
# IMPROVED CRYSTALLOGRAPHIC METHODS FOR RNA POLYMERASE II COMPLEXES USING INTRINSIC ANOMALOUS SCATTERING

by Peter Adrian Meyer

---

# IMPROVED CRYSTALLOGRAPHIC METHODS FOR RNA POLYMERASE II COMPLEXES USING INTRINSIC ANOMALOUS SCATTERING

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Peter Adrian Meyer
August 2010

# IMPROVED CRYSTALLOGRAPHIC METHODS FOR RNA POLYMERASE II COMPLEXES USING INTRINSIC ANOMALOUS SCATTERING

Peter Adrian Meyer, Ph. D.

Cornell University 2010

RNA Polymerase II (Pol II) is the central enzyme in eukaryotic mRNA transcription. In recent years, structural studies have provided insights into the mechanism of Pol II. A full understanding of the mechanisms of transcription will require structural insights into the numerous complexes of Pol II with transcription factors involved in in vivo transcription. However, these structural studies are hampered by the limited resolution produced by crystals of such complexes. Improved crystallographic methods designed to enhance structural data obtainable from such complexes have been developed taking advantage of the anomalous scattering due to intrinsic zinc ions in Pol II. These approaches have been validated using the known structure of 12 subunit Pol II, and were able to provide additional insights into its structure.

## BIOGRAPHICAL SKETCH

Peter Meyer was raised in Allendale, New Jersey where he attended Northern Highlands Regional High School, graduating in 1997. Throughout High School, he received an indirect introduction to biological research, on the applied side, while working at the Allendale Animal Hospital. Peter then attended the University of Chicago, graduating in 2001 with a Bachelor of Arts in Biological Sciences. During his time in Chicago, he worked in the laboratory of Anna DiRienzo assisting with research on human genetics. After a summer working as a canoe guide at Floodwood Mountain Reservation, in the Adirondacks, he arrived at Cornell University to start graduate studies. In summer of 2007, he followed Jianhua Fu to the Medical College of Wisconsin.

# ACKNOWLEDGMENTS

Events rarely have single causes, and this work is no exception. I would like to thank Jianhua Fu for his guidance and support in learning about crystallography, science, research and life in general. I would also like to express my gratitude to my committee members for their help and support; especially to Jeff Roberts for agreeing to serve as co-chair.

I'd like to thank all of the members of the Fu lab, Ping, Min-Cheng, Ajit and Man-Hee, helped make it a great place to learn and do research. I would also like to thank all of the members of the DiRienzo lab, especially Gustavo, Rocky, Angelika, Linda and Martha, where I got my first experience in scientific research as an under graduate.

My friends in Ithaca, especially Elysa, Patrick, Thi, Tom and Man-Hee, helped make Cornell a good place to live as well as work.

Most of all, I'd like to thank my family, especially my mother and sister, for more than can be said.

I would like to thank the National Institute of Health for partial financial support, through the Pre-Doctoral Training in Cellular and Molecular Biology grant GM07273, as well as National Institute of Health grant GM064651 to Jianhua Fu. Use of the National Synchrotron Light Source, Brookhaven National Laboratory, was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-98CH10886. This work is based upon research conducted at the Cornell High Energy Synchrotron Source (CHESS), which is supported by the National Science Foundation and the National Institutes of Health/National Institute of General Medical Sciences under NSF award DMR-0225180, using the Macromolecular Diffraction at CHESS (MacCHESS) facility, which is supported by award RR-01646 from the National Institutes of Health, through its

National Center for Research Resources. Use of the Advanced Photon Source at Argonne National Laboratory was supported by the U. S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1: INTRODUCTION

The research described in this work primarily concerns improvements to X-ray crystallographic data processing techniques necessitated by the problems posed by crystals of RNA Polymerase II complexes. This necessitates an introduction covering two disparate areas: a description of the role of RNA Polymerase II in biological processes to indicate the biological relevance of these crystallographic problems; and an overview of macro-molecular X-ray crystallography, in order to illustrate how these problems arose and how they were resolved.

## 1.1 The Need of Crystallographic Methods Improvement for the Analysis of RNA Polymerase II Complexes

Many structural studies have been reported on cellular RNA Polymerases, mainly but not exclusively on archael polymerase and *Saccharomyces cerevisiae* RNA Polymerase II (Pol II). X-ray structures of Pol II that correspond to several stages of the transcription reaction have been determined. However, despite the large number of transcription factors interacting with Pol II during the transcription process, few structures of Pol II with transcription factors are available. As of June 2009, X-ray structures of one TFIIB-Pol II (Bushnell et al., 2004) complex and several TFIIS-Pol II (Kettenberger et al., 2003; Wang et al., 2009) complexes have been published. An EM surface of the TFIIF-Pol II complex has also been determined (Chung et al., 2003).

The comparatively few transcription factor complex structures available do not indicate a lack of biological interest. Instead, this reflects the difficulty of determining such structures. The large asymmetrical protein complexes formed by Pol II yield crystals capable of providing only low-resolution diffraction data. The resolution limit of these

crystals give rise to difficulties in several fundamental steps of the crystallographic process: in particular, phase determination and structure refinement. The research described in this work is primarily focused on methods to reduce the difficulty of determining structures of complexes of RNA Polymerase II.

## 1.2 Biological Function of RNA Polymerase II

### 1.2.1 Overview of RNA Polymerase

RNA polymerases are the core enzymes responsible for the synthesis of RNA from sequences encoded in template DNAs. Cellular RNA polymerases are large multi-subunit proteins, ranging in composition and size from as small as 4 subunits and approximately 400 KDa in prokaryotes and as large as 17 subunits and approximately 690 KDa in eukaryotes. There are three main classes of eukaryotic RNA Polymerase (Pol), known as Pol I, Pol II and Pol III. Pol I is responsible for the transcription of ribosomal RNAs. Pol II transcribes precursors of messenger RNA (mRNA) and a variety of non-coding RNA moieties, including micro-RNAs and anti-sense RNAs. Pol III synthesizes transfer RNAs, 5S rRNA and U6 snRNA.

Ensuring appropriate and accurate production of pre-mRNAs is a key component of gene regulation. Transcription by Pol II is an integration point for the activities of numerous regulatory factors, and occurs in concert with the processing pathways necessary for the generation of mature mRNAs from raw transcripts.

### 1.2.2 General Transcription Factors

In addition to the numerous regulatory and processing factors required for *in vivo* transcription, Pol II requires the assistance of five additional protein factors to transcribe pre-mRNA from template DNA: TFIIF, TFIID, TFIIB, TFIIE and TFIIH

(Orphanides et al., 1996). The activities of these factors are required at different points in transcription, and some transcription factors play multiple roles in transcription. The conventional model for the assembly of Pol II and the basal transcription factors is one of ordered stepwise assembly in which a growing pre-initiation complex (PIC) accumulates additional transcription factors (Figure 1.1). The first step in this assembly process is the binding of the TATA-box binding protein (TBP) and TBP-associated factor (TAF) components of TFIID to the promoter region of the target gene. This is followed by binding of TFIIB, which stabilizes the TFIID-DNA interaction and assists with start site selection, forming a BD complex (Thomas and Chiang, 2006). The asymmetrical nature of TFIIB binding to the TFIID-DNA complex provides a basis for downstream directionality of transcription (Orphanides et al., 1996). Separately, TFIIF binds to Pol II, forming the Pol-F complex, which is recruited to the BD complex to form the DB-Pol-F complex. TFIIE, which assists with promoter melting, is the next factor recruited, resulting in the DB-Pol-EF complex. TFIIE also assists in recruiting the remaining basal transcription factor, TFIIH which provides the kinase and DNA helicase activities essential for transcription initiation. TFIIA, which is required for activated transcription, may also play a role in basal transcription. An alternative model for PIC formation was suggested by observations that Pol II complexes containing several general transcription factors could be purified from cells (Orphanides et al., 1996; Thomas and Chiang, 2006). In this model, a pre-assembled holoenzyme complex binds in a single step to the template DNA.

Post-translational modifications of several components of the Pol II machinery play significant roles throughout the transcription cycle. TFIIB auto-acetylates, which enhances its ability to recruit TFIIF-Pol II to the DB complex (Sims et al., 2004). The RAP74 subunit of human TFIIF is phosphorylated (Orphanides et al., 1996), although

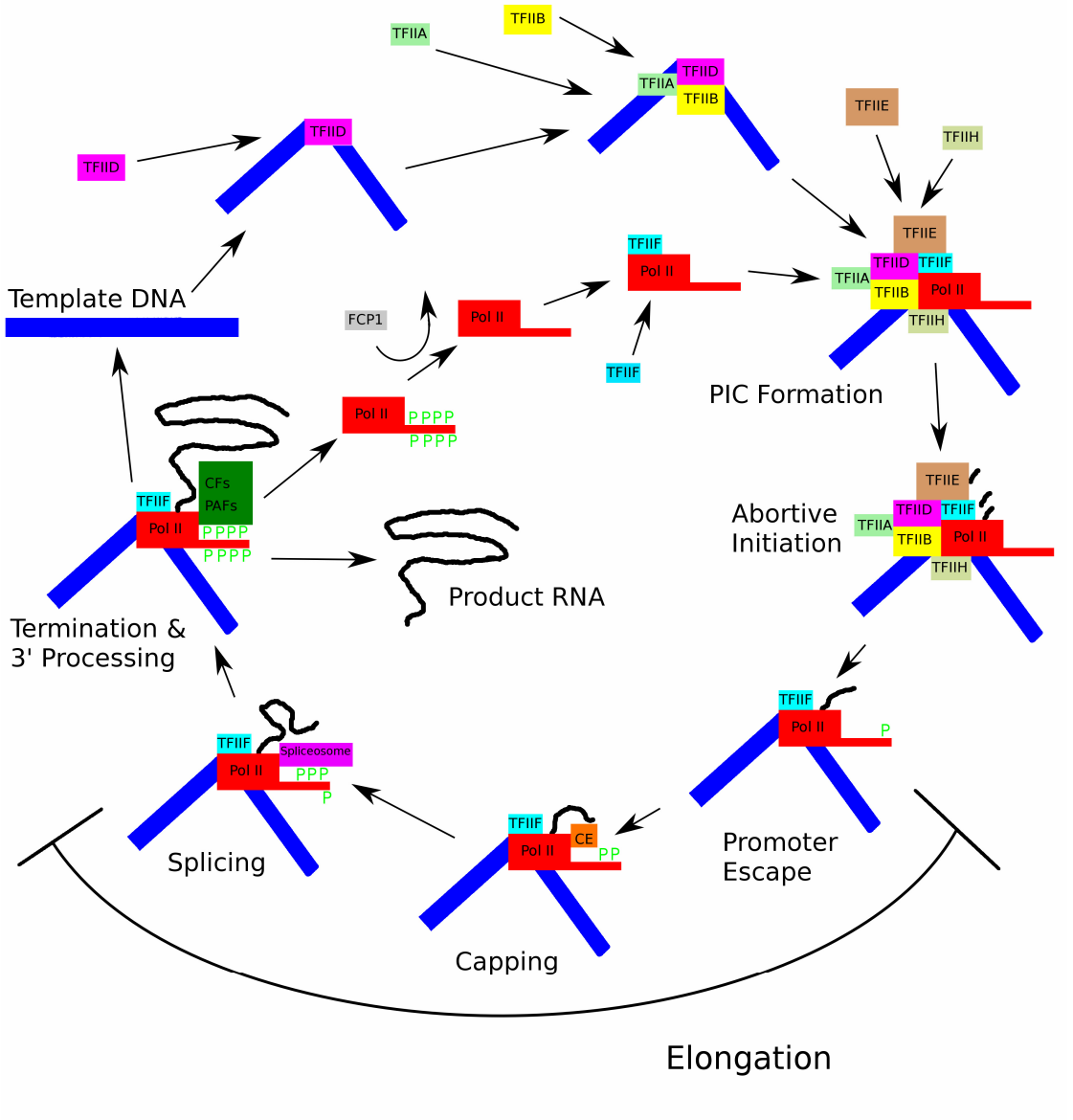


Figure 1.1: Pol II Complexes Formed During the Transcription Cycle

Pol II with phosphorylated CTD is illustrated by the presence of *green P*s on the tail of Pol II. DNA is shown as *blue bar*, RNA as *black curved lines*. Transcriptional initiation is illustrated following the stepwise-assembly model.

the functional role of these modifications is currently unclear. Rpb1, the largest subunit of Pol II, undergoes cycles of phosphorylation and de-phosphorylation during transcription; these modifications, discussed in more detail below, occur primarily in the carboxy-terminal domain (CTD) of this subunit, which consists of a series of conserved hepta-peptide repeats with consensus sequence $Y_1S_2P_3T_4S_5P_6S_7$ (Allison et al., 1988; Nonet et al., 1987; Saunders et al., 2006).

### 1.2.3 The Basal Transcription Cycle

Transcription does not proceed monotonously from PIC assembly through initiation and processive elongation to termination. Instead, transcription progresses through discrete stages of open complex formation, promoter escape, promoter-proximal pausing, productive elongation, and termination. The transitions between these stages require assistance from specific transcription factors, and are targeted by various regulatory factors.

The first step following the assembly of the PIC at the promoter is the separation of the template and non-template DNA strands into a structure known as the transcription bubble. The ATP-dependant helicase activity of TFIIH, enhanced by TFIIE, is required for DNA melting and bubble formation (Timmers, 1994). Upon melting, the template strand must be positioned such that the correct start site is located in the Pol II active site. Start site selection involves the B-finger domain of TFIIB (Li et al., 1994), and the Rpb9 subunit of Pol II (Sun et al., 1996). Following start site positioning, initiation of transcription may occur. Transcription of each nucleotide requires three distinct biochemical steps: nucleotide selection, catalysis, and translocation. In nucleotide selection, a ribonucleotide triphosphate enters the nucleotide addition site and base pairs with the template DNA; mismatched ribonucleotides are discriminated against in this

step. The second step is the formation of a phosphodiester bond between the product RNA and incoming ribonucleotide. Finally, the translocation step clears the active site for the next nucleotide by shifting the DNA-RNA hybrid by one base pair concurrent with the unwinding of one base pair at the downstream end of the hybrid and a rewinding of the template and non-template DNA strands at the upstream edge of the transcription bubble.

During early elongation, a series of short transcripts are often created and released in a process known as abortive elongation. The CTD, which is hypo-phosphorylated during PIC formation and initiation, becomes phosphorylated by the CDK7 subunit of TFIIH (Thomas and Chiang, 2006). Abortive elongation continues until the transcript reaches a length of 25-30 nucleotides (Pal and Luse, 2003). At this point, Pol II, accompanied by TFIIF and other elongation factors, escapes from the initiation scaffold and enters into the productive elongation stage. This transition from abortive initiation to processive elongation is highly regulated, and the biochemical mechanisms involved are currently being investigated. Similarly, the termination step of transcription is another highly regulated step. Although the many details remain to be determined, it is known that termination is coupled to events involved in 3'-end RNA processing (Howe, 2002; Richard and Manley, 2009).

### 1.2.4 Co-transcriptional RNA Processing

The production of mature messenger RNA and other Pol II transcripts requires several processing steps in addition to transcription: 5' cap addition, intron splicing, and 3' poly-A tail addition (Figure 1.1). The transcripts produced by Pol I and Pol III are not processed in this manner, implying that enzymes involved in pre-mRNA processing have a mechanism for distinguishing transcripts on the basis of the polymerase

complexes that produced them. This is currently believed to be accomplished by coupling of pre-mRNA processing reactions to Pol II transcription (Howe, 2002).

The first mRNA processing step is the formation of a 5' cap structure, which is required for efficient splicing, nuclear export, and mRNA stability (Proudfoot et al., 2002). Three enzymatic activities are required for cap formation: an RNA triphosphatase, a guanylyl-transferase and a methyltransferase. The enzymes responsible for the first two activities form a hetero-dimer in yeast (abbreviated CE), and are expressed as a single polypeptide in mammals (Ho et al., 1998). The methyltransferase activity is provided by a separate protein (Abd1 in *S. cerevisiae*), which functions separately from CE (Schroeder et al., 2004). Cap addition occurs only after 20-30 bases of RNA have been transcribed (Rasmussen and Lis, 1993) and may be correlated with the transition from abortive to productive elongation occurring at this point (Pal and Luse, 2003). Phosphorylation of the Pol II Rpb1 CTD, discussed further below, in required for efficient capping *in vivo* and enhances the activity of CE (Cho et al., 1997; Ho et al., 1998; McCracken et al., 1997; Yue et al., 1997).

Intron splicing is another processing reaction that is required for mRNA maturation, and occurs co-transcriptionally (Proudfoot, 2000). 3' RNA processing is also coupled to transcriptional termination, and the poly-A tail is required for the termination and release steps in this process (Zorio and Bentley, 2004).

One common theme of the various co-transcriptional RNA processing reactions is that each event requires interaction with Pol II and formation of a complex with Pol II, in order to insure appropriate functionality. The information regarding the structural characteristics of these complexes, as well as the regulated transcribing Pol II molecules

which are the substrate for their formation, could greatly assist the understanding of the biochemical mechanisms that underlie co-transcriptional RNA processing. A robust structural characterization of the core polymerase will be essential for the investigation of the higher-order complexes involved in these events.

### 1.2.5 Structural Organization of Pol II

A significant amount of structural data is available for RNA polymerase. The vast majority of this data has been provided by X-ray crystallography. The size of the protein currently precludes structure determination by nuclear magnetic resonance (NMR). Electron microscopy (EM) has been used in cases where X-ray structures are not available, but produces lower resolution information (up to 18 Å for Pol II). The available structures of cellular prokaryotic and eukaryotic RNA Polymerases have provided a wealth of information regarding the basic mechanism of transcription, as well as illustrating commonalities between the two enzyme systems.

Pol II is composed of 12 different subunits, named Rpb1 to Rpb12, with a total molecular mass of 511 KDa (Table 1.1 and Figure 1.2). Pol II shares five subunits (Rpb5, Rpb6, Rpb8, Rpb10 and Rpb12) with Pol I and Pol III (Woychik and Young, 1990; Woychik et al., 1990). The two largest Pol II subunits, Rpb1 and Rpb2, contribute approximately 65% of the total mass of the protein. Five of these subunits are homologous to prokaryotic polymerase subunits, at the sequence and structure levels. Among these, Rpb1 is homologous to the prokaryotic β' subunit, and Rpb2 is homologous to the prokaryotic β. There are two copies of the α subunit in prokaryotic RNA Polymerase, which are homologous to eukaryotic Pol II subunits Rpb3 and Rpb11 (Ebright, 2000; Sweetser et al., 1987). The ω subunit of the prokaryotic enzyme shows limited homology to Rpb6 of Pol II, based on sequence and the quaternary structures

Table 1.1: *Saccharomyces cerevisiae* RNA Polymerase II Subunit Composition

| | Mass (KDa) | Number of Residues | Mass Percentage |
|---|---|---|---|
| Intact Pol II | 511.7 | 4525 | 100.00% |
| | | | |
| Rpb1 | 191.0 | 1733 | 37.33% |
| Rpb2 | 138.0 | 1224 | 26.97% |
| Rpb3 | 35.0 | 318 | 6.84% |
| Rpb4 | 25.4 | 221 | 4.96% |
| Rpb5 | 25.0 | 215 | 4.89% |
| Rpb6 | 17.9 | 115 | 3.50% |
| Rpb7 | 19.0 | 171 | 3.71% |
| Rpb8 | 16.5 | 146 | 3.22% |
| Rpb9 | 14.3 | 122 | 2.79% |
| Rpb10 | 8.3 | 70 | 1.62% |
| Rpb11 | 13.6 | 120 | 2.66% |
| Rpb12 | 7.7 | 70 | 1.50% |

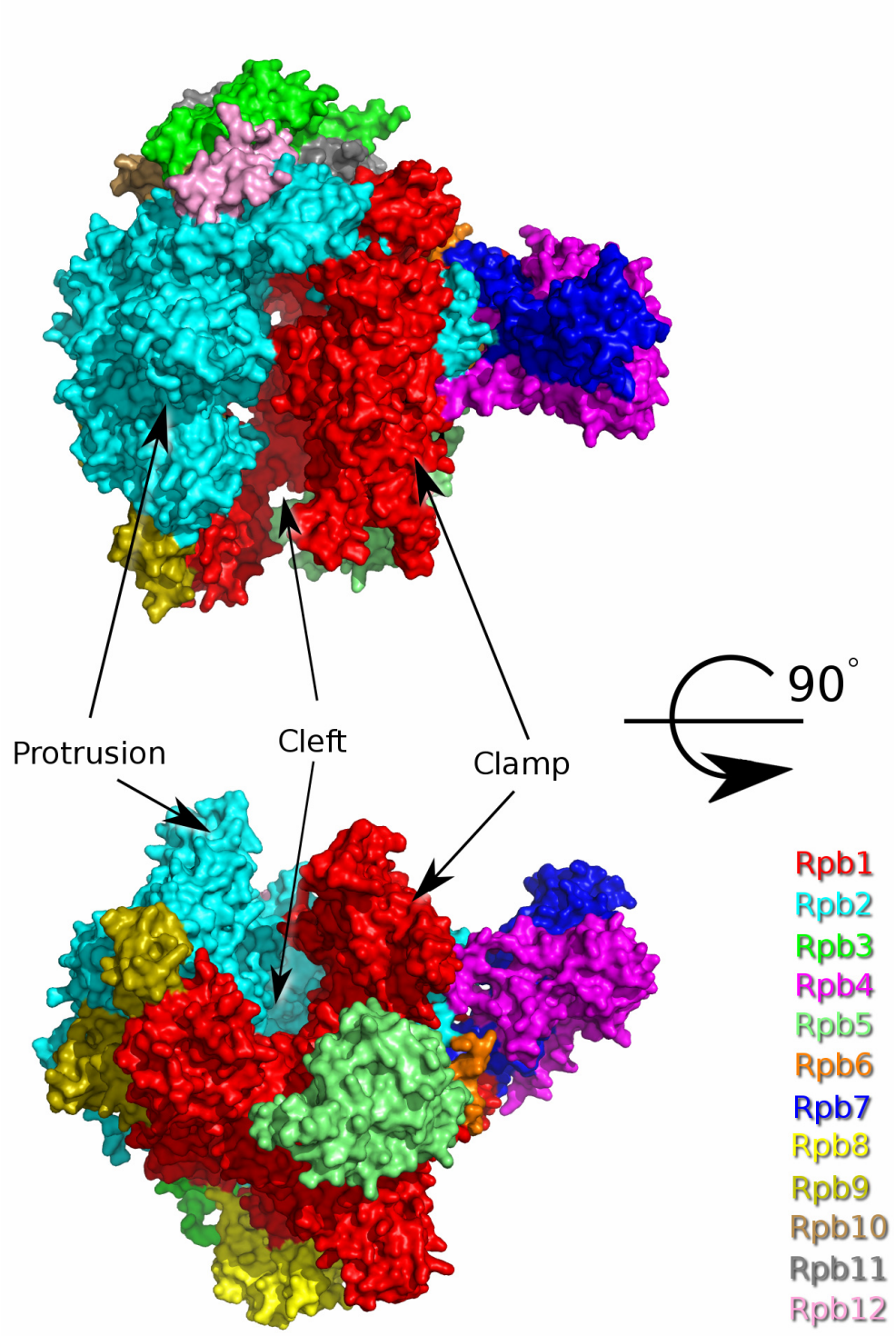


Figure 1.2: RNA Polymerase II Subunit Architecture

The locations of Pol II subunits are indicated by color coding, as shown in insert.

(Minakhin et al., 2001) (Figure 1.3). Two Pol II subunits, Rpb4 and Rpb7, are required for promoter dependent initiation and form a sub-complex dissociable from the main body of the polymerase (Edwards et al., 1991). The core Pol II, without Rpb4/7, is able to elongate RNA *in vitro* when provided with an appropriate substrate (Gnatt et al., 1997). The original Pol II structures were determined using crystals of the 10-subunit form (Cramer et al., 2001; Fu et al., 1999; Gnatt et al., 1997; Gnatt et al., 2001), and the highest resolution data for Pol II to date also comes from these crystals (Westover et al., 2004).

