

SUPPLEMENTAL MODULES (BIOCHEMISTRY)



Michael Blaber
Florida State University

Florida State University
Supplemental Modules (Biochemistry)

Michael Blaber

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (<https://LibreTexts.org>) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by [NICE CXone Expert](#) and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptations contact info@LibreTexts.org or visit our main website at <https://LibreTexts.org>.

This text was compiled on 05/01/2026

TABLE OF CONTENTS

Licensing

1: DNA

- 1.1: DNA as Genetic Material
- 1.2: Structure of DNA and RNA
- 1.3: Bacterial Restriction/Modification system
- 1.4: DNA Modifying Enzymes
- 1.5: DNA Replication - Introduction to Prokaryotic replication
- 1.6: DNA Supercoiling and Topoisomerases

2: Bacteria

- 2.1: The Bacteria - An Introduction
- 2.2: Bacterial Growth and Reproduction
- 2.3: Extrachromosomal elements, plasmids, selectable markers
- 2.4: Central Dogma and the Genetic Code
- 2.5: Gene and Operon
- 2.6: The lac Operon; CAP site; DNA footprinting
- 2.7: Transcriptional Regulation: Transcription termination, the trp operon
- 2.8: mRNA Translation

3. Biotechnology 1

- 3.1: Gel Electrophoresis
- 3.2: DNA Sequence Analysis
- 3.3: Polymerase Chain Reaction (PCR) and Cloning of PCR Products
- 3.4: Prokaryotic Expression Vectors
- 3.5: Protein Sequencing, Peptide Mapping, Synthetic Genes
- 3.6: cDNA and Genomic Libraries

4. Biotechnology 2

- 4.1: Protein Purification
- 4.2: M13 Phage
- 4.3: M13 Phage Display Libraries
- 4.4: SELEX (Selective Evolution of Ligands by Exponential Enrichment)
- 4.5: Protein-Protein Recognition Probed Using a Yeast Transcriptional Activator System
- 4.6: Molecular Imprinting

5. Lab Notes Part 1

- 5.1: Basic practices and techniques, Laboratory notebooks, Presentation of data, Sections of a scientific report
- 5.2: Spectrophotometry
- 5.3: Quantification of Protein Concentration
- 5.4: Chromatography
- 5.5: Gel Electrophoresis of Proteins
- 5.6: Introduction to Protein Purification

6. Lab Notes Part 2

- 6.1: Genetic Transformation (using bacteria and the pGLO plasmid)
- 6.2: Enzyme kinetics
- 6.3: Ligand binding
- 6.4: Restriction Mapping
- 6.5: Polymerase Chain Reaction (PCR)
- 6.6: Use of PC and internet for biochemical research

[Index](#)

[Glossary](#)

[Detailed Licensing](#)

Licensing

A detailed breakdown of this resource's licensing can be found in [Back Matter/Detailed Licensing](#).

CHAPTER OVERVIEW

1: DNA

- 1.1: DNA as Genetic Material
- 1.2: Structure of DNA and RNA
- 1.3: Bacterial Restriction/Modification system
- 1.4: DNA Modifying Enzymes
- 1.5: DNA Replication - Introduction to Prokaryotic replication
- 1.6: DNA Supercoiling and Topoisomerases

This page titled [1: DNA](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

1.1: DNA as Genetic Material

In the early 1900's many people thought that **protein** must be the **genetic material** responsible for inherited characteristics. One of the reasons behind this belief was the knowledge that proteins were quite complex molecules and therefore, they must be specified by molecules of equal or greater complexity (i.e. other proteins). DNA was known to be a relatively simple molecule, in comparison to proteins, and therefore it was hard to understand how a complex molecule (a protein) could be determined by a simpler molecule (DNA). What were the key experiments which identified DNA as the primary genetic material?

1928 F. Griffith

Background:

Diplococcus pneumoniae, or pneumococcus, is a nasty little bacteria which, when injected into mice, will cause pneumonia and **death** in the mouse. The bacteria contains a **capsular polysaccharide** on its surface which protects the bacteria from host defenses. Occasionally, variants (mutants) of the bacteria arise which have a defect in the production of the capsular polysaccharide. The mutants have two characteristics: 1) They are **avirulent**, meaning that without proper capsular polysaccharide they are unable to mount an infection in the host (they are destroyed by the host defenses), and 2) Due to the lack of capsular polysaccharide the surface of the mutant bacteria appears **rough** under the microscope and can be distinguished from the wild type bacteria (whose surface appears **smooth**).

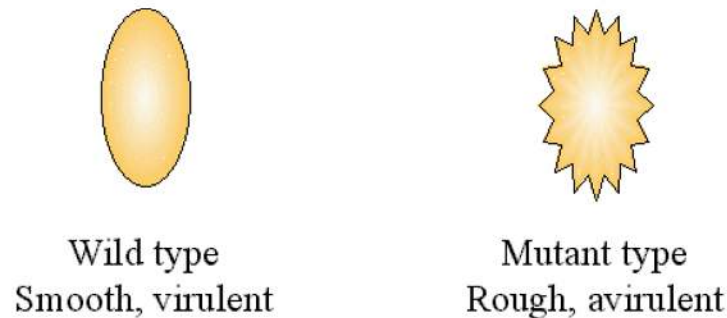


Figure 1.1.1: Wild type vs. Mutant type pneumococcus

The virulent smooth wild type pneumococcus can be **heat treated** and rendered **avirulent** (still appears smooth under the microscope however). Finally, there are several different *subtypes* of pneumococcus capsular polysaccharide (subtypes I, II and III). These subtypes are readily distinguishable from one another, and **each** can give rise to mutants lacking capsular polysaccharide (i.e. the avirulent rough type).

The experiments:

Controls:

- w.t. (smooth) + mouse = dead mouse
- mutant (rough) + mouse = live mouse
- heat treated w.t. (smooth) + mouse = live mouse

Combinations:

- heat treated w.t. (smooth) + mutant (rough) + mouse = **dead mouse**

In this case when the bacteria were recovered from the cold lifeless mouse they were smooth virulent pneumococcus (i.e. indistinguishable from wild type).

A closer look at what is going on, by keeping using, and keeping track of, different subtypes

- heat treated w.t. (smooth) type I + mutant(rough) type II + mouse = **dead mouse**

*In this case when the bacteria were isolated from the cold lifeless mouse they were smooth virulent **type I** pneumococcus.*

The overall conclusions from these experiments was that there was a "transforming agent" in the the heat treated type I bacteria which transformed the live mutant (rough) type II bacteria to be able to produce type I capsule polysaccharide.

Question

Was the "transforming agent" protein or DNA, or what?

1944 O.T. Avery

Background:

The experiment of Griffith could not be taken further until methods were developed to separate and purify DNA and protein cellular components. Avery utilized methods to extract relatively pure DNA from pneumococcus to determine whether it was the "transforming agent" observed in Griffith's experiments.

The experiment:

- w.t. (smooth) type I -> extract the DNA component
- mutant (rough) type II + type I DNA + mouse = dead mouse

Isolation of bacteria from the dead mouse showed that they were type I w.t. (smooth) bacteria

A more sophisticated experiment:

Purified type I DNA was divided into two aliquots. One aliquot was treated with **DNase** - an enzyme which non-specifically degrades DNA. The other aliquot was treated with **Trypsin** - a protease which (relatively) non-specifically degrades proteins.

- Type I DNA + DNase + mutant (rough) type II + mouse = live mouse
- Type I DNA + Trypsin + mutant (rough) type II + mouse = **dead mouse**

Conclusion:

The work of Avery provided strong evidence that the "transforming agent" was in fact DNA (and not protein). However, not everyone was convinced. Some people felt that a residual amount of protein might remain in the purified DNA, even after Trypsin treatment, and could be the "transforming agent".

1952 A.D. Hershey and M. Chase

Background:

T2 is a virus which attacks the bacteria *E. coli*. The virus, or *phage*, looks like a tiny lunar landing module:

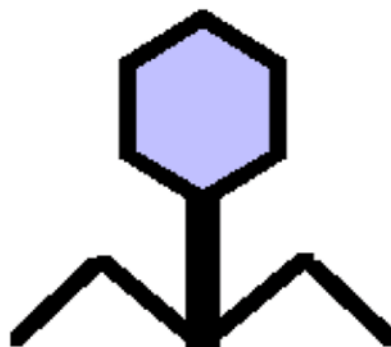


Figure 1.1.2: T2 phage

The viral particles adsorb to the surface of the *E. coli* cells. It was known that some material then leaves the phage and enters the cell. The "empty" phage particles on the surface cells can be physically removed by putting the cells into a blender and whipping them up. In any case, some 20 minutes after the phage adsorb to the surface of the bacteria the bacteria bursts open (lysis) and releases a multitude of progeny virus.

If the media in which the bacteria grew (and were infected) included ^{32}P labeled ATP, progeny phage could be recovered with this isotope incorporated into its DNA (normal proteins contain only hydrogen, nitrogen, carbon, oxygen, and sulfur atoms). Likewise, if the media contained ^{35}S labeled methionine the resulting progeny phage could be recovered with this isotope present only in its protein components (normal DNA contains only hydrogen, nitrogen, carbon, oxygen, and phosphorous atoms).

The experiment:

Phage were grown in the presence of either ^{32}P or ^{35}S isotopic labels.

1) *E. coli* were infected with ^{35}S labeled phage. After infection, but prior to cell lysis, the bacteria were whipped up in a blender and the phage particles were separated from the bacterial cells. The isolated bacterial cells were cultured further until lysis occurred. The released progeny phage were isolated.

Where the ^{35}S label went:

- Adsorbed phage shells 85%
- Infected cells (prior to lysis) 15%
- Lysed cell debris 15%
- Progeny phage <1%

2) *E. coli* were infected with ^{32}P labeled phage. The same steps as in 1) above were performed.

Where the ^{32}P label went:

- Adsorbed phage shells 30%
- Infected cells (prior to lysis) 70%
- Lysed cell debris 40%
- Progeny phage 30%

Conclusion:

The material which was being transferred from the phage to the bacteria during infection appeared to be mainly DNA. Although the results were not entirely unambiguous they provided additional support for the view that ***DNA was the "stuff" of genetic inheritance.***

This page titled [1.1: DNA as Genetic Material](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

1.2: Structure of DNA and RNA

The Double Helix

DNA (deoxyribonucleic acid) and RNA (ribonucleic acid) are composed of two different classes of nitrogen-containing bases: the **purines** and **pyrimidines**. The most commonly occurring purines in DNA are **adenine** and **guanine**:

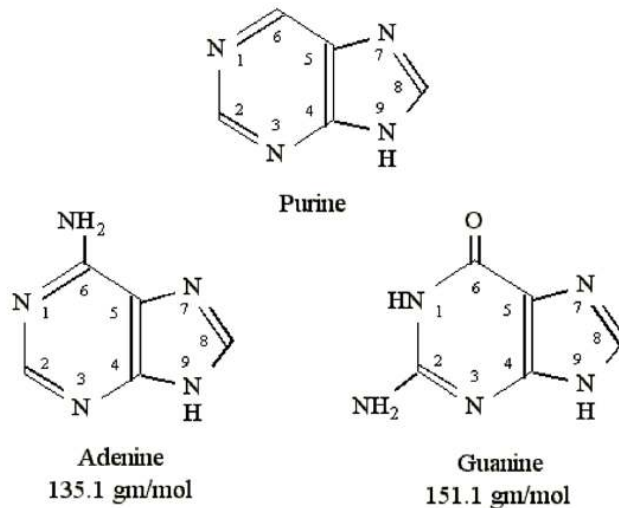


Figure 1.2.1: Purines

The most commonly occurring pyrimidines in DNA are **cytosine** and **thymine**:

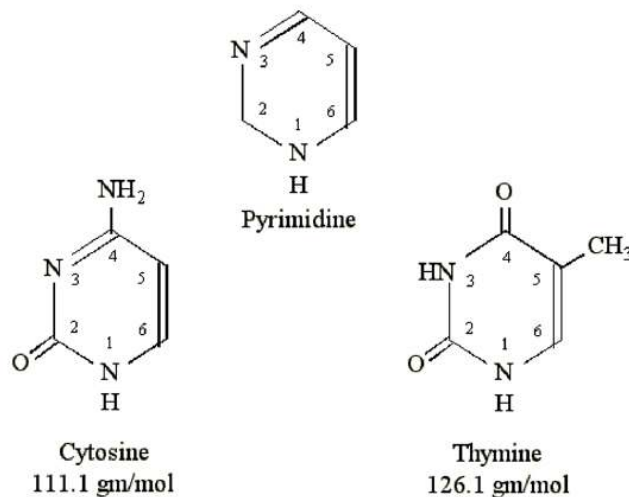


Figure 1.2.2: Pyrimidines

RNA contains the same bases as DNA with the exception of thymine. Instead, RNA contains the pyrimidine **uracil**:

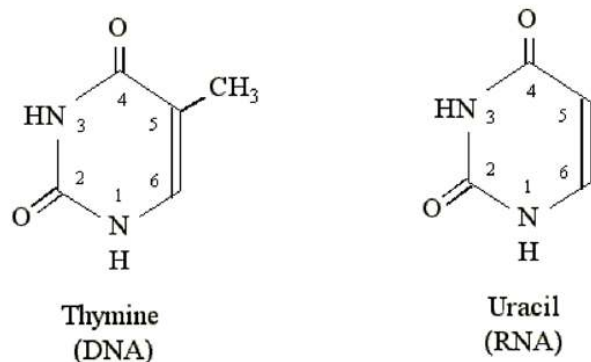


Figure 1.2.3: Thymine vs. Uracil

Adenine, guanine, cytosine, thymine and uracil are usually abbreviated using the single letter codes A, G, C, T and U, respectively.

Purines and pyrimidines can form chemical linkages with pentose (5-carbon) sugars. The carbon atoms on the sugars are designated 1', 2', 3', 4' and 5'. It is the 1' carbon of the sugar that becomes bonded to the nitrogen atom at **position N1 of a pyrimidine** or **N9 of a purine**. DNA precursors contain the pentose deoxyribose. RNA precursors contain the pentose ribose (which contains an additional OH group at the 2' position):

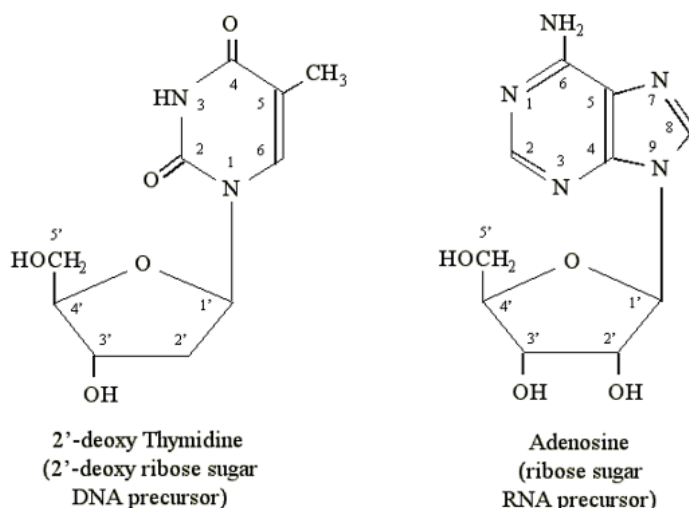


Figure 1.2.4: Nucleosides

Before a nucleoside can become part of a DNA or RNA molecule it must become complexed with a phosphate group to form a **nucleotide** (either a deoxyribonucleotide or ribonucleotide). Nucleotides can possess **1, 2 or 3 phosphate groups**, e.g. the nucleotides adenosine monophosphate (AMP), adenoside diphosphate (ADP) and adenosine triphosphate (ATP). The phosphate groups are attached to the **5' carbon** of the ribose sugar moiety. Beginning with the phosphate group attached to the 5' ribose carbon, they are labeled **a, b and g phosphate**. It is the **tri-phosphate nucleotide** which is incorporated into DNA or RNA.

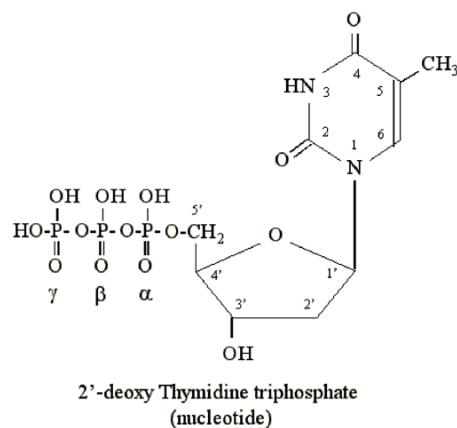


Figure 1.2.5: Nucleotide

DNA and RNA are simply long polymers of nucleotides called **polynucleotides**. Only the **a phosphate** is included in the polymer. It becomes chemically bonded to the **3' carbon** of the sugar moiety of another nucleotide:

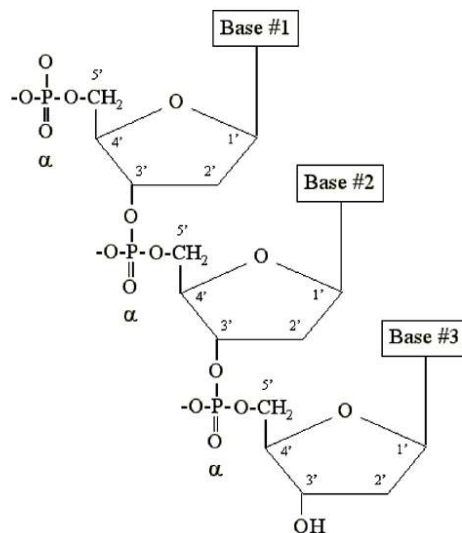


Figure 1.2.6: Polynucleotide

In other words, the polynucleotide is connected by a series of 5' to 3' phosphate linkages. Note the sequence of the bases in the above diagram. **Polynucleotide sequences are referenced in the 5' to 3' direction**. Typically, polynucleotides will contain a 5' phosphate and 3' hydroxyl terminal groups. The common representation of polynucleotides is as an **arrow** with the **5' end at the left and the 3' end at the right**.



Summary of terms:

Base	Nucleoside	Nucleotide	RNA (monophosphate)	DNA (monophosphate)	Code
Adenine	Adenosine	(Adenylic acid)	AMP	dAMP	A
Guanine	Guanosine	(Guanylic acid)	GMP	dGMP	G
Cytosine	Cytidine	(Cytidylic acid)	CMP	dCMP	C
Thymine	Thymidine	(Thymidylic acid)		dTMP	T
Uracil	Uridine	(Uridylic acid)	UMP		U

What is the structure of DNA? How is the structure related to function?

1950's

The primary chemical structure of polynucleotides was known (i.e. the 3'-5' phosphate linkage).

1951 E. Chargaff

The experiment:

Take DNA from a variety of species and hydrolyze it to yield individual pyrimidines and purines. Determine the relative concentrations of the A, T, C and G bases.

Result:

Although different species had uniquely different ratios of pyrimidines or purines, the relative concentrations of adenine always equaled that of thymine, and guanine equaled cytosine.

Chargaff's Law: A=T, G=C

1950's R.E. Franklin

X-ray diffraction studies of DNA fibers demonstrated that DNA adopted a highly ordered helical structure. Franklin concluded that two or more chains must coil around each other to form a helix. Some basic dimensions of the helix were calculated from the x-ray diffraction data.

1953 L. Pauling and R.B. Corey

Propose a three chain helical structure for DNA with the phosphate backbone in the center and the bases on the outside.

1953 J.D. Watson and F.H.C. Crick

Identified a hydrogen bonding arrangement between models of thymine and adenine bases, and between cytosine and guanine bases which fulfilled Chargaff's rule:

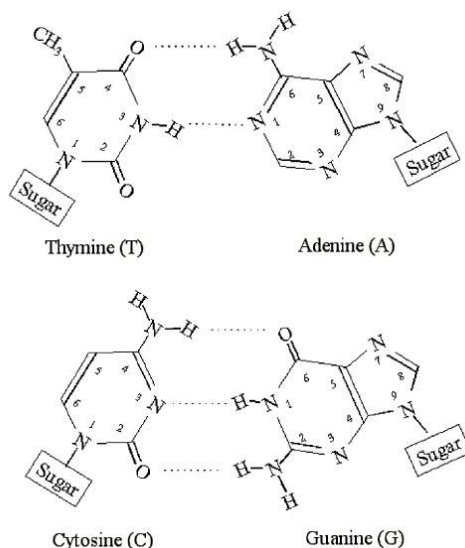


Figure 1.2.7: Chargaff's Rule Bonding

Note that the "TA" pair can overlay the "GC" pair with the bonds to the sugar groups in similar juxtaposition. In the "double helix" model of Watson and Crick the polynucleotide chains interact to form a double helix with the chains running in **opposite directions**. The bases are directed towards the center (and stack on top of one another) and the sugar backbones face the outside of the helix.

The Watson and Crick model had the following physical dimensions:

- 34 Å per helical repeat
- 10 base pairs per repeat (i.e. per turn of the helix)
- 3.4 Å inter-base stacking distance
- 20 Å diameter for the helical width

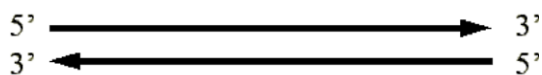
Physical characteristics of the model matched those determined by Rosalind Franklin's x-ray diffraction studies.

Consequences of the model for genetic information:

The Watson and Crick paper was an exercise in brevity (1 page only in Nature). The structure was so rich with implication that quite a bit could be written. The authors, however, chose only to say "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material".

1. If G always paired with C, and T always paired with A, then either strand could be regenerated from the **complementary information** in the other strand.
2. The basis of the complementarity was hydrogen bonding, i.e. non-covalent interactions which could be easily broken and re-formed.
3. The information which DNA carried was within the unique **base sequence** of the DNA.
4. From the general interior location of the bases, it would appear that the double helix would **have to dissociate** in order to access the information.
5. The non-equatorial location of the sugar moieties (see above) suggested that the DNA helix would have a **major groove** and a **minor groove**.

General notation of double stranded DNA:



This page titled [1.2: Structure of DNA and RNA](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

1.3: Bacterial Restriction/Modification system

The restriction/modification system in bacteria is a small-scale immune system for protection from infection by foreign DNA.

W. Arber and S. Linn (1969)

Plating efficiencies of bacteriophage lambda (1 phage) grown on *E. coli* strains C, K-12 and B, when plated on these bacteria:

E. coli strain on which parental phage had been grown	E. coli strain for plating phage		
	C	K-12	B
C	1	<10 ⁻⁴	<10 ⁻⁴
K-12	1	1	<10 ⁻⁴
B	1	<10 ⁻⁴	1

- The DNA of phage which had been grown on strains K-12 and B were found to have chemically modified bases which were methylated.
- Additional studies with other strains indicate that different strains had specific methylated bases.
- Typical sites of methylation include the **N₆ position of adenine**, the **N₄ position of cytosine**, or the **C₅ position of cytosine**.

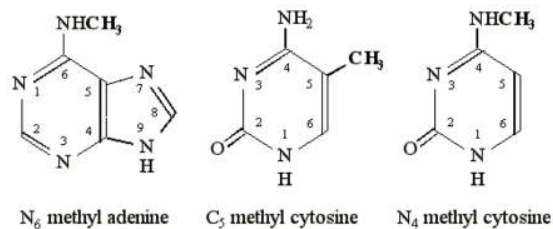


Figure 1.3.1: Methylation

- In addition, only a fractional percentage of bases were methylated (i.e. not every adenine was methylated, for example) and these occurred at very specific sites in the DNA.
- A characteristic feature of the sites of methylation, was that they involved palindromic DNA sequences.



Figure 1.3.2: EcoRI methylase specificity. Rubin and Modrich, 1977

- In addition to possessing a particular methylase, individual bacterial strains also contained accompanying specific endonuclease activities.
- The endonucleases cleaved at or near the methylation recognition site.

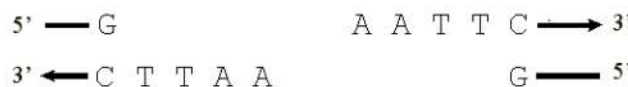


Figure 1.3.3: Cleavage at methylation sites

- These specific nucleases, however, would not cleave at these specific palindromic sequences if the DNA was methylated.

Thus, this combination of a specific methylase and endonuclease functioned as a type of immune system for individual bacterial strains, protecting them from infection by foreign DNA (e.g. viruses).

- In the bacterial strain EcoRI, the sequence GAATTC will be methylated at the internal adenine base (by the EcoRI methylase).
- The EcoRI endonuclease within the same bacteria will not cleave the methylated DNA.
- Foreign viral DNA, which is not methylated at the sequence "GAATTC" will therefore be recognized as "foreign" DNA and will be cleaved by the EcoRI endonuclease.
- Cleavage of the viral DNA renders it non-functional.

Such endonucleases are referred to as "restriction endonucleases" because they restrict the DNA within the cell to being "self".

The combination of restriction endonuclease and methylase is termed the "restriction-modification" system.

Of course, this type of protective system is beaten if the attacking phage was previously grown on the same strain as that which it is infecting. In this case the phage will have its DNA already methylated at the appropriate sequence, and will be recognized as "self" (see the table above). *E. coli* strain 'C' (above) is strain which has no known restriction-modification system.

We will discuss DNA replication later, but it should be mentioned that:

- replicating host DNA will initially have one strand (parental) methylated and the other (nascent strand) non-methylated.
- This is recognized as "self" and is not cleaved by the restriction endonuclease.

- It is subsequently methylated by the host methylase.

Structural and biochemical studies have indicated that for the common R/M systems (so called **type II**), the methylase recognizes and methylates one strand of the DNA duplex, whereas the restriction endonuclease recognizes both strands of the DNA (i.e. both strands must be non-methylated for recognition). It is able to do this because it is a homo-dimer protein.

Restriction endonucleases

Since different bacterial strains and species have potentially different R/M systems, their characterization has made available over 200 endonucleases with different sequence specific cleavage sites.

- They are one of the primary tools in modern molecular biology for the manipulation and identification of DNA sequences.
- Restriction endonucleases are commonly named after the bacterium from which it was isolated.

Examples of different restriction enzymes

Name	Source	Recognition Sequence	Comments
Alu I	<i>Arthrobacter luteus</i>	<pre> 5' ... A G C T ... 3' 3' ... T C G A ... 5' </pre>	"Four cutter". Leaves blunt ends to the DNA.
Bfa I	<i>Bacteroides fragilis</i>	<pre> 5' ... C T A G ... 3' 3' ... G A T C ... 5' </pre>	"Four cutter". Leaves 5' overhang.
Nci I	<i>Neisseria cinerea</i>	<pre> C 5' ... C C G G G ... 3' 3' ... G G C C C ... 5' G </pre>	"Five cutter". Middle base can be either cytosine or guanine. Leaves 5' overhang. <u>Different recognition sites may have non-complementary sequences.</u>
Eco R1	<i>Escherichia coli</i>	<pre> 5' ... G A A T T C ... 3' 3' ... C T T A A G ... 5' </pre>	"Six cutter". Leaves 5' overhang. Behaves like a "four cutter" ('star' activity) in high salt buffer. \$44 for 10,000 units.
Hae II	<i>Haemophilus aegyptius</i>	<pre> 5' ... Pu G C G C Py ... 3' 3' ... Py C G C G Pu ... 5' </pre>	"Six cutter". Pu is any purine, Py is any pyrimidine. Leaves 3' overhang.
EcoO109I	<i>Escherichia coli</i>	<pre> 5' ... Pu G G N C C Py ... 3' 3' ... Py C C N G G Pu ... 5' </pre>	"Seven cutter". Pu is any purine, Py is any pyrimidine, N is any base. Leaves 5' overhang. <u>Different recognition sites may have non-complementary sequences.</u>
Bgl I	<i>Bacillus globigii</i>	<pre> 5' ... GCCN NNNNGC ... 3' 3' ... CGGNNN NCCG ... 5' </pre>	"Six cutter with interrupted palindrome". Leaves 5' overhang. <u>Different recognition sites may have non-complementary sequences.</u>
Bsa HI	<i>Bacillus stearothermophilus</i>	<pre> 5' ... G Pu C G Py C ... 3' 3' ... C Py G C Pu G ... 5' </pre>	"Six cutter". Different recognition sites <u>will</u> be complementary.
Aat II	<i>Acetobacter aceti</i>	<pre> 5' ... G A C G T C ... 3' 3' ... C T G C A G ... 5' </pre>	"Six cutter" with 3' overhang. Same recognition sequence as Bsa HI, but different cleavage position.
Bpm I	<i>Bacillus pumilus</i>	<pre> 5' ... C T G G A G N₁₆ ... 3' </pre>	Non-palindrome, distal cleavage. Leaves 3' overhang. \$50 for 50 units.

		3' ... G A C C T C N₁₄ ... 5'	
Not I	<i>Nocardia otitidiscaviarum</i>	5' ... G C G G C C G C ... 3' 3' ... C G C C G G C G ... 5'	"Eight cutter". Leaves 5' overhang.
Bsm I	<i>Bacillus stearothermophilus</i>	5' ... G A A T G C N ... 3' 3' ... C T T A C G N ... 5'	"weird". Leaves 3' overhang.

- The utility of restriction endonucleases lies in their specificity and the frequency with which their recognition sites occur within any given DNA sample.
- If there is a 25% probability for a specific base at any given site, then the frequency with which different restriction endonuclease sites will occur can be easily calculated (0.25ⁿ):**

Nucleotide Specificity	Example	Frequency of Occurrence
Four	Alu I	256 (0.25 Kb)
Five	Nci I	1024 (1.0 Kb)
Six	EcoR I	4096 (4.1 Kb)
Seven	EcoO109I	16384 (16.4 Kb)
Eight	Not I	65536 (65.5 Kb)

Thus, on average, any given DNA will contain an Alu I site every 0.25 kilobases, whereas a Not I site occurs once about every 65.5 kilobases.

- Not I is therefore a very useful enzyme for isolating large regions of DNA, typically in research involving genomic DNA manipulations.
- Alu I would be expected to digest a DNA sample into lots of little pieces.

The assortment of DNA fragments would represent a specific "fingerprint" of the particular DNA being digested. Different DNA would not yield the same collection of fragment sizes. Thus, DNA from different sources can be either matched or distinguished based on the assembly of fragments after restriction endonuclease treatment. **These are termed "Restriction Fragment Length Polymorphisms", or RFLP's.** This simple analysis is used in various aspects of molecular biology as well as a law enforcement and genealogy. For example, genetic variations which distinguish individuals also may result in fewer or additional restriction endonuclease recognition sites.

Restriction endonucleases are supplied in various concentrations with activities that are based upon cleavage rates of "standard" DNA samples.

- One unit of activity is typically defined as the amount of enzyme required to digest (or "restrict") one microgram of reference DNA in one hour at 37 °C.**
- The reference DNA may actually have one or more recognition sites for the nuclease in question. DNA's used as "standard" samples may include phage λ DNA, or the plasmid pBR322.
- The endonuclease hydrolysis is a spontaneous reaction and does not, for example, require addition of ATP. Reaction buffers for restriction endonucleases usually contain a buffer component (typically 10 mM TRIS buffer around pH 8.0), magnesium salt (often 10 mM MgCl₂), a reducing agent (usually 1mM dithiothreitol, or DTT), a protective carrier protein (typically 100 ug/ml bovine serum albumin, or BSA), and salt (sodium chloride).
- The biggest determinant of enzyme activity is typically the ionic concentration (NaCl content) of the buffer.** Although there are hundreds of different restriction endonucleases, the majority of them can exhibit between 30-100% activity using a simple system of three buffers, containing either low (20 mM), medium (100 mM) or high (250 mM) salt (NaCl) concentrations in the above described buffer.

Enzyme digests are typically performed for 1-2 hours at 37 °C. However, quantitative digestion can sometimes only be achieved after extended incubation (i.e. overnight).

This page titled [1.3: Bacterial Restriction/Modification system](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

1.4: DNA Modifying Enzymes

Methylases

Just as the study of the bacterial restriction-modification system has provided a variety of specific endonucleases, there are also available a variety of specific DNA methylases.

- The recognition sequences of the methylases are the same as the associated endonucleases (e.g. EcoR1 methylase recognizes and methylates at the sequence "GAATTC").
- All methylases transfer the methyl group from S-adenosylmethionine (SAM) to a specific base in the recognition sequence, and SAM is a required component in the methylation reaction.
- Methylation of DNA usually has the effect of protecting the DNA from the related restriction endonuclease. However, there are methylases with minimal specificity. For example, Sss I methylase will methylate cytosine residues in the sequence 5' ... CG ... 3'. In this case, *the methylated DNA will be protected from a wide variety of restriction endonucleases*.
- Some restriction endonucleases will only cut DNA at their recognition sites if the DNA **is** methylated (e.g. Dpn I).
- Still other restriction endonucleases will cut **both** methylated and non-methylated DNA at their recognition sequences (e.g. BamH I).

dam and dcm Methylation

- The methylase encoded by the *dam* gene (dam methylase) transfers a methyl group from SAM to the N₆ position of the adenine base in the sequence 5' ... GATC ... 3'.
- The methylase encoded by the *dcm* gene (dcm methylase) methylates the internal cytosine base, at the C₅ position, in the sequences 5' ... CCAGG ... 3' and 5' ... CCTGG ... 3'.
- Almost all strains of *E. coli* commonly used in cloning have a *dam+dcm+* genotype. The point here is *not* that we particularly want our DNA to be methylated, but that to make a *dam-dcm-* host someone has to mutate the bacteria and isolate the correct mutant. That apparently has not been done for a lot of bacterial strains. Probably because the *dam* and *dcm* methylation affects only a small subset of potential restriction endonucleases

DNA isolated from *dam+dcm+* strains **will not actually be cut** by a modest subset of available restriction endonucleases:

Recognition sequence	Restriction enzyme	GATC	G ^{me} ATC
TGATCA	<i>Bcl I</i>	+	-
GATC	<i>Mbo I</i>	+	-
ATCGAT	<i>Cla I</i>	+	-
TCTAGA	<i>Xba I</i>	+	-
TCGA	<i>Taq I</i>	+	-
GAAGA	<i>Mbo II</i>	+	-
GGTGA	<i>Hph I</i>	+	-

DNA may have to be prepared from *E. coli* strains which are *dam-dcm-* in order to be cut by these enzymes.

DNA Polymerases

A wide variety of polymerases have been characterized and are commercially available. All DNA polymerases share two general characteristics:

1. They add nucleotides to the 3'-OH end of a primer

2. The order of the nucleotides in the nascent polynucleotide is template directed

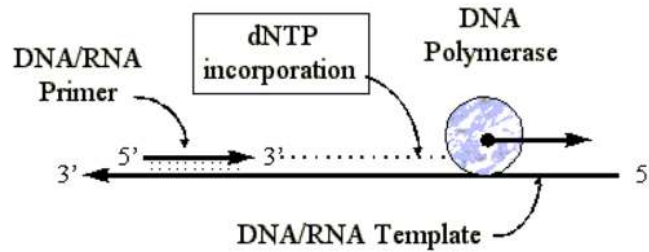


Figure 1.4.1: DNA Replication

In addition to the 5'→3' polymerase activity, polymerases can contain exonuclease activity. This exonuclease activity can proceed either in the 5'→3' direction, or in the 3'→5' direction.

- Exonuclease activity in the 3'→5' direction allows the polymerase to correct a mistake if it incorporates an incorrect nucleotide (so called "**error correction activity**"). It can also slowly degrade the 3' end of the primer.
- Exonuclease activity in the 5'→3' direction will allow it to degrade any other hybridized primer it may encounter. Without 5'→3' exonuclease activity, obstructing primers may or may not be physically displaced, depending on the polymerase being used.

Different polymerases have differing error rates of misincorporation, and different rates of polymerization.

E. coli DNA polymerase I	E. coli DNA polymerase I - Klenow Fragment	T4 DNA polymerase	T7 DNA polymerase	Taq DNA polymerase	M-MuLV Reverse Transcriptase
5'→3' exonuclease activity	*				*
3'→5' exonuclease activity	*	*	*	*	
Error Rate (x10 ⁻⁶)	9	40	<1	15	285
Strand Displacement		*			
Heat Inactivation	*	*	*	*	

Uses of polymerases

The various activities of the different polymerases lend them to a variety of applications. For example, restriction endonucleases can yield fragments of DNA with either 3' or 5' nucleotide "overhangs".

- In the case of 5' overhangs, the 5'→3' polymerase activity can fill these in to make **blunt ends**.
- In the case of 3' overhangs, the 3'→5' exonuclease activity present in some polymerases (especially T4 DNA polymerase) can "chew back" these ends to also make blunt-ended DNA fragments.

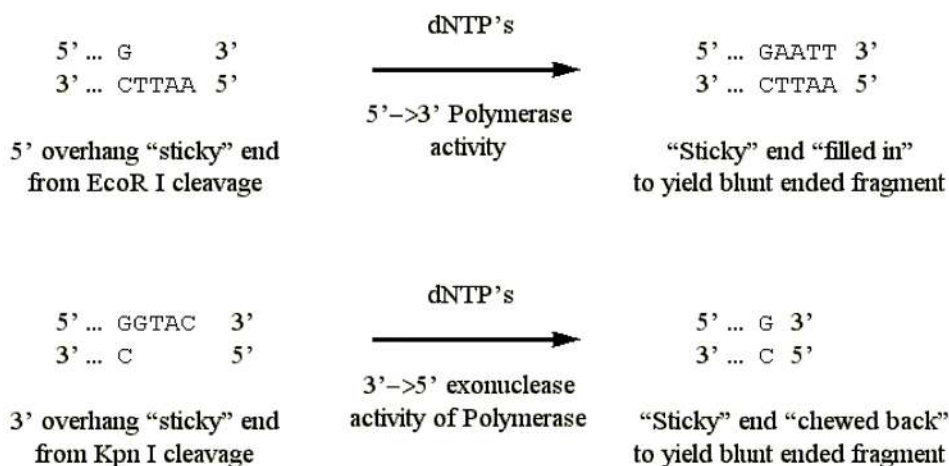


Figure 1.4.2: Polymerase activity

"Nick-translation"

This method is used to obtain highly radiolabeled single strand DNA fragments, which makes use of 5'→3' exonuclease activity present in some polymerases (*E. coli* DNA polymerase I, for example).

- In this method a DNA duplex of interest is "nicked" (i.e. one of the strands is cut; see DNase I).
- Then DNA pol I is added along with radiolabeled nucleotides. The 5'→3' exonuclease activity chews away the 5' end at the "nick" site and the polymerase activity incorporates the radiolabeled nucleotides. The resulting polynucleotide will be highly radiolabeled and will hybridize to the DNA sequence of interest.

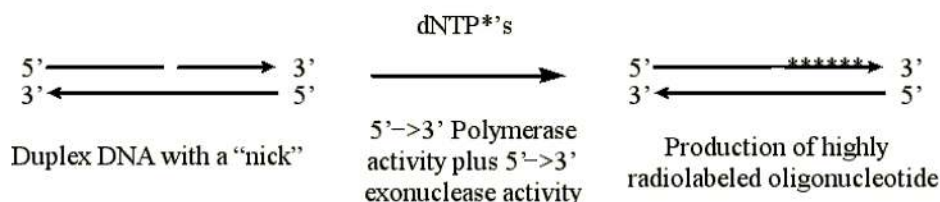


Figure 1.4.3: Nick-translation

- Thermostable polymerases have the ability to remain functional at temperature ranges where the DNA duplex will actually "melt" and become separated. This has allowed the development of the "**Polymerase Chain Reaction**" technique (PCR), which has had a profound impact on modern Biotechnology. We will discuss this method at a later date.
- The incorporation of dideoxy bases (i.e. no hydroxyl groups on either the 2' or 3' carbon of the ribose sugar) leads to **termination of the polymerase reaction**. This will be discussed in greater detail later. However, this chain termination by incorporation of dideoxynucleotides is the basis of the Sanger method of DNA sequencing, as well as therapies to try to inhibit viral replication.

Nucleases

Nuclease BAL-31

- This is an exonuclease (starts at the termini and works inward) which will degrade both 3' and 5' termini of double-stranded DNA. It will not make internal cleavages ("nicks"), however, it will degrade the ends of DNA at existing internal "nicks" (which create both 3' and 5' termini).
- The degradation of termini is not coordinated, meaning that the product is not 100% blunt-ended (even though the original duplex may have been blunt ended).
- Such "ragged" ends can be made blunt by filling in and chewing back by a suitable polymerase (e.g. T4 DNA polymerase). The unit definition of 1 unit is the amount of enzyme required to remove 200 base pairs from each end of duplex DNA in 10 minutes

at 30 °C.

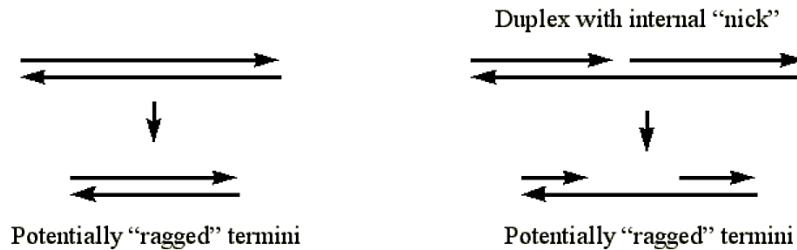


Figure 1.4.4: Nuclease BAL-31 activity

Exonuclease III

- Catalyzes the stepwise removal of nucleotides from the 3' hydroxyl termini of duplex DNA.
- The enzyme will attack the 3' hydroxyl at duplex DNA with blunt ends, with 5' overhangs, or with internal "nicks".
- Since duplex DNA is required, the enzyme **will not** digest the 3' end of duplex DNA where the termini are 3' overhangs.

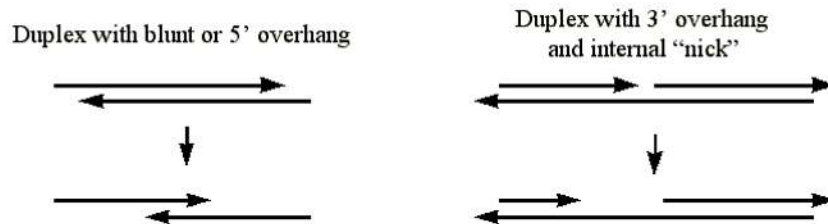


Figure 1.4.5: Exonuclease III Activity

Mung Bean Nuclease (isolated from mung bean sprouts)

- A single strand specific DNA and RNA endonuclease which will degrade single strand extensions from the ends of DNA and RNA leaving blunt ends.
- The single strand extensions can be either 5' or 3' extensions - both are removed and a blunt duplex is left.

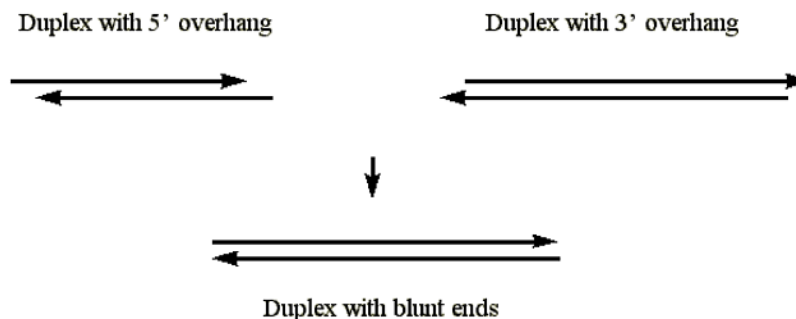


Figure 1.4.6: Mung Bean Nuclease activity

Deoxyribonuclease I (DNase I) from Bovine pancreas

- This enzyme hydrolyzes duplex or single DNA strands preferentially at the phosphodiester bonds 5' to pyrimidine nucleotides
- In the presence of Mg^{2+} ion, DNase I attacks each strand independently and produces nicks in a random fashion (useful for nick-translation)
- In the presence of Mn^{2+} ion DNase I cleaves both strands of DNA at approximately the same position (but leaving "ragged" ends)

Ligases

- Ligases catalyze the **formation** of a phosphodiester bond between juxtaposed 5' phosphate and 3' hydroxyl termini of nucleotides (potentially RNA or DNA depending on the ligase).
- In a sense, they are the opposite of restriction endonucleases, but they do not appear to be influenced by the local sequence, *per se*.
- Ligases require **either rATP or NAD⁺ as a cofactor**, and this contrasts with restriction endonucleases.

The following are different types of ligases and their characteristics.

T4 DNA ligase

- Isolated from bacteriophage T4.
- Will ligate the ends of duplex DNA or RNA.
- This enzyme will join blunt-end termini as well as ends with cohesive (complementary) overhanging ends (either 3' or 5' complementary overhangs).
- This enzyme will also repair single stranded nicks in duplex DNA, RNA or DNA/RNA duplexes. Requires **ATP** as a cofactor.

Taq DNA ligase

- This ligase will catalyze a phosphodiester bond between two adjacent oligonucleotides which are hybridized to a complementary DNA strand:

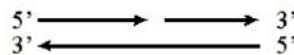


Figure 1.4.7: Taq DNA ligase activity

- The ligation is efficient only if the oligonucleotides hybridize perfectly with the template strand.
- The enzyme is active at relatively high temperatures (45 - 65 °C). Requires **NAD⁺** as a cofactor.

T4 RNA ligase

- Will catalyze formation of a phosphodiester bond between RNA/RNA oligonucleotides, RNA/DNA oligonucleotides, or DNA/DNA oligonucleotides.
- Requires **ATP** as a cofactor.
- This enzyme does not require a template strand.

T4 RNA ligase can be used for a variety of purposes including constructing RNA/DNA hybrid molecules.

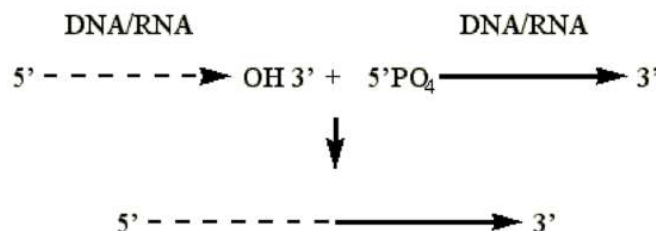


Figure 1.4.8: T4 RNA ligase activity

DNA ligase (E. coli)

- Will catalyze a phosphodiester bond between duplex DNA **containing cohesive ends**.
- It will **not** efficiently ligate blunt ended fragments.
- Requires **NAD⁺** as a cofactor.

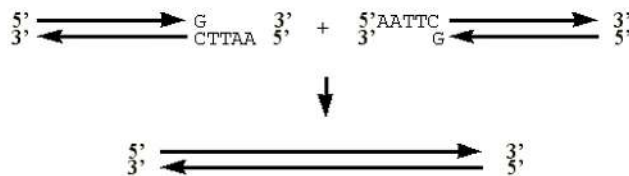


Figure 1.4.9: DNA ligase (*E. Coli*) activity

T4 polynucleotide kinase

- Catalyzes the transfer and exchange of a phosphate group from the **g** position of rATP (adenine ribose triphosphate nucleotide) to the 5' hydroxyl terminus of double stranded and single stranded DNA or RNA, **and** nucleoside 3' monophosphates.
- The enzyme will also remove 3' phosphoryl groups.
- **Oligonucleotides which are obtained from automated synthesizers lack a 5' phosphate group, and thus, cannot be ligated to other polynucleotides.**

T4 polynucleotide kinase can be used to phosphorylate the 5' end of such polynucleotides:

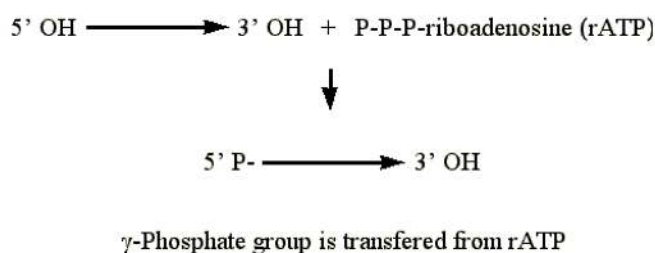


Figure 1.4.10: T4 polynucleotide kinase activity

Calf intestinal phosphatase (CIP)

- Catalyzes the removal of 5' phosphate groups from RNA, DNA and ribo- and deoxyribo- nucleoside triphosphates (e.g. ATP, rATP).
- CIP treated duplex **DNA cannot self ligate**.
- Hemi-phosphorylated duplexes will be ligated on one strand (the phosphorylated strand) and remain "nicked" on the other.

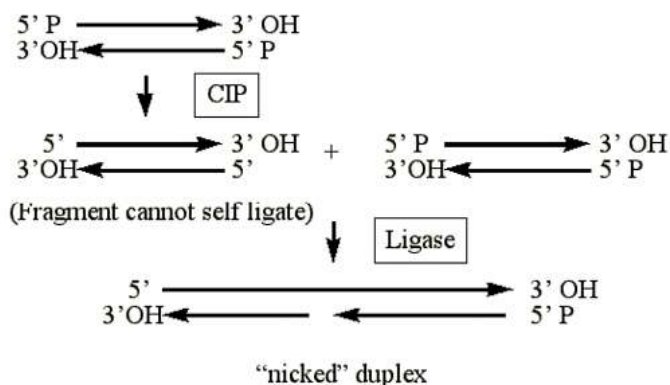


Figure 1.4.11: CIP activity

This page titled [1.4: DNA Modifying Enzymes](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

1.5: DNA Replication - Introduction to Prokaryotic replication

1953

- The DNA model of Watson and Crick suggested how genetic information might be replicated: either strand of the duplex can be used as a template to replicate the sequence information.
- But, was the replication conservative (i.e. the original parental strands remain together after replication) or semi-conservative (one parental strand pairs with one newly synthesized strand)?

The answer for prokaryotic organisms (i.e. lack a true membrane bound nucleus and cellular organelles; e.g. bacteria) came from the 1958 experiment of Meselson and Stahl.

1958

The nitrogen in ammonium salts in culture broth is incorporated into DNA bases. The most common isotope of nitrogen is ^{14}N . However, ^{15}N ammonium salts (a heavier isotope) can also be obtained.

- DNA from *E. coli* cells grown with ^{15}N ammonium salts will have a *higher density* than DNA grown in "normal" (^{14}N) ammonium salts.
- Such DNA will migrate differently on cesium chloride (CsCl_2) equilibrium density gradient centrifugation.
- The more dense DNA will migrate as a lower band (on this type of centrifugation the characteristic migration position is a function of density, and is *independent of DNA length*).

