EXAMINING FLEXIBLE BIOLOGICAL STRUCTURES

By

SETH ANDREW VILLARREAL

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Pharmacology

CASE WESTERN RESERVE UNIVERSITY

January, 2015
CASE WESTERN RESERVE UNIVERSITY
SCHOOL OF GRADUATE STUDIES

We hereby approve the thesis/dissertation of Seth Andrew Villarreal

Candidate for the degree of Doctor of Philosophy in Pharmacology*.

Committee Chair
Derek Taylor

Committee Member
Phoebe Stewart

Committee Member
Krzysztof Palczewski

Committee Member
Focco van den Akker

Committee Member
Jason Mears

Date of Defense
8th of October, 2014

*We also certify that written approval has been obtained for any proprietary material contained therein.
# Table of Contents

List of Tables ........................................................................................................ vi

Acknowledgements .............................................................................................. ix

Abstract ................................................................................................................ x

Chapter 1: Introduction .......................................................................................... 1

Examining Structure ......................................................................................... 2

Two EM Structures of Protein Complexes ....................................................... 4

Circadian Clock Systems ................................................................................... 5

Identification of *Synechococcus elongatus* Circadian Cycle ..................... 6

Kai A, B, and C as the Core Oscillator ......................................................... 8

Structural Aspects of the KaiABC System .................................................... 13

DNA Maintenance and Human Health ....................................................... 17

Types of Damage Affecting the Integrity of DNA ..................................... 18

Repairing Double Strand DNA Breaks ...................................................... 22

Components of the NHEJ Pathway .............................................................. 23

Chapter 2: CryoEM and Molecular Dynamics of the Circadian KaiB–KaiC Complex Indicates KaiB Monomers Interact with KaiC and Block ATP Binding

Clefts .................................................................................................................. 29

Abstract .............................................................................................................. 30

Keywords ............................................................................................................. 31

Graphical Abstract ............................................................................................ 31
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset Management Scripts</td>
<td>168</td>
</tr>
<tr>
<td>renamer.perl</td>
<td>168</td>
</tr>
<tr>
<td>dm3-transfer.s</td>
<td>169</td>
</tr>
<tr>
<td>Refinement Scripts</td>
<td>171</td>
</tr>
<tr>
<td>ctffind3-batchboxer-append-155kx-p1.s</td>
<td>171</td>
</tr>
<tr>
<td>master_dd_rnd1</td>
<td>174</td>
</tr>
<tr>
<td>f_rnd9_p1</td>
<td>187</td>
</tr>
<tr>
<td>f_rnd9_pr63</td>
<td>188</td>
</tr>
<tr>
<td>rescue-creator-20120509.s</td>
<td>188</td>
</tr>
<tr>
<td>VMD Scripts</td>
<td>198</td>
</tr>
<tr>
<td>vmd_setup_mngr.s</td>
<td>199</td>
</tr>
<tr>
<td>vmd_setup_before_autopsf.tcl</td>
<td>200</td>
</tr>
<tr>
<td>vmd_setup_after_autopsf.tcl</td>
<td>200</td>
</tr>
<tr>
<td>vmd_namde_v03.tcl</td>
<td>201</td>
</tr>
<tr>
<td>c1_non_bond_e.tcl</td>
<td>203</td>
</tr>
<tr>
<td>vmd_kabbc_pdb_to_chimera.s</td>
<td>203</td>
</tr>
<tr>
<td>output_ccc.tcl</td>
<td>204</td>
</tr>
<tr>
<td>Bibliography</td>
<td>205</td>
</tr>
</tbody>
</table>
List of Tables

Table 2.1 KaiB crystal structure statistics ........................................................... 40
Table 2.S1 Comparison of 16 KaiBC models ..................................................... 71
Table 2.S2. Robustness of KaiBC interface after molecular dynamics flexible fitting runs with different starting KaiB orientations ............................................. 72
Table 3.S1 Results of protein disorder predictor webservers for the human DNA-PKcs sequence ................................................................................................. 104
List of Figures

Figure 1.1 KaiABC interactions as the basis of the *S. elongatus* circadian system ........................................................................................................................... 12

Figure 1.2 Diagram of various types of DNA damage ........................................ 20

Figure 2.1 Crystal structure of *S. elongatus* KaiB ............................................... 38

Figure 2.2 CryoEM structure of KaiB-KaiC-Δ489.................................................. 43

Figure 2.3 KaiBC atomic model docked within the cryoEM density ................. 48

Figure 2.4 Interaction residues at the KaiBC interface ...................................... 50

Figure 2.5 KaiC mutation near the predicted KaiBC interface affects KaiB binding ........................................................................................................................... 52

Figure 2.S1 Sequence alignment for KaiB proteins from selected cyanobacteria ........................................................................................................................... 65

Figure 2.S2 Conformation of the KaiB tetramer ................................................. 66

Figure 2.S3 Interactions stabilizing the KaiB dimer-dimer interface................. 67

Figure 2.S4 Flexibility of N- and C-terminal tails in KaiB monomeric subunits ... 68

Figure 2.S5 Schematic outline of the image processing procedure followed to generate the KaiB-KaiC-Δ489 cryoEM structure ................................................. 69

Figure 2.S6 Strongest density within the KaiB-KaiC-Δ489 cryoEM structure aligns with α-helices in KaiCl ............................................................................................ 70

Figure 2.S7 Eigenimages calculated from the aligned side view particles images ............................................................................................................................... 70

Figure 3.1 3D Fourier Space refined structures of DNA-PKcs and two DNA-PKcs complexes ............................................................................................................. 81
Figure 3.2 Eigenimage sorting of DNA-PKcs/DNA with streptavidin cryoEM particle images ................................................................. 84
Figure 3.3 Eigenimage sorting of DNA-PKcs/DNA without streptavidin cryoEM particle images ............................................................... 86
Figure 3.4 Eigenimage sorting of DNA-PKcs/Ku/DNA with streptavidin cryoEM particle images ............................................................... 88
Figure 3.5 Distribution patterns of eigenimage sorted subset structures .......... 93
Figure 3.S1 Comparison of the cryoEM and crystal structures of DNA-PKcs... 101
Figure 3.S2 Gold standard Fourier Shell Correlation (FSC) curves ............... 102
Figure 3.S3 Selected class average images of a negative-stain DNA-PKcs/Ku/DNA dataset ................................................................................. 103
Figure 3.S4 Eigenimage sorting of uncomplexed DNA-PKcs cryoEM particle images .................................................................................. 103
Figure A3.1 Regions of the KaiBC intermediate structure .............................. 165
Figure A3.2 Image slices of the KaiB region from the KaiBC intermediate structure ................................................................................................. 166
Figure A3.3. Rotation correlation analysis of images slices from the KaiBC intermediate structure ........................................................................ 167
Acknowledgements

My family and Amy, who saw me through my PhD, from start to finish.

Mentors in my life goals: Ms DeYoung, Mrs Brown, Ms Stevens, Mr Rehak, Mr Patulak, and Professor Bob Rose.

My advisor, Phoebe Stewart, for her help and guidance in my growth into a scientist.

My Thesis Committee, Derek Taylor, Krzysztof Palczewski, Focco Van den Akker, and Jason Mears for their insights and critiques.

My former advisor, Anthony Forster, for the experience and opportunities I had within his lab.

The faculty and staff where I started my work, Vanderbilt University, and where I finished it, Case Western Reserve University, for their patience and knowledge.
Examining Flexible Biological Structures

Abstract

By

SETH ANDREW VILLARREAL

Cellular function is dependent on numerous pathways that carry out necessary regulatory or metabolic functions. Understanding the mechanisms within these pathways often requires overcoming issues of protein flexibility and sample heterogeneity. Cryoelectron microscopy (cryoEM) provides a technique for structure determination of proteins and complexes that can manage these difficulties. This thesis presents cryoEM structures from two projects: circadian rhythm within Synechococcus elongatus and the Non Homologous End Joining (NHEJ) pathway for DNA repair in humans. In S. elongatus circadian rhythm, three proteins, KaiA, KaiB and KaiC, use ATP to give rise to cyclical patterns of phosphorylation, signaling the time of day and subsequently triggering metabolic regulation. The protein-protein interactions between KaiB and KaiC are not well characterized. In my work, I present evidence based on a cryoEM structure of KaiBC and molecular simulations of how KaiB monomers bind a KaiC hexamer. This work supports a KaiBC interface in which KaiB binds the CII domain of KaiC, with each monomer of KaiB blocking access to one of six ATP binding
clefts on KaiC, providing greater insight into how the rhythmicity of this biological clock is achieved. In my second project, I examined complexes of the DNA protein kinase catalytic subunit (DNA-PKcs), in complex with DNA or DNA and the Ku heterodimer. In NHEJ, DNA-PKcs functions as a scaffold for the recruitment of subsequent components of the pathway, and as a kinase, performs phosphorylation for signaling and activation of other proteins. Examination of these DNA-PKcs complexes implicate the base of DNA-PKcs, separate from the kinase domain, as the binding site for DNA. DNA binding may be signaled by conformationally flexible HEAT repeats between the base and the kinase domain. Notably, the flexibility of DNA-PKcs may play a role in how DNA is protected after damage, and in the interactions between DNA-PKcs and other components of the repair process. The development of targeted therapeutics and manipulation of biological systems both benefit from an understanding of the underlying structure of proteins and their complexes, providing an impetus to characterize these systems.
Chapter 1: Introduction

Biological pathways within a cell are dependent on the structures of the individual components. Proteins, DNA, RNA, and other building blocks of cells are the basis for achieving the variety of biochemical reactions necessary for life. The interactions of biological molecules within a pathway provide the mechanisms by which cellular metabolism, and by extension, the biological processes of a multicellular organism, proceed over time. Proper protein structure is required for the diverse functional requirements of cellular processes and to ensure specificity within particular pathways, preventing overlap and crosstalk. Study of protein interactions is hampered by the same complexity that provides these interactions and this complexity limits generalization of protein preparation methods for structural analysis (Mészáros, Simon, & Dosztányi, 2011). The benefit of such efforts is insights in the mechanisms that drive cellular processes, and the opportunity for a better understanding of human health and improvement of therapeutic treatments.

Signaling and other cellular pathways are not static. Through their structures, proteins are capable of a diverse range of activities, including binding ligands, protein-protein interactions, phosphorylation, and DNA methylation. Protein complexes present challenges for structural studies, as there is often difficulty in purifying a stable complex. Identifying interactions within components of even a single system requires biochemical methods that can distinguish between random, non-relevant, associations and productive events within that system. Tightly bound complexes are more easily purified, while more unstable
complexes require less stringent purification methods, which increases the chance of contamination from non-interacting proteins. Nevertheless, characterizing protein-protein interactions is critical for an improved understanding of cellular metabolism and subsequent development of targeted treatments (Blow, 2009).

Examining Structure

Elucidating the structural architecture of protein based complexes is critical to understanding the signaling of information and the chemical modification of ligands. The characteristics of numerous proteins can make understanding their structure difficult. Protein folding, both self-induced and chaperone assisted, are controlled and limited by the sequence of the protein amino acid (AA) chain. The high number of possible sequences \(20^n\) and folding patterns for any particular protein length lead to the difficulty of a priori modeling of structure from protein sequence. However, various experimental methods can determine the structure of folded proteins at sufficiently high resolution (<3.5\(\AA\)) to characterize backbone and side chain structure, and interaction sites. Even at more moderate resolution (<20\(\AA\)), structures can provide beneficial information about the arrangement of proteins within a complex and the interaction of domains.

X-ray crystallography has been the major contributor to the understanding the structure of individual proteins. Stabilizing a protein into a crystal allows diffraction collection via exposure of the crystal to x-rays. However, in many cases proteins, or regions of proteins, are too flexible to form highly ordered
crystals. Problems with purification or crystallization tend to increase as the number of components in a complex increase. These issues make x-ray crystallography of protein complexes challenging.

Nuclear Magnetic Resonance (NMR) uses the intrinsic properties of certain isotopes to allow structure determination of small proteins, generally less than 40 kDa. Additionally, this method can provide range of motion information, providing information on the dynamics of the structure. However, NMR structure determination is usually not feasible for large protein complexes (>50 kDa).

Electron microscopy (EM) offers multiple techniques, based around single particle reconstruction or tomography, in order to work with amenable to protein complexes. While EM is capable of examining multiple structures, sample heterogeneity and protein flexibility can limit the resulting information to moderate resolution.

The difficulty in working with protein complexes is readily apparent by comparison of the over 80,000 total x-ray crystallography structures in 2014 to only ~650 examples of protein-protein complexes, as counted in 2006 (Radaev, Li, & Sun, 2006). Complexes crystalized out of solution need to be homogenous for effective x-ray diffraction, but each component can present individual issues that sum together to make the overall complex more difficult than any single component. Problems, including flexibility and heterogeneity, are compounded by the number of components in the complex.

Understanding the structure of both the individual protein and the complex allows direct comparisons between the bound and unbound states, revealing the
conformational changes that create the appropriate protein-protein interface. Often a structure of the unbound state allows recognition of the relevant portions of the protein that require further stabilization to crystallize a complex (Rasmussen et al., 2011; Rosenbaum et al., 2007). However, there is always a concern that modified proteins will lead to determination of protein complex structures that show unnatural conformations.

Two EM Structures of Protein Complexes

Single particle cryoEM was used to provide structural information on protein complexes in two projects of biological interest. The first complex is a key part of the circadian rhythm of *Synechococcus elongatus*. The second system includes complexes of the DNA Protein Kinase catalytic subunit (DNA-PKcs), which serves as a scaffold within a human DNA damage response pathway.

Circadian rhythm describes the biological processes by which an organism keeps track of time. Such systems provide a benefit through partial independence from environmental conditions, allowing an organism to minimize the effects of unusual conditions. In the case of *S. elongatus*, three proteins form the basis for this clock, KaiA, KaiB, and KaiC. Collaborations between the Stewart lab and the Egli and Johnson labs at Vanderbilt University have led to an improved understanding of this system in both protein-protein interactions and the biochemical signals of this system. Two of the proteins form the KaiBC complex, which signals the latter half of the 24 hour cycle. However, there is limited structural information available for this complex. X-ray crystallography has provided a structure of KaiC (Pattanayek et al., 2004), but efforts to examine the
complex of KaiBC by crystallography have not yielded similar results (private communication, Martin Egli). Clarification on how KaiB binds to KaiC is necessary in order to understand the role of this complex within the overall circadian rhythm system.

The DNA damage response describes the collective pathways by which an organism maintains the integrity of DNA. Complete cleavage of DNA is called a double strand break, and in humans, DNA-PKcs is one of the first proteins to bind the broken end of DNA. Only limited structural information is available on DNA-PKcs or how it interacts with DNA. Both x-ray crystallography and EM techniques have only been able to provide moderate resolution information about the structure of DNA-PKcs (Sibanda, Chirgadze, & Blundell, 2010; D. R. Williams, Lee, Shi, Chen, & Stewart, 2008). The flexibility of this protein is likely to prove challenging for x-ray crystallography, but cryoEM offers the chance to gain a greater understanding of how DNA-PKcs interacts with DNA to begin the repair process.

These complexes involved in DNA repair and circadian rhythm present challenges in the form of conformational changes and inherent flexibility, making them difficult systems for structural studies. Nevertheless, it is worth pursuing these studies to gain a greater understanding of their respective molecular mechanisms.

Circadian Clock Systems

Human health and function is affected by daily patterns of physiological changes. Activities and environmental stress can vary in their potential
physiological impact based on the timing of their occurrence. Modern study of these patterns began with the recognition of internal consistency for activity despite the removal of noted environment cues, suggesting the possibility of internal biological mechanisms providing an approximate time of day, or circadian rhythm (Aschoff, 1960; Pittendrigh, 1960). Study of these processes has proceeded from both emergent patterns of activity and study of the underlying biological mechanisms (Aschoff, 1984). Human behavior provides useful opportunities to examine the interplay between internal rhythms and the environment. In particular, shift work and changes in time zone both provide clues about the effect of disrupted circadian cycles on health (Haimov & Arendt, 1999; van Amelsvoort, Schouten, & Kok, 1999). Distinguishing behaviors influenced by circadian rhythm from the study of the underlying mechanism was facilitated by the identification of the core biological mechanism for the mammalian circadian rhythm. The suprachiasmatic nucleus (SCN) within the hypothalamus of the brain was shown to play a role in both physical activity (Stephan & Zucker, 1972) and hormone regulation (Moore & Eichler, 1972). Eventually, the SCN was demonstrated to be the master clock and pacemaker for mammalian circadian rhythm (Pando, Morse, Cermakian, & Sassone-Corsi, 2002; Ralph, Foster, Davis, & Menaker, 1990).

Identification of *Synechococcus elongatus* Circadian Cycle

Initially, inherent circadian rhythms were believed to require the existence of a nucleus and the greater complexity of eukaryotes, and that such clocks arose due to the benefits in metabolism over the generally longer lifetime of such
cells (Kippert, 1987). However, during the 1990’s, an interesting aspect of unicellular cyanobacteria was recognized. Namely that the simplicity of these organisms prevented the spatial separation of two incompatible biochemical pathways: oxygenic photosynthesis and nitrogen fixation, which is oxygen sensitive. Further temporal metabolic patterns emerged, including insensitivity to temperature fluctuation for periods of nitrogen fixation and photosynthesis (Huang, Tu, Chow, & Chen, 1990), the rates of amino acid uptake (T.-H. Chen, Chen, Hung, & Huang, 1991), and buildup of nitrogen fixation components (Schneegurt, Sherman, Nayar, & Sherman, 1994).

One particular cyanobacteria, *S. elongatus*, was especially useful for study. Luciferase reporters enabled rapid screening for the identification of clock controlled genes (Kondo et al., 1993), while subsequent screening complementation efforts eventually led to the localization of the *KaiABC* gene cluster as the likely basis for circadian rhythm in cyanobacteria (Ishiura et al., 1998). Starting at transcription, these efforts identified *KaiA* as individually transcribed, with the remaining two genes, *KaiBC*, under dicistronic transcription. Initially, this locus was tested for rescue ability against known clock cycle mutants. A plasmid with *KaiABC* was able to rescue over 50 period and arrhythmia mutants. In comparison, deletion or inactivation of any one of these three genes is sufficient to negate circadian rhythm within *S. elongatus*. Notably, despite the loss of rhythmicity, the mutants grew equivalently to WT under continuous lighting conditions. This suggested that these genes were solely
relevant to maintaining circadian rhythm, and were otherwise not involved in growth (Ishiura et al., 1998).

Moving to translation studies, Ishiura et al. also did a cursory examination of the KaiABC protein products, which are predicted to be of relatively small size, 284aa (KaiA), 102aa (KaiB), and 519aa (KaiC). Only KaiC presented recognizable functional motifs, possessing a Walker’s motif A in both the N and C terminal domains of the protein, a binding site for either ATP or GTP, and two putative catalytic carboxylate Glutamate residues, for phosphorylation. These proteins were recognized as playing a role in their own regulation.

Overexpression of KaiC from a second location and an external promoter led to repression of KaiBC, whereas overexpression of KaiA abolished rhythmicity. Based on the work in this paper, KaiABC was characterized as an essential part of a transcription-translation feedback loop, giving rise to a circadian rhythm in S. elongatus (Ishiura et al., 1998).

Kai A, B, and C as the Core Oscillator

The Kondo lab continued their work on the KaiABC proteins, looking for possible protein-protein interactions, surmising a high potential for protein-protein interactions due to the tendency of prokaryotes towards cluster organizations of genes with cooperative functions. Using a yeast two-hybrid system (Fields & Song, 1989), Iwasaki et al. were able to show the formation of homodimers of KaiB and KaiC, but the assay gave a background positive signal for KaiA, limiting assessment of homotypic interactions for that protein. This same assay also showed KaiA-KaiC and KaiB-KaiC interactions. While the sequences of KaiA and
KaiB did not possess recognizable features, the authors noted that KaiC possessed a curious repetition in its sequence, with the first half of the protein possessing 21% and 42% sequence identity and similarity, respectively, to the second half. Dubbed the CI and CII domains, a further analysis of individually expressed domains suggested that both were able to interact with KaiA and KaiB. Additionally, overexpression of either domain was sufficient to suppress KaiBC expression, and Iwasaki et al. noted mutations in both domains capable of preventing normal circadian oscillations. In the same paper, the Kondo lab demonstrated that KaiABC interactions allow KaiC to oscillate in its influence over the level of KaiBC transcription (Iwasaki, Taniguchi, Ishiura, & Kondo, 1999).

Two years later, the Johnson lab demonstrated that the degree of expression of KaiC has a direct correlation with the extent of induced phase shifting of the circadian cycle and that blocking translation and cell growth, via chloramphenicol, didn’t eliminate the overall circadian rhythm. Furthermore, incubation with ATP resulted in autophosphorylation of KaiC (Nishiwaki, Iwasaki, Ishiura, & Kondo, 2000). In recognition of the role of such a capability in circadian clock loops (Dunlap, 1999), Nishiwaki et al. hypothesized that KaiC autophosphorylation may play a role in the ability of KaiC to regulate KaiBC transcription or provide a signal to an unknown pathway (Nishiwaki et al., 2000).

The role of KaiA was originally characterized as a positive regulator of KaiBC expression, but Iwasaki et al., from the Kondo lab, followed up through examination of potential interactions between KaiA and the autokinase activity of
KaiC. In an *in vitro* assay using $^{32}$P, KaiA was found to dramatically enhance KaiC autophosphorylation. Furthermore, using 2D Thin Layer Chromatography, they identified KaiC as having both Serine and Threonine phosphorylations. Due to this KaiA-induced change on KaiC, the authors proposed that phosphorylation of KaiC affects its stability, and its ability to regulate the rate of KaiBC transcription (Iwasaki, Nishiwaki, Kitayama, Nakajima, & Kondo, 2002).

The role of KaiB remained poorly understood, despite insights on the other components in the circadian cycle. KaiA induced autophosphorylation events on KaiC, while KaiC played a role in limiting transcription of the *KaiBC* gene locus. KaiB is necessary to sustain circadian rhythm in *S. elongatus*, and Kitayama *et al.* examined temporal interactions to identify a potential role for KaiB within this system. The authors predicted that KaiB may affect the phosphorylation state of KaiC. Sidestepping the complexity of *in vivo* efforts, the Kondo lab found that, while simply mixing KaiA and KaiC promoted KaiC phosphorylation, the addition of KaiB significantly lowered the peak level of phosphorylation achieved. Through this work, the first inkling of the system of protein-protein interactions driving this circadian rhythm emerged. KaiC, influenced by its phosphorylation state, controls the circadian dependent transcription of genes. KaiA promotes autokinase activity and KaiB promotes autophosphatase activity (Kitayama, Iwasaki, Nishiwaki, & Kondo, 2003). As a result of this paper from the Kondo lab, the Johnson lab reexamined their previous paper and determined that the supposed KaiC degradation (Yao Xu, Mori, & Johnson, 2000), was more likely to represent
temporal changes in the phosphorylation state of KaiC (Yao Xu, Mori, & Johnson, 2003).

As the *S. elongatus* clock became increasingly characterized, the efficiency of this system could be compared to eukaryotic variants (Johnson, 2004). In multicellular organisms, such as rats, it is perhaps unsurprising that clusters of cells work together to signal the time, but the clocks in these cells individually are notably noisier than the averaged circadian signal (Karakashian & Schweiger, 1976; Welsh, Logothetis, Meister, & Reppert, 1995). *S. elongatus* might be expected to exhibit a similar pattern, with individual cyanobacterium showing imprecise rhythms while local groups of cells achieve a more precise rhythm through intercellular signals. However, using a particular strain of *S. elongatus* with rhythmically expressed luciferase (Katayama, Tsinoremas, Kondo, & Golden, 1999), an examination of individual cyanobacterium compared to clusters, under both standard and nonstandard lighting periods, showed that the precision of the *S. elongatus* circadian clock was intrinsic to each cell (Johnson, 2004).

Despite the role of KaiC in repression of *kaiBC* transcription, changing the promoter did not affect the patterns of oscillation (Nakahira et al., 2004). Further work, including an *in vitro* analysis of KaiC phosphorylation and dephosphorylation patterns, as well as *in vivo* studies with blocked transcription, showed that this oscillation was not reliant on transcription-translation feedback loops (Tomita, Nakajima, Kondo, & Iwasaki, 2005). The Kondo lab then expanded on this work by demonstrating that the KaiABC oscillator could
function *in vitro*, without requiring daily entrainment. Furthermore, even *in vitro*, the temperature tolerance remained consistent with *in vivo* measurements (Nakajima et al., 2005). The ability of KaiABC to maintain a circadian rhythm was eventually demonstrated to last at least 10 days *in vitro* (Ito et al., 2007).

An overall pattern of steps emerged for this oscillator. As shown in Figure 1.1, KaiA drives autophosphorylation of KaiC, initially at T432, and secondly at S431. At the halfway point, KaiB induces dephosphorylation in the same pattern.

![Figure 1.1 KaiABC interactions as the basis of the *S. elongatus* circadian system](image)
The three Kai proteins form the basis of this system. Unphosphorylated KaiC signals the morning. KaiC hexamers present the CI and CII domain, of which the CII domain binds KaiA. The KaiAC complex induces phosphorylation of KaiC, first at threonine 432 and subsequently at serine 431. Fully phosphorylated KaiC leads to recruitment of KaiB and then sequestration of KaiA. KaiB induces dephosphorylation, beginning with threonine 432 and then at serine 431. Unphosphorylated KaiC dissociates from KaiB. This diagram is based on a model describing the overall interactions of this pathway and does not show the actual protein-protein interactions of this system (Johnson et al., 2008). Scale bar is 50 Å.
of T432 followed by S431. Each step regulates the reaction of the subsequent step (Nishiwaki et al., 2007). The ATPase activity of KaiC is temperature dependent, but interactions within the ring structure of KaiC are proposed to activate or repress functionality in a way to compensate for these fluctuations, thus allowing KaiC to always ensure appropriate activity rates in the overall cycle (Murayama et al., 2011; Terauchi et al., 2007). The system of *S. elongatus* circadian rhythm shows both benefits for study and unique features compared to other model organisms. The KaiABC proteins are the basis of this system, and provide a straightforward and simple model for the study of circadian systems.

Structural Aspects of the KaiABC System

The capabilities and resiliency of the KaiABC circadian rhythm is a result of the functional aspects of the individual proteins as well as the protein-protein interactions of this system. In addition to the study of this system, structural analysis of the components began to clarify the physical features of the individual components. Initial structural assessment of KaiC used negative stain EM to demonstrate that KaiC formed hexameric rings, with the dual domain of a KaiC monomer giving rise to a double barrel shape around a central channel (Mori et al., 2002). Two years later, three papers provided a jump in understanding of KaiC. The Kondo lab made use of mass spectroscopy and alanine mutation to identify serine 431 and threonine 432 as two autophosphorylation sites of KaiC (Nishiwaki et al., 2004). The Egli lab provided a crystal structure of KaiC, which showed the ATP binding sites, the location of known mutations, and phosphorylation on threonine 432. The first KaiC crystal structure did not provide
information on the C-terminal tails of KaiC, suggesting a greater degree of flexibility relative to the rest of the structure (Pattanayek et al., 2004), but a more complete structure was later solved (Pattanayek et al., 2006). Further work from Vanderbilt University, jointly from the Johnson and Egli labs, examined potential phosphorylated residues, identifying a third critical residue, threonine 426, which is adjacent to serine 431 and may be phosphorylated, or shuttle a single phosphate between itself and serine 431. Furthermore, mutation of this residue affects circadian rhythmicity (Pattanayek et al., 2009; Yao Xu et al., 2004, 2009).

Of the three Kai proteins, only KaiC is known to undergo chemical modifications over the course of a circadian period. KaiC carries out both autokinase and autophosphatase activities, and even transiently generates ATP while undergoing dephosphorylation (Egli et al., 2012; Nishiwaki & Kondo, 2012).

The study of KaiA structure and its interactions with KaiC was facilitated by NMR and x-ray crystallography. An initial NMR study from the LiWang lab, at Texas A&M University, identified the C terminus of KaiA as binding to KaiC and stimulating autophosphorylation (S. B. Williams, Vakonakis, Golden, & LiWang, 2002). A continuation of this work led to an NMR structure of the C-terminal KaiA domain in complex with the C terminal tails of KaiC (Vakonakis & LiWang, 2004). Both of these analyses used KaiA from *Thermosynechococcus elongatus*, but there is high homology (61% sequence identity) with KaiA of *S. elongatus*. This same year, the crystal structure of KaiA was published, showing an entwined dimer with swapped domains, meaning that the N terminal domain of one chain
is paired with C terminal domain of the other (Ye, Vakonakis, Ioerger, LiWang, & Sacchettini, 2004).

Two papers in early 2008 provide an initial baseline for assessing KaiB-KaiC interactions. A SAXS analysis of the complexes at various points in the cycle, and comparison to the envelopes of the individual proteins, led to the conclusion that KaiB binds KaiC as a pair of dimers or a tetramer to prevent KaiA induced autokinase activity (Akiyama, Nohara, Ito, & Maéda, 2008). A collaborative paper from the Johnson, Stewart, and Egli labs a month later provided a host of information about the structure and interactions of KaiB and KaiC. A *T. elongatus* KaiB crystal structure, which included the C terminal tails, showed the tails as adopting multiple conformations, including extended from the protein and folded up against it. Examination of individually expressed domains of KaiC by negative stain EM noted that the CI domain retained the ability to form hexameric rings, while the CII domain did not, suggesting that the stable formation of the KaiC hexamer was dependent on the CI domain. Furthermore, KaiB solely interacts with the CII domain, when assayed by native PAGE gel with the isolated domains. However KaiB showed no affinity for the C terminal tails of KaiC, in contrast to KaiA, which the authors interpreted as reducing the likelihood of KaiB directly competing with KaiA for a binding site. Examining the full proteins by negative stain EM showed a triple layer structure, with a discernable double barrel shape for KaiC, and a third layer of density attributed to KaiB. While the resolution did not permit assignment of individual KaiB monomers, the volume of the ring was considered appropriate for two KaiB dimers, four subunits in total,
binding to the top of a KaiC hexamer, on opposite sides of the central channel. The link between the roles of KaiB and KaiA was predicted to be reliant upon conformational shifts in the ATP binding clefts of KaiC, induced by KaiB, and thus negating the effects of KaiA (Pattanayek et al., 2008).

The crystal structure of KaiB shows a tetramer, while EM structural data indicated that KaiB binds KaiC as a dimer. The Ishiura lab examined potential effects for the tetrameric oligomer in maintaining the circadian oscillations. They found that a deletion of the negatively charged C terminal residues of KaiB, which were known to weaken \textit{in vivo} circadian rhythms (Iwase et al., 2005), also prevented formation of the KaiB tetramer. While \textit{in vitro} experiments demonstrated that the core KaiABC oscillator functioned normally with the modified form of KaiB, the mutation \textit{in vivo} produced a weakened circadian cycle. Additionally this modified system showed reduced circadian linked gene expression, suggesting that understanding the role of KaiB may require assessing multiple oligomers of KaiB in their interaction with KaiC and other proteins (Murakami et al., 2012).

The LiWang lab sought to further evaluate the interaction between KaiB and KaiC. They expressed each KaiC domain individually, with a S431E mutation on the Cl domain, such that both individual domains could form hexamers. When either domain was mixed with KaiB, and examined by NMR (using $^{15}$N labeling), only the Cl domain experienced a change in spectra. Furthermore, they found that KaiB and only the Cl domain would form complexes in gel filtration chromatography. On the basis of this work, they concluded that the difficulty in
interpreting the SAXS and EM structures due to the similarity of these domains had led to a mistaken assignment of the KaiB binding location (Chang, Tseng, Kuo, & LiWang, 2012).

In response to these concerns, the Egli lab collected negative stain EM images that would more conclusively visualize the domain of KaiC providing the binding surface for KaiB. In order to distinguish between the two domains of KaiC, they attached a gold nanoparticle to the CII domain. The additional density was readily observable, and allowed comparison between images of KaiC and KaiBC. The resulting image analysis supported KaiB binding to the CII domain of KaiC (Pattanayek, Yadagirib, Ohi, & Egli, 2013). Due to the conflicting data on the nature of KaiB and KaiC interactions, an understanding of the mechanisms by which KaiB influences the phosphorylation state of KaiC remains elusive.

DNA Maintenance and Human Health

DNA lesions can occur in a variety of ways, and represent a threat to the continued survival of a cell, necessitating a repair response for reliably recovering from damage. In many cases, DNA exists as only a single copy within a cell, and often the sequence of DNA required for survival is a unique molecule within a cell. Repair of DNA lesions must be able to occur without access to the information contained within any particular section of DNA.

The direct relationship between genetic information and the development of cancer makes DNA lesions an inherent concern (Bartek, Bartkova, & Lukas, 2007). Damage to the DNA of a cell results in one of three outcomes: survival, replicative senescence, and death (Campisi, 2005; Hoeijmakers, 2001). While
survival may seem the most useful outcome, the imperfect nature of DNA repair means that a repaired double strand of DNA may not necessarily match the original sequence prior to being damaged. A cell can accumulate damage over time, and eventually this may even trigger uncontrolled proliferation and cancer. One difficult aspect of examining cancer is the number of possible variations in the genome. The $p53$ gene is the most commonly mutated gene in human cancer, but a number of mutations can affect disease progression, and the tendency towards ongoing mutation, make examination of the causes of any particular instance difficult (Giglia-Mari & Sarasin, 2003; Hussain & Harris, 2000).

Although cancer is widely recognized, there are a number of other diseases and disorders that arise from defects in the cellular DNA damage response. Many of these problems are genetic and inherited. Studying the consequences of genetic defects in the DNA repair process allows a way to investigate the mechanisms of this process and facilitates the development of more precise therapeutic treatments (Wijnhoven et al., 2005).

Types of Damage Affecting the Integrity of DNA

Over the course of a day, an average human cell experiences up to an estimated 10,000 instances of DNA lesions (Lindahl, 1993; Sander et al., 2005). While RNA and proteins are critical to the survival of a cell, in the majority of cases any single RNA sequence or protein can be degraded and replaced when it is damaged or malfunctioning. DNA must be maintained within a cell, and DNA damage must be repaired.
In most cases, the cellular DNA damage response is successful and a cell continues in its role within the body. However, there are other alternatives. Programmed senescence describes the accumulation of age-related cellular damage, including DNA lesions that trigger a cell to cease replicating. A senescent cell does not divide into two daughter cells and therefore avoids increasing the population of cells with those errors. A decisive response to irreparable DNA damage is cell death, which prevents the corruption of genetic information from causing any further damage to the organism (Hoeijmakers, 2009).

Ideally, DNA lesions can be repaired, and there are a variety of mechanisms in a cell that carry out such repairs. Broadly speaking, in order to restore DNA, the type of damage must be correctly identified and then the DNA must undergo a chemical reaction to recreate the appropriate structure. Shown in Fig. 1.2 are general categories of the structural changes found in DNA lesions. Three types of lesions involve only the bases of DNA, and are termed base adducts. A missing base leaving a gap, called an abasic site, results from destabilized bond between the base and the backbone. Base pair mismatches (commonly A/C or G/T pairings) occur as a result of damaged bases being replicated by the cell. One example is cytosine deamination, which converts a cysteine to a uracil, leading to a synthesis of a complementary A instead of a G. The third type of base lesion is a chemical modification to a base, such as UV induced linkage between thymine dimers or guanine methylation. These three
types of base adducts can be repaired from the complimentary chain of DNA, which possesses the appropriate base to ensure information is not lost.

Figure 1.2 Diagram of various types of DNA damage
Both the backbone of DNA (black) and the bases (blue) are susceptible to damage (red), which prevents the normal function of DNA. In undamaged DNA the backbones of both strands are continuous and the bases are appropriately paired. An abasic site is missing one half of a base pair. Base pair mismatch occurs when AT and GC pairing is not maintained, while in some cases the chemical structure of each strand of DNA is correct, the sequence information is wrong. Modified bases can arise from the formation of dimers between bases or chemical modification of a single base. Interstrand crosslinks prevent the unzipping of DNA, and are a covalent linkage between the two strands. Single and double strand breaks of DNA are gaps in one or both backbones.
A separate type of damage is interstrand crosslinks. Normally, each DNA strand is a distinct molecule, and the transient separation of the complimentary strands is an essential part of replication and transcription (Deans & West, 2011). A covalent linkage between the two strands can limit the requisite flexibility. This particular type of damage is frequently the result of lipid peroxidation (Kozekov et al., 2003; Stone et al., 2008), and factors such as a high fat diet and alcoholism can increase the instances of interstrand crosslinks (Brooks & Theruvathu, 2005; Folmer, Soares, Gabriel, & Rocha, 2003).