The most prominent feature of the Pol II structure is a large cleft in the center of the molecule (Figures 1.4 and 1.5). The polymerase active site, marked by two catalytic magnesium ions, is located at the base of this cleft. In addition to the magnesium ions, the tertiary structure is stabilized by eight non-catalytic zinc ions; however the C-terminal zinc site of Rpb9 has been implicated in start site selection (Hull et al., 1995). The internal face of the cleft is mainly formed by Rpb1 and Rpb2 (Figures 1.2 and 1.4). A domain referred to as the Clamp (Figure 1.5) forms one side of this cleft (Cramer et al., 2000), and has been observed in three conformational states, differing primarily in the distance from the clamp to the opposite side of the cleft. The most widely 'open' conformation has only been observed in 10-subunit Pol II structures in the absence of nucleic acids (Bushnell et al., 2002; Cramer et al., 2001; Kaplan et al., 2008; Wang et al., 2006; Westover et al., 2004). The presence of DNA/RNA, or presence of the Rpb4/7 sub-complex, is sufficient to shift the clamp into a more closed conformation. A 'collapsed' conformation has been observed only in the initial EM structures, with even less space available inside the channel (Craighead et al., 2002; Darst et al., 1991). The Protrusion and Stalk domains form the opposite side of the cleft. In 12-subunit

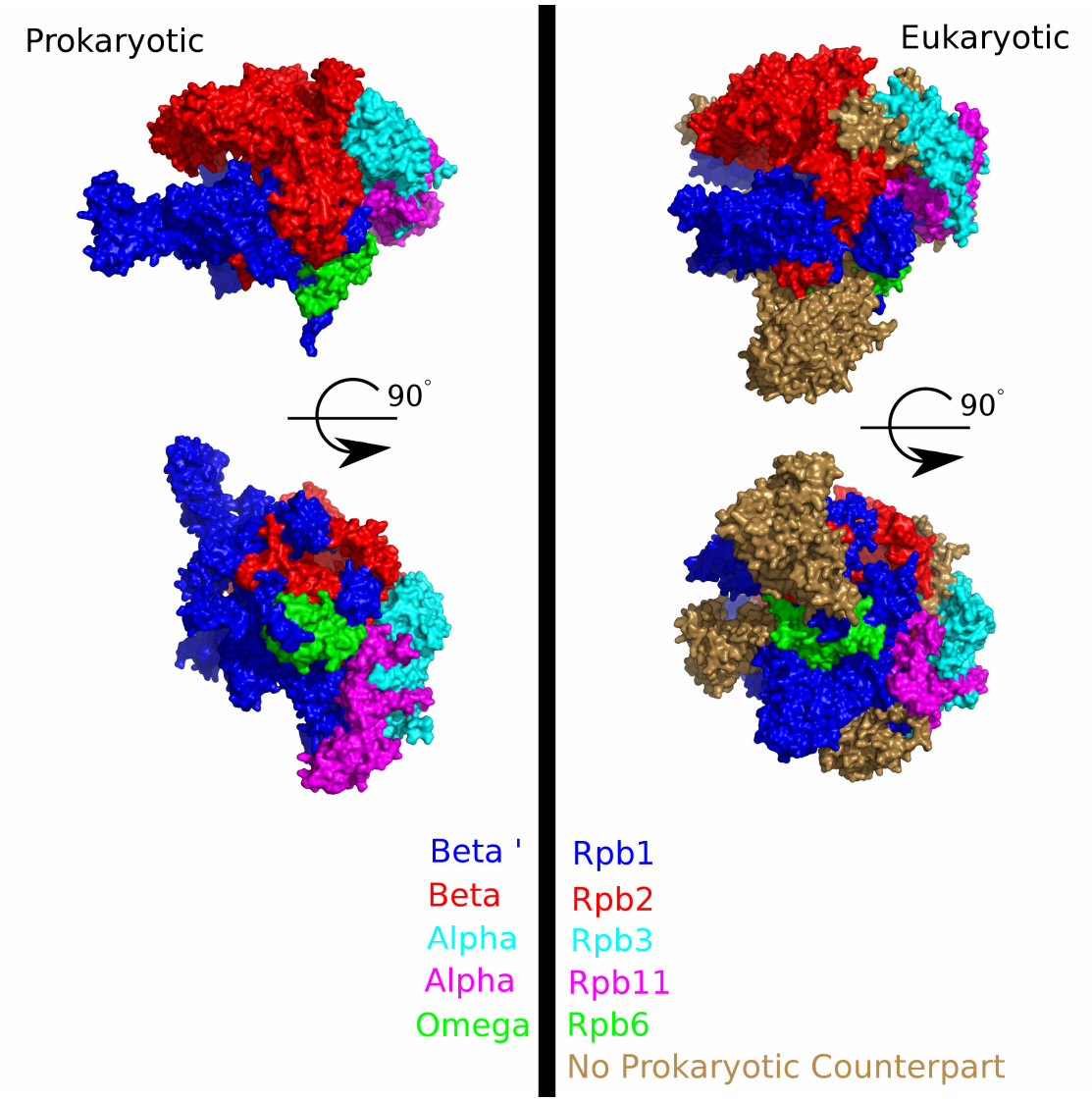


Figure 1.3: Structural Homology between Eukaryotic and Prokaryotic RNA

Polymerases

Conserved subunits are colored by homology, as indicated by pairs of names near center line. Eukaryotic subunits with no prokaryotic counterparts are in *sand*. Prokaryotic model: PDB ID 2PPB (Vassylyev et al., 2007), Eukaryotic model: PDB ID 3FKI (Meyer et al., 2009).

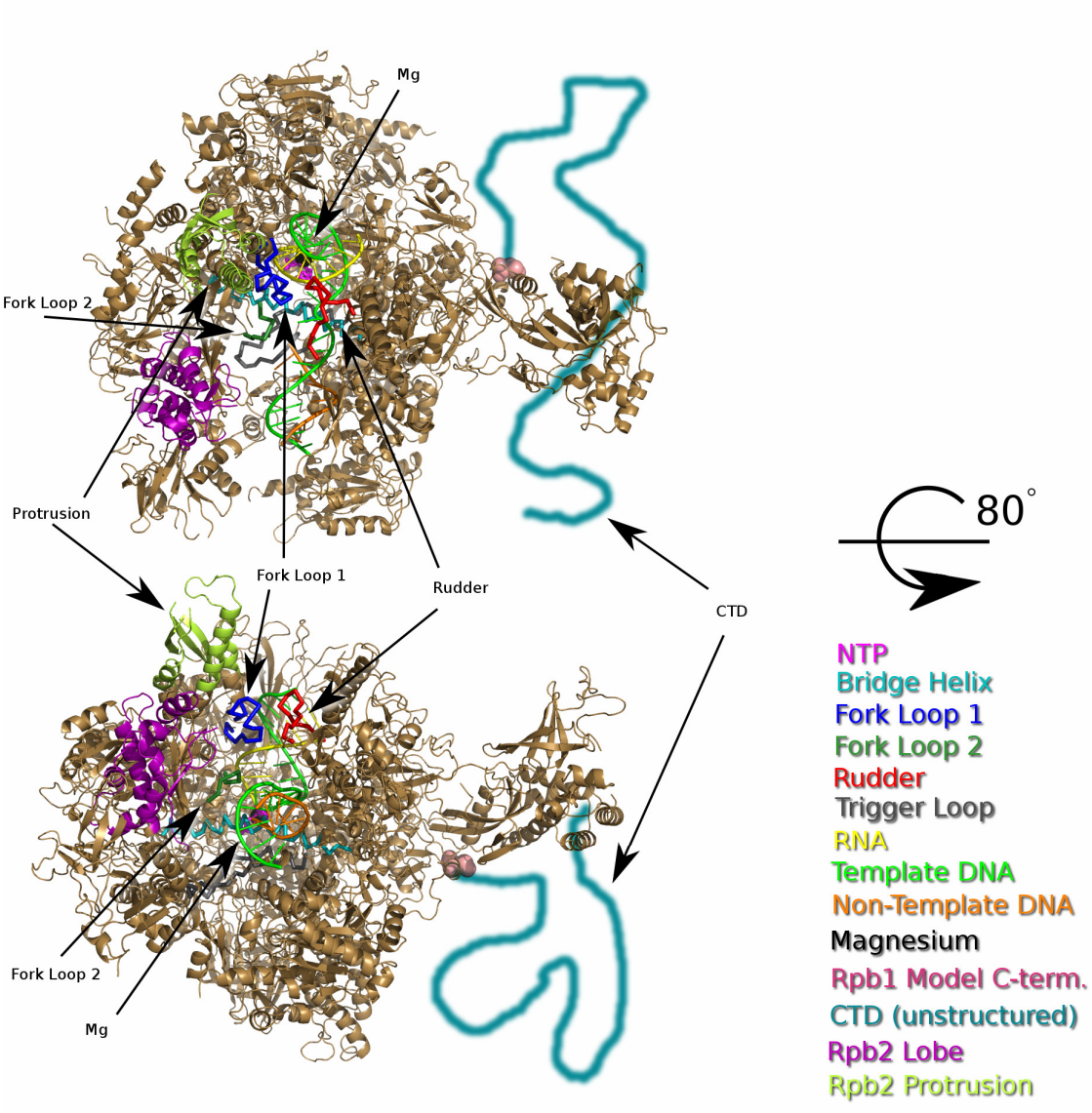


Figure 1.4: Structural Features of RNA Polymerase II

Notable structural features of Pol II, colored following inset.

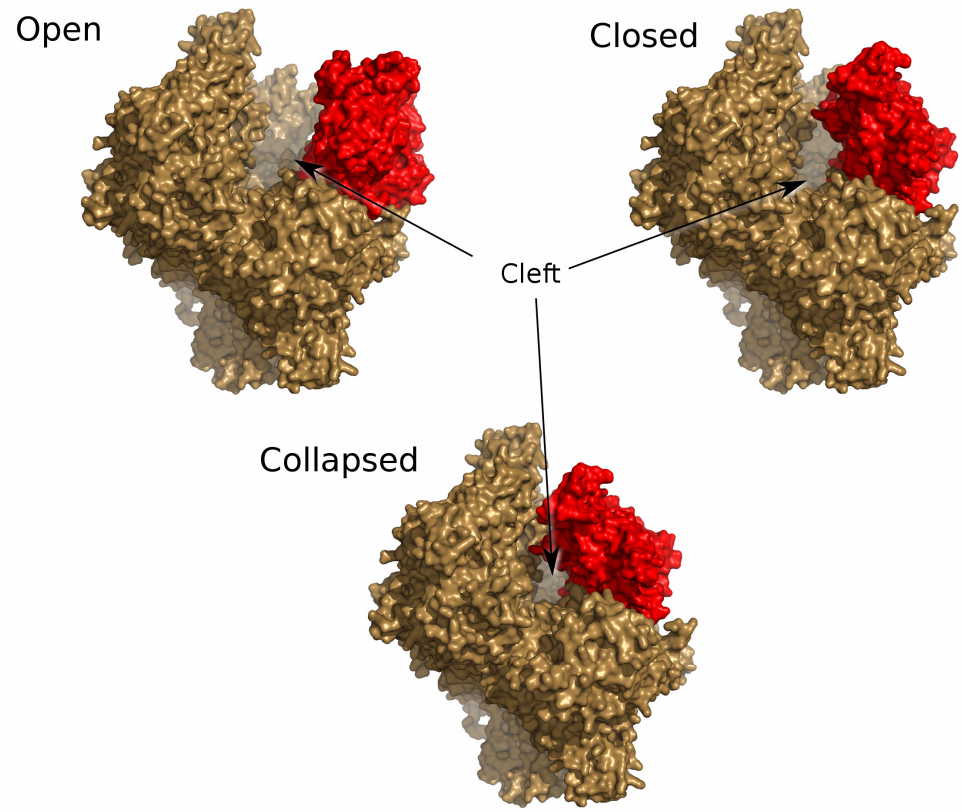


Figure 1.5: Conformations of the Clamp Domain

Clamp domain is shown in *red* throughout.
Open conformation from PDB ID 1I3Q (Cramer et al., 2001);
Closed conformation from PDB ID 3FKI (Meyer et al., 2009) with Rpb4/7 hidden;
Collapsed conformation modeled on basis of EM surfaces (Craighead et al., 2002; Darst et al., 1991).

structures, the Rpb4/7 sub-complex projects at an angle away from the Clamp side of the polymerase.

As shown in several structures of Pol II in complex with DNA and RNA, the DNA enters through a large open channel along the base of the cleft, with the template and non-template strands separating approximately two thirds of the way into this channel (Gnatt et al., 2001; Westover et al., 2004). A channel between Rpb1 and Rpb2 is well positioned to be the path of the nascent RNA transcript, after it separates from the DNA/RNA hybrid, and has been named the RNA exit channel (Cramer et al., 2000). This channel could provide an explanation for the extent of RNA protection observed biochemically (Gu et al., 1996). An opening below the active site is referred to as the secondary channel, and is thought to allow for the entry of nucleotide triphosphates and may also accommodate the transcript when backtracking, or reverse translocation, occurs (Wang et al., 2009).

Several regions within the cleft, referred to using the nomenclature originating from (Cramer et al., 2000) and (Cramer et al., 2001), have been implicated for specific activities in the transcription reaction (Figure 1.4). The Bridge helix, Rpb1 812-847, is conserved in archaeal and eukaryotic RNA polymerases, and has been implicated in translocation (Bar-Nahum et al., 2005; Brueckner and Cramer, 2008; Westover et al., 2004). In all eukaryotic RNA polymerase structures to date, the Bridge helix observed to be straight. However, in prokaryotic structures, the Bridge helix is seen as either straight (Temiakov et al., 2005) or kinked (Zhang et al., 1999) at residues corresponding to Rpb1 831 and 832 (Gnatt et al., 2001). The bending of the bridge helix is thought to play a role in translocation of template DNA (Gnatt et al., 2001). The downstream

region of template DNA in the transcription bubble crosses over the bridge helix. Above the template DNA, Fork Loop 1 (Rpb2 462-481) (FL1) makes contacts with a loop structure named Rudder (Rpb1 310-324); these two regions may contribute to preventing collapse of the transcription bubble and stabilization of the hybrid. An additional loop structure, Fork Loop 2 (Rpb2 502-509) (FL2) is found near the site where the double-stranded DNA separates into template and non-template strands. As will be discussed further in Chapter 3, FL2 is a mobile element. However, it is in position to sterically block the approach of the non-template strand towards the active site, as well as to hydrogen bond with the template strand (Gnatt et al., 2001; Kettenberger et al., 2004; Wang et al., 2006). The Trigger Loop (Rpb1 1070-1101), is located below the Bridge helix, and has been implicated in the correct positioning of incoming NTPs in the active site through hydrogen bond interactions with incoming nucleotides and the Bridge helix (Wang et al., 2006). The interaction of the Trigger Loop with the Bridge helix has also been implicated in the translocation step (Bar-Nahum et al., 2005; Brueckner and Cramer, 2008; Kaplan and Kornberg, 2008).

The C-terminal domain (CTD) of Rpb1 (unstructured region in Figure 1.4) mentioned earlier is composed of tandem repeats of a seven amino acid motif. The number of repeats in the CTD varies with species: 26-27 in *Saccharomyces cerevisiae*, 52 in human, 34 in *Caenorhabditis elegans* (Corden et al., 1985; Orphanides et al., 1996). This repeat motif is a unique feature of eukaryotic Pol II's and has a consensus sequence of $Y_1S_2P_3T_4S_5P_6S_7$. Several of these residues are potential phosphorylation sites, and the serine residues are known to be phosphorylated and de-phosphorylated *in vivo* (Phatnani and Greenleaf, 2006; Zhang and Corden, 1991). The phosphorylation state of the CTD is highly regulated, and correlated with the state of Pol II in the transcription cycle (Saunders et al., 2006). In general, the CTD of Pol II unengaged with DNA is in

the hypo-phosphorylated state (abbreviated as CTDa). During initiation and early elongation, serine residues at position 5 (Ser5) are preferentially phosphorylated by the general transcription factor TFIIH. As elongation progresses, serine residues at position 2 (Ser2) becomes increasingly phosphorylated. By the time transcription reaches termination, the CTD is hyper-phosphorylated (abbreviated as CTDo) (Hirose and Ohkuma, 2007). Phosphorylation of Ser7 has been implicated in snRNA expression, although not in mRNA transcription (Egloff et al., 2007). The threonine at position 4 is also a potential target for phosphorylation (Wong et al., 2007), although no information on its relevance to transcription is available. The isomerisation state of the proline residues within the CTD has also been implicated in transcriptional regulation (Shaw, 2007). When the CTD has been removed (abbreviated CTDb), Pol II remains competent for *in vitro* transcription in purified reconstituted systems. However, the CTD is required for *in vitro* transcription from systems reconstituted from cell extracts (Li and Kornberg, 1994). The phosphorylation state of the CTD has also been implicated in transcriptional termination (Gudipati et al., 2008). In addition, a minimum of 8 to 10 copies of the concensus CTD repeat is required for cell viability in yeast (Nonet et al., 1987; West and Corden, 1995). In the context of the complete Pol II, CTD phosphorylation states are referred to as Pol IIa, Pol IIo, and Pol IIb, respectively.

The CTD is known to interact physically with several transcription factors, and synthetic CTD peptides have been used in co-crystals with three transcription factors. In each of these, the serine-phosphorylated CTD (Fabrega et al., 2003; Meinhart and Cramer, 2004; Verdecia et al., 2000; Zhang et al., 2006) was observed in a different conformation, supporting the idea that it functions as a flexible binding platform for various regulatory factors. In all published Pol II structures to date, the CTD has been

maintained in the hypo-phosphorylated state (Pol IIa). However, no electron density has been observed for this region, suggesting that CTDa is disordered in the context of Pol II.

Although much is known about the structural basis of the core enzymatic activity of Pol II, considerably less is known regarding the structural basis of regulated transcription *in vivo*. Transcription *in vivo* is a highly regulated process involving numerous macromolecular complexes anchored on the core Pol II molecule. Biochemical studies have identified many regulatory components and in many cases elucidated their functions. The understanding of the mechanism based on these transcription complexes would be greatly deepened by structural knowledge of the relevant complexes. However, there are many obstacles to the determination of these structures. In this work, I described adaptations and improvements to crystallographic techniques for addressing the problems associated with X-ray structure determination in the *S. cerevisiae* Pol II system.

## 1.3 Crystallography Introduction

The purpose of macromolecular X-ray crystallography is to determine the structure of a molecule of biological interest. Structural knowledge of a biological molecule can provide insights into catalytic functions and relationships between proteins that are not apparent at the sequence level. The process of structure determination by x-ray crystallography involves crystallization, diffraction data collection, phase determination, and building and refining a model of the molecule of interest. In order to explain the ways these stages fit together, they be explained by working backwards

from the goal, a model of the molecule of interest, towards the starting point, the collection of diffraction images.

### 1.3.1 Conceptual Background

A brief conceptual overview of some of the mathematical basis underlying x-ray crystallography can be helpful to provide the groundwork for understanding the steps involved in structure determination. A crystal is composed of an arrangement of atoms that repeat periodically in three dimensions. Any crystal can be decomposed into two parts: 1) the lattice, which describes the periodicity of the crystal; and 2) the unit-cell, which consists of the contents of the repeating unit. The observations in a crystallographic experiment are a collection of images representing two-dimensional slices of the three-dimensional diffraction pattern arising from the crystal. The Fourier Transform (FT) is used to relate this diffraction pattern to the electron density produced by the atoms present in the crystal; this electron density is responsible for the scattering of X-ray photons during the diffraction experiment.

The Fourier series of a function is a discrete set of wave functions. Each component wave function has three terms: an amplitude, a phase, and a unique frequency. Under some circumstances, it is clearer to represent these components as complex numbers, with a real part and imaginary part. With an infinite number of terms, any periodic function can be exactly reproduced by its corresponding Fourier series. The Fourier Transform of a function can be thought of as the Fourier series with the discrete sum replaced by the integral of a complex continuous function. The domain of the Fourier Transform is often referred to as reciprocal, or inverse, space. The electron density within the crystal can be thought of as the convolution of the periodic crystal lattice with the electron density of a single copy of the unit-cell. Equivalently, the Fourier

Transform of a crystal is the product of the respective Fourier Transforms of the crystal lattice and the contents of a unit-cell. The discrete nature of the lattice function implies that the transform of the crystal is also discrete, as illustrated by the regularly spaced spots observed in a diffraction pattern.

A macromolecular crystal can be represented by a periodic function, at least to a first approximation, due to the crystal lattice. Any point in the unit-cell is equivalent to all other points that are related by a translation of one or more unit-cell lengths. This translational equivalence is the most basic symmetry element found in any crystal. The electron density in a crystal lacks imaginary components, so all diffraction patterns also exhibit a center of symmetry at the origin of the diffraction pattern. This center of symmetry, known as Friedel's law, is inexact in the presence of anomalous scattering, as discussed below. In the majority of cases, biological crystals have additional symmetry elements in which related points within the unit-cell are equivalent: related by a rotation or translation of less than one unit-cell length, or both. Crystallographic symmetry elements repeat over the entire crystal lattice, and reduce the volume of unique space in the crystal to a region known as the asymmetric unit. Non-crystallographic symmetry (NCS) elements, which are not related to lattice symmetry, can also be present within the asymmetric unit. NCS often arises due to internal symmetry in a macromolecule, such as a virus capsid or homo-multimeric membrane ion channel. Due to the chiral nature, or handedness, of biological molecules, symmetry elements containing a mirroring operator can not occur in real space, although they may be present in reciprocal space. Centering operators, or lattice centering, representing translational shifts of less than one unit-cell without a rotational component, can also occur in biological crystals.

The collection of crystallographic symmetry operators comprises the space group of the crystal, and produces a variety of effects. The presence of a symmetry operator can restrict the allowed angles between unit-cell edges, and may place restrictions on the lengths of the unit-cell edges. Symmetry operators also exhibit their effects in reciprocal space. The allowed phase values for some crystallographic indices can be restricted to two values (separated by 180 degrees) as a consequence of symmetry. These reflections are known as centric reflections. For other indices, symmetry related components of the structure factor cancel out, restricting the amplitude to zero. These reflections are considered to be systematically absent.

### 1.3.2 Model – Final Result of Diffraction Experiment

The goal of a crystallographic experiment is to produce a model of the molecule of interest. This model consists of coordinates of the atoms present in the unit-cell. Crystallographic models are essentially static: only very limited dynamic information, in the form of temperature factors, is available. In contrast to the coordinates, several different methods have been developed to model temperature factors. The temperature factor of an object is related to its mobility and positional disorder. An isotropic temperature factor models the mobility as uniform in all directions. In some cases, the thermal motion is greater in some directions than others; anisotropic temperature factors can be used for these cases. The modeled object can be a single atom, a group of atoms, or an entire molecule, depending on the type of temperature factor model that is used. Any regions that exhibit significant mobility are likely to vary considerably between different copies of the unit-cell, preventing their observation.

A model is not a direct outcome of a diffraction experiment. The X-ray photons interact only with electron clouds in the crystal, not the atomic centers whose coordinates are

used in the model. The measured amplitudes are combined with a source of phase information to allow the calculation of an electron density map by inversion of the Fourier Transform. The model is initially built into this map. The structure factors (amplitude and phase) that the model would produce can be calculated, and compared with their experimental values. The model is then refined against experimental observations, and updated in an iterative manner until the differences between the calculated observable values match their experimentally measured values sufficiently closely.

Data over-fitting and model bias can cause significant problems, both during the initial model building stage and its subsequent refinement. Over-fitting formally refers to a model refined with more parameters than justified by the available number of observations. Over fitting can result in a model with apparently good statistical qualities, but this is a result of fitting the errors rather than the data. This issue is mainly dealt with by the use of cross-validation, where a subset of diffraction data is not used during refinement, in order to produce unbiased statistics. The 'free' statistics should improve along with the 'working' statistics determined using reflections used for refinement; a worsening or lack of improvement of the 'free' statistics may indicate the occurrence of over fitting. In an illustrative example (Jones et al., 1991; Kleywegt and Jones, 1995), Jones and co-workers built and refined a model with the direction of the peptide chain in reverses from the known structure, and observed only a slight difference in resulting standard statistics, although the ‘free’ statistics were able to reveal the presence of problems.

Model bias occurs when a region appears in the maps due to its presence in the pre-existing model used to determine phase values, not its presence in the data. This can

occur because the phase terms have a stronger influence on the map than the amplitudes (Read, 1997). There are a variety of approaches for reducing the effects of model bias. One type of approach involves down-weighting map coefficients for which the phase is expected to be less accurate. This is estimated based on the agreement of the calculated amplitude with the observed amplitude, as well as the overall agreement between the set of calculated and observed amplitudes. The $\sigma_A$ weighting scheme (Read, 1986) is the weighting scheme that is most commonly used at present. The other type of approach consists of various types of omit maps, in which maps phased using models with and without a region are compared. Composite omit maps can be created where each region of the map (Bhat, 1988), or model, is removed in sequence. A final combined map is reconstituted using electron density from the omit region of each of these maps.

### 1.3.3 Electron Density Map

The calculation of an electron density map requires a set of phase terms corresponding to the set of amplitudes to be used. Since the phases of a diffraction pattern can not be measured directly, they have to be determined by other means. Several methods have been developed, each with different advantages and disadvantages. For small molecules, statistical constraints and the extremely high observation to parameter ratio allow direct determination of phase values. However, the requirements of these methods are far outside what can be provided by the majority of macromolecular crystals, Pol II crystals in particular. The remaining two approaches can be broadly divided into two classes: experimental phasing and model phasing.

#### 1.3.3.1 Maps Calculated Using Model Phasing

In this approach, values for amplitudes and phases are calculated from model (in conjunction with unit-cell and symmetry parameters). These phases can then be used to calculate a map using the observed amplitudes. If the model that has been used is sufficiently similar to the molecule of interest, this map can provide more information than is present in the model. This additional information is often revealed through the use of difference Fourier techniques. Since the phases used in map calculation have more influence on the electron density than the amplitude that are used, care must be taken to avoid introducing bias towards elements of the model that are not present in the data. The refinement process discussed earlier makes use of model phases as an essential component, and is also susceptible to model bias.

In order for the resulting phases to be useful, the model must be correctly positioned and oriented in the unit-cell. This process of searching and model phasing is known as Molecular Replacement (MR), and can be broken down into several stages. First, a model "sufficiently similar" to the contents of the crystal must be identified, although this is complicated by not knowing for certain if a model is sufficiently similar until after the process has succeeded. Models of proteins that are homologous to the target protein are frequently used, as are models representing a part of the molecules of interest (for example, one subunit of a two-component complex). Regions of the search model that are expected to differ, such as side chains, mobile loops, or non-conserved regions, are usually removed prior to searching. The temperature factor of the model is typically reset to a uniform value, in order to reduce the potential of introducing inaccurate temperature factor differential. Once a potential starting model or models have been identified, it has to be positioned in the unit-cell. This requires identifying three rotational parameters and three translational parameters for each copy of the

molecule in the asymmetric unit. For computational reasons, the search process is usually broken down into two stages: identification of potential rotational solutions followed by translational searching for each of the rotational solutions.

### 1.3.3.2 Maps Calculated Using Experimental Phasing

Experimental phasing is an umbrella term covering several different approaches for obtaining phases which do not require a starting model. Fundamentally, all of these techniques make use of small differences between two or more measurements of amplitudes related in a known manner. These small differences may be caused by addition of heavy atoms, as in Isomorphous Replacement, or they may be caused by X-ray fluorescence effects, as in Multiple-Wavelength Anomalous Diffraction (MAD), or a combination of these effects. Although the experimental methods for generating crystals capable of providing the necessary amplitude differences vary, the overall phase calculation procedures are very similar (Ramakrishnan and Biou, 1997).

Regardless of the experimental method used to generate amplitude differences, positions of the responsible scatters must be determined prior to phase calculation. These sites are often referred to as heavy atom sites, although they are not necessarily due to heavy atoms (for example, the "heavy atom" in question may be a change in scattering due to a change in wavelength or due to radiation damage). Once these sites are located, initial phases can be calculated. Experimental phases are mathematically ambiguous, in that two phase values are equally possible for a single source of phase information.

The original method used for generating amplitude differences for the purpose of phase determination was the process referred to as isomorphous replacement (IR). In this

approach, heavy atom compounds are soaked into a macromolecular crystal, and useful derivative crystals are created when the heavy atoms bind to discrete sites on the macromolecule without disturbing the crystal packing in the lattice. In the ideal case, the only change in the crystal is the binding of the heavy atoms. Other changes that occur, such as changes in crystal packing, unit-cell parameters or conformational changes, reduce the degree of isomorphism, or similarity, between the derivative crystal and the unmodified native crystal. If large enough changes occur, the two crystals will no longer be isomorphous. This results in the derivative being unusable for phasing, as the derivative structure factor is no longer equivalent to the native structure factor plus the heavy atom structure factor. Accordingly, one of the bottlenecks in IR is the identification of suitable derivative crystals. This process is referred to as Single Isomorphous Replacement (SIR), or Multiple Isomorphous Replacement (MIR), depending on the number of derivative crystals.

