Spring 2015

Exploring the Effect of Climate Change on Biological Systems

Nardos Sori
Old Dominion University

EXPLORING THE EFFECT OF CLIMATE CHANGE ON BIOLOGICAL SYSTEMS

by

Nardos Sori
B.S. May 2006, Old Dominion University - Norfolk, VA

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

CHEMISTRY AND BIOCHEMISTRY

OLD DOMINION UNIVERSITY
May 2015

Approved by:

Lesley Greene (Director)

Jing He (Member)

Patricia Pleban (Member)

Jennifer Poutsma (Member)

ABSTRACT

EXPLORING THE EFFECT OF CLIMATE CHANGE ON BIOLOGICAL SYSTEMS

Nardos Sori
Old Dominion University, 2015
Director: Dr. Lesley Greene

The present and potential future effect of global warming on the ecosystem has brought climate change to the forefront of scientific inquiry and discussion. For our investigation, we selected two organisms, one from cyanobacteria and one from a cereal plant to determine how climate change may impact these biological systems. The study involved understanding the physiological and adaptive responses at both the genetic and protein function levels to counteract environmental stresses. An increase in atmospheric carbon dioxide is a key factor in global climate change and can lead to alterations in ocean chemistry. Cyanobacteria are important, ancient and ubiquitous organisms that can aid in the study of the biological response to increasing carbon dioxide. Climate predictions estimate that by the year 2100 atmospheric carbon dioxide will exceed 700 ppm. In our first study, we looked at the transcriptional effect of high pCO_2 on the cyanobacteria, *Trichodesmium erythraeum*. Total RNA sequencing was used to quantify changes in gene expression in *T. erythraeum* grown under present day and projected pCO_2 concentrations for the year 2100. Two bioinformatics methods were used to analyze the transcriptional data. The results from this study indicate that a substantial number of genes are affected by high pCO_2. However, increased pCO_2 does not completely alter any one specific metabolic pathway.

As the climate shifts throughout the world, it becomes essential for crops to withstand weather changes. In our second study, we investigated the function of the temperature induced lipocalin (Tatil) from *Triticum aestivum*, which is proposed to help plants survive adverse conditions. This protein is part of a functionally diverse and divergent superfamily of proteins called the lipocalins; they share a common three-dimensional structure, which consists of an antiparallel β-barrel and a C-terminal α-helix. Lipocalins are found in various organisms with a wide range of functions such as pheromone activity, lipid

transport and coloration. Recently, proteins from wheat and Arabidopsis were identified as lipocalins through the elucidation of three structurally conserved regions. The study is particularly timely, as recent studies within the scientific community have shown that at higher temperatures wheat yields will decrease and production will decline by 6% for each 1°C increase. We analyzed the nature of conservation in a large group of sequentially divergent and functionally diverse lipocalins and identified seventeen highly conserved positions as well as built models of the native three-dimensional state of the wheat lipocalin. Based on these computational studies, the wild-type protein and three variants were chosen for a cellular localization study involving site-directed mutagenesis, a gene gun and a confocal microscope. The results provide support for the hypothesis that the L5 loop is involved in the association of the protein with the plasma membrane. We also developed an expression and purification system to produce the wild-type wheat lipocalin protein. Gel filtration chromatography eluted two different sized proteins. Based on the elution volume, one is believed to be the wheat lipocalin trimer while the other one is the monomer. Circular dichroism and fluorescence spectroscopy show that the biological characteristics of the two proteins are different. In the study, Tatil maintains its structure up to approximately 50°C (122°F). In summary, we provide experimental data to better understand mechanistically how microorganisms and plants adapt to environmental change. In cyanobacteria, we show that *T. erythraeum* adapts to pCO_2 increases by up- or down-regulating its genes. In plants, we provide insight into the way in which Tatil interacts with the plant cell membrane as part of its putative function to facilitate robustness in response to temperature increases. The study of Tatil is vital as this protein is believed to help plants tolerate oxidative stress and extreme conditions which broadens our understanding of plant sustainability in different environments.

To Jesus Christ, who had finished the work as He had started it.

ACKNOWLEDGMENTS

First and foremost, may God take all the glory for this work. I would like to thank Dr. Greene, without whom this work would not have been possible. She has continued to support and has challenged me to be the best that I can be. I would like to extend my thanks to the members of my dissertation committee, Dr. Jing He, Dr. Patricia Pleban, and Dr. Jennifer Poutsma for their time and valuable inputs. I would like to thank Dr. Janet Moloney, Dr. Agatha Munyani and Dr. Komala Ponniah for their help and advice. I am very grateful to Dr. Jean- Benoit Charron, who allowed me to conduct research at McGill University.

I would like to thank my parents and brothers for their unwavering support. There are no words to describe the gratitude I feel for my parents who have given everything to bring me to this point. I would like to thank my brothers, Yoseph, my protector, whose never ending patience and wisdom has carried me through the years, Dawit, whose loyalty has been steadfast and Gugssa, through whom God has changed the course of my life. I would like to thank Abeselom and Hiwot who have supported me throughout my graduate life. Seni, Zewdu and Betty, who through their prayers has made this possible. Finally, I would like to thank Jason Collins and Wassim Obeid, who have made invaluable contributions to my life.

NOMENCLATURE

AGP	α-1-acid glycoprotein
AOTF	Acousto optical tunable filter
ApoD	Apolioprotein D
Attil	Arabidopsis thaliana temperature-induced lipocalin
BBP	Bilin-binding protein
BLAST	Basic Local Alignment Search Tool
Blc	Bacterial lipocalin
BLG	β-lactoglobulin
CaMV	Cauliflower mosaic virus
CCM	CO_2 concentrating mechanisms
CD	Circular dichroism
CHLs	Chloroplastic lipocalins
COG	Clusters of Orthologous Groups
EC	Enzyme Commission
ENCAD	Energy calculation and dynamics
ESTs	Expressed sequence tags
GFP	Green fluorescent protein
GPI	Glycosylphosphatidylinositol
KEGG	Kyoto Encyclopedia of Genes and Genomes
MUSCLE	Multiple Sequence Comparison by Log-Expectation
OBP	Odorant-binding protein
PSSM	Position-specific scoring matrix
RBP	Retinol-binding Protein
RuBisCO	Ribulose-1,5- bisphosphate carboxylase/oxygenase
SCRs	Structurally conserved regions
SMTL	SWISS-MODEL template library
Tatil	Triticum aestivum temperature-induced lipocalin
TCA	Tricarboxylic acid cycle
TEV	Tobacco etch virus
TILs	Temperature-induced lipocalins
TIM	Triosephosphate isomerase
TL	Translational

VDE	Violaxanthin de-epoxidase
VEGP	Von Ebner's gland protein
WT	Wild-type

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

1.1 THE PAST, THE PRESENT, AND THE FUTURE

As the climate becomes more unpredictable and carbon dioxide (CO_2) concentration increases at an alarming rate, it becomes essential to study organisms at their current and projected future states. The study of how species respond to various climates aids our understanding of the effects the changing environment can have on biological systems. When assessing climate change, one can look at several indicators such as sea level rise, ocean acidification, global temperature change, ice loss, and changes in atmospheric gas concentration (Radic and Hock, 2011; Bellard et al., 2012). One of the factors that affect the climate are greenhouse gases which have the potential to trap heat. These gases include CO_2, nitrous oxide (N_2O), and methane (CH_4) and are found in nature and hydroflurocarbons, perfluorocarbons, and sulfur hexafluoride (Rigby et al., 2014). These gases are released as a result of industrial process . The concentration of greenhouse gases have increased dramatically in the last century (Farmer and Cook, 2013; Stocker et al., 2013; Delworth and Zeng, 2014). The greenhouse effect describes how greenhouse gases warm the earth. This effect keeps the atmospheric temperature at the right balance so that life on earth as we currently know it is supported. However, as the amount of gases unnaturally increase in the atmosphere, more heat is absorbed, and the earth's atmosphere warms up to potentially problematic levels (Montzka et al., 2011).

Even though the concentration of greenhouse gases may fluctuate naturally, the rate in which the amount is increasing suggests that there are other factors that are the driving force behind global warming (Parmesan et al., 2013; Trenberth et al., 2014). One major factor that has changed in recent years is the industrial revolution of the 1700's (Bachu and Adams, 2003). After 1750, the amount of natural fuels being used, which release gases into the atmosphere,

The dissertation journal model is *Cell*.

have risen considerably (Raynaud et al., 1993; Prather et al., 2012; Rockström et al., 2014). Since 1750 CO_2 has increased by 40% to 391 ppm, N_2O has increased by 20% to 324 ppb and methane has increased by 150% to 1803 ppb (Farmer and Cook, 2013). One of the essential greenhouse gases that has a high impact in trapping heat is CO_2 (Solomon et al., 2009).

1.1.1 CARBON DIOXIDE

CO_2 is essential for both the atmosphere and the carbon cycle. Starting from the top layer, the earth's atmosphere is divided into exosphere, thermosphere, mesosphere, stratosphere and troposphere (Seinfeld and Pandis, 2012). Even though CO_2 is only 0.04% of the atmosphere, it is one of the most important gases in the troposphere (Ahrens, 2011). CO_2 is also an important greenhouse gas as it keeps the earth warm by keeping infrared radiation within the stratosphere layer. CO_2 is unique in that the dipole covalent bonds of the molecule absorb infrared radiation and release it back into the atmosphere in all directions. As the amount of CO_2 increases, the amount of heat within the atmosphere also increases. Thus, a direct correlation between CO_2 concentration and global warming exists (Farmer and Cook, 2013; Humlum et al., 2013; Friedlingstein et al., 2014).CO_2 plays an important role in the global carbon cycle (Figure 1). The cycle works well as long as equilibrium is maintained. As the amount of CO_2 increases, the released gas has to be absorbed by one of the sinks such as the soil or ocean. The ocean is a great reservoir for the gas as CO_2 easily dissolves in water.

Estimates suggest that half of the CO_2 that was released as a result of fossil fuel burning has been absorbed by the ocean (Shaw et al., 2013). The CO_2 absorbed by the ocean results in ocean acidification, which is based on the change in pH as a result of CO_2 concentration. When CO_2 reacts with water, it forms carbonic acid, thus making the water more acidic. Acidity has a high impact on the ocean's ecosystem. For example, coral reefs are unable to grow and organisms such as oysters and clams become incapable of forming their protective outer layer (Pandolfi et al., 2011; Parker et al., 2011). Even though studies on how CO_2 increases and ocean acidification affects different organisms such as fish and crustaceans have been conducted (Kroeker et al., 2010), few studies have been done at the genetic level. To gain this understanding the whole transcriptome and genetic adaptation of an organism should be examined. Adaptation is defined by how organisms are able to change at their genetic level, at the physiological level or alter their behavior as a result of a force exerted by external factors. Studies have shown that animals such as birds and squirrels have made genetic changes to adapt to climate changes (Root et al., 2003; Bradshaw and Holzapfel, 2006; Van Asch et al., 2013). For example, Canadian red squirrels have altered their reproduction time to take advantage of the spruce cone yield timing which has changed throughout the years (Réale et al., 2003).

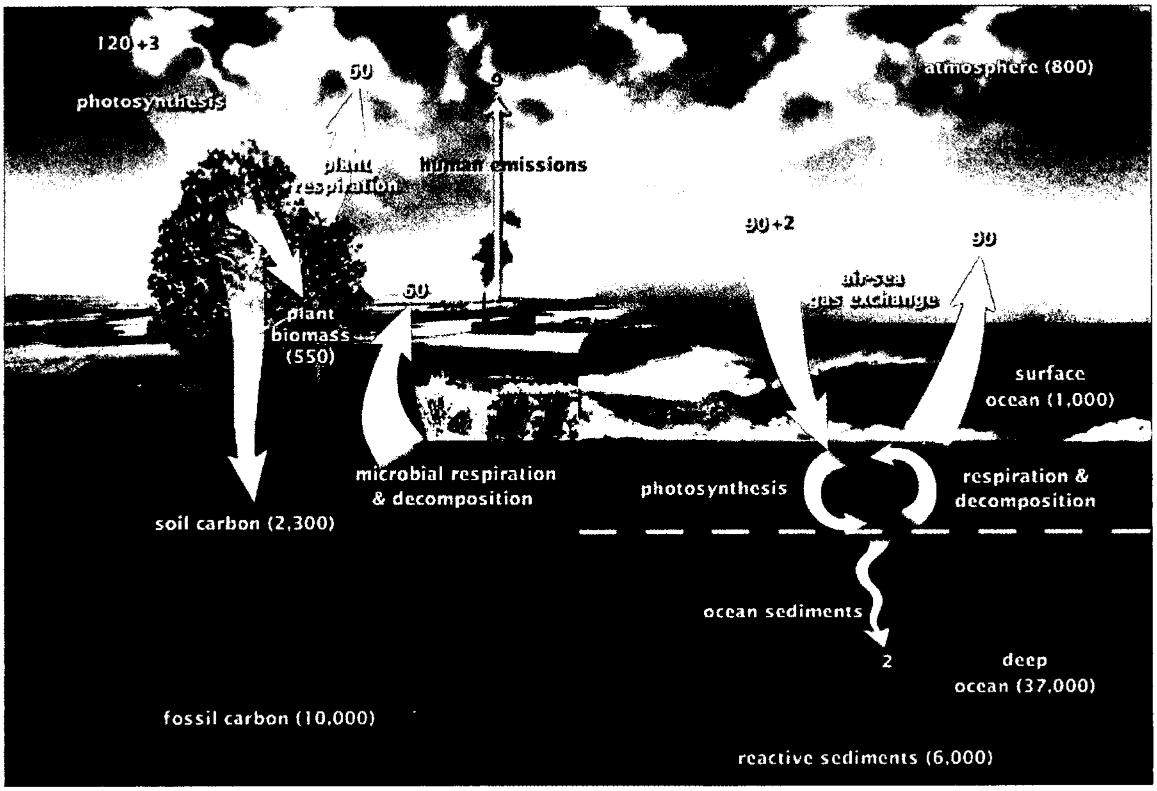


Figure 1. Global Carbon Cycle

A representation of the global carbon cycle.The numbers in parentheses show the main carbon reservoirs in gigatons. The numbers in red show the amount of carbon released as a result of human emissions. This figure is reproduced from (DOE, 2008).

1.1.2 CLIMATE CHANGE

The earth has a variety of climates. One type of climate classification is the Koppen-Geiger which is based on temperature and precipitation. According to this classification, climate is primarily grouped into tropical, dry, mild mid-latitude, cold mid-latitude and polar depending on the location (Stern et al., 2000). Researchers have proposed that the climate will start to shift from its current state as a result of global warming (Colwell et al., 2008). Changes will occur to the weather patterns and temperatures, as well as to the location and level of rain fall. As the global temperature increases, the earth could possibly develop new climate zones (Mahlstein et al., 2013). An example of climate type change was observed in the alpine tundra, which is classified under the polar climate group. From 1987 to 2006 the temperature has increased to 10°C in the western United States, surpassing the threshold for alpine tundra classification (Diaz and Eischeid, 2007). Current climate change is having a drastic effect on organisms, including plants (Root et al., 2003; Pauls et al., 2013; Poloczanska et al., 2013). Therefore, climate change may affect the location and methods used to grow plants which may lead to food scarcity. The growth of wheat, maize and barley have already decreased as a result of temperature increase (Lobell and Field, 2007). Thus, the study of how plants and other organisms can adapt to the changing climate becomes essential.

1.2 MODEL ORGANISMS

Two model systems, cyanobacteria and plants were chosen to study the effect of high CO_2 directly and indirectly. For cyanobacteria, *Trichodesmium erythraeum* was selected to examine how an increase in CO_2 has an effect at the transcriptional level. As a representative of plants, a lipocalin protein from *Triticum aestivum* (wheat) was chosen to understand how it can assist crops to withstand temperature changes. Interestingly, cyanobacteria and plants are ubiquitous organisms that contribute to the current atmospheric environment. While cyanobacteria were essential in facilitating the oxygenation of the earth's early atmosphere about 3 billion years ago (Kaufman, 2014), it also plays an important role in nitrogen and carbon fixation today. Even though the significance of plants as providers of food has been known for some time, the intricate workings of plant proteins are still being discovered.

1.2.1 CYANOBACTERIA

Cyanobacteria are photosynthetic prokaryotic organisms that can survive in a wide range of environments. Fossil records show that cyanobacteria have been present about 3.5 billion years with oxygen evolving cyanobacteria found at least 2.7 billion years ago (Whitton, 2012). Their ability to take electrons from water and produce O_2 transformed the environment so that different types of species can survive on earth (Kaufman, 2014). Even though they produce oxygen as a result of photosynthesis, some cyanobacteria also fix nitrogen thus playing an important role both in carbon and nitrogen cycles (Kumar et al., 2010).

Throughout the years, they have adapted to their changing habitat. They have several inherent abilities that allows them to adapt to various environments. They are able to withstand water stress and high salinity. They are also able to use low amounts of light for photosynthesis. However, the changing environment still affects how the organism grows. Cyanobacteria blooms, dense growth of the organism, are affected by several factors such as the amount of light they receive, temperature, salinity and nutrients (Paerl and Huisman, 2009; Paerl and Otten, 2013). For example, studies show an increase in water temperature result in large blooms of cyanobacteria such as the one seen on the *Lyngbya* species (Paul, 2008). Changes in rainfall patterns or the intensity of hurricanes also alters the growth of cyanobacteria.

The cell structure of cyanobacteria is critical to the species survival in a wide range of conditions as well as conduct photosynthesis. They have three different types of membranes, the outer membrane, the plasma membrane, and a thylakoid membrane (Bryant, 1995). Photosystem I and II, cytochrome b6f, NADH dehydrogenase, cytochrome oxidase, and ATP synthase are found within the thylakoid membrane. A glycocalyx is located on the top of the outer membrane and protects the cell from harm in different environments. Cyanobacteria have a peptidoglycan layer between the outer membrane and the plasma membrane.

Some cyanobacteria also conduct nitrogen fixation. Nitrogenase is an enzyme that converts atmospheric nitrogen to ammonia and is inhibited by oxygen (Berman-Frank et al., 2001). Because of the irreversible inhibition by oxygen, different cyanobacteria have adapted various ways to separate nitrogenase from the oxygen produced by photosynthesis. One of the adaptation mechanisms separates the time at which nitrogen fixation and photosynthesis occur. Another one is to confine nitrogen fixation to certain cells called heterocysts.

Many different types of cyanobacteria are found throughout the world. One way to classify them is based on their morphology. This method divides cyanobacteria into five distinct groups: Chroococcales, Pleurocapsales, Oscillarotiales, Nostocales and Stigonematales (Herrero and Flores, 2008). In the group Chroococcales, the cyanobacteria are mostly colonial and unicellular such as *Synechococcus* (Flores and Herrero, 2008). Pleurocapsales are composed of unicellular to pseudo-filamentous cyanobacteria. Oscillarotiales are filamentous organisms that do not form heterocysts. Nostocales are also filamentous but form heterocysts. The last group is Stigonematales which consist of filamentous nitrogn-fixing cyanobacteria. The cyanobacteria that we chose for our experiments is '*Trichodesmium*' which is found in the group Oscillarotiales, and is a non-heterocyst filamentous cyanobacteria. The cyanobacteria, *Trichodesmium erythraeum* received its name in 1830 when Ehrenberg observed a change in the color of the Red sea caused by this species (Bergman et al., 2013).

Trichodesmium is unique in that nitrogen fixation and oxygen production occurs within the same period of time and lacks specialized cells for nitrogen fixation. The organism is able to do that by varying the level of nitrogen fixation and photosynthesis during the day (Berman-Frank et al., 2001). In the morning, the level of photosynthesis is high while nitrogen fixation is low, and during midday it is reversed . *Trichodesmium* is found in the tropical, subtropical, Atlantic, Pacific and Indian oceans as well as the Caribbean and South China seas (Capone et al., 1997). The cyanobacteria are found in blooms, which consist of tuft and puff shaped colonies (Figure 2). The colonies consist of hundreds of trichomes, which can be associated to form filaments or remain as individuals (Finzi-Hart et al., 2009). The arrangment of the colonies is believed to help in the buoyancy of the cyanobacteria. The optimal temperature for *Trichodesmium* growth is 24-30°C; however, it can also survive in lower temperatures (Bergman et al., 2013). Cyanobacteria require various types of nutrients but nitrogen fixation is limited by the nutrients iron and phosphorous (Sañudo-Wilhelmy et al., 2001). We chose this important cyanobacteria to analyze the impact that high pCO_2 might have on its transcriptome. Chapter two discusses the difference in transcriptional change between cyanobacteria grown under the present day pCO_2 and the projected pCO_2 for the year 2100.

1.2.2 *TRITICUM AESTIVUM*

Triticum aestivum, known by the common name "wheat", is the most widely used crop throughout the world (Figure 4) (Lobell et al., 2012). It was one of the first domesticated crops 8,000 years ago (Carver, 2009). Even though wheat can be grown within a wide range of temperatures, the best temperature to grow the crop is about 25°C. Based on when the cereal is grown, wheat is divided into two categories, spring and winter wheat (Matz, 1991). Winter wheat is planted in the fall and becomes dormant in the winter months until the soil becomes warm in the spring. Even though, winter wheat can withstand cold temperatures, prolonged exposure can destroy the plant. On the other hand spring wheat is planted during spring and harvested in late summer or fall. Wheat growth is affected by many factors such as carbon dioxide, light, nitrogen, temperature and water. The production of wheat is adversely affected by lack of water, temperature and soil salinity (Acevedo et al., 2002; Saqib et al., 2013; O'Leary et al., 2014).

Wheat morphology is divided into root and shoot (Kirby, 2002). The plant has two types of root, seminal roots and nodal roots, which come from the lower nodes. The shoot consists of phytomers, which are repeating units each consisting of a potential leaf, a stem and a lateral bud. Several of these units are found at the stem of the shoot. One of the units makes up the stem which is an elongated internode while the other units remain short.

1.3 LIPOCALINS

1.3.1 ORIGIN OF THE NAME AND THE STRUCTURE

The term 'lipocalin' was first coined in 1987 by Syed Pervaiz and Keith Brew (Pervaiz and Brew, 1987) . The name was given to a superfamily of proteins that enclose lipophilic molecules, the same way that a calyx enfolds a flower. They compared four proteins, the retinol-binding protein (RBP), β-lactoglobulin (BLG), α2u-globulin and α1-microglobulin, based on their amino-acid sequence, disulfide bonds and three-dimensional (3D) structures. They found the proteins to be

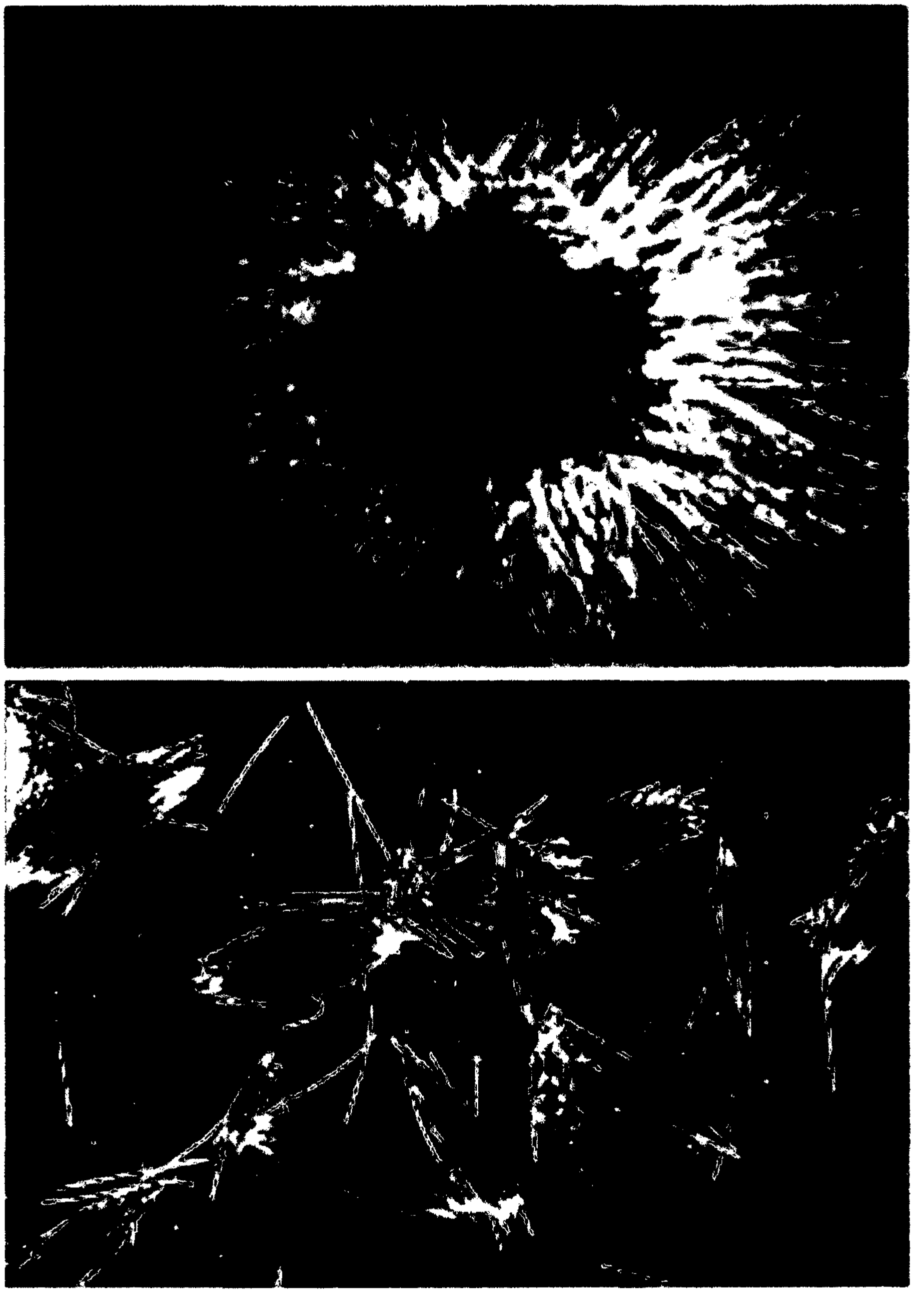


Figure 2. Two Morphological Forms of *Trichodesmium*

T. erythraeum is shown in two different forms. The top panel shows the puff form while the bottom panel shows the tuft form. Figure adapted from (Whitton, 2012).

Figure 3. Photo of Wheat

A picture of wheat taken by John Bedford, Old Dominion University, in Morocco (2013).

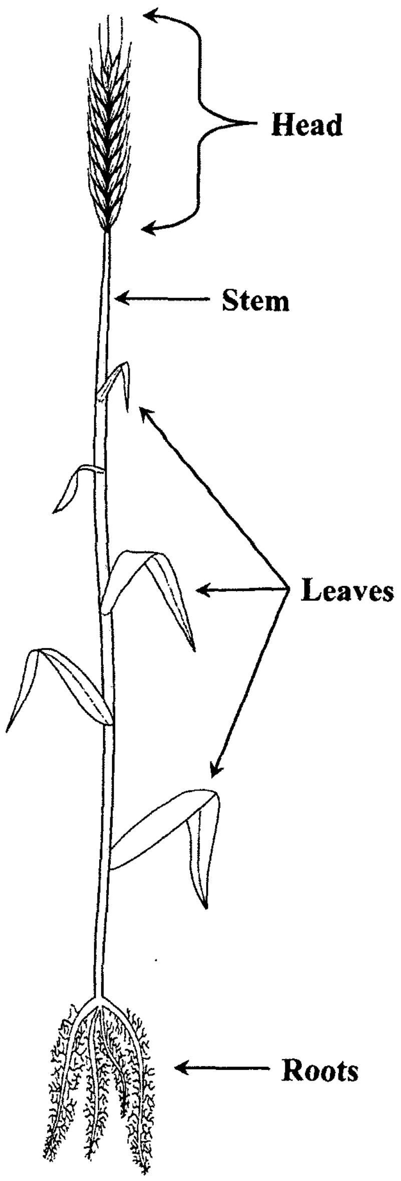


Figure 4. Basic Structure of the Wheat Plant
An illustration showing key components of wheat.

similar in structure and bind hydrophobic molecules. Since then more lipocalins have been discovered in various organisms and the function of the lipocalin proteins was determined to vary from ligand transport to insect coloration (Åkerström, 2006). Lipocalins can be found in humans (Meining and Skerra, 2012), boar (Spinelli et al., 2002), mouse (Timm et al., 2001) and bacteria (Krebs et al., 2013) just to mention a few and are essential for various organisms (Schlehuber and Skerra, 2005).

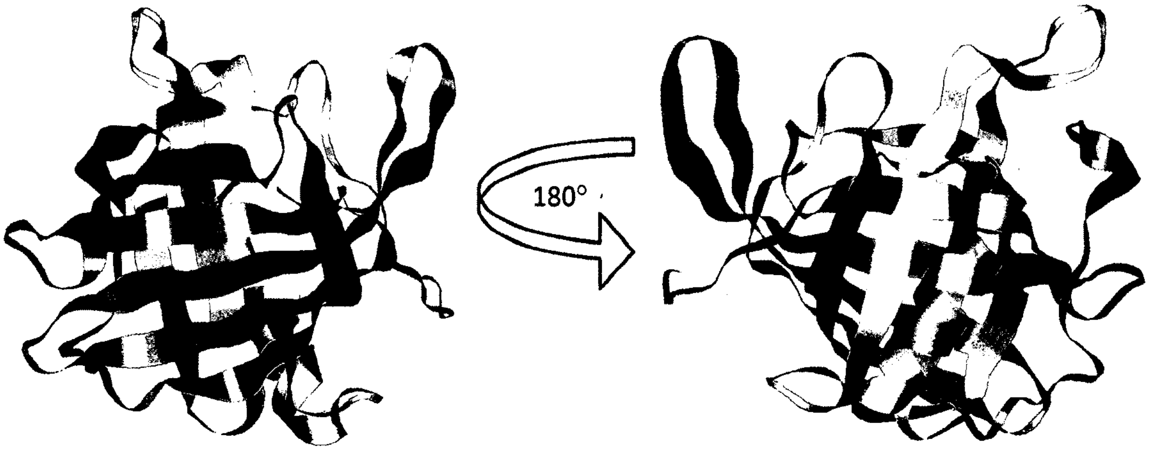
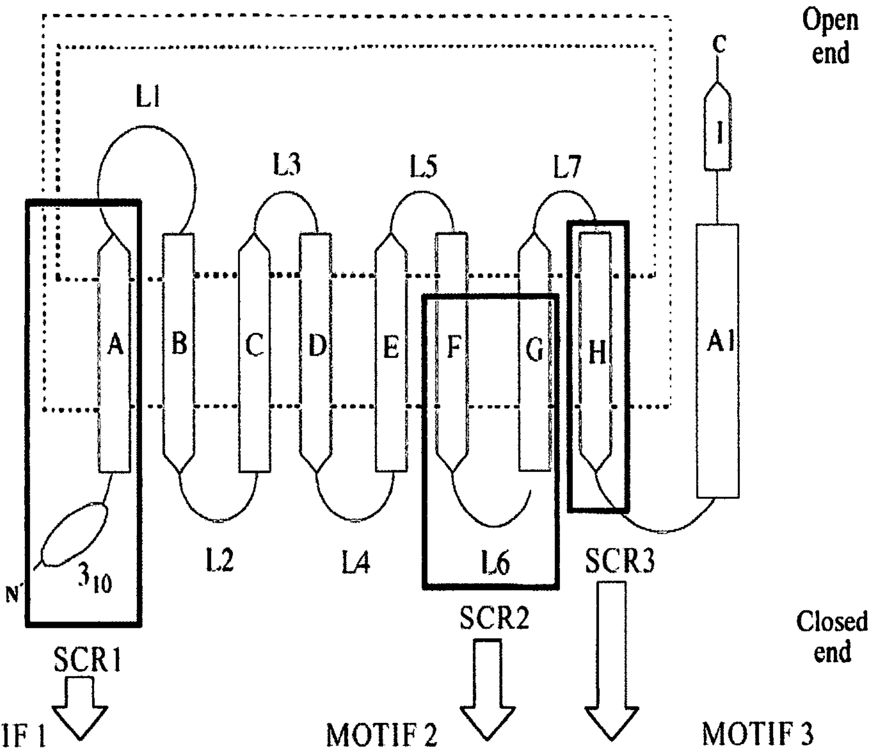


Figure 5. 3D Structure of a Retinol-Binding Protein
Ribbon drawing of the Retinol-binding protein lipocalin structure (PDB Code:1RBP). Color scheme going from the N-terminus in blue, to the C-terminus in red. This structure is visualized on Rasmol (version 2.7.5.2) (Sayle and Milner-White, 1995)

In 2002, Jean-Benoit Charron and co-workers from the University of Quebec at Montreal identified the first true plant lipocalin proteins in wheat and Arabidopsis. The proteins were named *Triticum aestivum* temperature-induced lipocalin (Tatil) and Arabidopsis thaliana temperature-induced lipocalin (Attil). A cyanobacteria lipocalin, *Gloeobacter Violaceus* was also discovered, showing that lipocalins have existed for a long period of time (Charron et al., 2005). The plant lipocalins were selected for further study as the protein helps plants withstand extreme temperatures and recover from various conditions. This aid prolongs the life of the plants. This property of the protein would be a tremendous benefit in growing plants in various environments without relying on perfect weather conditions. A detailed discussion of Tatil and its context within the lipocalin protein superfamily is discussed in chapter three of this dissertation.

Lipocalins contain motifs that are identified as structurally conserved regions (SCRs) (Flower, 1996). The sequences within these regions are used to identify if a protein belongs to the lipocalin superfamily. Lipocalins can either have two or three SCRs. The proteins that have all three of the SCRs are identified as kernel lipocalins, while the proteins that have two out of the three SCRs are identified as outlier lipocalins (Grzyb et al., 2006). Another way to identify lipocalins is through their structure. Lipocalins are unique in that even though their sequence identity is low, often falling below 20%, they have a common 3D structure (Greene et al., 2001). This character of lipocalins is different as most proteins with the same topology have high sequence identity. Lipocalins have an eight-stranded antiparallel β-barrel that is composed of two sheets and an α-helix at the C-terminus (Figure 5). The eight β-strands are labelled A-H and the connecting loops are labeled L1-L7 (Figure 6). The loops are mostly β-hairpins with L1 being an omega loop and closes the barrel. SCR1 is found on strand A, SCR2 spans strand F to G which includes loop 6. SCR3 is found on strand H and the loop connecting it to the α-helix. The three SCRs are found close to each other in the 3D structure which forms the base of the β-barrel (Figure 6).

Lipocalins are part of a large group of proteins that have the β-barrel architecture. Perhaps the most prominent β-barrel protein is the green fluorescent protein (GFP) (Yang et al., 1996). The protein is an 11 stranded β-barrel protein that was first found in a jellyfish. This protein has revolutionized the way cells can be studied because of the protein's ability to emit light under certain wavelengths. β-barrels are formed by either parallel or anti-parallel β-strands. They can be classified into two categories based on the hydrogen bond pattern between the strands (Nagano et al., 1999). If all the β-strands are involved in hydrogen bonds then the structure is considered a complete barrel. However, if hydrogen bonds are not formed by some of the strands then it is regarded as a distorted barrel. As hydrogen bonds are formed throughout the lipocalin barrel, the superfamily can be categorized as a complete barrel. The majority of the β-barrel proteins have anti-parallel β-strands. The exceptions are



	MOTIF 1
A1M	NFNISRIYGKWYNLAIG
RBP	NFDKARFSGTWYAMAKK
BBP	NFDWSNYHGKWWEVAKY
VEGP	TEENQDVSGTWYLKAAA
AGP	NATLDQITGKWFYIASA
OBP	NETLSWLSGKWFLIAVA

	MOTIF 2
A1M	SYVVHTNYDEYAIFL
RBP	HWIVDTDYDTYAVQY
BBP	FNVLSTDNKNYIIGY
VEGP	YIIPSSVEDHYIFYY
AGP	LILRDTKTYMLAFDV
OBP	RVLEKHGAIMLFFDL

	MOTIF 3
A1M	YGRAPQLRETLLQDF
RBP	FSRDPNGLPPEAQKI
BBP	LSRSKVLTGEAKTAV
VEGP	VGRDPEINQEALEDF
AGP	YADKPETTKEQLGEF
OBP	SARRPDIPPELREVF

Figure 6. 2D Lipocalin Structure Schematic

A general schematic of the lipocalin secondary structure and connectivity is shown. The three SCRs are annotated in black boxes. The three motifs found in various organisms is also shown in a black box and the highly conserved residues are highlighted in red. Protein code (A1M: α1-microglobulin, RBP: plasma retinol-binding protein, BBP: bilin-binding protein, VEGP: von Ebner's gland protein, AGP: α1-acid glycoprotein, OBP: odorant-binding protein). Figure adapted from (Flower, 1996).

TIM barrel proteins which gained their name from the structure of triosephosphate isomerase (Wierenga, 2001). The TIM fold is comprised of eight β-strands on the inside with eight α-helices on the outside. The TIM β-barrel protein which functions as an enolase has a $\beta\beta\alpha\alpha(\beta\alpha)_6$ topology and also contains anti-parallel β-strands outside the barrel (Figure 7) (Hosaka et al., 2003). The average number of amino acids in a TIM barrel is about 200 residues. These proteins are involved in various enzymatic activities. They are found in four of the five enzyme commission classes; none of them are ligases (Vega et al., 2003).

Another class of β-barrel proteins are membrane proteins. They are found only in specific areas, such as in gram negative bacteria, chloroplasts and mitochondria (Noinaj et al., 2013). These proteins perform various functions ranging from transport to receptors. Even though lipocalins have an average of eight β-strands, the number of β-strands can vary in membrane proteins. They can range from eight as seen in *Yersinia Pestis* to twenty-four strands as found in *E. coli* (Fairman et al., 2011) (Figure 8).

1.3.2 LIPOCALIN FUNCTIONS

Even though lipocalins have the same 3D structure they are involved in various functions. Lipocalins were first known for transporting lipophilic molecules. Yet, even then their function is varied. Apolipoprotein D and M are examples of lipocalins that are involved in the transportation of lipids (Drayna et al., 1986; Xu and Dahlback, 1999). A lipocalin in human plasma, RBP (Figure 5), transports vitamin A (Newcomer and Ong, 2000) and in another case, the bilin-binding protein in butterfly gives the insect a blue color (Huber et al., 1987). Another function of a lipocalin is seen in the grasshopper Lazarillo. The protein is a cell surface lipocalin expressed from embryogenesis to maturity and aids in the growth of axons (Sanchez et al., 1995; Ruiz et al., 2012). Lipocalins, such as rat and frog odorant-binding proteins transport odorant molecules to receptors (Flower, 1996). Odorant-binding molecules may also protect against insects and be involved in detoxification (Grzyb et al., 2006). The use of a common fold to achieve such functional diversity indicates a robust adaptability in evolution.

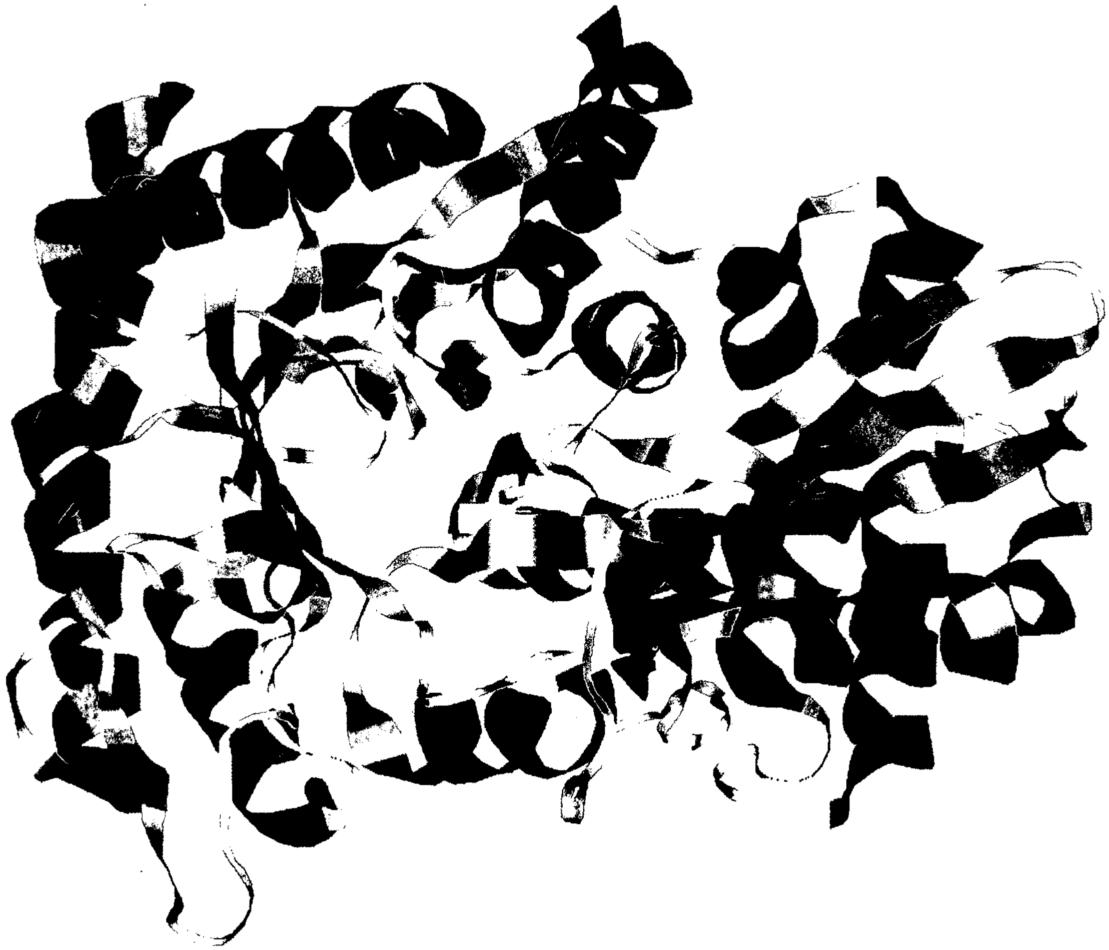


Figure 7. 3D Structure of an Enolase from *Enterococcus hirae*
The illustration shows one of the monomers that forms the dimer of an enolase. The β-barrel is displayed in yellow while α-helices are shown in pink. The loop regions are shown in white. The anti-parallel β-strands found beside the β-barrel can be seen on the right (PDB Code:1IYX). The structure is visualized as a ribbon drawing using Rasmol (version 2.7.5.2).

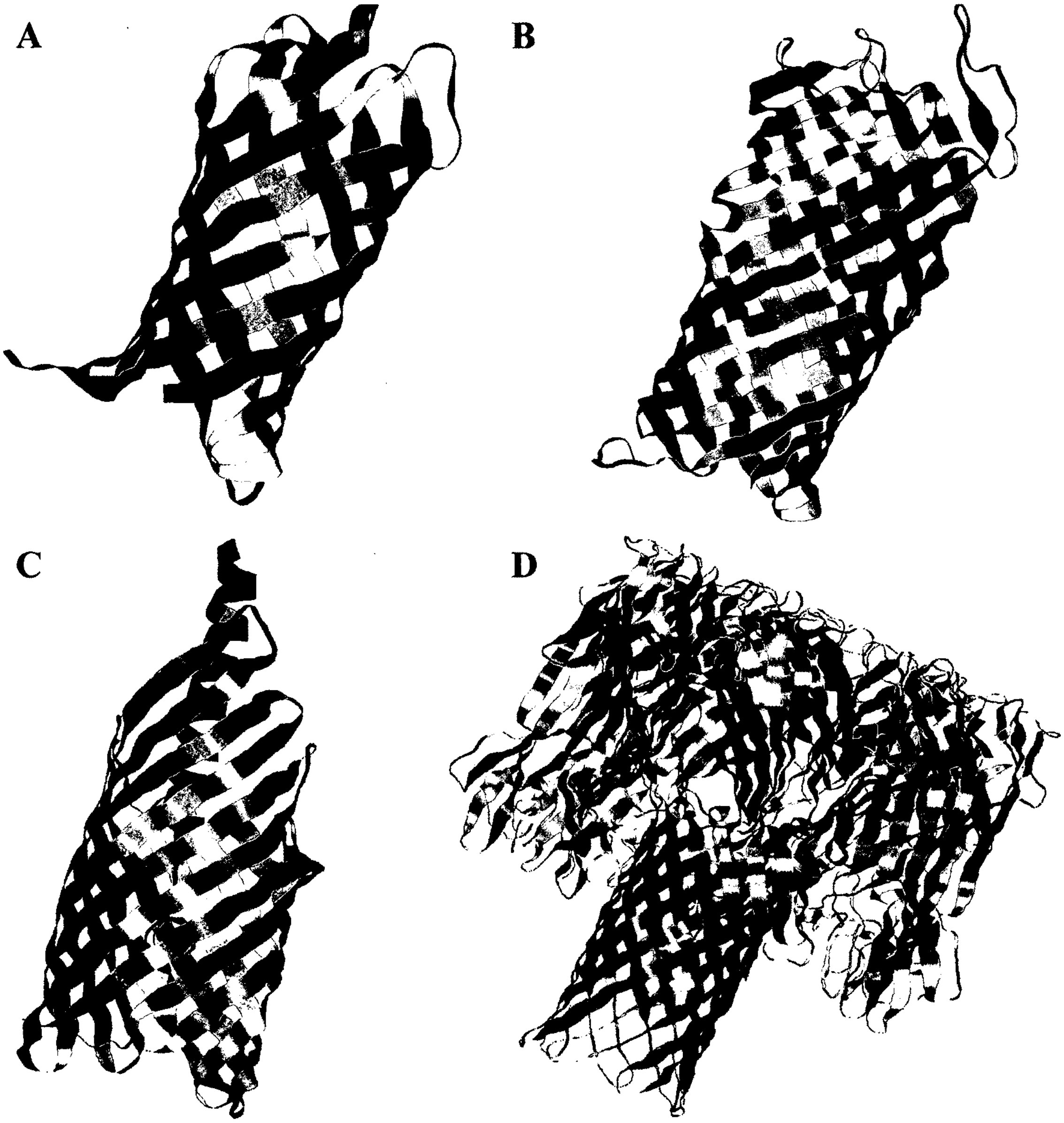


Figure 8. 3D Structure of Select β-barrel Proteins

Proteins that have β-barrel structures but with different numbers of β-strands. A. Neisserial surface protein A is a protein found in gram-negative bacteria. It has eight β-strands and is involved in the adhesion of the bacteria to a host cell (Vandeputte-Rutten et al., 2003) (PDB code:1P4T). B. A protease found in *E. coli* that has 10 β-strands (Vandeputte-Rutten et al., 2001) (PDB code:1I78). C. β-barrel that is composed of 12 β-strands. It is an autotransporter found in *E.coli* (PDB code:3AEH). D. Part of a heptameric transmembrane pore found in *Staphylococcus aureus* (PDB code:7AHL).

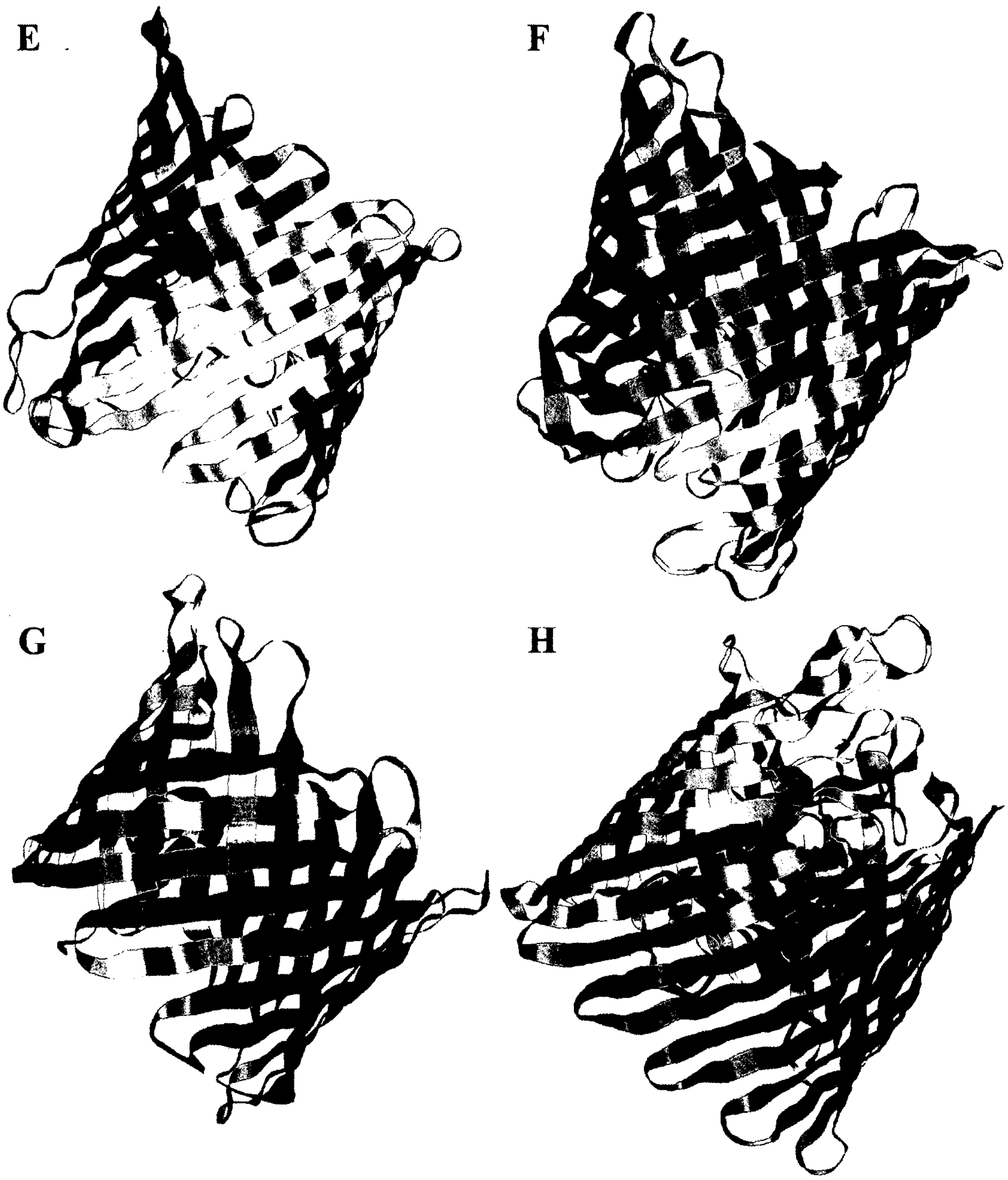


Figure 8. Continued

Diverse β-barrel proteins with different number of β-strands. E. Porin Omp32 protein which has 16 strands of β-strands (Zachariae et al., 2006) found in *Delftia acidolorans* (PDB code:2FGQ). F. A transporter protein seen in *Pseudomonas fluorescens* that has 18 β-strands (Sampathkumar et al., 2010) (PDB code:3JTY). G. A voltage dependent anion channel found in a mouse. It has 19 β-strands (Ujwal et al., 2008) (PDB code:3EMN). H. A transporter that is composed of 22 β-strands seen in gram-negative bacteria (Chimento et al., 2003) (PDB code:1NQE).

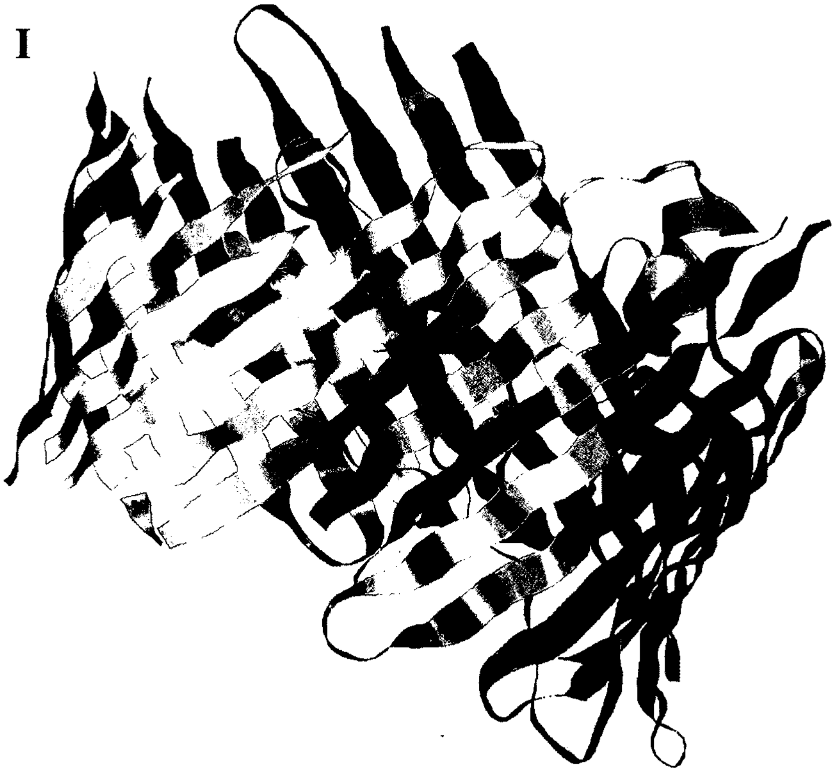


Figure 8. Continued

β-barrel protein with large number of β-strands. I. 24 β-stranded β-barrel of a bacterial outer membrane protein (Huang et al., 2009) (PDB code:3FIP).

