Structural and Biochemical Characterizations of the Symplekin-Ssu72-CTD Complex in Pre-mRNA 3’ end Processing

Kehui Xiang

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014
ABSTRACT

Structural and Biochemical Characterizations of the Symplekin-Ssu72-CTD Complex in Pre-mRNA 3’ end Processing

Kehui Xiang

RNA polymerase II (RNAP II) transcribes essentially all messenger RNAs (mRNAs) in eukaryotes. The C-terminal domain (CTD) of its largest subunit contains consensus heptad repeats Y$_1$S$_2$P$_3$T$_4$S$_5$P$_6$S$_7$. Dynamic post-translational modifications of the CTD regulate RNAP II transcriptional activity and also facilitate transcription-coupled RNA processing events. One important mark is phosphorylation at Ser5 position, whose level peaks during transcription initiation but gradually diminishes toward the 3’ end of genes. Ssu72 is a known CTD pSer5 phosphatase. Recent studies identified a binding partner of Ssu72, symplekin, which is an essential scaffold protein in pre-mRNA 3’ end processing. Little is known about the molecular function of symplekin and neither do we understand how the symplekin-Ssu72 interaction couples pre-mRNA 3’ processing to transcription.

We first determined the crystal structure of the symplekin-Ssu72-CTD phosphopeptide complex. The N-terminal domain of symplekin embraces Ssu72 with its HEAT-repeat motif, serving as a typical molecular scaffold. Strikingly, the CTD phosphopeptide bound to the active site of Ssu72 has the peptide bond between pSer5 and Pro6 in the cis configuration, distinct from all known CTD conformations, which were exclusively in trans. While it was generally believed that only the trans peptide bond is recognized by proline-directed serine/threonine phosphatases or kinases, our discovery demonstrates for the first time that Ssu72 targets the energetically less-favorable cis peptide bond. In addition, we found that the binding of symplekin and also the
presence of a proline cis-trans isomerase can stimulate the phosphatase activity of Ssu72 in vitro. The symplekin-Ssu72 interaction as well as the catalytic activity of Ssu72 is required in our transcription-coupled polyadenylation assay. Overall, our study has important implications for the regulation of RNAP II transcription by cis-trans isomerization of the CTD and will help us understand how CTD modifications influence the recruitment of pre-mRNA 3’ end processing factors in a transcription-coupled manner.

Recent studies showed that Ssu72 is also a phosphatase of CTD pSer7, which is involved in small nuclear RNA transcription and 3’ end processing. However, a pSer7 phosphatase activity appears to be inconsistent with our structure because pSer7 is followed by Tyr1’ of the next repeat rather than a proline, and it is unlikely for the pSer7-Tyr1’ peptide bond to be in cis configuration. To solve this conundrum, we determined the crystal structure of the pSer7 CTD peptide bound to Ssu72. Surprisingly, the backbone of the pSer7 CTD runs in an opposite direction compared with the pSer5 CTD, allowing a trans pSer7-Pro6 peptide bond to be accommodated in the active site. However, Ssu72 has a much lower affinity for pSer7 than pSer5 and several structural features are detrimental for the catalytic activity towards pSer7. Consistent with these observations, our in vitro assays showed that the dephosphorylation of pSer7 by Ssu72 is ~4000-fold lower than that of pSer5. This further characterization of Ssu72 not only presents the first phosphatase in the literature that recognizes peptide substrates in both directions but also provides a more comprehensive understanding on CTD regulation by phosphatases from a structural perspective.

Another protein, Rtr1, was recently suggested to function as a pSer5 phosphatase in a zinc-dependent fashion, separately or redundantly with Ssu72. We solved the crystal structure of Rtr1 and discovered a new type of zinc finger with no close structural homologs. Unexpectedly,
Rtr1 does not present any evidence of an active site and it lacks detectable phosphatase activity in all our assays. We believe that, based on our results, Rtr1 does not have catalytic ability but instead indirectly regulate the phosphorylation state of the CTD.

In summary, our studies on the symplein-Ssu72-CTD complex as well as Rtr1 have revealed several novel structural features that are essential for the CTD regulation at the atomic level. These results will also shed light on understanding the mechanism by which RNAP II transcription and RNA processing are coupled.
# Table of Contents

Table of Contents ........................................................................................................................................... i

Acknowledgements ........................................................................................................................................ iv

Dedication ................................................................................................................................................ vi

Preface ...................................................................................................................................................... vii

Chapter I

Delineating the structural blueprint of the pre-mRNA 3’ end processing machinery

Introduction .................................................................................................................................................. 2

Pre-mRNA cis elements ............................................................................................................................... 4

*Trans*-acting protein factors ...................................................................................................................... 6

Cleavage and Polyadenylation Specificity Factor (CPSF) ............................................................................ 7

Cleavage stimulation Factor (CstF) ............................................................................................................... 17

Cleavage Factor I (CFI) ............................................................................................................................... 24

Cleavage Factor II (CFII) ............................................................................................................................. 28

Symplekin .................................................................................................................................................... 31

RNA Polymerase II C-terminal domain (RNAP II CTD) ............................................................................ 34

Poly(A) Polymerase (PAP) ........................................................................................................................ 38

Poly(A) binding protein (PABPN1) ............................................................................................................. 41

Perspective .................................................................................................................................................. 42
Chapter II

Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex

Abstract........................................................................................................................................75

Main text......................................................................................................................................75

References ....................................................................................................................................79

Methods ......................................................................................................................................80

Supplementary Information .......................................................................................................82

Chapter III

An unexpected binding mode for a Pol II CTD peptide phosphorylated at Ser7 in the active site of the CTD phosphatase Ssu72

Abstract........................................................................................................................................98

Results and Discussion................................................................................................................98

Materials and methods...............................................................................................................101

References ....................................................................................................................................102

Supplementary Information .......................................................................................................104

Chapter IV

The yeast regulator of transcription protein Rtr1 lacks an active site and phosphatase activity

Abstract........................................................................................................................................107
Results .......................................................................................................................... 108

Discussion ......................................................................................................................... 110

Methods ............................................................................................................................ 110

References .......................................................................................................................... 111

Supplementary Information .............................................................................................. 113
Acknowledgements

I would not have arrived here and started this journey if it wasn’t for my parents, who have been guiding me, encouraging me and supporting me to explore the scientific world ever since I was born. I owe the greatest debt of gratitude to my father, Zuozhi Xiang and my mother, Jun Li.

I would like to thank my thesis advisors, Dr. Liang Tong and Dr. James L. Manley, who co-mentored me throughout my Ph.D. training at Columbia University. They not only taught me skills and gave me suggestions, but also set great examples as scientists, which inspired me in every way. Their collaboration created a healthy and comfortable environment that facilitated my research and enriched my experience. I am truly grateful for their support.

I would also like to thank my thesis committee members, Dr. Lawrence Chasin and Dr. Joachim Frank, for their critical comments and helpful advice on my research projects, as well as Dr. Christopher D. Lima for taking the time to examine my thesis and attend my defense.

I want to thank past and current members of the Tong Lab, especially Turgay Kilic for teaching me molecular cloning and protein purification, Song Xiang for assisting me with solving structures, Yun Bai for helpful advice on bench work, Christine Huang for being a wonderful “bay-buddy”, Jeong-Ho Chang for sharing his scientific wisdom, Ruiying Wang for patiently talking and listening to me, Vivien Wang for caring about and spending time with me, Jia Wei for helping me with the EM study, Shukun Luo for bringing in new ideas and techniques, Timothy Tran and Linda Yu for all the joy they created in the lab, as well as other people for being friendly labmates.

I also want to thank former and current members of the Manley Lab, particularly Jing-ping Hsin for sharing his research and life experience and giving me helpful advice, Takashi
Nagaike for his help on in vitro assays, Nishta Rao, Charlotte Logan and Denis Lazarev for HeLa nuclear extract, Shuang Feng for her support with experimental materials, Dafne Campigli Di Giammartino for discussions about biochemical assays and other people for allowing me to enjoy a colorful lab life.

During my graduate career, my friends outside my work have been very thoughtful and helpful. I would like to thank Lin Hong, who spent five years with me closely, for her understanding and support, and for her being an incredible companion. I would also like to thank my “basketball-buddies”, Junpu Wang, Wenbo Li, Yinsheng Guo and Cheng Chen, who spent a lot of time with me sweating in the gym. Last but not the least, I want to thank my BFF, Qi Yao, who is always there, whenever, wherever.
To all, who pursue science and knowledge
Preface

This thesis is divided into five chapters. Chapter I reviews current knowledge of pre-mRNA 3’ end processing from a structural perspective. Chapter II is a research article about structural studies of the symplekin-Ssu72-CTD phosphopeptide complex, provided in the original form published in Nature. Chapter III is a research article that describes extended findings of Chapter II, also provided in the original form published in Genes Development. Chapter IV is a brief communication research article about structural and biochemical studies of Rtr1, in the original form published in Nature Communications. Chapter V summarizes our current research progress on biochemical characterizations of CPSF-73, CPSF-100 and symplekin complex.
Chapter I: Delineating the structural blueprint of the pre-mRNA 3’ end processing machinery
Introduction

The central dogma dictates that RNA serves to deliver genetic information from DNA to protein. In eukaryotic cells, messenger RNA (mRNA) is transcribed by RNA polymerase II (RNAP II) and nascent mRNAs, termed pre-mRNAs, undergo extensive modifications to become mature mRNAs that are subsequently exported out of the nucleus and translated into proteins. The modifications, including 5’ end capping, intron splicing and 3’ end processing, frequently happen co-transcriptionally and all contribute to ensure proper and efficient gene expression, transcript and genome stability as well as accurate translation and mRNA turnover (Bentley, 2002; Hirose and Manley, 2000; Li and Manley, 2006; Moore and Proudfoot, 2009; Proudfoot et al., 2002).

3’ end processing of most pre-mRNAs involves a two-step reaction: an endonuclease cleaves the pre-mRNA and a poly (A) polymerase synthesizes a poly adenosine tail on the cleaved upstream product. The only known exceptions are metazoan histone pre-mRNAs, which are cleaved but not polyadenylated. The seemingly simple process involves intricate cis elements on the transcript and a massive and complex machinery consisting of more than 20 protein polypeptides in yeast (Zhao et al., 1999a) and as many as 80 in human cells (Shi et al., 2009). Pre-mRNA 3’ end processing is critical for many cellular events, from upstream coupled to transcription and splicing, and downstream to mRNA export, stability, translation and even decay (Mandel et al., 2008; Moore and Proudfoot, 2009). Defects in 3’ end processing can have catastrophic consequences for the cell and have been associated with a variety of human diseases (Danckwardt et al., 2008; Zhao et al., 1999a). Moreover, 3’ end processing can serve as a means for gene expression regulation through alternative polyadenylation (APA; Di Giammartino et al., 2011; Tian and Manley, 2013). Recently, several genome-wide studies have shown that APA,
which involves use of different cleavage sites in the pre-mRNAs, is widely utilized in modulating mRNA transcript levels in diverse cell types and development stages, such as tumor cells, lymphocytes, iPS cells and embryogenesis (Ji and Tian, 2009; Ji et al., 2009; Mayr and Bartel, 2009; Sandberg et al., 2008). The general mechanisms governing APA are still largely unknown, and fully deciphering these will require detailed knowledge of pre-mRNA 3’ end processing at the molecular level.

Early characterizations of mRNA 3’ end formation focused on genetic and biochemical studies, using a combination of in vivo and in vitro approaches. Sub-complexes, protein binding partners and individual factors were dissected layer by layer to unravel the complexity of the process (Colgan and Manley, 1997; Proudfoot, 2011; Zhao et al., 1999a). In the past decade, structural studies have thrived and played a major role in helping us interpret the molecular assembly and function of this machinery (Mandel et al., 2008; Yang and Doublié, 2011). From single protein to protein-protein complex, protein-RNA complex, and the entire assembly, the ultimate goal is to delineate the structural blueprint of the whole pre-mRNA 3’ end processing machinery with details at atomic resolution, not only mapping associations within the machinery and probing connections to other coupled nuclear events, but also understanding how the machinery operates and how it is regulated. In this review, we will summarize protein factors involved in pre-mRNA 3’ end processing, with emphasis on structural information and protein interaction networks, and how they translate into knowledge about the mechanistic function of the process. While both mammalian and yeast systems have been extensively studied as models in the past, we will primarily focus on the mammalian pre-mRNA 3’ end processing complex and only discuss the yeast system for comparison when necessary.
**Pre-mRNA cis elements**

For 3’ end processing to occur at the correct place, pre-mRNAs need to contain conserved cis elements embedded in the primary sequence with defined information to guide the protein factors. The cleavage site has no apparent consensus, though it is often proceeded by a CA dinucleotide (Sheets et al., 1990). Accurate positioning of the 3’ end processing complex requires a combination of upstream and downstream cis elements. The AAUAAA hexamer polyadenylation signal (PAS) was the first element identified in playing such a role (Proudfoot and Brownlee, 1976). The PAS is highly conserved in mammalian mRNAs (Beaudoing et al., 2000; Proudfoot, 1991; Tian et al., 2005). It is located between 10 to 35 nucleotides upstream of the cleavage site (Chen et al., 1995; Fitzgerald and Shenk, 1981; Hu et al., 2005) and is bound by the 3’ end processing factor CPSF (Bardwell et al., 1991; Bienroth et al., 1991; Gilmartin and Nevins, 1991; Murthy and Manley, 1992). A few other close sequence variants of the PAS have also been recognized, such as AUUAAA, UAUAAA, AGUAAA and UAUAUA (Beaudoing et al., 2000). Together they are responsible for directing more than 95% of polyadenylation in human mRNAs (Tian et al., 2005).

A second major positioning cis element exists downstream of the cleavage site, generally within 30 nucleotides (MacDonald et al., 1994). Both the sequence and the location are less conserved compared with the PAS. The downstream element (DSE) was initially described as a single element with two forms, GU-rich (Hart et al., 1985; McLauchlan et al., 1985) and U-rich (Chou et al., 1994; Gil and Proudfoot, 1987). Subsequent computational analyses attributed them to two independent DSEs, which have distinct prevalence across eukaryotic species (Hu et al., 2005; Salisbury et al., 2006). While the GU-rich DSE peaks at around 10 nucleotides 3’ end to the cleavage site, the U-rich DSE has a preferred position 15 nucleotides further downstream (Hu
et al., 2005; Salisbury et al., 2006). Both DSEs are likely recognized directly by the 3’ end processing factor CstF (Beyer et al., 1997; Deka et al., 2005; Pérez Cañadillas and Varani, 2003; Takagaki and Manley, 1997).

Multiple UGUA consensus sequences can be located within 40-100 nucleotides upstream of many cleavage sites (Hu et al., 2005). They provide a specific binding site for the essential 3’ end processing factor CFI (Venkataraman et al., 2005; Yang et al., 2010). The UGUA elements also serve as the primary sequence determinant for polyadenylation in some mRNAs that lack the PAS (Venkataraman et al., 2005). A tripartite mechanism has been proposed in that the PAS, the DSEs and the UGUA elements act cooperatively in directing poly(A) site recognition (Venkataraman et al., 2005; Yang and Doublié, 2011).

Besides the core components, some auxiliary cis elements have also been identified computationally and experimentally, most of which play regulatory roles (Millevoi and Vagner, 2010). Two U-rich elements can be located in the regions flanking the PAS (Hu et al., 2005). One precedes the cleavage site by 40-100 nucleotides while the other immediately follows the PAS, by an average of 5 nucleotides, both of which have been observed in many mRNAs (Levitt et al., 1989; Wilusz and Shenk, 1988). In addition, G-rich and C-rich elements were found either upstream or downstream but relatively distant from the cleavage site (Hu et al., 2005). Some of them can stimulate polyadenylation (Bagga et al., 1995; Yonaha and Proudfoot, 2000) but others have no functional role reported.

In yeast, mRNAs have a distinct primary structure in poly(A) definition (see reviews Mandel et al., 2008; Zhao et al., 1999a). The cleavage site usually follows a pyrimidine and a stretch of adenosines (Heidmann et al., 1994), the position of which is defined by an A-rich positioning element (PE) located 10-30 nucleotides upstream (Guo and Sherman, 1995). PE has
a similar sequence consensus to PAS but assumes a less essential role. Strangely, PE is not recognized by the yeast counterpart of CPSF but rather the CstF homolog CFIA. Another UA-rich efficiency element (EE) is present further upstream. It is bound by CFIB and is mainly responsible for improving processing efficiency. Moreover, some conserved upstream and downstream U-rich elements have also been identified to synergistically enhance the cleavage and polyadenylation (Dichtl and Keller, 2001; Graber et al., 1999). However, protein factors that associate with these cis elements remain elusive.

**Trans-acting protein factors**

Mammalian pre-mRNA 3’ end processing can be reconstituted in vitro with exogenous RNA substrates and cell nuclear extracts (Moore and Sharp, 1985). This assay system provided a powerful means to identify active trans-acting components involved in the cleavage and polyadenylation. Early studies were devoted to biochemical fractionation and characterization. Five major factors were identified and separated during a series of purification steps: Cleavage and Polyadenylation Specificity Factor (CPSF), Cleavage stimulation Factor (CstF), Cleavage Factor I (CFI), Cleavage Factor II (CFII) and Poly(A) Polymerase (PAP) (Christofori and Keller, 1988; Gilmartin and Nevins, 1989; Takagaki et al., 1988, 1989). All these factors, with the exception of PAP, are multi-subunit protein complexes. They constitute a mega-dalton machinery acting on a seemingly simple two-step enzymatic reaction. While polyadenylation only requires CPSF and PAP, all factors are indispensable for efficient cleavage of the pre-mRNA.

Some other essential members of the 3’ end processing machinery were discovered subsequently, including poly(A) binding protein PABPN1, the RNAP II largest subunit C-
terminal domain (CTD) and symplekin (Wahle, 1991a; McCracken et al., 1997; Hirose and Manley, 1998; Takagaki and Manley, 2000). A recent systematic proteomic study revealed even more proteins that are associated with the 3’ end processing complex (Shi et al., 2009). Some of them are bona fide components, some serve regulatory roles and others may assist in coupling to other processes.

Compared to the mammalian 3’ end processing complex, the yeast machinery is assembled in a different way. Proteins are found in 3 main factors: Cleavage and Polyadenylation Factor (CPF), Cleavage Factor IA (CFIA) and Cleavage Factor IB (CFIB) (see reviews Mandel et al., 2008; Zhao et al., 1999a). The yeast processing complex contains several components that do not exist in the mammalian system and it also lacks some proteins in the machinery of mammals. Despite the evolutionary divergence, evident conservation and similarity still prevail at the protein level. For example, CPF contains all the homologous proteins from CPSF, and CFIA are composed of members that share high homology with those in CstF and CFII.

In the following sections, we will describe each mammalian pre-mRNA 3’ end processing factor and also their respective yeast homologs in more detail. A list of protein-protein interactions that have been studied is provided in Table 1. The protein structures that have been reported are summarized in Table 2.

**Cleavage and Polyadenylation Specificity Factor (CPSF)**

CPSF defines the specificity of pre-mRNA 3’ end processing and is required for both cleavage and polyadenylation (Christofori and Keller, 1988; Gilmartin and Nevins, 1989; Takagaki et al., 1989). CPSF also plays important roles in transcription coupling, as it is
recruited by TFIID to the transcription initiation complex and then it accompanies RNAP II throughout the transcription process to the 3’ end (Dantonel et al., 1997). CPSF weakly recognizes the PAS, but the binding is enhanced by the presence of CstF (Bardwell et al., 1991; Bienroth et al., 1991; Gilmartin and Nevins, 1991; Murthy and Manley, 1992). Early purifications of CPSF revealed that it contained four polypeptides: CPSF-160, CPSF-100, CPSF-73 and CPSF-30, all of which were named according to their estimated molecular weight on electrophoresis gels (Bienroth et al., 1991; Murthy and Manley, 1992). A later biochemical study identified a bona fide member, Fip1(Kaufmann et al., 2004), and more recently a sixth component of CPSF was proposed, WDR33 (Shi et al., 2009).

The CPSF subunit that gained most attention recently is CPSF73, largely because it turns out to be the essential endonuclease for the cleavage reaction that researchers have been hunting for three decades (Dominski, 2010). The earliest clue came from a sequence analysis showing that CPSF-73 belongs to the metallo-β-lactamase (MβL) superfamily, whose members are mostly hydrolases dependent on metal ions (Callebaut et al., 2002). Yeast cells with mutations of the putative residues for zinc binding in Ysh1 (yeast homolog of CPSF-73) are lethal (Ryan et al., 2004). Furthermore, zinc chelators added into HeLa cell nuclear extract significantly reduced and even abolished 3’ end cleavage (Ryan et al., 2004). Despite all the evidence pointing at CPSF-73, the definitive piece didn’t arrive until the crystal structure of CPSF-73 was determined.

The N-terminal region of CPSF-73 (residue 1-460) contains a canonical MβL domain with a four-layer αβ/βα sandwich (Figure 1). There are additional strands at the carboxy-terminal region of this domain, which are also present in CPSF-100 and RNaseZ, likely a unique characteristic for nucleases (Mandel et al., 2006a). A novel β-CASP domain featuring β strands
surrounded by α helices is inserted like a cassette into the MβL domain. In the active site, two zinc atoms are octahedrally coordinated to essential motifs within the MβL domain, similar to the binding mode in RNase Z but not other canonical MβLs (Li de la Sierra-Gallay et al., 2005; Mandel et al., 2006a). However, the active site is buried at the interface between the β-CASP domain and the MβL domain, whose access to the RNA substrate is severely restricted. This likely explains why the bacterially expressed N-terminal domain of CPSF-73 displayed minimal nuclease activity in vitro. Strikingly, calcium cation was able to activate this protein, through a mechanism that is not understood but was speculated to involve conformational changes triggered by the cation (Mandel et al., 2006a).

In recent years, many bacterial and archaeal CPSF-73 homologs belonging to the β-CASP family have been well studied and a handful of crystal structures have been determined, shedding light on the catalytic mechanism of CPSF-73 (Ishikawa et al., 2006; Li de la Sierra-Gallay et al., 2008; Mir-Montazeri et al., 2011; Nishida et al., 2010; Silva et al., 2011) (Figure 1B). Bacterial RNase J has a similar overall domain architecture as CPSF-73 (Figure 1A). The additional C-terminal region that was missing in the CPSF-73 structure is connected to the MβL domain through a long linker, indicating a certain degree of flexibility. Importantly, this C-terminal domain mediates the dimerization of RNase J and is crucial for its nuclease activity (Li de la Sierra-Gallay et al., 2008). Archaeal β-CASP proteins have extra KH motifs at the N-terminus, which are responsible for RNA recognition (Figure 1A). They form dimers through their extreme C-terminal region within the MβL domain, unlike RNase J (Mir-Montazeri et al., 2011; Nishida et al., 2010; Silva et al., 2011). Based on these studies, another question arises: Does CPSF-73 display a similar mechanism by which dimerization through the C-terminal domain is required to cleave the RNA? The structural information is not yet available, nor do we
know if CPSF-73 can self-associate, but the fact that full-length CPSF-73 purified from HEK293 cells was not active (Kolev et al., 2008) suggests that CPSF-73 likely employs a different mechanism. In yeast, the C-terminal domain of Ysh1 is required for survival and both cleavage and polyadenylation (Zhelkovsky et al., 2006). Other studies indicate that CPSF-73 makes direct contact with CPSF-100 and symplekin through the C-terminal domain (Dominski et al., 2005a; Zhelkovsky et al., 2006). It is therefore tempting to hypothesize that CPSF-73 and CPSF-100 (a sequence homolog of CPSF-73, see below) form a heterodimer to perform catalysis, or a third protein such as symplekin may be required to activate the nuclease activity of CPSF-73 (Dominski, 2007, 2010).

