The Pennsylvania State University

The Graduate School

Department of Chemistry

THE EFFECTS OF MULTI-SITE PHOSPHORYLATION ON THE STRUCTURE AND FUNCTION OF THE CARBOXYL-TERMINAL DOMAIN OF THE RNA POLYMERASE II LARGE SUBUNIT

A Dissertation in

Chemistry

by

Eric Bryant Gibbs

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2017

The dissertation of Eric Bryant Gibbs was reviewed and approved* by the following:

Scott A. Showalter
Associate Professor of Chemistry
Associate Professor of Biochemistry and Molecular Biology
Dissertation Advisor
Chair of Committee

Philip C. Bevilacqua
Professor of Chemistry
Professor of Biochemistry and Molecular Biology

David Gilmour
Professor of Molecular and Cellular Biology

William G. Noid
Associate Professor of Chemistry

Thomas E. Mallouk
Head of the Chemistry Department
Evan Pugh University Professor of Chemistry,
Biochemistry and Molecular Biology, Physics,
and Engineering Science and Mechanics

*Signatures are on file in the Graduate School

ABSTRACT

Intrinsically disordered proteins (IDPs) are broadly defined as protein regions that do not cooperatively fold into spatially or temporally stable structures. A growing body of research supports the hypothesis that structural disorder renders IDPs uniquely capable of regulating key biological processes such as cellular signaling and transcription. Yet, this conformational plasticity often precludes the characterization of IDPs by traditional structural biology techniques. Advances in NMR spectroscopy, mass spectrometry, and small angle X-ray scattering presented here have enabled rigorous mechanistic studies of disordered proteins and regions, as exemplified by our investigation into the effects of multi-site phosphorylation on the structure and function of the carboxyl-terminal domain of the RNA polymerase II large subunit (CTD). We identify phosphorylation sites in the *Drosophila melanogaster* CTD that are targeted by the Positive Transcription Elongation Factor b (DmP-TEFb). We show that phosphorylation occurs primarily at Ser5 residues and that Tyr1 is necessary this specificity. Importantly, we demonstrate that Ser5 phosphorylation induces highly sequence-specific conformational switches in the CTD, which tune the apparent activity of CTD-interacting factors, using the CTD phosphatase Ssu72-Symplekin as an example. These studies highlight how regulation of IDPs can be mediated through cryptic sequence features and establish a foundation for elucidating the molecular basis of CTD regulation.

TABLE OF CONTENTS

List of Figures vi

List of Tables ix

Acknowledgements x

Chapter 1: Quantitative Biophysical Characterization of Intrinsically Disordered Proteins 1
- Introduction 2
- Defining Structure Amid Disorder 5
  - Nuclear Magnetic Resonance Spectroscopy (NMR) 6
  - Small-angle Scattering Methods 10
  - Single Molecule Förster Resonance Energy Transfer (FRET) 11
- Generation of IDP Structural Ensembles 13
- Thermodynamic Analysis of IDP Interactions 17
  - Isothermal Titration Calorimetry (ITC) 18
- The Kinetics of IDP Interactions 24
  - Resolving Induced Fit from Conformational Selection 25
  - Transition States 29
- Conclusion 32
- References 33

Chapter 2: Application of NMR to Studies of Intrinsically Disordered Proteins 44
- Introduction 45
- NMR chemical shift assignment strategies for IDPs 48
  - Innovations in proton-detected NMR 48
  - Heteronuclear direct-detection NMR 50
- NMR structural and dynamic constraints for IDPs 52
  - Structural constraints 53
  - Modelling IDP dynamics 54
- Monitoring post-translational modifications by NMR 56
  - Phosphorylation 57
  - Acetylation 59
- NMR studies of aqueous phase separation 60
  - Characterization of Liquid-Liquid Phase Separation (LLPS) 61
  - Hydrogel formation 63
- Applications of NMR to aggregation-prone IDPs and their amyloid fibers 64
- In cell NMR of IDPs 65
- Conclusions and prospectus 67
- References 69

Chapter 3: A NMR and MS Based Approach to Phosphorylation Site Identification for the RNA Pol II CTD 76
- Introduction 77
- Results & Discussion 78
  - Phosphorylation Site Identification of CTD2′ by LC-ESI-MS/MS: 80
  - Peptide Scoring Algorithms 82

Phosphorylation Site Identification of CTD2′ by NMR Spectroscopy 84
Cross-Validation 88
Conclusion 94
Materials & Methods 95
References 103

Chapter 4: DmP-TEFb Specificity by Probed by MALDI MS 105
Introduction 105
Results & Discussion 106
Conclusion 120
Materials & Methods 123
References 124

Chapter 5: Phosphorylation Induces Sequence-Specific Conformational Switches in the RNA Polymerase II C-Terminal Domain 125
Introduction 126
Results & Discussion 128
Transgenic Drosophila Reveal a Developmentally Necessary CTD Region 128
Hyper-Phosphorylation of Recombinant CTD2′ by P-TEFb 130
Hyper-Phosphorylation does not Alter the Global Dimensions of CTD2′ 135
Hyper-Phosphorylation of CTD2′ Induces Sequence-Dependent Proline Isomerization 136
Sequence-Dependent Proline Isomerization Modulates Apparent Ssu72 Activity 140
Conclusion 144
Materials & Methods 147
Supporting Information 157
References 166

Chapter 6: Preliminary Data for the CTD (aa1815-1887) 169
Introduction 169
Results & Discussion 169
Conclusion 175
Materials & Methods 175
References 178

Appendix A: Abbreviations 179

Appendix B: Theoretical and Observed MS2 Ions – Chapter 3 180

Appendix C: Theoretical and Observed MS2 Ions – Chapter 4 182

LIST OF FIGURES

Figure 1-1. A generalized free energy landscape to describe IDPs 4

Figure 1-2. Comparison of ^{1}H,^{15}N-HSQC and ^{13}C,^{15}N-CON spectra for IDPs 8

Figure 1-3. Ensemble models deposited in the Protein Ensemble Database (pE-DB) 15

Figure 1-4. The free energy surface of a “fuzzy complex”...20

Figure 1-5. Proposed IDP binding mechanisms...25

Figure 1-6. A generalized reaction scheme for IDP kinetic flux pathways 28

Figure 1-7. Phi value analysis for IDP folding upon binding 30

Figure 2-1. Mapping the phase diagram of intrinsic disorder...47

Figure 2-2. Residue-specific resolution is achievable in high-resolution spectra of intrinsically disordered regions 49

Figure 3-1. Phosphorylation site assignment of CTD2′ by LC-ESI-MS2 and NMR spectroscopy...79

Figure 3-2. Localization of phosphorylation sites using MS2 peptide spectra 81

Figure 3-3. Backbone resonance assignment of CTD2′ by ^{13}C Direct-Detect NMR..............85

Figure 3-4. Detection and assignment of phosphorylated CTD2′ by NMR.............................87

Figure 3-5. NMR chemical shift perturbations observed for CTD2′ following phosphorylation by DmP-TEFb 89

Figure 3-6. Assignment of pS1675 by NMR spectroscopy and site-directed mutagenesis..90

Figure 3-7. NMR assigned DmP-TEFb phosphorylation sites in CTD2′ verified by site-directed mutagenesis 91

Figure 3-8. Top-down sequencing of unphosphorylated and DmP-TEFb treated CTD2′ by MALDI-ISD. 93

Figure 4-1. DmP-TEFb activity and specificity probed by MALDI-MS2..............................107

Figure 4-2. MALDI-TOF-MS spectra of DmP-TEFb treated CTD peptides 108

Figure 4-3. MALDI-MS2 spectra of DmP-TEFb treated CTD peptides.................................109

Figure 4-4. MALDI-TOF-MS spectra of DmP-TEFb treated mutant CTD peptides 110

Figure 4-5. MALDI-MS2 spectra of DmP-TEFb treated mutant CTD peptides 111

Figure 4-6. Active sites of various CDKs 113

Figure 4-7. MALDI-TOF-MS spectra of DmP-TEFb treated Tyr mutant CTD peptides........115

Figure 4-8. MALDI-MS2 spectra of DmP-TEFb treated Tyr mutant CTD peptides 116

Figure 4-9. MALDI-MS2 spectra of a DmP-TEFb treated CTD peptide lacking Tyr1...........117

Figure 4-10. Multi sequence alignment of various CDKs. 119

Figure 4-11. Alternative binding modes and assembly states in CDK-cyclin complexes 122

Figure 5-1. A highly conserved region in the *Drosophila* CTD is essential for viability and is targeted for Ser5 phosphorylation by DmP-TEFb *in vitro* 129

Figure 5-2. ΔCTD2′ Ser5 phosphorylation probed by MS and NMR spectroscopy...............131

Figure 5-3. Phospho-sites and proline isomerization in hyper-pSer5 CTD2′ probed by NMR spectroscopy...134

Figure 5-4. Small angle X-ray scattering reveals no significant change in pair-wise distances within CTD2′ upon extensive serine 5 phosphorylation..................................136

Figure 5-5. Structural characterization of unphosphorylated and phosphorylated CTD2′ by NMR spectroscopy...138

Figure 5-6. The impact of serine 5 phosphorylation on the structure of the DmCTD 139

Figure 5-7. Structural switches in hyper-pSer5 CTD2′ modulate the apparent Ssu72 activity..141

Figure 5-8. The conserved conformation of CTD peptides recognized by Ssu72. 143

Figure 5-S1. Identification of a region of the Rpb1 CTD that is essential for rescuing lethality caused by RNAi-mediated depletion of endogenous Rpb1 157

Figure 5-S2. Kinetics of DmP-TEFb phosphorylation of CTD2′ measured by RT-NMR 158

Figure 5-S3. Small angle X-ray scattering reveals no significant changes in R_g for CTD2′upon serine 5 phosphorylation 159

Figure 5-S4. Phosphoserine pKa values determined by ^{31}P NMR spectroscopy.....................160

Figure 5-S5. Random coil conformations of CTD2′ heptads separate consecutive Ser5s or pSer5s by distances that exceed the Debye length...161

Figure 5-S6. Structural characterization of unphosphorylated CTD2′ by NMR spectroscopy...162

Figure 5-S7. Proline *cis-trans* isomerization in Hyper-pSer5 CTD2′ probed by NMR spectroscopy....................163

Figure 5-S8. Kinetics of Ssu72-Symplekin dephosphorylation of Hyper-pSer5 CTD2′.........163

Figure 6-1. 12% SDS-PAGE displaying the products of *in vitro* kinase treatments 170

Figure 6-2. Analysis of CTD_end phosphoforms by SDS-PAGE and LC-ESI-MS................171

Figure 6-3. Analysis of MAPK treated CTD_end by LC-ESI-MS2 using ETD activation....................172

Figure 6-4. Analysis of the proteolytic digests of MAPK phosphorylated CTD_end by SDS-PAGE....................173

Figure 6-5. Analysis of the unphosphorylated and DmP-TEFb treated CTD_end by NMR spectroscopy....................174

LIST OF TABLES

Table 1-1. Thermodynamics of IDP interactions with folded partners measured by ITC 23

Table 4-1. Model CTD peptides for analysis by MALDI-MS 107

Table 5-S1. Kinetic parameters for DmP-TEFb phosphorylation of CTD2′ 170

Table 5-S2. SAXS data collection and scattering derived parameters for unphosphorylated and hyper-pSer5 CTD2′...170

Table 5-S3. Kinetic parameters for Ssu72-Symplekin dephosphorylation of Hyper-pSer5 CTD2′...171

Table A-1. MALDI-ISD ion series observed for unphosphorylated CTD2′...........................175

Table A-2. MALDI-ISD ion series observed for DmP-TEFb phosphorylated CTD2′...........176

Table B-1. MS2 ion series for WT: precursor @ 1537 m/z (pSer2 State)...............................177

Table B-2. MS2 ion series for WT: precursor @ 1537 m/z (pSer5 State)...............................178

Table B-3. MS2 ion series for WT: precursor @ 1617 m/z (pSer2/pSer5 State).....................179

Table B-4. MS2 ion series for S2A: precursor @ 1521 m/z (pSer5 State)..............................180

Table B-5. MS2 ion series for S5A: precursor @ 1521 m/z (Internal pSer2 State).................181

Table B-6. MS2 ion series for S5A: precursor @ 1521 m/z (+1 heptad pSer2 State) 182

Table B-7. MS2 ion series for Y1A: precursor @ 1445 m/z (pSer2 State) 183

Table B-8. MS2 ion series for Y1A: precursor @ 1445 m/z (pSer5 State) 184

Table B-9. MS2 ion series for Y2A: precursor @ 1445 m/z (pSer2 State) 185

Table B-10. MS2 ion series for Y2A: precursor @ 1445 m/z (pSer5 State) 186

Table B-11. MS2 ion series for Y12A: precursor @ 1353 m/z (pSer2 State) 187

Table B-12. MS2 ion series for Y12A: precursor @ 1353 m/z (pSer5 State) 188

ACKNOWLEDGEMENTS

This dissertation is dedicated to my family, mentors, colleagues & friends.

Chapter 1

Quantitative Biophysical Characterization of Intrinsically Disordered Proteins

[This chapter was modified from a published manuscript: E. B. Gibbs, S. A. Showalter (2015) "Quantitative biophysical characterization of intrinsically disordered proteins" Biochemistry 54. 1314-26.]

Intrinsically disordered proteins (IDPs) are broadly defined as protein regions that do not cooperatively fold into a spatially or temporally stable structure. Recent research strongly supports the hypothesis that a conserved functional role for structural disorder renders IDPs uniquely capable of functioning in biological processes such as cellular signaling and transcription. Recently, the application of rigorous mechanistic biochemistry and quantitative biophysics to disordered systems has increased dramatically. For example, the launch of the Protein Ensemble Database (pE-DB) demonstrates that the potential now exists to refine models for the native state structure of IDPs using experimental data. However, rigorous assessment of which observables place the strongest and least biased constraints on those ensembles is now needed. Most importantly, the past few years have seen strong growth in the number of biochemical and biophysical studies attempting to connect structural disorder with function. From the perspective of equilibrium thermodynamics, there is a clear need to assess the relative significance of hydrophobic vs. electrostatic forces in IDP interactions, if it is possible to generalize at all. Finally, kinetic mechanisms that invoke conformational selection and/or induced fit are often used to characterize coupled IDP folding and binding, although application of these models is

typically built upon thermodynamic observations. Recently, the reaction rates and kinetic mechanisms of more intrinsically disordered systems have been tested through rigorous kinetic experiments. Motivated by these exciting advances, here we provide review and prospectus for the quantitative study of IDP structure, thermodynamics, and kinetics.

Introduction

The folding funnel hypothesis established a new perspective on the widely accepted view that the biological function of a protein is determined by its "native state." Pioneering investigations of heme protein dynamics and function revealed many decades ago that the native state of proteins must be characterized by multiple conformational substates in order to accommodate observed functions[1, 2]. In the modern view, we recognize that the natively-folded state exists as an ensemble of conformations sampled from an energy landscape where dynamic fluctuations between closely related conformers facilitate catalysis, macromolecular association, and other biological functions of cooperatively folded proteins.[3] Intrinsically disordered proteins (IDPs) represent the most extreme examples of this ensemble view, because they lack cooperatively-folded structure under native conditions and are best described by highly dynamic and heterogeneous conformational ensembles, yet they retain function. This leads to a broader paradigm, affirming that the native state determines biological function, regardless of whether folding occurs. Our awareness that function can arise from native protein disorder suggests a pressing need for quantitative biochemical and biophysical characterization of the mechanisms linking structure and function in this exciting class of proteins.

The relatively recent expansion of interest in disordered proteins among biochemists contrasts with their high prevalence in nature, particularly in eukaryotes. Within the human proteome, ~50% of all proteins are predicted to contain long disordered segments (≥30 residues)[4, 5] with enrichment to as much as 70% of all polypeptide sequence among transcription factors and signaling proteins.[6] In this context, structural plasticity facilitates multiple protein-protein interactions, placing IDPs at the "nodes" of large interaction networks.[7] Upon binding, IDPs often experience disorder-to-order transitions that tend toward desolvation and burial of a larger surface area than an equivalent cooperatively folded protein would bury, while requiring shorter amino acid sequences to do so.[8] These properties support the hypothesis that retaining disorder is an evolutionary strategy that facilitates complex function within a compact genome.[9] In contrast, while disorder may support complex function, it can also promote complex pathology.[10] In cells, the abundance and turnover of IDPs are under tight control,[11] and aberrant IDP regulation has been implicated in cancer pathways and neurodegenerative disease,[10] completing the rationale for seeking to understand the molecular properties of disordered proteins.

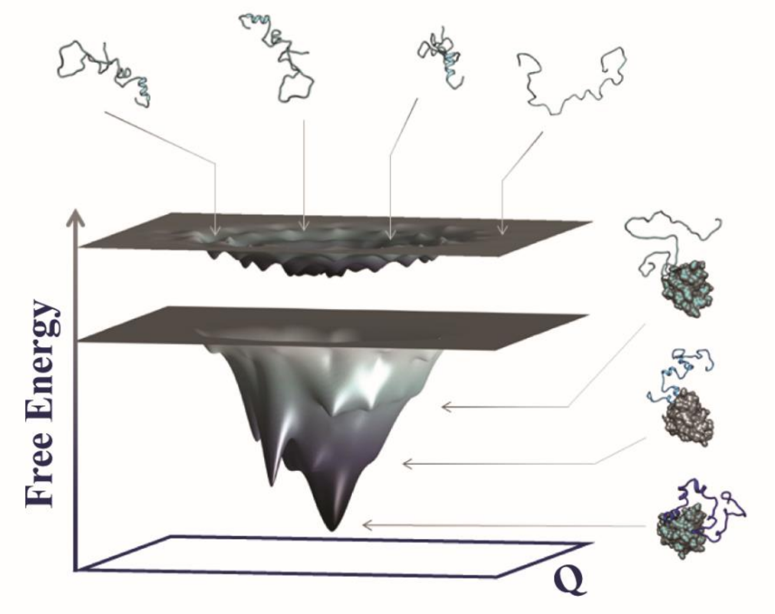


Figure 1-1. The free energy landscape hypothesis can be generalized to describe intrinsically disordered proteins. The free energy surface for an IDP in the unbound state can be characterized by a rugged, but relatively flat landscape (top surface); while IDP binding events funnel the free energy landscape, yielding a well-defined minimum (bottom surface).

For the reasons reviewed above, there is extraordinary current interest in elucidating the unique physicochemical properties and biological functions of IDPs. Presumably, unbound IDPs sample a diverse ensemble of conformations along a rugged free energy surface that lacks a significant bias towards any particular equilibrium structure (Figure 1-1, top surface).[4, 12] Conversely, IDP *interactions* appear to be governed by a funneled free energy surface that guides them toward a bound, folded structure (Figure 1-1, bottom surface), drawing strong parallels with the current view describing cooperatively folding proteins.[13, 14] Despite the appeal of mechanistic proposals built from these broad observations, remarkably few experimental studies have quantified the ensembles IDPs adopt in solution, the thermodynamics that govern their interactions, or the kinetics that describe their transitions between biologically relevant states. Expanding our understanding of the native disordered state will require quantitative descriptions of IDP conformational ensembles and interaction mechanisms, both of which are fertile grounds for modern biophysics. In the following, we review recent, ever intensifying efforts toward filling this gap and conclude with our perspectives on the future of reconciling pervasive conformational disorder with quantitative mechanistic insight for intrinsically disordered systems.

Defining Structure Amid Disorder

Certain 3D folds are repeated in proteins with specific functions; similarly, some biological functions are optimally performed by IDPs.[15-20] The unique functional properties of IDPs present a compelling reason for their study and suggest there is insight encoded in their structural properties, just as is the case for cooperatively folding domains.[6] Rigorous testing of this hypothesis requires that quantitative structural studies of IDPs be pursued. Of special interest, many recent studies have investigated the remarkable disorder-to-order transition that couples IDP folding to ligand binding,[21-27] suggesting a further biological rationale for IDP structure assessment. Among the many structure assessment tools available to the protein chemistry community, NMR, small angle x-ray scattering (SAXS), and single-molecule Förster resonance energy transfer (smFRET) have emerged as the clear leaders for IDP applications. We recognize the successful application of these methods to studies of pathogenic IDPs, but for the sake of brevity we have focused our discussion to IDPs that support regulation of normal cellular function; we direct interested readers to several in depth reviews of pathological IDP misfolding and aggregation.[28-30] Finally, we note that a variety of purely computational approaches to this problem have also generated significant advances in our field,[31-34] but we have elected to focus solely on experimental techniques for the sake of constraining the scope of this review. There is one necessary exception to this decision; computational models are required to evaluate the meaning of averaged experimental observables, due to the ensemble nature of IDP structure.[35] Therefore, a brief discussion of computational methods for IDP conformer generation will close this section.

Nuclear Magnetic Resonance Spectroscopy (NMR)

Solution NMR spectroscopy is by far the most widely applied method for studying the dynamic structural ensembles of IDPs.[36, 37] Despite the enormous potential seen in the examples we will review here, the general study of IDPs by NMR remains limited by the extremely poor ^{1}H-amide chemical shift dispersion typically observed in their spectra. To date, this has represented the most substantial barrier to broad applications in this field, because investigators are limited to working with those few IDPs whose spectra do show sufficient peak dispersion. Clearly, more systematically successful methods for NMR applications to IDPs are needed, given the biological imperative to better understand IDP structure-function relationships.

Within the protein NMR community, the ^{1}H,^{15}N-heteronuclear single quantum correlation (HSQC) experiment has emerged as the detection platform of choice for routine applications. The principal advantages to choosing the ^{1}H,^{15}N-HSQC are the low costs of sample preparation and this spectrum's simple structure, with one 2D-resonance per amino acid residue, except for proline. In principle, protein NMR strategies could be built around any 2D spectrum meeting these minimal requirements, as evidenced by the adoption of transverse relaxation optimized (TROSY) methods for large systems. Recently, ^{13}C-direct detection spectroscopy has re-emerged as a viable tool for studying proteins in solution.[38] For IDPs, the ^{15}N,^{13}C-CON spectrum, correlating the ^{13}C-carbonyl with the ^{15}N-amide of each peptide plane, is an especially effective choice of 2D-platform.[39-44] As illustrated for the disordered C-terminal tail of the phosphatase FCP1 in Figure 1-2A and B, the overwhelming peak overlap observed in the ^{1}H,^{15}N-HSQC is almost completely relieved through ^{15}N,^{13}C-CON detection. Perhaps even more significantly for IDPs, which tend to be enriched 1.7-1.8-fold in proline compared to cooperatively folding proteins,[45] the ^{15}N,^{13}C-CON spectrum contains a resonance for each peptide bond involving a proline nitrogen. The significance of this advantage is clearly seen for the C-terminal tail of the transcription factor Pdx1 (Figure 2C, D), for which 21% of the residues in the construct studied in our laboratory are prolines.[43]

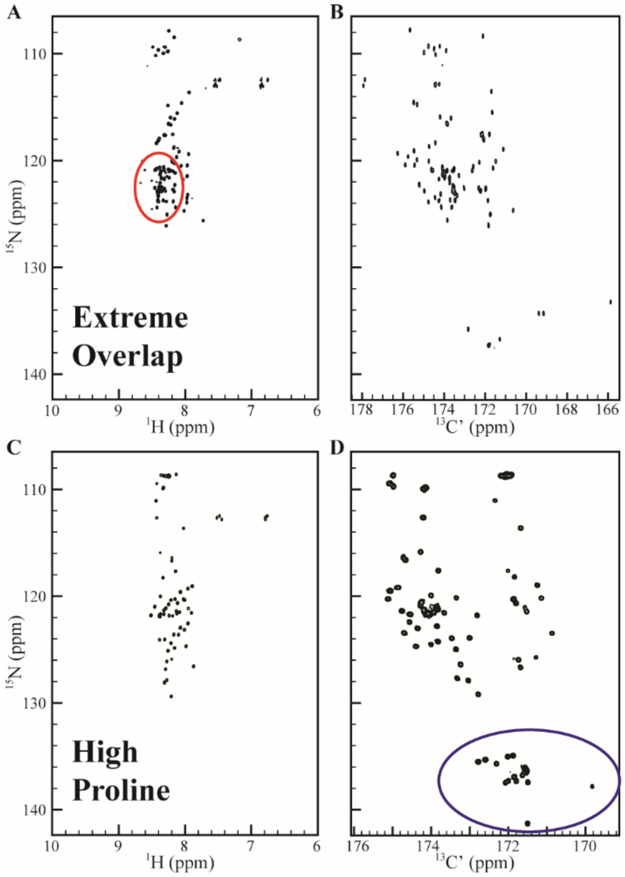


Figure 1-2. ^{1}H,^{15}N-HSQC spectra of IDPs generally suffer from poor chemical shift dispersion, which is typically relieved in the ^{15}N,^{13}C-CON spectrum, as demonstrated for the C-terminus of FCP1 (A and B, respectively). Additionally, the C-terminus of Pdx-1 (C,D) displays the power of the ^{15}N,^{13}C-CON for proline-rich disordered proteins.

The number of NMR studies published for IDPs has grown tremendously in recent years.[46] Intriguingly, chemical shifts, which represent one of the most direct NMR observables, have emerged as one of the most effective of the sparse structure constraints available for the refinement of disordered protein ensembles.[47] This is largely true because chemical shifts are indispensable probes for local secondary structure.[48] More importantly, regions of (partially ordered) secondary structure in the native IDP ensemble are strong candidates for establishing connections between the fine details of IDP structure and their biological functions. The binding elements of IDPs that undergo disorder-to-order transitions upon forming interactions are often partially or completely pre-formed in the apo-state, leading to their categorization as molecular recognition fragments (MoRFs).[49-51] When validated, these non-random structural features lend direct support to the hypothesis that native disorder provides a molecular pathway to biological fitness.

While the chemical shifts of backbone nuclei are powerful constraints on secondary structure, they are generally not sufficient to fully constrain the ensemble in the absence of other data. Measurement of residual dipolar couplings (RDCs) provides an effective means of complementing chemical shifts, particularly for the characterization of MoRFs. For example, the C-terminal region of the Sendai virus nucleoprotein was demonstrated to be α-helical through systematic RDC measurement.[52] More importantly, RDCs offer orthogonal constraints to chemical shifts because they are sensitive to long range interactions.[47] On a cautionary note, dipolar couplings are sensitive to dynamic averaging, up to the millisecond timescale, and therefore are subject to limitations when employed in isolation as a long range structural constraint for highly flexible systems.

One effective means to complement RDC constraints on global order in IDP ensembles is to measure the paramagnetic relaxation enhancement (PRE) generated through deliberate attachment of paramagnetic species to the polypeptide chain. PRE measurements are powerful reporters of transient tertiary interactions within the ensemble.[53-56] While incorporating PRE constraints is relatively straightforward for cooperatively folding domains, caution is needed for IDP applications because backbone and probe dynamics both heavily influence the intensity of induced effects, causing complications. Also, as discussed for RDC measurements, the convolution of effects from dynamics and low population conformational states can cause interpretation of PRE data to be ambiguous. Fortunately, certainty in the interpretation of PRE data can be improved through concurrent analysis of backbone spin relaxation, which helps to define the amplitude of local conformational dynamics in the ensemble.

Small-angle Scattering Methods

Small-angle X-ray scattering (SAXS) provides low resolution structural information for biomolecules in solution, which is often used synergistically with NMR constraints.[57-61] This is demonstrated by three remarkable examples from the intrinsically disordered segments in the Sendai virus phosphoprotein,[62] the tumor suppressor p53,[63] and the cyclin dependent kinase inhibitor Sic1.[64] SAXS curves directly provide information about the oligomeric state, size, and overall shape of a molecule. Using the Guinier approximation, the radius of gyration (Rg) can be extracted from SAXS curves,[58] while conformational flexibility can be assessed both qualitatively and quantitatively through Kratky analysis and application of the Porod-Debye law.[65] SAXS data can also be used as input for various computational procedures in order to quantitatively describe the distribution of conformers within the IDP's structural ensemble,[58] yielding valuable information about the 3-dimensional shape,[66] and the global level of compaction present.[67]

In addition, small-angle neutron scattering (SANS) has emerged as a complimentary method for the structural characterization of IDPs.[68-70] SANS is similar to SAXS in that it provides information about the overall shape of a molecule in solution. When combined with contrast variation (CV) or selective deuteration, the benefits for disordered proteins, especially those in complex, can be clearly seen. For example, CV-SANS has been used to show that the histone tails within the canonical H2A nucleosome intertwine with the surrounding DNA, while those within the H2A.B variant, which features outstretched DNA, turn inward toward the core, thus rationalizing the differences in the observed stabilities of the two forms.[71] Future studies may exploit CV-SANS to obtain useful distance restraints as input for ensemble generation of IDPs in the bound state. This would be especially advantageous for systems that are large by the standards of NMR spectroscopy or for highly dynamic complexes that elude detection, due to chemical exchange on the timescale of chemical shift evolution.

Single Molecule Förster Resonance Energy Transfer (FRET)

The final experimental technique reviewed here is single molecule FRET, which has also contributed substantially to our understanding of the conformational landscapes of disordered proteins.[72-74] Several notable studies have used smFRET derived distances to explore how the unique physicochemical properties of IDPs affect their dimensions in solution. For example, it was found that despite convergence in the dimensions of both IDPs and cooperatively folded proteins at high GdmCl concentrations, the polymer scaling laws of IDPs diverged significantly from folded proteins when the denaturant concentration was reduced.[75] The clearest interpretation of these results is that, under native conditions, IDPs adopt fundamentally different ensembles from the unfolded state of cooperatively folding proteins, due to differences in solvation of the relatively hydrophilic and charged amino acid sequences found in natively disordered proteins.[75] The prior result was achieved using the charged denaturant guanidinium-HCl; when the neutral denaturant urea is used instead, this result is not reproduced, highlighting the importance of repulsive electrostatic forces for establishing the dimensions of IDP ensembles and confirming predictions from polyampholyte theory.[76] For Sic1, polyampholyte behavior manifest in a similar denaturant/ionic strength dependence on scaling. Interestingly, investigation of Sic1 revealed three distinct subpopulations, each of which exhibited different sensitivities to electrostatic screening, thus highlighting the utility of single molecule techniques for resolving conformational subpopulations otherwise obscured by ensemble averaged measurements.[77] In addition, intramolecular distances derived from smFRET have been used to elucidate structure function relationships for IDPs. For example, compaction of the N-terminus of PAGE4, a stress-response protein overexpressed in prostate cancer cells, was observed following phosphorylation by HIPK1. This modification weakened PAGE4 affinity for c-Jun, suggesting a possible regulatory role for expression of c-Jun target genes.[78] As a final note, smFRET-based intramolecular distances have not yet been broadly applied in ensemble modeling schemes, but recent demonstration of their utility for modeling flexible single-stranded DNA[79] suggests smFRET distances will prove increasingly valuable for IDP ensemble generation in the future.

Generation of IDP Structural Ensembles

Spectroscopic data suggest that IDPs may adopt anything from a heterogeneous, but relatively compact, ensemble of structures to a denatured state highly enriched in native-like secondary structural features, reminiscent of molten globules.[80] Physically, IDPs are best described by ensemble states where the protein is able to interconvert, on some timescale(s), between multiple conformations on a rugged potential energy landscape (Figure 1). To rigorously assess the relationship between this landscape and function requires quantitative evaluation of the IDP conformers generated through application of experimental data as constraints. Particularly for the case of molecular recognition fragments, which are often enriched in secondary structure, examples now abound lending support to the hypothesis that the details of protein structure offer insights into function and that this is especially true for IDPs, for which our mechanistic understanding is still largely incomplete.

Since its inception in 1971, the protein data bank (www.pdb.org)[81] has provided an irreplaceable resource for the research community, aiming to make the products of structural biology freely accessible to all potential beneficiaries. Unfortunately for those who study disordered proteins, the conformational ensembles we generate are very often inappropriate for inclusion in the PDB. While it is now routine to refine IDP ensembles capable of reproducing average structural observables (e.g., NMR chemical shifts, SAXS radius of gyration), it generally is not the case that the individual conformers contributing to the structure sets are themselves unique. On the other hand, the sets of conformers generated in these efforts often lead to unique and important hypotheses, as was the case for Sic1,[64] creating an imperative to distribute the conformer sets generated both broadly and freely. Recently, a consortium of investigators has launched the Protein Ensemble Database, which seeks to facilitate biological data-mining and structural methods development through the broad distribution of structural coordinates and primary data for disordered protein ensembles.[82]

At the time this manuscript was published, there were 16 entries in the pE-DB, each of which is presented graphically in Figure 1-3. As can be seen, a holistic view of the deposited IDP ensembles often reveals unique structural properties, despite the often degenerate nature of individual conformers. For example, the bent pSic1 ensemble promotes ultra-sensitive binding to CDC4 (Figure 1-3A), whereas the rod-like Sic1 ensemble (Figure 1-3B) does not. Finally, the entries highlight the diversity of experimental techniques currently applied to ensemble generation; for example p15PAF (Figure 1-3D), generated from NMR & SAXS or the unbound p27 KID domain (Figure 1-3N), generated through molecular dynamics (MD) simulation. Although the pE-DB is still in its infancy, its establishment sets an important milestone for the study structural disorder.

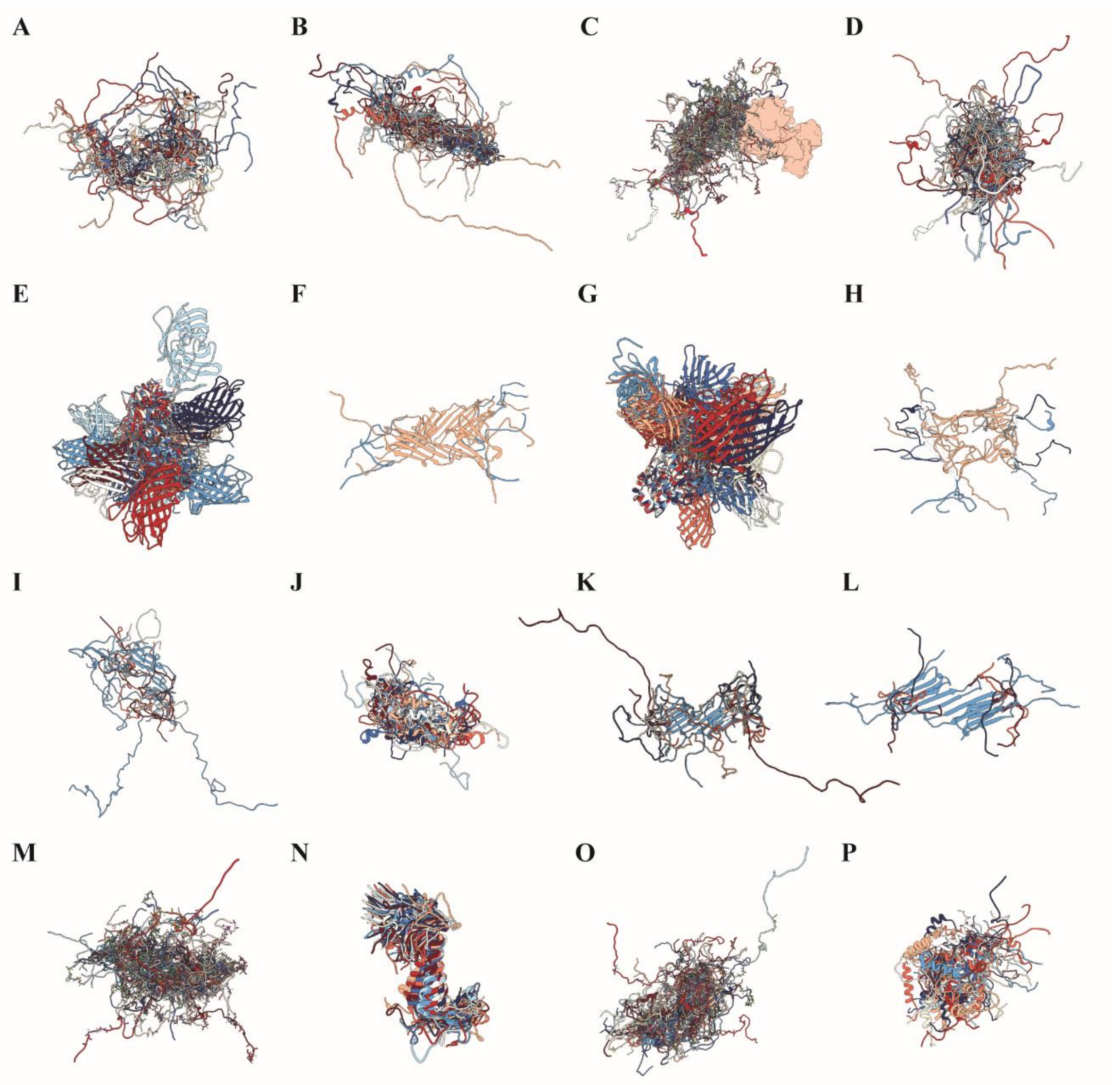


Figure 1-3. Ensemble models deposited in the Protein Ensemble Database (pE-DB). (A) 1AAA, (B) 9AAA, (C) 5AAC, (D) 6AAA, (E) 3AAA, (F) 8AAA, (G) 4AAA, (H) 3AAB, (I) 7AAA, (J) 8AAC, (K) 2AAB, (L) 1AAB, (M) 6AAC, (N) 2AAA, (O) 7AAC, (P) 5AAA.

Ensemble generation can be broadly separated into *de novo* strategies (purely computational) or strategies built upon experimental constraints. Most experimental approaches rely on generating a large pool of starting structures, from which experimental parameters are back calculated and compared to experimental data, resulting in conformer rejection or refinement based on some statistic, for example minimization of χ^2. Due to the sparse nature of ensemble averaged data, IDP ensembles can be sensitive to the structures used as input. Accordingly, starting pool generation has been accomplished by various techniques. Molecular dynamics simulation provides an attractive option for input conformer generation.[83, 84] However, there is still reason to be concerned that classical MD is not well suited for rigorous sampling of IDP potential energy surfaces and extremely long trajectories are likely required to reach convergence, if possible.[85] In this regard, advanced sampling techniques including replica exchange molecular dynamics (REMD)[33, 86] and accelerated molecular dynamics (AMD)[63, 87] may be advantageous. Monte Carlo (MC) simulations also provide an attractive option for generating large conformational ensembles.[88] For example, the ABSINTH implicit solvent force field, developed specifically for MC simulation of IDPs,[89] efficiently and accurately samples the conformational space spanned by highly disordered proteins.

The diversity of software packages available to implement experimentally constrained ensemble refinement strategies has also grown in recent years. For example, the Flexible-Meccano algorithm, and its more recent evolution into the ASTEROIDS package for ensemble refinement, has proven highly successful for modeling IDPs.[47] Although not formally intended for NMR applications during its initial development, the

TRaDES package provides an alternative for generating initial trial sets of structures.[90] In this capacity, TRaDES has been bundled with a set of algorithms collectively named ENSEMBLE that has also proven highly effective for IDP ensemble generation and refinement.[91] Finally, ensemble generation using Bayesian Statistics also provides an estimate of the uncertainty in the weights of each conformer.[86] Regardless of which method for ensemble refinement is applied, it is clear that our capacity to generate IDP ensembles and use them for rigorous hypothesis testing has rapidly matured. As their use grows and rigorous validation methods are established for quality control, there is reason to hope that quantitative structural biology of intrinsically disordered systems will become an expected tool in comprehensive biochemical investigations, just as it has for highly ordered systems.

Thermodynamic Analysis of IDP Interactions

Structural and dynamic descriptions of disordered states and bound complexes are necessary, but not sufficient, to understand the functional behavior of IDPs. Accordingly, there is considerable interest in quantifying the thermodynamic forces that govern IDP interactions. For example, NMR titrations have been employed to extract residue specific equilibrium dissociation constants (K_d) when multiple binding sites on an IDP are present.[92-94] In addition, various florescence based techniques, including tryptophan quenching,[95] fluorescence anisotropy,[94, 96] smFRET,[7] and fluorescence correlation spectroscopy (FCS)[97-99] have been applied to study IDP interactions. Fluorescence based methods are useful for accurately measuring the K_d when binding is extremely tight[7] or when one or more of the interacting species is prone to aggregation.[7, 94] Importantly, fluorescence based techniques are not restricted by any apparent limitation on the size of the IDP under investigation.[98] In a recent example, smFRET was used to study the interactions between the adenovirus E1A oncoprotein, an intrinsically disordered hub, and two of its partners, the Taz2 domain CBP and pRb.[7] This study demonstrated that either positive or negative cooperatively can be selected for, depending on the availability of E1A binding sites. Thus, E1A demonstrates that allosteric modulation of a disordered protein hub, through binary or ternary interactions, can influence population distributions and functional outcomes.

Isothermal Titration Calorimetry (ITC)

When applicable, isothermal titration calorimetry (ITC) is by far one of the most robust methods available for thermodynamic studies.[100] ITC provides a direct measurement of the binding enthalpy (ΔH) and equilibrium association constant (K_a), enabling the calculation of the Gibbs free energy (ΔG) through the known temperature and parametric determination of the binding entropy (ΔS). Significantly, when ΔH is determined by ITC in a temperature series, the constant-pressure heat capacity change (ΔCp) associated with the binding process is also experimentally accessible.

For many IDPs, binding is accompanied by a disorder-to-order transition, leading to the hypothesis that the unfavorable entropy loss incurred by conformational restriction of the IDP in the bound state must be offset by a favorable gain in enthalpy. It has been proposed that this entropy penalty may be mitigated in “fuzzy” IDP complexes, which retain some extent of disorder in the bound state (Figure 1-4).[101] Although the potential functional advantages available to fuzzy complexes are intriguing, the broader picture implicit in the hypothesis is that entropy losses incurred upon IDP binding promote reversibility. While it seems likely that the conformational entropy of the IDP chain will decrease upon binding, this point of view does not account for the role of solvent in mediating IDP association with binding partners. In other words, as with many other favorable protein folding phenomenon, a favorable change in solvent entropy near room temperature is likely to overwhelm the loss in chain entropy, making the functional value of any marginal stabilization brought about by "fuzziness" unclear. Fortunately, the detailed thermodynamic information conveyed through temperature-dependent ITC measurements provides exactly the experimental data needed to rigorously evaluate the energetics of coupled IDP folding and binding, even if the entropy estimates generated are indirect.

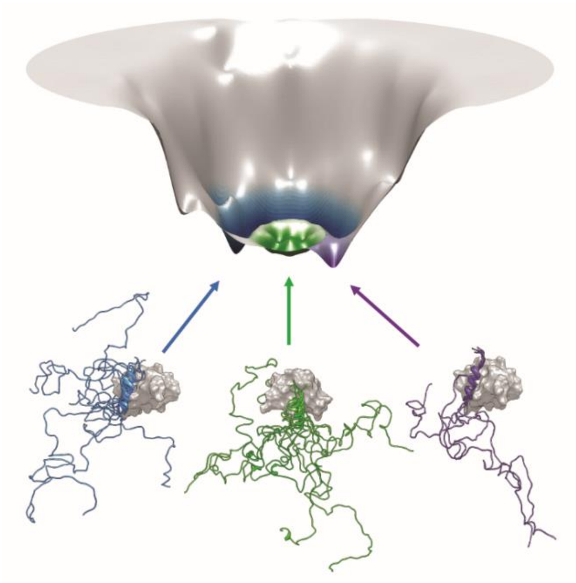


Figure 1-4. The free energy surface of a "fuzzy complex" is depicted as a funneled, yet wide free energy surface, where multiple distinct IDP conformations are sampled in the bound state.

Structural analysis of bound-IDPs has revealed significant hydrophobic character at many IDP binding interfaces, suggesting that hydrophobic forces play an important role in coupled folding and binding.[8, 102] One striking example is the Gcn4:Gal11/Med14 interaction, which is stabilized exclusively by three hydrophobic contacts.[103] Variable temperature ITC performed on this system revealed the signatures of apolar desolvation through strong temperature dependence in the observed binding enthalpy, resulting in a large negative heat capacity.[103] Far from being unique, this trend is also seen in the FCP1-Rap74 interaction where binding is endothermic at low temperatures, but transitions to be exothermic above room temperature.[104] Significantly, the calculated change in system entropy made a favorable contribution to FCP1 binding at all temperatures, indicating that the FCP1-Rap74 interaction is at least partially under entropic control, despite being accompanied by an increase in the extent of FCP1 folding.

In macromolecular assembly, structural disorder finds a delicate balance between conformational adaptability and induced stabilization. In fact, recent work has shown that IDPs often utilize cooperativity to assemble large macromolecular complexes.[105, 106] One prominent example is Nup159, an intrinsically disordered hub involved in nucleocytoplasmic transport that assembles with Dyn2 to support the central pore.[25] The binding of one Dyn2 dimer to two Nup159 molecules creates a bivalent scaffold. This structure, albeit weakly assembled, lowers the conformational entropy of the system, thus enabling two more Dyn2 dimers to bind in a cooperative manner, characterized by increasingly favorable enthalpic and entropic contributions. However, after binding the third Dyn2 dimer, binding becomes an enthalpically driven process, owing to the entropic cost of increased rigidity. Interestingly, although there are 6 Dyn2 binding sites on each Nup159 molecule, only 5 become occupied in the fully bound complex. Presumably, this feature is an evolutionary compromise that allows Nup159 to balance the unfavorable entropy of binding successive Dyn2 dimers with moderate affinity, and the stability required for its biochemical function.

In addition to solvation/desolvation effects, IDPs often rely on finely-tuned electrostatic interactions to achieve high specificity.[107] In these cases, ITC can aid in revealing the molecular origins of these phenomena, particularly in conjunction with site-directed mutagenesis of mechanistically significant charged residues.[108-111] For example, in studies of the tumor suppressor p53, ITC has been combined with site-directed mutagenesis to characterize the effects of oncogenic mutations on ASPP2 binding.[112] Similarly, post-translational modifications can lead to changes in electrostatics and ITC has been used to address the role of phosphorylation in binding mechanisms,[94, 113] as highlighted for binding to p300 Taz2.[114] These and other studies have generally shown that electrostatics contribute a modest enhancement to hydrophobic interactions, but extreme cases have been reported where binding is dominated by direct enthalpic contributions. For example, charge-charge complementarity is almost exclusively responsible for the interaction between the Kelch domain of Keap1, a hub protein involved in oxidative stress response, and the intrinsically disordered oncoprotein ProTα.[111] Systems that rely primarily on direct enthalpic contributions for complex formation tend to remain highly disordered in the bound state,[77, 94, 111, 115] underpinning the role of electrostatics in fuzzy complex formation and the order-to-disorder transitions observed for some IDPs upon binding.[115]

While hub proteins have generated significant attention, the thermodynamic profiles for a broad set of IDP interactions, determined from ITC, highlight the functional diversity available to this class of proteins (Table 1). The wide range of binding affinities observed parallels the multifaceted roles of IDPs in the cell. First, the extremely high binding affinity that supports inhibition of the gyrase poison CcdB by the antitoxin CcdA[26] demonstrates that not all IDP interactions are weak and easily reversible. It is especially instructive to compare the varying affinities of ProTα, NRF2, and WTX for the hub protein Keap1, as discussed above.[113] Finally, the delicate balance between entropy and enthalpy that facilitates the coupled folding and binding of IDPs is often masked in single temperature titrations. Therefore, it appears that the most prudent course for investigators is to perform titrations over a range of temperatures, in order to avoid generalizations regarding entropic penalties to binding that may lead to incorrect modelling of the data, or the generation of un-instructive hypotheses.

Table 1. Thermodynamics of IDP interactions with folded partners measured by ITC.

Folded	**IDP**	**K_d (μM)**	**ΔG (kcal/mol)**	**ΔH (kcal/mol)**	**TΔS (kcal/mol)**	**ΔCp (cal mol^{-1}K^{-1})**	**Ref.**
CcdB	CcdA	3.66E-6^{a}	-15.6	-35.5	-19.9	-630	*26*
eIF4E	4E-BP2	0.0032	-11.42^{b}	-8.81	2.61		*134*
Keap1	NRF2	0.023	-10.4	-16.96	-6.56		*113*
Cdk2–cyclin A	p27-KID	0.035	-11.6	-40.2	-28.6	-872	*117*
SBDS**	EFL1	0.0787	-9.86	-14	-4.09		*135*
RPP29^{*,c}	RPP21	0.105	-9.35	5.95	15.29	-1115	*27*
Keap1	WTX	0.25	-9.01	18.04	-9.03		*113*
IκBα*	Relα NLS	0.371	-8.6	-4.2	4.4	-400	*136*
Keap1	WTX pS286	1.5	-7.95	-10.83	-2.88		*113*
RAP74	FCP1	1.91	-7.797	-1.337	6.46	-240	*104*
NudE	IC (1-143)	2.2	-7.8	-4.2	3.6		*137*
Keap1	ProTα	2.6	-7.61	-14.8	-7.19		*111*
SEC3	mVβ8.2	12	-6.70	-4.88	1.82	-136	*138*
Sem-5 C-SH3	SosY	39.22	-6	-8.3	-2.3	-166	*139*
Pcf11 CID*	RNAPII pSer2CTD	180	-5.094^{a}	-9.394^{b}	-4.3		*140*

* Reported parameters are from data collected at 20 ^{o}C.

** Reported parameters are from data collected at 30 ^{o}C.

All remaining parameters are from data collected at 25 ^{o}C.

$^{a}\Delta G= -RTLnK_a$ $^{b}\Delta G= \Delta H – T\Delta S$, ^{c}Both interacting molecules are IDPs.

The Kinetics of IDP Interactions

Elucidating the mechanisms of IDP interactions, and distinguishing between competing kinetic mechanisms, requires knowledge of kinetic rate constants. To this end, NMR relaxation dispersion,[116] surface plasmon resonance,[117, 118] and stopped-flow spectroscopy[119-122] have all been applied. Common to many of these studies, IDP binding is often well described by an apparent two-state kinetic model, which features a linear dependence of k_{obs} on protein concentration, (Figure 1-5A).[23, 123] However, more complex multiphasic kinetics (three-state, and higher), which display a non-linear dependence of k_{obs} on protein concentration, have also been observed.[119, 124] While this behavior reveals that a conformational change takes place along the reaction coordinate, the nature of this transition is widely debated. Several theories have emerged, the dominant two being the conformational selection model, wherein a pre-formed bound state-like conformation is required for ligand recognition and binding (Figure 1-5B), and the induced fit model, where ligand recognition occurs in the disordered state prompting IDP folding (Figure 1-5C). Confirming our expectation that not all IDPs behave equivalently, there is abundant evidence in support of both of these limiting models among the set of proteins studied.[22, 116, 125] In contrast, direct experimental evidence to exclusively support either pathway, through rejection of the other, is often difficult to obtain. The similarity of these models and experimental designs to those used routinely in the protein folding and enzymology communities is no accident. However, IDPs do possess distinct physicochemical properties and so adaptation of methods and models is almost certain to be required in order to accurately represent the kinetic behavior of disordered proteins.

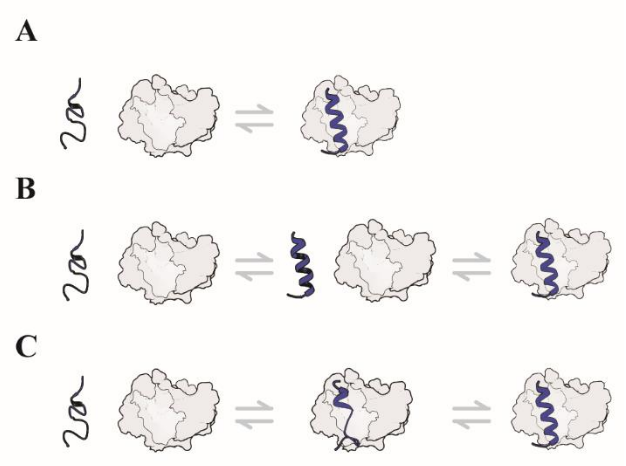


Figure 1-5. Three proposed IDP binding mechanisms include (A) one step binding (apparent two-state); (B) two step binding schemes (apparent three-state), where conformational change precedes binding (conformational selection); and (C) where conformational change follows binding (induced fit).

Resolving Induced Fit from Conformational Selection

In principle, kinetic measurements can be used to distinguish between the two limiting scenarios of conformational selection and induced fit. For example, Gianni et al. argue that by performing experiments in which the concentration of both the protein and ligand are varied separately, the induced fit mechanism will manifest itself as a hyperbolic dependence of k_{obs} on the concentration of both species.[124] However, if a fast conformational change precedes binding, k_{obs} will only display hyperbolic behavior when the species that undergoes conformational change is held constant; linear behavior will manifest when the concentration of the species undergoing a conformational change is increased. While such an experimental design is theoretically sound, it is often difficult to implement in practice, as high concentrations of both interacting species are required. Also, kinetic methods with unusually fast time resolution may be needed to detect fast folding of preformed structural elements, such as α-helicies. Another type of decisive experiment involves determining whether a binding reaction is diffusion limited. As Rogers et al. argue, reaching the diffusion limit would require all molecular collisions between cognate partners to result in binding, regardless of the particular conformation of the IDP (i.e., binding would proceed through induced fit).[126] Detailed kinetic investigation of the Bcl-1:PUMA interaction showed that the criteria necessary to define the diffusion limited reaction for folded proteins – the predictable dependence of k_a on solvent viscosity and temperature – may not be sufficient for disordered proteins, due to geometrical considerations.[126]

Synergistic models, which combine aspects of both mechanisms, have also been reported. For example, the extended conformational capture model builds upon a similar framework to the folding funnel hypothesis discussed above and highlights how the energy surfaces of both the IDP and the folded partner can influence each other and bias the IDP's binding trajectory.[14] In a natural extension of the induced fit/ conformational capture dichotomy discussed earlier in this section, flux based models acknowledge that conformational selection and induced fit pathways may both be present as limiting behaviors for most systems. From this new point of view, the flux of the reaction can be biased by both intrinsic and extrinsic factors, explaining why some systems have produced evidence in support of both models (Figure 1-6). For example, Greives et al. suggest that intra-chain dynamics within the IDP ensemble can shift the binding mechanism, with slower conformational transitions favoring conformational selection and fast inter-conversion rates favoring induced fit pathways.[127] In addition, Hammes et al. have suggested ligand concentration is an important determinant to increase flux through a particular pathway.[128] As an example of these mechanistic nuances, Daniels et al. showed that the *Bacillus subtilis* RNase P protein experienced varying levels of folding through either mechanism, based on the concentration of PPi, with lower and higher concentrations of ligand favoring conformational selection and induced fit pathways, respectively.[122] Thus, from a biological perspective, flux models provide an attractive rationale for kinetic control of signaling in response to environmental stimuli.