Figure 9. 3D Structure of the Bilin-Binding Protein
The lipocalin bilin-binding protein found in butterfly. The protein is a homotetramer with each chain given a different color (PDB code:1BBP). The structure is visualized as a ribbon drawing using Rasmol (Version 2.7.5.2).

1.3.3 MEMBRANE ASSOCATION OF LIPOCALINS

Homology studies of Tatil and Attil have shown that they are related to mammalian Apolioprotein D (ApoD), bacterial lipocalin Blc and insect Lazarillo protein (Charron et al., 2002). All three are membrane proteins that associate with the membrane through different methods (Yang et al., 1994; Ganfornina et al., 1995; Campanacci et al., 2004). One possibility for Attil and Tatil to associate with the plasma membrane is through a loop. Attil is known to localize in the plasma membrane (Charron et al., 2005); however, how it interacts with the membrane is unknown. We studied the possible mode of interaction of Tatil by mutating three amino acids found in the loop between β-strands E and F (loop 5) (chapter four). A gene gun was then used to insert an onion cell with the DNA for the wild-type (WT) and three variants. The localization of the protein in the cell was then analyzed by confocal microscopy. The gene gun technology is very valuable for introducing DNA or RNA into a system such as cells or tissues (Kim and Eberwine, 2010). For example, it has been used to generate expression in liver for gene therapy (Kuriyama et al., 2000), in inducing T cell responses for immunization (Cho et al., 2001) and to apply vaccines (O'Brien and Lummis, 2011).

1.4 METHODS OF ANALYSIS

1.4.1 TRANSCRIPTION AND GENOMICS

A great amount of information can be gained by studying an organisms whole transcriptome. Transcriptomics is a newly developed method of study that looks at mRNAs, non-coding RNAs and small RNAs to gain an understanding of the transcriptional level of genes (Wang et al., 2009). The method that we used to study the whole mRNA transcriptome is based on RNA-sequencing. In this method, mRNA is converted to a library of cDNA fragments, sequencing adaptors are then added to the cDNA, which are then sequenced. This methodology has been applied by researchers around the world to study a wide range of scientific problems such as Alzheimer's disease (Twine et al., 2011), cell lymphoma (Kridel et al., 2012), epilepsy (Okamoto et al., 2010) and stress induced changes in plants (Zeller et al., 2009). The Alzheimer study provides an example of this approach. Here, a whole transcripome analysis was performed on a patient's brain affected by Alzheimer's disease and compared with a normal brain. The results indicated that there were in fact transcriptonal differences between the two brains.

In this dissertation (chapter two) whole mRNA transcriptome analysis was conducted on *Trichodesmium* that was grown under current pCO_2 and projected pCO_2 in the year 2100. The transcriptional changes between the two conditions were analyzed to assess which biochemical pathways are affected and how Trichodesmium fares in high pCO_2.

1.4.2 BIOINFORMATICS

Bioinformatics is a computational approach applied to the study of important questions in the disciplines of biology and biochemistry. It is applied in this dissertation research to elucidate the determinants of protein structure, function, stability and folding. The computational results are used to guide the experimental research. In general, bioinformatics is also key to understanding the evolution of proteins and genes. Three central and fundamental bioinformatics approaches are applied in this research: searching for related sequences, generating structure-based superfamily alignments, and 3D model building.

1.4.3 PROTEIN BIOCHEMISTRY

The ability to understand protein structure, function and dynamics drastically improved with the advent of cloning and protein engineering technologies that have been significantly advanced since the 1990's. Proteins from all organisms can be expressed in one or more *in vivo* systems such as *E. coli* or, yeast and *in vitro* translation systems. After optimization, which is often not trivial, proteins are purified using column chromatography. Most of these basic methods have been around since the 60's and 70's but the technology has advanced significantly to improve experimental results and the speed of the experiments. After a successful expression and purification of a protein, the next step is characterization. Through different methods such as circular dichroisim and fluorescence spectroscopy, the stability and structure of the protein is characterized. This dissertation looks at the steps required for the purification and expression of Tatil. The characterization of this protein enables the study of crop survival in various environments.

1.5 SPECIFIC AIMS

This dissertation attempts to reveal by using two model systems how life can adapt to withstand global warming. To answer a question of this magnitude, two organisms were chosen, the cyanobacteria *Trichodesmium erythraeum* and the wheat plant *Triticum aestivum.* These organisms are ideal to study the adaptation of various species as they represent ones that contributed to the oxygenation of the earth's atmosphere and a crop that is widely used in the world. In one study, what happens to an organism at the transcriptional level under a high CO_2 concentration is examined and in the other, how plants adapt to withstand climate change is investigated. The studies conducted on these models are a stepping point to further understand how various species are impacted and the adaptation that organisms make to acclimate to the changing environment.

CHAPTER 2

EFFECT OF PRESENT AND FUTURE CO_2 LEVELS ON GENE EXPRESSION IN *TRICHODESMIUM ERYTHRAEUM*

2.1 BACKGROUND

CO_2 is emitted in the atmosphere naturally as well as by man-made ways and is essential to the environmental balance of the atmosphere and oceans (Schnoor, 2007). Atmospheric CO_2 concentration has increased approximately 40 percent in the last 300 years, mainly as a result of human activities (Prinn, 2003; Quéré et al., 2013; Stocker et al., 2013). Carbon cycle models project that in the year 2100, pCO_2 may increase between 700 to 1000 ppm (Booth et al., 2012; Khodayari et al., 2013; Zhang et al., 2014). As the concentration of CO_2 in the atmosphere increases, the amount of CO_2 absorbed by the ocean also increases, which leads to a rise in hydronium ion concentration.

$CO_2 + H_2O + CO_3^{2-} \rightarrow 2HCO_3^-$

This leads to a decrease in ocean pH and thus, ocean acidification (Zeebe et al., 2008). Researchers estimate that the pH will decrease from the current value of 8.10 to 7.82 by the year 2100 (Gattuso and Hansson, 2011). One of the organisms that will be affected by ocean acidification is cyanobacteria. Some species of cyanobacteria can grow in a wide range of pH conditions such as the *Synechocystis* species, which can grow between pH 7.5 and 10, while others are not able to withstand a pH change (Hinga, 2002; Summerfield and Sherman, 2008). The heterocystous cyanobacteria *Hapalosiphon* can be found in Flow Country, a region in Scotland in pH of 4.1- 4.5, while other types of cyanobacteria such as *Oscillatoria* and *Spirulina* grow in Lusatia, Germany in a pH of 2.9 (Whitton, 2012).

Cyanobacteria are important organisms. They played a key role in facilitating the rise of atmospheric oxygen about 2 billion years ago (Mulkidjanian et al., 2006). In addition, even though cyanobacteria are less than 1% of the marine system, they are responsible for the majority of nitrogen fixation (Kasting and

Siefert, 2002). One of the cyanobacteria, *T. erythraeum*, plays a key role in the carbon and nitrogen cycles of oligotrophic oceans (Breitbarth et al., 2007). *T. erythraeum* is an ideal organism for studying the effect of CO_2 as the cyanobacteria can illustrate the consequence of high CO_2 on photosynthesis and carbon fixation as well as nitrogen fixation (Capone et al., 1997).

Illumina sequencing-based transcriptome analysis methodology (mRNA-Seq) was conducted to give us a view on how *T. erythraeum* regulates gene expression in a high pCO_2 atmosphere. mRNA-Seq is a new technology that has several advantages, such as having a high sensitivity to low levels of gene expression readings, detecting genomic sequences that have not been established, and being highly reproducible (Wang et al., 2009). Even though the effect of CO_2 on *T. erythraeum* has been studied by looking at nitrogen and carbon fixation on limited set of genes (Kroeker et al., 2010), this research is the first time that whole mRNA sequencing has been conducted on *T. erythraeum* under challenge by external pressures. Another group very recently looked at the whole transcriptome of *T. erythraeum* to study transcriptional start sites (Pfreundt et al., 2014). The results showed that *Trichodesmium* has 6,080 transcriptional start sites and that approximately 40% of the promoters are involved in non-protein coding RNAs. In the present dissertation research a whole transcriptome study was done to gain an understanding of the underlying molecular effects of high pCO_2 on the cyanobacteria, which in turn gives an insight into how biological systems could be affected by the changing atmospheric environment.

Very few experiments studying the effect of elevated CO_2 on the whole transcriptome of an organism have been performed in general. One study looked at how a coral, *Acropora spp.* responds to high CO_2 concentration (Moya et al., 2012). The coral was grown under 380 ppm CO_2 and projected CO_2 concentrations of 750 ppm and 1000 ppm. The data showed that under high CO_2 concentration genes involved in oxidative metabolism are down-regulated whereas those involved in extracellular organic matrix synthesis are up-regulated. Long term exposure to high pCO_2 could impact the immune response of the coral and the organism may not have the ability to withstand external pressures. In another study, a plant *Arabidopsis thaliana* was grown under pCO_2 from 360-370 ppm and at 550 ppm (Miyazaki et al., 2004). Transcriptional changes were observed for the plants grown at two different conditions, for example the genes involved in stress response which were up-regulated. Studies of Aspen trees grown under 360 ppm and 560 ppm CO_2 concentrations for 12 years showed that a number of genes involved in the Calvin cycle, cell growth, cell division and hormone metabolism were up-regulated (Wei et al., 2013). The transcriptional changes resulted in an increase of biomass production in the trees grown under high pCO_2. In summary, investigation of the effect of high pCO_2 on the species just discussed have shown that CO_2 causes transcriptional changes as well as growth of the organism. Moreover, there is an indication that their ability to withstand additional stressors may be highly affected.

2.2 ILLUMINA mRNA SEQUENCING

The study of the transcriptome gives an indication of how changing the environment or various applied factors affects gene activity, thus the number of instruments used to study the transcriptome have been expanding. Illumina is one of the leading mRNA-seq instruments used in the field (Mardis, 2008; Ansorge, 2009; Xu et al., 2013). Other sequencing technologies include Roche/454, Life/APG and Helicos Biosciences. These next-generation sequencing instruments differ in template preparation, sequencing and data analysis (Metzker, 2010). In Roche/454 and Life/APG, emulsion PCR is used to amplify the template. In this method, after the cDNA is fragmented, adapters are attached to the DNA. The template is then denatured so that a single stranded DNA can anneal to DNA magnetic beads. An aqueous emulsion bounds each magnetic beads separately after which the template is amplified by PCR. The amplified templates are crosslinked to a glass surface in Life/APG or attached to PicoTiterplate wells in Roche/454 to proceed for sequencing.

In Illumina, with only a few basic steps, libraries of mRNA of interest can be prepared and run to look at the transcriptome changes under various conditions. Once mRNA is obtained, it is fragmented into several pieces and converted to cDNA (Figure 10). The cDNA fragment is further processed to prepare it for sequencing. The average fragment size is between 200 and 500 base pairs (Bronner et al., 2014). Adaptors, short nucleotide sequences, are first attached to both ends of the cDNA fragments so that they can easily be attached to flow cell channels, a glass slide with lanes and to serve as identification tags (Figure 11) (Shendure and Ji, 2008). The adapters that attach to the DNA contain different regions which aid in identification and sequencing of the fragment. One side of the adapter contains nucleotides complementary to the flow cell oligonucleotides, as well as nucleotides named as index 2, followed by a sequencing primer binding site 1. The other side has sequencing primer binding site 2, index 1, and nucleotides which are identical to the flow cell oligos. The indices are used to identify the DNA sequence in a future step.

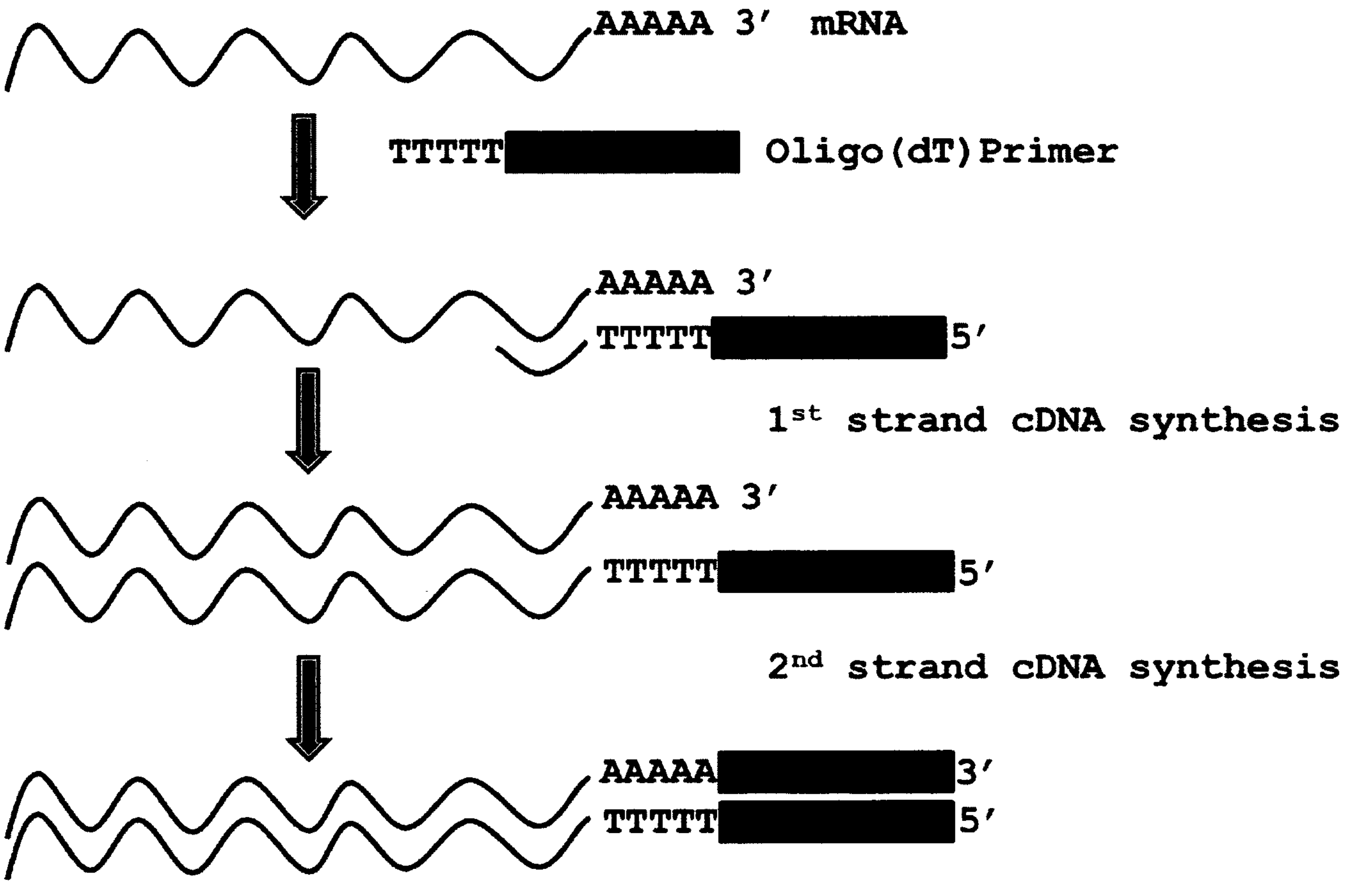


Figure 10. Conversion of mRNA to cDNA
The illustration shows how mRNA is converted to cDNA before the sample is run on the illumina. Information was adopted from (Wang et al., 2010).

The next step is increasing the concentration of the cDNA via cluster generation on the flow cell of the Illumina (Figure 12) (Metzker, 2010). The flow cell is coated with two types of oligonucleotides where one of the oligonucleotides is complementary to the adaptor that is attached to the cDNA. First, DNA fragments hybridize to the adapters on the flow cell. The strand then folds over and hybridizes to the second type of oligonucleotide on the flow cell.The next step is to create a cluster through bridge amplification where unlabeled nucleotides and enzymes are added to start the amplification (Fedurco et al., 2006; Bentley et al., 2008). With the strand as a template, the DNA polymerase and nucleotides are used to create complimentary stranded bridges. The double-stranded DNA denatures to create two complementary single strands. The process is repeated several times to create clusters for each fragment. Once the cluster is created, complementary strands are cleaved and washed off. The sequencing cycle starts by the addition of labeled reversible terminators, primers and DNA polymerase (Son and Taylor, 2011). The cluster is excited by a light source and fluorescence is emitted which identifies each nucleotide on the growing chain. All identical strands on the flow cell are read simultaneously.Once the sequencing is done, forward and reverse reads are paired creating contiguous sequences. The sequences are then aligned back to the reference genome and the level of transcriptome analyzed.

2.3 MATERIALS AND METHODS

The experiment was divided into three distinct parts. (1) *T. erythraeum* was grown under the two different pCO_2, and the C and N_2 fixation were studied in Dr. Muholland's lab by her Master's student Ivy Ozman at Old Dominion University (Section 2.3.1). (2) The RNA isolation, sequencing and initial bioinformatics analysis was done by Dr. Alice Hudder, Dr. Susan Lande and Dr. Adele Kruger at Wayne State University in collaboration with Dr. Lesley Greene (Section 2.3.2-2.3.3). (3) The metabolic analysis of the computational data was done by myself in Dr. Greene's lab at Old Dominion University (Section 2.3.4-2.3.5).

2.3.1 CELL CULTURE AND GROWTH

T. erythraeum isolate IMS101 was grown in YBCII media which is a specially made media that is suitable for this strain. The media was formulated on a

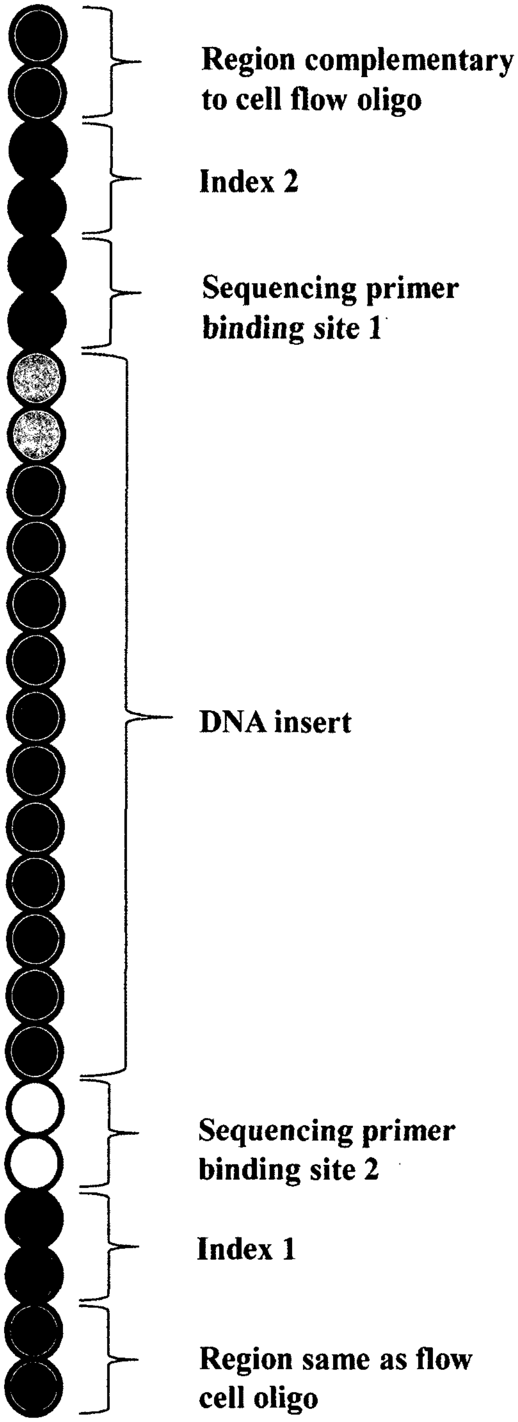

Region complementary to cell flow oligo
Index 2
Sequencing primer binding site 1
DNA insert
Sequencing primer binding site 2
Index 1
Region same as flow cell oligo

Figure 11. DNA Adaptors

This illustration shows the regions of the adaptor that is attached to the DNA fragments. Information was adapted from (Son and Taylor, 2011).

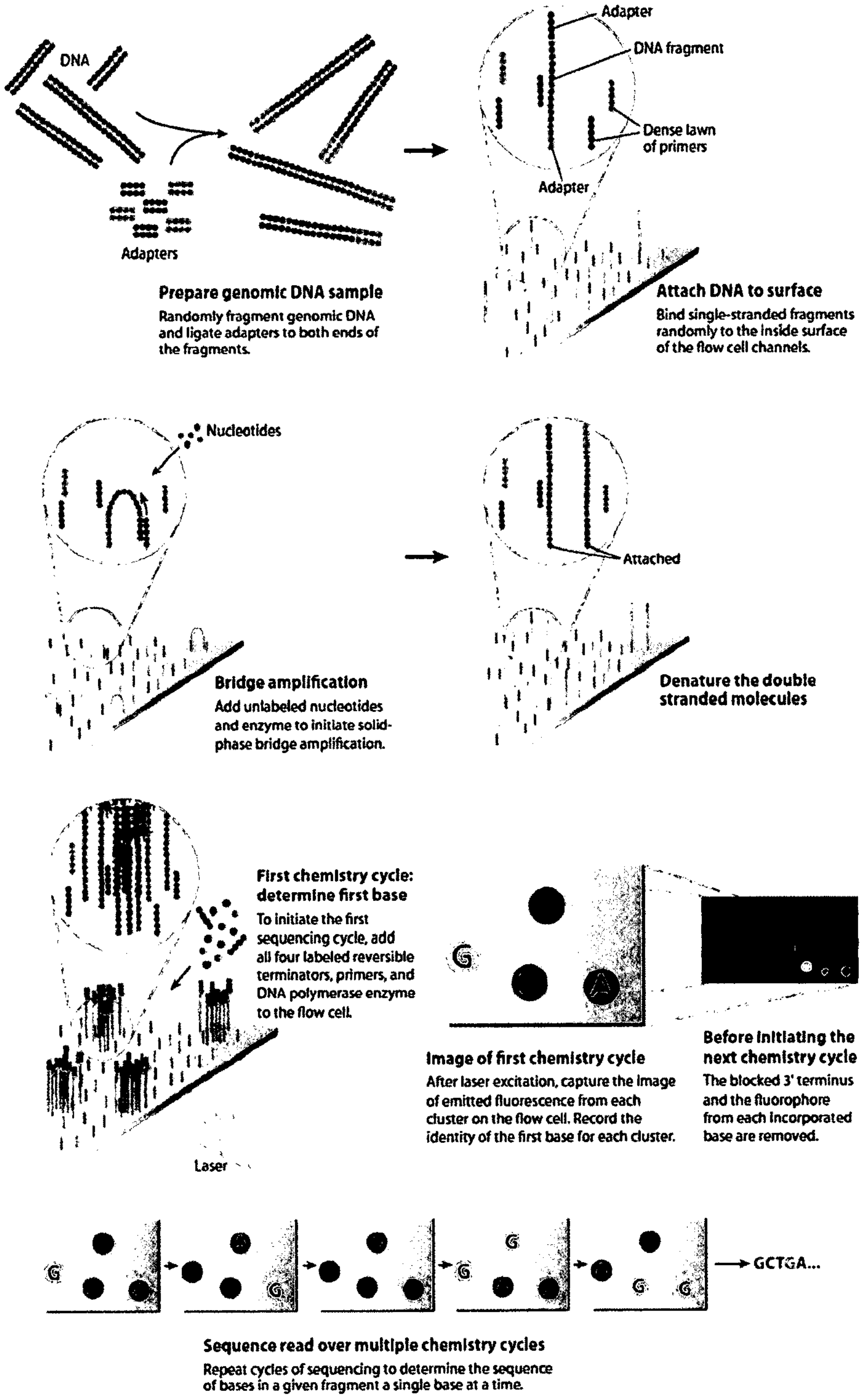

DNA
Adapters
Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
Adapter
DNA fragment
Dense lawn of primers
Adapter
Attach DNA to surface
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
Nucleotides
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Attached
Denature the double stranded molecules
First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.
Laser
Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.
Before initiating the next chemistry cycle
The blocked 3' terminus and the fluorophore from each incorporated base are removed.
GCTGA...
Sequence read over multiple chemistry cycles
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

Figure 12. Illumina Sequencing Approach

A genomic DNA sample is prepared by adding adapters. Cluster strands are created through bridge amplification. The sequence of the DNA is determined by detecting the fluorescently labeled nucleotides. This illustration is adapted from (Mardis, 2008).

medium designed by Ohki *et al* (Ohki et al., 1986). The composition of the media is similar to the content of tropical and subtropical water where *Trichodesmium* is naturally found.

The culture media was first filtered with 0.2 μm and pre-equilibrated to 386 ppm for present day CO_2 concentration and 750 ppm for projected CO_2 concentration for the year 2100. The cyanobacteria were grown at 30°C in a walk-in incubator. The concentration of CO_2 was maintained by bubbling custom air mixtures. Biomass of the cyanobacteria was checked by sampling an aliquot of the culture. The concentration of the inorganic carbon and the level of the pH was checked at each sampling point. The samples were filtered through Whatman GF/F glass fiber filters and frozen until mRNA sequencing. Further details of this work can be found in Ivy Ozman's Master's thesis entitled 'Implications of climate change for cyanobacteria over the Western Florida Shelf in the Gulf of Mexico', Old Dominion University, Norfolk, Virginia (2014).

2.3.2 mRNA SAMPLING

mRNA was purified from the total RNA by using a MICROBExpress bacterial mRNA purification kit (Ambion). The kit removes 16S and 23S ribosomal RNAs. The RNA is first incubated with the Capture Oligonucleotide Mix which contains a nucleotide that interacts with rRNAs. Then a nucleotide complimentary to the capture oligonucleotide and attached to a magnetic bead is added. A magnet is used to sequester all the ribosomal RNAs to the side of the container while the mRNA is left in the sample and can be pipetted out.

The purified mRNA was fragmented and cDNA was synthesized using SuperScript II (Invitrogen). End-repair of the fragments was performed with T4 DNA polymerase and Klenow (Illumina). This was followed by adenylation of the 3' ends in preparation for adapter ligation (Illumina). Adaptors with distinct sequences were added to the 3' and 5' end of the fragments and the samples that had the adaptors properly ligated were purified on Lonza's 1.2% agarose Recovery FlashGel (Basel, Switzerland). The Lonza system is unique in that the samples do not have to be cut out of the gel for purification. The gel has two tier wells, where the samples are loaded into the first well and after the run the sample of interest is recovered from the second well. The 150-200 bp fragments were recovered in 20μl recovery buffer (Lonza). Samples were then enriched with 15 cycles of a polymerase chain reaction (PCR). In the PCR reaction, the primers anneal to the adaptors that are found on the 3' and 5' end of the fragments. The libraries of cDNA that were generated were verified using a DNA 1000 chip on an Agilent Technologies 2100 Bioanalyzer. The bioanalyzer checks for the amount and size of each fragment. Fourteen picomoles were sequenced on Illumina's GAII sequencer. Each library was run on a separate lane. This experimental work was conducted in the Hudder laboratory and the Applied Genomics Technology Center at Wayne State University.

2.3.3 mRNA ANALYSIS

Novoalign (Novocraft Technologies, Malaysia) aligner was used to align the reads from the sequencer to the *T. erythraeum* IMS 101 genome. The number of reads mapping to each gene was determined. The level of expression for each gene was determined by taking the raw read count. The results were analyzed using two methods, with the second one being potentially more stringent than the first one. For the first analysis, genes that showed at least 40% change or greater between the samples grown under the different CO_2 concentrations were considered to be significant. In addition, a minimum of 10 reads in at least one of the two samples was required. For the second method of analysis, the number of total reads was normalized. Genes in both samples with a log-fold change between -1 and 1 were not considered to be differentially expressed. The genes that were above 1 or below -1 were considered to be up- or down-regulated, respectively. Thus, the data from the second analysis is different from that of the first one.

2.3.4 KEGG PATHWAY

The Kyoto Encyclopedia of Genes and Genomes (KEGG) was used to find the location of the transcribed genes in the metabolic pathway. KEGG is a bioinformatics database that consists of known genes and biochemical functions and contains drawn metabolic pathways using gene products (Kanehisa et al., 2012; Ogata et al., 1999). The pathway is organized in a hierarchal system where each level shows a detailed view of the pathway. The enzymes are identified based on their Enzyme Commission number (EC), where enzymes are classified based on the reaction they catalyze. Along with the EC number, KEGG also gives the locus tag of the gene used for transcription. The EC number and locus tag from the KEGG pathway was compared with the data from *T. erythraeum* gene expression to identify the location of the genes in a specific metabolic pathway. Based on the *T. erythraeum* gene database, the up- and down-regulated genes were identified and noted on KEGG pathways. The website address for KEGG is http://www.genome.jp/kegg/kegg2.html, where the metabolic paths are found under KEGG PATHWAY.

2.3.5 COG FUNCTIONS

Clusters of Orthologous Groups of proteins (COG) were used to classify the proteins in the sequenced genomes. COG categorizes the proteins according to their homologous relationship which predicts their biochemical activities (Tatusov et al., 1997, 2003). COG was used to categorize the up- and down-regulated genes to gain an idea of which genes are impacted by the high pCO_2. Information for COG functions can be found in http://img.jgi.doe.gov/w/doc/datasources.html.

2.4 RESULTS AND DISCUSSION

The study of the cyanobacteria transcriptome under various conditions opens an avenue to understand the effect of global warming. Elucidating how cyanobacterial gene expression changes in response to different factors provides a view of how carbon and nitrogen fixing organisms respond to those factors, which in turn, may affect different environmental cycles such as carbon.

Researchers have already determined that *T. erythraeum* growth rates remain the same or increase in a higher pCO_2 environment (Hutchins et al., 2007; Levitan et al., 2007). However, the whole gene expression picture of how *Trichodesmium* adjusts its gene expression to high pCO_2 was not understood. We compared our mRNA sequence data from *T. erythraeum* grown at present day and predicted future pCO_2 conditions to the *T. erythraeum* genome IMS101 (Markowitz et al., 2010), which contains 5,156 genes (7,750,108 bases) (Table 1). The data was analyzed in two ways: 1) statistical differences in gene expressions between the two pCO_2 concentrations were noted and 2) a normalization of the data was performed prior to comparison of the expression differences. The results with the different methods of analysis are discussed.

2.4.1 ANALYSIS 1

Approximately twenty-nine million reads were generated for each concentration of CO_2 (Table 1). Sequences were compared to known sequences in databases and analyzed for statistical significance. The genes had to show at least a 40% change between the two growth conditions to be further analyzed. The fold change was determined by dividing gene reads from 750 ppm count by 386 ppm count. In total there were 596 down-regulated genes (FC > 1.7) and 826 up-regulated genes (FC > 1.4) in the *T. erythraeum* cultures grown at 750 ppm pCO_2 compared to those grown at present day pCO_2 (Table 2). Considering only those genes represented by at least 10 reads, we identified 332 down-regulated and 484 up-regulated genes in the elevated pCO_2. Thus, a total of 816 genes were differentially affected by the change in CO_2 presence (Table 2).

Table 1. Run Statistics of RNA-Seq Analysis

	386 ppm pCO_2	**750 ppm pCO_2**
Number of reads	29,702,591	28,576,202
Reads aligned	14,892,388	11,531,763
Uniquely aligned	715,128	584,131
Mapped to 2-5 locations	14,173,880	10,944,520
Mapped to 6 or more locations	3,380	3,112
Failed QC	4,893,749	4,282,910

Of the 816 affected genes, 404 genes could be classified based on their COG functions (Figure 13). Of these, 161 genes were down-regulated and 243 genes were up-regulated. The majority of the genes (60%) were up-regulated, which suggests that these genes are aiding the survival of *T. erythraeum* at elevated pCO_2. Based on the gene expression data, the most affected by CO_2 concentration were involved in coenzyme metabolism (H), amino acid transport and metabolism (E) (Figure 14). The group of genes that were most down-regulated were those related to energy production and conversion (C)

(Figure 13). Genes involved in defense mechanisms (V) showed the least amount of change in gene expression with two genes down-regulated and one gene up-regulated (Figure 13). The highest level of fold changes were seen in the range of 0.45 to 0.55 while a small number of genes have a fold change range that is greater than 3 (Figure 15). This shows that the amount of transcriptional changes in *T. erythraeum* grown under the two pCO_2 is significant and the data gives us some insight into what happens to the cyanobacteria when the environment is modified.

Table 2. Annotation Statistics of RNA-Seq Analysis

Genes not represented	421
Genes mapped in 750 ppm pCO_2 but not 386 ppm pCO_2	158
Genes mapped in 386 ppm pCO_2 but not 750 ppm pCO_2	160
Down-regulated in 750 ppm pCO_2	596
Up-regulated in 750 ppm pCO_2	826
Not differentially regulated	2,995
TOTAL	5,156
Genes represented by at least 10 reads in at least one sample	816
Down-regulated in 750	332
Up-regulated in 750	484
Genes represented by at least 100 reads in at least one sample	97
Down-regulated in 750	60
Up-regulated in 750	37

Even though whole metabolic pathways were not completely up-regulated or down-regulated, certain genes in the pathways were differentially regulated so that cellular growth and carbon and nitrogen fixation are adjusted in high pCO_2.

mRNA-seq data showed that glucokinase and phosphoglycerate mutase, proteins involved in the breakdown of carbohydrates in the metabolic pathway of glycolysis, were down-regulated (Figure 16). *T. erythraeum* does not have a complete glycolytic pathway as it lacks the enzyme 6-phosphofructokinase used in the conversion of fructose-6-phosphate to fructose-1,6-bisphosphate common to most glycolytic pathways (Gomez-Baena et al., 2008; Pelroy et al., 1972). Thus, the pentose phosphate pathway may be utilized to break-down stored glycogen (Matthijs and Lubberding, 1988; Carpenter et al., 1992; Sandh et al., 2011). In this pathway, rather than conversion to fructose 6-phosphate, glucose 6-phosphate is converted to 6-δ phosphoglucono lactone to enter the pentose phosphate pathway. The substrates from the pentose phosphate pathway can then enter into the glycolytic pathway. Genes involved in the pentose phosphate pathway, glucose 6-phosphate dehydrogenase, lactonase and transaldolase, an enzyme which is involved in the interconversion of glyceraldehyde 3-phosphate and sedoheptulose 7-phosphate to erythrose 4-phosphate and fructose 6-phosphate respectively were also down-regulated (Figure 16). This change suggests that the breakdown of carbohydrates is slowed down under elevated pCO_2.

An enzyme in the Calvin cycle that plays an integral role in the incorporation of carbon by a cyanobacteria is ribulose-1,5- bisphosphate carboxylase/oxygenase (RuBisCO) (Chen and Spreitzer, 1989; Lin et al., 2014). RuBisCO converts ribulose-1,5 bisphosphate to 3-phosphoglycerate by assimilating CO_2 is up-regulated. The regulation of pentose-phosphate pathway and RuBisCO indicates that even though carbohydrate synthesis is increased the intracellular catabolism is slowed. The energy required for CO_2 fixation comes from ATP while NADPH originates from photosynthesis (Tang et al., 2011). However, the majority of the genes involved in the generation of ATP and NADPH were down-regulated in photosynthesis (Figure 17). Moreover, four genes are down-regulated and six genes are up-regulated in porphyrin and chlorophyll metabolism (Figure 18). This suggests that the amount of energy that is available for the fixation of CO_2 via the Calvin cycle may be reduced.

RuBisCO not only has an affinity for CO_2 but also for O_2. Because of the character of the enzyme, which binds indiscriminately to either O_2 or CO_2, cyanobacteria had to come up with method to overcome this problem. One way is

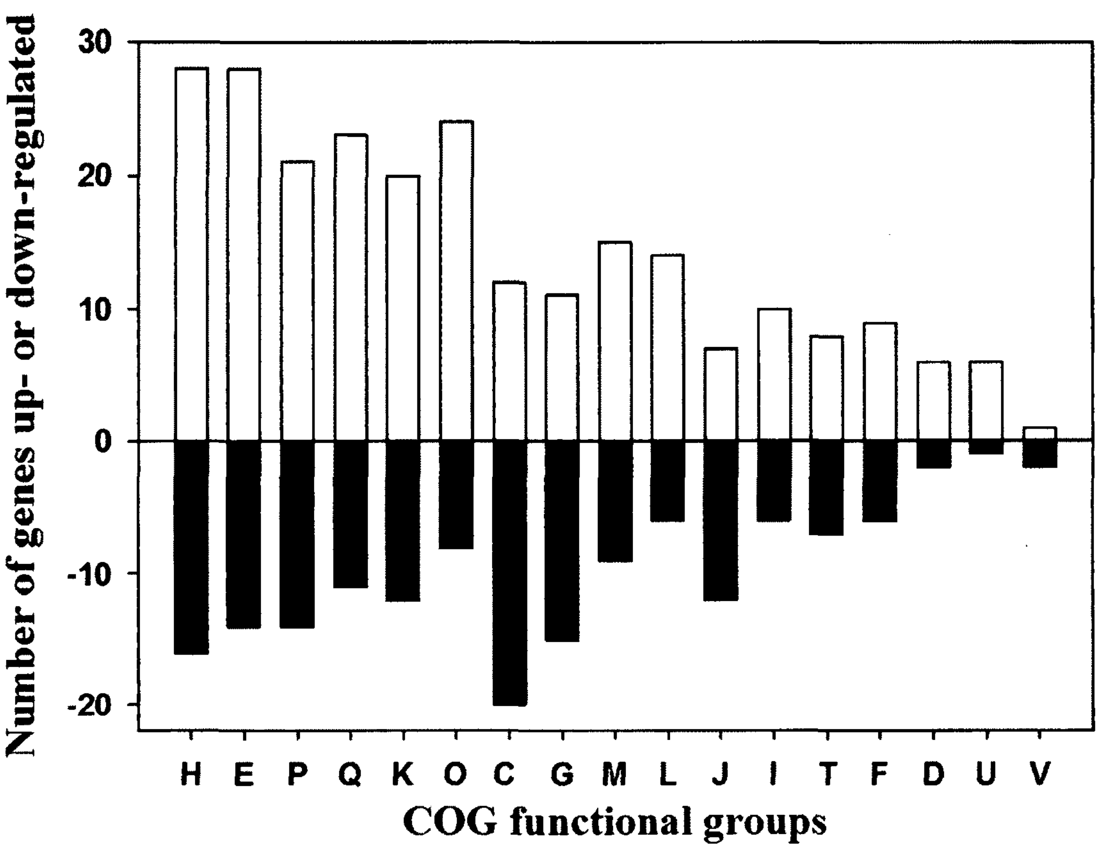


Figure 13. Analysis Method 1: Up- and Down-Regulation of *T. erythraeum* Gene Expression Results Organized by COG Functions

Up-regulated genes are shown in white bars and the down-regulated genes are shown as black bars. The codes and names of the COG functions are as follows: H=coenzyme metabolism; E=amino acid transport and metabolism; P=inorganic ion transport and metabolism; Q=secondary metabolites biosynthesis, transport and catabolism; K=transcription; O=posttranslational modification, protein turnover, chaperones; C=energy production and conversion; G=carbohydrate transport and metabolism; M=cell envelope biogenesis; L=DNA replication, recombination and repair; J=translation, ribosomal structure and biogenesis; I= lipid metabolism; T=signal transduction mechanisms; F=nucleotide transport and metabolism; D=cell division and chromosome partitioning; U=intracellular trafficking and secretion; V=defense mechanisms.

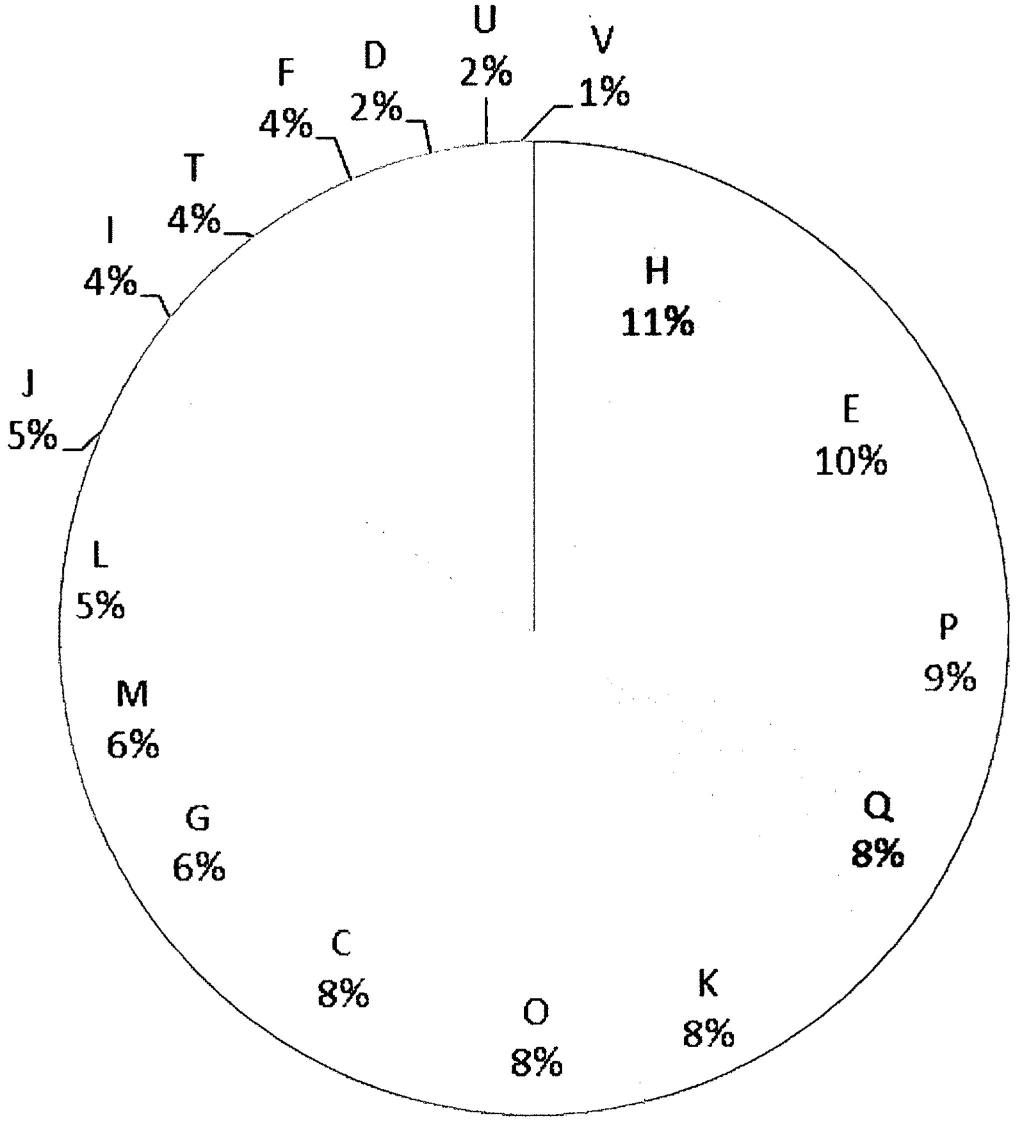


Figure 14. Analysis Method 1: Percentage of *T. erythraeum* Gene Expression Classified According to COG Functions

The codes and names of the COG functions are as follows: H=coenzyme metabolism; E=amino acid transport and metabolism; P=inorganic ion transport and metabolism; Q=secondary metabolites biosynthesis, transport and catabolism; K=transcription; O=posttranslational modification, protein turnover, chaperones; C=energy production and conversion; G=carbohydrate transport and metabolism; M=cell envelope biogenesis; L=DNA replication, recombination and repair; J=translation, ribosomal structure and biogenesis; I= lipid metabolism; T=signal transduction mechanisms; F=nucleotide transport and metabolism; D=cell division and chromosome partitioning; U=intracellular trafficking and secretion; V=defense mechanisms.

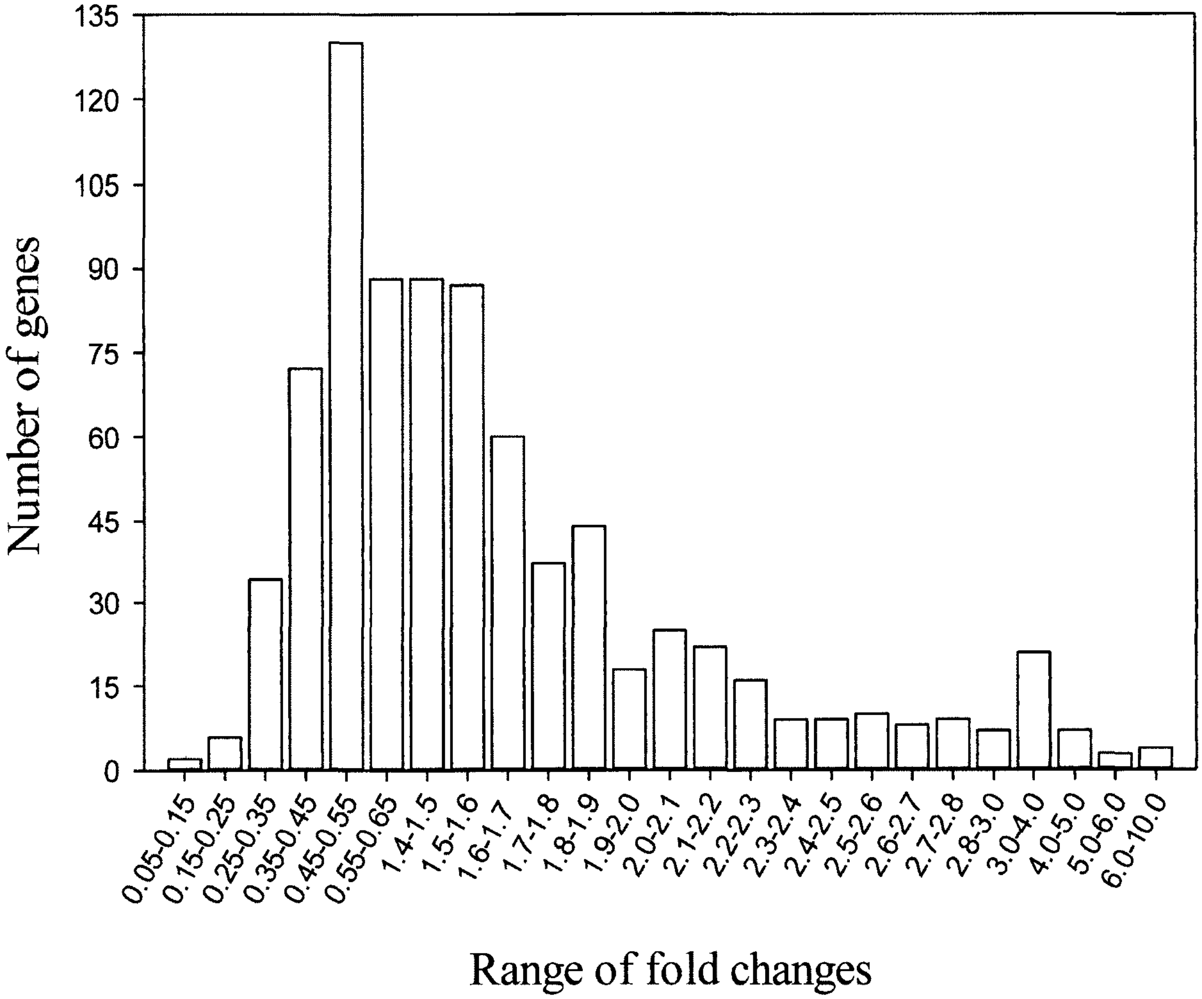


Figure 15. Regulated Genes Categorized by Transcriptional Activity
The genes are organized by fold changes indicating orders of magnitude different from the control.

through the CO_2 use of concentrating mechanisms (CCM) such as carboxysome, a compartment that contains RuBisCO with carboxysomal anhydrase and CO_2 and HCO_3^- transporters (Badger and Price, 2003). The up- or down-regulation of the genes involved in the pathway is unknown. However, the enzyme involved in the conversion of glycolate to glyoxylate, glyoxylate oxidase which is engaged by the O_2 is down-regulated (Figure 16). The result suggests that increased pCO_2 causes RuBisCO to preferentially employ CO_2 rather than O_2.

Gene expression differences were also observed for the tricarboxylic acid cycle (TCA). *T. erythraeum* was thought to have an incomplete tricarboxylic acid cycle as it lacks α-oxoglutarate dehydrogenase also known as α-ketoglutarate dehydrogenase (Pearce et al., 1969). However, recently research has shown two enzymes 2-oxoglutarate decarboxylase and succinic semialdehyde dehydrogenase are able to replace the function of α-oxoglutarate dehydrogenase (Zhang and Bryant, 2011). Two genes involved in the TCA cycle, fumarase, an enzyme that converts fumarate to malate, and succinyl-CoA synthase an enzyme which converts succinyl CoA to succinate, are up-regulated while citrate synthase, an enzyme that incorporates acetyl-CoA into the citric cycle, is down-regulated (Figure 16). The down-regulation of citrate synthase suggests that the amount of energy produced by *T. erythraeum* may be decreased. However, even though citrate synthase is down-regulated, valine, methionine, and isoleucine can be converted to succinyl CoA, thus allowing the cycle to continue. Moreover, under high CO_2 treatment, the methionine concentration is most likely increased as the enzyme which converts homocysteine to methionine, 5-methyltetrahydropteroyltriglutamate-homocysteine methyl transferase, is up-regulated (Figure 19).

The genes involved in oxidative phosphorylation that could be located are all down-regulated. For photosynthesis, most of the enzymes are down-regulated but four genes are up-regulated (Figure 17). Cyanobacteria use phycobilisomes as major light harvest complexes. Phycobilisomes associate with linker polypeptides that help to stabilize the structure so that the absorbance and energy transfer among phycobilisomes is maximized (Liu et al., 2005). PsbK protein is found in the photosystem II core and is not believed to be essential for the system's function, but it may enhance its activity. Research has shown that when the *psbK* gene was deleted in Synechocystis, the cyanobacteria grew more slowly (Ikeuchi et al., 1991). This suggests that the down-regulation of *psbK* does not affect *T. erythraeum* very much as the amount of cyanobacteria biomass is higher under the higher pCO_2 (Ozman, 2014).

The cytochrome b6f complex is part of the respiratory as well as the photosynthetic electron transport chain. It oxidizes plastoquinol and transfers electrons to either cytochrome c6 or plastocyanine (Schneider et al., 2007). Cytochrome b6f complex is composed of four major subunits with four additional small subunits. One of the small subunits *petG* is down-regulated under 750 ppm pCO_2. A study has shown that PetG is essential for the structure and function of cytochrome b6f complex (Schneider et al., 2007), thus the down-regulation of the gene in *T. erythraeum* may have a great effect on the photosynthetic ability of the cyanobacteria.

Another gene that is down-regulated that is involved in photosynthetic electron transport is *petJ*. A study showed that the deletion of this gene does not have an effect on photosynthesis in cyanobacteria (Kerfeld and Krogmann, 1998), which might suggest that the cyanobacteria down-regulates genes that are not essential for the survival of the organism. Under high light intensity, it is believed that photosystem II (PSII) produces excited oxygen species that oxidizes species and destroys PSII. It has been shown that PsbT protein that is part of PSII helps in the recovery of photodamaged PSII (Ohnishi and Takahashi, 2001). Even though the effect of light on *T. erythraeum* was not studied in our experiment, the down-regulation of *psbT* suggests that the cyanobacteria did not require it under our conditions. However, if the amount of light intensity were to change, cyanobacteria might not be able to recover from the photodamage.