The other major method used for generating amplitude difference for phasing involves exploiting the changes in the atomic scattering factor due to X-ray energy. Each element has characteristic electron resonance energies, and incident X-rays near these energies cause anomalous scattering effects due to atomic fluorescence. The real part of the atomic scattering factor drops sharply near a particular resonance energy, and recovers quickly above or below that energy. This results in a change between the structure factor at this energy, referred to as the inflection point wavelength, and that at a wavelength away from the resonance energy, known as the remote wavelength. In effect, this change in real scattering factor is equivalent to the change caused by the addition of heavy atoms, although it is generally smaller in magnitude. In this work, I will use the term “dispersive” to refer to amplitude differences, or phases, due to a change in the real component of the scattering factor, regardless of the source. If the

same crystal can be used for measurements at both wavelengths, then the possibility of non-isomorphism is eliminated, although radiation damage can still be a limiting factor.

In addition to changes in the real part of the scattering factor, X-ray energies slightly above the resonance energy, cause a large increase in the imaginary component of the scattering factor. This wavelength is referred to as the anomalous peak wavelength. In most cases, the imaginary component of atomic scattering is sufficiently small that it can be neglected. For cases where there are atoms with a sufficiently large imaginary component of scattering factor, Friedel's law no longer holds. Therefore, there is a measureable amplitude difference between a reflection with index (*h, k, l*) and that with index (*-h, -k, -l*). In the presence of meaningful anomalous signal, the Friedel pair is referred to as a Bijvoet pair. This difference can be exploited as a source of phase information independent from that derived from dispersive changes in the real scattering. For clarity, I will refer to amplitude difference or phase information due to changes in the imaginary component of the scattering factor as "anomalous". When phase information based on imaginary scattering alone is used, it is referred to as single anomalous scattering (SAS) or single anomalous dispersion (SAD). It can also be combined with SIR, referred to as SIRAS; or with MIR, referred to as MIRAS. When SAS is used in combination with the change in the real part scattering, it is referred to as multiple-wavelength anomalous dispersion (MAD). The dispersive phase information is independent of the anomalous phase information; therefore, MAD is capable of providing unambiguous phase values from a single crystal.

In principle, two independent sources of phase information allow unambiguous identification of the correct phase value. In practice, this is not necessarily the case due to the presence of error. Often maps calculated with multiple independent phase

sources may remain difficult to interpret. The next stage is to reduce phase errors by a process known as phase improvement, or density modification. These methods make use of additional information from various sources to improve the map. These procedures are frequently required to produce interpretable maps from experimental phases, although they are sometimes applied to model-phased maps as well. These methods use alternate sources of information as constraints on either the map, or phase set, and produce a phase set more consistent with these additional constraints.

One approach, known as solvent flattening, applies the constraint that the electron density should be featureless in the solvent region by assigning a constant value to all solvent regions of the map. Solvent flipping, a related procedure, works similarly, but uses a more involved procedure to determine the new value for a point with the solvent region. Either of these procedures requires the use of a solvent mask, which can be pre-existing or procedurally generated. Other constraints can be used as well. Non-crystallographic symmetry (NCS) and histogram matching both provide additional constraints on density with the protein region: NCS by constraining the density of related molecules to the same value and histogram matching by adjusting the overall density with the protein region to match that for previously well-determined values.

An additional factor that complicates experimental phase determination is the issue of handedness. Biological macromolecules are chiral; however, this information is lost due to the center of symmetry that is present in the diffraction pattern. The standard crystallographic statistics generated by a set of heavy atom sites in the incorrect hand will match those generated in the correct hand. For high resolution maps (better than 3 Å), this issue can be resolved because the correct hand can be recognized by visual inspection of secondary structural elements in the protein density. In lower resolution

maps, the level of detail available does not permit this distinction. Cross-anomalous difference maps can be used in these cases to determine handedness.

### 1.3.4 Data Reduction

Before any phasing procedure or map calculation can occur, a set of structure factor amplitudes is needed. These amplitudes are the experimental observations; however they are the result of several processing steps. The spot intensities observed in a diffraction image are directly related to the structure factor amplitudes used for phase determination and map calculation.

The vast majority of macromolecular diffraction experiments currently measure the diffraction by collecting a series of diffraction images of the crystal as it is rotating in an X-ray beam. The first step in producing amplitudes from images is to determine the parameters necessary for the prediction of spot locations, a process known as indexing. This process has been currently highly automated, but a conceptual understanding of the process greatly assists troubleshooting of cases in which the automated procedure fails. In order to predict the locations of diffraction spots, several parameters are necessary. Some of these parameters specify the experimental setup, such as the X-ray beam properties (e.g. wavelength, divergence), distance from the crystal to the detector, and rotational position of the goniometer. Others specify properties of the images, such as the position of the un-diffracted X-ray beam on the image, the number of pixels and size of the image. The orientation of the crystal and unit-cell parameters are also required; these are determined by the auto-indexing procedure. Errors in the input parameters produce varying degrees of error in the output unit-cell parameters. Large errors in wavelength or crystal-to-detector distance will allow indexing; however the lengths of the unit-cell edges will be incorrect by a constant factor. Errors in the direct beam

position produce the most significant problems during indexing. In the best cases, errors in the direct beam cause auto-indexing to fail, or produce solutions with obviously incorrect spot positions. More serious errors may result in a valid indexing solution, but one where the origin is inaccurate, meaning that the spots are predicted correctly, but intensities are assigned to incorrect (h, k, l) indices. Such errors do not manifest themselves until several steps later in the data reduction process. An initial space group is also assigned, based on a collection of symmetry operators consistent with the unit-cell parameters that are determined during indexing. Symmetry related reflections are not usually checked at this stage, so the initial space group may be incorrect. In addition, symmetry elements with a translational component (such as differentiating screw axes from rotation axes) are usually not distinguished at the indexing stage; however lattice centering operators can be detected.

The next stage of data reduction is to determine the total intensity for each reflection. This is done in two procedural steps: integration and merging. The integration process determines the total intensity for each predicted spot within each image. The total intensity for a reflection may be spread over several diffraction images, depending on mosaicity (three-dimensional spot width), beam divergence and the rotational range covered by the image. The crystal orientation, machine parameters, and unit-cell parameters, can be refined during integration to compensate for initial inaccuracies and allow for small shifts during data collection.

The merging process produces a single estimate of the intensity of a reflection from its observations as a spot which may cross consecutive images and spots observed at symmetry related positions. This is the first stage to explicitly use the space group in the form of symmetry operators, rather than as constraints on the cell parameters.

Accordingly, any symmetry element that is not present in the crystal will first be apparent at this stage, although the absence of a symmetry element that hasn't been assigned is generally not evident during the merging step. Errors in the origin of the reciprocal lattice during indexing will also be detectable at this point, as the observations that are combined during merging will not be truly equivalent. The merging process also needs to consider the phasing strategy to be used. If anomalous scattering is to be used, then the intensities of Friedel mates should not be merged, as this would result in a loss of anomalous signal. On the other hand, if the dataset is to be phased with molecular replacement, or used to generate dispersive phases, then the Friedel mates can be merged to increase the accuracy of the measurement.

The final stage of reducing diffraction images to structure factor amplitudes is to convert the intensities to amplitudes, and place the amplitudes on an approximate absolute scale. In order to place the data on an absolute scale, an estimate of the total scattering in the crystal, or equivalently the asymmetric unit, is necessary. This requires some information about the mass of the macromolecule as well as how many copies of the molecule are present. The mass of the macromolecule, in combination with the size of the unit-cell, is also used to determine the solvent percentage of the unit-cell.

### 1.3.5 Diffraction Data Collection

The main factor influencing the quality of data obtained during a diffraction experiment is the quality of the crystals. However, several experimental factors should be considered in order to allow extraction of the maximum possible amount of information from the available crystals. The typical X-ray diffraction experiment uses a single crystal rotating about one or more axes in a monochromatic X-ray beam. A single image generally covers an angular range of 0.5 to 2.0 degrees. The width of an image is

dictated by the mosaicity of the crystal. Larger mosaicities will result in fewer spots being fully recorded in a single image. The total angular range required to collect a complete dataset depends on the space group and the initial orientation of the crystal. A crystal with high symmetry may require only 60 to 90 degrees of data for a complete dataset. For crystals of lower symmetry, an angular range of 180 degrees would be required. Collection of a wider angular range than required for completeness can improve the dataset by increasing the multiplicity of observations for each reflection. However this improvement may be negated by damage caused to the crystal by prolonged radiation exposure. The phasing strategy to be used is another factor influencing data collection. For data to be phased using a model, or used to derived dispersive (or isomorphous) experimental phases, Friedel pairs can be considered equivalent. Data in which usable anomalous signal is expected must treat the equivalent Friedel pairs of reflections as non-equivalent Bijvoet pairs, requiring a doubling of the angular range required for a complete dataset. To maximize the measurement of the anomalous signal available, the Bijvoet mates should be measured with as short time between them as possible, in order to minimize amplitude differences due to X-ray source fluctuations, accumulating radiation damage, and other effects. In some cases, the crystal can be aligned so that both members of a Bijvoet pair are recorded on the same image. An alternative approach is to collect one or more images (recording spots at *h, k, l* indices) followed by images offset by 180 degrees (recording spots at *–h, -k, -l* indices). An analogous procedure can be used for MAD data collection, shifting wavelengths instead of, or in addition to, shifting the angular range. The wavelength used for data collection is particularly significant for anomalous or dispersive datasets. The anomalous scatterers may have slightly different absorbance edges when bound to a macromolecule than when in their elemental form, so the wavelengths should be chosen on the basis of an X-ray fluorescence scan of the experimental crystal.

For datasets where anomalous data is not of importance the wavelength used is of less importance, and is generally selected to minimize air scattering, or maximize the intensity available. For non-synchrotron sources, wavelength selection is determined by the anode in the X-ray generator (copper anodes producing 1.54 Å X-rays are most commonly used). Other experimental parameters, such as exposure time and crystal-to-detector distance, need to be optimized for data collection. Decreasing the distance between the detector and the crystal allows for recording of higher resolution spots, but reduces spot separation; in practice the resolution of the crystal usually dictates the choice for distance. The exposure time for an image should be chosen to maximize the signal to noise ratio for the spots, while minimizing the number of saturated pixels. In practice, it is sometimes not possible to do this within a single dataset. In these cases, multiple 'passes' can be made, with one selected to record high resolution spots, and another optimized for low resolution reflections.

## 1.4 Application of Crystallography to the Study of Pol II Structures

A variety of Pol II complexes play essential roles in transcription. As mentioned earlier, the limited structural knowledge of such complexes is due to the technical obstacles in determining Pol II complex structures. Although these difficulties are inherent to Pol II complexes, similar difficulties could be expected for structures of other large macromolecular complexes.

The first bottleneck to any structure determination is the production of material. The purification of Pol II alone, in either 10 subunit or 12 subunit forms, is a substantial undertaking. The number of subunits prevents the standard route of cloning into *E. coli* and over-expressing. The best current approach is to purify Pol II directly from a

suitable source, such as *S. cerevisiae*. For a complex, this difficulty is enhanced by requiring formation of a stoichiometric complex of Pol II and the additional factor. This is accomplished by either reconstitution with a separately purified factor, or purification of the complex directly from yeast. Once enough protein is available, the determination of good crystallization conditions presents the next obstacle. The search for conditions capable of producing diffraction quality crystals is hampered by the relatively limited quantities of protein that can be produced. Crystals available from good conditions can still present additional problems. Careful handling is required to prevent mechanistic damage and oxidation of the Pol II protein, which contains oxygen-sensitive zinc motifs. Data collection at cryogenic temperatures is essential, in order to prevent radiation damage during data collection. Crystals of Pol II require careful optimization of cryo-protectant conditions and freezing methodology, due to their high solvent content (up to 80%) and relatively weak lattice contacts. Even with extensive work on the steps that are required for producing the best possible crystals, the diffraction obtained from the best crystals of Pol II complexes (3.5 to 4.5 Å) would be typically considered low resolution by the standards used for work on smaller macromolecules.

Improvements in X-ray sources, data processing algorithms and software have dramatically simplified the process of determining a structure once crystals are available. For a typical protein crystal, this process can be highly automated, and the time required for data processing can range from a few weeks to a few minutes. However, as judged by structures deposited in the Protein Data Bank (Berman et al., 2000) (as of June 2009), the average structure has a resolution of approximately 2.2 Angstroms, and a molecular mass of approximately 70 KDa. By way of comparison, the average Pol II complex crystal has a molecular mass of 500 to 700 KDa and

diffracts to a resolution of up to 3.8 Angstroms. The relatively low quality of diffraction data that is available requires careful optimization of the diffraction data processing, just as with the earlier stages (protein purification, crystallization, and cryo-protection) in order to extract the maximum amount of structural information. Work in the Fu lab on complexes of Pol II with TFIIF or RNA capping enzyme revealed several cases in which improvements to data processing could assist the determination of Pol II complex structures. In order to ensure the effectiveness of these improvements, they were validated by testing them on diffraction data from the 12-subunit Pol II.

# CHAPTER 2: MULTI-CRYSTAL ZINC-MAD PHASING OF POL II

## 2.1 Introduction

### 2.1.1 Why Zn-MAD?

Following the initial determination of the 10-subunit Pol II structure (Cramer et al., 2000), additional Pol II structures have become available. There are structures of the full 12-subunit Pol II, both 10- and 12-subunit forms complexed with various nucleic acids and NTPs, and partial complexes with two different transcription factors, reviewed in (Cramer et al., 2008). By far, the vast majority of these structures were phased by molecular replacement using an earlier model. This approach bypasses many of the technical steps required for experimental phasing. The search model is by definition highly homologous to the target, since the same core Pol II molecule is present in the crystal, although significant conformational changes may be present due to variation in clamp position or other potentially mobile domains of the polymerase.

The potential limitations of this approach are illustrated by the structures of two complexes containing transcription factors that have been published. In each case, the resolution of data provided by complex crystals was limited: 3.8 Å for the TFIIS complex (Kettenberger et al., 2003; Wang et al., 2009) and 4.5 Å for the TFIIB complex (Bushnell et al., 2004). Possibly related to that limitation, only a portion of the transcription factor was able to be modeled. Similarly, work in our lab on complexes with other transcription factors was only able to produce crystals with relatively low resolution, and poor electron density for additional factors using model derived phases.

The general properties of crystallographic structure factors offer three supporting explanations for why model derived phasing exhibits limitations for phasing Pol II complexes. The first of these factors is the relative completeness of the model in a transcription factor complex. The effectiveness of model phases is a function of the accuracy and completeness of the model (Read, 1986; Srinivasan and Ramachandran, 1965). Assuming the accuracy of the Pol II model is unchanged, lower quality density would be expected for a 200 kDa transcription factor than a 500 Da nucleotide, as the known Pol II model represents less of the total mass in the former case. The second factor is the effect of phase error as a function of resolution. As illustrated by Wilson statistics and atomic scattering factors, the average amplitude of a structure factor tends to decrease with increasing resolution. Therefore, for a constant phase error, the root-mean-square (RMS) error between the true structure factor and model derived structure factor will be greater for a low resolution reflection (Figure 2.1). This increased RMS error in the structure factors translates directly into increased errors in the electron density map (Read, 1997). The third factor is also related to resolution: the number of reflections increases roughly cubicly with increasing resolution. Since the amplitudes associated with these reflections are the sole source of experimental information in the absence of experimental phases, as the resolution limit decreases there is simply less experimental information available.

These limitations suggest that experimental phases would be necessary in order to phase Pol II complexes containing large transcription factors. The insufficiency of model phasing was demonstrated by lack of additional protein density in co-crystals of TFIIF-Pol II and CE-Pol II complexes using phases from the Pol II model (unpublished results). In each of these cases, molecular replacement was able to correctly position the Pol II model with relative ease, but was not able to provide additional usable phase

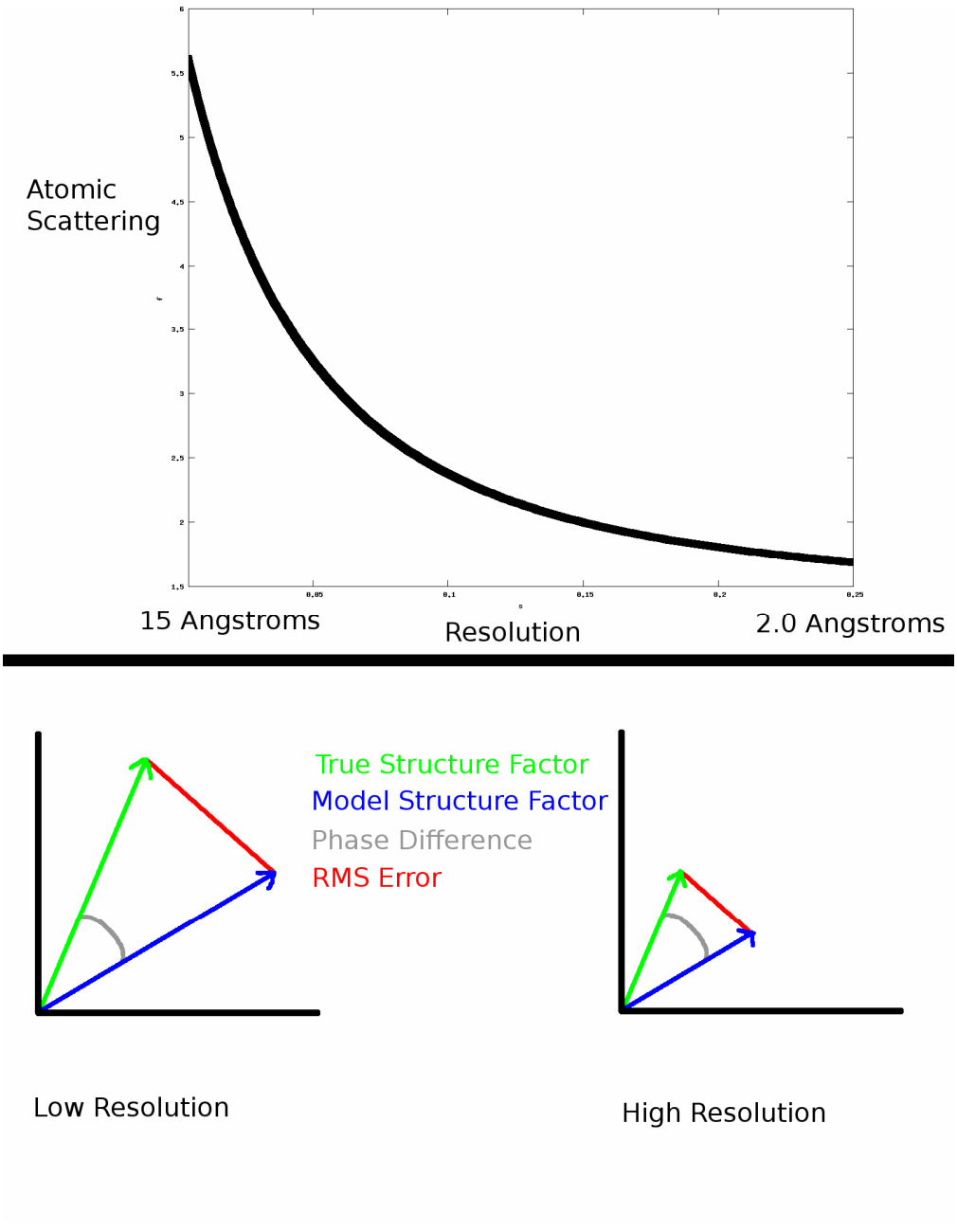


Figure 2.1: Average Structure Factor Amplitude and RMS Error at High and Low Resolutions

Distribution of average atomic scattering as a function of resolution is shown in *top panel*. This distribution for a protein would show a hump at 4 Å due to secondary structural elements. Harker diagrams for a low resolution, large amplitude reflection (*bottom left*) and high resolution, small amplitude reflection (*bottom right*) illustrate the RMS differences for a constant phase difference.

information regarding the non-Pol II portion of the complex. MIRAS was the phasing approach that was used the first Pol II crystals (Cramer et al., 2000; Fu et al., 1999). As discussed above, this approach requires the identification of suitable derivative crystals. The use of selenomethionine substituted proteins to generate anomalous signal for MAD or SAS is an approach that has become widely used. However, selenomethionine substition is toxic to *S. cerevisiae*, and the partially substituted proteins are unable to produce sufficient phasing power for large complexes. As an alternative, the use of intrinsic Zn ions present in Pol II for phase determination was investigated. As described below, this approach was effective.

### 2.1.2 Why Pol II Again?

As described above, Pol II contains 8 Zn ions, which are believed to stabilize the tertiary structure. Zinc produces a relatively weak anomalous signal, roughly comparable to selenium, which could be used as a potential source of experimental phase information. However, the ratio of anomalous Zinc to non-anomalous light atoms in Pol II is extremely low: approximately 1 Zinc per 60 kDa protein, corresponding to a Bijvoet ratio of 1.32%. For comparison, the median ratio for structures deposited in the Protein Data Bank (Berman et al., 2000) phased by means of Zn anomalous is 1 Zinc per approximately 16 kDa protein (Figure 2.2).

A multi-crystal approach was adapted in order to enhance the weak phasing signal. Early work on CE-Pol II co-crystals indicated that this approach had the potential to allow for unbiased phasing of Pol II complex crystals. In order to validate the potential of this approach, and identify important factors for producing the best possible experimental map, a control experiment was conducted to re-phase the 12-subunit Pol II using multi-crystal Zn-MAD.

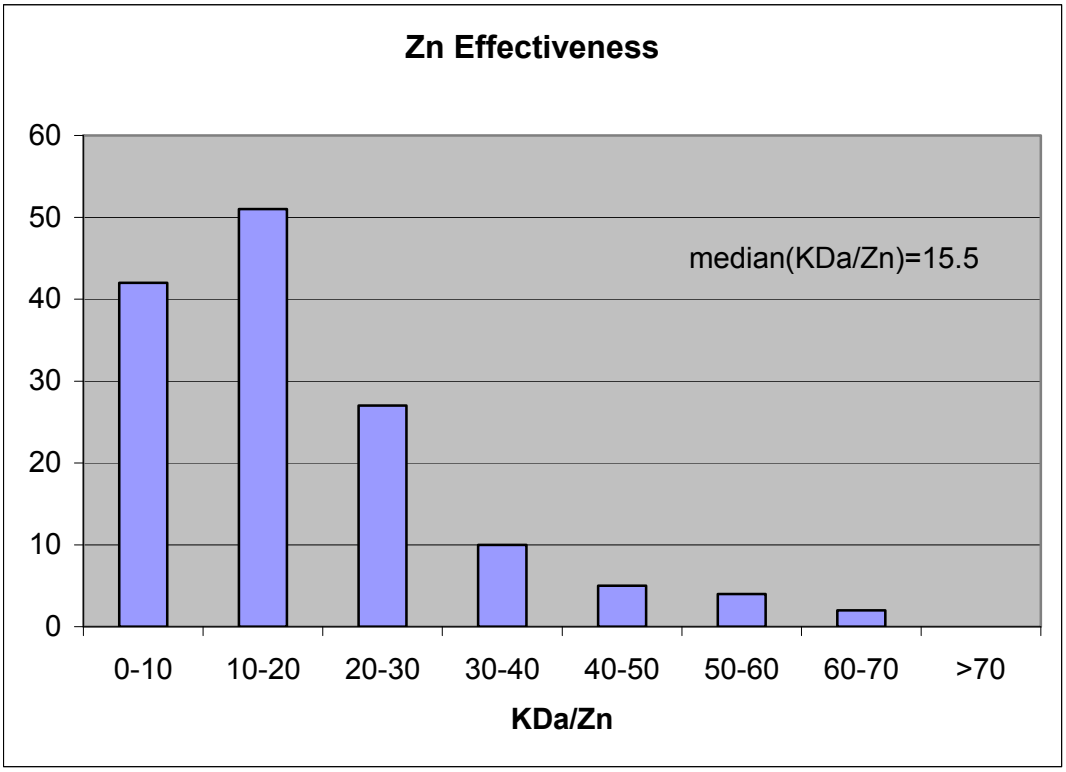


Figure 2.2: Comparative Effectiveness of Zn Anomalous Phasing

Histogram of molecular mass / number of Zn for structures phased using Zn anomalous. Entries were identified by reported phasing method, wavelength, and number of Zn sites according to RCSB Protein Data Bank (Berman et al., 2000) snapshot from Jan 5, 2009. The phasing effectiveness of Zn-MAD is 64.8 KDa/Zn for Pol II.

### 2.1.3 Multi-Crystal Approach Applied to Weak Anomalous Data

Multi-crystal approaches are used in several ways in crystallographic data processing. In order to clarify the advantages and disadvantages of multi-crystal phasing, it is helpful to provide some details regarding the distinction between multi-crystal phasing and other multi-crystal methods. Multi-crystal approaches differ in the stage of data processing at which combination occurs, which in turn places restrictions on the datasets suitable for combination.

One of the earliest and most common multi-crystal approaches addresses the tendency of crystals to deteriorate during data collection due to radiation damage. As a result of this, each crystal does not produce a dataset that is sufficiently complete. Multiple crystals are used to allow collection of a complete dataset with sufficient redundancy. As discussed earlier, the merging step combines multiple, possibly partial, observations of the intensity for a single reflection from different frames or equivalent indices into an estimate of the intensity (and associated error) for that reflection. The combination of multiple crystals at the reflection merging stage adds additional observations for these indices from the multiple datasets used. Multi-crystal merging requires the crystals used to be isomorphous; otherwise the final intensity will be a mixture of multiple intensities rather than an improved estimate of a single intensity. The end result of this approach is a more complete, ideally more accurate, or both, set of intensities than produced by any of the individual datasets used.

An alternative multi-crystal approach is electron density averaging using multiple crystals. As discussed earlier, density modification is the process of applying real-space constraints to a phase set in order to improve the phase values. In multi-crystal density

modification, maps from multiple crystals are used as simultaneous sources of real-space information in order to obtain phase improvement. This approach does not require that the crystals used be isomorphous; indeed, the data used are frequently from different crystal forms with differing space groups. The end result of this approach is a set of improved phases, in both crystal forms (Cowtan, 1994).

In yet another approach, that is multi-crystal phasing, datasets from individual crystals are combined at the phasing step (Abrahams and Leslie, 1996). This approach is somewhat analogous to MIR or MAD phasing, with the caveat that the input phase probability distributions are not fully independent. Like multi-crystal averaging, the end result is an improved set of phases. Similar to multi-crystal merging and MIR, the crystals used must be isomorphous, or the final result will be a degraded phase set rather than an improved one.

