University of Vermont
UVM ScholarWorks

Graduate College Dissertations and Theses | Dissertations and Theses

2009

# The Investigation of Cleavage Factor IM by Crystallographic and Biochemical Techniques

Molly Coseno
*University of Vermont*

Follow this and additional works at: https://scholarworks.uvm.edu/graddis

## Recommended Citation

Coseno, Molly, "The Investigation of Cleavage Factor IM by Crystallographic and Biochemical Techniques" (2009). *Graduate College Dissertations and Theses*. 55.
https://scholarworks.uvm.edu/graddis/55

This Dissertation is brought to you for free and open access by the Dissertations and Theses at UVM ScholarWorks. It has been accepted for inclusion in Graduate College Dissertations and Theses by an authorized administrator of UVM ScholarWorks. For more information, please contact scholarworks@uvm.edu.

# THE INVESTIGATION OF CLEAVAGE FACTOR $I_M$ BY CRYSTALLOGRAPHIC AND BIOCHEMICAL TECHNIQUES

A Dissertation Presented

by

Molly Coseno

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Microbiology and Molecular Genetics

February, 2009

Accepted by the Faculty of the Graduate College, The University of Vermont, in partial fulfillment of the requirements for the degree of Doctor of Philosophy, specializing in Microbiology and Molecular Genetics.

Dissertation Examination Committee:

Sylvie Doublié, Ph.D Advisor

Gregory M. Gilmartin, Ph.D

David S. Pederson, Ph.D

Jeffrey P. Bond, Ph.D

Stephen J. Everse, Ph.D Chairperson

Frances E. Carr, Ph.D Vice President for Research and Dean of Graduate Studies

Date: December 18, 2008

# ABSTRACT

RNA maturation involves several steps prior to export of the mRNA out of the nucleus and translation in the cytoplasm. Pre-mRNA 3'-end processing is one of such steps, and comprises the endonucleolytic cleavage and polyadenylation of the 3'-end of the pre-mRNA. These two steps involve more than 14 processing factors that coordinate multiple protein-protein and protein-RNA interactions necessary to coordinate efficient cleavage and polyadenylation. To date, many of these interactions have been investigated biochemically and require additional structural characterization both to confirm and highlight key residues involved in substrate contacts. Further structural characterization will also open investigation into the mechanism of 3'-end processing by providing structural insight into the coordination of multiple binding components. The cleavage factor $I_m$, CF $I_m$, is a component of the 3'-end processing machinery and plays an important role early, during endonucleolytic cleavage, and additionally to increase polyadenylation efficiency and regulate poly(A) site recognition. CF $I_m$ is composed of a small 25 kDa subunit, CF $I_m$25, and a large, either 58 kDa, 68 kDa, or 72 kDa subunit. The 25 kDa subunit of CF $I_m$ interacts with both the RNA and other processing factors such as the poly(A) polymerase, Clp1, and the larger subunit of CF $I_m$. It is our goal to crystallize CF $I_m$25 alone and in complex with one of its interacting partners to better understand CF $I_m$25 contributions to pre-mRNA 3'-end processing.

The structural investigation of CF $I_m$25 and its binding partners has accomplished four major objectives: 1) Characterized the crystal structure of CF $I_m$25 alone and bound to diadenosine tetraphosphate, 2) Provided insight into the oligomeric state of the CF $I_m$ complex, 3) Determined the binding properties of the Nudix domain of CF $I_m$25 and its function in 3'-end processing, 4) Further characterize the interactions between CF $I_m$25 and PAP, CF $I_m$68, and Clp1. These results demonstrate CF $I_m$25 is a dimer both in solution and in the crystal suggesting that it is likely to be a dimer in the CF $I_m$ complex. The nucleotide binging capability of CF $I_m$25 has no apparent role in 3'-end processing *in vitro* but may provide a function outside of 3'-end processing or may directly be involved in RNA recognition. The additional investigation of complex interactions with the 25 kDa subunit of CF $I_m$25 suggests that although these factors interact during the 3'-end processing event additional mechanisms may play a role in stabilizing those interactions.

# CITATIONS

Material from this dissertation has been published in the following form:

Coseno, M., Martin, G., Berger, C., Gilmartin, G., Keller, W., and Doublié, S. (2008) Crystal Structure of the 25kDa Subunit of Human Cleavage Factor $I_m$. *Nucleic Acids Research* 36 (10), 3474-3483.

# ACKNOWLEDGMENTS

Completing my graduate studies would not have been possible without the guidance of my mentor, Dr. Sylvie Doublié. She has imparted critical and independent thinking that has allowed me to progress into a young scientist. In addition, she is an inspiration to all women in science both for her success in crystallography and as a mother. I would also like to thank Dr. Mark Rould for his inspiring lectures on crystallography, enthusiasm, and willingness to take time out of his schedule to address experimental problems.

The past and present members of the Doublié lab have been a tremendous asset: Dr. Pierre Aller and Karl Zahn were instrumental in teaching me the structural programs and troubleshooting refinement issues. The days of unsuccessful research would have been grueling if not for Justin Meyette. We had some good times. I would also like to thank Qin Yang for his great smile, cheerfulness, and for being my in-lab graduate companion. Additionally, Dr. Frederick Faucher was extremely helpful when I was formatting my thesis and always available for crystallography questions.

Most importantly, I would like to thank my family and friends. My mom who is always available for guidance, support, and multiple laughs. J.P. and Mark, my two brothers, I could not do without and are the best bros a sister could ask for. My stepfather Jay is sincere and always full of encouragement. Lastly, Ray, my husband and best friend, is always available to cheer me up and make any stressful situation seem minimal in the context of life.

# TABLE OF CONTENTS


CITATIONS ..... ii
ACKNOWLEDGMENTS ..... iii
LIST OF TABLES ..... vi
LIST OF FIGURES ..... vii
CHAPTER 1: ..... 1
*Eukaryotic pre-mRNA 3'-end Processing* ..... *1*
Capping of the pre-mRNA ..... 3
The Splicing Reaction ..... 4
pre-mRNA 3'-end Cleavage and Polyadenylation Events ..... 5
Yeast pre-mRNA 3'-end Processing ..... 6
Mammalian Cleavage and Polyadenylation ..... 9
CF $I_m$ ..... 18
The Nudix domain of CF $I_m$25 ..... 22
Specific Aims ..... 25
CHAPTER 2: ..... 44
CONTRIBUTIONS ..... 45
*Crystal Structure of the 25 kDa Subunit of Human Cleavage Factor $I_m$* ..... *46*
ABSTRACT ..... 47
INTRODUCTION ..... 48
MATERIALS AND METHODS ..... 51
Protein purification ..... 51
Crystallization ..... 52
Crystallographic data collection ..... 53
Structure determination and refinement ..... 53
Oligomeric state determination ..... 54
Hydrolase activity assay ..... 55
Fluorescence measurements ..... 55
Cleavage and Polyadenylation assays ..... 56
RESULTS ..... 58
Crystal structure of human CF $I_m$25 ..... 58
Description of the structure ..... 58
Dimer interactions ..... 60
Metal binding ..... 61
Enzymatic assays and substrate binding ..... 61
Complex of CF $I_m$25 with $Ap_4A$ ..... 63
Chelation Affect of $Ap_4A$ ..... 65
Electrostatic surface and putative RNA and protein binding regions ..... 66
DISCUSSION ..... 67
FUNDING ..... 71
ACKNOWLEDGMENTS ..... 72
CHAPTER 3: ..... 83
*Crystallization Trials of CF $I_m$25 with Various Binding Partners* ..... *83*
Introduction: CF $I_m$ ..... 84
Proteolytic Degradation Occurs at the C-terminal Region of CF $I_m$68 ..... 85
UGUAN Containing RNAs do not Stabilize CF $I_m$ ..... 86
Complex Crystallization Attempts with CF $I_m$25 and 68 kDa Truncation Mutants ..... 87
Conclusions ..... 89
Introduction: CF $I_m$25 and PAP ..... 90
A Stable Complex is not Maintained by Gel Filtration Chromatography ..... 91
Complex is Stabilized during Electromobility Shift Assay ..... 93
Crystallization Attempts of the CF $I_m$25/PAP513 Complex ..... 93
Conclusions ..... 95

Introduction: Fip1 ..... 97
Purification and Expression Attempts of Fip1 ..... 98
Construct 1-355 of Fip1 Does not Stabilize Expression and Purification Attempts ..... 100
Point Mutation in Fip1 Constructs ..... 101
Conclusions ..... 102
Introduction: Clp1 ..... 103
Expression and Purification ..... 104
Complex Crystallization Trials with CF $I_m$25 and Clp1 ..... 105
Dynamic Light Scattering Experiments ..... 105
Conclusions ..... 106
Protein Purification ..... 122
RNA preparation ..... 125
Electromobility Shift Assays ..... 125
Oligomeric State Determination ..... 126
CHAPTER 4: ..... 127
*Future Directions and Conclusions* ..... *127*
Introduction ..... 128
Cleavage Factor $I_m$ ..... 128
Alternative methods ..... 129
Determining more suitable candidates ..... 130
Is CF $I_m$25's nucleotide binding affinity necessary for its function? ..... 131
Unresolved Questions ..... 132
Cleavage Factor $I_m$25 and Poly(A) polymerase ..... 133
Factor Interacting with Pap1 ..... 134
Alternative methods ..... 135
REFERENCES ..... 137

# LIST OF TABLES

Table 1: Yeast pre-mRNA 3'-end Processing components and mammalian homologues 29
Table 2: Human Nudix genes and hydrolases ..... 43
Table 3: Data Collection and Refinement Statistics ..... 82

# LIST OF FIGURES

Figure 1: Eukaryotic pre-mRNA 3'-end Processing ..... 28
Figure 2: Crystal structure of yeast Clp1/Pcf11 bound to ATP ..... 30
Figure 3: Crystal structure of Pap1 bound to oligo (A) ..... 31
Figure 4: Eukaryotic cleavage components ..... 32
Figure 5: Poly(A) tail elongation machinery ..... 33
Figure 6: HAT domain of murine CstF77 ..... 34
Figure 7: Structural similarities between CPSF73 and CPSF100 ..... 35
Figure 8: Interaction of the 25-kDa subunit of CF $I_m$ with PAP513 ..... 36
Figure 9: Crystal structure of bovine PAP513 in complex with 3'dATP ..... 37
Figure 10: Scheme of the direct interactions among processing factors ..... 38
Figure 11: Domain architecture of the CF $I_m$ subunits ..... 39
Figure 12: Signature fold of Nudix proteins ..... 40
Figure 13: Mechanism of substrate hydrolysis by $Ap_4A$ pyrophosphatase ..... 41
Figure 14: Alternative mechanism of GDP-mannose hydrolase ..... 42
Figure 15: Sequence alignment of CF $I_m25$ with Nudix proteins ..... 73
Figure 16: Gel filtration profile of CF $I_m25$ ..... 74
Figure 17: Domain organization of CF $I_m25$ ..... 75
Figure 18: CF $I_m25$ steady state tryptophan fluorescence experiments with ATP and $Ap_4A$ ..... 76
Figure 19: Complex of CF $I_m25$ and diadenosine tetraphosphate ..... 77
Figure 20: Comparison with other Nudix proteins ..... 78
Figure 21: Positions of DR1025 ATP and $Ap_4A$ in the CF $I_m25$ structure ..... 79
Figure 22: Poly(A) assays in the presence of additional nucleotide ..... 80
Figure 23: Surface representation of CF $I_m25$ ..... 81
Figure 24: Diagram of the Disordered Regions of CF $I_m68$ ..... 107
Figure 25: Dynamic Light Scattering of the CF $I_m$ complex ..... 108
Figure 26: Domain architecture of the 68kDa truncation mutants ..... 109
Figure 27: Gel filtration profile of CF $I_m25$ and N68 complex ..... 110
Figure 28: Dynamic light scattering of CF $I_m25$ and N68 complex ..... 111
Figure 29: Secondary structure prediction of SELEX 1 RNA ..... 112
Figure 30: Crystal of CF Im25 and N68 ..... 113
Figure 31: Architecture of bovine PAP and regions involved in binding CF $I_m25$ ..... 114
Figure 32: Truncation mutant of bovine PAP ..... 115
Figure 33: Gel shift of PAP513 and CF $I_m25$ complex ..... 116
Figure 34: Domain architecture of Fip1 and regions interacting with PAP513 ..... 117
Figure 35: Disorder prediction profile of human Fip1 ..... 118
Figure 36: SDS-page gel of Ni-NTA purified 1-355Fip1 before and after TEV cleavage ..... 119
Figure 37: Domain architecture of human Clp1 ..... 120
Figure 38: Dynamic light scattering of Clp1/CF $I_m25$ complex at 15°C ..... 121

# CHAPTER 1:

## Eukaryotic pre-mRNA 3'-end Processing

Eukaryotic pre-messenger RNA undergoes several steps of processing prior to export from the nucleus and translation in the cytoplasm. These steps include the addition of a 5' 7-methylguanosine ($m^7G$) cap, splicing of intronic messages, and lastly, cleavage followed by polyadenylation as shown in **Figure 1**. These processing steps are thought to be tightly coupled and occur co-transcriptionally by recruitment of processing factors to the actively transcribed RNA via interaction with the carboxy terminal domain (CTD) of RNA polymerase II (*1*). This has also been demonstrated *in vitro*, where it is possible to monitor transcription and the subsequent processing steps individually. In doing so it becomes evident that not only is processing tightly coupled to transcription but each processing event is capable of enhancing the subsequent event (*2, 3*). During transcription this occurs through interaction with the CTD. The CTD is composed of tandem heptad repeats with the consensus sequence, $Y^1S^2P^3T^4S^5P^6S^7$. Phosphorylation and dephosphorylation of specific residues within the CTD consensus sequence, in response to transcriptional checkpoints, is crucial for co-transcriptional processing by recruitment of pre-mRNA 3'-end processing factors to the site of active transcription (*4, 5*). This coupling event promotes rapid mRNA stabilization by protection of the initially transcribed pre-mRNA and additionally provides a platform to tether processing factors to each other.

## Capping of the pre-mRNA

The initial step in pre-mRNA 3'-end processing is capping. The capping process is orchestrated by phosphorylation of serine residues within the CTD heptad resulting in recruitment of the mammalian capping enzyme, CE. CE contains both the RNA triphosphatase and guanylyltransferase activities and was identified by crosslinking experiments to associate with the elongation complex at the promoter region. In yeast, the triphosphatase and guanylytransferase activities are contained on two separate genes (*6*). The 7-methyltransferase has a more distributive binding pattern and is associated with the elongation complex within the coding region as well as at the 5' end of the gene, as demonstrated by crosslinking experiments (*4*). This finding suggests that CE is at the transcriptional start site at the time of transcriptional initiation to cap the 5' end once available. The capping reaction begins with the removal of a phosphate from the 5' end of the transcript by the RNA 5' triphosphatase activity of the CE. The guanylyltransferase activity then adds a guanosine residue in an inverted position via a 5'-5' triphosphate linkage and the N7 position of guanosine is subsequently methylated by 7-methyltransferase. Upon methylation the $m^7G$ cap structure is bound by the cap binding complex (CBC) at the 5' end of the mRNA. The CBC bound cap plays an important role, preventing degradation of the RNA by 5'-3' exonucleases and is additionally important in enhancing translation in the cytoplasm along with other translation initiation factors (*6, 7*). Finally, *in vitro*, the cap structure has been shown to stimulate the next processing step, splicing, by enhancing the splicing reaction of capped vs. uncapped RNAs (*6*).

## The Splicing Reaction

Following the addition of the cap structure, pre-mRNA 3'-end processing continues with the removal of intronic RNA during the splicing process. The splicing machinery, also termed spliceosome, consists of five RNA protein complexes termed U snRNPs, small nuclear ribonuclear protein particles, along with multiple serine-rich (SR) proteins. The snRNPs consist of U1, U2, U4, U5, and U6 and include several transient splicing factors. There is some evidence to suggest that the splicing event is influenced, as is the case for the capping reaction, by phospho-CTD interactions (*1, 8*). Yeast Prp40p, a component of U1 snRNP, was shown to interact directly with the phosphorylated CTD (*9*). A similar role of phospho-CTD interactions in splicing has yet to be demonstrated in the mammalian spliceosome.

Splicing occurs in two steps and is controlled by recognition of sequence elements within the RNA. Splicing begins with 5' splice site cleavage, yielding a lariat structure. Second, cleavage occurs at the 3' splice site followed by exon ligation. Spliceosome assembly is initiated with the binding of U1 snRNP to the 5'splice site by recognition of the mammalian consensus sequence AG/GURAGU. SF1/mBBP recognizes the branchpoint sequence, YNYURAC, at the 3' end along with U2 auxiliary factor's, U2AF, binding at the pyrimidine tract and 3' AG splice site. U2 snRNP binds, although weakly.

These initial binding steps form the early (E) complex. The next step in splicing is the formation of the A complex, which occurs when U2snRNP binds at the branchpoint. The association of additional factors U4/U6-U5 snRNP forms the B

complex. Subsequent to B complex formation, a rearrangement of the spliceosome occurs to form the active catalytic C complex (*10*).

Along with snRNPs, serine/arginine-rich proteins (SR proteins) function in the splicing mechanism. SR splicing factors are involved in the initial recruitment of U1 snRNP and U2AF as well as in the later A and B complex transition and C complex (*10*). They have been shown to have a role in the regulation of 3' splice site selection by binding at an exon sequence and enhancing the splicing of the adjacent intron (*10*). Additionally, they can promote the recognition of alternative 5' splice sites and function in constitutive splice site selection by defining small exon borders amongst larger introns (*11*). Splicing has also been shown to couple events in RNA processing and is evident by the stimulation of the splicing event by poly(A) polymerase's (PAP) recruitment of a component of U2AF, U2AF 65, to the 3' splice site. The coupling of splicing and processing events is also made evident with the stimulation of polyadenylation via U2AF 65's interaction with the 59 kDa subunit of CF $I_m$ (*12*).

## pre-mRNA 3'-end Cleavage and Polyadenylation Events

The newly capped and spliced pre-mRNA in eukaryotes is processed in a two-step reaction consisting of endonucleolytic cleavage of the pre-mRNA followed by addition of a poly(A) tail at the 3' end of the upstream cleavage product (*13-15*). Many of the factors present in the mammalian pre-mRNA 3'-end processing machinery are conserved in the yeast system. Although I will be focusing primarily on the mammalian system, I will provide a brief summary of pre-mRNA 3'-end processing events in yeast.

## Yeast pre-mRNA 3'-end Processing

Much like the mammalian system, yeast pre-mRNA 3'-end processing involves recognition of sequences elements for regulation and efficient processing. These elements are slightly different in sequence context from the mammalian system and because of such differences these elements can be recognized by different factors.

The yeast sequence elements that dictate proper and efficient cleavage and polyadenylation are the AU-rich element or efficiency element (EE), the A-rich positioning element (PE), the cleavage site, and the U-rich elements that border the cleavage site. The efficiency element UAUAUA, with U at both the first and fifth positions is the most efficient for processing, although it is not critical for the cleavage reaction (*16, 17*). The PE is similar to the mammalian AAUAAA hexamer or polyadenylation signal (PAS) in dictating the proper 3'-end processing position. The optimal PE sequence context in yeast is AAUAAA or AAAAAA and is positioned 10-30 nucleotides upstream of the cleavage site. The cleavage site or poly(A) site consists of the sequence Py(A)n and functions along with the upstream sequence element (USE) and downstream sequence element (DSE) to specify proper cleavage site selection (*18*).

The yeast pre-mRNA 3'-end processing machinery consists of more than 20 processing factors. These factors can be broken down into smaller subcomplexes of which core homologs can be found in the mammalian system (**Table 1)**. *In vitro*, the yeast cleavage reaction requires cleavage factor IA (CF IA), cleavage factor IB (CF IB), and cleavage factor II (CF II). Multiple other protein factors co-purify with the cleavage

and polyadenylation machinery and are important for increasing the efficiency and linking transcription to 3' pre-mRNA processing (*13, 19*).

The crystal structure of components of CF IA, Clp1-ATP in complex with the Clp1 binding region of Pcf11 reveals structural interactions, which are likely to be present in the mammalian homologues **(Figure 2)**. A high structural similarity with the mammalian homologues is expected due to high sequence conservation within the Clp1/Pcf11 binding interface. The Clp1 component has a Walker A motif which harbors the ATP nucleotide binding site where the base selectivity is mediated by interaction of a glutamate residue from the amino terminal domain with the N6 position of adenine (*20*). Additionally, the structure of the RNA recognition motifs (RRM) of Hrp1, a CF IB factor, has been solved in complex with the AU-rich efficiency element (*21*). This structure depicts the recognition of a six tandem repeat of AU by two RRMs of Hrp1 arranged in parallel. The RRMs adopt a more closed conformation upon RNA binding creating a noticeable RNA binding cleft between the two RRMs. This is most apparent in solution, where the RRMs behavior alone resembles two independent rigid bodies that become tightly associated in RNA titration experiments in solution (*21*).

Pre-mRNA 3'-end processing concludes with the addition of the poly (A) tail by Pap1. Interestingly, the poly(A) tail length in yeast is significantly shorter than in the mammalian counterpart, averaging only 50-70 nucleotides (*17*). There are several crystal structures of Pap1 including a more recent crystal structure of Pap1 in complex with 25 residues of Fip1, a component of the complex PF 1. The Fip1 sequence is highly divergent across species, including the Pap1 binding region which was crystallized (*22*).

Fip1 is predicted to be highly flexible as suggested by POODLE plots of the protein's disordered regions and suggested by the small fragment crystallized (*23*). Other crystal structures of Pap1 were captured in the closed, cordycepin (3'dATP), and ATP and oligo(A) bound states (*24-26*). Pap1 exhibits a three domain polymerase structure composing a N-terminal, central, and a C- terminal domain distinct in Pap1 and involved in RNA primer binding. The initial structures of Pap1 attempted to address the domain movements of the protein necessary to facilitate ATP recognition and catalysis. In doing, so Balbo et al. 2007 described a closed conformation of Pap1 not previously described where flexibility depends on two hinge regions located in the middle domain. Additionally, superposition of ATP or ATP analog into this structure suggests possible base specificity through contacts with a neighboring asparagine residue (*24*). The most recent structure of a Pap1 ternary complex with ATP and oligo(A) was obtained by a mutation of one of the catalytic aspartates, D154A, capturing the enzyme in a closed conformation with RNA bound **(Figure 3)** (*25*). Overall, the ternary complex provides details of the path of the RNA and possible protein-RNA interactions although B-factors of the adenine residues close to the active site are elevated.

Due to similarities between the yeast and mammalian systems the structural data of yeast pre-mRNA 3'-end processing factors can provide further insight into the mechanism of mammalian 3'-end processing which will be introduced below.

## Mammalian Cleavage and Polyadenylation

Mammalian endonucleolytic cleavage of the pre-mRNA and addition of a poly(A) tail at the 3' end are additionally coupled as seen with capping and splicing. This coupling requires multiple protein-protein and protein-RNA interactions and has been shown to require ATP for the polymerization reaction, cleavage, and the recruitment of cleavage factors to the pre-mRNA in a substrate specific manner. In addition to the ATP requirement, creatine phosphate (CP) has also been proposed as a necessary factor for efficient pre-mRNA 3'-end processing. CP's role was initially thought to be necessary for replenishing the ATP pool in the nuclear extract due to contaminating ATPases. However, the lack of CP hydrolysis to compensate for a diminishing ATP pool suggests otherwise (*14*). CP may in fact play a role as an effector molecule for a component of the cleavage and polyadenylation machinery.

Despite its seemingly simplistic two-step mechanism, it has been estimated that more than 14 factors are required for mammalian pre-mRNA 3'-end processing and both cleavage and polyadenylation steps are crucial for transcript fidelity and translation. This is evident when disruption of polyadenylation occurs and results in the loss of cytoplasmic transport and a subsequent decrease in protein expression (*16*). Furthermore, addition of a poly(A) tail and binding of the poly(A) binding protein (PABP) promotes stability at the 3'-end, the primary target of degradation in the cytoplasm. The processing of the 3'-end also enhances the translation of the mRNAs by the presence of both the cap and poly(A) tail. Lastly, pre-mRNA 3'-end processing is crucial in the coupling of transcription with further processing events in which 3' processing factors interact with

the CTD of RNAP to control transcriptional initiation and to signal termination via poly(A) site strength and tail synthesis (*6*).

The factors that are necessary and sufficient to reconstitute cleavage and polyadenylation in a mammalian *in vitro* system are poly(A) polymerase (PAP), cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), cleavage factor $I_m$ (CF $I_m$), cleavage factor $II_m$ (CF $II_m$) and the nuclear poly(A) binding protein 1 (PABPN1) **(Figures 4 and 5**). Recruitment of the mammalian cleavage and polyadenylation machinery to the newly formed 3'-end of the pre-mRNA relies on the recognition of two conserved sequence elements: the highly conserved hexamer AAUAAA polyadenylation signal (PAS) recognized by CPSF and a GU/U-rich sequence downstream of the cleavage site recognized by CstF. Additionally a set of UGUA elements, a third sequence element found upstream of the cleavage site and not as universally conserved, is recognized by CF $I_m$ and functions in poly (A) site selection (*27, 28*). The AAUAAA sequence is found in 70% of sequenced human and mouse mRNAs of which the second most common sequence is AUUAAA (*16*). The importance of the high conservation in the sequence context of the hexamer is highlighted in α- and β-thalassemic patients where a point mutation occurs in the last position of the conserved hexamer resulting in diminished levels of mRNA (*29*). Of additional importance, is the positioning of the hexamer 10-30 nucleotides downstream of the poly(A) site. Deletions within this region have been shown to result in the formation of a new cleavage site, which reconstitutes the original length of the deleted portion. The GU/U rich DSE is positioned approximately 30 nucleotides from the cleavage site and has a lower sequence

conservation than the PAS hexamer. The remaining cleavage site element is found between the PAS hexamer and the DSE with an optimal cleavage site containing CA (*13, 16*).

The cleavage event is initiated upon recognition of the conserved hexamer, by the 160 kDa subunit of CPSF, and binding at the GU/U-rich DSE by CstF64 **(Figure 2)**. The recognition of these elements is enhanced by loading both CPSF and CstF onto the pre-mRNA during transcription at the promoter region. The stabilization of both CPSF and CstF on the RNA occurs through interactions at the 5' end of the RNA with the CBC. This interaction effectively links the 5'-end and the 3'-end of the pre-mRNA during the capping event and demonstrates crosstalk amongst processing factors. *In vitro* capping experiments with CBC depleted nuclear extracts have demonstrated CBC stabilization of 3'-end processing factors primarily affects cleavage, having no effect on polyadenylation. In addition, *in vitro* reconstitution of the cleavage activity with recombinant proteins does not fully restore the cleavage activity with the addition of the CBC. This suggests the link between the 5'-end and 3'-end is facilitated by additional processing factors through interaction with the CBC (*30*).

The additional coupling of the cleavage reaction to transcription is demonstrated by the increase in the efficiency of the cleavage reaction in the presence of the CTD (*31*). CPSF consists of five factors CPSF30kDa, 160kDa, 73kDa, 100kDa, and Factor Interacting with Pap1 (Fip1), all of which are required for cleavage and/or polyadenylation. It has been suggested that the 160 kDa subunit alone recognizes and binds at the hexamer, although this interaction is likely to be stabilized and strengthened

by its other components. Stabilization at the conserved hexamer by CPSF160 requires CstF binding and most likely CF $I_m$. The 30 kDa subunit of CPSF can be crosslinked to the hexamer RNA sequence and may also contribute to RNA binding via contacts with its five zinc finger domains and zinc knuckle. It has been previously suggested that the 30 kDa factor may not be a necessary component of CPSF due to CPSF30's weak association with its other components (*14*). The evidence stemmed from two published reports detailing the purified CPSF complex in which CPSF30 did not appear on an SDS-PAGE gel (*32, 33*). More than likely the association of the 30 kDa subunit with its other components is enhanced during the appropriate processing events and has a more transient interaction otherwise.

Like many of the processing factors, CstF is a multi-component complex consisting of a 50 kDa, 64 kDa, and 77 kDa subunits. The 77 kDa subunit is involved in bridging both the 64 kDa and 50 kDa subunits. The 77 kDa subunit from *Encephalitozoon cuniculi*, a protozoan parasite, crystallizes as a homodimer containing only α- helices distinct to the HAT-repeat family (*34*). The dimer interaction is also consistent with the crystal structure of the murine CstF77, which represents the entire 77 kDa structure except for an estimated 200 residues at the carboxy terminus **(Figure 6)** (*35, 36*). In addition, NMR structures for the N- terminal and C- terminal regions of the 64 kDa subunit have been solved (*36, 37*). The N-terminus of the 64 kDa subunit has the characteristic RRM fold that most closely resembles that of U1A (*36*). There is no structural data so far for the 50 kDa subunit of CstF. CstF50, aside from its role in 3' processing, provides a co-transcriptional link between the damage repair pathway

through its interaction with the repair proteins BRCA1/BARD1 and the CTD following cellular damage (*38, 39*).

CF $I_m$ is another component of the cleavage machinery and is an oligomer composed of a small 25 kDa subunit (CF $I_m$25, also referred to as CPSF5 or NUDT21) and a larger subunit of either 59 kDa, 68 kDa, or 72 kDa (*40*). Despite the presence of an RNA binding domain in the 68 kDa subunit, high affinity RNA binding requires both the 25 kDa and 68 kDa subunits and is facilitated by the 25 kDa amino terminal region and the arginine/serine rich region of the 68 kDa subunit (*41*). As presented so far, most of the cleavage factors associate with each other as well as the RNA. It can be suspected that following a weaker binding by individual factors to their recognition sequences, there is a rearrangement of the RNA, facilitating these factors to come into close contact. This rearrangement then allows for multiple protein-protein interactions and strengthens both binding to the recognition elements and complex formation. As of yet there is no data to indicate a specific change in the RNA structure upon complex formation. These events also are thought to occur in an orchestrated manner to increase efficiency of the reaction. This is suggested by the stimulation of the cleavage reaction following preincubation of CF $I_m$ with the RNA and subsequent addition of CstF, PAP, and CPSF (*40*)**.** The stimulation of cleavage only occurs when CF $I_m$ is preincubated with RNA, suggesting CF $I_m$ 's proficient binding to the RNA could then allow other cleavage factors to bind more efficiently. Alternatively the cleavage reaction occurs in a step wise order of component addition.