CPSF-73 is also involved in mammalian histone pre-mRNA 3’ end processing (Dominski et al., 2005b). UV-crosslinking experiments placed CPSF-73 in the vicinity of both the cleavage site and the downstream product in a U7 snRNP dependent manner, indicating CPSF-73 might be both an endonuclease and a 5’ to 3’ exonuclease in this case (Dominski et al., 2005b; Yang et al., 2009). Exonuclease activity was not detected with the N-terminal domain of CPSF-73 (Mandel et al., 2006a), though it has been observed with other β-CASP proteins (Dominski, 2007; Phung et al., 2013). Whether CPSF-73 is a genuine 5’to 3’ exonuclease is still under debate, but if it is, other partners might be required for this activity (Dominski, 2010).

Like CPSF-73, CPSF-100 also belongs to the β-CASP family and they share high sequence homology (Callebaut et al., 2002; Jenny et al., 1996) (Figure 1A). However, it lacks conserved motifs that are essential for catalytic activity and is thus considered to be inactive (Callebaut et al., 2002). The crystal structure of the N-terminal domain of yeast CPSF-100 homolog (Ydh1) has been determined. The domain organization of CPSF-100 is very similar to CPSF-73 and other β-CASP proteins, but equivalent motifs critical for zinc binding are missing
in CPSF-100, making it incapable of catalysis (Mandel et al., 2006a). The β-CASP domain in CPSF-100 is noticeably larger than that in CPSF-73. A segment in this region containing almost 200 residues was not observed in the structure, probably being proteolytically digested during crystallization (Mandel et al., 2006a, 2006b). All CPSF-100 proteins from different species have this highly hydrophilic region with variable length, but its biological significance is currently unknown.

The exact function of CPSF-100 is not clear, but several studies have shown its importance in pre-mRNA 3’ end processing. In yeast, Ydh1 is indispensable for cell viability and its mutants have negative effects on cleavage and polyadenylation (Kyburz et al., 2003; Preker et al., 1997). Ydh1 can also be UV-crosslinked to the pre-mRNA in a cis-element dependent manner and it may make direct contact with select RNA substrates (Dichtl and Keller, 2001; Zhao et al., 1997). GST pull-down assays indicate that Ydh1 interacts with the RNAP II CTD and Pcf11, raising its potential role in transcription coupled pre-mRNA processing (Kyburz et al., 2003). CPSF-100 is tightly associated with CPSF-73, likely through their C-terminal domains (Dominski et al., 2005a). CPSF-73, CPSF-100 and symplekin constitute a core complex in both pre-mRNA and histone pre-mRNA 3’ end processing (Sullivan et al., 2009). These results indicate that CPSF-100 may play crucial regulatory roles in modulating the nuclease activity of CPSF-73 through direct interaction or by recruiting other factors.

CPSF-30 is the smallest member in the CPSF complex. Earlier studies debated about whether it is part of CPSF, because it appeared less than stoichiometric in some preparations (Bienroth et al., 1991; Murthy and Manley, 1992) and its yeast homolog Yth1 did not appear to be in the same complex with other CPSF homologs (Zhao et al., 1997). Its identity was backed in later studies but its association with the rest members of CPSF seems weak (Jenny et al., 1994).
Despite this controversy, CPSF-30 is essential for cleavage and polyadenylation in both yeast and mammalian cells (Barabino et al., 1997, 2000).

**Figure 1.** Cleavage and Polyadenylation Specificity Factor (CPSF). (A) Domain organization of human CPSF subunits and two CPSF-73 homologs. (B) Structural comparison between CPSF-73 (PDB ID: 2I7T (Mandel et al., 2006a)) and its bacterial and archaeal homologs, RNase J (PDB ID: 3BK1 (Li de la Sierra-Gallay et al., 2008)) and MTH1203 (PDB ID: 2YCB (Silva et al., 2011)).
CPSF-30 consists of five CCCH zinc finger motifs and a CCHC zinc knuckle motif at the C-terminus that is not present in Yth1 (Barabino et al., 1997) (Figure 1A). Structures of these motifs in other proteins have been determined and characterized in RNA recognition (D’Souza and Summers, 2004; Hudson et al., 2004), strongly indicating that CPSF-30 may bind to the pre-mRNA. Indeed, CPSF-30 can be UV-crosslinked to RNA oligomers, with a preference for poly(U) sequences (Barabino et al., 1997; Jenny et al., 1994). Interestingly, a conserved U-rich element is often located next to the PAS (Hu et al., 2005), presenting a strong candidate for the CPSF-30 binding site in vivo. In yeast, Yth1 binds the pre-mRNA near the cleavage site (Barabino et al., 2000), and its RNA recognition ability can be impaired with the removal of either zinc finger motif 2 or 4 (Tacahashi et al., 2003).

The zinc fingers in CPSF-30 are also responsible for making contacts with other factors involved in pre-mRNA 3’ end processing. In fact, it is not uncommon that zinc finger motifs can mediate protein-protein interactions (Gamsjaeger et al., 2007). In the case of Yth1, the integrity of zinc finger 4 is crucial for binding to Fip1 as well as Ysh1 (Barabino et al., 1997, 2000; Tacahashi et al., 2003). These two interactions are aided respectively by the C-terminus and the N-terminus of Yth1 (Barabino et al., 2000; Tacahashi et al., 2003). Besides factors in the cleavage and polyadenylation complex, CPSF-30 was found to interact with the NS1A protein from influenza A virus, a mechanism employed by the virus to inhibit host pre-mRNA 3’ end processing (Nemeroff et al., 1998). The crystal structure of NS1A in complex with zinc finger 2 and 3 of CPSF-30 has been solved. The two zinc fingers show high structural similarity to other CCCH zinc finger proteins, in agreement with the proposed RNA recognition function (Das et al., 2008). CPSF-30 also binds to the body of RNAP II, likely responsible for the association of CPSF and RNAP II during transcription elongation (Nag et al., 2007).
The largest subunit, CPSF-160, can be UV-crosslinked to pre-mRNA in a PAS-dependent manner (Gilmartin and Nevins, 1989; Keller et al., 1991). Purified recombinant CPSF-160 protein was shown to bind RNA selectively but the affinity for the AAUAAA sequence was lower than intact CPSF, indicating other factors in CPSF may help stabilize the interaction (Murthy and Manley, 1995). The yeast homolog, Yhh1, is essential for both cleavage and polyadenylation (Dichtl et al., 2002).

CPSF-160 is composed of tandem WD40 repeats clustered into 3 major β-propellers (Neuwald and Poleksic, 2000) (Figure 1A). Though no structural information is available, CPSF-160 shares low sequence homology but similar domain architecture with DDB1, a scaffold protein for cullin-binding in E3 ubiquitination, whose structure has been well characterized in the literature (Angers et al., 2006; Li et al., 2006; Scrima et al., 2008). In fact, WD40 domain is one of the most abundant domains in eukaryotic proteomes and also the top protein interacting domain in human and yeast interactome databases (Stirnimann et al., 2010). Many β-propeller protein structures have been determined. They generally serve as scaffolds in protein-protein interactions (Stirnimann et al., 2010), but exceptions exist. For example, the WD40 domain of DDB2, a protein involved in DNA lesion repair, can recognize DNA (Scrima et al., 2008). This structural evidence explains the fact that CPSF-160 is involved in both protein-protein and protein-RNA interactions. In particular, CPSF-160 makes direct contacts with CstF-77 and interacts weakly with PAP (Murthy and Manley, 1995). The middle β-propeller domain of Yhh1 binds RNA (Dichtl et al., 2002).

CPSF-160 is also involved in transcription-coupled 3’ end processing. Yhh1 mutants gave rise to defects in transcription termination in yeast and Yhh1 interacts with phosphorylated RNAP II CTD in vitro and in vivo, though this interaction was not impaired in the mutants and
thus not the cause for the transcription termination defects (Dichtl et al., 2002). More importantly, CPSF-160 binds to several factors in TFIID, mainly accounting for CPSF being recruited to the RNAP II pre-initiation complex (Dantonel et al., 1997).

Fip1 was identified more than a decade later than the above mentioned CPSF members, though its yeast homolog had been known for a long time. It was shown that Fip1 could stably associate with all other components of CPSF. Immuno-depletion of Fip1 impairs both cleavage and polyadenylation, which can be rescued by CPSF (Kaufmann et al., 2004). In contrast, yeast Fip1 mutants affect polyadenylation but not cleavage (Preker et al., 1995). Human Fip1 is almost twice as large as its yeast counterpart and has a C-terminal extension containing two extra domains (Kaufmann et al., 2004) (Figure 1A). The N-terminal regions of the two proteins share similar domain organization but have relatively low sequence conservation (Kaufmann et al., 2004).

Yeast Fip1 was originally discovered in a yeast two-hybrid screen for binding partners of Pap1 (Preker et al., 1995). Their strong interaction tethers Pap1 to other 3’ end processing factors to ensure specificity in polyadenylation (Preker et al., 1995). A fragment of Fip1 (residue 80-105) has been co-crystallized with Pap1. In the complex structure, Fip1 binds to the C-terminal domain of Pap1, distant from the active site (Meinke et al., 2008). The binding of this peptide stabilizes Pap1, but it induces minimal structural changes and has little effect on the catalytic activity (Meinke et al., 2008). Mutations that disrupt this interaction cause lethality in yeast (Meinke et al., 2008). Interestingly, residues at the interface are not conserved in either Fip1 or Pap1 but the interaction between these two proteins has been observed across different eukaryotic species, including yeast, plants and human (Forbes et al., 2006; Meinke et al., 2008;
Preker et al., 1995). A more recent study showed that only the tethering but not the atomic details of these two proteins is essential for polyadenylation (Ezeokonkwo et al., 2011).

Human Fip1 has been shown to bind multiple partners. Besides PAP, Fip1 also binds to CPSF-30, CPSF-160 as well as CstF-77. These interactions are mainly mediated by its N-terminal region, which is homologous to its yeast counterpart, but the C-terminus may also play a role (Kaufmann et al., 2004). The C-terminal arginine-rich domain can also bind to RNA, particularly U-rich sequences (Kaufmann et al., 2004). This interaction likely mediates the Fip1-dependent stimulation of PAP in specific polyadenylation in vitro, because removal of the potential binding sites on the pre-mRNA abolishes the stimulatory effect (Kaufmann et al., 2004). In yeast, Fip1 inhibits Pap1 activity in non-specific polyadenylation, probably through a different mechanism due to its lack of the equivalent RNA binding domain (Helmling et al., 2001). In solution, Fip1 seems to be a largely disordered protein (Meinke et al., 2008), and its extended nature can provide scaffolding interactions with multiple proteins (Gunasekaran et al., 2003).

WDR33/WDC146 was proposed to be a bona fide CPSF component from a proteomic study (Shi et al., 2009). WDR33 is a 146kD protein that consists of an N-terminal WD40 domain, a middle collagen-like domain and a C-terminal GPR (glycine-proline-arginine) domain (Ito et al., 2001) (Figure 1A). It migrates at a position very close to CPSF-160 on SDS-gels, probably explaining why it was missed in earlier studies (Shi et al., 2009). Subsequent biochemical characterizations indicate WDR33 co-elutes with all CPSF members during gel filtration and is essential for cleavage in vitro (Shi et al., 2009). However, the exact function of this protein is not known.
Pfs2, the yeast homolog of WDR33, is not only required for cell survival but also critical for both cleavage and polyadenylation (Ohnacker et al., 2000). Pfs2 has a smaller size and only contains the WD40 domain (Ohnacker et al., 2000). As mentioned above, WD40 repeats are commonly involved in protein-protein interactions and indeed Pfs2 binds to many protein factors in the pre-mRNA 3’ end processing complex, including Rna14, Ysh1, Fip1 (Ohnacker et al., 2000) and Clp1 (Ghazy et al., 2012). Additionally, an S. pombe Pfs2 mutant was shown to have defects in transcription termination, suggesting a potential role in transcription-coupled polyadenylation (Wang et al., 2005).

**Cleavage stimulation Factor (CstF)**

CstF was initially purified as a factor not required for polyadenylation but significantly enhanced the cleavage reaction (Takagaki et al., 1989). However, CstF is now considered essential, and the initial studies likely reflected residual CstF in other fractions. CstF activity reflects in part the cooperative binding of CPSF and CstF to the pre-mRNA (Gilmartin and Nevins, 1991; Wilusz et al., 1990), in which CstF recognizes the DSEs (MacDonald et al., 1994). Like CPSF, CstF also associates with RNAP II during transcription elongation and facilitates transcription-coupled 3’ end processing (McCracken et al., 1997). Three proteins were found in the CstF complex: CstF-77, CstF-64 and CstF-50 (Gilmartin and Nevins, 1991; Takagaki et al., 1990).

A computational analysis identified the N-terminal region of CstF-77 as a HAT (half a TPR) domain (Figure 2A), because the sequence bears similar repetitive patterns to the TPR motif but lacks some conserved residues (Preker and Keller, 1998). The TPR (tetratrico-peptide repeat) is a 34-amino acid motif adopting helix-turn structure, which generally serve as
interacting modules with other proteins or themselves (Zeytuni and Zarivach, 2012). The crystal structure of the HAT domain of CstF-77 has been determined, revealing an intrinsic dimeric association (Bai et al., 2007; Legrand et al., 2007) (Figure 2B). The HAT domain dimerizes into an elongated shape, with each made up of a HAT-N domain and a HAT-C domain in a perpendicular organization (Bai et al., 2007). The HAT-C domain not only mediates CstF dimerization (Bai et al., 2007) but also makes contacts with other proteins, such as CPSF-160 (Bai et al., 2007) and CstF-50 (Takagaki and Manley, 2000). Subsequent biophysical and biochemical experiments confirmed the self-association of CstF-77, suggesting that the CstF complex assembles with two copies of each subunit (Bai et al., 2007). In fact, this is consistent with an earlier biochemical characterization (Takagaki and Manley, 2000) as well as genetic studies in flies (Benoit et al., 2002). Similar characteristics have also been observed in the fungal homolog of CstF-77, Rna14. The HAT domain of Rna14 from K. lactis dimerizes in mostly the same way as that of CstF-77, despite significant structural variations (Paulson and Tong, 2012). Disruption of this dimerization in yeast can severely impair both pre-mRNA cleavage and polyadenylation (Gordon et al., 2011).

CstF-77 interacts with both CstF-50 and CstF-64, in a way that bridges them since the other two factors make no direct contacts (Takagaki and Manley, 1994, 2000). The C-terminal part of CstF-77 contains a proline-rich region which is required for the binding to CstF-64 (Bai et al., 2007; Hockert et al., 2010; Takagaki and Manley, 2000). This interaction is both sufficient and necessary for the nuclear localization of CstF-64 (Hockert et al., 2010; Ruepp et al., 2011a) and may also facilitate RNA recognition by CstF-64 (see below). In yeast, Rna14 and Rna15 (CstF-64 homolog, see below) assemble through the same regions (Legrand et al., 2007; Moreno-Morcillo et al., 2011a; Paulson and Tong, 2012). With the dimerization of Rna14, they
constitute a $\alpha_2\beta_2$ tetramer in the shape of kinked rod (Noble et al., 2004). The complex structure of Rna14 hinge domain and Rna15 C-terminal domain has been obtained alone (Moreno-Morcillo et al., 2011a) and together with the HAT domain of Rna14 (Paulson and Tong, 2012) (Figure 2C). The two domains tether as an interlocked structure, through which they stabilize each other (Moreno-Morcillo et al., 2011a; Paulson and Tong, 2012). This formation requires cooperative folding between the two proteins and probably reflects their tight association in vivo (Moreno-Morcillo et al., 2011a). Additionally, there is a long linker connecting the HAT domain and the C-terminal domain of Rna14, making the two domains flexible and possibly functionally independent (Paulson and Tong, 2012).

CstF-64 was the first protein in the pre-mRNA 3’ end processing machinery shown to UV-crosslink to the RNA (Wilusz and Shenk, 1988). This effect, later found to be independent of the PAS, can be attributed to an RNA Recognition Motif (RRM) at its N-terminus (Takagaki et al., 1992) (Figure 2A). Further investigation revealed that the RRM can specifically recognize the U-rich DSE (MacDonald et al., 1994) and it selects GU-rich sequences in vitro (Takagaki and Manley, 1997). The structure of the RRM has been determined. By comparative analysis with other RRM structures, a binding pocket was identified at the surface of the central $\beta$-sheet to accommodate UU dinucleotides, whose presence could enhance the RNA-RRM interaction appreciably (Pérez Cañadillas and Varani, 2003). It was believed that with fine-tuning of contacts outside this pocket, the RRM is able to modulate its preference for a wide selection of Gs and Us while still discriminating against other RNA sequences (Pérez Cañadillas and Varani, 2003). This specific binding variability enables CstF-64 to recognize both U-rich and GU-rich DSEs. In fact, the two DSEs are in close distance (within 15 nucleotides) and might be bound by two copies of CstF-64 simultaneously, which are readily bridged by the long HAT domains of
CstF-77. This also supports the dimeric association of CstF, constructed around the CstF-77 HAT domain dimer.

The yeast homolog of CstF-64, Rna15, not only bears high sequence identity but also shares structural similarity in the RRM region. However, Rna15 also has some distinct characteristics. The RRM of Rna15 has a generally lower affinity for the RNA (Noble et al., 2004) though it preferentially binds to U-rich or GU-rich consensus in vitro (Pancevac et al., 2010; Takagaki and Manley, 1997). There are two RNA binding sites identified in the RRM structure of Rna15 (Pancevac et al., 2010) (Figure 2E). The primary site is located at a surface loop, whose interactions with the RNA are mediated by specific hydrogen bonding. Such a configuration would not form if A or C nucleotides were placed at the same location, largely explaining the specificity for GU-rich sequences (Pancevac et al., 2010). The second site is positioned at a canonical RRM-RNA contact region as in CstF-64 (Pancevac et al., 2010; Pérez Cañadillas and Varani, 2003). This interaction is facilitated predominantly by base-aromatic ring stacking and thus less specific. In a different scenario, with the presence of Rna14 or Hrp1 (a yeast CFI component that binds the EE, no mammalian homolog), the binding of the RRM of Rna15 to the RNA can be considerably enhanced and directed to the PE in yeast (Gross and Moore, 2001b; Noble et al., 2004). The structure of the Rna15-Hrp15-RNA complex gave a clear atomic view of how this recognition occurs (Figure 2F). The two RRMbs of Hrp1 not only specifically recognize the UA-rich RNA sequence but also interacts with Rna15, which brings the RRM of Rna15 adjacent to the A-rich PE element (Leeper et al., 2010). As a result of the cooperative binding, Rna15 is able to attach to this sequence in a canonical manner through the less specific site. In fact, the two distinct RNA binding modes of Rna15 can be well suited in the context of the yeast CFI complex. Recent studies have shown that there are two copies of Rna15
but only one copy of Hrp1 (Barnwal et al., 2012; Gordon et al., 2011). It is possible that one Rna15 in complex with Hrp1 associates with the EE while the other Rna15 alone recognizes a not-so-far U-rich consensus upstream of the cleavage site (Dichtl and Keller, 2001; Graber et al., 1999), with each involving different binding mechanisms (Leeper et al., 2010; Pancevac et al., 2010).

Following the RRM in CstF-64 is a hinge region that interacts with both CstF-77 and symplekin, in a mutually exclusive manner (Ruepp et al., 2011a; Takagaki and Manley, 2000, see below). The CstF-64-CstF-77 interaction is conserved between human and yeast. Interestingly, in the Rna14-Rna15 complex structure, the helix immediately preceding the hinge region from Rna15 displays high flexibility and can be locally unfolded (Moreno-Morcillo et al., 2011a). A similar characteristic of the same helix was also observed in both CstF-64 and Rna15 RRM structures (Pancevac et al., 2010; Pérez Cañadillas and Varani, 2003). This connecting helix maintains contacts with the central β-sheet of the RRM in a way that it blocks the RNA binding site (Pancevac et al., 2010; Pérez Cañadillas and Varani, 2003). Such a feature of the C-terminal α-helix extension is not uncommon in a variety of RRM structures (Dominguez et al., 2010; Nagai et al., 1990; Oubridge et al., 1994; Schellenberg et al., 2006; Selenko et al., 2003). It is likely that the partner association to the hinge region of CstF-64/Rna15 triggers unfolding of this helix, making the RRM accessible to RNA. In fact, it has been observed that the binding of Rna14 substantially boosts the Rna15-RNA interaction (Noble et al., 2004), which can be further augmented by the dimerization of Rna14 (Gordon et al., 2011).
**Figure 2.** Cleavage stimulation Factor (CstF). (A) Domain organization of human CstF subunits. (B) Crystal structure of the murine CstF-77 HAT domain (PDB ID: 2O0E (Bai et al., 2007)). (C) NMR structure of the Rna14 C-terminal Rna15-binding domain in complex with the hinge region of Rna15 (PDB ID: 2L9B (Moreno-Morcillo et al., 2011a)). (D) Crystal structure of the dimerization domain of CstF-50 (PDB ID: 2XZ2 (Moreno-Morcillo et al., 2011b)). (E) Crystal structure of the RRM domain of Rna15 with RNA bound at primary and secondary sites (PDB ID: 2X1A, 2X1F (Pancevac et al., 2010)). (F) NMR structure of the RRM of Rna15, two RRMs of Hrp1 and RNA ternary complex (PDB ID: 2KM8 (Leeper et al., 2010)).
Compared to the first two domains, the remaining part of CstF-64 is less well studied. The domain at the very C-terminus of CstF-64 is highly conserved in eukaryotes. It comprises a three α-helix bundle structure, whose integrity is required for pre-mRNA 3’end processing and the interaction with Pef11 (Gross and Moore, 2001a; Qu et al., 2007). In the middle lies a proline-glycine rich region (40%), inserted with a MEAR(A/G) pentapeptide repeat motif that is all α-helical (Richardson et al., 1999; Takagaki et al., 1992). This region does not exist in Rna15 and its function is currently unspecified.

CstF-50 is the only CstF subunit that does not have a yeast counterpart. This protein contains an N-terminal dimerization domain and seven WD40-repeats at its C-terminus (Takagaki and Manley, 1992, 2000) (Figure 2A). The dimerization domain is crucial for its self-association (Takagaki and Manley, 2000) and together with CstF-77 accounts for the hexameric architecture of the whole CstF complex. The crystal structure of this domain suggests a formation of 3 tandem α helices tethered to the other protomer, mediated primarily through a conserved hydrophobic core (Moreno-Morcillo et al., 2011b) (Figure 2D). A highly conserved region was also identified on the surface of the dimer and hence proposed to bind other proteins (Moreno-Morcillo et al., 2011b). In addition, the N-terminal domain is necessary for CstF-50 to interact with RNAP II CTD (McCracken et al., 1997), although they seem to only associate weakly if the rest of CstF-50 is removed (Moreno-Morcillo et al., 2011b). This interaction places CstF-50 in an important position in coupling transcription to 3’ end processing.