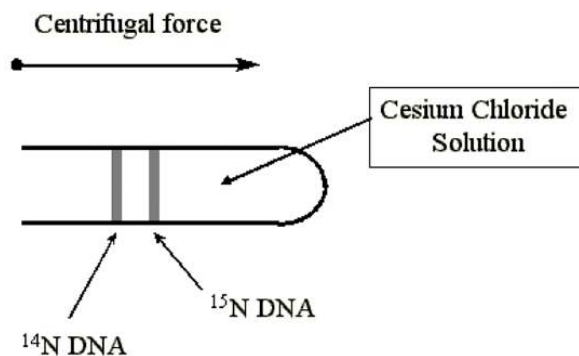


Figure 1.5.1: Density of ^{15}N and ^{14}N DNA

Meselson and Stahl reasoned that if they grew *E. coli* in ^{15}N salts then **switched media to ^{14}N salts** for additional rounds of replication, the mode of replication could be deduced from the density of the DNA.

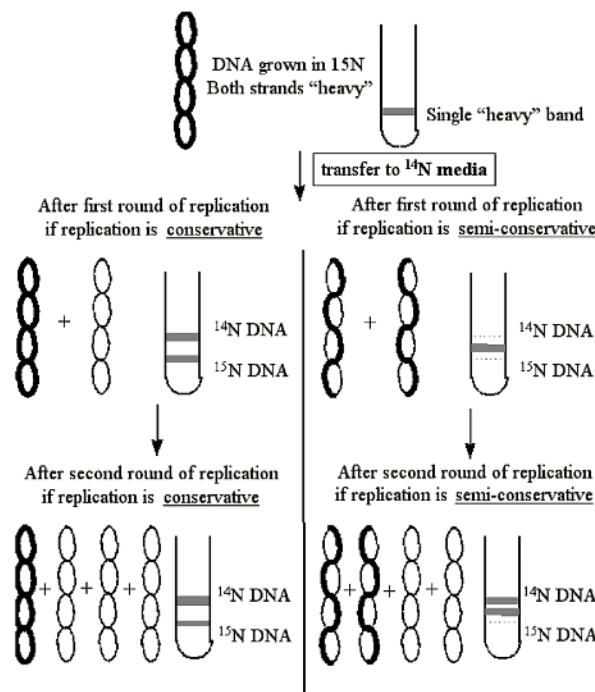


Figure 1.5.2: Meselson and Stahl experiment

- After switching to the ^{14}N media and allowing the cells to go through a round of replication a single band of intermediate density was observed (i.e. between ^{14}N and ^{15}N control DNA samples).
- After a second round of replication in ^{14}N media two bands were present in approximately equimolar amounts; one was intermediate in density and the other migrated as purely ^{14}N labeled DNA.

The results were consistent with a semi-conservative mode of replication for DNA. Evidence of semi-conservative replication of DNA has since been obtained with both plant and animal DNA.

DNA Replication in *E. coli*

E. coli DNA polymerase characteristics:

- Polymerase will only elongate an existing polynucleotide. It **cannot initiate polynucleotide formation**:

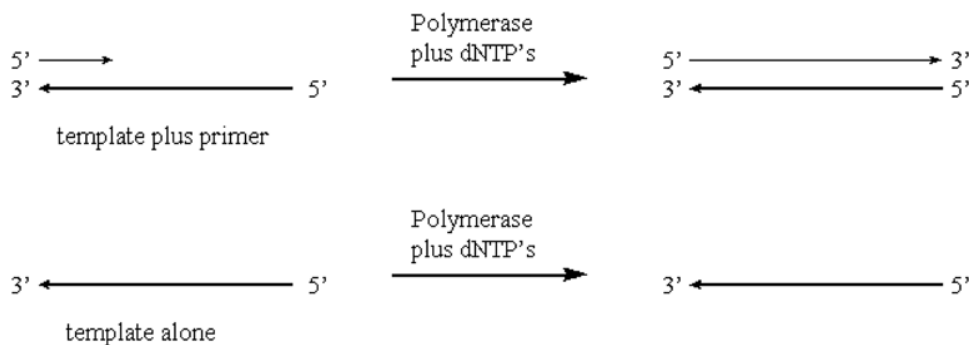


Figure 1.5.3: DNA Polymerase activity

- Polymerase will catalyze polymerization of nucleotides only in one direction ($5' \rightarrow 3'$) via a phosphodiester bond between a 3' hydroxyl and 5' phosphate group.
- DNA polymerase is unable to unwind duplex DNA to separate the two strands which need to be copied

E. coli genome is circular duplex DNA of approximately 4×10^6 base pairs (i.e. 4 Mb)

- The genome has a single origin of replication.
- **DNA duplication in *E. coli* begins at a specific site in the DNA called "oriC".**

OriC is a region of DNA approximately 240 nucleotides long.

- It contains repetitive 9-base pair and 13-base pair sequences (known as the '9-mer' and '13-mer' regions).
- These sequences are AT rich regions, which *melt at lower temperatures* than DNA containing GC pairs.
- These regions are postulated to help melt the DNA duplex in the oriC region for initiation of DNA replication.

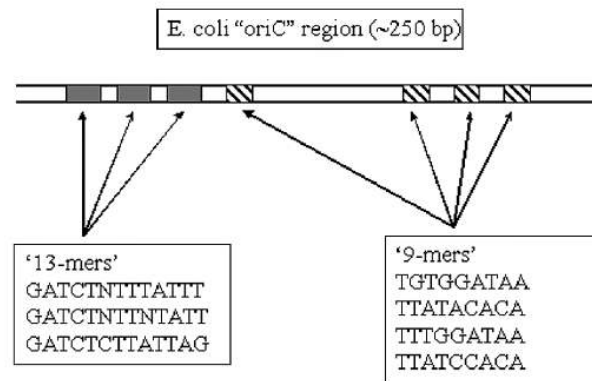


Figure 1.5.4: oriC region

The **dnaA** gene product: (dnaA protein)

- Strains of *E. coli* with mutations in the *dnaA* gene were able to grow at 30 °C, but not at 39-42 °C.
- However, if DNA synthesis was begun at 30 °C, and then the temperature was shifted to 42 °C, DNA synthesis continued until the genome was replicated (and the cell divided), but no new initiation of DNA synthesis was possible.

Conclusion: *Somehow the product of the dnaA gene (i.e. the dnaA protein) is required for initiation of DNA synthesis.*

Studies of purified dnaA protein:

- dnaA protein binds to the '9-mer' region in oriC and forming a multimeric complex with 10-20 protein subunits (i.e. at a single oriC region there will be bound 10-20 dnaA protein molecules).
- Binding requires ATP.
- Further addition of ATP was observed to result in a *melting and opening up* of the DNA duplex in the oriC region. This was determined by addition of *S1 nuclease* (like mung bean, but will also cut DNA at the site of an internal nick), which resulted in cleavage of DNA at the site of oriC.

The **dnaB** gene product: (dnaB protein)

The protein encoded by the *dnaB* gene appears to be essential for DNA replication. The *dnaB* protein has been identified as a helicase. A helicase moves along a DNA strand opening up the duplex to melt and separate the DNA strands.

- dnaB protein binds to the single stranded DNA in the general region of the oriC DNA segment.
- Binding requires ATP as well as the *dnaC* gene product (the *dnaC* protein).
- After helicase/*dnaC* binds to the DNA, the *dnaC* protein is released.
- Two helicases bind at the oriC region, one helicase on each strand of the DNA.

This stage represents the prepriming complex:

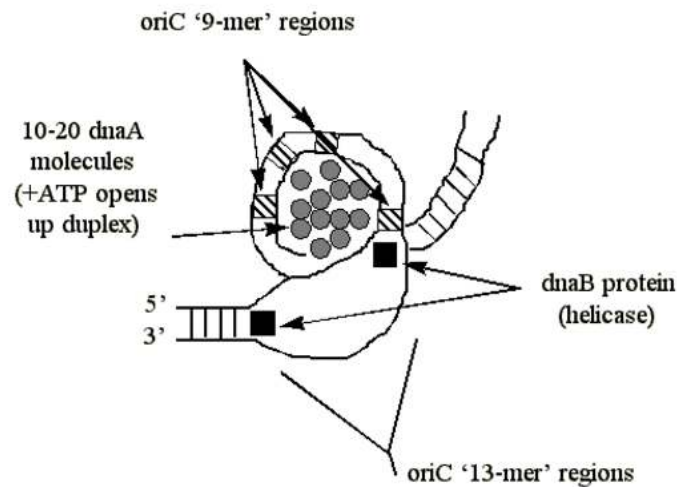


Figure 1.5.5: Prepriming complex

Separated strands in the oriC region are prevented from reannealing by the binding of *single-stranded binding protein* (ssb protein).

The dnaG gene protein:

The dnaG gene protein is called **primase**.

- Primase catalyzes synthesis of short RNA molecules that function as primers for DNA synthesis by *E. coli* DNA polymerase III (pol III).
- Primase binds to dnaB protein at oriC and forms a **primosome**.
- The primase within the primosome complex provides RNA primers for synthesis of both strands of duplex DNA.
- **Primase lays down tracks of pppAC(N)₇₋₁₀** (RNA).

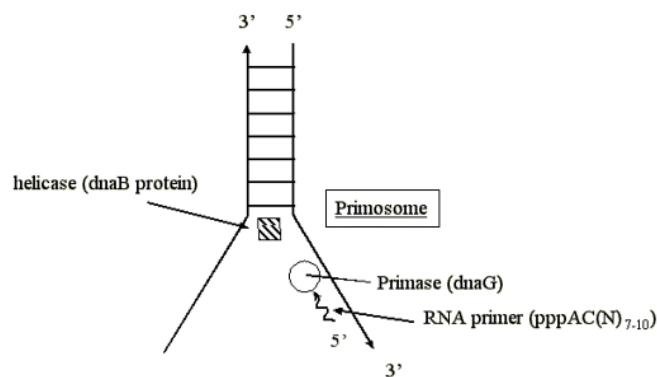


Figure 1.5.6: Primase activity

- After synthesis of the 9-12 mer RNA primer, DNA Pol III holoenzyme enters the replication fork and is able to utilize the RNA as a primer for DNA synthesis.
- As the replication fork opens up, the leading strand synthesis can continue, but **a gap develops in the lagging strand**:

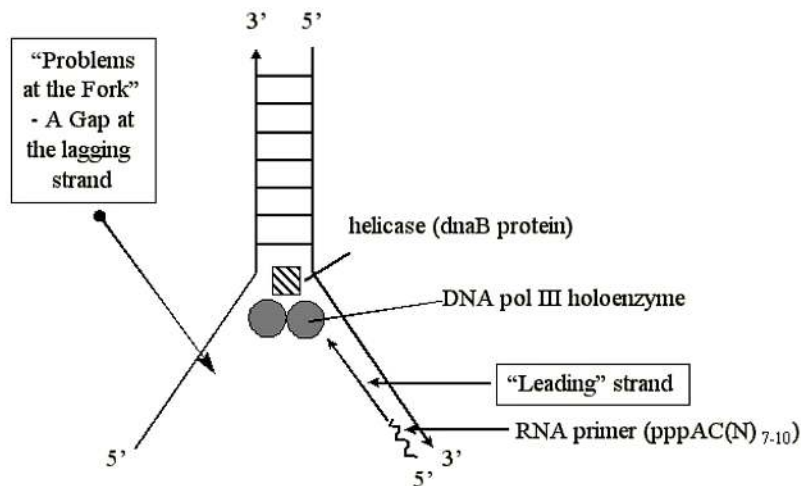


Figure 1.5.7: Lagging Strand

DNA Pol III is a large multicomplex enzyme (holoenzyme) which is somewhat dimeric in nature (there are two polymerase active sites). The two active polymerase sites in Pol III could actually function to synthesize both nascent strands at the fork. However, the synthesis of the lagging template strand would be in the opposed direction to the movement of the Pol III complex:

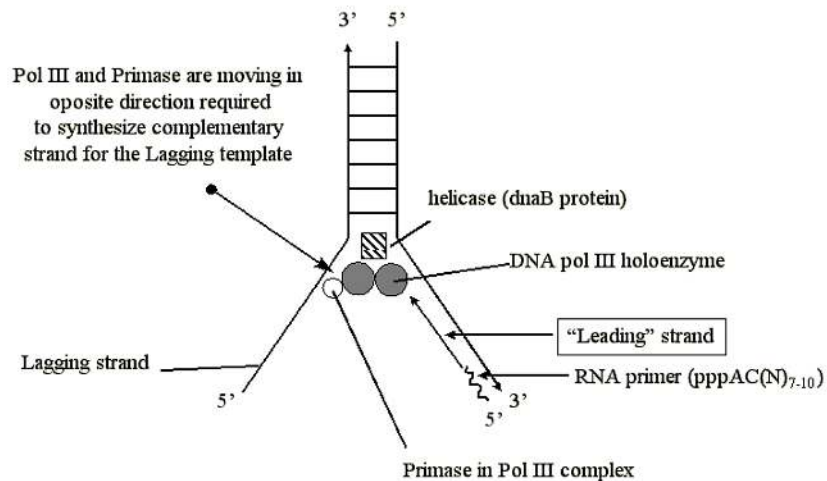


Figure 1.5.8: Pol III movement

Primase can bind to the Pol III complex, but the arrangement of the DNA strand as it passes through the Pol III/primase complex is quite unique. ***It forms a loop structure such that primase and the Pol III active site can accomplish discontinuous synthesis of the lagging template strand even though the general direction of the Pol III complex is opposite to the require direction of DNA synthesis:***

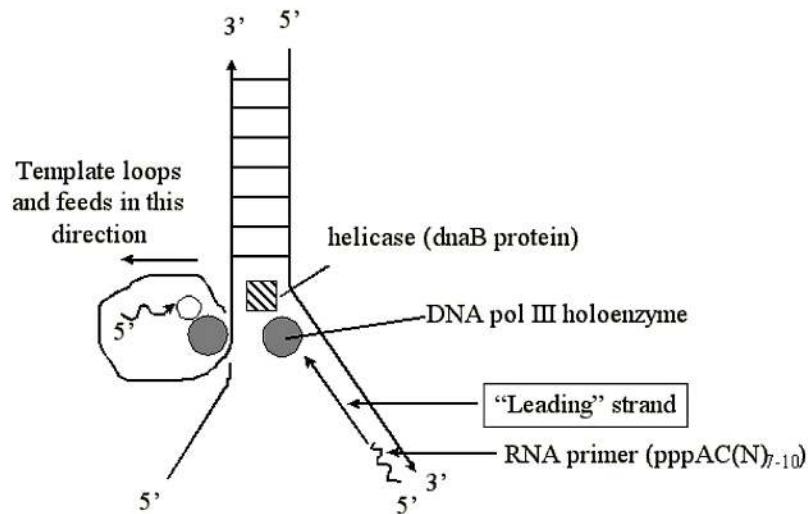


Figure 1.5.9: Loop for Pol III activity

After primase makes another primer on the lagging template, the adjacent Pol III active site can extend the primer (incorporating dNTP's) by utilizing the same loop structure and feeding the template through in the direction shown.

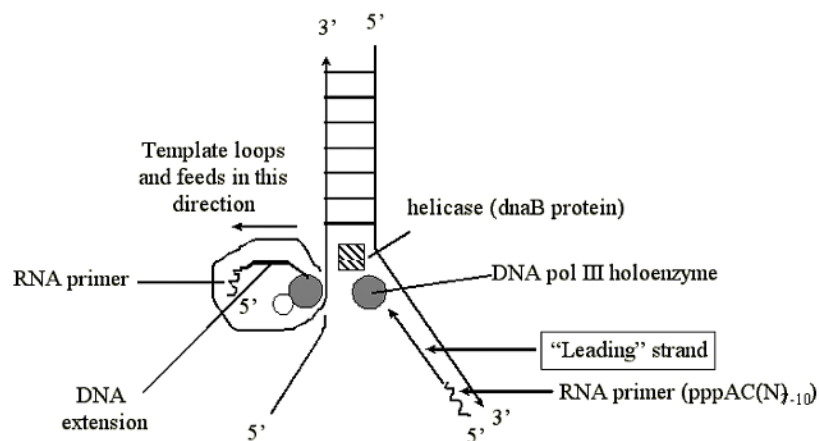


Figure 1.5.10: Synthesis of lagging strand

The lagging strand loop cannot be fed through the Pol III complex forever, and after a nascent DNA strand is synthesized the loop is released and a new one is formed using the opened template DNA further up the fork:

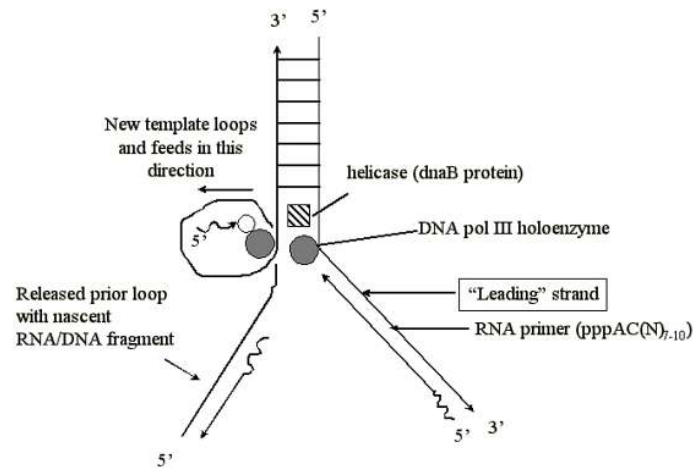


Figure 1.5.11: Continuation of lagging strand synthesis

As synthesis continues:

- there will be a single *continuous DNA strand* on the *leading strand*
- there will be a series of *short fragments* on the *lagging strand*, containing both RNA and DNA, called **Okazaki fragments**:

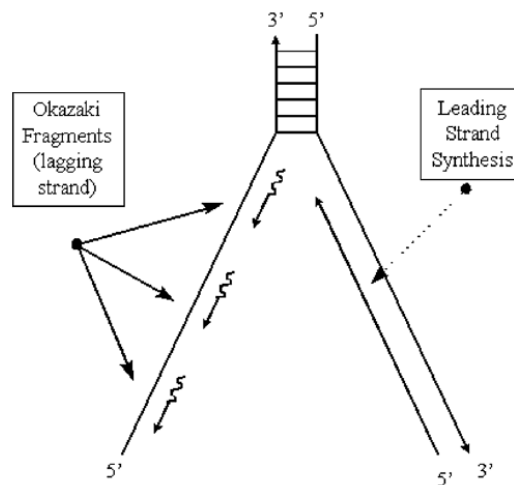


Figure 1.5.12: Okazaki Fragments

How are these RNA/DNA fragments converted into one long continuous DNA strand? The RNA could be removed by a polymerase which has 5'→3' exonuclease activity, however, *Pol III* lacks this activity.

- *DNA Pol I* does have 5'→3' exonuclease activity
- it can extend the DNA synthesis via nick-translation.
- The nick-translation activity results in degradation of the RNA primers.
- The end result is a series of "nicks" in the lagging strand, now *100% DNA*:

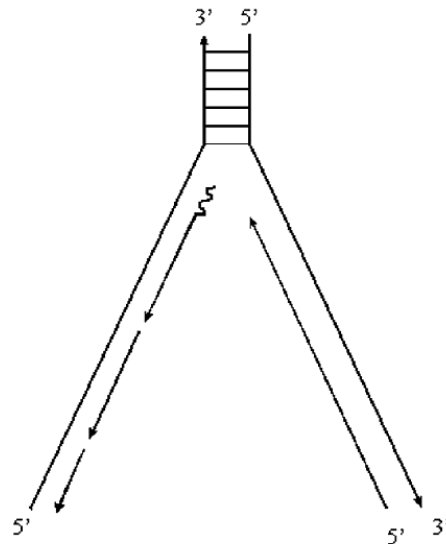


Figure 1.5.13: Nicks in lagging strand

- DNA Pol I leaves and DNA ligase then joins these discontinuous DNA fragments to form a continuous DNA duplex on the lagging strand.

Summary of steps in *E. coli* DNA Synthesis

1. dnaA protein melts duplex in oriC region.
2. dnaB (helicase), along with dnaC and ATP binds to replication fork (dnaC protein exits).¹ (**Pre-priming complex**)
3. Single strand binding protein (ssb protein) binds to separated strands of DNA and prevents reannealing.
4. Primase complexes with helicase, creates RNA primers (pppAC(N)₇₋₁₀) on the strands of the open duplex² (Primase+helicase constitute the **Primosome**).
5. After making the RNA primers, **DNA pol III holoenzyme comes in** and extends the RNA primer (laying down dNTP's) on the leading strand.
6. As the replication fork opens up (via helicase + ATP action) leading strand synthesis is an uninterrupted process, the lagging strand experiences a **gap**.
7. The gap region of the lagging strand can wind around one active site unit of the Pol III complex, and bound Primase **initiates an RNA primer in the gap region**³.
8. On the lagging strand, Pol III extends the RNA primer with dNTP's as the lagging template strand is looped through the Pol III complex
9. After synthesis of a nascent fragment the lagging strand loop is released and the single strand region further up near the replication fork is subsequently looped through the Pol III complex.
10. Steps 7-9 are repeated.
11. Meanwhile, Pol I removes the RNA primer regions of the Okazaki fragments via 5' to 3' exonuclease activity (nick translation)
12. Pol I exits and ligase joints the DNA fragments (on lagging strand).

Notes From Above

1. Polymerases are unable to open up duplex DNA, thus the requirement for helicase
2. Polymerases cannot replicate a DNA template in the absence of a primer (either DNA or RNA).
3. Polymerases extend a polynucleotide in the 5' to 3' direction only. Gaps at the 5' end must be filled by "upstream" discontinuous synthesis.

Properties of *E. coli* polymerases (Pol I, II and III)

	DNA Pol I	DNA Pol II	DNA Pol III
5'→3' Polymerase activity	*	*	*
3'→5' Exonuclease activity(proof reading)	*	*	*
5'→3' Exonuclease activity(nick translation)	*		
<u>Synthesis from:</u>			
Duplex DNA			
Primed single strand	*		
Primed single strand plus ssb protein	*		*
Chain elongation rate(<i>in vitro</i>) bp/min	600	?	30,000
Molecules/cell	400	?	10-20
Mutation Lethal?	*		*

Pol Functions:

- Pol I: gap filling during DNA synthesis and repair, removal of RNA primers
- Pol II: involved in DNA synthesis of damaged templates
- Pol III: functional polymerase at the replication fork

Pol III Structure and function

- A "holoenzyme" complex of 10 different polypeptides
- resultant molecular weight is greater than 600 KDa (i.e. it is a large complex).
- It is structurally an asymmetric dimer - it contains two copies of most of the polypeptides which comprise it, including two catalytic sites for nucleotide addition (i.e. polymerization).

The various protein subunits have a variety of functions:

1. Subunits for polymerase activity: α , ϵ , subunits
2. Subunits to dimerize the core polymerase (τ)
3. Subunits to increase processivity (i.e. to increase the ability to synthesize long stretches w/o releasing from the DNA template): θ subunits
4. Subunits to bind θ to DNA-primer substrate: (γ , δ , δ' , ϵ ,)

Termination of DNA replication

- Specific termination sites of DNA replication exist in *E. coli*.
- Termination involves the binding of the *tus* gene product (*tus* protein).
- This protein may act to prevent helicase from unwinding DNA (will therefore halt *pol III* and *pol I* action).
- DNA replication produces two interlocking rings which must be separated.
- This is accomplished via the enzyme *topoisomerase*.

ColE1 Plasmid

E. coli can contain a small extrachromosomal element called the ColE1 plasmid. This plasmid has the following general features:

- 6.4 Kb circular duplex DNA
- Autonomously replicating

- 10-15 copies per cell (i.e. per *E. coli* chromosome)

Although it is autonomously replicating, it does not contain an *oriC* type of sequence for initiation of replication, and does not undergo the same steps in replication.

- The plasmid produces (among other things) two RNA oligonucleotides (RNA I and RNA II)
- RNA II has complementarity to the ColE1 origin of replication, which contains an AT rich sequence
- The bound RNA II molecule can serve as a primer for polIII
- Since RNA II binds on one strand only, the replication of the ColE1 plasmid is **unidirectional**
- RNA I has complementarity with RNAII, and such a duplex RNA cannot serve as a replication primer, thus control of the plasmid **copy number** is achieved by the interaction between RNA I and RNA II.

This page titled [1.5: DNA Replication - Introduction to Prokaryotic replication](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

1.6: DNA Supercoiling and Topoisomerases

Unwinding of the helix during DNA replication (by the action of helicase) results in supercoiling of the DNA ahead of the replication fork.

- This supercoiling increases with the progression of the replication fork.
- If the supercoiling is not relieved, it will physically prevent the movement of helicase.

The topology of DNA can be described by **three** parameters:

1. **Linking Number (L)** An integer value. "Positive" is referenced as right-handed.
2. **Twist (T)** A real number (the "apparent" linkage number)
3. **Writhe (W)** A real number ("supercoils" in the DNA structure)

Consider closed circular DNA:

- **Linking number** is an integer value.
- It refers to the number of times the two strands of the duplex make a complete 360 degree turn.

For circularly closed DNA, like the *E. coli* genome, the linking number can only be changed if we do the following:

1. physically break the duplex
2. introduce (or remove) a 360 degree turn
3. ligate (covalently close) the break.

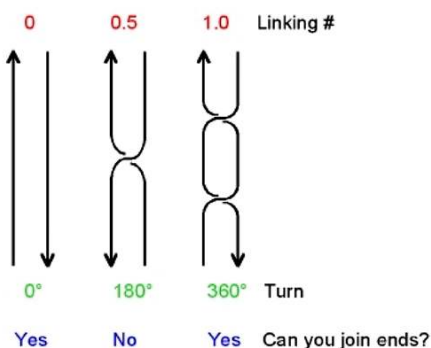


Figure 1.6.1: Turns to change the linking number

Rubber tubing "helix" experiment

Cut two lengths of 1/8" rubber tubing, each about 20" long. Insert a smaller piece of tubing, or piece of pipette tip in the ends to allow the ends to be connected. These two pieces of rubber tubing represent each strand of a DNA duplex. DNA can be ligated, or joined, when we have 5' (phosphate) and 3' (hydroxyl) ends. So we need a way to keep track of which end is which for each piece of tubing.

1. You can either write "5" and "3" on opposite ends of each piece and align them in opposite directions, or
2. On one piece of tubing, mark both ends with a sharpie. Only similarly colored ends can be ligated (see diagram above)

This will allow you to maintain correct strand "orientation" when you "ligate" the strands of the duplex.

- Introduce a Linking number = +2 (two 360° right handed twists into the duplex)
- then "ligate" the ends (make sure you maintain strand orientation, i.e. you connect the appropriate two strands).

Confirm that the correct linking number has been introduced by "melting" the duplex on one side and forcing all turns into a small region of the duplex (easy to count this way).

- Confirm that looking down the helix at the turns that they are "**right handed**" (does not matter which way you look down the helix).

Note that the "duplex" when held between thumb and forefinger and allowed to hang, prefers a "supercoiled" topology, as opposed to "relaxed" [Note: this is usually seen with very skinny tubing, larger diameter tubing may not readily adopt this topology].

- Confirm that the "supercoils" are actually **left handed** as you look down the supercoils (regardless of direction down the supercoils).
- The "supercoil" is most likely a full 360°, rather than 180°. In any case, hold the ends of the duplex so that a left handed 360° "supercoil" is present.
- Now, count how many times the strands of the "duplex" cross each other. In this conformation, the strands of the "duplex" will **not** actually cross each other (Note: you may have one strand crossing and then later uncrossing, for a net result of no crossing).

Thus, in response to the introduction of +2 Linking number, the "duplex" can adopt +2 (180°) "supercoils", such that the resulting **apparent linkage number** (i.e. "twist" value) of the two strands is zero

Note

A supercoil is considered positive, if it is "left handed").

- Remove one "positive" supercoil by unwinding by 180°.
- Now hold the ends of the "duplex" and count the **apparent linkage number** (i.e. "twists"). There will be a single right-handed "twist".
- **Thus, a single 180° "positive" supercoil has the effect of removing a single "positive" twist (i.e. reduces the apparent linkage number by 1).**

The **Writhe** number refers to the number of supercoils present.

- Although it may seem that the consequence of introducing **supercoiling** (Writhe) is changing the Linking number, **it is not**.
- The consequence of Writhe is that the **Twist (apparent linkage number)** is altered (increased or decreased).

DNA has a preferred "**Twist**" value (preferred apparent linkage number) **for a specified length** of DNA:

- Watson and Crick's model of the DNA duplex had 10 basepairs per turn.
- Under physiological conditions of salt (0.15 M NaCl) and temperature, **DNA prefers to adopt about 10.6 bp/turn.**
- **Writhe is introduced in the DNA to achieve this value for the "Twist" (apparent linkage number)**

For a given (fixed) Linkage number over a given length of DNA, the DNA can adopt either positive or negative supercoils to achieve a "twist" (apparent linkage number) such that there will be 10.6 basepairs/turn.

Linkage number does not change with supercoiling (it can only change by breaking the duplex)

- Writhe has the effect of changing the apparent Linkage number.
- One supercoil is defined as being able to change the apparent linkage number by +/- 1.

The twist value (apparent linkage number) for a given length of DNA is related to the number of base pairs per turn that the DNA wants to adopt:

Linkage Number = (size of DNA in base pairs)/(basepairs/turn) + Writhe

or

Linkage Number = #of Twists + Writhe

this is usually abbreviated as

Linkage = Twist + Writhe

L = T + W

For example, if we have a circularly closed DNA molecule with a length of 5300 base pairs, and a preferred conformation of 10.6 basepairs per turn, can it achieve this conformation without having to introduce any supercoiling (i.e. writhe)?

Apparent linkage number (Twist) = (5300 base pairs) / (10.6 base pairs/turn)

Apparent linkage number (Twist) = 500

In other words, in order to achieve the desired conformation of 10.6 bp/turn in the helix, exactly 500 turns are required over the length of 5300 base pairs.

$$\text{Linkage number} = 500 + \text{Writhe}$$

We can have integral values for the linkage number, and we can certainly introduce 500, which would require no Writhe at all:

$$500 = 500 + 0$$

What is the DNA molecule was 5200 base pairs?

$$\text{Apparent linkage number (Twist)} = (5200 \text{ base pairs}) / (10.6 \text{ base pairs/turn})$$

$$\text{Apparent linkage number (Twist)} = 490.6$$

We can introduce either 490 or 491 as a linkage number, but not 490.6. What happens if there is a linkage number of 490 in the DNA molecule?

$$\text{Linkage number} = \text{Twist} + \text{Writhe}$$

$$490 = 490.6 + \text{Writhe}$$

$$\text{Writhe} = -0.6$$

In this case, the DNA adopts a negative 0.6 supercoil (about 108° of a right-handed supercoil) which will increase the apparent linkage number from 490 to 490.6 (and achieve 10.6 basepairs per turn in the duplex).

How many basepairs per turn would there be in the DNA if the DNA was not able to adopt any supercoil structure for this length of DNA with a linkage number of 490?

$$\text{Linkage number} = \text{Twist} + \text{Writhe}$$

$$490 = \text{Twist} + 0$$

$$\text{Twist} = 490 \text{ turns}$$

$$\text{Twist} = (5200 \text{ base pairs}) / (\text{bp/turn}) = 490 \text{ turns}$$

$$\text{bp/turn} = 5200 \text{ base pairs} / 490 \text{ turns}$$

$$\text{bp/turn} = 10.61$$

There are slightly more than 10.6 basepairs per turn in the DNA

A small circularly closed genome

The Simian Virus 40 (SV40) genome is a circular, closed, double stranded DNA genome. For the purposes of this discussion, it has 5300 bases. We expect that under physiological conditions the DNA will exhibit 10.6 base pairs per turn (i.e. one Twist = 10.6 bp/turn). In this case, with no Writhe, the Linking number would be:

$$\text{Linking number} = 5300 \text{ bp} / (10.6 \text{ bp/turn}) + 0$$

$$\text{Linking number} = 500 \text{ turns}$$

i.e. we would expect 500 360° turns of the DNA strands over the length of the circular genome.

- This form (with 10.6 base pairs per turn) with no Writhe represents the "standard", or undistorted, DNA helix.
- This is also known as the "relaxed" form of DNA, and the duplex could physically be laid out **flat** on a surface because it needs no Writhe to achieve the preferred value of 10.6 basepairs per twist:

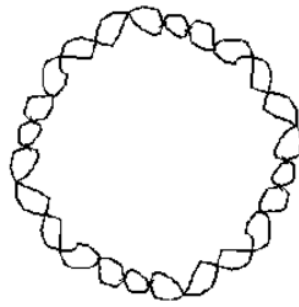


Figure 1.6.2: Standard DNA helix

However, when the replication of SV40 is initially completed it is observed that there remains an open duplex region in the DNA:

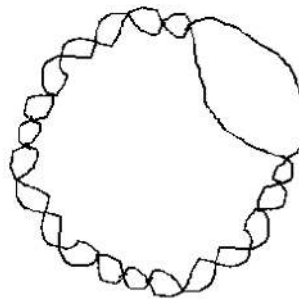


Figure 1.6.3: DNA with open region

The result is that there are about 475 turns of the helix within the duplex DNA (i.e. the Linking number = 475).

- The DNA is said to be **underwound**.
- An open area is energetically unfavorable.
- The covalently closed molecule **cannot** adjust for this by increasing the Linking number. That is, it cannot **spontaneously** break one or both strands of the duplex, introduce another 25 turns into the duplex (increase the Linking number by 25) and re-ligate the duplex.

The DNA has **three** choices:

1. It can adjust the number of **basepairs per turn** throughout the molecule from a desired 10.6 bp/turn to 11.2 bp/turn (i.e. 5300 bp/475 turns). (NOTE: an **increase** in the number of basepairs per turn will **decrease** the twist value; **underwound** DNA has a greater number of basepairs per turn).
2. The DNA can coil up into a "supercoil" topology and maintain the desired twist value (10.6) with the given linking number (475 in this case).
3. The duplex can exist with a twist of 10.6 bp/turn for most of the structure, and then have **a region with zero twist** (not necessarily a melted duplex). This is quite unfavorable due to the geometry required of bond angles.

Thus for the 5300 bp SV40 genome, with a Linking number of 475, to maintain a value of 10.6 bp/twist, a total of 25 negative supercoils (Writhe=25) are needed:

$$475 = (5300/10.6) + \text{Writhe}$$

$$-25 = \text{Writhe}$$

- That is, 25 **negative supercoils** (twenty five 180° turns of the DNA duplex, **right handed** as you look down the supercoiling).

Topoisomerases

The enzymes that control DNA topology are critical to DNA replication and transcription.

- As the replication fork opens up, the region of the duplex in front of the fork becomes **overwound** - i.e. it has *fewer basepairs per turn*.

- The linking number has not changed, but the length of DNA which contains all the turns is effectively shorter.
- To maintain 10.6 bp/turn in that region, *the DNA will adopt positive supercoils.*

For example, during the early stages of SV40 replication, the duplex around the origin of replication may initially melt (open up) a region of 750 bases. Since the Linkage number (500) is unchanged, it is effectively distributed over only:

$$5300 - 750 = 4550 \text{ bases.}$$

Assuming no supercoiling has been introduced:

$$500 = 4550 \text{ basepairs} / (X \text{ basepairs/twist}) + 0$$

$$= 9.1 \text{ basepairs/twist}$$

Thus, if no supercoiling is introduced, *the DNA must adopt a conformation of 9.01 base pairs/twist of the helix within the region ahead of the replication fork.*

- This is energetically unfavorable, and one option for the DNA is to adopt a **supercoiled** configuration to achieve 10.6 bp/twist:

$$500 = (4550/10.6) + \text{Writhe}$$

$$70.8 = \text{Writhe}$$

- *Thus, movement of the growing fork causes the DNA to adopt positive supercoils.*
- In this case the DNA has adopted 70.8 left handed supercoils (180° each).
- *Twist (=basepairs * [twist/basepair]) and Writhe are both real numbers.*

Type I Topoisomerase

- Type I topoisomerases cut one strand of the DNA (i.e. it "nicks" the DNA duplex).
- The 5' phosphate of the nicked strand is covalently attached to a tyrosine in the protein.
- The 3' end of the nick then passes once through the duplex.
- **The nick is then resealed, and the linkage number is change by a value of +1.**
- This can therefore result in the removal of a single negative supercoil.

In *E. coli*, type I topoisomerase *can only relieve negatively supercoiled DNA* (negative supercoiling is the end result of newly replicated DNA genome). In eukaryotes, type I topoisomerase can also relieve positively supercoiled DNA.

The *net result* of *E. coli* topoI can be diagrammed as follows:

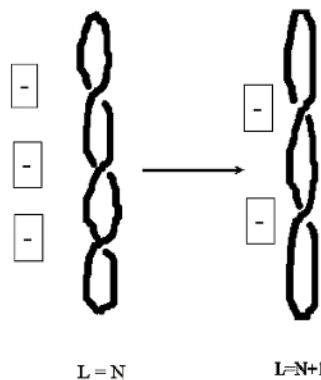


Figure 1.6.4: *E. Coli* Topoisomerase I

Type II Topoisomerases

- Type II topoisomerases actually cleave the duplex DNA in changing the linkage number.
- Type II topoisomerases can *convert a single positive supercoil into a negative supercoil.*
- **Thus the linkage number is reduced by two (-2) in a single step.**
- Type II topoisomerases are involved in both decatenation of daughter chromosomes, and **relieving the positive supercoiling ahead of the replication fork.**

- *E. coli* DNA **gyrase** is an example of a type II topoisomerase.

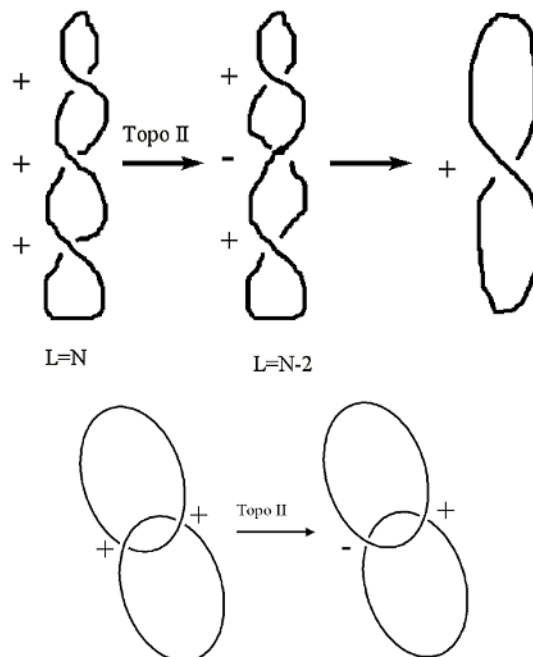


Figure 1.6.5: Type II Isomerase activity

This page titled [1.6: DNA Supercoiling and Topoisomerases](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

CHAPTER OVERVIEW

2: Bacteria

- 2.1: The Bacteria - An Introduction
- 2.2: Bacterial Growth and Reproduction
- 2.3: Extrachromosomal elements, plasmids, selectable markers
- 2.4: Central Dogma and the Genetic Code
- 2.5: Gene and Operon
- 2.6: The lac Operon; CAP site; DNA footprinting
- 2.7: Transcriptional Regulation: Transcription termination, the trp operon
- 2.8: mRNA Translation

This page titled [2: Bacteria](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

2.1: The Bacteria - An Introduction

Bacteria evolved some 600 million years ago, and were probably responsible for the production of the earth's atmosphere (cyanobacteria). Bacteria were discovered in the 17th century after the development of the microscope.

- Single cell organism
- Widely dispersed in the environment
- Invisible to naked eye, but discernible by their actions - milk sours, wounds become septic, meat putrefies, etc.
- *Prokaryotic* type cells (other organisms are *eukaryotic* type cells)

Major differences between prokaryotic and eukaryotic cells

Prokaryotic cells	Eukaryotic cells
No nuclear membrane: chromosome(s) in direct contact with cytoplasm	Chromosomes are enclosed in a double layered nuclear membrane
Simple chromosome structure	Complex chromosome structure; DNA associated with histone proteins
Cell division does not involve meiosis	Cell division involves mitosis and meiosis
If present, cell walls contain <i>peptidoglycan</i> , no cellulose or chitin	If present, cell walls contain <i>cellulose</i> or <i>chitin</i> , never peptidoglycan
No mitochondria or chloroplasts	Mitochondria usually present, chloroplasts in photosynthetic cells
Cells contain ribosomes of only one size	Cells contains two types of ribosomes, one in cytoplasm, and smaller type in mitochondria
Flagella, if present, have a <i>simple</i> structure	Flagella, if present, have <i>complex</i> structure

Note: bacteria are *microorganisms*, but not all microorganisms are bacteria. Algae, fungi, lichens, protozoa, viruses and subviral agents are all microorganisms (with most of these being eukaryotic type cells)

Bacterial activities

Pathogenic (disease causing) bacteria:

- *Cholera* (vomiting, profuse diarrhea)
- *Botulism* (muscle paralysis)
- *Tetanus* (uncontrollable contractions of skeletal muscle)
- *Staphylococcal* food poisoning (vomiting, diarrhea)
- *Shiga toxin* (verotoxin) (classic dysentery)
- *Typhoid* (septicemia: bacteria in blood, destruction of host tissue)
- *Oroya fever* (*Bartonella bacilliformis* - destroys red blood cells)
- *Endotoxic shock* (lipopolysaccharide cell wall component causes release of host inflammatory agents leading to shock and death)
- *Reactive arthritis* (response in some people to *Salmonella* infection)

Most bacteria do no harm to humans, and can be quite useful:

- Antibiotic production
- Enzyme additives for detergents
- Insecticides
- Production of biodegradable plastics
- "Biomining" - leaching of metals from low grade ores
- Uses in food industry
 - Butter
 - Cheese
 - Yogurt
 - Vinegar
 - Cocoa
 - Coffee
- Soil fertility

Classifying and Naming of bacteria

Differences between bacteria can include

- shape
- size

- structure
- chemical activities
- required nutrients
- form of energy required
- required environment
- reaction to certain dyes

Family, genus, species, strain

Bacteria in the same *Family*, in general would have:

- similar structure
- use the same form of energy
- react similarly to certain dyes

Bacteria in the same *Family* may be divided into different *Genera* based on differences in

- chemical activities
- nutrient requirements
- conditions for growth
- shape and size (to some extent)

Strains of bacteria are bacteria of the same species, but with some subtle difference (maybe a single mutational difference)

Latin binomial name

- Name of the genus, capitalized.
- Followed by name of species, lower case
- Italicized. Sometimes the genus name is abbreviated to a single letter (with a period)

Escherichia coli
E. coli

- Strain name follows, usually in parentheses. In the vernacular, the strain name is commonly used to identify the bacteria

E. coli (JM101)
E. coli (CJ236)
"Can I borrow some JM101?"

Some Characteristics of Bacteria

Shape

- Rounded or spherical cells - *cocci* (singular: *coccus*)
- Elongated or rod-shaped cells - *bacilli* (singular: *bacillus*)
- Rigid spirals - *spirilla* (singular: *spirillum*)
- Flexible spirals - *spirochetes* (singular: *spirochete*)

Note

There is a genus of bacteria called Bacillus. Some bacillus shaped bacteria belong to the genus *Bacillus*, some do not.

Size

- Bacteria are usually measured in micrometers (1×10^{-6} m)
- The smallest bacteria are about 0.2 micrometers (*Chlamydia*)
- The largest bacteria are about 600 micrometers (*Epulopiscium fishelsoni*. - inhabits the gut of a fish)
- "Average" bacteria are 1-10 micrometers (note: limit of resolution of the light microscope is about 0.2 micrometers)

A "generalized" bacterium:

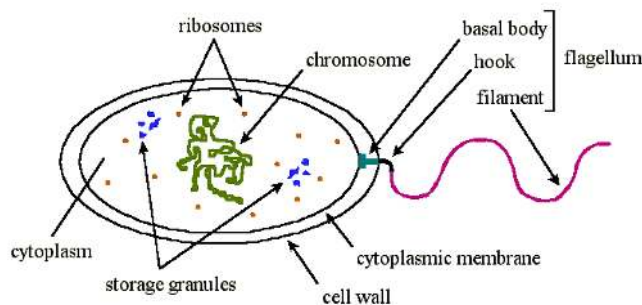


Figure 2.1.1: General bacterium diagram

- The cell's DNA is extensively folded to form a body called the *nucleoid*
- The *cytoplasm* fills the interior of the cells, and bathes the nucleoid
- *Storage granules* contain a reserve of nutrients - typically polymeric forms of b-hydroxybutyrate and phosphate. Poly-b-hydroxybutyrate is the basis of a biodegradable plastic (Biopol)

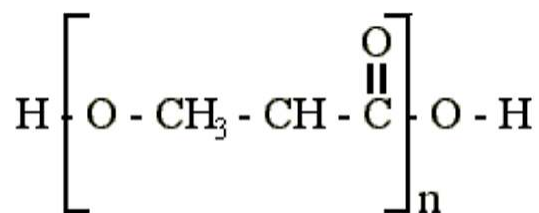
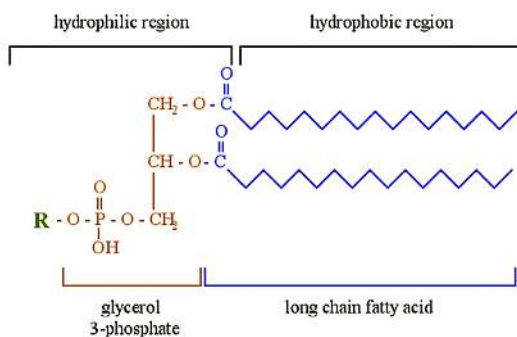


Figure 2.1.2: Poly-b-hydroxybutyrate

- The nucleoid, ribosomes, cytoplasm and storage granules are bounded by a membranous sac, the *cytoplasmic membrane* (*cell membrane*, or *plasma membrane*)
- The outermost layer is a tough *cell wall*. Together, the plasma membrane and cell wall are called the *cell envelope*
- The region between the plasma membrane and the cell wall is called the *periplasmic space*
- The *flagellum* is used for motility

Cytoplasmic membrane

- **lipid bilayer**, 7-8 nm thick, with protein molecules partly or completely embedded
- The inner and outer layers are *hydrophilic*, while the interior of the bilayer is *hydrophobic*
- In *E. coli* the main lipid is phosphatidylethanolamine; minor lipid components include phosphatidylglycerol and diphosphatidylglycerol



R = 'NH₃ - CH₂ - CH₃ in phosphatidylethanolamine
R = CH₂OH - CHOH - CH₃ in phosphatidylglycerol

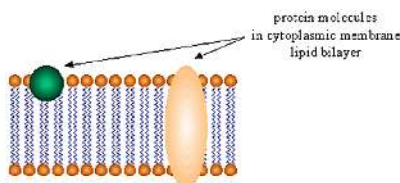


Figure 2.1.3: Phospholipid bilayer

The cytoplasmic membrane proteins include:

- enzymes involved in the synthesis of the cell wall peptidoglycan
- transport proteins (translocated ions and molecules across the cytoplasmic membrane)
- proteins of energy converting systems (ATPases and electron transport chains)
- "sensory" proteins, which detect changes in cell's external environment

The cytoplasmic membrane is not freely permeable to most molecules

- some small uncharged molecules (O_2 , CO_2 , NH_3 , H_2O) can freely pass through
- charged ions typically cannot pass across the membrane, and must be *transported* (with the expenditure of energy)

If the cell wall is removed, what remains of the cell is called the *protoplast*

- can survive (in a test tube) and carry out most normal cell processes
- quite sensitive to osmotic shock - if placed in pure water it will swell (as water enters the cell to balance the osmotic force) and rupture (*osmotic lysis*)
- In an intact cell the cell wall prevents the protoplast from swelling and undergoing osmotic lysis
- The cell wall also determines the shape of the bacteria - all protoplasts are spherical, regardless of the shape of the intact bacteria

The Cell Wall

Among the *Eubacteria* (Kingdom of all bacteria excluding the *archebacteria*, which are typically halophiles and thermophiles) **there are only two major types of cell wall**

- They can be identified by their reaction to certain dyes (characterized by Christian Gram in 1880's):

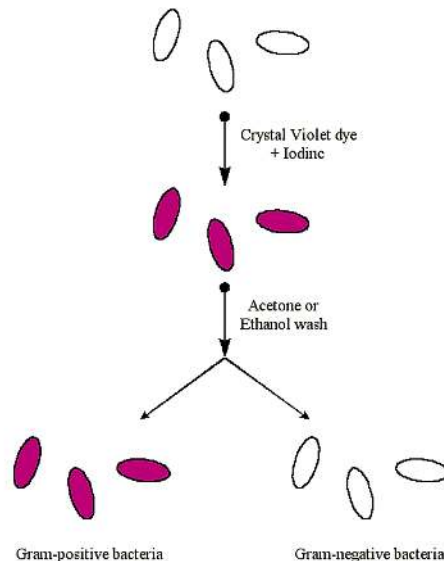


Figure 2.1.4: Gram positive and Gram negative bacteria

Gram positive type cell wall

- relatively thick (30-100 nm)
- 40-80% of the wall is made of a tough complex polymer called peptidoglycan (linear heteropolysaccharide chains cross-linked by short peptides)

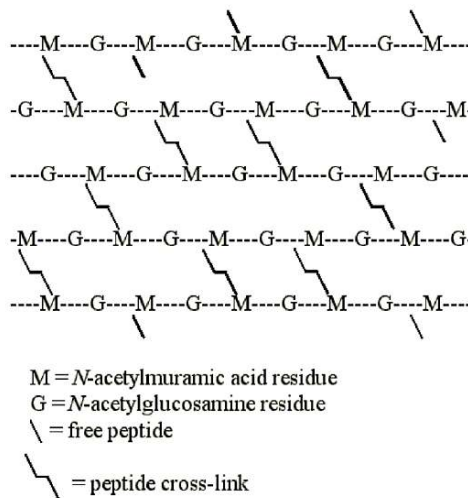


Figure 2.1.5: Gram positive cell wall

- The cell wall of a gram-positive cell is a multi-layered network which appears to be continually growing by the addition of new peptidoglycan at the inner face, with concomitant loss at the outer surface

Gram-negative type cell wall (e.g. *E. coli*)

- thinner than gram-positive type cell wall (only 20-30 nm thick)
- has distinctly layered appearance
- inner region consists of a monolayer of peptidoglycan
- outer layer of cell wall is *essentially a protein containing lipid bilayer*
 - inward facing lipids are phospholipids
 - outward facing lipids are macromolecules called lipopolysaccharides

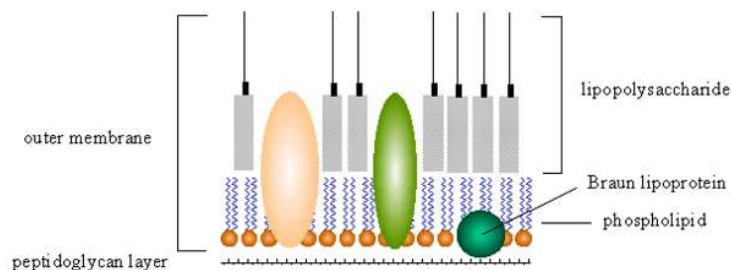


Figure 2.1.6: Gram negative cell wall

- half the mass of the outer membrane consists of proteins
 - the **Braun protein**, which is covalently linked to the peptidoglycan layer
 - transport proteins
 - porins** - molecules which span the outer membrane to create a 'pore' through the membrane. These pores allow certain molecules and ions to pass through the outer membrane (e.g. *ompC*, *ompF* proteins)
- adjacent outer lipopolysaccharides are held together by electrostatic interactions with divalent metal ions (Ca^{2+} , Mg^{2+})
 - the addition of chelating agents (e.g. EDTA) can disrupt these interactions and weaken the outer membrane
- lysozyme (produced by phage lambda, for example) can cleave the saccharide links in the inner peptidoglycan layer

How to lyse a gram-negative bacteria (e.g. *E. coli*):

- Add a chelating agent of divalent metals (e.g. EDTA) to disrupt outer membrane lipopolysaccharides
- Add lysozyme to break up peptidoglycan layer
- cell wall is now structurally weakened and cannot protect the protoplast from osmotic shock
- osmotically shock the cell to disrupt protoplast and release cytoplasmic contents (i.e. high osmotic shock using sucrose solution; low osmotic shock using pure water),
- or use mechanical shear/cavitation (French Press, Menton Gaulin press)

Human microflora: some of the bacteria commonly associated with the human body

Location	Genus
Colon	<i>Bacteroides, Clostridium, Escherichia, Proteus</i>
Ear	<i>Corynebacterium, Mycobacterium, Staphylococcus</i>
Mouth	<i>Actinomyces, Bacteriodes, Streptococcus</i>
Nasal passages	<i>Corynebacterium, Staphylococcus</i>
Nasopharynx	<i>Streptococcus, Haemophilus (e.g. H. influenzae)</i>
Skin	<i>Propionibacterium, Staphylococcus, Others (personal hygiene, environment)</i>
Urethra	<i>Acinetobacter, Escherichia, Staphylococcus</i>
Vagina (adult, pre-menopausal)	<i>Acinetobacter, Corynebacterium, Lactobacillus, Staphylococcus</i>

This page titled [2.1: The Bacteria - An Introduction](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

2.2: Bacterial Growth and Reproduction

Flagella and motility

- *monotrichous* flagella - the bacterial cell has a single flagella
- *peritrichous* flagella - the bacterial cell has several flagella which are located at various sites on the cell surface (e.g. *E. coli*)

Motility is due to the presence of one or more flagella.

