University of Tennessee Health Science Center
UTHSC Digital Commons

Theses and Dissertations (ETD) College of Graduate Health Sciences

12-2007

DNA Binding Specificity of Mu Transcription Factor C and Crystallization of C : DNA Complex

Karthik Shanmugantham
University of Tennessee Health Science Center

Follow this and additional works at: https://dc.uthsc.edu/dissertations

 Part of the Medical Sciences Commons

Recommended Citation

Shanmugantham, Karthik , "DNA Binding Specificity of Mu Transcription Factor C and Crystallization of C : DNA Complex" (2007). *Theses and Dissertations (ETD).* Paper 238. http://dx.doi.org/10.21007/etd.cghs.2007.0283.

This Dissertation is brought to you for free and open access by the College of Graduate Health Sciences at UTHSC Digital Commons. It has been accepted for inclusion in Theses and Dissertations (ETD) by an authorized administrator of UTHSC Digital Commons. For more information, please contact jwelch30@uthsc.edu.

DNA Binding Specificity of Mu Transcription Factor C and Crystallization of C : DNA Complex

Abstract

The lytic cycle of phage Mu is regulated by a transcriptional cascade consisting of early, middle and late transcription. The Mor protein is an activator of the middle promoter Pm and is encoded by the last gene of the early transcript. The C protein is an activator of the four late promoters Plys, PI, PP, and Pmom and is expressed from the middle transcript. Both Mor and C proteins bind an imperfect dyad-symmetry element just upstream and overlapping the –35 region of Pm and Plys respectively. The main aims of this study was, (1) To understand the binding specificity of C and determine a possible consensus sequence for C binding, and (2) To crystallize the C : DNA complex as a first step towards structure determination. In previous work, single base substitution mutations in Plys identified bases and positions important for C binding and activation. To get a consensus sequence for C binding, we tested additional candidate mutations within and flanking the C binding sequence. Wild-type C protein was used in gel mobility shift assays with annealed oligonucleotides containing mutations, insertions and deletions. The assay showed that, (1) mutation in positions –53, –52 and –32 did not affect C binding, (2) mutations flanking the IR spacer (–40, –41, –46, –47) influence C binding, and (3) insertion or deletion of a single base pair in the IR spacer abolished C binding. Mor and C proteins are the founding members of a new class of transcription factors. The Mor structure revealed that it has a classical DNA-binding HTH motif and a dimerization domain. Based on the structure it has been proposed that Mor has to undergo conformational changes to bind DNA. Modelling of C based on the Mor structure revealed that C might also have a dimerization domain and a HTH DNA binding motif. To see if any conformational changes occur in C when it binds DNA, co- crystallization of a C : DNA complex was undertaken. Preliminary structural analysis of the complex revealed that under the crystallization conditions used C protein is bound to its symmetrical binding site using two HTH motifs from two C dimers without inducing any conformational change in itself or the DNA.

Document Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Program

Microbiology and Immunology

Research Advisor

Martha M. Howe, Ph.D.

Keywords

Bacteriophage Mu, transcription, transcription factor, activator protein C protein DNA interaction, crystallography

Subject Categories

Medical Sciences | Medicine and Health Sciences

This dissertation is available at UTHSC Digital Commons: https://dc.uthsc.edu/dissertations/238

DNA Binding Specificity of Mu Transcription Factor C and Crystallization of C : DNA Complex

A Dissertation
Presented for
The Graduate Studies Council
The University of Tennessee
Health Science Center

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy
From The University of Tennessee

By
Karthik Shanmuganatham
December, 2007

Chapter 3 © 2007 by International Union of Crystallography.
All other material © 2007 by Karthik Shanmuganatham

Dedication

This dissertation is the culmination of all the sacrifices done by my loving parents Vasuki Shanmuganatham and Shanmuganatham and for that I dedicate this dissertation to them.

Acknowledgements

My heartfelt thanks goes to my mentor Dr. Martha M. Howe for providing me with an opportunity to work in her lab and giving her guidance and input in all my scientific endeavors. I am especially grateful to her for her unwavering support in my crystallography project and letting me learn from my own mistakes. I am grateful to my committee member Dr. Hee Won Park who has been a co-mentor, friend and continuous source of encouragement for me throughout my crystallography project. I would also like to thank my other committee members, Dr. Marko Radic, Dr. Rajendra Raghow, and Dr. James Patrick Ryan for their suggestions, help and advice during the course of my graduate study. I also wish to thank Ms. Manimekalai Ravichandran for being very tenacious and helping me crystallize the complex used in this dissertation.

I personally thank my best friends Jeetendra and Muthiah for guiding me through my tumultuous graduate study and to their invaluable friendship and support in my personal and public life. I would also like to thank my adoptive sister Ms. Himangi Jayakar for her continuous support and encouragement.

I am very thankful to my girlfriend Ms. Angi Beau who has been at my side with words of encouragement and motivation whenever I needed her for the good latter part of my graduate study.

Words cannot describe how to thank the two most influential people in my life, my parents Vasuki and Shanmuganatham. This dissertation would not be possible but for their love, and unwavering support in my abilities. In addition, I would like to give them my heartfelt thanks for all their sacrifices they have made for me.

Abstract

The lytic cycle of phage Mu is regulated by a transcriptional cascade consisting of early, middle and late transcription. The Mor protein is an activator of the middle promoter P_m and is encoded by the last gene of the early transcript. The C protein is an activator of the four late promoters P_{lys}, P_I, P_P, and P_{mom} and is expressed from the middle transcript. Both Mor and C proteins bind an imperfect dyad-symmetry element just upstream and overlapping the –35 region of P_m and P_{lys} respectively. The main aims of this study was, (1) To understand the binding specificity of C and determine a possible consensus sequence for C binding, and (2) To crystallize the C : DNA complex as a first step towards structure determination.

In previous work, single base substitution mutations in P_{lys} identified bases and positions important for C binding and activation. To get a consensus sequence for C binding, we tested additional candidate mutations within and flanking the C binding sequence. Wild-type C protein was used in gel mobility shift assays with annealed oligonucleotides containing mutations, insertions and deletions. The assay showed that, (1) mutation in positions –53, –52 and –32 did not affect C binding, (2) mutations flanking the IR spacer (–40, –41, –46, –47) influence C binding, and (3) insertion or deletion of a single base pair in the IR spacer abolished C binding.

Mor and C proteins are the founding members of a new class of transcription factors. The Mor structure revealed that it has a classical DNA-binding HTH motif and a dimerization domain. Based on the structure it has been proposed that Mor has to undergo conformational changes to bind DNA. Modelling of C based on the Mor

structure revealed that C might also have a dimerization domain and a HTH DNA binding motif. To see if any conformational changes occur in C when it binds DNA, co-crystallization of a C : DNA complex was undertaken. Preliminary structural analysis of the complex revealed that under the crystallization conditions used C protein is bound to its symmetrical binding site using two HTH motifs from two C dimers without inducing any conformational change in itself or the DNA.

Table of Contents

Chapter 1. Introduction 1

An overview of transcription 1

RNA polymerase 2
- The sigma (σ) subunit 2
- The alpha (α) subunit 5
- The β and β′ subunits 6
- Overall structure of RNAP 6
- Promoter 8
- RNAP-activator interactions 9

Transcription initiation 9

Regulation of transcription 12

X-ray crystallography 17
- Principles of X-ray crystallography 18
- Crystallization 22
- Methods for protein crystallization 23
- Screening methods 26
- Optimization 27
- Data collection and processing 27
- Phase angle determination 33
- Calculation of electron density map, refinement and model building 35
- Validation 36

Bacteriophage Mu 36

Chapter 2. Binding Specificity of Mu Transcription Activator C 45

Introduction 45

Materials and methods 49
- Media, chemicals and enzymes 49
- Oligodeoxyribonucleotides 50
- Bacterial strains and plasmids 50
- P_{sym} mutants 54
- Electrophoretic mobility shift assays 56

Results 56
P_{sym} mutants 56
Gel-shift assays 57
Discussion 60

Chapter 3. Expression, Purification, Crystallization and Preliminary X-ray analysis of C Protein Bound to P_{sym} DNA 75

Introduction 75

Materials and methods 76
Chemicals, enzymes and media 76
Bacterial strains and plasmids 77
Oligodeoxyribonucleotides 79
Crystallization 79
Wild-type C protein production 82
Expression and solubility test 82
Large-scale production 83
Cell lysis 83
Chromatography 84
Heparin-sepharose affinity chromatography 84
SP-sepharose cation exchange column 86
Phenyl-sepharose hydrophobic exchange column 86
Gel-filtration or size exclusion chromatography 87
Selenomethionine C protein production 88
Electrophoretic mobility shift assays 89
Preparation of C : DNA complex for crystallization 89
Crystallization 90
Screening, data collection and processing 91

Results 92
Protein purification: wild-type C protein 92
Characterization of the purified WT C protein by SDS PAGE, gel-shift assay and ESI-TOF mass spectrometry 97
Protein purification: selenomethionine C protein 97
Electrophoretic mobility shift assay 101
Complex formation 106
Crystallization of Mu C : DNA complex 110
Screening, data collection and processing 120

Discussion 123

Chapter 4. General Discussion.....131

Major findings..... 131
Binding specificity of Mu transcription activator C 131
Expression, purification, crystallization and preliminary X-ray analysis of C protein bound to P_{sym} DNA 132

Future directions 136
Estimation of dissociation constants (Kd) 136
Analytical ultracentrifugation (AUC) of C : DNA complex 136
Crystallization of truncated C : DNA complex..... 137
Crystallization of C with modified DNA..... 138

List of References.....140

Vita.....159

List of Tables

Table 2-1. Oligodeoxyribonucleotides used for P_{sym} promoter construction 50

Table 2-2. Oligodeoxyribonucleotides used for EMSA 51

Table 2-3. Grouping analysis of mutant promoters with C 58

Table 2-4. Summary of the relative binding efficiency of P_{sym} and P_{sym} mutants altered at –47, –46, –41 and –40 64

Table 3-1. Bacterial strains 79

Table 3-2. Oligodeoxyribonucleotides used for EMSA and crystallization 80

Table 3-3. Buffers for protein purification 85

Table 3-4. Diffraction data statistics of SeMet C : DNA complex crystals 122

List of Figures

Figure 1-1. Domain organization of σ^{70} into various conserved regions with their identified functions. 3

Figure 1-2. Illustration of the σ^{70} subunit and RNA polymerase 7

Figure 1-3. Diagramatic represenation of various stages in transcription initiation. 10

Figure 1-4. Variation of the classical HTH motif 14

Figure 1-5. Composition of a protein crystal 19

Figure 1-6. Different methods of protein crystallization 25

Figure 1-7. X-ray diffraction experiment 29

Figure 1-8. Importance of resolution 30

Figure 1-9. Transcriptional organization of bacteriophage Mu 38

Figure 1-10. Amino-acid sequence alignment of members of the Mor and C family of transcription activators. 41

Figure 1-11. Crystal structure of His-Mor 43

Figure 2-1. Mu C binding region in P_{sym} and the Mu late promoters 48

Figure 2-2. Two-plasmid transcription activation assay system 55

Figure 2-3. Gel-shift assay of P_{sym} with varying DNA length 59

Figure 2-4. Gel-shift assay for P_{sym} mutants altered at –47, –46, –41 and –40 61

Figure 2-5. Gel-shift assay for mutants altered at –53 and –52. 65

Figure 2-6. Gel-shift assay for IR spacer mutants 66

Figure 2-7. Gel-shift assay for mutants altered at -32 69

Figure 2-8. Gel-shift assay to test the relative binding affinity of C to P_{sym} and Mu late promoters 70

Figure 2-9. Summary of the gel-shift assay results 71

Figure 3-1. Circular plasmid map of pZZ41..... 78

Figure 3-2. Expression gel of C protein..... 93

Figure 3-3. WT C purification strategy I. 95

Figure 3-4. WT C purification strategy II..... 96

Figure 3-5. WT C purification. 98

Figure 3-6. WT C ESI-TOF..... 99

Figure 3-7. Electro-mobilty shift assay of C protein presenting different chromatographic fractions 100

Figure 3-8. Selenomethionine C ESI-TOF. 102

Figure 3-9. Gel-shift assay with purified WT C in C buffer with different pH..... 103

Figure 3-10. Gel-shift assay with purified WT C in C buffer with different NaCl concentration..... 104

Figure 3-11. Gel-shift assay with purified WT C in C buffer with different Mg^{2+} concentration..... 105

Figure 3-12. Gel-shift assay with purified WT C in C buffer with P_{sym} and P_{lys} probes of varying length 107

Figure 3-13. Gel-filtration chromatography of C : DNA complex..... 111

Figure 3-14. Comparative gel-filtration elution profile of WT C and C : DNA complex.....…...112

Figure 3-15. Complete list of P_{sym} and P_{lys} DNA used for crystallization. 113

Figure 3-16. WT C P_{sym} 24-mer co-crystal optimization 115

Figure 3-17. WT C P_{sym} 20-mer plus 2 base overlap co-crystal optimization..... 116

Figure 3-18. Different crystallization condition identified for WT C P_{sym} 20-mer plus 2 base overlap..... 118

Figure 3-19. Co-crystallization of native C and selenomethionine C with P_{sym}20-mer plus 2 base overlap in 1.4 M ammonium sulfate 16% ethylene glycol pH 5.7 119

Figure 3-20. Diffraction image of SeMet C : DNA complex at 3.1 Å collected at 17ID of the Advanced Photon Source 121

Figure 3-21. Cα-Polyalanine main chain model of C : DNA complex structure.. ... 127

Figure 3-22. Comparison of C and Mor HTH motif 129

List of Abbreviations

Ap	ampicillin
bp	base pair(s)
Cm	chloramphenicol
DNase I	deoxyribonuclease I
(d)NTP	(deoxy)ribonucleoside triphosphate
DTT	dithiothreitol
HTH	helix-turn-helix
IPTG	isopropyl-β-D-thiogalactopyranoside
PCR	polymerase chain reaction
RNAP	*Escherichia coli* RNA polymerase
SDS	sodium dodecyl sulfate
WT	wild-type
α CTD	the C-terminal domain of the α subunit of the RNA polymerase
α NTD	the C-terminal domain of the α subunit of the RNA polymerase
β-ME	β-mercaptoethanol

Chapter 1. Introduction

Gene expression is one of the important classes in the central dogma of molecular biology. In gene expression, the information from the DNA coding region is converted into a functional protein in a multi-step process. Transcription is the first step in gene expression in which the information in the DNA is transcribed by DNA-dependent RNA polymerase (RNAP) into messenger RNA (mRNA). The mRNA is then translated by the ribosomes to produce polypeptides. This process can be regulated or modulated in each step, giving the cell or the organism a control over its morphology, differentiation and function.

An overview of transcription

Transcription proceeds in the 5′ to 3′ direction using the DNA as template and is divided into three sequential stages; initiation, elongation and termination. Transcription initiation is a multi-step process characterized by binding of the sigma (σ) subunit of the RNAP holoenzyme to the promoter (P) to form a closed binary complex (RPc). Subsequent melting of the DNA strands along with a conformational change in the RNAP leads to the formation of an open complex (RPo). Formation of RPc and RPo are rate-limiting steps that can be modulated by activators and repressors. RPo in the presence of the four nucleotide triphosphates proceeds to an initiated complex (RP init) which can be engaged in an iterative abortive initiation process, generating and releasing short nascent RNA chains less than 10 nt long. Abortive initiation terminates and initiation is complete when RNAP breaks contacts with the promoter, releases the σ

subunit, and escapes as a productive elongation complex. Termination of transcription may occur when (1) the RNAP encounters a transcription terminator involving formation of a mRNA hairpin and (2) by a protein called Rho, leading to a process called Rho-dependent termination. Regulation of transcription is usually done by regulatory proteins, which can be (1) specificity factors (σ), (2) repressors and (3) activators. Regulation is only achieved when these regulatory proteins bind specific DNA sequences using different DNA-binding motifs.

RNA polymerase

In prokaryotes, two forms of RNAP are involved in transcription. One form of the RNAP, termed holoenzyme, is required for transcription initation and is specific to the promoters (Figure 1-1). The specificity is due to the σ subunit of the RNAP and is formed when σ binds the core. The second form called the core RNAP (Figure 1-1) binds DNA nonspecifically since it lacks the σ subunit; is incapable of initiating transcription from the promoter but capable of elongating an RNA transcript. The core enzyme is composed of five subunits ($\alpha 1$, $\alpha 2$, β, β', ω) with a combined molecular mass of ~ 400 kDa (Borukhov and Severinov, 2002).

The sigma (σ) subunit

Sigma is a DNA-binding specificity factor that recognizes the promoter upstream of the protein-coding region and is very important for transcription initiation (Darst *et al.*, 1998; Zhang *et al.*, 1999; Tahirov *et al.*, 2002). There are seven different species of the σ subunit identified in *Escherichia coli* (*E.coli*) (Gross *et al.*, 1998; Helmann and

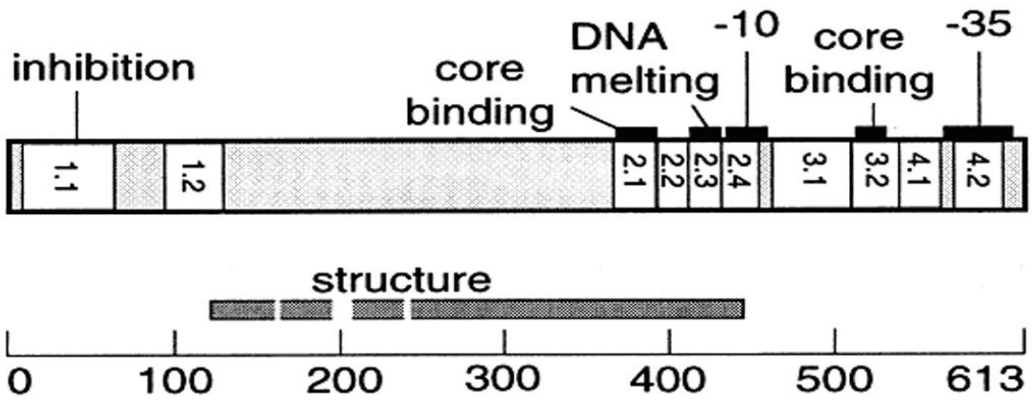


Figure 1-1. Domain organization of σ^{70} into various conserved regions with their identified functions. Reprinted with permission from Owens, J. T., Miyake, R., Murakami, K., Chmura, A.J., Fujita, N., Ishihama, A., and Meares. C.F. (1998) Mapping the σ^{70}subunit contact sites on *Escherichia coli* RNA polymerase with a σ^{70}-conjugated chemical protease *PNAS 95*: 6021-6026

Chamberlin, 1988; Ishihama, 1988). Each sigma is involved in transcription of a specific set of genes. Most of the housekeeping genes, which are expressed during the exponential growth phase, are transcribed by the holoenzyme containing σ^{70} (*rpoD* gene product), the principal sigma factor (Ishihama, 2000). The *rpoH* gene product σ^{32} is expressed when the bacterial cells are exposed to high temperatures in order to turn on transcription of genes required for surviving such temperatures. The *rpoS* gene product σ^{S} is expressed when the bacterial cells are in stationary phase. The *rpoF* gene product σ^{F} is expressed to turn on genes involved in flagellum synthesis. The *rpoE* gene product σ^{E} is expressed during heat shock, oxidative stress and for expression of extracytoplasmic proteins. The *fecI* gene product σ^{FecI} regulates the *fec* genes for iron dicitrate transport. The *rpoN* gene product σ^{54} is a distinct class of sigma, which is structurally and chemically different from members of the σ^{70} family. This sigma is present at all times in the bacterial cell and is important for turning on genes involved in nitrogen regulation. There are four flexible domains in σ^{70} referred to as $\sigma_{1.1}$, σ_2, σ_3 and σ_4 (Helmann and Chamberlin, 1988) (Figure 1-1). Each domain contains one or more highly conserved regions; 1.1 and 1.2, 2.1-2.4, 3.1-3.2 and 4.1- 4.2 (Malhotra *et al.*, 1996; Gross, 1998; Borukhov and Severinov, 2002; Campbell *et al.*, 2002; Murakami *et al.*, 2002; Vassylyev *et al.*, 2002; Murakami and Darst, 2003). The σ^{70} makes sequence-specific contacts within the –10 hexamer (consensus TATAAT) and –35 hexamer (consensus TTGACA) of the promoter using region 2.4 and region 4.2 only when σ^{70} is bound to the core RNAP. The σ^{70} cannot bind the promoter on its own since region 1.1 masks region 2.4 and region 4.2; and binding inhibition is only released after sigma binds the core RNAP. The σ^{70} can also direct transcription from extended –10 promoters that do not have a good –35 hexamer

(Helmann and Chamberlin, 1988). Region 2.5 (not shown) makes supplementary contact with the extended –10 motif (Barne *et al.*, 1997). Region 4 is important for contacting the –35 hexamer and interacting with the α-CTD and transcription activators.

The alpha (α) subunit

The α subunit is made up of 329 amino-acids and is structurally divided into two independent domains (α-NTD and α-CTD). The α-NTD (28 kDa) plays an important role in assembling the RNAP by providing the contact surfaces for α dimerization and for binding the β and β′ subunits (Kimura *et al.*, 1994; Kimura and Ishihama, 1995a, b, 1996) of the RNAP. A flexible linker made up of 13 amino-acids connects the α-NTD and α-CTD. This long unstructured linker allows the α-CTD to occupy different positions relative to the α NTD and RNAP (Blatter *et al.*, 1994; Busby and Ebright, 1994; Ebright and Busby, 1995). The α-CTD (8 kDa) plays a regulatory role by providing the contact surfaces for trans-acting regulatory protein factors and cis-acting DNA elements (Igarashi *et al.*, 1991; Igarashi and Ishihama, 1991; Gaal *et al.*, 1996; Murakami *et al.*, 1996). The α CTD binds the DNA minor groove upstream of the –35 hexamer using a helix-hairpin-helix (HhH) DNA binding motif (Jeon *et al.*, 1995; Gaal *et al.*, 1996; Shao and Grishin, 2000; Ross *et al.*, 2001); this motif is called the 265 determinant (Gaal *et al.*, 1996; Murakami *et al.*, 1996)**.** The α-CTD interacts with various transcription factors using the 287 determinant, which is made up of eight different amino-acid residues.

The β and β′ subunits

The core RNAP is made up of two α subunits, one β (150.6 kDa, *rpoB* gene product) and one β′ (155.2 kDa, *rpoC* gene product) subunit along with a smaller more-loosely bound subunit called omega (ω). The β and β′ subunits are the two largest subunits of the RNAP and form a catalytic core around a chelated Mg^{2+} using conserved regions A to I in β and region A to H in β′ (Figure 1-2). The β subunit contacts α1 while β′ makes contacts with α2 (Zhang *et al.*, 1999). The β and β′ subunits play a key role in transcription since they form the active site for polymerization and interact with activators, repressors, elongation factors and termination factors (Severinov *et al.*, 1994; Sharp *et al.*, 1999). The ω subunit is the smallest subunit and has a chaperone-like role in RNAP assembly (Mukherjee *et al.*, 1998).

Overall structure of RNAP

The bacterial RNAP holoenzyme structure has been solved from *Thermus aquaticus* and *Thermus thermophilus* (Murakami *et al.*, 2002; Vassylyev *et al.*, 2002). The core RNAP resembles the claw of a crab (Figure 1-2). The two sides of the claw are made up of β and β′, which also form the active site channel with a diameter of 27 Å. The entire length of the active-site channel is occupied by σ, and the negatively charged amino-acids of $\sigma_{1.1}$ and $\sigma_{3.2}$ occupy the upper and lower portions of the active-site channel, respectively (Mekler *et al.*, 2002; Murakami *et al.*, 2002; Vassylyev *et al.*, 2002). These charged amino-acids mask the positively-charged portion of the channel, thus narrowing the channel and preventing nonspecific binding of RNAP to DNA.

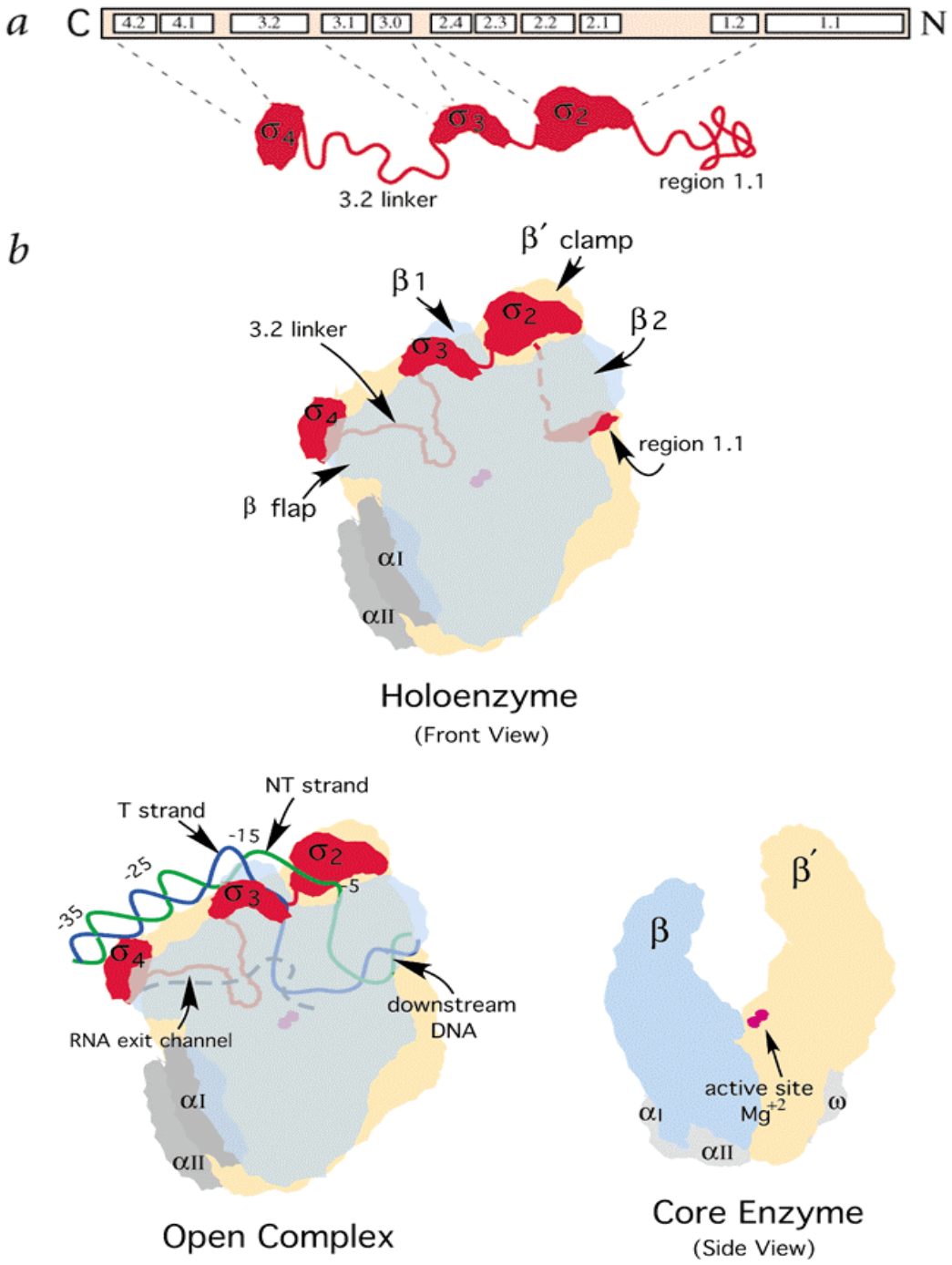


Figure 1-2. Illustration of the σ^{70} subunit and RNA polymerase. The σ^{70} subunit is colored red, the β subunit is blue, and β′ is pink. (a) A diagrammatic representation of σ^{70} depicting its conserved domain organization. (b) different forms of RNAP. Reprinted with permission from Hsu, L.M. (2002) Open season on RNA polymerase. *Nat Struct Biol* 9: 502-504.

Regions 2.4 and 4.2 of σ are solvent exposed and make specific contacts with the –10 and –35 hexamer of the promoter. Region 3.2 is a linker, connecting regions 3 and 4; it loops around the RNAP active-site channel and exits through the RNA-exit channel.

Promoter

The consensus sequence for promoters from *E. coli* has a –10 (5' TATAAT 3') and –35 hexamer (5' TTGACA 3') separated by a 16-18 bp spacer with 17 bp being the most optimal (Youderian *et al.*, 1982; Harley and Reynolds, 1987; Helmann and Chamberlin, 1988; Lisser and Margalit, 1993). The hexamers are only contacted by the σ subunit of the RNAP holoenzyme (Thomas *et al.*, 1996; Gross *et al.*, 1998). In some promoters an AT-rich region known as the UP element is present just upstream of the –35 hexamer; and this region contains two distinct subsites (proximal and distal), which are recognized by the α-CTD (Ross *et al.*, 1993; Gaal *et al.*, 1996; Estrem *et al.*, 1998). The promoter strength is a function of the entire promoter, with very strong promoters like the *rrn*B promoter having sequences close to the consensus. The start point of transcription (+1) in *E. coli* promoters is occupied by a purine nucleotide, with 'A' preferred over 'G' in the vast majority of the characterized promoters.

There is an additional class of promoter known as the "extended –10 promoter" in which the –35 recognition elements of σ are dispensable for transcription initiation. These promoters are distinguished by the presence of a new consensus sequence in the –10 region (TGnTATAAT) (Keilty and Rosenberg, 1987; Chan and Busby, 1989).

RNAP-activator interactions

Transcription from most bacterial promoters can be modulated by regulatory proteins. In most cases regulatory proteins bind within or upstream of the core promoter and may have direct contact with RNAP. Transcription activation involves protein-protein interactions or the presence of an activator protein in close proximity to the RNAP with an intact DNA duplex between the two binding sites (Kustu *et al.*, 1991; Ebright, 1993). An important contact point in RNAP for these activator proteins is the α-CTD. Based on the positioning of an activator protein on the promoter and contacts made by the α-CTD and σ-CTD the transcription complexes are classified as basal, UP element-dependent and activator dependent. These complexes are divided into Class I/II/III or mixed promoters (Busby and Ebright, 1994, 1999)

Transcription initiation

Transcription initiation is a complicated multi-step process that involves (1) promoter recognition by the RNAP to form a closed complex (2) formation of a competent initiation complex (open complex), followed by production of short abortive transcripts and (3) promoter clearance of RNAP as it enters the elongation phase. Due to the multiple steps involved, there are many targets for regulation. The sigma subunit of the RNAP holoenzyme recognizes the –10 and the –35 hexamers to form an inactive, unstable, closed complex (RPc) (Figure 1-3). The RPc extends from –55 to –10 in a sequence-specific manner in which the DNA is protected from DNAse and hydroxy radical cleavage by the holoenzyme (Pavletich and Pabo, 1991; Marmorstein *et al.*, 1992; Fairall *et al.*, 1993; Marmorstein and Harrison, 1994;. Raumann *et al.*, 1994)

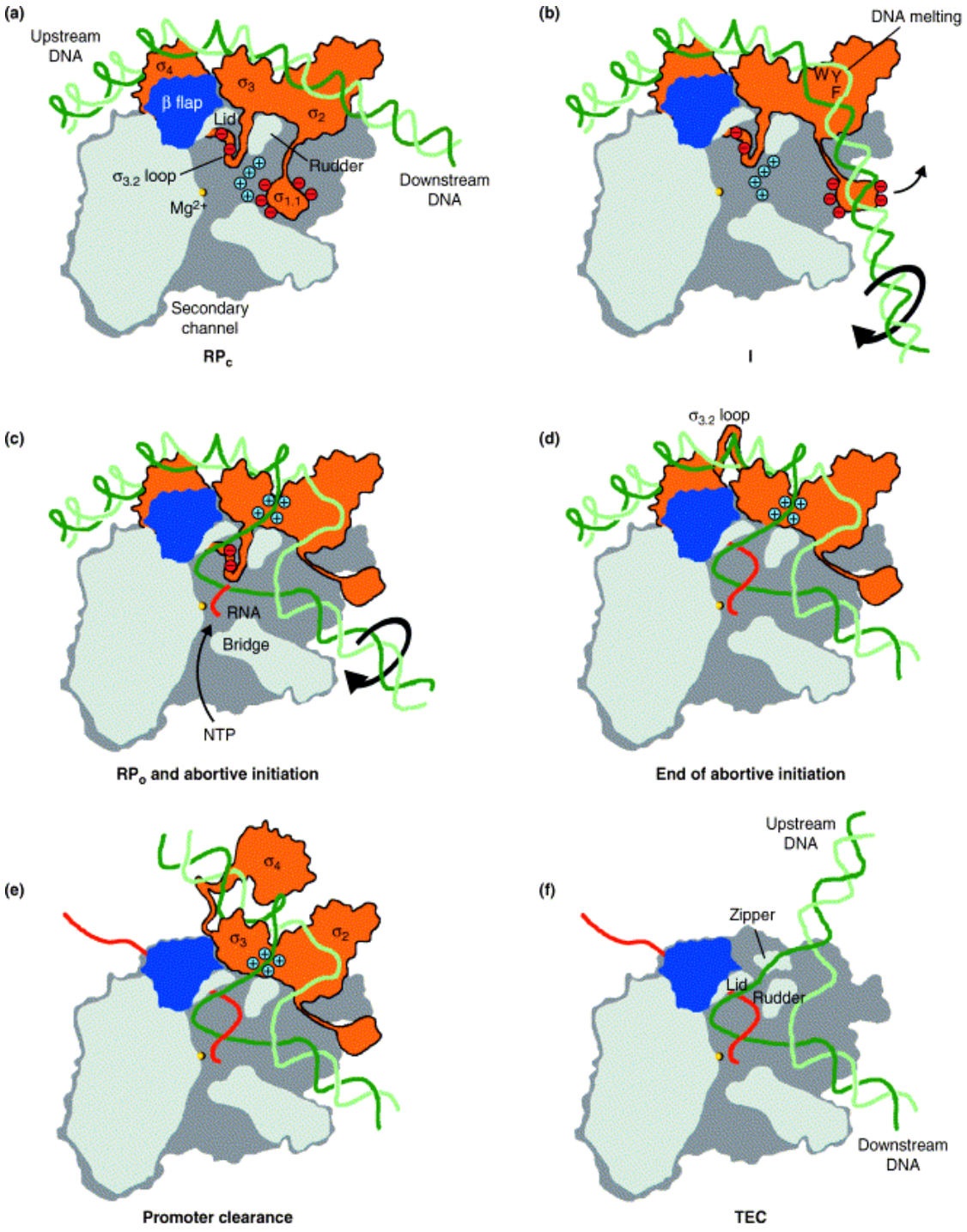


Figure 1-3. Diagrammatic representation of various stages in transcription initiation. (a) closed complex, (b) intermediate complex, (c) and (d) open complex leading to abortive initiation and subsequent ending of abortive initiation, (e) promoter clearance, and (f) transcription elongation complex (TEC). Reprinted with permission from Murakami, K.S., and Darst, S.A. (2003) Bacterial RNA polymerases: the wholo story. *Curr Opin Struct Biol* 13: 31-39.

(Figure 1-3). The RPc is sensitive to the DNA mimic heparin that competes with DNA for RNAP binding The second step is the transition to an intermediate complex (Buc and McClure, 1985) (Figure 1-3). The intermediate complex is a stressed complex formed due to a change in the conformation of the RNAP and DNA. The –35 and –10 regions of the DNA are rotated relative to each other by untwisting of the spacer DNA followed by wrapping of the DNA around RNAP, a process called open complex formation (RPo) (Figure 1-3). The open complex is resistant to heparin and base pairs between positions –9 to +2 in the promoter DNA are opened uni-directionally to form the transcription bubble (Saucier and Wang, 1972; Kirkegaard *et al.*, 1983). The template strand of the DNA is located near the active site of the RNAP and lies in the positively-charged active site channel (Fogh *et al.*, 1994). Upon addition of nucleotide triphosphates the transcriptionally competent open complex RPo starts to synthesize the RNA. During this process the $\sigma_{3.2}$ loop is present in the path of the extending RNA–DNA hybrid, which results in abortive initiation (Figure 1-3). This is a recurring process until the nascent RNA reaches a critical length of ~ten nucleotides. When the transcript is more than ten nucleotides, it is sufficiently long enough to exit from the RNA exit channel under the β flap, resulting in complete displacement of the $\sigma_{3.2}$ loop. This displacement and the continuing RNA synthesis leads to destabilization of the σ_4 interactions with the –35 region, allowing RNAP to break free from the promoter. A stable transcription elongation complex (TEC) is formed with release of the sigma subunit (Figure 1-3). In most promoters, the presence of RNAP and transcription activators is sufficient for RPo formation, while in others the addition of NTPs or formation of the first phosphodiester bond is required (Newlands *et al.*, 1991; Ohlsen and Gralla, 1992).

Regulation of transcription

Regulation of transcription in prokaryotes is based on the environment detected by the cell. This regulation is mainly achieved by regulatory proteins called transcription factors, which may or may not interact directly with the RNAP. Depending upon their mode of action, the transcription factor can be either a repressor (negative regulator) or an activator (positive regulator) (Ptashne, 2004). To interact with the RNAP the regulatory protein should first bind to its cognate DNA binding site using one of many different DNA binding motifs. The most common motif is the helix-turn-helix (HTH), which was first, identified in Cap, Cro and λ repressor proteins using structural and sequence similarities (Matthews *et al.*, 1982; Ohlendorf *et al.*, 1982; Sauer *et al.*, 1982; Steitz *et al.*, 1982; Ohlendorf *et al.*, 1983). In prokaryotes, this motif is commonly found in regulatory proteins (repressors and activators) and sigma factors (Landick *et al.*, 1984; Yura *et al.*, 1984; Gribskov and Burgess, 1986). In eukaryotes the HTH motif is seen in proteins that regulate development and differentiation, transcription factors (TFIIB/ TFIIE) and chromatin proteins (histone H1) (Schultz *et al.*, 1991; Wilson *et al.*, 1992 Brennan, 1993; Clark *et al.*, 1993; Ramakrishnan *et al.*, 1993; Swindells, 1995; Kodandapani *et al.*, 1996; Aravind and Koonin, 1999; Gajiwala and Burley, 2000). This motif has also been identified in proteins involved in DNA repair and RNA metabolism (Moore *et al.*, 1994; Wah *et al.*, 1997; Selmer and Su, 2002; Alfano *et al.*, 2004; Dong *et al.*, 2004). In addition to DNA binding, the HTH can also be adapted for mediating specific protein-protein interactions or can be part of a structural unit of a large enzymatic domain (Guo *et al.*, 1997; Wong *et al.*, 2000; Zheng *et al.*, 2002).

The HTH motif consists of three α helices (one, two and three), which are usually right, handed and interconnected to each other by linkers. Helix one is the scaffolding helix for positioning of helices two and three (recognition helix) which form the HTH motif, with the linker in between making a 120° turn (Figure 1-4). The linker is made up of three or four amino-acids and does not tolerate insertions or deletions, but the loop between helix one and helix two may vary depending upon which class of the HTH family the protein is in. The third helix "recognition helix" makes base- specific contacts in the major groove and forms the principal DNA-protein interface (Brennan and Matthews, 1989; Otting *et al.*, 1990; Brennan, 1993).

There are several conserved features in the HTH fold, and the most distinct is the 'shs' pattern (Aravind *et al.*, 2005) in the turn between helix two and three. The 's' denotes a small unbranched amino-acid like glycine and 'h' denotes a hydrophobic residue. The other conserved feature called the 'phs' pattern is present in helix two in which 'p' is a charged residue. The core tri-helical bundle is stabilized by localization of hydrophobic interactions between conserved hydrophobic residues in helix one, two and three to form a distinct hydrophobic core (Figure 1-4).

