Structural and Kinetic Characterization of RNA Polymerase II

C-Terminal Domain Phosphatase Ssu72

and Development of New Methods for NMR Studies of Large Proteins

by

Jonathan W. Werner-Allen

Department of Biochemistry
Duke University

Date:________________________

Approved:

___________________________

Pei Zhou, Supervisor

___________________________

John Cavanagh

___________________________

Arno L. Greenleaf

___________________________

Paul L. Modrich

___________________________

Leonard D. Spicer

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Biochemistry in the Graduate School
of Duke University

2011
ABSTRACT

Structural and Kinetic Characterization of RNA Polymerase II

C-Terminal Domain Phosphatase Ssu72

and Development of New Methods for NMR Studies of Large Proteins

by

Jonathan W. Werner-Allen

Department of Biochemistry
Duke University

Date:_______________________

Approved:

Pei Zhou, Supervisor

John Cavanagh

Arno L. Greenleaf

Paul L. Modrich

Leonard D. Spicer

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biochemistry in the Graduate School of Duke University

2011
Abstract

Ssu72 is a protein phosphatase that selectively targets phosphorylated serine residues at the 5th position (pS5) in the heptad repeats of the C-terminal domain (CTD) of RNA polymerase II, in order to regulate the CTD-mediated coupling between eukaryotic transcription and co-transcriptional events. The biological importance of Ssu72 is underscored by (1) the requirement of its activity for viability in yeast, and (2) the numerous phenotypes – affecting all three stages of the transcription cycle – that result from its mutation in yeast. Despite limited homology to the low molecular weight (LMW) subclass of protein tyrosine phosphatases (PTPs), several lines of evidence suggest that Ssu72 represents the founding member of a new class of enzymes, including its unique substrate specificity and an in vivo connection with the activity of proline isomerase Ess1.

The main focus of this work has been to structurally and kinetically characterize Ssu72, in order to define its relation to known enzyme families, to provide biochemical explanations for extant in vivo observations, and to allow future structure-guided investigations of its role in coordinating transcription with co-transcriptional events. To this end, we solved the structure of Ssu72 in complex with its pS5 CTD substrate, revealing an enzyme fold with unique structural features and a surprising substrate
conformation with the pS5-P6 motif of the CTD adopting the cis configuration. Together with kinetic assays, the structure provides a new interpretation of the role of proline isomers in regulating the CTD phosphorylation state, with broad implications for CTD biology.

The second goal of this dissertation has been to develop new methods for NMR studies of large proteins, which present unique challenges to conventional methods, including fast signal decay and severe signal degeneracy. The first of these new methods, the ‘just-in-time’ HN(CA)CO, improves the sensitivity of a common backbone assignment experiment. The next two methods, the 4-D diagonal-suppressed TROSY-NOESY-TROSY and the 4-D time-shared NOESY, were designed for use with sparse sampling techniques that allow the acquisition of high-resolution, high-dimensionality datasets. These efforts culminate with global fold calculations for large proteins, including the 23 kDa Ssu72, with accurate and unambiguous automated assignment of NOE crosspeaks. We expect that the methods presented here will be particularly useful as the NMR community continues to push toward higher molecular weight targets.
Contents

Abstract.................................................................................................................................iv

List of Tables............................................................................................................................x

List of Figures..........................................................................................................................xi

Acknowledgements..................................................................................................................xiii

1. Introduction to CTD Biology and Phosphatase Ssu72 ...................................................1
   1.1 The C-Terminal Domain of RNA Polymerase II.......................................................1
   1.2 The CTD Code...............................................................................................................3
      1.2.1 Phosphorylation States.........................................................................................3
      1.2.2 CTD Kinases and Phosphatases.........................................................................8
      1.2.3 Proline Isomers and Isomerases........................................................................11
   1.3 CTD Functions and Significance.................................................................................18
   1.4 The pS5 CTD Phosphatase Ssu72............................................................................26
      1.4.1 Role in the Transcription Cycle.........................................................................26
      1.4.2 Similarities to LMW PTPs....................................................................................29
      1.4.3 In Vivo Connection with Proline Isomerase Ess1.............................................33

2. Structural and Kinetic Studies of Phosphatase Ssu72 ...................................................37
   2.1 NMR Assignments and Catalytic Mutant Screening..............................................38
      2.1.1 Backbone and Sidechain Assignment.............................................................39
      2.1.2 Finding a Substrate-Trapping Mutant with NMR Titrations..........................41
2.2  Structure of Ssu72 in Complex with pS5 CTD Substrate ........................................46
2.2.1  Comparison with the LMW PTP Fold .................................................................48
2.2.2  Binding of a Cis Proline Substrate Conformation ............................................51
2.3  Ess1-Catalyzed Proline Isomerization Stimulates Ssu72 Activity ..........................54
2.3.1  Measurement of Substrate Proline Isomer Populations ....................................54
2.3.2  Ssu72 Phosphatase Assays with Ess1 .................................................................56
2.3.3  Reaction Timescale Affects Stimulation by Ess1 ..............................................59
2.4  Summary, Implications, and Future Directions ......................................................62
2.4.1  Substrate Specificity of Ssu72 ..............................................................................63
2.4.2  Regulation of Isomer-Specific Enzymes .............................................................64
2.4.3  Implications of Proline Isomerization in CTD Biology ......................................67
2.4.4  Comparison of Recent Ssu72 Structures .............................................................72
2.4.5  Proposals for Future Studies ................................................................................77
2.5  Materials and Methods ............................................................................................87
2.5.1  Molecular Cloning and Protein Expression and Purification .............................87
2.5.2  CTD Peptide Preparation and Characterization .................................................88
2.5.3  NMR ......................................................................................................................89
2.5.4  Crystallization and Structure Determination .......................................................90
2.5.5  Enzyme Kinetics with Peptide Substrate ............................................................91
2.5.6  Enzyme Kinetics with Full-Length Substrate ......................................................92
3.  Introduction to NMR Methods for Large Proteins ..................................................94
3.1  Challenges Posed by Large Proteins ....................................................................95
6.6.1 Sample Preparation................................................................. 160
6.6.2 Data Acquisition and Processing............................................. 161
6.6.3 Data Analysis........................................................................... 163
6.6.4 Structure Calculations.............................................................. 164
References......................................................................................... 168
Biography .......................................................................................... 188
List of Tables

Table 1. Crystallographic Data Collection and Refinement Statistics for the Ssu72-Substrate Complex ................................................................. 47

Table 2. Summary of Ssu72 Mutations Used in Prior Yeast Studies .................................................. 82

Table 3. Peak Assignment and Structure Determination Statistics for the CYANA Calculations of Ssu72 with TS NOESY Data ........................................................................ 157

Table 4. Peak Assignment and Structure Determination Statistics for the CYANA Calculations of HCAII with TS NOESY Data ................................................................. 158
List of Figures

Figure 1. The C-Terminal Domain of RNA Polymerase II ........................................... 2
Figure 2. CTD Phosphorylation States and the Transcription Cycle ............................... 4
Figure 3. Proline Isomerization .................................................................................. 12
Figure 4. Catalytic Mechanism of PTPs ....................................................................... 30
Figure 5. Assigned $^1$H/$^{15}$N-HSQC-TROSY Spectrum of Phosphate-Bound Drosophila Melanogaster Ssu72 ................................................................. 40
Figure 6. Phosphate Titrations of Active Site Mutants of Ssu72 ................................. 42
Figure 7. Phosphate- and Substrate-Induced Chemical Shift Perturbations for dSsu72 . 45
Figure 8. Crystals and Diffraction Data for the Ssu72-Substrate Complex .................... 46
Figure 9. Structure of pS5 CTD Phosphatase Ssu72 and Comparison with LMW PTP Lpt1 .............................................................................................................. 49
Figure 10. Binding of Cis Proline pCTD Substrate by Ssu72 ....................................... 52
Figure 11. Measurement of Proline Isomer Populations for Ssu72 Substrate ................ 55
Figure 12. Ess1 Stimulates pCTD Dephosphorylation by Ssu72 ................................. 57
Figure 13. Reaction Timescale Affects Stimulation of Ssu72 Activity by Ess1 .......... 60
Figure 14. A New Model for Ess1-Regulated Ssu72 Activity .................................... 68
Figure 15. Overlay of Apo Ssu72 Structures ................................................................. 73
Figure 16. Comparison of Ssu72 Structures .................................................................. 76
Figure 17. Structural Analysis of Ssu72 Mutations Used in Prior Yeast Studies .......... 83
Figure 18. Sparse Sampling Patterns ........................................................................... 102
Figure 19. Aliasing Artifacts and Removal Algorithms ................................................... 108
Figure 20. Pulse Sequence for the JIT TROSY-HN(CA)CO Experiment ......................... 118
Figure 21. Comparison of TROSY-HN(CA)CO Spectra .................................................. 121
Figure 22. Slices and Sensitivity Profiles of TROSY-HN(CA)CO Spectra ....................... 122
Figure 23. Pulse Sequence for the 4-D ds-TNT Experiment ............................................. 130
Figure 24. Diagonal Suppression with the ds-TNT Experiment ........................................ 131
Figure 25. Representative Planes from the 4-D ds-TNT Spectrum .................................. 134
Figure 26. Comparison of Peak Intensities in 3-D and 4-D ds-TNT Spectra ...................... 137
Figure 27. Comparison of 4-D ds-TNT Peaks with FFT and FFT-CLEAN Processing .... 138
Figure 28. Pulse Sequence for the 4-D TS NOESY Experiment .................................... 145
Figure 29. Artifact Removal in the 4-D TS NOESY Spectra for Ssu72 ............................ 148
Figure 30. Artifact Removal in the 4-D TS NOESY Spectra for HCAII ............................. 149
Figure 31. Peak Identification in the 4-D TS NOESY Spectra for Ssu72 ............................ 151
Figure 32. Peak Identification in the 4-D TS NOESY Spectra for HCAII ............................. 152
Figure 33. CYANA Structure Calculations Using TS NOESY Data ............................... 155
Acknowledgements

I still clearly remember receiving my first research objective as a graduate student, as delivered by my advisor, Pei Zhou: design a four-dimensional pulse sequence that collects amide-amide NOEs while suppressing diagonal signals. Having no prior experience writing pulse sequences, or with NMR at all for that matter, I knew I was in way over my head. Over the last seven years, a number of unfathomably patient, bright, and generous colleagues have helped me reach a level of acuity and understanding from which I can look back and think “That first project really wasn’t too hard.” Foremost is Pei, who has invested an enormous amount of time and energy mentoring my research endeavors, all the while prodding me to work hard, think clearly, and pursue new opportunities aggressively. No matter what path I choose from here, those habits will serve me well in the future, and I offer my heartfelt thanks.

Brian Coggins has been involved in nearly every stage of this dissertation, from providing early instruction on the basics of NMR theory to co-authoring our latest NMR methods manuscript. His contribution to sparse sampling methods and artifact removal algorithms is the platform from which our NMR methods work was launched. Moreover, I will miss our near daily discussions on politics, economics, theology, and, occasionally, science.
I’ve also been blessed to work and commiserate with a number of talented graduate students at Duke. Martha Bomar and I entered Pei’s lab at the same time, and endured the ups and downs of graduate research together. I worked closely with Jeffrey Boyles for three years, and benefited greatly from his knowledge of computer programming and NMR data analysis. When trying to understand frustratingly complicated kinetic problems, Louis Metzger offered his expertise in bleary-eyed, caffeine-driven, late-night talks. Chuljin Lee has tolerated my incessant questions about X-ray crystallography, given excellent advice at several critical junctures in this dissertation research, and celebrated with me during one of the most memorable moments in my graduate career: calculating the initial electron density map for the Ssu72-substrate complex and observing, for the first time, the cis proline substrate conformation.

I’ve had the pleasure of performing this work in the Zhou lab, alongside a dedicated and always entertaining band of researchers, who have created a supportive and stimulating environment that makes all the difference in the world. I thank Ling Jiang for teaching me how to operate the NMR spectrometers, Su Wang for performing some of the early kinetic assays with Ssu72, and Shan Jin for designing and characterizing Ssu72 catalytic mutants and extending our structural results toward in vivo studies in yeast. Additionally, none of the projects described in this dissertation would have been possible without the excellent NMR and X-ray crystallography
facilities at Duke University. The NMR Center staff, including Tony Ribeiro, Ron Venters, and Don Mika, deserves credit for providing timely access to their instruments and for working tirelessly to solve technical issues and instruct users. Nathan Nicely, at the X-ray Crystallography Shared Resource, was incredibly generous with his facility’s equipment and reagents and his time, helping with everything from robotic tray setting to diffraction collection.

I would also like to recognize the financial support I’ve received from the Henry Kamin fellowship, an award endowed by Dottie Kamin in remembrance of her late husband, a renowned biochemist and longtime professor in the Biochemistry Department at Duke. I like to think that, as an enzymologist, he too would have been excited by the unusual substrate specificity of Ssu72.

To steal a line from my brother’s dissertation: I couldn’t ask for better parents. They’ve supported me from the beginning of my education to this end, even making an effort to understand my graduate research and, of course, wondering why it’s taken so long to finish it. Their encouragement, along with the impressive academic achievements of my siblings Geoffrey and Amy, has continually inspired me to do my best.

Finally, to my fiancé, Cammy Contizano, who has stuck it out with me through five years of my graduate school career, always reminding me that there is a wild and wonderful world outside of the laboratory: thank you for everything.
1. Introduction to CTD Biology and Phosphatase Ssu72

Eukaryotic transcription is a highly complex biochemical process that is tightly coupled to many co-transcriptional events, like capping, splicing, and 3′-end processing of the nascent RNA, which are required to generate mature and functional transcripts. The central player coordinating these processes is the C-terminal domain (CTD) of Rpb1, the largest subunit of RNA polymerase II (RNAPII) (Buratowski 2009; Howe 2002; Phatnani and Greenleaf 2006).

1.1 The C-Terminal Domain of RNA Polymerase II

The CTD is a unique domain composed of a tandem array of heptad repeats with the consensus sequence Y1-S2-P3-T4-S5-P6-S7 (Figure 1) (Corden et al. 1985). The number of heptads varies across species and roughly correlates with genomic complexity, ranging from 26 repeats in yeast to 52 in humans (Allison et al. 1988; Corden 1990). The CTD is conserved among eukaryotes, and truncations result in a loss of viability in yeast (Nonet et al. 1987; Zehring et al. 1988). Structurally, the CTD has been shown to be largely disordered in solution (Noble et al. 2005), and although models have been proposed for higher order structure (Cagas and Corden 1995; Meinhart and
Cramer 2004), there is no evidence to date that the domain adopts any uniform, repetitive conformation. Indeed, the unique flexibility and structural plasticity of the CTD make it an ideal platform for interacting with numerous binding partners with distinct binding modes (Meinhart et al. 2005). In support of this model, previous structural studies of CTD-interacting proteins in complex with peptidyl CTD fragments show a variety of binding-induced CTD conformations (Fabrega et al. 2003; Meinhart and Cramer 2004; Verdecia et al. 2000). The length of the CTD is another important feature (Figure 1); in a fully-extended conformation, the CTD could reach 750-1500 Å.

**Figure 1. The C-Terminal Domain of RNA Polymerase II.** The CTD is shown in a fully-extended state to scale with the main body of the polymerase. Phosphorylation at the S2 and S5 positions of the consensus heptad repeat are the two best-characterized post-translational modifications.
from the main body of RNAPII (Meinhart et al. 2005), allowing the simultaneous interaction of multiple binding partners while they function on the template DNA or the nascent RNA transcript.

### 1.2 The CTD Code

In addition to the CTD’s length and structural plasticity, each heptad contains multiple sites for post-translational modification. The numerous possible patterns of these modifications within each repeat and distributions of patterns across the full domain create a ‘CTD code’ with a staggering potential complexity (Buratowski 2003). Over the last decade, a wealth of research has explored how the CTD code serves to link transcription with co-transcriptional processes in a temporal manner.

#### 1.2.1 Phosphorylation States

The timely recruitment of specific factors to the transcribing polymerase is mediated by CTD phosphorylation patterns that vary throughout the transcription cycle. The most-studied and best-characterized sites of phosphorylation are at the S2 and S5 positions of each heptad repeat (Figure 1). In the canonical model, RNAPII is recruited
to the promoter with an unphosphorylated CTD, and upon preinitiation complex formation, the CTD becomes heavily phosphorylated at the S5 position (pS5), attracting CTD-binding partners that influence initiation complex formation, mRNA capping, and the transition into elongation (Buratowski 2009). As the polymerase moves toward the 3’-end of genes, the level of pS5 declines while phosphorylation at S2 (pS2) increases, recruiting nuclear factors responsible for elongation, termination, and 3’-end formation.

Figure 2. CTD Phosphorylation States and the Transcription Cycle. In the canonical model, opposing gradients of pS5 and pS2 serve to recruit distinct sets of nuclear factors to the appropriate stage of the transcription cycle.
(Buratowski 2009). The CTD therefore undergoes two opposing gradients of phosphorylation, with pS5 and pS2 predominating early and late in the transcription cycle, respectively, and a mix including doubly-phosphorylated pS2/pS5 repeats in the center of the gene (Figure 2).

Recently, several studies have added layers of complexity to this simple model. First, the S7 position is now recognized as a functionally-relevant site of phosphorylation (pS7) (Chapman et al. 2007; Egloff et al. 2007). Although it appears dispensable for the expression of protein-coding genes, pS7 is required for the proper expression of small noncoding RNAs, presumably through recruitment of the Integrator complex (Egloff et al. 2007). Integrator is a large, multiprotein complex involved in small noncoding RNA 3’-end processing, and associates with the CTD by binding to a pS7/pS2 doubly-phosphorylated pattern (Baillat et al. 2005; Egloff et al. 2010). Despite this link to 3’-end processing, peak pS7 levels are generally observed at the 5’-end of genes, similar to pS5 (Mayer et al. 2010; Tietjen et al. 2010). The fate of pS7 in later stages of transcription appears to depend on the type of transcript, with high levels of pS7 observed throughout the length of protein-coding genes, but decreasing levels, similar to the pattern of pS5, observed toward the 3’-end of noncoding genes (Tietjen et al. 2010).

Phosphorylation is not the only covalent modification in the CTD code. Methylation of R1810 in the 31st heptad of human CTD was recently reported, and
appears to repress transcription of certain noncoding RNAs (Sims et al. 2011).

Glycosylation has also been reported (Kelly et al. 1993), although its biological significance remains unknown. Likewise, transcriptional functions have yet to be characterized for phosphorylation observed at other sites in the heptad repeats, including Y1 and T4 (Baskaran et al. 1993; Zhang and Corden 1991).

Finally, there are several indications that the CTD code is not uniform across the genome, with exceptions and variations depending on the class of gene and other factors. Three recent studies of genome-wide occupancy of different phosphorylated CTD (pCTD) patterns suggest that, overall, the canonical model holds for the majority of genes (Kim et al. 2010; Mayer et al. 2010; Tietjen et al. 2010), although peak pS2 levels are reached at ~600-1000 nucleotides downstream of the transcription start site, regardless of the gene length, as opposed to steadily increasing throughout transcription as implied by the gradient model (Kim et al. 2010; Mayer et al. 2010). However, numerous exceptions to this model were observed and appear to correlate with several variables including the class of gene (coding versus noncoding), gene length, promoter structure, and genetic region (intron versus exon). This added nuance is bolstered by the newly-discovered elements of the CTD code, pS7 and R1810 methylation, which are transcript-dependent modifications, affecting only the expression of noncoding RNAs (Egloff et al. 2007; Sims et al. 2011).
It is also important to note the limitations of our current understanding of the CTD code. The principle method used to characterize CTD phosphorylation states at different positions along template genes is chromatin immunoprecipitation (ChIP) with antibodies that recognize different heptad phosphorylation patterns. Although there are antibodies that target each major phosphorylation site (pS5 for example), their epitope recognition properties are in fact quite complicated, owing to phosphorylation at adjacent sites (pS2 and pS7) that can either enhance or inhibit binding (Chapman et al. 2007). Furthermore, these experiments provide only aggregate pictures of the CTD phosphorylation state that lack information about the density and distribution of phosphates along the length of the domain. For example, the phosphorylation pattern in the middle of coding genes may contain pS2/pS5/pS7 triply-phosphorylated heptads, or an evenly-distributed mix of singly-phosphorylated heptads. Alternatively, phosphorylation of specific sites may predominate at certain regions of the CTD. This is already known to be the case for pS7 in humans, which is naturally confined to the polymerase-proximal half of the CTD through the substitution of S7 by other amino acids in the distal repeats (Chapman et al. 2007).
1.2.2 CTD Kinases and Phosphatases

The CTD phosphorylation state is maintained by multiple kinases and phosphatases working in concert during the transcription cycle in a highly dynamic fashion. Underscoring this point, it was recently shown that promoter-distal pS7 marks are placed anew during transcription, rather than retained from earlier phosphorylation events (Tietjen et al. 2010).

CTD kinases function in the form of cyclin/cyclin-dependent kinase (CDK) pairs. The cyclins in these complexes are from a transcription-associated, non-cycling family, which are active throughout the cell cycle, and regulation of kinase activity is instead mediated by mechanisms like selective recruitment and binding to regulatory factors. The Kin28/Ccl1 pair (CDK7/CyclinH in metazoa) is part of the general transcription factor TFIIH and phosphorylates S5 and S7 during the formation of the preinitiation complex (Akhtar et al. 2009; Glover-Cutter et al. 2009; Kim et al. 2009; Komarnitsky et al. 2000; Schroeder et al. 2000). The Srb10/Srb11 pair (CDK8/CyclinC in metazoa) also phosphorylates S5 as part of the mediator complex and appears to be involved in the activation or repression of transcription, although its exact role remains a matter of debate (Galbraith et al. 2010; Meinhart et al. 2005). During elongation in *Saccharomyces cerevisiae*, phosphorylation of S2 is performed by two kinases: CTDK-I, which is
responsible for the majority of S2 phosphorylation in vivo (Lee and Greenleaf 1991; Sterner et al. 1995), and Bur1, which also has an elongation phase S7 kinase activity (Tietjen et al. 2010; Yao et al. 2000). Higher eukaryotes also have two S2 CTD kinases, P-TEFb and CDK12, which are the orthologues of Bur1 and CTDK-I, respectively (Bartkowiak and Greenleaf 2011). P-TEFb is composed of CDK9 and cyclinT (Marshall et al. 1996; Marshall and Price 1995), while the cyclin partner of CDK12 has not been firmly established (Bartkowiak et al. 2010; Chen et al. 2006; Chen et al. 2007).

Two CTD phosphatases, Fcp1 and Ssu72, have been discovered in yeast. Fcp1 is a conserved and essential phosphatase with preferential activity for pS2, and associates with the polymerase throughout the coding region of genes, presumably contributing to the gradual nature of increasing pS2 levels during elongation (Cho et al. 2001). In support of this point, Fcp1 mutants show elevated pS2 levels in coding regions (Cho et al. 2001). Fcp1 also plays a major role in recycling the CTD to an unphosphorylated state after transcription (Cho et al. 1999). Structurally, Fcp1 is a Y-shaped protein containing an N-terminal catalytic domain, a C-terminal breast cancer protein related C-terminal (BRCT) domain, and a helical insert domain (Ghosh et al. 2008). Fcp1 belongs to a superfamily of metal-dependent phosphotransferases that feature a signature DxDxT catalytic motif, in which the first aspartate acts as a nucleophile to attack the substrate phosphorous atom (Collet et al. 1998). Higher eukaryotes also have small CTD...
phosphatases (SCPs) that contain the Fcp1 catalytic domain but lack the BRCT domain (Yeo et al. 2003; Zhang et al. 2006). Unlike Fcp1, Scp1 has preferential activity toward pS5, and appears to function mainly as a suppressor of neuronal gene expression in nonneuronal tissue (Yeo et al. 2005). Plants have a similar set of enzymes, the CTD phosphatase-like proteins (CPLs), which also lack a BRCT domain, but display a strict specificity for the pS5 position (Koiwa et al. 2004).

Until recently, Ssu72 was the only known pS5-specific phosphatase in yeast. The biological and kinetic characteristics of Ssu72 are discussed in detail in Section 1.4. Briefly, Ssu72 is a component of the cleavage and polyadenylation factor (CPF) and, in ChIP assays, crosslinks predominantly to the 3’-end of genes (Ansari and Hampsey 2005; Nedea et al. 2003). Mutations of Ssu72 cause defective 3’-end processing as well as improper termination for a subset of RNA transcripts (Ganem et al. 2003; He et al. 2003; Nedea et al. 2003). The concentration of these effects on late stage transcriptional events suggests that there may be an additional phosphatase that facilitates the decrease in pS5 levels during elongation. The recently-discovered Rtr1 protein may fill that role. Rtr1 was initially connected to RNAPII and transcription through genetic interactions, and was later shown to interact directly with the CTD, binding preferentially to the pS5 form (Gibney et al. 2008; Mosley et al. 2009). ChIP assays show that Rtr1 localizes to the coding region of genes, and its deletion causes elevated pS5 levels to persist during
elongation (Mosley et al. 2009). Despite these provocative results, Rtr1 appears to be only weakly active toward full-length pCTD substrates in vitro (Mosley et al. 2009). Therefore, further biochemical evidence will be required to fully establish its role in transcription.

1.2.3 Proline Isomers and Isomerases

In addition to well-studied modifications like phosphorylation, the CTD code contains a second, noncovalent element for regulating the association of CTD-binding partners: proline isomers (Buratowski 2003; Lu et al. 2007). The peptide bond linking adjacent residues in a protein can adopt either the trans or cis conformation, which are related by a 180° rotation about the central N-C’ bond of the peptide plane, creating distinct and interconvertible backbone configurations (Figure 3). For all amino acids except proline, the trans isomer minimizes backbone steric repulsion and is far more energetically favorable. For proline, the unique, backbone-linked ring structure of its sidechain causes the cis and trans isomers to be closer in free energy, making the cis conformation thermodynamically feasible. In folded proteins, prolines are generally constrained to a single isomer state by the native structure, with the cis isomer occurring with a frequency of 5-6% (Pal and Chakrabarti 1999; Stewart et al. 1990); however, there
Figure 3. Proline Isomerization. Proline residues equilibrate slowly between cis and trans isomers with the cis isomer being less populated. The two isomers are represented here for the S5-P6 motif in a model CTD peptide with the central N-C’ bond highlighted in green. The proline isomerase Ess1/Pin1 catalyzes the interconversion of isomers in the CTD.

are several examples of structures in which populations of both isomers are observed for specific proline residues (Andreotti 2003). In short and unstructured peptides, the population of cis proline ranges from 5 to 40%, depending on the adjacent residues (Reimer et al. 1998).

Importantly, proline isomerization is an intrinsically slow interconversion that can be enzymatically regulated (Figure 3). Activation barriers for cis-to-trans conversion
have been measured at ~20 kcal/mol, with cis-to-trans conversion rates on the order of \( s^{-1} \) to min\(^{-1} \) (Grathwohl and Wuthrich 1981; Yaffe et al. 1997; Zhou et al. 2000), and given the lower energy of the trans isomer, the activation barrier and rate of the reverse conversion would be even higher and slower, respectively. Relative to the short timescale of cellular events, the uncatalyzed exchange of isomers effectively creates two structurally-distinct and kinetically-isolated epitopes, and in the case of isomer-specific biological processes, it would take many minutes for equilibrium populations to be restored. The regulation of proline isomerization is performed by peptidyl proline isomerases (PPIs), which can speed the interconversion of isomers by several orders of magnitude, allowing them to restore cis-trans equilibria at a biologically relevant timescale (Lu et al. 2007). This regulation gives PPIs control over the duration and amplitude of a variety of cellular events, in which proline isomerization acts as a molecular switch or timer (Andreotti 2003).

One potential site of regulatory cis-trans proline isomerization is the CTD. The regular spacing of proline residues is one of the most highly conserved features of the CTD, with each heptad containing two at the P3 and P6 positions. Interestingly, both are preceded by serine residues that are critical targets of phosphorylation. Phosphorylation of Ser-Pro motifs in non-CTD peptides has been shown to modestly stabilize the cis form and decrease the rate of isomerization (Schutkowski et al. 1998). A
study with a pS2 CTD peptide reported a \textit{cis} population of \(<30\%\) for the pS2-P3 motif, with very slow interconversion of \textit{cis} and \textit{trans} isomers, on the order of \(s^{-1}\) (Noble et al. 2005).

Currently, only one isomerase, Ess1 (Pin1 in humans), is known to target prolines in the CTD. Although Ess1/Pin1 is a member of the parvulin subfamily of PPIs, it is the only isomerase that specifically recognizes phosphorylated Ser/Thr-Pro motifs (Lu et al. 1996; Ranganathan et al. 1997). Its activity has been demonstrated on pCTD peptides in vitro, and is preferential for the pS5-P6 site over the pS2-P3 site (Gemmill et al. 2005; Verdecia et al. 2000). Structurally, Ess1/Pin1 contains an N-terminal WW binding domain and a C-terminal catalytic domain connected by a flexible linker. While the WW domain also specifically binds to phosphorylated Ser/Thr-Pro motifs and is essential in vivo (Lu et al. 2002; Verdecia et al. 2000), it likely plays a regulatory role, as it does not improve catalytic function on phosphorylated peptides in vitro (Eckerdt et al. 2005; Peng et al. 2009; Rippmann et al. 2000). Recently, NMR lineshape analysis was used to fully characterize the microscopic kinetics and thermodynamics of the Pin1 catalytic domain isomerization cycle (Greenwood et al. 2011).