- in peritrichous flagellate bacteria the flagella rotate independently of one another
 - 95% of the time the flagella rotate counterclockwise
 - 5% of the time the flagella switch directions and rotate clockwise
 - When the flagella are all rotating counterclockwise, the flagella are bundled together and the bacteria travels in a straight line (i.e. it swims)
 - When one flagella switches direction the bundle disassociates, and the bacteria tumbles
 - alternative swimming and tumbling results in a three-dimensional random walk
- chemoattractants and repellents can interact with receptor proteins in the cell envelope, which in turn influence the rate of tumbling when the cell is moving in a given direction

Growth and Reproduction

Essential requirements for growth include:

- supply of suitable nutrients
- source of energy
- water
- appropriate temperature
- appropriate pH
- appropriate levels (or absence) of oxygen

Nutrients

Cells need a source of:

- carbon
- nitrogen
- phosphorous
- sulfur
- other trace materials

Although a given bacteria typically uses a limited range of compounds, bacteria as a group can utilize a wide range of compounds as nutrients:

- sugars and carbohydrates
- amino acids
- sterols
- alcohols
- hydrocarbons
- methane
- inorganic salts
- carbon dioxide

Energy

Energy is needed for

- essential chemical reactions

- uptake of nutrients
- flagellar motility

Phototrophic vs chemotrophic bacteria

- *phototrophic* - energy derived from light source
- *chemotrophic* - energy is obtained by processing chemicals from the environment

Water

- 80% of the mass of typical bacteria is water
- Water is needed for growth and reproduction
- Dessication (extreme lack of water) is tolerated to different degrees by different bacteria

Temperature

- Growth proceeds most rapidly at the *optimum growth temperature* for a particular bacteria (and decreases as temperature is raised or lowered from this optimum)
- For any bacteria, there is a *minimum* and *maximum* temperature beyond which growth is not supported

Thermophilic bacteria

- *optimum* growth temperature is $>45^{\circ}\text{C}$
- occur in compost piles, hot springs and ocean floor hydrothermal vents
- *Pyrodictium* have an optimum growth temperature of 105°C

Mesophilic bacteria

- *optimum* growth temperatures between 15 and 45°C
- live in a wide range of habitats
- since the human body is $37\text{--}42^{\circ}\text{C}$, human pathogenic bacteria are *mesophiles*

Psychrophilic bacteria

- optimum at 15°C or below
- minimum temperature of 0°C , or less
- maximum temperature of 20°C
- occur in polar seas

pH

- most bacteria grow optimally near neutral pH (7.0)
- *acidophiles* have an optimum pH which is more acidic (*Thermoplasma acidophilum*, found in hot springs, prefers pH $0.8\text{--}3$, and will not grow at neutral pH)
- *alkalophiles* have an optimum at higher (alkaline) pH ranges (*Exiguobacterium aurantiacum*, found in natural alkaline lakes, prefers pH $8.5\text{--}9.5$)

Oxygen

- bacteria which must have oxygen for growth are termed *obligate aerobes*
- bacteria which can grow only in the absence of oxygen are termed *obligate anaerobes* (e.g. environment which are isolated from the atmosphere; e.g. river mud, and within the intestines, for example)
- bacteria which normally grow in the presence of oxygen, but which can manage to grow in its absence, are termed *facultative anaerobes*
- conversely, bacteria which normally grow anaerobically, but which can manage to grow in the presence of oxygen, are termed *facultative aerobes*

Inorganic ions

All bacteria need low concentrations of certain inorganic ions in order to function, e.g.

- iron for cytochromes (energy metabolism), and certain enzymes

- magnesium for cell wall stability
- manganese and nickel in metabolic enzymes
- high concentrations usually inhibit growth (e.g. salt has been used in the preservation of pork, beef and cod)
- some bacteria, *halophiles*, grow only in the presence of high concentrations of certain salts (e.g. sodium chloride).
Halobacteriaceae grow only in the presence of 3-4 M NaCl. This amount of salt is needed to maintain the structure of the cell wall and internal molecular assemblies (e.g. ribosomes).

Growth in a single cell (e.g. *Escherichia coli* - a gram negative bacillus)

The cycle of events in which a cell grows, and divides into two daughter cells, is called *the cell cycle*.

"Slow Growth"

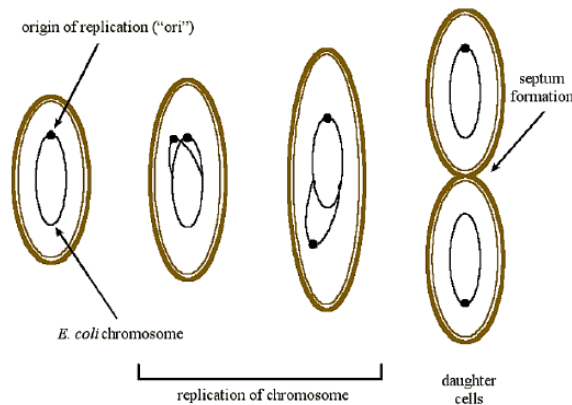


Figure 2.2.1: Replication for slow growth

- Replication begins at a specific place on the chromosome - the *origin* or "ori" region.
- During "slow growth" each new daughter cell contains exactly one chromosome, because a new round of chromosomal replication does not begin until after completion of cell division

The cell division cycle can be thought of as a linear sequence of three periods: *I*, *C* and *D*

- *C* is the period during which chromosomal replication occurs
- *D* is the period in which the septum forms, and *cell division occurs at the end of the D period*
- *I* is the period between each successive initiation of chromosomal replication

For the above type of slow growth, the relationship between these three periods is as follows:

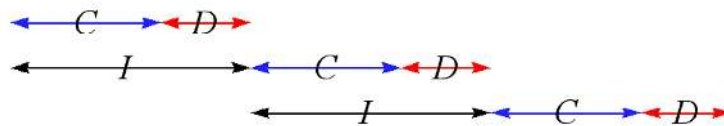


Figure 2.2.2: Periods of cell cycle for slow growth

- *I* is also known as the **doubling time** of the bacterial growth

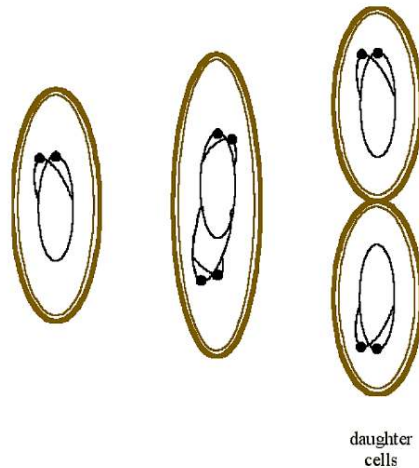


Figure 2.2.3: Replication for rapid growth

- A new round of chromosome replication begins *before* cell division occurs

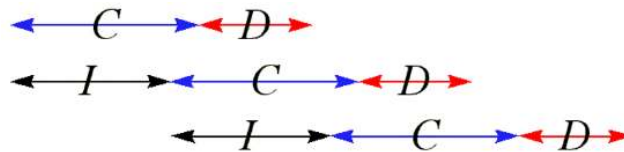


Figure 2.2.4: Cell cycle for rapid growth

- Each daughter cell has the equivalent of about 1 ½ chromosomes
- In rapidly growing *E. coli* cells the *C* period is about 42 minutes and *D* is about 25 minutes.
- The maximum doubling time for *E. coli* is about 20 minutes
- "**medium**": any solid or liquid specially prepared for bacterial growth
- "**culture**": a liquid or solid medium containing bacteria which have grown (or are growing) in or on that medium
- "**incubation**": the process of maintaining a particular temperature (and/or other desirable conditions) for bacterial growth
- "**innoculation**": the initial process of adding the cells to the medium

Growth on a solid medium

- Liquid solution of nutrients plus 1% *agar* (the preferred form of *Jello* in Asia. A polysaccharide extracted from seaweed; as opposed to Western *Jello* which is a protein extracted from the hooves of large farm animals) forms solid media

Typical Solid Media Recipe (1 liter). The media here is commonly called "Luria Broth", or "LB". It is named after one of the scientists who developed it (not "Dr. Broth").

- **Yeast extract** 5 g (nucleic acids, cofactors, inorganic salts, carbohydrates)
- **Tryptic digest of casein** (milk protein) 10g (peptides and amino acids)
- **NaCl** 10g (note: final concentration is thus 0.17 M, or close to *physiological*)
- **Agar** 10g
- **Water** (bring volume up to 1.0 liter)
- Autoclave, pour into petri dish, let cool

An individual bacterial cell will divide and eventually become a visible mass of cells known as a **colony**

If instead of a single cell, the solid media is initially populated with a large number of cells, *confluent growth* or a **lawn** of bacteria will be visible

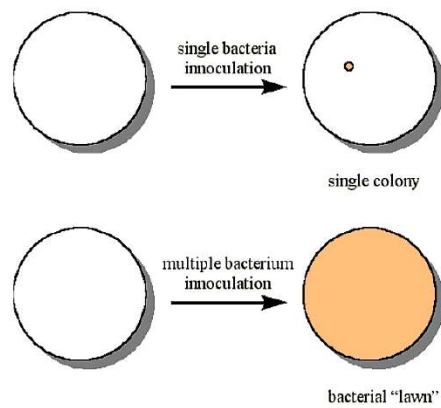


Figure 2.2.5: Growth on solid medium

"Streaking" solid media plates

- bacteria can be introduced onto solid media by a sterile transfer tool, such as a wire loop (nichrome wire), or autoclaved toothpick, which has been dipped into a bacterial culture
- such a transfer contains *thousands* of individual bacterial cells
- it is desirable to grow colonies from individual cells rather than from a large population

This is done to avoid the takeover of the strain by potential wild-type revertants

- "streaking" is a method to isolate individual cells for growth on solid media from an inoculation which originally contains thousands of cells

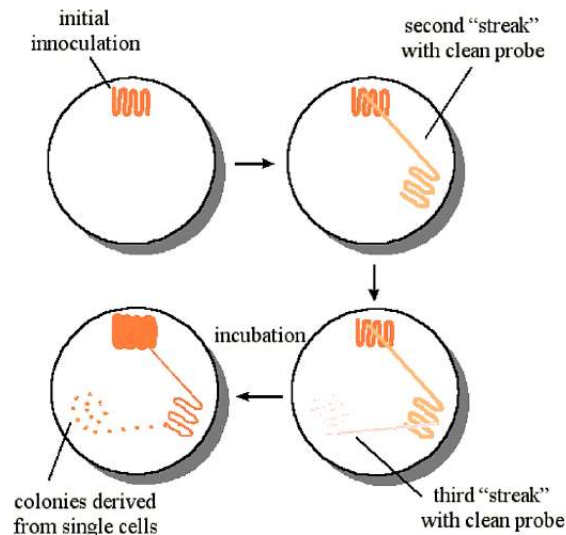


Figure 2.2.6: Streaking

Growth in liquid medium

- No agar
- Use Erlenmeyer flask, or fermenter
- If necessary, aerate, agitate
- progeny will be dispersed throughout medium (diffusion or locomotion)
- as the cell density increases, the media becomes *turbid*
- number of cells plotted versus time will yield a *growth curve*

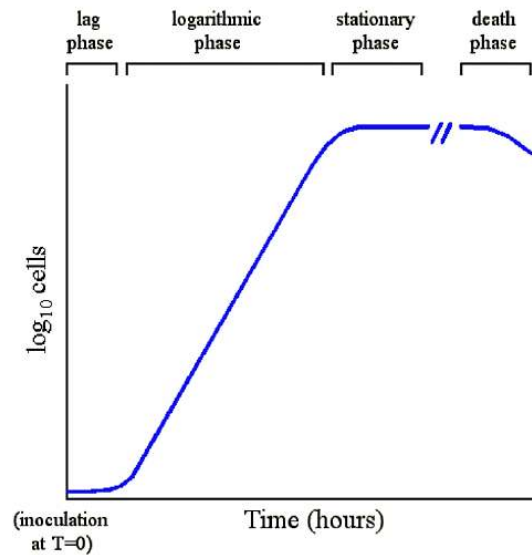


Figure 2.2.7: Growth curve in liquid medium

- "**lag phase**" after inoculation cells are becoming acclimated to the new environment (temp, nutrients, etc.)
- "**log phase**" cells have adapted and are dividing at a constant rate (i.e. the maximum for the species under the given conditions of temp, pH, nutrients, oxygen, etc.)
- "**stationary phase**" cell growth ceases as nutrients are exhausted and/or waste products build up in the media
- "**death phase**" number of viable (living cells) in the stationary phase culture decreases (usually due to toxicity of waste products)

Cell density can be conveniently monitored using the absorbance of visible light (usually at 600 nm)

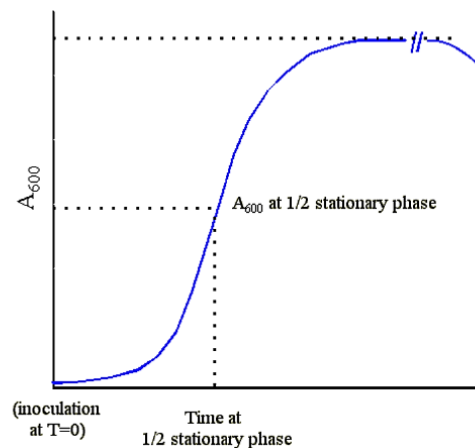


Figure 2.2.8: Cell density curve

Different media will result in different growth rates and different stationary phase densities

- Rich media will have short (<1.0 hour) doubling times and will result in higher cell densities at stationary phase
- Minimal media will exhibit slow growth (doubling times ~1.0 hour at 37°C) and low final densities

- Efficient agitation and aeration can increase final cell densities (fermenters may achieve higher densities than shaker flasks).

	Minimal	LB	Terrific
Phosphate salts	16g	-	12g
Ammonium salts	1g	-	-
Magnesium salts	0.1g	-	-
Glucose/glycerol	4g	-	4g
Sodium Chloride	0.5g	10g	-
Enzymatic digest of casein (milk protein)	-	10g	12g
Yeast extract	-	5g	24g
Approx doubling time (min)	60	45	30
Stationary phase density (A_{600})	3	7	15

Let's take a look at some raw data from a culture of *E. coli* growing in a fermenter. The absorbance at 600nm was recorded at various time intervals:

Time (minutes)	A_{600}
0	0.09
18	0.102
78	0.124
142	0.253
205	0.487
255	1.02
322	1.98
378	3.95
446	5.88
504	6.76
564	7.2

The growth curve looks like this:

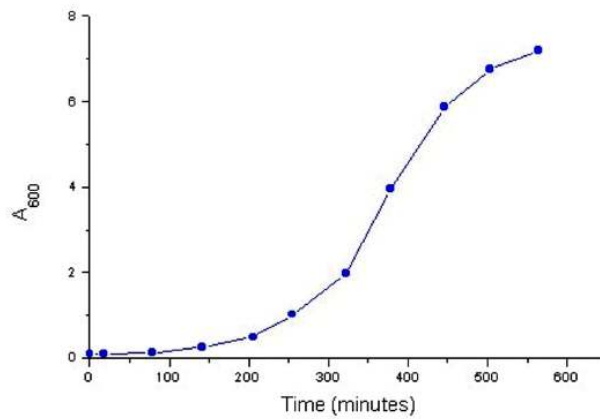


Figure 2.2.9: Sample growth curve

In order to calculate the doubling time we need to know the region of the growth curve for which the growth is *logarithmic*. We can evaluate this by plotting the absorbance data on a logarithmic ($\log_{10}$) scale. In this case, our data will look like this:

Time (minutes)	A ₆₀₀	$\log_{10}$ A ₆₀₀
0	0.09	-1.05
18	0.102	-0.99
78	0.124	-0.91
142	0.253	-0.60
205	0.487	-0.31
255	1.02	0.01
322	1.98	0.30
378	3.95	0.60
446	5.88	0.77
504	6.76	0.83
564	7.2	0.86

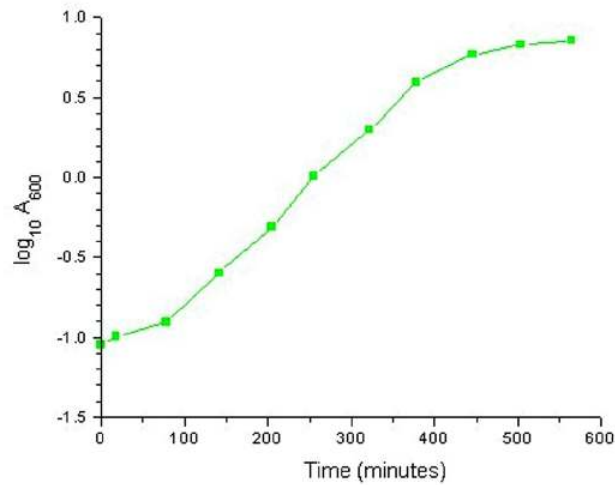


Figure 2.2.10: Sample logarithmic growth curve

In this case, the data appears to have an initial lag period followed by logarithmic growth until about 400 minutes, then the rate slows. In other words, the $\log_{10}$ curve appears linear over the time period 100 - 400 minutes, or so.

We can conveniently determine the doubling time by replotting the data over this period and converting the A_{600} into $\log_2$ values.

Note

$$\log_N(X) = \log_{10}(X) / \log_{10}(N)$$

Time (minutes)	A_{600}	$\log_{10}$ A_{600}	$\log_2$ A_{600}
0	0.09		-1.05
18	0.102	-0.99	
78	0.124	-0.91	-3.01
142	0.253	-0.60	-1.98
205	0.487	-0.31	-1.04
255	1.02	0.01	0.029
322	1.98	0.30	0.99
378	3.95	0.60	1.98
446	5.88	0.77	
504	6.76	0.83	
564	7.2	0.86	

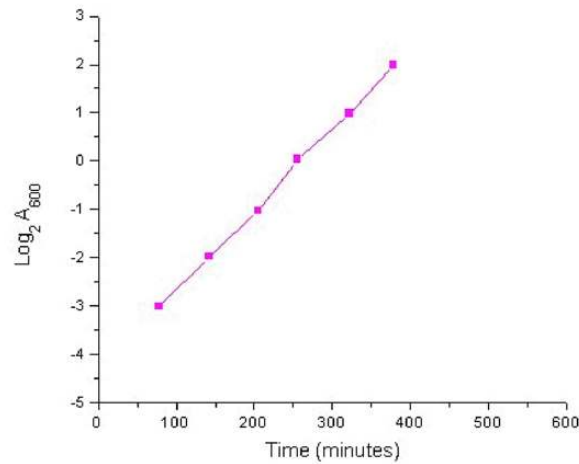


Figure 2.2.11: Sample $\log_2$ growth curve

If we plot a straight line through these points the slope will give us the rate of change of the $\log_2 A_{600}$ values as a function of time:

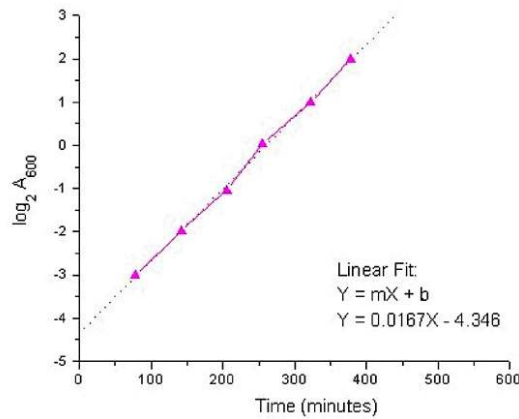


Figure 2.2.12: Linear fit for sample $\log_2$ growth curve

- Thus, for this culture, the growth (slope) can be described as a rate of $0.0167 \log_2 A_{600}/\text{min}$.
- The inverse of the slope will therefore tell us how many minutes it takes for the culture to increase its density by 1 $\log_2 A_{600}$ absorbance units.
- Since it is $\log_2$, a change of 1 absorbance units means the absorbance has doubled.

Therefore, the inverse of the slope for the $\log_2$ plot gives us the time it takes for the culture absorbance to double (i.e. the doubling time)

For this experiment, the doubling time is $1/0.0167$ or **59.9 minutes**. This suggests that the media is probably not very rich (maybe like minimal media), however, the final absorbance is higher than that expected for minimal media. So, possibly the media LB, but the growth was done at a lower temperature (i.e. leading to a slower growth rate).

Important parameters to determine:

- A_{600} at stationary phase
- A_{600} at $\frac{1}{2}$ stationary phase (*usually the time point chosen to stimulate bacteria to produce recombinant proteins*)
- Doubling time (plot $\log_2$ of A_{600} over logarithmic range of growth, take inverse slope)

This page titled [2.2: Bacterial Growth and Reproduction](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

2.3: Extrachromosomal elements, plasmids, selectable markers

- Including an origin of replication (i.e. the *E. coli* **oriC** region) into a circular DNA molecule is a mechanism to have an **extrachromosomal element** in the prokaryotic cell.

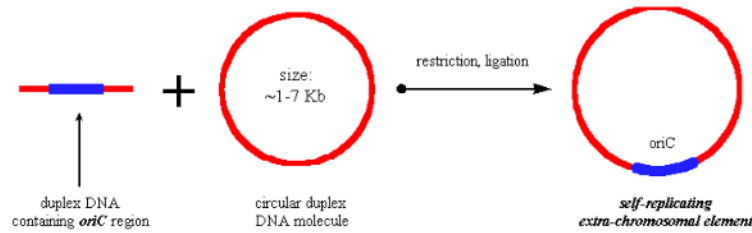


Figure 2.3.1: *oriC* Extrachromosomal element

- Such an extrachromosomal element is called a **plasmid**, or **vector**
- The plasmid uses the host cell machinery (i.e. polymerases, helicases, dNTP's etc.) to direct replication.

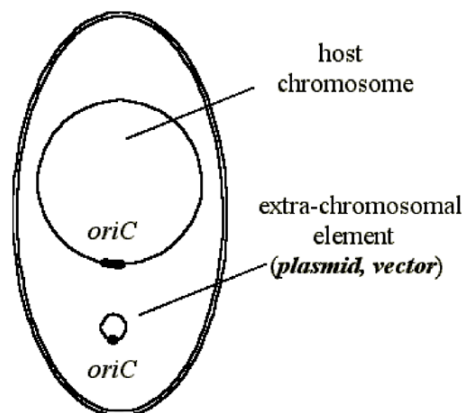


Figure 2.3.2: Vector/Plasmid

- However, since the added work of replicating the extrachromosomal element is a load on a cell, **it will be out-competed by other cells which do not contain the plasmid**.
 - Since in prokaryotic cells the segregation of plasmids is a random event, daughter cells can arise which do not contain the plasmid and these grow faster (out-compete) the parent cell.
 - In other words, in the absence of other pressures, after a period of time the population of cells in a culture will be those which have "lost" the plasmid.

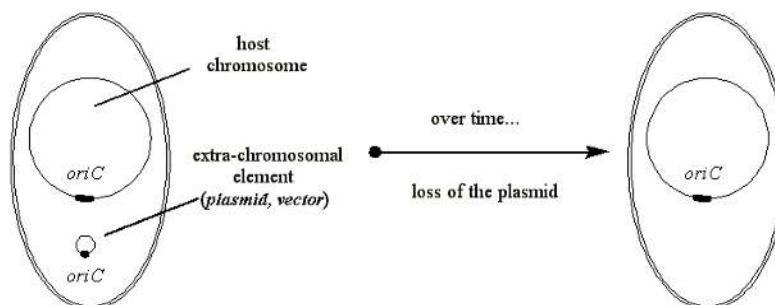


Figure 2.3.3: Loss of plasmid

- In organisms with more than one chromosome (eukaryotes) there are a variety of mechanisms to ensure that proper segregation of chromosomes occurs, i.e. to make sure that daughter cells contain equal numbers of all the chromosomes.
- One basic mechanism is that each chromosome contains **essential genes**, and if these are lost, the cell cannot survive.

Drug resistance

- By far the most common approach to the maintenance of plasmids is through the incorporation of drug resistance genes.
 - These are also known as **selectable markers**, i.e. we can select for their presence by including **antibiotics** in the growth media.

Ampicillin

- Ampicillin binds to and inhibits a number of enzymes in the bacterial membrane that are involved in the synthesis of the **gram negative** cell wall.
 - Therefore, proper cell replication cannot occur in the presence of ampicillin.
- The ampicillin resistance gene (**amp^r**) codes for an enzyme (***b-lactamase***) that is secreted into the periplasmic space of the bacterium where it catalyzes hydrolysis of the b-lactam ring of the ampicillin.
 - Thus, the gene product of the amp^r gene **destroys the antibiotic**.
- Over time the ampicillin in a culture medium or petri plate may be substantially destroyed by b-lactamase.
 - When this occurs, cell populations can arise which have "lost" the plasmid.

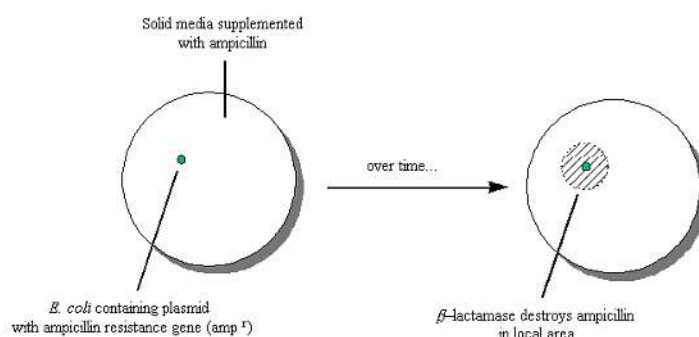


Figure 2.3.4: Ampicillin resistance

Tetracycline

- Tetracycline binds to a protein of the **30S** subunit of the ribosome and inhibits ribosomal translocation along the messenger RNA which codes for the protein (i.e. the drug interferes with normal translation or production of proteins).
 - The tetracycline resistance gene (tet^r) encodes a 399 amino acid outer membrane associated protein of *gram negative* cells that **prevents the antibiotic from entering the cell**.
- Thus, this drug resistance gene **does not destroy the antibiotic**. Pressure will be maintained throughout the cell culture process to keep the plasmid containing the drug resistant gene.

Chloramphenicol

- Chloramphenicol binds to the ribosomal **50S** subunit and inhibits protein synthesis.
- The chloramphenicol resistance gene (**Cm^r**) codes for a protein known as the **cat** protein.
 - The **cat** protein is a tetrameric cytosolic protein that, in the presence of acetyl coenzyme A, catalyzes the formation of hydroxyl acetoxy derivatives of chloramphenicol that are unable to bind to the ribosome.
 - Like with *ampicillin*, **the Cm^r gene product destroys the antibiotic**.
 - Additionally, the expression of **cat** protein is influenced (down regulated) by the presence of glucose in the media.

Kanamycin and neomycin

- Bind to ribosomal components and inhibits protein synthesis.
- The Kan^r gene codes for a protein which is secreted into the periplasmic space and interferes with the transport of these antibiotics into the cell.
 - Like tetracycline resistance, the Kan^r gene does not destroy the antibiotic.

Colicin E1

- This is a member of a general class of substances known as **bacteriocins**.
- *Colicin E1* causes lethal membrane changes in bacteria.
 - The drug resistance gene (*cea*) codes for a protein that interferes with the action of colicin in an unknown manner.

Prokaryotic plasmids

- In addition to making a plasmid by using the *E. coli* *OriC* region, there is a naturally occurring *E. coli* plasmid called the **ColE1 plasmid**
 - The ColE1 origin of replication is uni-directional (unlike *oriC*)
 - The replication from the ColE1 *ori* region does not require the associated proteins (e.g. *dnaA* protein) like *oriC*, (but it does require specific RNA molecules).
- Depending on the exact region of the ColE1 origin which is inserted into a circular dna molecule, the extrachromosomal element will be maintained with either a "low" or "high" copy number
- The **rop** gene near the ColE1 origin is involved in the regulation of replication.
 - If the ColE1 *ori* region includes this gene then the plasmid is maintained with an average copy number of 10-30 plasmids/cell. This is considered **low copy** number.
 - If the ColE1 *ori* region does not have the *rop* gene, then the resulting plasmid is maintained with an average copy number of 100-200 plasmids/cell. This is considered a **high copy number** type of plasmid.
- If the plasmid contains a gene coding for a protein (like drug resistance genes) copy number can influence the amount of such protein in the cell.

pBR322 (4.36 Kb)

- One of the original cloning plasmids.
- Constructed by ligating together:
 - the tetracycline resistance gene from plasmid pSC101
 - ColE1 and *rop* region from the ColE1 plasmid
 - the ampicillin resistance gene from the transposon Tn3
- Contains both **ampicillin** and **tetracycline** resistance genes (markers).
- Contains unique restriction sites inside and outside of these markers.
- Contains **rop** region near CoE1 *ori*, therefore, it has a low copy number (10-30)
- Numbering begins at the unique *EcoR* I restriction site (GAATTC). The first 'T' in this sequence is base number "1".

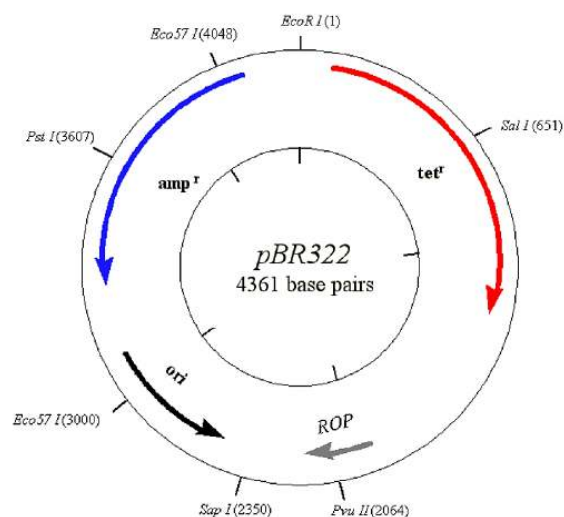


Figure 2.3.5: pBR322

The plasmid diagram:

- In the center is the plasmid name (usually starts with a lower case 'p') and the size in basepairs
- the inner ring provides ticks in 1 Kilobase (Kb) intervals to give an idea of the general location of parts of the plasmid
- the arrows indicate genes, markers, ori or replication, promoters, polylinkers, transcription terminators, and other important or functional items
- the outer ring usually indicates the location of unique, or limited number (usually <3), restriction endonuclease sites. Restriction enzymes which have more than three sites will not be indicated. Be aware that enzymes which do not cut at all will also not be listed!

PUC18/19 (2.69 Kb)

- Lacks the *rop* gene near the ColE1 ori region. Thus, this plasmid tends to accumulate in high copy number (100-200).
- This vector contains only the ampicillin resistance marker.
- This vector contains a **polylinker** region
 - A synthetic DNA sequence which contains a clustering of unique restriction enzyme recognition sites
 - Allows DNA fragments, generated by a variety of restriction endonuclease cleavages, to be inserted into the plasmid
 - pUC18 has the polylinker in one orientation
 - pUC19 has the same polylinker, but in the **opposite** orientation
 - Thus, **fragments with unique restriction sites on each end can be inserted in a specific orientation**
 - The *Pst* I site in the *amp^r* gene was mutated to remove it. The *Eco*R I site at position (1) was mutated to remove it. **This was done to make the restriction sites in the polylinker region unique**
- This vector also contains a *transcription promoter region* from the *lac operon*, which allows foreign genes to be inserted and transcribed/translated.
 - The polylinker region is just downstream (3') to the lac promoter
 - Inserted genes can be transcribed from this promoter

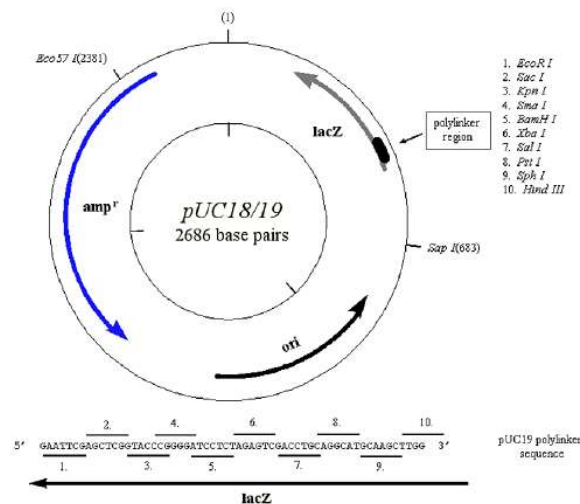


Figure 2.3.6: pUC18/19

This page titled [2.3: Extrachromosomal elements, plasmids, selectable markers](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

2.4: Central Dogma and the Genetic Code

The "Central Dogma"

We have seen how DNA, with the aid of specific polymerases and accessory proteins, is able to replicate. We have also seen how we can use this information to create autonomously replicating extra-chromosomal elements (i.e. **plasmids**). However, the real utility of such systems arises when we use them to create **proteins** of interest. To get to proteins we have to go through **RNA** first.

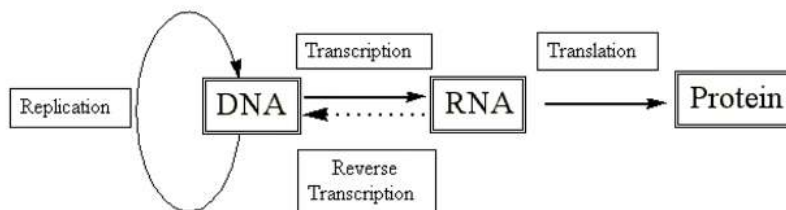


Figure 2.4.1: Central dogma

Structural features of RNA:

1. Similar to DNA except it contains a 2' hydroxyl group (**makes phosphodiester bond more labile than DNA**).
2. **Thymine** in DNA is replaced by **Uracil** in RNA

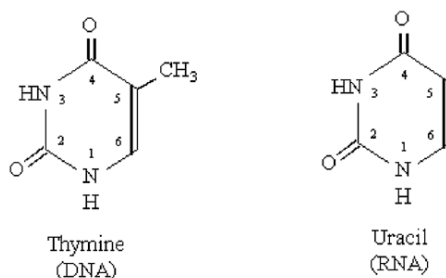


Figure 2.4.2: Thymine vs. Uracil

3. RNA's can adopt regular **three-dimensional structures** which allow them to function in the process of genetic expression (i.e. the production of proteins).

- This ability to adopt defined three dimensional structures which impart functionality places RNA in a unique class - somewhat akin to proteins, and different from DNA.
- For example certain RNA molecules, when folded, exhibit **catalytic capacities** (e.g. the cleavage of RNA molecules).
- The majority of RNA in cells is found in complexes with proteins. The most common example is ribosomes (involved in protein synthesis).

Transcription: the copying of DNA by an RNA polymerase to make RNA.

RNA polymerase:

- Can initiate a new nucleic acid strand given a template.
- DNA polymerases **cannot**; they require a **primer** (or more typically, an RNA polymerase to provide the primer).

Protein Synthesis

- Three kinds of RNA molecules perform different functions in the protein synthesizing apparatus:
 1. **Messenger RNA** (mRNA) encodes the genetic information copied from DNA in the form of a sequence of bases that specifies a sequence of amino acids
 2. **Transfer RNA** (tRNA) is part of the structural machinery which deciphers the mRNA code. They carry specific amino acids which are transferred to a nascent polypeptide according to the instructions contained within the mRNA.

3. **Ribosomal RNA** (rRNA) forms a complex with specific proteins to form the **ribosome** which is the key translational component
 - the ribosome complexes with mRNA and directs appropriate tRNA's and the synthesis of the polypeptide bond.

Translation:

The process by which the information contained within a mRNA is used to direct the synthesis of the corresponding polypeptide.

The Genetic Code

How is the information for a polypeptide sequence stored within an mRNA molecule? There are twenty different common amino acids, but only four different bases in RNA (A, C, G, and U).

Base Arrangement	Possible Combinations
1	$4^1=4$
2	$4^2=16$
3	$4^3=64$
4	$4^4=256$

A triplet arrangement would seem to be the minimum possible combination necessary to code for the 20 different amino acids. Although, there are obviously going to be a lot of codons "left over". Most amino acids are coded for by more than a single unique triplet, and therefore the genetic code is said to be **degenerate**.

Experiments which led to the solution of the genetic code:

Nirenberg and Matthei (1961): Nirenberg and Matthei worked with bacterial extracts which contained everything needed for translation, with the exception of mRNA. To this they added either poly A, poly U or poly C RNA. The proteins produced by the translation of these RNA's was determined (*poly G did not work, probably due to conformational problems*):

Poly U	Poly A	Poly C
Phe	Lys	Pro

Thus, the triplet UUU = Phe, AAA = Lys, and CCC = Pro.

Korana (1963): In a cell free extract system, Korana added mRNA with repeating nucleotide sequences. The sequence ...ACACACAC... resulted in a polypeptide with alternating **threonine and histidine** residues. But, was threonine coded by ACA, and histidine by CAC? Or vice versa? *To determine the answer to this, the mRNA sequence ...AACAACAACAAC... was tried.* There were three different possible reading frames for the translation of this mRNA:

- AAC AAC AAC
- ACA ACA ACA
- CAA CAA CAA

But CAC was not a possible triplet. This sequence was found to code for three different polypeptide chains: poly Asn, poly Thr, and poly Gln. Since no histidine was found, histidine was therefore coded for by the triplet CAC.

Nirenberg and Leder (1964): Nirenberg and Leder used a filter which would allow RNA triplets and charged tRNA's to pass through, but would prevent passage of larger ribosomes. Specific triplet RNA sequences would bind to ribosomes and cause the binding of the associated charged tRNA molecules (coded for by the specific triplet). In a given experiment, if a unique charged tRNA were radiolabeled (on the amino acid), then it could be determined whether that particular charged tRNA was associated for by the unique triplet. In this way, all 61 codons for amino acids were determined.

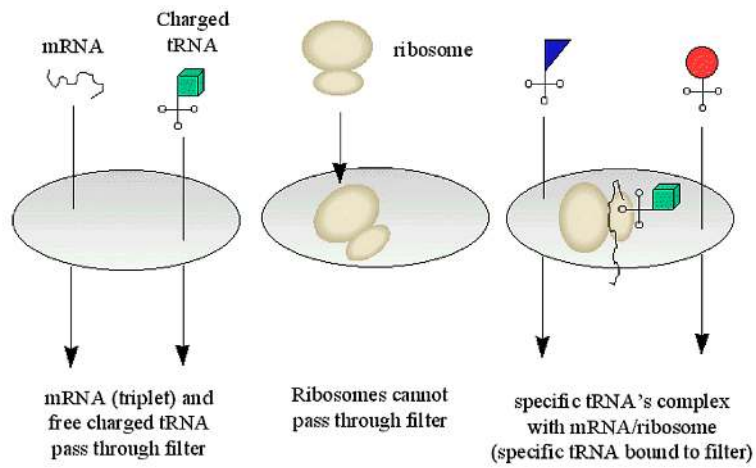


Figure 2.4.3: Nirenberg and Leder experiment

The genetic code

5' End (Start)	Second Position				3' End
	U	C	A	G	
U	Phe 0.24	Ser 0.34	Tyr 0.25	Cys 0.49	U
	Phe 0.76	Ser 0.37	Tyr 0.75	Cys 0.51	C
	Leu 0.02	Ser 0.02	Stop	Stop	A
	Leu 0.03	Ser 0.04	Stop	Trp 1.00	G
C	Leu 0.04	Pro 0.08	His 0.17	Arg 0.74	U
	Leu 0.07	Pro 0.00	His 0.83	Arg 0.25	C
	Leu 0.00	Pro 0.15	Gln 0.14	Arg 0.01	A
	Leu 0.83	Pro 0.77	Gln 0.86	Arg 0.00	G
A	Ile 0.17	Thr 0.35	Asn 0.06	Ser 0.03	U
	Ile 0.83	Thr 0.55	Asn 0.94	Ser 0.20	C
	Ile 0.00	Thr 0.04	Lys 0.74	Arg 0.00	A
	Met (start) 1.00	Thr 0.07	Lys 0.26	Arg 0.00	G
G	Val 0.51	Ala 0.35	Asp 0.33	Gly 0.59	U
	Val 0.07	Ala 0.10	Asp 0.67	Gly 0.38	C
	Val 0.26	Ala 0.28	Glu 0.78	Gly 0.00	A
	Val 0.16	Ala 0.26	Glu 0.22	Gly 0.02	G

Note

E. coli codon preferences are indicated.

All proteins in prokaryotes and eukaryotes begin translation with the *initiator* codon **AUG** (methionine). The three codons, **UAA**, **UGA** and **UAG** are **termination codons** (don't code for any amino acids but signal the end of the protein chain).

Note the apparent relative importance of the middle base in the codon triplet

Middle base in codon triplet

U	C	A	G
Phe Leu Ile Met Val	Ser Pro Thr Ala	Tyr His Gln Asn Lys Asp Glu Stop	Cys Trp Arg Ser Gly Stop
<i>Hydrophobic</i>	<i>Small/Polar</i>	<i>Charged/Polar</i>	<i>Polar</i>

Can common protein architectures be patterned by a simple quaternary pattern of residues?

The twenty common amino acids and their three-letter and single-letter acronyms:

Amino Acid	Three letter acronym	One letter acronym
Alanine	Ala	A
Cysteine	Cys	C
Aspartic Acid	Asp	D
Glutamic Acid	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	Thr	T
Valine	Val	V
Tryptophan	Trp	W
Tyrosine	Tyr	Y

This page titled [2.4: Central Dogma and the Genetic Code](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

2.5: Gene and Operon

Genes

The entire nucleic acid sequence that is necessary for the synthesis of a functional polypeptide or RNA molecule. Thus, a gene contains additional sequence information beyond that which codes for the amino acids in a protein or the nucleotides in an RNA molecule. The gene also contains the DNA necessary to get a particular transcript made.

Transcription control regions can be remote to the coding region (on the order of Kb's or 10's of Kb's away).

Most prokaryotic genes lack introns (intervening DNA sequence). In prokaryotes, genes which encode proteins with relationships in a metabolic pathway form *Operons* - which produce *polycistronic mRNA's*.

Definitions: Operon and Promoter

- An operon is in bacterial DNA, a cluster of contiguous genes transcribed from one promoter that gives rise to a polycistronic mRNA.
- A Promoter is a DNA sequence to which RNA polymerase binds prior to initiation of transcription - usually found just upstream of the transcription start site of a gene

e.g. *Trp Operon* - involved in the biosynthesis of the amino acid tryptophan:

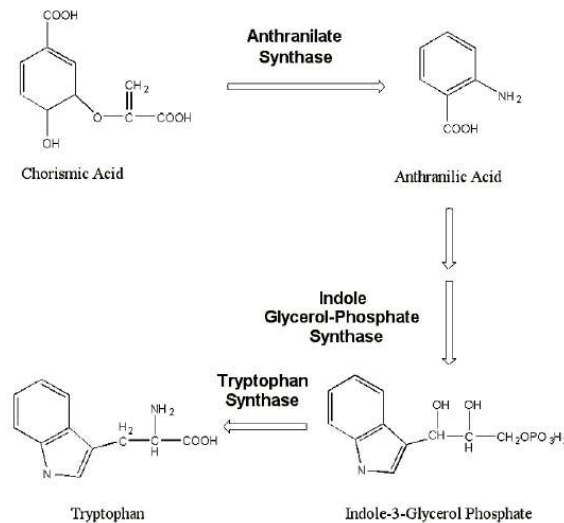


Figure 2.5.1: Chemical pathway of *trp* operon

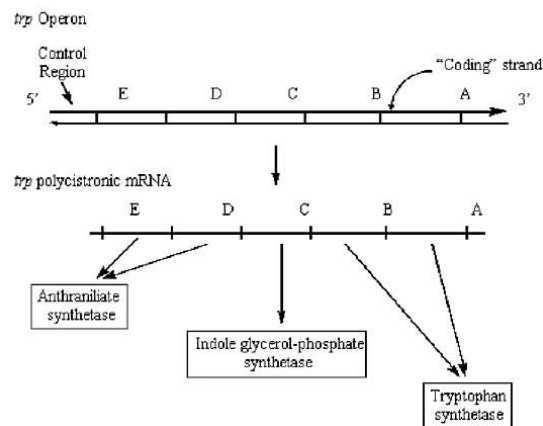


Figure 2.5.2: *Trp* operon in DNA/RNA

A consequence of the arrangement of bacteria genes into operons is that *the level of mRNA for each of the genes in the operon is exactly the same.*

Ribosomes transcribe from the start of each gene, not only from the first gene.

Another consequence of the arrangement of bacteria genes into operons is that an upstream mutation (i.e. possibly inhibiting transcription) can prevent "downstream" genes from being transcribed and expressed. Most eukaryotic transcription units produce *monocistronic* mRNA's, (i.e. they encode only one protein). There is a fundamental difference in the translation processes of prokaryotes and eukaryotes:

1. In prokaryotes ribosomes can bind at specific recognition sequences anywhere within the mRNA (called ribosome binding sites, or "Shine-Dalgarno" sites).
 2. In eukaryotes, ribosomes bind via the interaction with specifically modified 5' region (so called **5' cap site**) of mRNA molecules.
 3. Most eukaryotic mRNA's are therefore **monocistronic**.
- Mutations in simple eukaryotic transcription units affect only one protein.

Complex Eukaryotic Transcription Units

The primary RNA transcript encoded by complex transcription units can be *spliced* in more than one way. Because of the different processing possibilities, the *exons* (coding regions) in a single complex transcription unit can be linked in alternative ways, to yield *different mRNAs* and *different proteins*.

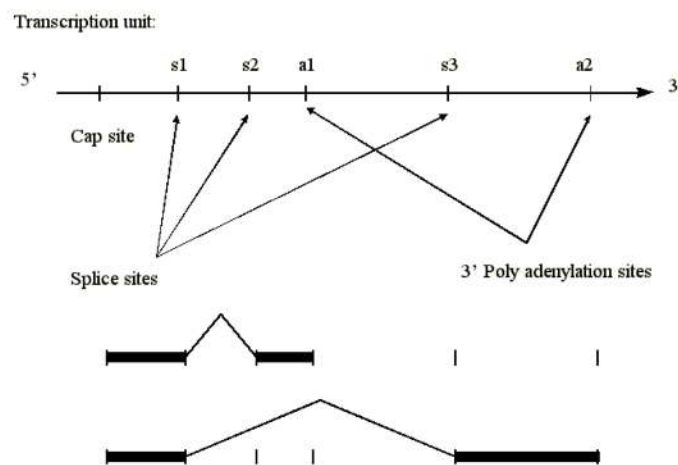


Figure 2.5.3: Complex transcription units

Transcriptional regulation

Successful survival requires adaptability and economy:

1. The ability to switch from metabolizing one substrate to another as environmental resources change
2. It would be an energetic waste to produce enzymes for a metabolic pathway which is not needed.

Induction versus Repression of Enzyme Synthesis

In *E. coli* certain enzymes are produced only when the cells are grown on certain substrates. This effect is called enzyme *induction*. For example, when cells are grown in the absence of a type of sugar known as a *b galactoside* (e.g. lactose) the cells contain very few molecules (~5 per cell) of the enzyme *b-galactosidase* (which cleaves lactose into glucose and galactose).

