **(D)** Crosslink map of the ITC. Crosslinks within Pol II were excluded for clarity. TFIIB and Pol II subunits are colour-coded as before[41] and TBP and TFIIF were coloured as in Figure 3. The map was generated with a MATLAB® script by coauthor Simon Neyer (see Supplementary Material 1).

3.1.5 TFIIF arm and charged helix

In the resulting model, the Tfg1 'arm' (a b-hairpin comprising yeast residues 146-153 and 319-338) extends from the TFIIF dimerization module, traversing between the Pol II protrusion and lobe domains into the active center cleft (Figure 3c). The arm forms 19 crosslinks in the cleft, consistent with detection of the arm in the human PIC by EM[37]. A mutation at the point where the arm extends from the dimerization module leads to shifts in the transcription start site[101].

A second extension from the TFIIF dimerization module, a negatively charged, flexible[93] α-helix at the beginning of the 'charged region' in Tfg1 (named here the 'charged helix', yeast residues 406-417) clashed with the Pol II lobe. A reorientation of the charged helix towards the jaws released this clash and explained 6 crosslinks of the charged helix to the Pol II cleft. The location of the charged helix explained a distinct EM difference density that was hypothesized to stem from the corresponding human TFIIF region[37].

Published mutational and kinetic data revealed important roles of the charged helix in transcription initiation and elongation[102,103]. These roles can now be rationalized due to the location near downstream DNA (Figure 3c). The charged helix apparently repels the downstream DNA from the lobe, positioning it along the clamp head on the opposite side of the cleft. This may help to stabilize melted DNA and to align the DNA template in the active site and account for the known role of the charged helix in stimulating initial RNA synthesis[104].

3.1.6 Conclusion

Our crosslinking data and detailed architectural model of the core yeast ITC agree with previous site-specific protein cleavage mapping of the yeast PIC[34,35,105]. Our model further agrees with structural analysis of human Pol II PIC intermediates by EM[37]. Thus the architecture of the core ITC is highly conserved between yeast and human. Domains in TFIIF and TFIIB adopt very similar locations on the Pol II surface in both species, although the position of the Tfg2 WH domain above the cleft may change slightly upon binding TFIIA and/or TFIIE or different DNA template sequences. Furthermore, two TFIIF motifs extending from the dimerization module, the arm and charged helix, adopt the same locations in the downstream cleft. Our results also indicate that the overall domain architecture of the initiation complex is generally maintained during the transition from a PIC to an ITC.

Finally, our core ITC model can explain the XL-MS data obtained recently with a complete yeast PIC[38]. In the latter study, 117 distance restraints were obtained for Pol II, TFIIB, TFIIF and TBP. Of these, only one crosslink disagrees with our model, which was derived from 472 crosslinks with 332 distance restraints. Apparently the published study[38] contains correct crosslink information, but conflicting electron microscopic results, which have apparently led to an alternative initiation complex model. With respect to the core ITC, the discrepancies are now resolved. Our results lead to a unified, highly conserved architecture of the core transcription initiation complex. The location of the remaining general transcription factors TFIIE and TFIIH differs to some extent in three published studies[36-38] and may be analyzed in the future.

Table 8: Observed lysine-lysine crosslinks in the yeast core Pol II ITC.
Provided is the number of crosslinks between certain parts of the ITC, referring to unique distance restraints.

ITC parts	Crosslinks (all)	Crosslinks (mapable)
All (*inter and intra crosslinks*)	472	328
inter crosslinks		
All	241	164
Pol II-Pol II	90	90
Pol II-TFIIF	84	40 (48)[1]
Pol II-Tfg1/2 dimerization module[2]	25 (29)	25 (29)
Pol II-Tfg1 WH	1	1
Pol II-Tfg2 WH	14 (17)	14 (17)
Pol II-TFIIB	27	23
Tfg1-Tfg2	33	11
TFIIB-Tfg2	6	0 (1)
TBP-Tfg2	1	0 (1)
intra crosslinks		
All	231	164
Pol II	104	104
TFIIF	98	32
Tfg1/2 dim.-module[2]	4	4
Tfg1 WH-model	18	18
Tfg2 WH-model	11	11
TFIIB	23	22
TBP	6	6

[1]Numbers in brackets include crosslinks that involve amino acids located no more than three residues away from residues within known structures.

[2]Residues from the charged region of Tfg1[42] (400-417) and N-terminal region (92-98) are also part of the dimerization model based on the human X-ray structure.[93]

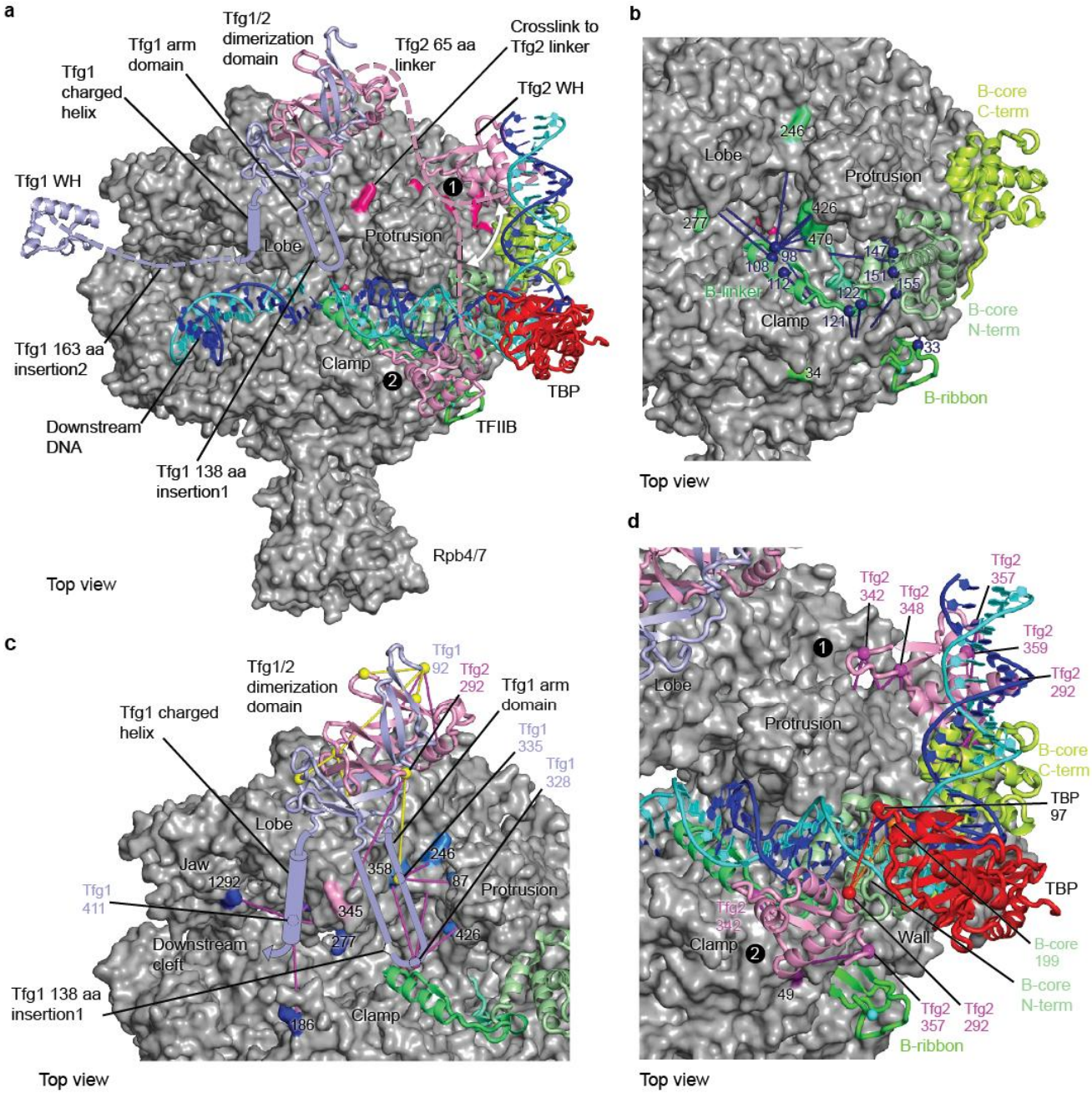

a
Tfg1/2 dimerization domain
Tfg1 arm domain
Tfg1 charged helix
Tfg2 65 aa linker
Crosslink to Tfg2 linker
Tfg2 WH
Tfg1 WH
Lobe
Protrusion
Clamp
TBP
TFIIB
Tfg1 163 aa insertion2
Downstream DNA
Tfg1 138 aa insertion1
Rpb4/7
Top view
b
B-core C-term
Lobe
Protrusion
B-linker
Clamp
B-core N-term
B-ribbon
Top view
c
Tfg1/2 dimerization domain
Tfg1 charged helix
Tfg1 arm domain
Lobe
Jaw
Protrusion
Downstream cleft
Tfg1 138 aa insertion1
Clamp
Top view
d
Lobe
Protrusion
B-core C-term
TBP
Clamp
Wall
B-core N-term
B-ribbon
Top view
e
Dimerization domain
Charged region
TFG1
N-term
Insertion1
Insertion2
WH
Arm
Charged helix
Insertion
predicted
Hs
Sc

Figure 3: Crosslinking-derived model of the yeast core ITC**.**
(A) Top view of the ITC, highlighting the locations of TBP (red), TFIIB (green), and TFIIF subunits Tfg1 (light blue) and Tfg2 (pink) on the Pol II surface. The TFIIF arm and charged helix elements are indicated as an antiparallel β-hairpin and α-helix, respectively. Alternative positions of the Tfg2 WH domain are indicated with black circled numbers (1, outside the cleft near upstream DNA as in the Pol II-TFIIF binary complex; 2, at the DNA bubble above the cleft). Mobile linkers are shown as dashed lines. **(B)** Pol II-TFIIB crosslinks (blue lines) viewed from the top as in (a) can be explained with the previously derived crystallographic TFIIB (B) core and ribbon domain locations[40,41]. **(C)** Location of the Pol II-TFIIF dimerization module (pink and yellow lines depict inter- and intra crosslinks, respectively). **(D)** The Tfg2 WH domain adopts two distinct locations. At position 1, the Tfg2 WH domain crosslinks to Pol II (pink lines), and at position 2, it crosslinks additionally to TFIIB and TBP (orange and red lines, respectively). **(E)** Domain organization of TFIIF subunit Tfg1 and location and conservation of the arm and charged helix elements. The charged helix was partially resolved in the X-ray structure[93] and is predicted to be longer[106]. Residues required for normal transcription initiation and elongation[102] are indicated as grey asterisks.

3.2 The RNA polymerase II CTD kinase complex subunit Ctk3 contains a non-canonical CTD-interacting domain

Wolfgang Mühlbacher, Andreas Mayer, Mai Sun, Michael Remmert, Alan C.M. Cheung, Jürgen Niesser, Johannes Soeding and Patrick Cramer. The RNA polymerase II CTD kinase complex subunit Ctk3 contains a non-canonical CTD-interacting domain. *Proteins* 2015; Accepted Article.

3.2.1 Architecture of Ctk3

In the final part of the thesis we investigate the structure of Ctk3. We predict a possible CTD-binding domain (CID) in the N-terminal region of Ctk3. The CID fold consists of eight α-helices in a right-handed superhelical arrangement and was thus far discovered in the three yeast proteins Pcf11, Nrd1, and Rtt103[77-79]. We use X-ray crystallography to show that Ctk3 indeed contains a CID fold. However, unlike other CID domains, this domain does not bind the Pol II CTD. We therefore refer to this domain as ‘CID-like domain’. In addition, we predict a three-helix bundle in the C-terminal region of Ctk3.

3.2.2 Prediction of a CID domain in Ctk3

We tried to predict a possible structure and function of Ctk3 by using HHblits[107], an iterative sequence search tool that represents both query and database sequences by profile hidden Markov models (HMMs). Such alignment methods are the most sensitive class of sequence search methods and the best choice for structure prediction and 3D homology modelling. We started with the protein sequence of Ctk3 and generated a profile HMM by performing two iterations of HHblits against the UniProt20 database, a clustered profile HMM database with 20% maximum pairwise sequence identity based on the UniProt sequence database (www.uniprot.org). The resulting profile HMM was then used for a final search against the PDB70 database, a clustered version of the protein data bank (PDB) filtered to 70% maximum pairwise sequence identity.

In this search, the best matches to Ctk3 were the three known CIDs present in the yeast genome. These reside in the transcription termination factors Rtt103, Pcf11, and Nrd1, and gave rise to predicted probabilities of 98%, 94%, and 93%, respectively. The matched alignment covered the entire CID, except for a weakly conserved C-terminal helix where the confidence values were lower. Based on these three alignments we generated a structural model of Ctk3 with the use of the MODELLER software[85]. The Ctk3 model showed a high conservation in the N-terminal five helices of the CID with an insertion between the first and second helix of approximately 20 amino acid residues in length.

The detected similarities strongly suggested that Ctk3 contains a CID, thus apparently representing a forth CID-containing protein in the yeast genome.

3.2.3 Crystal structure analysis of Ctk3 N-terminal domain

To clarify whether Ctk3 indeed contains the predicted CID fold, we prepared the putative CID-containing region of Ctk3 from *S. pombe* (Ctk3 (1-140), Figure 4) in recombinant form after overexpression in *E. coli* (Materials and Methods). The Ctk3 N-terminal domain was purified and crystallized by vapour diffusion (Materials and Methods). X-ray diffraction data were collected at the Swiss Light Source to a resolution of 2.0 Å (Table 9). The structure was determined by selenomethionine (SeMet) incorporation and multiwavelength anomalous diffraction (MAD) phasing, and the resulting model was refined to a free R-factor of 25.4% and showed very good stereochemistry (Materials and Methods). The overall fold of the Ctk3 N-terminal domain consists of eight α-helices in a right-handed superhelical arrangement (Figure 5A). As predicted, the fold closely resembles the known CID domains in the *S. cerevisiae* proteins Rtt103, Pcf11, and Nrd1, and in the human protein SCAF8 (Figure 5B). Superposition of the known CID structures revealed a difference in helix α2 of Ctk3 (residue 19-21), which is bent at its residue Pro27 towards helix α4 that flanks the proposed CTD-binding region of Ctk3 (Figure 5B).

Table 9: Crystallographic diffraction data and refinement statistics.

Data processing			
Space group	$P4_32_12$		
Unit cell dimensions (Å)	a=b= 51.3, c= 119.1		
	Peak	Inflection	Remote
Wavelength (Å)	0.97964	0.98012	0.97197
Resolution (Å)	47.1–2.00 (2.05–2.00)[a]	47.2–2.50 (2.57–2.50)	47.2–2.15 (2.21–2.15)
No. of observed reflections	157,934	80,837	126,934
No. of unique reflections	20,524	10,523	16,590
Completeness (%)	100 (99.9)	100 (100)	100 (100)
I/σ(I)	20.9 (4.9)	17.1 (5.1)	22.1 (5.0)
R_{merge} (%)	8.2 (32.8)	10.4 (32.9)	7.8 (32.4)
R_{meas} (%)	8.3 (48.8)	11.0 (46.7)	7.8 (47.4)
Wilson *B* value (Å^2)	27.8	33.3	31.4
Refinement			
Resolution range (Å)	47.12–2.0 (2.05–2.0)[a]		
No. of protein atoms	1194		
Amino acid residues	140		
Water molecules / ionic molecules (EPE ligands)[b]	67 / 2		
R-factor / free R-factor (%)	20.04 / 25.28		
B-factor average / macromolecules / ligands / solvent (Å^2)	33.20 / 31.50 / 74.00 / 42.70		
Ramachandran plot: Favoured / allowed / disallowed[c] (%)	99.3 / 100.0 / 0.0		
RMSD bond lengths / bond angles	0.008 Å / 1.047º		

[a]Values in parentheses refer to the highest resolution shell.
[b]4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid.
[c]As calculated using MolProbity [108].