The biological importance of CTD proline isomerization is unclear, but hints have been provided by studies of Ess1/Pin1 (Shaw 2007; Xu and Manley 2004). Overexpression of Pin1 inhibits transcription in vivo and both transcription and
RNAPII-mediated splicing in cell extracts, and this inhibition appears to be result of Pin1 activity early in the transcription cycle (Xu and Manley 2007). In contrast, Ess1 activity has been implicated in the initiation, elongation, and termination stages of transcription, suggesting the possibility of multiple, independent roles in transcription (Hani et al. 1999; Morris et al. 1999; Wilcox et al. 2004; Wu et al. 2003). Ess1 associates with the promoter and terminator regions of genes (Krishnamurthy et al. 2009), and a genome-wide analysis of mRNA expression in cells with impaired Ess1 activity found defective termination for a subset of transcripts, consisting mainly of small, noncoding RNAs (Singh et al. 2009).

For both Ess1 and Pin1, the transcription-related phenotypes associated with their altered activity levels appear to be the result of misregulation of the CTD phosphorylation state. Pin1 overexpression causes hyperphosphorylation of the CTD in vivo, while Pin1 knockouts accumulate hypophosphorylated CTD, indicating that Pin1 may inhibit CTD dephosphorylation (Xu et al. 2003). Consistent with this idea, Pin1 was reported to inhibit CTD phosphatase Fcp1 and stimulate kinase Cdc2/CyclinB (Xu et al. 2003). Conversely, growth defects in yeast carrying catalytically-impaired ess1 mutations are suppressed by Fcp1 overexpression and reduced levels of CTD kinases, including Kin28 and Srb10, suggesting that Ess1 stimulates CTD dephosphorylation (Wilcox et al. 2004; Wu et al. 2000). Interestingly, it was also found that alanine
substitutions at the S5 position in the CTD, but not the S2 position, suppressed the ess1 growth defects, which points to pS5-P6 as the major target of Ess1 activity, in agreement with its in vitro binding preference (Gemmill et al. 2005; Wilcox et al. 2004). Despite these intriguing findings, there is no clear explanation for the divergent effects of Pin1 and Ess1, nor is there a detailed mechanistic understanding of how catalyzed proline isomerization modulates the CTD phosphorylation state.

In addition to the CTD, Ess1/Pin1 targets many other proteins and helps regulate a diverse set of cellular processes, including cell cycle progression, immune response, neuronal function, and gene expression (Liou et al. 2011; Lu et al. 2007; Lu and Zhou 2007). Currently, there appear to be two potential mechanisms by which catalyzed proline isomerization influences downstream events. The first is by regulating the target’s phosphorylation state, as described above for the CTD of RNAPII, which then influences the activity and interactions of the target protein. The only example of this mechanism that has been characterized kinetically is the dephosphorylation of pSer-Pro motifs in Cdc25C and Tau proteins, where Pin1 stimulates the activity of phosphatase PP2A (Zhou et al. 2000). The second and most commonly observed consequence of Ess1/Pin1 activity in vivo is altered protein stability, presumably through the activation or inhibition of ubiquitin-mediated degradation. Given that phosphorylation of Ser/Thr-Pro motifs is a major factor in controlling ubiquitin-mediated proteolysis, it is very
possible that this mechanism is simply a byproduct of regulating phosphorylation states. However, based on the isomer-specific binding to phosphorylated Ser/Thr-Pro motifs of proteins that recruit targets for degradation (Orlicky et al. 2003), it has been proposed that Pin1/Ess1 participates in ubiquitin-mediated proteolysis by providing rapid access to the requisite cis/trans configuration of the target (Liou et al. 2011).

In line with its diverse physiological roles, Ess1/Pin1 has been linked to a number of diseases, including cancer, asthma, and Alzheimer’s disease (Liou et al. 2011; Lu and Zhou 2007). However, the therapeutic potential of targeting Ess1/Pin1 is limited by a poor understanding of the mechanism by which catalyzed proline isomerization regulates the phosphorylation and stability of its targets. This is due, in part, to the noncovalent nature of proline isomerization and the inability of current methods to directly visualize isomer states in vivo, and contrasts sharply with the wealth of knowledge accumulated over the last 50 years regarding protein phosphorylation and its mechanistic effects. It is likely that future structural and kinetic studies of Ess1/Pin1 targets will play a crucial role in providing insight into the regulatory roles of proline isomers and isomerization.
1.3 CTD Functions and Significance

The CTD code serves to recruit specific nuclear factors to the polymerase at the correct stage of transcription, facilitating a broad range of transcription-associated activities. These include activities directly related to the production of mature RNA transcripts – capping, splicing, 3’-end processing, termination, and export – and many others, like transcriptional activation, chromatin modification and remodeling, and DNA repair (Egloff and Murphy 2008; Meinhart et al. 2005; Phatnani and Greenleaf 2006). For some co-transcriptional processes, the specific nuclear factors at play and their coupling to CTD phosphorylation patterns have been characterized in great detail.

Of all the CTD-mediated co-transcriptional activities, 3’-end processing and termination are the most relevant for understanding the biological implications of Ssu72 phosphatase activity. The production of mature mRNAs involves a number of processing steps, including endonucleolytic cleavage and polyadenylation at the 3’-end of nascent transcripts (Shatkin and Manley 2000). This 3’-end processing has great functional importance, promoting the nuclear export, cytoplasmic stability, and efficient translation of mRNA, and its disruption has calamitous effects on cell growth and viability (Mandel et al. 2008; Zhao et al. 1999). Processing of 3’-ends is intimately coupled to transcription termination, which occurs when RNA synthesis ends and both
the polymerase and the nascent transcript dissociate from the DNA template (Kuehner et al. 2011). Termination also has great biological significance. For example, it prevents readthrough transcription that can lead to the misregulation of downstream genes through transcriptional interference (Gilmour and Fan 2008). This genomic partitioning may be most important in the compressed genome of *Saccharomyces cerevisiae*, which lacks the long intergenic regions of higher eukaryotes, but recent evidence that eukaryotic genomes are almost entirely transcribed suggests that it may play a critical role in higher eukaryotes as well (Amaral et al. 2008). Furthermore, termination promotes polymerase recycling and prevents extra long transcripts resulting from readthrough transcription that may prove harmful to the cell.

Processing of 3′-ends is catalyzed by a large and complex machinery that is physically and functionally coupled to the RNAPII CTD (Mandel et al. 2008; Millevoi and Vagner 2010; Shatkin and Manley 2000). The yeast machinery contains the cleavage and polyadenylation factor (CPF), cleavage factor IA (CFIA), and cleavage factor IB (CFIB); CPF can be further decomposed into two subcomplexes, cleavage factor II (CFII) and polyadenylation factor I (PFI). In vitro polyadenylation activity requires CPF, CFIA, CFIB, and the poly(A) polymerase Pap1, while in vitro cleavage requires only CFIA, CFIB, and CFII. Despite the relative simplicity of the cleavage and polyadenylation reactions, over 20 yeast proteins are required for proper 3′-end processing, many of
which are involved in auxiliary functions, like crosstalk with transcription and other co-
transcriptional processes and recognition of regulatory RNA sequence elements. While
the organization of the processing machinery in higher eukaryotes is slightly different,
most of the individual protein factors are well-conserved (Zhao et al. 1999). The RNAPII
CTD is not essential for either cleavage or polyadenylation, however, it strongly
stimulates both processing steps (Licatalosi et al. 2002). This functional coupling is
mediated by direct physical interactions with multiple processing factors, including CPF
subunits Yhh1/Cft1 and Cft2/Ydh1 and CFIA subunit Pcf11 (Barilla et al. 2001; Dichtl et
al. 2002b; Kyburz et al. 2003), and may serve to recruit the processing machinery to the
elongating polymerase through preferential binding to pS2 CTD, the predominant
phosphorylation pattern late in transcription, where it can then scan the emerging RNA
for regulatory sequences (Licatalosi et al. 2002; Meinhart and Cramer 2004).

Termination of transcription is intimately coupled to 3'-end processing. These
events are triggered by the transcription of regulatory sequences in the untranslated end
of the nascent RNA transcript, the poly(A) site for protein-coding genes in yeast for
example, that are recognized by components of the polymerase-associated processing
machinery (Zhao et al. 1999). The binding of these components induces polymerase
pausing, which appears to provide an important impetus for termination, and
endonucleolytic cleavage, followed by polyadenylation of the upstream cleavage
product (Gromak et al. 2006). Meanwhile, the downstream product is degraded by the 5’-3’ exoribonuclease Rat1, which is recruited to the 3’-end of genes in part through the interaction of its binding partner Rtt103 with pS2 CTD (Kim et al. 2004). In the aptly named ‘torpedo model,’ endonucleolytic cleavage would allow Rat1 to chew its way along the RNA tail of the stalled polymerase until the two collide (Connelly and Manley 1988). This collision, and possibly allosteric associations with components of the 3’-end processing machinery, would then induce the polymerase to release its DNA template (Luo et al. 2006).

While the 3’-end processing and termination mechanism described above is the canonical pathway for the majority of protein-coding mRNA transcripts, an alternative mechanism is used in yeast for some short mRNAs and most noncoding transcripts (Kuehner et al. 2011). These noncoding RNAs, including small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), and cryptic unstable transcripts (CUTs), are ~60-600 nucleotides long and play a variety of important roles in the cell (Matera et al. 2007). snRNAs and snoRNAs function in ribonucleoprotein complexes to process other classes of RNA transcripts, with snRNAs involved primarily with mRNA splicing and snoRNAs assisting with the endonucleolytic cleavage of ribosomal RNAs (rRNAs) and the chemical modification of nucleotides in transfer RNAs, rRNAs and snRNAs. In contrast to mRNA, snRNA and snoRNA are not polyadenylated, and their 3’-ends are
instead matured through endonucleolytic cleavage followed by exonucleolytic trimming. CUTs originate from intergenic regions throughout eukaryotic genomes and are rapidly degraded by the exosome (Wyers et al. 2005). Recent work suggests that CUTs are largely by-products of divergent transcription from bidirectional promoters, although some may play a role in transcriptional regulation (Neil et al. 2009).

The 3’-end processing and termination of small noncoding RNAs requires a distinct set of nuclear factors, including some members of the canonical mRNA pathway, and a complex of three unique and central players, the RNA-binding proteins Nrd1 and Nab3 and the helicase Sen1 (Lykke-Andersen and Jensen 2007). This pathway appears to be triggered in part by the cooperative binding of Nrd1 and Nab3 to specific downstream RNA sequence motifs (Carroll et al. 2007; Carroll et al. 2004). Instead of the Rat1 torpedo mechanism, termination in this pathway is dependent on the activity of Sen1, which is proposed to unwind the RNA-DNA hybrid in the active site of the polymerase (Steinmetz and Brow 1996). Termination is coupled to 3’-end processing through the association of the Nrd1 complex with the exosome and the exosome-activating complex TRAMP (Vasiljeva and Buratowski 2006). Both Nrd1 and TRAMP stimulate exosome activity, which is responsible for the trimming of snRNA and snoRNA 3’-ends and the complete degradation of CUTs (Houseley et al. 2006). In addition to its RNA-binding function, Nrd1 contains a CTD-interacting domain that
binds preferentially to p5 CTD, and like components of the mRNA pathway, this association is thought to recruit the Nrd1 complex to the elongating polymerase and thereby couple transcription to RNA processing events (Vasiljeva et al. 2008). The extent to which this pathway is conserved in higher eukaryotes is unclear, as mammalian snoRNA are encoded within introns and therefore do not require termination (Matera et al. 2007). Furthermore, the 3’-end processing of snRNA in higher eukaryotes is catalyzed by the Integrator complex, which has two subunits homologous to mRNA processing factors (Baillat et al. 2005). This homology and the fact that senataxin, the homolog of the yeast Sen1 helicase, does not appear to be required for snRNA termination (Suraweera et al. 2009) suggest that higher eukaryotes may utilize the torpedo mechanism or a unique and undiscovered pathway to terminate snRNA transcription.

It has become increasingly clear that in addition to RNA regulatory sequences, the phosphorylation state of the RNAPII CTD plays a critical role in determining which 3’-end processing and termination pathway is used to end transcription (Rondon et al. 2008). This is a particularly salient issue given that the machineries for both pathways localize at both mRNA and small noncoding RNA genes (Kim et al. 2006). One distinguishing characteristic of the two pathways is the preference of their CTD-binding components for different CTD phosphorylation patterns: while the Nrd1 complex
strongly favors pS5 CTD (Vasiljeva et al. 2008), components of the mRNA pathway, like Pcf11 and Rtt103, preferentially bind pS2 CTD (Kim et al. 2004; Licatalosi et al. 2002).

The potential biological implication of these distinct binding preferences was recently explored by Gudipati et al. with an elegant in vivo system (Gudipati et al. 2008). The authors constructed a series of yeast strains with a well-defined CUT sequence inserted into a host mRNA gene at variable distances from the promoter. While the CUT sequence lacked its own promoter, it contained an Nrd1-dependent terminator.

Intriguingly, when placed close to the host gene’s promoter, the CUT sequence efficiently directed Nrd1 termination and RNA degradation, but when positioned further toward the 3’-end of the host gene, the Nrd1-dependent terminator was bypassed and polyadenylated transcripts were produced. This study also reported that increasing transcriptional pS2 CTD levels, or decreasing pS5 levels, inhibited Nrd1-dependent termination. Thus, it appears that CTD phosphorylation plays a critical role in choosing termination pathways that, at least in certain cases, can superced the function of regulatory RNA motifs (Gudipati et al. 2008). These findings imply a model in which the CTD acts as a transcriptional position sensor, with the opposing gradients of pS5 and pS2 levels facilitating a binding competition between the Nrd1 complex and the mRNA processing machinery. In this model, the predominance of pS5 CTD early in elongation would allow the Nrd1 complex to scan potentially short or aberrant
transcripts for termination sequences, while the predominance of pS2 CTD later in transcription would promote canonical mRNA processing and discourage spurious Nrd1-dependent termination (Gudipati et al. 2008).

The RNAPII CTD is not known to play a direct role in human disease, perhaps owing to the redundancy of its heptad repeats, but many CTD-associated processes have significant medical implications. The most obvious example is DNA repair. Although the details of its connection to the CTD are poorly understood, it is interesting to note that DNA damage induces altered patterns of CTD phosphorylation and mutations of the yeast CTDK-1 kinase cause sensitivity to DNA damaging agents (Ostapenko and Solomon 2003). A second example with a better characterized connection to the CTD is RNA processing. Aberrant mRNA processing has been implicated in a number of human diseases, including cancer and Alzheimer’s disease (Kalnina et al. 2005; van Leeuwen et al. 2000), and the misregulation of alternative splicing is thought to account for ~15% of disease-causing mutations (Philips and Cooper 2000). The influence of CTD phosphorylation states on these important co-transcriptional processes, and the fact that many viruses have developed methods to regulate transcriptional activity by modulating the phosphorylation state of the CTD (Fraser and Rice 2005; Kim et al. 2002; Tamrakar et al. 2005), suggests that CTD kinases and phosphatases may provide
avenues to mitigate a number of disease pathways and may therefore prove to be promising therapeutic targets.

1.4 The pS5 CTD Phosphatase Ssu72

The CTD phosphorylation state is maintained by the concerted action of CTD kinases and phosphatases, including the pS5-specific Ssu72 phosphatase. Since its discovery in 1996, Ssu72 has been the subject of intense interest, with over 30 published biochemical and genetic studies of its activity and transcriptional role. These studies have revealed a number of intriguing results, including a limited homology to protein tyrosine phosphatases that seems at odds with its enzymatic activity toward phosphorylated serine residues, and a complicated biological function that influences all three stages of the transcription cycle.

1.4.1 Role in the Transcription Cycle

Prior studies have uncovered roles for Ssu72 in the initiation, elongation, and termination stages of transcription, and are based largely on in vivo and in vitro phenotypes resulting from mutations that were thought to, at least in part, impair its
enzymatic activity. A reexamination of the possible effects of these mutations in light of new structural information is presented in Section 2.4.5. Ssu72 was first identified through a genetic interaction with the general transcription factor TFIIB, in which it acted as an enhancer of a TFIIB mutation that conferred a downstream shift in transcription start site selection (Sun and Hampsey 1996; Wu et al. 1999). These two proteins were later shown to interact directly in vitro (Dichtl et al. 2002a; Wu et al. 1999), suggesting a potential role in transcription initiation. Ssu72 may also be involved in elongation as its mutation increases RNAPII pausing, decreases elongation efficiency, and suppresses the toxicity of 6-azauracil, a compound that slows transcription by reducing intracellular levels of GTP and UTP (Dichtl et al. 2002a; Reyes-Reyes and Hampsey 2007).

The role of Ssu72 in 3’-end processing and termination may be the most enigmatic aspect of its biological function. Ssu72 is an integral component of the cleavage and polyadenylation factor (CPF) complex (Nedea et al. 2003), where it is anchored through its association with the N-terminal domain of Pta1 (Ghazy et al. 2009; He et al. 2003), or symplekin in higher eukaryotes, and appears to influence both 3’-end processing and termination. In yeast, Ssu72 is required for both cleavage and polyadenylation (He et al. 2003; Krishnamurthy et al. 2004), although its phosphatase activity is not necessary for this function (Krishnamurthy et al. 2004). Instead, it appears
that the interaction of Ssu72 with Pta1 masks an otherwise inhibitory effect of the Pta1 N-terminal domain on mRNA 3'-end processing (Krishnamurthy et al. 2009). This function may be somewhat divergent in higher eukaryotes, as the N-terminal domain of symplekin inhibits only transcription-coupled polyadenylation, and both the activity and binding of Ssu72 are required to mitigate the inhibition (Xiang et al. 2010). Ssu72 is also involved in termination. Genome-wide expression analysis of yeast strains carrying Ssu72 mutations revealed readthrough transcription for snoRNAs and some mRNAs (Ganem et al. 2003; Nedea et al. 2003), suggesting that Ssu72 is involved mainly in Nrd1-dependent termination. On the other hand, several studies have noted defects in canonical mRNA termination (Dichtl et al. 2002a; Nedea et al. 2003; Steinmetz and Brow 2003), which may indicate roles for Ssu72 in both termination pathways that can be unevenly perturbed in an allele-specific manner.

Finally, Ssu72 is involved in gene looping, a proposed mechanism for transcription reinitiation that tethers the promoter and terminator regions of a gene in order to facilitate polymerase recycling (Ansari and Hampsey 2005; O'Sullivan et al. 2004). This mechanism is supported by the finding that initiation factor TFIIB and several 3'-end processing and termination factors, including Ssu72, occupy both the promoter and terminator regions (Ansari and Hampsey 2005; Calvo and Manley 2003; Singh and Hampsey 2007). Ssu72 is required for gene loop formation in a manner
dependent upon its catalytic activity, and its binding partner Pta1 is also required (Ansari and Hampsey 2005). The essential nature of its phosphatase activity suggests that Ssu72 may promote polymerase recycling simply by helping return the RNAPII CTD to a hypophosphorylated, initiation-competent form. However, it may also have a structural role in bridging factors involved in initiation and termination through its interactions with Pta1 and TFIIB. In line with this second role, TFIIB occupancy at the terminator, but not the promoter, is disrupted by depletion of Ssu72 in yeast (Singh and Hampsey 2007).

1.4.2 Similarities to LMW PTPs

The phosphatase activity of Ssu72 was originally surmised by the identification of the signature CX\textsubscript{5}R catalytic motif of protein tyrosine phosphatases (PTPs) (Ganem et al. 2003; Meinhart et al. 2003). PTPs constitute a superfamily of enzymes encoded by ~80 genes in humans and can be classified into four main families on the basis of their sequence, structure and function (Alonso et al. 2004). Despite a conserved active site structure and mechanism and similar folds, the PTP families share almost no sequence homology and have distinct topologies, suggesting that they are products of convergent evolution (Jackson and Denu 2001). The N-terminal positioning of the catalytic motif of
Ssu72 and its relatively small size (206 residues in yeast) are two defining characteristics of the low molecular weight (LMW) family of PTPs, and this homology is further supported by a similar arrangement of predicted secondary structure elements (Ganem et al. 2003; Meinhart et al. 2003). The LMW PTPs are 18 kDa enzymes that dephosphorylate many growth factor receptors in order to regulate a variety of processes including cell growth rate, adhesion, and cytoskeleton organization (Souza et al. 2009). Interestingly, although they are highly conserved from bacteria to humans, LMW PTPs are represented in the human genome by a single gene (Alonso et al. 2004).

PTPs utilize their signature CX5R motif in a highly conserved two-step mechanism of dephosphorylation (Figure 4). The overall fold of the PTP catalytic domain contains a central parallel β-sheet of at least 4 strands that is sandwiched

![Figure 4. Catalytic Mechanism of PTPs.](image)

Schematic representation of dephosphorylation with backbone groups shown in black, sidechains in red, and the substrate in blue. In addition to the active site loop, the reaction requires an aspartate located far from the catalytic motif in the primary sequence that acts as a general acid.
between α-helices (see Figure 9 in Section 2.2.1), and the CX3R catalytic motif is found in a substrate-binding loop (the P-loop in PTP parlance) at the C-terminal end of one of the central β-strands (Jackson and Denu 2001). In the first step of catalysis, the cysteine sidechain, which is held as a negatively charged thiolate, initiates dephosphorylation through nucleophilic attack of the substrate phosphorous atom to form a phosphoenzyme intermediate. The remainder of the active site loop is responsible for binding the substrate phosphate moiety and for stabilizing the intermediate, with the conserved arginine sidechain and the amide groups of the preceding five residues providing strong electrostatic interactions. In the second step, the intermediate is hydrolyzed to produce inorganic phosphate and the regenerated enzyme. In addition to the catalytic motif, the PTP mechanism requires a conserved aspartate residue which acts as a general acid in the reaction. This aspartate is distant in primary sequence to the catalytic motif, and is positioned near the active site on a flexible ‘aspartate loop.’ Initial kinetic investigations of Ssu72 demonstrated that, like PTPs, Ssu72 is active toward para-nitrophenylphosphate (pNPP), a commonly used phosphotyrosine substrate analog, although its activity is fairly weak (Meinhart et al. 2003). Mutation and inhibition studies also provided support for a conserved PTP mechanism, with mutation of the catalytic cysteine abolishing activity in vitro and conferring lethality in yeast, which underscores the biological significance of its phosphatase function (Ganem et al. 2003;
Meinhart et al. 2003). On the basis of these kinetic commonalities and predicted structural similarities to the LMW PTP family, it was originally proposed that the role of Ssu72 in transcription was mediated through dephosphorylation of phosphotyrosine residues (Ganem et al. 2003; Meinhart et al. 2003).

While Ssu72 bears some resemblance to LMW PTPs, it also has several unique features and characteristics. First, although other families of PTPs have less stringent substrate specificity, the activity of LMW PTPs is directed exclusively at phosphotyrosine residues (Alonso et al. 2004). In contrast, Krishnamurthy et al. showed that the in vivo target of Ssu72 activity is phosphoserine residues in the RNAPII CTD (Krishnamurthy et al. 2004). Interestingly, this activity is highly specific for the pS5 position over pS2 both in vivo and in vitro (Hausmann et al. 2005; Krishnamurthy et al. 2004), unlike the activities of CTD phosphatases Scp1 and Fcp1, which have only preferential activity for pS5 and pS2, respectively (Hausmann and Shuman 2002; Yeo et al. 2003). Second, Ssu72 is ~60 residues larger than the LMW PTP enzymes and contains several additional predicted secondary structure elements, including an inserted sequence of ~35 residues predicted to form two short β-strands followed by a long α-helix and a C-terminal extension of ~25 residues predicted to form a β-strand followed by an α-helix. Additionally, the ‘aspartate loop’ that contains the conserved catalytic aspartate residue is present in the Ssu72 primary sequence, but it is predicted to be
about half as long as its equivalent in the LMW PTP structure. Finally, aside from catalytic residues, there is no identifiable sequence conservation between Ssu72 and members of the LMW family of PTPs. Given these intriguing differences, the place of Ssu72 in the PTP superfamily remains an unanswered question, and it is possible that Ssu72 represents the founding member of a new family of PTPs.

1.4.3 In Vivo Connection with Proline Isomerase Ess1

Recently, two studies illuminated an unexpected in vivo connection between Ssu72 and Ess1, a proline isomerase that targets pSer-Pro motifs in the RNAPII CTD (see Section 1.2.3). Impaired Ess1 catalytic activity in yeast cells was found to cause a temperature sensitive phenotype and the accumulation of pS5 CTD (Krishnamurthy et al. 2009; Singh et al. 2009). Importantly, both defects were ameliorated by overexpression of Ssu72 (Krishnamurthy et al. 2009). Furthermore, a genome-wide analysis of mRNA expression in these cells showed readthrough transcription for a set of genes – mainly snoRNAs – with marked similarity to those previously identified in cells with impaired Ssu72 activity (Singh et al. 2009). In line with this effect on termination, Ess1 catalytic activity was shown to influence the recruitment of termination factors Nrd1 and Pcf11 to the 3’-end of snoRNA genes (Singh et al. 2009).
Impaired Ess1 activity caused Nrd1 occupancy at the terminator to increase, while Ess1 overexpression led to a decrease in occupancy. The opposite scenario was observed for Pcf11 recruitment, with impaired Ess1 activity causing decreased occupancy at the 3'-end.

The authors of both studies proposed a common model that explains these results in the following way (Krishnamurthy et al. 2009; Singh et al. 2009). At initiation, phosphorylation at the S5 position would cause the CTD to adopt predominantly the cis form of the pS5-P6 motif. Near the end of small noncoding genes, dephosphorylation of pS5 would be required to coordinate the exchange of Nrd1, which preferentially binds pS5 CTD (Vasiljeva et al. 2008), and Pcf11, which prefers pS2 CTD (Licatalosi et al. 2002), in order to facilitate termination via the Nrd1-dependent pathway. However, Ssu72 would have isomeric specificity for the less populated trans form of the pS5-P6 motif, and would require Ess1 activity to efficiently process the entire pool of available substrate. Consequently, reduced catalysis by Ess1 would lead to an increased level of pS5 CTD, causing an improper localization of pS5 CTD-bound Nrd1 and a concomitant blocking of pS2 CTD-mediated Pcf11 recruitment and resulting in readthrough transcription. This ordered competition between Nrd1 and Pcf11 for binding to the CTD is supported by the observation that overexpression of Pcf11 reduces the abnormally high occupancy of Nrd1 at snoRNA terminators in yeast strains with impaired Ess1
activity (Singh et al. 2009), and is consistent with the involvement of Pcf11 in the Nrd1-dependent termination pathway in addition to its function in the canonical pathway (Kim et al. 2006).

These studies add an interesting nuance to the prevailing model for Nrd1-dependent termination events (see Section 1.3). According to the mechanism proposed by Gudipati et al., Nrd1-dependent termination is facilitated by the predominance of pS5 CTD in early elongation, which allows Nrd1 to associate with the polymerase while scanning the nascent transcript for regulatory sequences (Gudipati et al. 2008). Later in elongation, pS2 CTD predominance would promote the association of mRNA processing factors and termination by the canonical pathway. Under this model, the elevated pS5 CTD levels caused by impaired Ess1 activity would be expected to encourage, rather than disrupt, Nrd1-dependent termination. It therefore seems likely that while pS5 CTD is required for the initial Nrd1-mediated recognition of termination sequences in small noncoding transcripts, a sudden drop in pS5 CTD levels catalyzed by Ssu72 is also necessary to coordinate Nrd1 release and Pcf11 recruitment by pS2 CTD, which may in turn promote transcript cleavage or some other critical step in Nrd1-dependent termination.

The proposed model also makes a bold prediction regarding CTD proline isomer states (Krishnamurthy et al. 2009; Singh et al. 2009). In order for Ess1 to effectively
regulate Ssu72 activity, the model supposes that pS5 CTD favors the cis configuration of the pS5-P6 motif, a point contradicted by studies of isomer populations for pSer-Pro motifs in synthetic peptides (Noble et al. 2005; Schutkowski et al. 1998). It is possible that such a conformation could be stabilized by a higher order CTD structure or the binding of specific factors, but there is no evidence to date in support of this idea. However, a predominant cis pS5-P6 configuration would allow Ssu72 to specifically dephosphorylate the trans isomer, in agreement with the fact that all reported enzymes that target pSer-Pro motifs with isomeric specificity recognize the trans configuration (see Section 2.4.2).
2. Structural and Kinetic Studies of Phosphatase Ssu72

The CTD phosphatase Ssu72 is a particularly intriguing enzyme. In order to define its relation to known enzyme families and to provide biochemical explanations for previous functional observations, we performed an extensive biophysical characterization using a variety of techniques. We expect that this work will promote future in vivo investigations of its role in coordinating transcription with co-transcriptional events, through the structure-guided design of mutants that perturb its function in a precise and well-defined manner. More generally, this work will help illuminate how the CTD code is controlled during the transcription cycle, and how proline isomers and isomerization are used to manipulate phosphorylation states during transcription and other cellular events.

Our early research efforts focused on the Ssu72 ortholog from Saccharomyces cerevisiae (ySsu72). Recombinant overexpression in Escherichia coli and standard purification techniques yielded large amounts of pure enzyme with activity toward the phosphotyrosine analog, para-nitrophenylphosphate (pNPP). Despite these promising results, the preparation of concentrated ySsu72 samples for structural studies caused severe precipitation, and extensive screening of buffers and solubility-enhancing additives failed to identify conditions permitting concentrations over ~100 µM. Attempts to find a smaller and more stable construct with limited proteolysis were also
unsuccessful. While the reasons for the observed precipitation remain unclear, the
$^1\text{H}/^{15}\text{N}$-HSQC-TROSY spectrum of ySsu72 showed very poor dispersion of amide
resonances, which may indicate aggregation or unfolding. To circumvent these
difficulties, we screened orthologs and identified the *Drosophila melanogaster* Ssu72
dSsu72) as a highly stable and soluble enzyme with significant sequence similarity to
ySsu72 (Figure 9). The excellent quality of the $^1\text{H}/^{15}\text{N}$-HSQC-TROSY spectrum of dSsu72
(Figure 5) showed that the enzyme is well-folded and signaled a promising target for
further structural characterization.