PSII also contains cytochrome (cyt) c-550, which is encoded by *psbV*. It has been shown that a cyanobacteria with cyt c-550 can still produce oxygen even with the absence of other extrinsic proteins (Shen et al., 1995). This supports our data where *psbV* is up-regulated but other genes are down-regulated and the cyanobacteria still appears to function properly. PsbP is a component of PSII that is found on the thylakoid membrane. Even though it is believed that PsbP optimizes oxygen evolution under calcium and chloride, the exact function of PsbP in cyanobacteria is unknown with some groups showing that there is no effect with mutant *psbP* while others observing growth defects and decrease in oxygen evolution (Bricker et al., 2012). Under 750 ppm, *psbP* is down-regulated

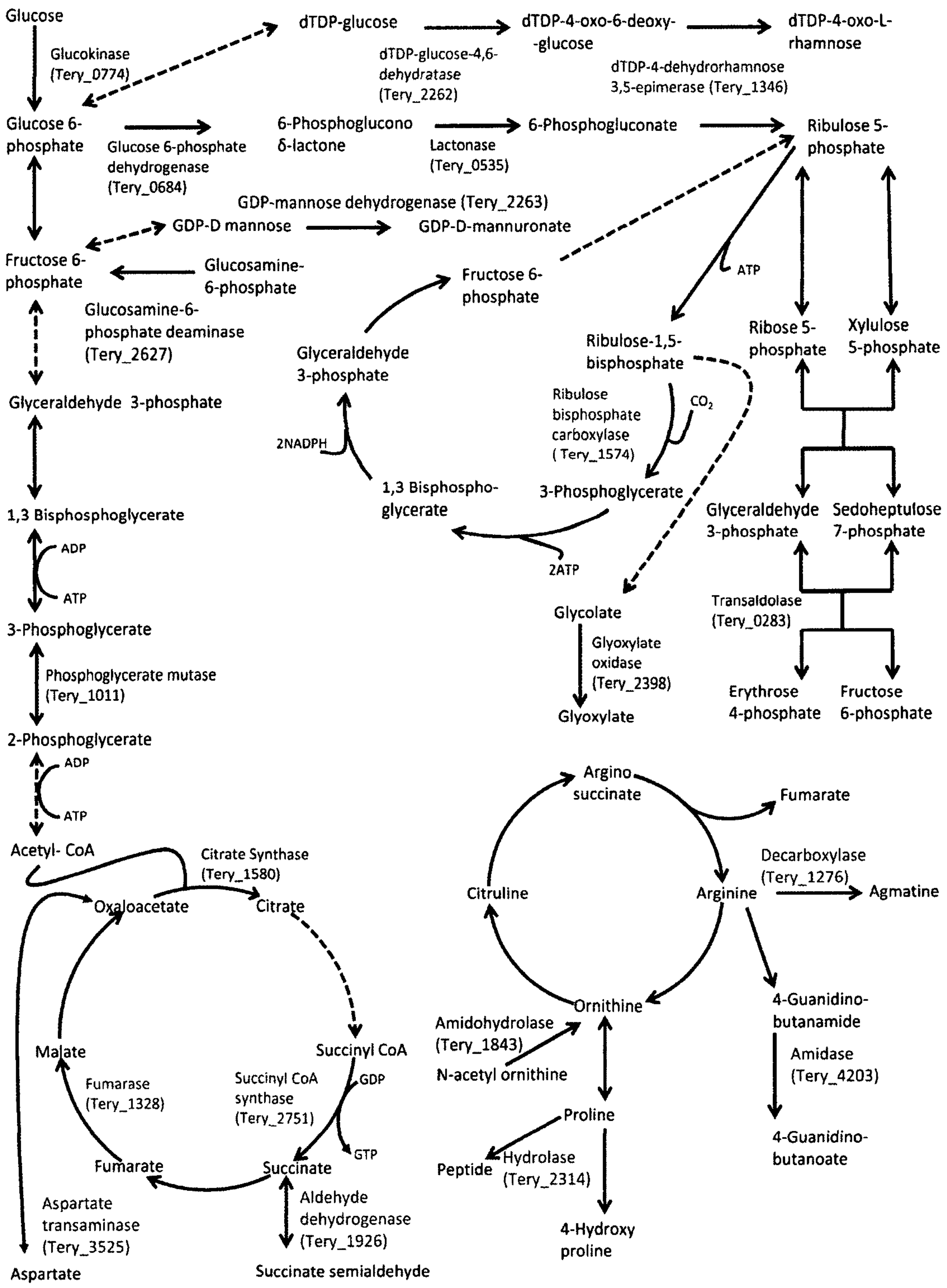


Figure 16. Metabolic Pathway: Glycolysis, TCA Cycle, Pentose Phosphate, Calvin Cycle, Urea Cycle

The red arrows show down-regulation and green arrows show up-regulation. Blue arrows show pathways that are not affected and broken arrows indicate that there are genes that are not shown in the figure.

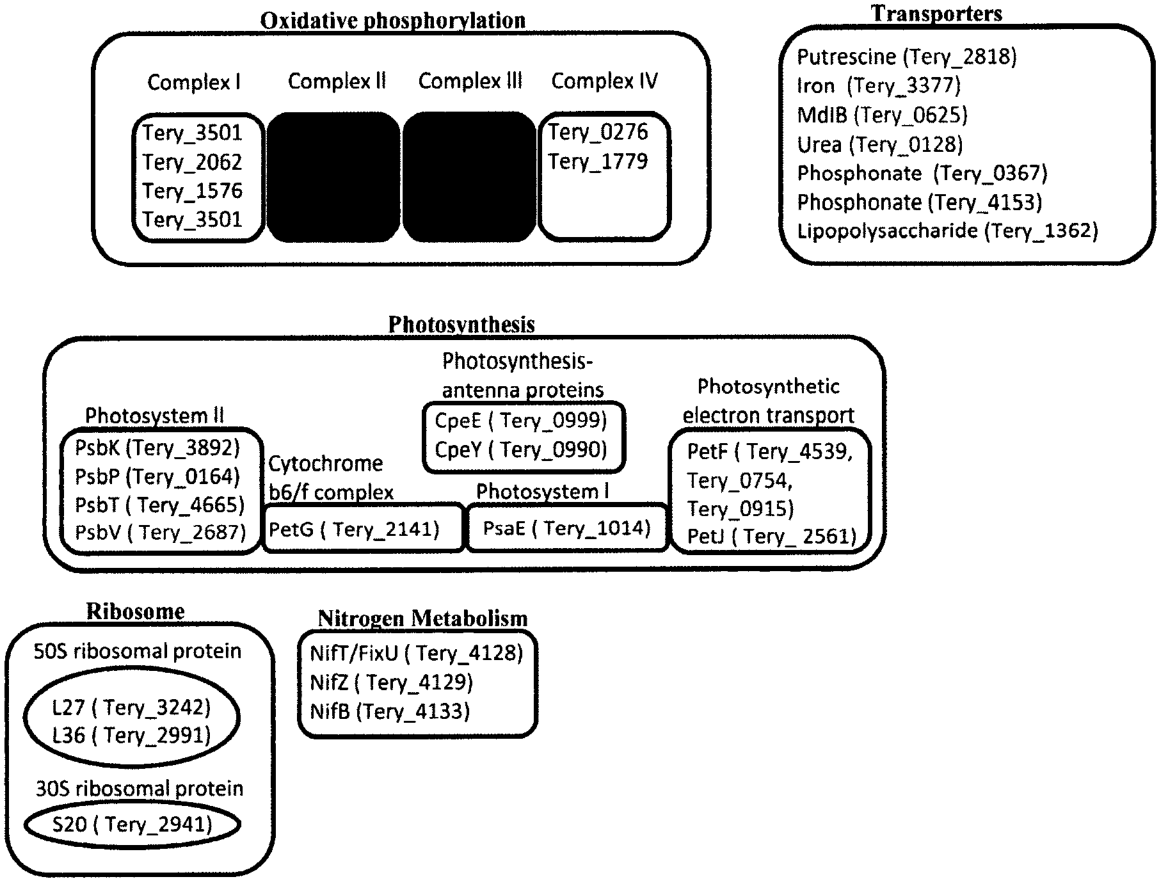


Figure 17. Metabolic Pathway: Oxidative phosphorylation, Transporters, Photosynthesis, Ribosome and Nitrogen Metabolism

Up-regulated and down-regulated genes in *T. erythraeum* are denoted as follows: red shows down-regulation and green shows up-regulation.

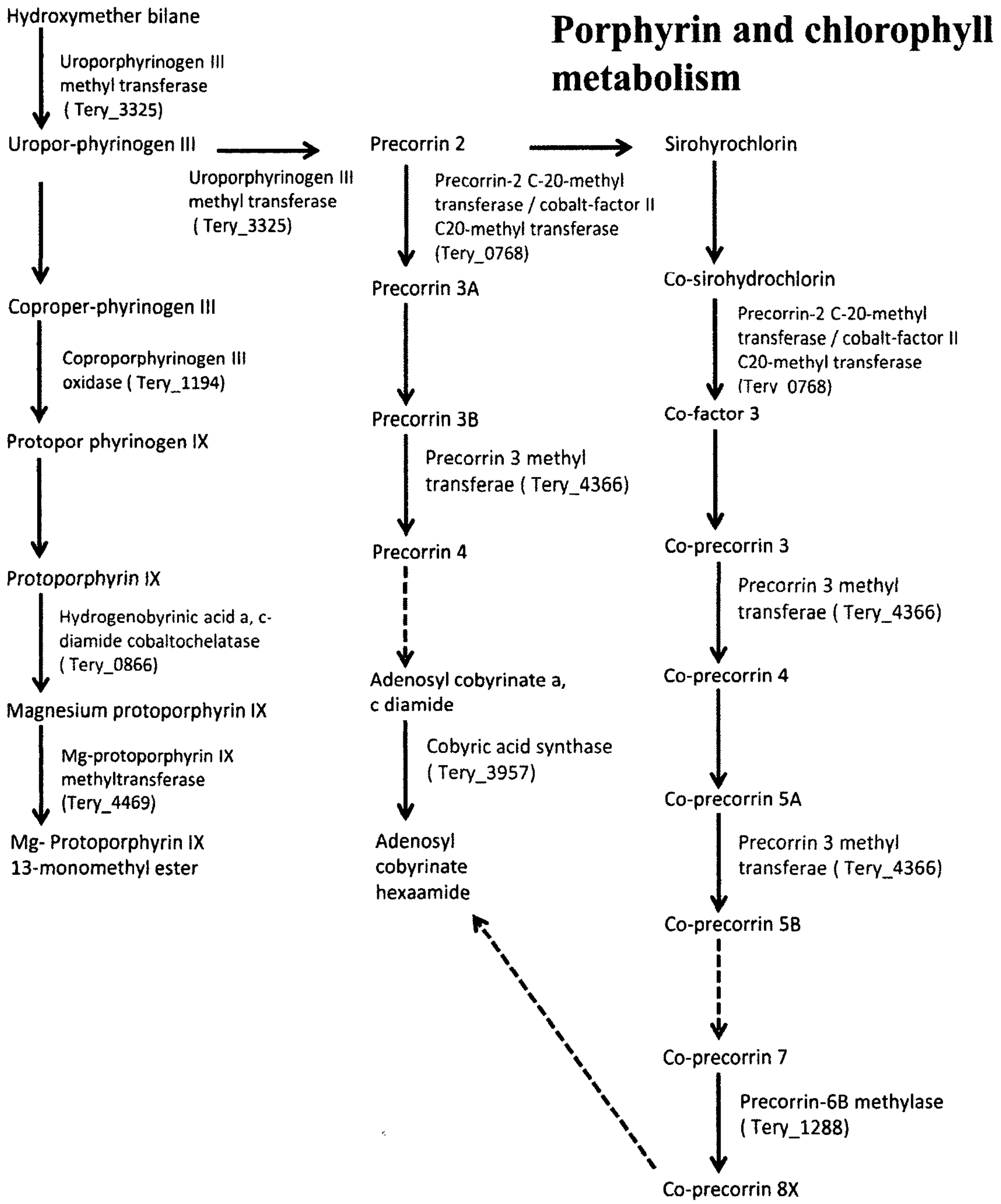

Porphyrin and chlorophyll metabolism
Hydroxymether bilane
Uroporphyrinogen III methyl transferase (Tery_3325)
Uropor-phyrinogen III
Uroporphyrinogen III methyl transferase (Tery_3325)
Precorrin 2
Sirohyrochlorin
Precorrin-2 C-20-methyl transferase / cobalt-factor II C20-methyl transferase (Tery_0768)
Precorrin 3A
Co-sirohydrochlorin
Coproper-phyrinogen III
Precorrin-2 C-20-methyl transferase / cobalt-factor II C20-methyl transferase (Terv 0768)
Coproporphyrinogen III oxidase (Tery_1194)
Precorrin 3B
Co-factor 3
Protopor phyrinogen IX
Precorrin 3 methyl transferae (Tery_4366)
Precorrin 4
Co-precorrin 3
Protoporphyrin IX
Precorrin 3 methyl transferae (Tery_4366)
Hydrogenobyrinic acid a, c-diamide cobaltochelatase (Tery_0866)
Adenosyl cobyrinate a, c diamide
Co-precorrin 4
Magnesium protoporphyrin IX
Cobyric acid synthase (Tery_3957)
Mg-protoporphyrin IX methyltransferase (Tery_4469)
Co-precorrin 5A
Mg- Protoporphyrin IX 13-monomethyl ester
Adenosyl cobyrinate hexaamide
Precorrin 3 methyl transferae (Tery_4366)
Co-precorrin 5B
Co-precorrin 7
Precorrin-6B methylase (Tery_1288)
Co-precorrin 8X

Figure 18. Metabolic Pathway: Porphyrin and Chlorophyll Metabolism Up-regulated and down-regulated genes of porphyrin and chlorophyll metabolism of *T. erythraeum*. Red arrows show down-regulation and green arrows show up-regulation. Blue arrows show pathways that are not affected and broken arrows indicate that there are genes that are not shown in the figure.

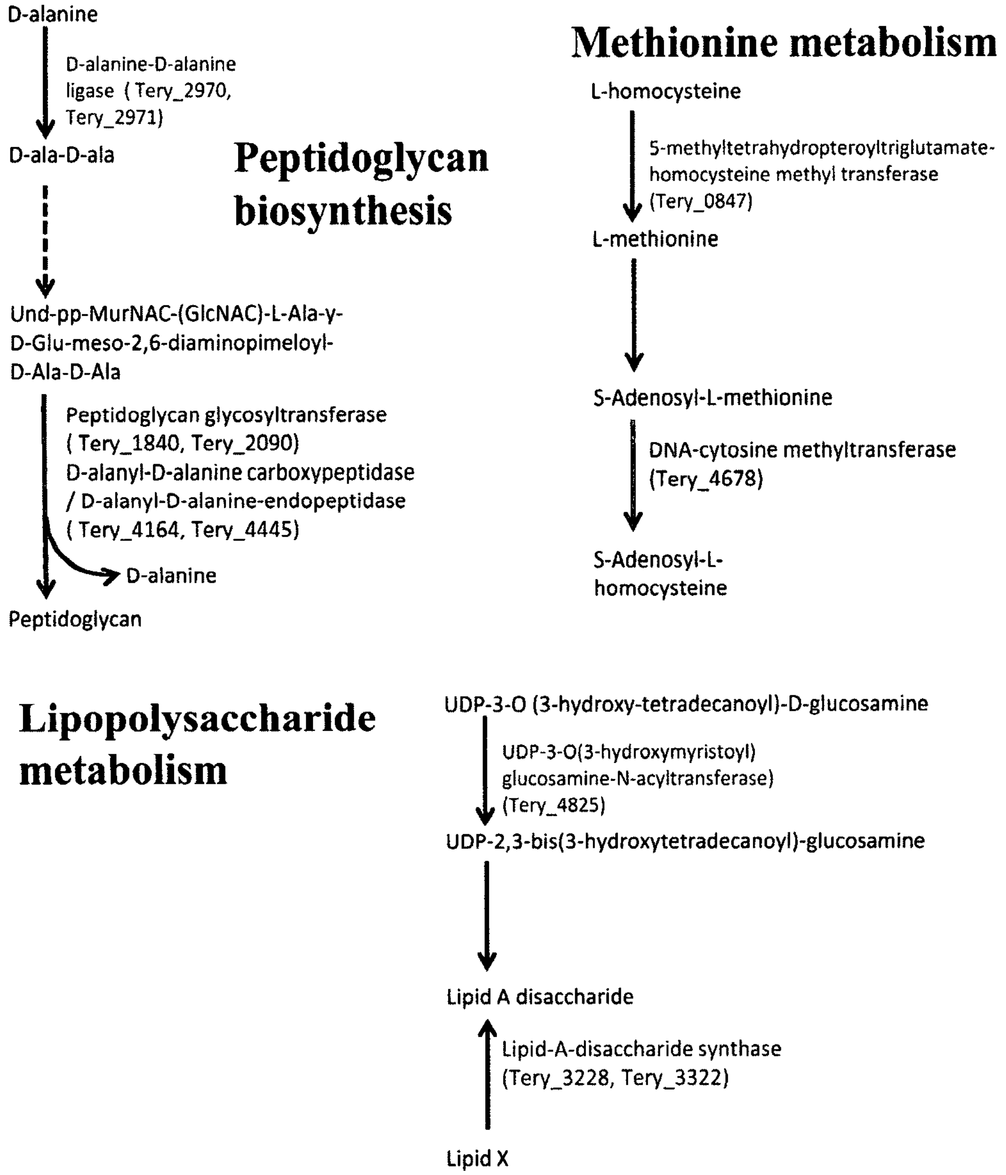

D-alanine
D-alanine-D-alanine ligase (Tery_2970, Tery_2971)
D-ala-D-ala
Peptidoglycan biosynthesis
Und-pp-MurNAC-(GlcNAC)-L-Ala-γ-D-Glu-meso-2,6-diaminopimeloyl-D-Ala-D-Ala
Peptidoglycan glycosyltransferase (Tery_1840, Tery_2090)
D-alanyl-D-alanine carboxypeptidase / D-alanyl-D-alanine-endopeptidase (Tery_4164, Tery_4445)
D-alanine
Peptidoglycan
Methionine metabolism
L-homocysteine
5-methyltetrahydropteroyltriglutamate-homocysteine methyl transferase (Tery_0847)
L-methionine
S-Adenosyl-L-methionine
DNA-cytosine methyltransferase (Tery_4678)
S-Adenosyl-L-homocysteine
Lipopolysaccharide metabolism
UDP-3-O (3-hydroxy-tetradecanoyl)-D-glucosamine
UDP-3-O(3-hydroxymyristoyl) glucosamine-N-acyltransferase) (Tery_4825)
UDP-2,3-bis(3-hydroxytetradecanoyl)-glucosamine
Lipid A disaccharide
Lipid-A-disaccharide synthase (Tery_3228, Tery_3322)
Lipid X

Figure 19. Metabolic Pathway: Peptidoglycan Biosynthesis, Methionine Metabolism, Lipopolysaccharide Metabolism

Up-regulated and down-regulated genes of peptidoglycan biosynthesis, methionine metabolism and lipopolysaccharide metabolism pathways in *T. erythraeum*. Red arrows show down-regulation and green arrows show up-regulation. Blue arrows show pathways that are not affected and broken arrows indicate that there are genes that are not shown in the figure.

which may suggest that the function of *psbP* may not be as essential as other genes.

It was found that the formation of nucleotides was potentially up-regulated under 750 ppm pCO_2. Four genes were found to be down-regulated and ten genes up-regulated in purine and pyrimidine metabolism (Figure 20). Even though it seemed that a number of genes were up-regulated in purine and pyrimidine metabolism, there were several enzymes that were involved in more than one reaction. As RuBisCO is up-regulated, the production of 3-phosphoglycerate is up-regulated which in turn increases ribose 5-phosphate, leading to the production of nucleotides. Production of uridine, cytidine, adenosine, thiamidine, and guanosine production was also potentially increased as the enzyme responsible, 5'nucleotidase was up-regulated (Figure 20).

We also wanted to look at how the gene transcriptional differences seen under the two pCO_2 concentrations would impact the overall conditions of the cyanobacteria. During all growth phases, the cultures grown under 750 ppm pCO_2 fixed more nitrogen and carbon than cultures grown under present day pCO_2 conditions. N_2 fixation rates increased 243%, 127% and 2% (Figure 21A), and carbon fixation rates increased 50%, 178% and 137% (Figure 21B) at 750 ppm pCO_2 for lag, exponential, and stationary growth phases, respectively (Ozman, 2014). Even though the cultures grown under 750 ppm pCO_2 fixed more nitrogen and carbon, the differences were not statistically significant ($p> 0.05$ for all growth phases) (Ozman, 2014). In cyanobacteria, there are sixteen N_2 fixation genes (nif) out of which eight are considered to be essential for the pathway: *nifE*, *nifH*, *nifD*, *nifK*, *nifU*, *nifB*, *nifK* and *nifS* (Latysheva et al., 2012). Of these, only *nifB* is down-regulated (Figure 17). *nifT* and *nifZ* are also down-regulated. However, a decrease in nitrogen fixation was not seen in the experimental data of the *T. erythraeum* culture grown under elevated pCO_2. Thus, the unaffected genes may compensate for those that are down-regulated.

Exponential growth rates were higher for the cultures grown at 750 ppm pCO_2, $\overline{\mu}_{386}$= 0.49 d-1 and $\overline{\mu}_{750}$= 0.62 d-1, than those grown at present day pCO_2 (Ozman, 2014). $\overline{\mu}$ symbolizes the growth of the cyanobacteria and d-1 is the unit for the growth rate. As RuBisCO is up-regulated under the 750 ppm of pCO_2, it suggests that the cyanobacteria are using CO_2 for growth. Cyanobacteria fixes CO_2 to produce glucose 6-phosphate. Through the pentose phosphate pathway

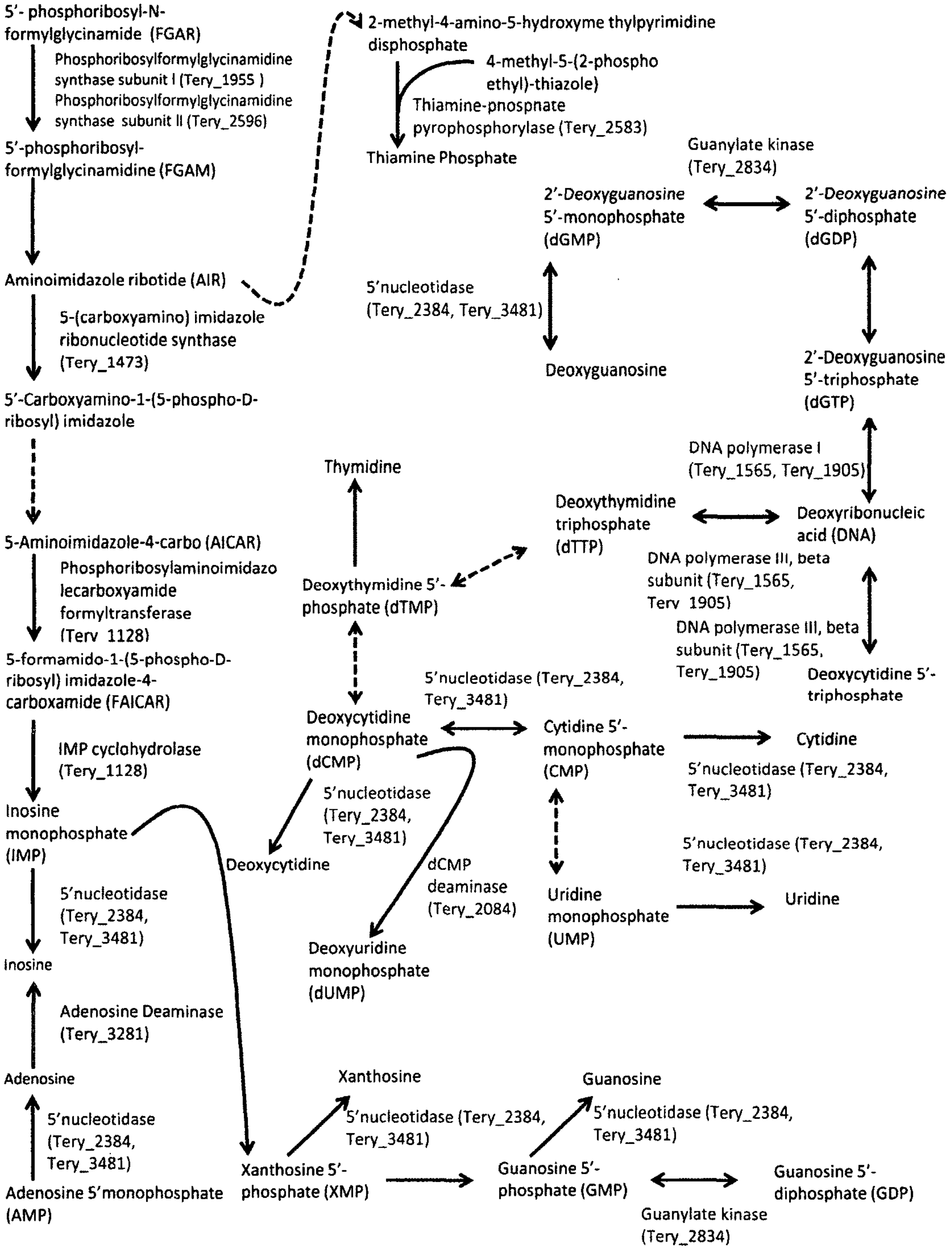

5'- phosphoribosyl-N-formylglycinamide (FGAR)
Phosphoribosylformylglycinamidine synthase subunit I (Tery_1955)
Phosphoribosylformylglycinamidine synthase subunit II (Tery_2596)
5'-phosphoribosyl-formylglycinamidine (FGAM)
Aminoimidazole ribotide (AIR)
5-(carboxyamino) imidazole ribonucleotide synthase (Tery_1473)
5'-Carboxyamino-1-(5-phospho-D-ribosyl) imidazole
5-Aminoimidazole-4-carbo (AICAR)
Phosphoribosylaminoimidazo lecarboxyamide formyltransferase (Terv 1128)
5-formamido-1-(5-phospho-D-ribosyl) imidazole-4-carboxamide (FAICAR)
IMP cyclohydrolase (Tery_1128)
Inosine monophosphate (IMP)
5'nucleotidase (Tery_2384, Tery_3481)
Inosine
Adenosine Deaminase (Tery_3281)
Adenosine
5'nucleotidase (Tery_2384, Tery_3481)
Adenosine 5'monophosphate (AMP)
2-methyl-4-amino-5-hydroxyme thylpyrimidine disphosphate
4-methyl-5-(2-phospho ethyl)-thiazole)
Thiamine-pnospnate pyrophosphorylase (Tery_2583)
Thiamine Phosphate
Guanylate kinase (Tery_2834)
2'-Deoxyguanosine 5'-monophosphate (dGMP)
2'-Deoxyguanosine 5'-diphosphate (dGDP)
5'nucleotidase (Tery_2384, Tery_3481)
Deoxyguanosine
2'-Deoxyguanosine 5'-triphosphate (dGTP)
DNA polymerase I (Tery_1565, Tery_1905)
Thymidine
Deoxythymidine triphosphate (dTTP)
Deoxyribonucleic acid (DNA)
DNA polymerase III, beta subunit (Tery_1565, Terv 1905)
DNA polymerase III, beta subunit (Tery_1565, Tery_1905)
Deoxythymidine 5'-phosphate (dTMP)
Deoxycytidine 5'-triphosphate
5'nucleotidase (Tery_2384, Tery_3481)
Deoxycytidine monophosphate (dCMP)
Cytidine 5'-monophosphate (CMP)
Cytidine
5'nucleotidase (Tery_2384, Tery_3481)
5'nucleotidase (Tery_2384, Tery_3481)
Deoxycytidine
dCMP deaminase (Tery_2084)
5'nucleotidase (Tery_2384, Tery_3481)
Uridine monophosphate (UMP)
Uridine
Deoxyuridine monophosphate (dUMP)
Xanthosine
Guanosine
5'nucleotidase (Tery_2384, Tery_3481)
5'nucleotidase (Tery_2384, Tery_3481)
Xanthosine 5'-phosphate (XMP)
Guanosine 5'-phosphate (GMP)
Guanosine 5'-diphosphate (GDP)
Guanylate kinase (Tery_2834)

Figure 20. Metabolic Pathway: Nucleotide Metabolism

Up-regulated and down-regulated genes of different metabolic pathways in *T. erythraeum*. Red arrows show down-regulation and green arrows show up-regulation. Blue arrows show pathways that are not affected and broken arrows indicate that there are genes that are not shown in the figure.

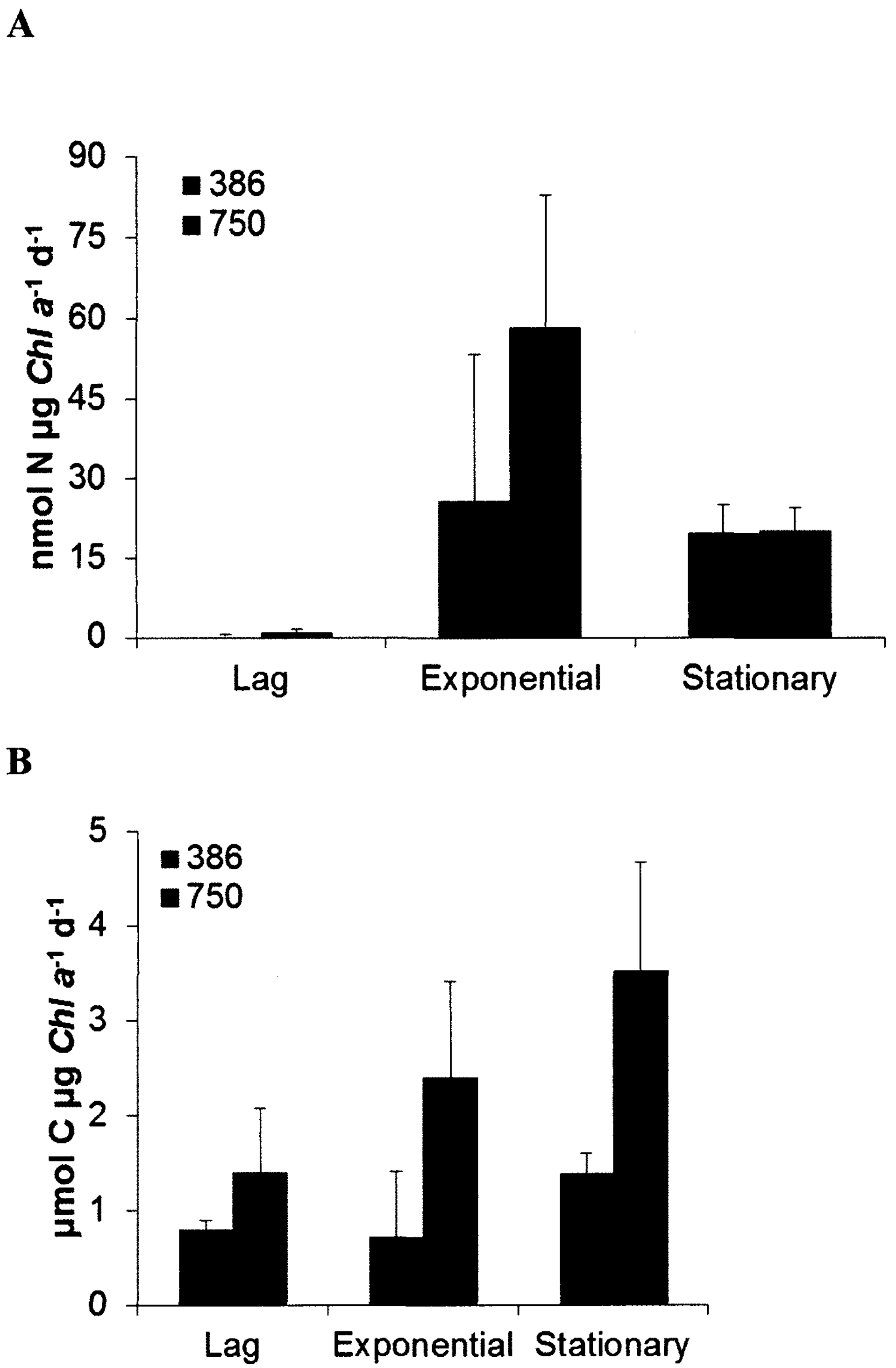


Figure 21. Average Nitrogen Fixation and Carbon Rates for *T. erythraeum*. Rate of change during lag, exponential and stationary growth phases acclimated to ambient (386 ppm pCO2) and 750 ppm pCO_2 (n=3). Figure adapted from (Ozman, 2014).

glucose 6-phosphate is broken down into glucose 5-phosphate and NADPH which are used for nucleotide biosynthesis (Thompson et al., 2011). However, the protein involved in the assembly of a peptidoglycan is down-regulated under high CO_2 concentration which may show that the integrity of the cell is weakened under high concentration. Moreover, an enzyme, DNA-cytosine methyltransferase, which converts S-adenosyl-L methionine to S-adenosyl L-homocysteine, and is involved in the protection of the cell from self-degradation through methylation is also down-regulated. This data suggests that even though the cyanobacterial biomass, number of cells, is increasing (Ozman, 2014), the ability of the *T. erythraeum* to protect the cell from external factors may be weakened under high pCO_2.

2.4.2 ANALYSIS 2

Under analysis 2 parameters, only 470 genes were up- or down-regulated with 268 genes up-regulated and 202 being down-regulated. Out of the 470 genes, only 101 genes had identifiable COG functions, with 55 genes up-regulated and 46 down-regulated. When we look at the distribution of genes using COG functions (Figure 22), we can see that there are certain categories where the genes are all up- or down-regulated. Genes in the energy production and conversion (C), cell division and chromosome partitioning (D), lipid metabolism (I), function unknown (S) and intracellular trafficking and secretion (U) catagories are all up-regulated; while genes in the translation, ribosomal structure and biogenesis (J) and defense mechanisms (V) categories are down-regulated. The results suggests that there are certain groups of proteins that are affected by high pCO_2 compared to others.

The distribution of genes using the COG function is also different compared to the first analysis (Figure 23). In the second analysis, the highest percentage of altered genes (19%) is in the category R, general function prediction only. In this section, the function of the proteins are unknown, and a general biochemical activity is given based on similar proteins. Under the first analysis, hypothetical proteins whose functions were unknown were not taken into consideration. The other categories with high percentage are DNA replication, recombination and repair (L) at 10%, coenzyme metabolism (H) at 8% and amino acid transport and metabolism (E) also at 8%. Thus, if the genes whose functions are unknown are

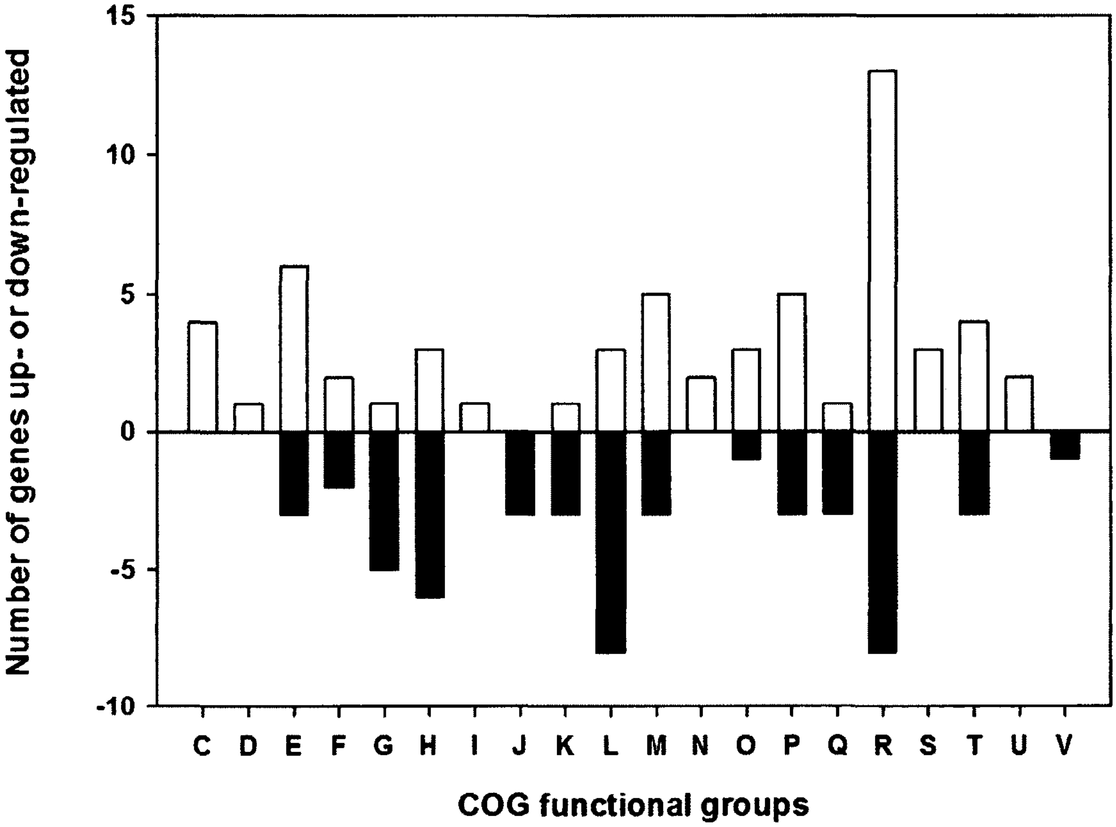


Figure 22. Analysis Method 2: Up- and Down-Regulated *T. erythraeum* Gene Expression Results Organized by COG Functions

The codes and names for the COG functions are as follows: C=energy production; D=cell cyclecontrol, cell division, chromosome partitioning; E=amino acid transport and metabolism; F= nucleotide transport and metabolism; G=carbohydrate transport and metabolism; H=coenzyme metabolism; I= lipid metabolism; J=translation; K=transcription; L=DNA replication, recombination and repair; M=cell envelope biogenesis; N=cell motility transport and metabolism; O=post-translational modification; P=inorganic ion transport and metabolism; Q=secondary metabolites biosynthesis, transport and catabolism; R=general function prediction only; S=function unknown; T=signal transduction mechanisms; U=intracellular trafficking and secretion; V=defense mechanisms.

removed, those in groups H, E and L would be the most effected which gives the same result as analysis 1.

We were unable to draw a condensed metabolic pathway as the genes were seen in various places performing a variety of functions. The genes that were up-regulated were mainly involved in cell-maintenance, energy production, ATP-binding cassette transporters, vitamin production and transferases. This suggests that under high pCO_2, *T. erythraeum* is up-regulating genes that are essential for survival but at the same time genes that are involved in replication and transcription, biosynthesis pathway are down-regulated. This alteration may weaken the integrity of the cell. These results might indicate that if the organism is subjected to other pressures such as pH and temperature changes, the cyanobacteria might not be able to survive.

2.5 CONCLUSION

Our experiment clearly shows that high pCO_2 has an effect on gene expression in *T. erythraeum*. Based on the first analysis parameters, gene expression changes on main metabolic pathways were seen. The main enzyme responsible for the incorporation of CO_2, RuBisCO is up-regulated, which indicates that the cyanobacteria takes in a higher amount of CO_2 compared to the control. The result suggests that CO_2 is being broken down for the growth of *T. erythraeum*, which is confirmed in the growth rates. The growth is also reflected in the purine and pyrimidine metabolism pathways, where the majority of the enzymes are up-regulated. Even though a second rigorous parameter was applied, the cyanobacteria were still observed to adjust to the high pCO_2 by regulating its genes. According to our data, *T. erythraeum* would not only survive, but may potentially thrive in the year 2100 under high pCO_2 if no other pressures are applied. However, we know that there are many factors other than CO_2 that might affect the organism. Whether or not *T. erythraeum* would survive additional changes to its environment such as pH needs to be studied in the future. How this scenario would affect the ecosystem is yet to be determined.

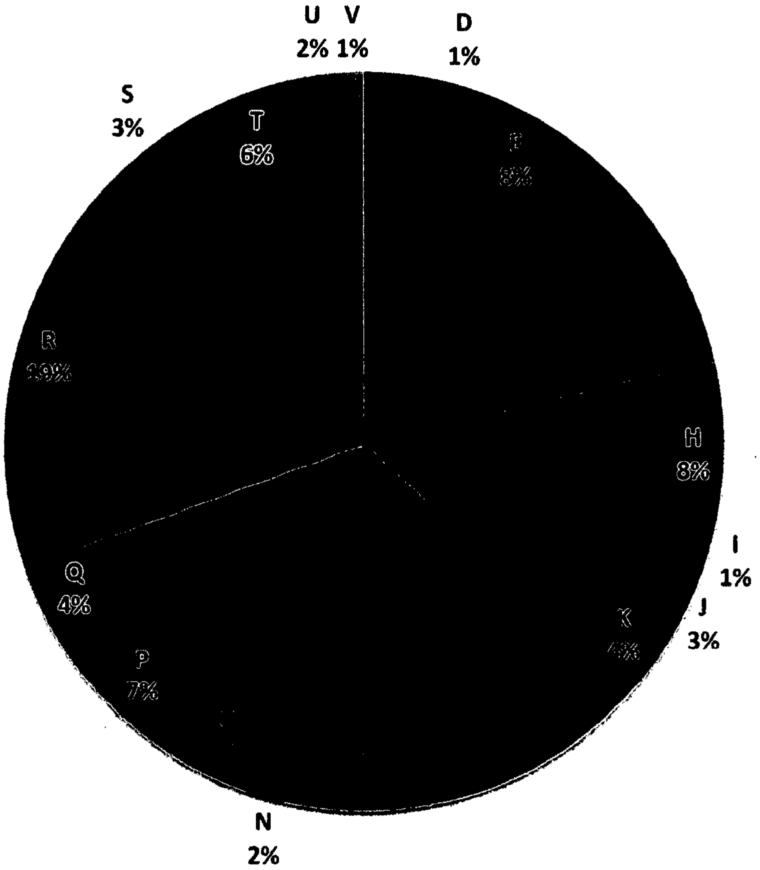


Figure 23. Analysis Method 2: Percentage of *T. erythraeum* Gene Expression Results for Genes Organized by COG Functions

C=energy production; D=cell cyclecontrol, cell division, chromosome partitioning; E=amino acid transport and metabolism; F= nucleotide transport and metabolism; G=carbohydrate transport and metabolism; H=coenzyme metabolism; I= lipid metabolism; J=translation; K=transcription; L=DNA replication, recombination and repair; M=cell envelope biogenesis; N=cell motility transport and metabolism; O=posttranslational modification; P=inorganic ion transport and metabolism; Q=secondary metabolites biosynthesis, transport and catabolism; R=general function prediction only; S=function unknown; T=signal transduction mechanisms; U=intracellular trafficking and secretion; V=defense mechanisms.

CHAPTER 3

BIOINFORMATICS ANALYSIS OF THE LIPOCALIN SUPERFAMILY

3.1 BACKGROUND

Proteins are the basis of life and the study of their structure and function can provide an insight into the inner workings of an organism. Historically, biological features of proteins have been studied experimentally. However, as the number of proteins being discovered have increased exponentially, the elucidation of protein structure and function through experiment has not been able to keep pace (Moult et al., 2014). Bioinformatics is the study of biological systems using computational approaches, to fill in the gaps in our knowledge, and direct future experimental studies (Lesk, 2002). The structure, folding and stability of a protein can be predicted using computers before the analysis is tested out experimentally. These studies are usually conducted by using the information gathered from related proteins.

Relationships among proteins can be established through evolutionary trees. Evolution is one way to learn about how proteins changed over time and achieved their current state. Through natural selection, organisms were chosen as a result of certain traits they possess to overcome environmental pressures (Patthy, 2008). In addition, evolution can occur via genetic drift, which is a random change in genetic information. The genetic difference could be the outcome of mutations, recombination, gene duplication or gene loss. In a protein, evolution points out the amino acids that may be essential for the structure, folding and stability as they are conserved in different organisms (Greene et al., 2003).

One of the ways to analyze a protein is by looking at its 3D structure. Protein 3D structures are more conserved than their sequences and functions, which makes structure a great tool for studying proteins (Sikosek et al., 2012; Ingles-Prieto et al., 2013). With respect to functions, either various protein folds can perform the same function or one particular fold can perform different tasks.

For example the proteins of the lipocalin superfamily have a common β-barrel fold but performs diverse functions.

One question bioinformatics attempts to answer is how kinetic and thermodynamic equilibrium drive proteins to fold to a given 3D structure. This is specially true in a superfamily of proteins where the sequence identity is low but still results in the same structure. Thus, the identification of the underlying forces that lead to folding is important. The protein folding problem looks at what governs the folding of a protein from the primary structure to its tertiary state (Dill and MacCallum, 2012). "Levinthal's Paradox" states that if a protein has 100 amino acids and each amino acid has three possible conformations, then there exists 3^{100}= $5x10^{47}$ different possible configurations for the protein (Zwanzig et al., 1992). If a protein randomly tries one conformation every 10^{-4} nanosecond, it will take 10^{27} years to cover the entire conformational space. This indicates that proteins may have a particular folding pathway that energetically directs them toward their preferred fold. Three main protein folding mechanisms have been proposed: the hydrophobic collapse model, the framework model and the nucleation-condensation model. The hydrophobic collapse model is based on the formation of a hydrophobic core by the association of nonpolar amino acids which is followed by the formation of the 3D structure (Dill et al., 1995). In the framework model, α-helices and β-structures are formed first which then associate to form the 3D and secondary structures (Baldwin, 1989; Ptitsyn et al., 1990; White et al., 2005). In the nucleation-condensation model, a small structural unit is formed, from which the nucleus grows to form the native structure (Fersht, 1995, 1997).

The nucleation-condensation folding mechanism may be used to explain how the 3D structure and amino acid residues are conserved evolutionarily. In this model certain amino acid residues interact to form the nucleus and as such they may be conserved in proteins with the same fold. The identification of the residues that form the essential interactions may give an indication about the kinetics and thermodynamics of protein folding. This type of analysis lets us look into the different types of superfamilies to understand which amino acids are essential for folding as well as look into how to engineer proteins to fold to a certain geometry.

We were interested in looking at the conservation of amino acids and folding of the lipocalin superfamily (Pervaiz and Brew, 1987). Even though the superfamily has been known since the mid 1980's, it was only in the 2000's that the protein was discovered in wheat which may give the plant an ability to withstand various temperatures (Charron et al., 2002, 2005). The ability of this protein to aid crops in withstanding stress made it a fascinating study. Understanding how Tatil folds and maintains its stability may also give an insight into how this protein can be applied to grow other crops at various temperatures. We used bioinformatics to locate the key determinants of structure, stability and folding by finding the conserved amino acids between Tatil and select related lipocalins. We also constructed a 3D model of Tatil as the structure of the protein has not been experimentally elucidated.

3.1.1 BASIC LOCAL ALIGNMENT SEARCH TOOL

The first step in gaining information about a protein for which the structure is unknown is to find homologous proteins. The inquiry can be accomplished by searching for other proteins that have high sequence identity with the protein of interest. One of the best tools used for homology searching is the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1997). Homology defines the relationship between two proteins when they arise from a common ancestor (Fitch, 2000). BLAST compares a protein or DNA sequences with a database and finds a region of similarity by constructing an alignment. Sequence alignments are classified into global and local alignments (Notredame et al., 2000). In the global alignment, two sequences are aligned to improve their overall alignment even though there might be stretches of low similarity. On the other hand, a local alignment looks for regions with the highest identity or similarity by sacrificing the overall alignment. BLAST uses local alignment and has a mathematical matrix for the sequences that are being compared. Positive scores are given for identical or conserved residues while a negative score is given for improbable substitutions. BLAST searches its database for stretch of segments that is similar to the aligned template sequence. Then the program checks if the alignment has a high enough score to be considered significant. Based on the scores, proteins similar to the input sequence are listed. The generated index can also be used to search for distantly related proteins. In our study, BLAST was used to find a protein whose structure has been experimentally solved and has a high sequence identity with Tatil.

3.1.2 PROTEIN DATA BANK

The Protein Data Bank (PDB) is a depository for structures of proteins that are gained through various techniques: X-ray crystallography, NMR spectroscopy, cryoelectron microscopy and theoretical modeling (Bernstein et al., 1977; Berman et al., 2000; Rose et al., 2013). Proteins placed in the PDB have a special identification so that they can be retrieved easily. The PDB code is alphanumeric which generally contains three or two letters and a number. The code has no special reference and any combination can be chosen as long as that identification number has never been used.

For each protein structure, certain information are found in the file irrespective of how the structure was gained. The file contains: source of the protein such as genus, species and variant of gene, the sequence of the protein, chemical structure of cofactors and prosthetic groups, name of all components of the structure, description of the characteristics of the structure, literature citations for the submitted structure and the 3D coordinates of the structure in the form of X, Y and Z. The PDB was used to obtain the structures of proteins that were related to Tatil.

3.1.3 STRUCTURE ALIGNMENT

A structural alignment is the comparison of structures based on the superposition of their 3D structures (Hasegawa and Holm, 2009). The structural alignment can provide information that cannot be gained from a sequence alignment alone, such as function. As 3D structures are more conserved than sequences, topology can be used to identify proteins from the same ancestor, even if the proteins have low sequence identity. The Dali server is a program that was used to identify and compare the lipocalin superfamily 3D structures (Holm and Rosenström, 2010). It contains the database PDB90, which contains all of the structures found in the PDB, except no two proteins have more than 90% sequence identity. Using the query protein, Dali looks for similar proteins based on the structures. Once the program finds approximately 100 structures, the generated list is used to find other related proteins. The result of a Dali study is a list of proteins, aligned both sequentially and structurally with the query protein.

3.2 HOMOLOGY MODELING

In the absence of a 3D structure, a model of the desired protein can be constructed under the right circumstances. In our case, since the structure of Tatil is unknown, we set out to build a model of the protein using various programs. We used homology modeling to solve the structure where a protein that is a close relative and has a high sequence identity to the one whose structure we want to solve is located.

Homology modeling is based on two main theories (Bourne and Weissig, 2003). The first is that the structure of a protein relies on the amino acid sequence (Anfinsen, 1973). The second is that the 3D structure is more evolutionarily conserved than amino acid sequences, thus proteins with high sequence identity should have the same fold pattern (Chothia and Lesk, 1986; Grishin, 2001). Homology modeling gives the best result when the relationship between the protein of interest and the related protein fall within the 'safe homology modeling zone' (Figure 24) (Sander and Schneider, 1991). The zone is determined by looking at the percentage of identical residues versus the number of aligned residues between the two sequences. If the percentage of identical residues and the number of aligned residues is low, the alignment falls in the 'twilight zone' and the likelihood that the model generated would be a true representative of the structure decreases. On the other hand, if the percentage of identical residues and number of aligned residues is high, it falls in the 'safe zone' and the model can be trusted.

Even though several kinds of homology programs such as SWISS-MODEL (Biasini et al., 2014) (Biasini et al., 2014) and MODELLER (Eswar et al., 2006) exist, the programs have similar multistep processes. The steps include the following: template recognition, initial alignment, alignment correction, backbone generation, loop modeling, side-chain modeling, model optimization and model validation (Figure 25) (Schwede and Peitsch, 2008; Szilagyi and Zhang, 2014).