- There is no need for this enzyme in the absence of lactose.
- If lactose is added to *E. coli*, in a very short amount of time there are approximately **5000 molecules** of b-galactosidase per cell (approximately ~1,000 fold induction).

- If lactose is removed from the media **synthesis** of b-galactosidase **stops**.

A similar but opposite situation occurs in regard to the synthesis of tryptophan (the biosynthetic enzymes are contained in the **trp operon**). In this case production of the enzymes for tryptophan biosynthesis are rapidly **shut down** if tryptophan is **present**, in a process called *repression*. Repression is a transcriptional regulatory mechanism for commonly required gene products. Induction is a transcriptional regulatory mechanism for gene products which may be required under unusual or infrequent situations.

This page titled [2.5: Gene and Operon](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

2.6: The lac Operon; CAP site; DNA footprinting

The *E. coli* lac operon

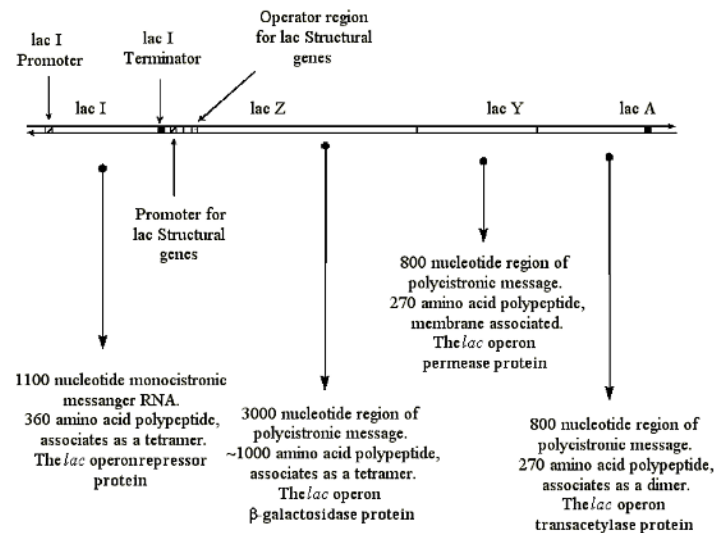


Figure 2.6.1: *lac* Operon

- *lac Z* codes for **b-galactosidase**, which is an enzyme that cleaves b-galactosides (e.g. lactose).
- *lac Y* codes for **permease**, which is involved in the transport of b-galactosides into the cell.
- *lac A* codes for **b-galactoside transacetylase**, which acetylates b-galactosides.
- A mutation in either *lac Z* or *lac Y* can lead to a *lac⁻* genotype, i.e. cells which cannot utilize b-galactosides as a nutrient.
- A *lac A⁻* mutant, lacking transacetylase activity, can still utilize b-galactosides (it is still *lac⁺* genotype). Its role in the metabolism of bgalactosides is not clear.

Promoter

a region of DNA involved in binding of RNA polymerase to initiate transcription.

Terminator

a sequence of DNA that causes RNA polymerase to terminate transcription.

- The cluster of three genes, *lac ZYA*, is transcribed into a single mRNA (**polycistronic** message) from a **promoter** just upstream from the *lac Z* gene.
- In the *absence* of an inducer the gene cluster is *not* transcribed.
- When an inducer is added (e.g. lactose, or the non-hydrolyzable analog isopropyl thiogalactoside - IPTG) transcription starts at a single promoter (**lac P**) and proceeds through the *lac ZYA* genes to a terminator sequence located downstream of the **lac A** gene.

Note

The *lac ZYA* mRNA has a half life of ~3 minutes, which allows induction to be reversed relatively rapidly (i.e. cells stop producing enzymes rapidly after induction stops).

Exercise 2.6.1

What molecule does the inducer (lactose) interact with to affect transcriptional regulation (i.e. induction of the *lac* operon)?

Answer

It is not b-galactosidase, permease or transacetylase, rather it is a separate protein called a repressor protein.

- The *lac* genes are controlled by a mechanism called **negative regulation**.
 - This means that *they are transcribed unless they are turned off by the regulator protein*.
 - A mutation that **inactivates the regulator protein** causes the *lacZYA* genes to be **continually expressed**.

*Since the function of the regulator protein is to prevent expression, it is called a **repressor protein**.*

There are two types of genes in the *lac* operon:

1. **Structural genes** - they code for enzymes required for some biochemical pathway (e.g. *lac Z*, *Y* and *A*).
2. **Regulator genes** - they code for proteins involved in regulation of structural genes.

*Mutations in **structural** genes typically affect the function of only that structural gene.*

*Mutations in **regulator** genes can affect the expression of all structural genes in an operon.*

***lac I* is the regulator gene of the *lac* operon.**

- This gene is located just upstream of the promoter region for the *lac* structural genes.
- The *lac I* gene has its own *promoter (constitutive)* and terminator.
- It makes a monocistronic message, and codes for one protein - the ***lac* repressor protein**.

*The crucial feature of the *lac* "control circuit" resides in the dual feature of the *lac I* repressor protein:*

1. It can prevent transcription
2. It can recognize and bind the small molecule inducer (lactose or IPTG)

Prevention of transcription by the *lac* repressor

- *lac* repressor (active as a **tetrameric** protein) binds to a sequence of DNA called the **operator** (*lac O* region).
 - The operator region lies between the *lac* promoter region (site of RNA polymerase binding and transcription initiation) and the *lac Z* gene.
 - The first 26 base pairs of the *lac Z* gene comprise the operator region.

When the repressor binds to the operator region, its presence prevents RNA polymerase from initiating transcription at the promoter.

- It is not that the repressor protein "blocks" the movement of RNA polymerase through the *lac Z* gene.
- Repressor binding and RNA polymerase binding (to the promoter) are mutually exclusive at the *lac* promoter/operator (*lac PO*) region.

How does the repressor/operator interaction change in the presence of the inducer molecule?

- The **inducer** can bind to the repressor to form a **repressor/inducer complex that no longer associates with the operator**.
 - The key feature of this interaction is that *the repressor protein has two binding sites*, one for the inducer and one for the operator.
 - When the inducer binds at its site, it changes the conformation of the repressor protein such that the operator binding site has a much reduced affinity for the DNA operator region.
 - This type of control is called **allosteric control**.
 - The result is that *when the inducer is added the repressor is converted to a form which releases from the operator*.

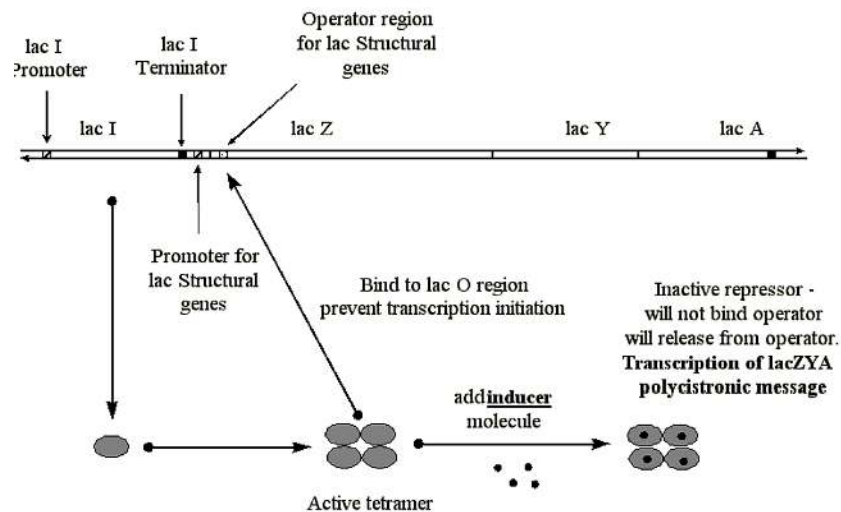


Figure 2.6.2: Inducer

Positive control of the lac operon is exerted by cAMP-CAP complex

- E. coli* prefers **glucose** over other carbon sources.

When glucose enters an E. coli cell it is utilized directly without induction of any new enzymes.

- When *E. coli* is grown on glucose, if another sugar (e.g. lactose) is added *the induction of enzymes to utilize the other sugar does not occur until the glucose is used up.*
- When *E. coli* is starved for glucose it synthesizes an unusual nucleotide: *cyclic 3'5' adenosine monophosphate* (cyclic AMP, or *cAMP*):

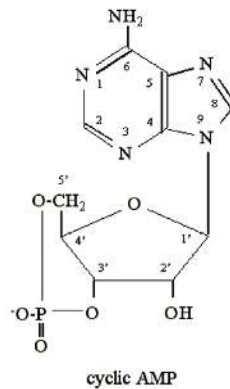


Figure 2.6.3: cAMP

- In bacteria an increase in the cAMP level seems to be an "alert" signal indicating a low glucose level:

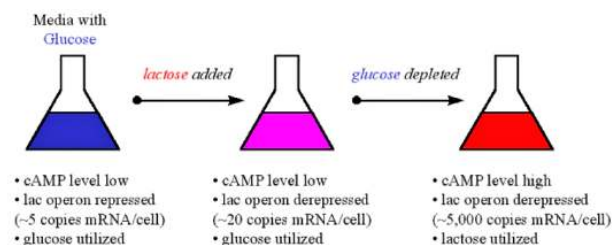


Figure 2.6.4: Interaction of cAMP level and lac operon

Dibutyl cAMP

- an *analogue* of cAMP which can pass through the *E. coli* membrane and into the cell.
- If this is added to media containing glucose and lactose it will result in the **induction** of the *lac* operon.
- Thus, it mimics the chemical message which tricks the *E. coli* to respond as though glucose levels were low.
- Mutants of *E. coli* have been isolated which cannot be induced to metabolize **any** sugar other than glucose. There were two general categories of mutants:
 1. Class I. *Defective in the enzyme adenylate cyclase*. These mutants are unable to make cAMP even when the glucose concentration is low.
 2. Class II. Lacks a particular protein known as **cAMP receptor protein** (CRP) or, also known as catabolite receptor protein (CRP).
- Maximum transcription from the *lac* operon requires the presence of a **cAMP/CRP** complex.
 - cAMP/CRP complex binds to a specific sequence in the *lac* control region called the "**CAP**" site.
 - The CAP site is just **upstream** from the RNA polymerase binding site.
 - Mutations in the CAP site that prevent cAMP-CRP binding also prevent high levels of expression of the *lac* operon.
- Thus, bound cAMP/CRP complex **activates** transcription (**positive control**), whereas bound *lac* repressor **inhibits** transcription (**negative control**).
 - cAMP/CRP complex has affinity for DNA, and RNA pol.
 - Enhances complex formation of RNA pol with the DNA promoter region.

Induction of the *lac* operon with lactose analogues

- The *lac* operon can be induced with *lactose*
 - b-galactosidase (*lacZ* gene product) metabolizes the lactose
 - When levels of lactose are reduced, the *lac* operon is again repressed by the *lac* repressor (*lacI* gene product)
- *Non-metabolized lactose analogues* can continually induce (i.e. de-repress) the *lac* operon
 - isopropyl b-thiogalactoside, or **IPTG**, is a non-metabolized lactose analogue

DNA "footprinting" experiments

- If a protein binds to a region of DNA, *it can protect that region of DNA from digestion by dnase* (**DNAse I**: an endonuclease at sites adjacent to pyrimidine nucleotides).
 - A fragment of DNA can be labeled at the 5' ends with ³²P and then the label can be preferentially removed from one end (i.e. the 3' end of a gene) by an appropriate restriction endonuclease.
 - If this DNA fragment, with a label at one specific end, forms a complex with a DNA binding protein the protein will protect the region of DNA that it binds to from DNAse I digestion.
 - The digestion is done so as to be incomplete, for the purposes of this discussion, imagine that each DNA molecule is cleaved only once. Furthermore, the site of cleavage is randomly chosen from the available sites.
 - Fragments of the DNA, separated and analyzed by size (using gel electrophoresis) after digestion will indicate the protected region:

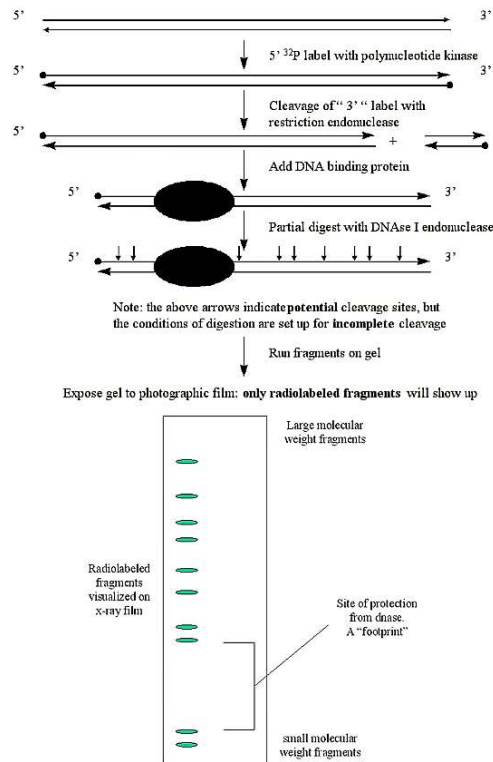


Figure 2.6.5: DNA footprinting

Results of footprinting experiments

lac DNA incubated with either cAMP/CAP protein, or RNA polymerase, or lac I repressor protein:

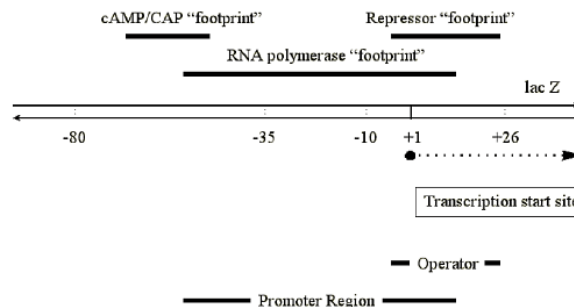


Figure 2.6.6: *lac* repressor with cAMP

RNA polymerase interacts with specific promoter sequences and produces a "footprint" over a region of ~70 base pairs.

- This protection was observed to be more obvious on one strand than the other (i.e. if the other strand was labeled the results did not show as much protection).
- This region of DNase protection included sites in the DNA from which mutagenesis experiments produced either "up" regulation or "down" regulation of promoter strength.
- These mutagenic "hot" spots affecting promoter strength were located at positions either -10 or -30 upstream from the transcription start site (position +1 in the above diagram):

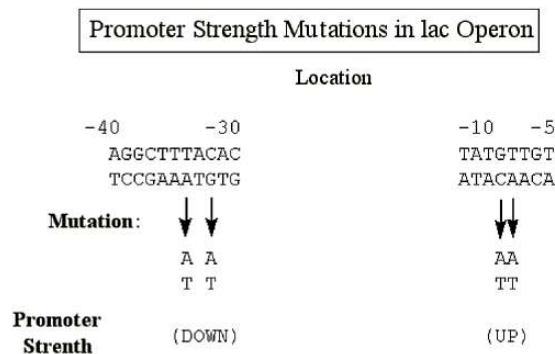


Figure 2.6.7: Promoter Strength Mutations

- Promoters can be classified according to their "strength".
- This refers to the relative **frequency of transcription initiation** (transcriptional initiation events per minute), and is related to the **affinity of RNA polymerase for the promoter region**.
- Many promoters in *E. coli* have been characterized and a "consensus" promoter sequence has been identified:

***E. coli* Consensus Promoter Strength at -35 and -10 Promoter Regions**

-40 -35 -(15-17 Base Pairs)- -10 -5
 XXXTTGACAT TATAATXX

Frequency of occurrence:
 >75%
 50-75%
 40-50%

Figure 2.6.8: Consensus Promoter Strength

Note

The lac promoter is a relatively weak promoter

RNA Polymerase

- E. coli* RNA polymerase is a holoenzyme comprised of subunits b', b, a (dimer) and σ^{70} .
 - The σ^{70} subunit is the subunit which binds to the promoter region, but is unable to initiate RNA synthesis.
 - After the σ^{70} subunit binds, the other subunits bind forming a function RNA polymerase.
 - After approximately 10 base pairs have been transcribed the σ^{70} subunit leaves and the **core polymerase** continues on.

The lac Operator Region

- The lac operator region is comprised of an (imperfect) **inverted repeat** region.
- Not surprisingly active repressor molecules are composed of a homodimer.
 - In the homodimer structure there are a pair of α -helix regions which insert into adjacent major grooves of the DNA.
 - The separation is approximately 34 angstroms apart.

This page titled 2.6: The lac Operon; CAP site; DNA footprinting is shared under a [not declared](#) license and was authored, remixed, and/or curated by Michael Blaber.

2.7: Transcriptional Regulation: Transcription termination, the trp operon

Transcription termination in prokaryotes

- Because protein coding regions are closely spaced in the genomes of microorganisms, independent control of neighboring genes is possible only if a **transcription termination site** lies between them (i.e. downstream genes must be independently regulated).
- There are two general mechanisms of transcription termination:
 - one requires the presence of a transcription termination protein called *rho*,
 - the other requires no associated proteins.

Rho-independent termination

- Rho*-independent transcription termination sequences have two general characteristic features:
 - A 3' stretch of T residues
 - A 'GC' rich interrupted palindrome just upstream of the 3' poly T region.

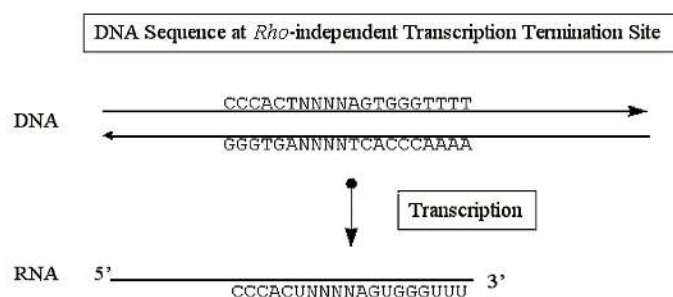


Figure 2.7.1: Rho-independent transcription termination site

The important features of Rho-independent transcription termination are as follows:

- The inverted repeat region can self-hybridize to form a "stem-loop" structure:

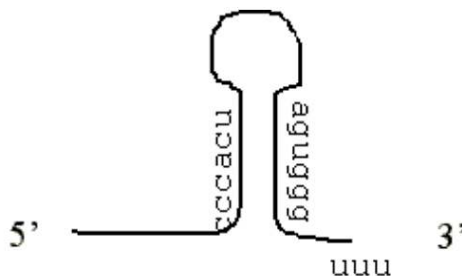


Figure 2.7.2: Stem-loop structure

- The GC rich stem loop structure interacts with RNA polymerase to cause it to pause.
- The short UUU region, which is base pairing with AAA sequence on the anti-sense DNA strand, has low thermal stability and melts - releasing the nascent RNA transcript.

Attenuation at the trp operon and premature mRNA chain termination

- The *trp* operon is regulated by the *trp* repressor protein.
 - Trp* repressor protein binds to tryptophan and undergoes a conformational change which allows it to bind to the *trp* operator region (downstream of the promoter).
 - This prevents transcription of the *trp* operon in the **presence** of tryptophan (i.e. tryptophan represses expression).
 - However the repression is somewhat incomplete and **there is a second mechanism which contributes to transcription repression**.

- When tryptophan is present, and the operon is repressed, the few transcripts which are made are actually **quite short** and do not encompass the entire polycistronic message of the *trp* operon.
- The short mRNA that is made comprises only a region called the leader sequence:

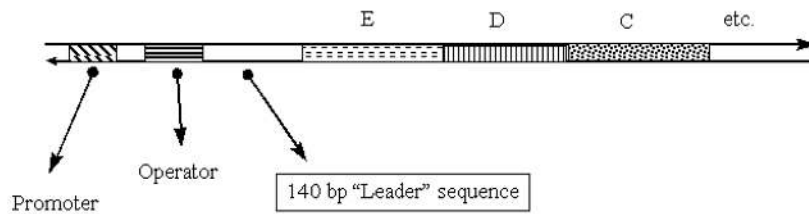


Figure 2.7.3: Leader sequence

- The leader region contains an attenuator sequence - a site where a choice is made between **elongation** of the growing *trp* transcript or (premature) **termination**.
- Attenuation** depends on the interplay between *ribosome binding* and translation of the nascent mRNA transcript and the formation of a particular *stem-loop structure* in the mRNA leader sequence.
 - Formation of this structure depends on the rate of ribosomal translation of the leader sequence (**the leader sequence contains an AUG start codon**).
 - Efficient translation of the leader sequence depends upon the concentration of charged tRNA's for the appropriate amino acids coded for by the mRNA leader sequence.*
- The leader sequence has the following characteristics:
 - It contains a ribosome binding site and an AUG start codon necessary for the initiation of translation.
 - The corresponding amino acid sequence coded for by the leader sequence contains several **tryptophan residues**.
 - It contains several **inverted repeat sequences** such that the mRNA can adopt different alternative structures.
 - One of these structures represents a *rho*-independent **termination site**.

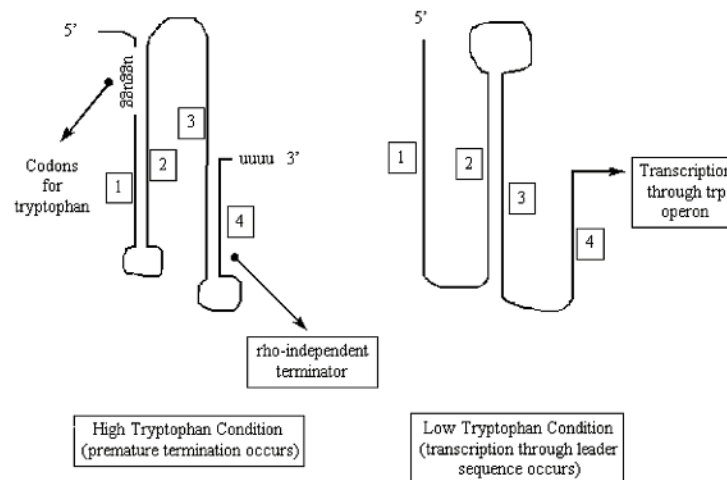


Figure 2.7.4: High vs. Low tryptophan

Under conditions of low tryptophan

- A **ribosome** translating the nascent *trp* polycistronic mRNA **stalls** at the codons for tryptophan in the leader mRNA sequence.
 - This **prevents** segments 1 and 2 of the nascent mRNA from forming a potential stem-loop structure, and frees up segment 2 to form a stem loop structure with segment 3.

- In this case, nascent transcription can proceed through, transcribing the entire *trp* operon.

Under conditions of high tryptophan

- A **ribosome** translating the nascent *trp* polycistronic mRNA does **not** stall at segment 1 (the region with codons coding for tryptophan).
 - In this case the stem-loop structure between 1 and 2 is **not prevented from forming**, and thus the stem-loop structure 3-4 **can form**.
 - This stem loop structure is essentially a ***rho*-independent transcription termination structure**.
 - The RNA polymerase transcribing the nascent message will terminate.

Rho-dependent transcription termination

- About half the termination sites in *E. coli* require an accessory protein called the ***rho* factor**.
- Many *rho*-dependent sites have been characterized from *E. coli* and **no obvious sequence similarity is present**.
- Rho associates with the nascent RNA transcript and this interaction activates an ATPase activity which appears to allow *rho* to **translocate** along the mRNA in the **3' direction**.
 - It may be that if the RNA polymerase pauses, *rho* can catch up and cause termination (i.e. bump RNA polymerase off the DNA template?).
 - Pausing of RNA polymerase is thought to be an important mechanism of rho-dependent transcription termination.

This page titled [2.7: Transcriptional Regulation: Transcription termination, the *trp* operon](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

2.8: mRNA Translation

How is the code contained in mRNA translated into a protein?

Structure and function of transfer RNA's

- **tRNA's** have **two** functions:
 1. To chemically link to a particular amino acid (**covalent**)
 2. To recognize a specific **codon in mRNA (non-covalent)** so that its attached amino acid can be added to a growing peptide chain

Amino-acyl tRNA synthetases

- Function is to "**charge**" tRNA molecules; i.e. *to chemically link a specific amino acid to its associated tRNA molecule.*

Amino Acids	Amino-acyl tRNA synthetases	tRNA's	Codons
20	20	30-40 (prokaryotes) 50 (eukaryotes)	61 (3 stop codons)

Conclusions:

1. There is one amino-acyl tRNA synthetase per amino acid (they are quite specific).
2. There is potentially more than one tRNA per amino acid.

Therefore, amino-acyl tRNA synthetases must be able to recognize more than one tRNA.

1. There is potentially more than one codon per tRNA.

Therefore each tRNA must be able to recognize more than one codon (there is not a unique tRNA for each codon).

Structure of tRNA's

- 70-80 nucleotides long
- Form a series of stem/loop secondary structures
- tRNA's are synthesized with the standard bases AGCU. However, after synthesis several bases may be modified:
 1. Uridylate may be **methyalted** to produce **Thymidylate**
 2. Uridylate may be **rearranged** to produce **pseudouridylate** (i.e. ribose attached to Carbon 5 instead of Nitrogen 1).
 3. Guanidylate may be **methyalted** at different positions.
- The amino acid is attached at the 3' end of the tRNA to either the 2' hydroxyl or the 3' hydroxyl.
 1. **Class I** amino-acyl tRNA synthetases attach their associated amino acids to the tRNA **2' hydroxyl** (NOTE: typically the hydrophobic amino acids)
 2. **Class II** amino-acyl tRNA synthetases attach their associated amino acids to the tRNA **3' hydroxyl** (NOTE: typically hydrophilic amino acids)

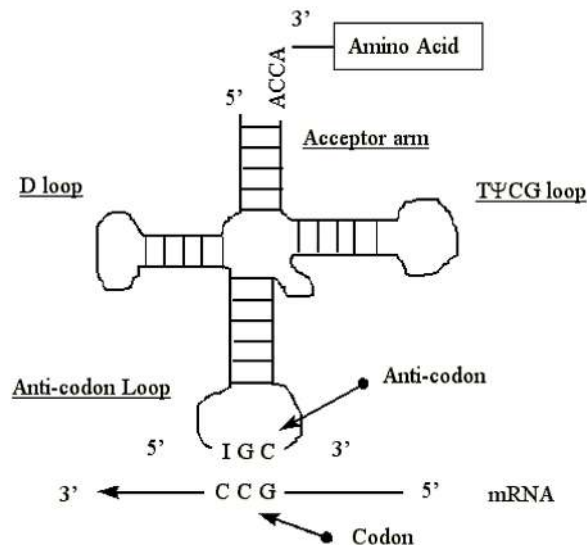


Figure 2.8.1: tRNA

- If perfect Watson-Crick base pairing were required at the codon/anti-codon triplet then **61 different tRNA's would be required**.
- We know this is not the case, therefore a single tRNA anti-codon must be able to recognize several different mRNA codon triplets.
- This greater recognition of tRNA is possible due to "wobble" basepair interactions at the third base in the codon/first base in the anti-codon:

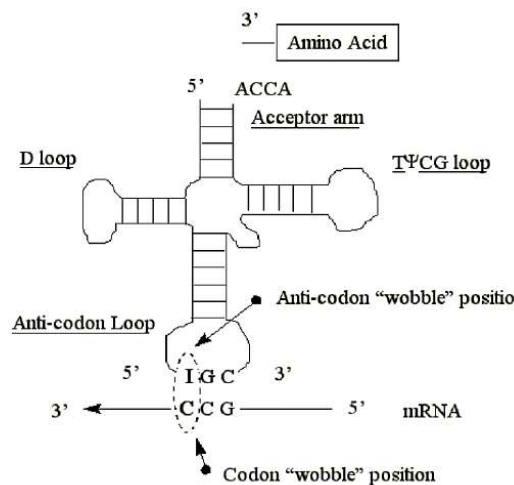


Figure 2.8.2: Codon wobble

Possible "wobble" codon base pairing (in addition to Watson-Crick):

1. **U - G**
2. **I - C**
3. **I - A**
4. **I - U**

- Where U, G, A and C can be in **either** the codon (mRNA) or anti-codon (tRNA)
- **I** (inosine) can be found in the **anti-codon**.

For example, the codons UUU and UUC are both recognized by the tRNA which has GAA in the anti-codon position (making either G - C, or G - U base pairings).

Recognition of amino acids by amino-acyl tRNA synthetases

- Appears to involve not only the anti-codon triplet but significant other contacts as well (mostly involving the acceptor stem region).

Ribosomes

- The mRNA with its encoded information and the individual tRNAs loaded with their amino acids are brought together by a mutual affinity for an RNA-protein complex called the **Ribosome**.
- The rate of protein synthesis by a ribosome is approximately **3-5 amino acids/minute**.
 - For example, a large protein (e.g. Titin, 30,000 amino acids) takes 2-3 hours to make.
- Ribosomes are composed of individual ribosomal RNA (rRNA) molecules and more than 50 accessory proteins, with a general prokaryotic organization of a **small subunit (30S)** and a **large subunit (50S)**.

Translation of mRNA to proteins

- Protein synthesis is usually considered in three steps:

1. **Initiation**
2. **Elongation**
3. **Termination**

AUG is the initiation signal in mRNA

- The **first** event of the initiation stage is the attachment of a free molecule of methionine (Met) to the end of a tRNA^{Met} by a specific aminoacyl-tRNA synthetase.
- There are at least **two** types of tRNA^{Met}:
 1. tRNA_i^{Met}: can **initiate** protein synthesis (at **AUG** met codon)
 2. tRNA^{Met}: can **incorporate** Met residues during on-going protein synthesis (at **AUG** met codon)
- Methionine tRNA synthetase attaches Methionine to both tRNA molecules.
- Only methionyl-tRNA_i^{Met} can bind to the small ribosomal subunit to begin the process of protein synthesis.
 - In bacteria, the amino group of the methionine in methionyltRNA_i^{Met} is **formylated**.
 - The Met-tRNA_i^{Met}, along with a protein-GTP complex and the small (30S) ribosomal subunit bind to the mRNA at a specific site, **near** the AUG initiation codon.

Initiation of protein synthesis

- In most prokaryotes an RNA component (16S rRNA) in the small rRNA subunit (30S) recognizes and hybridizes to a specific sequence on the mRNA called **the Shine-Dalgarno** sequence:

mRNA 5' -**UAAGGAGG** -(5-10 nucleotides)-**AUG** 3'
16S rRNA OH-**AUCCUCC** -(~1400 nucleotides)-5'

- The Shine-Dalgarno sequence is thus a **ribosome binding site** which is necessary for the initiation of translation.
 - Note that the ribosome does not bind at the AUG start codon, but 5-10 nucleotides **upstream**.
 - The Shine-Dalgarno sequence can be located **anywhere** within an mRNA.
- A series of **initiation factors**, **Met-tRNA_i^{Met}**, **mRNA** and the **30S (i.e. 16S component)** ribosomal subunit are necessary for formation of the 30S initiation complex.
- The large (50S rRNA) rRNA binds along with release of initiation factors 1 and 2, and hydrolysis of GTP, to form the 70S initiation complex:

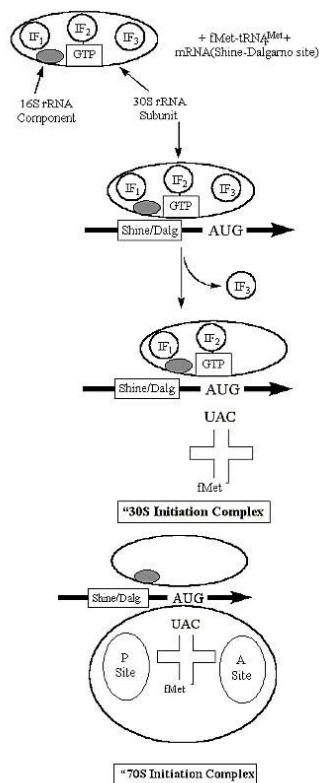


Figure 2.8.3: Initiation complexes

Elongation

1. In the first part of the **elongation step** of translation, the ribosome moves along the mRNA to position the fMet residue to the **P site** (peptidyl site) in the 50S subunit.
 - This allows the **second codon** of the mRNA to be positioned in the **A site** (amino acyl tRNA site).
2. The appropriate charged tRNA (with amino acid) specified by the second codon is positioned in the **A site** of the 50S subunit.
3. Next peptide bond formation is synthesized and the tRNA in the A site (which is covalently attached to the nascent polypeptide) is translocated to the **P site**.
 - This process requires **GTP and the G elongation factor** protein (prokaryotes).
4. The process is repeated.

Termination

1. When a stop codon is reached the **polypeptide** is hydrolyzed away from the last tRNA.
 - The peptide is released and the ribosome typically dissociates.
 - This process requires GTP and three different termination factors (TF's; only one required in Eukaryotes)

This page titled 2.8: mRNA Translation is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

CHAPTER OVERVIEW

3. Biotechnology 1

3.1: Gel Electrophoresis

3.2: DNA Sequence Analysis

3.3: Polymerase Chain Reaction (PCR) and Cloning of PCR Products

3.4: Prokaryotic Expression Vectors

3.5: Protein Sequencing, Peptide Mapping, Synthetic Genes

3.6: cDNA and Genomic Libraries

This page titled [3. Biotechnology 1](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

3.1: Gel Electrophoresis

Gel electrophoresis is used to characterize one of the most basic properties - molecular mass - of both polynucleotides and polypeptides. Gel electrophoresis can also be used to determine: (1) the **purity** of these samples, (2) **heterogeneity/extent** of degradation, and (3) subunit **composition**.

DNA

The most common gel electrophoresis materials for DNA molecules is **agarose** and **acrylamide**.

DNA agarose gels

The electrophoretic migration rate of DNA through agarose gels is dependent upon **four main parameters**:

1. The **molecular size** of the DNA. Molecules of linear duplex DNA travel through agarose gels at a rate which is inversely proportional to the log of their molecular weight.

$$M_r \propto 1/\log(Mw) \quad (3.1.1)$$

Example: Compare molecular mass vs. expected migration rate:

Molecular Mass (Da)	log (Molec. Mass)	1/log (Molec. Mass) i.e. relative Mr
100,000	5.0	0.20
50,000	4.7	0.21
10,000	4.0	0.25
5,000	3.7	0.27
1,000	3.0	0.33

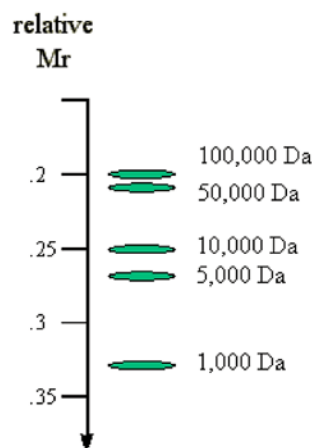


Figure 3.1.1: Relative migration rate with molecular mass

2. The **agarose concentration**. There is an inverse linear relationship between the logarithm of the electrophoretic mobility and gel concentration.

$$\log(M_r) \propto 1/[gel] \quad (3.1.2)$$

Gel %	1/Gel %	inv log(1/Gel %) (i.e. relative Mr)
2.0	0.50	3.2
1.5	0.67	4.6
1.0	1.00	10.0
0.5	2.00	100.0

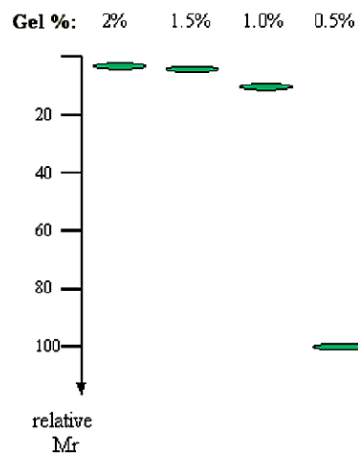


Figure 3.1.2: Relative migration rate with gel concentration

3. The conformation of the DNA.

- closed circular DNA (**form-I**) - typically supercoiled
- nicked circular (**form-II**)
- linear DNA (**form-III**)

These different forms of the *same* DNA migrate at different rates through an agarose gel. Almost always the **linear form** (form-III) migrates at the **slowest** rate of the three forms and supercoiled DNA (form-I) usually migrates the **fastest**.

4. The applied **voltage**.

- Typical value for running an agarose gel is **5 volts per cm** (length of gel).

Agarose (%)	Range of separation of linear DNA (in kilobases)
0.3	60 - 5
0.6	20 - 1
0.7	10 - 0.8
0.9	7 - 0.5
1.2	6 - 0.4
1.5	4 - 0.2
2.0	3 - 0.1

- Agarose gels are usually poured and run horizontally

Finally, the DNA being an acidic molecule, migrates towards the positively charged electrode (cathode).

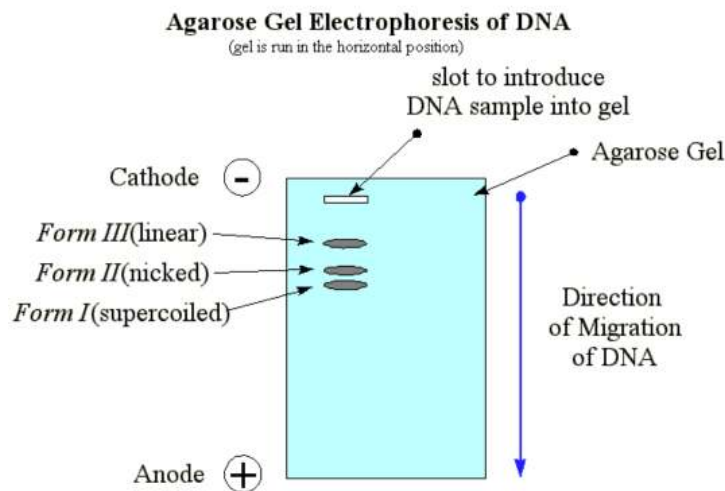


Figure 3.1.3: Gel electrophoresis setup

DNA acrylamide Gels

Acrylamide gels are useful for separation of **small DNA fragments** typically oligonucleotides <100 base pairs. These gels are usually of a *low acrylamide concentration* ($\leq 6\%$) and contain the non-ionic *denaturing agent* Urea (6M). The denaturing agent prevents secondary structure formation in oligonucleotides and allows a relatively accurate determination of molecular mass.

Gel Electrophoresis for Proteins

Gel electrophoresis of proteins almost exclusively utilizes **polyacrylamide**. The acrylamide solution usually contains **two components**: acrylamide and bis acrylamide. A typical value for the acrylamide:bis ratio is 19:1. The bis acrylamide is essentially a **cross-linking** component of the acrylamide polymer. The **total** acrylamide concentration in the gel affects the migration of proteins through the matrix (as with the concentration of agarose).

Protein gels are usually performed under **denaturing conditions** in the presence of the detergent **sodium dodecyl sulfate (SDS)**. The proteins are denatured by heat in the presence of SDS. The SDS binds, via hydrophobic interactions, to the proteins in an amount *approximately proportional to the size of the protein*. Due to the charged nature of the SDS molecule the proteins thus have a somewhat **constant charge to mass ratio** and migrate through the gel at a rate proportional to their molecular mass. The proteins migrate towards the **anode**.

Acrylamide (%)	Range of separation of Polypeptides (in kilodaltons)
8	200 - 25
10	100 - 15
12.5	70 - 10
15	60 - 6
20	40 - 4

Since the SDS treatment will dissociate non-covalent protein complexes, they may thus exhibit a much *lower than expected* molecular mass on SDS polyacrylamide gel electrophoresis (**SDS PAGE**). Protein PAGE gels are usually polymerized between two glass plates and run in the vertical direction.

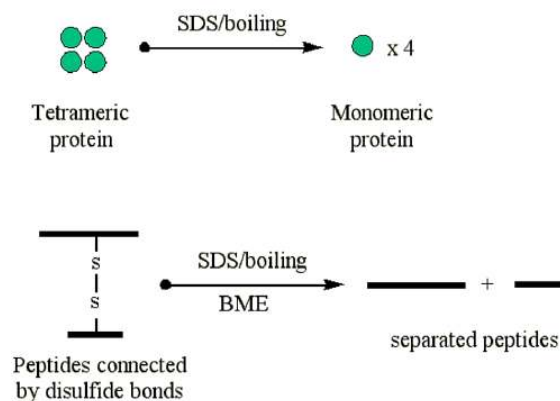


Figure 3.1.4: Effect of SDS treatment

PAGE may also be run in the presence of **reducing agents**, such as ***b*-mercaptoethanol (BME)**. BME is a reducing agent which will reduce any disulfide bonds (e.g. as exists between some pairs of cysteine residues in a protein). This helps to remove residual secondary structure in the SDS treated protein, but it may also allow the separation of polypeptide fragments from each other (i.e. their covalent interaction was entirely made up of one or more disulfide bonds). Thus, an apparently single protein may exhibit a set of small fragments under reducing PAGE conditions.

Stains for DNA

Ethidium

The most convenient method to visualize DNA in gel electrophoresis is staining with the fluorescent dye **ethidium bromide**. This compound contains a planar group that **intercalates** between the stacked bases of DNA. The orientation and proximity of ethidium with the stacked bases causes the dye to display an increased fluorescence compared to free dye (in solution). U.V. radiation at 254 nm is absorbed by the DNA and transmitted to the bound dye and the energy is re-emitted at 590 nm in the **red-orange** region of the spectrum.

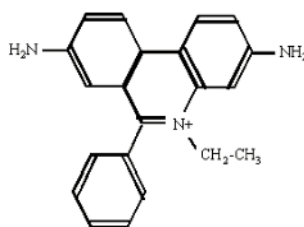


Figure 3.1.5: Ethidium

Ethidium bromide is usually prepared as a stock solution of 10 mg/ml in water, stored at room temp and protected from light. The dye is usually incorporated into the gel and running buffer, or conversely, the gel is stained after running by soaking in a solution of ethidium bromide (0.5 ug/ml for 30 min). The stain is visualized by irradiating with a UV light source (i.e. using a transilluminator) and photographing with polaroid film. The usual sensitivity of detection is better than 0.1 ug of DNA.

Because ethidium is a DNA intercalating agent, it is a powerful mutagen. Incorporation of ethidium in the DNA of living organisms (i.e. you and I) can cause (unwanted) mutations.

Note:

- The intercalation of ethidium bromide causes a reduction in the number of effective basepairs per twist.
- Thus, the DNA wants to adopt a conformation with fewer base pairs per twist.
 - The number of twists will increase for the given length of DNA
 - Writhe will decrease if the Linkage number remains unchanged
- The net result is that addition of ethidium bromide will introduce negative supercoils into the DNA

- *Addition of ethidium bromide to a DNA sample prior to electrophoresis can cause positively supercoiled DNA to migrate as relaxed (nicked, or form II) DNA*
- After replication DNA is typically *underwound* (negative supercoils), thus it will become *more* underwound with EtBr treatment (migration rate may actually *increase* due to more compact state)

Flourescence shadowing:

- DNA fragments resolved on polyacrylamide gels can also be visualized by the method of **UV shadowing**.
 - In this method the gel is placed on top of a fluorescent material, usually a flourescent TLC silica plate.
 - The gel is then illuminated by a UV light source.
 - DNA bands in the gel will block transmittance of the UV light to the substrate.
 - This will result in a dark area (i.e. non-fluorescing) area on the substrate.
 - This method is usually used in gel purification of oligonucleotides (i.e. the dark bands can be cut out with a razor blade and the oligonucleotide extracted)

Stains for Proteins

Coomassie brilliant blue

- Coomassie blue is a triphenylmethane **textile dye** which is able to stain proteins.
 - After a polyacrylamide gel is run it is usually "fixed" by placing in a 50% methanol/10 acetic acid solution for 30 minutes (to precipitate the proteins and prevent diffusion out of the gel).
 - The fixed gel is then soaked in a methanol/acetic acid solution containing 2.5 gm/liter of Coomassie blue.
 - Destaining of the background gel is accomplished by soaking in changes of 10% methanol/7 % acetic acid. This method can typically detect protein samples of 0.1 ug or greater.

Silver stain

- Silver staining is a method which utilizes a silver nitrate solution to stain proteins in an acrylamide gel.
- The method is similar in nature to the use of silver in photographic plates.
- The sensitivity is approximately **two orders of magnitude more sensitive than coomassie staining** (i.e. one can detect approximately 1ng of protein).
- Due to its high sensitivity this method is usually used to determine the presence of trace contaminants in protein samples.

Molecular Weight Standards

- Both DNA and protein gel electrophoresis utilize molecular weight standards to calibrate the size(s) of samples being analyzed
- DNA molecular weight standards will consist of a mixture of DNA fragments of known sizes (molecular mass)
 - One convenient DNA molecular weight standard is constructed by partial ligation of a 100 base pair fragment of duplex DNA
 - Partial ligation will result in formation of dimers (200 bp), trimers (300 bp) and so on, as well as some amount of the original (100 bp) fragment
 - This produces a DNA "ladder" after gel electrophoresis
 - Another DNA fragment standard might be a known DNA sequence, such as the plasmid pBR322, which has been digested with a four-cutter restriction endonuclease (e.g. Alu I).
 - This produces a variety of fragments with different sizes
 - It is easily reproduced (just cut more pBR322)
- Protein molecular weight markers usually consist of a mixture of a half dozen or so pure proteins with known molecular masses

Protein	Molecular Mass (Da)
Phosphorylase B	94,000
Bovine Serum Albumin	67,000
Ovalbumin	43,000
Carbonic Anhydrase	30,000
Soybean Trypsin Inhibitor	20,100
α -Lactalbumin	14,400

- In both DNA and protein gels, molecular weight markers are run in a lane at one end of the gel

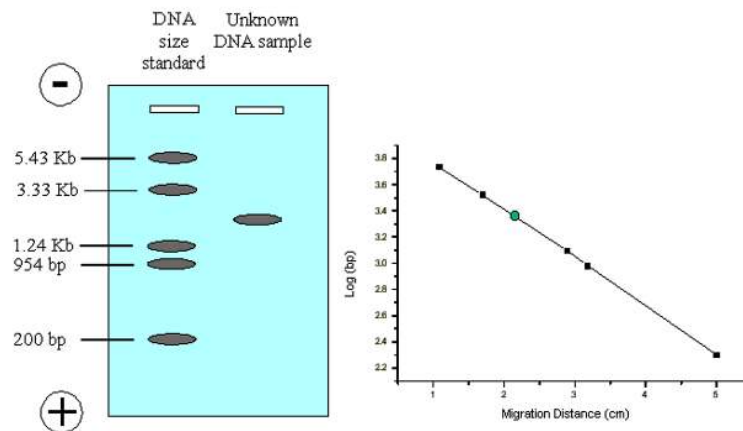


Figure 3.1.6: DNA size standards

This page titled [3.1: Gel Electrophoresis](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

3.2: DNA Sequence Analysis

DNA sequencing is most often accomplished using a procedure referred to by one of the following names:

1. Sanger sequencing
 2. Di-deoxy sequencing
 3. Chain termination sequencing
- Each of these refers to the same method:
 - the use of **di-deoxy base** incorporation in a **polymerization** reaction
 - leads to termination of primer extension (a method pioneered by Fred Sanger, now retired and happily puttering about in his garden).

The basic method involves:

1. annealing a **primer** 5' to a region of DNA we would like to sequence.
2. The primer is extended in the traditional manner (i.e. with DNA polymerase and the four dNTP's).
3. However, a small concentration of **di-deoxy bases** are included in the reaction mix.
 - Usually this is accomplished by having four separate reactions, into which one of the ddNTP's is added.
 - Thus, one tube would contain primer, template, DNA polymerase, the four dNTP's and ddATP, another tube would have the same thing but with ddTTP instead of ddATP, and so on for ddCTP and ddGTP.
4. During the reaction, the normal dNTP's are incorporated into the growing chain.
 - However, occasionally the DNA pol will incorporate a *dd*base into the growing primer.
 - When this happens, the primer **cannot** be extended any further (because the 3' *dd* base does not have an available 3' hydroxyl group).
 - *The resulting DNA fragment begins at the 5' end of the sequencing primer and ends at the site of dd base incorporation.*

Thus, in the reaction mixture containing the dd base ddATP, there will result an ensemble of fragments of varying lengths, each ending in with the ddA base (i.e. at all positions in the template where there was a complementary 'T' base).

The mixture containing ddCTP will have a different mix of fragments - they will contain ddC at the 3' ends (at positions in the template where 'G' bases were located).

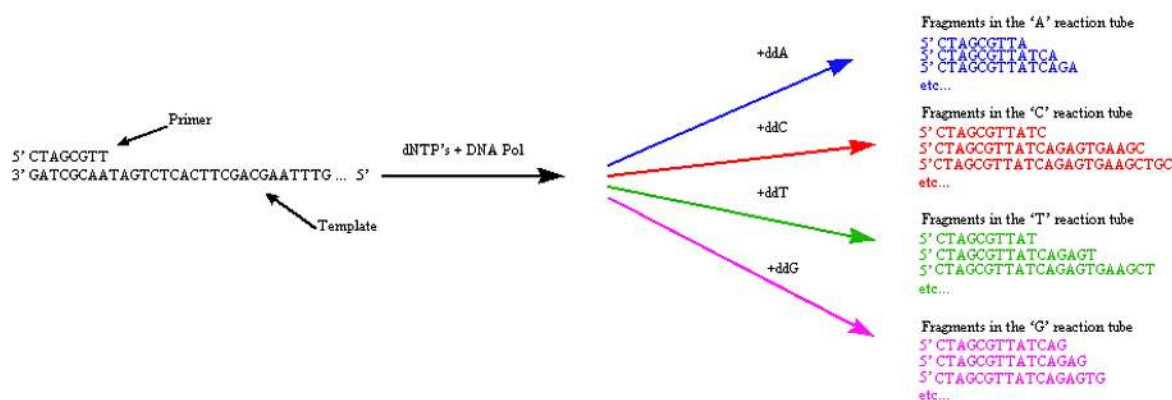


Figure 3.2.1: DNA fragments

- If the fragments from the 'A' reaction mix are run on a urea/acrylamide gel (typically 6%) the fragments will separate according to size.
- Likewise for the 'C', 'G' and 'T' reaction mix fragments.
 - If the four different reaction mixtures are run next to each other the fragment sizes can be **directly compared to one another**.
 - Note that the shortest expected fragment is the primer itself and the longer the fragment, the further from the primer the extension reaction went before termination.
- Consider a case where the template has a stretch of six 'A' bases in a row.