The two remaining categories of DNA lesions are single and double strand breaks (SSB and DSB) in the backbone of DNA (Iyama & Wilson, 2013). In the first case, a single backbone of a DNA double strand is broken. A random and uncontrolled single strand break is dangerous (Sander et al., 2005), as it allows base degradation from the site of the break (Thompson & West, 2000). A more dangerous, but rarer, form of damage is a double strand break (Bohgaki, Bohgaki, & Hakem, 2010). Unlike the ends of chromosomes, which are capped by protective proteins bound to a telomere (Rooney et al., 2003), an unprotected double stranded end of DNA is subject to degradation. If this lesion occurs without a sister chromatid the genetic information lost is likely irrecoverable (Khanna & Jackson, 2001). In both single and double strand break lesions, the damaged backbone of DNA presents an immediate issue that must be addressed by the cell to preserve the integrity of the overall strands of DNA.
Repairing Double Strand DNA Breaks

Despite the frequency and dangers to an organism, in most cases the DNA repair response of a cell is sufficient to recover all of the genomic information. The incidence of daily lesions would mean that even a 1% error rate would quickly render cells cancerous or trigger cell death pathways. There are a variety of mechanisms used by the DNA damage response pathways, and the method used depends both on the type of damage and the cellular process recognizing the lesion. In some cases, the simplest response is to ignore the damage. A damaged base within an intron need not affect a final protein sequence, but transcription stalling due to damage would limit the population of the protein. In that manner, there are several polymerases capable of bypassing base adduct lesions (Andersen, Xu, & Xiao, 2008). Base adducts are targeted by base-excision repair (BER), in which first the damaged base is removed, creating an abasic site. The remaining portion of the nucleotide can be cut away from the DNA strand, and the resulting gap filled in by DNA synthesis with pol β (Barnes & Lindahl, 2004; Caldecott, 2008; Slupphaug, Kavli, & Krokan, 2003). Larger forms of damage that distort the overall shape of a DNA helix fall under the category of nucleotide excision repair (NER) which cuts away the entire nucleotide and then uses a polymerase (δ, ε and/or κ) to fill in the resulting gap (Gillet & Schärer, 2006; Sugasawa, 2006).

Uncontrolled DNA breaks can affect either one or both strands of DNA. Such breaks can occur as part of the other types of DNA damage responses, such as when excising a chemically modified nucleotide. Through the
complementarity of DNA, the repair of SSBs generally results in the complete recovery of genetic information. If necessary, any damaged nucleotides can be excised through other pathways, and then DNA synthesis replaces missing nucleotides, allowing the remaining nick in the backbone of DNA to be repaired by DNA Ligase I (Hegde, Hazra, & Mitra, 2008).

DSBs create four unprotected ends of DNA within the cell. During the G1 growth phase, there is no sister chromatid, but homologous regions within the genome may be present. In comparison, the G2 phase allows use of a duplicated chain of DNA as a model to build a complete duplicate of the original sequence. The S-phase and M-phase of the cell cycle present indeterminate states between the presence and absence of sister chromatids for each chromosome, as the genome is either in the process of replication or separation into separate cells. For any individual break, a homologous section of DNA may or may not be available for reference. Homologous recombination (HR) describes the repair process using the sister chromatid as a template, while nonhomologous end joining (NHEJ) provides a repair pathway for DNA independent of a template. Finally, there is third option, microhomology-mediated end joining (MMEJ), which is more recently recognized and uses small regions of similar DNA, but not sister chromatids, to replace degraded ends of DNA (Symington & Gautier, 2011).

Components of the NHEJ Pathway

Many aspects of the NHEJ repair pathway remain poorly understood, as this system has proven difficult to characterize. The repair of double strand DNA breaks is initiated by the Ku heterodimer, which plays a critical role as a key
detection mechanism for DSB events. Forming a ring, Ku is a heterodimer that can bind DNA, sliding over the broken end of DNA within seconds of a break occurring (Britton, Coates, & Jackson, 2013; Downs & Jackson, 2004). Breaks in DNA are usually not sequence specific and they can occur in any location. Ku compensates for this by interacting with the backbone of DNA, rather than the bases, allowing the heterodimer a consistent binding surface for any DNA sequence (J. R. Walker, Corpina, & Goldberg, 2001). DSB breaks leave exposed ends of DNA, which unlike the telomere capped ends of chromosomes, are susceptible to degradation. Ku blocks various end processing enzymes, thereby protecting the broken ends of a DSB (Sun, Lee, Davis, & Chen, 2012). The majority of the remaining components of the NHEJ pathway directly interact with Ku. The process begins with the recruitment of DNA-PKcs (Uematsu et al., 2007), and includes the ligase factors XRCC4 and ligase IV (Costantini, Woodbine, Andreoli, Jeggo, & Vindigni, 2007; McElhinny, Snowden, McCarville, & Ramsden, 2000). In contrast to a sequential addition of factors, the NHEJ complex appears to assemble with Ku and DNA-PKcs as a scaffold for the collective stabilization of the remaining components in a dynamic and concerted manner, potentially providing the flexibility to work with a wide variety of lesions occurring during a double strand break, but hampering efforts to understand the binding order and other aspects of how the NHEJ complexes form during repair (Cottarel et al., 2013; Yano, Morotomi-Yano, Adachi, & Akiyama, 2009).

DNA-PKcs and Ku form a complex referred to as DNA-PK, which has a central role in the NHEJ pathway (Gottlieb & Jackson, 1993). DNA-PKcs is a
member of the phosphatidylinositol-3 (PI-3) kinase-like kinase family (PIKK), and contains both a PI-3 kinase domain, as well as numerous HEAT (Huntington-elongation-A-subunit-TOR) repeats (Hartley et al., 1995; Perry & Kleckner, 2003). Structural studies of DNA-PKcs revealed an overall shape with two broadly discernable parts (Sibanda et al., 2010; D. R. Williams et al., 2008). The C-terminus of the sequence contains the kinase domain, and forms a ‘crown’ atop the remaining portion of the protein, with the HEAT repeats forming two arms in a pincer shape. The x-ray crystal structure shows the position and shape of the kinase domain, with two arms extending from it (Sibanda et al., 2010). In comparison the cryoEM structure, from the Stewart lab, shows a similar volume in the area of the kinase domain, but a much larger volume for the base (D. R. Williams et al., 2008). This is likely due to the crystal structure being at a resolution (6.6Å) only sufficient to identify 46% of the backbone residues, which in the N terminus, were mainly the HEAT repeats in the arms of the structure (Sibanda et al., 2010). Due to the complexity of the NHEJ pathway, and the requirements to deal with numerous types of DNA lesions, the differences between the x-ray and cryoEM structures maybe represent areas of flexibility or the tendency of DNA-PKcs to undergo conformational changes.

The initial formation of DNA-PK results in translocation of Ku along the length of DNA, allowing activation of DNA-PKcs kinase activity (Yoo & Dynan, 1999). In comparison, DNA-PKcs in the unbound state has limited or no kinase activity (Hammarsten & Chu, 1998). Interestingly, the kinase domain, which is near the C-terminus, is affected by the N-terminus, and deletion of the N-terminal
region of the protein results in constitutive activity (Davis, Lee, & Chen, 2013; Meek, Lees-Miller, & Modesti, 2012). A functional kinase domain of DNA-PKcs is essential for proper function during the NHEJ process. Without that capability both DSB repair and V(D)J recombination fail (Kurimasa et al., 1999). Surprisingly, while DNA-PKcs can phosphorylate in vitro the canonical NHEJ components, including Ku (D. W. Chan, Ye, Veillette, & Lees-miller, 1999), DNA ligase IV (Y.-G. Wang, Nnakwe, Lane, Modesti, & Frank, 2004), and Artemis (Mahaney, Meek, & Lees-Miller, 2009; C. Wang & Lees-Miller, 2013), this phosphorylation activity is not required for NHEJ. Part of this may be explained through redundancy via ATM, which is in the same protein family as DNA-PKcs and may be capable of supporting defects in DNA-PKcs function. Still, these proteins are not interchangeable, and continuing assessment of the NHEJ pathway suggests that unique DNA-PKcs phosphorylations may be relevant to efficient DSB repair (Kusumoto-Matsuo et al., 2014).

DNA-PKcs autophosphorylates, resulting in inactivation and dissociation in vitro, but the physiological roles of at least 40 phosphorylation sites throughout the protein remain poorly understood (Dobbs, Tainer, & Lees-Miller, 2010; Douglas et al., 2007; Merkle et al., 2002; Olsen et al., 2010). Particular phosphorylation sites are better understood, such as Ser2056, an autophosphorylation site which is recognized as improving the efficiency of the NHEJ process (Cui et al., 2005). Thr2609 is phosphorylated by ATM and loss of this site by alanine substitution sensitizes cells to replication stress, and this mutation leads to an early death in mice (Uematsu et al., 2007; Yajima, Lee, &
Chen, 2006; S. Zhang et al., 2011). A functional role for Thr2609 has been proposed in connection to protein-protein interactions, which lead to recruitment of Artemis (Goodarzi et al., 2006). These two phosphorylation sites, Ser2056 and Thr2609, may be in opposition for proper DNA-PKcs function, with phosphorylated Ser2656 limiting DNA processing, and phosphorylated Thr2609 promoting it (Cui et al., 2005; Meek, Douglas, Cui, Ding, & Lees-Miller, 2007). Overactive processing could lead to excessive loss of genetic information, but the inability to remove damaged bases is a problem of the opposite extreme. Maintaining the balance between ensuring the removal of damaged bases and minimizing the loss of genetic information is a delicate process.

Nucleolytic activity is required for DSB repair of sites with chemical modification to one or both of the broken ends. This activity is required in many cases for double strand breaks, and nucleases, both Artemis and others, play an essential role in the repair of DNA. However, the number of possible proteins involved and the numerous types of potential damage complicates the activities and roles of these proteins. While a number of enzymes are recruited during NHEJ, not all are likely necessary, and one proposal for the selection of the appropriate repair factors is a guess and check process, in which a randomly selected processing enzyme is provided access to the damaged end of DNA, and then after the repair attempt, the end of DNA can either be ligated together or are passed back to the start of the prior step to again attempt to remove damage bases (Strande, Waters, & Ramsden, 2012). A more procedural method suggests that Ku and DNA-PKcs may recruit additional proteins based on the
complexity of the DNA lesions, with support for this idea coming from noted changes in the dynamics of NHEJ complex formation depending on the type of induced DNA damage (Reynolds et al., 2012).

The maintenance of DNA within an organism allows the reliable continuation and propagation of genes in spite of numerous types of damage that would create mutations or simply degrade the DNA. In humans, the DNA damage response is a complex process in which at least three different pathways, homologous recombination, microhomology-mediated end joining, and nonhomologous end joining, repair the wide variety of chemical modifications that can occur on DNA. Shared proteins and functions between NHEJ and the other two repair pathways have made understanding these systems a complex undertaking. These efforts remain important though, as a malfunction within a DNA repair pathways can lead to a variety of health disorders and a predisposition to cancer. The genetic underpinnings of such issues make treatments difficult, and are usually restricted to mitigation of symptoms and limiting the consequences on the disease. In comparison to the other pathways in the DNA damage response, NHEJ in particular is critical to human health because of its additional role in the adaptive immune system. Understanding the components of NHEJ and how they interact to fulfill their roles in DNA repair are critical for developing better treatment options and therapeutics for genetic diseases of these repair pathways and as potential targets in cancer treatments.
Chapter 2: CryoEM and Molecular Dynamics of the Circadian KaiB–KaiC Complex Indicates KaiB Monomers Interact with KaiC and Block ATP Binding Clefts

This chapter was published in the Journal of Molecular Biology in 2013

Abstract

The circadian control of cellular processes in cyanobacteria is regulated by a posttranslational oscillator formed by three Kai proteins. During the oscillator cycle KaiA serves to promote autophosphorylation of KaiC while KaiB counteracts this effect. Here we present a crystallographic structure of the wild-type *Synechococcus elongatus* KaiB and a cryoEM structure of a KaiBC complex. The crystal structure shows the expected dimer core structure and significant conformational variations of the KaiB C-terminal region, which is functionally important in maintaining rhythmicity. The KaiBC sample was formed with a C-terminally truncated form of KaiC, KaiC-Δ489, which is persistently phosphorylated. The KaiB-KaiC-Δ489 structure reveals that the KaiC hexamer can bind six monomers of KaiB, which form a continuous ring of density in the KaiBC complex. We performed cryoEM guided molecular dynamics flexible fitting simulations with crystal structures of KaiB and KaiC to probe the KaiBC protein-protein interface. This analysis indicated a favorable binding mode for the KaiB monomer on the CII end of KaiC, involving two adjacent KaiC subunits and spanning an ATP binding cleft. A KaiC mutation, R468C, which has been shown to affect the affinity of KaiB for KaiC and lengthen the period in a bioluminescence rhythm assay, is found within the middle of the predicted KaiBC interface. The proposed KaiB binding mode blocks access to the ATP binding cleft in the CII ring of KaiC, which provides insight into how KaiB might influence the phosphorylation status of KaiC.
Keywords

Circadian oscillator/ cryoEM/ MDFF/ protein-protein interface / Synechococcus elongatus

**Abbreviations:** cryoEM, cryo-electron microscopy; EM, electron microscopy; FSC, Fourier shell correlation; KaiABC, KaiA-KaiB-KaiC; KaiAC, KaiA-KaiC; KaiBC, KaiB-KaiC; KaiCI, KaiC N-terminal domain; KaiCII, KaiC C-terminal domain; MDFF, molecular dynamics flexible fitting; PTO, posttranslational oscillator; SAXS, small angle X-ray scattering; TTFL, translation-transcription feedback loop

**Graphical Abstract**

**Highlights**

- The cyanobacterial circadian oscillator is composed of three interacting proteins
- A KaiB crystal structure shows variation in functionally important C-terminal tails
- A KaiBC cryoEM structure shows six KaiB monomers per KaiC hexamer
- MDFF fitting of KaiB and KaiC structures indicates KaiB binds the CII side of KaiC
- KaiB binds near the ATP binding cleft of KaiC and may block ATP access
Introduction

The cyanobacterial clock is an elegant nanomachine that maintains remarkable synchronicity despite cell division and temperature fluctuations and can be entrained by light/dark cycles possibly involving light-driven changes in energy metabolism (Egli & Johnson, 2013; Rust, Golden, & O’Shea, 2011). *Synechococcus elongatus* contains one of the simplest clocks in which three proteins, KaiA, KaiB and KaiC in the presence of ATP, form a posttranslational oscillator (PTO) (Nakajima et al., 2005). *In vitro*, this clock ticks through daily cycles of KaiC phosphorylation and dephosphorylation for up to 10 days in the absence of a translation-transcription feedback loop (TTFL) (Ito et al., 2007). This clock is temperature compensated (Nakajima et al., 2005; Terauchi et al., 2007), which minimizes the role of daily environmental fluctuations, remains accurate during cell divisions, and has a built in ratcheting mechanism to ensure that the *in vitro* KaiABC reaction is unidirectional (Johnson, Egli, & Stewart, 2008; Johnson, Stewart, & Egli, 2011).

Study of this system, including crystal structures of individual proteins and evaluation of protein-protein interactions, has revealed salient features of the PTO and its relationship to a TTFL for modulation of transcriptional oscillation of clock regulated genes (Johnson et al., 2008). During the course of a single PTO cycle there is a strict order of phosphorylation and dephosphorylation of KaiC. KaiA promotes KaiC phosphorylation first at T432 and subsequently at S431 (Nishiwaki et al., 2007; Rust, Markson, Lane, Fisher, & O’Shea, 2007). KaiB is then recruited to the KaiA-KaiC complex, which sequesters KaiA, and allows the
circadian cycle to proceed (Brettschneider et al., 2010; Pattanayek et al., 2011; Qin et al., 2010; Rust et al., 2007). KaiC dephosphorylates, at least partially via ADP phosphotransferase activity (Egli et al., 2012; Nishiwaki & Kondo, 2012) pT432 first followed by pS431, which returns KaiC to the initial state of the cycle (Nakajima et al., 2005). Throughout the oscillation cycle, KaiC exists as a population differentiated by the ratios of various states (monomeric and hexameric KaiC, KaiAC, KaiBC, and KaiABC). Synchronicity is maintained in part by subunit exchange of the KaiC hexamer with the free monomers within the cell (Kageyama et al., 2006; Mori et al., 2007) and by KaiA sequestration (Qin et al., 2010; Rust et al., 2007; Van Zon, Lubensky, Altena, & Wolde, 2007). In vivo, the histidine kinase SasA relays the phosphorylation state of the PTO to transcription factors, including RpaA and RpaB, to influence global transcription rhythms (Hanaoka et al., 2012; Iwasaki et al., 2000; Takai et al., 2006).

The *S. elongatus* KaiC crystal structures of wild-type, mutants, and phospho-mimics reveal a hexamer, forming a “double-doughnut” of stacked N-terminal CI and C-terminal CII rings (KaiCI and KaiCII) with a central channel, and two main phosphorylation sites, S431 and T432, at the subunit interface (Pattanayek et al., 2006, 2009). Despite a high degree of homology between the two KaiC domains, they appear to have distinct functions. The CI domain may act as an input-independent timer and the CII domain may sense the ATP/ADP ratio in the cell (Phong, Markson, Wilhoite, & Rust, 2013). The KaiC hexamer binds 12 ATP molecules, one between each subunit interface in both the CI and CII rings (Pattanayek et al., 2004). The *S. elongatus* KaiA crystal structure shows
a dimer with a domain swapped arrangement (Ye et al., 2004). KaiA binds the KaiCII ring via the KaiC C-terminal tails, and stimulates KaiC autokinase activity (Egli et al., 2013; Kim, Dong, Carruthers, Golden, & LiWang, 2008; Pattanayek et al., 2006; Vakonakis & LiWang, 2004). KaiB interacts with the phosphorylated form of KaiC and facilitates the return of KaiC to the unphosphorylated state (Kitayama et al., 2003). In crystals and in solution KaiB forms a tetramer (Garces, Wu, Gillon, & Pai, 2004; Hitomi, Oyama, Han, Arvai, & Getzoff, 2005; Iwase et al., 2005; Pattanayek et al., 2008). The component protein structures provide a good starting point for biophysical characterization of the assemblies formed by this nanomachine during the PTO cycle.

The KaiBC complex has been studied by EM (Pattanayek et al., 2008), SAXS (Akiyama et al., 2008; Pattanayek et al., 2011), and most recently by EM with gold labeling of KaiC in complex with KaiB (Pattanayek et al., 2013). In the first SAXS study, a tetrameric state of KaiB was assumed from the existing crystal structures (Akiyama et al., 2008). In contrast, the first EM study of KaiBC clearly showed a third layer of density above the KaiC hexamer and with a shape appropriate for two KaiB dimers (Pattanayek et al., 2008). The finding that KaiB binds KaiC as a dimer was unexpected because all KaiB crystal structures show a tetrameric state (Garces et al., 2004; Hitomi et al., 2005; Iwase et al., 2005; Pattanayek et al., 2008). A more recent SAXS study of KaiB alone and in complex with KaiC confirmed a tetrameric state for KaiB in solution and a dimeric state in the KaiBC complex (Pattanayek et al., 2011). Modeling of this KaiBC SAXS envelope with a pair of KaiB dimers, a single tetramer, or a pair of
tetramers indicated a best match with two KaiB dimers bound to the KaiC hexamer, consistent with the EM-based model (Pattanayek et al., 2008).

The similarity of the two ends of the KaiC hexamer, the CI and CII rings, has made the question of where KaiB binds on the KaiC hexamer difficult to answer and has led to a continuing debate. In the Akiyama et al. SAXS study, it was supposed that KaiB binds to the CII ring of KaiC on the basis of the KaiC C-terminal tails, which were assigned within the low-resolution SAXS envelope on the same side of the complex as KaiB (Akiyama et al., 2008). A few months later a native PAGE assay of the individually expressed *S. elongatus* KaiC CI and CII domains indicated that *S. elongatus* KaiB binds CII but not CI (Pattanayek et al., 2008). In a more recent SAXS study by Pattanayek et al. the envelope for KaiC in solution showed the KaiC C-terminal tails protruding from the CII side of the hexamer (Pattanayek et al., 2011). In the KaiBC SAXS envelope of the same study, the KaiC C-terminal tails were docked between two KaiB dimers modeled on the CII side of the hexamer to produce a model that agreed with the SAXS envelope. The question of where the KaiB binding site is located on the KaiC hexamer was revisited by LiWang and coworkers with gel-filtration chromatography and NMR binding experiments (Chang et al., 2012). Their results indicated that a mutated form of *Thermosynechococcus elongatus* KaiB, which preferentially assembles as dimers in solution, formed detectable complexes with an individually expressed form of the KaiC CI domain, but not with the individually expressed KaiC CII (S431 mutant) domain. These results led them to propose that KaiB must bind to the CI ring of the intact KaiC hexamer.
Phong et al. have shown that KaiC with two catalytic mutations in the CI ring, E77Q and E78Q, does not bind KaiB (Phong et al., 2013). The location of these mutations is within the middle of the CI ring, rather than on the surface of KaiC, suggesting that their effect must be transferred to the KaiB binding site on KaiC. Most recently, EM studies of KaiB in complex with KaiC S431D mutant hexamers with C-terminal His6 tags labeled with nickel(II) nitrilotriacetic acid gold nanoparticles (Ni-NTA-Nanogold) showed that the dense gold particles are located on the same side as KaiB (Pattanayek et al., 2013). This work provides direct evidence for KaiB binding on the CII ring of the KaiC hexamer, which places KaiB close to the main KaiC phosphorylation sites, and near the KaiA binding site on the KaiC C-terminal tails.

Critical interactions between two loops of a KaiC subunit are disrupted upon KaiA binding (Egli et al., 2013). This local disruption is propagated to neighboring KaiC subunits, leading to enhanced flexibility within the KaiC CII ring and increased phosphorylation of KaiC. The action of KaiA in promoting KaiC phosphorylation is opposed by KaiB and may involve sequestration of KaiA in a ternary KaiABC complex, in which KaiA has an alternative binding mode to KaiC or KaiBC (Pattanayek et al., 2011). The molecular mechanisms underlying the action of KaiB on KaiC and the KaiABC complex are poorly understood. Knowledge of the protein-protein interface between KaiB and KaiC would help elucidate the molecular mechanisms underlying the action of KaiB during the PTO cycle. We present a crystal structure of wild-type *S. elongatus* KaiB and a cryoEM structure of a KaiBC complex. We used molecular dynamics flexible
fitting simulations to examine the protein-protein interface of KaiB relative to KaiC. We propose that six KaiB monomers interact with a KaiC hexamer at the KaiCII side and effectively block the ATP binding clefts on KaiCII.

Results

Crystal structure of *S. elongatus* KaiB

Crystal structures of the KaiB proteins from the cyanobacteria *Anabaena* (Garces et al., 2004), *Synechocystis* (Hitomi et al., 2005), and *T. elongatus* [T64C mutant (Iwase et al., 2005) and wild-type (Pattanayek et al., 2008)] were previously determined. In all cases the protein forms a dimer of dimers with the core (amino acids Y7-Y93; *S. elongatus* KaiB numbering) of monomeric subunits and the tetramer adopting very similar conformations in all four structures, consistent with highly conserved sequences across species in that region (Fig. 2.S1). Crystals of *S. elongatus* KaiB were grown from mixtures of the KaiB and KaiC proteins, and the structure was determined by molecular replacement at a resolution of 2.6 Å. The asymmetric unit in space group *C2* reveals a single tetramer (Fig. 2.S2) and, except for six N-terminal residues from subunit a, all 102 residues for subunits a to d could be built into the electron density. Examples of the quality of the final electron density are depicted in Fig. 2.1A,B, and selected data collection and refinement parameters are summarized in Table 2.1.
Figure 2.1 Crystal structure of *S. elongatus* KaiB

Quality of the final electron density (A) in the region of the C-terminal tail of subunit a, and (B) around residues P69 to I75 of subunit c. The Fourier 2Fo-Fc sum electron density map is drawn at the 1σ level. The KaiB dimer (c, green, and d, cyan, subunits) viewed (C) approximately along the non-crystallographic dyad and (D) rotated around the horizontal and the vertical to illustrate the wishbone-like arrangement of N- and C-terminal tails. N- and C-terminal portions of the ribbon are highlighted in black and red, respectively, with side chains of selected residues depicted in ball and stick mode.

The KaiB dimer resembles a trapezoid with a base dimension of ca. 60 Å, a height of 30 Å and a thickness (Y7-Y93 core) of ca. 30 Å (Fig. 2.S2A).

However, N- and C-terminal tails from subunits a/c and b/d (a/b, dimer 1, and c/d, dimer 2) jut out on opposite sides of the more or less flat KaiB tetramer, thus approximately doubling the thickness of the assembly (Fig. 2.S2B). Within the monomer residues Y7 and Y93 stitch together two β-strands six and eight residues from the N- and C-termini, respectively. These tyrosines also play a key

role in the stabilization of the dimer-dimer interface, where they form mostly hydrophobic contacts (Fig. 2.S3). Their importance in the formation of the dimer of dimers is supported by the fact that the KaiB Y7, 93A double mutant was found to exist predominantly in the dimeric form (Chang et al., 2012). N- and C-terminal tails extend in a wishbone-like arrangement from the Y7-Y93 core structure (Figs. 2.1CD, 2.S2B), and no direct association between the first six and the last eight residues appears to exist in the 102-residue monomeric subunits. While core residues are highly conserved in KaiB proteins from diverse cyanobacteria, the lengths and sequences of the C-terminal region vary widely (Fig. 2.S1). Clustered acidic residues represent a common feature, and Ishiura and coworkers previously demonstrated that glutamates and aspartates in the KaiB C-terminal tail play a critical role in maintaining rhythmicity (Iwase et al., 2005). However, without high-resolution structures of the binary KaiB-KaiC and/or the ternary KaiA-KaiB-KaiC complex, it has remained unclear how KaiB C-terminal residues influence clock period and amplitude. Thus, it is possible that the tail simply affects the dimer-tetramer equilibrium (Murakami et al., 2012), whereby only dimers interact with KaiC. Alternatively, the KaiB C-terminal region could somehow affect the course of KaiC dephosphorylation by interfering with the KaiC active site at subunit interfaces or serve to enhance the interaction between KaiB-KaiC complex and KaiA near the end of the daily clock cycle (Pattanayek et al., 2011). A superimposition of monomeric subunits from the S. elongatus KaiB crystal structure illustrates significant conformational variations of the C-terminal region (Fig. 2.S4).
Table 2.1 KaiB crystal structure statistics

<table>
<thead>
<tr>
<th>Data collection</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>C2</td>
</tr>
<tr>
<td>Cell dimensions $a, b, c$ [Å], $\beta$ [$^\circ$]</td>
<td>69.82, 116.12, 53.23, 98.3</td>
</tr>
<tr>
<td>Wavelength [Å]</td>
<td>0.98</td>
</tr>
<tr>
<td>Resolution [Å] (last shell)</td>
<td>30.00 – 2.60 (2.69 – 2.60)</td>
</tr>
<tr>
<td>Unique reflections (last shell)</td>
<td>12,670 (1,278)</td>
</tr>
<tr>
<td>Completeness [%] (last shell)</td>
<td>100.0 (99.9)</td>
</tr>
<tr>
<td>R-merge (last shell)</td>
<td>0.060 (0.267)</td>
</tr>
<tr>
<td>$I/\sigma(I)$ (last shell)</td>
<td>14.6 (2.1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Refinement</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Working set reflections</td>
<td>8,966 [$F \geq 3\sigma(F)$]</td>
</tr>
<tr>
<td>Test set reflections</td>
<td>423 (4.7%)</td>
</tr>
<tr>
<td>Protein atoms</td>
<td>3,173</td>
</tr>
<tr>
<td>Solvent atoms</td>
<td>99</td>
</tr>
<tr>
<td>R-work / R-free</td>
<td>0.25 / 0.30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Average B-factors [Å$^2$]</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein (all atoms)</td>
<td>78</td>
</tr>
<tr>
<td>Solvent</td>
<td>64</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R.m.s. deviations</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>bond lengths [Å]</td>
<td>0.003</td>
</tr>
<tr>
<td>bond angles [$^\circ$]</td>
<td>0.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ramachandran analysis [%]</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Most favored</td>
<td>81.9</td>
</tr>
<tr>
<td>Allowed</td>
<td>13.2</td>
</tr>
</tbody>
</table>

CryoEM structure of KaiB-KaiC complex formed with KaiC-Δ489

KaiB forms a complex with KaiC hexamers following autophosphorylation of KaiC at serine 431 (Murayama et al., 2011; Nishiwaki et al., 2007). During circadian oscillation, this site is transiently phosphorylated, and less than 30% of the KaiC hexamers are in complex with KaiB (Mori et al., 2007). To enrich for the
KaiB-KaiC (KaiBC) complex for cryoEM structure determination, we investigated complexes assembled with three mutant forms of KaiC. Two mutants, KaiC-S431E-T432E and KaiC-S431D, are phosphomimetic mutants while the third, KaiC-Δ489, is a truncation mutant lacking 30 amino acids from the C-terminus. This C-terminal truncation results in a mutant form of KaiC that is hyperphosphorylated (Kim et al., 2008) and forms a stable complex with KaiB (Qin et al., 2010). Micrographs of a negatively stained KaiBC sample formed with KaiC-S431E-T432E showed a high background level of a protein about the size of a KaiC subunit. This suggested that the KaiC-S431E-T432E hexamer might not be stable in the presence of KaiB. Native-PAGE analysis of this sample confirmed that approximately 10-30% of KaiC is in the monomeric form (Qin et al., 2010). Cryo-electron micrographs of the KaiBC sample formed with the second form of KaiC, KaiC-S431D, displayed a preferred orientation for the complex with the predominant view along the six-fold axis of the KaiC hexamer (top view). Cryo-electron micrographs of the KaiBC sample formed with the third form of KaiC, KaiC-Δ489, displayed the best properties for a cryoEM structural analysis, with reasonably uniform and well dispersed particles. Therefore we collected a cryoEM dataset of this third complex with a total of 195,226 particle images.

In order to select subsets of particle images with high KaiB occupancy an Eigenimage analysis was performed similar to that performed for GroEL by Clare et al. (Clare et al., 2012). This led to selection of side view particles and further sorting of the side view particle images with high KaiB occupancy (~60%) or low KaiB occupancy (~40%). Reference free image classification was performed on
both subsets followed by three-dimensional structure calculation without imposed symmetry using a 30 Å filtered representation of the KaiC crystal structure (aa 14-489) as the starting model. The high KaiB occupancy subset resulted in a class-average-based structure with a third layer of KaiB density, whereas the low KaiB occupancy subset resulted in a structure resembling KaiC without the third layer of density (Fig. 2.S5). Projections of the high KaiB occupancy class-average-based structure resemble the matching class averages of this subset (Fig. 2.2A). During refinement of the particle images in the high KaiB occupancy subset, it became apparent that the KaiBC structure displayed approximate 6-fold symmetry for both the KaiB and KaiC density layers, and therefore C6 symmetry was imposed for the remaining refinement rounds. The best structure generated from the high KaiB occupancy subset was used as a search model for one round of systematic parameter search against the full dataset (including top, side and tilted views). We reasoned that since 60% of the side view particle images sorted with the high KaiB occupancy subset, that probably ~60% of the full dataset also had high KaiB occupancy. Therefore we selected ~50% of the total dataset, those particles with the highest cross correlation values with the search model, to produce a final C6 symmetrized structure with a resolution of 16 Å at the FSC 0.5 threshold (13 Å at the FSC 0.143 threshold) (Fig. 2.2B).

Overall the cryoEM structure of KaiB-KaiC-Δ489 resembles the negative-stain EM KaiBC structure (Pattanayek et al., 2008) but with a more complete ring of density assigned to KaiB. Presumably the negative-stain structure represents an average of particle images with various levels of KaiB occupancy, and the
Eigenimage analysis of the cryoEM dataset served to select a more uniform subset of particle images. The previous negative-stain EM KaiBC structure formed with wild-type KaiC did not show the flexible C-terminal tails of KaiC, which protrude from the CII end of KaiC, probably because of a high degree of conformational flexibility. Both the negative-stain EM and cryoEM structures
display a double-doughnut shape recognizable as KaiC. The strong similarities between the two structures provide confidence that the truncation of KaiC did not adversely affect the binding location of KaiB or the overall conformation of the KaiBC complex. Both structures show a third layer of density assigned to KaiB, however the cryoEM structure shows KaiB density of the appropriate size for a KaiB monomer above each KaiC subunit (Fig. 2.2B).

Docking of KaiB and KaiC crystal structures into the KaiB-KaiC-Δ489 density

To gain insight into the structural basis for how KaiB might influence the phosphorylation status of KaiC as well as promote the release of KaiA from a KaiABC complex, we built molecular models for the KaiBC complex. Given the nearly symmetrical shape of the KaiC hexamer and the absence of helical density rods within the cryoEM structure, it was not possible to determine from the cryoEM structure alone whether KaiB binds to the N-terminal CI or C-terminal CII end of the hexamer. However, we noted that the strongest cryoEM density within KaiC was farthest from the third layer of KaiB density, which would be consistent with the α-helices of the CI domain at this position (Fig. 2.S6). We have previously presented indirect evidence that KaiB binds to the CII ring of KaiC (Pattanayek et al., 2008). A recent negative-stain EM study, using KaiB and gold-labeled full-length KaiC, indicates that KaiB binds KaiCII (Pattanayek et al., 2013). Nevertheless, given the difficulty in distinguishing the two ends of KaiB by cryoEM, we built models for the KaiBC complex with KaiB interacting with either the CI or the CII end of KaiC.
An atomic model for KaiC-Δ489 was made by computationally removing thirty C-terminal residues from the KaiC crystal structure (PDB: 3DVL, aa14-489). In the crystal structure, four copies of Ser-431 and all six copies of Thr-432 are phosphorylated. These phosphorylated residues were retained, and the remaining two copies of Ser-431 were computationally modified to the phosphorylated state with VMD 1.9 (Humphrey, Dalke, & Schulten, 1996). A fully phosphorylated form of KaiC was built because KaiB associates with phosphorylated KaiC more strongly than with non-phosphorylated KaiC (Nishiwaki et al., 2004; Yao Xu et al., 2004). In addition, a KaiC double phospho-mimic, S431D-T432D, has been shown to be more favorable for KaiBC complex formation than a KaiC mutant that mimics the unphosphorylated state, KaiC-S431A-T432A (Murakami et al., 2012).

Two molecular models for KaiC-Δ489 were refined into the cryoEM density map using cryoEM guided molecular dynamics based flexible fitting with MDFF and NAMD (Trabuco, Villa, Mitra, Frank, & Schulten, 2008; Trabuco, Villa, Schreiner, Harrison, & Schulten, 2009). Both possible orientations of KaiC were tested, with the KaiC CI ring next to KaiB and the KaiC CII ring next to KaiB. Similar potential energies were obtained after MDFF simulations with KaiC docked in either orientation in the absence of KaiB. During MDFF a standard potential energy function preserves correct stereochemistry, while application of an additional guiding force based on the cryoEM density map steers the coordinates into better agreement with the density. This hybrid cryoEM/MDFF approach has been applied to several macromolecular complexes (K.-Y. Chan,
Trabuco, Schreiner, & Schulten, 2012). In the case of a cryoEM structure of human adenovirus in complex with coagulation factor X, an MDFF analysis led to identification of key residues at the interface, and one of these residues was confirmed by mutagenesis to be critical for the interaction (Doronin et al., 2012).

For the KaiB monomer, we generated an atomic model lacking the C-terminal tail (cut after Y93) since the crystal structure of wild-type *S. elongatus* KaiB indicates that the tails are flexible (Fig. 2.S4). Also Murakami *et al.* have shown that truncation of the C-terminal tails of *T. elongatus* KaiB, which has 89% identity with *S. elongatus* KaiB, increases the binding affinity of KaiB for a hyperphosphorylated mimic of KaiC by a factor of 15 (Murakami et al., 2012). The KaiB monomer was manually docked into the third layer of cryoEM density with USCF Chimera (Pettersen et al., 2004). The visually best fitting model (Fig. 2.2B) positioned the two longest α-helices of KaiB against KaiC. Eight reasonably well fitting positions for KaiB were generated: four with the α-helices of KaiB against KaiC and rotated 90° with respect to each other, and four with the β-strand side of KaiB against KaiC and rotated 90° with respect to each other. All of the KaiB models were raised ~5 Å above the KaiC surface before the MDFF simulations.

