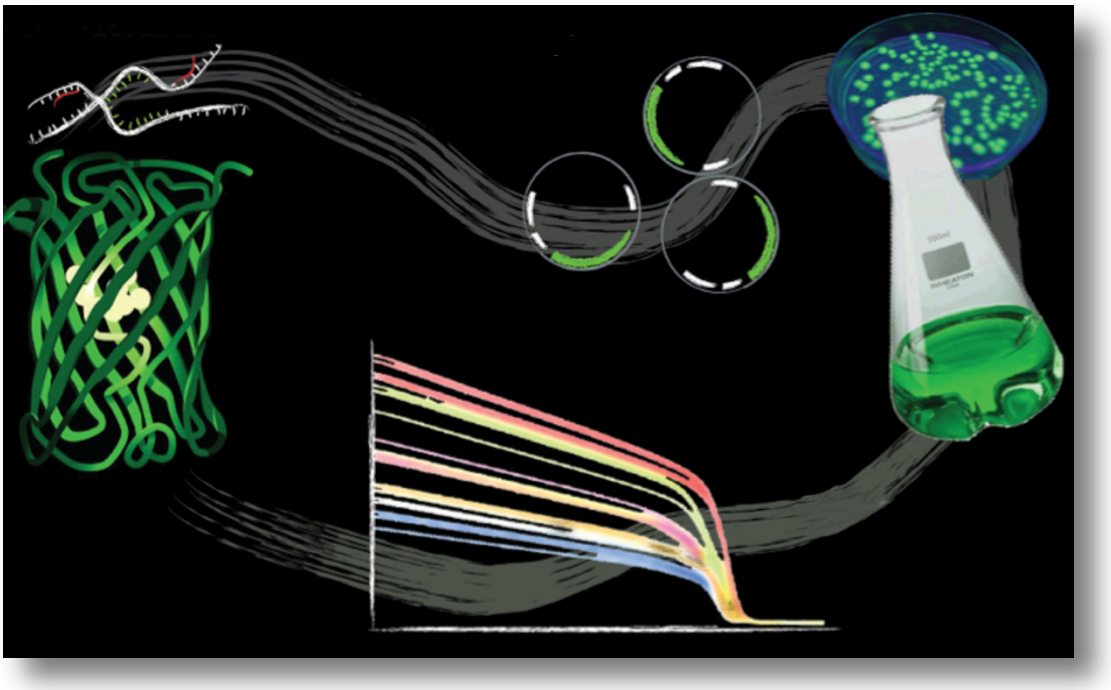
Ludwig-Maximilians Universität München
Faculty of Physics
Center for NanoScience (CeNS)
Chair of Applied Physics and Biophysics

Creation, Expression, Purification and Characterization of GFP

G4b GFP Expression and Melting Curves

Version 2017 English



Outline

1 | Introduction

Molecular biology is a cornerstone of modern biological science, and is closely tied with biochemistry and genetics. It is concerned with understanding and harnessing the activity of biomolecules – such as DNA and enzymes. Moreover, molecular biology employs innovative techniques that utilize biological components already developed in nature in order to carry out novel functions. This advanced practical course provides a first look into some of the fundamental techniques used by molecular biologists. Green Fluorescent Protein (GFP) serves not only as a model for any protein of interest in the techniques we will employ, but is also itself an invaluable tool in the biological sciences.

1.1 GFP – From a Jellyfish to Nobel Prize

Green Fluorescent Protein (GFP) is naturally found in the species of jellyfish *Aequorea victoria* (*Figure 1.1*) along with aequorin – another protein that is bound with a light-releasing luciferin prosthetic group. Together, they form a multi-step enzymatic process that results in **bioluminescence**. The jellyfish releases Ca^{2+} ions, which activates the catalytic aequorin and allow it to break down luciferin. This breakdown causes the luciferin group to temporarily enter an excited state. As it relaxes into the ground state, it releases the excess energy as blue light of λ = 465 nm. A nearby molecule of GFP will absorb some of this blue light, exciting its own chromophore. Then, as the GFP chromophore relaxes, it releases green light of λ = 509 nm.[1]

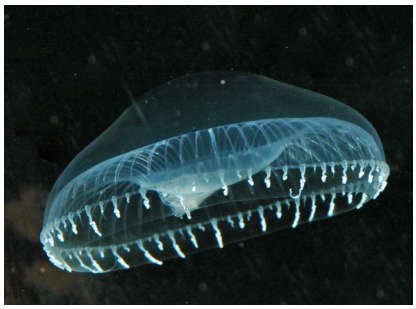


Figure 1.1: Photograph of *A. Victoria*. The bioluminescence of the jellyfish can be observed as a ring around its exterior. GFP was first discovered in this jellyfish.[2]

What are some useful properties of GFP? In contrast to many other proteins, GFP is a relatively "hands-off" protein. That is to say, it will naturally fold into its native state without the aid of chaperone proteins, it does not require any post-translational modifications, it has a relatively small size of 27kDa, it is predominantly monomeric, it is generally non-toxic for host organisms, and it auto-catalytically forms the chromophore that produces its fluorescent properties. This overall means that GFP can be easily fused with many proteins of interest and expressed in a number of organisms – from *E. coli* to mice. Importantly, it will also generally retain its fluorescent properties. This makes it a reliable, versatile protein with unique properties that enable detection.[3]

How can we harness this protein in research? While GFP is a rather single-function protein (absorbs blue light and emits green light), it has been employed in countless studies and is considered an essential tool in the biological sciences. To name a few examples, GFP can be fused

to another protein of interest to track its localization in a cell. It can also be used as an expression reporter for the presence of other biomolecules that are not so easily detected. Many derivatives of GFP have been developed to absorb and emit light across a wide spectrum (*Figure 1.2*). In much the same way that aequorin transfers its blue light to GFP which then emits green light, genetically-encoded light transfer systems using GFP and other fluorescent proteins are possible.[3]

In summary, GFP is used ubiquitously in molecular biology, cell biology, biochemistry and related fields. In fact, researchers Roger Y. Tsien, Osamu Shimomura and Martin Chalfie were awarded the 2008 Nobel Prize in Chemistry for the discovery and early engineering of GFP. While it might seem like an unlikely story, this small protein from a jellyfish has made vital contributions to science, and will likely remain an indispensable tool in many fields of research.

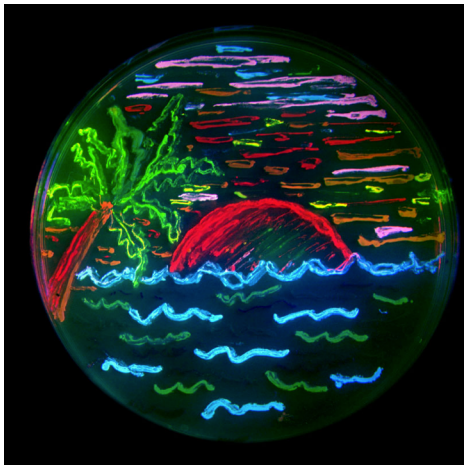


Figure 1.2: Fluorescent proteins expressed in living bacteria spread onto an agar plate. These proteins include GFP and some of its engineered variants. The display of colors demonstrates the range of fluorescent properties (e.g., λ_{ex} and λ_{em} maxima) of these proteins.[4]

1.2 Molecular Biology in this Practical Course

The techniques used in this practical course are a sampling of some of the most fundamental and important techniques used frequently by molecular biologists, biochemists, and biophysicists. Everything from assembly of the GFP gene construct to expression and finally isolation of the GFP protein will be covered. Additionally, as this course relates back to the study of physics in general, we will explore the unfolding behavior of GFP as a function of temperature. From a simple melting experiment, we will investigate the properties that determine folding and stability of GFP and proteins in general (*Figure 1.3*).

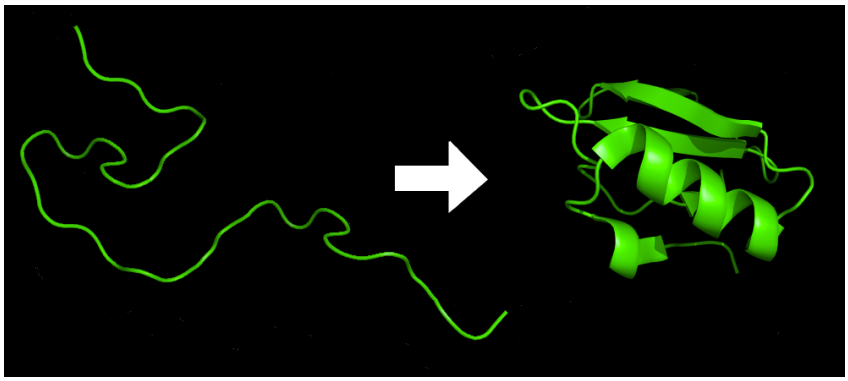


Figure 1.3: The unfolded and folded states of a protein. The process of protein folding is a complicated process, but is ultimately determined by the core principles of thermodynamics.[5]

2 | Biophysical View of GFP

A core principle of biophysics is that there is a direct connection between structure and function. The model protein of GFP is no exception. Its easily recognizable shape serves a specific purpose of scaffolding the light-sensitive chromophore at its core, and ultimately its structure is determined by the underlying process of protein folding. Thus, GFP is an excellent model protein for the investigation of several biophysical parameters.

2.1 Protein Folding – A Step-by-Step Process

The structure of a protein can be considered on several levels. The **primary structure** is the amino acid sequence. This is the most basic view of a protein, although the primary structure is often insufficient to predict the final conformation. Many proteins with dissimilar amino acid sequences have similar final conformations, while single point mutations in a sequence can dramatically change a protein's fold.

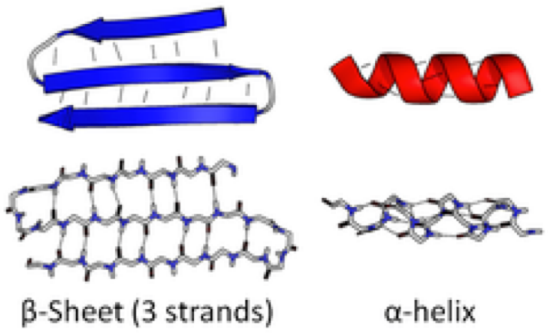


Figure 2.1: Canonical secondary structures. Two of the most easily-recognizable secondary structures are β-sheets and α-helices. These form spontaneously based on local non-covalent interactions between amino acid side-chains and backbones.[6]

As the growing protein polymer is produced by the ribosome, the amino acids will interact with their neighbors and form **secondary structures**. These include canonical structures such as **β-sheets and α-helices** (*Figure 2.1*). The driving forces of the formation of these structures are hydrogen bonds (H-bonds), ionic interactions, van der Waals interactions, and hydrophobic interactions. These non-covalent interactions occur between amino acid side chains as well as the amino-carboxy backbone. Importantly, secondary structures often form between neighboring amino acids. Therefore, stretches of the primary structure of a protein can sometimes predict the formation of canonical, well-characterized secondary structures. Simultaneously, a "secondary structure" might instead consist of a flexible linker region that does not adopt a locally stable or predictable shape.

The **tertiary structure** is the final, stabilized fold that forms when secondary structures or flexible regions interact with each other. A primary driving force behind this stage of folding is the fact that hydrophobic side-groups and the aqueous environment of a cell unfavorably interact with each other. In much the same way that oil and water don't mix, the hydrophobic groups are sequestered away from the hydrophilic environment. In other words, the protein folds in a way such that the hydrophobic side-groups are hidden in the core of the protein and shielded by an outer layer of hydrophilic side-groups that can favorably interact with water.

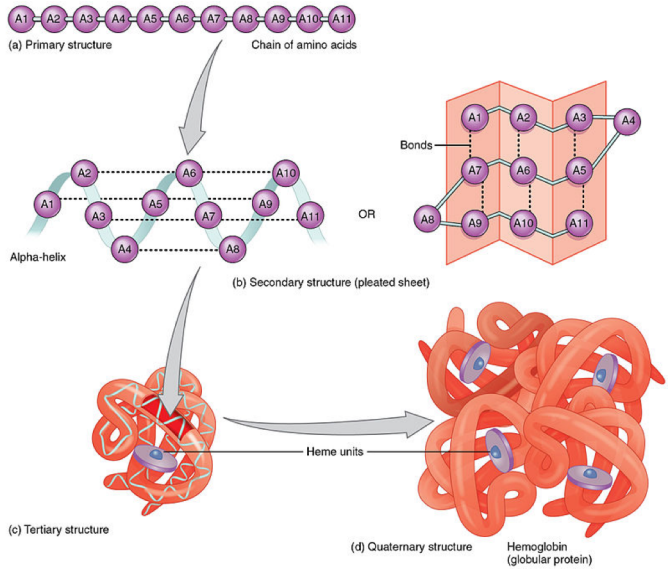


Figure 2.2: Formation of tetrameric hemoglobin. Each stage of protein folding results from interactions between the amino acids and their environment.[7]

Lastly, a **quaternary structure** is the assembly of multi-domain complexes from separately translated proteins. For example, the oxygen- and CO_2-carrying hemoglobin complex is formed as a heterotetramer from four distinct subunits (*Figure 2.2*). The interactions that drive the formation of quaternary structures are non-covalent interactions at the subunit interfaces, although the final structure is often covalently stabilized, e.g. via disulfide bonds.