2.1 NMR Assignments and Catalytic Mutant Screening

One of the initial challenges to any structural study of an enzyme and its
interaction with substrate is the need to generate a stable and static enzyme-substrate
complex. This is usually accomplished in one of two ways. First, the wild-type enzyme
can be studied in complex with a non-hydrolyzable substrate analog, provided that the
analog can be inexpensively purchased or synthesized in large amounts. Second, an
inactive form of the enzyme can be designed using targeted mutations and then studied
in complex with the natural substrate. For Ssu72, the second strategy was more feasible,
and NMR proved to be a critical tool for identifying mutations that disrupted the
catalytic activity of Ssu72 without perturbing its interaction with substrate.
2.1.1 Backbone and Sidechain Assignment

Backbone amide resonances of phosphate-bound *Drosophila melanogaster* Ssu72 (dSsu72) were assigned with a $^2$H/$^{13}$C/$^{15}$N-labeled protein sample using a suite of TROSY-based, 3-D triple-resonance experiments – HNCO, the ‘just-in-time’ HN(CA)CO (see Chapter 4), HNCA, HN(CO)CA, HN(CA)CB, and HN(COCA)CB – and the PACES algorithm to identify resonance connectivity (Coggins and Zhou 2003). Amide assignments were confirmed through analysis of a 4-D NH-NH diagonal-suppressed TROSY-NOESY-TROSY experiment (see Chapter 5), and through resonance connectivity in TROSY-based HN(CA)HA and HN(COCA)HA experiments collected with a $^{13}$C/$^{15}$N-labeled sample. Overall, every identifiable backbone amide peak in the $^1$H/$^{15}$N-HSQC-TROSY spectrum was assigned (Figure 5), representing 93% (177 of 190) of the non-proline residues in Ssu72. Residues without amide signals are found in flexible regions, including the N-terminus (residues 1-3) and the active site loop (residues 14-15 and 18-19). The perdeuterated NMR sample also contained protonated valine $\gamma$ methyls and leucine and isoleucine $\delta$ methyls, which were assigned with a 4-D HC(CCO)NH experiment.

Sidechain resonances were assigned using a 3-D HCCH TOCSY experiment, which worked well for shorter sidechains with efficient magnetization transfer. Longer
Figure 5. Assigned $^1$H/$^{15}$N-HSQC-TROSY Spectrum of Phosphate-Bound Drosophila Melanogaster Ssu72. Collected at 800 MHz and 30°C.

Sidechain assignments were confirmed and extended using intraresidue NOEs in a $^1$H/$^{13}$C-NOESY dataset. Aromatic ring $^1$H and $^{13}$C resonances were assigned using a structure-guided analysis of NOEs in both aromatic and aliphatic 3-D $^1$H/$^{13}$C-NOESY datasets, however, several of these ring resonances were broadened or missing. Methyl groups of valine and leucine residues were stereospecifically assigned using fractional $^{13}$C-labeling (Neri et al. 1989). As a result of backbone and sidechain resonance
assignment, ~82% of the protons in Ssu72 were assigned, including 95% of the backbone protons and 77% of sidechain protons, along with 80% of the $^{13}$C and $^{15}$N atoms.

### 2.1.2 Finding a Substrate-Trapping Mutant with NMR Titrations

Next, we used NMR titrations to explore the phosphate-binding properties of wild-type Ssu72 and mutant enzymes with alanine, serine, and aspartate replacements of the catalytic cysteine residue (C13). The catalytic cysteine is part of a conserved CX$_5$R motif also found in protein tyrosine phosphatases. The C13 sidechain is held as a negatively charged thiolate group and initiates dephosphorylation through nucleophilic attack of the substrate phosphorous atom, while the five following residues form the phosphate-binding active site loop. Together with the sidechain of the conserved arginine (R19), the backbone amides of the active site loop hold the substrate phosphate group in place with strong electrostatic interactions. Catalysis by Ssu72 also requires an aspartate residue (D144) on a flexible loop positioned near the active site (‘the aspartate loop’), which protonates the leaving phosphate to regenerate the enzyme. Surprisingly, we observed a large number of significant perturbations upon titration of wild-type Ssu72 with inorganic phosphate (Figures 6 and 7A). Perturbed residues cluster into three regions on the primary sequence (Figure 7A): near the active site loop (region 1), residues centered on D51 (region 2), and the predicted aspartate loop (region 3). Given
Figure 6. Phosphate Titrations of Active Site Mutants of Ssu72. The phosphate titration of WT dSsu72 is shown alongside identical titrations of C13D, C13S and C13A mutants. Brown to yellow coloring denotes $^1$H/$^{15}$N-HSQC-TROSY spectra for apo enzyme to a 25:1 molar ratio of phosphate to enzyme, respectively. Close-up views of peak perturbations for four residues in the WT and C13D titrations are shown below the full spectra.
the small size of the phosphate group and the large number of residues affected, it is likely that these perturbations are caused by conformational changes associated with phosphate binding, rather than direct interaction with phosphate.

We then tested C13A, C13S and C13D dSsu72 mutants for phosphate binding. While C13D dSsu72 exhibited phosphate-induced chemical shift perturbations similar to the WT enzyme (Figures 6 and 7B), neither C13A nor C13S showed any perturbations (Figure 6), suggesting that a negatively charged sidechain at this position is required for proper substrate binding. This idea is supported by the fact that no perturbations were observed in a titration of C13S dSsu72 with a synthetic pS5 CTD peptide. Notably, a catalytic cysteine to serine mutation in protein tyrosine phosphatase 1B (C215S PTP1B) has been reported to dramatically distort the apo conformation of the active site loop, causing it to flip open into the substrate-binding space (Scapin et al. 2001). It is possible that the loss of a negatively charged catalytic sidechain in the C13A and C13S mutants causes a similar distortion of the conformation and dynamics of the active site loop of Ssu72, rendering it unable to effectively bind substrate.

To finalize the design of a substrate-trapping mutant, we performed titrations with a synthetic substrate peptide. This 7-residue pS5 CTD peptide contains the minimal binding epitope required for optimal Ssu72 activity (Hausmann et al. 2005) with an additional residue at each end (Ac-PTpSPSYS-NH$_2$). In contrast to the C13S mutant, titration of C13D dSsu72 with the substrate peptide resulted in the loss of a
large number of peaks, which indicates an interaction in the intermediate exchange regime on the NMR timescale. The set of residues affected by the substrate peptide is very similar to the set of residues perturbed by phosphate titration, suggesting that phosphate binding alone causes an analogous conformational change (compare Figure 7B and C). Unfortunately, C13D dSsu72 was found to have residual activity, making it unsuitable for further structural studies; to eliminate this activity, we made an additional mutation of the catalytic aspartate residue (D144). Titration of C13D/D144N dSsu72 with substrate peptide revealed a set of perturbed residues that was nearly identical to the set of residues perturbed in the single mutant enzyme (compare Figure
7C and D), verifying that this double mutant is fully capable of proper substrate recognition.

2.2 *Structure of Ssu72 in Complex with pS5 CTD Substrate*

To obtain a high resolution model of the Ssu72-pCTD interaction, we used X-ray crystallography to determine the structure of C13D/D144N dSsu72 in complex with the same substrate peptide used in the NMR titrations described above, a synthetic, 7-residue pS5-CTD peptide with the sequence Ac-PTpSPSYS-NH₂ that contains the minimal binding epitope (Hausmann et al. 2005) plus an additional residue on each end. Extensive screening and optimization of crystallization conditions for the complex sample yielded 50 x 50 x 200 μm rhombic-shaped crystals that diffracted to 2.1 Å at the

![Figure 8. Crystals and Diffraction Data for the Ssu72-Substrate Complex. Crystals were grown by hanging drop vapor diffusion (left panel) and diffracted to 2.1 Å resolution at the SER-CAT 22-ID beamline at Argonne National Laboratory (right panel).](image)
Table 1. Crystallographic Data Collection and Refinement Statistics for the Ssu72-Substrate Complex.

<table>
<thead>
<tr>
<th><strong>Data Collection</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution range (Å)</td>
<td>50-2.10 (2.18-2.10)</td>
</tr>
<tr>
<td>Space group</td>
<td>R3</td>
</tr>
<tr>
<td>Unit cell dimensions (Å)</td>
<td>157.5, 157.5, 118.8</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.9 (99.8)</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>63898 (6361)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>3.9</td>
</tr>
<tr>
<td>R&lt;sub&gt;merge&lt;/sub&gt; (%)</td>
<td>6.1 (32.6)</td>
</tr>
<tr>
<td>I/I&lt;sub&gt;o&lt;/sub&gt;</td>
<td>23.7 (4.1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Refinement</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R&lt;sub&gt;work&lt;/sub&gt;/R&lt;sub&gt;free&lt;/sub&gt; (%)</td>
<td>20.90 / 24.11</td>
</tr>
<tr>
<td>Number of atoms&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>6360</td>
</tr>
<tr>
<td>Peptide</td>
<td>236</td>
</tr>
<tr>
<td>Water</td>
<td>658</td>
</tr>
<tr>
<td>Other molecules&lt;sup&gt;c&lt;/sup&gt;</td>
<td>31</td>
</tr>
<tr>
<td>B-factors (Å&lt;sup&gt;2&lt;/sup&gt;)&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>33.38</td>
</tr>
<tr>
<td>Peptide</td>
<td>36.53</td>
</tr>
<tr>
<td>Water</td>
<td>34.55</td>
</tr>
<tr>
<td>Other molecules&lt;sup&gt;e&lt;/sup&gt;</td>
<td>32.83</td>
</tr>
<tr>
<td>Rmsd from ideal geometry</td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.003</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>0.748</td>
</tr>
<tr>
<td>Ramachandran plot</td>
<td></td>
</tr>
<tr>
<td>Favored (%)</td>
<td>97.2</td>
</tr>
<tr>
<td>Allowed (%)</td>
<td>99.9</td>
</tr>
</tbody>
</table>

<sup>a</sup>Values in parenthesis are for the highest-resolution shell.

<sup>b</sup>R<sub>free</sub> was calculated with the 5% of the data randomly omitted from refinement.

<sup>c</sup>Nonhydrogen atoms. Riding hydrogens were used in refinement and are included in the deposited structure.

<sup>d</sup>Imidazole and a PEG fragment

<sup>e</sup>Only nonhydrogen atoms included.
SER-CAT facility at Argonne National Laboratory (Figure 8). Phasing was carried out by molecular replacement using the apo wild-type dSsu72 structure deposited by the Northeast Structural Genomics Consortium (pdb code 3FDF). Data collection and refinement statistics for the final model are shown in Table 1, and the coordinates were deposited in the Protein Data Bank under accession number 3P9Y.

2.2.1 Comparison with the LMW PTP Fold

The Ssu72 structure contains the LMW PTP fold and two sizable additions (Figure 9A-C). The conserved scaffold consists of a central, 4-stranded, parallel β-sheet flanked by two α-helices on one side and a third α-helix on the other, and is represented in Figure 9 by the structure of the LMW PTP Lpt1 in complex with its substrate pNPP (Wang et al. 2000). The first addition to the Ssu72 fold is a ~60 residue insertion after the second β-strand that forms a subdomain of two β-strands packed against three α-helices. This insert structure sits immediately adjacent to the active site loop and contributes to substrate binding (see Section 2.2.2). The corresponding region in the LMW PTP fold contains ~25 residues and forms a single α-helix that packs against the back of the central β-sheet and is tethered by long loops that interact with substrate. The second addition is a ~30 residue C-terminal extension that forms a helix-turn-strand motif with the β-strand bonding in an anti-parallel fashion to the central β-sheet. Another
Figure 9. Structure of pS5 CTD Phosphatase Ssu72 and Comparison with LMW PTP Lpt1. (A) Sequence alignment of Ssu72 orthologues and the LMW PTP Lpt1. Ssu72 proteins from *Drosophila melanogaster* (Dm Ssu72), human (Hs Ssu72) and *Saccharomyces cerevisiae* (Sc Ssu72) are shown with conserved and similar residues highlighted in orange and yellow, respectively. α-helices and β-strands from the structures of dSsu72 and Lpt1 (pdb code 1D1Q) are denoted with cylinders and arrows, respectively. The
secondary structure elements of Ssu72 that comprise the two additions to the LMW PTP scaffold are colored dark blue to match the structure in (B). Catalytic residues are highlighted in red. (B) Ribbon diagram of the dSsu72-pCTD complex. Additions to the LMW PTP fold are colored dark blue. The substrate peptide and catalytic residues C13 (mutated to D13), R19 and D144 (mutated to N144) are shown as sticks. (C) Ribbon diagram of the LMW PTP Lpt1. The Lpt1 structure shows a catalytic mutant bound to the substrate pNPP (pdb code 1D1Q). The substrate and catalytic residues C14 (mutated to A14), R20 and D133 are shown as sticks. (D-E) Active site close-ups of substrate-bound dSsu72 (D) and Lpt1 (E). The backbone of the phosphate-binding loops is shown as sticks along with important catalytic residues. Hydrogen bonds are represented by gray dashed lines.

distinguishing characteristic of Ssu72 is the length of its aspartate loop. The α-helix following the aspartate loop is one turn shorter in Ssu72, which allows a commensurate shortening of the aspartate loop and increases active site accessibility from the side opposite the insert subdomain.

The active site architecture of Ssu72 is very similar to LMW PTPs, providing structural support for a common mechanism of dephosphorylation. Figure 9D-E shows a close-up view of phosphate binding for the Ssu72-pCTD complex and the LMW PTP Lpt1-pNPP complex (Wang et al. 2000). The conformations of active site loops in these structures are nearly identical, with the backbone amide groups forming a series of hydrogen bonds to the substrate phosphate group and the guanidinium group of the catalytic arginine extending this hydrogen bonding to form a complete circle. Despite the different length of aspartate loops, the positions of the catalytic aspartate sidechains (D144N in the Ssu72 structure) are very similar. While the carboxylate of the Lpt1 aspartate is turned away from the active site, presumably due to electrostatic repulsion
with the substrate phosphate group, in the Ssu72 structure the catalytic D144N sidechain is pointed toward the phosphate. Another difference between the active sites is the position of the substrate phosphate. In the Lpt1 structure, the catalytic cysteine is mutated to alanine and this loss of negative charge allows the phosphate to bind deeper in the active site, close to the position predicted for the phosphoenzyme intermediate (Wang et al. 2000), while the phosphate group in the Ssu72 structure is held ~0.7 Å higher by the negative charge of the C13D sidechain.

2.2.2 Binding of a Cis Proline Substrate Conformation

The electron density for the pCTD peptide is excellent and extends almost completely through both termini (Figure 10A). While the substrate phosphate group is anchored in the active site with strong hydrogen bonds, the insert subdomain creates a deep and narrow recognition cleft that forces the C-terminal end of the CTD to extend away immediately from the phosphate in a nearly opposite direction. This constrained conformation is made possible by a cis isomer of the pS5-P6 peptide bond (ω = -2.42°) that provides an abrupt turn in the substrate backbone. The uniqueness of the cis pS5-P6 isomer and its pronounced effect on the substrate conformation are readily apparent by comparison with the structure of pS5 CTD bound to the CTD phosphatase Scp1 (pdb code 2GHT) in Figure 10C-D (Zhang et al. 2006). In the Scp1 complex, the trans isomer
Figure 10. Binding of Cis Proline pCTD Substrate by Ssu72. (A) Orthogonal views of pCTD conformation and binding. The enzyme structure is shown in tan with yellow highlighting for the active site loop, and residues that form the substrate recognition surface are labeled and shown as green sticks. Grey mesh represents Fo-Fc omit map density (contoured at 3.0σ) surrounding the substrate. (B) Schematic of the Ssu72-pCTD interaction. Hydrogen bonds and van der Waal contacts are shown as red and blue dashed lines, respectively. The green line represents an intramolecular hydrogen bond in the pCTD substrate. (C-D) The conformations of pS5 CTD peptides in complex with dSsu72 (C) and phosphatase Scp1 (D) are shown with the pS5-P6 peptide bond highlighted in magenta. For clarity, only residues P3 to S7 are shown. The cis isomer is stabilized by an intramolecular hydrogen bond (gray dashed line) and causes a severe kink in the CTD backbone.
of the pS5-P6 bond provides an extended backbone configuration, in contrast to the sharp kink produced by the cis pS5-P6 isomer.

To achieve its unusual isomeric specificity, Ssu72 must stabilize the high energy cis proline substrate conformation (Figure 10B). This is accomplished in part by facilitating the formation of an intramolecular hydrogen bond between the substrate T4 sidechain hydroxyl and the P6 backbone carbonyl. The T4 sidechain is positioned by its methyl group binding in a hydrophobic pocket formed by the sidechains of K44, P46 and P53 and the backbone of L45. Hydrogen bonding with the T4 hydroxyl forces the P6 residue to adopt the cis isomer, which is further stabilized by strong van der Waal interactions between the P6 ring and a hydrophobic patch of Ssu72 formed by the sidechains of P46, L82 and M85 and the backbone of N145. Between the T4 and P6 residues, the pS5 backbone amide forms a hydrogen bond with the K44 backbone carbonyl of Ssu72, while the pS5 sidechain sits in the active site as described above. At the N-terminal end of the substrate peptide, the P3 residue makes minor van der Waal contacts with the K44 sidechain of Ssu72. Toward the C-terminal end, additional isomer-specific binding energy is provided by an intermolecular hydrogen bond between the substrate S7 mainchain amide group and the D144N backbone carbonyl. While the sidechains of the S7 and S2 residues are largely solvent exposed, the phenol ring of the intervening Y1 sidechain forms an aromatic-amide stacking interaction with the A49-F50 peptide bond of Ssu72 (Toth et al. 2001), causing a reorientation of the loop
between the two β-strands of the insert subdomain in the complex compared to the apo structure (see Section 2.4.4).

2.3 *Ess1*-Catalyzed Proline Isomerization Stimulates Ssu72 Activity

Given that Ssu72 recognizes a *cis* pS5-P6 isomer configuration, the extremely slow interconversion of *cis-trans* proline (s⁻¹) may present a rate-limiting step for dephosphorylation and therefore a potential regulatory mechanism. To test this possibility, we first quantified the pS5-P6 isomer populations in the substrate peptide using natural abundance ¹H/¹³C-HSQC spectra, and then examined the effect of catalyzed proline isomerization on Ssu72 kinetics. Reactions were monitored by measuring phosphate release with a Malachite Green-based assay.

2.3.1 Measurement of Substrate Proline Isomer Populations

A previous study of peptides with pSer-Pro motifs showed that the *cis* isomer is less favored than the *trans* form, with populations of 12-20% depending on the adjacent residues (Schutkowski et al. 1998). Similar work with a pS2 CTD peptide reported a *cis* population under 30% for the pS2-P3 isomer (Noble et al. 2005). While these results suggest that phosphorylation of the S5-P6 CTD motif is unlikely to change the minority
status of the cis isomer, no direct measurements have been reported. To address this point, we measured the fractions of cis and trans pS5-P6 in our substrate, a synthetic peptide with the sequence Ac-PTpSPSYS-NH₂, which is the same peptide used for crystallization. A natural abundance $^1$H/$^{13}$C-HSQC spectrum shows two sets of signals

Figure 11. Measurement of Proline Isomer Populations for Ssu72 Substrate. (A-B) Natural abundance $^1$H/$^{13}$C-HSQC spectra were collected for the substrate peptide (A) and a shorter pCTD peptide with only one proline residue (B). Close-up views depict the proline β and γ region in each spectrum. Colored boxes indicate the range of chemical shifts from plus one to minus one standard deviation about the mean for each proline atom in either the cis (blue) or trans (green) form (Schubert et al. 2002).
for each proline Cβ and Cγ atom, corresponding to the cis and trans forms (Figure 11A). However, the substrate peptide contains two prolines, making it difficult to unambiguously assign peaks to the P3 and P6 residues. By collecting the same experiment with a shorter peptide containing only the P6 residue (Figure 11B), we were able to assign the P6 resonances and readily map them to the longer peptide. Populations were calculated using the average ratio of cis and trans peak volumes for the P6 Cβ and Cγ atoms. This gave a fractional population of 12.4% for the cis isomer, verifying that the pS5-P6 motif is predominantly in the trans configuration.

2.3.2 Ssu72 Phosphatase Assays with Ess1

To examine the effect of catalyzed proline isomerization on Ssu72 activity, we performed a series of kinetic assays with our synthetic substrate peptide. In a reaction with 60 μM pCTD and 4 μM WT dSsu72, less than half of the input substrate was dephosphorylated in 5 minutes. To speed the interconversion of proline isomers, we added increasing amounts of Ess1, a CTD proline isomerase that acts on pS5-P6 motifs. Ess1 significantly stimulated the activity of Ssu72, with a 2 μM Ess1 concentration causing a ~2-fold increase in reaction completion over 5 minutes (Figure 12A). The increase in Ssu72 activity was saturable (Figure 12B) and a concentration of 10 μM Ess1 was sufficient to maximize the completion percentage at all time points. To ensure that
Figure 12. Ess1 Stimulates pCTD Dephosphorylation by Ssu72. (A) The activity of Ssu72 was monitored with various concentrations of the proline isomerase Ess1. The maximum stimulation was ~2-fold under these conditions. In (A-D), each point represents the average from three independent reactions and error bars denote standard deviations. (B) Ess1 enhancement of Ssu72 activity is saturable. The last points of the reactions in (A) are plotted versus Ess1 concentration, with black circles representing reactions omitted from (A) for clarity. (C) pCTD dephosphorylation by Ssu72 reaches completion without Ess1. The ‘no Ess1’ reaction in (A) was monitored over 25 minutes,
and reaches ~90% completion in 20 minutes, 4-fold slower than the reaction with 2 μM Ess1. (D) Four catalytically-impaired Ess1 mutants fail to significantly stimulate Ssu72. Reactions contained 1 μM of each Ess1 protein. Residual enhancement corresponds to a 50-100-fold reduction of WT Ess1 activity (compare to (A)). (E-F) Ess1 stimulates Ssu72 dephosphorylation of full-length pCTD. The 26-repeat yCTD was hyperphosphorylated in vitro and used as a substrate in reactions monitored by Western blotting with pS5-specific antibodies. As a control, pS2 levels were also measured; the apparent increase in the pS2 signal is likely due to a higher affinity of the pS2 antibody for singly-phosphorylated pS2 heptads over doubly-phosphorylated pS2/pS5 heptads, as reported for other pS2-specific antibodies (Chapman et al. 2007). pS5 CTD levels were quantified with infrared imaging and the resulting substrate depletion curves are shown in (F). Each point represents the average of three independent reactions and error bars denote standard deviations.

Ssu72 was capable of completely dephosphorylating the pCTD substrate in the absence of Ess1, we followed the reaction without Ess1 over the course of 25 minutes (Figure 12C), which showed that ~90% of substrate is dephosphorylated in 20 minutes, 4-fold slower than the reaction with 2 μM Ess1. This behavior implies a kinetic effect on dephosphorylation, with Ess1-catalyzed isomerization improving the availability of cis proline substrate for Ssu72 until the rate of isomer interconversion exceeds the rate of pCTD dephosphorylation.

To confirm that catalyzed isomerization is the cause of Ssu72 stimulation, we performed identical experiments with four catalytically-impaired Ess1 mutants. Three of these mutations are in the catalytic domain of Ess1 (C120S, S122P and H164R), while the fourth disrupts phosphate binding in the WW domain (K68A). All four mutations significantly decreased the enhancement of Ssu72 activity (Figure 12D). To ensure that this result is not caused by reduced thermostability, we measured melting temperatures
for each Ess1 mutant by circular dichroism. Only the H164R mutation caused a large
decrease in stability (melting temperature of 32°C) that may contribute to its impaired
stimulation. Together, these studies demonstrate that Ess1 enhances Ssu72 activity by
catalyzing proline isomerization of the pCTD substrate.

Finally, to evaluate the effect of Ess1 on Ssu72 activity toward its natural
substrate, we performed in vitro reactions using the full-length, 26-repeat CTD from
Saccharomyces cerevisiae. A GST-yCTD-His fusion construct was hyperphosphorylated
with the yeast CTD kinase CTDK-1 (Phatnani et al. 2004) and dephosphorylation by
Ssu72 was monitored by measuring substrate depletion with a pS5-specific antibody in
reactions with and without 100 μM Ess1 (Figure 12E-F). As in the reactions with
synthetic pCTD substrate, Ess1 greatly facilitates dephosphorylation by Ssu72,
decreasing the time required to reach ~90% completion by 20-fold (from 20 minutes to 1
minute).

### 2.3.3 Reaction Timescale Affects Stimulation by Ess1

Reaction timescale is likely a critical determinant of the effect of Ess1-catalyzed
proline isomerization on dephosphorylation by Ssu72. For example, we observed only a
~2-fold enhancement of Ssu72-catalyzed product accumulation in reactions with
synthetic pCTD substrate, as a significant amount of non-catalyzed proline
isomerization occurs over the course of the five minute reactions (Figure 12A-B).

Notably, a similar level of stimulation was reported for the isomer-specific PP2A

![Graph A](image)

![Graph B](image)

![Graph C](image)

**Figure 13. Reaction Timescale Affects Stimulation of Ssu72 Activity by Ess1.** The kinetics in this figure are analogous to those in Figure 12A-C. Reactions were performed with 20 μM Ssu72 and measured at 15 second increments. In (A-C), each point represents the average from three independent reactions and error bars denote standard deviations. (A) The activity of Ssu72 was monitored with various concentrations of the proline isomerase Ess1. The maximum stimulation was 3.4-fold under these conditions. (B) Ess1 enhancement of Ssu72 activity is saturable. The last points of the reactions in (A) are plotted versus Ess1 concentration, with black circles representing reactions omitted from (A) for clarity. (C) pCTD dephosphorylation by Ssu72 reaches completion without Ess1. The ‘no Ess1’ reaction in (A) was monitored over 10 minutes, and reaches ~90% completion in 10 minutes, 8-fold slower than the reaction with 20 μM Ess1.
phosphatase described in Section 2.4.2, in reactions on the same timescale (Zhou et al. 2000). However, at a faster timescale the effect of Ess1 on Ssu72 activity should be greater, as the baseline of non-catalyzed proline isomerization is reduced. To verify this hypothesis, we performed kinetic assays with synthetic substrate peptide analogous to those describe above, but at a 4-fold faster timescale. Reactions were performed with 20 μM Ssu72 and measured at 15 second increments. Under these conditions, the maximum stimulation of product accumulation by Ess1 was 3.4-fold (Figure 13A-B), and the reaction without Ess1 required 10 minutes to reach ~90% completion, 8-fold slower than the reaction with 20 μM Ess1 (Figure 13C). Therefore, the shorter timescale significantly increased the stimulatory effect of catalyzed proline isomerization as expected, both in terms of product accumulation and average reaction rate. These reactions also effectively illustrate the ‘burst phase’ in the absence of catalyzed proline isomerization, which results from dephosphorylation of the initial cis proline substrate population. In the reaction without Ess1, the first ~14% of substrate is hydrolyzed in the first 15 seconds, followed by a slow and nearly linear rate of reaction that requires a full minute to hydrolyze the second 14% of substrate (15 seconds to 1.25 minutes).
2.4 Summary, Implications, and Future Directions

The CTD phosphatase Ssu72 is an unusual enzyme. Despite being built on the scaffold of a protein tyrosine phosphatase, its activity is directed at phosphorylated serine residues in the 5th position of the CTD heptad, with isomeric specificity for the cis configuration of the pS5-P6 peptide bond. Our complex structure also suggests a conserved catalytic mechanism, which agrees well with prior kinetic studies (Meinhart et al. 2003). Of the two additions to the LMW PTP scaffold, the insert subdomain plays a major role in the unique substrate specificity of Ssu72, contributing nearly all of the residues that form the non-phosphate substrate recognition surface and severely restricting access to the active site. Very recently, it was discovered that the second addition of Ssu72, the C-terminal extension, forms the binding site with symplekin (homologous to Pta1 in yeast), a scaffold protein in the cleavage and polyadenylation complex (see Section 2.4.4) (Xiang et al. 2010). Thus, the distinguishing structural features of Ssu72 mediate two of its integral functions and define its place in a unique subfamily of the LMW PTP enzymes.
2.4.1 Substrate Specificity of Ssu72

Prior kinetic work on Ssu72 has demonstrated a strict specificity for the pS5 position of the CTD heptad (Hausmann et al. 2005; Krishnamurthy et al. 2004), in contrast to CTD phosphatases Scp1 and Fcp1, which have only preferential activity for pS5 and pS2, respectively (Hausmann and Shuman 2002; Yeo et al. 2003). Interestingly, both potential sites are pSer-Pro motifs, meaning that two of the four residues recognized by Ssu72 are identical in the pS2 and pS5 substrates. Therefore, discrimination must be based on the T4 and distal Y1 positions. For a pS2 substrate, the Y1 position would be substituted by serine, causing a decrease in binding energy from the loss of the aromatic-amide stacking interaction. However, the T4 position is likely the more important determinant of substrate specificity. The T4 residue would be replaced by tyrosine in a pS2 substrate, which would disrupt substrate recognition in two major ways. First, the favorable hydrophobic interactions with the insert subdomain would be lost and the accommodation of the large phenol group would require a significant reconfiguration of the substrate backbone to prevent steric clashes. Second, this substrate would be unable to replicate the intramolecular hydrogen bond we observe between the T4 hydroxyl and the P6 backbone carbonyl, making the already unfavorable cis conformation even less energetically stable.
Although the exact degree of isomeric specificity of Ssu72 is unclear, our structural and kinetic studies demonstrate a strong preference for the cis pS5-P6 conformation. First, modeling the trans pS5-P6 isomer in the Ssu72-pCTD complex structure leads to large steric clashes between the pCTD backbone and the sidechains of the C-terminal α-helix of the insert subdomain. Also, if the activity toward cis and trans isomers were comparable, there would be no stimulation of dephosphorylation by Ess1. Finally, additional support for strong cis proline specificity comes from a kinetic study that measured activity toward a set of pS5 CTD substrates with an alanine substitution at each heptad position (Hausmann et al. 2005). These experiments showed that CTD positions T4, P6, and Y1 of the following heptad were critical for substrate recognition, a result that is consistent with the intermolecular contacts in our complex structure. Importantly, the largest decrease in Ssu72 activity was observed with the P6A mutant substrate, a ~30-fold reduction compared to wild-type. This likely reflects not only the loss of binding energy for the proline ring, but also the reduced propensity of the pS5-A6 peptide bond to adopt a cis configuration.