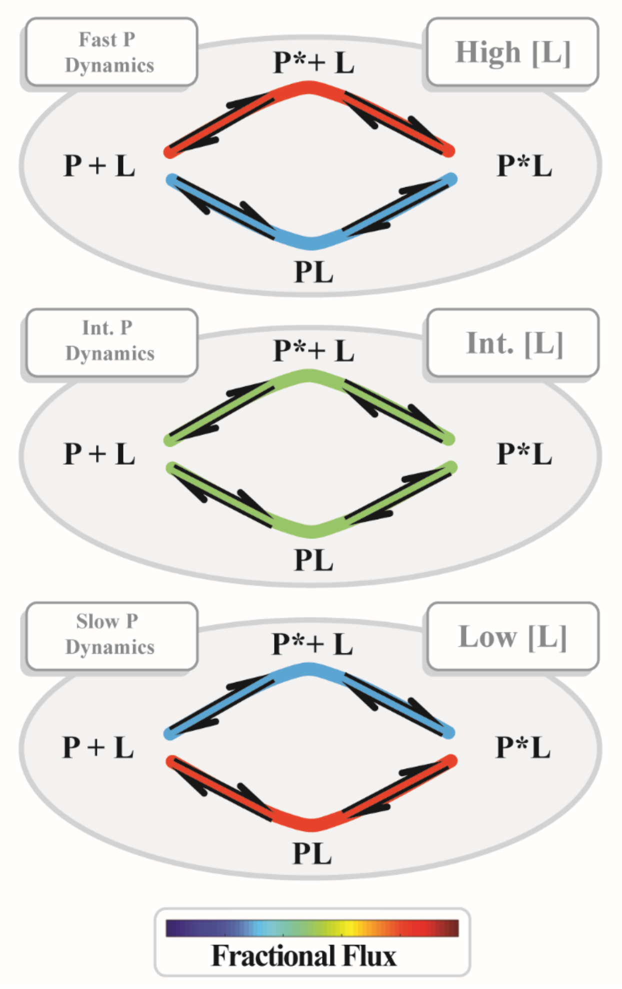


Figure 1-6. A generalized reaction scheme where flux is kinetically partitioned between the conformational selection and induced fit pathways, based on IDP conformational dynamics and ligand concentration.

Transition States

In the field of kinetic enzymology, identification of transition states and their description *vis* the substrate or product structure has been the key to broad progress, often leading to rational design of inhibitors and motivating the search for drug candidates. To this end, several groups have applied a protein engineering method known as the Φ-value analysis, which has also seen some success in describing the kinetics of IDP binding.[23, 24, 120, 129, 130] In Φ-value analysis, atomic-level structural information about the transition state is inferred by comparing the binding kinetics of the wild-type protein to a series of single point mutants to obtain the "Φ-value," which is calculated as the ratio of the change in activation energy for folding upon mutation ($\Delta\Delta G^{\dagger\dagger}$) and the change in equilibrium free energy upon mutation ($\Delta\Delta G^{eq}$).[131] For IDP interactions, it is common to introduce alanine-glycine substitutions or non-disruptive mutations to reduce the size of side chains, in order to gain insight into the secondary and tertiary structure of the transition state, respectively.

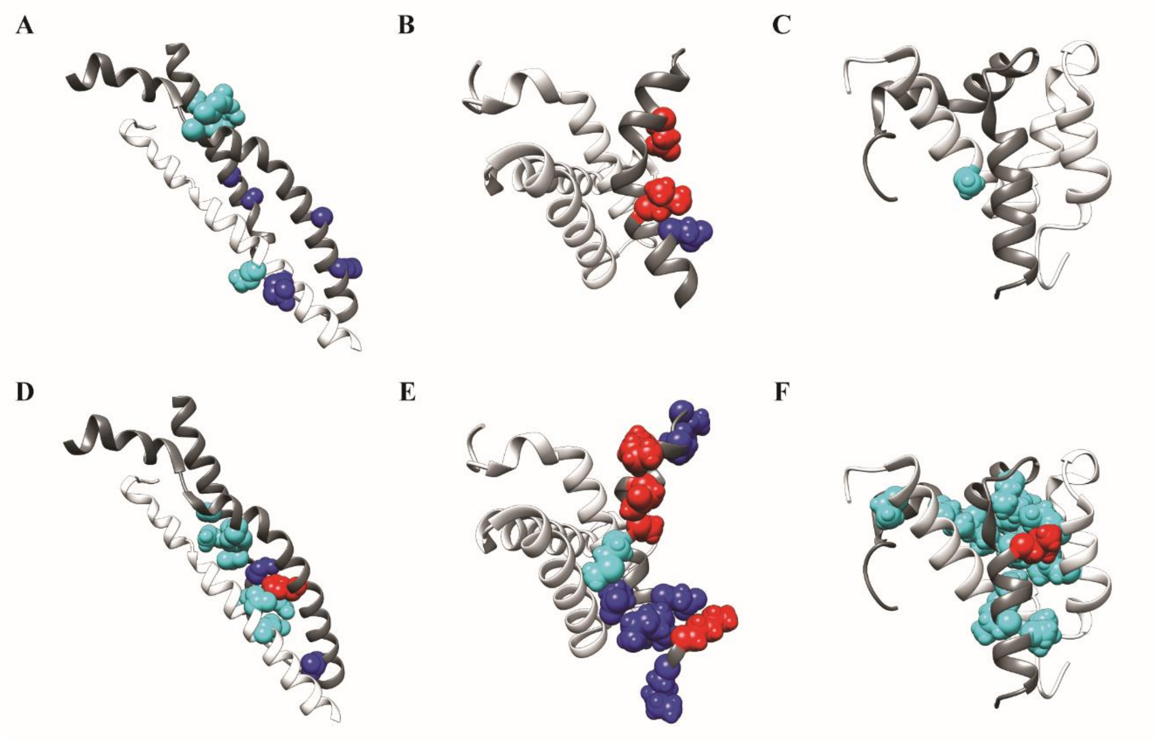


Figure 1-7. Secondary (A-C) and tertiary (D-F) Φ-values classified as weak (0< Φ <0.3) shown in cyan, medium (0.3< Φ <0.7) shown in blue, and strong (0.7< Φ <1) shown in red. Φ-values are mapped onto PDB structures for (A,D) erythroid α-spectrin (white)-β-spectrin (grey) (PDB:3LBX), (B,E) pwtKIX (white)-c-MYB* (grey) (PDB: 1SB0), and (C,F) ACTR (grey)-NCBD (white) (PDB: 1KBH).

Recently, Hill et al. used a detailed Φ-value analysis to characterize the formation of spectrin repeat domains (Figure 1-7A, D).[23] The analysis invoked a model wherein the preformed C-helix from α-spectrin acts as a template that guides the ensemble of transiently formed secondary structures in the A & B-helicies of β-spectrin toward the bound state, which is a fully folded triple-helical domain. This “templating mechanism” is an interesting example of synergistic binding, where both the presence of preformed structural elements and structural adaptation act in concert to accomplish folding and binding. In the coupled folding and binding of c-Myb to KIX (Figure 1-7B, D), Φ-values suggest c-Myb possesses a high degree of native-like structure in the transition state.[24] In contrast, Φ-value analysis of the ACTR:NCBD interaction (Figure 7C, E) suggested the presence of substantially more disorder in the transition state.[120] It is also interesting to consider that there is a positive correlation between preformed structure and binding kinetics for ACTR:NCBD,[125] while residual structure has little effect on the binding kinetics of c-Myb:KIX.[22] This suggests that, for IDPs possessing a large degree of bound-state structure in the transition state, pre-stabilization of these conformations can lower the energetic barrier of the rate limiting step for association. Importantly, these Φ-value analyses highlight the general applicability of methods traditionally used to study catalysis, ligand binding, or protein folding to the study of coupled folding and binding involving IDPs.

The most general conclusion to be drawn from the kinetic investigations conducted thus far is that IDPs are not monolithic as a class of proteins. Rather, these studies suggest that disordered proteins rely on a range of mechanisms to bind their partners, just as cooperatively folding proteins do. It is becoming increasingly apparent that IDPs may possess various pathways toward the bound state, which are often influenced by extrinsic factors, such as local ligand concentration.[119, 122] Furthermore, detailed structural knowledge can facilitate the design of experiments aimed at discriminating between limiting pathways, because the dynamics of conformational fluctuations within the ensemble often couple to kinetic outcomes.[127] Of significant biological interest, the kinetic mechanism describing IDP interactions is often influenced by the presence or absence of post-translational modifications.[132, 133] Clearly, this is a rich area of future growth for biochemists to explore.

Conclusion

In the past few decades, the field of intrinsically disordered proteins has increased dramatically, yielding several general conclusions to be drawn from the studies presented here. Most importantly, IDPs have much in common with their cooperatively-folded counterparts. Experimental methods and computational procedures for ensemble generation now enable routine modeling of disordered systems, facilitating hypothesis testing as in any other field of structural biology. Of pressing need now is a formal framework for ensemble validation. Rigorous assessment is needed to define input constraint combinations that cost-effectively produce the most unique ensembles, while also minimizing over-fitting. More significantly, standardized reporting practices would benefit the community. For example, the PDB has established data-reporting criteria for model deposition and the community has agreed to helpful norms regarding, e.g., the number of models to be included in NMR structure bundles. Similar guidelines for reporting IDP ensembles in the pE-DB or other databases would help investigators assess model quality for themselves.

Also of great significance, equilibrium thermodynamics experiments have helped to dispel the common misconception that structural disorder constrains interactions to a narrow range of affinities by imposing entropic penalties to binding. Indeed, IDPs interact over a broad range of affinities, utilize cooperativity to enhance stability, and rely on hydrophobic effects for coupled folding and binding, in strong analogy to protein folding. Most notably, recent applications of isothermal titration calorimetry have provided insight into the hydrophobic impetus for coupled-folding and binding. The continued application of variable temperature ITC will deepen our understanding of this process and may help to better understand the functional advantages of bound-state induced disorder and dynamic fuzziness.

Finally, kinetic studies have demonstrated that IDPs rely on a broad range of mechanisms to accomplish biological function. Although models for coupled folding and binding are typically built upon thermodynamic observations, mechanistic insights into these processes require detailed kinetic analyses. Recent studies have demonstrated that many of the experimental techniques used in enzymology or protein folding are directly transferrable to disordered systems, or require modest adaptation to the physicochemical norms of IDPs. Continued kinetic investigation will be necessary to elucidate the ways in which IDP interactions can be tuned to support intricate biochemical pathways. The prevalence of protein non-folding as an important regulatory mechanism in biology is well established, yielding a rich new class of proteins for biochemists to characterize quantitatively in the laboratory.

References

1. Nagel, R. L., Gibson, Q. H., and Charache, S. (1967) Relation between structure and function in Hemoglobin Chesapeake, *Biochemistry 6*, 2395-2402.
2. Austin, R. H., Beeson, K. W., Eisenstein, L., Frauenfelder, H., and Gunsalus, I. C. (1975) Dynamics of ligand binding to myoglobin, *Biochemistry 14*, 5355-5373.

3. Boehr, D. D., Nussinov, R., and Wright, P. E. (2009) The role of dynamic conformational ensembles in biomolecular recognition, *Nat. Chem. Biol. 5*, 789-796.
4. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E., and Babu, M. M. (2014) Classification of intrinsically disordered regions and proteins, *Chem. Rev. 114*, 6589-6631.
5. Oates, M. E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M. J., Xue, B., Dosztanyi, Z., Uversky, V. N., Obradovic, Z., Kurgan, L., Dunker, A. K., and Gough, J. (2013) D(2)P(2): database of disordered protein predictions, *Nucleic Acids Res. 41*, D508-516.
6. Tantos, A., Han, K. H., and Tompa, P. (2012) Intrinsic disorder in cell signaling and gene transcription, *Mol. Cell. Endocrinol. 348*, 457-465.
7. Ferreon, A. C., Ferreon, J. C., Wright, P. E., and Deniz, A. A. (2013) Modulation of allostery by protein intrinsic disorder, *Nature 498*, 390-394.
8. Gunasekaran, K., Tsai, C. J., Kumar, S., Zanuy, D., and Nussinov, R. (2003) Extended disordered proteins: targeting function with less scaffold, *Trends Biochem. Sci. 28*, 81-85.
9. Gunasekaran, K., Haspel, N., Tsai, C. J., Kumar, S., Wolfson, H., and Nussinov, R. (2003) Extended disordered proteins: An elegant solution to having large intermolecular interfaces, yet keeping smaller genome and cell sizes, *Biophys. J. 84*, 163a-163a.
10. Uversky, V. N., Dave, V., Iakoucheva, L. M., Malaney, P., Metallo, S. J., Pathak, R. R., and Joerger, A. C. (2014) Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases, *Chem. Rev. 114*, 6844-6879.
11. Gsponer, J., Futschik, M. E., Teichmann, S. A., and Babu, M. M. (2008) Tight regulation of unstructured proteins: from transcript synthesis to protein degradation, *Science 322*, 1365-1368.
12. Tompa, P. (2002) Intrinsically unstructured proteins, *Trends Biochem. Sci. 27*, 527-533.
13. Papoian, G. A., and Wolynes, P. G. (2003) The physics and bioinformatics of binding and folding-an energy landscape perspective, *Biopolymers 68*, 333-349.
14. Csermely, P., Palotai, R., and Nussinov, R. (2010) Induced fit, conformational selection and independent dynamic segments: an extended view of binding events, *Trends Biochem. Sci. 35*, 539-546.
15. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., and Dunker, A. K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins, *J. Mol. Biol. 323*, 573-584.
16. Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, J. L. (2008) Function and structure of inherently disordered proteins, *Curr. Opin. Struct. Biol. 18*, 756-764.
17. Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005) Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling, *J. Mol. Recognit. 18*, 343-384.

18. Fuxreiter, M., Tompa, P., Simon, I., Uversky, V. N., Hansen, J. C., and Asturias, F. J. (2008) Malleable machines take shape in eukaryotic transcriptional regulation, *Nature Chem. Biol. 4*, 728-737.

19. Liu, J., Perumal, N. B., Oldfield, C. J., Su, E. W., Uversky, V. N., and Dunker, A. K. (2006) Intrinsic disorder in transcription factors, *Biochemistry 45*, 6873-6888.

20. Tompa, P. (2014) Multisteric regulation by structural disorder in modular signaling proteins: an extension of the concept of allostery, *Chem. Rev. 114*, 6715-6732.

21. Wright, P. E., and Dyson, H. J. (2009) Linking folding and binding, *Curr. Opin. Struct. Biol. 19*, 31-38.

22. Rogers, J. M., Wong, C. T., and Clarke, J. (2014) Coupled folding and binding of the disordered protein PUMA does not require particular residual structure, *J. Am. Chem. Soc. 136*, 5197-5200.

23. Hill, S. A., Kwa, L. G., Shammas, S. L., Lee, J. C., and Clarke, J. (2014) Mechanism of assembly of the non-covalent spectrin tetramerization domain from intrinsically disordered partners, *J. Mol. Biol. 426*, 21-35.

24. Giri, R., Morrone, A., Toto, A., Brunori, M., and Gianni, S. (2013) Structure of the transition state for the binding of c-Myb and KIX highlights an unexpected order for a disordered system, *Proc. Natl. Acad. Sci. USA 110*, 14942-14947.

25. Nyarko, A., Song, Y., Novacek, J., Zidek, L., and Barbar, E. (2013) Multiple recognition motifs in nucleoporin Nup159 provide a stable and rigid Nup159-Dyn2 assembly, *J. Biol. Chem. 288*, 2614-2622.

26. Drobnak, I., De Jonge, N., Haesaerts, S., Vesnaver, G., Loris, R., and Lah, J. (2013) Energetic basis of uncoupling folding from binding for an intrinsically disordered protein, *J. Am. Chem. Soc. 135*, 1288-1294.

27. Xu, Y., Oruganti, S. V., Gopalan, V., and Foster, M. P. (2012) Thermodynamics of coupled folding in the interaction of archaeal RNase P proteins RPP21 and RPP29, *Biochemistry 51*, 926-935.

28. Ferreon, A. C., Moran, C. R., Gambin, Y., and Deniz, A. A. (2010) Single-molecule fluorescence studies of intrinsically disordered proteins, *Methods Enzymol. 472*, 179-204.

29. Drescher, M., Huber, M., and Subramaniam, V. (2012) Hunting the chameleon: structural conformations of the intrinsically disordered protein alpha-synuclein, *Chembiochem 13*, 761-768.

30. Trexler, A. J., and Rhoades, E. (2013) Function and dysfunction of alpha-synuclein: probing conformational changes and aggregation by single molecule fluorescence, *Molecular neurobiology 47*, 622-631.

31. Click, T. H., Ganguly, D., and Chen, J. (2010) Intrinsically disordered proteins in a physics-based world, *Int. J. Mol. Sci. 11*, 5292-5309.

32. Baker, C. M., and Best, R. B. (2014) Insights into the binding of intrinsically disordered proteins from molecular dynamics simulation, *Wiley Interdiscip. Rev. Comput. Mol. Sci. 4*, 182-198.

33. Ostermeir, K., and Zacharias, M. (2013) Advanced replica-exchange sampling to study the flexibility and plasticity of peptides and proteins, *Biochim. Biophys. Acta 1834*, 847-853.

34. Ganguly, D., and Chen, J. (2009) Atomistic details of the disordered states of KID and pKID. Implications in coupled binding and folding, *J. Am. Chem. Soc. 131*, 5214-5223.

35. Vendruscolo, M. (2007) Determination of conformationally heterogeneous states of proteins, *Curr. Opin. Struct. Biol. 17*, 15-20.

36. Mittag, T., and Forman-Kay, J. D. (2007) Atomic-level characterization of disordered protein ensembles, *Curr. Opin. Struct. Biol. 17*, 3-14.

37. Jensen, M. R., Ruigrok, R. W., and Blackledge, M. (2013) Describing intrinsically disordered proteins at atomic resolution by NMR, *Curr. Opin. Struct. Biol. 23*, 426-435.

38. Bermel, W., Bertini, I., Felli, I. C., Piccioli, M., and Pierattelli, R. (2006) ^{13}C-detected protonless NMR spectroscopy of proteins in solution, *Prog. Nucl. Magn. Reson. Spectrosc. 48*, 25-45.

39. O'Hare, B., Benesi, A. J., and Showalter, S. A. (2009) Incorporating ^{1}H-chemical shift determination into ^{13}C-direct detected spectroscopy of intrinsically disordered proteins in solution, *J. Magn. Reson. 200*, 354-358.

40. Showalter, S. A. (2009) NMR Assignment of the Intrinsically Disordered C-terminal Region of Homo sapiens FCP1 in the Unbound State, *Biomol. NMR Assign. 3*, 179-181.

41. Lawrence, C. W., Bonny, A., and Showalter, S. A. (2011) The disordered C-terminus of the RNA Polymerase II phosphatase FCP1 is partially helical in the unbound state, *Biochem. Biophys. Res. Commun. 410*, 461-465.

42. Lawrence, C. W., and Showalter, S. A. (2012) Carbon-Detected N-15 NMR Spin Relaxation of an Intrinsically Disordered Protein: FCP1 Dynamics Unbound and in Complex with RAP74, *J. Phys. Chem. Lett. 3*, 1409-1413.

43. Sahu, D., Bastidas, M., and Showalter, S. A. (2014) Generating NMR Chemical Shift Assignments of Intrinsically Disordered Proteins using Carbon-Detected NMR Methods, *Anal. Biochem. 449*, 17-25.

44. Showalter, S. A. (2014) Intrinsically Disordered Proteins: Methods for Structure and Dynamics Studies, *EMagRes 3*, 181-189.

45. Theillet, F.-X., Kalmar, L., Tompa, P., Han, K.-H., Selenko, P., Dunker, A. K., Daughdrill, G. W., and Uversky, V. N. (2013) The alphabet of intrinsic disorder, *Intrinsically Disord. Proteins 1*, e24360.

46. Jensen, M. R., Zweckstetter, M., Huang, J. R., and Blackledge, M. (2014) Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy, *Chem. Rev. 114*, 6632-6660.

47. Schneider, R., Huang, J. R., Yao, M., Communie, G., Ozenne, V., Mollica, L., Salmon, L., Jensen, M. R., and Blackledge, M. (2012) Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy, *Mol. Biosyst. 8*, 58-68.

48. Jensen, M. R., Salmon, L., Nodet, G., and Blackledge, M. (2010) Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts, *J. Am. Chem. Soc. 132*, 1270-1272.

49. Fuxreiter, M., Simon, I., Friedrich, P., and Tompa, P. (2004) Preformed structural elements feature in partner recognition by intrinsically unstructured proteins, *J. Mol. Biol. 338*, 1015-1026.

50. Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., and Uversky, V. N. (2006) Analysis of molecular recognition features (MoRFs), *J. Mol. Biol. 362*, 1043-1059.

51. Mészáros, B., Simon, I., and Dosztányi, Z. (2009) Prediction of protein binding regions in disordered proteins, *PLoS Comput. Biol. 5*, e1000376.

52. Jensen, M. R., Houben, K., Lescop, E., Blanchard, L., Ruigrok, R. W. H., and Blackledge, M. (2008) Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: Application to the molecular recognition element of Sendai virus nucleoprotein, *J. Am. Chem. Soc. 130*, 8055-8061.

53. Vise, P., Baral, B., Stancik, A., Lowry, D. F., and Daughdrill, G. W. (2007) Identifying long-range structure in the intrinsically unstructured transactivation domain of p53, *Proteins 67*, 526-530.

54. Wu, K. P., and Baum, J. (2010) Detection of transient interchain interactions in the intrinsically disordered protein alpha-synuclein by NMR paramagnetic relaxation enhancement, *J. Am. Chem. Soc. 132*, 5546-5547.

55. Iesmantavicius, V., Jensen, M. R., Ozenne, V., Blackledge, M., Poulsen, F. M., and Kjaergaard, M. (2013) Modulation of the intrinsic helix propensity of an intrinsically disordered protein reveals long-range helix-helix interactions, *J. Am. Chem. Soc. 135*, 10155-10163.

56. Esteban-Martin, S., Silvestre-Ryan, J., Bertoncini, C. W., and Salvatella, X. (2013) Identification of fibril-like tertiary contacts in soluble monomeric alpha-synuclein, *Biophys. J. 105*, 1192-1198.

57. Hennig, J., and Sattler, M. (2014) The dynamic duo: combining NMR and small angle scattering in structural biology, *Protein Sci. 23*, 669-682.

58. Sibille, N., and Bernado, P. (2012) Structural characterization of intrinsically disordered proteins by the combined use of NMR and SAXS, *Biochem. Soc. Trans. 40*, 955-962.

59. Schwalbe, M., Ozenne, V., Bibow, S., Jaremko, M., Jaremko, L., Gajda, M., Jensen, M. R., Biernat, J., Becker, S., Mandelkow, E., Zweckstetter, M., and Blackledge, M. (2014) Predictive atomic resolution descriptions of intrinsically disordered hTau40 and alpha-synuclein in solution from NMR and small angle scattering, *Structure 22*, 238-249.

60. Sterckx, Y. G., Volkov, A. N., Vranken, W. F., Kragelj, J., Jensen, M. R., Buts, L., Garcia-Pino, A., Jove, T., Van Melderen, L., Blackledge, M., van Nuland, N. A., and Loris, R. (2014) Small-angle X-ray scattering- and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin PaaA2, *Structure 22*, 854-865.

61. Huang, J. R., Warner, L. R., Sanchez, C., Gabel, F., Madl, T., Mackereth, C. D., Sattler, M., and Blackledge, M. (2014) Transient electrostatic interactions dominate the conformational equilibrium sampled by multidomain splicing factor U2AF65: a combined NMR and SAXS study, *J. Am. Chem. Soc. 136*, 7068-7076.

62. Bernado, P., Blanchard, L., Timmins, P., Marion, D., Ruigrok, R. W. H., and Blackledge, M. (2005) A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering, *Proc. Nat. Acad. Sci. USA 102*, 17002-17007.

63. Wells, M., Tidow, H., Rutherford, T. J., Markwick, P., Jensen, M. R., Mylonas, E., Svergun, D. I., Blackledge, M., and Fersht, A. R. (2008) Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain, *Proc. Nat. Acad. Sci. USA 105*, 5762-5767.

64. Mittag, T., Marsh, J., Grishaev, A., Orlicky, S., Lin, H., Sicheri, F., Tyers, M., and Forman-Kay, J. D. (2010) Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase, *Structure 18*, 494-506.

65. Rambo, R. P., and Tainer, J. A. (2011) Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law, *Biopolymers 95*, 559-571.

66. Marasini, C., Galeno, L., and Moran, O. (2013) A SAXS-based ensemble model of the native and phosphorylated regulatory domain of the CFTR, *Cell. Mol. Life. Sci. 70*, 923-933.

67. Bianconi, A., Ciasca, G., Tenenbaum, A., Battisti, A., and Campi, G. (2012) Temperature and solvent dependence of the dynamical landscape of tau protein conformations, *J. Biol. Phys. 38*, 169-179.

68. Budkevich, T. V., Timchenko, A. A., Tiktopulo, E. I., Negrutskii, B. S., Shalak, V. F., Petrushenko, Z. M., Aksenov, V. L., Willumeit, R., Kohlbrecher, J., Serdyuk, I. N., and El'skaya, A. V. (2002) Extended conformation of mammalian translation elongation factor 1A in solution, *Biochemistry 41*, 15342-15349.

69. Goldenberg, D. P., and Argyle, B. (2014) Minimal effects of macromolecular crowding on an intrinsically disordered protein: a small-angle neutron scattering study, *Biophys. J. 106*, 905-914.

70. Stadler, A. M., Stingaciu, L., Radulescu, A., Holderer, O., Monkenbusch, M., Biehl, R., and Richter, D. (2014) Internal nanosecond dynamics in the intrinsically disordered myelin basic protein, *J. Am. Chem. Soc. 136*, 6987-6994.

71. Sugiyama, M., Arimura, Y., Shirayama, K., Fujita, R., Oba, Y., Sato, N., Inoue, R., Oda, T., Sato, M., Heenan, R. K., and Kurumizaka, H. (2014) Distinct features of the histone core structure in nucleosomes containing the histone H2A.B variant, *Biophys. J. 106*, 2206-2213.

72. Mukhopadhyay, S., Krishnan, R., Lemke, E. A., Lindquist, S., and Deniz, A. A. (2007) A natively unfolded yeast prion monomer adopts an ensemble of collapsed and rapidly fluctuating structures, *Proc. Natl. Acad. Sci. USA 104*, 2649-2654.

73. Ferreon, A. C., Gambin, Y., Lemke, E. A., and Deniz, A. A. (2009) Interplay of alpha-synuclein binding and conformational switching probed by single-molecule fluorescence, *Proc. Natl. Acad. Sci. USA 106*, 5645-5650.

74. Ducas, V. C., and Rhoades, E. (2014) Investigation of intramolecular dynamics and conformations of alpha-, beta- and gamma-synuclein, *PLoS One 9*, e86983.

75. Hofmann, H., Soranno, A., Borgia, A., Gast, K., Nettels, D., and Schuler, B. (2012) Polymer scaling laws of unfolded and intrinsically disordered proteins

quantified with single-molecule spectroscopy, *Proc. Natl. Acad. Sci. USA 109*, 16155-16160.

76. Muller-Spath, S., Soranno, A., Hirschfeld, V., Hofmann, H., Ruegger, S., Reymond, L., Nettels, D., and Schuler, B. (2010) Charge interactions can dominate the dimensions of intrinsically disordered proteins, *Proc. Natl. Acad. Sci. USA 107*, 14609-14614.

77. Liu, B., Chia, D., Csizmok, V., Farber, P., Forman-Kay, J. D., and Gradinaru, C. C. (2014) The effect of intrachain electrostatic repulsion on conformational disorder and dynamics of the Sic1 protein, *J. Phys. Chem. B 118*, 4088-4097.

78. Mooney, S. M., Qiu, R., Kim, J. J., Sacho, E. J., Rajagopalan, K., Johng, D., Shiraishi, T., Kulkarni, P., and Weninger, K. R. (2014) Cancer/testis antigen PAGE4, a regulator of c-Jun transactivation, is phosphorylated by homeodomain-interacting protein kinase 1, a component of the stress-response pathway, *Biochemistry 53*, 1670-1679.

79. Kalinin, S., Peulen, T., Sindbert, S., Rothwell, P. J., Berger, S., Restle, T., Goody, R. S., Gohlke, H., and Seidel, C. A. (2012) A toolkit and benchmark study for FRET-restrained high-precision structural modeling, *Nat. Methods 9*, 1218-1225.

80. Uversky, V. N. (2002) Natively unfolded proteins: A point where biology waits for physics, *Protein Sci. 11*, 739-756.

81. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, *Nucleic Acids Res. 28*, 235-242.

82. Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A. K., Felli, I. C., Forman-Kay, J. D., Kriwacki, R. W., Pierattelli, R., Sussman, J., Svergun, D. I., Uversky, V. N., Vendruscolo, M., Wishart, D., Wright, P. E., and Tompa, P. (2014) pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins, *Nucleic Acids Res. 42*, D326-335.

83. Ball, K. A., Phillips, A. H., Wemmer, D. E., and Head-Gordon, T. (2013) Differences in beta-strand Populations of Monomeric Abeta40 and Abeta42, *Biophys. J. 104*, 2714-2724.

84. Xiang, S., Gapsys, V., Kim, H. Y., Bessonov, S., Hsiao, H. H., Mohlmann, S., Klaukien, V., Ficner, R., Becker, S., Urlaub, H., Luhrmann, R., de Groot, B., and Zweckstetter, M. (2013) Phosphorylation drives a dynamic switch in serine/arginine-rich proteins, *Structure 21*, 2162-2174.

85. Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S., and Shaw, D. E. (2012) Structure and dynamics of an unfolded protein examined by molecular dynamics simulation, *J. Am. Chem. Soc. 134*, 3787-3791.

86. Fisher, C. K., Huang, A., and Stultz, C. M. (2010) Modeling intrinsically disordered proteins with bayesian statistics, *J. Am. Chem. Soc. 132*, 14919-14927.

87. Mukrasch, M. D., Markwick, P., Biernat, J., von Bergen, M., Bernado, P., Griesinger, C., Mandelkow, E., Zweckstetter, M., and Blackledge, M. (2007) Highly populated turn conformations in natively unfolded Tau protein identified from residual dipolar couplings and molecular simulation, *J. Am. Chem. Soc. 129*, 5235-5243.

88. Vitalis, A., and Pappu, R. V. (2009) Methods for Monte Carlo simulations of biomacromolecules, *Annu. Rep. Comput. Chem. 5*, 49-76.

89. Vitalis, A., and Pappu, R. V. (2009) ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions, *J. Comput. Chem. 30*, 673-699.

90. Feldman, H. J., and Hogue, C. W. (2000) A fast method to sample real protein conformational space, *Proteins 39*, 112-131.

91. Marsh, J. A., and Forman-Kay, J. D. (2009) Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints, *J. Mol. Biol. 391*, 359-374.

92. Zhuo, Y., Ilangovan, U., Schirf, V., Demeler, B., Sousa, R., Hinck, A. P., and Lafer, E. M. (2010) Dynamic interactions between clathrin and locally structured elements in a disordered protein mediate clathrin lattice assembly, *J. Mol. Biol. 404*, 274-290.

93. Mittag, T., Orlicky, S., Choy, W. Y., Tang, X., Lin, H., Sicheri, F., Kay, L. E., Tyers, M., and Forman-Kay, J. D. (2008) Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor, *Proc. Natl. Acad. Sci. USA 105*, 17772-17777.

94. Lunde, B. M., Reichow, S. L., Kim, M., Suh, H., Leeper, T. C., Yang, F., Mutschler, H., Buratowski, S., Meinhart, A., and Varani, G. (2010) Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain, *Nat. Struct. Mol. Biol. 17*, 1195-1201.

95. Bozoky, Z., Krzeminski, M., Muhandiram, R., Birtley, J. R., Al-Zahrani, A., Thomas, P. J., Frizzell, R. A., Ford, R. C., and Forman-Kay, J. D. (2013) Regulatory R region of the CFTR chloride channel is a dynamic integrator of phospho-dependent intra- and intermolecular interactions, *Proc. Natl. Acad. Sci. USA 110*, E4427-4436.

96. Butz, M., Kast, P., and Hilvert, D. (2014) Affinity maturation of a computationally designed binding protein affords a functional but disordered polypeptide, *J. Struct. Biol. 185*, 168-177.

97. Rhoades, E., Ramlall, T. F., Webb, W. W., and Eliezer, D. (2006) Quantification of alpha-synuclein binding to lipid vesicles using fluorescence correlation spectroscopy, *Biophys. J. 90*, 4692-4700.

98. Milles, S., and Lemke, E. A. (2014) Mapping multivalency and differential affinities within large intrinsically disordered protein complexes with segmental motion analysis, *Angew. Chem. Int. Ed. Engl. 53*, 7364-7367.

99. Elbaum-Garfinkle, S., Cobb, G., Compton, J. T., Li, X. H., and Rhoades, E. (2014) Tau mutants bind tubulin heterodimers with enhanced affinity, *Proc. Natl. Acad. Sci. USA 111*, 6311-6316.

100. Ghai, R., Falconer, R. J., and Collins, B. M. (2012) Applications of isothermal titration calorimetry in pure and applied research--survey of the literature from 2010, *J. Mol. Recognit. 25*, 32-52.

101. Tompa, P., and Fuxreiter, M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions, *Trends Biochem. Sci. 33*, 2-8.

102. Espinoza-Fonseca, L. M. (2012) Aromatic residues link binding and function of intrinsically disordered proteins, *Mol. Biosyst. 8*, 237-246.

103. Brzovic, P. S., Heikaus, C. C., Kisselev, L., Vernon, R., Herbig, E., Pacheco, D., Warfield, L., Littlefield, P., Baker, D., Klevit, R. E., and Hahn, S. (2011) The acidic transcription activator Gcn4 binds the mediator subunit Gal11/Med15 using a simple protein interface forming a fuzzy complex, *Mol. Cell 44*, 942-953.

104. Lawrence, C. W., Kumar, S., Noid, W. G., and Showalter, S. A. (2014) Role of Ordered Proteins in the Folding-Upon-Binding of Intrinsically Disordered Proteins, *J. Phys. Chem. Lett. 5*, 833-838.

105. Hall, J., Karplus, P. A., and Barbar, E. (2009) Multivalency in the assembly of intrinsically disordered Dynein intermediate chain, *J. Biol. Chem. 284*, 33115-33121.

106. Slevin, L. K., Romes, E. M., Dandulakis, M. G., and Slep, K. C. (2014) The Mechanism of Dynein Light Chain LC8-mediated Oligomerization of the Ana2 Centriole Duplication Factor, *J. Biol. Chem. 289*, 20727-20739.

107. Mittag, T., Kay, L. E., and Forman-Kay, J. D. (2010) Protein dynamics and conformational disorder in molecular recognition, *J. Mol. Recognit. 23*, 105-116.

108. Demarest, S. J., Deechongkit, S., Dyson, H. J., Evans, R. M., and Wright, P. E. (2004) Packing, specificity, and mutability at the binding interface between the p160 coactivator and CREB-binding protein, *Protein Sci. 13*, 203-210.

109. Zhan, Y. A., Wu, H., Powell, A. T., Daughdrill, G. W., and Ytreberg, F. M. (2013) Impact of the K24N mutation on the transactivation domain of p53 and its binding to murine double-minute clone 2, *Proteins 81*, 1738-1747.

110. Krieger, J. M., Fusco, G., Lewitzky, M., Simister, P. C., Marchant, J., Camilloni, C., Feller, S. M., and De Simone, A. (2014) Conformational recognition of an intrinsically disordered protein, *Biophys. J. 106*, 1771-1779.

111. Khan, H., Cino, E. A., Brickenden, A., Fan, J., Yang, D., and Choy, W. Y. (2013) Fuzzy complex formation between the intrinsically disordered prothymosin alpha and the Kelch domain of Keap1 involved in the oxidative stress response, *J. Mol. Biol. 425*, 1011-1027.

112. Tidow, H., Veprintsev, D. B., Freund, S. M., and Fersht, A. R. (2006) Effects of oncogenic mutations and DNA response elements on the binding of p53 to p53-binding protein 2 (53BP2), *J. Biol. Chem. 281*, 32526-32533.

113. Cino, E. A., Killoran, R. C., Karttunen, M., and Choy, W. Y. (2013) Binding of disordered proteins to a protein hub, *Sci. Rep. 3*, 2305.

114. Feng, H., Jenkins, L. M., Durell, S. R., Hayashi, R., Mazur, S. J., Cherry, S., Tropea, J. E., Miller, M., Wlodawer, A., Appella, E., and Bai, Y. (2009) Structural basis for p300 Taz2-p53 TAD1 binding and modulation by phosphorylation, *Structure 17*, 202-210.

115. Kurzbach, D., Schwarz, T. C., Platzer, G., Hofler, S., Hinderberger, D., and Konrat, R. (2014) Compensatory adaptations of structural dynamics in an intrinsically disordered protein complex, *Angew. Chem. Int. Ed. Engl. 53*, 3840-3843.

116. Sugase, K., Dyson, H. J., and Wright, P. E. (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein, *Nature 447*, 1021-1025.

117. Lacy, E. R., Filippov, I., Lewis, W. S., Otieno, S., Xiao, L. M., Weiss, S., Hengst, L., and Kriwacki, R. W. (2004) p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding, *Nature Struct. Mol. Biol. 11*, 358-364.

118. Kapinos, L. E., Schoch, R. L., Wagner, R. S., Schleicher, K. D., and Lim, R. Y. (2014) Karyopherin-centric control of nuclear pores based on molecular occupancy and kinetic analysis of multivalent binding with FG nucleoporins, *Biophys. J. 106*, 1751-1762.

119. Chang, Y. C., and Oas, T. G. (2010) Osmolyte-induced folding of an intrinsically disordered protein: folding mechanism in the absence of ligand, *Biochemistry 49*, 5086-5096.

120. Dogan, J., Mu, X., Engstrom, A., and Jemth, P. (2013) The transition state structure for coupled binding and folding of disordered protein domains, *Sci. Rep. 3*, 2076.

121. Wang, N., Lodge, J. M., Fierke, C. A., and Mapp, A. K. (2014) Dissecting allosteric effects of activator-coactivator complexes using a covalent small molecule ligand, *Proc. Natl. Acad. Sci. USA 111*, 12061-12066.

122. Daniels, K. G., Tonthat, N. K., McClure, D. R., Chang, Y. C., Liu, X., Schumacher, M. A., Fierke, C. A., Schmidler, S. C., and Oas, T. G. (2014) Ligand concentration regulates the pathways of coupled protein folding and binding, *J. Am. Chem. Soc. 136*, 822-825.

123. Hemsath, L., Dvorsky, R., Fiegen, D., Carlier, M. F., and Ahmadian, M. R. (2005) An electrostatic steering mechanism of Cdc42 recognition by Wiskott-Aldrich syndrome proteins, *Mol. Cell 20*, 313-324.

124. Gianni, S., Dogan, J., and Jemth, P. (2014) Distinguishing induced fit from conformational selection, *Biophys. Chem. 189*, 33-39.

125. Iesmantavicius, V., Dogan, J., Jemth, P., Teilum, K., and Kjaergaard, M. (2014) Helical propensity in an intrinsically disordered protein accelerates ligand binding, *Angew. Chem. Int. Ed. Engl. 53*, 1548-1551.

126. Rogers, J. M., Steward, A., and Clarke, J. (2013) Folding and binding of an intrinsically disordered protein: fast, but not 'diffusion-limited', *J. Am. Chem. Soc. 135*, 1415-1422.

127. Greives, N., and Zhou, H. X. (2014) Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit, *Proc. Natl. Acad. Sci. USA 111*, 10197-10202.

128. Hammes, G. G., Chang, Y. C., and Oas, T. G. (2009) Conformational selection or induced fit: a flux description of reaction mechanism, *Proc. Natl. Acad. Sci. USA 106*, 13737-13741.

129. Bachmann, A., Wildemann, D., Praetorius, F., Fischer, G., and Kiefhaber, T. (2011) Mapping backbone and side-chain interactions in the transition state of a coupled protein folding and binding reaction, *Proc. Natl. Acad. Sci. USA 108*, 3952-3957.

130. Haq, S. R., Chi, C. N., Bach, A., Dogan, J., Engstrom, A., Hultqvist, G., Karlsson, O. A., Lundstrom, P., Montemiglio, L. C., Stromgaard, K., Gianni, S., and Jemth, P. (2012) Side-chain interactions form late and cooperatively in the binding

reaction between disordered peptides and PDZ domains, *J. Am. Chem. Soc. 134*, 599-605.

131. Fersht, A. R., and Sato, S. (2004) Phi-Value analysis and the nature of protein-folding transition states, *Proc. Nat. Acad. Sci. USA 101*, 7976-7981.

132. Borg, M., Mittag, T., Pawson, T., Tyers, M., Forman-Kay, J. D., and Chan, H. S. (2007) Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity, *Proc. Nat. Acad. Sci. USA 104*, 9650-9655.

133. Lee, C. W., Ferreon, J. C., Ferreon, A. C., Arai, M., and Wright, P. E. (2010) Graded enhancement of p53 binding to CREB-binding protein (CBP) by multisite phosphorylation, *Proc. Natl. Acad. Sci. USA 107*, 19290-19295.

134. Lukhele, S., Bah, A., Lin, H., Sonenberg, N., and Forman-Kay, J. D. (2013) Interaction of the eukaryotic initiation factor 4E with 4E-BP2 at a dynamic bipartite interface, *Structure 21*, 2186-2196.

135. Asano, N., Atsuumi, H., Nakamura, A., Tanaka, Y., Tanaka, I., and Yao, M. (2014) Direct interaction between EFL1 and SBDS is mediated by an intrinsically disordered insertion domain, *Biochem Biophys Res Commun 443*, 1251-1256.

136. Cervantes, C. F., Bergqvist, S., Kjaergaard, M., Kroon, G., Sue, S. C., Dyson, H. J., and Komives, E. A. (2011) The RelA nuclear localization signal folds upon binding to IkappaBalpha, *J. Mol. Biol. 405*, 754-764.

137. Nyarko, A., Song, Y., and Barbar, E. (2012) Intrinsic disorder in dynein intermediate chain modulates its interactions with NudE and dynactin, *J. Biol. Chem. 287*, 24884-24893.

138. Cho, S., Swaminathan, C. P., Bonsor, D. A., Kerzic, M. C., Guan, R., Yang, J., Kieke, M. C., Andersen, P. S., Kranz, D. M., Mariuzza, R. A., and Sundberg, E. J. (2010) Assessing energetic contributions to binding from a disordered region in a protein-protein interaction, *Biochemistry 49*, 9256-9268.

139. Ferreon, J. C., and Hilser, V. J. (2004) Thermodynamics of binding to SH3 domains: the energetic impact of polyproline II (PII) helix formation, *Biochemistry 43*, 7787-7797.

140. Noble, C. G., Hollingworth, D., Martin, S. R., Ennis-Adeniran, V., Smerdon, S. J., Kelly, G., Taylor, I. A., and Ramos, A. (2005) Key features of the interaction between Pcf11 CID and RNA polymerase II CTD, *Nat. Struct. Mol. Biol. 12*, 144-151.

Chapter 2

Application of NMR to Studies of Intrinsically Disordered Proteins

[This chapter is modified from a manuscript entitled “Application of NMR to Studies of Intrinsically Disordered Proteins”, which was in revision at the time this dissertation was written. Authors include E. B. Gibbs, E. Cook, and S. A. Showalter. Sections on IDP interactions and the application of PRE to IDP aggregation were written by E. Cook and removed accordingly. Data presented in figure 2-2 was collected by E. Cook and reproduced here with permission]

The prevalence of intrinsically disordered protein regions, particularly in eukaryotic proteins, and their clear functional advantages for signaling and gene regulation have created an imperative for high-resolution structural and mechanistic studies. NMR spectroscopy has played a central role in enhancing not only our understanding of the intrinsically disordered native state, but how that state contributes to biological function. While pathological functions associated with protein aggregation are well established, it has recently become clear that disordered regions also mediate functionally advantageous assembly into high-order structures that mediate formation of membrane-less sub-cellular compartments and even hydrogels. Across the range of functional assembly states accessed by disordered regions, post-translational modifications and regulatory macromolecular interactions that can also be investigated by NMR spectroscopy feature prominently. Here we will explore the many ways in which NMR has advanced our understanding of the physical-chemical phase space occupied by disordered protein regions and provide prospectus for the future role of NMR in this emerging and exciting field.

Introduction

The post-genomic era has been characterized by an explosion of new functional annotations and insights into biomolecular structure. In the past decade, one of the most groundbreaking transitions that has occurred as a result of this new data is the recognition that intrinsically disordered regions (IDRs) in proteins are both prevalent and functionally significant, particularly in eukaryotes [1]. Here we designate any segment of 30 or more contiguous amino acid residues in the primary structure of a protein that lack a temporally stable tertiary structure as an IDR, which is the acronym we will use in preference to intrinsically disordered protein (IDP), as the latter invokes the image of a protein that completely lacks stable tertiary structure throughout its entire length.

Just as the fraction of a given protein's structure found in its IDRs varies widely, so does the extent to which cooperative folding behavior manifests. A general picture has emerged with full disorder and full cooperativity existing primarily as limits, and increased pliability in the continuum between having been recently recognized as a path to allosteric regulation [2]. This contrasts with the usual understanding of IDRs that was prevalent until the current decade, in which IDRs were primarily associated with dysregulation, owing to the well-known capacity of certain IDRs to aggregate and/or form amyloid fibrils. It is now understood that not all higher-order IDR assembly is pathological, but rather that the phase space of proteins with substantial disordered regions spans from soluble forms through gels and liquid-like phases that constitute membrane-less organelles, which serve vital roles in normal cellular function. There is ample reason to extend any exploration of this phase behavior to the cellular environment [3]. Thus, the functional relevance of IDRs and our increased awareness of the complexity of their phase behavior necessitates quantitative evaluation of sequence-to-structure-to-function relationships in this ubiquitous class of proteins.

As a consequence of lacking a temporally stable tertiary structure, IDRs are incapable of contributing directly to enzymatic catalysis. Instead, these regions are uniquely well suited to mediate interactions with other proteins [4]. In analogy to the range of length scales discussed above, interaction motifs contained in IDRs can be large and multi-segmented, but short linear motifs within the primary structure are more prevalent, as are individual sites of post-translational modification (PTM) [5, 6]. Many of the earliest examples of functional roles associated with interaction motifs and PTMs embedded in IDRs came from analysis of transcriptional control, particularly in eukaryotes [7]. It is now clear that dynamically assembled cellular signaling complexes overwhelmingly utilize IDRs to mediate protein-protein interactions [8].

The rapid growth of research into IDRs was made possible by pioneering work within the bioinformatics community, as well as the forward-looking efforts of a small, visionary community of NMR spectroscopists. Recently, IDR research has matured to the point where quantitative biophysical and structural work is generally practical [9]. In this article, we provide an in-depth review of the specific role NMR has played in characterizing the structural and assembly space accessed by IDRs. (Figure 2-1). We will begin with a technical overview of the ways in which standard biomolecular NMR methodology can be successfully applied to IDRs, placing emphasis on the most recent developments. In order to showcase those areas in most pressing need of future investment, we will focus primarily on establishing direct functional roles for IDRs, which is a challenge the NMR community can help to address going forward.

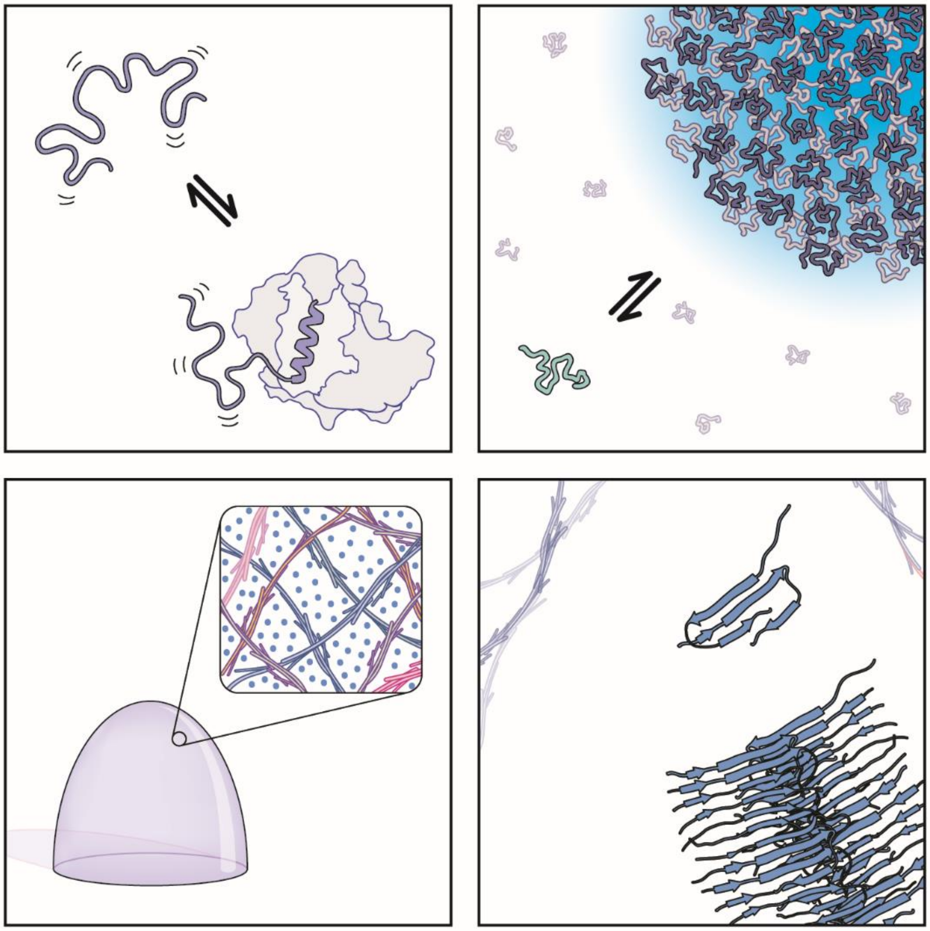


Figure 2-1. Protein disorder supports function across a wide range of native structures and length scales. From individual protein-protein interactions in solution through the support of aqueous phase separation into membrane-less organelles and even hydrogels, intrinsically disordered regions mediate the contacts that are required for biological function. Disorder is also well known for the role of misfolding in amyloid formation and its contribution to disease. NMR spectroscopy has made uniquely quantitative insights into the mechanisms of these diverse sets of interactions.

NMR chemical shift assignment strategies for IDPs

NMR has emerged as the preeminent method to investigate the molecular structure of IDRs, and yet these proteins also present substantial challenges for NMR spectroscopy, because their spectra are generally characterized by extremely low chemical shift dispersion. Reasonably comprehensive chemical shift assignments are a necessary pre-requisite to any biomolecular NMR investigation and it is assumed the reader is familiar with general and widely-adopted strategies. For a detailed contemporary review of the special challenges presented by IDRs and the methods that have been developed to work around them, the reader is referred to Brutscher et al. [10]. Here we will briefly present the emerging themes in the IDR literature that are of greatest interest.

Innovations in proton-detected NMR

Proton-detected NMR of IDRs is challenging because of the limited chemical shift dispersion often encountered in their spectra, as well as their tendency to possess proline residues in abundance. In the best cases, proton-detected ^{1}H,^{15}N-HSQC (or TROSY) spectra of IDRs are characterized by minimal chemical shift dispersion in the proton dimension, but their typically narrow lineshapes still allow for good peak resolution. By way of example, this is certainly the case for the C-terminal domain of the pancreatic transcription factor Pdx1. (Figure 2-2A). As will be discussed below, phosphorylation is an extremely common post-translational modification in IDRs and the characteristic downfield proton shift associated with phosphoserine and phosphothreonine places these resonances into a region of the ^{1}H,^{15}N-HSQC that is typically clear of resonances in IDRs, as is the case for the red pSer268 resonance of Pdx1 in Figure 2A. Given that many IDR samples are mass limited, particular for proteins that assemble into amyloid or other aggregates at high concentration, the sensitivity of proton-detected spectroscopy, often combined with high magnetic field strength, can be crucial for project execution.

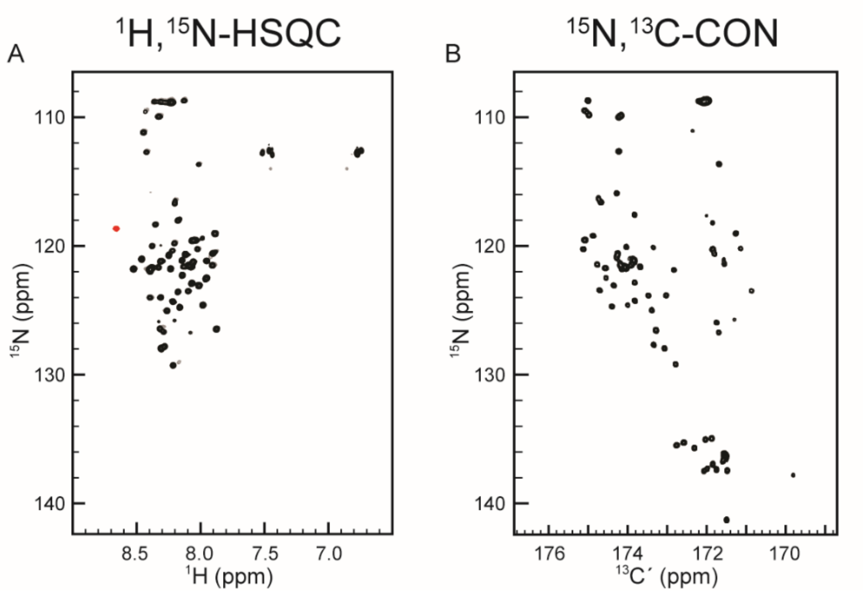


Figure 2-2. Residue-specific resolution is achievable in high-resolution spectra of intrinsically disordered regions. (A) Traditional proton-detected ^{1}H,^{15}N-HSQC spectrum of the intrinsically disordered C-terminal tail from the pancreatic transcription factor Pdx1 overlaid with the same spectrum acquired on a phosphorylated sample (grey spectrum). The resonance for phosphor-Ser268 shows a characteristic downfield shift in the proton dimension. (B) Relatively new carbon direct-detect methods enhance the resolution of disordered protein spectra, while facilitating the study of highly proline-enriched proteins like Pdx1, where the resonances between 134 – 142 ppm in the ^{15}N-dimension all correspond to peptide bonds including proline residues.