Similar to CPSF-160 and WDR33, the WD40 domain of CstF-50 functions as a binding platform. Truncation of a single repeat abrogates its interaction with CstF-77 (Takagaki and Manley, 2000). This domain may also serve as a regulatory adaptor. In particular, BARD1-BRCA1, a breast cancer related protein complex that functions in DNA damage repair (Irminger-
Finger and Leung, 2002), binds to the WD40 domain of CstF-50 upon DNA damage, and thereby strongly but transiently inhibits pre-mRNA 3’ end processing (Kleiman and Manley, 1999, 2001). It was proposed that this can repress aberrant mRNA 3’ end processing during transcription, probably at the damaged DNA positions (Kleiman and Manley, 2001).

**Cleavage Factor I (CFI)**

CFI is necessary for cleavage but not polyadenylation (Takagaki et al., 1989). ChIP experiments indicate that CFI associates early with the transcription elongation complex, alongside with CPSF and CstF, in facilitating transcription-coupled 3’ end processing (Venkataraman et al., 2005). At the polyadenylation site, CFI recognizes pre-mRNAs and stabilizes the binding of CPSF (Rüegsegger et al., 1996). Unlike other major 3’ end processing factors, CFI exists only in metazoans and does not have a yeast equivalent.

The CFI complex is assembled as a heterotetramer with a dimer of the small subunit CFI25 and another two copies of a combination of the large subunits, CFI59, CFI68 or CFI72 (Kim et al., 2010; Rüegsegger et al., 1996, 1998). CFI59 and CFI68 are encoded by two paralogous genes while CFI72 is an isoform of CFI68, possibly generated by alternative splicing (Ruepp et al., 2011b). The three large subunits may be functionally redundant because CFI68 alone is capable of reconstituting CFI activity with CFI25 in vitro (Rüegsegger et al., 1998). Meanwhile, all CFI subunits can be UV-crosslinked to pre-mRNAs, suggesting their critical roles in RNA recognition (Rüegsegger et al., 1996). SELEX experiments identified a binding consensus sequence for CFI: UGUAN (N: A > U > G or C) (Brown and Gilmartin, 2003), which was later shown to be an important cis element for poly(A) site definition (Hu et al., 2005; Venkataraman et al., 2005).
Figure 3. Cleavage Factor I (CFI). (A) Domain organization of human CFI subunits. (B) Crystal structure of the CFI25-CFI68-RNA complex (PDB ID: 3Q2T (Yang et al., 2011a)).

CFI25 encompasses a central Nudix domain (Dettwiler et al., 2004) (Figure 3A). The Nudix superfamily is widespread in all life kingdoms and characterized as mostly pyrophosphohydrolases that target nucleoside diphosphate linked to another moiety X (McLennan, 2006). These proteins generally perform housekeeping function by removing toxic metabolites in the cell (Bessman et al., 1996). Intriguingly, two signature glutamates that are key for metal coordination and enzymatic catalysis are missing in the Nudix motif of CFI25 (GX₅EX₇REUXEXGU, U: hydrophobic residue, X: any residue), which therefore distinguishes
CFI25 from most other Nudix proteins (McLennan, 2006). The crystal structure of CFI25 shows a core Nudix domain featuring a canonical α/β/α fold sandwich augmented with N- and C-terminal extensions (Coseno et al., 2008; Trésaugues et al., 2008) (Figure 3B). No metal ions were observed around the Nudix motif and subsequent biochemical assays failed to detect any enzymatic activity, suggesting that CFI25 is unlikely to be a hydrolase (Coseno et al., 2008; Trésaugues et al., 2008).

Despite the lack of hydrolase activity, CFI25 utilizes its Nudix domain as a platform for binding other proteins, nucleotides and more importantly, RNA. CFI25 interacts with CFI68, PAP and PABPN1 (Dettwiler et al., 2004). It can also bind ATP and Ap₄A, albeit with relatively low affinity (Coseno et al., 2008). Ap₄A was captured in a structure with CFI25, located in a binding pocket outside of the Nudix motif (Coseno et al., 2008). The crystal structure of the CFI25-RNA complex has been determined, providing insights into the mechanism by which CFI specifically recognizes the UGUA element (Yang et al., 2010). Unexpectedly, in addition to the Nudix domain, the N-terminal extension also makes direct contact with the RNA (Figure 3B). Interactions are achieved mainly through hydrogen bonds formed between RNA bases and the main chain or side chains of the protein. Watson-Crick/sugar-edge base interactions within the RNA also contribute to the binding specificity (Yang et al., 2010). Moreover, the dimeric nature of CFI25 enables it to bind two copies of UGUA elements simultaneously, providing an approach through which the poly(A) site can be accurately defined and regulated (see below).

The large subunit CFI68 is composed of an N-terminal RRM, a middle proline-rich region and a C-terminal RS domain with alternating arginine and serine residues (Rüegsegger et al., 1998) (Figure 3A). The domain organization resembles SR proteins involved in pre-mRNA splicing. In fact, CFI68 was shown to co-purify with the spliceosome (Awasthi and Alwine,
2003; Rappsilber et al., 2002; Zhou et al., 2002) and interacts with splicing factors, such as U2AF65 and other SR proteins (Dettwiler et al., 2004; Millevoi et al., 2006), suggesting its potential role in coordinating pre-mRNA splicing and 3’ end processing. CFI68 also interacts with the U7 snRNP and participates in histone mRNA processing (Ruepp et al., 2010).

The RRM of CFI68 interacts weakly with the RNA but it can be enhanced substantially with the cooperative binding of CFI25 (Dettwiler et al., 2004). In the crystal structure of the CFI25-CFI68-RNA complex, each CFI68 monomer maintains simultaneous interactions with the CFI25 dimer at the opposite side (Li et al., 2011; Yang et al., 2011a) (Figure 3B). The presence of CFI68 causes little structural change to the CFI25 dimer, nor does it affect the RNA binding specificity (Li et al., 2011; Yang et al., 2011a). Two UGUA RNA sequences are each bound to the monomer of CFI25 in an antiparallel fashion. The connecting loop sequence, though not included in the structure, is likely to be stabilized by the RRM of CFI68 (Li et al., 2011; Yang et al., 2011a). Interestingly, the interaction between the RRM and the RNA substrate does not seem to be mediated through the canonical RNA binding site on the central β-sheet, as mutations of the conserved putative binding residues had modest effects on the affinity (Li et al., 2011; Yang et al., 2011a). Further mutagenesis analysis revealed that the looping of the RNA is facilitated by residues in a cleft adjacent to the CFI25-CFI68 interface (Yang et al., 2011a). More importantly, based on the structural model, a RNA looping mechanism directed by CFI has been proposed for poly(A) selection in APA (Yang et al., 2011a, 2011b). Simply put, the CFI complex can take advantage of different UGUA elements surrounding multiple PAS elements and loop out an entire AAUAAA region, resulting in polyadenylation occurring at a distal site (Yang et al., 2011a, 2011b). This model is consistent with the evidence from several studies indicating that
down-regulation of CFI level in cells correlates with widespread usage of proximal poly(A) sites (Kim et al., 2010; Kubo et al., 2006; Martin et al., 2012; Sartini et al., 2008).

**Cleavage Factor II (CFII)**

Like CFI, CFII is also only required for cleavage but dispensable for polyadenylation (Takagaki et al., 1989; de Vries et al., 2000). CFII is arguably the least well-characterized factor in the mammalian 3’ end processing complex, partly because its exact components remain poorly defined. In an early purification, more than 15 proteins were found in the CFII fraction, although most of them may be cross contamination or only play stimulatory roles. Only two proteins, Pcf11 and Clp1 co-eluted with the CFII activity (de Vries et al., 2000). Both of these have yeast homologs that are essential for yeast pre-mRNA 3’ end formation (Amrani et al., 1997a; Minvielle-Sebastia et al., 1997). However, there is no evidence indicating that these two proteins alone are capable of reconstituting the CFII activity.

Most of our understanding on CFII comes from parallel studies in yeast. Both Pcf11 and Clp1 are part of the CFIA complex, which also includes Rna14 and Rna15 (CstF homologs). Pcf11 plays a pivotal role in the assembly of CFIA as it is the only component in the complex that makes direct contacts with all other members. The interacting domains have been mapped to the middle and C-terminal regions (Amrani et al., 1997a; Gross and Moore, 2001a; Qu et al., 2007; Sadowski et al., 2003) (Figure 4A). Two highly conserved zinc fingers were identified flanking the C-terminal Clp1-interating domain but their function remains unknown. There is also a stretch of 20 consecutive glutamines preceding the middle Rna14-Rna15 binding domain, likely to serve as a molecular linker to the N-terminal region (Sadowski et al., 2003).
Figure 4. Cleavage Factor II (CFII). (A) Domain organization of the yeast homologs of CFII subunits. (B) Crystal structure of the CID of Pcf11 in complex with a CTD peptide (PDB ID: 1SZA (Meinhart and Cramer, 2004)). (C) Conformation of the CTD peptide bound to the CID of Pcf11. (D) Crystal structure of the Clp1-Pcf11-ATP complex (PDB ID: 2NPI (Noble et al., 2007))

The N-terminus of Pcf11 features an RNAP II CTD interacting domain (CID), which comprises eight α-helices arranged in a right-handed superhelical formation (Meinhart and Cramer, 2004; Noble et al., 2005) (Figure 4B). The CID interacts with both unphosphorylated and phosphorylated CTD, but it has a higher affinity with the latter (Barillà et al., 2001; Licatalosi et al., 2002; Sadowski et al., 2003). Surprisingly, the CID-CTD interaction is not necessary for pre-mRNA 3’ end processing, but it is essential for proper transcription termination (Barillà et al., 2001; Sadowski et al., 2003). The CID also weakly binds to RNA (Hollingworth et
al., 2006) and bridges the CTD to the pre-mRNA in a way that the transcription elongation complex can be dismantled (Zhang et al., 2005), supporting a role for Pcf11 in coupling transcription to 3’ end processing.

Human Pcf11 is twice as large as its yeast homolog and only shares sequence homology at its N-terminal CID (de Vries et al., 2000). The remainder of human Pcf11 has not been characterized. Despite apparent differences in the primary sequence, the function of Pcf11 is evolutionarily conserved. Knockdown of Pcf11 in HeLa cells resulted in deficiency in both pre-mRNA cleavage as well as transcription termination and Pcf11 may also be required for degradation of the 3’ end product following cleavage (West and Proudfoot, 2008).

Clp1 consists of a central ATPase domain and two smaller domains at its N- and C-terminus (Noble et al., 2007) (Figure 4A). The primary sequence of yeast Clp1 reveals a conserved Walker A/P loop motif (de Vries et al., 2000), which is typically involved in ATP/GTP binding or catalysis (Walker et al., 1982). Indeed, an ATP molecule was bound to Clp1 in the crystal structure (Figure 4D). However, ATPase activity could not be detected for Clp1, either on its own or with its binding partners (Noble et al., 2007). Several recent studies showed that certain mutations introduced in the ATP-binding pocket could perturb the Clp1-Pcf11 interaction and in turn cause defects in 3’ end processing and transcription termination (Ghazy et al., 2012; Haddad et al., 2012; Holbein et al., 2011), although some of these mutations did not affect ATP binding. In the complex structure, the interface between Clp1 and Pcf11 is distant from the ATP-binding site. It is possible that ATP acts as a molecular switch that can regulate or be affected by the binding partners of Clp1 (Ghazy et al., 2012). Nonetheless, further investigation is needed to understand the function of Clp1 ATP binding capability.
While yeast Clp1 does not possess catalytic activity, the human homolog is an active 5’-OH polynucleotide kinase (Weitzer and Martinez, 2007a). The enzymatic activity is required in tRNA splicing and can not be complemented by yeast Clp1 (Ramirez et al., 2008; Weitzer and Martinez, 2007a, 2007b). Although human Clp1 bears high sequence homology with its yeast counterpart, it might have acquired a diverged role during evolution. In pre-mRNA 3’ end processing, Clp1 interacts with both CPSF and CFI and likely functions to tether them to CFII (de Vries et al., 2000).

**Symplekin**

Symplekin was initially identified as a tight junction plaque protein but it is ubiquitously expressed in all cell types (Keon et al., 1996). Its association with the pre-mRNA 3’ end processing complex was uncovered in a study looking for binding partners of CstF-64 (Takagaki and Manley, 2000). Sequence alignment indicated that symplekin bears low primary structure homology to an essential yeast cleavage and polyadenylation factor Pta1 (Ghazy et al., 2009; Preker et al., 1997; Takagaki and Manley, 2000; Zhao et al., 1999b). Pta1 interacts with various 3’ end processing proteins, including Ysh1 (Zhelkovsky et al., 2006), Ydh1 (Kyburz et al., 2003), Fip1 (Ghazy et al., 2009), Pcf11 (Kyburz et al., 2003), Clp1 (Ghazy et al., 2009), Pap1 (Ghazy et al., 2009), and Ssu72 (He et al., 2003). Pta1 (and also symplekin) is thus recognized as a scaffold protein composed of assorted protein-interacting modules, coordinating the assembly of the machinery and tethering relevant components to increase functional efficiency.

Secondary structure analysis suggests that symplekin consists of all α helices. Based on the sequence conservation, partner binding region and homology to Pta1, symplekin can be divided into four distinctive domains: an N-terminal Ssu72 (see below) binding domain (NTD),
CstF-64 binding domain, CPSF-73 binding domain and a C-terminal domain that is not present in Pta1 and has no known function (Ghazy et al., 2009; Xiang et al., 2010) (Figure 5A).

Symplekin NTD contains seven pairs of anti-parallel α helices arranged in the shape of an arc (Kennedy et al., 2009; Xiang et al., 2010). The overall fold is reminiscent of the ARM or HEAT repeats, which is typically involved in protein-protein interactions (Andrade et al., 2001), in agreement with its suggested scaffolding function. Similar to Pta1, symplekin NTD interacts with Ssu72, an RNAP II CTD phosphatase (Ghazy et al., 2009; Xiang et al., 2010). The crystal structure of the symplekin-Ssu72 complex has been determined. Ssu72 binds to symplekin NTD at the concave surface, mediated by ion-pair, hydrogen-bonding as well as hydrophobic interactions (Xiang et al., 2010) (Figure 5B). Interestingly, although the active site of Ssu72 is remote from the contact surface, the binding of symplekin NTD can stimulate the phosphatase activity of Ssu72, likely through stabilization of the enzyme and/or allosteric regulation (Xiang et al., 2010). These results suggest that symplekin may not simply be a passive scaffold, but can also play a regulatory role in the 3’ end processing complex. In fact, many scaffold proteins in other systems have been reported to assume similar modulating function (Gold et al., 2006; Good et al., 2011; Reyes-Turcu et al., 2009).

The N-terminal domain of Pta1 is not required for in vitro pre-mRNA 3’ end processing but the addition of it can inhibit polyadenylation (Ghazy et al., 2009). Symplekin NTD displays a similar inhibitory effect when polyadenylation is coupled to transcription (Xiang et al., 2010). Ssu72 was able to alleviate the reduction while a catalytically inactive mutant failed, suggesting that the enzymatic activity of Ssu72 is indispensible for efficient transcription-coupled polyadenylation. A model was proposed in that symplekin brings Ssu72 to remove
phosphorylation of Ser5 residues in RNAP II CTD, so that other 3’ end processing factors can be recruited to promote polyadenylation (Xiang et al., 2010).

Symplekin interacts with the hinge region of CstF-64, competitively with CstF-77 (Takagaki and Manley, 2000; Ruepp et al., 2011a). A CstF-64 mutant whose interaction with symplekin was abolished maintained its association with CstF-77 (Ruepp et al., 2011a). This mutant did not affect general pre-mRNA cleavage and polyadenylation but impaired histone pre-mRNA 3’ end processing (Ruepp et al., 2011a). It appears that CstF-64 associates exclusively with either CstF-77 or symplekin in two separate pre-mRNA processing complex and perform distinct functions. Interestingly, the equivalent region of Pta1 does not interact with Rna15 but instead binds to Pti1 (Ghazy et al., 2009; Zhelkovsky et al., 2006), an Rna15 paralog involved in small nucleolar RNA 3’ end processing (Dheur et al., 2003).

Symplekin also interacts with CPSF-73. The binding region was deduced based on sequence conservation with Pta1 and the fact that the C-terminal domain of Pta1 interacts with the C-terminal domain of Ysh1 (Zhelkovsky et al., 2006; Ghazy et al., 2009). Symplekin tightly associates with CPSF-73 as well as CPSF-100 (Hofmann et al., 2002; Sullivan et al., 2009). They comprise a shared stable core complex for both general and histone pre-mRNA 3’ end processing (Sullivan et al., 2009). As a consequence, it has been speculated that symplekin may regulate the nuclease activity of CPSF-73 through direct interactions (Dominski, 2010; Sullivan et al., 2009), but further investigation is necessary to test this hypothesis.

Besides general mRNA 3’ end processing, symplekin participates in a variety of important cellular processes. Symplekin provides the scaffold in histone mRNA 3’ end processing (Kolev and Steitz, 2005). It is also required for CPEB (Cytoplasmic polyadenylation element binding protein)-mediated cytoplasmic polyadenylation (Barnard et al., 2004).
Furthermore, symplekin has been implicated in maintaining mitotic fidelity by supporting microtubule dynamics (Cappell et al., 2010) as well as promoting tumorigenesis in human intestinal epithelial cells (Buchert et al., 2010).

**RNA polymerase II largest subunit C-terminal domain (RNAP II CTD)**

The largest subunit of RNAP II contains an extended C-terminal domain that separates it from the globular core structure (Cramer et al., 2001). The CTD consists a stretch of consensus repeats Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7 (Allison et al., 1985; Corden et al., 1985). The number of the repeats varies across eukaryotic organisms, in particular 52 in vertebrates and 26 in yeast (Chapman et al., 2008; Liu et al., 2010). The flexible and repetitive nature of the CTD enables it to provide a perfect platform for the binding of a variety of protein factors, through which RNAP II transcription can be coordinated with other nuclear events (reviewed in Bartkowiak et al., 2011; Hsin and Manley, 2012; Zhang et al., 2012; Heidemann et al., 2013).

The CTD was first linked to 3’ end processing in a study showing that the truncation of the CTD from RNAP II resulted in defects in cleavage and polyadenylation in vivo (McCracken et al., 1997). The CTD was shown to be essential for efficient 3’ end processing in vitro (Hirose and Manley, 1998). Exactly how the CTD promotes mRNA 3’ end formation is still not well understood. A platform role has been proposed since a number of 3’ end processing factors have been observed binding to the CTD, such as Ydh1 (Kyburz et al., 2003), Yhh1 (Dichtl et al., 2002), CstF-77/Rna14 (McCracken et al., 1997; Sadowski et al., 2003), CstF-50 (McCracken et al., 1997), Pcf11 (Barillà et al., 2001; Licatalosi et al., 2002; Meinhart and Cramer, 2004; Sadowski et al., 2003) and perhaps Rna15 (Sadowski et al., 2003) and Pta1 (Rodriguez et al., 2000). The association between the CTD and some 3’ end processing factors can be traced back
to the promoter region and is likely maintained through elongation to the 3’ end (Dantonel et al., 1997; Nag et al., 2007; Kim et al., 2010; Mayer et al., 2010, 2012).

The ability of the CTD to orchestrate transcription to 3’ end processing and other mRNA processing events is primarily a result of its structural diversity, which can be summarized in three ways. First, the largely unfolded repeats create significant variations in the 3D structure. The free CTD was initially proposed to form β-turns (Suzuki, 1989) and this conformation has also been observed in both NMR experiments (Cagas and Corden, 1995; Harding, 1992) and crystal structures (Meinhart and Cramer, 2004). Later studies however indicate that the CTD adopts a non-uniform and overall disordered architecture in solution (Bienkiewicz et al., 2000; Noble et al., 2005), allowing various interactions formed with other proteins through an induced fit mechanism (see review Meinhart et al., 2005). Second, post-translational modifications on signature residues can affect CTD recognition of other factors by altering its chemical structure (see reviews Jasnovidova and Stefl, 2013; Meinhart et al., 2005). It is now clear that all five residues with a hydroxyl group (Tyr1, Ser2, Thr4, Ser5 and Ser7) can be phosphorylated, generating a combinatorial CTD code. This code not only changes dynamically throughout the transcription cycle but it also patterns in a gene-specific manner (see reviews: Egloff et al., 2012; Heidemann et al., 2013; Hsin and Manley, 2012; Zhang et al., 2012). In addition to phosphorylation, other modifications have also been identified on the CTD, such as glycosylation (Ranuncolo et al., 2012), methylation (Sims et al., 2011) and ubiquitination (Li et al., 2007), further complicating the CTD code. Third, the peptide bonds of Ser2-Pro3 and Ser5-Pro6 are subject to cis-trans isomerization, which can produce drastic conformational changes to the CTD structure, allowing selective binding of processing factors (see review Jasnovidova and
Stefl, 2013). All three layers together make the CTD so structurally versatile and elaborate that it is able to coordinate related processes to transcription in a coupled manner.

Although structural information of the complete CTD is not available, a number of pieces ranging from less than one repeat to nearly three repeats have been captured associated with CTD binding proteins. These CTD structures present immensely diverse conformations and modifications (see reviews Jasnovidova and Stefı́l, 2013; Meinhart et al., 2005). Here we will only briefly discuss two that are closely related to pre-mRNA 3’ end processing.

The first is the Pcf11-pSer2 CTD complex (Figure 4B). In this structure, the bound CTD adopts a β-turn conformation (Meinhart and Cramer, 2004), which however is likely formed via induced fit through the binding of Pcf11, because NMR experiments suggest that a similar CTD peptide exists as a dynamic unfolded ensemble in solution (Noble et al., 2005). It had been suggested that phosphorylation at Ser2 can enhance binding of Pcf11 (Barillà et al., 2001; Sadowski et al., 2003), but the phosphate group of Ser2 does not make direct contacts with Pcf11 and therefore unlikely contributes to the interaction (Figure 4C). Instead, it forms hydrogen bonds within the CTD, in a way that indirectly stabilizes the β-turn structure (Meinhart and Cramer, 2004). By iterating the observed CTD repeat, Meinhart et al. deduced a complete CTD model as a compact β-spiral. While Ser2 phosphorylation can be readily accommodated in the model, Ser5 phosphorylation would open up the spiral and induce a more extended structure. With dynamic phosphorylation and de-phosphorylation, the CTD conformations would be altered and cycled, so that the spatial and temporal control of mRNA processing factor binding during transcription can be achieved.
Figure 5. Other mammalian pre-mRNA 3’ end processing factors. (A) Domain organization of human symplekin, PAP and PABPN1. (B) Crystal structure of the symplekin-Ssu72-CTD peptide complex (PDB ID: 3O2Q (Xiang et al., 2010)). (C) Conformation of the CTD peptide in the active site of Ssu72. (D) Crystal structure of the RRM dimer of PABPN1 (PDB ID: 3B4M (Ge et al., 2008)). (E) Crystal structure of the yeast Pap1 in complex with ATP and oligo(A) sequence (PDB ID: 2Q66 (Balbo and Bohm, 2007)).
The second one is the Ssu72-pSer5 CTD complex (Figure 5B). Surprisingly, the CTD captured in the active site of Ssu72 has the peptide bond between pSer5 and Pro6 in the cis configuration (Figure 5C). This contrasts to all earlier known CTD structures, which were exclusively in trans (Xiang et al., 2010; Werner-Allen et al., 2011). The substrate-binding pocket of Ssu72 has a confined space so that only the CTD with pSer5-Pro6 in cis configuration can be accommodated. The selectivity of Ssu72 nonetheless severely limits its substrate availability, because less than 20% of the total population of the pSer5-Pro6 peptide bond exists in the cis configuration and the natural cis-trans conversion is rather slow (Werner-Allen et al., 2011; Xiang et al., 2010). Therefore, peptidyl-prolyl isomerases (Pin1 in human and Ess1 in yeast) can promote de-phosphorylation of the CTD by accelerating the cis-trans conversion, which presents a higher level regulation of the CTD function (Krishnamurthy et al., 2009; Singh et al., 2009; Xiang et al., 2010; Werner-Allen et al., 2011).