In template recognition and initial alignment, a program is used to identify proteins with high sequence identity to the query protein. From these, one protein is chosen and aligned. The second step is alignment correction where the result from the first step is streamlined to give a better alignment (Venselaar et al., 2010). One method involves performing multiple sequence alignments of related proteins so that a better idea about evolutionary insertions and deletions is

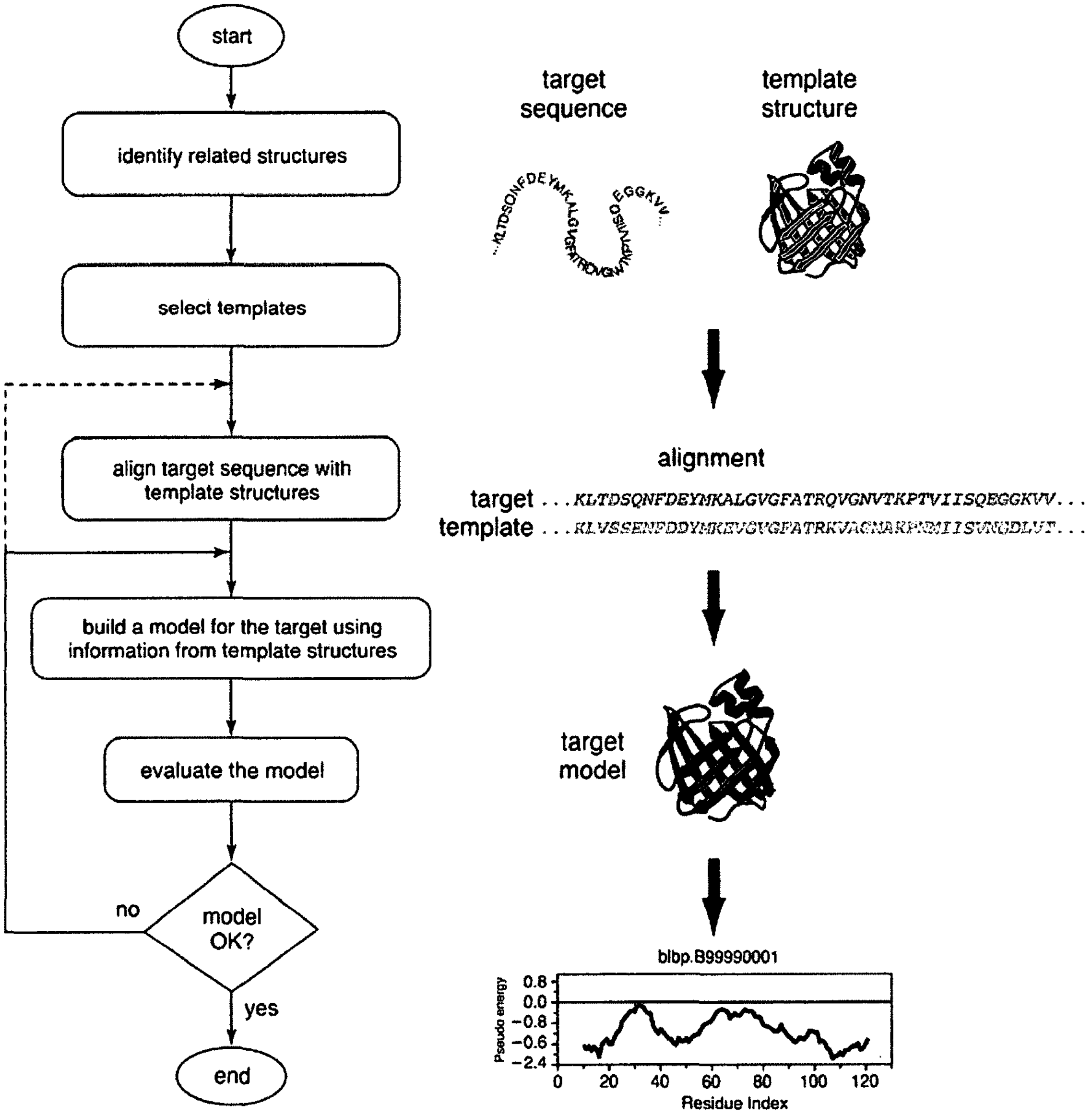

start
identify related structures
select templates
align target sequence with template structures
build a model for the target using information from template structures
evaluate the model
model OK?
no
yes
end
target sequence
template structure
alignment
target ...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
template ...KLVSSENFDDYMKEVGVGFATRKVAGMAKPNMIISVNGDLVT...
target model
blbp.B99990001
Pseudo energy
0.8
0.0
-0.8
-0.6
-2.4
0
20
40
60
80
100
120
Residue Index

Figure 24. Comparative Protein Structure Modeling Process

The illustration shows the steps that are taken to build a 3D structure based on homology. This schematic was reproduced from (Marti-Renom et al., 2000).

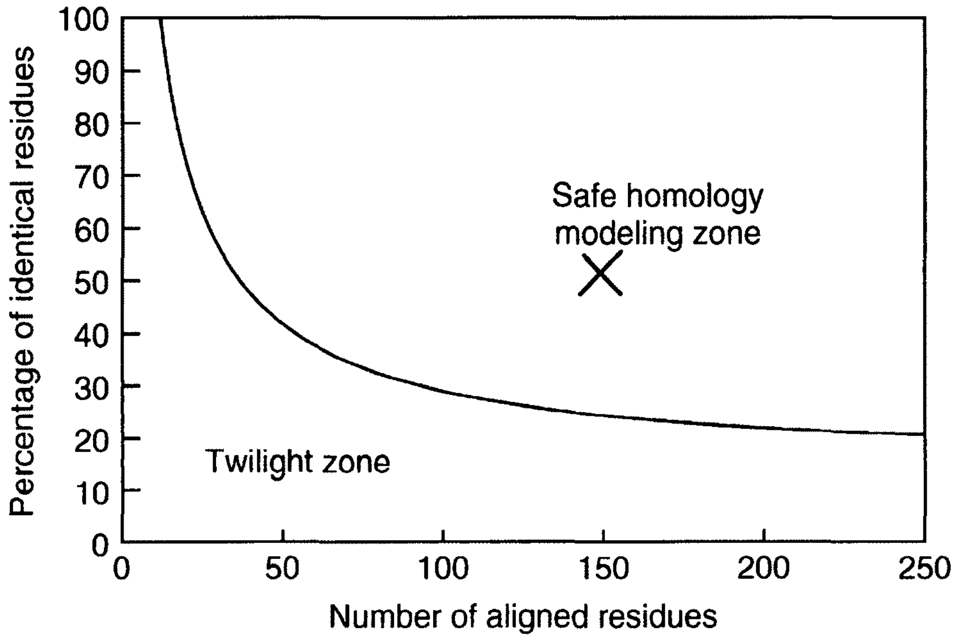


Figure 25. Safe Homology Modeling Zone
An illustration that shows twilight and safe homology modeling zones when two protein sequences are aligned. Figure is adapted from (Krieger et al., 2003).

obtained. Once the alignment is done, the backbone of the target protein is generated from that of the chosen protein. In the alignment, if the residues are exactly alike, the backbone with the side chain is copied from the template structure to the target structure. However, if the residues are different, only the backbone coordinates are used. If there are poorly aligned regions, another template with better alignment for that area can be used in what is called multiple template modeling.

When the backbone is generated, the loops are treated one of two ways. Either the residues found in the loop region are not included when forming the backbone or a continuous backbone is formed and the residues found in the loop region are removed. Thus, to model a loop, information from a previous 3D structure whose loops are similar to the query are copied or the loop can be generated using *ab initio* prediction programs, where the loop conformation is checked using energy (Fiser et al., 2000; Jamroz and Kolinski, 2010). For side-chain modeling, if the sequence identity of the query with the template is high, the necessary coordinates can be copied to the model. The modeling works because most structurally conserved proteins have similar $\chi1$- angles (torsion angle about the Cα-Cβ bond). For those residues with no sequence identity, rotamers can be predicted based on the backbone itself or they can be predicted based on the knowledge of other similar 3D structure side chains (Francis-Lyon and Koehl, 2014). The last step in homology modeling is model validation where the miscalculations in the structure are estimated (Schwede, 2013). The assessment can be done either by calculating the model energy based on a force field or by comparing the normality analysis (statistical test) of the model with a known structure. The model energy method tests whether the bond lengths and bond angles are within normal ranges. Out of the several homology modeling methods available, the ones used to model Tatil were: SWISS-MODEL, CPH model, Robetta server, MODELLER and the EsyPred program. These are described in the next sections.

3.2.1 SWISS- MODEL

SWISS-MODEL is a fully automated 3D modeling program that uses sequence homology to generate structures (Schwede et al., 2003; Arnold et al., 2006; Biasini et al., 2014). The program has a template collection named SWISS-MODEL template library (SMTL). The library collects information from the PDB about structures that have been solved experimentally. SMTL organizes the data according to the information provided by the PDB. A template is searched in the SMTL by using BLAST and HHblits, a statistical model based on hidden Markov model. The SWISS-MODEL algorithm performs several steps to choose a template for the 3D structure: sequence identity, sequence similarity, HHblits score, agreement between predicted secondary structure of target and template, agreement between predicted solvent accessibility among target and template. After alignment, ProMod-II, a protein modeling tool, is used to generate a 3D structure. If ProMod-II does not give a good result for the loop region, MODELLER is used in its place. The quality of the model generated is scored using QMEAN. It compares geometrical features of the model such as pairwise atomic distances and torsion angles to the statistical distribution found in the template.

3.2.2 CPHMODEL

In CPHmodel, a position-specific scoring matrix (PSSM) is generated for an amino acid sequence by performing up to five iterations using PSI-BLAST (Nielsen et al., 2010). PSI-BLAST stands for position specific iterative BLAST (Altschul et al., 1997). The search for homologous protein is called iteration which is based on the BLAST program. The result from the first iteration is used to build a multiple alignment for sequences with a certain e-value threshold. E-value depicts the probability that a similar sequence was found by chance. PSI-BLAST then calculates the PSSM conservation pattern from the alignment. The score is used to search for other sequences by using the pattern.

In the CPHmodel, after each iteration, the PSSM is used to search for a template in the PDB. If a template is found with an e-value of $<10^{-5}$, a PSSM is also generated for the possible template. The template is then aligned with the query, where the PSSM of every aligned position is the average of the template and the query. The alignment is accepted if the e-value is $<10^{-5}$ and percent identity of the alignment is greater than 30%. The backbone coordinates of the template are then used to construct the query 3D structure. The next step is to add missing atoms. A program, Segmod is used to build unknown segments of a structure by using known protein X-ray structures (Levitt, 1992). The energy calculation and dynamics (ENCAD) simulation program, is used to refine the 3D model (Levitt et al., 1995).

3.2.3 ROBETTA SERVER

The Robetta server is an automated program that predicts the 3D structure of a protein based on the information gathered from its amino acid sequences. The first part of the program is domain assignment (Kim et al., 2004). Robetta assigns domains for the input amino-acid sequence using Ginzu. To find homologous proteins Ginzu searches different databases such as BLAST and FFAS03 for each domain. BLAST is used first and then a list of other databases is followed. The programs are used step-by-step until all the sequence domains have been assigned.

Comparative modeling is used to build the 3D structure for the part of the sequence for which a homology protein has been found. If no domain is found for part of the input sequence, then a search is performed in Pfam-A using HMMER, which applies a hidden markov model. After the domains have been assigned for a homologous template, Robetta server uses either the *de novo* or template based modes to generate the model. In *de novo* modeling, a CASP-4 Rosetta protocol is used to generate various types of possible structures. Out of the generated structures, the ones that have non-protein-like conformations are removed. From the remaining structures, models with the lowest energy are chosen. For the template-based modeling, the input sequence is aligned with the homologous protein where the 3D structure is known. Robetta uses its own method for alignment called K*sync. It aligns two proteins by combining sequence and structural information by dynamic programming. Based on the alignment, a model is generated. The variable region model generated by the Rosetta *de novo* method are combined with the template by selecting a model with the lowest energy.

3.2.4 MODELLER

MODELLER searchs the PDB for proteins that have sequences similar to the input protein (Šali and Blundell, 1993; Fiser et al., 2002; Eswar et al., 2006). The next step is to align proteins with the target protein to find a template with a high sequence identity. The alignment is done by applying ALIGN2D which uses a global dynamic programming algorithm to calculate scores for the alignment. After the ideal template is found, spatial constraints and CHARMM energy terms for stereochemistry are used to form the backbone. The constraint information such as atom-atom distances and dihedral angles is gained from the template. Modeller then uses optimization based approaches to model the loops. The position of nonhydrogen atoms in the loop are optimized via molecular dynamics with restraints on the bond lengths and angles in molecular dynamics. The final step is to assess the model using various evaluation systems to check if the model stays within the constraints given.

3.2.5 ESYPRED PROGRAM

EsyPred is a fully automated program that involves four stages (Lambert et al., 2002). The first step is to identify structural templates for the desired protein. The templates are found by running the input sequence in the non-redundant databank of the National Center for Biotechnology Information. The program chooses the lowest e-value within four iterations with 0.0001 being the cutoff. Various programs such as ClustaW and Match-Box are used to align the template with the query. The programs generate several pairwise alignments between the template and the query, out of which the best ten alignments are selected. A database is then built which contains information about each position in the alignment. If the different programs align a particular amino acid similarly, that position receives a high score. The scores are then used to build a final alignment. In the final step a 3D model is made using the homology modeling routine of MODELLER.

3.3 CONSERVATION ANALYSIS

Lipocalins have low sequence identities; however, their 3D structure is highly conserved. The topology suggests that even though the residues are mutated throughout evolution there are certain amino acids within the sequence that are conserved, which are essential for determining the structure, kinetics and thermodynamic stability (Greene and Brew, 1995; Brew and Greene, 1997; Greene et al., 2003). Before calculating the conservation of each position in the sequence, a multiple sequence or sequence-based structure alignment must be performed. The alignment includes proteins from different organisms and with various functions. In the alignment, amino acids that are conserved across species and function indicate that they may be essential for structure, stability and folding, irrespective of the source of the protein. A conservation analysis was used to identify amino acids that are proposed to be key determinants for the lipocalin superfamily.

3.4 MATERIALS AND METHODS

3.4.1 CONSTRUCTION OF A SUPERFAMILY ALIGNMENT

The Dali program was used for the structural alignment of the lipocalin superfamily. The protein with the highest sequence identity with Tatil was used as the basis for a search of similar structures in Dali. From this generated list, seventeen lipocalin structures from different organisms and with various functions were selected and aligned (Table 1). The alignment was also analyzed by eye to verify the direction of the side chains.

3.4.2 MODELING

Five different programs were used to generate the 3D structure of Tatil. The programs were SWISS-MODEL, CPH model, EsyPred, Robetta and MODELLER. In SWISS-MODEL, the sequence of Tatil was supplied and a list of template sequences was generated. The template with the highest sequence identity was chosen to generate a model. The chosen template was 2ACO (Figure 26), a bacterial lipocalin with a 40% sequence identity with Tatil (Figure 28). In the CPH model, the sequence was entered and the program chose the template and built the model without further input from the user. For EsyPred, the lipocalin sequence 2ACO was provided to the server, where the option of inserting the template information is given. MODELLER is a command driven program where the Tatil alignment was used to generate the 3D model. In the last program Robetta, the sequence of the lipocalin was entered. The server searched for a template for alignment. Once the server selected a template, it proceeded to construct the 3D model of Tatil.

3.4.3 CONSERVATION ANALYSIS

Conservation analysis of the lipocalin superfamily was calculated using modified Shannon's entropy,

S(i) =$\sum_{j=1}^{m} -Pj(i)ln[Pj(i)](i = 1, 2, ..., 20)$

where Pj(i) is the fractional occurrence of an amino acid type in location i and m is the number of amino acid types. The conservation was the calculated using C(i)=1-(S(i))/(ln(m)). The program used to count the amino acids at each position of the alignment for input into our equations was coded by the He lab (Old Dominion University, Norfolk, VA).

3.4.4 HYDROPATHY ANALYSIS

Hydropathy analysis of the lipocalin superfamily was based on the Kyte and Doolittle hydrophobicity plot and used the following formula:

Hydrophobicity = sum of the number of each amino acid at a given position * hydrophobicity of that amino acid.

3.4.5 LONG-RANGE INTERACTION NETWORK

A list of all the atom contacts within the Tatil model was generated with the Contact Program in the CCP4 software (Potterton et al., 2003). The contact file was used as input to calculate the long-range interactions using a script that was coded by Dr. L. Greene and J. Pothen called degLR (Greene and Higman, 2003). A long-range interaction is considered a contact between amino acids that are 7 residues apart in the primary structure but are within 5 angstroms in the 3D structure. The program Pajek, was used to generate a 2D schematic of the long-range interactions (Batagelj and Mrvar, 1998).

3.5 RESULTS AND DISCUSSION

A set of 3D structures of Tatil (Figure 29) was generated using homology modeling. Different programs were used to corroborate the proposed structure of Tatil. When possible, 2ACO was requested for the template unless the program chose its own. The Robetta server chose 1QWD which is also a bacterial lipocalin. All five gave the same general structure of a lipocalin with eight anti-parallel β-strands forming a β-barrel. The difference was in the C-terminal α-helix and the number of amino acids used to construct the 3D structure. Even though Tatil has 190 amino acids, some programs did not use use all the amino acids to construct the models. In the models generated, the SWISS-MODEL structure contained only 166 amino acids, the CPH model 158 amino acids, EsyPred 167 amino acids, Robetta and MODELLER 190 amino acids. Robetta and MODELLER generated five various models each with varying C-terminal structure. Although all the models have one C-terminal α-helix, Modeller and Robetta models added another loop and α-helix after that. A possible explanation for the strand after the C-terminal α-helix could be that Tatil is a dimer like the bacterial lipocalin 2ACO and the strand is used to connect to another monomer. This can be confirmed when a crystal structure of Tatil is obtained. We gave the model a temporary protein code 9TAT and will deposit the structure in the PDB, although the code may be changed by the database curators.

Table 3. The Function of Select Lipocalin Superfamily Members

PDB Code	Name	Function	Source	State
9TAT	Temperature-induced lipocalin	Temperature-induced lipocalins are concentrated in the plasma membrane (Charron et al., 2005). Increase in the expression of the lipocalin is seen at high and low temperatures. In a knock-out line of the lipocalin, accumulation of hydrogen peroxide and sensitivity to light and freezing was increased (Charron et al., 2008).	Plant	To be determined
2ACO	Bacterial lipocalin	A membrane protein that faces the periplasmic space in a gram-negative bacteria. It is expressed when bacteria is under stress during high osmolarity (Campanacci et al., 2006).	Bacteria	Dimer
1BBP	Bilin-binding protein	BBP is a blue pigment protein that is a result of complexation with biliverdin, which is a product of the porphyrin biosynthetic pathway. The lipocalin is responsible for the color of the insect and might contribute in photoprotection and photoreception (Huber et al., 1987).	Butterfly	Tetramer

Table 3. Continued

PDB Code	Name	Function	Source	State
1GKA	Beta-crustacyanin	Binds carotenoid astaxanthin (AXT). The binding and unbinding of AXT gives the lobster different colors, which helps with camouflage in various environments (Cianci et al., 2002).	Lobster	Dimer
1AQB	Retinol-binding protein	It carries retinol in the plasma. It is attached to thyroxine-binding transthyretin when in the plasma (Zanotti et al., 1998).	Pig	Monomer
3QKG	α1-macroglobulin	A protein found in most tissues and is associated with immunosuppression, neutrophil chemotaxis and inhibition of antigen-stimulated interleukin, heme and tryptophan metabolism (Meining and Skerra, 2012).	Human	Monomer
4GE1	Amine binding protein	It binds biogenic amines such as serotonin, norepinephrine and epinephrine. It is an inhibitor of biogenic amine-mediated platlet activation and smooth-muscle contraction (Xu et al., 2013).	Rhodnius prolixus (bugs)	Monomer

Table 3. Continued

PDB Code	Name	Function	Source	State
1DFV	Neutrophil gelatinase associated lipocalin (NGAL)	Expressed in epithelial cells during inflammation and neoplastic transformation (Goetz et al., 2000). NGAL is believed to be involved in transporting iron, induction of apoptosis in cytokine-dependent neutrophils and in suppression of bacterial growth.	Human	Monomer, dimer
1EPA	Epididymal retinoic acid binding protein	Found in the lumen of the epididymis, which is a site of sperm maturation. It binds both all trans and 9-cis retinoic acid (Newcomer, 1993).	Rat	Dimer
1I04	Urinary protein	Binds different types of pheromones such as 2-sec-butyl-4,5-dihydrothiazole and dehydro-exo-brevicomin that affect estrus, puberty and inter-male aggression (Cavaggioni and Mucignat-Caretta, 2000).	Mouse	Monomer
2WEX	Apolipoprotein M	Found in high density lipoprotein (HDL). ApoM is believed to have the ability to prevent the initiation or acceleration of the deposition of lipids in the arterial lumen (Sevvana et al., 2009).	Human	Monomer

Table 3. Continued

PDB Code	Name	Function	Source	State
2OFM	Apo nitrophorin 4	Nitrophorins are lipocalins that transport nitric oxide(NO). Nitrophorins take NO and deliver it to a victim's tissue (Amoia and Montfort, 2007).	Rhodnius rolixus	Monomer
1BJ7	Allergen bos d 2	Bos d 2 is made by the apocrine sweat glands (Mantyjarvi et al., 2000). The lipocalin is an allergen that induces an immune response leading to the production of IgE antibodies.	Bovine	Monomer
1BEB	Beta-lactoglobulin	A major whey protein found in the milk of many mammals (Brownlow et al., 1997). Even though the exact function is not known, it binds to fatty acids as well as retinol.	Bovine	Dimer
1E5P	Aphrodisin	It stimulates male copulatory behavior. The vomeronasal organ is the receptor of the aphrodisin lipocalin (Vincent et al., 2001).	Hamster	Monomer
1XKI	Tear lipocalin	It has broad ligand specificity. It binds fatty acids, fatty alcohols, phospholipids, glycolipids, cholesterol, retinol, arachidonic acid, isoprostanes (Breustedt et al., 2005).	Human	Monomer

Table 3. Continued

PDB Code	Name	Function	Source	State
2R74	Trichosurin	A protein from the milk whey of the possum. The lipocalin is hypothesized to bind phenolic compounds in a detoxification pathway (Watson et al., 2007).	Possum	Dimer

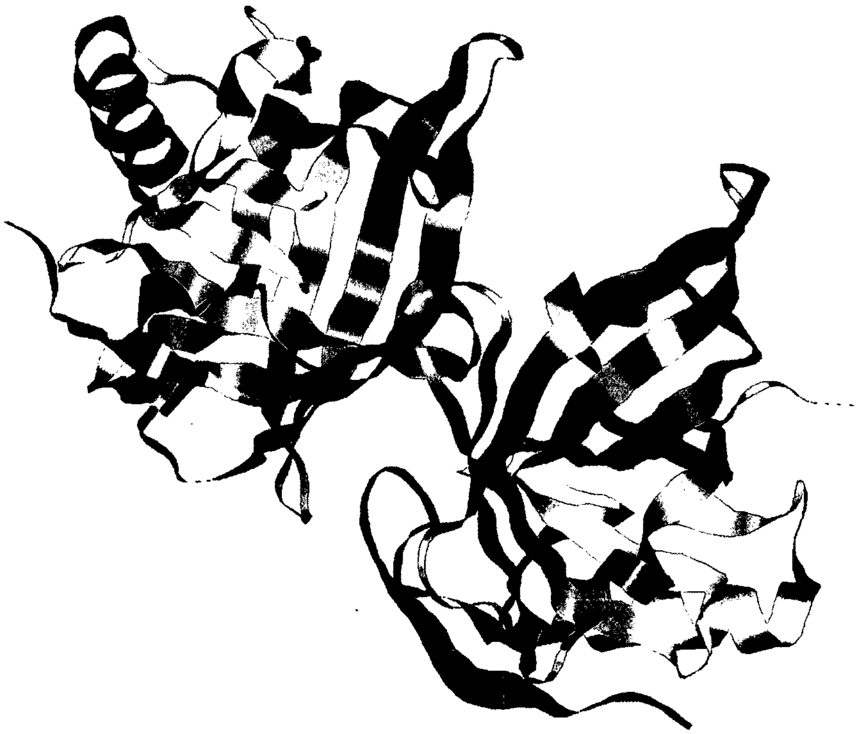


Figure 26. 3D Structure of a Bacterial Lipocalin
The structure of a bacterial lipocalin dimer from *E. coli* (PDB code:2ACO). This structure is visualized on Rasmol (version 2.7.5.2).

Conservation analysis of a protein superfamily enables us to find the amino acids that may be responsible for the stability and folding of a protein. We were able to construct a structural-based sequence alignment of the lipocalin superfamily (Figure 27) using different organisms with various functions (Table 3). In the first step, related proteins to Tatil with known 3D structures were identified using BLAST. The bacterial lipocalin (2ACO), which has a high sequence identity of 40% with Tatil was used to search Dali to find proteins with similar structures. Even though the search yielded a large number of lipocalins, we were only interested in the ones where the sequence alignment percentage is very low (Table 3). In addition, lipocalins with a similar function but in different organisms were not considered. Furthermore, we tried to make sure that the sequence similarity among the lipocalins stayed below 35% (Table 4). Using these qualifications we identified sixteen lipocalins including the bacterial lipocalin. After we constructed the 3D structure of Tatil, we also added it to the alignment which made the total number of aligned proteins seventeen (Figure 27).

The structural alignment of the lipocalin superfamily was used to calculate conserved positions by applying a modified Shannon's entropy equation. We were able to identify seventeen highly conserved and sixteen moderately conserved residues in Tatil. We propose that these amino acids are key to determining the structure, folding kinetics and thermodynamic stability (Figure 30). Positions that has a conservation (C(i)) between 0.40 and 0.49 were considered to be moderately conserved while positions higher than 0.50 were taken to be highly conserved.

```
                                                                    * * * * *SCR1* * * * *
9TAT (Wheat lipocalin)           2 ----------AAKKSGSE--MGV---VLG-L--DVARY-M--GRWYEIASF-----P-- 32
2ACO (Blc lipocalin)            10 HLESTSLYKKAGSTPPRG--VTV---VNN-F--DAKRY-L--GTWYEIARF-----D-- 50
1BBP (Bilin binding protein)     2 --------NVYHDGACPE--VKP---VDN-F--DWSNY-H--GKWWEVAKY-----P-- 34
1GKA (β-crustacyanin)            2 -----KIPNFVVPGKCAS--VDRNKLWAE-QTPNRNSY-A--GVWYQFALT-----N-- 42
1AQB (Retinol binding protein)   1 ----------ERDCRVSS--FRV---KEN-F--DKARF-S--GTWYAMAKK-----D-- 31
3QKG (α-1 microglobulin)         8 ----------------DN--IQV---QEN-F--NISRI-Y--GKWYNLAIGSTSPWL-- 37
4GE1 (Amine binding protein)     2 ----------SGCST-----VDT---VKD-F--NKDNFFT--GSWYITHYK-----L-- 30
1DFV (Neutrophil gelatinase)     5 ---------SDLIPAPPLSKVPL---QQN-F--QDNQF-Q--GKWYVVGLA-------- 37
1EPA (Retinoic acid binding)     3 --------------------------VKD-F--DISKF-L--GFWYEIAFA-----S-- 22
1I04 (Urinary protein)          19 EEAS------------------S---TGRNF--NVEKI-N--GEWHTIILA-----S-D 45
2WEX (Major Apolipoprotein M)   18 ------------SHMNQCPEHSQ---LTT-GKEFPEVH-L--GQWYFIAGAAPTKEE-- 59
2OFM (Apo nitrophorin 4)         1 -------------ACTKN--AIA---QTG-F--NKDKY-FNGDVWYVTDYL-----DLE 32
1BJ7 (Allergen lipocalin)        7 ------------------------------I--DPSKI-P--GEWRIIYAA------AD 24
1BEB (β-lactoglobulin)           5 ---------------------QT---MKG-L--DIQKV-A--GTWYSLAMA-----ASD 28
1E5P (Aphrodisin)                3 ----------------------------------FAEL-Q--GKWYTIVIA-----A-D 18
1XKI (Tear lipocalin)           12 ------------------------------------DV-S--GTWYLKAMT-------- 23
2R74 (Trichosurin)              24 ---------------------------------------L--RHWHTVVLA-----S-S 36
                                                                            β1
```

Figure 27. Lipocalin Superfamily Alignment

An alignment of seventeen lipocalins from different organisms with various functions. Amino acid sequences highlighted in red indicate the highly conserved residues while the ones highlighted in blue show the moderately conserved residues as determined by the conservation calculation. The PDB code and name of the lipocalin are listed. The residue numbers are shown on the left- and right-side of the alignment. The secondary elements in Tatil are shown with arrows and lines. The location of the SCRs in Tatil are also shown at the top of the alignment.

```
9TAT (Wheat lipocalin)        33 N-------FFQPRDGRDTRATYELM-ED-GATVHVLNETWS-----KGKRDFIEGTAY--  76
2ACO (Blc lipocalin)          51 H-------RFERG-LEKVTATYSLR-DD-G-GLNVINKGYNPD---RGMWQQSEGKAY--  94
1BBP (Bilin binding protein)  35 N-------SVEKY-GKCGWAEYTPE--G-K-SVKVSNYHVIH-----GKEYFIEGTAY--  75
1GKA (β-crustacyanin)         43 N-------PYQLI-EKCVRNEYSFD-G--K-QFVIKSTGIAY----DGNLLKRNGKLY--  84
1AQB (Retinol binding protein)32 P------EGLFL--QDNIVAEFSVD-EN-G-HMSATAKGRVRLLNNWDVCADMVGTFT--  78
3QKG (α-1 microglobulin)      38 K-------KIMDR-MTVSTLVLGEG-AT-EAEISMTSTRWRK-----GVCEETSGAYE--  80
4GE1 (Amine binding protein)   31 G----DSTLEVGD-KNCTKFLHQKT-AD-G-KIKEVFSNYNPN---AKTYSYDISFAKVS  79
1DFV (Neutrophil gelatinase)  38 --GNAILREDKDP-QKMYATIYELK-ED-K-SYNVTSVLFR-----KKKCDYWIRTFV--  84
1EPA (Retinoic acid binding)  23 -----------KE-EKMGAMVVELK-E--N-LLALTTTYYSE-----DHCVLEKVTAT--  68
1I04 (Urinary protein)        46 KREKIEDNGNFR----LFLEQIHVL--E-N-SLVLKFHTVR-----DEECSELSMVAD--  90
2WEX (Major Apolipoprotein M) 60 L-------ATFDP-VDNIVFNMAAG-SAPM-QLHLRATIRMKD----GLCVPRKWIYH-- 103
2OFM (Apo nitrophorin 4)      33 P-------DDVPK-RYCAALAAGTA-SG-K--LKEALYHYDPK---TQDTFYDVSELQ--  75
1BJ7 (Allergen lipocalin)     25 NKDKIVEGGPLR----NYYRRIECI-NDCE-SLSITFYLKD-----QGTCLLLTEVAK--  71
1BEB (β-lactoglobulin)        29 ISLLDAQSAPLR----VYVEELKPT-PE-G-DLEILLQKWENG---E--CAQKKIIAE--  74
1E5P (Aphrodisin)             19 NLEKIEEGGPLR----FYFRHIDCY-KNCS-EXEITFYVIT-----NNQCSKTTVIGY--  65
1XKI (Tear lipocalin)         24 V-------NLE----SVTPMTLTTL-EG-G-NLEAKVTM-------SGRCQEVKAVLE--  69
2R74 (Trichosurin)            37 DRSLIEEEGPFR----NFIQNITVE--S-G-NLNGFFLTRK-----NGQCIPLYLTAF--  81
                                                ━━━━━━━━►      ━━━━━━━━━━━►   ━━━━━━━━━━━►
                                                    β2              β3              β4
```

Figure 27. Continued

```
                                                                              * * * *SCR 2* * * *
9TAT (Wheat lipocalin)          77 KADPASEEAKLKVKFY------V------PPFLPIIPVV----GDYWVLYVDDDYQYAL 119
2ACO (Blc lipocalin)            95 FTG-APTRAALKVSFF------G-------------PFY----GGYNVIALDREYRHAL 129
1BBP (Bilin binding protein)    76 PVG-DSKIGKIYHKLT------Y----------GGVTKE----NVFNVLSTDN-KNYII 115
1GKA (β-crustacyanin)           85 PNP-F-GEPHLSIDYE------N-------------SFA----APLVILETD-YSNYAC 117
1AQB (Retinol binding protein)  79 DTE-D--PAKFKMKYWGVASFLQ-------------KGN----DDHWIIDTDY-DTYAV 116
3QKG (α-1 microglobulin)        81 KTD-T--DGKFLY------------------HKSKWNIT----MESYVVHTN-YDEYAI 113
4GE1 (Amine binding protein)    80 DFD-G-NNGKYTAKNV------I----VEKDGRKIDERT----LQVSYIDTD-YSKYSV 121
1DFV (Neutrophil gelatinase)    85 PGC-Q--PGEFTLGNI------K-------------SYPGLTSYLVRVVSTN-YNQHAM 120
1EPA (Retinoic acid binding)    69 EGD-G--PAKFQVT-R------L-------------SGK----KEVVVEATDYL-TYAI  99
1I04 (Urinary protein)          91 KTE-K--AGEYSVTYD---------------------GF----NTFTIPKTD-YDNFLM 120
2WEX (Major Apolipoprotein M)  104 LTE-G--STDLRT------------------------------MKTELFSSS-CPGGIM 133
2OFM (Apo nitrophorin 4)        76 VES----LGKYTANFK------KVDKNGNVKVAVTAGNY----YTFTVMYAD--DSSAL 118
1BJ7 (Allergen lipocalin)       72 RQE----GYVYVLEFY---------------------GT----NTLEVIHVS--ENMLV  99
1BEB (β-lactoglobulin)          75 KTK-I--PAVFKIDAL---------------------NE----NKVLVLDTD-YKKYLL 104
1E5P (Aphrodisin)               66 LKG-NG---TYETQFE---------------------GN----NIFQPLYIT--SDKIF  93
1XKI (Tear lipocalin)           70 KTD-E--PGKYTAD----------------------GGK----HVAYIIRSHVK-DHYI  98
2R74 (Trichosurin)              82 KTE-E--ARQFKLNYY---------------------GT----NDVYYE-SSKPNEYAK 111
                                              β5                              β6          β7
```

Figure 27. Continued

```
                                                               * * * * *SCR3* * * * *
9TAT (Wheat lipcalin)           120 VGEP-------R---------RKSLWILCRKT-HIEEEVYNQLLEKAKEEG- 153
2ACO (Blc lipocalin)            130 VCGP-------D---------RDYLWILSRTP-TISDEVKQEMLAVATREG- 163
1BBP (Bilin binding protein)    116 GYYC-------KYDEDKKG-HQDFVWVLSRSK-VLTGEAKTAVENYLIGSPV 158
1GKA (β-crustacyanin)           118 LYSC-------I-DYNFGY-HSDFSFIFSRSA-NLADQYVKKCEAAFKNIN- 158
1AQB (Retinol binding protein)  117 QYSC-------RLQNLDGTCADSYSFVFARDPHGFSPEVQKIVRQRQEELC- 160
3QKG (α-1 microglobulin)        114 FLTKKFSRHHGP---------TITAKLYGRAP-QLRETLLQDFRVVAQGVG- 154
4GE1 (Amine binding protein)    122 VHVC-----------DPAAPDYYLYAVQSRTE-NVKEDVKSKVEAALGKVG- 160
1DFV (Neutrophil gelatinase)    121 VFFK---------KVSQNR-EYFKITLYGRTK-ELTSELKENFIRFSKSLG- 160
1EPA (Retinoic acid binding)    100 IDIT----------SLVAGAVHRTMKLYSRSL-DDNGEALYNFRKITSDHG- 139
1I04 (Urinary protein)          121 AHLI---------NEKDGE-TFQLMGLYGREP-DLSSDIKERFAQLCEEHG- 160
2WEX (Major Apolipoprotein M)   134 LNET-------G------Q-GYQRFLLYNRSP-HPPEKCVEEFKSLTSCLD- 169
2OFM (Apo nitrophorin 4)        119 IHTC---------LHKGNKDLGDLYAVLNRNK-DAA--AGDKVKSAVSAAT- 157
1BJ7 (Allergen lipocalin)       100 TYVE-NYDGERI---------TKMTEGLAKGT-SFTPEELEKYQQLNSERG- 139
1BEB (β-lactoglobulin)          105 FCME----------NSAEPEQSLVCQCLVRTP-EVDDEALEKFDKALK--A- 142
1E5P (Aphrodisin)                94 FTNK--------NXDRAGQ-ETNXIVVAGKGN-ALTPEENEILVQFAHEKK- 134
1XKI (Tear lipocalin)            99 FYSE-----------GEGK-PVRGVKLVGRDPKNNLE-ALEDFEKAAGARG- 138
2R74 (Trichosurin)              112 FIFY---------NYHDGK-VNVVANLFGRTP-NLSNEIKKRFEEDFMNRG- 151
                                    β7                         β8              α
```

Figure 27. Continued

```
9TAT (Wheat lipocalin)          154 YDVAKLHKT---PQS-----------D----------------------- 166
2ACO (Blc lipocalin)            164 FDVSKFIWV---QQP-----------G----------------------S 177
1BBP (Bilin binding protein)    159 VDSQKLVYS---DFS-----------E----------------AACKVN- 178
1GKA (β-crustacyanin)           159 VDTTRFVKT---VQG-----------S------------SCPYDTQKTV- 181
1AQB (Retinol binding protein)161 LA-RQYRII---THN-------GYCD------------------------ 175
3QKG (α-1 microglobulin)        155 IPEDSIFTM---ADR-----------G------------------ECVP- 171
4GE1 (Amine binding protein)    161 LKLSGLFDA--TTLG-----------NKCQYDDETLQKLLKQSFPNYEK- 196
1DFV (Neutrophil gelatinase)    160 LPENHIVFP---VPI-----------D------------------QCID- 177
1EPA (Retinoic acid binding)    140 FSETDLYIL---KHDLTCVKVLQSAA------------------------ 162
1I04 (Urinary protein)          161 ILRENIIDL---SNA-----------N------------------RCLQ- 177
2WEX (Major Apolipoprotein M)   170 SK--AFLLT---PRN-----QEACEL------------------------ 185
2OFM (Apo nitrophorin 4)        158 LEFSKFIST---KEN-----------N--------CAYDNDSLKSLLTK- 184
1BJ7 (Allergen lipocalin)       140 VPNENIENL---IKT------DNCPP------------------------ 156
1BEB (β-lactoglobulin)          143 LPMHIRLSFNPTQLE-----------E--------------------QC- 160
1E5P (Aphrodisin)               135 IPVENILNI---LAT------DTCPE------------------------ 151
1XKI (Tear lipocalin)           139 LSTESILIP---RQS----------------------------------- 150
2R74 (Trichosurin)              152 FRRENILDI---SEV-----------D--------------------HC- 166
```

Figure 27. Continued

```
2ACO    HHHHHHLESTSLYKKAGSTPPRGVTVVNNFDAKRYLGTWYEIARFDHRFE-RGLEKVTATYSLRDDGG-LNVINKGYNPD
9TAT              MAAKKSGSE----MGVVLGLDVARYMGRWYEIASFPNFFQPRDGRDTRATYELMEDGATVHVLNETWS--

2ACO    RGMWQQSEGKAYFTG-APTRAALKVSFFGPFY-------GGYNVIALDREYRHALVCGPDRDYLWILSRTPTISDEVKQE
9TAT    KGKRDFIEGTAYKADPASEEAKLKVKFYVPPFLPIIPVVGDYWVLYVDDDYQYALVGEPRRKSLWILCRKTHIEEEVYNQ

2ACO    MLAVATREGFDVSKFIWVQQPGS
9TAT    LLEKAKEEGYDVAKLHKTPQSDPPPESDAAPTDSKGTWWFKSLFGK
```

Figure 28. Sequence alignment of 2ACO and 9TAT Proteins

A sequence alignment of the wheat (9TAT) with bacterial (2ACO) lipocalin is shown. The highlighted residues show the amino acids that are identical in both proteins.

In SCR1, there are four highly conserved positions and three moderately conserved positions (Figure 27) (Table 6). Three of the highly conserved positions, Gly at position 23, Trp at position 25 and Tyr at position 25 are used to qualify whether or not a protein is a lipocalin (Figures 6 and 27). We located a fourth conserved position Ala at position 29 in the same region. In SCR2, four highly conserved positions and one moderately conserved position were identified. Generally, in SCR2 there are four residues that are used to classify a protein as a lipocalin. Surprisingly, in the structural alignment, only two of the positions were identified as highly conserved. In SCR3, two highly conserved positions and six moderately conserved positions were identified. The SCR3 region of a lipocalin is identified partially by the presence of an Arg at position 133. This position was highly conserved as was an Ile at position 130. Other highly and moderately conserved positions were found in the protein. The conserved residues can be seen on the secondary structure of the proposed model of Tatil (Figure 32).

The number of conserved positions that were found are high even though the sequence similarity among lipocalins is low. The clustered location of the conserved residues may indicate their importance for the stabilization of the protein (Greene et al., 2001; Greene and Higman, 2003; Schueler-Furman and Baker, 2003). When we look at the conserved positions in the 3D structure of the Tatil model, it can be seen that the residues have clustered in certain areas particularly at the bottom of the β-barrel (Figure 32). The results imply that the conserved positions maybe essential for the folding and stability of lipocalins. We also looked at salt bridges within the Tatil model structure to observe if the conserved positions that were identified relate to the amino acids that form salt bridges. Salt bridges are established between amino acids with opposite charges within close proximity for electrostatic attraction. We ran a program called ESBRI that searched for salt bridges within a protein (Costantini et al., 2008). ESBRI looks for carboxyl oxygen atoms in Asp or Glu and nitrogen atoms in Arg, Lys or His which are within 4 Angstroms apart. The program found four salt bridges (Table 5). The amino acids involved in the formation of the salt bridges are not the same as the ones that were identified in the conservation analysis. The association between Arg41 and Asp42 is unlikely as the amino acids are found next to each other in the same strand (β-strand B). The other salt bridges that

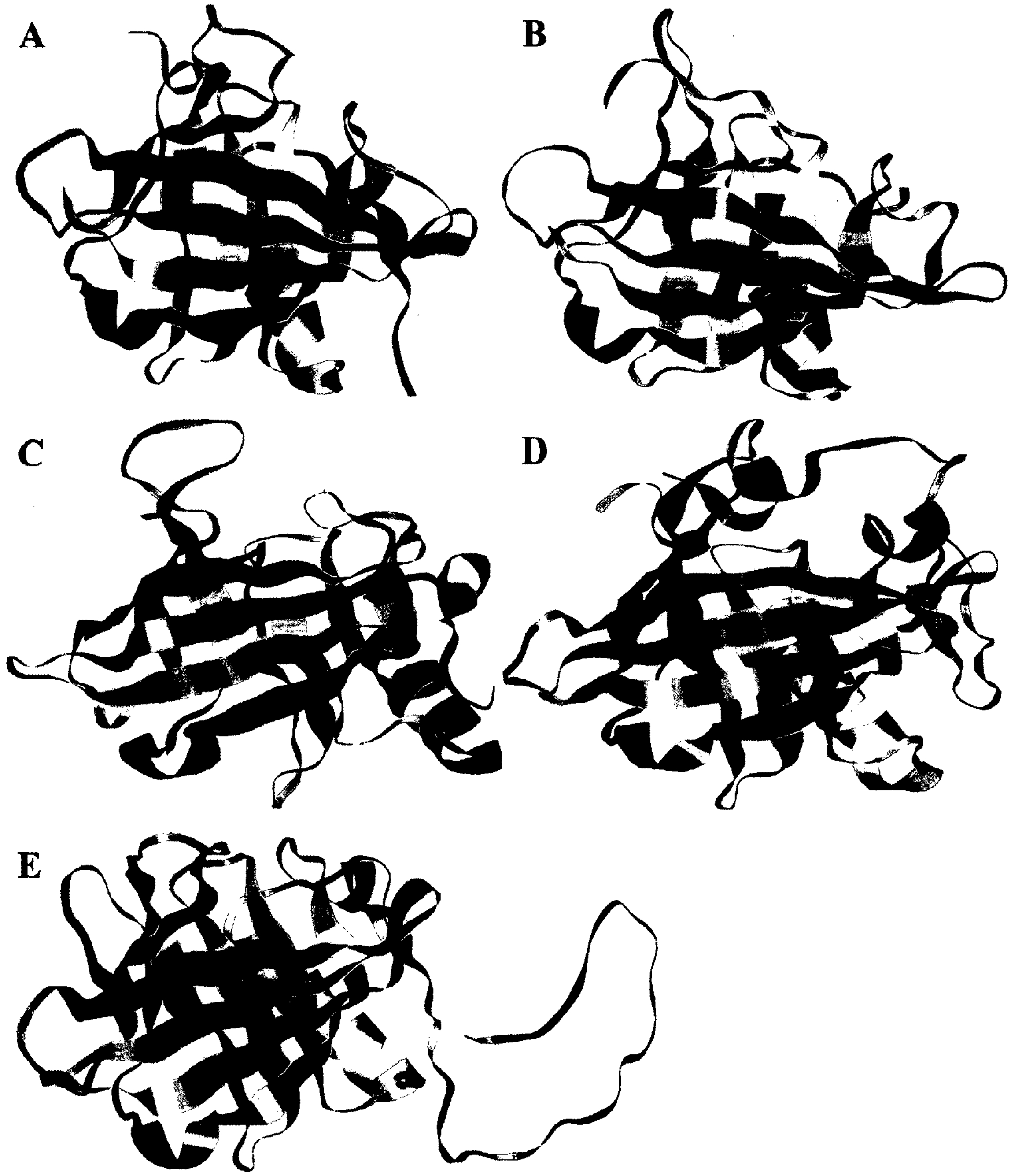


Figure 29. Proposed 3D Structures of Tatil

The 3D structures of Tatil using five different homology modeling programs are shown. A. SWISS-MODEL. B. CPH program. C. EsyPred program. D. Robetta program. E. MODELLER program.

were identified included Arg44 (β-strand B) with Glu61(β-strand C), His57 (β-strand C) with Glu72 (β-strand D) and His136 (loop region between strand H and the C-terminal helix) with Asp166 (C-terminal helix). These salt bridges might give additional support to the stabilization of the Tatil protein.

Table 4. Sequence Identity of the Lipocalin Superfamily

2ACO	100.00																
9TAT	40.13	100.00															
1AQB	21.52	21.35	100.00														
1BBP	25.48	22.35	15.79	100.00													
1BEB	16.56	17.09	21.25	17.50	100.00												
1BJ7	15.23	20.00	10.32	16.77	18.71	100.00											
1DFV	20.37	15.38	12.94	18.56	16.77	18.83	100.00										
1E5P	12.84	14.00	17.88	14.09	16.11	35.33	19.33	100.00									
1EPA	22.00	19.51	20.00	15.00	24.52	16.34	21.29	11.92	100.00								
1GKA	20.00	17.06	16.47	22.62	16.25	11.54	14.20	10.00	13.84	100.00							
1I04	13.25	11.36	17.37	15.50	15.62	25.32	18.82	22.82	18.99	11.43	100.00						
1XKI	15.00	22.22	17.50	11.39	25.17	15.33	15.23	12.24	22.01	12.42	19.48	100.00					
2OFM	13.17	16.02	16.95	20.96	14.91	16.77	15.88	14.77	20.12	12.50	13.71	16.05	100.00				
2R74	12.96	11.18	15.43	14.91	18.47	22.37	20.86	21.23	16.56	20.99	33.13	24.83	9.32	100.00			
2WEX	16.15	12.20	18.35	18.12	16.67	16.99	13.69	13.91	18.30	11.11	13.61	19.33	12.43	18.12	100.00		
3QKG	18.87	17.65	17.34	12.21	18.52	14.10	26.19	18.12	24.69	11.11	20.96	32.92	15.34	14.72	14.55	100.00	
4GE1	16.17	16.04	19.77	17.96	19.25	16.23	17.86	14.57	21.95	22.62	13.64	18.52	25.97	14.81	14.04	15.22	100.00
	2ACO	**9TAT**	**1AQB**	**1BBP**	**1BEB**	**1BJ7**	**1DFV**	**1E5P**	**1EPA**	**1GKA**	**1I04**	**1XKI**	**2OFM**	**2R74**	**2WEX**	**3QKG**	**4GE1**

Table 5. Salt Bridge

Residue 1	Residue 2	Distance
ARG 41	ASP 42	3.94
ARG 44	GLU 61	3.85
HIS 57	GLU 72	3.49
HIS 136	ASP 166	3.79

The average hydropathy of the lipocalin superfamily was also calculated using the Kyte and Dolittle hydropathy scale (Figure 31). The highly and moderately conserved positions were found in both hydrophobic as well as hydrophilic regions with seventeen residues in the hydrophobic region and sixteen residues in the hydrophilic area. The average hydropathy profile of the lipocalin superfamily indicates that not all conserved residues are hydrophobic and not all hydrophobic residues are conserved (Figure 31).

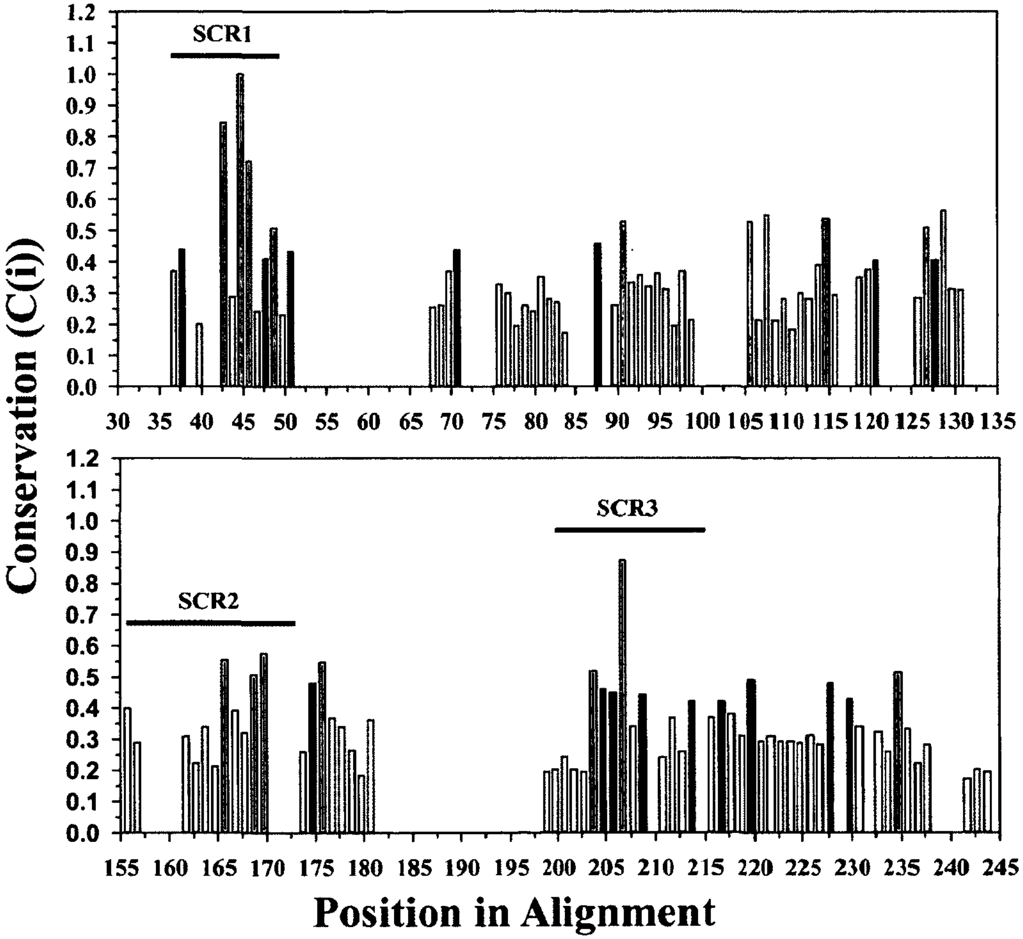


Figure 30. Conservation Analysis of the Lipocalin Superfamily

Conserved positions of the amino acids in the lipocalin superfamily. Red shows the highly conserved positions while blue shows the moderately conserved postions. The location of the SCRs are indicated by a black line.