In addition to the factors previously mentioned, the cleavage reaction requires CF $II_m$ and PAP to proceed. CF $II_m$ is the least understood of the cleavage factors and presently does not have all of its components identified and purified to homogeneity (*14*). Initial characterization of CF $II_m$ divides it into two subcomponents, characterized by two separate functions: cleavage stimulatory activity and facilitated protein-protein interactions. The first subcomplex of CF $II_m$ interacts with CPSF and CF $I_m$ and fractionates with the Clp1 protein. The other subcomplex of CF $II_m$ is thought to be involved in stimulating the cleavage event (*6*). The recruitment of PAP to the assembled cleavage complex involves interactions with CPSF to increase binding at the hexamer and is likely the predominant mechanism for coupling cleavage to polyadenylation. This positions PAP at the poly(A) site following cleavage. The cleavage event occurs 20-30 bases downstream of the conserved hexamer. The 73 kDa subunit of CPSF is a possible candidate as the endonuclease in this event. The endonuclease activity of CPSF73, based on *in vitro* data, is dependent upon the presence of calcium for efficient cleavage and is sequence independent. The specificity of the cleavage event *in vivo*, by CPSF73 could be directed by the preferential binding of CstF64 and CPSF160 to either the GU/U-rich or PAS site. Additionally, there is no described requirement in the processing reactions for calcium and may not be necessary *in vivo* (*42*). Further investigation to characterize CPSF73's involvement as the endonuclease in the cleavage process should be evaluated *in vitro* with CPSF73 depleted nuclear extracts as well as in the presence of the cleavage components to restore the specificity of CPSF73 and the calcium requirement.

In conjunction with the biochemical investigation of the endonuclease activity of CPSF73, Mandel et al. 2006 solved the crystal structure of both the 73 kDa subunit and the yeast 100 kDa subunit (Ydh1p) of CPSF **(Figure 7)**. CPSF73 and CPSF100 both belong to the metallo-β-lactamase superfamily (MBL) of zinc dependent hydrolases. CPSF73 has the structurally conserved metallo active site, 4 of the 5 conserved MBL motifs, and a β-CASP domain. Interestingly, CPSF100 which reflects a similar structural layout, lacks the coordinating ions necessary for catalysis, suggesting a nonfunctional MBL enzyme. The active site of CPSF73 coordinates two zinc ions and a sulfate ion. Appropriately, with other members of the metallo superfamily, CPSF73 has a hydroxide ion positioned near the sulfate, which mimics the potential substrate binding position. The hydroxide ion is a potential nucleophile for the hydrolysis reaction and is a strong indication of a potential endonuclease activity (*42*). Follow up research has added further insight into the role of the endonuclease with the investigation of histone mRNA processing (*43*). Point mutations of critical MBL residues based on the structure and sequence conservation in both CPSF100 and CPSF73 are inhibitory to the endonuclease activity necessary for creating the histone new 3'-end (*43*)**.** This recent finding suggests both the 100 kDa and 73 kDa function together in the 3' –end endonuclease cleavage of all pre-mRNAs. This is due to core conservation between histone 3'-end maturation and the factors involved in establishing the new 3'-end of all other pre-mRNAs, aside from proteins involved in the adenylate tail synthesis.

As mentioned previously, cleavage generally occurs on the 3' side of an A, with a preference for a CA at the site of cleavage. Once the endonuclease has cleaved the pre-

mRNA, CstF and CF $II_m$ are released with the downstream cleavage fragment allowing polyadenylation to proceed. The cleavage reaction is tightly coupled to the recruitment of the remaining factors in the last processing step, polyadenylation.

The polymerization of the adenosine tail requires at minimum PAP, PABPN1, and CPSF. CPSF160 and PABPN1 confer an increase in binding and specificity to the otherwise, nonspecific and distributive RNA polymerase, PAP. Fip1 in conjunction with CPSF160 functions to recruit PAP to the poly(A) site to stimulate polymerase activity by binding U-rich USEs within the pre-mRNA. The recruitment of PAP for efficient polyadenylation also requires interaction with CF $I_m$ in conjunction with Fip1 and CPSF160, where CF $I_m$ association with Fip1 is involved in PAP stimulation (*28, 44*). hFip1 interacts via its C-terminal arginine and arginine/glutamic acid region with the RBD and N-terminal region of PAP (*44*). In contrast, yeast Fip1 alone inhibits Pap1's processivity and switches Pap1 to a distributive mode *in vitro*. However, this inhibition can be alleviated by contributions from subunits of CPF for Pap1 stimulation (*45*).

The polymerization activity is regulated both *in vivo* and *in vitro* by multiple phosphorylation states within the carboxyl-terminal domain of PAP. *In vivo*, this regulation coincides with the activity of the cell cycle where phosphorylation of the C-terminal region of PAP decreases enzyme activity. The kinase most likely to be responsible for modulation of the phosphorylation state of PAP is mitosis promoting factor (MPF). This is suggested by the several consensus sequences contained within the C-terminal region of PAP that are specific for MPF binding. In addition, MPF has been shown to phosphorylate PAP both *in vivo* and *in vitro* (*14*). The processing complex can

additionally be modulated by the acetylation of lysines 635/644/730/734 of PAP and lysine 23 of CF $I_m$25, by the Creb Binding Protein (CBP), as shown by immunoprecipitation experiments with an acetylated-lysine antibody and subsequent identification by mass spectrometry (*46*). *In vivo*, however, acetylation of lysines 635/644/730/734 of PAP and lysine 23 of CF $I_m$25 has been shown to disrupt the interaction between CF $I_m$25 and PAP in COS-7 cells treated under acetylation conditions and cotransfected with CF $I_m$25, CF $I_m$68, and PAP (*46*). Interestingly, the carboxy terminal region of PAP is not crucial for maintaining CF $I_m$25 interaction *in vitro*, as demonstrated by pull-down experiments with recombinant CF $I_m$25 and a carboxy terminal truncated PAP **(Figure 8)** (*41*). This finding suggests a transient interaction could exist between PAP and CF $I_m$ 25, which does not require the C-terminal domain and remains an area that needs to be further investigated. In addition, an indirect or direct interaction between PAP and CF $I_m$ occurs during processing, via CF $I_m$25 and other processing components. This interaction stabilizes the processing complex whereby acetylation may play a role in the regulation of the processing complex through the disruption of CF $I_m$ and PAP.

*In vitro*, the polymerase activity of PAP increases 30 -100 fold with the addition of manganese in comparison to magnesium. Manganese functions to stimulate poly(A) addition by increasing the binding of PAP to 3'-ends of the RNA (*47, 48*). The polymerization event *in vivo* involves the binding of two magnesium ions to the conserved aspartic acid residues within the active site of PAP. Metal binding allows coordination of the incoming ATP and stabilization of the transition state following AMP

addition. The crystal structure of bovine PAP provides insight into the active site mechanism of polyadenylation and the polymerases ability to discriminate against other nucleotides to only incorporate adenosine into a growing adenylate tail (*48*). A carboxy terminal truncation mutant of PAP (PAP513) bound to cordycepin triphosphate was crystallized lacking the C-terminal 226 residues **(Figure 9)**. Importantly, this mutated construct can still be stimulated by CPSF and PABPN1 and additionally PAP513 can interact only in the presence of oligo(A) with PABPN1 (*49, 50*). PAP belongs to the superfamily of nucleotidyl transferases and as such contributes three conserved aspartic acid residues and a helical turn motif that have been shown to be important for phosphoryl transfer amongst family members. The three aspartic acid residues and a helix-turn motif are involved in coordination of the bound nucleotide. Selectivity for ATP over competing nucleotides is via N7 contacts to one of the catalytic aspartic acids and stabilization of the adenine ring via hydrophobic interactions with two neighboring valine residues (*50*). The binding of PAP to RNA is relatively weak and nonspecific and only has maximum processivity of polymerization in conjunction with PABPN1 and CPSF. PABPN1 is also involved in tail length control and restricts the polyadenylated tail length to approximately 250 nucleotides in higher eukaryotes whereby, upon completion the mRNA can be efficiently transported out of the nucleus into the cytoplasm (*13, 14*).

### CF $I_m$

The coupling of preceding steps, such as transcription, capping and splicing in pre-mRNA 3'-end processing is important for the efficiency of both cleavage and

polyadenylation. This relies on the crosstalk of multiple protein-protein assemblages that initially form during the cleavage step and rearrange subsequently to function in polyadenylation. Much of what we know about complex formation of processing factors on the pre-mRNA and its regulation has been investigated biochemically. A structural characterization of individual and multicomponent processing factors will elucidate the domain interactions important for the pre-mRNA 3'-end processing mechanism that cannot be deciphered by biochemical means. We chose the structural investigation of the 25 kDa subunit of CF $I_m$ (CF $I_m$25) because of its role in cleavage and stimulatory role in polyadenylation. The 25 kDa subunit of CF $I_m$ interacts with multiple players of the cleavage and polyadenylation machinery, directly or as a complex with its larger subunit **(Figure 10)**. In addition, CF $I_m$25 plays an important role in promoting complex recognition and regulation of poly(A) site selection.

As mentioned previously, CF $I_m$ is an oligomer composed of a small 25 kDa subunit (CF $I_m$25, also referred to as CPSF5 or NUDT21) and a larger subunit of either 59 kDa, 68 kDa, or 72 kDa (*40*). Homologs of CF $I_m$ subunits are also found in cDNA databases of *Caenorabditis elegans*, *Drosophila melanogaster*, and plants. The three larger subunits share substantial sequence homology. They are encoded on two different genes and the 72 kDa subunit is a splice variant of the 68 kDa polypeptide. CF $I_m$ can be reconstituted *in vitro* from the 25- and 68 kDa subunits (*40, 51*). The 68 kDa subunit has a domain architecture similar to splicesomal SR-proteins. The 68 kDa subunit contains an amino terminal RNP domain, central proline rich region, and a carboxyl terminal arginine and serine rich region (RS) **(Figure 11)**. The interaction of CF $I_m$ 68 with

components of the human splicesome suggests a link between 3' pre-mRNA processing and splicing (*52*). It has been demonstrated that the 25 kDa, 59 kDa, and 68 kDa, subunits individually bind RNA in the absence of other processing factors as demonstrated by UV crosslinking (*40*). More recently, and as mentioned previously, assays involving the 68 kDa and 25 kDa subunits suggest both are necessary for efficient RNA binding. RNA binding is facilitated primarily by the amino terminal region, amino acids 1-160 of CF $I_m$ 25, and the RS region of CF $I_m$ 68, as demonstrated with pull-down assays using truncation mutants of both proteins (*41*). Both the cleavage and polyadenylation steps involve a CF $I_m$ complex rather than individual components of CF $I_m$, which suggests a cooperative RNA binding function. In addition to characterizing the RNA binding potential of CF $I_m$, Dettwiler et al. 2004 demonstrated the portions of CF $I_m$25 that are sufficient to interact with other processing factors. The entire CF $I_m$ 25 kDa subunit is necessary for the interaction with CF $I_m$ 68 kDa subunit and likely occurs through the Nudix region (**nu**cleoside **di**phosphate linked to another moiety-**X)** of CF $I_m$ 25 based on the charge distribution. The interaction of the 25 kDa subunit with the 68 kDa subunit relies on the RRM region of the 68 kDa subunit. Recently, CF $I_m$, has also been implicated in a role in the regulation of poly(A) site recognition by preferential binding to UGUAN sequences contained within the poly(A) site region (*28*). The binding of CF $I_m$ to UGUAN containing sequence elements around the poly(A) site suggests an additional mechanism to select for the noncanonical hexamer sequences and alternative poly(A) sites that exist. SELEX analysis identified short UGUAN containing RNAs with high affinity for recombinant CF $I_m$. The addition of these UGUAN

containing RNAs to cleavage and polyadenylation reactions inhibited the poly (A) site cleavage by sequestering CF $I_m$ in the nuclear extract (*27*). CF $I_m$ is also involved in stabilizing CPSF at the conserved hexamer through interactions with the 160 kDa subunit (*40*). *In vitro*, binding assays with recombinant proteins suggest only a weak interaction between CF $I_m$25 and CPSF160 (*41*). This interaction may require the 68 kDa subunit or be an indirect interaction and require other components for stabilizing the 160 kDa subunit of CPSF. *In vitro*, CF $I_m$ enhances the rate of poly (A) site cleavage, and has been shown to interact with PAP and PABPN1 via its 25 kDa subunit (*41*). Furthermore CF $I_m$ also interacts with splicing factors, indicating a role in communicating between different RNA processing complexes (*12*).

As mentioned earlier CF $I_m$ facilitates multiple interactions among mammalian mRNA processing members, not only in cleavage and polyadenylation, but also in splicing. Despite its apparent role in the mammalian system and similarities to the yeast processing machinery, there is no homolog of CF $I_m$ subunits in yeast. However the strong sequence conservation in higher eukaryotes of the 25 kDa CF $I_m$ subunit suggests CF $I_m$25 may have significant contributions outside of pre-mRNA 3'-end processing. In accordance with this, CF $I_m$25 was confirmed by MALDI mass spectrometry to be associated as a component of a large RNA complex with Rae1, an mRNA export protein. This RNA dependent complex was necessary for mitotic spindle assembly and links CF $I_m$25 with the RNA export machinery (*53*). This suggests CF $I_m$25 may potentiate other RNA export processing steps that have yet to be discovered and could be facilitated by the same interaction surfaces necessary for pre-mRNA 3'-end processing.

## The Nudix domain of CF $I_m$25

Biochemical data suggest the 25 kDa subunit of CF $I_m$ facilitates many of the key interactions between CF $I_m$ and other processing factors through interactions via the Nudix Domain of CF $I_m$25. The structural characterization of CF $I_m$25 will allow the investigation of the function of a Nudix domain present in this protein and validate the biochemical interactions previously described.

The first Nudix protein to be characterized enzymatically and structurally was *Escherichia coli* MutT (*54*). Nudix proteins are generally characterized as housekeeping enzymes due to their role in hydrolysis of substrates described as **nu**cleoside **di**phosphate linked to another moiety **X,** many of which are potentially toxic molecules (*55, 56*). Nudix proteins have a conserved Nudix fold consisting of an α/β/α sandwich **(Figure 12)**. This fold is also common to isopentenyl diphosphate isomerases and the carboxyl terminus of Mut-Y type DNA glycosylases (*55*). Within the Nudix fold the consensus sequence of the Nudix box is $GX_5EX_7REUXEEXGU$, where U is a hydrophobic residue and X is any residue, and folds into a loop - α helix - loop structure (*54*). The glutamate residues found within this Nudix box are important for divalent metal binding and catalysis. The most common bound divalent metal of physiological relevance is magnesium (*56*). Members of the Nudix family vary in their substrate specificity, position of residues involved in substrate coordination, and the number of metals required for catalysis. In the case of MutT, two divalent ions are necessary for catalysis: one ion

is involved in coordination of the substrate to the β- and γ- phosphate, while the second divalent metal is coordinated by the enzyme (*55*).

Nudix proteins can recognize and hydrolyze a large range of substrates such as nucleotide sugars, oxidized and nonmutagenic nucleoside di- and triphosphates, RNA caps, and dinucleoside polyphosphates. Residues and motifs outside of the Nudix box, which together forms part of the α/β/α sandwich of the Nudix domain, determine the substrate specificity of each Nudix member. Notable differences in substrate binding exist not only with the recognized substrate, but also with the positioning of the nucleotide's base with in the active site (*57*). The catalytic base is generally a glutamate residue that can be found within or outside of the catalytic sequence and orchestrates hydrolysis by substitution at phosphorous **(Figure 13)** (*55*).

Nudix proteins are found in a wide range of organisms and likewise have a wide spectrum of substrate specificities. This enables them to play a versatile role in protection, regulation, and signaling. In bacteria, the number of Nudix proteins correlates linearly with the size of the genome as well as the environmental growth conditions. Bacteria with a considerably larger genome have more Nudix proteins to promote genome conservation and allow for the break down of metabolite intermediates found in harsher environmental conditions (*56*). This is not consistent in higher eukaryotes where the genome contains a large amount of non-coding DNA and would result in an overestimation of Nudix enzymes. To date, the human genome alone encompasses 24 Nudix enzymes and 5 pseudogenes **(Table 2)**. Most of the human Nudix enzymes are

found associated with a housekeeping function involving sanitizing the toxic nucleotide pool with the exception of Dcp2, X29, and CF $I_m$25.

The proteins Dcp2, X29, and CF $I_m$25, all overlap in their ability to bind RNA and participate in RNA metabolism. Dcp2 functions in the Dcp1/Dcp2 decapping mRNA complex to generate m7GDP and 5'phosphorylated mRNA. This is an essential function for 3' to 5' and 5' to 3' mRNA decay and specialized decay pathways such as nonsense mediated decay (*58*). X29 is involved in decapping the U8 snoRNA releasing m7GDP (*56*). CF $I_m$ 25, unlike both Dcp2 and X29, has no apparent role in mRNA decapping, but has a function in binding mRNA to facilitate in 3' pre-mRNA processing events as mentioned. In addition, it is the only Nudix protein that is associated as a complex with its counterpart the 59 kDa, 68 kDa, or 72 kDa larger subunit for function in cleavage and polyadenylation. Interestingly, the Nudix box in CF $I_m$25 lacks two of the four conserved glutamate residues, three of which were shown to be important for catalysis (*59*). Recently, a GDP-mannose hydrolase found in enterobacteria and *Vibrio* species has been described and similarly lacks two of the conserved glutamate residues of the catalytic box. This substitution imparts an alternative hydrolase mechanism whereby nucleophilic substitution occurs at a carbon instead of a phosphorous **(Figure 14)** (*60, 61*). This makes it plausible that CF $I_m$ 25 may retain hydrolase activity, albeit with an alternative substitution mechanism.

## Specific Aims

This project is aimed at elucidating the crystallographic structures of factors important for pre-mRNA 3'-end processing. Currently, there are a limited number of crystallographic structures of pre-mRNA 3'-end processing factors. This ensemble of structures deposited in the Protein Data Bank include several partial structures and very few complex structures of processing factors. Crystallographic techniques will allow a validation of regions important for protein-protein and protein-RNA interactions that have previously been reported biochemically. In addition, structural data will provide insight into new functionally and/or mechanistically important protein residues essential for pre-mRNA 3'-end processing.

The specific processing factor that is the focus of this project is the 25 kDa subunit of CF $I_m$. CF $I_m$25 is involved in early steps of 3'-end processing and contributes to both protein-protein and protein-RNA interactions. CF $I_m$ 25's involvement with multiple factors and with the RNA allows for several candidates for a complex crystal structure. Interestingly, many of the biochemical interactions with CF $I_m$ 25 are contradictory to each other which made it important to obtain the crystal structures of the protein alone and in complex with a ligand or protein binding partner to validate the available biochemical data.

Aim 1 is to crystallize the structure of the apo human CF $I_m$25 protein. The crystal structure of CF $I_m$25 is expected to provide insight into the function of the Nudix domain of CF $I_m$25 and its potential role in 3'-end processing. The CF $I_m$25 structure is predicted to harbor the classic Nudix $\alpha/\beta/\alpha$ fold but lacks two of the four glutamic acid

residues important for catalytic activity and metal binding. The structural investigation will help further define the domains important for substrate and protein interactions and elucidate how the structural organization may allow for its versatility in multiple processing steps. The function and activity of the Nudix domain can be addressed additionally by biochemical assays. Biochemical assays will be used to determine if CF $I_m$25 can bind or hydrolyze a nucleotide moiety via its Nudix domain. The outcome of the previous biochemical data will determine if a nucleotide bound structure is worth pursuing.

Aim 2 is to determine a complex structure of the 25 kDa subunit of CF $I_m$ (CF $I_m$25) with a binding partner relevant to 3' pre-mRNA processing. Possible binding partners of interest are the 68 kDa subunit of CF $I_m$, PAP, and Clp1. Obtaining a complex of CF $I_m$ 25 and its counterpart the 68 kDa subunit will address and validate which regions of both proteins are necessary to mediate complex formation. Pull-down assays have previously demonstrated that the RRM of the 68 kDa subunit is sufficient to bind CF $I_m$ 25. Previous data from the literature describing the interaction between PAP and CF $I_m$ 25 are more controversial. The interaction between PAP and CF $I_m$ 25 has been proposed to involve the Nudix region of CF $I_m$25 and a C-terminal truncated PAP. Additionally, data have been presented that detail this interaction involving CF $I_m$ 25 and the C-terminal region of PAP. Recently, a mechanism has also been described to disrupt the polyadenylation machinery involving the acetylation at the carboxy terminus of PAP and amino terminus of CF $I_m$ 25. A complex structure of PAP and CF $I_m$ 25 will be necessary to investigate the conflicting biochemical data presented so far. Lastly, a

structure of the complex between CF $I_m$ 25 and hClp1 will be pursued. Clp1p is essential for yeast pre-mRNA 3'-end processing. In humans however, hClp1 plays an important role in tRNA splicing and as a siRNA kinase but interestingly is not able to complement a Clp1p knockout in yeast (*62*). This raises questions about hClp1 importance to 3'-end processing which can be addressed by capturing a complex structure containing CF $I_m$25 and hClp1.

Aim 3 is to determine the crystal structure of hFip1. Recently, hFip1 has been included as a subunit of CPSF although the stoichiometry of Fip1 to the remaining four subunits appears to be less than a 1:1 ratio. Despite its transient association with the other substituents of CPSF, Fip1 is important for the polyadenylation step of pre-mRNA 3'-end processing to stimulate PAP. In addition, aim 3 also involves determining a complex structure of Fip1 with PAP.

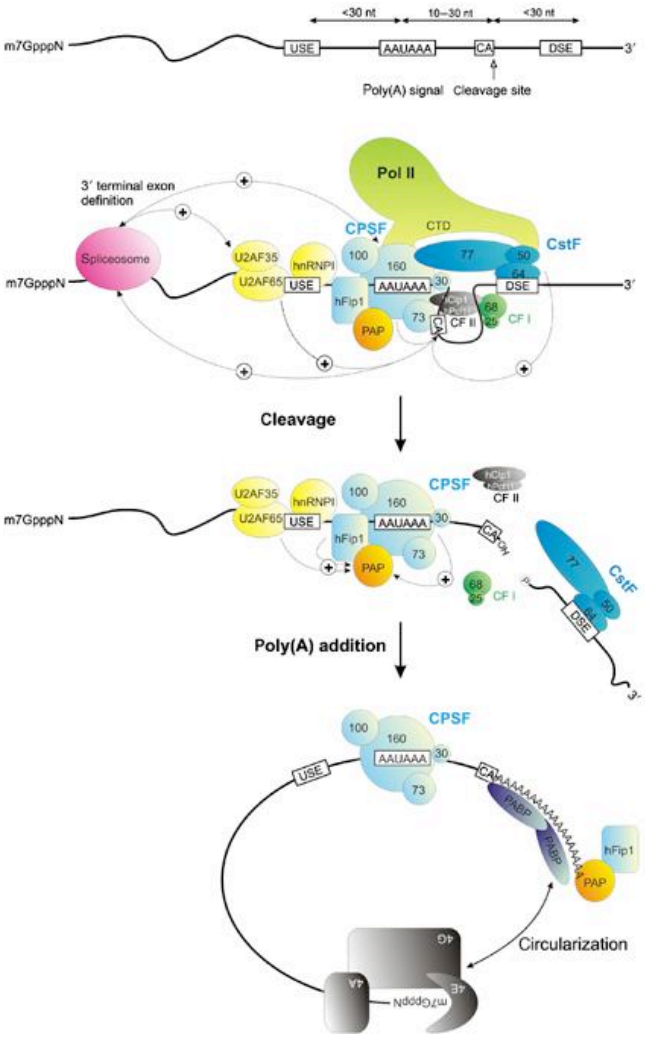



**Figure 1: Eukaryotic pre-mRNA 3'-end Processing**

Eukaryotic pre-mRNA 3'-end processing events include capping at the 5'-end, splicesomal cleavage of intronic mRNA, pre-mRNA 3'-end cleavage followed by polyadenylation. These processing events are orchestrated by the recognition of cis-acting sequence elements to recruit processing factors to the site of processing. The removal of the intronic message allows cleavage factors to be recruited to the mRNA. Cleavage occurs at a CA dinucleotide as shown and this presents the 3' end of the pre-mRNA for PAP binding and PABPN1 stabilization for polyadenylation. Adapted from Danckwardt et al. 2008 (*63*).

| Factor | Subunit(kDa) | Gene | Mammalian Homologue[a] |
|---|---|---|---|
| CF IA | 76 | RNA14 | CsfF-77 |
| | 72 | PCF11 | - |
| | 50 | CLP1 | - |
| | 38 | RNA15 | CstF-64 |
| CF IB | 73 | HRP1 | - |
| CF IIy | 150 | YHH1/CFT1 | CPSF-160 |
| | 105 | YDH1 | CPSF-100 |
| | 100 | YSH1/BRR5 | CPSF-73 |
| | 90 | unknown | - |
| PF 1 | 150 | YHH1/CFT1 | CPSF-160 |
| | 105 | YDH1 | CPSF-100 |
| | 100 | YSH1/BRR5 | CPSF-73 |
| | 85 | PTA1 | - |
| | 64 | PAP1 | poly(A) polymerase |
| | 58 | PFS1 | - |
| | 55 | FIP1 | hFIP1 |
| | 53 | PFS2 | - |
| | 26 | YTH1 | CPSF-30 |

**Table 1: Yeast pre-mRNA 3'-end Processing components and mammalian homologues**

Yeast pre-mRNA 3'-end processing complexes and their individual substituents are listed, along with their corresponding mammalian homologues. [a] Mammalian homologues may exist for additional yeast processing factors but have not been linked to mammalian processing reactions. This figure was adapted from Wahle et al. 1999 (*14*).



**Figure 2: Crystal structure of yeast Clp1/Pcf11 bound to ATP**

Crystal Structure of yeast Clp1 with Pcf11 fragment (residues 454-563) bound to ATP (PDB 2NP1). The Pcf11 fragment is represented in yellow and the Clp1 protein is colored according to domain architecture. The N-terminal region of Clp1 is represented in red and sequesters the ATP within the Central region, colored green. The Central domain harbors the nucleotide binding motif or Walker A motif. Lastly, the C-terminal region is represented in blue and shows no structural similarity amongst PDB entries (*20*).

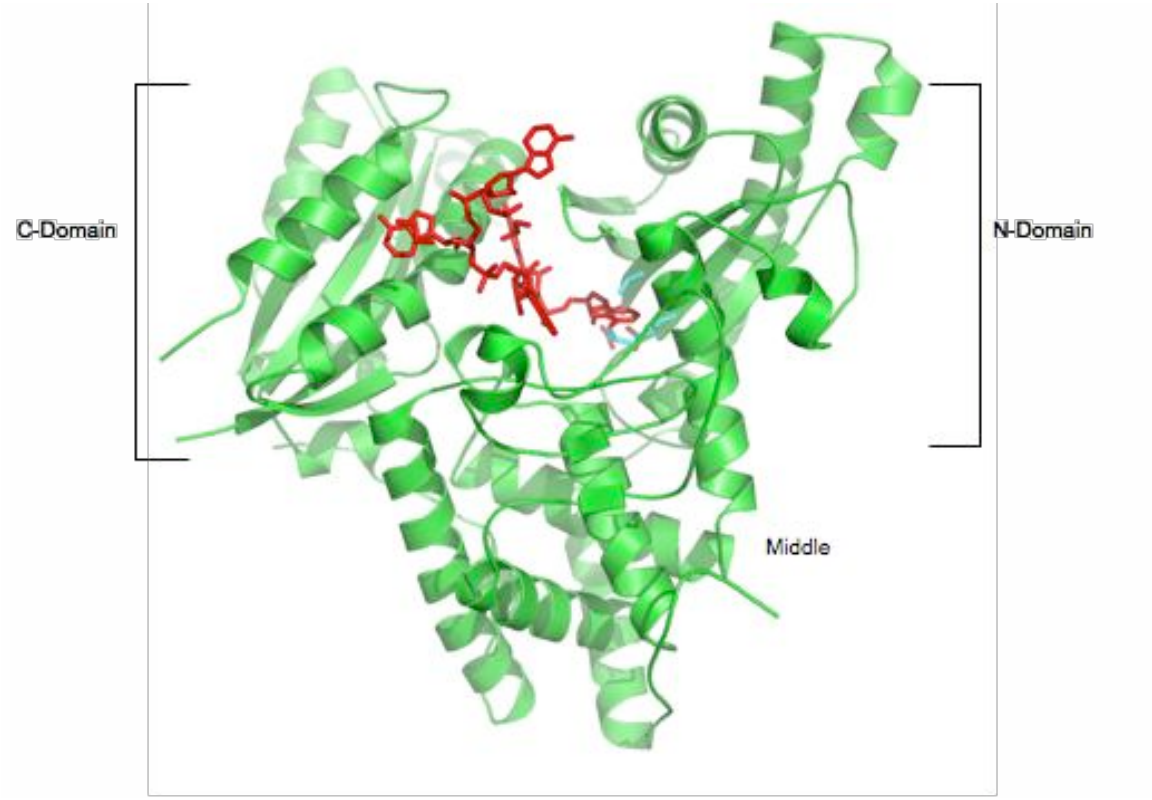



**Figure 3: Crystal structure of Pap1 bound to oligo (A)**

The yeast Pap1 crystal structure is presented in green bound with a 5mer oligo(A) (PDB 2Q66). In order to capture the RNA bound Pap1 structure aspartate 154 (D154), one of the catalytic residues, was mutated to an alanine. The remaining catalytic aspartate residues, D100 and D102, are colored in cyan in addition to the D154A mutation. The MgATP is not represented for the sake of clarity (*25*).