**Poly(A) polymerase (PAP)**

PAP catalyzes the second step of pre-mRNA 3’ end processing: polyadenylation. In the presence of Mn$^{2+}$ in vitro, PAP is capable of nonspecifically polyadenylating RNA primers on its own but it exhibits distributive feature (Raabe et al., 1991; Wahle, 1991b). Under physiological condition where Mg$^{2+}$ is used, polyadenylation by PAP occurs in a PAS-dependent manner. CPSF is required to confer processivity and specificity (Takagaki et al., 1988). This specific polyadenylation can be further boosted by PABPN1 (discussed below). In mammals, PAP is also in most cases indispensable for the cleavage reaction (Christofori and Keller, 1988; Takagaki et al., 1988).
PAP belongs to the DNA polymerase β family (Edmonds, 1990). The first 500 residues of PAP are conserved throughout eukaryotic species (Martin and Keller, 1996; Martin et al., 2000). Crystal structures of bovine PAP and yeast PAP (Pap1) have been determined, revealing a three-globular-domain organization (Bard et al., 2000; Martin et al., 2000) (Figure 5E). A large open central cleft harboring the active site is encompassed by the three domains. Upon substrate binding, the cleft closes as the N- and C-terminal domains interact (Balbo and Bohm, 2007), suggesting an induced-fit mechanism for the enzyme, which is consistent with earlier experiments (Balbo et al., 2005, 2007; Martin et al., 2004). Vertebrate PAP has a C-terminal extension of ~20 kD that does not exist in lower eukaryotes and is not essential for polyadenylation activity (Martin and Keller, 1996) (Figure 5A). This extension is enriched with serines and threonines, which are targets for modulating PAP activity and are subject to various post-translational modifications, including phosphorylation (Colgan et al., 1996), acetylation (Shimazu et al., 2007), sumoylation (Vethantham et al., 2008) and PARylation (Di Giammartino et al., 2013).

PAP has been shown to interact with many protein factors in the pre-mRNA 3’ end processing complex, such as CPSF-160 (Murthy and Manley, 1995) and CFI25 (Dettwiler et al., 2004). In yeast, Pap1 binds to both the N-terminal domain of Pta1 and Yhh1 through its own N-terminal domain (Ezeokonkwo et al., 2012). The C-terminal domain of Pap1 interacts with Fip1 (Helmling et al., 2001; Meinke et al., 2008; Preker et al., 1995), with a complex structure reported (discussed above).
Figure 6. Mammalian pre-mRNA 3’ end processing machinery. (A) A wire map of the interaction network of mammalian pre-mRNA 3’ end processing machinery, also a diagram summary of Table 1. Thick double lines represent interactions observed in both mammalian and yeast systems. Solid lines represent interactions studied only in mammals while dashed lines represent interactions studied only in yeast. (B) A visionary model of mammalian pre-mRNA 3’ end processing machinery. Cis-elements are highlighted in boxes on the pre-mRNA. Red arrow indicates the cleavage site.
Poly(A) binding protein (PABPN1)

The first poly(A) binding protein was identified in the cytoplasm (Blobel, 1973). Cytoplasmic poly(A) binding proteins (PABPC) exist in all eukaryotes (Adam et al., 1986) and play essential roles in translation control and mRNA decay (Kühn and Wahle, 2004; Mangus et al., 2003). The identity of the nuclear poly(A) binding protein (PABPN1) involved in pre-mRNA 3’ end processing was not unearthed until almost two decades later (Nemeth et al., 1995; Wahle, 1991a). PABPN1 serves as another specificity factor in addition to CPSF for PAP in PAS-dependent polyadenylation (Wahle, 1991a). The presence of either CPSF or PABPN1 only provides moderate processivity, but together they promote rapid poly(A) elongation to approximately 200-300nt in length (Bienroth et al., 1993; Wahle, 1995), which matches average newly synthesized poly(A) tails in vivo (Brawerman, 1981). PABPN1 not only ensures the proper length of the poly (A) tail, but it may also be a key regulator in APA (Jenal et al., 2012).

Mammalian PABPN1 has a molecular weight of ~33 kD and a domain architecture very different from PABPC (Figure 5A). The N-terminal domain is acidic and rich in glutamates and may function to prevent undesirable contacts between PAP and PABPN1, as removal of this domain can enhance their interactions (Kerwitz et al., 2003). The middle region contains a coiled coil (CC) domain that is required for stimulation of PAP (Kerwitz et al., 2003). Immediately following the CC is a canonical RRM, whose structure indicates that it forms a dimer (Ge et al., 2008) (Figure 5D). This is compatible with the earlier observation that the C-terminal domain (arginine-rich) of PABPN1 can also self-associate. In fact, in the absence of poly(A) and at elevated concentrations, PABPN1 is prone to aggregation into oligomers (Keller et al., 2000; Nemeth et al., 1995).
Each PABPN1 can recognize a ~10nt poly(A) sequence specifically (Nemeth et al., 1995). Both the RRM motif and the C-terminal domain are required for RNA binding (Kühn et al., 2003). As polyadenylation proceeds, PABPN1 coats the poly(A) tail sequentially, forming a linear filament or a spherical particle up to 21nm in diameter (Keller et al., 2000). This structure is thought to restrict CPSF to the PAS while facilitating the physical interaction between CPSF and PAP over the newly synthesized poly(A) sequence until a desired length is reached (Kühn et al., 2009).

In yeast, PABPN1 has a sequence homolog Rbp29, but this protein localizes in the cytoplasm and has a very different function (Winstall et al., 2000). The true functional counterpart of PABPN1 is still under debate. Two candidates are Pab1 and Nab2. Pab1 associates with CFIA and can regulate poly(A) tail length either on its own or through the poly(A) nuclease PAN-dependent pathway (Amrani et al., 1997b; Brown and Sachs, 1998; Minvielle-Sebastia et al., 1997). However, Pab1 resembles PABPC and is predominantly cytoplasmic, arguing it is equivalent to PABPN1. The nuclear protein Nab2 is involved in mRNA export and has also been shown to control poly(A) synthesis in vitro and in vivo (Dheur et al., 2005; Hector et al., 2002; Viphakone et al., 2008). In contrast to PABPN1, Nab2 does not have an RRM and instead binds to the RNA though its zinc finger (Anderson et al., 1993). Nevertheless, both Pab1 and Nab2 negatively modulate poly(A) length, which is opposite to the stimulatory effect of PABPN1.

**Perspective**

Since the start of the second millennium when the first protein structure of a component of the pre-mRNA 3’ end processing complex was reported (PAP; Bard et al., 2000; Martin et al.,
structural studies in this field have flourished and advanced our knowledge of this intricate process at an ever-increasing pace. They have been proven to be an invaluable source of knowledge, when combined with hypothesis-driven biochemical and functional characterizations, to help us understand the mechanistic details of the machinery. A handful of structures of protein factors have been determined and analyzed (Table 2), but compared to the complicated machinery, what we know is only the tip of the iceberg. Many key proteins have not yet been structurally characterized. For example, how CPSF-160 recognizes the PAS and at the same time orchestrates interactions with other protein factors is still elusive. It will be of great interest to obtain its structure and dissect its binding specificity. More importantly, past studies have focused on individual proteins or even domains rather than looking at the bigger picture. To understand in detail how the polyadenylation complex is assembled and functions as a whole, we will need a more complete structural blueprint (Figure 6). In recent years, several complex structures have been determined, such as CFI complex (Li et al., 2011; Yang et al., 2011a), symplekin-Ssu72-CTD complex (Xiang et al., 2010) and Rna14-Rna15 complex (Moreno-Morcillo et al., 2011a; Paulson and Tong, 2012), which substantially facilitated the molecular mapping of protein interconnections. Nevertheless, structural information on protein complexes in the 3’ end processing machinery is still limited and confined within sub-complexes. How different sub-complexes associate and coordinate in the recruitment process and also the enzymatic reactions is largely unclear. On the other hand, an EM study examining the whole 3’ end processing complex has provided us an unparalleled first look at the overall architecture and may provide a powerful future approach (Shi et al., 2009). Finally, the coupling of 3’ end processing to other nuclear events inevitably raises the complexity, but also opens up an interesting and emerging direction in which analyses can be performed on bridging factors from
a structural perspective, so as to visualize how coupling is achieved and also how cleavage and polyadenylation can affect or be affected by other processes.
Table 1. Protein interactions studied in both mammalian and yeast pre-mRNA 3’ end processing machinery

<table>
<thead>
<tr>
<th>Protein</th>
<th>partner</th>
<th>source</th>
<th>method</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPSF-73 (C-terminus)</td>
<td>CPSF-100</td>
<td>H. sapiens / M. musculus</td>
<td>Yeast two hybrid (weak)</td>
<td>(Dominski et al., 2005a)</td>
</tr>
<tr>
<td>Ysh1</td>
<td>Ydh1</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Kyburz et al., 2003)</td>
</tr>
<tr>
<td>Ysh1</td>
<td>Yth1 (ZF4 &amp; N-terminus)</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Tacahashi et al., 2003)</td>
</tr>
<tr>
<td>Ysh1</td>
<td>Pfs2</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Ohnacker et al., 2000)</td>
</tr>
<tr>
<td>Ysh1</td>
<td>Pcf11</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Ghazy et al., 2012; Kyburz et al., 2003)</td>
</tr>
<tr>
<td>Ysh1</td>
<td>Clp1</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Kyburz et al., 2003)</td>
</tr>
<tr>
<td>Ysh1 (C-terminus)</td>
<td>Pta1 (C-terminus)</td>
<td>S. cerevisiae</td>
<td>Yeast two hybrid, GST pull-down</td>
<td>(Ghazy et al., 2009; Zhelkovsky et al., 2006)</td>
</tr>
<tr>
<td>Ydh1</td>
<td>Yhh1</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Kyburz et al., 2003)</td>
</tr>
<tr>
<td>Ydh1</td>
<td>Pcf11</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Kyburz et al., 2003)</td>
</tr>
<tr>
<td>Ydh1</td>
<td>Pta1</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Ghazy et al., 2009; Kyburz et al., 2003)</td>
</tr>
<tr>
<td>Ydh1</td>
<td>RNAP II CTD (phosphorylated or unphosphorylated)</td>
<td>S. cerevisiae</td>
<td>Yeast two hybrid &amp; GST pull-down</td>
<td>(Kyburz et al., 2003)</td>
</tr>
<tr>
<td>CPSF-30</td>
<td>RNAP II body</td>
<td>H. sapiens</td>
<td>Immuno-pull-down</td>
<td>(Nag et al., 2007)</td>
</tr>
<tr>
<td>Protein</td>
<td>Domain</td>
<td>Organism</td>
<td>Interaction Method</td>
<td>Reference(s)</td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>----------</td>
<td>-------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Yth1 (C-terminus)</td>
<td>Fip1 (residue 206-220)</td>
<td>S. cerevisiae</td>
<td>Yeast two hybrid &amp; immune-pull-down</td>
<td>(Barabino et al., 1997, 2000; Helmling et al., 2001)</td>
</tr>
<tr>
<td>Yth1 ZF4</td>
<td>Fip1</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Tacahashi et al., 2003)</td>
</tr>
<tr>
<td>CPSF-160 (residue 400-973)</td>
<td>CstF-77 (HAT-C, residue 242-549)</td>
<td>H. sapiens</td>
<td>Immuno-pull-down, yeast two-hybrid</td>
<td>(Bai et al., 2007; Murthy and Manley, 1995)</td>
</tr>
<tr>
<td>CPSF-160</td>
<td>PAP</td>
<td>H. sapiens</td>
<td>Immuno-pull-down</td>
<td>(Murthy and Manley, 1995)</td>
</tr>
<tr>
<td>CPSF-160</td>
<td>TAF$<em>{II}$20, TAF$</em>{II}$55, TAF$_{II}$100</td>
<td>H. sapiens</td>
<td>His-pull-down</td>
<td>(Dantonel et al., 1997)</td>
</tr>
<tr>
<td>Yhh1</td>
<td>Pcf11</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Kyburz et al., 2003)</td>
</tr>
<tr>
<td>Yhh1</td>
<td>Clp1</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Kyburz et al., 2003)</td>
</tr>
<tr>
<td>Yhh1</td>
<td>Rna14</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Ghazy et al., 2012; Kyburz et al., 2003)</td>
</tr>
<tr>
<td>Yhh1</td>
<td>RNAP II CTD phosphorylated</td>
<td>S. cerevisiae</td>
<td>Yeast two hybrid &amp; GST pull-down</td>
<td>(Dichtl et al., 2002)</td>
</tr>
<tr>
<td>Yhh1</td>
<td>Pap1 N-terminal domain</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Ezeokonkwo et al., 2012)</td>
</tr>
<tr>
<td>Fip1 (N-terminal acidic &amp; conserved domain)</td>
<td>PAP</td>
<td>H. sapiens</td>
<td>GST pull-down</td>
<td>(Kaufmann et al., 2004)</td>
</tr>
<tr>
<td>Fip1 (N-terminal conserved domain)</td>
<td>CPSF-30</td>
<td>H. sapiens</td>
<td>GST pull-down</td>
<td>(Kaufmann et al., 2004)</td>
</tr>
<tr>
<td>Fip1</td>
<td>CPSF-160</td>
<td>H. sapiens</td>
<td>GST pull-down</td>
<td>(Kaufmann et al., 2004)</td>
</tr>
<tr>
<td>Protein</td>
<td>Interaction</td>
<td>Species</td>
<td>Method</td>
<td>Reference</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td>---------</td>
<td>--------</td>
<td>-----------</td>
</tr>
<tr>
<td>Fip1 (N-terminal acidic &amp; conserved domain)</td>
<td>CstF-77</td>
<td>H. sapiens</td>
<td>GST pull-down</td>
<td>Kaufmann et al., 2004</td>
</tr>
<tr>
<td>Fip1 (residue 80-105)</td>
<td>Pap1 C-terminal domain</td>
<td>S. cerevisiae</td>
<td>Yeast two hybrid, immune-pull-down, in vitro reconstitution &amp; complex structure</td>
<td>Helmling et al., 2001; Meinke et al., 2008; Preker et al., 1995</td>
</tr>
<tr>
<td>Fip1</td>
<td>Rna14</td>
<td>S. cerevisiae</td>
<td>Immuno-pull-down</td>
<td>Preker et al., 1995</td>
</tr>
<tr>
<td>Fip1</td>
<td>Pta1</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>Ghazy et al., 2009</td>
</tr>
<tr>
<td>Pfs2</td>
<td>Rna14</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>Ohnacker et al., 2000</td>
</tr>
<tr>
<td>Pfs2</td>
<td>Fip1</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>Ohnacker et al., 2000</td>
</tr>
<tr>
<td>Pfs2</td>
<td>Clp1</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>Ghazy et al., 2012</td>
</tr>
<tr>
<td>CstF-77 (proline-rich domain)</td>
<td>CstF-64 (hinge domain, residue 101-195)</td>
<td>H. sapiens</td>
<td>Far-western, His pull-down</td>
<td>Bai et al., 2007; Hockert et al., 2010; Takagaki and Manley, 2000</td>
</tr>
<tr>
<td>CstF-77</td>
<td>RNAP II CTD</td>
<td>H. sapiens</td>
<td>GST pull-down (weak)</td>
<td>McCracken et al., 1997</td>
</tr>
<tr>
<td>Rna14</td>
<td>RNAP II CTD</td>
<td>S. cerevisiae</td>
<td>Yeast two hybrid, GST pull-down (weak)</td>
<td>Sadowski et al., 2003</td>
</tr>
<tr>
<td>Rna14 (C-terminal proline-rich domain)</td>
<td>Rna15 (hinge domain)</td>
<td>S. cerevisiae /K lactis.</td>
<td>GST &amp; His pull-down, complex structure</td>
<td>(Gross and Moore, 2001a; Legrand et al., 2007; Moreno-Morcillo et al., 2011a; Paulson and Tong, 2012)</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>----------------------</td>
<td>-------------------------</td>
<td>---------------------------------------</td>
<td>-----------------------------------------------------------------</td>
</tr>
<tr>
<td>Rna14</td>
<td>Pcf11</td>
<td>S. cerevisiae</td>
<td>Yeast two hybrid, GST pull-down</td>
<td>(Amrani et al., 1997a; Gross and Moore, 2001a)</td>
</tr>
<tr>
<td>Rna14 (residue 201-240)</td>
<td>Hrp1 (RRM)</td>
<td>S. cerevisiae</td>
<td>GST pull-down, NMR</td>
<td>(Barnwal et al., 2012; Gross and Moore, 2001a)</td>
</tr>
<tr>
<td>CstF-64 (hinge domain)</td>
<td>symplekin (residue 390-500)</td>
<td>H. sapiens</td>
<td>Far-western</td>
<td>(Ruepp et al., 2011a; Takagaki and Manley, 2000)</td>
</tr>
<tr>
<td>Rna15</td>
<td>Pcf11</td>
<td>S. cerevisiae</td>
<td>Yeast two hybrid, GST &amp; MBP pull-down</td>
<td>(Amrani et al., 1997a; Gross and Moore, 2001a; Qu et al., 2007)</td>
</tr>
<tr>
<td>Rna15</td>
<td>Hrp1</td>
<td>S. cerevisiae</td>
<td>In vitro reconstitution &amp; complex structure</td>
<td>(Leeper et al., 2010)</td>
</tr>
<tr>
<td>Rna15</td>
<td>RNAP II CTD</td>
<td>S. cerevisiae</td>
<td>Yeast two hybrid</td>
<td>(Sadowski et al., 2003)</td>
</tr>
<tr>
<td>CstF-50 (WD40 domain)</td>
<td>RNAP II CTD</td>
<td>H. sapiens</td>
<td>GST pull-down</td>
<td>(McCracken et al., 1997)</td>
</tr>
<tr>
<td>CstF-50 (WD40 repeats)</td>
<td>CsF-77 (C-terminal HAT domain)</td>
<td>H. sapiens</td>
<td>Far-western</td>
<td>(Takagaki and Manley, 2000)</td>
</tr>
<tr>
<td>CstF-50 (WD40 repeats)</td>
<td>BARD1</td>
<td>H. sapiens</td>
<td>Yeast two hybrid, 48mmune-pull-down</td>
<td>(Kleiman and Manley, 1999)</td>
</tr>
<tr>
<td>CF125</td>
<td>CF168 (RRM domain)</td>
<td>H. sapiens</td>
<td>GST pull-down, in vitro reconstitution, complex structure</td>
<td>(Dettwiler et al., 2004; Li et al., 2011; Rüegsegger et al., 1998; Yang et al., 2011a)</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------</td>
<td>------------</td>
<td>----------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>CF125</td>
<td>PAP</td>
<td>H. sapiens</td>
<td>GST pull-down</td>
<td>(Dettwiler et al., 2004)</td>
</tr>
<tr>
<td>CF125</td>
<td>PABPN1</td>
<td>H. sapiens</td>
<td>GST pull-down</td>
<td>(Dettwiler et al., 2004)</td>
</tr>
<tr>
<td>CF168 (RS domain)</td>
<td>SR proteins (hTra2β, SRp20, 9G8)</td>
<td>H. sapiens</td>
<td>Yeast two hybrid, co-IP</td>
<td>(Dettwiler et al., 2004)</td>
</tr>
<tr>
<td>Pcf11</td>
<td>RNAP II CTD</td>
<td>S. cerevisiae</td>
<td>Yeast two hybrid, GST pull-down, complex structure</td>
<td>AM (Barillà et al., 2001; Licatalosi et al., 2002; Meinhart and Cramer, 2004; Sadowski et al., 2003)</td>
</tr>
<tr>
<td>Pcf11</td>
<td>Pta1</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Ghazy et al., 2009; Kyburz et al., 2003)</td>
</tr>
<tr>
<td>hClp1</td>
<td>CFI</td>
<td>H. sapiens</td>
<td>GST pull-down</td>
<td>(de Vries et al., 2000)</td>
</tr>
<tr>
<td>hClp1</td>
<td>CPSF-100</td>
<td>H. sapiens</td>
<td>GST pull-down</td>
<td>(de Vries et al., 2000)</td>
</tr>
<tr>
<td>Clp1</td>
<td>Pcf11</td>
<td>S. cerevisiae</td>
<td>GST pull-down, in vitro reconstitution, complex structure</td>
<td>(Gordon et al., 2011; Gross and Moore, 2001a; Noble et al., 2007)</td>
</tr>
<tr>
<td>Clp1</td>
<td>Pta1</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Ghazy et al., 2009, 2012)</td>
</tr>
<tr>
<td>Clp1</td>
<td>Ssu72</td>
<td>S. cerevisiae</td>
<td>GST pull-down</td>
<td>(Ghazy et al., 2012)</td>
</tr>
<tr>
<td>Protein</td>
<td>Interaction Partner</td>
<td>Species</td>
<td>Technique</td>
<td>Reference</td>
</tr>
<tr>
<td>---------</td>
<td>---------------------</td>
<td>--------------------------</td>
<td>-------------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>Symplekin (N-terminal domain, residue 30-340)</td>
<td>Ssu72</td>
<td>H. sapiens, D. melanogaster</td>
<td>In vitro reconstitution, complex structure</td>
<td>(Xiang et al., 2010; Luo et al., 2013)</td>
</tr>
<tr>
<td>Pta1 (N-terminal domain, residue 1-300)</td>
<td>Ssu72</td>
<td>S. cerevisiae</td>
<td>Yeast two hybrid, GST pull-down</td>
<td>(Ghazy et al., 2009; He et al., 2003)</td>
</tr>
<tr>
<td>Pta1</td>
<td>Pap1</td>
<td>S. cerevisiae</td>
<td>Yeast two hybrid, GST pull-down</td>
<td>(Ezeokonkwo et al., 2012; Ghazy et al., 2009)</td>
</tr>
<tr>
<td>Pta1</td>
<td>RNAP II CTD phosphorylated</td>
<td>S. cerevisiae</td>
<td>GST pull-down (from complex mixture)</td>
<td>(Rodriguez et al., 2000)</td>
</tr>
<tr>
<td>PAP</td>
<td>PABPN1 (N-terminal truncation)</td>
<td>B. taurus</td>
<td>GST pull-down</td>
<td>(Kerwitz et al., 2003)</td>
</tr>
</tbody>
</table>
Table 2. Reported structures of protein factors in mammalian or yeast pre-mRNA 3’ end processing complex