Table 6. Summary of all conserved position conservation analysis

Amino Acid	Secondary Structure	Amino Acid Conservation	Hydropathy
21 Y	β-Strand 1	Moderately	Hydrophobic
23 G	β-Strand 1	Highly	Hydrophilic
25 W	β-Strand 1	Highly	Hydrophilic
26 Y	β-Strand 1	Highly	Hydrophilic
28 I	β-Strand 1	Moderately	Hydrophobic
29 A	β-Strand 1	Highly	Hydrophobic
31 F	Loop 1	Moderately	Hydrophobic
37 P	Loop 1	Moderately	Hydrophilic
53 G	Loop 2	Moderately	Hydrophilic
56 V	β-Strand 3	Highly	Hydrophobic
66 G	Loop 3	Highly	Hydrophilic
68 R	β-Strand 4	Highly	Hydrophobic
75 A	β-Strand 4	Highly	Hydrophobic
79 D	Loop 4	Moderately	Hydrophilic
85 A	β-Strand 5	Highly	Hydrophilic
86 K	β-Strand 5	Moderately	Hydrophilic
87 L	β-Strand 5	Highly	Hydrophobic
108 V	β-Strand 6	Highly	Hydrophobic
111 V	β-Strand 6	Highly	Hydrophobic
112 D	β-Strand 6	Highly	Hydrophilic
117 Y	β-Strand 7	Moderately	Hydrophilic
118 A	β-Strand 7	Highly	Hydrophilic
130 I	β-Strand 8	Highly	Hydrophobic
131 L	β-Strand 8	Moderately	Hydrophobic
132 C	β-Strand 8	Moderately	Hydrophilic

Table 6. Continued

Amino Acid	Secondary Structure	Amino Acid Conservation	Hydropathy
133 R	Loop connecting β-Strand 8 and C-terminal helix	Highly	Hydrophilic
135 T	Loop connecting β-Strand 8 and C-terminal helix	Moderately	Hydrophobic
139 E	C-terminal helix	Moderately	Hydrophilic
142 R	C-terminal helix	Moderately	Hydrophilic
145 L	C-terminal helix	Moderately	Hydrophobic
153 G	C-terminal helix	Moderately	Hydrophilic
154 G	C-terminal helix	Moderately	Hydrophilic
159 L	C-terminal helix	Highly	Hydrophobic

The folding of a protein from a primary structure to its 3D structure requires the association of residues. The interactions among the amino acids are important for the stability and folding of a protein. The association of amino acids in a protein can be categorized into long-range and short-range interactions (Greene and Higman, 2003). Long-range interactions describe the contact between residues that are close in tertiary structure but are far apart in the primary structure. On the other hand, short-range interactions are amino acids that are close to each other in the tertiary as well as in the primary structure. Research has shown that long-range interactions are important in the β-class of proteins (Gromiha and Selvaraj, 1999; Gromiha et al., 2004). The long-range interactions in Tatil were generated to view the association among the residues (Figure 33). We looked at the long-range interactions that are ten or more residues apart in the primary structure but are within seven angstroms in the tertiary structure. The network shows that the majority of amino acids are connected by one or more

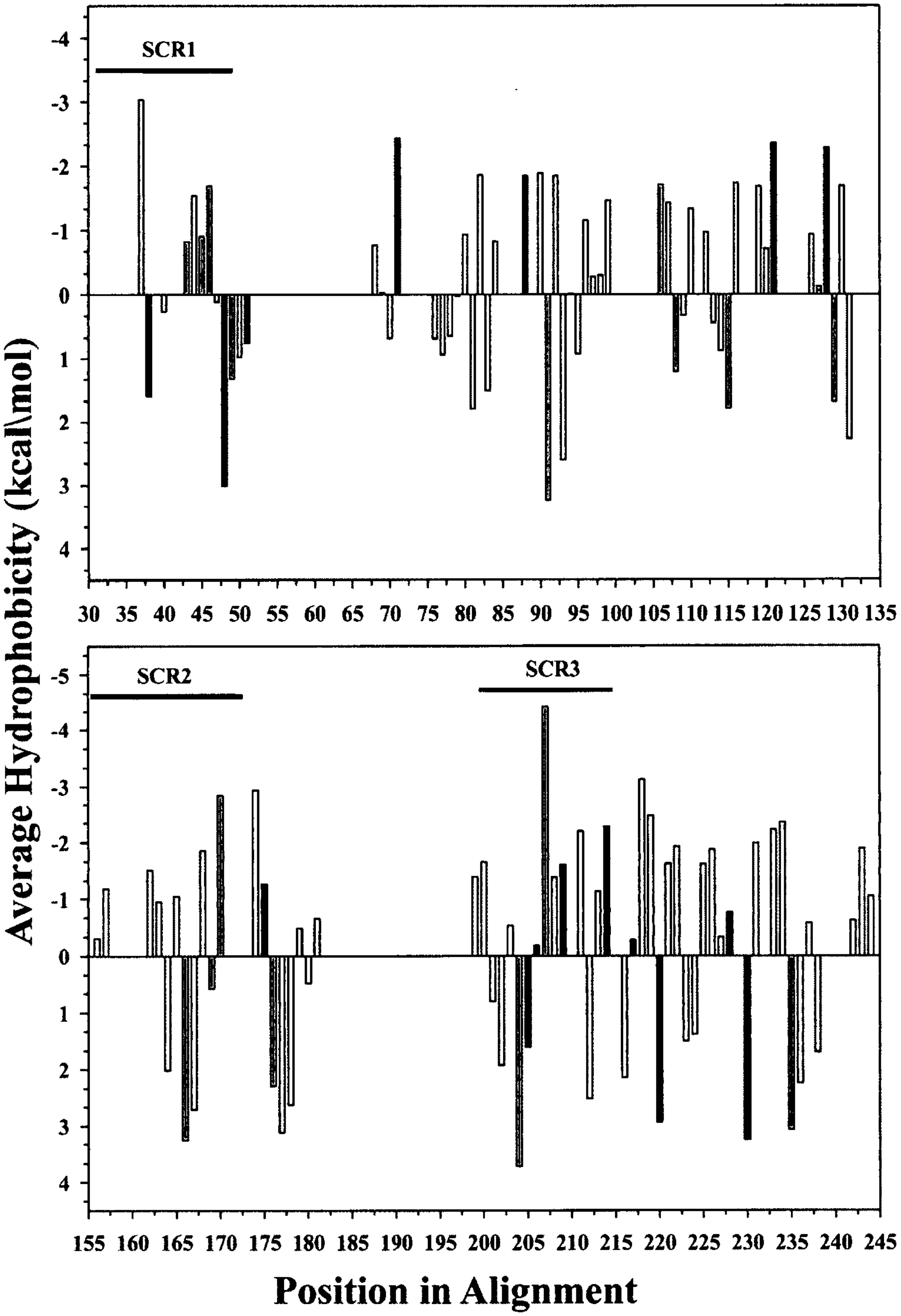


Figure 31. Hydropathy Analysis of the Lipocalin Superfamily

Hydropathy analysis of the conserved positions of the amino acids in the lipocalin superfamily. Red shows the highly conserved positions while blue shows the moderately conserved postions. The location of the SCRs are indicated by a black line.

long-range interactions. However, a small number of the conserved amino acids are not involved in the long-range network, residue numbers 37, 53, 79, 139, and 153 from the moderately conserved positions and 25, 66, 68, 111 and 112 from the highly conserved positions. Twelve out of the seventeen highly conserved residues and eleven out of the sixteen moderately conserved residues are seen in the network. Out of the conserved residues that were not involved in the long-range interactions, only two positions were hydrophobic (111 and 112). The rest of the positions were hydrophilic indicating that the hydrophobic positions are essential for the long-range interactions. The result shows that there are conserved residues that are evolutionarily conserved but not participate in any kind of long-range interactions. The conserved residues that are involved in the long-range interactions might be essential for the protein folding and stability.

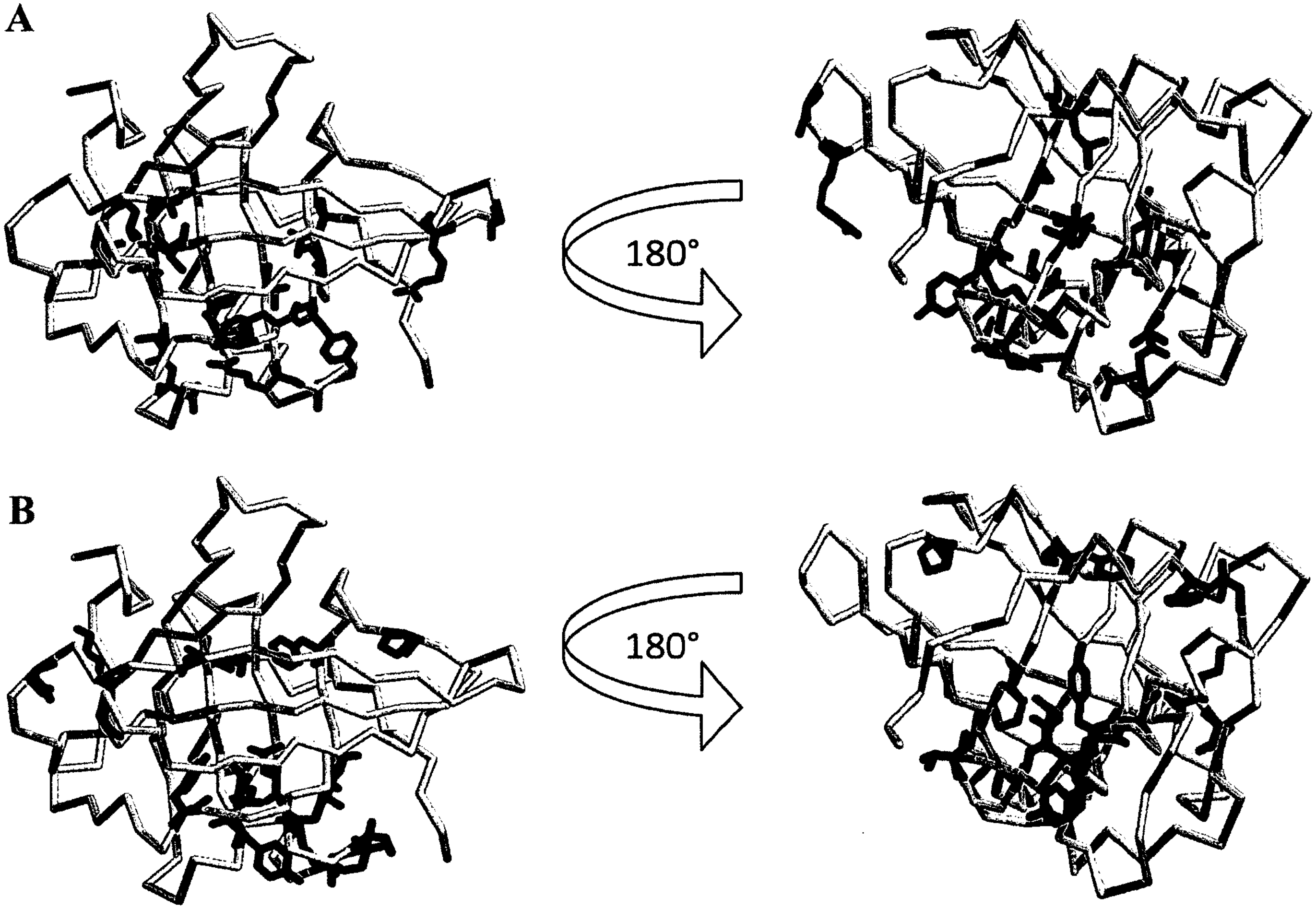


Figure 32. Location of Highly and Moderately Conserved Residues

The figure shows the positions of conserved residues in the Tatil SWISS-MODEL. A. Highly conserved residues highlighted in red. B. Moderately conserved residues highlighted in blue.

3.6 CONCLUSION

Bioinformatics is an essential component in the analysis of proteins. We used homology programs to construct a proposed structure for Tatil. The 3D models suggests that the protein has a lipocalin fold. The model was used to study the function and localization of the protein in a cell (Chapter 4). We also aligned lipocalins with various functions from different organisms to study the lipocalin superfamily. The conservation analysis showed the positions that were conserved throughout evolution. The study also showed a number of positions that are conserved which might be important for the stability and folding of the protein. Long-range interactions of Tatil also showed the presence of significant number of interactions among the amino acids. The bioinformatics study gives an insight into which amino acids may be important in the strucutre, stability and folding of the protein.

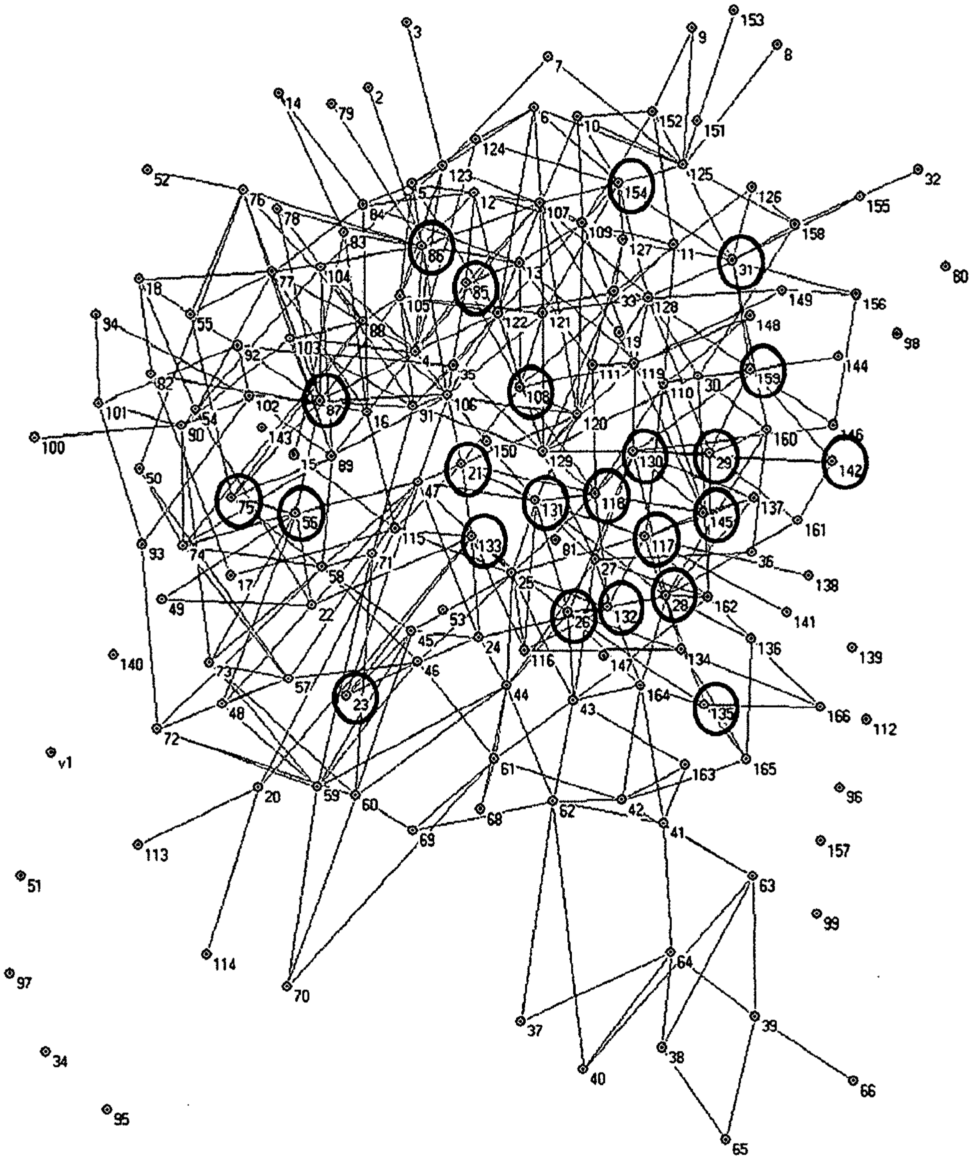


Figure 33. Long-Range Interactions of 9TAT

The illustration shows the long-range interaction between amino acids in Tatil. Orange dots represent amino acids while the grey color shows the interaction among the amino acids. The highly conserved positions are circled in red and the moderately conserved positions are circled in blue. This network was drawn with Pajek.

CHAPTER 4

STABILITY AND CELLULAR LOCALIZATION OF TATIL

4.1 BACKGROUND

Food, without question, is an essential part of our survival. Throughout the world, different types of food from vegetables to cereals are grown to sustain human life. As the number of people increase, it becomes essential to grow enough food to feed the whole population. It is projected that in the year 2100, the world's population would increase from the current 7.2 billion to 9.6 - 12.3 billion (Gerland et al., 2014). However, the changing climate condition is making food production difficult. Thus, food sustainability is one of the 21st century's greatest problems.

In a 2005 paper, the potential impact of climate change on different parts of the world such as North America, South America, Asia and Africa was reviewed, based on climate change projections (Parry et al., 2005). The study showed that within the coming years, crop yields would dramatically decrease and the effect of global temperature increases would have the greatest impact on Africa and Asia. One of the crops that would be greatly impacted is wheat, which is one of the most important crops in the world (Curtis and Halford, 2014). Its production not only has an effect on food levels but also on economics. Thus, the growth and maintenance of crops, especially wheat, is essential.

In the search for food sustainability, plant lipocalins can be one possible avenue to help crops grow in different conditions. Research has shown that the overproduction of the temperature-induced lipocalin in Arabidopsis helps the plant withstand stress that is caused by freezing and heat shock (Charron et al., 2008). Moreover, the protein associates with the plasma membrane even though the exact method of association was not determined. Therefore, the study of the plant lipocalin structure, function and its interaction with cellular membranes can help us understand how this protein can enhance a crop resilience to climate changes. As wheat is a major crop that is grown throughout the world, the study of the temperature induced lipocalin in wheat is essential to aid the growth of wheat in various climates.

Using bioinformatics and experimental approaches we investigated whether or not the wheat lipocalin is localized to the plasma membrane. The bioinformatics studies were used to model the possible structure of the wheat lipocalin (chapter 3) and identify amino acids that may associate with the plasma membrane. In the experimental section, a gene gun coupled with confocal microscopy was used to confirm the association of lipocalin with the membrane and to determine the role of three select amino acids identified using bioinformatics. Next, the wheat lipocalin was structurally characterized. The protein was expressed, purified and initially studied with respect to structure, stability and multimeric state.

4.1.1 PLANT LIPOCALINS

In 1996, Bugos and Yamamoto identified a domain in Violaxanthin de-epoxidase (VDE) as a lipocalin (Hieber et al., 2000). Even though this claim has been disputed and the protein reclassified as lipocalin-like (Åkerström, 2006), the research marked the first time a possible lipocalin was identified in plants. VDE is found in the xanthophyll cycle, which is involved in the protection of plants from excess light. Zeaxanthin, a carotenoid involved in the cycle, is produced in the presence of excess light. The amount of light harvested is decreased by zeaxanthin, which dissipates the energy as heat. Under strong light, a decrease in pH in the lumen results in the activation of VDE, which is converted to zeaxanthin to release the energy as heat.

In 2002, two types of proteins were identified as true plant lipocalins (Charron et al., 2002). The expressed sequence tags (ESTs) database of plants was searched for sequence identity and similarity to the lipocalins. ESTs are fragments of cDNA from expressed part of the genome (Parkinson, 2009).The search yielded the temperature-induced lipocalin in *Triticum aestivum* (Tatil-1) and *Arabidopsis thaliana* (Attil) (Charron et al., 2002). The conserved amino acids that are necessary for identification as a lipocalin were located within the SCR regions in Tatil-1 and Attil (Table 7). The Tatil-1 sequence was used to search the GenBank databases and found proteins that are related to Tatil-1. The sequence of these proteins was found to be similar to human apolipoprotein D (ApoD), *Escherichia coli* bacterial lipocalin (Blc), and the American grasshoper

Lazarillo protein (Table 8) (Charron et al., 2005). Based on various analyses such as bioinformatics predictions, structural features and phylogenetic relationships, the results were divided into two groups: temperature-induced lipocalins (TILs), and chloroplastic lipocalins (CHLs) (Charron et al., 2005). Two types of TILs were identified, TIL-1 and TIL-2 which show some differences in the N-terminus (Figure 34). It was shown that CHL has 25% identity and 35% similarity with Tatil-1. TIL genes form proteins that range from 179 to 201 amino acids, while CHL genes encode proteins that range from 328 to 340 amino acids. The molecular mass for TIL ranges from 19 to 23 kDa, while the CHLs molecular mass range from 36 to 39 kDa.

Table 7. Location of the Conserved Residues in the SCRs

	Tatil-1	Attil
SCR1	GLDVARYMGRWYEIASF [Residues(15-31)]	GLNVERYMGRWYEIASF [Residues(12-28)]
SCR2	YWVLYVDDDYQYALV [Residues(105-119)]	YWVLYIDPDYQHALI [Residues(101-115)]
SCR3	ILCRKTHIEEEVNQL [Residues(129-144)]	ILSRTAQMEEETYKQL [Residues(125-140)]

a [a]The table shows the location of the structurally conserved residues found in wheat (Tatil-1) and Arabidopsis (Attil). The numbers in the bracket indicate the amino acid numbers within the protein while the underlined residues are the conserved amino acids that identified the proteins as lipocalins (Charron et al., 2002).

Table 8. Tatil-1 Identity and Similarity with Related Proteins

	Identical (%)	Similarity (%)
ApoD	29	46
Blc	31	54
Lazarillo	23	40

a [a]The table shows the identity and similarity of Tatil-1 with three of the related lipocalin proteins ApoD from human, Blc from *E.coli* and lazarillo from grasshopper (Charron et al., 2002).

```
Tatil-1   MAAKKSGSEMGVVLGLDVARYMGRWYEIASFPNFFQPRDGRDTRATYELMEDGATVHVLN
Tatil-2   MAA------MKVVRNLDLERYMGRWYEIACFPSRFQPKDGANTRATYTLGPDGA-VKVLN

Tatil-1   ETWSKGKRDFIEGTAYKADPASEEAKLKVKFYVPPFLPIIPVVGDYWVLYVDDDYQYALV
Tatil-2   ETWTDGRRGHIEGTAFRADPAGDEAKLKVRFYVPPFLPVFPVTGDYWVLHVDDAYQFALV

Tatil-1   GEPRRKSLWILCRKTHIEEEVYNQLLEKAKEEGYDVAKLHKTPQSDPPPESDAAPTDSKG
Tatil-2   GQPSRNYLWILCRQPQMDEGVYEELVERAKEEGYDVSKLRKTPHPEPTPESQDAPKDG-G

Tatil-1   TWWFKSLFGK
Tatil-2   LWWIKSLFGK
```

Figure 34. Alignment of Tatil-1 with Tatil-2

Alignment of wheat temperature-induced lipocalins representative of TIL-1 and TIL-2 proteins constructed in Multiple Sequence Comparison by Log-Expectation (MUSCLE). The red color shows the amino acid residues that are identical between the proteins.

4.1.2 FUNCTION OF TEMPERATURE-INDUCED LIPOCALIN

To investigate where Tatil-1 and Attil are found within a cell, a homology search was performed. Homology describes the characteristics that arise as a result of originating from a common ancestor (Lesk, 2002). The results showed that Tatil-1 and Attil share significant similarity with three other lipocalins: the human ApoD, Blc, and the Lazarillo protein (Charron et al., 2002). As all three of the lipocalins are membrane proteins, it is proposed that Tatil-1 and Attil are also membrane associating proteins (Charron et al., 2005). Moreover, Blc, ApoD, and Lazarillo lipocalins all appear to be expressed in response to conditions that cause membrane stress, which suggests the biological role of Tatil-1 and Attil might be to assist plants to withstand stress conditions.

The possible functions of TILs were investigated by subjecting spring wheat, winter wheat, Arabidopsis and several other plants to various conditions. These included low temperature (4°C) and heat-shock treatments (45°C for 1 hour) (Charron et al., 2008). Northern blots, a biological technique used to detect RNA, showed that under low temperature and heat-shock treatments the expression of Tatil-1 increased 10-fold while water-stressed plants only showed a 3.5 fold increase in expression. In Arabidopsis, Attil increased 6-fold and 9-fold at low-temperature and during heat-shock treatments, respectively. The study showed that Tatil-1 and Tatil-2 accumulates in winter tolerant as well as spring wheat.

Further experiments were conducted on Arabidopsis to gain an understanding of the function of the protein (Charron et al., 2008). Four different lines of Arabidopsis were grown: WT, a line that expresses Attil 4-fold (OEX), a line that expresses Attil 2-fold (Comp) and a line that does not express Attil (KO). Under normal conditions the KO line grew the same way as the WT, while the OEX and Comp lines showed a delay in flowering and the green leaf color persisted (stay-green phenotype) for a longer period of time. Western blot analysis of the WT plant in different stages of the life cycle showed that the level of Attil expression is the same throughout. When the plants were subjected to a freezing test at -6°C, only 50% of the KO survived while more than 90% of the Comp and OEX lines survived. Under the freezing test, damage and necrosis were observed on the WT and KO plants, while no damage was seen on the Comp and OEX plants. When the plants were kept in the cold for 7 days, the amount of

Attil increased in all plants except KO plants. The study suggests that the increase in Attil amount is a result of the temperature drop. To test if the different Arabidopsis lines were able to withstand reactive oxygen species, the plants were treated with the herbicide oxidant paraquat. Under this treatment the OEX plants were able to tolerate the herbicide, while the KO plants showed more necrosis than the WT. To study if light affects the different lines of Arabidopsis, the plants were analyzed under various light conditions (Charron et al., 2008). Whether they were grown in the dark or light, no difference in the Attil concentration was observed. When the KO plants that were grown in the dark were placed under a light, they were unable to store chlorophyll and did not survive. However, when WT, OEX and Comp were transferred from dark to light conditions, they were able to survive. The result suggests that the TILs might also be essential for the adaption of the plant to sudden changes in light conditions. To further the understanding of the function of the TILs, the ability of Arabidopsis to withstand nonlethal heat stress called basal thermotolerance and the ability to withstand lethal heat stress called acquired thermotolerance were studied (Chi et al., 2009). Results indicate that TILs may be necessary for both basal and acquired thermotolerance in young plants of Arabidopsis.

It is all possible that TILs could be chaperones. A chaperone protein interacts with a non-native protein to prevent aggregation and promote the proper folding of the protein (Slavotinek and Biesecker, 2001). These proteins play an essential role in response to stress such as heat and salinity. Chaperones are classified based on its molecular size, location within the cell and its function. If a chaperone is unable to fold the protein to its native state the protein is tagged so that it can be degraded. A class of chaperone that is well known is the chaperonin. The class is divided into two groups depending on the presence of co-chaperonin. Group I chaperonins need the co-chaperonins while group II do not.

4.1.3 LOCALIZATION OF TATIL ON THE PLASMA MEMBRANE

Based on the sequence alignment and lipid binding prediction programs, the possible mode of interaction of plant lipocalins with membranes was investigated. Lipocalins are known to associate with membranes in 3 ways. First, the plant lipocalin can associate with a plasma membrane through the N-terminal signal peptide. However, the alignment of Tatil and Attil with the ApoD, Blc and lazarillo lipocalins suggest that they do not possess an N-terminal signal peptide (Charron et al., 2002). The second method of attachment could be through a glycosylphosphatidylinositol (GPI) anchor. A third form of attachment for lipocalins is through a hydrophobic loop, as seen in ApoD (Bishop et al., 1995). In ApoD, the protein may interact with the plasma membrane through the hydrophobic loop between β-strands G and H (Bishop, 2000; Oakley et al., 2012).

In 2005, it was shown that plant lipocalin proteins do not have a GPI anchor (Charron et al., 2005). Further, a localization study was done to observe where TIL-1 fused with GFP would be located in a cell. Attil was cloned in a pAVA321 vector which has GFP. The plasmid DNA was coated onto M17 tungsten particles and delivered onto an onion epidermis. The localization of the lipocalin was then successfully tracked using the GFP. GFP is a naturally fluorescent protein that is found in Aequorea jellyfish (Shimomura et al., 1962). It is an 11-stranded β-barrel with the chromophore buried in the center of the barrel (Figure 35) (Sayle and Milner-White, 1995). The diameter of the cylinder is about 30 Å and it is 40 Å long (Chalfie and Kain, 2006). The fluorophore of GFP is a Ser-Tyr-Gly sequence, which is modified to a (4-(p-hydroxybenzylidene)-imidazolidin-5-one structure (Tsien, 1998). Cyclization between Ser64 and Gly67 forms an imidazolin-5-one intermediate. The Tyr is then oxygenated by O_2. The β-barrel of GFP surrounds the fluorophore. In the study, Attil:GFP was seen on the plasma membrane. In a control with just GFP, the protein was seen throughout the cell.

In this dissertation research, we wanted to investigate if Tatil-1 also localized on the cell membrane and identify the amino acids that are responsible for the localization. Since we focused only on Tatil-1, we refer it as Tatil for the remainder of this dissertation. The GFP that is used for the present localization study of Tatil in plants is a variant: Ser65Thr. This mutation gives a higher fluorescence intensity when compared to the WT. In our work, GFP was used to track the localization of Tatil.

Programs that predict transmembrane proteins or regions scan for a patch or patches of hydrophobic sequences (Zvelebil and Baum, 2008). The patch is identified by calculating the hydrophobicity of individual amino acids within the sequence. The program that we used to check if Tatil is a transmembrane protein is the PRED-TMR algorithm (Pasquier et al., 1999). The program is fully

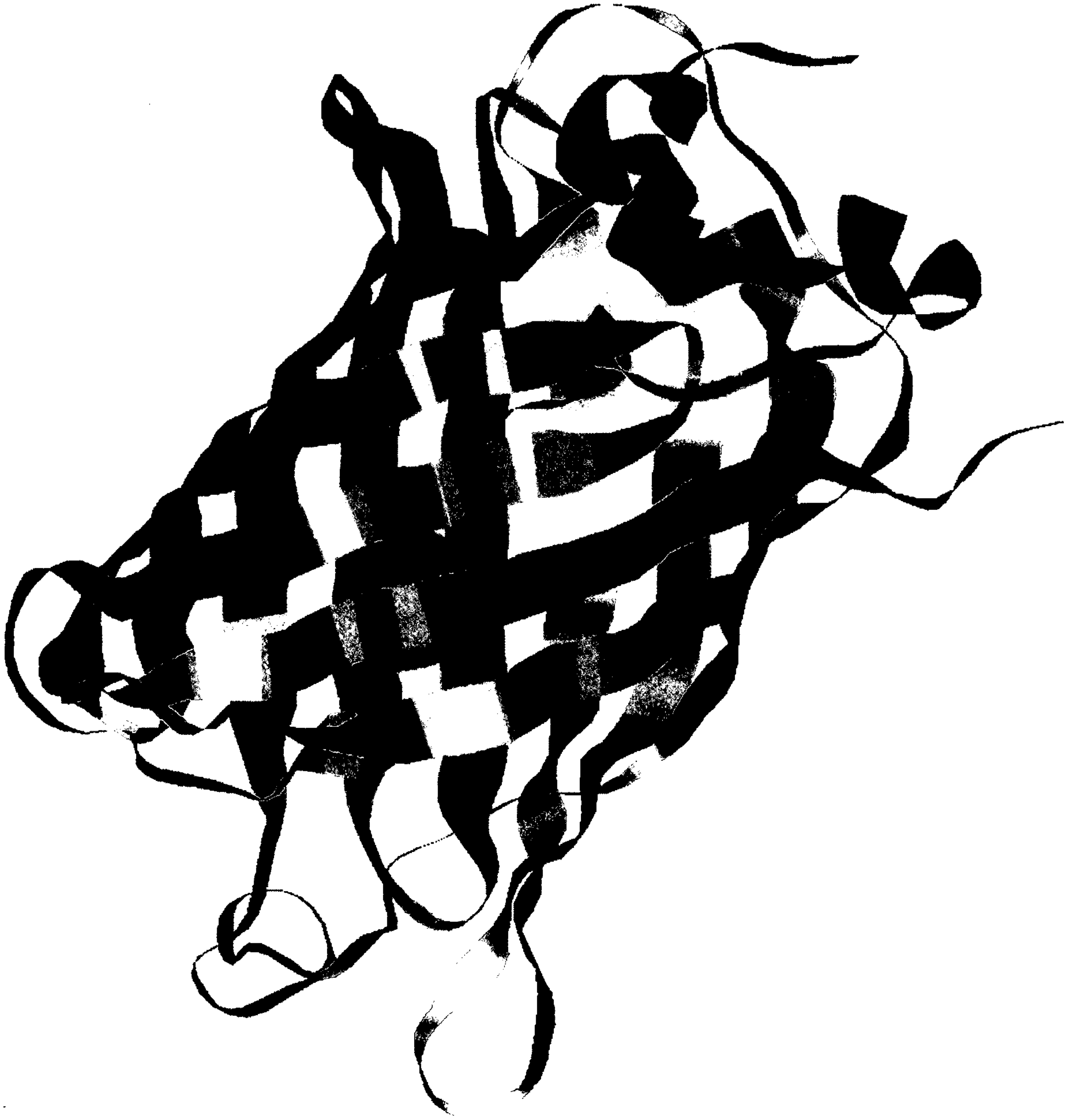


Figure 35. 3D Structure of Green Fluorescent Protein

A ribbon drawing of GFP is illustrated using Rasmol (PDB Code:1GFL). The color scheme indicates the flow of the polypeptide chain going from blue (N-terminus) to red (C-terminus).

automated and only requires the input of the sequence of interest. Based on the homology models, Tatil it is thought to associate with a membrane through the hydrophobic loop between β-strands E and F (Charron et al., 2002).

4.1.4 AMINO ACID SOLVENT ACCESSIBILITY

Amino acid solvent accessibility methods can be used to study how proteins associate with biological membranes. By looking at the secondary structure and the solvent accessibility of certain amino acids, we may learn about the interactions between different amino acids and about protein function. One way to predict solvent accessibility is by looking at the hydrophobicity of the amino acids that are buried or exposed. Certain programs look at the sequence and secondary or tertiary structure to predict which amino acids are solvent accessible. These programs can look at each amino acid separately or look at a patch of amino acids (Rost and Sander, 1994). Alignment of different proteins can also be used to predict solvent accessibility, as it is preserved throughout evolution within a family (Rost, 1996).

ASAView algorithim was used to calculate the solvent accessibility of Tatil (Ahmad et al., 2004). This program is unique in that it not only calculates solvent accessibility but also shows the results in a spiral form for easy visualization. ASAView uses a different algorithim to calculate solvent accessibility of each amino acid if the structure is known. It applies an absolute surface area calculation using a linear neural network rather than assigning buried or exposed thresholds. The program was used to select the possible amino acids responsible for the association of the protein with a plasma membrane, since they should have high solvent accessibility.

4.1.5 HYDROPATHY SCALE

In a protein structure, a distinct area of hydrophobicity or hydrophilicity can be observed. The protein character within a given region can provide information on a protein's function. Thus, exploring the side chain characters of a protein sequence is imperative. The analysis is based on a hydropathy scale, where each amino acid is assigned a value that shows its propensity towards either hydrophobicity or hydrophilicity. The scale shows the average hydrophobicity of the amino acids. Even though there are a variety of hydropathy scales, a highly reliable method is the Kyte and Doolittle method (Kyte and Doolittle, 1982). In the Kyte and Doolittle method, the average number of residues used to determine hydrophobicity is between 5 and 9. Regions with positive values are considered hydrophobic while negative values are hydrophilic. A hydropathy scale was used to identify the hydrophobic regions within Tatil. As transmembrane regions tend to be hydrophobic, the scale gave additional information about Tatil function.

4.1.6 pAVA321 VECTOR

In order to conduct studies to investigate if and how Tatil associates with a lipid membrane, the cDNA that codes for Tatil was inserted in a vector that contains GFP called pAVA321. The cloning vector has a dual 35S promoter from the cauliflower mosaic virus (CaMV), the translational enhancer sequence of the tobacco etch virus (TEV), a modified GFP cDNA, and the 35S polyadenylation signal from CaMV (Von Arnim et al., 1998). The GFP is based on mGFP4 where the expression of GFP is restored in plants by modifying the codon of WT-GFP. The GFP has a mutation S65T, so that it can give a specific absorption (488 nm) and emission peak (511 nm (Cubitt et al., 1995). The vector pAVA321 has restriction enzymes NcoI at the amino terminal and BglII and XbaI at the carboxy terminal (Figure 36) which are used for cloning proteins.

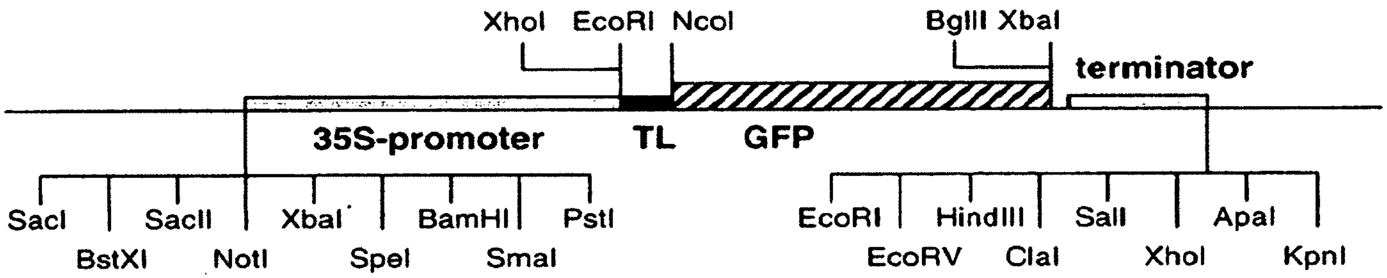


Figure 36. pAVA Vector

The structure for the cloning vector pAVA321 with cloning sites NcoI, BglII and XbaI. The location of the translational enhancer (TL) and the green fluorescent protein (GFP) is shown (Von Arnim et al., 1998).

4.1.7 GIBSON ASSEMBLY

The Gibson Assembly is a new introduction to the arsenal of a recombinant DNA technology. It is a method used to join DNA molecules in a few steps (Gibson et al., 2009). The DNA molecules that are to be combined need to have overlapping sequences of at least twelve nucleotides.The cloning kit consists of a master mix and NEB 5-alpha competent *E. coli* cells. The exonuclease in the master mix cuts up the nucleotide from the 5' end of the DNA molecules creating a complementary overhang between them. DNA ligase attaches nicks in the assembly, annealing the DNA molecules.

4.1.8 HELIOS GENE GUN

The helios gene gun is a physical method used to bombard M17 tungsten particles carrying Tatil cDNA into onion skin for cell transformation. The gene gun can be used for both transient and stable transfection. In transient transfection, the gene is expressed in high amounts for a short period of time and the gene does not get assimilated into the host chromosome (Sambrook et al., 2001). On the other hand, in stable transfection, the target gene gets assimilated into the host chromosome. To deliver the cDNA into the cell, there are two choices for the microcarrier of the DNA, gold or tungsten. Tungsten is toxic for animal cell lines but not for plants, while gold is inert and not toxic. Gold is smooth while tungsten is jagged which can cause damage to cell walls. Gold is more dense and has more inertia than tungsten.

The blueprint of the flowing helium gun, a type of helios gene gun, used in this experiment was developed in 1992 based on physical bombardment (Finer et al., 1992) (Figure 37). The concept behind the gene gun is to accelerate the delivery of DNA tungsten coated particles using a stream of low pressure helium into cells for gene expression and transformation. The flowing helium gun has the same concept as a gene gun that can be commercially bought. The helium gun contains a vacuum chamber to accelerate the particles more efficiently and protect the cell from damage. A time relay solenoid is attached to the top of the chamber to control the amount of helium delivered.

The vacuum chamber was made from a steel plate and measured 16.5 x 16.5 x 30.5 cm (length x height x width). The door was made using 2.5 cm thick

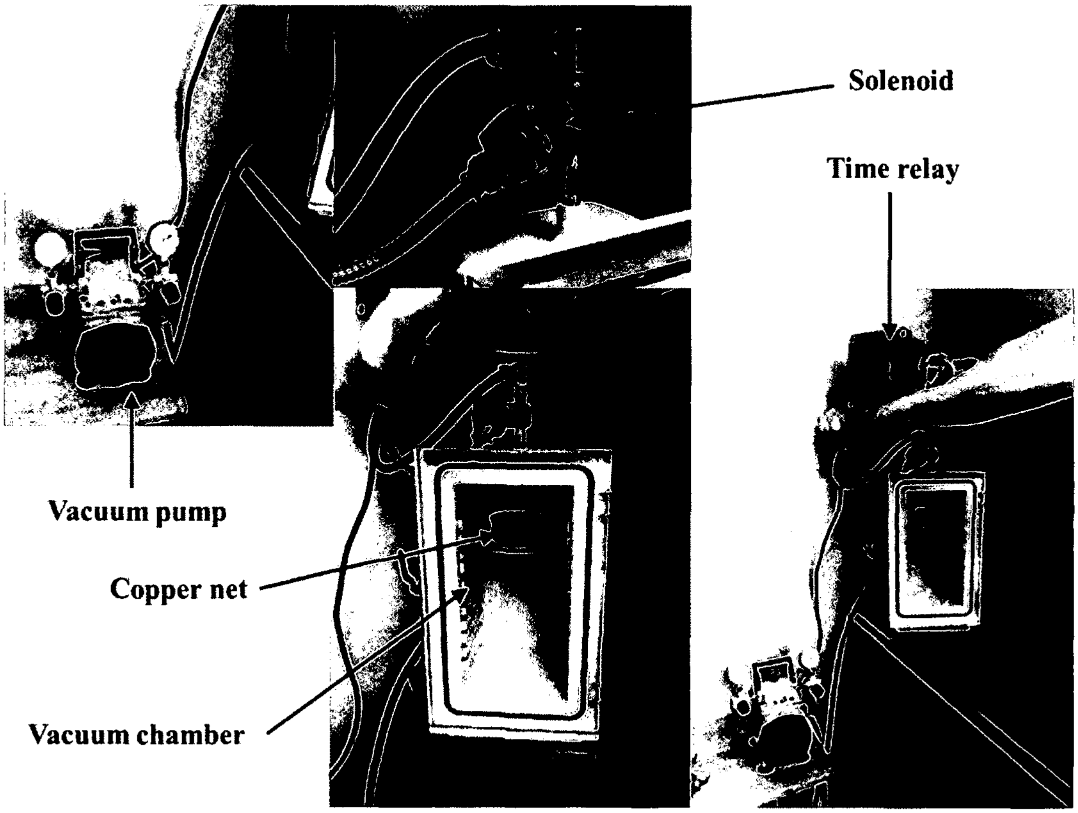


Figure 37. Flowing Helium Gene Gun

A picture is shown of the gene gun that was used for our experiment. The image shows the vacuum chamber with the solenoid attached on the top.

plexiglass and 6.4 mm thick rubber gasket to seal the door shut for the vacuum. The door was attached to the chamber with two hinges. The vacuum pump was attached to the chamber by a hole that is drilled on the side. A valve on the side of the chamber was used to open and close the connection to the vacuum pump. A two-way solenoid was connected to the top of the chamber and attached to a time relay. A helium tank was connected to the other opening of the solenoid. The solenoid was attached to the vacuum chamber through the compression fitting. A Leur-lok syringe adaptor was attached to the fitting. Inside the vacuum chamber, grooves were constructed to fit a number of pexiglasses which hold a copper net to filter out aggregated particles and to hold the sample at different distances from the helium inlet. This system was built by Jean-Benoit Charron at McGill University Canada.

4.1.9 FLUORESCENT MICROSCOPE

In the fluorescent microscope, an excitation light from a lamp is brought to the sample by reflecting it off a dichroic mirror (Lichtman and Conchello, 2005; Chandler and Roberson, 2009). Fluorescent molecules in the specimen absorb the light, then emit light with a longer wavelength. The phenomena is based on Stokes law which states that when electrons go from an excited state to the ground state, there is a loss of vibrational energy. Thus, the emission spectrum is shifted to longer wavelength than the excitation spectrum. The emitted light passes through the objective lens and the dichroic mirror so that the image can be seen through the eyepiece. The dichroic mirror reflects light with shorter wavelengths while transmitting light with longer wavelengths. The emitted light that passes through the dichroic mirror passes through a barrier filter next. The barrier filter blocks any light that was not emitted.

4.1.10 CONFOCAL MICROSCOPE

A confocal microscope is another instrument that can be used to view fluorescent molecules. It has an Acousto optical tunable filter (AOTF) which consists of either a tellurium dioxide or quarz anisotropic crystal, to which a piezoelectric transducer and acoustic absorber are attached (Paddock, 1999; Claxton et al., 2006) (Figure 38). The crystal has optical properties that are altered in the presence of an acoustic wave. The AOTF controls the wavelength and amplitude of light from a laser. Changing the frequency of the transducer applied to the crystal changes the period of the diffraction index, and therefore the wavelength of light that is diffracted. Changing the amplitude of the wave changes the intensity of the light. The acoustic and optic waves move through the crystal at different angles, thus the diffracted and the undiffracted beams are separated when they come out of the crystal. The undiffracted beams are blocked and the diffracted beams enter the confocal scan head. A set of scanning mirrors and objective lens are used to scan the specimen with an excitation beam. The mirrors are controlled by a galvanometer and can be adjusted using a computer. The emitted light passes through the dichroic mirror as well as a confocal pinhole to reach a series of photomultiplier detectors. The signal from the photomultipliers is then amplified by dynode voltage. A confocal microscope was used to visualize the location of Tatil within the onion cells.

4.2 PROTEIN EXPRESSION SYSTEMS

There are several fundamental cell lines that are used to express proteins. The main ones are bacterial, mammalian, insect and yeast (Yokoyama, 2003; Chen, 2012; Mattanovich et al., 2012). The most widely used is the bacterial system as the components needed for protein expression are relatively inexpensive and bacteria can grow to high density within a short amount of time (Terpe, 2006). The main bacterial host that is used is *Escherichia coli*. Expressed proteins are found in the cytoplasm or in periplasmic space of *E. coli*. Problems can arise in protein expression as a result of differences between the codon usage and the type of codon required in the desired protein cDNA. An example of codon usage difference can be seen when a eukaryotic protein is expressed in a prokaryotic cell. For example, *E. coli* have a bias in the type of codon they use for each amino acid. Thus, the tRNAs available for translation are different for each organism. If the sequence of the protein requires codons that are not frequently used by *E. coli* then the translation is not going to be effective. The need for rare tRNAs causes translational stalling, premature translation termination, translation frameshift and amino acid mis-incorporation (Kurland and Gallant, 1996). For *E. coli*, the codons AGG, AGA, CUA, AUA, CCC, CGA are rarely used, which makes expressing proteins with these codons difficult. This problem can be solved by using hosts that supply the rare tRNAs.

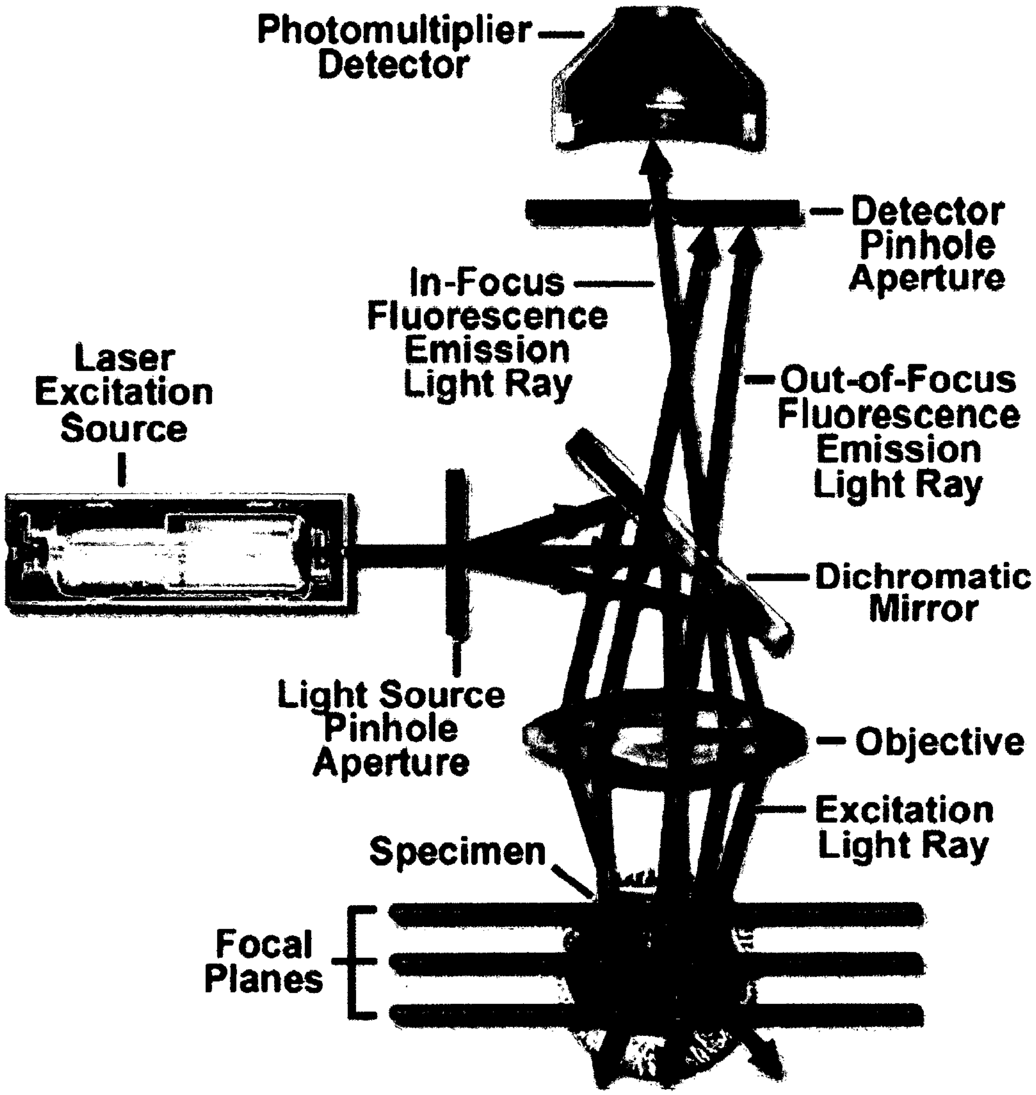

Photomultiplier Detector
Detector Pinhole Aperture
In-Focus Fluorescence Emission Light Ray
Laser Excitation Source
Out-of-Focus Fluorescence Emission Light Ray
Dichromatic Mirror
Light Source Pinhole Aperture
Objective
Excitation Light Ray
Specimen
Focal Planes

Figure 38. Schematic Diagram of a Confocal Microscope
Instrument components and the optical pathway in a confocal microscope. Figure adapted from (Claxton et al., 2006).

One of the *E. coli* cells that is commonly used in experiments is BL21(DE3). This strain contains a prophage λDE3 that expresses T7 polymerase from the lacUV5 promoter when isopropyl β-D-1-thiogalactopyranoside (IPTG) is added. Before induction of lacUV5, little to no T7 RNA polymerase and protein of interest is found in the cells. However, when the lacUV5 is induced with IPTG, the T7 polymerase is synthesized which in turn synthesizes the protein of interest.

4.2.1 AFFINITY TAG

One way to purify proteins from crude extracts is to use an affinity tag. Tags allow purification of specific proteins and removal of cellular debris without applying any additional steps (Lichty et al., 2005). Affinity tags can be classified into three classes based on the tag and the protein of interest. The first category consists of a protein or peptide tag that binds to a ligand. A good example of this is a polyhistidine tag that binds to an immobilized metal (Terpe, 2003). The metals are transition metals such as: Co^{2+},Ni^{2+}, Cu^{2+} and Zn^{2+}. Histidine has a strong interaction with the immobilized metal ions by forming coordination bonds using the imidazole ring. During purification, proteins that have histidine tags attach to the metal ions, while those that lack the tag pass through the column. The protein of interest then can be eluted by changing the pH of the column or by using a high concentration of imidazole to displace the protein. In the second category, a peptide tag binds to a protein in a chromatography resin. An example is the camodulin tag (Stofko-Hahn et al., 1992). Since the resin also contains camodulin, a protein that has the tag can easily be purified. In the third category, an antibody is in the resin and can bind to a specific protein (Swinnen et al., 2007). In the expression of Tatil, we chose to use a polyhistidine tag on the N-terminus. The tag can be added by incorporating the vectors pTrcHis and pET14b. As the vectors contain the sequences for enterokinase and thrombin binding sites next to the tag, the histidines can be removed after purification.

4.2.2 MONITORING THE STRUCTURE AND THERMODYNAMIC STABILITY OF PROTEINS

Circular dichroism (CD) uses plane polarized light to study the structural changes of a protein (Greenfield, 2007). Plane polarized light is made up of circularly polarized light with one rotating counter-clockwise (left handed, L) and one clockwise (right handed, R). The differential absorption of the components is referred to as CD. If L and R are absorbed equally, the combination of L and R would generate radiation polarized in the original plane. However, if L and R are absorbed unequally then the radiation generated would be called elliptically polarized.

The CD instrument measures the difference in absorbance between the two different polarized lights (L and R). CD can be measured in three different ways (Kelly et al., 2005). The first is modulation, where the incident light is continuously switched between L and R components. The second one is direct subtraction, where the absorbances of L and R are taken separately and then subtracted from each other. The third method is ellipsometric, where the ellipticity of the transmitted radiation is measured.

A CD signal is observed when a chromophore is chiral. Chirality of the chromophore can be a result of the chromophore having a chiral structure, being linked to a chiral center in the molecule or being found in an asymmetric environment as a result of the 3D structure. In proteins, the peptide bonds, aromatic amino acid side chains, disulfide bonds and non-protein cofactors are capable of absorbing lights and are, thus, chromophores. Secondary structure composition of a protein can be gathered from the peptide bond region. Absorption for secondary structures is seen at 240 nm or below. Absorbance at 260-320 nm is a result of aromatic amino acids (Sreerama et al., 1999). Each of the aromatic amino acids have their own characteristic wavelength. Tryptophan has a peak at 290 nm with fine structure between 290 and 305 nm. Tyrosine has a peak between 275 and 282 nm with a shoulder at a longer wavelength. Phenylalanine has a peak between 255 and 270 nm. In near-UV CD (260 nm-320 nm), the shape and magnitude of the peaks for a given protein is determined by various factors, such as the number and type of aromatic amino acids, the nature of their environment and their spatial disposition in the protein.