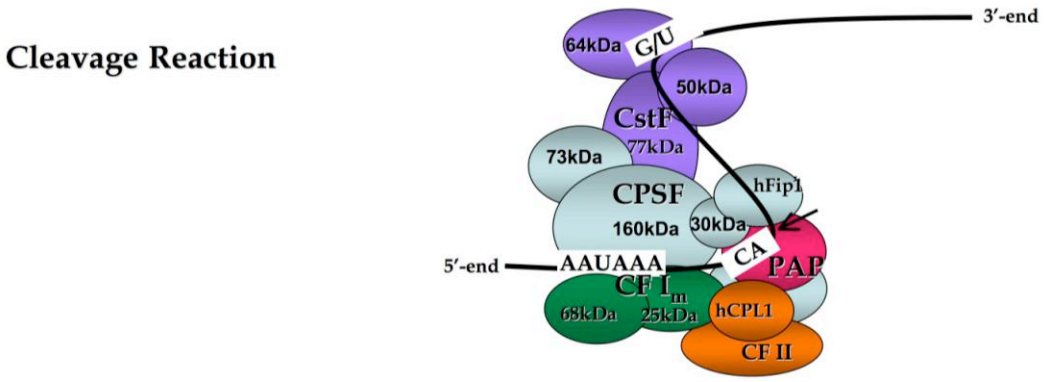



**Figure 4: Eukaryotic cleavage components**

The core cleavage complex involves the factors depicted above. The recognition of the conserved hexamer by CPSF160 and additional binding of CstF64 at the downstream G/U rich element initiates cleavage at the 3'-end. Further stabilization at the conserved hexamer is facilitated by CF $I_m$ interactions with CPSF160 and in conjunction with Fip1, CF $I_m$, recruits PAP to the cleavage complex. Following cleavage complex assembly endonucleolytic cleavage occurs at the CA dinucleotide.

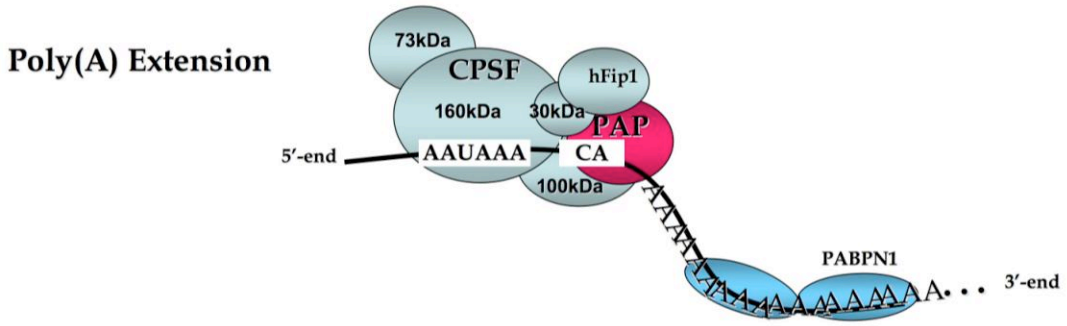



**Figure 5: Poly(A) tail elongation machinery**

Poly(A) tail elongation is facilitated by the factors CPSF, PAP, and PABPN1. PABPN1 is necessary for PAP processivity during tail elongation.

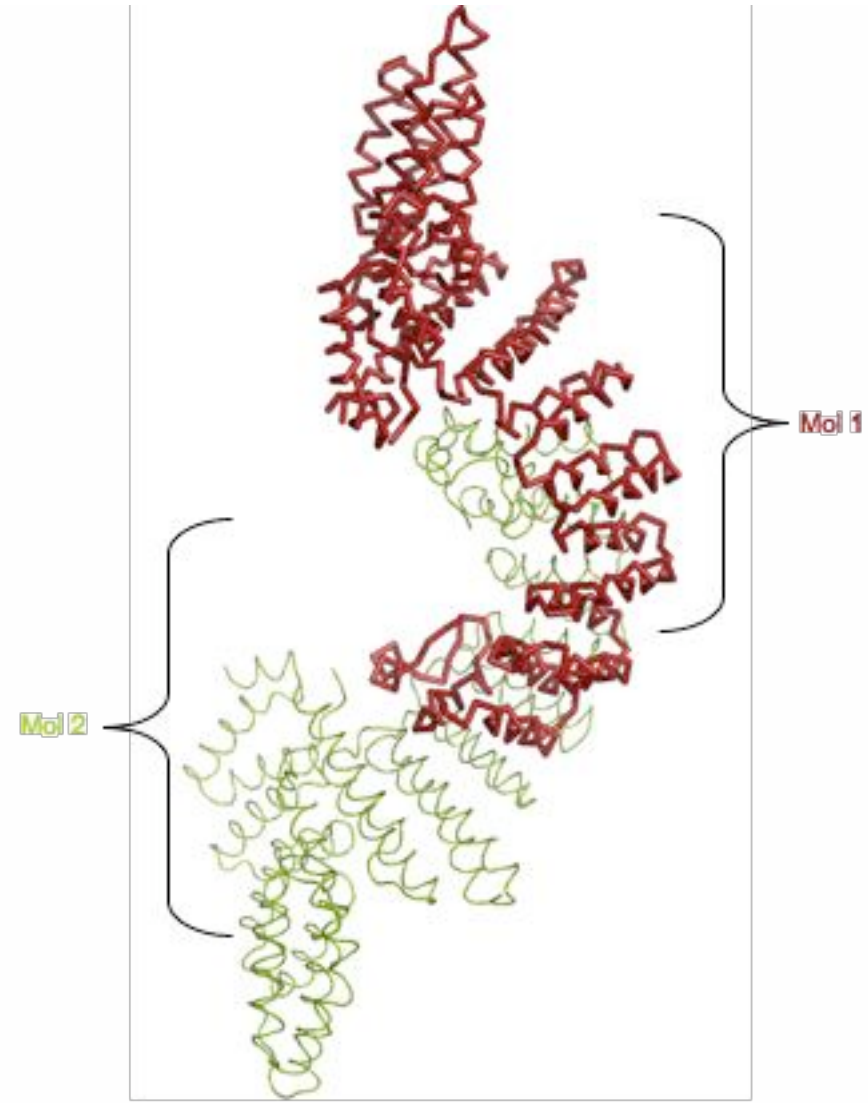



**Figure 6: HAT domain of murine CstF77**

Ribbon diagram of CstF77 entire HAT domain shown as a dimer and colored according to molecule number, molecule 1 (Mol1) is red and molecule 2 (Mol2) is yellow/green. CstF77 is a dimer in the crystal structure as well as in solution and dimer formation is mediated through the C-terminal HAT region (PDB 200E). CstF77 HAT domain is believed to be involved in coordination of protein-protein interactions (*35*).

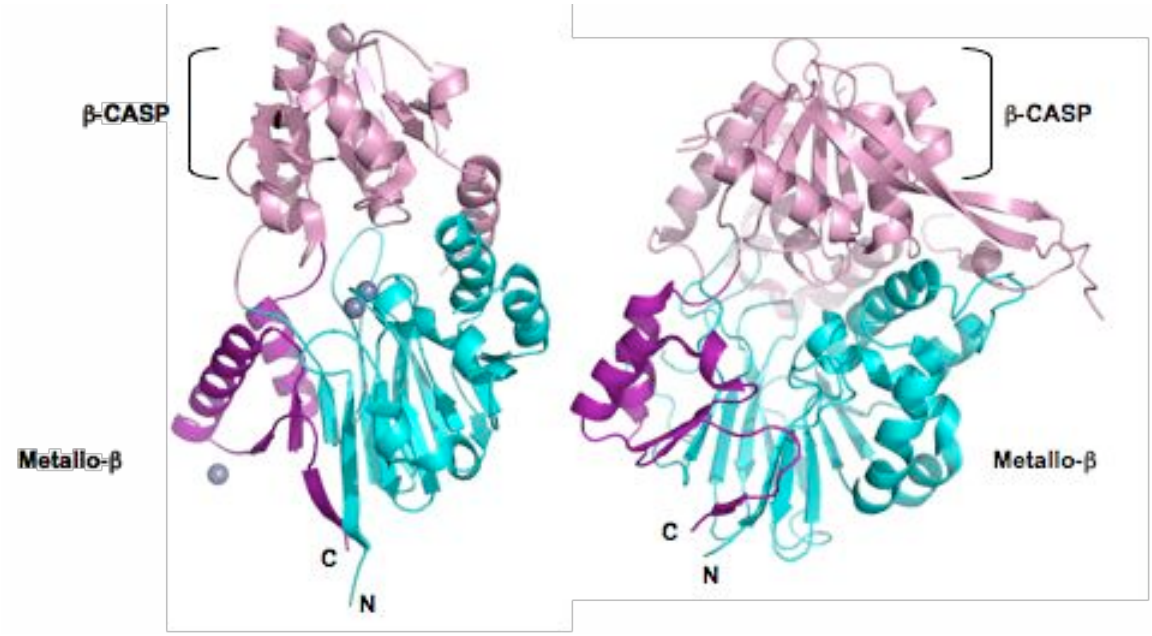



**Figure 7: Structural similarities between CPSF73 and CPSF100**

Crystal structures of human CPSF73 and CPSF100 share common folds, both belonging to the metallo-β-lactamase superfamily and additionally composed of a β-CASP domain (PDB 2I7V and 2I7X). The metallo-β-lactamase domain is represented in cyan and purple (Metallo-β), the β-CASP is shown in pink, and grey spheres represent zinc atoms. Additional helices are apparent in the Metallo-β domain colored in cyan and are thought to be specific to RNA/DNA processing nucleases (*42*).

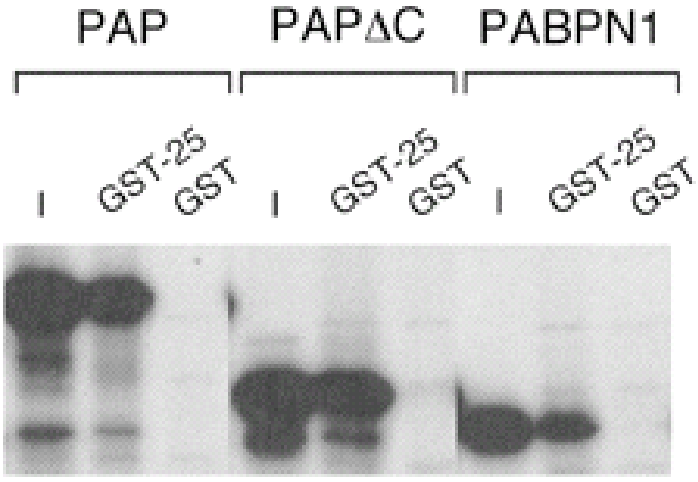



**Figure 8: Interaction of the 25-kDa subunit of CF $I_m$ with PAP513**

GST-CF $I_m$25 is able to interact directly with *in vitro* translated [$^{35}$S]-methionine labeled full length PAP, C-terminally truncated PAP (PAP513 or PAPΔC) and PABPN1 via pull-down experiments (*41*). This evidence contradicts previous data demonstrating the C-terminal region of PAP is necessary to mediate interaction with CF $I_m$25 (*64*). This figure was adapted from Dettwiler et al. 2004.

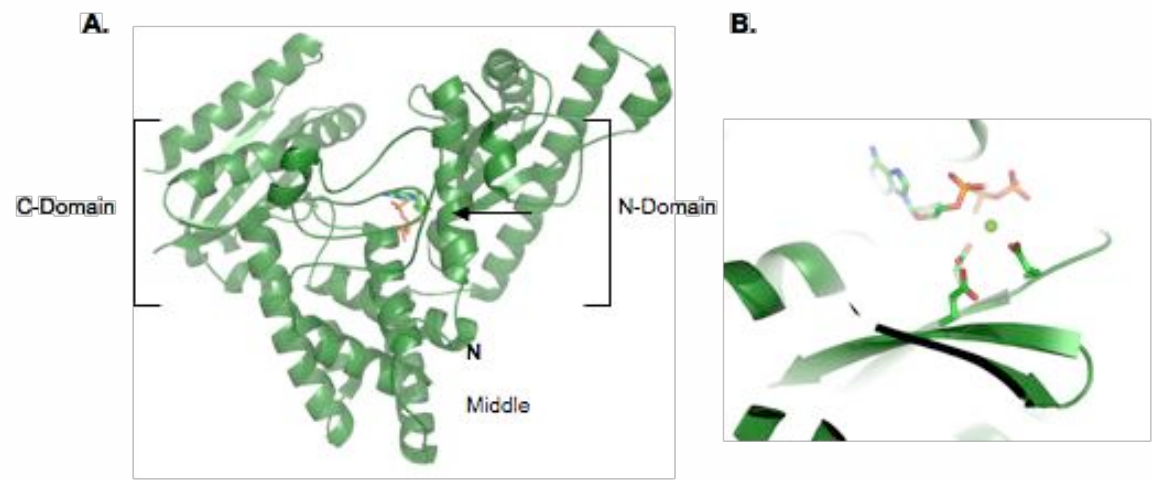



**Figure 9: Crystal structure of bovine PAP513 in complex with 3'dATP**

A. Representation of the C, Middle, and N Domains as previously shown in the yeast Pap1 structure. The arrow indicates the position of 3'dATP (cordycepin). B. The catalytic triad consisting of aspartic acid 113, 115, and 167 are involved in coordination of the bound magnesium and 3'dATP (*48*).

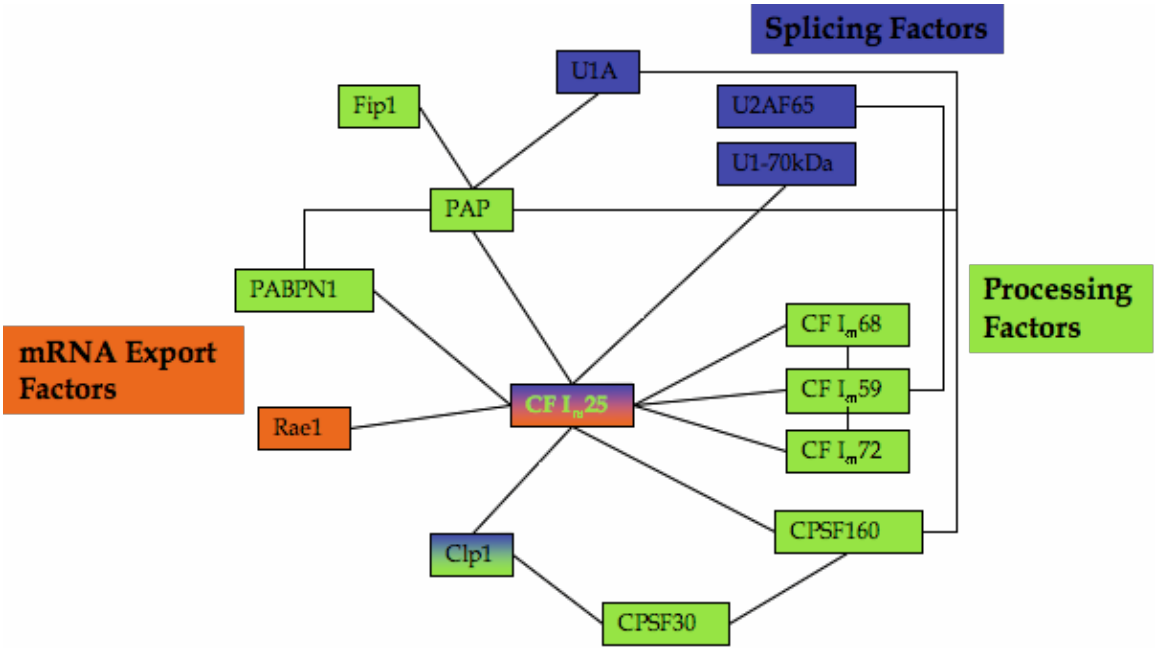



**Figure 10: Scheme of the direct interactions among processing factors**

Multiple interactions among polyadenylation pre-mRNA 3'-end processing factors are coupled to splicing and mRNA export (*12, 41, 44, 64-66*).

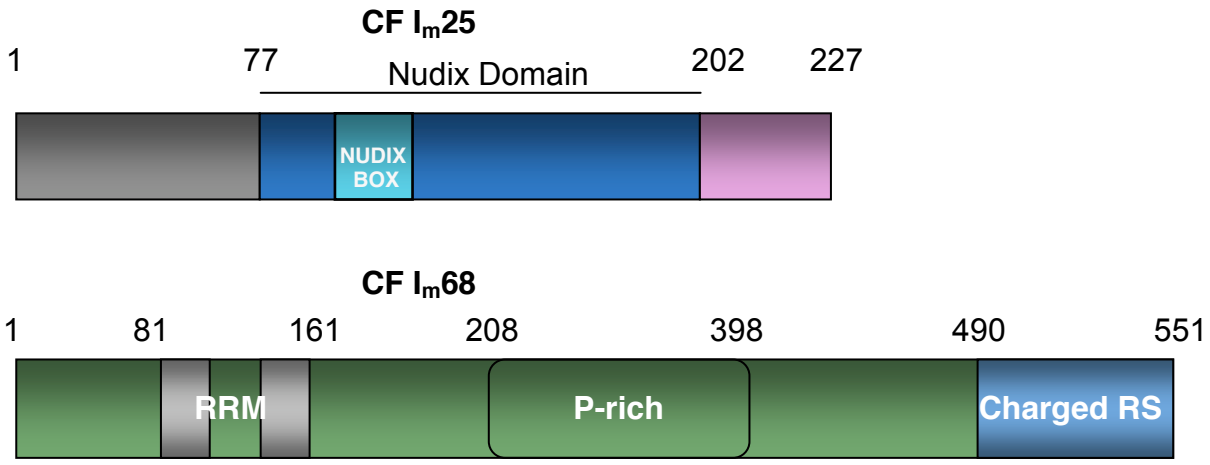



**Figure 11: Domain architecture of the CF $I_m$ subunits**

CF $I_m$25 is composed of a N-terminal domain involved in RNA binding interactions, a nucleoside diphosphate linked to another moiety X domain (Nudix) located centrally, and a C-terminal domain. The 68 kDa subunit has an N-terminal RNA recognition motif (RRM) involved in protein-protein interactions, central proline-rich (P-rich) region, and a C-terminal charged arginine and serine domain reminiscent of SR proteins.

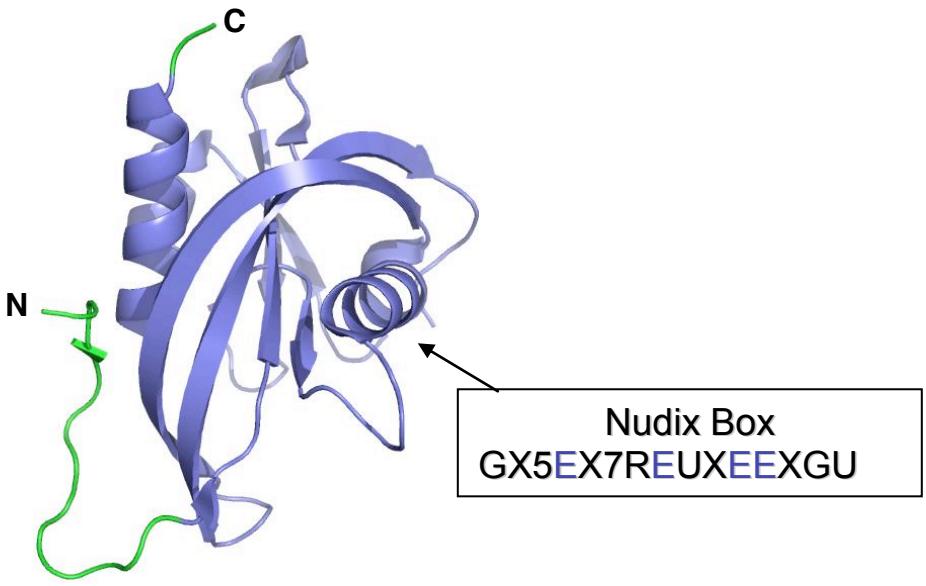



**Figure 12: Signature fold of Nudix proteins**

MutT was the first Nudix protein to be characterized and demonstrates the classic α/β/α sandwich of the Nudix fold (2AZW). Within the Nudix fold is the conserved Nudix box and has the representative sequence. The arrow depicts the position of the Nudix box (*67*).

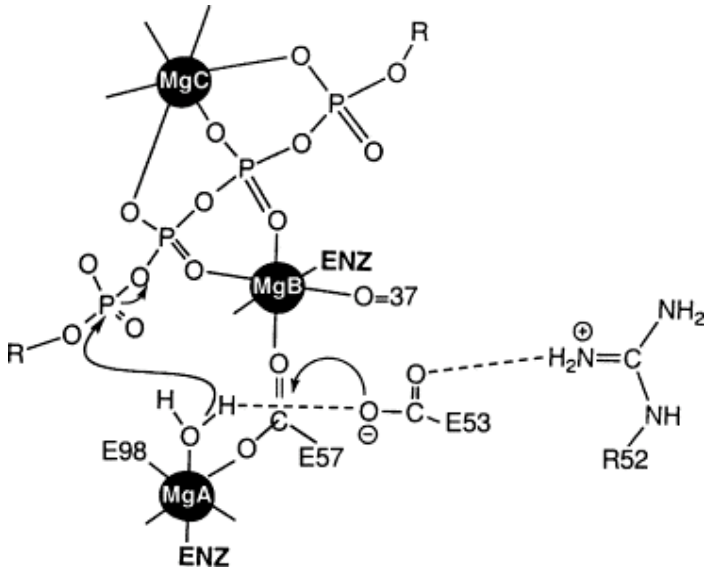



**Figure 13: Mechanism of substrate hydrolysis by $Ap_4A$ pyrophosphatase**

Mechanism of substrate hydrolysis with nucleophilic substitution occurring at a phosphorous by the Ap4A pyrophosphatase enzyme from *B. bacilliformus*. Catalysis of the substrate, diadenosine tetraphosphate (Ap4A), requires three divalent metals, two of which are bound by the enzyme (ENZ). This figure was adapted from Xia et al. 2005 (*55*).

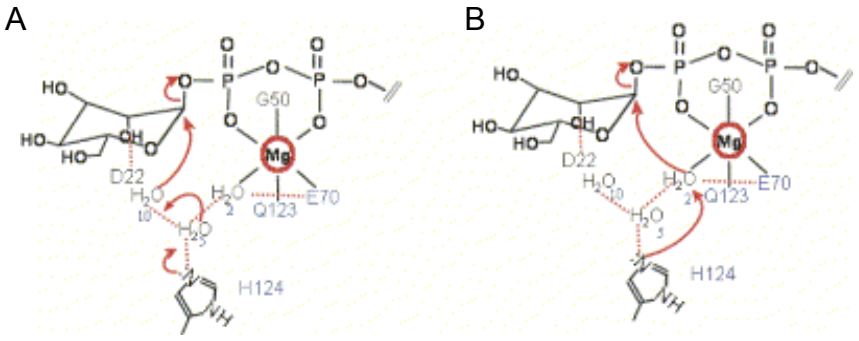



**Figure 14: Alternative mechanism of GDP-mannose hydrolase**

(A and B) Schematic diagram of two possible mechanisms of nucleophilic substitution occurring at a carbon instead of a phosphorous by the *E. coli* enzyme GDP-mannose hydrolase. Both mechanisms are based on structural and biochemical data. This figure was adapted from Bianchet et al. 2004 (*68*).

| Name | Synonyms | Known Substrates |
|---|---|---|
| NUDT1 | MTH1 | 8-oxo-(d)GTP,8-OH-(d)ATP,2-OH-(d)ATP |
| NUDT2 | APAH1 | NpnN (n>4) e.g Ap4A,Gp4G,Ap5A,Ap6A |
| NUDT3 | DIPP1 | PP-InsP5,[PP]2-InsP4,Ap6A,Ap5A |
| NUDT4 | DIPP2α, DIPP2β | PP-InsP5,[PP]2-InsP4,Ap6A,Ap5A |
| NUDT4P1 | | unknown |
| NUDT4P2 | | unknown |
| NUDT5 | ADPRibase-II | ADP-sugars,8-oxo-(d)GDP |
| NUDT6 | GFG | (ADP-ribose)4 |
| NUDT7 | | CoA,CoA esters,CoASSCoA,3'-dephospho-CoA |
| NUDT8 | | (CoA and derivatives) |
| NUDT9 | ADPRibase-I, ADPRibase-m | ADP-ribose,IDP-ribose |
| NUDT9L1 | TRPM2 | ADP-ribose |
| NUDT9P1 | C10orf98 | unknown |
| NUDT10 | DIPP3α | PP-InsP5,[PP]2-InsP4,Ap6A,Ap5A |
| NUDT11 | DIPP3β | PP-InsP5,[PP]2-InsP4,Ap6A,Ap5A |
| NUDT12 | | NAD(P)H,NAD(P)+,Ap2A,FAD,ADP-ribose,GDP-glucose |
| NUDT13 | | NAD(P)H,Ap2A |
| NUDT14 | UGPP | UDP-glucose,ADP-ribose,ADP-glucose,GDP-glu |
| NUDT15 | MTH2 | 8-OH-(d)ATP,dGTP,(DHNTP) |
| NUDT15P1 | | unknown |
| NUDT15P2 | | unknown |
| NUDT16 | hX29 | capped snRNAs |
| NUDT16L1 | syndesmos | unknown |
| NUDT17 | | unknown |
| NUDT18 | | unknown |
| NUDT19 | hRP2 | (CoA and derivatives) |
| NUDT20 | hDcp2 | capped mRNAs |
| NUDT21 | CFIm25 | Ap4A, ATP |
| NUDT22 | | unknown |

**Table 2: Human Nudix genes and hydrolases**

Summary of the human Nudix genes and hydrolases, including known pseudogenes. The characterized hydrolases listed are presented with their recognized substrates. This figure was adapted from McLennan 2006 (*56*).

# CHAPTER 2:

# CONTRIBUTIONS

My contributions to the production of this paper include: protein expression and purification, crystallization experiments, determination of the oligomeric state, structure solution, steady state fluorescence experiments, and contribution to the writing of the manuscript. Dr. Mark Rould contributed to troubleshooting the fluorescence experiments and helpful suggestions for the refinement of $Ap_4A$. Additionally, Dr. Greg Gilmartin contributed to helpful discussion on the nucleotide binding properties of CF $I_m25$ and suggested further experiments to determine how Ap4A binding may play a role in 3'-end processing.

**Crystal Structure of the 25 kDa Subunit of Human Cleavage Factor $I_m$**

Molly Coseno, Georges Martin[1], Christopher Berger[2], Gregory Gilmartin, Walter Keller[1], Sylvie Doublié*

Departments of Microbiology and Molecular Genetics and [2]Molecular Physiology and Biophysics, University of Vermont, Burlington, Vermont 05405, USA

[1] Department of Cell Biology, Biozentrum, University of Basel, CH-4056 Basel, Switzerland

*** Corresponding author**:

Sylvie Doublié, Dept. of Microbiology and Molecular Genetics, The Markey Center for Molecular Genetics, University of Vermont, Stafford Hall, 95 Carrigan Dr., Burlington, VT 05405-0068.Tel.: 802-656-9531; Fax: 802-656-8749; E-mail: sdoublie@uvm.edu.

**Running title**: Crystal structure of human CF $I_m$25

**Keywords**: RNA processing, CPSF5, cleavage factor, X-ray crystallography, diadenosine tetraphosphate

Coseno, M., Martin, G., Berger, C., Gilmartin, G., Keller, W., and Doublié, S. (2008) **Crystal Structure of the 25kDa Subunit of Human Cleavage Factor $I_m$**. *Nucleic Acids Research* 36 (10), 3474-3483.

## ABSTRACT

Cleavage Factor $I_m$ is an essential component of the pre-messenger RNA 3' end processing machinery in higher eukaryotes, participating in both the polyadenylation and cleavage steps. Cleavage factor $I_m$ is an oligomer composed of a small 25 kDa subunit (CF $I_m$25) and a variable larger subunit of either 59 kDa, 68 kDa, or 72 kDa. The small subunit also interacts with RNA, poly(A) polymerase, and the nuclear poly(A) binding protein. These protein-protein interactions are thought to be facilitated by the Nudix domain of CF $I_m$25, a hydrolase motif with a characteristic $\alpha/\beta/\alpha$ fold and a conserved catalytic sequence or Nudix box. We present here the crystal structures of human CF $I_m$25 in its free and diadenosine tetraphosphate ($Ap_4A$) bound forms at 1.85Å and 1.80Å, respectively. CF $I_m$25 crystallizes as a dimer and presents the classical Nudix fold. Results from crystallographic and biochemical experiments suggest that CF $I_m$25 makes use of its Nudix fold to bind but not hydrolyze ATP and $Ap_4A$. The complex and apo protein structures provide insight into the active oligomeric state of CF $I_m$ and suggest a possible role of nucleotide binding in either the polyadenylation and/or cleavage steps of pre-messenger RNA 3' end processing.

# INTRODUCTION

Pre-messenger RNA 3'-end processing in eukaryotes is a two-step reaction consisting of endonucleolytic cleavage of the pre-mRNA followed by addition of a poly(A) tail at the 3' end of the upstream cleavage product (*13-15*). The coupling of these processing reactions relies on multiple protein-protein and protein-RNA interactions. The factors that are necessary and sufficient to reconstitute cleavage and polyadenylation in a mammalian *in vitro* system are poly(A) polymerase (PAP), cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), cleavage factor $I_m$ (CF $I_m$), cleavage factor $II_m$ (CF $II_m$) and the nuclear poly(A) binding protein 1 (PABPN1). The recruitment of the mammalian polyadenylation machinery to the pre-mRNA relies on the recognition of conserved sequence elements, such as the highly conserved hexamer AAUAAA recognized by CPSF and a U rich sequence downstream of the cleavage site recognized by CstF. Additionally a set of UGUA elements, a third sequence element found upstream of the cleavage site and not as universally conserved, is recognized by CF $I_m$ (*27, 28*). Upon completion of the processing steps the mRNA can be efficiently transported out of the nucleus into the cytoplasm. Homologs of CF I subunits are also found in cDNA databases of *Caenorabditis elegans*, *Drosophila melanogaster*, and plants.

CF $I_m$ is an oligomer composed of a small 25 kDa subunit (CF $I_m$25, also referred to as CPSF5 or NUDT21) and a larger subunit of either 59 kDa, 68 kDa, or 72 kDa (*40*). The

three larger subunits share substantial sequence homology. They are encoded on two different genes and the 72 kDa subunit is a splice variant of the 68 kDa polypeptide. CF $I_m$ can be reconstituted *in vitro* from the 25- and 68 kDa subunits (*40*). CF $I_m$, in addition to having a role in regulation of poly(A) site recognition, is also involved in stabilizing CPSF at the conserved hexamer, enhances the rate of poly (A) site cleavage *in vitro*, and has been shown to interact with PAP and PABPN1 via its 25 kDa subunit (*41*). Furthermore $CFI_m$ also interacts with splicing factors, indicating a role in communicating between different RNA processing complexes (*12*).