- Automated sequencers can thus read **further** than with manual methods.
- Since a single lane is used per sample (as compared to four lanes with the radiolabeled method) many more samples can be analyzed and the **throughput** is greater
- The acrylamide gels used for sequence analysis are typically 50 cm to 100 cm long.
 - In manual sequencing the four reaction mixes are loaded and the gel is run for approximately 2 hours then the samples are reloaded on another part of the gel and the gel run is continued. A third set of samples may be loaded after another 2 hours.
 - The gel is stopped after the dye front of last sample loaded has just reached the bottom of the gel.
 - Thus, the short fragments can be visualized in the last load, medium fragments in the second load and the long fragments will be visualized in the first set of reaction mixtures loaded.
 - Manual sequencing can resolve on the order of 400 bases of continuous sequence. Automated sequencers can routinely provide twice this amount of information.
 - Automated sequencers use the same types of glass plates
 - The continuous running of the gel (and dye identification) means that typically 400-700 bases or more can be read
 - Automatic software will interpret the dye signals into a sequence
 - nuances of the sequencing chemistry and expert knowledge can be programmed into the sequence analysis software (e.g. the software can compensate for the "smile" of the gel)

This page titled [3.2: DNA Sequence Analysis](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

3.3: Polymerase Chain Reaction (PCR) and Cloning of PCR Products

PCR is an *in vitro* technique for the amplification of a region of DNA which lies between two regions of known sequence. PCR amplification is achieved by using oligonucleotide primers. These are typically short, single stranded oligonucleotides which are complementary to the outer regions of known sequence.



Figure 3.3.1: Complementary oligonucleotide primers

The oligonucleotides serve as **primers** for DNA polymerase and the denatured strands of the large DNA fragment serves as the **template**. This results in the synthesis of new DNA strands which are complementary to the parent template strands. These new strands have **defined 5' ends** (the 5' ends of the oligonucleotide primers), whereas the 3' ends are potentially ambiguous in length.

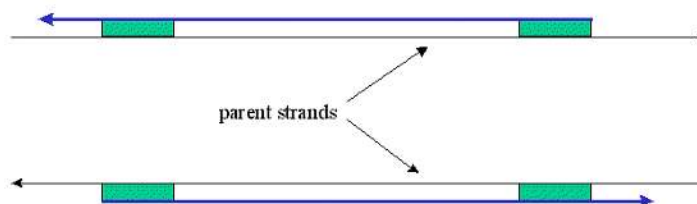


Figure 3.3.2: Synthesis of new DNA strands

The oligonucleotide directed synthesis of daughter DNA strands can be repeated if the new duplex is denatured (by heating) and additional primers are allowed to anneal (by cooling to an appropriate temperature).

The steps of:

1. **Template denaturation**
2. **Primer annealing**
3. **Primer extension**

comprise a single "cycle" in the PCR amplification methodology.

After each cycle the newly synthesized DNA strands can serve as templates in the next cycle.



Figure 3.3.3: Product of elongation reaction as template

Note that half of the newly synthesized strands from the second round of replication have 5' and 3' termini which are defined by the annealing location of the primer oligonucleotides.

Summary of products at the end of each PCR cycle:

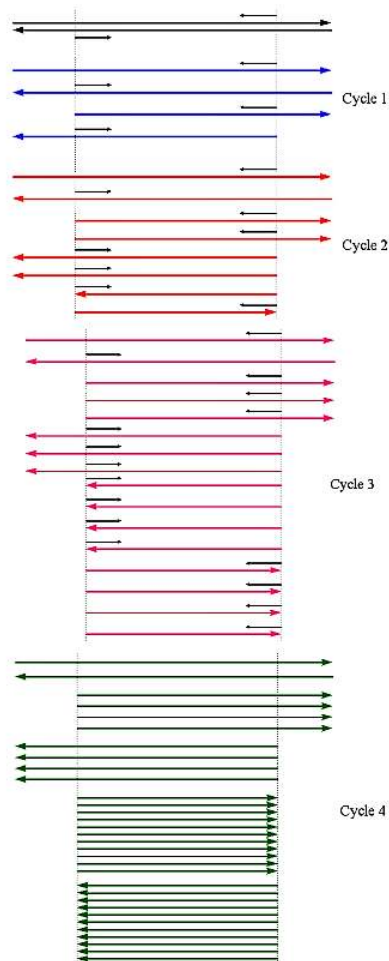


Figure 3.3.4: PCR products

The total of each of the different types of oligonucleotide fragments for each cycle can be summarized as follows:

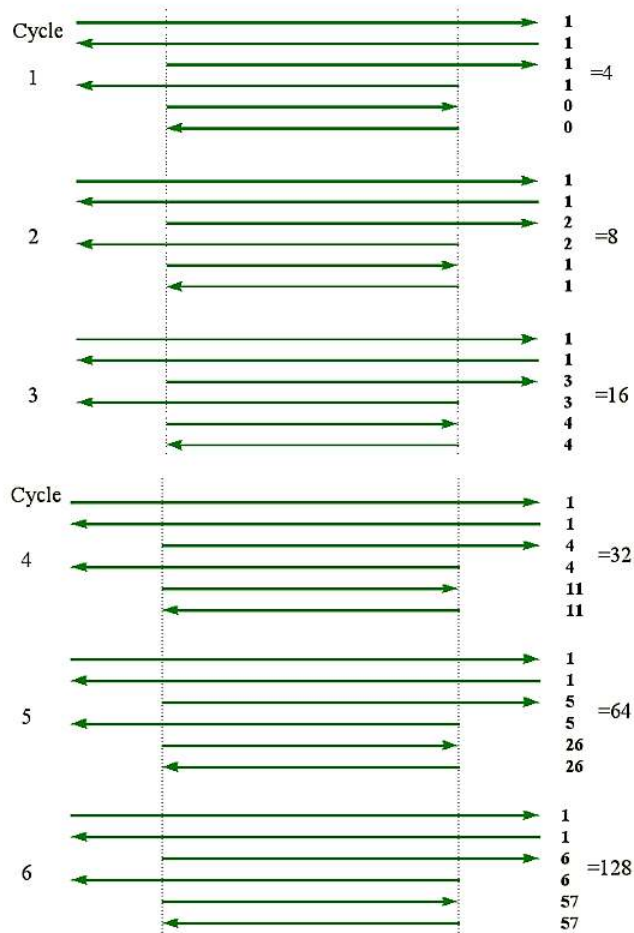


Figure 3.3.5: Types of fragments for each cycle

- The amplification of the fragments follows the following pattern:
 - There is always 1x copy of each of the original templates - the PCR as outlined never reproduces the full length template
 - There will be 'n'x copies of each fragment with an indeterminate length (where 'n' is the number of cycles). The fragments of indeterminate length have one end defined by a PCR primer and the other end is of indeterminate length
 - There will be $(2^n - (n + 1))$ x copies of each fragment of defined length (i.e. each end defined by the two PCR primers).
- For example, after 6 cycles we will have
 - 1 x copy of each original template
 - 6 x copies of each fragment of indeterminate length
 - $(64 - (6 + 1)) = 57$ x copies of each fragment of defined length
- In the above case, the desired PCR product will be a duplex of the defined length fragment. The question is: how many will be produced?
 - The original templates will not necessarily anneal with one another
 - Likewise, the templates of indeterminate length will not necessarily anneal with one another
 - Furthermore, as the number of cycles proceeds, the defined length fragments far outnumber the others
 - Therefore, the original template, and fragments of indeterminate length are most likely going to hybridize with defined length fragments**
 - Our expected **amplification** would therefore be (using the 6 cycle PCR example)

$57 - (6 + 1) = 50$ x copies of each defined length fragments

(i.e. 50 x duplexes of defined length)

- The expected **amplification** of the desired defined length product with respect to the original template concentration (x) can thus be represented by the formula:

$$[(2^n - (n + 1)) - (n + 1)]x \quad (3.3.1)$$

or

$$(2^n - 2(n + 1)) \quad (3.3.2)$$

(this is often abbreviated to a simple rule of thumb for the **amplification**: $(2^n - 2n) x$)

The interpretation of this formula is that

- For a given number of cycles 'n' we make ' $2^n x$ ' **total possible duplexes**
- For a given number of cycles there will be ' $2(n+1)$ ' (or $2n$ in our approximation) x ' duplexes which are formed from either the original template, or a fragment of indeterminate length, along with a fragment of defined length (and represent an undesired product)
- Thus, the total concentration of desired product (duplexes with a length defined by the PCR primers) will be $(2^n - 2(n+1)) x$ (where x is the concentration of the original duplex)
- The theoretical amplification value is never achieved in practice. Several factors prevent this from occurring, including:
 1. Competition of complementary daughter strands with primers for reannealing (i.e. two daughter strands reannealing results in no amplification).
 2. Loss of enzyme activity due to thermal denaturation, especially in the later cycles
 3. Even without thermal denaturation, the amount of enzyme becomes limiting due to molar target excess in later cycles (i.e. after 25 - 30 cycles too many primers need extending)
 4. Possible second site primer annealing and non-productive priming

Note

PCR was invented in 1985 by Kary Mullis, working for Cetus corporation somewhere near Berkeley, California. The original method of PCR used the Klenow fragment of *E. coli* DNA polymerase I. This enzyme, however, denatures at temperatures lower than that required to denature most template duplexes. Thus, after each cycle, fresh enzyme had to be added to the reaction. This was quite tedious. In addition to this problem with the enzyme, the samples had to be moved from one temperature bath to another to allow the individual steps of denaturation, annealing and polymerization, which all required different temperatures. This was pretty tedious too.

Thus, **two main advances** allowed the process to be **automated**, these advances were:

1. The use of **thermostable DNA polymerases**, which resisted denaturation (inactivation) at high temperatures. Thus, an initial aliquot of polymerase could last throughout the numerous cycles of the protocol. The first thermostable DNA polymerase to be used was isolated from the bacterium *Thermus aquaticus*. It was isolated from a hot spring in Yellowstone National Park where it lived happily (i.e. it replicated its DNA) at temperatures in excess of 85 °C
2. The development of temperature baths which could shift their temperatures up and down rapidly and in an automated programmed manner. These are known as **thermal cyclers** or **PCR machines**.

These two developments led to the automation of PCR. The PCR process is covered by patents owned by Hoffmann-La Roche Inc. (a faceless conglomerate you can trust).

Thermal cycling parameters

- The thermal cycling parameters are critical to a successful PCR experiment. The important steps in each cycle of PCR include:
 1. **denaturation of template**
 2. **annealing of primers**
 3. **extension of the primers**

A representative temperature profile for each cycle might look like the following:

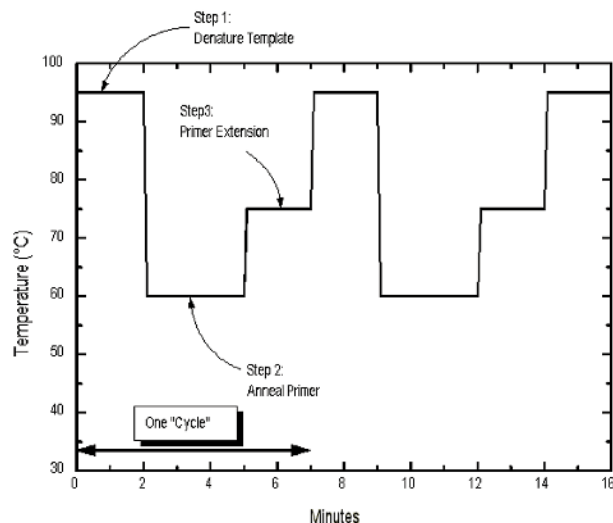


Figure 3.3.6: Temperature for PCR cycle

Template denaturation

- The *initial* denaturation of template is accomplished at 95-100 °C.
 - Supercoiled plasmids are tougher to melt and may require boiling for several minutes, or may be initially denatured by using base (NaOH, followed by pH neutralization).
 - Denaturation during the PCR experiment (i.e. second cycle onward) is usually accomplished at temperatures of 92-95 °C (usually empirically determined).

Primer annealing temperature

- Primer annealing temperature is an important parameter in the success of the PCR experiment.
- The annealing temperature is characteristic for each oligonucleotide:
 - it is a function of the length and base composition of the primer as well as the ionic strength of the reaction buffer.
 - Estimates of the annealing temperature can be calculated in several different ways.
 - These calculated annealing temperatures are a **starting point** for the PCR experiment, but ideal annealing temperatures are determined empirically.

Primer extension

- Primer extension is usually performed at 72 °C, or the *optimum temperature of the DNA polymerase*.
- The length of time of the primer extension steps can be increased if the region of DNA to be amplified is long, however, for the majority of PCR experiments an extension time of 2 minutes is sufficient to get complete extension.

Number of cycles

- The number of cycles is usually between 25 and 35.
 - More cycles mean a greater yield of product.
 - However, with increasing number of cycles the greater the probability of generating various artifacts (e.g. mispriming products).
 - It is unusual to find procedures which have more than 40 cycles.

Choice of Polymerases for PCR

- One of the important advances which allowed development of PCR was the availability of thermostable polymerases.
- This allowed initially added enzyme to survive temperature cycles approaching 100 °C.

Thermostable DNA polymerases and their sources

DNA Polymerase	Natural or recombinant	Source
Taq	Natural	<i>Thermus aquaticus</i>
Amplitaq®	Recombinant	<i>T. aquaticus</i>
Amplitaq (Stoffel fragment)®	Recombinant	<i>T. aquaticus</i>
Hot Tub™	Natural	<i>Thermus flavus</i>
Pyrostase™	Natural	<i>T. flavus</i>
Vent™	Recombinant	<i>Thermococcus litoralis</i>
Deep Vent™	Recombinant	<i>Pyrococcus GB-D</i>
Tth	Recombinant	<i>Thermus thermophilus</i>
Pfu	Natural	<i>Pyrococcus furiosus</i>
ULTma™	Recombinant	<i>Thermotoga maritima</i>

- Properties of DNA polymerases used in PCR

	Taq/Amplitaq®	Stoffel fragment	Vent™	Deep Vent™	Pfu	Tth	ULTma™
95 °C half-life	40 min	80 min	400 min	1380 min	>120 min	20 min	>50 min
5'3' exo	+					+	
3'5' exo			+	+	+		+
Extension rate (nt/sec)	75	>50	>80	?	60	>33	?
RT activity	Weak	Weak	?	?	?	Yes	?
Resulting ends	3' A	3' A	>95% blunt	>95% blunt	blunt	3' A	blunt
Strand displacement			+	+			
M.W. (kDa)	94	61	?	?	92	94	70

Buffers and MgCl₂ in PCR reactions

A typical reaction buffer for PCR would something like:

- 10 mM Tris, pH 8.3
- 50 mM KCl
- 1.5 mM MgCl₂
- 0.01% gelatin
- The MgCl₂ concentration in the final reaction mixture is usually between 0.5 to 5.0 mM, and the optimum concentration is determined empirically (typically between 1.0 - 1.5 mM). **Mg²⁺ ions:**
 - form a soluble complex with dNTP's which is essential for dNTP incorporation
 - stimulate polymerase activity
 - increase the T_m (melting temperature) of primer/template interaction (i.e. it serves to stabilize the duplex interaction)

Generally,

- **low** Mg²⁺ leads to low yields (or no yield) and
- **high** Mg²⁺ leads to accumulation of nonspecific products (mispriming).

Primers

Primer design

- Generally, primers used are 20 - 30 mer in length. This provides for practical annealing temperatures (of the high temperature regimen where the thermostable polymerase is most active).
- Primers should avoid stretches of polybase sequences (e.g. poly dG) or repeating motifs - these can hybridize with inappropriate register on the template.
- Inverted repeat sequences should be avoided so as to prevent formation of secondary structure in the primer, which would prevent hybridization to template
- Sequences complementary to other primers used in the PCR should be avoided so as to prevent hybridization between primers (particularly important for the 3' end of the primer)
- If possible the 3' end of the primer should be rich in G, C bases to enhance annealing of the end which will be extended
- The distance between primers should be less than 10 Kb in length. Typically, substantial reduction in yield is observed when the primers extend from each other beyond ~3 Kb.

Melting temperature (T_m) of primers

- The T_m of primer hybridization can be calculated using various formulas. The most commonly used formula is:

$$T_m = [(number\ of\ A + T\ residues) \times 2^\circ C] + [(number\ of\ G + C\ residues) \times 4^\circ C] \quad (3.3.3)$$

- This formula was determined originally from oligonucleotide hybridization assays, which were performed in 1 M NaCl, and appears to be accurate in lower salt conditions only for primers less than about 20 nucleotides in length.
- The common wisdom is that the T_m is more like 3-5 °C lower than the value calculated from this formula.

$$T_{ms} = 81.5 + 16.6(\log_{10}[J^+]) + 0.41(\%G + C) - (600/l) \quad (3.3.4)$$

- Where $[J^+]$ = the **molar** concentration of monovalent cations (e.g. Na^+ from NaCl), and l = the length of oligonucleotide. (%G+C) is the actual percentage value and not a fractional representation (i.e. the value to insert for a primer which had 90 % G+C content would be "90" and not "0.90").
- This formula is reportedly useful for primers of 14 to 70 bases in length.

$$T_{mp} = 22 + 1.46([2 \times (G + C)] + (A + T)) \quad (3.3.5)$$

- This formula is reportedly useful for primers of 20-35 bases in length.
- The calculated annealing temperature is only a **reference temperature** from which to initiate experiments.
 - The actual annealing temperature may be 3-12 °C higher than the calculated T_m .
 - The actual annealing temperature condition should be determined empirically.
 - The highest annealing temperature which gives the best PCR product should be used.
- Examples of T_m , T_{ms} and T_{mp} calculations (0.05 M K^+)

	G	A	T	C	T_m	T_{ms}	T_{mp}
15 mer	3	5	2	5	46	42	56
20 mer	6	5	4	5	62	52	67
30 mer	8	6	8	8	92	62	89

Calculating primer concentrations

- The molar concentration of a primer can be calculated based on the absorbance of the primer at 260 nm (A_{260}) and the **molar extinction coefficient** for the primer at this wavelength.
- The molar extinction coefficient for the primer can be calculated by knowing the sequence of the primer and then summing the molar extinction values for the individual bases which comprise the primer.
- The individual bases have the following molar extinction coefficients at 260 nm:
 - A 1.0 molar solution of dT has a value of 8,400 absorbance units at 260 nm.

- A 1.0 molar solution of dA has a value of 15,200 absorbance units at 260 nm.
- A 1.0 molar solution of dG has a value of 12,010 absorbance units at 260 nm.
- A 1.0 molar solution of dC has a value of 7050 absorbance units at 260 nm.
- For example, the primer 5' TAGC 3' would have a molar extinction coefficient of 42,660 at 260 nm. Likewise, a 10 micromolar solution of this primer would give an absorbance of 0.427 at 260 nm.

Cloning PCR Products

Introduction of restriction sites

- It is possible to introduce restriction site sequences into PCR products by having these sequences incorporated into the **5' end** of the PCR primer(s).

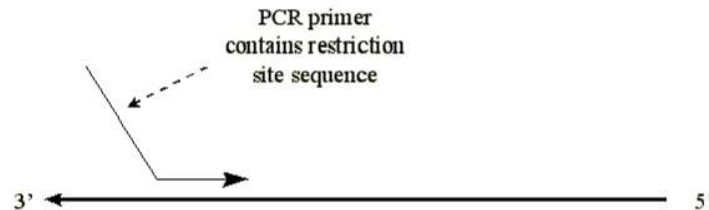


Figure 3.3.7: Restriction site sequence introduction

- The short restriction site sequence on the 5' end of the PCR primer will not hybridize, but as long as the 3' hybridizing region is long enough (i.e. its T_m is high enough; ~20 mer) the primer will specifically bind to the appropriate site.
- The PCR product will thus have an additional DNA sequence at the 5' end which will contain the endonuclease restriction site.
- A similar or different restriction site sequence can be added via the other PCR primer.
- If the other primer has a different restriction sequence then the PCR fragment can be inserted in a directional dependent manner in a host plasmid.

The potential problems with this method include:

- There is no easy way to prevent internal sites containing similar restriction sequences from being cut when the end of the PCR product are cut
- Restriction sequences are inverse repeat sequences, thus the potential exists for primer dimer association and resultant non-productive annealing

Generation of half sites

- This method is similar to the method of introducing restriction sites, described above.
- The primary difference is that instead of the primer containing the entire restriction site sequence (say the six nucleotides of a six cutter) it will contain only the last three (and the other PCR primer will contain the complementary sequence for the first three).

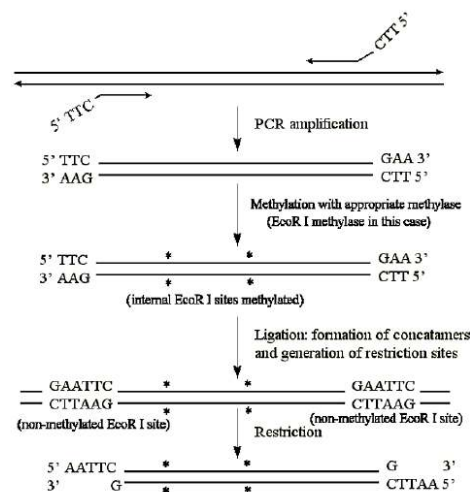


Figure 3.3.8: Half sites

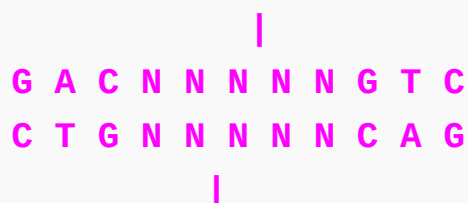
The advantages of this method are:

- Typically internal restriction sites cleave with much greater efficiency (i.e. some sites if located at the ends of linear DNA never cut well at all)
- There is no need to gel purify linker fragments after digestion
- The DNA can be methylated (the half sites will not be). After concatenation the linkers will be cut but internal restriction sites will not
- A disadvantage is that the same restriction site is incorporated into both ends so the PCR fragment cannot be ligated into a host vector in an orientation dependent manner.
- Also, in this method 3' A overhang cannot be tolerated.

Blunt end ligation

- As stated earlier some thermostable DNA polymerases add a single dA residue onto the 3' end of the PCR product.
- There are three choices to be made when attempting to subclone without the use of added restriction sites within the primers:
 - Use a DNA polymerase which leaves the 3' strand blunt (e.g. Vent) and do a blunt end ligation (i.e. host vector was opened up with blunt cutting restriction endonuclease)
 - "Fix" the 3' A overhang by chewing back with Pol I, dNTP's.
 - Use the 3' A overhang to anneal and ligate to a "T" vector - a vector which has a single dT overhang on its 3' ends.
- "T" vectors can be made by opening up a vector, via a blunt cutting restriction endonuclease and ligating in a specific linker.
- The linker contains a restriction sequence for a restriction endonuclease which recognizes an interrupted palindrome and cuts in the internal region with a single 3' overhang.
- The linker will contain two copies of the restriction recognition sequence, the first with a sequence in the interrupted palindrome which leaves a 3' T overhang and the second with a sequence in the interrupted palindrome which leaves another 3' T overhang
- This may sound confusing, so here's an example of how to make a T vector:

Ahd I restriction endonuclease:



Ahd I will cut this sequence to produce:

G A C N N N	N N G T C
C T G N N	N N N C A G

We could design an oligonucleotide with two Ahd I restriction sequences, with slightly different sequences in the interrupted region of the palindrome, to give:

G A C N N	<u>I</u>	N N G T C	G A C N N
C T G N N	<u>A</u>	N N C A G	C T G N N

If this this were inserted into a vector, and the vector then was cut with Ahd I, it would have the following sequence at the ends of the linearized vector:

-G A C N N <u>I</u>	N N G T C-
-C T G N N	<u>I</u> N N C A G-

In other words, a 3' 'T' overhang at both ends of the vector. A PCR product, with 3' A overhangs could thus be inserted into such a 'T' vector

Adding promoters, ribosome binding sites, start codons, and stop codons

- The ability to add unique sequences to the 5' ends of PCR primers allows for short control elements to be **directly incorporated**.
- These can include a start codon or stop codon (3 bases), a promoter (~30 nucleotide region) or a ribosome binding site (~8 bases).

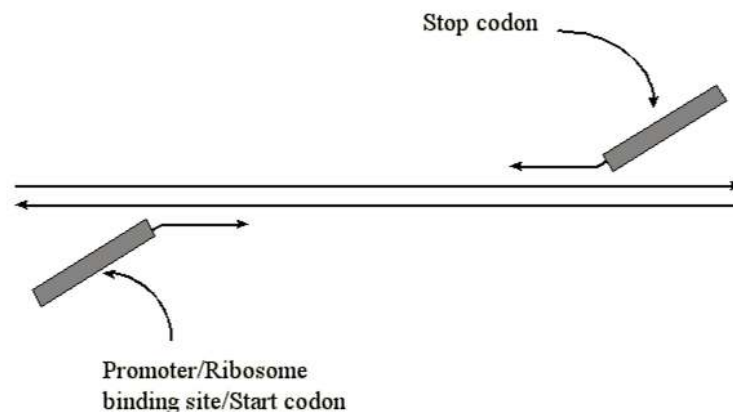


Figure 3.3.9: Inserting promoter/ribosome/binding site/start codon/stop codon

PCR Mutagenesis

Gene fusion

- This method is useful for joining overlapping regions of a large gene, or for the construction of **chimeric** genes.

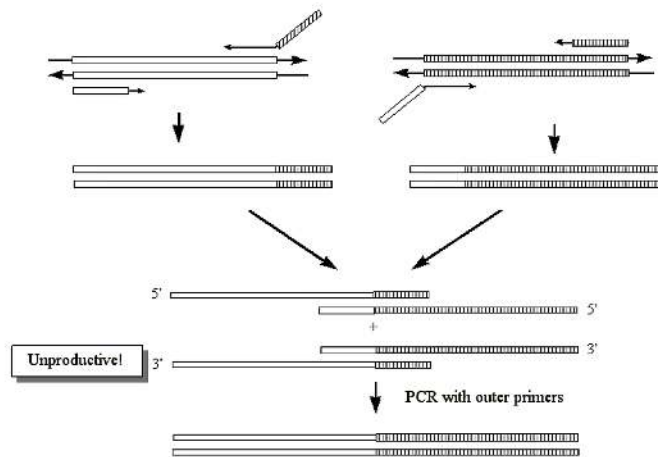


Figure 3.3.10: Gene Fusion

Creation of deletions within a gene

- A very similar methodology can be used within a single gene for the production of a mutant gene containing a specific deletion:

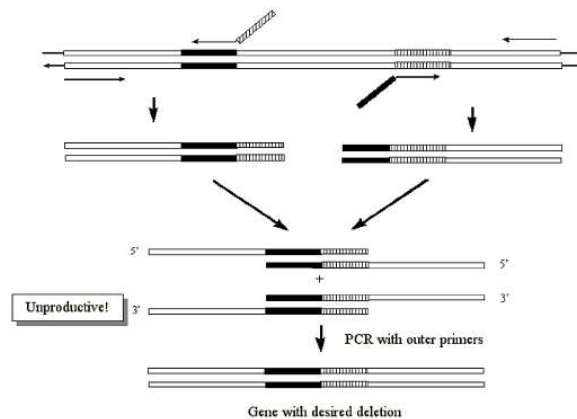


Figure 3.3.11: Deletions

- If the gene is contained within circular DNA (i.e. a plasmid) deletions can be constructed in a single PCR reaction with a single set of primers (this type of methodology is also known as **"inverse" PCR**).

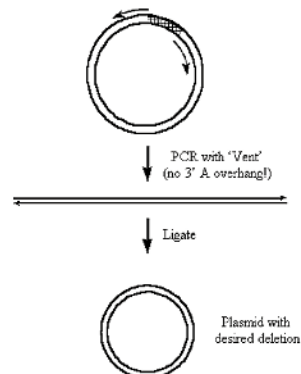


Figure 3.3.12: Inverse PCR

Generation of point mutation(s) - i.e. base substitution mutations

- The generation of base substitutions can proceed along a similar route as with the deletion mutations.
- However, in this case the primers are mutagenic - there will be a mismatch, or mismatches, between the primer and target sequences.
- The mutagenic oligo will have a lower than expected T_m due to this mismatch(es).

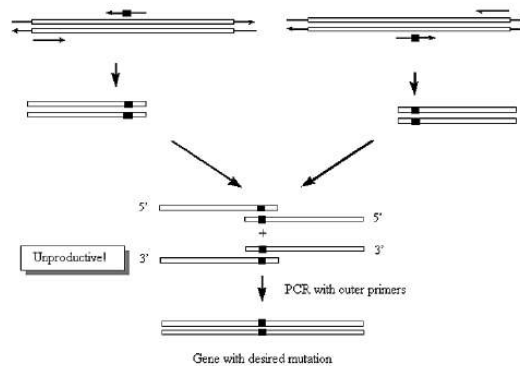


Figure 3.3.13: Point mutation

Introduction of base substitutions via asymmetric PCR:

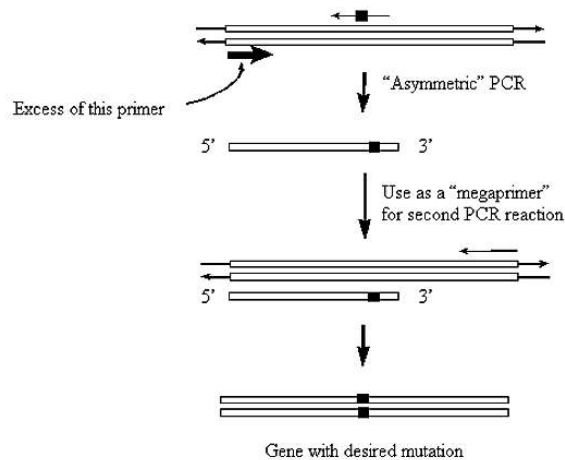


Figure 3.3.14: Substitution with asymmetric PCR

Insertion mutagenesis

- Short insertions (~1-6 basepairs) can be incorporated directly into a PCR primer, either internally, or at the 5' end.
- If the template DNA is linear and the desired site of insertion is not at the end of the template, then the entire gene (plus insertion) can be produced using asymmetric PCR or overlapping PCR (i.e. shown above).

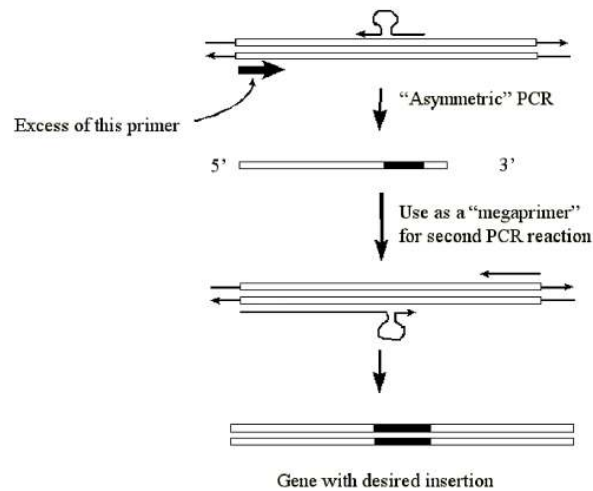


Figure 3.3.15: Insertion mutagenesis

- Large insertions can be accomplished by using a template (the desired insertion) for PCR with the primers having 5' sequences which are complementary to the region of insertion in the desired gene:

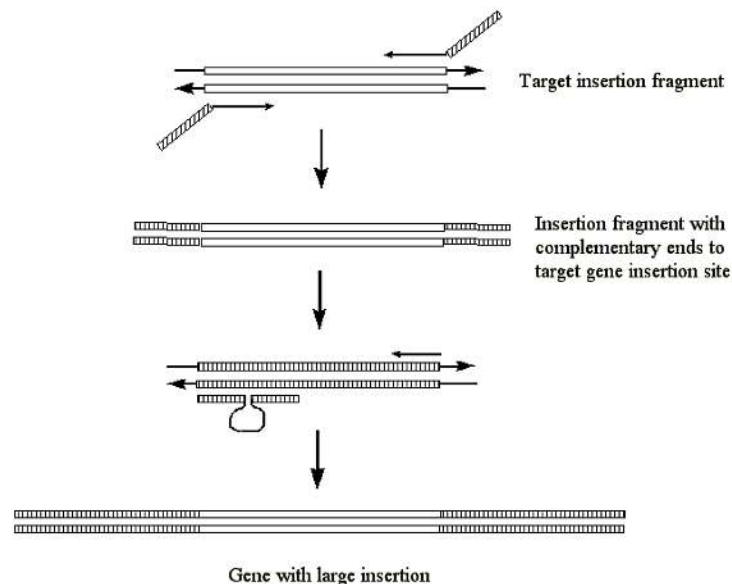


Figure 3.3.16: Large insertion

"Random" mutagenesis with PCR

- The PCR protocol can be modified so as to introduce mutations at random positions in the target DNA.
- The principle behind the mutagenesis is **misincorporation** of bases at "random" positions.
- Misincorporation by Taq polymerase, for example, can be achieved by adding Mn^{2+} to the reaction buffer, and decreasing the concentration of one of the four dNTP's.
 - At the sites in the template where the reduced base should be incorporated, there will be an increased probability of misincorporation.
 - Thus, the choice of base with diminished concentration determines the sites in the template which will potentially be mutated.
 - The misincorporated base is more or less random.

- The ideal Mn^{2+} concentration to add varies between 0.1 to 0.5 mM and is determined empirically. The relative concentrations of bases is 1 mM for each base, except the reduced base, which is typically present at a 1:5 or 1:10 ratio (i.e. 0.2 to 0.1 mM).

This page titled [3.3: Polymerase Chain Reaction \(PCR\) and Cloning of PCR Products](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

3.4: Prokaryotic Expression Vectors

The first vector we will consider is the *pUC* family of vectors

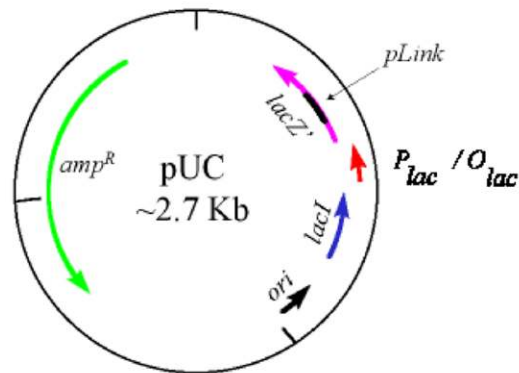


Figure 3.4.1: *pUC* vector

Although it is not typically used for the expression of recombinant proteins, it has all the necessary elements of an expression vector:

1. **An origin of replication.** The *pUC* family of vectors are *high copy* vectors. They have a ColE1 origin of replication, but a deletion of the *rop* replication regulatory region.
2. **A drug resistance marker.** The *pUC* vectors contain a gene for *ampicillin resistance* (b-lactamase).
3. **An inducible promoter.** This family of vectors contains the *lac* promoter (P_{lac}) along with the associated *lac* Operator region (O_{lac})

These vectors contain a couple of other elements which add utility to the vector:

1. **The *lacI* gene.** This gene codes for the *lac* repressor protein. Being a high copy plasmid the host *lac* repressor levels may not be sufficient to efficiently repress the *lac* operator on the plasmids. The plasmid produces *lac* repressor to augment the host levels.
2. **The *lacZ'* gene.** This gene product is transcribed from the *lac* promoter and produces an amino-terminal fragment of the b-galactosidase protein.
3. **A polylinker (multiple cloning site).** This short stretch of DNA is located just downstream from the *lac* promoter and after the first few codons of the *lacZ'* gene. It is a short stretch of nucleotides which contains a variety of restriction endonuclease sites.

a-complementation

The *lac* promoter of the *pUC* plasmid can be induced using the lactose analogue *isopropylthiol-b-D-galactoside*. ("IPTG").

- This will result in the expression of the *lacZ'* gene.
- This gene codes for an amino-terminal fragment of the b-galactosidase protein
- By itself, this fragment is *non-functional* (it will not hydrolyze lactose or other b-galactosides)

However, b-galactosidase is an interesting protein:

- If we express the carboxy-terminal fragment of the protein it is also non-functional
- But, if we combine both peptide fragments they can combine to yield a **functional** b-galactosidase protein
- This is termed **a-complementation**

The *lacZDM15* mutation in a bacteria means that its **genome** has suffered a deletion of the amino-terminal region of the *lacZ* gene.

- Such a bacteria has a *non-functional* b-galactosidase protein
- If this type of bacteria contains an extra-chromosomal element which expresses the *lacZ'* peptide, it can complement the *lacZDM15* protein to produce functional b-galactosidase protein (*a-complementation*)

The galactoside, *5-bromo-4-chloro-3-indoyl-b-D-galactoside* ("X-gal"), can be hydrolyzed by b-galactosidase and produces a **dark blue** color.

- *lacZ*DM15 bacteria in the presence of lactose (or IPTG) and X-gal will appear a normal white-yellow color.
- The same bacteria which harbor the above pUC plasmid will be **dark blue** in color when IPTG and X-gal are present in the media (due to a-complementation and a functional b-galactosidase protein)

The pUC polylinker region

The pUC polylinker region is a short stretch of DNA which is actually inserted *within* the *lacZ'* gene, just downstream from the start codon:

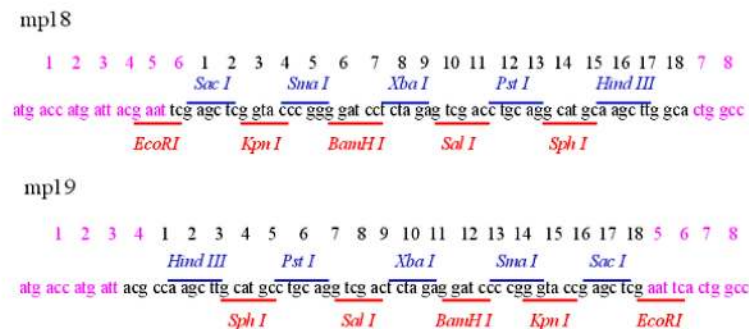


Figure 3.4.2: pUC Polylinker region

Note

In the above diagram the codon numbers for the wild-type b-galactosidase are given in magenta (in this case the initiator methionine is not present on purified protein and the first codon is considered to be the ACC (threonine) residue). The codons from the polyliner (which interrupt the b-galactosidase gene) are given in black.

- The different available plink regions are all a multiple of 3 nucleotides in length, therefore the *lacZ'* gene remains *in-frame*.
- The resulting *lacZ'*/plink peptide *can still function in a-complementation*.

What would happen to the *lacZ'* peptide if we inserted a DNA fragment into the plink region?

- The insertion of a random piece of DNA into the plink region will result in *three possible different reading frames* after the insert (depending on the number of nucleotides in the inserted fragment)
- There is a 33% (one in three) possibility that the downstream *lacZ'* gene will be translated in the correct reading frame
- There is also the possibility that even if the reading frame is correct that the insertion of a large peptide at the start of the *lacZ'* protein will prevent a-complementation.
- Therefore, *most of the time*, a DNA fragment inserted into the plink region will *abolish b-galactosidase function*
- DNA inserts in the plink region can therefore be identified by growing the host bacteria on media containing IPTG and X-gal and looking for white-yellow colonies (with the caveat that some blue colonies may in fact contain an insert)

DNA coding for a gene of interest can be inserted in the plink region in-frame with the *lacZ'* reading frame.

- The protein coded for by this gene can be expressed by induction with IPTG (i.e. using the *P_{lac}*)
- The expressed protein will have as its amino-terminal sequence the first few amino acids of the b-galactosidase gene (in the absence of any other manipulations/mutagenesis). This is also known as a (short) *fusion protein*.
- Although the *lac* promoter is considered to be relatively weak, the high copy number plasmid may result in a useful level of expression of the protein of interest

The pET vector system (Novagen, Inc.)

The pET vector looks like this:

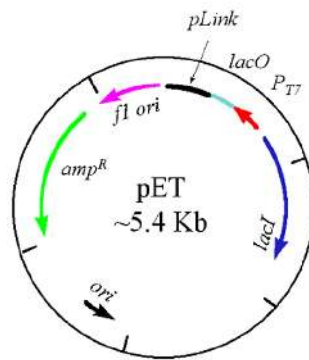


Figure 3.4.3: pET Vector

It has the following important elements:

- Ampicillin resistance marker
- ColE1 origin of replication
- f1 origin of replication (allows single stranded vector to be produced when co-infected with M13 helper phage)
- *lacI* gene (*lac* repressor protein)
- T7 transcription promoter (specific for phage T7 RNA polymerase)
- *lac* operator region 3' to the T7 promoter
- multiple cloning site (polylinker region) downstream of the T7 promoter

The pET vector is a little different from the pUC vector: pUC uses the *lac* promoter and pET uses a promoter from phage T7

- The phage T7 promoter is stronger than the *lac* promoter
- Phage T7 RNA polymerase will specifically recognize the T7 promoter region and will not efficiently transcribe from other promoters
- The T7 promoter will not be efficiently transcribed by *E. coli* RNA polymerase

Where will the phage T7 RNA polymerase come from?

The pET system involves not only an expression vector, but also a *genetically engineered host bacteria*. The host bacteria for the pET vector is typically *E. coli* strain **BL(DE3)**

- This strain has integrated into its chromosome the gene for T7 RNA polymerase
- The T7 RNA polymerase in the host genome is constructed such that it is under the control of a *lac* promoter and operator
- Thus, induction by the lactose analogue, IPTG, causes the host to produce T7 RNA polymerase
- The *E. coli* host genome also carries the *lacI* (repressor) gene

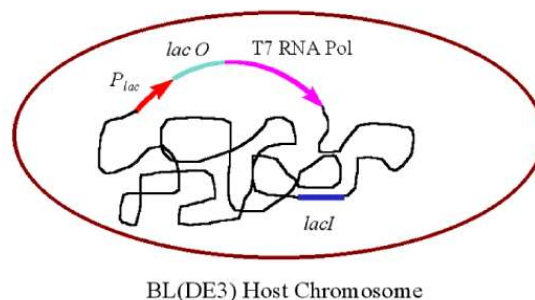


Figure 3.4.4: E.Coli BL(DE3) chromosome

Thus, induction by IPTG results in:

- Derepression of T7 RNA Pol gene on host chromosome with subsequent production of this polymerase
- Derepression of target gene under *lac O* regulation

- Transcription of target gene by T7 RNA Pol

Thus, the system couples a strong promoter with tight regulation (i.e. extremely low level of expression in the repressed state)

The pET vector itself is available with several different polylinker sequences. They contain the same restriction sites, but differ in the *reading frame* leading into the pLink region:

pET 21a

1	2	3	4	5	6	7	8	9	10	11	12	13	14
atg	gct	agc	atg	act	ggt	gga	cag	caa	atg	ggt	cgc	<u>gga</u>	<u>tcc</u>
met	ala	ser	met	thr	gly	gly	gln	gln	met	gly	arg	gly	arg

→

pET 21b

1	2	3	4	5	6	7	8	9	10	11	12	13	14
atg	gct	agc	atg	act	ggt	gga	cag	caa	atg	ggt	<u>cg</u>	<u>gat</u>	<u>cga</u>
met	ala	ser	met	thr	gly	gly	gln	gln	met	gly	arg	asp	pro

→

pET 21c

1	2	3	4	5	6	7	8	9	10	11	12	13	14
atg	gct	agc	atg	act	ggt	gga	cag	caa	atg	ggt	<u>cg</u>	<u>atc</u>	<u>cga</u>
met	ala	ser	met	thr	gly	gly	gln	gln	met	gly	arg	ile	arg

→
plink

Figure 3.4.5: Polylinker sequences

- The ggatcc site (BamH I restriction endonuclease site) is the first restriction site in the polylinker.
- The polylinker region is thus available in the three different possible reading frames
- A gene of interest can thus be cloned to be *in-frame* with the transcribed region downstream from the T7 promoter. Again, like the pUC vector the expressed protein will be a fusion protein (with 12-13 amino acids at the amino-terminal end of the polypeptide)

The pTrcHis vector system (Invitrogen, Inc.)

The pTrcHis vector looks like this:

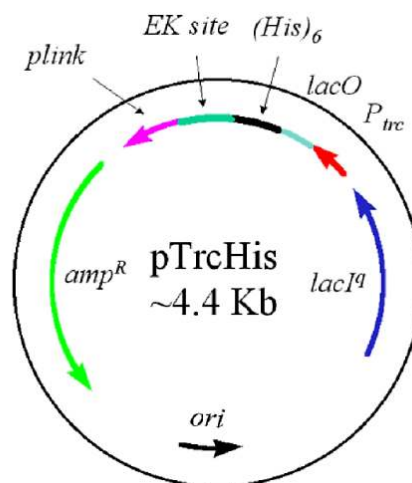


Figure 3.4.6: pTrcHis Vector

The important elements are:

- Ampicillin resistance marker
- ColE1 origin of replication
- *lacI^q* gene. This is a *lac* repressor mutation which is *upregulated* (produces more *lac* repressor than normal)
- *Trc* promoter. This is a hybrid of the *lac* and *trp* promoters which is a stronger promoter in comparison to the *lac* promoter
- The *lac* operator region downstream of *P_{trc}*. This allows regulation by IPTG.
- An initial stretch of 6 histidine residues in the amino terminal region of the translated protein, also known as a "*His tag*" region.
- An enterokinase (*EK*) cleavage recognition sequence (*asp asp asp asp lys* , with cleavage after the *lys* residue)
- A polylinker region downstream from the EK site

Purpose of the His tag

- Histidine residues can coordinate to form a metal (Ni^{2+}) binding site in a protein.
- The stretch of six His residues forms such a binding site.
- Proteins with a His tag have affinity for *metal chelating resins*, and this characteristic can be used to *selectively purify* such proteins.

Purpose of the EK cleavage site

- Although the His tag can allow rapid and selective purification of a cloned protein, the presence of these his residues may prevent normal function of the protein
- The His tag can be cleaved away from the protein by introducing a specific recognition sequence for an endopeptidase
- The sequence "asp asp asp asp lys" is recognized and cleaved by enterokinase. This sequence is not common and it is doubtful that the protein of interest contains another such sequence

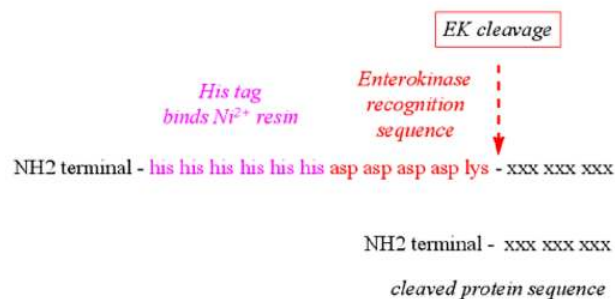


Figure 3.4.7: Enterokinase cleavage

- His tags/EK sites are typically introduced at the amino terminus of proteins, but can be introduced at the carboxyl terminus as well

Expression hosts

One of the important characteristics of expression hosts (bacteria) is that they allow the expressed protein to accumulate.

- Some expressed proteins may not be folded correctly. Although they may be refolded later to yield functional protein, misfolded proteins may be rapidly proteolyzed in the host bacterium
- Host proteases may selectively cleave precursor forms of expressed proteins to produce mature active forms, which may not be desirable.
- *OmpT* and *lon* genes code for proteases which can degrade expressed proteins.
- Two common mutations in *E. coli* to eliminate the action of host proteases includes the *OmpT⁻* and *lon⁻* strains.
- *E. coli* strain BL(DE3) is both *OmpT⁻* and *lon⁻*

This page titled [3.4: Prokaryotic Expression Vectors](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

3.5: Protein Sequencing, Peptide Mapping, Synthetic Genes

Historically, some important disease states were identified as being caused by the lack of an important **protein**, or the presence of a dysfunctional mutated form of a protein.

- For example, **diabetes**, types of **dwarfism** and **hemophilia** were found to be due to deficiencies in **insulin**, **growth hormone** and **clotting factor VIII**, respectively.

These diseases could be treated by injecting **supplemental** doses of purified, or partially purified, preparations of these proteins.

- These proteins were isolated from natural materials, e.g. pig (insulin), human cadaver pituitaries (human growth hormone) or blood fractions pooled from normal donors (factor VIII).
- In most cases, even if the protein was found in relatively abundant supply, **the cost of production was substantial**.

More often than not, interesting bioactive properties were associated with proteins which could be isolated only in **minute quantities** (e.g. the blood clot dissolving protein **tissue plasminogen activator**).

Also, **non-human proteins** typically elicited **an immune response** when injected into humans, thus the **human form** of a protein was the only useful form.

- If the protein were not readily available from blood, or urine, it would prove impractical to obtain **adequate starting material** for production.
- Unfortunately, if the material were derived from human sources, the possibility existed for the spread of **human disease** (e.g. hepatitis and the AIDS virus).

If the genetic information for these proteins could be isolated, and then transcribed and translated in an easily scaleable biological system, potentially large amounts of protein could be obtained - and hopefully, relatively cheaply.

With the development of "molecular biology", i.e.

- the structure of DNA,
- the elucidation of the genetic code,
- the identification of transcriptional promoters and ribosome binding sites,
- the isolation of restriction endonucleases,
- the identification of the origin of DNA replication
- the development of plasmids with selectable markers, and
- the culturing of *E. coli*,

the possibility existed in the mid 1970's to put it all together and produce relatively large amounts of **any human protein** for therapeutic use.

*How would you go about the process of producing large amounts of some important human protein? (i.e. **protein purification**)*

The starting point is typically **an assay for a functionality of interest**. For example, we may have a hemophiliac whose blood does not clot. However, we find that if we take a sample of his blood and add to it a small amount of blood from a "normal" individual, the hemophiliac's blood will now clot. This will be the basis for our **assay**.

Using this assay, we will fractionate normal blood using various means - chemical precipitation (with ethanol, or ammonium sulfate), and then various liquid chromatography steps, etc.

- Along the way we will follow where our clotting activity is going.
- Hopefully, at some point we will be unable to fractionate it further and will have a **pure protein**.

Once we have a pure protein we can begin to characterize it with regard to its **amino acid sequence**. From there we can ultimately get the gene for the protein and express it.