<table>
<thead>
<tr>
<th>Protein</th>
<th>Note</th>
<th>Source</th>
<th>PDB</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPSF-73</td>
<td>Residue 1-460</td>
<td>H. sapiens</td>
<td>2I7T, 2I7V</td>
<td>(Mandel et al., 2006a)</td>
</tr>
<tr>
<td>Ydh1</td>
<td>Residue 1-720</td>
<td>S. cerevisiae</td>
<td>2I7X</td>
<td>(Mandel et al., 2006a)</td>
</tr>
<tr>
<td>CPSF-30</td>
<td>Zinc finger 2 &amp; 3, in complex with influenza NS1A</td>
<td>H. sapiens</td>
<td>2RHK</td>
<td>(Das et al., 2008)</td>
</tr>
<tr>
<td>CstF-77</td>
<td>HAT domain</td>
<td>M. musculus</td>
<td>2OND, 20OE</td>
<td>(Bai et al., 2007)</td>
</tr>
<tr>
<td>CstF-77</td>
<td>HAT domain</td>
<td>E. cuniculi</td>
<td>2UY1</td>
<td>(Legrand et al., 2007)</td>
</tr>
<tr>
<td>CstF-64</td>
<td>RRM domain, residue 1-111</td>
<td>H. sapiens</td>
<td>1P1T</td>
<td>(Pérez Cañadillas and Varani, 2003)</td>
</tr>
<tr>
<td>CstF-64</td>
<td>C-terminal domain, residue 531-577</td>
<td>H. sapiens</td>
<td>2J8P</td>
<td>(Qu et al., 2007)</td>
</tr>
<tr>
<td>CstF-50</td>
<td>Residue 1-65</td>
<td>D. melanogaster</td>
<td>2XZ2</td>
<td>(Moreno-Morcillo et al., 2011b)</td>
</tr>
<tr>
<td>Rna14</td>
<td>HAT domain</td>
<td>K. lactis</td>
<td>4E6H, 4E85</td>
<td>(Paulson and Tong, 2012)</td>
</tr>
<tr>
<td>Rna14 &amp; Rna15 complex</td>
<td>C-terminal domain &amp; hinge domain</td>
<td>S. cerevisiae</td>
<td>2L9B</td>
<td>(Moreno-Morcillo et al., 2011a)</td>
</tr>
<tr>
<td>Rna14 &amp; Rna15 complex</td>
<td>Full length &amp; hinge domain</td>
<td>K. lactis</td>
<td>4EBA</td>
<td>(Paulson and Tong, 2012)</td>
</tr>
<tr>
<td>Rna15</td>
<td>RRM domain, residue 16-111</td>
<td>S. cerevisiae</td>
<td>2X1B</td>
<td>(Pancevac et al., 2010)</td>
</tr>
<tr>
<td>Rna15 &amp; RNA complex</td>
<td>RRM domain</td>
<td>S. cerevisiae</td>
<td>2X1A, 2X1F</td>
<td>(Pancevac et al., 2010)</td>
</tr>
<tr>
<td>Rna15 &amp; Hrp1 in complex with a RNA</td>
<td>RRM domain in Rna15, 2 RRM domains in Hrp1</td>
<td>S. cerevisiae</td>
<td>2KM8</td>
<td>(Leeper et al., 2010)</td>
</tr>
<tr>
<td>CFI25</td>
<td>Residue 21-227, apo form or with a sulfate</td>
<td>H. sapiens</td>
<td>2CL3, 2J8Q</td>
<td>(Trésaugues et al., 2008)</td>
</tr>
<tr>
<td>CFI25</td>
<td>Full length, apo form or with Ap4A</td>
<td>H. sapiens</td>
<td>3BAP, 3BHO</td>
<td>(Coseno et al., 2008)</td>
</tr>
<tr>
<td>CFI25 in complex with RNA</td>
<td>Full length, with two RNA sequences</td>
<td>H. sapiens</td>
<td>3MDI, 3MDG</td>
<td>(Yang et al., 2010)</td>
</tr>
<tr>
<td>Complex</td>
<td>Description</td>
<td>Species</td>
<td>Structure IDs</td>
<td>References</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td>---------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>CFI125 &amp; CFI68 in complex with RNA</td>
<td>RRM domain of CFI68, H. sapiens</td>
<td>3Q2S, 3Q2T, 3P5T, 3P6Y</td>
<td>(Li et al., 2011; Yang et al., 2011a)</td>
<td></td>
</tr>
<tr>
<td>Clp1 in complex with Pcf11</td>
<td>Residue 454-563 of Pcf11, ATP bound</td>
<td>S. cerevisiae</td>
<td>2NPI</td>
<td>(Noble et al., 2007)</td>
</tr>
<tr>
<td>Pcf11</td>
<td>CID domain</td>
<td>S. cerevisiae</td>
<td>1S9Z, 2BFO</td>
<td>(Meinhart and Cramer, 2004; Noble et al., 2005)</td>
</tr>
<tr>
<td>Pcf11 in complex with RNAP II CTD</td>
<td>CID domain of Pcf11, residue 1-140; CTD phosphorylated at Ser2</td>
<td>S. cerevisiae</td>
<td>1SZA, 2BFO</td>
<td>(Meinhart and Cramer, 2004)</td>
</tr>
<tr>
<td>Sympolekin</td>
<td>N-terminal HEAT domain, residue 19-271</td>
<td>D. melanogaster</td>
<td>3GS3</td>
<td>(Kennedy et al., 2009)</td>
</tr>
<tr>
<td>Sympolekin</td>
<td>N-terminal HEAT domain, residue 1-395</td>
<td>H. sapiens</td>
<td>3O2T, 3ODR, 3ODS</td>
<td>(Xiang et al., 2010)</td>
</tr>
<tr>
<td>Sympolekin &amp; Ssu72 in complex with CTD</td>
<td>N-terminal HEAT domain of symplekin, CTD phosphorylated at Ser5</td>
<td>H. sapiens</td>
<td>3O2S, 3O2Q</td>
<td>(Xiang et al., 2010)</td>
</tr>
<tr>
<td>Sympolekin &amp; Ssu72 in complex with CTD</td>
<td>N-terminal HEAT domain of symplekin, CTD phosphorylated at Ser5, or both Thr4 and Ser5</td>
<td>D. melanogaster</td>
<td>4IMJ, 4IMI</td>
<td>(Luo et al., 2013)</td>
</tr>
<tr>
<td>Pap1</td>
<td>Complex with 3’-dATP</td>
<td>S. cerevisiae</td>
<td>1FA0, 201P, 2HHP</td>
<td>(Balbo et al., 2007; Bard et al., 2000)</td>
</tr>
<tr>
<td>Pap1 in complex with MgATP and RNA</td>
<td>D154A mutant</td>
<td>S. cerevisiae</td>
<td>2Q66</td>
<td>(Balbo and Bohm, 2007)</td>
</tr>
<tr>
<td>Pap1 in complex with Fip1</td>
<td>Fip1 residue 80-105</td>
<td>S. cerevisiae</td>
<td>3C66</td>
<td>(Meinke et al., 2008)</td>
</tr>
<tr>
<td>Protein</td>
<td>Description</td>
<td>Species</td>
<td>PDBs</td>
<td>Reference</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td>---------</td>
<td>------</td>
<td>-----------</td>
</tr>
<tr>
<td>PAP</td>
<td>Residue 1-513, in complex with 3'-dATP and Mn$^{2+}$ or Mg$^{2+}$</td>
<td>B. taurus</td>
<td>1F5A, 1Q78, 1Q79</td>
<td>(Martin et al., 2000, 2004)</td>
</tr>
<tr>
<td>PABPN1</td>
<td>RRM domain</td>
<td>H. sapiens</td>
<td>3B4D, 3B4M</td>
<td>(Ge et al., 2008)</td>
</tr>
</tbody>
</table>
References


Martin, G., Gruber, A.R., Keller, W., and Zavolan, M. (2012). Genome-wide analysis of pre-
mRNA 3’ end processing reveals a decisive role of human cleavage factor I in the regulation of
3’ UTR length. Cell Reports 1, 753–763.

transitions of the general RNA polymerase II transcription complex. Nat. Struct. Mol. Biol. 17,
1272–1278.


Mayr, C., and Bartel, D.P. (2009). Widespread shortening of 3’UTRs by alternative cleavage and

McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S.D.,
Wickens, M., and Bentley, D.L. (1997). The C-terminal domain of RNA polymerase II couples

YGTGTTYY located downstream from the AATAAA signal is required for efficient formation
of mRNA 3’ termini. Nucleic Acids Res. 13, 1347–1368.

143.


Structure of yeast poly(A) polymerase in complex with a peptide from Fip1, an intrinsically


Millevoi, S., Loulergue, C., Dettwiler, S., Karaa, S.Z., Keller, W., Antoniou, M., and Vagner, S.
(2006). An interaction between U2AF 65 and CF I(m) links the splicing and 3’ end processing
machineries. EMBO J. 25, 4854–4864.

major yeast poly(A)-binding protein is associated with cleavage factor IA and functions in


Chapter II: Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex
Symplekin (Pta1 in yeast) is a scaffold in the large protein complex that is required for 3′-end cleavage and polyadenylation of eukaryotic messenger RNA precursors (pre-mRNAs)\(^1\). It also participates in transcription initiation and termination by RNA polymerase II (Pol II)\(^2\). Symplekin mediates interactions between many different proteins in this machinery\(^3\)–\(^7\), although the molecular basis for its function is not known. Here we report the crystal structure at 2.4 Å resolution of the amino-terminal domain (residues 30–340) of human symplekin in a ternary complex with the Pol II carboxy-terminal domain (CTD) Ser 5 phosphatase Ssu72 (refs 7, 10–17) and a CTD Ser 5 phosphopeptide. The N-terminal domain of symplekin has the ARM or HEAT fold, with seven pairs of anti-parallel α-helices arranged in the shape of an arc. The structure of Ssu72 has some similarity to that of low-molecular-mass phosphotyrosine protein phosphatase\(^8\)–\(^10\), although Ssu72 has a unique active-site landscape as well as extra structural features at the C terminus that are important for interaction with symplekin. Ssu72 is bound to the concave face of symplekin, and engineered mutations in this interface can abolish interactions between the two proteins. The CTD peptide is bound in the active site of Ssu72, with the pSer 5-Pro 6 peptide bond in the cis configuration, which contrasts with all other known CTD peptide conformations\(^11\)–\(^13\). Although the active site of Ssu72 is about 25 Å from the interface with symplekin, we found that the symplekin N-terminal domain stimulates Ssu72 CTD phosphatase activity in vitro. Furthermore, the N-terminal domain of symplekin inhibits polyadenylation in vitro, but only when coupled to transcription. Because catalytically active Ssu72 overcomes this inhibition, our results show a role for mammalian Ssu72 in transcription-coupled pre-mRNA 3′-end processing.

Human symplekin contains 1,274 amino-acid residues (Fig. 1a) and its sequence is well conserved in higher eukaryotes (Supplementary Fig. 1). In comparison, symplekin shares only weak sequence similarity with yeast Pta1 (ref. 1) (Supplementary Fig. 2), and Pta1 lacks the C-terminal 500 residues of symplekin (Fig. 1a). Symplekin and Pta1 do not have any recognizable homology with other proteins. Predictions of secondary structure suggest the presence of an all-helical segment in the N-terminal region of symplekin and Pta1 (Fig. 1a and Supplementary Figs 1 and 2). Recent studies in yeast suggested that the N-terminal segment of Pta1 is important for interaction with Ssu72 (ref. 9). Ssu72 is required for pre-mRNA 3′-end cleavage in yeast\(^9\), although its phosphatase activity is not necessary for this function\(^9\). The catalytic activity of Ssu72 may instead be important for Pol II transcription and termination and for gene looping\(^17\). Ssu72 is highly conserved in the eukaryotes (Supplementary Fig. 3), but so far no evidence exists implicating mammalian Ssu72 in 3′-end processing.

To determine the structure of a symplekin–Ssu72–CTD phosphopeptide ternary complex, residues 30–360 of human symplekin and full-length human Ssu72 were overexpressed and purified separately. The two proteins were mixed, with Ssu72 in slight molar excess, and the symplekin–Ssu72 complex was purified by gel-filtration chromatography. This procedure also demonstrated strong interactions between the two proteins, consistent with observations on their yeast counterparts\(^7\). The decamer CTD phosphopeptide used in this study, Ser-Tyr 1-Ser 2-Pro 3-Thr 4-pSer 5-Pro 6-Ser 7-Tyr-Ser 8, where Ser 5 is phosphorylated, contained an entire CTD heptad repeat as well as a serine residue from the previous repeat and Tyr-Ser from the following repeat. To prevent hydrolysis, the active-site nucleophile Cys 12 of Ssu72 was mutated to Ser in the ternary complex. We have also determined the crystal structure of the symplekin–Ssu72 (wild-type) binary complex and the structures of the symplekin N-terminal domain alone (for residues 30–395 or 1–395). All the structures are in excellent agreement with the crystallographic data and the expected geometric parameters (Supplementary Table 1).

Figure 1 | Structure of the human symplekin–Ssu72–CTD phosphopeptide ternary complex. a, Domain organization of human symplekin and Saccharomyces cerevisiae Pta1. The domains are suggested by secondary-structure predictions, and the middle region of Pta1 is suggested by functional studies\(^7\). b, Schematic drawing of the structure of human symplekin–Ssu72–CTD phosphopeptide ternary complex, in two views. The N-terminal domain of symplekin is shown in cyan, and Ssu72 in yellow. The CTD phosphopeptide is shown as a stick model (in green for carbon atoms). c, Overlay of the structures of the N-terminal domain of human (in cyan) and Drosophila (in grey) symplekin\(^22\). Drosophila symplekin lacks the last two pairs of helical repeats (boxed). All structure figures were produced with PyMOL (http://www.pymol.org).
The structures show that residues 30–340 of symplekin (Symp-N) form seven pairs of antiparallel β-helices, whereas residues 1–29 and 341–395 are disordered (Supplementary Fig. 1). The pairs of helices are arranged in an arc, with the first helix in each pair, the αA helix, being located on the convex face of the arc, and the αB helix on the concave face (Fig. 1b). Most of the loops connecting the helices are short, except for that linking helices αB and αA, with 31 residues (Fig. 1b). The overall fold of Symp-N is found in many other proteins, including those with the ARM or HEAT repeats. A small, two-stranded antiparallel β-sheet (β2A and β2B) is located near the active site (Fig. 2a). The αD helix is in a different conformation in Ssu72 and also contributes to phosphopeptide binding. Finally, Ssu72 contains an extra helix (αG) and a β-strand (β5S) at the C terminus, which are essential for interactions with symplekin (Fig. 1b).

Our structure of Ssu72 contains a central five-stranded β-sheet (β1–β5) that is surrounded by helices on both sides (Fig. 2a and Supplementary Fig. 4). The closest structural homologue of Ssu72 is the low-molecular-mass phosphotyrosine protein phosphatase (Fig. 2b), as suggested previously18,19, even though the two proteins share only 16% sequence identity. However, our studies show that Ssu72 possesses three unique structural features compared with this other phosphatase (Fig. 2b), which are formed by highly conserved residues (Supplementary Fig. 3) and have important functions. A small, two-stranded antiparallel β-sheet (β2A and β2B) is located near the active site (Fig. 2a). The αD helix is in a different conformation in Ssu72 and also contributes to phosphopeptide binding. Finally, Ssu72 contains an extra helix (αG) and a β-strand (β5S) at the C terminus, which are essential for interactions with symplekin (Fig. 1b).

Resolution for the CTD phosphopeptide, contoured at 3σ, shows detailed interactions between the CTD phosphopeptide and the active site of Ssu72. Ion-pair and hydrogen-bonding interactions are indicated by red dashed lines. Molecular surface of the active-site region of Ssu72. The CTD phosphopeptide is shown in a stick model.

Figure 2 | Recognition of the CTD phosphopeptide by human Ssu72. a, Schematic drawing of the structure of human Ssu72–CTD phosphopeptide complex. b, Overlay of the structures of Ssu72 (yellow) and low-molecular-mass phosphotyrosine protein phosphatase (grey)18,19. Arrows indicate unique structural features in Ssu72. Stereo versions of a and b are given in Supplementary Fig. 4. c, Two views of the omit Fc – Fd electron density at 2.4 Å.

©2010 Macmillan Publishers Limited. All rights reserved
with the pSer-Pro peptide bond in the cis configuration, in contrast to all other known CTD phosphatases (Supplementary Information).

Our observation of a cis configuration for the CTD also provides a different interpretation for the role of the peptidyl-prolyl isomerase Pin1 (Ess1 in yeast) in regulating Pol II transcription. Previous reports have interpreted the opposite effect of Pin1/Ess1 in yeast Ssu72 (refs 24, 25) as evidence for the existence of a trans configuration of the pSer-Pro peptidyl bond in the phosphopeptide of the CTD (Supplementary Information). However, our crystal structure indicates that the opposite must be true. Our in vitro phosphatase assays show that Pin1 strongly stimulates the phosphatase activity of Ssu72 (Supplementary Information and Supplementary Fig. 5), consistent with its specificity for the cis configuration.

The active site of Ssu72 is located at the bottom of a narrow groove (Fig. 2e), one wall of which is formed by the small b-sheet (b2A and b2B) and the loop linking the two strands (Fig. 2d). This severely limits the possible conformation of the CTD, ensuring that only the cis configuration of the pSer-Pro peptide bond can be accommodated in the active site. In fact, the Thr4-pSer5 peptide bond is π-stacked with the Pro6-Ser7 peptide bond (Fig. 2d), suggesting a highly restrained conformation for the CTD phosphopeptide in this region. Residues Thr4, pSer5 and Pro6 of the same repeat as well as Tyr1 of the following repeat have interactions with the enzyme (Fig. 2d and Supplementary Information), explaining the preference for pSer5 by Ssu72 and consistent with results from earlier biochemical studies on yeast Ssu72 (ref. 14).

The phosphate group of the peptide is bound deepest in the structure, having extensive ion-pair and hydrogen-bonding interactions with the active site region of Ssu72, especially for the phosphate atom of Ser5 (Fig. 2b, Supplementary Information). The catalytic nucleophile of Ssu72, Ser5, is hydrogen-bonded to the main-chain carbonyl of Lys43 (in b2A). The catalytic nucleophile of Ssu72, Cys12, is located directly below the phosphate group and can be in the correct position for the in-line nucelophilic attack on the phosphorus atom to initiate the reaction (Supplementary Fig. 7). The side chain of Asp143 is located 3.5 Å from the Oγ atom of Ser5, consistent with its role as the general acid to protonate the leaving group. There are some conformational changes in the active-site region of Ssu72, especially for the b2A–b2B loop, on binding of the CTD phosphopeptide (Supplementary Fig. 7), although this loop seems to be flexible and can assume different conformations in the various structures.

In the structures of the binary and ternary complexes, Ssu72 is bound to the concave face of Symp-N (Fig. 1b). About 950 Å² of the surface area of each protein is buried in the interface of this complex, which involves helices α3B–α6B of Symp-N (Fig. 3a and Supplementary Fig. 5), and helix α9E, the following αE–β4 loop, helix αG and strand β5 of Ssu72 (Supplementary Fig. 3). In addition, residue Arg206, at the tip of the long loop connecting helices α4B and α5A of Symp-N, is also located in the interface (Fig. 3a). Ion-pair, hydrogen-bonding and hydrophobic interactions make contributions to the formation of this complex (Supplementary Information). In particular, the side chains of Val191 and Phe193 of Ssu72 (in strand β5) establish hydrophobic interactions with those of Lys185 (α4B) and Ile251 (α5B) of symplekin in the centre of this interface (Fig. 3a). In addition, the side-chain hydroxyl group of Thr190 (β5) of Ssu72 is hydrogen-bonded to the side chain of Asn300 (α6B) of symplekin. The relative positions of Symp-N and Ssu72 seem to be somewhat variable among the different structures (Fig. 3a).

The catalytic nucleophile of Ssu72, Ser5, is hydrogen-bonded to the main-chain carbonyl of Lys43 (in b2A). The catalytic nucleophile of Ssu72, Cys12, is located directly below the phosphate group and can be in the correct position for the in-line nucelophilic attack on the phosphorus atom to initiate the reaction (Supplementary Fig. 7). The side chain of Asp143 is located 3.5 Å from the Oγ atom of Ser5, consistent with its role as the general acid to protonate the leaving group. There are some conformational changes in the active-site region of Ssu72, especially for the b2A–b2B loop, on binding of the CTD phosphopeptide (Supplementary Fig. 7), although this loop seems to be flexible and can assume different conformations in the various structures.

In the structures of the binary and ternary complexes, Ssu72 is bound to the concave face of Symp-N (Fig. 1b). About 950 Å² of the surface area of each protein is buried in the interface of this complex, which involves helices α3B–α6B of Symp-N (Fig. 3a and Supplementary Fig. 5), and helix α9E, the following αE–β4 loop, helix αG and strand β5 of Ssu72 (Supplementary Fig. 3). In addition, residue Arg206, at the tip of the long loop connecting helices α4B and α5A of Symp-N, is also located in the interface (Fig. 3a). Ion-pair, hydrogen-bonding and hydrophobic interactions make contributions to the formation of this complex (Supplementary Information). In particular, the side chains of Val191 and Phe193 of Ssu72 (in strand β5) establish hydrophobic interactions with those of Lys185 (α4B) and Ile251 (α5B) of symplekin in the centre of this interface (Fig. 3a). In addition, the side-chain hydroxyl group of Thr190 (β5) of Ssu72 is hydrogen-bonded to the side chain of Asn300 (α6B) of symplekin. The relative positions of Symp-N and Ssu72 seem to be somewhat variable among the different structures (Fig. 3a).

The catalytic nucleophile of Ssu72, Ser5, is hydrogen-bonded to the main-chain carbonyl of Lys43 (in b2A). The catalytic nucleophile of Ssu72, Cys12, is located directly below the phosphate group and can be in the correct position for the in-line nucelophilic attack on the phosphorus atom to initiate the reaction (Supplementary Fig. 7). The side chain of Asp143 is located 3.5 Å from the Oγ atom of Ser5, consistent with its role as the general acid to protonate the leaving group. There are some conformational changes in the active-site region of Ssu72, especially for the b2A–b2B loop, on binding of the CTD phosphopeptide (Supplementary Fig. 7), although this loop seems to be flexible and can assume different conformations in the various structures.