4.3 MATERIALS AND METHODS

Based on the bioinformatics studies in chapter 2 three amino acids were selected to be mutated for the membrane localization study. The hydrophobicity and solvent accessibility of the amino acids were confirmed using the Expasy server (Wilkins et al., 1999) and ASAView (Ahmad et al., 2004), respectively. The

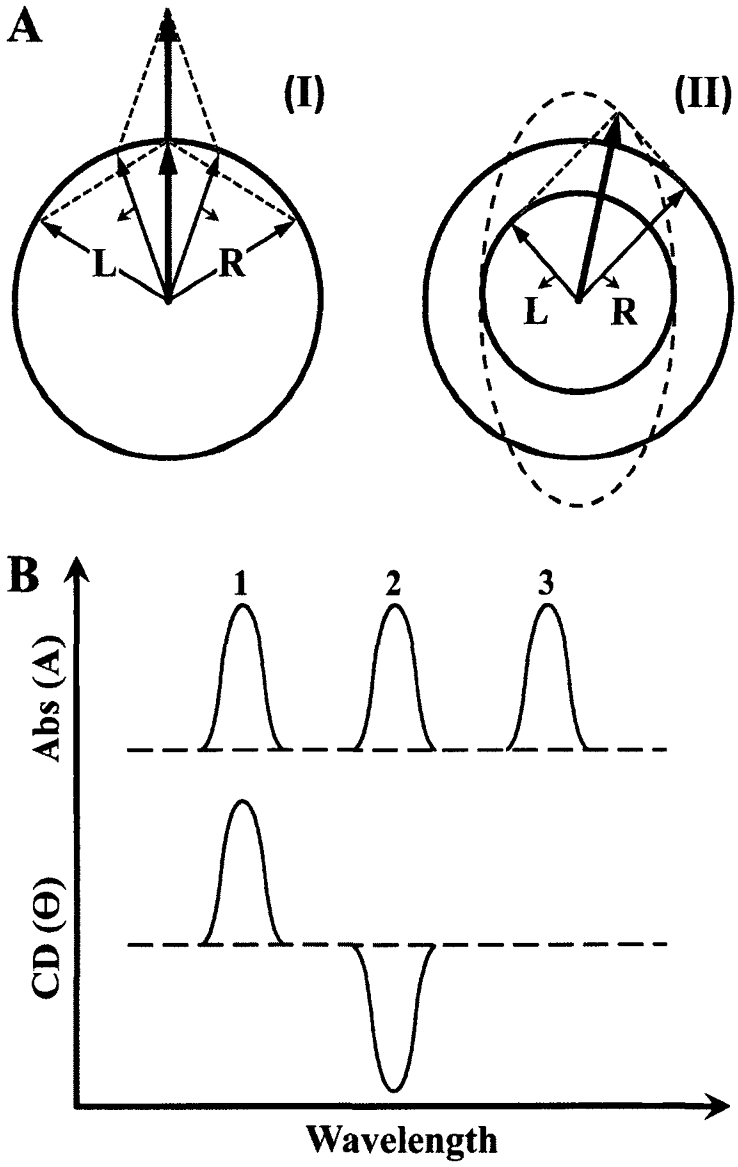


Figure 39. An Illustration of the CD Effect

A. A schematic showing left (L) and right (R) component of a plane polarized light. It shows the combination of L and R when they have the same magnitude (I) and when the components have different amplitude (II). B. Picture illustrates the relationship between absorption and CD spectra where band 1 shows a positive CD spectrum with L is absorbed more than R. Band 2 shows a negative CD spectrum with R absorbed more than L. Band 3 shows a spectrum when L and R have equal amplitude (Collins, 2015).

interaction of the amino acids was studied using a gene gun and onion cells. The experiments conducted in sections 4.3.2-4.3.7 were conducted in the Charron Laboratory (McGill University, Canada)

4.3.1 PLASMID FRAGMENTS

The double-stranded cDNA encoding Tatil was initially designed and then synthesized by Integrated DNA Technologies, Inc. (Coralville, Iowa) for direct insertion into the vector pAVA321. The fragments of Tatil also consisted of 12 or 25 extra sequences that are part of the pAVA321 vector with the restriction enzymes BglII and XbaI flanking the 5' and 3' ends of the fragments, respectively. Subsequently, three different cDNAs were designed, each with mutations that resulted in a select codon change which corresponds to an amino acid change (Table 9).

Table 9. Wild-Type and Variants cDNA of Tatil

	DNA Sequence
Wild-type	GATGAACTATACAGATCTATGGCGGCCAAGAAGAGCGGGAGCGAAATGGGCGTGGTGCTGGGGCTGGACGTTGCGCGGTACATGGGGCGGTGGTACGAGATCGCGTCCTTCCCCAACTTCTTCCAGCCGCGCGACGGGCGCGACACGCGGGCGACCTACGAGCTCATGGAGGACGGCGCCACGGTGCACGTGCTCAACGAGACGTGGAGCAAAGGGAAGCGCGACTTCATCGAGGGCACCGCCTACAAGGCCGACCCGGCCAGCGAGGAGGCCAAGCTCAAGGTCAAGTTCTACGTCCCGCCCTTCCTCCCCATCATCCCCGTCGTCGGCGACTACTGGGTCCTCTATGTCGACGACGACTACCAGTATGCCCTCGTCGGCGAGCCCCGCCGGAAAAGCCTATGGATCCTGTGCAGGAAGACGCACATCGAGGAGGAGGTGTACAACCAGCTGCTGGAGAAGGCCAAGGAGGAAGGCTACGACGTGGCCAAGCTGCACAAGACGCCGCAGAGCGACCCGCCGCCGGAGAGCGATGCCGCGCCCACCGACAGCAAAGGGACCTGGTGGTTCAAGTCGCTCTTTGGTAAAATCTAGAGTCCGCAAAAAT

Table 9. Continued

Mutation	DNA Sequence
Phe96Gly	ATGGCATGGATGAACTATACAGATCTATGGCGGCCAAGAAGAGCGGGAGCGAAAT GGGCGTGGTGCTGGGGCTGGACGTTGCGCGGTACATGGGGCGGTGGTACGAG ATCGCGTCCTTCCCCAACTTCTTCCAGCCGCGCGACGGGCGCGACACGCGGGC GACCTACGAGCTCATGGAGGACGGCGCCACGGTGCACGTGCTCAACGAGACGT GGAGCAAAGGGAAGCGCGACTTCATCGAGGGCACCGCCTACAAGGCCGACCC GGCCAGCGAGGAGGCCAAGCTCAAGGTCAAGTTCTACGTCCCGCCCGGCCTCC CCATCATCCCCGTCGTCGGCGACTACTGGGTCCTCTATGTCGACGACGACTACC AGTATGCCCTCGTCGGCGAGCCCCGCCGGAAAAGCCTATGGATCCTGTGCAGG AAGACGCACATCGAGGAGGAGGTGTACAACCAGCTGCTGGAGAAGGCCAAGGA GGAAGGCTACGACGTGGCCAAGCTGCACAAGACGCCGCAGAGCGACCCGCCG CCGGAGAGCGATGCCGCGCCCACCGACAGCAAAGGGACCTGGTGGTTCAAGT CGCTCTTTGGTAAAATCTAGAGTCCGCAAAAATCACCAGTC

Table 9. Continued

Mutation	DNA Sequence
Pro98Gly	ATGGCATGGATGAACTATACAGATCTATGGCGGCCAAGAAGAGCGGGAGCGAAAT GGGCGTGGTGCTGGGGCTGGACGTTGCGCGGTACATGGGGCGGTGGTACGA GATCGCGTCCTTCCCCAACTTCTTCCAGCCGCGCGACGGGCGCGACACGCG GGCGACCTACGAGCTCATGGAGGACGGCGCCACGGTGCACGTGCTCAACGA GACGTGGAGCAAAGGGAAGCGCGACTTCATCGAGGGCACCGCCTACAAGGC CGACCCGGCCAGCGAGGAGGCCAAGCTCAAGGTCAAGTTCTACGTCCCGCC CTTCCTCGGCATCATCCCCGTCGTCGGCGACTACTGGGTCCTCTATGTCGACG ACGACTACCAGTATGCCCTCGTCGGCGAGCCCCGCCGGAAAAGCCTATGGAT CCTGTGCAGGAAGACGCACATCGAGGAGGAGGTGTACAACCAGCTGCTGGAG AAGGCCAAGGAGGAAGGCTACGACGTGGCCAAGCTGCACAAGACGCCGCAG AGCGACCCGCCGCCGGAGAGCGATGCCGCGCCCACCGACAGCAAAGGAC CTGGTGGTTCAAGTCGCTCTTTGGTAAAATCTAGAGTCCGCAAAAATCACCAGTC

Table 9. Continued

Mutation	DNA Sequence
Ile99Gly	ATGGCATGGATGAACTATACAGATCTATGGCGGCCAAGAAGAGCGGGAGCGAAATGGGCGTGGTGCTGGGGCTGGACGTTGCGCGGTACATGGGGCGGTGGTACGAGATCGCGTCCTTCCCCAACTTCTTCCAGCCGCGCGACGGGCGCGACACGCGGGCGACCTACGAGCTCATGGAGGACGGCGCCACGGTGCACGTGCTCAACGAGACGTGGAGCAAAGGGAAGCGCGACTTCATCGAGGGCACCGCCTACAAGGCCGACCCGGCCAGCGAGGAGGCCAAGCTCAAGGTCAAGTTCTACGTCCCGCCCTTCCTCCCCGGCATCCCCGTCGTCGGCGACTACTGGGTCCTCTATGTCGACGACGACTACCAGTATGCCCTCGTCGGCGAGCCCCGCCGGAAAAGCCTATGGATCCTGTGCAGGAAGACGCACATCGAGGAGGAGGTGTACAACCAGCTGCTGGAGAAGGCCAAGGAGGAAGGCTACGACGTGGCCAAGCTGCACAAGACGCCGCAGAGCGACCCGCCGCCGGAGAGCGATGCCGCGCCCACCGACAGCAAAGGGACCTGGTGGTTCAAGTCGCTCTTTGGTAAAATCTAGAGTCCGCAAAAATCACCAGTC

Amino acid three-letter codes: Gly=Glycine, Phe=Phenylalanine, Pro=Proline, Ile=Isoleucine. The codon for glycine is GGC. The color-coding for the highlighted regions are as follows: green (restriction enzyme site BglII), red (restriction enzyme site XbaI) cyan (mutated protein for glycine).

Table 10. Gibson Assembly Reaction Setup

Gibson Assembly	Volume and Concentration
Variant Fragment	2-10 µl (0.02-0.5 pmols) (2 fold excess of vector)
Linearized Vector	2-10 µl (0.02-0.5 pmols)
Gibson Assembly Master Mix (2x)	10 µl
Deionized H_2O	0-6 µl
Total Volume	20 µl

4.3.2 GIBSON ASSEMBLY

The pAVA vector was linearized using BglII and XbaI restriction enzymes (New England Biolabs, Ipswich, MA). The fragments were then inserted into the vector pAVA321 using Gibson assembly (New England Biolabs, Ipswich, MA). picomol of the pAVA vector and Tatil fragment were calculated based on their length and weight. The following formula was used to convert from ng to pmol.

pmols = (weight in ng)*(1,000)/(base pairs * 650 daltons)

The samples were then incubated in a thermocycler (Biorad, Hercules, CA) at 50°C for 15 minutes and placed on ice until transformation. The transformation process was done using NEB 5-alpha competent *E.coli* cells using 2 µl of the sample based on Gibson Assembly Transformation Protocol. The cells were plated on 100 µg/ml ampicillin plates and incubated at 37°C for 16 hrs (Fisher Scientific, Isotemp 650D gravity convection incubator).

4.3.3 PLASMID DNA EXTRACTION

An overnight culture of 250 ml of Luria broth with 100 µg/ml of ampicillin was used for each control and variant. After transformation a colony was taken from the selection plate and added to the broth and placed in a 37°C shaking incubator. Plasmid DNA was extracted using the plasmid DNA extraction Maxi prep kit (Biobasic Canada, Markham, Ontario).

4.3.4 PREPARATION FOR BOMBARDMENT

0.7μm tungsten microcarrier (Bio-rad, Hercules, CA) was suspended in 1 ml of ethanol (Commerical alcohols, Toronto, Ontario). Six washes of the tungsten particles were done by adding 1 ml ethanol and centrifuging at a Vmax of 14,000 rpm on Bio-Rad model 16 K microcentrifuge. After the last centrifuge, the pellet was resuspended in 1 ml of ethanol and kept at -20˚C.

On the day of the bombardment, the tungsten beads were sonicated (Bioruptor standard, Model B01010003) for 1 minute to break up any aggregation. 2 μl of plasmid DNA at a concentration of 2 μg/μl was placed in a 1.5 ml centrifuge tube. After adding 2 μl of 95% ethanol, it was vortexed for 10 seconds. Then, 8 μl of 95% ethanol was added to the sample and vortexed for an additional 10 seconds. 20 μl of the tungsten beads were then added to the tube and vortexed for 15 seconds. The sample was then ready to bombard the onion tissue.

The plates for the onion tissues were prepared using 3% agar (Amresco, Solon, Ohio) in 100 mm Petri dishes. The plates were kept in a 4ºC fridge until use.

4.3.5 BOMBARDMENT CONDITIONS

The solenoid timer was set at 50 msec and the helium pressure was set from 85 to 100 PSI. The vacuum pump was also turned on at this time. 8 μl of the DNA coated beads were spread on top of Tungsten M-17 Microcarrier (Biorad, Hercules, CA) and dried until the smell of ethanol went away. The carrier was then fastened into the gun. The onion tissue was placed on the 3% agar plate and positioned on the shelf that is two slots below the level of the macro-screen. Once the door is closed, the vacuum is applied. When the gauge on the vacuum pump indicated 25 Hg, a pulse of helium was delivered by pressing the 'RESET' button. The plate containing the bombarded cell was then wrapped in aluminum foil and incubated for 24 hrs.

4.3.6 FLUORESCENT MICROSCOPE

Onion cells were checked for the localization of Tatil using a Zeiss SteREO Discovery V20 operated with a SYCOP touch panel. Pictures were taken with a Zeiss axiocam MRC camera. An X-cite series 120Q lumen dynamics was used for spectral fluorescence excitation of GFP.

4.3.7 CONFOCAL MICROSCOPE

After the cells were checked for transformation under the fluorescence microscope, they were taken to the Institute of Parasitology at McGill University for analysis using a confocal microscope. Cells were viewed using a Bio-Rad confocal Radiance 2100 with a Nikon Microscope E800.

4.3.8 EXPRESSION VECTORS

Tatil was originally cloned into a pTrcHis vector (Life technologies, Grand Island, NY). However, after expression our colleagues at McGill University were unable to cutoff the histag using enterokinase. As a result, Tatil was cloned into a pET14b vector (EMD Millipore, Billerica, MA), which has a thrombin cleavage site that based on our experience is more amenable to enzymatic activity (rDNA/IBC# 13-020).

4.3.9 COMPETENT CELLS

Protein expression of Tatil was first attempted using *E. coli* BL21(DE3). This *E. coli* cell line did not yield significant amounts of protein. Thus, another cell line, Rosetta (DE3)pLysS (EMD Millipore, Billerica, MA), was used. This cell improved the yields from approximately 1 mg/L to 2.5 mg/L (rDNA/IBC# 13-020).

4.3.10 CHROMATOGRAPHY

Hispur Ni-NTA resin (Thermo Scientific, Rockford, IL) was used for the affinity column and Sephadex G-75 resin (Sigma, St. Louis, MA) for gel filtration. 25 ml of the Ni-NTA resin was loaded into a 1.5 cm X 30 cm column. The resin was equilibrated with 50 mM Tris-HCl (pH 8.0), 300 mM NaCl, and 2 mM β-mercaptoethanol. Once the lysate was run through the column, a buffer containing 50 mM Tris-HCl (pH 8.0), 300 mM NaCl, 20 mM imidazole and 2 mM β-mercaptoethanol was used to elute any proteins that were loosely attached to the resin. The concentration of imidazole was raised to 250 mM in a gradient fashion to elute the protein of interest. For the second part of the purification, a G-75 resin was first dissolved in H_2O. The resin was left overnight to swell. It was poured into a column with a radius of 1.3 cm and 62 cm in height. Once the resin settled, it was equilibrated with 50 mM Tris-HCl (pH 8.0), 300 mM NaCl and 2 mM β-mercaptoethanol. The gel filtration column was then calibrated using myoglobin (17.6 kDa), deoxyribonuclease (31 kDa) and albumin (66 kDa) to gain an idea of the multimeric state of the wheat lipocalin.

4.3.11 CIRCULAR DICHROISM

CD spectra were obtained using the Jasco J-815 spectropolarimeter (Jasco, Easton, MD). The cuvette pathlength was 0.1 cm and 1 cm for far- and near UV CD, respectively. The protein buffer was 20 mM Tris-HCl (pH 8.0), 150 mM NaCl and 2 mM β-mercaptoethanol. The concentration of the protein was 0.22 mg/ml for peak one and 0.14 mg/ml for peak two. The experiment was done in triplicates.

4.3.12 THERMAL UNFOLDING

Thermal unfolding of the wheat lipocalin was done using a Jasco J-815 spectropolarimeter (Jasco, Easton, MD). Far- and near-UV CD were done using 0.1 cm and 1 cm cuvettes respectively. The protein buffer was 20 mM Tris HCl (pH 8.0), 150 mM NaCl and 2 mM β-mercaptoethanol. Concentration of the protein was 0.22 mg/ml for peak one and 0.14 mg/ml for peak two. The experiment was done in triplicates.

4.3.13 FLUORESCENCE

Fluorescence data was obtained using a Cary Eclipse fluorescence spectrophotometer (Varian, Palo Alto, CA). The concentration of the samples were adjusted to 0.05 mg/ml. The protein buffer was 20 mM Tris HCl, 150 mM NaCl and 2 mM β-Mercaptoethanol (pH 8.0). The excitation and emission wavelengths were set at 290 nm and 340 nm, respectively. The experiment was done in triplicates.

4.4 RESULTS AND DISCUSSION

4.4.1 LOCALIZATION OF AMINO ACIDS IN LIPOCALINS

Recently, the first true plant lipocalins with all three structurally conserved regions were identified from wheat and Arabidopsis (Charron et al., 2005). In the absence of a 3D structure, a model of the desired protein was constructed. We generated 3D structures of the wheat lipocalin (Chapter 3) using five homology methods. These are: SWISS-MODEL, CPH, ESYPred, Robetta and Modeller program. The resulting models had the same basic structure of eight anti-parallel β-strands with some variety at the N- and C-terminus. The models showed that the structure has a large loop between β-strands E and F.

Before we conducted the localization experiment, we analyzed the loops closely in order to make informed decisions on target residues for mutagenesis. The sequence of Tatil was run through the PRED-TMR algorithm, a transmembrane program (Pasquier et al., 1999). It identified amino acids from 90 to 120 as being a potential transmembrane region and this region is found between β-strands E and F also called loop 5 (Figure 40). This result provide additional evidence that the loop may be responsible for cell membrane association. Moreover, a protein analysis tool which is based on the Kyte and Doolittle hydrophobicity scale identified the loop between β-strands E and F as hydrophobic (Wilkins et al., 1999) (Figure 41), which strongly suggests that this loop interacts with a membrane. The plot shows that the loop has a high hydrophobicity score when compared to the rest of the structure. The solvent accessibility program, ASAView, (Figure 43) which looks at the accessible surface area identified several amino acids which are accessible to a solvent, out of which we picked 7 amino acids from loop 5, Pro94, Phe96, Pro98, Ile99, Ile100, Pro101 and Val103 (Ahmad et al., 2004). Only four hydrophobic residues were as solvent accessible as these 7 and they are spread out over the rest of the protein. The high presence of proline in the loop suggests that the loop is structured (Krieger et al., 2005). By looking at the data mentioned above and by comparing the sequence with ApoD, amino acids Phe96, Pro98 and Ile99 were chosen to be mutated to glycine. Glycine was selected as it is the simplest form of amino acid and removes a target side chain without a deletion of the position.

This ensures that only the interactions of that specific amino acid are studied and the loop remains intact. If Tatil associates with the cell membrane as a result of these specific amino acids then the mutation will disrupt the interaction and the protein will not localize on the plasma membrane.

A biolistic process is an ideal method to analyze if Tatil interacts with the plasma membrane through specific amino acids. The method employs high velocity to transfer substances into cells (Sanford, 2000). The process takes a shorter amount of time and the localization of the protein can viewed by following GFP *in vivo*. As strong evidence show that Tatil localizes on the plasma membrane, it might follow a specific path once the protein is translated. In eukaryotes, mRNA is translated to proteins by ribosomes in the cytosol (Figure 42). Thus. initially the Tatil will appear throughout the cell. Once Tatil is translated, it may be directed to the plasma membrane via a targeting sequence.

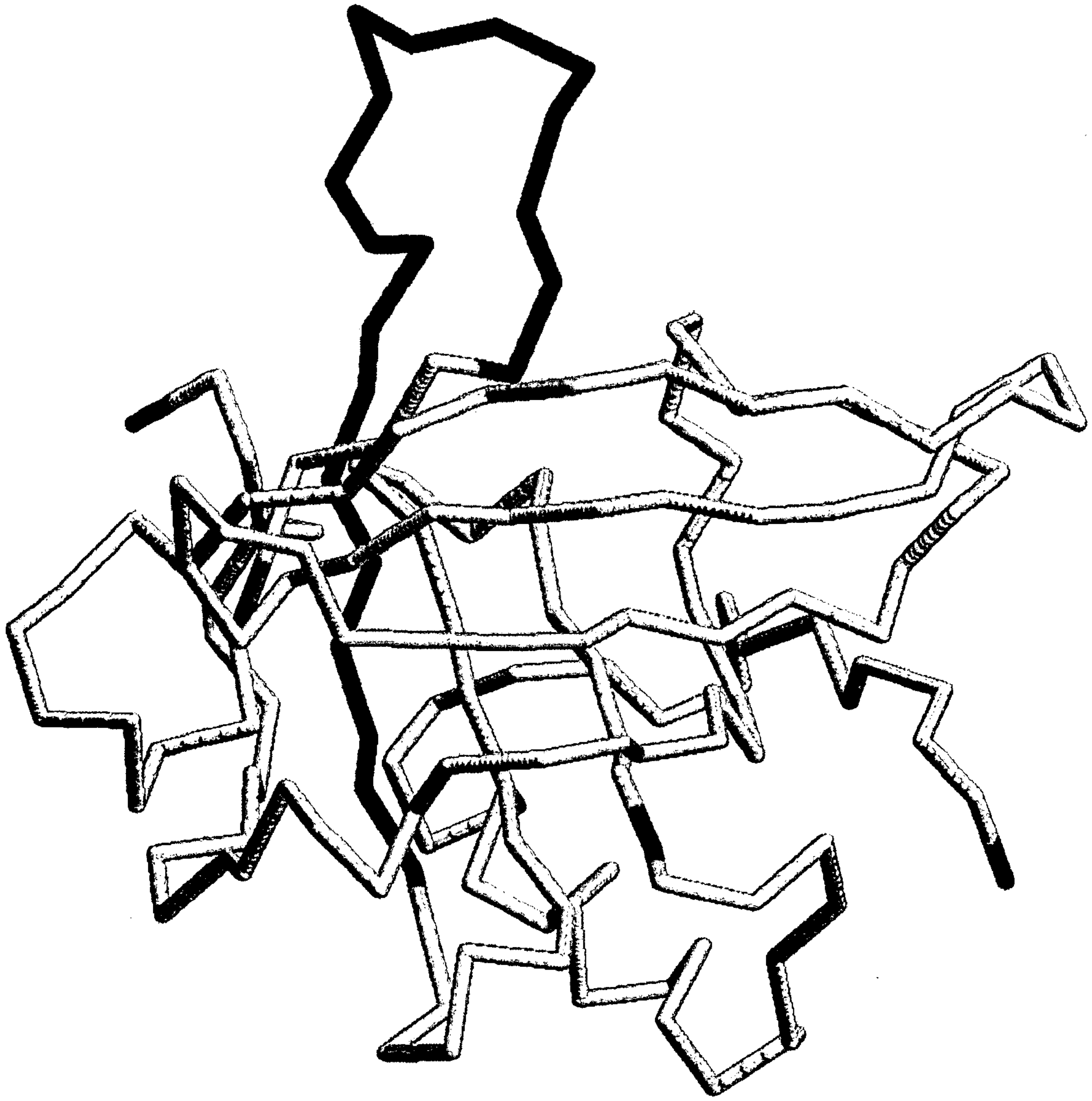


Figure 40. Putative Transmembrane Loop in the 3D Model Structure of Tatil The model was constructed using the SWISS-MODEL program using bacterial lipocalin (2ACO) as the homology template. The model is given the code 9TAT for this dissertation. The N-terminus is colored blue and the C-terminus is colored red, with the green color indicating the location of a cysteine. The purple color shows the loop of the protein that is thought to be associating with the membrane (residues 91-110).

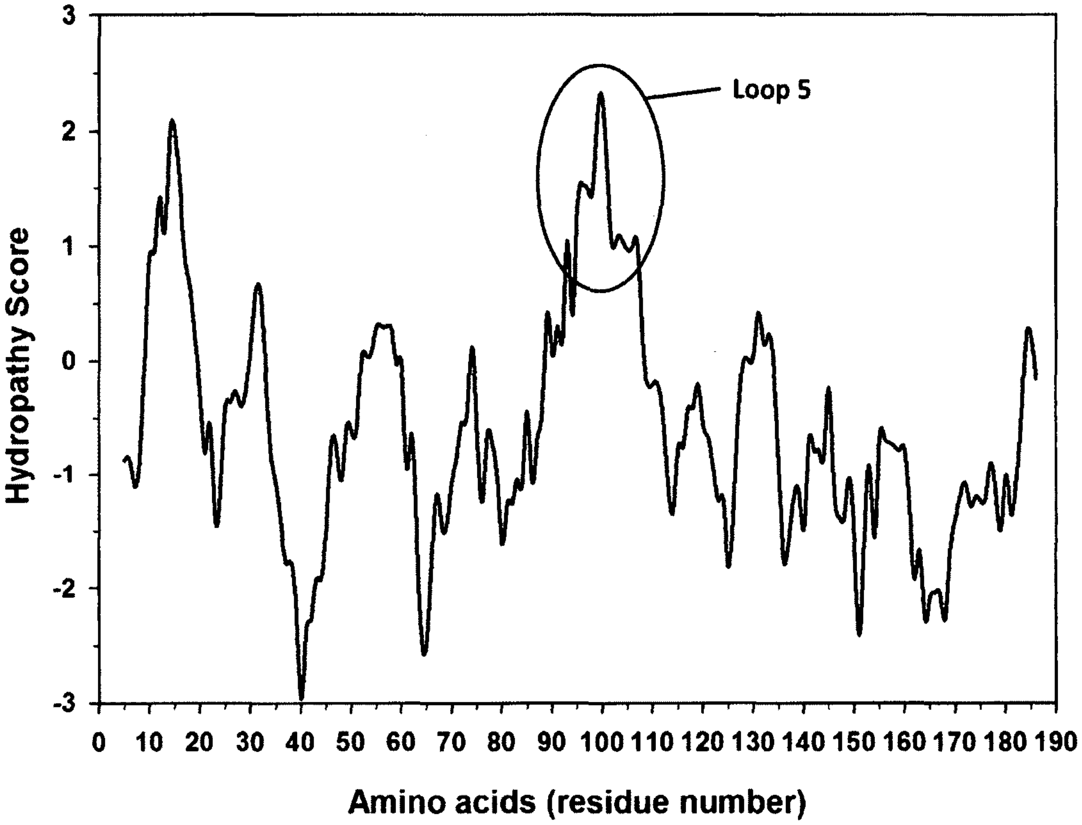


Figure 41. Hydrophobicity Scale

A hydrophobicity plot of the amino acid residues of 9TAT based on the Kyte and Doolittle scale (Wilkins et al., 1999) (http://web.expasy.org/protscale/). The sequence calculation was conducted by averaging over 9 residues.

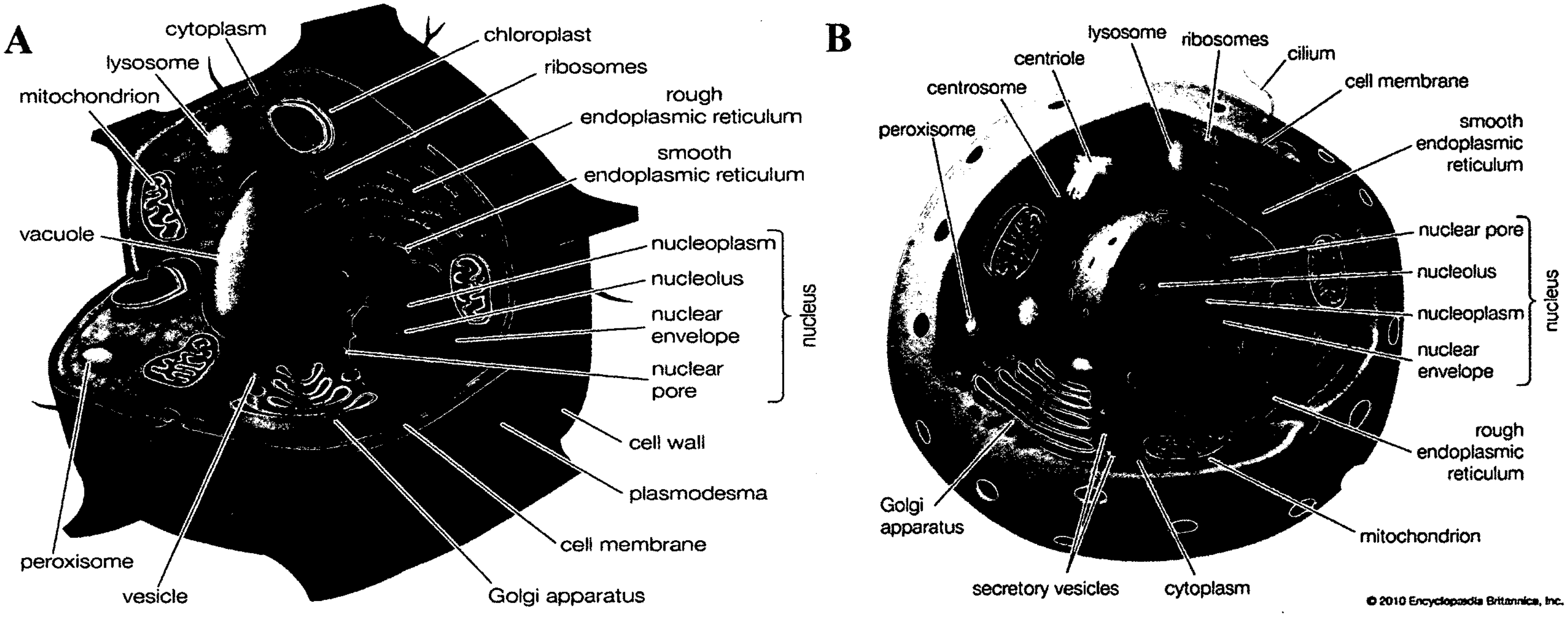


Figure 42. Illustration of Plant Cell and Animal Cell

Differences between plant and animal cell. A. Figure of a plant cell. B. Figure of an animal cell. Figures are adapted from Encyclopaedia Britannica (Britannica, 2015).

In the early stages of the experiment, 10-16 hrs after transformation, GFP is observed within the nucleus as well as the whole cell as expected. However, after 24 hrs, Tatil can be clearly seen localized on the plasma membrane (Figure 44 C). The nuclear pore is estimated to be between 40-60 kDa (Gorlich and Mattaj, 1996), whereas the size of the GFP is 28 kDa (Von Arnim et al., 1998) and Tatil is 22 kDa (Charron et al., 2005). The combined size of GFP and Tatil is about 50 kDa, which suggests that the protein can move easily in and out of the nucleus. Thus, the localization of Tatil on the plasma membrane may be the result of the association of the amino acids from the protein with membrane. As GFP is attached at the N-terminus of Tatil, wherever green fluorescence is observed that is where Tatil is localized.

To compare the results with the variants, bombardment was also done with two controls, WT Attil and pAVA321 which expresses just the GFP. As Attil was shown previously to localize on the plasma membrane, it serves as a good control to compare to the results with Tatil (Charron et al., 2005). Onion is an ideal plant for the study of the localization of Tatil using gene gun bombardment as the onion epidermis can easily be peeled off and the cell components can be seen under the microscope. The localization experiment with just pAVA shows that GFP is found throughout the cell (Figure 44 A). On the other hand, the localization of WT Tatil is found on only the plasma and nucleus membrane (Figure 44 C). Amino acids Phe96, Pro98 and Ile99 were individually mutated to glycine

When the onion cell was bombarded with Phe96Gly:GFP, it showed that the protein is not fully localized to the plasma membrane (Figure 45 C). Even though fluorescence can be seen on the plasma membrane, GFP can also be observed in the nucleus and throughout the cell. The intensity of the fluorescence in the nucleus seen in Phe96Gly:GFP is less when compared to the localization of GFP alone. The data implies that the mutation at Phe96 did affect how the protein interacts with the plasma membrane but did not completely remove its ability to associate with the membrane.

The localization of Pro98Gly:GFP was also different when compared to the WT Tatil and different from Phe96Gly (Figure 46 C). GFP was observed on the plasma membrane as well as in the nucleus. However, the intensity of the fluorescence seen for Phe96Gly indicates that this mutation had a larger effect on disrupting membrane association. When Ile99Gly:GFP was bombarded onto an

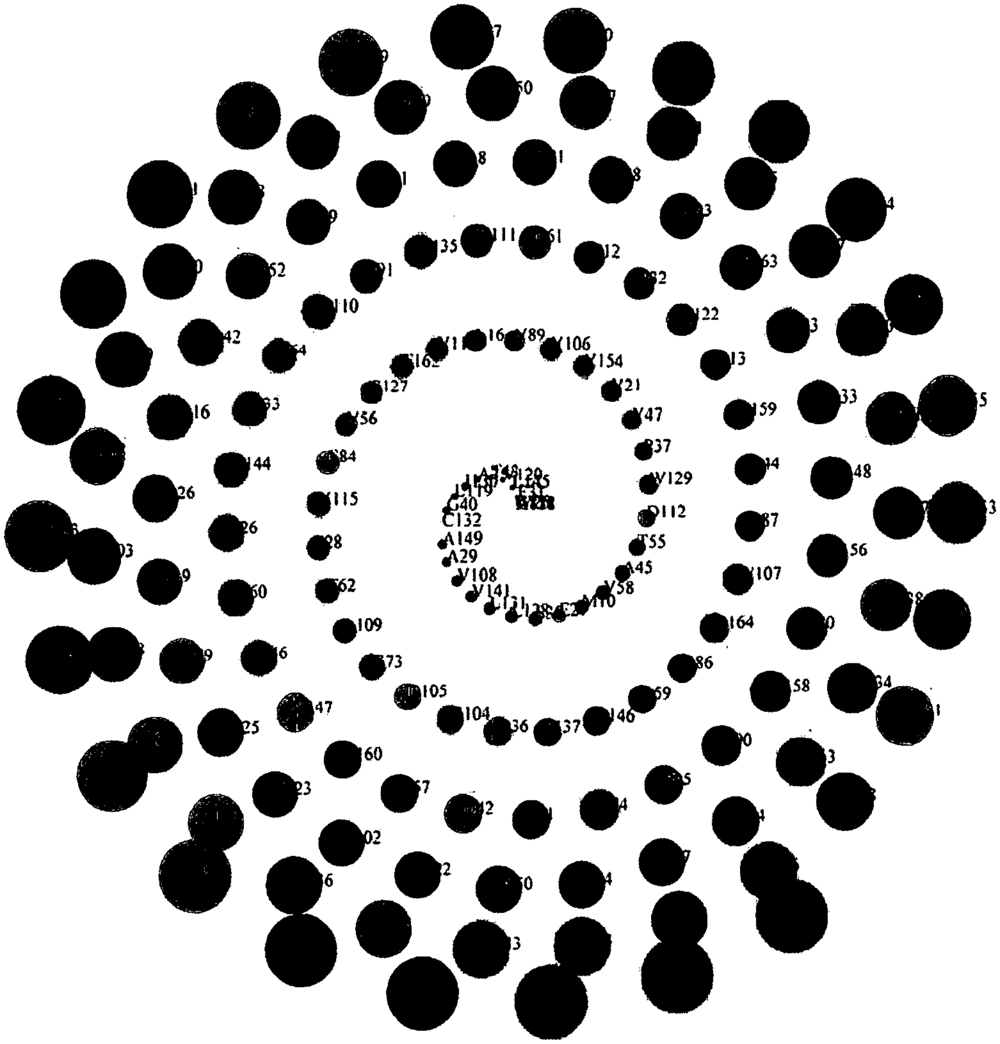


Figure 43. Accessible Surface Area

The solvent accessible area of the 9TAT protein was calculated using a model constructed with the SWISS-MODEL. The picture shows the most solvent accessible in the outer circle with the less solvent accessible residues found in the inner circle. The color scheme is as follows: grey shows hydrophobic residues, red shows negatively charged residues, blue shows positively charged residues and green shows polar neutral residues.

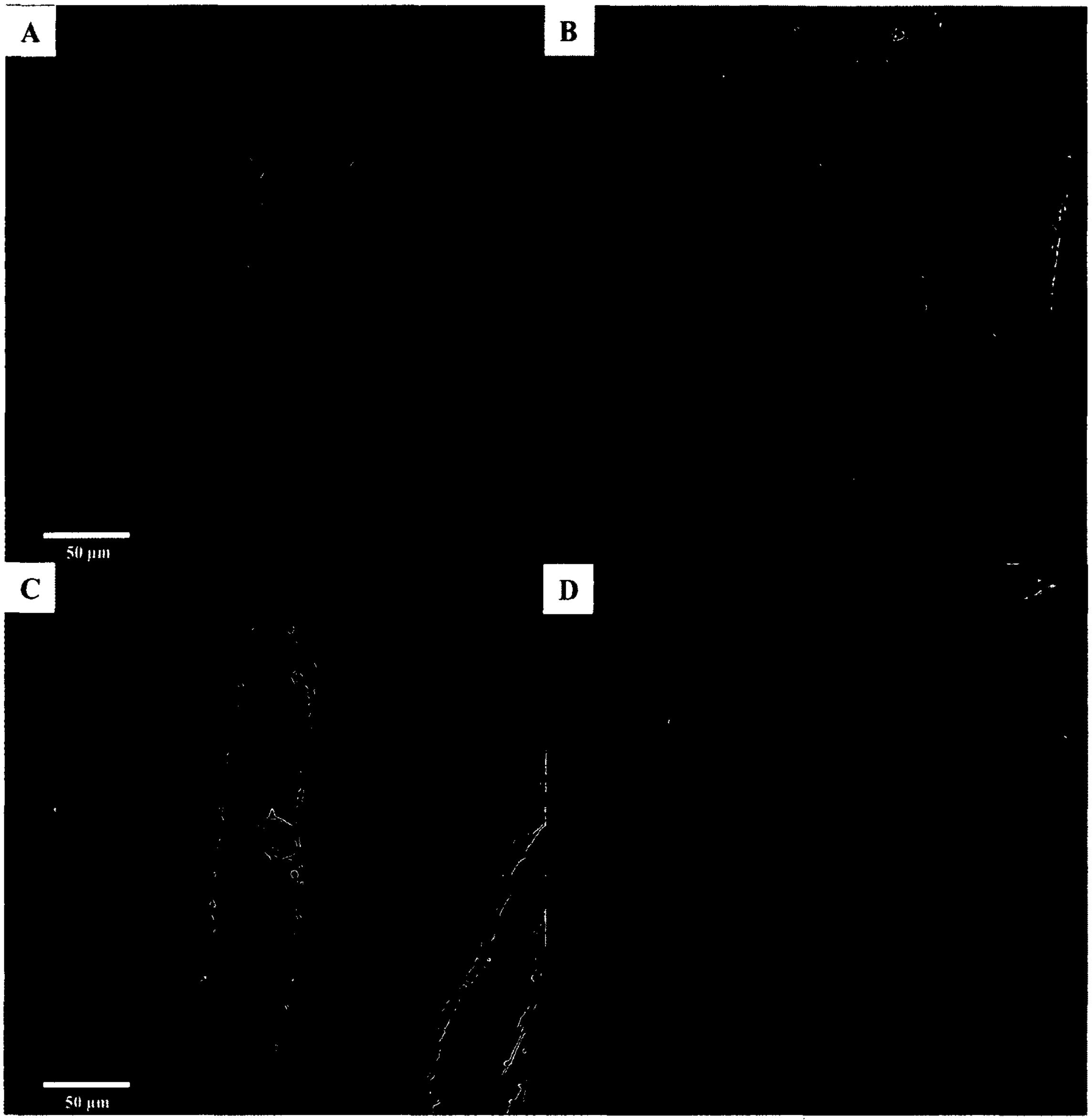


Figure 44. Cellular Localization of pAVA and Tatil
A. Transient expression of GFP. B. Bright field picture of the cell for GFP. C. Transient expression of WT Tatil:GFP. D. Bright field picture of the cell for WT Tatil:GFP. Scale bars are shown in the lower left corner of each panel.

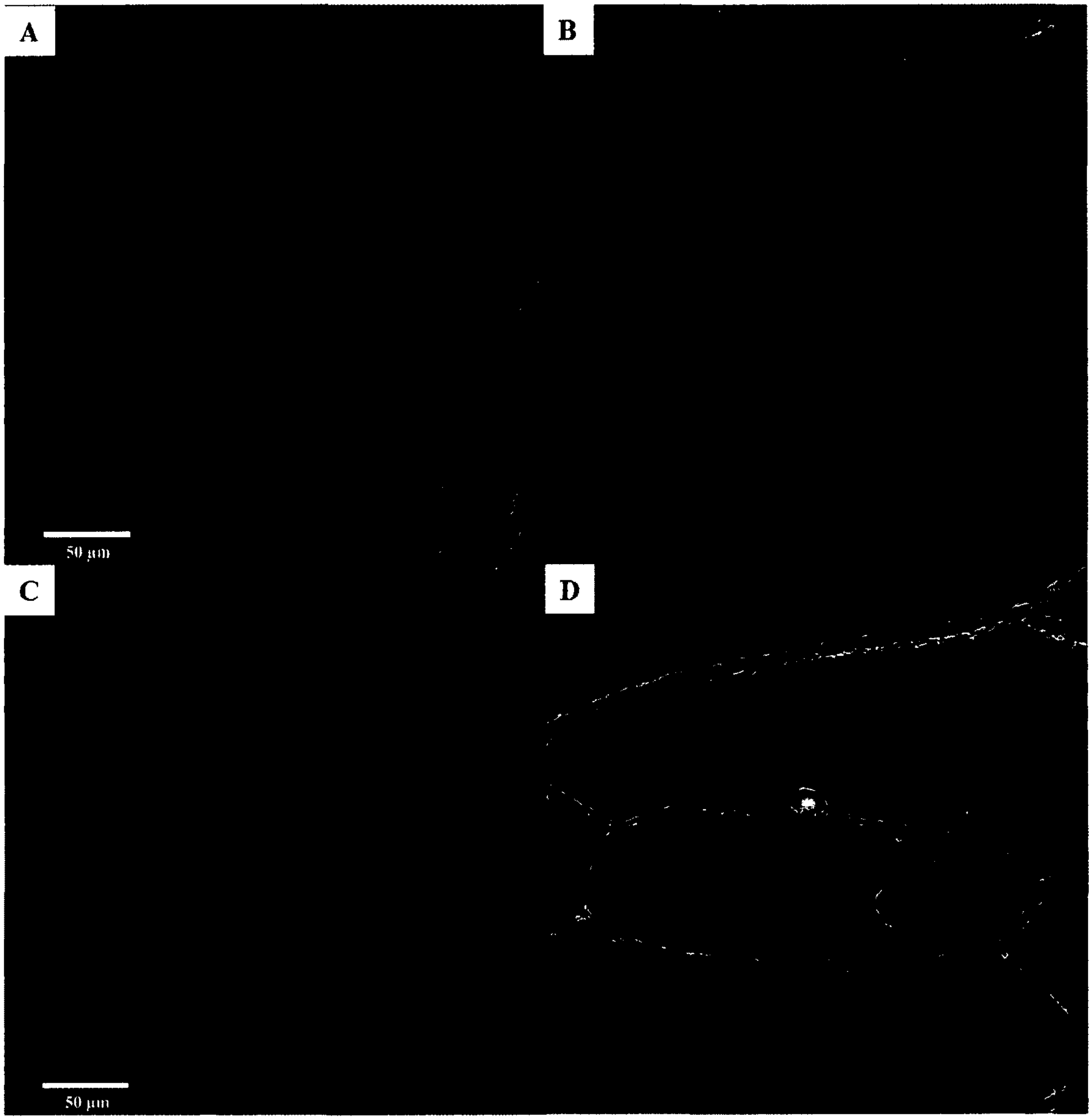


Figure 45. Cellular Localization of pAVA and Phe96Gly

A.Transient expression of WT Tatil:GFP. B. Bright field picture of the cell for WT Tatil:GFP. C. Transient expression of Phe96Gly:GFP. D. Bright field picture of the cell for Phe96Gly:GFP. Scale bars are shown in the lower left corner of each panel.

onion cell, fluorescence showed that it is found throughout the cell (Figure 47C). The localization of the Ile99Gly:GFP is the same as GFP alone, which indicates that Ile99Gly is not attached to the plasma membrane. When WT Tatil is compared to the three mutations (Figure 48), we can see that the intensity of GFP within the nucleus increases as we progress from the WT to Ile99Gly. This suggests that the loop is inserted in the plasma membrane not in a vertical manner but sideways where amino acids Phe96 and Pro98 are partially attached to the membrane. However, amino acid Ile99 may be fully associating with the plasma membrane.

4.4.2 STRUCTURAL CHARACTERIZATION

The first step in the structural characterization of a protein is to express and purify the protein. The expression and purification of Tatil involved several steps to find the right conditions for the protein. Tatil was first cloned into the pTrcHis vector which has a N-terminal polyhistidine (6xHis) tag. The vector contains an enterokinase recognition sequence which allows the enzyme to cut off the tag. A His-tag was chosen as it is allows the purification of the protein with an immobilized metal affinity column. However, after the expression of Tatil, our collaborators at McGill University were unable to cut off the tag and recover a significant amount of protein to conduct our experiments. Also, as the cost of enterokinase is very expensive, it was not feasible to continue to use the pTrcHis vector. Thus, Tatil was cloned into pET14b. This vector also contains an N-terminal His-tag but the enzyme required to cut off the tag is the less-expensive thrombin rather than enterokinase.

E. coli BL21(DE3) was first used to express Tatil. The expression level of Tatil in this cell was not substantial and it took 22 L of Luria broth to get 50 to 60 mg of protein. Moreover, a time trial in BL21(DE3) showed that the protein was not expressed in high levels. The problem with low expression could be a result of BL21(DE3) not supplying the rare codons which are required for this eukaryotic protein. To supplement the codons that the protein needs, Rosetta(DE3) pLysS was used. The time trial showed that Tatil was expressed in higher levels and has a difference in expression levels before and after induction with IPTG (Figure 49).

After the time-trial was done, we did three batches of large expressions. Each batch of 6 L media gave us about 16 mg of protein. After purification, the protein

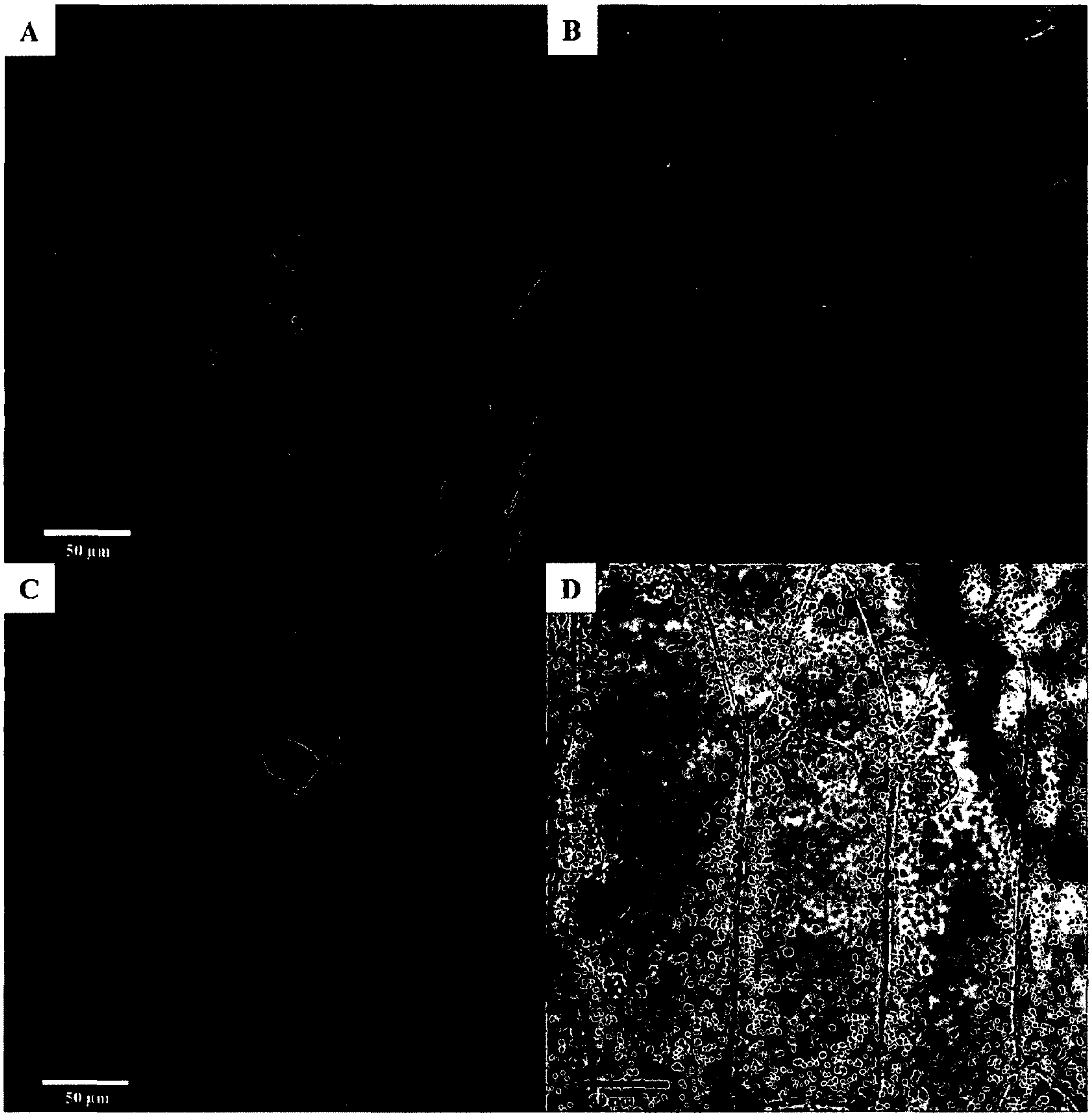


Figure 46. Cellular Localization of Tatil and Pro98Gly

A.Transient expression of WT Tatil:GFP. B. Bright field picture of the cell for WT Tatil:GFP. C. Transient expression of Pro98Gly:GFP. D. Bright field picture of the cell for Pro98Gly:GFP. Scale bars are shown in the lower left corner of each panel.

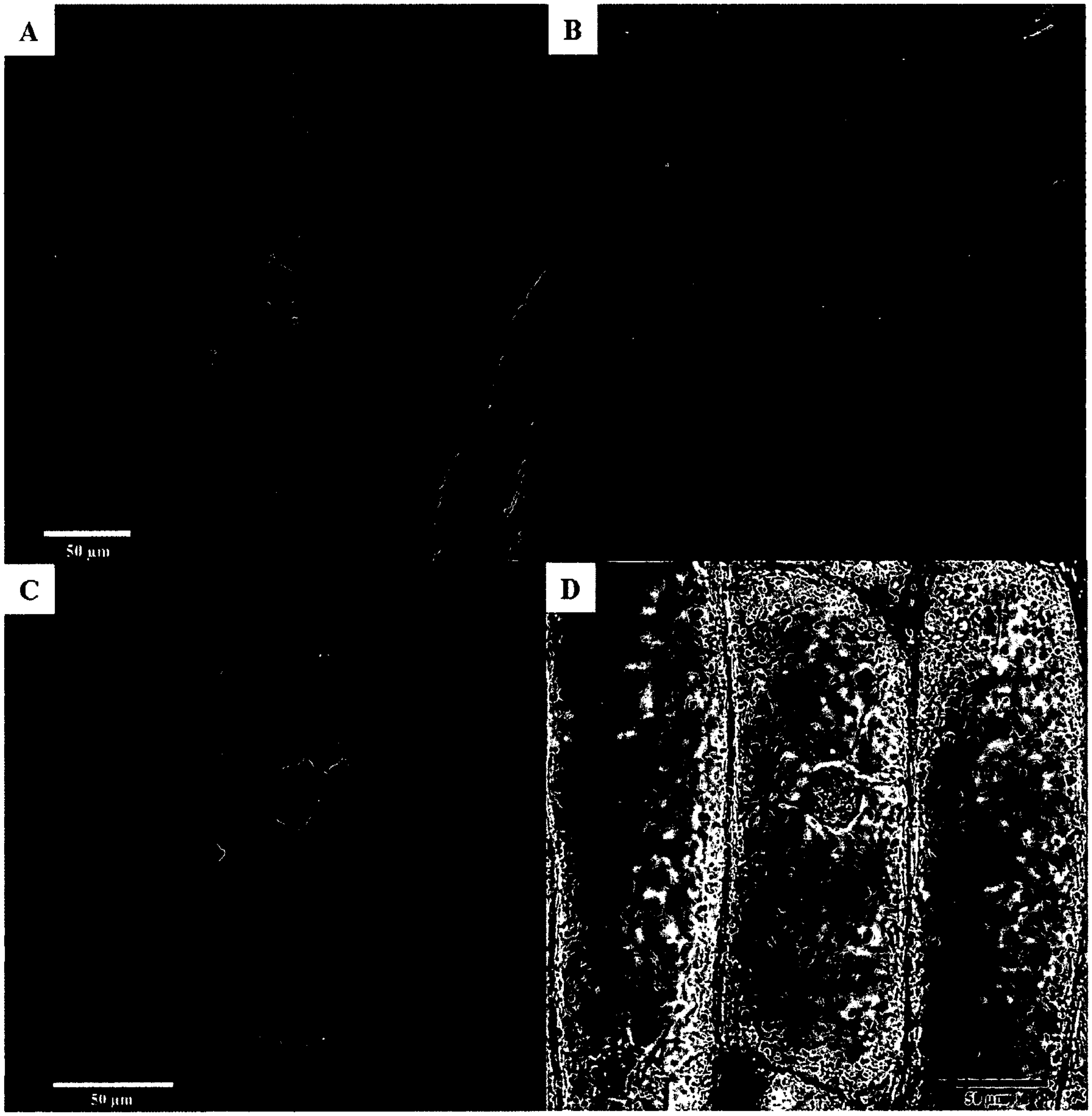


Figure 47. Cellular Localization of Tatil and Ile99Gly
A.Transient expression of WT Tatil:GFP. B. Bright field picture of the cell for WT Tatil:GFP. C. Transient expression of Ile99Gly:GFP. D. Bright field picture of the cell for Ile99Gly:GFP. Scale bars are shown at the bottom of each panel.