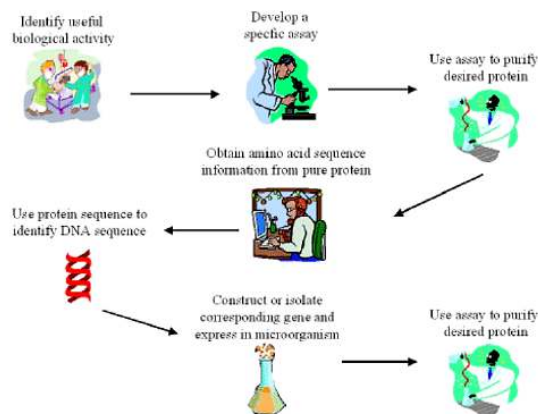


Figure 3.5.1: Protein production

N-terminal peptide sequence analysis

Polypeptides can be sequenced from their amino-terminus by automated procedures based upon the **Edman degradation** reaction:

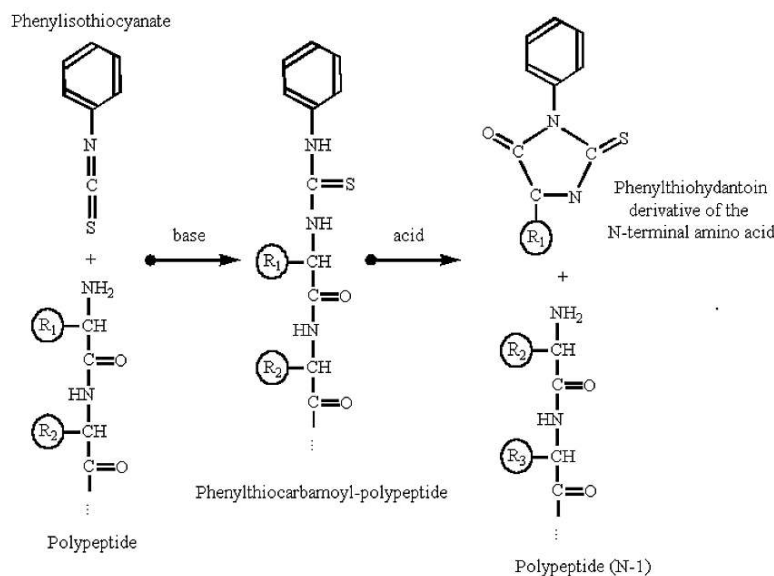


Figure 3.5.2: Edman degradation

- Note that with Edman chemistry only the N-terminal residue is attached and removed, the rest of the polypeptide remains intact after the reaction.
- The new amino terminal group (previously the second amino acid in the polypeptide chain) is now available for another round of reactions. Thus, the method can be **automated**.
- The amino acid side chain of the phenylthiohydantoin derivative can be identified using liquid chromatography. Modern amino acid sequencers can probably sequence on the order of two to three dozen cycles (amino acids) of a polypeptide.
- **Note that the reaction requires a free amino group on the N-terminal of the protein.** If the amino-terminal residue is methylated or formylated then the reaction will not proceed (and the polypeptide is said to have a "blocked" N-terminal).

C-terminal peptide sequence analysis

C-terminal peptide sequence analysis is not as well developed as amino terminal analysis.

- The method usually makes use of **non-specific carboxypeptidases**.

- Carboxypeptidases will **sequentially** hydrolyze polypeptides from the **carboxy-terminus** end. The released amino acid can be identified using liquid chromatographic methods, and the remaining polypeptide is available for further reactions.
- Various carboxypeptidases are available, usually they are not entirely non-specific (i.e. they have certain preferences):

Name	Source	Specificity
Carboxypeptidase A	Bovine Pancreas	Aromatics, aliphatics (hydrophobics)
Carboxypeptidase B	Pig Pancreas	Arginine, Lysine, Ornithine
Carboxypeptidase P	Penicillium	Generally non-specific
Carboxypeptidase Y	Yeast	Aromatic, aliphatics

Sometimes the choice of which carboxypeptidase to use is based upon the expected sequence information. In these types of experiments:

1. samples are taken at different time points during the digestion
2. free amino acids are separated from polypeptides
3. the released amino acids are identified via amino acid analysis (liquid chromatography).

C-terminal analysis is usually only accurately for identification of the last half-dozen residues or so in a polypeptide.

Peptide Mapping

One of the obvious problems with protein sequencing is that even if the N-terminal is not "blocked" only limited sequence information can be obtained from an intact polypeptide (i.e. only about two dozen from the N-terminal and half a dozen from the C-terminal).

How can sequence information for the entire polypeptide be obtained?

One method is that of **peptide mapping**. Peptide mapping makes use of **proteolytic cleavages** of the polypeptide to produce smaller polypeptides. These smaller polypeptides can then be isolated from one another and subject to sequence analysis.

How do we order the different sequences which we obtain?

One of the easiest ways is to repeat the experiment, but with a protease with a different specificity, and in this way obtain **overlapping sequence information**.

Name	Source	Specificity
Chymotrypsin	Bovine Pancreas	Cleavage after Tyr, Phe and Trp; some cleavage after Leu, Met and Ala
Bromelain	Pineapple	Cleavage after Lys, Ala and Tyr
Trypsin	Bovine Pancreas	Cleavage after Arg, less after Lys
V8 protease	<i>Staphylococcus aureus</i>	Cleavage after Glu, less after Asp

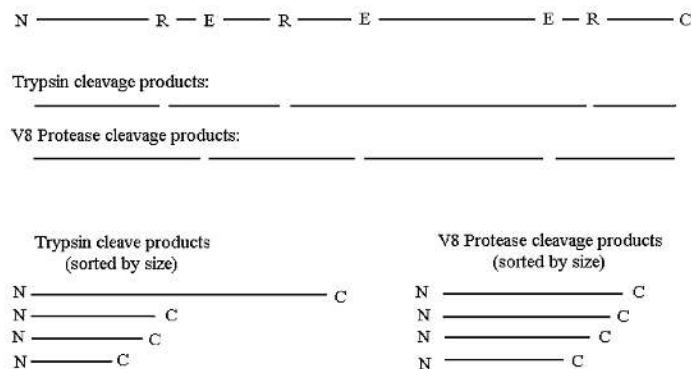


Figure 3.5.3: Overlapping cleavage products

Overlapping sequence information can allow you to align the peptides in the correct order and determine the sequence of the original large polypeptide (i.e. protein).

One problem which can arise deals with **Cysteine** residues and the nature of any **covalent disulfide bridges** in the protein.

- Any "peptide" mobilities (on either liquid chromatographic or PAGE analyses) which split into two smaller peptides after treatment with a reducing agent (such as b-ME) *indicate the presence of a cysteine mediated disulfide bond*.
- Upon sequencing these peptides should each contain a cysteine residue. If each peptide has only one cysteine then the disulfide bond assignment is unambiguous.

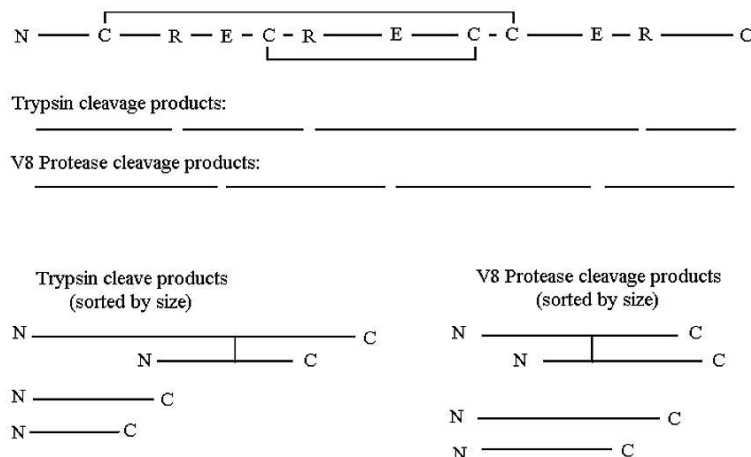


Figure 3.5.4: Cysteine residues in cleavage products

Corresponding genetic information

Once we have partial, or complete, peptide sequence information we can begin to identify and isolate the corresponding **genetic information**. *This is the main goal*. Once we have the corresponding genetic information it may be possible to produce relatively large amounts of the desired polypeptide.

Back translation

Since we know the genetic code, we can back translate any polypeptide sequence into a corresponding genetic sequence.

- Thus, from the amino acid sequence we could synthesize an **artificial gene** which would code for the protein of interest.

- Since many amino acids are coded for by more than one codon, there is potential **ambiguity** with regard to the original exact genetic sequence.

Amino Acid	Number of Codons
Met, Trp	1
Phe, Tyr, His, Gln, Asn, Lys, Asp, Glu, Cys	2
Ile	3
Val, Pro, Thr, Ala, Gly	4
Leu, Arg, Ser	6

However, making sure we back translate in such a way as to faithfully duplicate the original genetic sequence may not be critical - a correct protein sequence is the overall goal.

In fact, if we are attempting to express the protein in another organism (say expressing a mammalian gene in a bacterial system) we may actually prefer to choose a **codon bias** appropriate for the **expression host organism**.

Synthetic genes for small proteins are a reasonable way to proceed; this is one way in which **human insulin** has been expressed in bacterial systems.

- However, automated synthesis of DNA oligonucleotides is practical for polymer lengths of approximately 60-90 bases or less (about 20-30 amino acids).
- Furthermore, the method of construction of synthetic genes typically calls for overlapping complementary oligonucleotides (to be ligated into a single duplex DNA gene "cassette").

Thus, many oligonucleotides are required for even a single small synthetic gene.

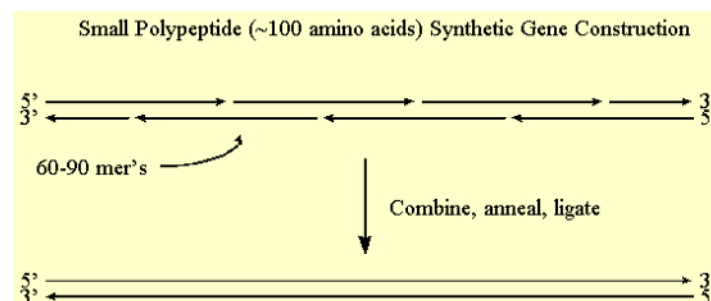


Figure 3.5.5: Synthetic gene construction

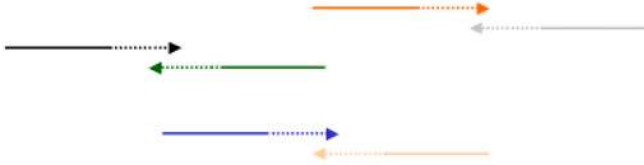
One way to improve upon the above method of synthetic gene construction is with a *direct PCR* approach. This method does not utilize ligase, or even oligonucleotides that butt together. Instead, with this method many (~100) different overlapping oligonucleotides are simultaneously used in a PCR reaction. Their sequence complementarity can be represented as follows:



The entire set of oligonucleotides may not line up to give the entire gene, but that is alright. We will do multiple rounds of PCR with the idea that *some* complementary oligo's will anneal and be extended and will lead, bit by bit, to construction of a contiguous synthetic gene:



On the next PCR cycle, some of these extended fragments will anneal with others:



These will be extended via the PCR and can go on to anneal with other larger PCR fragments. Eventually, the entire gene will be constructed. However, since the efficiency of construction of the full-length gene is probably not going to be very good, we need to conduct a subsequent PCR experiment to amplify the full-length gene (using outer primers). The principle features of this method are summarized as follows:

- Many (as many as 1-2 hundred) overlapping oligo's are combined in a single PCR reaction
- The oligos are designed to be as long as possible (~100mers) with limited overlap (~20 bases)
- The full-length gene is constructed in an initial (low yield) PCR experiment
- This full length gene is amplified with a subsequent typical PCR experiment using outer primers.
-

This page titled [3.5: Protein Sequencing, Peptide Mapping, Synthetic Genes](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

3.6: cDNA and Genomic Libraries

Isolation of corresponding genetic information

Instead of synthesizing a desired gene, can we use the amino acid information to directly isolate the corresponding genetic information?

- There are two general sources of genetic information:

1. **Genomic DNA**
2. **mRNA**

If we are considering **genomic DNA** from eukaryotes, then there are a couple of things to consider:

1. The coding region for a gene of interest may be *interrupted* by one or more **intron** regions, and thus the complete coding region could be quite long.
2. To a first approximation, it does not matter which tissue we use to isolate the genomic information, i.e. the genomic content is the same in all tissues.

If we are considering mRNA from eukaryotes, we may realize the following advantages:

1. Introns will be spliced out and *the mRNA will contain a contiguous coding region*.
2. Tissue specific expression of the protein of interest may allow us to isolate appropriate mRNA at enhanced levels, i.e. *in tissues where the protein is expressed the mRNA levels are considerably higher than the corresponding genomic levels (there are many more molecules of mRNA than copies of the gene)*.

Libraries

A "library" is a convenient storage mechanism of genetic information.

- They are typically either "genomic" or "cDNA" (i.e. mRNA in DNA form) genetic information.
- Deduced genetic sequences from corresponding polypeptide information can be used to identify specific genetic information within a library.

cDNA library construction

"Reverse transcription" is a mechanism whereby genetic information contained in mRNA is converted back into a double stranded DNA form.

The enzyme responsible for this is an RNA dependent DNA polymerase called **reverse transcriptase**.

- Reverse transcriptases have traditionally been isolated from viruses whose genome is actually in an RNA form and must be converted to duplex DNA.
- These viruses typically carry a functional reverse transcriptase along with their mRNA genetic component when they infect cells.
- One of the most common commercially available reverse transcriptases is **Moloney murine leukemia virus (MMLV)**.
- This **RNA dependent DNA polymerase** (as will all polymerases) add nucleotides to a nascent polynucleotide in the 5' to 3' direction using RNA as the **template**. It does not contain any 3'→5' exonuclease (proofreading) activity.

MMLV will use mRNA as a template, but requires a **primer** (it can extend a DNA primer but cannot synthesize one).

- One of the really neat things about eukaryotic mRNA's is the presence of the 3' **poly A tracks**.

Therefore we can use poly dT as a single primer for a variety of different eukaryotic mRNA's.

cDNA Library Construction: First Strand Synthesis

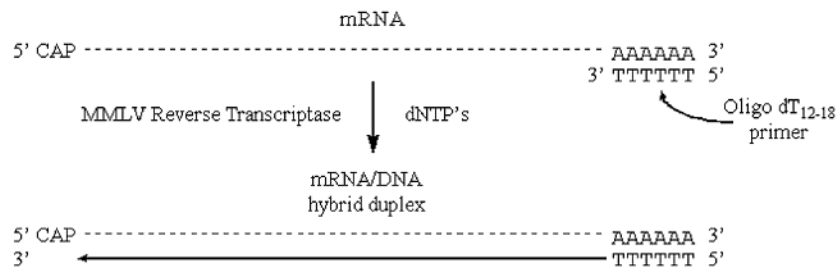


Figure 3.6.1: First strand synthesis

*Note that we have produced **complementary DNA (or cDNA)** to the original mRNA strand.*

If we can introduce "nicks" into the RNA half of this DNA/RNA duplex then the situation would be very similar to that observed in "lagging strand" synthesis of prokaryotic genomic DNA.

- Nicks in the RNA half of the molecule can be introduced via the action of the enzyme RNase H.
- This enzyme exhibits endonucleolytic cleavage of the RNA moiety of RNA/DNA hybrids, as well as 5'→3' and 3'→5' exoribonuclease activity.
- In other words, it will nick the RNA and then proceed to digest back in both directions:

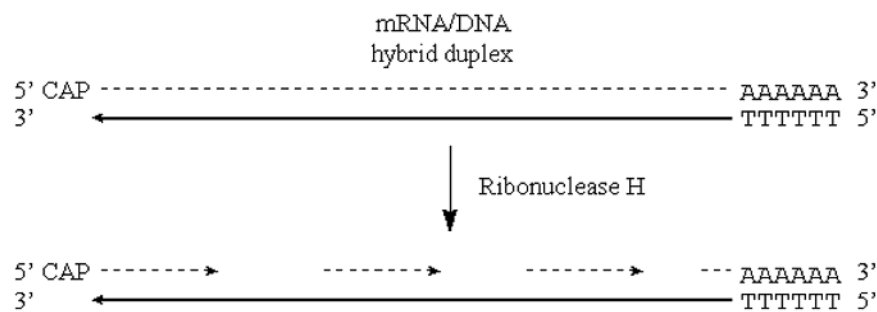


Figure 3.6.2: Nicks in mRNA

- These RNA fragments can now serve as primers for DNA synthesis by *E. coli* Pol I. This enzyme will also translate the "nicks" to effectively remove the RNA primers:

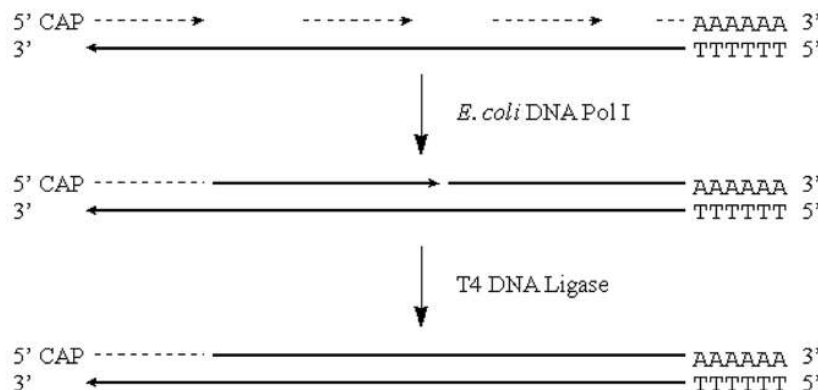


Figure 3.6.3: DNA synthesis

Note that we will potentially have either a residual 5' RNA cap region, or a gap at the 5' end of the original mRNA strand.

Insertion of cDNA into plasmid.

To complete our construction of a useful cDNA library we need a way to *maintain and propagate* our cDNA.

- We can accomplish this by inserting the cDNA into an appropriate **plasmid**.
- There are two classical ways of accomplishing this feat:

1. Homopolymeric tailing
2. Linker addition

Homopolymeric tailing

Terminal transferase is an unusual DNA polymerase found only in a type of eukaryotic cell called a *prelymphocyte*.

- In the presence of a divalent cation the enzyme catalyzes the addition of dNTP's to the 3'-hydroxyl termini of DNA.
- When the nucleotide to be added is a purine, Mg^{2+} is the cation used.
- When the nucleotide to be added is a pyrimidine, Co^{2+} is used.
- Depending on the reaction conditions, anywhere from three to several thousand bases will be added.

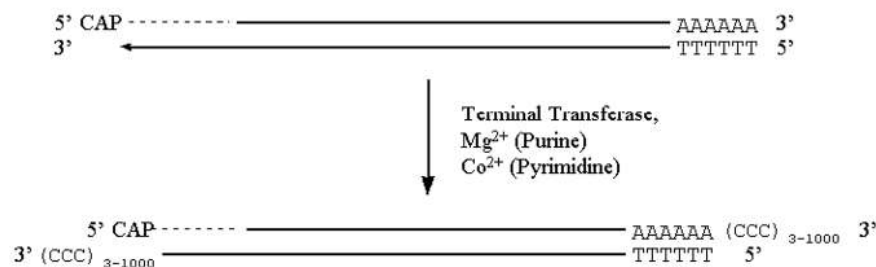


Figure 3.6.4: Terminal transferase activity

How can we insert this into a plasmid?

- If we cut our plasmid and also treat it with terminal transferase, except now we add the **complementary base** to the one we added to our cDNA, we can anneal and ligate the cDNA into the plasmid.

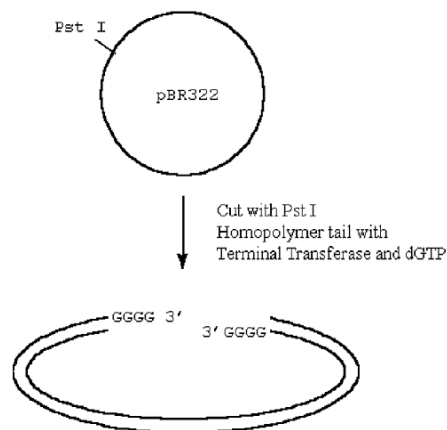


Figure 3.6.5: Ligating cDNA into the plasmid

- The utility of inserting the C-tailed cDNA insert into a G-tailed Pst I site in the vector is as follows:

1. The Pst I recognition sequence and cleavage site is

5' C T G C A G 3'

3' G A C G T C 5'
2. Cleavage of this site by Pst I, followed by G-tailing will produce

5' C T G C A (G)_n G 3'

3' G (G)_n A C G T C 5'

Thus, the Pst I recognition cleavage site is regenerated and the C-tailed cDNA insert can be excised again from the library vector by cleaving with Pst I.

Linkers

An alternate method to insert cDNA fragments into a library vector is through the addition of "linkers".

- Linkers are short oligonucleotides (~18 to 24 mers) which are typically *palindromic* and contain a single or repeated restriction *endonuclease recognition sequence*.
- The palindromic nature allows the linker oligonucleotide to *self-hybridize* to form a blunt ended *duplex*.
- If the ends of the cDNA fragments are blunt, then *the linker can be ligated to both ends to introduce useful terminal restriction sites*.

The steps in linker addition are as follows:

1. Treatment of cDNA with S1 nuclease (to remove possible 5' cap mRNA fragment remaining in cDNA duplex)
2. Convert potential "ragged" ends to blunt by treatment with Pol I (will fill in 5' overhangs and chew back 3' overhangs)
3. Methylate cDNA at potential internal Eco RI sites by treatment with Eco RI methylase (plus S-adenosyl methionine)
4. Ligate linkers to blunt, methylated cDNA using T4 DNA ligase
5. Cut linkers with Eco RI restriction endonuclease
6. Remove linker fragments from cDNA fragments by agarose gel electrophoresis
7. Ligate cDNA to vector DNA fragment (opened up by Eco RI restriction endonuclease)

*Period of time between first man-powered flight and landing on the moon (1902-1969):
67 years*

*Period of time between discovery of structure of DNA and determination of the sequence of the entire human genome (1953-2010?)
57 years (?)*

Editorial Note

This textbook was published in 1998. The Human Genome Project was completed in 2003.

Genomic DNA libraries

Size of some genomes and chromosomes:

Comparative Sequence Sizes	(Bases)
(yeast chromosome 3)	350 Thousand
Escherichia coli (bacterium) genome	4.6 Million
Largest yeast chromosome now mapped	5.8 Million
Entire yeast genome (completed 5/96)	15 Million
Smallest human chromosome (Y)	50 Million
Largest human chromosome (1)	250 Million
Entire human genome	3 Billion

- The human genome contains approximately 50,000 unique genes within 3-4 billion base pairs of DNA, scattered about in 23 pairs of chromosomes.

Fragmentation of genomic DNA for library construction

Restriction endonuclease digestion

- A six-cutter (e.g. Eco RI) will cut on average every 4.1 Kb. Complete digestion of human DNA with this type of enzyme will result in approximately 1×10^6 unique fragments.
- What is the probability of finding a clone within a given library?

The exact probability of having any given DNA sequence in the library can be calculated from the equation

$$N = \frac{\ln(1-P)}{\ln(1-f)}$$

P is the desired probability

f is the fractional proportion of the genome in a single recombinant

N is the necessary number of recombinants

For example, how large a library (i.e. how many clones) would you need in order to have a 99% probability of finding a desired sequence represented in a library created by digestion with a 6-cutter?

$$N = \frac{\ln(1 - 0.99)}{\ln(1 - (4096/3 \times 10^9))}$$

$$N = 3.37 \times 10^6 \text{ clones}$$

Thus, from this type of analysis we can see that we need a technology which will allow us to achieve the following:

1. *Stable insertion of relatively large DNA fragments into our cloning vector*
 2. *High efficiency of insertion and the ability to handle large numbers of clones*
- For example, when plating *E. coli* colonies on a 3" petri plate, the maximum practical density to allow isolation of individual colonies is about 100-200 colonies per plate.
 - If we were to try to plate our library of 3.37×10^6 in such a way would need about **22,500 plates**.
 - Not only that, but such large DNA fragments are not well tolerated in typical *E. coli* cloning vectors such as pBR322.

Bacteriophage lambda vectors are commonly used for construction of genomic libraries

Bacteriophage λ is an *E. coli* phage with a type of icosahedral phage particle which contains the viral genome:

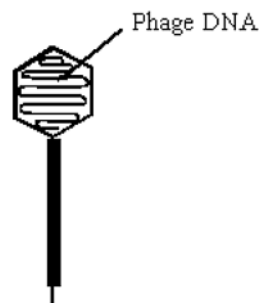


Figure 3.6.6: Bacteriophage λ

- During replication, the phage DNA is produced in a concatameric form, which is cleaved by appropriate endonucleases to allow packaging of a single genome within the phage capsid.
- It was found that internal regions of the phage genome, which were not essential to phage replication, could be removed and replaced with DNA of interest.
- This hybrid DNA could be efficiently packaged, and form an infective phage.

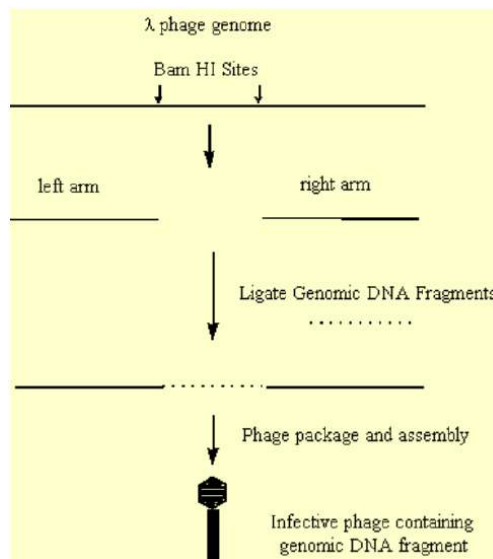


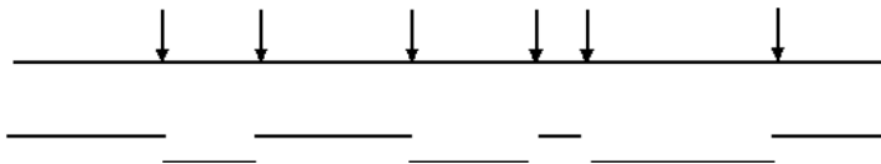
Figure 3.6.7: Creation of ineffective phage

The advantages of this type of system vs plasmids like pBR322 are:

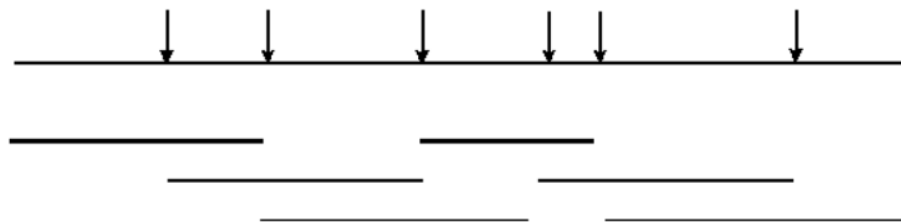
1. The phage genome is able to package efficiently with DNA inserts as large as 20 Kb.
2. Furthermore, the packaged phage are highly infectious and infect *E. coli* at a much higher efficiency than plasmid transformation methods.

Incomplete Digestion of Genomic DNA will allow identification of sequence overlaps

Complete digestion with an endonuclease will result in a library containing *no overlapping fragments*:



However, *incomplete digestion* will result in a library containing overlapping fragments:



- Thus, the sequence information obtained from *one clone* will allow the isolation of clones containing *neighboring (overlapping) sequence information*.
- This can allow large contiguous stretches of sequence information to be obtained ("**Chromosome Walking**").

Probing libraries

Once a library (cDNA or genomic) has been constructed we want to be able to identify clones which contain DNA of interest.

- For example, from protein sequence information we can *deduce possible stretches of the corresponding DNA sequence* (there will however be ambiguity due to the degeneracy of codons).
- If we can synthesize an oligonucleotide complementary to our DNA sequence of interest we can use it to specifically hybridize to the appropriate clone in our library (i.e. to **probe** our library).

In standard methodologies the oligonucleotide is *phosphorylated* at the 5' end with radiolabeled $g^{32}P$ -ATP and *T4 polynucleotide kinase*.

- The probe is then incubated with individual phage plaques which have been fixed onto nitrocellulose and their DNA denatured by treatment with base.
- If the plaque contains complementary DNA to the probe sequence, the probe will hybridize.
- If the nitrocellulose (containing many individual plaques) is exposed to x-ray film, *only those plaques with hybridized probe will show up (as a dark spot)*:

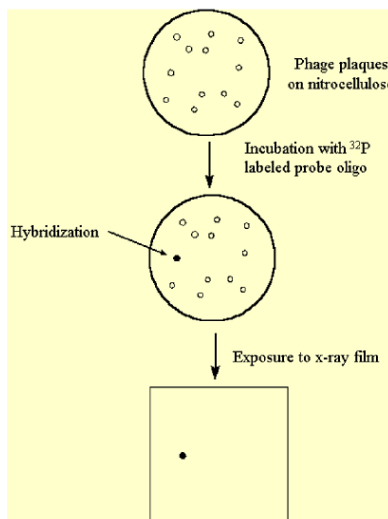


Figure 3.6.8: Radiolabeled plaque

Note that it's important to keep track of the orientation of the nitrocellulose in relationship to the x-ray film (usually radioactive ink is used to identify the nitrocellulose orientation).

False positives

If we are designing DNA probes from protein sequence information we will have *possible ambiguity* in our deduced DNA sequence used for the design of the probe.

- Usually 14-24mer oligonucleotides are used as probes, a 14-24mer probe means we need a stretch of 5-8 *amino acids* in the polypeptide.
- Given the choice, the best amino acid sequences to look for in a polypeptide are those with *low codon degeneracy* (see above).
- Thus, we would look for a short stretch of polypeptide sequence hopefully containing **Met or Trp**, and with the remaining amino acids comprising either **Phe, Tyr, His, Gln, Asn, Lys, Asp, Glu or Cys**.
- Regions including **Leu, Arg or Ser** are to be avoided (6 codons each).

During oligonucleotide synthesis multiple bases will be incorporated at ambiguous positions.

- Thus our probe will actually be a *mixture of oligonucleotides*.
- The higher the degeneracy, the greater the possibility of "false positives", i.e. clones which hybridize but are unrelated to the actual sequence we want.
- Positive clones are sequenced and the deduced amino acid sequence is compared to our polypeptide sequence information to identify correct clones.

Antibodies (Immunoglobulins)

If the particular vector, or phage, used to construct a cDNA library contains a **promoter** region upstream of the insertion site we may be able to screen for desired clones by looking for *expression of the protein of interest*.

- In this case, we need an assay which is both *sensitive* (we will not be producing a lot of protein) and *specific* (we want to minimize any false positives).
- One of the best assays, which is both sensitive and specific, makes use of **antibodies**.

Antigen, antibody, epitope

One of the defense mechanisms of vertebrates is the ability to distinguish between **self** and **non-self** molecules.

- Thus, if a foreign molecule (either from another species or sometimes from another individual within a species) invades a vertebrate organism, the immune system functions to learn to identify that molecule.
- In future invasions by the same molecule, the organism mounts a defense against it by producing specific **antibodies** which recognize and bind to the foreign **antigen**.
- When antibodies bind to antigen certain white blood cells (macrophages and monocytes) recognize the invading body as foreign and respond by destroying it.

Antibodies are 'Y' shaped molecules which contain two identical heavy chains, and two identical light chains.

- The stem of the 'Y' comprises the *Fc (constant) domain*, and the 'arms' of the 'Y' comprise the *Fab (variable) domains*.
- Antigens bind to the *complementarity-determining regions (CDR's)* located at the ends of the Fab domains.

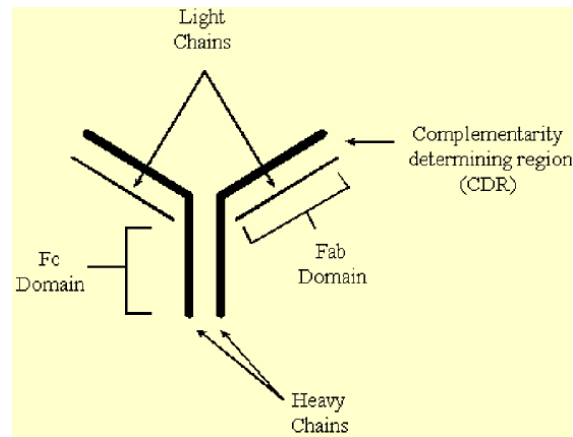


Figure 3.6.9: Antibody structure

Antibodies are synthesized by B lymphocytes. Each B lymphocyte is capable of producing a single type of antibody directed against a specific structural determinant, or **epitope**, on an antigen.

- Thus, an immune response to a protein antigen may result in a population of B lymphocytes each producing antibodies which recognize a different structural determinant of the foreign protein.
- An epitope may be a *contiguous region of 5 or 6 amino acids* in the foreign polypeptide, or the epitope may comprise a half dozen or so amino acids brought in juxtaposition in the native protein, yet *widely spaced in the polypeptide sequence*.
- Thus, some antibodies will recognize native *and* denatured forms of a foreign protein equally well, while other antibodies may only recognize one or the other.

If the protein of interest has been purified it can be used to *induce an immune response in a host animal*.

- Typical host animals include mouse, chicken, rabbit, goat, sheep, horse and occasionally, human.
- After an initial immunization, followed by one or more booster shots, the B lymphocytes of the host animal may produce antibodies directed against the antigen.
- The antibodies can be purified from blood samples withdrawn from the animal. Such preparations of antibodies are said to be **polyclonal**.
- This refers to the fact that the antibodies present are from a collection of different B lymphocytes and thus will recognize a *variety of different epitopes* on the antigen protein.
- The ability to isolate antibodies from blood samples means that the host animal does not need to be destroyed.
- Of course, the size of the animal determines how much antibodies one can obtain. For example, a rabbit can provide 5 mls of blood every two weeks, a mouse provides significantly less, while a horse can provide quite a bit more.

An antibody isolated from a *single B lymphocyte cell population* is termed **monoclonal**.

- It recognizes a *single epitope* on the antigenic protein.

- Antibody producing B lymphocytes can be isolated from the spleen or from lymph nodes. However, they have a *finite life span* in culture, i.e. they will undergo a certain number of cell divisions and then die.
- These cells can, however, be fused with immortal (cancerous myeloma) lymphocytes to produce a **hybridoma** cell.
- Such a cell is *immortal* like the myeloma, and produces a specific antibody from the B lymphocyte. The ability to grow indefinitely in culture allows the isolation of useful amounts of specific **monoclonal antibodies**.

Sometimes immunizing with the protein of interest is problematic: appropriate amounts of purified material cannot be produced, or the protein is itself toxic at the dosage level necessary to produce an immune response.

- If partial sequence information is known, then large amounts of polypeptides representing short fragments of the protein, can be synthesized and used to immunize the animal.
- Often these polypeptides are covalently attached to a carrier protein (typically serum albumin) to enhance the antigenic response.
- Antibodies produced against such peptides will recognize only epitopes within the polypeptide. Thus, even polyclonal antibodies would be quite limited in their epitope recognition.

As with radiolabeled oligonucleotides, antibodies can be used to identify library clones which contain a cDNA of interest. *This method would of course rely upon a host vector or phage which contains a promoter upstream from the site of insertion of the genomic DNA.*

- Antibodies can be used to screen viral plaques or plasmid clones which have been bound to nitrocellulose.
- Bound antibodies can be identified using radiolabeled protein A (which binds to immunoglobulins) or via a second antibody (which, like protein A, can recognize general immunoglobulins) which has a dye or dye releasing enzyme covalently attached.

This page titled [3.6: cDNA and Genomic Libraries](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

CHAPTER OVERVIEW

4. Biotechnology 2

4.1: Protein Purification

4.2: M13 Phage

4.3: M13 Phage Display Libraries

4.4: SELEX (Selective Evolution of Ligands by Exponential Enrichment)

4.5: Protein-Protein Recognition Probed Using a Yeast Transcriptional Activator System

4.6: Molecular Imprinting

This page titled 4. Biotechnology 2 is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

4.1: Protein Purification

Assays, Specific Activity, Initial Fractionation

A successful protein purification procedure can be nothing short of amazing. Whether you are starting off with a recombinant protein which is produced in *E. coli*, or trying to isolate a protein from some mammalian tissue, you are typically starting with gram quantities of a complex mixture of protein, nucleic acids, polysaccharide, etc. from which you may have to extract milligram (or microgram!) quantities of desired protein at high purity, and hopefully with high yield.

The first step in any purification is the development of a *specific assay* for the protein of interest. The specific assay can be based upon some *unique characteristic* of the protein of interest

- Enzymatic activity
- Immunological activity
- Physical characteristics (e.g. molecular mass, spectroscopic properties, etc.)
- Biological activity
- Ideally, an assay should be
 - *Specific* (you don't want a false positive)
 - *rapid* (you don't want to wait a week for the results)
 - *sensitive* (you don't want to consume all your sample in order to assay it)
 - *quantitative* (you need an accurate way to measure the quantity of your protein at each step in the purification)

Western Blotting

Antibodies can be used in a method called *Western blotting*, which is useful for determining levels of protein expression and for assaying proteins during purification. This method usually involves the following steps:

1. A protein sample is subjected to polyacrylamide gel electrophoresis.
2. After this the gel is placed over a sheet of nitrocellulose and the protein in the gel is electrophoretically transferred to the nitrocellulose.
3. The nitrocellulose is then soaked in gelatin to "block" its ability to non-specifically bind proteins.
4. The nitrocellulose is then incubated with the specific antibody for the protein of interest.
5. The nitrocellulose is then incubated with a second antibody which is specific for the first antibody. For example, if the first antibody was raised in rabbits, the second antibody might be termed "goat anti-rabbit immunoglobulin". What this means is that rabbit immunoglobulins were used to elicit an antibody response in goats. The goat antibodies (polyclonal) will include those which recognize the conserved region in the rabbit antibodies. Since the Fc region is conserved, it will bind to any and all rabbit antibodies, including those on the nitrocellulose paper.
6. The second antibody will typically have a covalently attached enzyme which, when provided with a chromogenic substrate, will cause a color reaction.
7. Thus the molecular weight and amount of the desired protein can be characterized from a complex mixture (e.g. crude cell extract) of other proteins.

In a variation of the above, the protein sample may be blotted directly on a nitrocellulose paper (called a *dot blot*) without first running a gel. This may be desirable if, for example, the antibody is monoclonal and recognizes an epitope which is dependent upon native structure (which would be destroyed upon running an SDS PAGE).

In addition to their varied uses, antibodies can also be used to purify proteins.

- If relatively large amounts of an antibody can be obtained, they can be covalently attached to a chromatography resin (e.g. sephadex beads).
- If a crude cell extract is run over such a column, only the protein of interest should bind, and everything else will flow through.
- The bound protein can then be eluted. This is typically achieved by moderately low pH conditions (using acetic acid). As long as the protein of interest is not irreversibly denatured by such conditions, the method will work quite well.
- One potential pitfall involves that of monoclonal antibodies being utilized to purify mutant proteins. The regions of the protein comprising the epitope cannot be modified without destroying the ability of the antibody to bind. Thus, the use of monoclonal antibodies in a purification scheme may preclude its use in purifying certain mutants.

Protein purification can be thought of as a series of *fractionation steps* designed so that:

- The protein of interest is found almost exclusively in one fraction (and with good yield)
- A significant amount of the contaminants can be found in a different fraction

During purification you will need to monitor several parameters, including:

1. *Total sample volume*
2. *Total sample protein* (can be estimated by A_{280} ; 1.4 ~ 1.0 mg/ml)
3. *Units of activity of desired protein* (based on specific assay)

This basic information will allow you to keep track of the following information during each step of purification:

1. *% yield* for each purification step
2. *Specific activity* of the desired protein (units/mg total protein)
3. *Purification enhancement* of each step (e.g. "3.5x purification")

In designing a purification scheme you typically have to balance *purification* with *yield*.

- For example, it may be relatively straightforward to obtain 90% pure material with good yield.
- However, it may be difficult to improve that purity an additional few percentile with good yield.
- *The planned application of the purified protein determines the target purity.*
- If the protein is to be used to determine amino acid sequence information, maybe 90% is acceptable. However, if the material is to be used in clinical trials, 99.99+% may be the target purity.

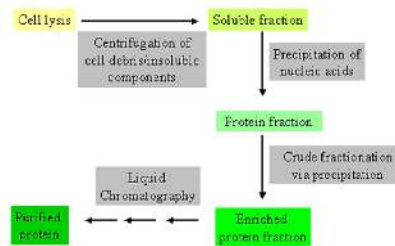


Figure 4.1.1: Purification steps

- It is extremely helpful to have some information not only on the general physical and chemical characteristics of the protein you are trying to purify, but also on the *contaminating* components.
- For example, many *E. coli* proteins are generally *low molecular weight* (<50,000 Da) and somewhat *acidic* in isoelectric point

Usually the initial steps in purification make use of general physical and/or chemical differences between soluble proteins and other cell components.

- For example, soluble proteins can be separated from general cellular debris, and intact cells, by *centrifugation*.
- Thus, cells are physically disrupted (via homogenization or a cell press) to allow release of cell contents. This is then followed by centrifugation to separate generally soluble components from those which are insoluble.
- It is at this point that data collection begins in order to monitor the purification.

Nucleic acids can sometimes be readily removed from the sample by the addition of large cationic compounds such as *polyethylene imine*, or *streptomycin sulfate*.

- The nucleic acids bind to these compounds via electrostatic interactions and the complex precipitates and can be removed via centrifugation.
- The same general result can be obtained by mixing in ion exchange resins which are *anion exchangers* (i.e. the resins contain cationic groups) and then filtering or centrifuging to remove. As with either method, it should be confirmed that the desired protein is *not* bound as well.

Crude fractionations of proteins can be achieved by adding various quantities of precipitants such as *ammonium sulfate*, or *polyethylene glycol* (PEG).

- For this type of purification step an initial experiment is performed to monitor the fraction of overall protein, as well as desired protein, remaining in solution (and pellet) as a function of precipitant concentration.

Ammonium Sulfate (% saturated)	0	10	20	30	40	50	60	70	80	90
Sample A_{280}	1000	900	600	200	100	75	50	40	25	20
Activity assay(units)	200	200	200	190	170	100	30	5	0	0

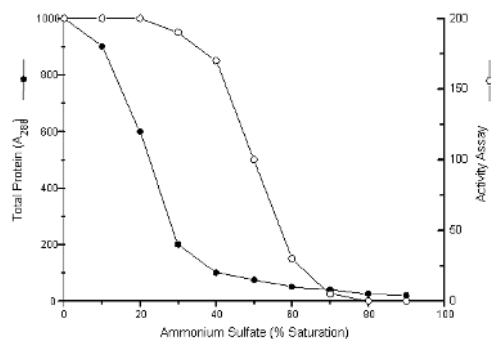


Figure 4.1.2: Protein activity as a function of precipitant concentration

- In this particular example we are in luck: at around 30% ammonium sulfate we can precipitate about 80% of the total protein concentration in our sample, yet our activity assay for our desired protein indicates that about 95% of our desired protein is still soluble.
- At 80% ammonium sulfate all of our desired protein has precipitated. Thus, from these results we would do the following:
 - Add ammonium sulfate to our sample to a concentration of 30% saturation
 - Centrifuge and discard the pellet
 - Add ammonium sulfate to 80% saturation
 - Centrifuge and keep the pellet. Resuspend the pellet in buffer to solubilize the protein.
- We would expect about a 5-fold purification with about 95% yield.

Column Chromatography - Ion exchange; Dialysis and Concentration

Column chromatography

After initial fractionation steps the typical procedure is to move to *column chromatography*.

- In column chromatography we have a glass tube (column) which is filled with a material ("resin") which has certain physical/chemical characteristics.
- These characteristics allow it to interact in various ways with different proteins.
- Some common types of chromatographic resins include:

1. Ion exchange
2. Hydrophobic
3. Gel filtration
4. Affinity

Ion exchange

Ion exchange resins contain **charged groups**.

- These may be *acidic* in nature (in which case the resin is a *cation* exchanger)
- or *basic* (in which case it is an *anion* exchanger).
- Cation and anion exchangers may be broken down further into *weak* and *strong* exchangers (reflecting binding affinity).

Type of exchanger	Functional group	Common name
Weak cation exchanger	carboxymethyl	CM cellulose/sephadex
Strong cation exchanger	sulfoethyl	SP sephadex
Weak anion exchanger	diethylaminoethyl	DE cellulose/sephadex
Strong anion exchanger	quaternary amine	QAE sephadex

- Usually, samples are loaded under **low ionic strength** conditions and bound material is eluted using either a step or gradient elution of buffer with **higher ionic strength**.
- Generally speaking, a protein will bind to a cation exchange resin if the buffer pH is **lower than the isoelectric point (pI)** of the protein, and will bind to an anion exchange resin if the pH is **higher than the pI**.

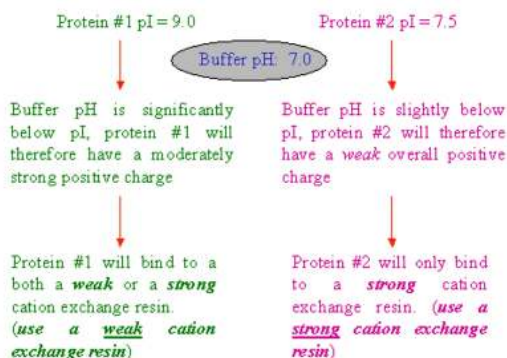


Figure 4.1.3: Protein binding to resins

- Knowledge of the pI of the protein is therefore helpful in designing a purification protocol using ion exchange resins (however, you can always simply try different resins to see which works best).

Generally speaking, ion exchange columns are short and fat in dimensions.

Elution of proteins from ion exchange resins

- Proteins bound to ion exchange resins are bound via non-covalent ionic (salt-bridge) interactions. We can compete for these ionic binding sites on the resin with other ionic groups, namely, **salts**
- There are two general types of methods when eluting with a salt solution: 1. **Gradient** elution and 2. **Step** elution
- A gradient elution refers to a smooth transition of salt concentration (from low to high) in the elution buffer. Weakly binding proteins elute first, and stronger binding proteins elute last (i.e. they require higher salt concentrations in the buffer to compete them off the column)
- A gradient salt concentration can be made using a **gradient maker**. In its simplest form, this consists of two containers (**must be the same shape**) connected by a siphon (or tube at the bottom). One container contains the low salt buffer, and the other contains high salt buffer. The buffer is withdrawn from the low salt container:

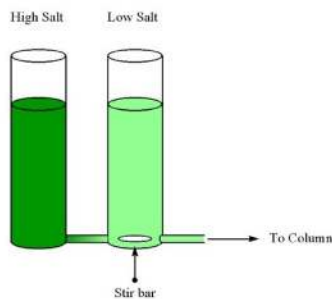


Figure 4.1.4: Gradient maker

- This will produce a linear gradient from low to high salt concentrations over the total volume of the gradient

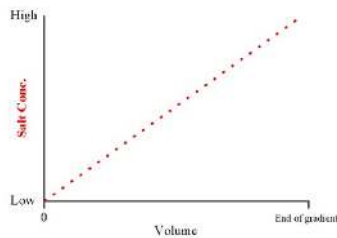


Figure 4.1.5: Salt concentration and volume

- If we know the concentration range of salt over which a protein of interest will elute we can simply elute with a buffer containing that concentration of salt. This is known as a **step elution**.
- Step elutions are generally faster to run, and elute the protein in a smaller overall volume than with gradient elutions. They generally work best when contaminants elute at a significantly different salt concentration than the protein of interest

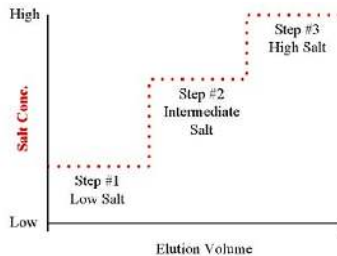


Figure 4.1.6: Step elution

Note that after ion exchange chromatography the protein of interest will be in a buffer with a potentially high salt concentration. This must be taken into account before proceeding with the next step in the purification scheme

Dialysis

- After an ammonium sulfate precipitation step, or an ion exchange chromatography step, the protein of interest may be in a high salt buffer. This may be undesirable for several reasons. How do we get rid of salt in our sample?
- One of the most common methods is that of **dialysis**
- The method of dialysis makes use of **semi-permeable membranes**. In the simplest example, this membrane is manufactured in the form of tubing (looking much like a sausage casing)
- The main feature of this membrane is that it is porous. However, the pore size is such that while small salt ions can freely pass through the membrane, larger protein molecules cannot (i.e. they are retained). Thus, dialysis membranes are characterized by the molecular mass of the smallest typical globular protein which it will retain.
- This is commonly referred to as the **cutoff** of the tubing (e.g. Spectrapore #6 dialysis tubing has a cutoff of 1,000 Daltons, meaning that a 1,000 Dalton protein will be retained by the tubing but that smaller molecular mass solutes will pass through the tubing)
- Dialysis proceeds by placing a high salt sample in dialysis tubing (i.e. the dialysis "bag") and putting it into the desired low salt buffer:

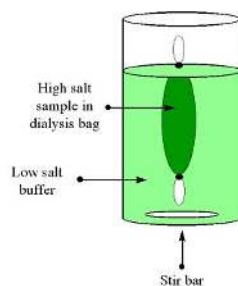


Figure 4.1.7: Dialysis

- Over time the concentration of low molecular mass solutes within the bag, and in the low salt buffer, will come achieve equilibrium. In practical terms (for the above case) salt molecules will diffuse out of the bag into the low salt buffer:

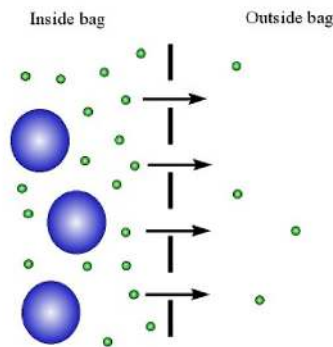


Figure 4.1.8: Salt diffusion

- At equilibrium the salt concentration of the sample can be calculated as follows:

$$\frac{(\text{sample volume}) \times (\text{sample salt concentration}) + (\text{buffer volume}) \times (\text{buffer salt concentration})}{\text{total volume}} = \text{final salt concentration} \quad (4.1.1)$$

Note

Often the buffer salt concentration is 0 M

- The buffer volume for the dialysis is a function of the required final concentration of salt in the sample

Example 4.1.1:

Dialysis example

We have a 10ml protein sample from an ion exchange column elution pool which contains 1.0M NaCl. For our next step in the purification we can have no more than 1mM NaCl in the sample.

$$\begin{aligned} \frac{(\text{sample vol}) \times (\text{sample salt conc}) + (\text{buffer vol}) \times (\text{buffer salt conc})}{\text{total vol}} &= \text{final salt conc} \\ \frac{(0.01\text{L}) \times (1.0\text{M}) + (\text{buffer vol}) \times (0)}{\text{total vol}} &= 0.001\text{M} \\ \frac{(0.01\text{L}) \times (1.0\text{M})}{0.001\text{M}} &= \text{total vol} \\ 10\text{ L} &= \text{total vol} \end{aligned}$$

Therefore, the required buffer volume would be (total vol - sample vol) = 9.990 L (or ~ 10 L)

- Thus, if we dialyzed 10mls of sample (with 1.0M NaCl conc) in 10 L of water after equilibrium the NaCl concentration in the sample would be 1.0 mM.
- Note that in the above example this would commonly be referred to as a "1:1,000" dialysis.
- Suppose that we don't want to make up 10 L of buffer? We can actually achieve the same results with two sequential "1:32" dialyses (i.e. the square root of the 1:1,000 dialysis - in other words, two sequential 1:32 dialyses is equivalent to a single 1:1,000 dialysis):

First dialysis versus 310 ml of buffer: sample NaCl conc will be $(10 \times 1.0) / (320) = 31\text{ mM}$

Second dialysis versus 310 ml of buffer: sample NaCl conc will be $(10 \times 0.031) / (320) = 0.97\text{ mM}$

Thus, instead of making 10 L of buffer, we could make only 620 ml and achieve the same results with two dialysis steps

- In this case, removing the salt would take twice as long, i.e. we need to perform two dialysis steps. How long does dialysis take?