All eight positions of KaiB were tested with the two orientations for KaiC, with the CI or CII side interacting with KaiB, in 100ps MDFF simulations. The nonbonded interaction energies between a KaiB monomer and the KaiC hexamer at the ends of the simulations are reported in Table 2.S1 for all of the models. Within the set of 8 models that have KaiB interacting with KaiCII there is
one most favorable orientation for KaiB (model 1), which has a nonbonded interaction energy of -365 kcal/mol and a buried accessible surface area between a KaiB monomer and a KaiC hexamer of 2,143 Å² (14% of the accessible surface of KaiB and KaiC; Table 2.S1). The second best model with KaiB interacting with KaiCII has a less favorable nonbonded interaction energy (-226 kcal/mol). Among the set of 8 models that have KaiB interacting with KaiCI there are three potentially favorable models (9, 11, and 13) based on the nonbonded interaction energy (-305 to -367 kcal/mol). However, visual examination of these models showed that KaiB interacted predominantly with a highly flexible loop on the CI surface of KaiC (aa112-121) and did not form a large interaction surface with KaiC. All of the models with KaiB interacting with KaiCI had buried accessible areas of less than 1,418 Å² (<9% of the accessible surface of KaiB and KaiC; Table 2.S1). Consideration of both the nonbonded interaction energy and the buried accessible surface area, led us to choose model 1 as the most likely KaiBC model consistent with both the cryoEM density and the EM gold labeling study showing that KaiB binds to the CII side of KaiC.³⁵

The robustness of the KaiBC interface indicated by model 1 was tested by rotating the KaiB monomer plus or minus 10° and re-running the MDFF simulations. In both cases the long α-helices of KaiB returned to essentially the same positions relative to the KaiCII surface and reformed the majority of the same residue-residue interactions with KaiC (Table 2.S2). The robustness of the protein interface provides support for the selection of model 1 as a reasonable model for the KaiBC interaction.
KaiB monomers form a continuous ring in the KaiBC complex

The selected MDFF refined model for KaiBC has a KaiB monomer interacting predominantly with one subunit of the KaiC hexamer and partially covering one ATP binding cleft on the KaiCII surface (Figs. 2.3 and 2.4). During the MDFF simulation the KaiB dimerization loop (aa45-60) and adjacent short α-helix (aa61-67) changed conformation to better fit the cryoEM density. In the process, the short α-helix moved into density corresponding to a neighboring

Figure 2.3 KaiBC atomic model docked within the cryoEM density
(A) The best MDFF-refined model within the KaiBC cryoEM density (transparent gray). A KaiC hexamer (gold and brown ribbons) is shown with one KaiB monomer (green ribbon). The KaiB dimerization loop and short α-helix (aa45-69) are shown in blue. The asterisks mark the positions of the N- and C-terminal tails of KaiB missing from the docked core model of KaiB (aa7-93). Twelve ATP molecules and 6 Mg^{2+} ions are shown in space filling representation and colored by element. (B) An MDFF-refined model with six KaiB monomers (green and cyan) forming a ring on the CII side of KaiC. The KaiB dimerization loops and short α-helices were removed to limit steric clashes before the MDFF simulation with all six KaiB monomers.
KaiB subunit (Fig. 2.3A). The cryoEM structure indicates that KaiB binds KaiC as a monomer. It seems reasonable that there would be a conformational change of the KaiB dimerization loop after binding of KaiB to KaiC; however this is difficult to model in the absence of a high-resolution structure of a KaiB monomer. Therefore, we removed the dimerization loop and short α-helix (aa45-69) of KaiB before building a KaiBC model with six KaiB monomers. An MDFF simulation with six KaiB monomers docked on the KaiC hexamer indicates that a model with six KaiB monomers bound to the CI side of KaiC is consistent with the cryoEM density (Fig. 2.3B). The simulation also indicates that neighboring KaiB monomers are closely packed and form a continuous ring around the top of KaiC, further blocking the ATP binding clefts on the KaiCII surface. In fact the modeled KaiB orientation places the N- and C-termini of each KaiB monomer in a position to interact with the neighboring KaiB monomer and potentially stabilize the KaiB ring (Fig. 2.3B).
Key interaction residues at the KaiBC interface

The MDFF simulations indicate that the top ten interaction residues on both KaiB and KaiC are mostly charged residues on the α-helices of KaiB and the CII surface of KaiC (Fig. 2.4). On KaiC the top interaction sites include residues on two adjacent KaiC subunits spanning an ATP binding cleft on the KaiCII surface. On KaiB the top interaction residues are in the two long α-helices or within 4 residues of these helices. The MDFF-refined coordinates for the KaiBC interaction indicate that the core of one KaiB monomer sits mostly on one KaiC subunit, and also overlaps with the adjacent clockwise KaiC subunit when

Figure 2.4 Interaction residues at the KaiBC interface
(A) Top views of the interacting KaiC and KaiB surfaces. The top ten most strongly interacting residues for both KaiC (K475-chainA, D464-chainA, E272-chainA, E444-chainF, E448-chainA, R269-chainA, R446-chainF, S473-chainA, E448-chainA, M471-chainA) and KaiB (R82, R22, K25, E34, R74, N29, N19, K84, T17, I30) as indicated by the MDFF simulation are shown in space filling representation. The ATP molecules are shown in ball-and-stick representation and Mg2+ ions are magenta spheres. KaiB and KaiC are colored as in Figure 2.3A. (B) Side views of interacting surfaces.
viewed from the KaiCII surface (Fig. 2.4A). Although the KaiB N- and C-termini were not included during the MDFF simulation, they are positioned in such a manner that they could interact with the adjacent counterclockwise KaiC subunit.

If the predicted binding interface on KaiC is correct, then mutations on the KaiCII surface should affect KaiB binding. We have previously mutated KaiC residue R468 into a Cys (KaiC-R468C) (Yao Xu et al., 2003). Native gel electrophoresis shows that KaiB binds to both wild-type KaiC and KaiC-R468C; however the affinity of KaiB for KaiC is slightly higher with the R468C mutant (Fig. 2.5). Bioluminescence rhythm assays of the circadian clock performed with a cyanobacterial strain lacking the endogenous kaiC gene and expressing wild-type KaiC or KaiC-R468C indicate a long period (~55 hrs) for KaiC-R468C (Xu et al, 2003). If the R468C mutation increases the stability of the KaiBC interaction such that KaiB forms a more stable ring on KaiC, this would be consistent with a longer period phenotype as it would presumably take longer for KaiB to dissociate. KaiC residue 468 is within the middle of the predicted KaiBC interface and between the two α-helices of KaiB (Fig. 2.5C). The MDFF based model of the KaiBC interaction presented here will be useful to guide for future mutagenesis experiments to probe the KaiBC interface.
A high-resolution structure of the KaiBC complex would be helpful in understanding the mechanistic underpinnings of the cyanobacterial circadian clock. In the absence of a crystal structure for the complex, we pursued a cryoEM structure of the complex and analyzed it with atomic structures for KaiB and KaiC and MDFF guided molecular dynamics. We used a truncated form of KaiC (KaiC-Δ489) in an attempt to make more uniform KaiBC complexes for a cryoEM structural study. However, the KaiB-KaiC-Δ489 complexes display compositional heterogeneity with variation in KaiB occupancy, and they may have conformational heterogeneity as well. Even the KaiC hexamer alone does not possess exact six-fold symmetry, as noted in the crystal structure (Pattanayek et al., 2004, 2006). The dynamic nature of KaiC subunit exchange, which occurs at a higher rate when KaiB is bound, could lead to even greater asymmetry and structural variability for the KaiBC complex. All of these factors may have contributed to the determination of a cryoEM structure at only moderate (16 Å) resolution, despite a relatively large dataset of ~195,000 particle images.
The cryoEM structure was analyzed by MDFF simulations with 16 different starting orientations for KaiB. One KaiBC model was selected on the basis of the calculated nonbonded interaction energy and the buried accessible surface area. In the selected model the two long α-helices of KaiB form the main interaction surface with KaiC. These helices are accessible in both the dimeric and tetrameric forms of KaiB (Fig. 2.S3). Therefore, while KaiB may initially contact KaiC as a dimer, the relatively strong predicted nonbonded interactions between the α-helices of both KaiB monomers with the CII surface of KaiC could lead to conversion of the dimer into two neighboring monomers bound to KaiC. Each monomer of KaiB would then form the same interactions with adjacent KaiC subunits. The MDFF simulations also indicate that a significant percentage (14%) of the total accessible surface area of KaiB and KaiC is buried when KaiB binds to the CII side of KaiC. This was not the case for the simulations that involved docking KaiB on the CI side of KaiC. Therefore, although KaiB may transiently bind to multiple sites on KaiC, this cryoEM and MDFF study supports the earlier EM gold labeling study of KaiB in complex with KaiC indicating that KaiB binds to KaiCII (Pattanayek et al., 2013).

The cryoEM and MDFF based model for the KaiBC complex offers an explanation for the finding that truncation of the C-terminal tails of *T. elongatus* KaiB increases the binding affinity of KaiB for KaiC (Murakami et al., 2012). In the selected KaiBC model, the N- and C-terminal tails of KaiB are in a position to interact with a neighboring KaiB monomer bound to KaiC. However, this also means that if fewer than six KaiB monomers are bound to KaiC the long C-
terminal tails of KaiB (aa94-102 for *S. elongatus*, aa95-108 for *T. elongatus*) could block access to the remaining KaiC subunits. The KaiBC model also provides a rationale for how KaiB promotes the dephosphorylation of KaiC. The model shows that one KaiB monomer binds to two KaiC subunits and covers one ATP binding cleft on KaiCII. A complete ring of six bound KaiB monomers would cover all of the ATP binding clefts on this side of KaiC. The dephosphorylation mechanism of KaiC has been shown to involve a reversal of the phosphorylation reaction which places the phosphates from KaiC back onto ADP (Egli et al., 2012; Nishiwaki & Kondo, 2012). Covering the ATP binding clefts would presumably block the entry of ADP and the release of ATP and stall the dephosphorylation reaction.

In summary, our goal was to gain more detailed structural information on the interaction of KaiB with KaiC. The cryoEM structure shows that KaiB forms a continuous ring of density on one side of KaiC. The MDFF analysis indicates that the ring of KaiB density is consistent with six monomers bound to the CII side of a KaiC hexamer. The cryoEM and MDFF results led us to propose an atomic model for the KaiBC complex with each of the six KaiB monomers covering an ATP binding cleft on the CII side of KaiC. The biological function of KaiB is to push KaiC toward a dephosphorylated state in the cyanobacterial posttranslational circadian oscillator by opposing the action of KaiA. The KaiBC model presented here suggests a possible way in which KaiB could influence the ATP binding pockets within the CII ring of KaiC.
Materials and Methods

Expression and purification of wild-type KaiB and KaiC-Δ489

The GST-KaiB and GST-KaiC fusion proteins were expressed in *E. coli* and purified by affinity chromatography followed by GST cleavage with PreScission protease (GE Healthcare) by gel filtration chromatography as described in (Nishiwaki et al., 2004) and (Mori et al., 2007). The KaiC-Δ489 and KaiC-R468C proteins were expressed as GST-fusion proteins in *E. coli* DH5α cells and purified by affinity chromatography with glutathione agarose beads (Thermo Scientific) following GST cleavage with recombinant GST-human rhinovirus 3C protease and ion exchange chromatography.

X-ray crystallography

Crystals for *S. elongatus* KaiB were grown from droplets containing 6.5 mg/mL KaiB-KaiC complex, 20 mM Tris (pH 7.8), 100 mM NaCl, 5 mM MgCl₂, 1 mM ATP and 2 mM BME. The reservoir solution was 100 mM sodium acetate, 500 mM sodium formate and 5% glycerol (v/v). Crystals grew within a week at room temperature using the hanging vapor diffusion technique. They were mounted in nylon loops, cryo-protected in mother liquor containing 25% glycerol (v/v) and then flash frozen in liquid nitrogen. X-ray diffraction data were collected on the 21-ID-D beamline of the Life Sciences Collaborative Access Team (LS-CAT) at the Advanced Photon Source, Argonne National Laboratory (Argonne, IL). Data were integrated and scaled with the program HKL2000 (Otwinowski & Minor, 1997). The structure was determined by the Molecular Replacement technique using the program MOLREP ("The CCP4 suite: programs for protein..."
crystallography.," 1994; Vagin & Teplyakov, 1997) and the *T. elongatus* KaiB structure with PDB: 2QKE (dimer) as the search model. Initial refinement was carried out with the program BUSTER-TNT (Blanc et al., 2004) and N- and C-terminal residues for the four subunits were gradually built into the electron density, followed by additional rounds of refinement with the program PHENIX (Adams et al., 2010). The program COOT (Emsley & Cowtan, 2004) was used for manual rebuilding. Water molecules were added gradually and isotropic/TLS refinement was continued with the program PHENIX with intermittent adjustments of side chain torsion angles. A summary of crystal data, data collection and refinement parameters is provided in Table 2.1. Illustrations were generated with the program UCSF Chimera (Pettersen et al., 2004).

Cryo-electron microscopy

KaiB-KaiC-Δ489 complexes were formed for electron microscopy studies as previously described for wild-type *S. elongatus* KaiB and KaiC (Mori et al., 2007). KaiB-KaiC-Δ489 samples were applied to freshly prepared EM grids with homemade holey carbon film and to C-flat grids (Protochips, Inc.). After blotting with filter paper the sample grid was plunged into ethane slush cooled by liquid nitrogen with either a homebuilt plunger or a Vitrobot (FEI Company). CryoEM images were collected on an FEI Polara microscope (300kV, FEG) operated at liquid nitrogen temperature and using a Gatan UltraScan 4000 CCD camera. The majority of the data (159,952 particles) was collected with an absolute magnification of 254,669X at 300kV. An additional smaller set of data (35,274
particles) was collected at 267,493X at 200kV. The defocus range for the entire dataset was -1 to -7 µm.

Image Processing

Individual particle images (195,226 total) were picked manually using in-house scripts that utilize EMAN subroutines (Ludtke, Baldwin, & Chiu, 1999), and the stacks were subsequently binned to a pixel size of 2.32 Å for the majority of the dataset (2.21 Å for the smaller subset). At the start of refinement, the initial estimates for microscope defocus and astigmatism parameters were calculated with CTFFIND3 (Mindell & Grigorieff, 2003). The images were filtered, normalized and band-pass filtered between 140 and 10 Å for MSA classification in IMAGIC (Van Heel, Harauz, Orlova, Schmidt, & Schatz, 1996). During the MSA-CLASSIFY step the worst 15% of the images were ignored. The majority of the dataset (159,952 particles) was classified into 2,000 class averages from which a subset of 360 class averages representing side views was selected. This subset was subjected to an Eigenimage analysis following a procedure applied to GroEL (Clare et al., 2012). First the selected side view class averages were rotationally and translationally aligned forming a reference set. Then the particle images that formed the side view class averages (21,937 particles) were extracted and rotationally and translationally aligned to the reference set. The aligned particle images underwent a round of MSA classification. One of the resulting eigenimages (Fig. 2.S7) described the presence or absence of a third layer of KaiB in the particle images. This eigenimage was used for sorting the 255kX side view particle images into two groups, one with high KaiB occupancy
(12,527 particles) and the other with low KaiB occupancy (9,410 particles). A similar Eigenimage analysis was performed on the data set collected with a magnification of 267kX, and this resulted in smaller subgroups of side view particle images with high KaiB occupancy (425 particles) and low KaiB occupancy (496 particles).

The larger 255kX subgroups underwent MSA classification resulting in 1,200 and 900 class averages for the high and low KaiB occupancy sets respectively. Each set of class averages was processed with the 3D Fourier Space programs developed by Rubinstein that incorporate the gold standard refinement scheme (www.sickkids.ca/research/rubinstein) (Benlekbir, Bueler, & Rubinstein, 2012). The initial starting model for both sets was a 30 Å filtered representation of the KaiC-Δ489 hexamer (aa 14-489). A third layer of KaiB density emerged in the structure for the set with high KaiB occupancy but not for the set with low KaiB occupancy (Fig. 2.S5). Although no symmetry was imposed, the third layer of KaiB density appeared approximately 6-fold symmetric in the high KaiB occupancy structure (Fig. 2.S5). The refined class-average-based structure with high KaiB occupancy had a gold standard resolution of 29 Å at the FSC 0.5 threshold (21 Å at the 0.143 threshold). This class-average-based structure was used as the starting model for a systematic parameter search of the 12,527 particles in the high KaiB occupancy subgroup. After five rounds of refinement the KaiB density still appeared approximately 6-fold symmetric (Fig. 2.S5). Therefore, the structure at the end of round five was rotated so that the 6-fold axis was along the z-axis, and a second systematic parameter search was
performed. The cross correlation (cc) values of the resulting parameter file were converted into Frealign style phase residuals (phase residual = 100 – cc value), and a C6 symmetrized map was calculated using FREALIGN (Grigorieff, 2007) in iflag=0 mode. Additional rounds of refinement were performed using the 3D Fourier Space programs and with C6 symmetric odd and even maps calculated with FREALIGN at the end of each round. This refined particle-image-based structure based on 11,415 particles had a gold standard resolution of 18 Å at the FSC 0.5 threshold (16 Å at the FSC 0.143 threshold).

The refined particle-image-based structure was used as the starting model for a systematic parameter search with all 159,952 particles collected with a magnification of 255kX and all 35,274 particles collected at 267kX. The voxel size of the search model was appropriately adjusted for the 267kX data set with PROC3D in EMAN (Ludtke et al., 1999). The selected particles from the two data sets with different magnifications were combined into one map using FREALIGN with the relmag parameter (relative magnification correction factor). Three different relmag values were tried, 0.99, 1.0 and 1.01. The best FSC curve was found with a relmag value of 1.0. The final map shown in Fig. 2.2 is C6 symmetrized, based on 98,932 particles (51% of 195,226 total particles), sharpened with a B-factor of -1000 Å² and filtered with the Cref filter (Rosenthal & Henderson, 2003). The resolution of the final map is 16 Å at the FSC 0.5 threshold and 13 Å at the FSC 0.143 threshold. The lowest spatial frequency used during 3D Fourier Space refinement was 140 Å. The highest spatial frequency was started at 30 Å in the initial round and maintained at two
resolution shells below the resolution limit of the map, as assessed by Fourier shell correlation with the 0.143 criterion. The cryoEM structure has been deposited in the EM Data Bank EMD: 5672.

MDFF Simulations

In preparation for MDFF simulations an atomic model was generated for a C-terminally truncated form of *S. elongatus* KaiC-Δ489 starting with the crystal structure of the KaiC hexamer (PDB: 3DVL) and removing coordinates for the C-terminal tails (aa490-519). Two copies of Ser-431 were computationally modified to the phosphorylated state with VMD 1.9 (Humphrey et al., 1996) so that all Ser-431 and all Thr-432 residues were phosphorylated for the MDFF simulations. All twelve ATP molecules were retained in the structure for the simulations. An atomic model was generated for an *S. elongatus* KaiB monomer from the X-ray crystallographic coordinates. Given the observed flexibility for the *S. elongatus* KaiB C-terminal tails, the MDFF simulations were performed with a computationally C-terminally truncated form the KaiB monomer containing residues 7-93 from chain A of the crystal structure.

The Molecular Dynamics Flexible Fitting (MDFF) plugin for NAMD 2.8 and VMD 1.9 (Humphrey et al., 1996; Trabuco et al., 2008, 2009) was used to perform molecular dynamics simulations guided by the cryoEM density map. The MDFF simulations were performed with implicit solvent, the CHARMM force field, and a gscale factor of 0.3, which describes the strength of the external potential derived from the EM density map. The KaiC-Δ489 hexamer coordinates were docked into the cryoEM density in two different orientations, with the CI or CII
ring toward KaiB, with UCSF Chimera (Pettersen et al., 2004). MDFF simulations (100ps each) were performed for both KaiC orientations, first in the absence of KaiB. Ten residues at the C-terminal end of KaiC moved into the KaiB density ring when KaiC was oriented with the CII ring toward KaiB, so these residues (aa480-489) were computationally removed before the MDFF simulations with KaiB.

Eight different orientations of the KaiB monomer, four with the KaiB helices oriented toward KaiC and four with the KaiB β-stands oriented toward KaiC, were tested with each KaiC orientation. The four helix-down KaiB orientations and the four strand-down KaiB orientations were related by 90°. Visual assessment with UCSF Chimera (Pettersen et al., 2004) of the KaiBC models and the cryoEM density indicated that the helix-down orientations more closely resembled the cryoEM density than the strand-down orientations. In addition, models with other KaiB monomer surfaces docked next to KaiC did not fit the cryoEM density as well. The initial KaiBC coordinates were docked into the cryoEM density with UCSF Chimera (Pettersen et al., 2004) leaving a small ~5 Å gap between KaiB and KaiC. 100ps MDFF simulations were performed followed by 2,000 steps of minimization. After the selection of one KaiBC model, two additional 100ps MDFF simulations were performed with KaiB rotated by + or -10 degrees from its selected binding orientation and raised ~5Å above the KaiC surface so that the protein interface would be reformed during the simulation. Finally, one MDFF simulation was performed with 6 docked KaiB monomers in the selected KaiB binding orientation, with KaiB raised ~5Å above the KaiC
surface, and with the KaiB dimerization loop and short helix (aa45-69) computationally removed to avoid steric clashes with neighboring KaiB monomers. The intermolecular nonbonded interaction energies between KaiC and KaiB subunits were evaluated with the NAMD Energy plugin in VMD (Humphrey et al., 1996) and the buried accessible surface area was measured with UCSF Chimera (Pettersen et al., 2004). The MDFF simulations were performed on the Case Western Reserve University High Performance Computing Cluster.

KaiC R468C mutant binding assays

The KaiC-R468C mutant was first described by (Yao Xu et al., 2003). The binding reactions between KaiB and KaiC proteins were carried out at 3.4 μM of the wild-type or KaiC-R468C with various concentrations (0-10 μM) of the wild-type KaiB in 20 mM Tris-HCl, pH 8, 150 mM NaCl, 5 mM MgCl₂, 1 mM ATP and 0.5 mM EDTA at 30°C for 4 hours. Native polyacrylamide gel electrophoresis was performed as previously described (Mori et al., 2002).

Accession numbers

Final coordinates and structure factors for the KaiB crystal structure have been deposited in the Protein Data Bank with accession number 4KSO.

Acknowledgements

We would like to thank Mr. Said K. Sidiqi for help with protein expression and purification, Dr. Yao Xu for providing the expression construct for KaiC R468C, and Dr. Zdzislaw Wawrzak, Northwestern University, for assistance with
X-ray diffraction data collection and processing. Use of the Advanced Photon Source was supported by the U. S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. Use of the LS-CAT Sector 21 was supported by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor for the support of this research program (Grant 085P1000817). This work was supported by the National Institute of Health grants GM067152 (to CHJ), GM073845 (to ME), and GM081646 (to PLS).
Supplemental Information

CryoEM and Molecular Dynamics of the Circadian KaiB–KaiC Complex Indicates KaiB Monomers Interact with KaiC and Block ATP Binding Clefts

Seth A. Villarreal,1 Rekha Pattanayek,2 Dewight R. Williams,3,5 Tetsuya Mori,4 Ximing Qin,4 Carl H. Johnson,4 Martin Egli,2 and Phoebe L. Stewart1*

1Department of Pharmacology and Cleveland Center for Membrane and Structural Biology, Case Western Reserve University, Cleveland, OH 44106, USA
2Department of Biochemistry, Vanderbilt University, School of Medicine, Nashville, TN 37232, USA
3Department of Molecular Physiology and Biophysics, Vanderbilt University, School of Medicine, Nashville, TN 37232, USA
4Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA
5Present address: Department of Pathology and Laboratory Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA
Figure 2.S1 Sequence alignment for KaiB proteins from selected cyanobacteria

1. KaiB *Synechococcus elongatus* PCC 7942, gi|22002534|; 2. KaiB *Thermosynechococcus elongatus* BP-1, gi|22294203|; 3. KaiB *Synechocystis* sp. PCC 6803, gi|16332221|; 4. KaiB *Anabaena variabilis* ATCC 29413, gi|75700965|; 5. KaiB *Prochlorococcus marinus* str. MIT 9313, gi|33863686|. The alignment was performed with the program MULTALIN (Combret, Blanchet, Geourjon, & Deléage, 2000; Corpet, 1988). Residues conserved for 90% or more (upper-case letters): 76 (63.33%). Residues conserved for 50% and less than 90% (lower-case letters): 13 (10.83%). Residues conserved less than 50% (white space): 25 (20.83%). IV conserved positions (!): 3 (2.50%). LM conserved positions ($): 0. FY conserved positions (%): 0. NDQEBZ conserved positions (#): 3 (2.50%).

Numbering of residues in the present contribution refers to the KaiB protein from *T. elongatus* (M1 to E108), i.e. *T. elongatus* Y94 corresponds to *S. elongatus* Y93, Y8 corresponds to Y7, etc.
Figure 2.52 Conformation of the KaiB tetramer
The KaiB tetramer viewed (A) roughly along the non-crystallographic dyad relating the two dimers, and (B) from the side, after rotating around the horizontal by 90°. Subunits are colored blue (a), yellow (b), green (c) and cyan (d), and ribbons at the N- and C-terminal ends are colored in black and red, respectively. The drawing was generated with the program UCSF Chimera (Pettersen et al., 2004).
Figure 2.S3 Interactions stabilizing the KaiB dimer-dimer interface
Tyrosines 7 and 93 near the N- and C-terminal ends, respectively (side chains highlighted in red), engage in an edge-to-face stacking interaction that ties together strands of the antiparallel β-sheet in the α/β sandwich fold adopted by KaiB monomeric subunits. The Tyr pairs help stabilize the dimer-dimer interface by forming mostly hydrophobic interactions with E34 and F35 as well as L59 and T63 (side chains highlighted in black; a/d and b/c subunit interfaces). The Y7/93A KaiB double mutant with additional truncations at the N- and C-terminal ends (residues <6 and >93, respectively) exists predominantly in the dimeric form (Chang et al., 2012). The drawing was generated with UCSF Chimera (Pettersen et al., 2004).
Figure 2.S4 Flexibility of N- and C-terminal tails in KaiB monomeric subunits
Superimposition of the a (blue), b (yellow), c (green) and d (cyan) subunits (residues Y8-Y94) viewed (A) across α-helices and the β-sheet and (B) rotated around the vertical by 90°, illustrating the conformational variations in the N- and C-terminal regions. Side chain carbons of N-terminal Met and C-terminal Phe residues are highlighted in black and red, respectively (please note that the first six residues of subunit “a” were not visible in the electron density). The drawing was generated with UCSF Chimera (Pettersen et al., 2004).
Figure 2.55 Schematic outline of the image processing procedure followed to generate the KaiB-KaiC-Δ489 cryoEM structure. The Eigenimage analysis was performed following a procedure applied to GroEL (Clare et al., 2012). See Materials and Methods section for additional details.
Figure 2.S6 Strongest density within the KaiB-KaiC-Δ489 cryoEM structure aligns with α-helices in KaiCI
(A) CryoEM structure displayed with two isosurface levels: contoured for 100% of the expected volume for six KaiB monomers and a KaiC hexamer (transparent gray) and contoured to show just the strongest density (solid blue). (B) CryoEM structure (transparent gray) with the atomic model for KaiC-Δ489 (based on PDB ID 3DVL) after a 100ps MDFF simulation docked with the CII end of KaiC next to KaiB. Alpha-helices are in cyan and ATP molecules are in space filling representation. Note that the strongest density at the base of the KaiB-KaiC-Δ489 structure (A) is consistent with α-helices at the bottom of the CI domain in this KaiC orientation.

Figure 2.S7 Eigenimages calculated from the aligned side view particles images
The first five eigenimages are shown for the dataset collected with a magnification of 255kX. The circled eigenimage was used for sorting the side view particles images into two groups, one with high KaiB occupancy and the other with low KaiB occupancy. This eigenimage predominantly indicates the presence of the KaiB density layer as well as some orientational variation.
Table 2.S1 Comparison of 16 KaiBC models

<table>
<thead>
<tr>
<th>Model</th>
<th>KaiC interaction side</th>
<th>Nonbonded interaction energy between a KaiB monomer and KaiC hexamer (kcal/mol)(^a)</th>
<th>Buried accessible surface area between a KaiB monomer and KaiC hexamer (Å(^2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(^b)</td>
<td>CII</td>
<td>-365</td>
<td>2,143</td>
</tr>
<tr>
<td>2</td>
<td>CII</td>
<td>-199</td>
<td>2,055</td>
</tr>
<tr>
<td>3</td>
<td>CII</td>
<td>-226</td>
<td>1,968</td>
</tr>
<tr>
<td>4</td>
<td>CII</td>
<td>-90</td>
<td>1,653</td>
</tr>
<tr>
<td>5</td>
<td>CII</td>
<td>-161</td>
<td>1,626</td>
</tr>
<tr>
<td>6</td>
<td>CII</td>
<td>-87</td>
<td>1,697</td>
</tr>
<tr>
<td>7</td>
<td>CII</td>
<td>-116</td>
<td>1,841</td>
</tr>
<tr>
<td>8</td>
<td>CII</td>
<td>-74</td>
<td>1,613</td>
</tr>
<tr>
<td>9</td>
<td>CI</td>
<td>-367</td>
<td>1,331</td>
</tr>
<tr>
<td>10</td>
<td>CI</td>
<td>-227</td>
<td>1,170</td>
</tr>
<tr>
<td>11</td>
<td>CI</td>
<td>-343</td>
<td>1,418</td>
</tr>
<tr>
<td>12</td>
<td>CI</td>
<td>-121</td>
<td>1,110</td>
</tr>
<tr>
<td>13</td>
<td>CI</td>
<td>-305</td>
<td>1,219</td>
</tr>
<tr>
<td>14</td>
<td>CI</td>
<td>-143</td>
<td>982</td>
</tr>
<tr>
<td>15</td>
<td>CI</td>
<td>-174</td>
<td>1,029</td>
</tr>
<tr>
<td>16</td>
<td>CI</td>
<td>-49</td>
<td>1,170</td>
</tr>
</tbody>
</table>

\(^a\)Note that negative values for nonbonded interaction energies are favorable

\(^b\)Selected KaiBC model
Table 2.S2. Robustness of KaiBC interface after molecular dynamics flexible fitting runs with different starting KaiB orientations

The top ten interaction residues at the end of the MDFF simulation are listed in order of their contribution to the total nonbonded interaction energy between KaiB and KaiC (highest contribution on the left). Interaction residues in common between model 1 and the models with KaiB rotated +/- 10° are highlighted in yellow. For the KaiC interaction residues, the chain is indicated after the residue number.

A. KaiB interaction residues

<table>
<thead>
<tr>
<th>Model 1</th>
<th>R82</th>
<th>R22</th>
<th>K25</th>
<th>E34</th>
<th>R74</th>
<th>N29</th>
<th>N19</th>
<th>K84</th>
<th>T17</th>
<th>I30</th>
</tr>
</thead>
<tbody>
<tr>
<td>KaiB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rotated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+10°</td>
<td>R22</td>
<td>K66</td>
<td>R82</td>
<td>E32</td>
<td>E34</td>
<td>E83</td>
<td>K25</td>
<td>N29</td>
<td>R74</td>
<td>I30</td>
</tr>
<tr>
<td>rotated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-10°</td>
<td>R22</td>
<td>R74</td>
<td>R82</td>
<td>K84</td>
<td>K25</td>
<td>E34</td>
<td>N29</td>
<td>T26</td>
<td>D45</td>
<td>N19</td>
</tr>
</tbody>
</table>

B. KaiC interaction residues

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>KaiB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rotated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rotated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 3: CryoEM and Image Sorting for Flexible Protein/DNA Complexes

This chapter was published in the Journal of Structural Biology in 2014

Abstract

Intrinsically disordered regions of proteins and conformational flexibility within complexes can be critical for biological function. However, disorder, flexibility, and heterogeneity often hinder structural analyses. CryoEM and single particle image processing techniques offer the possibility of imaging samples with significant flexibility. Division of particle images into more homogenous subsets after data acquisition can help compensate for heterogeneity within the sample. We present the utility of an eigenimage sorting analysis for examining two protein/DNA complexes with significant conformational flexibility and heterogeneity. These complexes are integral to the non-homologous end joining pathway, and are involved in the repair of double strand breaks of DNA. Both complexes include the DNA-dependent protein kinase catalytic subunit (DNA-PKcs) and biotinylated DNA with bound streptavidin, with one complex containing the Ku heterodimer. Initial 3D reconstructions of the two DNA-PKcs complexes resembled a cryoEM structure of uncomplexed DNA-PKcs without additional density clearly attributable to the remaining components. Application of eigenimage sorting allowed division of the DNA-PKcs complex datasets into more homogeneous subsets. This led to visualization of density near the base of the DNA-PKcs that can be attributed to DNA, streptavidin, and Ku. However, comparison of projections of the subset structures with 2D class averages indicated that a significant level of heterogeneity remained within each subset. In summary, image sorting methods allowed visualization of extra density near the
base of DNA-PKcs, suggesting that DNA binds in the vicinity of the base of the molecule and potentially to a flexible region of DNA-PKcs.

Keywords

Cryo-electron microscopy; cryoEM; single particle reconstruction; flexibility; heterogeneity; eigenimage sorting; protein/DNA complexes

Introduction

There is growing recognition of the importance of structural flexibility and intrinsic disorder in proteins for performing essential biological functions (Babu, Kriwacki, & Pappu, 2012; Mittag, Kay, & Forman-Kay, 2010; Uversky, 2011). Conformational flexibility is an issue that experimental structural biology is beginning to address. Often flexible regions are absent from crystal structures. NMR methods are being developed to quantify intrinsically disordered proteins and provide ensemble descriptions of flexible regions (Jensen, Ruigrok, & Blackledge, 2013). In general, cryoEM single particle image reconstruction requires a significant level of structural homogeneity in the dataset. Small intrinsically disordered regions within a symmetrical assembly, such as an icosahedral virus, can be modeled with molecular dynamics simulations (Flatt, Kim, Smith, Nemerow, & Stewart, 2013). Inherently flexible complexes such as the ribosome, which stochastically fluctuates among multiple conformations, can be induced into intermediate states, which can be studied by cryoEM and x-ray crystallography (Frank, 2012). Various software approaches are being developed for dealing with heterogeneity within cryoEM datasets. These include a bootstrapping approach (Penczek, Frank, & Spahn, 2006), a Bayesian statistics
approach (Scheres, 2012), and several methods based on Multivariate Statistical Analysis (MSA) (Orlova & Saibil, 2010; Van Heel et al., 2000).

CryoEM single particle studies of macromolecular complexes can be complicated by flexibility within the components, heterogeneity in the composition of the complex, and low signal-to-noise ratios. The first two factors lead to a loss of density for flexible regions through averaging and the last factor can result in overfitting and refinement on noise artifacts. The so-called gold standard refinement approach provides a degree of protection from overfitting and the subsequent misinterpretation of the resulting structure (Scheres & Chen, 2012). This refinement approach involves processing half datasets separately (e.g. odd and even particle images), calculating separate maps for the odd and even particle images, and using these half maps for additional rounds of refinement while keeping the odd and even datasets independent (odd map with odd particle images, and even map with even particle images). The resolution is still estimated by comparison of the odd and even maps as normal, but the odd and even maps are not influenced by each other during refinement. In this manner, random features based solely on noise within either half map should not be reinforced in the merged full map, overfitting is limited, and more realistic estimates are obtained for the resolution.