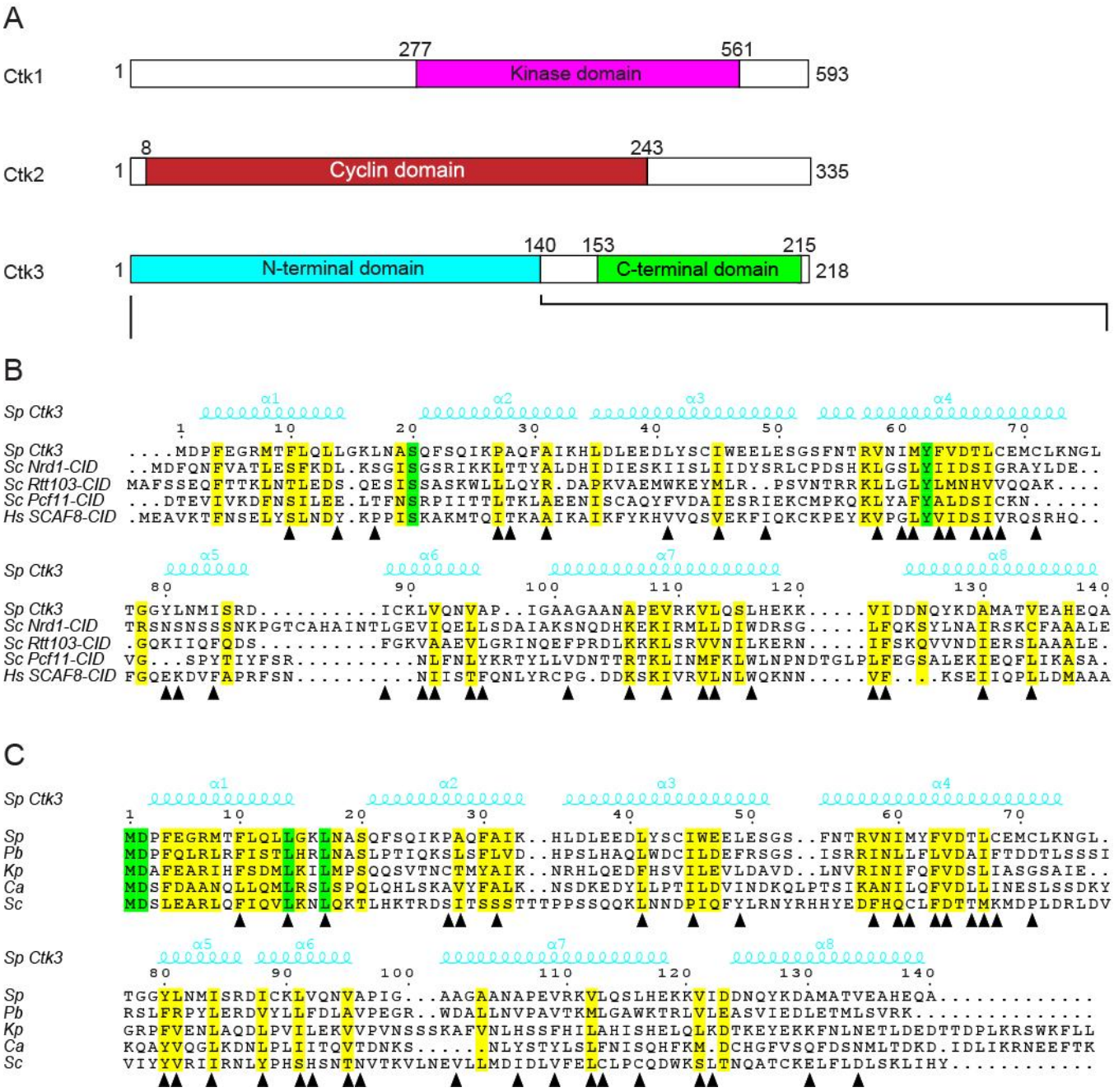


Figure 4: CTDK-I domains and conservation of Ctk3 N-terminal domain.

(**A**) Domain architecture of the three *S. pombe* CTDK-I subunits Ctk1 (Lsk1), Ctk2 (Lsc1), Ctk3 (Lsg1). (**B**) Structural alignment of *S. pombe* Ctk3 (Lsg1) N-terminal domain (1-140) to known CID domains. Conserved residues are colored yellow, invariant residues green. Sequences were ordered from highest (top) to lowest conservation, relative to *S. pombe* (*Sp, Schizosaccharomyces pombe; Sc, Saccharomyces. cerevisiae; Hs, Homo sapiens*). α-Helices are indicated above the alignment. Residues forming the hydrophobic core are depicted with black triangles.

(**C**) Alignment of Ctk3 N-terminal domains in different yeast species (*Pb, Paracoccidioides brasiliensis; Kp, Komagataella pastoris; Ca, Candida albicans*). Alignments were performed by ClustalW2 and the results were visualized by ESPript 3.0, using default values and "%Equivalent" as similarities depiction parameters [109,110].

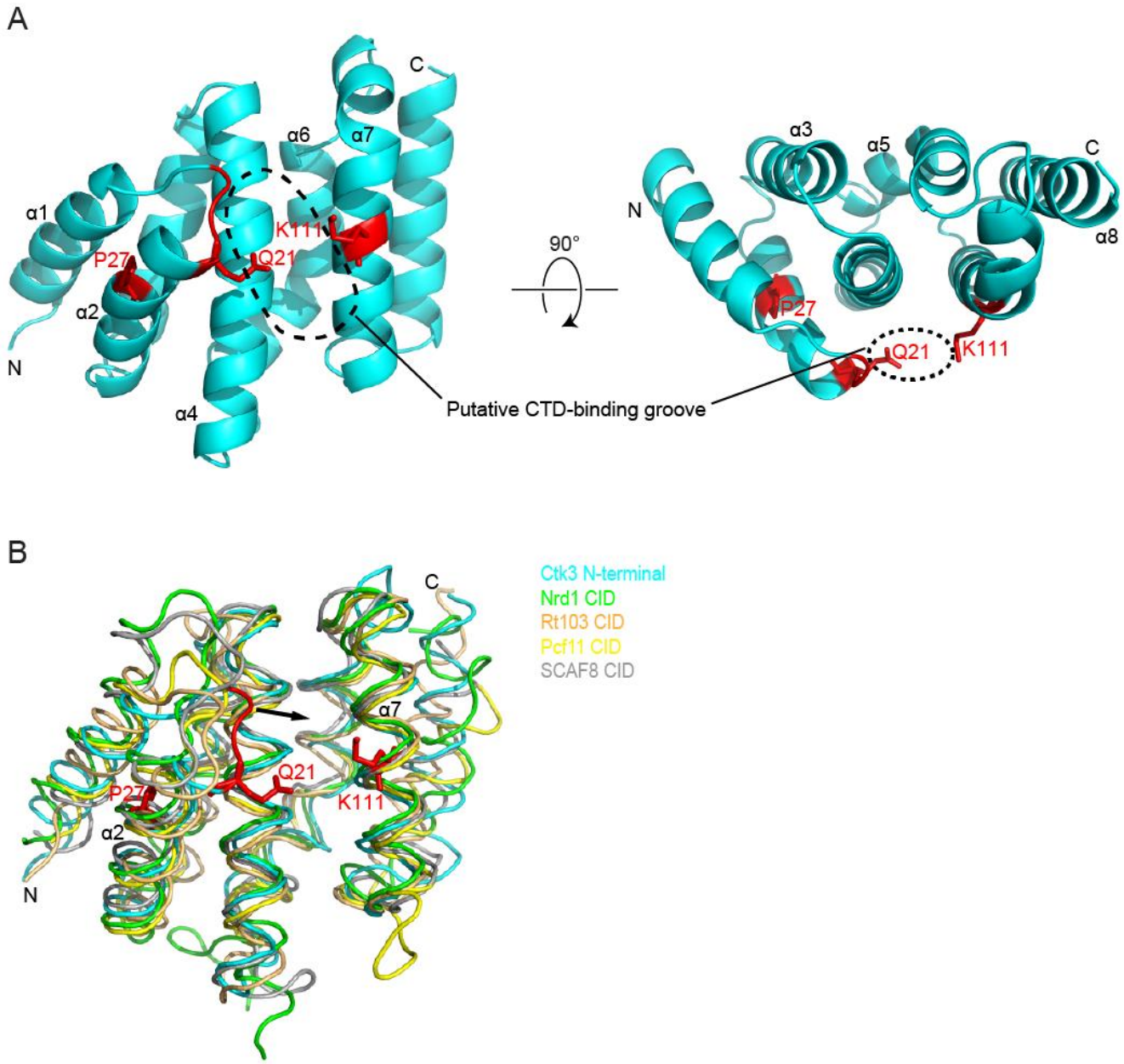


Figure 5: Crystal structure of the Ctk3 N-terminal domain at 2.0 Å resolution.
(**A**) Two views of a ribbon representation of the Ctk3 N-terminal domain, related by a 90 degree rotation around a horizontal axis. Residues that partially occupy the space of the surface region corresponding to the CTD-binding groove in canonical CID domains are depicted in red. The CTD-binding groove is indicated with black dashed lines and is corresponding to the CID in *Sc* Pcf11 [77]. (**B**) Superposition of known CID structures onto the Ctk3 N-terminal domain reveals its similarity to the CID fold. The key structural difference between the Ctk3 CID-like domain and the canonical CIDs is shown in red (part of helix α2 in Ctk3). The N- and C-termini of the protein domains are indicated.

3.2.4 The Ctk3 N-terminal domain has a non-canonical surface

To analyze whether the CID fold in Ctk3 also shares surface properties with canonical CID domains, we analyzed the conservation and electrostatic properties of the domain surface. We first colored the molecular surface of our structure according to conservation of amino acid residues over species (Figure 6A, alignment Figure 4C). Comparison with the surface of CID domains in Rtt103, Pcf11, and Nrd1 revealed that the Ctk3 N-terminal domain has a distinct surface conservation. Whereas canonical CID domains show a very high conservation of the CTD-binding groove between helices α2, α4 and α7, Ctk3 only shows weak and partial conservation in this region (Figure 6A). Moreover, Ctk3 residues Q21 and K111 partially obstruct the region corresponding to the CTD-binding groove in CID domains (Fig 2A).

The Ctk3 N-terminal domain also differs from canonical CID domains with respect to its surface charge distribution. To determine the charge distribution on the surface of the structure we used the APBS Tool via PyMOL (Figure 6B)[111]. For the canonical CID structures in Rtt103, Pcf11 and Nrd1, the conserved CTD-binding groove is positively charged, as required for binding to the negatively charged, phopshorylated CTD. In contrast, the corresponding region of the Ctk3 N-terminal domain is mainly positively charged. Taken together, analysis of the surface properties of the Ctk3 domain revealed that the putative CTD-binding groove was not conserved and differently charged, arguing against a CTD-binding function.

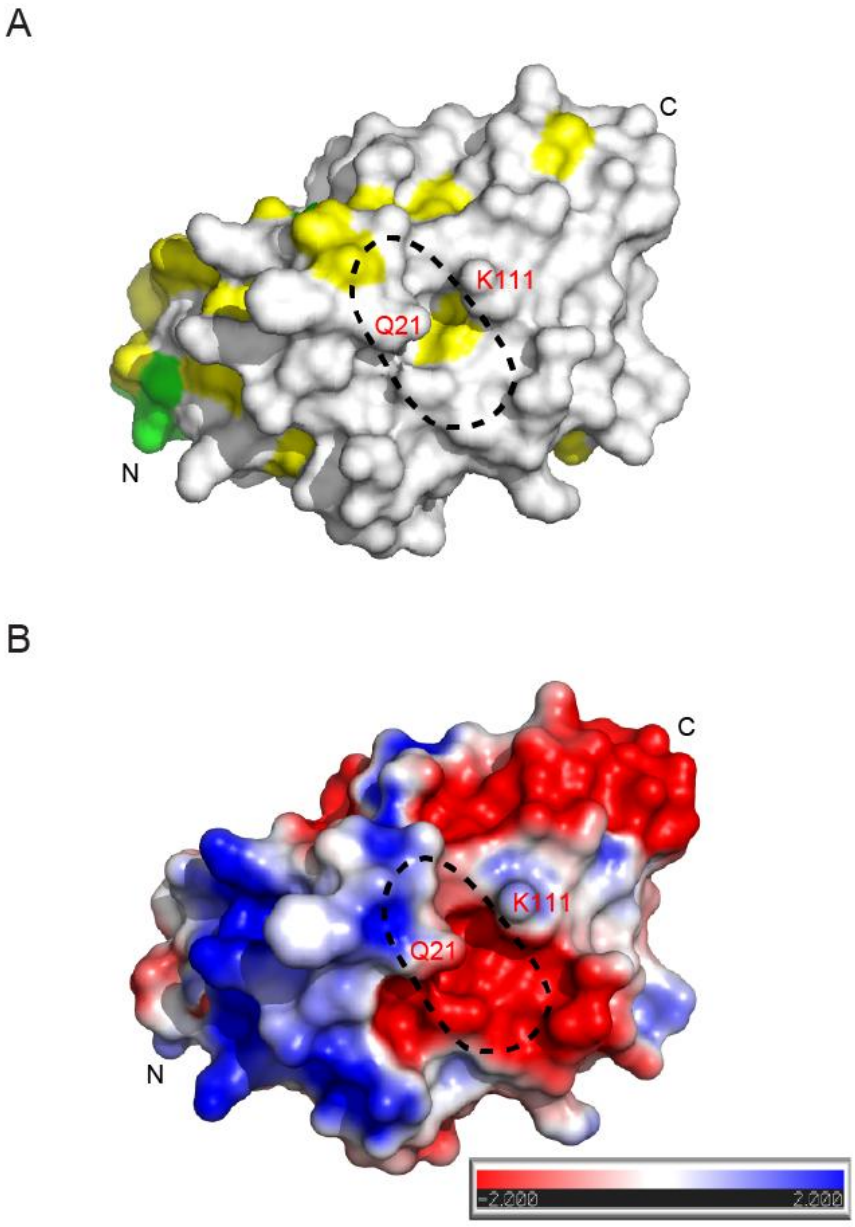


Figure 6: Surface properties of the Ctk3 N-terminal domain structure**.**

(**A**) Lack of a conserved CTD-binding groove in the Ctk3 N-terminal domain. The area corresponding to the CTD-binding groove in the CID domain of Pcf11 is indicated by a black dashed line [77]. Surface model generated by ESPript 3.0 [110] and colored according to conservation as in Figure 4C. The alignment in Figure 1C was used as an input for ESPript. Residues that partially occupy the space of the surface region corresponding to the CTD-binding groove in CID domains are labeled in red. (**B**) Electrostatic surface potentials of Ctk3 N-terminal domain. Positive and negative charges are in blue and red, respectively. Surface potentials were calculated with APBS [111].

3.2.5 The Ctk3 N-terminal domain does not bind CTD-derived petides

Structural analysis of Ctk3 revealed the fold of a CID domain, but also surface features that clearly differ from canonical CID domains and therefore question whether Ctk3 binds directly the CTD. To test CTD binding, we used a fluorescence-based phosphopeptide interaction assay in solution. We monitored changes in fluorescence anisotropy of fluorescently labeled CTD peptides upon addition of increasing amounts of purified Ctk3 N-terminal domain. The peptides were based on a di-heptad sequence (Figure 7A), which represents the functional unit of the Pol II CTD[112] and comprise different phosphorylation sites, mimicking various phosphorylation states of the CTD that are adopted during the transcription cycle.

We tested binding of CTD peptides phosphorylated at positions Tyr1, Ser2, Ser5, or Ser7, or phosphorylated at both Ser2 and Ser5, and also used a non-phosphorylated di-heptad peptide as a control (Figure 7A). We did not observe any binding of the Ctk3 N-terminal domain to any of the CTD-based peptides under the conditions tested (Figure 7B). As a positive control we reproduced binding of purified *S. pombe* Pcf11-CID to one of the Ser2-phosphorylated peptide, which demonstrated that the assay was working very well under our conditions. The data obtained from the positive control titration could be fit to the Hill equation as expected (Figure 7B). Taken together, we were unable to detect any binding of the Ctk3 N-terminal domain to any of the CTD-based peptides, strongly indicating that the domain does not directly bind the CTD *in vitro*.