The "**native state**" of a protein is more or less synonymous with a correct, functional fold of the protein. For some proteins, this happens spontaneously as they are translated by ribosomes. For others, a chaperone protein is required to assist folding. Chaperone proteins function by physically preventing other proteins from adopting incorrect conformations as they stabilize into a tertiary structure. Therefore, a chaperone limits the possible conformational states another protein can occupy, and increases the likelihood that it folds into the correct state.[8][9]

2.2 Energetics of Protein Folding (Qualitative)

The formation of secondary structures can be considered a largely **enthalpic** process (i.e., the formation of bonds or bond-like interactions that stabilize atoms within the molecule). Because these bonds lead to a minor decrease in energy (and therefore increase in stability), they form spontaneously and favorably. Simultaneously, the tertiary structure of a protein forms largely as an **entropic** process; the entropy gained from shielding the hydrophobic groups from water molecules is highly favorable compared to the unfolded state of the protein.

Hydrophobic side-groups are labeled as "hydrophobic" because they do not possess polar or charged groups that can form H-bonds with water. Thus, the water molecules will exclude the hydrophobic groups in favor of hydrophilic groups or other water molecules. Again, think of what happens when you combine oil and water – they naturally separate and form an interface with the smallest possible surface area.

Before the tertiary folding occurs, hydrophobic side-groups are exposed to the aqueous environment of the cell. Water readily forms H-bonds with other molecules, but this is not possible with hydrophobic side-groups. If a hydrophobic side-group is exposed to water, it causes the water to become ordered, and decreases the number of microstates that the whole system (protein and

environment) can occupy. This overall leads to an unfavorable local *reduction* of entropy. Consequently, the protein folds in such a way that hydrophobic side-groups are shielded from water (i.e., sequestered in the core of the protein), thereby preventing the water from becoming ordered. Compared to the state of the protein before the formation of the tertiary structure, this incurs a favorable *increase* of entropy.

Overall, the main driver of tertiary folding is shielding the hydrophobic side-groups from water molecules. Additionally, by providing a surface of hydrophilic side-groups, the water molecules can now favorably interact with the protein by forming H-bonds. This provides a minor contribution to the stability of the fold on an enthalpic level.

However, it is important to address the other side of the protein folding coin: a stably folded protein has fewer degrees of freedom than a protein that resembles a freely-jointed chain, and therefore has an unfavorable *reduction* of entropy. Protein folding is a balancing act between several competing forces. It is important to remember that while protein folding is energetically unfavorable in some regards, in others it is favorable. Ultimately, the sum total of energy lost and gained by the system in different microstates – or the **change in Gibbs free energy** – will determine a protein's final state. In fact, for many small peptides that do not have hydrophobic residues, their "final state" is that of an extremely unstructured flexible chain that can occupy many microstates.

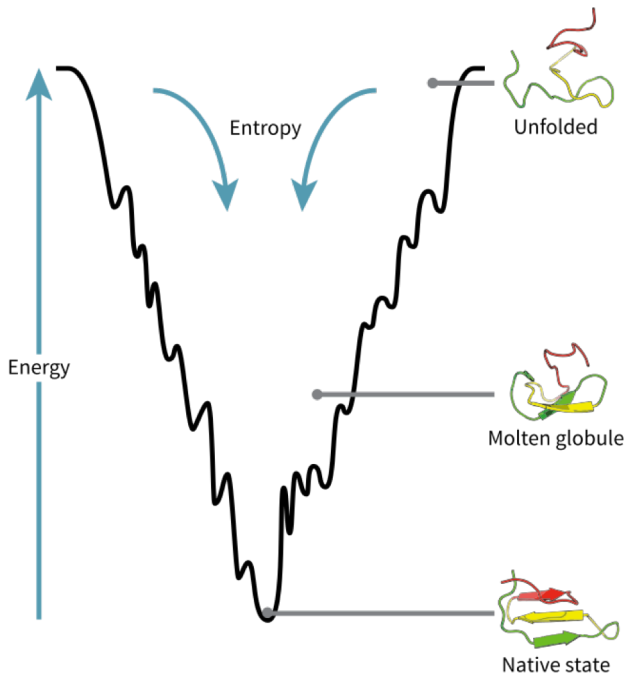


Figure 2.3: Energy landscape of protein folding. The thermodynamic processes that govern protein folding can be abstracted as an energy funnel that drives the conformation of the protein into increasingly lower energy states. The increase in entropy from shielding the hydrophobic core of the protein from the aqueous environment is a key force behind folding.[10]

Protein folding is often abstracted as an **energy landscape**, which allows you to consider not only the state that a protein is in, but many different states that it could occupy. Given the sequence of a protein and the environment it exists in, there will be microstates of higher and lower energies. Thermodynamic processes drive the protein to adopt conformations of lower energy states until it reaches an energetic minimum, which is for a correctly folded protein its native state (*Figure 2.3*).

Just as thermodynamics determine how a protein will fold, they likewise play a part in protein unfolding. Changes to the protein's environment – such as temperature, pH, salt concentration, or hydrophobicity – effectively alter the playing field. Introducing something to the system that makes it more favorable for the protein's hydrophobic core to interact with the surroundings decreases the entropic force that drives protein folding in the first place. For example, an amphiphilic

detergent in solution can bond to the hydrophobic side-chains. Consequently, the protein is free to adopt new microstates. If it finds a new energy state in a local energetic minimum such that it can no longer revert to the native state, the unfolded protein is considered to be "**denatured**".[8]

2.3 Energetics of Protein Folding (Quantitative)

The **free energy** of the current state of a system relative to nearby microstates determines its **stability**. If a nearby microstate would confer a comparative reduction in free energy, the system will spontaneously adopt that new conformation. In a general sense, this is the principle that governs protein folding and unfolding.

The native state of a protein is a conformational state that was achieved as a result of the unfolded peptide chain exploring microstates of increasingly lower free energy, until finally it reached a significant local energy minimum and structural stability. The denatured (i.e., unfolded or incorrectly folded) state of a protein results in two ways. Either the protein was never able to achieve its native state, or its native state was disrupted. By strategically altering the environment – e.g., by increasing the temperature of the system – we can denature proteins, and consequently investigate the thermodynamics of protein folding.[8]

A protein may be considered to occupy a native folded state N, or an unfolded / denatured state U. The concentration of proteins in N and U states can be related with the equilibrium constant K_{eq}:

(Equation 1)

$$K_{eq} = \frac{[U]}{[N]} = \frac{\alpha}{1 - \alpha}$$

where α is the fraction of total molecules in state U, which is further defined as:

(Equation 2)

$$\alpha = \frac{[U]}{[U] + [N]}$$

The Gibbs free energy ΔG of a system is a measure of the stability of the system's current state relative to other possible states it could occupy. The free energy is a function of the enthalpy ΔH, the entropy ΔS and temperature T of the system defined as:

(Equation 3)

$$\Delta G = \Delta H - T\Delta S$$

The Gibbs free energy is additionally related back to the equilibrium constant and the Boltzmann constant k_B with:

(Equation 4)

$$\Delta G = -k_B T \cdot \ln K_{eq}$$

From Equation 3 and Equation 4, the equilibrium constant can be expressed as a function of temperature:

(Equation 5)

$$\ln K_{eq} = \frac{-\Delta G}{k_B T} = \frac{-\Delta H}{k_B} \cdot \frac{1}{T} + \frac{\Delta S}{k_B}$$

The relationship between $1/T$ and $\ln K_{eq}$ is linear, with slope $m = -\Delta H/k_B$ and y-intercept $\Delta S/k_B$. This is known as a Van 't Hoff plot.[11] Experimentally, this is accomplished by changing the temperature of a protein sample and determining the equilibrium constant for many points, and then applying a linear fit. In this way, we can graphically relate temperature and protein folding as well as calculate the change in enthalpy and entropy of this process (*Figure 2.4*).

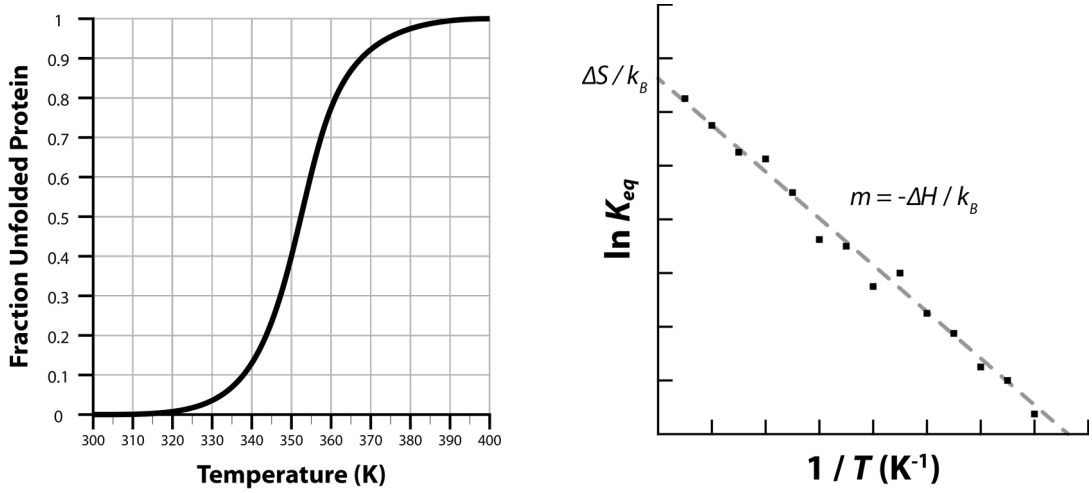


Figure 2.4: Protein denaturation and Van 't Hoff plot. The fraction of the protein that is in an unfolded state increases with increasing temperature. The ratio of unfolded and folded protein yields the equilibrium constant. By graphing the relationship of temperature and the equilibrium constant, it is possible to determine the ΔH and ΔS of a process.

For most proteins, determining the relative fractions of native and unfolded molecules is tricky. In the case of GFP, we have an extremely accessible reporter of correctly folded protein – its fluorescence!