Recent innovations in both sample preparation and spectral acquisition have opened new possibilities for efficient chemical shift assignment of IDRs. On the sample side, there is interest in modifying solvent conditions to enhance the efficiency of return to equilibrium, thus accelerating the rate of transient accumulation [11]. It is also possible to use selective amino acid labelling in order to reduce spectral complexity [12]. On the acquisition side, the problems traditionally presented by proline residues have been circumvented by using long range correlation methods that rely on modified Carr-Purcell transfer schemes [13]. In addition, many IDR samples are not stable for long periods of time and so sequential data acquisition using multiple chemical sub-types of a specific nucleus has emerged as an efficient way to acquire multiple 3D spectra in essentially the same amount of spectrometer time traditionally required to acquire one [14]. Finally, as will be discussed further below, NUS methods that accelerate data acquisition by enhancing resolution and/or signal-to-noise per unit of spectrometer time invested have seen wide adoption in the IDR community [15].

Heteronuclear direct-detection NMR

It is well recognized that NMR spectra of IDPs tend to suffer from low chemical shift dispersion in the amide ^{1}H-dimension, leading to poor spectral dispersion. While this problem can be mitigated through the application of ultra-high field for many systems, its ubiquitous nature has also motivated the development of alternative detection modalities. Of particular interest, ^{13}C direct-detect NMR has emerged as a leading alternative to proton-detection for IDPs, owing to the excellent chemical shift dispersion and narrow linewidths generally encountered [10, 16]. The benefits of carbon-detection can clearly be seen in the ^{15}N,^{13}C-CON spectrum of Pdx1-C displayed in Figure 2-2B, which features both high resonance dispersion and the inclusion of resonances corresponding to proline residues. Given that proline is one of the most over-represented amino acid types in IDP sequences, this second advantage should not be under-estimated. Perhaps the most significant driver of ^{13}C direct-detect methods development has been the steady increase in availability of cryogenically cooled probes with enhanced carbon sensitivity, which is the key technological advance needed for this spectroscopy. One barrier to wide adoption of carbon-detection strategies remains the relatively low sensitivity of detection – most commercially purchased cryogenic probes remain inverse-probes optimized for proton sensitivity, not carbon. In contrast, carbon-detection on commercially available cryogenic probes is generally tolerant of a wide range of solution conditions including high salt, basic pH [16], the presence of stabilizing co-solutes and crowding agents, and even cellular extract [17]. As analysis of high-order assembly and IDR behavior in cellular or cell-like environments is an area of intense current interest, the tolerance of carbon direct-detect experiments to diverse co-solute environments is likely to emerge as one of its most valuable features.

Even with the enhanced spectral dispersion generally provided by ^{13}C direct-detection, the tendency of IDRs to feature low sequence complexity, often with repetitive sequence motifs recurring throughout the primary structure, has driven innovation aimed at further enhancing spectroscopic resolution. Spectral editing to generate amino acid-specific sub-spectra of uniformly isotope labeled samples can dramatically accelerate chemical shift assignment [18], and new software specifically tailored to IDR applications has been developed to automate amino acid type prediction [19]. In addition, recent advances in non-uniform sampling and other resolution enhancement techniques have made hyper-dimensional NMR possible, with demonstrated benefits for carbon-detected applications to IDRs [20, 21].

The advantages of heteronuclear direct-detect spectroscopy have more recently stimulated interest in nitrogen-detected NMR as an alternative for IDP applications. The original proof of concept that traditionally carbon-detected experiments could be converted to nitrogen detection offers exciting possibilities to further improve linewidths in ^{15}N,^{13}C-CON and ^{15}N,^{13}C-CAN spectra [22]. More recently, ^{15}N direct-detect TROSY spectra have been reported [23, 24], suggesting future opportunities for growth in this area. Unlike ^{13}C direct-detection biomolecular NMR, ^{15}N direct-detect spectroscopy remains practical only on custom-built cryogenic probes at the time of writing. Thus, broad adoption of this new technology remains a frontier for the future and there is an imperative for hardware manufacturers to put nitrogen-optimized probes into broader distribution.

NMR structural and dynamic constraints for IDPs

While IDRs are by definition disordered compared to their cooperatively folded counterparts, their structures in solution are in many cases far from random. NMR methods amenable to the determination of structural and dynamic constraints for IDPs have been extensively reviewed [25], so this article will only provide a broad overview and focus on recent developments.

Structural constraints

Disordered protein regions tend to be depleted in long-range tertiary contacts, making the structural information encoded in the chemical shift itself, which is particularly well suited to assign α-helical secondary structure, a key indicator of IDR ensemble properties [26]. It has been observed that a significant fraction of IDRs possess (partially formed) α-helical segments that are disproportionately involved in mediating protein-protein interactions [27, 28]. Identification of secondary structures other than α-helix, or alternatively the positive identification of a lack of defined secondary structure, is often more easily made through the collection of large scalar coupling datasets [29]. For proline-rich proteins, carbon direct-detect strategies readily facilitate secondary structure assignment [18, 30]. As the quantity of available secondary structure information for IDRs has grown, a tendency has emerged to reach the conclusion that if secondary structure is present, it must be important for function; but a growing number of studies cast doubt on the generality of this assumption. For example, it is the length of the disordered linker in the transcription factor Notch that correlates with strength of transcriptional activation, not the presence or absence of helical segments in its IDR [31]. This counter-example underscores the urgency of coupling NMR investigations to mechanistic and functional investigations in order to arrive at a complete picture of the system.

In addition to chemical shifts, residual dipolar couplings (RDCs) have emerged as strong constraints on IDR ensembles [32]. Historically, RDCs have been used most aggressively to constrain the structures of IDRs that that possess α-helical segments. In an excellent recent counter-example to this trend, Janowska and Baum used RDCs to demonstrate structural differences between α-synuclein and β-synuclein that correlate with the propensity of the respective proteins to form amyloid fibrils [33]. Most importantly, this study provided a direct mechanistic rationale for a propensity to dementia in patients harboring a specific missense mutation in β-synuclein, which RDCs showed stimulates a transition to a monomeric solution structure more reminiscent of α-synuclein in the point mutant.

Although IDRs by definition lack a temporally static tertiary structure, their ensembles very often feature dynamic tertiary features, such as transient long-range contacts of even molten globule-like structures. While chemical shifts and RDCs both provide helpful constraints on IDR ensembles, neither measurement is well suited to establish transient long range contacts and low-population tertiary structures. In contrast, the phenomenon of paramagnetic relaxation enhancement (PRE) has emerged as a reliable constraint on long-range structure for IDRs [25]. As with RDC measurements, caution must be exercised because the IDR ensemble can feature multiple transient contacts, yielding mutually incompatible constraints, or masking minor states. Considerable effort has been put into solving this problem, including the recent establishment of a protocol to combine RDC measurements with PRE data for validation using a strategy that explicitly accounts for the potential impact of tagging and anisotropic alignment on IDR structure [34].

Modelling IDP dynamics

Conformational dynamics are an integral component of the ensemble description of IDRs, but of course their extensive dynamics, compared to cooperatively folded domains, produces many of the same challenges as those discussed above for structure constraint. For example, α-helical structure and long-range tertiary contacts both increase local contact density and therefore have the same effect on measured T2 times. [35] Of course, the influence of helical structure can be independently verified through assessment of chemical shifts, thus resolving the ambiguity. One excellent example of this synergy is a recent study of the disordered C-terminal domain from the Nipah virus nucleoprotein, in which ^{15}N spin-relaxation and carbon chemical shift data were combined to describe the sampling of α-helical conformations in a region known to mediate protein-protein interactions [36].

As discussed above, carbon direct-detect NMR offers an alternative to traditional proton-detect biomolecular NMR that often provides enhanced spectral dispersion, which is especially critical for quantitative spin relaxation measurements. The past several years have seen the development of carbon direct-detect experiments suitable to measure spin relaxation of carbon [37], nitrogen [38], and hydrogen nuclei in IDPs [39].

One limitation that has previously hindered the wide-spread analysis of IDR dynamics by NMR is the fact that widely accepted formalisms, such as the Lipari-Szabo Model free description of biomolecular dynamics [40], are not strictly applicable to disordered proteins. As has become increasingly common in the folded-protein community, utilization of the heteronuclear NOE as a reporter of backbone flexibility, without downstream modelling, has gained wide acceptance [31, 41]. In addition, while spectral density mapping [42, 43] has never overtaken Lipari-Szabo analysis in the cooperatively folded protein community, it has emerged as a rigorous means to connect spin relaxation and conformational dynamics in IDRs [44, 45]. Collection of massive spin-relaxation data sets spanning magnetic field strengths from 9.4 – 23.5 T enabled direct spectral density mapping that demonstrated a wide range of dynamic timescales sampled by the disordered Engrailed transcription factor [46]. For partially ordered segments of IDRs, such as the α-helical segment of the Sendai virus nucleoprotein's disordered tail, dynamics on the 10-100 ns timescale can be critically important. By collecting temperature-dependent spin relaxation data sets at multiple fields, the Blackledge group was able to gain quantitative access to dynamics occurring in Sendai virus nucleoprotein on this timescale [47].

Monitoring post-translational modifications by NMR

Post-translational modifications (PTMs) allow proteins to undergo rapid and reversible changes in structure and function in response to internal and external stimuli. This process serves key roles in transcription, translation, cell cycle control, and the immune response. PTMs tend to be more prevalent in disordered regions [48], which stems from the inherent amino acid sequence bias toward chemically labile side chains in IDRs and their structural plasticity, which provides greater access to modifying enzymes.

The conformational heterogeneity of IDRs can facilitate interactions with multiple binding partners. In this context, PTMs serve as an important regulator of IDR function and can effect binding through a variety of mechanisms [49]. Recent studies, employing various NMR techniques, have shown that PTMs can influence IDR binding by promoting changes in dynamics [50], charge-charge interactions [51], and secondary structure propensities [52, 53]. The effects of PTMs on the structural properties of IDPs vary widely [53-55]; however, the covalent modification of amino acid side chains often produces characteristic perturbations to nearby spin systems [56], leading NMR to become one of the foremost techniques applied within the PTM community (along with mass spectrometry). Here we will focus on NMR as a tool to investigate the role of phosphorylation and acetylation specifically.

Phosphorylation

One of the most well studied PTMs is phosphorylation, which most commonly occurs on the side chains of serine, threonine, and tyrosine residues. Addition of a phosphate group to Ser/Thr sidechains strongly perturbs nearby ^{1}HN, ^{13}Cα/β resonances and assignments are typically performed using standard 2D ^{1}H,^{15}N correlation spectra and ^{1}H-detect triple resonance experiments [57]. However, methods exploiting the ^{31}P-^{13}C scalar coupling have also been developed [58].

Multi-site phosphorylation, in particular, is prevalent in IDPs that are involved in signal transduction. And time resolved analysis of modified amino acid resonances using real-time NMR (RT-NMR) allows for the extraction of kinetic rate constants, providing a basis for mechanism determination [57, 59, 60]. For example, Cordier, et al. used a combination of RT-NMR to decipher the ordered and distributive phosphorylation of the tumor suppressor PTEN by CKII and GSK3β [61], which occurred in two independent cascades. And more recently, Mylona et al. [59] examined the distributive multi-site phosphorylation of the transcription factor Elk-1 by ERK2, which partitioned into fast (F), intermediate (I), and slow (S) sites. Using a mouse embryonic fibroblast model system, the authors went on to show that mutation of F and I sites has antagonizing effects on target gene expression, while mutation of S sites dramatically increases target gene expression and cell proliferation, suggesting that multi-site phosphorylation of Ets-1 by ERK2 effectively becomes self-limiting in the absence of antagonizing phosphatase activity. These examples highlight the range of mechanisms used in IDR signaling, and also the utility of RT-NMR for mechanistic studies of IDR multi-site phosphorylation.

In addition to the regulation of IDRs in the context of healthy cells, multi-site phosphorylation is a hallmark of viral infection. Bioinformatics studies have identified intrinsic disorder as a prevalent feature of viral proteomes [62]. In this context, intrinsic disorder maximizes the functional repertoire of the viral proteome when genome size becomes limiting. Viral proteins must be able to interact with a range of viral and host molecules over the course of the viral life cycle, all while avoiding detection by the host. As such, viruses often hijack the host regulatory machinery and phosphorylation is used to modulate interactions with host proteins. Recently, several groups have utilized NMR to study the unstructured proteins encoded by the Hepatitis C Virus [60, 63-66]. The Nonstructural Protein 5A (NS5A), in particular has roles in genome replication and virion assembly and is regulated through multi-site phosphorylation [67]. Investigations into the phosphorylation of NS5A by CKII have been conducted using RT-NMR, revealing phosphorylation events at both canonical and non-canonical sites [60, 66]. Interestingly, these modifications proceeded at markedly different rates and resulted in long-range chemical shift perturbations to residues in remote transiently formed helices and low complexity PPII motifs [60]. Each affected region contains positively charged residues, suggesting that phosphorylation induces structural and/or dynamic changes to NS5A mediated through charge-charge interactions. Similarly, using ^{13}C direct-detect NMR, we were able to show that phosphorylation of T2332 by PKA alters the dynamics of NS5A at an adjacent PPII motif and directly monitor c-Src SH3 binding at this site, thus characterizing a PPI that is important for virus replication [64]. Thus, NMR offers a framework for the identification and characterization of novel NS5A phosphorylation events, thus increasing our understanding of how viruses utilize PTMs to maximize function while minimizing genome size.

Acetylation

Lysine side-chain acetylation is a common and reversible post-translational modification, recognized by bromodomains and other proteins, that regulates protein-protein interactions and through them multiple cellular functions [68]. Particularly in eukaryotic organisms, N-terminal acetylation is also employed to control protein structure and sub-cellular localization, as is the case for α-synuclein, which has recently been observed in cells using NMR [69, 70] (see Section 8 below). The classic and best understood example of protein regulation through lysine acetylation is the epigenetic marking of histone tails. Several years ago, Dose et al. demonstrated the ability to study histone acetylation and de-acetylation kinetics in situ using high-resolution NMR [71]. Excitingly, these studies enabled the investigators to demonstrate using cellular extract as the NMR "solvent" that histone deacetylase activity is far more abundant than previously thought [71]. Finally, many proteins that are post-translationally modified accept multiple modifications in the course of tightly regulated and often competitive signaling events. The exquisite site-resolution inherent to NMR spectroscopy has allowed simultaneous detection of both phosphorylation and acetylation on histone protein H3 [72, 73], demonstrating that it is feasible to quantify the mechanism of co-regulation by distinct PTMs on individual disordered proteins.

NMR studies of aqueous phase separation

Following the wide-spread acceptance that intrinsic protein disorder can mediate signal transduction throughout the cell, an increasing body of work has also linked structural disorder to sub-cellular compartmentalization via the formation of membrane-less organelles. Recent studies suggest that these bodies form through liquid-liquid phase separation (LLPS), mediated by a variety of protein-protein and protein-nucleic acid interactions. While not unique to disordered proteins, LLPS is driven in part by low complexity (LC) amino acid sequences, suggesting that the sequence characteristics which promote structural disorder may also promote phase separation. In addition, mutations to LC regions have been implicated in several human diseases, suggesting that phase separation and/or aberrant regulation of the assemblages within liquid-like bodies may play a role in pathology. Unsurprisingly, NMR has played an active role in current efforts focused on elucidating the molecular mechanisms that underlie disorder-mediated LLPS.

Characterization of Liquid-Liquid Phase Separation (LLPS)

In principle, LC sequences may engage in many types of intermolecular interactions, including multivalent, π-π, cation-π, and dipole-dipole interactions [74]. And in analogy to the previously discussed SLiMs, multiple LC sequences are often found embedded within larger disordered domains. One leading question is how the linear arrangement of LC motifs relates to interactions that drive phase separation. Chemical shift perturbations have permitted the identification of interaction sites [75-77] involving LC domains and the extraction of site-specific dissociation constants [78] for those interactions, both in dilute and condensed droplet states, revealing the subtle ways in which LC domains engage in intermolecular interactions during droplet assembly. For example, in the liquid-liquid phase separated FUS LC, chemical shift perturbations indicated that self-interactions were randomly distributed amongst multiple LC motifs throughout the domain [75]; by contrast, the C-terminus of TDP43 localized PPIs to specific LC sequences [76]. Finally, secondary chemical shifts have shown that some IDRs undergo local folding upon assembly in the droplet state [76], demonstrating that many of the same guiding principles required to understand IDR-mediated interactions in dilute phases are also relevant to the LLPS state.

In many other ways, the physical state of proteins in the droplet state can be substantially different from that of dilute solution. In particular, during LLPS the density of proteins, RNAs and other biomolecules increases dramatically, which can lead to drastic increases in the local viscosity. Simultaneously, the open nature of the membrane-less organelle permits the rapid diffusion of molecules across the phase boundary, thus setting these bodies apart from other assemblages like hydrogels, aggregates or amyloids. Further, differences in the dynamic behavior of liquid-like bodies have been observed, potentially reflecting the relevant time scales for biological processes. Hence, there is much interest in connecting the bulk viscometric and diffusive properties of liquid-like bodies to IDR dynamics on the molecular scale. To this end, a variety of NMR dynamics measurements have been utilized, including pulse field gradient (PFG) diffusion [91], nuclear spin relaxation [75, 76, 78], paramagnetic relaxation enhancement [76], and relaxation dispersion [75, 76].

While the molecular motions of IDRs engaging in transient intermolecular interactions within the liquid-like state often complicates measurement and interpretation of dynamic parameters, the careful analysis of NMR dynamic parameters measured in dilute, ligand bound and phase separated states has provided key insights for several systems. In a recent study of NPM1, a homopentameric Nucleolar protein containing acidic LC motifs that mediate phase separation through interactions with Arg-rich proteins and rRNAs, nuclear spin relaxation was used to compare the dynamic behavior of the folded pentameric core and disordered LC domains during phase separation. through measurement of the of R_1 and the transverse ^{15}N chemical-shift anisotropy (CSA)/dipole-dipole cross-correlation relaxation rates (η_{xy}), Mitrea et. al. were able to separate the contributions from the core and the LC motifs obtaining local average correlation times ($\tau_{c,local}$) [78]. Under dilute conditions the LC domain experienced fast local motions on the ns timescale which slowed considerably at saturating Arg-rich peptide concentrations. Similarly, Burke et. al. showed that upon phase separation, backbone nuclei in FUS experienced significant decreases in R_1 with increases in R_2 and heteronuclear NOEs, suggesting that the average reorientational motions of the protein were hindered within the liquid droplet state [75].

These studies demonstrate the feasibility of applying NMR to investigate IDRs in protein rich liquid droplet-like states and highlight how LLPS can alter the structural properties and timescales of intramolecular motions. Undoubtedly, NMR will continue to play an important role in elucidating the molecular determinants of large scale organization within the liquid-like bodies that form membrane-less organelles.

Hydrogel formation

In addition to forming liquid-liquid phase separated states, LC sequences have long been known to form hydrogels *in vitro*. Over the years, these assemblies have been widely investigated for their material properties, therapeutic applications, and to gain insights into the process of fibrilization [80-82]. The semi-solid nature of hydrogels makes solid-state NMR uniquely suited to investigate the structural and dynamic properties of peptides [83,84] and proteins [85-87] in the hydrogelated state. Particularly for IDR hydrogels, the combined use of through-space (cross-polarization; CP) and through-bond (scalar coupling) magnetization transfer schemes permits the detection of highly rigid and dynamic regions of the assembly, respectively. For example, in a comparative study of nucleoporin phenylalanine-glycine repeat domains (FG NUPs), IDRs from the nuclear pore complex that form hydrogels with distinct molecular sieve-like properties, these techniques were applied to probe the molecular details of hydrogel assembly. Surprisingly, key differences could be observed between the *S. cerevisiae* Nsp1 FG hydrogel, which assembles through β-sheet rich fibers [85], and the *Xenopus* Nup98 FG hydrogel, which assembled through structures with a reduced β-sheet propensity that could be modulated by glycosylation, thus explaining their relative differences in transport factor permeability. Further, Dannatt et. al. used high-resolution solid-state NMR (ssNMR) with magic-angle spinning (MAS) to study *E. coli* ssDNA-binding protein (SSB) in hydrogels and in hydrated solids produced from centrifugal sedimentation [87]. Comparison of cross-polarization and scalar-coupling based heteronuclear correlation experiments (CP-HSQC and J-HSQC, respectively), collected on SSB in the hydrogelated state and as a hydrated solid in complex with ssDNA, enabled the assignment of the highly dynamic C-terminal IDR. And, through ^{15}N $R_{1\rho}$ measurements the authors were able to probe transient interactions between an acidic protein interaction motif within the IDR and the DNA binding groove, thus confirming a long-standing model for SSB self-inhibition.

Applications of NMR to aggregation-prone IDPs and their amyloid fibers

The aberrant folding or aggregation of IDPs has been implicated in the progression of several neurodegenerative diseases including, dementia (tau protein), Alzheimer’s disease (Aβ), and Parkinson’s (α-synuclein). Correlated with the loss of neuronal function is the intracellular accumulation of protein deposits, either as large amorphous oligomeric species or as long and highly ordered fibrous assemblies known as amyloids. While the precise identity of the pathological species in these diseases are widely debated, the elucidation of the molecular mechanisms underlying disease progression has placed a strong imperative on studying the assembly of IDPs into insoluble forms. The role of NMR in establishing the nature of the assembled states and the mechanisms of their formation has been the subject of several recent and extensive reviews [88-91]. For completeness, we provide a brief overview of the application of NMR spectroscopy to aggregation prone IDPs here.

Recent efforts have focused on connecting the kinetics of aggregation to molecular descriptions of aggregation prone IDPs in their monomeric states. The use of real-time NMR provides a complimentary technique to traditional Thioflavin T fluorescence assays [29,92]. In particular, the high sensitivity and reduced spectral complexity offered by direct detection of the ^{19}F nucleus has enabled monitoring of conformational changes and fibrillation kinetics for several amyloid proteins [92,93].

In addition, many of the solution state techniques discussed in earlier sections have been exploited to provide molecular descriptions of secondary structure [29, 33, 94], tertiary contacts [95] and intramolecular dynamics [92, 96]. For example, Bai et al. characterized the initial ensemble structures of α-synuclein in the presence or absence of 150 mM NaCl using an ensemble approach combining real-time ^{19}F NMR, ^{19}F & ^{15}N nuclear spin relaxation, PRE and PFG measurements. This revealed the presence of two distinct conformations, with a more compact and rigid ensemble with faster fibrilization kinetics dominating in the absence of NaCl [92].

In cell NMR of IDPs

Toward the broad goal of understanding protein structure and function within the complex cellular environment, recent developments have enabled the high-resolution

NMR measurements of proteins in live cells [3, 97]. It has been recognized that the cellular environment, which slows molecular tumbling and provides myriad non-specific interactions, can cause excessive line broadening for cooperatively folded proteins. By contrast, the dynamic properties of IDPs and IDRs provide enhanced spectral quality in cellular environments [3, 98]. In-cell NMR, thus provides an exceptional opportunity to study the structural properties of IDPs under native conditions where the effects of molecular crowding and interactions with endogenous factors on structure and dynamics are not known a priori.

Recently, several groups have investigated the effects of cellular environments of the structure and dynamics of α-Synuclein [69, 70, 99]. By comparing the rates of amide proton exchange measured in buffer and in *E. coli*, Smith et al. showed that α-Synuclein retains a level of structural disorder comparable to unstructured peptides in the bacterial cytoplasm [70]. More recently, Theillet, et. al. used in-cell NMR and EPR techniques to characterize the structure and dynamics of α-Synuclein in multiple non-neuronal and neuronal cell lines. This revealed that α-Synuclein predominantly exists in a monomeric, disordered, and relatively compact conformation, which shields hydrophobic residues in the amyloidogenic NAC domain, suggesting that large conformational rearrangements may proceed oligomerization [69].

One limitation of in-cell methods is the need for protein overexpression, which may saturate the pool of interaction partners effectively rendering the contributions from key cellular interactions a minor fraction of the total detectable signal. To address this issue, several groups have leveraged DNP MAS NMR [100-102], where microwave irradiation is used to transfer polarization of the large electron spin to relatively insensitive nuclei, thus obtaining considerable signal enhancements. Although these experiments are performed in cellular lysates, the boost in sensitivity permits measurement of the target proteins at endogenous protein concentrations. In a recent study by Frederick, et al., this technique was used to interrogate the yeast prion Sup35, revealing that a region that is highly disordered in purified samples assumed β-sheet structure in lysates, presumably through interactions with chaperone proteins [101]. This lead the authors to conclude that PPIs within cellular milieu are important for defining the native structure of the IDR.

Conclusions and prospectus

Twenty years ago, biologically appropriate protein function was almost exclusively thought to be associated with the achievement of a spatially and temporally stable tertiary structure by the polypeptide chain. Since then, a dramatically growing body of evidence has led to a re-formulation of the protein structure-function paradigm to assert that all polypeptide chains, under native conditions, demonstrate native functions that arise from their unique amino acid sequence and the structural ensemble it imparts. Owing to its unique ability to not only tolerate, but quantitatively monitor molecular conformational dynamics, NMR spectroscopy has played a leading role in developing our modern understanding of the physical-chemical properties of intrinsically disordered protein regions, and how these properties are uniquely leveraged for function in biology. Both low sequence complexity and abundant internal dynamics do tend to reduce the chemical shift dispersion of disordered protein samples, leading to challenges applying high-resolution NMR spectroscopy to their study. Fortunately, the growth of interest in protein disorder among the biophysical community has corresponded in time with the development of enhanced hardware and data processing techniques that have made NMR spectroscopy of these exciting systems far more practical. As we have outlined in this review, the outcome of these improvements is that NMR is now poised to contribute fundamental mechanistic knowledge, particularly when paired with biochemical, cellular, and/or organism-level studies. For example, when studying post-translational modifications, the specificity of modifying enzymes in vitro is always a concern. A few groups have developed cell based strategies to recombinantly express isotopically enriched IDPs complete with site-specific PTMs [55]. This area is fertile ground for future investigations into the structural and functional impact of PTMs in disordered regions.

Equal to NMR's role in generating insight into protein chemistry and biology is its potential to establish clear mechanistic understanding of the new regulatory and functional roles recently attributed to liquid-liquid phase separation and the formation of "membrane-less" organelles in cells. As in dilute aqueous solution, the molecular motions accessed by IDRs in liquid droplet states are complex and may require the measurement and deconvolution of dynamics that span many timescales. Although the application of NMR dynamic measurement to IDPs that undergo LLPS is still in the early stages, the few examples in the literature demonstrate the feasibility of these techniques and highlight the need for careful experimental design and analysis.

References

1. P. Tompa, (2012) Intrinsically disordered proteins: a 10-year recap, Trends Biochem. Sci. 37 509-16.
2. V. Munoz, L. A. Campos, M. Sadqi, (2016) Limited cooperativity in protein folding, Curr. Opin. Struct. Biol. 36 58-66.
3. E. Luchinat, L. Banci, A (2016) Unique Tool for Cellular Structural Biology: In-cell NMR, J. Biol. Chem. 291 3776-84.
4. P. Tompa, E. Schad, A. Tantos, L. Kalmar (2015) Intrinsically disordered proteins: emerging interaction specialists, Curr. Opin. Struct. Biol. 35. 49-59.
5. P. Tompa, N. E. Davey, T. J. Gibson, M. M. Babu, (2014) A million peptide motifs for the molecular biologist, Mol. Cell 55. 161-9.
6. R. B. Berlow, H. J. Dyson, P. E. Wright (2015) Functional advantages of dynamic protein disorder, FEBS Lett. 589. 2433-40.
7. M. Fuxreiter, P. Tompa, I. Simon, V. N. Uversky, J. C. Hansen, F. J. Asturias (2008) Malleable machines take shape in eukaryotic transcriptional regulation, Nat. Chem. Biol. 4. 728-737.
8. P. E. Wright, H. J. Dyson (2015) Intrinsically disordered proteins in cellular signaling and regulation, Nat. Rev. Mol. Cell Biol. 16. 18-29.
9. E. B. Gibbs, S. A. Showalter (2015) Quantitative biophysical characterization of intrinsically disordered proteins, Biochemistry 54. 1314-26.
10. B. Brutscher, I. C. Felli, S. Gil-Caballero, T. Hosek, R. Kummerle, A. Piai, et al. (2015) NMR Methods for the Study of Instrinsically Disordered Proteins Structure, Dynamics, and Interactions: General Overview and Practical Guidelines, Advances in experimental medicine and biology 870. 49-122.
11. N. A. Oktaviani, M. W. Risor, Y. H. Lee, R. P. Megens, D. H. de Jong, R. Otten, et al. (2015) Optimized co-solute paramagnetic relaxation enhancement for the rapid NMR analysis of a highly fibrillogenic peptide, J. Biomol. NMR 62. 129-42.
12. A. Dubey, R. V. Kadumuri, G. Jaipuria, R. Vadrevu, H. S. Atreya (2016) Rapid NMR Assignments of Proteins by Using Optimized Combinatorial Selective Unlabeling, Chembiochem 17. 334-40.
13. Y. Yoshimura, N. V. Kulminskaya, F. A. Mulder (2015) Easy and unambiguous sequential assignments of intrinsically disordered proteins by correlating the backbone 15N or 13C' chemical shifts of multiple contiguous residues in highly resolved 3D spectra, J. Biomol. NMR 61. 109-21.
14. N. Goradia, C. Wiedemann, C. Herbst, M. Gorlach, S. H. Heinemann, O. Ohlenschlager, et al. (2015) An approach to NMR assignment of intrinsically disordered proteins, Chemphyschem 16. 739-46.
15. C. Wiedemann, N. Goradia, S. Hafner, C. Herbst, M. Gorlach, O. Ohlenschlager, et al., (2015) HN-NCA heteronuclear TOCSY-NH experiment for (1)H(N) and (15)N sequential correlations in ((13)C, (15)N) labelled intrinsically disordered proteins, J. Biomol. NMR 63. 201-12.

16. M. Bastidas, E. B. Gibbs, D. Sahu, S. A. (2015) Showalter, A primer for carbon-detected NMR applications to intrinsically disordered proteins in solution, Con. Magn. Reson. A 44. 54-66.
17. C. W. Lawrence, A. Bonny, S. A. Showalter, (2011) The disordered C-terminus of the RNA Polymerase II phosphatase FCP1 is partially helical in the unbound state, Biochem. Biophys. Res. Commun. 410. 461-465.
18. D. Sahu, M. Bastidas, S. A. Showalter, (2014) Generating NMR Chemical Shift Assignments of Intrinsically Disordered Proteins using Carbon-Detected NMR Methods, Anal. Biochem. 449. 17-25.
19. A. Piai, L. Gonnelli, I. C. Felli, R. Pierattelli, K. Kazimierczuk, K. Grudziaz, et al., (2016) Amino acid recognition for automatic resonance assignment of intrinsically disordered proteins, J. Biomol. NMR 64. 239-53.
20. P. Dziekanski, K. Grudziaz, P. Jarvoll, W. Kozminski, A. Zawadzka-Kazimierczuk, (2015) (13)C-detected NMR experiments for automatic resonance assignment of IDPs and multiple-fixing SMFT processing, J. Biomol. NMR 62. 179-90.
21. S. Zerko, W. Kozminski (2015) Six- and seven-dimensional experiments by combination of sparse random sampling and projection spectroscopy dedicated for backbone resonance assignment of intrinsically disordered proteins, J. Biomol. NMR 63. 283-90.
22. K. Takeuchi, G. Heffron, Z. Y. Sun, D. P. Frueh, G. Wagner (2010) Nitrogen-detected CAN and CON experiments as alternative experiments for main chain NMR resonance assignments, J. Biomol. NMR 47. 271-82.
23. K. Takeuchi, H. Arthanari, I. Shimada, G. Wagner (2015) Nitrogen detected TROSY at high field yields high resolution and sensitivity for protein NMR, J. Biomol. NMR 63. 323-31.
24. K. Takeuchi, H. Arthanari, M. Imai, G. Wagner, I. Shimada (2016) Nitrogen-detected TROSY yields comparable sensitivity to proton-detected TROSY for non-deuterated, large proteins under physiological salt conditions, J. Biomol. NMR 64. 143-51.
25. R. Schneider, J. R. Huang, M. Yao, G. Communie, V. Ozenne, L. Mollica, et al., (2012) Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy, Mol. Biosyst. 8. 58-68.
26. M. R. Jensen, L. Salmon, G. Nodet, M. Blackledge, (2010) Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts, J. Am. Chem. Soc. 132. 1270-2.
27. K. Van Roey, B. Uyar, R. J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, et al., (2014) Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation, Chem. Rev. 114. 6733-78.
28. A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker, et al., (2006) Analysis of molecular recognition features (MoRFs), J. Mol. Biol. 362. 1043-59.
29. J. Roche, Y. Shen, J. H. Lee, J. Ying, A. Bax, (2016) Monomeric Abeta(1-40) and Abeta(1-42) Peptides in Solution Adopt Very Similar Ramachandran Map Distributions That Closely Resemble Random Coil, Biochemistry 55. 762-75.

30. A. Piai, E. O. Calcada, T. Tarenzi, A. D. Grande, M. Varadi, P. Tompa, et al., (2016) Just a Flexible Linker? The Structural and Dynamic Properties of CBP-ID4 Revealed by NMR Spectroscopy, Biophys. J. 110. 372-81.
31. K. P. Sherry, S. E. Johnson, C. L. Hatem, A. Majumdar, D. Barrick, (2015) Effects of Linker Length and Transient Secondary Structure Elements in the Intrinsically Disordered Notch RAM Region on Notch Signaling, J. Mol. Biol. 427. 3587-97.
32. L. Salmon, M. Blackledge, (2015) Investigating protein conformational energy landscapes and atomic resolution dynamics from NMR dipolar couplings: a review, Rep. Prog. Phys. 78. 126601.
33. M. K. Janowska, J. Baum, (2016) The loss of inhibitory C-terminal conformations in disease associated P123H beta-synuclein, Protein Sci. 25. 286-94.
34. F. N. Newby, A. De Simone, M. Yagi-Utsumi, X. Salvatella, C. M. Dobson, M. Vendruscolo, (2015) Structure-Free Validation of Residual Dipolar Coupling and Paramagnetic Relaxation Enhancement Measurements of Disordered Proteins, Biochemistry 54. 6876-86.
35. J. A. Marsh, J. D. Forman-Kay, (2011)Ensemble modeling of protein disordered states: Experimental restraint contributions and validation, Proteins.
36. L. Baronti, J. Erales, J. Habchi, I. C. Felli, R. Pierattelli, S. Longhi, (2015) Dynamics of the intrinsically disordered C-terminal domain of the nipah virus nucleoprotein and interaction with the x domain of the phosphoprotein as unveiled by NMR spectroscopy, Chembiochem 16. 268-76.
37. I. Bertini, I. C. Felli, L. Gonnelli, M. V. V. Kumar, R. Pierattelli, High-Resolution Characterization of Intrinsic Disorder in Proteins: Expanding the Suite of (13)C-Detected NMR Spectroscopy
38. C. W. Lawrence, S. A. Showalter, (2012) Carbon-Detected N-15 NMR Spin Relaxation of an Intrinsically Disordered Protein: FCP1 Dynamics Unbound and in Complex with RAP74, J. Phys. Chem. Lett. 3. 1409-1413.
39. T. Hosek, S. Gil-Caballero, R. Pierattelli, B. Brutscher, I. C. Felli, (2015) Longitudinal relaxation properties of (1)H(N) and (1)H(alpha) determined by direct-detected (13)C NMR experiments to study intrinsically disordered proteins (IDPs), J. Magn. Reson. 254. 19-26.
40. G. Lipari, A. Szabo, (1982) Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 1. Theory and Range of Validity, J. Am. Chem. Soc. 104. 4546-4559.
41. L. Deshmukh, R. Ghirlando, G. M. Clore, (2015) Conformation and dynamics of the Gag polyprotein of the human immunodeficiency virus 1 studied by NMR spectroscopy, Proc. Natl. Acad. Sci. U.S.A. 112. 3374-9.
42. J. W. Peng, G. Wagner, (1992) Mapping of the spectral densities of N-H bond motions in eglin c using heteronuclear relaxation experiments, Biochemistry 31. 8571-86.
43. J. W. Peng, G. Wagner, (1995) Frequency spectrum of NH bonds in eglin c from spectral density mapping at multiple fields, Biochemistry 34. 16733-52.
44. M. L. Gill, R. A. Byrd, A. G. Palmer, (2016) Dynamics of GCN4 facilitate DNA interaction: a model-free analysis of an intrinsically disordered region, Physical Chemistry Chemical Physics 18. 5839-5849.

45. V. To, E. Dzananovic, S. A. McKenna, J. O'Neil, (2016) The Dynamic Landscape of the Full-Length HIV-1 Transactivator of Transcription, Biochemistry 55. 1314-1325.
46. S. N. Khan, C. Charlier, R. Augustyniak, N. Salvi, V. Dejean, G. Bodenhausen, et al., (2015) Distribution of Pico- and Nanosecond Motions in Disordered Proteins from Nuclear Spin Relaxation, Biophys. J. 109. 988-99.
47. A. Abyzov, N. Salvi, R. Schneider, D. Maurin, R. W. Ruigrok, M. R. Jensen, et al., (2016) Identification of Dynamic Modes in an Intrinsically Disordered Protein Using Temperature-Dependent NMR Relaxation, J. Am. Chem. Soc. 138. 6240-51.
48. F. X. Theillet, A. Binolfi, T. Frembgen-Kesner, K. Hingorani, M. Sarkar, C. Kyne, et al., (2014) Physicochemical properties of cells and their effects on intrinsically disordered proteins (IDPs), Chem. Rev. 114. 6661-714.
49. A. Bah, J. D. Forman-Kay, (2016) Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications, J. Biol. Chem. 291. 6696-705.
50. H. Shimojo, A. Kawaguchi, T. Oda, N. Hashiguchi, S. Omori, K. Moritsugu, et al., (2016) Extended string-like binding of the phosphorylated HP1alpha N-terminal tail to the lysine 9-methylated histone H3 tail, Scientific reports 6. 22527.
51. M. Schwalbe, H. Kadavath, J. Biernat, V. Ozenne, M. Blackledge, E. Mandelkow, et al., (2015) Structural Impact of Tau Phosphorylation at Threonine 231, Structure 23. 1448-58.
52. Y. Huang, M. K. Yoon, S. Otieno, M. Lelli, R. W. Kriwacki, (2015) The activity and stability of the intrinsically disordered Cip/Kip protein family are regulated by non-receptor tyrosine kinases, J. Mol. Biol. 427. 371-86.
53. A. Bah, R. M. Vernon, Z. Siddiqui, M. Krzeminski, R. Muhandiram, C. Zhao, et al., (2015) Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch, Nature 519. 106-9.
54. E. W. Martin, A. S. Holehouse, C. R. Grace, A. Hughes, R. V. Pappu, T. Mittag, (2016) Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation, J. Am. Chem. Soc.
55. K. P. Wall, M. Pagratis, G. Armstrong, J. L. Balsbaugh, E. Verbeke, C. G. Pearson, et al., (2016) Molecular Determinants of Tubulin's C-Terminal Tail Conformational Ensemble, ACS Chem. Biol.
56. F. X. Theillet, C. Smet-Nocca, S. Liokatis, R. Thongwichian, J. Kosten, M. K. Yoon, et al., (2012) Cell signaling, post-translational protein modifications and NMR spectroscopy, J. Biomol. NMR 54. 217-36.
57. F. Cordier, A. Chaffotte, N. Wolff, (2015) Quantitative and dynamic analysis of PTEN phosphorylation by NMR, Methods 77-78. 82-91.
58. L. P. McIntosh, H. S. Kang, M. Okon, M. L. Nelson, B. J. Graves, B. Brutscher, (2009) Detection and assignment of phosphoserine and phosphothreonine residues by (13)C-(31)P spin-echo difference NMR spectroscopy, J. Biomol. NMR 43. 31-7.
59. A. Mylona, F. X. Theillet, C. Foster, T. M. Cheng, F. Miralles, P. A. Bates, et al., (2016) Opposing effects of Elk-1 multisite phosphorylation shape its response to ERK activation, Science 354. 233-237.
60. Z. Solyom, P. Ma, M. Schwarten, M. Bosco, A. Polidori, G. Durand, et al., (2015) The Disordered Region of the HCV Protein NS5A: Conformational Dynamics, SH3 Binding, and Phosphorylation, Biophys. J. 109. 1483-96.

61. F. Cordier, A. Chaffotte, E. Terrien, C. Prehaud, F. X. Theillet, M. Delepierre, et al., (2012) Ordered phosphorylation events in two independent cascades of the PTEN C-tail revealed by NMR, J. Am. Chem. Soc. 134. 20533-43.
62. B. Xue, D. Blocquel, J. Habchi, A. V. Uversky, L. Kurgan, V. N. Uversky, et al., (2014) Structural disorder in viral proteins, Chem. Rev. 114. 6880-911.
63. G. Gupta, H. Qin, J. Song, (2012) Intrinsically unstructured domain 3 of hepatitis C Virus NS5A forms a "fuzzy complex" with VAPB-MSP domain which carries ALS-causing mutations, PLoS One 7. e39261.
64. D. G. Cordek, T. J. Croom-Perez, J. Hwang, M. R. Hargittai, C. V. Subba-Reddy, Q. Han, et al., (2014) Expanding the proteome of an RNA virus by phosphorylation of an intrinsically disordered viral protein, J. Biol. Chem. 289. 24397-416.
65. M. Dujardin, V. Madan, R. Montserret, P. Ahuja, I. Huvent, H. Launay, et al., (2015) A Proline-Tryptophan Turn in the Intrinsically Disordered Domain 2 of NS5A Protein Is Essential for Hepatitis C Virus RNA Replication, J. Biol. Chem. 290. 19104-20.
66. E. Secci, E. Luchinat, L. Banci, (2016) The Casein Kinase 2-Dependent Phosphorylation of NS5A Domain 3 from Hepatitis C Virus Followed by Time-Resolved NMR Spectroscopy, Chembiochem 17. 328-33.
67. D. Ross-Thriepland, M. Harris, (2015) Hepatitis C virus NS5A: enigmatic but still promiscuous 10 years on!, J Gen Virol 96. 727-38.
68. C. Choudhary, C. Kumar, F. Gnad, M. L. Nielsen, M. Rehman, T. C. Walther, et al., (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions, Science 325. 834-40.
69. F. X. Theillet, A. Binolfi, B. Bekei, A. Martorana, H. M. Rose, M. Stuiver, et al., (2016) Structural disorder of monomeric alpha-synuclein persists in mammalian cells, Nature 530. 45-50.
70. A. E. Smith, L. Z. Zhou, G. J. Pielak, (2015) Hydrogen exchange of disordered proteins in Escherichia coli, Protein Sci. 24. 706-13.
71. A. Dose, S. Liokatis, F. X. Theillet, P. Selenko, D. Schwarzer, (2011) NMR profiling of histone deacetylase and acetyl-transferase activities in real time, ACS Chem. Biol. 6. 419-24.
72. S. Liokatis, A. Dose, D. Schwarzer, P. Selenko, (2010) Simultaneous detection of protein phosphorylation and acetylation by high-resolution NMR spectroscopy, J. Am. Chem. Soc. 132. 14704-5.
73. S. Liokatis, R. Klingberg, S. Tan, D. Schwarzer, (2016) Differentially Isotope-Labeled Nucleosomes To Study Asymmetric Histone Modification Crosstalk by Time-Resolved NMR Spectroscopy, Angew. Chem. Int. Ed. Engl. 55. 8262-5.
74. C. P. Brangwynne, P. Tompa, R. V. Pappu, (2015) Polymer physics of intracellular phase transitions, Nat Phys 11. 899-904.
75. K. A. Burke, A. M. Janke, C. L. Rhine, N. L. Fawzi, (2015) Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II, Mol. Cell 60. 231-41.
76. A. E. Conicella, G. H. Zerze, J. Mittal, N. L. Fawzi, (2016) ALS Mutations Disrupt Phase Separation Mediated by alpha-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain, Structure 24. 1537-49.

77. D. M. Mitrea, C. R. Grace, M. Buljan, M. K. Yun, N. J. Pytel, J. Satumba, et al., (2014) Structural polymorphism in the N-terminal oligomerization domain of NPM1, Proc. Natl. Acad. Sci. U.S.A. 111. 4466-71.
78. D. M. Mitrea, J. A. Cika, C. S. Guy, D. Ban, P. R. Banerjee, C. B. Stanley, et al., (2016) Nucleophosmin integrates within the nucleolus via multi-modal interactions with proteins displaying R-rich linear motifs and rRNA, eLife 5.
79. T. J. Nott, E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowietz, et al., (2015) Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles, Mol. Cell 57. 936-47.
80. X. Du, J. Zhou, J. Shi, B. Xu, (2015) Supramolecular Hydrogelators and Hydrogels: From Soft Matter to Molecular Biomaterials, Chem. Rev. 115. 13165-307.
81. L. M. De Leon Rodriguez, Y. Hemar, J. Cornish, M. A. Brimble, (2016) Structure-mechanical property correlations of hydrogel forming beta-sheet peptides, Chem Soc Rev 45. 4797-824.
82. M. Kato, S. L. McKnight, (2016) Cross-beta Polymerization of Low Complexity Sequence Domains, Cold Spring Harb Perspect Biol.
83. K. Nagy-Smith, E. Moore, J. Schneider, R. Tycko, (2015) Molecular structure of monomorphic peptide fibrils within a kinetically trapped hydrogel network, Proc. Natl. Acad. Sci. U.S.A. 112. 9816-21.
84. S. R. Leonard, A. R. Cormier, X. Pang, M. I. Zimmerman, H. X. Zhou, A. K. Paravastu, (2013) Solid-state NMR evidence for beta-hairpin structure within MAX8 designer peptide nanofibers, Biophys. J. 105. 222-30.
85. C. Ader, S. Frey, W. Maas, H. B. Schmidt, D. Gorlich, M. Baldus, (2010) Amyloid-like interactions within nucleoporin FG hydrogels, Proc. Natl. Acad. Sci. U.S.A. 107. 6281-5.
86. A. A. Labokha, S. Gradmann, S. Frey, B. B. Hulsmann, H. Urlaub, M. Baldus, et al., (2013) Systematic analysis of barrier-forming FG hydrogels from Xenopus nuclear pore complexes, EMBO J. 32. 204-18.
87. H. R. Dannatt, M. Felletti, S. Jehle, Y. Wang, L. Emsley, N. E. Dixon, et al., (2016) Weak and Transient Protein Interactions Determined by Solid-State NMR, Angew. Chem. Int. Ed. Engl. 55. 6638-41.
88. T. K. Karamanos, A. P. Kalverda, G. S. Thompson, S. E. Radford, (2015) Mechanisms of amyloid formation revealed by solution NMR, Prog. Nucl. Magn. Reson. Spectrosc. 88-89. 86-104.
89. B. Eftekharzadeh, B. T. Hyman, S. Wegmann, (2016) Structural studies on the mechanism of protein aggregation in age related neurodegenerative diseases, Mech Ageing Dev 156. 1-13.
90. J. J. Sabbagh, C. A. Dickey, (2016) The Metamorphic Nature of the Tau Protein: Dynamic Flexibility Comes at a Cost, Front Neurosci 10. 3.
91. C. Wang, C. Zhao, D. Li, Z. Tian, Y. Lai, J. Diao, et al., (2016) Versatile Structures of alpha-Synuclein, Front Mol Neurosci 9. 48.
92. J. Bai, K. Cheng, M. Liu, C. Li (2016) Impact of the alpha-Synuclein Initial Ensemble Structure on Fibrillation Pathways and Kinetics, J. Phys. Chem. B 120. 3140-7.

93. Y. Suzuki, J. R. Brender, M. T. Soper, J. Krishnamoorthy, Y. Zhou, B. T. Ruotolo, et al., (2013) Resolution of oligomeric species during the aggregation of Abeta1-40 using (19)F NMR, Biochemistry 52, 1903-12.
94. B. Eftekharzadeh, A. Piai, G. Chiesa, D. Mungianu, J. Garcia, R. Pierattelli, et al., (2016) Sequence Context Influences the Structure and Aggregation Behavior of a PolyQ Tract, Biophys. J. 110. 2361-6.
95. E. Akoury, M. D. Mukrasch, J. Biernat, K. Tepper, V. Ozenne, E. Mandelkow, et al., (2016) Remodeling of the conformational ensemble of the repeat domain of tau by an aggregation enhancer, Protein Sci. 25. 1010-20.
96. L. Lim, Y. Wei, Y. Lu, J. Song, (2016) ALS-Causing Mutations Significantly Perturb the Self-Assembly and Interaction with Nucleic Acid of the Intrinsically Disordered Prion-Like Domain of TDP-43, PLoS Biol 14. e1002338.
97. G. M. Clore, J. Iwahara, (2009) Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes, Chem. Rev. 109. 4108-39.
98. M. K. Janowska, K. P. Wu, J. Baum, (2015) Unveiling transient protein-protein interactions that modulate inhibition of alpha-synuclein aggregation by beta-synuclein, a pre-synaptic protein that co-localizes with alpha-synuclein, Scientific reports 5. 15164.
99. L. M. de Lau, M. M. Breteler, (2006) Epidemiology of Parkinson's disease, Lancet Neurol 5. 525-35.
100. M. Hashimoto, E. Rockenstein, M. Mante, M. Mallory, E. Masliah, (2001) beta-Synuclein inhibits alpha-synuclein aggregation: a possible role as an anti-parkinsonian factor, Neuron 32. 213-23.
101. L. Barbieri, E. Luchinat, L. Banci, (2016) Characterization of proteins by in-cell NMR spectroscopy in cultured mammalian cells, Nat. Protoc. 11. 1101-11.
102. M. Popovic, D. Sanfelice, C. Pastore, F. Prischi, P. A. Temussi, A. Pastore, (2015) Selective observation of the disordered import signal of a globular protein by in-cell NMR: the example of frataxins, Protein Sci. 24. 996-1003.
103. A. Binolfi, A. Limatola, S. Verzini, J. Kosten, F. X. Theillet, H. M. Rose, et al., (2016) Intracellular repair of oxidation-damaged alpha-synuclein fails to target C-terminal modification sites, Nature communications 7. 10251.
104. U. Akbey, H. Oschkinat, (2016) Structural biology applications of solid state MAS DNP NMR, J. Magn. Reson. 269. 213-24.
105. K. K. Frederick, V. K. Michaelis, B. Corzilius, T. C. Ong, A. C. Jacavone, R. G. Griffin, et al., (2015) Sensitivity-enhanced NMR reveals alterations in protein structure by cellular milieus, Cell 163. 620-8.
106. T. Viennet, A. Viegas, A. Kuepper, S. Arens, V. Gelev, O. Petrov, et al., (2016) Selective Protein Hyperpolarization in Cell Lysates Using Targeted Dynamic Nuclear Polarization, Angew. Chem. Int. Ed. Engl. 55. 10746-50.

Chapter 3

A NMR and MS Based Approach to Phosphorylation Site Identification for the RNA Pol II CTD

[MS/MS data presented in this chapter was collected by T. Laremore, Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, Pennsylvania. Expression and purification of DmP-TEFb was performed by B. Portz]

Intrinsically disordered proteins (IDPs) have become a focus of recent biophysical studies due to their prominent roles in transcription and human disease. Many IDPs are regulated through multi-site phosphorylation. However, the accurate identification of IDP phosphorylation sites can be complicated by their inclusion in proline rich, low complexity amino acid sequences, which challenge conventional bioanalytical techniques. The carboxyl-terminal domain of the RNA polymerase II largest subunit (CTD) represents an extreme example of such a system; it is composed of tandem repeats of the consensus sequence (YSPTSPS), wherein multiple potential phosphorylation sites are flanked by proline. We identified phosphorylation sites in the *Drosophila melanogaster* CTD that are targeted by the Positive Transcription Elongation Factor b (DmP-TEFb) using mass spectrometry and NMR spectroscopy. We show that phosphorylation occurs primarily at Ser5 residues. The approach outlined here may provide a general workflow for the identification of phosphorylation sites within proline rich low complexity domains.

Introduction

In eukaryotic cells, transcription of all protein coding genes is performed by RNA polymerase II. The carboxyl-terminal domain of Pol II's largest subunit (CTD) is essential for the function of the polymerase *in vivo*[1]. This intrinsically disordered domain (IDD), consisting of multiple tandem repeats of the consensus sequence ($Y^1S^2P^3T^4S^5P^6S^7$), serves as a platform for the docking and assembly of mRNA processing factors. Tight control of factor association and removal is regulated by CTD specific kinases, phosphatases, and peptidyl-prolyl isomerases, which generate dynamic patterns of post-translational modifications (PTMs) collectively referred to as the "CTD code"[2, 3]. A wealth of biochemical and genomic studies have highlighted the importance of specific CTD modifications for discrete steps of the transcription cycle, with pSer5 linked to promoter proximal pausing and mRNA capping, and pSer2 linked to elongation, co-transcriptional splicing, and 3′ processing[4].

While an increasing number of studies support the CTD code hypothesis, a knowledge gap currently exists between the broader functional outcomes of CTD phosphorylation and the underlying molecular mechanisms of CTD regulation. Indeed, the CTD's highly repetitive amino acid sequence makes disambiguation of individual repeats extremely difficult, complicating backbone resonance assignment by NMR[5, 6]. And the identification of phosphorylation sites within yeast and mammalian CTDs by tandem mass spectrometry (MS2) required extensive mutagenesis[7, 8]. Methodologies enabling the accurate assignment of CTD phosphorylation states, would therefore provide new opportunities to address structure function relationships for the CTD. Here we leveraged novel ^{13}C Direct-Detect NMR spectroscopy, in combination with mass spectrometry to establish a protocol for the accurate identification of phosphorylation sites within proline rich, low complexity IDPs. To illustrate the workflow, we mapped sites within the *Drosophila melanogaster* CTD that are targeted by DmP-TEFb *in vitro*.

Results & Discussion

Our model system for this study, referred to here as CTD2′, spanned a large portion of the *D. melanogaster* CTD (Rpb1 residues 1657-1739). (Figure 3-1A) This construct was chosen for several reasons: CTD2′ has the highest sequence conservation among model eukaryotes (Chapter 5); the transcription termination factor Pcf11 was shown to preferentially associate within this region *in vitro* [9]; and as shown in chapter 5, CTD2′ contains a region that is essential for the function of the polymerase *in vivo*. Based on the literature, CTD2′ contains 24 sites that can potentially be phosphorylated by transcriptional cyclin dependent kinases (CDKs), including serine residues located at the 2, 5 and 7 positions of multiple repeats [2, 10-12]. (Figure 3-1B) We, therefore wanted to determine which sites could be targeted *in vitro* by DmP-TEFb.