There are a number of variations to the classical HTH which may include (1) a longer turn with different angles with and without rearrangements of the three helix bundle (Assa-Munt *et al.*, 1993; Clark *et al.*, 1993; Donaldson *et al.*, 1994; Fogh *et al.*, 1994; Harrison *et al.*, 1994; Liang *et al.*, 1994; Schumacher *et al.*, 1994; Vuister *et al.*, 1994), (2) a different conformation as in the *c-myb* proto-oncogene (Ogata *et al.*, 1992) and hepatocyte transcription factor 1fb1/hnf1 (Finney, 1990) and (3) topological variation in the HTH in which the HTH may be completely alpha helical or beta strands may form

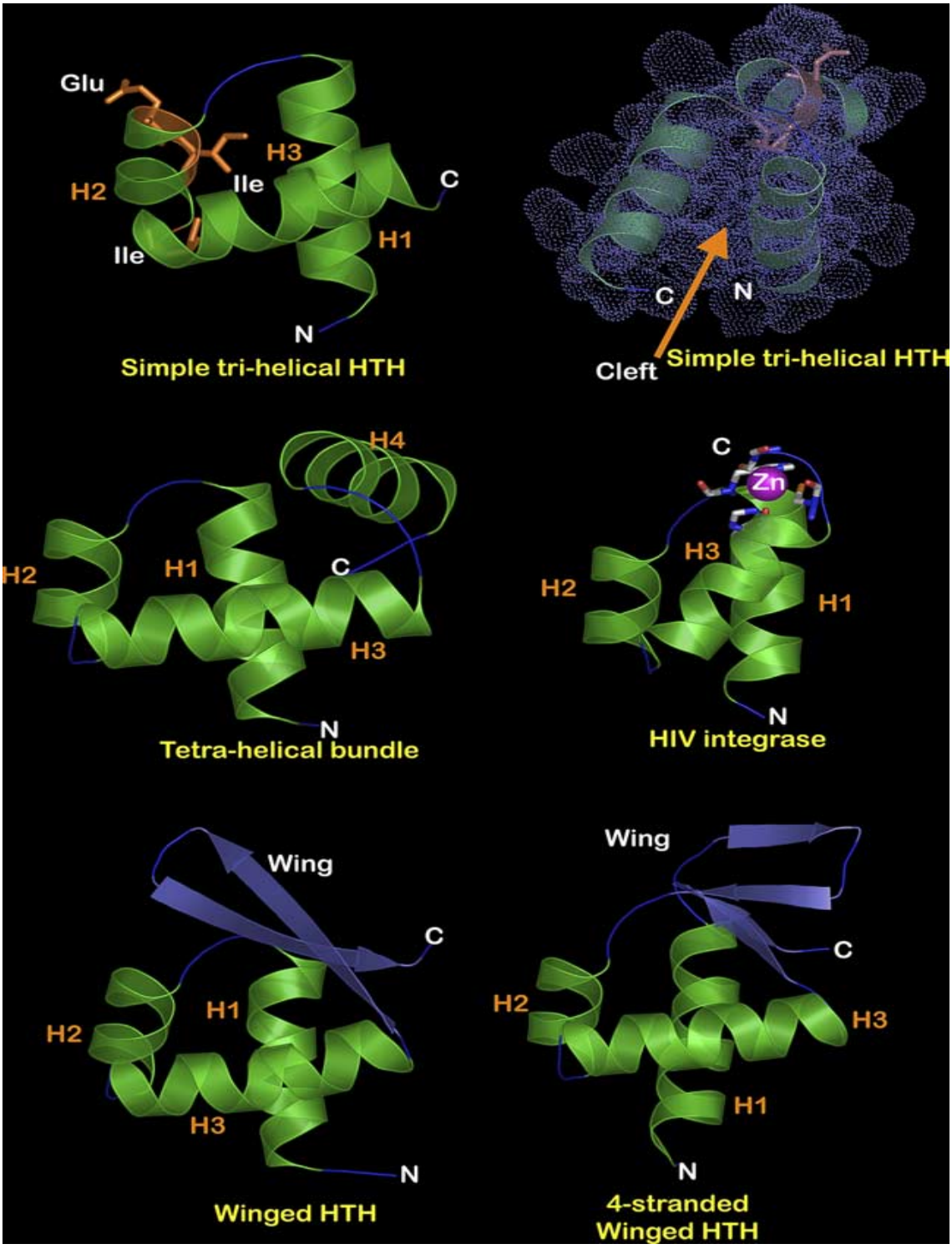


Figure 1-4. Variations of the classical HTH motif. Reprinted with permission from Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., and Iyer, L.M. (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev* 29: 231-262.

an antiparallel beta sheet which may interrupt the motif by preceding or following the helices involved in DNA binding (Brennan, 1993; Lai *et al.*, 1993: Clubb *et al.*, 1994). Based on the distinctive features the HTH-containing proteins can be classified into two major structural classes and a very distinct third class that has drastic alterations to the HTH core.

In class I of the HTH-containing proteins in addition to the basic tri-helix there may be tight packing of the recognition helix by means of Zn^{++} as seen in the retroviral integrase (Figure 1-4) (Cai *et al.*, 1997). In certain proteins such as AraC, TetR, TrpR, there is an additional C-terminal helix, which forms a tetra-helical version of the HTH motif (Figure 1-4). In some eukaryotic basal-transcription factors, the tetra-helical HTH is further modified to form a multi-helical HTH. There are also rare modifications of the basic tri-helix in which there are additional N-and C-terminal helices with different packing than that seen normally in tetra-helical forms, e.g. KorB-like HTH (Khare *et al.*, 2004) and Fihb-like HTH (Campos *et al.*, 2001)

The class II HTH is mainly composed of the winged HTH (wHTH) as shown in Figure 1-4. The wHTH was based on the structure of the hepatocyte nuclear factor (HNF)3/forkhead DNA binding domain (Clark *et al.*, 1993; Feng *et al.*, 1994). In addition to the HTH, there are two beta-strand hairpins that make DNA backbone contacts. These hairpins follow the HTH region and provide additional charged residues to interact with the minor groove bases adjacent to the major groove (Brennan, 1993; Clark *et al.*, 1993; Swindells, 1995). There are also several modifications of the beta sheet, which may include three-or four-strand versions (Figure 1-4). In prokaryotes, the two-or three-strand versions predominate.

The third class of HTH includes proteins from both class I and II but the proteins are highly modified, as observed for the Met I - Arc family of transcription factors. This family is also referred to as the ribbon-helix-helix (RHH family). This RHH family is usually a dimer and has a C-terminal HTH which is similar to the canonical HTH (Gomis-Ruth *et al.*, 1998). The N-terminal region forms a strand that is used for dimerization as well as making contacts in the DNA major groove (Gomis-Ruth *et al.*, 1998). One interesting aspect of the N-terminus is that a single mutation can convert the sheet to a helix, which then forms a motif similar to the conventional HTH.

In the wHTH of the bacterial transcription regulator MerR, helix one of the HTH motif is not present. The topoisomerase II family has two copies of the wHTH domain, and in one there is a beta sheet as well as a wHTH insertion between helix one and helix two; the beta sheet contacts the DNA by forming a brace around the DNA (Grishin, 2000). The Oct1-Pou DNA complex consists of two DNA binding domains; (1) the Pou-specific domain in which the HTH is similar to that of the lambda repressor (Assa-Munt *et al.*, 1993; Dekker *et al.*, 1993) and (2) the Pou homeodomain which is similar to the eukaryotic homeodomain (Kissinger *et al.*, 1990; Wolberger *et al.*, 1991 Qian *et al.*, 1994). Both domains are tethered by only a linker (Klemm *et al.*, 1994) with no other protein-protein interactions between them.

The Hin recombinase protein contains an HTH motif as well as N-and C-terminal arms that contact the minor groove (Feng *et al.*, 1994). The important difference is the position of the HTH, which is midway between that seen for the lambda repressor and the homeodomain. The N-terminal arm is similar to the HTH recognition arm since it makes base-specific contacts. The HTH motif of the PurR repressor is different from the

classical HTH because the PurR HTH motif is formed from helix one and two, with its orientation reversed. This orientation reversion is also seen in the solution structure of the LacI repressor bound to DNA as well as in the Tet repressor (Hinrichs *et al.*, 1994). The PurR repressor also uses the fourth helix known as the 'hinge helix' to make base-specific interactions in the minor groove (Schumacher *et al.*, 1994). Dimerization of the PurR monomer is facilitated by the hinge helix making hydrophobic interactions.

X-ray crystallography

A complete understanding of biology will require three-dimensional (3D) structures of individual components as well as macromolecular complexes. Structures of biological macromolecules allow us to study how macromolecules interact with each other, what their mechanism of action is at the atomic level, and how the structure can be used for structure based-drug development (Smyth and Martin, 2000). Three dimensional structures are obtained from a wide array of techniques such as X-ray crystallography, nuclear magnetic resonance (NMR), cryo electron microscopy (CEM) and atomic force microscopy (AFM) to name a few.

X-ray crystallography is often used to obtain detailed information on the structure and function of biological molecules. In bio-molecular X-ray crystallography, the first step is to purify the macromolecule (e.g. protein, DNA) to high purity. The purified macromolecule is then crystallized. This process can be extremely problematic since protein crystals are difficult to grow, difficult to transport, and deteriorate rapidly. Several hundred must be grown in order to have one or two high-quality crystals. The crystals obtained are screened by exposing them to an X-ray beam to obtain a diffraction

pattern, which is processed to reveal information about the crystal packing symmetry and the size of the repeating unit in the crystals. Then the intensity of the diffraction pattern is used to determine the "structure factors" from which the electron density map is calculated. The quality of the map is improved using established refinement techniques so that a structure can be built using an available protein and/or nucleic acid sequence. The resulting structure from the map is refined so that it is in a very thermodynamically-favored and accurate conformation.

Principles of X-ray crystallography

In crystallography, crystals are used because X-ray diffraction from a single molecule is weak and would not be detected above the background noise. A crystal by definition is a homogeneous solid mass having a high degree of internal order with a specific overall chemical composition. Due to this internal order the scattered X-ray waves add up in phase, raising the signal to a measurable level. Thus, for the crystal to act as an amplifier it should be of the highest quality (Figure 1-5), which is measured in terms of diffraction quality. A good crystal has a clear diffraction pattern with a good distribution of peak spots. X-rays have several characteristic wavelengths, and the one used for crystallography is called CuKα, which has a wavelength of 1.5418Å. This wavelength is used to study molecular structure because it is similar to the distance between two carbon atoms in a molecule. The consequence for an X-ray beam hitting a crystal is scattering of the X-ray beam as a function of incident and scattered angle, polarization and energy of wavelength, i.e. the result is the diffraction pattern, which is an array of spots from the electron density of all the atoms. The position and intensity of

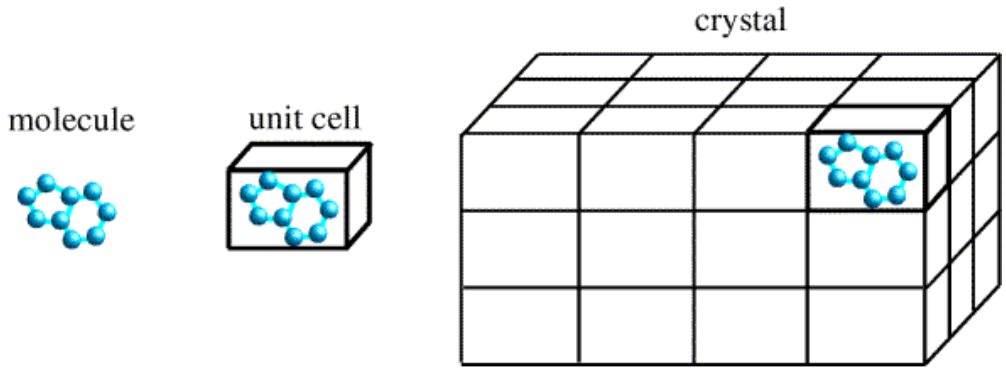


Figure 1-5. Composition of a protein crystal. Reprinted with permission from Randy, J.R. University of Cambridge. http://www-structmed.cimr.cam.ac.uk/Course/Overview/Overview.html. Accessed April 6, 2007.

each spot is very important since it contains the information about the size of the unit cell and where the atoms lie with respect to each other in 3D space. The spots are two-dimensional and represent a wave with amplitude and a relative phase. In reality, only the numbers of photons that are reflected from the crystal are measured and any information about the relative phase of the different diffractions is lost; this is called the (phase problem). The problem arises because the X-rays start out with a dispersed phase, and deconvolution of the phase in the detector where the spots are collected is presently impossible. The diffraction pattern of a crystal depends on a number of factors. The size of the crystal is important because the scattering of the X-ray is directly proportional to the number of unit cells and how ordered they are in the crystal. The total number of unit cells in a crystal is directly proportional to the total volume of the crystal, so a simple doubling of the crystal dimensions will increase the diffraction by eight fold. The total number of unit cells also depends on the crystal size, which is directly correlated with the size of the protein, how many protein molecules are present in the asymmetric unit, and finally the symmetry within the crystal.

To solve a new structure one must obtain the phase angle of the structure to be determined. As mentioned earlier, the phase angles are lost during the data collection process so they cannot be determined directly. However, they can be resolved indirectly through two approaches. The traditional, more conservative approach is to obtain a crystal similar, to the one being studied, except that a few atoms have either been added or replaced. If the replaced atoms are heavy, i.e. having a larger atomic number, they will disturb the diffraction pattern, a process called anomalous scattering. This perturbation may be enough to deduce the positions of the heavy atoms, and from knowing the

positions it may be possible to deduce the phase angles. This technique is called multiple isomorphous replacement (MIR). A similar technique called multiple-wavelength anomalous dispersion (MAD) can also be used. In MAD, data is collected from two different wavelengths corresponding to the peak and inflection of the heavy atoms present in the crystal. This is done by tuning the X-ray source. The difference in the perturbation of the diffraction pattern gives the same kind of information as in MIR. A more recent method is called molecular replacement (MR). MR takes advantage of a previously solved structure that is structurally or chemically similar to the structure being solved. The solved structure is used as a search model to guess the orientation and position of the molecules in the unit cell of the test structure. The phases thus obtained can be used to generate the electron density map. Finally, if the data generated from the native crystal is of very high resolution i.e. better than 1.6 Å or 160 picometers, the structure can be solved directly.

The phase obtained is then used to build the initial model in the electron density map. The electron density map contains the entire peak at each of the atomic positions and more typically tubes of electron density of the atoms that are bonded together. This is because proteins are such dynamic molecules that the unit cells are not aligned properly in the crystal, which is very noticeable when finer details are examined. A detailed protein structure is dependent upon how many atoms can be resolved from one another. Thus, higher resolution is critical if finer details are required.

The initial model built by fitting an individual molecule to the electron density is refined, taking into account the thermal motion of the atoms and the best fit to the observed diffraction data. The process of refining generates a new set of phases and a

new electron density map. This process of refining and model building is carried out until there is the highest possible correlation between the diffraction data and the model. The R-factor (residual or agreement factor) is simply the average functional error of the calculated amplitude to the observed amplitude. Amplitude is the magnitude or "intensity" of the X-ray waves. The final solved structure should have an R-factor in the range of 15% to 25%.

Crystallization

The first and most important rate-limiting step in macromolecular structure determination by X-ray crystallography is the crystallization of the macromolecule of interest. It is important to produce a crystal big enough (> 0.1 mm in its smallest dimension) in which the molecules are arranged in precise order so that the X-ray beam will be able to diffract and create a well-defined pattern of reflections that can be used for creating the electron density map. Crystallization involves the separation of macromolecules from the liquid phase into the solid phase; it involves two important processes; “nucleation” and “crystal growth”. Both these processes occur only in a supersaturated solution where the concentration of macromolecules is far beyond their equilibrium solubility value. Crystallization conditions are chosen to promote crystal formation as compared to random aggregation of the protein (referred to as precipitation). Nucleation in a solution occurs when there is a sufficiently elevated solute concentration in a particular region to form stable clusters of molecules. The stability of the clusters is determined by a variety of factors including the concentration, temperature and presence of contamination. This process is very important since at this point of crystal growth,

atoms in the macromolecules are internally arranging in an ordered fashion. Nucleation is followed by either crystal growth or more nucleation, and this is dependent upon what the condition favors. The result of the crystallization trial might be a single crystal or multiple crystals that vary in size and shape. Even if supersaturation is attained, the requirements for nucleation and crystal growth are different. At a high supersaturation level both nucleation and growth occurs; at a lower level only the growth of the crystal is seen. Supersaturation that facilitates both nucleation and crystal growth can be obtained by a number of ways, (1) cooling the protein solution, (2) reducing the solubility of the macromolecule by the addition of a second solvent, (3) increasing the concentration of the solvent, (4) increasing the protein concentration, (5) changing the pH (6) changing the crystallization technique, and (7) addition of an additive. Crystals grow until the solution is no longer at supersaturation and the solid-liquid equilibrium is reached.

Methods for protein crystallization

There are many methods for protein crystallization, such as vapor diffusion, batch crystallization and the dialysis method to name a few. The vapor diffusion method is by far the most effective and popular. This technique involves equilibration in a closed chamber of a drop containing the protein, protein/DNA complex, buffers and precipitants at a lower concentration insufficient to precipitate, with a much larger reservoir containing precipitant and a dehydrating agent. This equilibration supersaturates the protein in the drop by removing solvent, and if the condition is right, nucleation may occur (Weber, 1991)**.** Vapor diffusion is a convenient method because it uses a small volume, set-up is straightforward, and the results of each experiment can be quickly

checked. Vapor diffusion can be done in either a hanging drop or a sitting drop. In the hanging drop method the protein drop is placed on a coverslip, which is then inverted and used to seal the reservoir in a linbro plate (Figure 1-6). The sitting drop method uses a much smaller sample than the hanging drop method. In the sitting drop method, the drop is placed on a depression in either a microbridge in a linbro plate or a glass plate and put into a closed system to equilibrate against a larger reservoir (Figure 1-6). In batch crystallization all components including protein, buffer, and precipitants are combined together as shown in Figure 1-6. In this method, supersaturation required for nucleation is achieved on mixing and, if the condition is right, nucleation may occur. Since supersaturation is achieved very rapidly in this method, a large number of nuclei may be rapidly formed which may lead to many small poor-quality crystals. Alternatively, rapid nucleation may begin with a select few crystals that grow slowly and form better diffracting crystals.

In the dialysis method the concentration of the macromolecules remains constant, but due to diffusion, the solution composition is altered considerably (Figure 1-6). This method can be used for crystallization of protein at low and high ionic strength. This is a very versatile method of crystallization because the protein and other components can be maintained at a constant level while changing the pH and precipitant concentration. This technique can be employed for proteins known to have lower solubility at low ionic strength since, when the concentration of the ions are reduced, the macromolecules tend to stabilize themselves through interaction with other molecules (McPherson, 1990); this phenomenon is traditionally referred to as “salting in”. Depending on the availability of the protein, one of the following dialysis methods can be used; (1) equilibrium dialysis

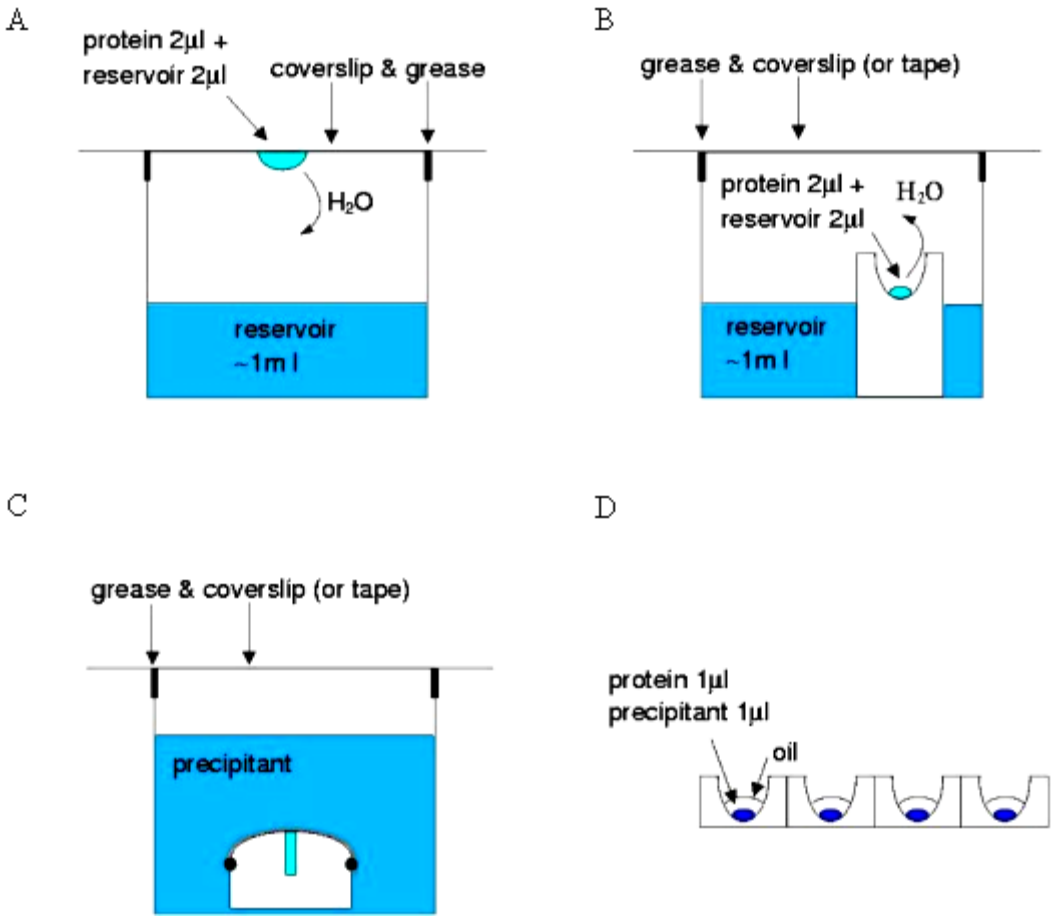


Figure 1-6. Different methods of protein crystallization. (A) hanging drop method, (B) sitting drop method, (C) dialysis method, and (D) batch method. Reprinted with permission from Airlie, J.M. University of Cambridge. http://www-structmed.cimr.cam.ac.uk/Course/Crystals/ Theory/methods.html. Accessed April 6, 2007.

with dialysis tubes, (2) equilibrium dialysis with micro-dialysis (3) and equilibrium dialysis with acrylamide plugs.

Screening methods

A good crystal is identified through a process called screening. There are many factors, such as pH, temperature, precipitant, ionic strength, and mono and/or divalent ions that influence crystallization. Screening is a search within these variables for an optimum crystallization condition to obtain a good diffraction-quality crystal. There are many screening methods, including full factorial, incomplete factorial, random, and sparse matrix. In the full factorial method all the parameters are taken into consideration and sampled. The problem with this method is that it is time consuming, since it requires a large number of crystallization trials and involves a large amount of sample. In the incomplete factorial screening method, a small number of factors are chosen rationally and tested evenly and efficiently. The effect of each factor is scored and evaluated. The results obtained are used to restrict the search criteria for optimal crystal growth. This method is a powerful tool since it reduces the number of crystallization trials required to identify the different variables (Carter and Carter, 1979). Random screening is similar to incomplete factorial screening except that the parameters selected for screening are purely random. The most commonly used method of screening is the sparse matrix, which was designed by utilizing conditions that have worked previously for other macromolecules (Jancarik and Kim, 1991).

Optimization

The conditions, which favor nucleation and crystallization are used for preliminary optimization. In this step the buffer, pH, precipitant, protein concentration, crystallization method and additives are compared with each other by always keeping one of the parameters constant and varying another. From these comparisons, the important parameter that reproducibly drives the nucleation, crystallization and reproducibility can be identified. The parameters identified from preliminary optimization are refined until an optimal crystallization condition is found. The goal of optimization of crystallization condition is to get a good diffraction-quality crystal independent of its size and morphology.

If the outcome of the optimization does not yield a good crystal, then the best possible way to obtain a structure is to try the following; (1) a ligand protein complex; this can help because the binding of the ligand orders the region of the protein that bind it or the ligand may bring a subdomain of the structure together and reduce flexibility, changing the surface properties of the protein, or finally causing a conformational change in the protein, (2) different constructs; this strategy is based on constructing sequential N- and C-terminal truncations of the gene of interest so that during crystallization the flexible features of the protein are not present and (3) different protein species such as point mutants.

Data collection and processing

The crystals once prepared are harvested and mounted for diffraction analysis, since the most important requirement is that the crystal diffracts to high resolution. The

diffraction quality of the crystal can be (1) no diffraction, (2) weak diffraction to10 Å (3) promising diffraction 3.5-6 Å, and (4) good diffraction that is better than 2.8 Å. There are several methods for mounting, and one is to place the crystal in a drop of oil or cryoprotectant using a nylon loop, which is immediately flash frozen in liquid nitrogen. By freezing the crystal in cryoprotectants, the radiation damage incurred during data collection is greatly reduced. Freezing reduces the thermal motion within the crystal, which may give rise to a good diffracting crystal with better quality data. The harvested crystals are mounted on a diffractometer that is coupled to an X-ray generator, which can have a stationary anode (circa 2kW DC input), rotating anode (circa 14kW DC input) or a synchrotron (Figure 1-7). The crystals can be screened by either exposing the crystals in a mounted capillary tube at room temperature or exposing the cryo-mounted crystal in a stream of liquid N_2 at 100K (Hope, 1990). The X-rays diffract by interacting with the electrons in the crystal, and the diffrrcation is recorded on a charge-coupled device detector, which is then scanned into a computer.

Successive images are collected as the crystals are rotated within the X-ray beam. For determination a structure with atomic detail, high resolution data should be collected with a well ordered array of spots towards the edge of the diffraction image. A resolution of 3 Å is sufficient to detect amino-acid side chains in the electron density (Figure 1-8).

The amount of data required for structure determination depends on several variables (1) crystallographic symmetry: Only 60° of diffraction data is sufficient if the crystal has a high crystal symmetry; (2) non-crystallographic symmetry (NCS) the amount of symmetry present in the asymmetric unit i.e. how many identical units are present. A dimeric protein might exhibit a high level of NCS, and a high quality structure

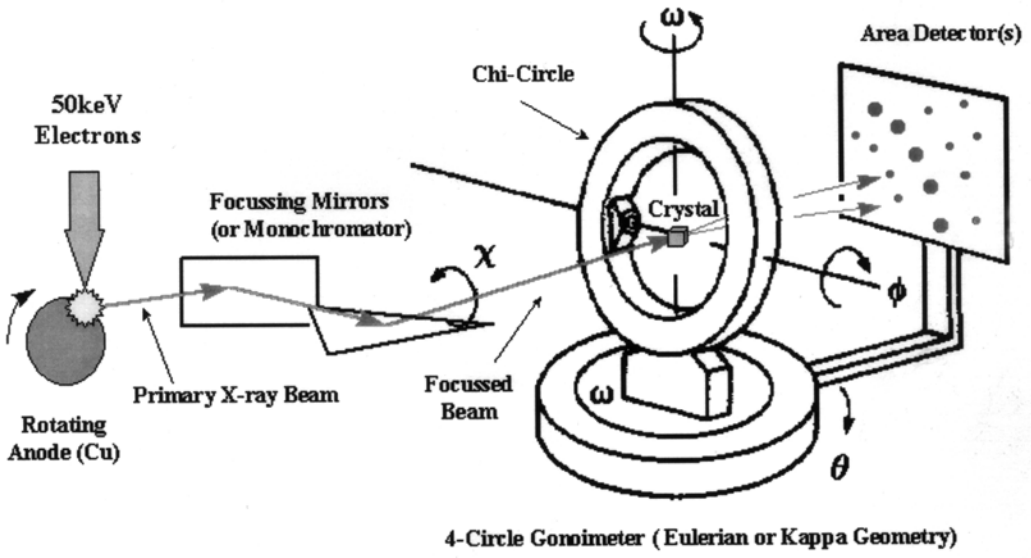


Figure 1-7. X-ray diffraction experiment. Reprinted with permission from Bernhard, R. http://ruppweb.dyndns.org/ Accessed April 6, 2007.

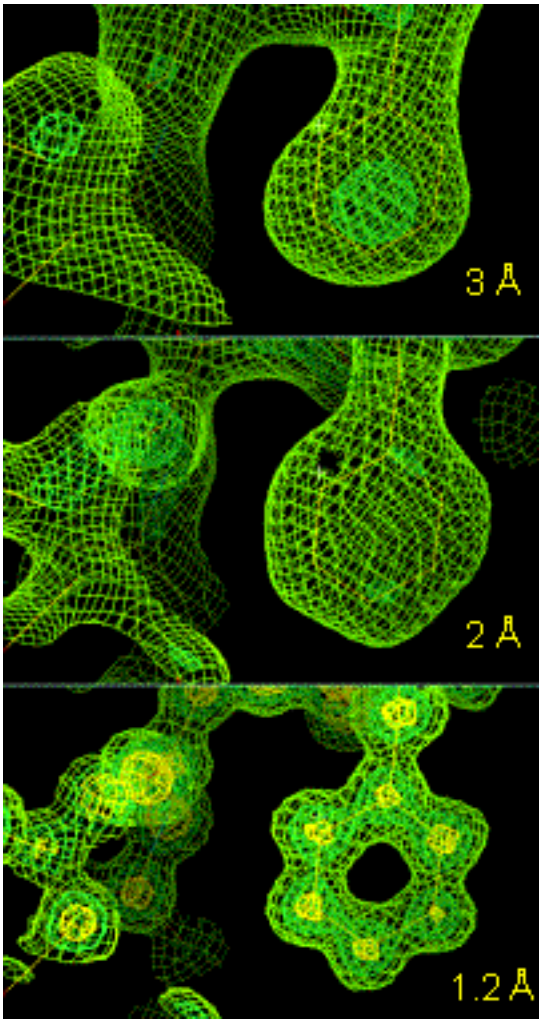


Figure 1-8. Importance of resolution. The top panel shows an amino-acid sidechain at 3Å resolution, the second and third panels show the same amino-acid at 2 and 1.2 Å. Reprinted with permission from Bernhard, R. http://ruppweb.dyndns.org/. Accessed April 6, 2007.

can be obtained from an incomplete dataset because of the compensation by averaging, (3) availability of molecular replacement; If a related structure has already been solved and only gaps need to be filled in the final structure, then an incomplete dataset is sufficient for structure determination, and (4) the resolution limit required. The amount of diffraction data required to determine a structure increases exponentially with the resolution.

The initial diffraction pattern analysis gives enough information on the spot spacing on the detector for the desired crystal to detector distance. This is important since the general appearance of the diffraction pattern is due to the crystal symmetry, whereas the spot is dependent only on the unit cell dimension. Variation in the intensities of each spot defines the structure, which is extracted during data processing. The crystal to detector distance is also used to determine the oscillation range for each exposure. Oscillation can be from as little as 0.25° to 2°. The total number of spots on the diffraction pattern is controlled by the oscillation range of the exposure. All these parameters are taken into consideration for collecting the maximum amount of data possible for each image and the onset of spot overlap.

Data is processed mathematically by many different algorithms. The first and foremost step of data processing is to determine relatively accurately the unit cell dimension and the crystal system to which the data belongs. This part of the data is obtained from the first image and this determines what the subsequent crystal orientation should be. After a complete dataset is obtained, indexing (Kabsch, 1988) is done. Each spot in the diffraction pattern is given an index, which is assigned three integers *h*, *k* and *l.* Indexing is the process of calculating a prediction of what the diffraction image will

look like from the unit cell dimensions and orientation, and then comparing by fitting the real image with the predicted image. After indexing, the intensity of each spot has to be measured. Presently a program called DENZO (Otwinowski, 1997) performs both indexing and intensity measurement. For structure determination the collection of diffraction images in the dataset must be related to each other. However, when the data is collected, due to the change in the intensity of the synchtron beam in different frames, and due to the size variation in the crystal in different orientations the spot intensity varies with the amount of beam traversed. So, through a process called scaling, all the diffraction images within a dataset are processed in relation to each other. By monitoring the scaling process and its statistics, spots or whole images can be rejected or reprocessed to preserve the quality of the data. SCALEPACK is one of the widely used programs for scaling (Otwinowski, 1997)**.** The output file from SCALEPACK contains the index of each individual spot and its measured intensity; this data has to be listed in numerical order according to index, which is done by the CCP4 suite (CCP4, 1994)

As mentioned earlier, the intensity of the spot is determined by the amplitude of the wave and the phase, the difference is expressed as an angle between them. The CCP4 suite is one of several programs that can deduce the amplitude. From the amplitude and the phase, a key parameter called the structure factor can be deduced. The structure factor is a complex number that contains the amplitude and phase of a wave. From the structure factor the arrangement of the atoms in the unit cell can be calculated. To calculate the structure factor the phase angle and, as mentioned earlier, there is no method currently to deconvolute the data in the detector; this is called the phase problem.

Phase angle determination

There are several methods to get the phase angle; they may be classified as the heavy atom method, molecular replacement and direct phasing (*Ab Initio*).

Heavy atoms are normally used where there are no closely-related structures available. This is an experimental phasing method based on the perturbation of the structure and diffraction pattern caused by the heavy atom. The classical method is the isomorphous replacement method (MIR) (Dickerson, 1961; Otwinowski, 1991; Hengming, 1997)**.** In MIR, a nearly identical crystal (isomorphous) to the test crystal is used and a few atoms are either replaced or added, mostly by adding heavy atoms. The electron in the heavy atom scatters the X-ray and significantly perturbs the diffraction pattern in phase with one another. Due to this, different atoms contribute to the scattering intensity and are directly proportional to the number of electrons present in the crystal. This technique requires two datasets: one native and a second heavy atom derivative dataset. The heavy atoms typically used are mercury, platinum or gold. By comparing the differences that arise between the native and heavy atom derivative and carrying out subsequent refinement (Terwilliger and Eisenberg, 1987) the positions of the heavy atoms can be deduced along with the amplitudes, the structure factor.

Multiple wavelength anomalous dispersion (MAD) is a technique similar to MIR but one in which the wavelength of the X-ray is changed to cause anomalous scattering and/or perturbation from the heavy atom crystal. The information gathered is similar to MIR but presently MAD is the method of choice for solving protein structures because all the data can be obtained from a single crystal and the phase angle information obtained is more accurate than that from MIR (Hendrickson, 1991; Smith, 1991; Hendrickson,

1997). In MAD the heavy atoms electrons absorb X-rays of a particular wavelength and re-emit them after a delay, thereby inducing the shift in phase in all the reflections. This is known as the anomalous dispersion effect. This shift, when analyzed, can give the phase angle. The drawback of MAD is that it requires X-ray excitation at a very specific wavelength near the absorption edge of the heavy atom used, so it is necessary to use synchrotron radiation. The most poplar way to use MAD is to introduce seleno-methionine (SeMet) in place of methionine residues in the protein (Ogata, 1998). This method is popular because methionine naturally occurs in proteins at a level of about 2%, and SeMet-substituted proteins are structurally isomorphous to the native protein. Furthermore, both prokaryotic and eukaryotic cells can be grown in media substituted with SeMet. The other experimental method of phase determination is a modification of MAD called single wavelength anomalous diffraction (SAD) (Wang, 1985). SAD phasing requires only a single dataset and is made up of (1) finding anomalous scatterers, (2) evaluating initial phases, and (3). phase improvement by density modification algorithms (Hauptman, 1982; La Fortelle, 1997; Cowtan, 1999; Langs *et al.*, 1999; Terwilliger, 2000).

Molecular replacement (MR) (Rossmann, 1990) is a technique, which is used when a closely-related structure is available. The structure has to be either chemically or structurally similar for it to be used in MR. In this technique the structure factors and the phase are calculated or borrowed from the already known coordinate files and applied to the data to be solved. Before the phase can be applied to the test structure a process called, rotation function (Rossman, 1972; Rossmann, 1990) has to be performed. In rotation function the model structure has to be placed in the unit cell in the same exact

position and orientation as the new protein molecule. The rotation function is followed by the translation function (Taylor, 1959; Fujinaga, 1987) which moves the repositioned data through the unit cell to fit the new molecule precisely. Once the model has been positioned through multiple refinements, the position and orientation is improved. After refinement (Brunger, 1992) the phases from the model structure are applied and, with the new amplitudes, the structure factor for the test data is calculated. The most commonly used program for MR is called AmoRe (Navaza, 1997). The drawback of MR is that there is a severe bias towards the model, but MR is very useful for doing ligand-binding studies or studying molecules with small differences such as mutations.

Ab Initio phasing can be done if there is a very high resolution dataset available. The cut-off for this type of phasing is data should be better than 1.6 Å or 160 picometers. The limiting factors for this method are data quality and processing power.

Calculation of electron density map, refinement and model building

Once the initial phases have been obtained, an electron density map can be calculated. The map has the three dimensional contours into which the structure will be built. The quality of the map depends on the spacing of the unit cell edges. The map is used preliminarily to derive portions of the structure that will give rise to a new set of phases and a new electron density map. This process is repeated until an error term called R-free has stabilized to a satisfactory value. This process is called refinement (Brunger, 1992). Through refinement, an electron density map of sufficient quality can be determined, and this map can be used for model building. Model building is presently done through a computer graphics program such as “O” which displays the map” (Jones

et al., 1991). Using the protein sequence, residues are inserted into the map to commence model building. Care is taken to minimize the energy of conformation by using a dictionary of data on bond length and angles within the constraints of the map. The structure that has been built is judged through the standard crystallographic R-factor, which is simply the average fractional error in the calculated amplitude compared to the observed amplitude. For a good structure, the R-factor should be in the range of 15% to 25%. The final output file is usually in the format of a protein data bank, known as a PDB file.

Validation

At a given resolution, there are three to four parameters for each atom. These parameters describe the position and mobility of the atom. Due to the lack of enough data, diffraction data is usually supplanted with restraints on geometry, which maintains the bond length, angles and close contacts in a reasonable range. Therefore, with this information it is very easy to overfit the data; thereby disagreeing with the observed data. Usually this problem is circumvented by using only part of the data for refinement and then the rest of the data is used to verify how well the refinement has performed. This process is called cross-validation (Brunger, 1992). Another tool for validation is to use the Ramachandran plot, which defines the mainchain and side chain torsion angles.

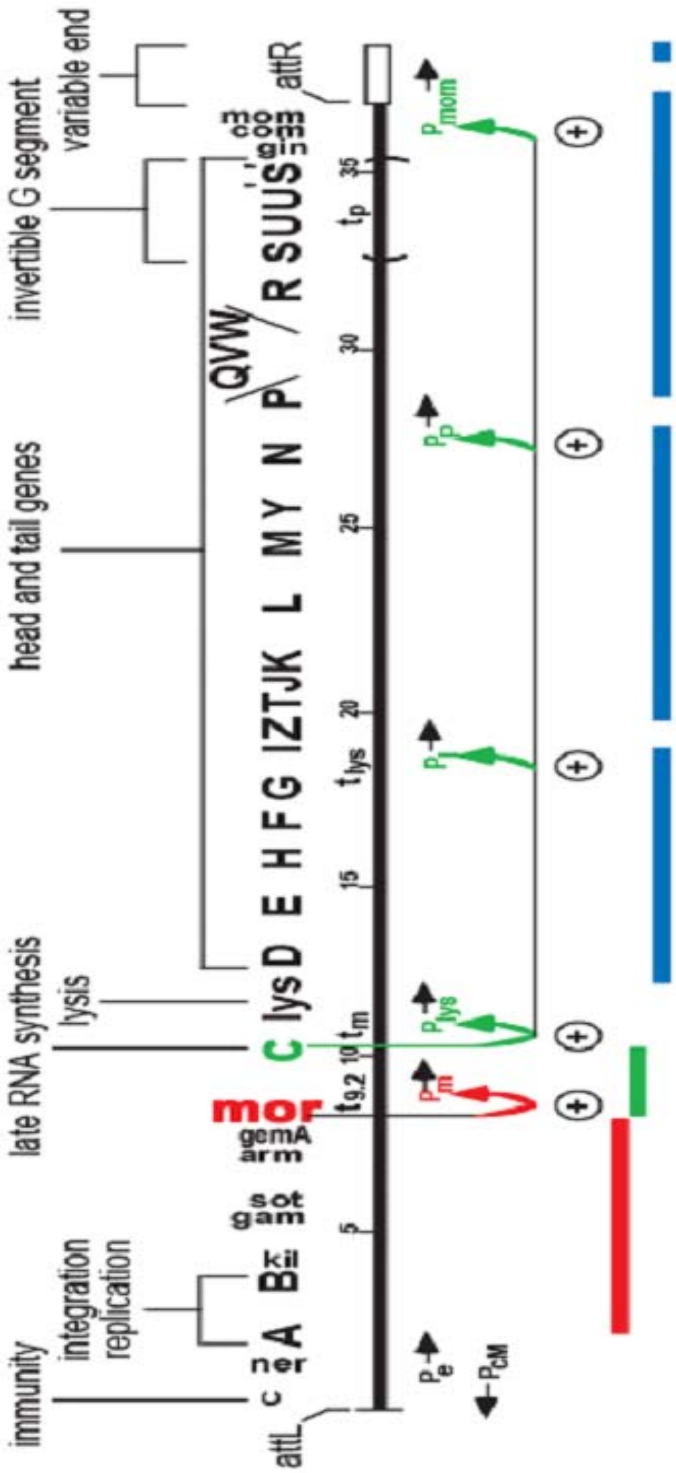
Bacteriophage Mu

Mu is a temperate phage of *Escherichia coli* K-12 and other enteric bacteria (Paolozzi, 2006; Symonds, 1987). The Mu genome encodes ~ 45 genes, which are

necessary for the phage replication, growth, morphogenesis, DNA modification and cell lysis. Following entero-bacterial infection, depending on the environment, nutrition and host cell conditions, Mu can form a lysogen or enter a lytic mode of replication for production of progeny phage particles. The lytic cycle is tightly regulated by a transcriptional cascade, which is divided into early, middle, and late phases (Figure 1-9) (Stoddard and Howe, 1989; Marrs and Howe, 1990)**.** Transcription from the early promoter (P_e) is entirely independent of replication and *de novo* protein synthesis. Host encoded RNAP is sufficient to carry out transcription from P_e since P_e resembles the consensus promoter sequence. The middle operon regulator, Mor, is produced from the last gene in the early transcript (Mathee and Howe, 1990)**.** Mor activates the middle promoter P_m, and C protein is encoded in the last gene of the middle transcript. The C protein activates transcription from the four late promoters P_{lys}, P_I, P_P, and P_{mom} (Margolin *et al.*, 1989).

The middle and late promoters each have a –10 hexamer but lack a detectable –35 hexamer (Margolin *et al.*, 1989). These promoters have a Mor or C binding site just upstream of the –35 region and centered on – 43.5 (Chiang and Howe, 1993; Artsimovitch and Howe, 1996; Ramesh and Nagaraja, 1996; Sun *et al.*, 1997). These properties are typical for promoters under positive regulation. Transcription from the middle and late promoters are catalyzed by the host-encoded RNAP containing σ^{70} (Hattman *et al.*, 1985; Margolin *et al.*, 1989) and are dependent on DNA replication and *de novo* protein synthesis. A homology search with the amino-acid sequence of Mor identified the Mu late promoter activator C, bacterial regulator RdgB, and 12 proteins from other Mu-like prophages as sequence homologues (Kumaraswami *et al.*, 2004).

Figure 1-9. Transcriptional organization of bacteriophage Mu. The Mu genome is shown as a dark horizontal line above which the location of each Mu gene is given. The promoters and the direction in which they promote transcription are shown below the genome. The early (red) transcript encodes the Mor protein, which activates middle transcription (green). The middle transcript encodes the C protein, which activates late transcription (blue).

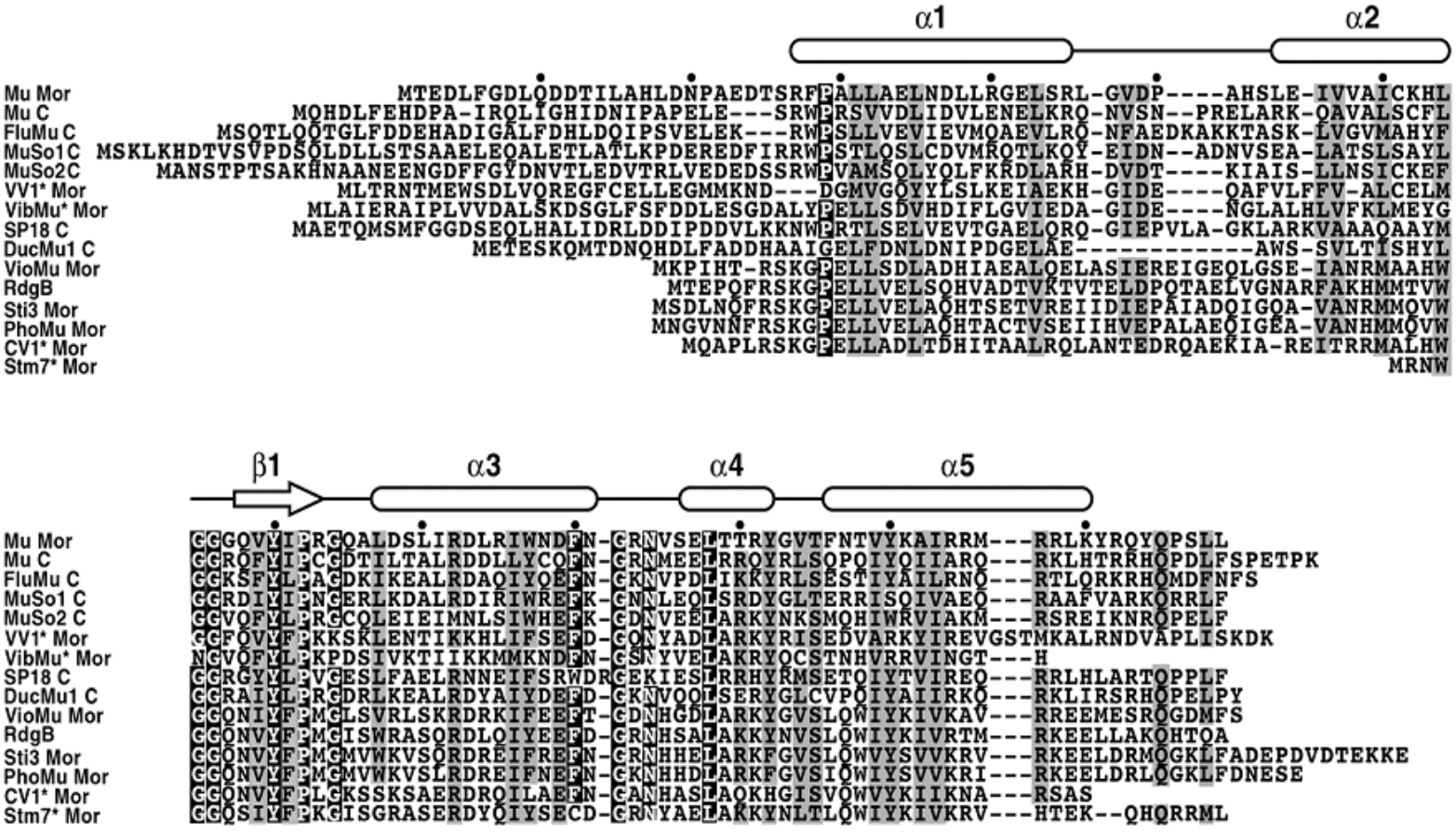

immunity
integration
replication
late RNA synthesis
lysis
head and tail genes
invertible G segment
variable end
attL
attR
cI
cro
A
B
kil
gam
sot
mor
gemA
arm
C
lysD
E
H
F
G
I
Z
T
J
K
L
M
Y
N
P
Q
V
W
R
S
U
U
S
o3
g60
g33
5
10
15
20
25
30
35
$t_{9.2}$
t_m
t_{lys}
t_p
P_e
P_{cM}
P_m
P_{lys}
P_t
P_p
P_{mom}

BLASTP alignment (Altschul *et al.*, 1997) of Mor and C amino-acid sequences revealed that Mor and C share a high degree of amino-acid sequence similarity with each other, (38% identical amino- acids and 55% chemically similar amino-acids), but not with other known transcription activators (Figure 1-10). Thus, they identify a new family of transcription factors (Mathee and Howe, 1990) (Figure 1-10)**.** Mor and C proteins form dimers in solution (Artsimovitch and Howe, 1996; Ramesh and Nagaraja, 1996) and bind an imperfect dyad-symmetry element. Mutational and biochemical analysis of P_m indicate that the –10 hexamer and the Mor binding regions are important for activation (Artsimovitch and Howe, 1996). These analyses also showed that the bases downstream of the Mor binding site have an important role in DNA distortion and a possible role in transcription activation. The AT- rich region just upstream of the Mor binding site binds the α-CTD and may function as an UP-like element (Artsimovitch and Howe, 1996; Ma and Howe, 2004). *In vitro* transcription assays have indicated that the C-terminal domains of both the alpha and sigma subunits of RNAP are required for transactivation of P_m by Mor.

A working model as to how Mor activates transcription in P_m has been proposed. According to the model Mor binds to the promoter as a dimer, recruits the RNAP and interacts with both α and σ^{70} subunits leading to transcription initiation (Artsimovitch and Howe, 1996).

The crystal structure of Mor identified certain structural components which participate in dimerization, DNA binding and transcription activation (Figure 1-11). Mor has an N-terminal dimerization domain and a C-terminal HTH DNA-binding domain (Kumaraswami *et al.*, 2004). The two N-terminal helices of two monomers

Figure 1-10. Amino-acid sequence alignment of members of the Mor and C family of transcription activators. Based on the Mor crystal structure, the locations of secondary structures (alpha α and beta β) predicted for the family is indicated above the sequence. Letters on a black background represent identical amino-acids and those on the shaded grey background represent chemically similar residues. The dots above the sequence indicate a 10 amino-acid sequence interval. Reprinted with permission from Kumaraswami, M., Howe, M.M., and Park, H.W. (2004) Crystal structure of the Mor protein of bacteriophage Mu, a member of the Mor/C family of transcription activators. *J Biol Chem* 279: 16581-16590.


α1
α2
Mu Mor
Mu C
FluMu C
MuSo1 C
MuSo2 C
VV1* Mor
VibMu* Mor
SP18 C
DucMu1 C
VioMu Mor
RdgB
Sti3 Mor
PhoMu Mor
CV1* Mor
Stm7* Mor
β1
α3
α4
α5

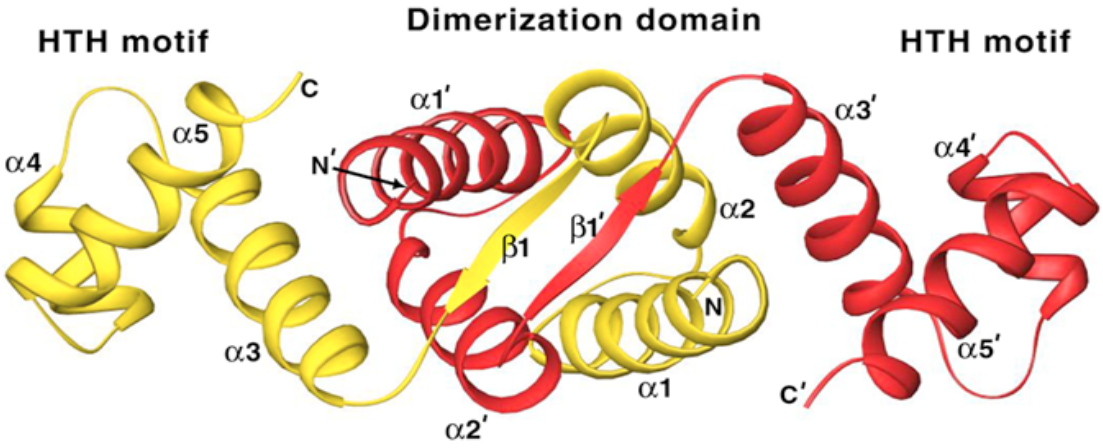


Figure 1-11. Crystal structure of His-Mor. The ribbon illustration shows two individual Mor monomers (yellow and red) dimerizing using alpha helices one and two and two interacting beta strands. The DNA binding HTH motif is made up of helix three, four and five. Reprinted with permission from Kumaraswami, M., Howe, M.M., and Park, H.W. (2004) Crystal structure of the Mor protein of bacteriophage Mu, a member of the Mor/C family of transcription activators. *J Biol Chem* 279: 16581-16590.

intertwine with each other to form a single, central, dimerization domain.

The two flanking HTH domains are proposed to bind to two adjacent major grooves. Since the predicted DNA binding residues of Mor are too far apart to fit into two adjacent major grooves, DNA binding may be associated with conformational changes in the Mor dimer and the DNA (Kumaraswami *et al.*, 2004). Based on the Mor crystal structure and amino-acid alignment it has been predicted that C has a similar HTH DNA-binding motif and dimerization motif (Kumaraswami *et al.*, 2004) and that DNA binding by C may also involve conformational changes.

Chapter 2. Binding Specificity of Mu Transcription Activator C

Introduction

In transcriptional regulation, understanding the sequence-specific binding of the transcription regulator is quite important. Sequence specificity of a protein is the ability to recognize its cognate binding site in the presence of non-specific DNA. In addition, specificity has to be defined in terms of the number of base pairs and sequence required to define a unique binding site in the whole genome. The molecular mechanism of identifying the individual base pairs is primarily through complementary hydrogen-bonding through the major and minor grooves of the DNA and appropriately positioned specific amino-acids in the DNA-binding motif of the protein (von Hippel and McGhee, 1972; Luger *et al.*, 1997; Luscombe *et al.*, 2001). When recognition of DNA by a regulatory protein is not possible, secondary mechanisms like single vs double strand and DNA secondary structure (B form vs Z form) comes into play. This secondary recognition is primarily influenced by base-pair composition, electrostatic potential and stereochemistry (von Hippel and Berg, 1986). This mechanism only gives the regional binding specificity and does not give any information about specific base-pair amino-acid specificity, which is mostly provided by hydrogen bonding interactions. Following sequence specific reading by the protein of the probable hydrogen bond donors and acceptors, the specificity is further influenced by the overall affinity of the protein to the specific site in the DNA. Since there is a detectable decrease in the affinity from the correct site to the incorrect sites with decreasing homology, a relative comparison of affinity can be made. This is important because if there is an increased occurrence of the

wrong site, which can compete for the free protein, the occupancy of the correct site will be reduced. It is important to understand that more sequence specific binding energy is lost for every wrong base-pair when compared to a consensus pair because at least one hydrogen bond is broken and not replaced, thus specificity may also be attributed to the unfavorable effect of incorrect contacts.

In most regulatory proteins the main determinants of specificity are the combination of contributions from the correct and incorrect base pairs, and in many, such as the Lac repressor, in addition to specific binding, there is also a requirement for non-specific electrostatic interaction between the amino-acid sidechains and the DNA phosphate backbone (Winter *et al.*, 1981). But in certain isomerization states of the Lac repressor, T4 DNA polymerase (Fairfield *et al.*, 1983) and *E. coli* RNAP the binding mode will be predominantly or wholly from electrostatic interaction rather than specific sequence interaction. This is to facilitate the proteins "sliding" over the DNA backbone.

When the protein interacts with the DNA, due to the flexible nature of both macromolecules, there is a high possibility of structural distortion to facilitate specific binding interactions within an energetically available conformation. Conformational changes in a complex are common; they may improve or degrade potential hydrogen bonding between the DNA and the protein.

Temperate phage Mu (Symonds *et al*, 1987; Paolozzi, 2006) must have a mechanism to determine when to have its prophage excise from the host DNA and become lytic. The lytic cycle is tightly regulated by a transcriptional cascade designated as early, middle and late. All the lytic genes are arranged left to right in the genome and the transcription proceeds in this direction. Most proteins required for the head and tail

morphogenesis are encoded from the late transcripts originating from the four late promoters P_{lys}, P_I, P_P, and P_{mom} (Margolin *et al.*, 1989). The C protein is directly responsible for activating transcription from these late promoters. In the well studied promoters, P_{lys} and P_{mom}, transcription starts downstream of a conserved sequence (Hattman *et al.*, 1985; Margolin *et al.*, 1989). Bolker *et al*. (1989) confirmed the finding that the C protein binds to a conserved sequence just upstream and overlapping the –35 region using MPE Fe(II) footprinting.

To define the sequence required for activation Chiang and Howe (1993) first made sequential deletions of P_{lys}, with the results suggesting that sequences upstream of –60 and downstream of +8 were dispensable for promoter activity. Chiang and Howe (1993) also made point mutations in P_{lys} which revealed that sequence from –50 to –35 was required for C-dependent activation (Figure 2-1). In P_{mom}, experiments from the Hattman lab showed that the C target site was between –33 and –52 (Balke *et al.*, 1992; Gindlesperger and Hattman, 1994; Sun *et al.*, 1997). By comparing the four late promoters along with the results of the point mutations Chiang and Howe (1993) proposed that C recognizes an inverted hexa-nucleotide repeat separated by a tetra-nucleotide spacer. In P_{mom}, footprinting analysis led to the development of a consensus sequence which contained an inverted tetra nucleotide repeat separated by a GC-rich spacer and, most importantly, these data revealed that C interacts asymmetrically with this conserved sequence 5'…TTAT-N_6-ATAACC… 3'(Ramesh and Nagaraja, 1996).

Mutational analysis done by Zhao (1999) within the C protein binding site revealed that mutations made in the symmetry elements generally reduced C binding, but the effect was more pronounced when the mutations were in the promoter-proximal

C DNase I footprint

	-58	-57	-56	-55	-54	-53	-52	-51	-50	-49	-48	-47	-46	-45	-44	-43	-42	-41	-40	-39	-38	-37	-36	-35	-34	-33	-32	-31	-30	-29
P_{lys}	C	G	G	T	T	A	T	T	T	C	C	T	G	T	C	A	C	C	A	T	A	A	T	C	C	C	G	C	A	C
P_{mom}	C	A	G	A	T	C	G	A	T	T	A	T	G	C	C	C	C	A	A	T	A	A	C	C	A	C	A	C	T	C
P_I	C	T	C	C	A	G	T	A	C	T	C	A	A	A	T	A	G	C	A	T	A	A	C	C	C	C	A	G	A	T
P_P	C	A	G	T	T	A	C	A	G	T	T	A	A	C	T	G	C	C	A	T	A	A	C	C	C	C	G	G	A	C

Figure 2-1. Mu C binding region in P_{sym} and the Mu late promoters. The heavy bar represents the region protected from DNase I cleavage by C at P_{lys} and P_{mom}. The C binding sites defined by mutational analysis of P_{lys} and P_{mom} are indicated by a pair of inverted arrows.

half of the dyad-symmetry. In contrast, when the point mutations in the distal half of P_{lys} made the sequence more symmetrical to the proximal half, C binding increased with concomitant increased promoter activity (Chiang and Howe, 1993; Zhao, 1999). It was also found that mutations outside the dyad-symmetry element reduced C-dependent activation without affecting C binding. Interestingly, while selecting for increased *in vivo* DNA binding at P_{lys}, Jiang (1999) made the distal half of P_{lys} more symmetrical to the proximal half, resulting in a new C repressible promoter called P_{rep} . On closer examination of P_{rep} using gel-shift it was observed that C protein bound about twice as strongly at P_{rep} as compared to wild- type P_{lys}. This symmetrical C binding sequence 5'…**ATTATG**ACTC**CATAAT**… 3' was called P_{sym}. The goal of this project is to define the binding specificity of C with emphasis on optimization of the binding sequence.

Materials and methods

Media, chemicals and enzymes

MacConkey-lactose plates contained 40 g/L of MacConkey agar base and were supplemented with 0.5% lactose (Difco). Ampcillin (Sigma) and chloramphenicol (Sigma) were used at 50 and 34 μg/ml respectively. The G50 Probe Quant™ Sephadex column was from GE Healthcare Bio-Sciences AB. Radiolabelled [γ-^{32}P] ATP (3000 Ci/mmol) was from Perkin Elmer Life Sciences, *Eco*RI, *Bam*HI and T4 polynucleotide kinase were from New England Biolabs. All dNTPS were purchased from Promega. Acrylamide, bisacrylamide, tetramethylethylenediamine (TEMED) and ammonium persulfate (APS) were purchased from BioRad. The QIAquick spin purification kit was

from Qiagen. Automated DNA sequencing was performed by the Molecular Resource Center of The University of Tennessee Health Science Center.

Oligodeoxyribonucleotides

Tables 2-1 and 2-2 include the oligodeoxyribonucleotides used in cloning and electro-mobility shift assays. Synthesis of the oligodeoxyribonucleotides was done by Integrated DNA Technologies, Inc (IDT).

Bacterial strains and plasmids

Plasmids containing P_{sym} mutant promoters (pKK) were constructed in strain MH 13312 (*mcrA* Δ*proAB-lac thi gyrA endA hsdR relR supE44 recA; F' (pro*$^{+}$ *lacI*^{Q1} Δ*lacZY*) (Artsimovitch and Howe, 1996). The plasmid pLC1 is a promoter-less *lac*Z fusion cloning vector and a Δ*lacY* derivative of pRS415 Chiang and Howe (1993). Plasmid pZZ41 contains the Mu *C* gene under the control of a T7 promoter. It also contains a

Table 2-1. Oligodeoxyribonucleotides used for P_{sym} promoter construction.

Primer	Sequence	Comments
KAR 2	CGGAATTCCCGCCGGTTATATTANN ACTCNNTAATCC	Top strand primer used in P_{sym} to generate mutations at positions –47, –46,–41, –40; with 5' EcoRI site
KAR 3	ACGGGATCCCCAATTCTCTGATGGC AGT	Bottom strand primer to complete synthesis of mutants with 5' BamHI

Table 2-2. Oligodeoxyribonucleotides used for EMSA.

Primer	Sequence	Comments
KAR 75	GTTATATTATGACTCCATAATCCCGC	P_{sym} 26-mer Top strand WT
KAR 76	GCGGGATTATGGAGTCATAATATAAC	P_{sym} 26-mer Bottom strand WT
KAR 77	CGGTTATATTATGACTCCATAATCCC GCAC	P_{sym} 30-mer Top strand WT
KAR 78	GTGCGGGATTATGGAGTCATAATATA ACCG	P_{sym} 30-mer Bottom strand WT
KAR 79	CGGTTCTATTATGACTCCATAATCCCG CAC	P_{sym} 30-mer Top strand –53 C
KAR 80	GTGCGGGATTATGGAGTCATAATAGA ACCG	P_{sym} 30-mer Bottom strand –53 C
KAR 81	CGGTTTTATTATGACTCCATAATCCCG CAC	P_{sym} 30-mer Top strand –53 T
KAR 82	GTGCGGGATTATGGAGTCATAATAAA ACCG	P_{sym} 30-mer Bottom strand –53 T
KAR 83	CGGTTGTATTATGACTCCATAATCCC GCAC	P_{sym} 30-mer Top strand –53 G
KAR 84	GTGCGGGATTATGGAGTCATAATACA ACCG	P_{sym} 30-mer Bottom strand –53 G
KAR 85	CGGTTACATTATGACTCCATAATCCC GCAC	P_{sym} 30-mer Top strand –52 C
KAR 86	GTGCGGGATTATGGAGTCATAATGTA ACCG	P_{sym} 30-mer Bottom strand –52 C
KAR 87	CGGTTAAATTATGACTCCATAATCCC GCAC	P_{sym} 30-mer Top strand –52 A
KAR 88	GTGCGGGATTATGGAGTCATAATTTA ACCG	P_{sym} 30-mer Bottom strand –52 A

Table 2-2 (continued).

Primer	Sequence	Comments
KAR 89	CGGTTAGATTATGACTCCATAATCCC GCAC	P_{sym} 30-mer Top strand –52 G
KAR 90	GTGCGGGATTATGGAGTCATAATCT AACCG	P_{sym} 30-mer Bottom strand –52 G
KAR 91	CGGTTATATTATGTCACCATAATCCC GCAC	P_{sym} 30-mer with P_{lys} IR Top strand
KAR 92	GTGCGGGATTATGGTGACATAATAT AACCG	P_{sym} 30-mer with P_{lys} IR Bottom strand
KAR 93	CGGTTATTTCCTGACTCCATAATCCC GCAC	P_{lys}30-mer with P_{sym} IR Top strand
KAR 94	GTGCGGGATTATGGAGTCAGGAAAT AACCG	P_{lys}30-mer with P_{sym} IR Bottom strand
KAR 95	CGGTTATATTATAACTCCATAATCCC GCAC	P_{sym} 30-mer Top strand –46 A
KAR 96	GTGCGGGATTATGGAGTTATAATAT AACCG	P_{sym} 30-mer Bottom strand –46 A
KAR 97	CGGTTATATTATGACTCGATAATCCC GCAC	P_{sym} 30-mer Top strand –41 G
KAR 98	GTGCGGGATTATCGAGTCATAATAT AACCG	P_{sym} 30-mer Bottom strand –41 G
KAR 99	CGGTTATATTATGACTCAATAATCCC GCAC	P_{sym} 30-mer Top strand –41 A
KAR 100	GTGCGGGATTATTGAGTCATAATATA ACCG	P_{sym} 30-mer Bottom strand –41 A
KAR 101	CGGTTATATTAAAACTCCATAATCCC GCAC	P_{sym} 30-mer Top strand –46 A –47 A

Table 2-2 (continued).

Primer	Sequence	Comments
KAR 102	GTGCGGGATTATGGAGTTTTAATATA ACCG	P_{sym} 30-mer Bottom strand –46 A -40 A
KAR 103	CGGTTATATTATAACTCCTTAATCCC GCAC	P_{sym} 30-mer Top strand –46 A –40T
KAR 104	GTGCGGGATTAAGGAGTTATAATAT AACCG	P_{sym} 30-mer Bottom strand –46 A –40 T
KAR 105	CGGTTATATTATTACTCGATAATCCC GCAC	P_{sym} 30-mer Top strand –46 T –41 G
KAR 106	GTGCGGGATTATCGAGTAATAATAT AACCG	P_{sym} 30-mer Bottom strand –46 T –41 G
KAR 107	CGGTTATATTATCACTCGATAATCCC GCAC	P_{sym} 30-mer Top strand –46 C –41 G
KAR 108	GTGCGGGATTATCGAGTGATAATAT AACCG	P_{sym} 30-mer Bottom strand –46 C –41 G
KAR 109	CGGTTATATTATGACTCCATAATCCC CCAC	P_{sym} 30-mer Top strand –32 C
KAR 110	GTGGGGGATTATGGAGTCATAATAT AACCG	P_{sym} 30-mer Bottom strand –32 C
KAR 111	CGGTTATATTATGACTCCATAATCCC ACAC	P_{sym} 30-mer Top strand –32 A
KAR 112	GTGTGGGATTATGGAGTCATAATAT AACCG	P_{sym} 30-mer Bottom strand –32 A
KAR 113	CGGTTATATTATGACTCCATAATCCC TCAC	P_{sym} 30-mer Top strand –32 T
KAR 114	GTGAGGGATTATGGAGTCATAATAT AACCG	P_{sym} 30-mer Bottom strand –32 T

modified P_{lacUV5} operator and promoter called P_{lac}SYN downstream from the T7 promoter in front of the Mu *C* gene (Zhao, 1999).

P_{sym} mutants

Degenerate mutations were introduced in positions –47, –46, –41 and –40 using degenerate top strand primers. The top and bottom primers (Kar 2 and 3) included P_{sym} sequence from–51 to –36 with *Eco*RI and *Bam*HI restriction sites at the 5′ ends. The promoters fragments were made by using primers Kar 2 and Kar 3 with the P_{sym} promoter in plasmid as template (MH 14714). The PCR products were purified using a QIAquick spin purification kit (Qiagen), digested with *Eco*RI and *Bam*HI and cloned into corresponding *Eco*RI and *Bam*HI sites in pLC1 in front of the promoter-less *lacZ* gene. The ligation mix was transformed into MH 16823 a derivative of MH13312 containing the pZZ41 plasmid. Competent cells were prepared by the $CaCl_2$ method (Mandel and Higa, 1970; Mandel *et al.*, 1990). After transformation with the pZZ41 library, mixture was then plated onto LB plates containing appropriate ampicillin and chloramphenicol. The resulting clones were transferred onto three fresh MacConkey-lactose indicator plates with different IPTG concentration (33 μM, 100 μM and 300 μM), ampicillin and chloramphenicol (two-plasmid transcription activation assay system). The plates were incubated at 32° C and colony color development was recorded every 3 hrs after 12 hrs of incubation. The promoter activity of the mutants was determined by comparing the intensity and the timing of color development over the incubation period relative to the activity of wild-type P_{sym}. The Figure 2-2 shows how the two-plasmid transcription activation assay system works.

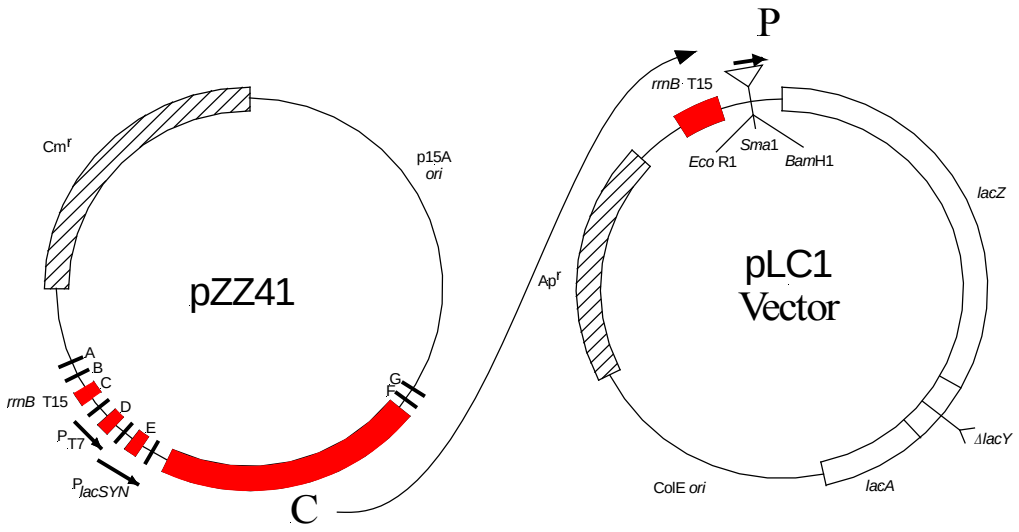


Figure 2-2. Two-plasmid transcription activation assay system. Expression of C protein occurs when IPTG releases the repression of P_{lacSYN}. The expressed C then transactivates the P_{sym} promoter. Transcription from the P_{sym} promoter leads to the synthesis of β-galactosidase from the *lacZ* gene.

Electrophoretic mobility shift assays

The ability of C protein to bind mutant C-binding sites was tested by electrophoretic mobility shift assay (EMSA) or simply, band shift assay (Carey, 1991). A 30-bp probe was generated by taking 100 ng of top or bottom strand oligonucleotide and end-labelling with γ-^{32}P ATP (3000 Ci/mmol) with polynucleotide kinase buffer. The labeled oligos were then annealed with 300 to 500 ng of the complimentary bottom or top strand by placing the mixture in a 100° C heat block for 2 minutes and then switching off the block for it to cool to room temperature. The annealed probe was purified using a G50 Probe Quant™ Sephadex column as per manufacturer's recommendation. The purified labeled probe (100-150 cpm) was incubated at 25° C for 30 min with and without purified wild-type C in 20 µl of buffer C (25 mM Hepes, pH 7.0, 75 mM NaCl, 5% glycerol, 4.5 mM $MgCl_2$, 1 mM EDTA, 10 mM DTT). After incubation, the binding reaction was loaded on to a 8% non-denaturing acrylamide gel in 1X TBE buffer (Tris Base, Boric acid and EDTA) and subjected to electrophoresis for 2 to 3 hrs at 4° C at 10V/cm. The gel was then blotted onto Whatman filter paper and exposed to Kodak Biomax™ MR without screen at -70° C overnight.

Results

P_{sym} mutants

Degenerate primers were used to introduce mutations at positions –47, –46, –41 and –40. The phenotype of each promoter mutant was estimated from its color development on MacConkey-lactose plates. The plate phenotype for representative

mutants was done at least thrice with P_{sym} as wild-type control. Table 2-3 shows the plate phenotype for all the mutants. Candidates for further testing using a gel-shift assay were identified and grouped as (1) high activity mutants, (2) moderate activity mutants, and (3) mutants with no activity based on their plate phenotype. At position –47 comparison of the WT base in all four late promoters and the plate phenotype of these mutants show that –47 T was most preferred. Mutants with A or G at position –46 had greater activity than mutants with either G or C; this was also seen in all Mu late promoters. At –41, C was preferred in the mutants as it was in the natural Mu late promoters. At position –40, A was consistently preferred. Based on these preferences, P_{sym} containing –47 T, –46 G or A, –41 C and –40 A should have good transcription activity, and this prediction was validated by a single mutant with high activity in the plate assay Table 2-3.

Gel-shift assays

To obtain an optimized C-binding sequence and to test candidate mutations it was necessary first to determine a minimal probe length. Oligonucleotide probes of varying length containing P_{sym} sequence were annealed and were used to test the influence of length on C binding (Figure 2-3). The assay revealed that probe length of less than 26 bp reduced C binding. Therefore, a 30-mer probe length was chosen since it gave optimal binding with minimal influence in C binding. The binding ability of the mutants was scored by visually comparing the shifted band with the wild-type promoter. A score of "++++" was given if there was partial shift at 1x WT C concentration and complete shift between 2x and 4x. A single 1x is the difference in C protein required to shift the DNA relative to the WT promoter.

Table 2-3. Grouping analysis of mutant promoters with C.

Genotype with high activity[a]	Genotype with moderate activity	Genotype with no detectable activity	
–47–46–41–40[b]	–47–46–41–40	–47–46–41–40	–47–46–41–40
T A C A	T A T A	T T G A	A T A A
	A A C A	G A T A	A C A A
	T G G A	T A C T	T C A A
	T G A A	T G T T	C C G T
	C T C A	A A A A	C C T G
		T T G G	A C A T
		T A A G	C A G A
		A T A G	C T G T
		T T A C	T T A G
		A T G G	T T C T
		T A G C	G C G T
		A T T T	T C G A
		G T T T	T C G T
		C T T T	T C G C
		A G A T	T T G G
		T T A G	T A T G
		T T A C	C A A A
		T T T C	G A A T
		G T A T	C G C T
		T C T C	G T T A

[a] *In vivo* transcription activity for each mutant promoter was determined by plate phenotyping.

[b] The bases at position –47–46–41–40 on the top strand are listed. The wild-type bases for P_{lys}, P_I, P_P, and P_{mom} are TGCA, AACA, AACA and TGAA respectively.

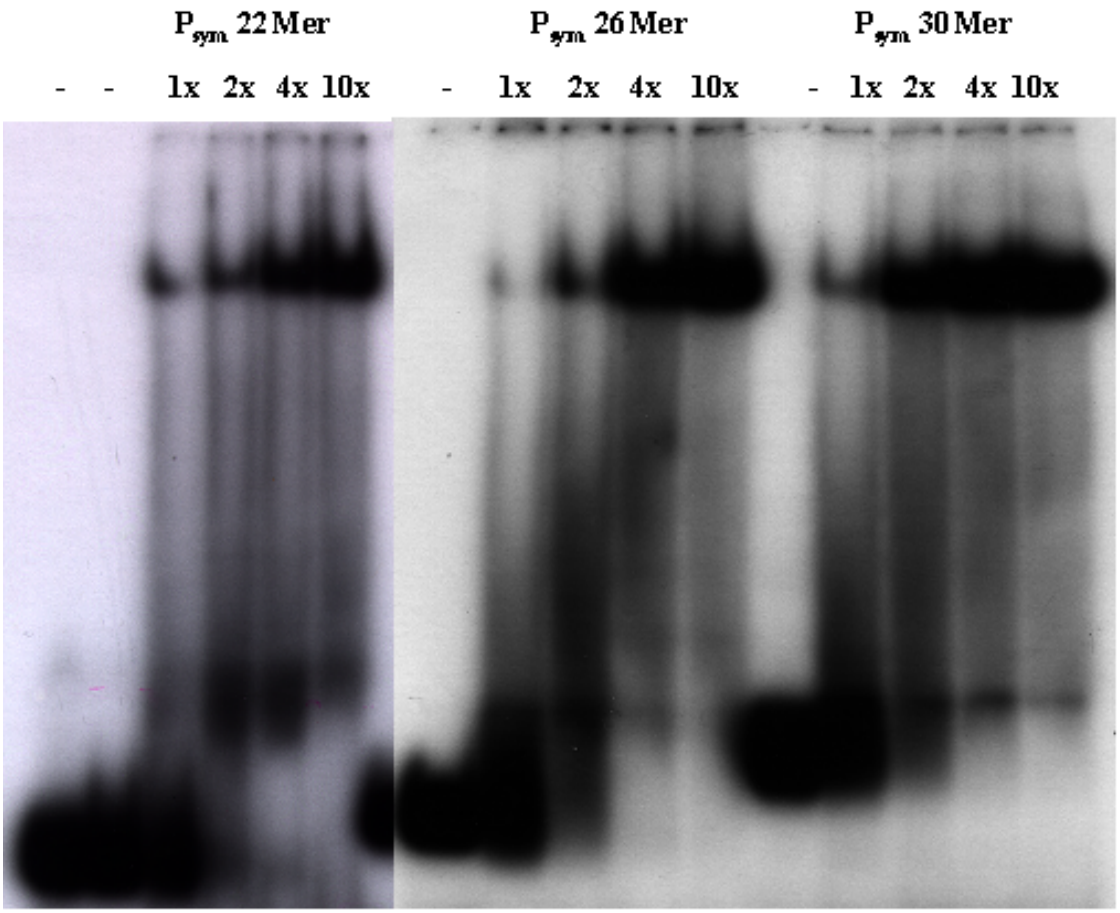


Figure 2-3. Gel-shift assay of P_{sym} with varying DNA length. Annealed P_{sym} probes were incubated with WT C protein [0 ng (-), 20 ng (1x), 40 ng (2x), 80 ng (4x), 200 ng (10x)].

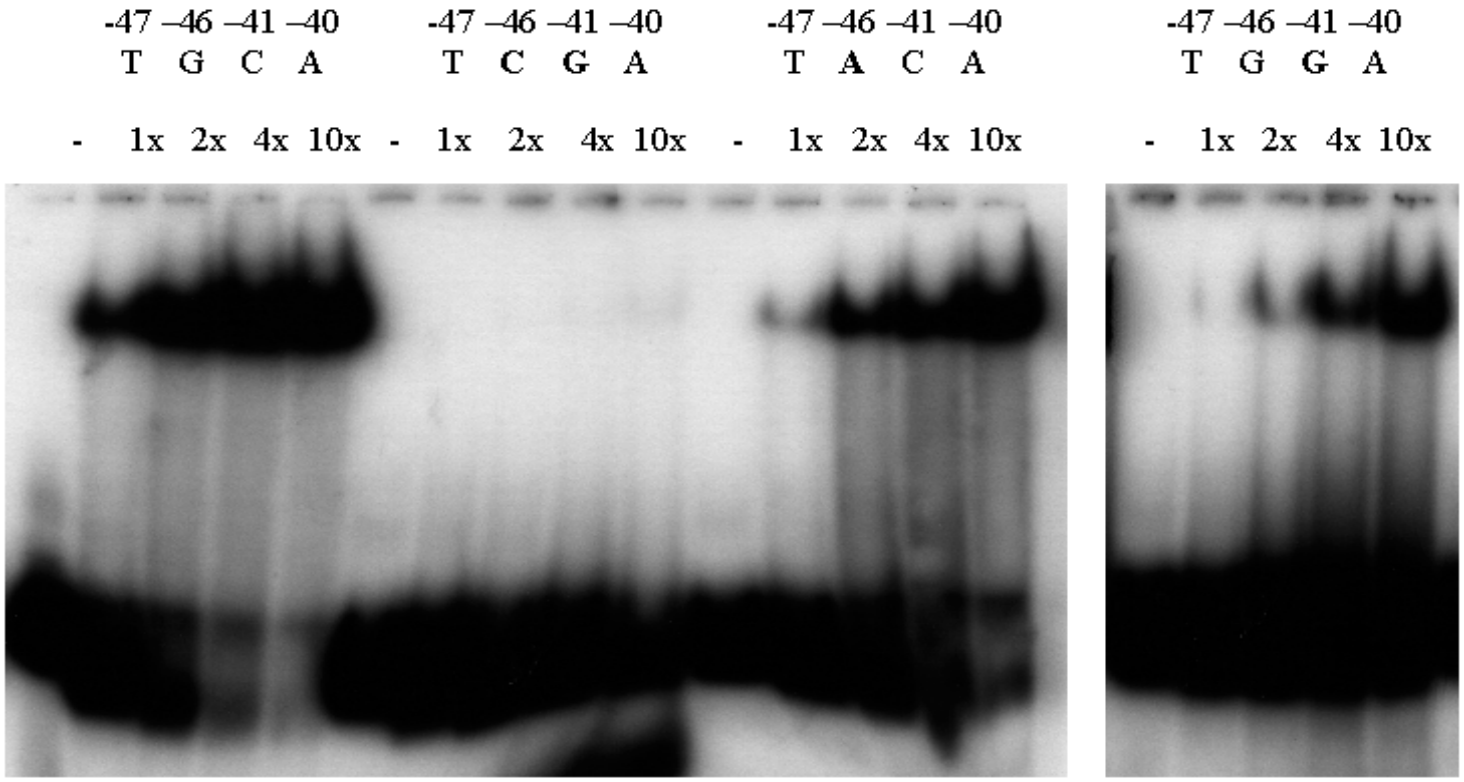
Gel-shift was used to test C binding to all four natural Mu late promoters, P_{sym} and mutant P_{sym} promoters. For P_{sym} there was almost a 20-30% shift at 20 ng and almost 100% shift was seen at 80 ng. The gel-shift results and its summary for all the mutants tested are presented in Figure 2-4, Table 2-4, Figure 2-5, Figure 2-6, Figure 2-7, Figure 2-8 and Figure 2-9.

Mutations that greatly influence C binding were restricted to the region between –51 and–36. Mutations tested outside of this region had very little or no effect on C binding. For mutations flanking the inverted repeat spacer (IR) all mutations tested except –46 A reduced C binding. Mutations located within the 4-bp spacer had the most effect in C binding. By switching the IR between P_{sym} and P_{lys}, it was found that the P_{sym} IR in the P_{lys} context significantly reduced C binding. When P_{sym} IR was changed at only one base (–43 or –45) rather than multiple bases, C binding was not reduced as significantly. Lastly, when the length of the IR spacer was changed either by insertion or deletion, C binding was completely abolished.

Discussion

Previous studies with P_{lys} involved site-directed mutagenesis, deletion mapping, footprinting and gel-shift analysis. Those assays were done to investigate C binding as well as C-dependent transcription activation. These previous studies involved full-length promoters and His-tagged C protein. The present study was done to extend and validate the previous binding analyses by using shorter C-binding sequences and WT-C protein. The binding assays in this study were aimed at primarily delineating the importance of the sequences within and flanking the IR spacer. The secondary aspect of this project was

Figure 2-4. Gel-shift assay for P_{sym} mutants altered at –47, –46, –41 and –40. Labeled wild-type and mutant P_{sym} 30-mer probes were incubated with 0 ng (-), 20 ng (1x), 40 ng (2x), 80 ng (4x), 200 ng (10x) with WT C protein. The bases at positions –47, –46, –41 and –40 are listed above. The wild-type bases are TGCA.


-47 –46 –41 –40
T G C A
- 1x 2x 4x 10x
-47 –46 –41 –40
T C G A
- 1x 2x 4x 10x
-47 –46 –41 –40
T A C A
- 1x 2x 4x 10x
-47 –46 –41 –40
T G G A
- 1x 2x 4x 10x

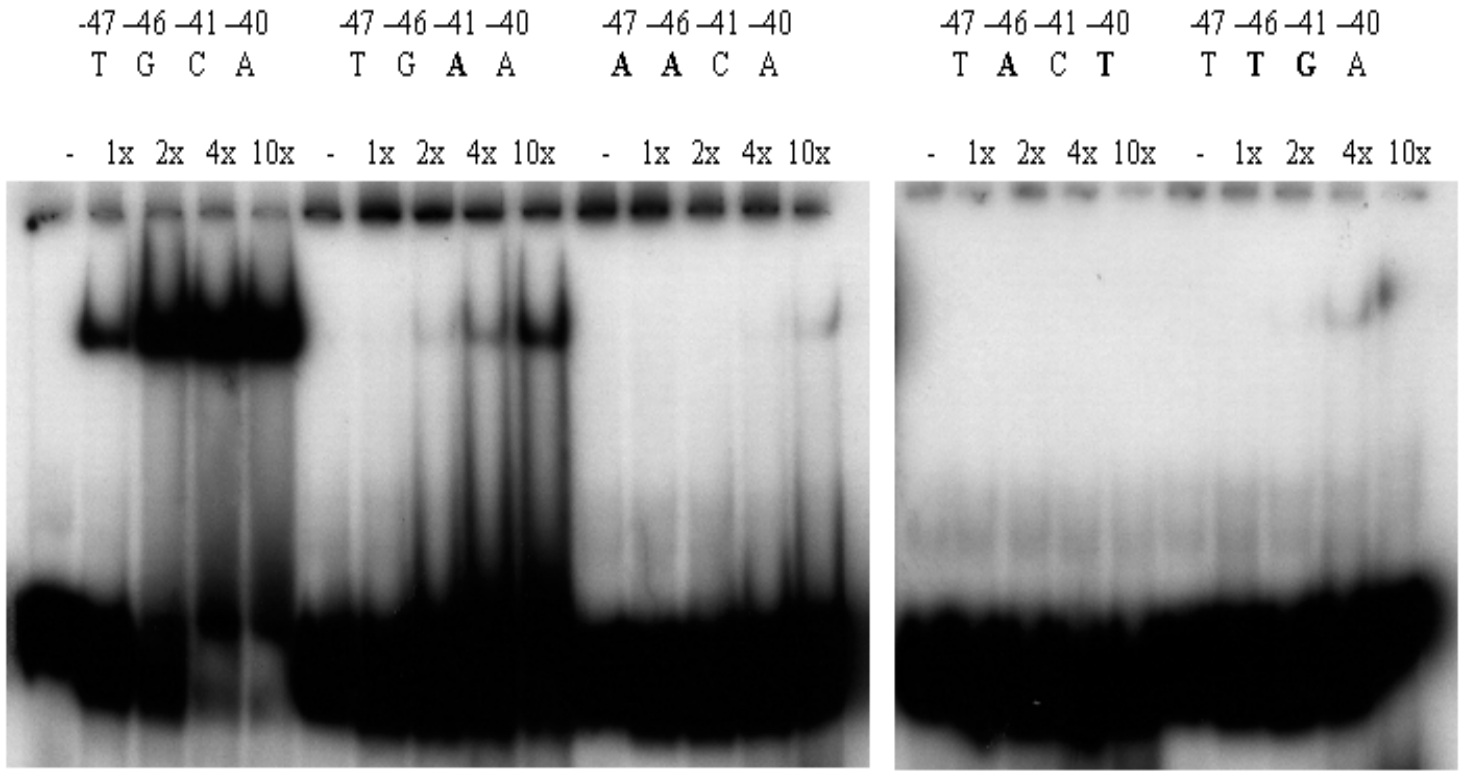


Figure 2-4 (continued).

Table 2-4. Summary of the relative binding efficiency of P_{sym} and P_{sym} mutants altered at –47, –46, –41 and –40.

–47	–46	–41	–40	Binding Efficiency
T	G	C	A	++++
T	A	C	A	+++$\underline{+}$
T	G	G	A	+++
T	G	A	A	++
A	A	C	A	$\underline{+}$
T	T	G	A	$\underline{+}$
T	C	G	A	-
T	A	C	T	-

The binding abilities of the P_{sym} mutants were scored as "+", relative to wild-type binding of "++++."

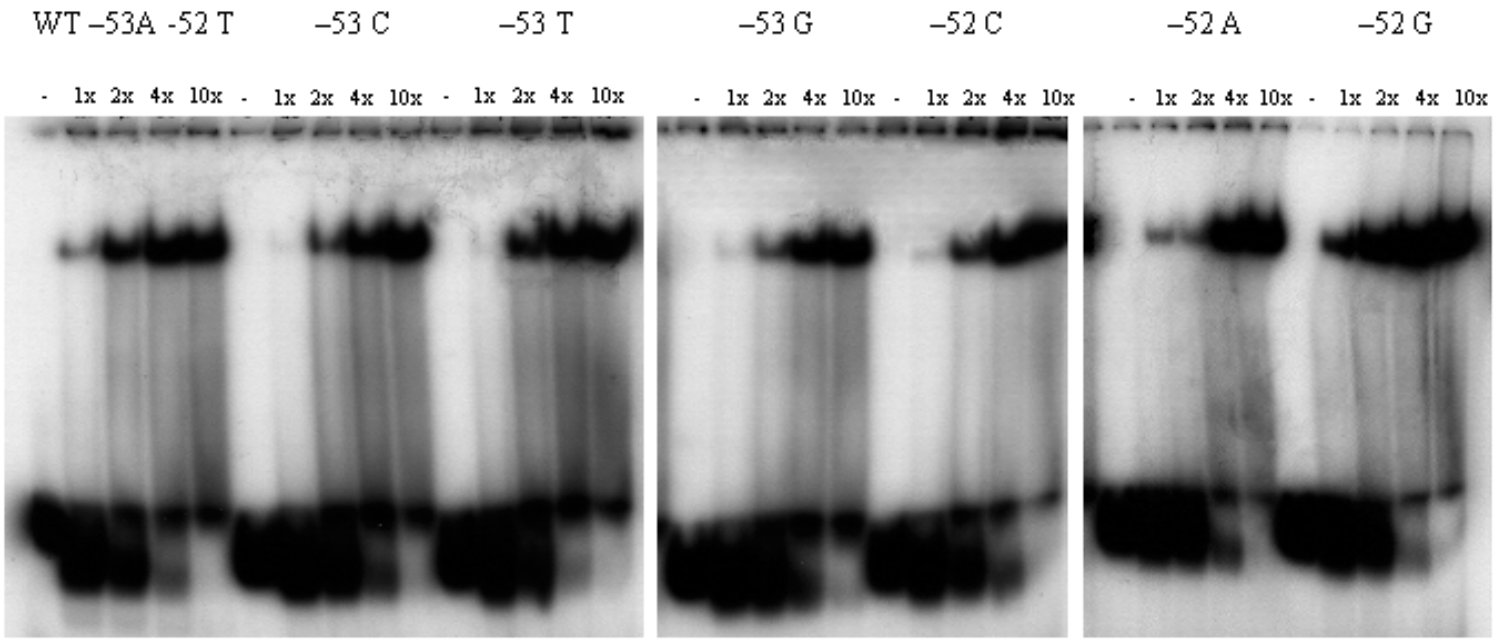
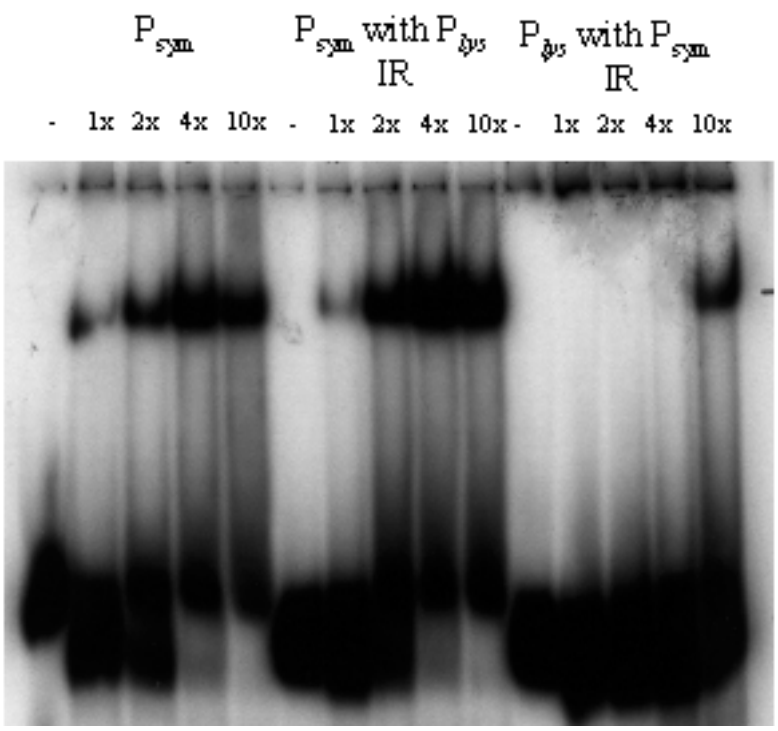


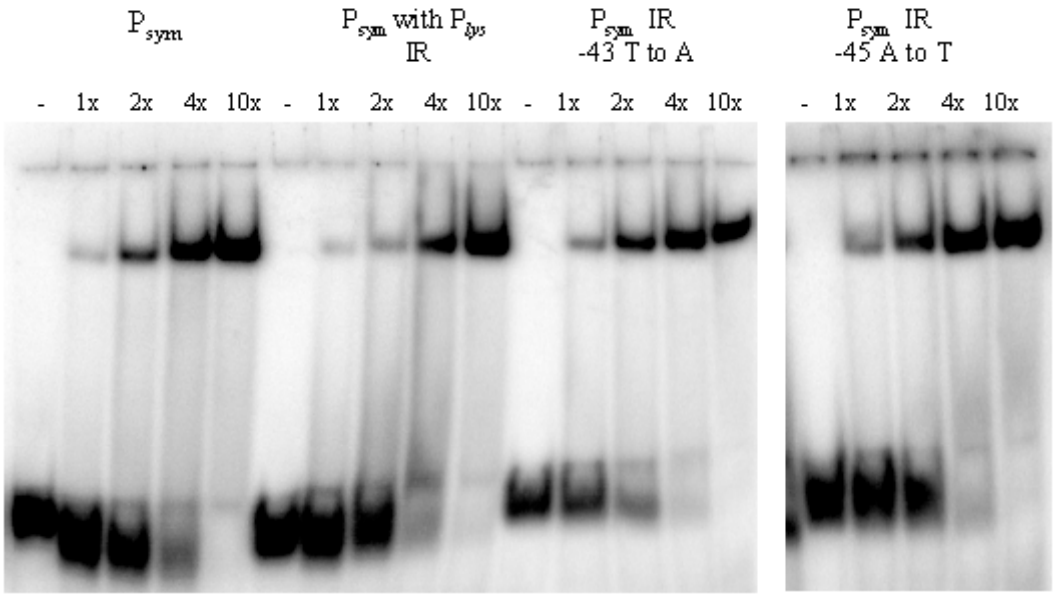
Figure 2-5. Gel-shift assay for mutants altered at –53 and –52. Labeled wild- type and mutant P_{sym} 30-mer probes were incubated with 0 ng (-), 20 ng (1x), 40 ng (2x), 80 ng (4x), 200 ng (10x) with WT-C protein. The bases present at positions –53 and –54 are shown above.

Figure 2-6. Gel-shift assay for IR spacer mutants. Labeled wild-type and mutant P_{sym} 30-merprobes were incubated with 0 ng (-), 20 ng (1x), 40 ng (2x), 80 ng (4x), 200 ng (10x) with WT-C protein. The mutant bases at positions –43 and –45 are shown above. (A) Gel-shift of P_{sym} with $P_{\textit{lys}}$ IR and $P_{\textit{lys}}$ with P_{sym} IR, (B) and (C) P_{sym} IR spacer mutants.

A


P_{sym}
P_{sym} with P_{lys} IR
P_{lys} with P_{sym} IR
- 1x 2x 4x 10x - 1x 2x 4x 10x - 1x 2x 4x 10x

B


P_{sym}
P_{sym} with P_{lys} IR
P_{sym} IR -43 T to A
P_{sym} IR -45 A to T
- 1x 2x 4x 10x - 1x 2x 4x 10x - 1x 2x 4x 10x - 1x 2x 4x 10x

C

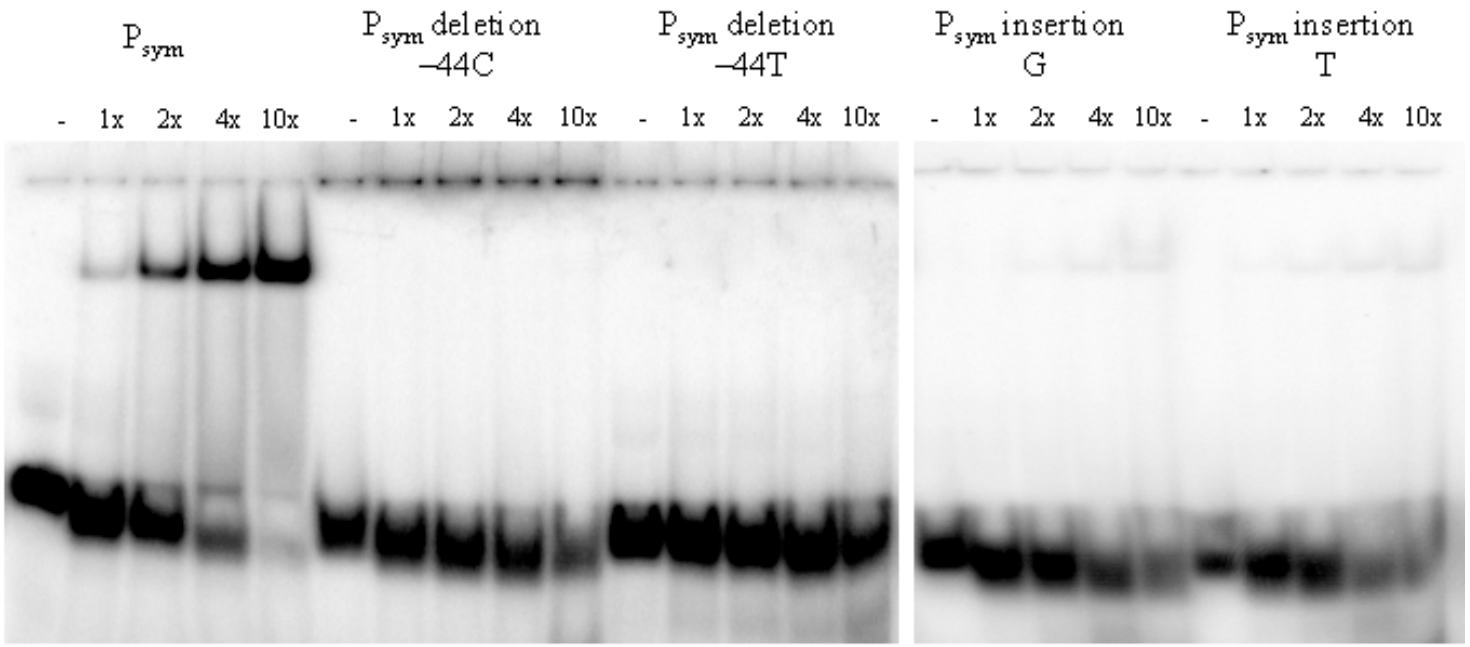


Figure 2-6 (continued).

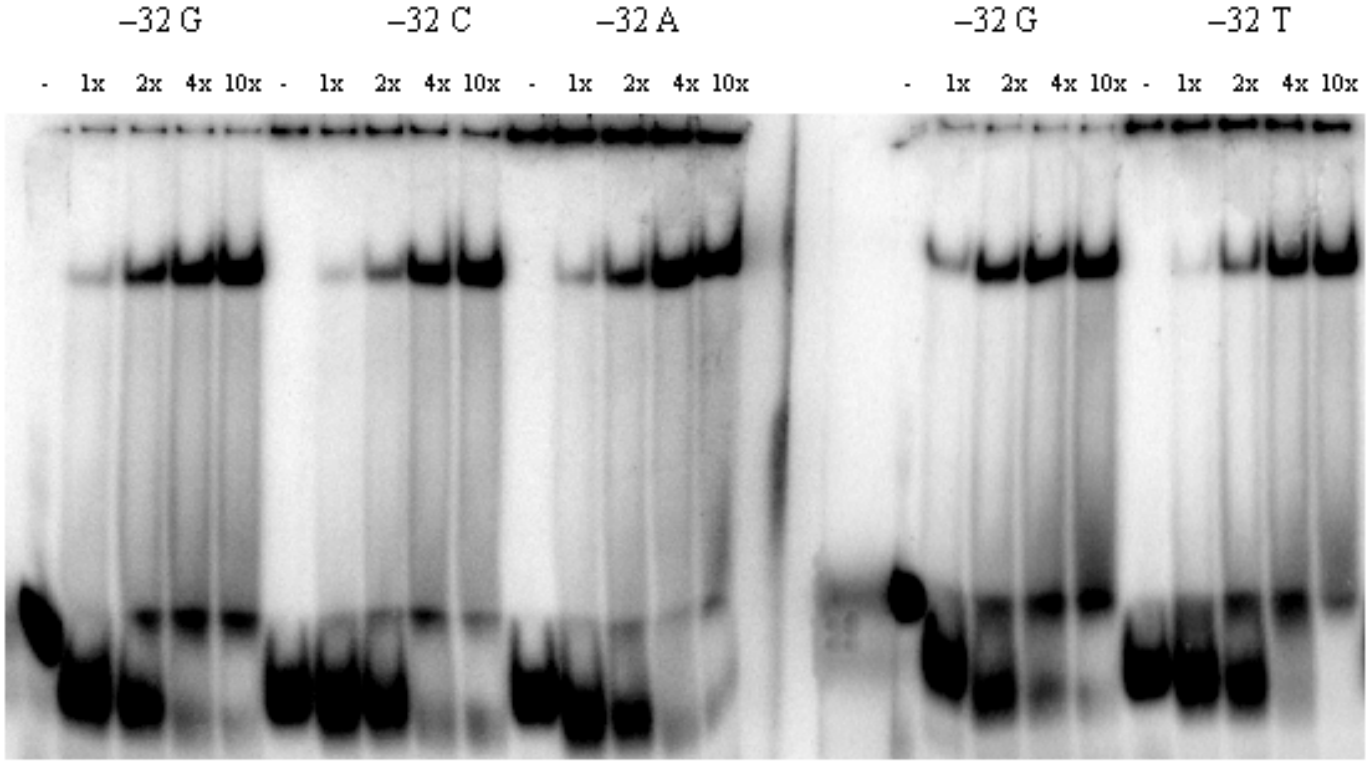


Figure 2-7. Gel-shift assay for mutants altered at –32. Labelled wild-type and mutant P_{sym} 30-merprobes were incubated with 0 ng (-), 20 ng (1x), 40 ng (2x), 80 ng (4x), 200 ng (10x) with WT C protein. The altered bases at position –32 are shown

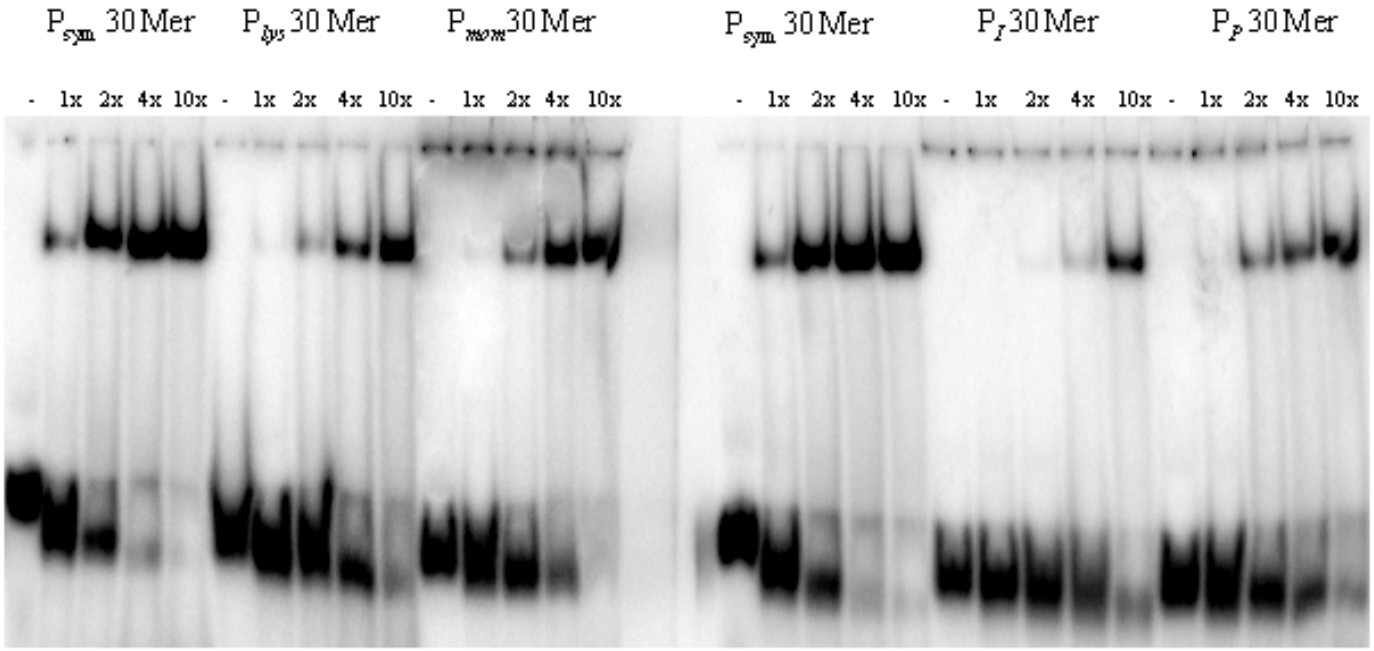
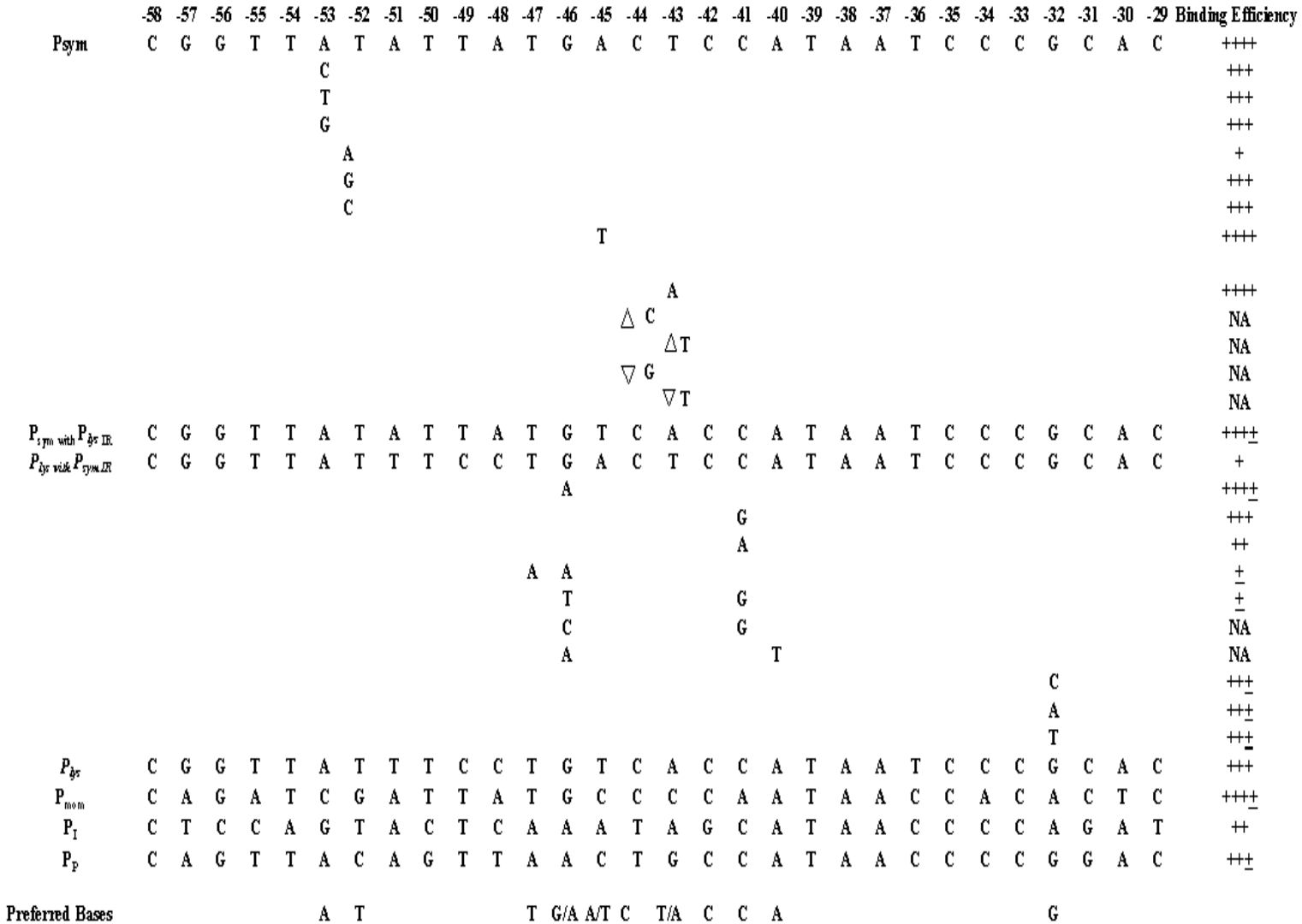


Figure 2-8. Gel-shift assay to test the relative binding affinity of C to P_{sym} and Mu late promoters. Labeled P_{lys}, P_I, P_P, P_{mom} and P_{sym} 30-merprobes were incubated with 0 ng (-), 20 ng (1x), 40 ng (2x), 80 ng (4x), 200 ng (10x) of WT C protein.

Figure 2-9. Summary of the gel-shift assay results. The top strand of the C footprint region in P_{sym} from –58 to –29 is shown on the top of the figure. The candidate mutants tested in the assay are listed at their respective positions below the P_{sym} sequence. The symbol '▲' refers to a deletion in the sequence and '▼' refers to an insertion in the sequence. The abbreviation "NA" refers to "no activity".


-58 -57 -56 -55 -54 -53 -52 -51 -50 -49 -48 -47 -46 -45 -44 -43 -42 -41 -40 -39 -38 -37 -36 -35 -34 -33 -32 -31 -30 -29 Binding Efficiency
Psym C G G T T A T A T T A T G A C T C C A T A A T C C C G C A C ++++
C +++
T +++
G +++
A +
G +++
C +++
T ++++
A ++++
△ C NA
△T NA
▽ G NA
▽T NA
$P_{sym\ with}P_{dyr\ IR}$ C G G T T A T A T T A T G T C A C C A T A A T C C C G C A C +++±
$P_{dyr\ with}P_{sym\ IR}$ C G G T T A T T T C C T G A C T C C A T A A T C C C G C A C +
A +++±
G +++
A ++
A ±
T ±
G ±
C NA
A NA
C ++±
A ++±
T ++±
P_{dyr} C G G T T A T T T C C T G T C A C C A T A A T C C C G C A C +++
P_{mom} C A G A T C G A T T A T G C C C C A A T A A C C A C A C T C ++++±
P_I C T C C A G T A C T C A A A A T A G C A T A A C C C C A G A T ++
P_P C A G T T A C A G T T A A C T G C C A T A A C C C C G G A C ++±
Preferred Bases A T T G/A A/T C T/A C C A G

to identify possible additional flanking sequences that influence C binding.

In the present study, there is enough evidence suggesting that length of the C-binding sequence is important for efficient C-binding. When the length of the C-binding sequence was gradually reduced from a 30-mer to 18-mer, C binding was also reduced even though the core sequences known to be required for binding were intact. This result suggests that the bases and phosphodiester backbone adjacent to the core sequences may be required for stabilizing the C : DNA interactions.

The imperfect dyad-symmetrical C-binding site (–36 to –51) was identified by extensive biochemical analyses. In these assays, single mutations at positions –52 or –53 had little effect on C-binding (Zhao, 1999). But Mo (2004) found that both positions showed increased C binding if –53 and –52 had a T. In this study, a conclusion can be reached that for these positions no bases are particularly preferred for C binding. The possible explanation as to why differences were noticed in previous assays may be attributed to how the experiment was set up. It is known that C-binding to its DNA is pH dependent and both previous studies used an optimal and a sub-optimal pH for C-binding assays.

Mutations in the sequences flanking the IR (–47, –46, –41, –40) in the present study confirmed and extended the previous results (Zhao, 1999) to show that symmetrical base pairs (T-A, –47/–40) and (G-C,–46/–41) are important binding of C. Gel-shift analysis of the seven candidate mutants isolated by plate phenotyping showed only one mutant having almost WT activity. This mutant has WT bases at all positions except –46 A. Previously in gel-shift assays it was shown that –46 A is a down mutant. Therefore, it is not surprising that the present study did validate the previous study that the best

binding site for C is a perfect inverted repeat.

The most surprising mutants that completely abolished or greatly diminished C binding were in the IR spacer sequence found between the inverted repeats. By switching the P_{sym} IR with P_{lys} IR and vice-versa, a profound effect in C-binding was noticed. The binding effect was first thought to be a combined sequence effect since P_{lys} was already known to bind C less effectively than P_{sym}. So, to identify the effect of the spacer sequence, point mutants were generated at positions –43 and –45 of the P_{sym} IR. Both mutations did not reduce C binding dramatically, which suggested that the two mutations in the IR might disrupt possible C : DNA backbone interactions or change the local DNA architecture, which may destabilize C binding. However, G or T insertion in the spacer (or) deletion of C or T within the IR spacer completely abolished C binding. This effect may be due to the repositioning of the major groove in relation to the dyad symmetry since it is known that the C-binding site needs to be centered on –43.5. By taking into account the previous data and the present data a possible consensus sequence can be derived with the following characteristics; (1) the dyad symmetry element should be perfectly symmetrical (2) the IR spacer length requirement is absolute.

The next chapter will describe the results of C : P_{sym} co-crystallization experiments done in order to obtain a three-dimensional structure which might reveal possible protein – DNA interactions and any associated conformational changes.

Chapter 3. Expression, Purification, Crystallization and Preliminary X-ray Analysis of C Protein Bound to P_{sym} DNA*

Introduction

Escherichia coli K-12 and other enteric bacteria are hosts for bacteriophage Mu. Mu is a temperate phage that randomly inserts its 39-kb double-stranded linear DNA into the host DNA and may enter a lysogenic or a lytic life cycle depending on the host environment. The lysogenic cycle is mainly under the control of the Mu c repressor protein. When repression is released, the phage starts replicating and proceeds through the lytic cycle. The lytic cycle is tightly controlled by a transcriptional cascade; early, middle and late (Goosen, 1987; Stoddard and Howe,1989; Marrs and Howe, 1990). Early transcription starts from P_e and does not require *de novo* protein synthesis or DNA replication. Middle transcription is dependent upon Mor, an activator protein expressed from the Mu early transcript (Stoddard and Howe, 1989; Mathee and Howe, 1990; Marrs and Howe, 1990). The middle transcript codes for the C protein, the activator of the four late promoters P_{lys}, P_I, P_P, and P_{mom}. The C protein binds a dyad-symmetry element just upstream of the –35 region from –52 to –32 on all four late promoters (Margolin *et al.*, 1989; Sun *et al.*, 1997; Zhao, 1999). It has been shown that a C dimer (Ramesh and Nagaraja, 1996; De *et al.*, 1997; Zhao, 1999) is able to bind the dyad-symmetry element.

The C protein (140 amino-acids, 16.5 kDa monomer) is a close homologue of the Mu Mor protein. Both proteins share high sequence similarity with each other (Figure

* Modified with permission. Shanmuganatham, K. K., Ravichandran, M., Howe, M. M., and Park, H.W. (2007). Crystallization and preliminary X-ray analysis of phage Mu activator protein C in a complex with promoter DNA. *Acta Cryst.F* 63, 620-623.

1-10, Chapter 1). Secondary structural analysis of Mor and C proteins using the algorithm of Dodd and Eagan (Dodd and Egan, 1987) revealed that both proteins contain a C-terminal HTH DNA-binding domain (Mathee and Howe, 1990). The crystal structure of His-Mor revealed that a Mor monomer is made up of an N-terminal dimerization domain and a C-terminal HTH DNA-binding domain (Kumaraswami *et al.*, 2004). Dimerization between the two monomers occurs by intertwining the N-terminal helices (Figure 1-10 Chapter 1). The two flanking HTH DNA-binding domains are proposed to bind two adjacent major grooves. Since the predicted DNA-binding residues of Mor are too far apart to fit into two adjacent major grooves, DNA binding may be associated with a conformational change in the Mor dimer. Since no protein : DNA complex structure has been determined for either Mor or C protein, structural analysis of Mu C protein bound to a synthetic late promoter P_{sym} (Jiang, 1999a) was undertaken. The binary complex structure will provide a direct test of the predicted amino-acid-DNA interactions and associated protein and DNA conformational changes that have been proposed for DNA-bound Mor and C proteins, based on the structure of Mor in the absence of DNA (Kumaraswami *et al.*, 2004). The objective of this study is to crystallize and solve the structure of the Mu C : DNA complex.

Materials and methods

Chemicals, enzymes and media

Standard bacterial cell growth and protein over-expression were done in Luria-Bertani (LB) medium (Sambrook *et al*, 1989) containing chloramphenicol (Cm) at 34

μg/ml. Chloramphenicol, EDTA, Hepes, tris base, glycine, dithiothreitol (DTT), magnesium chloride ($MgCl_2$), sodium chloride (NaCl), sodium dodecyl sulphate (SDS), boric acid, M9 minimal media, amino-acids L-lysine, L-phenylalanine, L-threonine, L-isoleucine, L-leucine, L-valine and L-selenomethionine were purchased from Sigma. Isopropyl-β–D-thiogalacto-pyranoside (IPTG) was obtained from American Bioorganics. Ready Gel™ precast gels, acrylamide, bisacrylamide, low molecular weight precision plus protein standard, Bio-Safe™ Coomassie, tetramethylethylenediamine (TEMED), ammonium persulfate (APS) and Bradford reagent were purchased from BioRad. Glycerol was obtained from Fisher Scientific. Prepacked Hi-Trap Heparin column, SP Sepharose Fast Flow (FF) resin, Phenyl Sepharose FF column, Superdex-75 26/60 and 16/160 gel-filtration columns and G50 Probe Quant™ Sephadex column were from GE Healthcare Bio-Sciences AB. YM30 ultrafiltration membranes and Amicon concentrators were from Amicon Bioseparation. Radiolabelled [γ-^{32}P] ATP (3000 Ci/mmol) was from Perkin Elmer life sciences; T4 polynucleotide kinase was from Promega and New England Bioloab. The automated DNA sequencing was performed by the Molecular Resource Center of The University of Tennessee Health Science Center. Electrospray mass spectrometry was performed by the Hartwell center for Biotechnology and Bioinformatics of St. Jude Children's Research Hospital.

Bacterial strains and plasmids

Plasmid pZZ41 (Figure 3-1) containing the Mu C gene under the control of a T7 promoter was constructed by (Zhao, 1999) for efficient protein expression. The bacterial strains used for protein expression is listed in Table 3-1.

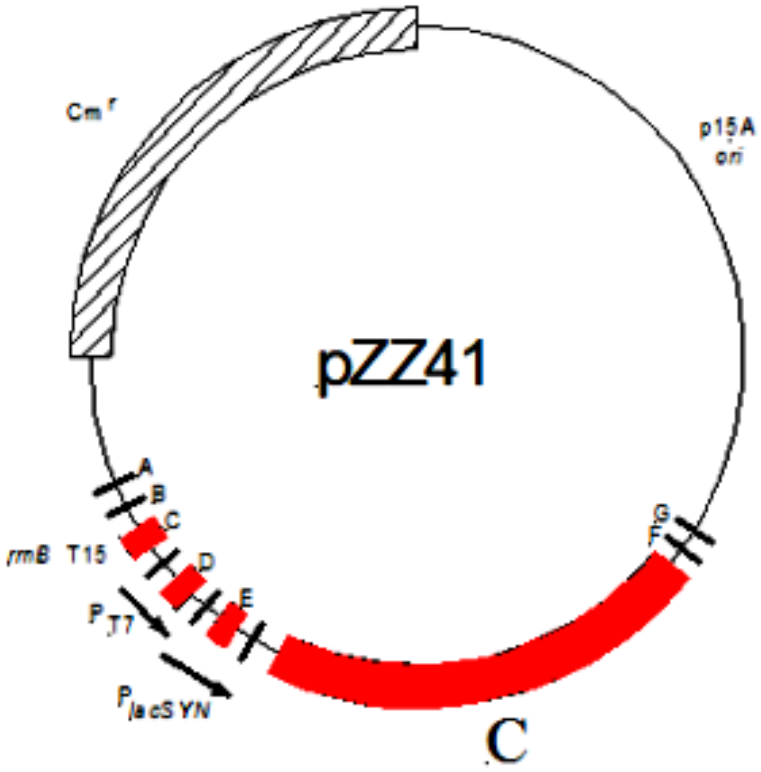


Figure 3-1. Circular plasmid map of pZZ41. Plasmid pZZ41 is a pACYC derivative containing the P_{T7} and P_{lacSYN} promoters upstream of the *C* gene.

Table 3-1. Bacterial strains.

Strain	Strain genotype	Reference/ derivation
JM109DE3	mcrA ΔproAB-lac thi gyrA endA hsdR relR supE44 recA; F' (traΔ36 lacI^{Q} ΔlacM15 pro^{+}) λDE3[a]	Yanish-Perron *et al*., 1985
MH13355	mcrA ΔproAB-lac thi gyrA endA hsdR relR supE44 recA; F' (pro^{+} lacI^{Q1} ΔlacZY); λDE3[a]	Artsimovitch and Howe, 1996
MH13312	mcrA ΔproAB-lac thi gyrA endA hsdR relR supE44 recA; F' (pro^{+} lacI^{Q1} ΔlacZY)	Artsimovitch and Howe, 1996

[a]λ DE3 is a derivative of λ D69 containing the T7 RNA polymerase gene under control of the IPTG-inducible P_{lacUV5}; it also carries *imm*21 and Δ*nin*5 mutations. Yanisch-Perron, C., Vieira, J., and Messing, J. (1985) Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* 33: 103-119. Artsimovitch, I., and Howe, M.M. (1996) Transcription activation by the bacteriophage Mu Mor protein: analysis of promoter mutations in Pm identifies a new region required for promoter function. *Nucleic Acids Res* 24: 450-457.

Oligodeoxyribonucleotides

Table 3-2 includes the oligodeoxyribonucleotides used in electro-mobility shift assays (EMSA) and crystallization. Synthesis of the oligodeoxyribonucleotides was done by Integrated DNA Technologies, Inc. (IDT) on commercial nucleic acid synthesizers (Model ABI394) using phosphoramidite chemistry (Caruthers *et al.*, 1983).

Crystallization

The 24-well polystyrene Linbro plates for hanging and sitting drop methods, square siliconized coverslips, forceps, anodized cryo loop tools for crystal manipulation and sealing tapes were purchased from Hampton Research. Crystal screen II, Lite screen, PEG/Ion Screen and I Additive screen were purchased from Hampton Research. Wizard

Table 3-2. Oligodeoxyribonucleotides used for EMSA and crystallization.

Primer	Sequence	Comments
KAR 12	ATTATGACTCCATAATCC	Top strand P_{sym} 18-mer
KAR 13	GGATTATGGAGTCATAAT	Bottom strand P_{sym} 18-mer
KAR 14	TTCCTGTCACCATAATCC	WT P_{lys} 18-mer
KAR 15	GGATTATGGTGACAGGAA	Bottom strand P_{lys} 18-mer
KAR 16	TTTTATTATGACTCCATAATCCCG	Top strand primer from –55 to–32 of P_{sym} 24-mer
KAR 17	CGGGATTATGGAGTCATAATAAAA	Bottom strand P_{sym} 24-mer
KAR 18	TTATTATGACTCCATAATCC	Top strand primer from –53 to –34 of P_{sym} 20-mer
KAR 19	GGATTATGGAGTCATAATAA	Bottom strand P_{sym} 20-mer
KAR 20	ATTTCCTGTCACCATAATCC	Top strand P_{lys} 20-mer sequence from –53 to –34
KAR 21	TAAAGGACAGTGGTATTAGG	Bottom strand P_{lys} 20-mer
KAR 22	TTATTTCCTGTCACCATAATCCCG	Top strand P_{lys} 24-mer from –53 to –32
KAR 23	CGGGATTATGGTGACAGGAAATAA	Bottom strand P_{lys} 24-mer
KAR 24	ATATTATGACTCCATAATCC	Top strand P_{sym} 20-mer from –53 to –34.
KAR 25	GGATTATGGAGTCATAATAT	Bottom strand P_{sym} 20-mer
KAR 26	TATATTATGACTCCATAATCCC	Top strand P_{sym} 22-mer from –54 to –33.
KAR 27	GGGATTATGGAGTCATAATATA	Bottom strand P_{sym} 22-mer
KAR 28	TTGTATTATGACTCCATAATCCCA	Top strand P_{sym} 24-mer with –53G mutation.
KAR 29	TGGGATTATGGAGTCATAATACAA	Bottom strand P_{sym} 24-mer with –53 G mutation.
KAR 30	TGTATTATGACTCCATAATCCC	Top strand P_{sym} 22-mer with –53G mutation
KAR 31	GGGATTATGGAGTCATAATACA	Bottom strand P_{sym} 22-mer with –53 G Mutation

Table 3-2 (Continued).

Primer	Sequence	Comments
KAR 32	GTATTATGACTCCATAATCC	Top strand P_{sym} 20-mer
KAR 33	GGATTATGGAGTCATAATAC	Bottom strand P_{sym} 20-mer
KAR 34	TATTATGACTCCATAATC	Top strand P_{sym} 18-mer
KAR 35	GATTATGGAGTCATAATA	Bottom strand P_{sym} 18-mer
KAR 36	GTATTATGACTCCATAATCCGG	Top strand P_{sym} 20 plus 2 base overlap at the 3′ end
KAR 37	GGATTATGGAGTCATAATACCC	Bottom strand P_{sym} 20 plus 2 base overlap at the 3′ end
KAR 61	GTATTATGACTCCATAATCCG	Top strand P_{sym} 21-mer i.e. 20 mer with 1 base overlap at the 3′ end.
KAR 62	GGATTATGGAGTCATAATACC	Bottom strand P_{sym} 21-mer i.e. 20 mer with 1 base overlap.
KAR 63	AGATTATGATATCATAATCTG	P_{sym} 21-mer, symmetrical sequence with 1 base overlap at the 3′ end
KAR 64	AGATTATGATATCATAATCTC	Bottom strand P_{sym} 21-mer, symmetrical sequence with 1 base overlap at the 3′ end
KAR 75	GTTATATTATGACTCCATAATCCCGC	Top strand P_{sym} 26-mer
KAR 76	GCGGGATTATGGAGTCATAATATAA C	Bottom strand P_{sym} 26-mer
KAR 77	CGGTTATATTATGACTCCATAATCCC GCAC	Top strand P_{sym} 30-mer
KAR 78	GTGCGGGATTATGGAGTCATAATAT AACCG	Bottom strand P_{sym} 30-mer
KAR 126	CGGTTATTTCCTGTCACCATAATCCC GCAC	Top strand P_{lys} 30-mer
KAR 127	GTGCGGGATTATGGTGACAGGAAAT AACCG	Bottom strand P_{lys} 30-mer

[a] The oligodeoxyribonucleotide sequences are written from 5’ to 3’.

Screen I and II were purchased from Emerald Biosciences, the customized screens SGC and Redwing were from the Structural Genomics Consortium SGC. Additional solutions for refinement were made from chemicals of minimum purity ACS grade and were stored in 4° C or room temperature based on their chemical properties and manufacturer's recommendations.

Wild-type C protein production

Expression and solubility test

The expression and solubility of the WT C protein produced from pZZ41 was tested as follows. Plasmid pZZ41 was freshly transformed into JM109DE3 and a single colony was inoculated into 10 ml of LB with chloramphenicol at 34 μg /ml and grown at 37° C overnight. The next morning the cells were used to inoculate a 100 ml culture, which was grown at 37° C at 225 rpm until the OD_{600} reached 0.4 to 0.5. At this point a 1ml sample was taken and kept on ice to serve as an un-induced control. The rest of the culture was induced with 1 mM IPTG. Every hour a 1ml sample was taken for use as an induced sample. After 3 hours, the uninduced and induced cell cultures were centrifuged at 6000 RCF (Sorvall GSA 6000), and the cell pellets were kept on ice before proceeding to cell lysis. The pellet was resuspended in 1 ml of lysis buffer. The resuspended cells were lysed by sonication on ice using continuous cycle (4 times X 3 min Duty cycle 40 and Output control 50). The lysed cells were centrifuged at 6000 RCF for 30 min to separate the supernatant from the pellet. The pellets were resuspended in 1 ml of buffer C. Then 30 μl of the sonicated supernatant and the resuspended pellet were loaded on to

10-20 Ready Gel™ precast gels and electrophoresed in 1X tris glycine buffer for one hr at 10V/cm and then stained with Bio-Safe™ Coomassie dye. The expression of WT C was also tested by varying the induction temperature from 16 to 32° C, expressing the protein in different expression hosts and in different media, including Terrific broth (Tartof, 1987)

Large-scale production

Based on the results of the expression tests, large-scale production was done with 8-20 liters of LB medium supplemented with 34 μg /ml chloramphenicol. One hundred milliliters of overnight culture were used to inoculate 800 ml of LB in a 2-liter flask. The cells were grown at 37° C shaking at 225 rpm until the OD_{600} reached 0.4 to 0.5. At this stage, protein expression was induced with 1 mM IPTG. After three hours of expression, the cells were harvested by centrifugation at 6000 RCF for 15 min (Sorvall GSA 6000); the pellets were either processed immediately or frozen in liquid nitrogen and stored at -70° C. The final protein yield depended mainly on the OD and culture volume induced.

Cell lysis

Fresh or thawed cell pellets were resuspended in lysis buffer and the cells were lysed by using a Microfluidics microfluidizer HC-8000. The cell suspension was passed through the microfluidizer twice to promote efficient cell lysis. Once lysed the cell suspension was subjected to centrifugation at 20,000 RCF for 30 min at 4° C (Sorvall SS34) to remove cell debris. Aliquots of the supernatant and insoluble fraction were saved for further analysis.

Chromatography

Wild-type C and protein : DNA complexes were purified using fast performance liquid chromatography (FPLC) in an ÄKTA FPLC machine (GE Healthcare Bio-Sciences AB). All buffers used hereafter were degassed by stirring under a vacuum. The over-expressed C protein was purified with a four-step chromatography procedure as described below. The composition of the protein buffer used for purification is listed in Table 3-3.

Heparin-sepharose affinity chromatography

Heparin is a sulphated polysaccharide that mimics the binding properties of DNA and can be used as a first step to purify DNA binding proteins. In addition, heparin has a high density charge on its surface, so it can also be used as an ion exchanger. Due to the latter property, the ionic strength of the buffer used should be low, and the pH should be near pH 7.

The lysed samples were first filtered through a 0.45 μm filter to remove cell debris. This is very important because an unfiltered sample will clog the column. The column, a prepacked 5-ml Hi-trap heparin FF column (GE Healthcare Bio-Sciences AB) was equilibrated with 100 ml of Heparin buffer A before the samples were loaded. To ensure maximum binding of the samples loaded onto the column, the flow rate was adjusted to 2ml/min and the flow through (FT) was kept for analysis. Elution of the bound protein was done using a linear gradient from zero to 500 mM NaCl over a 200 ml volume with a flow rate of 2 ml/min, and 5 ml fractions were collected. C protein eluted over a range from 200 to 300 mM NaCl, with the peak elution about 250 mM NaCl.

Table 3-3. Buffers for protein purification.

Buffer name	Buffer composition
Cell lysis buffer	25 mM Hepes, pH 7.0, 50 mM NaCl, 5% glycerol, 1 mM $MgCl_2$, 1 mM EDTA, 1 mM DTT
Hi-Heparin affinity chromatography buffer A	Heparin buffer A: 25 mM Hepes, pH 7.0, 50 mM NaCl, 5% glycerol, 1 mM $MgCl_2$, 1 mM EDTA, 1 mM DTT
Hi-Heparin affinity chromatography buffer B	25 mM Hepes, pH 7.0, 500 mM NaCl, 5% glycerol, 1 mM $MgCl_2$, 1 mM EDTA, 1 mM DTT
SP-Sepharose Cation exchange chromatography buffer A	25 mM Hepes, pH 7.0, 100 mM NaCl, 5% glycerol, 1 mM $MgCl_2$, 1 mM EDTA, 1 mM DTT
SP-Sepharose Cation exchange chromatography buffer B	25 mM Hepes, pH 7.0, 500 mM NaCl, 5% glycerol, 1 mM $MgCl_2$, 1 mM EDTA, 1 mM DTT
Phenyl Sepharose Hydrophobic chromatography buffer A	25 mM Hepes, pH 7.0, 1.5 M NaCl, 5% glycerol, 1 mM $MgCl_2$, 1 mM EDTA, 1 mM DTT
Phenyl Sepharose Hydrophobic chromatography buffer A	25 mM Hepes, pH 7.0, 0 mM NaCl, 5% glycerol, 1 mM $MgCl_2$, 1 mM EDTA, 1 mM DTT
Gel-filtration buffer A for protein purification	25 mM Hepes, pH 7.0, 150 mM NaCl, 5% glycerol, 1 mM $MgCl_2$, 1 mM EDTA, 10 mM DTT
Gel-filtration buffer B for complex purification	25 mM Hepes, pH 7.0, 75 mM NaCl, 5% glycerol, 1 mM $MgCl_2$, 1 mM EDTA, 10 mM DTT
Buffer C	25 mM Hepes, pH 7.0, 75 mM NaCl, 5% glycerol, 4.5 mM $MgCl_2$, 1 mM EDTA, 10 mM DTT

SP-sepharose cation exchange column

Ion exchange chromatography is based on the electrostatic properties of the protein that bind to the charged surface group on the resin. For efficient electrostatic binding the total ionic strength has to be low (~50 mM). The above eluted proteins fractions were pooled and diluted with C buffer without NaCl so that the final salt concentration was 100 mM. This step is crucial since a minimum concentration of 100 mM NaCl is required for C to bind to the column. The sample was then loaded onto a pre-equilibrated open SP-Sepharose cation exchange column (50 ml). The flow-through was kept for gel analysis. For elution, a linear gradient of 100 to 500 mM NaCl over a 200-ml volume was used with a flow rate of 5 ml/min and fraction volume of 5 ml. The C protein eluted from 200 to 300 mM NaCl with the peak elution about 275 mM NaCl.

Phenyl-sepharose hydrophobic exchange column

Here separation of biomolecules is based on their hydrophobicity. When the biomolecules in a polar solvent are applied to a hydrophobic matrix, they establish a strong interaction with it. Elution is done by gradually reducing the polarity of the solution.

This step was used to remove some of the contaminating protein not removed through the previous two purification steps. The pooled SP-Sepharose fraction was diluted with 5 M NaCl so that the final salt concentration was 1.5 M NaCl. The diluted pooled samples were then loaded onto a pre-equilibrated the Phenyl-Sepharose FF hydrophobic column (5 ml) (GE Healthcare Bio-Sciences AB). This high salt concentration was required for the protein to bind the matrix. The bound protein was

eluted with a linear gradient of 1.5 M to 0 mM NaCl over 110 ml with a flow rate of 2ml/min with a fraction volume of 5ml was used. C protein eluted from around 700 to 150 mM NaCl.

Gel-filtration or size-exclusion chromatography

Gel-filtration chromatography (GFC) is based on the size of the protein and the sieving properties of the gel-filtration matrix. Usually the GFC matrix has an exclusion limit based on which proteins of different sizes can be separated. High molecular weight protein elutes first and the protein with the lowest molecular weight elutes last.

Gel-filtration was used as a final polishing step and to exchange the protein buffer during C protein purification and for purification of the protein DNA complex. This procedure was done on both an analytical and preparative scale. Analytical runs were done with Superdex 75 16/60 (GE Healthcare Bio-Sciences AB) and preparative runs in Superdex 75 26/60(GE Healthcare Bio-Sciences AB).

The pooled fractions from the hydrophobic separation step were first filtered through a 0.22 μM filter to remove protein aggregates and then loaded onto a column pre-equilibrated with GF buffer A or B. Wild-type C usually migrates as a single peak of 32 kDa (the calculated dimer mass is 33 kDa) whereas the purified C : DNA complex migrates around 47 kDa.

The peak C fractions or C: DNA fractions were pooled and concentrated to 30 mg/ml using YM30 Amicon concentrators at room temperature (Amicon Bioseparation, Bedford, Massachusetts, USA) and stored at –70°C.

The purity of the purified protein was visually examined on a Ready Gel™ precast

gels stained with Bio-Safe™ Coomassie. The concentration of C protein and the C : DNA complex was measured by the Bradford method (Bradford, 1976).

Selenomethionine C protein production

The following modifications were introduced into the basic C protein expression protocol described above to incorporate selenomethionine (SeMet) into C protein during expression. This modification is based on the metabolic inhibition method described by Van Duyne *et al*. (1993). A single colony was inoculated into 150 ml of LB medium containing 34 μg/ml chloramphenicol and grown overnight at 37° C. The next day the cells were collected and resuspended in 2L M9 minimal medium (Symonds, 1987) with 34 μg/ ml chloramphenicol and grown for another 12 hrs or overnight. This culture was used to seed 8 L of minimal medium and grown at 37° C. When the cells reached an OD_{600} of 0.6, the following amino-acids (all from Sigma) were added: 800 mg each of L-lysine, L-phenylalanine, and L-threonine, as well as 400 mg each of L-isoleucine, L -leucine, L-valine and L-selenomethionine. After the culture was shaken for 15 min, protein expression was induced by adding IPTG to a final concentration of 1 mM, and the culture was grown for 12 hr at 37° C. Cell lysis and purification of the SeMet C protein were performed as described above for wild-type C.

Electrospray mass spectrometry (ESI-TOF) was used to confirm the incorporation of two SeMet residues in place of the two naturally occurring methionine residues in C protein.

Electrophoretic mobility shift assays

The ability of C to bind double-stranded DNA of different lengths and optimum binding conditions were determined by electrophoretic mobility shift assay (EMSA) (Carey, 1991). One hundred nanograms of top or bottom stand oligonucleotides used for crystallization were end labeled with T4 polynucleotide kinase γ^{32} P ATP (3000 Ci/mmol) in polynucleotide kinase buffer. The labeled oligos were then annealed with 300 to 500 ng of the complementary bottom or top strand oligos by placing the mixture on a 100° C heat block for 2 min and then switching the block off to let it cool to room temperature. The annealed probes were purified using a G50 Probe Quant™ Sephadex column as per the manufacturers recommendation. The purified labeled probes were incubated at 25° C for 30 min with and without purified wild-type C in 20 µl of buffer C. After incubation, the binding reaction was loaded on to a 8% non-denaturing acrylamide gel in 1X TBE buffer and subjected to electrophoresis for 2 to 3 hrs at 4° C at 10V/cm. The gel was the blotted onto Whatman filter paper and exposed to Kodak Biomax™ MR without a screen at -70° C overnight.

Preparation of C : DNA complex for crystallization

Oligonucleotides used for crystallization were obtained as separate top and bottom strands from IDT (Integrated DNA Technologies, Coralville, IA). Equimolar amounts of the top and bottom strand were dissolved in buffer C, mixed and annealed in a thermal cycler. Efficient annealing was achieved by first incubating the oligodeoxyribonucleotides at 5° C above the predicted T_m of an oligonucleotide for 5 min

and next at 5° C below the predicted T_m for another 5 min; then rapidly cooling to room temperature.

Protein : DNA complexes were formed by mixing WT C protein with double stranded DNA in a 1:1 molar ratio of C dimer to DNA. The formed complex was purified by using a Superdex 75 26/60 size-exclusion column (GE Healthcare Bio-Sciences AB) pre-equilibrated with GF buffer B. After elution the purified complex was concentrated to 25-30 mg/ml using YM30 Amicon concentrators at RT (Amicon Bioseparation, Bedford, MA) and stored at -70° C.

Crystallization

Crystallization trials with the complex were done by both the hanging and the sitting drop methods as described above. The concentrated C : DNA complex at 25-30 mg/ml was centrifuged at 13,000 rpm for 5 min in a tabletop centrifuge to remove aggregates, precipitates and dust. Depending on the method used for crystallization, the amount of complex solution mixed with the reservoir solution was varied (1 µl: 1 µl for hanging drop and 0.5 µl: 0.5 µl for sitting drop). In case of the sitting drop method, the plates were sealed with sealing tape; in case the of hanging drop the wells were sealed with siliconized cover slips.

Crystallization was attempted at different temperatures including 18° C, 22° C and 37° C. During the first week of the crystallization trial, the trays were examined every day; later they were examined once or twice a week. Initial screens were done with commercial screens such as Wizard screen™ I & II (Emerald Bioscience), Crystal screen™, Crystal screen™ lite, PEG/Ion screen (Hampton Research) and customized

screens SGC / Redwing (Structural Genomics Consortium SGC). Initial crystallization conditions were fine tuned using very methodical grid screening in which the primary precipitant, additives, pH, temperature and the length of the DNA used in C : DNA complexes were varied.

Screening, data collection and processing

The crystals were mounted in cryo loops after transiently dipping in a cryo protectant and flash frozen in liquid nitrogen. The quality of the crystals obtained from different C : DNA complexes and conditions was determined by X-ray diffraction analysis at the home source by comparing the diffraction patterns as well as the resolution cutoff. The crystal that showed good pattern and resolution was refrozen in liquid nitrogen. A SeMet peak data set was collected from the using a CCD image plate detector with synchrotron radiation of wavelength 0.9791 Å at beam line 17ID of the Advanced Photon Source (Argonne National Laboratory, Chicago, U.S.A). The distance between the crystal and detector was 300 mm; a total of 360 oscillation images were recorded with exposure times of 10 seconds. The diffraction data was indexed, processed, and scaled with DENZO and SCALEPACK programs in the HKL package(Otwinowski, 1997). The positions of the two SeMet residues were determined by the SOLVE program (Terwilliger and Berendzen, 1996). Preliminary model building and refinement was done with program O (Jones *et al.*, 1991)and REFMAC5 (CCP4, 1994).

Results

Many protocols have been developed previously to express and purify C protein for structural studies (Nagaraja *et al.*, 1988; Bolker *et al.*, 1989; Gindlesperger and Hattman, 1994; Ramesh *et al.*, 1994). A N- terminal truncation version of C protein was expressed in *E coli* BL21DE3 as an insoluble pellet and purified from it using a high salt extraction procedure followed by a specific immuno-affinity chromatography. Additionally, protocols for small scale production of his-tagged C protein in *E coli* using the soluble C fraction have also been developed (Jiang, 1999; Mo, 2004).

Nevertheless, purification of native C protein for structural studies; a prerequisite for structure determination of protein DNA complexes involving WTC and its cognate DNA have not been described to date; the results presented here is the first such procedure. In the following section protein purification from the soluble fraction, complex formation and first crystallization trials and data collection will be described.

Protein purification: wild-type C protein

The *C* gene which was cloned into pACYC expression vector was assayed for expression by transforming into JM109DE3.The C protein expresses very poorly with at most 1-2 mg per litre of culture and is expressed as 50% soluble and 50% insoluble (Figure 3-2). Since WT C is a DNA binding protein a Heparin affinity chromatography was used as the first capture step, for intermediate purification a cation exchange chromatography and a hydrophobic exchange chromatography was used. As a final purification and buffer exchange step, gel-filtration/size exclusion was used.

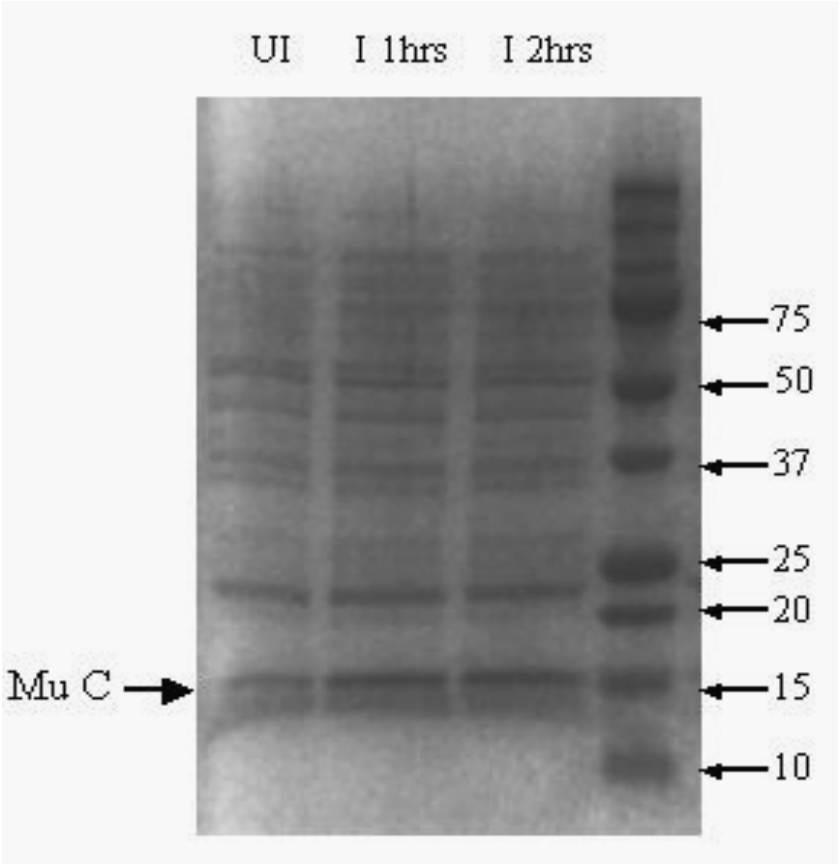


Figure 3-2. Expression gel of C protein. After over-expression the C protein was visualized by SDS-PAGE (8-20%) stained with Bio Safe™ coomassie Blue R250**.** Lanes, UI uninduced, I 1hrs C protein expression after one hour IPTG induction, I 2hrs C protein expression after two hour IPTG induction, and low molecular weight protein marker.

After expression, the cells were either used right away or frozen in -70° C. When protein was to be purified the cell pellet were lysed by using a microfludizer in 200 to 300 ml of C buffer and this process was repeated twice for effective cell lysis. Cell debris and the over expressed proteins were separated by centrifugation. The cell pellet and supernatant was kept in 4° C for subsequent SDS-PAGE analysis.

The C protein and other DNA binding protein in the supernatant were isolated from the Heparin affinity column using fast performance liquid chromatography (FPLC). Since C protein has low binding kinetics with heparin the lysed sample had to be applied a couple of times with a low flow rate for optimum binding to the column. The bulk contaminants not removed from the Heparin affinity chromatography was removed using a cation exchange column and a hydrophobic column as described in the materials and methods The peak fractions from cation and hydrophobic columns were fractionated in 5ml tubes and pooled in preparation for the next stage of purification. Since C protein from the hydrophobic column was eluted out using a huge salt gradient, the protein storage buffer was exchanged using a gel-filtration column (Figures 3-3 and 3-4). After GFC typically the purity was ~95% as visualized from SDS-PAGE. Freezing the protein even in its storage buffer for an extended period denatured the protein as shown by the presence of large amounts of precipitates after thawing the stored protein. The detailed purification procedure is as described in the materials and methods.

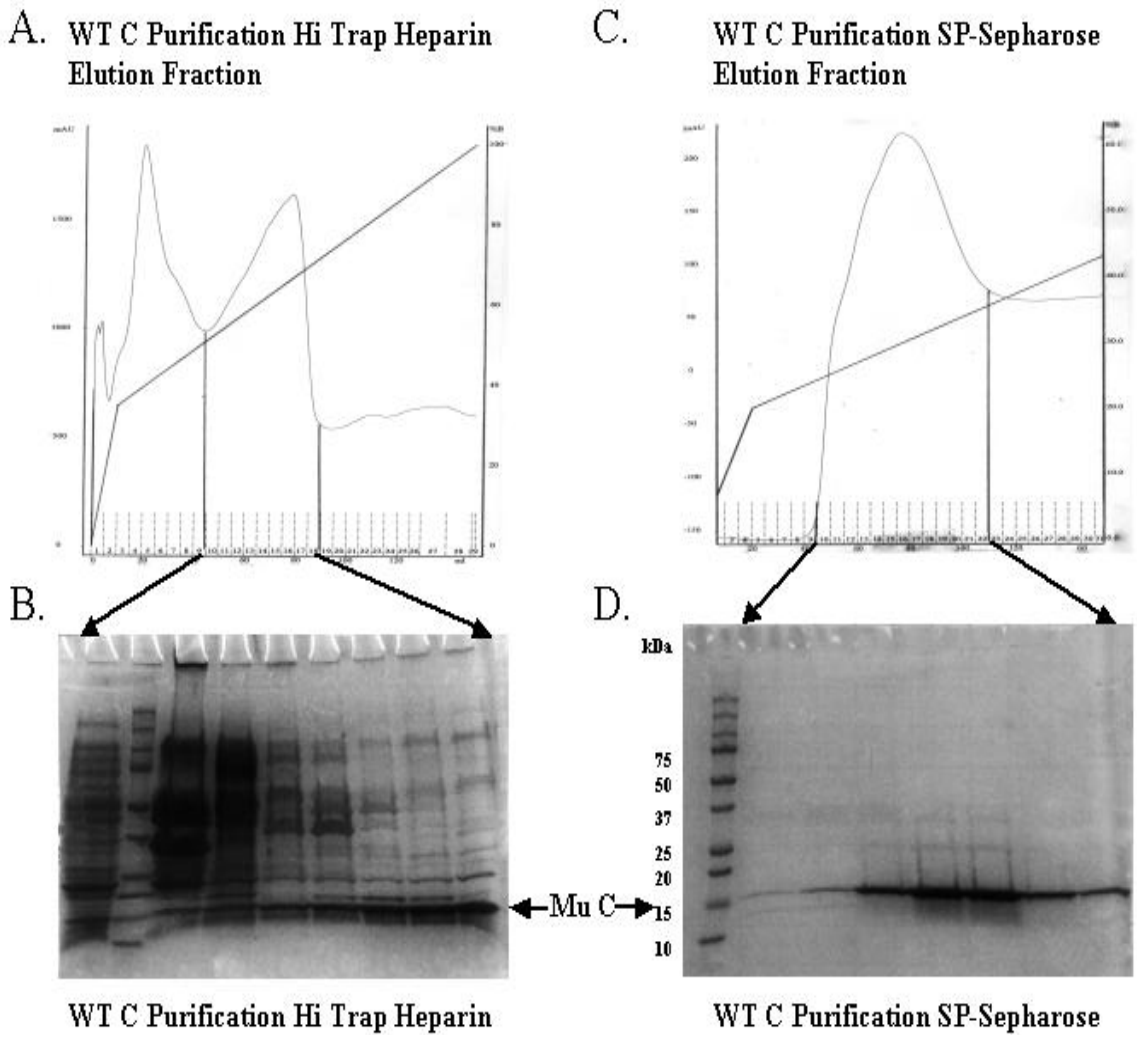


Figure 3-3. WT C purification strategy I. (A) chromatographic elution profile of the hi-trap heparin run, (B) SDS-PAGE (8-20%) of the elution fraction of hi-trap heparin stained with Bio Safe™ coomassie, (C) chromatographic elution profile of the SP-sepharose, and (D) SDS-PAGE (8-20%) of the elution fraction of SP-sepharose stained with Bio Safe™ Coomassie.

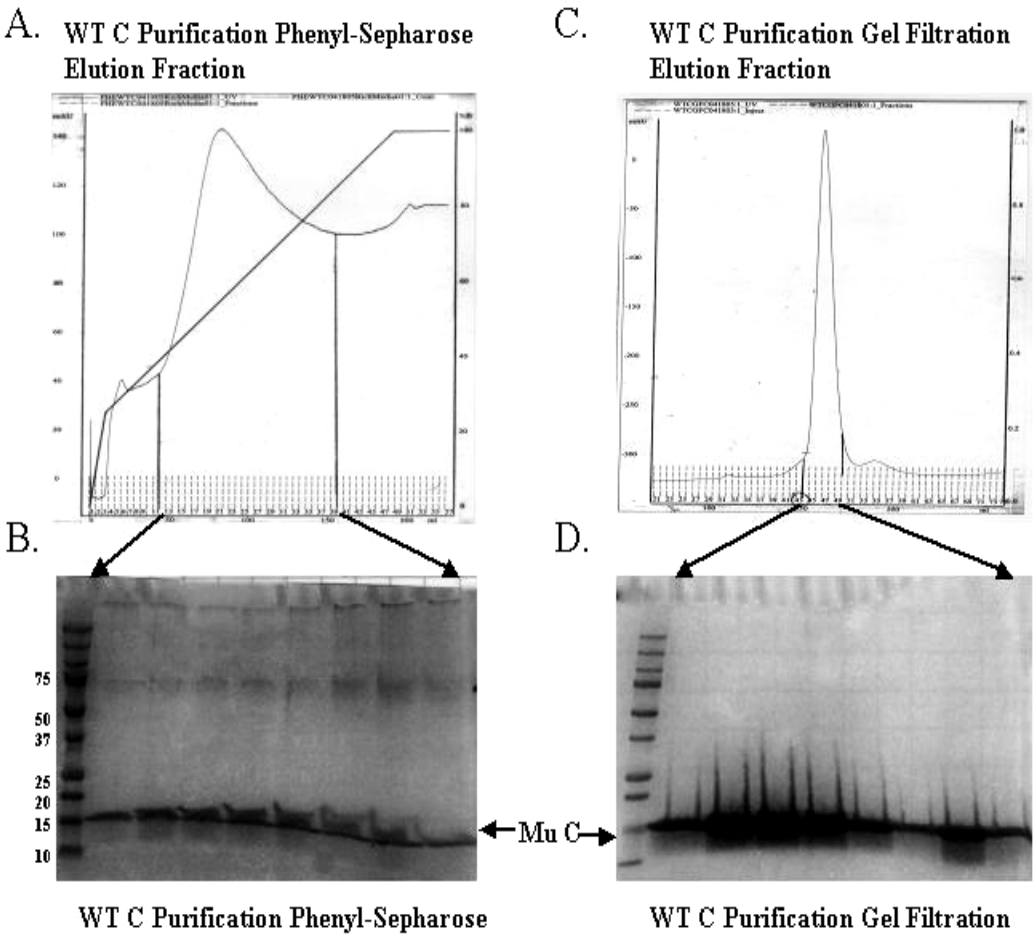


Figure 3-4. WT C purification strategy II. (A) chromatographic elution profile of Phenyl-sepharose run, (B) SDS-PAGE (8-20%) of the elution fraction of Phenyl-sepharose stained with Bio Safe™ coomassie, (C) chromatographic elution profile of the GFC and (D) SDS-PAGE (8-20%) of the elution fraction of GFC stained with Bio Safe™ coomassie.

Characterization of the purified WT C protein by SDS PAGE, gel-shift assay and ESI-TOF mass spectrometry

The SDS-PAGE was used throughout the purification process to monitor the purity, molecular weight and the integrity of C (aggregation and degradation). After each FPLC column, representative sample from the pooled fractions were collected and examined on a SDS-PAGE (Figure 3-5).

Molecular weight of the purified protein was obtained by Electrospray ionization (ESI)- time of flight (TOF) mass spectrometry (MS). The observed mass of 16515.14 da is the monomeric molecular weight of C (Figure 3-6).

To test if the purification scheme affected the biological activity (binding) of C, electro-mobility shift assays (EMSA) was carried out. EMSA was done using a 40-mer P_{sym} fragment (Figure 3-7). Samples containing C fractions were taken from each stage of purification and assayed for binding at 100, 400 and 800ng total protein concentration in a 20µl. binding reaction. The binding was carried out at RT for 20 min before electrophoresis. The C protein present in different fractions was able to shift almost the same amount of probes at a given concentration suggesting that the biological activity of C protein was not compromised during purification.

Protein purification: selenomethionine C protein

Induction and over-expression of selenomethionine C was done in *E. coli* JM109DE3 containing pT7-P_{lacSYN} C expression plasmid. Selenomethionine incorporation into the expressed C protein and its purification was done as per protocol described in the materials and method section. Since the metabolic inhibition method was used to incorporate of selenomethionine into the C protein, the amount of protein

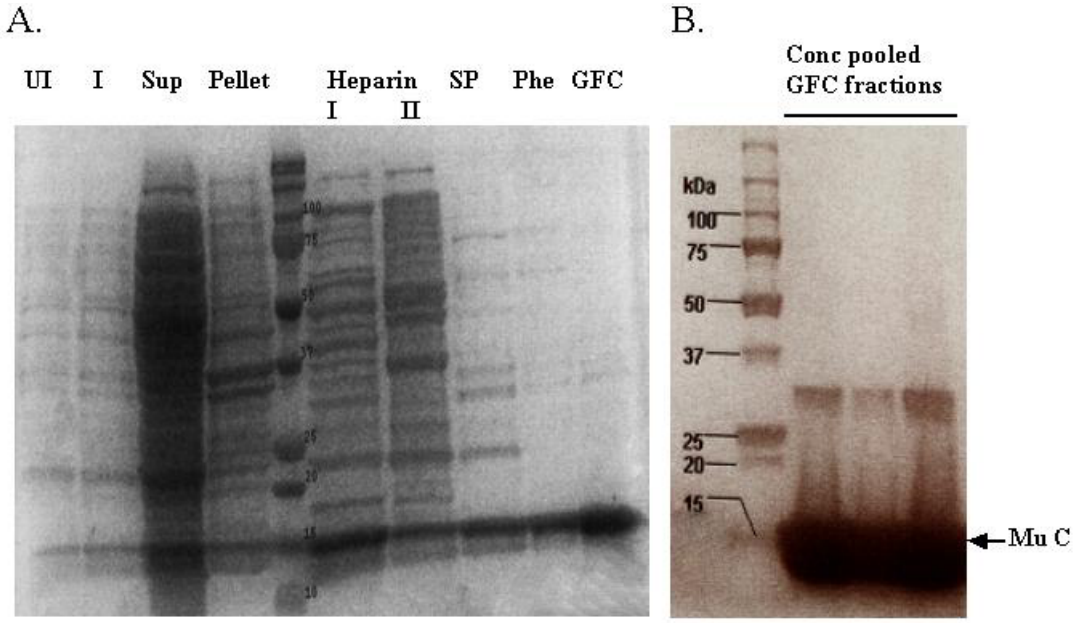


Figure 3-5. WT C purification. (A) SDS-PAGE (8-20%) stained with Bio Safe™ Coomassie Blue R250. Lanes are uninduced (UI), Induced (I), Supernatant (Sup), Pellet (P), Heparin I and II, SP-sepharose (SP), Phenyl-sepharose (Phe) and Gel-filtration column (GFC) Chromatographic elution profile of Phenyl-sepharose run, and (B) SDS-PAGE (8-20%) of concentrated GFC fraction stained with Bio Safe™ coomassie

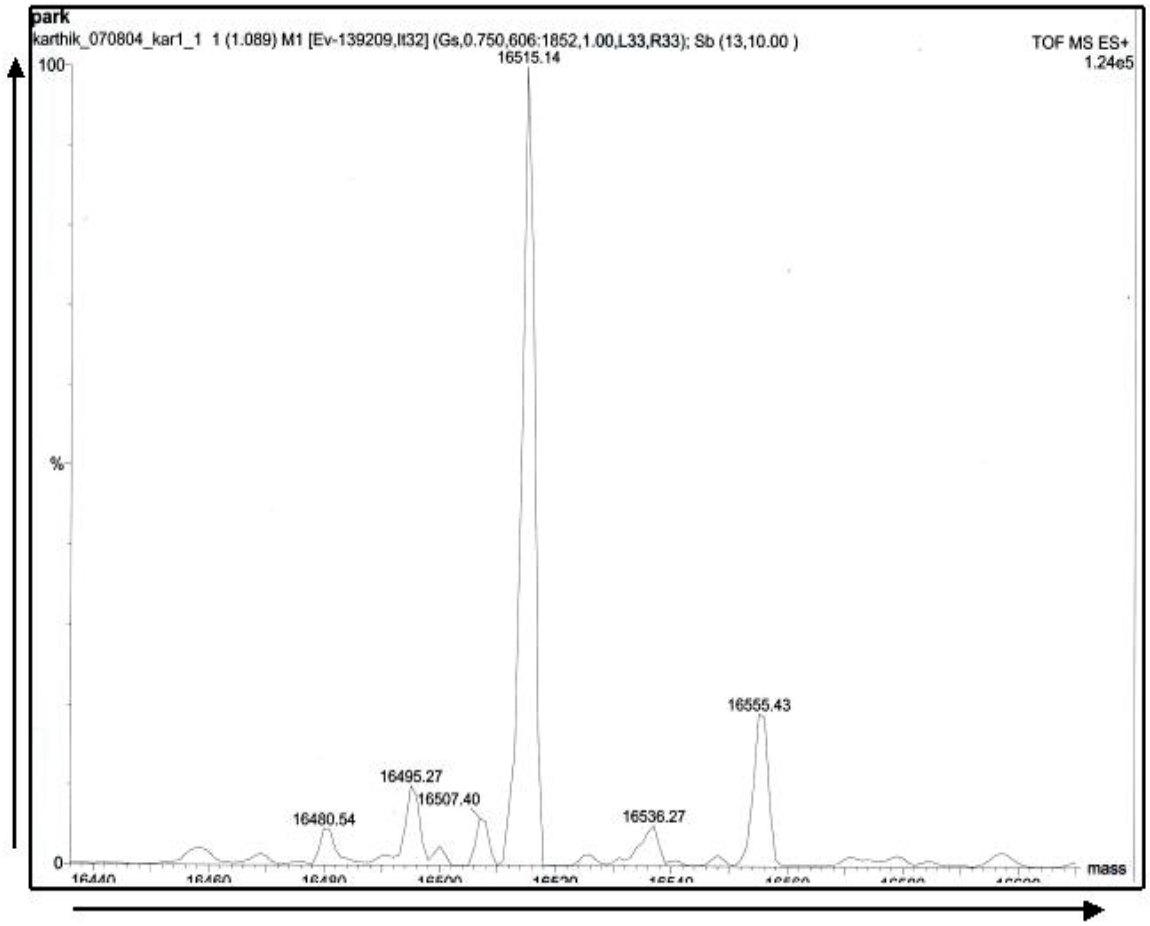


Figure 3-6. WT C ESI-TOF. The ESI mass spec profile of purified C protein. The X-axis denotes the molecular weight (da) range and the Y –axis denotes the percent of the protein at a given molecular weight. The observed peak mass of 16515.14 da refers to one monomer of C.

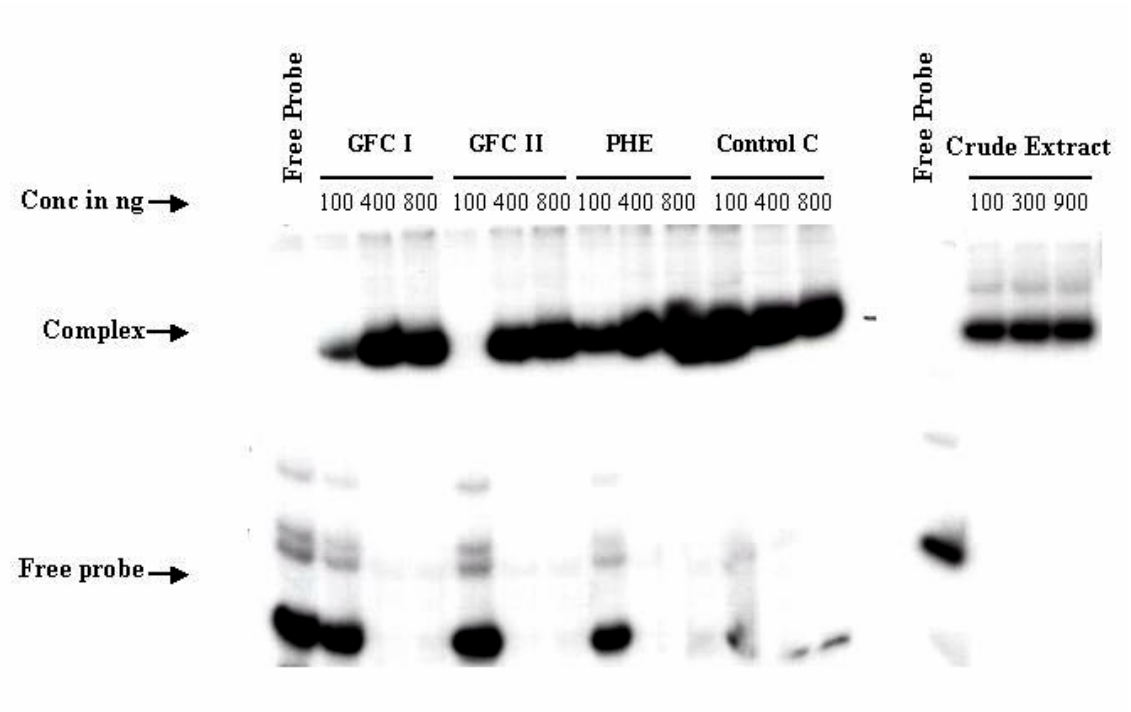


Figure 3-7. Electro-mobilty shift assay of C protein presenting different chromatographic fractions. P_{sym} 40-mer probes were incubated with different C fraction [100, 400 and 800ng total protein] in 20µl C buffer pH 7.0 at RT for 20min and electrophrosed on 10% a native acrylamide gel.

expressed was less than two fold when compared to wild-type protein expression. The final purified SeMet C protein was approximately 90% pure as determined visually from SDS-PAGE. Molecular weight of the SeMet C was determined by Electrospray ionization (ESI)- time of flight (TOF) mass spectrometry (MS) (Figure 3-8). The observed mass of 16608.37 da is the molecular weight of a C monomer with two selenium atoms incorporated without any further modification (Figure 3-8).

Electrophoretic mobility shift assay

Crystallization was to be attempted with GFC purified C : DNA complexes made from DNA of varying lengths and sequences. Thus, it was necessary to identify suitable binding conditions using gel-shifts in which the complexes were stable enough to withstand the harsh GFC environment. Parameters including, pH, salt, magnesium and varying DNA length and sequence were tried

Gel-shifts done with varying pH showed that pH 7 and 8.0 gave maximum C binding whereas pH 5.0 and 6.0 had minimal effect in C binding (Figure 3-9). Since there was little difference between pH 7.0 and 8.0, pH 7.0 buffer, which is near the physiological pH was chosen for complex formation.

Gel-shifts revealed that the C : DNA complex became unstable if the salt (NaCl) concentration increased from 150mm to 250mM. Based on this assay, 75mM NaCl concentration was chosen, since GFC requires a slightly higher salt concentration to prevent nonspecific binding of protein or complex to the resin (Figure 3-10).

In the gel-shift assays to test the effect of increasing magnesium concentration (Figure 3-11), it was found that there was no noticeable difference in the stability of the

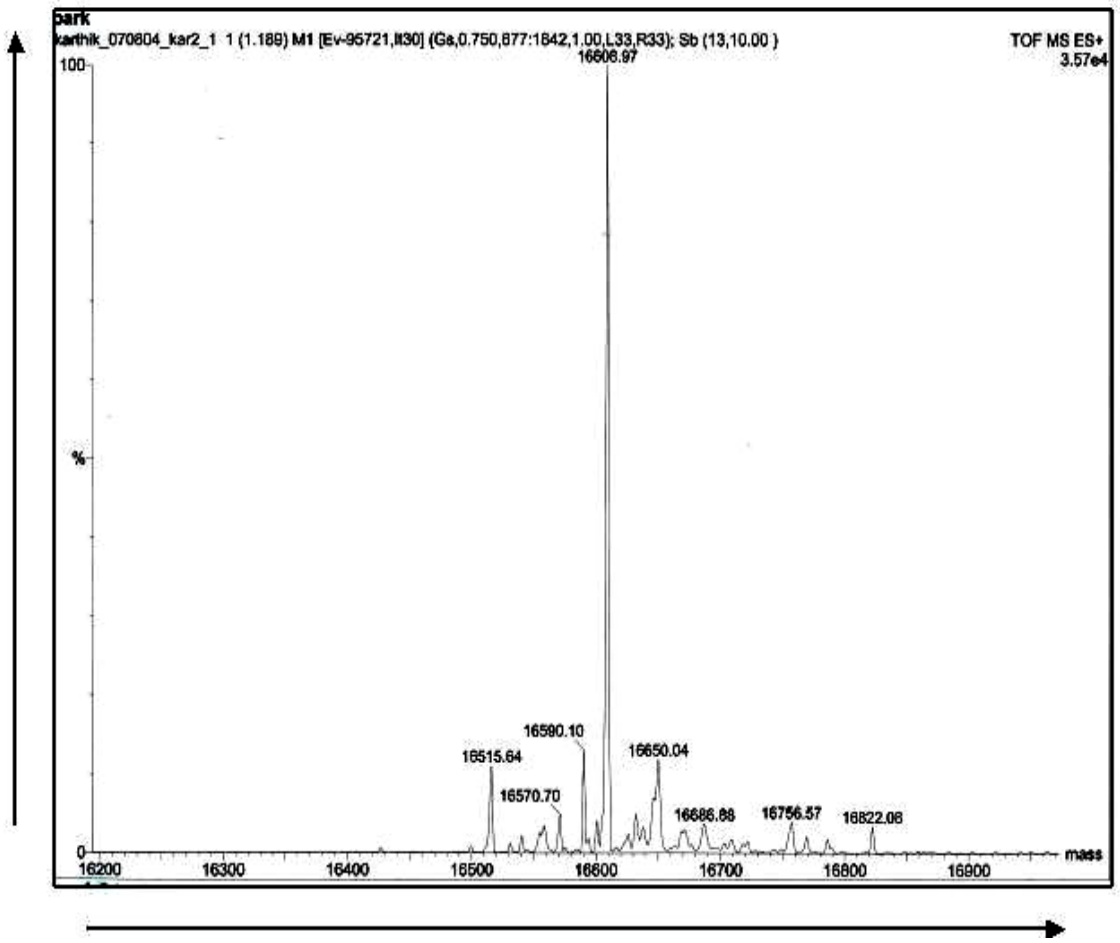


Figure 3-8. Selenomethionine C ESI-TOF. The ESI mass spec profile of purified SeMet C protein. The X -axis denotes the molecular weight (da) range and the Y–axis denotes the percent of the protein at a given molecular weight. The peak observed mass of 16608.97 da refers to difference of two selenomethionine incorporation to one monomer of C.

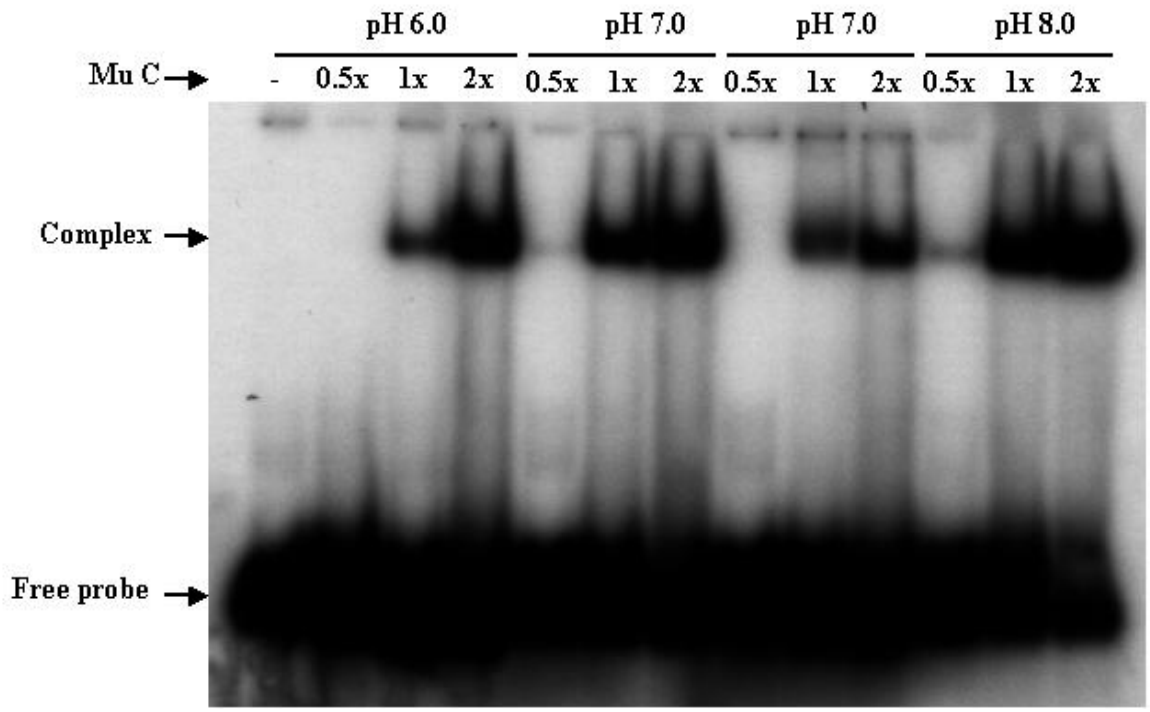


Figure 3-9. Gel-shift assay with purified WT C in C buffer with different pH. P_{sym}30-mer probes were prepared by annealing end labeled Kar77 with Kar 78. Purified annealed probes were incubated with WT C protein [0ng (-), 20ng (1x), 40ng (2x)] in 20µl C buffer with different pH at RT for 20min. Electrophoresis was done on 8% native acrylamide gel. The pH of the buffers used is listed on top.

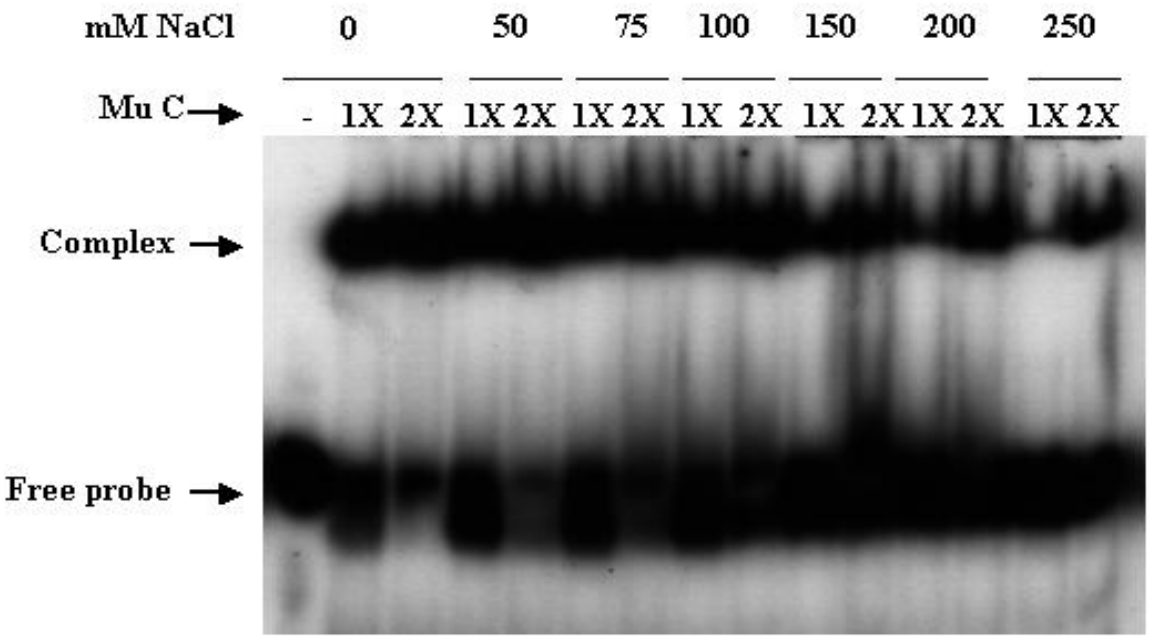


Figure 3-10. Gel-shift assay with purified WT C in C buffer with different NaCl concentration. P_{sym}30-mer probes were prepared by annealing end labeled Kar 77 with Kar 78. Purified annealed probes were incubated with WT C protein [0ng (-), 20ng (1x), 40ng (2x)] in 20µl C buffer pH 7.0 at RT for 20min. Electrophoresis was done on 8% native acrylamide gel. The NaCl concentration in the buffer used is listed on top of each panel.

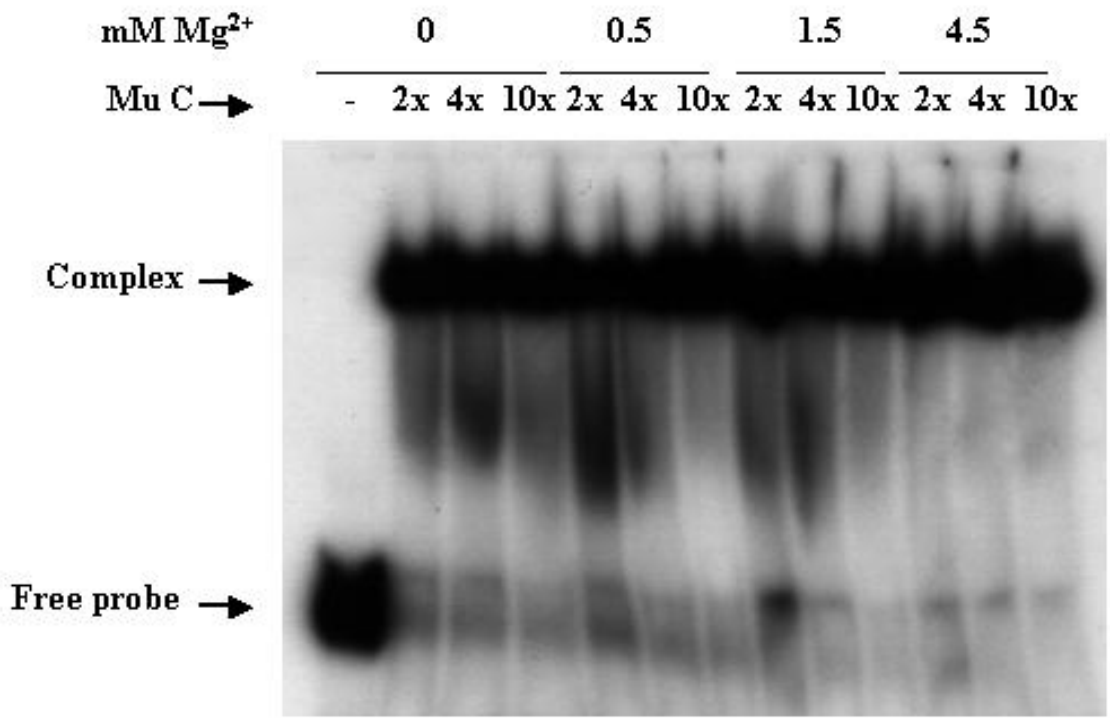


Figure 3-11. Gel-shift assay with purified WT C in C buffer with different Mg^{2+} concentration. P_{sym}30-merprobes were prepared by annealing end labeled Kar 77 with Kar 78. Purified annealed probes were incubated with WT C protein [0ng (-), 20ng (1x), 40ng (2x)] in 20µl C buffer pH 7.0 at RT for 20min. Electrophoresis was done on 8% native acrylamide gel. The Mg^{2+} concentration in the buffer used is listed on top of each panel.

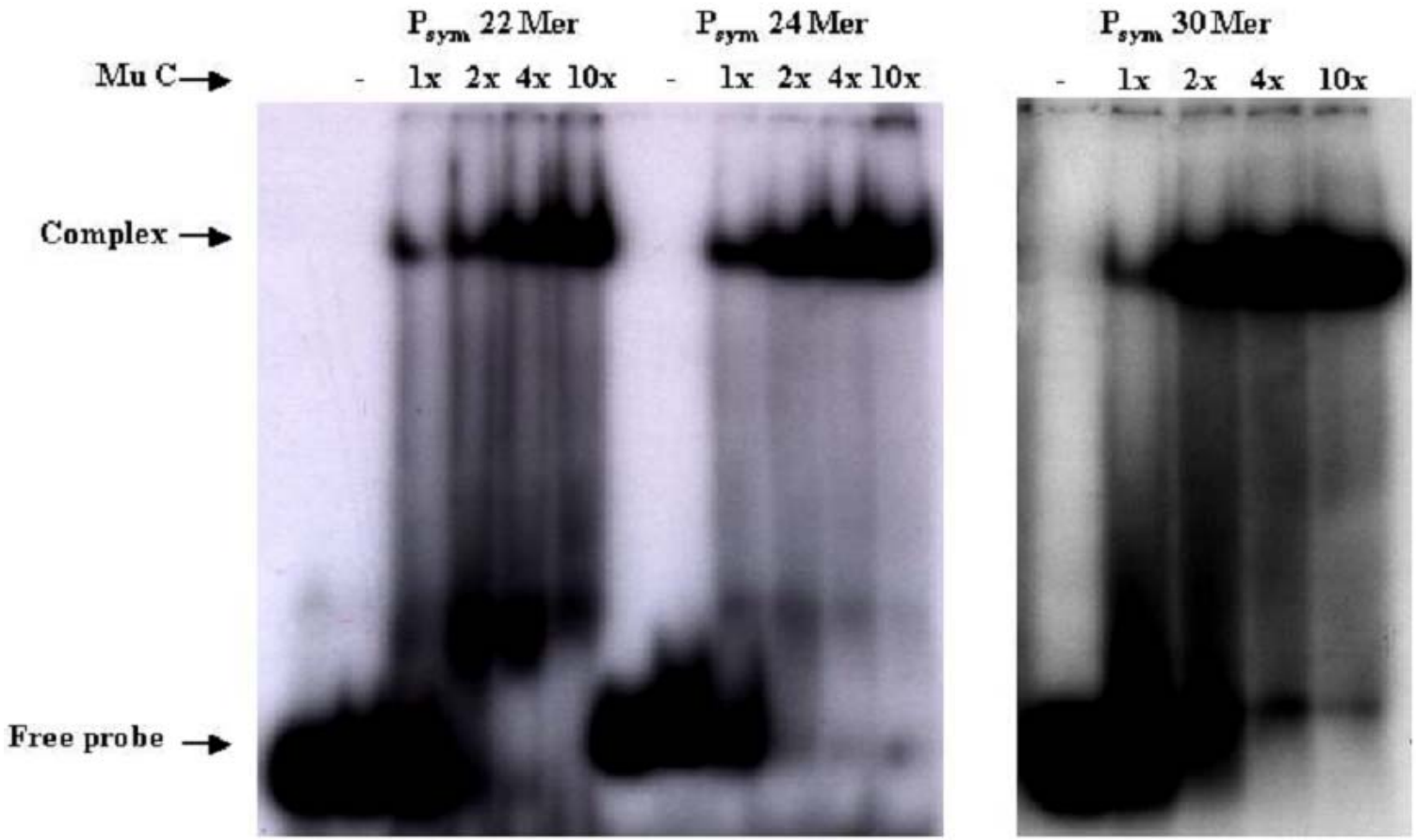
complex with regard to magnesium concentration. So a minimal concentration of 1mm $MgCl_2$ was used in complex formation concentration since it has been suggested previously that Mg^{2+} may be important for C protein conformational changes (De *et al.*, 1998).

To increase the probability of complex crystallization, P_{sym} and P_{lys} of varying length were to be tried. Since, multiple combinations in length were possible, a subpopulation of DNA was tested in gel-shift assays (Figure 3-12). The assay showed that C was able to bind a P_{sym} 24-mer effectively but in P_{lys} the length had to be around 30 bp. Binding and stability of the complex was significantly reduced when the length was reduced from 22-mer to 18-mer. Therefore for crystallization trials it was decided that a minimum of 18 bp and a maximum of 24 bp of DNA would be tested for both P_{sym} and P_{lys}.

Complex formation

The GFC purification of Mu C: DNA complexes suggested that the stability of the complex was not compromised and could be used in crystallization trials. Additionally, purification showed that, the binding condition identified through gel-shift assay was optimum and the purified protein was functional. For complex formation, WT C or selenomethionine C (SeMet C) was used at a concentration of 20-30mg in 10-12ml binding volume, and the binding was carried out by slow addition of the annealed oligos to the protein. This was done to avoid high local concentration of the DNA and to prevent precipitation. The binding reaction was carried out for 20 minutes at RT before gel-filtration. All binary complexes used for crystallization were purified using a

Figure 3-12. Gel-shift assay with purified WT C in C buffer with P_{sym} and $P_{\textit{lys}}$ probes of varying length. P_{sym}30-mer probes were prepared by annealing end labeled Kar 77 with Kar 78. Purified annealed probes were incubated with WT C protein [0ng (-), 20ng (1x), 40ng (2x)] in 20µl C buffer pH 7.0 at RT for 20min. Electrophoresis was done on 8% native acrylamide gel. The Mg^{2+} concentration in the buffer used is listed on top of each panel.


P$_{sym}$ 22 Mer
P$_{sym}$ 24 Mer
P$_{sym}$ 30 Mer
Mu C
- 1x 2x 4x 10x - 1x 2x 4x 10x
- 1x 2x 4x 10x
Complex
Free probe

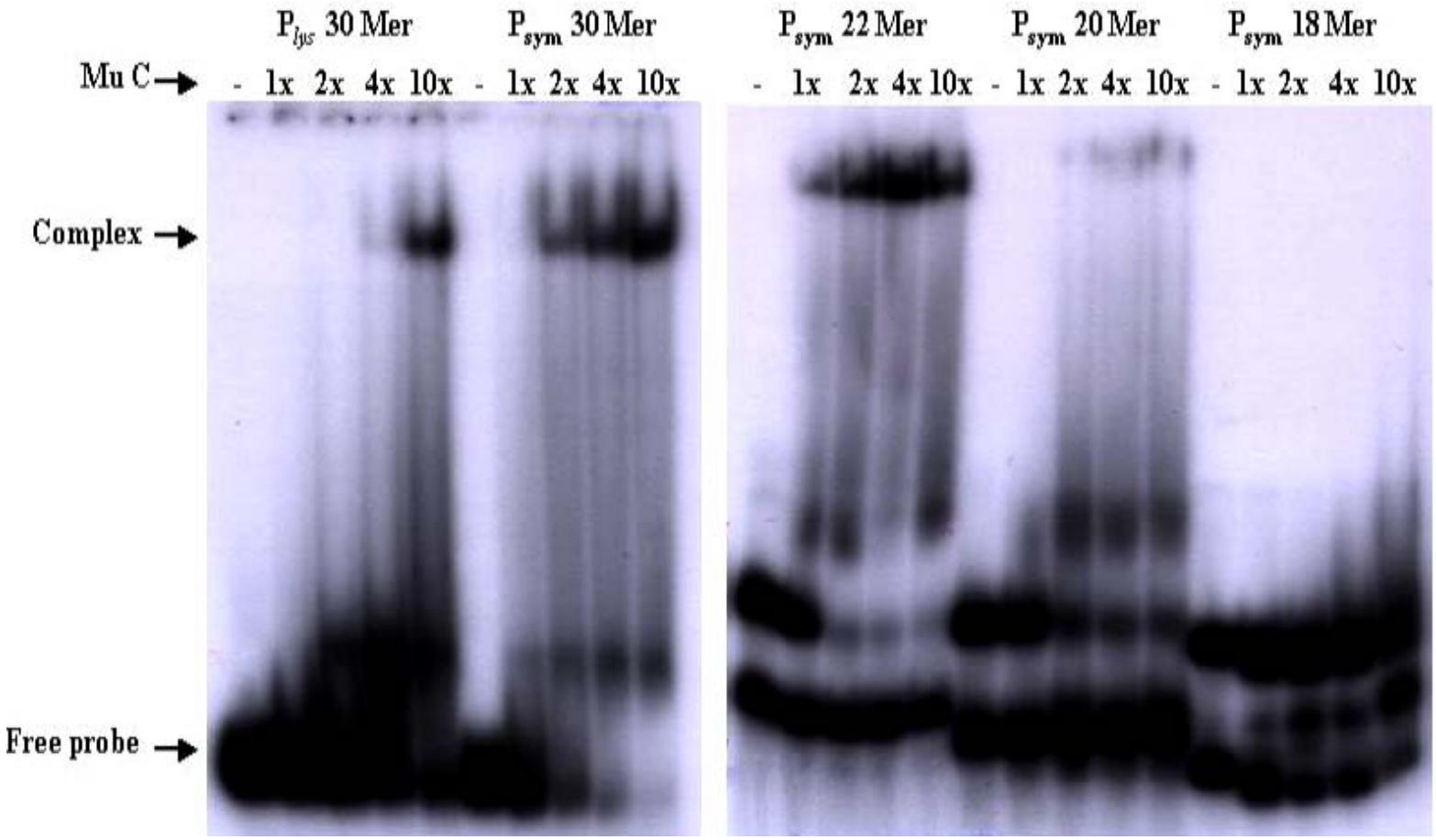


Figure 3-12 (continued).

gel-filtration column because, (1) to exchange the protein buffer with the complex buffer (25mM Hepes, pH 7.0, 75mM NaCl, 5% glycerol, 1mM $MgCl_2$, 1mM EDTA, 10mM DTT), and (2) to separate the binary complex from free DNA and /or free protein. The amount of free DNA and/ or protein varied every time the complex was purified, this variation was largely seen when different protein preparations were used, and the amount of precipitation formed during complex formation.

The C : DNA binary complex eluted out at the expected size of 46-49 kDa (Figure 3-13) Elution fraction were tested by SDS PAGE electrophoresis and Bradford assay for the presence of C protein. By comparing the shift in the GFC elution profile (Figure 3-14) as well as the shift in the UV absorbance value for the DNA, the DNA in the complex was identified as the binding DNA. The purified binary complex could be concentrated up to 30-35 mg/ml without much precipitation.

Crystallization of Mu C : DNA complex

Crystallization trial of the binary complex was initially done with WT C : DNA complex since SeMet C : DNA complex production was laborious and the purified complex yield extremely low. Binary complex crystallization was done with a wide assortment of P_{sym} and $P_{\textit{lys}}$ DNA, which varied in length and sequence. (Figure 3-15) Initial screenings were done with a 24-mer P_{sym} C complex at 20-30 mg/ml, using sitting drop vapor diffusion method. Depending on the availability of the complex six to seven different commercial screens were used (1µl:1µl complex to reservoir with a total of 200-500 drops /complex) in different incubation temperature (4° C,18° C, 28° C and 37° C) The screens used were Hampton Index & Index HT™ screen, Hampton Natrix screen,

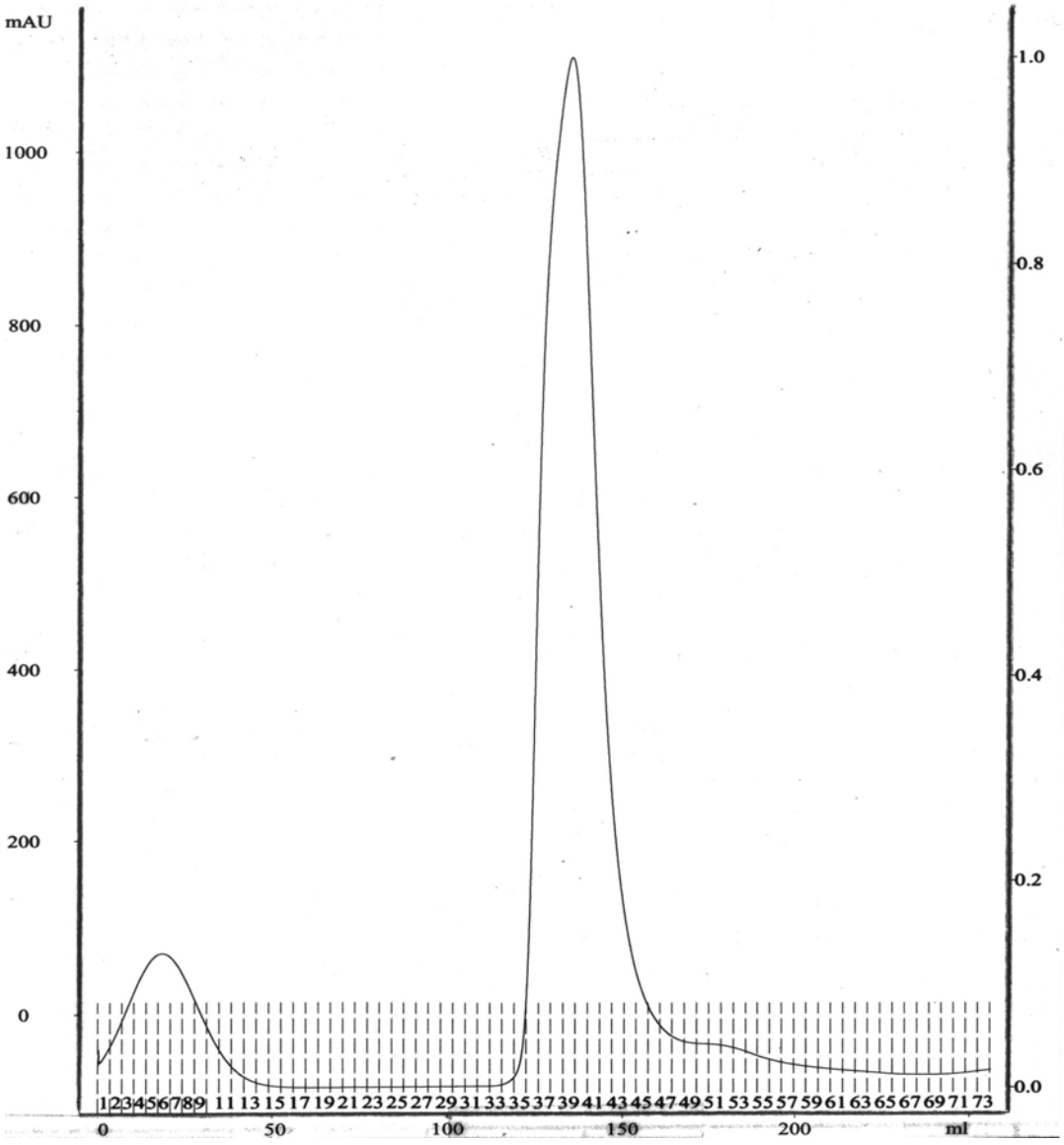


Figure 3-13. Gel-filtration chromatography of C: DNA complex. The final GFC elution profile after complex formation. The UV absorbance at 260 shown in the Y-axis confirmed the presence of DNA.

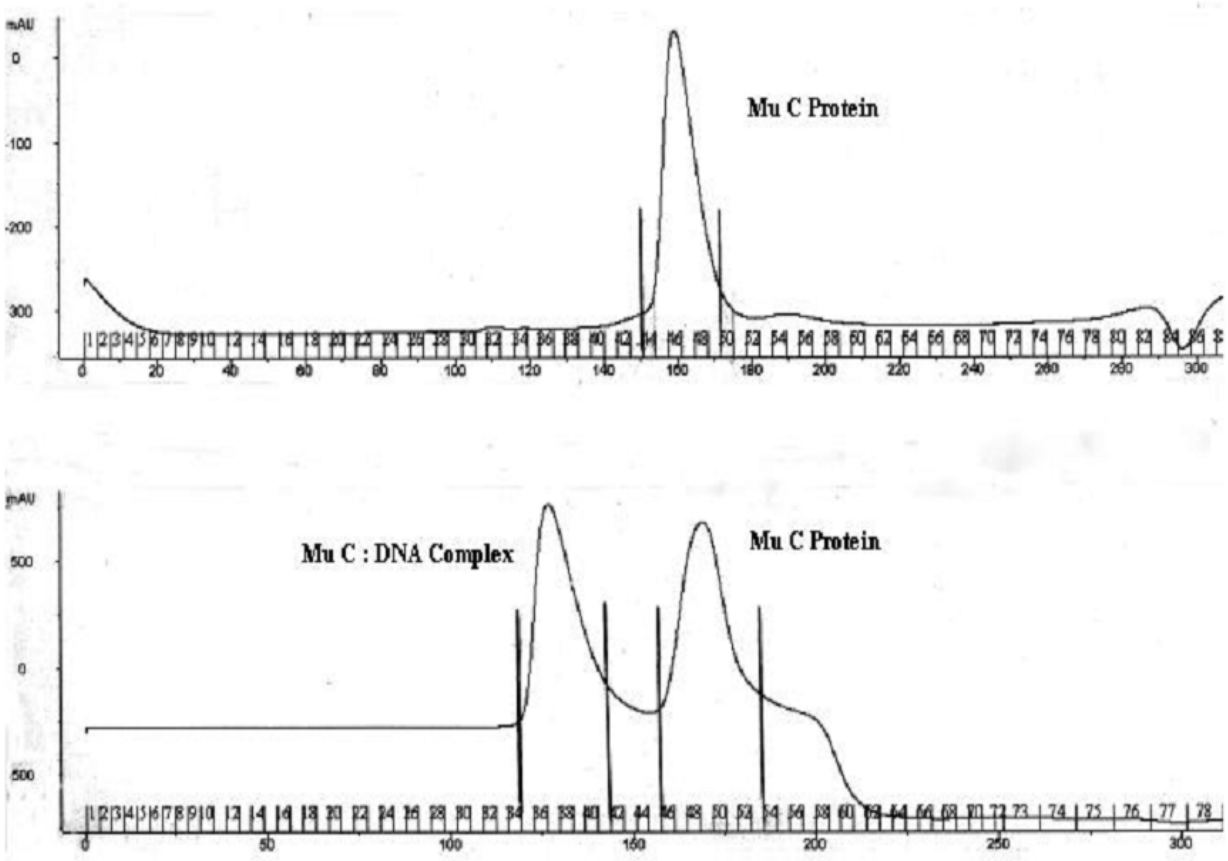


Figure 3-14. Comparative gel-filtration elution profile of WT C and C : DNA complex. The top panel show the position were dimeric C protein elutes during GFC . The bottom panel shows the gradual shift to a higher molecular weight species indicating complex formation in the presence of DNA.

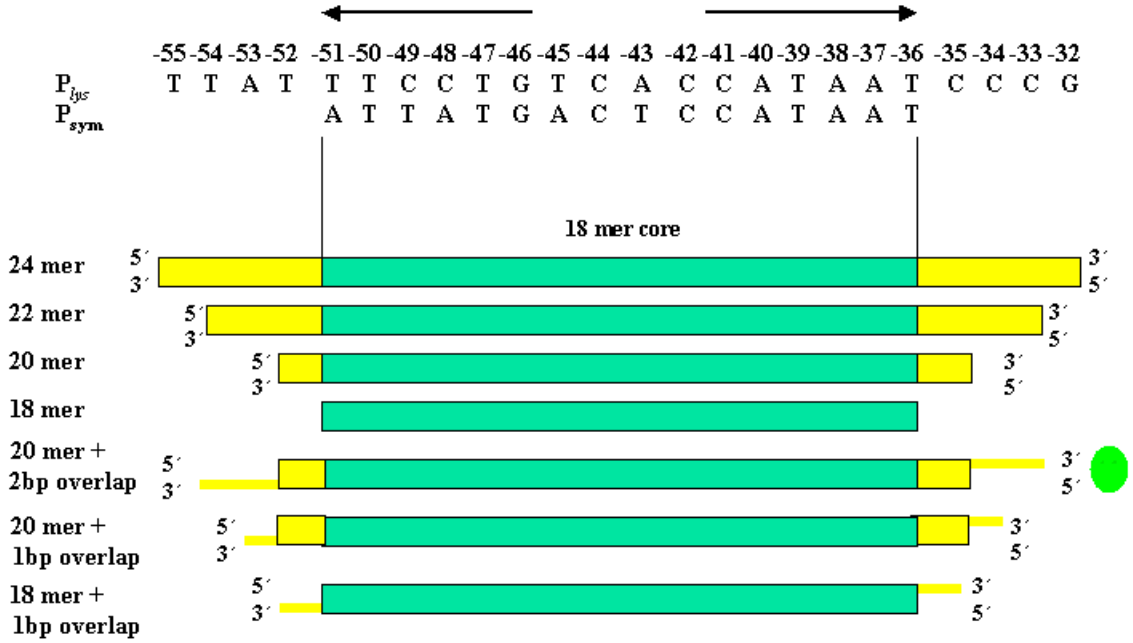


Figure 3-15. Complete list of P_{sym} and P_{lys} DNA used for crystallization. The minimum DNA length used for crystallization was a 18-mer (green shade). This formed the core for the entire longer DNA used. The sequence for the 18-mer and the rest of the flanking sequence for P_{sym} and P_{lys} is shown on the top panel. The green dot next to the 20mer + 2bp overlap was the sequence used with C protein which gave diffracting crystal.

Hampton Lite screen, Hampton Crystal Screen I and II, Hampton Peg/Ion screen and Emerald wizard screen I and II. Micro crystals, which were birefringent were only seen in crystallization trials done at 37° C and 18° C. Crystallization trials done at lower temperature had only clear drop or precipitation. The microcrystals grew from many different conditions and the most promising condition was from Emerald wizard screen. These crystals crystallized rapidly having a hollow tubular shaped morphology surrounded by precipitates and protein skin. A similar condition also from Emerald wizard screen was identified for a complex made from P_{sym}20-mer plus 2 base overlap. This condition gave a drop full of sperulites within 24 hrs. Since complex production was a laborious task with minimal yield, it was decided to refine first the P_{sym}24-mer complex since the crystals obtained were physically better looking micro-crystals. Refinement was started for this complex because the crystals were too small for X-ray crystallography analysis. The Figure 3-16. shows the various stage of refinement for this complex. At the end of refinement, these crystals were tested for diffraction using an in house X-ray source. A wide array of mounting techniques were used to test the crystals, including,(1) varying the cryoprotectant, (2) testing crystal of different dimensions, and (3) performing X-ray analysis at room temperature. These techniques revealed that the quality of the crystals was poor and cannot be improved any further.

Crystallization of the complex made from P_{sym}20-mer plus 2 base overlap was then undertaken since earlier crystallization trials with this complex was giving sperulites. After several round of refinements the condition was optimized from 0.1 M Na Citrate pH 6.5 and 20% w/v PEG 3000, 18° C to 0.1 M Na Citrate pH 5.7, 14% w/v PEG 3000 with additive Benzamidine HCl 8% w/v. The Figure 3-17 shows the

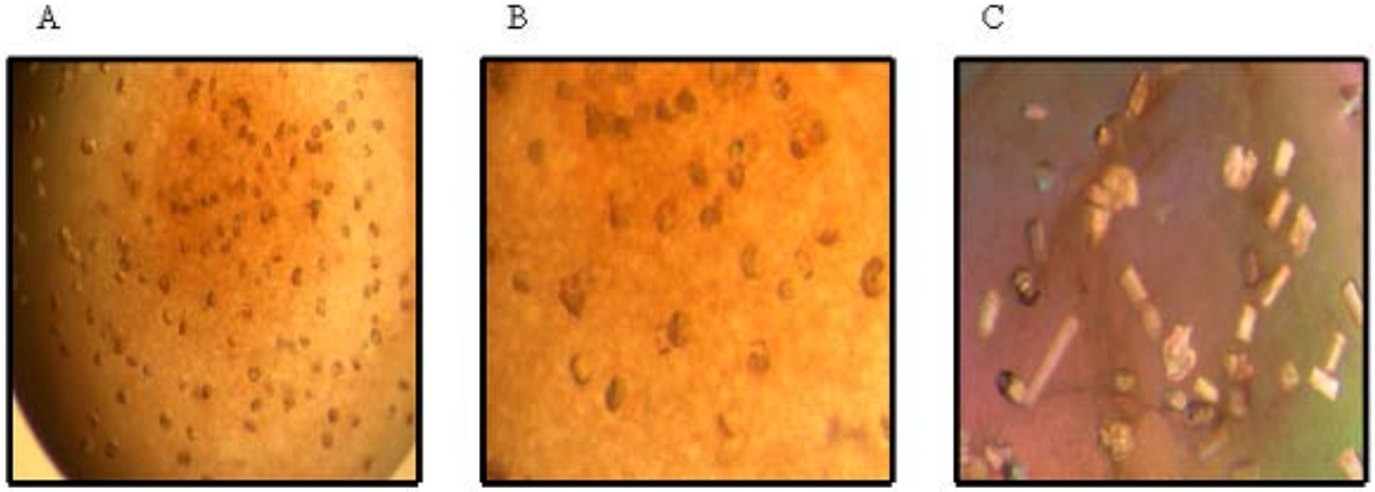


Figure 3-16. WT C P_{sym} 24-mer co-crystal optimization. (A) preliminary condition from Wizard screen (B) same crystals after a couple of round of refinements (C) the final refined condition showing long tubular shaped crystals which are birefringent.

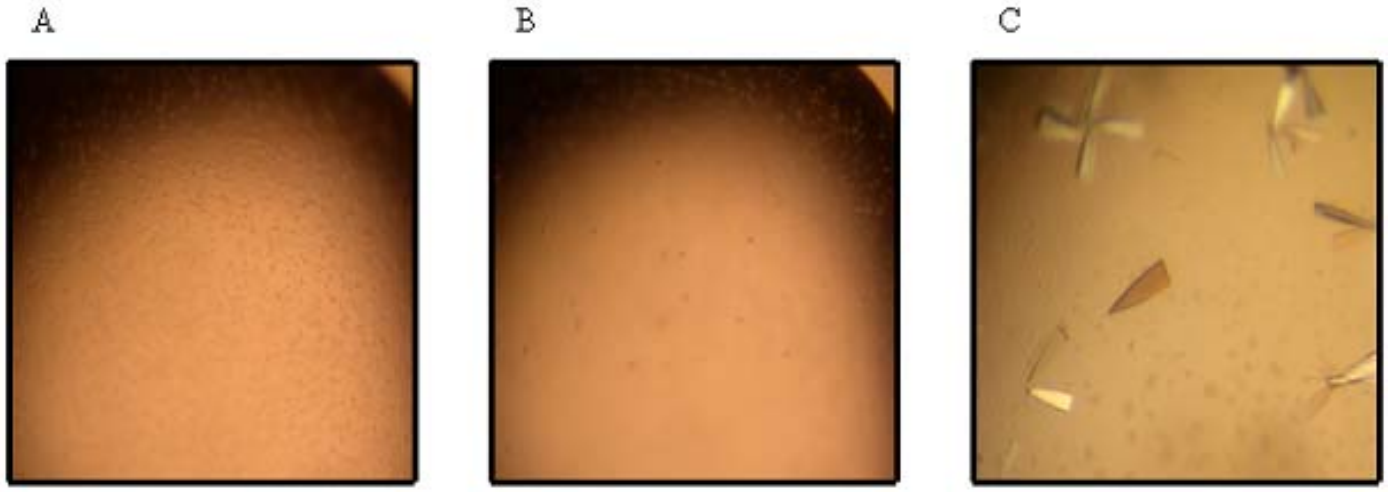


Figure 3-17. WT C P_{sym}20-mer plus 2 base overlap co-crystal optimization. (A) preliminary condition (B) same crystals after a couple of rounds of refinements (C) the final refined condition showing birefringent triangular plate like crystal.

improvement of the crystals during various stages of refinement. The final crystal obtained was a fragile highly birefringent triangular plate like crystal.

The final refined crystals were tested for diffraction using an in house X-ray source and multiple datasets were collected from advance photon source (APS, Argonne National Laboratory). The resolution of the data collected was ~2.8 Å but due to the fragility of the crystals and twinning the data collected was very mosaic and could not be processed further. Since no more refinement could be done with the present condition it was decided that new condition/s had to be identified to improve the quality of the the crystals and thus the quality of the data. To identify new conditions crystallization trials was done in Structural Genomics Consortium (SGC, Toronto) using hanging and sitting drop method in 24 well as well as 96 well plates. Screening was done only with two customized in house screens called SGC and Redwing. Multiple conditions were obtained in which this complex crystallized and Figure 3-18 shows some of those conditions, which gave good quality crystals.

The crystals which grew in 1.4 M ammonium sulfate 16% ethylene glycol pH 5.7 gave better diffracting crystals with low mosacity when it was analyzed using the in house X-ray source. Multiple dataset were collected and processed but due to the lack of a structural homologue, the complex structure could not be solved by molecular replacement. So to get the phase angle information required to solve this complex structure a selenomethionine C: P_{sym}20-mer plus 2 base complex was prepared and crystallized in the same condition (Figure 3-19)

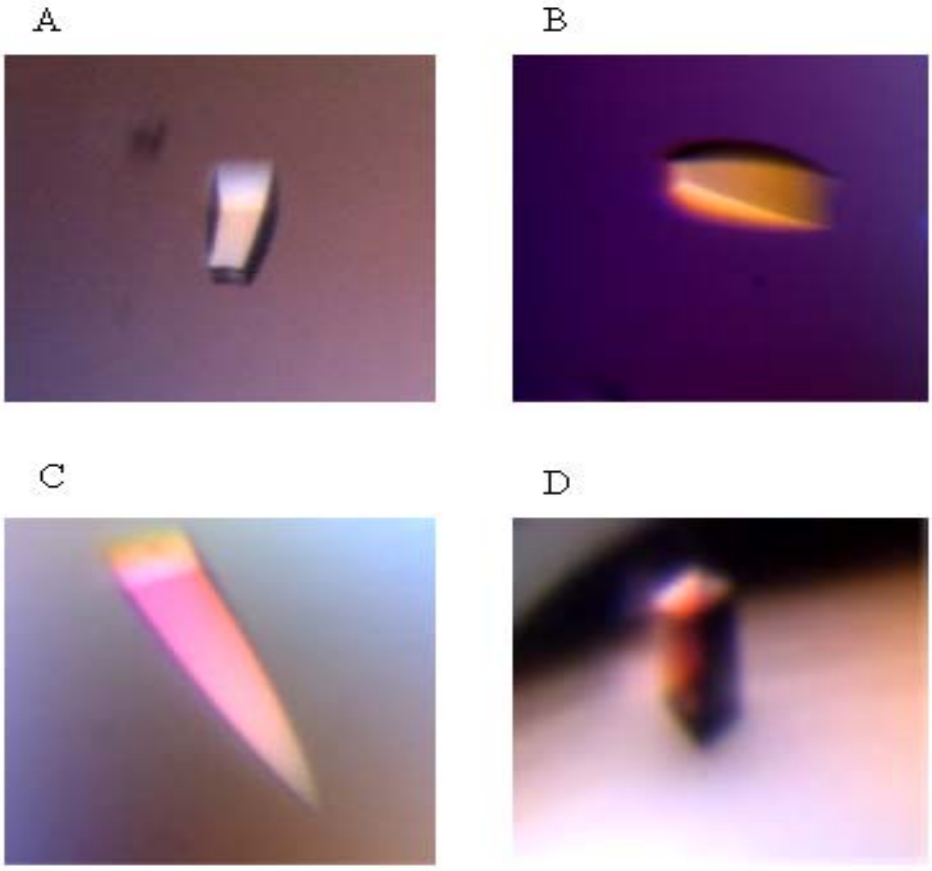


Figure 3-18. Different crystallization condition identified for WT C P_{sym}20-mer plus 2 base overlap. (A) 1.5 M ammonium sulfate 15% glycerol pH 5.7, (B) 1 M ammonium sulfate 35% glycerol pH 5.7, 10 mm DTT, (C) 1.4 M ammonium sulfate 15% ethylene glycol pH 5.7, and (D) 1.4 M ammonium sulfate 16% ethylene glycol pH 5.7.

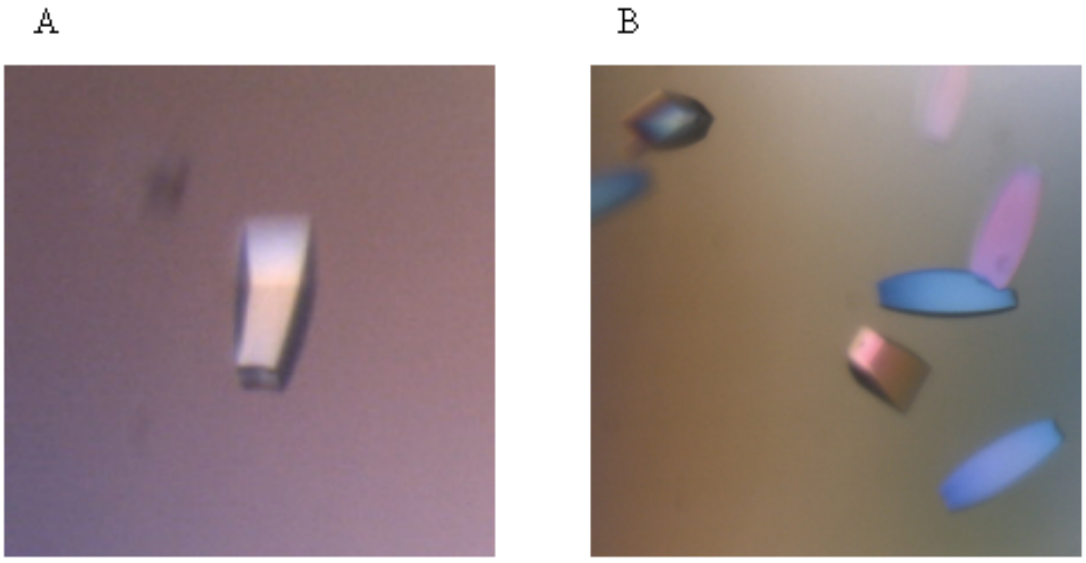


Figure 3-19. Co-crystallization of native C and selenomethionine C with P_{sym}20-mer plus 2 base overlap in 1.4 M ammonium sulfate 16% ethylene glycol pH 5.7. (A) native C : DNA co-crystals (B) Selenomethionine C : DNA co-crystals.

Screening, data collection and processing

A single crystal that diffracted to approximately 4Å at the home source was taken to 17ID beam-line of the Advanced Photon Source (Argonne National Laboratory, Chicago, U.S.A) A single wavelength SAD experiment was conducted using a CCD image plate detector; a fluorescence scan near the selenium edge was carried out to obtain the correct wavelength for data collection. A complete data set was collected at wavelength 0.9791. The distance between the crystal and detector was 300 mm and a total of 360 oscillation images were recorded with exposure times of 10 seconds (Figure 3-20).

The diffraction data was indexed, processed, and scaled with DENZO and SCALEPACK programs in the HKL package (Otwinowski, 1997). The SeMetC : DNA complex was crystallized in $P4_3$ spacegroup with unit cell parameter a= 68.9Å and c=187.6 and there were two copies of the complex present in asymmetric unit (Table 3-4). Structure determination was based on the incorporation of the SeMet in the C protein and obtaining the phase angle using the single wavelength dataset by applying the SAD technique. The positions of the two SeMet residues were determined by the SOLVE program (Terwilliger and Berendzen, 1999). An electron density map at a resolution of 3.1Å was calculated from the phases obtained and at this resolution a straight B form DNA was clearly visualized. To fit the DNA in the electron density map as a rigid body a blunt end DNA model the DNA used for crystallization (P_{sym} 20 with 2 base overlap) was made using the model it server (Vlahovicek *et al.*, 2003) http://www.icgeb.trieste.it/dna). This model was then used to fit the DNA in the electron density using the program "O"(Jones *et al.*, 1991). Rigid body modification was then done using REFMAC5

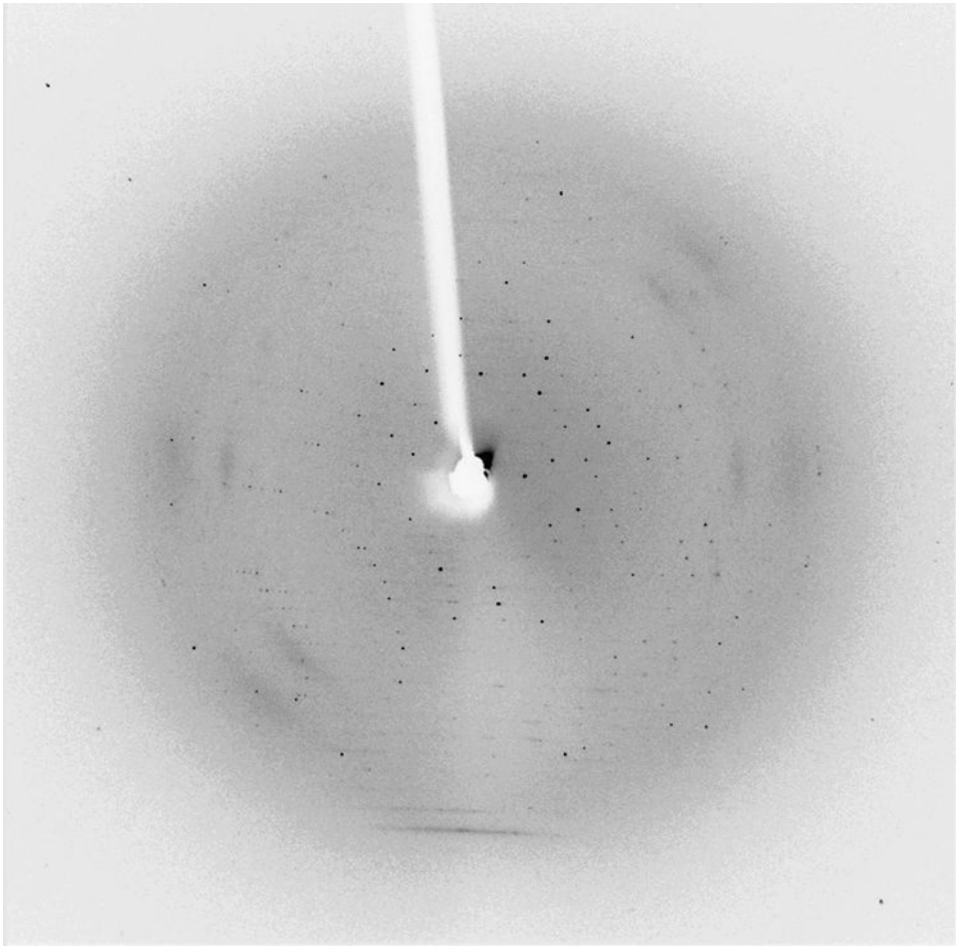


Figure 3-20. Diffraction image of SeMet C : DNA complex at 3.1 Å collected at 17ID of the Advanced Photon Source.

Table 3-4. Diffraction data statistics of SeMet C : DNA complex crystals.

Data set	SeMet C : DNA Complex
X-ray source	APS (Beamline 17ID)
Wavelength	0.9791 Å
Detector	CCD
Crystal Parameters	
Space Group	$P4_3$
Unit cell Parameters	
a	68.9 Å
c	187.6 Å
Data statistics	
Resolution	3.1 Å
No. of unique reflections	15,849
Completeness (%)	99.7 (100.0)
Rsym[a]	0.073 (0.299)
Average I/σ (I)	14.7 (6.4)

(CCP4, 1994) which resulted in the reduction of the R value from ~65% to ~47%. Further refinement were hindered at this point since the model DNA was blunt ended and also only 21 bp out of 22 bp of DNA used was fitting in the unit cell. So to fit the modelled DNA in the electron density map it was split into two 10-mers. The split 10-mer DNA was superimposed in the electron density according to the pseudo dyad symmetry of the DNA and rigid body modification was done, but the R factor did not improve anymore. Since the resolution of the dataset was low and the phases obtained through SAD not accurate enough only the C-α polyalanine model for the protein could be built.

Discussion

Structural studies in C protein have been hampered due to its tendency to aggregate at high concentration and no one has been able to produce a functional protein at a high enough concentration. The present study describes for the first time in detail; (1) the over-expression and purification of milligram quantities of homogenous, functional C protein, (2) the crystallization of a C : DNA complex, and (3) preliminary structural information of how C protein interacts with the DNA.

The C protein expression vector (pZZ41) used in this study was based on the T7 based expression system(Studier *et al.*, 1990) and was constructed by Zhao, 1999. This vector has a synthetic promoter/operator called P_{lacSYN} downstream from the T7 promoter in front of the Mu *C* gene for efficient repression and to prevent leaky protein expression before induction. The efficient repression was found to be extremely important because leaky expression of C protein led to a high C concentration within the cell, which is toxic and prevented C over-expression when the gene was induced.

The expression host JM109 DE3 was the only T7 expression host found suitable for C over-expression. In other expression host like BL21DE3, C protein expressed in large quantities but all of them were insoluble. It is possible that a rapid rate of C expression will have resulted in misfolding causing the protein to accumulate in the insoluble fraction. It is possible that JM109 DE3 may have provided an optimum balance between expression and folding resulting in more than basal level of expression but in just enough quantities to be purified. To overcome this problem large culture volume was utilized for purification.

The over-expressed C protein was found in both the soluble and insoluble fractions. Since the primary aim was to obtain a fully functional and properly folded C protein only the soluble fraction was utilized. The protocol used in this study is unique because it minimizes the time required to efficiently purify large quantities of functional C protein. This protocol avoids the use of dialysis between subsequent steps, which usually take a longtime, which in turn may affect the activity of the protein. This procedure utilized the heparin affinity chromatography as the first step to enrich the protein since C is a DNA-binding protein. This step enriched every DNA binding proteins present in the soluble fraction so a second cation exchange chromatography was used to remove some of the contaminating proteins. This step took advantage of electrostatic surface potential of C, which has many charged residues in its N-and C-terminus. During the course of developing this purification scheme, it was found that C protein was very stable at high salt concentration so the hydrophobic interaction chromatography (HIC) was utilized. The HIC is important because it removed some major contaminating protein co-purifying with C. These contaminating proteins could not

be removed by other procedure like ROTOFOR and only HIC was able to remove most of them. Since HIC used a high salt extraction process and C protein was to be used for co-crystallization, size exclusion was utilized as the final step to do buffer exchange and to remove some high molecular contaminant.

The functional activity of the purified protein during each purification step was assayed by gel-shift using a labeled P_{sym} promoter fragments and specific binding of the protein was maintained throughout the purification scheme.

The primary aim of this study was to crystallize the C protein with its cognate DNA thus the choice of the DNA used for crystallization became very important. Previous protein DNA co-crystallizations have shown that, precise length and composition of the oligonucleotides used for crystallization is the most critical variable that must be determined for every new protein. Therefore, I initially looked at the well-characterized Mu late promoter P_{lys} for crystallization, but previous successful co-crystallization mostly involved the use of symmetrical DNA binding sequence for dimeric protein like C, so I decided to use P_{sym}. The P_{sym} promoter has a perfect inverted symmetry between its binding sites and has a high affinity for C protein. The second critical variable for co-crystallization is the length of the DNA. In the crystal, the DNA has a strong tendency to stack ends to end, so a precise length of the DNA fragment and the nature of the stacking interactions determines the crystalline order and subsequently the unit cell size. Since the length of the DNA fragment is determined by the size of the minimal binding site necessary for tight complex formation, I began complex formation with a 18-mer, which is the minimal double stranded sequence required for C binding and progressively increased the length till it reached a 24-mer. To increase the probability of

end-to-end stacking of oligonucleotides within the crystal, I tried changing the composition of the DNA ends by adding a single or double complimentary overhanging bases in the 5′ end.

A stable C : DNA complex was a pre-requisite for successful co-crystallization. This was achieved by purification the complex after it was made using size exclusion chromatography. This procedure is efficient in removing unbound protein or DNA that might hinder crystallization. A successful complex purification was also suggestive of a very stable complex formation with the expected stoichometry.

Crystallization of the complex was done in both hanging and sitting drop method involving many different conditions. The appropriate crystal for structure analysis was identified through a series of screening.

The preliminary complex structure of the C protein bound to the P_{sym} promoter is the first ever for C protein as well as for the proteins of the Mor/ C family of transcription factors. The preliminary Cα- polyalanine main chain model reveals that C is a dimer and has a HTH motif in the C terminus. However, due to the lack of resolution the critical protein-DNA interactions, side chain contacts and C dimerization interface could not be visualized. The most interesting and bizarre find in the complex structure pertains to how C is bound to the major groove. In the crystal structure, the two symmetrical sites in the DNA is occupied by two C dimers and not one (Figure 3-21). This is a complete first as to how a dimeric protein binds its symmetrical binding site. Since there is no precedent for this mode of binding coupled with the low resolution of the structure, interpretation of the present structure is difficult. However, few rational hypotheses can explain as to why C follows this mode of DNA interactions. (1) Mor protein is closest homologue of C

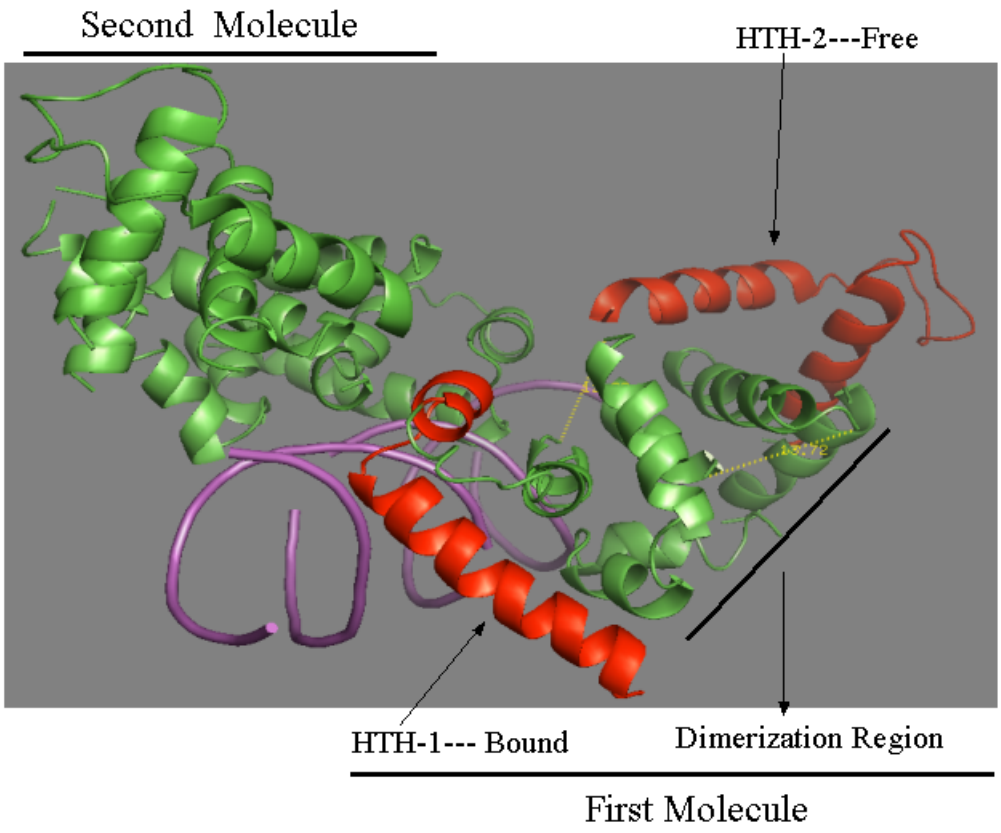
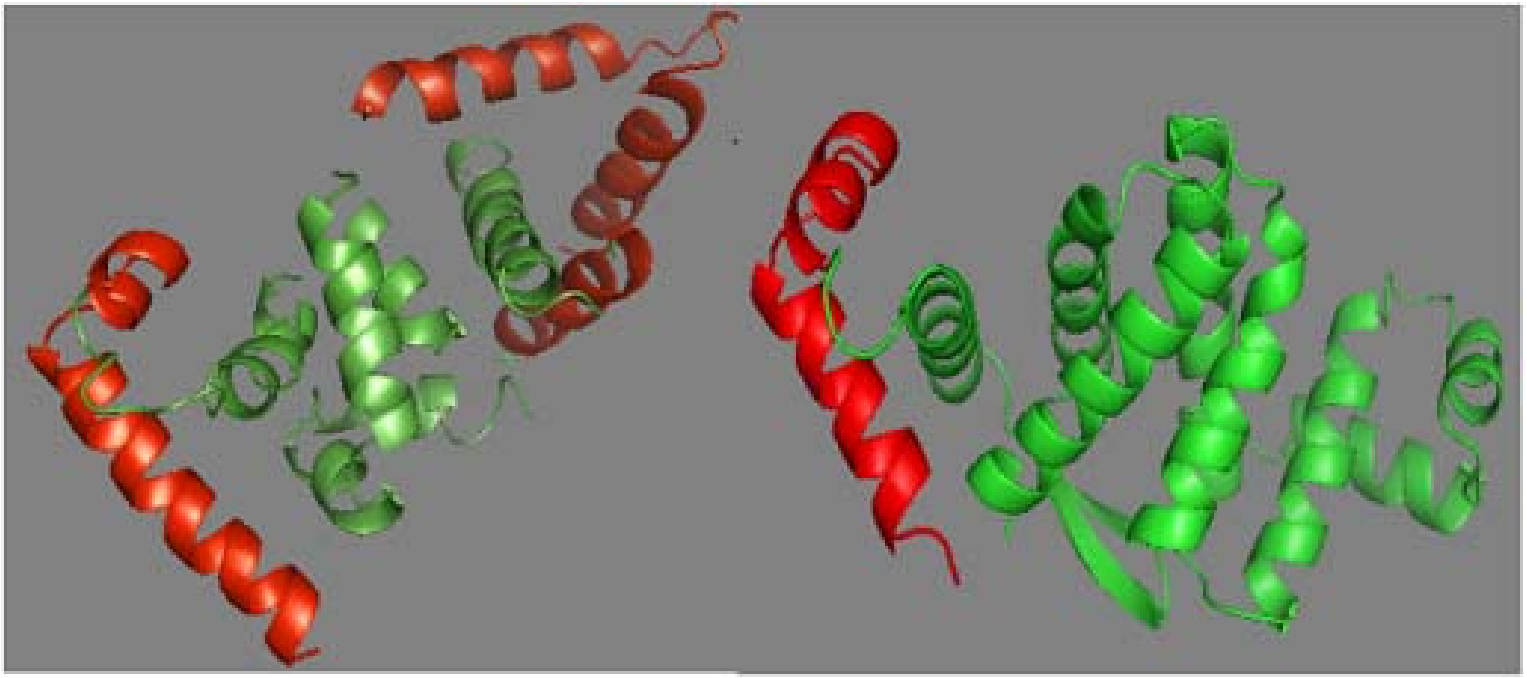


Figure 3-21. Cα-Polyalanine main chain model of C : DNA complex structure. The structure reveals that two dimer of C bind two subsequent major grooves in the same phase. C interacts with DNA using HTH from two different monomers of two different C dimer.

protein. The HTH motif of the Mor when compared to the C HTH, look structurally identical (Figure 3-22). In addition, Mor HTH is structurally very similar to the Trp R HTH. In Trp R, it is shown that, at a low concentration, a single dimer binds to its operator site and if the concentration of the protein is increased, a second dimer may bind. By cautiously extrapolating this data we can hypothesize that at a high concentration like that used in crystallization there will be sufficient competition for the binding sites that two different dimer may bind subsequent major groove, (2) In the present study, I have used a perfect symmetrical sequence within the C binding area. This symmetry may cause the protein to bind with the same affinity and not discriminate two different dimer molecules when compared to an asymmetric sequence like P_{lys} were if one C dimer bind the stronger proximal half the distal half may favor binding the C dimer already bound to proximal half in-order to stabilize the interaction. or the roundabout is the second C dimer cannot bind the weaker distal half since one half of the symmetry is already occupied, (3) In gel-shift assays done by (De *et al.*, 1997) using a mixture of C and a protein A-C fusion, it was shown that there is a possibility that a tetramer could bind. But in the assay no appropriate molecular weight was used to substantiate that, a tetramer could bind. Subsequent gel-shift and footprinting studies done by Sun and Hattman (1998) showed that C exhibited strong co-operativity in binding consistent with the binding of a tetramer,(4) The present structure may be a intermediate stage and crystallization may have only captured a snapshot of the whole DNA binding process. (5) The last hypotheses is that this might be nothing but an artifact of crystallization.

Figure 3-22. Comparison of C and Mor HTH motif. The ribbon figure shows structural similarity of the C and Mor HTH. The bottom panel shows the primary sequence alignment of Mor and C protein along with their secondary structure prediction based on Mor structure. The alignment shows the chemically identical (black shades) or similar amino-acids (grey shades) in Mor and C protein.


C HTH
Mor HTH
α1
α2
Mu Mor
Mu C
β1
α3
α4
α5
Mu Mor
Mu C

Chapter 4. General Discussion

This chapter recapitulates the findings from the studies described in this work and recommends ideas for further research to understand the binding specificity and mode of C : DNA interactions.

Major findings

Binding specificity of Mu transcription activator C

Transcription from the late promoters requires the activator protein C. The C protein binds the promoter in as a site-specific manner. Of the late promoters, P_{lys} and P_{mom} have been studied in detail. In P_{lys} the C-binding site is an imperfect dyad symmetry extending from –51 to –36. The binding site consists of two imperfect hexamer repeat (proximal and distal, 5'…TTCCTG***TCAC***CATAAT…3') separated by a four base spacer. Mutational analysis within the binding site showed that mutation in the distal half reduced C binding to varying degrees; whereas mutation in the proximal half caused severe C binding defect (Zhao, 1999). A strong C binding site (P_{sym}) was developed from P_{lys} were the hexamer of the distal half-site perfectly matched the wild type proximal half-site and separated by a four base spacer (5'…ATTATG***ACTC***CATAAT…3') (Jiang, 1999). Gel-shift experiments done with only the C–binding sequence from P_{sym} promoter have demonstrated that in addition core- binding sequences, flanking sequences from –58 to –52 and -36 to –29 are required for C to stabilize its interactions with the core-binding sequences. It addition the analysis also showed that positions –52, -53 and –32 do

not influence the C-binding specificity. In P_{mom} genetic and biochemical analysis have shown that C bind P_{mom} asymmetrically and based on these analyses a consensus binding sequence for C was developed (5'…TTAT…N_6…ATAACC…3') (Gindlesperger and Hattman, 1994; Ramesh and Nagaraja, 1996; Sun *et al.*, 1997). Curiously, the consensus sequence developed from P_{mom} has a six base spacer and whose contribution to C-binding was deemed minimal. However, in this study it is shown if the spacer length is not a rigid four base pair C-binding is completely abolished. Additionally I have shown that the bases flanking the IR spacer (–47, –46, –41 and –40) should be a symmetrical pair (–47 T and –40 A and –46 G and –41 C) for efficient C-binding. By merging the data from the present study with the data already available a conclusion can be drawn about the C-binding requirements and specificity; (1) A minimum length of 30 bp is required for efficient C-binding (2) within the 30 bp, the core binding site should be a perfect hexameric dyad symmetry and (3) The spacer between the hexamer should be no more or less than four bases. From the results of this study, I was able to derive information regarding C-binding specificity and compared it to the binding specificity proposed by Gindlesperger & Hattaman.

Expression, purification, crystallization and preliminary X-ray analysis of C protein bound to P_{sym} DNA

Previously, structural studies in C have been hampered due to the lack of a suitable protocol describing how to purify milligram quantities of soluble C protein. In this study, large quantities of functional C protein were purified to near homogeneity, crystallized it with P_{sym} DNA and preliminary structural information of how C binds its DNA was derived. The wild-type C protein expressed in JM109DE3 was purified using a

four- step chromatography procedure. Each step in the protocol was incorporated to enrich the C protein, remove contaminant proteins, decrease aggregation and increase solubility.

Crystallization was performed with purified WT C : DNA complex from P_{sym} and P_{lys} which varied in sequence and length using various commercially available screens in two different crystallization methods (hanging and sitting drop). Diffraction data was collected from a single crystal (0.1(L) x 0.1 (B) x 0.5 (H) μM), which was crystallized using the sitting drop method in 1.4 M ammonium sulfate 16% ethylene glycol pH 5.7. Due to the lack of a structural homologue to do molecular replacement (MR) a SeMet C : DNA complex was crystallized under the same WT condition and diffraction data was collected at 3.1 Å resolution. The phase angles were obtained using SAD phasing. A preliminary Cα- polyalanine main chain model was built based on the available phase angle information. The preliminary structure shows the general architecture of the protein-DNA complex. In the complex, two C-dimers are interacting without conformational changes with two adjacent major grooves on one face of the C-binding site. This structure is not consistent to structures of other DNA binding dimeric proteins and conflicts the gel-shifts results in Chapter 2. Gel-shifts have shown that IR spacer deletion in P_{sym} were detrimental to C binding which is suggestive that the binding defect will occur only if a single C dimer is occupying its binding site. If two dimers of C protein were binding independently there will not be any binding defect because the two C dimer molecules will be able to overcome the change in orientation of the binding sites caused by the deletion. Furthermore in the deletion gel-shifts a barely detectable level of binding was noticed which is suggestive that a single C dimer may be trying to

compensate for the change in orientation in the C binding site caused by deletion.

Presently it is very hard to draw a conclusion if there are two C dimers per DNA or one dimer per DNA since there are considerable differences in how gel-shift and crystallography are done. In-order to facilitate C : DNA crystallization a symmetrical C binding site (P_{sym}) was used instead of wild-type P_{lys}. In P_{lys} the proximal C binding is a stronger binding site (Jiang, 1999) when compared to distal binding site and this arrangement of binding site may favor a stepwise binding in which a single C dimer may bind the stronger binding site first and through a series of conformational changes may bind the weak binding site thereby stabilizing the complex. Since P_{sym} has two strong binding sites, two C dimers may bind the binding sites with the same binding affinity thereby preventing the second C molecule in the either dimer to bind as seen in the crystal structure. Additionally the type of binding seen in the complex crystal structure may not be seen *in-vivo* since in 2004, Mo found out that there is an UP like element just upstream of the C binding site which is required for α-CTD binding. If two C dimers are present as seen in the crystal structure α-CTD binding to its binding site may be physically hindered. This argument is validated by the footprinting assays done by Zhao (1999), who by changing the order of addition was able to show that a single C molecule may bind to its binding site first and then recruit the RNA polymerase.

The present structure may not reflect of how C interacts with its cognate DNA *invivo* since there is no precedence for this mode of binding and the available biochemical evidence does not validate this structure. There a few possible explanations as to why we have the present structure (1) as mentioned in the introduction, in crystallography crystals can only be obtained if higher concentration of macromolecules

are used. In protein DNA complex crystallization, crystal formation is usually driven by either the protein or DNA and if either of the macromolecules are in higher concentration than others abnormal crystals may form due to physico chemical variations normally not seen *invivo*, (2) in-order to drive crystallization a symmetrical DNA was used instead of the wild-type and due to the presence of two strong binding site the step wise transition by binding the stronger site first and then binding the weaker side is prevented. Since both sites have high affinity for C proteins, protein molecules binding to either side cannot displace each other, this combined with the high concentration might favor trapping the macromolecules as seen in the crystal and (3) the present structure may be an artifact of crystallization.

The present structure has raised numerous questions in addition to the question regarding the mode of interaction. Due to the poor resolution and phase angle problems only an initial polyalanine could be built, with no further information available of how certain amino-acids interact with each other as well as the DNA. Additionally, out of the 22 bp of DNA used in the crystallization only 21 bp is visible in the structure. In addition, there is no enough information if the structure is made up of dimers of dimers or independent tetramer.

Even though there is a lot of ambiguity with regard to the structure, this study has overcome substantial problems that were hindering structural studies in C and has taken a major step towards understanding the transcriptional activation mechanism of C.

Future directions

The present study has not addressed certain key questions with regard to binding specificity and many questions needs to be addressed with regard to the preliminary C-DNA structure. For the purpose of discussion, I will only be suggesting few ideas, which might address these questions.

Estimation of dissociation constants (Kd)

It has been shown qualitatively that, P_{sym} is a stronger binding promoter than P_{lys} but quantitatively this has not been shown. By estimating the dissociation constant, the strength of binding (or affinity) between the DNA and the C protein can be measured. The dissociation constant can be measured by a number of ways but the easiest is by doing gel-shifts. In gel-shifts the dissociation constant can be obtained by (a) titration of a low concentration of standard amount of DNA with increasing quantities of C protein and (b) titration of a high concentration of DNA with increasing quantities of C protein. Both these methods involve measuring the amount of free and bound DNA using densitometry or phosphorimaging. In both methods, it is very important to measure accurately the protein and DNA concentrations. The dissociation constant obtained through this method is not absolute but can used in comparative studies involving Mu late promoters.

Analytical ultracentrifugation (AUC) of C : DNA complex

The preliminary C-DNA crystal structure shows that two dimers of C are binding to a single binding site. Since there is no precedence for this mode of binding, further biochemical and biophysical investigations need to be done to validate the structure.

Analytical ultra-centrifugation is one of the methods, which can be used to test the structure. Analytical ultra-centrifugation can do two different experiments. (1) sedimentation velocity experiment and (2) sedimentation equilibrium experiments These method can reveal (a) The native molecular mass. AUC is presently the best method to determine the native molecular weight of the protein accurately, (b) Stoichiometry. High quality AUC data can easily determine if the native protein is a monomer or a multimer, (c) Assembly models. The assembly of a protein complex (eg. C : DNA complex) can be calculated from the determined molecular mass of the protein and DNA. It is even possible to follow the assembly when the different partners are added to the mixture one by one. In addition, the binding of protein to a ligand like DNA can be analyzed using sedimentation velocity methods because the DNA and the protein differ greatly in their sedimentation coefficients, (d) Conformation & shape. The conformation of a protein and as well as its macromolecular interactions can be studied using the sedimentation and diffusion coefficients obtained from the sedimentation velocity experiment. The overall conformation and shape of the protein or the protein complex can be compared with the crystal structure to assess the applicability of these macromolecules in solution, and (e) Association. The sedimentation equilibrium method is a very sensitive method to study relatively weak associations constants (K_a).

Crystallization of truncated C : DNA complex

The information available from the present C : DNA complex structure is very limited due to its low resolution. In order to understand how C recognizes it binding sequence a higher resolution structure is needed. Resolution of the structure is mostly

dependent on the crystal quality. The crystal quality can be improved by changing many different variables either in the protein or in its ligand. During the course of C : DNA crystallization only the length and the sequence of the DNA used in the complex was varied. An alternative approach to improve the resolution is to use a modified or truncated C protein in crystallization. Truncation of a protein may substantially improve the quality of the crystal because the flexible region in the protein may interfere with proper crystal formation thereby reducing its quality. Drawback to this approach is to identify which terminus and how many residues to truncate. This problem can be overcome if regions important for protein functions have been identified or if a protein similar to the protein of interest has already been solved. The Mor protein is the closest homologue of C. In the crystal structure of Mor, 26 residues at the N terminus and 9 residues at the C terminus are not visible (Kumaraswami *et al.*, 2004). The absence of these residues in the crystal structure is suggestive that they are flexible and disordered. By comparing the primary amino-acid sequence of C with Mor the number of amino-acid that could be truncated in the N and C terminus can be identified. Since a number of N and C terminal truncation combinations are possible, a rational approach in truncation is needed.

Crystallization of C with modified DNA

The quality of the crystal can be improved if better phase angles can be obtained. The phases angle for the present C : DNA complex was obtained through Single wavelength anomalous dispersion method (SAD). In SAD phasing a single wavelength is used to calculate the phases so there is phase ambiguity which may result in poor quality

structure. These ambiguities are not usually seen in structures solved with multi-wavelength anomalous dispersion method (MAD) that uses two wavelengths to calculate the phase angles. Due to the fragility of the C : DNA complex crystal heavy atom soaks needed for MAD phasing could not be performed. To overcome this problem DNA in the complex offers a simple and direct way to incorporate heavy atoms for MAD phasing. Heavy atoms like bromine and iodine can be incorporated in the oligonucleotides during its synthesis. Typically, 5-bromo-deoxyuridine and 5-iodo-deoxyuridine are introduced as isomorphous substitutions for thymidine and 5-bromo-deoxycytosine and 5-iodo-deoxycytosine as nearly isomorphous substitutions for deoxycytosine. The main drawback in using this approach is that the modified bases may interfere with protein binding. Therefore, before crystallization the modified oligos need to be tested for protein binding using gel-shift assay.

.

List of References

Airlie, J. M. University of Cambridge. http://www-structmed.cimr.cam.ac.uk/Course/Crystals/Theory/methods.html. Accessed April 6, 2007.

Alfano, C., Sanfelice, D., Babon, J., Kelly, G., Jacks, A., Curry, S., and Conte, M.R. (2004) Structural analysis of cooperative RNA binding by the La motif and central RRM domain of human La protein. *Nat Struct Mol Biol* **11**: 323-329.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.

Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., and Iyer, L.M. (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev* **29**: 231-262.

Aravind, L., and Koonin, E.V. (1999) DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res* **27**: 4658-4670.

Artsimovitch, I., and Howe, M.M. (1996) Transcription activation by the bacteriophage Mu Mor protein: analysis of promoter mutations in Pm identifies a new region required for promoter function. *Nucleic Acids Res* **24**: 450-457.

Assa-Munt, N., Mortishire-Smith, R.J., Aurora, R., Herr, W., and Wright, P.E. (1993) The solution structure of the Oct-1 POU-specific domain reveals a striking similarity to the bacteriophage lambda repressor DNA-binding domain. *Cell* **73**: 193-205.

Bachmann, B.J. (1987) Derivations and genotypes of some mutant derivatives of Escherichia coli K-12,. In *Escherichia coli and Salmonella: Cellular and Molecular Biology,*. Vol. II. F. C. Neidhardt, R.C.I., J. L. Ingraham, E. C. C. Lin, K. B. Low (ed). Washington, DC: American Society of Microbiology, pp. 1190-1219.

Balke, V., Nagaraja, V., Gindlesperger, T., and Hattman, S. (1992) Functionally distinct RNA polymerase binding sites in the phage Mu mom promoter region. *Nucleic Acids Res* **20**: 2777-2784.

Barne, K.A., Bown, J.A., Busby, S.J., and Minchin, S.D. (1997) Region 2.5 of the *Escherichia coli* RNA polymerase sigma70 subunit is responsible for the recognition of the 'extended-10' motif at promoters. *Embo J* **16**: 4034-4040.

Bernhard, R. http://ruppweb.dyndns.org/. Accessed April 6, 2007.

Blatter, E.E., Ross, W., Tang, H., Gourse, R.L., and Ebright, R.H. (1994) Domain organization of RNA polymerase alpha subunit: C-terminal 85 amino-acids constitute a domain capable of dimerization and DNA binding. *Cell* **78**: 889-896.

Bolker, M., Wulczyn, F.G., and Kahmann, R. (1989) Role of bacteriophage Mu C protein in activation of the mom gene promoter. *J Bacteriol* **171**: 2019-2027.

Borukhov, S., and Severinov, K. (2002) Role of the RNA polymerase sigma subunit in transcription initiation. *Res Microbiol* **153**: 557-562.

Brennan, R.G. (1993) The winged-helix DNA-binding motif: another helix-turn-helix takeoff. *Cell* **74**: 773-776.

Brennan, R.G., and Matthews, B.W. (1989) The helix-turn-helix DNA binding motif. *J Biol Chem* **264**: 1903-1906.

Brunger, A.T. (1992) *X-plor manual.* New Haven: Yale University Press.

Buc, H., and McClure, W.R. (1985) Kinetics of open complex formation between *Escherichia coli* RNA polymerase and the lac UV5 promoter. Evidence for a sequential mechanism involving three steps. *Biochemistry* **24**: 2712-2723.

Busby, S., and Ebright, R.H. (1994) Promoter structure, promoter recognition, and transcription activation in prokaryotes. *Cell* **79**: 743-746.

Busby, S., and Ebright, R.H. (1999) Transcription activation by catabolite activator protein (CAP). *J Mol Biol* **293**: 199-213.

Cai, M., Zheng, R., Caffrey, M., Craigie, R., Clore, G.M., and Gronenborn, A.M. (1997) Solution structure of the N-terminal zinc binding domain of HIV-1 integrase. *Nat Struct Biol* **4**: 567-577.

Campbell, E.A., Muzzin, O., Chlenov, M., Sun, J.L., Olson, C.A., Weinman, O., Trester-Zedlitz, M.L., and Darst, S.A. (2002) Structure of the bacterial RNA polymerase promoter specificity sigma subunit. *Mol Cell* **9**: 527-539.

Campos, A., Zhang, R.G., Alkire, R.W., Matsumura, P., and Westbrook, E.M. (2001) Crystal structure of the global regulator FlhD from *Escherichia coli* at 1.8 A resolution. *Mol Microbiol* **39**: 567-580.

Carey, J. (1991) Gel retardation. *Methods Enzymol* **208**: 103-117.

Carter, C.W., Jr., and Carter, C.W. (1979) Protein crystallization using incomplete factorial experiments. *J Biol Chem* **254**: 12219-12223.

Caruthers, M.H., Beaucage, S.L., Becker, C., Efcavitch, J.W., Fisher, E.F., Galluppi, G., Goldman, R., deHaseth, P., Matteucci, M., McBride, L., and *et al*. (1983) Deoxyoligonucleotide synthesis via the phosphoramidite method. *Gene Amplif Anal* **3**: 1-26.

CCP4 (1994) The CCP4 suite: programs for protein crystallography. *Acta Cryst D* **50**: 760-763.

Chan, B., and Busby, S. (1989) Recognition of nucleotide sequences at the Escherichia coli galactose operon P1 promoter by RNA polymerase. *Gene* **84**: 227-236.

Chiang, L.W., and Howe, M.M. (1993) Mutational analysis of a C-dependent late promoter of bacteriophage Mu. *Genetics* **135**: 619-629.

Chuprina, V.P., Rullmann, J.A., Lamerichs, R.M., van Boom, J.H., Boelens, R., and Kaptein, R. (1993) Structure of the complex of lac repressor headpiece and an 11 base-pair half-operator determined by nuclear magnetic resonance spectroscopy and restrained molecular dynamics. *J Mol Biol* **234**: 446-462.

Clark, K.L., Halay, E.D., Lai, E., and Burley, S.K. (1993) Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364**: 412-420.

Clubb, R.T., Omichinski, J.G., Savilahti, H., Mizuuchi, K., Gronenborn, A.M., and Clore, G.M. (1994) A novel class of winged helix-turn-helix protein: the DNA-binding domain of Mu transposase. *Structure* **2**: 1041-1048.

Cowtan, K. (1999) Error estimation and bias correction in phase-improvement calculations. *Acta Cryst D* **55**: 1555-1567.

Darst, S.A., Polyakov, A., Richter, C., and Zhang, G. (1998) Structural studies of *Escherichia coli* RNA polymerase. *Cold Spring Harb Symp Quant Biol* **63**: 269-276.

De, A., Paul, B.D., Ramesh, V., and Nagaraja, V. (1997) Use of protein A gene fusions for the analysis of structure-function relationship of the transactivator protein C of bacteriophage Mu. *Protein Eng* **10**: 935-941.

De, A., Ramesh, V., Mahadevan, S., and Nagaraja, V. (1998) Mg2+ mediated sequence-specific binding of transcriptional activator protein C of bacteriophage Mu to DNA. *Biochemistry* **37**: 3831-3838.

Dekker, N., Cox, M., Boelens, R., Verrijzer, C.P., van der Vliet, P.C., and Kaptein, R. (1993) Solution structure of the POU-specific DNA-binding domain of Oct-1. *Nature* **362**: 852-855.

Dickerson RE, K.J., Strandberg BE (1961) The phase problem and isomorphous replacement methods in protein structurs. In *Computing methods and the phase problem in X-ray crystal analysis*. Pepinsky R, R.J., Speakman JC (ed). Oxford: Pergamon Press, pp. 236-251.

Dodd, I.B., and Egan, J.B. (1987) Systematic method for the detection of potential lambda Cro-like DNA-binding regions in proteins. *J Mol Biol* **194**: 557-564.

Donaldson, L.W., Petersen, J.M., Graves, B.J., and McIntosh, L.P. (1994) Secondary structure of the ETS domain places murine Ets-1 in the superfamily of winged helix-turn-helix DNA-binding proteins. *Biochemistry* **33**: 13509-13516.

Dong, G., Chakshusmathi, G., Wolin, S.L., and Reinisch, K.M. (2004) Structure of the La motif: a winged helix domain mediates RNA binding via a conserved aromatic patch. *Embo J* **23**: 1000-1007.

Ebright, R.H. (1993) Transcription activation at Class I CAP-dependent promoters. *Mol Microbiol* **8**: 797-802.

Ebright, R.H., and Busby, S. (1995) The *Escherichia coli* RNA polymerase alpha subunit: structure and function. *Curr Opin Genet Dev* **5**: 197-203.

Estrem, S.T., Gaal, T., Ross, W., and Gourse, R.L. (1998) Identification of an UP element consensus sequence for bacterial promoters. *Proc Natl Acad Sci U S A* **95**: 9761-9766.

Fairall, L., Schwabe, J.W., Chapman, L., Finch, J.T., and Rhodes, D. (1993) The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature* **366**: 483-487.

Fairfield, F.R., Newport, J.W., Dolejsi, M.K., and von Hippel, P.H. (1983) On the processivity of DNA replication. *J Biomol Struct Dyn* **1**: 715-727.

Feng, J.A., Johnson, R.C., and Dickerson, R.E. (1994) Hin recombinase bound to DNA: the origin of specificity in major and minor groove interactions. *Science* **263**: 348-355.

Finney, M. (1990) The homeodomain of the transcription factor LF-B1 has a 21 amino-acid loop between helix 2 and helix 3. *Cell* **60**: 5-6.

Fogh, R.H., Ottleben, G., Ruterjans, H., Schnarr, M., Boelens, R., and Kaptein, R. (1994) Solution structure of the LexA repressor DNA binding domain determined by 1H NMR spectroscopy. *Embo J* **13**: 3936-3944.

Fujinaga, M.R., R.J (1987) Experiences with a new translation-function program. *J appl Crystallogr* **20**: 517-521.

Gaal, T., Ross, W., Blatter, E.E., Tang, H., Jia, X., Krishnan, V.V., Assa-Munt, N., Ebright, R.H., and Gourse, R.L. (1996) DNA-binding determinants of the alpha subunit of RNA polymerase: novel DNA-binding domain architecture. *Genes Dev* **10**: 16-26.

Gajiwala, K.S., and Burley, S.K. (2000) Winged helix proteins. *Curr Opin Struct Biol* **10**: 110-116.

Gindlesperger, T.L., and Hattman, S. (1994) In vitro transcriptional activation of the phage Mu mom promoter by C protein. *J Bacteriol* **176**: 2885-2891.

Gomis-Ruth, F.X., Sola, M., Acebo, P., Parraga, A., Guasch, A., Eritja, R., Gonzalez, A., Espinosa, M., del Solar, G., and Coll, M. (1998) The structure of plasmid-encoded

transcriptional repressor CopG unliganded and bound to its operator. *Embo J* **17**: 7404-7415.

Goosen, N.P.v.d.P. (1987) Regulation of transcription. In *Phage Mu*. Symonds, N., Toussaint, A., van de Putte, P., and Howe, M.M (ed). New York: Cold Spring Harbor Laboratory Press, pp. 41-52.

Gribskov, M., and Burgess, R.R. (1986) Sigma factors from E. coli, B. subtilis, phage SP01, and phage T4 are homologous proteins. *Nucleic Acids Res* **14**: 6745-6763.

Grishin, N.V. (2000) Two tricks in one bundle: helix-turn-helix gains enzymatic activity. *Nucleic Acids Res* **28**: 2229-2233.

Gross, C.A., Chan, C., Dombroski, A., Gruber, T., Sharp, M., Tupy, J., and Young, B. (1998) The functional and regulatory roles of sigma factors in transcription. *Cold Spring Harb Symp Quant Biol* **63**: 141-155.

Guo, F., Gopaul, D.N., and van Duyne, G.D. (1997) Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature* **389**: 40-46.

Harley, C.B., and Reynolds, R.P. (1987) Analysis of *E. coli* promoter sequences. *Nucleic Acids Res* **15**: 2343-2361.

Harrison, C.J., Bohm, A.A., and Nelson, H.C. (1994) Crystal structure of the DNA binding domain of the heat shock transcription factor. *Science* **263**: 224-227.

Hattman, S., Ives, J., Margolin, W., and Howe, M.M. (1985) Regulation and expression of the bacteriophage mu mom gene: mapping of the trans-activation (dad) function to the C region. *Gene* **39**: 71-76.

Hauptman, H. (1982) On integrating the techniques of direct methods and isomorphous replacement. I. The theoretical basis. *Acta Cryst A* **38**: 289-294.

Helmann, J.D., and Chamberlin, M.J. (1988) Structure and function of bacterial sigma factors. *Annu Rev Biochem* **57**: 839-872.

Hendrickson, W.A. (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**: 51-58.

Hendrickson, W.A.O., C.M (1997) Phase Determination from Multi-wavelength Anomalous Diffraction Measurements. *Methods Enzymol* **276**: 494-523.

Hengming K (1997) Overview of isomorphous replacement phasing. *Methods Enzymol* **276**: 448-461.

Hinrichs, W., Kisker, C., Duvel, M., Muller, A., Tovar, K., Hillen, W., and Saenger, W. (1994) Structure of the Tet repressor-tetracycline complex and regulation of antibiotic resistance. *Science* **264**: 418-420.

Hope, H. (1990) Crystallography of biological macromolecules at ultra-low temperature. *Annu Rev Biophys Biophys Chem* **19**: 107-126.

Hsu, L.M. (2002) Open season on RNA polymerase. *Nat Struct Biol* **9**: 502-504.

Igarashi, K., Hanamura, A., Makino, K., Aiba, H., Mizuno, T., Nakata, A., and Ishihama, A. (1991) Functional map of the alpha subunit of *Escherichia coli* RNA polymerase: two modes of transcription activation by positive factors. *Proc Natl Acad Sci U S A* **88**: 8958-8962.

Igarashi, K., and Ishihama, A. (1991) Bipartite functional map of the *E.coli* RNA polymerase alpha subunit: involvement of the C-terminal region in transcription activation by cAMP-CRP. *Cell* **65**: 1015-1022.

Ishihama, A. (1988) Promoter selectivity of prokaryotic RNA polymerases. *Trends in Genetics* **4(10)**: 282-286.

Ishihama, A. (2000) Functional modulation of *Escherichia coli* RNA polymerase. *Annu Rev Microbiol* **54**: 499-518.

Jancarik, J., and Kim, S.H. (1991) Sparse matrix sampling: a screening method for crystallization of proteins. *Journal of Applied Crystallography* **24**: 409-411.

Jeon, Y.H., Negishi, T., Shirakawa, M., Yamazaki, T., Fujita, N., Ishihama, A., and Kyogoku, Y. (1995) Solution structure of the activator contact domain of the

RNA polymerase alpha subunit. *Science* **270**: 1495-1497.

Jiang, Y. (1999) Ph.D. dissertation: Mutational analysis of C protein: the late gene activator of bacteriophage Mu. The University of Tennessee Health Science Center, Memphis.

Jones, T.A., Zou, J.Y., Cowan, S.W., and Kjeldgaard, M. (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Cryst A* **47 (Pt 2)**: 110-119.

Kabsch, W. (1988) Automatic indexing of rotation diffraction patterns. *Journal of Applied Crystallography* **21**: 67-71.

Keilty, S., and Rosenberg, M. (1987) Constitutive function of a positively regulated promoter reveals new sequences essential for activity. *J Biol Chem* **262**: 6389-6395.

Khare, D., Ziegelin, G., Lanka, E., and Heinemann, U. (2004) Sequence-specific DNA binding determined by contacts outside the helix-turn-helix motif of the ParB homolog KorB. *Nat Struct Mol Biol* **11**: 656-663.

Kimura, M., Fujita, N., and Ishihama, A. (1994) Functional map of the alpha subunit of *Escherichia coli* RNA polymerase. Deletion analysis of the amino-terminal assembly domain. *J Mol Biol* **242**: 107-115.

Kimura, M., and Ishihama, A. (1995) Functional map of the alpha subunit of *Escherichia coli* RNA polymerase: insertion analysis of the amino-terminal assembly domain. *J Mol Biol* **248**: 756-767.

Kimura, M., and Ishihama, A. (1996) Subunit assembly *in vivo* of *Escherichia coli* RNA polymerase: role of the amino-terminal assembly domain of alpha subunit. *Genes Cells* **1**: 517-528.

Kirkegaard, K., Buc, H., Spassky, A., and Wang, J.C. (1983) Mapping of single-stranded regions in duplex DNA at the sequence level: single-strand-specific cytosine methylation in RNA polymerase-promoter complexes. *Proc Natl Acad Sci U S A* **80**: 2544-2548.

Kissinger, C.R., Liu, B.S., Martin-Blanco, E., Kornberg, T.B., and Pabo, C.O. (1990) Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: a framework for understanding homeodomain-DNA interactions. *Cell* **63**: 579-590.

Klemm, J.D., Rould, M.A., Aurora, R., Herr, W., and Pabo, C.O. (1994) Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell* **77**: 21-32.

Kodandapani, R., Pio, F., Ni, C.Z., Piccialli, G., Klemsz, M., McKercher, S., Maki, R.A., and Ely, K.R. (1996) A new pattern for helix-turn-helix recognition revealed by the PU.1 ETS-domain-DNA complex. *Nature* **380**: 456-460.

Kumaraswami, M., Howe, M.M., and Park, H.W. (2004) Crystal structure of the Mor protein of bacteriophage Mu, a member of the Mor/C family of transcription activators. *J Biol Chem* **279**: 16581-16590.

Kustu, S., North, A.K., and Weiss, D.S. (1991) Prokaryotic transcriptional enhancers and enhancer-binding proteins. *Trends Biochem Sci* **16**: 397-402.

La Fortelle, E.de., and Bricogne, G. (1997) Maximum-likelihood heavy-atom parameter refinement for the MIR and MAD methods. *Methods Enzymol* **276**: 472-494.

Lai, E., Clark, K.L., Burley, S.K., and Darnell, J.E., Jr. (1993) Hepatocyte nuclear factor 3/fork head or "winged helix" proteins: a family of transcription factors of diverse biologic function. *Proc Natl Acad Sci U S A* **90**: 10421-10423.

Landick, R., Vaughn, V., Lau, E.T., VanBogelen, R.A., Erickson, J.W., and Neidhardt, F.C. (1984) Nucleotide sequence of the heat shock regulatory gene of E. coli suggests its protein product may be a transcription factor. *Cell* **38**: 175-182.

Langs, D.A., Blessing, R.H., and Guo, D. (1999) Progress on the direct-methods solution of macromolecular structures using single-wavelength anomalous-dispersion (SAS) data. *Acta Cryst A* **55**: 755-760.

Liang, H., Olejniczak, E.T., Mao, X., Nettesheim, D.G., Yu, L., Thompson, C.B., and Fesik, S.W. (1994) The secondary structure of the ets domain of human Fli-1 resembles that of the helix-turn-helix DNA-binding motif of the *Escherichia coli* catabolite gene activator protein. *Proc Natl Acad Sci U S A* **91**: 11655-11659.

Lisser, S., and Margalit, H. (1993) Compilation of *E. coli* mRNA promoter sequences. *Nucleic Acids Res* **21**: 1507-1516.

Liu, T., DeRose, E.F., and Mullen, G.P. (1994) Determination of the structure of the DNA binding domain of gamma delta resolvase in solution. *Protein Sci* **3**: 1286-1295.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**: 251-260.

Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res* **29**: 2860-2874.

Ma, J., and Howe, M.M. (2004) Binding of the C-Terminal Domain of the {alpha} Subunit of RNA Polymerase to the Phage Mu Middle Promoter. *J. Bacteriol.* **186**: 7858-7864.

Malhotra, A., Severinova, E., and Darst, S.A. (1998) Mapping the σ^{70}subunit contact sites on *Escherichia coli* RNA polymerase with a σ^{70}-conjugated chemical protease *PNAS* **95**: 6021-6026.

Mandal, N., Su, W., Haber, R., Adhya, S., and Echols, H. (1990) DNA looping in cellular repression of transcription of the galactose operon. *Genes Dev* **4**: 410-418.

Mandel, M., and Higa, A. (1970) Calcium-dependent bacteriophage DNA infection. *J Mol Biol* **53**: 159-162.

Margolin, W., Rao, G., and Howe, M.M. (1989) Bacteriophage Mu late promoters: four late transcripts initiate near a conserved sequence. *J Bacteriol* **171**: 2003-2018.

Marmorstein, R., Carey, M., Ptashne, M., and Harrison, S.C. (1992) DNA recognition by GAL4: structure of a protein-DNA complex. *Nature* **356**: 408-414.

Marmorstein, R., and Harrison, S.C. (1994) Crystal structure of a PPR1-DNA complex: DNA recognition by proteins containing a Zn2Cys6 binuclear cluster. *Genes Dev* **8**: 2504-2512.

Marrs, C.F., and Howe, M.M. (1990) Kinetics and regulation of transcription of bacteriophage Mu. *Virology* **174**: 192-203.

Mathee, K., and Howe, M.M. (1990) Identification of a positive regulator of the Mu middle operon. *J Bacteriol* **172**: 6641-6650.

Matthews, B.W., Ohlendorf, D.H., Anderson, W.F., and Takeda, Y. (1982) Structure of the DNA-binding region of lac repressor inferred from its homology with cro repressor. *Proc Natl Acad Sci U S A* **79**: 1428-1432.

McPherson, A. (1990) Current approaches to macromolecular crystallization. *Eur J Biochem* **189**: 1-23.

Mekler, V., Kortkhonjia, E., Mukhopadhyay, J., Knight, J., Revyakin, A., Kapanidis, A.N., Niu, W., Ebright, Y.W., Levy, R., and Ebright, R.H. (2002) Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex. *Cell* **108**: 599-614.

Mo, Y. (2004) Ph.D. dissertation: RNA Polymerase recruitment, Dimerization, and Alpha CTD Dependent Activation by Members of the Mor/C Family of Transcription Activators. The University of Tennessee Health Science Center, Memphis.

Moore, M.H., Gulbis, J.M., Dodson, E.J., Demple, B., and Moody, P.C. (1994) Crystal structure of a suicidal DNA repair protein: the Ada O6-methylguanine-DNA methyltransferase from *E. coli*. *Embo J* **13**: 1495-1501.

Mukherjee, A., Cui, Y., Ma, W., Liu, Y., Ishihama, A., Eisenstark, A., and Chatterjee, A.K. (1998) RpoS (sigma-S) controls expression of rsmA, a global regulator of secondary metabolites, harpin, and extracellular proteins in Erwinia carotovora. *J Bacteriol* **180**: 3629-3634.

Murakami, K.S., and Darst, S.A. (2003) Bacterial RNA polymerases: the wholo story. *Curr Opin Struct Biol* **13**: 31-39.

Murakami, K., Fujita, N., and Ishihama, A. (1996) Transcription factor recognition surface on the RNA polymerase alpha subunit is involved in contact with the DNA enhancer element. *Embo J* **15**: 4358-4367.

Murakami, K.S., Masuda, S., and Darst, S.A. (2002) Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 A resolution. *Science* **296**: 1280-1284.

Nagaraja, V., Hecht, G., and Hattman, S. (1988) The phage Mu 'late' gene transcription activator, C, is a site-specific DNA binding protein. *Biochem Pharmacol* **37**: 1809-1810.

Navaza J, S.P. (1997) AmoRe: an automated molecular replacement program package. *Methods Enzymol* **276**: 581-594.

Newlands, J.T., Ross, W., Gosink, K.K., and Gourse, R.L. (1991) Factor-independent activation of *Escherichia coli* rRNA transcription. II. Characterization of complexes of rrnB P1 promoters containing or lacking the upstream activator region with *Escherichia coli* RNA polymerase. *J Mol Biol* **220**: 569-583.

Ogata, C.M. (1998) MAD phasing grows up. *Nat Struct Biol* **5 Suppl**: 638-640.

Ogata, K., Hojo, H., Aimoto, S., Nakai, T., Nakamura, H., Sarai, A., Ishii, S., and Nishimura, Y. (1992) Solution structure of a DNA-binding unit of Myb: a helix-turn-helix-related motif with conserved tryptophans forming a hydrophobic core. *Proc Natl Acad Sci U S A* **89**: 6428-6432.

Ohlendorf, D.H., Anderson, W.F., Fisher, R.G., Takeda, Y., and Matthews, B.W. (1982) The molecular basis of DNA-protein recognition inferred from the structure of cro repressor. *Nature* **298**: 718-723.

Ohlendorf, D.H., Anderson, W.F., and Matthews, B.W. (1983) Many gene-regulatory proteins appear to have a similar alpha-helical fold that binds DNA and evolved from a common precursor. *J Mol Evol* **19**: 109-114.

Ohlsen, K.L., and Gralla, J.D. (1992) Interrelated effects of DNA supercoiling, ppGpp, and low salt on melting within the *Escherichia coli* ribosomal RNA rrnB P1 promoter. *Mol Microbiol* **6**: 2243-2251.

Otting, G., Qian, Y.Q., Billeter, M., Muller, M., Affolter, M., Gehring, W.J., and Wuthrich, K. (1990) Protein--DNA contacts in the structure of a homeodomain--DNA complex determined by nuclear magnetic resonance spectroscopy in solution. *Embo J* **9**: 3085-3092.

Otwinowski, Z. (1991) *Isomorphous replacement and anamolous scattering.* Proceedings of the CCP4 study weekend, 25-26 January 1991, edited by W. Wolf, P.R. Evans and A.G.W. Leslie, Warrington: Daresbury Laboratory, pp. 80-85.

Otwinowski, Z., and Minor, W. (1997) Processing of X-ray diffraction data collected in the oscillation mode. *Methods Enzymol* **276**: 307-326.

Owens, J. T., Miyake, R., Murakami, K., Chmura, A.J., Fujita, N., Ishihama, A., and Meares. C.F. (1998) Mapping the σ^{70}subunit contact sites on *Escherichia coli* RNA polymerase with a σ^{70}-conjugated chemical protease *PNAS* **95**: 6021-6026.

Paolozzi, L.G., P. (2006) *The Bacteriophages*. Calendar, R. (ed). New York: Oxford University Press, pp. 469-496.

Pavletich, N.P., and Pabo, C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **252**: 809-817.

Ptashne, M. (2004) *Genetic Switch: Phage Lambda Revisited.* New York: Cold Spring Harbor Laboratory Press.

Qian, Y.Q., Furukubo-Tokunaga, K., Resendez-Perez, D., Muller, M., Gehring, W.J., and Wuthrich, K. (1994) Nuclear magnetic resonance solution structure of the fushi tarazu homeodomain from Drosophila and comparison with the Antennapedia homeodomain. *J Mol Biol* **238**: 333-345.

Ramakrishnan, V., Finch, J.T., Graziano, V., Lee, P.L., and Sweet, R.M. (1993) Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature* **362**: 219-223.

Ramesh, V., De, A., and Nagaraja, V. (1994) Overproduction and purification of C protein, the late gene transcription activator from phage Mu. *Protein Expr Purif* **5**: 379-384.

Ramesh, V., and Nagaraja, V. (1996) Sequence-specific DNA binding of the phage Mu C protein: footprinting analysis reveals altered DNA conformation upon protein binding. *J Mol Biol* **260**: 22-33.

Randy, J.R. University of Cambridge. http://www-structmed.cimr.cam.ac.uk/Course/Overview/Overview.html. Accessed April 6, 2007.

Raumann, B.E., Rould, M.A., Pabo, C.O., and Sauer, R.T. (1994) DNA recognition by beta-sheets in the Arc repressor-operator crystal structure. *Nature* **367**: 754-757.

Ross, W., Ernst, A., and Gourse, R.L. (2001) Fine structure of *E. coli* RNA polymerase-promoter interactions: alpha subunit binding to the UP element minor groove. *Genes Dev* **15**: 491-506.

Ross, W., Gosink, K.K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K., and Gourse, R.L. (1993) A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science* **262**: 1407-1413.

Rossman M. (1972) *The Molecular replacement method.* New York: Gordon and Breach.

Rossmann, M. (1990) The molecular replacement method. *Acta Cryst A* **46**: 73-82.

Saucier, J.M., and Wang, J.C. (1972) Angular alteration of the DNA helix by E. coli RNA polymerase. *Nat New Biol* **239**: 167-170.

Sauer, R.T., Yocum, R.R., Doolittle, R.F., Lewis, M., and Pabo, C.O. (1982) Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature* **298**: 447-451.

Schultz, S.C., Shields, G.C., and Steitz, T.A. (1991) Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science* **253**: 1001-1007.

Schumacher, M.A., Choi, K.Y., Zalkin, H., and Brennan, R.G. (1994) Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. *Science* **266**: 763-770.

Selmer, M., and Su, X.D. (2002) Crystal structure of an mRNA-binding fragment of Moorella thermoacetica elongation factor SelB. *Embo J* **21**: 4145-4153.

Severinov, K., Kashlev, M., Severinova, E., Bass, I., McWilliams, K., Kutter, E., Nikiforov, V., Snyder, L., and Goldfarb, A. (1994) A non-essential domain of *Escherichia coli* RNA polymerase required for the action of the termination factor Alc. *J Biol Chem* **269**: 14254-14259.

Shanmuganatham, K. K., Ravichandran, M., Howe, M.M., and Park, H.W. (2007). Crystallization and preliminary X-ray analysis of phage Mu activator protein C in a complex with promoter DNA. *Acta Cryst F* **63**, 620-623.

Shao, X., and Grishin, N.V. (2000) Common fold in helix-hairpin-helix proteins. *Nucleic Acids Res* **28**: 2643-2650.

Sharp, M.M., Chan, C.L., Lu, C.Z., Marr, M.T., Nechaev, S., Merritt, E.W., Severinov, K., Roberts, J.W., and Gross, C.A. (1999) The interface of sigma with core RNA polymerase is extensive, conserved, and functionally specialized. *Genes Dev* **13**: 3015-3026.

Siebenlist, U. (1979) RNA polymerase unwinds an 11-base pair segment of a phage T7 promoter. *Nature* **279**: 651-652.

Smith, J.L. (1991) Determination of three-dimensional structure by multiwavelength anomalous diffraction. *Current Opinion in Structural Biology*: 1002-1011.

Smyth, M.S., and Martin, J.H.J. (2000) X-ray crystallography *Mol Pathol* **53**: 8-14.

Steitz, T.A., Ohlendorf, D.H., McKay, D.B., Anderson, W.F., and Matthews, B.W. (1982) Structural similarity in the DNA-binding domains of catabolite gene activator and cro repressor proteins. *Proc Natl Acad Sci U S A* **79**: 3097-3100.

Stoddard, S.F., and Howe, M.M. (1989) Localization and regulation of bacteriophage Mu promoters. *J Bacteriol* **171**: 3440-3448.

Studier, F.W., Rosenberg, A.H., Dunn, J.J., and Dubendorff, J.W. (1990) Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol* **185**: 60-89.

Sun, W., and Hattman, S. (1998) Bidirectional transcription in the mom promoter region of bacteriophage Mu. *J Mol Biol* **284**: 885-892.

Sun, W., Hattman, S., and Kool, E. (1997) Interaction of the bacteriophage Mu transcriptional activator protein, C, with its target site in the mom promoter. *J Mol Biol* **273**: 765-774.

Swindells, M.B. (1995) Identification of a common fold in the replication terminator protein suggests a possible mode for DNA binding. *Trends Biochem Sci* **20**: 300-302.

Symonds, N., Toussaint, A., van de Putte, P., and Howe, M.M (1987) *Phage Mu.* Cold Spring, New York: Cold Spring Harbor Laboratory.

Tahirov, T.H., Temiakov, D., Anikin, M., Patlan, V., McAllister, W.T., Vassylyev, D.G., and Yokoyama, S. (2002) Structure of a T7 RNA polymerase elongation complex at 2.9 A resolution. *Nature* **420**: 43-50.

Tartof, K.D., Hobbs, C.A. (1987) Improved media for growing plasmid and cosmid clones. *Bethesda Res. Lab* **Focus 9**.

Taylor CA, M.K. (1959) An improved method for determining the relative positions of molecules. *Acta Cryst*: 101-105.

Terwilliger, T. (2000) Maximum-likelihood density modification. *Acta Cryst D* **56**: 965-972.

Terwilliger, T.C., and Berendzen, J. (1996) Correlated phasing of multiple isomorphous replacement data. *Acta Cryst D* **52**: 749-757.

Terwilliger, T.C., and Berendzen, J. (1999) Discrimination of solvent from protein regions in native Fouriers as a means of evaluating heavy-atom solutions in the MIR and MAD methods. *Acta Cryst D* **55**: 501-505.

Terwilliger, T.C., Eisenberg, D. (1987) Isomorphous replacement- effects of error on the phase probability-distribution. *Acta Cryst A* **43**: 6-13.

Thomas M, R., Jr., William, S.R., Maria, L.C., McQuade, K.L., and Schlax, P.J. (1996) *Escherichia coli* RNA Polymerase (Es70), Promoters, and the Kinetics of the Steps of Transcription Initiation. In *Escherichia Coli and Salmonella: Cellular and Molecular Biology*. Neidhardt, F.C. (ed). Washington, D.C.: ASM Press.

Vassylyev, D.G., Sekine, S., Laptenko, O., Lee, J., Vassylyeva, M.N., Borukhov, S., and Yokoyama, S. (2002) Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A resolution. *Nature* **417**: 712-719.

Vlahovicek, K., Kajan, L., and Pongor, S. (2003) DNA analysis servers: plot.it, bend.it, model.it and IS. *Nucleic Acids Res* **31**: 3686-3687.

von Hippel, P.H., and Berg, O.G. (1986) On the specificity of DNA-protein interactions. *Proc Natl Acad Sci U S A* **83**: 1608-1612.

von Hippel, P.H., and McGhee, J.D. (1972) DNA-protein interactions. *Annu Rev Biochem* **41**: 231-300.

Vuister, G.W., Kim, S.J., Orosz, A., Marquardt, J., Wu, C., and Bax, A. (1994) Solution structure of the DNA-binding domain of Drosophila heat shock transcription factor. *Nat Struct Biol* **1**: 605-614.

Wah, D.A., Hirsch, J.A., Dorner, L.F., Schildkraut, I., and Aggarwal, A.K. (1997) Structure of the multimodular endonuclease FokI bound to DNA. *Nature* **388**: 97-100.

Wang, B.C. (1985) Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol* **115**: 90-112.

Weber, P.C. (1991) Physical principles of protein crystallization. *Adv Protein Chem* **41**: 1-36.

Wilson, K.P., Shewchuk, L.M., Brennan, R.G., Otsuka, A.J., and Matthews, B.W. (1992) Escherichia coli biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin-and DNA-binding domains. *Proc Natl Acad Sci U S A* **89**: 9257-9261.

Winter, R.B., Berg, O.G., and von Hippel, P.H. (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The *Escherichia coli* lac repressor--operator interaction: kinetic measurements and conclusions. *Biochemistry* **20**: 6961-6977.

Wolberger, C., Vershon, A.K., Liu, B., Johnson, A.D., and Pabo, C.O. (1991) Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell* **67**: 517-528.

Wong, H.C., Mao, J., Nguyen, J.T., Srinivas, S., Zhang, W., Liu, B., Li, L., Wu, D., and Zheng, J. (2000) Structural basis of the recognition of the dishevelled DEP domain in the Wnt signaling pathway. *Nat Struct Biol* **7**: 1178-1184.

Yanisch-Perron, C., Vieira, J., and Messing, J. (1985) Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* **33**: 103-119.

Youderian, P., Bouvier, S., and Susskind, M.M. (1982) Sequence determinants of promoter activity. *Cell* **30**: 843-853.

Yura, T., Tobe, T., Ito, K., and Osawa, T. (1984) Heat shock regulatory gene (htpR) of *Escherichia coli* is required for growth at high temperature but is dispensable at low temperature. *Proc Natl Acad Sci U S A* **81**: 6803-6807.

Zhang, G., Campbell, E.A., Minakhin, L., Richter, C., Severinov, K., and Darst, S.A. (1999) Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 A resolution. *Cell* **98**: 811-824.

Zhao, Z. (1999) Ph.D. dissertation: Effects of P_{lys} promoter sequence on transcription activation. The University of Tennessee Health Science Center, Memphis.

Zheng, N., Schulman, B.A., Song, L., Miller, J.J., Jeffrey, P.D., Wang, P., Chu, C., Koepp, D.M., Elledge, S.J., Pagano, M., Conaway, R.C., Conaway, J.W., Harper, J.W., and Pavletich, N.P. (2002) Structure of the Cul1-Rbx1-Skp1-F box Skp2 SCF ubiquitin ligase complex. *Nature* **416**: 703-709.

Vita

Karthik Shanmuganatham was born in Chennai, Tamilnadu, India, on November 1, 1974. He entered the Tamilnadu Veterinary and Animal Sciences University in August 1992 and in May1998 received a Bachelors in Veterinary Science degree. From the fall of 1998 to the summer of 2000 he was enrolled in the physiology master's program in the Tamilnadu Veterinary and Animal Sciences University.

In the fall of 2000 he entered the graduate program in the Department of Microbiology and Immunology, The University of Tennessee Health Science Center, Memphis from where he will receive his Ph.D under the supervision of Dr. Martha M. Howe.