2.4 GFP Structure and Fluorescence

The relationship between protein structure and function is easily observable in GFP. This fluorescent protein has two major features: a barrel formed of anti-parallel β-sheets, aptly named a "beta barrel"; and a chromophore in the center of the barrel that forms autocatalytically from three amino acid side-chains ($Serine_{65}$, $Tyrosine_{66}$ and $Glycine_{67}$). While the chromophore is directly responsible for the protein's fluorescence, the chromophore cannot form unless the beta barrel has

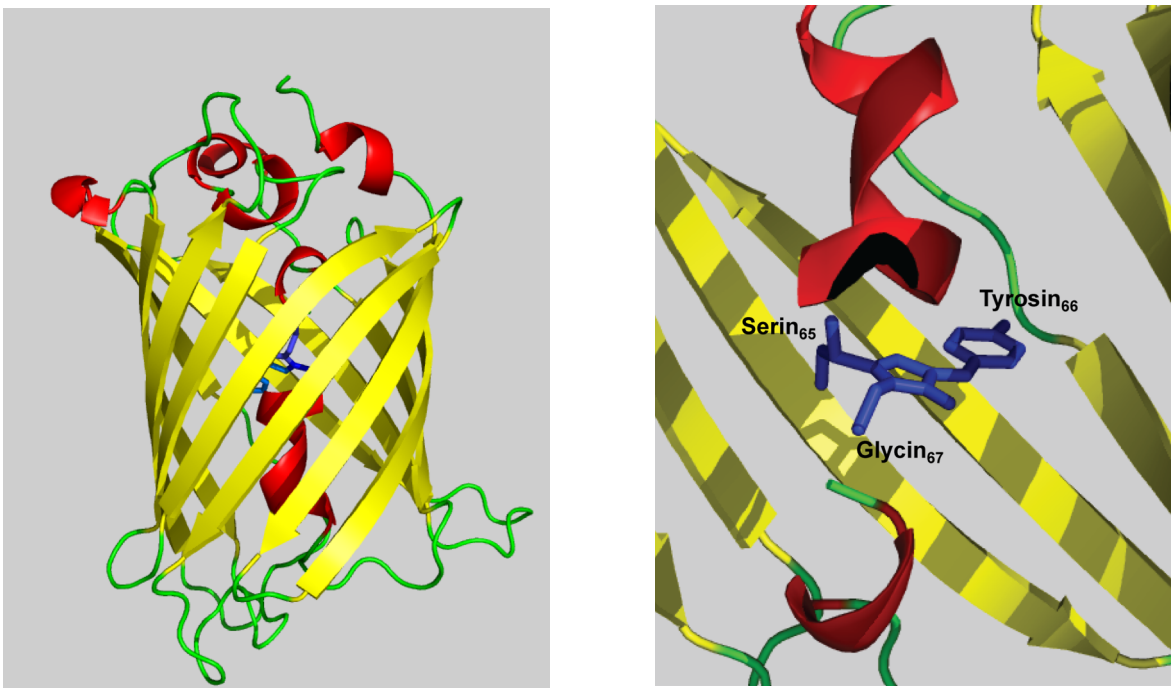


Figure 2.5: Structure of wildtype GFP from *A. victoria* (PBD ID: 1GFL). The complete view of GFP (left) and a magnified view of the central chromophore (right). The beta barrel structure (yellow) scaffolds and protects the chromophore (blue), which forms autocatalytically from three key amino acids that project from a central α-helix (red). The flexible linker regions (green) hinder the intrusion of other molecules.[12]

folded correctly to bring the amino acids in proximity. Moreover, the beta barrel is a very stable structure that is able to protect the chromophore, and the flexible loops that extend from the barrel edges protect the chromophore from other molecules that could interfere with fluorescence (*Figure 2.5*). Hence, each structural element of GFP plays a crucial role in its fluorescence.[12]

So-called wildtype GFP (wtGFP) that is found naturally in *A. victoria* possesses some attributes that are non-ideal for laboratory research. For example, although the chromophore forms autocatalytically, it matures slowly and photobleaches quickly. The protein's excitation spectrum also possesses two peaks, with only the minor peak contributing to fluorescence. Since its discovery, several key point mutations have been introduced to the sequence of GFP that confer increased stability and improved fluorescent properties. For example, the first major improvement to the protein was a single S65T mutation of the chromophore that generally improved its fluorescent characteristics and specifically altered its excitation spectrum to a single major peak at 488 nm. An additional F64L point mutation in the scaffold improved folding efficiency at 37°C, and this variant was termed enhanced GFP (EGFP).[3] More recently, a series of mutations to further enhance the protein's folding and maturation were introduced. This variant – termed superfolder GFP (sfGFP)[3][13] – is the GFP variant that we will use in this practical course. These improved variants of GFP have been extensively studied, and *Figure 2.6* shows a comparison of the fluorescent properties of wtGFP and sfGFP.

Additionally, key mutations may further effect the fluorescent properties of GFP. Just as certain key mutations can impact the stability of the protein, its excitation- and emission- spectra

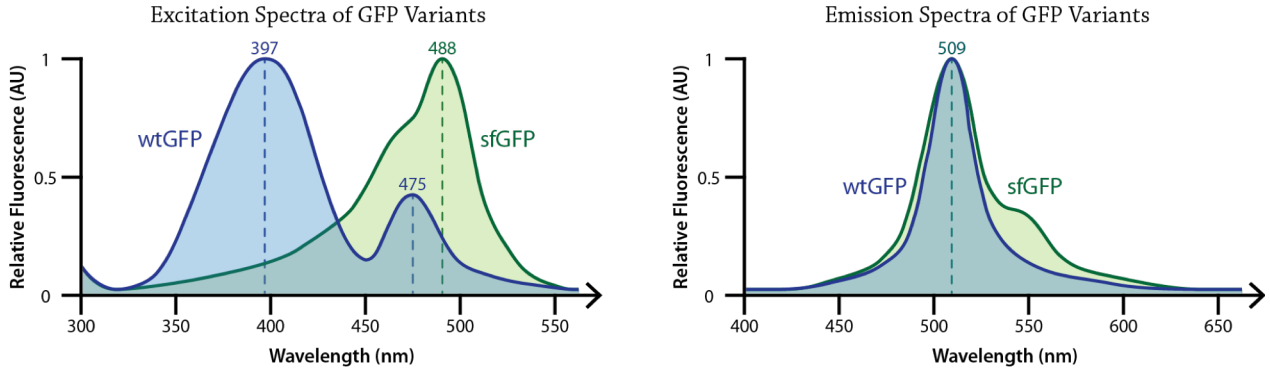


Figure 2.6: Excitation and emission spectra of wtGFP and sfGFP. The mutations present in sfGFP alter the excitation spectrum from a major peak at 397 nm and a minor peak at 475 nm to a single major peak at 488 nm with a minor shoulder. The emission spectrum is not dramatically altered, and still retains a major peak at 509 nm.

can likewise be significantly shifted with strategic mutations. For example, blue fluorescent protein (BFP) contains a Y66H mutation, which replaces tyrosine$_{66}$ in the chromophore with a histidine residue. The chromophore still forms autocatalytically, but its fluorescent properties have been altered such that it now has a broad excitation peak with a maximum at 380 nm and an emission maximum at 448 nm.[3] Mutations outside of the chromophore can also shift the fluorescence spectra. For example, yellow fluorescent protein (YFP) contains a critical mutation T203Y that causes Pi-Pi stacking of the chromophore with the introduced tyrosine residue. This results in a red-shift in both the excitation and emission spectra to respectively 514 nm and 527 nm.[14]

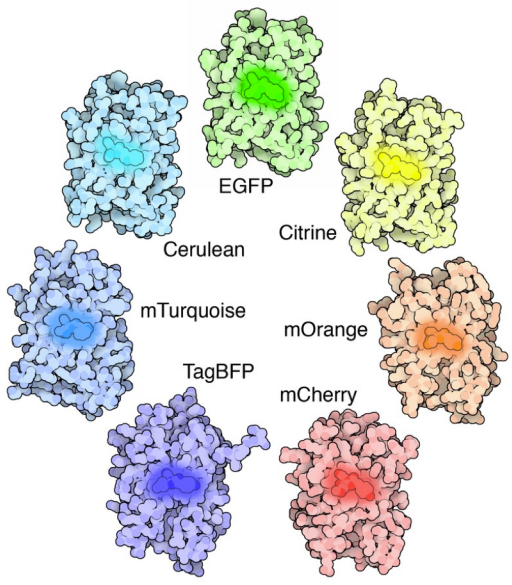


Figure 2.7: A palette of fluorescent proteins. A small handful of naturally-occurring fluorescent proteins have undergone rigorous engineering, yielding variants with improved fluorescent properties and shifted spectra. Most proteins in the blue spectrum are based on GFP, and most in the red spectrum on DsRed.[3][17]

Interestingly, GFP is not the only naturally-found fluorescent protein that has undergone significant engineering. A protein termed DsRed was isolated from the *Discosoma* genus of coral, and it fluoresces in the red spectrum with an excitation maximum of 558 nm and an emission maximum of 583 nm. The structure of DsRed is very similar to that of GFP, specifically its autocatalytically-formed chromophore embedded in a beta barrel. From this fluorescent protein, many variants in with improved properties or shifted spectra in the red range have been developed.[15][16] Meanwhile, most fluorescent proteins in the purple to green range are derived from GFP.[3][16] Modern researchers now have a colorful palette of fluorescent proteins to choose from (*Figure 2.7*).

How is this useful for understanding protein folding in the practical course? Since GFP can only fluoresce if it is correctly folded, we can use its fluorescence as a reporter for the folded state of the protein! Specifically, we can calculate the equilibrium constant by first calculating the concentration of (correctly folded and thus fluorescent) GFP.

The Beer-Lambert law[18] relates the absorption of light to the properties of the material through which the light travels:

(Equation 6)

$$A = \varepsilon \cdot c \cdot l$$

with
A: absorption at a wavelength λ (no units)
ε: molar extinction coefficient at a wavelength λ ($M^{-1} \cdot cm^{-1}$)
c: concentration of the material (M)
l: pathlength (cm)

The **extinction coefficient** for a protein can be determined experimentally, or approximated with algorithmic software. It also depends on the specific wavelength λ of the measurement. In molecular biology, this law is employed frequently for determining the concentration of DNA or protein in a sample. In the case of sfGFP, we will use its absorption at 488 nm, which is also its fluorescent excitation maximum.

Measuring the absorption at 488 nm will inform us of the concentration of folded sfGFP in a sample. At higher temperatures, we would expect that a fraction of the sfGFP in solution denatures, thereby destabilizing the chromophore and decreasing the absorption at 488 nm. Returning to Equations 1 – 5, we then use the concentrations of folded and unfolded sfGFP in solution to calculate K_{eq} for a given temperature. By performing this simple measurement and subsequent calculations to create a Van 't Hoff plot over a range of temperatures, we can investigate the relationship between protein stability and temperature.

3 | Molecular Biology

Now that we have reviewed some of the core principles of physics that are essential for this practical course, the next question is: how will we acquire the GFP? It is possible to isolate the protein from the source organism that natively produces it, but for several reasons, this is not ideal. For one, the jellyfish *A. victoria* is not readily accessible for research. Additionally, the wildtype GFP has poor stability, maturation and fluorescence compared to several engineered variants. Therefore, it is both easier and ultimately more useful to produce our own GFP! This section contains an overview of the principles of molecular biology that we will use, starting with a review of how proteins are produced in cells.

3.1 From Genes to Proteins

The Central Dogma of molecular biology essentially states that DNA makes RNA, and RNA makes protein. While recent advances have shown that this notion is heavily simplified and leaves out critical processes, it is still valid, especially for gaining a fundamental understanding of the process of protein production.

An overview of the Central Dogma is shown in *Figure 3.1*. In cells, DNA carries the genes that code for proteins. During cell division, each strand of the DNA double-helix splits and is replicated once by DNA polymerase, thereby giving each daughter cell a copy of the genome. In order to make proteins, individual genes are transcribed into mRNA by RNA polymerase, and the RNA is then translated into protein by ribosomes.[20]

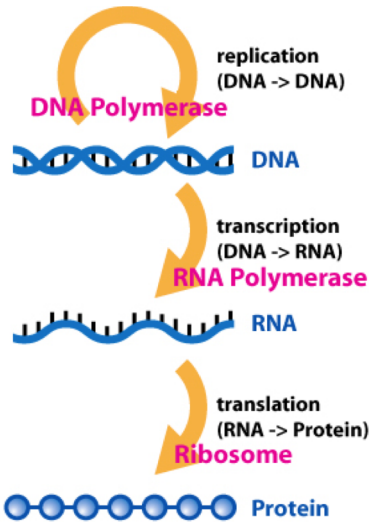


Figure 3.1: The Central Dogma in its simplest form.[19]

In this model, genomic DNA is the critical starting information. A DNA construct serves as an easily-programmable blueprint for proteins production, which can be manipulated using modern molecular biology techniques. Indeed, in this practical course we will design and build a DNA construct from which our protein will be created.

3.2 Plasmids and DNA Manipulation

DNA exists in many forms. Genomic DNA – which is often millions of basepairs long – contains a multitude of genes and regulatory regions. In this practical course, we will handle something much simpler: plasmids.

Plasmids are relatively small, circular DNA molecules that are separate from genomic DNA. They are replicated and regulated independently, and are readily transferable to bacteria cells.[21] Unlike genomic DNA – which contains thousands of genes and regulatory sequences – plasmids are much simpler. Plasmids used as vectors for protein expression in bacteria contain several essential components: an origin of replication, a selection marker, the expression cassette for the gene of interest (in this case, the gene for GFP), and the elements of regulation and expression.[22] The components are summarized below, and a schematic of a hypothetical cloning vector map is shown in *Figure 3.2*.