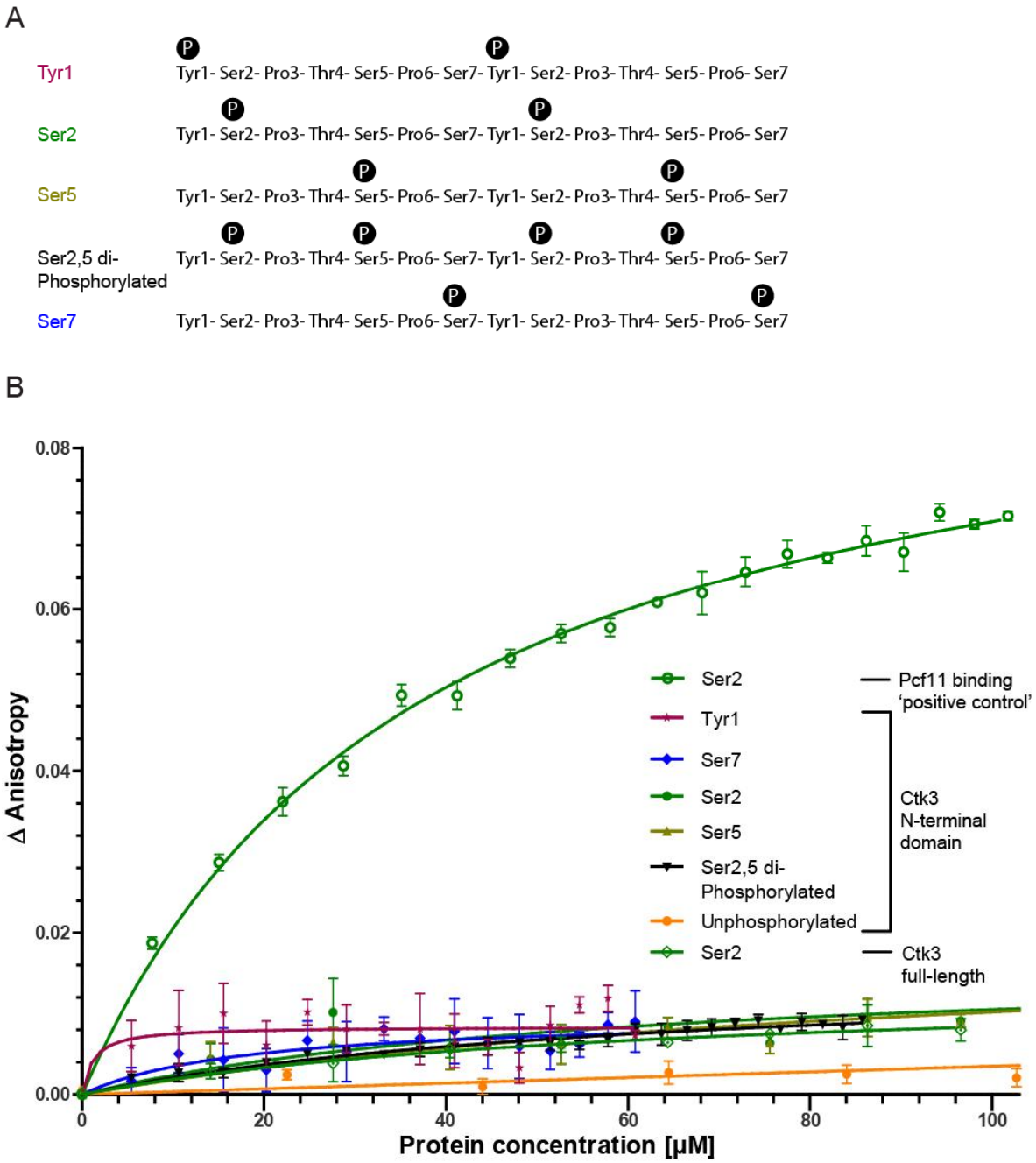


Figure 7: Ctk3 N-terminal domain does not bind CTD phosphopeptides.

(A) CTD-derived diheptad repeat peptides with phosphorylation positions indicated. From top to bottom, sequences of Tyr1-, Ser2-, Ser5-, Ser2,5- and Ser7-phosphorylated peptides are given. **(B)** Fluorescence anisotropy titration curves, using purified Ctk3 N-terminal domain (residues 1-140) and fluorescently labeled peptides shown in panel A. All measurements based on technical triplicate data. Plotted are the mean and standard deviation. As a positive control, we monitored binding of purified Pcf11-CID to Ser2-phosphorylated CTD peptide (green[113]). Ser2-phosphorylated binding was additionally measured with full-length Ctk3 (residues 1-218).

3.2.6 Ctk3 contains a highly conserved C-terminal bundle domain

In order to gain structural insights also into the C-terminal region of Ctk3, we returned to bioinformatic analysis. *S. pombe* Ctk3 residues 153-215 were annotated as a possible conserved domain in the PFAM database (http://pfam.xfam.org/). This region shows even higher sequence conservation than the N-terminal domain of Ctk3 (Figure 8A). We used the prediction algorithm PSIPRED[114] to assign potential secondary structure to this region. This revealed three helical stretches in this region that we named putative helices α1, α2, and α3 (Figure 8A). Throughout different yeast species, helices α2 and α3 were only weakly conserved, whereas helix α1 was highly conserved, showing 163 yeast species comprising helix α1 (Ctk3 residues 153-183) in a BLAST search (conservative expect threshold of 1e-6)[115]. We generated a three-dimensional model for the Ctk3 C-terminal region with the I-TASSER software[116], which uses structural templates from the Protein Data Bank. The model with the highest score was a three-helix bundle (Figure 8B). Analysis of the model revealed a conserved putative hydrophobic core, supporting the accuracy of the prediction, and a conserved surface patch on helix α1 that may be involved in interactions with other regions of CTDK-I.

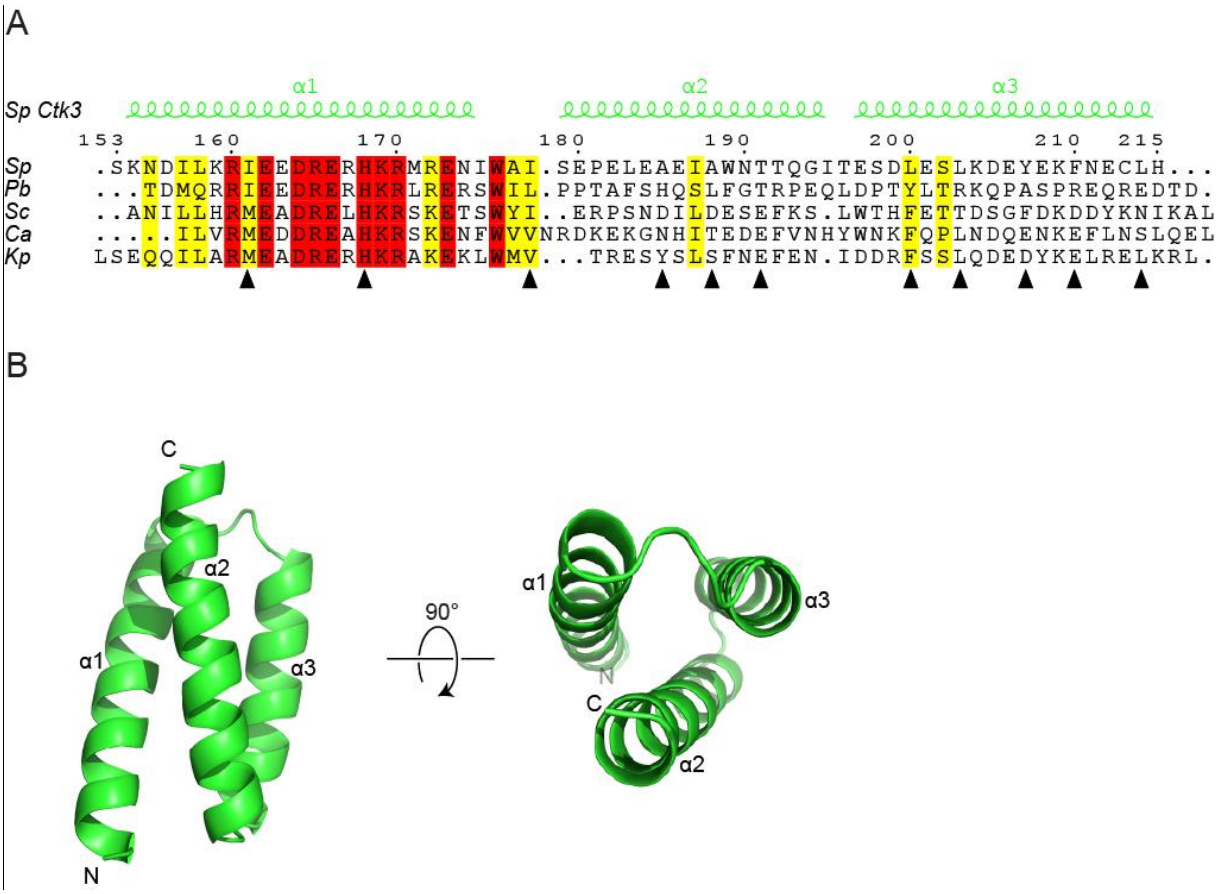


Figure 8: The Ctk3 C-terminal region contains a predicted helical bundle domain.

(**A**) Sequence alignment and secondary structure prediction of Ctk3 C-terminal region comparing different yeast species. Conserved and invariant residues colored yellow and red, respectively. Sequences ordered from highest conservation (on top) to lowest, relative to *S. pombe* Ctk3 C-terminal domain (153-215). (*Sp*, *Schizosaccharomyces pombe*; *Pb*, *Paracoccidioides brasiliensis*; *Sc*, *Saccharomyces cerevisiae*; *Ca*, *Candida albicans*; *Kp*, *Komagataella pastoris*). α-Helices are indicated above the alignment. Residues forming the hydrophobic core are depicted with black triangles. (**B**) Two views of the predicted Ctk3 C-terminal domain model obtained with I-TASSER [116]. The N- and C-termini of the protein domain are indicated.

3.2.7 Conclusion

Here we predicted that Ctk3 consists of an N-terminal CTD-interacting domain (CID) and a C-terminal helical bundle domain. X-ray crystallography confirmed the N-terminal CID fold despite the low sequence conservation, consistent with the general observation that the three-dimensional structure of proteins is more conserved than their sequence. The Ctk3 N-terminal domain however differs from canonical CID domains in its surface conservation and charge distribution. In particular, the putative CTD-binding groove between helices α2, α4, and α7 is not conserved and lacks positively charged residues. Consistent with this, the N-terminal domain of Ctk3 failed to bind CTD phosphopeptides in solution. The CID fold resembles the fold of VHS (Vsp27p/Hrs/STAM) domains[117], as detected by DALI[118]. Thus the N-terminal domain in Ctk3 could have arisen from a CID domain that lost its CTD-binding surface groove, or from a VHS domain of unknown function.

The C-terminal domain of Ctk3 is more conserved than its N-terminal domain, especially at its putative α1 helix. It remains unclear whether the C-terminal domain of Ctk3 associates stably with the N-terminal domain or whether it forms a distinct domain connected with a short linker. Consistent with a two-domain, flexible structure, our attempts to crystallize full-length Ctk3 failed. The predicted C-terminal domain is essential for cell growth in *S. cerevisiae,* and a truncation that removes most of the C-terminal domain destabilizes the interaction of Ctk3 with Ctk2 *in vitro*, whereas interaction with Ctk1 was maintained[67]. From these results and considerations, a topological model of CTDK-I emerges with Ctk1 and Ctk2 forming a canonical kinase-cyclin pair, as observed for pTEF-b[119] and Ctk3 being anchored to Ctk2 via its C-terminal helical bundle domain (Figure 9). In this model, the newly found N-terminal domain in Ctk3 remains available for other, currently unknown functions.

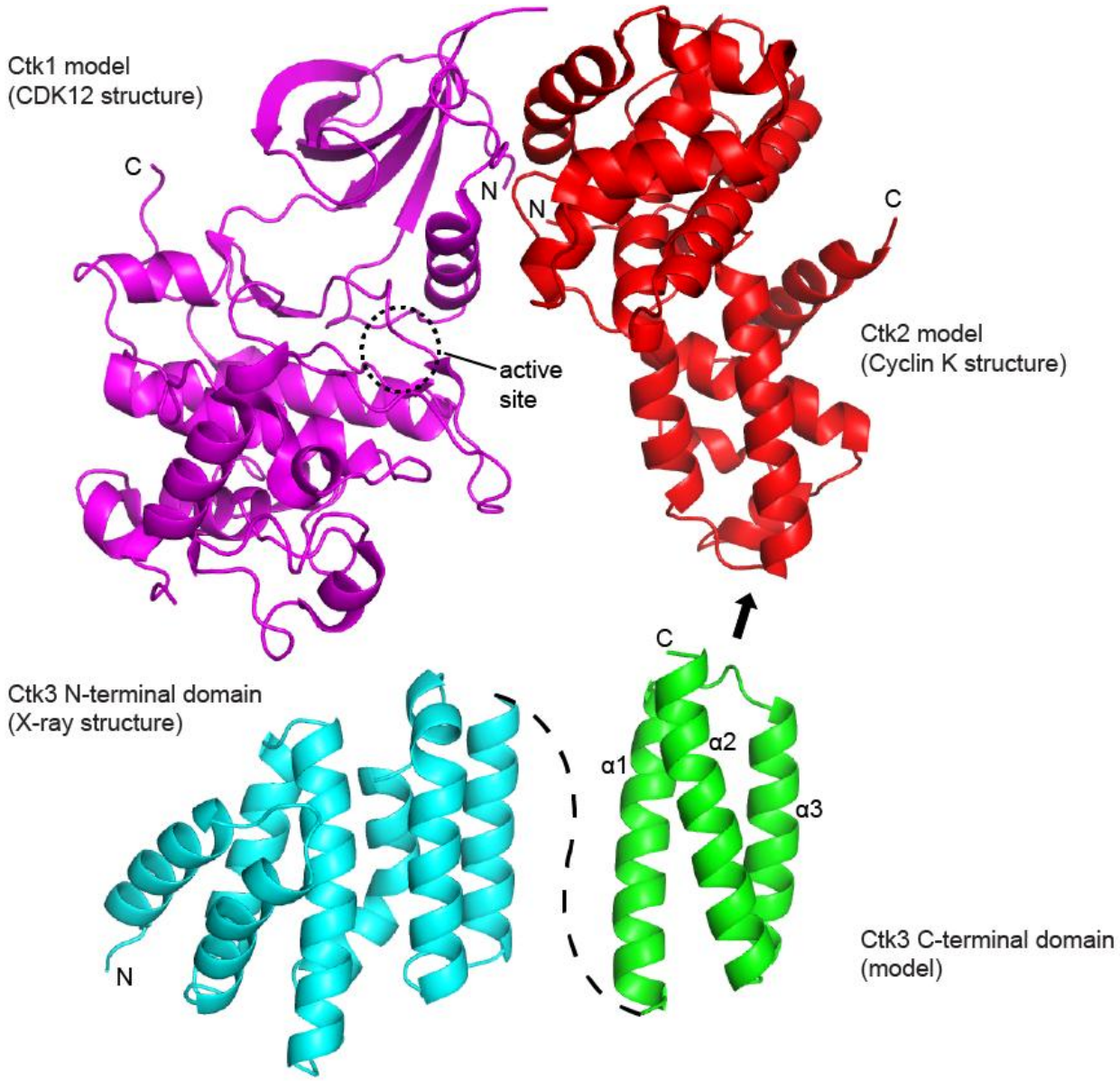


Figure 9: Overview of CTDK-I architecture.

Relative size of the structure of mammalian Ctk1-Ctk2 counterpart CDK12–Cyclin K (PDB-code 4UN0) and *S. pombe* Ctk3 N-terminal domain structure (this work) and predicted C-terminal domain (this work). Arrow indicate putative interaction of the C-terminal domains with Ctk2 (compare text). The active center of the CDK12 kinase structure is indicated with a black dashed circle. The N- and C-termini of the proteins are labeled.

4 Outlook

4.1 The architecture of Pol II initiation complexes

Most of the fundamental mechanisms of transcriptional regulation are conserved from yeast to human, assigning the *S. cerevisiae* model system a central role in understanding how gene expression is controlled in all eukaryotes[13]. Protein crosslinking of the core Pol II initiation complex underlines this statement even more as it shows the conserved architecture. Our proposed model of the conserved architecture of the ITC was further validated by a later study that combined crosslinking with single-particle cryo-EM[120]. Still, many questions regarding the topology of the core initiation complex remain: Where do the Tfg1 insertion and the WH domains of TFIIF reside, and does the insertion domain get structured due to binding to Pol II? Are the Tfg1 arm and the charged helix contacting the DNA strand, and what function does the negative charge of the charged helix have? Do both positions of the Tfg2 WH also exist *in vivo*, and does the Tfg2 WH bind upstream promoter DNA? Further experiments comprising both structural and biochemical methods are needed to address these questions. On a structural level, the highest potential for success will be the combination of X-ray crystallography, single-particle cryo-EM, mass spectrometric approaches (protein/DNA crosslinking-MS, native-MS, and hydrogen-deuterium exchange) and computational based methods. Combining diverse structural restrains deriving from different methods is challenging, but will be rewarding, as new insights can be gained and misinterpretations can be minimized.