Components within the non-homologous end joining (NHEJ) pathway have to be adaptable enough to act on a wide variety of deleterious DNA end substrate configurations. In addition to repairing deleterious breaks, this pathway is also critical for repairing programmed double-strand DNA breaks necessary for
generation of diversity in the adaptive immune system. The DNA-dependent protein kinase catalytic subunit (DNA-PKcs) has a central regulatory role in NHEJ repair and is a large (469 kDa) serine/threonine protein kinase belonging to the phosphatidylinositol-3-OH kinase (PI3K)-related protein family. DNA-PKcs functions together with Ku, which is a heterodimer of Ku70 and Ku86. When a double-strand break occurs, Ku recognizes and binds broken DNA ends. Next DNA-PKcs is recruited to the break site by the Ku-DNA complex via interaction with the C-terminal region of Ku86. The interaction between DNA-PKcs and DNA leads to autophosphorylation of DNA-PKcs, which is presumed to regulate the access of the DNA end to other NHEJ components. DNA-PKcs has predicted disordered regions, and we speculate that NHEJ complexes may use intrinsically disordered regions to enable recognition of a diverse array of damaged ends during the repair process. In addition, conformational flexibility may facilitate recruitment of other factors in the NHEJ pathway, including enzymes with nuclease, polymerase and ligase activities (Lieber, 2010).

Numerous structural studies have provided only moderate resolution structures of DNA-PKcs (Chiu, Cary, Chen, Peterson, & Stewart, 1998; Leuther, Hammarsten, Kornberg, & Chu, 1999; Rivera-Calzada et al., 2005; Sibanda et al., 2010; D. R. Williams et al., 2008). A crystal structure of human DNA-PKcs at 6.6 Å resolution revealed helices throughout the molecule but did not permit fitting of the primary sequence into the density (Sibanda et al., 2010). Coordinates were only determined for 46% of the backbone residues and these coordinates were not assigned to specific residues within the sequence. Loops
between secondary structural elements are missing. In addition, since specific residues were not assigned there is a possibility that large regions of protein are missing from the crystal structure due to flexibility. Sibanda et al. proposed that the kinase domain is localized within the head region after showing a good superposition for a region in the head of the 6.6 Å resolution crystal structure with the atomic structure of the homologous PI3K-γ kinase (E. H. Walker, Perisic, Ried, Stephens, & Williams, 1999). Crystallization of DNA-PKcs was performed in complex with a C-terminal fragment of Ku86 (Ku80ct194, aa 539-732), but it was not possible to identify the Ku80ct194 fragment within the overall structure. EM studies have revealed the head and the base of the molecule (Chiu et al., 1998; Leuther et al., 1999; Rivera-Calzada et al., 2005; D. R. Williams et al., 2008), as well as a large central channel with a helical protrusion recessed from the opening of the channel that has been proposed as a possible DNA binding site (D. R. Williams et al., 2008).

The EM and x-ray structures of DNA-PKcs differ mainly in the size of the base of the molecule (Supplemental Fig. 3.S1), which is composed in part by conformationally flexible HEAT repeats (Sibanda et al., 2010). The base of the EM structure appears substantially larger than that of the x-ray structure. Although the resolution of the crystal structure (6.6 Å) is similar to the nominal resolution of the EM structure (D. R. Williams et al., 2008) (7 Å), the EM resolution was estimated without using the gold standard refinement approach and therefore may be an overestimate. The differences between the EM and x-ray structures might be explained if there is conformational variability in the base
of DNA-PKcs that is of sufficient magnitude to smear the crystallographic density and not the EM density at more moderate resolution. Alternatively, the two structures might represent different conformations, with the cryoEM structure representing the more prevalent conformation in solvent and without crystal packing forces.

Crystal structures of DNA-free and DNA-bound forms of Ku have been determined using full-length Ku70 and a truncated form of Ku86 (aa1-565), missing 167 C-terminal residues (J. R. Walker et al., 2001). The DNA-binding core of Ku is formed by Ku70 and the majority of Ku86, excluding the extended C-terminal portion of Ku86. Double-stranded DNA binds within a preformed ring of the Ku DNA-binding core. NMR structures have been determined for portions of the C-terminal region of Ku86, including 592-709aa (Harris et al., 2004) and 566-710aa (Z. Zhang et al., 2004). Neither of the NMR studies revealed the structure of the final ~20 residues at the C-terminus of Ku86, which are known to mediate a specific interaction with DNA-PKcs (Falck, Coates, & Jackson, 2005; Gell & Jackson, 1999). A small angle x-ray scattering (SAXS) analysis of full-length Ku with and without DNA indicates that the Ku86 C-terminal region forms a flexible arm that extends up to ~100 Å away from the DNA-binding core (Hammel et al., 2010).

Here we present the results of an eigenimage sorting analysis applied to cryoEM datasets of two NHEJ complexes including DNA-PKcs, double-stranded DNA, and Ku. Our original goal was to localize the binding site on DNA-PKcs for double-stranded DNA within these complexes. The samples were formed with
DNA biotinylated at one end and complexed with streptavidin, so as to block multiple DNA-PKcs and Ku proteins from associating with a single DNA molecule. Considerable structural flexibility within DNA-PKcs, as well as heterogeneity within the NHEJ complexes, hampered our efforts to localize the DNA binding site. However, eigenimage sorting allowed us to constrain the region of DNA-PKcs where DNA and Ku are likely to bind.

Results

Initial Structures of DNA-PKcs/DNA and DNA-PKcs/Ku/DNA

DNA-PKcs/DNA and DNA-PKcs/Ku/DNA complexes were formed with DNA biotinylated at one end, allowing one streptavidin tetramer to bind, and limiting access of DNA-PKcs to only one end of the DNA. The length of the double-stranded DNA molecules differed in the two complexes with 18bp in the DNA-PKcs/DNA complex and a longer 35bp DNA molecule in the DNA-PKcs/Ku/DNA complex to allow for binding both Ku and DNA-PKcs. CryoEM datasets were collected for DNA-PKcs/DNA and DNA-PKcs/Ku/DNA (35,176 and 19,825 particle images, respectively). Image processing of the DNA-PKcs/DNA and DNA-PKcs/Ku/DNA datasets was performed with the 3D Fourier Space software package, (Benlekbir et al., 2012) which incorporates the so-called gold standard refinement process. This resulted in structures for the two complexes that resemble uncomplexed DNA-PKcs (Fig. 3.1). The resolution of these
structures measured 15.0 Å and 21.4 Å, respectively, using the Fourier Shell Correlation (FSC) 0.5 threshold (Supplemental Fig. 3.S2).

For comparison, we re-processed the cryoEM dataset of DNA-PKcs with 3D Fourier Space using the portion of the dataset collected with a single magnification (284,173 particles). With gold standard refinement of this partial dataset, we calculated a DNA-PKcs structure at 13.6 Å resolution by the FSC 0.5 threshold (Fig. 3.1A, Supplemental Fig. 3.S2). All three structures showed similar

Figure 3.1 3D Fourier Space refined structures of DNA-PKcs and two DNA-PKcs complexes
Structures of (A) DNA-PKcs, (B) DNA-PKcs/DNA with streptavidin, and (C) DNA-PKcs/Ku/DNA with streptavidin are shown sharpened with a B-factor of -1,000 Å², filtered with the Cref filter (Rosenthal & Henderson, 2003) and with an additional low-pass filter at their respective FSC 0.5 thresholds (13.6 Å, 15.0 Å, and 21.4 Å). Structures are contoured at 100% of the expected volume for DNA-PKcs. Scale bar represents 50 Å.
overall features, albeit at different resolutions. No clear density for DNA, streptavidin or Ku was observed in the structures of the complexes. There appeared to be no large conformational changes in DNA-PKcs when in complex with DNA or Ku/DNA; however, we cannot rule out small conformational changes. Both complex structures included >99% of the particle images, and it is possible that they represent heterogeneous populations with multiple conformations or partial DNA or Ku occupancy. The homogeneity of the complex samples was assessed by negative-stain EM (Supplemental Fig. 3.S3). Visual inspection of particle images and class averages led to an estimate of >30% of each population forming the desired complex.

During refinement of DNA-PKcs/DNA and DNA-PKcs/Ku/DNA, we noted that the resolutions indicated by the FSC curves were influenced by the density mask used. 3D Fourier Space allows the calculation of a custom mask for each complex with a user-specified density threshold. For consistency, we calculated FSC curves for all three structures with uniform outer radial masks of 80 Å.

Intrinsically disordered regions within DNA-PKcs

To scan for potential regions of disorder within DNA-PKcs, the human DNA-PKcs sequence (4,128 aa) was submitted to several intrinsic protein disorder web servers, GLOB Plot 2, IUPred, and DisEMBL (Dosztányi, Csizmok, Tompa, & Simon, 2005; Linding, Russell, Neduva, & Gibson, 2003; Linding, Jensen, et al., 2003). The results indicate multiple regions with predicted disorder throughout the sequence and the percentage of predicted disordered residues ranged from 4% to 39% (Supplemental Table 3.S1). Flexibility, particularly in the
base of DNA-PKcs, may help to explain the observed differences between the cryoEM and x-ray structures of DNA-PKcs (Fig. 3.1). Disordered regions or conformational flexibility also offers a possible explanation for why the initial structures of DNA-PKcs/DNA and DNA-PKcs/Ku/DNA did not show density for Ku or DNA (Fig. 3.1). If DNA and Ku bind to an intrinsically disordered region of DNA-PKcs, it would be difficult to reconstruct density for these components in the complexes.

Eigenimage sorting of the DNA-PKcs/DNA dataset

In order to help overcome sample heterogeneity and select for a more homogeneous subset of DNA-PKcs/DNA particle images, an eigenimage analysis was performed similar to that performed for GroEL (Clare et al., 2012). Three preliminary rounds of sorting were used to remove particle images displaying close neighboring DNA-PKcs molecules. In a fourth round of sorting the first eigenimage for the DNA-PKcs/DNA complex displayed the predominant features of DNA-PKcs in side view (Fig. 3.2A). The second and third eigenimages indicated either variations in the DNA-PKcs orientation within the dataset or possibly conformational changes in DNA-PKcs upon DNA binding. The fourth eigenimage displayed a strong feature beyond the base of DNA-PKcs, which we attribute to biotinylated DNA and streptavidin. Selection for images that agreed with eigenimage 4 resulted in a subset with 6,222 particle images.
Figure 3.2 Eigenimage sorting of DNA-PKcs/DNA with streptavidin cryoEM particle images

(A) The first four eigenimages of the DNA-PKcs/DNA dataset are shown with the first eigenimage representing the predominant side view of DNA-PKcs. The eigenimage marked with an asterisk represents extra density at the base of DNA-PKcs and was used for sorting the particles images into a more homogeneous subset. (B) Subset structure of DNA-PKcs/DNA (transparent gray) shown at ~140% of the expected volume of the complex and with the cryoEM structure of DNA-PKcs (Fig. 3.1A, magenta) docked inside. (C) Selected class averages (top) and matching projections of the subset structure (bottom) with the weak extra density at the base computationally strengthened prior to projection. The density at the base of DNA-PKcs is indicated by an arrow for the first class average. (D) Model of the complex with 18bp DNA (red) protruding into the base of DNA-PKcs (transparent gray) and simulated density for a streptavidin tetramer at the distal end of DNA (cyan). Scale bars represent 50 Å.
This subset was processed with reference-free image classification followed by 3D Fourier Space refinement of the class average images. This resulted in a DNA-PKcs/DNA structure that displays a weak density lobe below the base of DNA-PKcs (Fig. 3.2B). Comparison of class averages and matching projections show that the average position for streptavidin observed in the subset structures matches the two-dimensional information in the class averages (Fig. 3.2C). We observed that the density values within the weak density lobes are approximately 1/10\(^{th}\) that of the values within DNA-PKcs. We interpret this to mean that there is still significant heterogeneity in the position of the extra density at the base within the subset after the eigenimage sorting procedure. If we assume that the extra density represents a streptavidin tetramer (53 kDa) bound at the distal end of DNA, there must be considerable variability in the position of both DNA and streptavidin relative to DNA-PKcs. A model of the complex (Fig. 3.2D) indicates that if streptavidin were rigidly positioned with respect to DNA-PKcs and reconstructed as well as DNA-PKcs it should appear larger than the observed extra density at the base.

A control dataset of DNA-PKcs/DNA without streptavidin (17,323 particle images) was also collected. A similar eigenimage sorting analysis indicated no significant density at the base of DNA-PKcs (Fig. 3.3). This result supports the idea that the weak density observed below the base of DNA-PKcs in the DNA-PKcs/DNA-streptavidin structure corresponds to streptavidin (53 kDa) rather than the 18bp dsDNA in the complex (~12 kDa).
One further control was performed on a subset of ~35,000 cryoEM particle images of uncomplexed DNA-PKcs. Following a similar eigenimage sorting analysis with this DNA-PKcs dataset resulted in eigenimages similar to the first three eigenimages of DNA-PKcs/DNA-streptavidin (Fig. 3.2A) and all of the eigenimages of DNA-PKcs/DNA without streptavidin (Fig. 3.3A). The first eigenimage of uncomplexed DNA-PKcs displayed the predominant features of the molecule and the following eigenimages showed either variation in the DNA-PKcs orientation within the dataset or possibly conformational changes in DNA-PKcs (Supplemental Fig. 3.S4). None of the uncomplexed DNA-PKcs eigenimages indicated density beyond the base of molecule. Attempts to sort the DNA-PKcs particle images based on the first six eigenimages resulted in highly overlapping subsets of particles rather than discrete subsets. This suggests that

Figure 3.3 Eigenimage sorting of DNA-PKcs/DNA without streptavidin cryoEM particle images
(A) The first six eigenimages of the DNA-PKcs/DNA without streptavidin dataset are shown with the first eigenimage representing the predominant side view of DNA-PKcs. None of the eigenimages represent extra density at the base of DNA-PKcs. (B) Representative class averages. Scale bar represents 50 Å.
the majority of the variation observed in these eigenimages corresponds to variation in orientation.

Eigenimage sorting of the DNA-PKcs/Ku/DNA dataset

The DNA-PKcs/Ku/DNA dataset also underwent an eigenimage sorting procedure. This resulted in four eigenimages that reflected additional density at the base of DNA-PKcs (Fig. 3.4A). Sorting produced four subsets of particle images ranging from 4,573 to 5,235 particles and 3D structures generated from the subsets showed extra density lobes below the base of DNA-PKcs (Fig. 3.4B). These lobes are further away from DNA-PKcs and spread over greater distances (160 Å in x, 90 Å in y, and 140 Å in z) than the weak density observed for the DNA-PKcs/DNA complex, consistent with the longer 35bp DNA molecule in the DNA-PKcs/Ku/DNA complex.
Figure 3.4 Eigenimage sorting of DNA-PKcs/Ku/DNA with streptavidin cryoEM particle images
(A) The first six eigenimages of the DNA-PKcs/Ku/DNA dataset are shown. The eigenimages with asterisks represent extra density at the base of DNA-PKcs and were used for sorting the particles images into four subsets. (B) Subset structures of DNA-PKcs/Ku/DNA (transparent gray) shown at ~125% of the expected volume of the complex and with the cryoEM structure of DNA-PKcs (Fig. 3.1A, magenta) docked inside. Perpendicular views are shown. (C) Selected class averages (top) and matching projections of the subset structures (bottom) with the weak extra density at the base computationally strengthened prior to projection. The multi-lobed density at the base of DNA-PKcs in the eigenimage 4 class averages is indicated by arrows. (D) Model of the complex with 35bp DNA (red) protruding into the base of DNA-PKcs (transparent gray) with simulated density for a streptavidin tetramer at the distal end of DNA (cyan). (E) Model of the complex with 35bp DNA (red) protruding into the base of DNA-PKcs (transparent gray) with simulated densities for the Ku heterodimer (yellow and magenta) and a streptavidin tetramer at the distal end of DNA (cyan). Scale bars represent 50 Å.
Class averages for one of the four DNA-PKcs/Ku/DNA subsets (Eigen 4) showed multi-lobed density at the base of DNA-PKcs of approximately the correct size for the DNA-binding core of Ku plus DNA (Fig. 3.4C, Eigen 4). The class averages with multi-lobed density appear similar to those of negatively-stained DNA-PKcs/Ku/DNA complexes (Supplemental Fig. 3.S3 and (Spagnolo, Rivera-Calzada, Pearl, & Llorca, 2006)). However, the cryoEM class averages indicated some variability in the orientation of the multi-lobes with respect to DNA-PKcs. As only one subset of the cryoEM images showed multi-lobed density, perhaps only ~25% of the particle images represent complexes containing Ku. This is roughly consistent with our estimate from a negative-stain EM analysis that at least 30% of the particle images represent complexes (Supplemental Fig. 3.S3). However, Ku is known to translocate along the DNA after associating with DNA-PKcs (De Vries, Van Driel, Bergsma, Arnberg, & Van der Vliet, 1989; Paillard & Strauss, 1991; Yoo & Dynan, 1999). Therefore, it is also possible that Ku is present in a higher percentage of the complexes but in a highly variable position with respect to DNA-PKcs.

The majority of the class averages of DNA-PKcs/Ku/DNA, those sorted with eigenimages 3, 5 and 6, can be modeled with streptavidin at the distal end of 35bp DNA assuming variability in the position of DNA relative to DNA-PKcs (Fig. 3.4D). One subset of class averages, those sorted with eigenimage 4, can be modeled more closely with Ku and streptavidin (Fig. 3.4E). However, the size of the multi-lobed density in the class averages (Fig. 3.4C, Eigen 4) is smaller
than would be expected based on the model of Ku plus DNA (Fig. 3.4E), suggesting heterogeneity or variability for Ku and DNA within the classes.

It is notable that the structure based on the Eigen 4 subset has less density at the base of the molecule than the structures based on the other three Eigen subsets (Fig. 3.4B). This is despite the fact that the class averages for the Eigen 4 subset show more density at the base of molecule than the other Eigen subsets (Fig. 3.4C). We interpret this to mean that when the DNA-PKcs/Ku/DNA complex is imaged without a solid carbon support, as is the case for our cryoEM dataset, the main DNA-binding core of Ku can adopt a variety of positions with respect to DNA-PKcs. This variability could be enabled by flexible linkages of both DNA and Ku to DNA-PKcs. We suspect from our cryoEM results on the DNA-PKcs/DNA complex that DNA may to bind to a flexible region of DNA-PKcs. The connection between Ku and DNA-PKcs is thought to be mediated by the final ~20 residues at the C-terminus of Ku86 (Falck et al., 2005; Gell & Jackson, 1999) and these residues are at the distal end of a long (100 Å) flexible arm of Ku86 (Hammel et al., 2010). The combination of two flexible linkages within the DNA-PKcs/Ku/DNA complex, while possibly providing a functional advantage, presents a challenge for cryoEM structure determination.

Discussion

There is a developing consensus that a significant fraction of proteins and protein domains are intrinsically disordered and that they are functional when they are disordered (Tompa, 2011). Image processing provides a tool to help overcome flexibility that can hamper an understanding of the molecular
architecture of complexes by cryoEM methods. This structural study of the DNA-PKcs/DNA and DNA-PKcs/Ku/DNA complexes was motivated by the desire to localize the binding site for double-stranded DNA on the DNA-PKcs molecule. The results presented here suggest that there is some level of intrinsic flexibility within these complexes making it difficult to visualize the DNA binding site.

Structures have been determined for several NHEJ components, but not for many of the complexes involved in the NHEJ pathway (Ochi et al., 2010). Understanding the molecular architecture of NHEJ complexes would help us to interpret the molecular mechanisms underlying double-strand break repair. Toward this goal we collected cryoEM datasets of DNA-PKcs/DNA and DNA-PKcs/Ku/DNA. The complexes were both formed with biotinylated DNA complexed with streptavidin, which helped to localize the distal end of DNA. Different lengths of double-stranded DNA were used to form the two complexes, with 18bp in DNA-PKcs/DNA and 35bp in DNA-PKcs/Ku/DNA. Class average images of both complexes showed extra density at the base of DNA-PKcs. However, visualizing this density at the base in three-dimensional reconstructions was challenging and required application of an eigenimage sorting procedure (Figs. 3.2 and 3.4).

Spagnolo et al. have previously published both a negative-stain and a cryoEM structure of a DNA-PKcs/Ku/DNA complex formed with a DNA molecule containing a 35 base pair duplex segment with a 19-base Y-structure at one end (Spagnolo et al., 2006). The negative-stain structure was based on 14,239 particle images and the cryoEM structure was based on a more limited dataset of
~4,000 particle images. The resolutions of the two structures were reported as ~25 Å for negative-stain and ~30 Å the cryoEM structure. Overall the shape of DNA-PKcs in the negative-stain Spagnolo et al. structure agrees well our cryoEM structures of the complex both before and after eigenimage sorting (Fig. 3.1C and Fig. 3.4B). In addition, the negative-stain Spagnolo et al. structure shows density recognizable as Ku at the base of DNA-PKcs, which is not seen in our cryoEM structures except as weak density below the base of DNA-PKcs in the eigenimage sorted structures (Fig. 3.4B). We interpret this difference as due either to a stabilization effect of negative-staining on a solid carbon support film or to formation of a more uniform complex with the specific Y-shaped DNA molecule used by Spagnolo et al. In contrast to their negative-stain structure, the cryoEM Spagnolo et al. structure shows less recognizable features for DNA-PKcs. Although the atomic structure of DNA-Ku (PDB:1JEY) is shown docked into their cryoEM structure, the shape of the density attributed to Ku is less recognizable than in their negative-stain EM structure (Spagnolo et al., 2006).

The cryoEM results presented here suggest that the broken end of DNA binds DNA-PKcs in the vicinity of the base of the molecule. We previously proposed that the free end of double-stranded DNA might enter the central channel of DNA-PKcs, as the size of the channel is large enough to accommodate double-stranded DNA (D. R. Williams et al., 2008). Although the central channel of DNA-PKcs may play a role in the repair pathway, the additional density we observed in subset structures of the DNA containing complexes was within a cone-shaped region near the base of DNA-PKcs (Fig.
Multiple intrinsically disordered regions are predicted for the DNA-PKcs sequence (Supplemental Table 3.S1). Given that the persistence length of dsDNA has been measured as 35-53nm (Brinkers, Dietrich, de Groote, Young, & Rieger, 2009; Marek et al., 2005; Rivetti, Guthold, & Bustamante, 1996) and that the dsDNA molecules used to form the DNA-PKcs complexes were significantly shorter than the persistence length (6nm for the DNA-PKcs/DNA complexes and 12nm for the DNA-PKcs/Ku/DNA complex), we surmise that at least part of the flexibility of the complexes is due to flexibility in the protein region where DNA binds to DNA-PKcs.

The inherent flexibility observed in DNA-PKcs complexes may be required for proper function. Given the high level of structural flexibility observed, it might
not be feasible to collect large enough datasets that could be sorted into sufficiently homogeneous subsets to resolve the position of DNA in the complexes. Additional experimental methods for restricting the structural flexibility of the complexes, perhaps with crosslinking agents, would be useful. Nevertheless, cryoEM with computational sorting methods offers a way to visualize flexible complexes and provide constraints for model building of molecular interactions.

Methods

NHEJ complex formation

The DNA-PKcs/DNA and DNA-PKcs/Ku/DNA samples were generously provided by Dr. Michael R. Lieber and Dr. Noriko Shimazaki. DNA-PKcs was isolated from HeLa cells by sequential column chromatography. Full-length Ku70 and Ku86 were expressed in insect cells. The DNA-PKcs/DNA sample was formed with 18bp linear DNA biotinylated at one end and conjugated with streptavidin. A control sample of the DNA-PKcs/DNA sample was also formed without streptavidin. The 18 bp DNA was made by annealing YM-27 and YM-28, which yields blunt-ended dsDNA: YM-27 5' XAG GCT GTG TCC TCA GAG G -3', X=biotin; YM-28 is complementary and without a biotin. The DNA-PKcs/Ku/DNA sample was formed with 35bp linear DNA also biotinylated at one end and conjugated with streptavidin. The 35bp DNA was made by annealing YM-18 and YM-7 to yield blunt-ended dsDNA: YM-18 5' XAG GCT GTG TTA AGT ATC TGC GCT CGC CCT CAG AGG -3', X=biotin; YM-7 is complementary.
and without a biotin. All complexes were prepared for cryoEM analysis in the absence of ATP.

CryoEM and preliminary image processing

CryoEM grids were prepared with 2 to 3μl of sample per grid, at a protein concentration of 10 to 20 nmol. Two different types of grids were used, homemade holey carbon film grids and prefabricated Protochip C-flat grids (Protochips, Inc.). CryoEM grids were frozen in liquid ethane using a custom built cryo-plunger. Cryo-electron micrographs were collected on an FEI Polara (300 kV, FEG) with the sample cooled to liquid nitrogen temperature. A Gatan UltraScan 4000 CCD camera was used for recording images with absolute magnifications of 254,669X for DNA-PKcs/DNA and 198,939X for DNA-PKcs/Ku/DNA. The underfocus values of the micrographs ranged from 1 μm to 6 μm.

Particle images were selected from micrographs using the Boxer subroutine of EMAN (Ludtke et al., 1999). Datasets were generated containing 35,176 particle images of DNA-PKcs/DNA with streptavidin, 17,323 particle images of DNA-PKcs/DNA without streptavidin, and 19,825 particle images of DNA-PKcs/Ku/DNA with streptavidin. The particle images were binned to have pixel sizes of 4.7 Å and 2.35 Å for DNA-PKcs/DNA and 4.5 Å and 2.26 Å for DNA-PKcs/Ku/DNA. The images with coarser pixels were used for initial refinement, and those with finer pixels were used for final refinement. Values for the microscope defocus and astigmatism were determined using CTFFIND3 (Mindell & Grigorieff, 2003).
IMAGIC classification and eigenimage sorting

For the eigenimage sorting procedure, the particles images of each dataset were filtered, normalized and band-pass filtered between 150 and 30 Å for MSA classification in IMAGIC (Van Heel et al., 1996). During the MSA-CLASSIFY step, the worst 15% of the images were ignored. After two rounds of classification, the particle images were subjected to an eigenimage analysis following a procedure applied to GroEL (Clare et al., 2012) and the circadian KaiBC complex (Villarreal et al., 2013). For the DNA-PKcs/DNA with streptavidin dataset, three rounds of sorting served to remove particle images with two or more DNA-PKcs molecules. This reduced the dataset from 35,176 to 13,225 particles images. The fourth round of sorting resulted in one eigenimage that described the presence or absence of density at the base of DNA-PKcs in the particle images. This eigenimage was used for sorting of the remaining dataset into a subset of 6,222 particle images that agreed with this eigenimage. This subset was subjected to an additional round of reference-free classification. Following a similar procedure for the DNA-PKcs/DNA without streptavidin dataset resulted in no eigenimages that showed density at the base of DNA-PKcs.

The eigenimage sorting procedure was also performed on the DNA-PKcs/Ku/DNA dataset and resulted in four eigenimages with density at the base of DNA-PKcs and subsets of 5,235, 4,757, 4,573, and 4,722 particles images. As with the DNA-PKcs/DNA datasets each of these subsets was subjected to an additional round of reference-free classification. The 3D Fourier Space v2 programs developed by Rubinstein that incorporate the gold standard refinement
scheme (Benlekbir et al., 2012) (www.sickkids.ca/research/rubinstein) were used to produce structures for the DNA-PKcs/DNA with streptavidin subset and the DNA-PKcs/Ku/DNA subsets from the final class average images.

3D Fourier Space refinement

Determination and refinement of the orientational parameters and three-dimensional reconstruction was performed with 3D Fourier Space v2 (Benlekbir et al., 2012). The cryoEM structure of DNA-PKcs (D. R. Williams et al., 2008) filtered to 30 Å resolution was used as the initial input map. Processing the DNA-PKcs/DNA-streptavidin and DNA-PKcs/Ku/DNA-streptavidin datasets with 3D Fourier Space produced structures of DNA-PKcs/DNA at 15 Å resolution; and DNA-PKcs/Ku/DNA at 21 Å resolution using the FSC 0.5 threshold criterion and with an outer mask radius of 80 Å. The maps are shown in Fig. 3.1B and C after they were sharpened with a B-factor of −1,000 Å² and filtered with the Cref filter (Rosenthal & Henderson, 2003) and with an additional low-pass filter at the FSC 0.5 threshold (15 or 21 Å, respectively).

Re-processing of 284,173 particle images from the cryoEM dataset of uncomplexed DNA-PKcs collected at one magnification (254kX) and binned to have a pixel size of 1.8 Å (D. R. Williams et al., 2008) was performed with 3D Fourier Space. One global search round followed by five refinement rounds resulted in a structure with a resolution of 13.6 Å at the FSC 0.5 threshold using an outer radial mask of 80 Å. A subset of 153,545 out of 284,173 particles (54%) was included in the final map. The map was sharpened with a B-factor of −1,000
Å² and filtered with the Cref filter (Rosenthal & Henderson, 2003) and with an additional low-pass filter at 13.6 Å (Fig. 3.1A).

Molecular graphics figures were produced with UCSF Chimera (Pettersen et al., 2004). The cryoEM structures of DNA-PKcs/DNA-streptavidin and DNA-PKcs/Ku/DNA-streptavidin, as well as the re-processed structure of DNA-PKcs, as determined with the 3D Fourier Space programs have been deposited in the EM Data Bank with accession numbers EMD: 5832, EMD: 5833 and EMD: 5831 respectively.

Intrinsic Disorder Prediction

Three protein disorder predictor webservers, GLOB Plot 2 (Linding, Russell, et al., 2003), IUPred (Dosztányi et al., 2005), and DisEMBL (Linding, Jensen, et al., 2003), were used to evaluate the human DNA-PKcs sequence (4,128aa, UniProt P78527 PRKDC_HUMAN).

Acknowledgements

The authors gratefully thank Michael Lieber and Noriko Shimazaki for generously supplying the DNA-PKcs/DNA and DNA-PKcs/Ku/DNA samples and for numerous insightful discussions. We appreciatively thank Dewight Williams for his effort in cryoEM data acquisition, preliminary image processing, and critical evaluation of the data. We also thank John Rubinstein for his assistance with implementation of his 3-D Fourier Space software package, Justin Flatt for his help with image processing, and the High Performance Computing Cluster.
(HPCC) at Case Western Reserve University for computational support. This work was supported by the National Institutes of Health [CA140538 to P.L.S.].
Supplemental Information

CryoEM and Image Sorting for Flexible Protein/DNA Complexes

Seth A. Villarreal\textsuperscript{a}, Phoebe L. Stewart\textsuperscript{a,*}

\textsuperscript{a} Department of Pharmacology and Cleveland Center for Membrane and Structural Biology, 10900 Euclid Ave., Case Western Reserve University, Cleveland, OH 44106, US
Figure 3.S1 Comparison of the cryoEM and crystal structures of DNA-PKcs
The 3D Fourier Space refined cryoEM structure of DNA-PKcs is shown contoured at 100% of the expected volume (transparent gray). The crystal structure of DNA-PKcs with the C-terminal region of Ku86 (Sibanda et al., 2010) is shown filtered to 7 Å resolution, contoured to reveal α-helical rods (red), and docked within the cryoEM density with UCSF Chimera (Pettersen et al., 2004). Scale bar represents 50 Å.
Figure 3.S2 Gold standard Fourier Shell Correlation (FSC) curves
FSC curves are shown for (A) DNA-PKcs, (B) DNA-PKcs/DNA, and (C) DNA-PKcs/Ku/DNA processed with the 3D Fourier Space package (Benlekbir et al., 2012). For resolution assessment, all maps were calculated with an outer radial mask of 80 Å. The FSC 0.5 and 0.143 threshold levels are indicated.
Figure 3.S3 Selected class average images of a negative-stain DNA-PKcs/Ku/DNA dataset
A dataset of 2,840 particle images was processed with IMAGIC reference-free image classification (Van Heel et al., 1996). A selected subset of class averages is shown with DNA-PKcs in a side view and with recognizable density for Ku (red arrow, first class average). Visual inspection of the full set of class average images indicated that ~30% of the averages show recognizable density for Ku at the base of DNA-PKcs. A similar percentage of desired complexes (~30%) was found for the DNA-PKcs/DNA with streptavidin sample. Scale bar represents 50 Å.

Figure 3.S4 Eigenimage sorting of uncomplexed DNA-PKcs cryoEM particle images
The first six eigenimages of the DNA-PKcs dataset are shown with the first eigenimage representing the predominant side view of DNA-PKcs. Scale bar represents 50 Å.
Table 3.S1 Results of protein disorder predictor webservers for the human DNA-PKcs sequence

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Percent Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLOB Plot 2(^a)</td>
<td>9%</td>
</tr>
<tr>
<td>IUPred(^b)</td>
<td>5%</td>
</tr>
<tr>
<td>DisEMBL(^c) (loops/coils)</td>
<td>39%</td>
</tr>
<tr>
<td>DisEMBL(^c) (hot-loops)</td>
<td>13%</td>
</tr>
<tr>
<td>DisEMBL(^c) (remark-465)</td>
<td>4%</td>
</tr>
</tbody>
</table>

\(^a\)GLOB Plot 2 (Linding, Russell, et al., 2003)

\(^b\)IUPred (Dosztányi et al., 2005)

\(^c\)DisEMBL (Linding, Jensen, et al., 2003)
Chapter 4: Conclusions

Understanding cellular processes includes mapping out the interacting components, but also the structural characteristics of the involved proteins that enable those required interactions.

The circadian system of *S. elongatus* relies upon the interactions of the three Kai proteins, which use ATP to drive a cyclical pattern of phosphorylation changes to signal time. The protein-protein interactions between KaiB and KaiC drive the latter half of this cycle, with the binding of KaiB to KaiC inducing dephosphorylation on KaiC (Ishiura et al., 1998; Iwasaki et al., 1999; Nishiwaki et al., 2000). My study on this system provides evidence of how KaiB binds to KaiC as six monomers, at the CII domain of KaiC above the ATP binding clefts. Each monomer of KaiB forms an interaction with KaiC that involves residues of two KaiC subunits. This binding surface bridges the KaiC ATP binding cleft formed by the interface of the subunits. The role of KaiB is the promotion of the autophosphatase activity of Kaic, to drive KaiC towards the unphosphorylated state and complete a 24 hour circadian cycle. In carrying out this process, KaiB may be directly affecting the stability of the KaiC ATP binding cleft. KaiC dephosphorylation uses ADP to accept the phosphate, and so this work led to the hypothesis that KaiB promotes this activity by stabilizing the KaiC ATP binding cleft in a conformation favorable for ADP binding (Villarreal et al., 2013).

The second system I examined is involved in a human DNA damage response pathway. Non Homologous End Joining repairs double strand DNA breaks, providing a cell with a way to minimize potential loss of DNA, remove
chemically damaged bases, and reconnect broken DNA strands. DNA-PKcs fulfills a critical role within this pathway as a scaffold, but understanding the structure of DNA-PKcs and DNA-PKcs complexes has proven challenging (Downs & Jackson, 2004; Uematsu et al., 2007; Yano et al., 2009). Through single particle cryoEM and image sorting techniques, I determined structures of DNA-PKcs in complexes with DNA and Ku, which provide information about how DNA interacts near the base of DNA-PKcs in the initial stages of the NHEJ DNA repair pathway. In contrast to the hypothesis going into this project, that DNA-PKcs binds DNA within the cavity of the protein, the structures of the two complexes show additional density distant from the base of the protein. This unexpected finding led to the interpretation that DNA in complex with DNA-PKcs extends from the base of the protein. The proximity of this location to the HEAT repeat connection between the upper and lower portions of DNA-PKcs suggests a possible signaling mechanism, via these HEAT repeats, to regulate DNA-PKcs kinase activity. As DNA-PKcs plays a central role as a scaffold and regulatory kinase in the NHEJ repair pathway, understanding the initial stages of this pathway provides a framework with which to pursue a more detailed understanding of the repair process (Villarreal & Stewart, 2014). Understanding the structure of the proteins and complexes that are intrinsic to both of these biological systems, circadian rhythm and NHEJ DNA repair, allows for a better understanding of their underlying mechanisms.
KaiBC Research

Advancing the understanding of the natural world requires continued refinement of ideas and assumptions. In the case of *S. elongatus* circadian rhythm there have been several published papers presenting contradictory information about the nature of the protein-protein interaction of KaiB and KaiC. Circadian rhythm is useful to an organism as it allows prediction of repeating cycles. In the case of *S. elongatus*, the mutually opposed pathways of nitrogen fixation and photosynthesis do not have the benefit of organelle separation as in eukaryotes. Furthermore, both are required for survival (Huang et al., 1990; Schneegurt et al., 1994). Ensuring that the photosynthetic pathway remains active during periods of time in which this cyanobacteria is likely to be exposed to light provides a selective benefit to the organism. The cycling pattern of phosphorylations and dephosphorylations on KaiC is the basis for the circadian rhythm in *S. elongatus* (Dunlap, 1996; Ishiura et al., 1998). Understanding the mechanisms driving this system provides insight into how this system achieves the required rhythmicity.