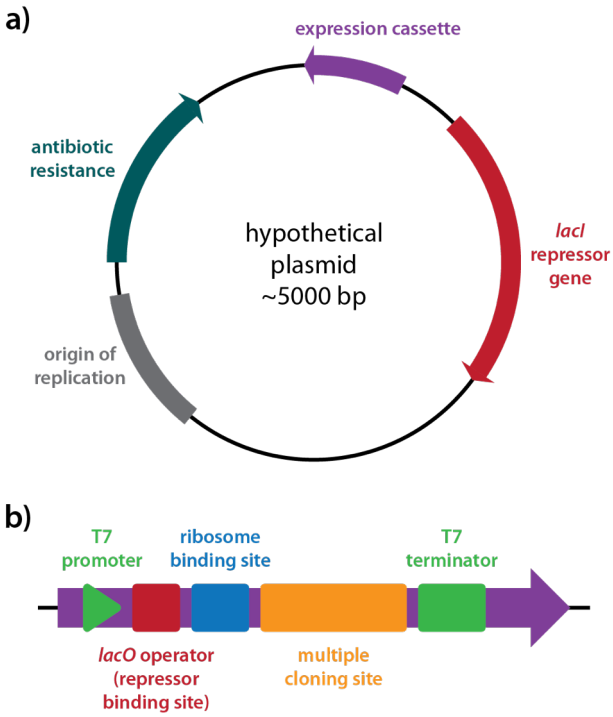


Figure 3.2: Vector map of the core elements of an expression plasmid. a) Overview showing the major regions in the plasmid. b) Detailed view of the expression cassette.

The origin of replication is a region in the plasmid that (as the name implies) is the starting point of DNA replication. This is essential for passing on a copy of the plasmid to daughter cells during cell division.[22]

A selection marker is a gene contained in the plasmid that confers some kind of survival advantage to the organisms that possess it. A standard marker for bacteria is a gene for **antibiotic resistance**. This means that a bacterium that contains the plasmid will also produce a protein that can break down a specific antibiotic, and therefore allow the bacterium to survive and proliferate in selective growth medium. On the other hand, bacteria that do not possess the plasmid will die from the antibiotic. Since bacteria have a natural tendency to eject a foreign plasmid if there is no survival pressure to keep it, this ensures that the bacterial culture produces a high density of bacteria with the plasmid, and therefore higher yields of the protein of interest.[22]

The expression cassette is a stretch of DNA with a combination of elements that are necessary to initiate and regulate expression of the protein of interest. First and foremost, the **multiple cloning site (MCS)** contains many short DNA sequences that are recognized by certain **restriction enzymes**. These enzymes cut DNA when they encounter a specific DNA sequence. Notably, there are over 300 commercially available restriction enzymes, each with specific DNA recognition sequences. The reaction produces sticky overhangs of single-stranded bases, which hybridize to their complementary sequence with very high off-rates due to the short sequence length (usually 2 – 8 nucleotides long). Therefore, the DNA ends are unstable, and DNA fragments can be separated from the vector via gel electrophoresis. The isolated vector can then be combined with a new piece of

DNA that possesses the same sticky overhang sequences and the gene of interest. The digested vector can be ligated to these new pieces of DNA via reaction with **DNA ligase**. This allows molecular biologists to move genes into new vectors, or rearrange segments of their gene of interest at will. A graphical representation of the restriction enzyme cloning strategy is shown in *Figure 3.3*. **The T7 promoter and terminator sequences** are necessary for the binding and release of T7 RNA polymerase. This polymerase comes from the T7 bacteriophage, and is normally not found in *E. coli*. However, the strains of bacteria used for protein expression often have extra genes inserted into their genomic DNA that improve their qualities as host organisms, including a gene that produces the enzyme T7 RNA polymerase. This polymerase transcribes the DNA sequence into RNA, which can then be translated by ribosomes into a protein. However, ribosomes need their own recognition sequence, hence the presence of the **ribosome binding site** upstream of the gene of interest.[22]

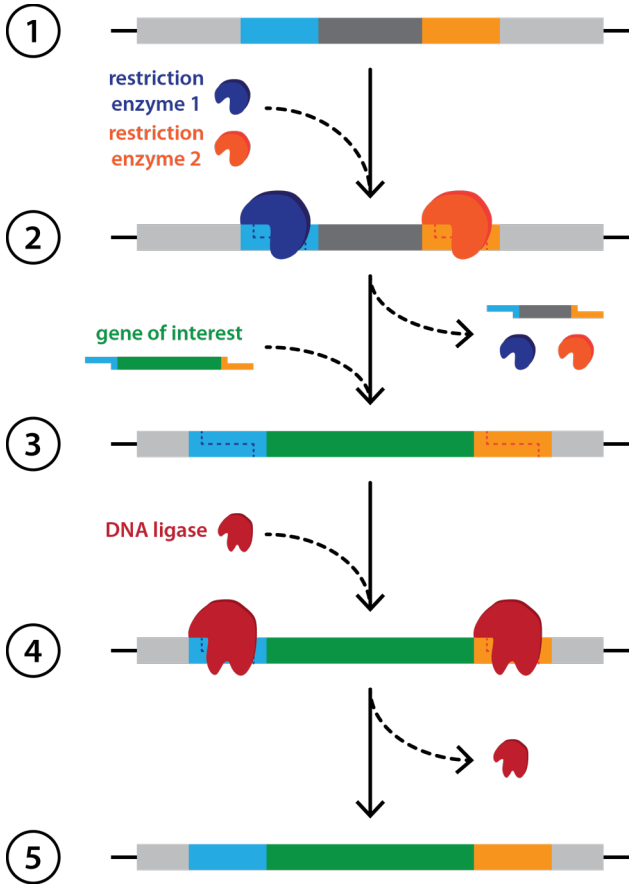


Figure 3.3: Workflow of restriction enzyme subcloning within the MCS of a plasmid. (1) The starting vector contains two restriction sites. (2) Digestion with the corresponding enzymes separates the vector from an unused portion of the MCS. (3) The digested vector is separated from the DNA fragment and restriction enzymes, and a new fragment containing the gene of interest and matching digested restriction sites is introduced. (4) DNA ligase covalently ligates the vector and the gene sequence, producing a stable vector that contains the gene of interest. (5) This new plasmid can now be transformed into bacteria for protein expression.

The elements of regulation and expression include the ***lacO* operon sequence** in the expression cassette as well as the ***lacI* repressor gene**. The repressor gene produces a protein that binds to the operon sequence in the plasmid and prevents the production of the gene of interest by physically blocking T7 RNA polymerase from transcribing the gene. However, the repressor is released from the *lacO* operon sequence if an inducer molecule is introduced to the bacterium. Therefore, it is possible to turn the expression of the gene of interest on or off depending on the contents of the bacteria growth medium.[23] This system to regulate the protein of interest is called **inducible expression**. On the other hand, the *lacI* repressor protein is always expressed, which is called **constitutive expression**.[24] A graphical representation of the *lac* regulation system often used in plasmids is shown in *Figure 3.4*.

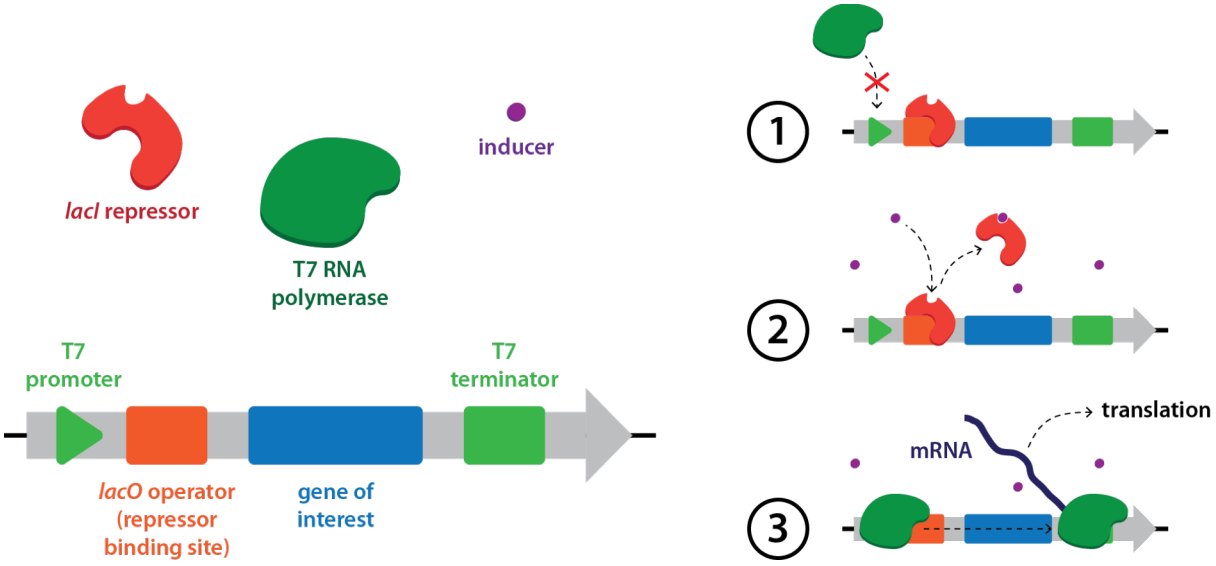


Figure 3.4: The *lac* system of protein expression regulation. (1) The *lacI* repressor binds to a consensus *lacO* sequence in the MCS, which prevents T7 RNA polymerase from binding to its promoter and transcribing the gene of interest. (2) An inducer molecule binds to the repressor and releases it from the DNA. (3) Now T7 RNA polymerase can transcribe the gene of interest into mRNA, which is then translated into protein by ribosomes.

3.3 Protein Expression and Purification

The production of the protein of interest occurs in the cells of a host organism. For relatively simple proteins, bacteria are often chosen as such a host. A primary advantage of using bacteria over other host organisms are that they grow quickly and cheaply. Moreover, it is a straightforward process to transform them with the cloning vector and extract their cell contents. However, not all proteins can be easily expressed in such a minimal organism as a bacterium. Other host organisms – such as yeast, fruit flies, *C. elegans* worms, or insects – may be used instead.[25]

Fortunately for us, GFP is extremely easy to reliably produce in bacteria (which again is one of the reasons why it is so popular for research purposes). Aside from the gene of interest, the components in our plasmid have been specifically tailored for expression in a modified strain of *E. coli*. This strain is a fast-growing bacterium whose genomic DNA is enhanced with the T7 expression system, and it will efficiently overexpress our protein of interest. We will use the *lac* system of protein expression to regulate when our bacteria go into production mode.[25]

How is this complicated *lac* regulation of protein production useful? Wouldn't it be better for the host bacteria to always produce our protein, so that we can acquire more of it? It turns out that overproduction of the protein of interest is often harmful for the bacteria. In most cases, the protein of interest remains in the host cells. If the protein is a toxin or an enzyme, it can do a lot of damage to the host organisms. Therefore, the strategy of inducible expression is standard for most proteins, even relatively harmless ones like GFP. The bacterial culture is allowed to grow to a certain

density, and then protein production is induced by adding an inducer that frees the repressor from the operon sequence (*Figure 3.5*).[23][24]

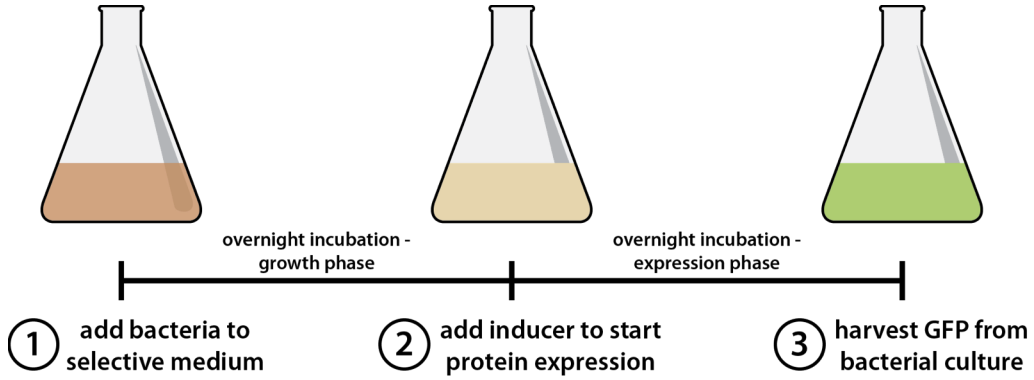


Figure 3.5: Bacterial culture preparation for protein expression. (1) A small sample of bacteria containing a plasmid with the gene for GFP is added to antibiotic expression growth medium. (2) The culture is allowed to grow overnight until it reaches an optimal cell density. At that point, the inducer molecule is added to the medium, which initiates production of GFP. (3) After further incubation with the inducer present, the culture is harvested and GFP is purified from the cell lysate.