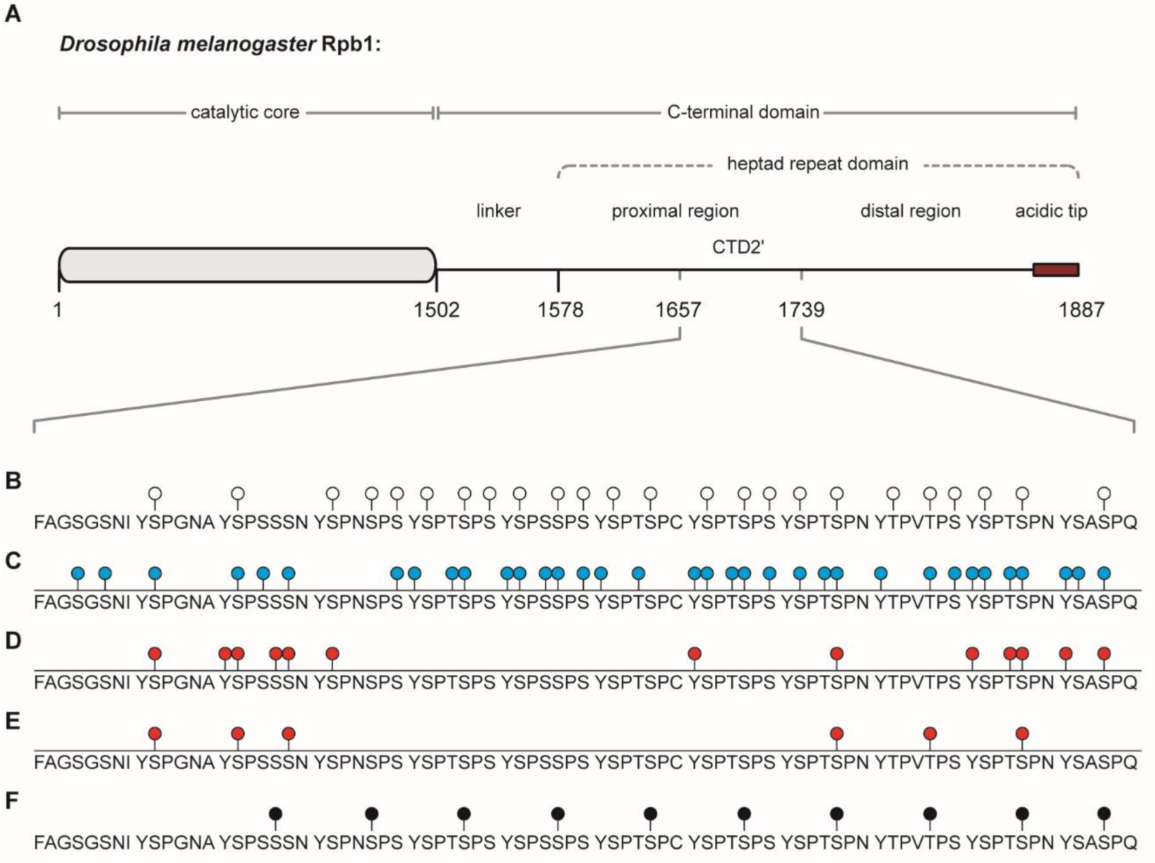

A
Drosophila melanogaster Rpb1:
catalytic core
C-terminal domain
heptad repeat domain
linker
proximal region
distal region
acidic tip
CTD2'
1
1502
1578
1657
1739
1887
B
FAGSGSNI YSPGNA YSPSSSN YSPNSPS YSPTSPS YSPSSPS YSPTSPC YSPTSPS YSPTSPN YTPVTPS YSPTSPN YSASPQ
C
FAGSGSNI YSPGNA YSPSSSN YSPNSPS YSPTSPS YSPSSPS YSPTSPC YSPTSPS YSPTSPN YTPVTPS YSPTSPN YSASPQ
D
FAGSGSNI YSPGNA YSPSSSN YSPNSPS YSPTSPS YSPSSPS YSPTSPC YSPTSPS YSPTSPN YTPVTPS YSPTSPN YSASPQ
E
FAGSGSNI YSPGNA YSPSSSN YSPNSPS YSPTSPS YSPSSPS YSPTSPC YSPTSPS YSPTSPN YTPVTPS YSPTSPN YSASPQ
F
FAGSGSNI YSPGNA YSPSSSN YSPNSPS YSPTSPS YSPSSPS YSPTSPC YSPTSPS YSPTSPN YTPVTPS YSPTSPN YSASPQ

Figure 3-1. Phosphorylation site assignment of CTD2′ by LC-ESI-MS2 and NMR spectroscopy. (A) Schematic representation of the *Drosophila melanogaster* Rpb1 showing the location of CTD2′. (B) The amino acid sequence of CTD2′ is shown, where the open markers represent putative phosphorylation sites based on the documented activities of transcriptional CDKs. Phosphorylation sites determined by the SEAQUEST (C), PhosphoRS (D), and AScore (E) algorithms are shown, where the blue markers represent SEAQUEST scores above 0.4, red markers represent sites mapped at the 99% confidence interval and the peptide coverage is denoted by the black bar. (F) Sites assigned unambiguously by NMR spectroscopy are denoted by the black markers.

Phosphorylation Site Identification of CTD2′ by LC-ESI-MS/MS

Tandem mass spectrometry is one of the most widely used techniques for the identification of protein phosphorylation sites owing to its high sensitivity, straightforward approach, and potential for high-throughput analysis. The accurate localization of phosphorylation sites is contingent upon: (i) the generation of a sufficient number of unique peptides by proteolytic digest to map the polypeptide sequence, and (ii) the efficient ionization of the resulting fragments to produce "site-determining" ions, which display characteristic mass shifts for the addition of the phosphate group. Ionization can be accomplished by several methods [13], including collision induced dissociation (CID), where translational energy from peptide collisions in the gas phase induce fragmentation into b and y-type ions, and electron transfer dissociation (ETD), which utilizes hydrogen radicals to cleave the peptide backbone producing c and z-type ions. (Figure 3-2A). CID is a high energy fragmentation method that that efficiently cleaves the peptide backbone. However, for phosphoproteins, phosphate bond cleavage presents a competing pathway and neutral loss ions formed by the loss of H_3PO_4 may dominate, thus reducing the information content in the acquired spectra. (Figure 3-2B). In these cases, ETD may be advantageous as it often preserves the covalent linkages to phosphate groups. (Figure 3-2C). For phosphorylated CTD2′, proteolysis was performed by chymotrypsin, which specifically cleaves c-terminal to Tyr residues, producing peptides that could be uniquely mapped to the amino acid sequence. And dual activation by CID and ETD provided overlapping sets of peptide spectra, thus maximizing the amount and type of ions available for phospho-site determination.

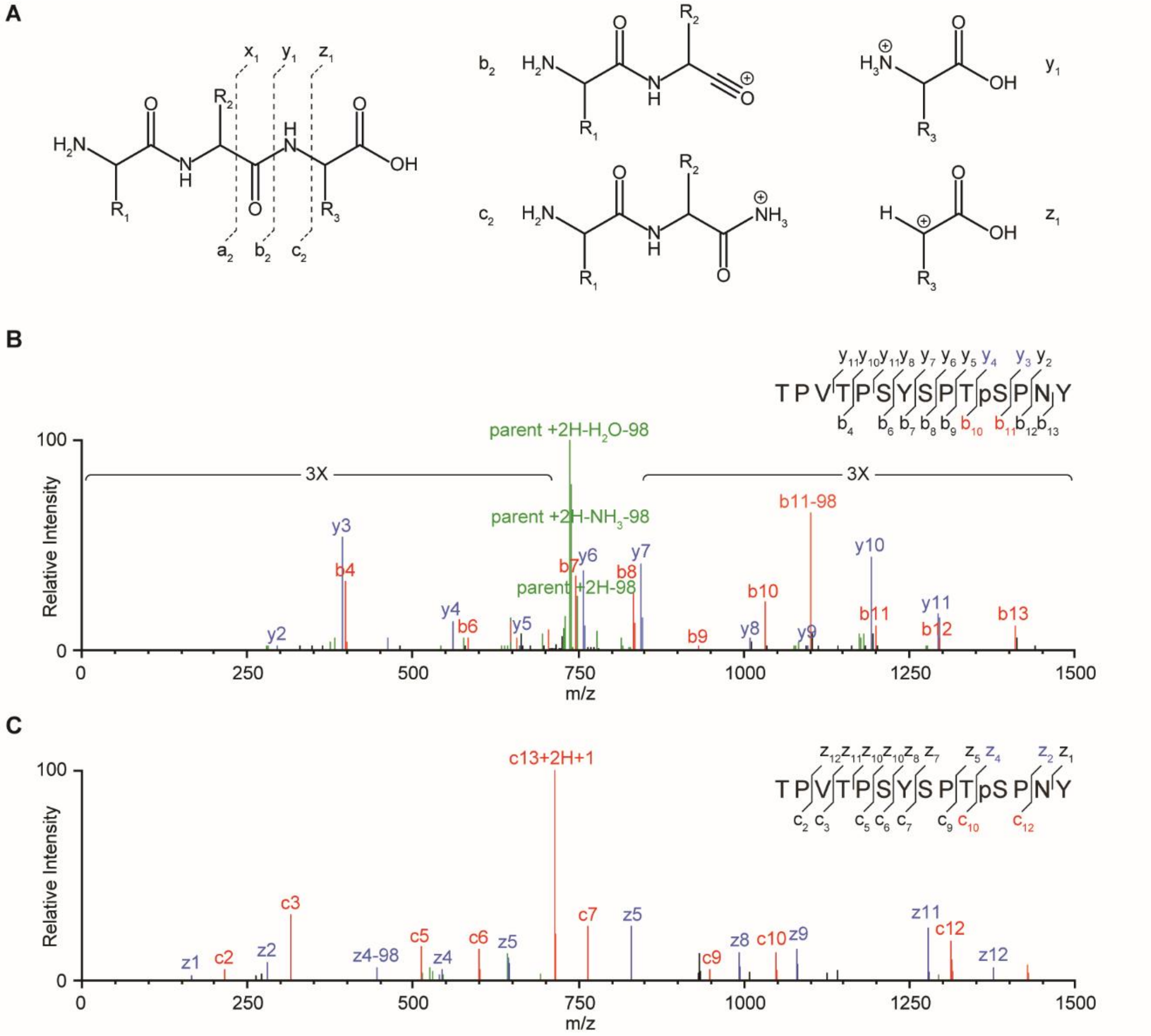


Figure 3-2. Localization of phosphorylation sites using MS2 peptide spectra. (A) Cleavage of the peptide backbone in the gas phase produces characteristic ion types, including b & y ions produced by CID and the c & z ions produced by ETD. (B) MS2 spectrum of a peptide from DmP-TEFb phosphorylated CTD2′ produced by CID. (C) MS2 spectrum of a peptide from DmP-TEFb phosphorylated CTD2′ produced by ETD. In both examples, pS1731 is localized by the presence of “site-determining” ions, which are highlighted in the diagrams above each spectrum.

Peptide Scoring Algorithms

In a typical MS2 experiment, thousands of individual peptide fragment spectra can be collected. Therefore, analysis of MS2 data sets is generally performed with the aid of peptide identification algorithms. Processing of the unphosphorylated and DmP-TEFb phosphorylated CTD2′ MS2 data sets was performed using several algorithms, including the SEAQUEST [14], PhosphoRS [15], and Ascore [16] algorithms. (Figure 3-1,C-E). An initial SEAQUEST peptide search was performed producing the pattern shown in Figure 3-1C. The SEAQUEST algorithm generates theoretical MS2 spectra using the single letter amino acid codes contained within protein sequence databases, e.g. NCBI. The experimental spectrum is then compared to theoretical spectra of the same precursor ion mass and scored based on a cross-correlation function. This SEQUEST score is calculated for all possible phosphorylation patterns within the a given peptide and the difference between the highest and second highest score (the delta correlation or deltaCn) provides a measure of the ambiguity of the assignment, with deltaCn scores >0.1 considered “good hits”. Since the scores reflect the best match to theoretical spectra, they may not contain the site-determining ions necessary for localization of the phosphorylated residues, but will nonetheless be reported. SEAQUEST can however, rapidly identify target peptide spectra in large data sets, and thus provides a good first step in the processing pipeline.

Following the SEAQUEST search, two probability based algorithms were used independently, PhosphoRS and Ascore. (Figure 3-1D-E). The Ascore algorithm starts by dividing the mass range of an experimental spectrum is into windows of 100 m/z. For each window, the top n most intense peaks are compared to the theoretical site-determining b and y-ions corresponding to each possible phospho-isoform. The analysis is repeated for n=1 to 10 ions per window and a peptide score is calculated based on a cumulative binomial probability, using the number of matches and the number of attempts. The ambiguity score, or Ascore, is the difference between the top two best possible solutions, where a score ≥19 is analogous to a confidence interval of >99%. Thus, Ascore measures the likelihood that the difference in site-determining ions for two phospho-isoforms were matched by random chance. PhosphoRS uses a similar approach, with the exceptions that the optimal number of peaks in each 100 m/z window are determined independently for each window, and all ion types can be analyzed, e.g. CID and ETD. As these two algorithms only consider site-determining ions, they provide a more accurate means to identify phosphorylation sites. For CTD2′, this reduced the number of hits, resulting in the identification of predominantly serine phosphorylation, which was expected based on the documented specificity of P-TEFb [10].

These results demonstrate that LC-ESI-MS2 can be used to localize phosphorylation sites within a native CTD sequence provided enough variation is present to uniquely identify target peptides and enough site-determining ions are present to localize phosphorylated residues. However, the lack of hits for the internal heptads in the phospho-site maps produced from the PhosphoRS and Ascore algorithms were of concern. And, the lack of apparent specificity in the resulting phosphorylation patterns was striking, particularly for the PhosphoRS map, which contained several sites with little similarity to known transcriptional CDK peptide recognition motifs, e.g., pTyr1. We therefore, sought orthogonal techniques to cross-validate or MS2 results.

Phosphorylation Site Identification of CTD2′ by NMR Spectroscopy

NMR spectroscopy is arguably the most powerful technique for the characterization of IDPs as it provides near atomic resolution and is amenable to flexible systems [17,18]. The interpretation of NMR spectra requires the unambiguous assignment of the observed resonances, a non-trivial task for IDPs like CTD2′, which produce characteristically condensed spectra as a result of conformational heterogeneity and low sequence complexity. Therefore, our assignment strategy utilized ^{13}C Direct-Detect NMR, which greatly improves resonance dispersion for IDPs and permits the detection of proline residues [17]. The work-flow proceeded through assignment of the ^{13}C,^{15}N-CON. To obtain unambiguous starting points, 2D carbon-detected amino acid selective spectra (CAS) were collected, including the CAS-HACACON, which provides resonances for a specific amino acid type (residue i), and the CAS-HACANCO, which provides resonances for residue i and the following residue (i+1) [19]. (Figure 3-3A). In Figure 3A, S1696 (asterisked) is easily identified using the pair of CAS-Ser spectra as it is located within a polyserine motif (YSPSS*PS). From these starting points, the 3D (HACA)N(CA)CON experiment was collected to obtain correlations from the ^{15}N$_H$ chemical shift of each residue to the ^{15}N$_H$ chemical shift of the preceding residue, generating the i to i - 1 connectives needed for the walk along the backbone type assignment [17]. (Figure 3-3B,C). Corroboration of these assignments, was obtained using the 3D (HACA)N(CA)NCO spectrum, which provides correlations from the ^{15}N$_H$ of residue i to those in residues i-1 and i+1. (Figure 3-3B,D) Assignments were then transferred onto the ^{1}H,^{15}N-HSQC via standard HNCO and HNCACO spectra to obtain the corresponding ^{1}H$_N$ chemical shifts.

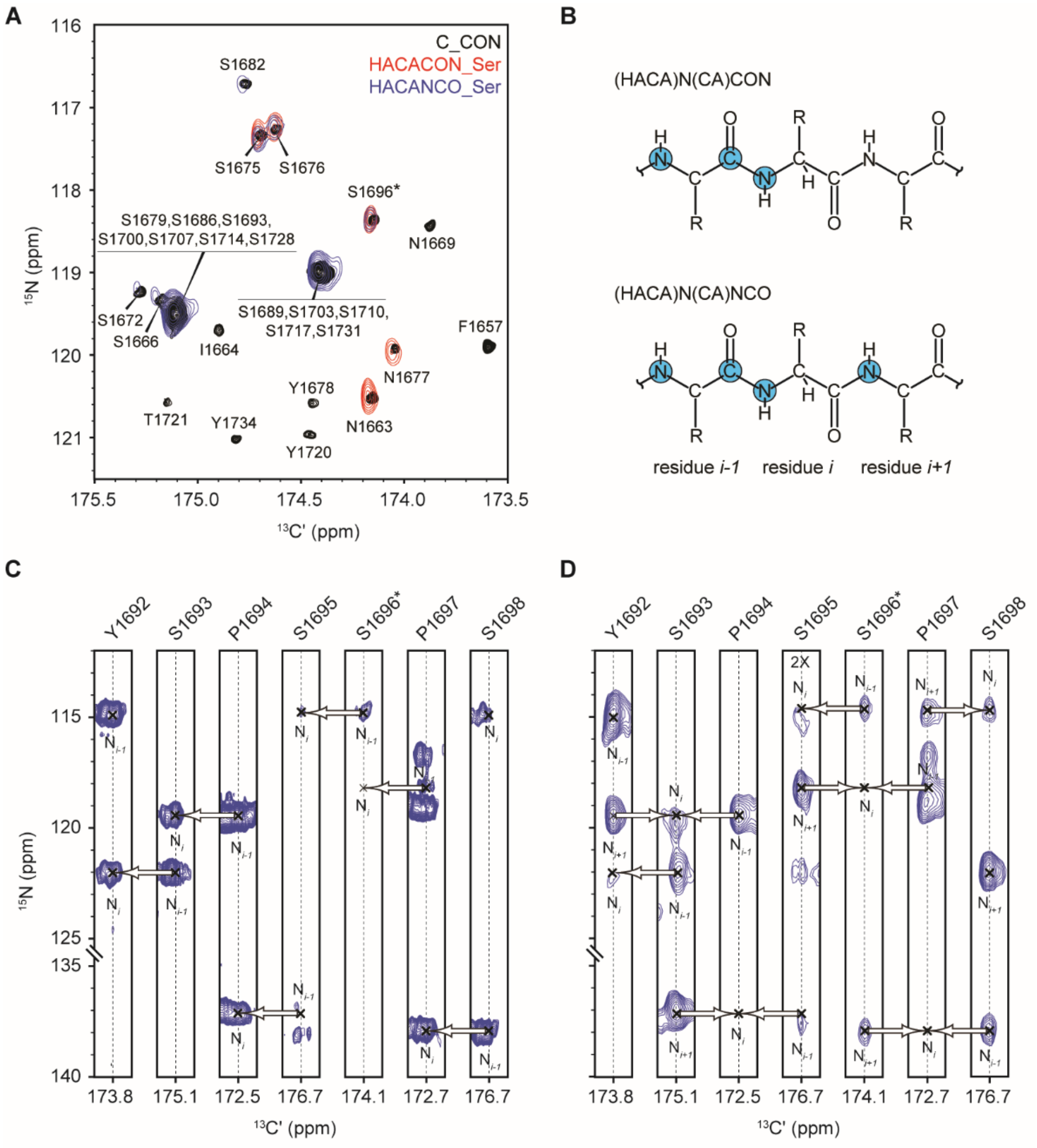


Figure 3-3. Backbone resonance assignment of CTD2′ by ^{13}C Direct-Detect NMR. (A) Zoomed in region of 2D ^{13}C,^{15}N-CON (black) with overlaid HACACON_Ser (red), and HACANCO_Ser (blue). Peaks common to all three spectra correspond to residues that are located in polyserine stretches. (B) Schematic displaying resonances observed by the 3D (HACA)N(CA)CON (top) and 3D (HACA)N(CA)NCO (bottom). Strips from the (C) 3D (HACA)N(CA)CON and (D) 3D (HACA)N(CA)NCO demonstrate the walk along the backbone assignment of unphosphorylated CTD2′.

Due to its high sensitivity to changes in local chemical environments, NMR has emerged as one of the premier methods for identifying phosphorylation sites [20]. Through isotopic enrichment, samples can be prepared such that they provide a variety of intramolecular probes for phosphorylation. (Figure 3-4A). The ^{31}P nucleus is a sensitive probe owing to its spin ½, near 100% natural abundance and considerable gyromagnetic ratio. Thus, ^{31}P NMR can provide a fast and straightforward means to detect protein-incorporated phosphate groups, which resonate downfield in a 1D ^{31}P spectrum. (Figure 3-4B). However, the presence of multiply phosphorylated species and/or conformational exchange makes disambiguation of individual resonances difficult. By contrast, 2D heteronuclear correlation spectroscopy can more easily resolve resonances corresponding to phosphorylated residues. The covalent attachment of the phosphate group effects the local chemical environments of nearby spin systems inducing characteristic chemical shift perturbations (CSPs) [21]. The presence of the highly charged phosphate group strongly effects the amide-proton resulting in downfield shifts in $^{1}H_{N}$ that can be observed in ^{1}H,^{15}N-HSQC spectra [22,23]. (Figure 3-4C). Similar perturbations can be observed in ^{13}C,^{15}N-CON spectra, although it should be noted that the directionality and amplitude of these shifts may vary as $^{13}C'$ resonances may be more sensitive to local reconfiguration of the protein backbone. (Figure 3-4D). Additionally, sidechain resonances, including the $^{13}C_{\beta}$ and $^{1}H_{\beta}$ chemical shifts can provide sensitive reporters of phosphorylation [23]. (Figure 3-4E,F).

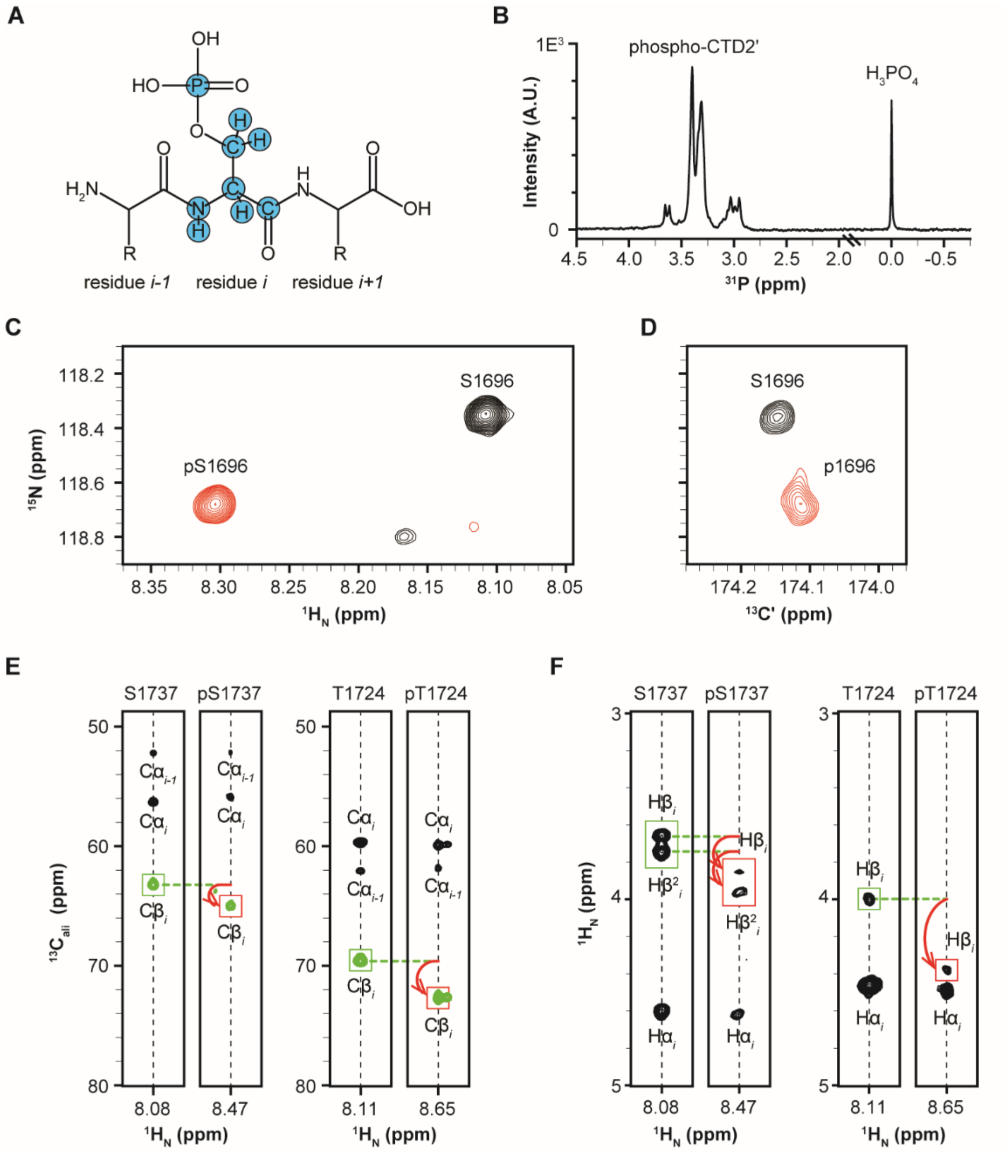


Figure 3-4. Detection and assignment of phosphorylated CTD2′ by NMR. (A) 1D ^{31}P spectrum of phospho-CTD2′ shows characteristic phosphoprotein resonances located downfield of an internal phosphoric acid standard. (B) 2D ^{1}H,^{15}N-HSQC and (C) ^{13}C,^{15}N-CON spectra of unphosphorylated (black) and phosphorylated (red) CTD2′ demonstrates chemical shift perturbation of backbone amide and carbonyl resonances upon phosphorylation of S1703. (D) 3D HNCACB and (E) DIPSI-HSQC spectra show phosphorylation induced chemical shift perturbation of side chain carbon and proton resonances, respectively.

When quantified and plot as a function of residue number, CSP plots can aid in the visualization of phosphorylation sites. For CTD2′, the greatest extent of CSP was observed for Ser/Thr residues located at the 5 position of the heptad repeats. (Figure 3-5A) In addition, several residues flanking these sites also experienced large CSPs (Figure 3-5B). However, these perturbations could be attributed to conformational changes rather than phosphorylation. The effects of phosphorylation on the structure of CTD2′ is discussed in greater detail in chapter 5. In total, 10 *bona-fide* DmP-TEFb phosphorylation sites were identified in CTD2′, resulting the pattern shown in Figure 3-1F. These were distributed evenly throughout the polypeptide sequence and overwhelmingly corresponded to pSer5 marks.

Cross-Validation

Comparison of the phosphorylation sites identified by MS2 and NMR demonstrated good agreement between the two methods for several sites within the DmP-TEFb phosphorylated CTD2′. Further, NMR revealed five sites within the interior of the polypeptide sequence that were unable to be identified by MS2. By contrast, the least agreement was obtained for sites within the N-terminal portion, including pSer2 marks in the second and third repeats, (pS1666 and pS1672, respectively), and a pSer mark distributed within the polyserine motif of the YSPSSSN heptad (p1675 or pS1676). For the latter site, careful inspection of ^{1}H-Detect NMR spectra permitted the resolution of pS1675, which displayed a characteristic downfield shift in the $^{13}C_{\beta}$ resonance as observed in the 3D HNCACB spectrum. (Figure 3-6A). Corroboration of this assignment was obtained by site-directed mutagenesis. (Figure 3-6,7). Upon phosphorylation, the S1676A mutant displayed downfield shifted resonances similar to that of the phosphorylated wild type CTD2′ (Figure 3-6B,C), while the S1675A mutant lacked these perturbed resonances. (Figure 3-6D).

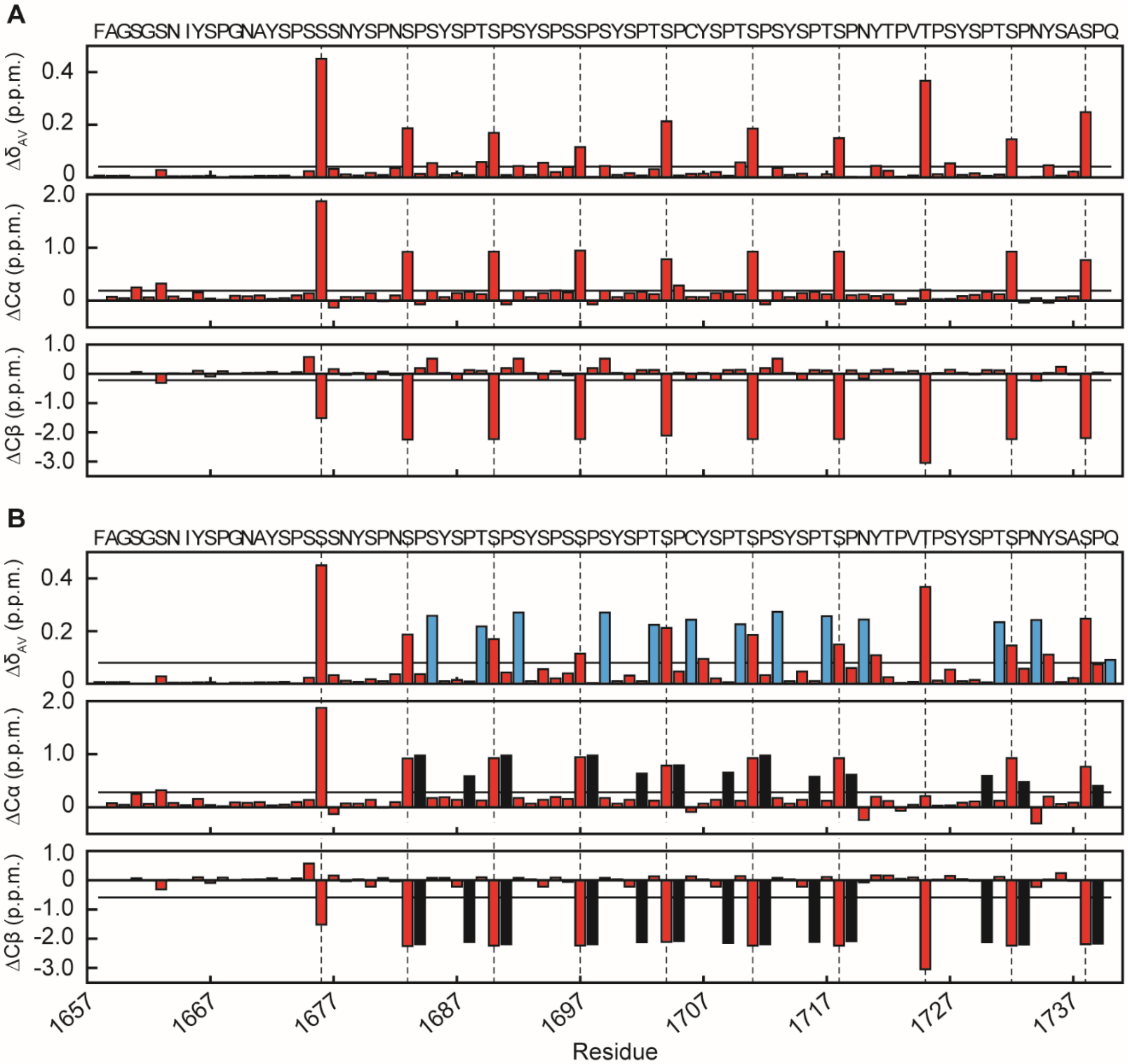


Figure 3-5. NMR chemical shift perturbations observed for CTD2′ following phosphorylation by DmP-TEFb. (A) Average changes to backbone chemical shifts are shown for the all *trans*-proline state of DmP-TEFb phosphorylated CTD2′ (top) with the change in $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ chemical shifts shown below. (B) Phosphorylation of Ser5 induces isomerization of adjacent prolines in CTD2′ as

indicated by perturbation of backbone chemical shifts for residues in the 4 and 7 position of corresponding heptad repeats (blue bars) and the $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ resonances of proline residues (black bars).

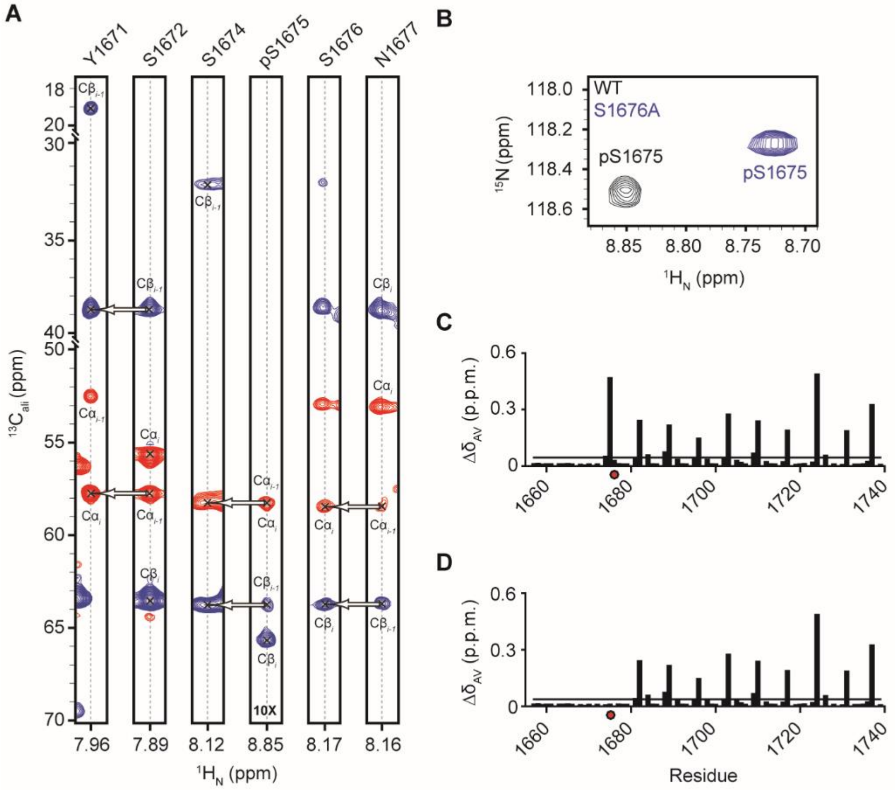


Figure 3-6. Assignment of pS1675 by NMR spectroscopy and site-directed mutagenesis. (A) Strips from the 3D HNCACB spectrum displays a downfield shifted $^{13}C_{\beta}$ resonance for pS1675 and relatively unperturbed side chain resonances for adjacent serine residues. (B) Comparison of ^{1}H,^{15}N-HSQC spectra for phosphorylated CTD2′ (black) and the S1676A mutant (blue) displays similar downfield shifted ^{1}H resonances (C) In spectra of the phosphorylated wild type and S1676A mutant, similar patterns of chemical shift perturbations are observed throughout the polypeptide chain. (D) Assignment of pS1675 is confirmed by the loss of resonances in S1675A mutant spectra.

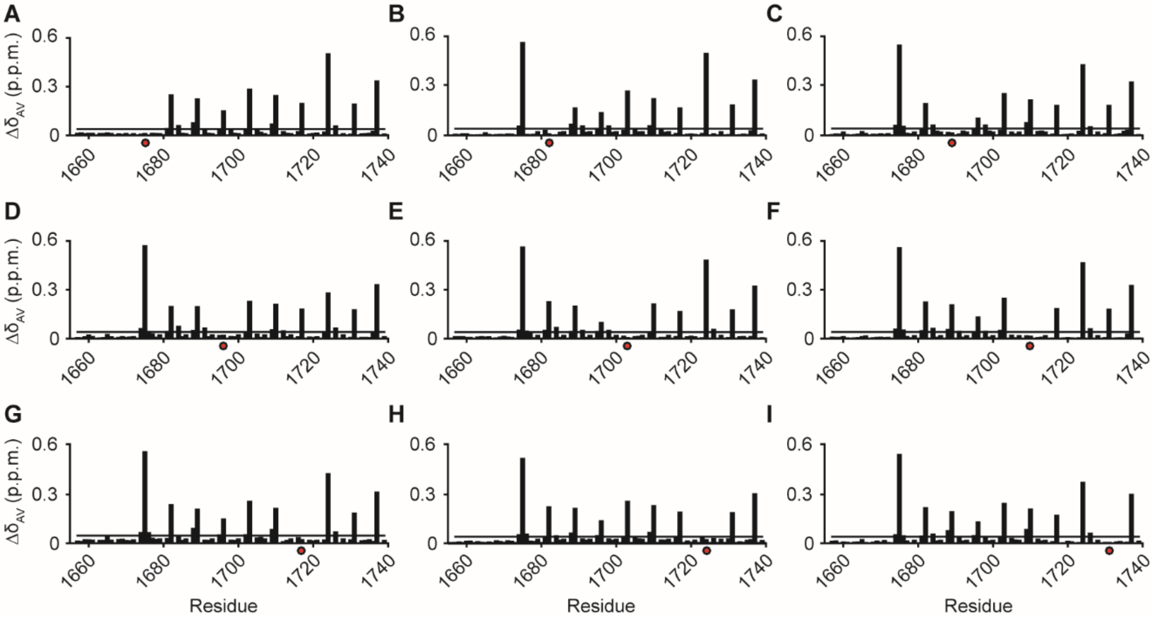


Figure 3-7. NMR assigned DmP-TEFb phosphorylation sites in CTD2′ verified by site-directed mutagenesis. The CSP patterns observed for the CTD2′ mutants following phosphorylation demonstrate the loss of specific perturbed resonances, consistent with the removal of the targeted phosphorylation sites. In each, the site of mutation is indicated by the red dot. (A) S1675A, (B) S1682A, (C) S1689A, (D) S1696A, (E) S1703A, (F) S1710A, (G) S1717A, (H) S1724, (I) S1731A

Cross-validation of the remaining phosphorylation sites assigned by NMR was obtained in similar fashion, resulting in the CSP plots shown in figure 3-7. In order to corroborate the MS2 assignment of pS1666 and pS1672, S1666A and S1672A point mutants were also tested, however, no change in the patterning of the resulting CSPs could be detected (data not shown), suggesting that pS1666 and pS1672 populate states below the limit of detection by ^{1}H-detect NMR (<5%). Interestingly, these two sites are located in heptad repeats that lack Ser5s. Using matrix-assisted laser desorption ionization mass spectrometry (MALDI-TOF-MS), we were able to detect modest amounts of Ser2 phosphorylation (~20%) within the context of a synthetic consensus CTD peptide with the Ser5 mutated to Ala (SPSYSPTAPSYSPT). (Chapter 4). Therefore, to confirm the identification of pS1666 and pS1672, we subjected DmP-TEFb phosphorylated CTD2′ to top down sequencing by MALDI using in-source decay (MALDI-ISD). In analogy to ETD activation, ISD relies on hydrogen radicals to induce peptide fragmentation (c and y-type ions). This results in a peptide ladder of N and C-terminal fragments that can be mapped to the polypeptide sequence. (Figure 3-8A). Based on this mode of fragmentation, ISD preserves labile modifications like phosphoserine [24]. And in analogy to MS2, phospho-site localization by MALDI-ISD relies on identification of "site-determining" ions. For DmP-TEFb phosphorylated CTD2′, top down sequencing by MALDI-ISD produced a c-ion series enabling the localization of pS1666 and pS1672, thus confirming their initial identification by LC-ESI-MS2. (Figure 3-8B).

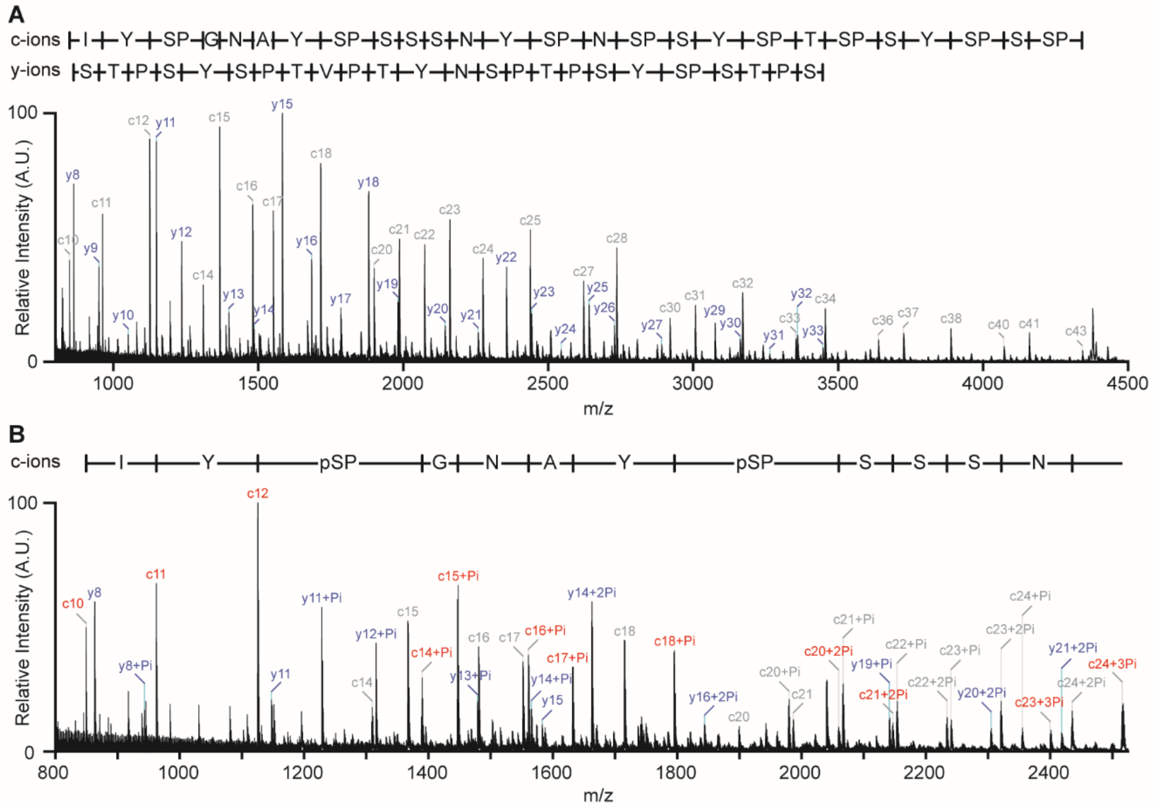


Figure 3-8. Top-down sequencing of unphosphorylated and DmP-TEFb phosphorylated CTD2′ by MALDI-ISD. (A) ISD spectrum of the unphosphorylated CTD2′ demonstrates nearly complete coverage of the polypeptide sequence by N-terminal (c-ions) and C-terminal (y-ions) fragmentation. (B) ISD spectrum of DmP-TEFb phosphorylated CTD2′ displaying the localization of pS1666 and pS1672.

Conclusion

The combined use of NMR and MS for the identification of phosphorylation sites has been well documented [25,26]. As demonstrated here for the DmP-TEFb phosphorylated CTD2′, the combination of ^{13}C-Direct Detect NMR and LC-ESI-MS2 provides an attractive option for mapping phospho-sites in proline rich, low complexity and or repeat containing IDPs. In this case, initial MS2 screens identified several phosphorylation sites with a high degree of confidence, albeit with relatively less coverage within the interior repeats, presumably due to inefficiencies in proteolytic digest. Heteronuclear NMR spectroscopy, aided by direct detection of ^{13}C′ nucleus enabled the backbone resonance assignment of CTD2′, thus permitting the identification of an additional five DmP-TEFb phosphorylation sites. As we showed, backbone and sidechain chemical shifts provide sensitive probes for phosphorylation. However, as a note of caution, alterations in secondary structure as a result of phosphorylation were observed for CTD2′. And similar effects have been observed for other IDPs [27]. Therefore, cross-validation through the analysis of multiple distinct chemical shifts, or through implementation of orthogonal bioanalytical techniques, provides a route to phospho-site validation. As a final note, substoichiometric phosphorylation may pose significant challenges, particularly if the sub-state populations are close to the limits of detection. MALDI-ISD therefore provides an expedient path to cross-validation. However, it should be noted that the increasing complexity of observed phospho-isoforms as well as the loss in efficiency of ISD above 3.5 kDa places practical limitations on phosphorylation site identification for extensively modified proteins [28].

Materials & Methods

Recombinant expression and purification of CTD2':

A synthetic E. coli codon optimized plasmid for the *Drosophila melanogaster* Rpii215 gene (encoding amino acids 1503-1887) was purchased from GeneArt (Thermo Fisher Scientific). C-terminal domain residues (1657-1739) was subcloned by PCR amplification, digestion with XhoI (NEB) and XmaI (NEB), and subsequent ligation into the pET49b+ expression vector (Novagen) using T4 DNA ligase (NEB). The resulting construct contained N-terminal GST & polyhistidine tags. Alanine mutants were produced by site-directed mutagenesis using the QuickChange Lightning Site Directed Mutagenesis kit (Agilent). All primers were designed using the online Quickchange Primer Design tool.

Protein overexpression was performed in 500 ml batch cultures of E. coli BL21 DE3 cells grown to an optical density of 0.8 at 37 ^{o}C and induced using 0.5 mM IPTG. Expression proceeded for 3 hours at 37 ^{o}C. Cell lysis was performed by sonication on ice in lysis buffer (50 mM Tris/HCl pH 7.5, 500 mM NaCl, 20 mM Imidaozole, 2.5 mM β-mercaptoethanol, 10X EDTA-free protease inhibitor cocktail (Calbiochem), and 10 units of RNAse free DNAse (NEB)). The crude lysate was then centrifuged at 4 ^{o}C for 40 minutes at 11,500 x g. The cleared supernatant was passed over HisPur Ni^{2+}-NTA resin (Thermo Fisher Scientific), contaminants were removed by using 5 column volumes of wash buffer (50 mM Tris/HCl pH 7.5, 500 mM NaCl, 20 mM imidaozole, 0.1% Triton-1000, 2.5 mM β-mercaptoethanol) and purified protein was eluted using elution buffer (50 mM Tris/HCl pH 7.5, 500 mM NaCl, 200 mM imidaozole, 2.5 mM β-mercaptoethanol). Recombinant His-tagged HRV 3C protease was added to the elution and the mixture was subjected to overnight dialysis against 50 mM Tris/HCl pH 7.5, 300 mM NaCl, 2.5 mM β-mercaptoethanol at 4 ^{o}C in 1 kDa MWCO dialysis membranes (Spectrapor), thus proteolytically cleaving the GST and 6XHis tags (resulting in an N-terminal non-native GPG) and diluting the concentration of Imidazole. The protein was passed over a Ni^{2+}-NTA column to remove the protease and non-specifically bound contaminants. And a final purification was performed by size exclusion chromatography (SEC) in 80 mM Imidazole pH 6.5, 50 mM KCl, 2.5 mM β-mercaptoethanol using P-10 resin (BioRad).

Recombinant expression and purification of DmP-TEFb:

Sf9 cells were grown in suspension at 27 ^{o}C to 1.5 million cells/ml and infected with 1/10 culture volume *D. mel* P-TEFb virus (generous gift from J.T. Lis). Infection was carried out at 27 ^{o}C at a shaker speed of 75 rpm for 72 hours. Following lysis in 50 mM HEPES pH 7.5, 500 mM NaCl, 10% glycerol, 1% Nonidet P-40, 2.5 mM imidizole and 2.5mM β-mercaptoethanol and protease inhibitors, by dounce homogenization, lysates were ultra-centrifuged at 4°C for 30 minutes at 100,000 x g. Cleared supernatant was passed over TALON resin (Clontech) and bound protein was washed using 5 column volumes of 50 mM HEPES pH 7.5 500 mM NaCl 10% glycerol 1% Nonidet P-40 10 mM imidizole 2.5 mM β-mercaptoethanol. DmP-TEFb was eluted with 50 mM HEPES pH 7.5 500 mM NaCl 10 % glycerol 1% Nonidet P-40, 200 mM imidazole, 2.5 mM β-mercaptoethanol, and flash frozen.

In vitro kinase reactions:

Kinase reactions were performed in 50 mM Tris/HCl pH 7.5, 50 mM NaCl, 12 mM MgCl2, 2 mM DTT, 12 mM ATP with 50-100 µg/ml DmP-TEFb and 100 µM CTD2′. Reactions were carried out at 24^{o}C and allowed to proceed to completion (~16 hours), after which 20 mM EDTA was added to ensure reaction termination. Following kinase treatment, an additional SEC purification with P-10 resin (BioRad) in 80 mM Imidazole pH 6.5, 50 mM KCl, 2.5 mM β-mercaptoethanol was performed.

Chymotrypsin digest:

A chymotrypsin stock solution at 1 µg/µL was prepared with sequencing-grade chymotrypsin (Thermofisher) in 1 mM hydrochloric acid, diluted 50-fold with chilled 50 mM aqueous solution of triethylammonium bicarbonate (TEAB), and added to the CTD solution in 1:1 vol:vol ratio. Proteolysis proceeded overnight at 37 ^{o}C. Samples were acidified with a 1% aqueous solution of formic acid (FA), speed-vacuumed to dryness, and re-dissolved in 15 µL of 4% acetonitrile (ACN) containing 0.1% FA for the nano-LC MS2 analysis.

Nano-LC MS2:

3 µL of digested peptide solution was loaded onto an Acclaim PepMap100 trapping column (100 µm × 2 cm, C18, 5 µm, 100 Å, Thermo) at a flow rate of 20 µL/min using 4% aqueous acetonitrile (ACN), 0.1% formic acid (FA) as a mobile phase. The peptides were separated on an Acclaim PepMap RSLC column (75 µm × 15 cm, C18, 2 µm, 100 Å, Thermo) with a 90-min 4% - 60% linear gradient of aqueous acetonitrile containing 0.1%

FA. The gradient was delivered by a Dionex Ultimate 3000 nano-LC system (Thermo) at 300 nL/min.

An LTQ Orbitrap Velos ETD mass spectrometer (Thermo) was set to acquire data using the following data-dependent parameters. A full FT MS scan at R 60,000 over 350 – 2000 *m/z* range was followed by 5 FT MS2 scans with CID activation and 5 FT MS2 scans with ETD activation on most intense precursors at R 7,500. Only the precursors with charge states +2 and higher were selected for MS2 based on the FT master scan preview; the charge-state-dependent ETD time and monoisotopic precursor selection were enabled; the isolation window was 5 *m/z,* and the minimum precursor signal was set at 10000 counts for both CID and ETD. The ETD activation time was 100 ms. Polysiloxane ion, *m/z* 445.12003 was used as a lock mass.

Data analysis:

Mass spectra were processed using Proteome Discoverer 1.3 (P.D. 1.3, Thermo). The expressed CTD2′ sequence was appended to a database containing 45443 sequences (*Drosophila Melanogaster, Escherichia Coli*, common contaminants). The workflow was split into pipelines for CID and for ETD; and the following search parameters were used for both pipelines: enzyme chymotrypsin with 3 missed cleavages, precursor mass tolerance 30 ppm, fragment mass tolerance 0.8 Da, Met oxidation and Ser, Thr, Tyr phosphorylation as dynamic modifications, Cys carbamidomethylation as static modification. The CID ion series weights for b and y were set to 1; and the ETD ion series weights were 1 for c and z ions and 0.25 for b and y ions. The ETD pipeline included a non-fragment filter node which was set to remove precursor peak, charge-reduced precursor peaks, and peaks due to the neutral loss from charge-reduced precursors prior to the SEQUEST search. The PhosphoRS analysis was performed in Proteome discoverer using a peptide mass tolerance of 0.5 Da and a maximum peak depth of 8. The Ascore analysis was performed using the exported search files in Scaffold PTM (Proteome Software) using the default settings.

MALDI-TOF-MS:

MALDI TOF mass spectra were acquired on an Ultraflextreme instrument (Bruker, Billerica, MA). The instrument was calibrated using a protein mixture containing bovine insulin, MW 5733.5, bovine ubiquitin, MW 8564.8, bovine RNAse A, MW 13682.2, equine heart cytochrome C, MW 12359.9, and equine heart myoglobin, MW 16951.3 (all from Sigma); a 20 mg/mL solution of 2,5-Dihydroxybenzoic acid & 2-Hydroxy-5-methoxybenzoic acid (Super-DHB, Sigma) in 50% aqueous acetonitrile containing 0.1% Phosphoric acid (Sigma) and 0.1% Trifluoroacetic acid (Thermofisher) was used as the matrix for both the calibrants and the CTD samples. All samples for the MALDI TOF MS were prepared by mixing 1 µL of 100 µM protein solution in 0.1% Formic acid and 1 µL of the matrix solution and applying 1 µL of this mixture to a polished stainless steel target. Top down sequencing of intact CTD2′ and phosphorylated CTD2′ by ISD was performed in LIFT mode using the factory-configured instrument parameters for the 800 -5,000 *m/z* range.

NMR Spectroscopy:

CTD2′ samples were expressed as described in M9 minimal media enriched with ^{15}N-NH_4Cl and/or ^{13}C-D-Glucose (Cambridge Isotope Laboratories). Following purification, samples were buffer exchanged into 80 mM imidazole pH 6.5, 50 mM KCl, 10% glycerol, 2 mM DTT and 10% D_2O with Amicon Ultra-15 3,000 NMWL centrifugal filters (Merck Millipore Ltd.). NMR Spectra were collected at the Lloyd Jackman NMR facility at the Pennsylvania State University on Bruker Avance-III spectrometers operating at proton frequencies of 500, 600 or 850 MHz equipped with TCI single axis gradient cryoprobes (^{1}H/^{13}C/^{15}N/^{2}H) with enhanced sensitivity for ^{1}H and ^{13}C. Phosphorous experiments were collected on a Bruker Ascend spectrometer operating at a proton frequency of 500 MHz an equipped with a nitrogen cooled broadband (BBO) probe. Specific acquisition parameters are as follows:

2D ^{13}C-Detect spectra:

For routine HACACON spectra, 16 transients were collected with 1024(C)x256(N) points, sweep widths of 12x34 and a recycle delay of 1.3s. For high resolution C_CON spectra, 32 scans were collected with 1024(C)x512(N) points and sweep widths of 10x22 ppm, and 1.3s recycle delays. CAS-HACACON spectra were acquired by collecting 32 scans with 1024(C)x256(N) points at sweep widths of 20x40 ppm and 1.3s recycle delays. And CAS-HACANCO spectra were acquired by collecting 64 scans with 1024(C)x256(N) points at sweep widths of 20x40 ppm and 1.3s recycle delays.

3D ^{13}C-Detect spectra:

For the 3D (HACA)N(CA)CON spectrum, 16 transients were acquired with 1024(C)x64(N)x128(N) complex points, sweep widths of 12x34x34 ppm, and a recycle delay of 1s. For the 3D (HACA)N(CA)NCO spectrum, 32 transients were acquired using 30% non-uniform sampling, with 1024(C)x64(N)x128(N) complex points, sweep widths of 12x34x34 ppm, and a recycle delay of 1s.