2.4.2 Regulation of Isomer-Specific Enzymes

To the best of our knowledge, Ssu72 is the first example of an enzyme with specificity for the cis isomer of proline. Isomeric specificity for trans pSer-Pro motifs has
been previously established for two enzymes. The serine/threonine phosphatase PP2A dephosphorylates \textit{trans} pSer-Pro motifs in its substrate Cdc25C and Tau proteins and shows increased activity in reactions that include Pin1 (Zhou et al. 2000). Likewise, the serine/threonine kinase ERK2 phosphorylates \textit{trans} Ser-Pro motifs in the RNase T1 substrate and is also stimulated by proline isomerase activity (Weiwd et al. 2004). Importantly, catalyzed proline isomerization only increases the activity of these \textit{trans}-specific enzymes toward (p)Ser-Pro motifs that adopt the \textit{cis} conformation in their protein’s structure. For example, ERK2 phosphorylates two sites in RNase T1, \textit{cis} Ser54-Pro55 and \textit{trans} Ser72-Pro73; however, only phosphorylation at Ser54-Pro55 is enhanced by the addition of PPIs. This illustrates a crucial point: catalyzed proline isomerization is an effective regulatory mechanism of phosphorylation/dephosphorylation only when the target motif exists predominantly in the opposite isomeric configuration required for catalysis. In this case, the isomer-specific enzyme can only process a small fraction of the total substrate, and the extremely slow, uncatalyzed \textit{cis-trans} isomerization prevents re-equilibration of the non Processed fraction. In the opposite scenario, the majority of substrate is already in the correct isomeric form and can be readily processed, with catalyzed proline isomerization having little influence.

The effect of proline isomerization on CTD-associated activities depends greatly on the CTD conformation during the transcription cycle. All structural studies of CTD-binding proteins have revealed binding that is, in fact, isomer specific, but preferential
for trans (p)Ser-Pro motifs. In the case of Pcf11, the CTD-interacting domain
discriminates in favor of three consecutive trans prolines, including one pS2-P3 motif
(Noble et al. 2005). In the structure of CTD phosphatase Scp1 in complex with pS2/pS5
CTD, both the pS2-P3 and pS5-P6 motifs adopt the trans configuration, with binding to
either cis form seemingly prevented by steric clashes (Zhang et al. 2006). For both of
these proteins, the regulation of binding or catalysis by proline isomerization would
demand a CTD conformation with a minority trans population. This is especially true
for non-enzymatic proteins like Pcf11, which may only require one correctly configured
binding site out of dozens of CTD repeats. However, our measurements with a pS5
CTD peptide and prior work with a pS2 CTD peptide show that cis proline is the less
populated isomer (Noble et al. 2005). Given the lack of evidence for any stable full-
length CTD structure, it seems likely that cis prolines constitute the isomeric minority in
vivo as well, and that proteins with specific recognition of cis proline are the major
targets of regulation by isomerization.

The timescale of transcriptional events is another critical determinant of the effect
of Ess1 on CTD-associated activities. In general, proline isomerization should have
maximum influence on fast biological processes in which isomer populations do not
have sufficient time to naturally re-equilibrate. As discussed in Section 2.3.3, non-
catalyzed proline isomerization provided a substantial baseline supply of cis proline
substrate for Ssu72 in our original 5 minute reactions (Figure 12). In contrast, at the
faster timescale of the 1.25 minute reactions (Figure 13), the amount of non-catalyzed isomerization was smaller, thereby allowing a substantially stronger stimulation of Ssu72 activity by Ess1. While the exact in vivo timescale of Ssu72 activity is unknown, the effect of impaired Ess1 activity on Nrd1 localization and Pcf11 recruitment is largely confined to the 3'-end of genes (Singh et al. 2009), suggesting that pS5 CTD dephosphorylation by Ssu72 occurs mainly during termination, which may take only tens of seconds (Femino et al. 1998). This scenario is well-illustrated by the kinetic assays with full-length pCTD in Figure 12E-F, where the reaction without Ess1 takes 20 minutes to dephosphorylate ~90% of the initial pCTD substrate, while the reaction with Ess1 reaches ~90% completion in 1 minute, on par with the estimated timescale of transcription termination.

2.4.3 Implications of Proline Isomerization in CTD Biology

Our kinetic experiments provide an excellent interpretation of recent in vivo studies of Ess1 (Krishnamurthy et al. 2009; Singh et al. 2009). We show that the stimulation of Ssu72 activity by Ess1-catalyzed cis-trans interconversion of pS5-P6 motifs in the CTD heptad is severely reduced with catalytically impaired Ess1 mutants, which explains the in vivo accumulation of pS5 CTD in ess1 cells. Interestingly, overexpression of Ssu72 suppressed this phenotype in an A144T ess1 strain, but not a H164R ess1 strain.
(Krishnamurthy et al. 2009). It is possible that unequal levels of Ssu72 expression or different residual activities or thermostabilities of the Ess1 mutants are responsible for this discrepancy. In contrast, overexpression of Ssu72 rescues the temperature sensitivity of both ess1 strains (Krishnamurthy et al. 2009), suggesting that improper

![Figure 14. A New Model for Ess1-Regulated Ssu72 Activity.](image)

RNAP II is shown near termination of a small noncoding RNA transcript. For clarity, only one CTD heptad is labeled. The upper (gray) and lower (light blue) panels represent the \textit{trans} and \textit{cis} forms, respectively, of the pS5-P6 motif. Dephosphorylation cannot proceed through the \textit{trans} form (red cross) due to the \textit{cis} proline substrate specificity of Ssu72. Instead, Ess1 catalyzes the interconversion of isomers which is the rate limiting step for Ssu72 activity. The minor pathway is indicated with gray arrows, with the dashed line representing non-catalyzed isomerization that occurs very slowly compared to transcripational events. The change in CTD phosphorylation facilitates proper termination by promoting the release of Nrd1 and recruiting Pcf11.
regulation of CTD phosphorylation is the main cause of this phenotype. As described in Section 1.4.3, the proposed proline isomer-based regulatory scheme that explains these results requires \textit{trans} specific activity by Ssu72 (Krishnamurthy et al. 2009; Singh et al. 2009). Our results compel a simple but unexpected adjustment of this model: reversing the isomeric specificity of Ssu72 (Figure 14).

Ssu72 provides an exciting validation of proline isomers as a critical component of the CTD code, but also raises many new questions about how this regulatory mechanism operates. First, it is unclear whether the regulation of Ssu72 is a general property of RNAPII transcription or a gene-specific one. Genome-wide expression analyses of catalytically-impaired Ess1 and Ssu72 mutants show altered transcription levels for a limited number of genes, mainly those downstream of small noncoding RNA transcripts (Ganem et al. 2003; Singh et al. 2009). This likely results from improper release of pS5 CTD-bound Nrd1 and concomitant blocking of Pcf11 recruitment, a defect specific to genes that undergo Nrd1-dependent transcription termination (Singh et al. 2009). As described in Section 1.3, pS5 CTD predominates across these small noncoding RNA transcripts while longer genes see decreasing pS5 levels toward the 3’ end, and this difference in phosphorylation states appears to dictate whether termination occurs by the Nrd1-dependent pathway or the standard mRNA pathway (Gudipati et al. 2008). It is possible then that Ssu72 plays a role in the termination of all RNAPII-transcribed
genes, but that defects in its activity or its regulation by Ess1 are only apparent for genes that require a sudden drop in pS5 levels for proper termination.

It is also unknown whether catalyzed proline isomerization regulates a specific phase of transcription. Genetic studies indicate a role for Ess1 activity in transcription initiation, elongation, and termination, which is consistent with a broad distribution of Ess1 crosslinking along snoRNA genes (Singh et al. 2009; Wilcox et al. 2004). However, another study showed that Ess1 occupies the promoter and terminator regions of the PMA1 and PHO5 genes (Krishnamurthy et al. 2009). Importantly, it is possible that defective regulation of Ssu72 activity at termination has a deleterious impact on the recycling of RNAPII and its subsequent initiation. This would account for the effects of Ssu72 and Ess1 on both termination and initiation and would also explain the requirement of Ssu72 activity for gene looping (Ansari and Hampsey 2005).

Another interesting question is whether isomerization regulates only CTD phosphorylation or also the association of an undiscovered subset of CTD-binding factors with cis proline recognition. A large amount of genetic evidence ties proline isomerase activity to the function of CTD kinases and phosphatases and the proper maintenance of the CTD phosphorylation state, including a report that Pin1 stimulates CTD phosphorylation by cdc2/cyclin B in vitro (Xu et al. 2003). While the mechanism of this effect has not been elucidated yet, cdc2/cyclin B may represent another example of a cis proline-specific enzyme that is regulated by proline isomerization. The impact of
impaired Ess1/Pin1 activity on non-enzymatic CTD-binding proteins, like Nrd1 and Pcf11, is likely an indirect result of misregulated phosphorylation. However, it is possible that there are other proteins that specifically recognize cis proline motifs in the CTD and whose association with the transcribing polymerase is regulated by Ess1/Pin1 in order to fine-tune the recruitment of their co-transcriptional activities.

Finally, it is tempting to speculate on the cellular ubiquity of proline isomerization as a mechanism for regulating phosphorylation events. The potential reach of this mechanism is underscored by the pervasive influence of Ess1/Pin1, which is specific for phosphorylated Ser/Thr-Pro motifs, on a diverse set of biological processes, although the effect of proline isomerization on its target proteins is still poorly understood and may not be related to the regulation of phosphorylation states (see Section 1.2.3). Aside from the CTD, isomer-based regulation of phosphorylation has been established for three target proteins, Cdc25C, Tau, and RNase T1 (Weiwad et al. 2004; Zhou et al. 2000), and the disparate functions of these targets provides another indication of a general regulatory strategy. While the biological breadth of this mechanism remains unknown, there is an unquestionably sufficient capacity for a widespread impact, as over 25% of sites identified in global phosphorylation studies are Ser/Thr-Pro motifs and several major classes of Ser/Thr kinases and phosphatases have proline-directed activity (Ubersax and Ferrell 2007). Therefore, it will be exciting to see...
how many other proteins recognize proline motifs with biologically-relevant isomeric specificity and how their functions are altered by catalyzed isomerization.

**2.4.4 Comparison of Recent Ssu72 Structures**

In a relatively short span of time, several Ssu72 structures were reported, including the one featured in this dissertation. The Northeast Structural Genomics Consortium solved the structure of apo *Drosophila melanogaster* Ssu72 in two crystal forms. An orthorhombic form with four molecules in the asymmetric unit diffracted to 3.2 Å resolution and was used as a search model for molecular replacement in our structural study, and a monoclinic form with eight molecules in the asymmetric unit diffracted to 3.3 Å resolution. The large number of structures present in these two crystals provides an opportunity to investigate possible sites of conformational heterogeneity, with the important caveat that any observed structural variations may result from crystal packing and therefore may not reflect natural motions of the enzyme in solution. A structural alignment of the apo dSsu72 structures reveals a very uniform core fold, with only two sites of local structural differences (Figure 15). The first is in the loop between the β-strands of the insert subdomain, which contacts the aromatic ring of the pCTD substrate residue Y1 in our complex structure. The second is in the loop...
preceding the final β-strand of the core fold and is far from both the active site and the substrate-binding interface. Both sites also exhibit elevated crystallographic B-factors.

The structure of *Drosophila melanogaster* Ssu72 in complex with the inhibitor vanadate was reported by Zhang and co-workers at 2.35 Å resolution (Zhang et al. 2011). Vanadate is a vanadium-based oxyanion compound that inhibits cysteine-based phosphatases by forming a complex that mimics the phosphoenzyme intermediate of the dephosphorylation reaction (Lindqvist et al. 1994). In its complex with Ssu72, vanadate binds covalently to the active site cysteine nucleophile and adopts a trigonal

---

**Figure 15. Overlay of Apo Ssu72 Structures.** Apo structures of *Drosophila melanogaster* Ssu72 from the asymmetric units of two different crystal forms are aligned globally. The two main sites of local conformational heterogeneity are the loop between the insert subdomain β-strands (red arrow) and the loop before the final β-strand of the core fold (orange arrow). Structures are colored by B-factor, with a spectrum of blue to red denoting low to high values. Three structures with uniformly high B-factors were omitted for clarity.
bipyramidal geometry. This study also reported kinetic data for the wild-type and mutant enzymes, including three mutations, P46A, L82A, and M85A, of residues that form the binding pocket for the ring moiety of the substrate cis proline residue (P6) in our complex structure. Interestingly, only the P46A and M85A mutants were found to have significantly impaired enzymatic activity toward synthetic pCTD peptide substrate (Zhang et al. 2011).

Finally, Xiang et al. solved a 2.4 Å resolution structure of human Ssu72 (hSsu72) in a ternary complex with pCTD substrate and the N-terminal domain of symplekin, the higher eukaryote homolog of yeast Pta1 that anchors Ssu72 to the cleavage and polyadenylation complex (Xiang et al. 2010). The Ssu72 fold and cis proline substrate conformation are consistent with our findings, and the authors also used analogous enzymatic assays to demonstrate a similar kinetic enhancement of dephosphorylation by catalyzed proline isomerization. The N-terminal domain of symplekin forms an arc-shaped arrangement of seven pairs of anti-parallel α-helices, which constitutes a HEAT domain, a common structural motif often involved in mediating protein-protein interactions. The concave face of symplekin binds to Ssu72 through the C-terminal extension of its core LMW PTP fold, assigning a functional role for this unique structure feature. The symplekin binding site is ~25 Å from the active site of Ssu72 and the binding of symplekin does not cause any major changes in the Ssu72 structure; however, the N-terminal domain of symplekin was found to stimulate the phosphatase activity of
Ssu72 toward pCTD peptide substrate in vitro, either through a general stabilization of the enzyme’s structure or some other undetermined allosteric mechanism, possibly involving protein dynamics (Xiang et al. 2010).

A structural alignment of our *Drosophila melanogaster* Ssu72-pCTD complex with the apo and vanadate complex forms from the same species and the analogous human complex reveals a remarkable conformational homogeneity across binding states and species (Figure 16). As observed in the multiple apo form structures (Figure 15), the two main sites of local variations are (1) the loop between the \( \beta \)-strands of the insert subdomain and (2) the loop preceding the final \( \beta \)-strand of the core fold. The structural details of the active sites, including the conformations of the phosphate-binding loops and positions of the catalytic residues, are also highly consistent. In the vanadate complex, the vanadate molecule sits ~0.6 Å lower in the active site compared to the pCTD substrate’s phosphate group, in closer resemblance to the phosphoenzyme intermediate state, while the catalytic aspartate sidechain is turned away from the active site, presumably to minimize electrostatic repulsion with the vanadate molecule. The pCTD conformations in the human and *Drosophila melanogaster* complexes have two minor discrepancies. First, the N-terminal proline residue (P3) is oriented differently which is unsurprising given its very minor contacts with the enzyme (see Section 2.2.2). Second, although the phosphate group positions are very similar, the sidechain of the phosphorylated serine residue (pS5) is fit to different rotamers, which likely reflects the
Figure 16. Comparison of Ssu72 Structures. The structure of the dSsu72-pCTD complex described in this dissertation is aligned with apo dSsu72 (A), the dSsu72-vanadate complex (B), and the hSsu72-pCTD complex (C). In (A-C), the dSsu72-pCTD complex is colored blue. Left panels show the full structures in ribbon format, while right panels feature close-ups of the active site and substrate-binding interface with the active site loop and catalytic residues shown as sticks. The catalytic cysteine and aspartate residues are labeled according to the wild-type Drosophila melanogaster sequence.
insufficient resolution of the models’ electron density to unambiguously define its conformation.

2.4.5 Proposals for Future Studies

Despite the recent findings detailed in this chapter, many aspects of Ssu72’s function have yet to be fully elucidated, including several that would benefit from further structural studies. For example, symplekin is reported to significantly stimulate in vitro pS5 CTD dephosphorylation by Ssu72; however, the Ssu72-symplekin complex structure shows that the symplekin binding site is far from the active site of Ssu72 (~25 Å) and that complex formation does not cause any significant changes in the Ssu72 structure (Xiang et al. 2010), suggesting that dynamics may play a dominant role in the allosteric mechanism. Ssu72 has also been reported to interact with the general transcription factor TFIIB (Dichtl et al. 2002a; Wu et al. 1999), and this interaction may contribute to the physical coupling of initiation factors with the 3’-end processing machinery during gene looping (Singh and Hampsey 2007). NMR and our reported assignments for Ssu72 should provide important tools for mapping protein-protein interaction sites and for understanding how these binding events affect the structure and dynamics of Ssu72 in order to modulate its role in transcription.
While the majority of biological studies of Ssu72 have focused on the yeast enzyme, all of the structural information reported to date is for the *Drosophila melanogaster* and human orthologs. The Ssu72 primary sequence is highly conserved among eukaryotes (Figure 17), with 43% identity between yeast and *Drosophila melanogaster*, which suggests a high degree of structural similarity. However, this sequence conservation is not evenly distributed; the C-terminal half of the yeast protein is significantly more divergent and includes an 8 residue insertion in the loop preceding the C-terminal extension to the core LMW PTP scaffold, which is likely located near or at the binding interface with Pta1. Therefore, structural characterization of the yeast Ssu72 may provide important species-specific insights into its biological function. Our initial investigations of yeast Ssu72 were thwarted by aggregation at the high concentrations required for structural work, which may have resulted from protein instability or unfolding. However, it is possible that this obstacle may be overcome by the addition of vanadate during purification and final concentration of samples. Vanadate is a vanadium-based phosphate analog that inhibits LMW PTPs by binding to the catalytic cysteine nucleophile in a manner that mimics the intermediate of the dephosphorylation reaction (Lindqvist et al. 1994). Recently, vanadate binding was found to significantly improve the thermostability of several Ssu72 orthologs (Zhang et al. 2011). Additionally, incorporation of vanadate stabilized crystals of the apo *Drosophila melanogaster* enzyme, improving the resolution and lowering thermal factors (Zhang et al. 2011). Although
vanadate binding would preclude interaction with pCTD substrate, it may at least provide an avenue for determining the structural features of yeast Ssu72 and their relation to higher eukaryotic counterparts.

There is also an important enzymatic point that remains unsettled. Our kinetic assays demonstrate that Ssu72 has a strong preference for the cis pS5-P6 isomer of its pCTD substrate, and modeling of our Ssu72-pCTD complex structure with a trans pS5-P6 motif leads to large intermolecular steric clashes (see Section 2.4.1). Together, these findings suggest that Ssu72 is able to dephosphorylate only the cis proline form of its substrate; however, we were unable to measure its isomeric specificity directly. A potential tool for assessing this specificity is provided by the recent development of conformationally-locked analogs of cis and trans pSer-Pro motifs (Wang et al. 2004). These synthetic compounds can be incorporated into longer peptides, where they adopt single, inconvertible isomer states that closely mimic the structures of their natural proline counterparts. In support of this structural conservation, phosphorylated peptides containing cis and trans proline-locked analogs effectively inhibit proline isomerase Pin1 in a competitive manner (Wang et al. 2004). Furthermore, crystal structures of Pin1 in complex with the inhibitors show that they bind in the active site with an orientation similar to the natural substrate (Etzkorn et al. 2005). Assaying the activity of Ssu72 toward pS5 CTD peptides containing cis and trans proline-locked analogs at the P6 position would allow for a direct evaluation of its isomeric specificity.
An even more exciting possible application of these compounds is the identification of additional pCTD-interacting proteins that have cis proline specificity. For example, the cis proline analog could be incorporated into a phosphorylated CTD peptide at either the P3 or P6 position for use as bait in pull-down experiments with yeast nuclear extract, with the equivalent trans proline analog-containing peptide serving as a control, similar to the strategy used previously to identify novel pCTD-interacting proteins that recognize different phosphorylation patterns (Phatnani et al. 2004). This would provide a simple screen for nuclear factors that recognize cis proline residues in the context of the CTD.

Although Ssu72 has been a target of biological studies for over 15 years, its exact role in transcription remains unclear and key questions are left unanswered. For example, does Ssu72 affect only the termination stage of transcription? And is its phosphatase activity required for the proper transcription and co-transcriptional processing of all gene classes or only small noncoding genes? Most of the extant functional studies of Ssu72 have relied on analyzing the phenotypes of yeast strains harboring mutant ssu72 alleles and were performed before structural information was available. Therefore, it is worthwhile to revisit these mutations and place each of them in the context of the recently reported structures of Ssu72 and its complexes with pCTD substrate and symplekin. This endeavor is subject to two main limitations. First, several of the previously studied ssu72 mutations have not been genetically mapped, and
therefore cannot be interpreted structurally. Second, structural information exists only for the human and *Drosophila melanogaster* Ssu72 orthologs (see Section 2.4.4). However, Ssu72 is highly conserved from yeast to humans (Figure 17), particularly in regions proximal to the active site, with near complete conservation of substrate-binding residues across species. This similarity suggests that the existing structures provide an adequate model for rationalizing the effect of mutations on the enzymatic properties of yeast Ssu72. As described above, the yeast Ssu72 sequence is more divergent in the C-terminal extension responsible for binding to Pta1, which correlates with the weak sequence similarity of Pta1 to its ortholog symplekin (Xiang et al. 2010). Nonetheless, secondary structure predictions show that Pta1 shares the all-helical N-terminal domain of symplekin that interacts with Ssu72, and in support of a common domain function, the N-terminal 300 residues of Pta1 are sufficient for binding to Ssu72 (Ghazy et al. 2009). As such, the structure of the human Ssu72-symplekin complex likely reflects the broad features of the yeast Ssu72-Pta1 interaction accurately, although the atomic details of the complex interface and the effects of mutations on binding may differ.

A summary of the Ssu72 mutations that have been studied in yeast and their associated phenotypes is shown in Table 2, with uncharacterized mutant strains included for completeness. In addition to the degron-tagged *ssu*72 strain and N-terminal and C-terminal extensions, the *ssu*72 alleles include single amino acid mutations at three sites, G33, G42, and R129. Figure 17 shows the position of the
Table 2. Summary of Ssu72 Mutations Used in Prior Yeast Studies.

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Yeast Strain</th>
<th>Phenotype</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-terminal duplication</td>
<td>ssu7-1</td>
<td>Interacts genetically with TFIIB mutants affecting start site selection.</td>
<td>(Sun and Hampsey 1996)</td>
</tr>
<tr>
<td>G33A</td>
<td>G33A ssu72</td>
<td>Causes readthrough of both poly(A) and Nrd1-dependent terminators; does not impair cleavage or polyadenylation.</td>
<td>(Steinmetz and Brow 2003)</td>
</tr>
<tr>
<td>G42V</td>
<td>G42V ssu72</td>
<td>Causes readthrough of Nrd1-dependent terminator.</td>
<td>(Steinmetz and Brow 2003)</td>
</tr>
<tr>
<td>G42D</td>
<td>ssu7-3319</td>
<td>Interacts genetically with the Cak1, Bur1 and Ctk1 kinases.</td>
<td>(Ganem et al. 2003; Ganem et al. 2006)</td>
</tr>
<tr>
<td>R129A</td>
<td>ssu72-2</td>
<td>Impairs catalytic activity and disrupts interactions with TFIIB and Pta1; Causes readthrough of both poly(A) and Nrd1-dependent terminators; reduces elongation efficiency and in vitro 3'-end cleavage activity.</td>
<td>(Dichtl et al. 2002a; He et al. 2003; Kim et al. 2006; Pappas and Hampsey 2000; Reyes-Reyes and Hampsey 2007)</td>
</tr>
<tr>
<td>C-terminal TAP-tagged</td>
<td>n/a</td>
<td>Causes readthrough of both poly(A) and Nrd1-dependent terminators.</td>
<td>(Nedea et al. 2003)</td>
</tr>
<tr>
<td>degron-tagged Ssu72</td>
<td>ssu72-td</td>
<td>Contains heat-inducible degradation tag at N-terminus. Reduces in vitro 3'-end cleavage and polyadenylation activity; abolishes gene looping and TFIIB crosslinking to the terminator.</td>
<td>(Ansari and Hampsey 2005; Govind et al. 2007; He et al. 2003; Krishnamurthy et al. 2004; Singh and Hampsey 2007)</td>
</tr>
<tr>
<td>unknown</td>
<td>ssu72-ts69</td>
<td>Causes readthrough of both poly(A) and Nrd1-dependent terminators.</td>
<td>(Ganem et al. 2003)</td>
</tr>
<tr>
<td>unknown</td>
<td>ssu72-3</td>
<td>Causes readthrough of Nrd1-dependent terminators; reduces in vitro 3'-end cleavage activity.</td>
<td>(He et al. 2003; Kim et al. 2006)</td>
</tr>
<tr>
<td>unknown</td>
<td>ssu72-7</td>
<td>Reduces in vitro 3'-end cleavage activity.</td>
<td>(He et al. 2003)</td>
</tr>
</tbody>
</table>

The R129A mutation was designed before the phosphatase activity of Ssu72 was discovered and was meant to target a signature motif found in ATP-dependent RNA helicases (Pappas and Hampsey 2000). It has since become the most studied Ssu72 point mutant. The R129 residue in human Ssu72 is located on the face of the enzyme that interacts with symplekin (Figure 17), with its side chain packed into the interior of the corresponding residues in the structure of the human ortholog of Ssu72. The following examination of the possible structural consequences of these mutations and their potential effects on the functions of Ssu72 illustrates the difficulty of interpreting resulting phenotypes.
Figure 17. Structural Analysis of Ssu72 Mutations Used in Prior Yeast Studies. (A) Sequence alignment of Ssu72 orthologs. Conserved and similar residues are highlighted in orange and yellow, respectively, and catalytic residues are highlighted in red. Asterisks denote residues in human Ssu72 that interact with substrate (red) or symplekin (purple), while blue crosses denote sites of mutations in the yeast protein that have been studied in vivo. (B) Mapping of interaction sites and mutations on the structure of Ssu72. Human Ssu72 (tan ribbons) is shown in its tertiary complex with pS5 CTD substrate (green sticks) and symplekin (gray ribbons). The symplekin structure is omitted from the right panel for clarity. Residues that interact with substrate and symplekin are highlighted to correspond with the asterisk colors in (A). Mutations studied in ssu72 yeast strains are colored blue and labeled with the corresponding yeast residue numbers.
enzyme. Although the mutation is far from the active site and substrate-binding surface, purified GST-tagged R129A ySu72 was reported to have <40% of wild-type phosphatase activity toward pNPP substrate (Reyes-Reyes and Hampsey 2007). While the location of the R129 residue suggests that its mutation may disrupt binding to Pta1, the equivalent mutation in human Su72 does not impair its interaction with symplekin (Xiang et al. 2010). However, as described above, it is possible that this protein-protein interface is not perfectly conserved between yeast and higher eukaryotes, and consistent with this idea, the R129A mutation was found to impair Pta1 binding in pull down experiments with in vitro-translated proteins (Dichtl et al. 2002a). Phenotypes of the R129A su72 strain may therefore be the result of reduced Su72 catalytic activity, impaired binding between Su72 and Pta1, or some combination of both.

The second mutation site, the G42 residue, is invariant across Su72 orthologs and is located near the active site. While it does not directly contact substrate, the residue’s backbone amide group forms a hydrogen bond with the backbone carbonyl of one of the active site loop residues (A16 in the yeast ortholog) that cradle the substrate phosphate group. Two spontaneously arising mutations have been identified at this site, G42V and G42D. Based on the known structures, the valine or aspartate sidechain would likely extend toward the substrate-binding groove, possibly affecting substrate recognition. G42D ySu72 is described in the literature as a gain-of-function mutant with increased catalytic activity (Ganem et al. 2003; Ganem et al. 2006); however, this
conclusion is based on experiments measuring pNPP hydrolysis of soluble crude extracts from *E. coli* expressing either wild-type or mutant forms of Ssu72, without standardization of enzyme concentrations. Finally, the site of another spontaneously arising mutation, G33A, is a well conserved residue located far from both the active site and substrate-binding cleft and the Pta1-binding interface. The known structures suggest that the mutant alanine sidechain would project into solvent, making it difficult to decipher how this mutation alters the enzyme’s stability, kinetics, or interaction with Pta1.

The structural analysis of characterized Ssu72 mutations used in prior yeast studies suggests that none cleanly disrupt a single aspect of the enzyme’s function, which complicates the interpretation of resulting phenotypes in two major ways. First, the two functions of Ssu72, its pS5 CTD phosphatase activity and its interaction with the N-terminal domain of scaffolding protein Pta1, may have distinct and independent biological roles. For example, the N-terminal domain of Pta1 inhibits mRNA 3’-end processing, and this inhibition is masked by binding to Ssu72 in a manner that does not require its phosphatase activity (Ghazy et al. 2009; He et al. 2003; Krishnamurthy et al. 2004). The fact that extant mutations likely affect both functions of Ssu72 makes it impossible to cleanly separate the phenotypes associated with each function. A second complication is that the cellular levels of Ssu72 and Pta1 are tightly coupled. Degron-mediated degradation of Ssu72 in yeast results in the loss of Pta1 (He et al. 2003), and
degron-mediated degradation of Pta1 leads to the loss of Ssu72 (Ghazy et al. 2009; Krishnamurthy et al. 2004). Moreover, deletion of only the N-terminal domain of Pta1 also results in the loss of Ssu72, suggesting that the in vivo stability of both proteins depends on their interaction (Ghazy et al. 2009). This relationship has the potential to distort conclusions drawn from previous mutant studies, as phenotypes associated with degron-tagged ssu72 strains may result from the loss of either Ssu72, Pta1, or both. Additionally, any mutation in Ssu72 that affects its stability or its interaction with Pta1 may lead to a concomitant loss of both proteins and a scrambling of phenotypes.