Protein purification functions on the principle that no two types of protein are exactly alike. Therefore, the differences in their size, charge, stability, binding affinity, etc., can be exploited to separate the extracted cell contents.

A standard means of separating biological components is **affinity chromatography.** In this strategy, the unpurified contents of a cell are pumped through a purification column that is packed with a special resin. This resin can have different properties depending on what kinds of molecules need to be purified. The column will bind a certain subset of molecules, while the remainder flow out. The bound molecules are then eluted by introducing a buffer that competes for binding to the column. As the sample flows out of the column, it passes through a UV/Vis spectrophotometer, which allows us to determine which fraction of the eluate contains our protein of interest. Hence, it is possible to separate the contents of the cell according to their biophysical or biochemical characteristics.[26]

If an unmodified protein is to be extracted from its native organism (for example, wildtype GFP from *A. victoria*), there is often no straightforward strategy to cleanly isolate such a protein, as it might strongly resemble many other proteins that are also expressed. However, in our plasmid where the gene of interest is under our control, we can add features to a protein that will improve this process. Specifically, affinity tags are often added to the ends of proteins that enable specific, quick purification.

An **affinity tag** is a small peptide sequence added to a protein of interest, and it often has unique properties that can be exploited for purification by affinity chromatography. For example, the *Strep*-tag II peptide will bind with high affinity to *Strep*-Tactin. Therefore, if a *Strep*-tag II is added to the protein of interest, it can be easily purified with a column whose resin contains *Strep*-Tactin. Similarly, the FLAG-tag epitope binds with high affinity to its corresponding antibody. In addition to

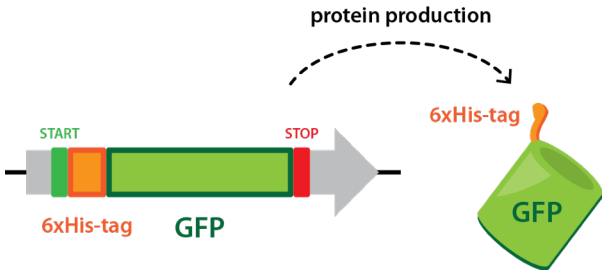


Figure 3.6: Protein expressed with a 6xHis-tag. Within the MCS, a gene (for example, GFP) is modified with the sequence for the purification tag. The resulting peptide is expressed from the START codon to the STOP codon with both components covalently linked as a single protein.

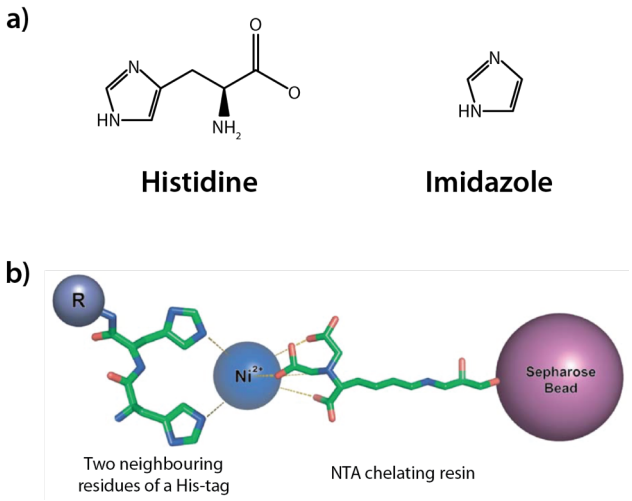


Figure 3.7: Structure of histidine and formation of the coordination compound. a) The side-group of histidine is imidazole, and the two molecules are both able to bind to the Ni-NTA resin. b) The NTA resin chelates Ni^{2+} ions, which together form a coordination compound with histidine residues in the 6xHis-tag.[29] The protein is eluted with the addition of competing imidazole.

short peptide tags, larger protein domains – such as Glutathione-S-transferase or Maltose binding protein – can similarly be used for purification.[27]

One of the most frequently used affinity tags – which we will also use in this practical course – is a **6xHis-tag**. It consists of a series of six histidine amino acid residues on one of the termini of the protein of interest (*Figure 3.6*). The corresponding column contains a resin with nitrilotriacetic acid (NTA). This compound chelates nickel ions (Ni^{2+}), effectively immobilizing them in the resin of the column. Hence, this subset of affinity chromatography is known as **immobilized metal ion affinity chromatography (IMAC)**. The immobilized Ni^{2+} forms coordination compounds with the NTA resin and two of the histidine side-groups in the 6xHis-tag. The bound 6xHis-tag is eluted with high concentrations of **imidazole**, which is structurally very similar to histidine and competes with the histidine residues for binding to the Ni-NTA resin (*Figure 3.7*).[28]

The formation of the coordination compound can occur with any histidine residue on the surface of any protein. However, the 6xHis-tag can form up to 3 coordination compounds with the column resin, which gives it a much higher affinity for the column than most untagged proteins. Consequently, with low concentrations of imidazole, contaminating proteins will for the most part flow out of the column. Only with the addition of high concentrations of imidazole does the 6xHis-tag protein elute. The strategy of 6xHis-tag purification by Ni-IMAC is summarized in *Figure 3.8*.

We can design, express, and purify proteins with relative ease using these molecular biology techniques. Although this practical course focuses on GFP, the procedures described here are widely applicable for many genes and proteins. This practical course can therefore be considered a model workflow that you might use for any number of proteins of interest.

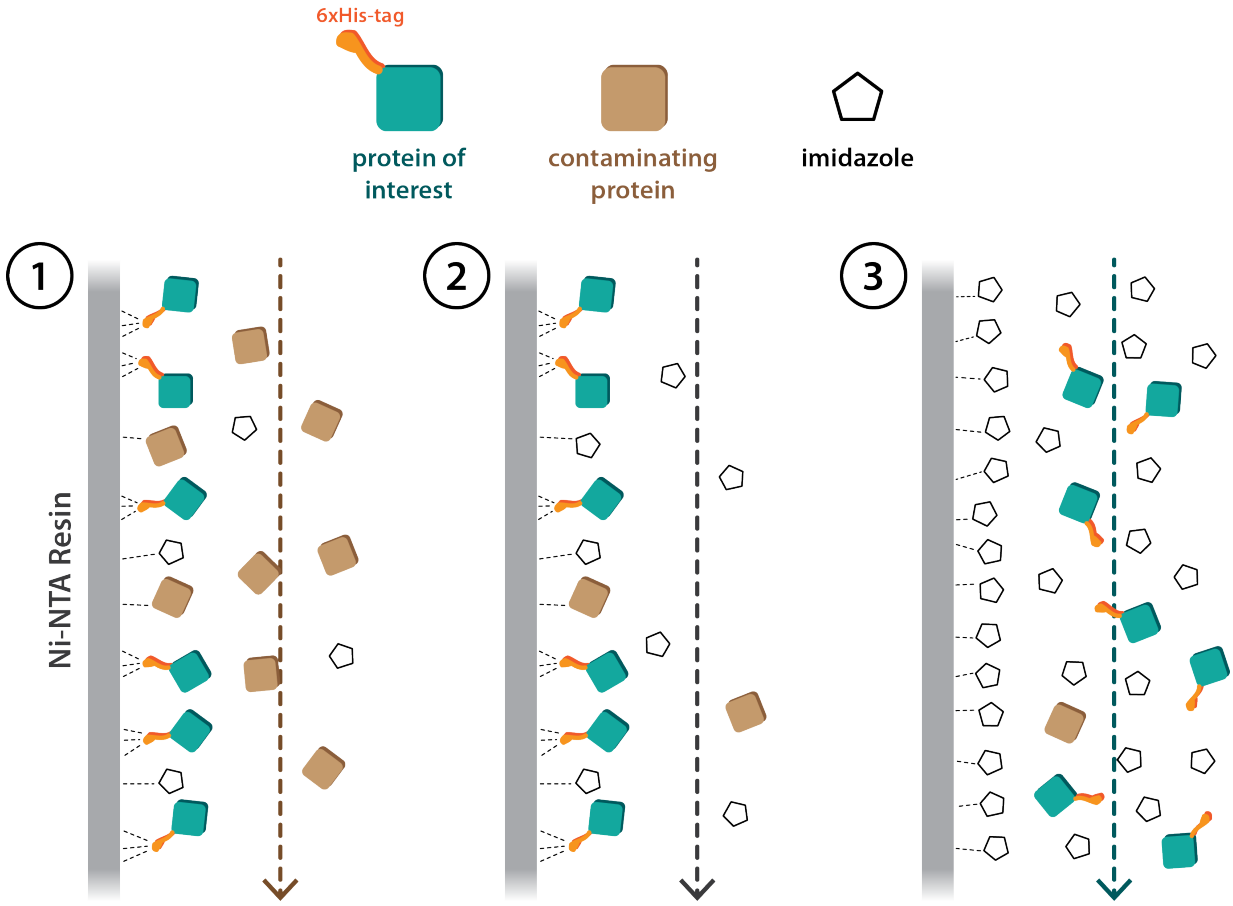


Figure 3.8: Protein purification via 6xHis-tag. (1) The sample containing the protein of interest as well as contaminants from the cell lysate are loaded into the Ni-NTA column. At low concentrations of imidazole, the 6xHis-tag binds strongly to the resin. Some of the contaminating protein binds weakly to the column, but most flows through. (2) The column is washed with low concentrations of imidazole, causing some of the weakly-bound contaminants to unbind from the column and flow through. (3) A buffer with a high concentration of imidazole is applied to the column. The imidazole out-competes all of the protein bound to the column, and the protein of interest as well as any remaining contaminants are eluted. If purification was successful, the fractions contain mostly the protein of interest and relatively little other protein from the cell lysate.

4 | Methods and Techniques

The following section outlines some of the other important techniques used in this practical course, as well as further technical information on some of the techniques already mentioned in previous sections.

4.1 Absorption and Fluorescence Spectroscopy

Most biological compounds – not only fluorescent molecules such as GFP – have predictable absorption spectra. With a known extinction coefficient, the concentration of a molecule can be calculated from the measured absorption at a critical wavelength using the Beer-Lambert law (Equation 6).

Two wavelengths that are commonly used to determine the concentration of nonfluorescent biomolecules in solution are **260 nm** and **280 nm**. Generally speaking, DNA has a peak absorption at 260 nm and proteins have a peak at 280 nm. The source of the absorption of DNA in this wavelength range are the aromatic groups of the nucleobases, which absorb strongly at 260 nm and weakly at 280 nm. Double-stranded DNA forms a predictable double-helix shape at room temperature under a wide range of buffer conditions, and most long DNA molecules have approximately equal ratios of all four nucleobases. Therefore, the concentration of most DNA samples can be determined with the average mass extinction coefficient of ε_{260} = 0.020 $(\mu g/mL)^{-1}$ cm^{-1}. Additionally, the ratio of the absorption at 260 nm and 280 nm can also hint at the ratio of DNA to protein in solution. This is extremely useful for determining (for example) the success of DNA purification from a miniprep.[30]

Determining protein concentration from the absorption spectrum is more challenging, which is caused by several factors. Firstly, the primary sources of absorption at 260 nm and 280 nm are the side chains of only four out of twenty amino acids: the aromatic groups of phenylaline, tyrosine and tryptophan, and the possible disulfide bonds formed by cysteines. Therefore, the absorption of a protein is strongly sequence-dependent, and cannot necessarily be predicted by the sheer number of amino acids it contains. Secondly, the fold of a protein has a huge impact on its absorption. Local interactions between aromatic side chains and neighboring amino acids can impact their absorption, causing a deviation from expected absorption properties. Lastly, the buffer and ambient conditions can have dramatic effects on the structure of a protein or the chemical state of the side chains in many amino acids.[31] Despite advances in protein structure prediction algorithms, it is still impossible to efficiently and correctly predict the final structure of most proteins with only the amino acid sequence and buffer conditions as starting information. Therefore, the absorbance spectrum of a natively folded protein must often be determined experimentally – in contrast to DNA, whose absorption spectrum is reliably calculable based solely on nucleotide sequence.

Despite these difficulties, it *is* possible to predict the molar extinction coefficient of a protein using models and algorithms. Many available online tools, namely ProtParam from ExPASy,[32] can

calculate an estimated extinction coefficient at 280 nm and other useful parameters based on the amino acid sequence of a protein.