## 2.2 Results on Zn-MAD Phasing of Pol II

### 2.2.1 Zinc Signal is Sufficient for Locating Zinc-ions

Determining the location of anomalous scatterers, such as Zn, is the first step of MAD phasing. X-ray fluorescence scans clearly showed the presence of Zn in these crystals (Figure 2.3). Difference Fourier is a standard technique used to allocate anomalous scatterers when a prior phase source is available. Using this approach, peaks were observed in anomalous and dispersive difference maps phased with phases from the Pol II model. These peaks corresponded with the previously determined positions of zinc atoms in the Pol II structure (Figure 2.4). This observation, in combination with the X-ray fluorescence scans of the crystals (Figure 2.2), indicated that the zinc sites had been successfully located. As anticipated, the dispersive difference maps showed weaker peaks, and more noise, than anomalous difference maps (Figure 2.4).

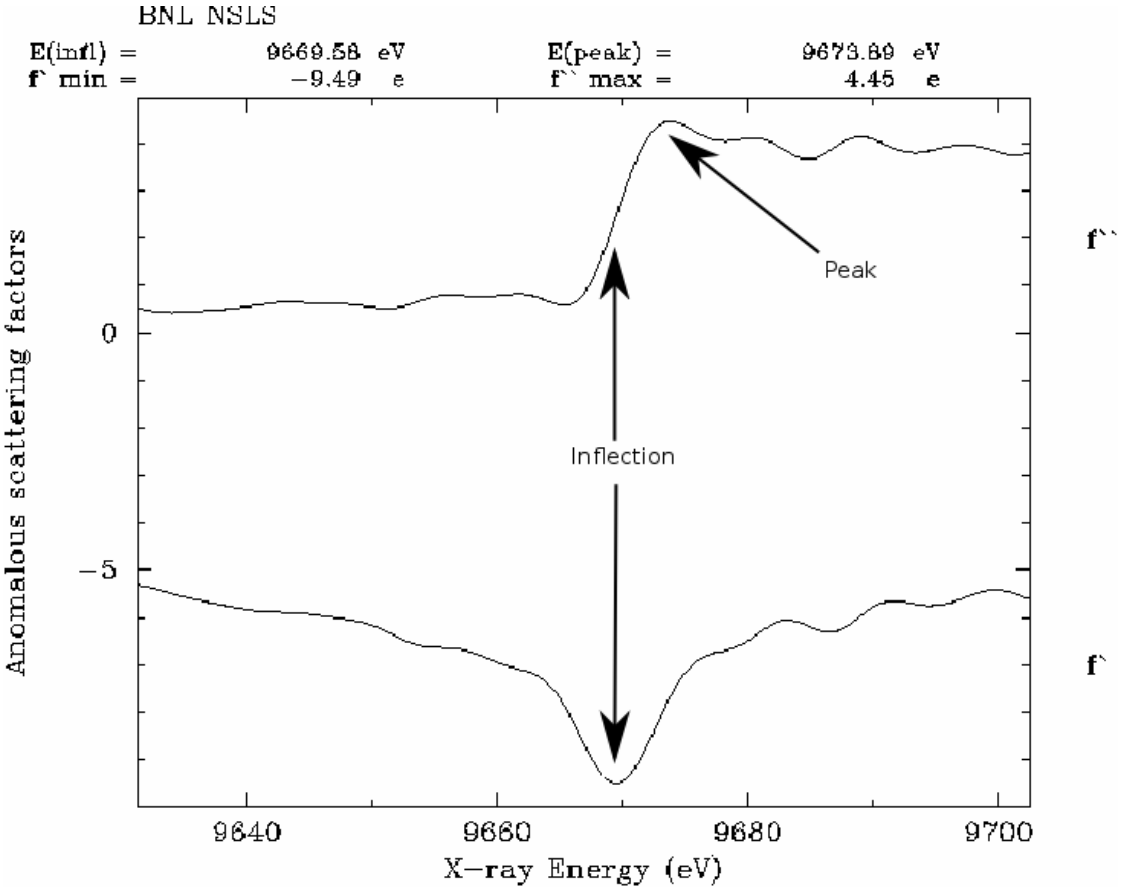


Figure 2.3: Representative X-ray Fluorescence Scan

The fluorescence scan shows the presence of measurable zinc anomalous signal in one of the 12-subunit Pol II crystals, as indicated by the significant absorption at the inflection energy. Arrows indicate the *inflection* energy (where real scattering due to anomalous effects is smallest) and *peak* energy (where imaginary scattering is maximized) for the K-edge of Zn. f' is the real component of scattering due to anomalous effects, and f'' is the imaginary component of scattering due to anomalous effects.

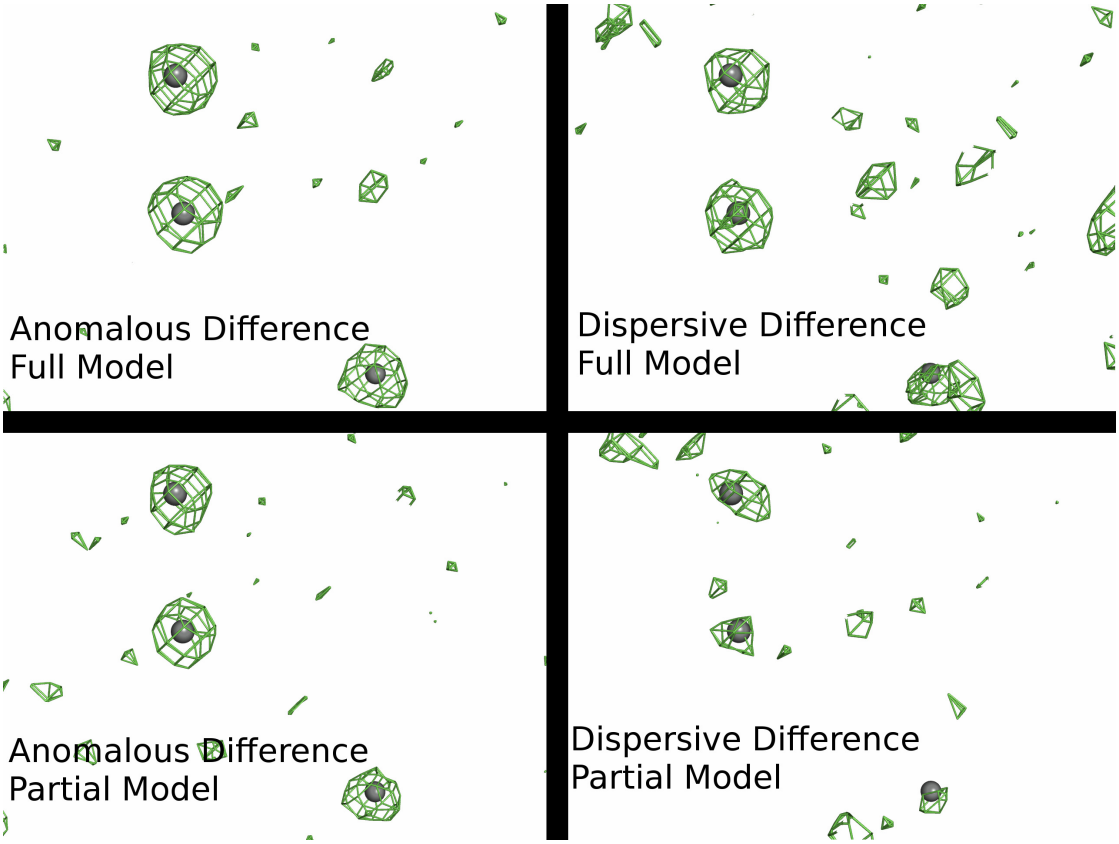


Figure 2.4: Representative Model Phased Anomalous and Dispersive Difference Maps

Three of eight zinc atoms are shown as *grey spheres*. All maps are shown in *green mesh*, contoured at 3.0 σ. Anomalous difference Fourier maps were phased using either full model (*upper left*) and partial model (*lower left*). Dispersive difference Fourier maps were phased using full model (*upper right*) or partial model (*lower right*). The partial model used for anomalous difference Fourier was the 12-subunit model described in the text with Rpb1 and Rpb2 removed. The partial model used for dispersive difference Fourier was the same 12-subunit model with Rpb1 removed**.** The closest distance between any two zinc atoms, or symmetry-related equivalents, is 16 Å.

In order to exclude the possibility that these peaks were due to bias in the model, the model was manually placed into an arbitrary position in the asymmetric unit. Difference maps calculated using the incorrectly placed model did not show any peaks. Similarly, the presence or absence of zinc atoms in the model that was used for phasing did not affect the presence of peaks in the difference maps, excluding the possibility that their presence was residual effects often associated with difference Fourier maps.

Although the peaks observed in these difference maps were sensitive to the orientation and position of the model, they were relatively insensitive to its completeness. The peaks were still observed in difference maps phased using a model from which the Rpb1 and Rpb2 subunits had been removed, corresponding to a model that was only approximately 40% complete (Figure 2.4, lower panels).

The location of Zn sites in model phased difference maps avoids a potential requirement for the use of heavy-atom clusters. In the original phasing of the first Pol II crystal, the use of heavy atom clusters was required in order to initiate experimental phasing, as the background noise from the polymerase prevented the location of discrete heavy atoms. In addition, potential problems with the handedness of the data were also avoided by using model phased difference maps.

The use of difference Fourier techniques was essential for location of the zinc positions. Peaks corresponding to zinc ions were not detectable in either anomalous or dispersive difference Patterson maps. Similarly, SAPI (Hao et al., 2003) was unable to locate the zinc sites using a tangent-formula approach.

### 2.2.2 Single Crystal Phasing

Before phase combination, datasets from each of the individual crystals were phased individually. For one of these crystals, it was possible to collect a full MAD dataset. The quality of difference Fourier maps roughly correlated with the phasing statistics and quality of solvent flattened maps from individual crystals, with SAS maps generally being of higher quality than dispersive maps. As expected, the single crystal MAD phases were of higher quality than the component SAS or dispersive phases (Figure 2.5). The effects of radiation damage on dispersive data were illustrated by the behavior of a dataset that was not collected in wedges relative to the other dispersive maps. Collecting the remote and inflection as two consecutive datasets produced a worse map in comparison to datasets where remote and inflection images were collected in wedges of 20 or 40 images.

### 2.2.3 Multi-Crystal Map Agrees with the Known Model

The single crystal maps produced were of varying quality, and showed density in physically impossible regions (for example, isolated islands with no connection to the remainder of the lattice) and contained regions of the model with no corresponding experimental density. In order to improve upon these results, the following multi-crystal approach was investigated. Starting with the best diffracting crystal as a base, the remaining datasets were combined at the phase calculation step. Datasets that improved the solvent flattened map were kept; otherwise the additional dataset was discarded. Somewhat surprisingly, not all datasets resulted in a combined map after addition to the phase set. During later projects, alternative methods to determine an optimal phase set were investigated, as described in the discussion. The final experimental map was produced using a set of 2 anomalous datasets and 2 dispersive datasets from a total of 3 crystals, with a resolution of 4 Angstroms (Table 2.1). This

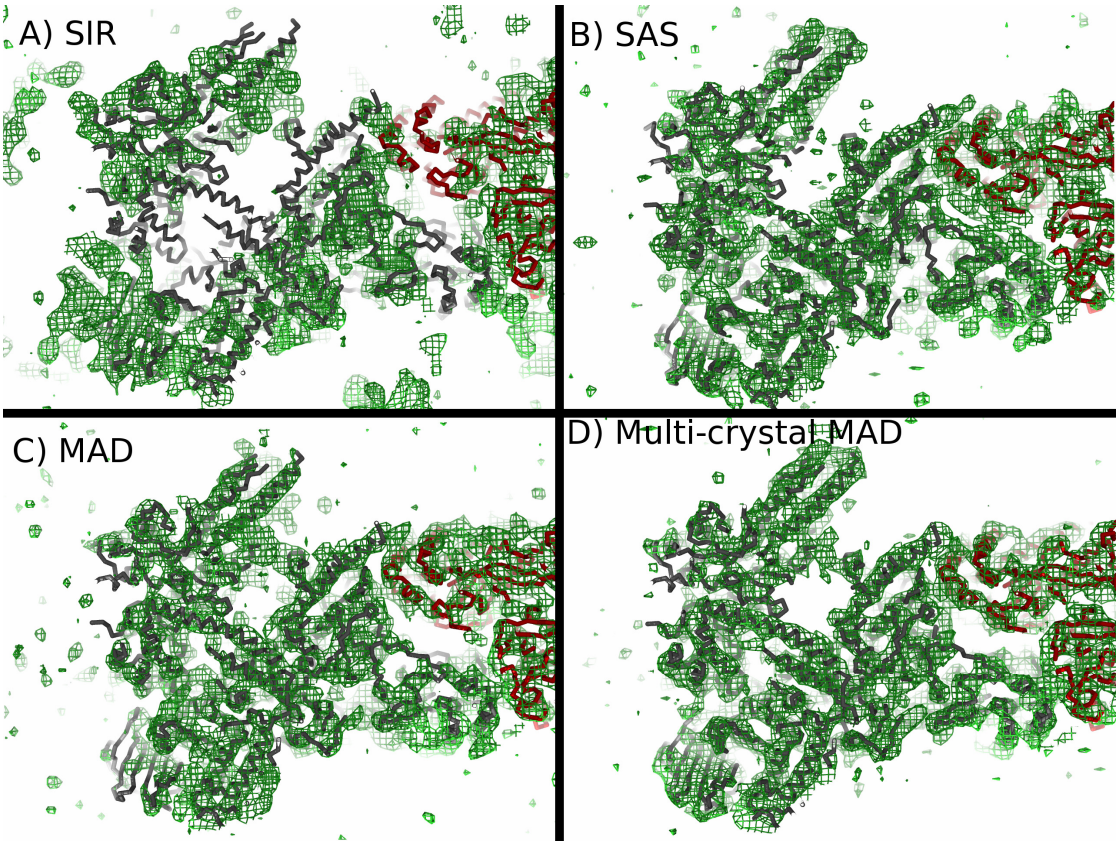


Figure 2.5: Experimental Maps

Representative solvent-flattened experimental maps, contoured at 1.0 σ, are shown as *green mesh*. Pol II model is shown as *grey ribbon*, symmetry related model shown as *red ribbon*.

Table 2.1: Merging and Phasing Statistics for Pol II Datasets

All datasets were in space-group $C222_1$, with one molecule per asymmetric unit. Dataset naming scheme based on crystal and drop ID according to Limbro plate (e.g. A3X7 is crystal 7 from the well at row 3 column A). Suffix (inf, rmt or pk) denotes X-ray energy for Zn inflection, high-energy remote, or peak wavelength, respectively. Merging statistics are from SCALA, and phasing statistics are from PHASIT. Datasets phased using dispersive differences are labeled SIR in the Type column. Datasets phased using anomalous are labeled SAS in the Type column.

| Dataset | Λ | a | b | c | I/sigma | completeness | multiplicity | resolution | Rsym |
|---|---|---|---|---|---|---|---|---|---|
| A4X4inf | 1.282 | 221.95 | 394.36 | 280.92 | 5.5 (2.2) | 99.8 (99.8) | 5.9 (6.0) | 81-7.65 | 0.117 (0.296) |
| A4X4rmt | 1.279 | 220.23 | 392.52 | 280.80 | 5.8 (2.5) | 99.9 (100.0) | 6.9 (7.1) | 112 – 6.9 | 0.109 (0.271) |
| A4X6inf | 1.283 | 220.85 | 394.45 | 282.08 | 4.1 (2.4) | 99.8 (99.8) | 13.2 (13.6) | 182-7.2 | 0.154 (0.295) |
| A4X6rmt | 1.265 | 221.13 | 394.38 | 281.72 | 4.3 (2.5) | 99.8 (99.9) | 12.4 (12.8) | 91-7.3 | 0.148 (0.280) |
| A3X7pk | 1.282 | 220.56 | 391.38 | 280.39 | 7.5 (3.3) | 99.6 (99.6) | 13.7 (14.1) | 182-6.5 | 0.084 (0.235) |
| A3X7pk-ANISO | 1.282 | 220.56 | 391.38 | 280.39 | 4.7 (2.0) | 85.5 (64.1) | 9.3 (3.6) | 182-4.15 | 0.143 (0.369) |
| A3X7inf | 1.283 | 220.22 | 391.27 | 280.45 | 6.7 (2.6) | 99.2 (99.2) | 8.8 (9.0) | 182-7.3 | 0.104 (0.290) |
| A3X7rmt | 1.276 | 220.22 | 391.27 | 280.45 | 7.5 (2.9) | 99.2 (99.3) | 8.8 (9.1) | 182-7.2 | 0.093 (0.274) |
| A2X10pk | 1.282 | 220.69 | 394.33 | 281.31 | 7.3 (2.4) | 100.0 (100.0) | 11.5 (11.9) | 182-6.2 | 0.090 (0.298) |
| A2X11inf | 1.283 | 222.12 | 394.37 | 281.80 | 6.9 (2.6) | 82.2 (83.5) | 4.6 (4.7) | 91-6.6 | 0.088 (0.278) |
| A2X11rmt | 1.276 | 222.12 | 394.37 | 281.80 | 6.8 (2.9) | 82.6 (83.9) | 4.7 (4.8) | 87-6.5 | 0.092 (0.256) |
| master(ANISO) | | | | | 3.8 (1.6) | 85.2 (63.9) | 26.8 (3.6) | 158-4.15 | 0.165 (0.453) |
| master(ISO) | | | | | 4.1 (2.7) | 100.0 (100.0) | 67.5 (27.7) | 158-6.5 | 0.153 (0.266) |

map agreed almost completely with the pre-existing model, demonstrating that the zinc anomalous signal is sufficient to phase a protein of this size (Figure 2.5).

### 2.2.4 Determination of the Solvent Mask is a Critical Step

The Pol II crystals used typically diffracted to resolutions of 8 to 6 Angstroms. However, one crystal used in this study allowed the collection of a 4 Angstrom dataset. During initial processing of the data, it was treated as a 6.5 Angstrom dataset. Once an optimal phase set had been identified, the higher resolution data was incorporated. Surprisingly, the solvent flattened map produced from this phase set was worse than the comparable map at 6.5 Angstroms. Comparison of the solvent masks used showed that the higher resolution phase set did not produce a reasonable solvent mask when using the standard spherical averaging procedure (Wang, 1985). After conversion of the lower resolution mask to the appropriate grid suitable for solvent flattening at the higher resolution, density modification procedures were able to produce a good experimental map (Figure 2.5 D).

### 2.2.5 Simulation Suggests Multi-Crystal Zn-MAD Should be Effective for Complexes up to ~1 MDa

In order to further investigate the limits of this phasing approach, a numerical experiment was conducted. Simulated data was generated for Pol II by calculating structure factors incorporating appropriate f' and f'' values for all atoms, and generating errors. The number of Zn atoms was varied from a maximum of 8 (as in native Pol II) to 1. Multiple simulated crystals were created by independent repetition of the error generation procedure, up to a maximum of 10 for each number of Zn atoms. As anticipated, the number of datasets required to produce a good protein map increased as the number of Zn sites decreased. Assuming 4 isomorphous crystals supplying a full

MAD dataset, or the equivalent number of SAS and dispersive datasets from more crystals, as a practical limitation, it was found that 4 to 5 zinc sites would be sufficient for phasing Pol II (Figure 2.6 and Table 2.2).

Somewhat unexpectedly, the contribution of the non-Zn atoms (light atoms) to the total anomalous scattering exhibited a substantial effect on the simulation results. During initial testing of the simulation procedure, the anomalous contribution of the light atoms was set to zero. Under these conditions, a single simulated crystal with a single Zn atom was able to produce a high quality map. Under more realistic conditions, where anomalous scattering from the light atoms were set to theoretical values, a greater number of Zn sites and crystals were required.

### 2.2.6 The Experimental Map Shows Previously Un-modeled Regions of Pol II

The final experimental map agreed very closely with the existing Pol II model. However, additional density was observed at several regions. Closer inspection showed that this density corresponded to regions where the model had gaps. A preliminary poly-alanine model was built for three of these regions: fork loop 1 in the Rpb1 subunit, part of the Rpb2 protrusion (B437-B446), and part of Rpb4 (D113-D117). Initial attempts to refine the model using the MLKF (Murshudov et al., 1997) and MLHL (Pannu et al., 1998) targets in CNS (Brünger et al., 1998) were unsuccessful. These regions, along with a workable refinement scheme that was later implemented, are discussed in more detail in chapter 3. The model containing the newly built regions, along with master amplitudes and experimental phases, was deposited in the Protein Data Bank as entry 2B8K (Meyer et al., 2006).

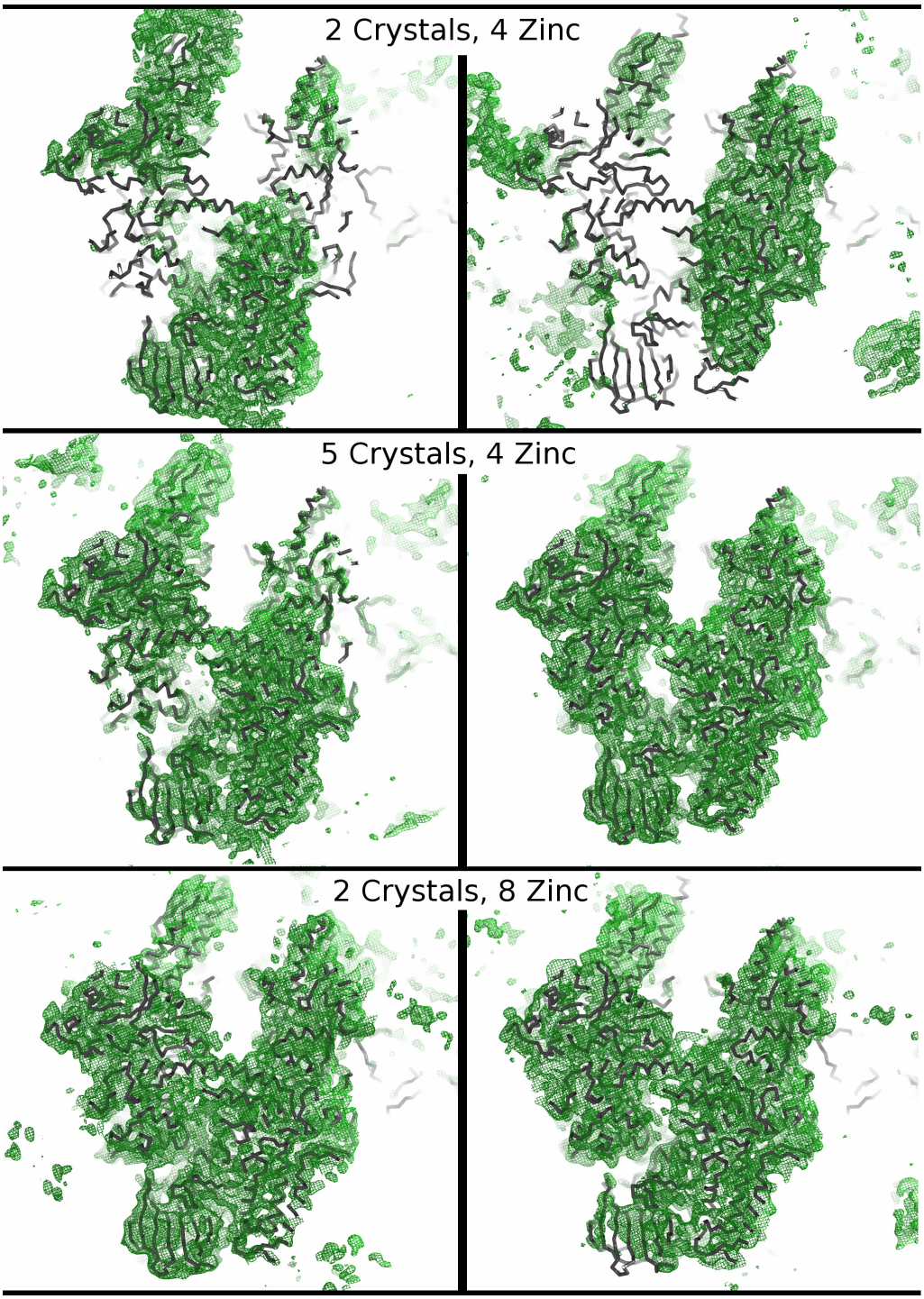


Figure 2.6: Maps from Simulation

Duplicates of maps calculated using simulated multi-crystal datasets showing examples of unsuccessful (two crystals, 4 zinc) and successful (5 crystals, 4 zinc; 2 crystals, 8 zinc) simulation runs. Models are shown as *grey ribbon*; maps contoured at 1 σ are shown as *green mesh*.

Table 2.2: Simulation Summary Statistics

*Bold* entries represent conditions where the simulated map matched the model. *Underlined* entries represent conditions where there were slight differences between the simulated map and model. Values are the average of two independent simulation runs.

A. Correlation coefficients (C.C.) between Pol II Fc map and maps from simulated data

| Map C.C. | # Crystals = 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| # Zn = 8 | 0.579 | **0.660** | 0.600 | **0.669** | **0.759** | 0.598 | **0.729** | 0.588 | **0.764** | **0.754** |
| 7 | 0.327 | 0.641 | **0.691** | 0.586 | 0.689 | **0.741** | 0.625 | 0.581 | 0.617 | **0.756** |
| 6 | 0.241 | 0.528 | 0.546 | 0.649 | 0.555 | **0.693** | 0.616 | 0.644 | 0.652 | 0.615 |
| 5 | 0.423 | 0.549 | 0.591 | 0.433 | 0.576 | 0.550 | 0.512 | **0.680** | 0.564 | **0.716** |
| 4 | 0.401 | 0.457 | 0.634 | 0.521 | **0.700** | 0.566 | 0.631 | 0.661 | **0.722** | 0.652 |
| 3 | 0.326 | 0.273 | 0.432 | 0.374 | 0.345 | 0.708 | 0.364 | 0.426 | 0.556 | 0.573 |
| 2 | 0.436 | 0.264 | 0.241 | 0.424 | 0.406 | 0.212 | 0.427 | 0.578 | 0.526 | 0.517 |
| 1 | 0.183 | 0.151 | 0.329 | 0.305 | 0.390 | 0.369 | 0.366 | 0.414 | 0.309 | 0.461 |

B. Real space R-factors (RSR) between the Fc map and maps from simulated data

| RSR | # Crystals = 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| # Zn = 8 | 0.557 | **0.512** | 0.554 | **0.502** | **0.438** | 0.556 | **0.455** | 0.575 | **0.445** | **0.433** |
| 7 | 0.720 | 0.522 | **0.479** | 0.569 | 0.491 | **0.444** | 0.545 | 0.586 | 0.554 | **0.429** |
| 6 | 0.774 | 0.603 | 0.591 | 0.517 | 0.601 | **0.496** | 0.544 | 0.532 | 0.530 | 0.556 |
| 5 | 0.664 | 0.586 | 0.571 | 0.673 | 0.579 | 0.594 | 0.628 | **0.502** | 0.593 | **0.474** |
| 4 | 0.670 | 0.662 | 0.524 | 0.610 | **0.473** | 0.583 | 0.553 | 0.513 | **0.472** | 0.531 |
| 3 | 0.724 | 0.761 | 0.669 | 0.715 | 0.737 | 0.477 | 0.728 | 0.686 | 0.599 | 0.577 |
| 2 | 0.634 | 0.756 | 0.782 | 0.666 | 0.683 | 0.813 | 0.678 | 0.566 | 0.619 | 0.622 |
| 1 | 0.797 | 0.809 | 0.712 | 0.726 | 0.678 | 0.697 | 0.699 | 0.667 | 0.742 | 0.643 |

## 2.3 Discussion

The quality of the final experimental map demonstrates that there is sufficient anomalous signal from the intrinsic Zn atoms to produce a high quality phase set that is not susceptible to model bias. Although the use of an existing model is essential in the initial location of the phasing sites, errors in the initial protein model do not affect the experimentally phased map. In addition, the increased effects of phase error on low resolution reflections discussed in the introduction of this chapter do not affect the difference Fourier techniques used for Zn site location. This is due to the fact that the anomalous difference ( $|F^+ - F^-|$) and dispersive difference ($|F_{remote} - F_{inflection}|$) do not show the same resolution dependence as the native amplitudes discussed in the introduction of this chapter. The effects of model completeness on anomalous and dispersive difference maps show that this technique of heavy atom site location requires only approximately 50% of model to be used, suggesting that the Zn sites could be located in Pol II complexes where Pol II represents only half of the total mass. The simulation results indicate that the size limitation for Zn-MAD phasing depends largely on the number of compatible crystals from which data can be collected, and that complexes where Pol II represented 50-60% of the total mass should be within practical limits. Therefore, this phasing approach should be effective for phasing Pol II complexes with a total mass of up to 1 MDa using the 8 Zn sites present in native Pol II.