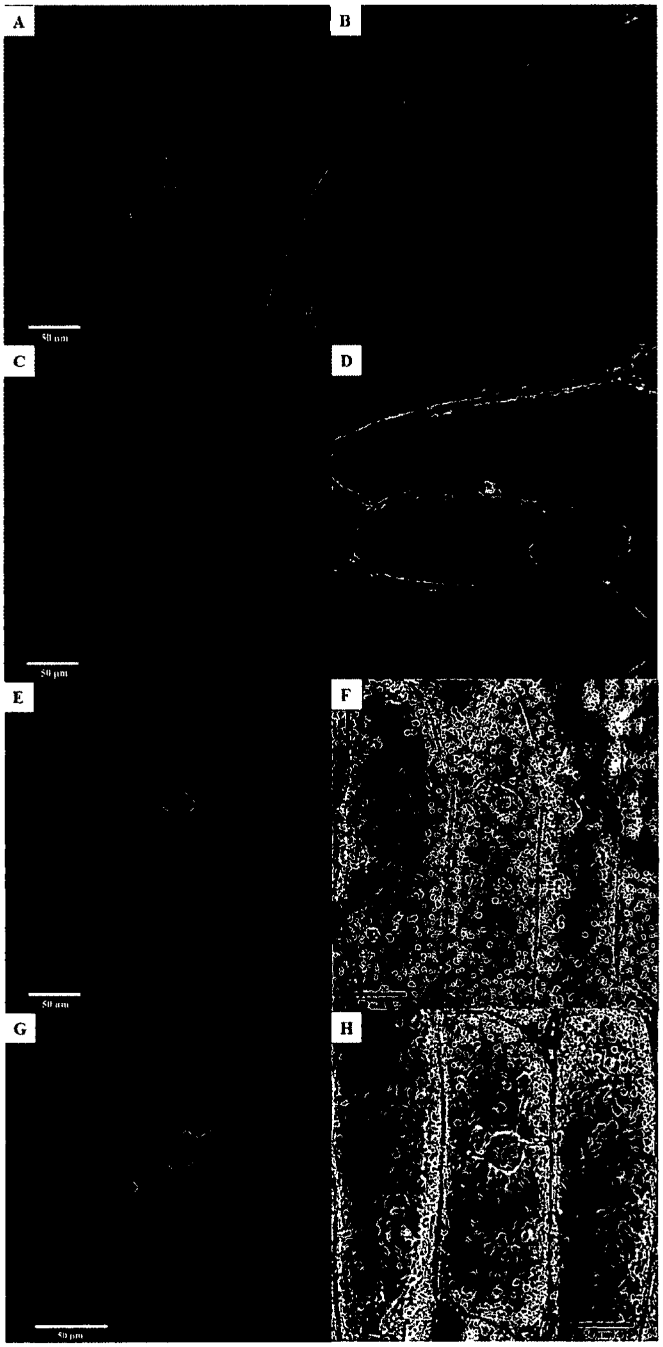


Figure 48. Cellular Localization of Various Proteins

A.Transient expression of WT Tatil:GFP. B. Bright field picture of the cell for WT Tatil:GFP. C. Transient expression of Phe96Gly:GFP. D. Bright field picture of the cell for Phe96Gly:GFP . E. Transient expression of Pro98Gly:GFP. F. Bright field picture of the cell for Pro98Gly:GFP. G. Transient expression of Ile99Gly:GFP. H. Bright field picture of the cell for Ile99Gly:GFP. Scale bars are shown at the bottom of each panel.

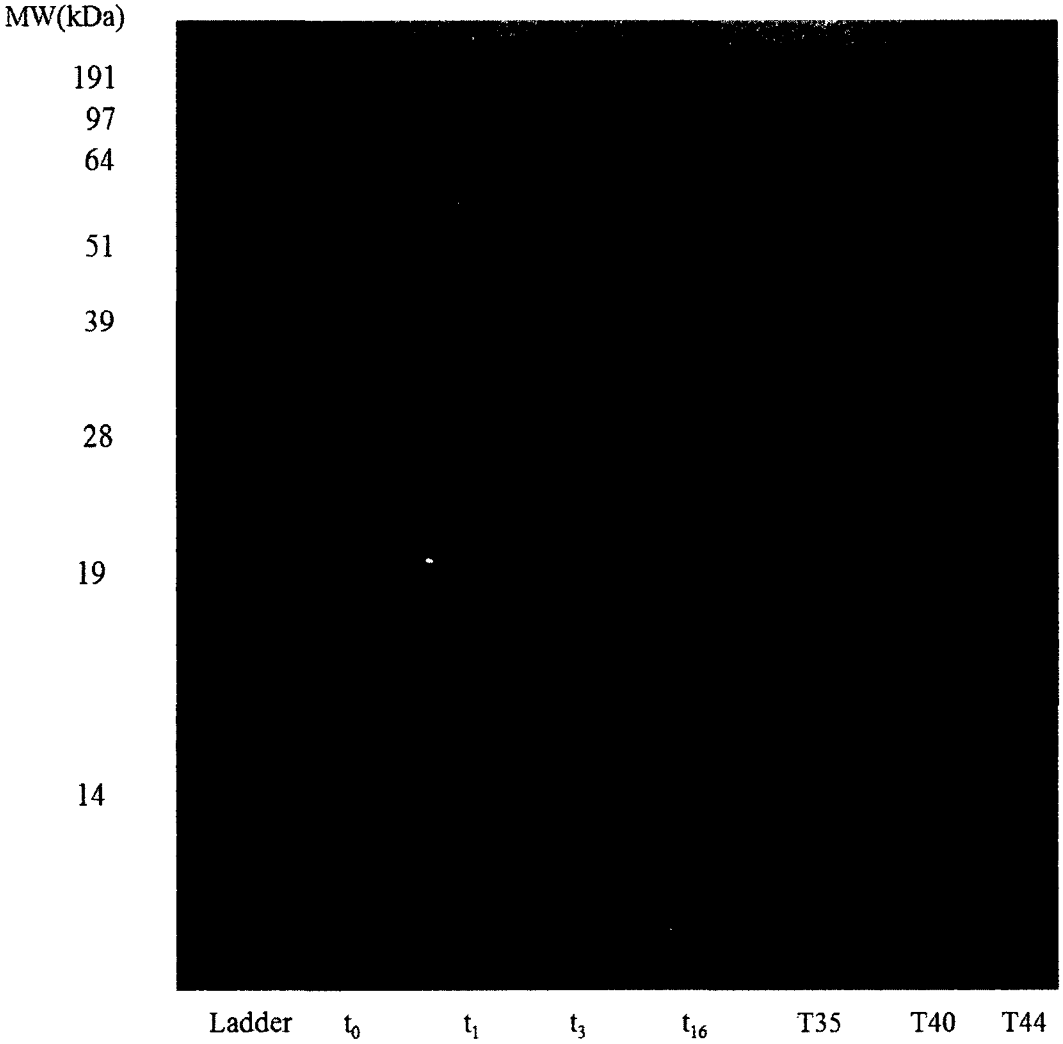


Figure 49. Gel Picture of Tatil Time-Trial and Protein Purification

The SDS-PAGE gel picture shows the time trial of Tatil in the Rosetta (DE3) pLysS cell line. t0 shows Tatil expression before induction with IPTG; t1, expression after 1 hour of induction, t3, 3 hrs after induction and t16, 16 hrs after induction. T35, T40 and T44 show the Tatil expression levels after purification of the protein using a Nickel affinity chromatography column.

was dialyzed in water and lyophilized. The dialysis water was changed twice a day for 3 days. In the last day, the protein started to aggregate. After lyopholization, we tried to dissolve the protein in a buffer that is suitable for the use of thrombin (20mM Tris-HCl, 150 mM NaCl, pH 8.0). However, the protein would not completely dissolve. We took the protein and added 6M Guanidine-HCl to denature the protein. The solution was filtered using a 0.45 micron filter to remove any aggregates. The protein was added to a dialysis tube and placed in a buffer (20mM Tris-HCl, 150 mM NaCl, pH 8.0) to refold the protein. The protein did not refold properly and aggregated. We took another batch of protein and tried to refold the protein by adding it drop wise to a buffer (20mM Tris-HCl, 150 mM NaCl, pH 8.0 with 10% glycerol). One batch was done at room temperature whereas the other one was performed at 4°C. The protein did not aggregate while it was being added into the buffer. However, after 16 hrs the protein aggregated. The result showed that Tatil is not stable in water.

Another batch of 9 L of Tatil was expressed and purified using a Nickel affinity chromatography column. A concentrator was used to bring the volume down from 170 ml to 3.2 ml. The imidazole present in the purification buffer was exchanged with 50 mM Tris-HCl, 300 mM NaCl, pH 8.0. After the buffer exchange, no aggregation was observed. The result showed that Tatil cannot be placed in unbuffered solutions such as H_2O without causing aggregations. Ideal purification methods cannot be predicted, only discovered experimentally. For example the human serum RBP, which is related to Tatil, can be placed in H_2O without any problems. The yield of protein was still low, 10.8 mg from 9 L of media. After Tatil was concentrated down, thrombin was added and incubated for 24 hrs at 4°C to cut the histag off. Tatil was then run on a G-75 gel filtration chromatography column for further purification and also to gain an idea of the state of the protein, such as whether it is a monomer or a dimer. The filtration showed two peaks with the first one coming out around 66 kDa and the second one around 22 kDa (Figure 50). The size of the proteins were established by calibrating the G-75 column using myoglobin (17.6 kDa), deoxyribonuclease (31 kDa) and albumin (66 kDa). Absorbance of the protein was taken at 280 nm and 330 nm (Figure 50). The absorbance at 280 nm detects the amount of protein that is present, while the 330 nm absorbance indicates if there is any aggregation. The results show that there is no aggregation as the absorbance at

330 nm is very low. As the molecular weight of Tatil is 21.77 kDa, the gel filtration results suggest that the protein can exist as a trimer as well as a monomer.

An SDS-PAGE gel was run to look at the purity and size of the two peaks. The denatured gel showed that the two peaks are on the same lane below the 28 kDa mark (Figure 51). However, the first peak is not as pure as the second peak. The samples were heated to two different temperatures 80°C and 100°C, to discern if the other lanes in peak one were a result of heating the sample before loading it into the gel. Nonetheless, the samples from both temperatures look identical, which shows that heat is not the cause of the fragments seen on the gel. With further confirmation of size, it is possible that Tatil would be the first lipocalin found to be a trimer.

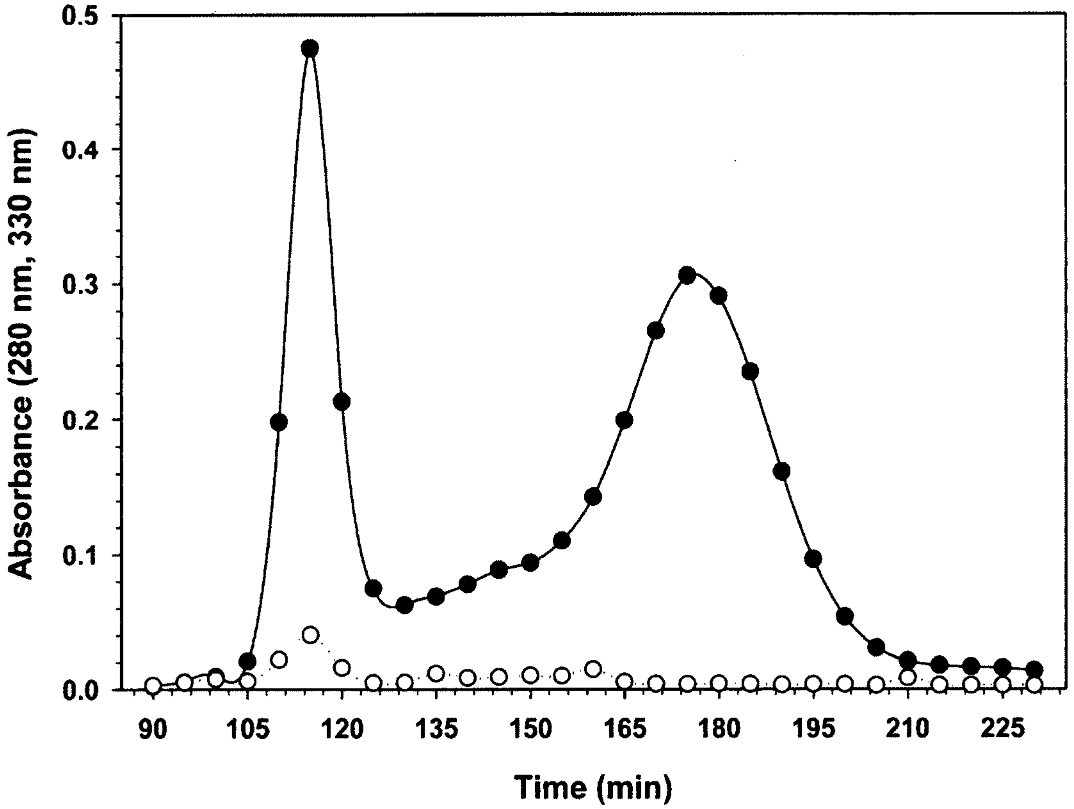


Figure 50. Size Exclusion Chromatography
Peaks obtained when Tatil was run through a G-75 column. The filled circles show the absorbance at 280 nm while the unfilled circles display the 330 nm absorbance.

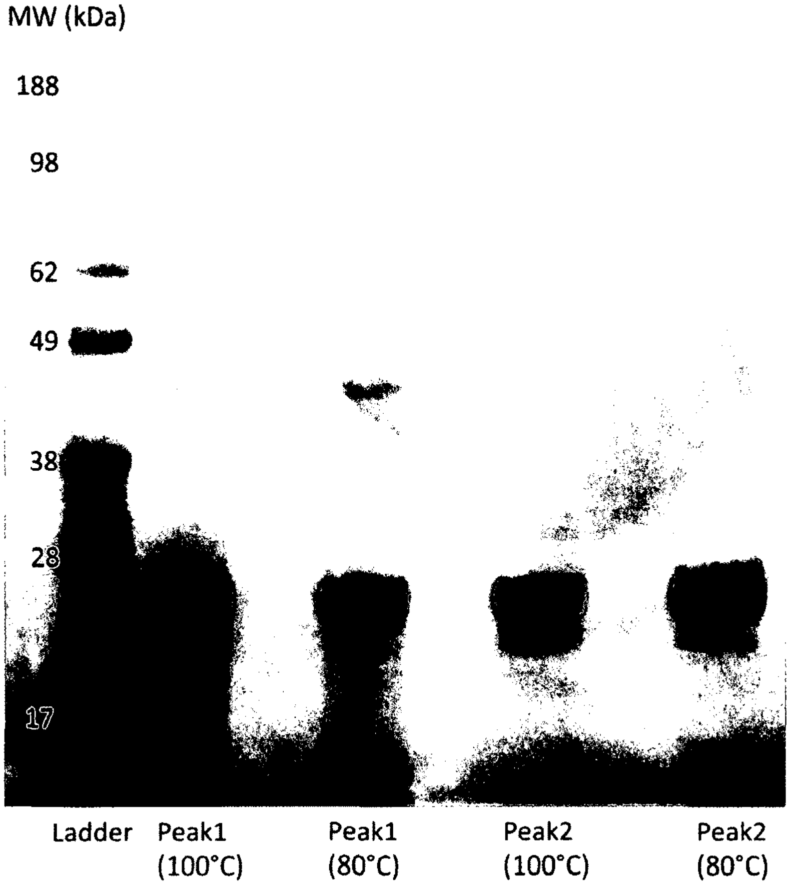


Figure 51. SDS-PAGE of Tatil from Size Exclusion Chromatography
The Image shows the two peaks of Tatil purified using gel filtration chromatography after being run on a denaturing gel. The samples were heated at two different temperatures (80°C and 100°C). The dark circle in the middle of the image is a result of the camera that was used to take a picture of the gel.

CD and fluoresence studies were done to study the structure of Tatil. The far-UV CD shows the secondary structure of Tatil (Figure 52 A). The spectra reveals that there is a difference between peak one and two. While peak one indicates that the protein has α-helices based on the increase in molar ellipticity seen around 208 nm and 220 nm, peak two has the typical spectra of a β-sheet. The difference between the two peaks could be attributed to the C-terminal structure of Tatil. The structure models were not able to agree on the protein topology after the eighth β-strand. Some of the models, like the SWISS-MODEL, had a short α-helix in the structure. The other models, such as Robetta, gave a long α-helix. Formation of the trimer may occur through the interaction of α-helix at the C-terminus, and that could give an α-helical signature in the far-UV CD, which is seen in peak one. The near-UV CD signature of Tatil in peak one and peak two is similar (Figure 52 B). The near-UV CD shows a signature for the tertiary structure of Tatil. Peak one has a Phe signal around 260 nm and a Trp signal in the 290 nm region. Peak two also has the Trp signal but lacks the Phe signal seen in peak one.

Thermal unfolding monitored by CD of peak one and peak two was conducted to look at the stability of the protein. The thermal unfolding was monitored from 20°C to 90°C. At 95°C peak one aggregated while peak two aggregated at 90°C. Peak one lost the minima that is seen at 208 nm and only maintained the one at 220 nm as the temperature increased (Figure 53 A). On the other hand, peak two lost the minimum at 208 nm with increasing temperature (Figure 53 B). The thermal unfolding of the two peaks is clearly different. In the first peak, as the temperature increased the signal goes from the negative (-1.2e+7) to the positive (-8.0e+6) (Figure 54 A). However, in the second peak, the change is not consistent as the 50°C and 60°C spectra are slightly lower than the 20°C spectrum but the spectra at temperatures higher than 70°C are above the 20°C spectra. At 50°C and 90°C, both peak one and peak two's secondary structure change however, the proteins do not completely unfold (Figure 55). Examining peak one and peak two, the data indicates that the trimer is more stable at higher temperatures compared to the monomer as the trimer still maintains its β-sheet signature.

The thermal unfolding fluorescence spectra are different when peak one and two are compared (Figure 56). In peak one the spectra shows a straight line going down from 20°C to 100°C, while peak two raises in intensity around 50°C and then drops down. That may be suggestive of an intermediate. Also, most unusually there is no baseline. What we expected to see was two sigmoidal curves for peak one and one sigmoidal curve for peak two. For a trimer, the first sigmoidal curve would show the dissociation of the protein and the second one shows the unfolding of the protein. A monomer would show one sigmoidal curve as the protein unfolds. However, the sigmoidal curves were not observed for peak one and two of Tatil. Moreover, thermal unfolding spectra show that the peaks do not unfold at 100°C (Figure 57) which further supports the data that

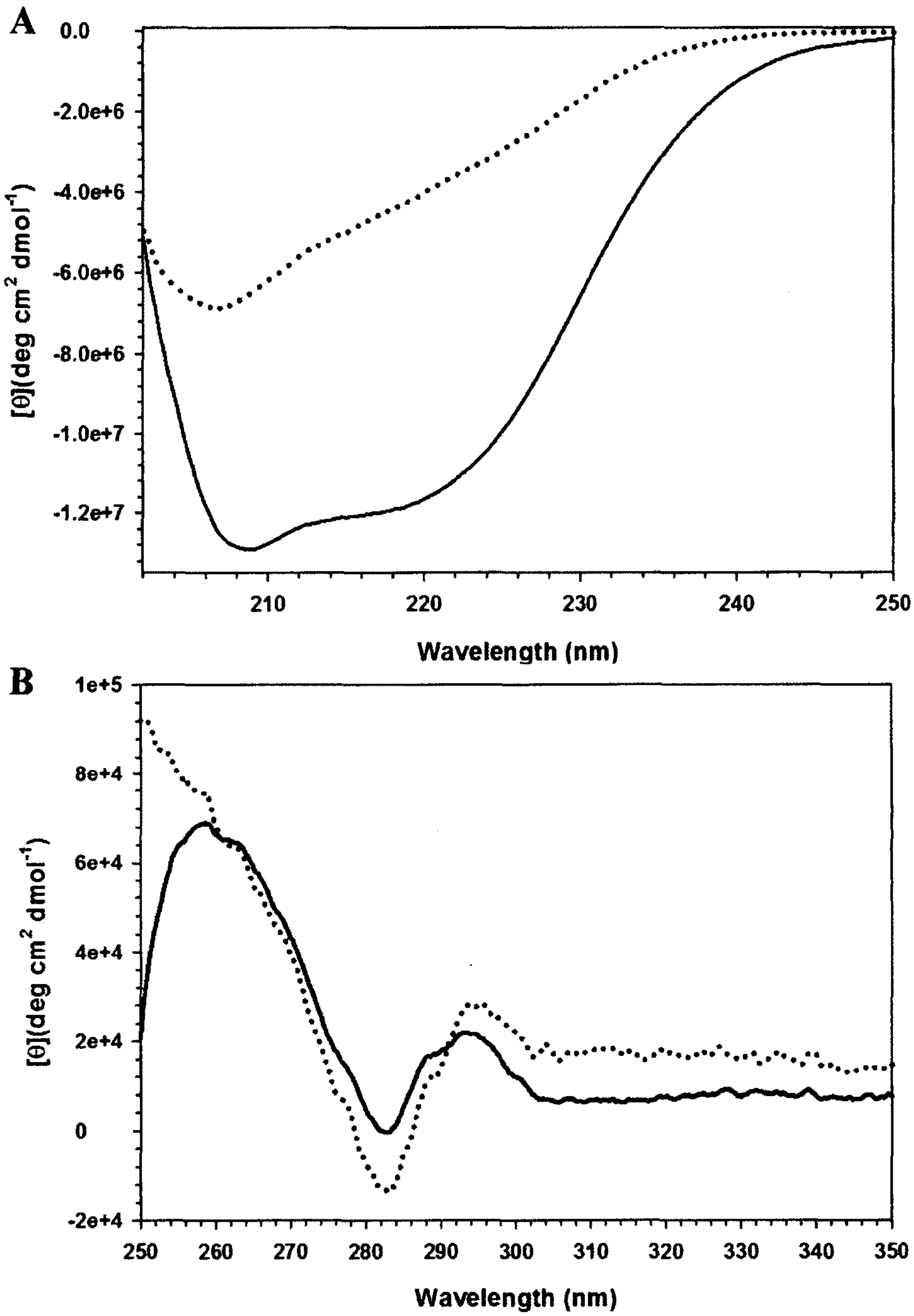


Figure 52. Circular Dichroism Spectra of Tatil

Spectra of the protein in peak one and peak two from the gel-filtration purification step. A. Far-UV CD of Tatil. B. Near-UV CD of Tatil. The solid line shows peak one while the long dash displays peak two. The cuvette pathlength was 0.1 cm and 1 cm for far- and near-UV CD, respectively. Each spectrum was an average of 28 scans for far-UV CD and 15 scans for near-UV CD.

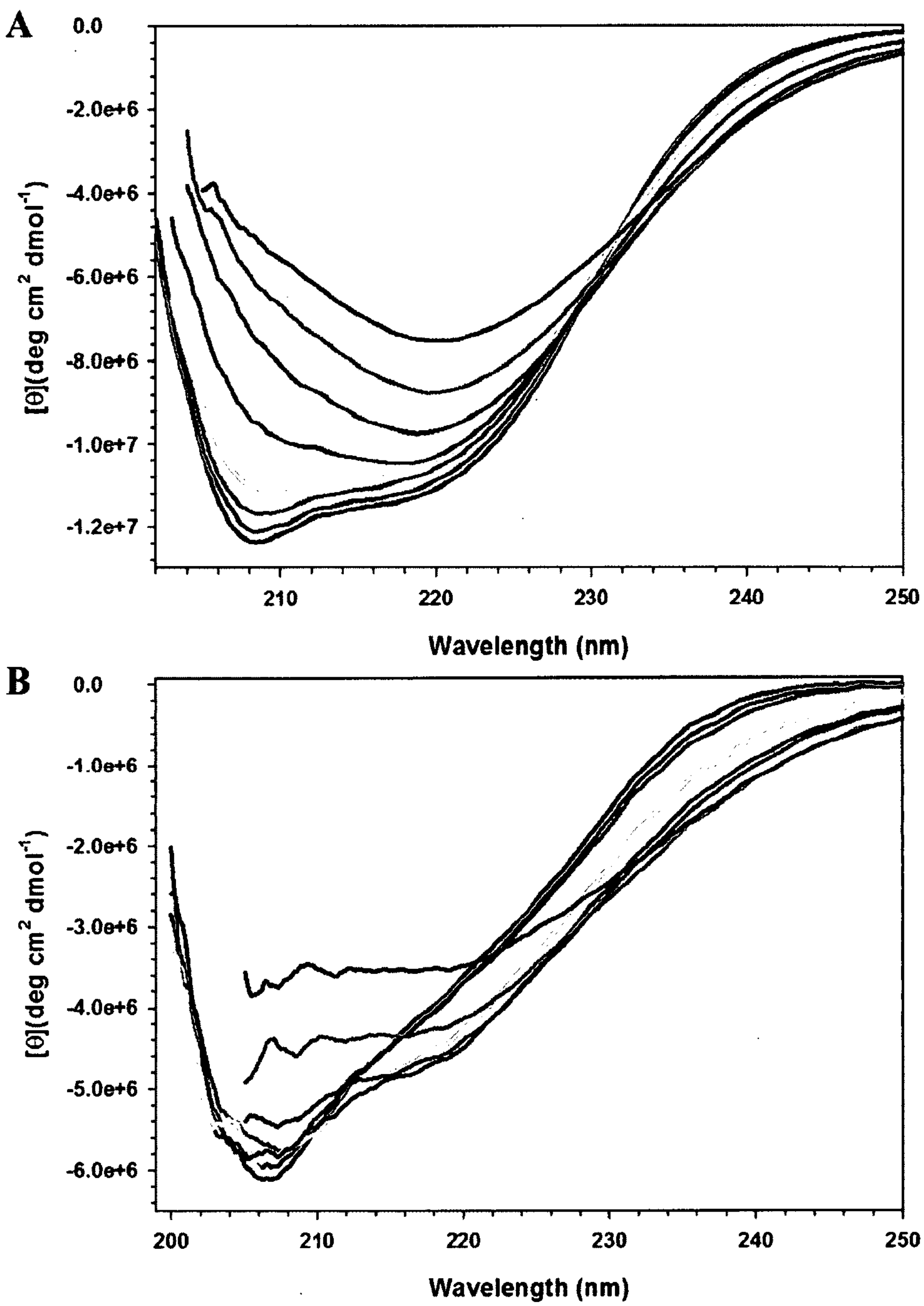


Figure 53. Thermal Unfolding Circular Dichroism Spectra of Tatil

Spectra showing the thermal unfolding of Tatil from 20°C to 90°C as monitored by far-UV CD. A. A graph illustrating the thermal unfolding of peak one. B. The graph shows the thermal unfolding of peak two. The color coding scheme in the graph is assigned as follows: Black - 20°C, Red - 30°C, Green - 40°C, Yellow - 50°C, Blue - 60°C, Pink - 70°C, Cyan - 80°C, Dark red - 90°C. The cuvette pathlength was 0.1 cm.

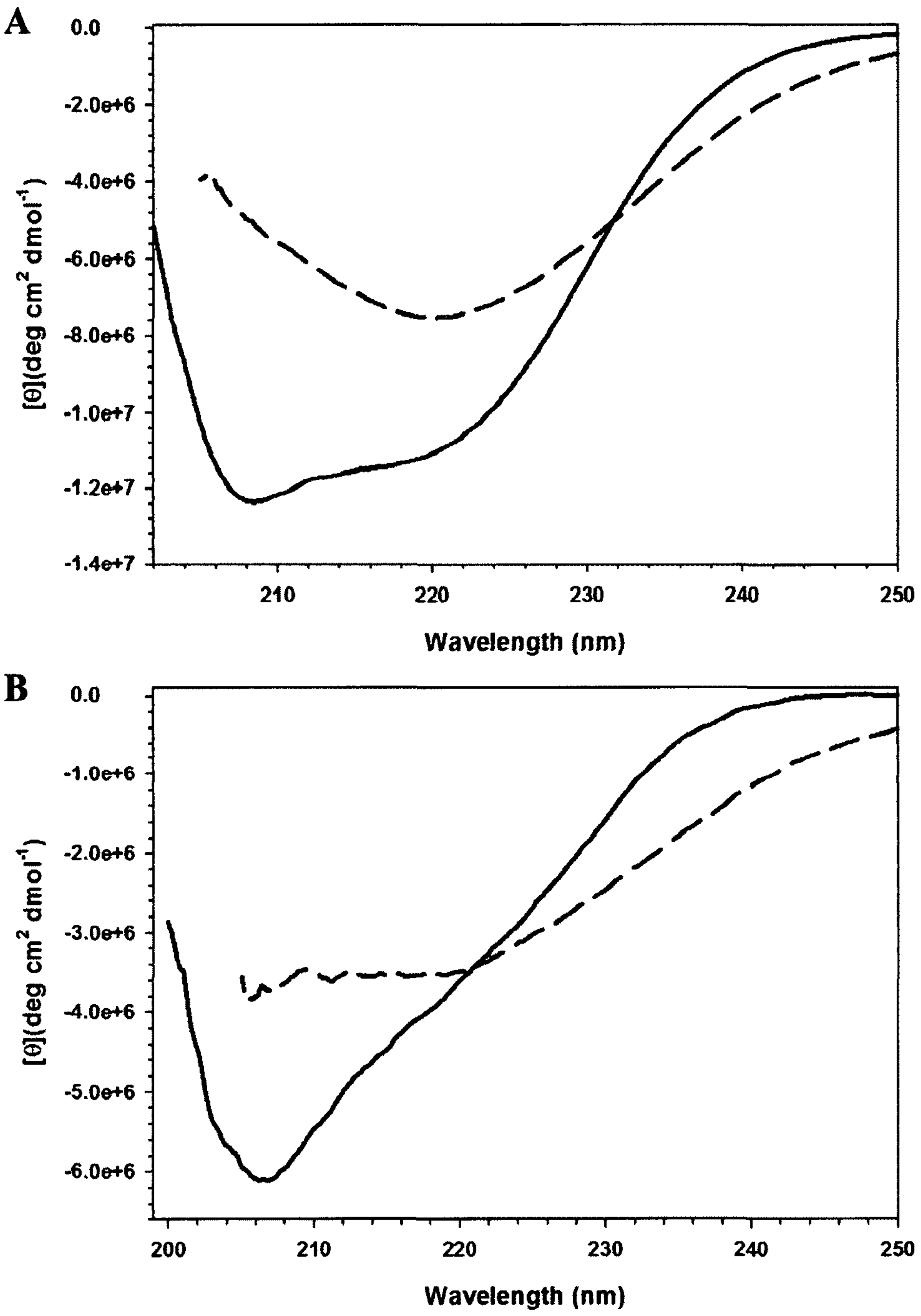


Figure 54. Thermal Unfolding Circular Dichroism Spectra of Tatil at Various Temperatures

Spectra showing the thermal unfolding of Tatil at 20°C and 90°C as monitored by far-UV CD. A. A graph illustrating the thermal unfolding of peak one. B. The graph shows the thermal unfolding of peak two. The solid line shows 20°C while the long dash displays 90°C. The cuvette pathlength was 0.1 cm.

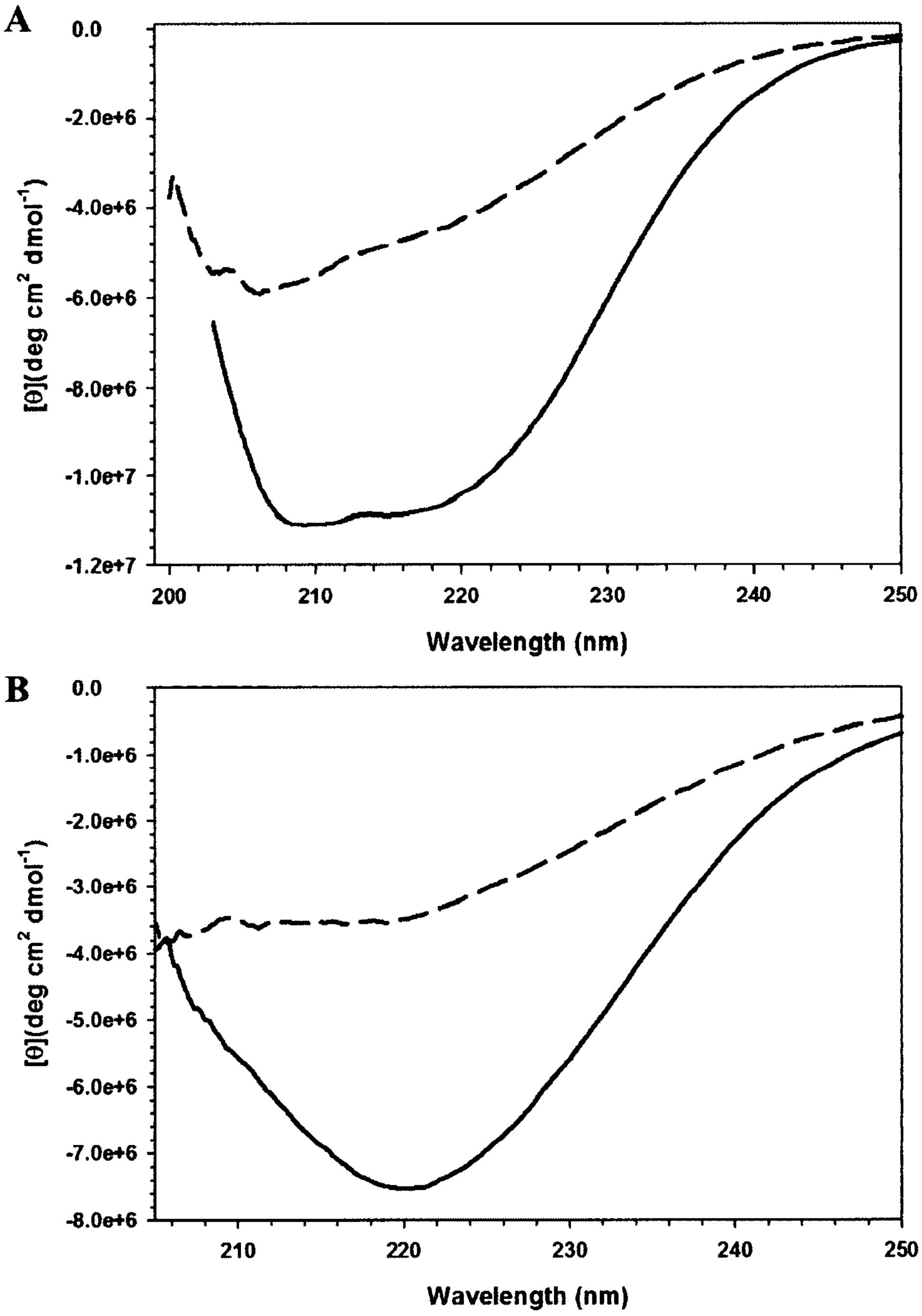


Figure 55. Comparison of the Thermal Unfolding Circular Dichroism Spectra of Tatil at Various Temperatures

Spectra showing the thermal unfolding of Tatil at 50°C and 90°C as monitored by far-UV CD. A. A graph illustrating the thermal unfolding of peak one and peak two at 50°C. B. The graph shows the thermal unfolding of peak one and peak two at 90°C. The solid line shows peak one while the long dash displays peak two. The cuvette pathlength was 0.1 cm.

Tatil enables the plant to withstand high temperatures. At 100°C, 0.05 mg/ml of the protein didn't aggregate when inspected visually, which is opposite of what happened at a higher concentration of 0.22 mg/ml for peak one and 0.14 mg/ml for peak two. This suggests that Tatil aggregates at higher concentrations *in vitro*.

4.5 CONCLUSION

Global warming has become one of the main problems of the 21st century. As climates have started to shift and weather patterns have varied, the maintenance of crops in the changing environment is crucial. In a contribution to this study we have looked at wheat lipocalin. TIL proteins have been proposed to help plants withstand various temperatures as well as oxidative stress. The study of this important protein might help to sustain plants in locations that they would not normally grow. Moreover, the plants might resist outside pressures and grow to maturity to provide sustenance. Based on the bioinformatics study, the wheat lipocalin loop between β-strands E and F was believed to be the mode of interaction with the plasma membrane. Using various programs such as ASAview, three amino acids (Phe96, Pro98, Ile99) were chosen to be mutated for localization studies. The cDNA of the wild-type and variant wheat lipocalins were transfected onto onion cells using a gene gun and observed by confocal microscopy. The study shows that these amino acids might be responsible for the association of the protein with the membrane. Furthermore, we were also able to express and purify the wheat lipocalin. The result suggests that wheat lipocalin can be found in a trimer as well as in a monomeric state. Near and far-UV CD show that the two states are different and indicate that they have variations in structures. Thermal unfolding monitored by far-UV CD shows that the protein is stable which supports the data that Tatil aids plants to tolerate temperatures even at 45° C (Charron et al., 2005).

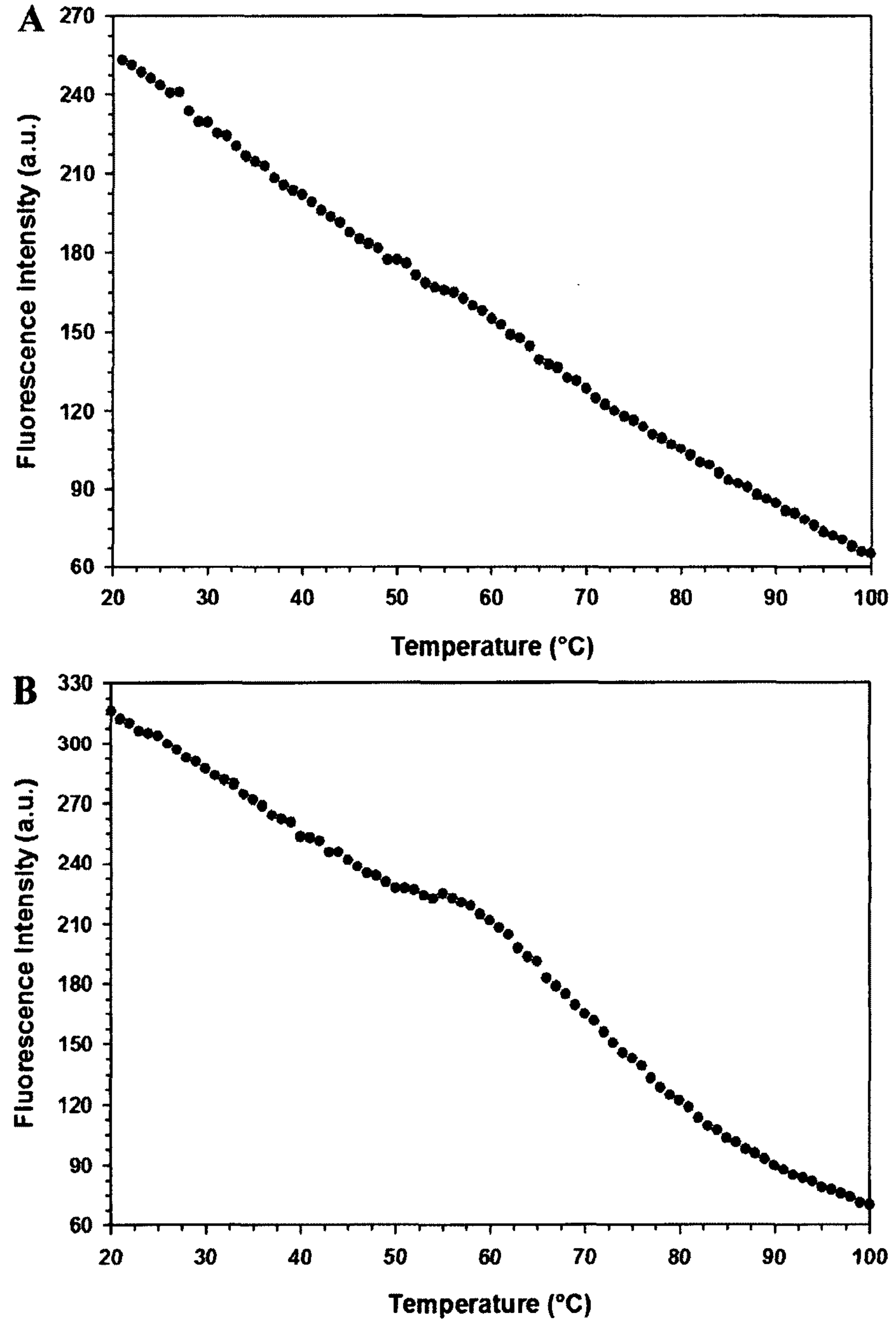


Figure 56. Thermal Unfolding of Tatil Monitored by Fluorescence Spectroscopy

Fluorescence spectra of Tatil. A. This graph shows the fluorescence spectra of peak one. B. The figure illustrates the spectra of peak two. The data represents an average of three studies.

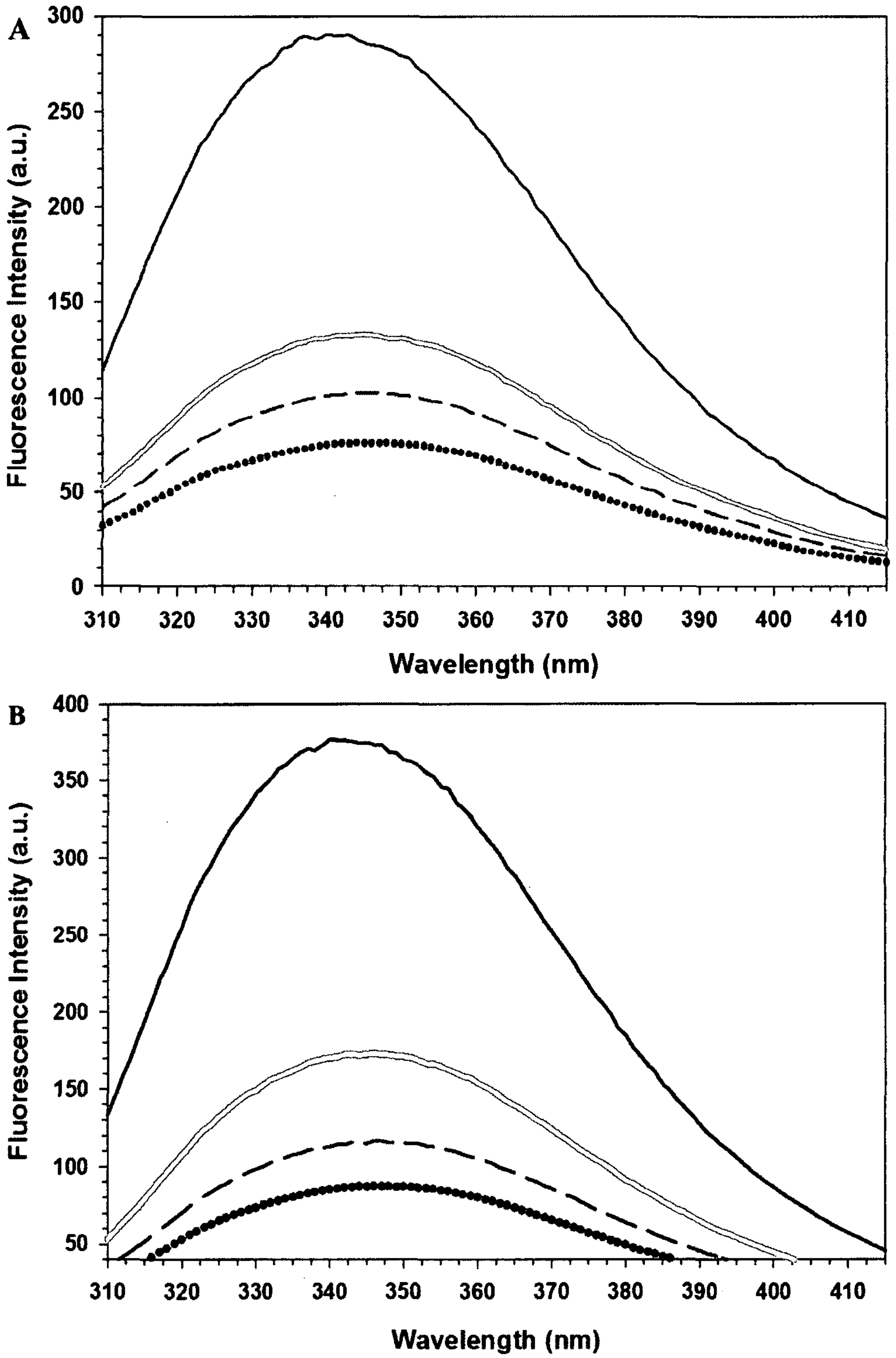


Figure 57. Fluorescence Thermal Unfolding of Tatil at Various Temperatures Fluorescence spectra of Tatil at different temperatures. A. Fluorescence spectra of peak one. B. Fluorescence spectra of peak two. The different temperatures in the figures were assigned as follows: solid black line: 20°C; solid gray line: 80°C; long dash line: 90°C; circle: 100°C.

CHAPTER 5

SUMMARY AND FUTURE WORK

This dissertation presents studies that were performed to investigate select questions that we face when it comes to global warming. The current change in climate affects all species from mankind to organisms at the microscopic level. Even though the change cannot be denied, the type of effect that global warming will have on the environment is still being debated. Thus, we explored two spectrums of the question. The first is how a change in the CO_2 conditions affect a microorganism at the genetic level. The increase in atmospheric CO_2 concentration has been dramatic since the industrial revolution. In just our lifetimes we have observed the climate changes. Consequently, how does that affect living things. The second question that we explored is how plants can adapt to the change. We looked at the functional and structural characteristics of a protein in wheat that is believed to contribute to its withstanding temperature change.

Chapter II studies an essential cyanobacteria, *T. erythraeum*, under current pCO_2 conditions of 386 ppm and the projected pCO_2 in the year 2100, 750 ppm. Cyanobacteria are ubiquitous organisms that may have been involved in the oxygenation of the earth's atmosphere. We wanted to observe the transcriptional changes that occur in an ancient organism that has been on the earth for more than 2 billion years. Moreover, this cyanobacteria makes a great contribution to the nitrogen and carbon cycles. These abilities make the organism an essential component in the study of the impact of high pCO_2. The difference in metabolic behavior was analyzed for *T. erythraeum* grown under the two pCO_2 conditions. Transcriptional results reveal that there are several genes that are up-regulated and down-regulated, indicating that cyanobacteria are able to survive under high pCO_2. However, the data also suggests that the cyanobacteria might not be able to withstand and fight off other external pressures.

Further investigations where the conditions of the cyanobacteria growth are varied should be performed. In our study, cyanobacteria were grown in an environment where everything was kept constant except the change in CO_2

concentration. However, these organisms do not grow in restrictive conditions in nature. Thus, the transcriptional changes to *T. erythraeum* at high pCO_2 conditions and other varied conditions, such as nutrients and pH, should be studied. A change in nutrients would indicate if the cyanobacteria would be able to regulate their genes to thrive in high pCO_2. For example, the biomass of *T. erythraeum* increased under the projected pCO_2 conditions. An analysis of how the organism responds to a change in nutrient supply would give a clearer picture of the environmental condition. One of the experiments can involve the supply of iron. As *Trichodesmium* blooms require iron for survival, a change in the supply under different CO_2 conditions might affect transcription. Another important study needed would assess how change in pH as well as combined changes in pH and CO_2 conditions affect metabolic pathways. As ocean pH decreases and pCO_2 increases, the effect that it would have on gene transcription would give an insight into the long term impact. As our analysis has shown that the cell wall of *Trichodesmium* might be affected, an inspection of the cell wall using microscopy such as confocal and transmission electron microscopes would reveal if the integrity of the wall is altered by high pCO_2 and low pH. The study could show if the cyanobacteria would be able to survive in these conditions. Another essential study is to look at the impact that a change in temperature would have on *Trichodesmium*. With the increase of the concentration of CO_2, the temperature also rises. The study of the effect of temperature on *Trichodesmium* will also give an in-depth understanding of what happens to the organism as the current environment changes. The amount of metabolites when cyanobacteria is grown under various environmental conditions would also shed more light on the impact of the changing climate. Even though, mRNA sequencing gives us an indication about the genes that are affected, the experiment does not give information about the amount of metabolites produced. One way to quantify the metabolites is through metabolic fingerprinting. The technology allows to measure metabolite differences in organisms grown in different conditions. Moreover, studies done on other cyanobacteria with the conditions mentioned above would give an overall picture to what would happen to the species.

In an attempt to explore how proteins aid plants to survive temperature change, we combined bioinformatics and experimental studies to look at the wheat lipocalin. In chapter III a proposed 3D model of the wheat lipocalin was generated using five different programs. The model shows that the protein has the lipocalin fold of a β-barrel with eight antiparallel β-strands and a C-terminal helix. We also looked at the lipocalin superfamily by aligning proteins from different organisms and with various functions. The plant lipocalin from wheat was added to the structural alignment and used to find the positions that are conserved throughout evolution. The conservation analysis elucidated the amino acids that might be essential in the stability and folding of Tatil. A long-range interaction network in Tatil also shows which amino acids interact to stabilize the structure. The study in this chapter provided the foundation for our investigation of the function and structure of Tatil. In the future, the determinants of stability and folding of Tatil can be confirmed by using different variants of Tatil. The amino acids that were identified to be highly and moderately conserved can be mutated and compared with WT Tatil to observe if there would be any changes in folding. The next step in this work is to confirm the 3D structure of Tatil. One way the structure of the protein can be studied is through X-ray crystallography. Before establishing the 3D structure of Tatil, several steps need to be done. As we have expressed and purified the protein, the next step is to find the solubility range of Tatil for which no precipitation occurs. After that, the right conditions for crystal formation have to be determined by setting up various concentrations of the protein in different solutions. Then, to obtain a structure, the crystal that is formed needs to be well diffracting in an X-ray beam. The pattern produced by the beam is analyzed to find the structure of the protein.

In chapter IV, we looked at the localization of Tatil on the plasma membrane of an onion cell. We identified the possible amino acids that associate with the cell membrane. The amino acids were initially identified based on the bioinformatics studies that were done in chapter III. By looking at the solvent accessibility and hydrophobicity of Tatil, three amino acids were picked for the experimental mutagenesis study. Using a gene gun and confocal microscope, the location of Tatil variants within a cell were analyzed. The change in fluorescence intensity of the WT Tatil in comparison to the Phe96Gly, Pro98Gly and Ile99Gly variants show a difference in how the residues associate with the plasma membrane. The results suggest that the protein is associating with the plasma membrane in a diagonal manner. Moreover, we also determined conditions for the expression and purification of Tatil. Gel filtration chromatography showed the elution of two peaks suggesting that the wheat lipocalin may exist in two states, a trimer and a monomer. We were also able to characterize the two peaks using far- and near-UV CD. The result shows that they have different properties. Solving the structure of these two forms would clarify the findings. If confirmed, Tatil would be the first lipocalin to be shown to be a trimer. As the amount of protein produced in the current condition is low, other expression systems such as yeast can be investigated. Vectors other than pET14 can also be considered to solve the problem of low expression. Furthermore, the structure, folding and stability of the loop variants should be checked to see how they compare with the WT protein. The stability of the variants will give an additional confirmation that the amino acid residues are responsible for the localization of the protein at the plasma membrane.