Much of what we know about complex formation of processing factors on the pre-mRNA and its regulation has been investigated biochemically. Structural characterization of individual and multicomponent processing factors will elucidate the domain interactions important for the 3' pre-mRNA processing mechanism that cannot be deciphered by biochemical means. The structure of the 25 kDa subunit of CF $I_m$ (CF $I_m$25) will help further define the domains important for substrate and protein interactions.

The structure of CF $I_m$25 will also allow us to investigate the function of a Nudix domain present in this protein. The first Nudix protein to be characterized enzymatically and structurally was *Escherichia coli* MutT (*54*). Nudix proteins are generally characterized as housekeeping enzymes due to their role in hydrolysis of substrates described as **nu**cleoside **di**phosphate linked to another moiety **X,** many of which are potentially toxic molecules (*55, 56*). Nudix proteins have a conserved Nudix fold consisting of an α/β/α

sandwich. Within the Nudix fold the consensus sequence of the Nudix box is $GX_5EX_7REUXEEXGU$, where U is a hydrophobic residue and X is any residue, and folds into a loop - α helix - loop structure (*54*). Interestingly, the Nudix box in CF $I_m25$ lacks two of the four conserved glutamate residues, three of which were shown to be important for catalysis **(Figure 15)** (*59*).

In this study we present the crystal structure of human CF $I_m25$ alone and in complex with diadenosine tetraphosphate ($Ap_4A$). CF $I_m25$ crystallizes as a dimer, which is also the oligomeric state of the protein in solution. The CF $I_m25$ structure presents the classic Nudix α/β/α fold and harbors residues outside of the Nudix core that could potentiate ligand binding. Structural and biochemical evidence suggests that CF $I_m25$ binds, but does not hydrolyze, mono and di-adenosine nucleotides.

## MATERIALS AND METHODS

### *Protein purification*

The construction of the plasmid expressing the 25 kDa subunit of human cleavage factor $I_m$ (CF $I_m$) with an N-terminal 6xHis tag was previously described (*41*). The cDNA of CF $I_m$25 was subsequently cloned into a Gateway® vector with a dual 6xHis-maltose binding protein (MBP) affinity tag provided by Dr. David S. Waugh (National Cancer Institute, Frederick, MD) (*69*). The following primers were used in the polymerase chain reaction (PCR): 5' GAG AAC CTG TAC TTC CAG GGT ATG TCT GTG GTA CCG CCC 3', 5' GGG GAC CAC TTT GTA CAA GAA AGC TGG GTT ATT AGT TGT AAA TAA AAT TGA A 3', and 5' GGGG ACA AGT TTG TAC AAA AAA GCA GGC TCG GAG AAC CTG TAC TTC CAG3'. CF $I_m$25 (227 amino acids) was then expressed in Rosetta (DE3) pLysS cells (Novagen) and grown in LB medium for 24 hours at 25°C following induction with 0.4 mM IPTG. The protein was purified on Nickel NTA beads (Qiagen), followed by a Tobacco etch virus (Tev) protease cleavage step, and cation exchange on a Resource S column (GE Healthcare): Cells were lysed at 4ºC by sonication in a buffer containing 20 mM Tris-HCl pH 8.0, 200 mM NaCl, and a protease inhibitor tablet (Roche). The lysate was centrifuged at 12,000 x g and incubated with Ni-NTA beads (Qiagen) for 1 hour at 4°C. Protein was eluted with lysis buffer containing 100-500 mM imidazole. Pooled fractions were dialyzed into 10% (v/v) glycerol, 20 mM Tris-HCl pH 7.5, and 50 mM KCl. CF $I_m$25 was cleaved from the MBP tag by the addition of equimolar amounts of Tev protease. MBP and Tev were then separated from

CF $I_m25$ by elution with a 50 mM-1 M KCl gradient at pH 7.5 on a Resource S column (GE Healthcare). The selenomethionyl protein was produced by inhibiting methionine biosynthesis in *Escherichia coli* (*70*) and purified as described above. Proteins were concentrated to about 6-20 mg/ml (Millipore Amicon Ultra-15), flash frozen and stored at -80ºC.

*Crystallization*

CF $I_m25$ crystals were initially obtained with the sitting drop method in a 96-well tray format. Sitting drops were set up with a 925 PC Workstation (Gilson) by mixing 0.6 µl of protein (16 mg/ml) with 0.6 µl reservoir solution (25% w/v PEG 3350, 0.2 M $MgCl_2$, 0.1 M Tris-HCl pH 8.5) (Hampton Research Index Screen, condition 85) and equilibrated against 160 µl reservoir buffer. Subsequently, crystals were obtained by streak seeding hanging drops with a protein concentration of 6 mg/ml. Hanging drops were set up by mixing 1 µl of protein with 1 µl of reservoir solution under the same crystallization conditions. Trigonal crystals grew to a maximum length of 200 µm in space group $P3_121$ ($P3_221$) with unit cell parameters $a = b = 80.11$ Å, $c = 72.21$ Å and $\gamma = 120º$. There is one molecule per asymmetric unit with an estimated solvent content of 52%. Crystals of the complex with $Ap_4A$ were obtained with the hanging drop method. The hanging drops were set up manually by mixing 1 µl of protein with 1 µl of reservoir solution (25% w/v PEG 3350, 0.025 M $MgCl_2$, 0.1 M Tris-HCl pH 7.5). The drops were streak seeded after a 24 h incubation period. When the crystals reached at least 100 µm, $Ap_4A$ at a final

concentration of 44 mM was added directly to the drop. After six hours the soaked $Ap_4A$ crystals were cryoprotected by the addition of 1 µl of 25% (w/v) PEG 3350 and 50% (v/v) glycerol to the 2 µl hanging drop prior to flash cooling in liquid nitrogen.

*Crystallographic data collection*

Multiple wavelength anomalous diffraction (MAD) data were collected at beamline 23-ID-D (Advanced Photon Source at Argonne National Laboratory) on a MAR m300 CCD detector. One complete selenomethionyl MAD data set was collected on one crystal at the peak, inflection, and high-energy remote wavelengths to a maximum resolution of 1.85 Å. Data were collected at 1.80 Å resolution on the $Ap_4A$ complex at beamline X12B (National Synchrotron Light Source) on a Quantum-4 CCD (ADSC) detector. The data from three $Ap_4A$-soaked crystals were merged to increase redundancy. Diffraction data were processed and scaled with DENZO and SCALEPACK (*71*). Data collection statistics are summarized in Table 1.

*Structure determination and refinement*

The program SOLVE (*72*) identified three of the four selenium sites (The N-terminal methionine is disordered or missing). AutoSHARP (*73*) was then used for refinement of the selenium parameters. The space group was judged to be $P3_121$ and not the enantiomorphic $P3_221$, based on the map quality and continuity. The phasing information was then used in RESOLVE (*72*) for density modification and iterative model building.

70% of the model was built by RESOLVE. The remaining residues were built manually using the program COOT (*74*). Residues 1-20 and 132-135 were omitted from the model because of poorly defined density.

Iterative rounds of refinement including simulated annealing, energy minimization, and B-factor refinement were done with CNS (*75*). Each refinement round was followed by rebuilding in COOT. A composite simulated annealing omit map was generated in CNS to validate the model and build the remaining side chains. Water molecules were added with CNS and COOT. The quality of the model was evaluated with PROCHECK (*76*). All non-glycine residues fall within either the most favored or additionally allowed regions of the Ramachandran plot. The refined model of the unliganded protein provided phases to calculate an isomorphous difference Fourier (Fo-Fo) map between the unliganded protein and the complex with $Ap_4A$ (*77*). The $R_{cross}$ on amplitudes between the two data sets is 0.149, indicating good isomorphism between the two crystals. The resulting map showed clear density for one adenine base and three phosphates of $Ap_4A$. The complex with $Ap_4A$ was refined with CNS (*75*). The refinement statistics for both structures are reported in **Table 3**. Figures were drawn with PyMOL (*78*).

*Oligomeric state determination*

Size exclusion chromatography was performed with a Superdex 75 column (GE Healthcare). The protein sample or molecular mass standards were applied to the Superdex 75 column and eluted with 10% glycerol, 20mM Tris pH 7.5, and 50mM KCl. Standard proteins (Sigma) were: lysozyme (MW 14.4 kDa), egg albumin (45 kDa), bovine serum albumin (67 kDa), and the void volume was determined with blue dextran (GE Healthcare). Dynamic light scattering was also used to confirm the oligomeric state of CF $I_m25$ (Dynapro, Wyatt). The buried surface area (2,800 $Å^2$) was calculated with AREAIMOL (*79*) and with CNS (*75*) (2,700 $Å^2$).

*Hydrolase activity assay*

The standard reaction mixture contained 50 mM Tris-HCl, pH 7.5 or 8.5, 50 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 20 μM CF $I_m25$, 2 U calf intestinal phosphatase (New England Biolabs) or 1 U *S. cerevisiae* pyrophosphatase (Sigma), and 2 mM or 4 mM substrate in a total volume of 50 μl. The putative substrates tested for CF $I_m25$ hydrolase activity include ATP, $Ap_4A$ (Sigma), and $m^7G$ (5')ppp(5')G cap structure analog (New England Biolabs). The reactants were combined at 4°C then incubated at 25°C for 30 and 60 minutes. The reaction was terminated by the addition of 250 μl of 20 mM EDTA and the liberated orthophosphate was determined by the colorimetric assay of Ames and Dubin (*80*) (*54*). The limit of detection of this assay is 5 μM of orthophosphate.

*Fluorescence measurements*

Steady state tryptophan fluorescence was measured with a Quantamaster fluorimeter (Photon Technology International, South Brunswick, NJ) as described (*81*) with a WG320 cut-off emission filter. CF $I_m25$ contains four tryptophans and only three are built in the structure: Trp148 and Trp149 are within 10-15 Å of the active site and Trp139 is located within 20Å of the active site. The fourth tryptophan, Trp13, is located in the disordered portion of the amino terminus. The tryptophan emission spectrum was measured by excitation of the samples at 295 nm and collecting the emitted fluorescence at 90° to the incident light over the range 300-400 nm. The slit widths were set at a resolution of 1 nm for excitation and 4 nm for emission. Fluorescence measurements of all protein samples were performed using a microcuvettte with a magnetic stir bar in 20 mM Tris-HCl pH 7.5, 50 mM KCl, and 25 mM $MgCl_2$ at 25°C for the protein alone and in the presence of increasing amounts of nucleotide. All fluorescence measurements were corrected for Raman scatter and background fluorescence and represent experiments performed in triplicate and then normalized and averaged. The ATP fluorescence data were fit with a single hyperbola (y = a*x/(b+x)) with a $K_d$ of 1.53 ± 0.18 mM (1.17 – 1.89 mM at 95% confidence interval). The $Ap_4A$ data were fitted with a single hyperbola (y = a*x/(b+x)) with a $K_d$ of 2.44 ± 0.49 μM (1.46 – 3.43 μM at 95% confidence interval).

*Cleavage and Polyadenylation assays*

Lamin RNA substrate was capped and uniformly labeled by SP6 RNA polymerase transcription of a PCR-amplified human DNA. *In vitro* poly(A) cleavage and poly(A) addition assays using dialyzed and undialyzed nuclear extract from HeLa cells were carried out as described previously with the addition of diadenosine tetraphosphate ($Ap_4A$) in increasing amounts (*27*). Increasing amounts of $Ap_4A$ were added in addition to stoichiometric amounts of $MgCl_2$.

## RESULTS

### *Crystal structure of human CF $I_m$25*

The original N-terminal His-tagged plasmid of human CF $I_m$25 (*41*) did not express to high enough levels for structural studies. We therefore inserted the coding sequence of CF $I_m$25 into a dual HisMBP vector (*69*). With the resulting expression vector, 1.5 mg of protein could be purified from 1 L of culture. The structure was solved to a resolution of 1.85 Å by multiple wavelength anomalous diffraction of the selenomethionyl protein variant. Residues 21-131 and 136-227 are visible in the electron density map. A complex with $Ap_4A$ was also obtained and refined to a resolution of 1.80 Å.

### *Description of the structure*

Human CF $I_m$25 is composed of 227 residues, with a calculated molecular weight of 26 kDa. CF $I_m$25 elutes as a dimer in gel exclusion chromatography with an apparent molecular weight of ~53 kDa **(Figure 16)**. The dimeric state of CF $I_m$25 has also been confirmed by dynamic light scattering. In the crystal structure, dimer formation relates two monomers by a two-fold crystallographic axis. The Nudix domain is located in the middle of the protein and spans residues 77-202 **(Figure 17)**. The Nudix fold comprises two mixed β sheets flanked by two helices (α2 and α3). The Nudix box is located in helix α2 (residues 109-131) and the preceding loop. The region immediately preceding the Nudix domain (residues 21-76) folds into two distinct segments: residues 21-32 form

an extended loop structure that projects away from the globular domain of the protein (residues 1-20 are not visible in the electron density map). The extended loop is stabilized via crystal packing interactions with a symmetry related monomer and this region, which has been shown to participate in RNA binding, is likely to adopt a different fold *in vivo* (*41*). Lys23 has previously been shown to undergo acetylation, a modification which modulates the interaction of CF $I_m25$ with PAP (*46*). The second segment of the N-terminal region (residues 33-76) contributes two short β strands and a long helix. Within this region an additional modification, this time phosphorylation, was reported for Tyr40 (*82*). We do not see either modification because the protein was expressed in *E. coli*. The C-terminal region (residues 203-227) is composed of two helices (α4 and α5) and a short beta strand (β10) comprising the final four residues.

Structural homologs of CF $I_m25$ were searched with the DALI server (*83*) and significant structural homology was found, but only with other Nudix proteins, including *Deinococcus radiodurans* DR1025 (Z score = 11.2, PDB ID code 1SJY, (*59*)), $Ap_4A$ hydrolase (Z score = 10.0, PDB ID code 1KT9, (*57*)), ADP ribose pyrophosphatase (Z score = 9.3; PDB ID code 1V8M; (*84*)) and the mRNA decapping enzyme DCP2P (Z score = 9.0; PDB ID code 2A6T; (*85*)). The highest Z-score (33.6 with associated rmsd of 0.6 Å) was actually that of another crystal structure of CF $I_m25$ for which coordinates were deposited but had no publication associated with it (Structural Genomics Consortium, Karolinska Institute; PDB ID code 2CL3).

*Dimer interactions*

Dimer formation buries an interface of about 2,700 $Å^2$ (*75, 79*). The buried surface is extensive and represents 13% of the surface of the homodimer. The dimer interface is maintained by approximately 20 residues per monomer, which participate in both hydrogen bonding and non-polar contacts. The monomer interactions are facilitated by the amino terminal extension (residues 21-32) and by helix α5, the loop preceding helix α5, helix α6, beta strand β6, the loop linking β6 and β7, and beta strand β10.

The main hydrogen bonding contributions are between the amino terminal extension, beta strand β6, and beta strand β10. These involve residues Thr32 of chain A and Asp142 of chain B, Ser220 of chain A and Asn147 of chain B as well as the reverse monomer interactions. The loop preceding helix α5 and the loop between β6 and β7 contribute to dimer formation through hydrophobic contacts involving residues Pro159 and Tyr160 from chain A and Tyr202 and Phe199 from chain B. Additional interactions are provided by stacking helix α5 from one monomer on to strand β6 of the opposing monomer in a tail-to-tail fashion allowing hydrogen bonding between residues in the loop following helix α5. These hydrogen bonding interactions include Arg221 and its symmetry mate and the carboxyl group of Ser220 with Asn147 from the opposing monomers.

*Metal binding*

CF $I_m25$ is missing the second and fourth conserved glutamates of the Nudix box (residues 124 and 128), which are replaced by a leucine and isoleucine, respectively. These glutamate residues are very often involved in metal binding (*59*). We therefore set out to investigate whether CF $I_m25$ was still capable of binding metal via the two remaining carboxylates. Although the CF $I_m25$ crystals were obtained in the presence of 200 mM $MgCl_2$, there was no identifiable $Mg^{2+}$ bound in the electron density map. Magnesium is a light atom, which is usually not easily identifiable in electron density maps and can often be mistaken for a water molecule. We therefore used metals that are more electron dense such as $MnCl_2$ and $GdCl_3$ to identify putative metal site(s) (*86*). $Mn^{2+}$ and $Gd^{3+}$ present the added advantage that they are anomalous scattering atoms, which should allow unambiguous identification of metal sites. We were unable to identify binding of either metal, regardless of whether the metal was co-crystallized or soaked into pre-existing crystals. Since Nudix enzymes require a divalent cation for catalysis (*54*), these experiments suggested that it is unlikely that CF $I_m25$ functions as a hydrolase. This finding called for further investigation of CF $I_m25$'s binding to potential substrates (see below).

*Enzymatic assays and substrate binding*

We next set out to investigate whether CF $I_m25$ binds a substrate. We used three different methods: a colorimetric assay to measure the release of inorganic phosphate (*80*), co-

crystallization/crystal soaking with putative substrates, and steady state tryptophan fluorescence experiments.

A colorimetric assay to measure the release of inorganic phosphate, and thus the potential hydrolytic activity of CF $I_m25$ on nucleotides, was performed (*80*). The following putative substrates were tested: ATP, $Ap_4A$, and the 7mG(5')ppp(5')G cap analog. $Ap_4A$ was chosen because the two enzymes found to be most structurally similar to CF $I_m25$ bind $Ap_4A$ (*57, 59*). ATP and the cap analog were tested because of their prominent role in RNA processing. The assay indicated that, within the limits of detection of the assay, none of the putative substrates tested were hydrolyzed.

Although CF $I_m25$ does not seem to possess a hydrolytic activity, it could still bind nucleotides. Several putative ligands were either co-crystallized or soaked into the crystals: ATP, GTP, ADP, GDP, diadenosine triphosphate ($Ap_3A$), $Ap_4A$, AMP, 7mG(5')ppp(5')G cap analog, NAD+, and GDP-mannose. In addition to testing these nucleotides, which are known substrates for Nudix enzymes, we also attempted to co-crystallize CF $I_m25$ with a 21mer RNA derived from the PAP γ cDNA sequence (*27, 28*). Of all the putative ligands tested, only $Ap_4A$ bound to the crystal (see below). The original crystallization conditions contained 200 mM $MgCl_2$ and 25% (w/v) polyethylene glycol (PEG) 3350. Because nucleotides can precipitate in the presence of high concentrations of PEG and magnesium (*87*), care was taken to modify the crystallization conditions so as to decrease or completely eliminate $MgCl_2$ and therefore lessen the risk

of the nucleotide precipitating out of solution. Conditions with NaCl or even no salt could be used in lieu of $MgCl_2$ to grow unliganded crystals. Even when the divalent cation was omitted from the crystallization solution crystals did not form with any of the ligands tested (with the notable exception of $Ap_4A$), demonstrating that the lack of binding was not due to the nucleotide falling out of solution.

Binding studies were then performed employing the intrinsic tryptophan fluorescence properties of CF $I_m25$. The following Nudix ligands were tested: ATP, GTP, $Ap_4A$, inositol hexaphosphate (IP6), 7mG(5')ppp(5')G cap analog, ADP, and AMP. Dissociation constants ($K_d$) measured by titrating the ligand concentration were estimated for ATP and $Ap_4A$ **(Figure 18A and 18B, respectively)**. The ATP data were fit with a single hyperbola curve, indicating one binding affinity with a $K_d$ value of 1.53 ± 0.18 mM. The $Ap_4A$ binding data were fit to a single hyperbola curve and represent one binding site per monomer, resulting in a $K_d$ of 2.44 ± 0.49 μM.

*Complex of CF $I_m25$ with $Ap_4A$*

Co-crystallization experiments were performed in order to resolve the ligand binding interactions of CF $I_m25$. Of all the molecules screened only $Ap_4A$ was captured in a co-crystal **(Figure 19A)**. The $Ap_4A$ co-crystal structure was determined by using the phases from the unliganded structure. The resulting unbiased isomorphous difference Fourier (Fo-Fo) map revealed distinct density for one of the adenine bases and three of the four phosphates of $Ap_4A$, in the cavity of CF $I_m25$ **(Figure 19B)**. The rmsd between the bound

and apo structures is low (0.41 Å), indicating that only very small changes take place upon substrate binding (*75*). The binding site residues Arg63, Arg150, Gln157, and Lys172 are found outside of the Nudix box region, within 3 Å of the triphosphate moiety, and are involved in coordinating the triphosphate moiety **(Figure 19B)**. The binding site is composed of residues from helix α1, beta strand β6, the loop linking β6 and β7, and beta strand β**7.** In CF $I_m$25, Arg63 and Arg150 are highly conserved across species while Gln157 and Lys172 are moderately conserved. The majority of hydrogen bonding interactions with $Ap_4A$ involve the oxygens of the β- and γ phosphates. The ordered adenine base of the $Ap_4A$ molecule stacks with Phe103, a residue contained within the Nudix domain. We note that the position of the γ phosphate of $Ap_4A$ coincides with the position of a sulfate ion reported for CF $I_m$25 (Structural Genomics Consortium, Karolinska Institute; PDB ID code 2J8Q).

CF $I_m$25 has a core structure similar to that of the *D. radiodurans* Nudix protein DR1025 (*59*) (PDB ID code 1SU2) **(Figure 20)**. Variations between these two structures arise from an extension in the loop linking β6 and β7, a shortening of the loop following β4 and an additional α-helix (helix α1) in the CF $I_m$25 structure. The additional α-helix in CF $I_m$25 plays a role in sequestering the substrate binding pocket from solvent exposure. When comparing the positions of the bound nucleotides between CF $I_m$25 and DR1025 we also see variations in substrate fit. Superposition of the ATP bound in DR1025 on to the CF $I_m$25 structure shows that both the base and phosphate tail would clash with protein residues in helix α1 and beta strand β5 in the CF $I_m$25 structure **(Figure 21)**.

Commonly, variations in the nucleotide binding region of Nudix proteins occur due to differences in the side chains and motifs and contribute to the substrate specificity of the Nudix protein. Interestingly, both DR1025 and CF $I_m25$ stabilize the adenine base via stacking interactions with a phenylalanine residue found outside of the Nudix box region. Base stacking interactions via a Tyr or Phe located 17 amino acids downstream of the Nudix box is commonly found in all of the Nudix $Ap_4A$ hydrolases, including DR1025 (*56*). Phe103 stabilizes the adenine base in the $Ap_4A$ bound CF $I_m25$ structure but is located six residues upstream of the Nudix box region.

*Chelation Affect of $Ap_4A$*

Poly(A) cleavage and addition reactions were carried out in the presence of increasing amounts of $Ap_4A$ to investigate the potential role $Ap_4A$ in pre-mRNA 3'-end processing. Poly(A) cleavage and addition assays were conducted with lamin RNA substrate and increasing concentrations of $Ap_4A$ or ATP from 0.625mM – 5mM. Both assays demonstrated a decrease in 3'-end processing at the concentration of 1.25mM Ap4A or ATP. This inhibition of processing could be alleviated by the stoichiometric addition of $MgCl_2$ and demonstrates the mechanism of inhibition is via chelation of catalytically important metals **(Figure 22)**. This suggests that if Ap4A does play a role in 3'-end processing events it can not be deciphered by utilizing the *in vitro* decoupled cleavage and poly(A) addition assays.

*Electrostatic surface and putative RNA and protein binding regions*

The electrostatic surface potential of CF $I_m25$ was calculated with GRASP **(Figure 23A)** (*88*). The amino terminal region and residues 78-160 of the Nudix domain were shown to participate in RNA binding (7). We observe a good correlation between the location of the putative RNA binding residues and that of the positively charged residues on the surface of the protein **(Figure 23B)**.

## DISCUSSION

In our crystal structure CF $I_m25$ is a homodimer. A dimeric state for CF $I_m25$ is consistent with dynamic light scattering (DLS) and gel filtration experiments performed with the 25 kDa subunit both unliganded and with ATP or $Ap_4A$. The dimeric structure of CF $I_m25$ suggests that the active form of CF $I_m$ may be a heterotrimer composed of a CF $I_m25$ homodimer and either the 59 kDa, 68 kDa, or 72 kDa subunit. This is consistent with interaction studies in *C. elegans* (DIP interaction database accession DIP:25083N) where the CF $I_m25$ homolog (Uniprot accession Q93716) was found to interact with itself and with the CF $I_m68$ homolog (Uniprot accession Q18937). Alternatively, the complex with the larger subunit of CF $I_m$ could be a heterotetramer. A third possibility is that CF $I_m25$ monomerizes upon binding the larger subunit.

The Nudix box of CF $I_m25$ lacks two of the four glutamates important for catalytic function and metal binding. Our structural and biochemical characterization suggests that CF $I_m25$ is able to bind but not hydrolyze nucleotide substrates. In the dinucleotide bound CF $I_m25$ structure, the dinucleotide is found outside of the Nudix box and lies deeper within the active site compared to ATP in DR1025. The fluorescence binding data for CF $I_m25$, indicate a binding affinity in the low micromolar range for $Ap_4A$ indicative of a potential role as a signaling molecule and a weak, albeit physiologically relevant, binding affinity for ATP. Even though CF $I_m25$ binds both $Ap_4A$ and ATP, we have so far only been able to obtain a co-crystal complex with $Ap_4A$. It is entirely possible that

the crystallization conditions we have explored to date for the protein-ATP complex are not compatible with the formation of a crystal lattice and that further exploration of the crystallization space might yield the desired conditions. We note that pre-formed crystals of CF $I_m25$ dissolve upon addition of ATP, indicating a possible conformation change upon binding of the nucleotide.

To our knowledge, CF $I_m25$ is the first example of a Nudix protein binding and not hydrolyzing a nucleotide substrate. This loss of function/gain of a regulatory role is not unprecedented in evolution. A similar loss of function was reported for the *Lactobacillus lactis* ATP phosphoribosyl transferase (ATP-PRT) regulatory subunit, His Z (*89*). ATP-PRT functions to initiate the biosynthesis of histidine and requires both the HisZ subunit and HisG, the catalytic subunit, for activity. The regulatory subunit, HisZ, resembles the catalytic domain of functional histidyl-tRNA synthetases (HisRS) and utilizes its fold, not for catalysis, but for binding of histidine to monitor histidine levels. Although there were no metals evident in either the free or bound structures of CF $I_m25$ and no identifiable hydrolytic activity in our colorimetric assay it remains possible that the Nudix box of CF $I_m25$ could potentiate hydrolysis of $Ap_4A$ and ATP upon interaction of CF $I_m25$ with one of the larger CF $I_m$ subunits or additional binding partners. Subsequent *in vitro* poly(A) cleavage and addition assays in the presence of $Ap_4A$ suggests a possible role in processing maybe more evident with a reconstituted cleavage and poly(A) addition reaction. A reconstituted system would allow addition of only the components necessary for both the cleavage and polyadenylation assays. Alternatively, this result could suggest

CF $I_m25$'s binding specificity is related to an alternative function outside of pre-mRNA 3'-end processing and additionally may validate a loss of function model.

$Ap_4A$ belongs to the family of diadenosine oligophosphates, $Ap_nA$, which were first discovered forty years ago (*90, 91*). $Ap_4A$ is composed of two adenosines and four phosphates linked in 5' to 5' phosphodiester linkages. The role of $Ap_nA$s in higher eukaryotes has remained elusive. Recently, $Ap_nA$s have been suggested to play a role as putative extra- and intracellular signaling molecules. $Ap_4A$ itself is involved in the cellular stress response, inhibition of $K_{ATP}$ channels, stimulation of DNA replication and repair, as well as influencing other essential cellular processes in eukaryotes (*90, 91*). Presently, the only link between $Ap_4A$ and 3'-end processing was described in yeast where stimulation of primer independent synthesis by yeast poly(A) polymerase was observed in the presence of dinucleoside polyphosphates, including $Ap_4A$ (*92*). There is no CF $I_m25$ homolog in yeast but this observation still suggests a potential role for dinucleoside polyphosphates as signaling molecules during RNA processing events. The concentration of ATP and $Ap_4A$ can fluctuate within the cell in response to cellular stress or growth. Preliminary experiments on the role of $Ap_4A$ during the polyadenylation step of 3' end processing resulted in a non-competitive inhibition of polyadenylation in a poly (A) extension assay with mammalian PAP (results not shown). This suggests that under conditions of high concentrations of $Ap_4A$ such as stress, $Ap_4A$ can bind to a site in PAP, other than the ATP binding site, to inhibit polyadenylation.

The residues 81-160 of the Nudix domain play a dual role, binding RNA and stabilizing protein-protein interactions. Additionally, the amino terminus (1-76) participates in RNA binding (Figure 2) (*41*). A stable interaction between CF $I_m25$ and CF $I_m68$, unlike that involving PAP or PABPN1, requires the entire CF $I_m25$ protein. CF $I_m68$ must contact regions outside of the known RNA binding region of CF $I_m25$ to promote complex formation and stimulate pre-mRNA 3' end processing. Interestingly, CF $I_m25$ has a patch of negatively charged surface residues that runs the length of the dimer interface (Figure 7A). This charged region is composed of residues primarily from beta strand β3 and helix α4 and could potentiate protein-protein interactions with CF $I_m25$'s other binding partners, possibly CF $I_m68$. Also intriguing is the observation that a binding interaction between the substrate RNA and the 25 kDa subunit occurs in the absence of a putative RNA binding domain. This suggests another mechanism of RNA recognition, possibly through homodimer formation. A CF $I_m25$ homodimer may enhance the binding potential to the RNA substrate compared to a monomer interaction by increased surface area interactions. A definite answer regarding the oligomeric state of the 25 kDa subunit and the RNA binding mechanism will have to await a CF $I_m25$ structure with RNA bound.

The results reported here suggest that CF $I_m25$ is unable to hydrolyze nucleotides or dinucleotides even though it harbors a classic Nudix fold. The Nudix domain of CF $I_m25$ may instead facilitate protein-protein interactions, as suggested by the large distribution of charged residues in the electrostatic surface representation of the Nudix domain. This

charge distribution of the Nudix domain correlates well with results from pull-down experiments with PAP and PABPN1 (*41*).