To gain further knowledge on how initiation is regulated, the architecture of more complete initiation complexes need to be addressed. Therefore, the addition of GTFs (transcription factor-IIA, -IID, -IIE, IIH) and the Mediator complex to the core initiation complex are the next logical steps. As crystallization of such big and potentially flexible complexes can be difficult, single-particle cryo-EM starts to play an increasingly important role, especially with recent improvements such as the (introduction of the direct detector[121], marking a milestone in generating high resolution data.

Based on these high resolution restrains, target sites could be identified for mutational analysis, addressing central questions, like: What are the functions of the GTFs modulated; how initiation complexes can form on TATA-less promotors; which additional, unknown factors bind to the initiation complex or how are coactivators placed within pre-initiation complexes.

4.2 The transcription elongation promoting CTDK-I complex

For future work, the arrangement of the CTDK-I complex needs to be further investigated to reveal the interplay of the three subunits, which is unique among CDK complexes. The heterodimer Ctk2-Ctk3 represents a promising target for structural determination by X-ray crystallography, as it was shown to form a stable complex in gel filtration chromatography[67].

A high resolution structure of Ctk2-Ctk3 could provide insights which address serval central questions: 1) how does the CTDK-I complex approach the Pol II CTD? 2) Which mechanism does allow Ctk3 to act as a co-cyclin and why can't Ctk2 fulfill the cyclin functions on its own? 3) Does the non-canonical CID fold of Ctk3 fulfill an unknown purpose? and 4) Does the highly conserved C-terminal domain of Ctk3 interact with Ctk2 as proposed[67].

One general biological question regarding the CTDK-I complex is its role in connecting gene transcription with cell cycle events. In contrast to most CDKs which bind different cyclins throughout the cell cycle[122], the binding of Ctk1 is specific to cyclin Ctk2 and the co-cyclin Ctk3[67]. This unique difference compared with other CDKs could lead to the discovery of a new mechanism in the regulation of CTKs.

References

1 Weinzierl, R. *Mechanisms of Gene Expression: Structure, Function and Evolution of the Basal Transcriptional Machinery*. (World Scientific Publishing Company).

2 Vannini, A. & Cramer, P. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Mol. Cell* **45**, 439-446 (2012).

3 Cramer, P. *et al.* Structure of eukaryotic RNA polymerases. *Annual review of biophysics* **37**, 337-352, doi:10.1146/annurev.biophys.37.032807.130008 (2008).

4 Jacquier, A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet.* **10**, 833-844, doi:10.1038/nrg2683 (2009).

5 Asin-Cayuela, J. & Gustafsson, C. M. Mitochondrial transcription and its regulation in mammalian cells. *Trends Biochem. Sci.* **32**, 111-117, doi:10.1016/j.tibs.2007.01.003 (2007).

6 Cheetham, G. M. & Steitz, T. A. Structure of a transcribing T7 RNA polymerase initiation complex. *Science* **286**, 2305-2309 (1999).

7 Heidemann, M., Hintermair, C., Voss, K. & Eick, D. Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. *Biochim. Biophys. Acta* **1829**, 55-62, doi:10.1016/j.bbagrm.2012.08.013 (2013).

8 Eick, D. & Geyer, M. The RNA polymerase II carboxy-terminal domain (CTD) code. *Chemical reviews* **113**, 8456-8490, doi:10.1021/cr400071f (2013).

9 Hahn, S. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat. Struct. Mol. Biol.* **11**, 394-403, doi:10.1038/nsmb763 (2004).

10 Svejstrup, J. Q. The RNA polymerase II transcription cycle: cycling through chromatin. *Biochim. Biophys. Acta* **1677**, 64-73, doi:10.1016/j.bbaexp.2003.10.012 (2004).

11 Liu, X., Bushnell, D. A., Wang, D., Calero, G. & Kornberg, R. D. Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism. *Science* **327**, 206-209, doi:10.1126/science.1182015 (2010).

12 Shandilya, J. & Roberts, S. G. The transcription cycle in eukaryotes: from productive initiation to RNA polymerase II recycling. *Biochim. Biophys. Acta* **1819**, 391-400, doi:10.1016/j.bbagrm.2012.01.010 (2012).

13 Hahn, S. & Young, E. T. Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics* **189**, 705-736, doi:10.1534/genetics.111.127019 (2011).

14 Sikorski, T. W. & Buratowski, S. The basal initiation machinery: beyond the general transcription factors. *Curr. Opin. Cell Biol.* **21**, 344-351, doi:10.1016/j.ceb.2009.03.006 (2009).

15 Liu, X., Bushnell, D. A. & Kornberg, R. D. RNA polymerase II transcription: structure and mechanism. *Biochim. Biophys. Acta* **1829**, 2-8, doi:10.1016/j.bbagrm.2012.09.003 (2013).

16 Biddick, R. & Young, E. T. Yeast mediator and its role in transcriptional regulation. *C. R. Biol.* **328**, 773-782, doi:10.1016/j.crvi.2005.03.004 (2005).

17 Vojnic, E. *et al.* Structure and VP16 binding of the Mediator Med25 activator interaction domain. *Nat. Struct. Mol. Biol.* **18**, 404-409, doi:10.1038/nsmb.1997 (2011).

18 Lee, T. I. & Young, R. A. Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.* **34**, 77-137, doi:10.1146/annurev.genet.34.1.77 (2000).

19 Thomas, M. C. & Chiang, C. M. The general transcription machinery and general cofactors. *Crit. Rev. Biochem. Mol. Biol.* **41**, 105-178, doi:10.1080/10409230600648736 (2006).

20 SK, B. - X-ray crystallographic studies of eukaryotic transcription initiation factors. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **351**, 483-489 (1996).

21 Buratowski, S. Progression through the RNA polymerase II CTD cycle. *Mol. Cell* **36**, 541-546, doi:10.1016/j.molcel.2009.10.019 (2009).

22 Zhou, Q., Li, T. & Price, D. H. RNA polymerase II elongation control. *Annu. Rev. Biochem.* **81**, 119-143, doi:10.1146/annurev-biochem-052610-095910 (2012).

23 Nechaev, S. & Adelman, K. Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. *Biochim. Biophys. Acta* **1809**, 34-45, doi:10.1016/j.bbagrm.2010.11.001 (2011).

24 Saunders, A., Core, L. J. & Lis, J. T. Breaking barriers to transcription elongation. *Nat. Rev. Mol. Cell Biol.* **7**, 557-567, doi:10.1038/nrm1981 (2006).

25 Kettenberger, H., Armache, K. J. & Cramer, P. Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage. *Cell* **114**, 347-357 (2003).

26 Cheung, A. C. & Cramer, P. Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature* **471**, 249-253, doi:10.1038/nature09785 (2011).

27 Aranda, A. & Proudfoot, N. Transcriptional termination factors for RNA polymerase II in yeast. *Mol. Cell* **7**, 1003-1011 (2001).

28 Lykke-Andersen, S., Mapendano, C. K. & Jensen, T. H. An ending is a new beginning: transcription termination supports re-initiation. *Cell cycle (Georgetown, Tex.)* **10**, 863-865 (2011).

29 Mischo, H. E. & Proudfoot, N. J. Disengaging polymerase: terminating RNA polymerase II transcription in budding yeast. *Biochim. Biophys. Acta* **1829**, 174-185, doi:10.1016/j.bbagrm.2012.10.003 (2013).

30 Yudkovsky, N., Ranish, J. A. & Hahn, S. A transcription reinitiation intermediate that is stabilized by activator. *Nature* **408**, 225-229, doi:10.1038/35041603 (2000).

31 Buratowski, S., Hahn, S., Guarente, L. & Sharp, P. A. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* **56**, 549-561 (1989).

32 Grunberg, S. & Hahn, S. Structural insights into transcription initiation by RNA polymerase II. *Trends Biochem. Sci.* **38**, 603-611, doi:10.1016/j.tibs.2013.09.002 (2013).

33 Vannini, A. & Cramer, P. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Molecular cell* **45**, 439-446, doi:10.1016/j.molcel.2012.01.023 (2012).

34 Chen, H. T., Warfield, L. & Hahn, S. The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex. *Nat. Struct. Mol. Biol.* **14**, 696-703, doi:10.1038/nsmb1272 (2007).

35 Eichner, J., Chen, H. T., Warfield, L. & Hahn, S. Position of the general transcription factor TFIIF within the RNA polymerase II transcription preinitiation complex. *EMBO J.* **29**, 706-716, doi:10.1038/emboj.2009.386 (2010).

36 Grunberg, S., Warfield, L. & Hahn, S. Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening. *Nat. Struct. Mol. Biol.* **19**, 788-796, doi:10.1038/nsmb.2334 (2012).

37 He, Y., Fang, J., Taatjes, D. J. & Nogales, E. Structural visualization of key steps in human transcription initiation. *Nature* **495**, 481-486, doi:10.1038/nature11991 (2013).

38 Murakami, K. *et al.* Architecture of an RNA polymerase II transcription pre-initiation complex. *Science (New York, N.Y.)* **342**, 1238724, doi:10.1126/science.1238724 (2013).

39 Cheung, A. C. & Cramer, P. A movie of RNA polymerase II transcription. *Cell* **149**, 1431-1437, doi:10.1016/j.cell.2012.06.006 (2012).

40 Kostrewa, D. *et al.* RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature* **462**, 323-330, doi:10.1038/nature08548 (2009).

41 Sainsbury, S., Niesser, J. & Cramer, P. Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. *Nature* **493**, 437-440, doi:10.1038/nature11715 (2013).

42 Chen, Z. A. *et al.* Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **29**, 717-726, doi:10.1038/emboj.2009.401 (2010).

43 Napolitano, G., Lania, L. & Majello, B. RNA polymerase II CTD modifications: how many tales from a single tail. *Journal of cellular physiology* **229**, 538-544, doi:10.1002/jcp.24483 (2014).

44 Corden, J. L. Transcription. Seven ups the code. *Science* **318**, 1735-1736, doi:10.1126/science.1152624 (2007).

45 Egloff, S., Dienstbier, M. & Murphy, S. Updating the RNA polymerase CTD code: adding gene-specific layers. *Trends Genet.* **28**, 333-341, doi:10.1016/j.tig.2012.03.007 (2012).

46 Schroeder, S. C., Schwer, B., Shuman, S. & Bentley, D. Dynamic association of capping enzymes with transcribing RNA polymerase II. *Genes Dev.* **14**, 2435-2440 (2000).

47 Cho, E. J., Kobor, M. S., Kim, M., Greenblatt, J. & Buratowski, S. Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain. *Genes Dev.* **15**, 3319-3329, doi:10.1101/gad.935901 (2001).

48 Jona, G., Wittschieben, B. O., Svejstrup, J. Q. & Gileadi, O. Involvement of yeast carboxy-terminal domain kinase I (CTDK-I) in transcription elongation in vivo. *Gene* **267**, 31-36 (2001).

49 Lee, J. M. & Greenleaf, A. L. Modulation of RNA polymerase II elongation efficiency by C-terminal heptapeptide repeat domain kinase I. *J. Biol. Chem.* **272**, 10990-10993 (1997).

50 Ho, C. K. & Shuman, S. Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme. *Mol. Cell* **3**, 405-411 (1999).

51 Cismowski, M. J., Laff, G. M., Solomon, M. J. & Reed, S. I. KIN28 encodes a C-terminal domain kinase that controls mRNA transcription in Saccharomyces cerevisiae but lacks cyclin-dependent kinase-activating kinase (CAK) activity. *Mol. Cell. Biol.* **15**, 2983-2992 (1995).

52 Feaver, W. J., Svejstrup, J. Q., Henry, N. L. & Kornberg, R. D. Relationship of CDK-activating kinase and RNA polymerase II CTD kinase TFIIH/TFIIK. *Cell* **79**, 1103-1109 (1994).

53 Keogh, M. C., Cho, E. J., Podolny, V. & Buratowski, S. Kin28 is found within TFIIH and a Kin28-Ccl1-Tfb3 trimer complex with differential sensitivities to T-loop phosphorylation. *Mol. Cell. Biol.* **22**, 1288-1297 (2002).

54 Borggrefe, T., Davis, R., Erdjument-Bromage, H., Tempst, P. & Kornberg, R. D. A complex of the Srb8, -9, -10, and -11 transcriptional regulatory proteins from yeast. *J. Biol. Chem.* **277**, 44202-44207, doi:10.1074/jbc.M207195200 (2002).

55 Hengartner, C. J. *et al.* Temporal regulation of RNA polymerase II by Srb10 and Kin28 cyclin-dependent kinases. *Mol. Cell* **2**, 43-53 (1998).

56 Marshall, N. F. & Price, D. H. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J. Biol. Chem.* **270**, 12335-12338 (1995).

57 Cheng, B. & Price, D. H. Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. *J. Biol. Chem.* **282**, 21901-21912, doi:10.1074/jbc.M702936200 (2007).

58 Keogh, M. C., Podolny, V. & Buratowski, S. Bur1 kinase is required for efficient transcription elongation by RNA polymerase II. *Mol. Cell. Biol.* **23**, 7005-7018 (2003).

59 Zhou, K., Kuo, W. H., Fillingham, J. & Greenblatt, J. F. Control of transcriptional elongation and cotranscriptional histone modification by the yeast BUR kinase substrate Spt5. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 6956-6961, doi:10.1073/pnas.0806302106 (2009).

60 Liu, Y. *et al.* Phosphorylation of the transcription elongation factor Spt5 by yeast Bur1 kinase stimulates recruitment of the PAF complex. *Mol. Cell. Biol.* **29**, 4852-4863, doi:10.1128/mcb.00609-09 (2009).

61 Qiu, H., Hu, C. & Hinnebusch, A. G. Phosphorylation of the Pol II CTD by KIN28 enhances BUR1/BUR2 recruitment and Ser2 CTD phosphorylation near promoters. *Mol. Cell* **33**, 752-762, doi:10.1016/j.molcel.2009.02.018 (2009).

62 Wood, A., Schneider, J., Dover, J., Johnston, M. & Shilatifard, A. The Bur1/Bur2 complex is required for histone H2B monoubiquitination by Rad6/Bre1 and histone methylation by COMPASS. *Mol. Cell* **20**, 589-599, doi:10.1016/j.molcel.2005.09.010 (2005).

63 Mosley, A. L. *et al.* Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. *Mol. Cell* **34**, 168-178, doi:10.1016/j.molcel.2009.02.025 (2009).

64 Krishnamurthy, S., He, X., Reyes-Reyes, M., Moore, C. & Hampsey, M. Ssu72 Is an RNA polymerase II CTD phosphatase. *Mol. Cell* **14**, 387-394 (2004).

65 Yao, S., Neiman, A. & Prelich, G. BUR1 and BUR2 encode a divergent cyclin-dependent kinase-cyclin complex important for transcription in vivo. *Mol. Cell. Biol.* **20**, 7080-7087 (2000).

66 Bartkowiak, B. *et al.* CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes Dev.* **24**, 2303-2316, doi:10.1101/gad.1968210 (2010).

67 Hautbergue, G. & Goguel, V. Activation of the cyclin-dependent kinase CTDK-I requires the heterodimerization of two unstable subunits. *J. Biol. Chem.* **276**, 8005-8013, doi:10.1074/jbc.M010162200 (2001).

68 Sterner, D. E., Lee, J. M., Hardin, S. E. & Greenleaf, A. L. The yeast carboxyl-terminal repeat domain kinase CTDK-I is a divergent cyclin-cyclin-dependent kinase complex. *Mol. Cell. Biol.* **15**, 5716-5724 (1995).

69 Karagiannis, J., Bimbo, A., Rajagopalan, S., Liu, J. & Balasubramanian, M. K. The nuclear kinase Lsk1p positively regulates the septation initiation network and promotes the successful completion of cytokinesis in response to perturbation of the actomyosin ring in Schizosaccharomyces pombe. *Molecular biology of the cell* **16**, 358-371, doi:10.1091/mbc.E04-06-0502 (2005).