A useful rule of thumb is that for most types of dialysis tubing the dialysis is 80% complete after four hours

- One consequence of dialysis to watch out for is that while salt ions are moving out of the bag, water molecules are moving into the bag. Thus the volume of sample may actually increase (the bag will swell) and, therefore, the protein concentration will decrease
- In the extreme case, the bag may actually swell to the point of rupture. Therefore, it is a good idea not to fill the bag completely, but leave a void to allow for potential swelling.

Concentration

- What if our protein sample is too dilute for our needs? How can we concentrate our samples?
- One common method is, again, to use a semi-permeable membrane for this purpose.
- A very simple method is to place our sample in a dialysis bag and coat it with a high molecular weight solute which can readily be dissolved by the buffer.
- For example, polyethylene glycols and polyvinyl pyrrolidones can have very large molecular masses (i.e. 20,000 Da). These compounds are also readily dissolved in water. If our sample in a dialysis bag is coated with dry forms of the above polymers, water will leave the dialysis bag (it can go through the pores) and hydrate the polymers. The result is a decrease in volume of buffer in the dialysis bag (the protein will be concentrated).
- In another variation, the semi-permeable membrane is manufactured into a flat disk and placed at the bottom of a container which holds our sample. In one method the container is pressurized and forces buffer out of the container (protein is retained and is concentrated). In another method, the vessel is centrifuged and the centripetal force achieves the same goal as pressure in the previous example.

For both dialysis and concentration, it is essential that the membrane does not interact with the protein (i.e. has no affinity for, and will not bind, the protein)

- With the pressure type concentrators, **dialysis and concentration can be achieved in tandem**. For example, the sample can be concentrated and then buffer added to the sample. The sample is then concentrated again. Every time buffer is added the salt concentration is reduced. After repeated cycles of this, the salt concentration is at the desired level and the sample is concentrated to the desired final volume.

Gel Filtration, Affinity and Hydrophobic resins; Preparation of Resin, Plumbing

Gel filtration

Gel filtration does not rely on any chemical interaction with the protein, rather it is based on a **physical** property of the protein - that being the *effective molecular radius* (which relates to mass for most typical globular proteins).

- Gel filtration resin can be thought of as beads which contain pores of a defined size range.
- Large proteins which cannot enter these pores pass around the *outside* of the beads.
- Smaller proteins which can enter the pores of the beads have a longer, tortuous path before they exit the bead.
- Thus, a sample of proteins passing through a gel filtration column will *separate based on molecular size*: the big ones will elute first and the smallest ones will elute last (and "middle" sized proteins will elute in the middle).

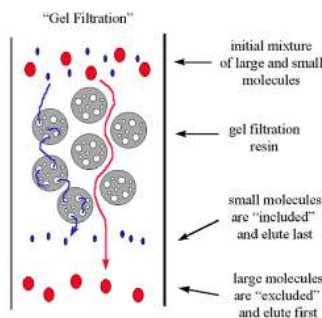


Figure 4.1.9: Gel filtration

- If your protein is unusually "small" or "large" in comparison to contaminating proteins then gel filtration may work quite well.

Where will a protein elute in a gel filtration experiment?

- There are two extremes in the separation profile of a gel filtration column.
- There is a critical molecular mass (large mass) which will be **completely excluded** from the gel filtration beads. All solutes in the sample which are equal to, or larger, than this critical size will behave identically: they will all eluted in the excluded volume of the column
- There is a critical molecular mass (small mass) which will be **completely included** within the pores of the gel filtration beads. All solutes in the sample which are equal to, or smaller, than this critical size will behave identically: they will all eluted in the included volume of the column
- Solutes between these two ranges of molecular mass will elute between the excluded and included volumes

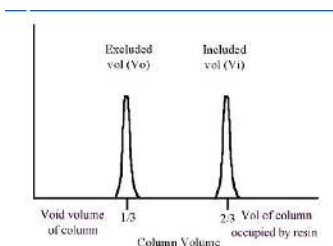


Figure 4.1.10: Excluded vs. included volume

As a general rule of thumb, the excluded volume (V_0) is approximately equal to one third of the column volume, the included volume is approximately equal to two thirds of the column volume

- In gel filtration the resolution is a function of column length (the longer the better)
- However, one drawback is related to the maximum sample volume which can be loaded. The larger the volume of sample loaded, the more the overlap between separated peaks. Generally speaking, the sample size one can load is limited to about 3-5% of the total column volume.
- Thus, gel filtration is best saved for *the end stages of a purification*, when the sample can be readily concentrated to a small volume.
- Gel filtration can also be used to remove salts from the sample, due to its ability to separate "small" from "large" components.
- Finally, gel filtration can be among the most "gentle" purification methods due to the lack of chemical interaction with the resin.

Affinity chromatography

Affinity chromatography is a general term which applies to a wide range of chromatographic media. It can be basically thought of as some inert resin to which has been attached some compound which has a *specific affinity* for your protein of interest.

- Thus, a specific antibody attached to an inert resin would be a type of affinity chromatography.
- Other examples might include: a protease inhibitor attached to some matrix, designed to bind a specific protease
- a cofactor bound to some matrix, designed to bind to a particular enzyme
- a metal ion bound to a matrix, designed to chelate a protein with a metal binding site, and so on.

In each case, the type of resins used and the method of attachment may vary, as will the method of elution. One generalization regarding method of elution is that the bound ligand can be competed off of the column's functional group by including in the elution buffer a high concentration of the free functional group. For example, if the functional group of the column is a cofactor, then the bound protein can be competed off the column by passing a buffer containing a high concentration of cofactor (or cofactor analog) through the column.

Other methods of elution include changing the buffer conditions such that the protein is no longer in the native state (since it is the native state which confers the structure required for the specific binding interaction). This can be achieved by changing pH or by adding denaturing agents such as urea or guanidine.

With affinity chromatography, typically the purification achieved in a single step can be dramatic - on the order of several thousand fold. Single step purifications with specific affinity columns are not unheard - in fact it is an ideal goal of purification - *a matrix which recognizes only the protein of interest and none other*.

Hydrophobic resins

Hydrophobic resins contain a non-polar functional group, such as an alkane or aromatic group.

- Many proteins are able to sequester such groups on their surface and this exclusion from solvent provides the basis of the binding energy (i.e. the "hydrophobic effect").
- This interaction is enhanced by increasing ionic strength, such that proteins may *bind under high salt conditions and elute under low salt conditions*.
- As such these columns may be used to not only provide purification, but to *desalt* samples (for example after an initial ammonium sulfate precipitation).
- It is usually not possible to predict in advance which particular resin will bind a given protein, this is usually determined empirically. However, the longer the alkane, or the larger the aromatic compound, the stronger the binding typically will be.

Due to the nature of hydrophobic interactions and ionic strength, hydrophobic chromatography and ion exchange chromatography can be conveniently used sequentially. For example, after ion exchange the protein is in high salt conditions, thus it can be loaded directly onto a hydrophobic column. Conversely, a hydrophobic column is eluted in low salt, which is a requirement for binding to an ion exchange resin.

A distinction should be noted between hydrophobic interaction chromatography and reverse phase chromatography

- Hydrophobic interaction chromatography is performed in aqueous solvent conditions and changes in ionic strength are used to elute the column. The protein typically binds in the native state via hydrophobic groups located on the surface of the protein. The native state is retained during the elution conditions
- Reverse phase chromatography utilizes a hydrophobic solvent (typically acetonitrile) and the binding of a ligand is a function of the phase partition between the hydrophobic nature of the solvent and column functional group. Proteins are typically denatured in such solvents and bind due to the hydrophobic nature of the entire polypeptide sequence. Since the majority of hydrophobic groups are located in the core of globular proteins, the binding is related to the denaturation of the protein and the accessibility of these groups to the column functional groups. Proteins can be purified using reverse phase chromatography, but usually must be refolded in some way to regain functionality (i.e. the native state)

Preparation of resins

The steps in preparing a chromatographic resin typically involve:

1. Hydration of resin
 2. Decanting fines
 3. Equilibrating the resin and preparing a slurry
 4. Degassing the slurry
- Resins come either dry or preswollen. If they are dry they need to be hydrated. This is usually accomplished by mixing the dry resin with buffer and letting it hydrate slowly overnight (or faster at higher temperatures)
 - After the resin has hydrated and settled, very fine particles will settle at the top. These "fines" slow the flow rate of the packed resin. The settled resin is therefore carefully decanted to discard these fines.
 - The resin is then equilibrated in the buffer to be used for the analysis. Equilibration usually involves pH'ing the resin, or buffer exchanges. Never use a stir bar when pH'ing the resin (it can mechanically shear the resin and produce fines), rather stir the resin slurry with a stir rod.
 - After the equilibrated resin has settled, an equal volume of buffer is added to produce a 50% *slurry* of resin. This is usually "thin" enough to allow air bubbles to escape when packing the column.
 - Finally, the slurry is degassed prior to packing the column. This will help minimize the formation of air bubbles.

Packing the column

Low pressure columns are typically packed using gravity.

- Add a small amount of buffer to the bottom of the column.
- Place a *packing reservoir* on the top of the column. Since we will be using a 50% slurry we will have a volume which is 2x the column volume and its best to pour the resin in all at one time. Thus, the packing reservoir should have a volume equal to, or greater, than the column volume.
- Carefully pour the resin slurry into the packing reservoir/column, avoiding the introduction of air bubbles as much as possible
- Let the column sit for about 5 minutes to allow large air bubbles to escape
- Open the column valve at the bottom and allow the column to pack under gravity
- Note the top of the resin bed. It will move down as the column packs. When the column is packed the top of the resin bed will no longer move down.

Plumbing

Chromatography systems may be run using only gravity and a beaker to collect the appropriate fraction. Most common systems, however, will include the following:

- A pump. Usually a peristaltic pump with variable flow rate and a communications port for a controller. *The pump is usually set up to push buffer through the column, rather than sucking buffer out of the column (which can cause a low pressure condition with production of air bubbles)*
- A detector. This is typically a UV (A_{280}) detector. Most detectors are of the two-cell type - meaning that you can have plain buffer as a blank in the detector while analyzing your column fractions. The detector sends the absorbance information to a chart recorder to be displayed (printed)
- A fraction collector. This allows you to collect fractions either by number of drops (~30 per ml) or by time. In conjunction with a controllable pump, time collection translates to volume. The fraction collector will typically have an communications port to output a signal when it changes fractions and to receive commands from the detector/controller on some sophisticated systems.
- A chart recorder. This will print a continuous trace of the detector output and the fraction collector event marker (signalling when a fraction changes). Fractions can also be read individually on a UV spectrophotometer if a chart recorder is unavailable.

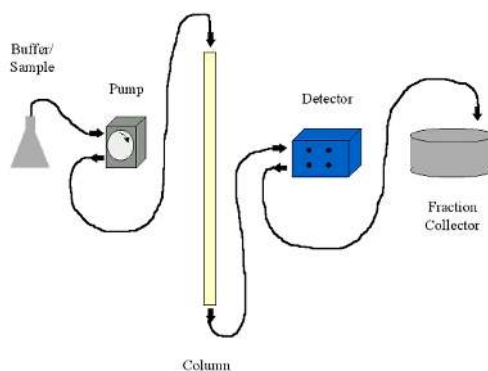


Figure 4.1.11: Chromatography setup

- If a gravity system is used, a safety loop should be installed to prevent the column from drying up if the buffer is used up when the column is unattended

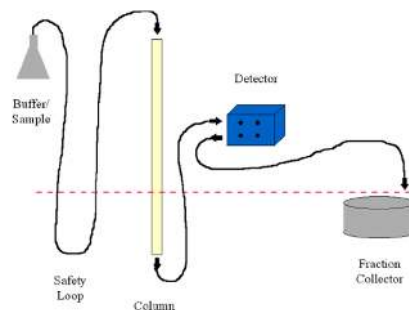


Figure 4.1.12: Safety loop

Note that the bottom of the safety loop is lower than the outlet to the fraction collector.

Running the Experiment, Resolving Peaks

The following represents an example of a low pressure liquid chromatography (ion exchange resin) experiment.

Sample:

- Volume = 90 mls
- $A_{280} = 1.8$
- Total $A_{280} = 162$

Column:

- DE-52 (diethylaminoethyl cellulose; anion exchange)
- Size = 1.0 x 12.7 cm
- Volume = $\pi \times (\text{radius})^2 \times \text{height} = 40 \text{ mL}$

Fraction Collector:

- 10 mls / fraction (~300 drops/fraction)

The chromatogram for this experiment looked like this:

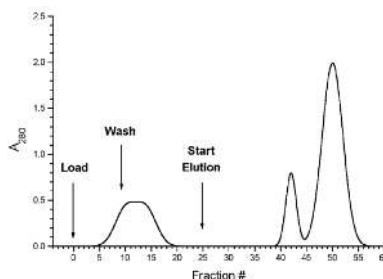


Figure 4.1.13: Chromatogram

The following events took place during this chromatography run:

1. Note the tick marks on the chromatogram.
 - The "event" marker from the fraction collector notifies the chart recorder when a tube change takes place.
 - The experiment begins with the tick next to the '0' on the x-axis ("tick 0"); this indicates the start of fraction (tube) number 1.

- The next tick mark ("tick 1") indicates the end of fraction number 1, and the start of fraction number 2.
 - Thus, *fractions span the gap between the tick marks.*
2. The sample loading is begun at tick 0.
3. Sometime during fraction 5 we begin to notice the absorbance of the column effluent increasing
- It has taken about (5 fractions x 10mls per fraction) or 50 mls from the start of loading until the detector notes any absorbance.
 - This compares well with the fact that the column volume is about 40 mls and there is some volume associated with the tubing going in and out of the column.
 - Thus, this 'delay' from sample load to sample detection is the *dead volume* of the system
4. Obviously, some material is not binding to the resin during the loading step. This is the *flow-through*. Is this some component of the sample which does not have affinity for the resin, or, does it represent that we have exceeded the capacity of the resin?
- If we have exceeded the capacity of the resin, then the flow-through will have an A_{280} similar to the sample being loaded
 - Also, prior to exceeding the capacity, the flow-through will have some characteristic A_{280} which will then transition to another A_{280} (that of the loaded sample), resulting in a *double-plateau* chromatogram.
 - In the above experiment the flow-through plateaus around $A_{280}=0.5$ or about 25% of the absorbance of the load. This would seem to indicate that a component, or component(s), representing one quarter of our sample, does not have affinity for the resin in the column
5. Around fraction 9 we begin to wash the column
- This makes sense because 9 fractions x 10mls per fraction = 90 mls have loaded and this is equivalent to our original sample volume (i.e. all the sample has loaded)
 - The column is typically washed using the same buffer conditions in the protein sample
6. Around fraction 14 we note the A_{280} begins to decrease
- This makes sense given that we determined the dead volume of the system to be approximately 50 mls or 5 fractions. Thus, a wash which was begun at fraction 9 is observed to decrease the absorbance around fraction 14
 - We continue washing the column until the A_{280} approaches 0 (*baseline*). In other words, all of the non-binding material in the sample has been washed away
7. After the A_{280} comes back down to baseline we begin our *elution* protocol. In this particular experiment we will use a linear gradient of increasing salt (NaCl) concentration (in wash buffer) to compete off the material bound to the ion exchange resin.
8. Our elution has produced two peaks: a small peak centered around fraction 42 and a larger peak centered around fraction 50
- We will have to assay each peak (and the flow through) to find out where our protein of interest has gone
 - The two *elution peaks* are fairly well resolved. We could combine fractions 40-44 and call that "peak 1", and combine fractions 46-55 and call that "peak 2".
9. *Is there any material left on the column?* The integrated areas (i.e. summing the A_{280} 's of each fraction in a pool) of the flow-through, peak 1 and peak 2 are as follows
- Flow through: 4
 - Peak 1: 2
 - Peak 2: 10
 - This gives a total integrated area of 16. Each fraction is 10 mls, so this gives a total $A_{280} = 16 \times 10 = 160$ which is quite close to the total A_{280} of our loaded sample.
 - In other words, it looks like our chromatogram is accounting for all the components in our original sample.
10. If our protein of interest was actually peak 1 (and if our yield was 100%), then this column has provided an eight fold purification ($2 \times 10 / 162$).

Resolving peaks

- Contaminating peaks will not necessarily be completely separated from the peak which contains our protein of interest
- In the following picture there are two components being resolved, and they are present in equimolar amounts (thus, the starting purity is 50%). The yield and purity are listed for the situation where we were to pool each peak by splitting at the midpoint between them (in this particular example the yield and purity are identical in each case)

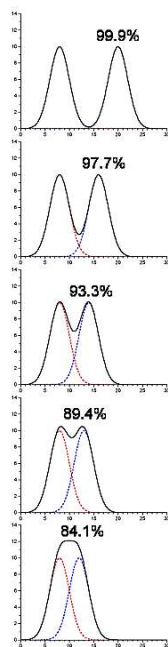


Figure 4.1.14: Contaminating peaks

- This gives you some idea of the amount of cross-contamination in each peak as a function of their separation from one another.
- Software to fit gaussians to a chromatogram can provide this type of information

Pooling for purity verses yield

Do we try to pool to maximize yield or to maximize purity?

Usually, you will probably be pooling fractions in such a way as to maximize the recovery of your protein of interest. However, you always have the option of pooling to increase purity, and if you have lots of protein to work with this may allow you to achieve the desired purity with fewer steps. Here's an example of how it's done:

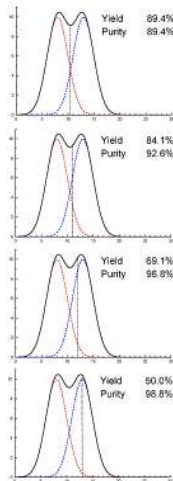


Figure 4.1.15: Yield vs. purity

- These are all the same chromatogram, however, we can pool them differently to get better purity (at the expense of yield)
- The blue peak is the peak of interest and it is not resolved from a contaminating peak (in red).
- The vertical line represents the left-most fraction we use to pool the peak (we pool all fractions to the right of the vertical line to get our protein of interest)
- In the last panel we see that we can achieve about 98.8% purity if we are willing to part with half our protein!

Monitoring the purification

How do you know when you are finished purifying a protein?

There are several criteria. One criteria is that we cannot improve upon the *specific activity* of our sample. This value refers to the *functional activity of our sample in relationship to the total protein concentration of the sample*.

- In the initial stages of purification this value will be low (not much activity in relationship to the total amount of protein).
- This value will increase after each purification step as we *remove* other proteins from the sample.
- At some point the specific activity will *plateau*, and by definition, *if it is pure we cannot increase the specific activity*.

- There may be a published value for the specific activity which we can compare ours to.

Also, each step of the purification should be monitored by gel electrophoresis.

- In the initial stages of purification we will probably see a variety of bands, of various molecular weights, on our gel.
- After the different purification steps, we should see the disappearance of certain bands concomitant with the increasing concentration of a certain band (or bands) representing our protein.
- If we have successfully purified our protein (and if it is a single polypeptide) we should arrive at a constant specific activity and *a single band on a gel*.
- Analytical methods like HPLC or densitometer scanning of a stained gel can give us a quantitative idea of the purity of our final sample.

The following chart represents the typical data one would monitor during a purification:

Step	Total protein (mg)	Total activity (units)	Specific activity (units/mg)	Purification	% Yield
Crude cell lysate	5500	6600	1.2		
30-70% Ammonium sulfate cut	1020	5910	5.8	4.8	89.5
DEAE Sephadex pool	187	5070	27.1	4.7	85.8
CM Sephadex pool	102	4420	43.3	1.6	87.2
Phenyl Sepharose pool	56	3930	70.2	1.6	88.9
Gel Filtration pool	32	2970	92.8	1.3	75.6
Affinity resin type #1 pool	5.8	2520	434.5	4.7	84.8
Affinity resin type #2 pool	5.3	2390	450.9	1.0	94.8
Total purification	376				
Total yield (%)	36				

This page titled [4.1: Protein Purification](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

4.2: M13 Phage

The bacteriophage known as "M13" forms the basis of cloning systems designed to easily introduce mutations into genes inserted into the phage genome. It also has been used in various "phage display" methodologies and "combinatorial" DNA and peptide libraries.

M13 infection and replication

M13 is a filamentous *bacteriophage* which infects *E. coli* host. The M13 genome has the following characteristics:

- Circular *single-stranded* DNA
- 6400 base pairs long
- The genome codes for a total of 10 genes (named using Roman numerals I through X)

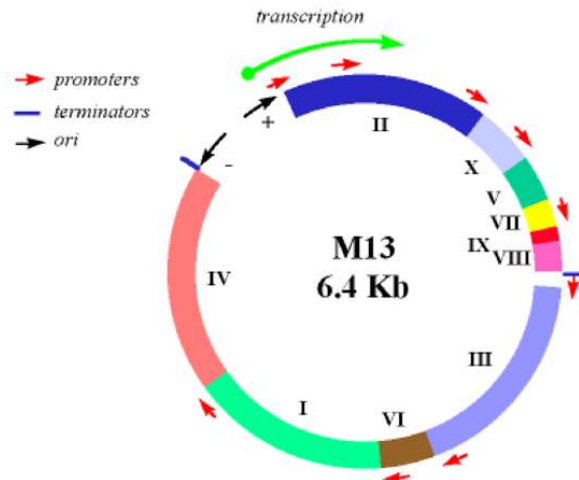


Figure 4.2.1: M13 genome

- *Gene VIII* codes for the major structural protein of the bacteriophage particles
- *Gene III* codes for the minor coat protein

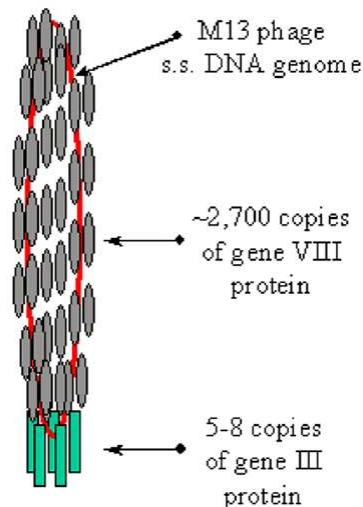


Figure 4.2.2: Gene III and gene VIII

- The gene VIII protein forms a tubular array of approx. 2,700 identical subunits surrounding the viral genome

- Approximately five to eight copies of the gene III protein are located at the ends of the filamentous phage (i.e. genome plus gene VIII assembly)
- Allows binding to bacterial "sex" pilus
 - Pilus is a bacterial surface structure of *E. coli* which harbor the "F factor" extrachromosomal element

Infection

- Single strand genome (designated '+' strand) attached to pilus enters host cell
 - Major coat protein (gene VIII) stripped off
 - Minor coat protein (gene III) remains attached
- Host components convert single strand (+) genome to double stranded circular DNA (called the replicative or "RF" form)
- Transcription begins
 - Series of promoters
 - Provides a gradient of transcription such that gene nearest the two transcription terminators are transcribed the most
 - Two terminators
 - One at the end of gene VIII
 - One at the end of gene IV
 - Transcription of all 10 genes proceeds in same direction

Amplification of viral genome

- Gene II protein introduces 'nick' in (+) strand
- Pol I extends the (+) strand using *strand displacement* (and the '-' strand as template)
- After one trip around the genome the gene II protein nicks again to release a completed (linear) '+' genome
 - Linear (+) genome is circularized
- During first 15-20 minutes of DNA replication the progeny (+) strands are converted to double stranded (RF) form
 - These serve as additional templates for further transcription
- Gene V protein builds up
 - This is a *single stranded DNA binding protein*
 - Prevents conversion of single (+) strand to the RF form
- Now get a buildup of circular single stranded (+) DNA (M13 genome)

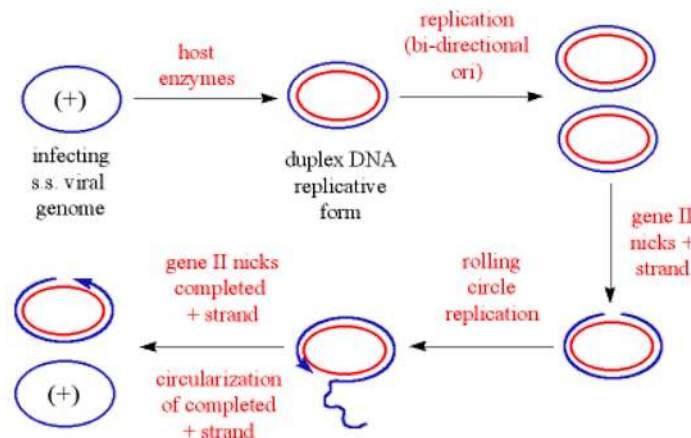


Figure 4.2.3: Amplification of genome

Phage packaging

- Major coat protein (Gene VIII) present in *E. coli* membrane
- M13 (+) genome, covered in ss binding protein - Gene V protein, move to cell membrane

- Gene V protein stripped off and the major coat protein (Gene VIII) covers phage DNA as it is extruded out
 - Packaging process is therefore *not linked to any size constraint* of the M13 genome
 - Length of the filamentous phage is determined by size of the DNA in the genome
 - Inserts of up 42 Kb have been introduced into M13 genome and packaged (7x genome size)
- ~8 copies of the Gene III protein are attached at the end of the extruded genome

Development of M13 into a cloning vector

M13 was developed into a useful cloning vector by inserting the following elements into the genome:

- a gene for the *lac* repressor (*lac I*) protein to allow regulation of the *lac* promoter
- the operator-proximal region of the *lac* Z gene (to allow for a-complementation in a host with operator-proximal deletion of the *lac* Z gene).
- a *lac* promoter upstream of the *lac* Z gene
- a polylinker (multiple cloning site) region inserted several codons into the *lac* Z gene

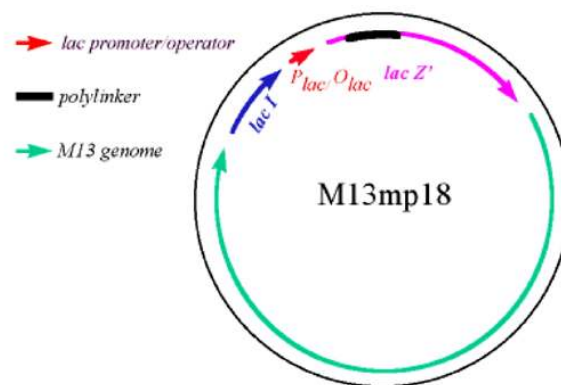


Figure 4.2.4: Insertions into genome

- The vectors were named according to the specific polylinker region they contained
- The vectors were typically constructed in pairs, with the polylinker regions in opposite orientations

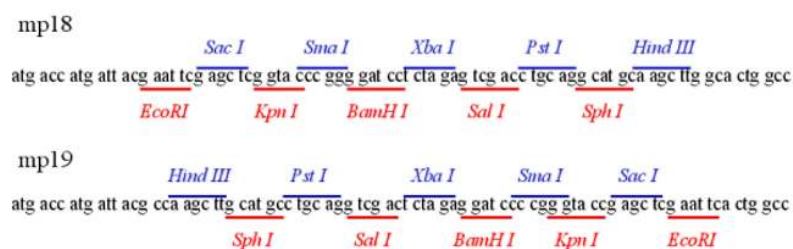


Figure 4.2.5: Polylinker regions

Cloning into M13mp vectors

- The RF (double stranded) form of the M13 phage can be isolated and treated just like any other plasmid
- The polylinker region can be "opened" using restriction endonucleases appropriate for accepting the fragment of interest
- The fragment is ligated into the plink region
- The availability of inverse oriented plink's (e.g. mp18, mp19) means that inserted fragments with non-complementary ends can be inserted in either orientation

Single stranded forms of the phage

The ability to isolate a single stranded form of the phage has advantages in both sequencing and mutagenesis.

- Single stranded DNA template can be read further than double stranded template

An efficient mutagenesis method (the "Kunkel" method) was developed using the single stranded form of the phage.

- The M13mp vector with insert is first grown in a mutant *E. coli* host (e.g. *CJ236*) which would occasionally incorporate uracil into the DNA instead of thymidine.
 - *E. coli* normally synthesizes an enzyme (uracil-N-glycosidase) that removes uracil residues in DNA. However, in *ung*⁻ strains, the uracil is not removed
 - The level of uracil mis-incorporation into DNA is enhanced in strains which have a deficiency in dUTPase. This enzyme converts dUTP to dUDP, and therefore, in *dut*⁻ strains the levels of dUTP are elevated and enhance misincorporation of dUTP into host DNA
 - The *E. coli* strain *CJ236* has a genotype which includes *dut*⁻/*ung*⁻ features
- A mutagenic primer would be annealed to this single strand template (e.g. to produce a point mutation)
- The primer is extended using the four dNTP's, and is ligated to produce duplex DNA
- The duplex DNA is inserted into a different host *E. coli* (e.g. *JM101*) which recognizes and excises (degrades) uracil containing DNA (i.e. a strain which is *ung*⁺)
- The parent (wild type) strand is preferentially degraded and the mutagenic strand is replicated.
- The phage progeny typically have a high incidence (80-90%) of the desired mutation.

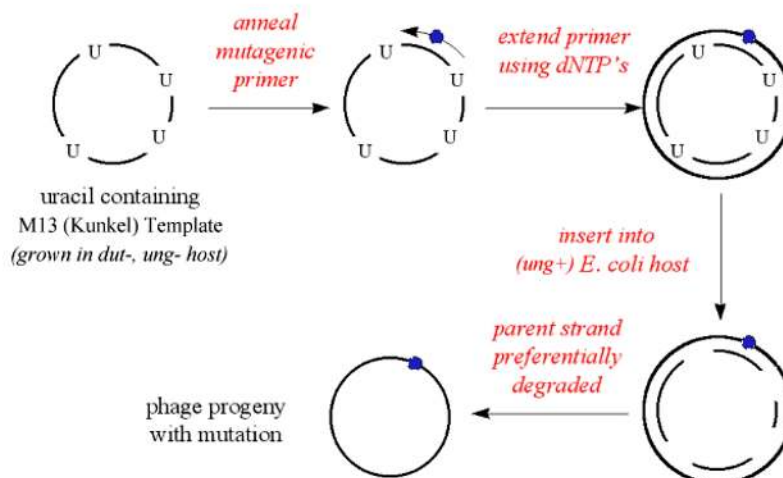


Figure 4.2.6: Kunkel method

This page titled [4.2: M13 Phage](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

4.3: M13 Phage Display Libraries

Microbial systems developed for experimental molecular evolution

A protein or peptide is introduced at the amino-terminus of the major coat protein (Gene VIII) via PCR-based mutagenesis

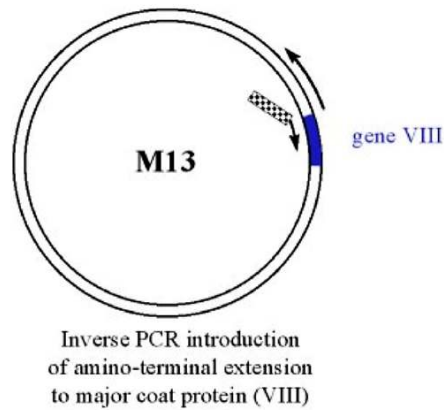


Figure 4.3.1: Introduction of protein at gene VIII

The linear PCR product is amplified to regenerate an intact circular viral genome whose gene VIII protein now has an amino terminal extension. The mutagenized M13 is introduced into *E. coli* cells. It is naked DNA (not packaged with viral coat proteins) so it is introduced into *E. coli* like a plasmid (i.e. via electroporation or using CaCl_2 competent *E. coli* cells).

- The extruded viral progeny will:
 - Be somewhat longer as the genome is larger but should be packaged o.k.
 - Express the mutant gene VIII (major coat) protein as part of the filamentous phage assembly

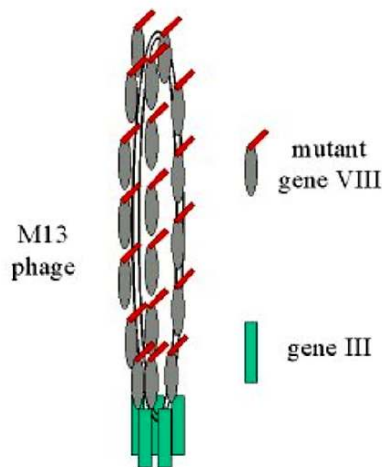


Figure 4.3.2: Mutated M13 progeny

Selection

The amino-terminal peptide may have affinity for a specific ligand, possibly an antibody. We can use this binding affinity to purify the mutant phage. We can construct a chromatography column where the chromatography matrix has attached to it the ligand of interest. If we pass our phage through this column, the mutant phage containing the mutant gene VIII protein will bind to, and be immobilized, on the column

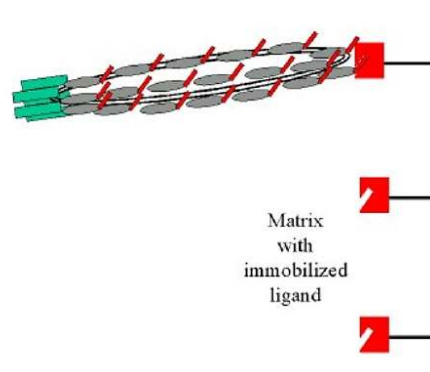


Figure 4.3.3: Immobilized ligand

- Any wild-type phage will not bind, and will pass through the column
- The bound phage can be eluted and used directly to reinfect *E. coli* to propagate viruses with the mutant gene VIII protein

The mutant gene VIII protein which has been purified carries along with it the DNA which codes for itself. In other words, we have specifically selected for the DNA which codes for the gene VIII of interest (i.e. the mutant)

Changing binding specificity

Suppose we have a slightly different ligand attached to the matrix...



Our phage with mutant gene VIII protein will not bind very well to this different ligand. If we introduce the appropriate amino acid changes into our mutant gene VIII protein, we may be able to improve the binding affinity (specificity) and get the phage to bind to the new ligand

We do not know what specific amino acid changes to make, so we will introduce random changes into the amino terminal extension of the gene VIII protein

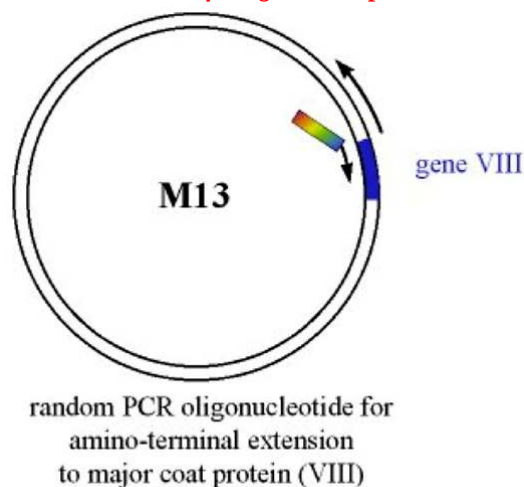


Figure 4.3.4: Random changes in amino-terminal extension

- After PCR and ligation, this will produce a **library** of phage genomes, each with a potentially *different* sequence for the amino terminal extension of the gene VIII protein
- Generally speaking, when these are introduced into *E. coli* the efficiency is typically so low that, at most, each successful transformation results in only one phage genome per *E. coli*

- Progeny phage from our library will, therefore, have only one type of gene VIII protein per phage (a single phage will have the gene VIII protein which it's DNA codes for).
- If we pass the collection of progeny phage from our library over our ligand column, we will selectively bind those phage whose mutant gene VIII protein has affinity for the ligand

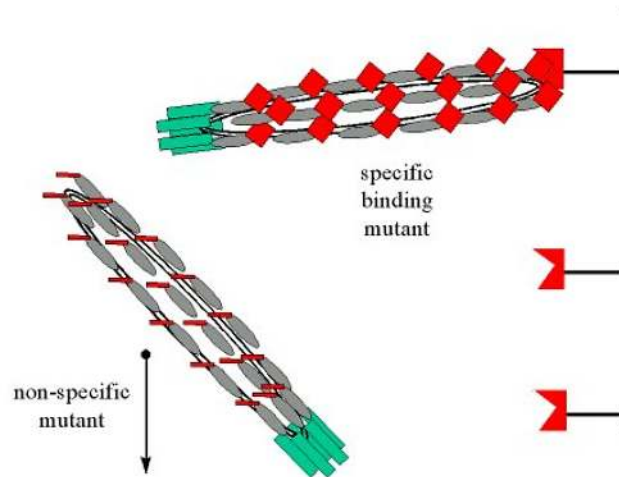


Figure 4.3.5: Specific binding

- If we isolate the bound phage we can re-infect *E.coli*, and isolate the progeny phage
This will result in a second library which will be enriched for mutant gene VIII sequences with the desired ligand binding specificity
- In this first selection we may have many other phage which might be weak, or generally non-selective, ligand binders. So, the process is repeated until we have a library of specific high-affinity mutants (this is sometimes called "Bio-mining" or "Bio-panning")
- **We can then sequence the DNA and find out what sequence(s) in the gene VIII mutant provide high binding specificity to our new ligand**

Practical limitations

- Over what length of DNA/protein can we randomize?

Example: a random stretch of 10 amino acids

- 20^{10} possible different unique polypeptide sequences. Each amino acid is 110 g/mol or 1.8×10^{-22} g per amino acid. This would mean that a complete random library of 10mer peptides would have a mass of 1.8×10^{-9} g.

Example: a random stretch of 20 amino acids

- 20^{20} possible different unique polypeptide sequences. Each amino acid is 110 g/mol or 1.8×10^{-22} g per amino acid. This would mean that a complete random library of 20mer peptides would have a mass of **18.9 Kg!**

This does not even consider the significantly larger mass of the virus particle of which it would be a part. This type of calculation demonstrates that generating random DNA sequence libraries is limited to very short stretches of about half a dozen or so codon positions.

Random Mutagenesis

We could randomly introduce all four bases at each position (NNN) while trying to synthesize a mutagenic random codon, but what would this give us?

Leu, Arg, Ser (6 codons/64)	9.4%
Val, Pro, Thr, Ala, Gly (4 codons/64)	6.3%
Ile (3 codons/64)	4.7%
Phe, Tyr, His, Gln, Asn, Glu, Asp, Cys (2 codons/64)	3.1%
Met, Trp (1 codon/64)	1.6%
Stop (3 codons/64)	4.7%

- Thus, not all amino acids are present in equal amounts, and we will have 4.7% stop codons
- We can improve this by using all four bases at the first two positions, and an equal mix of G and C bases at the third (wobble) position. This would result in 32 possible codons and the following distribution of amino acids and codons:

Phe, Tyr, Cys, Trp, His, Gln, Ile, Asn, Met, Lys, Asp, Glu, Stop (1 codon/32)	3.1%
Gly, Val, Ala, Pro, Thr (2 codons/32)	6.3%
Leu, Arg, Ser (3 codons/32)	9.4%

Another general problem with the above method is that of **non-specific** binding

- This is due in part to the large number of mutant gene VIII proteins (~2500) on the virus surface.
- Ligand binding may be due to the interaction of two or more gene VIII proteins contributing to binding the ligand (i.e. "polyvalent" binding)
- In this case, the isolated mutant peptide may show weak or no binding affinity at all

How can we minimize the number of mutant proteins per phage particle?

Approach #1 - Mutate gene III protein instead of gene VIII

- While there are ~2,500 copies of the gene VIII (major coat protein) per phage, there are 5 copies of the gene III (minor coat protein) per phage

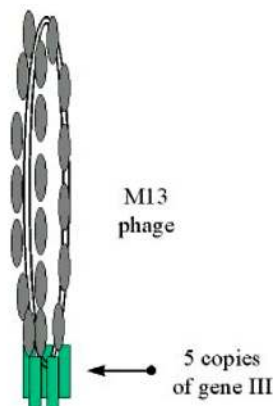


Figure 4.3.6: Mutating gene III

- We can introduce the mutagenic sequence to the amino-terminal of the gene III protein and thereby limit the number of copies on the phage particle

- this may still result in polyvalent binding

Approach #2 - Supplement with wild-type gene III protein during phage assembly

- If the M13 phage contains a second copy of gene III which we do not mutate then when the phage assembles (as it extrudes through the outer membrane) the five gene III proteins it picks up will be a mixture of mutant and wild type (in roughly equal amounts). In this case, the progeny phage will have, on average, 2-3 copies of the mutant gene III protein
- If we can increase the expression of the wild type gene III protein, to the point where it is present in four-fold excess over the mutant, then the progeny phage will, on average, contain 1 copy of the mutant gene III protein (and ligand binding will, by definition, be mono-valent)

How can we do this?

- We can put a strong promoter in front of the wild type gene III
- We could place the wild type gene III near one of the terminators (after gene IV, for example). This position would result in higher expression of the wild type gene due to the cumulative effects of the upstream promoters

Variation

Expression of mutagenic protein as a fusion of an outer membrane protein of *E.coli* itself

- We can attach our mutagenic sequence to a gene for an outer membrane protein (e.g. *OmpT*, *LamB*)
- The surface of the *E. coli* cell itself displays the mutagenic peptide
- Outer membrane proteins may be present at a level of 10,000 molecules or more
- Our ligand column will selectively bind specifically binding *E. coli* cells. These can be eluted and cultured and the process repeated to select for the mutant sequence with binding affinity

Expression of mutagenic protein as a fusion with a DNA binding protein

- We can attach our mutagenic sequence to a gene for a protein which binds DNA (e.g. *lac* repressor)
- The mutagenic *lac* repressor will be encoded by a plasmid which itself contains two *lac* operator sequences
- The mutant repressor molecule expressed by the plasmid will bind strongly to the *lac* operator regions on the very same plasmid
- The cells are lysed (the repressor remains bound to its associated plasmid DNA) and the ligand binding step is performed to enrich for ligand binding sequences
- The associated plasmid DNA is isolated and transformed back into *E. coli* for amplification and another round of selection

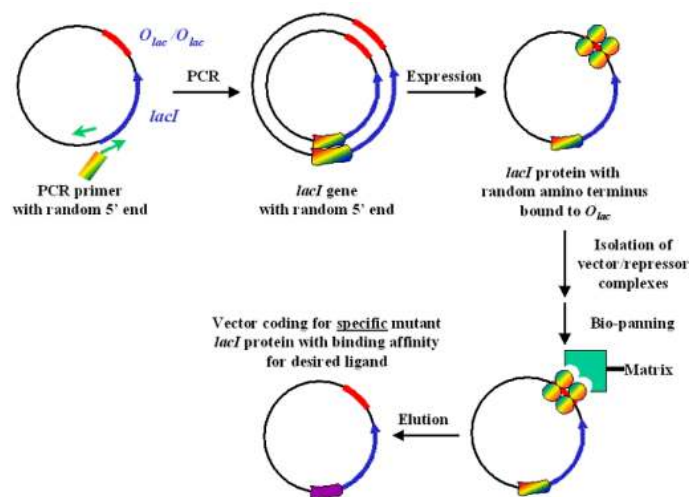


Figure 4.3.7: Fusing mutagenic protein and binding protein

This page titled 4.3: M13 Phage Display Libraries is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

4.4: SELEX (Selective Evolution of Ligands by Exponential Enrichment)

Tuerk and Gold, *Science* (1990) 249:505-510

Phage T4 genome encodes a DNA polymerase (T4 DNA pol)

The mRNA for T4 DNA pol contains a sequence which has binding affinity for T4 DNA pol. This sequence overlaps the Shine/Dalgarno sequence and binding of T4 DNA pol prevents the mRNA from being transcribed (and producing more T4 DNA pol). Thus, the expression of T4 DNA pol is self-regulating (**autogenous regulation**). The structure of the mRNA in this region forms a stem-loop structure

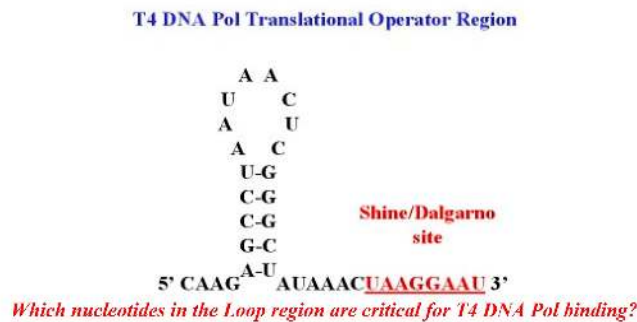


Figure 4.4.1: T4 DNA polymerase mRNA

The experimental design:

In simple terms, generate a library of mRNA molecules with a random nucleotide sequence in the loop region. This library will have a size of 4⁸ or 65,536 uniquely different sequences (if only the loop region is randomized). Identify those sequences with high affinity for T4 DNA pol and determine their sequence.

Template Construction

A "synthetic gene" duplex DNA template is constructed from five primers:



Figure 4.4.2: Template construction

- The synthetic gene is 110 nucleotides in length. It is composed of oligonucleotides #3, #4 and #5. Oligonucleotides #1 and #2 can be thought of as "bridging" oligos which allow #3, #4 and #5 to be ligated
- Oligonucleotide #4 contains the 5' end of the T4 DNA Pol gene corresponding to the mRNA Shine/Dalgarno sequence and the T4 DNA pol binding step/loop. It is synthesized to contain **random bases** at the 8 base long loop structure of the T4 DNA pol binding domain.
- Oligonucleotide #1 contains the sequence for T7 RNA polymerase promoter region. Transcription can be driven from this promoter by addition of T7 RNA polymerase and ribonucleotides.

in vitro transcription

- The 110 nucleotide gene (double-stranded DNA form) can be transcribed *in vitro* by T7 RNA pol
- Transcription produces a 92 nucleotide long RNA molecule. There will be a library of potentially 65,536 unique RNA molecules

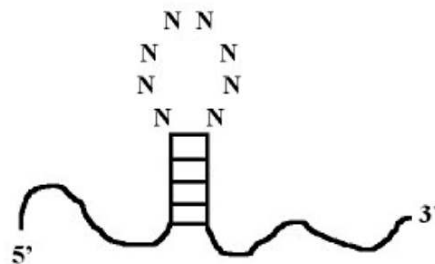


Figure 4.4.3: In vitro transcription

Selection by binding to T4 DNA pol

Those RNA transcripts with a sequence which confers binding specificity for T4 DNA pol are **selected for** by incubating the RNA library with T4 DNA pol which has been immobilized on nitrocellulose filters. The filters are washed to remove non-specifically bound RNA molecules. Tight (specific) binding RNA molecules are eluted from the T4 DNA pol protein and are collected

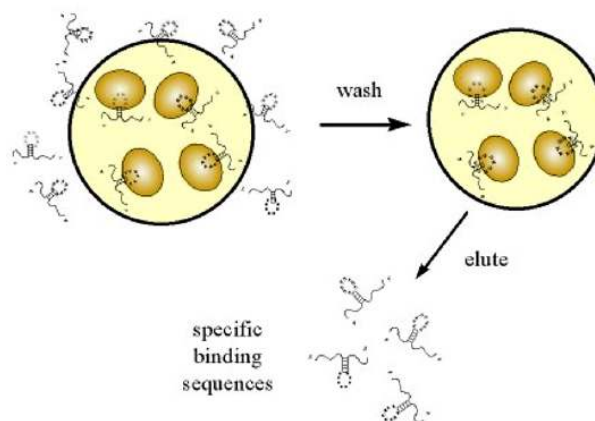


Figure 4.4.4: Specific binding

cDNA construction

The eluted RNA molecules are converted to single strand cDNA using oligonucleotide #5 as a primer, and adding Reverse transcriptase and dNTP's. Duplex DNA is produced from the single strand cDNA by PCR using oligonucleotides #1 and #5.

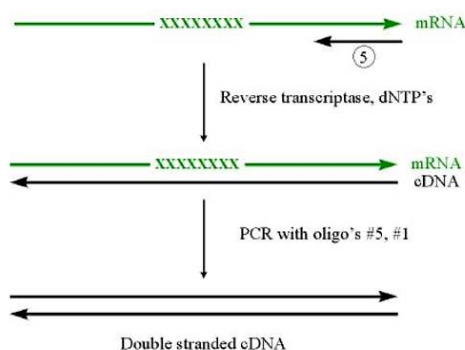


Figure 4.4.5: cDNA Construction

In vitro transcription

- In vitro transcription produces a (smaller) library of RNA molecules enriched for sequences with binding affinity for T4 DNA pol
- the process of selection and enrichment is continued until a collection of high affinity binding sequences has been produced

Experimental results

1. Library construction

- The original "synthetic gene" library was sequenced *en mass*
- A "ladder" of bases at the random position indicated that the library started out being more or less a completely random collection of sequences in this region (i.e. there did not appear to be any sequence bias in the collection of molecules)

2. Samples from the sequential rounds of selection and amplification were saved for sequencing analysis

- The results indicated that as the selection and amplification process continued certain bases were observed at specific positions in the 8 base mutagenized region

Wild type seq: A A U A A C U C
 Mutant seq: A ^{AU} _{GC} A A C ^{UC} _{CU}

Figure 4.4.6: Mutant sequence

The wild type sequence could be found within this heterogenous mixture of tight binding sequences. 20 clones from this final library were sequenced.