For our experiments in the practical course, we will use a **nanodrop** to measure DNA A_{260} and Protein A_{280}. For the GFP Melting curves, we will use a **spectrophotometer** that is capable of simultaneously heating and measuring the absorbance. Values relevant to absorption spectroscopy of DNA and proteins in this practical course are shown in *Table 1* and *Table 2*. The biochemical and biophysical parameters of GFP variants are summarized in *Table 3*. An example of protein parameter calculation from ProtParam is shown in *Figure 4.1*.

Table 1: Absorption Parameters of Purified DNA at 260 nm and 280 nm[30]

ε_{260} (mL μg^{-1} cm^{-1})	ε_{280} (mL μg^{-1} cm^{-1})	A_{260} / A_{280}
0.020	0.011	≈1.8

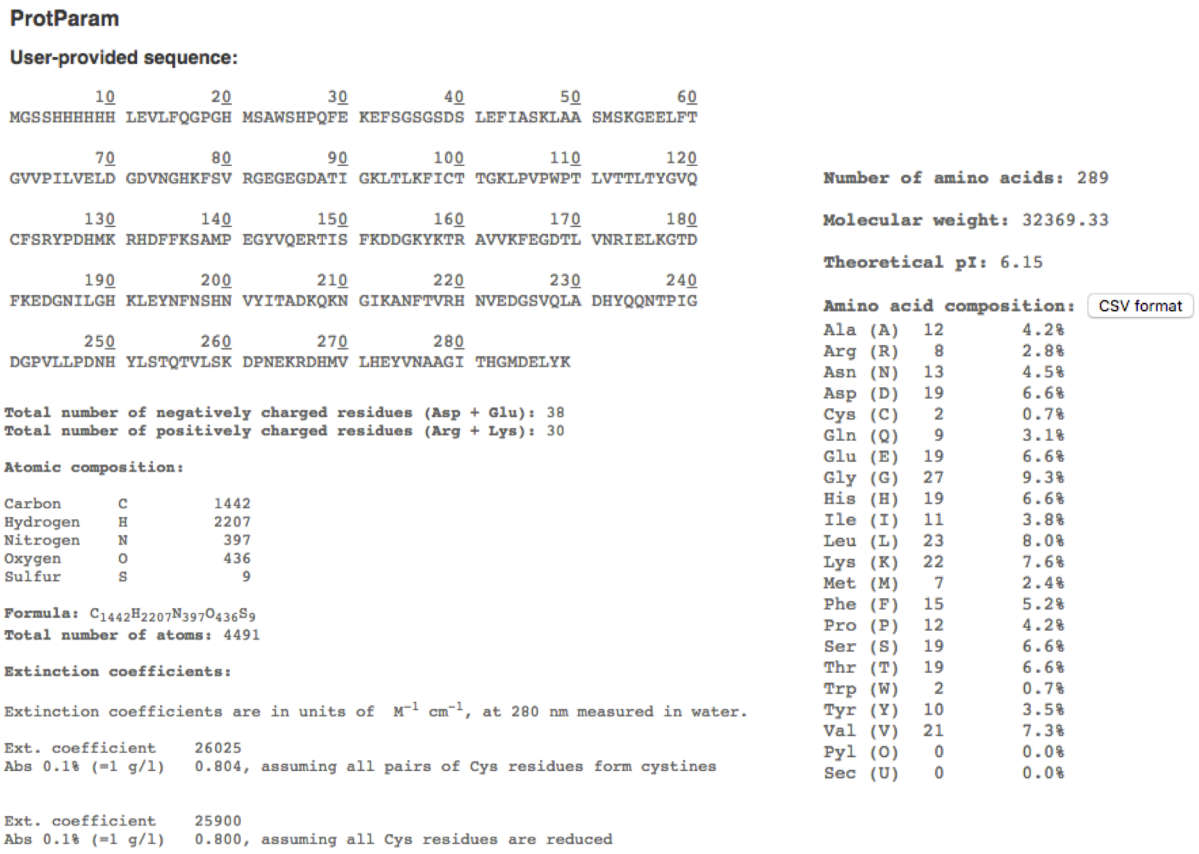
Table 2: Absorption Parameters of Amino Acids at 280 nm and Nearest Local Maximum[31]

Amino Acid	ε_{280} (M^{-1} cm^{-1})	Nearest Local Maximum (nm)	ε_{Max} (M^{-1} cm^{-1})
Cysteine (disulfide)	110	-	-
Phenylalanine	0.7	257	200
Tryptophan	5,559	279	5,600
Tyrosine	1,197	275	1,400

Table 3: Fluorescent Parameters of GFP Variants[3][13][33][34]

Parameter	wtGFP	sfGFP	sfGFP (Praktikum)
Number of Amino Acids	238	238	289
Molar Mass (g/mol or Da)	26766	26792	32370
Excitation Maxima (nm)	397 and 475	488	488
Emission Maximum (nm)	509	509	509
Quantum Yield	0.79	0.65	0.65
ε_{280} (M^{-1} cm^{-1})	22,015*	20,400*	25,900*
ε_{397} (M^{-1} cm^{-1})	25,000	-	-
ε_{475} (M^{-1} cm^{-1})	9,500	-	-
ε_{488} (M^{-1} cm^{-1})	-	83,300	83,300

* calculated with ProtParam http://web.expasy.org/cgi-bin/protparam/protparam



ProtParam

User-provided sequence:

```
        10         20         30         40         50         60
MGSSHHHHHH LEVLFQGPGH MSAWSHPQFE KEFSGSGSDS LEFIASKLAA SMSKGEELFT

        70         80         90        100        110        120
GVVPILVELD GDVNGHKFSV RGEGEGDATI GKLTLKFICT TGKLPVPWPT LVTTLTYGVQ

       130        140        150        160        170        180
CFSRYPDHMK RHDFFKSAMP EGYVQERTIS FKDDGKYKTR AVVKFEGDTL VNRIELKGTD

       190        200        210        220        230        240
FKEDGNILGH KLEYNFNSHN VYITADKQKN GIKANFTVRH NVEDGSVQLA DHYQQNTPIG

       250        260        270        280
DGPVLLPDNH YLSTQTVLSK DPNEKRDHMV LHEYVNAAGI THGMDELYK
```

Total number of negatively charged residues (Asp + Glu): 38
Total number of positively charged residues (Arg + Lys): 30

Atomic composition:

Carbon	C	1442
Hydrogen	H	2207
Nitrogen	N	397
Oxygen	O	436
Sulfur	S	9

Formula: $C_{1442}H_{2207}N_{397}O_{436}S_9$
Total number of atoms: 4491

Extinction coefficients:

Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient	26025
Abs 0.1% (=1 g/l)	0.804, assuming all pairs of Cys residues form cystines
Ext. coefficient	25900
Abs 0.1% (=1 g/l)	0.800, assuming all Cys residues are reduced

Number of amino acids: 289

Molecular weight: 32369.33

Theoretical pI: 6.15

Amino acid composition: CSV format

Ala (A)	12	4.2%
Arg (R)	8	2.8%
Asn (N)	13	4.5%
Asp (D)	19	6.6%
Cys (C)	2	0.7%
Gln (Q)	9	3.1%
Glu (E)	19	6.6%
Gly (G)	27	9.3%
His (H)	19	6.6%
Ile (I)	11	3.8%
Leu (L)	23	8.0%
Lys (K)	22	7.6%
Met (M)	7	2.4%
Phe (F)	15	5.2%
Pro (P)	12	4.2%
Ser (S)	19	6.6%
Thr (T)	19	6.6%
Trp (W)	2	0.7%
Tyr (Y)	10	3.5%
Val (V)	21	7.3%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%

Figure 4.1: Examples of the calculation output from ProtParam. The amino acid sequence of the sfGFP construct that we will use in this practical course has been entered into the online calculator. ProtParam's algorithms produce useful estimated parameters.[34]

4.2 Bacterial Transformation

Transformation is the process by which a plasmid is incorporated into bacteria. The cells we will use in this practical course have been treated beforehand with a series of solutions that make them susceptible to absorbing foreign DNA (i.e., chemical competence). The buffer usually contains a high concentration of divalent salts, which makes the outer membrane of the cells unusually permeable. Aliquots of the competent cells are stored frozen at -80°C and are thawed only once they are used for transformation.

As the cells thaw on ice, a **plasmid** is introduced and enters the cells through the permeabilized membranes. Next, the cells are quickly heat-shocked and transferred to ice. Some of the cells have now successfully incorporated a copy of the plasmid.

The bacteria are incubated in nonselective medium for up to 1 hour. Cells that have a copy of the plasmid begin production of the protein for antibiotic resistance, and those without the plasmid simply grow. Following incubation, the cells are plated onto an agar growth plate that contains the **selective antibiotic**. Only the transformed cells with a copy of the plasmid can survive in these conditions. After overnight incubation, the plate is dotted with bacterial colonies that each

stemmed from one transformed bacterial cell. Samples from individual colonies can then be selected and grown in liquid medium.[35]

4.3 Centrifugation

Centrifugation is a simple but versatile technique in which an **angular momentum** is applied to a sample, thereby inducing an exponential increase in the effective gravitational force. It is mainly used for separating the contents of a liquid suspension according to molecular characteristics, such as size or density. Additionally, filtration columns often take advantage of centrifugation to quickly pull the contents through the filter.[36] We will use centrifugation to separate living bacteria cells from their growth medium; separate the soluble fraction containing our protein of interest from debris after cell lysis; and quickly purify DNA with a Miniprep kit.

4.4 DNA Isolation by Miniprep

Much like harnessing bacteria as living factories to produce our protein of interest, we also use bacteria to replicate plasmids. A single bacterium with a plasmid of interest can multiply and produce millions of daughter cells overnight, each with a copy of the plasmid. Extracting plasmids from bacteria is easily accomplished with the assistance of a Miniprep kit.

First, a small culture of transformed bacteria grows overnight in selective liquid medium, and the next day is harvested and processed with the kit. The cells are **lysed** – i.e., exposed to buffer that strongly disrupts the outer cell membrane, essentially bursting the cell and causing the contents to flow into the medium. The lysate from the burst cells is exposed to a series of buffers that degrade RNA, and cause precipitation of protein, lipids and genomic DNA. Centrifugation is used to separate the precipitate and the soluble fraction (which contains the plasmid). The soluble fraction is run through a filtration column that captures DNA under specific buffer conditions. The DNA is captured in the column, washed several times, and eluted with water into a fresh tube. The result is a sample of **purified plasmid** that can be used in molecular biology applications, DNA sequencing, biochemical reactions, cell transformation and cloning.[37]

4.5 Enzymatic Alteration of DNA

Subcloning – the process of rearranging sequences of DNA to create a desired plasmid construct – requires enzymatic alteration of DNA, which is easily accomplished through many commercially-available enzyme systems. The enzymatic reactions that we will utilize in the practical course are restriction digestion, ligation, and exponential replication via PCR. Each of these processes is accomplished through different enzymes that also have their own optimized buffer and incubation conditions.

4.6 Gel Electrophoresis

As with protein purification, we can exploit the biophysical characteristics of different proteins or DNA strands to separate them in a **gel matrix** in order to analyze and quantify the molecules. In contrast to absorption spectroscopy – which can only tell us the total mass concentration of protein or DNA in a sample – gel electrophoresis enables separation and analysis of individual components. This technique is useful for analyzing DNA samples after **restriction digest** or **PCR** to verify that the desired reactions have occurred. Additionally, it is indispensable after protein purification to check that the protein of interest has indeed been expressed and purified. Countless other applications that are not used in this practical course also benefit from gel electrophoresis, making it an essential tool in molecular biology.

There are many strategies to separate biomolecules, but in this practical course (and also generally speaking), we use the **charge** of molecules to separate them by **size**. Samples are loaded into the pockets of a gel, whose dense matrix hinders the migration of molecules. An **electrical current** is then applied to the matrix such that negatively-charged molecules are pulled towards the positively-charged anode. Due to the sieving nature of the gel matrix, smaller molecules travel faster through the gel pores and larger molecules slower, producing a gradient according to size. In addition to the samples of interest, gels are often loaded with a commercially-available **ladder** to assist with determining the size of the molecules (*Figure 4.2*). After electrophoresis, the gel is visualized and analyzed. This may require additional staining, although modern strategies often employ a gel whose matrix already has the staining molecule to expedite the process.[40]

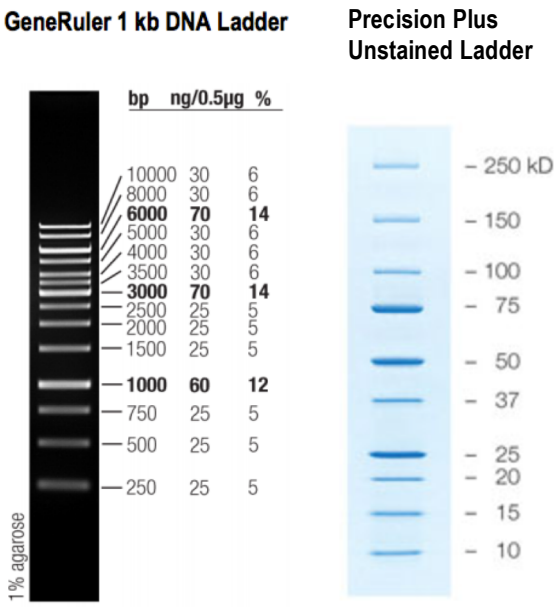


Figure 4.2: Ladder standards for DNA (GeneRuler, Thermo Scientific)[38] and proteins (Precision Unstained, Bio-Rad).[39] Ladders provide a frame of reference for samples of unknown size. Note that typically DNA size is measured in base pairs (bp), while protein size is measured in kilo Daltons (kD or kDa), which is defined as 1 Da = 1 g/mol.