Several potential problem areas in multi-crystal phasing were also identified. The small amplitude differences due to Zn anomalous scattering require that care be taken to minimize errors in the measurement of amplitude differences. Dispersive data worked best when collected in wedges covering a relatively small angular range, to minimize any variations due to radiation damage or X-ray source fluctuations.

Anomalous data collected using wedges offset by 180 degrees similarly minimized differences between Bijvoet pairs. The shift in X-ray energy, or angular offset, introduces a delay between images collected at the edges of each wedge, during which the alternate wedge is collected. When reducing the diffraction data, the results obtained using the integration program MOSFLM (Leslie, 1992) were improved when the images were integrated in chronological, rather than rotational, order. This was rationalized as avoiding large jumps in the parameters modeling the setup of the X-ray station, which could lead to increased errors. Due to the way MOSFLM handles the orientation matrix, improved results were obtained by storing the orientation matrix after integrating a single wedge, and reloading this orientation matrix for proceeding to the next angular wedge.

The second problem area is the identification of a compatible set of individual datasets for phase combination. Some of the collected datasets made the combined map worse rather than better upon combination. This effect has been attributed to non-isomorphism between different crystals. Differences in the unit-cell parameters of different crystals were not helpful for detecting non-isomorphism, nor were merging R-factors. As is widely known, unit-cell parameters determined by auto-indexing at low resolution are not determined very precisely. This was observed by indexing the same dataset using spots from different images: variations of as much as 5 Å were observed in these cases. Cross-crystal dispersive difference maps were evaluated as another source of information, but also proved insufficiently sensitive. In the work on Pol II, the only method that was able to determine if a phasing dataset would improve the combined phase set or not was to check the solvent flattened maps. During later work on crystals of Pol II complexes, attempts were made to improve methods for identifying an optimal phase combination by other means. These efforts were

motivated partially to reduce the time required, and partially to avoid any potential subjectivity when comparing the maps. Cluster analysis was one approach investigated. This was done by using the QT clustering algorithm (Heyer et al., 1999) to cluster un-flattened single crystal maps, using real space statistics (real-space R-factor and map correlation coefficient) as a distance metric. The clustering approach was able to cluster by phasing source (SAS or dispersive), but was ultimately unhelpful for identifying an optimal phase combination. This failure was attributed to the presence of the ambiguous 'noise' peak in the phase probability distributions. During work on the CE-Pol II complex, phase probability distributions were compared directly. This was done by phasing each dataset individually, selecting a small subset (5 to 10) of well-phased reflections present in all datasets, and visually comparing plots of their phase probability distributions. This approach was able to indentify the same phase set as determined by another experimenter (Man Hee Suh) by visual map comparison.

The observation that combination of some datasets degrades, rather than improves, the quality of the combined phases was explained by the degree of overlap between their respective phase probability distributions. Experimental phase distributions are bi-modal. When combining two independent phase distributions (for example single crystal MAD or MIR) each probability distribution would have a peak at or near the true phase which would be reinforced. The second, or noise, peak of these distributions would not overlap, resulting in a lower phase probability in the region of the incorrect peaks. When combining probability distributions from the same type of phasing (for example, zinc SAS), overlap would occur at the true peak and at the noise peak. Phases from incompatible crystals would have little or no overlap, producing a flatter combined probability distribution. One possible theoretical objection to multi-

crystal phasing is that it combines probability distributions that are not fully independent. While this is a statistically valid objection, combination of non-independent information occurs relatively often in crystallography. For example, the phases combined in single crystal MAD are not fully independent: knowing the location of the peaks in one probability distribution allows one to predict that the other probability distribution will be higher in those regions, although it will only be higher at the true peak.

The third problem area that was identified was that the solvent flattening procedure depends critically on the use of a reasonably accurate solvent mask. When this was not the case, as in initial solvent flattening of the high resolution phase set, the map was of significantly poorer quality.

Although the focus of this work was to determine the feasibility of this method for phasing Pol II complexes using intrinsic zinc, it is not limited to such complexes. An analogous approach would be useable for determining experimental phases for other complexes with weak anomalous scatterers in cases where a partial structure is available.

The mechanistic implications of the newly built regions are discussed along with their refinement in Chapter 3.

## 2.4 Materials and Methods

### 2.4.1 Data Collection and Reduction

Diffraction data were collected at National Synchrotron Light Source (NSLS) beamline X25. X-ray fluorescence scans were used to confirm the presence of zinc

signal in the crystals, and identify appropriate wavelengths for data collection. The fluorescence scans were repeated during data collection, and wavelength was adjusted as needed. Anomalous datasets were collected in wedges of 20 degrees alternating with the same angular range that was offset by 180 degrees. Dispersive datasets were collected in wedges of 20 or 45 degrees, alternating between inflection and remote wavelengths, with the exception of the crystal named A2X11 (Table 2.1), which was collected in two passes.

Diffraction data were integrated and indexed using MOSFLM 6.23 (stand-alone version)(Leslie, 1992). Indexing was performed using spots from three non-consecutive images (*i*, *i*+45 degrees, *i*+90 degrees). Integration was performed following each wedge in the order in which it had been collected. MOSFLM was modified to allow storing of the orientation matrix at arbitrary times (originally, this version of MOSFLM would only store orientation matrices after indexing or cell refinement) to prevent loss of orientation parameters due to program crashes during integration. This facility was used to store the orientation matrix after integration of a wedge, and reload after integration of the alternate (other wavelength, or 180 degree offset) wedge. As discussed in the results section, this was found to produce improved results. With the CCP4-6.1.0 (Collaborative Computational Project, 1994) release, this facility was incorporated into the official version of MOSFLM (Leslie, 1992). Merging was performed in SCALA (Evans, 2006) from CCP4-4.2.2. Statistics for all datasets are shown in Table 2.1.

### 2.4.2 Zinc Site Location

Anomalous and dispersive difference maps were calculated in PHASES (Furey and Swaminathan, 1997) using model derived phases. Molecular replacement (MR) using

AMORE (Navaza, 1994), with either 12-subunit models (PDB IDs 1NIK(Bushnell and Kornberg, 2003) or 1NT9 (Armache et al., 2003)) or the 10-subunit model (PDB ID 1I6H (Gnatt et al., 2001) with nucleic acids removed), positioned the model in the same location of the unit-cell, allowing for symmetry relations and alternative origins available in $C222_1$. A hybrid model, consisting of the core 10 subunits from 1NIK and Rpb4/Rpb7 model from 1NT9, was used. The B-factor of the model was reset to the Wilson B-factor of the low resolution master dataset. This model was subjected to rigid-body refinement in CNS 1.1 (Brünger et al., 1998) against the native amplitudes from the low resolution master dataset, using the rigid-body domains as listed in Table 2.3 and visualized in Figure 2.7. Matthew's coefficient calculations showed that a single copy of Pol II in the asymmetric unit would correspond to a solvent content of 79%, and two copies would correspond to a solvent content of 59%. However, no MR solutions for a second molecule were found without substantial steric clashes.

### 2.4.3 Master Dataset and Reference Scaling

Two master datasets were produced, differing primarily in which anomalous dataset from crystal A3X7 was included, one limited to 6.5 Å dataset, and the other to 4.0 Å. Otherwise, the master datasets included all datasets collected. Each wedge of data was included in merging as an individual run within SCALA (from the CCP4-4.2.2 distribution), requiring minor modification of SCALA to deal with the amount of data input.

The appropriate master dataset was used as a reference for scaling of data from individual crystals prior to phasing. For dispersive datasets, the inflection was scaled to the master dataset; followed by scaling the appropriate remote dataset to the inflection, using CMBISO of PHASES in both cases. Anomalous datasets were also

Table 2.3: Definitions of Rigid Body Domains of 12-Subunit RNA Polymerase II

| Domain | Chain | Residues |
|---|---|---|
| Shelf | A | 808-876, 1058-1141, 1275-1395 |
| | F | 69-155 |
| Clamp | A | 1-346, 1396-1436 |
| | B | 1151-1244 |
| Jaw-Lobe | A | 1142-1274 |
| | B | 218-405 |
| | I | 1-39 |
| Core | C | All |
| | J | All |
| | K | All |
| | L | All |
| | A | 347-807, 1437-1733 |
| | B | 1-217, 406-1150 |
| | I | 1-122 |
| Rpb47 | D | All |
| | G | All |
| Rpb8foot | H | All |
| Rpb1foot | A | 877-1057 |
| Rpb5 | E | All |

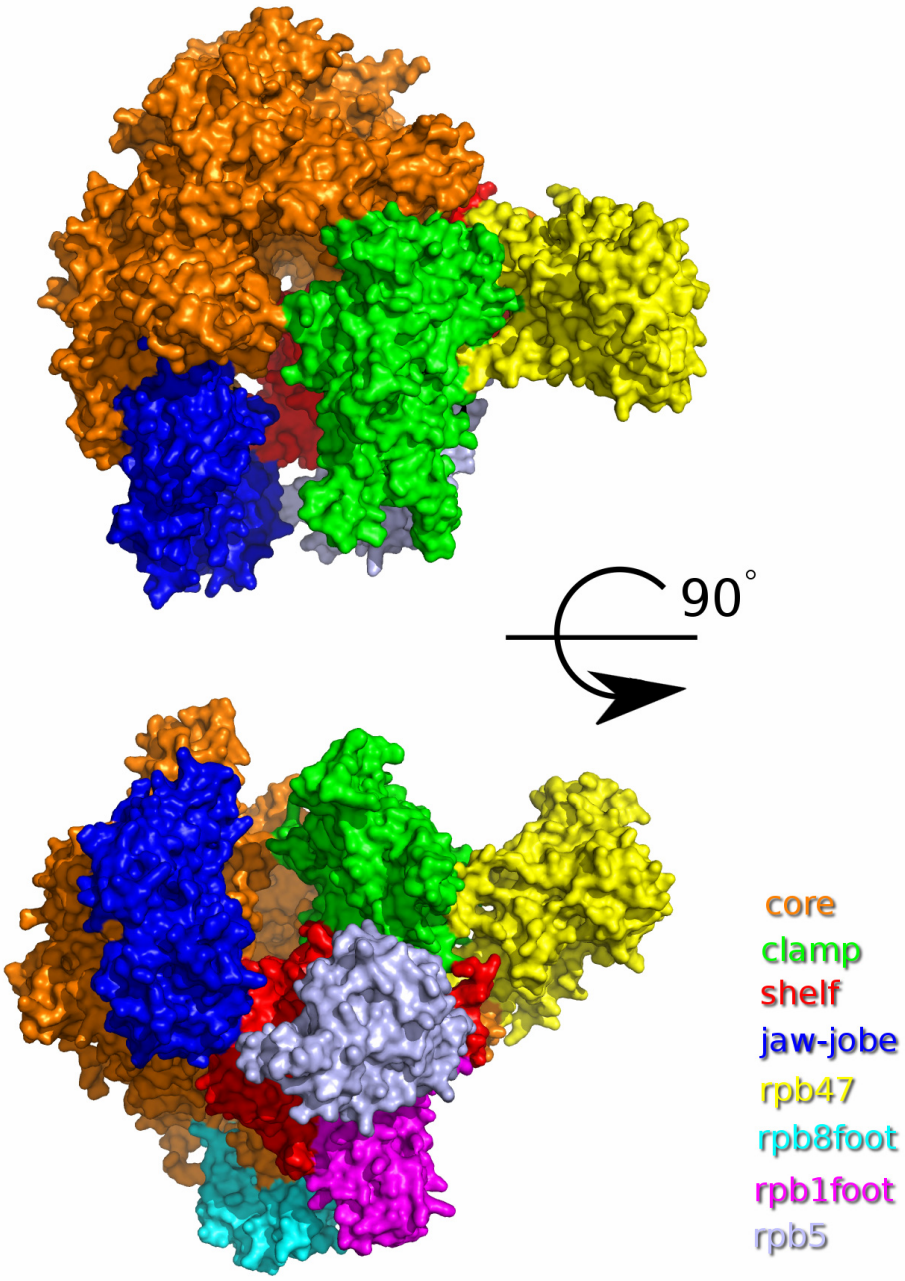


Figure 2.7: Rigid Body Domains of RNA Polymerase II

Location of rigid-body domains (listed in Table 2.3) displayed according to the color code in insert.

scaled to the master dataset using CMBANO, also of PHASES. The master amplitudes used for reference scaling were removed from reflection data files prior to phasing.

Although procedurally identical, the two-step scaling process is important because the stages are logically distinct. Internal scaling for dispersive datasets was required to ensure that the amplitude difference due to wavelength change were sufficiently accurate. Placing all datasets on a uniform scale before phasing was also required due to the method used to pick the 'native' amplitude produced by PHASIT. In PHASIT, the first amplitude seen by the program is used as the output 'native', or phased, amplitude. Variations in scale between individual datasets could produce additional noise. Although the scale factor between individual datasets is refined during phasing, numerical refinement is often less robust than one might like when using low resolution data.

### 2.4.4 Phase Calculation and Density Modification

Experimental phases were calculated using PHASIT of PHASES. Dispersive datasets were treated as SIR, with the inflection playing the role of the native, and inflection acting as the derivative (Ramakrishnan and Biou, 1997). Anomalous datasets were treated as native anomalous. Zinc sites for each derivative were treated as additional atom types, with $\Delta(f')$ and f'' values determined from X-ray fluorescence scans of each crystal. No external phase information was used during phase refinement. For multi-crystal phasing, data from each crystal was input as a new derivative, requiring minor changes of PHASIT to allow for the large number of additional atom types.

Solvent flattening was performed using PHASES for single crystal datasets, and each multi-crystal phase set as well. Un-flattened maps were not visually compared, due to their tendency to resemble solid blocks of electron density. A limited grid search for solvent flattening parameters solvent percentage and sphere radius was conducted, settling on a sphere radius of 13.0 Å and solvent percentage of 75%. Solvent flipping (Abrahams and Leslie, 1996) was implemented in PHASES by a modification of BNDRY. This was found to produce slightly worse, when using with γ-correction (Abrahams, 1997), to severely worse, without γ-correction, maps than solvent flattening. This approach was not used further.

### 2.4.5 Use of High-Resolution Data

Once a good phase combination from the multiple crystals had been identified using the available low-resolution dataset, the 4.0 Å SAS data from crystal A3X7 was additionally incorporated. The master dataset and reference scaling described above was repeated. The solvent mask generated by low resolution solvent flattening was converted to the high resolution grid, and used for solvent flattening without further modification. Although experimental phases covered the entire resolution range used, centric reflections beyond the resolution range covered by multiple crystals could not be phased by SAS, which was the only phase source available. For these reflections, phase extension was performed using the existing solvent mask.

### 2.4.6 Simulation Procedures

Standard structure factor calculation programs generally neglect the anomalous scattering factor f'', and do not allow for changes in f'. In order to produce calculated data incorporating these effects, I modified the FCAL2 program from B. C. Wang's group and used it to calculate and output $F^{(+)}$ and $F^{(-)}$ for all reflections within the

experimental resolution range. Anomalous scattering factors (f' and f'') for zinc atoms at inflection, peak and remote wavelengths were set according to values from X-ray fluorescence scans. Anomalous scattering factors for the remaining atom types, none of which were near an absorption edge at the wavelengths used, were calculated at the same wavelengths using CROSSEC (Cromer, 1983) of CCP4. This anomalous scattering information was incorporated into a parameter file, and three versions of the program were compiled (one for each wavelength with the appropriate anomalous scattering factors). Native amplitudes at each wavelength were generated by averaging $F^{(+)}$ and $F^{(-)}$. Eight sets of ideal anomalous data were calculated using the Pol II model containing one to eight Zn sites.

The next step was the simulation of errors for all of the simulated anomalous amplitudes. In order to produce simulated data of similar statistical characteristics to the experimental data, additional care was taken to insure the distribution of errors in the simulated data matched that in the experimental data. In preparation for error simulation, the simulated amplitudes ($F_{sim}$) were scaled to the multi-crystal master amplitudes, and $\sigma_{F,sim}$ was set to $\sigma_{F,master}$. $F_{sim}$ and $\sigma_{F,sim}$ were converted to intensities ($I_{sim}$ and $\sigma_{I\_master}$), and the Box-Muller procedure (Box and Muller, 1958) was used to apply normally distributed errors, as opposed to the uniformly distributed numbers typically obtained from random number generation routines. After errors had been applied to $I_{sim}$, the data were reconverted to amplitudes. This error generation step was repeated for each wavelength of each simulated crystal used, to produce independent errors. The number of crystals was varied from 1 to 10 for each set of simulated data.

For each fixed number of crystals and number of zinc sites, the simulated data were phased and subjected to solvent flattening using the same parameters as with the experimental setup, with the exception of placing the Zn sites in a different alternative origin in order to avoid accidentally mixing simulated and experimental data. The maps were visually checked in PYMOL (DeLano, ). Real space statistics (map correlation coefficient (Lunin and Woolfson, 1993) and real-space R-factor (Drenth, 1999)) were calculated for each map, relative to the $F_C$ map, as an additional means of comparison. This procedure of simulation, processing and evaluation was repeated twice for additional consistency.

# CHAPTER 3: REFINEMENT OF POL II MODEL USING ZINC ANOMALOUS SCATTERING AS ADDITIONAL DATA

## 3.1 Introduction

The end goal of X-ray crystallography is the establishment of a model that accurately reflects the structure of the molecule(s) in the crystal. An initial model can be generated by several different approaches: manual tracing based on visual inspection of the electron density map, automated model building programs, or by adapting of a model from another source for part(s) or all of the molecules in question. This preliminary model generally provides only an approximate match to the data, and whenever possible is optimized, or refined, in order to better match the available data. The refinement of a model against the observed data, and its subsequent validation, is the final stage of crystallographic analysis. For the majority of X-ray structures that are determined at moderate to high resolutions, this process has become relatively standardized and over time has received a decreasing amount of attention and description.

The refinement of structures at low resolutions, however, presents distinct problems that have as yet not been as clearly resolved as work done at higher resolutions (e.g. better than 3.0 Å). Low resolution maps, even those of high quality, by definition do not provide the same level of detail available at higher resolutions. As discussed in Chapter 2, the number of experimental observations available is substantially reduced at lower resolutions, and the effects of phase error are relatively more severe. Crystals of larger macromolecules often diffract weakly, and produce data only to low

resolutions. In addition, a large macromolecule by definition contains more atoms than a smaller one, resulting in a concomitant increase in the number of parameters to be refined. A survey of published Pol II models indicates that the standard refinement protocols tend not to produce high quality models at low resolution. Table 3.1 shows some relevant statistics from representative Pol II models. On the other hand, structures determined at high resolution, for example 1Y14 (Armache et al., 2005) and 1TWF (Westover et al., 2004), have reasonable values for geometric and refinement statistics (Table 3.1 B). At lower resolutions, these statistical measures show a trend of reduced model quality. In some cases, severe geometrical distortions are visually apparent (one example shown in Figure 3.1).

There are several potential pitfalls in refinement, particularly when using low resolution data. The most significant of these is over-fitting the model. In order for a model to represent a robust solution to the optimization problem that is refinement, the formulation used must be over-determined, with many more experimental observations than the parameters of the model to be determined. With the exception of ultra-high resolution (better than 1.0 Å) crystals, this condition is not met by the number of amplitudes that can be collected from protein crystals. Fortunately, protein structures are composed of a relatively limited number of building blocks (amino acid residues, small molecule ligands, ions, etc) whose structures have previously been established. The structures of these components are, with few exceptions, known to high accuracy, and the geometric parameters (bond angles and lengths, chiral centers, etc) of these components can be used as constraints on the model during refinement, effectively increasing the number of observations. Even when making use of geometric observations, the observation to parameter ratio drops substantially as resolution decreases (Table 3.1A and Table 3.2). Crystallographic refinement

Table 3.1A: Observation to Parameter Ratios for Representative RNA Polymerase Models

Ratio 1 = observations (reflections) / parameters
Ratio 2 = observations (reflections + restraints) / parameters

| PDB ID | Resolution (Å) | Description | Ratio 1 | Ratio 2 |
|---|---|---|---|---|
| 1Y14 | 2.3 | Rpb4/7 alone | 2.161 | 3.891 |
| 1TWF | 2.3 | 10-subunit Pol2 + UTP | 2.085 | 3.889 |
| 2E2J | 3.5 | 10-subunit Pol2 elongation + GMPCP | 0.720 | 2.498 |
| 1WCM | 3.8 | 12-subunit Pol2 | 0.984 | 2.805 |
| **3FKI** | 3.8 | 12-subunit Pol2 (anomalous included in refinement) | 2.083 | 3.689 |
| 2B8K | 4.15 | unrefined 12-subunit RNA Polymerase II | 0.622 | 0.804 |

"R_Tested", rms bond, angle and chiral values are from REFMAC5. Ramanchandran and Cβ deviations are from MOLPROBITY. Real space statistics ("CC,full", "CC,masked", "RSR,full" and "RSR,masked" were calculated as described in text. Final model from this refinement is PDB ID **3FKI**, indicated in *bold*.

Table 3.1B: Geometric Statistics for Representative RNA Polymerase Models

| PDB ID | rms Bond | rms Angle | rms Chiral | Ramachandran Favored (%) | Ramachandran Outliers (%) | Rotamer Outliers (%) | Cβ Deviations >0.25 Å |
|---|---|---|---|---|---|---|---|
| 1Y14 | 0.007 | 1.378 | 0.100 | 94.49 | 1.07 | 2.87 | 0 |
| 1TWF | 0.008 | 1.311 | 0.089 | 89.55 | 1.68 | 4.86 | 0 |
| 2E2J | 0.012 | 1.431 | 0.119 | 85.41 | 3.52 | 10.89 | 17 |
| 1WCM | 0.009 | 1.538 | 0.103 | 72.46 | 7.46 | 8.55 | 3 |
| **3FKI** | 0.005 | 0.855 | 0.057 | 87.84 | 2.87 | 3.74 | 2 |
| 2B8K | 0.010 | 1.732 | 0.111 | 72.22 | 7.69 | 8.64 | 7 |

Table 3.1C: Refinement Statistics for Representative RNA Polymerase Models

| PDB ID | R_Reported | $R_{free}$_Reported | R_Tested | CC, full | CC, masked | RSR, full | RSR, masked |
|---|---|---|---|---|---|---|---|
| 1TWF | 0.247 | 0.294 | 0.310 | 0.688 | 0.774 | 0.500 | 0.200 |
| 1WCM | 0.257 | 0.285 | 0.296 | 0.698 | 0.801 | 0.501 | 0.182 |
| 1Y14 | 0.228 | 0.274 | 0.240 | 0.778 | 0.838 | 0.409 | 0.157 |
| 2E2J | 0.242 | 0.306 | 0.313 | 0.657 | 0.762 | 0.516 | 0.202 |
| **3FKI** | 0.291 | 0.315 | NA | 0.774 | 0.848 | 0.446 | 0.148 |

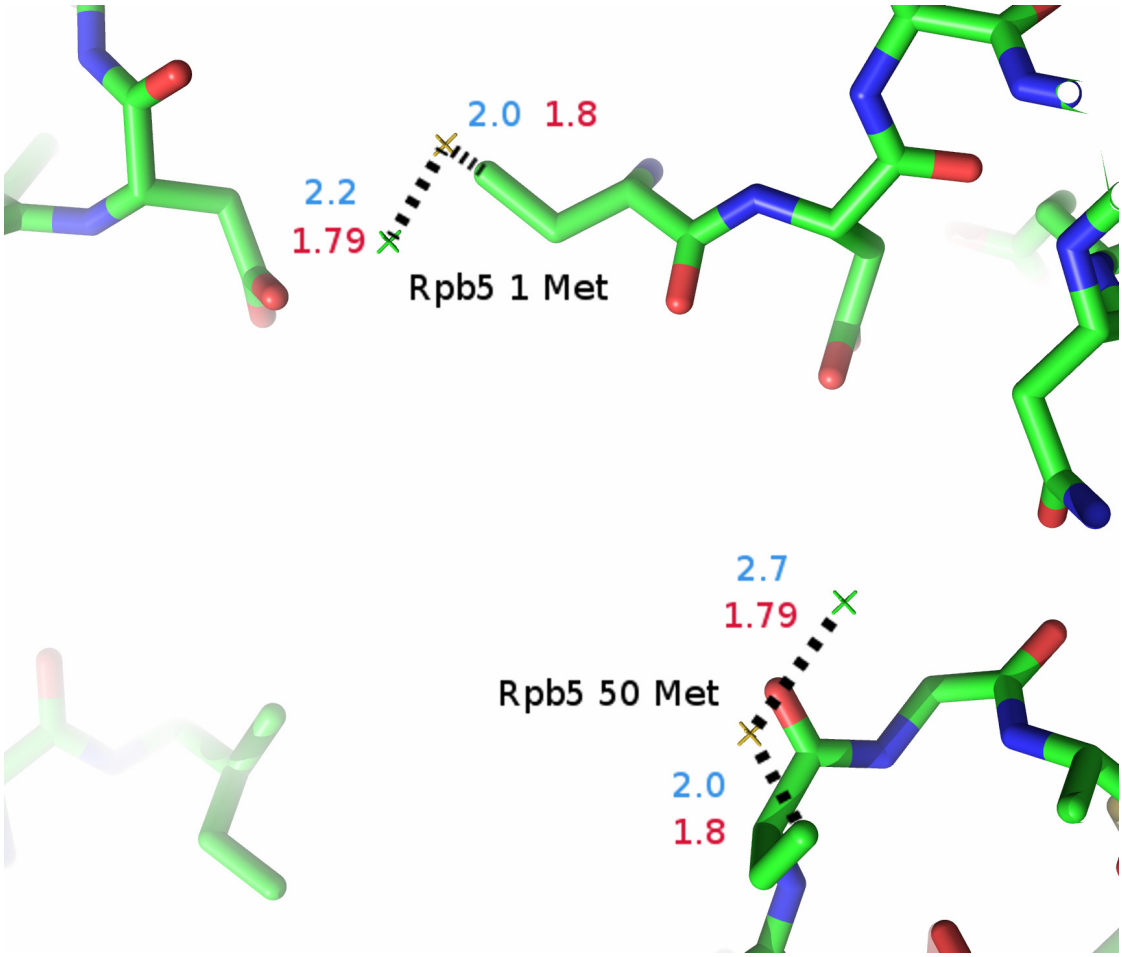


Figure 3.1: Examples of Geometric Distortions Observed in Poorly Refined Models

Measured bond distances are indicated as *blue text*, dictionary values for bond distances in standard amino-acids (from CCP4-6.1.0 monomer library) are shown in *red text*. Close up of residues E1MET and E50MET from PDB entry 1TWC. Model shown in *stick* representation: oxygen in *red*, nitrogen in *blue*, carbon in *red*, sulfur in *yellow*. Atoms with large distance deviations are shown as *crosses*.