^{1}H-Detect spectra:

Routine HSQC spectra were collected with 4-16 scans, 2048(H)x256(N) complex points, sweep widths of 12x22 ppm, and recycle delays of 1-1.3s. For the HNCO spectrum, 8 transients were acquired using 40% non-uniform sampling, with 2048(H)x64(N)x256(C) complex points, sweep widths of 12x32x22 ppm, and a recycle delay of 1.3s. For the HNCACO spectrum, 32 transients were acquired using 30% non-uniform sampling, with 2048(H)x64(N)x256(C) complex points, sweep widths of 12x32x22 ppm, and a recycle delay of 1.3s. HNCACB spectra, were acquired with 32 scans with 2048(H)x64(N)x180(C) complex points, sweep widths of 12x32x75 ppm, and a recycle delay of 1.3s using 40% non-uniform sampling. For CBCACONH spectra, 16 scans with 2048(H)x64(N)x152(C) complex points were acquired, at sweep widths of 12x32x75 ppm, and a recycle delay of 1.3 s using 40% non-uniform sampling. DIPSI-HSQC spectra were collected with 8 scans, 2048(H)x64(N)x256(H) complex points, sweep widths of 12x32x12 ppm, and a recycle delay of 1.3 s using 40% non-uniform sampling.

^{31}P-Detect spectra:

For ^{31}P spectra, a 1D sequence with power gated decoupling (WALTZ-16) using a 30^{o} flip angle was used. Spectra were collected using 2046 scans, sweep widths of 50 ppm and recycle delays of 2s.

Data processing and analysis:

^{13}C chemical shifts were referenced to a DSS standard and ^{31}Pchemical shifts were referenced to a coaxial phosphoric acid standard. Spectra were processed in Topspin 3.2 (Bruker) and analyzed in Sparky or Mnova (Mestrelab research). NUS spectra were processed using the MDD algorithm. Average chemical shift perturbations for ΔCTD2′ upon phosphorylation were calculated using:

$$\Delta\delta_{\mathrm{AV}} = [\frac{1}{n}\left\{\sum_{1}^{n}(\alpha\,\Delta\delta)^{2}\right\}]^{1/2}$$

Where n is the number of chemical shifts, Δδ is the difference in chemical shift between the unphosphorylated and phosphorylated species, and α is the scaling parameter (0.102 for ^{15}N & 0.251 for ^{13}C).

References

1. Zehring, W. A., Lee, J. M., Weeks, J. R., Jokerst, R. S., Greenleaf, A. L. (1988) The C-terminal repeat domain of RNA polymerase II largest subunit is essential in vivo but is not required for accurate transcription initiation in vitro. Proc. Natl. Acad. Sci. U. S. A. 85: 3698-3702.
2. Jeronimo, C., Bataille, A. R., & Robert, F. (2013) The Writers, Readers, and Functions of the RNA Polymerase II C-Terminal Domain Code. Chemical Reviews. 113: 8491-8522.
3. Buratowski, S. The CTD code. Nature Structural Biology. 10: 679-680. (2003).
4. Corden, J. L. (2013) RNA Polymerase II C-Terminal Domain: Tethering Transcription to Transcript and Template. Chemical Reviews. 113: 8423–8455.
5. Cagas, P. M., & Corden, J. L. (1995) Structural Studies of a Synthetic Peptide Derived From the Carboxyl-Terminal Domain of RNA Polymerase II. Proteins: Structure, Function, and Genetics. 21: 149-160.
6. Noble, C. G., Hollingworth, D., Martin, S. R., Ennis-Adeniran, V., Smerdon, S. J., Kelly, G. Taylor, I. A., & Ramos, A. (2005) Key features of the interaction between Pcf11 CID and RNA polymerase II CTD. Nature Structural & Molecular Biology. 12: 144-151
7. Schuller, R. et al. (2016) Heptad-Specific Phosphorylation of RNA Polymerase II CTD. Mol Cell 61, 305-14
8. Suh, H. et al. (2016) Direct Analysis of Phosphorylation Sites on the Rpb1 C-Terminal Domain of RNA Polymerase II. Mol Cell 61, 297-304
9. Zhang, Z., & Gilmour, D. S. (2006) Pcf11 is a termination factor in Drosophila that dismantles the elongation complex by bridging the CTD of RNA polymerase II to the nascent transcript. Molecular Cell. 21: 65-74.
10. Czudnochowski, N., Bösken, C. A., & Geyer, M. (2012) Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition. 3:842, 1-12.
11. Greifenberg, A. K., Honig, D., Pilarova, K., Duster, R., Bartholomeeusen, K., Bösken, C. A., Anand, K., Blazek, D & Geyer, M. (2016) Structural and Functional Analysis of the Cdk13/Cyclin K Complex. Cell Reports 14, 320–331.
12. Bösken, C. A., Farnung, L., Hintermair, C., Schachter, M. M., Vogel-Bachmayr, K., Blazek, D., Anand, K., Fisher, R. P., Eick, D., & Geyer, M. (2013) The structure and substrate specificity of human Cdk12/Cyclin K. Nature Communications. 5:3505.
13. Boersema, P. J., Mohammed, S., & Heck, A. J. R. (2009) Phosphopeptide fragmentation and analysis by mass spectrometry. J. Mass. Spectrom. 44, 861–878.
14. Eng, J.K., McCormack, A.L. & Yates, J.R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5: 976.

15. Taus,T., Kocher, T., Pichler, P., Paschke, C., Schmidt, A., Henrich, C., & Mechtler, K. (2011) Universal and Confident Phosphorylation Site Localization Using phosphoRS. J. Proteome Res. 10, 5354–5362.
16. Beausolei, S. A., Villen, J., Gerber, S. A., Rush, J. & Gygi, S. P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nature Biotechnology. 24:10, 1285-1292.
17. Bastidas, M., Gibbs, E.B., Sahu, D. & Showalter, S.A. (2015) A primer for carbon-detected NMR applications to intrinsically disordered proteins in solution. Concepts in Magnetic Resonance Part A 44, 54-66
18. Gibbs, E.B. & Showalter, S.A. (2015) Quantitative biophysical characterization of intrinsically disordered proteins. Biochemistry 54, 1314-26
19. Sahu, D., Bastidas, M. & Showalter, S.A. (2014) Generating NMR Chemical Shift Assignments of Intrinsically Disordered Proteins using Carbon-Detected NMR Methods. Anal Biochem 449, 17-25
20. F.-X. Theillet, C. Smet-Nocca, S.Liokatis, R. Thongwichian, J. Kosten, M. K. Yoon, R. W. Kriwacki, I. Landrieu, G. Lippens, P. Selenko, (2012) J. Biomol. NMR. 54,217 –236.
21. Bienkiewicz, E. A., & Lumb, K. J. (1996) Random-coil chemical shifts of phosphorylated amino acids. J. Biol. NMR. 15: 203-206.
22. Secci, E., Luchinat, E., Banci, L. (2016) The Casein Kinase 2-Dependent Phosphorylation of NS5A Domain 3from Hepatitis CVirus Followed by Time-Resolved NMR Spectroscopy ChemBioChem, 17,328 –333.
23. Cordier, F., Chaffotte, A. Wolff, N., (2015) Quantitative and dynamic analysis of PTEN phosphorylation by NMR. Methods. 77–78. 82–91.
24. Lennon, J. J., & Walsh, K. A. (1999) Locating and identifying posttranslational modifications by in-source decay during MALDI-TOF mass spectrometry. Protein Science. 8:2487–2493.
25. Selenko, P., Frueh, D. P., Elsaesser, S. J., Haas, W., Gygi S. P., & Wagner G. (2008) In situ observation of protein phosphorylation by high-resolution NMR spectroscopy. Nature Structural & Molecular Biology. 15: 321-329.
26. Prabakaran, S., Everley, R. A., Landrieu, I., Wieruszeski, J. M., Lippens, G., Steen, H., & Gunawardena, J. (2011) Comparative analysis of Erk phosphorylation suggests a mixed strategy for measuring phospho-form distributions. Molecular Systems Biology. 7: 482.
27. Bah, A. et al. (2015) Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. Nature. 519, 106-9
28. Debois, D., Smargiasso, N., Demeure, K., Asakawa, D., Zimmerman, T. A., Quinton, L., & De Pauw, E. (2013) MALDI In-Source Decay, from Sequencing to Imaging Top Curr Chem. 331: 117–142.

Chapter 4

DmP-TEFb Specificity by Probed by MALDI MS

[MS/MS data presented in this chapter was collected by E. Gibbs & T. Laremore, at the Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, Pennsylvania. Expression and purification of DmP-TEFb was performed by B. Portz]

The positive transcription elongation factor b (P-TEFb) promotes transcription elongation through phosphorylation of the C-terminal domain of the RNA Polymerase II large subunit (CTD). This process is not well understood, partly due to difficulties in determining P-TEFb specificity toward the CTD. A simple assay was developed to identify the substrate specificity of DmP-TEFb *in vitro*. This demonstrated that DmP-TEfb preferentially phosphorylates Ser5 within the CTD heptad, and surprisingly, revealed that Tyr1 is required for this specificity.

Introduction

The positive transcription elongation factor b (P-TEFb) promotes transcription elongation through phosphorylation of the C-terminal domain of the RNA Polymerase II large subunit (CTD)., which is composed of tandem heptad repeats of the consensus sequence ($Y^1S^2P^3T^4S^5P^6S^7$). This process is not well understood, partly due to difficulties in determining P-TEFb specificity toward the CTD. Studies of CDK2-cyclinA established the canonical substrate recognition motif for CDKs as S/TPXK/R, where X is any amino acid [1].However, the substrate preferences of the transcriptional CDKs are less clear. Since the initial discovery of P-TEFb, many conflicting reports have been published regarding its specificity, where Ser2, Ser5, and Ser7 activity have all been observed [2]. Recently, a chemical genetic screen identified over one hundred putative targets of human P-TEFb *in vivo*, revealing a variety of substrate peptide motifs, many bearing little similarity to the CTD heptad [3]. Thus, the *in vivo* specificity of P-TEFb toward the CTD remains a hotly contested issue. To understand the specificity of DmP-TEFb (CDK9-cyclinT1) under our *in vitro* conditions, we designed a small peptide library modeled around the consensus heptad sequence. (Table 4-1). Each peptide consisted of two repeats, arranged to provide nominally one phospho-site per peptide. Peptides were kinase treated *in vitro* and the reaction products were analyzed by MALDI mass spectrometry.

Results & Discussion

Analysis of the WT peptide by MALDI-TOF-MS demonstrated complete phosphorylation of the substrate by DmP-TEFb, resulting in a primarily mono-phosphorylated product. (Figure 4-2A) MALDI tandem MS spectra (MS2) acquired for the mono-phosphorylated precursor ion at 1537 m/z revealed fragment ions consistent with a pSer5 state and a pSer2 state. (Figure 4-3A) Using the peak intensities of the various b and y-ions the populations of these two species could be quantified, revealing 83% pSer5 and 17% pSer2. (Figure 4-1A) A di-phosphorylated product at 1617 m/z, corresponding to a pSer2/pSer5 species was also observed, however, this species was only present to ~1% of the total population. (Figure 4-2A,3B) Similarly, a doubly phosphorylated species could not be generated using a peptide pre-phosphorylated at Ser5. (Figure 4-1B) Thus, DmP-

TEFb can place one phosphorylation mark per heptad and primarily targets the Serine 5 residue.

Table 4-1. Model CTD peptides for analysis by MALDI-MS.

Peptide	AA Sequence	Molecular Mass (Da)
WT	SPSYSPTSPSYSPT	1457.51
CTDS2A	SPSYAPTSPSYSPT	1441.51
CTDS5A	SPSYSPTAPSYSPT	1441.51
CTDS5P	SPSYSPTSPpSYSPT	1537.60
CTDY1A	SPSASPTSPSYSPT	1365.42
CTDY2A	SPSYSPTSPSASPT	1365.42
CTDY12A	SPSASPTSPSASPT	1273.32

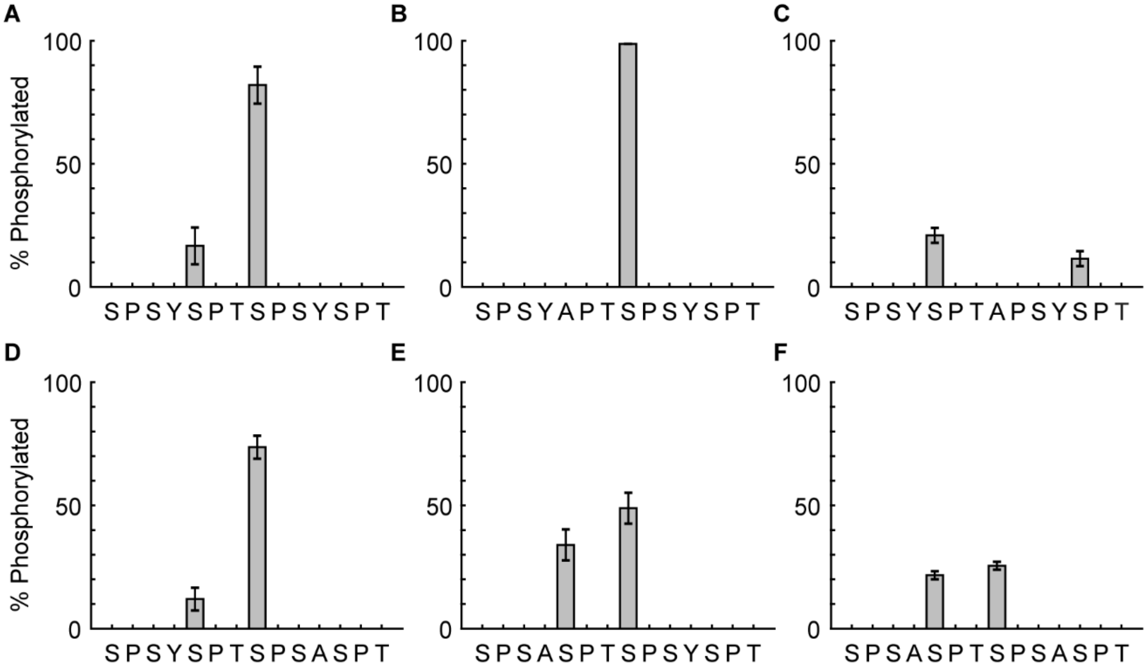


Figure 4-1. DmP-TEFb activity and specificity probed by MALDI-MS2. DmP-TEFb produced mixtures of mono-phosphorylated species *in vitro*. Using MS2, individual phosphoisoforms were identified and quantified. Shown here are the percentages of each phosphorylated residue; where the error bars represent the S.E.M (A) DmP-TEFb preferentially targets the Ser5 residue. (B) Mutation of Ser2 results in an exclusively pSer5 state. (C) By contrast, mutation Ser5 of disrupts DmP-TEFb specificity resulting in a mixture of pSer2 species. (D) Mutation of Tyr in the following

repeat does not significantly alter DmP-TEFb activity or specificity. (E) However, mutation of the internal Tyr1 results in a substantial loss of specificity. (F) And mutation of both Tyr residues results in a loss of activity and specificity.

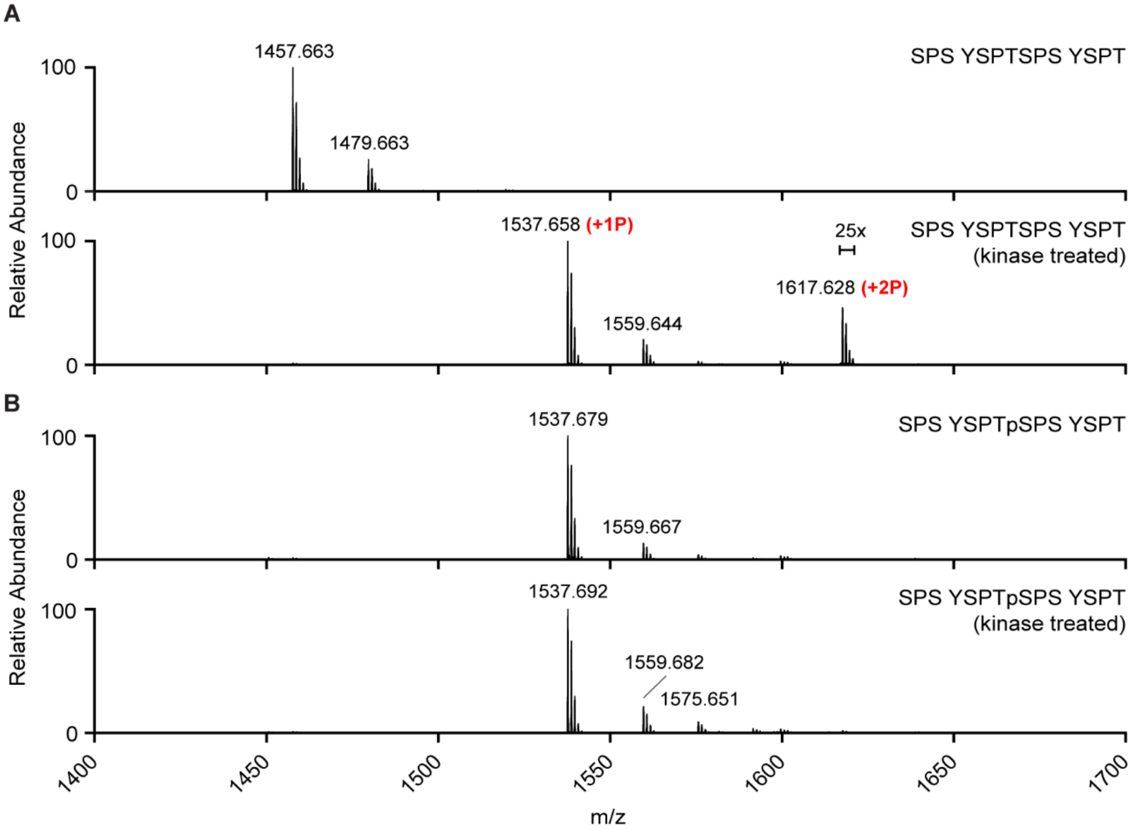


Figure 4-2. MALDI-TOF-MS spectra of DmP-TEFb treated CTD peptides. (A) The WT peptide in its unphosphorylated (top) and kinase treated states (bottom) demonstrates that DmP-TEFb generates a primarily mono-phosphorylated product. (B) The pre-phosphorylated S5P peptide in its unphosphorylated (top) and kinase treated states (bottom) shows that DmP-TEFb cannot efficiently generate a doubly phosphorylated species.

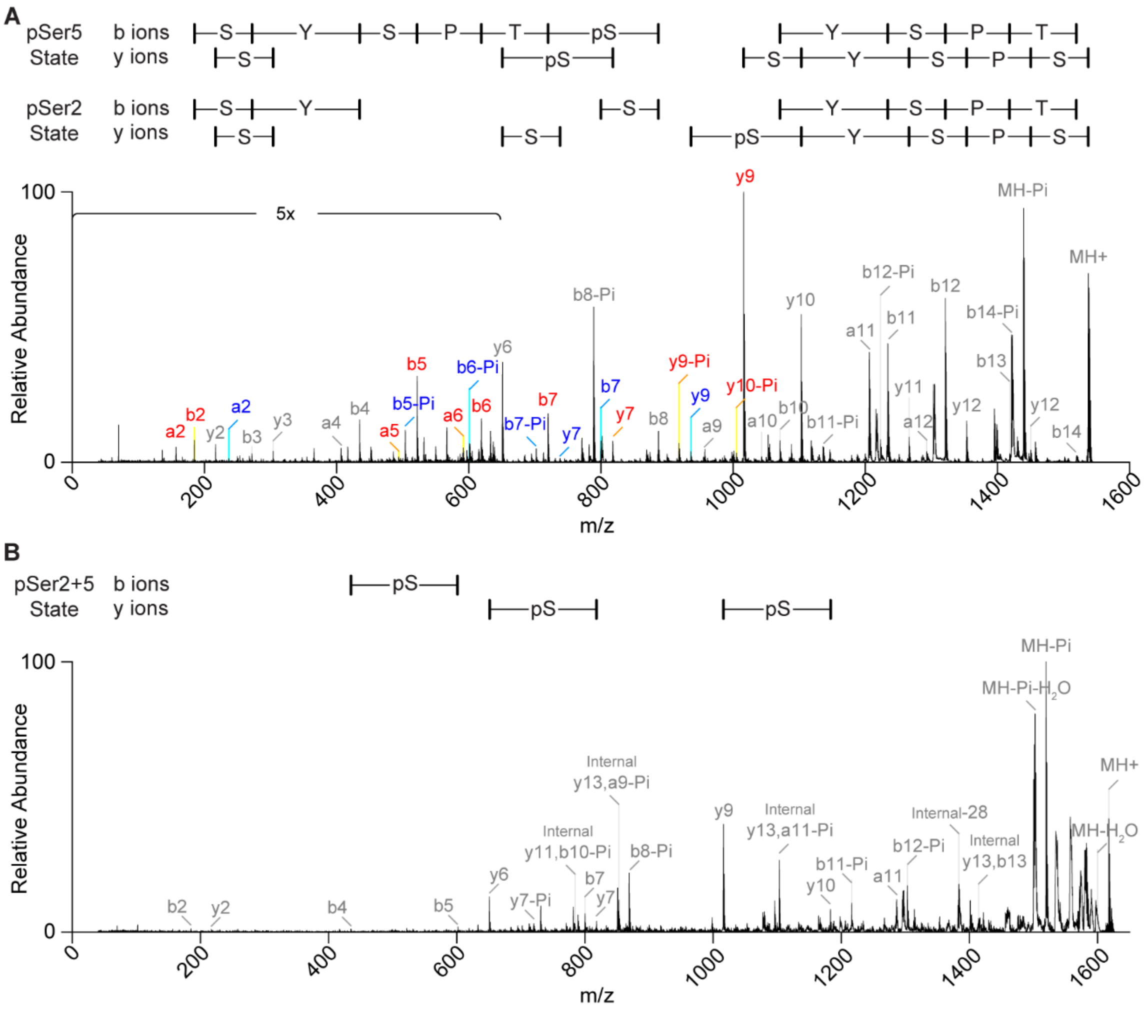


Figure 4-3. MALDI-MS2 spectra of DmP-TEFb treated CTD peptides. (A) Peptide fragment spectra of the kinase treated WT peptide (precursor @ 1537 m/z) reveals a mixture of mono-phosphorylated pSer2 and pSer5 species following phosphorylation by DmP-TEFb. (B) The kinase treated WT peptide (precursor @ 1617 m/z) shows a minor pSer2/pSer5 di-phosphorylated species. Red and blue labeled peaks are consistent with the dominant and minor species, respectively. Grey labeled peaks are consistent with all states.

As a cross check, alanine mutant peptides were also tested. As expected, mutation of Ser2 to alanine had no impacts on the reaction efficiency. (Figure 4-4A) And analysis by MS2 showed that an exclusively pSer5 species was generated. (Figure 4-1B,5A) Interestingly, the Ser5 to Ala mutant generated a mixture of pSer2 species, albeit with less efficiency (~ 30%). (Figure 4-1C,4B,5B) Thus, in the absence of the preferred target, DmP-TEFb can phosphorylate Ser2 above basal levels. However, this is accompanied by a loss in both activity and specificity.

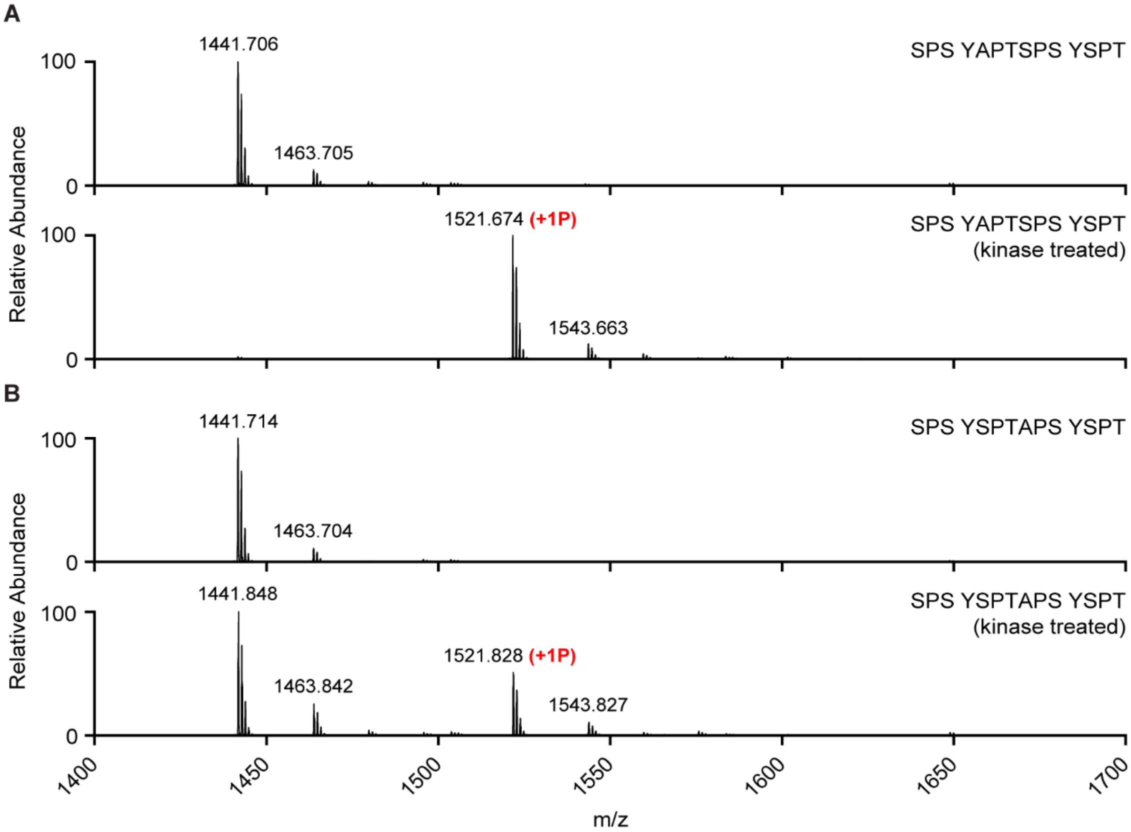


Figure 4-4. MALDI-TOF-MS spectra of DmP-TEFb treated mutant CTD peptides. (A) The S2A peptide in the unphosphorylated (top) and kinase treated states (bottom) demonstrates that DmP-TEFb generates a primarily mono-phosphorylated product with WT efficiency. (B) The S5A peptide in its unphosphorylated (top) and kinase treated states (bottom) shows that DmP-TEFb generates a primarily mono-phosphorylated product with reduced efficiency.

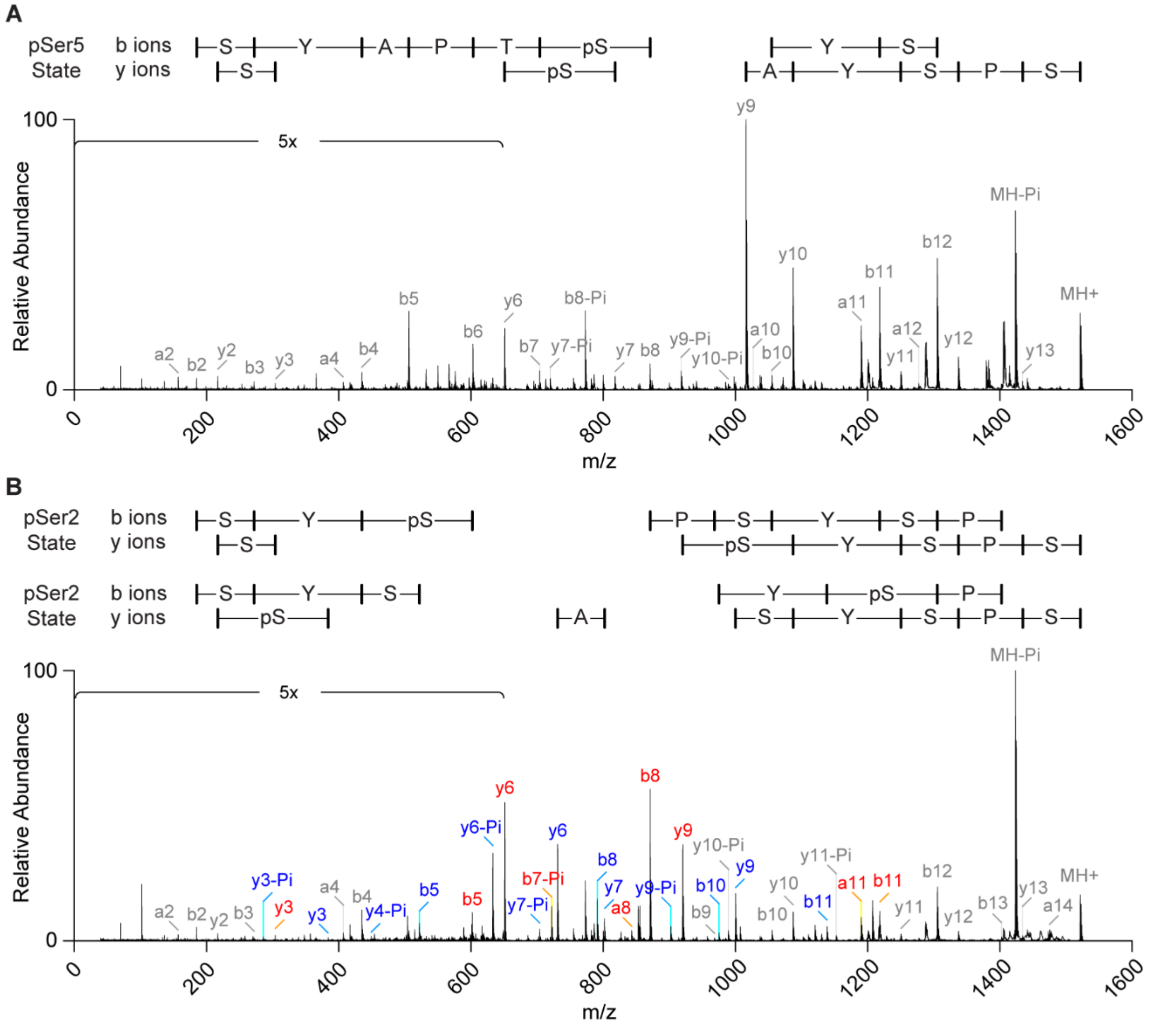


Figure 4-5. MALDI-MS2 spectra of DmP-TEFb treated mutant CTD peptides. (A) Peptide fragment spectra of the kinase treated S2A peptide (precursor @ 1521 m/z) shows an exclusively pSer5 state. (B) By contrast the kinase treated S5A peptide (precursor @ 1521 m/z) contains a mixture of pSer2 species. Red and blue labeled peaks are consistent with the dominant and minor species, respectively. Grey labeled peaks are consistent with all states.

This simple assay demonstrates that under our *in vitro* conditions DmP-TEFb preferentially targets the Ser5 position of the CTD heptad repeat. We wondered if it could also provide any insight into to the preferred orientation of the CTD heptad within the active site of DmP-TEFb. One long standing difficulty in broadly understanding the structural basis for CTD kinase specificity has been the lack of crystallographic information on the bound conformations of CTD substrate peptides. However, many have drawn inference from the structure of CDK2-CyclinA bound to the optimal substrate peptide (HHASPRK). (Figure 4-6) This structure reveals the importance of the proline residue +1 to the target serine and the lysine residue in the +3 position. Any residue besides Pro at the +1 position would be energetically unfavorable due to an unsatisfied backbone amide hydrogen bond. Further, the *trans*-conformation of the proline directs the +3 lysine side chain toward the pThr moiety within the activation T-loop [4]. This observation led Czudnochowski et al. to hypothesize that since a tyrosine residue occupies this position in the CTD heptad (S_5PSY_1 in this context), it would play an important role in defining the specificity of CDK9-CyclinT, toward the CTD [5].

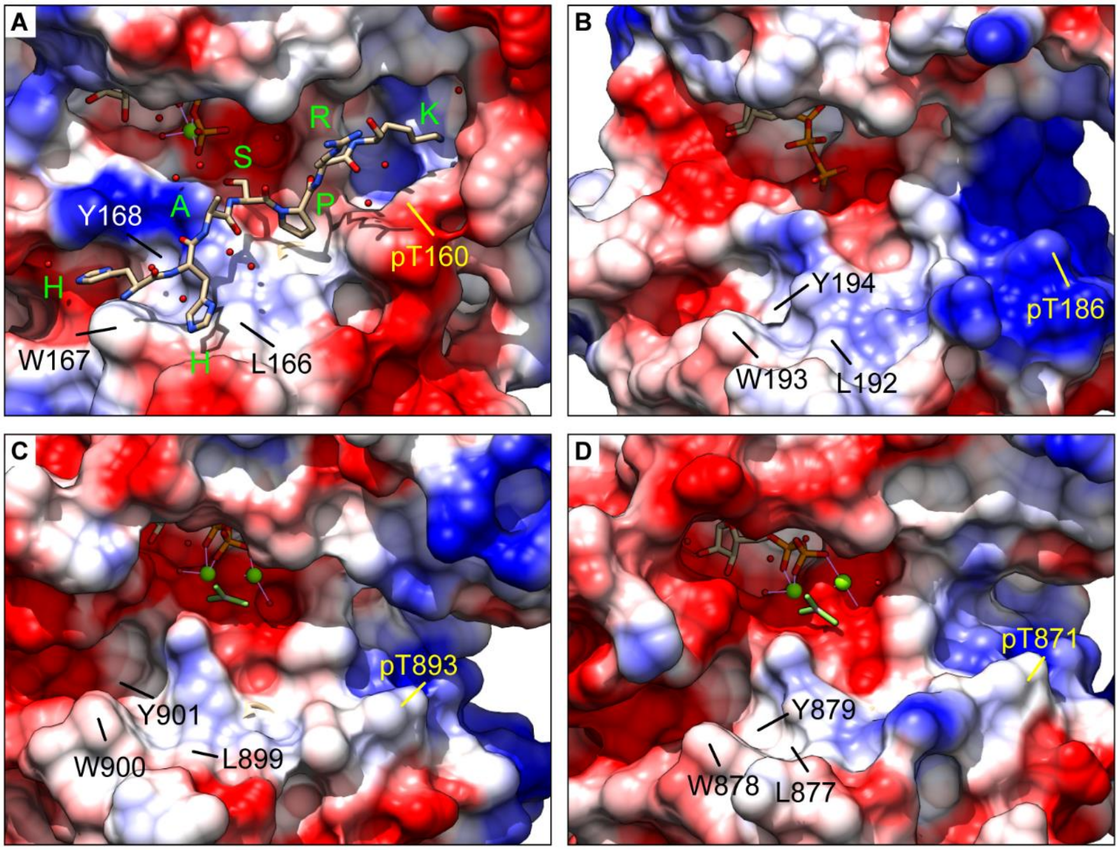


Figure 4-6. Active sites of various CDKs. The active sites of CDK2-CyclinA (PDB: 1QMZ) with the bound optimal peptide substrate (HHASPRK) (A), CDK9-CyclinT (PDB: 3BLQ) (B), CDK12-CyclinK (PDB: 4NST) (C), and CDK13-CyclinK (PDB: 5EFQ) (D) are shown in surface representation. Colors indicate the surface charge calculated using a Coulombic potential in Chimera, where the most intense red and blue areas correspond to -7 and +7 kcal/mol*e, respectively. In each panel, the activating pThr residue and several conserved active site residues are labeled.

To test this possibility, we assayed the ability of DmP-TEFb to phosphorylate CTD peptides containing various Tyr to Ala mutations. Surprisingly, mutation of the Tyr1 residue in the +1 heptad had only minimal effects on the reaction efficiency (~85 % completion). (Figure 4-7A). And essentially the same products were generated as with the wild-type peptide. (Figure 4-1D,8A) By contrast, mutation of the Tyr1 residue within the internal repeat severely disrupted DmP-TEFb specificity, resulting in almost equal amounts of pSer2 and pSer5 species (~40 % pSer2, 60% pSer5). (Figure 4-1E,7B,8B) And, mutation of both Tyr residues to Ala, dramatically reduced the reaction efficiency (~47%) and produced an equal mixture of pSer2 and pSer5 (46% pSer2, 54% pSer5). (Figure 4-1F,7C, 9) Thus, DmP-TEFb specificity toward the CTD is enforced by the Tyr1 residue of the heptad repeat. This suggests that Tyr1 effectively sets the register of the CTD within the active site of P-TEFb.

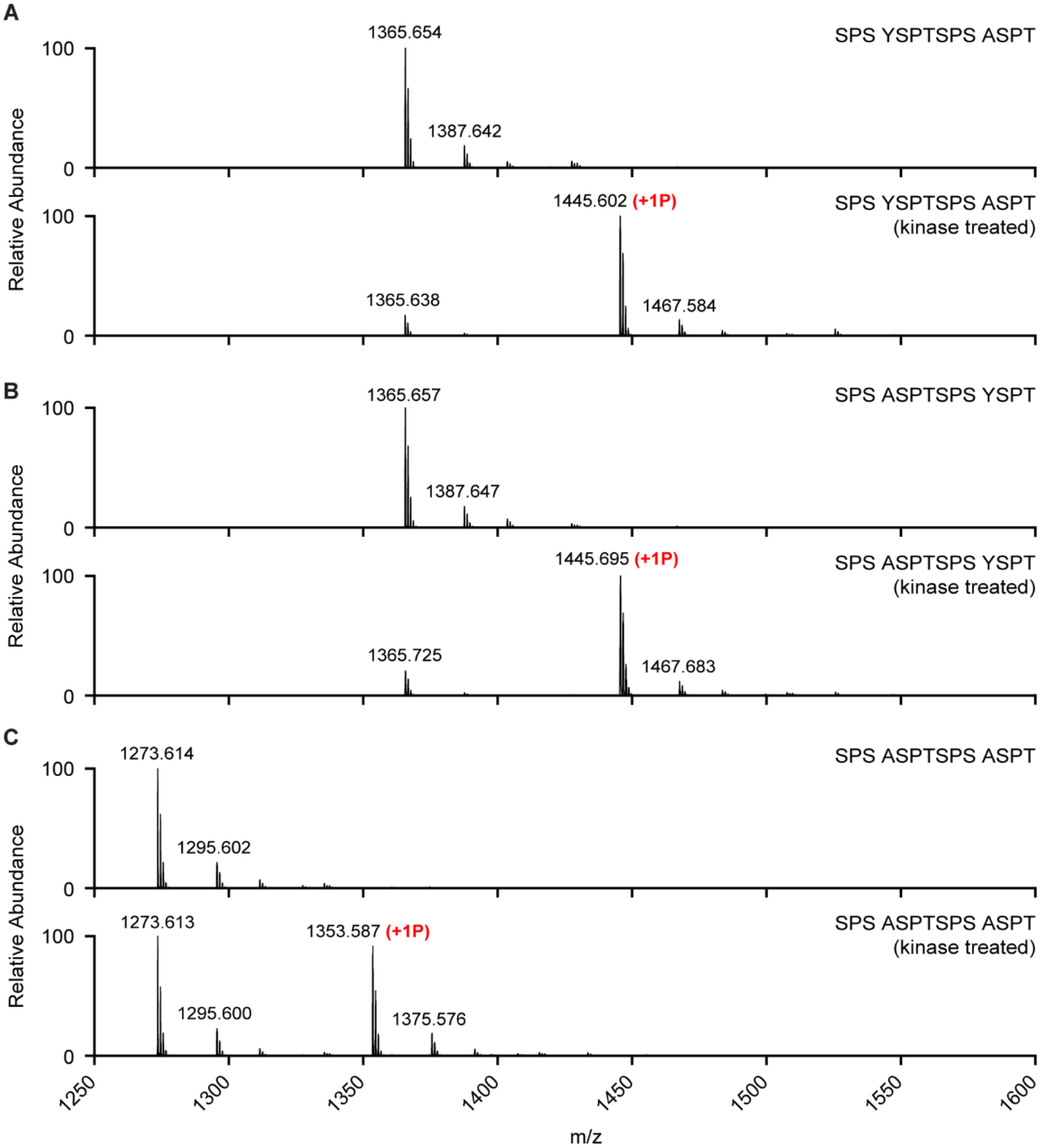


Figure 4-7. MALDI-TOF-MS spectra of DmP-TEFb treated Tyr mutant CTD peptides. (A) The Y2A peptide and (B) Y1A peptide in unphosphorylated (top) and kinase treated states (bottom) demonstrate that DmP-TEFb generates primarily mono-phosphorylated products with slightly less efficiency upon single Y-to-A substitutions. (C) The Y12A peptide in its unphosphorylated (top) and kinase treated states (bottom) demonstrates that DmP-TEFb generates a mono-phosphorylated product with severely reduced efficiency when both Tyr residues are mutated to Ala.

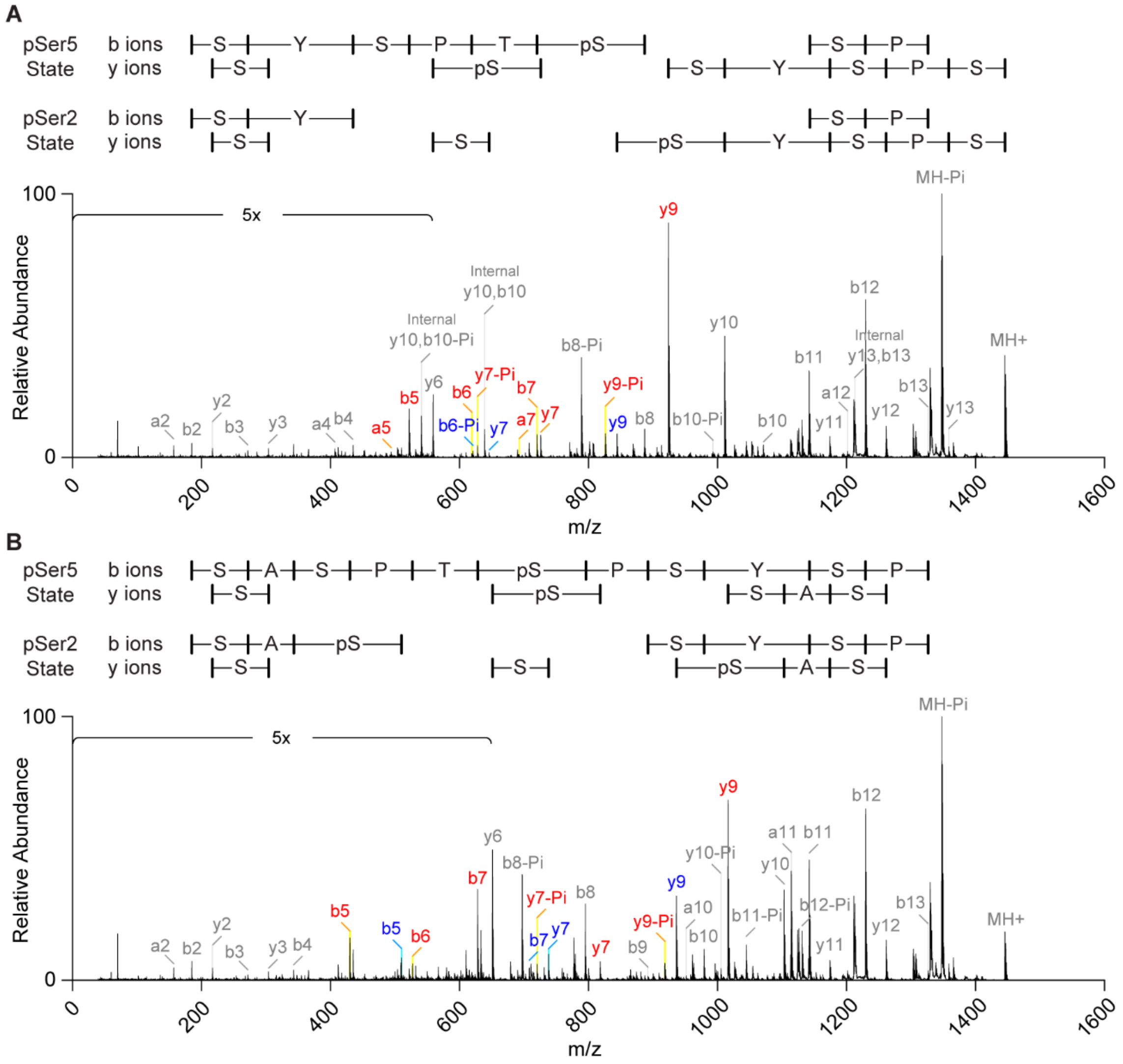


Figure 4-8. MALDI-MS2 spectra of DmP-TEFb treated Tyr mutant CTD peptides. Peptide fragment spectra of the kinase treated Y2A (precursor @ 1445 m/z) (A) and Y1A peptide (precursor @ 1445 m/z) (B) show the loss of DmP-TEFb specificity upon Tyr mutation. In both cases, mixtures of pSer2 and pSer5 phosphoisoforms were observed. However, mutation of Tyr1 within the internal heptad motif has a more deleterious effect. Grey labeled peaks are consistent with all states. Red and blue labeled peaks are consistent with the dominant and minor species, respectively.

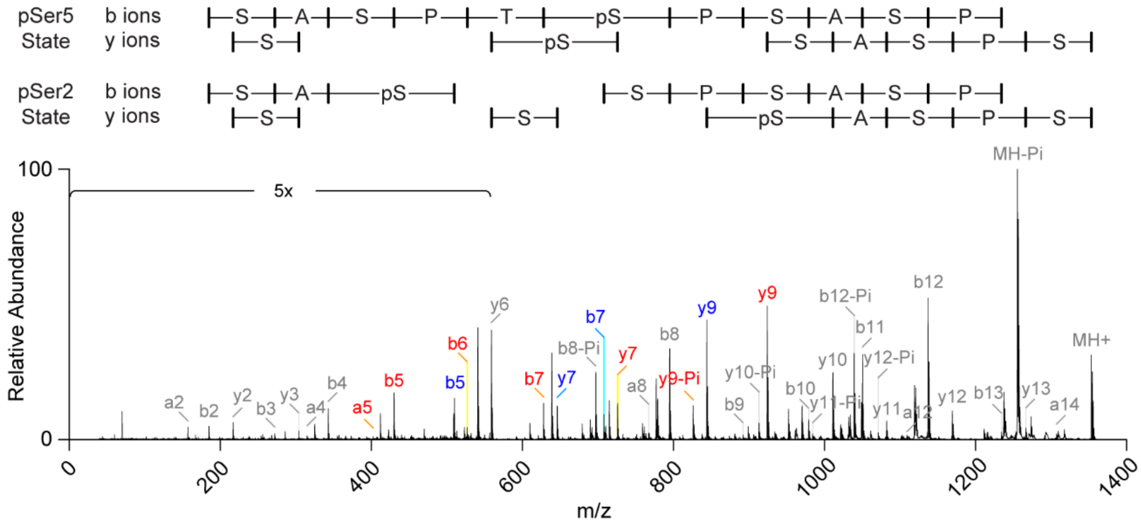
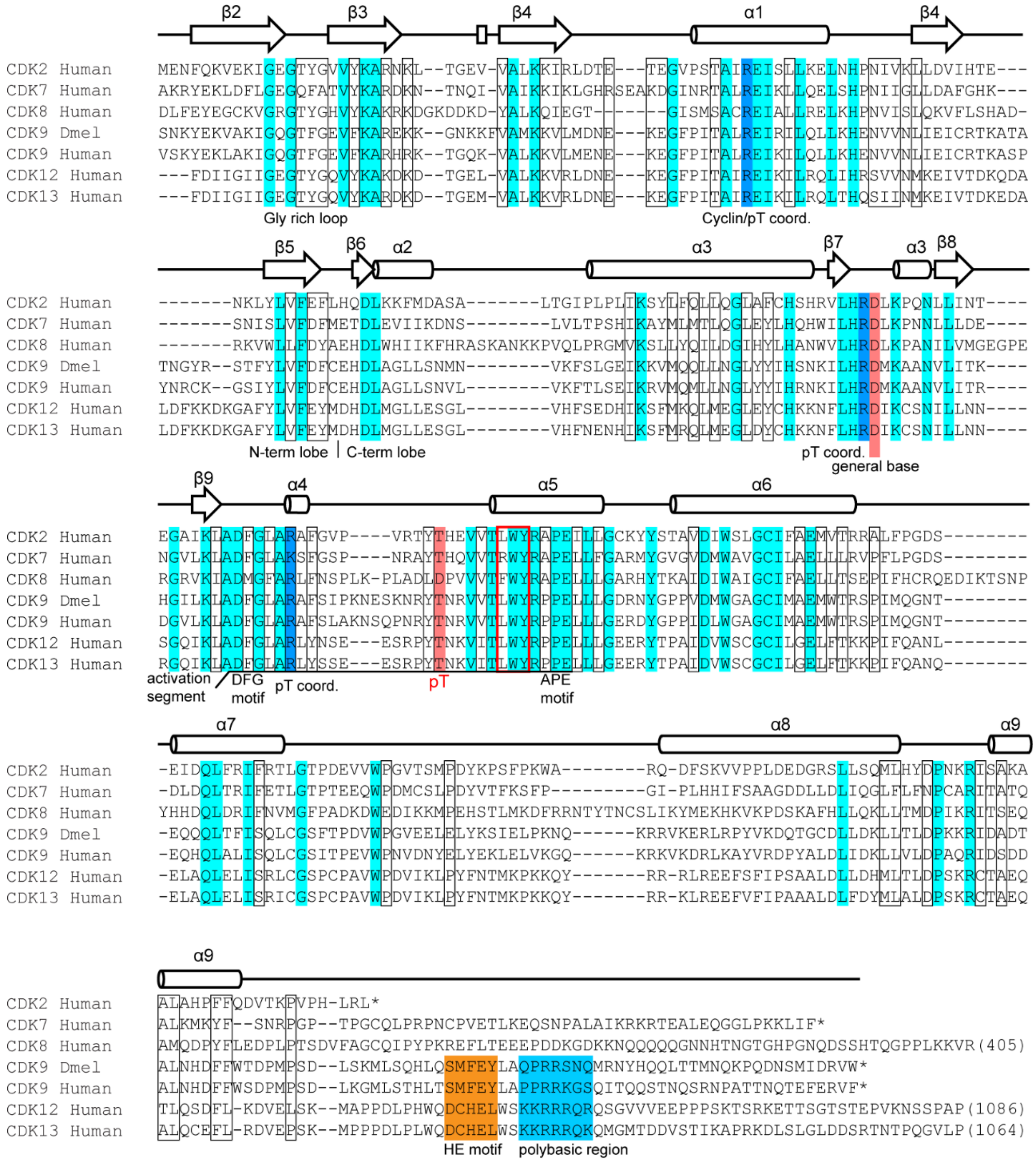


Figure 4-9. MALDI-MS2 spectra of a DmP-TEFb treated CTD peptide lacking Tyr1. Peptide fragment spectra of the kinase treated Y12A peptide (precursor @ 1353 m/z) shows that the mutation of both Tyr residues causes a severe loss in substrate specificity. Grey labeled peaks are consistent with all states. Red and blue labeled peaks are consistent with the dominant and minor species, respectively.