In the structures of the binary and ternary complexes, Ssu72 is bound to the concave face of Symp-N (Fig. 1b). About 950 Å² of the surface area of each protein is buried in the interface of this complex, which involves helices α3B–α6B of Symp-N (Fig. 3a and Supplementary Fig. 5), and helix α9E, the following αE–β4 loop, helix αG and strand β5 of Ssu72 (Supplementary Fig. 3). In addition, residue Arg206, at the tip of the long loop connecting helices α4B and α5A of Symp-N, is also located in the interface (Fig. 3a). Ion-pair, hydrogen-bonding and hydrophobic interactions make contributions to the formation of this complex (Supplementary Information). In particular, the side chains of Val191 and Phe193 of Ssu72 (in strand β5) establish hydrophobic interactions with those of Lys185 (α4B) and Ile251 (α5B) of symplekin in the centre of this interface (Fig. 3a). In addition, the side-chain hydroxyl group of Thr190 (β5) of Ssu72 is hydrogen-bonded to the side chain of Asn300 (α6B) of symplekin. The relative positions of Symp-N and Ssu72 seem to be somewhat variable among the different structures (Fig. 3a).
the binary and ternary complexes (Supplementary Information and Supplementary Fig. 8). The symplekin–Ssu72 interaction is located about 25 Å from the active site of Ssu72 (Fig. 1b). However, phosphatase assays measuring the hydrolysis of a p-nitrophenyl phosphate (pNPP) model substrate\(^\text{11,12}\) showed that Symp-N stimulated Ssu72 activity (Fig. 3b), and maximal activation was achieved when the two proteins were at a 1:1 molar ratio. To assess whether this stimulation also occurs with a natural substrate, we first used the decamer CTD phosphopeptide in the assay, monitoring the release of inorganic phosphate, and observed a similar stimulation (Supplementary Fig. 6). We next prepared a glutathione S-transferase (GST)–CTD fusion protein that had been phosphorylated on Ser2 and Ser5 with HeLa nuclear extract\(^\text{13}\). As demonstrated by western blotting with a pSer 5-specific antibody, Ssu72 dephosphorylated this protein on Ser5, in a manner that was also stimulated by Symp-N (Fig. 3c). Ssu72 was specific for dephosphorylating pSer 5, because Ser2 phosphorylation, as monitored by a pSer 2-specific antibody, was not affected (data not shown).

Our data indicate that the symplekin–Ssu72 interaction activated Ssu72 phosphatase, probably through a stabilization of the Ssu72 structure and/or an allosteric mechanism. This is consistent with previous studies on the R129A mutant (Ssu72-2) of yeast Ssu72, equivalent to Arg 126 in human Ssu72 (Supplementary Fig. 3). This mutant shows a twofold decrease in catalytic activity compared with wild-type Ssu72 and produces a severe growth defect at the non-permissive temperature\(^\text{14}\). Arg 126 is far from the active site and is in fact near the interface with symplekin (Fig. 3a). However, it does not contribute directly to interactions with symplekin, and the R126A mutation did not disrupt interaction with Symp-N (data not shown).

To assess the importance of individual residues for the stability of the symplekin–Ssu72 complex, we introduced mutations in the interface and characterized their effects on the complex by using gel-filtration chromatography and phosphatase assays. The presence of wild-type Ssu72 gave rise to a clear shift in the peak for Symp-N from a gel-filtration column (Fig. 3d), corresponding to the formation of the symplekin–Ssu72 complex. Ssu72 was present in twofold molar excess in this experiment, and only half of this protein was incorporated into the complex (Fig. 3d), demonstrating a 1:1 stoichiometry for the complex. Mutation of a symplekin residue in the interface, K185A (Fig. 3a), essentially abolished the interaction with wild-type Ssu72 (Fig. 3e), and mutation of three Ssu72 residues in the interface, T190A/V191A/F193A, abolished the interaction with wild-type symplekin. The chromatographic behaviour of the mutants alone was similar to that of the wild-type protein (Fig. 3e), suggesting that the mutations did not disrupt the structure of the proteins. This was also confirmed by the crystal structure of the K185A mutant (data not shown). Consistent with the gel-filtration data, the symplekin K185A mutant failed to stimulate Ssu72 phosphatase activity, and the T190A/V191A/F193A mutant of Ssu72 could not be stimulated by wild-type Symp-N (Fig. 3b).

We next wished to assess the functional importance of the symplekin–Ssu72 interaction with respect to 3’-end formation. Given the roles of their yeast counterparts in both transcription and polyadenylation, we used a transcription-coupled 3’-end processing assay\(^\text{15}\). HeLa nuclear extract was preincubated with increasing concentrations of Symp-N, which led to a pronounced inhibition of polyadenylation (Fig. 4a), similar to the effect observed earlier with the yeast Pta1 N-terminal domain in a transcription-independent assay\(^\text{16}\). Transcription, as measured by the accumulation of unprocessed pre-mRNA, was not affected (Fig. 4b). RNase protection assays showed that 3’-end cleavage was also not affected (Supplementary Fig. 9), indicating that Symp-N affects only the polyadenylation step 3’-end formation. Inclusion of purified Ssu72 during the preincubation with Symp-N blocked the inhibition, whereas Ssu72 alone had no effect (Fig. 4c). The K185A mutation in Symp-N abolished this inhibitory effect, whereas the T190A/V191A/F193A mutant of Ssu72 failed to overcome the inhibition by wild-type Symp-N (Fig. 4c). These results provide strong evidence that the inhibitory effect of Symp-N reflects its interaction with Ssu72, and thus implicates Ssu72 in mammalian 3’-end processing. In contrast with studies in yeast\(^\text{7,27}\), the catalytically inactive C125 mutant of Ssu72 failed to overcome this inhibition (Fig. 4c), and Symp-N had no detectable effect on transcription-independent polyadenylation (Fig. 4d, and data not shown). Together, these results indicate that Ssu72 phosphatase activity is required for polyadenylation of pre-mRNAs, but only when processing is coupled to transcription.

Our finding that a CTD phosphopeptide is bound to Ssu72 with the pSer-Pro peptide bond in the cis configuration indicates the existence of a novel CTD conformation. Although Ssu72 has been well studied in yeast and has functions in transcription and 3’-end processing, essentially nothing was known about its mammalian counterpart. In fact, whereas the yeast enzyme is a stable component of the polyadenylation machinery and is required for processing, mammalian Ssu72 has not been found associated with polyadenylation factors and was not detected in a recent proteomic analysis of the assembled polyadenylation complex\(^\text{78}\). Consistent with this, our results provide evidence that in mammals Ssu72 is only necessary for polyadenylation when processing is coupled to transcription.
transcription. A parsimonious model is that symplekin recruits Ssu72 to the transcription complex and activates its phosphatase activity, which promotes polyadenylation. Conceivably, this occurs by facilitating the recruitment of poly(A) polymerase, known for many years to be only weakly associated with other 3'-end processing factors9-13, to the complex. Given that the CTD is necessary for efficient 3'-end formation in mammalian cells14, and that CTD pSer5 is the only known target of Ssu72, CTD pSer5 dephosphorylation may well be important in facilitating polyadenylation during transcription.

METHODS SUMMARY
Crystallography. The N-terminal domain of human symplekin and full-length human Ssu72 were overexpressed separately in Escherichia coli and purified. The symplekin–Ssu72 complex was purified by gel filtration of a mixture of the two proteins. Crystals were obtained by the sitting-drop vapour-diffusion method, and the structures were determined by the selenomethionyl single-wavelength anomalous diffraction method and the molecular replacement method.

In vitro transcription-coupled polyadenylation assay. Transcription-coupled polyadenylation was performed with a DNA construct containing GAL4-binding sites upstream of the adenovirus E4 core promoter and SV40 late poly(A) site downstream. Recombinant symplekin and Ssu72 proteins were preincubated with HeLa nuclear extract before transcription was started by adding the DNA template and purified GAL4–VP16. RNA products were purified, separated into nonpolyadenylated and polyadenylated fractions and analysed on 5% denaturing template and purified GAL4–VP16. RNA products were purified, separated into nonpolyadenylated and polyadenylated fractions and analysed on 5% denaturing gel. Radioactivity was detected with a phosphorimager. Assays were repeated multiple times with consistent results.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 16 March; accepted 3 August 2010.
Published online 22 September 2010.

1. Takagaki, Y. & Manley, J. L. Complex protein interactions within the human Ssu72 function but is dispensible for pre-mRNA 3'-end processing to transcription. A parsimonious model is that symplekin recruits Ssu72 to the transcription complex and activates its phosphatase activity, which promotes polyadenylation. Conceivably, this occurs by facilitating the recruitment of poly(A) polymerase, known for many years to be only weakly associated with other 3'-end processing factors9-13, to the complex. Given that the CTD is necessary for efficient 3'-end formation in mammalian cells14, and that CTD pSer5 is the only known target of Ssu72, CTD pSer5 dephosphorylation may well be important in facilitating polyadenylation during transcription.

METHODS SUMMARY
Crystallography. The N-terminal domain of human symplekin and full-length human Ssu72 were overexpressed separately in Escherichia coli and purified. The symplekin–Ssu72 complex was purified by gel filtration of a mixture of the two proteins. Crystals were obtained by the sitting-drop vapour-diffusion method, and the structures were determined by the selenomethionyl single-wavelength anomalous diffraction method and the molecular replacement method.

In vitro transcription-coupled polyadenylation assay. Transcription-coupled polyadenylation was performed with a DNA construct containing GAL4-binding sites upstream of the adenovirus E4 core promoter and SV40 late poly(A) site downstream. Recombinant symplekin and Ssu72 proteins were preincubated with HeLa nuclear extract before transcription was started by adding the DNA template and purified GAL4–VP16. RNA products were purified, separated into nonpolyadenylated and polyadenylated fractions and analysed on 5% denaturing template and purified GAL4–VP16. RNA products were purified, separated into nonpolyadenylated and polyadenylated fractions and analysed on 5% denaturing gel. Radioactivity was detected with a phosphorimager. Assays were repeated multiple times with consistent results.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 16 March; accepted 3 August 2010.
Published online 22 September 2010.

1. Takagaki, Y. & Manley, J. L. Complex protein interactions within the human Ssu72 function but is dispensible for pre-mRNA 3'-end processing to transcription. A parsimonious model is that symplekin recruits Ssu72 to the transcription complex and activates its phosphatase activity, which promotes polyadenylation. Conceivably, this occurs by facilitating the recruitment of poly(A) polymerase, known for many years to be only weakly associated with other 3'-end processing factors9-13, to the complex. Given that the CTD is necessary for efficient 3'-end formation in mammalian cells14, and that CTD pSer5 is the only known target of Ssu72, CTD pSer5 dephosphorylation may well be important in facilitating polyadenylation during transcription.

METHODS SUMMARY
Crystallography. The N-terminal domain of human symplekin and full-length human Ssu72 were overexpressed separately in Escherichia coli and purified. The symplekin–Ssu72 complex was purified by gel filtration of a mixture of the two proteins. Crystals were obtained by the sitting-drop vapour-diffusion method, and the structures were determined by the selenomethionyl single-wavelength anomalous diffraction method and the molecular replacement method.

In vitro transcription-coupled polyadenylation assay. Transcription-coupled polyadenylation was performed with a DNA construct containing GAL4-binding sites upstream of the adenovirus E4 core promoter and SV40 late poly(A) site downstream. Recombinant symplekin and Ssu72 proteins were preincubated with HeLa nuclear extract before transcription was started by adding the DNA template and purified GAL4–VP16. RNA products were purified, separated into nonpolyadenylated and polyadenylated fractions and analysed on 5% denaturing template and purified GAL4–VP16. RNA products were purified, separated into nonpolyadenylated and polyadenylated fractions and analysed on 5% denaturing gel. Radioactivity was detected with a phosphorimager. Assays were repeated multiple times with consistent results.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 16 March; accepted 3 August 2010.
Published online 22 September 2010.

1. Takagaki, Y. & Manley, J. L. Complex protein interactions within the human Ssu72 function but is dispensible for pre-mRNA 3'-end processing to transcription. A parsimonious model is that symplekin recruits Ssu72 to the transcription complex and activates its phosphatase activity, which promotes polyadenylation. Conceivably, this occurs by facilitating the recruitment of poly(A) polymerase, known for many years to be only weakly associated with other 3'-end processing factors9-13, to the complex. Given that the CTD is necessary for efficient 3'-end formation in mammalian cells14, and that CTD pSer5 is the only known target of Ssu72, CTD pSer5 dephosphorylation may well be important in facilitating polyadenylation during transcription.

METHODS SUMMARY
Crystallography. The N-terminal domain of human symplekin and full-length human Ssu72 were overexpressed separately in Escherichia coli and purified. The symplekin–Ssu72 complex was purified by gel filtration of a mixture of the two proteins. Crystals were obtained by the sitting-drop vapour-diffusion method, and the structures were determined by the selenomethionyl single-wavelength anomalous diffraction method and the molecular replacement method.

In vitro transcription-coupled polyadenylation assay. Transcription-coupled polyadenylation was performed with a DNA construct containing GAL4-binding sites upstream of the adenovirus E4 core promoter and SV40 late poly(A) site downstream. Recombinant symplekin and Ssu72 proteins were preincubated with HeLa nuclear extract before transcription was started by adding the DNA template and purified GAL4–VP16. RNA products were purified, separated into nonpolyadenylated and polyadenylated fractions and analysed on 5% denaturing template and purified GAL4–VP16. RNA products were purified, separated into nonpolyadenylated and polyadenylated fractions and analysed on 5% denaturing gel. Radioactivity was detected with a phosphorimager. Assays were repeated multiple times with consistent results.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 16 March; accepted 3 August 2010.
Published online 22 September 2010.

1. Takagaki, Y. & Manley, J. L. Complex protein interactions within the human Ssu72 function but is dispensible for pre-mRNA 3'-end processing to transcription. A parsimonious model is that symplekin recruits Ssu72 to the transcription complex and activates its phosphatase activity, which promotes polyadenylation. Conceivably, this occurs by facilitating the recruitment of poly(A) polymerase, known for many years to be only weakly associated with other 3'-end processing factors9-13, to the complex. Given that the CTD is necessary for efficient 3'-end formation in mammalian cells14, and that CTD pSer5 is the only known target of Ssu72, CTD pSer5 dephosphorylation may well be important in facilitating polyadenylation during transcription.

METHODS SUMMARY
Crystallography. The N-terminal domain of human symplekin and full-length human Ssu72 were overexpressed separately in Escherichia coli and purified. The symplekin–Ssu72 complex was purified by gel filtration of a mixture of the two proteins. Crystals were obtained by the sitting-drop vapour-diffusion method, and the structures were determined by the selenomethionyl single-wavelength anomalous diffraction method and the molecular replacement method.

In vitro transcription-coupled polyadenylation assay. Transcription-coupled polyadenylation was performed with a DNA construct containing GAL4-binding sites upstream of the adenovirus E4 core promoter and SV40 late poly(A) site downstream. Recombinant symplekin and Ssu72 proteins were preincubated with HeLa nuclear extract before transcription was started by adding the DNA template and purified GAL4–VP16. RNA products were purified, separated into nonpolyadenylated and polyadenylated fractions and analysed on 5% denaturing template and purified GAL4–VP16. RNA products were purified, separated into nonpolyadenylated and polyadenylated fractions and analysed on 5% denaturing gel. Radioactivity was detected with a phosphorimager. Assays were repeated multiple times with consistent results.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 16 March; accepted 3 August 2010.
Published online 22 September 2010.
METHODS

Protein expression and purification. Residues 30–360 of human symplekin were subcloned into the pET28a vector (Novagen). The recombinant protein carries a hexahistidine tag at the N terminus. The plasmids were transformed into E. coli BL21(DE3) Star cells. After induction with 0.5 mM isopropyl β-D-thiogalactoside, the cells were allowed to grow at 20°C for 16 h, concentrated by centrifugation, and lysed by sonication. Soluble symplekin was purified by Ni²⁺-nitrilotriacetate (Qiagen) and gel-filtration (Sephacryl S-300; GE Healthcare) chromatography. Purified protein was concentrated to 15 mg ml⁻¹ in a buffer containing 20 mM Tris-HCl pH 7.5, 200 mM NaCl, 5 mM DTT, and 5% (v/v) glycerol, flash-frozen in liquid nitrogen and stored at −80°C. The N-terminal His-tag was not removed for crystallization.

The purified sample was concentrated to 0.1 mg ml⁻¹ in a buffer containing 100 mM Tris-HCl pH 8.5, 200 mM NaCl, 5 mM Dithiothreitol (DTT) and 5% (v/v) glycerol, and stored at −80°C. The selenomethionyl anomalous diffraction method was used to determine the Se sites upstream of an E4 core promoter and an SV40 late poly(A) site downstream. The Se atoms were located with the program BnP,36 regularized multivariate anomalous-diffraction transformation (OMAT),37 and Coot.38

The Se-Sysm→Ser mutation in rat symplekin (residues 32–338) was generated and verified by sequencing. The mutant proteins were expressed and purified by following the same protocol as that for the wild-type protein.

Analytical gel-filtration experiments were carried out on a Superose 12 10/30 column (GE Healthcare), with a buffer containing 20 mM Tris-HCl pH 7.5 and 200 mM NaCl. Symplekin (210 µg) and Ssu72 (380 µg) were mixed and dialyzed to a final volume of 500 µl with the gel-filtration buffer. The mixture was incubated on ice for 1 h before being loaded on the column. The proteins were also run separately on the column to determine their migration behaviour alone.

Ssu72 CTD peptide phosphatase assays. Reaction mixtures (50 µl) containing 50 mM Tris-HCl pH 7.5, 50 mM NaCl, 1.5 mM MgCl₂, 0.5 mM EDTA, 5% (v/v) glycerol and the indicated amount of Ssu72 and GST–Pin1 were incubated at 10°C. Phosphate release was determined with measuring A₂₅₄ and comparing it with a phosphate standard curve.

To study the effect of Pin1 on Ssu72, reaction mixtures (50 µl) containing 50 mM Tris-HCl pH 7.5, 50 mM NaCl, 1.5 mM MgCl₂, 0.5 mM EDTA, 5% (v/v) glycerol and the indicated amount of Ssu72 and GST–Pin1 were incubated at 10°C. Phosphate release was determined with the malachite green reagent (BIOMOL Research Laboratories). Phosphate release was determined by measuring A₅₄₀ and comparing it with a phosphate standard curve.

In vitro transcription-coupled polyadenylation assay. The DNA construct used for the transcription-coupled, polyadenylation assay contained GALA-binding sites upstream of an E4 core promoter and an SV40 late poly(A) site downstream. Transcription-coupled polyadenylation was performed at 30°C for 1 h in 20 µl reaction mixtures containing 10 µl of nuclear extract, 100 ng of Gal4–VP16, recombinant proteins (symplekin and Ssu72), 1 mM HEPES pH 7.9, 0.5 mM CaCl₂, 10 mM MgCl₂, 20 mM creatine phosphate (di-tris), 2.4% (v/v) Hepes pH 7.5, 0.5 mM EDTA, 11.2 mM DTT and 0.33 mM phenylmethylsulphonyl fluoride. Recombinant symplekin and Ssu72 proteins were preincubated with nuclear extract (30 min at 23°C) before transcription was started by adding the DNA templates. The reaction was stopped by adding proteinase K. RNA products were separated into non-polyadenylated and polyadenylated fractions by oligo(dT) selection; thereafter 2% of non-polyadenylated and 100% of polyadenylated fractions were analysed on a 3% denaturing gel. Radioactivity was detected with a Phosphorimage.

RNAase protection assay. To make the probe, pGSVL-A was linearized with Sall and transcribed with T7 RNA polymerase (Promega) for 2 h at 37°C, uniformly labelled with [α-³²P]UTP. The RNA was gel-purified and reconstituted in hybridization buffer containing 40 mM PIPES pH 6.4, 400 mM NaCl, 1 mM EDTA and 80% (v/v) formamide.

Transcription-coupled polyadenylation was performed as mentioned above except that 0.5 µl of each of ATP, GTP, UTP, and CTP, but not GTP-U (0.5 µl), were used in the reaction mixture. The RNA:DNA mixture was resuspended in 20 µl of Turbo DNase buffer (Ambion) with 1 U of Turbo DNase (Ambion) and incubated for 1 h at 37°C. The DNA was removed by the DNA templates. The remaining RNA products were hybridized overnight with the probe (5× 10⁵ c.p.m.) at 42°C. Each reaction was then incubated for 45 min with 14 µg of RNase A and 0.7 µg of RNase T1 at 30°C. The final RNA products were resolved on a 10% denaturing gel. Radioactivity was detected with a Phosphorimage.


Comparison of Ssu72 with other CTD phosphatases

The structure of Ssu72 has a different backbone fold compared to other CTD phosphatases with known structures, such as Scp1 and Fcp1. Scp1 prefers pSer⁵ CTD as the substrate⁴³, while Fcp1 prefers pSer² CTD (ref. 44). The catalysis of both enzymes requires Mg²⁺, in contrast to Ssu72. The active sites of these enzymes have pockets that can accommodate the trans configuration of the CTD. In fact, the structure of Scp1 in complex with a CTD phosphopeptide shows that the cis configuration is incompatible with its active site. Therefore, Ssu72 and Scp1 may have mutually exclusive preferences for their CTD substrate.

Pin1 stimulates phosphatase activity of Ssu72

The phosphatase activity of Ssu72 towards the 10-mer CTD phosphopeptide was monitored through the release of inorganic phosphate. ³¹P NMR experiments suggest that the cis configuration of the peptide is present in roughly 20% abundance, similar to results on other phosphopeptides (~10%)⁴⁵. Pin1 is expected to facilitate the trans-cis isomerization and thereby increases the apparent phosphatase activity of Ssu72 towards this peptide. To demonstrate this simulation, we incubated the phosphopeptide (500 µM concentration) with Ssu72 (200 µM), and determined the effect of 10 µM Pin1 (purified as a GST-fusion protein) on this reaction. The reactions were carried out at 10°C to reduce the spontaneous isomerization between the two configurations. The results of the assays show a 4-fold stimulation of the phosphatase activity of Ssu72 (Supplementary Fig. 6), consistent with its specificity for the cis configuration of the substrate.

Interactions between the CTD phosphopeptide and Ssu72

The phosphate group of the pSer⁵ residue is recognized by bi-dentate ion-pair interactions with the side chain of Arg18 and extensive hydrogen-bonding interactions with the main-chain amides of residues 13-18 (in the β1-α loop) (Fig. 2d). The main-chain amide of pSer⁵ is hydrogen-bonded to the main-chain carbonyl of Lys43 (β2A). The side chain of Pro⁶ is positioned near two hydrophobic residues, Pro45 (β2A-β2B loop) and Met84 (αD helix). One face of the Tyr⁴ residue from the next heptad repeat is positioned against residues Ala48-Pro49 (β2A-β2B loop), while the other face is exposed. The main-chain amide of this Tyr residue may be hydrogen-bonded to the side chain of Thr⁴. In contrast, residues Pro³ and Ser² do not have interactions with the enzyme. In addition, residues of the phosphopeptide that are disordered in the current structure, SY⁴S⁶⁵ and S⁵, probably do not have strong interactions with Ssu72 either.