Table 3.2: Ideal Observation to Parameter Ratios for Different Refinement Approaches

*Bold* entries highlight observation-to-parameter ratio at 3 Å. *Underlined* entries represent equivalent values for alternate refinement formulations. The refined Pol II model was used for determining the number of parameters; different columns show different B-factor formulations. "B_atomic" used individual isotropic B-factors. "TLS" used 20 TLS groups with a single global B-factor. "B_over" used a single isotropic B-factor. The number of reflections was determined using all possible reflections for the experimental unit-cell and space-group.

| **Reflections / Parameters** | | | | | | |
|---|---|---|---|---|---|---|
| Target: | Normal Amplitudes | | | Normal plus Anomalous Amplitudes | | |
| Resolution (Å) | B_atomic | TLS | B_over | B_atomic | TLS | B_over |
| 1.00 | 52.49 | 69.84 | 69.99 | 105.06 | 139.77 | 140.06 |
| 1.50 | 15.62 | 20.78 | 20.82 | 31.23 | 41.54 | 41.63 |
| 2.00 | 6.62 | 8.81 | 8.83 | 13.23 | 17.61 | 17.64 |
| 2.50 | 3.40 | 4.53 | 4.54 | 6.80 | 9.05 | 9.07 |
| 2.80 | 2.43 | 3.23 | 3.24 | 4.85 | 6.45 | 6.47 |
| **3.00** | **1.98** | 2.63 | 2.63 | 3.95 | 5.26 | 5.27 |
| 3.25 | 1.56 | 2.07 | 2.08 | 3.11 | 4.14 | 4.15 |
| 3.50 | 1.25 | 1.66 | 1.67 | 2.50 | 3.32 | 3.33 |
| 3.80 | 0.98 | 1.30 | 1.30 | 1.96 | 2.60 | 2.61 |
| 4.00 | 0.84 | 1.12 | 1.12 | 1.68 | 2.23 | 2.24 |
| 4.25 | 0.70 | 0.93 | 0.94 | 1.40 | 1.87 | 1.87 |
| 4.50 | 0.59 | 0.79 | 0.79 | 1.18 | 1.57 | 1.58 |
| 4.75 | 0.50 | 0.67 | 0.67 | 1.01 | 1.34 | 1.34 |
| 5.00 | 0.43 | 0.58 | 0.58 | 0.87 | 1.15 | 1.15 |
| 6.00 | 0.25 | 0.34 | 0.34 | 0.50 | 0.67 | 0.67 |
| 7.00 | 0.16 | 0.21 | 0.21 | 0.32 | 0.43 | 0.43 |
| 8.00 | 0.11 | 0.14 | 0.14 | 0.22 | 0.29 | 0.29 |
| | | | | | | |
| **(Reflections and Restraints ) / Parameters** | | | | | | |
| Target: | Normal Amplitudes | | | Normal plus Anomalous Amplitudes | | |
| Resolution (Å) | B_atomic | TLS | B_over | B_atomic | TLS | B_over |
| 1.00 | 54.55 | 72.58 | 72.73 | 107.11 | 142.50 | 142.80 |
| 1.50 | 17.67 | 23.51 | 23.56 | 33.28 | 44.28 | 44.37 |
| 2.00 | 8.67 | 11.54 | 11.57 | 15.29 | 20.34 | 20.38 |
| 2.50 | 5.46 | 7.26 | 7.28 | 8.86 | 11.78 | 11.81 |
| 2.80 | 4.48 | 5.96 | 5.98 | 6.91 | 9.19 | 9.21 |
| 3.00 | **4.03** | 5.36 | 5.38 | 6.01 | 7.99 | 8.01 |
| 3.25 | 3.61 | 4.81 | 4.82 | 5.17 | 6.88 | 6.89 |
| 3.50 | 3.30 | 4.40 | 4.41 | 4.55 | 6.06 | 6.07 |
| 3.80 | 3.03 | 4.04 | 4.04 | 4.01 | 5.34 | 5.35 |
| 4.00 | 2.89 | 3.85 | 3.86 | 3.73 | 4.97 | 4.98 |
| 4.25 | 2.76 | 3.67 | 3.68 | 3.46 | 4.60 | 4.61 |
| 4.50 | 2.65 | 3.52 | 3.53 | 3.24 | 4.31 | 4.32 |
| 4.75 | 2.56 | 3.41 | 3.41 | 3.06 | 4.07 | 4.08 |
| 5.00 | 2.49 | 3.31 | 3.32 | 2.92 | 3.89 | 3.89 |
| 6.00 | 2.31 | 3.07 | 3.08 | 2.56 | 3.40 | 3.41 |
| 7.00 | 2.22 | 2.95 | 2.95 | 2.37 | 3.16 | 3.17 |
| 8.00 | 2.16 | 2.88 | 2.88 | 2.27 | 3.02 | 3.03 |

algorithms incorporating such constraints, in the form of a chemical dictionary of known monomers, have been developed, allowing improved refinement of moderate resolution structures. However, for crystals diffracting to lower resolutions, such as the 12-subunit Pol II (up to 3.8 Å), the traditional treatment is insufficient. Additional measures need to be taken in order to ensure meaningful refinement results.

In order to overcome this problem with refining the Pol II model, an alternative refinement approach was investigated. The main feature of this approach was to maximize the observation to parameter ratio to the extent possible at the available resolution. This was made possible by two technical advances: 1) the development of a refinement target directly incorporating $F^{(+)}$ and $F^{(-)}$ (Skubák et al., 2004), and 2) the use of TLS groups for modeling temperature factors (Howlin et al., 1989; Schomaker and Trueblood, 1968) in place of atomic temperature factors.

The direct use of SAS ($F^{(+)}$ and $F^{(-)}$ ) data in the target function (referred to here as the FPFM target) results in a large increase in the number of experimental observations usable in refinement. As discussed earlier, the amplitudes used in refinement usually neglect anomalous scattering. Treating each member of a Bijvoet pair as an independent observation increases the number of observations by roughly twice as many. Although $F^{(+)}$ and $F^{(-)}$ are typically correlated, the anomalous difference which is implicitly incorporated into the refinement target is not correlated with the mean amplitude. For crystals in a space group containing centric reflections, the phase restriction at these amplitudes constrains the amplitudes of both members of the Bijvoet pair to be equal even in the presence of significant anomalous scattering. In addition, if the dataset used in refinement is not fully complete, there will most likely be some reflections where only one member of the Bijvoet pair has been observed.

Therefore, in practice the increase in the number of effective observations for refinement is slightly less than a complete doubling. This approach has been proven to be effective in test cases (Skubak et al., 2004; Skubák et al., 2005), but had not been evaluated for large structures such as Pol II. In particular, diffraction data available were limited to lower resolution than present in test cases; additionally, the available anomalous signal from Zn present in Pol II data was weaker than those in published test cases.

The use of Translation-Libration-Screw Rotation, or TLS (Schomaker and Trueblood, 1968), groups assisted in the improvement of the observation to parameter ratio by decreasing the number of parameters used in the model. As discussed in Introduction, a crystallographic model consists of the positions of the atoms in the unit-cell as well as their degree of mobility, as described by their temperature factors. The use of atomic temperature factors for individual atoms would have increased the number of parameters by approximately 30,000. However, use of a single overall temperature factor would not accurately reproduce the varying degrees of thermal motion of the different parts, or domains, of Pol II. As a compromise approach, TLS groups were used as group anisotropic temperature factors. This allowed for improved modeling of thermal motion within molecule while using only approximately 200 temperature factor parameters.

## 3.2 Results

### 3.2.1 Low-Resolution Refinement of 12-Subunit Pol II With the Aid of Zn SAS Data

Using this approach, the observation-to-parameter ratio for the 12-subunit Pol II model at 3.8 Å is comparable to that obtained for the 10-subunit Pol II model at 2.3 Å (Westover et al., 2004), which is the highest resolution dataset publicly available as of summer 2009 (Table 3.1 A). Consistent with the improvement in this ratio, geometric and stereo-chemical statistics showed improvement relative to other 12-subunit models reported at comparable, or better, resolution (Table 3.1 B). These improvements were confirmed by comparison of the electron density maps calculated using the refined model with those calculated using the unrefined model or other published Pol II models (Figures 3.2 and 3.3). The improved map quality also manifested in an improvement in real space statistics (map correlation coefficient and real-space R-factor); particularly for the protein-occupied region (masked statistics) (Table 3.1 C).

The reciprocal space statistics ($R_{work}$ and $R_{free}$) improved during the course of refinement. However, these R-factors did not show improvement in comparison with the published R-factors for the other Pol II structures. These statistics were recalculated for the previously published structures, using the deposited models and structure factor amplitudes. These calculations showed a general trend for the recalculated R-factors to be higher than those reported in the literature (Table 3.1 C). This trend was confirmed using R-factors calculated by the Uppsala Electron Density Server (Kleywegt et al., 2004), which also returned values higher than those reported.

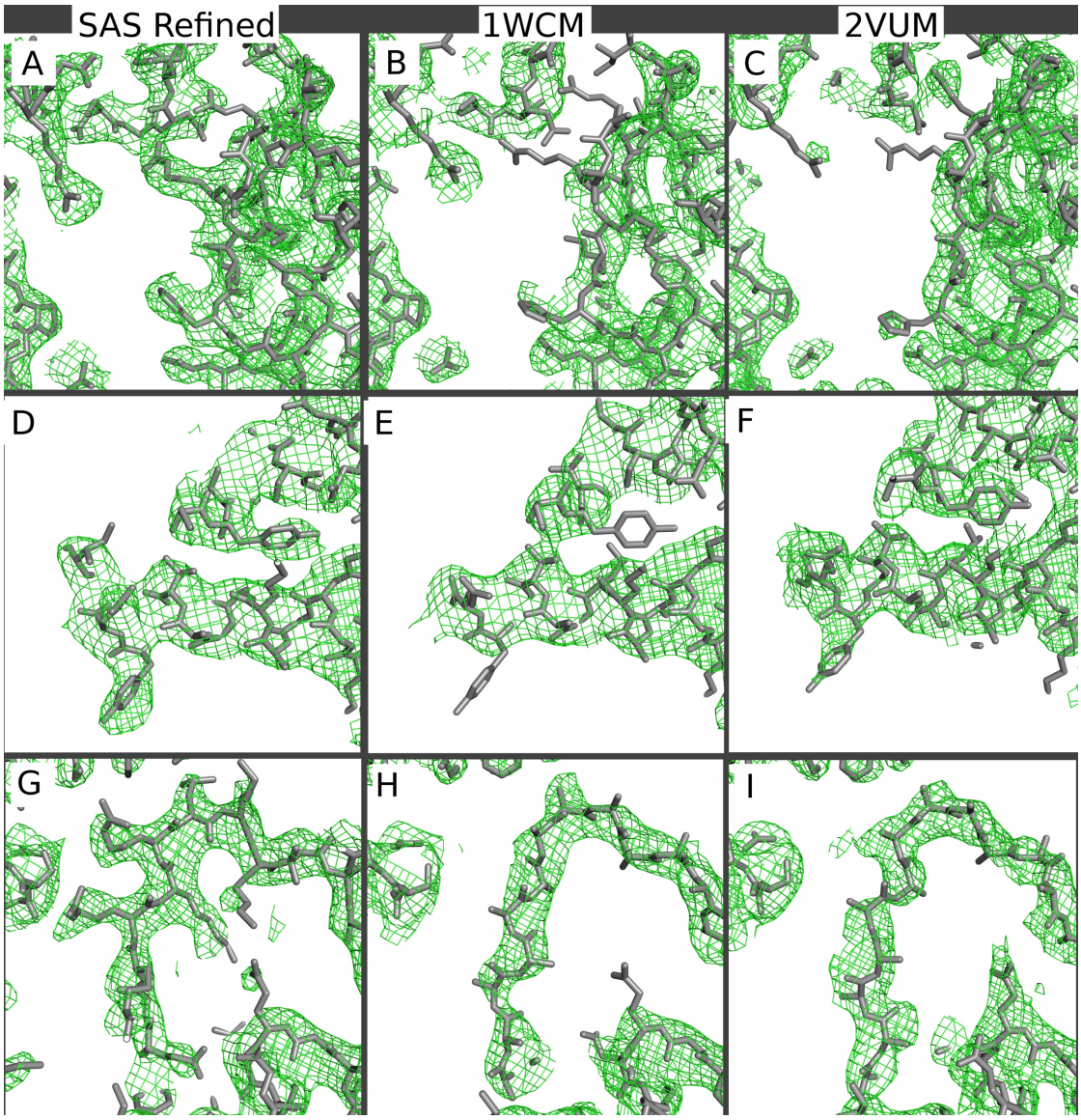


Figure 3.2: Comparison of Refined $2F_o$-$F_c$ Maps

$\sigma_A$ weighted $2F_O$-$F_C$ composite omit maps (Bhat, 1988) contoured at 1 σ are shown as *green mesh*, respective Pol II models shown as *grey sticks*. 1WCM (3.8 Å) and 2VUM (3.4 Å) were refined without anomalous data. Top row (A, B and C) are centered on one interface between Rpb1 and Rpb2 (A63 and B884). Middle row (D, E and F) are centered on Rpb2 (B666 and B679). Bottom row (G, H and I) centered on a loop in Rpb4 (D9 through D20).

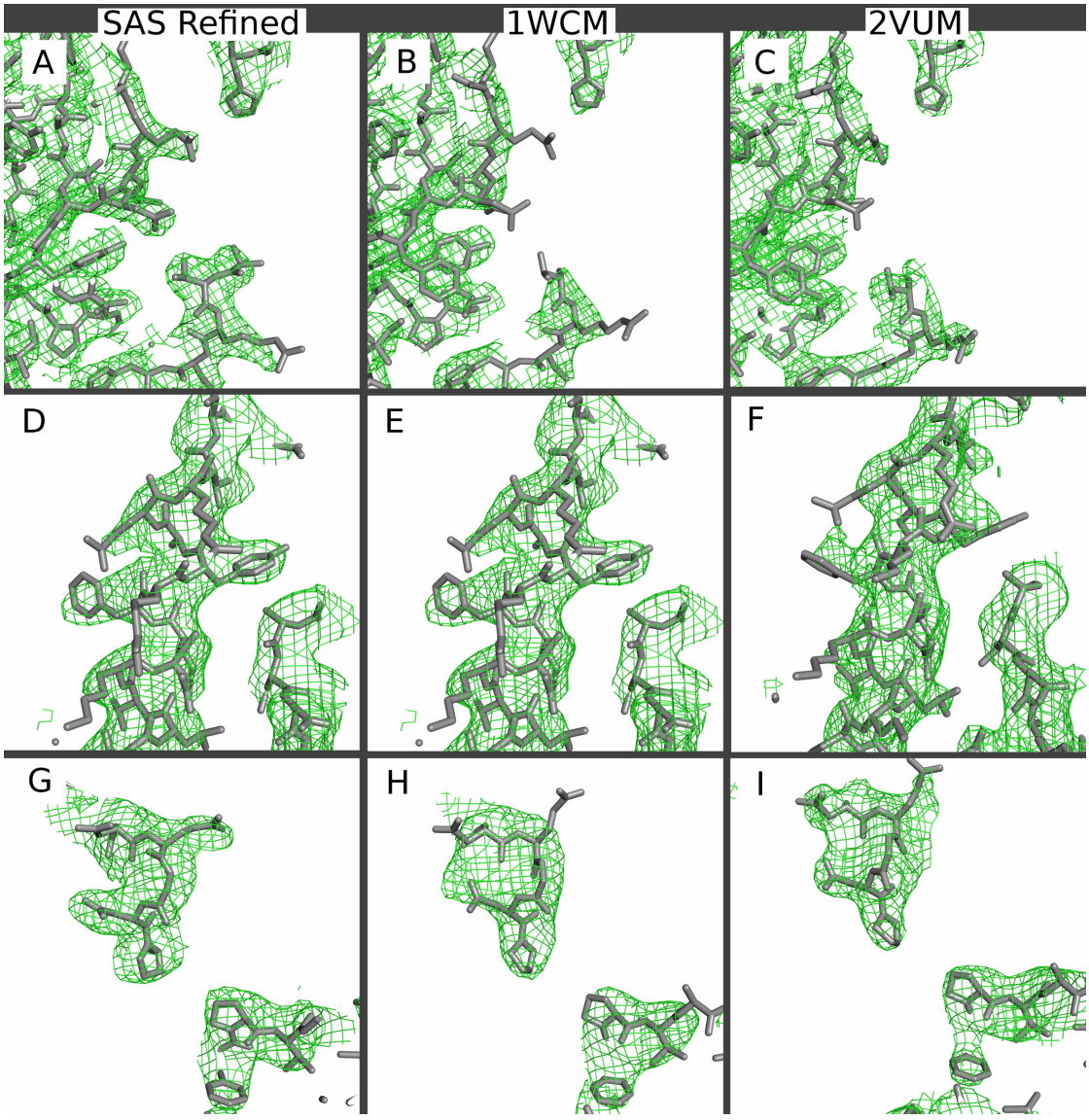


Figure 3.3: Comparison of Refined $2F_o$-$F_c$ Maps

Maps and models are displayed as in Figure 3.2. Top row (A, B, C) centered on Rpb8 (H75 and H137). Middle row (D, E and F) centered on Rpb2 (B429 and B431). Bottom row (G, H and I) centered on Rpb3 and Rpb8 (C218 and H48).

### 3.2.2 Biological Implications from the Improved 12-Subunit Pol II Structure

As discussed in Chapter 2, the experimental map allowed modeling of three additional regions of 12-subunit Pol II, but these preliminary models were not refined in the initial study. In the course of SAS-assisted refinement, additional density was observed in difference Fourier maps which allowed building of several additional regions. The majority of these newly modeled regions are relatively small, and have no biological function ascribed to them. Nonetheless, several of these regions have been implicated in transcriptional roles (Figure 3.4).

As mentioned in Chapter 2, Fork Loop 1 (Figure 1.4) was initially modeled on the basis of the multi-crystal Zn-MAD experimental map (Figure 3.4). The backbone conformation shifted substantially during the course of the refinement, with the refined backbone 2.5 to 5 Å further away from the Rudder element (Figure 3.5). However, as a result of the side chains in the refined model, and absent in the previous unrefined model, the average distance between Fork Loop 1 and the rudder is approximately the same. Although Fork Loop 1 has been observed in several other Pol II structures containing nucleic acids bound to the polymerase, this model is the first definition of Fork Loop 1 in free Pol II.

Fork Loop 2 (Figure 1.4) was not observed in the multi-crystal Zn-MAD experimental map. However, omit density for this region appeared as the refinement progressed, allowing it to be modeled.

The Protrusion domain (Figure 1.4) is one of two stalk-like projections in Rpb2 that rise from the active site cleft. The experimental multi-crystal map revealed electron

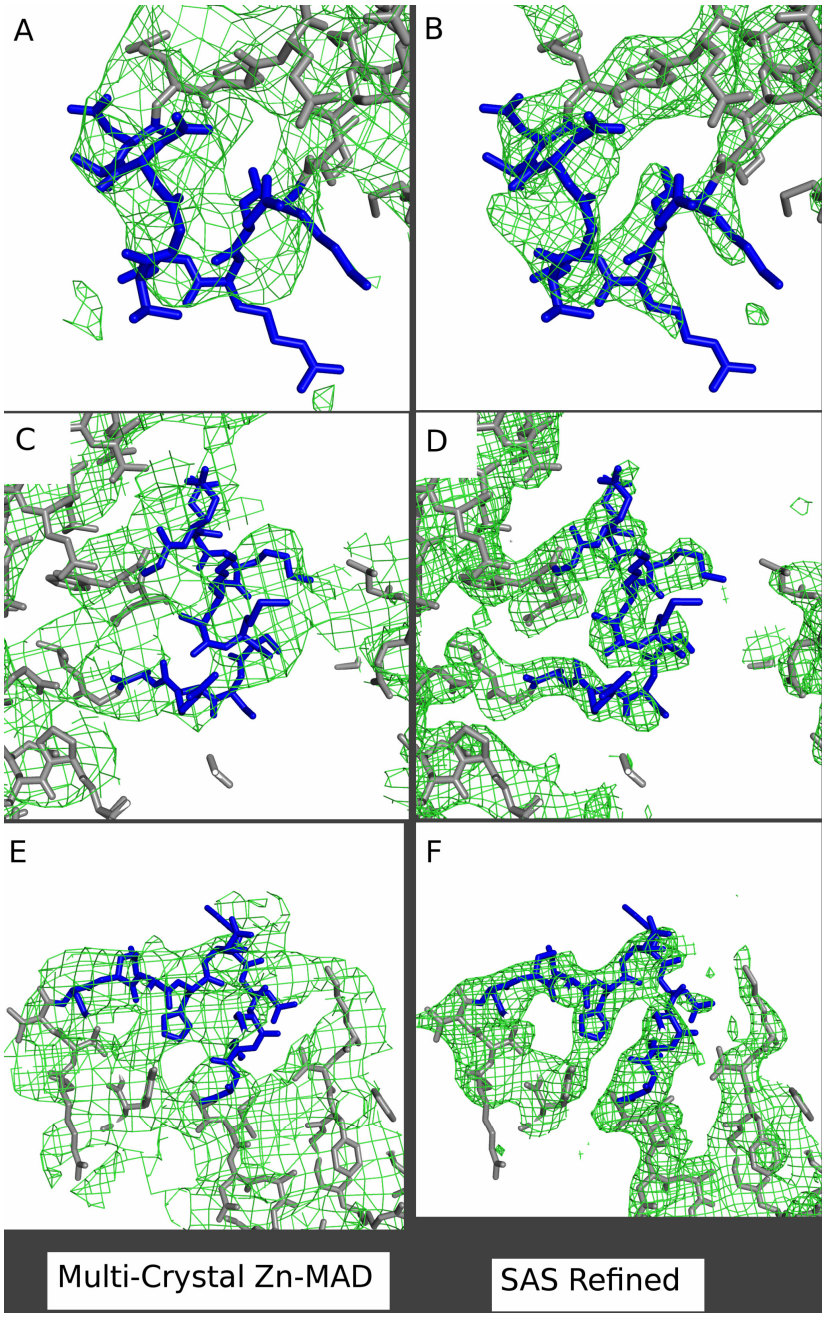


Figure 3.4: Experimental and Model Phased Maps for Regions with Biological Implications

Clamp Top of Rpb1 (A187-A195) (experimental, A; model phased, B); Fork Loop 1 of Rpb2 (B462-B481) (experimental, C; model phased, D); Protrusion of Rpb2 (B437-B446) (experimental, E; model phased, F). Existing models are shown in *grey*, new regions in *blue*, electron density contoured at 1.0 σ in *green*.

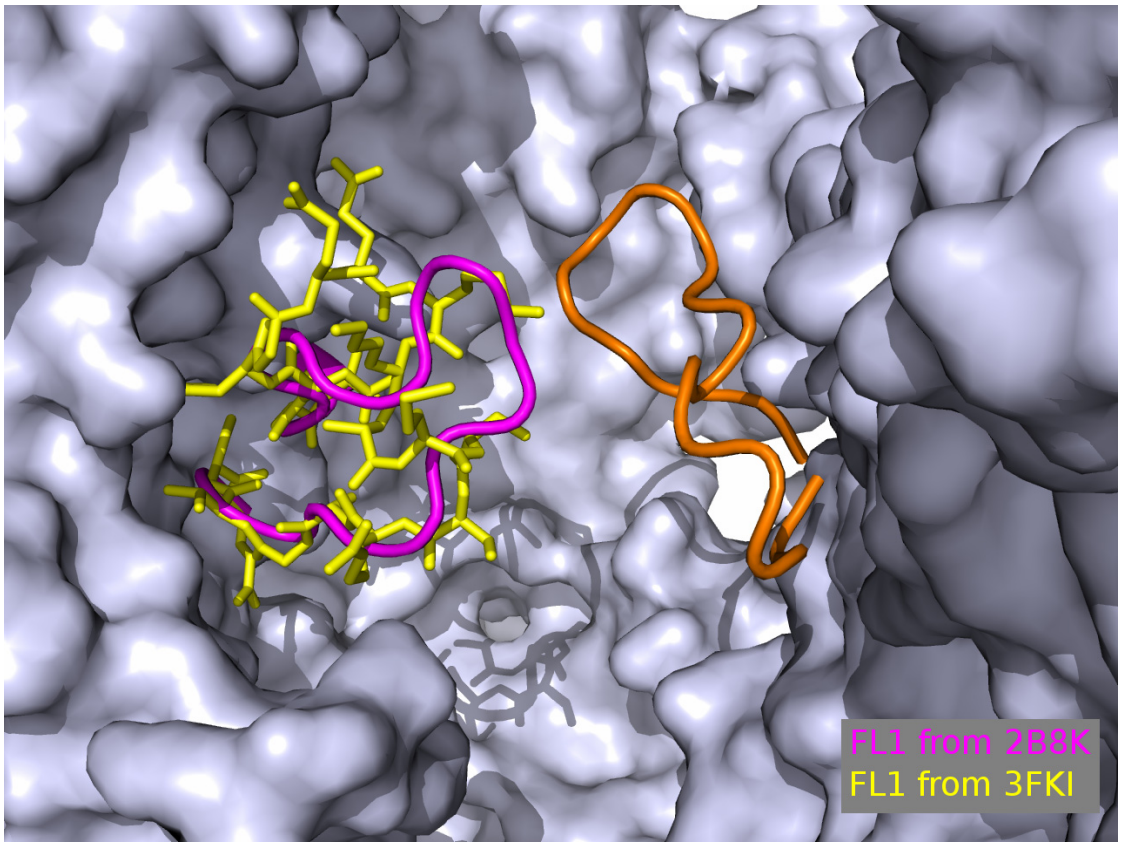


Figure 3.5: Change in Fork Loop 1 Conformation Before and After Refinement

Unrefined Fork Loop 1 Model as *Magenta Ribbon* (2B8K)
Refined Fork Loop 1 Model as *Yellow Sticks* (3FKI)
Rudder shown as *Orange Ribbon* (3FKI)

density for this region, which allowed for initial modeling of a portion of the Protrusion. Further modeling was made possible by maps based on the refined model (Figure 3.4). This region was not observed in other published Pol II structures.