In order to understand how the CTD could be accommodated in the active site of P-TEFb, one can consider the published crystal structures of various CDKs. Comparing the structures of CDK2-CyclinA to CDK9-CyclinT, CDK12-CyclinK and CDK13-CyclinK, three CDKs known to have Ser5 activity, it is apparent that the surface around the pThr residue in the T-Loop has the opposite charge. (Figure 4-6) Indeed, these positive patches were hypothesized to account for the enhancement in Ser5 activity observed for CTD substrates containing pSer7. However, beyond the potential for favorable electrostatic interactions, it is not clear what advantage the Tyr1 residue could provide in this region. By contrast, a highly conserved feature of the activation segment is a LWY motif (Figure 4-10), which creates a hydrophobic groove in the active site. In CDK2-CyclinA, the -2 His residue of the substrate sits in this groove, while the -3 His forms contacts with W167. If the CTD heptad were to assume a similar register in the active site of CDK9, with Ser5 as the target residue, Pro3 could be accommodated in this groove and Tyr1 would be in close proximity to W193. Indeed Tyr/Trp interactions have been shown to provide significant stabilization [6]. And in this context favorable interactions, e.g. π-π stacking, between Tyr1 and W199 in *Drosophila melanogaster* CDK9, may thus account for proper alignment of the substrate within the active site.



```
CDK2 Human  MENFQKVEKIGEGTYGVVYKARNKL--TGEV-VALKKIRLDTE---TEGVPSTAIREISLLKELNHPNIVKLLDVIHTE---
CDK7 Human  AKRYEKLDFLGEGQFATVYKARDKN--TNQI-VAIKKIKLGHRSEAKDGINRTALREIKLLQELSHPNIIGLLDAFGHK---
CDK8 Human  DLFEYEGCKVGRGTYGHVYKAKRKDGKDDKD-YALKQIEGT------GISMSACREIALLRELKHPNVISLQKVFLSHAD-
CDK9 Dmel   SNKYEKVAKIGQGTFGEVFKAREKK--GNKKFVAMKKVLMDNE---KEGFPITALREIRILQLLKHENVVNLIEICRTKATA
CDK9 Human  VSKYEKLAKIGQGTFGEVFKARHRK--TGQK-VALKKVLMENE---KEGFPITALREIKILQLLKHENVVNLIEICRTKASP
CDK12 Human ---FDIIGIIGEGTYGQVYKAKDKD--TGEL-VALKKVRLDNE---KEGFPITAIREIKILRQLIHRSVVNMKEIVTDKQDA
CDK13 Human ---FDIIGIIGEGTYGQVYKARDKD--TGEM-VALKKVRLDNE---KEGFPITAIREIKILRQLTHQSIINMKEIVTDKEDA

CDK2 Human  -------NKLYLVFEFLHQDLKKFMDASA-------LTGIPLPLIKSYLFQLLQGLAFCHSHRVLHRDLKPQNLLINT----
CDK7 Human  -------SNISLVFDFMETDLEVIIKDNS--------LVLTPSHIKAYMLMTLQGLEYLHQHWILHRDLKPNNLLLDE----
CDK8 Human  -------RKVWLLFDYAEHDLWHIIKFHRASKANKKPVQLPRGMVKSLLYQILDGIHYLHANWVLHRDLKPANILVMGEGPE
CDK9 Dmel   TNGYR--STFYLVFDFCEHDLAGLLSNMN--------VKFSLGEIKKVMQQLLNGLYYIHSNKILHRDMKAANVLITK----
CDK9 Human  YNRCK--GSIYLVFDFCEHDLAGLLSNVL--------VKFTLSEIKRVMQMLLNGLYYIHRNKILHRDMKAANVLITR----
CDK12 Human LDFKKDKGAFYLVFEYMDHDLMGLLESGL--------VHFSEDHIKSFMKQLMEGLEYCHKKNFLHRDIKCSNILLNN----
CDK13 Human LDFKKDKGAFYLVFEYMDHDLMGLLESGL--------VHFNENHIKSFMRQLMEGLDYCHKKNFLHRDIKCSNILLNN----

CDK2 Human  EGAIKLADFGLARAFGVP----VRTYTHEVVTLWYRAPEILLGCKYYSTAVDIWSLGCIFAEMVTRRALFPGDS--------
CDK7 Human  NGVLKLADFGLAKSFGSP----NRAYTHQVVTRWYRAPELLFGARMYGVGVDMWAVGCILAELLLRVPFLPGDS--------
CDK8 Human  RGRVKIADMGFARLFNSPLK-PLADLDPVVVTFWYRAPELLLGARHYTKAIDIWAIGCIFAELLTSEPIFHCRQEDIKTSNP
CDK9 Dmel   HGILKLADFGLARAFSIPKNESKNRYTNRVVTLWYRPPELLLGDRNYGPPVDMWGAGCIMAEMWTRSPIMQGNT--------
CDK9 Human  DGVLKLADFGLARAFSLAKNSQPNRYTNRVVTLWYRPPELLLGERDYGPPIDLWGAGCIMAEMWTRSPIMQGNT--------
CDK12 Human SGQIKLADFGLARLYNSE---ESRPYTNKVITLWYRPPELLLGEERYTPAIDVWSCGCILGELFTKKPIFQANL--------
CDK13 Human RGQIKLADFGLARLYSSE---ESRPYTNKVITLWYRPPELLLGEERYTPAIDVWSCGCILGELFTKKPIFQANQ--------

CDK2 Human  -EIDQLFRIFRTLGTPDEVVWPGVTSMPDYKPSFPKWA--------RQ-DFSKVVPPLDEDGRSLLSQMLHYDPNKRISAKA
CDK7 Human  -DLDQLTRIFETLGTPTEEQWPDMCSLPDYVTFKSFP---------GI-PLHHIFSAAGDDLLDLIQGLFLFNPCARITATQ
CDK8 Human  YHHDQLDRIFNVMGFPADKDWEDIKKMPEHSTLMKDFRRNTYTNCSLIKYMEKHKVKPDSKAFHLLQKLLTMDPIKRITSEQ
CDK9 Dmel   -EQQQLTFISQLCGSFTPDVWPGVEELELYKSIELPKNQ------KRRVKERLRPYVKDQTGCDLLDKLLTLDPKKRIDADT
CDK9 Human  -EQHQLALISQLCGSITPEVWPNVDNYELYEKLELVKGQ------KRKVKDRLKAYVRDPYALDLIDKLLVLDPAQRIDSDD
CDK12 Human -ELAQLELISRLCGSPCPAVWPDVIKLPYFNTMKPKKQY-------RR-RLREEFSFIPSAALDLLDHMLTLDPSKRCTAEQ
CDK13 Human -ELAQLELISRICGSPCPAVWPDVIKLPYFNTMKPKKQY-------RR-KLREEFVFIPAAALDLFDYMLALDPSKRCTAEQ

CDK2 Human  ALAHPFFQDVTKPVPH-LRL*
CDK7 Human  ALKMKYF--SNRPGP--TPGCQLPRPNCPVETLKEQSNPALAIKRKRTEALEQGGLPKKLIF*
CDK8 Human  AMQDPYFLEDPLPTSDVFAGCQIPYPKREFLTEEEPDDKGDKKNQQQQQGNNHTNGTGHPGNQDSSHTQGPPLKKVR(405)
CDK9 Dmel   ALNHDFFWTDPMPSD--LSKMLSQHLQSMFEYLAQPRRSNQMRNYHQQLTTMNQKPQDNSMIDRVW*
CDK9 Human  ALNHDFFWSDPMPSD--LKGMLSTHLTSMFEYLAPPRRKGSQITQQSTNQSRNPATTNQTEFERVF*
CDK12 Human TLQSDFL-KDVELSK--MAPPDLPHWQDCHELWSKKRRRQRQSGVVVEEPPPSKTSRKETTSGTSTEPVKNSSPAP(1086)
CDK13 Human ALQCEFL-RDVEPSK--MPPPDLPLWQDCHELWSKKRRRQKQMGMTDDVSTIKAPRKDLSLGLDDSRTNTPQGVLP(1064)
```

Figure 4-10. Multi sequence alignment of various CDKs. Regions of high conservation (identical) and moderate conservation are denoted by cyan bars and black boxes, respectively. Conserved active site residues in close proximity to the bound substrate in the CDK2-CyclinA structure are highlighted by the red box. Structural multi sequence alignment was generated using PROMALS3D webserver using the amino acid sequences and the structures shown in Figure 4-6 as input. Secondary structure is shown for human CDK9 (PDB: 3BLQ)

Conclusion

In addition to the features within the active sites, interactions between the kinase domains and the cyclins are crucial for the activity and specificity of CDKs. This has been shown extensively for the cell cycle CDKs. For example, in complex with cyclinA, CDK2 has a wide range of substrates, including those involved in S phase DNA replication, while the CDK2-cyclinE complex targets a subset of these factors, including p27KIP1 [4,7]. Beyond kinase activation, cyclins contain docking domains that aid in substrate recruitment. In cell cycle CDKs, these small hydrophobic patches are ~40 Å from the active site, and interact specifically with substrates containing RXL motifs [1]. Hence, substrates may interact exclusively with the active site or in bi-dentate fashion as shown for CDK2-cyclinA-CDC6. (Figure 4-11A) In this way, enhanced catalytic efficiency can be achieved by increasing the local substrate concentration [8] or by properly orientating the substrate in the kinase active site [9]. CDK9 interacts with several T-type cyclins including T1, T2a and T2b as well as cyclin K (Figure 4-11B). It is enticing to speculate that CDK9 substrate recognition may be regulated in a similar manner.

In addition to interactions with distal binding sites, it is also likely that P-TEFb specificity can be modulated through the association of regulatory factors. Binding of HIV Tat protein to P-TEFb has been shown to induce conformational changes to the substrate binding surface of CDK9 [10], which could in principle alter the substrate specificity. (Figure 4-11C). Furthermore, in the context of healthy cells, P-TEFb is found in a variety of active assembly states, including in complex with the bromodomain-containing protein 4 (BRD4) and as member of a family of elongation complexes [11]. These include the little elongation complexes, identified in Drosophila melanogaster, and the super elongation complexes (SECs), composed of members of the eleven-nineteen Lys-rich Leukemia (ELL) and AF4/FMR2 (AFF) protein families. The AFF proteins are intrinsically disordered, and act as scaffolds for SEC assembly through folding upon binding mechanisms. And, while the association of AFF4 with CDK9-cyclinT does not significantly alter the conformation of CDK9 (Figure 4-11D), HIV-1 Tat was shown to bind the P-TEFb:AFF4 complex with ~10-fold higher affinity than P-TEFb alone, and in a partially overlapping binding site [12]. This suggests that Tat binds in a distinct conformation within viral SECs, and may also hint at a general mechanism by which regulatory factors may modulate the specificity of P-TEFb through interactions with the regulatory cyclin.

The study presented here comes amidst several recent reports demonstrating the preferential specificity of P-TEFb toward the CTD Serine 5, *in vitro*. However, it is becoming increasingly clear that *in vivo*, P-TEFb specificity may be determined largely by its constituents and co-factors. Elucidating the in vivo specificity of P-TEFb will thus require the identification and characterization of its various regulatory complexes. And in all likelihood, P-TEFb specificity, like many other components of the Pol II regulatory machinery, is varied systematically in order to regulate cellular, developmental, and/or gene specific transcriptional processes.

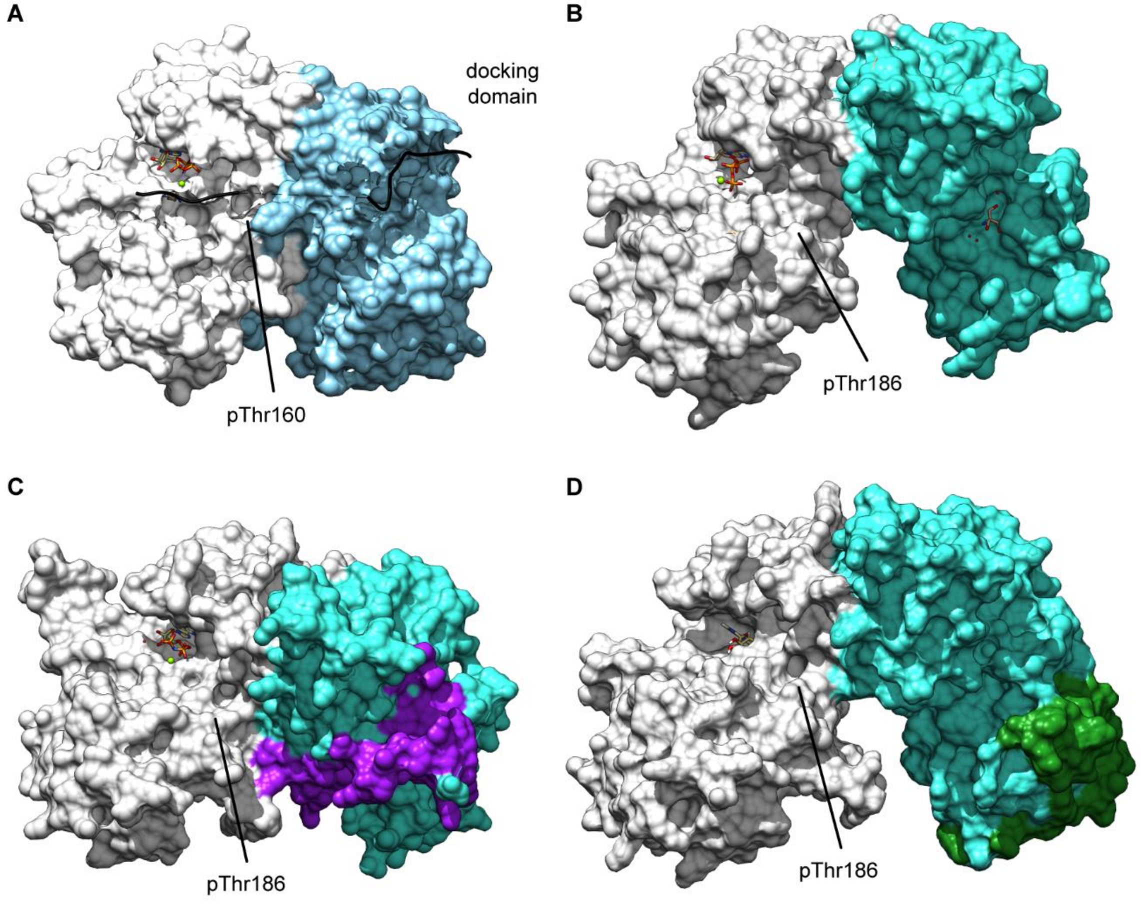


Figure 4-11. Alternative binding modes and assembly states in CDK-cyclin complexes. (A) CDK2 (white) is shown in complex with cyclinA (blue); The CDC6 peptide (black) binds the active site and distal docking domain in a bi-dentate manner. (PDB: 2CCI), (B) CDK9 (grey) is shown in complex with cyclinT1 (turquoise) (PDB: 3BLQ), where it assumes a characteristic open conformation. (C) However, binding of HIV-1 Tat (magenta) disrupts the active site conformation of CDK9 (PDB: 3MIA). (D) By contrast, AFF4 (green) binding induces little change to the conformation of CDK9 (PDB: 4IMY).

Materials & Methods:

CTD peptides were synthesized at the Tufts University Core Facility by FastMoc chemistry on an ABI 431 Peptide Synthesizer. Subsequent purification was performed by reversed phase HPLC. Purity was determined by Mass Spectrometry (>90%). Kinase reactions were performed in 50 mM Tris/HCl pH 7.5, 50 mM NaCl, 12 mM $MgCl_2$, 2 mM DTT, 12 mM ATP with 80 µg/ml DmP-TEFb and 500 µM peptide. Reactions were carried out at 24^{o}C and allowed to proceed to completion (~16 hours), after which 20 mM EDTA was added to ensure reaction termination. All samples were purified by Pierce C18 spin columns (Thermo Scientific), eluted into MS grade 70% acetonitrile (ACN), 0.1% formic acid (FA) in water and speed vacuumed to dryness. Samples were resuspended in 0.1% FA and mixed 1:1 with a 10 mg/mL solution of 4-chloro-α-cyanocinnamic acid (ClCCA, Sigma) in 50% aqueous acetonitrile containing 2.5% trifluoroacetic acid (Thermofisher). All samples were spotted on a polished stainless steel plate.

MALDI TOF MS and MS2 mass spectra were acquired on an Ultraflextreme instrument (Bruker, Billerica, MA). MS spectra were acquired in reflector positive detection mode (500-5000 *m/z* range). MS2 spectra were acquired in LIFT mode targeted for the masses of mono- and di-phosphorylated precursor ions. The instrument was calibrated using peptides from trypsin digested Bovine Serum Albumin (Sigma). MS2 spectra were assigned using Bruker BioTools peptide/protein analysis software, using the known sequence and S/T phosphorylation as a variable modification. Theoretical peptide ions were determined using the MS-product module in Protein Prospector.

References:

1.) Ubersax, J. E. & Ferrell, J. E. (2007) Mechanisms of specificity in protein phosphorylation. Nature Reviews Molecular Cell Biology. 8, 530–541.

2.) Dos Santos Paparidis, N. F., Duvale, C. Canduri, F. (2016) The emerging picture of CDK9/P-TEFb: more than 20 years of advances since PITALRE. Molecular BioSystems. DOI: 10.1039/c6mb00387g

3.) Sansó, M. Levin, R. S. Lipp, J. J. Wang, V. Y.-F., Greifenberg, A. K., Quezada, E. M., Ali, A., Ghosh, A., Larochelle, S., Rana, T. M., Geyer, M., Tong, L., Shokat, K. M. & Fisher, R. P. (2016) P-TEFb regulation of transcription termination factor Xrn2 revealed by a chemical genetic screen for Cdk9 substrates. Genes Dev. 2016 30: 117-131.

4.) Brown, N. R., Noble, M. E. M., Endicott, J. A. & Johnson, L. N. (1999) The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. Nature Cell Biology. 1. 438-443.

5.) Czudnochowski, N., Bösken, C. A., & Geyer, M. (2012) Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition. 3:842, 1-12.

6.) Wu, L., McElheny, D., Takekiyo, T., & Keiderling, T. A. (2010) Geometry and Efficacy of Cross-Strand Trp/Trp, Trp/Tyr, and Tyr/Tyr Aromatic Interaction in a β-Hairpin Peptide. Biochemistry.49, 4705–4714

7.) Nigg, E. A. (1993) Targets of cyclin-dependent protein kinases. Curr. Opin. Cell Biol. 5, 187–193.

8.) Takeda. D. Y., Wohlschlegel, J.A., & Dutta, A. (2001) A Bipartite Substrate Recognition Motif for Cyclin-dependent Kinases. The Journal of Biological Chemistry. 3:276. 1993–1997.

9.) Cheng, K.-Y., Noble, M. E. M., Skamnaki, V., Brown, N. R., Lowe, E. D., Kontogiannis, L., Shen, K., Cole, P. A., Siligardi, G., & Johnson, L. N. (2006) The Role of the Phospho-CDK2/Cyclin A Recruitment Site in Substrate Recognition. The Journal of Biological Chemistry. 32:281. 23167–23179

10.) Tahirov, T. H., Babayeva, N. D., Varzavand, K., Cooper, J. J., Sedore, S. C., & Price, D. H. (2010) Crystal structure of HIV-1 Tat complexed with human P-TEFb. Nature. 465:10. 747-751.

11.) Luo, Z., Lin, C., & Shilatifard, A. (2012) The super elongation complex (SEC) family in transcriptional control. Nature Reviews Molecular Cell Biology. 13. 543-547.

12.) Schulze-Gahmen, U., Upton, H., Birnberg, A., Bao, K., Chou, S., Krogan, N. J., Zhou, Q., Alber, T. (2013) The AFF4 scaffold binds human P-TEFb adjacent to HIV Tat. eLife. e00327. 1-14.

Chapter 5

Phosphorylation Induces Sequence-Specific Conformational Switches in the RNA Polymerase II C-Terminal Domain

[This chapter is modified from a manuscript entitled "Phosphorylation Induces Sequence-Specific Conformational Switches in the RNA Polymerase II C-Terminal Domain" that was under revision at the time this dissertation was written. Authors include: E. B. Gibbs, F. Lu, B. Portz, M. J. Fisher, B. P. Medellin, T. N. Laremore, Y. J. Zhang, D. S. Gilmour, and S. A. Showalter. Data presented in Figure 5-1 & S1 was collected and analyzed by F. Lu & D. S. Gilmour. MS data presented in Figure 5-2 was collected by T. Laremore. DmP-TEFb was provided by B. Portz and M. J. Fisher, Ssu72 was provided by B. P. Medellin and Y. J. Zhang.]

The Carboxyl-Terminal Domain (CTD) of the RNA polymerase II (Pol II) large subunit cycles through multiple phosphorylation states that correlate with progression through the transcription cycle and regulate nascent mRNA processing. Structural analyses of yeast and mammalian CTD have been hampered by their repetitive sequences. Here we identify a region of the *Drosophila melanogaster* CTD that is essential for Pol II function *in vivo* and capitalize on natural sequence variations within it to facilitate structural analysis. Mass spectrometry and NMR spectroscopy reveal that hyper-Ser5 phosphorylation transforms the local structure of this essential region via proline isomerization. The sequence context of this switch tunes the apparent activity of the CTD phosphatase Ssu72, suggesting a mechanism for the selective recruitment of *cis*-proline specific regulatory factors that may synergize with CTD phosphorylation to augment gene regulation in developmentally complex organisms.

Introduction

The carboxyl-terminal domain (CTD) of Rpb1, the largest subunit in RNA polymerase II, is an essential regulator of eukaryotic gene expression. This intrinsically disordered region (IDR), consisting of multiple tandem repeats of the consensus sequence ($Y^1S^2P^3T^4S^5P^6S^7$), acts as a scaffold for the recruitment of factors required for transcription, mRNA biogenesis, and modification of the chromatin structure [1]. Tight control over the spatial and temporal recruitment of CTD-associated factors is regulated at the molecular level in part by CTD specific kinases and phosphatases, which generate dynamic patterns of post-translational modifications (PTMs) collectively referred to as the "CTD code" [2]. While much is known about how heptads matching the consensus sequence contribute to gene expression [1], heptads that deviate from the consensus are found in all eukaryotes whose sequences are known. The number and complexity of these non-consensus heptads roughly correlate with developmental complexity [3]. Expression of genes involved in multicellularity were affected by mutating non-consensus heptads of mouse cells in culture [4], despite data indicating that the non-consensus heptads are not essential for the viability of individual cells [5]. Likewise, deletion of a small region encompassing several non-consensus heptads caused severe developmental defects and growth retardation in mice [6], demonstrating that non-consensus heptads may contribute to development and cellular differentiation.

Despite the wealth of information detailing CTD function, little is known about the molecular basis of the CTD code. The established view is that the intrinsically disordered nature of the CTD renders its structure and interactions non-specific, to be dictated predominantly by its post-translational modification status [7]. Prior structural studies, focusing on short CTD peptides composed entirely of consensus heptad repeats, revealed turn and coil structures that have been extrapolated to represent the entirety of the CTD [7,8]. The repetitive amino acid sequence comprising the CTD has been a major obstacle in studying its structure because it prevents heptad-specific interpretation of both mass spectra and NMR spectra. Therefore, recent mass spectrometry-based investigations have resorted to introducing mutations that facilitate analysis of specific regions of the CTD [9,10]. However, it is unclear how best to interpret these results molecularly because the impacts of mutations on IDP structure are often difficult to predict. Here, we turn to *Drosophila*, which is unique among commonly used model organisms in that only 2 of its 42 heptads precisely match the consensus sequence. This feature has allowed us to use NMR and mass spectrometry to precisely monitor PTM patterns and local structural features in the context of a natural CTD sequence.

Results & Discussion

Transgenic *Drosophila* Reveal a Developmentally Necessary CTD Region

To identify a functionally significant region of the CTD for our structural analysis, we tested the ability of ectopically expressed derivatives of Rpb1 to rescue the lethality caused by inhibiting expression of endogenous Rpb1. When a transgenic fly line expressing RNAi against endogenous Rpb1 in response to the GAL4 activator is mated to a fly line that ubiquitously expresses GAL4, no adult progeny are produced. This lethality is overcome by co-expressing a wild-type version of Rpb1 that has been rendered resistant to the RNAi through synonymous substitutions in the Rpb1 coding sequence (Fig. 5-1A). To identify regions of the CTD essential for development of an adult fly, several deletions were made in the CTD of the RNAi-resistant form of Rpb1 and tested for their ability to rescue the lethality associated with depleting the endogenous Rpb1 (Fig. 5-1B, Fig. 5-S1A,B). Western blot analysis with antibody against the ectopically expressed Rpb1 indicated that each derivative was expressed at comparable levels in tissues derived from pupae, indicating that each Rpb1 variant was equally stable (Fig. 5-1C, 5-S1C). The CTDΔ2 mutant was the only one that failed to produce adult flies (Fig. 5-1D, Fig. Table 5-S1). Strikingly, this region of the *Drosophila* CTD is the most highly conserved among higher eukaryotes (Fig. 5-1E, Fig. 5-S1A), suggesting functional necessity.

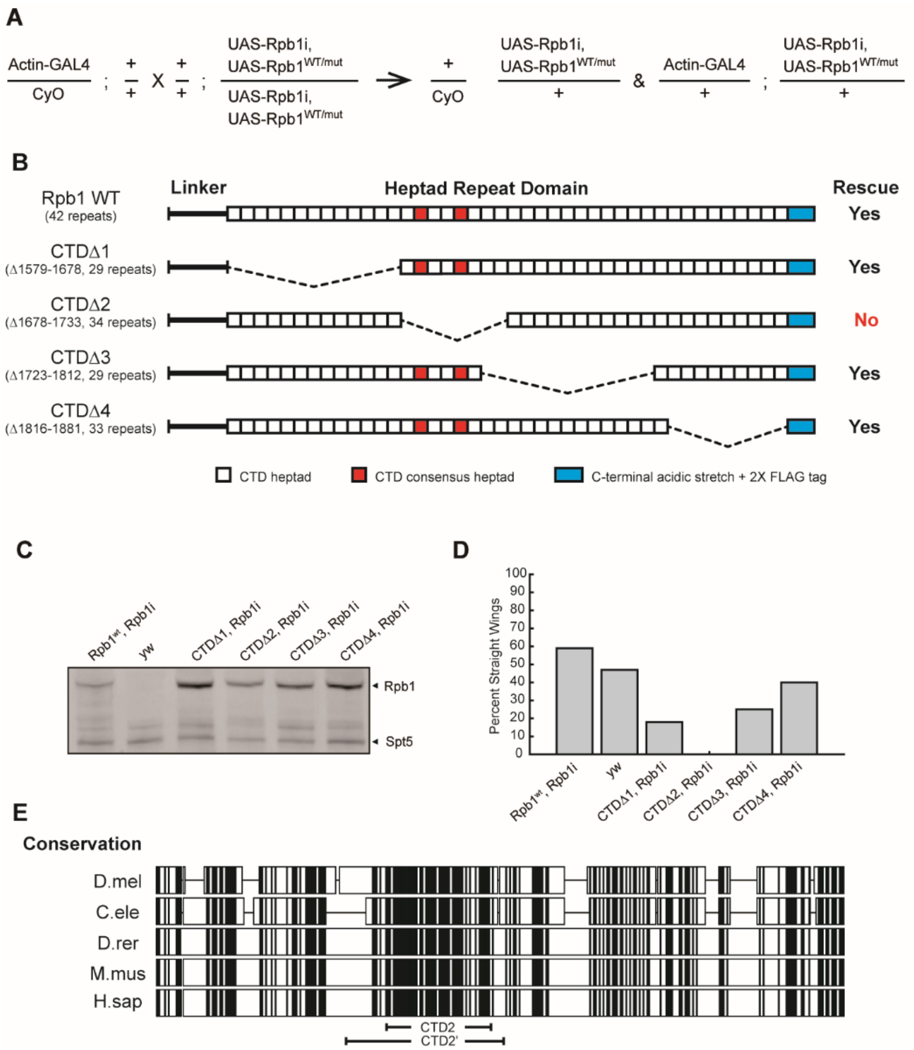


Figure 5-1. A highly conserved region in the *Drosophila* CTD is essential for viability and is targeted for Ser5 phosphorylation by DmP-TEFb *in vitro*. (A) Schematic of the rescue assay: Ubiquitous expression of shRNA targeting Rpb1 (UAS-Rpb1i) by the Actin-GAL4 transgene is lethal and results in no straight-winged progeny. Rescue as a result of co-expression of an Rpb1 derivative (UAS-Rpb1$^{WT/mut}$) is indicated by the presence of straight-winged adults among the progeny. (B) Four internal deletion mutants of Rpb1 were expressed under Actin-GAL4 control to test for rescue of lethality caused by ubiquitous depletion of endogenous Rpb1 (*Actin-GAL4/+; UAS-Rpb1i/+*). Only CTDΔ2 failed to rescue. (D) Western blot analysis of the expression of Rpb1

derivatives: Tissues were collected from late pupae derived from the same genetic cross as described in (A). Late pupae were analyzed because these were produced by each of the matings, including the one with CTDΔ2. Ectopic Rpb1 expression was detected with FLAG antibody. Detection of Spt5 served as a loading control. (E) Evolutionary conservation of the CTD across different species. Identical amino acids are denoted by black bars. The CTDΔ2 deletion and recombinant protein CTD2′ (Black lines, bottom) contain a highly conserved region.

Hyper-Phosphorylation of Recombinant CTD2′ by P-TEFb

Having identified a region of the *Drosophila* CTD that is essential for normal development, we next sought to incorporate the region removed in the CTDΔ2 mutant into a recombinant construct displaying the properties necessary for high resolution structural biology. We selected a CTD construct, CTD2′, containing residues 1657-1739 of the Rpb1 polypeptide sequence (Fig. 5-2A). In order to explore how phosphorylation impacts the structure of the CTD, we used *D. melanogaster* Positive Transcription Elongation Factor b (DmP-TEFb) to hyper-phosphorylate CTD2′ *in vitro* [11]. Analysis of hyper-phosphorylated CTD2′ by MS revealed successful incorporation of up to 12 phosphates per polypeptide (Fig. 5-2C). Phospho-site identification by tandem mass-spectrometry (MS/MS) led to the conclusion that our *in vitro* phosphorylation protocol predominantly generates high levels of Ser5 phosphorylation (Fig. 5-2A,B). Preferential phosphorylation *in vitro* of Ser5 over other amino acids in the CTD by human P-TEFb has been previously observed [12,13].

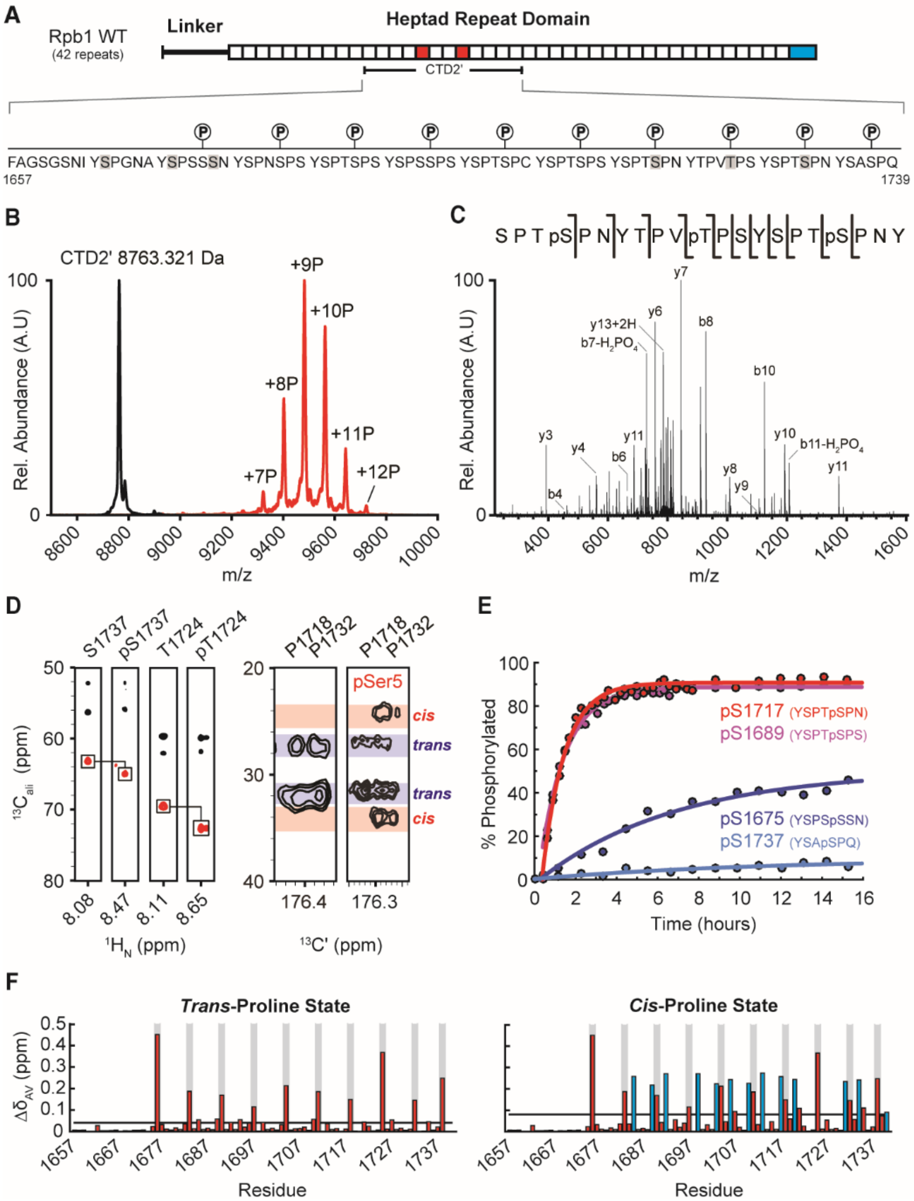


Figure 5-2. ΔCTD2′ Ser5 phosphorylation probed by MS and NMR spectroscopy**.** (A) Amino acid sequence of CTD2′ displaying 99% confidence phospho-site assignments by MS/MS (gray boxes) and by NMR (P-symbols). MS/MS peptide coverage was complete. (B) Linear positive MALDI TOF MS of unphosphorylated CTD2′ (black) and hyper-pSer5 CTD2′ (red) (C) Representative spectrum from Nano-LC MS/MS analysis of hyper-pSer5 CTD2′. (D) Representative strips from 3D HNCACB spectra of unphosphorylated and hyper-pSer5 CTD2′ showing perturbation upon phosphorylation (left) and strips from 3D CCCON spectra of unphosphorylated and hyper-pSer5

CTD2′ showing pSer5-induced *trans*-to-*cis* isomerization of Pro6 (P1718 and P1732; right). (E) Representative kinetic traces for CTD2′ phosphorylation monitored by RT-NMR. (F) Average chemical shift perturbations for CTD2′ upon phosphorylation for the *trans*-proline enriched (red) and *cis*-proline enriched states (blue). Gray bars indicate pSer/pThr5 residues and the black line denotes the average perturbation.

NMR spectroscopy was employed to augment and cross-check our MS-based phospho-site assignment, leading to the pattern depicted in Fig. 5-2A. NMR is well suited to this task due to its high sensitivity to the local chemical environments of individual residues, but backbone resonance assignment of disordered and repeat-containing polypeptides like the CTD is often complicated by extreme signal overlap [14-16] (Fig. 5-3A). To circumvent this limitation, we and others have developed ^{13}C Direct-Detect NMR spectroscopy [17], which vastly improves resonance dispersion for IDRs (Fig. 5-3B) and also provides direct measurement of proline resonances (Fig. 5-3D) [18,19]. This allowed unambiguous backbone resonance assignment of CTD2′, including assignment of the proline residues, which comprise 23% of the polypeptide sequence. Assignments were mapped onto ^{1}H,^{15}N correlation spectra through standard triple resonance experiments. Certain amide proton and side-chain carbon resonances (Fig. 5-2D,F & 5-3C) experienced downfield chemical shifts that were consistent with phosphorylation [20,21]. Correlated with these shifts, several proline residues adjacent to phosphorylation sites showed carbon chemical shift changes characteristic of *trans*-to-*cis* isomerization of the peptide plane (Fig. 5-2D,F & 5-3D). In total, 10 *bona-fide* phospho-sites were identified. Real-time NMR (RT-NMR) permitted the kinetic measurement of CTD2′ phosphorylation, revealing that for the seven internal heptads containing Ser5-Pro6 pairs, phosphorylation proceeded at similar apparent rates and reached comparable levels upon saturation (90%) (Fig. 5-2E, 5-S2, Table 5-S2) , with incomplete phosphorylation of the terminal repeats and of Thr5 in YTPVTPS. The observed rates are consistent with a distributive mechanism similar to that observed for the human P-TEFb [12]. Thus, the *in vitro* phosphorylation reactions produce a nearly complete hyper-pSer5 state, in agreement with our MS data (Fig. 5-2A-C,F).

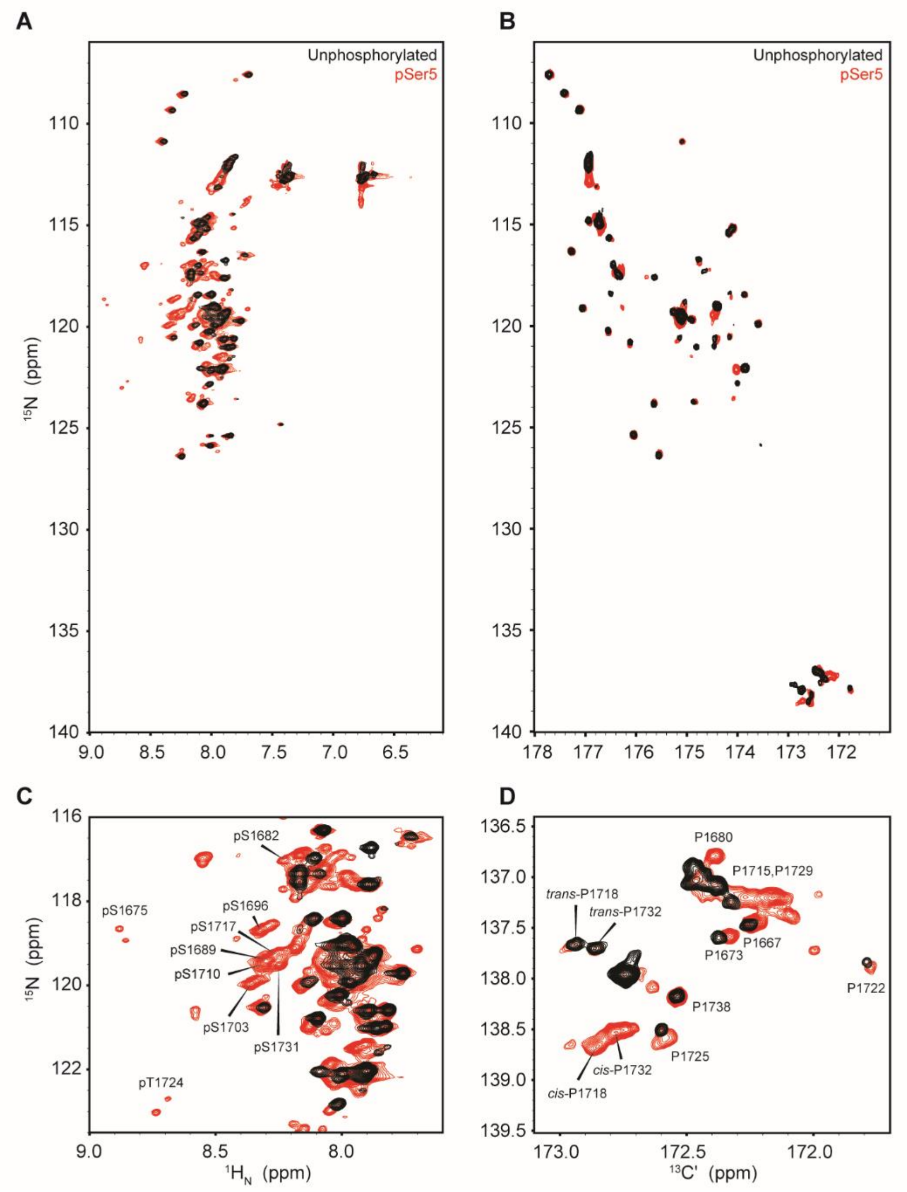


Figure 5-3. Phospho-sites and proline isomerization in hyper-pSer5 CTD2′ probed by NMR spectroscopy. (A) 2D ^{1}H-^{15}N correlation spectra of unphosphorylated (black) and hyper-pSer5 (red) CTD2′ (B) 2D ^{13}C′-^{15}N correlation spectra of unphosphorylated (black) and hyper-pSer5 (red) CTD2′ (C) Annotation of pSer5/pThr5 resonances in the downfield region of the 2D ^{1}H-^{15}N correlation spectrum (D) Proline region from the 2D ^{13}C′-^{15}N correlation spectrum of unphosphorylated (black) and hyper-pSer5 (red) CTD2′ with some proline resonances annotated.

Hyper-Phosphorylation does not Alter the Global Dimensions of CTD2′

We next turned to a detailed investigation of the effects hyper-phosphorylation has on the structure of the CTD2′ region. In order to test how compact CTD2′ is in solution, which may impact its accessibility to CTD binding factors, we collected small angle X-ray scattering (SAXS) data on the unphosphorylated and hyper-pSer5 states (Fig. 5-4, Table 5-S3). In the unphosphorylated state, CTD2′ displayed an average R_g of 28.0 ± 0.7 Å, while hyper-pSer5 CTD2′ displayed a similar average R_g of 28.3 ± 0.3 Å using the Guinier approximation (Fig. 5-S3). For comparison, the ubiquitous CTD interaction domain (~140 amino acids) has an R_g ~17 Å; in contrast, the nucleosome core particle (~800 amino acids and 146 DNA base-pairs) has an R_g ~41 Å, which demonstrates that CTD2′ is relatively expanded in solution. Similarly, pair-wise distance distributions revealed no significant increase in the maximum dimension (D_{max}) upon phosphorylation (Fig. 5-4B). These results are in good agreement with the dimensions predicted for an excluded volume random coil with the same number of monomers as CTD2′ [22]. Independently, ^{31}P NMR spectroscopy revealed that the phosphates in the CTD were in the -2 charge state under our experimental conditions (Fig. 5-S4), and yet charge-charge repulsions did not appear to impact the dimensions of the CTD. Incorporation of this data into a model for random-coil structure demonstrates that the median pSer5-pSer5 distance (approximately 18Å) is likely to be greater than the Debye screening length under our experimental conditions (Fig. 5-S5), which accounts for the lack of chain expansion upon hyper-phosphorylation. In summary, our SAXS data demonstrate that the region of the CTD encompassed by CTD2′ experiences no significant phosphorylation-induced change in structure on the nanometer length scale. This suggests that the functional purpose of hyper-phosphorylation is not to effect a dramatic expansion of the CTD in order to modulate access by interacting factors, but rather to change the pattern of binding motifs displayed on an already expanded structure.

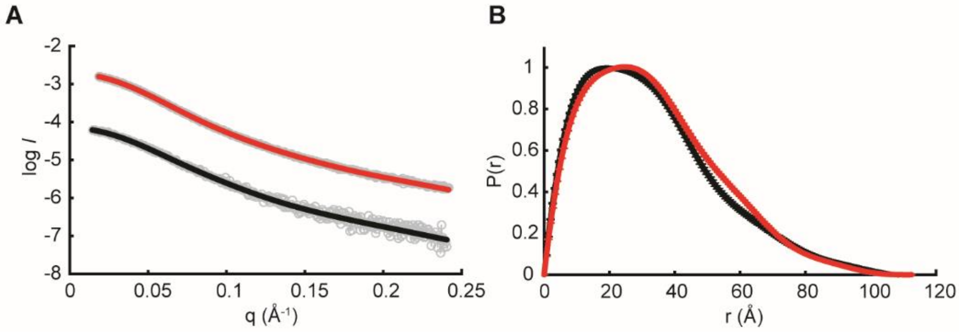


Figure 5-4. Small angle X-ray scattering reveals no significant change in pair-wise distances within CTD2′ upon extensive serine 5 phosphorylation. (A) Raw scattering data for unphosphorylated CTD2′ (gray circles, bottom) and hyper-pSer5 CTD2′ (gray circles, top). Fits for unphosphorylated CTD2′ and hyper-pSer5 CTD2′ are shown superimposed on the raw data (solid black and red lines, respectively). (B) Representative pair-wise distance distributions for unphosphorylated CTD2′ (black) and hyper-pSer5 CTD2′ (red) calculated using the autoGNOM function in Primus qt.

Hyper-Phosphorylation of CTD2′ Induces Sequence-Dependent Proline Isomerization

On the local scale of several to tens of amino acids, which is the size of the motifs most CTD-interacting domains recognize, phosphorylation has been shown to strongly perturb backbone dihedral angles [23-25], suggesting that phosphorylation could induce local structural perturbations in the CTD. Therefore, we used NMR to probe the backbone conformation of CTD2′. Secondary structure populations for the unphosphorylated state of CTD2′ were calculated from chemical shifts using the secondary structure calculation program δ2D (Fig. 5-S6A), revealing small populations of extended β/PPII character and a strong propensity for random coil like conformations, consistent with our SAXS results and previous solution studies of CTD peptides [15]. Further, proline C_β and C_γ chemical shifts demonstrated a strong preference for the *trans*-proline conformation (~95% *trans*-proline isomer) (Fig. 5-5A, 5-S6B). Thus, in the unphosphorylated state CTD2′ is highly disordered and contains nearly all *trans*-proline.

Surprisingly, in the hyper-pSer5 state, CTD2′ displayed a 2-fold enrichment of *cis*-proline content averaged over all 19 proline residues (Fig. 5-5B, 5-S6C). Further, where isomerization occurred, peak splitting into two sets of NMR resonances was observed for Thr4, Ser7, Cys7 and Asn7 residues and associated large chemical shift perturbations (blue bars, Fig. 5-2F). The presence of two sets of assignable resonances suggests a chemical exchange process on the millisecond time scale or slower. To confirm exchange between *cis*- and *trans*-proline isomers, we collected ^{15}N ZZ-exchange, which permits the quantitative observation of conformational exchange on the ms-s timescale, in the presence of the *Drosophila* prolyl isomerase Dodo [26] (Fig 5-S7). Exchange peaks were observed for all Pro6 residues adjacent to pSer5, but not for Pro3, consistent with Dodo specificity for the pSer5-Pro6 pair [27]. Interestingly, no exchange could be observed for the pThr5-Pro6 pair in the YTPVpTPS sequence context, suggesting that this heptad is essentially *trans*-locked on the 100 ms timescale, even in the presence of a prolyl isomerase. Thus, in the hyper-pSer5 state, CTD2′ experiences slow exchange between *trans*- and *cis*-proline isomers. In this exchange regime, peak intensities correspond to the populations of *cis*- and *trans*-proline species, which allowed us to estimate the magnitude of *cis*-proline within each heptad (Fig. 5-6A). Within repeats of YSPTpSPS and the similar cysteine-containing repeat of YSPTpSPC, *cis*-Pro6 content was enriched 3-fold (to ~15%) by Ser5 phosphorylation. Further, Pro3 within all heptads containing pSer5 showed a modest enrichment of *cis*-proline content by ~5%, suggesting some non-local effects. Strikingly, Pro6 within heptads containing Asn7 (YSPTpSPN) showed a 6-fold (~35%) enrichment in *cis*-proline. Thus, the proline *trans*-to-*cis* switch is sequence context-dependent and modulated by both phosphorylation and deviations from the consensus heptad sequence.

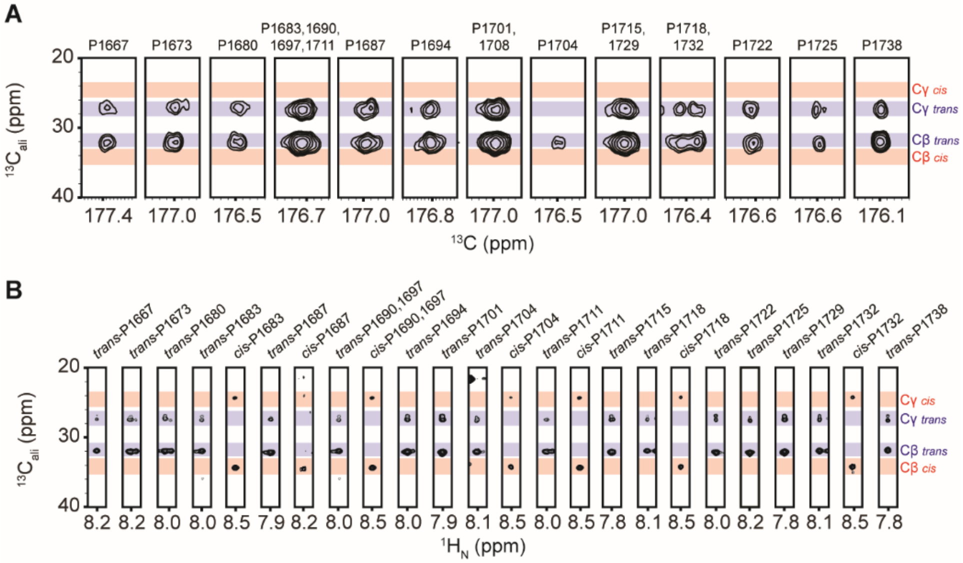


Figure 5-5. Structural characterization of unphosphorylated CTD2′ by NMR spectroscopy. (A) Cβ and Cγ chemical shifts from the 3D CCCON spectrum of unphosphorylated CTD2′ demonstrate that when resolved, individual proline side chain resonances show a nearly all trans state. Blue and red bars represent the range of chemical shifts (mean ± S.D.) for prolines in the trans and cis conformation, respectively. (B) Cβ and Cγ chemical shifts from the 3D CCCONH spectrum of hyper-pSer5 CTD2′ reveal dramatic trans to cis conformational switches in response to pSer5.

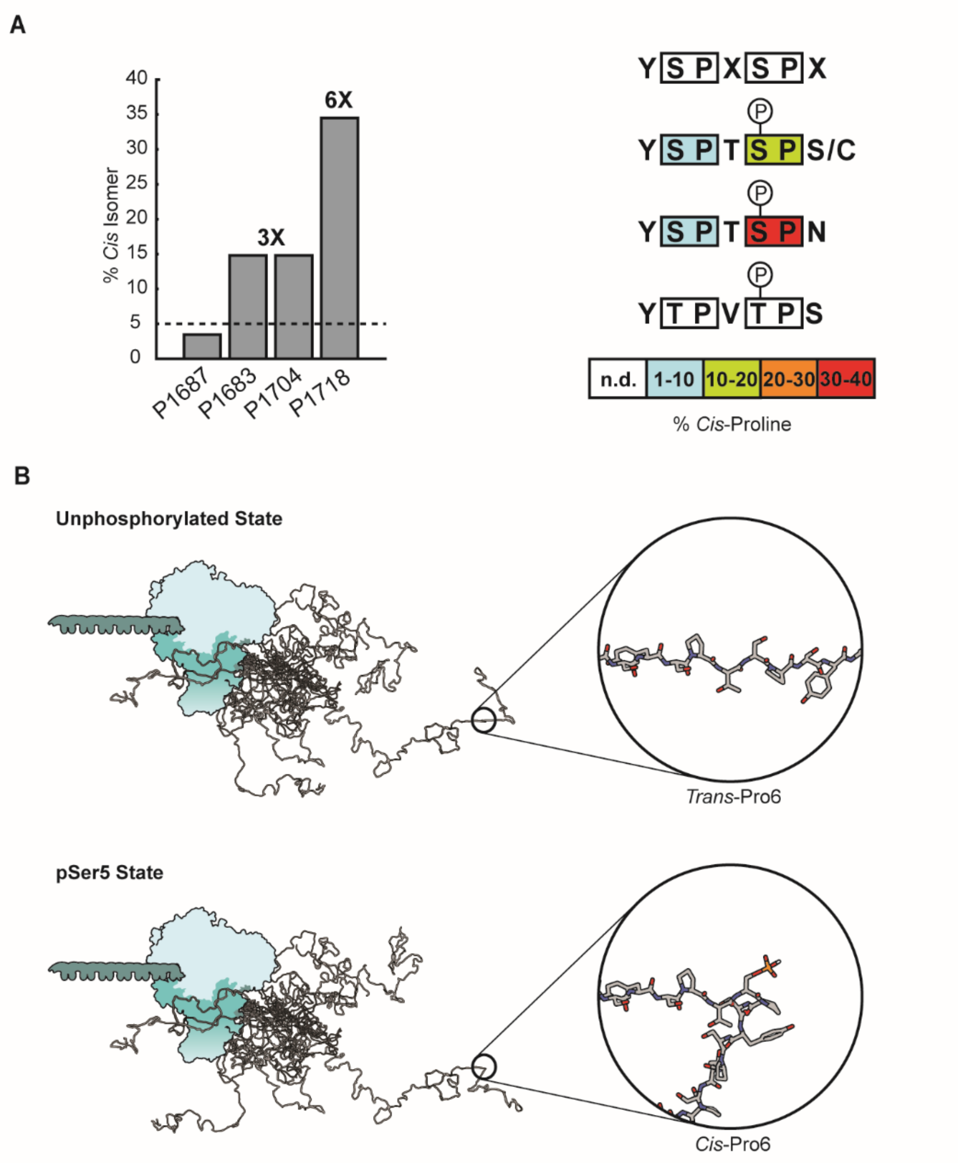


Figure 5-6. The impact of serine 5 phosphorylation on the structure of the DmCTD**. (A)** Percentage of *cis*-proline for several proline residues determined from peak intensities in 2D NMR correlation spectra of hyper-pSer5 CTD2′, where the dotted line denotes the average percentage of *cis*-proline in the unphosphorylated state (left). This is depicted schematically for various heptad sequences in CTD2′ (right) (B) Model for the effect of Ser5 phosphorylation on the structure of the CTD. In the unphosphorylated state the CTD exists in an ensemble of conformational states that favor prolines in the *trans* conformation (top) Hyper-pSer5

Sequence-Dependent Proline Isomerization Modulates Apparent Ssu72 Activity

The observation that deviations from the consensus heptad sequence modulate the extent to which Pro6 *cis-trans* equilibria are affected by pSer5 suggests that in response to uniform phosphorylation patterns, non-consensus heptads may impart an additional layer of specificity for CTD interacting factors. The CTD phosphatase Ssu72 has been shown to exhibit activity toward heptads containing the pSer5-*cis*Pro6 dipeptide pair [28]. Thus, we hypothesized that non-consensus heptads within the hyper-pSer5 CTD modulate the apparent Ssu72 activity through pSer5 induced Pro6 isomerization. To test this possibility, we followed the dephosphorylation of hyper-pSer5 CTD2′ by *D. melanogaster* Ssu72-symplekin using RT-NMR spectroscopy (Fig. 5-7). Loss in NMR peak intensities relative to the zero-time point were observed for all heptads containing pSer5-Pro6 dipeptide pairs (Fig. 5-7A, 5-S8). Interestingly, in each heptad sequence context, different apparent rates of dephosphorylation were observed (Fig. 5-7B, Table 5-S4). For pSer5 residues within the region flanked by the two consensus heptads, similar apparent rates were observed, suggesting that small deviations from the consensus motif (YSPTpSPC or YSPSpSPS) do not dramatically alter Ssu72 activity. However, minimal Ssu72 activity was observed for pS1682 (YSPNpSPS) and no dephosphorylation could be observed for pS1675 (YSPSpSSN), consistent with the requirements of Thr4 and Pro6 for Ssu72 activity [28,29]. Strikingly, Ssu72 exhibited nearly 3-fold apparent activity enhancement toward pSer5 residues within the Asn7 heptads, relative to the consensus motifs, strongly suggesting that the higher propensity for *cis*-Pro6 within these motifs increases the apparent Ssu72 activity.

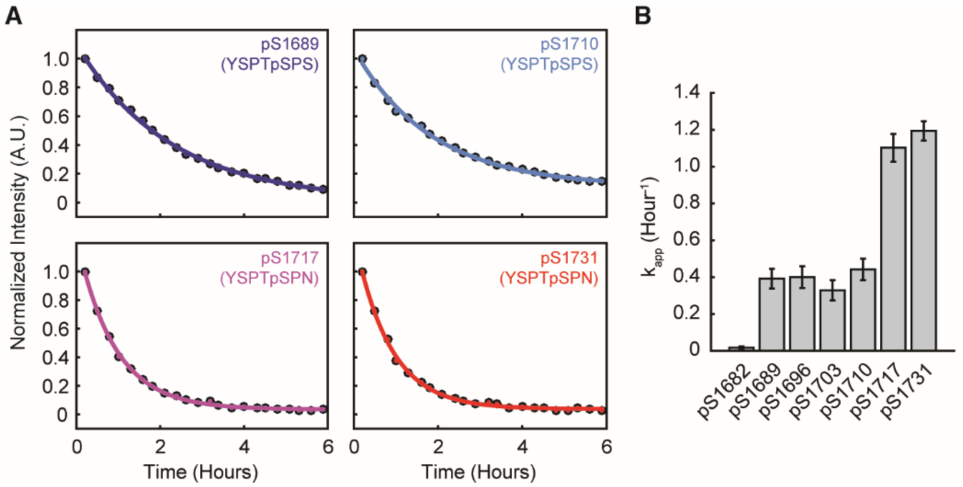


Figure 5-7. Structural switches in hyper-pSer5 CTD2′ modulate the apparent Ssu72 activity. (A) Representative kinetic traces of Ssu72 dephosphorylation of pSer5 in CTD2′ monitored by RT-NMR. (B) Apparent rate constants forpSer5 dephosphorylation reveal heptad specific apparent Ssu72 activities. The highest apparent Ssu72 activities are observed for pSer5 residues within heptads containing Asn7.