The interaction of CF $I_m$ with the RNA substrate and with PAP stimulates the rate of polyadenylate tail synthesis. This may be facilitated by the binding of ATP to CF $I_m$25. Preliminary data (S. Dettwiler and W. Keller, unpublished) showed that CF $I_m$25 interacts with hClp1, a protein shown to bind ATP which is involved in 3' pre-mRNA processing and tRNA splicing (*20, 93*). This observation, in conjunction with the fact that hClp1 has recently been shown to function as an siRNA kinase and a kinase that phosphorylates the 5' end of the 3' splicing product in human tRNA splicing (*94*), suggests a possible link between CF $I_m$25's binding of ATP and protein-protein cross talk. Recently, CF $I_m$25 has also been shown to be associated in a large RNP complex with Rae1, an mRNA export protein, in the nucleation and stabilization of microtubules during spindle assembly (*53*). This interaction is via CF $I_m$25 association with an RNA component of the RNP complex and suggests a role for CF $I_m$25 in mRNA export via direct RNA association that may be influenced by the concentration of intracellular nucleotides or dinucleotides.

## FUNDING

This work was supported by National Institutes of Health grant GM62239 to SD and a Department of Energy Experimental program to Stimulate Competitive Research predoctoral fellowship to MC. Work in the laboratory of WK was supported by the University of Basel and the Swiss National Science Foundation. The beamline at

Advanced Photon Source at Argonne National Laboratory (GM/CA-CAT) has been funded in whole or in part with federal funds from National Cancer Institute Grant Y1-CO-1020 and National Institute of General Medical Science Grant Y1-GM-1104. The beamline at the National Synchrotron Light Source, Brookhaven National Laboratory, was supported by the U.S. Department of Energy, Office of Science, Office of Basic Sciences, under Contract No. DE-AC02-98CH10886.

## ACKNOWLEDGMENTS

We are grateful to Dr. David Waugh (NIH) for his generous gift of the dual His-MBP vector and the Tev protease plasmid. We thank Justin Meyette for helpful discussions concerning protein purification and Dr. Mark A. Rould, Dr. Pierre Aller, Dylan Murray, and Karl Zahn for help with data collection and refinement, and Dr. Mark A. Rould for help with troubleshooting the fluorescence experiments. Coordinates and structure factor amplitudes have been deposited in the Protein Data Bank with entry codes 3BAP and 3BHO for the unliganded and $Ap_4A$ bound structures, respectively.

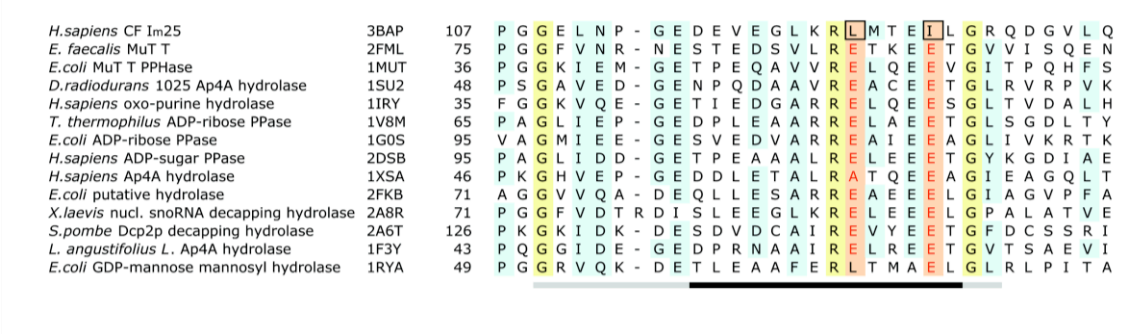



**Figure 15: Sequence alignment of CF $I_m25$ with Nudix proteins**

ClustalW sequence alignment of CF $I_m25$ and Nudix proteins of known structure (*95*). PDB ID codes are shown to the right of the enzyme names followed by the residue number of the first amino acid. The position of the Nudix box is indicated below the alignment as a grey/black bar where the black part marks the position of helix α2. Residues are on a light blue background if over 70% conserved or hydrophobic or hydrophilic and are on yellow background if invariant. The two catalytic glutamates conserved in most Nudix enzymes are displayed in red font on light orange background. L124 and I128, which are found in place of the conserved glutamates, are boxed in the CF $I_m25$ sequence. The abbreviations are defined as: PPHase for pyrophosphohydrolase and PPase for pyrophosphatase.

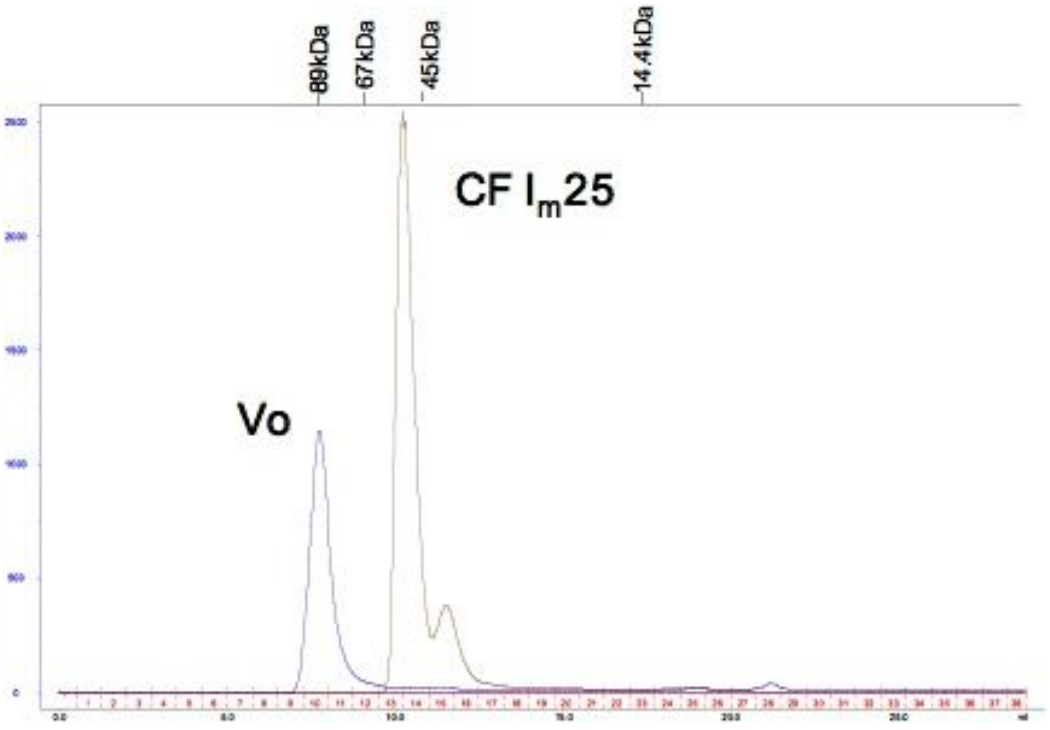



**Figure 16: Gel filtration profile of CF $I_m$25**

Gel filtration profile of CF $I_m$25 corresponding to an estimated molecular weight of 53 kDa. Vo is the elution volume of the blue dextran. Proteins were loaded onto a Superdex 75 gel filtration column (GE Healthcare), and protein elution was followed at $A_{280}$ nm. The elution volume of protein standards (bovine serum albumin, 67 kDa; egg albumin, 45 kDa; lysozyme, 14.4 kDa) and blue dextran is indicated on top of the graph.

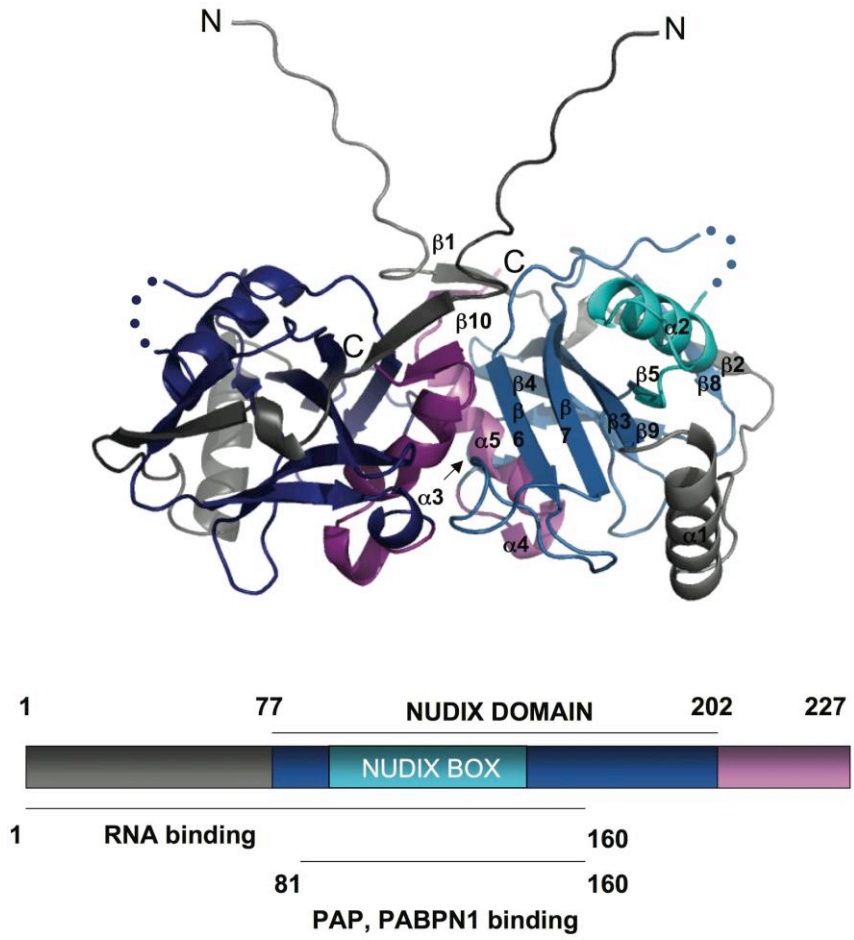
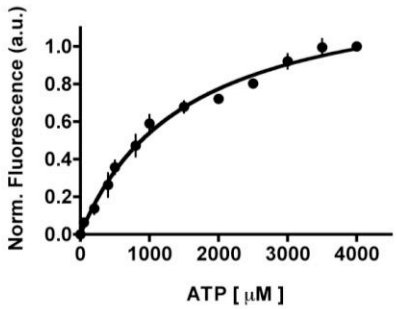



**Figure 17: Domain organization of CF $I_m25$**

Ribbon diagram of the CF $I_m25$ dimer comprising residues 21-131 and 136-227. The ribbon color scheme corresponds to that of the domain architecture shown below. The second monomer on the left is shown in light grey. The secondary structure numbering is based on a DSSP analysis (*96*).

A.




B.

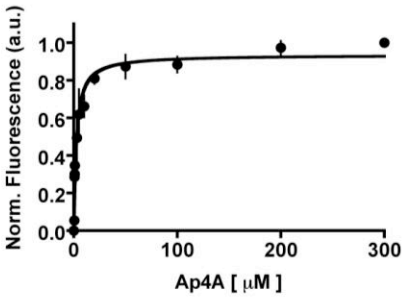



**Figure 18: CF $I_m25$ steady state tryptophan fluorescence experiments with ATP and $Ap_4A$**

All data are represented as normalized and averaged experiments done in triplicate. Error bars that are not represented lie within the symbol.
A. ATP steady-state tryptophan fluorescence data are fit with a single hyperbola, with $K_d$ = 1.53 mM.
B. $Ap_4A$ steady-state tryptophan fluorescence data are fit with a single hyperbola, with $K_d$ = 2.44 μM.

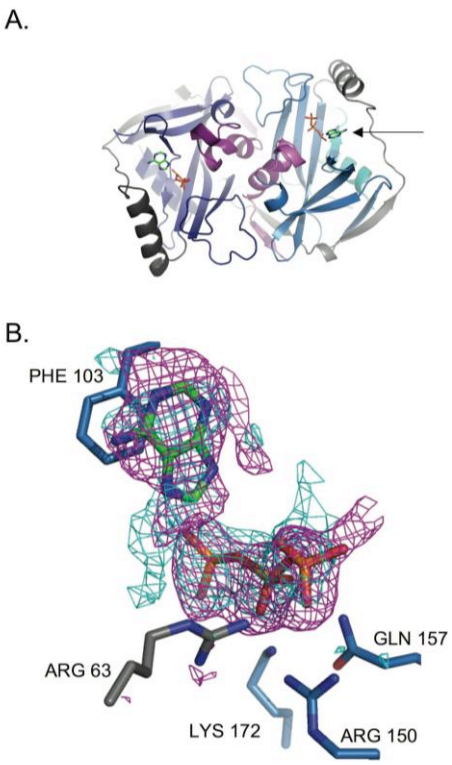



**Figure 19: Complex of CF $I_m25$ and diadenosine tetraphosphate**

A. The arrow points to the position of $Ap_4A$ within the CF $I_m25$ dimer shown rotated 120° as compared to Figure 2.
B. The ordered moieties of $Ap_4A$ (adenine base and triphosphate) are superimposed on an original, unbiased 1.80Å (Fo-Fo) isomorphous difference Fourier map contoured at 2.6σ (cyan) and a simulated annealing omit map contoured at 2.7σ (magenta). The adenine base of CF $I_m25$ is stabilized through base stacking interaction with Phe103. Arg63, Arg150, Gln157 and Lys172 contact the oxygens of the β and γ phosphates.

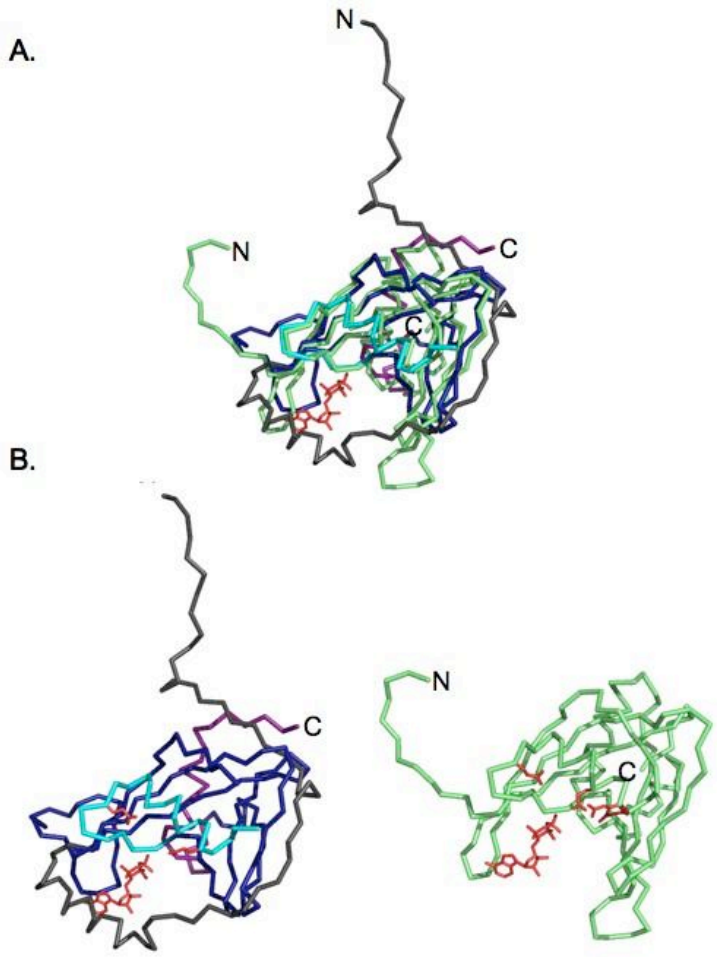



**Figure 20: Comparison with other Nudix proteins**

A. Superposition of the Nudix regions between DR1025 (PDB ID code 1SU2) (*59*) and CF $I_m25$ in their monomeric form. The superposition was calculated over residues 12-145 in DR1025 and 77-202 in CF $I_m25$. DR1025 is shown in light green and its bound substrate (ATP) is colored in red. CF $I_m25$ is colored according to its domain architecture as shown in Figure 2.

B. Side-by-side representation of the monomeric forms of the substrate bound DR1025 and CF $I_m25$ with ATP from the DR1025 structure and the conserved glutamate residues both shown in red.

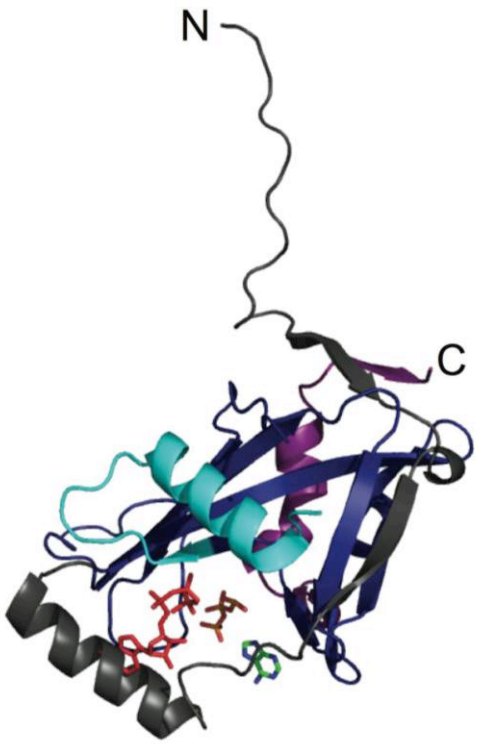



**Figure 21: Positions of DR1025 ATP and $Ap_4A$ in the CF $I_m25$ structure**

Positions of the DR1025 ATP shown in red and CF $I_m25$ $Ap_4A$ colored in orange (phosphates) and blue/green (adenine base) within the CF $I_m25$ structure. Ligand position was determined by superposition of the Nudix region over residues 12-145 in DR1025 (PDB ID code 1SU2) and 77-202 in CF $I_m25$. Only one CF $I_m25$ monomer is shown.

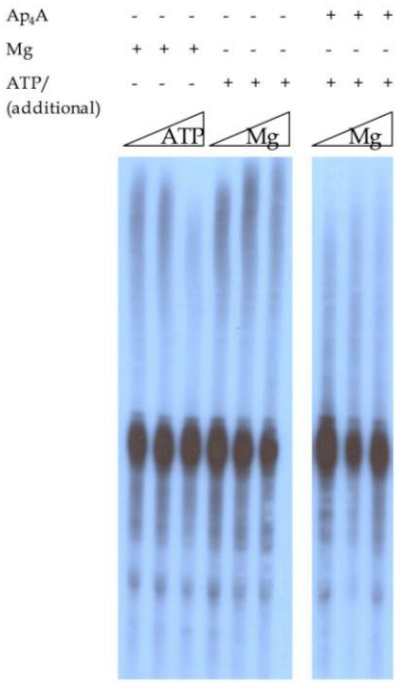



**Figure 22: Poly(A) assays in the presence of additional nucleotide**

Polyadenylation assays are presented with increasing amounts of either ATP or magnesium (Mg) with a triangle and constant or in addition to amounts, are presented by +/-. The addition of Ap4A or ATP inhibits poly(A)adenylation, but can be restored by stoichiometric amounts of Mg.

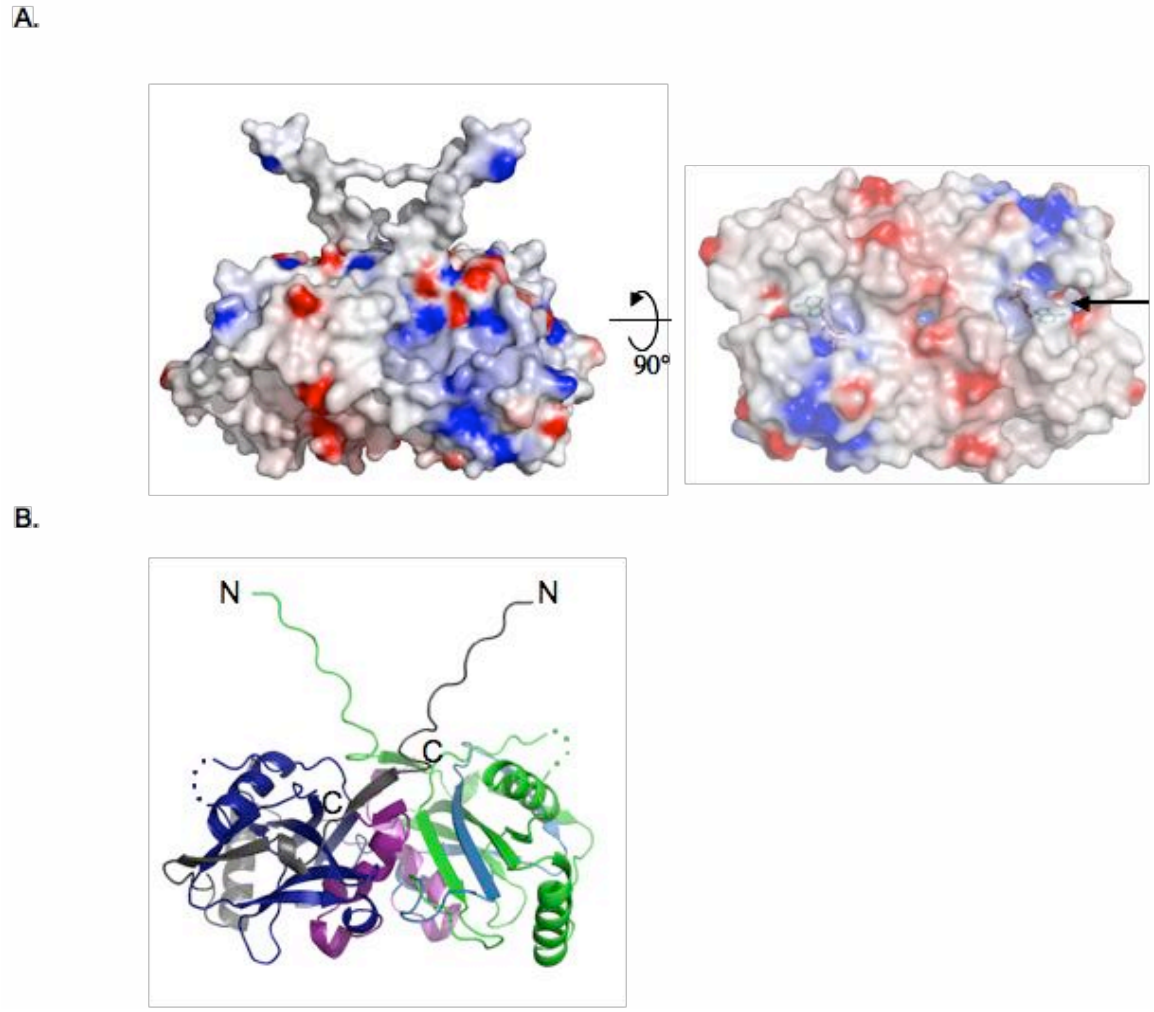



**Figure 23: Surface representation of CF $I_m25$**

A. Electrostatic potential surface representation of dimeric CF $I_m25$ calculated by the program GRASP (32). The surface is shown in the same orientation as in Figure 2 (left) and rotated 90° (right). The surface on the right was made semi-transparent to show the bound $Ap_4A$ (shown with arrow). The areas of negative charge are depicted in red, whereas the positively charged regions are colored in blue.
B. Ribbon diagram of dimeric CF $I_m25$ in the same orientation as in A. with residues 21-160 corresponding to the RNA binding region highlighted in green.

| Data collection | Se-Peak | Se-Edge | Se-Remote | $Ap_4A$ complex |
|---|---|---|---|---|
| Wavelength (Å) | 0.97923 | 0.97939 | 0.97166 | 0.9 |
| Resolution (Å) | 30-1.82 (1.89-1.82)[a] | 30-1.85 (1.92-1.85)[a] | 30-1.90 (1.97-1.90)[a] | 15-1.80 (1.86-1.80)[a] |
| Number of reflections | | | | |
| Measured | 157128 | 148933 | 136117 | 543233 |
| Unique | 23262 | 22468 | 20608 | 25257 |
| Redundancy | 6.8 (5.6) | 6.6 (5.2) | 6.6 (5.2) | 21.6 (6.7) |
| Completeness (%) | 97.7 (85.9) | 96.5 (80.6) | 95.7(79.9) | 99.8 (98.2) |
| $R_{merge}$ (%)[b] | 8.6 (51.7) | 7.3 (57.7) | 7.0 (52.8) | 13.8 (26.0) |
| I/σ | 16.09 (2.2) | 19.7 (2.3) | 19.3 (2.4) | 66.0 (4.4) |
| Selenium sites | 3 | | | |
| $R_{cullis}$[c] | .788 | .886 | .908 | |
| Phasing power | 1.13 | .736 | .815 | |
| Overall mean FOM | 0.33/0.72[d] | | | |

| | Se-Peak | $Ap_4A$ |
|---|---|---|
| PDB ID Code | 3BAP | 3BHO |
| Number of atoms | | |
| Protein (non-hydrogen) | 1649 | 1663 |
| Water | 167 | 169 |
| $R_{work}$[e] | 0.212 | 0.202 |
| $R_{free}$[e] | 0.234 | 0.225 |
| r.m.s.d. bond distances (Å) | 0.005 | 0.004 |
| r.m.s.d. bond angles (°) | 1.320 | 1.269 |
| Ramachandran plot | | |
| Most Favored Region (%) | 92.0 | 92.9 |
| Additionally Allowed (%) | 8.0 | 7.1 |
| Generously allowed (%) | 0 | 0 |
| Disallowed (%) | 0 | 0 |
| Average B factor all atoms (Å$^2$) | | |
| Wilson B-factor | 33.3 | 31.4 |
| Protein only | 38.7 | 33.0 |
| Water only | 50.2 | 44.7 |
| Ligand only | 58.7 | 50.4 |

[a] High resolution shell is shown in parentheses.

**Table 3: Data Collection and Refinement Statistics**

# CHAPTER 3:

## Crystallization Trials of CF $I_m25$ with Various Binding Partners

## Introduction: CF $I_m$

CF $I_m$ is recruited early to the pre-messenger RNA during pre-mRNA 3'-end processing to assemble the cleavage complex. Additionally, CF $I_m$ has been shown to regulate poly(A) site recognition by binding UGUAN-containing sequence elements in the vicinity of the poly(A) site (*27*). The binding of CF $I_m$ to UGUAN-containing RNA sequences suggests an additional mechanism to select for noncanonical hexamer sequences and alternative poly(A) sites. This has subsequently been demonstrated by the recognition by CF $I_m$ of human poly(A) sites lacking a conserved hexamer in a sequence specific manner (*28*). The recent crystal structure of CF $I_m$25 validates previous biochemical research regarding the RNA binding region and potential regions thought to be involved in protein-protein interactions with alternative processing factors. In addition, the recent finding that CF $I_m$25 is a dimer both in solution and in the crystal suggests an alternative stoichiometry of CF $I_m$ than previously described (*97*). The CF $I_m$ complex contains a dimer of the smaller 25 kDa subunit and a monomer or dimer of either the 59 kDa, 68 kDa, or 72kDa larger subunit.

The stoichiometry of the CF $I_m$ complex can be further investigated crystallographically by the pursuit of a complex structure containing the 25 kDa subunit and the 68 kDa subunit. Additionally, obtaining a complex of CF $I_m$ bound to a UGUAN containing sequence will investigate CF $I_m$'s regulation of pre-mRNA 3'-end processing events by recognition of UGUAN-containing RNA.

*Proteolytic Degradation Occurs at the C-terminal Region of CF $I_m$68*

Initial attempts to express and purify the CF $I_m$ components involved expression in insect cells and purified using a dual baculovirus vector. The dual vector allows for the expression of cDNA for both the 68 kDa and 25 kDa CF $I_m$ subunits. The complex was purified following expression using two chromatographic steps: CF $I_m$ components were purified away from insect cell lysate using Ni-NTA and anion exchange chromatography techniques. The eluted protein following the anion exchange column had a multiple banding pattern visible by SDS-PAGE at the apparent molecular weight of the 68 kDa subunit. This banding pattern increased over time and was thought to be due to the degradation at the C-terminal region of CF $I_m$68. The C-terminal region is predicted to be disordered by disorder prediction programs such as PONDR and Poodle (*23, 98*) making it a likely candidate for proteolytic degradation **(Figure 24)**. The 68 kDa subunit of CF $I_m$ is composed of an amino terminal RRM, central proline rich region, and a carboxy terminal alternating arginine/serine region (RS). We expected the amino terminus of the 68 kDa to be resistant to proteolysis due to the presence of CF $I_m$ 25 during our purification because of previously described interactions between the 68 kDa RRM region and CF $I_m$25 (*41*). The degradation of the 68 kDa C-terminal end was confirmed by Western blot analysis using an antibody which recognized the amino terminal hexahistidine tag of the 68 kDa subunit. The Western blot exhibits the similar banding pattern seen on the SDS-PAGE gel indicating degradation of the C-terminus of the 68 kDa subunit. Additionally, the protein complex required a concentration step prior to setting up crystallization screens to obtain a millimolar protein concentration range

suitable for crystallization. This process resulted in the precipitation of CF $I_m$ at concentrations higher than 2 mg/ml. Stabilization of CF $I_m$ could not be achieved by adjusting the salt concentration or the buffer pH as demonstrated by a greater than 20% polydispersity in Dynamic Light Scattering (DLS) experiments **(Figure 25)**. Crystallographic screens were still attempted with the CF $I_m$ complex despite the heterogeneity of the CF $I_m$68 C-terminal region and poor solubility of CF $I_m$ at high protein concentrations.

*UGUAN Containing RNAs do not Stabilize CF $I_m$*

Recombinant CF $I_m$ has been shown to bind with high affinity to UGUAN-containing RNAs as demonstrated by SELEX analysis (*27*). In this manner a UGUAN-containing RNA can be used to stabilize the degradation at the C-terminal region of the 68 kDa subunit and allow for a stable complex suitable for crystallization. To utilize this technique crystallographically a UGUAN-containing RNA of minimal length and bound by CF $I_m$ was first investigated by electromobility shift assays (EMSA). CF $I_m$'s ability to bind the short UGUAN-containing RNA by EMSA was compared to the 60mer SELEX 1 RNA to determine binding potential (*27*).