As mentioned in Introduction, the Clamp domain (Figure 1.4) is located on the opposite side of the cleft from the Protrusion domain. Weak additional density was observed for this region in the multi-crystal Zn-MAD experimental map, although it was not modeled during that study. The omit density map after refinement allowed modeling of this region (Figure 3.4), which was absent in previously published Pol II structures.

## 3.3 Discussion

### 3.3.1 Crystallographic Discussion

The improvements in model and map quality demonstrate that including anomalous data in the refinement target function produces a Pol II model of higher quality than those refined using traditional refinement targets at comparable resolution. The 12-subunit Pol II structure refined by this method is of comparable quality to the 10-subunit Pol II structures refined at higher resolution (3.4 Å). The improvement in model quality was also illustrated by the observation that solvent flattening of the unrefined model phases, using the final mask used for multi-crystal phasing, did not show any of the additional regions of density that were observed in either the experimental map or maps phased with the refined model.

In some ways this improved result is expected, as increasing the observation-to-parameter ratio is expected to result in a more accurate determination of the model

parameters, which in turn typically results in a higher-quality model (and therefore more accurate model-derived phases). However, at the onset of this experiment it was not known if this improvement would be obtained, for several reasons. The increase in the number of observations depends on the presence of a sufficient Zn anomalous signal within a single anomalous dataset. As discussed in Chapter 2, a single SAS dataset was not sufficient for experimental phasing of Pol II. Insufficient anomalous signal would have meant that the number of independent observations did not in fact increase in the data used for refinement.

In addition, earlier efforts to incorporate model derived phases in more traditional methods were unhelpful. Direct phase combination of experimental and unrefined-model phases did not produce improved maps, despite adjustment of combination weights. Earlier attempts to incorporate phase information into refinement by using the MLHL refinement target (Pannu et al., 1998) also failed to produce any useful results in the case of 12-subunit Pol II, although others have reported using this approach successfully for low resolution refinement (DeLaBarre and Brunger, 2006). One common factor shared by both of these approaches is that the experimental phase information was used in the form of Hendrickson-Lattman (HL) phase coefficients (Hendrickson and Lattman, 1970). In contrast, by combining SAS-phasing directly with information from the model (Skubák et al., 2004), there is no intermediate representation of the experimental phase information. The HL representation of phase probabilities is only exact in the case of a uni- or bi-modal phase probability distribution. In the case of combination of a uni-modal model phase probability distribution and a bi-modal phase probability distribution from SAS, the HL representation would no longer be exact. It is possible that avoiding this intermediate

approximation allowed successful use of both model and experimental information through the FPFM target function.

In the absence of a theoretical understanding of the degree of over-determination, it is not straightforward to determine what the minimum resolution is required for valid results for a given refinement formulation. However, comparison of observation-to-parameter ratios for different refinement formulations at different resolutions provides an empirical way to estimate the minimum resolution required for meaningful refinement using different approaches. For a given structure, the number of available observations varies according to the resolution of the data, the number of restraints, and the decision to use anomalous amplitudes. The number of parameters varies mainly with the type of B-factor parameterization used. Table 3.2 lists the observation-to-parameter ratio for several different formulations across a wide range of resolutions using the SAS refined model, and assuming all possible reflections within the resolution limit were available for refinement. Taking isotropic atomic B-factor refinement against native amplitudes at 3.0 Å as a point of comparison (Table 3.2, upper section), the refinement approach described here could be expected to produce meaningful results for data limited to 4.75 Å (Table 3.2, lower section).

## 3.3.2 Biological Discussion

### 3.3.2.1 Fork Loop 1

Fork Loop 1 is composed of residues 461-480 in Rpb2 (the second largest Pol II subunit). Fork Loop 1 is located above the floor of the cleft, and approximately two-thirds of the way back towards the wall at the end of the cleft (Figure 1.4). It interacts with the Rudder domain of Rpb1 310-324 (Figure 3.6). Fork Loop 1 and the Rudder

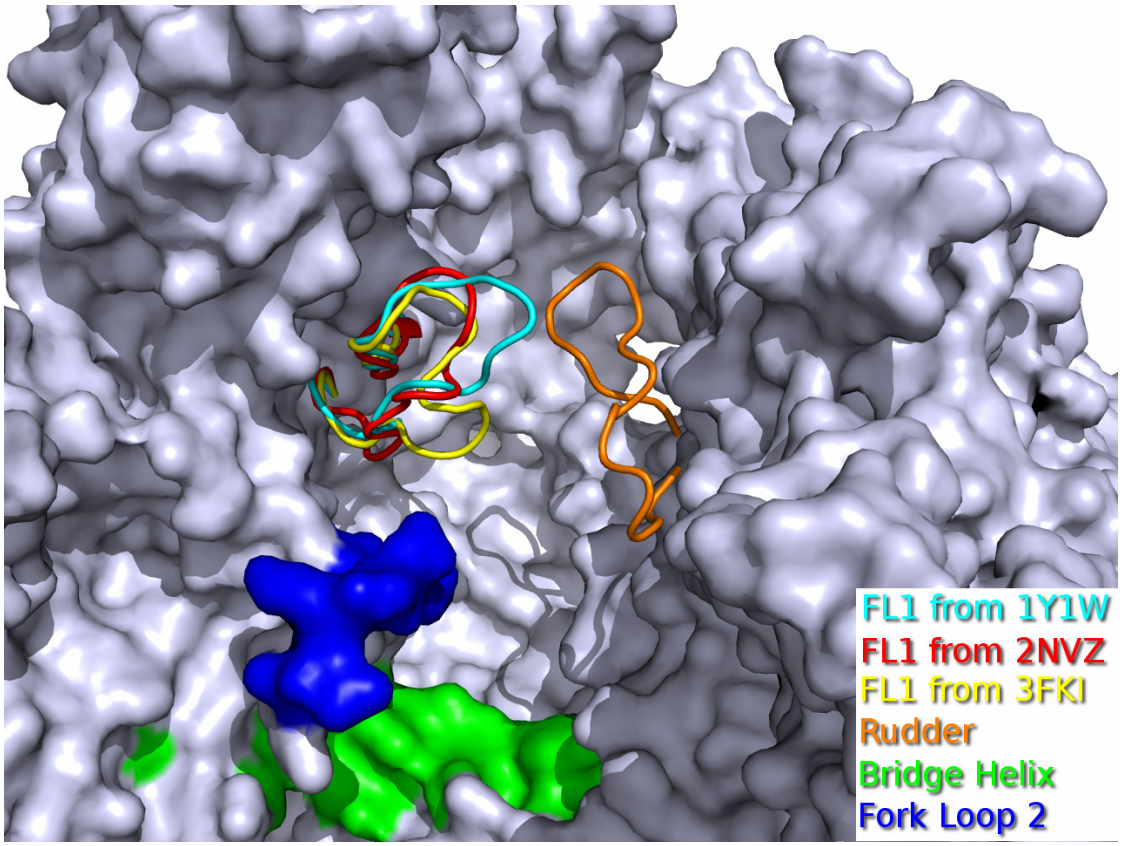


Figure 3.6: Representative Fork Loop 1 Conformations

The open conformation is represented by 2NVZ, shown in *red*. The closed conformation is represented by 1Y1W, shown in *cyan*. The free conformation, after SAS refinement, is labeled by PDB ID 3FKI, shown in *yellow*. Rudder (*orange ribbon*), Bridge Helix (*green surface*), and Fork Loop 2 (*blue surface*) are also shown for structural context and orientation.

both interact with the DNA/RNA hybrid in the barrel region of the cleft, suggesting that they are responsible for stabilizing the binding of the hybrid, and may play a role in maintaining strand separation. Biochemical experiments have shown that deletion

of Fork Loop 1 results in a loss of transcriptional activity in mammalian Pol II (Jeronimo et al., 2004). However, deletion studies in *Pyrococcus furiosus* indicated that Fork Loop 1 is not required for transcriptional initiation or elongation (Naji et al., 2008). In addition, prokaryotic RNA Polymerases lack a region corresponding to Fork Loop 1. This discrepancy could be explained by comparison of the sizes of the gap between the Rudder in prokaryotic and eukaryotic structures. In prokaryotic RNA Polymerase, the size of the gap between the rudder and opposing side of the cleft is approximately 3.6-4 Å. This compares to approximately 3.9 Å in Pol II when Fork Loop 1 is localized, versus approximately 7 Å if Fork Loop 1 were not present. Under the relatively reasonable assumption that the size of the DNA/RNA hybrid does not vary significantly between organisms, the lack of a requirement for a Fork Loop 1 element in some species could be attributed to the narrower opening between the Rudder and the opposite side of the cleft.

In previous structures of Pol II, Fork Loop 1 has only been observed in the presence of nucleic acids, suggesting that this region is only ordered in the presence of DNA/RNA. The observation of this region in free Pol II indicates that this is not always the case.

While the Rudder confirmation is relatively constant in available Pol II structures, Fork Loop 1 varies between two main conformations, differing mainly in their distance from the rudder (Figure 3.6). There appears to be no relationship between the

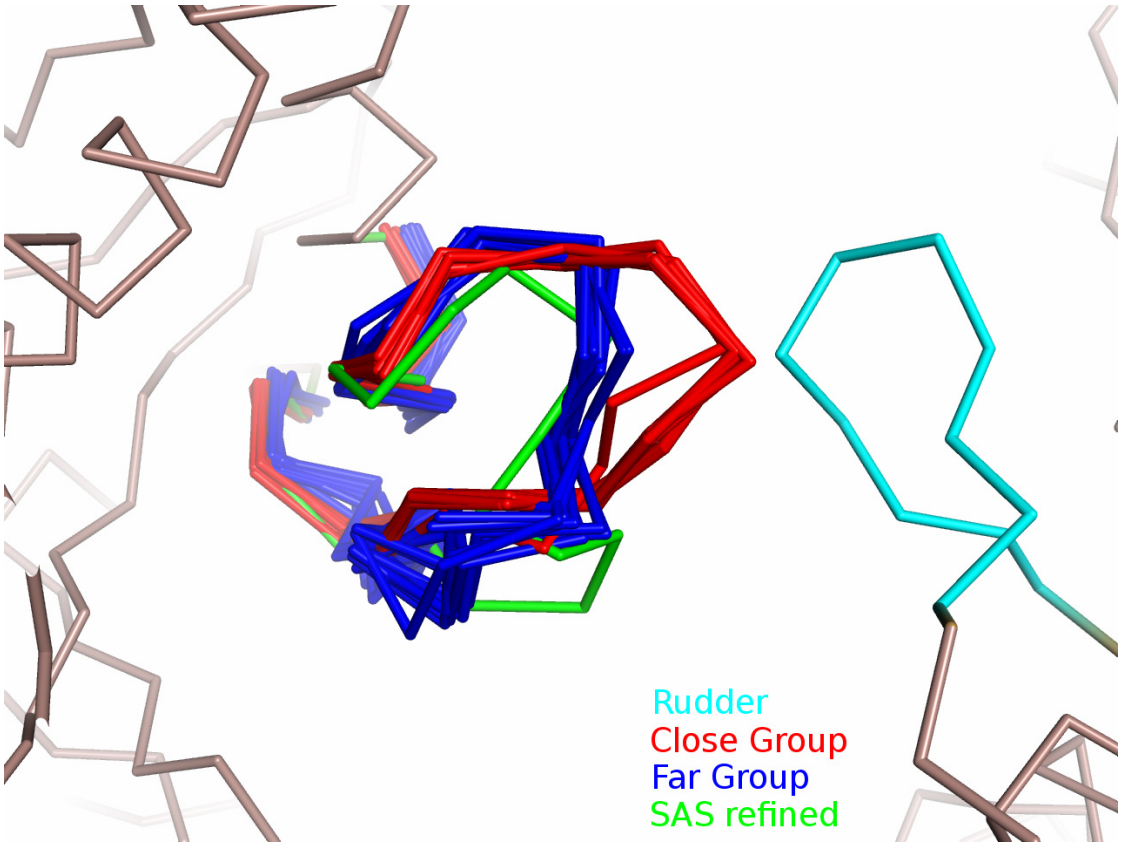


Figure 3.7: Grouping of Fork Loop 1 Conformations

PDB ID and authors for structures with Fork Loop 1 localized (Rudder shown as *cyan ribbon*)
Close Conformations (*red*):
2R92, E.LEHMANN,F.BRUECKNER,P.CRAMER
2R93, E.LEHMANN,F.BRUECKNER,P.CRAMER
1Y77, H.KETTENBERGER,K.-J.ARMACHE,P.CRAMER
2VUM, F.BRUECKNER,P.CRAMER
2JA6, F.BRUECKNER,U.HENNECKE,T.CARELL,P.CRAMER
1Y1W, P.CRAMER,H.KETTENBERGER,K.-J.ARMACHE
2R7Z, G.E.DAMSMA,A.ALT,F.BRUECKNER,T.CARELL,P.CRAMER
2JA7, F.BRUECKNER,U.HENNECKE,T.CARELL,P.CRAMER
2JA5, F.BRUECKNER,U.HENNECKE,T.CARELL,P.CRAMER
2JA8, F.BRUECKNER,U.HENNECKE,T.CARELL,P.CRAMER
Far Conformations (*blue*):
1R9T, K.D.WESTOVER,D.A.BUSHNELL,R.D.KORNBERG
1SFO, K.D.WESTOVER,D.A.BUSHNELL,R.D.KORNBERG
2E2H, D.WANG,D.A.BUSHNELL,K.D.WESTOVER,C.D.KAPLAN,R.D.KORNBERG
2E2I, D.WANG,D.A.BUSHNELL,K.D.WESTOVER,C.D.KAPLAN,R.D.KORNBERG
2E2J, D.WANG,D.A.BUSHNELL,K.D.WESTOVER,C.D.KAPLAN,R.D.KORNBERG
2NVQ, D.WANG,D.A.BUSHNELL,K.D.WESTOVER,C.D.KAPLAN,R.D.KORNBERG
2NVT, D.WANG,D.A.BUSHNELL,K.D.WESTOVER,C.D.KAPLAN,R.D.KORNBERG
2NVX, D.WANG,D.A.BUSHNELL,K.D.WESTOVER,C.D.KAPLAN,R.D.KORNBERG
2NVZ, D.WANG,D.A.BUSHNELL,K.D.WESTOVER,C.D.KAPLAN,R.D.KORNBERG
2YU9, D.WANG,D.A.BUSHNELL,K.D.WESTOVER,C.D.KAPLAN,R.D.KORNBERG

SAS Refined Conformation (*green*)

Fork Loop 1 conformation and nucleotides bound, pH, or crystallization conditions. However, there is a clear relationship between which research group deposited the structure, and which conformation Fork Loop 1 was present in (Figure 3.7). Fork Loop 1 observed in free Pol II is further from the Rudder, and slightly closer to the bottom of the cleft, than either of conformations observed in engaged Pol II. This suggests that a conformational change may be required for hybrid engagement. Given its location and contacts with the Rudder (Figure 3.6), Fork Loop 1 must undergo conformational rearrangement in order to open a path for single-stranded template DNA to reach the active site during open complex formation, and to allow release of the template DNA during termination. The conformational variability of Fork Loop 1 and its absence due to delocalization in the majority of free Pol II structures reported to date, indicate that this region has some degree of structural flexibility.

#### 3.3.2.2 Fork Loop 2

Fork Loop 2 consists of residues 502-510 in Rpb2, and is located in front of and below Fork Loop 1, and forms part of the Cleft floor underneath the incoming DNA duplex. This region has been observed previously in some but not all Pol II structures with DNA and RNA present; and not in structures of free Pol II. Fork Loop 2 is located at the point where the double stranded DNA unwinds into single-stranded template and non-template; because of this, it is thought to be involved in the formation and maintenance of the transcription bubble (Gnatt et al., 2001). Two conformations of Fork Loop 2 have been previously observed (Kettenberger et al., 2004; Wang et al., 2006), differing mainly in the direction in which the loop bends. The conformation observed in free Pol II refined with the SAS data appears to be an intermediate between them (Figure 3.8), supporting the idea that this region is highly mobile.

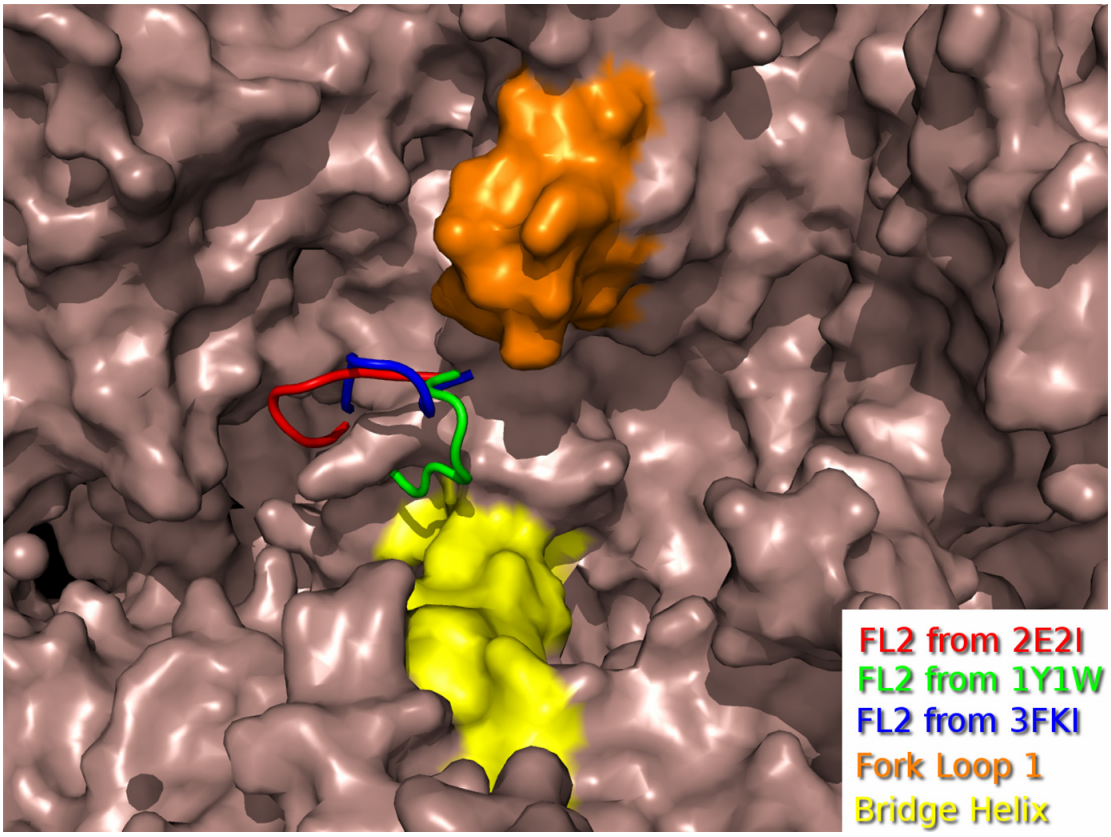


Figure 3.8: Representative Fork Loop 2 Conformations

Three Fork Loop 2 conformations: *red* and *green* ribbons show two engaged conformations; *Blue* ribbon shows the nucleic acid-free conformation from the SAS-refined model. Fork Loop 1 (*orange surface*) and Bridge Helix (*yellow surface*) are shown for orientation and context.

### 3.3.2.3 Protrusion

The Protrusion domain is a region of Rpb2 forming one of four stalk-like projections emanating from the top of Pol II (Figure 1.4). Several stretches of residues within this region that are absent in existing Pol II models. Additional density was also observed in this region in the experimental multi-crystal map, as well as multi-crystal maps of other Pol II complexes (Ceg1/Cet1-Pol II and Spt5-Pol II, unpublished results). An initial poly-alanine model was built for some of the missing residues, as described in Chapter 2. The model for this region was completed by the addition of side-chains, and subsequent refinement. Additional residues were later added on the basis of omit maps as refinement progressed (Figure 3.4 E and F; Figure 3.9 B). However, some segments of this region remain difficult to model, due to unclear electron density (not shown).

Residues in this region have recently been implicated in interacting with TFIIF (Chen et al., 2007) based on cross-linking data. The newly built region has two cross-linking sites nearby, 18-30 Å away (Figure 3.9 A). In addition, cryo-EM reconstructions of the Pol II-TFIIF complex also exhibited density nearby this region which was attributed to TFIIF (Chung et al., 2003).

The Protrusion loops, including the newly modeled region, appear to exhibit a range of conformational variability in existing structures, as shown by its absence from previous structures. For 12-subunit crystals, a portion of this variation may be due the lack of neighboring molecules in the crystal lattice to stabilize this region of the protein (Armache et al., 2003; Bushnell and Kornberg, 2003; Meyer et al., 2006). The distance between symmetry related molecules in both crystal forms of 10-subunit Pol

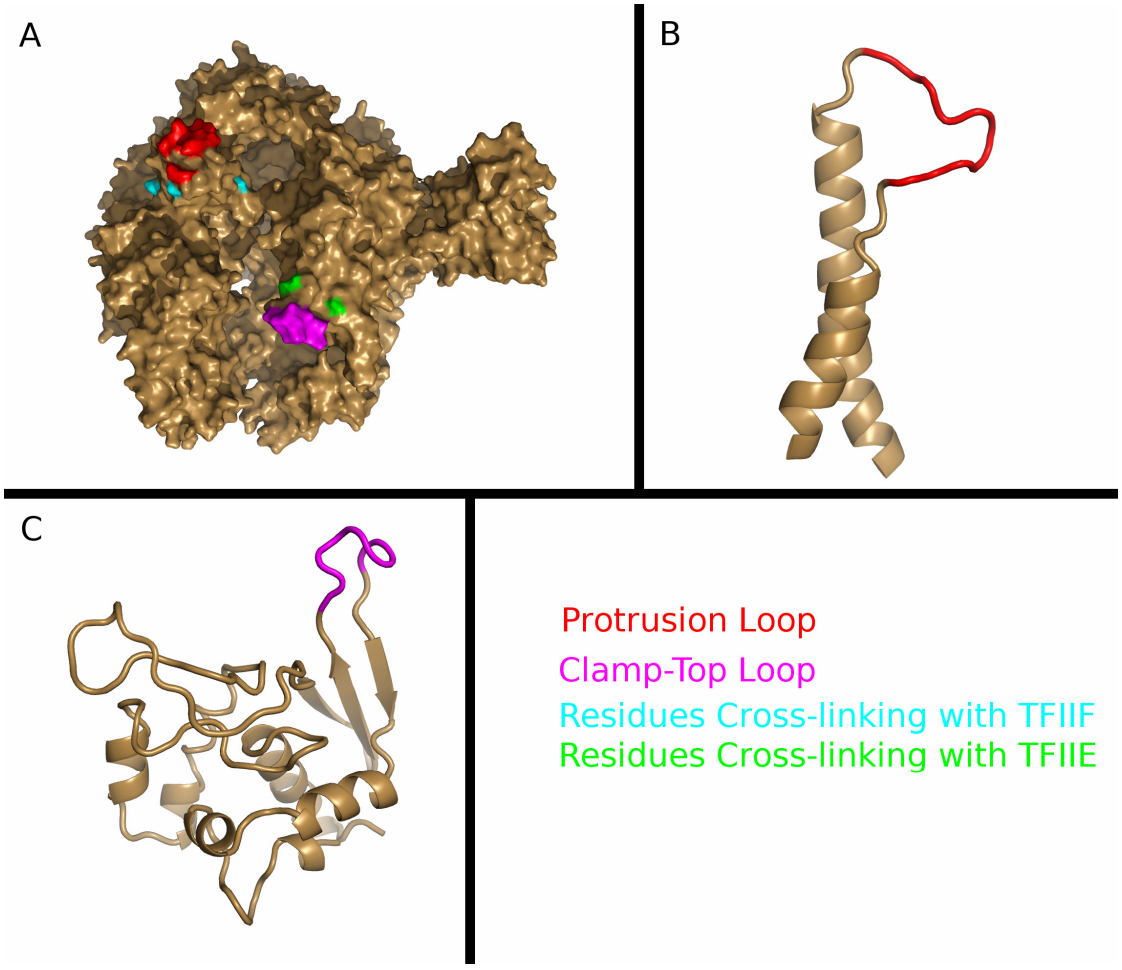


Figure 3.9: Conformations of Protrusion and Clamp-Top Loops

Panel A shows the positions of residues implicated in interactions with TFIIF (*cyan*) and TFIIE (*green*) relative to the loops that were revealed after SAS refinement. The conformations of the newly modeled loops are shown in more detail: *Red* region in Panel B for the Protrusion Loop; *Magenta* region in Panel C for the Clamp-Top Loop.

II is substantially smaller. Despite this, the Protrusion loops were also disordered in the higher-resolution 10-subunit Pol II structures (Cramer et al., 2000; Cramer et al., 2001). The conformation of the newly modeled Protrusion Loop, as well as that of the additional loop for which the density did not permit modeling, is likely to change upon interaction with TFIIF.

#### 3.3.2.4 Clamp-Top

The Clamp domain forms the other side of the cleft opposite to the Protrusion (Figures 1.1 and 1.4). The Clamp contacts the DNA and RNA extensively (Gnatt et al., 2001; Kettenberger et al., 2004). This region has been known to be mobile since the earliest structural results on Pol II. Additional density for the loop element at the top of the Clamp, Clamp-Top Loop which has been missing in previous structures, was modeled during this refinement (Figure 3.4 A and B; Figure 3.9 C). The position of this loop makes it unlikely that it is directly involved in interactions with nucleic acids. However, residues surrounding this loop have been implicated in interaction with TFIIE (Chen et al., 2007). As with the Protrusion loops, the absence of the Clamp-Top Loop in previous structures suggests that this region has a degree of conformational flexibility greater than that of the Clamp domain as a whole.