To understand how residues flanking the pSer5-Pro6 pair affect pSer5 dephosphorylation by Ssu72, we analyzed the conformation of phosphoryl CTD peptides upon Ssu72 binding. Several structures have been published for *Drosophila* or human Ssu72 bound to CTD peptides of different phosphorylation states including *Drosophila* Ssu72, *Drosophila* Ssu72-Symplekin, and human Ssu72-Symplekin, bound to pSer5 CTD (PDB codes: 3P9Y [16], 4IMJ [30], and 3O2Q [31], respectively), and *Drosophila* Ssu72-Symplekin bound to a CTD peptide with Thr4/Ser5 doubly phosphorylated (PDB code: 4IMI [30]). The superimposition of these structures reveals that all known phosphoryl CTD peptides adopt a tight turn facilitated by *cis*-proline upon binding to Ssu72-Symplekin (Fig. 5-8A). This tight turn is stabilized by three intra-molecular hydrogen bonds formed by the hydroxyl side chain of Thr4 with the main chain carboxylate of Ser7 (2.8 Å) and the amide group of Tyr1 in the following repeat (3.3 Å), as well as the main chain carboxylate of Thr4 and Pro6 (3.2 Å). Due to the high conservation of the tight turn configuration, the identity of the residues flanking Pro6 can be altered for effective Ssu72 recognition as long as the intra-molecular hydrogen bond network is maintained. For example, our NMR results demonstrate that Ssu72 dephosphorylates YSPSpSPS or YSPTpSPC with similar efficacy as it does consensus heptads; molecular modeling suggests that in all three of these heptad motifs, the intra-molecular hydrogen bond network can be conserved even as the sequence varies (Fig. 5-8B). On the other hand, little to no Ssu72 activity was observed upon the replacement of Thr4 by Asn. We attribute this loss of activity to the need for Asn4 to adopt an alternative rotameric state to avoid steric clashes, which is incompatible with forming two of the hydrogen bonds that stabilize the Ssu72 recognition conformation (Fig. 8C). We have shown previously that the phosphorylation of Thr4, which also disrupts these two hydrogen bonds, reduces Ssu72 activity four-fold [30].

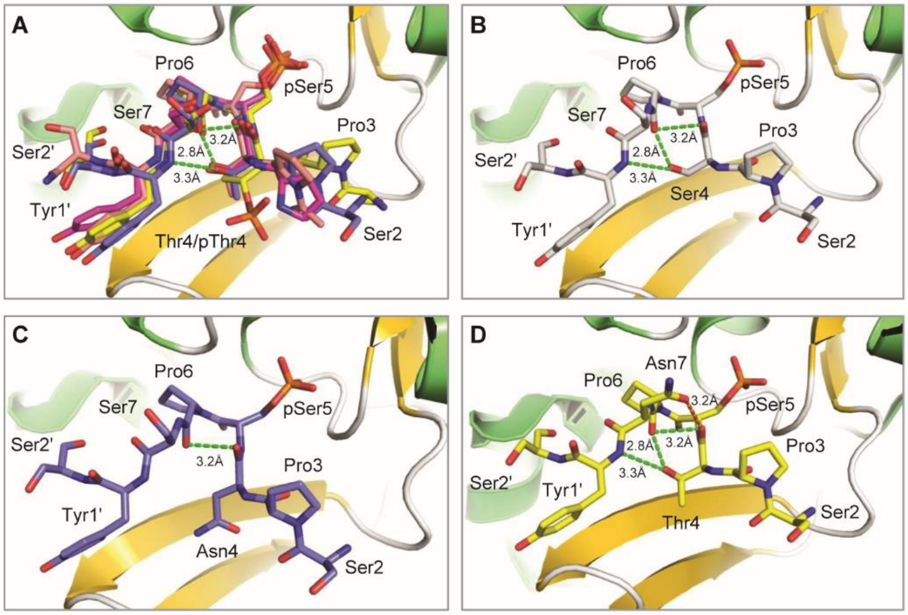


Figure 5-8. The conserved conformation of CTD peptides recognized by Ssu72. (A) Ssu72 is shown as a ribbon diagram with α-helices in green and β-strands in gold. CTD peptides are shown as colored sticks with carbon atoms shown in different colors: PDB code 4IMI (yellow), 4IMJ (blue), 3P9Y (salmon) and 3O2Q (magenta). The intra-molecular hydrogen bonds are shown in green dash lines. The CTD residues are numbered based on consensus sequence and the following repeat residues are labeled with a prime. (B) The intra-molecular hydrogen bond network can be maintained even when Thr4 is replaced by Ser. (C) The replacement of Thr4 by Asn loses two intra-molecular hydrogen bonds. (D) An additional intra-molecular hydrogen bond can be formed (orange dashed line) when Ser7 is replaced by Asn.

In the context of the present study, the most significant observation of our NMR analysis of Ssu72-catalyzed CTD2′ dephosphorylation is that Ssu72 shows its greatest activity toward pSer5-CTD heptads containing Asn7, which we have also shown are the heptads most highly enriched in *cis*-proline among those observed in this region of the CTD. For this heptad, molecular modeling suggests that the side chain of Asn7 could form an additional intra-molecular hydrogen bond with the carboxylate group of Thr4 (Fig. 5-8D). With all possible isomeric states of the Asn sidechain, the most favorable configuration is within 3.2 Å away from the backbone carboxylate of Thr4, which further strengthens the tight turn conformation needed for Ssu72 recognition and makes the *cis*-Pro6 more energetically favorable. Taken together, this data strongly suggests that in the presence of uniform phosphorylation patterns, non-consensus CTD heptads encode cryptic structural switches to fine tune the specificity of CTD interacting factors.

Conclusion

Based on these results, we propose a model in which hyper-pSer5 does little to alter the global-scale structure of the CTD2′ region of CTD. Instead, dramatic structural rearrangements occur on the single heptad scale, driven by sequence context-dependent proline *trans*-to-*cis* isomerizations (Fig. 5-6B). This discovery predicts multiple potential mechanistic outcomes in the context of the CTD code.

A general conclusion from these observations is that these features allow the CTD to transduce homogenous post-translational modifications into structurally and functionally diverse responses. The diversity of CTD sequences across eukaryotes has been recently acknowledged [3], though the functional significance of many non-consensus heptads has not been widely investigated. Non-consensus CTD repeats may expand the repertoire of available post-translational modifications, thus increasing the complexity of signaling through the CTD [4,32-35]. However, the prevalence of non-consensus motifs in the CTDs of many eukaryotes, *e.g. D. melanogaster*, suggests that they must also maintain the ability to support general transcription, though the specificity of CTD kinases and phosphatases toward the myriad non-consensus motifs present in these systems have not been widely explored. In our minimal system, we observed that non-consensus motifs that contained small deviations from the consensus responded similarly to DmP-TEFb and Ssu72. While, other more cryptic variants like YSPNSPS produced drastically different outcomes depending on the modifying enzyme present, and severe degradation of the heptad sequence rendered some repeats resistant to modification by both enzymes. This suggests that the relationship between consensus-conservation and functional specialization may lie on a continuum. In this view, small sequence deviations may be tolerated by the majority of regulating enzymes, imparting only modest differences in modification kinetics and patterning. By contrast, large fluctuations in sequence space may only support a sub-set of regulatory factors. While the unique functions these and other non-consensus motifs serve during transcription will need to be determined empirically, these observations provide support for the emerging hypothesis that variation in CTD sequence enables differential gene regulation in the context of normal development, or at the level of individual genes [32,33]. In this context, our model leads to the prediction that conserved heptads maintain the ability to attract a wide range of factors involved in basic cellular processes, while non-consensus heptads enhance spatial control of interacting factor recruitment, thus creating more tightly regulated transcriptional programs in higher eukaryotes.

Our observation that non-consensus heptads within the DmCTD containing Asn7 show a preponderance for *cis*-pro6 in response to pSer5 is striking. These heptads are conserved from yeast to human and tend to cluster in a region just beyond the consensus repeats [7]. Our results suggest that these regions would populate a high degree of *cis*-proline when pSer5 marks are prevalent, as in early phases of the transcription cycle. This could restrict the binding of many known pSer5-*trans*-Pro6 CTD interacting factors or alternatively, favor the association of pSer5-*cis*-Pro6 specific factors. Ssu72 is a pSer5-*cis*-Pro6 specific phosphatase known to promote the transition from hyper-pSer5 to hyper-pSer2 during transcript elongation [28]. And in line with this model, our results demonstrate that Asn7 heptads can increase the apparent Ssu72 activity. Recent work has revealed that most CTD heptads are phosphorylated only once at any given time [9,10], suggesting ordered erasing of the pSer5 mark prior to writing of the pSer2 mark. Our results suggest that one role of Asn7 repeats may be to pre-prime this particular region of the CTD for pSer5 dephosphorylation by Ssu72, thus assisting in the dynamic spatiotemporal control of the pSer5-to-pSer2 transition. Thus, the strategy employed here highlights the general feasibility of applying quantitative biophysical techniques to obtain mechanistic insights into the CTD code.

Materials & Methods

Fly procedures

Sequences encoding RNAi resistant Rpb1^{WT} or Rpb1 derivatives with a double FLAG-tag at the C-terminus were subcloned into the pUASt-attB vector, followed by transformation into the attP site on chromosome 3 in the *PhiC31 attP 86Fb* fly line [36]. *UAS-Rpb1i* and *yw*; *Actin-GAL4/CyO,* were obtained from the Bloomington Stock Center (lines 36830 and 4414, respectively). Rpb1i-resistance of the ectopically expressed Rpb1 variants was achieved by changing the part of the coding sequence of Rpb1 that corresponds to the 21nt RNAi recognition sequence (sense strand: AACGGTGAAACTGTCGAACAA) to AAC*C*GT*C*AA*GTTGAGC*AACAA. The *UAS-Rpb1i, UAS-Rpb1* lines were generated by routine matings and meiotic recombination. The lethality test was done by mating virgin female *yw*; *Actin-GAL4/CyO* to male *yw*; *UAS-Rpb1i, UAS-Rpb1*$^{WT/mut}$. Animals were raised at 21°C. Rescue was confirmed by the emergence of straight-winged adults among the progeny (*Actin-GAL4/+* ; *UAS-Rpb1i, UAS-Rpb1*$^{WT/mut}$/+).

Western blot analysis for ectopic expression of Rpb1 was done by dissecting late pupae from the pupal case and then homogenizing and boiling the tissue in LDS sample buffer (Invitrogen). Equal numbers of male and female late pupae were selected and pupae of the genotype *yw*; *Actin-GAL4/+*; *UAS-Rpb1i, UAS-Rpb1*$^{WT/mut}$/+ were distinguished from the *yw*; *CyO/+*; *UAS-Rpb1i, UAS-Rpb1*$^{WT/mut}$/+ counterpart by the intensity of red pigment in the eye (the *UAS-Rpb1* and *Actin-GAL4* transgenes have white gene markers). For western blotting, tissues equivalent to 0.3 pupae were loaded into each lane on a 3-8% Tris-acetate SDS-PAGE gel (Life Technologies). Transgenic Rpb1 expression was detected with rabbit anti-Flag antibody (1:3000; Genscript). Spt5 was detected with rabbit anti-Spt5 (1:3000). The blot was subsequently probed with goat anti-rabbit IgG (1:3000; Alexa Fluor 488) and visualized with a Typhoon (GE Healthcare).

Protein Expression and Purification

CTD2´: A synthetic gene for the *Drosophila melanogaster* RPB1 Carboxyl-terminal domain was purchased from GeneArt (Thermo Fisher Scientific). A region corresponding to residues (1657-1739) was amplified by PCR, cut with XhoI (NEB) and XmaI (NEB), and ligated into the pET49b+ expression vector (Novagen) using T4 DNA ligase (NEB) to produce a construct containing GST & His tags. Protein expression was performed in *E. coli* BL21 DE3 cells. 500 ml batch cultures were grown to an optical density of 0.8 at 37 ^{o}C at which point, cells were induced using 0.5 mM IPTG and allowed to incubate at 37 ^{o}C for 3 hours. Following lysis by sonication on ice in lysis buffer (50 mM Tris/HCl pH 7.5, 500 mM NaCl, 20 mM Imidaozole, 2.5 mM β-mercaptoethanol, 10X EDTA-free protease inhibitor cocktail (Calbiochem), and 10 units of RNAse free DNAse (NEB)), samples were centrifuged at 4 ^{o}C for 40 minutes at 11,500 x g. Cleared supernatant was passed over HisPur Ni^{2+}-NTA resin (Thermo Fisher Scientific) and bound protein was washed of contaminants using 5 column volumes of wash buffer (50 mM Tris/HCl pH 7.5, 500 mM NaCl, 20 mM imidaozole, 0.1% Triton-1000, 2.5 mM β-mercaptoethanol). Protein was eluted using elution buffer (50 mM Tris/HCl pH 7.5, 500 mM NaCl, 200 mM imidaozole, 2.5 mM β-mercaptoethanol). The GST and 6XHis tags were removed by adding recombinant His-tagged HRV 3C protease to the protein (resulting in an N-terminal non-native GPG) and dialyzing the mixture overnight against 50 mM Tris/HCl pH 7.5, 300 mM NaCl, 2.5 mM β-mercaptoethanol at 4 ^{o}C. The protein was then passed over the Ni^{2+}-NTA column to remove the protease and non-specifically bound contaminants. A final purification was then performed by size exclusion chromatography in 80 mM Imidazole pH 6.5, 50 mM KCl, 2.5 mM β-mercaptoethanol using P-10 resin (BioRad).

DmP-TEFb: Sf9 cells were grown in suspension at 27 ^{o}C to 1.5 million cells/ml and infected with 1/10 culture volume *D. mel* P-TEFb virus (generous gift from J.T. Lis). Infection was carried out at 27 ^{o}C at a shaker speed of 75 rpm for 72 hours. Following lysis in 50 mM HEPES pH 7.5, 500 mM NaCl, 10% glycerol, 1% Nonidet P-40, 2.5 mM imidizole and 2.5mM β-mercaptoethanol and protease inhibitors, by dounce homogenization, lysates were centrifuged at 4°C for 30 minutes at 100,000 x g. Cleared supernatant was passed over TALON resin (Clontech) and bound protein was washed using 5 column volumes of 50 mM HEPES pH 7.5 500 mM NaCl 10% glycerol 1% Nonidet P-40 10 mM imidizole 2.5 mM β-mercaptoethanol. DmP-TEFb was eluted with 50 mM HEPES pH 7.5 500 mM NaCl 10 % glycerol 1% Nonidet P-40, 200 mM imidazole, 2.5 mM β-mercaptoethanol, and flash frozen. Kinase activity toward CTD2′ was confirmed by auto-radiography.

DmDodo: A synthetic gene corresponding to residues (1-166) of *D. mel* Dodo was purchased from GeneArt (Thermo Fisher Scientific). The gene was cut with XhoI (NEB) and XmaI (NEB) and ligated into the pET47b+ expression vector (Novagen) using T4 DNA ligase (NEB) to produce a His-tagged construct. Protein expression was performed in *E. coli* BL21 DE3 cells. 500 ml batch cultures were grown to an optical density of 0.8 at 37 ^{o}C at which point, cells were induced using 0.5 mM IPTG and allowed to incubate at 37 ^{o}C for 3 hours. Following lysis by sonication, samples were centrifuged at 4 ^{o}C for 40 minutes at 11,500 x g and purified by affinity chromatography using HisPur Ni^{2+}-NTA resin (Thermo Fisher Scientific) as previously described. Following removal of the His-tag using His-tagged HRV 3C protease, a final purification was performed by gel filtration in 50 mM Tris/HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 2.5 mM β-mercaptoethanol using a sephacryl S-100 Hi-prep 16/60 size exclusion column on an Äkta FPLC (GE).

Kinase Reactions

Kinase reactions were generally carried out in 50 mM Tris/HCl pH 7.5, 50 mM NaCl, 10 mM $MgCl_2$, 2 mM DTT, 12 mM ATP with 80 µg/ml DmP-TEFb and 100 µM CTD2′. Reactions were carried out at 24°C and allowed to proceed to completion (~16 hours), after which 20 mM EDTA was added to ensure reaction termination.

Mass Spectrometry

MALDI-TOF-MS; MALDI TOF mass spectra of intact unphosphoryalated CTD2′ and phospho CTD2′ were acquired on an Ultraflextreme instrument (Bruker, Billerica, MA) in linear positive detection mode using factory-configured instrument parameters for 5,000 -20,000 *m/z* range. The instrument was calibrated using a protein mixture containing bovine insulin, MW 5733.5, bovine ubiquitin, MW 8564.8, bovine RNAse A, MW 13682.2, equine heart cytochrome C, MW 12359.9, and equine heart myoglobin, MW 16951.3 (all from Sigma); a 20 mg/mL solution of super-DHB (2,5-Dihydroxybenzoic acid and 2-hydroxy-5-methoxybenzoic acid, Sigma) in 50% aqueous acetonitrile containing 0.1 % o-Phosphoric acid (EMD Millipore) and 0.1 % trifluoroacetic acid (Thermofisher) was used as the matrix for both the calibrants and the CTD samples. The CTD samples for the

MALDI TOF MS were prepared by mixing 1 µL of 100 µM protein solution in water and 1 µL of the matrix solution and applying 1 µL of this mixture to a polished stainless steel target. Mass spectra were acquired by summing 2000-3000 shots at a 1000 Hz laser repetition rate, average calibration error was 521 ppm.

Chymotrypsin digest; A 1 µg/µL stock solution of sequencing-grade chymotrypsin (Thermofisher) prepared in 1 mM hydrochloric acid was diluted 50-fold with chilled 50 mM aqueous solution of triethylammonium bicarbonate (TEAB) and added to the CTD solution in 1:1 vol:vol ratio. The proteolysis was allowed to proceed overnight at 37 ^{o}C. Samples were acidified with a 1% aqueous solution of formic acid (FA), dried down, and re-dissolved in 15 µL of 4% acetonitrile (ACN) containing 0.1% FA for the nano-LC MS2 analysis.

Nano-LC MS2 ; 3 µL of digested peptide solution was loaded onto an Acclaim PepMap100 trapping column (100 µm × 2 cm, C18, 5 µm, 100 Å, Thermo) at a flow rate of 20 µL/min using 4% aqueous acetonitrile (ACN), 0.1% formic acid (FA) as a mobile phase. The peptides were separated on an Acclaim PepMap RSLC column (75 µm × 15 cm, C18, 2 µm, 100 Å, Thermo) with a 90-min 4% - 60% linear gradient of aqueous acetonitrile containing 0.1% FA. The gradient was delivered by a Dionex Ultimate 3000 nano-LC system (Thermo) at 300 nL/min.

An LTQ Orbitrap Velos ETD mass spectrometer (Thermo) was set to acquire data using the following data-dependent parameters. A full FT MS scan at R 60,000 over 350 – 2000 *m/z* range was followed by 5 FT MS2 scans with CID activation and 5 FT MS2 scans with ETD activation on most intense precursors at R 7,500. Only the precursors with charge states +2 and higher were selected for MS2 based on the FT master scan preview; the charge-state-dependent ETD time and monoisotopic precursor selection were enabled; the isolation window was 5 *m/z*, and the minimum precursor signal was set at 10000 counts for both CID and ETD. The ETD activation time was 100 ms. Polysiloxane ion, *m/z* 445.12003 was used as a lock mass.

Data analysis; The mass spectra were processed using Proteome Discoverer 1.3 (P.D. 1.3, Thermo). The expressed CTD sequence was appended to a database containing 45443 sequences (*Drosophila Melanogaster, Escherichia Coli*, common contaminants). The workflow was split into pipelines for CID and for ETD; and the following search parameters were used for both pipelines: enzyme chymotrypsin with 3 missed cleavages, precursor mass tolerance 30 ppm, fragment mass tolerance 0.8 Da, Met oxidation and Ser, Thr, Tyr phosphorylation as dynamic modifications, Cys carbamidomethylation as static modification. The CID ion series weights for b and y were set to 1; and the ETD ion series weights were 1 for c and z ions and 0.25 for b and y ions. The ETD pipeline included a non-fragment filter node which was set to remove precursor peak, charge-reduced precursor peaks, and peaks due to the neutral loss from charge-reduced precursors prior to the SEQUEST search. Resulting search files were exported into Scaffold PTM (Proteome Software) where phospho-site probabilities were determined using the Ascore algorithm [37].

NMR Spectroscopy

CTD2′ samples were expressed as described in M9 minimal media enriched with ^{15}N-NH$_4$Cl and/or ^{13}C-D-Glucose (Cambridge Isotope Laboratories). Following purification, samples were buffer exchanged in Amicon Ultra-15 3,000 NMWL centrifugal filters (Merck Millipore Ltd.). Typically, 80 mM imidazole pH 6.5, 50 mM KCl, 10% glycerol, 2 mM DTT and 10% D$_2$O was used for NMR experiments. For ^{15}N ZZ-exchange experiments, 20 mM MES pH 6.5, 50 mM KCl, 10% glycerol, 2 mM DTT and 10% D$_2$O was used. To obtain the desired pH range during ^{31}P experiments, 80 mM Citrate pH 4.0/5.0/5.5, 80 mM imidazole pH 6.2/6.5, and 80 mM Tris/HCl pH 7.2/8.3, containing 50 mM KCl, 10% glycerol, 2 mM DTT and 10% D$_2$O were used.

NMR Spectra were collected at the Lloyd Jackman NMR facility at the Pennsylvania State University on Bruker Avance-III spectrometers operating at proton frequencies of 500, 600 or 850 MHz equipped with TCI single axis gradient cryoprobes (^{1}H/^{13}C/^{15}N/^{2}H) with enhanced sensitivity for ^{1}H and ^{13}C. Phosphorous-detect experiments were performed on a 500 MHz Bruker Avance-III-HD spectrometer equipped with a broadband (BBO) Prodigy CryoProbe. Chemical shift assignments were made using ^{13}C-Direct Detect methods developed in-house [17,38,39], as well as standard ^{1}H-Detect triple resonance experiments. ^{13}C & ^{31}P chemical shifts were referenced to DSS and phosphoric acid standards, respectively. 2D Spectra were processed in Topspin 3.2 (Bruker) & NMRPipe and analyzed in Sparky. 1D spectra were processed in Topspin 3.2 (Bruker) and analyzed in Mnova (Mestrelab Research). Average chemical shift perturbations for ΔCTD2′ upon phosphorylation were calculated by:

$$\Delta\delta_{\mathrm{AV}} = [\frac{1}{3}\{(\Delta\delta_{\mathrm{HN}})^2 + (0.102\,\Delta\delta_{\mathrm{N}})^2 + (0.251\,\Delta\delta_{\mathrm{C'}})^2\}]^{1/2}$$

where $\Delta\delta_{HN}$, $\Delta\delta_{N}$, and $\Delta\delta_{C'}$ are the differences in ^{1}H, ^{15}N, and ^{13}C′ chemical shift between the unphosphorylated and phosphorylated species, respectively.

RT-NMR Kinetics and Data Processing

DmP-TEFb kinase reactions were performed in 50 mM HEPES pH 6.8, 50 mM KCl, 20 mM $MgCl_2$, 12 mM ATP, 2 mM DTT and 10% D_2O with 250 µM CTD2′ and ~100 µg/mL DmP-TEFb. Standard ^{1}H, ^{15}N-HSQC spectra were collected at 850 MHz with 1024(H) x 256(N) points, 4 scans and a recycle delay of 0.8s for total acquisition times of ~16 minutes for the first 8 hours, after which 16 scans were collected. To measure slow sites 16 scans were acquired for each experiment over the entire time course. As recalibration of the instrument was required following enzyme addition, the first data point was acquired in an effective dead-time of ~20 minutes. Phosphatase reactions were performed in 80 mM imidazole pH 6.5, 50 mM KCl, 2 mM DTT and 10% D_2O with 1.0 mM hyper-pSer5 CTD2′ and ~10 µg of *D. mel* Ssu72-Symplekin complex. Ssu72-Symplekin was prepared following previously published protocols[28]. For the phosphatase reactions, spectra were collected as described using 4 scans throughout the time course. Spectra were processed in Mnova (Mestrelab Research). Extracted peak intensities for pSer/pThr resonances and effected resonances of neighboring residues were plotted as a function of time and fit in MATLAB. Single exponential decays and build-up curves were fit as irreversible first-order reactions and intermediate species were fit as two consecutive irreversible first-order reactions by the method of non-linear least squares using:

$$y = y_0 + S_0 e^{-kapp\ t}$$

$$y = y_0 + S_0(1 - e^{-kapp\ t})$$

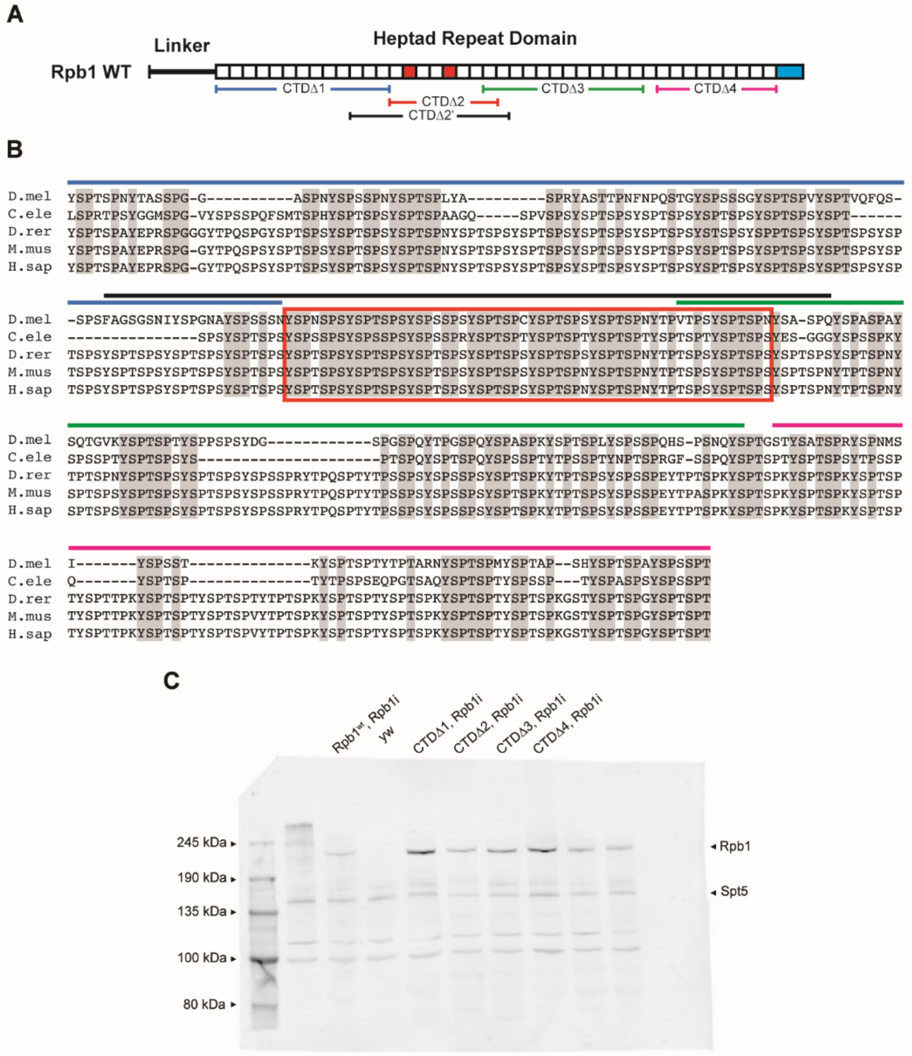
$$y = \frac{k_1\ S_0}{k_2 - \ k_1}(e^{-k1\ t} - e^{-k2\ t})$$

Where possible, k_{app} was calculated as the average between a given pSer resonance and the resonances of neighboring residues. Reported errors represent the 95% confidence intervals, or the propagation of error where k_{app} represents an average.

Small-Angle X-Ray Scattering

CTD2′ samples were expressed as described in LB media and purified as described. Following purification, samples were buffer exchanged into 80 mM Tris/HCl pH 7.5, 50 mM KCl, 10% glycerol, 5 mM DTT**.** SAXS data was collected at the Cornell High Energy Synchrotron Source (CHESS) on the G1 beamline. Incident radiation was produced at 9.963 keV with a flux of $8\text{x}10^{11}$ photons s^{-1} at 51 mA, providing a q-space range of 0.007-0.7 $Å^{-1}$. Scattering from a silver behenate standard was used for q-axis mapping. Data collection was performed using dual Pilatus 100K-S detectors. Reduction of the 2D images to 1D scattering profiles was performed using BioXtas Raw. Scattering profiles and uncertainties were computed as the average and standard deviation of three exposures, with each exposure comprising 20 one second frames. Solvent blanks were collected immediately before and after each protein sample exposure by measuring the scattering from the spin column flow through from each sample, and solvent subtraction was performed using equivalent numbers of frames. Data were collected at protein concentrations from 4-11 mg/mL for both unphosphoryalted and pSer5 CTD2′. No signs of aggregation, inter-particle effects, or radiation damage were observed. Average radius of gyration ($R_{g)}$ values were determined for each sample using the Guinier approximation with $qR_g \leq 0.8$, as suggested for disordered systems like CTD2′.[40] Guinier fitting and pair-wise distance distribution calculations were performed using the method of non-linear least squares in MATLAB and the auto-GNOM function in Primus qt, respectively.

Supporting Information



```
D.mel  YSPTSPNYTASSPG-G----------ASPNYSPSSPNYSPTSPLYA---------SPRYASTTPNFNPQSTGYSPSSSGYSPTSPVYSPTVQFQS-
C.ele  LSPRTPSYGGMSPG-VYSPSSPQFSMTSPHYSPTSPSYSPTSPAAGQ-----SPVSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPT------
D.rer  YSPTSPAYEPRSPGGGYTPQSPGYSPTSPSYSPTSPSYSPTSPNYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSTSPSYSPPTSPSYSPTSPSYSP
M.mus  YSPTSPAYEPRSPG-GYTPQSPSYSPTSPSYSPTSPSYSPTSPNYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSP
H.sap  YSPTSPAYEPRSPG-GYTPQSPSYSPTSPSYSPTSPSYSPTSPNYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSP

D.mel  -SPSFAGSGSNIYSPGNAYSPSSSNYSPNSPSYSPTSPSYSPSSPSYSPTSPCYSPTSPSYSPTSPNYTPVTPSYSPTSPNYSA-SPQYSPASPAY
C.ele  ---------------SPSYSPTSPSYSPSSPSYSPSSPSYSPSSPRYSPTSPTYSPTSPTYSPTSPTYSPTSPTYSPTSPSYES-GGGYSPSSPKY
D.rer  TSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPNYTPTSPSYSPTSPSYSPTSPSYSPTSPNY
M.mus  TSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPNYSPTSPNYTPTSPSYSPTSPSYSPTSPNYTPTSPNY
H.sap  TSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPNYSPTSPNYTPTSPSYSPTSPSYSPTSPNYTPTSPNY

D.mel  SQTGVKYSPTSPTYSPPSPSYDG------------SPGSPQYTPGSPQYSPASPKYSPTSPLYSPSSPQHS-PSNQYSPTGSTYSATSPRYSPNMS
C.ele  SPSSPTYSPTSPSYS---------------------PTSPQYSPTSPQYSPSSPTYTPSSPTYNPTSPRGF-SSPQYSPTSPTYSPTSPSYTPSSP
D.rer  TPTSPNYSPTSPSYSPTSPSYSPSSPRYTPQSPTYTPSSPSYSPSSPSYSPTSPKYTPTSPSYSPSSPEYTPTSPKYSPTSPKYSPTSPKYSPTSP
M.mus  SPTSPSYSPTSPSYSPTSPSYSPSSPRYTPQSPTYTPSSPSYSPSSPSYSPTSPKYTPTSPSYSPSSPEYTPASPKYSPTSPKYSPTSPKYSPTSP
H.sap  SPTSPSYSPTSPSYSPTSPSYSPSSPRYTPQSPTYTPSSPSYSPSSPSYSPTSPKYTPTSPSYSPSSPEYTPTSPKYSPTSPKYSPTSPKYSPTSP

D.mel  I-------YSPSST--------------KYSPTSPTYTPTARNYSPTSPMYSPTAP--SHYSPTSPAYSPSSPT
C.ele  Q-------YSPTSP--------------TYTPSPSEQPGTSAQYSPTSPTYSPSSP---TYSPASPSYSPSSPT
D.rer  TYSPTTPKYSPTSPTYSPTSPTYTPTSPKYSPTSPTYSPTSPKYSPTSPTYSPTSPKGSTYSPTSPGYSPTSPT
M.mus  TYSPTTPKYSPTSPTYSPTSPVYTPTSPKYSPTSPTYSPTSPKYSPTSPTYSPTSPKGSTYSPTSPGYSPTSPT
H.sap  TYSPTTPKYSPTSPTYSPTSPVYTPTSPKYSPTSPTYSPTSPKYSPTSPTYSPTSPKGSTYSPTSPGYSPTSPT
```

Figure 5-S1. Identification of a region of the Rpb1 CTD that is essential for rescuing lethality caused by RNAi-mediated depletion of endogenous Rpb1 (A) Schematic representation of the C-terminal heptad repeat domain of Rpb1 showing the deleted regions from the respective transgenic strains and the recombinant CTD2′ construct. (B) Evolutionary conservation of the CTD across different species: Drosophila melanogaster (D. mel); Caenorhabditis elegans (C. ele); Danio rerio (D. rer); Mus musculus (M. mus); and Homo sapien (H. sap). Conserved amino acids are highlighted in gray. CTDΔ1, CTDΔ3, CTDΔ4 deletions are denoted by the blue, green and magenta

bars, respectively. CTDΔ2 deletion (red box) contains a highly conserved region. The recombinant CTD2′ construct is denoted by the black line. (C) Uncropped image showing the western blot analysis of the expression of Rpb1 derivatives from Fig. 1c. The lanes shown in Fig. 1c are labeled for clarity.

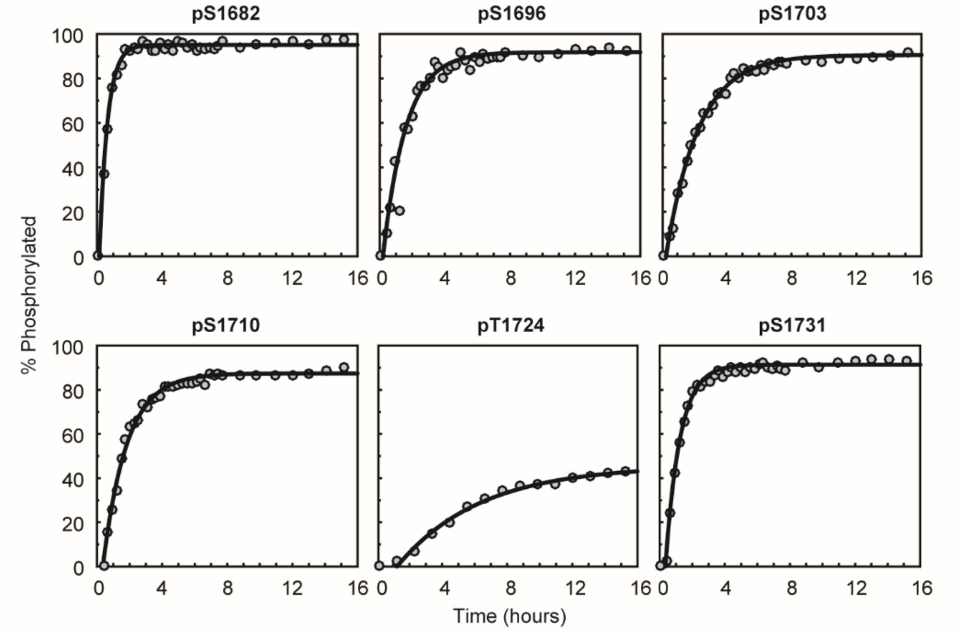


Figure 5-S2. Kinetics of DmP-TEFb phosphorylation of CTD2′ measured by RT-NMR. Peak intensities for unphosphorylated and phosphorylated species were extracted from 2D ^{1}H,^{15}N correlation spectra and the percentage of phosphorylation was plotted as a function of time. Fits were performed as described in the materials & methods.

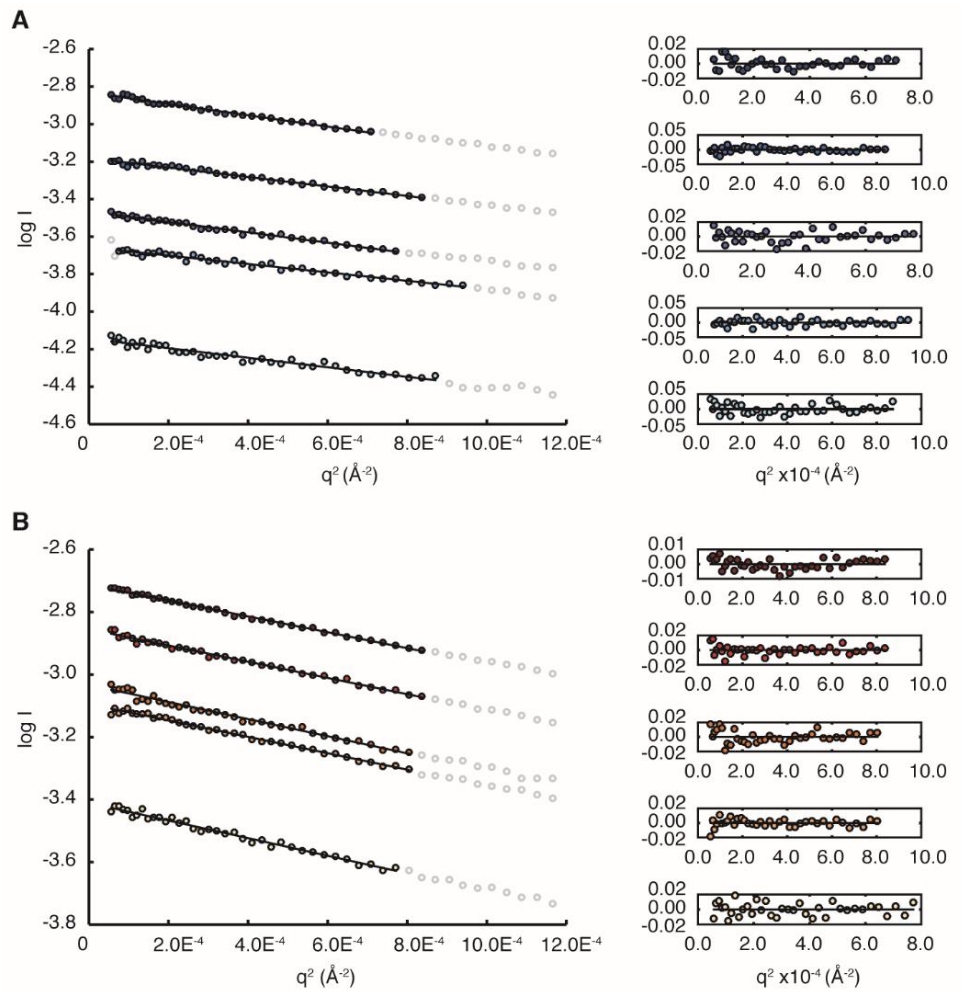


Figure 5-S3. Small angle X-ray scattering reveals no significant changes in R_g for CTD2′upon serine 5 phosphorylation. (A) Guinier fits for unphosphorylated CTD2′ collected at protein concentrations of 4.0, 5.9, 7.0, 8.8, and 11.4 mg/ml (left) with residuals (right). (B) Guinier fits for hyper-pSer5 CTD2′ collected at protein concentrations of 4.1, 6.8, 7.2, 8.7, 10.0 mg/ml (left) with residuals (right). Data over the range of $qR_g < 0.8$ was fit to the Guinier approximation in MATLAB using the method of non-linear least squares.

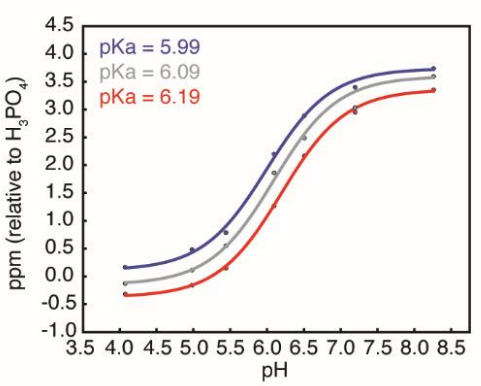


Figure 5-S4. Phosphoserine pKa values determined by ^{31}P NMR spectroscopy. pKa values for hyper-pSer5 CTD2′ were determined by non-linear least squares fitting in MATLAB. The chemical shifts of well resolved phosphoserine resonances (δ) were plotted as a function of pH and fit to $\delta = [\delta^{2-} (10^{pH-pKa}) + \delta^{-}] / [1 + 10^{pH-pKa}]$, where δ^{-} and δ^{2-} are the chemical shifts of the mono-anionic and di-anionic phosphate species, respectively. Data points and best fit lines are shown for three phosphoserine resonances, representing fitted pKa values at the acidic (blue) and basic (red) ends of the range, as well as a representative intermediate fitted value (gray).

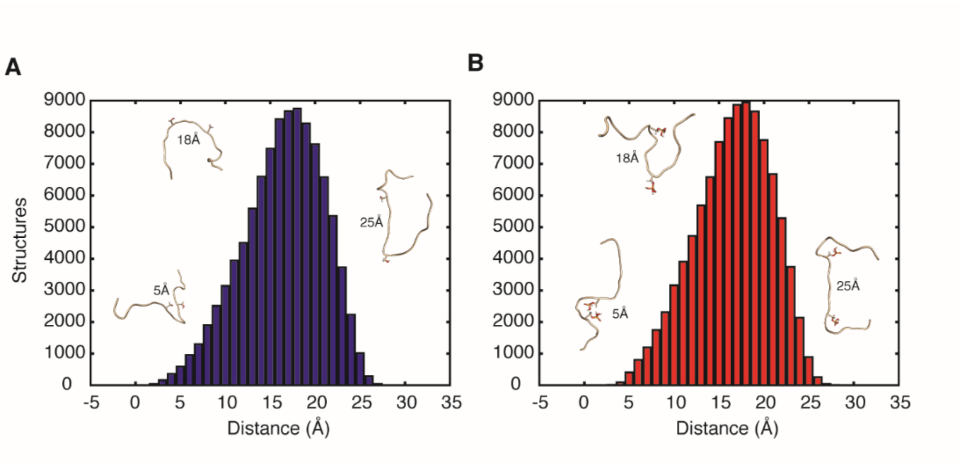


Figure 5-S5. Random coil conformations of CTD2′ heptads separate consecutive Ser5s or pSer5s by distances that exceed the Debye length. (A) Histogram showing the distribution of inter-Ser5-Ser5 distances computed from 100,000 random coil structures of YSPTSPSYSPTSSPSYSPTSPCYSPTSPS generated using traDES. (B) Histogram showing the distribution of inter-pSer5-pSer5 distances computed from 100,000 random coil structures of YSPTpSPSYSPTSpSPSYSPTpSPCYSPTpSPS generated using traDES. Distances were calculated from Oγ atoms in Ser5/pSer5 residues located within the two central heptads of each construct. Three representative structures are shown as insets to each panel.

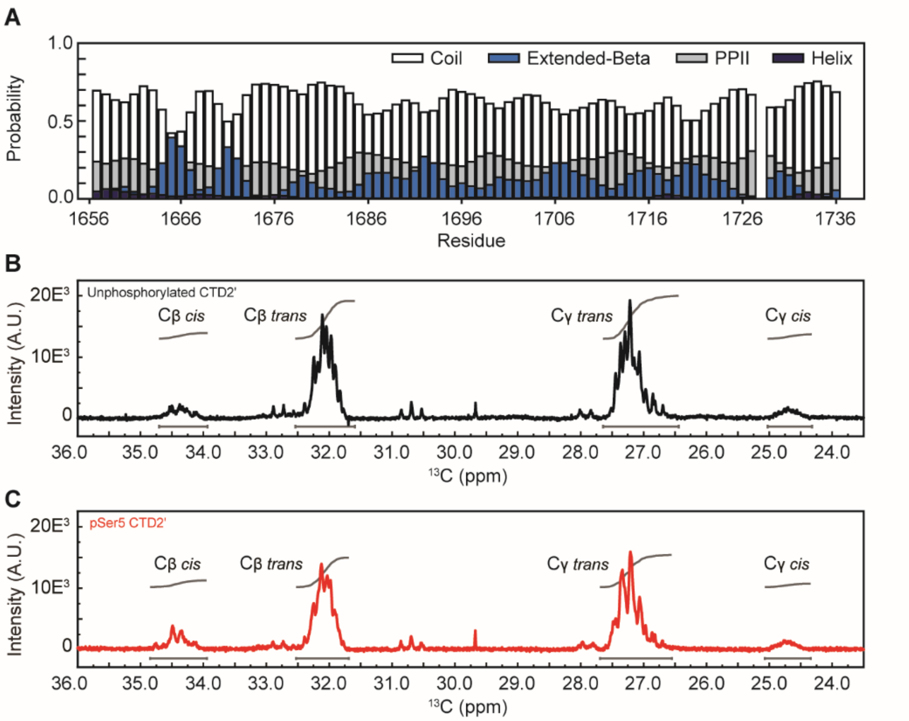


Figure 5-S6. Structural characterization of unphosphorylated CTD2′ by NMR spectroscopy. (A) Secondary structure populations for unphosphorylated CTD2′ determined with Δ2D using NMR chemical shifts (B) Proline Cβ and Cγ chemical shifts from the ^{13}C spectrum of unphosphorylated CTD2′. These peaks represent the population of *cis* and *trans* isomers averaged over all proline residues (95% *trans*, 5% *cis*) (C) In the hyper-pSer5 state, the population of *cis*-proline is enriched ~2-fold.

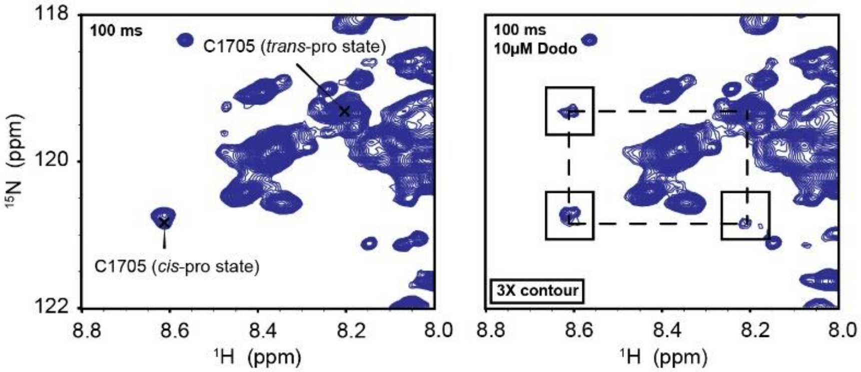


Figure 5-S7. Proline *cis-trans* isomerization in Hyper-pSer5 CTD2′ probed by NMR spectroscopy. 2D H-N correlation spectra for the measurement of ^{15}N ZZ-exchange collected on 1 mM hyper-pSer5 CTD2′, in the absence of Dodo, at a relaxation delay of 100 ms (left). In the presence of 10 μM Dodo, ZZ exchange pairs can be observed. (right)

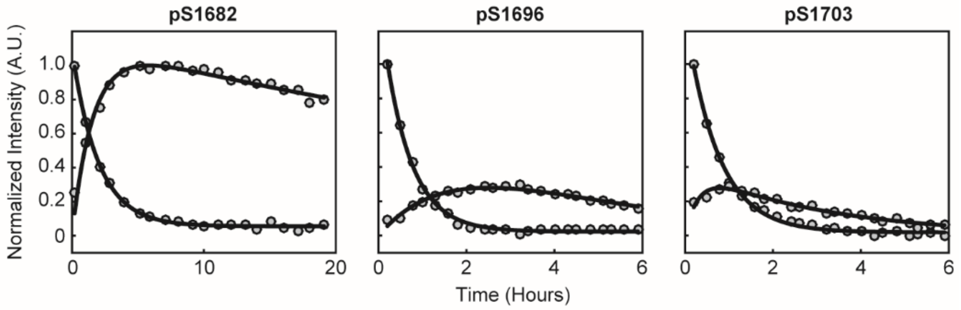


Figure 5-S8. Kinetics of Ssu72-Symplekin dephosphorylation of Hyper-pSer5 CTD2′. measured by RT-NMR. Extracted peak intensities from 2D ^{1}H,^{15}N correlation spectra were plotted as a function of time and fit as described in online methods.

Table 5-S1. Results of the rescue assay. Numbers of progeny with particular phenotypes and the percentages of straight-winged progeny are calculated as shown in (Fig. 1d). Rpb1i corresponds to the Gal4-activated, UAS-Rpb1i transgene. Rpb1WT and CTDΔ1 to Δ4 correspond to the Gal4-activated UAS-Rpb1 transgenes.

Transgenic Strain	**curly wing females**	**curly wing males**	**straight wing females**	**straight wing males**
yw	99	49	68	62
Rpb1^{wt}, Rpb1i	53	18	37	67
CTDΔ1, Rpb1i	38	15	9	3
CTDΔ2, Rpb1i	39	27	0	0
CTDΔ3, Rpb1i	25	27	7	10
CTDΔ4, Rpb1i	36	34	13	34

Table 5-S2. Kinetic parameters for DmP-TEFb phosphorylation of CTD2′.

Residue:	**Sequence**	**k_{app} (hour^{-1})**
S1675	YSPSSSN	0.15 ± 0.03
S1682	YSPNSPS	1.39 ± 0.20
S1689	YSPTSPS	0.78 ± 0.09
S1696	YSPSSPS	0.71 ± 0.06
S1703	YSPTSPC	0.76 ± 0.09
S1710	YSPTSPS	0.68 ± 0.05
S1717	YSPTSPN	0.78 ± 0.11
T1724	YTPVTPS	0.26 ± 0.02
S1731	YSPTSPN	0.95 ± 0.11
S1737	YSASPQ	0.09 ± 0.08

Table 5-S3. SAXS data collection and scattering derived parameters for unphosphorylated and hyper-pSer5 CTD2′. [†] Reported as the average ± S.E.M., [‡] Dmax shown for 4.0 mg/ml sample, * Dmax shown for 10.0 mg/ml sample.

Data Collection Parameters:	**CTD2′**	**pSer5 CTD2′**
Instrument	CHESS G1 Station	CHESS G1 Station
Beam diameter:	250 µm × 250 µm	250 µm × 250 µm
Wavelength (Å)	1.244	1.244
E (keV)	9.963	9.963
qRange ($Å^{-1}$)	0.007- 0.7	0.007- 0.7
Flux (photons/s)	8 x 10ˆ11 @ 51 ma	8 x 10ˆ11 @ 51 ma
Exposure Time (s)	60	60
Concentration (mg ml^{-1})	4-11	4-11
Temperature (K)	296	296
Structural Parameters:		
[†]*R*g (Å) (Guinier)	28.01 ± 0.69	28.27 ± 0.26
*R*g (Å) [P(r)]	27.11 ± 0.25	27.52 ± 0.84
D_{max} (Å)	111.13‡	112.48*
Software Used:		
Primary data reduction	BioXtasRAW	BioXtasRAW
Guinier fitting	MATLAB	MATLAB
P(r) calculations	PRIMUSqt/GNOM	PRIMUSqt/GNOM

Table 5-S4. Kinetic parameters for Ssu72-Symplekin dephosphorylation of Hyper-pSer5 CTD2′.

Residue:	**Sequence**	**k_{app} ($hour^{-1}$)**
pS1675	YSPSpSSN	N.D.
pS1682	YSPNpSPS	0.02 ± 0.01
pS1689	YSPTpSPS	0.39 ± 0.05
pS1696	YSPSpSPS	0.40 ± 0.06
pS1703	YSPTpSPC	0.33 ± 0.05
pS1710	YSPTpSPS	0.44 ± 0.06
pS1717	YSPTpSPN	1.10 ± 0.07
pT1724	YTPVpTPS	N.D.
pS1731	YSPTpSPN	1.19 ± 0.05
pS1737	YSApSPQ	N.D.

References

1. Eick, D. & Geyer, M. (2013) The RNA polymerase II carboxyl-terminal domain (CTD) code. Chem Rev 113, 8456-90

2. Buratowski, S. (2003) The CTD code. Nat Struct Biol 10, 679-80

3. Yang, C. & Stiller, J.W. (2014) Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain. Proc Natl Acad Sci U S A 111, 5920-5

4. Simonti, C.N. et al. (2015) Evolution of lysine acetylation in the RNA polymerase II C-terminal domain. BMC Evol Biol 15, 35

5. Chapman, R.D., Conrad, M. & Eick, D. (2005) Role of the mammalian RNA polymerase II C-terminal domain (CTD) nonconsensus repeats in CTD stability and cell proliferation. Mol Cell Biol 25, 7665-74

6. Litingtung, Y. et al. (1999) Growth retardation and neonatal lethality in mice with a homozygous deletion in the C-terminal domain of RNA polymerase II. Mol Gen Genet 261, 100-5

7. Corden, J.L. (2013) RNA polymerase II C-terminal domain: Tethering transcription to transcript and template. Chem Rev 113, 8423-55

8. Meinhart, A., Kamenski, T., Hoeppner, S., Baumli, S. & Cramer, P. (2005) A structural perspective of CTD function. Genes & Development 19, 1401-1415

9. Schuller, R. et al. (2016) Heptad-Specific Phosphorylation of RNA Polymerase II CTD. Mol Cell 61, 305-14

10. Suh, H. et al. (2016) Direct Analysis of Phosphorylation Sites on the Rpb1 C-Terminal Domain of RNA Polymerase II. Mol Cell 61, 297-304

11. Ni, Z., Schwartz, B.E., Werner, J., Suarez, J.R. & Lis, J.T. (2004) Coordination of transcription, RNA processing, and surveillance by P-TEFb kinase on heat shock genes. Mol Cell 13, 55-65

12. Czudnochowski, N., Bosken, C.A. & Geyer, M. (2012) Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition. Nat Commun 3, 842

13. Liang, K. et al. (2015) Characterization of human cyclin-dependent kinase 12 (CDK12) and CDK13 complexes in C-terminal domain phosphorylation, gene transcription, and RNA processing. Mol Cell Biol 35, 928-38

14. Cagas, P.M. & Corden, J.L. (1995) Structural studies of a synthetic peptide derived from the carboxyl-terminal domain of RNA polymerase II. Proteins 21, 149-60

15. Noble, C.G. et al. (2005) Key features of the interaction between Pcf11 CID and RNA polymerase II CTD. Nat Struct Mol Biol 12, 144-51

16. Werner-Allen, J.W. et al. (2011) cis-Proline-mediated Ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72. J Biol Chem 286, 5717-26

17. Bastidas, M., Gibbs, E.B., Sahu, D. & Showalter, S.A. (2015) A primer for carbon-detected NMR applications to intrinsically disordered proteins in solution. Concepts in Magnetic Resonance Part A 44, 54-66

18. Gibbs, E.B. & Showalter, S.A. (2015) Quantitative biophysical characterization of intrinsically disordered proteins. Biochemistry 54, 1314-26

19. Felli, I.C. & Pierattelli, R. (2014) Novel methods based on (13)C detection to study intrinsically disordered proteins. J Magn Reson 241, 115-25

20. Bienkiewicz, E.A. & Lumb, K.J. (1999) Random-coil chemical shifts of phosphorylated amino acids. J Biomol NMR 15, 203-6

21. Cordier, F., Chaffotte, A. & Wolff, N. (2015) Quantitative and dynamic analysis of PTEN phosphorylation by NMR. Methods 77-78, 82-91

22. Kohn, J.E. et al. (2004) Random-coil behavior and the dimensions of chemically unfolded proteins. Proc Natl Acad Sci U S A 101, 12491-6

23. Zor, T., Mayr, B.M., Dyson, H.J., Montminy, M.R. & Wright, P.E. (2002) Roles of phosphorylation and helix propensity in the binding of the KIX domain of CREB-binding protein by constitutive (c-Myb) and inducible (CREB) activators. Journal of Biological Chemistry 277, 42241-42248

24. Xiang, S. et al. (2013) Phosphorylation drives a dynamic switch in serine/arginine-rich proteins. Structure 21, 2162-74

25. Bah, A. et al. (2015) Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. Nature 519, 106-9

26. Maleszka, R., Hanes, S.D., Hackett, R.L., de Couet, H.G. & Miklos, G.L. (1996) The Drosophila melanogaster dodo (dod) gene, conserved in humans, is functionally interchangeable with the ESS1 cell division gene of Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 93, 447-51

27. Hanes, S.D. (2014) The Ess1 prolyl isomerase: traffic cop of the RNA polymerase II transcription cycle. Biochim Biophys Acta 1839, 316-33

28. Mayfield, J.E. et al. (2015) Chemical Tools To Decipher Regulation of Phosphatases by Proline Isomerization on Eukaryotic RNA Polymerase II. ACS Chem Biol 10, 2405-14

29. Hausmann, S., Koiwa, H., Krishnamurthy, S., Hampsey, M. & Shuman, S. (2005) Different strategies for carboxyl-terminal domain (CTD) recognition by serine 5-specific CTD phosphatases. J Biol Chem 280, 37681-8

30. Luo, Y. et al. (2013) novel modifications on C-terminal domain of RNA polymerase II can fine-tune the phosphatase activity of Ssu72. ACS Chem Biol 8, 2042-52

31. Xiang, K. et al. (2010) Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex. Nature 467, 729-33

32. Schroder, S. et al. (2013) Acetylation of RNA polymerase II regulates growth-factor-induced gene transcription in mammalian cells. Mol Cell 52, 314-24

33. Sims, R.J., 3rd et al. (2011) The C-terminal domain of RNA polymerase II is modified by site-specific methylation. Science 332, 99-103

34. Voss, K. et al. (2015) Site-specific methylation and acetylation of lysine residues in the C-terminal domain (CTD) of RNA polymerase II. Transcription 6, 91-101

35. Dias, J.D. et al. (2015) Methylation of RNA polymerase II non-consensus Lysine residues marks early transcription in mammalian cells. Elife 4.