70 Karagiannis, J. & Balasubramanian, M. K. A cyclin-dependent kinase that promotes cytokinesis through modulating phosphorylation of the carboxy terminal domain of the RNA Pol II Rpb1p sub-unit. *PloS one* **2**, e433, doi:10.1371/journal.pone.0000433 (2007).

71 Sukegawa, Y., Yamashita, A. & Yamamoto, M. The fission yeast stress-responsive MAPK pathway promotes meiosis via the phosphorylation of Pol II CTD in response to environmental and feedback cues. *PLoS Genet.* **7**, e1002387, doi:10.1371/journal.pgen.1002387 (2011).

72 Lee, J. M. & Greenleaf, A. L. CTD kinase large subunit is encoded by CTK1, a gene required for normal growth of Saccharomyces cerevisiae. *Gene Expr* **1**, 149-167 (1991).

73 Viladevall, L. *et al.* TFIIH and P-TEFb coordinate transcription with capping enzyme recruitment at specific genes in fission yeast. *Mol. Cell* **33**, 738-751, doi:10.1016/j.molcel.2009.01.029 (2009).

74 Coudreuse, D. *et al.* A gene-specific requirement of RNA polymerase II CTD phosphorylation for sexual differentiation in S. pombe. *Curr. Biol.* **20**, 1053-1064, doi:10.1016/j.cub.2010.04.054 (2010).

75 Hautbergue, G. & Goguel, V. The yeast C-type cyclin Ctk2p is phosphorylated and rapidly degraded by the ubiquitin-proteasome pathway. *Mol. Cell. Biol.* **19**, 2527-2534 (1999).

76 Lidschreiber, M., Leike, K. & Cramer, P. Cap completion and C-terminal repeat domain kinase recruitment underlie the initiation-elongation transition of RNA polymerase II. *Mol. Cell. Biol.* **33**, 3805-3816, doi:10.1128/mcb.00361-13 (2013).

77 Meinhart, A. & Cramer, P. Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. *Nature* **430**, 223-226, doi:10.1038/nature02679 (2004).

78 Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S. & Meinhart, A. The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. *Nat. Struct. Mol. Biol.* **15**, 795-804, doi:10.1038/nsmb.1468 (2008).

79 Lunde, B. M. *et al.* Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain. *Nat. Struct. Mol. Biol.* **17**, 1195-1201, doi:10.1038/nsmb.1893 (2010).

80 Sydow, J. F. *et al.* Structural basis of transcription: mismatch-specific fidelity mechanisms and paused RNA polymerase II with frayed RNA. *Mol. Cell* **34**, 710-721, doi:10.1016/j.molcel.2009.06.002 (2009).

81 Treutlein, B. *et al.* Dynamic architecture of a minimal RNA polymerase II open promoter complex. *Molecular cell* **46**, 136-146, doi:10.1016/j.molcel.2012.02.008 (2012).

82 Jennebach, S., Herzog, F., Aebersold, R. & Cramer, P. Crosslinking-MS analysis reveals RNA polymerase I domain architecture and basis of rRNA cleavage. *Nucleic acids research* **40**, 5591-5601, doi:10.1093/nar/gks220 (2012).

83 Herzog, F. *et al.* Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science* **337**, 1348-1352, doi:10.1126/science.1221483 (2012).

84 Walzthoeni, T. *et al.* False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nat. Methods* **9**, 901-903, doi:10.1038/nmeth.2103 (2012).

85 Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815, doi:10.1006/jmbi.1993.1626 (1993).

86 Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *Journal of Applied Crystallography* **26**, 795-800, doi:doi:10.1107/S0021889893005588 (1993).

87 Terwilliger, T. SOLVE and RESOLVE: automated structure solution, density modification and model building. *Journal of synchrotron radiation* **11**, 49-52 (2004).

88 Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature protocols* **3**, 1171-1179, doi:10.1038/nprot.2008.91 (2008).

89 Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta crystallographica. Section D, Biological crystallography* **60**, 2126-2132, doi:10.1107/s0907444904019158 (2004).

90 Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta crystallographica. Section D, Biological crystallography* **66**, 213-221, doi:10.1107/s0907444909052925 (2010).

91 Engel, C., Sainsbury, S., Cheung, A. C., Kostrewa, D. & Cramer, P. RNA polymerase I structure and transcription regulation. *Nature* **502**, 650-655, doi:10.1038/nature12712 (2013).

92 Armache, K. J., Mitterweger, S., Meinhart, A. & Cramer, P. Structures of complete RNA polymerase II and its subcomplex, Rpb4/7. *J. Biol. Chem.* **280**, 7131-7134, doi:10.1074/jbc.M413038200 (2005).

93 Gaiser, F., Tan, S. & Richmond, T. J. Novel dimerization fold of RAP30/RAP74 in human TFIIF at 1.7 A resolution. *J. Mol. Biol.* **302**, 1119-1127, doi:10.1006/jmbi.2000.4110 (2000).

94 Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365**, 512-520, doi:10.1038/365512a0 (1993).

95 Chasman, D. I., Flaherty, K. M., Sharp, P. A. & Kornberg, R. D. Crystal structure of yeast TATA-binding protein and model for interaction with DNA. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 8174-8178 (1993).

96 Brent, M. M., Anand, R. & Marmorstein, R. Structural basis for DNA recognition by FoxO1 and its regulation by posttranslational modification. *Structure (London, England : 1993)* **16**, 1407-1416, doi:10.1016/j.str.2008.06.013 (2008).

97 Kilpatrick, A. M., Koharudin, L. M., Calero, G. A. & Gronenborn, A. M. Structural and binding studies of the C-terminal domains of yeast TFIIF subunits Tfg1 and Tfg2. *Proteins*, doi:10.1002/prot.23217 (2011).

98 Tan, S., Conaway, R. C. & Conaway, J. W. Dissection of transcription factor TFIIF functional domains required for initiation and elongation. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 6042-6046 (1995).

99 Yan, Q., Moreland, R. J., Conaway, J. W. & Conaway, R. C. Dual roles for transcription factor IIF in promoter escape by RNA polymerase II. *J. Biol. Chem.* **274**, 35668-35675 (1999).

100 Cabart, P., Ujvari, A., Pal, M. & Luse, D. S. Transcription factor TFIIF is not required for initiation by RNA polymerase II, but it is essential to stabilize transcription factor TFIIB in early elongation complexes. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 15786-15791, doi:10.1073/pnas.1104591108 (2011).

101 Khaperskyy, D. A., Ammerman, M. L., Majovski, R. C. & Ponticelli, A. S. Functions of Saccharomyces cerevisiae TFIIF during transcription start site utilization. *Mol. Cell. Biol.* **28**, 3757-3766, doi:10.1128/mcb.02272-07 (2008).

102 Funk, J. D., Nedialkov, Y. A., Xu, D. & Burton, Z. F. A key role for the alpha 1 helix of human RAP74 in the initiation and elongation of RNA chains. *J. Biol. Chem.* **277**, 46998-47003, doi:10.1074/jbc.M206249200 (2002).

103 Zhang, C., Zobeck, K. L. & Burton, Z. F. Human RNA polymerase II elongation in slow motion: role of the TFIIF RAP74 alpha1 helix in nucleoside triphosphate-driven translocation. *Mol. Cell. Biol.* **25**, 3583-3595, doi:10.1128/mcb.25.9.3583-3595.2005 (2005).

104 Ren, D., Lei, L. & Burton, Z. F. A region within the RAP74 subunit of human transcription factor IIF is critical for initiation but dispensable for complex assembly. *Mol. Cell. Biol.* **19**, 7377-7387 (1999).

105 Fishburn, J. & Hahn, S. Architecture of the yeast RNA polymerase II open complex and regulation of activity by TFIIF. *Mol. Cell. Biol.* **32**, 12-25, doi:10.1128/mcb.06242-11 (2012).

106 Soding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research* **33**, W244-248, doi:10.1093/nar/gki408 (2005).

107 Remmert, M., Biegert, A., Hauser, A. & Soding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173-175, doi:10.1038/nmeth.1818 (2012).

108 Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375-383, doi:10.1093/nar/gkm216 (2007).

109 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).

110 Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320-324, doi:10.1093/nar/gku316 (2014).

111 Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10037-10041, doi:10.1073/pnas.181342398 (2001).

112 Stiller, J. W. & Cook, M. S. Functional unit of the RNA polymerase II C-terminal domain lies within heptapeptide pairs. *Eukaryot. Cell* **3**, 735-740, doi:10.1128/ec.3.3.735-740.2004 (2004).

113 Mayer, A. *et al.* CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science* **336**, 1723-1725, doi:10.1126/science.1219651 (2012).

114 Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195-202, doi:10.1006/jmbi.1999.3091 (1999).

115 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410, doi:10.1016/s0022-2836(05)80360-2 (1990).

116 Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**, 7-8, doi:10.1038/nmeth.3213 (2014).

117 Lohi, O., Poussu, A., Mao, Y., Quiocho, F. & Lehto, V. P. VHS domain -- a longshoreman of vesicle lines. *FEBS Lett.* **513**, 19-23 (2002).

118 Holm, L. & Rosenstrom, P. in *Nucleic Acids Res.* Vol. 38 W545-549 (2010).

119 Baumli, S., Hole, A. J., Wang, L. Z., Noble, M. E. & Endicott, J. A. The CDK9 tail determines the reaction pathway of positive transcription elongation factor b. *Structure* **20**, 1788-1795, doi:10.1016/j.str.2012.08.011 (2012).

120 Plaschka, C. *et al.* Architecture of the RNA polymerase II-Mediator core initiation complex. *Nature*, doi:10.1038/nature14229 (2015).

121 Kuhlbrandt, W. Cryo-EM enters a new era. *eLife* **3**, e03678, doi:10.7554/eLife.03678 (2014).

122 Morgan, D. O. Cyclin-dependent kinases: engines, clocks, and microprocessors. *Annu. Rev. Cell. Dev. Biol.* **13**, 261-291, doi:10.1146/annurev.cellbio.13.1.261 (1997).

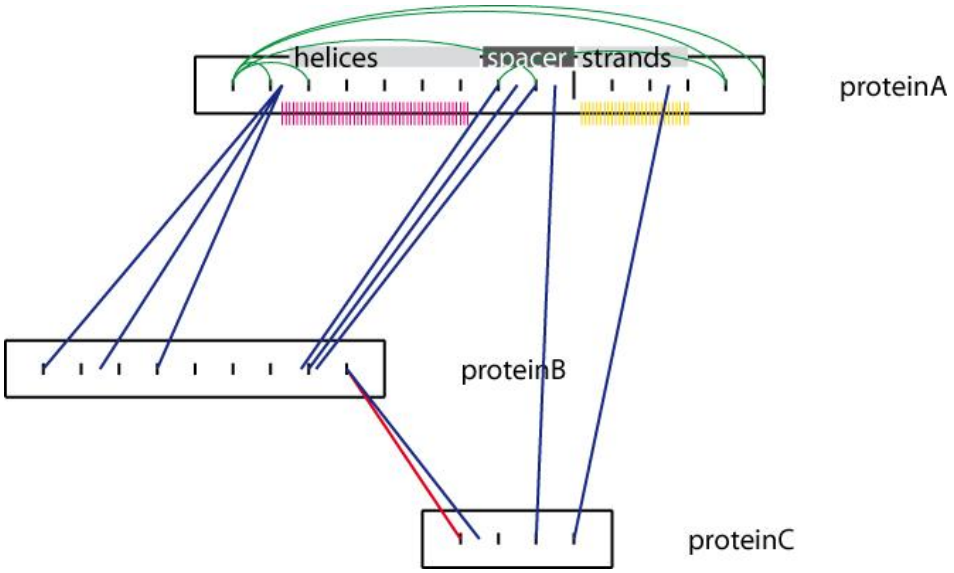
Appendix

Supplementary Material 1

Text and figures of the section Supplementary Material 1 were adapted from Simon Neyer, who programmed the MATLAB based script which was used as a basis to generate Figure 2.

BiClAn

BiClAn is a MATLAB® based tool to visualize Bivalent Crosslinking Analysis. It simplifies cross-link interpretation by converting a list of cross-links to a vector based diagram. Additionally, it is possible to export a list without redundant distance restraints.



Supplementary Figure 1: Example of a crosslink map.

In the following, BiClAn is explained by the trimeric complex proteinA/B/C. Intra and inter cross-links are shown as green and blue lines, respectively, while possible candidates are colored in red. For proteinA, domains are highlighted. Magenta and yellow lines represent α-helices and β-strands, respectively.

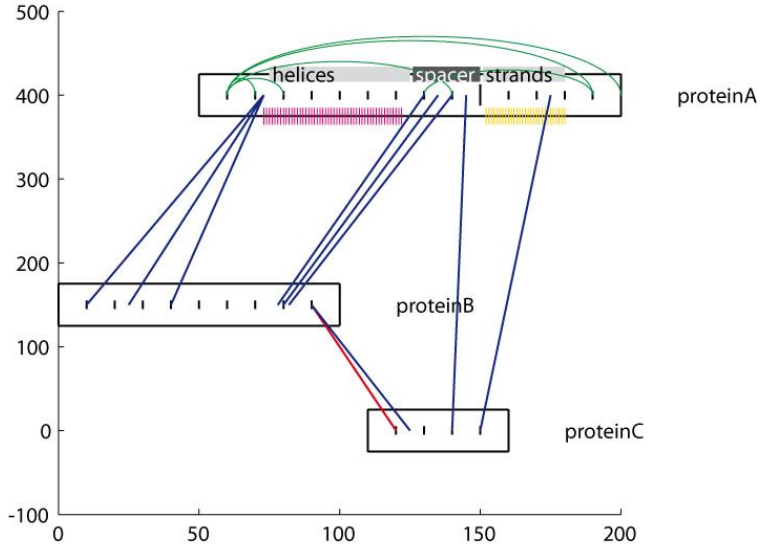
Data preparation and input files

As input files, tab delimited *.txt files have to be placed in the folder ‘input’. The names of these files should reflect what is contained in the variable file within the main script (BiClAn). The input files are formatted in the way that the third and fourth column state the position of the cross-links of the subunits given in column one and two, respectively. The fifth column allows one to mark a cross-link, e.g. as a candidate and allows to draw it in a different manner (Supplementary Figure 3). Please note, that inter and intra cross-links have to be provided in separate input files. The protein names may not include spaces. This is true for all input files.

The variable default loads a tab delimited file which sets the name of the proteins and allows the subunits to be treated separately (Supplementary Figure 3). Please note that you have to give the name of the proteins in descending alphabetical order (as they are called in file). Two additional rows are needed at the end of the document. The easiest way to generate this file is to use a spreadsheet program and save as tab delimited *.txt.

The number of amino acids is given in the second column. Although no axes are displayed, the diagram is an x-y-coordinate system (Supplementary Figure 2). The value in column eight and nine sets the x- and y-value, respectively. Whether an operation has (‘1’) or has not (‘0’) to be executed for a specific subunit this can be set in the residual columns.

Before drawing distance restraints, the user has to choose which input files to use. As shown in Supplementary Figure 4, this is done by handing over the index of the input file to variable ‘i’. By executing these functions one by one, cross-links from different experiments can be drawn in different colors to make them distinguishable.



Supplementary Figure 2: x-/y- position of protein boxes.