## 3.4 Materials and Methods

The same diffraction datasets used in the phasing experiment described in Chapter 2 were used for refinement. Anomalous and normal amplitudes from the best diffracting crystal (A3X7 in Table 2.1) were used for refinement of the model after reprocessing in HKL2000 (Otwinowski and Minor, 1997) which allowed recovery of data to 3.8 Å.

In preparation for refinement with SAS data from the bound Zn ions, the anomalous site parameters (positional and occupancy) were initially refined in BP3 (Pannu and Read, 2004) against the single crystal SAS amplitudes. The unrefined 12-subunit Pol II model (Meyer et al., 2006) was subjected to geometry regularization, removal of incomplete amino acids, and replacement of poly-alanine stretches with the correct sequence. Temperature factors (atomic B-factors) of the model were all reset to a uniform value of 77, following the Wilson B-factor of the normal amplitudes.

Refinement proceeded according to the standard practice of cycles of reciprocal space refinement alternating with manual model adjustment and real-space refinement. The positional parameters for the Zn sites were refined in concert with the polymerase model using REFMAC5D (Skubak et al., 2004). The constant for relative weighting of X-ray and geometry terms as defined in REFMAC (Murshudov et al., 1997) was determined manually, as the auto-weighting scheme produced severe geometry distortions (comparable to those shown in Figure 3.1), although with improved R-factors. Weighting terms for real-space refinement in COOT (Emsley and Cowtan, 2004) were determined similarly. In the final stages of refinement, problematic regions were identified using ADIT (Berman et al., 2000), MOLPROBITY (Lovell et al., 2003), and COOT.

In the initial stage, only positional parameters and an overall temperature factor were refined. Once this had converged, positional parameters were further refined with TLS groups. To select TLS groups, two different approaches were evaluated. The first approach was to use the known rigid-body domains of Pol II (Table 2.3 and Figure 2.7) as TLS groups. The second approach was to conduct several cycles of refinement of atomic temperature factors only, without positional refinement. The

resulting model was then used as input to the TLS Motion Determination server (Painter and Merritt, 2006), which suggested TLS groups for each subunit of Pol II. Possibly due to the restriction of TLS groups to individual peptides, this approach did not perform any better than the rigid-body groups. As such, the first approach was chosen for defining the TLS groups used in later refinement. These TLS groups were expanded as needed to account for the regions that had increased flexibility, which was manifested as negative thermal atomic displacements and regions of negative $F_O$-$F_C$ density.

# CHAPTER 4: SUMMARY AND FUTURE DIRECTIONS

## 4.1 Summary

The progress made through this work resolves two crystallographic problems hampering the determination of Pol II complex structures. In the first, multi-crystal phasing using intrinsic Zn present in Pol II allows relatively straightforward determination of high quality experimental phases; circumventing the issue of model bias in model-derived phases at low resolutions. In the second, enhancement of the observation-to-parameter ratio in refinement, was made possible by incorporating anomalous amplitudes directly in refinement and using of TLS groups. At low resolutions, this approach allows for the determination of a higher quality model than could be achieved using standard refinement methodologies.

Although the problems addressed here pertain to Pol II crystals, and especially crystals of Pol II-containing complexes, they are not necessarily limited the Pol II system. Large macromolecular complexes play essential roles in many biological processes. Crystallographic structural models of these complexes will be essential for illuminating their mechanisms. However, crystals of large complexes often diffract poorly, and producing high-resolution crystals of such complexes can be difficult. The inherent internal dynamics of many macromolecular complexes, in conjunction with our current understanding of bio-molecular crystallization, suggest that the process of producing high-resolution crystals of large macromolecules will not become substantially easier in the immediate future. The strategies adopted in this work suggest that meaningful biological results can be obtained even with relatively weakly-diffracting crystals, which may be more easily produced.

The use of anomalous data arising from the Zn ions intrinsic to the Pol II system was essential for the main results discussed in this work. However, the essential component is the presence of an anomalous signal arising from a large macromolecule, not that the anomalous signal is due to Zn as opposed to that from iron, selenium, bromine, iodine or sulfur. Therefore these approaches can be applied to any macromolecular crystal with a weak anomalous signal, whether due to intrinsic metal ions, derivatization with heavy atom compounds, or selenomethionine substitution.

## 4.2 Future Directions

These results provide a starting point for three lines of research. This work was conducted with the primary goal of resolving crystallographic problems presented by low-resolution diffraction data from Pol II complexes. Therefore, the most straightforward avenue of research opened by this work is to apply these methods to novel complexes of Pol II, such as complexes of Pol II with elongation factor Spt5 and RNA capping enzyme.

The remaining potential directions concern the methods themselves. Both of the technical approaches discussed here, multi-crystal phasing and optimizing observation-to-parameter ratio in refinement, proved effective in dealing with low-resolution diffraction data from Pol II crystals, distinguishing them from the substantially longer list of ineffective approaches. The overall philosophy guiding both approaches described in this work was to extract the maximum amount of information available from the available data. Still, further improvements are possible in both phasing and refinement techniques.

For multi-crystal phasing, the main remaining limitations are in the finding of anomalous scattering sites and identification of an optimal combination of phasing datasets. The location of scattering sites for crystals of Pol II complexes is likely to remain straightforward in the near future, as the current approach of using model-phased anomalous and dispersive difference maps should prove effective for much larger complexes than are currently contemplated as targets. The identification of an optimal, or at least sufficient, combination of phasing datasets remains as one of the more time-consuming steps in the process. An extension of the probability plotting method discussed in Chapter 2 could potentially reduce the time required. As incompatible phasing datasets are currently attributed to crystal non-isomorphism, alternative methods to detect non-isomorphism at low resolutions may facilitate this process. The elimination of non-isomorphous datasets from the multi-crystal phase set is preferable to allowing them to degrade the final result, but means that some experimental information is being discarded. It may be possible to compensate for non-isomorphism in multi-crystal phasing by incorporation of a cross-crystal transformation matrix, similar to that used in multi-crystal averaging, into the phase calculations. If successful, such a procedure could potentially expedite multi-crystal phasing by both reducing the number of datasets that need to be collected, and reducing the time spend processing that data.

Several questions remain regarding the refinement of models at low resolutions, and additional improvements may be possible. The incorporation of anomalous amplitudes ($F^{(+)}$ and $F^{(-)}$) into refinement proved crucial in the refinement of Pol II at 3.8 Å. In contrast, the incorporation of SAS or MAD phases in refinement as a probability distribution (MLHL (Pannu et al., 1998)) did not allow for successful refinement of the model, despite a roughly equivalent increase in the number of

observations. Although this difference in performance could be due to either allowing the refinement of the zinc sites during model refinement or the use of un-approximated phase information in the refinement target, it remains an unsettled question. Additionally, it may be possible to further improve the observation-to-parameter ratio used in refinement. It should be possible to incorporate information from additional sets of amplitudes into refinement. For example, incorporating datasets collected at inflection and remote, in addition to peak wavelength in a MAD experiment would provide an additional increase in the number of independent observations usable for model refinement.

The third potential direction for this research was inspired by an observation from the simulation carried out to investigate phasing efficiency, as described in Chapter 2: in the absence of anomalous scattering from the light atoms in the protein, a single zinc atom was sufficient to allow phasing. Although the anomalous scattering due to a single light-atom is very small at the Zn absorption edge, the combined effect from tens of thousands of these individual atoms can be significant. This combined effect can be sufficient to interfere with the phasing process. In effect, the errors present in current experimental phases are a combination of random experimental errors and calculation errors due to the non-random anomalous scattering from the "non-anomalous" atoms in the protein which are currently unaccounted for. Therefore, compensating for this residual anomalous scattering, even partially, could minimize the errors in experimental phase values due to the neglect of anomalous scattering from the light atoms. One potential approach to this problem is to incorporate an inverse transformed likelihood residual map (Fortelle and Bricogne, 1997) into the heavy-atom (or anomalous-scatterer) structure factor used during phase calculation.

# REFERENCES

Abrahams, JP. (1997). Bias reduction in phase refinement by modified interference functions: introducing the gamma correction. Acta Crystallogr. D Biol. Crystallogr. *53*, 371-376.

Abrahams, JP. and Leslie, AG. (1996). Methods used in the structure determination of bovine mitochondrial F1 ATPase. Acta Crystallogr. D Biol. Crystallogr. *52*, 30-42.

Allison, LA., Wong, JK., Fitzpatrick, VD. et al. (1988). The C-terminal domain of the largest subunit of RNA polymerase II of Saccharomyces cerevisiae, Drosophila melanogaster, and mammals: a conserved structure with an essential function. Mol. Cell. Biol. *8*, 321-329.

Armache, K., Kettenberger, H. and Cramer, P. (2003). Architecture of initiation-competent 12-subunit RNA polymerase II. Proc. Natl. Acad. Sci. U.S.A. *100*, 6964-6968.

Armache, K., Mitterweger, S., Meinhart, A. et al. (2005). Structures of complete RNA polymerase II and its subcomplex, Rpb4/7. J. Biol. Chem. *280*, 7131-7134.

Bar-Nahum, G., Epshtein, V., Ruckenstein, AE. et al. (2005). A ratchet mechanism of transcription elongation and its control. Cell *120*, 183-193.

Berman, HM., Westbrook, J., Feng, Z. et al. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235-242.

Bhat, TN. (1988). Calculation of an OMIT map. Journal of Applied Crystallography *21*, 279-281.

Box, G. and Muller, M. (1958). A Note on the Generation of Random Normal Deviates. The Annals of Mathematical Statistics *29*, 610-611.

Brueckner, F. and Cramer, P. (2008). Structural basis of transcription inhibition by alpha-amanitin and implications for RNA polymerase II translocation. Nat. Struct. Mol. Biol. *15*, 811-818.

Brünger, AT., Adams, PD., Clore, GM. et al. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905-921.

Bushnell, DA. and Kornberg, RD. (2003). Complete, 12-subunit RNA polymerase II at 4.1-A resolution: implications for the initiation of transcription. Proc. Natl. Acad. Sci. U.S.A. *100*, 6969-6973.

Bushnell, DA., Cramer, P. and Kornberg, RD. (2002). Structural basis of transcription: alpha-amanitin-RNA polymerase II cocrystal at 2.8 A resolution. Proc. Natl. Acad. Sci. U.S.A. *99*, 1218-1222.

Bushnell, DA., Westover, KD., Davis, RE. et al. (2004). Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms. Science *303*, 983-988.

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

Cho, EJ., Takagi, T., Moore, CR. et al. (1997). mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. Genes Dev. *11*, 3319-3326.

Chung, W., Craighead, JL., Chang, W. et al. (2003). RNA polymerase II/TFIIF structure and conserved organization of the initiation complex. Mol. Cell *12*, 1003-1013.

Collaborative Computational Project, N4. (1994). The CCP4 suite: programs for protein crystallography. Acta Crystallographica Section D *50*, 760-763.

Corden, JL., Cadena, DL., Ahearn, JMJ. et al. (1985). A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. Proc. Natl. Acad. Sci. U.S.A. *82*, 7934-7938.

Cowtan, K. (1994). 'dm': An automated procedure for phase improvement by density modification. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography *31*, 34-38.

Craighead, JL., Chang, W. and Asturias, FJ. (2002). Structure of yeast RNA polymerase II in solution: implications for enzyme regulation and interaction with promoter DNA. Structure *10*, 1117-1125.

Cramer, P., Armache, K., Baumli, S. et al. (2008). Structure of eukaryotic RNA polymerases. Annu Rev Biophys *37*, 337-352.

Cramer, P., Bushnell, DA. and Kornberg, RD. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science *292*, 1863-1876.

Cramer, P., Bushnell, DA., Fu, J. et al. (2000). Architecture of RNA polymerase II and implications for the transcription mechanism. Science *288*, 640-649.

Cromer, DT. (1983). Calculation of anomalous scattering factors at arbitrary wavelengths. Journal of Applied Crystallography *16*, 437.

Darst, SA., Edwards, AM., Kubalek, EW. et al. (1991). Three-dimensional structure of yeast RNA polymerase II at 16 A resolution. Cell *66*, 121-128.

DeLaBarre, B. and Brunger, AT. (2006). Considerations for the refinement of low-resolution crystal structures. Acta Crystallographica Section D *62*, 923-932.

DeLano, W. (). The PyMOL Molecular Graphics System. , .

[Drenth1999] Drenth, J. Principles of Protein X-Ray Crystallography, Second Edition. . Springer, 1999.

Ebright, RH. (2000). RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. J. Mol. Biol. *304*, 687-698.

Edwards, AM., Kane, CM., Young, RA. et al. (1991). Two dissociable subunits of yeast RNA polymerase II stimulate the initiation of transcription at a promoter in vitro. J. Biol. Chem. *266*, 71-75.

Egloff, S., O'Reilly, D., Chapman, RD. et al. (2007). Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. Science *318*, 1777-1779.

Emsley, P. and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126-2132.

Evans, P. (2006). Scaling and assessment of data quality. Acta Crystallogr. D Biol. Crystallogr. *62*, 72-82.

Fabrega, C., Shen, V., Shuman, S. et al. (2003). Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II. Mol. Cell *11*, 1549-1561.

Fortelle, EDL. and Bricogne, G. (1997). [27] Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *276*, 472-494.

Fu, J., Gnatt, AL., Bushnell, DA. et al. (1999). Yeast RNA polymerase II at 5 A resolution. Cell *98*, 799-810.

Furey, W. and Swaminathan, S. (1997). [31] PHASES-95: A program package for processing and analyzing diffraction data from macromolecules. *277*, 590-608.

Gnatt, A., Fu, J. and Kornberg, RD. (1997). Formation and crystallization of yeast RNA polymerase II elongation complexes. J. Biol. Chem. *272*, 30799-30805.

Gnatt, AL., Cramer, P., Fu, J. et al. (2001). Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. Science *292*, 1876-1882.

Gu, W., Wind, M. and Reines, D. (1996). Increased accommodation of nascent RNA in a product site on RNA polymerase II during arrest. Proc. Natl. Acad. Sci. U.S.A. *93*, 6935-6940.

Gudipati, RK., Villa, T., Boulay, J. et al. (2008). Phosphorylation of the RNA polymerase II C-terminal domain dictates transcription termination choice. Nat. Struct. Mol. Biol. *15*, 786-794.

Hao, Q., Gu, YX., Yao, JX. et al. (2003). SAPI: a direct-methods program for finding heavy-atom sites with SAD or SIR data. Journal of Applied Crystallography *36*, 1274-1276.

Hendrickson, WA. and Lattman, EE. (1970). Representation of phase probability distributions for simplified combination of independent phase information. Acta Crystallographica Section B *26*, 136-143.

Heyer, LJ., Kruglyak, S. and Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. Genome Res. *9*, 1106-1115.

Hirose, Y. and Ohkuma, Y. (2007). Phosphorylation of the C-terminal domain of RNA polymerase II plays central roles in the integrated events of eucaryotic gene expression. J Biochem *141*, 601-608.

Ho, CK., Schwer, B. and Shuman, S. (1998). Genetic, physical, and functional interactions between the triphosphatase and guanylyltransferase components of the yeast mRNA capping apparatus. Mol. Cell. Biol. *18*, 5189-5198.

Ho, CK., Sriskanda, V., McCracken, S. et al. (1998). The guanylyltransferase domain of mammalian mRNA capping enzyme binds to the phosphorylated carboxyl-terminal domain of RNA polymerase II. J. Biol. Chem. *273*, 9577-9585.

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

Howlin, B., Moss, DS. and Harris, GW. (1989). Segmented anisotropic refinement of bovine ribonuclease A by the application of the rigid-body TLS model. Acta Crystallogr., A, Found. Crystallogr. *45 ( Pt 12)*, 851-861.

Hull, MW., McKune, K. and Woychik, NA. (1995). RNA polymerase II subunit RPB9 is required for accurate start site selection. Genes Dev. *9*, 481-490.

Jeronimo, C., Langelier, MF., Zeghouf, M. et al. (2004). RPAP1, a novel human RNA polymerase II-associated protein affinity purified with recombinant wild-type and mutated polymerase subunits. Mol. Cell. Biol. *24*, 7043-7058.

Jones, TA., Zou, JY., Cowan, SW. et al. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr., A, Found. Crystallogr. *47 ( Pt 2)*, 110-119.

Kaplan, C. and Kornberg, R. (2008). A bridge to transcription by RNA polymerase. J. Biol. *7*, 39.

Kaplan, CD., Larsson, K. and Kornberg, RD. (2008). The RNA polymerase II trigger loop functions in substrate selection and is directly targeted by alpha-amanitin. Mol. Cell *30*, 547-556.

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

Kettenberger, H., Armache, K. and Cramer, P. (2004). Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS. Mol. Cell *16*, 955-965.

Kleywegt, GJ. and Jones, TA. (1995). Where freedom is given, liberties are taken. Structure *3*, 535-540.

Kleywegt, GJ., Harris, MR., Zou, JY. et al. (2004). The Uppsala Electron-Density Server. Acta Crystallogr. D Biol. Crystallogr. *60*, 2240-2249.

Leslie, AG. (1992). Recent changes to the MOSFLM package for processing file and image plate data. Joint CCP4 + ESF-EAMCB Newsletter on Protein Crystallography *26*, .

Li, Y. and Kornberg, RD. (1994). Interplay of positive and negative effectors in function of the C-terminal repeat domain of RNA polymerase II. Proc. Natl. Acad. Sci. U.S.A. *91*, 2362-2366.

Li, Y., Flanagan, PM., Tschochner, H. et al. (1994). RNA polymerase II initiation factor interactions and transcription start site selection. Science *263*, 805-807.

Lovell, SC., Davis, IW., Arendall, W. B., 3rd et al. (2003). Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins *50*, 437-450.

Lunin, VY. and Woolfson, MM. (1993). Mean phase error and the map-correlation coefficient. Acta Crystallogr. D Biol. Crystallogr. *49*, 530-533.

McCracken, S., Fong, N., Rosonina, E. et al. (1997). 5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. Genes Dev. *11*, 3306-3318.

Meinhart, A. and Cramer, P. (2004). Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. Nature *430*, 223-226.

Meyer, PA., Ye, P., Suh, M. et al. (2009). Structure of the 12-subunit RNA polymerase II refined with the aid of anomalous diffraction data. J. Biol. Chem. *284*, 12933-12939.

Meyer, PA., Ye, P., Zhang, M. et al. (2006). Phasing RNA polymerase II using intrinsically bound Zn atoms: an updated structural model. Structure *14*, 973-982.

Minakhin, L., Bhagat, S., Brunning, A. et al. (2001). Bacterial RNA polymerase subunit omega and eukaryotic RNA polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly. Proc. Natl. Acad. Sci. U.S.A. *98*, 892-897.

Murshudov, GN., Vagin, AA. and Dodson, EJ. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr. D Biol. Crystallogr. *53*, 240-255.

Naji, S., Bertero, MG., Spitalny, P. et al. (2008). Structure-function analysis of the RNA polymerase cleft loops elucidates initial transcription, DNA unwinding and RNA displacement. Nucleic Acids Res. *36*, 676-687.

Navaza, J. (1994). AMoRe: an automated package for molecular replacement. Acta Crystallographica Section A *50*, 157-163.

Nonet, M., Sweetser, D. and Young, RA. (1987). Functional redundancy and structural polymorphism in the large subunit of RNA polymerase II. Cell *50*, 909-915.

Orphanides, G., Lagrange, T. and Reinberg, D. (1996). The general transcription factors of RNA polymerase II. Genes Dev. *10*, 2657-2683.

Otwinowski, Z. and Minor, W. (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode. Methods in Enzymology *276*, 307-326.

Painter, J. and Merritt, EA. (2006). TLSMD web server for the generation of multi-group TLS models. Journal of Applied Crystallography *39*, 109-111.

Pal, M. and Luse, DS. (2003). The initiation-elongation transition: lateral mobility of RNA in RNA polymerase II complexes is greatly reduced at +8/+9 and absent by +23. Proc. Natl. Acad. Sci. U.S.A. *100*, 5700-5705.

Pannu, NS. and Read, RJ. (2004). The application of multivariate statistical techniques improves single-wavelength anomalous diffraction phasing. Acta Crystallogr. D Biol. Crystallogr. *60*, 22-27.

Pannu, NS., Murshudov, GN., Dodson, EJ. et al. (1998). Incorporation of prior phase information strengthens maximum-likelihood structure refinement. Acta Crystallogr. D Biol. Crystallogr. *54*, 1285-1294.

Phatnani, HP. and Greenleaf, AL. (2006). Phosphorylation and functions of the RNA polymerase II CTD. Genes Dev. *20*, 2922-2936.

Proudfoot, N. (2000). Connecting transcription to messenger RNA processing. Trends Biochem. Sci. *25*, 290-293.

Proudfoot, NJ., Furger, A. and Dye, MJ. (2002). Integrating mRNA processing with transcription. Cell *108*, 501-512.

Ramakrishnan, V. and Biou, V. (1997). [31] Treatment of multiwavelength anomalous diffraction data as a special case of multiple isomorphous replacement. *276*, 538-557.

Rasmussen, EB. and Lis, JT. (1993). In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. Proc. Natl. Acad. Sci. U.S.A. *90*, 7923-7927.

Read, RJ. (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Crystallographica Section A *42*, 140-149.

Read, RJ. (1997). [7] Model phases: Probabilities and bias. *277*, 110-128.

Richard, P. and Manley, JL. (2009). Transcription termination by nuclear RNA polymerases. Genes Dev. *23*, 1247-1269.

Saunders, A., Core, LJ. and Lis, JT. (2006). Breaking barriers to transcription elongation. Nat. Rev. Mol. Cell Biol. *7*, 557-567.

Schomaker, V. and Trueblood, KN. (1968). On the rigid-body motion of molecules in crystals. Acta Crystallographica Section B *24*, 63-76.

Schroeder, SC., Zorio, DAR., Schwer, B. et al. (2004). A function of yeast mRNA cap methyltransferase, Abd1, in transcription by RNA polymerase II. Mol. Cell *13*, 377-387.

Shaw, PE. (2007). Peptidyl-prolyl cis/trans isomerases and transcription: is there a twist in the tail?. EMBO Rep. *8*, 40-45.

Sims, RJ3., Mandal, SS. and Reinberg, D. (2004). Recent highlights of RNA-polymerase-II-mediated transcription. Curr. Opin. Cell Biol. *16*, 263-271.

Skubak, P., Murshudov, GN. and Pannu, NS. (2004). Direct incorporation of experimental phase information in model refinement. Acta Crystallogr. D Biol. Crystallogr. *60*, 2196-2201.

Skubák, P., Murshudov, GN. and Pannu, NS. (2004). Direct incorporation of experimental phase information in model refinement. Acta Crystallogr. D Biol. Crystallogr. *60*, 2196-2201.

Skubák, P., Ness, S. and Pannu, NS. (2005). Extending the resolution and phase-quality limits in automated model building with iterative refinement. Acta Crystallogr. D Biol. Crystallogr. *61*, 1626-1635.

Srinivasan, R. and Ramachandran, G. (1965). Probability distribution connected with structure amplitudes of two related crystals. V. The effect of errors in the atomic coordinates on the distribution of observed and calculated structure factors. Acta Crystallographica *19*, 1008-1014.

Sun, ZW., Tessmer, A. and Hampsey, M. (1996). Functional interaction between TFIIB and the Rpb9 (Ssu73) subunit of RNA polymerase II in Saccharomyces cerevisiae. Nucleic Acids Res. *24*, 2560-2566.

Sweetser, D., Nonet, M. and Young, RA. (1987). Prokaryotic and eukaryotic RNA polymerases have homologous core subunits. Proc. Natl. Acad. Sci. U.S.A. *84*, 1192-1196.

Temiakov, D., Zenkin, N., Vassylyeva, MN. et al. (2005). Structural basis of transcription inhibition by antibiotic streptolydigin. Mol. Cell *19*, 655-666.

Thomas, MC. and Chiang, C. (2006). The general transcription machinery and general cofactors. Crit. Rev. Biochem. Mol. Biol. *41*, 105-178.

Timmers, HT. (1994). Transcription initiation by RNA polymerase II does not require hydrolysis of the beta-gamma phosphoanhydride bond of ATP. EMBO J. *13*, 391-399.

Vassylyev, DG., Vassylyeva, MN., Zhang, J. et al. (2007). Structural basis for substrate loading in bacterial RNA polymerase. Nature *448*, 163-168.

Verdecia, MA., Bowman, ME., Lu, KP. et al. (2000). Structural basis for phosphoserine-proline recognition by group IV WW domains. Nat. Struct. Biol. *7*, 639-643.

Wang, BC. (1985). Resolution of phase ambiguity in macromolecular crystallography. Meth. Enzymol. *115*, 90-112.

Wang, D., Bushnell, DA., Huang, X. et al. (2009). Structural basis of transcription: backtracked RNA polymerase II at 3.4 angstrom resolution. Science *324*, 1203-1206.

Wang, D., Bushnell, DA., Westover, KD. et al. (2006). Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. Cell *127*, 941-954.

West, ML. and Corden, JL. (1995). Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. Genetics *140*, 1223-1233.

Westover, KD., Bushnell, DA. and Kornberg, RD. (2004). Structural basis of transcription: nucleotide selection by rotation in the RNA polymerase II active center. Cell *119*, 481-489.

Wong, Y., Lee, T., Liang, H. et al. (2007). KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. Nucleic Acids Res. *35*, W588-94.

Woychik, NA. and Young, RA. (1990). RNA polymerase II: subunit structure and function. Trends Biochem. Sci. *15*, 347-351.

Woychik, NA., Liao, SM., Kolodziej, PA. et al. (1990). Subunits shared by eukaryotic nuclear RNA polymerases. Genes Dev. *4*, 313-323.

Yue, Z., Maldonado, E., Pillutla, R. et al. (1997). Mammalian capping enzyme complements mutant Saccharomyces cerevisiae lacking mRNA guanylyltransferase and selectively binds the elongating form of RNA polymerase II. Proc. Natl. Acad. Sci. U.S.A. *94*, 12898-12903.

Zhang, G., Campbell, EA., Minakhin, L. et al. (1999). Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 A resolution. Cell *98*, 811-824.

Zhang, J. and Corden, JL. (1991). Identification of phosphorylation sites in the repetitive carboxyl-terminal domain of the mouse RNA polymerase II largest subunit. J. Biol. Chem. *266*, 2290-2296.

Zhang, Y., Kim, Y., Genoud, N. et al. (2006). Determinants for dephosphorylation of the RNA polymerase II C-terminal domain by Scp1. Mol. Cell *24*, 759-770.

Zorio, DAR. and Bentley, DL. (2004). The link between mRNA processing and transcription: communication works both ways. Exp. Cell Res. *296*, 91-97.