Interface between symplekin and Ssu72

The interface between symplekin N-terminal domain and Ssu72 involves ion-pair, hydrogen-bonding, and hydrophobic interactions. The ion-pair interactions include those between the side chains of Lys185 (symplekin) and Glu127 (Ssu72), Lys299 and Glu113 and Asp117, Arg206 (in the long loop in symplekin) and Glu129, and Lys127 and Glu169 (Fig. 3a). The hydrogen-bonding interactions include those between the side chains of Asn300 (symplekin) and Thr190 (Ssu72), and Ser292 and Glu180. A large
number of residues make important contributions to the surface area burial in this
interface, and they are highlighted in Supplementary Fig. 1 for symplekin and
Supplementary Fig. 3 for Ssu72. Residues with greater than 50 Å² surface area burial
in symplekin are Leu131, Lys185, Arg206, Ile251, Ser292, Ser296, and Lys299, and those
in Ssu72 are Glu129, Cys131, Glu169, Asn170, Leu188, and Phe193.

In the crystals of the binary and ternary complexes, a second possible interface
exists between symplekin and Ssu72, where Ssu72 is bound to the convex face of the N-
terminal domain (Supplementary Fig. 10). The amount of buried surface area in this
interface is smaller, about 830 Å². Three hydrophobic residues, Leu150 and Phe153 of
Ssu72 and Met116 of symplekin, are located in the center of this interface
(Supplementary Fig. 10). Our mutagenesis studies, on M116A mutant in symplekin and
L150A/F153A mutant in Ssu72, showed however that this interface is not important for
symplekin-Ssu72 interactions. Nonetheless, the structural observations provide additional
evidence that the surface of symplekin may be capable of interacting with many different
protein partners, in agreement with the scaffold function of this protein.

Conformational plasticity in the symplekin-Ssu72 complex

The relative positions of symplekin N-terminal domain and Ssu72 show some
degree of variation among the structures that we have studied. These include the ternary
complex, a binary complex in the same space group (P2₁) as the ternary complex, and a
binary complex in a different space group (P2₁2₁2). There are two copies of the
complexes in the monoclinic space groups, affording a total of five complexes for
comparison (Supplementary Fig. 8). Clear differences in the position of symplekin
relative to Ssu72 are observed in the two ternary complexes. Interestingly, only one of the
two complexes contained the CTD phosphopeptide in the Ssu72 active site, while the
other complex in the same crystal contained only a phosphate group, although neither
active site was directly involved in crystal packing. This suggests that the position of
symplekin N-terminal domain might allosterically affect binding in the active site of
Ssu72, which may be consistent with our observation of the stimulatory effect of
symplekin on Ssu72.

There are no major conformational changes in the N-terminal domain of symplekin
upon the formation of the complex with Ssu72 (Supplementary Fig. 11). The rms
distance for equivalent Cα atoms in the structures of the N-terminal domain alone and in
the complex is 1.0 Å, although the arc for the N-terminal domain has a slightly higher
curvature in the complex with Ssu72. Similarly, no major conformational differences are
observed between the structure of human Ssu72 in the complex with symplekin and that
of Drosophila Ssu72 alone (PDB entry 3FDF, Supplementary Fig. 7). The rms distance
for equivalent Cα atoms of the two structures is 1.0 Å, and the two proteins share 61%
amino acid sequence identity (Supplementary Fig. 3).

Implications for the structure of the Pta1-Ssu72 complex

Residues that make important contributions to the symplekin-Ssu72 interface are
generally conserved or show conservative variations among symplekin (Supplementary
Fig. 1) and Ssu72 (Supplementary Fig. 3) homologs. While Pta1 has much weaker
sequence homology with symplekin, some of the symplekin residues in the interface with
Ssu72 may be conserved in Pta1. For example, three consecutive predicted helices in
Ptα1 (between residues 90 and 180) each contains a highly conserved Lys residue in the middle (Supplementary Fig. 2), one of which may be equivalent to the Lys185 residue in symplekin. At the same time, it could also be possible that the exact binding mode of yeast Ssu72 to Ptα1 may be somewhat different from that observed here for the human proteins due to the large sequence variability between Ptα1 and symplekin.

References:

### Table 1. Data collection and refinement statistics

<table>
<thead>
<tr>
<th></th>
<th>Symplekin (30-395)</th>
<th>Symplekin-Ssu72 complex</th>
<th>Symplekin-Ssu72(C12S)-CTD complex</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data collection</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Space group</td>
<td>$P2_1$</td>
<td>$P2_1;2;2$</td>
<td>$P2_1$</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a$, $b$, $c$ (Å)</td>
<td>41.6, 63.0, 62.3</td>
<td>99.4, 113.6, 59.1</td>
<td>67.0, 97.6, 105.0</td>
</tr>
<tr>
<td>$\alpha$, $\beta$, $\gamma$ (°)</td>
<td>90, 90.6, 90</td>
<td>90, 90, 90</td>
<td>90, 98.7, 90</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>30–1.4 (1.5–1.4)*</td>
<td>30–2.5 (2.6–2.5)</td>
<td>30–2.4 (2.5–2.4)</td>
</tr>
<tr>
<td>$R_{merge}$</td>
<td>2.9 (49.2)</td>
<td>6.7 (30.6)</td>
<td>7.8 (40.5)</td>
</tr>
<tr>
<td>$I/\sigma I$</td>
<td>40.4 (1.7)</td>
<td>15.7 (3.1)</td>
<td>11.8 (2.4)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99 (95)</td>
<td>100 (99)</td>
<td>100 (99)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>3.6 (2.5)</td>
<td>4.1 (3.6)</td>
<td>3.3 (3.3)</td>
</tr>
<tr>
<td><strong>Refinement</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>30–1.4</td>
<td>30–2.5</td>
<td>30–2.4</td>
</tr>
<tr>
<td>No. reflections</td>
<td>59,384</td>
<td>22,566</td>
<td>49,442</td>
</tr>
<tr>
<td>$R_{work}$/$R_{free}$</td>
<td>16.0 / 19.2</td>
<td>21.1 / 26.6</td>
<td>18.4 / 24.1</td>
</tr>
<tr>
<td>No. atoms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>2,409</td>
<td>4,014</td>
<td>8,078</td>
</tr>
<tr>
<td>Ligand/ion</td>
<td>4</td>
<td>5</td>
<td>54</td>
</tr>
<tr>
<td>Water</td>
<td>332</td>
<td>76</td>
<td>250</td>
</tr>
<tr>
<td>B-factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>15</td>
<td>41</td>
<td>36</td>
</tr>
<tr>
<td>Ligand/ion</td>
<td>17</td>
<td>33</td>
<td>44</td>
</tr>
<tr>
<td>Water</td>
<td>35</td>
<td>22</td>
<td>32</td>
</tr>
<tr>
<td>R.m.s deviations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.015</td>
<td>0.017</td>
<td>0.016</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>1.5</td>
<td>1.7</td>
<td>1.6</td>
</tr>
</tbody>
</table>

One crystal was used for the collection of each data set.  
*Highest resolution shell is shown in parenthesis.
Fig. 1. Sequence alignment of the N-terminal domain of symplekin. The secondary structure elements in the structure of human symplekin are indicated (S. S.). Residues in human symplekin that have less than 25% exposed surface area are colored in magenta, and those not observed in our structure are shown in italic. The red dots highlight those residues in the interface with Ssu72 (with more than 15 Å² contribution to the surface area burial).
Fig. 2. Sequence alignment of the putative N-terminal domain of fungal Pta1 homologs. The predicted helices are indicated. The low sequence conservation with symplekin makes it difficult to produce a reliable alignment of their sequences together, and therefore two separate alignments are shown.
Fig. 3. Sequence alignment of eukaryotic Ssu72. The catalytic nucleophile, Cys12, is shown in red, and residues identical among the sequences are highlighted in magenta. The red dots highlight those residues in the interface with symplekin.
Fig. 4. (Top). Structure of the C12S mutant of Ssu72 in complex with a CTD phosphopeptide. (Bottom). Overlay of the structures of wild-type human Ssu72 (in yellow) and the bovine low-molecular-weight phosphotyrosine protein phosphatase (in gray). The active site residues of Ssu72 and a sulfate ion are shown as stick models. Red arrows point to unique structural features of Ssu72. Produced with the program PyMOL (www.pymol.org).
Fig. 5. Overlay of the bound conformations of CTD peptides to various proteins, Ssu72 (green), CTD Ser⁵ phosphatase Scp1 (cyan), WW domain of Pin1 prolyl isomerase (yellow), and RNA guanylyltransferase Cgt1 (pink). The direction of the peptide backbone in the Ssu72 complex is indicated in red, while that of the others is in black, going in the opposite direction.
Fig. 6. (Left) Pin1 stimulates the phosphatase activity of Ssu72. The reaction velocity of Ssu72 towards the 10-mer CTD phosphopeptide was determined in the absence and presence of Pin1, at 10°C to reduce spontaneous cis-trans isomerization. A four-fold stimulation was observed. (Right) Symplekin N-terminal domain stimulates the phosphatase activity of Ssu72 towards the 10-mer CTD peptide. In both assays, the release of inorganic phosphate was monitored, with the BIOMOL reagent.
Fig. 7. (Top). Overlay of the structure of human Ssu72 in the complex with symplekin and CTD phosphopeptide (in yellow) with that of Ssu72 in complex with symplekin (in gray). A change in the side chain rotamer of Cys12 will bring it into the correct location for catalysis, as seen with Ser12. (Bottom). Overlay of the structure of human Ssu72 in the complex with symplekin (in yellow) with that of *Drosophila* Ssu72 alone (PDB entry 3FDF, in gray).
**Fig. 8.** Overlay of the structures of symplekin-Ssu72 complexes. The ternary complex is shown in cyan for symplekin, yellow for Ssu72 and green for the CTD phosphopeptide. The binary complex in the same crystal (with a phosphate in the Ssu72 active site) is shown in pink. The two complexes in the other $P_2_1$ crystal are shown in gray, and the binary complex in the $P_2_12_12$ crystal is shown in orange. The overlay is based on the structure of Ssu72 only.
Fig. 9. Symplekin N-terminal domain does not affect transcription-coupled 3'-end cleavage. RNase protection was carried out after transcription-coupled polyadenylation. Positions of the unprocessed RNA (pre-mRNA, 140nt) and the cleaved product (80nt) are indicated.
Fig. 10. Another possible symplekin-Ssu72 interface in the crystal of the complex. Symplekin is shown in cyan, Ssu72 in yellow. Ssu72 is bound to the convex face of the symplekin structure. The side chains of Leu150 and Phe153 in Ssu72 and Met116 in symplekin are shown as stick models.
Fig. 11. Overlay of the structure of the N-terminal domain of symplekin alone (in gray) and in complex with Ssu72 (in cyan). The panel on the right is related to that on the left by a 90° rotation around the vertical axis.
Chapter III: An unexpected binding mode for a Pol II CTD peptide phosphorylated at Ser7 in the active site of the CTD phosphatase Ssu72
An unexpected binding mode for a Pol II CTD peptide phosphorylated at Ser7 in the active site of the CTD phosphatase Ssu72

Kehui Xiang, James L. Manley, and Liang Tong

Department of Biological Sciences, Columbia University, New York, New York 10027, USA

Ssu72, an RNA polymerase II C-terminal domain (CTD) phospho-Ser5 (pSer5) phosphatase, was recently reported to have pSer7 phosphatase activity as well. We report here the crystal structure of a ternary complex of the N-terminal domain of human symplekin, human Ssu72, and a 10-mer pSer7 CTD peptide. Surprisingly, the peptide is bound in the Ssu72 active site with its backbone running in the opposite direction compared with a pSer5 peptide. The pSer7 phosphatase activity of Ssu72 is ~4000-fold lower than its pSer5 phosphatase activity toward a peptide substrate, consistent with the structural observations.

Supplemental material is available for this article.

Received June 18, 2012, revised version accepted August 28, 2012.

Results and Discussion

To understand the structural basis for how Ssu72 functions as a pSer7 phosphatase, we determined the crystal structure at 2.2 Å resolution of a ternary complex of a human symplekin N-terminal domain (NTD, residues 30–360), human Ssu72 (C12S mutant), and a 10-mer CTD peptide phosphorylated at Ser7 (Ser2–Pro3–Thr4–Ser5–Pro6–Ser7–Tyr1’–Ser2–Pro3’–Thr4’), with the prime indicating the next repeat of the CTD (Fig. 1A). Symplekin is a scaffold protein in the pre-mRNA 3’ end processing machinery (Takagaki and Manley 2000) and is known to interact with and enhance the activity of Ssu72 (Xiang et al. 2010). It was included here because it is important for the crystallization of human Ssu72 (Xiang et al. 2010). Crystals of the symplekin NTD–Ssu72 binary complex were soaked with the phosphopeptide, following the same protocol as used earlier for the pSer5 substrate (Xiang et al. 2010). Electron density was observed for most of the peptide based on the crystallographic analysis, except for one residue at either end of the peptide (Fig. 1B). In addition, the side chain of Tyr1’ has weak electron density.

Surprisingly, the pSer7 peptide is bound in the Ssu72 active site with its peptide backbone running in the opposite direction compared with the pSer5 peptide (Fig. 1C). As a result, the binding sites for Pro6 and pSer7 are equivalent to those for Pro6 and pSer5 in the pSer5 peptide (Xiang et al. 2010). Residues Ser5–Pro6–pSer7–Tyr1’ assume the conformation of a type I reverse turn, so that the overall shapes of the pSer7 and pSer5 peptides are similar when bound to Ssu72 (Fig. 1C). However, such a turn is wider than the cis proline structure, and a difference was observed in the conformation and position of Pro6 in the two peptides. On the other hand, the Pro6–pSer7 peptide bond and, in fact, all of the peptide bonds in the pSer7 peptide are in the trans configuration. The Tyr1’ residue in the pSer5 substrate is important for phosphatase activity [Hausmann et al. 2005] and shows interactions with Ssu72 (Fig. 1D). Equivalent interactions are absent in the pSer7 substrate complex (Fig. 1D).

Recognition of both orientations of peptide substrates or ligands has rarely been observed with other proteins. Reported examples include the SH3 domain (Feng et al. 1994; Lim et al. 1994), stromelysin-1 (Becker et al. 1995), thioredoxin (Qin et al. 1996), and the adaptor protein SbpB of the AAA+ protease ClpXP (Levchenko et al. 2005). As far as we are aware, this is the first time that a protein phosphatase has been shown to bind both orientations of its peptide substrate.
Figure 1. The Pol II CTD pSer7 peptide has a novel binding mode in the active site of Ssu72. (A) Schematic drawing of the structure of the ternary complex of the human symplekin NTD (cyan), human Ssu72 (C12S mutant, yellow), and the pSer7 CTD peptide (green). (B) Omit 2Fo – Fc electron density for the pSer7 peptide at 2.2 Å resolution, contoured at 0.8σ. (C) Overlay of the binding modes of the pSer7 peptide (in green) and the pSer5 peptide (in gray) (Xiang et al. 2010) in the active site of Ssu72. The directions of the peptide backbone are indicated by the arrows. (D) Overlay of the active site region of Ssu72 in the pSer7 complex (yellow) and the pSer5 complex (in gray). (E) Close-up of the active site region showing the structural differences in the pSer residues and Asp143. All of the structure figures were produced with PyMOL (http://www.pymol.org).
The pSer7 residue is anchored in the catalytic site and shows interactions with Ssu72 similar to those of the pSer5 residue (Fig. 1D, Xiang et al. 2010). Most of the Ssu72 residues in the active site have the same conformation in the pSer7 and pSer5 peptide complexes. One notable difference is in the side chain of Asp143, the general acid of the phosphatase reaction. The oxygen atom that should donate a proton to the leaving group (Og of Ser7) is instead hydrogen-bonded to the main chain amide of Ser7 (Fig. 1E). In contrast, such a hydrogen bond is not possible with the pSer5 substrate, as the pSer5–Pro6 peptide bond does not have an amide hydrogen. Moreover, the Cp–Og bond of Ser7 is at a right angle to the Asp143 side chain (Fig. 1E), possibly due to the reversal of the backbone direction, and therefore the Og atom may not be in the optimal conformation for receiving the proton from Asp143. Both of these differences may be detrimental for the phosphatase activity with the pSer7 substrate.

To understand how Ssu72 distinguishes between pSer2, pSer5, and pSer7 in the same CTD repeat, we determined the crystal structure at 2.0 Å resolution of another ternary complex using a 10-mer CTD peptide phosphorylated at all three positions, Ser7–Tyr1–pSer2–Pro3–Thr4–pSer5–Pro6–pSer7–pTyr1. The structure reveals binding of pSer5 in the catalytic site, while there is no evidence for pSer7 binding into this site (Fig. 2A). Instead, the phosphate group on Ser7 is exposed to the solvent, having no interactions with Ssu72 (Fig. 2B). The phosphate group on Ser2 may have ionic interactions with the side chain of Arg114 (Fig. 2B), although the electron density for pSer2 is rather weak (Fig. 2A). Overall, the binding mode of this triply phosphorylated peptide is similar to that of the pSer5 peptide reported earlier (Fig. 2B; Xiang et al. 2010). A conformational difference in the Pro3 residue of the peptide was observed, possibly linked to the interaction of pSer2 with the enzyme. We used another peptide with pSer5 in the center, Ser2–Pro3–Thr4–pSer5–Pro6–pSer7–Tyr1–pSer2–Pro3–Thr4–pSer5, and the observed electron density at 2.6 Å resolution was also consistent with pSer5 binding in the active site (data not shown). These observations indicate that Ssu72 prefers the binding of pSer5 and suggest that Ssu72 has higher affinity for pSer5 in the active site than for pSer7. This may also be detrimental for the phosphatase activity with the pSer7 substrate.

We next characterized the pSer7 phosphatase activity of Ssu72 using in vitro assays. We first used an assay that monitored the release of free phosphate from CTD phosphopeptides, which we previously used successfully with the pSer5 peptide (Xiang et al. 2010). Very low activity was observed in this assay using the 10-mer pSer7 peptide as substrate. Appreciable phosphate production was observed only after overnight incubation of purified His-tagged Ssu72 (Supplemental Fig. 1) with the phosphopeptide (40 μM phosphate was released after 12 h of incubation of 200 μM Ssu72 with 2 mM peptide at room temperature). The low activity did not appear to be affected by the presence of the symplekin NTD (results not shown). Parallel experiments with the 10-mer pSer5 peptide (Xiang et al. 2010) showed that the pSer5 phosphatase activity was ~4000-fold higher than the pSer7 phosphatase activity (160 μM phosphate was released after 45 min of incubation of 5 μM Ssu72 with 1 mM peptide at room temperature) (Fig. 3A, left panel).

A comparison, purified yeast Fcp1 (Supplemental Fig. 1) displayed high activity with both the pSer5 and pSer7 peptides (Fig. 3A, right panel). While the physiological role of Fcp1 in pSer7 dephosphorylation is unclear (e.g., see Bataille et al. 2012), this result at a minimum indicates that the pSer7 peptide used in our assays can be effectively dephosphorylated in vitro, strengthening the conclusion that Ssu72 activity toward this peptide substrate is low.

We next tested whether the pSer7 peptide can compete with the pSer5 peptide for binding to Ssu72 and thereby inhibit its dephosphorylation. We included up to 20-fold higher concentration of the pSer7 peptide (2 mM) than the pSer5 peptide (0.1 mM) in the assay but did not observe any appreciable effect on the dephosphorylation of the pSer5 peptide (results not shown). This supports our structural observation that the affinity of Ssu72 for the pSer5 peptide is substantially higher than that of the pSer7 peptide.

Phosphorylation of CTD Tyr1 in yeast was recently reported to inhibit the recruitment of Pol II termination factors (Mayer et al. 2012). We tested the activity of Ssu72 toward a pSer7, pTyr1 doubly phosphorylated peptide, Ser2–Pro3–Thr4–Ser5–Pro6–pSer7–pTyr1–Ser2–Pro3–Thr4–Ser5, and found that Tyr1 phosphorylation had essentially no effect on pSer7 dephosphorylation even though pTyr1 directly follows pSer7 in the primary sequence. This is
toward pSer7 and pSer5 in the GST-CTD substrate.

stimulates Ssu72 phosphatase activity toward both pSer7 and pSer5. GST-CTD substrate phosphorylated with Cdk7. (Phosphatase activity for human Ssu72 and yeast Fcp1 toward the pSer5 (Fig. 3B), consistent with results reported earlier demonstrating comparable activity toward pSer7 and pSer5 (Bataille et al. 2012; Zhang et al. 2012). Moreover, the symplekin NTD stimulated the pSer7 phosphatase activity of Ssu72 (Fig. 3C), as observed earlier with the pSer5 substrate recognition by Ssu72 and explain how the enzyme interaction for Ssu72 function in yeast (Ghazy et al. 2009) and may help to ensure that Ssu72 activity is constrained until the appropriate time during the transcription cycle. Overall, our results demonstrate a novel mode of substrate recognition by Ssu72 and explain how the enzyme can dephosphorylate both pSer5 and pSer7.

Materials and methods

Protein expression, purification, and crystallization

The experimental protocols followed those described earlier (Xiang et al. 2010). Briefly, human symplekin NTD and human Ssu72, both His-tagged, were expressed in Escherichia coli and purified separately using nickel affinity and gel filtration chromatography. The symplekin NTD–Ssu72 complex was puriﬁed by gel ﬁltration after mixing the two proteins. Crystals of this binary complex were transferred to a soaking solution containing 100 mM Tris (pH 8.5), 10 mM NaCl, 25% (w/v) PEG 3000, and either the pSer7 peptide (10 mM) for 48 h or the pSer2, pSer5, and pSer7 triple phosphorylated peptide (30 mM) for 16 h. The crystals were then transferred to the same soaking solution supplemented with 25% (v/v) ethylene glycol for 30 sec and flash-frozen in liquid nitrogen. His-tagged yeast Fcp1 (168-606) was over-expressed in E. coli BL21 (DE3) Star cells and puriﬁed by nickel afﬁnity and gel ﬁltration chromatography, following a published protocol (Kamenski et al. 2004).

Data collection and structure determination

X-ray diffraction data were collected on an ADSC charge-coupled device at the X29A beamline of National Synchrotron Light Source (NSLS). The diffraction images were processed and scaled with the HKL package (Otwinski and Minor 1997). The structure reﬁnement was carried out with the program Refmac [Murshudov et al. 1997] and manual rebuilding of the model was performed with the program Coot (Emsley and Cowtan 2004). The statistics on the structure reﬁnement are summarized in Table 1.
Table 1. Summary of crystallographic information

<table>
<thead>
<tr>
<th>Structure</th>
<th>pSer7 peptide complex</th>
<th>pSer2, pSer5, pSer7 peptide complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution range</td>
<td>50 Å-2.2 Å (2.28 Å-2.2 Å)</td>
<td>50 Å-2.0 Å (2.07 Å-2.0 Å)</td>
</tr>
<tr>
<td>Number of observations</td>
<td>248,943</td>
<td>320,952</td>
</tr>
<tr>
<td>Redundancy</td>
<td>3.7 [3.7]</td>
<td>3.6 [3.5]</td>
</tr>
<tr>
<td>Rmerge</td>
<td>6.7% (38.0%)</td>
<td>5.3% (44.5%)</td>
</tr>
<tr>
<td>I/σf</td>
<td>19.7 [3.7]</td>
<td>22.5 [2.5]</td>
</tr>
<tr>
<td>Number of reflections</td>
<td>63,783</td>
<td>83,323</td>
</tr>
<tr>
<td>Completeness</td>
<td>99.7% (99.1%)</td>
<td>96.1% (89.0%)</td>
</tr>
<tr>
<td>Free R factor</td>
<td>19.7% (22.7%)</td>
<td>19.9% (22.5%)</td>
</tr>
<tr>
<td>RMSD in bond lengths</td>
<td>0.018 Å</td>
<td>0.016 Å</td>
</tr>
<tr>
<td>RMSD in bond angles</td>
<td>1.6*</td>
<td>1.5*</td>
</tr>
</tbody>
</table>

*The numbers in parentheses are for the highest-resolution shell.