To determine the position of the boxes representing complex subunits, x- and y-values have to be provided in the default file. They correspond to the middle left edge of the box. It is recommended to set these values randomly in the first place. For a second round, optimized positions can be determined by moving the boxes in a vector based graphics suite.

demo_inter1.txt

Z0Z0Z0_protein_A	Y0Y0Y0_protein_B	23	10	
Z0Z0Z0_protein_A	Y0Y0Y0_protein_B	23	25	
Z0Z0Z0_protein_A	Y0Y0Y0_protein_B	23	40	
Z0Z0Z0_protein_A	Y0Y0Y0_protein_B	80	78	
Z0Z0Z0_protein_A	Y0Y0Y0_protein_B	85	80	
Z0Z0Z0_protein_A	Y0Y0Y0_protein_B	90	82	
Z0Z0Z0_protein_A	X0X0X0_protein_C	95	30	
X0X0X0_protein_C	Z0Z0Z0_protein_A	40	125	
Y0Y0Y0_protein_B	X0X0X0_protein_C	90	10	candidate
Z0Z0Z0_protein_A	Y0Y0Y0_protein_B	23	25	candidate

demo_inter2.txt

Y0Y0Y0_protein_B	X0X0X0_protein_C	90	15	
Y0Y0Y0_protein_B	X0X0X0_protein_C	90	15	candidate

demo_intra.txt

Z0Z0Z0_protein_A	Z0Z0Z0_protein_A	10	20
Z0Z0Z0_protein_A	Z0Z0Z0_protein_A	10	30
Z0Z0Z0_protein_A	Z0Z0Z0_protein_A	10	150
Z0Z0Z0_protein_A	Z0Z0Z0_protein_A	10	140
Z0Z0Z0_protein_A	Z0Z0Z0_protein_A	10	90
Z0Z0Z0_protein_A	Z0Z0Z0_protein_A	80	140

default.txt

	A	B	C	D	E	F	G	H	I
1	proteinC	50	1	1	0	0	0	110	0
2	proteinB	100	1	1	0	0	0	0	150
3	proteinA	150	1	1	1	1	1	50	400
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
	name	*# aa*	*box?*	*inter?*	*intra?*	*SecStruc?*	*domain?*	*x*	*y*

proteinA_sec.txt

```
CCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEECCCCCCCCCCCCCCCCCCCCC
```

proteinA_dom.txt

alpha_helices	helices	25	75
spacer spacer		76	100
beta-strands	strands	101	130

Supplementary Figure 3: Input files.

All files required to generate the demo diagram are displayed. As with the others, default.txt is a tab delimited file. In this figure a screenshot from a spreadsheet program is shown to highlight the meaning of each column. However, it was saved as *.txt later on.

```
%%%%%%%%%%%%%%%%%%%
%%% input files %%%
%%%%%%%%%%%%%%%%%%%
file = {'demo_inter1'   % 1
    'demo_inter2'       % 2
    'demo_intra'        % 3
    }
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%%% intra cross-links %%%
%%%%%%%%%%%%%%%%%%%%%%%%
i=[3]
param.loop='g';      %color of intra cross links
BiClIntra(param,sorted,file,intra,x,y,i)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% inter cross links from selected subunits to selected subunits only %%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
i=[]
param.candidate='r';  % color and style of candidates
param.valid='b';    % color of valid cross links
param.width=1;
BiClInter1(param,sorted,file,inter,x,y,i,conc)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% inter cross links from selected subunits to all subunits %%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
i=[1 2]
param.candidate='r';  % color and style of candidates
param.valid='b';    % color of valid cross links
param.width=1;BiClInter2(param,sorted,file,inter,x,y,i,conc)
```

Supplementary Figure 4: Script excerpt.

Parts of the BiClAn script are shown to demonstrate how to choose which dataset (loaded in file) has to be drawn. The variable ‘i’ sets the index of the respective input file. In this example intra cross-links of 'demo_intra' are drawn. Inter cross-links from selected subunits to all subunits are drawn in case of 'demo_inter1' and 'demo_inter1'.

Domains and secondary structure

As optional features, BiClAn offers to display the domain architecture and secondary structure information of subunits. In this case input files have to follow a specific nomenclature. Starting with the name as given in default, _sec or _dom is used as suffix for a tab-delimited *.txt file. H, E and C represent helices, strands and regions neither forming helices nor strands, respectively. Domain information is given in four columns. Column one will be neglected by BiClAn, but may not contain spaces. The second column states the name of the domain, whose boundaries are set in column three and four.

Output files

The generated MATLAB figure can be saved as a pixel based image (e.g. *.png or *.jpg) as well as a vector based graphic (e.g. *.eps or *.ai). By changing the size of the window displaying your diagram you also change the size of the saved image. If the figures are saved as a graphic, clipping masks might be added. It is possible to remove those in a vector graphic program.

The variable sorted contains (for each input file individually) a table of unique cross-links. The name of the involved subunits is represented in column one and two by the indices as indicated in default (Supplementary Figure 3). The respective amino acid number is given in column three and four. Whether a cross-link was labeled as a candidate is displayed in column three.

Support

We are more than happy to help, if you encounter problems while executing this script. Please, feel free to contact us in this case: simon.neyer@mpibpc.mpg.de

Supplementary Material 2

Supplementary Table 1: Inter protein-protein crosslinks.

Link Type	Protein 1	Linked residue 1	Domain protein 1	Protein 2	Linked residue 2	Domain protein 2	M/z
Inter	Rpb1	938	Foot	Rpb10	59	Tail	643,196
Inter	Rpb1	368	Active site	Rpb11	37	Dimerization	1189,395
Inter	Rpb1	461	Active site	Rpb11	20	Dimerization	569,828
Inter	Rpb1	323	Clamp core	Rpb2	471	Fork	454,761
Inter	Rpb1	323	Clamp core	Rpb2	471	Fork	606,014
Inter	Rpb1	323	Clamp core	Rpb2	934	Wall	400,63
Inter	Rpb1	403	Dock	Rpb2	886	Wall	534,472
Inter	Rpb1	403	Dock	Rpb2	1102	Hybrid binding	728,376
Inter	Rpb1	403	Dock	Rpb2	1102	Hybrid binding	530,784
Inter	Rpb1	695	Funnel	Rpb2	972	Wall	616,756
Inter	Rpb1	773	Funnel	Rpb2	510	Fork	772,906
Inter	Rpb1	830	Cleft	Rpb2	507	Fork	747,411
Inter	Rpb1	830	Cleft	Rpb2	510	Fork	747,41
Inter	Rpb1	938	Foot	Rpb2	813	Hybrid binding	637,547
Inter	Rpb1	1102	Cleft	Rpb2	507	Fork	513,067
Inter	Rpb1	1112	Cleft	Rpb2	507	Fork	888,215
Inter	Rpb1	15	Clamp core	Rpb5	171	Assembly	743,909
Inter	Rpb1	129	Clamp head	Rpb5	161	Assembly	487,021
Inter	Rpb1	129	Clamp head	Rpb5	161	Assembly	415,437
Inter	Rpb1	129	Clamp head	Rpb5	161	Assembly	691,724
Inter	Rpb1	129	Clamp head	Rpb5	171	Assembly	694,405
Inter	Rpb1	129	Clamp head	Rpb5	171	Assembly	442,665
Inter	Rpb1	129	Clamp head	Rpb5	171	Assembly	737,103
Inter	Rpb1	129	Clamp head	Rpb5	171	Assembly	506,497
Inter	Rpb1	132	Clamp head	Rpb5	161	Assembly	519,044
Inter	Rpb1	132	Clamp head	Rpb5	171	Assembly	442,665
Inter	Rpb1	132	Clamp head	Rpb5	171	Assembly	553,078
Inter	Rpb1	934	Foot	Rpb5	20	Jaw	538,712
Inter	Rpb1	1003	Foot	Rpb5	166	Assembly	741,43
Inter	Rpb1	1003	Foot	Rpb5	197	Assembly	620,114
Inter	Rpb1	1003	Foot	Rpb5	197	Assembly	826,484
Inter	Rpb1	1350	Cleft	Rpb5	201	Assembly	824,917

Continuation of table Supplementary Table 1.

Link Type	Protein 1	Linked residue 1	Domain protein 1	Protein 2	Linked residue 2	Domain protein 2	M/z
Inter	Rpb1	1350	Cleft	Rpb5	201	Assembly	1099,553
Inter	Rpb1	15	Clamp core	Rpb6	72	Assembly	672,626
Inter	Rpb1	1003	Foot	Rpb6	72	Assembly	698,398
Inter	Rpb1	1003	Foot	Rpb6	76	Assembly	745,745
Inter	Rpb1	977	Foot	Rpb8	136	ß-barrel	1126,109
Inter	Rpb1	1246	Jaw	Rpb9	20	Jaw	731,756
Inter	Rpb1	1246	Jaw	Rpb9	20	Jaw	557,276
Inter	Rpb1	1246	Jaw	Rpb9	20	Jaw	835,412
Inter	Rpb1	1246	Jaw	Rpb9	20	Jaw	832,212
Inter	Rpb1	49	Clamp core	Tfg1	284	Insertion 1	544,297
Inter	Rpb1	1262	Jaw	Tfg1	411	Charged region	621,842
Inter	Rpb1	1262	Jaw	Tfg1	411	Charged region	828,787
Inter	Rpb1	49	Clamp core	Tfg2	342	WH domain	497,78
Inter	Rpb1	49	Clamp core	Tfg2	357	C-terminal	703,729
Inter	Rpb1	176	Clamp head	Tfg2	179	Insertion	562,905
Inter	Rpb1	34	Clamp core	TFIIB	9	B-ribbon	436,259
Inter	Rpb1	34	Clamp core	TFIIB	9	B-ribbon	581,345
Inter	Rpb1	49	Clamp core	TFIIB	9	B-ribbon	482,017
Inter	Rpb1	49	Clamp core	TFIIB	98	B-linker	572,053
Inter	Rpb1	49	Clamp core	TFIIB	112	B-linker	517,041
Inter	Rpb1	49	Clamp core	TFIIB	121	B-linker	478,763
Inter	Rpb1	49	Clamp core	TFIIB	121	B-linker	638,015
Inter	Rpb1	49	Clamp core	TFIIB	155	B-core N-term. Cycl	618,316
Inter	Rpb10	59	Tail	Rpb2	813	Hybrid binding	731,404
Inter	Rpb10	59	Tail	Rpb2	813	Hybrid binding	513,956
Inter	Rpb10	68	Tail	Rpb3	149	Domain2	440,499
Inter	Rpb10	68	Tail	Rpb3	149	Domain2	469,256
Inter	Rpb10	68	Tail	Tfg2	290	Linker	663,696
Inter	Rpb11	20	Dimerization	Rpb1	644	Pore 1	676,374
Inter	Rpb11	20	Dimerization	Rpb1	644	Pore 1	760,02
Inter	Rpb11	26	Dimerization	Rpb1	637	Pore 1	811,449
Inter	Rpb11	26	Dimerization	Rpb1	644	Pore 1	763,214
Inter	Rpb11	26	Dimerization	Rpb1	644	Pore 1	1185,636
Inter	Rpb11	26	Dimerization	Rpb1	644	Pore 1	811,448
Inter	Rpb11	26	Dimerization	Rpb1	644	Pore 1	1014,058
Inter	Rpb11	26	Dimerization	Rpb1	644	Pore 1	760,018
Inter	Rpb11	26	Dimerization	Rpb1	644	Pore 1	949,773
Inter	Rpb11	37	Dimerization	Rpb3	160	Domain2	659,565

Continuation of table Supplementary Table 1.

Link Type	Protein 1	Linked residue 1	Domain protein 1	Protein 2	Linked residue 2	Domain protein 2	M/z
Inter	Rpb11	37	Dimerization	Rpb3	253	Dimerization	676,175
Inter	Rpb12	28	Zinc Ribbon	Rpb10	68	Tail	701,595
Inter	Rpb12	37	Zinc Ribbon	Rpb10	68	Tail	712,102
Inter	Rpb12	37	Zinc Ribbon	Rpb3	149	Domain2	629,074
Inter	Rpb12	37	Zinc Ribbon	Rpb3	149	Domain2	838,429
Inter	Rpb2	227	Lobe	Rpb1	1093	Cleft	543,569
Inter	Rpb2	507	Fork	Rpb1	1093	Cleft	467,031
Inter	Rpb2	507	Fork	Rpb1	1093	Cleft	622,371
Inter	Rpb2	886	Wall	Rpb1	372	Active site	904,472
Inter	Rpb2	892	Wall	Rpb1	343	Clamp core	623,522
Inter	Rpb2	1102	Hybrid binding	Rpb1	323	Clamp core	472,462
Inter	Rpb2	1102	Hybrid binding	Rpb1	343	Clamp core	534,282
Inter	Rpb2	1102	Hybrid binding	Rpb1	343	Clamp core	481,463
Inter	Rpb2	1183	Clamp	Rpb1	34	Clamp core	920,13
Inter	Rpb2	191	Protrusion	Rpb10	68	Tail	748,376
Inter	Rpb2	191	Protrusion	Rpb10	68	Tail	1035,842
Inter	Rpb2	801	Hybrid binding	Rpb10	59	Tail	816,439
Inter	Rpb2	801	Hybrid binding	Rpb10	59	Tail	847,656
Inter	Rpb2	864	Wall	Rpb12	58	Zinc Ribbon	470,515
Inter	Rpb2	191	Protrusion	Rpb3	149	Domain2	694,106
Inter	Rpb2	191	Protrusion	Rpb3	149	Domain2	925,137
Inter	Rpb2	1057	Hybrid binding	Rpb3	199	Loop	1039,907
Inter	Rpb2	1057	Hybrid binding	Rpb3	199	Loop	1088,145
Inter	Rpb2	87	Protrusion	Tfg1	335	Dimerization	507,028
Inter	Rpb2	246	Lobe	Tfg1	335	Dimerization	634,35
Inter	Rpb2	270	Lobe	Tfg1	411	Charged region	515,055
Inter	Rpb2	277	Lobe	Tfg1	411	Charged region	728,181
Inter	Rpb2	277	Lobe	Tfg1	411	Charged region	970,568
Inter	Rpb2	358	Lobe	Tfg1	328	Dimerization	739,8
Inter	Rpb2	358	Lobe	Tfg1	328	Dimerization	924,495
Inter	Rpb2	358	Lobe	Tfg1	328	Dimerization	638,016
Inter	Rpb2	358	Lobe	Tfg1	328	Dimerization	765,42
Inter	Rpb2	358	Lobe	Tfg1	328	Dimerization	680,876
Inter	Rpb2	358	Lobe	Tfg1	335	Dimerization	691,571
Inter	Rpb2	358	Lobe	Tfg1	335	Dimerization	864,213
Inter	Rpb2	358	Lobe	Tfg1	335	Dimerization	597,826
Inter	Rpb2	358	Lobe	Tfg1	335	Dimerization	717,189
Inter	Rpb2	606	External 2	Tfg1	61	N-terminal	687,586

Continuation of table Supplementary Table 1.

Link Type	Protein 1	Linked residue 1	Domain protein 1	Protein 2	Linked residue 2	Domain protein 2	M/z
Inter	Rpb2	606	External 2	Tfg1	61	N-terminal	726,611
Inter	Rpb2	606	External 2	Tfg1	61	N-terminal	638,52
Inter	Rpb2	87	Protrusion	Tfg2	163	Insertion	465,513
Inter	Rpb2	99	Protrusion	Tfg2	286	Linker	703,97
Inter	Rpb2	133	Protrusion	Tfg2	249	Linker	621,088
Inter	Rpb2	133	Protrusion	Tfg2	249	Linker	660,113
Inter	Rpb2	148	Protrusion	Tfg2	148	Insertion	874,428
Inter	Rpb2	177	Protrusion	Tfg2	245	Linker	996,508
Inter	Rpb2	191	Protrusion	Tfg2	279	Linker	827,659
Inter	Rpb2	246	Lobe	Tfg2	148	Insertion	675,163
Inter	Rpb2	246	Lobe	Tfg2	249	Linker	586,656
Inter	Rpb2	344	Lobe	Tfg2	179	Insertion	623,374
Inter	Rpb2	344	Lobe	Tfg2	179	Insertion	435,759
Inter	Rpb2	344	Lobe	Tfg2	179	Insertion	405,647
Inter	Rpb2	606	External 2	Tfg2	245	Linker	610,342
Inter	Rpb2	606	External 2	Tfg2	245	Linker	813,455
Inter	Rpb2	606	External 2	Tfg2	245	Linker	649,367
Inter	Rpb2	606	External 2	Tfg2	245	Linker	688,393
Inter	Rpb2	606	External 2	Tfg2	279	Linker	683,855
Inter	Rpb2	865	Wall	Tfg2	279	Linker	797,905
Inter	Rpb2	934	Wall	Tfg2	286	Linker	415,735
Inter	Rpb2	228	Lobe	TFIIB	98	B-linker	544,893
Inter	Rpb2	228	Lobe	TFIIB	98	B-linker	680,865
Inter	Rpb2	277	Lobe	TFIIB	108	B-linker	754,943
Inter	Rpb2	277	Lobe	TFIIB	108	B-linker	1006,256
Inter	Rpb2	426	Protrusion	TFIIB	108	B-linker	510,049
Inter	Rpb2	451	Protrusion	TFIIB	147	B-core N-ter. cyclin	575,836
Inter	Rpb2	458	Protrusion	TFIIB	147	B-core N-ter. cyclin	1003,534
Inter	Rpb2	458	Protrusion	TFIIB	147	B-core N-ter. cyclin	784,924
Inter	Rpb2	470	Fork	TFIIB	98	B-linker	904,123
Inter	Rpb2	892	Wall	TFIIB	33	B-ribbon	636,74
Inter	Rpb2	892	Wall	TFIIB	33	B-ribbon	573,976
Inter	Rpb3	154	Domain2	Rpb10	68	Tail	605,324
Inter	Rpb3	160	Domain2	Rpb2	934	Wall	607,003
Inter	Rpb3	137	Domain2	Tfg2	357	C-terminal	794,182
Inter	Rpb4	142	Tip-associated	Rpb3	149	Domain2	643,339
Inter	Rpb5	171	Assembly	Rpb1	129	Clamp head	480,878
Inter	Rpb5	201	Assembly	Rpb1	938	Foot	415,731

Continuation of table Supplementary Table 1.