DNA is easily analyzed in this method. Because double-stranded DNA adopts a predictable helical structure and is uniformly negatively-charged from the **sugar-phosphate nucleotide backbone**, gel electrophoresis of DNA in non-denaturing **agarose** is straightforward.[41] An example of such a gel is shown in *Figure 4.3*. DNA may also be easily excised from a gel in a **gel extraction**.

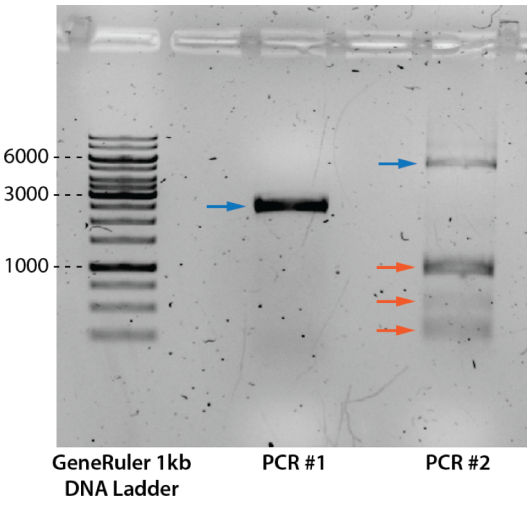


Figure 4.3: Agarose gel analyzing PCR products. The ladder enables determination of the sizes of other DNA molecules in the gel. Desired PCR products are indicated with the blue arrows, by-products from incorrect primer annealing are shown with the orange arrows. PCR #1 consists of primarily of the desired product at 2200 bp, while PCR #2 contains several by-products in addition to the desired product at 5500 bp.

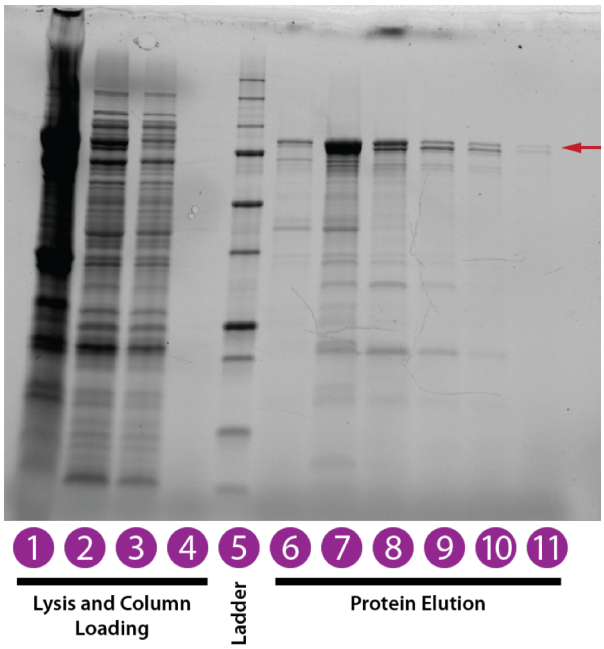


Figure 4.4: SDS-PAGE gel analyzing protein samples from purification of Pfu DNA Polymerase. (1) pellet and (2) supernatant after lysis, (3) column flowthrough, (4) column wash, (5) protein standard ladder, (6) – (11) elution fractions. The major purified product is indicated with the red arrow, and has an estimated mass between 75 – 100 kDa. The polymerase construct has an expected mass of 95 kDa. The optimal fraction that would be selected for experimentation is shown in lane (7), which has the highest concentration of the polymerase.

This is especially useful when a PCR reaction failed to produce only one DNA product, or DNA fragments from a restriction digest must be isolated.

It is likewise possible to separate proteins by size. However, protein is far less uniform than DNA. While the sugar-phosphate backbone of DNA is the primary contributor to its negative charge, the charge of a protein depends on the amino acid side-chains. Consequently, two proteins that are the same size might actually travel very differently in the electric gradient due to a difference in net charge of the molecule. Moreover, natively-folded proteins are often too compact to be sieved by a gel matrix. A solution to both of these complications is to analyze the proteins in denaturing conditions with the assistance of **sodium dodecyl sulfate** (**SDS**).

SDS is an anionic surfactant, meaning it is negatively-charged and contains both hydrophilic and hydrophobic groups. If a protein sample is mixed with SDS and heated to denaturing temperatures, the hydrophobic carbon tails of SDS molecules bind favorably to the exposed hydrophobic core of the protein and stabilizes a rod-like conformation of the protein. The strong negative charge of SDS also imparts a net negative charge to the protein. The negatively-charged protein:SDS complex can now be run on a denaturing gel that likewise contains SDS and a gel matrix made of acrylamide, also known as SDS-polyacrylamide gel electrophoresis (**SDS-PAGE**). Due to the denaturing conditions and interaction of SDS with the protein, the proteins will reliably run according to size.[42] An example of SDS-PAGE analysis of purification of recombinant Pfu DNA Polymerase is shown in *Figure 4.4*.

There exist many variations of agarose and acrylamide gels that are optimal for different molecules. For example, small DNA fragments or natively folded proteins can be analyzed on a non-denaturing PAGE gel, and RNA is often analyzed with urea gels. This diversity of applications makes gel electrophoresis a versatile fundament of modern molecular biology.

4.7 GFP Melting Curves

In order to systemically heat and measure the absorbance of GFP samples, this practical course uses a spectrophotometer that is also capable of heating the sample. We will prepare samples of GFP and measure the absorbance of the samples over a range of temperatures, which will inform us of the fraction of GFP molecules that are **fluorescent** (and therefore correctly folded) or **non-fluorescent** (and therefore denatured) using the Beer-Lambert law (see *Equation 6*). Furthermore, we will be able to calculate the enthalpy and entropy of GFP folding using *Equations 1 – 5*. This will allow us to investigate the thermodynamic processes that govern the folding and stability of GFP, the principles of which may be generalized to all proteins. Refer to *Table 3* for biophysical parameters of GFP that are relevant for these calculations.

4.8 Immobilized Metal Ion Affinity Chromatography (IMAC)

As earlier described, this practical course makes use of a 6xHis-tag and corresponding Ni-NTA column in **immobilized metal ion affinity chromatography** (**IMAC**). As molecules flow out of the column, they are measured in real time by a **UV/Vis spectrophotometer**. This allows us to monitor the purification and observe which eluted fractions likely contain the protein of interest based on the absorption intensity in the UV/Vis spectrum. In the case of GFP, it is actually possible to see with the naked eye which fractions contain the protein, as they will be bright green! However, most

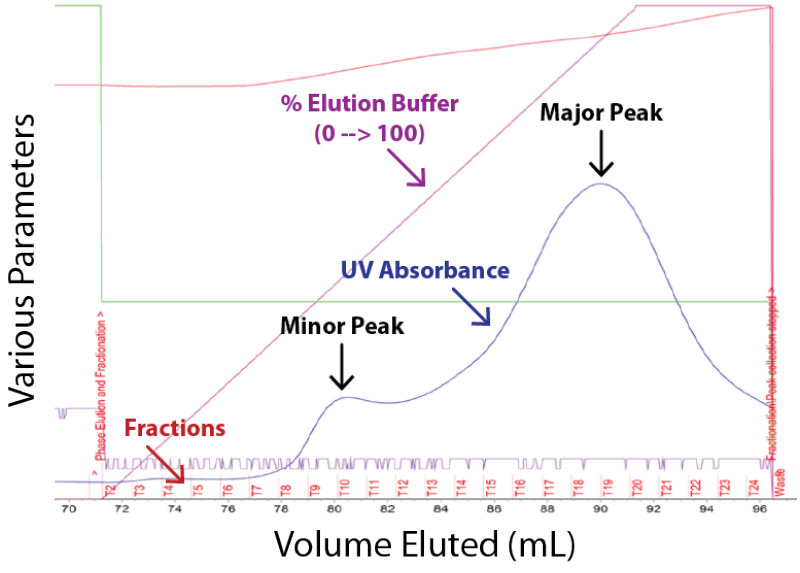


Figure 4.5: Chromatograph from Ni-IMAC purification. The UV absorption of the sample that flows out from the column (blue) indicates the relative concentration of protein in each fraction. At a lower percentage of elution buffer (purple) – and therefore a lower concentration of imidazole – a minor peak is observed. This is mostly contaminating protein that was non-specifically bound to the column. Later, a second major peak is observed at higher imidazole concentrations. These fractions (red) contain high concentrations and purity of the protein of interest.

proteins are colorless and do not afford such an advantage, and therefore must be detected with chromatography. It is also important to note that the absorption at 260 nm and 280 nm can only report the total amount of absorbing material in solution, and gives no indication of purity. Therefore, it is essential to also analyze the elution fractions via gel electrophoresis, which *can* report the relative purity of the protein of interest. An example chromatograph from Ni-NTA IMAC is shown in *Figure 4.5*.

4.9 Polymerase Chain Reaction (PCR)

Artificial synthesis of large amounts of DNA is time-consuming and costly. Even with well-developed modern techniques, for many purposes it is impractical to rely only on commercially-available synthesized DNA. On the other hand, if you already have a small amount of DNA of a desired sequence, it is very easy to exponentially replicate it via **polymerase chain reaction** (**PCR**).[43]

During the early stages of cell division, **DNA polymerase** (**DNAP**) replicates the entire genome. The DNA double helix is unzipped and duplicated, producing two copies of the same DNA. PCR works on the same principle, but substitutes some of the elements of *in vivo* replication for ones tailored to an artificial *in vitro* setting. While cells use helicase to unzip double-stranded genomic DNA and make it accessible to DNAP, PCR makes use of **thermal cycling** to melt DNA at high temperatures and re-anneal it at lower temperatures. Moreover, replication of genomic DNA in a cell begins at an origin of replication (which is also an essential component of plasmids). With PCR, DNAP spontaneously replicates DNA if it encounters a double-stranded primer followed by a stretch of single-stranded DNA.

A starting sample of double-stranded template DNA can be a plasmid, genomic DNA, or shorter linear DNA. The buffered reaction mixture also contains a thermostable DNAP, free nucleotides, and short DNA primers that have been specifically designed to anneal to select regions of the template DNA sequence. The template DNA is melted by incubation at high temperatures, usually around 96°C.

The reaction mixture is then cooled, enabling the primers to hybridize to their complementary sequences in the template strands. The temperature of this cooling step depends on the annealing temperature of the primers, which is determined by their DNA sequence and length as well as the ambient buffer conditions. There exist many online calculators that can predict a primer's annealing temperature based on its sequence and buffer conditions. Generally speaking, the annealing temperature is in the range of 65°C.

Finally, the reaction is heated slightly to 72°C – the optimal catalytic temperature for many thermostable polymerases, although replication does occur at lower temperatures. In fact, it is crucial that a small amount of replication occurs during the previous step after primer annealing; the primers are extended slightly and thereby remain hybridized to the template DNA even at 72°C. The duration of the 72°C DNA extension step depends on the length of the DNA to be replicated, but is usually in the range of 1 – 3 minutes.