Recent investigations of the structure and kinetics of Ssu72, including the work described in this chapter, should improve our understanding of the biological importance of Ssu72 through the structure-guided design of mutants that perturb its function in a precise and well-defined manner. In particular, mutations that impair the catalytic activity of Ssu72 without disrupting its interaction with Pta1 should be particularly useful. At minimum, genome-wide analysis of mRNA expression levels and distributions of CTD phosphorylation patterns in these ssu72 strains would help to more clearly characterize the timing of Ssu72 activity in the transcription cycle and the breadth of its function across different gene classes, while in vitro assays for cleavage and polyadenylation could be used to confirm the independence of its catalytic activity for 3′-end processing. These second generation in vivo experiments will likely be
essential for providing a more detailed and comprehensive picture of the role of Ssu72 in coupling transcription with co-transcriptional events.

2.5 Materials and Methods

2.5.1 Molecular Cloning and Protein Expression and Purification

The full-length WT *Drosophila melanogaster* Ssu72 gene was PCR-amplified from cDNA (OpenBiosystems, clone ID: RE29729), digested, and ligated into a pET15b vector (EMD Biosciences, Inc.) between the Nde1 and BamH1 restriction sites. Point mutants of Ssu72 and Ess1 were prepared using the QuikChange Site-Directed Mutagenesis Kit (Stratagene). The correct sequences for all constructs were verified by DNA sequencing. Both Ssu72 and Ess1 were overexpressed as N-terminal 6X His-tagged constructs in *Escherichia coli* Rosetta 2 cells (EMD Biosciences, Inc.). For Ssu72, bacterial cells were induced overnight with 0.5 mM IPTG at 25°C; for Ess1, cells were induced for 3 hours with 1.0 mM IPTG at 37°C. Proteins were purified with Ni²⁺-NTA resin (Qiagen) and His tags were removed by thrombin digestion. A second Ni²⁺-NTA column was used to remove uncleaved protein, followed by a size exclusion column (Superdex 200, GE Healthcare) equilibrated with 25 mM Tris-HCl, 25 mM KCl buffer at pH 8.0. Samples of dSsu72 for kinetic experiments were exchanged into a similar buffer with 10% glycerol.
and 2 mM DTT. Crystallography samples were stored in 25 mM HEPES, 25 mM KCl, 2 mM DTT buffer at pH 8.0. Ess1 samples were exchanged into kinetics buffer. All proteins samples used for crystallography and kinetics were flash frozen in liquid nitrogen and stored at -80°C. Contrary to a previous report (Meinhart et al. 2003), dSsu72 was found to be monomeric by analytical ultracentrifugation in a variety of buffer conditions. Proteolytic removal of the Ssu72 His tag leaves three extra residues N-terminal to the natural protein sequence, which are disordered and do not contribute peaks to the $^1$H/$^{15}$N-HSQC-TROSY spectrum. In this dissertation, residue numbering corresponds to the natural protein sequence. Circular dichroism experiments for Ess1 mutants were performed in a cacodylate buffer system and thermal denaturation was monitored at 222 nm.

2.5.2 CTD Peptide Preparation and Characterization

Lyophilized pCTD peptide with the sequence Ac-PTpSPSYS-NH$_2$ was purchased at >98% purity (CPC Scientific, Inc.). Stocks were made at neutral pH with concentrations of about 100 mM by resuspension in distilled water. Initial estimates of concentration were calculated using the Abs$_{280nm}$. Final concentrations were obtained by measuring the peptide’s phosphate content using the base hydrolysis method and the Malachite green assay with 25 μL sample volumes (Ekman and Jager 1993). Standard
curves were constructed from 10 points measured in triplicate; linear fits to these points were excellent with $R^2$ values above 0.98. To measure proline isomer populations, a natural abundance $^{1}H/^{13}C$-HSQC experiment was collected overnight with a 10 mM peptide sample in 25 mM Na$_2$PO$_4$, 50 mM KCl, 100% D$_2$O buffer at pH 6.5. Proline resonances were assigned by collecting the same experiment with a shorter, single-proline peptide (TpSPSY).

2.5.3 NMR

Isotopically enriched dSsu72 samples were prepared from cells grown in M9 minimal media with $^{15}$N-NH$_4$Cl and $^{13}$C-glucose as the sole nitrogen and carbon sources, respectively. Perdeuterated Ssu72 was expressed in D$_2$O M9 minimal media with $^{15}$N-NH$_4$Cl and $^2$H/$^{13}$C-glucose, with the addition of 85 mg [3-$^2$H] $^{13}$C-α-ketoisovalerate and 50 mg [3,3-$^2$H$_2$] $^{13}$C-α-ketobutyrate ~1 hour prior to induction for selective protonation of ILV methyl groups (Goto et al. 1999). A sample prepared from 10% $^{13}$C-glucose M9 minimal media was used to stereospecifically assign valine and leucine methyl groups (Neri et al. 1989). All isotopes were purchased from Cambridge Isotope Laboratories, Inc. NMR samples for resonance assignment were prepared by extensive buffer exchange into 25 mM sodium phosphate pH 8.0, 25 mM KCl, 2 mM DTT, with either 5% or 100% D$_2$O, and brought to a final protein concentration of ~1 mM. NMR samples for
titrations with inorganic phosphate and substrate peptide were prepared in 25 mM Tris-HCl, 25 mM KCl, 2 mM DTT and 5% D$_2$O. NMR data were collected on 600 and 800 MHz Varian Inova spectrometers equipped with triple-resonance, cryogenically-cooled probes at 30°C. FIDs were processed with NMRPIPE (Delaglio et al. 1995) and datasets were analyzed with CARA (Keller 2004).

2.5.4 Crystallization and Structure Determination

The Ssu72-pCTD complex was prepared by incubating a 5:1 molar ratio of pCTD peptide (4.17 mM) to C13D/D144N dSsu72 (0.83 mM) for 30 minutes on ice. The complex was crystallized by hanging drop vapor diffusion; 1 μL of protein sample was mixed with 1 μL of 22% (w/v) PEG monomethyl ether 550, 100 mM imidazole pH 6.5, and 150 mM DL-malic acid and suspended over a reservoir of the same solution. Equilibration at 4°C for ~5 days produced 50 x 50 x 200 μm rhombic-shaped crystals, which were cryoprotected by dipping into a solution of mother liquor with 25% ethylene glycol before flash-cooling in liquid nitrogen. Data were collected at the SER-CAT facility at Argonne National Laboratory and processed with HKL2000 (Otwinowski and Minor 1997). Initial phases were calculated with Phaser (McCoy et al. 2007) using the crystal structure of apo WT dSsu72 deposited by the Northeast Structural Genomics Consortium (pdb code 3FDF) as a search model. The final structure was obtained by
iterative cycles of manual model-building with Coot (Emsley and Cowtan 2004) and refinement with Phenix using riding hydrogens (Adams et al. 2010). The crystal contains four molecules in the asymmetric unit; clear density was observed for four imidazole molecules which mediate intermolecular packing interactions, and a PEG fragment was fit into a chain of weak density at another packing interface. The excellent quality of electron density for the substrate peptide allows an unambiguous interpretation of its binding mode (Figure 10A). Analysis of the final refined structure with Molprobity (Davis et al. 2007) gave a geometry score of 1.35 (99th percentile) and a clash score of 4.27 (99th percentile). Full statistics for the dSsu72-pCTD structure determination are shown in Table 1, and the final coordinates were deposited in the Protein Data Bank under accession number 3P9Y. In Figure 9, the Lpt1 sequence is aligned with Ssu72 based on its topological similarity using the DALI server (Hasegawa and Holm 2009).

2.5.5 Enzyme Kinetics with Peptide Substrate

Reactions with synthetic pCTD substrate were followed by measuring phosphate release with a standard Malachite Green reagent (Baykov et al. 1988). For experiments with WT Ess1, 350 μL reactions were set up in 50 mM HEPES, 20% glycerol, 1 mM EDTA, 2 mM DTT buffer at pH 6.5 (kinetics buffer) with 60 μM pCTD peptide and
varying amounts of Ess1. After incubating for 10-15 minutes, each reaction was started by the addition of 4 μM WT dSsu72 (or 20 μM for the faster timescale reactions in Figure 13). 50 μL aliquots were taken each minute and mixed with 50 μL of Malachite Green reagent in a 96-well plate, which was placed in the dark for 30 minutes to allow for color development. Color was quantified by measuring Abs\textsubscript{630nm} with a SpectraMax microplate reader (Molecular Devices). Absorbance values were converted to phosphate content using a standard curve made with Na\textsubscript{2}PO\textsubscript{4} solutions of known concentrations. The same protocol was followed for experiments with Ess1 mutants, with reactions containing 1 μM of each mutant.

2.5.6 Enzyme Kinetics with Full-Length Substrate

Unphosphorylated GST-yCTD-His fusion protein was overexpressed in *Escherichia coli*, purified with glutathione and Ni\textsuperscript{2+}-NTA resin, and dialyzed into kinase buffer (25 mM HEPES pH 7.6, 10 mM MgCl\textsubscript{2}) at 4°C. Substrate was hyperphosphorylated using the CTDK-1 kinase as described previously (Phatnani et al. 2004), and purified over a second glutathione column. The activity of Ssu72 was monitored at room temperature with reactions containing ~4 μg of hyperphosphorylated GST-yCTD-His substrate in dephosphorylation reaction buffer (50 mM Tris-HCl pH 6.5, 10 mM NaCl, 10 mM MgCl\textsubscript{2}, 0.5 mM DTT and 0.1 mM EDTA). For
assays with catalyzed proline isomerization, 100 μM Ess1 was preincubated with the reaction mixture for ~20 minutes. Reactions were initiated by the addition of 100 μM dSsu72, and 25 μL aliquots were removed at the specified time points and quenched with SDS loading buffer. Aliquots were resolved on SDS-PAGE gels, and proteins were then transferred to nitrocellulose membranes for Western blot analysis. Primary antibodies were rat monoclonal anti-pS5 (3E8) from D. Eick (Chapman et al. 2007) and rabbit polyclonal (affinity-purified) anti-pS2 (S2) from Bethyl Laboratories, Inc. Infrared dye-labeled secondary antibodies (LI-COR Biosciences) were used to detect and quantify pS5 and pS2 levels with the Odyssey imaging system (LI-COR Biosciences). The plots in Figure 12F represent averages from three independent reactions, with error bars denoting standard deviations.
3. Introduction to NMR Methods for Large Proteins

Nuclear magnetic resonance (NMR) spectroscopy has become one of the most important tools in structural biology and is unique in its ability to probe molecular dynamics and protein structure at atomic resolution in a solution environment. Since the first protein structure was solved by NMR in 1985 (Williamson et al. 1985), rapid advancements at nearly every step of the structure determination process have improved sensitivity, making it possible to study ever larger macromolecular targets (Tugarinov et al. 2004). Isotopic labeling strategies allow the selective observation of defined segments of proteins, specific residue types, or even stereospecific protons (Kainosho and Guntert 2009). Pulse sequences incorporate spin state selective coherence transfers that minimize relaxation losses (Pervushin 2000). Residual dipolar coupling and paramagnetic relaxation enhancement provide alternatives to the nuclear Overhauser effect (NOE) for measuring structural constraints (Bax 2003). Advances in NMR instrumentation, including higher field magnets, cryogenically-cooled probes, and probes for direct detection of heteronuclei, boost the sensitivity of data collection (Felli and Brutscher 2009). These and other improvements turned what was a sensitivity-limited technique into one limited in many cases by spectral resolution and dimensionality, as constrained by experimental measurement times, and helped spark the development of methods for optimizing the speed and efficiency of data collection.
The experiments described in the three chapters following this introduction address current limitations of both sensitivity and resolution/dimensionality that frustrate studies of large proteins. The next chapter introduces a ‘just-in-time’ HN(CA)CO pulse sequence that alleviates the sensitivity impediment of a common backbone assignment experiment. The last two chapters (1) describe novel 4-D NOESY experiments designed for high-resolution, high-dimensionality data collection with sparse sampling methods and (2) evaluate the spectral effects of processing algorithms that remove aliasing artifacts. In light of the wide scope of recent advances in the field of protein NMR, this introduction focuses solely on topics that are most relevant to the experiments detailed in subsequent chapters.

3.1 Challenges Posed by Large Proteins

Large macromolecules present two major challenges to NMR studies (Foster et al. 2007). First, the slower tumbling of large proteins in solution gives rise to faster transverse relaxation due to enhanced spin-spin interactions, which impairs the efficiency of coherence transfers and increases the linewidths of observed signals. Thus, fast transverse relaxation decreases both the sensitivity and resolution of NMR spectra, weakening and broadening peaks such that they become uninterpretable or unobservable for large proteins. Second, larger proteins contain more atoms and
therefore more observable signals that reside in a constrained range of frequencies, leading to spectral crowding. Spectral crowding can hinder the assignment of chemical shifts to individual atoms in large proteins, which is a crucial prerequisite for structural studies, but is particularly problematic in NOESY experiments, where through-space magnetization transfer creates numerous crosspeaks for each proton in the target protein, potentially leading to thousands of signals recorded in a single spectrum. Even in spectra which spread out these signals in three dimensions, there can still be substantial degeneracy. This signal overlap precludes the complete and unambiguous assignment of NOE crosspeaks, which provide the interproton distance constraints for structure calculations, and presents a serious deterrent to successful structure determination by NMR. Together, these challenges have restricted the application of classical NMR techniques to protein targets smaller than ~20 kDa.

3.2 Selective Labeling Strategies

Classical NMR studies generally utilize protein samples that are fully protonated and uniformly $^{13}$C- and $^{15}$N-labeled, permitting the observation of all protons and their attached heteronuclei. For large proteins, this high density of protons leads to daunting spectra complexity and severe signal loss due to proton-related relaxation pathways (Foster et al. 2007). A simple solution is to replace a subset of protons with deuterons
(Gardner and Kay 1998; Sattler and Fesik 1996). Achieving uniform deuteration is straightforward; bacterial expression is performed in D2O media with 2H/13C-glucose and 15N-NH4Cl as the sole carbon and nitrogen sources, respectively. Protons are then reincorporated at amide groups by transferring the protein sample to protonated solvent, although highly stable proteins may require denaturation and refolding in order to fully exchange amide groups buried in the protein core. This approach yields 2H/13C/15N-labeled samples with selectively-protonated amide groups, which are particularly useful for backbone assignment experiments that transfer magnetization from the amide proton through different sets of backbone carbon atoms. However, these samples are limited in their potential to provide structural information, as long-range contacts between amide protons are generally found only in β-sheets. An easy remedy is to selectively protonate samples at defined positions of specific residue types commonly found in protein cores by adding labeled metabolic precursors to bacterial cultures prior to induction. For example, the most popular strategy uses [3-2H]13C-α-ketoisovalerate and [3,3-2H2]13C-α-ketobutyrate to protonate the Ile δ1, Leu δ1/δ2, and Val γ1/γ2 (ILV) methyl positions in otherwise deuterated samples, which adds observable spins that are typically found in the hydrophobic core and can therefore provide useful long-range distance constraints (Gardner et al. 1997; Goto et al. 1999). In addition to these standard labeling schemes, a number of alternative methods have been developed including fractional deuterium labeling, amino acid-selective 13C- or 15N-
labeling, segmental labeling, and more recently, stereo-array isotope labeling, which employs in vitro protein synthesis and labeled versions of all 20 amino acids to yield samples with each carbon and nitrogen atom having at most one observable proton attached (Kainosho and Guntert 2009).

3.3 Spin State Selective Experiments

Another major methodological advance in protein NMR was the development of transverse relaxation-optimized spectroscopy (TROSY), a spin manipulation technique that reduces transverse relaxation and therefore sharpens linewidths (Pervushin 2000; Pervushin et al. 1997). In a conventional $^1$H/$^{15}$N-HSQC experiment, the scalar coupling between $^1$H and $^{15}$N nuclei during chemical shift evolution leads to the splitting of each peak in both dimensions, creating a pattern of four multiplet components. Each component arises from a unique magnetization transfer pathway in which the partner of the nucleus being frequency-encoded is in either the $\alpha$ or $\beta$ spin state, for example, $^1$H$\alpha$ or $^1$H$\beta$ during $^{15}$N evolution and $^{15}$N$\alpha$ or $^{15}$N$\beta$ during $^1$H evolution. Importantly, the four pathways have different relaxation properties due to constructive or destructive interference of the two main relaxation mechanisms – dipole-dipole interactions and chemical shift anisotropy – during each chemical shift evolution period. Normally, spin states are mixed during chemical shift evolution by decoupling, which eliminates peak
splitting and averages the relaxation properties of the four pathways. However, a key realization was that for large proteins it is beneficial to select the transfer pathway that is least susceptible to relaxation, yielding only the slowly relaxing, or TROSY, component that has a significantly sharpened linewidth (Pervushin et al. 1997). The TROSY spin state selective transfer was originally conceived for amide groups and was subsequently incorporated into a number of backbone assignment and NOESY experiments that utilize $^1\text{H}/^{15}\text{N}$-HSQC transfers (Konrat et al. 1999; Xia et al. 2000; Yang and Kay 1999b). The TROSY principle has also been extended to aromatic and methyl groups (Pervushin et al. 1998a; Tugarinov et al. 2003). The combination of deuteration and TROSY has substantially lifted the size limit of classical NMR techniques, paving the way for structural studies of proteins as large 82 kDa (Tugarinov et al. 2005a; Tugarinov et al. 2002).

### 3.4 Sparse Sampling Techniques for High-Dimensionality Data Collection

Even with the advent of selective labeling strategies and TROSY methods, spectral crowding remains a serious obstacle to structural studies of large proteins. Peak overlap is most pronounced in NOESY spectra, which can contain thousands of signals. The canonical NOE-based method of NMR structure determination requires that a high
percentage of NOE crosspeaks are identified and unambiguously assigned in order to provide a complete network of interproton distance constraints that will yield accurate and well-converged ensembles in subsequent structure calculations. Unfortunately, this is a challenging task using conventional 3-D NOESY experiments. In these experiments, the first two dimensions typically frequency encode the proton and attached heteronucleus of the NOE donor group prior to the NOE mixing period, while the third and directly-detected dimension frequency encodes either the proton or heteronucleus of each NOE acceptor group. Each NOE crosspeak must then be assigned to its corresponding acceptor group based on a single chemical shift. Even in sparsely-protonated samples of large proteins, there is significant degeneracy of potential acceptor groups when defined by a single chemical shift, particularly for certain group types, like methyls, which are usually poorly-dispersed. This leads to assignment ambiguity that severely complicates the analysis of NOESY spectra and may result in the incorrect assignment of NOE crosspeaks. A more pernicious problem is the overlap of NOE crosspeaks – either with other crosspeaks or with diagonal signals – that effectively eliminates important distance constraint information. In certain 3-D spectra, like methyl-methyl NOESY for example, overlap with diagonal signals alone can prevent the identification and assignment of a substantial percentage of NOE crosspeaks (see Section 6.4 and Tables 3 and 4).
A simple and effective solution to the problem of spectral crowding is to spread out signals further in higher-dimensionality experiments. In a 4-D NOESY experiment for example, the chemical shifts of both the proton and heteronucleus of each acceptor group are encoded, which decreases peak overlap and also reduces the number of assignment possibilities for each NOE crosspeak, provided that the experiment is collected at high enough resolution in each dimension. Unfortunately, collecting high-resolution, high-dimensionality spectra is made impractical by the measurement times required by conventional NMR sampling.

Conventional sampling of NMR experiments involves systemically increasing the chemical shift evolution period of each indirectly-encoded nucleus by a fixed amount of time, called the Nyquist interval, creating a regular grid of sampling points in the time domain (Figure 18A). The Nyquist interval in a given dimension is determined by the range of observed frequencies, or spectral width, and represents the lowest sampling rate required to unambiguously define the frequencies of the signals recorded in that dimension. Therefore, conventional sampling prescribes the minimum pattern needed to uniquely determine the position of peaks in multidimensional NMR spectra. However, a fundamental limitation of conventional sampling is that the number of sampling points grows exponentially as the dimensionality of experiments increases. For example, a 3-D experiment collected with 64-point resolution in each indirect dimension requires $64^2 \times (4,096)$ sampling points, whereas a 4-D experiment collected with
64-point resolution in each indirect dimension requires $64^3 (262,144)$ sampling points. At each sampling point, two measurements, or free induction decays (FIDs), are required for quadrature detection, the process of encoding phase information, in each of the indirect dimensions, and at least 4 scans are typically collected per FID. This gives a total of 16 scans per sampling point for a 3-D experiment, and 32 scans per point for a 4-D experiment. At ~2 seconds per scan, the conventionally-sampled 3-D experiment requires ~36 hours of acquisition time on the NMR spectrometer. In contrast, the 4-D experiment requires ~194 days. Setting aside sample stability concerns and other practical issues like instrument availability, the need to refill magnet coolant reservoirs limits uninterrupted acquisition to 12 days at most, an order of magnitude less than the time required for collecting high-resolution 4-D spectra. This discrepancy illustrates why 4-D datasets have traditionally been collected with very low resolution in at least one indirect dimension, thus limiting their ability to resolve the problems of peak overlap and assignment ambiguity in NOESY experiments.

**Figure 18. Sparse Sampling Patterns.** The distribution of points in the time domain is shown for different sampling schemes, including conventional (A), radial (B), concentric ring (C), and random (D). Axes denote indirect dimensions.
In order to speed up the acquisition of higher-dimensionality experiments, a number of alternative patterns for sampling the time domain of higher-dimensionality experiments have been devised. These patterns are collectively referred to as sparse sampling, as they use a smaller number of points than required by conventional Nyquist sampling, thereby providing a concomitant reduction in acquisition time. The first sparse pattern introduced in the NMR field was radial sampling (Figure 18B), which places points in the time domain along spokes that meet at the origin (Coggins et al. 2010). This approach is equivalent to sampling the time domain in polar coordinates instead of Cartesian coordinates (Coggins and Zhou 2006; Kazimierczuk et al. 2006; Marion 2006). Experimentally, radial sampling is accomplished by jointly, rather than independently, incrementing two or more of the indirect chemical shift evolution periods. This general concept was first introduced back in 1981 in accordion spectroscopy, although its implementation involved sampling an experimental dimension measuring chemical exchange, manifest as lineshapes of the resulting signals, as opposed to chemical shift (Bodenhausen and Ernst 1981). By the projection-slice theorem, a conventional Fourier transform (FT) of each spoke in a radially-sampled dataset yields a projection, or sum of signals along the perpendicular direction, at the same angle in the frequency domain (Nagayama et al. 1978). This principle has been exploited by a great variety of methods to extract information from projected peaks, including (1) geometric formulas to back-calculate peak positions in the full high-
dimensionality spectrum (Hiller et al. 2005) and (2) analysis of peak splittings on projections collected with quadrature detection in specific subsets of the jointly-evolved dimensions in order to deduce and assign chemical shifts (Kim and Szyperski 2003; Szyperski et al. 1993). Alternatively, a number of algorithms have been developed that reconstruct the full high-dimensionality spectrum using multiple radial spokes (Kupce and Freeman 2003, 2004; Venters et al. 2005). An important realization was that the standard FT modified by an appropriate weighting function could be used to generate linear reproductions of full high-dimensionality spectra (Coggins and Zhou 2006; Kazimierczuk et al. 2007; Marion 2006), and that the characteristic patterns of noise in these spectra represent the reinforcement of aliasing artifacts introduced by the regularity of radial sampling (see Section 3.5).

Following the introduction of radial sampling, a number of other sampling patterns were designed for fast data collection and spectral processing with the FT and alternative algorithms (Kazimierczuk et al. 2011). These include concentric ring or shell sampling (Figure 18C), random sampling (Figure 18D), and spiral sampling (Barna and Laue 1987; Barna et al. 1987; Coggins and Zhou 2008; Kazimierczuk et al. 2006). The distribution of sampling points on these patterns is usually weighted using an exponential or cosine function, biasing the measurement of low-resolution information found close to the origin of the time domain, in order to improve sensitivity and to impart a lineshape in the frequency domain. While the FT can be used to compute
spectra from datasets sampled between points defined by the conventional Nyquist grid, or off-grid sampling, use of the computationally-efficient fast Fourier transform (FFT) requires that points lie on the Nyquist grid. For this reason, points are typically distributed along the desired sparse sampling pattern and then moved to their nearest Nyquist grid positions.

Radial sampling and its sparse sampling successors allow for a significant reduction in the number of experimental measurements compared to conventional sampling, and therefore can dramatically decrease the acquisition time of high-dimensionality spectra. The exact amount of time savings depends on the number of points sampled, which is arbitrary with two main caveats. First, increased undersampling leads to more aliasing artifacts (see below), and second, the number of points sampled must provide sufficient sensitivity for the experiment being acquired and the system under study. Even with these restrictions, the time savings can be well over an order of magnitude. For example, Chapter 6 of this dissertation describes the application of sparse sampling to high-resolution 4-D NOESY of selectively-labeled 23 kDa and 29 kDa test proteins, using only 1.2% of the points required by conventional sampling at equivalent resolution. The spectra are collected in only 4 days and have sufficient sensitivity to permit the identification of near complete sets of interproton distance constraints under 5 Å.
3.5. Aliasing Artifacts and Removal Algorithms

Sparse sampling provides a powerful tool for collecting high-resolution, high-dimensionality spectra that can mitigate the problems of peak overlap and assignment ambiguity for large proteins; however, this benefit comes at the cost of introducing aliasing artifacts (Kazimierczuk et al. 2011). As described above, conventional sampling utilizes the minimum number of measurements required to precisely define the frequencies of observed signals. Sparse sampling purposely fails to meet this standard and thus creates uncertainty in the frequency measurement of each signal, which is manifest as aliasing artifacts in the resulting spectrum. The form of these artifacts depends on the sampling pattern. In general, regular patterns of sampling points lead to regular and reinforcing patterns of artifacts for each peak. For example, in the case of grid sampling with points spaced further than the Nyquist interval, the aliasing artifacts of a peak appear as duplicate peaks at regular intervals in the spectrum. This duplication reflects the fact that, due to the decreased sampling rate, multiple signals now fit the frequency measurements equally well. In the case of radial sampling, artifacts form low level ridges centered on each peak and extending away in directions perpendicular to the spokes of sampling points in the time domain. This artifact pattern reflects the inability of each radial spoke to measure signal modulation perpendicular to its direction in the time domain (Coggins et al. 2010). These regular patterns of artifacts
can complicate spectral analysis, and so newer sampling schemes generally distribute points in the time domain with some degree of randomization; for example, in the concentric ring pattern (Figure 18C), the relative orientation of rings is randomized in order to prevent sampling points from aligning along spokes (Coggins and Zhou 2007). This inhibits the reinforcement of aliasing artifacts and instead spreads them evenly throughout the spectrum where they take the appearance of thermal noise (Figure 19A-B). While the pattern of aliasing artifacts is determined by the sampling scheme, their intensity depends on the extent of undersampling, with lower sampling density giving rise to stronger artifacts. By the same principle, sparse sampling patterns that are populated with the number of points required by conventional sampling, or more points for certain patterns, yield artifact-free spectra (Coggins and Zhou 2007).

Another important characteristic of aliasing artifacts is that the level of artifacts generated by each signal is proportional to the signal’s intensity. This is particularly problematic for spectra that have a high dynamic range of signals, as the aliasing artifacts produced by strong signals can obscure weaker signals, even when the artifacts are distributed as apparent noise by the sampling pattern (Figure 19A-B). High dynamic range is best exemplified by NOESY experiments, as they often have at least 100-fold intensity differences between strong diagonal peaks, which result from magnetization not transferred during the NOE mixing period, and NOE crosspeaks, which provide the interproton distance constraint information necessary for successful structure
**Figure 19. Aliasing Artifacts and Removal Algorithms.** 1-D trace of a simulated spectrum containing two peaks with high dynamic range collected with either conventional sampling (A) or randomized sparse sampling (B). The high level of apparent noise in (B) that prevents observation of the weaker peak is the result of aliasing artifacts generated by the strong peak. The general approach of artifact removal algorithms is shown in (C-E). First, the strong peak is identified and its artifact pattern is calculated based on its intensity, position, and the sampling scheme (C). Next, the peak and its artifacts are subtracted from the experimental spectrum shown in (B) to reveal smaller peaks (D). This process is continued iteratively, and finally, uncorrupted peaks are added back to yield a spectrum with reduced artifacts (E).

determination (see Chapter 6). The potential loss of crosspeaks presents a serious drawback to the application of sparse sampling techniques to NOESY experiments, where high-resolution, high-dimensionality data are needed most. In fact, sparsely-sampled high-resolution 4-D NOESY has been employed in only a small number of structure determination studies, and has always been used in conjunction with nonlinear algorithms for spectral processing instead of the conventional FT (Hiller et al. 2009; Tugarinov et al. 2005b).
There are two potential solutions to the problem posed by the high dynamic range in NOESY experiments with diagonal peaks. The first is to remove the diagonal signals experimentally. This can be accomplished by designing experiments that suppress the diagonal signals as demonstrated in Chapter 5 of this dissertation, which details a novel pulse sequence that suppresses diagonal signals in 4-D amide-amide NOESY spectra using TROSY-based spin manipulation. However, this strategy is not applicable to all NOESY experiments. An alternative approach is to collect two spectra for each NOESY experiment, one with both crosspeaks and diagonal peaks and a second with only diagonal peaks, and then to subtract the two to produce a diagonal-free spectrum (Wu et al. 2004). Unfortunately, this requires an undesirable doubling of the time required for data collection.