One of the difficulties of working with circadian rhythm in *S. elongatus* arises from the balanced interactions of the components, making purification of a single complex difficult. The ticking of this clock is not discrete, and the KaiB-KaiC interactions proved particularly difficult to assess. The three Kai proteins within each *S. elongatus* cell exist as populations. The temporal signal is based on the average phosphorylation state of KaiC. As any particular KaiC hexamer may not be providing an accurate temporal signal, this averaging avoids any
particular out of sync signal from completely disrupting the circadian cycle (Nishiwaki et al., 2007). However, while this characteristic benefits *S. elongatus* in its normal environment, it makes purifying individual complexes difficult, as only the phosphorylation state of the individual KaiC hexamer dictates the complexes that are most favorable. Purifying a particular complex is difficult because at any point in the cycle there is not a homogenous sample of a single complex. Information on the protein-protein interactions between KaiC and KaiA was determined first, and the study of KaiB-KaiC interactions lagged and has been controversial.

In chapter two of this thesis, research is presented on a cryoEM structure of the KaiBC complex, stabilized by deletion of the KaiC C-terminal tails. In this cryoEM structure, the third ring of density was assigned to KaiB, and appeared continuous around the top of KaiC. Using molecular simulations of the individual crystal structures of KaiB and KaiC, in conjunction with the density map of the complex, allowed me to predict the binding surface between the two proteins. The most favorable binding surface was created by 6 KaiB monomers on the CII domain of KaiC (Villarreal et al., 2013).

There is more recent evidence that lends support to this model of KaiB-KaiC interactions. The Heck lab used tandem mass spectroscopy (MS) to make stoichiometric measurements of the complexes formed during *in vitro* cycling, and additionally used hydrogen-deuterium exchange MS (HDX-MS) to identify the probable binding surfaces between KaiB and KaiC. Their results showed KaiB interacting with KaiC as a monomer, and that the binding of KaiB to KaiC is
cooperative, such that the binding of an initial monomer of KaiB leads directly a total binding of six monomers. Furthermore, the HDX-MS supported KaiB binding to the CII domain of KaiC (Snijder et al., 2014). While this work may not settle all debate, the basis of scientific understanding is characterization, and further information will help to eventually understand this circadian system.

Future Work on S. elongatus Circadian Rhythm

The binding surface of KaiC for KaiB, indicated by my work, is proximate to the ATP binding cleft. This work is in agreement with known work on the binding of KaiB inducing dephosphorylation on KaiC (Pattanayek et al., 2013). Furthermore, the resiliency of the in vitro cycling and overall stability of the in vivo system was proposed to be due in part to KaiC monomer-hexamer exchange during the dephosphorylation phase of KaiC (Kageyama et al., 2006; Mori et al., 2007). This additional aspect of the KaiABC system increases its reliability and may represent an additional function of KaiB. If KaiB promotes KaiC monomer-hexamer subunit exchange in addition to promoting dephosphorylation, then it may do so by modulating the interface between the CII domains of the KaiC subunits. One potential method to examine this proposed interaction would be isotope tagging a sample of KaiC and then mixing with unmarked KaiC, with and without added KaiB. As a baseline, unphosphorylated KaiC samples could be mixed and then after a set time examined to determine the fraction of exchange between tagged and untagged KaiC hexamers. Varying the starting mixtures would allow testing of various influences on KaiC hexamer subunit exchange. Purified KaiA and KaiC in a premixed tube provides a highly phosphorylated
state of KaiC. Mixing in KaiB, either in the native state or with modifications, would then possibly adjust the rate of subunit exchange. Testing various mutations to either KaiB or KaiC would allow separation of function by different portions of the structure. It is known that the KaiB C-terminal tails influence the binding affinity between KaiB and KaiC. These tails may interfere with the interface between adjacent KaiC subunits, and may increase the rate of subunit exchange to facilitate dephosphorylation of the overall population of KaiC.

During the crossover period between conditions favorable to phosphorylation of KaiC due to KaiA, to the later portion of the cycle with KaiB induced dephosphorylation of KaiC, there is a KaiABC complex with all three of these proteins interacting (Mori et al., 2007). Understanding the process by which KaiB sequesters the function of KaiA would further clarify this step in the circadian cycle. However, difficulties arise in generating structural data on this complex. The issues are similar to those that affected the study of the KaiBC complex, with the added complication of the multiple possible interactions between KaiA and KaiB. One option for gathering further structural information about these interactions would be to collect cryoEM images of KaiABC complexes stabilized in vitreous ice from the time point during of the circadian cycle of their highest percentage of the overall population of complexes, and perform subsequent computational sorting of the collected dataset to select for the KaiABC complex. Due to the nature of this circadian system, unmodified KaiA, KaiB, and KaiC proteins are not amenable to forming a stable three component complex through just mixing the components. Examining a particular
multi-protein complex in purified form requires artificially manipulating this system, such as the KaiCΔ489 mutation used for formation of a more stable KaiBC complex in my cryoEM study. However, these changes can create misleading or artificial structural changes that are difficult to distinguish from the physiologically relevant structure. A possible experiment to ascertain structural information of a near native state KaiABC complex would be to use single particle cryoEM and sorting analysis. By creating an enriched subset from a heterogeneous dataset, information from undesired complexes can be minimized, allowing insight into the structure of the triple protein complex. Such structural information could help determine how KaiB sequesters KaiA, preventing further phosphorylation. A possible outcome is that the binding interface of KaiBC prevents the formation of the proper conformation of the KaiC C terminal tails during their interaction with KaiA, preventing the pro-phosphorylation effects of KaiA.

**DNA-PKcs Research**

The DNA damage response encompasses a variety of pathways, and plays a daily role in preventing cell death due to both internal cell processes and environmental conditions. In the case of NHEJ, the repair process is initiated by Ku binding, and subsequently, DNA-PKcs is thought to provide the hub around which the repair process occurs. An understanding of the protein-protein interactions based around DNA-PKcs, and the signaling functions of DNA-PKcs as a kinase, will enhance our understanding of the NHEJ pathway and help to identify therapeutic targets.
In chapter 3, the DNA and DNA-PKcs cryoEM structural data suggests that DNA interacts at the base of DNA-PKcs. Previously, the large open middle of DNA-PKcs was suggested as a possible binding location for DNA. Two complexes were examined to gain an understanding of the initial stage of the NHEJ repair pathway. Even a two component complex of DNA-PKcs and DNA presented a reconstruction hampered by flexibility, temporarily precluding the identification of the DNA binding location from this dataset. Image sorting provided a more conformationally consistent subset that allowed visualization in 3D of the streptavidin tag at the end of the DNA. This complex, and a second complex of DNA-PKcs, Ku, and DNA analyzed in a similar fashion, both suggested that the base of DNA-PKcs bound DNA in the vicinity of the HEAT repeat connection between the base of the protein and the kinase containing upper portion. While presenting information about the location of DNA in complex with DNA-PKcs, the interpretation of these structures also raised the question of the activation of the kinase domain on DNA-PKcs. Noting that there are HEAT repeats adjacent to the DNA binding area, in a HEAT repeat connection between the base and kinase region, suggests that the HEAT repeats may serve to relay a signal after DNA binding at the base of DNA-PKcs. The process by which DNA is protected and presented for repair requires further evaluation of this interface as it is an important aspect of effectively repairing DNA while minimizing loss of genetic information through base degradation.
Future Work on DNA-PKcs and NHEJ

Application of both x-ray crystallography and cryoEM techniques have so far only provided moderate resolution information with which to interpret the overall structure of DNA-PKcs (Sibanda et al., 2010; D. R. Williams et al., 2008). Despite this limited information set, a degree of interpretation has been possible. The crystal structure lacks 54% of the residues comprising this full protein and thus it was not possible to trace the polypeptide chain. The identified sections of DNA-PKcs consist of mainly the kinase domain and two arms below the kinase domain, formed from α-helical HEAT repeats (Sibanda et al., 2010). In contrast, the cryoEM structure shows a similar kinase domain, but the extended arms are shifted by ~10Å compared to the x-ray structure. In addition, the cryoEM structure shows these arms connecting to a large volume of density, forming a base for the overall protein (D. R. Williams et al., 2008). While the shift in the arms may be due to consequences from crystal packing in x-ray crystallographic structure or the plunging process in cryoEM, the presence and absence of a significant volume of density in the two structures is an example of the difficulty in working with this protein. Examining DNA-PKcs in complex with additional components does not alleviate the structural difficulties inherent in working with DNA-PKcs. Despite these issues, additional understanding of the protein-protein interactions involved in DNA-PKcs complexes would provide insight in the presentation of DNA for removal of damaged bases. To further clarify the role of DNA-PKcs as a scaffold for the NHEJ process, better information about the structure of DNA-PKcs and its complexes is needed.
Further efforts with purified DNA-PKcs may allow either x-ray crystallography or cryoEM to provide increased detail into the structure of this protein. However, these techniques may also fall short. Stabilization of DNA-PKcs through modification may facilitate further understanding of this protein’s structure. Stabilizing antibodies have been used to adjust the flexibility of other proteins and make them more amenable to crystallization efforts, such as in the example of the $\beta_2$ adrenergic receptor (Chung et al., 2011; Rasmussen et al., 2007, 2011). In the case of DNA-PKcs, generating antibodies to sections of the protein outside of the kinase domain may provide this desired stability. Furthermore, using cryoEM to quickly screen the location of these antibodies would provide a measure of selection over both how they are oriented, and ensuring that the overall structure remains similar to the known characteristics of DNA-PKcs. Using the stabilized DNA-PKcs for crystallization may provide higher resolution diffraction patterns, allowing the full backbone of the protein to be traced. As an alternative to working with the entire protein, sections of the protein may be stable when expressed as individual domains. Using NMR or x-ray crystallography to assess individually expressed domains of DNA-PKcs may allow a piecemeal assembly of a structure of the overall protein. Furthermore, understanding how DNA-PKcs manages the complexes involved in NHEJ has been hampered by the flexibility of the protein. Even if not all of the individual domains are amenable to structural techniques, understanding portions of DNA-PKcs structure in detail be helpful. One example is the location of the Ku C-terminal binding site on DNA-PKcs, which remains unclear. Ku and DNA-PKcs
bind DNA, leading to activating phosphorylations. Clarifying the activation signals of this complex is an important step in understanding the NHEJ pathway.

CryoEM has demonstrated the ability to work with complexes involving DNA-PKcs, and as such offers a possible method to further analyze a more stable complex. Partially exposed DNA while bound to DNA-PKcs offers the chance to place a more proximate tag to the complex for easier visualization. Biotinylation and subsequent labeling with streptavidin provides a reliable tag with which to mark DNA, but the link between biotin and DNA allows streptavidin freedom to move about in three dimensional space relative to the biotinylated end of DNA. This reduces the likelihood of collective single particle images showing streptavidin in the same location relative to the complex, leading to a noisier signal for streptavidin. Exposed DNA allows for proteins with an appropriate helix-turn-helix motif to bind the major grooves of DNA, one possible protein is the lac repressor protein (Lewis, 2005). Rather than using a tag bound via a flexible linker, a DNA binding protein tag would have the advantage of being directly adjacent to the bound DNA and may provide a more reliable location for DNA in complex with DNA-PKcs. Understanding more clearly how DNA binds DNA-PKcs would be helpful to further clarify the regulation of DNA repair with the NHEJ pathway.

Summary

Understanding the basic pathways required by living cells remains a complex process. While mapping out the components of a pathway provides a broad overview of the system, structural information of the complexes involved is
helpful for understanding the molecular mechanisms. Structural techniques are necessary to understand the protein-protein interfaces that comprise many of the interactions required for normal function. In the case of *S. elongatus* circadian rhythm, a single cell can keep track of time, using both entrainment signals from the environment and more importantly, a circadian rhythm that limits the detrimental effects of atypical environmental conditions. KaiC undergoes phosphorylations induced by KaiA, and the fully phosphorylated KaiC hexamer binds KaiB leading to dephosphorylation. This cyclical process forms the basis of circadian rhythm in *S. elongatus*. KaiB and KaiC form a complex starting the progression into the latter half of this 24 hour cycle, as KaiC returns to the fully dephosphorylated state. The cyclical process of KaiC phosphorylation is critical to the function of this biological clock, and understanding how the KaiBC complex accomplishes its role in this process provides insight into this circadian system and the structural mechanisms which may be used by other clocks.

The second project in this body of work examined two flexible complexes of DNA-PKcs involved in DNA break repair. Preventing the accumulation of mutations is a critical aspect to limiting the development of cancer, and congenital mutations in these pathways lead to a number of disorders. Understanding how DNA-PKcs manages the wide variety of DNA lesions that can occur during double strand breaks requires knowledge of the structural architecture of this complex. The peculiar shape of DNA-PKcs and the role it plays in protecting and presenting DNA during repair remains an area worthy of study. As a therapeutic target, understanding the functional portions of this
protein remains a beneficial avenue of research. Leveraging the most appropriate tools of structural biology to further characterize the mechanisms sustaining living organisms allows greater opportunities to comprehend the broader environment and benefit human health.
Role of Circadian Clocks in Human Health

Human health and function is affected by daily patterns of physiological changes. Activities and environmental stress can vary in their potential physiological impact based on the timing of their occurrence. Modern study of these patterns began with the recognition of internal consistency for activity despite the removal of noted environment cues, suggesting the possibility of internal biological mechanisms providing an approximate time of day, or circadian rhythm (Aschoff, 1960; Pittendrigh, 1960). Study of these processes has proceeded from both emergent patterns of activity and study of the underlying biological mechanisms (Aschoff, 1984). Simple exhaustion or jetlag can play a significant role in detrimental effects on capability, which is readily documented in industry (Folkard & Tucker, 2003; Kerin & Aguirre, 2005). Distinguishing behaviors influenced by circadian rhythm from the study of these two directions was facilitated by the identification of the core biological mechanism for the mammalian circadian rhythm. The suprachiasmatic nucleus (SCN) within the hypothalamus of the brain was shown to play a role in both physical activity (Stephan & Zucker, 1972) and hormone regulation (Moore & Eichler, 1972). Eventually, the SCN was demonstrated to be the master clock and pacemaker for mammalian circadian rhythm (Pando et al., 2002; Ralph et al., 1990).

The recognition of many of the components of these biological mechanisms facilitates study of their effects upon human health, particularly in
the area of metabolic disease. Human behavior provides useful opportunities to examine the interplay between internal rhythms and the environment. In particular, shift work and changes in time zone both provide clues about the effect on disrupted circadian cycles on health (Haimov & Arendt, 1999; van Amelsvoort et al., 1999).

These studies provide information about possible effects resulting from changes in an individual’s actions and environment. Understanding the genetic underpinnings of these interactions allows more nuanced characterization of these health effects. Towards that goal, studies of human polymorphisms occurring within proteins of the circadian clock have been implicated in weight gain during development and adulthood (Garaulet et al., 2009; Goumidi et al., 2013). Studies in model organisms offer the opportunity for controlled environments and models for the analysis of individual variables, both environmental and genetic. Comparisons between sleep trends in humans and controlled sleep deprivation studies in mice show notable agreement in tendencies towards glucose metabolism (Barclay et al., 2012; Spiegel, Tasali, Leproult, & Van Cauter, 2009). Mice studies of mutations in circadian transcription factors have been demonstrated to lead to hyperphagy and obesity (Turek et al., 2005), as well as more specific effects in lipid metabolism (Grimaldi et al., 2010).

Timing Mechanisms in Humans and Other Model Organisms

Understanding the human circadian systems has been a significant driver of research in both humans and model organisms. The characterization of these
biological processes benefits from the ability to manipulate the genetics and environment of model organisms. Through these studies, much has been learned about the biology of circadian rhythms. The complexity of the organism plays a notable role in the types of studies that are readily conducted. Long term studies of humans is difficult and expensive, frequently relying on self-reporting or databases, with limited biological tests. Circadian disruptions commonly occur as part of a job or activity, voluntarily engaged in for a variable amount of time. For example, one analysis of jet lag over time was limited to a group in Iceland for which their health records already existed, and only required permission to access (Rafnsson, Tulinius, Jónasson, & Hrafnkelsson, 2001). Furthermore, the study itself noted limitations in accounting for environmental effects such as increased exposure to radiation at higher altitudes. In comparison, studies in a model organism allow careful control of the entire environment, creating jet lab effects that occur reliably and precisely for a sample population of genetically and environmentally controlled mice (Davidson et al., 2006). The types of studies that are possible depend significantly on the model organism. Study of humans is limited to assessments of environmental and preexisting genetic factors over the course of uncontrolled conditions or limited time within a controlled environment. By comparison, mammals such as mice offer greatly increased opportunities for more finely tuned disruptions. However, consummate with the degree of similarity to human systems, is the cost of maintaining and modifying a model organism. Simpler systems are often lower cost and facilitate experimentation upon a particular pathway or environmental cue. A study of shiftworkers noted increased
mortality from coronary heart disease and diabetes, positively correlated with the duration of shiftwork over the course of a career in paper and pulp manufacturing (Karlsson, Alfredsson, Knutsson, Andersson, & Torén, 2005). In comparison, Drosophila melanogaster, allowed researchers to test multiple light conditions over the lifespans of the fruit flies, and demonstrate a significant difference in lifespan between a group in normal conditions (24 hour days, with light and dark periods) as compared to groups under abnormal lighting conditions (Pittendrigh & Minis, 1972). The ability to modify an organism allows for complementary comparisons, of normal environmental conditions, but genetically disrupted circadian rhythms. Select mutant strains of Arabidopsis thaliana have unusual periods in their circadian rhythms, and so, allow growth under nonstandard conditions matching their unusual circadian periods (Dodd et al., 2005). The benefits of using plants in this case was the ability show the metabolic benefits to the matching environmental conditions with the internal circadian rhythm of the plants. Simple questions about underlying biological patterns can still remain, such as the accuracy of assuming the role of the circadian rhythms in providing a benefit to fitness. The selection pressure for development of circadian rhythms within an organism has been suggested as arising from either intrinsic (Pittendrigh, 1993) or extrinsic (Beaver et al., 2002; Michael et al., 2003) sources. Unicellular models provide a rapid means to examine multigenerational trends in conjunction with perturbations from normal functions. Synechococcus elongatus, a cyanobacteria, provides a simple clock, with multiple circadian period affecting mutations and mutations in which circadian rhythms are nonfunctional. In a
comparison of growth rates, population variations of *S. elongatus* provide a good measure of reproductive fitness, and allow straightforward assessment of the possibility of selection pressure being entirely intrinsic (Woelfle, Ouyang, Phanvijhitsiri, & Johnson, 2004). Through this work on *S. elongatus*, it was demonstrated that the metabolic benefits are, at least in part, based on their predictive ability for environmental fluctuations.

There are numerous differences between model organisms and humans, even for the pathways that give rise to circadian rhythm. The three taxonomic kingdoms of life possess few overlapping proteins within the area of circadian rhythm (Van Ooijen et al., 2013). The previously described benefits of studying diverse model organisms remain.

Small mammals, in particular mice and rats, but also other rodents, possess many attractive characteristics as a substitute for studies on humans. The initial localization of the circadian oscillator was made by back tracing retinohypothalamic axons in several mammals (Hendrickson, Wagoner, & Cowan, 1972; Moore & Lenn, 1972) and subsequent ablation of the SCN (Stephan & Zucker, 1972). As previously mentioned, the isolation of the SCN in rats then demonstrated both the loss of circadian rhythmicity outside of the SCN, and the continuation of oscillation within the tissue (Inouye & Kawamura, 1979). Further work continued on both components of these systems, and environmental inputs. Here an issue arose demonstrating difficulties in translating work on any model organism to humans. Multiple small mammals use light cues from the eyes to entrain the circadian oscillator of the SCN to the environmental
light dark cycle (R. J. Nelson & Zucker, 1981), as demonstrated through blinding and subsequent loss of environmental entrainment of circadian rhythm through light. However, even with this information, a human study led to the conclusion that extraocular circadian photoreception was possible (Campbell, 1998). While eventually demonstrated as inaccurate (Wright & Czeisler, 2002), the ease of modification to ensure loss of pathways in mice and other small mammals has shown utility in direct comparisons of human circadian systems (Lucas et al., 2001).

While mice and other rodents are amenable to mutation and genetic models, *Drosophila melanogaster*, has provided numerous insights in the genes and proteins that underpin the circadian systems. The ease of mutant generation facilitated the discovery of *period*, the first clock mutant, over 40 years ago (Konopka & Benzer, 1971). *D. melanogaster* also provided the first complete list of components for the oscillations of a circadian loop. A previously known gene, *clock*, had been characterized as necessary for the function of circadian rhythm in multiple organisms (King et al., 1997), and was examined for potential binding to the *period* promoter. In this work, the Kay lab, showed that CLOCK increases the transcription of *period*, which in turn leads to inhibition of CLOCK (Darlington, 1998). The alternating patterns of promotion and inhibition provided the variance in activity levels of these proteins, giving rise to oscillations.

The third common eukaryotic model organism for the study of circadian rhythm is the fungus *Neurospora*, benefiting from the ease of genetic analysis as a haploid organism, simple growth requirements, and relative ease of
modification compared to mammals and insects (Lakin-Thomas, Bell-Pedersen, & Brody, 2011). The initial identification of temporal rhythm occurred in 1959 (Pittendrigh, Bruce, Rosensweig, & Rubin, 1959), beginning lines of research that continue into the present (Hong et al., 2014; Lakin-Thomas et al., 2011). In combination with *Drosophila*, the genetic studies in *Neurospora* led to recognition of genetic components as the primary driver of circadian rhythm. The first gene identified in *Neurospora* was *frequency*, using chemical mutagenesis to create three different mutants with unique periods of oscillation (Feldman & Hoyle, 1973). Additional studies have identified further genes and components in this pathway, with similarities to other eukaryotic systems (Loros & Dunlap, 2001). A current area of research in this model organism is the method of temperature compensation for maintenance of oscillation periods (Gooch et al., 2008).

**Identification of *Synechococcus elongatus* Circadian Cycle**

The study of eukaryotic systems alone may have hampered the understanding the evolutionary changes of clock systems. While such model organisms are of significant use in understanding human biology, their similarity also bring a corresponding degree of cellular complexity. Initially, inherent circadian rhythms were believed to require the existence of a nucleus and the greater complexity of eukaryotes, and that such clocks arose due to the benefits in metabolism over the generally longer lifetime of such cells (Kippert, 1987). At this time, an interesting aspect of unicellular cyanobacteria was recognized. Namely that the simplicity of these organisms prevented the spatial separation of two incompatible biochemical pathways: oxygenic photosynthesis and nitrogen
fixation, which is oxygen sensitive. The initial clarity on this conundrum showed only that these processes occurred during different stages of the cell cycle (Mitsui et al., 1986). Further temporal metabolic patterns emerged, including insensitivity to temperature fluctuation for periods of nitrogen fixation and photosynthesis (Huang et al., 1990), the rates of amino acid uptake (T.-H. Chen et al., 1991), and buildup of nitrogen fixation components (Schneegurt et al., 1994).

One particular cyanobacteria was especially useful for study. *S. elongatus*, is amenable to a number of techniques for genetic manipulation (Golden, Brusslan, & Haselkorn, 1987; Golden, 1988), including native susceptibility to transformation (Golden & Sherman, 1984) and conjugation with *Escherichia coli* for larger segments of DNA (Elhai & Wolk, 1988). Work proceeded quickly in *S. elongatus* compared to eukaryotic systems. Luciferase reporters enabled rapid screening for the identification of clock controlled genes (Kondo et al., 1993), leading to identification of widespread circadian influence across the *S. elongatus* genome, as well as circadian phenotypes, with clock periods ranging from 16 to 60 hours or arrhythmia (Kondo & Ishiura, 1994). While the basis for the oscillator proved time consuming to find (Golden, Ishiura, Johnson, & Kondo, 1997), screening complementation efforts eventually led to the localization of the *KaiABC* gene cluster as the likely basis for circadian rhythm in cyanobacteria (Ishiura et al., 1998). This particular paper by Ishiura *et al.*, jointly from the Kondo lab at Nagoya University and the Johnson lab at Vanderbilt University also
identified several important aspects of in the protein products from this gene cluster, naming it “kai”, the Japanese word for “cycle”.

Starting at transcription, these efforts identified KaiA as individually transcribed, with the remaining two genes, KaiBC, under dicistronic transcription. Initially, this locus was tested for rescue ability against known clock cycle mutants. A plasmid with KaiABC was able to rescue over 50 period and arrhythmia mutants. In comparison, deletion or inactivation of any one of these three genes is sufficient to negate circadian rhythm within S. elongatus. Notably, despite the loss of rhythmicity, the mutants grew equivalently to WT under continuous lighting conditions. This suggested that these genes were solely relevant to maintaining circadian rhythm, and were otherwise not involved in growth (Ishiura et al., 1998).

Moving to translation, Ishiura et al. also did a cursory examination of the KaiABC protein products, which are predicted to be of relatively small size, 284aa (KaiA), 102aa (KaiB), and 519aa (KaiC). Only KaiC presented recognized functional motifs, possessing a Walker’s motif A in both the N and C terminal domains of the protein, for either ATP or GTP, and two putative catalytic carboxylate Glutamate residues. While there was a limited knowledge of function for these three proteins, they were recognized as playing a role in their own regulation. Overexpression of KaiC from a second location and an external promoter led to repression of KaiBC, whereas overexpression of KaiA abolished rhythmicity (Ishiura et al., 1998). In particular, the authors note that KaiC may play a role as a “state variable”, similar to frq in Drosophila (Dunlap, 1996).
Based on the work in this paper, KaiABC was characterized as an essential part of a transcription-translation feedback loop, giving rise to a circadian rhythm in *S. elongatus* (Ishiura et al., 1998).

**Kai A, B, and C as the Core Oscillator**

The Kondo lab continued their work on the KaiABC proteins, looking for possible protein-protein interactions, surmising a high potential for protein-protein interactions due to the tendency of prokaryotes towards cluster organizations of genes with cooperative functions. Using a yeast two-hybrid system (Fields & Song, 1989), Iwasaki *et al.* were able to show the formation of homodimers of KaiB and KaiC, but the assay gave a background positive signal for KaiA, limiting assessment of homotypic interactions for that protein. This same assay also showed KaiA-KaiC and KaiB-KaiC interactions. While the sequences of KaiA and KaiB did not possess recognized features, KaiC possessed a curious repetition in its sequence, with the first half of the protein possessing 21% and 42% sequence identity and similarity, respectively to the second half. Dubbed the CI and CII domains, a further analysis of individually expressed domains suggested that both were able to interact with KaiA and KaiB. Additionally, overexpression of either domain was sufficient to suppress *KaiBC* expression, and Iwasaki *et al.* noted mutations in both domains capable of preventing normal circadian oscillations. In the same paper, the Kondo lab demonstrated that KaiABC interactions allow KaiC to oscillate in its influence over the level of *KaiBC* transcription and may form the basis of a transcription-translation feedback loop (Iwasaki et al., 1999).
Two years later, the Johnson lab demonstrated that the degree of expression of KaiC has a direct correlation with the extent of induced phase shifting of the circadian cycle. Furthermore, the authors found that blocking translation and cell growth, via chloramphenicol, didn’t eliminate the overall circadian rhythm. Expression of KaiC continues even under conditions limiting translation to 75-80%, suggesting it has both preferred expression status and is independent of growth (Yao Xu et al., 2000).

Both domains of KaiC possess the same predicted ATP-/GTP-binding domain. An *in vitro* expression system and a binding assay showed that KaiC primarily bound ATP in the CI domain, and this binding could be blocked by a point mutation. Furthermore, incubation with ATP resulted in autophosphorylation of KaiC (Nishiwaki et al., 2000). In recognition of the role of such a capability in circadian clock loops (Dunlap, 1999), Nishiwaki *et al.* hypothesized that KaiC autophosphorylation may play a role in the ability of KaiC to regulate *KaiBC* transcription or provide a signal to an unknown pathway (Nishiwaki *et al.*, 2000).

While KaiC was emerging as the most complex of the three Kai proteins, further work in the Kondo lab identified an additional interacting protein as necessary for maintaining normal circadian function. SasA, a histidine kinase, interacts with KaiC. Furthermore, the region of interaction, the N-terminus, shows 26% identity and 60% similarity to KaiB. However, despite the similarity between SasA and KaiB, SasA does not form homodimers. Loss of SasA leads to arrhythmia, but notably this effect occurs over several days and requires bright light conditions. In the reverse direction, overexpression of SasA suppressed the
KaiBC promoter. However, while both absence and abundance of SasA is a blow to the oscillatory mechanism, it is not strictly necessary to give rise to periodicity of KaiABC level (Iwasaki et al., 2000).

The role of KaiA was simply characterized as a positive regulator of KaiBC expression, but Iwasaki et al., from the Kondo lab, followed up through examination of potential interactions between KaiA and the autokinase activity of KaiC. In an in vitro assay using $^{32}$P, KaiA was found to dramatically enhance KaiC autophosphorylation. Furthermore, using 2D Thin Layer Chromatography, they identified KaiC as having both Serine and Threonine phosphorylations. Due to this KaiA-induced change on KaiC, the authors proposed that phosphorylation of KaiC affects its stability, and its ability to regulate the rate of KaiBC transcription (Iwasaki et al., 2002).

Identification of several possible types of KaiABC complexes provided a coarse box in which to identify different relevant interactions at various points through a clock cycle. Using immunoblotting and co-immunoprecipitation, Kageyama et al. provided an overview of the in vivo dynamics of the general KaiC interaction trends at points throughout a 24 hour cycle. During the day, early in the subjective time period, there are smaller discrete complexes, with the two largest complexes showing weights appropriate for a possible KaiC hexamer and a KaiA dimer. Over the last 12 hours, the night portion, larger complexes form, which the authors interpret as various components using a KaiC hexamer as a scaffold. While the relatively larger size of a KaiC hexamer allowed a more straightforward discrimination between complexes that could, and those that
could not, contain it, the predictions of this work in regards to KaiA, KaiB, and SasA were more limited. Multiple possible interactions were postulated, and finally, the authors noted that protein accumulation from \textit{KaiBC} transcription occurred several hours prior to peak association for the larger complexes. While the chemical or structural reason for the delay was unclear from these results, the delay itself was proposed as containing an important temporal step in achieving 24 periodicity (Kageyama et al., 2003)

The role of KaiB remained poorly understood, despite insights on the other components in the circadian cycle. KaiA induced phosphorylation events on KaiC, while KaiC played a role in limiting transcription of the \textit{KaiBC} gene locus. KaiB is necessary to sustain circadian rhythm in \textit{S. elongatus}, and Kitayama et al. examined temporal interactions to identify a potential role for KaiB within this system. Noting that KaiA induces KaiC phosphorylation, and that KaiC association with KaiA occurs prior to association with KaiB, the authors predicted that KaiB may be affecting the phosphorylation state of KaiC. Sidestepping the complexity of \textit{in vivo} efforts, the Kondo lab found that, while simply mixing KaiA and KaiC promoted KaiC phosphorylation, the addition of KaiB significantly lowered the peak level of phosphorylation achieved. Through this work, the first inkling of the system of protein-protein interactions driving this circadian rhythm emerged. KaiC, influenced by its phosphorylation state, controls the circadian dependent transcription of genes. KaiA promotes phosphorylation and KaiB promotes dephosphorylation, while SasA provides an unclear, but stabilizing role in maintaining the overall cycle (Kitayama et al., 2003). As a result of this paper
from the Kondo lab, the Johnson lab reexamined their previous paper and
determined that the supposed KaiC degradation (Yao Xu et al., 2000), was more
likely to represent temporal changes in the phosphorylation state of KaiC (Yao
Xu et al., 2003).

As the *S. elongatus* clock became increasingly characterized, the
efficiency of this system could be compared to eukaryotic variants (Johnson,
2004). In multicellular organisms, such as rats, it is perhaps unsurprising that
clusters of cells work together to signal the time, but the clocks in these cells
individually are notably noisier than the averaged circadian signal (Karakashian &
Schweiger, 1976; Welsh et al., 1995). Such mechanisms can provide a higher
accuracy than the individual cells of a group through the process of summing the
signals of individual cells and nonlinear phenomenon, such as a threshold for
signal activation in an individual cell (Enright, 1980). *S. elongatus* might be
expected to exhibit a similar pattern, with individual bacterium showing imprecise
rhythms while local groups of cells achieve a more precise rhythm through
intercellular signals. However, using a particular strain of *S. elongatus* with
rhythmically expressed luciferase (Katayama et al., 1999), an examination of
individual bacterium compared to clusters, under both standard and nonstandard
lighting periods, showed that the precision of the *S. elongatus* circadian clock
was intrinsic to each cell (Johnson, 2004).

The functional pathway for this clock remained an area of continued
interest. An interesting feature emerged, that a particular gene promoter of *kaiBC*
was not necessary to maintain rhythmicity. Despite the role of KaiC in repression
of kaiBC transcription, changing the promoter did not affect the patterns of oscillation (Nakahira et al., 2004). However, since KaiC is part of the superfamily of bacterial DNA recombinases (Leipe, Aravind, Grishin, & Koonin, 2000), this was interpreted as KaiC modulating global transcription level through interaction with DNA, allowing for the expected mechanism of a transcription-translation feedback loop (Nakahira et al., 2004). However, an in vitro analysis of KaiC phosphorylation and dephosphorylation patterns, as influenced by the presence or absence of KaiA, as well as in vivo studies with blocked transcription, showed that this oscillation was not reliant on transcription translation feedback loops, at least for over 2 days (Tomita et al., 2005). The Kondo lab then expanded on this work by demonstrating that the KaiABC oscillator could function in vitro, without requiring daily entrainment. Combining KaiA, KaiB, and KaiC, in a 1:1:4 ratio (by weight), with ATP, allowed three measurable cycles of phosphorylation and dephosphorylation under continuous lighting conditions. Furthermore, even in vitro, the temperature tolerance remained consistent with in vivo measurements (Nakajima et al., 2005). The ability of KaiABC to maintain a circadian rhythm was eventually demonstrated to last at least 10 days in vitro (Ito et al., 2007). The resiliency of the in vitro cycling and overall stability of the in vivo system was proposed to be due in part to KaiC monomer-hexamer exchange during the dephosphorylation phase of KaiC (Kageyama et al., 2006; Mori et al., 2007).

An overall pattern of steps emerged for this oscillator. KaiA drives autophosphorylation of KaiC, initially at T432, and secondly at S431. At the halfway point, KaiB induces dephosphorylation in the same pattern of T432
followed by S431. Each step regulates the reaction of the subsequent step (Nishiwaki et al., 2007). The ATPase activity of KaiC is temperature dependent, but interactions within the ring structure of KaiC are proposed to activate or repress functionality in a way to compensate for these fluctuations, thus allowing KaiC to always ensure appropriate activity rates in the overall cycle (Murayama et al., 2011; Terauchi et al., 2007). The system of *S. elongatus* circadian rhythm shows both benefits for study and unique features compared to other model organisms. The KaiABC proteins are the basis of this system, and provide a straightforward and simple model for the study of circadian systems.

**Structural Aspects of the KaiABC System**

The capabilities and resiliency of the KaiABC circadian rhythm is a result of the functional aspects of the individual proteins as well as the protein-protein interactions of this system. In addition to the study of this system, structural analysis of the components began to clarify the physical features of the individual components. Initial structural assessment of KaiC used negative stain EM to demonstrate that KaiC formed hexameric rings, with the dual domain of a KaiC monomer giving rise to a double barrel shape around a central channel (Mori et al., 2002). Two years later, three papers provided a jump in understanding of KaiC. The Kondo lab made use of mass spectroscopy and alanine mutation to identify serine 431 and threonine 432 as two autophosphorylation sites of KaiC (Nishiwaki et al., 2004). The Egli lab provided a crystal structure of KaiC, and in addition, showing ATP binding, the location of known mutations, and phosphorylation on threonine 432. Regrettably though, the crystal structure did
not provide information on the C-terminal tails of KaiC, suggesting a greater
degree of flexibility relative to the rest of the structure (Pattanayek et al., 2004),
but a complete structure was later solved (Pattanayek et al., 2006). Further work
from Vanderbilt University, jointly from the Johnson and Egli lab, examined
potential phosphorylated residues, identifying a third critical residue, threonine
426, which is adjacent to serine 431 and may be phosphorylated, or shuttle a
single phosphate between itself and serine 431. Furthermore, mutation of this
residue adjusts circadian rhythmicity (Pattanayek et al., 2009; Yao Xu et al.,
2004, 2009). Of the three Kai proteins, only KaiC is known to undergo chemical
modifications over the course of circadian period. KaiC carries out both
autokinase and autophosphatase activities, and even transiently generates ATP
while undergoing dephosphorylation (Egli et al., 2012; Nishiwaki & Kondo, 2012).