36. Bischof, J., Maeda, R.K., Hediger, M., Karch, F. & Basler, K. (2007) An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. Proc Natl Acad Sci U S A 104, 3312-7

37. Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J. & Gygi, S.P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 24, 1285-92

38. O'Hare, B., Benesi, A.J. & Showalter, S.A. (2009) Incorporating 1H-chemical shift determination into 13C-direct detected spectroscopy of intrinsically disordered proteins in solution. Journal of Magnetic Resonance 200, 354-358

39. Sahu, D., Bastidas, M. & Showalter, S.A. (2014) Generating NMR Chemical Shift Assignments of Intrinsically Disordered Proteins using Carbon-Detected NMR Methods. Anal Biochem 449, 17-25

40. Perez, Y., Gairi, M., Pons, M. & Bernado, P. (2009) Structural characterization of the natively unfolded N-terminal domain of human c-Src kinase: insights into the role of phosphorylation of the unique domain. J Mol Biol 391, 136-48

Chapter 6

Preliminary Data for CTD (aa1815-1887)

With the proposal to apply biophysical techniques to study the *Drosophila melanogaster* CTD came the need to choose an appropriate construct. Initial studies were performed on a stretch of the CTD spanning Rpb1 amino acids 1815-1887. This construct, referred to as CTD_end, was used for the development of many of the protocols described in earlier chapters, including the MS/MS and NMR spectroscopy described in chapter 3. These preliminary studies are discussed here.

Introduction

The rationale for selecting CTD_end was several fold. Limited proteolysis experiments by Bede Portz revealed a protease sensitive region of the *Drosophila* CTD, which when cleaved produced an approximately 10 kDa fragment containing the native c-terminus of the domain (data not shown). Further, the Eick group demonstrated that the last repeat of the human CTD, or "acidic-tip", is important for preserving the stability of the domain *in vivo* [1], thus highlighting the biological significance of the region.

Results and Discussion

The use of CTD_end for methods development was advantageous as it possesses several key features, including a native casein kinase II (CKII) recognition motif within the acidic tip (EESED). One well known difficulty associated with identifying the CTD by

SDS-PAGE is its inability to be stained by common protein stains, e.g. coomassie brilliant blue [2]. Phosphorylation of the CTD by CKII, using radio labelled ATP, has therefore been used to identify the CTD on PAGE gels [2]. Figure 6-1, Lane 1. This known kinase substrate was advantageous during the development of purification protocols, and offered a seemingly straight forward path toward structural characterization of a CTD phosphoform. In addition to CKII, several kinases were tested for activity toward CTD_end, including the p42 mitogen activated protein kinase (MAPK), and the human and *Drosophila* P-TEFb. Figure 6-1.

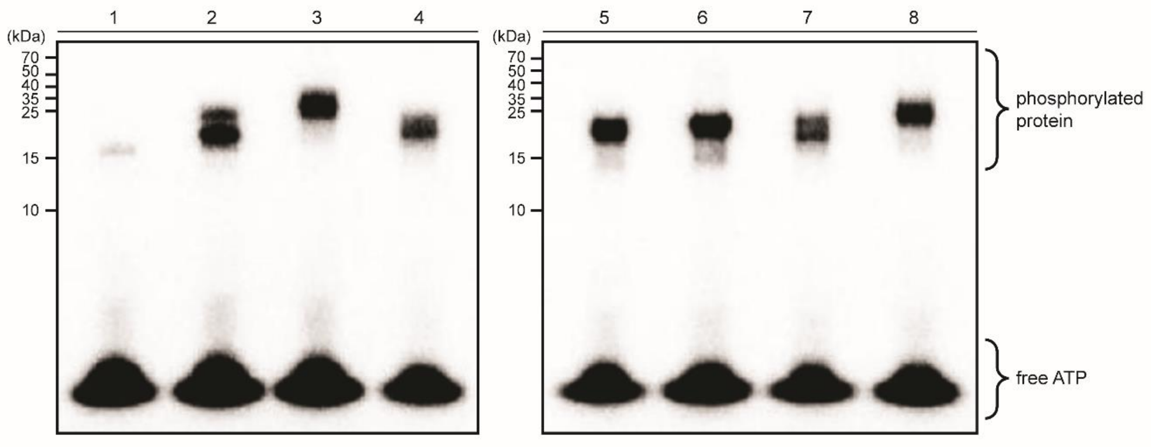


Figure 6-1. 12% SDS-PAGE displaying the products of *in vitro* kinase treatments. (lane 1) CKII treated CTD_end (lane 2) DmP-TEFb treated CTD_end (lane 3) DmP-TEFb treated CTD2′ (lane 4) CTD_end treated with CKII, MAPK, and DmP-TEFb (lane 5) MAPK treated CTD_end (lane 6) MAPK treated CTD2′ (lane 7) CTD_end treated with MAPK, and DmP-TEFb (lane 8) CTD2′ treated with MAPK, and DmP-TEFb.

Several hyperphosphorylated versions of CTD_end were analyzed by LC-ESI-MS revealing distinct phosphoisoform distributions, where the maximum number of incorporated phosphates were 2, 9, and 8, for DmP-TEFb, human P-TEFb, and MAPK, respectively. Figure 6-2. Analysis of the CKII phosphorylated CTD_end by MS was not possible as this modification led to precipitation of the protein.

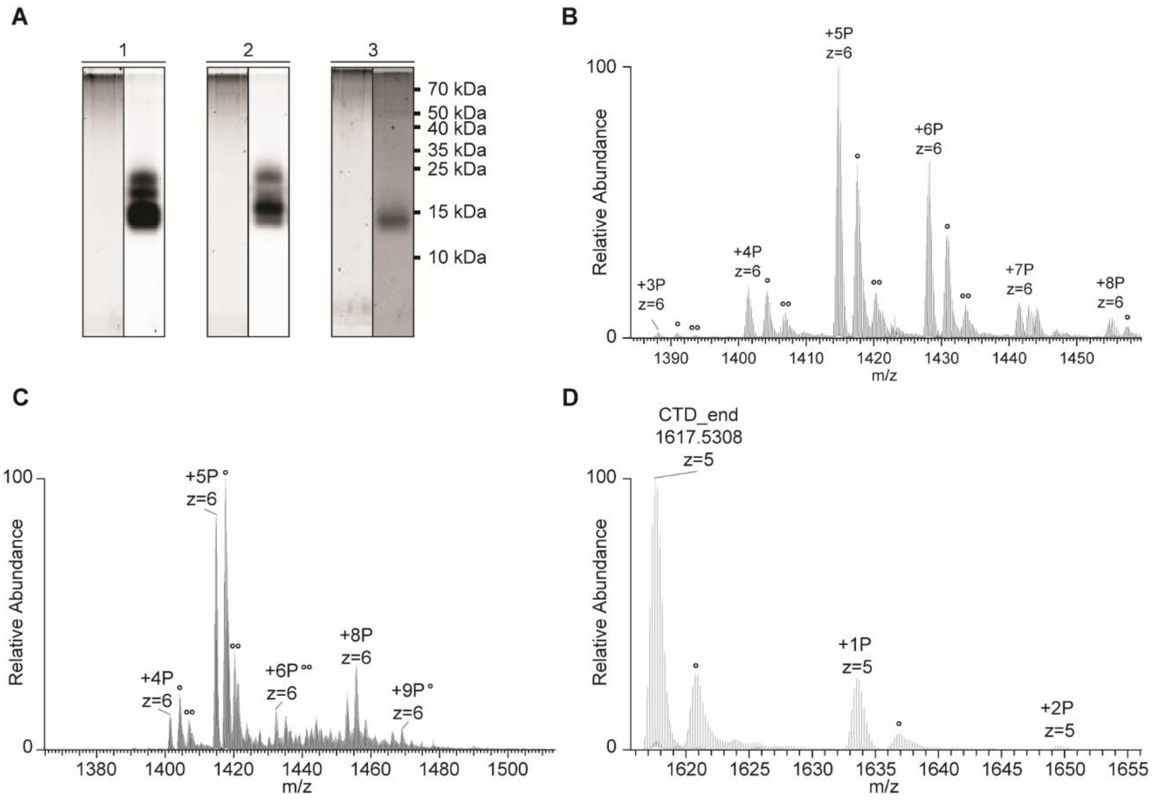


Figure 6-2. Analysis of CTD_end phosphoforms by SDS-PAGE and LC-ESI-MS. (A) 12% SDS-PAGE gels showing the products of *in vitro* kinase treatment of CTD_end by (1) MAPK (2) human P-TEFb (3) DmP-TEFb. Each gel was stained with SYPRO ruby total protein stain (left) and Pro-Q diamond phosphoprotein stain (right). These samples were analyzed in parallel using MS, where B-D corresponds to 1-3. In each MS spectrum the number of phosphates is shown and the black dots represent Methionine oxidation.

Protocols for MS/MS based phospho-site identification were also developed using CTD_end. The construct possesses two native Arginines and one Lysine residue, the C-termini of which are targets for proteolysis by trypsin, a commonly used protease for MS/MS sample preparation. However, using trypsin, the identification of phospho-sites within the MAPK hyperphosphorylated CTD_end was impeded by incomplete coverage of the polypeptide sequence at the N-terminus and the complex nature of the phosphorylated C-terminal fragments, which contained too many phosphates to accurately localized using peptide identification algorithms (see chapter 3). Figure 6-3A.

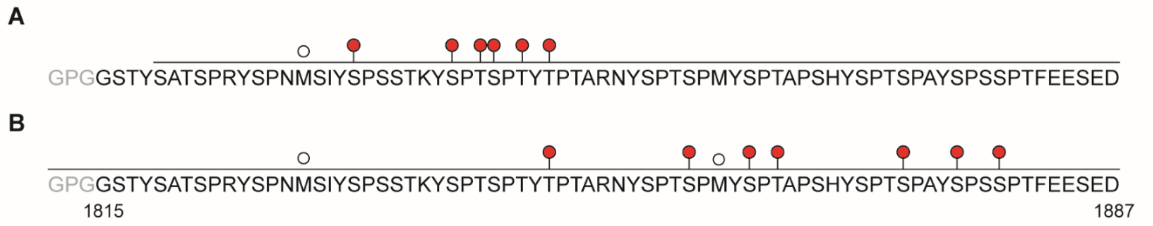

A
GPGGSTYSATSPRYSPNMSIYSPSSTKYSPTSPTYTPTARNYSPTSPMYSPTAPSHYSPTSPAYSPSSPTFEESED
B
GPGGSTYSATSPRYSPNMSIYSPSSTKYSPTSPTYTPTARNYSPTSPMYSPTAPSHYSPTSPAYSPSSPTFEESED
1815
1887

Figure 6-3. Analysis of MAPK treated CTD_end by LC-ESI-MS2 using ETD activation. Samples were digested with either trypsin (A) or chymotrypsin (B). PTMs were identified using the ZCore algorithm, where the open circles represent Methionine oxidation and the red markers indicate phosphorylation sites identified at the 99% confidence interval.

Reasoning that these problems stemmed from inefficient fragmentation of the proteolytic fragments in the ion-trap, we hypothesized that better results could be obtained using a protease that could generate smaller proteolytic fragments. As an alternative, we turned to chymotrypsin, which specifically cleaves peptide bonds at the C-termini of Tyr, Phe, Trp and Leu, and should in principle cleave at each heptad. Digestions of MAPK phosphorylated CTD_end were performed using several concentrations of chymotrypsin and analyzed by SDS-PAGE. Figure 6-4. The results showed that proteolysis of MAPK phosphorylated CTD_end by chymotrypsin produces smaller fragments relative to those generated through trypsin digest, many that are approximately the mass of a single heptad. And analysis of the chymotrypsin digested phospho-CTD_end by LC-ESI-MS/MS yielded complete coverage of the polypeptide sequence and revealed several C-terminal phosphorylation sites. Taken as a whole, the MS2 data for CTD_end demonstrates that MAPK produces a heterogeneous mixture of pSer2, pThr4, and pSer5 phosphorylation marks. Figure 6-3B.

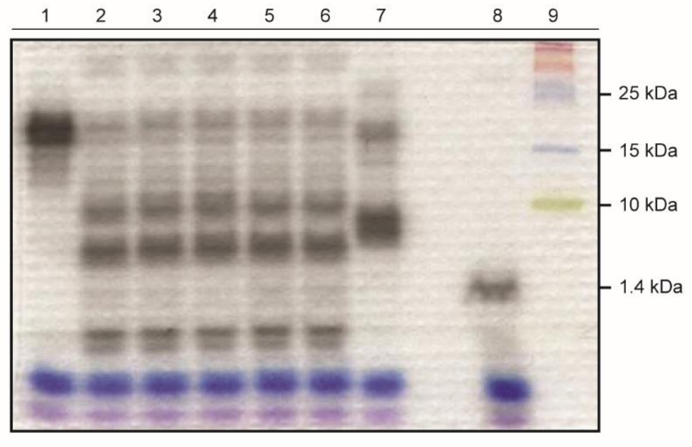


Figure 6-4. Analysis of the proteolytic digests of MAPK phosphorylated CTD_end by SDS-PAGE. The lanes correspond to (1) an undigested control, (2-6) overnight digests using 0.5-8.0 μg/ul of chymotrypsin, and (7) 6.6 ng/ul Trypsin. Molecular mass markers include lane (8) a monophosphorylated CKII substrate peptide and (9) a 10-260 kDa protein ladder.

Having established a route to generation of multiple phosphoforms of CTD_end, we sought to test the hypothesis that phosphorylation alters the conformational ensemble of the *D. mel* CTD. To this end, we turned to NMR spectroscopy. Figure 6-5. Upon phosphorylation, chemical shift perturbations were observed in a ^{1}H,^{15}N-HSQC spectrum, consistent with phosphate incorporation. Figure 6-5A Using ^{13}C Direct-Detect NMR spectroscopy several resonances were assigned, particularly in the C-terminal acidic tip. Figure 6-5B-C.

However, due to difficulties with protein stability at the high concentrations need for the (HACA)N(CA)CON and (HACA)N(CA)NCO experiments described in chapter 4 (~1 mM), complete back bone resonance assignments were not obtained. And studies of CTD_end were eventually put on hold as the focus shifted to CTD2′.

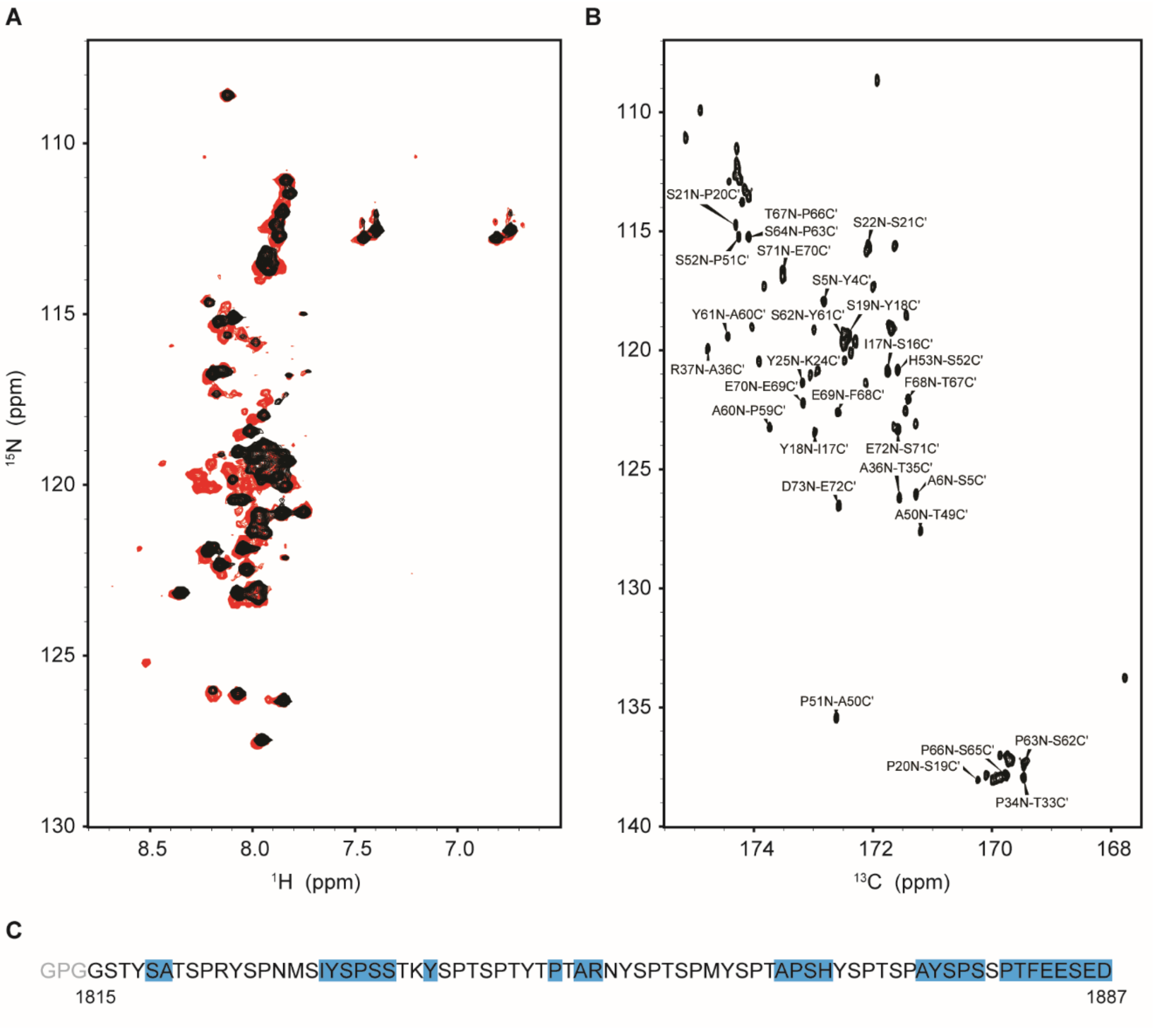


Figure 6-5. Analysis of the unphosphorylated and DmP-TEFb treated CTD_end by NMR spectroscopy. (A) Overlaid ^{1}H,^{15}N-HSQC spectra of the unphosphorylated (black) and phosphorylated (red) forms of CTD_end show characteristic chemical shift perturbations induced by phosphorylation (B) The ^{13}C,^{15}N-CON spectrum of the unphosphorylated CTD_end is displayed with the assigned resonances labeled. (C) The amino acid sequence of CTD_end is shown with the assigned resonances highlighted in blue.

Conclusion

The region of the *Drosophila* CTD encompassed by CTD_end is of interest due to its sequence conservation to regions of the human CTD that have been shown to maintain the stability of the CTD *in vivo* [1], and that are hypothesized to mediate the assembly of the polymerase into liquid-like droplet states [3]. The CTD_end construct proved useful for the development of MS and NMR based protocols for phosphorylation site identification and characterization. However, it was ultimately abandoned to pursue studies of the CTD2′ construct. The preliminary studies outlined here may therefore serve as a starting point for further investigation of this region of the *D. mel* CTD.

Materials & Methods

Protein Expression and Purification:

A synthetic gene for the *Drosophila melanogaster* RPB1 Carboxy-terminal domain was purchased from GeneArt (Thermo Fisher Scientific). A region corresponding to residues (1815-1887) was amplified by PCR, cut with XhoI (NEB) and XmaI (NEB), and ligated into the pET49b+ expression vector (Novagen) using T4 DNA ligase (NEB) to produce a construct containing GST & His tags. Protein expression was performed in *E. coli* BL21 DE3 cells. 500 ml batch cultures were grown at 37 ^{o}C in Luria-Bertani (LB) medium supplemented with 30μg/mL kanamycin. At an optical density at 600 nm (OD600) of 0.8, cells were induced using 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG) and allowed to incubate at 37 ^{o}C for 3 hours. Following lysis by sonication on ice in lysis buffer (50 mM Tris/HCl pH 7.5, 500 mM NaCl, 20 mM Imidaozole, 2.5 mM β-mercaptoethanol, 10X EDTA-free protease inhibitor cocktail (Calbiochem), and 10 units of RNAse free DNAse (NEB)), samples were centrifuged at 4 ^{o}C for 40 minutes at 11,500 x g. Cleared supernatant was passed over HisPur Ni^{2+}-NTA resin (Thermo Fisher Scientific) and bound protein was washed of contaminants using 5 column volumes of wash buffer (50 mM Tris/HCl pH 7.5, 500 mM NaCl, 20 mM imidaozole, 0.1% Triton-1000, 2.5 mM β-mercaptoethanol). Protein was eluted using elution buffer (50 mM Tris/HCl pH 7.5, 500 mM NaCl, 200 mM imidaozole, 2.5 mM β-mercaptoethanol). The GST and 6XHis tags were removed by adding recombinant His-tagged HRV 3C protease to the protein (resulting in an N-terminal non-native GPG) and dialyzing the mixture overnight against 50 mM Tris/HCl pH 7.5, 300 mM NaCl, 2.5 mM β-mercaptoethanol at 4 ^{o}C. The protein was then passed over the Ni^{2+}-NTA column to remove the protease and non-specifically bound contaminants. For MS experiments, further purification was performed by size exclusion chromatography in 80 mM Imidazole pH 6.5, 50 mM KCl, 2.5 mM β-mercaptoethanol using P-10 resin (BioRad). Samples were then buffer exchanged into kinase buffer, kinase treated, and subsequently purified by strong anion exchange chromatography using an Acrosep-Q column (Pall) on a BioLogic DuoFlow FPLC (biorad) in 50 mM Tris HCL pH 7.5, 50 mM NaCl, 1 mM EDTA with a NaCl gradient of 50 mM to 500 mM.

Kinase Assays:

For the samples displayed in Figure 1, kinase treatment was performed overnight at 30 ^{o}C in kinase buffer (50 mM Tris HCl pH 7.5, 50 mM NaCl, 10 mM $MgCl_2$, 2 mM

ATP, and γ^{32}P-ATP [~1 μCi]) with 1 unit of enzyme (as specified by manufacturer) and 100 μM CTD.

Protease Assays:

For the samples displayed in Figure, MAPK reactions were performed overnight at 30 °C in 80 mM Imidazole pH 7.2, 50 mM KCl, 10 mM $MgCl_2$, 3 mM ATP, and trace γ^{32}P-ATP with 1 unit of enzyme (as specified by manufacturer) and 100 μM CTD_end. CKII reactions were performed overnight at 30 °C in 50 mM Tris HCl pH 7.5, 50 mM NaCl, 10 mM $MgCl_2$, 750 μM ATP, and trace γ^{32}P-ATP with 1 unit of enzyme (as specified by manufacturer) and 100 μM CKII substrate peptide (RRRADDSDDDDD) (Enzo life sciences). Following kinase treatment, excess γ^{32}P-ATP was removed by size exclusion chromatography using Micro Bio Spin 6 columns (Biorad). Lyophilized trypsin and chymotrypsin stocks were reconstituted in 1 mM HCl, and working solutions were prepared in imidazole buffer at concentrations of 0.5 μg/ml and 40 μg/ml, respectively. Digestion reactions were prepared in imidazole buffer using 20 μl of hyperphosphorylated CTD_end 7 ng/ml trypsin and a series of chymotrypsin concentrations (0.5, 1, 2, 4, 8 μg/ml). Trypsin and chymotrypsin digestions proceeded overnight at 37 °C and 25 °C, respectively.

LC-ESI-MS and MS2:

All mass spectrometry was performed at the Proteomics Core facility within the Huck Institutes of Life Sciences at the Pennsylvania State University. Proteolytic digests proceeded overnight using either trypsin or chymotrypsin. LC-MS/MS of hyperphosphorylated dCTD_end constructs was performed on a Thermo LTQ OrbitTrap Velos using dual ETD-CID activation. Data were processed in Proteome Discoverer (Thermo Fisher Scientific, San Jose, CA, USA; version 1.3.0.339) using the ZCore algorithm with a fragment ion mass tolerance of 0.50 Da and a parent ion tolerance of 1.00 Da. Oxidation of Met and phosphorylation of Ser, Thr and Tyr as variable modifications.

NMR Spectroscopy:

For routine C_CON spectra, 16 transients were collected with 1024(C)x256(N) points, sweep widths of 12x34 and a recycle delay of 1.3s. For the spectrum in Figure. 6-5B, the buffer used was 50 mM Tris HCl pH 7.5, 50 mM NaCl. Routine HSQC spectra were collected with 4-16 scans, 2048(H)x256(N) complex points, sweep widths of 12x22 ppm, and recycle delays of 1-1.3s. For the spectra in Figure. 6-5A, the buffer used was 80 mM Imidazole pH 6.5, 50 mM KCl, 10 % glycerol.

References

1. Chapman, R.D., Palancade, B., Lang, A., Bensaude, O., & Eick, D., The last CTD repeat of the mammalian RNA polymerase II large subunit is important for its stability. Nucleic Acids Res. 2004; 32(1): 35–44.
2. Dahmus M.E. (1981) Phosphorylation of eukaryotic DNA-dependent RNA polymerase. Identification of calf thymus RNA polymerase subunits phosphorylated by two purified protein kinases, correlation with in vivo sites of phosphorylation in HeLa cell RNA polymerase II. J. Biol. Chem., 256, 3332–3339
3. Kwon, I., Kato, M., Xiang, S., Wu, L., Theodoropoulos, P., Mirzaei, H., Han, T., Xie, S., Corden, J. L., & McKnight, S.L. (2013) Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. Cell. 155(5):1049-60.

Appendix A

Abbreviations

IDP, intrinsically disordered protein; NMR, nuclear magnetic resonance; SAXS, small angle X-ray scattering; smFRET, single-molecule Förster resonance energy transfer; HSQC, heteronuclear single quantum correlation; TROSY, transverse relaxation optimized spectroscopy; MoRFs, molecular recognition fragments; RDC, residual dipolar coupling; PRE, paramagnetic relaxation enhancement; Rg, radius of gyration; SANS, small angle neutron scattering; CV, contrast variation; CV-SANS, contrast variation small angle neutron scattering; GdmCl, guanidinium-HCl; pE-DB, Protein Ensemble Database; MD, molecular dynamics; REMD, replica exchange molecular dynamics; AMD, accelerated molecular dynamics; MC, Monte Carlo; K_d, dissociation constant; K_a, association constant; FCS, fluorescence correlation spectroscopy; ITC, isothermal titration calorimetry; ΔH, change in binding enthalpy; ΔG, change in Gibbs free energy; ΔS, change in binding entropy; ΔCp, constant-pressure heat capacity change; k_{obs}, observed rate constant; PPi, diphosphate; $\Delta\Delta G^{\dagger\dagger}$, change in activation energy for folding upon mutation; $\Delta\Delta G^{eq}$, change in equilibrium free energy upon mutation, LC-ESI-MS; liquid chromatography electrospray ionization mass spectrometry, MALDI; matrix assisted laser deabsorption ionization, TOF; time of flight, FT; Fourier transform, CDK; cyclin dependent kinase.

Appendix B

Theoretical and Observed MS2 Ions – Chapter 3

Table A-1. MALDI-ISD ion series observed unphosphorylated CTD2′

c	Theoretical (m/z)	Observed (m/z)	Intensity (A.U.)	y	Theoretical (m/z)	Observed (m/z)	Intensity (A.U.)
10	850.369	849.31	7966	8	863.3894	863.335	14088
11	963.453	962.524	11614	9	950.4214	950.465	7319
12	1126.516	1125.675	18671	10	1051.4691	1051.579	1971
14	1310.601	1309.783	6276	11	1148.5218	1148.671	17226
15	1367.623	1366.81	19517	12	1235.5539	1235.714	9561
16	1481.666	1480.854	12549	13	1398.6172		
17	1552.703	1551.894	12032	14	1485.6492	1485.814	2465
18	1715.766	1714.97	16006	15	1582.702	1582.881	19430
20	1899.851	1899.083	6937	16	1683.7497	1683.925	7889
21	1986.883	1986.131	8943	17	1782.8181	1784.971	3736
22	2073.915	2073.183	7856	18	1879.8708	1880.078	12956
23	2160.947	2160.236	9317	19	1980.9185	1981.138	4262
24	2274.99	2274.305	6271	20	2143.9819	2144.211	2248
25	2438.053	2437.406	7396	21	2258.0248	2258.289	1678
27	2622.138	2621.504	4172	22	2355.0775	2355.386	5624
28	2736.181	2735.553	5666	23	2442.1096	2442.423	2484
30	2920.266	2919.596	1973	24	2543.1573		
31	3007.298	3006.593	2449	25	2640.21	2640.547	3051
32	3170.361	3169.565	2889	26	2727.242	2727.579	1742
33	3354.446	3353.462	891	27	2890.3054	2890.6	732
34	3455.494	3454.383	2050	28	2977.3374		
36	3639.578	3638.144	697	29	3074.3902	3074.629	1815
37	3726.61	3724.991	961	30	3161.4222	3161.586	837
38	3889.674	3887.638	1036	32	3359.5226	3359.513	1012
40	4073.758	4071.133	441	33	3446.5547	3446.407	490
41	4160.79	4157.836	802				
43	4344.875	4341.126	306				

Table A-2. MALDI-ISD ion series observed DmP-TEFb phosphorylated CTD2′

c	Theoretical (m/z)	Observed (m/z)	Intensity (A.U.)	y	Theoretical (m/z)	Observed (m/z)	Intensity (A.U.)
10	850.369	849.289	11516	8	863.3894	863.316	14282
11	963.453	962.508	15737	8+pi	943.3557	943.379	3278
12	1126.516	1125.663	23904	11	1148.522	1147.649	5193
14	1310.601	1309.745	2972	11+Pi	1228.488	1228.636	13475
15	1367.623	1366.794	13801	12+Pi	1315.52	1315.675	9755
14+pi	1390.567	1389.753	6395	13+Pi	1478.584	1478.743	4122
15+pi	1447.589	1446.767	17055	14+Pi	1565.449	1565.778	3673
16	1481.666	1480.838	8814	15	1582.702	1582.803	2222
17	1552.703	1551.886	8865	14+2Pi	1662.668	1662.848	13594
15+pi	1561.632	1560.815	9236	16+2Pi	1843.682	1843.864	2225
17+pi	1632.669	1631.86	8524	19+Pi	2140.851	2141.099	2443
18	1715.766	1714.971	10871	20+2Pi	2304.918	2304.189	1671
18+pi	1795.732	1794.944	9515	21+2Pi	2418.961	2418.27	1275
20	1899.851	1899.051	1901	26+3Pi	2968.144	2967.42	651
20+pi	1979.817	1979.064	4303				
21	1986.883	1986.121	2247				
20+2Pi	2059.783	2059.027	1872				
21+pi	2066.849	2066.112	5345				
21+2pi	2146.815	2146.084	2128				
22+pi	2153.881	2153.165	3690				
22+2pi	2233.848	2233.147	2413				
23+pi	2241.917	2240.207	2096				
23+2pi	2321.883	2320.199	3580				
24	2355.959	2354.265	1488				
23+3pi	2401.849	2400.179	1502				
24+2pi	2435.926	2434.249	2631				
24+3pi	2515.892	2515.324	3306				
25+pi	2519.023	2518.39	1152				
25+2pi	2598.989	2597.345	2503				
25+3Pi	2678.955	2677.302	863				
27+2pi	2783.074	2781.417	1303				

Appendix C

Theoretical and Observed MS2 Ions – Chapter 4

Table B-1. MS2 ion series for WT: precursor @ 1537 m/z (pSer2 State)

		WT: precursor @ 1537 m/z (pSer2 State)					
		b			b-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	1	---			---		
P	2	185.0921			---		
S	3	272.1241			---		
Y	4	435.1874			---		
pS	5	602.1858			504.2089	504.069	2214
P	6	699.2385			601.2617	601.075	1590
T	7	800.2862	800.138	7222	702.3093	702.166	4723
S	8	887.3183	887.16	10416	789.3414	789.186	54190
P	9	984.371			886.3941		
S	10	1071.403	1071.234	7359	973.4262	973.275	1630
Y	11	1234.466	1234.33	38767	1136.49	1136.346	6437
S	12	1321.498	1321.404	55729	1223.522	1223.396	7863
P	13	1418.551	1418.54	2641	1320.574	1321.404	55729
T	14	---			---		
		y			y-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	14	---			---		
P	13	1450.577	1450.574	4127	1352.601	1353.436	13487
S	12	1353.525	1353.436	13487	1255.548		
Y	11	1266.493	1266.364	8456	1168.516		
pS	10	1103.429	1103.274	50069	1005.452	1005.28	3336
P	9	936.4309	936.275	3937	---		
T	8	839.3781			---		
S	7	738.3305	738.157	1395	---		
P	6	651.2984	651.137	35160	---		
S	5	554.2457			---		
Y	4	467.2136			---		
S	3	304.1503	304.042	857	---		
P	2	217.1183	217.035	1307	---		
T	1	120.0655			---		

	Theoretical	Observed	Intensity
MH	1537.609	1537.662	68623
MH-H_2O	1519.599	1519.694	2617
MH-H_3PO_4	1439.633	1439.613	90148

Table B-2. MS2 ion series for WT: precursor @ 1537 m/z (pSer5 State)

		WT: precursor @ 1537 m/z (pSer5 State)					
		b			b-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	1	---			---		
P	2	185.0921	185.02	1543	---		
S	3	272.1241	272.001	617	---		
Y	4	435.1874	435.042	3168	---		
S	5	522.2195	522.062	6066	---		
P	6	619.2722	619.105	3052	---		
T	7	720.3199	720.175	16858	---		
pS	8	887.3183	887.16	10416	789.3414	789.186	54190
P	9	984.371			886.3941		
S	10	1071.403	1071.234	7359	973.4262	973.275	1630
Y	11	1234.466	1234.33	38767	1136.49	1136.346	6437
S	12	1321.498	1321.404	55729	1223.522	1223.396	7863
P	13	1418.551	1418.54	2641	1320.574	1321.404	55729
T	14	---			---		
		y			y-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	14	---			---		
P	13	1450.577	1450.574	4127	1352.601	1353.436	13487
S	12	1353.525	1353.436	13487	1255.548		
Y	11	1266.493	1266.364	8456	1168.516		
S	10	1103.429	1103.274	50069	1005.452	1005.28	3336
P	9	1016.397	1016.254	94291	918.4203	918.281	7181
T	8	919.3445			821.3676		
pS	7	818.2968	818.126	6935	720.3199	720.175	16858
P	6	651.2984	651.137	35160	---		
S	5	554.2457			---		
Y	4	467.2136			---		
S	3	304.1503	304.042	857	---		
P	2	217.1183	217.035	1307	---		
T	1	120.0655			---		

	Theoretical	Observed	Intensity
MH	1537.609	1537.662	68623
MH-H_2O	1519.599	1519.694	2617
MH-H_3PO_4	1439.633	1439.613	90148

Table B-3. MS2 ion series for WT: precursor @ 1617 m/z (pSer2/pSer5 State)

		WT: precursor @ 1617 m/z (pSer2/pSer5 State)					
		b			b-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	1	---			---		
P	2	185.0921	184.899	299	---		
S	3	272.1241			---		
Y	4	435.1874	434.856	282	---		
pS	5	602.1858	601.832	452	504.2089		
P	6	699.2385			601.2617		
T	7	800.2862	800.039	1712	702.3093		
pS	8	967.2846			869.3077		
P	9	1064.337			966.3605		
S	10	1151.369			1053.393		
Y	11	1314.433			1216.456	1216.337	2257
S	12	1401.465			1303.488	1303.353	3566
P	13	1498.518			1400.541		
T	14	---			---		
		y			y-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	14	---			---		
P	13	1530.544			1432.567		
S	12	1433.491			1335.514		
Y	11	1346.459			1248.482		
pS	10	1183.396	1183.206	1740	1085.419		
P	9	1016.397	1016.175	9392	918.4203		
T	8	919.3445			821.3676		
pS	7	818.2968	818.045	920	720.3199		
P	6	651.2984	651.045	3131	---		
S	5	554.2457			---		
Y	4	467.2136			---		
S	3	304.1503			---		
P	2	217.1183	216.99	239	---		
T	1	120.0655			---		

	Theoretical	Observed	Intensity
MH	1617.576	1617.616	9062
MH-H_2O	1599.565	1599.8	1088
MH-H_3PO_4	1519.599	1519.657	20888
MH-H_3PO_4-H_2O	1501.599	1501.431	19907

Table B-4. MS2 ion series for S2A: precursor @ 1521 m/z (pSer5 State)

		S2A: precursor @ 1521 m/z (pSer5 State)					
		b			b-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	1	---			---		
P	2	185.0921	185.012	635	---		
S	3	272.1241	271.985	404	---		
Y	4	435.1874	435.049	995	---		
A	5	506.2245	506.072	4336	---		
P	6	603.2773	603.08	2470	---		
T	7	704.325	704.144	5255	---		
pS	8	871.3233	871.141	6850	773.3464	773.182	21315
P	9	968.3761			870.3992		
S	10	1055.408	1055.228	3655	957.4312		
Y	11	1218.472	1218.354	26305	1120.495		
S	12	1305.504	1305.436	34456	1207.527		
P	13	1402.556			1304.579		
T	14	---			---		
		y			y-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	14	---			---		
P	13	1434.583	1434.5	2204	1336.606		
S	12	1337.53	1337.461	8620	1239.553		
Y	11	1250.498	1250.393	5221	1152.521		
A	10	1087.434	1087.308	32239	989.4575		
P	9	1016.397	1016.257	76279	918.4203	918.281	5522
T	8	919.3445			821.3676		
pS	7	818.2968	818.125	3648	720.3199	720.16	3108
P	6	651.2984	651.123	16754	---		
S	5	554.2457			---		
Y	4	467.2136			---		
S	3	304.1503			---		
P	2	217.1183			---		

T	1	120.0655			---		

	Theoretical	Observed	Intensity
MH	1521.615	1521.658	21701
$MH-H_3PO_4$	1423.638	1423.655	48878

Table B-5. MS2 ion series for S5A: precursor @ 1521 m/z (Internal pSer2 State)

		S5A: precursor @ 1521 m/z (Internal pSer2 State)					
		b			$b-H_3PO_4$		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	1	---			---		
P	2	185.0921	185.063	1288	---		
S	3	272.1241	272.036	362	---		
Y	4	435.1874	435.12	3145	---		
pS	5	602.1858	602.121	2661	504.2089		
P	6	699.2385			601.2617		
T	7	800.2862			702.3093		
A	8	871.3233	871.272	70085	773.3464		
P	9	968.3761	968.407	1483	870.3992		
S	10	1055.408	1055.386	4993	957.4312		
Y	11	1218.472	1218.457	12477	1120.495		
S	12	1305.504	1305.588	24335	1207.527		
P	13	1402.556	1402.73	1580	1304.579		
T	14	---			---		
		y			$y-H_3PO_4$		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	14	---			---		
P	13	1434.583	1434.711	2124	1336.606		
S	12	1337.53	1337.624	4147	1239.553		
Y	11	1250.498	1250.581	3136	1152.521		
pS	10	1087.434	1087.435	12822	989.4575		
P	9	920.436	920.399	45177	---		
T	8	823.3832			---		
A	7	722.3355	722.268	16427	---		
P	6	651.2984	651.23	64883	---		
S	5	554.2457			---		
Y	4	467.2136			---		
S	3	304.1503	304.079	506	---		
P	2	217.1183			---		
T	1	120.0655			---		

	Theoretical	Observed	Intensity
MH	1521.615	1521.837	21542
MH-H_3PO_4	1423.638	1423.832	125433

Table B-6. MS2 ion series for S5A: precursor @ 1521 m/z (+1 heptad pSer2 State)

		S5A: precursor @ 1521 m/z (+1 heptad pSer2 State)					
		b			b-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	1	---			---		
P	2	185.0921			---		
S	3	272.1241			---		
Y	4	435.1874	435.12	3145	---		
S	5	522.2195	522.147	1692	---		
P	6	619.2722			---		
T	7	720.3199			---		
A	8	791.357	791.299	19683	---		
P	9	888.4098			---		
S	10	975.4418	975.418	3757	---		
Y	11	1138.505	1138.52	6868	---		
pS	12	1305.504	1305.588	24335	1207.527		
P	13	1402.556	1402.73	1580	1304.579		
T	14	---			---		
		y			y-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	14	---			---		
P	13	1434.583	1434.711	2124	1336.606		
S	12	1337.53	1337.624	4147	1239.553		
Y	11	1250.498	1250.581	3136	1152.521	1152.605	2532
S	10	1087.434	1087.435	12822	989.4575	989.462	4888
P	9	1000.402	1000.39	21893	902.4254	902.409	6898
T	8	903.3496			805.3727		
A	7	802.3019	802.252	10523	704.325	704.271	5537
P	6	731.2648	731.206	44715	633.2879	633.234	8370
S	5	634.212			536.2351		
Y	4	547.18			449.2031	449.116	440
pS	3	384.1166	384.033	317	286.1397	286.096	377
P	2	217.1183			---		
T	1	120.0655			---		

	Theoretical	Observed	Intensity
MH	1521.615	1521.837	21542
MH-H_3PO_4	1423.638	1423.832	125433

Table B-7. MS2 ion series for Y1A: precursor @ 1445 m/z (pSer2 State)

		Y1A: precursor @ 1445 m/z (pSer2 State)					
		b			b-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	1	---			---		
P	2	185.0921	185.035	1541	---		
S	3	272.1241	272	454	---		
A	4	343.1612	343.124	820	---		
pS	5	510.1596	509.12	1820	412.1827		
P	6	607.2123			509.2354		
T	7	708.26	708.161	4912	610.2831		
S	8	795.292			697.3151	697.248	44334
P	9	892.3448	892.282	1838	794.3679		
S	10	979.3768	979.331	12165	881.3999		
Y	11	1142.44	1142.414	47046	1044.463	1044.458	14729
S	12	1229.472	1229.5	70158	1131.495	1131.505	19731
P	13	1326.525	1326.611	3931	1228.548		
T	14	---			---		
		y			y-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	14	---			---		
P	13	1358.551			1260.574		
S	12	1261.498	1261.53	15413	1163.522		
A	11	1174.466	1174.485	8368	1076.49		
pS	10	1103.429	1103.425	36138	1005.452	1005.427	4761
P	9	936.4309	936.391	35180	---		
T	8	839.3781			---		
S	7	738.3305	738.252	10936	---		
P	6	651.2984	651.227	55814	---		
S	5	554.2457			---		
Y	4	467.2136			---		
S	3	304.1503	304.087	776	---		
P	2	217.1183	217.078	1051	---		
T	1	120.0655			---		

	Theoretical	Observed	Intensity
MH	1445.583	1445.689	20493
MH-H_3PO_4	1347.606	1347.711	116017

Table B-8. MS2 ion series for Y1A: precursor @ 1445 m/z (pSer5 State)

		Y1A: precursor @ 1445 m/z (pSer5 State)					
		b			b-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	1	---			---		
P	2	185.0921	185.035	1541	---		
S	3	272.1241	272	454	---		
A	4	343.1612	343.124	820	---		
S	5	430.1932	430.116	3522	---		
P	6	527.246	527.139	1345	---		
T	7	628.2937	628.206	7696	---		
pS	8	795.292			697.3151	697.248	44334
P	9	892.3448	892.282	1838	794.3679		
S	10	979.3768	979.331	12165	881.3999		
Y	11	1142.44	1142.414	47046	1044.463	1044.458	14729
S	12	1229.472	1229.5	70158	1131.495	1131.505	19731
P	13	1326.525	1326.611	3931	1228.548		
T	14	---			---		
		y			y-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	14	---			---		
P	13	1358.551			1260.574		
S	12	1261.498	1261.53	15413	1163.522		
A	11	1174.466	1174.485	8368	1076.49		
S	10	1103.429	1103.425	36138	1005.452	1005.427	4761
P	9	1016.397	1016.376	75917	918.4203	918.391	7567
T	8	919.3445			821.3676		
pS	7	818.2968	818.229	7515	720.3199	720.263	6811
P	6	651.2984	651.227	55814	---		
S	5	554.2457			---		
Y	4	467.2136			---		
S	3	304.1503	304.087	776	---		
P	2	217.1183	217.078	1051	---		
T	1	120.0655			---		

	Theoretical	Observed	Intensity
MH	1445.583	1445.689	20493
MH-H_3PO_4	1347.606	1347.711	116017

Table B-9. MS2 ion series for Y2A: precursor @ 1445 m/z (pSer2 State)

		Y2A: precursor @ 1445 m/z (pSer2 State)					
		b			b-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	1	---			---		
P	2	185.0921	185	1088	---		
S	3	272.1241	272.027	611	---		
Y	4	435.1874	435.047	899	---		
pS	5	602.1858			504.2089		
P	6	699.2385			601.2617		
T	7	800.2862			702.3093		
S	8	887.3183	887.158	10234	789.3414	789.209	38281
P	9	984.371			886.3941		
S	10	1071.403	1071.285	4427	973.4262		
A	11	1142.44			1044.463		
S	12	1229.472	1229.397	57854	1131.495		
P	13	1326.525	1326.547	1861	1228.548		
T	14	---			---		
		y			y-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	14	---			---		
P	13	1358.551			1260.574		
S	12	1261.498	1261.433	11444	1163.522		
Y	11	1174.466	1174.364	7515	1076.49		
pS	10	1011.403	1011.291	44615	913.4262		
P	9	844.4047	844.28	9106	---		
T	8	747.3519			---		
S	7	646.3042	646.153	1905	---		
P	6	559.2722	559.12	24005	---		
S	5	462.2195			---		
A	4	375.1874			---		
S	3	304.1503	304.051	700	---		
P	2	217.1183	217.046	714	---		
T	1	120.0655			---		

	Theoretical	Observed	Intensity

MH	1445.583	1445.591	41266
MH-H$_3$PO$_4$	1347.606	1347.593	102530

Table B-10. MS2 ion series for Y2A: precursor @ 1445 m/z (pSer5 State)

		Y2A: precursor @ 1445 m/z (pSer5 State)					
		b			b-H$_3$PO$_4$		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	1	---			---		
P	2	185.0921	185	1088	---		
S	3	272.1241	272.027	611	---		
Y	4	435.1874	435.047	899	---		
S	5	522.2195	522.111	3931	---		
P	6	619.2722	619.144	1569	---		
T	7	720.3199	720.164	8678	---		
pS	8	887.3183	887.158	10234	789.3414	789.209	38281
P	9	984.371			886.3941		
S	10	1071.403	1071.285	4427	973.4262		
A	11	1142.44			1044.463		
S	12	1229.472	1229.397	57854	1131.495		
P	13	1326.525	1326.547	1861	1228.548		
T	14	---			---		
		y			y-H$_3$PO$_4$		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	14	---			---		
P	13	1358.551			1260.574		
S	12	1261.498	1261.433	11444	1163.522		
Y	11	1174.466	1174.364	7515	1076.49		
S	10	1011.403	1011.291	44615	913.4262		
P	9	924.371	924.251	88309	826.3941	826.27	9487
T	8	827.3183			729.3414		
pS	7	726.2706	726.114	8316	628.2937	628.162	4585
P	6	559.2722	559.12	24005	---		
S	5	462.2195			---		
A	4	375.1874			---		
S	3	304.1503	304.051	700	---		
P	2	217.1183	217.046	714	---		
T	1	120.0655			---		

	Theoretical	Observed	Intensity
MH	1445.583	1445.591	41266

MH-H_3PO_4	1347.606	1347.593	102530

Table B-11. MS2 ion series for Y12A: precursor @ 1353 m/z (pSer2 State)

		Y12A: precursor @ 1353 m/z (pSer2 State)					
		b			b-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	1	---			---		
P	2	185.0921	185.045	530	---		
S	3	272.1241	272.024	1015	---		
A	4	343.1612	343.059	5007	---		
pS	5	510.1596	510.007	6487	412.1827		
P	6	607.2123			509.2354		
T	7	708.26	708.082	20696	610.2831		
S	8	795.292	795.114	71455	697.3151	697.15	55287
P	9	892.3448	892.24	3810	794.3679		
S	10	979.3768	979.236	15502	881.3999	881.257	5128
A	11	1050.414	1050.259	63775	952.4371	952.314	25450
S	12	1137.446	1137.365	111725	1039.469	1039.369	69679
P	13	1234.499	1234.452	5676	1136.522		
T	14	---			---		
		y			y-H_3PO_4		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	14	---			---		
P	13	1266.525			1168.548		
S	12	1169.472	1169.41	22234	1071.495	1071.407	5326
A	11	1082.44	1082.334	14412	984.4633	984.35	5010
pS	10	1011.403	1011.285	52610	913.4262	913.293	13502
P	9	844.4047	844.248	97699	---		
T	8	747.3519			---		
S	7	646.3042	646.12	27299	---		
P	6	559.2722	559.093	89800	---		
S	5	462.2195			---		
A	4	375.1874			---		
S	3	304.1503	304.031	1415	---		
P	2	217.1183	217.052	2651	---		
T	1	120.0655			---		

	Theoretical	Observed	Intensity
MH	1353.557	1353.561	74204
MH-H_3PO_4	1255.58	1255.575	219498

Table B-12. MS2 ion series for Y12A: precursor @ 1353 m/z (pSer5 State)

		Y12A: precursor @ 1353 m/z (pSer5 State)					
		b			$b-H_3PO_4$		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	1	---			---		
P	2	185.0921			---		
S	3	272.1241			---		
A	4	343.1612			---		
S	5	430.1932	430.043	7509	---		
P	6	527.246	527.052	1969	---		
T	7	628.2937	628.122	29558	---		
pS	8	795.292	795.114	71455	697.3151	697.15	55287
P	9	892.3448	892.24	3810	794.3679		
S	10	979.3768	979.236	15502	881.3999	881.257	5128
A	11	1050.414	1050.259	63775	952.4371	952.314	25450
S	12	1137.446	1137.365	111725	1039.469	1039.369	69679
P	13	1234.499	1234.452	5676	1136.522		
T	14	---			---		
		y			$y-H_3PO_4$		
		Theoretical	Observed	Intensity	Theoretical	Observed	Intensity
S	14	---			---		
P	13	1266.525			1168.548		
S	12	1169.472	1169.41	22234	1071.495	1071.407	5326
A	11	1082.44	1082.334	14412	984.4633	984.35	5010
S	10	1011.403	1011.285	52610	913.4262	913.293	13502
P	9	924.371	924.23	108146	826.3941	826.25	27931
T	8	827.3183			729.3414		
pS	7	726.2706	726.1	28880	628.2937		
P	6	559.2722	559.093	89800	---		
S	5	462.2195			---		
A	4	375.1874			---		
S	3	304.1503			---		
P	2	217.1183			---		
T	1	120.0655			---		

	Theoretical	Observed	Intensity
MH	1353.557	1353.561	74204
$MH-H_3PO_4$	1255.58	1255.575	219498

VITA

Eric Bryant Gibbs

Education

Doctor of Philosophy **2012 – 2017**
The Pennsylvania State University
Department of Chemistry
University Park, PA
Dissertation Advisor: Scott A. Showalter

Bachelor of Arts **2005-2012**
Temple University
Department of Chemistry
Philadelphia, PA
Research Advisors: Daniel Strongin, Jaqueline Tanaka

Publications

1. Gibbs, E.B., Cook, E.C., Showalter, S.A. Application of NMR to studies of intrinsically disordered proteins. (Submitted)
2. Gibbs, E.B., Lu, F., Portz, B., Fisher, M.J., Laremore, T.N., Gilmour, D.S., & Showalter, S.A. 2016. Phosphorylation Induces Sequence-Specific Conformational Switches in the RNA Polymerase II CTD. (Submitted)
3. Portz, B., Lu, F., Gibbs, E.B., Showalter, S.A., & Gilmour, D.S. Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. (Submitted)
4. Gibbs, E. B., Showalter, S.A., 2016. Quantification of Compactness and Local Order in the Ensemble of the Intrinsically Disordered Protein FCP1. *J. Phys. Chem. B.* 120(34), 8960-8969
5. Gibbs, E.B., Showalter, S.A., 2015. Quantitative Biophysical Characterization of Intrinsically Disordered Proteins. *Biochemistry* 54, 1314–1326.
6. Bastidas, M., Gibbs, E.B., Sahu, D., Showalter, S.A., 2015. A primer for carbon-detected NMR applications to intrinsically disordered proteins in solution. *Concepts Magn. Reson*