The second solution is to employ algorithms that remove aliasing artifacts as part of spectral processing following data collection. Artifact removal algorithms exploit the fact that aliasing artifacts can be robustly and accurately predicted with very limited a priori information. In fact, the only parameters needed to calculate the artifacts generated by a given peak are the peak’s intensity and spectral position and the sampling pattern used for data collection. These algorithms therefore operate using variations on a common mode of action illustrated in Figure 19. Following FT computation of the sparsely-sampled spectrum, the most intense peak is identified, and its artifact pattern is calculated based on the parameters described above. The peak and
its artifacts are then subtracted from the spectrum, often revealing weaker peaks that were masked by the artifact noise. This process can be repeated iteratively until some stopping criteria, usually based on the residual noise level or the ability to confidently identify more peaks for processing, is reached. Uncorrupted peaks are then added back to yield a spectrum with reduced artifact levels (Figure 19).

The classical implementation of this process is an algorithm called CLEAN that was developed in the radioastronomy community (Högbom 1974). The detection of celestial objects at high resolution with radio waves requires mixing signals from a collection of small telescopes, which alone have low resolution but are relatively inexpensive to construct, in order to produce images with resolution equivalent to a telescope the size of the entire collection. This technique is called aperture synthesis, or synthesis imaging, and utilizes regular arrays of telescopes on the ground, along with the movement of the array in space due to the rotation of the Earth, to make measurements at regular grid points. There are many phenomena that can create gaps in these measurements, including equipment malfunctions and obstruction of radio sources by the Moon, that lead to artifacts in the final image, making it difficult or impossible to interpret (Högbom 1974). Thus, CLEAN was devised to remove these artifacts and reconstruct celestial images as if they had been measured with fully-sampled grids. The analogy to sparsely-sampled NMR spectra, where points in the time domain are purposely left unmeasured to reduce acquisition time, is straightforward.
However, the earliest applications of CLEAN in the field of NMR had alternative aims, including replacing unwanted dispersive lineshapes and eliminating signal truncation artifacts (Keeler 1984; Shaka et al. 1984).

In the past few years, the CLEAN algorithm and related strategies have become a widely recognized way to effectively remove aliasing artifacts from sparsely-sampled NMR spectra. The first application of the CLEAN principle was a method tailored specifically for radially-sampled experiments that attempted to eliminate the characteristic ridge patterns of aliasing artifacts found in reconstructed spectra (Kupce and Freeman 2005). This method first identified the tallest peak in a spectrum and subtracted it from each projection, and then reconstructed a new spectrum using the modified projections. The process was continued in an iterative fashion by finding the next highest peak, and the final spectrum was generated by reintroducing the subtracted peaks into the already reconstructed spectrum, thereby eliminating the ridge artifacts associated with their reconstruction. Later, two more general and classical versions of CLEAN designed for use with arbitrary sampling schemes were proposed by Kazimierczuk and co-workers and by our laboratory (Coggins and Zhou 2008; Kazimierczuk et al. 2007). The original implementation of the method of Kazimierczuk et al. required user-provided peak parameters as input to replicate a full set of artificial signals with assumed lineshapes (Kazimierczuk et al. 2007). The artifacts of all signals were then calculated and subtracted in a single step from the experimental spectrum. In
the most recent version, peaks are identified automatically and modeled with exponentially-decaying functions in the time domain (Stanek and Kozinski 2010). This contrasts with our implementation of CLEAN, described in the following paragraph, which treats each peak as a set of nondecaying components, and therefore does not require modeling each signal’s lineshape with an analytical equation. A more rudimentary strategy based on the CLEAN principle has also been developed to reduce aliasing artifacts during reconstruction of spectra with forward maximum entropy (Hyberts et al. 2009).

Our implementation of CLEAN is an iterative process modeled closely off of the standard version used in radioastronomy and other fields that performs many cycles of the following steps: (1) identifying the peak with highest intensity, (2) producing a replica, or component, of the peak, (3) calculating the artifact pattern of the component based on the sampling scheme, and (4) subtracting a fraction of the component and its artifacts from the spectrum (Coggins and Zhou 2008). Importantly, each peak is treated as a set of discrete points in the frequency domain, such that components represent nondecaying signals in the time domain. This avoids the need to model full peak shapes – a process that requires assumptions about lineshapes and is likely to fail in cases of overlapped or distorted peaks – and is a key advantage of our implementation. We first applied our CLEAN algorithm to high-resolution 4-D TOCSY (Coggins and Zhou 2008), and later to diagonal-suppressed 4-D amide-amide NOESY as described in Chapter 5 of
this dissertation. In both cases, CLEAN was highly successful at removing aliasing artifacts, with reductions in the apparent noise, which consists of both thermal noise and the aliasing artifacts, of up to 4.5-fold. Furthermore, we showed that CLEAN faithfully reproduces peak intensities and lineshapes (see Chapter 5), a crucial property for NOESY spectra, where any systematic distortion of signals can degrade the quality of interproton distance constraints and lead to bias in subsequent structure calculations.

Despite the promising capability of our CLEAN algorithm, we found that its removal of aliasing artifacts was not sufficiently thorough for use in the most challenging cases, like NOESY spectra with diagonal signals. This limitation in achievable dynamic range results from the fractional subtraction of peaks by the conventional algorithm, which allows each peak to be reduced down only to the level of apparent noise, at which point the residual signal can no longer be confidently distinguished. Thus, CLEAN leaves behind “stumps” for each peak at the level of apparent noise, along with the artifacts generated by the residual intensity. This problem is circumvented in versions of CLEAN that use fitted lineshapes, as they subtract the entire peak from the spectrum, including the peak’s stump and its artifacts (Stanek and Kozminski 2010). To address this limitation, we developed a new algorithm called SCRUB, for Scrupulous CLEANing to Remove Unwanted Baseline Artifacts. SCRUB utilizes the basic CLEAN premise for removing artifacts, but is designed to process peaks in batches, beginning with those that are strongest and well above the
noise level, and applies special criteria to identify and include points adjacent to each peak’s center, so that the full peak shape is subtracted rather than only the strongest point. A set of tests is run during each cycle to ensure that the points selected for processing represent true signals and not noise or artifacts. Finally, after validating the selected batch of peaks, they can be confidently driven down to the baseline with subtractive artifact removal. The cycle is then repeated iteratively on weaker sets of peaks. Like its CLEAN predecessor, SCRUB uses nondecaying signals to decompose each peak, thereby avoiding the pitfalls of fitting signal lineshapes, and retains the linearity of peak information. A comparison of SCRUB and CLEAN using simulated sparsely-sampled data demonstrates the substantial advantage of the new algorithm, with SCRUB reducing aliasing artifacts 250-fold further than CLEAN. The practical benefit of the new method is showcased in Chapter 6 of this dissertation, where SCRUB is applied to high-resolution 4-D time-shared NOESY spectra with diagonal peaks collected with only 1.2% of the points required by conventional sampling and succeeds in suppressing artifacts as far as possible: to the level of thermal noise. The reduced aliasing noise permits the identification of weak NOE crosspeaks, facilitating the assignment of near complete sets of interproton distance constraints under 5 Å for selectively-protonated samples of 23 kDa and 29 kDa test proteins.
4. The ‘Just-In-Time’ HN(CA)CO Experiment for Backbone Assignment of Large Proteins with High Sensitivity

As described in the previous chapter, a number of advances in NMR spectroscopy have vastly expanded our ability to study large proteins, with global fold determination being reported for a single chain protein of 82 kDa (Tugarinov et al. 2005a). The first step in nearly all NMR studies is the assignment of backbone amide resonances. This two step process begins by using backbone assignment experiments, also called triple-resonance experiments, to correlate each amide peak to the backbone carbon chemical shifts of its residue and the preceding one. Amides can then be connected sequentially through these backbone carbon chemical shifts, and finally, matched to the protein’s sequence by exploiting the distinct and characteristic ranges of backbone carbon chemical shifts for different residue types. Due to signal degeneracy in the highly complex spectra of large proteins, the unambiguous assignment of backbone resonances can only be achieved through the use of all three pairs (C_{\alpha}, C_{\beta} and CO) of connectivities (Coggins and Zhou 2003). Among the suite of commonly used backbone triple-resonance experiments, the HN(CA)CO experiment is the least sensitive due to two inherent sensitivity limitations. First, during the $^{13}\text{C}_{\alpha}$-$^{13}\text{CO}$ transfer where $^{13}\text{C}_{\alpha}$ magnetization is transverse, passive $^{13}\text{C}_{\alpha}$-$^{13}\text{C}_{\beta}$ coupling degrades the observable signal. Second, when $^{13}\text{CO}$ magnetization is transverse, fast relaxation of carbonyl coherence
caused by the large carbonyl chemical shift anisotropy (CSA) leads to substantial signal loss that becomes more severe with increasing protein size and at higher magnetic field. This chapter describes a novel ‘just-in-time’ (JIT) HN(CA)CO experiment that combines the advantages of its HMQC- and HSQC-based predecessors by removing \( ^{13}\text{C}_\alpha^{13}\text{C}_\beta \) coupling uniformly and minimizing \( ^{13}\text{CO} \) relaxation losses, thus offering high sensitivity for all resonances regardless of residue type.

### 4.1 Previous Implementations of the HN(CA)CO Experiment

Different approaches have been used to improve the sensitivity of the HN(CA)CO experiment. A constant time heteronuclear multiple quantum correlation (CT-HMQC) \( ^{13}\text{C}_\alpha^{13}\text{CO} \) transfer places \( ^{13}\text{C}_\alpha \) magnetization on the transverse plane for 28 ms to refocus the undesired \( ^{13}\text{C}_\alpha^{13}\text{C}_\beta \) coupling (Engelke and Rüterjans 1995). However, this method also includes a constant time carbonyl evolution period of 11 ms, during which \( ^{13}\text{CO} \) coherence is susceptible to degradation via rapid transverse relaxation. The alternative to the HMQC-based method employs a heteronuclear single quantum correlation (HSQC) \( ^{13}\text{C}_\alpha^{13}\text{CO} \) transfer sequence (Clubb et al. 1992). While this eliminates the need for a constant time \( ^{13}\text{CO} \) evolution period, \( ^{13}\text{C}_\alpha^{13}\text{C}_\beta \) coupling is not refocused and leads to over 60% signal loss. This drawback can be partially resolved by using selective \( ^{13}\text{C}_\alpha \) pulses during the \( ^{13}\text{C}_\alpha^{13}\text{CO} \) INEPT transfers for simultaneous \( ^{13}\text{C}_\alpha \)

However, due to the chemical shift degeneracy of $^{13}\text{C}_\alpha$ and $^{13}\text{C}_\beta$ nuclei, it has not been possible to eliminate the passive $^{13}\text{C}_\alpha$-$^{13}\text{C}_\beta$ couplings uniformly, leading to severe signal attenuation of peaks from certain residue types, particularly Ser and Leu (Matsuo et al. 1996; Tugarinov et al. 2002; Yang and Kay 1999b). Even with pulse sequence modifications to optimize the collection of Ser signals, a large percentage of expected Ser peaks are still lost (Tugarinov et al. 2002). Additionally, residues with $^{13}\text{C}_\alpha$ chemical shifts outside the refocusing window of the selective $^{13}\text{C}_\alpha$ pulse, such as Gly, Val and Pro, also experience significant signal attenuation (Matsuo et al. 1996; Yang and Kay 1999b).

4.2 The ‘Just-In-Time’ Approach

In order to overcome the sensitivity limitations of previous versions of the HN(CA)CO experiment, we designed a novel ‘just-in-time’ (JIT) HN(CA)CO experiment that removes $^{13}\text{C}_\alpha$-$^{13}\text{C}_\beta$ coupling uniformly and minimizes $^{13}\text{C}_\beta$ relaxation losses, thus offering high sensitivity for resonances of all residue types. The pulse sequence of the JIT TROSY-HN(CA)CO experiment is shown in Figure 20. It begins with coherence transfer from $^1\text{HN}$ to $^{15}\text{N}$ and $^{15}\text{N}$ to $^{13}\text{C}_\alpha$. As in the $^{13}\text{C}_\alpha$-$^{13}\text{CO}$ transfer element of the CT-
Figure 20. Pulse Sequence for the JIT TROSY-HN(CA)CO Experiment. Narrow and wide bars represent 90° and 180° pulses, respectively. All pulses are applied along the x-axis unless noted otherwise. 90° selective water pulses are indicated by shaped bars. The 13C carrier frequency is initially set to 57 ppm and shifted to 175 ppm between points a and b. Shaped aliphatic pulses represent 375 μs off-resonance REBURP pulses centered at 44 ppm. The remaining 13Cα pulses are selective rectangular pulses with maximum excitation at 57 ppm and null excitation at 175 ppm. Selective sinc-shaped 13CO pulses are applied at 175 ppm with null excitation at 57 ppm. An empirically optimized phase shift of θ = 28° is applied to the last 90° 13Cα pulse to correct for Bloch-Siegert effects. The positioning of the two asterisked 13CO 180° pulses is synchronized with the carbonyl evolution. The delays are τ1 = 2.4 ms, τ2 = 12.5 ms, τ3 = 1/4JCαCO = 4.5 ms, τ4 = 1/4JCαCβ - 1/4JCαCO = 2.6 ms, τ5 = 1/4JCαCβ = 7.1 ms, and τ6 = 2.8 ms. The phase cycling is φ1 = [2x, 2(-x)], φ2 = [x, -x], φ3 = [4y, 4(-y)], φ4 = [8x, 8(-x)], φ5 = [x], and φrec = [x, -x, -x, x, -x, x, -x]. States-TPPI in F1 is achieved by incrementing φ2 and φrec. The sensitivity enhanced gradient TROSY scheme requires the inversion of φ5 and G2 for hypercomplex data collection in F2. WALTZ-16 is used for 2H decoupling. Gradient durations and field strengths are G1 = (2 ms, 20.42 G/cm), G2 = (0.2 ms, 20.36 G/cm), G3 = (0.5 ms, 11.64 G/cm), G4 = (0.9 ms, 21.03 G/cm), G5 = (0.5 ms, 9.60 G/cm), G6 = (0.9 ms, 18.99 G/cm), G7 = (0.9 ms, 19.81 G/cm), G8 = (0.7 ms, 12.05 G/cm), and G9 = (0.7 ms, 12.86 G/cm).

HMQC experiment, 13Cα magnetization is kept transverse for a constant period of 1/JCαCβ to refocus 13Cα-13Cβ coupling. However, the 13CO chemical shift is recorded in real time by synchronizing the movement of the two 180° 13CO pulses during the 13Cα-13CO transfer (marked by asterisks in Figure 20) with respect to the 90° 13CO flip pulses flanking the 13CO evolution period. This keeps the effective 13Cα-13CO coupling time at
$1/2|\text{J}_{\text{C}_{\alpha}\text{CO}}|$ and allows the real time build up of carbonyl coherence before flipping $^{13}\text{CO}$ magnetization into the transverse plane ‘just in time’ for frequency labeling. The JIT $^{13}\text{C}_{\alpha}$-$^{13}\text{CO}$ transfer scheme limits the maximum $^{13}\text{CO}$ evolution time to ~11 ms. Although this restricts the digital resolution of the $^{13}\text{CO}$ dimension, the fast $^{13}\text{CO}$ transverse relaxation in large proteins makes longer acquisition time of limited benefit. $^{15}\text{N}$ and $^{1}\text{HN}$ shifts are collected using a sensitivity enhanced TROSY element with suppression of anti-TROSY signals (Nietlispach 2005). A similar $^{13}\text{C}_{\alpha}$-$^{13}\text{CO}$ transfer scheme has been utilized previously as part of a (HCA)CONH experiment (Folmer and Otting 2000). However, the rapid transverse relaxation of $^{13}\text{C}_{\alpha}$ nuclei in protonated samples has restricted the use of the (HCA)CONH experiment to small- to medium-sized proteins. Because the $^{13}\text{CO}$ CSA relaxation is more severe in large proteins, the signal enhancement of the JIT $^{13}\text{C}_{\alpha}$-$^{13}\text{CO}$ transfer is most pronounced for large deuterated proteins.

4.3 Sensitivity Analysis and Comparison with Previous Implementations

The JIT $^{13}\text{C}_{\alpha}$-$^{13}\text{CO}$ transfer provides significant sensitivity enhancement over the CT-HMQC-HN(CA)CO experiment. At the maximum $^{13}\text{CO}$ resolution attainable in the JIT experiment, the average time for $^{13}\text{CO}$ magnetization in the transverse plane is
reduced by 58% compared to the HMQC-based sequence, therefore avoiding substantial signal loss from fast carbonyl relaxation. This sensitivity gain increases for large proteins at high magnetic fields, where the rate of transverse $^{13}$CO relaxation is particularly detrimental. Compared to the HSQC-based experiment with selective $^{13}$C$_\alpha$ refocusing pulses, the JIT experiment is slightly less sensitive for residues experiencing complete $^{13}$C$_\alpha$ inversion and proper $^{13}$C$_\alpha$-$^{13}$C$_\beta$ decoupling, due to an additional delay of 11 ms experienced by transverse $^{13}$C$_\alpha$ magnetization. The signal loss should be minimal because of the relatively slow relaxation rate of transverse $^{13}$C$_\alpha$ magnetization in deuterated proteins (Farmer and Venters 1999). However, a significant signal gain is achieved for the remaining residues as the JIT experiment does not rely on selective $^{13}$C$_\alpha$ pulses to refocus $^{13}$C$_\alpha$ magnetization or to remove $^{13}$C$_\alpha$-$^{13}$C$_\beta$ coupling.

To demonstrate the advantages of our JIT TROSY-HN(CA)CO experiment, we collected 3-D spectra of a 0.9 mM $^2$H/$^{13}$C/$^{15}$N-labeled sample of the 29 kDa protein human carbonic anhydrase II (HCAII) on an 800 MHz Varian spectrometer at 15°C. The correlation time of HCAII was estimated to be 23 ns, corresponding approximately to a 39 kDa protein at room temperature. Two control spectra employing the CT-HMQC-HN(CA)CO and the HSQC-HN(CA)CO with selective $^{13}$C$_\alpha$ refocusing were also collected for sensitivity comparisons. For each experiment, 26, 64, and 512 complex points were collected in the $^{13}$CO, $^{15}$N, and $^1$HN dimensions, respectively, with a total acquisition time of 32 hours. 2-D $^1$HN-$^{13}$CO projections from each experiment are shown.
**Figure 21. Comparison of TROSY-HN(CA)CO Spectra.** 2-D $^1$HN-$^{13}$CO projections are shown for the JIT HN(CA)CO experiment and CT-HMQC-based (Engelke and Rüterjans 1995) and HSQC-based (Matsuo et al. 1996; Yang and Kay 1999b) controls. Due to the lack of $^{13}$C$_\alpha$-$^{13}$C$_\beta$ coupling, glycine peaks appear with negative intensity (in red) in the JIT and CT-HMQC-based experiments.

in Figure 21. Two sets of sensitivity comparisons were calculated using peak heights and 3-D integrated peak volumes (reported in parenthesis). On average, the JIT experiment provides a 42% (59%) sensitivity gain over the CT-HMQC-based control. In the HSQC-based experiment, the nonuniform excitation profile of the selective $^{13}$C$_\alpha$ inversion pulses leads to a significant variation of sensitivity that depends on the $^{13}$C$_\alpha$ and $^{13}$C$_\beta$ chemical shifts of each residue (Figure 22). Effective inversion of $^{13}$C$_\alpha$ magnetization occurs from 45 to 65 ppm; residues with $^{13}$C$_\alpha$ shifts inside and $^{13}$C$_\beta$ shifts outside this window experience effective $^{13}$C$_\alpha$ refocusing and $^{13}$C$_\alpha$-$^{13}$C$_\beta$ decoupling during the $^{13}$C$_\alpha$-$^{13}$CO INEPT transfers. For these residues, the JIT experiment averages 16% (19%) less sensitivity than the HSQC-based sequence. For residues with incomplete
Figure 22. Slices and Sensitivity Profiles of TROSY-HN(CA)CO Spectra. (A) Slices from the 3-D JIT, CT-HMQC-based (Engelke and Rüterjans 1995), and HSQC-based (Matsuo et al. 1996; Yang and Kay 1999b) HN(CA)CO spectra. Asterisks indicate peaks from residues not properly refocused by the selective $^{13}$C pulses in the HSQC-based control: intraresidue correlations from Ser 48 and Gly 232 and an interresidue correlation from Ser 56/Leu 57. 1-D traces are taken at the centers of peak intensity in the $^1$HN dimension. The negative intensity of peaks from glycine residues in the JIT and CT-HMQC-based experiments is caused by the absence of $^{13}$C-$^{13}$C coupling. (B-C) Relative sensitivities of the JIT (black), CT-HMQC-based (gray), and HSQC-based (white) experiments calculated with peak heights versus $^{13}$C$_\alpha$ and $^{13}$C$_\beta$ chemical shifts. The sensitivity profiles of the HSQC-based experiment are related to the excitation profile of its selective $^{13}$C$_\alpha$ inversion pulses. Signal loss is observed for residues at the edges of the $^{13}$C$_\alpha$ distribution (regions $\alpha$ and $\beta$) due to the incomplete refocusing of $^{13}$C$_\alpha$ magnetization and for residues with $^{13}$C$_\beta$ shifts between 45 to 65 ppm (region $\gamma$) due to improper $^{13}$C$_\alpha$-$^{13}$C$_\beta$ decoupling.
$^{13}\text{C}_{\alpha}$ refocusing (Figure 22B, regions $\alpha$ and $\beta$) or improper $^{13}\text{C}_{\beta}$ decoupling (Figure 22C, region $\gamma$), the JIT spectrum shows, on average, a 106% (104%) sensitivity gain for intra and interresidue peaks over the HSQC-based experiment. For HCAII, the 70 peaks that fall into this group include mainly Ser and Gly residues and a smaller number of Leu, Val, and Pro residues. While the JIT spectrum shows a consistent sensitivity advantage for these residues, the degree of sensitivity enhancement varies substantially. For the resonances shown in Figure 22A (S48, S56, and G232), the selective $^{13}\text{C}_{\alpha}$ inversion pulses in the HSQC-based experiment lead to a nearly complete loss of signal intensity.

4.4 Summary

In conclusion, we have introduced the JIT TROSY-HN(CA)CO pulse sequence and demonstrated its high and uniform sensitivity compared to HMQC- and HSQC-based controls. While previous HN(CA)CO experiments have not succeeded in both uniformly refocusing detrimental $^{13}\text{C}_{\alpha}$-$^{13}\text{C}_{\beta}$ coupling and minimizing $^{13}\text{CO}$ relaxation losses, the JIT $^{13}\text{C}_{\alpha}$-$^{13}\text{CO}$ transfer scheme overcomes both of these limitations. In particular, the real time $^{13}\text{CO}$ chemical shift evolution should make the JIT experiment especially beneficial for NMR studies of large proteins at high magnetic fields, where the reduction of adverse $^{13}\text{CO}$ CSA relaxation is critical. Given the importance of carbonyl data in obtaining unambiguous sequential assignment for backbone resonances, we
believe that our high sensitivity JIT HN(CA)CO experiment will provide a powerful new tool for extending NMR methodology to large proteins and protein complexes.
5. Fast Acquisition of High-Resolution 4-D Amide-Amide NOESY with Diagonal Suppression

Amide-amide NOESY provides valuable distance constraints for calculating global folds of large proteins, which are especially important for proteins with high \( \beta \)-sheet content. In extreme cases, such as \( \beta \)-barrel proteins, amide-amide constraints alone can be sufficient to define the overall topology. Amide-amide NOESY is also useful for defining secondary structure elements and for confirming and extending backbone amide assignments. Protein deuteration and the development of TROSY have greatly improved the sensitivity of the amide-amide NOESY experiment by minimizing signal loss from transverse relaxation (Farmer and Venters 1999; LeMaster and Richards 1988; Pervushin 2000; Pervushin et al. 1997), a major obstacle for NMR studies of large proteins (see Sections 3.2 and 3.3).

The spectral complexity of NOESY experiments with large proteins presents another serious challenge for structural studies. In 3-D NOESY experiments, the large number of proton signals can create severe resonance overlap, frustrating assignment efforts and reducing distance constraint accuracy. While this problem can be largely eliminated by recording high-resolution 4-D NOESY spectra, the measurement times required for conventional Nyquist sampling of individual dimensions have restricted the acquisition of 4-D NOESY data to very low resolution. However, the increased
sensitivity of NMR experiments benefiting from improved probe technology, protein deuteration, and TROSY methods has made it possible to apply sparse sampling methods for more efficient collection of high-resolution 4-D spectra with reduced peak overlap and assignment ambiguity. Indeed, significant efforts have been devoted to the development and optimization of sparse sampling patterns in the time domain as well as methods for reconstructing spectral information from sparsely-sampled datasets (Coggins et al. 2010; Kazimierczuk et al. 2010). Recently, our lab developed a sparse sampling scheme – randomized concentric shell sampling (RCSS) – for collecting high-resolution 4-D NMR spectra, which gives greater sensitivity and fewer artifacts than the radial sampling method (Coggins and Zhou 2008). Using RCSS data collection, we recorded a high-resolution 4-D HCCH-TOCSY spectrum for protein G’s B1 domain in ~21 hours, or 1.2 % of the time required for conventional sampling of the same experiment at equivalent resolution.

The inherent drawback of using sparse sampling schemes is the appearance of aliasing artifacts resulting from the violation of the Nyquist theorem (Coggins and Zhou 2006; Sarty 2003). Importantly, the level of aliasing artifacts for each signal is proportional to the signal amplitude. As discussed in Section 3.5, this property causes a major problem in NOESY spectra, where the dynamic range of signal amplitudes is very large, as aliasing artifacts from strong diagonal peaks may overshadow weak NOE crosspeaks – even with optimized arrangements of sampling points that reduce the
reinforcement of these artifacts. A simple solution to this problem is to suppress the diagonal signals. Among the methods of diagonal suppression (Harbison et al. 1985; Meissner and Sørensen 1999; Pervushin et al. 1999; Wu et al. 2004), the approach based on TROSY spin state selection is the most attractive for large proteins, as it has the added benefit of reducing transverse relaxation losses. Such an approach has been demonstrated in the 3-D NH-NH NOESY experiment (Meissner and Sørensen 2000; Xia et al. 2000; Zhu et al. 1999), but these implementations cannot be extended directly to 4-D experiments, as they lack a polarization transfer element that permits nitrogen chemical shift evolution after the NOE mixing period. This chapter describes a diagonal-suppressed $^1$H/$^{15}$N-TROSY-NOESY-$^1$H/$^{15}$N-TROSY (ds-TNT) pulse sequence with a novel single transition-to-single transition polarization transfer element that allows transverse relaxation optimized chemical shift evolution in all four dimensions. We use the ds-TNT pulse sequence with RCSS data collection and FFT-CLEAN processing to record a high-resolution 4-D spectrum that features well-suppressed diagonal peaks, low artifact levels, and reduced signal overlap and retains the linearity of NOE crosspeak information over its entire dynamic range.
5.1 The 4-D NH-NH ds-TNT NOESY Experiment

The pulse sequence for the 4-D ds-TNT experiment (Figure 23) can be separated into four main sections: the first TROSY selection with frequency labeling in $t_1$ (N1) and $t_2$ (H1), the NOE mixing period, a ‘reverse’ single transition-to-single transition polarization transfer (reverse ST2-PT), and the second TROSY selection with frequency labeling in $t_3$ (N2) followed by detection in $t_4$ (H2). The experiment begins with a sensitivity enhanced gradient selection of TROSY magnetization with suppression of the anti-TROSY coherence pathway (Nietlispach 2005). Inversion of phases $\phi_2$ and $\phi_{\text{rec}}$ and gradient G2 controls coherence pathway selection while incrementing phase $\phi_3$ (and phase $\phi_4$ for water suppression) controls cosine-sine selection for the first two dimensions. Ignoring the $t_3$ (N2) and $t_4$ (H2) dimensions, recording the sensitivity enhanced signals in the first two dimensions yields the following pattern of frequency labeling:

\[
\begin{align*}
S_{11} &= \cos(\omega_{N1}t_1 + \omega_{H1}t_2) \\
S_{12} &= \sin(\omega_{N1}t_1 + \omega_{H1}t_2) \\
S_{21} &= \cos(-\omega_{N1}t_1 + \omega_{H1}t_2) \\
S_{22} &= \sin(-\omega_{N1}t_1 + \omega_{H1}t_2) \\
\end{align*}
\]

which can be resolved into quadrature components by standard procedures (Xia et al. 2000).
After the first spin state selective transfer, the TROSY coherence of spin 1 (which is $N_\alpha$ polarized) enters the NOE mixing period and is transferred to magnetization in spin 2, creating both $N_\alpha$ and $N_\beta$ polarization:

$$N^1_\alpha H^1_z \xrightarrow{\text{NOE mixing}} N^1_\alpha H^1_z \ (\text{diagonal}) + N^1_\alpha H^2_z \ (\text{crosspeak})$$

$$= N^1_\alpha H^1_z \ (\text{diagonal}) + N^1_\alpha N^2_\alpha H^2_z \ (\text{crosspeak}) + N^1_\alpha N^2_\beta H^2_z \ (\text{crosspeak}) \quad (5.2)$$

The $N_\beta$ polarization of the spin 2 coherence created during the mixing period ($N^1_\alpha N^2_\beta H^2_z$) provides the means for selectively suppressing the diagonal coherence, which is still in the $N_\alpha$ state. Before eliminating the diagonal signals, however, the nitrogen chemical shift of spin 2 must be recorded to permit four dimensional data collection. This requires a specially-designed reverse ST2-PT that simultaneously achieves four coherence transfers:

$$N^1_\alpha H^1_z \rightarrow N^1_\alpha H^1_z$$
$$N^1_\alpha \rightarrow H^1_\alpha$$
$$N^2_\alpha H^2_z \rightarrow N^2_\alpha H^2_\alpha$$
$$N^2_\beta H^2_z \rightarrow -N^2_\mu H^2_\beta \quad (5.3)$$

This yields the following coherence terms immediately before the 90° $^{15}$N pulse that starts t3:

$$\rightarrow N^1_\alpha H^1_\alpha \ (\text{diagonal}) + H^1_\alpha N^2_\alpha H^2_z \ (\text{crosspeak}) - H^1_\alpha N^2_\beta H^2_z \ (\text{crosspeak}) \quad (5.4)$$

Thus, the desired magnetization – the $N_\beta$ polarized coherence of spin 2 ($N^1_\alpha N^2_\beta H^2_z$) – is transferred to $H_\beta$ polarized coherence ($H^1_\alpha N^2_\beta H^2_z$) to set up TROSY collection in t3.
Figure 23. Pulse Sequence for the 4-D ds-TNT Experiment. Narrow and wide bars represent 90° and 180° pulses, respectively. All pulses are applied along the x-axis unless noted otherwise. 90° selective water pulses are indicated by shaped bars. The delays are δ = 2.4 ms, Δ = 2.75 ms, and τm = 200 ms. Phase cycling is φ1 = [2x, 2(-x)], φ2 = [-x], φ3 = [y], φ4 = [-y], φ5 = [x, -x], φ6 = [x], φ7 = [y], and φrec = [x, -x, -x, x]. Inversion of φ3 (and φ4 for water suppression) at even numbered lattice points in t introduces a frequency shift of sw2/2 to the H1 dimension in order to center the amide signals while leaving the transmitter frequency on water. This was found to provide better water suppression than the alternative: shifting the transmitter frequency before t (H1) and applying an off-resonance water flip-back pulse (for φ4). Axial peaks are removed by setting (φ1 + 180°, φrec + 180°) and (φ5 + 180°, φrec + 180°) at even numbered lattice points in F1 and F3, respectively. Hypercomplex data collection for the two SE TROSY elements requires inversion of φ2, φrec, and G2 for the F1 dimension and inversion of φ6, φ7 and G2’ for the F3 dimension. Cosine-sine selection for the F1/ F2 dimensions is controlled by incrementing phase φ3 (and phase φ4 for water suppression). Gradient durations and field strengths are G1 = (2 ms, -18.38 G/cm), G2 = (0.2 ms, 18.32 G/cm), G1’ = (2 ms, 20.42 G/cm), G2’ = (0.2 ms, -20.36 G/cm), G3 = (0.5 ms, 11.64 G/cm), G4 = (1 ms, -18.99 G/cm), G5 = (0.5 ms, 17.77 G/cm), G6 = (0.5 ms, 15.72 G/cm), G7 = (1 ms, 23.89 G/cm), G8 = (0.5 ms, 10.01 G/cm), G9 = (0.5 ms, 10.82 G/cm), G10 = (1 ms, 21.03 G/cm), G11 = (0.5 ms, 10.41 G/cm), G12 = (0.5 ms, 14.09 G/cm). A small refocusing gradient is applied during t2 (Gb) to suppress water radiation damping.