The readout of KaiC signals proved to be an area of difficulty for analysis.
In August 2011, two papers came out, assigning contradictory binding locations
of SasA to KaiC. NMR titration experiments mixed the binding domain of SasA
with either the isolated CI or CII domain of KaiC, providing evidence that SasA
selectively binds to the CI domain of KaiC (Chang, Kuo, Tseng, & LiWang, 2011).
An alternative method, Surface Plasmon Resonance, a year later favored the
assignment of SasA binding to CI, but this too used separate domains of KaiC.
Conversely, a negative stain EM study of full length KaiC showed that SasA
competed with KaiB for binding and a SAXS study in the same paper identified
KaiB as binding to the CII domain of KaiC (Pattanayek et al., 2011), consistent
with previous results for KaiB binding (Pattanayek et al., 2008).
The study of KaiA structure and binding proved to be the least contentious of these three proteins. An initial NMR study from the LiWang lab, at Texas A&M University, identified the C terminus of KaiA as binding to KaiC and stimulating autophosphorylation (S. B. Williams et al., 2002). Continuing with this work led to an NMR structure of this KaiA domain binding to the C terminal tails of KaiC (Vakonakis & LiWang, 2004). Both of these analyses used KaiA from *Thermosynechococcus elongatus*, but the 61% sequence identity for the domain led to application of these results to *S. elongatus*. This same year, the crystal structure of KaiA was published, showing an entwined dimer with swapped domains, meaning that the N terminal domain of one chain is paired with C terminal domain of the other (Ye et al., 2004).

The structure and protein-protein interactions of KaiB proved more problematic. Work on these proteins in *Anabaena*, another cyanobacteria species, presented crystal structures of KaiA and KaiB. The sequence and structure to the binding regions of KaiA between Anabaena and *S. elongatus* are similar. Furthermore, a helix turn helix motif of KaiB is comparable to the KaiA binding interface, and the Pai lab of the University of Toronto suggested that KaiA and KaiB may compete for a binding site on KaiC, with the relative affinities modulated by the phosphorylation state of KaiC (Garces et al., 2004). Structural work on KaiB continued with two further structures of KaiB. In *T. elongatus* (Iwase et al., 2005) and *Synechocystis* (Hitomi et al., 2005), the crystal structures of KaiB both showed a pair of dimers, although both were limited in not resolving the C terminal tails of any of the subunits. Both papers suggested that positively
charged residues may play a role in binding, but Hitomi et al. further suggested that the tetramer of KaiB may play a role in binding KaiC (Hitomi et al., 2005). Based on the known information about KaiA binding the CII domain of KaiC, and further work using fluorescence anisotropy, the Liwang lab proposed that KaiB induces sequestration of KaiA by hindering conformational changes in the CII domain of KaiC at its interface with KaiA (Kim et al., 2008).

Two papers in early 2008 provide an initial baseline for assessing KaiB-KaiC interactions. A SAXS analysis of the complexes at various points in the cycle, and in comparison to the envelopes of the individual proteins led to the conclusion of KaiB binding as a pair of dimers or a tetramer to KaiC to prevent further KaiA induced autokinase activity (Akiyama et al., 2008). A collaborative paper from the Johnson, Stewart, and Egli labs a month later provided a host of information about the structure and interactions of KaiB and KaiC. A T. elongatus KaiB crystal structure, which included the C terminal tails, showed the tails as adopting multiple conformations, including extended from the protein and folded up against it. Examination of individually expressed domains of KaiC by negative stain EM noted that the CI domain retained the ability to form hexameric rings, while the CII domain did not, suggesting that the stable formation of the KaiC hexamer was dependent on the CI domain. Furthermore, KaiB solely interacts with the CII domain, when assayed by native PAGE gel with the isolated domains. However KaiB showed no affinity for the C terminal tails of KaiC, in contrast to KaiA, thus reducing the likelihood of KaiB directly competing with KaiA for a binding site. Examining the full proteins by negative stain EM showed
a triple layer structure, with a discernable double barrel shape for KaiC, and a third layer of density attributed to KaiB. While the resolution did not permit assignment of individual KaiB monomers, the volume of the ring was considered appropriate for two KaiB dimers binding to the top of KaiC, on opposite sides of the central channel. The link between the roles of KaiB and KaiA are predicted to be reliant upon conformational shifts in the ATP binding clefts of KaiC, induced by KaiB, and thus negating the effects of KaiA (Pattanayek et al., 2008).

Further work by these labs used SAXS as an alternative assessment to further evaluate their prior claims. Collecting SAXS envelopes for the individual proteins showed structures as expected, with KaiB showing a dimer and KaiC showing an extension for the C terminal tails. In agreement with their previously published work, they showed KaiAC and KaiBC envelopes. Notably, the KaiBC envelope did not show mass on both sides of the KaiC, in agreement with KaiB binding on the same domain as the KaiC C terminal tails. In addition, negative stain EM of the KaiABC complex provided information about how KaiB and KaiA simultaneously bind to KaiC. A triple layer structure is observed with an extension from one side, which was interpreted as KaiB shifting the KaiA dimer away from the central channel of KaiC. While the 3D density map showed a single protrusion, examination of the collected micrographs showed multiple KaiA dimers alongside KaiC when the concentration of KaiA was increased (Pattanayek et al., 2011).

The crystal structure of KaiB shows a tetramer, while structural data indicated that KaiB binds KaiC as a dimer or monomer. The Ishiura lab examined
potential effects for the tetramer oligomer in maintaining the circadian oscillations. They found that a deletion of the negatively charged C terminal residues of KaiB, which were known to weaken in vivo circadian rhythms (Iwase et al., 2005), also prevented formation of the KaiB tetramer. While in vitro experiments demonstrated that the core KaiABC oscillator functioned normally with the modified form of KaiB, the mutation in vivo produced a weakened circadian cycle. Additionally this modified system showed reduced circadian linked gene expression, suggesting that understanding the role of KaiB may require assessing multiple oligomers of KaiB in their interaction with KaiC and other proteins (Murakami et al., 2012).

Based on their previously published work supporting SasA binding to the CI domain of KaiC (Chang et al., 2011), the LiWang lab questioned how KaiB could compete for binding with SasA, if KaiB bound the CII domain of KaiC. Furthermore, they noted differences in the published literature as to which KaiC domain provided the binding surface for SasA. To further evaluate these studies, they expressed the KaiC domain individually, with a S431E mutation on the CI domain, such that both individual domains could form hexamers. When either domain was mixed with KaiB, and examined by NMR, only the CI domain experienced a change in spectra. Furthermore, they found that KaiB and only the CI domain would form complexes in gel filtration chromatography. On the basis of this work, they concluded that the difficulty in interpreting the SAXS and EM structures due to the similarity of these domains had led to a mistaken assignment of the KaiB binding location.
In response to these concerns, the Egli lab sought images via negative stain EM that could more conclusively visualize the domain of KaiC providing the binding surface for KaiB. In order to distinguish between the two domains of KaiC, they attached a gold nanoparticle to the CII domain. The additional density was readily observable, and allowed comparison between images of KaiC and KaiBC. The resulting image analysis supported KaiB binding to the CII domain of KaiC (Pattanayek et al., 2013).

Further work from the Johnson, Egli, and Stewart labs provided more information. In addition to a complete crystal structure of *S. elongatus* KaiB, a cryoEM structure presented another possible binding pattern for KaiB and KaiC. At higher resolution, the third ring of density appeared continuous around the top of KaiC. Using molecular simulations of the individual crystal structures and the density map of the complex allowed prediction of the binding surface between the two proteins. The most favorable binding surface was created by 6 KaiB monomers on the CII domain of KaiC (Villarreal et al., 2013).

Alternative methods were used by the Heck lab to further assess this debate. The use of tandem mass spectroscopy allowed a stoichiometric measurement of the complexes formed during *in vitro* cycling, while hydrogen-deuterium exchange MS (HDX-MS) provided information about the residues involves in KaiB binding KaiC. The results of this work led to claim that KaiB both interacts with KaiC as a monomer, and that the binding of KaiB to KaiC is cooperative, such that the binding of an initial monomer of KaiB leads directly a total binding of six monomers. Furthermore, the HDX-MS supported KaiB binding
to the CII domain of KaiC (Snijder et al., 2014). While this work may not settle all debate, the basis of scientific understanding is characterization, and further information will help to eventually understand this circadian system.

**Summary**

The complexity of the cyanobacterial oscillator provides a rich area of research, and an underlying model organism that is much easier to work with compared to multicellular organisms. Every circadian oscillator provides an internal signal to an organism, so as to allow metabolism regulation in the transient absence of external signals. The majority of multicellular organisms, including humans, rely on a master clock based on a transcription-translation feedback loop. Cyanobacteria possess a similar mechanism, but do not rely on it. Humans have circadian clocks in a majority of tissues, which are still undergoing characterization. Two eukaryotic organisms were recently shown to possibly possess circadian rhythm in the absence of transcription. Both the alga *Ostreococcus tauri* and humans, via red bloods cells that naturally lack transcription, were recently shown to possess metabolic oscillations (O’Neill & Reddy, 2011; O’Neill et al., 2011). The study of known systems such as *S. elongatus* provides a model for the study of more complex feedback loops and a comparison to any that may yet be found.
Appendix 2: Overview of DNA Damage and Repair

Genetic Information Within an Organism

DNA is the main method of information storage within an organism. From mitosis of prokaryotes to gestation of embryos in mammals, DNA is the primary medium by which coding and regulatory information is passed on to subsequent generations. During the lifespan of an organism, transcription of RNA from DNA allows translation of proteins and regulatory RNAs. Lastly, protein interaction with DNA also provides a source of regulatory information. Without DNA, a cell is incapable of survival. DNA lesions can occur in a variety of ways, and represents a threat to the continued survival of a cell, necessitating a repair response for reliably recovering from damage. In many cases, DNA exists as only a single copy within a cell, and in many cases the sequence of DNA required for survival is a unique molecule within a cell. Repair of DNA lesions must be able to occur without access to the information contained within any particular section of DNA.

Friedrich Miescher discovered what would be called DNA in 1868, while working at the University of Tübingen, Germany. He identified the substance as neither protein nor lipid, and noted an unusually high proportion of phosphorous in the compound. Due to its location in the nucleus, he called it nuclein, and hypothesized that it was critical to the function of the nucleus, noting an increase during cell division (Dahm, 2008). While suggesting the importance of DNA, Friedrich Miescher discarded the possibility of his nuclein transmitting traits across generations, and this discovery waited until 1944, at which time Avery et al. provided evidence that DNA carried this information, rather than the proteins
of a cell (Avery, Macleod, & McCarty, 1944). This concept was later confirmed using bacterial phages, by distinguishing between the role of protein and DNA in infection (Hershey & Chase, 1952). Research on DNA noted the fixed ratios between the bases of adenine (A) and thymine (T), as well as between cytosine (C) and guanine (G) (Chargaff, Vischer, Doniger, Green, & Misani, 1949; Chargaff, 1951). Shortly thereafter, the famous paper from Watson and Crick provided the structure of DNA as a double helix with phosphate-sugar backbones spiraling around A, T, C, and G bases (Watson & Crick, 1953b). Furthermore, they proposed the sequence of bases as the repository of genetic information and that the pairing between bases as the method of information transfer during cell division, such that each single strand recreates the complementary one (Watson & Crick, 1953a). This work formed the core of what was described by Francis Crick as the Central Dogma, DNA to RNA to protein, and the basis for genetic inheritance (Crick, 1955).

Health Issues Arising from DNA Lesions

The direct relationship between genetic information and the development of cancer makes DNA lesions an inherent concern (Bartek et al., 2007). Ideally, lesions are repaired without loss of genetic information, and the cell continues fulfilling its particular role. However, the DNA damage response is not perfect, and genetic abnormalities accumulate. Over time, a cell may no longer be able to function correctly, and there are various processes to allow an organism to manage these issues (Harper & Elledge, 2007). Broadly speaking, humans maintain populations of dividing cells to replace such losses. An individual cell
therefore ought to choose between three choices: survival, replicative senescence, and death (Campisi, 2005; Hoeijmakers, 2001). While survival is the most useful outcome, when DNA repair completes successfully, that outcome may be the most dangerous over time as the number of DNA damage response events increases and the number of repair errors accumulate. Due to the imperfect nature of DNA repair a repaired double strand of DNA does not necessarily match the original sequence prior to being damaged. A cell can accumulate damage over time, and eventually this may even trigger uncontrolled proliferation and cancer. One difficult aspect of examining cancer is the number of possible variations in the genome. The p53 gene is the most commonly mutated gene in human cancer, but a number of mutations can affect disease progression, and the tendency towards ongoing mutation, make examination of the causes of any particular instance difficult (Giglia-Mari & Sarasin, 2003; Hussain & Harris, 2000).

The DNA damage response encompasses a variety of pathways, and plays a role in both inducing cancer and preventing it. Oncogenes and tumor suppressors describe pro- and anti- cancer genes respectively, but their role in disease is a modification of normal function. Oncogenes promote cell division and growth, leading to increasing numbers of cells. As mutations constitutively activate oncogenes, the daughter cells also possess these mutations and offer further likelihood for any individual cell within this population to undergo additional mutations (Albertson et al., 2003). These genes however do play important roles in normal metabolism, allowing wound healing and cell
replacement. Conversely, tumor suppressors counter these effects, limiting excessive cell growth, and even triggering cell death to prevent excess growth (Deshpande et al., 2005; Rackley et al., 1993). Cancer results from changes in DNA, but due to the variety of possible changes, it can be described by characteristic aspects of the carcinogenic cells (Hanahan & Weinberg, 2000, 2011). All of these characteristics involve mutation in DNA: sustaining proliferative signaling (Lemmon & Schlessinger, 2010), evading growth suppressors (Sherr & McCormick, 2002), activating invasion and metastasis, enabling replicative immortality (Artandi & DePinho, 2010), resisting cell death (Willis & Adams, 2005), and inducing angiogenesis (Raica et al., 2009). These functions are coopted from normal cellular activities, an adjustment of function that can be triggered by mutation. In the case of tumor suppressors, a loss of an activating site via mutation may contribute to cancer. Alternatively, in the case of an oncogene, an activating phosphorylation site may become constitutively active via mutation, such as a phosphomimetic residue, which also then might contribute to cancer. Failure to maintain genomic integrity during the DNA damage response can introduce mutations that push a cell closer to becoming cancerous in a variety of ways (Soh et al., 2009).

Although cancer is widely recognized, there are a number of diseases and disorders that arise from defects in the cellular DNA damage response. Many of these problems are genetic and inherited. Studying the consequences of genetic defects in the DNA repair process allows a more stable look into the
mechanisms of this process facilitating the development of more precise therapeutic treatments (Wijnhoven et al., 2005).

Hypersensitivity to UV radiation can be caused by a recessive disorder in one of eight genes (XPA through G and polymerase η) required to repair the DNA lesion created UV radiation (Lehmann, McGibbon, & Stefanini, 2011). Termed Xeroderma pigmentosa, this relatively rare disease frequently produces symptoms of severe sunburn, prolonged recovery from sunburn, and increased freckles (Hirai et al., 2006; Kleijer et al., 2008). While freckles are a common trait, in this case the increased number are formed through numerous individual hyperplasia of melanocytes, and are a failure of the melanocytes to appropriately proliferate. Furthermore, an individual with Xeroderma pigmentosa is more likely to develop skin cancers. The condition can be managed in part by avoiding exposure to the sun, but instances of cancer can still occur and require prompt treatment due to their highly proliferative nature (Bradford et al., 2011).

Even simply prolonged durations of lesions, despite an otherwise functioning DNA damage response, can lead to severe consequences. Individuals with Cockayne syndrome have never had a reported case of cancer. The tradeoff is harsh though, as individuals cells are slow to repair and more frequently trigger cell death responses (Hoeijmakers, 2009). The overall functioning of the DNA repair pathways are insufficient, and an individual undergoes progeria (accelerated aging) over a shortened life expectancy of less than two decades. Furthermore, neuronal growth fails shortly after birth, leading to varying degrees of mental retardation (Rapin et al., 2006). In contrast to
cancer, in which the accumulation of DNA repair events that cause mutagenesis through failure to correctly replace the original genetic information, Cockayne syndrome slows down the DNA damage response and leads to the death of a cell (Andressoo & Hoeijmakers, 2005).

In some cases, a particular gene is critical and loss of function can lead to negative effects throughout multiple pathways. This is the case in Ataxia telangiectasia, involving mutation of the gene for the ATM (Ataxia telangiectasia mutated) protein. Patients with this disease have increased sensitivity to ionizing radiation, chromosomal instability, and a predisposition to cancer (Rotman & Shiloh, 1998). Mouse studies have clarified broader aspects of the role of ATM. ATM deficient cells divide at a slow rate, and after a relatively small number of divisions undergo growth arrest, reaching a state similar to senescence (Barlow et al., 1996; Yang Xu & Baltimore, 1996). The DNA of these cells is rather different in quality, exhibiting a large number of chromosomal breaks (Elson et al., 1996). Furthermore, such mice are infertile (Yang Xu et al., 1996) and possess an abnormal immune system with reduced numbers of lymphocytes, possibly due to fewer successful instances of V(D)J recombination within the G0/G1 phase of cell growth during the maturation of B and T cells in the immune system (Lin & Desiderio, 1995). ATM mutations have been described as blocking the caretaker function of this protein (Kinzler & Vogelstein, 1997), and the central role of ATM in the DNA damage response has provided numerous avenues for research into the maintenance of DNA (Nowak, 1995).
Types of Damage Affecting the Integrity of DNA

Cells are susceptible to harm through a variety of insults, both external and internal. Even DNA itself is not inherently stable, slowly breaking down under physiological conditions (Lindahl, 1993). Both cell permeable chemicals and photons of sufficient energy are capable of inducing DNA breaks or modifications, necessitating repair of damaged bases or lesions in the DNA backbone (Hughes & Reynolds, 2005; Martin, 2007; Wilson, Bohr, & McKinnon, 2008). Additionally, cellular processes can be inherently hostile to DNA. For example, the metabolic processes of a cell inherently produce reactive oxygen and nitrogen species that are nonselective in their targets, allowing for the potential of self-induced DNA damage (Harman, 1956; Kirkwood, 2005).

Unintentional chemical changes can occur frequently, and for DNA, the consequences can be serious. Over the course of a day, an average human cell experiences up to an estimated 10,000 instances of DNA lesions (Lindahl, 1993; Sander et al., 2005). While RNA and proteins are critical to the survival of a cell, in a majority of cases any single RNA sequence or protein can be degraded and replaced when it is damaged or malfunctioning. DNA must be maintained within a cell, and DNA damage repaired as necessary. In most cases, the cellular DNA damage response is successful and a cell continues in its role within the body. However, there are other alternatives. Programmed senescence describes the accumulation of age-related cellular damage, including DNA lesions that trigger a cell to cease replicating. A senescent cell does not divide into two daughter cells and therefore avoids increasing the population of cells with those errors. A
decisive response to irreparable DNA damage is cell death, which prevents the corruption of genetic information from causing any further damage to the organism (Hoeijmakers, 2009).

Ideally, DNA lesions can be repaired, and there are a variety of mechanisms in a cell that carry out such repairs. Broadly speaking, in order to restore DNA, the type of damage must be correctly identified and then the DNA must undergo a chemical reaction to recreate the appropriate structure. Shown in Fig. 1 are general categories of the structural changes found in DNA lesions. Three types of lesions involve only the bases of DNA, and are termed base adducts. A missing base leaving a gap, called an abasic site, results from destabilized bond between the base and the backbone. Base pair mismatches (commonly A/C or G/T pairings) occurs as a result of damaged bases being replicated by the cell. One example is cytosine deamination, which converts a cysteine to a uracil, leading to a synthesis of a complementary A instead of a G. The third type of base lesion is a chemical modification to a base, such as UV induced linkage between thymine dimers or guanine methylation. In comparison between a base pair mismatch and a modified base, the first may be chemically normal, despite the inaccuracy of the genetic information, while a modified base will be chemically distinct from standard DNA. These three types of base adducts can be repaired from the complimentary chain of DNA, which possesses the appropriate base to ensure information is not lost.

A separate type of damage is interstrand crosslinks. Each DNA strand is a distinct and non-covalently linked molecule, and their transient separation is an
essential part of replication and transcription (Deans & West, 2011). A covalent linkage between the two strands can limit the requisite flexibility. This particular type of damage is frequently the result of lipid peroxidation (Kozekov et al., 2003; Stone et al., 2008), and was recognized as increasing the odds of cancer, in particular leukemias, as a result of chemical warfare agents (Pechura & Rall, 1993). In addition, more mundane factors such as a high fat diet and alcoholism can lead to interstrand crosslinks (Brooks & Theruvathu, 2005; Folmer et al., 2003).

The two remaining categories of DNA lesions are single and double strand breaks (SSB and DSB) in the backbone of DNA (Iyama & Wilson, 2013). In the first case, a single backbone of a DNA double strand is broken. In many ways, this is analogous to normal cellular processes in which a single strand is snipped by topoisomerase I to allow release of torsion stress during supercoiling of DNA (Moukharskaya & Verschraegen, 2012). However, a random break is dangerous because it is uncontrolled (Sander et al., 2005), and can allow for base degradation at the site of the break (Thompson & West, 2000). A more dangerous, but rarer, form of damage is a double strand break (Bohgaki et al., 2010). Unlike the ends of chromosomes, which are capped by protective proteins bound to a telomere (Rooney et al., 2003), an unprotected end of DNA is subject to degradation, and if this lesion occurs outside of cell division and without a sister chromatid the loss of genetic information is likely irrecoverable (Khanna & Jackson, 2001). In both single and double strand break lesions, the damaged backbone of DNA presents an immediate issue in preserving the integrity of the overall strands of DNA.
Maintaining DNA

Despite the frequency and dangers to an organism, in most cases the DNA repair response of a cell is sufficient to recover all of the genomic information. The incidence of daily lesions would mean that even a 1% error rate would quickly render cells cancerous or trigger cell death pathways. There are a variety of mechanisms used by the DNA damage response pathways, and the method used depends both on the type of damage and the cellular process recognizing the lesion. In some cases, the simplest response is to ignore the damage. A damaged base within an intron need not affect a final protein sequence, but transcription stalling due to damage would limit the population of the protein. In that manner, there are several polymerases capable of bypassing base adduct lesions (Andersen et al., 2008). While continued transcription is critical, the tradeoff of ignoring damage is a part of the noted accumulation of mutation in somatic tissue and the associated connection with aging (Vijg, Busuttil, Bahar, & Dollé, 2005).

Base adducts are targeted by base-excision repair (BER), in which first the damaged base is removed, creating an abasic site. The remaining portion of the nucleotide can be cut away from the DNA strand, and the resulting gap filled in by DNA synthesis with pol β (Barnes & Lindahl, 2004; Caldecott, 2008; Slupphaug et al., 2003). Larger forms of damage that distort the overall shape of a DNA helix fall under the category of nucleotide excision repair (NER) which functions mechanically similar to BER, cutting away the entire nucleotide and then also using a polymerase (δ, ε and/or κ) to fill in the resulting gap (Gillet &
Schärer, 2006; Sugasawa, 2006). There are two variants of NER, global genome NER (GG-NER) and transcription coupled NER (TC-NER). These two pathways differ in their priority of activation and importance of repair. Large portions of the genome are transcriptionally inactive, allowing GG-NER to perform continuously, steadily checking for lesions. Initialization of the repair process includes the gene products that are mutated in Xeroderma pigmentosa (Gillet & Schärer, 2006; Sugasawa, 2006). The second option, TC-NER is triggered when RNA polymerases are stalled during transcription. Mutations in the signaling genes involved in this process can give rise to Cockayne syndrome, for which the CSA and CSB genes are named (Fousteri & Mullenders, 2008; Hanawalt, 2002). While GG-NER is aimed towards the integrity of the overall genome, TC-NER ensure that the actively used portions are repaired quickly. Both BER and NER pathways are involved in transcription, as well as being capable of locating and repairing DNA lesions. In total, they deal with the majority of oxidative damage and bulky lesions occurring upon DNA.

Uncontrolled DNA breaks can affect either one or both strands of DNA. Such breaks occur as even part of the other portions of the DNA damage response, such as when excising a chemically modified nucleotide. Through the complementarity of DNA, the repair of SSBs generally results in the complete recovery of genetic information. If necessary, any damaged nucleotides can be excised through other pathways, and then DNA synthesis replaces missing nucleotides, allowing the remaining nick in the backbone of DNA to be repaired by DNA Ligase I (Hegde et al., 2008). Conversely, DSBs create two unprotected
of DNA that are unconnected within the cell. During the G1 growth phase, there is no sister chromatid, but homologous regions within the genome may be present. In comparison, the G2 phase allows use of a duplicated chain of DNA as a model to build a complete duplicate of the original sequence. The S-phase and M-phase of the cell cycle present indeterminate states between the presence and absence of sister chromatids for each chromosome, as the genome is either in the process of replication or separation into separate cells. For any individual break, a homologous section of DNA may or may not be available for reference. Homologous recombination (HR) describes the repair process using the sister chromatid as a template, while nonhomologous end joining (NHEJ) provides a repair pathway for DNA independent of a template. Finally, there is third option, microhomology-mediated end joining (MMEJ), which is more recently recognized and uses small regions of similar DNA, but not sister chromatids, to recreate degraded end of DNA during a double strand break (Symington & Gautier, 2011).

The process of HR for restoring a DSB actually begins with resection of the 5’ to 3’ strands of DNA at each broken end (Krogh & Symington, 2004). The first section of HR repair is finding the homologous sequence of DNA, and the now single strand DNA (ssDNA), while bound to Rad51 and BRCA 1 & 2, can invade and pair with the template strand (West, 2003). This pairing occurs for both ends of DNA at the DSB, allowing for DNA polymerase to extend both strand of the damaged DNA (Pâques & Haber, 1999). The intertwined strands of the damaged DNA and the template DNA is called a Holliday Junction, and allows the complete synthesis of missing DNA from the template (Constantinou,
Davies, & West, 2001; Whitby & Dixon, 1997). The recovery of genetic information does not preclude large scale changes in the genome. The resolution of the Holliday Junctions is imperfect, and can lead to crossovers or recombination between the repaired and template DNA strands (Boddy et al., 2001; X.-B. Chen et al., 2001; Shrivastav, De Haro, & Nickoloff, 2008).

MMEJ was recognized after HR and NHEJ, during studies of transformation of plasmids into yeast that lacked effective NHEJ activity (Boulton & Jackson, 1996; Schär, Herrmann, Daly, & Lindahl, 1997). Initial characterizations of the role of this pathway noted the overlap between some of the processes of NHEJ, but that this alternative pathway was an order of magnitude kinetically slower. Attempts to classify this pathway suggested that it may not be relevant under normal biological conditions, such that it is a rarely used backup system (Lieber, Ma, Pannicke, & Schwarz, 2004; H. Wang et al., 2003). Complications have arisen in separating out the components in this pathway, leading to suggestions that there may be multiple uncharacterized pathways for DNA DSBs that operate independently of HR and NHEJ (Bennardo, Cheng, Huang, & Stark, 2008). A proposed pathway for MMEJ suggests that restriction of the broken DNA ends occurs in a similar fashion to HR (Decottignies, 2007; Lee & Lee, 2007). A similar, but not necessarily identical sequence, of approximately 8-22 bases provides the template to synthesize the new DNA (Daley & Wilson, 2005), and the repaired strands of DNA are ligated together with DNA ligases I and II (Liang et al., 2008). Through the differences permitted in the microhomology template compared to the original sequence,
MMEJ is associated with a higher rate of mutation, but offers an alternative to the other repair pathways to help maintain the overall requirements of genomic integrity (McVey & Lee, 2008).

The remaining major repair pathway is NHEJ, and this repair process does not use a template to recover the lost genetic information while repairing DNA and is the main pathway of DSB repair in vertebrates (Lieber, 2010). The Ku70/80 heterodimer (Ku), which was identified in 1981 (Mimori et al., 1981), binds a broken end of a DSB. This complex then activates the DNA-dependent Protein Kinase catalytic subunit (DNA-PKcs), which is the main protein that activates and controls the remainder of the NHEJ process (Dynan & Yoo, 1998). Due to the variety of possible DNA lesions, the Artemis endonuclease is involved in any necessary nucleolytic processing to provide appropriately clean ends of DNA for ligation (Ma, Pannicke, Schwarz, & Lieber, 2002). The loss of genetic information is an undesirable outcome, and so while nucleolytic activity is useful, overhangs left behind after removal of damaged bases offer the opportunity to avoid further loss by synthesis using the overhang as a template. Family X polymerases (including polymerases μ and λ, as well as terminal deoxynucleotidyl transferase) fill in these gaps to leave blunt ended strands of DNA (Yamtich & Sweasy, 2010). The repair process is completed by ligated the ends of DNA together with DNA Ligase 4 (Ogiwara & Kohno, 2011).

Variability in the Immune System

In addition to DNA repair, DNA-PKcs and other components of the NHEJ pathway play an important functional role in ensuring variability within the
adaptive immune system. In humans, and numerous other vertebrates, both the innate and the adaptive immune system serve as a block to bacterial and viral diseases. The innate immune system provides immediate protection by recognizing common traits of diseases, while the adaptive immune system relies on a random generation of antibodies to create ligands for unknown pathogens. The overall process is to locate non-host proteins and use them as ligand targets for randomly generated antibodies. If an appropriate antibody is found, then its production is increased and maintained as the host response to the pathogen. A critical aspect of the adaptive immune system is that the pathogen is already active in the host during the antibody generation process, and presumably the innate immune system was not able to appropriately prevent or remove the pathogen. The adaptive immune system must first identify whatever is bypassing the innate immune system, and then appropriately respond. There are several disorders associated with this process, because the immune system is intended to locate and respond to unknown proteins, and these targets are a small fraction within the overall population of the host proteins. Any randomly created antibody is more likely to bind to a host protein, leading to an autoimmune response, and therefore developing an appropriate antibody is a delicate process (Peng, Yang, & Racke, 2014).

There are numerous cell types involved in the adaptive immune system, and their critical roles are based on the generation and use of antibodies. Succinctly, through random uptake of cellular debris, an antigen is acquired and presented to Helper T cells. If activated, the Helper T cells stimulate activation of
B cells and Cytotoxic T cells. If a B cell recognizes an antigen in the blood, as well as being stimulated by a Helper T cell, then the B cell replicates and begins producing antibodies. The released antibodies are distributed throughout the body and bind to antigen, which can include free virus or infected cells displaying viral proteins. The bound antibodies are a signal to destroy the antibody and the bound antigen, commonly by engulfment from phagocytic cells. The Cytotoxic T cells react based on stimulus from Helper T cells, but are not greatly involved with B cells (Peng et al., 2014).

The generation of antibody and T cell receptor diversity relies on NHEJ. Containing all the possible number of unique genes within the cellular genome is not a reasonable option, and instead the necessary diversity to reliably develop an adaptive immune response relies on recombination of three distinct sections of the immunoglobulin gene. The variable (V), diversity (D), and joining (J) regions each contain a number of segments that can be combined to provide a much larger degree of diversity compared to full genes of equivalent size within those regions (Alt, Zhang, Meng, Guo, & Schwer, 2013). Each developing immune cell produces a single unique recombined sequence. The process of creating this degree of diversity relies on NHEJ components and pathways (Reth & Nielsen, 2014). Unlike damage induced DSBs, the DNA segments in V(D)J recombination undergo programmed DSBs. This process involves joining a segment from each section into a continuous piece and removal of intervening segments. Each joining event is cleaved by the RAG1/2 recombinases, leading to two blunt ends of DNA from the discarded segments and two hairpins. The
blunt ends are on the segments cut from the genome and rejoined to facilitate disposal. Artemis is used to open the hairpins, and allow random sequence changes at the opening via the terminal nucleotide transferase protein. Finally, the ends of DNA are ligated together through the NHEJ pathway, leading to a continuous sequence (Bednarski & Sleckman, 2012a, 2012b; Ma et al., 2002).

A variety of disorders are related to defects in the ability of immune cells to undergo V(D)J recombination, and a subset of these are due to problems with the components in NHEJ pathway. Severe Combined Immunodeficiency (SCID) results in a failure to appropriately generate B and T cells. This can range from a reduction through to a complete absence of mature cells of these types. This genetic disorder does affect both DSB repair and the immune system. However, there are multiple pathways for DSB repair, and while a lack of NHEJ leads to increased radiosensitivity, patients can still compensate through the remaining DSB repair pathways. In contrast, V(D)J recombination relies on the NHEJ pathway, and partial or complete loss of function mutations can render the adaptive immune system inoperative (Dvorak & Cowan, 2010).

Components of the NHEJ Pathway

Ku initiates the NHEJ pathway, and plays a critical role as a key detection mechanism for DSB events. Forming a ring, the Ku heterodimer can bind DNA, sliding over the broken end of DNA within seconds of a break occurring (Britton et al., 2013; Downs & Jackson, 2004). Breaks in DNA are not sequence specific, they can occur in any location. Ku compensates for this by interacting with the backbone of DNA, rather than the bases, allowing the heterodimer a consistent
binding surface (J. R. Walker et al., 2001). DSB breaks leave exposed ends of DNA, which unlike the telomere ends of chromosomes, are not protected by a shelterin complex and are susceptible to degradation. Ku blocks various end processing enzymes, including exonuclease I, thereby protecting the resulting ends of a DSB (Sun et al., 2012)

The majority of the remaining components of the NHEJ pathway directly interact with Ku, which begins with the recruitment of DNA-PKcs (Uematsu et al., 2007), and includes the ligase factors XRCC4 and ligase IV (Costantini et al., 2007; McElhinny et al., 2000). Additionally, Ku also participates in the actions of the end processing factors to remove damaged nucleotides (Davis & Chen, 2013). In contrast to a sequential addition of factors, the NHEJ complex appears to assemble with Ku and DNA-PKcs as a scaffold for the collective stabilization of the remaining components in a dynamic and concerted process (Cottarel et al., 2013; Yano et al., 2009).

DNA-PKcs is recruited shortly after Ku to the site of a DSB, and these three components form a complex referred to as DNA-PK, which is the basis for the remaining portions of the NHEJ pathway (Gottlieb & Jackson, 1993). DNA-PKcs is a member of the phosphatidylinositol-3 (PI-3) kinase-like kinase family (PIKK), and contains both a PI-3 kinase domain, as well as numerous HEAT (Huntington-elongation-A-subunit-TOR) repeats (Hartley et al., 1995; Perry & Kleckner, 2003). Structural studies have revealed a structure with two broadly discernable parts. The C terminus contains the kinase domain, and forms a ‘crown’ atop the remaining portion of the protein, with the HEAT repeats forming
two arms in a pincer shape. Both an x-ray crystal structure and cryoEM structure agree on the position and shape of the kinase domain, but the cryoEM structure, from the Stewart lab, shows a much larger volume for the base. This is likely due to the crystal structure being at a resolution only sufficient to identify 46% of the backbone residues, which in the N terminus, were mainly the HEAT repeats in the arms of the structure (Sibanda et al., 2010; D. R. Williams et al., 2008). Due to the complexity of the NHEJ pathway, and the requirements to deal with numerous types of DNA lesions, the differences between the x-ray and cryoEM structures maybe represent areas of flexibility or other conformational changes. The interactions between the components of DNA-PK proved difficult to characterize. Eventually, research showed that DNA-PKcs can bind DNA independently of Ku, but the C terminus of Ku enhances this interaction (Gell & Jackson, 1999; Weterings et al., 2009). Furthermore, both broken ends of the DSB must be brought together as part of the repair process, and a DNA-PK complex on each end may come together to form a synaptic complex (DeFazio, Stansel, Griffith, & Chu, 2002; Weterings & Van Gent, 2004).