This dissertation combines bioinformatics with experimental work to answer a fundamental question about global warming. Even though the questions that arise as a result of climate change are wide in its scope, we approached the problem by looking at two fundamental biological molecules, RNA and proteins, that can be responsive to global warming. In the study of *T. erythraeum* and the wheat lipocalin we have contributed to a better understanding of how gene expression changes under high pCO_2 and how a lipocalin protein may associate with the cell membrane to help plants survive climate change.

REFERENCES

Acevedo, E., Silva, P., and Silva, H. (2002), Wheat growth and physiology. Bread wheat improvement and production. FAO, Rome , 53–89.

Ahmad, S., Gromiha, M., Fawareh, H., and Sarai, A. (2004), ASAView: database and tool for solvent accessibility representation in proteins. BMC Bioinformatics *5*, 51.

Ahrens, C.D. (2011), Essentials of meteorology : an invitation to the atmosphere (Belmont, CA: Brooks/Cole).

Šali, A. and Blundell, T.L. (1993), Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol *234*, 779–815.

Altschul, S.F., Madden, T.L., Schäffer, 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.

Amoia, A.M. and Montfort, W.R. (2007), Apo-nitrophorin 4 at atomic resolution. Protein Science *16*, 2076–2081.

Anfinsen, C.B. (1973), Principles that govern the folding of protein chains. Science *181*, 223–230.

Ansorge, W.J. (2009), Next-generation DNA sequencing techniques. New biotechnol *25*, 195–203.

Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006), The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics *22*, 195–201.

Bachu, S. and Adams, J.J. (2003), Sequestration of CO_2 in geological media in response to climate change: capacity of deep saline aquifers to sequester CO_2 in solution. Energy Convers Manage *44*, 3151–3175.

Badger, M.R. and Price, G.D. (2003), CO_2 concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution. J Exp Bot *54*, 609–22.

Baldwin, R.L. (1989), How does protein folding get started? Trends Biochem Sci *14*, 291–294.

Batagelj, V. and Mrvar, A. (1998), Pajek-program for large network analysis. Connections *21*, 47–57.

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., and Courchamp, F. (2012), Impacts of climate change on the future of biodiversity. Ecol Lett *15*, 365–377.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., and Bignell, H.R. (2008), Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Bergman, B., Sandh, G., Lin, S., Larsson, J., and Carpenter, E.J. (2013), *Trichodesmium–a* widespread marine cyanobacterium with unusual nitrogen fixation properties. FEMS Microbiol Rev *37*, 286–302.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000), The protein data bank. Nucleic Acids Res *28*, 235–242.

Berman-Frank, I., Lundgren, P., Chen, Y.B., Küpper, H., Kolber, Z., Bergman, B., and Falkowski, P. (2001), Segregation of nitrogen fixation and oxygenic photosynthesis in the marine cyanobacterium *Trichodesmium*. Science *294*, 1534–1537.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977), The protein data bank. Eur J Biochem *80*, 319–324.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T.G., Bertoni, M., and Bordoli, L. (2014), SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res *42*, 252–258.

Bishop, R.E. (2000), The bacterial lipocalins. Bba-Protein Struct M *1482*, 73–83.

Bishop, R.E., Penfold, S.S., Frost, L.S., Holtje, J.V., and Weiner, J.H. (1995), Stationary phase expression of a novel *Escherichia coli* outer membrane lipoprotein and its relationship with mammalian apolipoprotein D. Implications for the origin of lipocalins. J Biol Chem *270*, 23097–23103.

Booth, B.B., Jones, C.D., Collins, M., Totterdell, I.J., Cox, P.M., Sitch, S., Huntingford, C., Betts, R.A., Harris, G.R., and Lloyd, J. (2012), High sensitivity of future global warming to land carbon cycle processes. Environ Res Lett *7*.

Bourne, P.E. and Weissig, H. (2003), Structural bioinformatics. Methods of biochemical analysis (Hoboken, NJ: Wiley-Liss).

Bradshaw, W.E. and Holzapfel, C.M. (2006), Evolutionary response to rapid climate change. Science *312*, 1477–1478.

Breitbarth, E., Oschlies, A., and LaRoche, J. (2007), Physiological constraints on the global distribution of *Trichodesmium* - effect of temperature on diazotrophy. Biogeosciences *4*, 53–61.

Breustedt, D.A., Korndorfer, I.P., Redl, B., and Skerra, A. (2005), The 1.8-A crystal structure of human tear lipocalin reveals an extended branched cavity with capacity for multiple ligands. J Biol Chem *280*, 484–93.

Brew, K. and Greene, L. (1997), Evolution, folding and flexibility. Protein Eng *10*, 44–44.

Bricker, T.M., Roose, J.L., Fagerlund, R.D., Frankel, L.K., and Eaton-Rye, J.J. (2012), The extrinsic proteins of Photosystem II. Bba-Bioenergetics *1817*, 121–142.

Britannica, Encyclopedia. (2015), Golgi apparatus gallery. Retrieved from http://www.britannica.com/EBchecked/media/115496/.

Bronner, I.F., Quail, M.A., Turner, D.J., and Swerdlow, H. (2014), Improved protocols for illumina sequencing. Curr Protoc Hum Genet , 18–2.

Brownlow, S., Morais Cabral, J.H., Cooper, R., Flower, D.R., Yewdall, S.J., Polikarpov, I., North, A.C., and Sawyer, L. (1997), Bovine beta-lactoglobulin at 1.8 A resolution-still an enigmatic lipocalin. Structure *5*, 481–95.

Bryant, D. (1995), The molecular biology of cyanobacteria (Norwell, MA: Kluwer Academic).

Campanacci, V., Bishop, R.E., Blangy, S., Tegoni, M., and Cambillau, C. (2006), The membrane bound bacterial lipocalin Blc is a functional dimer with binding preference for lysophospholipids. FEBS Lett *580*, 4877–83.

Campanacci, V., Nurizzo, D., Spinelli, S., Valencia, C., Tegoni, M., and Cambillau, C. (2004), The crystal structure of the *Escherichia coli* lipocalin Blc suggests a possible role in phospholipid binding. FEBS Lett *562*, 183–188.

Capone, D.G., Zehr, J.P., Paerl, H.W., Bergman, B., and Carpenter, E.J. (1997), *Trichodesmium*, a globally significant marine cyanobacterium. Science *276*, 1221–1229.

Carpenter, E.J., Capone, D.G., and Rueter, J.G. (1992), Marine pelagic cyanobacteria: *Trichodesmium* and other diazotrophs (Norwell, MA: Kluwer Academic).

Carver, B.F. (2009), Wheat : science and trade (Ames, Iowa: Wiley-Blackwell).

Cavaggioni, A. and Mucignat-Caretta, C. (2000), Major urinary proteins, alpha(2U)-globulins and aphrodisin. Biochim Biophys Acta *1482*, 218–28.

Chalfie, M. and Kain, S. (2006), Green fluorescent protein : properties, applications, and protocols. Methods of biochemical analysis (Hoboken, N.J.: Wiley-Interscience).

Chandler, D.E. and Roberson, R.W. (2009), Bioimaging : current concepts in light and electron microscopy (Sudbury, MA: Jones and Bartlett).

Charron, J.B., Breton, G., Badawi, M., and Sarhan, F. (2002), Molecular and structural analyses of a novel temperature stress-induced lipocalin from wheat and Arabidopsis. FEBS Lett *517*, 129–32.

Charron, J.B., Ouellet, F., Pelletier, M., Danyluk, J., Chauve, C., and Sarhan, F. (2005), Identification, expression, and evolutionary analyses of plant lipocalins. Plant Physiol *139*, 2017–28.

Charron, J.B.F., Ouellet, F., Houde, M., and Sarhan, F. (2008), The plant apolipoprotein D ortholog protects *Arabidopsis* against oxidative stress. Bmc Plant Biol *8*, 86.

Chen, R. (2012), Bacterial expression systems for recombinant protein production:*E. coli* and beyond. Biotechnol Adv *30*, 1102–1107.

Chen, Z. and Spreitzer, R. (1989), Chloroplast intragenic suppression enhances the low CO_2/O_2 specificity of mutant ribulose-bisphosphate carboxylase/oxygenase. J Biol Chem *264*, 3051–3053.

Chi, W.T., Fung, R.W.M., Liu, H.C., Hsu, C.C., and Charng, Y.Y. (2009), Temperature-induced lipocalin is required for basal and acquired thermotolerance in *Arabidopsis*. Plant Cell Environ *32*, 917–927.

Chimento, D.P., Mohanty, A.K., Kadner, R.J., and Wiener, M.C. (2003), Substrate-induced transmembrane signaling in the cobalamin transporter BtuB. Nat Struct Mol Biol *10*, 394–401.

Cho, J.H., Youn, J.W., and Sung, Y.C. (2001), Cross-priming as a predominant mechanism for inducing CD8+ T cell responses in gene gun DNA immunization. J Immunol *167*, 5549–5557.

Chothia, C. and Lesk, A.M. (1986), The relation between the divergence of sequence and structure in proteins. EMBO J *5*, 823.

Cianci, M., Rizkallah, P.J., Olczak, A., Raftery, J., Chayen, N.E., Zagalsky, P.F., and Helliwell, J.R. (2002), The molecular basis of the coloration mechanism in lobster shell: beta-crustacyanin at 3.2-angstrom resolution. P Natl Acad Sci *99*, 9795–9800.

Claxton, N.S., Fellers, T.J., and Davidson, M.W. (2006), Laser scanning confocal microscopy. Department of Optical Microscopy and Digital Imaging, Technical report, http://www. olympusconfocal. com/theory/LSCMIntro.pdf .

Collins, J.C. (2015), Investigations into protein folding and misfolding. Dissertation.

Colwell, R.K., Brehm, G., Cardelus, C.L., Gilman, A.C., and Longino, J.T. (2008), Global warming, elevational range shifts, and lowland biotic attrition in the wet tropics. Science *322*, 258–61.

Costantini, S., Colonna, G., and Facchiano, A.M. (2008), ESBRI: a web server for evaluating salt bridges in proteins. Bioinformation *3*, 137–8.

Cubitt, A.B., Heim, R., Adams, S.R., Boyd, A.E., Gross, L.A., and Tsien, R.Y. (1995), Understanding, improving and using green fluorescent proteins. Trends Biochem Sci *20*, 448–455.

Curtis, T. and Halford, N.G. (2014), Food security: the challenge of increasing wheat yield and the importance of not compromising food safety. Ann Appl Biol *164*, 354–372.

Delworth, T.L. and Zeng, F.R. (2014), Regional rainfall decline in Australia attributed to anthropogenic greenhouse gases and ozone levels. Nat Geosci *7*, 583–587.

Diaz, H.F. and Eischeid, J.K. (2007), Disappearing "alpine tundra" Koppen climatic type in the western United States. Geophys Res Lett *34*, L18707.

Dill, K.A., Bromberg, S., Yue, K., Chan, H.S., Ftebig, K.M., Yee, D.P., and Thomas, P.D. (1995), Principles of protein folding—a perspective from simple exact models. Protein Sci *4*, 561–602.

Dill, K.A. and MacCallum, J.L. (2012), The protein-folding problem, 50 years on. Science *338*, 1042–1046.

DOE, U. (2008), Carbon cycling and biosequestration integrating biology and climate through systems science : report from the March 2008 workshop.

Drayna, D., Fielding, C., McLean, J., Baer, B., Castro, G., Chen, E., Comstock, L., Henzel, W., Kohr, W., Rhee, L., and et al. (1986), Cloning and expression of human apolipoprotein D cDNA. J Biol Chem *261*, 16535–16539.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M., Eramian, D., Shen, M., Pieper, U., and Sali, A. (2006), Comparative protein structure modeling using Modeller. Curr Protoc Protein Sci , 5.6. 1–5.6. 30.

Fairman, J.W., Noinaj, N., and Buchanan, S.K. (2011), The structural biology of β-barrel membrane proteins: a summary of recent reports. Curr Opin Struc Biol *21*, 523–531.

Farmer, G.T. and Cook, J. (2013), Climate change science: Volume 1: The physical climate (New York: Springer).

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., and Turcatti, G. (2006), BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res *34*, e22.

Fersht, A.R. (1995), Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. P Natl Acad Sci *92*, 10869–10873.

Fersht, A.R. (1997), Nucleation mechanisms in protein folding. Curr Opin Struc Biol *7*, 3–9.

Finer, J.J., Vain, P., Jones, M.W., and McMullen, M.D. (1992), Development of the particle inflow gun for DNA delivery to plant cells. Plant Cell Rep *11*, 323–328.

Finzi-Hart, J.A., Pett-Ridge, J., Weber, P.K., Popa, R., Fallon, S.J., Gunderson, T., Hutcheon, I.D., Nealson, K.H., and Capone, D.G. (2009), Fixation and fate of C and N in the cyanobacterium *Trichodesmium* using nanometer-scale secondary ion mass spectrometry. P Natl Acad Sci *106*, 6345–6350.

Fiser, A., Do, R.K.G., and Šali, A. (2000), Modeling of loops in protein structures. Protein Sci *9*, 1753–1773.

Fiser, A., Feig, M., Brooks, C.L., and Sali, A. (2002), Evolution and physics in comparative protein structure modeling. Accounts Chem Res *35*, 413–421.

Fitch, W.M. (2000), Homology: a personal view on some of the problems. Trends Genet *16*, 227–231.

Flores, F.G. and Herrero, A. (2008), The cyanobacteria: molecular biology, genomics, and evolution (Norfolk, UK: Horizon Scientific Press).

Flower, D.R. (1996), The lipocalin protein family: structure and function. Biochemical J *318 (Pt 1)*, 1–14.

Francis-Lyon, P. and Koehl, P. (2014), Protein side-chain modeling with a protein-dependent optimized rotamer library. Proteins *82*, 2000–2007.

Friedlingstein, P., Andrew, R., Rogelj, J., Peters, G., Canadell, J., Knutti, R., Luderer, G., Raupach, M., Schaeffer, M., and van Vuuren, D. (2014), Persistent growth of CO_2 emissions and implications for reaching climate targets. Nat Geosci *7*, 709–715.

Ganfornina, M.D., Sánchez, D., and Bastiani, M.J. (1995), Lazarillo, a new GPI-linked surface lipocalin, is restricted to a subset of neurons in the grasshopper embryo. Development *121*, 123–134.

Gattuso, J.P. and Hansson, L. (2011), Ocean acidification (Oxford, England: Oxford University Press).

Gerland, P., Raftery, A.E., Sevcikova, H., Li, N., Gu, D.A., Spoorenberg, T., Alkema, L., Fosdick, B.K., Chunn, J., Lalic, N., Bay, G., Buettner, T., Heilig, G.K., and Wilmoth, J. (2014), World population stabilization unlikely this century. Science *346*, 234–237.

Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., and Smith, H.O. (2009), Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods *6*, 343–345.

Goetz, D.H., Willie, S.T., Armen, R.S., Bratt, T., Borregaard, N., and Strong, R.K. (2000), Ligand preference inferred from the structure of neutrophil gelatinase associated lipocalin. Biochemistry *39*, 1935–41.

Gomez-Baena, G., Lopez-Lozano, A., Gil-Martinez, J., Lucena, J.M., Diez, J., Candau, P., and Garcia-Fernandez, J.M. (2008), Glucose uptake and its effect on gene expression in *prochlorococcus*. Plos One *3*, e3416.

Gorlich, D. and Mattaj, I.W. (1996), Nucleocytoplasmic transport. Science *271*, 1513–1519.

Greene, L. and Brew, K. (1995), Conserved residues in the lipocalins suggest a folding motif. Protein Eng *8*, 100.

Greene, L.H., Chrysina, E.D., Irons, L.I., Papageorgiou, A.C., Acharya, K.R., and Brew, K. (2001), Role of conserved residues in structure and stability: Tryptophans of human serum retinol-binding protein, a model for the lipocalin superfamily. Protein Sci *10*, 2301–2316.

Greene, L.H., Hamada, D., Eyles, S.J., and Brew, K. (2003), Conserved signature proposed for folding in the lipocalin superfamily. FEBS lett *553*, 39–44.

Greene, L.H. and Higman, V.A. (2003), Uncovering network systems within protein structures. J Mol Biol *334*, 781–791.

Greenfield, N.J. (2007), Using circular dichroism spectra to estimate protein secondary structure. Nat Protoc *1*, 2876–2890.

Grishin, N.V. (2001), Fold change in evolution of protein structures. J Struct Biol *134*, 167–185.

Gromiha, M.M., Pujadas, G., Magyar, C., Selvaraj, S., and Simon, I. (2004), Locating the stabilizing residues in (α/β) 8 barrel proteins based on hydrophobicity, long-range interactions, and sequence conservation. Proteins *55*, 316–329.

Gromiha, M.M. and Selvaraj, S. (1999), Importance of long-range interactions in protein folding. Biophys Chem *77*, 49–68.

Grzyb, J., Latowski, D., and Strzałka, K. (2006), Lipocalins – a family portrait. J Plant Physiol *163*, 895–915.

Hasegawa, H. and Holm, L. (2009), Advances and pitfalls of protein structural alignment. Curr Opin Struc Biol *19*, 341–348.

Herrero, A. and Flores, E. (2008), The cyanobacteria : molecular biology, genomics, and evolution (Norfolk, UK: Caister Academic Press).

Hieber, A.D., Bugos, R.C., and Yamamoto, H.Y. (2000), Plant lipocalins: violaxanthin de-epoxidase and zeaxanthin epoxidase. Biochim Biophys Acta *1482*, 84–91.

Hinga, K.R. (2002), Effects of pH on coastal marine phytoplankton. Mar Ecol Prog Ser *238*, 281–300.

Holm, L. and Rosenström, P. (2010), Dali server: conservation mapping in 3D. Nucleic Acids Res *38*, W545–W549.

Hosaka, T., Meguro, T., Yamato, I., and Shirakihara, Y. (2003), Crystal structure of *Enterococcus hirae* enolase at 2.8 Å resolution. J Biochem *133*, 817–823.

Huang, Y., Smith, B.S., Chen, L.X., Baxter, R.H., and Deisenhofer, J. (2009), Insights into pilus assembly and secretion from the structure and functional characterization of usher PapC. P Natl Acad Sci *106*, 7403–7407.

Huber, R., Schneider, M., Epp, O., Mayr, I., Messerschmidt, A., Pflugrath, J., and Kayser, H. (1987), Crystallization, crystal structure analysis and preliminary molecular model of the bilin binding protein from the insect *Pieris brassicae*. J Mol Biol *195*, 423–434.

Humlum, O., Stordahl, K., and Solheim, J.E. (2013), The phase relation between atmospheric carbon dioxide and global temperature. Global Planet Change *100*, 51–69.

Hutchins, D.A., Fu, F.X., Zhang, Y., Warner, M.E., Feng, Y., Portune, K., Bernhardt, P.W., and Mulholland, M.R. (2007), CO_2 control of *Trichodesmium* N_2 fixation, photosynthesis, growth rates, and elemental ratios: Implications for past, present, and future ocean biogeochemistry. Limnol Oceanogr *52*, 1293–1304.

Ikeuchi, M., Eggers, B., Shen, G., Webber, A., Yu, J., Hirano, A., Inoue, Y., and Vermaas, W. (1991), Cloning of the psbK gene from *Synechocystis* sp. PCC 6803 and characterization of photosystem II in mutants lacking PSII-K. J Biol Chem *266*, 11111–11115.

Ingles-Prieto, A., Ibarra-Molero, B., Delgado-Delgado, A., Perez-Jimenez, R., Fernandez, J.M., Gaucher, E.A., Sanchez-Ruiz, J.M., and Gavira, J.A. (2013), Conservation of protein structure over four billion years. Structure *21*, 1690–1697.

Jamroz, M. and Kolinski, A. (2010), Modeling of loops in proteins: a multi-method approach. Bmc Struct Biol *10*, 5.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012), KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res *40*, D109–114.

Kasting, J.F. and Siefert, J.L. (2002), Life and the evolution of Earth's atmosphere. Science *296*, 1066–1068.

Kaufman, A.J. (2014), Early Earth: Cyanobacteria at work. Nat Geosci *7*, 253–254.

Kelly, S.M., Jess, T.J., and Price, N.C. (2005), How to study proteins by circular dichroism. Bba-Proteins Proteom *1751*, 119–139.

Kerfeld, C.A. and Krogmann, D.W. (1998), Photosynthetic cytochromes c in cyanobacteria, algae, and plants. Annu Rev Plant Phys *49*, 397–425.

Åkerström, B. (2006), Lipocalins. Molecular biology intelligence unit (Georgetown, TX: Landes Bioscience).

Khodayari, A., Wuebbles, D.J., Olsen, S.C., Fuglestvedt, J.S., Berntsen, T., Lund, M.T., Waitz, I., Wolfe, P., Forster, P.M., Meinshausen, M., Lee, D.S., and Lim, L.L. (2013), Intercomparison of the capabilities of simplified climate models to project the effects of aviation CO_2 on climate. Atmos Environ *75*, 321–328.

Kim, D.E., Chivian, D., and Baker, D. (2004), Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res *32*, W526–W531.

Kim, T.K. and Eberwine, J.H. (2010), Mammalian cell transfection: the present and the future. Anal Bioanal Chem *397*, 3173–3178.

Kirby, E. (2002), Botany of the wheat plant. Bread wheat: improvement and production. Food and Agriculture Organization of the United Nations (FAO), Rome .

Krebs, J.E., Lewin, B., Goldstein, E.S., and Kilpatrick, S.T. (2013), Lewin's essential genes (Burlington, MA: Jones Bartlett Publishers).

Kridel, R., Meissner, B., Rogic, S., Boyle, M., Telenius, A., Woolcock, B., Gunawardana, J., Jenkins, C., Cochrane, C., and Ben-Neriah, S. (2012), Whole transcriptome sequencing reveals recurrent NOTCH1 mutations in mantle cell lymphoma. Blood *119*, 1963–1971.

Krieger, E., Nabuurs, S.B., and Vriend, G. (2003), Homology modeling. Method Biochem Anal *44*, 509–524.

Krieger, F., Möglich, A., and Kiefhaber, T. (2005), Effect of proline and glycine residues on dynamics and barriers of loop formation in polypeptide chains. J Am Chem Soc *127*, 3346–3352.

Kroeker, K.J., Kordas, R.L., Crim, R.N., and Singh, G.G. (2010), Meta-analysis reveals negative yet variable effects of ocean acidification on marine organisms. Ecol Lett *13*, 1419–1434.

Kumar, K., Mella-Herrera, R.A., and Golden, J.W. (2010), Cyanobacterial Heterocysts. Cold Spring Harb Perspect Biol *2*.

Kuriyama, S., Mitoro, A., Tsujinoue, H., Nakatani, T., Yoshiji, H., Tsujimoto, T., Yamazaki, M., and Fukui, H. (2000), Particle-mediated gene transfer into murine livers using a newly developed gene gun. Gene Ther *7*, 1132–1136.

Kurland, C. and Gallant, J. (1996), Errors of heterologous protein expression. Curr Opin Biotech *7*, 489–493.

Kyte, J. and Doolittle, R.F. (1982), A simple method for displaying the hydropathic character of a protein. J Mol Biol *157*, 105–132.

Lambert, C., Léonard, N., De Bolle, X., and Depiereux, E. (2002), ESyPred3D: Prediction of proteins 3D structures. Bioinformatics *18*, 1250–1256.

Latysheva, N., Junker, V.L., Palmer, W.J., Codd, G.A., and Barker, D. (2012), The evolution of nitrogen fixation in cyanobacteria. Bioinformatics *28*, 603–606.

Lesk, A.M. (2002), Introduction to bioinformatics (Oxford; NY: Oxford University Press).

Levitan, O., Rosenberg, G., Setlik, I., Setlikova, E., Grigel, J., Klepetar, J., Prasil, O., and Berman-Frank, I. (2007), Elevated CO_2 enhances nitrogen fixation and growth in the marine cyanobacterium *Trichodesmium*. Global Change Biol *13*, 531–538.

Levitt, M. (1992), Accurate modeling of protein conformation by automatic segment matching. J Mol Biol *226*, 507–533.

Levitt, M., Hirshberg, M., Sharon, R., and Daggett, V. (1995), Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comput Phys Commun *91*, 215–231.

Lichtman, J.W. and Conchello, J.A. (2005), Fluorescence microscopy. Nat Methods *2*, 910–919.

Lichty, J.J., Malecki, J.L., Agnew, H.D., Michelson-Horowitz, D.J., and Tan, S. (2005), Comparison of affinity tags for protein purification. Protein Expres Purif *41*, 98–105.

Lin, M.T., Occhialini, A., Andralojc, P.J., Parry, M.A.J., and Hanson, M.R. (2014), A faster Rubisco with potential to increase photosynthesis in crops. Nature *513*, 547–550.

Liu, L.N., Chen, X.L., Zhang, Y.Z., and Zhou, B.C. (2005), Characterization, structure and function of linker polypeptides in phycobilisomes of cyanobacteria and red algae: an overview. Biochim Biophys Acta *1708*, 133–42.

Lobell, D.B. and Field, C.B. (2007), Global scale climate - crop yield relationships and the impacts of recent warming. Environmental Research Letters *2*.

Lobell, D.B., Sibley, A., and Ortiz-Monasterio, J.I. (2012), Extreme heat effects on wheat senescence in India. Nat Clim Chang *2*, 186–189.

Mahlstein, I., Daniel, J.S., and Solomon, S. (2013), Pace of shifts in climate regions increases with global temperature. Nat Clim Chang *3*, 739–743.

Mantyjarvi, R., Rautiainen, J., and Virtanen, T. (2000), Lipocalins as allergens. Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology *1482*, 308–317.

Mardis, E.R. (2008), Next-generation DNA sequencing methods. Annu Rev Genom Hum G *9*, 387–402.

Markowitz, V.M., Chen, I.M., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Anderson, I., Lykidis, A., Mavromatis, K., Ivanova, N.N., and Kyrpides, N.C. (2010), The integrated microbial genomes system: an expanding comparative analysis resource. Nucleic Acids Res *38*, D382–D390.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000), Comparative protein structure modeling of genes and genomes. Annu Rev Bioph Biom *29*, 291–325.

Mattanovich, D., Branduardi, P., Dato, L., Gasser, B., Sauer, M., and Porro, D. (2012), Recombinant protein production in yeasts (New York, NY: Springer). 329–358.

Matthijs, H. and Lubberding, H. (1988), Dark respiration in cyanobacteria. Biochemistry of the algae and cyanobacteria. (Oxford, UK: Clarendon Press).

Matz, S.A. (1991), The chemistry and technology of cereals as food and feed (McAllen, TX: Van Nostrand Reinhold).

Meining, W. and Skerra, A. (2012), The crystal structure of human α1-microglobulin reveals a potential haem-binding site. Biochem J *445*, 175–182.

Metzker, M.L. (2010), Sequencing technologies - the next generation. Nat Rev Genet *11*, 31–46.

Miyazaki, S., Fredricksen, M., Hollis, K.C., Poroyko, V., Shepley, D., Galbraith, D.W., Long, S.P., and Bohnert, H.J. (2004), Transcript expression profiles of *Arabidopsis thaliana* grown under controlled conditions and open-air elevated concentrations of CO_2 and of O_3. Field Crops Res *90*, 47–59.

Montzka, S.A., Dlugokencky, E.J., and Butler, J.H. (2011), Non-CO_2 greenhouse gases and climate change. Nature *476*, 43–50.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2014), Critical assessment of methods of protein structure prediction (CASP)—round X. Proteins *82*, 1–6.

Moya, A., Huisman, L., Ball, E.E., Hayward, D.C., Grasso, L.C., Chua, C.M., Woo, H.N., Gattuso, J.P., Foret, S., and Miller, D.J. (2012), Whole transcriptome analysis of the coral *acropora millepora* reveals complex responses to CO_2 driven acidification during the initiation of calcification. Mol Ecol *21*, 2440–2454.

Mulkidjanian, A.Y., Koonin, E.V., Makarova, K.S., Mekhedov, S.L., Sorokin, A., Wolf, Y.I., Dufresne, A., Partensky, F., Burd, H., Kaznadzey, D., Haselkorn, R., and Galperin, M.Y. (2006), The cyanobacterial genome core and the origin of photosynthesis. P Natl Acad Sci *103*, 13126–13131.

Nagano, N., Hutchinson, E., and Thornton, J.M. (1999), Barrel structures in proteins: automatic identification and classification including a sequence analysis of TIM barrels. Protein Sci *8*, 2072–2084.

Newcomer, M.E. (1993), Structure of the epididymal retinoic acid binding protein at 2.1 Å resolution. Structure *1*, 7–18.

Newcomer, M.E. and Ong, D.E. (2000), Plasma retinol binding protein: structure and function of the prototypic lipocalin. Bba-Protein Struct M *1482*, 57–64.

Nielsen, M., Lundegaard, C., Lund, O., and Petersen, T.N. (2010), CPHmodels-3.0—remote homology modeling using structure-guided sequence profiles. Nucleic Acids Res .

Noinaj, N., Kuszak, A.J., Gumbart, J.C., Lukacik, P., Chang, H., Easley, N.C., Lithgow, T., and Buchanan, S.K. (2013), Structural insight into the biogenesis of β-barrel membrane proteins. Nature *501*, 385–390.

Notredame, C., Higgins, D.G., and Heringa, J. (2000), T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol *302*, 205–217.

Oakley, A.J., Bhatia, S., Ecroyd, H., and Garner, B. (2012), Molecular dynamics analysis of apolipoprotein-D- lipid hydroperoxide interactions: Mechanism for selective oxidation of met-93. Plos One *7*.

O'Brien, J.A. and Lummis, S.C. (2011), Nano-biolistics: a method of biolistic transfection of cells and tissues using a gene gun with novel nanometer-sized projectiles. BMC Biotechnol *11*, 66.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999), KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res *27*, 29–34.

Ohki, K., Rueter, J.G., and Fujita, Y. (1986), Cultures of the pelagic cyanophytes *Trichodesmium erythraeum* and *Trichodesmium thiebautii* in synthetic medium. Mar Biol *91*, 9–13.

Ohnishi, N. and Takahashi, Y. (2001), PsbT polypeptide is required for efficient repair of photodamaged photosystem II reaction center. J Biol Chem *276*, 33798–33804.

Okamoto, O.K., Janjoppi, L., Bonone, F.M., Pansani, A.P., da Silva, A.V., Scorza, F.A., and Cavalheiro, E.A. (2010), Whole transcriptome analysis of the hippocampus: toward a molecular portrait of epileptogenesis. BMC genomics *11*, 230.

O'Leary, G.J., Christy, B., Nuttall, J., Huth, N., Cammarano, D., Stöckle, C., Basso, B., Shcherbak, I., Fitzgerald, G., and Luo, Q. (2014), Response of wheat growth, grain yield and water use to elevated CO_2 under a Free Air CO_2 Enrichment (FACE) experiment and modelling in a semi-arid environment. Global Change Biol .

Ozman, I. (2014), Implications of climate change for cyanobacteria over the Western Florida Shelf in the Gulf of Mexico. Master's thesis.

Paddock, S.W. (1999), Confocal laser scanning microscopy. Biotechniques *27*, 992–1007.

Paerl, H.W. and Huisman, J. (2009), Climate change: a catalyst for global expansion of harmful cyanobacterial blooms. Environmental Microbiol *1*, 27–37.

Paerl, H.W. and Otten, T.G. (2013), Harmful cyanobacterial blooms: causes, consequences, and controls. Microb Ecol *65*, 995–1010.

Pandolfi, J.M., Connolly, S.R., Marshall, D.J., and Cohen, A.L. (2011), Projecting coral reef futures under global warming and ocean acidification. Science *333*, 418–422.

Parker, L.M., Ross, P.M., and O'Connor, W.A. (2011), Populations of the Sydney rock oyster, *Saccostrea glomerata*, vary in response to ocean acidification. Mar Biol *158*, 689–697.

Parmesan, C., Burrows, M.T., Duarte, C.M., Poloczanska, E.S., Richardson, A.J., Schoeman, D.S., and Singer, M.C. (2013), Beyond climate change attribution in conservation and ecological research. Ecol Lett *16*, 58–71.

Parry, M., Rosenzweig, C., and Livermore, M. (2005), Climate change, global food supply and risk of hunger. Philos T R Soc B *360*, 2125–38.

Pasquier, C., Promponas, V., Palaios, G., Hamodrakas, J., and Hamodrakas, S. (1999), A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm. Protein Eng *12*, 381–385.

Patthy, L. (2008), Protein evolution (Malden, MA: Blackwell Science).

Paul, V.J. (2008), Global warming and cyanobacterial harmful algal blooms (Springer). 239–257.

Pauls, S.U., Nowak, C., Bálint, M., and Pfenninger, M. (2013), The impact of global climate change on genetic diversity within populations and species. Mol Ecol *22*, 925–946.

Pearce, J., Leach, C.K., and Carr, N.G. (1969), The incomplete tricarboxylic acid cycle in the blue-green alga *Anabaena variabilis*. J Gen Microbiol *55*, 371–8.

Pelroy, R.A., Rippka, R., and Stanier, R.Y. (1972), Metabolism of glucose by unicellular blue-green algae. Arch Mikrobiol *87*, 303–22.

Pervaiz, S. and Brew, K. (1987), Homology and structure-function correlations between alpha-1-acid glycoprotein and serum retinol-binding protein and its relatives. Faseb J *1*, 209–214.

Pfreundt, U., Kopf, M., Belkin, N., Berman-Frank, I., and Hess, W.R. (2014), The primary transcriptome of the marine diazotroph *Trichodesmium erythraeum* IMS101. Nature *4*.

Poloczanska, E.S., Brown, C.J., Sydeman, W.J., Kiessling, W., Schoeman, D.S., Moore, P.J., Brander, K., Bruno, J.F., Buckley, L.B., and Burrows, M.T. (2013), Global imprint of climate change on marine life. Nat Clim Chang *3*, 919–925.

Potterton, E., Briggs, P., Turkenburg, M., and Dodson, E. (2003), A graphical user interface to the CCP4 program suite. Acta Crystallogr D *59*, 1131–1137.

Prather, M.J., Holmes, C.D., and Hsu, J. (2012), Reactive greenhouse gas scenarios: Systematic exploration of uncertainties and the role of atmospheric chemistry. Geophys Res Lett *39*.

Prinn, R.G. (2003), The cleansing capacity of the atmosphere. Annu Rev Env Resour *28*, 29–57.

Ptitsyn, O., Pain, R.H., Semisotnov, G., Zerovnik, E., and Razgulyaev, O. (1990), Evidence for a molten globule state as a general intermediate in protein folding. FEBS Lett *262*, 20–24.

Quéré, C.L., Andres, R.J., Boden, T., Conway, T., Houghton, R., House, J.I., Marland, G., Peters, G.P., van der Werf, G., and Ahlström, A. (2013), The global carbon budget 1959–2011. Earth Syst Sci Data *5*, 165–185.

Radic, V. and Hock, R. (2011), Regionally differentiated contribution of mountain glaciers and ice caps to future sea-level rise. Nat Geosci *4*, 91–94.

Réale, D., Berteaux, D., McAdam, A., and Boutin, S. (2003), Lifetime selection on heritable life-history traits in a natural population of red squirrels. Evolution *57*, 2416–2423.

Raynaud, D., Jouzel, J., Barnola, J.M., Chappellaz, J., Delmas, R.J., and Lorius, C. (1993), The ice record of greenhouse gases. Science *259*, 926–934.

Rigby, M., Prinn, R.G., O'Doherty, S., Miller, B.R., Ivy, D., Muhle, J., Harth, C.M., Salameh, P.K., Arnold, T., Weiss, R.F., Krummel, P.B., Steele, L.P., Fraser, P.J., Young, D., and Simmonds, P.G. (2014), Recent and future trends in synthetic greenhouse gas radiative forcing. Geophysical Res Lett *41*, 2623–2630.

Rockström, J., Brasseur, G., Hoskins, B., Lucht, W., Schellnhuber, J., Kabat, P., Nakicenovic, N., Gong, P., Schlosser, P., and Costa, M.M. (2014), Climate change: the necessary, the possible and the desirable Earth league climate statement on the implications for climate policy from the 5th IPCC assessment. Earth's Future *12*, 606–611.

Root, T.L., Price, J.T., Hall, K.R., Schneider, S.H., Rosenzweig, C., and Pounds, J.A. (2003), Fingerprints of global warming on wild animals and plants. Nature *421*, 57–60.

Rose, P.W., Bi, C., Bluhm, W.F., Christie, C.H., Dimitropoulos, D., Dutta, S., Green, R.K., Goodsell, D.S., Prlić, A., and Quesada, M. (2013), The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res *41*, D475–D482.

Rost, B. (1996), [31] PHD: Predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol *266*, 525–539.

Rost, B. and Sander, C. (1994), Conservation and prediction of solvent accessibility in protein families. Proteins *20*, 216–226.

Ruiz, M., Wicker-Thomas, C., Sanchez, D., and Ganfornina, M.D. (2012), Grasshopper Lazarillo, a GPI-anchored Lipocalin, increases Drosophila longevity and stress resistance, and functionally replaces its secreted homolog NLaz. Insect Biochem Molec *42*, 776–789.

Sambrook, J., Russell, D.W., and Russell, D.W. (2001), Molecular cloning: a laboratory manual (Cold Spring Harbor, New York: Cold spring harbor laboratory press).

Sampathkumar, P., Lu, F., Zhao, X., Li, Z., Gilmore, J., Bain, K., Rutter, M.E., Gheyi, T., Schwinn, K.D., and Bonanno, J.B. (2010), Structure of a putative BenF-like porin from *Pseudomonas fluorescens* Pf-5 at 2.6 Å resolution. Proteins *78*, 3056–3062.

Sanchez, D., Ganfornina, M.D., and Bastiani, M.J. (1995), Developmental expression of the lipocalin Lazarillo and its role in axonal pathfinding in the grasshopper embryo. Development *121*, 135–147.

Sander, C. and Schneider, R. (1991), Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins *9*, 56–68.

Sandh, G., Ran, L., Xu, L., Sundqvist, G., Bulone, V., and Bergman, B. (2011), Comparative proteomic profiles of the marine cyanobacterium *Trichodesmium erythraeum* IMS101 under different nitrogen regimes. Proteomics *11*, 406–419.

Sanford, J.C. (2000), The development of the biolistic process. In Vitro Cell Dev-Pl *36*, 303–308.

Saqib, M., Akhtar, J., Abbas, G., and Nasim, M. (2013), Salinity and drought interaction in wheat (*Triticum aestivum L.*) is affected by the genotype and plant growth stage. Acta Physiol Plant *35*, 2761–2768.

Sañudo-Wilhelmy, S.A., Kustka, A.B., Gobler, C.J., Hutchins, D.A., Yang, M., Lwiza, K., Burns, J., Capone, D.G., Raven, J.A., and Carpenter, E.J. (2001), Phosphorus limitation of nitrogen fixation by *Trichodesmium* in the central Atlantic Ocean. Nature *411*, 66–69.

Sayle, R.A. and Milner-White, E.J. (1995), RASMOL: biomolecular graphics for all. Trends Biochem Sci *20*, 374–376.

Schlehuber, S. and Skerra, A. (2005), Lipocalins in drug discovery: from natural ligand-binding proteins to anticalins. Drug Discov Today *10*, 23–33.

Schneider, D., Volkmer, T., and Rogner, M. (2007), PetG and PetN, but not PetL, are essential subunits of the cytochrome b6f complex from *Synechocystis* PCC 6803. Res Microbiol *158*, 45–50.

Schnoor, J.L. (2007), The IPCC fourth assessment. Environ Sci Technol *41*, 1503.

Schueler-Furman, O. and Baker, D. (2003), Conserved residue clustering and protein structure prediction. Proteins *52*, 225–35.

Schwede, T. (2013), Protein modeling: what happened to the “protein structure gap”? Structure *21*, 1531–1540.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C. (2003), SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res *31*, 3381–3385.

Schwede, T. and Peitsch, M.C. (2008), Computational structural biology methods and applications (Tuck Link, Singapore: World Scientific).

Seinfeld, J.H. and Pandis, S.N. (2012), Atmospheric chemistry and physics from air pollution to climate change (Hoboken, NJ: Wiley).

Sevvana, M., Ahnstrom, J., Egerer-Sieber, C., Lange, H.A., Dahlback, B., and Muller, Y.A. (2009), Serendipitous fatty acid binding reveals the structural determinants for ligand recognition in apolipoprotein M. J Mol Biol *393*, 920–36.

Shaw, E.C., Mcneil, B.I., Tilbrook, B., Matear, R., and Bates, M.L. (2013), Anthropogenic changes to seawater buffer capacity combined with natural reef metabolism induce extreme future coral reef CO_2 conditions. Glob Change Biol *19*, 1632–1641.

Shen, J.R., Burnap, R.L., and Inoue, Y. (1995), An independent role of cytochrome c-550 in cyanobacterial photosystem-II as revealed by double-deletion mutagenesis of the *psbo* and *psbv* genes in *Synechocystis* Sp Pcc-6803. Biochemistry *34*, 12661–12668.

Shendure, J. and Ji, H. (2008), Next-generation DNA sequencing. Nat Biotechnol *26*, 1135–1145.

Shimomura, O., Johnson, F.H., and Saiga, Y. (1962), Extraction, purification and properties of Aequorin, a bioluminescent protein from luminous Hydromedusan, *Aequorea*. J Cell Compar Physl *59*, 223–239.

Sikosek, T., Bornberg-Bauer, E., and Chan, H.S. (2012), Evolutionary dynamics on protein bi-stability landscapes can potentially resolve adaptive conflicts. PLoS Comput Biol *8*.

Slavotinek, A.M. and Biesecker, L.G. (2001), Unfolding the role of chaperones and chaperonins in human disease. Trends Genet *17*, 528–535.

Solomon, S., Plattner, G.K., Knutti, R., and Friedlingstein, P. (2009), Irreversible climate change due to carbon dioxide emissions. P Natl Acad Sci *106*, 1704–1709.

Son, M.S. and Taylor, R.K. (2011), Preparing DNA libraries for multiplexed paired-end deep dequencing for Illumina GA sequencers. Curr Protoc Microbiol , 1E–4.

Spinelli, S., Vincent, F., Pelosi, P., Tegoni, M., and Cambillau, C. (2002), Boar salivary lipocalin. Eur J Biochem *269*, 2449–2456.

Sreerama, N., Manning, M.C., Powers, M.E., Zhang, J.X., Goldenberg, D.P., and Woody, R.W. (1999), Tyrosine, phenylalanine, and disulfide contributions to the circular dichroism of proteins: circular dichroism spectra of wild-type and mutant bovine pancreatic trypsin inhibitor. Biochemistry *38*, 10814–10822.

Stern, H., de Hoedt, G., and Ernst, J. (2000), Objective classification of Australian climates. Aust Meteorol Mag *49*, 87–96.

Stocker, T.F., Qin, D., Plattner, G.K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and P.M (2013), IPCC fifth assessment report. Weather *68*, 310–310.

Stofko-Hahn, R.E., Carr, D.W., and Scott, J.D. (1992), A single step purification for recombinant proteins characterization of a microtubule associated protein (MAP 2) fragment which associates with the type II cAMP-dependent protein kinase. FEBS Lett *302*, 274–278.

Summerfield, T.C. and Sherman, L.A. (2008), Global transcriptional response of the alkali-tolerant cyanobacterium Synechocystis sp strain PCC 6803 to a pH 10 environment. Appl Environ Microb *74*, 5276–5284.

Swinnen, K., Krul, A., Van Goidsenhoven, I., Van Tichelt, N., Roosen, A., and Van Houdt, K. (2007), Performance comparison of protein A affinity resins for the purification of monoclonal antibodies. J Chromatogr B *848*, 97–107.

Szilagyi, A. and Zhang, Y. (2014), Template-based structure modeling of protein–protein interactions. Curr Opin Struc Biol *24*, 10–23.

Tang, K.H., Tang, Y.J., and Blankenship, R.E. (2011), Carbon metabolic pathways in phototrophic bacteria and their broader evolutionary implications. Front Microbiol *2*, 165.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., and Natale, D.A. (2003), The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*, 41.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997), A genomic perspective on protein families. Science *278*, 631–7.

Terpe, K. (2003), Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. Appl Microbiol Biot *60*, 523–533.

Terpe, K. (2006), Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. Appl Microbiol Biot *72*, 211–222.

Thompson, L.R., Zeng, Q., Kelly, L., Huang, K.H., Singer, A.U., Stubbe, J., and Chisholm, S.W. (2011), Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. P Natl Acad Sci *108*, E757–E764.

Timm, D.E., Baker, L.J., Mueller, H., Zidek, L., and Novotny, M.V. (2001), Structural basis of pheromone binding to mouse major urinary protein (MUP-I). Protein Sci *10*, 997–1004.

Trenberth, K.E., Dai, A., van der Schrier, G., Jones, P.D., Barichivich, J., Briffa, K.R., and Sheffield, J. (2014), Global warming and changes in drought. Nat Clim Chang *4*, 17–22.

Tsien, R.Y. (1998), The green fluorescent protein. Annu Rev Biochem *67*, 509–544.

Twine, N.A., Janitz, K., Wilkins, M.R., and Janitz, M. (2011), Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. PloS one *6*, e16266.

Ujwal, R., Cascio, D., Colletier, J.P., Faham, S., Zhang, J., Toro, L., Ping, P., and Abramson, J. (2008), The crystal structure of mouse VDAC1 at 2.3 Å resolution reveals mechanistic insights into metabolite gating. P Natl Acad Sci *105*, 17742–17747.

Van Asch, M., Salis, L., Holleman, L.J., Van Lith, B., and Visser, M.E. (2013), Evolutionary response of the egg hatching date of a herbivorous insect under climate change. Nat Clim Chang *3*, 244–248.

Vandeputte-Rutten, L., Bos, M.P., Tommassen, J., and Gros, P. (2003), Crystal structure of Neisserial surface protein A (NspA), a conserved outer membrane protein with vaccine potential. J Biol Chem *278*, 24825–24830.

Vandeputte-Rutten, L., Kramer, R.A., Kroon, J., Dekker, N., Egmond, M.R., and Gros, P. (2001), Crystal structure of the outer membrane protease OmpT from *Escherichia coli* suggests a novel catalytic site. EMBO J *20*, 5033–5039.

Vega, M.C., Lorentzen, E., Linden, A., and Wilmanns, M. (2003), Evolutionary markers in the (β/α) 8-barrel fold. Curr Opin Chem Biol *7*, 694–701.

Venselaar, H., Joosten, R.P., Vroling, B., Baakman, C.A., Hekkelman, M.L., Krieger, E., and Vriend, G. (2010), Homology modelling and spectroscopy, a never-ending love story. Eur Biophys J *39*, 551–563.

Vincent, F., Löbel, D., Brown, K., Spinelli, S., Grote, P., Breer, H., Cambillau, C., and Tegoni, M. (2001), Crystal structure of aphrodisin, a sex pheromone from female hamster. J Mol Biol *305*, 459–469.

Von Arnim, A., Deng, X.W., and Stacey, M. (1998), Cloning vectors for the expression of green fluorescent protein fusion proteins in transgenic plants. Gene *221*, 35–43.

Wang, L., Li, P., and Brutnell, T.P. (2010), Exploring plant transcriptomes using ultra high-throughput sequencing. Brief Funct Genomics *9*, 118–128.

Wang, Z., Gerstein, M., and Snyder, M. (2009), RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet *10*, 57–63.

Watson, R.P., Demmer, J., Baker, E.N., and Arcus, V.L. (2007), Three-dimensional structure and ligand binding properties of trichosurin, a metatherian lipocalin from the milk whey of the common brushtail possum Trichosurus vulpecula. Biochem J *408*, 29–38.

Wei, H.R., Gou, J.Q., Yordanov, Y., Zhang, H.X., Thakur, R., Jones, W., and Burton, A. (2013), Global transcriptomic profiling of aspen trees under elevated [CO_2] to identify potential molecular mechanisms responsible for enhanced radial growth. J Plant Res *126*, 305–320.

White, G.W., Gianni, S., Grossmann, J.G., Jemth, P., Fersht, A.R., and Daggett, V. (2005), Simulation and experiment conspire to reveal cryptic intermediates and a slide from the nucleation-condensation to framework mechanism of folding. J Mol Biol *350*, 757–775.

Whitton, B.A. (2012), Ecology of Cyanobacteria II (Netherlands: Springer).

Wierenga, R. (2001), The TIM-barrel fold: a versatile framework for efficient enzymes. FEBS Lett *492*, 193–198.

Wilkins, M.R., Gasteiger, E., Bairoch, A., Sanchez, J.C., Williams, K.L., Appel, R.D., and Hochstrasser, D.F. (1999), Protein identification and analysis tools in the ExPASy server (Humana Press). 531–552.

Xu, N. and Dahlback, B. (1999), A novel human apolipoprotein (apoM). J Biol Chem *274*, 31286–90.

Xu, X., Zhang, Y., Williams, J., Antoniou, E., McCombie, W.R., Wu, S., Zhu, W., Davidson, N.O., Denoya, P., and Li, E. (2013), Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets. BMC bioinformatics *14*, S1.

Yang, C.Y., Gu, Z.W., Blanco-Vaca, F., Gaskell, S.J., Yang, M., Massey, J.B., Gotto, A.M.J., and Pownall, H.J. (1994), Structure of human apolipoprotein D: locations of the intermolecular and intramolecular disulfide links. Biochemistry *33*, 12451–12455.

Yang, F., Moss, L.G., and Phillips, G.N. (1996), The molecular structure of green fluorescent protein. Nat Biotechnol *14*, 1246–1251.

Yokoyama, S. (2003), Protein expression systems for structural genomics and proteomics. Curr Opin Chem Biol *7*, 39–43.

Zachariae, U., Klühspies, T., De, S., Engelhardt, H., and Zeth, K. (2006), High resolution crystal structures and molecular dynamics studies reveal substrate binding in the porin Omp32. J Biol Chem *281*, 7413–7420.

Zanotti, G., Panzalorto, M., Marcato, A., Malpeli, G., Folli, C., and Berni, R. (1998), Structure of pig plasma retinol-binding protein at 1.65 A resolution. Acta Crystallogr D Biol Crystallogr *54*, 1049–52.

Zeebe, R.E., Zachos, J.C., Caldeira, K., and Tyrrell, T. (2008), Carbon emissions and acidification. Science *321*, 51–52.

Zeller, G., Henz, S.R., Widmer, C.K., Sachsenberg, T., Rätsch, G., Weigel, D., and Laubinger, S. (2009), Stress-induced changes in the *Arabidopsis thaliana* transcriptome analyzed using whole-genome tiling arrays. Plant J *58*, 1068–1082.

Zhang, Q., Wang, Y.P., Matear, R., Pitman, A., and Dai, Y. (2014), Nitrogen and phosphorous limitations significantly reduce future allowable CO_2 emissions. Geophys Res Lett *41*, 632–637.

Zhang, S. and Bryant, D.A. (2011), The tricarboxylic acid cycle in cyanobacteria. Science *334*, 1551–3.

Zvelebil, M.J. and Baum, J.O. (2008), Understanding bioinformatics (New York, NY: Garland Science).

Zwanzig, R., Szabo, A., and Bagchi, B. (1992), Levinthal's paradox. P Natl Acad Sci *89*, 20–22.

APPENDIX A

LIST OF FIGURES WITH PERMISSIONS OBTAINED FROM PUBLISHERS

Figure 1	Figure 2
Figure 12	Figure 24
Figure 25	Figure 36
Figure 38	Figure 42

VITA

Nardos Sori
Department of Chemistry and Biochemistry
Old Dominion University
Norfolk, VA 23529

- **Education**
  - 2015 Ph.D. in Chemistry, Old Dominion University, Norfolk, VA 23529
  - 2006 B.S. in Biochemistry, Old Dominion University, Norfolk, VA 23529
- **Publications**
  - A window into the year 2100: The effect of high pCO2 on the gene expression in the cyanobacteria *Trichodesmium erythraeum*. (*In progress*)
  - Initial structural characterization of the wheat temperature induced lipocalin and insight into its localization on the cell membrane. (*In progress*)
- **Presentations**
  - 28th Annual Symposium of the Protein Society meeting in San Diego, California. “Bioinformatics and network analysis of lipocalin superfamily”, July 27-30, 2014.
  - 246th ACS National Meeting in Indianapolis, Indiana. “Key determinants of lipocalin protein evolution, stability, and folding”, September 8-12, 2013.
  - 244th ACS National Meeting in Philadelphia, PA. “Window into the year 2100: The effect of high pCO2 on gene expression in the cyanobacteria *Trichodesmium erythraeum*”, August 19-23, 2012.
- **Award**
  - Teaching assistant of the year in chemistry, Old Dominion University, April 2012