(RMSD) Root mean square deviation.

CTD peptide phosphatase assays

Reaction mixtures ([5 µL containing 30 mM HEPES (pH 7.0), 100 mM NaCl, 10 mM MgCl₂, 500 µM ATP and 2 mM CTD peptide, and 5 µM to 200 µM Ssu72 were incubated for 45 min (for pSer5 peptide) or 12 h (for pSer7 peptide) at room temperature and then quenched by adding 0.5 µL of malachite green reagent. Assays with Fep1 were performed similarly, using 5 µM Fep1 and 500 µm peptides. Phosphate release was determined by measuring 

GST-CTD phosphatase assays

Purified GST-CTD fusion protein was phosphorylated in vitro by Cdk7 complex as described [Glover-Cutter et al. 2009]. GST phosphatase assays were performed in a total volume of 20 µL under the standard phosphate condition (50 mM Bis-Tris at pH 6.5, 20 mM KCl, 10 mM MgCl₂, 5 mM DTT) containing 200 ng of phosphorylated GST-CTD and the indicated amount of Ssu72 or Fep1. Reaction mixtures were incubated for 1 h at 30°C and stopped by adding 5 µL of 5x SDS loading buffer, and 2.5 µL from each reaction was resolved on an 8% SDS-PAGE gel. pSer7 levels were detected using a Western blot using the 4E12 antibody (Millipore), and pSer5 levels were detected using the 3E8 (Millipore) and 4E12 antibodies, respectively, and quantitated with LI-COR. The same experiment was also performed at room temperature, with similar results.

pSer7 peptide inhibition of pSer5 peptide dephosphorylation

Ssu72 (10 µM) was preincubated with 100 µM ATP, 200 µM MgCl₂, 1 mM, or 2 mM pSer7 CTD peptide in a buffer containing 50 mM HEPES (pH 7.0), 100 mM NaCl, and 20 mM MgCl₂ for 5 min at room temperature. A final concentration of 100 µM pSer5 CTD peptide was added in each reaction. The reaction was incubated for 45 min at 37°C and stopped by adding malachite green reagent. After another 30 min, OD500 was recorded for each reaction to determine phosphate release.

Acknowledgments

We thank Patrick Cramer for the yeast Fep1 expression plasmid, Stephanie Larochelle and Robert P. Fisher for Cdk7, Incuclovirix, Eng-Ping Hsin and David Zhang for discussions, and Neil Whalen, Stuart Myers, Rick Jackimowicz, and Howard Robinson for access to the X29A beamline at the NSLS. This research was supported in part by grants from the NIH to L.T. (GM077175) and J.L.M. (GM028983).

References


Supplementary Information for

An unexpected binding mode for a Pol II CTD peptide phosphorylated at Ser7 in the active site of the CTD phosphatase Ssu72

Kehui Xiang, James L. Manley, Liang Tong

1Department of Biological Sciences
Columbia University
New York, NY10027, USA

Correspondence information for Liang Tong
Phone: (212) 854-5203, FAX: (212) 865-8246
E-mail: ltong@columbia.edu
Fig. 1. SDS gels of purified wild-type human Ssu72 and yeast Fcp1.
Chapter IV: The yeast regulator of transcription protein Rtr1 lacks an active site and phosphatase activity
The activity of RNA polymerase II (Pol II) is controlled in part by the phosphorylation state of the C-terminal domain (CTD) of its largest subunit. Recent reports have suggested that yeast regulator of transcription protein, Rtr1, and its human homologue RPAP2, possess Pol II CTD Ser5 phosphatase activity. Here we report the crystal structure of Kluyveromyces lactis Rtr1, which reveals a new type of zinc finger protein and does not have any close structural homologues. Importantly, the structure does not show evidence of an active site, and extensive experiments to demonstrate its CTD phosphatase activity have been unsuccessful, suggesting that Rtr1 has a non-catalytic role in CTD dephosphorylation.

1 Department of Biological Sciences, Columbia University, New York, 10027, USA. Correspondence and requests for materials should be addressed to L.T. (email: ltong@columbia.edu).
The activity of RNA polymerase II (Pol II) is regulated by the phosphorylation state of the C-terminal domain (CTD) of its largest subunit, which contains the consensus heptapeptide repeat Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7 (YSPYTPS). While all the hydroxyl groups in this repeat can become phosphorylated, phosphorylation of Ser2 and Ser5 has received the most attention so far. For example, phosphorylated Ser5 (pSer5) is found at the promoter and during the initiation and early stages of Pol II transcription, and it mediates the recruitment of the 5′-end capping machinery and other factors. pSer5 is then dephosphorylated, and Ser2 becomes phosphorylated during the elongation and termination stages of transcription, which also facilitates 3′-end processing of the pre-mRNA transcript. In keeping with the importance of Ser2 and Ser5 phosphorylation, the kinases and phosphatases involved have been well studied. The zinc ion is coordinated around one side of the structure, but there are no prominent features in this region of the protein (see below). The main-chain carbonyl oxygen atoms of the first two ligands, e.g., in the structure of Kluyveromyces lactis Rtr1 (KlRtr1) at 2.5 Å resolution (Table 1, Supplementary Fig. S1). The bacterial growth media was supplemented with zinc during protein expression, and the purified protein had nearly stoichiometric amounts of zinc. Fluorescence scans were carried out on the crystal at the absorption edges of Zn, Ni, Co and Fe, and anomalous signals were observed only at the Zn edge, which were used to solve the structure. The full-length protein (residues 1–211) was used for crystallization, but only residues 1–152 were observed in the structure. SDS gels of the crystals showed that the C-terminal segment of KlRtr1 was removed by proteolysis during crystallization (Supplementary Fig. S2). The sequence conservation for these C-terminal residues is much weaker among Rtr1 and RPAP2 homologues (Fig. 1).

The structure of KlRtr1 (residues 1–152) contains five anti-parallel β-helices (αα-ΩΩ, Fig. 2a). A long loop (residues 70–112) connects helices ΩΩ and ΩΩ, and the three strictly conserved Cys residues (73, 78 and 111) are located in this loop. Residues 88–100 at the tip of this loop are disordered in the current structure (Fig. 2a), and this loop may have also been proteolysed in some of the Rtr1 molecules in the crystal (Supplementary Fig. S2). The fourth ligand to the zinc ion, His115 or Cys, is in the first turn of helix ΩΩ. This helix is followed by another long loop (residues 126–152) that wraps around one side of the structure (Fig. 2a).

We also obtained crystals of Ashbya gossypii Rtr1 (AgRtr1), although the best diffraction data set extended only to 3.5 Å resolution. By combining the structural information from KlRtr1 and primary phase information from Zn anomalous signals, we were able to observe another helix in the C-terminal region of AgRtr1, likely equivalent to residues 154–166 of KlRtr1 (Fig. 1). This helix is projected away from the rest of the protein and is stabilized by crystal packing interactions in the AgRtr1 crystal (Supplementary Fig. S3), suggesting that the C-terminal region of Rtr1 may function independently of the N-terminal region.

Rtr1 is a new type of zinc finger protein. The zinc ion is coordinated in a tetrahedral fashion by the four conserved ligands (Fig. 2a and b). The zinc-binding site is located near the surface, but there are no prominent features in this region of the protein (see below).
Phosphatase activity for Rtr1 is consistent with our structure, which lack detectable CTD phosphatase activity. Purified Rtr1/RPAP2 lack detectable CTD phosphatase activity. We next attempted to demonstrate Pol II CTD pSer5 phosphatase activity with our purified protein samples of His-tagged Rtr1 from K. lactis, A. gossypii, S. cerevisiae and several other fungal species. To test the possibility that Rtr1 in the absence of zinc or in complex with a different metal ion could be catalytically active, we used protein samples purified from E. coli cells that were grown without additional zinc in the medium. The zinc occupancy ranged between 10 and 80%, but other metal ions could be present in these samples. We also expressed and purified residues 1–332 of human RPAP2, as His-tagged and GST-fusion proteins, and full-length RPAP2 as a GST-fusion protein. Despite extensive efforts and using a large collection of different substrates (Supplementary Table S1), we failed to detect CTD pSer5 phosphatase activity for Rtr1 and RPAP2 under any of the conditions tested, while robust activity was observed with Ssu72. The substrates used include a CTD peptide Ser7-Pro5-Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Tyr1-Ser2 phosphorylated at the Ser5, Ser5 and Ser7, or Ser2, Ser5 and Ser7 positions, as it has been shown that RPAP2 requires Ser7 phosphorylation to interact with the CTD. These assays monitored the release of free phosphate from the reactions. We also used as the substrate the entire CTD purified as a GST-fusion protein and either phosphorylated with HeLa cell nuclear extract or purified Cdk7. The reactions were monitored by western blotting with an antibody specific for pSer5. We again failed to detect activity with multiple Rtr1/RPAP2 samples and under a variety of conditions (Fig. 3a and b and results not shown). In addition, we soaked Rtr1 crystals with CTD phosphopeptides, but were not able to identify any binding based on the crystallographic analysis. The structure of Rtr1 lacks an apparent active site. The lack of phosphatase activity for Rtr1 is consistent with our structure, which

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Summary of crystallographic information.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>K. lactis Rtr1</strong></td>
<td><strong>A. gossypii Rtr1</strong></td>
</tr>
<tr>
<td>Space group</td>
<td>P2₁2₁2₁</td>
</tr>
<tr>
<td>Cell dimensions a, b, c (Å)</td>
<td>45.2, 88.7, 94.9</td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90, 90, 90</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>1.2833</td>
</tr>
<tr>
<td>Resolution range (Å)*</td>
<td>50–2.5 (2.6–2.5)</td>
</tr>
<tr>
<td>Rmerge (%)</td>
<td>5.5 (35.5)</td>
</tr>
<tr>
<td>Compleness (%)</td>
<td>99 (100)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>3.7 (3.8)</td>
</tr>
<tr>
<td>Refinement</td>
<td></td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>50–2.5</td>
</tr>
<tr>
<td>No. of reflections</td>
<td>12,890</td>
</tr>
<tr>
<td>Rwork/Rfree (%)</td>
<td>22.2/27.2</td>
</tr>
<tr>
<td>No. of atoms</td>
<td>Protein</td>
</tr>
<tr>
<td></td>
<td>Ligand/ion</td>
</tr>
<tr>
<td></td>
<td>Water</td>
</tr>
<tr>
<td>R-factors</td>
<td>Protein</td>
</tr>
<tr>
<td></td>
<td>Ligand/ion</td>
</tr>
<tr>
<td></td>
<td>Water</td>
</tr>
<tr>
<td>R.m.s. deviations</td>
<td>Bond lengths (Å)</td>
</tr>
<tr>
<td></td>
<td>Bond angles (°)</td>
</tr>
</tbody>
</table>

Abbreviations: MAD, multiple-wavelength anomalous diffraction; R.m.s. deviation, root-mean-square deviation; RAD, single-wavelength anomalous diffraction.

The numbers in parentheses are for the highest resolution shell. One crystal was used for the K. lactis MAD data set, and one crystal for the A. gossypii SAD data set.

Cys73 and Cys78, are hydrogen bonded to the guanidinium group of Arg67, one of the few other conserved residues among these proteins (Fig. 1). The hydrogen bonds stabilize the two Cys residues as well as the loop connecting them, which is hydrophobic in nature and contributes to the formation of the hydrophobic core of Rtr1 (Fig. 2a). The zinc ion therefore seems to have a structural role, likely stabilizing the overall conformation of Rtr1.

We have identified Rtr1 as a new type of zinc finger protein. A search through the Protein Data Bank with Dalilite did not identify any close structural homologues, with the highest Z score being 3.3. The overall fold of Rtr1 is somewhat reminiscent of the HEAT repeats, although the positions of the helices are different compared with these repeats. The topology of the zinc ligands in Rtr1, with the first three located in loops and the last one in the beginning of a helix, has some similarity to the A20 family of zinc finger proteins, as illustrated by the structure of Rabex-5, a guanine nucleotide exchange factor for Rab5 that also binds monouquitin. The overall conformations of the protein backbone near the first two zinc ligands are similar between Rtr1 and Rabex-5 (Fig. 2c). On the other hand, the loop connecting them, and especially the loop connecting to the third ligand, has different numbers of residues (the zinc ligands in Rabex-5 have the motif C-X₃-C-X₁₁-C-X₂-C). In addition, the orientation of the helix containing the fourth ligand differs by nearly 90° between the two structures. Finally, the zinc-binding site in Rtr1 is part of a much larger structure, with 150 residues (Fig. 2a), whereas the zinc finger domain of Rabex-5 contains only 35 residues.

Purified Rtr1/RPAP2 lack detectable CTD phosphatase activity. We next attempted to demonstrate Pol II CTD pSer5 phosphatase activity with our purified protein samples of His-tagged Rtr1 from K. lactis, A. gossypii, S. cerevisiae and several other fungal species. To test the possibility that Rtr1 in the absence of zinc or in complex with a different metal ion could be catalytically active, we used protein samples purified from E. coli cells that were grown without additional zinc in the medium. The zinc occupancy ranged between 10 and 80%, but other metal ions could be present in these samples. We also expressed and purified residues 1–332 of human RPAP2, as His-tagged and GST-fusion proteins, and full-length RPAP2 as a GST-fusion protein. Despite extensive efforts and using a large collection of different substrates (Supplementary Table S1), we failed to detect CTD pSer5 phosphatase activity for Rtr1 and RPAP2 under any of the conditions tested, while robust activity was observed with Ssu72. The substrates used include a CTD peptide Ser7-Pro5-Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Tyr1-Ser2 phosphorylated at the Ser5, Ser5 and Ser7, or Ser2, Ser5 and Ser7 positions, as it has been shown that RPAP2 requires Ser7 phosphorylation to interact with the CTD. These assays monitored the release of free phosphate from the reactions. We also used as the substrate the entire CTD purified as a GST-fusion protein and either phosphorylated with HeLa cell nuclear extract or purified Cdk7. The reactions were monitored by western blotting with an antibody specific for pSer5. We again failed to detect activity with multiple Rtr1/RPAP2 samples and under a variety of conditions (Fig. 3a and b and results not shown). In addition, we soaked Rtr1 crystals with CTD phosphopeptides, but were not able to identify any binding based on the crystallographic analysis.

The structure of Rtr1 lacks an apparent active site. The lack of phosphatase activity for Rtr1 is consistent with our structure, which
Pol II from the cytoplasm into the nucleus
RPAP2 may have an important role in the assembly and transport of Pol II, and therefore it is likely that Rtr1 could mediate interactions among various proteins in the Pol II complex. Given the evidence that Rtr1/RPAP2 is required for recruitment and/or activation of the actual phosphatase that dephosphorylates pSer5 during the early stages of Pol II transcription, the identity of that phosphatase remains to be determined.

Methods

Protein expression and purification. Full-length Rtr1 from K. lactis, A. gossypii, S. cerevisiae and the N-terminal region (residues 1–332) of human RPAP2 were cloned into the pET28a vector (Novagen) and overexpressed in E. coli BL21 (DE3) Star cells at 20°C by the addition of 0.5 mM isopropyl-β-D-thiogalactopyranoside at OD600 of 0.6. To help the zinc enrichment in the protein, 0.1 mM ZnSO4 was added to the culture 1 h before induction. The expression constructs introduced hexa-histidine tags at the N terminus of the proteins.

The soluble proteins were purified by Ni-NTA (Qiagen) with the elution buffer containing 20 mM Tris (pH 7.5), 200 mM NaCl and 250 mM imidazole, followed by gel filtration (Superdex S-300, GE Healthcare) chromatography in a running buffer of 20 mM Tris (pH 7.5), 200 mM NaCl and 2 mM DTT. The proteins were concentrated to 20 mg ml−1 in a buffer containing 20 mM Tris (pH 7.5), 200 mM NaCl, 2 mM DTT and 5% (v/v) glycerol, flash frozen in liquid nitrogen and stored at −80°C.

The GST fusion RPAP2 proteins were made by cloning the full-length human RPAP2 or the N-terminal region (residues 1–332) into pGEX-4T-3 (GE Life Sciences) vector and overexpressed in E. coli BL21 (DE3) Star cells with the same protocol as mentioned above. The proteins were purified by glutathione Sepharose 4 Fast Flow (GE Healthcare) with 20 mM reduced glutathione in a buffer of 20 mM Tris (pH 7.5), 200 mM NaCl and 5% (v/v) glycerol.

Protein crystallization. Crystals of KIRtr1 were obtained with the sitting-drop vapour diffusion method at 20°C. The reservoir solution contained 100 mM NaCl, 100 mM Tris (pH 9.5) and 20% (v/v) PEG8000. The crystals belong to space group P212121, with one molecule in the asymmetric unit. Crystals of AgRP1 were obtained with the sitting-drop vapour diffusion method at 20°C. The reservoir contained 120 mM Tris (pH 7.6) and 16% (v/v) ethanol. The crystals belong to space group P4222, with one molecule in the asymmetric unit. Crystals were cryo-protected by the respective reservoir solutions supplemented with 25% (v/v) ethylene glycol and flash frozen in liquid nitrogen for data collection at 100 K.

Data collection and structure determination. X-ray diffraction data were collected on an ADSC charge-coupled device at the X29A beamline of National...
1. References

2. Synchrotron Light Source (NSLS). The diffraction images were processed and scaled with the HKL package. The atomic coordinates of the KlRtr1 structure have been deposited in the Protein Data Bank under the accession code 4FSC. Multiple-wavelength anomalous diffraction data sets were collected on the KlRtr1 crystal to 3.5 Å resolution, at the zinc absorption edge (in Å), and remote (1.2652 Å) wavelengths. The data processing statistics are summarized in Table 1. The Zn sites were located with the program CNS and Refmac5 against the data set at the peak wavelength. The statistics on the structure refinement are summarized in Table 1. For the AgI-R1 crystal, a single-wavelength anomalous diffraction data set to 3.5 Å resolution was collected at the zinc peak wavelength. An electron density map was obtained based on the single-wavelength anomalous diffraction data, which showed clear indications for several helices. The atomic model of KlRtr1 could be readily positioned into the density, revealing an extra helix in the C-terminal region in the AgI-R1 structure. Refinement of this structure model was not carried out due to the limited resolution.

3. CTD peptide phosphatase assays. Reaction mixtures (25μl) in the standard phosphate condition (10mM Bis-Tris (pH 6.5), 20 mM KCl, 10 mM MgCl₂, 1 mM EDTA) containing 0.5μM CTD peptide, R1R, RPA2 or SnuT2 at 1μM, 5μM or 20μM concentration were incubated at 30°C. Time-point samples were taken and quenched by adding 0.5 ml of malachite green reagent (BIOMOL Research Laboratories, Plymouth Meeting, PA). Phosphate release was determined by measuring A₅₃₅ and comparing it with a phosphate standard curve.

4. CTD phosphatase assay. Purified GST–CDT fusion protein was phosphorylated in vitro by Hela cell nuclear extract as described, or by Gfp6 complex as described. CTD phosphatase assays were performed in a total volume of 20μl in the standard phosphate condition containing 200μg phosphatase GST–CDT and indicated amount of R1R, RPA2 or SnuT2. Reactions were incubated for 1 h at 30°C, stopped by adding 5μl 5x SDS loading buffer and 2.5 μl from each reaction was resolved on an 8% SDS-PAGE gel. Phospho level was detected by western blot using H14 antibody (Covance).

5. All CTD peptide phosphatase assays and CTD phosphatase assays were additionally performed in different conditions by varying pH (5.5–8.5) or KCl concentration (20–500 mM) or additives (1–10 mM ZnCl₂ or MnCl₂ or MgCl₂) based on the standard phosphate condition.

References


Acknowledgements
We thank Amber Mosley and Lawrence Chasin for discussions; Stephane Larochelle and Robert P. Fisher for Cdk7 baculovirus; Shona Murphy for sharing a preprint on RPAP2; Neil Whalen, Stuart Myers and Howard Robinson for setting up the X29A beamline at the NSLS. This research is supported by grants from the NIH to LT (GM007175) and JLM (GM028983).

Author contributions
K.X. carried out all experiments. J.L.M. analysed the phosphatase data. L.T. conceived the project, analysed the data and wrote the paper. All authors commented on and contributed to the manuscript.

Additional information
Accession codes: The atomic coordinates of the KlRtr1 structure have been deposited in the Protein Data Bank under the accession code 4FC8.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Xiang, K. et al. The yeast regulator of transcription protein Rtr1 lacks an active site and phosphatase activity. Nat. Commun. 3:946 doi: 10.1038/ncomms1947 (2012).
Supplementary Information

The yeast regulator of transcription protein Rtr1 lacks an active site and phosphatase activity

Kehui Xiang,1 James L. Manley,1 Liang Tong1

1Department of Biological Sciences
Columbia University
New York, NY10027, USA

Correspondence information for Liang Tong
Phone: (212) 854-5203, FAX: (212) 865-8246
E-mail: lttong@columbia.edu
Supplementary Figure S1. A stereo view of the final 2F$_\text{e}$–F$_\text{e}$ electron density map at 2.5 Å resolution for helices αA and αB of KIRtr1, contoured at 1σ.
Supplementary Figure S2. An SDS gel of purified KIRtr1 and KIRtr1 crystals. Proteolysis during crystallization removed the C-terminal segment of KIRtr1. Proteolysis in the 87-101 region (αD-αE loop) may have also occurred in some of the KIRtr1 molecules in the crystal.
Supplementary Figure S3. The C-terminal helix in the AgRtr1 structure is stabilized by crystal packing interactions. Two views of the crystallographic dimer are shown.
Supplementary Table S1 Summary of attempted phosphatase assays

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Substrates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fungal Rtr1 (S. cerevisiae, K. lactis, A. gossypii, V. polyspora, Z. rouxii, D. hansenii, C. glabrata), full-length proteins with His-tag</td>
<td>CTD peptide with phosphates on Ser5, Ser5 and Ser7, or Ser2, Ser5 and Ser7 (peptides were produced by chemical synthesis)</td>
</tr>
<tr>
<td></td>
<td>GST-CTD phosphorylated by HeLa cell nuclear extract</td>
</tr>
<tr>
<td></td>
<td>GST-CTD phosphorylated by Cdk7</td>
</tr>
<tr>
<td>Human RPAP2 (residues 1-332), His-tagged or GST-fusion protein; full-length RPAP2, GST-fusion</td>
<td>GST-CTD phosphorylated by HeLa cell nuclear extract</td>
</tr>
<tr>
<td></td>
<td>GST-CTD phosphorylated by Cdk7</td>
</tr>
</tbody>
</table>

1 Reactions with the peptide substrates were monitored using the BioMol reagent. Reactions with GST-CTD were monitored by Western blot with the H14 antibody.