The initial formation of DNA-PK results in translocation of Ku along the length of DNA, allowing activation of DNA-PKcs kinase activity (Yoo & Dynan, 1999), as compared to the unbound state in which DNA-PKcs has limited or no kinase activity (Hammarsten & Chu, 1998). Interestingly, the kinase domain is affected by the N terminus, and that deletion of that section of the protein results in constitutive activity (Davis et al., 2013; Meek et al., 2012). A functional kinase domain of DNA-PKcs is essential for proper function during the NHEJ process,
without that capability both DSB repair and V(D)J recombination fail (Kurimasa et al., 1999). Surprisingly, while DNA-PKcs can phosphorylate in vitro the canonical NHEJ components, including Ku (D. W. Chan et al., 1999), DNA ligase IV (Y.-G. Wang et al., 2004), and Artemis (Mahaney et al., 2009; C. Wang & Lees-Miller, 2013), this phosphorylation activity is not required for NHEJ. Part of this may be explained through redundancy via ATM, which is in the same protein family as DNA-PKcs and may be capable of supporting defects in DNA-PKcs function. Still, these proteins are not interchangeable, and continuing assessment of the NHEJ pathway suggests that unique DNA-PKcs phosphorylations may be relevant to efficient DSB repair, including the Werner syndrome protein (Kusumoto-Matsuo et al., 2014). DNA-PKcs autophosphorylates, resulting in inactivation and dissociation in vitro, with at least 40 phosphorylation sites throughout the protein with functional roles that remain poorly understood (Dobbs et al., 2010; Douglas et al., 2007; Merkle et al., 2002; Olsen et al., 2010). Particular phosphorylation sites are better understood, such as Ser2056, an autophosphorylation site which is recognized as improving the efficiency of the NHEJ process (Cui et al., 2005). In comparison, Thr2609 is phosphorylated by ATM and loss of this site by alanine substitution sensitizes cells to replication stress, and this mutation leads to an early death in mice (Uematsu et al., 2007; Yajima et al., 2006; S. Zhang et al., 2011). A functional role for Thr2609 is in connection to protein-protein interactions, leading to recruitment of Artemis (Goodarzi et al., 2006). These two phosphorylation sites may be in opposition for function, with phosphorylated Ser2656 limiting DNA processing, and phosphorylated Thr2609 promoting it (Cui
et al., 2005; Meek et al., 2007). Overactive processing could lead to excessive loss of genetic information, but the inability to remove damaged bases is a problem of the opposite extreme. Maintaining the balance between ensuring the removal of damaged bases and minimizing the loss of genetic information is a delicate process, one in which the variety of activity affecting phosphorylation sites on DNA-PKcs may allow control over this subtle process.

Nucleolytic activity is required for DSB in which there is chemical modification to one or both of the broken ends. This activity is required in many cases for double strand breaks, and nucleases, both Artemis and others, play an essential role in the repair of DNA. However, the number of possible proteins involved and the types of potential damage complicates the activities and roles of these proteins. While a number of enzymes are recruited during NHEJ, not all are likely necessary, and one proposal for the selection of the appropriate repair factors is a guess and check process, in which a randomly selected processing enzyme is provided access to the damaged end of DNA, and then after the repair attempt, the end of DNA can either be ligated together or are passed back to the start of the prior step to again attempt to remove damage bases (Strande et al., 2012). A more procedural method suggests that Ku and DNA-PKcs may recruit additional proteins based on the complexity of the DNA lesions, with support for this idea coming from noted changes in the dynamics of the NHEJ complex formation being affected by the type of induced DNA damage (Reynolds et al., 2012).
Conclusions

The maintenance of DNA within an organism allows the reliable continuation and propagation of genes in spite of numerous types of damage that would create mutations or simply degrade the DNA. In humans, the DNA damage response is a complex process in which at least three different pathways, homologous recombination, microhomology-mediated end joining, and nonhomologous end joining, repair the wide variety of chemical modifications that can occur on DNA. This last pathway, NHEJ, also plays a critical role in the adaptive immune system, in V(D)J recombination. Shared proteins and functions between NHEJ and the other two repair pathways have made understanding these systems a complex undertaking. These efforts remain important though, as a malfunction within a DNA repair pathways can lead to a variety of health disorders and a predisposition to cancer. The genetic underpinnings of such issues make treatments difficult, and are usually restricted to mitigation of symptoms and limiting the consequences on the disease. In comparison to the other pathways in the DNA damage response, NHEJ in particular is critical to human health because of its role in the adaptive immune system. Understanding the components of NHEJ and how they fulfill the necessary roles for this pathway is a critical part of developing better treatment options and therapeutics.
Appendix 3: KaiBC Symmetry Tests

During refinement of the KaiBC single particle dataset, a visual inspection of an asymmetric structure midway through the refinement process showed pseudo six fold symmetry. In addition to this assessment, an analysis via rotational correlation was used to examine this structure more rigorously. The published KaiC crystal structure shows pseudo six fold symmetry, and the KaiC portion of the KaiBC cryoEM structure presented in chapter 2 displays pseudo six fold symmetry even without imposing symmetry. However the KaiB crystal structure shows a pair of dimers. The question arose whether or not it was appropriate to apply six fold symmetry to the KaiB portion of the KaiBC cryoEM structure. Single particle EM data refinement benefits from increased resolution when appropriate symmetry restrictions can be applied during the refinement process. In order to further support the visual inspection of the overall symmetry of the KaiBC complex, the following symmetry tests were performed.

Image slices from an intermediate structure without imposed symmetry were taken orthogonal to the central channel, which is along the six fold symmetry axis of KaiC. Using EMAN, each image slice was rotated and a correlation between the original and rotated slices was calculated. A graph of these correlations for each slice, in 1° increments of rotation, provides a way to assess the rotational symmetry of each slice. Each set of correlations was normalized to its average, allowing comparison between slices. The image slices were selected from three areas of the intermediate structure: density attributable solely to KaiC, overlapping areas of KaiB and KaiC, and portions of the structure
resulting solely from KaiB. The determinations for these sections was made with the KaiB crystal structure, docked into the smaller third ring of density (Fig. A3.1).

Six slices were taken from the intermediate structure, three sequential slices from the edge of the KaiB density at the KaiC interface (slices 94, 95, 96), and the remaining three slices are spaced through the remaining KaiB density. The individual image slices are shown in Fig. A3.2, and numbered according to their z-axis height in the intermediate structure. The results of the rotation correlation analysis for these chosen slices are shown in Fig. A3.3. The three KaiB slices closest to the KaiC interface show six peaks, suggesting that six fold symmetry would be an appropriate restriction for the refinement process. However, the KaiB slices further from the KaiC interface exhibit noisier curves. These slices show two identifiable peaks, indicating at least two fold symmetry, and noisier sections with broader peaks that are difficult to interpret. Partial occupancy of KaiB would tend to weaken the observed symmetry for the KaiB section of the structure. The overall correlation throughout the KaiB slices shows weakening symmetry moving away from the KaiC interface, suggesting that the KaiB region is relatively noisy compared to the KaiC region of the structure. However, the observation of six fold symmetry for the KaiB slices close to the KaiC interface supports the interpretation of KaiB as possessing pseudo six fold symmetry.
Figure A3.1 Regions of the KaiBC intermediate structure
This figure shows the intermediate structure divided into sections, with the KaiB portion in teal, and the KaiC portion in blue. The gap between the two sections indicates the interface between KaiB and KaiC, and is one slice thick. The gap is the location and height of slice 94, with all the slices marked by arrows. Scale bar is 50Å.
Figure A3.2 Image slices of the KaiB region from the KaiBC intermediate structure. These 2D slices were used in the rotation correlation analysis, with each slice being compared to rotations of itself. Each slice is numbered according to the z-height at which the slice was extracted from the full intermediate structure.
Figure A3.3. Rotation correlation analysis of images slices from the KaiBC intermediate structure
Each image was compared to rotations of that image, and the resulting correlations were normalized to the average for that particular image. Peaks show higher correlation, and angles of rotation close to 0° show the highest correlation as these rotations as they approximate a perfect match of no rotation. The three slices closest to KaiC (slices 94, 95, and 96) show six peaks by inspection, suggesting six fold symmetry. The remaining three slices (slices 99, 104, and 109) show two clear peaks, and a less clear overall symmetry pattern.
Appendix 4: Scripts

Both chapter 2 and 3 of this work made use of computer programming scripts to manage files, rapidly pass information to various programs, and to organize data. In many cases, a short script would be written for a single purpose, but other scripts that served a recurring purpose were created to be more readily modified for a wider variety of actions. Select scripts that served a complex purpose or were frequently used throughout my work are included in this appendix. There are three broad categories: data management, cryoEM density map refinement, and molecular simulation scripts.

Dataset Management Scripts

These scripts facilitate the naming, movement, and other 'housekeeping' aspects of working with files and datasets.

renamer.perl

A set of files is renamed. While a simple process for a single file, several hundred or thousands of image files are occasionally renamed while transferring between operating systems or in preparation for archiving. The current variables in this script leads to replacement of a “.” with a “-“, so as to avoiding naming convention issues within a linux filesystem.

#!/usr/bin/perl
#by Robert Kim
#20110505
#Subsequently edited by Seth Villarreal

#The goal of this script is to rename all of an ordered list of files according
#to a standard convention

#There are 3 sections to this script, the first and last require editing for each unique run
# Start value
$increment = 0;

# End value
$total = 552;

while ($increment <= $total) {
    # This section determines the necessary number of zeros for a file
    $increment++;
    if ($increment > 0 && $increment < 10) {
        $counter = "000".$increment;
    } elsif ($increment > 9 && $increment < 100) {
        $counter = "00".$increment;
    } elsif ($increment > 99 && $increment < 1000) {
        $counter = "0".$increment;
    } elsif ($increment > 999 && $increment < 10000) {
        $counter = $increment;
    }

    # The renaming bit, with the parts shown in pink requiring updates
    # according to the appropriate naming convention
    $orighandle = 'Ad5_PBGYAR.'.$counter.'.dm3';
    $finalhandle = 'Ad5_PBGYAR-'.$counter.'.dm3';
    system("mv $orighandle $finalhandle");

    # This critical section functions by first identifying a file that
    # matches a particular convention, and then creating a new name
    # which allows the pair of names to be used to issue a system command
    # in this case, the mv command.
}

dm3-transfer.s

During the initial steps of particle picking, not all collected dm3 files will provide useful particle images. These dm3 files may be moved to a new location for further evaluation. This script first identifies a dm3, then looks for a
corresponding box file, signifying that particles were found within the dm3 file. If no box file is found the dm3 files is copied to a new location.

```csh
#!/bin/csh

#Identifies dm3 files without a corresponding box file, copying such dm3 files to a new directory
#------------------------------------------------------------------------------------------------------------------------------
# Written by Seth Villarreal 20120507, with codes adapted from the scripts of Dewight R Williams, Susan D. Saban and Jian Shi
#------------------------------------------------------------------------------------------------------------------------------

#Set the source directories
#set d_root = `pwd`
set s_root = "/kappa3/pstewart/dnapkcs_ku_dna"                  # Top level source directory
set s_dm3 = $s_root"/dm3"                                       # dm3 file source sub-directory
set s_box = $s_root"/box/"                                      # box file source sub-directory

set d_root = `pwd`                                              # Top level destination directory
set d_dm3 = $d_root"/dm3"                                       # dm3 file destination directory

cd $s_dm3

echo "Copying dm3 files without corresponding box files"
echo "from "$s_dm3" to "$d_dm3
echo " "

#dm3list = `ls *.dm3`

foreach s_dm3_root (*dm3)
  echo "Starting on file "$s_dm3_root
  set root_name = `basename $s_dm3_root .dm3`
  set box_name = "$root_name"_bin4.box"
  if (! -e $s_box/$box_name) then
    cp $s_dm3_root $d_dm3
    echo "$s_dm3_root" transferred to "$d_dm3
  endif
end

echo " "
echo "Transfer Complete"
```
Refinement Scripts

A variety of scripts provide a repeatable and precise control over the refinement process. Some of the largest scripts in Appendix 2 are used for controlling Frealign, and other programs used in refinement. While this section covers the scripts used, processing methods are discussed in greater detail in the methods sections of the corresponding chapter for each project.

cffind3-batchboxer-append-155kx-p1.s

Several discrete steps in the initial stage of the refinement process are handled by this single script. There are three discrete sections: selecting the individual particles from each dm3 file, creating a list of the number of particles found on each dm3 file, and using CTFFIND3 to get defocus and angle of astigmatism values.

#!/bin/tcsh
source /programs/EMAN/eman.cshrc

# 1. Box particles into separate imagic stacks
# 2. Count the particle number in the IMAGIC file picked by boxer, having the same name with dm3 file
# 3. Use CTFFIND3 to get the defocus and angle of astigmatism
# Written by Seth A Villarreal, with codes adapted from the scripts of DRW, Susan D. Saban, and Jian Shi

set divider_num = 1

set bin_factor = 6
set box_bin_factor = 3
set AP = 3.523
set bbscale = 0.5
set box_size = 256
set magnification = 155000
set absmag = 198939 # For the Polara: 115kX-147600 155kX-198939 200kX-256695 310kX-397878
set stack_name = "part_"$divider_num"_stack_dkd_"$magnification"_box"$box_size"_bin"$bin_factor
set ifilmin = 30000

set d_root = `pwd`
set d_mrc = ${d_root}"/mrc-155000-"$divider_num
set d_img = ${d_root}"/parts"
set d_box = ${d_root}"/../box-256/box"
set ctffind3_prog = "/programs/ctffind3/ctf_v3_1/ctffind3.exe"

#Bin number multiplied by (14 for the T12 and 15 for the TF30)
@dstep = $bin_factor * 15 # This converts the bin factor to dstep

# Currently the CTFFIND3 mrc variable has a hard coded bin number ! ! ! !

# CTFFIND3 Constants
# Card 3 = contains 5 variables: CS(mm), HT(kV), AmpCnst, XMAG, DStep (um)
set c3a = 2.26;set c3b = 300;set c3c = 0.07;set c3d = $absmag;set c3e = $dstep

# Card 4 = Contains 6 variables: Box (not related to box files), ResMin(A), ResMax(A), dFMin(A), dFMax(A), FStep
# Do not change Box (c4a) above 512. It must also be even.
set c4a = 256;set c4b = 50;set c4c = 15;set c4d = 10000;set c4e = 80000;set c4f = 1000
# This is size that CTFFIND3 divides up a micrograph into during processing

set total_particle_num = 0

echo "The Angstrom to Pixel ration is ":$AP
echo "bin factor is ":$bin_factor

echo "Working directory ":$d_root /
$d_root/"part_"$divider_num"prefrealign_"$magnification.log
echo "#These are the card10 lines for the $stack_name stack" > $d_root/"part_"$divider_num"card10-"$magnification.txt
cd $d_root

# Loop Start

foreach bin_full_path ($d_mrc/DNA*$magnification*.mrc)
# Batchboxer Section

# Loop Start

foreach bin_full_path ($d_mrc/DNA*$magnification*.mrc)
# Batchboxer Section


echo "Cutting particles for:
Bin Name: $bin_full_path
Box Name: $d_box/$box_name

Hed Name: $d_img/$part_name"

echo "Using a scale of $bbscale and a newsize of $box_size"

batchboxer input=$bin_full_path dbbox=$d_box/$box_name scale=$bbscale newsize=$box_size output=$d_img/$part_name
echo "Batchboxer Completed"

set n_part = `iminfo $d_root/parts/$part_name | awk '/contains/ {print $3;} '`
set npart = `cat $d_box/$box_name | wc -l`
@ total_particle_num = $npart + $total_particle_num

echo "$part_name contains $n_part particles and its box file had $npart particles" >> $d_root/"part_"$divider_num"prefrealign_"$magnification.log

# CTFFIND3 Section

##############################################################################
###########
cd $d_mrc
set c1a = $root_name"bin"$bin_factor".mrc" #Card 1 = Input image file name
set rootname = `basename $c1a .mrc` #Card 2 = Output diagnostic file name
set c2a = "delete_me_"$rootname".mrc" #Provided a value within the foreach loop

# This If check looks if a ctffind3 log exists, NOT if the log is Complete
if (! -e ctffind3_$root_name"bin"$bin_factor"_delete_me.log") then
    echo "Start of CTFFIND3 with "$c1a
    #create instance of ctffind3 script
    #If an older version of the script exists, it will be erased
    echo "#!/bin/tcsh" > ctffind3_$root_name"bin"$bin_factor"_delete_me.s"
    echo "$ctffind3_prog << END01" >> ctffind3_$root_name"bin"$bin_factor"_delete_me.s"
    echo $c1a >> ctffind3_$root_name"bin"$bin_factor"_delete_me.s"
    echo $c2a >> ctffind3_$root_name"bin"$bin_factor"_delete_me.s"
    echo $c3a $c3b $c3c $c3d $c3e >> ctffind3_$root_name"bin"$bin_factor"_delete_me.s"
    echo $c4a $c4b $c4c $c4d $c4e $c4f >> ctffind3_$root_name"bin"$bin_factor"_delete_me.s"
    echo "END01"  >> ctffind3_$root_name"bin"$bin_factor"_delete_me.s"
    chmod 755 ctffind3_$root_name"bin"$bin_factor"_delete_me.s"
    ./ctffind3_$root_name"bin"$bin_factor"_delete_me.s"
    cffind3_$root_name"bin"$bin_factor"_delete_me.log"
endif

set dfmid1 = `more ctffind3_$root_name"bin"$bin_factor"_delete_me.log" | awk '/Final Values/ {print $1}'`
set dfmid2 = `more ctffind3_$root_name"bin"$bin_factor"_delete_me.log" | awk '/Final Values/ {print $2}'`
set angast = `more ctffind3_$root_name"bin"$bin_factor"_delete_me.log" | awk '/Final Values/ {print $3}'`

# Card 10
#NIN, ABSMAGPIN, IFILMIN, DFMID1N, DFMID2IN, ANGASTIN, MORE
Numerous rounds of refinement are frequently required while using frealign. A master script provides a single source with which to set up a refinement round, both the copying and creation of necessary files, and starting steps in a particular refinement round. Each round requires the creation of log files, control scripts for individual refinement processes, and the copying of output information or files. This master script creates individual scripts to carry out refinement in parallel, supplying necessary variables regarding a particular dataset, and monitoring steps in the refinement round such that each round can be fully completed without user input after the initial set up steps. For each round of refinement, this master script must be modified to reflect the name of the current refinement round, as well as other variables that change from round to round. Creating subscripts and output logs for various stages of a refinement round facilitate interpretation of results and checking for errors. While the overall
length of this master script is ponderous, it greatly facilitates the management
and processing for each round of refinement.

#!/bin/tcsh -f
echo " "
echo "MASTER OUTPUT"
echo " "

# EDIT the following lines about the previous round
set TOP = /kappa3/villarsa/dnapkcs_dna/bin8
set XPREV = rnd0
set XROUND = rnd1
set CHECKLOG1 = $TOP/master/master_dd_$XPREV".log"
set PREV_ROUND = $XPREV
set PREVPARTICLES_DS1 = $TOP/DNA-PKcs_DNAsa_bin8_64b_norm.mrc

cd $TOP
echo "Ref point 1"
# EDIT the following lines about the current round
set ROUND = $XROUND
set REFN_HIGH_RESN = 20
set REFN_LOW_RESN = 75
set CALC_MAP_RESN = 15
set FSTP_LEVEL = 0
set DEF_REFN = F
set MAG_REFN = F
# NOTE ONLY WHOLE NUMBERS ARE CURRENTLY SUPPORTED FOR THRESHOLD VALUES
# Thresh will be set automatically below

echo "Ref point 2"
### THE REST OF THE PARAMETERS ARE SET AUTOMATICALLY
# SET fine step sizes
if ( $FSTP_LEVEL == 0 ) then
  set FSTP_ANG = 0.1
  set FSTP_SH = 0.01
endif
if ( $FSTP_LEVEL == 1 ) then
  set FSTP_ANG = 0.05
  set FSTP_SH = 0.01
endif
if ( $FSTP_LEVEL == 2 ) then
  set FSTP_ANG = 0.025
  set FSTP_SH = 0.01
endif
if ( $FSTP_LEVEL == 3 ) then
  set FSTP_ANG = 0.01
set FSTP_SH = 0.005
endif
if ( $FSTP_LEVEL == 4 ) then
    set FSTP_ANG = 0.005
    set FSTP_SH = 0.0025
endif
if ( $FSTP_LEVEL == 5 ) then
    set FSTP_ANG = 0.0025
    set FSTP_SH = 0.001
endif
if ( $FSTP_LEVEL == 6 ) then
    set FSTP_ANG = 0.001
    set FSTP_SH = 0.0005
endif
if ( $FSTP_LEVEL == 7 ) then
    set FSTP_ANG = 0.0005
    set FSTP_SH = 0.00025
endif
if ( $FSTP_LEVEL == 8 ) then
    set FSTP_ANG = 0.00025
    set FSTP_SH = 0.0001
endif
if ( $FSTP_LEVEL == 9 ) then
    set FSTP_ANG = 0.0001
    set FSTP_SH = 0.00005
endif
if ( $FSTP_LEVEL == 10 ) then
    set FSTP_ANG = 0.00005
    set FSTP_SH = 0.000025
endif
if ( $FSTP_LEVEL == 11 ) then
    set FSTP_ANG = 0.000025
    set FSTP_SH = 0.00001
endif
if ( $FSTP_LEVEL == 12 ) then
    set FSTP_ANG = 0.00001
    set FSTP_SH = 0.000005
endif
if ( $FSTP_LEVEL == 13 ) then
    set FSTP_ANG = 0.000005
    set FSTP_SH = 0.0000025
endif

# EDIT the following line to set the maximum number of check points
# Note checking is done every 5min, so NCHECKS=288 is 288 check points
# and checking will be done for up to 24 hours (5min *288)
# NCHECKS=900 is over 3 days
# This sets the maximum wait time before starting current round,
as well as maximum time for refinement, and maximum time for iflag0
set NCHECKS = 9000

# Determine if previous master run has finished
echo "Determining if previous master run has finished"
set i = 1
while ($i <= $NCHECKS )
  if (! -e $CHECKLOG1 ) then
    echo "Previous master round not done yet " `date`
    sleep 5m
  endif
  @ i = $i + 1
end

set j = 1
while ($j <= $NCHECKS )
  set DONE = `tail -1 $CHECKLOG1 | grep 'DONE master' | wc -l`
  if ( $DONE == 0 ) then
    echo "Previous master round not done yet " `date`
    sleep 5m
  endif
  @ j = $j + 1
end

echo "Proceeding with refinement " `date`

echo "Setting threshes" `date`
set XTHR = `ls -l|grep 3D|grep $PREV_ROUND|awk '{print $9}'|cut -f1 -d '.'|cut -c 3-4`
@ XTHR_M1 = $XTHR - 1
@ XTHR_P1 = $XTHR + 1
@ XTHR_P2 = $XTHR + 2
echo $XTHR_M1
echo $XTHR
echo $XTHR_P1
echo $XTHR_P2

set THR_MAP1 = $XTHR_M1
set THR_MAP2 = $XTHR
set THR_MAP3 = $XTHR_P1
set THR_MAP4 = $XTHR_P2

# The following lines are set automatically for the Previous Round
set MAP_NAME = `ls -l|grep 3D |grep $PREV_ROUND|awk '{print $9}''
set PREVMAP = $TOP/orig/$MAP_NAME
set MAP_NAME_BASE = `ls -l|grep 3D|grep $PREV_ROUND|awk '{print $9}'|cut -f1 -d '.'`
set PREV_THR = `ls -l|grep 3D|grep $PREV_ROUND|awk '{print $9}'|cut -f1 -d '.'|cut -f2 -d '-'`
echo "previous best map was $PREVMAP"
echo "previous phase residual cutoff was $PREV_THR"

#set PREV_PARA_DS1 = $TOP/$PREV_ROUND/iflag0/"dad_b64_bin8_"$PREV_ROUND".para"

# The following lines are set automatically for the Current Round
set FNB_NEW_PARA_DS1 = "dad_b64_bin8_"$ROUND

set FNB_NEW_PARASH_DS1 = "dad_b64_bin8_"$ROUND

set NEWMAP_THR1 = $TOP/$ROUND/iflag0/"dad_3D_map64_"$ROUND"-pr"$THR_MAP1".mrc"
set NEWMAP_THR2 = $TOP/$ROUND/iflag0/"dad_3D_map64_"$ROUND"-pr"$THR_MAP2".mrc"
set NEWMAP_THR3 = $TOP/$ROUND/iflag0/"dad_3D_map64_"$ROUND"-pr"$THR_MAP3".mrc"
set NEWMAP_THR4 = $TOP/$ROUND/iflag0/"dad_3D_map64_"$ROUND"-pr"$THR_MAP4".mrc"

# The following parameters probably will remain the same
set RELMAG = 1.0
set THR_REFN = 90.0

# Automatically set some additional variables
set PARTICLE_STACK_DS1 = `basename $PREVPARTICLES_DS1`
set FN_NEWMAP_THR1 = `basename $NEWMAP_THR1`
set FN_NEWMAP_THR2 = `basename $NEWMAP_THR2`
set FN_NEWMAP_THR3 = `basename $NEWMAP_THR3`
set FN_NEWMAP_THR4 = `basename $NEWMAP_THR4`
set FN_PREVMAP = `basename $PREVMAP`

#set FN_PREV_PARA_DS1 = `basename $PREV_PARA_DS1`

# Create new directories and files for current round
echo "Creating new dirs and files for current round"
cd $TOP
mkdir $ROUND
cp $PREVMAP $ROUND
cd $ROUND
mkdir iflag0
cp $PREVMAP iflag0
mv iflag0/$FN_PREVMAP iflag0/$FN_NEWMAP_THR1
cp $PREVMAP iflag0

178
mv iflag0/$FN_PREVMAP iflag0/$FN_NEWMAP_THR2
cp $PREVMAP iflag0
mv iflag0/$FN_PREVMAP iflag0/$FN_NEWMAP_THR3
cp $PREVMAP iflag0
mv iflag0/$FN_PREVMAP iflag0/$FN_NEWMAP_THR4
cd $TOP
#cp $PREV_PARA_DS1 $ROUND
cd $ROUND

echo "Ref point 3"
# Create frealign script names
echo "Creating frealign script names"
set refn_scr1 = "f_"$ROUND"_p1"
set refn_scr2 = "f_"$ROUND"_p2"
set refn_scr3 = "f_"$ROUND"_p3"
set refn_scr4 = "f_"$ROUND"_p4"
set refn_scr5 = "f_"$ROUND"_p5"
set refn_scr6 = "f_"$ROUND"_p6"
set refn_scr7 = "f_"$ROUND"_p7"
set refn_scr8 = "f_"$ROUND"_p8"
set refn_scr1_log = "f_"$ROUND"_p1.log"
set refn_scr2_log = "f_"$ROUND"_p2.log"
set refn_scr3_log = "f_"$ROUND"_p3.log"
set refn_scr4_log = "f_"$ROUND"_p4.log"
set refn_scr5_log = "f_"$ROUND"_p5.log"
set refn_scr6_log = "f_"$ROUND"_p6.log"
set refn_scr7_log = "f_"$ROUND"_p7.log"
set refn_scr8_log = "f_"$ROUND"_p8.log"

set iflag0_thr_scr1 = "f_"$ROUND"_pr$THR_MAP1
set iflag0_thr_scr2 = "f_"$ROUND"_pr$THR_MAP2
set iflag0_thr_scr3 = "f_"$ROUND"_pr$THR_MAP3
set iflag0_thr_scr4 = "f_"$ROUND"_pr$THR_MAP4
set iflag0_thr_scr1_log = "f_"$ROUND"_pr$THR_MAP1.log"
set iflag0_thr_scr2_log = "f_"$ROUND"_pr$THR_MAP2.log"
set iflag0_thr_scr3_log = "f_"$ROUND"_pr$THR_MAP3.log"
set iflag0_thr_scr4_log = "f_"$ROUND"_pr$THR_MAP4.log"

echo "Ref point 4"
# Create frealign scripts for parts 1 thru 8
echo "Creating frealign scripts for eight refinement parts"
set n = 1
while ($n <= 8 )
  if ( $n == 1 ) then
    set refn_scr = $refn_scr1
    set IFIRST = 1
    set ILAST = 2500
if ( $n == 2 ) then
    set refn_scr = $refn_scr2
    set IFIRST = 2501
    set ILAST =  5000
endif
if ( $n == 3 ) then
    set refn_scr = $refn_scr3
    set IFIRST = 5001
    set ILAST =  7500
endif
if ( $n == 4 ) then
    set refn_scr = $refn_scr4
    set IFIRST = 7501
    set ILAST = 10000
endif
if ( $n == 5 ) then
    set refn_scr = $refn_scr5
    set IFIRST = 10001
    set ILAST = 12500
endif
if ( $n == 6 ) then
    set refn_scr = $refn_scr6
    set IFIRST = 12501
    set ILAST = 15000
endif
if ( $n == 7 ) then
    set refn_scr = $refn_scr7
    set IFIRST = 15001
    set ILAST = 17500
endif
if ( $n == 8 ) then
    set refn_scr = $refn_scr8
    set IFIRST = 17501
    set ILAST = 20319
endif

# !/bin/csh -f "$refn_scr
echo "/programs/frealign707/frealign/bin/frealign707_64lm_pad64.exe << END1" >> $refn_scr
echo "M,-3,SMAG_REFN,"DEF_REFN","DEF_REFN",F,0,F,F,T,T" >> $refn_scr
echo "80.0,0.0,4.7,0.07,1.0,100.0,60.0,20.0,20,20" >> $refn_scr
#echo $FSTP_ANG "$FSTP_SH >> $refn_scr
echo "1 1 1 1 1" >> $refn_scr
echo "$IFIRST "$ILAST >> $refn_scr
echo "$RELMAG 120.0 10.0 "$THR_REFN 2.26 300.0 0.0 0.0" >> $refn_scr
echo "$CALC_MAP_RESN","REFN_LOW_RESN","REFN_HIGH_RESN",0.0" >> $refn_scr
echo "../$PARTICLE_STACK_DS1" >> $refn_scr
echo "proj_"$n".mrc" >> $refn_scr
# echo $FN_PREV_PARA_DS1 >> $refn_scr
cat ..//rnd0/dad_card10.txt >> $refn_scr
echo $FNB_NEW_PARA_DS1"_$n".para" >> $refn_scr
echo $FNB_NEW_PARASH_DS1"_$n".shift" >> $refn_scr
echo "-100.0 120.0 10.0 "$THR_REFN" 2.26 300.0 0.0 0.0" >> $refn_scr
echo $MAP_NAME_BASE".mrc" >> $refn_scr
echo "ffweigh_"$n".mrc" >> $refn_scr
echo "ffqf_"$n".mrc" >> $refn_scr
echo "ffamp_"$n".mrc" >> $refn_scr
echo "ffpha_"$n".mrc" >> $refn_scr
echo "ffpoi_"$n".mrc" >> $refn_scr
echo "END1" >> $refn_scr

@ n = $n + 1
end
chmod 755 $refn_scr1
chmod 755 $refn_scr2
 chmod 755 $refn_scr3
chmod 755 $refn_scr4
chmod 755 $refn_scr5
chmod 755 $refn_scr6
 chmod 755 $refn_scr7
 chmod 755 $refn_scr8

echo "Ref point 5"
# Create frealign scripts for iflag0 runs
echo "Creating frealign scripts for iflag0 runs"
set o = 1
while ($o <= 4 )
  if ( $o == 1 ) then
    set iflag0_scr = $iflag0_thr_scr1
    set FN_NEWMAP = $FN_NEWMAP_THR1
    set THR_MAP = $THR_MAP1
  endif
  if ( $o == 2 ) then
    set iflag0_scr = $iflag0_thr_scr2
    set FN_NEWMAP = $FN_NEWMAP_THR2
    set THR_MAP = $THR_MAP2
  endif
  if ( $o == 3 ) then
    set iflag0_scr = $iflag0_thr_scr3
    set FN_NEWMAP = $FN_NEWMAP_THR3
    set THR_MAP = $THR_MAP3
  endif
  if ( $o == 4 ) then
    set iflag0_scr = $iflag0_thr_scr4
    set FN_NEWMAP = $FN_NEWMAP_THR4
set THR_MAP = $THR_MAP4
dendif

echo "#!/bin/csh -f" > $iflag0_scr
echo "/programs/frealign707/frealign/bin/frealign707_64lm_pad64.exe << END1" >> $iflag0_scr

@ o = $o + 1
end

chmod 755 $iflag0_thr_scr1
chmod 755 $iflag0_thr_scr2
chmod 755 $iflag0_thr_scr3
chmod 755 $iflag0_thr_scr4

mv $iflag0_thr_scr1 iflag0
mv $iflag0_thr_scr2 iflag0
mv $iflag0_thr_scr3 iflag0
mv $iflag0_thr_scr4 iflag0

echo "Ref point 6"
# Run eight refinement parts
echo "Starting eight refn rounds"
./$refn_scr1 & ./$refn_scr1_log &
./$refn_scr2 & ./$refn_scr2_log &
./$refn_scr3 & ./$refn_scr3_log &
./$refn_scr4 & ./$refn_scr4_log &
# Determine when eight refinement rounds are done  
echo "Determining when eight refn rounds are done"

set p = 1
while ($p <= 8 )
    if ($p == 1) then
        set CHECK_REFN_LOG = $refn_scr1_log
    endif
    if ($p == 2) then
        set CHECK_REFN_LOG = $refn_scr2_log
    endif
    if ($p == 3) then
        set CHECK_REFN_LOG = $refn_scr3_log
    endif
    if ($p == 4) then
        set CHECK_REFN_LOG = $refn_scr4_log
    endif
    if ($p == 5) then
        set CHECK_REFN_LOG = $refn_scr5_log
    endif
    if ($p == 6) then
        set CHECK_REFN_LOG = $refn_scr6_log
    endif
    if ($p == 7) then
        set CHECK_REFN_LOG = $refn_scr7_log
    endif
    if ($p == 8) then
        set CHECK_REFN_LOG = $refn_scr8_log
    endif

    set q = 1
    while ($q <= $NCHECKS )
        if (! -e $CHECK_REFN_LOG ) then
            echo "Eight refinement parts not done yet " `date`
            sleep 5m
        endif
        @ q = $q + 1
    end

    set r = 1
    while ($r <= $NCHECKS )
        set DONE = `tail $CHECK_REFN_LOG | grep -i 'NORMAL TERMINATION' | wc -l`
        if ( $DONE == 0 ) then
            echo "Eight refinement parts not done yet " `date`
            sleep 5m
        endif
        @ r = $r + 1
    end
echo "Eight refinement parts not done yet" `date`
  sleep 5m
endif
  @ r = $r + 1
end
  @ p = $p + 1
end

echo "Eight refinement parts are done" `date`
echo ""

# Clean up after refinement
rm -f ff**.mrc
rm -f *3D*.mrc

# Create combined parameter file for stack
echo "Creating combined parameter files"

cat $FNB_NEW_PARA_DS1"_p1.para" >> temp1.txt
cat $FNB_NEW_PARA_DS1"_p2.para" >> temp1.txt
cat $FNB_NEW_PARA_DS1"_p3.para" >> temp1.txt
cat $FNB_NEW_PARA_DS1"_p4.para" >> temp1.txt
cat $FNB_NEW_PARA_DS1"_p5.para" >> temp1.txt
cat $FNB_NEW_PARA_DS1"_p6.para" >> temp1.txt
cat $FNB_NEW_PARA_DS1"_p7.para" >> temp1.txt
cat $FNB_NEW_PARA_DS1"_p8.para" >> temp1.txt

cat temp1.txt | grep '^' > $FNB_NEW_PARA_DS1".para"

rm -f temp*.txt

# Cp combined parameter file to iflag0 dir
echo "Copying combined parameter files to iflag0 dir"
cp $FNB_NEW_PARA_DS1".para" iflag0

# Run 4 iflag0 jobs
echo "Starting 4 iflag0 jobs"
  cd iflag0
  ./iflag0_thr_scr1 > & ./iflag0_thr_scr1_log &
  ./iflag0_thr_scr2 > & ./iflag0_thr_scr2_log &
  ./iflag0_thr_scr3 > & ./iflag0_thr_scr3_log &
  ./iflag0_thr_scr4 > & ./iflag0_thr_scr4_log &

# Determine when 4 iflag0 jobs are done
echo "Determining when first 3 iflag0 jobs are done"

set p = 1
while ($p <= 4 )
if ($p == 1) then
    set CHECK_REFN_LOG = $iflag0_thr_scr1_log
endif
if ($p == 2) then
    set CHECK_REFN_LOG = $iflag0_thr_scr2_log
endif
if ($p == 3) then
    set CHECK_REFN_LOG = $iflag0_thr_scr3_log
endif
if ($p == 4) then
    set CHECK_REFN_LOG = $iflag0_thr_scr4_log
endif

set q = 1
while ($q <= $NCHECKS )
    if (! -e $CHECK_REFN_LOG ) then
        echo "Four iflag0 jobs not done yet " `date` 
        sleep 5m
    endif
    @ q = $q + 1
end

set r = 1
while ($r <= $NCHECKS )
    set DONE = `tail $CHECK_REFN_LOG | grep -i 'NORMAL TERMINATION' | wc -l`
    if ( $DONE == 0 ) then
        echo "Four iflag0 jobs not done yet " `date` 
        sleep 5m
    endif
    @ r = $r + 1
end
@ p = $p + 1
end

echo "Four iflag0 jobs are done " `date` 
echo " "
rm -f ff*.mrc

tail -n 10 $iflag0_thr_scr1_log | grep Average > tmp_thr1.txt
tail -n 10 $iflag0_thr_scr2_log | grep Average > tmp_thr2.txt
tail -n 10 $iflag0_thr_scr3_log | grep Average > tmp_thr3.txt
tail -n 10 $iflag0_thr_scr4_log | grep Average > tmp_thr4.txt

awk '{print $4}' tmp_thr1.txt > fsc_thr1.txt
awk '{print $4}' tmp_thr2.txt > fsc_thr2.txt
awk '{print $4}' tmp_thr3.txt > fsc_thr3.txt
awk '{print $4}' tmp_thr4.txt > fsc_thr4.txt
set in1 = `cut -f2 -d '.' fsc_thr1.txt`
set in2 = `cut -f2 -d '.' fsc_thr2.txt`
set in3 = `cut -f2 -d '.' fsc_thr3.txt`
set in4 = `cut -f2 -d '.' fsc_thr4.txt`

set inn1 = `echo $in1 | sed 's/0*//'`
set inn2 = `echo $in2 | sed 's/0*//'`
set inn3 = `echo $in3 | sed 's/0*//'`
set inn4 = `echo $in4 | sed 's/0*//'`

echo "Average FSC values as integers for each iflag0 run"
echo $inn1
echo $inn2
echo $inn3
echo $inn4

set best = 1
set bestfsc = $inn1

if ( $in2 > $bestfsc ) then
    set best = 2
    set bestfsc = $inn2
endif

if ( $in3 > $bestfsc ) then
    set best = 3
    set bestfsc = $inn3
endif

if ( $in4 > $bestfsc ) then
    set best = 4
    set bestfsc = $inn4
endif

echo ""
echo "Best assigned to:"
echo $best
echo ""

if ( $best == 1 ) then
    cp "dad_3D_map64_"$ROUND"-pr"$THR_MAP1".mrc" ../../orig
endif

if ( $best == 2 ) then
    cp "dad_3D_map64_"$ROUND"-pr"$THR_MAP2".mrc" ../../orig
endif
if ( $best == 3 ) then
    cp "dad_3D_map64 "$ROUND"-pr"$THR_MAP3".mrc" ../../../orig
endif

if ( $best == 4 ) then
    cp "dad_3D_map64 "$ROUND"-pr"$THR_MAP4".mrc" ../../../orig
endif

# Done master script
echo "DONE master script for round " $ROUND

f_rnd9_p1

A script created by the preceding master script. This script carries
refinement in the orientation of a particle image relative to a density map.