Link Type	Protein 1	Linked residue 1	Domain protein 1	Protein 2	Linked residue 2	Domain protein 2	M/z
Inter	Rpb5	171	Assembly	Rpb6	72	Assembly	423,249
Inter	Rpb6	72	Assembly	Rpb5	166	Assembly	546,342
Inter	Rpb7	73	Tip	Rpb6	72	Assembly	633,849
Inter	Rpb9	93	Zinc Ribbon	Rpb1	1093	Cleft	778,197
Inter	Rpb9	93	Zinc Ribbon	Rpb2	227	Lobe	847,638
Inter	Rpb9	77	Zinc Ribbon	Tfg1	60	N-terminal	550,78
Inter	TBP	97	core-TBP	Tfg2	290	Linker	487,267
Inter	Tfg1	411	Charged region	Rpb1	186	Clamp head	432,502
Inter	Tfg1	706	WH domain	Rpb1	1217	Jaw	655,37
Inter	Tfg1	23	N-terminal	Rpb2	655	External 1	736,075
Inter	Tfg1	23	N-terminal	Rpb2	813	Hybrid binding	653,342
Inter	Tfg1	60	N-terminal	Rpb2	606	External 2	808,66
Inter	Tfg1	61	N-terminal	Rpb2	606	External 2	808,659
Inter	Tfg1	61	N-terminal	Rpb2	606	External 2	678,351
Inter	Tfg1	61	N-terminal	Rpb2	606	External 2	758,87
Inter	Tfg1	61	N-terminal	Rpb2	606	External 2	1011,492
Inter	Tfg1	61	N-terminal	Rpb2	652	External 1	612,321
Inter	Tfg1	89	N-terminal	Rpb2	606	External 2	758,581
Inter	Tfg1	89	N-terminal	Rpb2	622	External 2	692,96
Inter	Tfg1	328	Dimerization	Rpb2	87	Protrusion	567,314
Inter	Tfg1	328	Dimerization	Rpb2	87	Protrusion	756,081
Inter	Tfg1	328	Dimerization	Rpb2	246	Lobe	714,73
Inter	Tfg1	328	Dimerization	Rpb2	246	Lobe	600,586
Inter	Tfg1	328	Dimerization	Rpb2	426	Protrusion	463,873
Inter	Tfg1	328	Dimerization	Rpb2	426	Protrusion	579,589
Inter	Tfg1	328	Dimerization	Rpb2	426	Protrusion	547,565
Inter	Tfg1	328	Dimerization	Rpb2	426	Protrusion	611,614
Inter	Tfg1	61	N-terminal	Rpb9	77	Zinc Ribbon 2	777,1
Inter	Tfg1	61	N-terminal	Tfg2	245	Linker	620,932
Inter	Tfg1	61	N-terminal	Tfg2	245	Linker	524,071
Inter	Tfg1	61	N-terminal	Tfg2	245	Linker	654,84
Inter	Tfg1	61	N-terminal	Tfg2	245	Linker	693,863
Inter	Tfg1	61	N-terminal	Tfg2	249	Linker	490,242
Inter	Tfg1	61	N-terminal	Tfg2	279	Linker	728,353
Inter	Tfg1	61	N-terminal	Tfg2	286	Linker	467,637
Inter	Tfg1	61	N-terminal	Tfg2	290	Linker	499,055
Inter	Tfg1	89	N-terminal	Tfg2	94	Dimerization	816,817
Inter	Tfg1	89	N-terminal	Tfg2	94	Dimerization	952,99

Continuation of table Supplementary Table 1.

Link Type	Protein 1	Linked residue 1	Domain protein 1	Protein 2	Linked residue 2	Domain protein 2	M/z
Inter	Tfg1	91	N-terminal	Tfg2	99	Dimerization	798,652
Inter	Tfg1	91	N-terminal	Tfg2	99	Dimerization	921,787
Inter	Tfg1	91	N-terminal	Tfg2	127	Dimerization	854,829
Inter	Tfg1	108	Dimerization	Tfg2	148	Insertion	880,46
Inter	Tfg1	184	Insertion 1	Tfg2	148	Insertion	862,208
Inter	Tfg1	184	Insertion 1	Tfg2	156	Insertion	869,197
Inter	Tfg1	284	Insertion 1	Tfg2	174	Insertion	481,01
Inter	Tfg1	335	Dimerization	Tfg2	163	Insertion	451,763
Inter	Tfg1	335	Dimerization	Tfg2	179	Insertion	623,678
Inter	Tfg1	616	Insertion 2	Tfg2	156	Insertion	440,907
Inter	Tfg2	279	Linker	Rpb10	68	Tail	803,413
Inter	Tfg2	342	WH domain	Rpb10	59	Tail	746,016
Inter	Tfg2	348	WH domain	Rpb10	59	Tail	746,014
Inter	Tfg2	148	Insertion	Rpb2	87	Protrusion	574,558
Inter	Tfg2	164	Insertion	Rpb2	87	Protrusion	575,057
Inter	Tfg2	164	Insertion	Rpb2	344	Lobe	705,067
Inter	Tfg2	164	Insertion	Rpb2	344	Lobe	454,664
Inter	Tfg2	172	Insertion	Rpb2	344	Lobe	568,078
Inter	Tfg2	206	Dimerization	Rpb2	344	Lobe	537,328
Inter	Tfg2	206	Dimerization	Rpb2	344	Lobe	768,136
Inter	Tfg2	235	Linker	Rpb2	606	External 2	982,679
Inter	Tfg2	245	Linker	Rpb2	246	Lobe	482,53
Inter	Tfg2	245	Linker	Rpb2	246	Lobe	417,447
Inter	Tfg2	342	WH domain	Rpb2	813	Hybrid binding	611,928
Inter	Tfg2	348	WH domain	Rpb2	813	Hybrid binding	732,634
Inter	Tfg2	348	WH domain	Rpb2	813	Hybrid binding	764,658
Inter	Tfg2	348	WH domain	Rpb2	813	Hybrid binding	1019,207
Inter	Tfg2	359	C-terminal	Rpb3	149	Domain2	443,503
Inter	Tfg2	127	Dimerization	Tfg1	126	Dimerization	607,668
Inter	Tfg2	127	Dimerization	Tfg1	126	Dimerization	729,001
Inter	Tfg2	127	Dimerization	Tfg1	389	Dimerization	716,133
Inter	Tfg2	148	Insertion	Tfg1	335	Dimerization	747,405
Inter	Tfg2	156	Insertion	Tfg1	335	Dimerization	756,728
Inter	Tfg2	156	Insertion	Tfg1	616	Insertion 2	528,887
Inter	Tfg2	164	Insertion	Tfg1	284	Insertion 1	581,306
Inter	Tfg2	164	Insertion	Tfg1	335	Dimerization	748,07
Inter	Tfg2	206	Dimerization	Tfg1	335	Dimerization	569,578
Inter	Tfg2	235	Linker	Tfg1	61	N-terminal	961,247

Continuation of table Supplementary Table 1.

		Linked	Domain		Linked	Domain	
Link Type	Protein 1	residue 1	protein 1	Protein 2	residue 2	protein 2	M/z
Inter	TFIIB	98	B-linker	Rpb2	246	Lobe	523,773
Inter	TFIIB	98	B-linker	Rpb2	426	Protrusion	567,062
Inter	TFIIB	98	B-linker	Rpb2	471	Fork	641,992
Inter	TFIIB	98	B-linker	Rpb2	864	Wall	758,364
Inter	TFIIB	108	B-linker	Rpb2	426	Protrusion	478,025
Inter	TFIIB	151	B-core N-ter. cy	Rpb2	864	Wall	717,331
Inter	TFIIB	9	B-ribbon	Rpb7	27	Tip	650,364
Inter	TFIIB	155	B-core N-ter. cy	Tfg2	286	Linker	641,817
Inter	TFIIB	161	B-core N-ter. cy	Tfg2	179	Insertion	451,247
Inter	TFIIB	199	B-core N-ter. cy	Tfg2	279	Linker	890,491
Inter	TFIIB	199	B-core N-ter. cy	Tfg2	286	Linker	524,064
Inter	TFIIB	199	B-core N-ter. cy	Tfg2	286	Linker	698,417
Inter	TFIIB	199	B-core N-ter. cy	Tfg2	290	Linker	563,335

Supplementary Table 2: Intra protein-protein crosslinks.

Link Type	Protein 1	Linked residue 1	Domain protein 1	Protein 2	Linked residue 2	Domain protein 2	M/z
Intra	Rpb1	101	Clamp head	Rpb1	143	Clamp head	640,71
Intra	Rpb1	143	Clamp head	Rpb1	186	Clamp head	574,075
Intra	Rpb1	143	Clamp head	Rpb1	186	Clamp head	613,1
Intra	Rpb1	368	Active site	Rpb1	461	Active site	907,509
Intra	Rpb1	372	Active site	Rpb1	403	Dock	1290,352
Intra	Rpb1	372	Active site	Rpb1	403	Dock	968,011
Intra	Rpb1	403	Dock	Rpb1	343	Clamp core	616,828
Intra	Rpb1	431	Dock	Rpb1	343	Clamp core	644,593
Intra	Rpb1	644	Pore 1	Rpb1	461	Active site	969,538
Intra	Rpb1	644	Pore 1	Rpb1	461	Active site	727,408
Intra	Rpb1	695	Funnel	Rpb1	789	Funnel	721,18
Intra	Rpb1	705	Funnel	Rpb1	689	Funnel	706,155
Intra	Rpb1	705	Funnel	Rpb1	689	Funnel	898,507
Intra	Rpb1	705	Funnel	Rpb1	1093	Cleft	912,173
Intra	Rpb1	705	Funnel	Rpb1	1093	Cleft	547,707
Intra	Rpb1	705	Funnel	Rpb1	1286	Cleft	978,177
Intra	Rpb1	705	Funnel	Rpb1	1286	Cleft	587,31
Intra	Rpb1	705	Funnel	Rpb1	1286	Cleft	765,911
Intra	Rpb1	773	Funnel	Rpb1	1093	Cleft	837,685
Intra	Rpb1	830	Cleft	Rpb1	1093	Cleft	737,155
Intra	Rpb1	830	Cleft	Rpb1	1102	Cleft	1043,918
Intra	Rpb1	880	Foot	Rpb1	620	Pore 1	967,99
Intra	Rpb1	991	Foot	Rpb1	938	Foot	740,942
Intra	Rpb1	1102	Cleft	Rpb1	1093	Cleft	502,809
Intra	Rpb1	1112	Cleft	Rpb1	773	Funnel	1007,295
Intra	Rpb1	1112	Cleft	Rpb1	1093	Cleft	877,958
Intra	Rpb1	1112	Cleft	Rpb1	1102	Cleft	1231,659
Intra	Rpb1	1132	Cleft	Rpb1	705	Funnel	1107,838
Intra	Rpb1	1132	Cleft	Rpb1	1205	Jaw	1103,306
Intra	Rpb1	1132	Cleft	Rpb1	1286	Cleft	743,99
Intra	Rpb1	1246	Jaw	Rpb1	1217	Jaw	800,421
Intra	Rpb1	1246	Jaw	Rpb1	1221	Jaw	933,255

Continuation of supplementary Table 2.

Link Type	Protein 1	Linked residue 1	Domain protein 1	Protein 2	Linked residue 2	Domain protein 2	M/z
Intra	Rpb11	37	Dimerization	Rpb11	20	Dimerization	681,576
Intra	Rpb11	55	Dimerization	Rpb11	88	Dimerization	704,985
Intra	Rpb11	88	Dimerization	Rpb11	55	Dimerization	846,465
Intra	Rpb12	28	Zinc Ribbon	Rpb12	49	Zinc Ribbon	901,08
Intra	Rpb2	99	Protrusion	Rpb2	164	Protrusion	1046,512
Intra	Rpb2	148	Protrusion	Rpb2	87	Protrusion	1096,034
Intra	Rpb2	228	Lobe	Rpb2	257	Lobe	1013,244
Intra	Rpb2	228	Lobe	Rpb2	270	Lobe	552,926
Intra	Rpb2	228	Lobe	Rpb2	507	Fork	625,113
Intra	Rpb2	246	Lobe	Rpb2	426	Protrusion	626,354
Intra	Rpb2	246	Lobe	Rpb2	426	Protrusion	470,018
Intra	Rpb2	277	Lobe	Rpb2	228	Lobe	904,027
Intra	Rpb2	277	Lobe	Rpb2	471	Fork	704,908
Intra	Rpb2	358	Lobe	Rpb2	246	Lobe	846,949
Intra	Rpb2	358	Lobe	Rpb2	246	Lobe	677,761
Intra	Rpb2	358	Lobe	Rpb2	246	Lobe	703,384
Intra	Rpb2	358	Lobe	Rpb2	344	Lobe	831,961
Intra	Rpb2	426	Protrusion	Rpb2	246	Lobe	401,835
Intra	Rpb2	426	Protrusion	Rpb2	246	Lobe	502,042
Intra	Rpb2	426	Protrusion	Rpb2	471	Fork	460,015
Intra	Rpb2	426	Protrusion	Rpb2	471	Fork	427,99
Intra	Rpb2	451	Protrusion	Rpb2	470	Fork	978,21
Intra	Rpb2	458	Protrusion	Rpb2	471	Fork	714,375
Intra	Rpb2	458	Protrusion	Rpb2	864	Wall	801,654
Intra	Rpb2	458	Protrusion	Rpb2	864	Wall	833,678
Intra	Rpb2	470	Fork	Rpb2	246	Lobe	613,324
Intra	Rpb2	470	Fork	Rpb2	423	Protrusion	656,612
Intra	Rpb2	470	Fork	Rpb2	426	Protrusion	875,146
Intra	Rpb2	470	Fork	Rpb2	426	Protrusion	656,612
Intra	Rpb2	507	Fork	Rpb2	471	Fork	425,993
Intra	Rpb2	864	Wall	Rpb2	934	Wall	507,752
Intra	Rpb2	864	Wall	Rpb2	934	Wall	676,666
Intra	Rpb2	865	Wall	Rpb2	451	Protrusion	657,162
Intra	Rpb2	865	Wall	Rpb2	934	Wall	704,361
Intra	Rpb2	876	Wall	Rpb2	451	Protrusion	657,162
Intra	Rpb2	876	Wall	Rpb2	451	Protrusion	821,2
Intra	Rpb2	886	Wall	Rpb2	864	Wall	842,415
Intra	Rpb2	886	Wall	Rpb2	934	Wall	495,501

Continuation of supplementary Table 2.