CF $I_m$ was unable to form a stable complex in gel shift experiments utilizing short UGUAN containing RNAs from the 3'UTR of PAP gamma and 3'UTR of PAP alpha. However, both a 20mer and a 40mer oligo, whose design was based on the SELEX 1 sequence, demonstrated weak binding in comparison to a high affinity shift seen with the 60mer SELEX RNA. Improvements to the shifting potential of CF $I_m$ were not achieved

with changes in the gel shift conditions. Therefore, having found two RNAs capable of forming a complex with CF $I_m$, albeit weak in comparison to the 60mer SELEX RNA, crystallization trials were initiated. However as mentioned previously this requires concentration of the protein sample prior to setting up crystallization screens. The short SELEX RNAs were included in the concentration step with the hope of increasing the protein solubility. As seen previously, CF $I_m$ precipitated out of solution at high concentrations regardless of the presence of RNA, confirming CF $I_m$ cannot be stabilized by the addition of RNA to achieve high protein concentrations necessary for crystallographic screening.

*Complex Crystallization Attempts with CF $I_m$25 and 68 kDa Truncation Mutants*

Prior attempts to crystallize CF $I_m$ as a complex alone or in the presence of RNA were unsuccessful due to protein solubility issues at high concentrations. This led us to focus on a minimum complex crystal structure containing the CF $I_m$25 subunit and the minimal region of the 68 kDa subunit necessary to maintain complex formation. Two 68 kDa truncation mutants, the 68 kDa RRM (68RRM) and the 68 kDa N-terminus (N68), were investigated for on their capacity to interact with the 25 kDa subunit (*41*). The 68RRM retains only the RRM region, (81-160) necessary to facilitate protein-protein interactions with pre-mRNA 3'-end processing factors including CF $I_m$25. The N68 construct contains amino acids 1-160 representing the N-terminal region in addition to the RRM **(Figure 26)**. It should be noted that both constructs lack the RS region and therefore are unable to interact with RNA (*41*).

The N68 construct was a primary focus due to its ability to maintain an interaction with CF $I_m25$ while encompassing the amino terminal region. The complex association and stoichiometry between the N68 mutant and CF $I_m25$ was evaluated using DLS and gel filtration experiments. Gel filtration experiments showed separate elution profiles for the N68 and CF $I_m25$ when combined at equal millimolar concentrations **(Figure 27)**. Additionally, the N68 construct with an empirical molecular weight of 17.6 kDa elutes prior to CF $I_m$ 25 suggesting an oligomeric state for the N68. Further assessment of the oligomeric state of the N68 by DLS also suggests an aggregated state represented with a high polydispersity and molecular weight. The high polydispersity of N68 cannot be resolved by addition of CF $I_m$ 25 in a ratio of 1:1 or 1:2 as evaluated by DLS or upon addition of CF $I_m25$ and $Ap_4A$ (**Figure 28)**. Crystallization trials were attempted, although the N68 and CF $I_m25$ did not demonstrate a stable complex by either gel filtration or DLS. Attempts were made under the assumption that a high concentration suitable for crystallization may drive complex association.

Along with the N68 characterization, the 68RRM construct was expressed and purified. This construct did not yield much protein suitable for crystallization trials and attempts to improve the expression and purification were not successful. Despite the low protein yield, 68RRM/CF $I_m25$ complex crystallization trials were preformed. Several crystallization conditions were screened and a single crystal was obtained with N68/CF $I_m25$ that did not diffract. Effort to improve the crystal conditions by varying the precipitant concentration and pH were not successful.

## Conclusions

The CF $I_m$ complex can be reconstituted *in vitro* with the 25 kDa subunit and the 68 kDa subunit. The interactions between the two subunits is via the RRM region of CF $I_m$68 and the Nudix region of CF $I_m$25 (unpublished data (*41*). The CF $I_m$25 protein is represented both in the crystal structure and in solution as a dimer with a large dimer interface suggestive of a CF $I_m$ complex containing a dimer of CF $I_m$25 and a monomer or dimer of the 68 kDa subunit. Attempts at addressing complex formation crystallographically with the both full-length constructs of CF $I_m$25 and CF $I_m$68 were not successful. This was certainly due to degradation at the C-terminal region of the 68 kDa subunit and protein solubility problems at high protein concentrations that could not be resolved by the addition of RNA. Increased stability between the two subunits was seen with a larger RNA substrate as suggested by the stable interaction with the SELEX 60 mer RNA in EMSA experiments. However, this RNA length presents problems during crystallization trials due to unbound regions that may remain flexible and hinder crystal packing. Analysis of the secondary fold of the UGUAN-containing RNAs suggests a secondary structure representing a hairpin fold where structure complexity increases based on length of the RNA **(Figure 29)**. This suggests the shorter RNA may not have the secondary structure complexity necessary for CF $I_m$ complex recognition.

Pursuit of a minimal crystal structure containing the CF $I_m$25 with two different CF $I_m$68 N-terminal region constructs was also not successful. Inability to obtain a diffracting complex crystal was impeded by poor expression and purification of the 68RRM and inability to optimize crystal conditions for the complex containing the N68

construct. In general, RRM domains have a rigid fold making them suitable candidates for crystallographic trials. The inability to obtain a crystal utilizing this construct suggests more crystallization conditions need to be screened. To do so requires improvement of the expression and purification system by cloning the RRM construct into a better optimal expression vector.

The N68/CF $I_m$25 complex yielded a crystal that did not diffract, could not be used for seeding techniques, and whose conditions could not be improved **(Figure 30)**. These results suggest that the N-terminal residues 1-80 of the 68 kDa are not well ordered in comparison to the RRM region. In agreement with this, secondary structure predictions of CF $I_m$68 place the N-terminal residues within the disordered regions of the plot. Although CF $I_m$25 has been shown to interact with the RRM region of CF $I_m$68 this interaction does not encompass the additional N-terminal amino acids 1-80 and alone these residues do not fold into a rigid structure suitable for crystallization.

To conclude, obtaining a complex structure of CF $I_m$ relies on optimizing the expression and purification of the 68RRM and additionally, it may be necessary to include other pre-mRNA 3'-end processing components in hopes of stabilizing subunit interactions.

## Introduction: CF $I_m$25 and PAP

CF $I_m$25 is recruited early to the pre-mRNA to recruit, stabilize, and regulate pre-mRNA 3'-end processing by facilitating multiple protein-protein interactions. One such

interactions includes the proposed binding of CF $I_m25$ with poly(A) polymerase (PAP), which can be described as controversial given the opposing biochemical data available. Initial findings by Kim *et al.* 2001 describe a minimal interaction between murine PAP and CF $I_m25$ comprising the C-terminal 69 residues of PAP and N-terminal 60 residues of CF $I_m25$, in the absence of RNA (*64*). Similar domain interactions have recently been characterized involving modifications to the C-terminal region of PAP and N-terminus of CF $I_m25$. Acetylation of lysines 635/644/730/734 of PAP and lysine 23 of CF $I_m25$ disrupts the 3'-end processing machinery indicating that acetylation can regulate pre-mRNA 3'-end processing (*46*). In contrast, a C-terminally truncated PAP (PAP513) was shown to still maintain interaction with CF $I_m25$ in pull-down assays **(Figure 31 and 32)** (*41*). This discrepancy can be investigated biochemically by determining the oligomeric state of the complex and crystallographically by obtaining a complex crystal structure of PAP513 and CF $I_m25$.

*A Stable Complex is not Maintained by Gel Filtration Chromatography*

The stoichiometry and oligomeric state of the CF $I_m25$ and PAP513 complex can be evaluated by both gel filtration experiments and DLS. As mentioned previously, CF $I_m25$ forms a stable dimer in solution and is expected, based on the large dimer interface, to maintain dimerization upon binding other 3' processing components (*97*). Investigation of the estimated molecular weight of the complex by gel filtration experiments was determined by varying the protein ratio of CF $I_m25$ to PAP513. Complex association was performed with an excess of CF $I_m25$ in order to obtain a

complex with either a 1:1 or 1:2 stoichiometry off the gel filtration column. The elution profile estimates the molecular weight of the complex to be 80 kDa on a Superdex 200 gel filtration column (GE Healthcare). The elution of CF $I_m25$ alone yields an apparent molecular weight of 95 kDa on the same column, which encompasses a wide molecular weight range and is therefore less exact at low molecular weights. This finding suggests the overlap in the elution profile may represent the co-elution of the individual proteins based on a similar molecular weight of 58 kDa for PAP and 52 kDa for a dimer of CF $I_m25$. This hypothesis can be further explored by observing the elution profile of CF $I_m25$ alone and the complex on a Superdex 75 column (GE Healthcare). The Superdex 75 has a narrower molecular weight range and therefore should represent a better estimate of the individual proteins molecular weight versus a complex molecular weight. The estimated molecular weight of the CF $I_m25$ dimer is 51 kDa and the complex elutes at approximately the same position indicating that both CF $I_m25$ and PAP are co-eluting based on their individual apparent molecular weights. Furthermore, the addition of ATP or $Ap_4A$ did not shift the elution profile to a larger molecular weight indicative of a complex formation.

Complex formation can be evaluated by dynamic light scattering experiments, which can also provide an estimate of the oligomeric state of each protein in solution. CF $I_m25$ and PAP513 incubated in a 2:1 ratio display a polydispersity in the range of 25% and a molecular weight of 80-90 kDa over a temperature range of 5°C - 25°C. Interestingly, the DLS intensity reading for the complex is unexpectedly low for a solution at 0.4 mg/mL concentration, suggesting that precipitation might have occurred

during sample preparation. Despite possible sample preparation issues, the DLS results complement the gel filtration data and imply that the interaction between the two proteins is not direct, the interaction involves a monomer of CF $I_m25$ with PAP513, or the experimental conditions are not suitable to maintain a complex interaction.

### *Complex is Stabilized during Electromobility Shift Assay*

The CF $I_m25$ and PAP513 complex association was additionally addressed using electromobility gel shift assays (EMSA). Incubation of PAP513 with CF $I_m25$ in the presence of $^{32}$P ATP resulted in a slight supershift when compared to PAP513 alone **(Figure 33)**. This experiment was performed once and should be repeated to confirm the supershift. This finding, however, suggests that CF $I_m25$ can interact with PAP and does so in the presence of ATP but provides little information on stoichiometry or affinity.

### *Crystallization Attempts of the CF $I_m25$/PAP513 Complex*

Crystallization trials of the CF $I_m25$ and PAP513 complex were set up at a 1:1 and 2:1 protein ratio and concentrated to approximately 3 mg/mL. In addition to crystallization trials of the complex alone, CF $I_m25$ and PAP513 were incubated with a nucleotide or nucleotide moiety to stabilize putative flexible binding regions. CF $I_m25$

and PAP513 were eluted over a gel filtration column following complex incubation with ATP, Ap4A, or 3'dATP. The eluted fractions were pooled individually based on the stoichiometry estimated by the staining intensity on an SDS-PAGE and concentrated. A complex was also attempted using a PAP513 mutant crosslinked to a 15mer oligo(A). The crosslink is made via an alanine to a cysteine mutation on residue 152. The PAP mutant, in addition to the mutation of the crosslinking residue, has all cysteines mutated to serines with the exception of residue 118 and 204, which are mutated to valines.

Possible complex crystals of CF $I_m25$ and the crosslinked PAP/RNA were obtained in several of the conditions from the Hampton Peg Ion Screen (Hampton Research). The crystals had a similar morphology to PAP513 crystals, resembling needles and the calculated volume of either a 1:1 or 1:2 complex crystal would be permitted in the orthorhombic space group of PAP513. To verify the crystal contained both CF $I_m25$ and PAP, the crystals were dissolved and run on an SDS-PAGE gel. The presence of only the crosslinked PAP on the gel confirmed the crystal did not contain a complex. Additionally, the crosslinked PAP/RNA was also able to crystallize in these conditions alone. One other crystal was obtained from the incubation of CF $I_m25$, PAP513, and $Ap_4A$ in the condition 0.1M sodium iodide, 28% (w/v) PEG 3350, and 10% (v/v) glycerol. This crystal did not diffract, did not produce crystal seeds following a liquid nitrogen thaw into similar conditions, and could not be reproduced.

## Conclusions

The above results concerning the CF $I_m25$ and PAP513 complex confirm that these proteins are able to interact and this interaction can be facilitated in the absence of the C-terminal region of PAP. However, there are still many unanswered questions regarding the stoichiometry of the complex and complex affinity that were not verified by DLS or gel filtration experiments. The gel filtration and DLS results provide evidence both for and against complex association. Results show a possible complex may be composed of a monomer of CF $I_m25$ and a monomer of PAP513 or the apparent molecular weight is reflective of the co-elution of the individual proteins, CF $I_m25$ as a dimer and PAP as a monomer. The complex association and affinity of the complex may be regulated by the presence of a nucleotide and/or substrate, such as RNA. This regulation could be reflective of the events during 3'-end processing where a weak interaction is apparent when processing components or substrates are absent. A direct or high affinity binding may additionally involve the N-terminal region of CF $I_m25$ and the C-terminal of PAP and explain why modification of these regions by acetylation results in disruption of the processing machinery.

In order to obtain a complex consisting of both CF $I_m25$ and PAP513 it may be necessary to determine the binding affinity in the presence of nucleotide or substrate via gel shift experiments. This could identify a high affinity state that could then be pursued crystallographically. An attempt at obtaining a complex may also be successful with additional 3'-end processing components to further stabilize CF $I_m25$ and PAP. The interaction of CPSF160 with PAP has been demonstrated to be important in PAP

stimulation in addition to the functions of Fip1 and CF $I_m$ suggesting CPSF160 may be a suitable candidate (*44*).

## Introduction: Fip1

Factor Interacting with Pap1 (Fip1) was identified as the fifth subunit of CPSF and in conjunction with CPSF160, Fip1 plays an essential role in the recruitment of PAP to the poly(A) site (*44*). PAP binding to the pre-mRNA is relatively weak and nonspecific and therefore requires recruitment to the poly(A) site by multiple protein-protein and protein-RNA interactions. This recruitment is facilitated by Fip1's recognition of U-rich USEs along with hexamer recognition by CPSF160. Additionally, Fip1 alone has been shown to stimulate PAP in a U-rich sequence dependent manner of RNAs containing both canonical and non-canonical hexamers (*44*). The additional involvement of CF $I_m$ is necessary for PAP recognition and recruitment in a noncanonical sequence context. The binding of CF $I_m$ to UGUAN-containing sequence elements within the RNA contributes to the recognition of the poly(A) site by PAP and recruitment of both PAP and Fip1 for poly(A) synthesis (*28*).

Fip1 is composed of a N-terminal acidic region known to interact with PAP, a central conserved region followed by a proline-rich region, and a C-terminal region composed of both an alternating arginine and aspartate region and an arginine region. Both the N-terminal and C-terminal region contribute to RNA binding and interestingly, these regions are highly susceptible to phosphorylation making Fip1 the most phospho-modified core processing component **(Figure 34)** (*44, 99*). The phosphorylation of Fip1 may be necessary to regulate pre-mRNA 3'-end processing events through interactions with PAP, the RNA or other processing factors. This can be investigated by solving the crystal structure of Fip1 alone and in complex with PAP. A crystal structure will confirm

the regions necessary for PAP interaction and provide insight into how modifications to the PAP-Fip1 interface may affect processing.

*Purification and Expression Attempts of Fip1*

The Fip1 construct consisting of the full-length protein with an amino terminal hexahistidine tag was initially used to express and purify the protein in *E. coli*. Attempts to purify the Fip1 construct using various nickel bead resins, expression using the autoinduction protocol (*100*), and dual transformation with the addition of the PAP513 plasmid were unsuccessful. There were no observable changes among various nickel bead slurries to the purification or improvement in the expression of Fip1 with the Studier autoinduction protocol. The purification of the full-length construct resulted in no obvious Fip1 band when observed by SDS-PAGE. The inclusion of the PAP513 plasmid during the transformation step did not stabilize Fip1 expression and resulted in either PAP513 expression or no protein expression.

Due to expression and purification problems in bacteria, we turned to an insect cell expression system. There was no protein visible by SDS-PAGE following purification from either SF9 or HighFive cells. This led us to believe the baculovirus stock containing the Fip1 cDNA needed to be further amplified to allow for efficient viral infection and expression from Fip1. Multiple rounds of infection were performed to amplify the virus, followed by subsequent screening for Fip1 expression of the amplified virus. This, however, did not improve the insect cell expression of Fip1 and resulted in

the pursuit of a more stable Fip1construct utilizing a maltose binding protein (MBP) with a dual affinity tag.

The MBP vector can be used to improve the solubility and stability of proteins expressed in bacteria (*101*). Full-length Fip1 was cloned into the MBP dual vector. The vector consists of an amino terminal MBP tag containing an internal hexahistidine tag and a tobacco etch cleavage site following the MBP tag and before the cloning site. Expression in *E. coli* and purification yielded multiple bands resembling Fip1 fragments. To ensure that the multiple fragments were the Fip1 construct the tag was removed via a Tobacco etch virus (TEV) protease cleavage step and purified away from the MBP tag using batch nickel beads. This resulted in multiple Fip1 bands in the flow-through fraction confirming all fragments were Fip1. This elution pattern suggests the C-terminal region of Fip1 is prone to proteolytic degradation which correlates with the full length protein disorder prediction plots **(Figure 35)**. Alternatively, the multiple banding pattern could represent alternative spliced forms of Fip1 and/or various forms of modified Fip1, both of which have been previously described (*44, 99*).

Despite the degradation, Fip1 fragments were incubated with PAP513 to identify which fragments maintained a PAP interaction. Fragment interaction was determined by incubation with the amino hexahistidine tagged PAP513 and purified by Ni-NTA beads and additionally by co-elution on a gel filtration column. None of the fragments of Fip1 co-purified with PAP513 off the batch nickel column while the gel filtration column gave an estimated molecular weight of 69 kDa for the complex. The apparent molecular weight of PAP513 alone is 48 kDa suggesting a 20 kDa size fragment of Fip1 maintains

an interaction with PAP based on the gel filtration experiment. Due to the small size of the Fip1 fragment and the expression and purification issues with full-length Fip1 a truncation mutant of Fip1 known to interact with PAP was cloned.

*Construct 1-355 of Fip1 Does not Stabilize Expression and Purification Attempts*

Previous biochemical data demonstrate a region encompassing residues 1-355 and containing the acidic and conserved region of Fip1 is able to retain an interaction with PAP513 in pull-down experiments (*44*). This Fip1 truncation (1-355Fip1) was constructed using the dual MBP vector previously described. In comparison to the full length, the 1-355Fip1 resulted in more protein visible by SDS-PAGE during the Ni-NTA purification step **(Figure 36)**. However, 1-355Fip1 also displayed a similar multiple banding pattern, which was previously shown to be the result of degradation. The interaction with PAP513 was investigated by both gel filtration and co-elution on an affinity column after the MBP tag was removed. This resulted in a similar outcome as described for the full length fragments, where PAP513 did not co-elute with 1-355Fip1 using affinity chromatography (Nickel resin). The gel filtration results gave an estimated molecular weight of 242 kDa and 157 kDa for the complex. The approximate molecular weight of 1-355Fip1 alone is 40 kDa. This molecular weight suggests a large complex of PAP513 with multiple 1-355Fip1 fragments that was confirmed by the multiple banding pattern on SDS-PAGE.

1-355Fip1 did not demonstrate stable expression or purification based on the multiple banding pattern and consequently a C-terminal hexahistidine tag was added to the dual MBP-1-355Fip1 (1-355Fip1C) vector with the hope to preclude degradation at the C-terminus. Multiple trials to purify 1-355Fip1C were not successful and multiple Fip1 fragments were still apparent. Lastly, to address instability issues Fip1 was cloned into a dual expression vector with PAP513 with the hope of increasing its expression and stability. The 1-355Fip1 and PAP513 constructs were cloned into the pET Duet vector (Invitrogen) with and without the dual His-MBP tag. Expression and purification attempts did not show an improvement from previous Fip1 trials.

*Point Mutation in Fip1 Constructs*

A single point mutation introducing a tryptophan in the place of a glycine at residue 293 was identified in all of the Fip1 constructs. To rule out the possibility that the point mutation prevented the proper expression and purification of Fip1 all of the Fip1 constructs were corrected and confirmed by sequence analysis. This resulted in corrections to MBP-1-355Fip, MBP-Fip1 full length, and the pET duet expression vector. Subsequent expression and purification protocols were performed on the corrected Fip1 constructs and analyzed for an improvement in the yield or stability of Fip1 protein. Changes to residue 293 did not show a noticeable improvement when compared to the previous trials and suggests Fip1 is susceptible to degradation that can not be stabilized by PAP513 interactions.

## Conclusions

Fip1 interacts with PAP and other pre-mRNA 3'-end processing components via its N-terminal domain and additionally makes contacts to the RNA with its C-terminal arginine rich region (*44*). These regions cooperate in binding U-rich USE's and establishing interactions with its processing partners for PAP recruitment and recognition of the poly(A) site (*28, 44*). In this respect, the ability to establish dual contacts requires a flexible domain architecture representative of Fip1 and noted by a highly disordered prediction plot. The flexible domain architecture of Fip1 makes it a difficult candidate for crystallographic trials due its susceptibility to proteolytic degradation. Additionally, the different isoforms of Fip1 and multiple phosphorylated states make it difficult to establish a purification scheme yielding one major Fip1 band. The difficulty in optimizing the expression and purification of Fip1 was evident in the described data above and surprisingly, Fip1 stability was not increased with the inclusion of PAP513. This suggests it may be necessary to include all Fip1 binding partners to create a more rigid complex across the length of Fip1 or construct a very small Fip1 fragments that still maintains binding interactions.

The recent crystal structure of yeast poly(A) polymerase (Pap1) in complex with a 25 amino acid fragment of yeast Fip1 also makes note of the disordered nature of Fip1 (*22*). The yeast Fip1 fragment shows a high propensity to be disordered in the absence of Pap1. This is likely to be more severe in the human Fip1 due to the presence of the acidic C-terminal extension and confirmed by the above results.

## Introduction: Clp1

Human Clp1 was originally identified as an essential component of cleavage factor II (CF $II_m$) in pre-mRNA 3'-end processing and is able to interact with both CPSF and CF $I_m$. The interaction of Clp1 with CF $I_m$ is through direct interaction with the Nudix domain of CF $I_m$25 demonstrated by pull-down assays (unpublished, (*93*). Recent data suggest Clp1 functions as an RNA-specific 5'-kinase crucial in both tRNA maturation and phosphorylation of interfering RNAs. The 5'-kinase activity has been demonstrated *in vitro* for the phosphorylation of siRNAs and subsequent incorporation into the RNA-induced silencing complex. The association of Clp1 with the tRNA splicing endonuclease complex and its 5'-kinase activity has been demonstrated *in vitro* and *in vivo* confirming Clp1's function in tRNA splicing (*62, 94, 102*).

In yeast, the Clp1 homologue similarly plays a role in pre-mRNA 3'-end processing and is additionally essential for cell viability. Interestingly, the human Clp1 is not able to complement a yeast Clp1 knockout and likewise the yeast Clp1 lacks *in vitro* 5'-kinase activity previously established for the human protein. This suggests that although these homologous proteins may share a similar domain architecture they may have very different functions **(Figure 37)**. In order, to address the differences between the human and yeast Clp1 a crystal structure of the human Clp1 is necessary to compare domain architecture and substrate interactions; a crystal structure of Clp1 in complex with CF $I_m$25 will be pursued as well.

*Expression and Purification*

The initial trials to express and purify human Clp1 with a hexahistidine affinity tag utilized a bacterial expression system. Multiple cell lines including Rosetta, BL21, and pLysE cells were used in parallel with several different growth conditions (Autoinduction method, low induction temperature, short induction period) and resulted in minimal or no expression of Clp1. The pLysE cells showed the highest Clp1 expression level following purification. However, the band corresponding to the molecular weight of Clp1 could not be confirmed by mass spectrometry methods. The inability to obtain protein confirmation suggests the expression of Clp1 is unstable.

In addition, the cDNA of full length and an amino terminal truncation mutant of Clp1 were cloned into the His-MBP dual vector system in an attempt to increase the expression level. The truncation mutant (N43Clp1) was constructed based on the estimated 43 N-terminal residues of Clp1 predicted to be disordered by disorder prediction programs and the disordered regions of the yeast Clp1 structure, which lacks the first 18 residues. Neither the full length or N43Clp1 dual MBP vectors showed an improved expression level. The poor expression of N43Clp1 may be due to the truncation itself, which interrupts a seven-stranded β sandwich.

The poor expression using the bacterial system led us to pursue expression of Clp1 using insect cells. The expression and purification of a hexahistidine tagged human Clp1 showed an increased expression level and the Clp1 protein was confirmed by mass spectrometry. Clp1 expressed in insect cells has a very clean elution following Ni-NTA purification with a yield of approximately 1 mg/L.

*Complex Crystallization Trials with CF $I_m$25 and Clp1*

The complex between CF $I_m$25 and Clp1 was formed by incubation of Clp1 with CF $I_m$25 in a 1:2 ratio with the addition of a 10-fold excess of ATP. The initial buffers consisting of 0.02 M Tris-HCl pH 7.5 and 200 mM KCl or 0.2 M Tris-HCl pH7.5 and 0.2 M Ammonium sulfate showed very low solubility of the complex at concentrations between 1.5 and 5 mg/mL. To increase the solubility, buffer screens were examined looking for a decrease in protein precipitation at concentrations between 2.5 and 3 mg/mL. Hepes buffer at 0.1 M or Tris-HCl buffer pH 7.8 and 150 mM NaCl increase the solubility of the Clp1 complex and show promising precipitant aggregation using the crystallization conditions previously described for the yeast Clp1 (*20*).

*Dynamic Light Scattering Experiments*

Dynamic light scattering experiments were carried out on the Clp1 and CF $I_m$25 complex to investigate putative changes in the oligomeric state with the addition of ATP **(Figure 38)**. Experiments conducted with and without ATP show two distinct peaks: one of very high molecular weight and polydispersity, which is more prevalent upon addition of ATP. The small peak has an ideal polydispersity ranging from 18-25% and an average molecular weight of 125 kDa. These DLS data suggest that there may be two distinct oligomeric states of the complex or each peak is representative of each individual protein.

Additionally, ATP addition does not seem to increase the solubility of the complex as previously thought.

To address the possibility that Clp1 and CF $I_m25$ may not be forming a complex, DLS experiments were performed with Clp1 alone. The DLS results show a very large molecular weight and high polydispersity peak much like the larger peak observed with the complex.

## Conclusions

The investigation of the Clp1 and CF $I_m25$ complex has provided initial results that are promising for obtaining a crystal structure of a complex. The complex is more soluble in 0.1 M Hepes pH 7.0 or 0.1 M Tris-HCl pH 7.8 and shows promising spherulite formation in the crystallization conditions reported for the yeast enzyme (0.1M Hepes pH 7.0 and 14-20% (v/v) PEG 8000) (*20*). Additionally, dynamic light scattering assays with and without ATP indicate that the addition of ATP may not stabilize the complex. Future experiments involving screening around the yeast crystal conditions by varying the pH and gel filtration assays to confirm complex association will prove valuable.




**Figure 24: Diagram of the Disordered Regions of CF $I_m$68**

The domain architecture is shown above the disorder prediction profile (Poodle-S) of CF $I_m$68. A high probability of disorder is represented by a the red line above the 0.5 mark. Most of the C-terminal region of CF Im68 is predicted to be disordered (*98*).

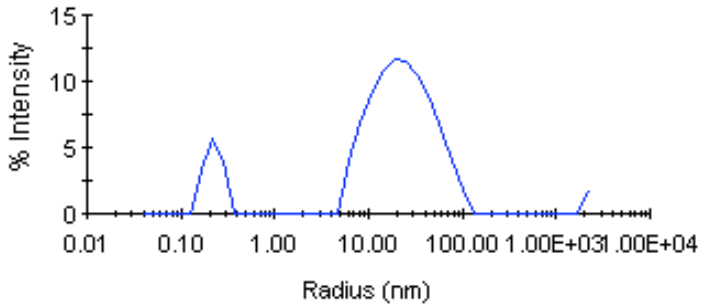



**Figure 25: Dynamic Light Scattering of the CF $I_m$ complex**

Dynamic light scattering with CF $I_m$ demonstrates a large oligomeric complex, represented by the second peak, with a estimated molecular weight of 8625 kDa (represented by the radius of the curve) and a high polydispersity of 74%.

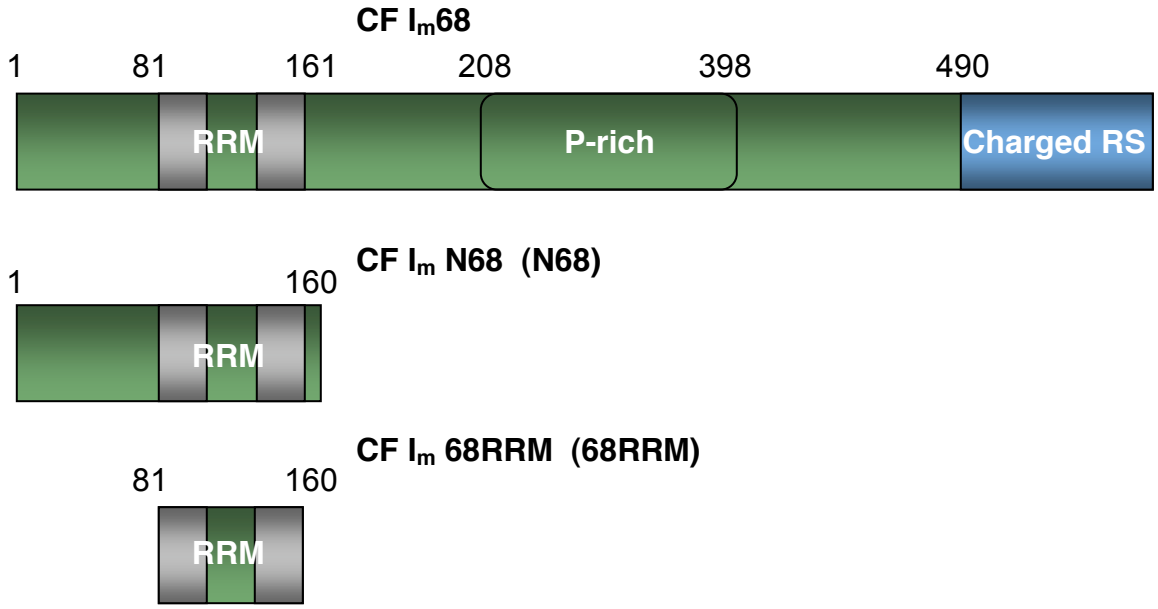



**Figure 26: Domain architecture of the 68kDa truncation mutants**

Complex association of CF $I_m$25 and CF $I_m$68 was investigated using truncation mutants comprising the N-terminal region of CF Im68 (N68 and residues 1-160) and only the RNA recognition motif (RRM) (RRM68 and residues 81-160).

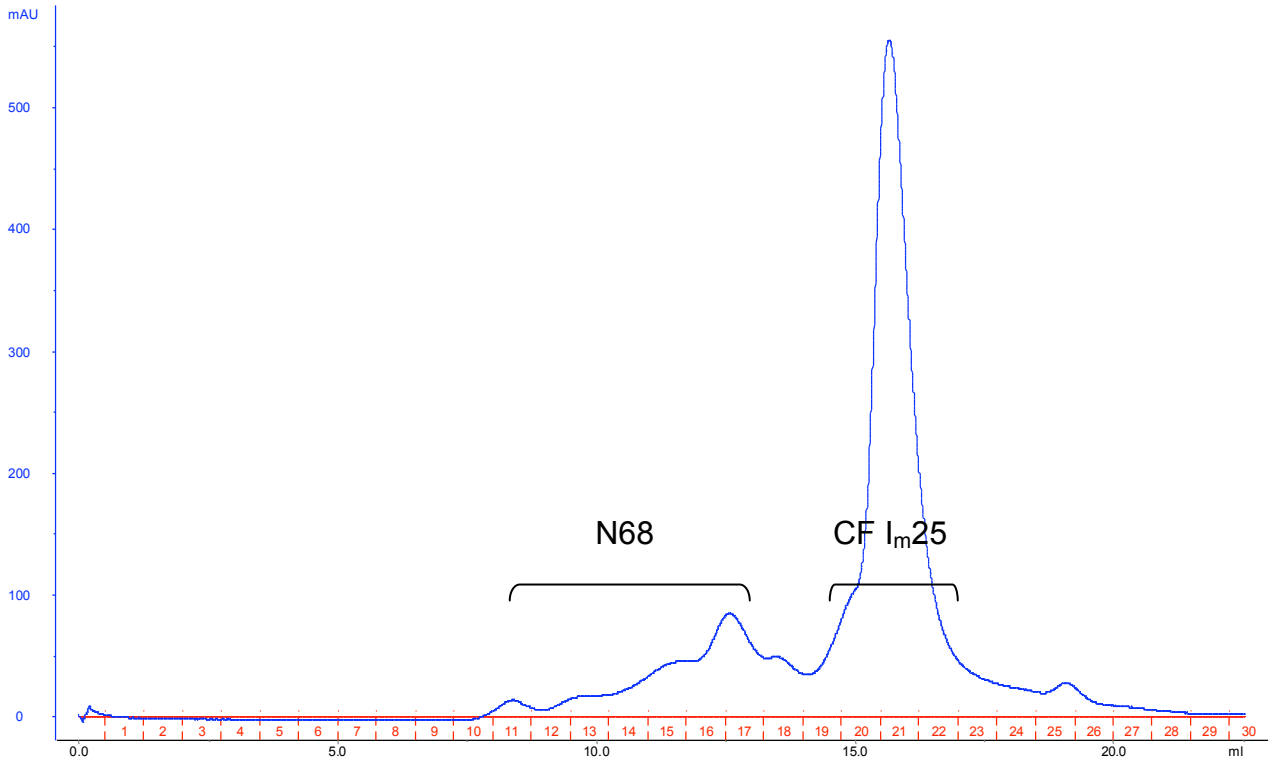



**Figure 27: Gel filtration profile of CF $I_m$25 and N68 complex**

N68 and CF $I_m$25 elute separately off a Superdex 200 (GE Healthcare) gel filtration column. The N68 protein elutes prior to CF $I_m$25 suggesting it forms a larger oligomer. The blue line represents absorbance at 280nm.

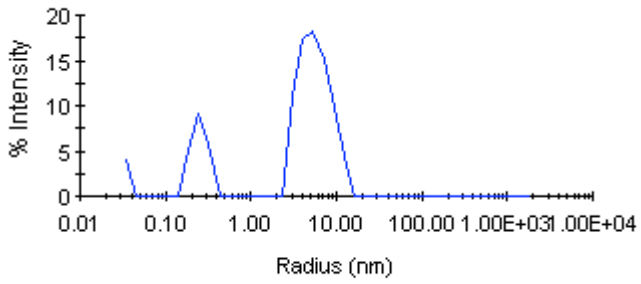



**Figure 28: Dynamic light scattering of CF $I_m25$ and N68 complex**

DLS data demonstrate a high polydispersity (41%) for the complex suggesting it may be a poor candidate for crystallization experiments.

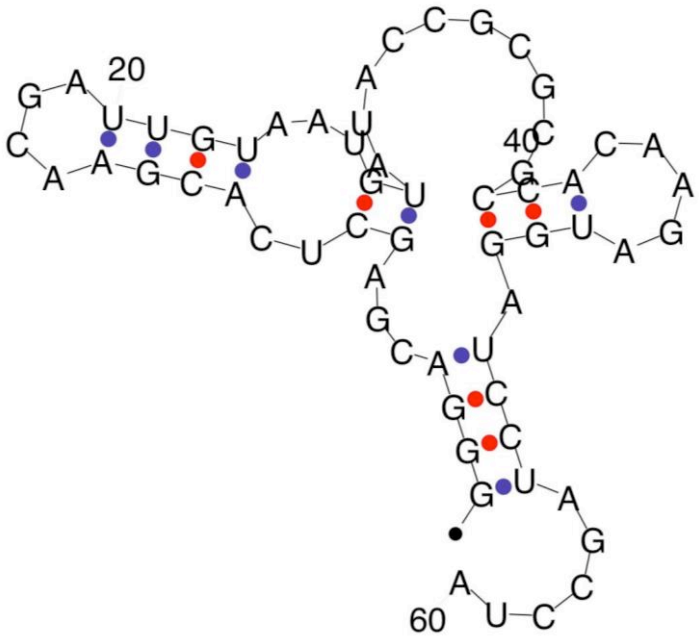



**Figure 29: Secondary structure prediction of SELEX 1 RNA**

An example of a secondary structure prediction (MFOLD) of 60mer SELEX 1 RNA with a minimal ΔG = -3.41 (initially -4.9). The fold prediction suggests a minimal energy state is maintained with a hairpin loop structure (*103*).



**Figure 30: Crystal of CF Im25 and N68**

Crystal of the complex set up in a 1:2 ratio of N68 to CF $I_m25$ at 2.5 mg/mL in the condition: 0.1 M Magnesium acetate, 10% (w/v) PEG, and 10% (v/v) glycerol at 12°C.

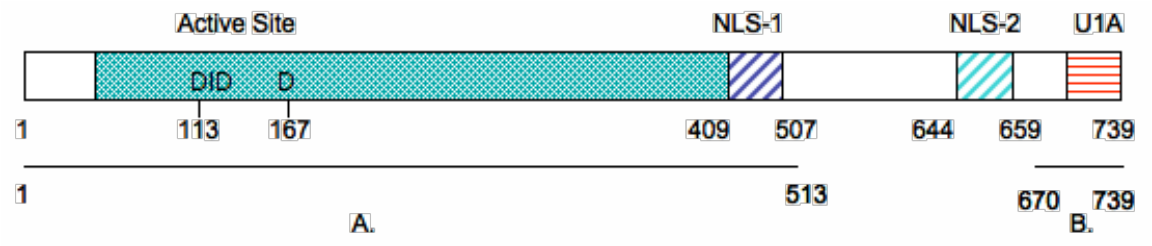



**Figure 31: Architecture of bovine PAP and regions involved in binding CF $I_m25$**

The domain architecture of full length bovine PAP is presented with the cyan region indicating the ordered residues in the PAP513 structure, the active site aspartic acid residues are indicated (D), as well as the positions of the nuclear localization signals (NLS-1 and NLS-2)(*50*). A. Represents the region in the C-terminal truncated bovine PAP (PAP513) that maintains interaction with CF $I_m25$(*41*). B. Represents previous data suggesting the C-terminal residues 670-739 of murine PAP are necessary for CF $I_m25$ interaction(*64*).

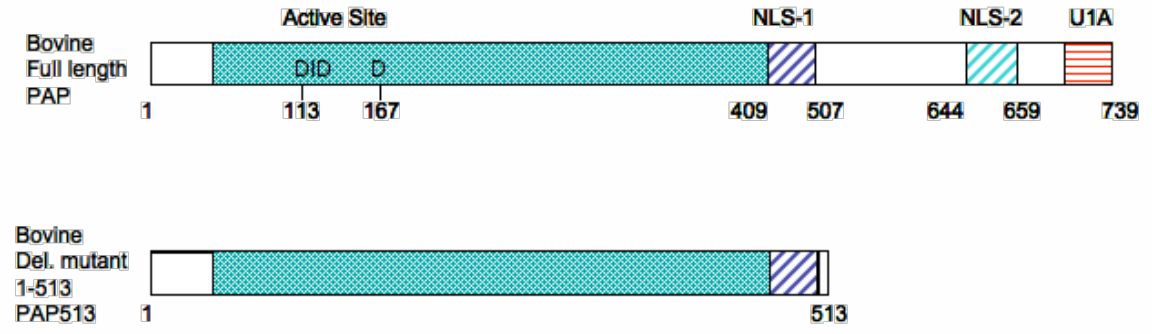



**Figure 32: Truncation mutant of bovine PAP**

The domain architecture of the full length and truncated bovine PAP (PAP513) protein. PAP513 contains the active site region necessary for polymerization.




**Figure 33: Gel shift of PAP513 and CF $I_m$25 complex**

Both CF $I_m$25 and PAP513 are capable of binding $^{32}$P-labeled ATP independently, evident by the change in mobility upon protein addition. In addition, CF $I_m$25 and PAP513 can form a complex as indicated by a small shift in comparison to the supershift seen with PAP513 alone (indicated with an arrow).

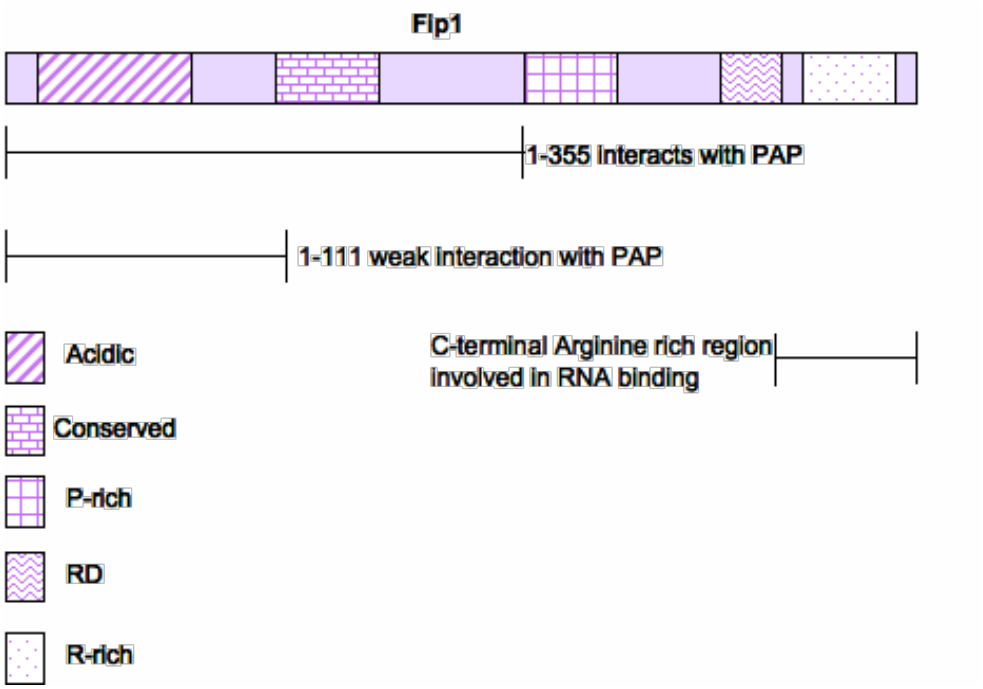



**Figure 34: Domain architecture of Fip1 and regions interacting with PAP513**

The human Fip1 protein is composed of an acidic N-terminus (Acidic), central conserved region (Conserved), proline rich region (P-rich), and a C-terminal region composed of both an alternating charged domain (RD) and an arginine rich domain (R-rich). The regions of Fip1 that are involved in PAP513 interactions are presented as described (*44*).

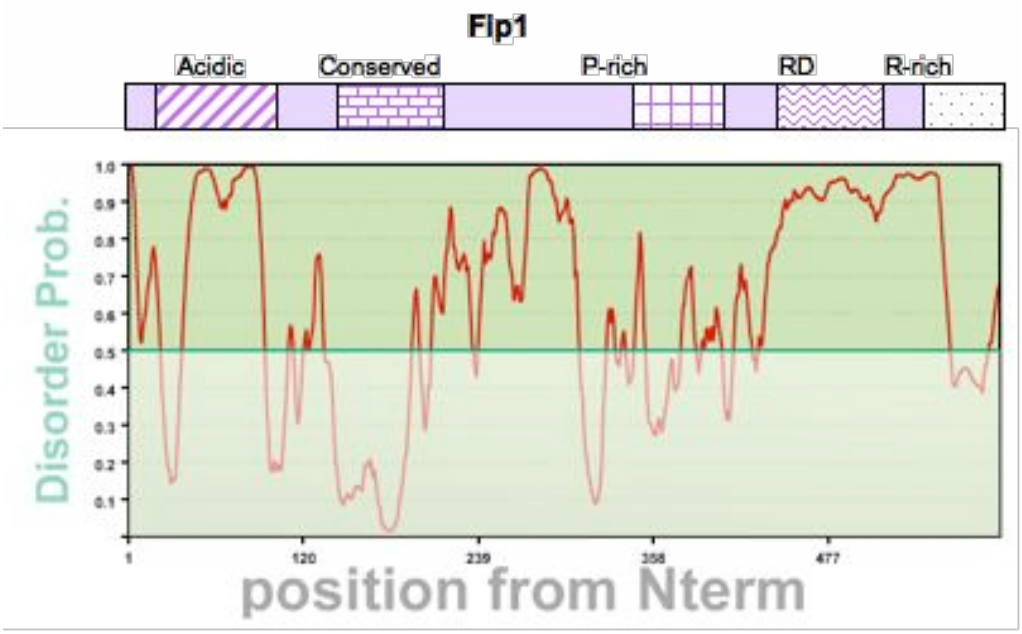



**Figure 35: Disorder prediction profile of human Fip1**

The disorder prediction profile of Fip1 (POODLE) shows multiple regions over the entire length of the protein have a high probability of being disordered (*98*).

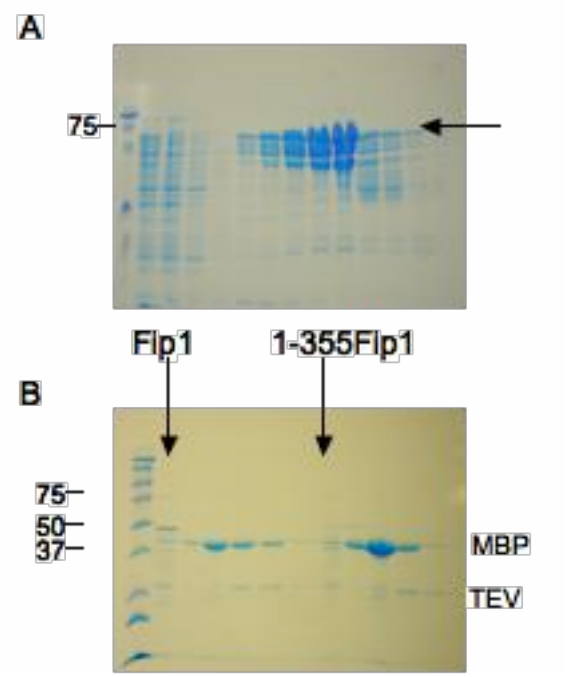



**Figure 36: SDS-page gel of Ni-NTA purified 1-355Fip1 before and after TEV cleavage**

A. Ni-NTA purification of MBP-1-355Fip1 (arrow indicates protein of interest). B. Ni-NTA purification following TEV cleavage of both the full length MBP-Fip1 and the MBP-1-355Fip1. The cleaved protein elutes in the flow through fraction indicated by the arrows and displays a multiple banding pattern due to proteolytic degradation.

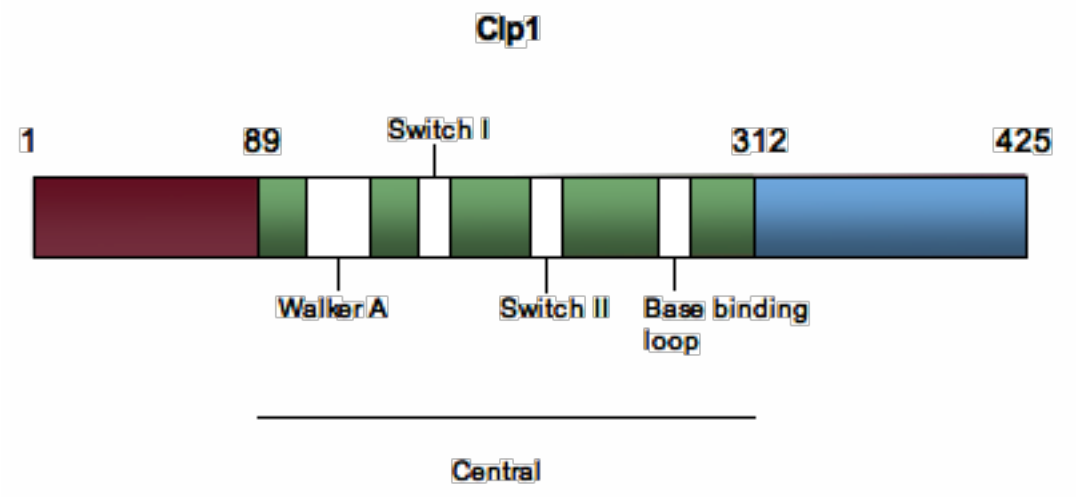



**Figure 37: Domain architecture of human Clp1**

Human Clp1 is composed of an N-terminal domain as shown in red, a central region involved in nucleotide binding, and a C-terminus shown in blue.

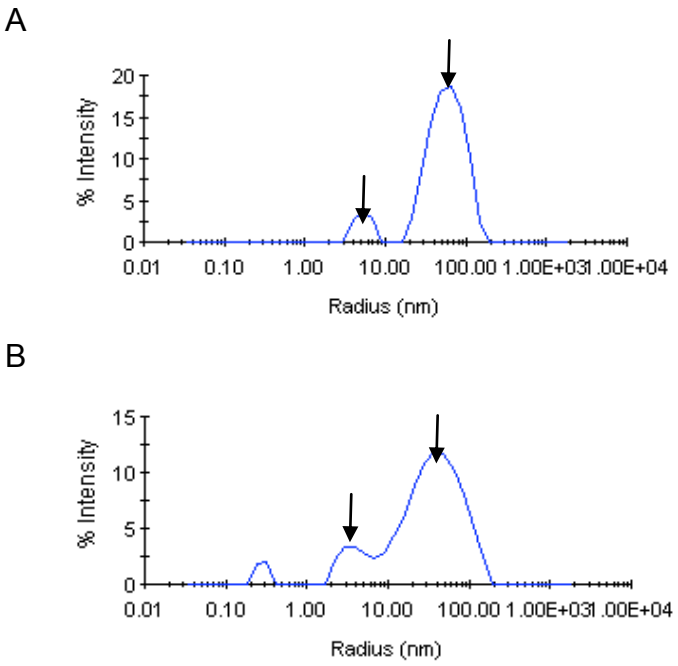



**Figure 38: Dynamic light scattering of Clp1/CF $I_m$25 complex at 15°C**

A. DLS of the Clp1/CF $I_m$25 complex in the conditions 0.1 M Hepes pH 7.0 and 150 mM NaCl. The polydispersity (pD) of the first peak is 21% and molecular weight (MW) is 169 kDa and the pD for the second peak is 46.8% with a MW of 55958 kDa. B. DLS of the Clp1/CF $I_m$25 complex and ATP in the same conditions. The pD of the first peak is 37% and MW is 96 kDa and the pD for the second peak is 68.6% with a MW of 28970 kDa. Two distinct oligomeric states are apparent in both experiments however, the polydispersity increases upon addition of ATP.

**Materials and Methods**

*Protein Purification*

The baculovirus dual expression vector containing the cDNA of the 25 kDa subunit and 68 kDa subunit of human cleavage factor $I_m$ (CF $I_m$) was a gift of Dr. G. Gilmartin (University of Vermont) and the purification has been previously described (*27*). The construction of the plasmid expressing the 25 kDa subunit of CF $I_m$ with a dual 6xHis-maltose binding protein (MBP) affinity tag was previously described (*97*). The 68 kDa RRM (81-160 residues; 68RRM) and the N-terminal (1-160 residues; N68) constructs with amino terminal glutathione affinity tags was a gift from Dr. W. Keller (Biozentrum, Basel, Switzerland) (*41*). The 68RRM and N68 were expressed in Rosetta 2 (DE3) pLysS cells (Novagen) and grown in LB medium for 24 hours following induction with 0.4 mM IPTG. Cells were lysed at 4°C by sonication in a buffer containing 1 x phosphate buffered saline pH 7.3 and a protease inhibitor tablet (Roche). The lysate was centrifuged at 12,000 x g prior to loading on a GST affinity column (GE Healthcare). The GST tag was removed by digestion with Tobacco etch virus (TEV) protease cleavage step and a second GST-affinity column (GE Healthcare) removed undigested material. Further purification required a final Capto S column, cation exchange (GE Healthcare). Protein concentration was determined by Bradford assay and combined in a 1:1 or 1:2 mM ratio with CF $I_m$25. Proteins were concentrated to approximately 1-4mg/mL (Millipore Amicon Ultra-15), flash frozen and stored at -80°C.

The bovine PAP513 plasmid was a gift from Dr. W. Keller and purified as previously described (*50*). The crosslinked PAP513 to an oligo(A) 15mer was a gift

from Qin Yang (University of Vermont). Proteins were concentrated to 3-6 mg/mL at a ratio of 1:1 or 1:2 (PAP513:CF $I_m25$) with or without nucleotide.

The construction of the plasmid expressing the human Fip1 protein with a hexahistidne affinity tag and human Fip1 baculovirus was a gift from Dr. W. Keller and purified as previously described (*44*). The cDNA of human Fip1 was cloned into the dual maltose binding protein affinity tag vector as described (*101*). The primers used to construct the 1-355Fip1 by insertion of a stop codon are: 5'CAATTTTAGCAAATAACCTCCGTTTTTCCCTCCAGG3' and 5'CCTGGAGGGAAAGGATATTTGCTAAAATTG3'. Construction of the dual MBP vector required the primers 5'GGGGACCACTTTGTACAAGAAAGCTGGGTTATTATTCTGCAGGTGTAGCTTCGGA3', 5'GAGAACCTGTACTTCCAGGGTATGTCGGCCGGCGAGGTC3', 5'GGGGACAAGTTTGTACAAA3' and 5'TCCAGGTGCATCAAGGTCTAC3'. The tryptophan to glycine point mutation at amino acid 293 in Fip1 was corrected with the primers 5'GCCAATTCAAGCGTTGGGAAGTGGCAGGATCG3' and 5'CGATCCTGCCACTTCCCAACGCTTGAATTGGC3'. Lastly, the duet vector was constructed with the following primers for Fip1: 5'TGATATCAATGTCGGCCGGCGAG3' and 5'GTGGTACCTTATTTGCTAAAATTGTTGTC3' and for PAP513 5'ATGGTCGACGATGGCTCACCAT3' and 5'CGACTAGTCTAGACGCCTTCTGTTGAATG3'. The MBP fusion and duet proteins were transformed into Rosetta DE3 pLysS cells (Novagen), lysed in 20 mM Tris-HCl pH

8.0, 200 mM NaCl, and 25 mM Imidazole pH 7.9, and purified by Nickel-NTA column (Qiagen). The eluted fractions were dialyzed into 25 mM Tris-HCl pH 7.5, 100 mM KCl, and 10% (v/v) glycerol and subsequently cleaved by TEV protease at a 1:20 ratio and loaded on a second Nickel-NTA column. Lastly, Fip1 was purified on an anion exchange mono Q column by increasing ionic strength from a start buffer of 25 mM Tris-HCl pH 7.5, 100 mM KCl, and 10% (v/v) glycerol.

The Clp1 hexahistidine affinity tagged construct was a gift from Dr. W. Keller (*93*). The Clp1 cDNA was cloned into a dual MBP vector by Justin Meyette as previously described (*69*). The Clp1 baculovirus was a gift from Dr. G. Gilmartin and expressed in SF9 insect cells (Gibco). All Clp1 constructs were lysed in 50 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM PMSF, and protease inhibitor tablet (Roche), centrifuged at 10,000 x g and purified using Ni-NTA affinity resin (Qiagen). The Clp1 protein was eluted with a step gradient of increasing imidazole concentration from 50, 100, 250, and 500mM imidazole. The eluted protein fractions containing Clp1 were dialyzed into 0.1 M Hepes pH 7.0, 150 mM NaCl, and 1 mM DTT or 20 mM Tris-HCl pH 7.8, 150 mM NaCl, and 0.5 mM TCEP. To form the complex Clp1 and CF $I_m25$ were incubated in a 1:2 ratio on ice with a 10-fold excess of ATP and concentrated to approximately 2.5 mg/mL with an Amicon Ultra 4 (Millipore).

*RNA preparation*

The DNA oligonucleotides used to synthesize the 60mer SELEX RNA is described in (*27*). The remaining primers used to generate the 40mer SELEX, PAP alpha and PAP gamma sequences are: 5'GGCGCGCGGTATACATTA 3', 5' ATTTAGGTGACACTATAGACATGATGTATGTAGTGTCT 3', 5' ATTTAGGTGACACTATAGCTGTACTCCTTGTATTTTT 3', ATTTAGGTGACACTATAGTTGTAAACAGATGATGTATT 3', 5' ATTTAGGTGACACATATAGACCATGTAATTCTTGTAATA 3', and 5' ATTTAGGTGACACTATAGTGTAAATGAAGTCGTATGAT 3'. Each oligonucleotide was annealed and transcribed by T7 RNA polymerase (Epicentre), and the transcripts were gel purified.

*Electromobility Shift Assays*

Gel shift assays were performed as previously described in 25 mM or 50 mM KCl Buffer D (20 mM HEPES pH 7.9, 10% (v/v) glycerol, 0.2 mM EDTA, 0.5 mM DTT, 0.1 mM PMSF) and may also include $^{32}$P ATP (*27*). The RNA-protein complexes were resolved on a 3% non-denaturing polyacrylamide gel at 4°C. The $^{32}$P ATP -protein complexes were resolved on a 6% non-denaturing polyacrylamide gel at 4°C.

*Oligomeric State Determination*

Size exclusion chromatography was performed with a Superdex 75 column and a Superdex 200 column (GE Healthcare). The protein sample or molecular mass standards were applied to the columns as previously described (*97*). Dynamic Light Scattering experiments were conducted at a protein concentration ranging from 0.3 – 0.4 mg/mL over the temperature range of 5°C - 25°C (Dynapro, Wyatt). At each temperature point three sets of 40 measurements were made to accurately measure the molecular weight and polydispersity. The standard protein buffer consisted of 10% (v/v) glycerol, 50 mM KCl, 20 mM Tris-HCl pH 7.5 - 8.0. Adjustments to this buffer according to polydispersity included CHES pH 9.0, Citric Acid pH 5.0, HEPES pH 7.0, 0–15% glycerol, 50 – 200 mM KCl or NaCl, and addition of UGUAN RNA, $Ap_4A$, or ATP.

# CHAPTER 4:

## Future Directions and Conclusions

## Introduction

The maturation of the pre-mRNA 3'-end appears deceptively simplistic when broken down into stepwise events entailing endonucleolytic cleavage followed by poly(A) addition. However, the complexity of the pre-mRNA 3'-end processing mechanism becomes apparent with the multiple factors involved and their established link to steps preceding and following, cleavage and polyadenylation. Factors recruited early to the pre-mRNA affect the efficiency of downstream processing events and may also be involved in multiple stages of processing not yet established, but suggested by biochemical data. It, therefore, becomes important to decipher the protein interfaces necessary for coupling these interactions and provide insight into how cleavage and polyadenylation factors are coordinated on the RNA and within the processing machinery. This can be addressed crystallographically by obtaining complex structures to define the regions necessary to facilitate protein-protein or protein-RNA interactions and additionally can aid in establishing new recognition surfaces. The crystallographic investigation of the protein complexes presented can be used to reevaluate the minimal interactions necessary for complex association and readdress suitable complex candidates.

### *Cleavage Factor $I_m$*

The pursuit of a CF $I_m$ complex structure was not successful due to degradation of the C-terminal region of the 68 kDa subunit and intrinsic flexible regions as suggested by disorder prediction plots. In order to further investigate a CF $I_m$ complex, improvement

to the expression and purification method of the 68 kDa truncation mutants is necessary. This would involve recloning into a more efficient vector, such as the His-MBP dual vector or a minimal hexahistidine tagged system. The involvement of CF $I_m$25 in mitotic spindle assembly separate from the 68 kDa subunit demonstrates that CF $I_m$25 is involved in other complexes and functions outside of pre-mRNA 3'-end processing (*53*). Understanding what mechanisms drive the interaction between CF $I_m$25 and CF $I_m$68 will be important in successfully readdressing a crystal complex.

### *Alternative methods*

Pull-down assays have demonstrated that the 68 kDa RRM region is sufficient in maintaining an interaction with the entire CF $I_m$25 protein (*41*). It would be worthwhile to investigate the minimal interaction between CF $I_m$25 and the truncation mutants of CF $I_m$68 via limited proteolysis experiments. The limited proteolysis assays will determine if there are additional flexible regions not stabilized through complex interactions and can be determined by additional bands on an SDS-PAGE following proteolytic cleavage. The exact position of proteolytic digestion can be determined by mass spectrometry. This information can then be used to reconstruct minimal expression vectors for both the 25 kDa and 68 kDa subunits and improve the complex crystallization process to eliminate potential flexible regions that prevent improper crystal packing interactions.

Investigating the affinity of complex formation between the 25 kDa and 68 kDa subunit can be defined by complex association at increasing salt concentrations. This

would suggest whether these subunits are suitable candidates for crystallization and can be addressed using pull down assays as previously described (*41*). Intolerance to high concentrations of salt would suggest a weak interaction between the two subunits and may result from a requirement of additional binding partners.

### *Determining more suitable candidates*

The data presented in Chapter 3 suggests the interactions of various pre-mRNA 3'-end processing factors described biochemically may require additional stabilizing factors to allow for crystallographic investigation. However, alternative binding partners capable of forming a complex with CF $I_m$25 can be investigated by pulldown assays using nuclear extract. These experiments have been preformed previously by the Gilmartin lab and demonstrated the association of CF $I_m$25 with chromatin remodeling factors (Brown, unpublished data). This experiment can be repeated with increasing amounts of salt to determine a tight binding complex. In addition, it would be necessary to determine if the interactions are direct interactions between CF $I_m$25 and the associated protein. This can partially be addressed by RNase treatment of the complex fraction. Loss of complex association following RNase treatment suggests RNA is necessary to facilitate binding. The direct association with CF $I_m$25 can then only be determined by direct interaction with the recombinant protein by pull-down, dynamic light scattering, or gel filtration experiments.

A brute force method can also be applied to obtain a complex crystal structure. This would involve combining all the expressed and purified proteins mentioned in

Chapter 3 and setting up crystallization trails under the assumption that the most likely candidates to interact will crystallize as a complex. This might prove to identify the minimal components necessary to form a complex.

### *Is CF $I_m$25's nucleotide binding affinity necessary for its function?*

The results from Chapter 2 demonstrate that CF $I_m$25 is able to bind to $Ap_4A$ and additionally to ATP. However, the nucleotide binding potential of CF $I_m$25 does not have an affect on *in vitro* cleavage and polyadenylation assays. To determine the necessity of nucleotide binding for the function of CF $I_m$25 *in vitro*, residues residing within the binding pocket such as arginine 63 (R63), arginine 150 (R150), and lysine 172 (K172) can be mutated to methionines (M) to disrupt the hydrogen binding potential with the β and γ phosphates and prevent nucleotide binding. Initial *in vitro* assays of the recombinant mutant can be used to verify an inability to bind nucleotides by steady state fluorescence experiments. Following confirmation, the mutant CF $I_m$25 can be used to reconstitute wild type CF $I_m$25 in CF $I_m$25 knockdown cells. Knockdown of endogenous CF $I_m$25 can be achieved by siRNA targeting in Hela cells and subsequently reconstituted with the mutant plasmid under a constitutive promoter. Changes affecting pre-mRNA 3'-end processing can be assayed between wild type and mutant CF $I_m$25 by sequencing of the two poly(A) polymerases (PAPOLG and PAPOLA) mRNA pools in Hela nuclear extracts, to reflect variation in the total amount of polyadenylated mRNA of a noncanonical and canonical poly(A) site. Previous work has demonstrated that CF $I_m$ is necessary for efficient processing of pre-mRNA containing noncanonical hexamers, in

the case of PAPOLG, and additionally enhances the 3'-end processing of canonical pre-mRNA, PAPOLA (*28*). In this manner the importance of nucleotide binding to pre-mRNA 3'-end processing *in vitro* can be further assessed.

Chapter 2 establishes the binding affinity of CF $I_m25$ to ATP and $Ap_4A$ and demonstrates the addition of $Ap_4A$ has no affect on *in vitro* cleavage and polyadenylation assays. Henceforth, the role of the Nudix domain of CF $I_m25$ in nucleotide binding is unclear. The conserved Nudix domain of CF $I_m25$ may be important for RNA binding and/or recognition. This can be addressed partially by the experiment described above. If variations in 3' processed canonical and noncanonical pre-mRNA are apparent this may be due to the inability of CF $I_m25$ to bind and recognize UGUAN containing RNAs. The importance of these residues, R63, R150 and K172, to facilitate RNA binding can then be addressed by gel shift assays using the SELEX 1 60mer RNA as previously described (*27*). The SELEX 1 RNA contains the UGUAN element recognized by CF $I_m$ and capable of forming a complex by gel shift. To test this hypothesis mutation of key residues in the CF $I_m25$ subunit of the CF $I_m$ dual baculovirus vector need to be constructed. Investigating the ability of the mutant CF $I_m25$ to bind the SELEX RNA would also suggest any changes in the ability of CF $I_m$ to regulate poly(A) site selection. Additionally, it would be interesting to see if the proteins CF $I_m25$ associates with via pull-down assays change with respect to its nucleotide binding potential.

## *Unresolved Questions*

CF $I_m25$ is a dimer in both the crystal structure and in solution as presented in Chapter 2. Due to the large dimer interface, 2,700 $Å^2$, between CF $I_m25$ monomers it is expected CF $I_m25$ remains a dimer when associated with other pre-mRNA 3'-end processing factors however, confirmation of the oligomeric state of the CF $I_m$ complex has yet to be determined. Estimating the stoichiometry of the CF $I_m$ subunits by DLS and gel filtration using the 68 kDa truncation mutants with CF $I_m25$ was not successful and could be reevaluated by analytical ultracentrifugation. Analytical ultracentrifugation could estimate the sample portion involved in complex formation and portion that is self aggregating. Knowing the stoichiometry of the CF $I_m$ complex provides information for setting up future crystallization trials and improves the success rate of obtaining a complex crystal. In doing so, the assumption can be made that the CF $I_m25$ protein maintains the same oligomeric state with other binding partners which again can be helpful in future crystallization attempts.

### *Cleavage Factor $I_m25$ and Poly(A) polymerase*

Evidence for the interaction of CF $I_m25$ with PAP513 was suggested by a supershift upon the addition of CF $I_m25$ in a gel shift assay presented in Chapter 3. This experiment alone does not suffice in establishing a tight interaction between both proteins and must be repeated. The PAP domains necessary to maintain an interaction with CF $I_m25$ is disputed in the literature and presented both as requiring the C-terminal region of PAP and maintaining the interaction despite a truncation to the C-terminal region of PAP. Our evidence so far does not conclusively clarify this issue. It is apparent that PAP513

alone with CF $I_m25$ does not demonstrate on a gel filtration column what would be expected of a tightly associating complex as mentioned in Chapter 3. We have shown CF $I_m25$ and PAP513 elute off of a gel filtration column at their individual, although overlapping, estimated molecular weights. To better address a complex between CF $I_m25$ and PAP513 the stoichiometry of the complex needs to be determined by ultracentrifugation methods. In addition, the direct interaction of CF $I_m25$ and PAP513 needs to be confirmed with the recombinant proteins, which has not been conclusive so far. The interaction could require the addition of the 68 kDa truncation mutant, RNA, Fip1 or ATP for efficient binding. Evidence for potential binding interactions and complex association has been demonstrated comprising CF $I_m$, PAP, and Fip1 in the context of a poly(A) site selection of noncanonical hexamer containing RNA (*28*). Further investigation of the minimal PAP513 and CF $I_m25$ complex can be addressed by increasing the 3'-end processing components until a complex molecular weight is achieved on a gel filtration column or confirmed by pull down assay. Upon confirmation of a complex, limited proteolysis experiments will confirm the stability of the complex and the minimal fragments necessary for subsequent crystallization techniques. Based on previous results the expression and purification of Fip1 was problematic but may be stabilized upon the addition of the 25 kDa subunit along with PAP513.

*Factor Interacting with Pap1*

The pursuit of a Fip1 crystal structure both alone and in complex with PAP513 was unsuccessful due to proteolytic degradation of the flexible regions of Fip1. A minimal Fip1 complex with PAP513 can be achieved by similar means as the recent complex of the yeast Pap1 and Fip1 fragment. The sequence of Fip1 is highly divergent from species to species except for the conserved region comprising residues 135-204 (based on the human sequence). The minimal interaction between the human homologues has been predicted to be maintained by amino acids 1-111 of Fip1 and additionally, full-length Fip1 interacts with the RBD (residues 366-513) and N-terminal (53-214) of PAP by pull-down experiments (*44*). Increased binding to PAP, in comparison to residues 1-111 of Fip1, is achieved with residues 1-355 of Fip1 as previously mentioned and presented in Chapter 3. Reconstitution of the complex comprising fragment 1-111 of Fip1 and PAP513 can then be subjected to limited proteolysis to determine if a smaller fragment of either protein can maintain complex interactions

### *Alternative methods*

Surface entropy reduction can be applied to optimize the crystallization techniques of any above mentioned protein complexes. This technique is based on minimizing high entropy surface residues such as lysines, glutamines, and glutamates that require more energy to bury these residues by forced crystal-crystal contacts (*104*). These surface residues can be predicted by the web-tool SERp and suggest more suitable, low entropy residues in their place (*105*). SER predictions on N68 suggest residues 20-23 (sequence EEFN**QE**A**E**), 138-139 (EASS**KK)**, and residue 146**K** are high entropy

residues and can be substituted by an alanine to improve crystallization odds. Additionally, Fip1 1-355 shows several high entropy clusters at residue ranges 21-23 (**EEE),** 245-247 (**EKE**TA), and 52-54 (ENEVERP**EEE**NASA). This technique may prove useful in future crystallization trials of difficult protein complexes following adequate optimization of protein expression and purification procedures.

# REFERENCES


(1) Bird, G., Zorio, D. A., and Bentley, D. L. (2004) RNA polymerase II carboxy-terminal domain phosphorylation is required for cotranscriptional pre-mRNA splicing and 3'-end formation. *Mol Cell Biol 24*, 8963-9.

(2) Maniatis, T., and Reed, R. (2002) An extensive network of coupling among gene expression machines. *Nature 416*, 499-506.

(3) Calvo, O., and Manley, J. L. (2003) Strange bedfellows: polyadenylation factors at the promoter. *Genes Dev 17*, 1321-7.

(4) Komarnitsky, P., Cho, E. J., and Buratowski, S. (2000) Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev 14*, 2452-60.

(5) Bentley, D. L. (2005) Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr Opin Cell Biol 17*, 251-6.

(6) Proudfoot, N. J., Furger, A., and Dye, M. J. (2002) Integrating mRNA processing with transcription. *Cell 108*, 501-12.

(7) Cougot, N., van Dijk, E., Babajko, S., and Seraphin, B. (2004) 'Cap-tabolism'. *Trends Biochem Sci 29*, 436-44.

(8) Millhouse, S., and Manley, J. L. (2005) The C-terminal domain of RNA polymerase II functions as a phosphorylation-dependent splicing activator in a heterologous protein. *Mol Cell Biol 25*, 533-44.

(9) Morris, D. P., and Greenleaf, A. L. (2000) The splicing factor, Prp40, binds the phosphorylated carboxyl-terminal domain of RNA polymerase II. *J Biol Chem 275*, 39935-43.

(10) Graveley, B. R. (2000) Sorting out the complexity of SR protein functions. *Rna 6*, 1197-211.

(11) Manley, J. L., and Tacke, R. (1996) SR proteins and splicing control. *Genes Dev 10*, 1569-79.

(12) 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*, 4854-64.

(13) Zhao, J., Hyman, L., and Moore, C. (1999) Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev 63*, 405-45.

(14) Wahle, E., and Rüegsegger, U. (1999) 3' End processing of pre-mRNA in eukaryotes. *FEMS Microbiol Rev 23*, 277-95.

(15) Minvielle-Sebastia, L., and Keller, W. (1999) mRNA polyadenylation and its coupling to other RNA processing reactions and to transcription. *Curr Opin Cell Biol 11*, 352-7.

(16) Mandel, C. R., Bai, Y., and Tong, L. (2008) Protein factors in pre-mRNA 3'-end processing. *Cell Mol Life Sci 65*, 1099-122.

(17) Keller, W., and Minvielle-Sebastia, L. (1997) A comparison of mammalian and yeast pre-mRNA 3'-end processing. *Curr Opin Cell Biol 9*, 329-36.

(18) Dichtl, B., and Keller, W. (2001) Recognition of polyadenylation sites in yeast pre-mRNAs by cleavage and polyadenylation factor. *Embo J 20*, 3197-209.

(19) Licatalosi, D. D., Geiger, G., Minet, M., Schroeder, S., Cilli, K., McNeil, J. B., and Bentley, D. L. (2002) Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II. *Mol Cell 9*, 1101-11.

(20) Noble, C. G., Beuth, B., and Taylor, I. A. (2007) Structure of a nucleotide-bound Clp1-Pcf11 polyadenylation factor. *Nucleic Acids Res 35*, 87-99.

(21) Perez-Canadillas, J. M. (2006) Grabbing the message: structural basis of mRNA 3'UTR recognition by Hrp1. *Embo J 25*, 3167-78.

(22) Meinke, G., Ezeokonkwo, C., Balbo, P., Stafford, W., Moore, C., and Bohm, A. (2008) Structure of yeast poly(A) polymerase in complex with a peptide from Fip1, an intrinsically disordered protein. *Biochemistry 47*, 6859-69.

(23) Li, X., Romero, P., Rani, M., Dunker, A. K., and Obradovic, Z. (1999) Predicting Protein Disorder for N-, C-, and Internal Regions. *Genome Inform Ser Workshop Genome Inform 10*, 30-40.

(24) Balbo, P. B., Toth, J., and Bohm, A. (2007) X-ray crystallographic and steady state fluorescence characterization of the protein dynamics of yeast polyadenylate polymerase. *J Mol Biol 366*, 1401-15.

(25) Balbo, P. B., and Bohm, A. (2007) Mechanism of poly(A) polymerase: structure of the enzyme-MgATP-RNA ternary complex and kinetic analysis. *Structure 15*, 1117-31.

(26) Bard, J., Zhelkovsky, A. M., Helmling, S., Earnest, T. N., Moore, C. L., and Bohm, A. (2000) Structure of yeast poly(A) polymerase alone and in complex with 3'-dATP. *Science 289*, 1346-9.

(27) Brown, K. M., and Gilmartin, G. M. (2003) A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor Im. *Mol Cell 12*, 1467-76.

(28) Venkataraman, K., Brown, K. M., and Gilmartin, G. M. (2005) Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes Dev 19*, 1315-27.

(29) Higgs, D. R., Goodbourn, S. E., Lamb, J., Clegg, J. B., Weatherall, D. J., and Proudfoot, N. J. (1983) Alpha-thalassaemia caused by a polyadenylation signal mutation. *Nature 306*, 398-400.

(30) Flaherty, S. M., Fortes, P., Izaurralde, E., Mattaj, I. W., and Gilmartin, G. M. (1997) Participation of the nuclear cap binding complex in pre-mRNA 3' processing. *Proc Natl Acad Sci U S A 94*, 11893-8.

(31) Hirose, Y., and Manley, J. L. (1998) RNA polymerase II is an essential mRNA polyadenylation factor. *Nature 395*, 93-6.

(32) Murthy, K. G., and Manley, J. L. (1992) Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus. *J Biol Chem 267*, 14804-11.

(33) Gilmartin, G. M., Fleming, E. S., Oetjen, J., and Graveley, B. R. (1995) CPSF recognition of an HIV-1 mRNA 3'-processing enhancer: multiple sequence contacts involved in poly(A) site definition. *Genes Dev 9*, 72-83.

(34) Legrand, P., Pinaud, N., Minvielle-Sebastia, L., and Fribourg, S. (2007) The structure of the CstF-77 homodimer provides insights into CstF assembly. *Nucleic Acids Res 35*, 4515-22.

(35) Bai, Y., Auperin, T. C., Chou, C. Y., Chang, G. G., Manley, J. L., and Tong, L. (2007) Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors. *Mol Cell 25*, 863-75.

(36) Qu, X., Perez-Canadillas, J. M., Agrawal, S., De Baecke, J., Cheng, H., Varani, G., and Moore, C. (2007) The C-terminal domains of vertebrate CstF-64 and its yeast orthologue Rna15 form a new structure critical for mRNA 3'-end processing. *J Biol Chem 282*, 2101-15.

(37) Perez Canadillas, J. M., and Varani, G. (2003) Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. *Embo J 22*, 2821-30.

(38) Kleiman, F. E., and Manley, J. L. (1999) Functional interaction of BRCA1-associated BARD1 with polyadenylation factor CstF-50. *Science 285*, 1576-9.

(39) Mirkin, N., Fonseca, D., Mohammed, S., Cevher, M. A., Manley, J. L., and Kleiman, F. E. (2008) The 3' processing factor CstF functions in the DNA repair response. *Nucleic Acids Res 36*, 1792-804.

(40) Rüegsegger, U., Beyer, K., and Keller, W. (1996) Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors. *J Biol Chem 271*, 6107-13.

(41) Dettwiler, S., Aringhieri, C., Cardinale, S., Keller, W., and Barabino, S. M. (2004) Distinct sequence motifs within the 68-kDa subunit of cleavage factor Im mediate RNA binding, protein-protein interactions, and subcellular localization. *J Biol Chem 279*, 35788-97.

(42) Mandel, C. R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J. L., and Tong, L. (2006) Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature 444*, 953-6.

(43) Kolev, N. G., Yario, T. A., Benson, E., and Steitz, J. A. (2008) Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3'-end maturation. *EMBO Rep 9*, 1013-8.

(44) Kaufmann, I., Martin, G., Friedlein, A., Langen, H., and Keller, W. (2004) Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *Embo J 23*, 616-26.

(45) Helmling, S., Zhelkovsky, A., and Moore, C. L. (2001) Fip1 regulates the activity of Poly(A) polymerase through multiple interactions. *Mol Cell Biol 21*, 2026-37.

(46) Shimazu, T., Horinouchi, S., and Yoshida, M. (2007) Multiple histone deacetylases and the CREB-binding protein regulate pre-mRNA 3' end processing. *J Biol Chem 282*, 4470-8.

(47) Wahle, E. (1991) Purification and characterization of a mammalian polyadenylate polymerase involved in the 3' end processing of messenger RNA precursors. *J Biol Chem 266*, 3131-9.

(48) Martin, G., Moglich, A., Keller, W., and Doublie, S. (2004) Biochemical and structural insights into substrate binding and catalytic mechanism of mammalian poly(A) polymerase. *J Mol Biol 341*, 911-25.

(49) Kerwitz, Y., Kuhn, U., Lilie, H., Knoth, A., Scheuermann, T., Friedrich, H., Schwarz, E., and Wahle, E. (2003) Stimulation of poly(A) polymerase through a direct interaction with the nuclear poly(A) binding protein allosterically regulated by RNA. *Embo J 22*, 3705-14.

(50) Martin, G., Keller, W., and Doublie, S. (2000) Crystal structure of mammalian poly(A) polymerase in complex with an analog of ATP. *Embo J 19*, 4193-203.

(51) Rüegsegger, U., Blank, D., and Keller, W. (1998) Human pre-mRNA cleavage factor Im is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits. *Mol Cell 1*, 243-53.

(52) Zhou, Z., Licklider, L. J., Gygi, S. P., and Reed, R. (2002) Comprehensive proteomic analysis of the human spliceosome. *Nature 419*, 182-5.

(53) Blower, M. D., Nachury, M., Heald, R., and Weis, K. (2005) A Rae1-containing ribonucleoprotein complex is required for mitotic spindle assembly. *Cell 121*, 223-34.

(54) Bessman, M. J., Frick, D. N., and O'Handley, S. F. (1996) The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed, "housecleaning" enzymes. *J Biol Chem 271*, 25059-62.

(55) Mildvan, A. S., Xia, Z., Azurmendi, H. F., Saraswat, V., Legler, P. M., Massiah, M. A., Gabelli, S. B., Bianchet, M. A., Kang, L. W., and Amzel, L. M. (2005) Structures and mechanisms of Nudix hydrolases. *Arch Biochem Biophys 433*, 129-43.

(56) McLennan, A. G. (2006) The Nudix hydrolase superfamily. *Cell Mol Life Sci 63*, 123-43.

(57) Bailey, S., Sedelnikova, S. E., Blackburn, G. M., Abdelghany, H. M., Baker, P. J., McLennan, A. G., and Rafferty, J. B. (2002) The crystal structure of diadenosine tetraphosphate hydrolase from Caenorhabditis elegans in free and binary complex forms. *Structure 10*, 589-600.

(58) Cohen, L. S., Mikhli, C., Jiao, X., Kiledjian, M., Kunkel, G., and Davis, R. E. (2005) Dcp2 Decaps m2,2,7GpppN-capped RNAs, and its activity is sequence and context dependent. *Mol Cell Biol 25*, 8779-91.

(59) Ranatunga, W., Hill, E. E., Mooster, J. L., Holbrook, E. L., Schulze-Gahmen, U., Xu, W., Bessman, M. J., Brenner, S. E., and Holbrook, S. R. (2004) Structural studies of the Nudix hydrolase DR1025 from *Deinococcus radiodurans* and its ligand complexes. *J Mol Biol 339*, 103-16.

(60) Frick, D. N., Townsend, B. D., and Bessman, M. J. (1995) A novel GDP-mannose mannosyl hydrolase shares homology with the MutT family of enzymes. *J Biol Chem 270*, 24086-91.

(61) Legler, P. M., Massiah, M. A., Bessman, M. J., and Mildvan, A. S. (2000) GDP-mannose mannosyl hydrolase catalyzes nucleophilic substitution at carbon, unlike all other Nudix hydrolases. *Biochemistry 39*, 8603-8.

(62) Ramirez, A., Shuman, S., and Schwer, B. (2008) Human RNA 5'-kinase (hClp1) can function as a tRNA splicing enzyme in vivo. *Rna 14*, 1737-45.

(63) Danckwardt, S., Hentze, M. W., and Kulozik, A. E. (2008) 3' end mRNA processing: molecular mechanisms and implications for health and disease. *Embo J 27*, 482-98.

(64) Kim, H., and Lee, Y. (2001) Interaction of poly(A) polymerase with the 25-kDa subunit of cleavage factor I. *Biochem Biophys Res Commun 289*, 513-8.

(65) Lutz, C. S., Murthy, K. G., Schek, N., O'Connor, J. P., Manley, J. L., and Alwine, J. C. (1996) Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro. *Genes Dev 10*, 325-37.

(66) Awasthi, S., and Alwine, J. C. (2003) Association of polyadenylation cleavage factor I with U1 snRNP. *Rna 9*, 1400-9.

(67) Zhang, R., Zhou, M., Moy, S., Collart, F., Joachimiak, A. . (2006) The 1.9A crystal structure of teh MutT/nudix family from Enterococcus faecalis.

(68) Gabelli, S. B., Bianchet, M. A., Azurmendi, H. F., Xia, Z., Sarawat, V., Mildvan, A. S., and Amzel, L. M. (2004) Structure and mechanism of GDP-mannose glycosyl hydrolase, a Nudix enzyme that cleaves at carbon instead of phosphorus. *Structure 12*, 927-35.

(69) Tropea, J. E., Cherry, S., Nallamsetty, S., Bignon, C., and Waugh, D. S. (2007) A Generic Method for the Production of Recombinant Proteins in *Escherichia coli* Using a Dual Hexahistidine-Maltose-Binding Protein Affinity Tag., in *Macromolecular Crystallography Protocols* (Doublié, S., Ed.) pp 1-20, Humana Press, Totowa.

(70) Doublié, S. (2007) Production of Selenomethionyl Proteins in Prokaryotic and Eukaryotic Systems, in *Macromolecular Crystallography Protocols* (Doublié, S., Ed.) pp 91-108, Humana Press, Totowa.

(71) Otwinowski, Z., and Minor, W. (1997) Processing of X-ray Diffraction Data Collected in Oscillation Mode, in *Macromolecular Crystallography, Part A* pp 307-326, Academic Press.

(72) Terwilliger, T. C. (2000) Maximum-likelihood density modification. *Acta Crystallogr D Biol Crystallogr 56*, 965-72.

(73) Vonrhein, C., Blanc, E., Roversi, P., and Bricogne, G. (2007) Automated Structure Solution with autoSHARP, in *Macromolecular Crystallography Protocols* (Doublié, S., Ed.) pp 215-230, Humana Press, Totowa.

(74) Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr 60*, 2126-32.

(75) Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr 54*, 905-21.

(76) Laskowski, R. A., McArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst 26*, 283-291.

(77) Rould, M. A. (2006) The same but different: isomorphous methods for phasing and high-throughput ligand screening. *Methods Mol Biol 364*, 159-82.

(78) DeLano, W. L. (2002), San Carlos, CA, USA.

(79) CCP4. (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr 50*, 760-3.

(80) Ames, B. N., and Dubin, D. T. (1960) The role of polyamines in the neutralization of bacteriophage deoxyribonucleic acid. *J Biol Chem 235*, 769-75.

(81) Yengo, C. M., Chrin, L. R., Rovner, A. S., and Berger, C. L. (2000) Tryptophan 512 is sensitive to conformational changes in the rigid relay loop of smooth muscle myosin during the MgATPase cycle. *J Biol Chem 275*, 25481-7.

(82) Rush, J., Moritz, A., Lee, K. A., Guo, A., Goss, V. L., Spek, E. J., Zhang, H., Zha, X. M., Polakiewicz, R. D., and Comb, M. J. (2005) Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat Biotechnol 23*, 94-101.

(83) Holm, L., and Sander, C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res 25*, 231-4.

(84) Yoshiba, S., Ooga, T., Nakagawa, N., Shibata, T., Inoue, Y., Yokoyama, S., Kuramitsu, S., and Masui, R. (2004) Structural insights into the *Thermus thermophilus* ADP-ribose pyrophosphatase mechanism via crystal structures with the bound substrate and metal. *J Biol Chem 279*, 37163-74.

(85) She, M., Decker, C. J., Chen, N., Tumati, S., Parker, R., and Song, H. (2006) Crystal structure and functional analysis of Dcp2p from *Schizosaccharomyces pombe*. *Nat Struct Mol Biol 13*, 63-70.

(86) Gabelli, S. B., Bianchet, M. A., Bessman, M. J., and Amzel, L. M. (2001) The structure of ADP-ribose pyrophosphatase reveals the structural basis for the versatility of the Nudix family. *Nat Struct Biol 8*, 467-72.

(87) Logan, D. T., Andersson, J., Sjoberg, B. M., and Nordlund, P. (1999) A glycyl radical site in the crystal structure of a class III ribonucleotide reductase. *Science 283*, 1499-504.

(88) Nicholls, A., Sharp, K. A., and Honig, B. (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins 11*, 281-96.

(89) Champagne, K. S., Sissler, M., Larrabee, Y., Doublie, S., and Francklyn, C. S. (2005) Activation of the hetero-octameric ATP phosphoribosyl transferase through subunit interface rearrangement by a tRNA synthetase paralog. *J Biol Chem 280*, 34096-104.

(90) Kisselev, L. L., Justesen, J., Wolfson, A. D., and Frolova, L. Y. (1998) Diadenosine oligophosphates (Ap(n)A), a novel class of signalling molecules? *FEBS Lett 427*, 157-63.

(91) Baxi, M. D., and Vishwanatha, J. K. (1995) Diadenosine polyphosphates: their biological and pharmacological significance. *J Pharmacol Toxicol Methods 33*, 121-8.

(92) Sillero, M. A., De Diego, A., Osorio, H., and Sillero, A. (2002) Dinucleoside polyphosphates stimulate the primer independent synthesis of poly(A) catalyzed by yeast poly(A) polymerase. *Eur J Biochem 269*, 5323-9.

(93) de Vries, H., Ruegsegger, U., Hubner, W., Friedlein, A., Langen, H., and Keller, W. (2000) Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. *Embo J 19*, 5895-904.

(94) Weitzer, S., and Martinez, J. (2007) The human RNA kinase hClp1 is active on 3' transfer RNA exons and short interfering RNAs. *Nature 447*, 222-6.

(95) Pearson, W. R., and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A 85*, 2444-8.

(96) Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*, 2577-637.

(97) Coseno, M., Martin, G., Berger, C., Gilmartin, G., Keller, W., and Doublie, S. (2008) Crystal structure of the 25 kDa subunit of human cleavage factor Im. *Nucleic Acids Res 36*, 3474-83.

(98) Shimizu, K., Hirose, S., and Noguchi, T. (2007) POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. *Bioinformatics 23*, 2337-8.

(99) Ryan, K., and Bauer, D. L. (2008) Finishing touches: post-translational modification of protein factors involved in mammalian pre-mRNA 3' end formation. *Int J Biochem Cell Biol 40*, 2384-96.

(100) Studier, F. W. (2005) Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif 41*, 207-34.

(101) Nallamsetty, S., Austin, B. P., Penrose, K. J., and Waugh, D. S. (2005) Gateway vectors for the production of combinatorially-tagged His6-MBP fusion proteins in the cytoplasm and periplasm of Escherichia coli. *Protein Sci 14*, 2964-71.

(102) Paushkin, S. V., Patel, M., Furia, B. S., Peltz, S. W., and Trotta, C. R. (2004) Identification of a human endonuclease complex reveals a link between tRNA splicing and pre-mRNA 3' end formation. *Cell 117*, 311-21.

(103) Walter, A. E., Turner, D. H., Kim, J., Lyttle, M. H., Muller, P., Mathews, D. H., and Zuker, M. (1994) Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad Sci U S A 91*, 9218-22.

(104) Cooper, D. R., Boczek, T., Grelewska, K., Pinkowska, M., Sikorska, M., Zawadzki, M., and Derewenda, Z. (2007) Protein crystallization by surface entropy reduction: optimization of the SER strategy. *Acta Crystallogr D Biol Crystallogr 63*, 636-45.

(105) Goldschmidt, L., Cooper, D. R., Derewenda, Z. S., and Eisenberg, D. (2007) Toward rational protein crystallization: A Web server for the design of crystallizable protein variants. *Protein Sci 16*, 1569-76.