(N2). This also converts the $N^1_\alpha$ coherence term of spin 2 (transferred during NOE mixing) to $H^1_\alpha$ coherence so that it does not interfere with frequency labeling.

Additionally, the diagonal magnetization ($N^1_\alpha H^1_\alpha$) and the $N_\alpha$ polarized coherence of spin 2 ($N^1_\alpha N^2_\alpha H^2_\alpha$) are converted to anti-TROSY coherence for t3.
Diagonal signals are suppressed by the second TROSY transfer element, which selects for the TROSY coherence pathway of spin 2 while suppressing the anti-TROSY pathways of spin 2 and spin 1. This transfer uses the ST2-PT method to convert the desired crosspeak magnetization to $N^1_{\beta}N^2_{\alpha}H^{2}_{+-}$ for direct detection (Pervushin et al. 1998b), with coherence pathway selection controlled by inversion of phases $\phi_6$ and $\phi_7$ and gradient $G_2'$. While the ST2-PT scheme is more susceptible to signal loss during coherence transfer compared to the Nietlispach technique used in the first TROSY transfer (Nietlispach 2005), the ST2-PT method proved to be better at suppressing signals from sidechain amino groups.

To test the effectiveness of diagonal suppression in our experiment, we used a $^2\text{H}/^{13}\text{C}/^{15}\text{N}$-labeled sample of the 29 kDa human carbonic anhydrase II (HCAII) to collect

**Figure 24. Diagonal Suppression with the ds-TNT Experiment.** 2-D amide-amide NOESY spectra were collected for the 29 kDa HCAII protein using an experiment without diagonal suppression (A) and the ds-TNT experiment (B).
2-D (H1, H2) planes with the ds-TNT pulse sequence and a standard $^1$H/$^{15}$N-HSQC-NOESY-$^1$H/$^{15}$N-HSQC pulse sequence which does not suppress diagonal signals. As shown in Figure 24, the ds-TNT experiment dramatically reduces the diagonal peaks down to the level of medium-sized crosspeaks.

5.2 High-Resolution Data Collection with Sparse Sampling and CLEAN

To collect a 4-D ds-TNT spectrum at high resolution, we employed cosine-weighted randomized concentric shell sampling (RCSS) – an extension of the 3-D concentric ring sampling (CRS) method (Coggins and Zhou 2007), which was developed as an alternative to radial sampling with better sensitivity and fewer artifacts (Coggins and Zhou 2008). In the RCSS method, points are set on a series of evenly spaced shells, where the distribution of points on each shell is uniform. The distribution of points across shells can be adjusted based on the properties of the signal being collected. In order to prevent the reinforcement of artifacts generated by regular patterns of sampling points, each shell is rotated randomly about all three dimensions. Finally, the sampling pattern is adapted to a fine Cartesian grid to enable rapid processing by the fast Fourier transform (FFT) without causing a significant increase in artifact levels (Coggins and Zhou 2008).
The CLEAN algorithm was used to reduce aliasing artifacts produced by RCSS data collection. As described in Section 3.5, our implementation of CLEAN is an iterative process that (1) finds the strongest peak in a spectrum, (2) generates a replica, or component, of the peak, (3) calculates the artifact pattern of the component based on knowledge of the sampling scheme, and (4) subtracts the component and its sampling artifacts from the spectrum (Coggins and Zhou 2008). While CLEAN has been traditionally applied with fitted lineshapes for each component, the use of nondecaying components (which treats each peak as a set of discrete points) is equally effective and eliminates the need to estimate linewidths. The algorithm is terminated when the noise level has stabilized or when all peaks with significant intensity have been removed. Artifact-free components are then convolved with an abridged point response function derived from the sampling pattern and added back to yield a spectrum with reduced sampling artifacts.

High-resolution 4-D data collection for the ds-TNT pulse sequence was tested with a 1 mM sample of $^{2}$H/$^{13}$C/$^{15}$N-labeled C13S Ssu72, a catalytically inactive mutant of a 23 kDa phosphatase. Data was collected on a Varian INOVA 800 MHz spectrometer equipped with a triple-resonance cold probe with Z-axis gradients. For the 4-D ds-TNT experiment, the RCSS sampling pattern contained 3,189 points distributed with cosine-weighting over 64 shells and digitized on a 64 x 64 x 64 grid. The final weighting of each grid-adapted sampling point was calculated as its Voronoi volume. Simple modification
Figure 25. Representative Planes from the 4-D ds-TNT Spectrum. Corresponding strips are from 3-D ds-TNT control spectra collected with conventional sampling. Residual diagonal signals are boxed in green. Panel (A) shows crosspeaks from residue I176 of C13S Ssu72 to D173 (1), N174 (2), D175 (3), D177 (5) and E178 (4). Panel (B) contains crosspeaks from two overlapped residues, L72 and N92. Blue numbers denote crosspeaks from L72 to Y69 (8), R70 (10), D71 (4), E73 (7), S74 (1) and K75 (5) and purple numbers correspond to crosspeaks from N92 to D90 (6), R91 (3), R93 (9) and R94 (2).

of the ds-TNT pulse sequence code allowed sampling from an explicit schedule of evolution times. The maximum evolution times were 0.0138 sec for H1 (4650 Hz spectral width) and 0.0237 sec for N1 and N2 (2700 Hz spectral width). The total measurement time was 48 hours. After FFT computation of the 4-D spectrum, the CLEAN algorithm was run to reduce sampling artifacts with a loop gain setting of 10%, a stopping threshold of 1%, and a 500 iteration maximum run for each cube as previously described (Coggins and Zhou 2008).

Representative examples from the FFT-CLEAN-processed 4-D ds-TNT spectrum are shown in Figure 25, including F1/F2 planes taken at the F3/F4 positions of residue
I176 and the overlapped residues L72/N92. The strips are from the corresponding positions in conventionally collected 3-D control spectra (see below). Residual diagonal signals are boxed in green. The pattern of residual diagonal peaks is generated by two processes: (1) incomplete suppression of the anti-TROSY coherence pathway by the spin state selective transfer elements and (2) $^{15}$N spin state relaxation during the NOE mixing period. Consistent with the results in Figure 24, the diagonals are well suppressed in all cases, with residual intensities matching medium-sized NOESY crosspeaks.

5.3 Linearity Testing in the 4-D Dataset

While RCSS data collection and FFT-CLEAN processing have been applied successfully to a 4-D HCCH-TOCSY experiment, NOESY presents a new challenge as the preservation of inherent distance constraint information requires a linear reproduction of peak shapes and intensities over a large dynamic range. In order to test for this linearity in our 4-D ds-TNT spectrum, we also collected 3-D (N1, N2, H2) and (H1, N2, H2) control spectra using conventional sampling, with spectral widths identical to those used in the 4-D experiment. The maximum evolution times for the 3-D (N1, N2, H2) experiment were 0.0237 sec (64 complex points) for N1 and N2 with a total measurement time of 32 hours (4 scans per FID). For the 3-D (H1, N2, H2) experiment, the maximum evolution times were 0.0103 sec (48 complex points) for H1 and 0.0237 sec.
(64 complex points) for N2 with a total measurement time of 48 hours. The resolution of
the H1 dimension was extended to 0.0138 sec (64 complex points) with linear prediction.
Both experiments were processed with a cosine window function and zero-filled to 128
points in each indirect dimension. A set of 376 crosspeaks from the (H1, N2, H2) 3-D ds-
TNT spectrum was used for comparisons with the 4-D dataset. For comparisons
between 3-D and 4-D spectra, peak intensity was found to be a more reliable measure
than peak volume due to increased signal overlap in the 3-D spectrum. About 12% of
the 3-D peaks correspond to two or more 4-D peaks that overlap in the 3-D spectrum,
and in these cases, a summation of the 4-D peak intensities was used. The linearity of
CLEAN processing was assessed with both peak intensities and volumes using a set of
564 non-overlapping peaks assigned in the 4-D ds-TNT spectrum.

To gauge the effects of RCSS data collection on the 4-D ds-TNT NOESY
experiment, we compared the intensities of the 376 crosspeaks in the 3-D (H1, N2, H2)
ds-TNT control with their counterparts in the FFT-processed 4-D spectrum. The results
are plotted in Figure 26. The comparison of 3-D and 4-D peak intensities shows a very
strong and consistent linear relationship that can be fit to a linear regression (y-intercept
set at zero) with an R^2 value of 0.9655. This result is consistent with the observation by
Kazimierczuk and co-workers that the FFT is a linear process and does not distort the
crosspeak information in a systematic fashion (Kazimierczuk et al. 2007). We next
evaluated the effect of the CLEAN procedure on the 4-D dataset. Noise levels for all 128
Figure 26. Comparison of Peak Intensities in 3-D and 4-D ds-TNT Spectra. Peak intensities from 376 NOESY crosspeaks were calculated from the conventionally-sampled 3-D (H1, N2, H2) ds-TNT dataset and the RCSS 4-D ds-TNT dataset with FFT processing. The data were fit to a linear model (y = kx) with an R² value of 0.9655. Relative percentage errors with respect to the 3-D peak intensities are plotted in the lower panel.

(N1, H1, N2) cubes in the 4-D spectrum were measured with and without artifact removal. This analysis showed a range of noise reductions with maximum benefits for cubes with strong signals. The largest decrease in apparent noise was 22%. We then tested the CLEAN procedure for retention of NOESY distance information with the set of 564 crosspeaks from the 4-D ds-TNT spectrum. The volume and intensity of each peak was calculated before and after the application of CLEAN and the results are plotted in Figure 27. The best linear fit to the peak volume data gives a slope of 1.00 and
Figure 27. Comparison of 4-D ds-TNT Peaks with FFT and FFT-CLEAN Processing. 564 NOESY crosspeaks from the RCSS 4-D ds-TNT dataset were used to assess the effect of artifact reduction with the CLEAN algorithm on peak intensities and volumes. In panel (A), peak intensities are plotted with and without CLEAN processing, giving a linear fit of \( y = 0.9995x \) with an \( R^2 \) value of 0.9994. In panel (B), peak volumes are plotted with and without CLEAN processing, giving a linear fit of \( y = 0.9990x \) with an \( R^2 \) value of 0.9967. For (A-B), relative percentage errors between FFT and FFT-CLEAN processing are plotted in the lower graphs.

an \( R^2 \) value of 0.99, demonstrating that CLEAN does not systematically distort signal information for the ds-TNT experiment.
5.4 Summary and Discussion

Amide-amide NOESY is one of the central experiments used to calculate global folds for large proteins and to verify and extend the assignment of backbone amide groups. However, the analysis of amide-amide NOESY data from high molecular weight proteins is hindered by the large number of signals that leads to resonance overlap and assignment ambiguity. Increasing the dimensionality of these datasets while maintaining high digital resolution is the most straightforward solution to both problems. The diagonal-suppressed NH-NH TROSY-NOESY-TROSY (ds-TNT) experiment presented here is designed for structural studies of large protein targets; accordingly, it has two features that should make it exceptionally beneficial to these studies.

The first key feature of the ds-TNT experiment is TROSY spin state selection in all four dimensions. As the molecular weight of NMR targets increases, TROSY optimization becomes more important, both for minimizing relaxation losses and for reducing peak overlap by sharpening linewidths. Suppression of diagonal signals is the second advantage of the ds-TNT experiment. Importantly, this suppression is achieved by manipulating the spin state selective transfers which are already used for TROSY selection, eliminating the loss in sensitivity or increase in spectrometer time inherent to other methods (Harbison et al. 1985; Wu et al. 2004). As shown in Figures 24 and 25, the
ds-TNT experiment reduces the strength of diagonal signals to that of medium-sized NOESY crosspeaks. This diagonal suppression allows the identification of crosspeaks that lie near diagonal peaks, increasing the completeness of NOESY assignment and potentially improving the accuracy of structure calculations; however, the most compelling reason for diagonal suppression comes from the application of sparse sampling schemes: the reduction of sampling artifacts. As the artifacts generated by each peak are proportional to its intensity, strong diagonal peaks produce strong artifact patterns that can obscure weak NOESY crosspeaks. Therefore, diagonal suppression makes the ds-TNT experiment particularly attractive for high-dimensionality data collection with many of the sparse sampling methods.

RCSS data collection and CLEAN processing have been demonstrated previously on 4-D HCCH-TOCSY (Coggins and Zhou 2008), but NOESY experiments present unique challenges to these methods, as the distance constraint information inherent in peak intensities and volumes must be faithfully reproduced to avoid bias in structure calculations. The comparison of crosspeak intensities for the 4-D ds-TNT spectrum and 3-D control (Figure 26) shows that essential peak information is replicated with consistent accuracy in the sparsely-sampled dataset. Furthermore, the strong linear relationship between 4-D peak intensities and volumes before and after CLEAN processing (Figure 27) demonstrates that the removal of aliasing artifacts does not systematically alter distance constraint information. Taken together, we conclude that
RCSS data collection and FFT-CLEAN processing can be applied to the 4-D ds-TNT experiment without sacrificing the accuracy of subsequent structure calculations. The advantages of high-resolution 4-D NOESY data can be clearly appreciated in the example planes of Figure 25 where signal overlap in the control 3-D ds-TNT spectra would preclude the complete and accurate assignment of crosspeaks.

The feasibility of applying sparse sampling techniques to collect high-resolution 4-D ds-TNT spectra is particularly exciting in light of the time savings over conventional sampling methods. A 4-D ds-TNT spectrum recorded conventionally with 64-point resolution in each indirect dimension (after two-fold linear prediction) would require nearly 3 weeks; the 4-D ds-TNT spectrum presented here was recorded with 64-point resolution in only 48 hours, a more than 10-fold reduction in spectrometer time. The success in applying RCSS data collection and FFT-CLEAN processing to obtain high-resolution 4-D NH-NH NOESY will likely be very helpful in extending NMR structure determination to high molecular weight targets.
6. Global Fold Calculations with High-Resolution 4-D NOESY

Protein structure determination by NMR has been steadily expanding to medium- and large-sized targets, which present unique challenges, including severe resonance overlap caused by large numbers of proton signals. As described in Section 3.2, one common strategy for overcoming this obstacle is to use labeling schemes that yield selectively-protonated samples – typically containing only amide and Ile δ1, Leu, and Val (ILV) methyl protons in an otherwise deuterated background – in order to simplify NOESY spectra while maintaining sufficient long-range NOE information for an initial structure determination (Gardner et al. 1997; Goto et al. 1999). The global folds derived from these sparse distance constraints constitute a critical first step in the calculation of a high-resolution structure, by greatly reducing the number of possible assignments of crosspeaks in NOESY spectra collected with fully-protonated samples. However, the low density and limited redundancy of observable NOE information in sparsely-protonated samples makes it critical to identify and unambiguously assign a high percentage of the NOE crosspeaks. Unfortunately, this is a challenging task using conventional 3-D NOESY experiments, as either the donor or acceptor group of each NOE crosspeak is encoded with only a single frequency. Even in sparsely-protonated samples of large proteins, there is significant degeneracy of signals along a single
dimension, particularly in the poorly-dispersed methyl region, which precludes complete and unambiguous assignment of crosspeaks in 3-D NOESY experiments. A simple and effective solution to this problem is to collect high-resolution 4-D NOESY data.

Despite the potential advantage of high-resolution 4-D NOESY, its use has been limited to a small number of structure determination studies (Hiller et al. 2009; Tugarinov et al. 2005a). The major obstacle to its wide implementation is the prolonged measurement times (many weeks) required by conventional Nyquist sampling of individual dimensions (see Section 3.4). However, recent work on the development and optimization of sparse sampling patterns in the time domain as well as methods for reconstructing spectral information from sparsely-sampled datasets has made it possible to collect high-resolution 4-D spectra in a fraction of the time required by Nyquist sampling (Coggins et al. 2010; Kazimierczuk et al. 2010). The inherent drawback of not fulfilling the Nyquist criteria is the appearance of aliasing artifacts (Coggins and Zhou 2006; Sarty 2003). Importantly, the level of the aliasing artifacts generated by each signal is proportional to the signal’s intensity. This becomes particularly problematic in NOESY spectra with diagonal signals, as the dynamic range between the intensities of diagonal signals and weak NOE crosspeaks is very large. In these spectra, aliasing artifacts from diagonal signals, which appear as noise for datasets collected with
randomized sampling patterns, can obfuscate critical NOE information (Hiller et al. 2009).

Recently, we and others have shown that aliasing artifacts can be removed with a processing method called CLEAN that iteratively identifies the strongest signal in a spectrum and subtracts its artifact pattern (Coggins and Zhou 2008; Högbom 1974; Stanek and Kozminski 2010). While our original implementation of CLEAN was successfully applied to 4-D TOCSY (Coggins and Zhou 2008) and diagonal-suppressed 4-D NOESY (see Chapter 5), it was insufficient for the most challenging applications, like 4-D NOESY with diagonals. In this chapter, we present the first application of ultrasparse sampling and SCRUB, an optimized artifact removal algorithm described in Section 3.5, to 4-D NOESY containing diagonal peaks, with the goal of efficiently determining global folds.

6.1 The 4-D Time-Shared NOESY Experiment

In order to collect all the possible NOE information for ILV methyl- and amide-protonated proteins, we designed a 4-D time-shared (TS) NOESY experiment that simultaneously records amide-amide, methyl-methyl, amide-methyl, and methyl-amide NOEs in $^1$H/$^{15}$N-HSQC-NOESY-$^1$H/$^{15}$N-TROSY, $^1$H/$^{13}$C-HSQC-NOESY-$^1$H/$^{13}$C-HSQC, $^1$H/$^{15}$N-HSQC-NOESY-$^1$H/$^{13}$C-HSQC, and $^1$H/$^{13}$C-HSQC-NOESY-$^1$H/$^{15}$N-TROSY spectra,
respectively (Figure 28). This is made possible by implementing concatenated $^{13}$C and $^{15}$N coherence transfer elements and joint $^{13}$C/$^{15}$N evolution periods that provide optimal resolution for both nuclei (Parella and Nolis 2010). Our experiment is conceptually similar to a previous implementation (Frueh et al. 2006), but features sensitivity-enhanced coherence transfers both before and after the NOE transfer period to provide optimal sensitivity.

Figure 28. Pulse Sequence for the 4-D TS NOESY Experiment. Narrow and wide bars represent 90° and 180° pulses, respectively. All pulses are applied along the x-axis unless noted otherwise. 90° selective water pulses are indicated by short, shaped bars. All shaped carbon pulses are 281 μs off-resonance Isnob2 pulses (centered at 175 ppm with a bandwidth of 30 ppm) that refocus carbonyl coupling during nitrogen chemical shift evolution (Kupce et al. 1995). The delays are $\tau_1 = 1.7$ ms ≈ 1/4$J_{CH}$, $\tau_2 = 0.7$ ms ≈ 1/4$J_{NH}$ – 1/4$J_{CH}$, $\tau_3 = 2.4$ ms ≈ 1/4$J_{NH}$, $\tau_4 = 1.05$ ms ≈ 1/4$J_{NH}$ – 1/4$J_{CH}$, $\tau_5 = 2.75$ ms ≈ 1/4$J_{NH}$, $\tau_6 = 1.34$ ms ≈ 1/8$J_{NH}$, $\tau_{m} = 200$ ms, and $\Delta = 200$ μs. Time increments are set to $\Delta t_{1a} = 1/s_{WN} – 1/s_{WC}$, $\Delta t_{1b} = 1/s_{WC}$, $\Delta t_2 = 1/s_{WH}$, $\Delta t_{3a} = 1/s_{WN} – 1/s_{WC}$, and $\Delta t_{3b} = 1/s_{WC}$. The proton inversion pulse at point c is required to refocus $J_{NH}$ scalar coupling in the two $\Delta$ delays (points a-b and c-d), while proton decoupling during $t_{1a}$ (points b-c) and $t_{1b}$ (points d-e) evolution is achieved by the proton inversion pulses centered in these periods. The shaped bar marked ‘Me’ represents a 587 μs off-resonance G3 inversion pulse (centered at -2.0 ppm with a bandwidth of 8.0 ppm) that selectively refocuses methyl proton coherence during $t_{3b}$ without affecting water and amide signals (Emsley and Bodenhausen 1990). The $^{15}$N carrier frequency is shifted 45 Hz downfield during the NOE mixing period to re-center nitrogen signals on the TROSY component. The spin-state-selective element between points f and g is used for active suppression of the $^1$H-$^{15}$N anti-TROSY component (Yang and Kay 1999a). Carbon decoupling during acquisition is achieved by using a WURST-
40 sequence with a field strength of 8.0 kHz (Kupce and Freeman 1995). Phase cycling is
\( \phi_1 = [2x, 2(-x)], \phi_2 = [x], \phi_3 = [-x], \phi_4 = [x], \phi_5 = [x], \phi_6 = [x, -x], \phi_7 = [x], \phi_{rec} = [x, -x, -x, x]. \)
Inversion of \( \phi_4 \) (and \( \phi_5 \) for water suppression) at even numbered lattice points in \( t_2 \) introduces a frequency shift of \( sw_t/2 \) to the H1 dimension in order to center the amide signals while leaving the transmitter frequency on water. Axial peaks are removed by setting \( (\phi_1 + 180^\circ, \phi_{rec} + 180^\circ) \) and \( (\phi_6 + 180^\circ, \phi_{rec} + 180^\circ) \) at even numbered lattice points in F1 and F3, respectively. Hypercomplex data collection for the two time-shared, sensitivity-enhanced coherence transfers requires inversion of \( \phi_3, \phi_{rec}, \) and G2 for the F1 dimension and inversion of \( \phi_7 \) and G2’ for the F3 dimension. Cosine-sine selection for the F1/ F2 dimensions is controlled by incrementing phase \( \phi_4 \) (and phase \( \phi_5 \) for water suppression). For the gradient selection of nitrogen and carbon coherence pathways, nitrogen single quantum coherence is encoded with the sum of G1N and G1C (or G1’N and G1’C in the second transfer) whereas carbon single quantum coherence is encoded only by G1C. Therefore, the duration of these gradients and the decoding gradient G2 are set such that \( \tau_{G1C} = 4\tau_{G2} \) and \( \tau_{G1C} + \tau_{G1N} = 10\tau_{G2} \), while the field strengths are optimized empirically, with the G1C gradient calibrated first. Gradient durations and field strengths are G1N = (1.2 ms, -18.34 G/cm), G1C = (0.8 ms, 18.51 G/cm), G2 = (0.2 ms, 18.38 G/cm), G1’N = (1.2 ms, -20.38 G/cm), G1’C = (0.8 ms, 20.56 G/cm), G2’ = (0.2 ms, 20.42 G/cm), G3 = (0.5 ms, 11.64 G/cm), G4 = (1 ms, -18.99 G/cm), G5 = (0.5 ms, 17.77 G/cm), G6 = (0.5 ms, 15.72 G/cm), G7 = (0.5 ms, 10.01 G/cm), G8 = (0.7 ms, 19.81 G/cm), G9 = (1 ms, -18.99 G/cm), G10 = (0.5 ms, 10.41 G/cm), and G11 = (0.5 ms, 14.09 G/cm). A small refocusing gradient is applied during \( t_2 \) (Gb) to suppress water radiation damping. In order to separate NOESY pathways originating from methyl and amide protons during data processing, two sets of FIDs are collected for each set of \( (t_{1a} + t_{1b}, t_2, t_{3a} + t_{3b}) \) delays, with \( \phi_2 = [y] \) in the second set of FIDs to selectively invert methyl signals during the first coherence transfer (see Section 6.6.2).

6.2 Application of Sparse Sampling and Artifact Removal with SCRUB

We tested the 4-D TS NOESY experiment on two ILV methyl- and amide-protonated proteins, the 23 kDa phosphatase Ssu72 and 29 kDa human carbonic anhydrase II (HCAII), using a cold probe-equipped 800 MHz spectrometer. To collect
high-resolution 4-D datasets, we employed cosine-weighted randomized concentric shell sampling, a pattern designed to prevent the reinforcement of aliasing artifacts, with points distributed on a grid with 64-point resolution in each indirect dimension. For each dataset, 3189 complex points were sampled, representing just 1.2% of the measurements required by conventional Nyquist sampling at equivalent resolution. This ultrasparse sampling allowed each 4-D dataset to be collected in only 4 days. The four simultaneously-recorded NOESY pathways were separated and processed with either the fast Fourier transform (FT) alone or the FT followed by artifact removal with SCRUB. This produced four 4-D spectra for each protein with dimensions (N1, HN1, N2, HN2), (C1, HC1, C2, HC2), (N1, HN1, C2, HC2) and (C1, HC1, N2, HN2), where label numbers 1 and 2 denote NOE donor and acceptor groups, respectively, and the directly detected dimension is HN2/HC2. The final resolution of each indirect dimension is 128 points.

We next assessed the extent of aliasing artifacts in the FT-processed 4-D TS spectra and the ability of SCRUB to reduce them. Importantly, the directly detected dimension is not sparsely-sampled, and aliasing artifacts at each point in the frequency domain along this dimension are independent of one another. Therefore, we measured noise levels at every HN2/HC2 point, each of which comprises a 3-D spectra defined by the three indirect dimensions. As expected, certain regions of the FT-processed methyl-methyl and amide-amide spectra have severely elevated noise levels due to the aliasing
artifacts generated by strong diagonal signals (Figures 29 and 30). The level of aliasing artifacts at each HN2/HC2 point is determined by the number and intensities of signals at the corresponding proton frequency, so that the pattern of noise along the direct dimension, or noise profile, is roughly correlated to the summed 1-D proton projection of the appropriate HSQC spectra (green traces in Figures 29 and 30). With SCRUB

Figure 29. Artifact Removal in the 4-D TS NOESY Spectra for Ssu72. Noise levels were measured at each HN2/HC2 point in the amide-amide (A) and methyl-methyl (C) NOESY spectra after FT or SCRUB processing. Green traces show summed 1-D proton projections from $^1$H/$^{15}$N- or $^1$H/$^{13}$C-HSQC spectra. (B, D) Representative (N1, HN1) or (C1, HC1) planes for each spectrum processed with the FT (right panel) or SCRUB (left panel) were chosen at the points in the direct dimension marked by asterisks in (A) and (C) and plotted at identical contour levels. 1-D slices along the green dashed lines are shown above each panel and highlight long-range crosspeaks that are lost to aliasing noise in the FT-processed spectra. Numbered crosspeaks in (B) are from R35 HN to (1) S36 HN, (2) V11 HN, (3) V34 HN, (4) V9 HN, and (5) A10 HN. Numbered crosspeaks in (D) are from L169 δ1 to (1) I173 δ1, (2) I108 δ1, (3) L169 δ2, and (4) L138 δ2. The same analysis for the HCAII dataset is presented in Figure 30.
Figure 30. Artifact Removal in the 4-D TS NOESY Spectra for HCAII. This figure is analogous to Figure 29. Noise levels were measured at each HN2/HC2 point in the amide-amide (A) and methyl-methyl (C) NOESY spectra after FT or SCRUB processing. Green traces show summed 1-D proton projections from $^1$H/$^{15}$N- or $^1$H/$^{13}$C-HSQC spectra. (B, D) Representative (N1, HN1) or (C1, HC1) planes for each spectrum processed with the FT (right panel) or SCRUB (left panel) were chosen at the points in the direct dimension marked by asterisks in (A) and (C) and plotted at identical contour levels. 1-D slices along the green dashed lines are shown above each panel and highlight long-range crosspeaks that are lost to aliasing noise in the FT-processed spectra. Numbered crosspeaks in (B) are from K126 HN to (1) G128 HN, (2) T87 HN, (3) T125 HN, and (4) Y127 HN. Numbered crosspeaks in (D) are from L250 $\delta_2$ to (1) I255 $\delta_1$ and (5) L250 $\delta_2$ and from L184 $\delta_1$ to (2) V49 $\gamma_2$, (3) V49 $\gamma_1$, and (4) L184 $\delta_2$. 

processing, however, there was dramatic reduction in the aliasing artifacts, with noise levels at each HC2/HN2 point decreased to, or close to, the baseline experimental noise. Maximum noise reductions of 7.37-fold and 11.39-fold were observed for the HCAII amide-amide and methyl-methyl spectra, respectively. The noise profiles of the FT-
processed amide-methyl and methyl-amide spectra, which do not contain diagonal peaks, were significantly more uniform, and SCRUB processing effectively suppressed artifacts in these spectra as well.