#!/bin/csh -f
/programs/frealign707/frealign/bin/frealign707_256lm_lopr_fstp_pad256.exe << END1
M,1,F,F,F,F,0,F,F,T,T
80.0,0.0,4.7,0.07,1.0,100.0,60.0,20.0,20,20
0.1 0.01
1 1 1 1
1 2500
C1
1.0 120.0 10.0 90.0 0.0 0.0 0.0
11,70,13,0.0
../DNA-PKcs_DNAsa_bin8_64b_norm.mrc
proj_1.mrc
/kappa3/villarsa/dnapkcs_dna/bin8/rand8/flag0/dd_b64_bin8_rnd8.para
dd_b64_bin8_rnd9_p1.para
dd_b64_bin8_rnd9_p1.shift
-100.0 120.0 10.0 90.0 0.0 0.0 2.26 300.0 0.0 0.0
dd_3D_map64_rnd8-pr64.mrc
ffweigh_1.mrc
ffqf_1.mrc
ffamp_1.mrc
ffpha_1.mrc
ffpoi_1.mrc
END1
f_rnd9_pr63

A script created by the previously described master script. After adjusting the orientation of particle images during a refinement round, new maps are generated. This script creates one such map.

#!/bin/csh -f
/programs/frealign707/frealign/bin/frealign707_256lm_lopr_fstp_pad256.exe << END1
M,0,F,F,F,F,0,F,F,T,T
80.0,0.0,4.7,0.07,1.0,100.0,60.0,20.0,20,20
0.1 0.01
1 1 1 1
1 20319
C1
1.0 120.0 10.0 63.0 2.26 300.0 0.0 0.0
11,70,13,0.0
../../DNA-PKcs_DNAsa_bin8_64b_norm.mrc
proj_ds1_pr63.mrc
dd_b64_bin8_rnd9.para
dd_b64_bin8_rnd9_if0_pr63.para
dd_b64_bin8_rnd9_if0_pr63.shift
0.0 120.0 10.0 63 2.26 300.0 0.0 0.0
dd_3D_map64_rnd9-pr63.mrc
ffweigh_pr63.mrc
ffqf_pr63.mrc
ffamp_pr63.mrc
ffpha_pr63.mrc
ffpoi_pr63.mrc
END1
rescue-creator-20120509.s

During refinement, each particle is given an initial center, and then during subsequent rounds the center is slightly adjusted to test for a better in the density map. In some cases, it is beneficial to reevaluate the center of a particle against a coarse sampling from all possible rotations. This script creates numerous independent scripts for testing different centers for each particle image.

#!/bin/tcsh
echo " "
echo "Rescue Round Script Creator"
echo "Created by Seth Villarreal"

set CURRENT_RND = 2
set PREV_RND = 1
set THR = 52
    #Desired Threshold
set COMPLEX = "dd_b64_bin8_rnd1"
    #Keeping track of what is being rescued, also used in naming output para
set MAP_NAME = "dad_3D_map64_rnd0-pr53.mrc"
    #map name used in the rescue rounds
set PARA_NAME = "dd_b64_bin8_rnd1.para"
    #para file used in the rescue rounds
set STACK = "/.\.\/DNA-PKcs_DNAsa_bin8_64b_norm.mrc"
    #location AND name of stack
set PARTICLE_COUNT = 35176
    #Number of Particles
set FREALIGN_VERSION = "/programs/frealign707/frealign/bin/frealign707_64lm_hipr_var_pad64.exe" #Set the
Frealign location

set DIR = "rnd"$CURRENT_RND
set TOP = "/kappa3/villarsa/dnapkcs_dna/bin8/"$DIR"/
    #Working directory
set MAP_LOC = "../orig/"
    #location of the rescue round map
set PARA_LOC = "../rnd$PREV_RND"/iflag0/
    #location of the parafile

mkdir $TOP
cd $TOP

set finalnum = 20     #Number of scripts created using the
loop. This is sequential
while ($o <= $finalnum )    #External Script Variables Loop
    ########################################
    if ( $o == 1 ) then    #First
        set X = 0    #Horizontal (X axis) Offset
        set Y = 2.5    #Vertical (Y axis) Offset
        set XX = "0"
        set YY = "2_5"
        mkdir $XX"-"$YY
        cp $MAP_LOC$MAP_NAME $XX"-"$YY
    end
    #Copying over the input map
cp $PARA_LOC$PARA_NAME $XX""$YY #Copying over the input para file
cd $TOP/$XX""$YY
endif
if ( $o == 2 ) then                     #Second
  set X = 2.5           #Horizontal (X axis) Offset
  set Y = 2.5           #Vertical (Y axis) Offset
  set XX = "2_5"        #Labeling for X axis
  set YY = "2_5"        #Labeling for Y axis
  mkdir $XX""$YY
  cp $MAP_LOC$MAP_NAME $XX""$YY         #Copying over the input map
  cp $PARA_LOC$PARA_NAME $XX""$YY #Copying over the input para file
  cd $TOP/$XX""$YY
endif
# Section Break: Out to (2.5,2.5)
if ( $o == 3 ) then                     #Third
  set X = 0             #Horizontal (X axis) Offset
  set Y = 5             #Vertical (Y axis) Offset
  set XX = "0"          #Labeling for X axis
  set YY = "5"          #Labeling for Y axis
  mkdir $XX""$YY
  cp $MAP_LOC$MAP_NAME $XX""$YY         #Copying over the input map
  cp $PARA_LOC$PARA_NAME $XX""$YY #Copying over the input para file
  cd $TOP/$XX""$YY
endif
if ( $o == 4 ) then                     #Fourth
  set X = 2.5           #Horizontal (X axis) Offset
  set Y = 5             #Vertical (Y axis) Offset
  set XX = "2_5"        #Labeling for X axis
  set YY = "5"          #Labeling for Y axis
  mkdir $XX""$YY
  cp $MAP_LOC$MAP_NAME $XX""$YY         #Copying over the input map
  cp $PARA_LOC$PARA_NAME $XX""$YY #Copying over the input para file
  cd $TOP/$XX""$YY
endif
if ( $o == 5 ) then                     #
  set X = 5             #Horizontal (X axis) Offset
  set Y = 2.5           #Vertical (Y axis) Offset
  set XX = "5"          #Labeling for X axis
  set YY = "2_5"        #Labeling for Y axis
  mkdir $XX""$YY
  cp $MAP_LOC$MAP_NAME $XX""$YY         #Copying over the input map
  cp $PARA_LOC$PARA_NAME $XX""$YY #Copying over the input para file
  cd $TOP/$XX""$YY
endif
if ( $o == 6 ) then                     #
  set X = 5             #Horizontal (X axis) Offset
  set Y = 5             #Vertical (Y axis) Offset
  set XX = "5"          #Labeling for X axis

set YY = "5"                        #Labeling for Y axis
mkdir $XX"-"$YY
cp $MAP_LOC$MAP_NAME $XX"-"$YY  #Copying over the input map
cp $PARA_LOC$PARA_NAME $XX"-"$YY #Copying over the input para file
cd $TOP/$XX"-"$YY

endif

# Section Break: Out to (5,5)

if ( $o == 7 ) then  #
    set X = 0       #Horizontal (X axis) Offset
    set Y = 7.5     #Vertical (Y axis) Offset
    set XX = "0"    #Labeling for X axis
    set YY = "7_5"  #Labeling for Y axis
    mkdir $XX"-"$YY
cp $MAP_LOC$MAP_NAME $XX"-"$YY  #Copying over the input map
cp $PARA_LOC$PARA_NAME $XX"-"$YY #Copying over the input para file
cd $TOP/$XX"-"$YY
endif

if ( $o == 8 ) then  #
    set X = 2.5    #Horizontal (X axis) Offset
    set Y = 7.5    #Vertical (Y axis) Offset
    set XX = "2_5" #Labeling for X axis
    set YY = "7_5" #Labeling for Y axis
    mkdir $XX"-"$YY
cp $MAP_LOC$MAP_NAME $XX"-"$YY  #Copying over the input map
cp $PARA_LOC$PARA_NAME $XX"-"$YY #Copying over the input para file
cd $TOP/$XX"-"$YY
endif

if ( $o == 9 ) then  #
    set X = 7.5    #Horizontal (X axis) Offset
    set Y = 2.5    #Vertical (Y axis) Offset
    set XX = "7_5" #Labeling for X axis
    set YY = "2_5" #Labeling for Y axis
    mkdir $XX"-"$YY
cp $MAP_LOC$MAP_NAME $XX"-"$YY  #Copying over the input map
cp $PARA_LOC$PARA_NAME $XX"-"$YY #Copying over the input para file
cd $TOP/$XX"-"$YY
endif

if ( $o == 10 ) then  #
    set X = 5      #Horizontal (X axis) Offset
    set Y = 7.5    #Vertical (Y axis) Offset
    set XX = "5"   #Labeling for X axis
    set YY = "7_5" #Labeling for Y axis
    mkdir $XX"-"$YY
cp $MAP_LOC$MAP_NAME $XX"-"$YY  #Copying over the input map
cp $PARA_LOC$PARA_NAME $XX"-"$YY #Copying over the input para file
cd $TOP/$XX"-"$YY
endif

if ( $o == 11 ) then  #


set X = 7.5    #Horizontal (X axis) Offset
set Y = 5    #Vertical (Y axis) Offset
set XX = "7_5"    #Labeling for X axis
set YY = "5"    #Labeling for Y axis
mkdir $XX"-"$YY
cp $MAP_LOC$MAP_NAME $XX"-"$YY  #Copying over the input map
cp $PARA_LOC$PARA_NAME $XX"-"$YY #Copying over the input para file
cd $TOP/$XX"-"$YY

if ( $o == 12 ) then  
set X = 7.5    #Horizontal (X axis) Offset
set Y = 7.5    #Vertical (Y axis) Offset
set XX = "7.5"    #Labeling for X axis
set YY = "7.5"    #Labeling for Y axis
mkdir $XX"-"$YY
cp $MAP_LOC$MAP_NAME $XX"-"$YY  #Copying over the input map
cp $PARA_LOC$PARA_NAME $XX"-"$YY #Copying over the input para file

cd $TOP/$XX"-"$YY
endif

if ( $o == 13 ) then  
set X = 0    #Horizontal (X axis) Offset
set Y = 10    #Vertical (Y axis) Offset
set XX = "0"    #Labeling for X axis
set YY = "10"    #Labeling for Y axis
mkdir $XX"-"$YY
cp $MAP_LOC$MAP_NAME $XX"-"$YY  #Copying over the input map
cp $PARA_LOC$PARA_NAME $XX"-"$YY #Copying over the input para file

cd $TOP/$XX"-"$YY
endif

if ( $o == 14 ) then  
set X = 2.5    #Horizontal (X axis) Offset
set Y = 10    #Vertical (Y axis) Offset
set XX = "2_5"    #Labeling for X axis
set YY = "10"    #Labeling for Y axis
mkdir $XX"-"$YY
cp $MAP_LOC$MAP_NAME $XX"-"$YY  #Copying over the input map
cp $PARA_LOC$PARA_NAME $XX"-"$YY #Copying over the input para file

cd $TOP/$XX"-"$YY
endif

if ( $o == 15 ) then  
set X = 10    #Horizontal (X axis) Offset
set Y = 2.5    #Vertical (Y axis) Offset
set XX = "10"    #Labeling for X axis
set YY = "2_5"    #Labeling for Y axis
mkdir $XX"-"$YY
cp $MAP_LOC$MAP_NAME $XX"-"$YY  #Copying over the input map
cp $PARA_LOC$PARA_NAME $XX"-"$YY #Copying over the input para file

cd $TOP/$XX"-"$YY
endif
cd $TOP/$XX""-"$YY
endif
if ( $o == 16 ) then
    set X = 5 #Horizontal (X axis) Offset
    set Y = 10 #Vertical (Y axis) Offset
    set XX = "5"
    set YY = "10"
    mkdir $XX""-"$YY
    cp $MAP_LOC$MAP_NAME $XX""-"$YY #Copying over the input map
    cp $PARA_LOC$PARA_NAME $XX""-"$YY #Copying over the input para file
    cd $TOP/$XX""-"$YY
endif
if ( $o == 17 ) then
    set X = 10 #Horizontal (X axis) Offset
    set Y = 5 #Vertical (Y axis) Offset
    set XX = "10"
    set YY = "5"
    mkdir $XX""-"$YY
    cp $MAP_LOC$MAP_NAME $XX""-"$YY #Copying over the input map
    cp $PARA_LOC$PARA_NAME $XX""-"$YY #Copying over the input para file
    cd $TOP/$XX""-"$YY
endif
if ( $o == 18 ) then
    set X = 7.5 #Horizontal (X axis) Offset
    set Y = 10 #Vertical (Y axis) Offset
    set XX = "7_5"
    set YY = "10"
    mkdir $XX""-"$YY
    cp $MAP_LOC$MAP_NAME $XX""-"$YY #Copying over the input map
    cp $PARA_LOC$PARA_NAME $XX""-"$YY #Copying over the input para file
    cd $TOP/$XX""-"$YY
endif
if ( $o == 19 ) then
    set X = 10 #Horizontal (X axis) Offset
    set Y = 7.5 #Vertical (Y axis) Offset
    set XX = "10"
    set YY = "7_5"
    mkdir $XX""-"$YY
    cp $MAP_LOC$MAP_NAME $XX""-"$YY #Copying over the input map
    cp $PARA_LOC$PARA_NAME $XX""-"$YY #Copying over the input para file
    cd $TOP/$XX""-"$YY
endif
if ( $o == 20 ) then
    set X = 10 #Horizontal (X axis) Offset
    set Y = 10 #Vertical (Y axis) Offset
    set XX = "10"
    set YY = "10"
    mkdir $XX""-"$YY
    cp $MAP_LOC$MAP_NAME $XX""-"$YY #Copying over the input map
    cp $PARA_LOC$PARA_NAME $XX""-"$YY #Copying over the input para file
    cd $TOP/$XX""-"$YY
endif
cp $MAP_LOC$MAP_NAME $XX"-"$YY #Copying over the input map
cp $PARA_LOC$PARA_NAME $XX"-"$YY #Copying over the input para file
cd $TOP/$XX"-"$YY
endif

# Section Break: Out to (10,10)
#########################################################

echo "Inner loop, 4 scripts"
set n = 1
echo "Identifying values of zero. Also, currently o is "$o
if ( $X != 0 ) then
  if ( $Y != 0 ) then
    echo "No zeros found and XX is "$XX" and YY is "$YY
while ($n <= 4 ) #Internal Script Creator Loop
  if ( $n == 1 ) then
    set X1 = $X
    set X2 = $Y
    set NX = $XX
    set NY = $YY
  endif
  if ( $n == 2 ) then
    set X1 = "."$X
    set X2 = $Y
    set NX = "m"$XX
    set NY = $YY
  endif
  if ( $n == 3 ) then
    set X1 = "."$X
    set X2 = "."$Y
    set NX = "m"$XX
    set NY = "m"$YY
  endif
  if ( $n == 4 ) then
    set X1 = $X
    set X2 = "."$Y
    set NX = $XX
    set NY = "m"$YY
  endif

  set scrpt_nam = $COMPLEX"_rescue_"$NX"-"$NY
  #set scrpt_nam = "f_rnd2_"$NX"-"$NY".pbs"  
  echo "#!/bin/tcsh" >> $scrpt_nam
  #echo "#PBS -M justin.w.flatt@vanderbilt.edu" >> $scrpt_nam
  #echo "#PBS -m bae" >> $scrpt_nam
  #echo "#PBS -l nodes=1:ppn=1:nehalem" >> $scrpt_nam
  #echo "#PBS -l mem=500mb">> $scrpt_nam
  #echo "#PBS -l walltime=60:00:00" >> $scrpt_nam
  #echo "#PBS -l cput=60:00:00" >> $scrpt_nam
  #echo "#PBS -o f_rnd2_"$NX"-"$NY".out" >> $scrpt_nam
#echo "#PBS -j oe" >> $scrpt_nam
echo "cd "$TOP$XX"-$YY" >> $scrpt_nam
echo $FREALIGN_VERSION" << END1" >> $scrpt_nam
echo "M 3 F F F 0 F T T" >> $scrpt_nam
echo "80.0 0.0 4.7 0.07 1.0 100.0 60.0 20.0 20 20" >> $scrpt_nam
echo "1 1 1 1" >> $scrpt_nam
echo "1 "$PARTICLE_COUNT >> $scrpt_nam
echo "C1" >> $scrpt_nam
echo "1.0 120.0 10.0 "$THR".0 2.26 300.0 0.0 0.0" >> $scrpt_nam
echo "15.0 75.0 20.0 0.0" >> $scrpt_nam
echo "$STACK" >> $scrpt_nam
echo "$PARA_NAME" >> $scrpt_nam
echo "$X1" "$X2" >> $scrpt_nam
echo "$COMPLEX"_rescue_"$NX"."$NY".para" >> $scrpt_nam
echo "$COMPLEX"_rescue_"$NX"."$NY".shift" >> $scrpt_nam
echo "-100.0 120.0 10.0 "$THR".0 2.26 300.0 0.0 0.0" >> $scrpt_nam
echo ".$MAP_NAME" >> $scrpt_nam
echo "fweigh.mrc" >> $scrpt_nam
echo "fqf.mrc" >> $scrpt_nam
echo "famp.mrc" >> $scrpt_nam
echo "fpha.mrc" >> $scrpt_nam
echo "fpoi.mrc" >> $scrpt_nam
echo "END1" >> $scrpt_nam

@ n = $n + 1  #Counter for the Internal Script Creator Loop
chmod 755 $scrpt_nam

else
  echo "Zero found for Y variable and XX is "$XX" and YY is "$YY"
  while ($n <= 4 )
    if ( $n == 1 )
      set X1 = $X
      set X2 = $Y
      set NX = $XX
      set NY = $YY
    endif
    if ( $n == 2 )
      set X1 = "-$X"
      set X2 = $Y
      set NX = "$XX"
      set NY = "$YY"
    endif
    n = $n + 1
  endwhile
@end

#This is required by the Rescue Manager script
set NY = $YY
endif
if ( $n == 3 ) then
  set X1 = $Y
  set X2 = $X
  set NX = $YY
  set NY = $XX
endif
if ( $n == 4 ) then
  set X1 = $Y
  set X2 = "-"$X
  set NX = $YY
  set NY = "m"$XX
endif
set scrpt_nam = $COMPLEX"_rescue_"$NX"-"$NY
#set scrpt_nam = "f_rnd2_"$NX"-"$NY".pbs"
echo "#!/bin/tcsh" >> $scrpt_nam
#echo "#PBS -M justin.w.flatt@vanderbilt.edu" >> $scrpt_nam
#echo "#PBS -m be" >> $scrpt_nam
#echo "#PBS -l nodes=1:ppn=1:nehalem" >> $scrpt_nam
#echo "#PBS -l mem=500mb" >> $scrpt_nam
#echo "#PBS -l walltime=60:00:00" >> $scrpt_nam
#echo "#PBS -l cput=60:00:00" >> $scrpt_nam
#echo "#PBS -o f_rnd2_"$NX"-"$NY".out" >> $scrpt_nam
#echo "#PBS -j oe" >> $scrpt_nam
echo "cd "$TOP""$XX"-"$YY" >> $scrpt_nam
echo $FREALIGN_VERSION" << END1" >> $scrpt_nam
echo "M 3 F F F F 0 F F T T" >> $scrpt_nam
echo "80.0 0.0 4.7 0.07 1.0 100.0 60.0 20.0 20 20" >> $scrpt_nam
echo "1 1 1 1 1" >> $scrpt_nam
echo "$PARTICLE_COUNT" >> $scrpt_nam
echo "C1" >> $scrpt_nam
echo "1.0 120.0 10.0 "$THR".0 2.26 300.0 0.0 0.0" >> $scrpt_nam
echo "15.0 75.0 20.0 0.0" >> $scrpt_nam
$STACK >> $scrpt_nam
echo "proj.mrc" >> $scrpt_nam
echo $PARA_NAME >> $scrpt_nam
echo "$X1" "$X2" >> $scrpt_nam
echo $COMPLEX"_rescue_"$NX"-"$NY".para" >> $scrpt_nam
echo $COMPLEX"_rescue_"$NX"-"$NY".shift" >> $scrpt_nam
echo "-100.0 120.0 10.0 "$THR".0 2.26 300.0 0.0 0.0" >> $scrpt_nam
#echo "$MAP_NAME" >> $scrpt_nam
#echo "fweigh.mrc" >> $scrpt_nam
#echo "fqf.mrc" >> $scrpt_nam
#echo "famp.mrc" >> $scrpt_nam
#echo "fpha.mrc" >> $scrpt_nam
#echo "fpoi.mrc" >> $scrpt_nam
#echo "END1" >> $scrpt_nam
@ n = $n + 1  #Counter for the Internal Script Creator Loop
chmod 755 $scrpt_nam

echo "./"$XX"-"$YY"/"$scrpt_nam" >& "$XX"-"$YY"/"$scrpt_nam".log
&" >> $TOP"../master/"$DIR"-rescue-script-start-file.s"        #This is required by the Rescue
Manager script
end
endif
else
    echo "Zero found for X variable and XX is "$XX" and YY is "$YY
while ($n <= 4 )
    #echo "Currently n is "$n
    if ( $n == 1 ) then
        set X1 = $X
        set X2 = $Y
        set NX = $XX
        set NY = $YY
    endif
    if ( $n == 2 ) then
        set X1 = "."$Y
        set X2 = $X
        set NX = "m"$YY
        set NY = $XX
    endif
    if ( $n == 3 ) then
        set X1 = $Y
        set X2 = $X
        set NX = $YY
        set NY = $XX
    endif
    if ( $n == 4 ) then
        set X1 = $X
        set X2 = "."$Y
        set NX = $XX
        set NY = "m"$YY
    endif
    set scrpt_nam = $COMPLEX"_rescue_"$NX"-"$NY
#set scrpt_nam = "f_rnd2_"$NX"-"$NY".pbs"
    echo "/bin/tcsh" >> $scrpt_nam
    echo "#PBS -M justin.w.flatt@vanderbilt.edu" >> $scrpt_nam
    echo "#PBS -m bae" >> $scrpt_nam
    echo "#PBS -l nodes=1:ppn=1:nehalem" >> $scrpt_nam
    echo "#PBS -l mem=500mb">> $scrpt_nam
    echo "#PBS -l walltime=60:00:00" >> $scrpt_nam
    echo "#PBS -l cput=60:00:00" >> $scrpt_nam
    echo "#PBS -o f_rnd2_"$NX"-"$NY".out" >> $scrpt_nam
    echo "#PBS -j oe" >> $scrpt_nam
echo "cd "$TOP"-$XX"-$YY >> $scrpt_nam
echo "$REALIGN_VERSION" << END1 >> $scrpt_nam
echo "M 3 F F F F F F" >> $scrpt_nam
echo "80.0 0.0 4.7 0.07 1.0 100.0 60.0 20.0 20 20" >> $scrpt_nam
echo "1 1 1 1 1" >> $scrpt_nam
echo "$PARTICLE_COUNT" >> $scrpt_nam
echo "C1" >> $scrpt_nam
echo "1.0 120.0 10.0 "$THR".0 2.26 300.0 0.0 0.0" >> $scrpt_nam
echo "15.0 75.0 20.0 0.0" >> $scrpt_nam
echo "$STACK" >> $scrpt_nam
echo "proj.mrc" >> $scrpt_nam
echo "$PARA_NAME" >> $scrpt_nam
echo "$X1" "$X2" >> $scrpt_nam
echo "$COMPLEX"_rescue_"$NX"-"$NY".para >> $scrpt_nam
echo "$COMPLEX"_rescue_"$NX"-"$NY".shift >> $scrpt_nam
echo ".-100.0 120.0 10.0 "$THR".0 2.26 300.0 0.0 0.0" >> $scrpt_nam
echo "$MAP_NAME" >> $scrpt_nam
echo "fweigh.mrc" >> $scrpt_nam
echo "famp.mrc" >> $scrpt_nam
echo "fpha.mrc" >> $scrpt_nam
echo "fpoi.mrc" >> $scrpt_nam
echo "END1" >> $scrpt_nam

@ n =$n + 1 #Counter for the Internal Script Creator Loop
chmod 755 $scrpt_nam

end
endif
@ o =$o + 1 #Counter for the External Script Variables Loop

echo "./"$XX"-"$YY"/"$scrpt_nam" & "$XX"-"$YY"/"$scrpt_nam".log & >> $TOP"../master/"$DIR"-rescue-script-start-file.s" #This is required by the Rescue Manager script

end

chmod 755 $TOP"../master/"$DIR"-rescue-script-start-file.s"

VMD Scripts

An important part of the interpretation of the KaiBC interface, from chapter 2 of this work, relied on molecular simulations of the individual crystal structures of the proteins as guided by the cryoEM density map. There are two key
programs for carrying out molecular dynamics flexible fitting (MDFF) simulations. The first is Visualizing Molecular Dynamics (VMD), which serves as an interface program for the second, which is the Not (just) Another Molecular Dynamics (NAMD) simulations package (M. T. Nelson et al., 1996; Phillips et al., 1998). VMD provides the mechanisms for setting up a MDFF simulation, while NAMD carries out the simulation.

vmd_setup_mngr.s

  Working with VMD led to repetitions in a number of areas, such as the reuse of the cryoEM density map and set variables. In order to facilitate the setup of many similar simulations, a set of scripts was used to automate as much of this process as possible. This particular script carries out several goals for setting up a simulation. The first removes files that may be left over from a failed or aborted setup. The second starts a subscript, which starts VMD and loads an initial file. The script then pauses while a user makes various modifications, if necessary, within VMD, before the script finishes preparing the files for use in a MDFF simulation.

#!/bin/tcsh -f

# This script REQUIRES the vmd_setup_before_autopsf.tcl and vmd_setup_after_autopsf.tcl subscripts to exist
# and be stored in ~/scripts

# Additionally, this script REQUIRES that copies of the appropriate par_all27_prot_lipid_na.inp and mdff_template.namd files be stored in ~/scripts/data_files

rm -f par_all27_prot_lipid_na.inp
rm -f mdff_template.namd

cmp -f top_all27_prot_na_patch_phos_atp_h20.rtf
cp ~/scripts/data_files/top_all27_prot_na_patch_phos_atp_h20.rtf ./

199
vmd -e ~/scripts/vmd_setup_before_autopsf.tcl
vmd -e ~/scripts/vmd_setup_after_autopsf.tcl
rm -f par_all27_prot_lipid_na.inp
rm -f mdff_template.namd
cp ~/scripts/data_files/par_all27_prot_lipid_na.inp ./
cp ~/scripts/data_files/mdff_template.namd ./
cp ~/scripts/data_files/generic-namd.pbs ./

vmd_setup_before_autopsf.tcl

This single line comprises the entirety of the first of two subscripts used by
the prior script during the setup of a MDFF simulation. Within VMD, various
modifications to the starting conditions of a structure may be necessary, and this
line loads up the default structure in preparation for this process.

mol new [glob kaibc.pdb]

vmd_setup_after_autopsf.tcl

The second of two scripts used by the previously described
vmd_setup_mngr.s script, creates numerous files required by NAMD. These files
describe the individual atoms, their characteristics and bonds, as well the matrix
of values resulting from the density map used to influence the conformational
changes in the crystal structure during a simulation.

#mol new PDB_BASENAME.pdb

#AutoPSF (SP2, THP2, Regenerate angles/dihedrals)

mdff gridpdb -psf [glob kaibc*_autopsf.psf] -pdb [glob kaibc*_autopsf.pdb] -o vmd_autopsf-
grid.pdb
mol new [glob kaibc*_autopsf.psf]
mol addfile [glob kaibc*_autopsf.psf]
ssrestraints -psf [glob kaibc*_autopsf.psf] -pdb [glob kaibc*_autopsf.pdb] -o extrabonds.txt -
hbonds
cispeptide restrain -o extrabonds-cispeptide.txt
chirality restrain -o extrabonds-chirality.txt
After a simulation completes, analysis of the KaiBC interface includes an assessment of the nonbonded energy between the interacting proteins. This script accesses VMD to quickly load up the resulting files from a simulation, divide them into sections, separating KaiB from KaiC, and then calculate the predicted energies for the interface.

```tcl
mol new [glob kaibc*_autopsf.psf]
mol addfile [glob kaibc*_autopsf.pdb]
mol addfile run-step2.dcd
waitfor all

#namdenergy -vdw -elec -nonb -ofile nonbonde_p7.txt -par par_all27_prot_lipid_na.inp -sel [atomselect top "segname P7"] [atomselect top "segname P1 P2 P3 P4 P5 P6"]
#namdenergy -vdw -elec -nonb -ofile nonbonde_p7-r7-44_70-93.txt -par par_all27_prot_lipid_na.inp -sel [atomselect top "segname P7 and (resid 7 to 44 or resid 70 to 93)"] [atomselect top "segname P1 P2 P3 P4 P5 P6"]
#exit

set u 7; # Constant Segname
set v 74; # Constant Resid

set w 1; # Variable Segname Start
set y 6; # Variable Segname End

set x 93; # Variable Resid Start
set z 94; # Variable Resid End
```
#while {$w < $y} {
    while {$x < $z} {
        nAMDenergy -vdw -elec -nonb -ofile "nonbonde_p$w-r$x-to-kaic.txt" -par
        par_all27_prot_lipid_na.inp -sel [atomselect top "segname P$w and resid
        $x"]  
        set x [expr {$x + 1}] 
    }
    #set x 14
    #set w [expr {$w + 1}]
#

    set u 7;# Constant Segname
    set v 74;# Constant Resid

    set w 6;# Variable Segname Start
    set y 7;# Variable Segname End

    set x 14;# Variable Resid Start
    set z 480;# Variable Resid End

    while {$w < $y} {
        while {$x < $z} {
            nAMDenergy -vdw -elec -nonb -ofile "nonbonde_p$w-r$x-to-p$u.txt" -par
            par_all27_prot_lipid_na.inp -sel [atomselect top "segname P$w and resid
            $x"]  
            set x [expr {$x + 1}] 
        }
        set x 14
        set w [expr {$w + 1}]
    }

    exit

    set u 7;# Constant Segname
    set v 74;# Constant Resid

    set w 1;# Variable Segname Start
    set y 6;# Variable Segname End

    set x 7;# Variable Resid Start
    set z 93;# Variable Resid End

    while {$w < $y} {
        while {$x < $z} {
            nAMDenergy -vdw -elec -nonb -ofile "nonbonde_p$w-r$x-to-p$w.txt" -par
            par_all27_prot_lipid_na.inp -sel [atomselect top "segname P$w and resid
            $x"]  
            set x [expr {$x + 1}] 
        }
        set x 14
        set w [expr {$w + 1}]
    }

set x [expr {$x + 1}]
}
set x 14
set w [expr {$w + 1}]
}
c1_non_bond_e.tcl

In addition to calculating the nonbonded energy on a per residue basis, the overall energy of interaction could be rapidly calculated. This script is separate from the prior one so as to allow for rapid assessment of results.

mol new [glob vmd*_autopsf.psf]
mol addfile [glob vmd*_autopsf.pdb]
mol addfile run-step8.dcd
namdenergy -vdw -elec -nonb -ofile non_bond_e.txt -par par_all27_prot_lipid_na.inp -sel [atomselect top "segname P1 P2"] [atomselect top "segname P5 P6 P7 P8 P9 P10"]
exit

vmd_kaibc_pdb_to_chimera.s

The NAMD output files of the final structures do not have the header information or the structure separation information present in the original input files. This short script takes the output structure from NAMD and restores the minimally necessary separation information to allow an interpretation in Chimera.

#!/bin/tcsh -f

set alpha = "final_frame.pdb"
set beta = "chimera_final_frame.pdb"

head -n 1 $alpha > $beta
echo "model 1" >> $beta
head -n 1558 $alpha | tail -n 1557 >> $beta
echo "endmdl\nmodel 2" >> $beta
head -n 3115 $alpha | tail -n 1557 >> $beta
echo "endmdl\nmodel 3" >> $beta
head -n 48757 $alpha | tail -n 45642 >> $beta
echo "endmdl\nmodel 4" >> $beta
head -n 51871 $alpha | tail -n 3114 >> $beta
echo "endmdl\nend" >> $beta
A cross correlation coefficient was examined as a possible method for examining the results of a NAMD simulation. This script generates these values, examining the conformational changes in a structure over simulated time as compared to the guiding density map.

```tcl
mol new [glob vmd*_autopsf.psf]
mol addfile [glob vmd*_autopsf.pdb]
mol addfile run-step8.dcd
mdff check -ccc -map ../../../align_bc_3d_map_rnd4all_pr66.mrc -res 5 -spacing 1.11 -cccfile ccc-output.txt
exit
```


Brinkers, S., Dietrich, H. R. C., de Groote, F. H., Young, I. T., & Rieger, B. (2009). The persistence length of double stranded DNA determined using


Pâques, F., & Haber, J. E. (1999). Multiple Pathways of Recombination Induced by Double-Strand Breaks in Saccharomyces cerevisiae Multiple Pathways of Recombination Induced by Double-Strand Breaks in Saccharomyces cerevisiae, 63(2).


as a major circadian timing mediator in cyanobacteria. *Proceedings of the National Academy of Sciences*, 103(32), 12109–14. doi:10.1073/pnas.0602955103