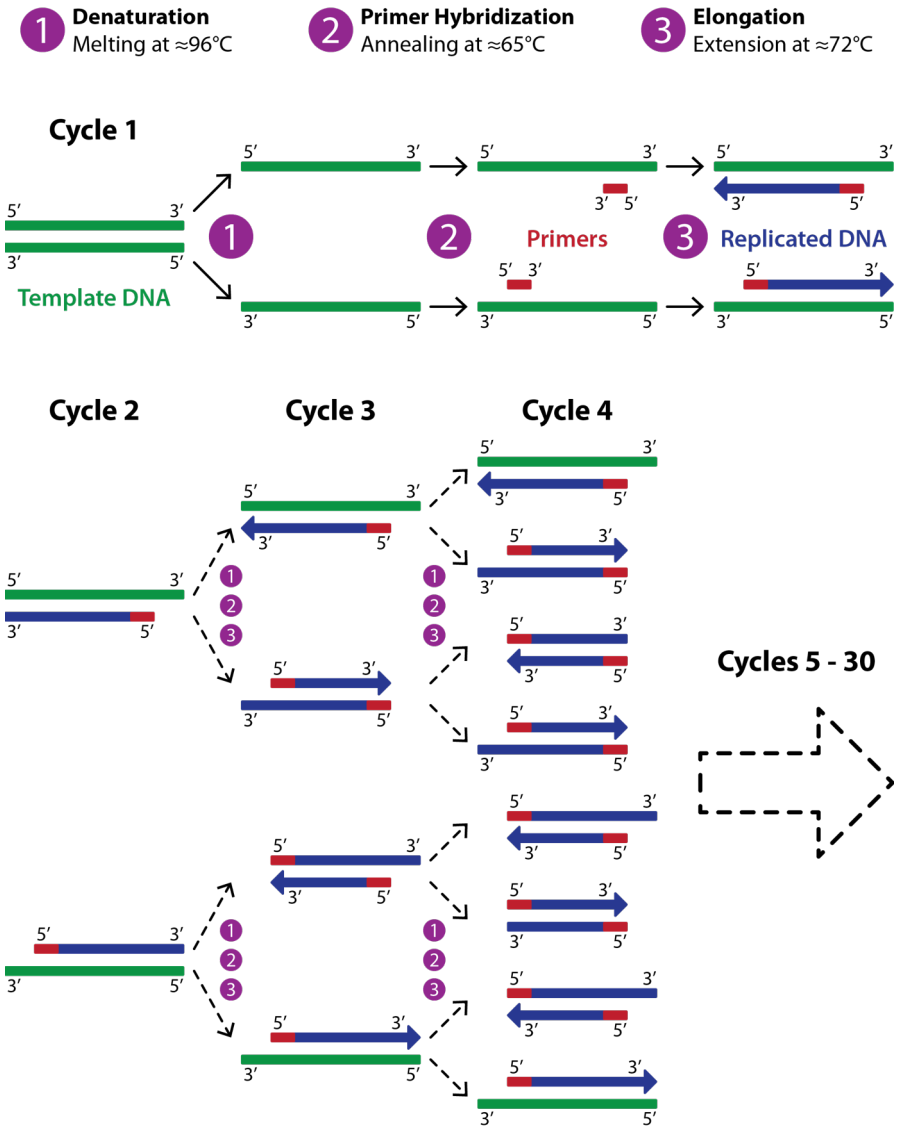


Figure 4.5: Polymerase chain reaction (PCR). Double-stranded template DNA melts, and primers anneal to the complementary sequence in each strand. DNAP replicates each strand, producing two copies for each template DNA molecule. This is repeated for 30 or more cycles, enabling the exponential amplification of a stretch of DNA.

These three heat steps are repeated 30 or more times, yielding an **exponential increase** in DNA of the sequence of interest. *Figure 4.5* shows a graphical summary of PCR, and how large amounts of DNA can be created from a single copy of template DNA. Afterwards, the reaction mixture is analyzed by gel electrophoresis to check that PCR was successful and that only the desired product (rather than a series of byproducts, as shown in *Figure 4.3*) has been amplified.

PCR amplification is additionally capable of DNA sequence modification. With specific primer design strategies, PCR can introduce point mutations, insert a sequence, splice two sequences together, or delete a sequence.[44] This versatile technique has many uses beyond simple DNA amplification, as is explored in this practical course. We will use PCR not only to amplify the GFP gene of interest, but also to add restrictions sites to its termini that will enable ligation into an expression plasmid.

5 | Workflow in Brief

Below is a summary of the workflow that one would use in a "real life" scenario to subclone, express, purify and test GFP. Normally this process occurs over days or even weeks if there are technical problems with subcloning and / or purification. Since the practical course condenses all of these steps into one day with overlapping protocols, it may be helpful to review the workflow in the correct chronological sequence.

Day 1

1. Acquire GFP sequence from commercial source or as a gene insert in a plasmid
2. PCR-amplify GFP gene and add restriction sites
3. Analyze PCR with gel electrophoresis
4. Digest PCR-amplified GFP gene and expression plasmid with restriction enzymes
5. Analyze restriction digests with gel electrophoresis
6. Ligate digested GFP gene and expression plasmid
7. Transform ligated DNA into subcloning bacteria, and grow colonies overnight

Day 2

1. Pick transformed bacteria colonies
2. Grow liquid cultures for miniprep overnight

Day 3

1. Miniprep overnight cultures and acquire purified plasmids
2. Check plasmids on a gel and send samples for commercial sequencing

Day 4

1. Analyze sequenced DNA and compare to desired sequence
2. If the GFP gene is successfully inserted into a plasmid, transform a sample of that miniprep to expression bacteria and grow colonies overnight

Day 5

1. Pick a colony of transformed expression bacteria
2. Grow in a liquid expression culture until a high cell density is reached
3. Add inducer to begin GFP expression
4. Incubate overnight for extended GFP expression

Day 6

1. Harvest GFP expression culture
2. Purify 6xHis-tagged GFP construct with Ni-NTA IMAC
3. Check elution fractions on SDS-PAGE gel and identify optimal fraction(s)
4. Perform melting curves with a sample from purification
5. Aliquot remaining protein and store at -80°C

6 | Follow-up Questions

Before the laboratory component of the practical course can begin, you will be given a short oral examination to confirm a basic understanding of the background information that is necessary to understand the various procedures and experiments of the course. Some of the following questions (as well as some new ones that are not shown below) will be presented to you. You should be able to demonstrate familiarity with the most important concepts, but you do not need to be an expert! Students are also encouraged to ask for clarification on anything in the practical course.

Where does GFP originally come from?

What are primary, secondary, tertiary and quaternary protein structures?

What are some of the biophysical forces that drive protein folding?

What are some conditions or forces that could cause a folded protein to denature?

How has wildtype GFP been improved or changed? What are some of the advantages or features of these variants compared to wildtype GFP?

How can the concentration of DNA or protein in solution be determined by light absorption?

How can we calculate the entropy and enthalpy of protein folding using GFP melting curve measurements?

What is the Central Dogma of molecular biology?

Why do we use plasmids for protein expression? Why do we use a gene for antibiotic resistance?

How does Ni-NTA IMAC work?

What are some uses of gel electrophoresis?

How does PCR amplification of DNA work?

7 | References

[1] Wikipedia contributors. "Aequorea victoria." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 27 Jun. 2017. Web. 11 Jul. 2017.

[2] By Sierra Blakely, Attribution, https://commons.wikimedia.org/w/index.php?curid=4091901

[3] Wikipedia contributors. "Green fluorescent protein." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 12 May. 2017. Web. 11 Jul. 2017.

[4] By Nathan Shaner - transferred from en:Image:FPbeachTsien.jpg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1752941

[5] By No machine-readable author provided. DrKjaergaard assumed (based on copyright claims). - No machine-readable source provided. Own work assumed (based on copyright claims)., Public Domain, https://commons.wikimedia.org/w/index.php?curid=1967109

[6] By Thomas Shafee - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=56202452

[7] By OpenStax College - Anatomy & Physiology, Connexions Web site. http://cnx.org/content/col11496/1.6/, Jun 19, 2013., CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=30131163

[8] Wikipedia contributors. "Protein folding." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 1 Jun. 2017. Web. 13 Jul. 2017.

[9] Wikipedia contributors. "Protein secondary structure." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 4 May. 2017. Web. 12 Jul. 2017.

[10] By Thomas Splettstoesser (www.scistyle.com) - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=28353539

[11] Wikipedia contributors. "Van 't Hoff equation." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 13 May. 2017. Web. 26 Jul. 2017.

[12] Zimmer, M. Green fluorescent protein (GFP): applications, structure, and related photophysical behavior. *Chem Rev*. 2002, 102(3):759-81.

[13] Pédeleacq, J. D., Cabantous, S., Tran, T., Terwilliger, T. C., Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nature Biotech.* 2006, 24(1):79-88.

[14] Wikipedia contributors. "Yellow fluorescent protein." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 30 Nov. 2016. Web. 19 Jul. 2017.

[15] Contributeurs à Wikipedia, 'DsRed', *Wikipédia, l'encyclopédie libre,* 25 juin 2017, 08:29 UTC, <http://fr.wikipedia.org/w/index.php?title=DsRed&oldid=138447632> [Page consultée le 19 juillet 2017]

[16] Wikipedia contributors. "MCherry." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 13 Nov. 2016. Web. 19 Jul. 2017.

[17] By David Goodsell - RCSB Protein Data Bank Molecule of the Month, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=39535373

[18] Wikipedia contributors. "Beer–Lambert law." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 18 Jun. 2017. Web. 18 Jul. 2017.

[19] By Dhorspool at en.wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=15183788

[20] Wikipedia contributors. "Central dogma of molecular biology." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 24 Jun. 2017. Web. 21 Jul. 2017.

[21] Wikipedia contributors. "Plasmid." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 22 Jun. 2017. Web. 21 Jul. 2017.

[22] Wikipedia contributors. "Cloning vector." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 3 Jun. 2017. Web. 24 Jul. 2017.

[23] Wikipedia contributors. "Lac operon." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 16 Jul. 2017. Web. 31 Jul. 2017.

[24] Wikipedia contributors. "Gene expression." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 10 May. 2017. Web. 31 Jul. 2017.

[25] Wikipedia contributors. "Protein production." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 4 Jul. 2017. Web. 31 Jul. 2017.

[26] Wikipedia contributors. "Affinity chromatography." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 31 Jul. 2017. Web. 31 Jul. 2017.

[27] Wikipedia contributors. "Protein tag." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 21 Jul. 2017. Web. 31 Jul. 2017.

[28] Wikipedia contributors. "Polyhistidine-tag." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 12 Jul. 2017. Web. 31 Jul. 2017.

[29] Bolanos-Garcia, V. M., Davies, O. R. Structural analysis and classification of native proteins from E. coli commonly co-purified by immobilized metal affinity chromatography. *Biochim. Biophys. Acta*. 2006, 1760(9):1304-13.

[30] Wikipedia contributors. "Nucleic acid quantitation." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 18 Jul. 2017. Web. 1 Aug. 2017.

[31] Simonian, M. H. 2002. Spectrophotometric Determination of Protein Concentration. Current Protocols in Cell Biology. 15:3B:A.3B.1–A.3B.7.

[32] Online protein parameter calculator: web.expasy.org/protparam/

[33] Patterson, G., Knobel, S., Sharif, W., Kain, S., Piston, D. Use of the Green Fluorescent Protein and Its Mutants in Quantitative Fluorescene Microscopy. *Biophysical J.* 1997, 73: 2782-90.

[34] http://web.expasy.org/cgi-bin/protparam/protparam

[35] Wikipedia contributors. "Transformation (genetics)." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 23 May. 2017. Web. 9 Aug. 2017.

[36] Wikipedia contributors. "Centrifugation." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 9 Jul. 2017. Web. 1 Aug. 2017.

[37] Wikipedia contributors. "Plasmid preparation." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 4 Jun. 2017. Web. 9 Aug. 2017.

[38] https://tools.thermofisher.com/content/sfs/manuals/MAN0013004_GeneRuler_1kb_DNALadder_250ug_UG.pdf

[39] http://www.bio-rad.com/de-de/sku/1610363-precision-plus-protein-unstained-protein-standards-strep-tagged-recombinant-1-ml

[40] Wikipedia contributors. "Gel electrophoresis." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 2 Aug. 2017. Web. 9 Aug. 2017.

[41] Wikipedia contributors. "Agarose gel electrophoresis." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 3 Aug. 2017. Web. 9 Aug. 2017.

[42] Wikipedia contributors. "Polyacrylamide gel electrophoresis." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 21 Jul. 2017. Web. 9 Aug. 2017.

[43] Wikipedia contributors. "Polymerase chain reaction." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 1 Aug. 2017. Web. 11 Aug. 2017.

[44] Wikipedia contributors. "Overlap extension polymerase chain reaction." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 11 Feb. 2017. Web. 11 Aug. 2017.