6.3 Peak Identification Tests

The success of SCRUB processing in removing artifact noise from the diagonal-containing TS spectra should significantly aid in the identification of NOE crosspeaks. To quantitatively measure this benefit, we picked and assigned all crosspeaks in the full SCRUB-processed TS datasets for both test proteins, and filtered them based on the noise levels of the FT- or SCRUB-processed datasets. A peak was deemed identifiable if its intensity was at least 3 times the noise level at its resident HC2/HN2 point. Each peak was then matched to its assigned interproton distance based on a reference crystal structure. Figures 31 and 32 show the resulting histograms for the methyl-methyl and amide-amide crosspeaks. For HCAII, 87% (1.94) and 97% (1.94) of the interproton distances 4.5 Å or shorter were observed as NOE crosspeaks in the SCRUB-processed methyl-methyl and amide-amide spectra, respectively, with numbers in parentheses indicating the redundancy of the NOE information (peaks per unique interproton distance). In the FT-processed spectra, however, both the data completeness and redundancy dropped significantly, to 57% (1.56) in the methyl-methyl spectrum and
Figure 31. Peak Identification in the 4-D TS NOESY Spectra for Ssu72. Removal of aliasing artifacts promotes peak identification in the amide-amide (A, B) and methyl-methyl (C, D) 4-D TS NOESY spectra. NOE crosspeaks deemed identifiable in spectra processed with either the FT or SCRUB were matched to their corresponding interproton distances, and unique distances are plotted as histograms in (A) and (C). Each interproton distance was counted as observed if either corresponding NOE (Hi → Hj or Hj → Hi) was identifiable. ‘All expected’ refers to the total number of interproton distances calculated from a reference crystal structure (PDB code 3P9Y). The histograms in (B) and (D) show the percent of observed distances relative to all expected. The small number of very long distances (>8.0 Å) observed in the methyl-methyl spectra likely results from incorrect rotamer assignments in the reference crystal structure due to the insufficient resolution of the model (2.1 Å) for precisely defining ILV sidechain orientations. For example, the two longest distances – 9.31 Å between I173 δ1 and L176 δ2 and 8.55 Å between V43 γ2 and L45 δ2 – are substantially shorter in a second crystal structure (PDB code 3FDF, 3.2 Å resolution) – 6.99 Å and 7.87 Å, respectively – as the result of different rotamer selections for residues I173 and L45. The same analysis for the HCAII dataset is presented in Figure 32.
Figure 32. Peak Identification in the 4-D TS NOESY Spectra for HCAII. Removal of aliasing artifacts promotes peak identification in the amide-amide (A, B) and methyl-methyl (C, D) 4-D TS NOESY spectra. NOE crosspeaks deemed identifiable in spectra processed with either the FT or SCRUB were matched to their corresponding interproton distances, and unique distances are plotted as histograms in (A) and (C). Each interproton distance was counted as observed if either corresponding NOE (Hi → Hj or Hj → Hi) was identifiable. ‘All expected’ refers to the total number of interproton distances calculated from a reference crystal structure (PDB code 2ILI). The histograms in (B) and (D) show the percent of observed distances relative to all expected. Short interproton distances that were not identified in the SCRUB-processed methyl-methyl spectrum are the result of signal degeneracy. For example, the four unidentified distances under 3.5 Å (out of 53 total) correspond to intraresidue methyl-methyl NOE crosspeaks for residues V109, V121, L143, and L223 that overlap with diagonal signals. The same analysis for the Ssu72 dataset is presented in Figure 31.
69% (1.74) in the amide-amide spectrum, over the same range of interproton distances. As shown in Figures 31 and 32, longer interproton distances – which correspond to weaker NOE crosspeaks – are preferentially lost in the FT-processed datasets. Many of these lower-intensity peaks provide long-range contacts that are crucial for successfully defining a target’s topology (see examples in Figures 29 and 30). In the amide-methyl and methyl-amide datasets for both test proteins, nearly all peaks in the SCRUB-processed spectra were also identifiable in the FT-processed spectra, a result that is consistent with the lower artifact noise levels observed in these diagonal-less datasets.

6.4 Structure Calculations with CYANA

To illustrate the impact of the high-resolution, high-dimensionality TS NOESY data on structure determination, we used CYANA (Herrmann et al. 2002) to calculate the global folds of Ssu72 and HCAII. The only input constraints were (1) dihedral restraints derived from the backbone chemical shifts (Shen et al. 2009) and (2) peak lists from all four TS NOESY spectra. Calculations were performed with either manually-assigned peak lists or unassigned peak lists and the automated assignment protocol of CYANA. The extra dimensionality of the TS dataset afforded by ultrasparse sampling and SCRUB processing should be especially beneficial for automated assignment efforts, as it significantly reduces the number of assignment possibilities for each crosspeak. To
test this idea, auto-assignment calculations were run with either 4-D peak lists or
simulated 3-D peak lists, which were generated by removing the first indirect
heteronuclear dimension (C1/N1) of the 4-D peak lists. For each setup, five independent
calculations were performed. Ensembles with the median RMSD of backbone atoms in
the core domains (RMSD_{bb,core}) are shown in Figure 33, and evaluations of automated
peak assignment correctness and ensemble convergence and accuracy are presented in
Tables 3 and 4.

For both test proteins, global fold calculations based on manually-assigned TS
NOESY crosspeaks yielded accurate and well-converged structural ensembles, a result
in line with the excellent completeness of observed interproton distances. For HCAII,
the ensembles from five independent calculations have a mean RMSD_{bb,core} of 0.904 ±
0.084 Å and a bias to the reference crystal structure over the same set of atoms of 1.634 ±
0.075 Å. For Ssu72, the mean RMSD_{bb,core} is 0.923 ± 0.128 Å and the bias is 2.076 ± 0.126
Å. Both structure calculations also demonstrate the limitations of the ILV methyl- and
amide-protonated labeling strategy. For example, the first 21 residues of HCAII lack an
ILV residue and do not feature any long-range amide-mediated contacts to the main
body of the enzyme, and hence are not converged in the structure calculations. In the
case of Ssu72, the enzyme contains a substrate-binding subdomain which lacks long-
range ILV methyl- and amide-mediated contacts to the core domain. While the
Figure 33. CYANA Structure Calculations Using TS NOESY Data. Ensembles for Ssu72 (top row) and HCAII (bottom row) were calculated using manually-assigned peaks (B, F), auto-assigned 4-D peaks (C, G), and auto-assigned simulated 3-D peaks (D, H). Reference crystal structures are shown in A (PDB code 3FDF) and E (PDB code 2ILI). Insets in B and C show the alignment of the substrate-binding subdomain of Ssu72, which is not constrained to the main phosphatase domain by the TS NOESY data. The ensembles in D and H are aligned over the largest segments of converged structure. The unconverged N-terminus of HCAII (21 residues) is omitted from F-H for clarity.

The subdomain structure converges (insets in Figure 33), its orientation relative to the core of the enzyme is poorly defined.

The results from structure calculations with automated assignment of the TS NOESY peak lists highlight the advantage of the 4-D datasets. For both targets, automated assignment of the 4-D peak lists by CYANA was remarkably accurate with low levels of assignment ambiguity. For HCAII, 98% of all crosspeaks were assigned...
correctly, including 97% of the critical long-range NOEs, and only 13% of crosspeaks were assigned ambiguously. A similar level of success was observed in the Ssu72 calculations, with 97% of crosspeaks assigned correctly, including 94% of the long-range NOEs, and only 11% assigned ambiguously. These calculations produced high-quality structural ensembles, with mean RMSDs_{bb,core} of 0.958 ± 0.081 Å and 1.009 ± 0.158 Å for HCAII and Ssu72, respectively. The biases of the ensembles were also close to those observed in ensembles calculated with manually-assigned peak lists, indicating that the automated assignment did not introduce any systematic structural distortions. The results of automated assignment with simulated 3-D peak lists, however, were strikingly worse. In the HCAII calculation, 84% of all crosspeaks were assigned correctly, including only 69% of the long-range NOEs, and 44% were assigned ambiguously. For Ssu72, 82% were assigned correctly, including only 60% of the long-range NOEs, with 49% assigned ambiguously. In contrast to the calculations with 4-D peak lists, the CYANA runs with 3-D peak lists produced structures that either failed to converge or converged poorly to inaccurate conformations. For Ssu72, the C-terminal portion of the central β-sheet converges, but the N-terminus is largely unrestrained. Similarly, small portions of the HCAII structures converge accurately along the central β-sheet, however, the remainder of the protein is pulled into a vastly non-native conformation.
Table 3. Peak Assignment and Structure Determination Statistics for the CYANA Calculations of Ssu72 with TS NOESY Data.

<table>
<thead>
<tr>
<th></th>
<th>manual</th>
<th>4-D automated</th>
<th>3-D automated</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Peak Assignment Statistics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all spectra</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct – all(^b)</td>
<td>1563</td>
<td>1517 (174)</td>
<td>1280 (621)</td>
</tr>
<tr>
<td>correct – long-range (^b)</td>
<td>490</td>
<td>461 (89)</td>
<td>296 (185)</td>
</tr>
<tr>
<td>unassigned</td>
<td>0</td>
<td>28</td>
<td>27</td>
</tr>
<tr>
<td>inconsistent (^c)</td>
<td>0</td>
<td>18</td>
<td>256 (129)</td>
</tr>
<tr>
<td>amide-amide</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct – all(^b)</td>
<td>703</td>
<td>692 (19)</td>
<td>618 (216)</td>
</tr>
<tr>
<td>correct – long-range (^b)</td>
<td>91</td>
<td>84 (0)</td>
<td>55 (21)</td>
</tr>
<tr>
<td>unassigned</td>
<td>0</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>inconsistent (^c)</td>
<td>0</td>
<td>1</td>
<td>69 (50)</td>
</tr>
<tr>
<td>methyl-methyl</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct – all(^b)</td>
<td>307</td>
<td>300 (75)</td>
<td>181 (128)</td>
</tr>
<tr>
<td>correct – long-range (^b)</td>
<td>201</td>
<td>195 (55)</td>
<td>110 (85)</td>
</tr>
<tr>
<td>unassigned</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>inconsistent (^c)</td>
<td>0</td>
<td>6</td>
<td>124 (79)</td>
</tr>
<tr>
<td>amide-methyl</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct – all(^b)</td>
<td>321</td>
<td>302 (21)</td>
<td>275 (131)</td>
</tr>
<tr>
<td>correct – long-range (^b)</td>
<td>123</td>
<td>114 (9)</td>
<td>81 (43)</td>
</tr>
<tr>
<td>unassigned</td>
<td>0</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>inconsistent (^c)</td>
<td>0</td>
<td>3</td>
<td>37</td>
</tr>
<tr>
<td>methyl-amide</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct – all(^b)</td>
<td>232</td>
<td>223 (59)</td>
<td>206 (146)</td>
</tr>
<tr>
<td>correct – long-range (^b)</td>
<td>75</td>
<td>68 (25)</td>
<td>50 (36)</td>
</tr>
<tr>
<td>unassigned</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>inconsistent (^c)</td>
<td>0</td>
<td>8</td>
<td>26</td>
</tr>
</tbody>
</table>

**Ensemble Convergence and Accuracy** \(^d\)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean RMSD(_{bb,cov}) (Å)</td>
<td>0.923 ± 0.128</td>
<td>1.009 ± 0.158</td>
<td>8.784 ± 1.305</td>
</tr>
<tr>
<td>Mean RMSD(_{bb,full}) (Å)</td>
<td>1.291 ± 0.095</td>
<td>1.296 ± 0.114</td>
<td>9.242 ± 1.514</td>
</tr>
<tr>
<td>Mean bias(_{bb,cov}) (Å)</td>
<td>2.076 ± 0.126</td>
<td>2.926 ± 0.262</td>
<td>25.461 ± 4.462</td>
</tr>
<tr>
<td>Mean bias(_{bb,full}) (Å)</td>
<td>2.413 ± 0.133</td>
<td>3.107 ± 0.198</td>
<td>25.252 ± 4.148</td>
</tr>
</tbody>
</table>

\(^a\) Assignment statistics are presented for a single CYANA structure calculation that produced an ensemble with the median RMSD\(_{bb,cov}\) out of five independent calculations.

\(^b\) Numbers in parentheses indicate peaks with ambiguous assignments.

\(^c\) Numbers in parentheses indicate peaks with diagonal assignments. ‘Inconsistent’ denotes peaks with automated assignments that differ from the manual ones; however, for the 4-D peak lists, the automated assignment is also compatible with the reference crystal structure in all cases.

\(^d\) Mean RMSD and bias were calculated with five ensembles – produced by independent CYANA runs and containing five structures each – over all non-hydrogen backbone atoms (bb) in the converged portion of either the full protein (full, residues 5-37, 97-195) or residues in secondary structure elements (core). Bias represents the average pairwise RMSD to the reference crystal structure (PDB code 3FDF).
Table 4. Peak Assignment and Structure Determination Statistics for the CYANA Calculations of HCAII with TS NOESY Data.

<table>
<thead>
<tr>
<th></th>
<th>manual</th>
<th>4-D automated</th>
<th>3-D automated</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Peak Assignment Statistics</strong>&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Spectra</td>
<td>1818(178)</td>
<td>1778(233)</td>
<td>1529(674)</td>
</tr>
<tr>
<td></td>
<td>698</td>
<td>678(126)</td>
<td>485(277)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>32</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>8</td>
<td>229(102)</td>
</tr>
<tr>
<td>amide-amide</td>
<td>879</td>
<td>853(16)</td>
<td>738(222)</td>
</tr>
<tr>
<td></td>
<td>271</td>
<td>261(8)</td>
<td>175(64)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>25</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>102(51)</td>
</tr>
<tr>
<td>methyl-methyl</td>
<td>346</td>
<td>341(142)</td>
<td>252(170)</td>
</tr>
<tr>
<td></td>
<td>228</td>
<td>224(97)</td>
<td>156(116)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>5</td>
<td>92(51)</td>
</tr>
<tr>
<td>amide-methyl</td>
<td>330</td>
<td>323(30)</td>
<td>300(135)</td>
</tr>
<tr>
<td></td>
<td>114</td>
<td>110(11)</td>
<td>90(52)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>methyl-amide</td>
<td>263</td>
<td>261(45)</td>
<td>239(147)</td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>83(10)</td>
<td>64(45)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>14</td>
</tr>
</tbody>
</table>

**Ensemble Convergence and Accuracy**<sup>d</sup>

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean RMSD&lt;sub&gt;bb,core&lt;/sub&gt; (Å)</td>
<td>0.904 ± 0.084</td>
<td>0.958 ± 0.081</td>
<td>3.128 ± 1.272</td>
</tr>
<tr>
<td>Mean RMSD&lt;sub&gt;bb,full&lt;/sub&gt; (Å)</td>
<td>1.340 ± 0.109</td>
<td>1.774 ± 0.272</td>
<td>9.579 ± 2.648</td>
</tr>
<tr>
<td>Mean bias&lt;sub&gt;bb,core&lt;/sub&gt; (Å)</td>
<td>1.634 ± 0.075</td>
<td>1.836 ± 0.300</td>
<td>8.731 ± 3.430</td>
</tr>
<tr>
<td>Mean bias&lt;sub&gt;bb,full&lt;/sub&gt; (Å)</td>
<td>2.455 ± 0.084</td>
<td>3.331 ± 0.185</td>
<td>16.689 ± 2.177</td>
</tr>
</tbody>
</table>

<sup>a</sup> Assignment statistics are presented for a single CYANA structure calculation that produced an ensemble with the median RMSD<sub>bb,core</sub> out of five independent calculations.

<sup>b</sup> Numbers in parentheses indicate peaks with ambiguous assignments.

<sup>c</sup> Numbers in parentheses indicate peaks with diagonal assignments. 'Inconsistent' denotes peaks with automated assignments that differ from the manual ones; however, for the 4-D peak lists, the automated assignment is also compatible with the reference crystal structure in all cases.

<sup>d</sup> Mean RMSD and bias were calculated with five ensembles – produced by independent CYANA runs and containing five structures each – over all non-hydrogen backbone atoms (bb) in the converged portion of either the full protein (full, residues 22-260) or residues in secondary structure elements (core). Bias represents the average pairwise RMSD to the reference crystal structure (PDB code 2ILI).
6.5 Summary

In conclusion, we have demonstrated that SCRUB processing can be successfully applied to sparsely-sampled NOESY experiments that feature large dynamic ranges of signals in order to rescue weak NOE crosspeaks from aliasing artifact noise. In our analysis, we were routinely able to identify NOE crosspeaks with intensities 100-fold less than diagonal signals. For example, peaks 1 to 3 in Figure 29B and peaks 2 and 4 in Figure 30B are between 90-fold and 100-fold weaker than the strongest diagonal signals at their resident HC2 points. Coupling ultrasparse sampling with the TS NOESY technique maximizes the efficiency of data collection, allowing us to collect all the 4-D data necessary for global fold calculations – at a resolution unachievable with conventional methods – in only 4 days. As shown here, this high-resolution, high-dimensionality data has an enormous practical benefit in automated NOE assignment procedures. For proteins like HCAII and Ssu72, where there is sufficient long-range distance constraint information available in sparsely-protonated samples, it is possible to go from NMR sample to global fold in a matter of days.

Several other methods have been developed for processing sparsely-sampled NMR data, including two – coupled-multidimensional decomposition (Co-MDD) and forward maximum entropy reconstruction (FM) – that have been applied to 4-D NOESY experiments (Hiller et al. 2009; Hyberts et al. 2009). Both methods reconstruct the full
time domain dataset by iterative fitting against the frequency domain spectrum. It should be noted that these procedures require sampling rates of at least 12% of the Nyquist grid, or 10 times the ultrasparse sampling rate used in this study (1.2%). Although high sampling rates may be required for low-sensitivity applications, as we show in this study, much lower rates are sufficient for observing nearly complete sets of interproton distances through 4.5 Å in selectively-protonated proteins of up to 29 kDa. Due to the advantages afforded by 4-D NOESY data, we expect that the methods presented here will be particularly useful as NMR structure determination continues to push toward higher molecular weight targets.

6.6 Materials and Methods

6.6.1 Sample Preparation

Full-length human carbonic anhydrase II (HCAII) was expressed and purified as described previously (Venters et al. 1995). The expression and purification protocol for Ssu72 is described in Section 2.5.1. The Ssu72 construct contains an active site mutation (C13S) that abolishes its phosphatase activity (see Section 2.1.2). Perdeuterated proteins were expressed in D₂O M9 minimal media with ¹⁵N-NH₄Cl and ³H/¹³C-glucose, with the addition of 85 mg [3-²H] ¹³C-α-ketoisovalerate and 50 mg [3,3-²H] ¹³C-α-ketobutyrate ~1
hour prior to induction for selective protonation of ILV methyl groups (Goto et al. 1999). Samples prepared from 10% 13C-glucose M9 minimal media were used to stereospecifically assign valine and leucine methyl groups (Neri et al. 1989). All isotopes were purchased from Cambridge Isotope Laboratories, Inc. Final sample conditions were 25 mM Tris-HCl pH 8.0, 25 mM KCl, 2 mM dithiothreitol, and 5% D2O for Ssu72, and 100 mM potassium phosphate pH 6.8, 2 mM dithiothreitol, and 10% D2O for HCAII. All samples contained ~1 mM protein.

### 6.6.2 Data Acquisition and Processing

4-D time-shared (TS) NOESY datasets were collected at 30°C for Ssu72 and 25°C for HCAII on an 800 MHz Varian Inova spectrometer equipped with a triple-resonance, cryogenically-cooled probe. To collect the datasets at high resolution, we used randomized concentric shell sampling (RCSS) with 3189 points distributed with cosine-weighting over 64 shells and digitized on a 64 × 64 × 64 grid (Coggins and Zhou 2008). Modification of the TS NOESY pulse sequence code allowed sampling from an explicit schedule of evolution times. Maximum evolution times were 0.0119 s for HN1/HC1 (5400 Hz spectral width), 0.0237 s for N1 and N2 (2700 Hz spectral width), and 0.0162 s for C1 and C2 (3950 Hz spectral width). Signals in the first proton dimension are frequency-shifted in order to center the amide region while leaving the transmitter
frequency on water (see Figure 28). Therefore, the indirect proton sweep width covers from water to the edge of the amide proton spectrum, and methyl protons are aliased. With a recycle delay of 1.4 s and 4 scans per FID, the total acquisition time for each 4-D dataset was 98 hours.

The first step in processing the 4-D TS NOESY datasets is to separate overlapping NOESY pathways (Frueh et al. 2006). Magnetization pathways detected on different proton types (e.g. amide-methyl and amide-amide NOESY) are readily separated by setting the sweep width of the directly detected dimension to cover both amide and methyl proton signals and then extracting the appropriate region when computing the Fourier transformation. However, pathways detected on the same proton type (e.g. amide-methyl and methyl-methyl) overlap due to the aliasing of methyl protons in the second indirect dimension. To overcome this issue, two interleaved datasets are collected for each set of frequency labeling delays, with magnetization pathways originating on methyl groups selectively inverted in the second. Addition and subtraction of the two datasets selects pathways originating solely from amide and methyl groups, respectively. The second processing step is to resolve the sensitivity enhanced signals recorded in the first two dimensions into quadrature components as described elsewhere (Xia et al. 2000).

Time domain datasets were then processed with either the fast Fourier transform (FT) alone, or the FT followed by artifact removal with SCRUB. Cosine-weighted
Voronoi volumes were used to determine the contribution of each sampling point in the FT. All spectra were calculated at 128 point resolution in each indirect dimension (Coggins et al. 2010; Coggins and Zhou 2006), and direct dimensions were zero-filled to give a final resolution of ~128 points after extracting the appropriate region of proton signals. SCRUB processing times for the amide-amide, amide-methyl, methyl-amide, and methyl-methyl spectra were 5.5 hours, 1 hour, 20 minutes, and 53 hours, respectively, for the Ssu72 dataset, and 10 hours, 30 minutes, 10 minutes, and 36 hours, respectively, for the HCAII dataset. A detailed description of the SCRUB algorithm is provided in Section 3.5.

6.6.3 Data Analysis

The 4-D TS NOESY spectra were analyzed with NMRView (Johnson 2004) and peaks were assigned manually without ambiguous assignments. Each peak was compared to the noise level at its resident HC2/HN2 point in either the FT- or SCRUB-processed spectra as described in Section 6.3, and peaks with intensities over 3 times the noise level were deemed identifiable. All peaks used in structure calculations were identifiable in the SCRUB-processed spectra. Each identifiable peak was then matched to its corresponding interproton distance using crystal reference structures 2ILI (1.05 Å resolution) and 3P9Y (2.1 Å resolution) for HCAII (Fisher et al. 2007) and Ssu72 (see
Section 2.2), respectively. Stereospecific assignments of leucine and valine methyls were used to improve the accuracy of distance matching, but were not used as restraints in structure calculations. Each distance was counted as observed if either corresponding NOE crosspeak ($\text{Hi} \rightarrow \text{Hj}$ or $\text{Hj} \rightarrow \text{Hi}$) was identifiable. Interproton distances involving methyl protons were calculated based on pseudoatom positions. Interproton distances involving amide groups without observable peaks in the $^1\text{H}/^{15}\text{N}$-HSQC-TROSY spectra of Ssu72 and HCAII were not included in the ‘all expected’ numbers. Although the Ssu72 sample used in this study is in the apo form, we chose to use the substrate-bound crystal structure (3P9Y) to calculate interproton distances rather than the available apo form crystal structure (3FDF), which is at much lower resolution (3.2 Å). While the backbone conformations of the two states are nearly identical (see Section 2.4.4), substantial discrepancies were observed for a number of methyl-methyl interproton distances due to different rotamer fitting for certain ILV sidechains. The higher resolution substrate-bound model should allow for better rotamer fitting for these sidechains and therefore more accurate interproton distances.

6.6.4 Structure Calculations

All structure determinations were performed with CYANA 3.0 (Herrmann et al. 2002) with five independent calculations for each setup. Dihedral angle restraints were
derived from backbone carbon chemical shifts using TALOS+ (Shen et al. 2009); 292 and 390 dihedral restraints were calculated for Ssu72 and HCA2, respectively. Upper distance limits were calculated by CYANA based on peak volumes with the average distance limit set to 4.5 Å and lower and upper cutoffs set to 2.0 Å and 7.0 Å, respectively. For calculations with manually-assigned peak lists, CYANA was run with default parameters and 20,000 torsion angle dynamics steps. 100 structures were calculated for each run with the 5 lowest energy structures represented in the final ensemble.

For calculations with automated assignment of peak lists, significant modifications to the default CYANA protocol were required to account for the sparse distance constraint information inherent to global fold calculations. CYANA uses seven cycles of combined automated NOE assignment and structure calculation, and in each cycle potential peak assignments are scored on several criteria, including how well they are fulfilled in the ensemble from the previous cycle. This structure-based probability is controlled by the ‘violation cutoff’ variable, which decreases from cycle to cycle. Due to the limited amount of distance constraint information available in selectively-protonated samples, global fold calculations converge relatively slowly over the seven cycles of the CYANA calculation, causing large numbers of potential assignments to be prematurely discarded based on the default violation cutoff values. To overcome this issue, the structure-based probabilities were relaxed (by setting the violation cutoffs to 10 times...
their default values in the noeassign.cya macro). A second criterion used by CYANA for judging peak assignment possibilities is network anchoring, which measures each potential peak assignment’s fit into the network formed by the assignment possibilities of all other peaks. In general, network anchoring probabilities were unexpectedly low in the global fold calculations, presumably due to the low redundancy and limited amount of available distance constraint information, which led to many peaks being left unassigned, particularly in the amide-amide datasets. Therefore, the weight of the network anchoring probability was decreased relative to the other assignment criteria (by setting the variable alignfactor to 10 in noeassign.cya), and the overall probability threshold for assignment acceptance was lowered (by setting the quality variable to 0.30 in noeassign.cya). The chemical shift tolerance for determining peak assignment possibilities was set to 0.02 ppm in the direct methyl dimension (HC2), and the tolerances in all other dimensions were scaled to this value based on digital resolutions, giving tolerances of 0.06 ppm for HN2, 0.08 ppm for HC1/HN1, 0.24 ppm for C1/C2, and 0.41 ppm for N1/N2. Simulated 3-D peak lists were generated by removing the first heteronuclear dimension (C1/N1) of the 4-D peak lists. 3-D peaks in the amide-amide and methyl-methyl NOESY spectra that overlap with the diagonal (and therefore have the diagonal assignment within the chemical shift tolerances) were discarded by CYANA. 100 structures were annealed per cycle using 20,000 torsion angle dynamics steps, and the 10 lowest energy structures were chosen for structure-based probability
scoring of assignment possibilities in the subsequent cycle. The final cycle calculated
100 structures with the 5 lowest energy structures represented in the final ensemble.
Assignment accuracy was judged by comparing peak lists outputted after the last cycle
of NOE assignment by CYANA (name-cycle7-ref.peaks) with the corresponding
manually-assigned lists.
References


Howe, K.J. (2002). RNA polymerase II conducts a symphony of pre-mRNA processing activities. Biochim Biophys Acta 1577, 308-324.


Wilcox, C. B., Rossettini, A., and Hanes, S. D. (2004). Genetic interactions with C-terminal domain (CTD) kinases and the CTD of RNA Pol II suggest a role for ESS1 in


Biography

Jonathan W. Werner-Allen was born on June 18, 1982 in Evanston, Illinois. After graduating from Rochester Adams High School in Rochester Hills, Michigan in 2000, he matriculated at Dartmouth College in Hanover, New Hampshire. In 2004, he earned his Bachelor of Arts degree in Biophysical Chemistry, with minors in Math and Music. In the fall of 2004, Jon entered the Duke University Biochemistry Department’s graduate program, where he was awarded the Kamin Fellowship, and later affiliated with the Structural Biology and Biophysics program. He joined Pei Zhou’s lab in the spring of 2005, and began research into the structure and kinetics of the Ssu72 phosphatase and new methods for fast NMR of large proteins. He has published four first author papers on his graduate work to date, and has presented at numerous conferences and forums, including an invited talk at the 2011 Keystone Symposium on the Frontiers of NMR in Biology.

A complete listing of his publications follows:


