STRUCTURAL STUDIES OF A PROTEIN INVOLVED IN DNA REPLICATION

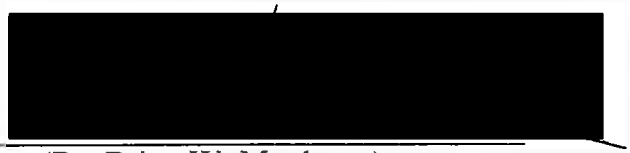
by

JEFFREY RICHARD SALERNO

A THESIS

Presented to the Department of Chemistry
and the Honors College of the University of Oregon
in partial fulfillment of the requirements for the degree of
Bachelor of Science

June 1999

APPROVED BY: 

(Dr. Brian W. Matthews)

An Abstract of the Thesis of

Jeffrey Richard Salerno for the degree of Bachelor of Science

in the Department of Chemistry to be taken June 1999

Title: STRUCTURAL STUDIES OF A PROTEIN INVOLVED IN DNA REPLICATION

Approved: _____
Dr. Brian W. Matthews

The DNA replication mechanism of Bacteriophage T4 is composed of seven proteins, the most important of which are the helicase, primase, and polymerase enzymes. The helicase enzyme unwinds double stranded DNA and the polymerase synthesizes double stranded DNA from each single stranded template. The primase enzyme, the focus of this study, synthesizes short RNA primers which allow DNA polymerase to begin synthesizing DNA daughter strands. In vitro, primase and helicase combine with single stranded DNA and ATP to form a complex called the primosome. Crystals of a truncated form of Bacteriophage T4 primase have been obtained. Four crystal forms have been characterized and a full data set has been obtained for one of the crystal forms.

ACKNOWLEDGEMENTS

The author would like the members of the thesis committee for their time commitment and patience. I would especially like to thank Dr. Ingo P. Korndörfer for his boundless patience and willingness to help me understand and communicate the concepts and ideas within this thesis. I would also like to thank Dr. Brian W. Matthews for the opportunity to do such exciting work under his mentorship and his commitment to junior researchers like myself. I would also like to thank all the members of the Matthews and Remmington Labs and Debra Jing of the von Hippel lab for providing the pDJi5 construct. This research was funded in part by the Howard Hughes Medical Institute.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Serendipity has played a key role in many great scientific discoveries throughout history. Perhaps the most widely told story of happenstance discovery is that of penicillin. In 1929 Alexander Fleming was busy trying to prove that an agent in his own saliva had antibacterial properties, a hypothesis that was later proven correct. To finance his research Fleming cured cases of syphilis among prominent London artists (Ho, 1999). He returned from vacation to find that his bacterial cultures of *staphylococci* had all been killed off by a fungal growth of *Penicillium rubrum.* Fleming was able to show that the active agent was deadly to bacteria but had no effect on human white blood cells. For his dumb luck, and the wisdom to pursue his findings, Fleming was knighted and awarded the Nobel Prize (Williams, 1969). Serendipity also played an important role in my research involving the Bacteriophage T4 primase enzyme.

Before my brief story of serendipity I would like to explain the goal of my research and some basic concepts of biology. The immediate goal of my research is to determine the three dimensional structure of primase, an enzyme involved in DNA replication. The three dimensional structures of proteins suggest much about their function and evolution. Thus protein structure is an active area of research.

Central Dogma

The central dogma of biology, proposed by Francis Crick in 1958 (Perutz, 1990), states that all life stores its genetic information as DNA or Deoxyribonucleic Acid. When organisms, or cells, or sometimes even viruses reproduce they replicate their DNA. A molecule of DNA is divided into segments called genes and the cell uses the information in each gene to synthesize a specific protein. Each protein in turn serves a function in the cell. For example, the cells in your eyes are continually replicating. Each time they replicate, their DNA is replicated as well. In addition to replicating, DNA serves as the template with which your cells synthesize a number of proteins, one of which is the protein rhodopsin. Rhodopsin is involved in the conversion of light into electrical pulses which travel down your optic nerve, allowing you to see (Voet and Voet, 1990). The process of DNA replication and the concept of DNA coding for proteins that confer function are two concepts of utmost importance to my research.

Proteins

Proteins are made up of individual units called amino acids (Figure 1), which are linked to each other to form a chain (Figure 2). The backbone of this chain is invariant, yet from each amino acid unit a side chain protrudes. This side chain, commonly denoted by R, can be any one of 20 commonly found

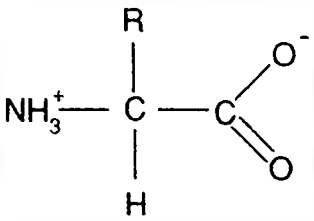


Figure 1: Structure of an amino acid.

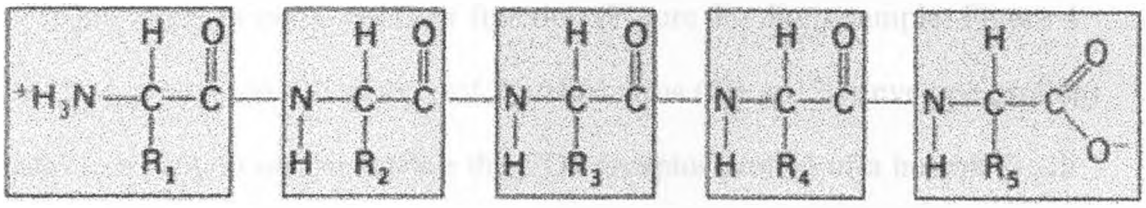


Figure 2: Primary strucure of a protein. A chain of linked amino acids.

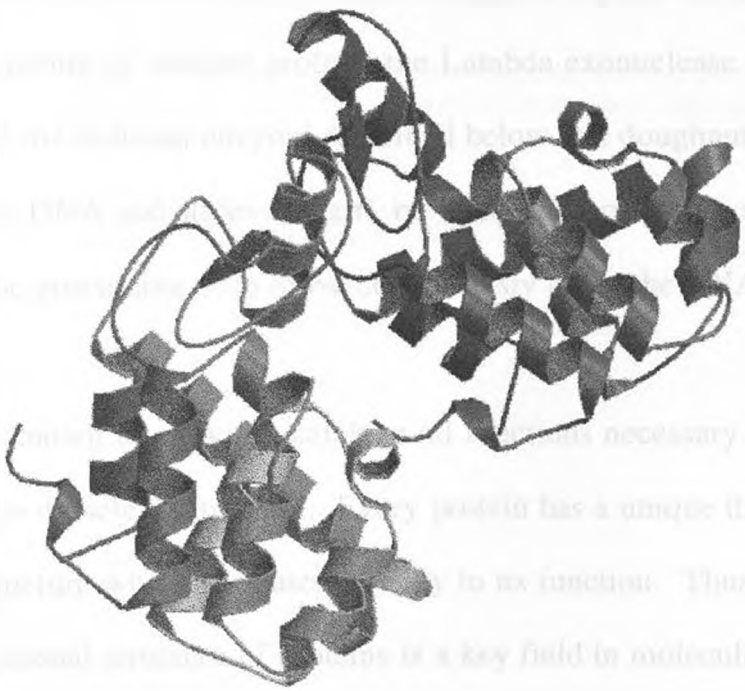


Figure 3: Three dimensional structure of a fully folded protein, T4 Lysozyme.

chemical structures. Table 1 lists the sequence of these side chains for pDJi5, the protein that I've been researching. Each letter corresponds to one amino acid.

Up to this point proteins seem fairly bland and dry. In fact, these linear chains of amino acids fold into amazingly complex and often beautiful patterns that allow them to carry out their function (Figure 3). For example, Figure 4 depicts a protein on the surface of the HIV virus (the gp120 envelope protein) bound to a protein on the surface the CD4 receptor protein of a human T-cell (Kwong et al., 1998). This interaction is the first step in the process of a cell being infected with HIV. The docking, or recognition reaction, is highly dependent on the structure of both proteins. Figure 5 depicts the three dimensional structure of another protein, the Lambda exonuclease (Kovall et al., 1997). Like the helicase enzyme described below this doughnut-shaped protein encircles DNA and slides along it, breaking its bonds. The shape allows the enzyme to be processive or to move continuously along the DNA like a motor.

Proteins known as enzymes catalyze all functions necessary for life, from food digestion to muscle contraction. Every protein has a unique three dimensional structure which is related directly to its function. Thus the study of the three dimensional structure of proteins is a key field in molecular biology.

Table I. Aligned protein sequence of pDJi5 and gp61 with regions of homology between *E. coli* , T7, T4, T3, and P4 underlined.

```
50
pDJi5 ..... ..... ..... ..... .....
gp61  MSSIPWIDNE FAYRALAHLP KFTQVNNSST FKLRFRCPVC GDSKTDQNKA

100
pDJi5 ..... ..... ..... ..... .....
gp61  RGWYYGDNNE GNIHCYNCNY HAPIGIYLKE FEPDLYREYI FEIRKEKGKS

150
pDJi5 ..... ..... ..... ..... .....
gp61  RPIEKPKELP KQPEKKIIKS LPSCVRLDKL AEDHPIIKYV KARCIPKDKW

200
pDJi5 ..... ..... ..... ..... MNEFAYRALK
gp61  KYLWFTTEWP KLVNSIAPGT YKKEISEPRL VIPIYNANGK AESFQGRALK

250
pDJi5 KDAPQKYITI KAYPEATKIY GVERVKDGDV YVLEGPIDSL FIENGIAITG
gp61  KDAPQKYITI EAYPEATKIY GVERVKDGDV YVLEGPIDSL FIENGIAITG

300
pDJi5 GQLDLEVVPF KDRRVWVLDN EPRHPDTIKR MTKLVDAGER VMFWDKSPWK
gp61  GQLDLEVVPF KDRRVWVLDN EPRHPDTIKR MTKLVDAGER VMFWDKSPWK

350
pDJi5 SKDVNDMIRK EGATPEQIME YMKNNIAQGL MAKMRLSKYAKIHHHHHH*
gp61  SKDVNDMIRK EGATPEQIME YMKNNIAQGL MAKMRLSKYAKI*......
```

Figure 4: Three dimensional structure of the HIV gp120 envelope protein (red and green) bound to the human CD4 receptor protein (blue and green).

Figure 5: Three dimensional structure of Lambda Exonuclease.

DNA Replication

Possibly one of the best examples of a molecular structure suggesting aspects of its function is Watson and Crick's double helical model of DNA. In 1953, after years spent deliberating in Cambridge's Eagle pub, James Watson and Francis Crick proposed a model for the structure of DNA (Wright, 1999). It resembles a twisted ladder, the sides composed of alternating sugar and phosphate molecules. Each "rung" is composed of a nucleic acid pair, either Cytosine (C) and Guanine (G) or Adenine (A) and Thymine (T). The entire molecule can be separated like a zipper, splitting each rung, or nucleic acid pair, in half. Because of the above pairing relationships each half carries the full set of information needed to reconstruct the entire molecule. Thus from the three dimensional structure of DNA it was possible for Watson and Crick to make conjectures about how the molecule replicates itself.

The biochemistry of viruses has been studied extensively as a model system for a number of biochemical processes. The advantage of studying biochemical processes in viruses is their simplicity. Viruses have evolved to reproduce and move to a new host quickly. Therefore their biochemistry is simple and efficient. Surprisingly, however, there is a great deal of homology between the chemical processes in the life cycle of a virus and that of bacteria, plants, and even animals.

Bacteriophage T4 is a virus that preys on bacteria. Some viruses, like HIV, use the enzymes of the host organism to replicate its DNA. However,

Bacteriophage T4 uses enzymes encoded from its own genome to replicate its DNA. This DNA replication system, made up of only a few enzymes, is well characterized and follows the general model for DNA replication that applies to all organisms.

The three most important enzymes Bacteriophage T4 uses to duplicate its genome are gp61 (primase), gp41 (helicase), and gp43 (polymerase). gp stands for gene product. For example, gp61 is the protein produced from gene number 61 on the viral genome. The genes were discovered and numbered before the protein they coded for was identified.

Each enzyme has a specific task and certain constraints on its activity. For example, helicase slides along the double stranded DNA, cleaving it like a zipper, into two pieces of single stranded DNA. Polymerase moves along the single stranded DNA and assembles nucleotides (sugar, phosphate, and one of the four bases), leaving double stranded DNA behind it. However, polymerase cannot begin synthesizing DNA on a single stranded template. It must start from a section of double stranded primer as in Figure 6. The primase enzyme, the focus of my study, synthesizes these primers.

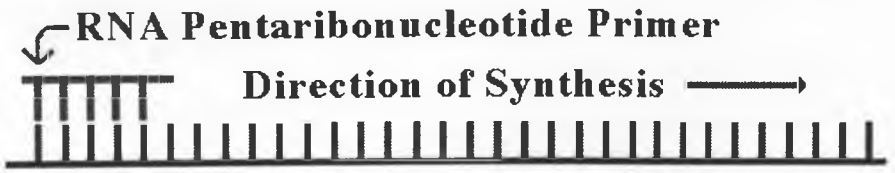


Figure 6: Primed section of single stranded DNA.

Much is known about the T4 primase enzyme and how it interacts with DNA and the other enzymes listed above. It seems that T4 helicase forms a hexamer (possibly a six membered ring) in the presence of ATP, a small molecule that enzymes consume for energy (Dong et al., 1995). This hexamer in turn binds one primase molecule and the single stranded DNA molecule, forming a primosome. This primosome is thought to remain intact throughout the process of DNA replication (Dong et al., 1996). A schematic representation of the primosome is shown in Figure 7.

The activities, or tasks performed by primase, have been characterized by various genetic and biochemical tests. Primase synthesizes pentaribonucleotides, or five membered segments of RNA, a molecule similar to DNA (Cha et al., 1986). These primers allow polymerase to begin synthesizing double stranded DNA from the primed single stranded template (Hinton et al., 1987). It is believed that these primers are efficiently synthesized by the primase enzyme only in the presence of the helicase enzyme. The primase must be a part of the primosome to be fully active (Hinton et al., 1987).

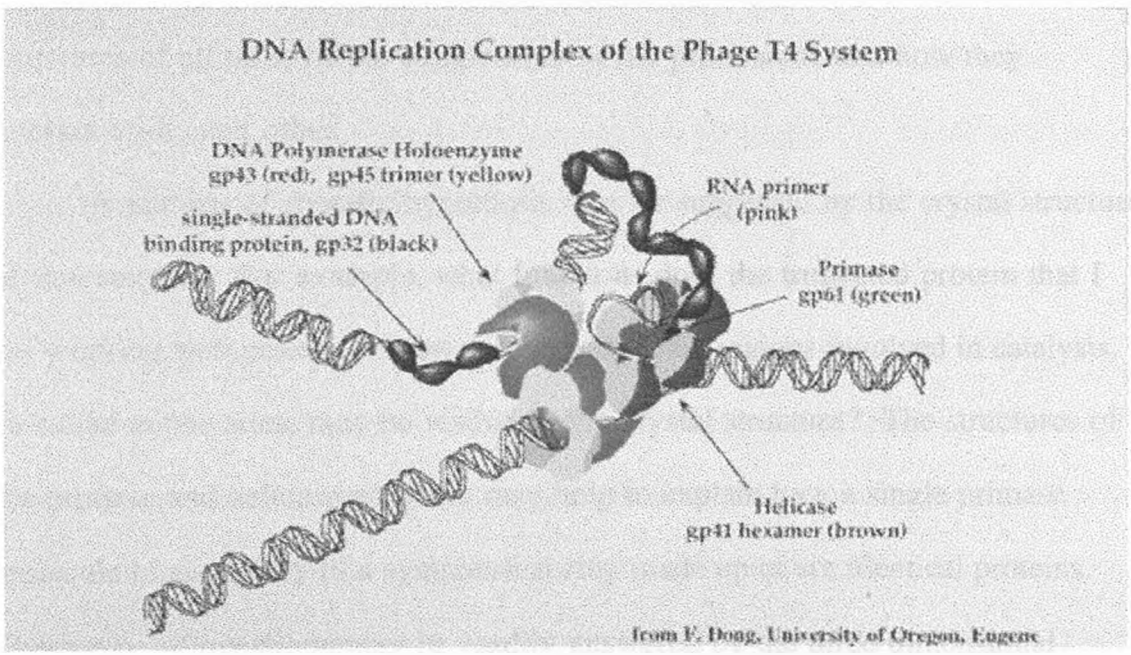


Figure 7: Schematic diagram of the primosome.

CHAPTER II: RESEARCH QUESTION

The immediate goal of this research is to determine the three dimensional structure of a truncated version of the primase enzyme. This data may be used to determine a hypothetical mechanism of action for the primase enzyme. Also, this fragment of the primase enzyme is a component of the larger enzyme complex, the primosome. The long term goal of this research is to elucidate the structures of all three of the components of the primosome and how they interact with each other.

A number of specific hypothesis may be suggested by the crystal structure of this enzyme. For example, what functions does the truncated protein that I am working with possess? More importantly what regions involved in catalysis, so called active sites, may be visible in the crystal structure? The structures of the primase and helicase enzymes may help to explain how a single primase molecule binds tightly to a symmetrical ring made up of six identical proteins. These type of questions may be readily answered by the three dimensional structure of the protein or may require further structural study on the complete enzyme or a complex of helicase and primase.

The motivation for studying primase and DNA replication systems is more esoteric than practical. There is only one known disease that involves the primase enzyme and it is implausible that a cure could be found once the structure is known. It is unlikely that the structure of primase would lead to a genetically engineered organism, like the "flavor saver" tomato, that would be

beneficial to industry. Rational drug design and protein engineering using the information contained in a protein structure are two common motivations for studying protein structure. However, in this case there is little immediate practical impetus for studying this enzyme.

Every cell of every living thing uses primase enzymes to replicate itself. Even viruses, which are so simple that few consider them as a form of life, use primase to replicate their genetic information. Primase is one of the major enzymes that participate directly in the propagation of genetic information and in the continuation of life. Therefore its study is of fundamental scientific interest.

In addition to basic interest in concepts like the propagation of life, the three dimensional structure of this protein will contribute to the overall body of knowledge on protein folding. Scientists still don't understand fully what forces cause proteins to fold into and retain their complex three-dimensional structures. There have been many attempts to predict, from the primary structure of a protein, its final three dimensional conformation. Not surprisingly these algorithms were quite successful at predicting the structures of proteins for which the structure was already known but were abysmal at predicting novel structures. The need for an accurate protein folding model is even more apparent with the impending determination of the complete sequence of the human genome along with the genomes of bacteria, yeast, *Drosophila*, and other plants and animals. A folding model could be used to predict structures

and possibly functions of proteins from genes for which only the sequence is known. Therefore nearly every novel protein structure contributes to our understanding of the molecular dynamics that govern protein folding.

CHAPTER III: pDJI5 PROTEIN

The gene for the primase enzyme was cloned, or spliced, into a molecule of DNA that can be read by *E. coli* bacteria (Hinton et al., 1985). This DNA can be manipulated using various genetic techniques, including Polymerase Chain Reaction or PCR. The most commonly known use for PCR is the amplification of small segments of DNA from blood stains found at crime scenes. Through PCR it is possible to create multiple copies of a segment of DNA selected from a large molecule. Primers are added, much like those synthesized by primase, to define the region one wants to amplify. For example, if this paper were a DNA sequence one could design primers that enclosed this sentence and replicate it, much like a computer's cut and paste function.

Molecular biologists use PCR to introduce truncations, additions or mutations into proteins. For example, the sequence of the primase protein as found in nature, its native form, is listed in Table 1 as gp61. The sequence of the protein I've been working with is labeled pDJi5. The sequences are aligned to show their similarities. Note that the first 190 amino acids are missing, or truncated, from pDJi5. Also note that the first few amino acids of pDJi5 are different, or mutated, from the native sequence. There is also an addition of six histidine or H amino acids at the end of the pDJi5 protein.

These modifications were all made using PCR and are the beginning of my short story of serendipity. Debra Jing of the von Hippel lab was using PCR to make a 6 amino acid truncation from the beginning of the native protein. PCR is

a finicky reaction and often the primers bind to segments that were unexpected and thus a few different sized pieces of DNA are amplified. This was the case in Debra's PCR reaction that produced the expected reaction product of length 1050 base pairs and an unexpected reaction product with a length of 420 base pairs. The shorter, unexpected segment was the gene for my pDJi5 protein. So, the subsequent purification and crystallization of this protein all followed an unwanted reaction product.

Relevance of a Truncation

Work on a truncated protein may seem pointless, but in fact it can be enlightening. Molecular biologists often study truncated proteins to isolate specific functions of an enzyme. Because catalytic functions are often grouped within the protein a truncated protein often exhibits some of the catalytic functions of the full length protein. For example, the so called Klenow Fragment of the polymerase enzyme still synthesizes DNA but no longer has a an error-correcting function that is present in the native enzyme.

Structural biologists often study truncated proteins because they are easier to work with than the full length protein. The smaller truncated proteins tend to crystallize more readily, an integral step to determining the structure of the protein. Luckily there have not been any cases where the truncated protein differs in structure from the corresponding section of the full length protein.

Thus the structure of a truncated protein often provides a snapshot of a section of the full length protein.

A number of observations suggested that a truncated version of the primase protein might be more amenable to structural study. Attempts were made to crystallize the full length primase protein with no success. It was observed that the full length protein degrades into a number of distinct fragments when stored at 4°C for prolonged periods. Digestion with an enzyme, trypsin, that cleaves the protein backbone also yields distinct products. These two facts suggest that the protein is composed of a number of tightly folded domains with linker regions that are exposed to proteolysis. A number of different constructs, or modified genes, were made, using the PCR reaction. All of these constructs have different truncations and histadine tag conformations (see Figure 8). A similar approach proved successful in determining the structure of the C-terminal fragment of the primase enzyme from the bacterium *Bacillus stearothemophilus* (Pan et al., 1999).

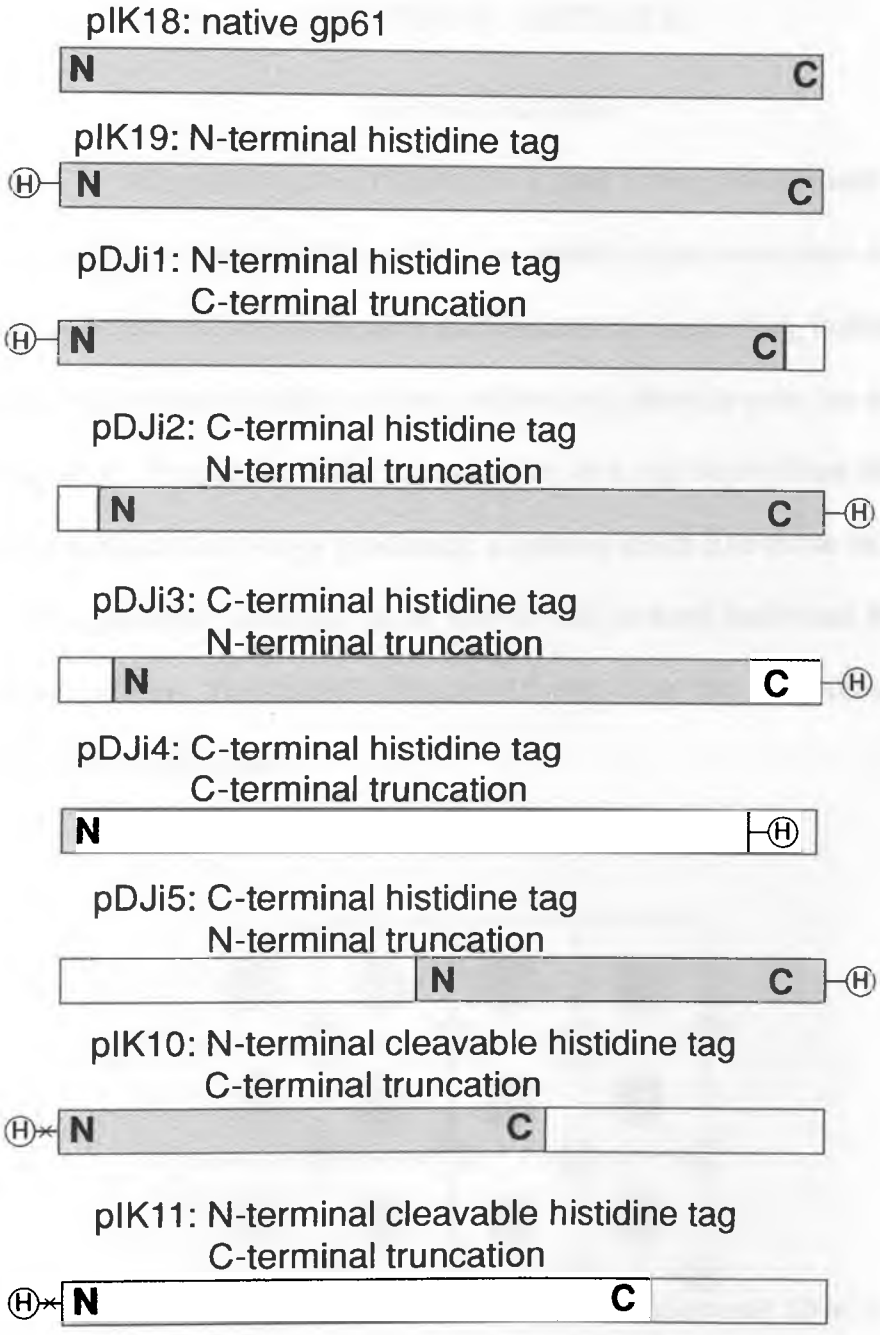


Figure 8: Primase constructs.

CHAPTER IV: METHODS

Crystallography

The protein structures in Figures 3, 4, and 5 were determined using a technique called x-ray crystallography. A crystal of any substance is composed of layers in which all the molecules are oriented in a repeating fashion. For example, in a crystal of table salt the sodium and chloride ions are arranged in a repeating array (Figure 9). When exposed to an x-ray beam these layers of molecules diffract the x-rays producing a pattern much like those in Figures 10 and 11. It is possible, although in no way trivial, to work backward from diffraction patterns like those in Figures 10 and 11 to the three dimensional structure of the molecule.

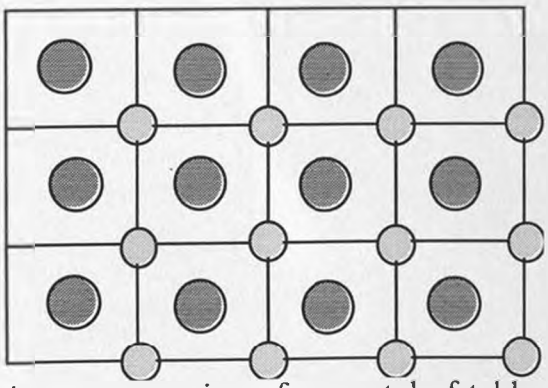


Figure 9: Schematic representation of a crystal of table salt (NaCl).

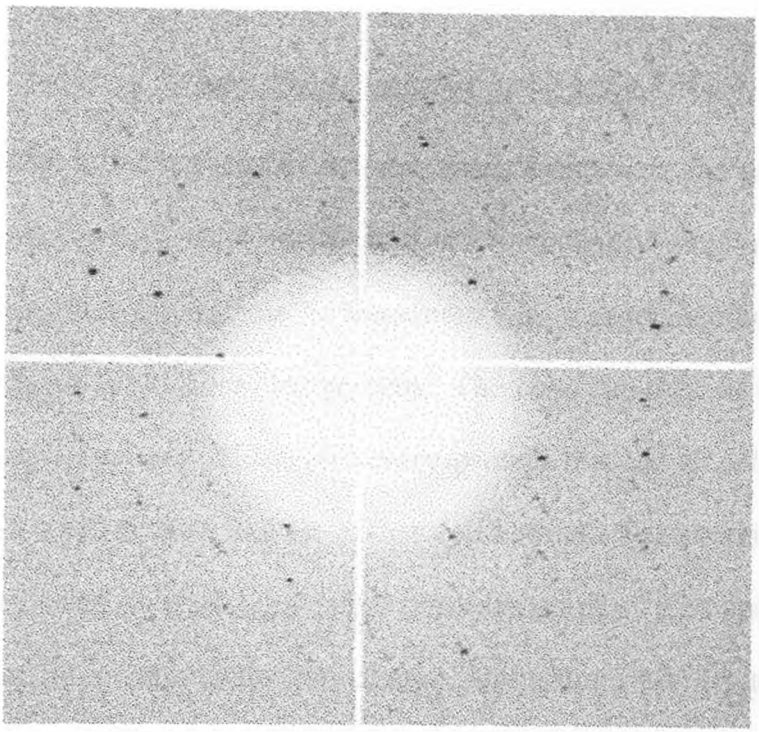


Figure 10: Diffraction pattern
from Crystal Form I.

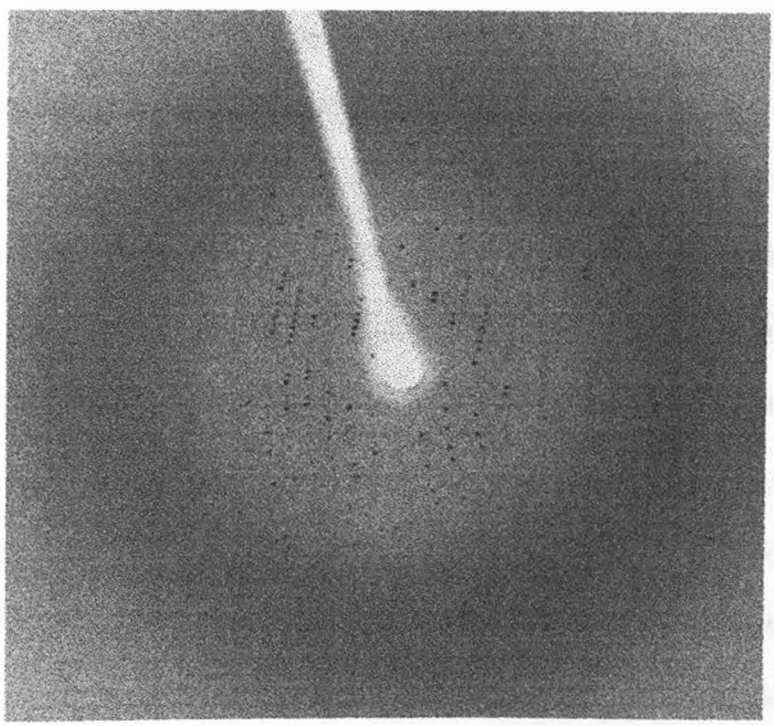


Figure 11: Diffraction Pattern
from Crystal Form II.

Expression and Purification

To obtain a protein crystal one must first synthesize and purify the protein. *E. coli* bacteria serve as the production plants for most molecular biology applications. *E. coli* can be implanted with a gene encoding the desired protein and they will produce that protein. The cells can then be broken apart and the protein extracted. However, every protein that the *E. coli* normally synthesizes in order to live will be in that cell extract. The protein of interest can then be isolated by a process called chromatography. The desired protein can be isolated according to its size, charge, and affinity (ability to bind to something). The pDJi5 protein has been cloned with six histidine (H) amino acids at its C-terminus. Histadines have an affinity for, or ability to bind nickel ions. Thus pDJi5 is applied to a column containing a nickel matrix. pDJi5 binds to this matrix while most of the other proteins do not.

The next step in purification involves a column that allows larger molecules to flow through it quickly but slows the motion of smaller molecules. This seems counterintuitive but is due to an ingenious column design. The column is packed with microscopic beads containing small holes and passageways. Small molecules enter these passageways and take a meandering route through the column matrix. Larger molecules, like proteins, are too large to fit into the passageways and thus take a more direct route through the column. Through this method of size exclusion chromatography it is possible to purify pDJi5 to homogeneity.

Technical Methodology

The conditions of expression and purification often have an effect on the quantity, purity, and chemical state of the resulting protein. Often, in the case of a novel protein, a complicated purification scheme must developed through a series of experiments. The following methods section describes the cloning, expression, and purification of the pDJi5 protein in great technical detail.

The above mentioned PCR product was obtained from Debra Jing of the von Hippel lab and subcloned into a pET11a expression vector (Novagen). The construct was electrotransformed into the cell strain BL21[DE3] and inoculated in 50mls LB medium with 100μg/ml ampicillin. This culture was incubated overnight at 37°C with vigorous shaking (~300rpms). The culture was again inoculated into 4.2L of LB medium with 100μg/ml ampicillin and incubated at 37°C with stirring (~900rpms) and aeration (10L/min) until the OD_{600} was ~0.8. Expression was then induced by the addition of 0.1mM IPTG. The cells were incubated for another 2 hours before harvest.

From here on the cell lysate was kept at 4°C. The cell slurry was spun down at 30,000 g and the cell pellet resuspended in ~200mLs sonication buffer (20mM Tris/HCl, 0.3M NaCl, 1mM BME, pH 8.0). The resultant suspension was sonicated using a Fisher Scientific Sonic Dismembrator 500 for 10 minutes at power 7 (0.3sec on, 0.7sec off). The solution was then spun down twice for 20 minutes at 30,000g and the supernate recovered.

Glycerol to 10% and imidazole to 5mM were added to the cell lysate. This solution was applied to a column containing Biorad Ni-NTA column material. The column was washed with ~400mLs of wash buffer (10% glycerol, 20mM Tris/HCl, 0.3M NaCl, 1mM BME, pH 8.0) containing 5mM imidazole. The column was then eluted with an imidazole gradient (0 to 250mM in Wash Buffer) and fractions were collected. The fractions containing protein were concentrated using a Centricon (Amicon) to between 2 and 5mg/mL. 2mL aliquots were applied to a Hi-Load Superdex 75 sizing column (Pharmacia) using a Pharmacia FPLC pump and control unit. Fractions containing the desired protein were combined and concentrated using a Centricon to the concentration required for crystallization. Protein purity was monitored using a 20% SDS PAGE phastgel (Pharmacia).

Crystallization

Crystallizing a large molecule is perhaps one of the most frustrating endeavors in molecular biology. There are a number of unpredictable factors that operate on a molecular scale that influence crystallization. A slightly impure protein solution, dust, or even a contamination by small molecules can disrupt or facilitate crystal formation. As a result there is a mystical air and a sizable amount of superstition and even voodoo involved in crystallization of proteins. Some people wear lucky socks and some contend that the phase of the

moon has an effect but it goes without saying that simply crystallizing a protein is a major step toward determining its structure.

The most common method for crystal growth is called the hanging drop vapor diffusion method. A small vial or well is filled with a concentrated well solution containing a buffer to maintain pH, a precipitant molecule, and sometimes a salt to help solubilize the protein. A drop of this solution is placed on a glass cover slip and a drop of protein solution is mixed with it. The cover slip is then inverted and placed over the well creating the hanging drop (see Figure 12). Since the precipitant concentration is higher in the well than in the drop the force of diffusion draws water from the drop back into the well. Concentrations in the drop increase and molecules tend to aggregate, hopefully in crystalline form.

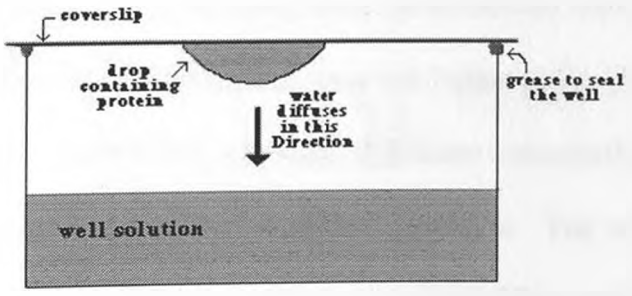


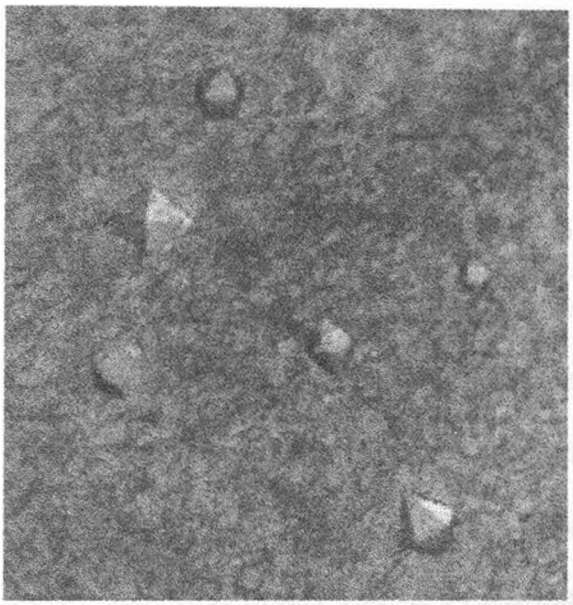
Figure 12: Hanging drop vapor diffusion crystallization setup.

Obtaining crystals of high enough quality to collect x-ray diffraction data is often an arduous process. Crystallization screens must be used to test the

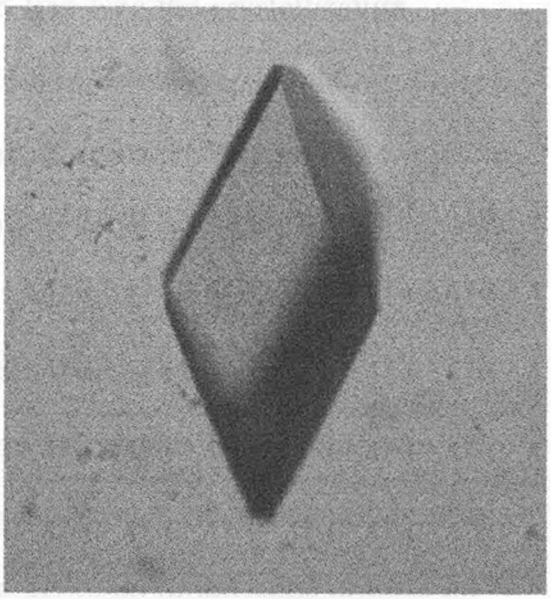
protein molecule's behavior in a number of different crystallization solutions. Often the correct combination must be found from dozens of buffers, dozens of precipitants, and nearly endless combinations of additives and salts. In addition to the identity, the concentrations of each of these species can influence crystallization. A method called sparse matrix crystallization screening was used to determine possible crystallization conditions. Sparse matrix screening involves testing a number of widely varying conditions in hopes of identifying one or more that induce crystallization. Often these conditions produce crystals that are too poor to collect x-ray diffraction on but serve as a lead to identifying crystallization conditions that will.

Four different crystal forms of pDJi5 have been identified and characterized (Figure 13). All of the crystal forms were grown using the hanging drop vapor diffusion method at a temperature of 4°C. All of the crystals formed within three to five days. Each hanging drop contained an initial protein concentration between 10 and 20 mg/mL in wash buffer (10% glycerol, 20mM Tris/HCl, 0.3M NaCl, 1mM BME, pH 8.0). All drops contained 5μL protein solution, 5μL well solution, and 1μL additive if present. The well solutions contained approximately 0.1M Mes pH6.0, 30% PEG 200, and 5% PEG 3.4K.

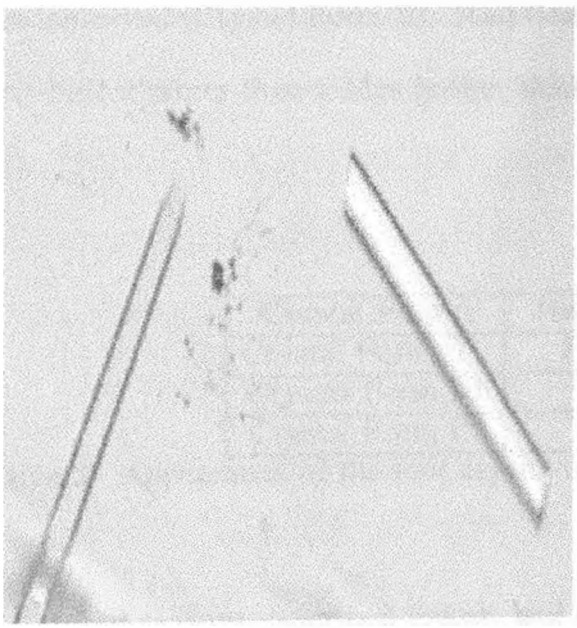
Using sparse matrix crystallization screening an initial crystallization condition was identified that produced thin threadlike needles. These crystals were not of sufficient quality to collect x-ray diffraction data but indicated that it was possible to crystallize this molecule.



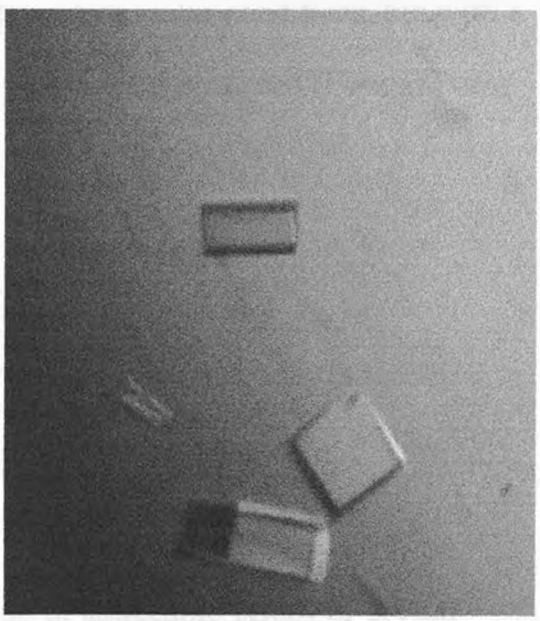
Crystal Form I



Crystal Form II



Crystal Form III



Crystal Form IV

Figure 13: Four crystal forms of pDJi5

A number of attempts were made to fine tune the crystallization conditions by varying pH, protein concentration, additives, and the concentration and size of the various PEGs. A commercially available additive screen (Additive Screen I and II, Hampton Research) was used to identify small molecules that may induce crystallization. A number of the additives produced hexagonal bipyramidal crystals, later dubbed Crystal Form I.

In addition to the hexagonal bipyramidal crystals larger, hexagonal needle shaped crystals were observed (Crystal Form III). These needle shaped crystals appeared under similar conditions and often in the same drop as the hexagonal bipyrmidal crystals. Under similar conditions another crystal form was observed (Crystal Form II). And finally under similar conditions, but with a Tris buffer rather than a Mes buffer, thick plates were observed (Crystal Form IV).

Crystal Form I	Hexagonal Bipyramidal
Crystal Form II	Hexagonal Needles
Crystal Form III	Trigonal
Crystal Form IV	Thick Plates

Table 2: Appearance of the four crystal forms.

For reasons described below, and due to noticeable issues of crystal quality, Crystal Form II is the most promising of the four. However, this crystal form is also difficult to reproduce. A number of different techniques have been used in an attempt to reproduce Crystal Form II. For example, the nucleation of

crystals is often a barrier to crystallization. An analogous process is the formation of freezing rain. Freezing rain is super-cooled water that is in fact cold enough to freeze but has nothing to crystallize onto. Thus it falls in the liquid form and freezes, or crystallizes onto whatever it hits. In a similar manner the nucleation of protein crystals can often be initiated by a seeding process, similar to cloud seeding.

Protein crystals can be seeded in three ways. The first involves striking a hair through a hanging drop of a crystallization setup. Hair is composed of ordered proteins, much like a protein crystal. Particles transferred from the hair into the drop may provide a nucleation site. Protein crystals can also be used to seed the growth of other crystals. A crystal can be broken into microscopic pieces in a solution and added to the hanging drop. This technique, known as micro-seeding, was used in an attempt to reproduce Crystal Form II. In addition crystals that have stopped growing can be placed in a new hanging drop, a technique known as macro-seeding.

For reasons discussed below in the phasing section Crystal Form II must be reproducible in order to solve the structure of the pDJi5 protein. So far all attempts at fine screening and crystal seeding have proved unsuccessful.

Data Collection

The general schematic for data collections is diagrammed in Figure 14. A monochromatic, or single wavelength beam of x-rays is directed through the protein crystal. The diffracted x-rays strike a film cassette or electronic area detector. The crystal is then rotated and another image recorded. Figures 10 and 11 show diffraction patterns from Crystal Forms I and II.

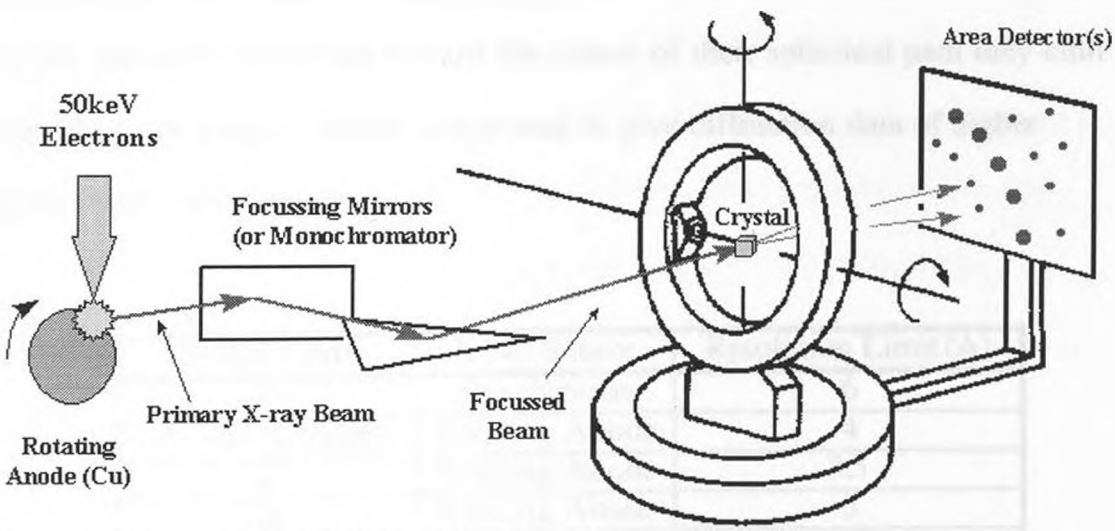


Figure 14: X-ray data collection apparatus.

The maximum angle that x-rays are diffracted to is analogous to the resolution (measured in Angstrom units) with which the final structure can be determined. The resolution of x-ray data is similar to that of a printer or microscope. To be of any use diffraction data must be near to or below 3Å. The resolution of the crystal depends on a number of factors including the quality of

the crystal, its size, and the intensity of the x-ray beam. Table 3 shows the diffraction limits and x-ray sources for the four crystal forms.

Keep in mind that the x-ray source and its intensity affect the resolution limit. The conventional means of generating x-rays is to fire an electron beam at a rotating copper anode (Figure 14). This produces x-rays of wavelength 1.54Å. In recent years protein crystallographers have utilized another, more powerful x-ray source, the synchrotron, to collect diffraction data. A synchrotron is a type of particle accelerator that propels atomic particles in a circular fashion. As the particles accelerate toward the center of their spherical path they emit high intensity x-rays. These x-rays tend to give diffraction data of higher quality and resolution.

Crystal Form	X-ray Source	Resolution Limit (Å)
I	Synchrotron	6
I (larger crystals)	Rotating Anode	4
II	Rotating Anode	3.5
III	Rotating Anode	5
IV	Rotating Anode	6

Table 3: Diffraction Limits of the four crystal forms

A crystallographic data set consists of a measurement of the positions and intensities of a finite number of reflections. These reflections lie in an indexed plane, which has the coordinates hkl. The outer limits of the coordinates measured in a data set is analogous to the diffraction limit. The

completeness of a data set is a measure of the portion of reflections within the stated resolution limit for which intensity data has been measured.

A number of data sets have been measure for crystal forms I and II. The data set for crystal form I was collected at the Advanced Light Source at Lawrence-Berkeley Labs. This variable wavelength source was set to a constant wavelength of 1.0Å. Fourteen exposures were taken, each one degree oscillation of the crystal, using a quantum4 area detector. Data from three different crystals of Crystal Form I was collected in house using a Cu-Kα source and an R-axis4 area detector. Crystal Forms III and IV were also characterized in house. One image with diffraction to 4Å resolution was taken in house on a larger crystal of Crystal Form II. This indicates a large improvement in the diffraction quality of this crystal form since the data collection at the synchrotron source.

CHAPTER V: RESULTS

Characterization of Crystal Forms

Unit Cell

A number of different sets of information must be obtained to solve the crystal structure of a protein. First, the dimensions of the unit cell must be determined. The unit cell is defined as the smallest volume of the crystal that can be used to reproduce the entire crystal by translation. Figure 15 depicts a popular M.C. Escher tessellation (Ernst, 1994) with a unit cell marked. Notice that you can slide, or translate, the unit cell to the right, left, up, or down and it will allow you to construct the entire pattern. The unit cell of the salt crystal in Figure 9 is defined by a black square. Note that this unit cell has one complete salt molecule, composed of a chloride and sodium ion, in each unit cell. The axes of the crystallographic unit cell are labeled a, b, and c and the angles between them α,β,γ. (Figure 16).

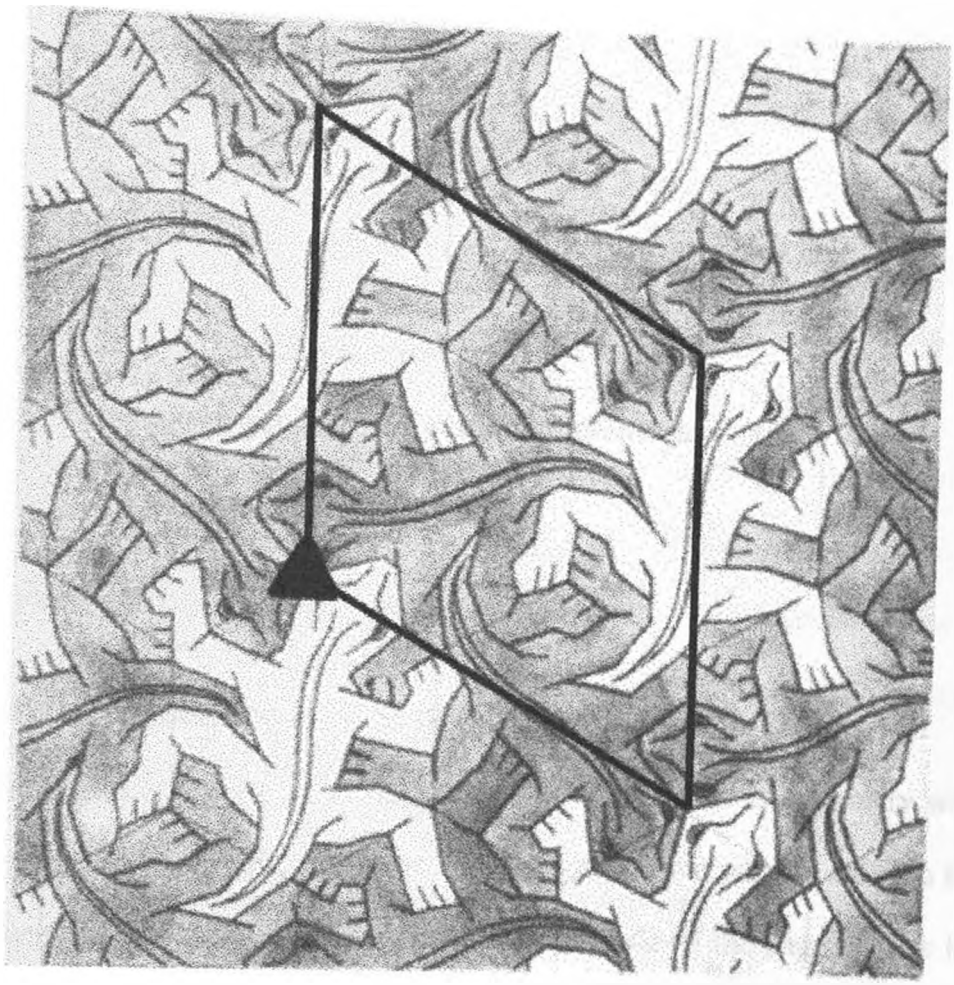


Figure 15: MC Escher's "Reptiles" with a unit cell and three-fold rotation axis marked.

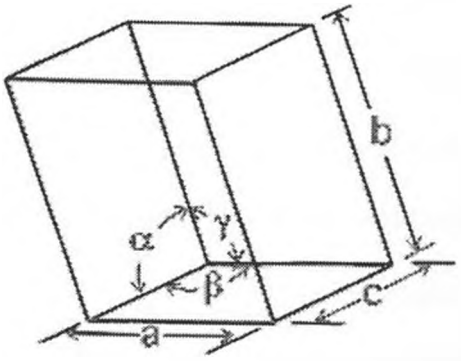


Figure 16: Convention for labeling unit cell dimensions.

Space Group

The orientation and position of the molecules within the unit cell determines what is known as the space group. Often there is symmetry within the unit cell. For example, in Figure 15 the lizards have what is called a three-fold symmetry, which runs perpendicular to the plane of the page. If one lizard were to be rotated 120° about the triangle in Figure 15 it would overlay the next lizard. Another 120° rotation would position it onto the third lizard. Thus if you had the structure of one lizard and knew that there was a three-fold axis in the unit cell you could easily construct the other two lizards.

In the unit cell of a protein crystal these kinds of symmetry elements are also applicable, yet they are often in three dimensions and involve more complicated symmetrical operations. The space groups of crystals are general written in short-hand with an upper case letter (P in this case) followed by three

numbers. The numbers each represent a symmetry element, which in this case is always a rotation axis. A 2 represents a two-fold rotation axis, a 3 represents a three-fold rotation axis and so on. There are three places available for symmetry elements, one in each of the three dimension. Often, in simple space groups there are less than three symmetry elements and a one is used as a place holder to indicate that there is no symmetry element in that dimension.

Often the rotation axis will have a subscript attached to it, indicating that the rotation axis is also a screw axis. Figure 17 depicts the difference between a two-fold axis of rotation and a two-fold screw axis. Note that the screw axis is equivalent to the combination of two symmetry operators: a rotation followed by a translation.

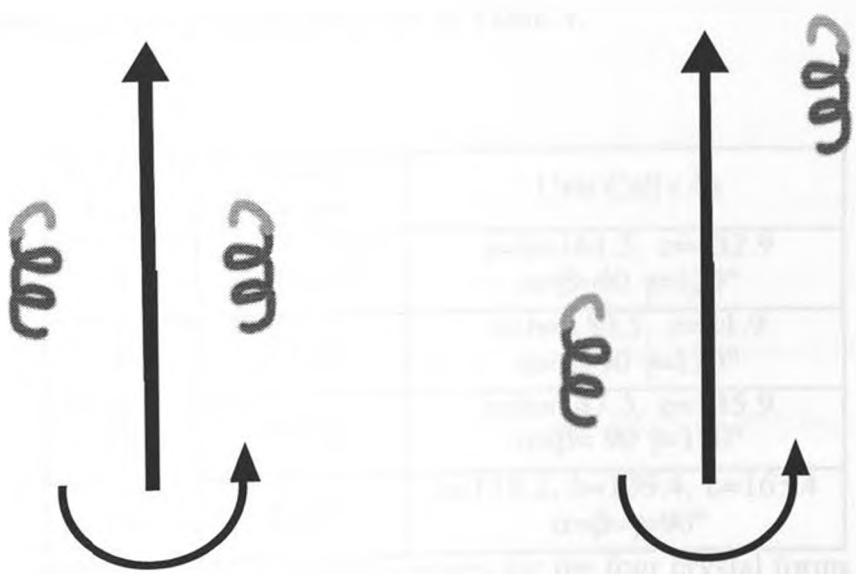


Figure 17: Two-fold rotation axis (left) and two-fold screw axis (right).

Data Processing

The unit cell and space group can both be determined from the diffraction pattern of a crystal. Notice that the diffraction spots in Figures 10 and 11 lie more or less along a grid. The unit cell length is inversely proportional to the spacing of these grid lines. The angles relating these grid line are directly related to the angles relating the edges of the unit cell.

The space group, like the unit cell, can be determined from the diffraction pattern of a crystal. The conformation of spots in the diffraction pattern indicates the symmetry of the crystal.

Diffraction images of each crystal form were indexed using the program Denzo and spacc groups were assigned from the resulting list of likely candidates (CCP4, 1994). Unit cell parameters were refined using Scalepack (CCP4, 1994) and are presented below in Table 4.

Crystal Form	Space Group	Unit Cell (Å)
I	$P6_{(1/5)}22$	a=b=161.3, c=432.9 $\alpha=\beta=90$ $\gamma=120°$
II	$P3_221$	a=b=150.5, c=81.9 $\alpha=\beta=90$ $\gamma=120°$
III	P6(?)	a=b=147.3, c=235.9 $\alpha=\beta=90$ $\gamma=120°$
IV	P222	a=118.2, b=139.4, c=163.4 $\alpha=\beta=\gamma=90°$

Table 4: Cell dimensions and space groups for the four crystal forms.

Data collected from Crystal Form I and II were indexed using Denzo and scaled using Scalepack. Data from three different crystals of Crystal Form II were scaled together using Scalepack (CCP4, 1994). Values for completeness and R_{merge}, a measure of the quality of data, are presented Table 5.

Crystal Form	Completeness (%)	R_{merge}
I	54.4 (30.0 - 6.0Å) 48.8 (7.0 - 6.0Å)	0.19
II	81.9 (30.0 - 3.5Å) 78.8 (3.6 - 3.5Å)	0.12

Table 5: Data collection statistics for Crystal Forms I and II.

These two data sets, one for Crystal Form I and one for Crystal Form II, are composed of lists of intensities and positions for a large number of indexed reflections. These two data sets were used for all subsequent calculations.

It is interesting to note that the dimensions of the unit cells of Crystal Forms I, II, and IV are geometrically related. For example, all of the α and β angles are 90°. Even more interesting is the fact that in each crystal form one of the cell edges is nearly an integer multiple of 80Å. For example, the cell edges a and b of Crystal Form II are equal to twice this multiple of 80Å. This may suggest that there are similar packing arrangements underlying each of the crystal forms.

Matthews Coefficient

Asymmetric Unit

The unit upon which the above symmetry elements operate often consists of more than one molecule. This unit is called the asymmetric unit and is defined as the smallest volume of the unit cell which can be used to recreate the entire unit cell by symmetry operations. The asymmetric unit may contain a fraction of molecules if each molecule has its own symmetry or it may contain a large number of molecules. The number of asymmetric units per unit cell is defined by the space group. However, the number of molecules per asymmetric unit must be determined in another method.

Matthews Coefficient Calculation

To estimate the number of molecules per asymmetric unit a simple calculation can be done. The molecular weight of the protein can be determined from its primary structure as inferred from the genetic sequence. The volume of the asymmetric unit in $Å^3$ can be determined from the volume of the unit cell and the identity of the space group. The coefficient V_m, also known as the Matthews parameter, is defined as the ratio between the volume of a single molecule and the molecular weight of the protein. Intuitively V_m represents the crystal density in volume of unit cell per unit of protein molecular weight. The most commonly observed value for V_m for protein crystals is near $2.15 Å^3$/dalton. From the same data the solvent content, in percent, can be calculated by assuming a

protein density of 1.35g/mL. A typical value for a solvent content of a protein crystal is around 43% (Matthews, 1968). The table below displays calculated V_m values and solvent contents for two different numbers of molecules per asymmetric unit.

crystal form	Number of molecules per asymmetric unit	Vm	Solvent content (%)
I	16	2.7	54
	24	1.8	31
II	6	2.5	50
	8	1.9	34
III	14	2.9	57
	22	1.8	32
IV	12	2.7	55
	16	1.1	40

Table 6: Matthews Coefficient calculations for the four crystal forms.

The asymmetric unit is the space in which crystallographers operate. Once the contents of the asymmetric unit are known the entire unit cell may be reconstructed by applying the symmetry elements. Thus the complexities involved in determining the structure escalate as the size and number of molecules in the asymmetric unit increases. For this and other reasons outlined above Crystal Form II is the most viable for structural determination.

Phase Determination

It is impossible to calculate the position of every atom in a protein molecule directly from the diffraction pattern. First, the phase problem must be overcome as discussed below. Second, the position of every atom in the protein molecule must be deduced from the calculated electron density. For example, in Figure 18 the net-like map represents an electron density calculated from x-ray diffraction data. A model of the protein backbone and the conformation of its side chains has been deduced from the electron density map.

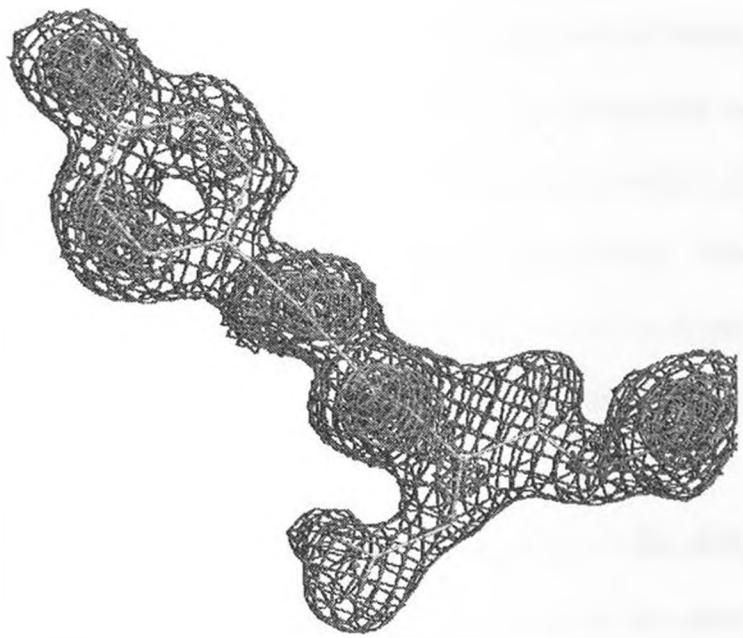


Figure 18: Electron density map with a model of the protein backbone and its side chains.

The following equation is used to calculated a three-dimensional map of the electron density, symbolized by the Greek letter ρ. The electron density, ρ at a coordinate in the unit cell, xyz is expressed as

$$\rho_{(x,y,z)} = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}\left|F_{(h,k,l)}\right|\exp\left[-2\pi i(hx+ky+lz-\alpha_{(h,k,l)})\right]$$

The upper case V represents the volume of the unit cell, which has been determined (18.6e5 $Å^3$ for crystal for II). The remainder of the expression is a sum over all the indices hkl, the integer values that crystallographers use to identify diffraction spots. So, for each diffraction spot with coordinates hkl the following operation is performed. The F value, which is the square root of the intensity of each spot, is multiplied by the exponential factor of a rather complicated function. This function involves the imaginary quantity i, which is the square-root of -1. It also involves a sum of the products of each hkl coordinate multiplied by the corresponding xyz variable. This function is a Fourier transformation. These mathematical functions of sins and cosines can be used to recreate any periodic function such as the electron density in a unit cell.

The final variable, α, represents the phase of the diffracted x-rays that form each spot. It is impossible to directly measure the phase of diffracted x-rays so a number of techniques are employed by crystallographers to solve what is called the phase problem.

The two most popular techniques for determining phases are Multiple Anomalous Dispersion (MAD) and Multiple Isomorphous Replacement (MIR). MIR, or Multiple Isomorphous Replacement is the most common form of phasing. It involves collecting a number of sets of data from heavy atom derivatives. These derivatives are native crystals with a heavy atom bound in a specific and consistent spot within the unit cell. Heavy atoms, like argon, gold, or mercury have a larger number of electrons and therefore diffract x-rays with greater intensity than the standard biological atoms carbon, oxygen, and sulfur. The deviations in the intensities of the spots between the different heavy atom data sets and the native data set can be used to determine the structure factor α for each reflection.

MAD phasing, which has only recently been developed, involves the use of just one heavy atom derivative. A number of data sets are collected with different wavelength x-rays. The differences between these data sets can be used to determine the structure factors for each reflection. MAD phasing is a recent development made possible by the high intensity and tunable wavelengths of the synchrotron x-ray sources.

The standard method of obtaining a heavy atom derivative involves a technique known as heavy atom soaks. Protein crystals, as described above, are usually around 50% solvent. This high solvent content allows medium sized molecules like ATP, heavy atom complexes, and even small pieces of DNA to diffuse into the crystal. Simply soaking a protein crystal in a dilute heavy atom

solution is the standard method used to obtain a heavy atom derivative. This method of obtaining a heavy atom has not been tried due to the irreproducibility problem encountered with Crystal Form II.

Attempts have been made to co-crystallize a heavy atom derivative of the pDJi5 protein. One of the additives that induces crystallization is barium chloride. Barium is a heavy atom which if bound at a distinct site within the unit cell, could be used as a heavy atom derivative. A number of attempts have been made to produce diffraction quality crystals with this barium additive with limited success.

A relatively new method of obtaining a heavy atom derivative was also attempted with pDJi5. One of the twenty standard amino acids is methionine, which has an internal sulfur atom. A synthetic analog of methionine called seleno-methionine has been synthesized with the heavy atom selenium at the position sulfur usually occupies. It is possible to trick bacterial cells into incorporating this methionine analog into the proteins that they produce. The seleno-methionine protein then folds and can be purified and crystallized just like the native protein. A number of attempts have been made to crystallize a seleno-methionine derivative of the protein with limited success.

Patterson Function

Structural Information Without Phasing

A number of calculations can be made in the absence of structure factors. One is called the Patterson function, which is generally calculated in uvw coordinates. Again the indices of each reflection are specified with the hkl index.

$$P_{(u,v,w)} = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|F_{hkl}|^2 \cos[2\pi(hu+kv+lw)]$$

Intuitively the Patterson map is a plot of all interatomic vectors within the unit cell. The vectors retain their direction and magnitude but are overlaid onto the origin of the Patterson map. For example, a simple tri-atomic molecule and its corresponding Patterson vectors are shown in Figure 19.

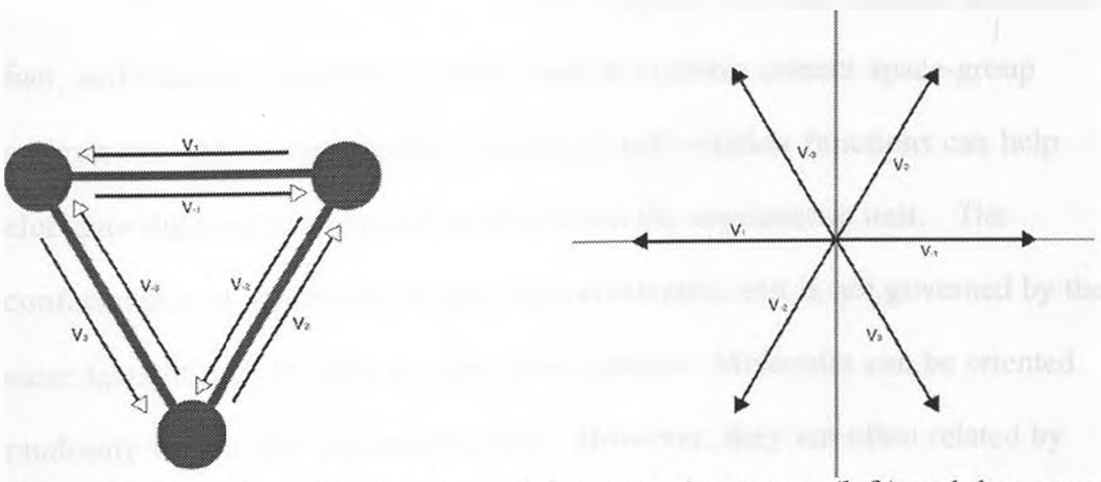


Figure 19: Triatomic molecule with interatomic vectors (left) and those same vectors in Patterson space (right).

Analysis of Patterson Functions

It is clear that the Patterson function for a protein with a large number of asymmetric units per unit cell, molecules per asymmetric unit, and atoms per molecule would be an indecipherable mess. Therefore calculations involving Patterson functions often involve comparing two Patterson functions to determine similarities or differences. For example, a self-rotation function calculation is analogous to isolating a volume of the Patterson map and rotating it about a number of different axes. If the rotated area happens to overlay perfectly with the portion of unit cell it is rotated upon, a peak is seen. For example a two-fold axis relates the contents of the unit cell in Crystal Form II. A self-rotation function would produce a peak corresponding to this axis of rotation.

Crystallographic axes are clearly apparent on self-rotation functions. In fact, self-rotation functions can be used to confirm correct space-group assignment. More importantly, however, self-rotation functions can help elucidate the conformation of molecules in the asymmetric unit. The conformation of molecules within the asymmetric unit is not governed by the same laws of packing that govern space groups. Molecules can be oriented randomly within the asymmetric unit. However, they are often related by symmetry elements that do not extend beyond the asymmetric unit. These non-crystallographic symmetry elements are necessary pieces of information in the process of structure determination.

Self-rotation functions for Crystal Form I were calculated for 90°, 120°, and 180° rotation axes corresponding to two, three and four fold axes of symmetry respectively (Figures 20, 21, and 22). Self rotation functions are a bit difficult to interpret. They are a projection of a three coordinate system onto a spherical axis. The abc coordinate system used is the same system used to define the dimensions of the unit cell. The c axis of the unit cell runs perpendicular to the paper through the center of the spherical projection. As we would expect the c and a axis are separated by a 90° angle. A 120° rotation (the γ angle between the a and b axes) around the edge of the sphere brings us to the b axis.

Crystallographic Symmetry

Let us first consider the crystallographic features of the Patterson maps. Crystallographic features are those that manifest themselves from the symmetry of the space group. For example, consider the search for three-fold axis using the 120° self-rotation function (Figure 21, Crys3 peak). The crystallographic three-fold axis, whose identity as a screw axis is irrelevant for this construction, is apparent in the middle of the spherical map.

Now consider the crystallographic symmetry features in the search for two-fold rotation axes (Figure 22, Crys2 peaks). As we expect from the space group there is a strong two-fold symmetry element parallel to the b axis. Because of its position at the edge of the sphere two peaks, one for each end of

the axis, separated by a 180° angle, are visible on the self-rotation function. However, you will note that there are three pairs of these strong two-fold peaks around the edge of the sphere. These are easily identified as crystallographic symmetry elements after the application of the three-fold axis parallel to c. This three-fold axis creates a three-fold symmetry in the Patterson, explaining all three sets of large peaks around the edge of the sphere (Figure 21, Crys2 peaks).

Non-Crystallographic Symmetry

The remainder of the peaks in the 90°, 120° and 180° self-rotation functions are due to non-crystallographic symmetry within the asymmetric unit. In the two-fold rotation axis search there are 12 peaks in the plane of the a and b axes (Figure 22, NCS1 peaks). These peaks are due to a single non-crystallographic two-fold axis in the ab plane. The combination of the two and three-fold crystallographic axes creates twelve distinct peaks on the self rotation function.

The most distinct features of the 120° self rotation search are the three peaks labeled NCS2 (Figure 21). These non-crystallographic three-fold rotation axis are about 10° toward the c axis from the ab plane. This NCS element, as all others in this crystal system, is privy to the crystallographic three-fold. Thus there are three peaks from a single three-fold axis.

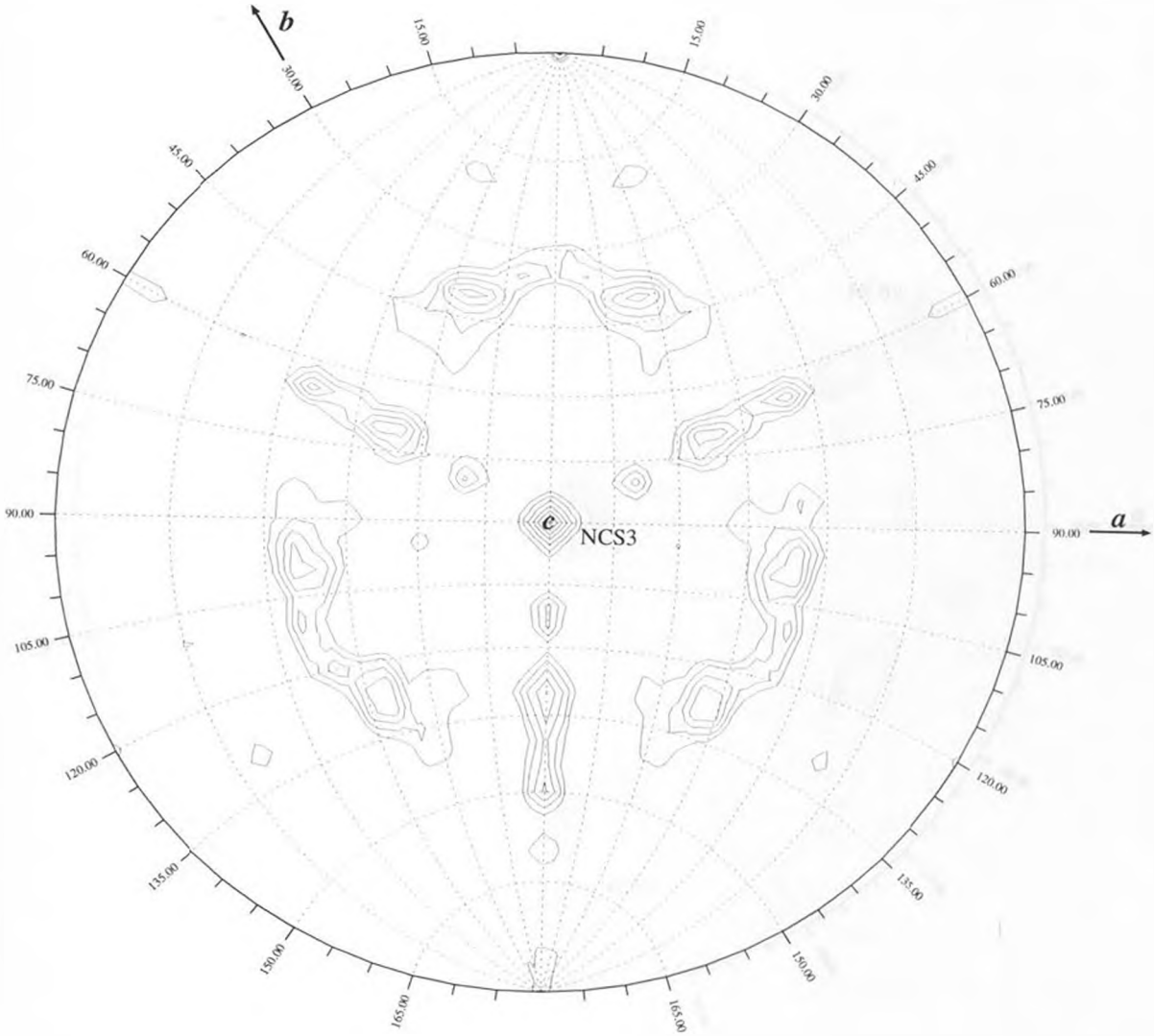


Figure 20: 90° self rotation function for Crystal Form II.

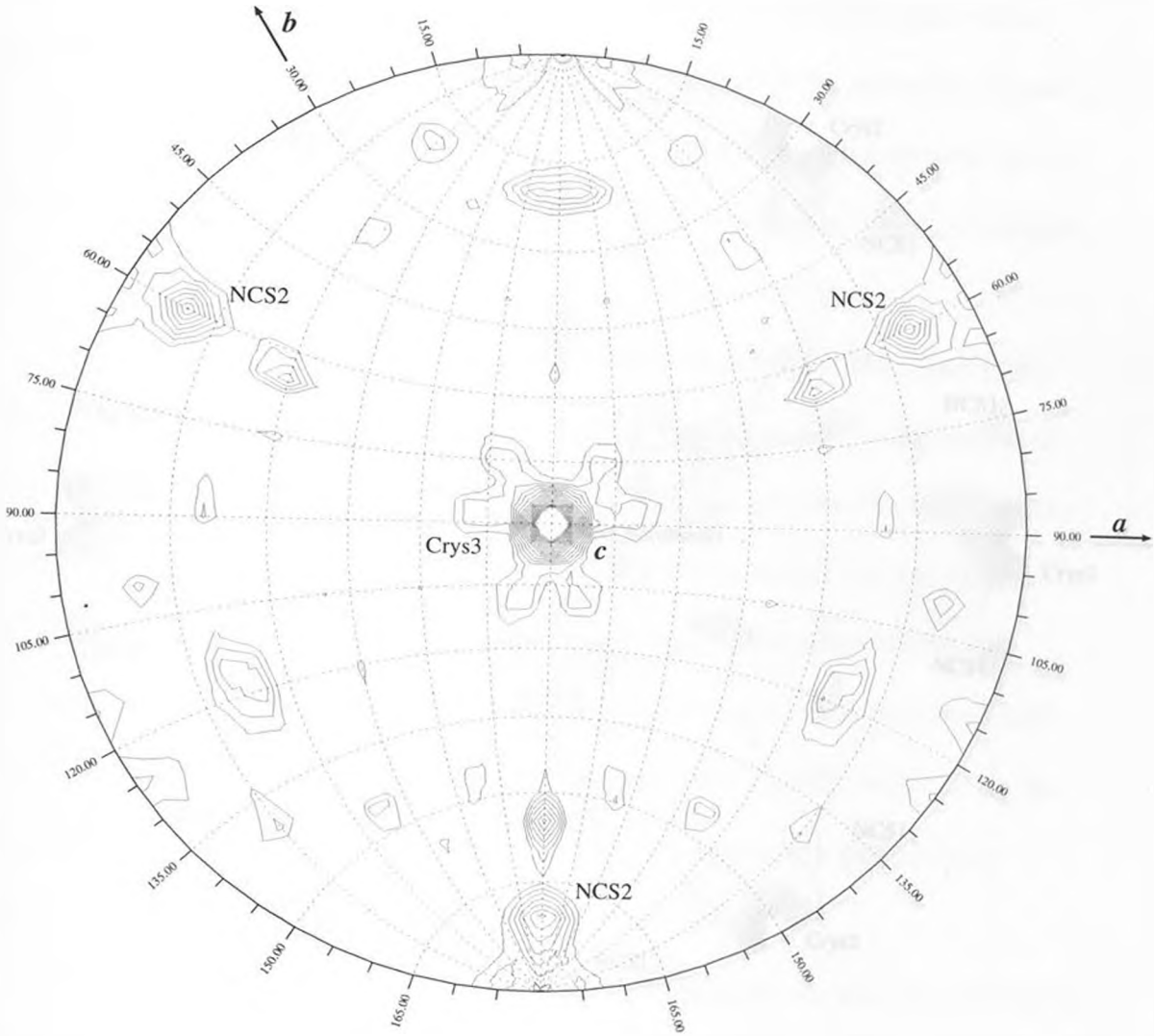


Figure 21: 120° self rotation function for Crystal Form II.

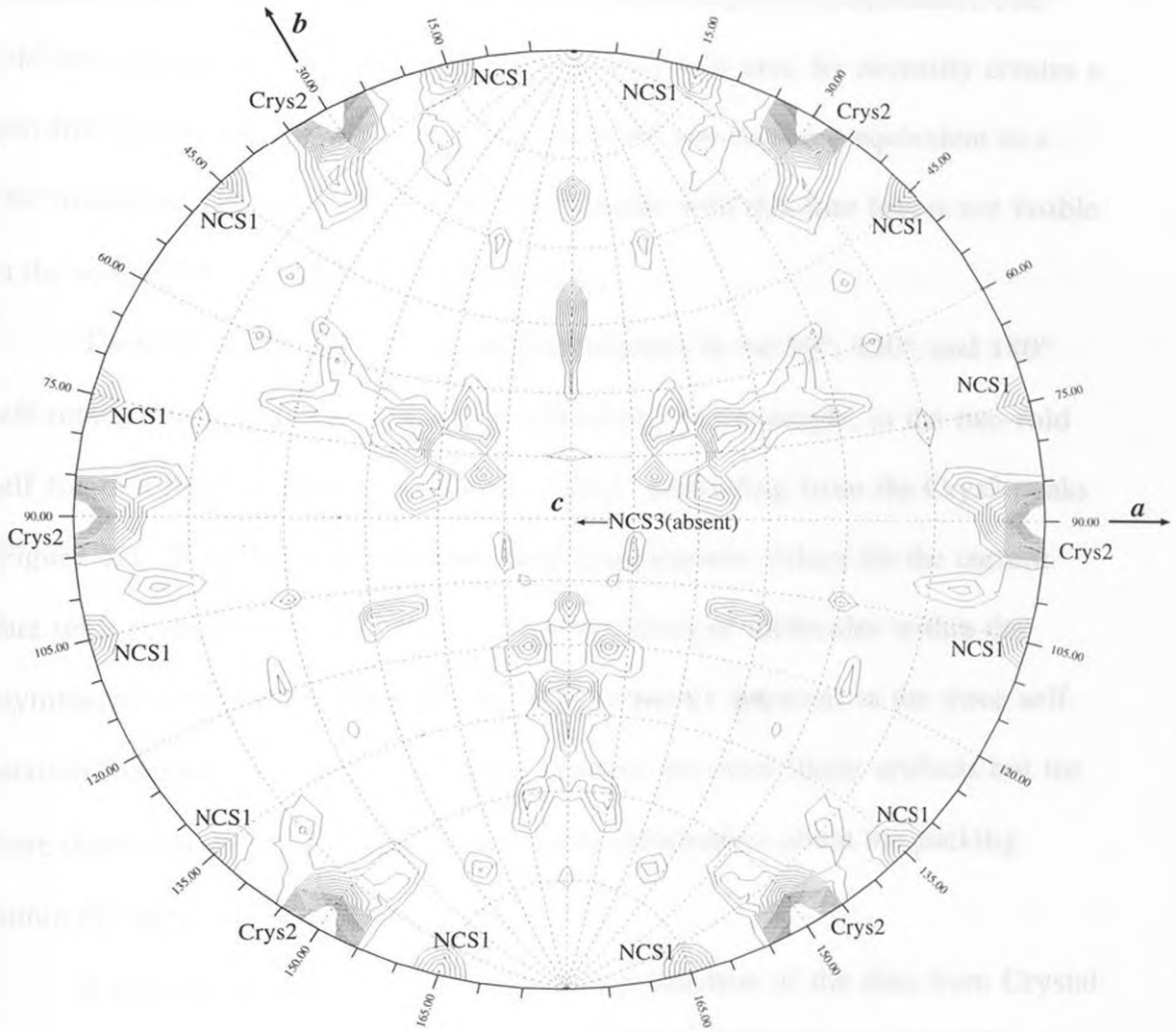


Figure 22: 180° self rotation function for Crystal Form I.

The most distinct feature of the 90° self-rotation search is also the most confusing NCS element. The peak labeled NCS3 (Figure 20) indicates a four fold axis parallel to the c-axis. However, a four fold axis, by necessity creates a two-fold symmetry as well. Two successive 90° rotations are equivalent to a 180° rotation. The two-fold that should coincide with this four fold is not visible in the search for two-fold axes (Figure 22).

There are a number of less distinct features in the 90°, 120°, and 180° self-rotation functions that remain inexplicable. For example, in the two-fold self rotation function there are distinct "arms" protruding from the Crys2 peaks (Figure 22). With the low resolution and completeness values for the current data set it is impossible to arrive at a conformation of molecules within the asymmetric unit that explains all the NCS elements apparent in the three self-rotation functions. A number of these features are most likely artifacts but the more distinct ones have provided invaluable information about the packing within the asymmetric unit.

A translation function for the Patterson function of the data from Crystal Form I was also calculated. Translation functions help to identify molecules that are related by translation. Molecules are most often related by translation when two two-fold axes, a non-crystallographic axis and a crystallographic axis, fall parallel to each other. Two successive 2-fold rotations have the same effect as a translation. No significant peaks were visible in the translation function.

A cross-translation function calculation was also preformed between the Patterson functions for Crystal Form I and Crystal Form II. This calculation was preformed to test the hypothesis that a similar packing arrangement underlies each crystal form. No significant peaks were observed suggesting that the two crystal forms have completely different packing arrangements.

Self-rotation functions were also preformed for Crystal Form I. However, due to the poor quality of the data only the crystallographic symmetry elements were visible.

CHAPTER VI: PROGRESS REPORT

The general procedure for determining the structure of a novel protein through x-ray crystallography is a complicated one. The gene for the protein must first be cloned into an expression system and expressed. A purification scheme must be devised and implemented to obtain a homogeneous mixture of protein. The protein then must be crystallized, usually through a process of screening and fine tuning likely crystallization conditions. Once suitable crystals have been obtained a native data set must be collected.

The next obstacle the crystallographer must overcome is the phase problem. After the phase information using the MAD, MIR, or MR methods is determined an electron density map may be calculated. The crystallographer must then build a molecular model into the calculated electron density and perform successive refinements of that model. Only then is the final structure considered solved.

The ultimate goal of my research of the primase enzyme was to determine the three dimensional structure of a truncated version of the protein. Due to a problem of crystal reproducibility I was unable to progress beyond the phase problem of x-ray structural determination. However, I was able to clone, purify, and crystallize a portion of the T4 primase enzyme, a feat that has never before been done. I was able to characterize four different classes of primase crystals and to collect a data set of 4Å resolution on one of them. I was able to analyze the elements of NCS within the asymmetric unit using the Patterson function.

Although this work is incomplete, it provides an ideal starting point for someone else to continue working toward determining the structure of this protein. I believe that crystal form II should be reproducible with sufficient crystallization screening. In addition, Crystal Form I may become viable for x-ray structure determination. These crystals continue to grow larger and diffract to higher resolution with further screening.

REFERENCES

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

Cha, T.A., Alberts, B.M. (1986). Studies of the DNA helicase-RNA primase unit from bacteriophage T4. *J.Biol.Chem.*. **261**, 7001-7010.

Dong, F., and von Hippel, P.H. (1996). The ATP-activated hexameric helicase of bacteriophage T4 (gp41) forms a stable primosome with a single subunit of T4-coded primase (gp61). *J. Biol Chem.*. **271**, 19625-19631.

Dong, F., Gogol, E.P., and von Hippel, P.H. (1995). The phage T4-coded DNA replication helicase (gp41) forms a hexamer upon activation by nucleoside triphosphate. *J. Biol. Chem.*. **270**, 7462-7473.

Ernst, Bruno. The Magic Mirror of M.C. Escher. Barnes & Noble Books, New York.

Hinton, D.M., Nossal, N.G. (1985). Bacteriophage T4 replication protein 61. *J.Biol. Chem.*. **262**, 12858-12865.

Hinton, D.M., Nossal, N.G. (1987). Bacteriophage T4 primase-helicase. *J. Biol. Chem.*. **262**, 10873-10878.

Ho, D. (1999). Alexander Fleming. *Time* vol. 153, no.12, 117-123.

Jing, D. (1998) PhD Dissertation. University of Oregon.

Kovall, R.H. & Matthews, B.W. (1997) Toroidal Structure of 1-Exonuclease. *Science*. **277**, 1824.

Kwong, P.D., Wyatt, R., Robinson, J., Sweet, R.W., Sodroski, J., and Hendrickson, W.A. (1998). Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature.* **393**, 648-652.

Matthews, B.W. (1968). Solvent Content of Protein Crystals. J. Mol. Biol.. **33**, 491-497.

Pan, H., Bird, L.E., and Wigley, D.B. Structural and biochemical studies of DNA primase from *Bacillus stearothermophilus*. Presented at the Keystone Symposium of Molecular Biology Winter 1999. Taos, NM.

Perutz, Max, (1992). Protein Structure: new approaches to disease and therapy. W. H. Freeman and Company, New York.

Villiams, T.I. (1969). A Biographical Dictionary of Scientists. Wiley-Interscience, New York.

Voet, D. & Voet. J.G. (1990). Biochemistry. John Wiley & Sons, New York.

Wright, R. (1999). Watson & Crick. *Time.* vol. 153, no.12, 172-176.