Link Type	Protein 1	Linked residue 1	Domain protein 1	Protein 2	Linked residue 2	Domain protein 2	M/z
Intra	Rpb2	892	Wall	Rpb2	864	Wall	674,134
Intra	Rpb2	892	Wall	Rpb2	934	Wall	736,127
Intra	Rpb2	892	Wall	Rpb2	934	Wall	589,102
Intra	Rpb2	892	Wall	Rpb2	934	Wall	640,93
Intra	Rpb2	979	Hybrid binding	Rpb2	1102	Hybrid binding	518,878
Intra	Rpb2	987	Hybrid binding	Rpb2	979	Hybrid binding	675,62
Intra	Rpb2	987	Hybrid binding	Rpb2	1102	Hybrid binding	538,687
Intra	Rpb2	987	Hybrid binding	Rpb2	1102	Hybrid binding	673,106
Intra	Rpb2	987	Hybrid binding	Rpb2	1102	Hybrid binding	592,524
Intra	Rpb2	987	Hybrid binding	Rpb2	1102	Hybrid binding	475,089
Intra	Rpb2	1102	Hybrid binding	Rpb2	471	Fork	408,412
Intra	Rpb2	1102	Hybrid binding	Rpb2	864	Wall	597,544
Intra	Rpb2	1102	Hybrid binding	Rpb2	886	Wall	468,436
Intra	Rpb2	1102	Hybrid binding	Rpb2	934	Wall	445,032
Intra	Rpb2	1102	Hybrid binding	Rpb2	972	Wall	615,584
Intra	Rpb2	1102	Hybrid binding	Rpb2	972	Wall	546,505
Intra	Rpb2	1102	Hybrid binding	Rpb2	979	Hybrid binding	572,715
Intra	Rpb2	1174	Clamp	Rpb2	1188	Clamp	850,595
Intra	Rpb3	15	Dimerization	Rpb3	137	Domain2	1191,863
Intra	Rpb3	165	Domain2	Rpb3	160	Domain2	453,262
Intra	Rpb3	165	Domain2	Rpb3	253	Dimerization	674,397
Intra	Rpb4	121	Tip-associated	Rpb4	60	Tip-associated	881,734
Intra	Rpb4	142	Tip-associated	Rpb4	60	Tip-associated	760,166
Intra	Rpb5	45	Jaw	Rpb5	20	Jaw	610,81
Intra	Rpb5	45	Jaw	Rpb5	20	Jaw	514,468
Intra	Rpb5	45	Jaw	Rpb5	20	Jaw	642,832
Intra	Rpb5	171	Assembly	Rpb5	161	Assembly	486,271
Intra	Rpb5	171	Assembly	Rpb5	166	Assembly	417,849
Intra	Rpb5	197	Assembly	Rpb5	201	Assembly	493,539
Intra	Rpb6	76	Assembly	Rpb6	128	Assembly	561,321
Intra	TBP	83	core-TBP	TBP	97	core-TBP	473,866
Intra	TBP	151	core-TBP	TBP	133	core-TBP	830,423
Intra	TBP	151	core-TBP	TBP	133	core-TBP	623,069
Intra	TBP	151	core-TBP	TBP	133	core-TBP	655,093
Intra	TBP	156	core-TBP	TBP	133	core-TBP	1039,181
Intra	TBP	156	core-TBP	TBP	133	core-TBP	779,638
Intra	Tfg1	61	N-terminal	Tfg1	89	N-terminal	587,799
Intra	Tfg1	89	N-terminal	Tfg1	61	N-terminal	794,181

Continuation of supplementary Table 2.

Link Type	Protein 1	Linked residue 1	Domain protein 1	Protein 2	Linked residue 2	Domain protein 2	M/z
Intra	Tfg1	91	N-terminal	Tfg1	61	N-terminal	885,413
Intra	Tfg1	120	Dimerization	Tfg1	389	Dimerization	531,556
Intra	Tfg1	126	Dimerization	Tfg1	120	Dimerization	555,72
Intra	Tfg1	184	Insertion 1	Tfg1	267	Insertion 1	1014,788
Intra	Tfg1	267	Insertion 1	Tfg1	284	Insertion 1	733,392
Intra	Tfg1	288	Insertion 1	Tfg1	279	Insertion 1	1265,589
Intra	Tfg1	406	Charged region	Tfg1	411	Charged region	773,744
Intra	Tfg1	421	Charged region	Tfg1	406	Charged region	1111,911
Intra	Tfg1	421	Charged region	Tfg1	406	Charged region	834,183
Intra	Tfg1	591	Insertion 2	Tfg1	522	Insertion 2	519,774
Intra	Tfg1	591	Insertion 2	Tfg1	529	Insertion 2	488,602
Intra	Tfg1	591	Insertion 2	Tfg1	579	Insertion 2	522,613
Intra	Tfg1	599	Insertion 2	Tfg1	719	WH domain	599,718
Intra	Tfg1	690	WH domain	Tfg1	706	WH domain	491,768
Intra	Tfg1	690	WH domain	Tfg1	706	WH domain	655,356
Intra	Tfg1	706	WH domain	Tfg1	690	WH domain	424,836
Intra	Tfg1	706	WH domain	Tfg1	690	WH domain	530,794
Intra	Tfg1	706	WH domain	Tfg1	714	WH domain	430,65
Intra	Tfg1	714	WH domain	Tfg1	706	WH domain	499,035
Intra	Tfg1	719	WH domain	Tfg1	529	Insertion 2	537,286
Intra	Tfg1	719	WH domain	Tfg1	690	WH domain	460,845
Intra	Tfg1	719	WH domain	Tfg1	690	WH domain	486,464
Intra	Tfg1	719	WH domain	Tfg1	706	WH domain	611,318
Intra	Tfg1	719	WH domain	Tfg1	706	WH domain	489,254
Intra	Tfg1	719	WH domain	Tfg1	706	WH domain	458,034
Intra	Tfg1	719	WH domain	Tfg1	706	WH domain	429,23
Intra	Tfg1	719	WH domain	Tfg1	706	WH domain	604,317
Intra	Tfg1	719	WH domain	Tfg1	714	WH domain	492,278
Intra	Tfg1	719	WH domain	Tfg1	714	WH domain	517,898
Intra	Tfg1	719	WH domain	Tfg1	714	WH domain	819,793
Intra	Tfg1	719	WH domain	Tfg1	714	WH domain	615,097
Intra	Tfg1	728	C-terminal	Tfg1	690	WH domain	486,465
Intra	Tfg1	728	C-terminal	Tfg1	714	WH domain	492,278
Intra	Tfg2	80	N-terminal	Tfg2	245	Linker	593,339
Intra	Tfg2	127	Dimerization	Tfg2	99	Dimerization	717,386
Intra	Tfg2	141	Dimerization	Tfg2	148	Insertion	823,691
Intra	Tfg2	141	Dimerization	Tfg2	148	Insertion	1055,219
Intra	Tfg2	142	Dimerization	Tfg2	179	Insertion	451,505

Continuation of supplementary Table 2.

Link Type	Protein 1	Linked residue 1	Domain protein 1	Protein 2	Linked residue 2	Domain protein 2	M/z
Intra	Tfg2	147	Insertion	Tfg2	164	Insertion	503,27
Intra	Tfg2	147	Insertion	Tfg2	179	Insertion	535,541
Intra	Tfg2	148	Insertion	Tfg2	142	Dimerization	410,024
Intra	Tfg2	148	Insertion	Tfg2	163	Insertion	692,055
Intra	Tfg2	148	Insertion	Tfg2	164	Insertion	795,412
Intra	Tfg2	148	Insertion	Tfg2	164	Insertion	838,112
Intra	Tfg2	148	Insertion	Tfg2	164	Insertion	503,268
Intra	Tfg2	148	Insertion	Tfg2	174	Insertion	704,384
Intra	Tfg2	148	Insertion	Tfg2	174	Insertion	528,539
Intra	Tfg2	148	Insertion	Tfg2	174	Insertion	454,252
Intra	Tfg2	148	Insertion	Tfg2	179	Insertion	713,721
Intra	Tfg2	148	Insertion	Tfg2	179	Insertion	713,719
Intra	Tfg2	148	Insertion	Tfg2	179	Insertion	535,541
Intra	Tfg2	148	Insertion	Tfg2	186	Insertion	507,025
Intra	Tfg2	156	Insertion	Tfg2	163	Insertion	701,377
Intra	Tfg2	156	Insertion	Tfg2	164	Insertion	847,434
Intra	Tfg2	156	Insertion	Tfg2	174	Insertion	574,557
Intra	Tfg2	163	Insertion	Tfg2	174	Insertion	653,71
Intra	Tfg2	164	Insertion	Tfg2	142	Dimerization	544,8
Intra	Tfg2	164	Insertion	Tfg2	174	Insertion	747,745
Intra	Tfg2	164	Insertion	Tfg2	179	Insertion	714,383
Intra	Tfg2	164	Insertion	Tfg2	179	Insertion	429,032
Intra	Tfg2	172	Insertion	Tfg2	142	Dimerization	544,799
Intra	Tfg2	172	Insertion	Tfg2	173	Insertion	561,061
Intra	Tfg2	172	Insertion	Tfg2	179	Insertion	568,063
Intra	Tfg2	172	Insertion	Tfg2	185	Insertion	431,839
Intra	Tfg2	172	Insertion	Tfg2	186	Insertion	539,546
Intra	Tfg2	174	Insertion	Tfg2	163	Insertion	558,991
Intra	Tfg2	179	Insertion	Tfg2	163	Insertion	611,024
Intra	Tfg2	185	Insertion	Tfg2	179	Insertion	446,253
Intra	Tfg2	206	Dimerization	Tfg2	174	Insertion	758,78
Intra	Tfg2	279	Linker	Tfg2	245	Linker	579,823
Intra	Tfg2	279	Linker	Tfg2	249	Linker	537,536
Intra	Tfg2	279	Linker	Tfg2	286	Linker	509,281
Intra	Tfg2	279	Linker	Tfg2	290	Linker	548,552
Intra	Tfg2	286	Linker	Tfg2	249	Linker	524,303
Intra	Tfg2	290	Linker	Tfg2	284	Linker	581,348
Intra	Tfg2	297	WH domain	Tfg2	290	Linker	531,558

Continuation of supplementary Table 2.

Link Type	Protein 1	Linked residue 1	Domain protein 1	Protein 2	Linked residue 2	Domain protein 2	M/z
Intra	Tfg2	330	WH domain	Tfg2	341	WH domain	874,157
Intra	Tfg2	330	WH domain	Tfg2	341	WH domain	655,869
Intra	Tfg2	335	WH domain	Tfg2	279	Linker	663,105
Intra	Tfg2	335	WH domain	Tfg2	290	Linker	558,319
Intra	Tfg2	335	WH domain	Tfg2	297	WH domain	646,112
Intra	Tfg2	335	WH domain	Tfg2	357	C-terminal	782,098
Intra	Tfg2	335	WH domain	Tfg2	359	C-terminal	739,4
Intra	Tfg2	348	WH domain	Tfg2	316	WH domain	944,493
Intra	Tfg2	348	WH domain	Tfg2	316	WH domain	755,793
Intra	Tfg2	348	WH domain	Tfg2	341	WH domain	657,876
Intra	Tfg2	356	C-terminal	Tfg2	179	Insertion	637,622
Intra	Tfg2	356	C-terminal	Tfg2	342	WH domain	537,049
Intra	Tfg2	356	C-terminal	Tfg2	359	C-terminal	535,293
Intra	Tfg2	356	C-terminal	Tfg2	359	C-terminal	567,317
Intra	TFIIB	33	B-ribbon	TFIIB	9	B-ribbon	497,033
Intra	TFIIB	98	B-linker	TFIIB	108	B-linker	531,781
Intra	TFIIB	98	B-linker	TFIIB	112	B-linker	711,37
Intra	TFIIB	98	B-linker	TFIIB	112	B-linker	533,782
Intra	TFIIB	98	B-linker	TFIIB	161	B-core N-ter. cyclin	597,805
Intra	TFIIB	108	B-linker	TFIIB	121	B-linker	584,319
Intra	TFIIB	112	B-linker	TFIIB	121	B-linker	586,986
Intra	TFIIB	116	B-linker	TFIIB	98	B-linker	591,556
Intra	TFIIB	116	B-linker	TFIIB	108	B-linker	502,52
Intra	TFIIB	116	B-linker	TFIIB	108	B-linker	534,544
Intra	TFIIB	116	B-linker	TFIIB	108	B-linker	427,837
Intra	TFIIB	116	B-linker	TFIIB	161	B-core N-ter. cyclin	600,567
Intra	TFIIB	132	B-core N-ter. cy	TFIIB	98	B-linker	920,706
Intra	TFIIB	155	B-core N-ter. cy	TFIIB	9	B-ribbon	726,357
Intra	TFIIB	155	B-core N-ter. cy	TFIIB	98	B-linker	635,056
Intra	TFIIB	155	B-core N-ter. cy	TFIIB	121	B-linker	722,019
Intra	TFIIB	155	B-core N-ter. cy	TFIIB	217	B-core N-ter. cyclin	804,059
Intra	TFIIB	155	B-core N-ter. cy	TFIIB	217	B-core N-ter. cyclin	603,296
Intra	TFIIB	161	B-core N-ter. cy	TFIIB	9	B-ribbon	676,684
Intra	TFIIB	161	B-core N-ter. cy	TFIIB	112	B-linker	542,791
Intra	TFIIB	161	B-core N-ter. cy	TFIIB	121	B-linker	672,35
Intra	TFIIB	211	B-core N-ter. cy	TFIIB	217	B-core N-ter. cyclin	778,436
Intra	TFIIB	217	B-core N-ter. cy	TFIIB	161	B-core N-ter. cyclin	453,035

Supplementary Table 3: Protein crosslinks, originating from the cross-reactivity. (formation of N-hydroxyl succinimide ester).

Link Type	Protein 1	Linked residue 1	Domain protein 1	Protein 2	Linked residue 2	Domain protein 2	M/z
Inter	Rpb1	15	Clamp core	Rpb4	4	Tip-associated	1091,89
Inter	Rpb2	889	Wall	Rpb1	369	Active site	904,472
Inter	Rpb7	29	Tip	Rpb4	4	Tip-associated	987,151
Inter	Tfg1	47	N-terminal	Rpb2	606	External 2	885,17
Inter	Tfg1	54	N-terminal	Rpb2	606	External 2	885,168
Inter	Tfg1	428	Charged region	Rpb9	20	Jaw	757,373
Inter	Tfg1	430	Charged region	Rpb9	20	Jaw	676,829
Inter	Tfg1	433	Charged region	Rpb9	20	Jaw	676,829
Inter	Tfg2	164	Insertion	Rpb2	1123	Hybrid binding	696,033
Inter	TFIIB	65	B-reader	Rpb2	864	Wall	686,296
Inter	TFIIB	67	B-reader	Rpb2	864	wall	686,295

List of abbreviations

Å	Angstrom
ATP	Adenosine triphosphate
Amp	Ampicillin
BRE	TFIIB recognition element
BLAST	Basic Local Alignment Search Tool
Ca	Candida albicans
C-terminus	Carboxy-terminus
CDK	Cyclin-dependent kinase
CID	C-terminal domain
cITC	core Initially Transcribing Complex
CPE	core promoter elements
CV	Column volume
CTD	Carboxy-terminal domain
CTDK-I	CTD kinase I
CTK1	CTD kinase subunit alpha
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DPE	Downstream promoter element
DSS	Di-succinimidyl-suberate
DTT	1,4-dithio-D,L-threitol
E. coli	Escherichia coli
EF	E longation factors

EM	Electron microscopy
EDTA	Ethylenediaminetetraacetic acid
EPE ligand	b4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
GTF	General transcription factor
His-tag	Histidine-tag
HMMs	Hidden Markov models
Hs	Homo sapiens
IPTG	Isopropyl-b-d-thiogalactoside
Inr	Initiator sequence
I-TASSER	Iterative Threading ASSEmbly Refinement
Kan	Kanamycin
Kp	Komagataella pastoris
MAD	Multi-wavelength anomalous diffraction
mRNA	messenger ribonuecleic acid
N-terminus	Amino-terminus
NEB	New England Biolabs
OD_{600}	Optical density at 600 nm
ORF	Open reading frame
pA	Polyadenylation
PAGE	Polyacrylamide gel electrophoresis
Pb	Paracoccidioides brasiliensis
PCR	Polymerase Chain Reaction
PDB	Protein data bank
PEG	Polyethylene glycol
PIC	Pre-initiation complex
Pol II	RNA Polymerase II

pTEF-b	positive transcription elongation factor
RMSD	Root-mean-square deviation
RNA	Ribonuecleic acid
Sc	Saccharomyces cerevisiae
SDS	Sodium dedocyl sulfate
SeMet	selenomethionine
Sm	Saccharomyces mikatae
Sp	Schizosaccharomyces pombe
TBP	Tata-box binding protein
TCEP	Tris(2-carboxyethyl)phosphin
TEC	Transcription elongation complex
TFII	Transcription factor II
WH	winged helix
XL-MS	Crosslinking coupled to mass spectrometry

List of figures

List of tables