Wild type seq:	A A U A A C U C	9/20 clones
	A A U G A C U C	1/20 clones
	A A U A A C U U	1/20 clones
	A G C A A C C U	8/20 clones
	A G C G A C C U	1/20 clones

Figure 4.4.7: Mix of tight binding sequences

- There were essentially only two different sequences present: the known wild type sequence and another different sequence (which varied at four positions)
- Both had similar tight binding characteristics (the minor sequences identified exhibited weaker binding)
- This alternative sequence may have a different stem-loop structure than the wild-type

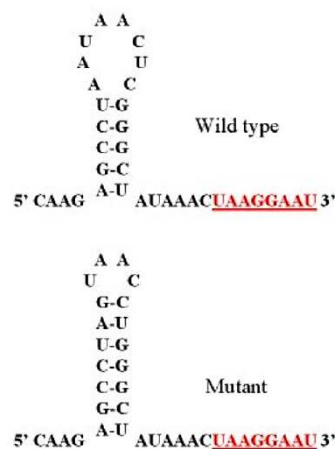


Figure 4.4.8: *Change in stem loop structure*

Conclusions

1. The SELEX method successfully identified the wild type sequence and an alternative (previously unknown) sequence with high affinity T4 DNA pol binding properties
2. The results suggest that the sequence of the wild type translational operator is pretty much optimized for binding (unable to improve further?)
3. The SELEX method should simplify the study of interactions between
 - transcriptional activators and repressors and transcriptional complexes at promoter sites
 - replication accessory proteins and DNA pols at origins of replication
 - ribosomes and translational repressors at ribosome binding sites
4. Results may be relevant to DNA/protein binding interactions in certain cases (i.e. if interactions do not involve 2' hydroxyl group and if a structure which DNA can adopt is involved)
5. Instead of beginning with all possible sequences, start with a smaller library and combine replication with error prone polymerases to generate additional sequences

This page titled [4.4: SELEX \(Selective Evolution of Ligands by Exponential Enrichment\)](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

4.5: Protein-Protein Recognition Probed Using a Yeast Transcriptional Activator System

Fields, S. and Ok-kyu, S. *Nature* (1989) **340**:245-246

Background

The yeast genome contains Upstream Activator Sequences (UAS's)

UAS are located 5' to the coding region of a gene and are regulatory regions - they bind transcriptional activator proteins. These transcriptional activator proteins, when bound to UAS, stimulate transcription (i.e. they may contain binding sites for RNA polymerase)

Yeast gene, protein nomenclature

Wild type genes are designated with capital letters in italics (e.g. *GAL4*). Mutant alleles of the genes are indicated in lowercase italic (e.g. *gal4*) and encoded proteins are indicated by roman type, with the first letter capitalized (e.g. Gal4)

Growth of yeast in galactose containing media

Incubation of wild type yeast in galactose results in >1,000 fold increase in the mRNA level for enzymes involved in galactose metabolism. This increase in mRNA levels is not observed in *gal4* mutants. The Gal4 protein is not one of the enzymes involved in galactose metabolism, rather, it appears to be a *transcriptional activator protein* which specifically activates the transcription of genes involved in the galactose metabolic pathway. Gal4 protein binds to a specific 17 nucleotide long region of DNA located in the 5' region of genes involved in galactose metabolism. These regions are termed UAS_{GAL}. Binding of Gal4 to UAS_{GAL} increases transcription from the nearby promoter

Gal4 Protein

- Single polypeptide chain which contains two domains:

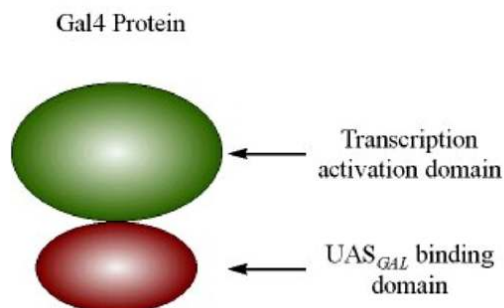


Figure 4.5.1: Gal4 protein domains

These domains are functionally separable in the polypeptide chain (Figure 4.5.2). The amino-terminal (1-74) domain by itself can bind to UAS_{GAL} sequences but cannot activate transcription. The carboxyl-terminal (738-881) domain contains the activating region but cannot activate transcription because it fails to localize to the UAS_{GAL} region (i.e. near the promoter region)

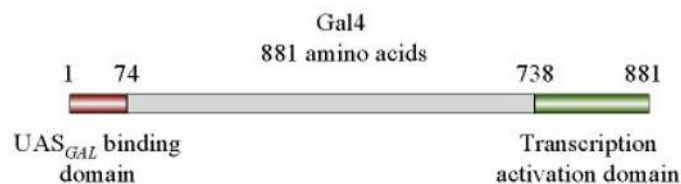


Figure 4.5.2: Gal4 polypeptide chain

Experimental Design

Imagine we have two proteins (**X** and **Y**), which normally form a stable complex (XY)

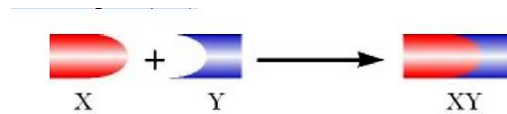


Figure 4.5.3: Complex

If we make a Gal4(1-74)- **X** fusion protein and a **Y** -Gal4(738-881) fusion protein we can potentially form a functional Gal4 transcription activator (i.e. a Gal4 activator region in complex with a UAS_{GAL} region)

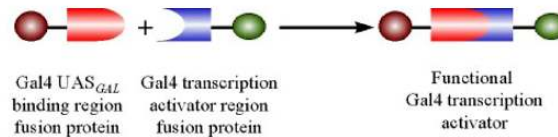


Figure 4.5.4: Formation of Gal4 transcription activator

This fusion protein complex can act as a functional transcriptional activator for the galactose metabolic enzymes

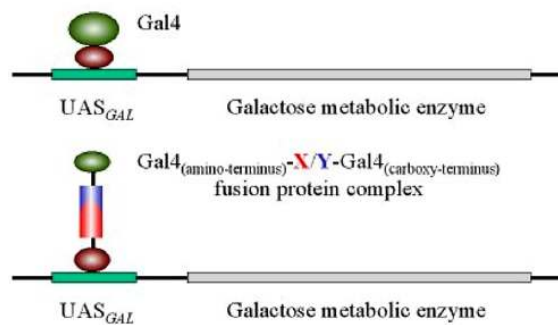


Figure 4.5.5: Protein complex acting as transcription activator

To test this hypothesis, the authors used the following system

- The **protein complex** involved the two proteins SNF1 and SNF4. The SNF1 protein is a serine-threonine specific protein kinase, and the SNF4 protein is known to be physically associated with SNF1 and is essential for function
- 5 different gene constructs involving Gal4 and SNF1 and SNF4 were made:

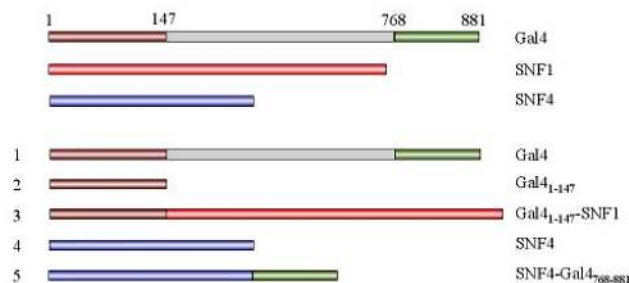


Figure 4.5.6: Gene constructs with Gal4, SNF1, SNF4

- These gene constructs were inserted into two different yeast plasmids:

- All constructs containing amino terminal fusions of *GAL4* were put into plasmids containing the *HIS3* gene (coding for a biosynthetic enzyme necessary for the production of the amino acid **histidine**).
- All constructs containing carboxy terminal fusions of *GAL4* were put into plasmids containing the *LEU2* gene (coding for a biosynthetic enzyme necessary for the production of the amino acid **leucine**)
- These plasmids were inserted into a host yeast (GGY1::171) which had the *GAL4* gene deleted, the *LACZ* gene deleted, the *HIS3* and *LEU2* genes mutated to be non-productive and which also contained a *GAL1-LACZ* gene fusion (i.e. a LacZ protein driven from the *GAL1* promoter which has a UAS_{GAL} . This host is thus *DGAL4*, *DLACZ*, *his3*, *leu2*, *gal1-lacZ*.
- Thus, selective pressure for the amino terminal *GAL4* fusions can be maintained by eliminating histidine in the growth media (yeast relies on the *HIS3* gene on the plasmid to make histidine)
- Likewise, selective pressure for the carboxy terminal *GAL4* fusions can be maintained by eliminating leucine in the growth media (yeast relies on the *LEU2* gene on the plasmid to make leucine)
- The host can harbor both kinds of plasmids (in which case the media would lack both histidine and leucine)
- The expression of the fusion constructs on the plasmid is actually regulated by galactose. Adding galactose results in expression of the fusion constructs. However, since the host is *DGAL4* all the other galactose metabolic genes (e.g. *GAL1*) will not be up-regulated in response to added galactose.

The presence of a functional Gal4 protein in the GGY1::171 yeast host will result in the following:

- Expression of the LacZ protein (contained in the host *GAL1-LACZ* gene fusion) will occur. Thus, in the presence of XGAL substrate the yeast colony will turn blue
- **All the host genes which contain an UAS_{GAL} will be up-regulated. This will include all the wild-type genes involved in galactose metabolism (e.g. *GAL1*)**

Experimental Results

Plasmid	LacZ activity
1. None	<1
2. <i>GAL4</i> (1-881)	4,000
3. <i>GAL4</i> (1-147)	<1
4. <i>GAL4</i> (1-147)-SNF1	<1
5. SNF4	<1
6. SNF4- <i>GAL4</i> (768-881)	<1
7. <i>GAL4</i> (1-147)-SNF1; SNF4- <i>GAL4</i> (768-881)	180
8. <i>GAL4</i> (1-147)-SNF1; SNF4	7
9. <i>GAL4</i> (1-147); SNF4- <i>GAL4</i> (768-881)	<1

- Functional Gal4 protein resulted in expression of the host *GAL1-LACZ* gene
- Co-expression of plasmids containing *GAL4*(1-147)-SNF1; SNF4-*GAL4*(768-881) fusion proteins resulted in expression of the host *GAL1-LACZ* gene, indicating the presence of functional Gal4 protein (i.e. complex formation of the Gal4 fusion proteins)
- The level of LacZ protein was less, however, indicating that the SFN1- and SFN4- Gal4 fusion protein complex was a less efficient as a transcription activator than wild type Gal4 protein
- There was some low level LacZ protein activity with the *GAL4*(1-147)-SNF1; SNF4 protein complex. This suggests that the SNF4 may have some minor affinity for RNA polymerase

Conclusions

This method may prove useful for identifying proteins which can form stable complexes (i.e. identifying specific protein:protein interactions). For example, if we want to know which proteins bind to a certain receptor we can make a *GAL4*(1-147)-receptor

fusion protein and screen a cDNA library for potential binding proteins by constructing cDNA-*GAL4*(768-881) fusion constructs. Potential binding proteins from such a library can be screened by looking for blue colonies with XGAL substrate, or selected for by growing the yeast in media with galactose as the carbon source.

This page titled [4.5: Protein-Protein Recognition Probed Using a Yeast Transcriptional Activator System](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

4.6: Molecular Imprinting

Mosbach, K. TIBS (1994) 19: 9-14

Molecular recognition is the foundation of many biological processes. A major emphasis in biotechnology is the development of **synthetic recognition molecules** and systems which are specific for a ligand of interest. One branch of molecular design involves the synthesis of molecules with **specific stereochemical structure** designed to recognize and interact with a particular ligand. In contrast, **molecular imprinting** is a method of constructing a molecule, with specific recognition properties, by having the **ligand itself** direct the assembly of the desired structure

Technical aspects

The general approach of the technique is to combine the desired "imprint" molecule with a solution of small monomeric molecules which has several desired features. They have a variety of functional groups (e.g. hydrogen bond donors, acceptors, possibly acidic or basic groups, or aliphatic or aromatic groups). One or more of these groups can specifically interact with the imprint molecule (e.g. they can form specific hydrogen bonds). The monomeric molecules can be made to polymerize to form a rigid, or semi-rigid polymer. The imprint molecule is soluble in the monomer solution. The monomers, either in solution or during the polymerization process, do not form covalent bonds with the imprint molecule (the interaction is strictly non-covalent)

After polymerization of the monomers (in the presence of the imprint molecule) the general features of the polymer are as follows:

1. The polymer is **porous enough** to allow the imprint molecule to diffuse out (and likewise to diffuse in)
2. The polymer is **rigid enough** that the organization of monomeric molecules in response to the imprint molecule is retained.
3. Therefore, the stereochemistry (i.e. size, shape and orientation of appropriate functional groups) of the pocket where the imprint molecules were located is retained.

These sites in the polymer constitute an "induced molecular memory" capable of selectively recognizing the imprint molecule

The typical polymer production technique has been to produce a bulk imprinted polymer and then grind it up to produce small particles (25 mm). Such small beads are appropriate for liquid chromatography methods (i.e. can be used as chromatographic resins and packed in a column). "Suspension" polymerization can yield beads directly from the polymerization process

Most of the reported molecular imprinted polymers reported have been made from monomers of acrylates, styrenes or silicates, and the imprinting step performed in organic solvents. Imprinting in organic solvents (e.g. low dielectric constant solvents) enhances hydrogen bonding (i.e. partial electrostatic) interactions. However, it can limit imprint molecules to those that are soluble in such solvents.

Example 4.6.1

Imprinting of the amino acid L-phenylalanine with methacrylic acid monomers and the enantioselective binding of the imprint molecule.

- Step I. "Imprinting"

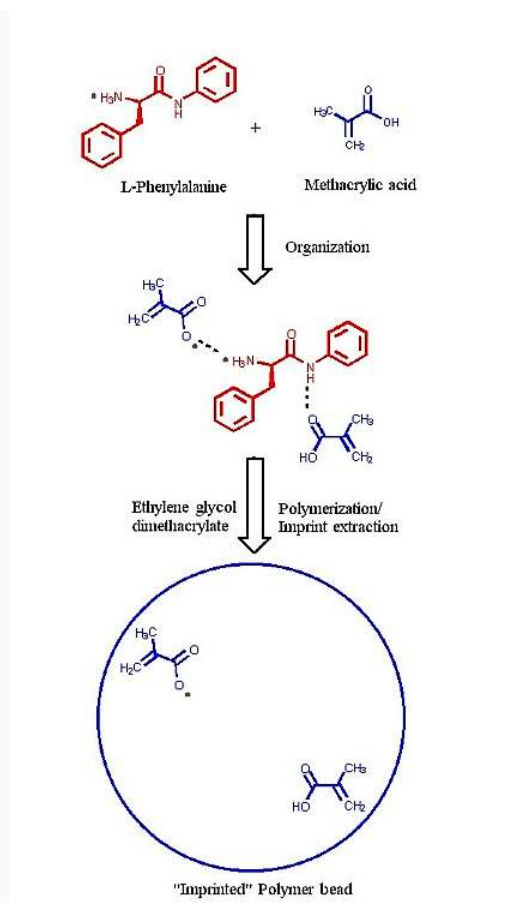


Figure 4.6.1: Imprinting

- Step II. Use of the imprinted polymer as a chromatographic resin to preferentially achieve the enantiomeric selection of L-phenylalanine from a racemic mixture of L- and D-phenylalanine

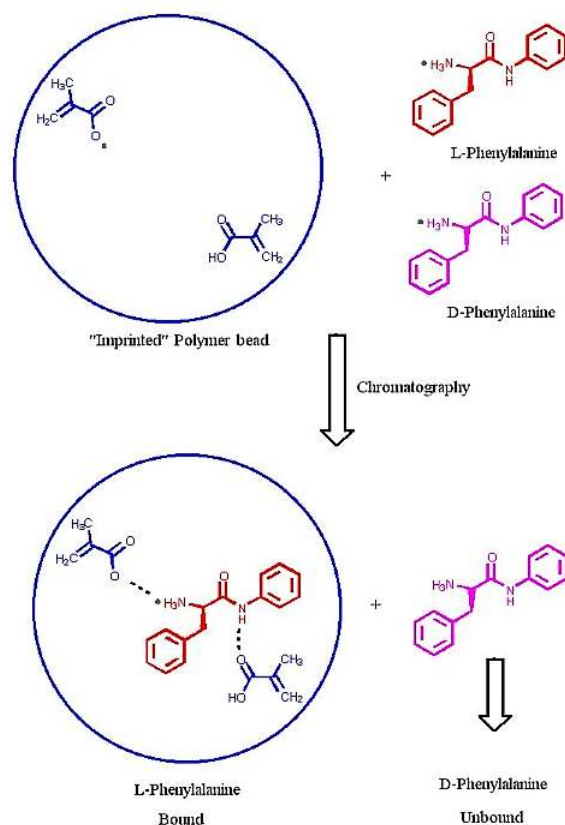


Figure 4.6.2: Enantiomeric selection

Applications

Three main areas of application of MIP's include chromatographic matrices for difficult (e.g. enantiomeric) separations, catalytically active polymers or enzyme mimics, and Bio-sensor devices.

Separations

There are about 500 drugs on the market which are optically active, about 90% of these are administered as racemic mixtures. The FDA now requires that for optically active drugs both enantiomers must be treated as separate substances in pharmacokinetic and toxicology profiling. There is a great need for preparative methods of enantiomer separations

Other separations

MIP's made against particular alkaloids or opiates can discriminate between the imprinted compounds and closely related compounds

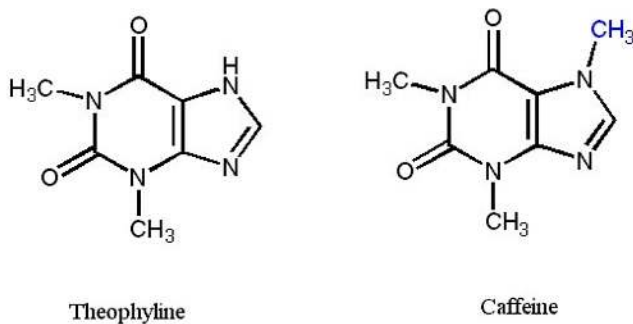


Figure 4.6.3: Closely related compounds that can be separated with MIP's

Imprinted polymers as enzyme mimics

If imprints are prepared against *transition state analogs*, the MIP will stabilize the transition state and result in enhanced catalytic rates. Increased catalytic rates (turnover) in comparison to substrate turnover rates in solution have been observed with polymers imprinted with transition state analogs. However, they are generally poor catalysts. This is most likely due to the fact that specific stereochemical placement of catalytic groups was not attempted. Thus, such MIP's represent an appropriate framework within which to place catalytic groups to achieve efficient catalysis

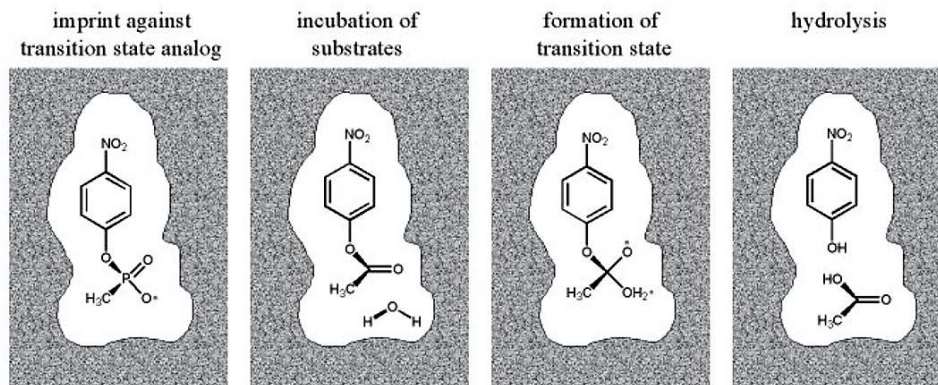


Figure 4.6.4: Imprints as enhancer for catalytic rates

Imprinted polymers as substrate selective sensors

"biosensors" are devices which integrate specific receptor molecules with signal transduction devices. The transducer may convert the binding energy of the MIP/analyte into changes in protons, light absorbance, heat, etc. This physical/chemical signal is amplified and output to a measuring device. In principle, the analyte can be any molecule which can be specifically imprinted. For example, an environmental toxin.

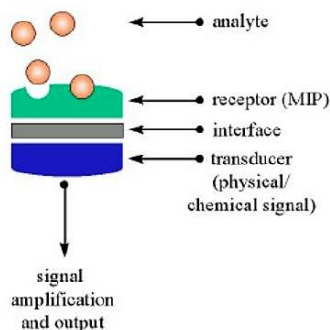


Figure 4.6.5: Imprinted polymer as biosensor

Conclusions

Although most work with MIP have involved small molecules, in principle it should be possible to extend MIP utility to include biomolecule recognition (i.e. proteins, nucleic acids). Such MIP should exhibit similar specificities. If MIP can mimic biomolecule binding sites, they may prove useful alternatives for the screening of new inhibitors, or antagonists. One potential limitation is that of "mutagenesis". Can a polymer which has already been imprinted be "mutagenized" (chemically modified?) to produce novel/enhanced recognition characteristics? Another limitation is that the exact stereochemical structure of the imprint is not known. Therefore, the structural information of a useful imprint is not available.

This page titled [4.6: Molecular Imprinting](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

CHAPTER OVERVIEW

5. Lab Notes Part 1

5.1: Basic practices and techniques, Laboratory notebooks, Presentation of data, Sections of a scientific report

5.2: Spectrophotometry

5.3: Quantification of Protein Concentration

5.4: Chromatography

5.5: Gel Electrophoresis of Proteins

5.6: Introduction to Protein Purification

This page titled [5. Lab Notes Part 1](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

5.1: Basic practices and techniques, Laboratory notebooks, Presentation of data, Sections of a scientific report

There are two main concerns in a biochemical laboratory environment, they are:

- Safety
- Accurate record keeping

Safety

A biochemical laboratory has safety concerns that include not only issues of chemical safety, but also issues of biohazard safety. In both cases, your safety depends upon a knowledge of the materials and procedures you are working with, as well as appropriate protective dress.

Safety issues will be covered in detail in the syllabus, and in the lab by the lab director, however, a few key points will be covered here:

1. Read the relevant chapter of the lab manual BEFORE arriving for the lab. To reinforce this practice, the labs will start with a short quiz related to the lab for that day. The quiz will start PROMPTLY at the beginning of class, will last approximately 10 minutes, and there will be NO MAKEUP QUIZZES for late arrivals. The beginning of lab time is used to organize the lab, and includes relevant safety information. Therefore, it is essential for all concerned that students show up promptly for the labs.
2. Dress appropriately. No open toed shoes (i.e. sandals), shorts, or bare midriffs are permitted in a laboratory setting (OSHA guidelines). YOU WILL BE ASKED TO LEAVE IF YOU ARE NOT APPROPRIATELY DRESSED. Lab coat and goggle guidelines will be addressed by the lab manager.
3. Lab attitude and behavior. Both safety and a successful learning experience are keenly associated with behavior and attitude. A cavalier attitude will result in accidents and experimental failures. Thinking about what you are doing, and asking questions, will result in success and a long life. One of the biggest problems in a laboratory setting is equipment damage due to a lack of knowledge about how to use such equipment. DO NOT USE ANY EQUIPMENT WITHOUT FIRST BEING INSTRUCTED ON HOW TO USE IT. Due to the expense of scientific equipment, labs can only function through the use of shared equipment, thus, everyone using such equipment needs to take care of it.

Accurate Record Keeping

There are typically two levels of record keeping for scientists involved in biochemical research:

1. The laboratory notebook
2. The research report

It is important to understand the purpose of each type of record keeping, and to adhere to the practices associated with each one.

The laboratory notebook

The lab notebook can be thought of as your "working record" or "experimental log". It is a sort of formalized diary, and has a specific purpose and associated formalism with regard to how to use and maintain it.

- For purposes of invention, discovery and patents, it is the primary record used to establish ownership of an invention and dates of discovery. To be useful for this legal process, certain guidelines must be followed.
- For the purpose of the scientific process it provides the necessary detail so that experiments can be repeated and the results replicated and confirmed
- For the purpose of reporting results, the lab notebook is the repository of all primary data and results

To meet the above needs, the lab notebook is:

- **bound** (no loose leaf) and
- **pages are numbered sequentially.**

· Furthermore, entries in the lab notebook are **always made in ink (black or blue)**, never pencil. If something is written that is incorrect, it can be crossed out.

In this way, an examiner of the notebook can determine whether the lab notebook is **intact** or whether pages or information have been **removed**.

Entries into the notebook for a given experiment always begin with:

- Writing the **date**
- Stating the **purpose** of the experiment

The remainder of the notebook section that covers the experiment will include any and all relevant information associated with the experiment, including:

- A description of the materials, equipment and reagents used
- A description of the methods used
- Calculations used (a truly essential aspect of trying to repeat the experiment and identifying errors)
- Tabulation of results collected
- Conclusions, hypotheses, notes, etc.

The key point is that you include enough information so that someone else can use your notebook to replicate the experiment and obtain the same results you did. In this regard, the notebook does not have to have perfect penmanship, but it must be legible and understandable by others. Diagrams are often essential. In industrial (and some academic) settings, it is a requirement that another person in the lab sign off on your notebook that they have **read, and can understand**, your notebook entry for a particular experiment (this is associated with the legal function of the notebook).

The notebook must not contain any loose bits of paper. If an experiment includes a photograph or some type of printout from an instrument, this must be firmly affixed to the notebook (i.e. glued & stapled). Any loose paper can be lost, and with it, some key information. The best approach, if you have a data printout from an instrument, is to copy it by hand into your notebook. Sometimes this is not practical, so the data must be glued & stapled. It is also not a good idea to have information beneath such added items. In some industrial settings, the lab notebooks are Xeroxed for archival purposes. In this case, Xeroxing a page will miss information if it is written underneath a photo, etc.

The first few pages of the lab notebook should be reserved for a table of contents. This is associated with the function of being able to replicate an experiment. Part of this requirement is the ability to locate a particular experiment in a given notebook. Researchers can generate dozens of notebooks involving hundreds of experiments, and a table of contents is essential. Furthermore, a researcher will often have two or more experiments on-going at one time. It is common for a notebook entry to start on one page and then reference a later page for continuation of the experiment (sort of like some magazine or newspaper articles). If appropriate, notebook pages should indicate both the referring page number, and the continuation page number.

The research report

The lab notebook is your working journal, and is not intended as a formal research report. The formal research report has **specific sections and organization**, and is a **final, polished document** relating to your research. These sections include the following (in specific order):

- Title
- Introduction
- Materials and Methods
- Results
- Discussion
- References

In many cases, such reports also have an ***abstract*** (a short, ~250 word report that precedes the Introduction and details the essential aspects and results). However, this will not be required in your own reports.

Each section of the report has a specific function:

Title

- Typically a **single sentence** that explains to the reader what the experiment is all about

Introduction

- One or more paragraphs that explain the purpose of the experiment, and provide any **relevant background information** so that the reader is prepared to follow the narrative of the report. In published research reports the introduction also lays out a logical argument as to the importance of the experiment

Materials and Methods

- A detailed "recipe", describing instrumentation, materials and supplies, and procedures used in performing the experiment. The "M&M" section is detailed enough so that the experiment can be repeated and the results confirmed by others. An essential aspect of the scientific method is that results are meaningless unless they can be confirmed by others.

Results

- Statements of fact regarding the data collected, with associated errors or deviations (e.g. all relevant values and/or numbers relating to your data). This is often represented using tables, and sometimes, figures.

Discussion

- An **interpretation** of the results, and a discussion of its meaning, including a discussion of error (where relevant). In published research reports, the discussion will also put the work in the context of other published work in the same field.

References

- If you used any texts, primary literature, or web sites in preparing your experiment or report, these should be referenced. There are a variety of formats, but the reference generally includes author, title, volume, pages and year (also publisher, if it is a text). For web sites, author and URL. A key aspect is to avoid **plagiarism**.

An addition to having the above sections, all research reports have a required **format**. The format of a report includes details of:

- Type of font, line spacing, page breaks, margins, etc.
- Page length and/or word number requirements
- Limitations on number of tables, figures, use of color, etc.
- Title page format
- Formatting of references

This course is no different, and **there is a required format for research reports**, and these can be found in the [syllabus](#).

Common mistakes in writing the research report

Although it may seem straightforward, writing an effective research report can take years of practice. There are several aspects that can prove difficult. Here are a few things to consider:

- Do not put Results into the Materials and Methods section. The M&M section may have equations, but will not have any data listed.
- Do not put comments associated with the Discussion section into the Results section. Sometimes it is difficult to know what goes where, but here is a general guideline: information in the Results section can be thought of as statements of fact (i.e. your data), whereas, interpretation and hypotheses are part of the Discussion section. Someone might argue with your ideas in the Discussion section, but there should be no argument about information in the Results section.
- Error associated with your data is presented in the Results section as well. Do not get hung up on semantics. A "discussion of error" goes into the Results section.
- Do not duplicate statements or paragraphs in different sections. If you have said something in the Introduction, do not repeat it in the Results or Discussion section.
- Do not duplicate data. If you present data in a table, do not also present it as a figure. However, having said this, it is important to recognize that, even with scientists, people can often grasp meaning easier with a figure as opposed to looking at numbers in a table. Often, selected data is extracted and compared using an appropriate figure.

- **YOU MUST FOLLOW THE FORMATTING GUIDELINES FOR THE RESEARCH REPORT.** In the "real world" failure to follow such guidelines will result in immediate rejection of the report.

An example of a research report is **given here**.

This page titled [5.1: Basic practices and techniques, Laboratory notebooks, Presentation of data, Sections of a scientific report](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

5.2: Spectrophotometry

The human eye responds to electromagnetic radiation within a range of wavelengths between 400-750 nm (i.e. the "visible spectrum"). Samples of light that contain a continuous spectrum of all wavelengths between 400-750nm will be perceived by the brain as "white light" (e.g. the sun). Light comprising a specific wavelength within this range is perceived by the brain as being "colored":

- ~400nm = "violet"
- ~450nm = "blue"
- ~500nm = "green"
- ~550nm = "yellow"
- ~600nm = "orange"
- ~700nm = "red"

Often objects appear colored because of their **absorption** of light within selective regions of the visible spectrum. The light from such objects that strikes our eyes (whose color we perceive) is composed of those wavelengths that the object **DID NOT** absorb.

For example, plant leaves contain two photosynthetic pigments: chlorophyll A and chlorophyll B. The chlorophyll A molecule has the ability to absorb light with a wavelength in the range of 430 and 660 nm; the chlorophyll B molecule absorbs light with a wavelength in the range of 450 and 640 nm. Thus, these two pigments in leaves absorb violet/blue and orange/red wavelengths of light (the energy that these photons represent is transferred to the chlorophyll molecules). Light with wavelengths in the range of 500-600nm is not absorbed by either molecule. Thus, after sunlight interacts with leaves the wavelengths that remain (and that our eye can perceive) are green-yellow. Therefore, plants are "green" because they don't absorb green light.

- What would an object look like if it contained a pigment that absorbed all wavelengths from 400-750nm? It would be "black"
- This is the same situation as being in a closed room with the lights turned out. **Without light there is no color.**

If we had a sample of leaf juice containing chlorophyll A and B we could quantitate the concentration of chlorophyll by the amount of violet/blue and orange/red light it absorbed. We **COULD NOT** quantitate it by the amount of green light transmitted (it would be essentially the same amount regardless of the amount of chlorophyll):

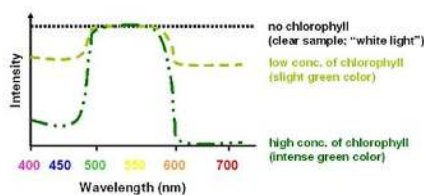


Figure 5.1.1: Chlorophyll light transmission

Even though the sample is "greener" with higher concentrations of chlorophyll, this is not due to an increase in the intensity of the green wavelength of light; it is due to a reduction in the violet/blue and orange/red wavelengths of light. It is this reduction that we can use to quantitate the concentration of chlorophyll. Although the sample appears green we cannot quantitate it by monitoring the green wavelength of light.

Light, Energy and Molecular Structure

Light is strange stuff, it is a radiative form of energy transfer, and in order to understand its properties we have to consider it to have both wave and particle properties. The different wavelengths of light differ in the energy they carry; the energy is directly proportional to the frequency of the light (the higher the frequency, the higher the energy):

$$E \propto \nu$$

The energy unit is the **erg** (1Joule = 10^7 ergs) and the frequency unit is **Hertz** (i.e. cycles per second, or just units of sec^{-1}). The constant of proportionality is Planck's constant, ***h***, with a value and units of 6.6×10^{-27} erg sec

$$E = hn$$

$$\text{erg} = (\text{erg sec})(\text{sec}^{-1})$$

The frequency (*n*) of light is inversely proportional to the wavelength (*l*):

$$n \propto 1/\lambda$$

Frequency has units of sec^{-1} and wavelength has units of meters. The constant of proportionality is *c* (the speed of light) with a value and units of $3 \times 10^8 \text{m/sec}$:

$$n = c(1/\lambda) = c/\lambda$$

This relationship is true in a vacuum; however, "optically dense" materials can slow the speed of light requiring the following correction to the above equation:

$$n = c/nl$$

where *n* is the "refractive index" of the material and has a value > 1

Therefore, the energy of light is inversely proportional to the wavelength:

$$E = (hc)/(nl)$$

The "photon" is a corpuscular (particle) description of light that is the carrier of the "quanta" of energy defined in the above equation. When light is absorbed by a molecule it transfers its quanta of energy, and the photon ceases to exist. Where did the energy go?

- The atomic bonds, chemical structure and electrons of a given molecule have **specific** excited states and vibrational modes.
- These excited states and vibrational modes have defined energy levels above the ground state.
- The molecule can change from the ground state to an excited state (or vibrational mode) upon the absorption of a quanta of energy exactly equal to the difference between the ground and excited states.
- **Thus, molecules can absorb the energy associated with specific wavelengths of light, and the light is consumed in the process.**

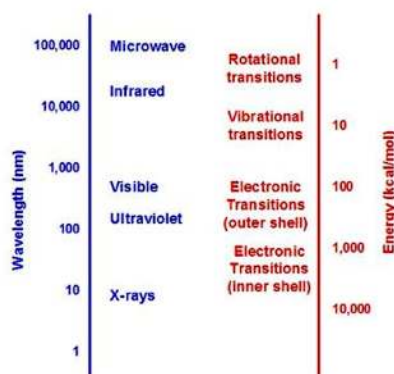


Figure 5.1.2: Scale of light

The atomic bonds, chemical and electronic structure of molecules are unique properties and differ from one type of molecule to another.

- Thus, the energetic differences between the ground and excited states differs from one type of molecule to another.
- Therefore, the ability to interact and absorb light, and the specific wavelengths of light absorbed differ from one type of molecule to another

The characteristic absorption pattern of the different wavelengths of light is unique for each type of molecule and is a type of **molecular "fingerprint"** that can be used to **identify and quantitate** molecules

In the example of chlorophyll given above the intensity of light passing through the sample is termed the "light **transmittance**" of the sample:

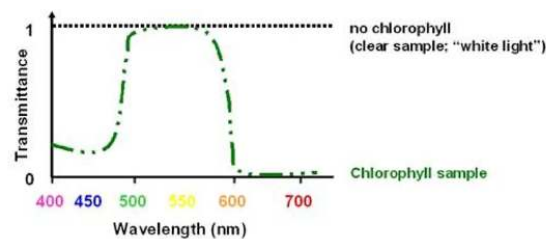


Figure 5.1.3: Light transmittance of chlorophyll

- Transmittance is a dimensionless number that varies from 1 (full transmittance) to 0 (no transmittance - complete absorption)

It is important to **note the distinction between transmittance and absorption**: they are **inversely proportional**

- A sample with high absorbance has a low transmission of light

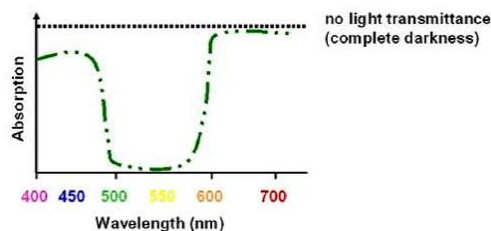


Figure 5.1.4: Chlorophyll absorption

How are transmittance (T) and absorption (A) quantified?

- The intensity of light shining on a sample is termed the incident light, I_0
- The intensity of light measured after passing through a sample is the transmitted light, I
- Transmittance, T , is defined as the ratio I/I_0 and will vary between 0 and 1:

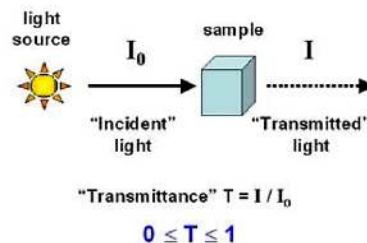


Figure 5.1.5: Transmittance

How is the transmittance affected by the sample concentration and physical size (i.e. path length)?

- If the transmittance is reduced due to the absorption by the sample, then the higher the concentration (c), the lower the transmittance. In other words, the transmittance is inversely proportional to the concentration: (Beer's law)

$T \propto 1/c$

- Similarly, the thicker the sample (i.e. the greater the **path length** of the light through the sample), the lower the transmittance. Thus, the transmittance is also inversely proportional to the path length (l): (Lambert's law)

$T \propto 1/l$

Thus, we expect the equation relating T to c and l to take the general form of:

$$T \propto 1/cl$$

Transmittance turns out to **decrease exponentially** with increases in the concentration and path length, thus the equation has the form:

$$\log T \propto 1/cl$$

$$-\log T \propto cl$$

The constant of proportionality is the extinction coefficient ϵ :

$$-\log T = \epsilon cl$$

Since log values are dimensionless, the units of ϵ would appear to be inverse concentration and inverse distance (e.g. $M^{-1}cm^{-1}$)

The $-\log T$ term on the left can be written as $\log(1/T)$ and identifies the **inverse** relationship between the transmittance T and the ϵcl term. Since absorbance and transmittance are inversely related, **the ϵcl term would seem to be a convenient definition for the absorbance (A):**

$$\log(1/T) = \epsilon cl = A$$

(the Beer-Lambert Law)

Values for ϵ and l

- Path length l is usually in units of **cm**
- Molar extinction coefficient ϵ has units of $M^{-1} cm^{-1}$ and is a constant of proportionality that relates the absorption of **molar** solutions
- Mass extinction coefficient $\epsilon^{1\%}$ refers to the absorbance of a 1% by mass solution. Typically this refers to an aqueous solution that we can take to have a density of 1000g/L. A 1% by mass aqueous solution would therefore refer to the dissolution of 10g/L, or a **10mg/ml** solution of the molecule of interest.
- Since the absorbance of a molecule is a function of the wavelength (i.e. the absorption is not equal for every wavelength) the extinction coefficient must also reference a wavelength. This is typically done using a subscript:

$$\epsilon^{1\%}_{280nm} = 14.5 g^{-1} L cm^{-1}$$

• In this case **a 10mg/ml solution of the molecule will have an absorbance reading of 14.5** (dimensionless units) at $l = 280nm$ (the absorption at other wavelengths may not be known). The units of concentration are g/L, thus ϵ will have dimensions of $g^{-1} L cm^{-1}$.

The relationship between changes in A , T and c

- The direct relationship between A and c means that there is a linear relationship between absorbance and concentration. If you double the concentration the absorbance will double, etc.
- The inverse log relationship between transmittance and absorbance can be stated as:

$$T = 1/10^A$$

- A doubling of the concentration will therefore result in a 10-fold reduction in the transmittance. Instrumentation for absorbance measurements (**spectrophotometers**) actually measure transmittance, and naturally become less accurate at low values of transmittance. Thus, the higher the absorbance reading the less accurate it is. **Most such instrumentation is inaccurate at absorbance readings > 1.5** (this works out to be at transmittance $< 3\%$ of full transmittance)

Design of Spectrophotometers

Spectrophotometers are precision instruments, however, conceptually they involve a relatively few number of parts. A simple design would look something like this:

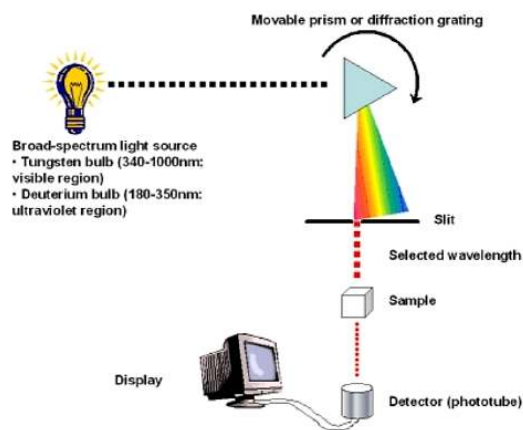


Figure 5.1.6: Spectrophotometer

- A tungsten lamp is used to produce wavelengths of light spanning the visible range, while a deuterium lamp is used to produce light spanning the ultraviolet range.
- The movable prism or diffraction grating is adjusted to direct the wavelength of interest towards the sample
- This is a "single beam" spectrophotometer and the reference and sample data are collected separately (the reference sample is used to determine the maximum transmission value (effectively I_0))

A modification of this design, a "dual beam" spectrophotometer, can permit simultaneous measurement of I and I_0 :

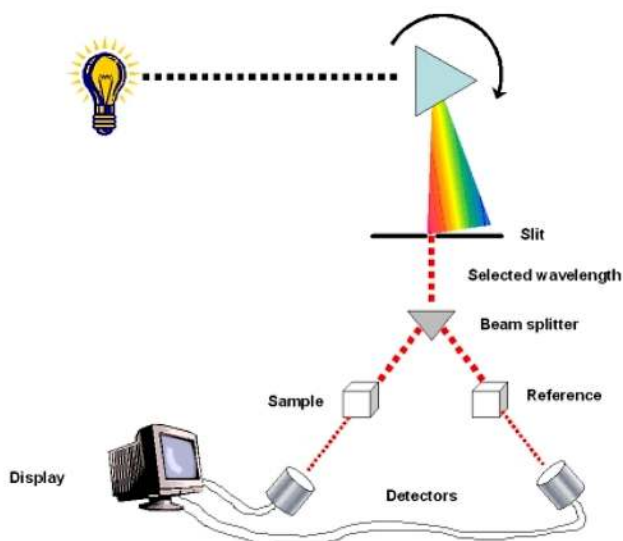


Figure 5.1.7: Dual beam spectrometer

Both the single and dual beam detectors allow the investigator to monitor the absorbance properties for a **single wavelength** of light (the wavelength selected by the prism/slit settings). A modification that incorporates an **array of detectors** can permit simultaneous measurement of a spectrum of wavelengths:

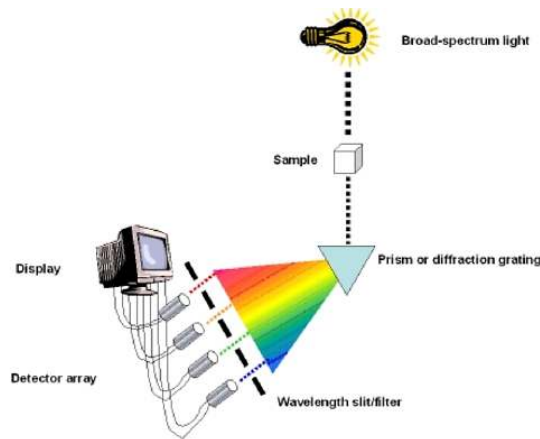


Figure 5.1.8: Simultaneous measurement of a spectrum

Absorbance spectra of biological molecules

Proteins

Proteins do not absorb in the visible wavelength unless they have a prosthetic group (e.g. Fe^{2+}) or an unnatural amino acid. However, the amino acids tryptophan, tyrosine and cysteine absorb light in the UV wavelength:

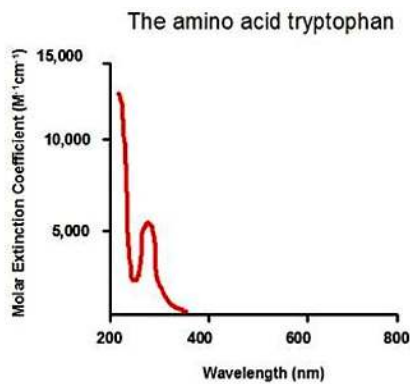


Figure 5.1.9: Tryptophan absorption

- Tryptophan has a peak of absorption at 280nm in the UV range
- This is a useful wavelength to quantitate the absorption of tryptophan
- Since the absorption is proportional to concentration, this is a useful way to quantitates protein concentration (for proteins containing Trp)

Nucleic acids

The aromatic rings in the bases of nucleic acids also absorb in the UV range:

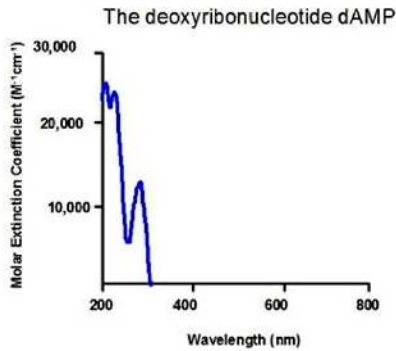


Figure 5.1.10: Nucleic acid absorption

- Each DNA and RNA base has a slightly different absorption spectrum
- 260 or 280nm is a typically useful wavelength to monitor concentration of nucleic acids

Note that samples of nucleic acids and proteins can both absorb at 280nm, therefore, samples of biological molecules should be **pure** in order to quantitate using UV absorption spectroscopy.

This page titled [5.2: Spectrophotometry](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

5.3: Quantification of Protein Concentration

Review: the Beer-Lambert Law

$$\log(1/T) = \epsilon cl = A$$

$$T = 1/10^A$$

Values for ϵ and l

- Path length l is usually in units of **cm**. (note: most spectrophotometers are designed to accept 1cm wide cuvettes)
- Molar extinction coefficient ϵ has units of $M^{-1} \text{ cm}^{-1}$ and is a constant of proportionality that relates the absorption of **molar** solutions
- Mass extinction coefficient $\epsilon^{1\%}$ refers to the absorbance of a 1% by mass solution. Typically this refers to an aqueous solution that we can take to have a density of 1000g/L. A 1% by mass aqueous solution would therefore refer to the dissolution of 10g/L, or a **10mg/ml** solution of the molecule of interest.
- Since the absorbance of a molecule is a function of the wavelength (i.e. the absorption is not equal for every wavelength) the extinction coefficient must also reference a wavelength. This is typically done using a subscript:

$$\epsilon^{1\%}_{280\text{nm}} = 14.5 \text{ g}^{-1} \text{ L cm}^{-1}$$

• In this case a **10mg/ml solution of the molecule will have an absorbance reading of 14.5** (dimensionless units) at $l = 280\text{nm}$ (the absorption at other wavelengths may not be known). The units of concentration are g/L, thus ϵ will have dimensions of $\text{g}^{-1} \text{ L cm}^{-1}$.

Why is it important to be able to quantitate protein concentration in a sample?

An important application of "Biotechnology" is the production of proteins as commercial products. Such products might have pharmaceutical applications (e.g. insulin, human growth hormone, tissue plasminogen activator, erythropoietin, blood clotting factor VIII.), industrial applications (e.g. subtilisin (an enzyme in detergents), 2,5-diketo-D-gluconate reductase (an enzyme in vitamin C production), as materials (e.g. silk protein in textiles, barnacle adhesion protein as a glue). In these cases, there are various aspects of successful production that require quantitation:

- How much of the protein can be produced (i.e. what is the efficiency of production)?
- How pure is the protein that is produced (industrial applications may require 90% pure, pharmaceutical applications may require 99.999% pure)

Such proteins may be isolated from natural sources (e.g. blood clotting factor VIII may be extracted from human blood), or they may be produced recombinantly (e.g. E. coli bacterial cells can be genetically engineered to produce human growth hormone). In both cases, it may be necessary to purify the protein using a series of fractionation steps. We will go into more detail about such fractionation steps in a later lecture, but the general idea is that a heterogeneous mixture of molecules can be fractionated based upon some physical property of the molecules. The following are properties that can be used to fractionate a heterogeneous mixture of biomolecules:

- Molecular mass (i.e. "big" molecules can be separated from "small" molecules)
- pKa (i.e. "acidic" molecules can be separated from "basic" molecules)
- Hydrophobicity (i.e. non-polar molecules can be separated from polar molecules)

For such fractionation steps involving proteins, we need to keep track of how much of the contaminating proteins went into one fraction and how much of our desired protein went into the other fraction. Although the details are somewhat more complicated than this simple description, **it is important to be able to quantitate protein concentration to be able to effectively purify a protein of interest**.

Once a protein is pure, it may be of considerable economic interest to be able to quantify the yield (and, therefore, be able to determine how much it cost to produce a given mass of protein). For example, the only source for human growth hormone (to treat small stature) used to be to extract it from human pituitary glands harvested from the brains of cadavers. Suffice it to say, this made the protein extremely expensive. Furthermore, the isolation from human tissues meant that the sample could also be potentially contaminated with human pathogens (hepatitis, CJD, AIDS, etc.). With the advent of genetic engineering, the production of human

growth hormone by bacterial cells (i.e. *E. coli*) meant that relative large quantities could be produced far cheaper (and with no threat of human pathogens).

Why not just weigh the protein?

- Most samples are typically quantities of milligrams or even micrograms, not grams, and thus, it is difficult to transfer and measure such small amounts
- Water is present in proteins, and it is extremely difficult to remove all the water (some water molecules hydrogen bond extremely tightly to proteins). Thus, the mass measurement would include some waters, and would increase the apparent mass of the protein

Absorbance spectra of biological molecules

Proteins

Proteins do not absorb in the visible wavelength unless they have a prosthetic group (e.g. Fe^{2+}) or an unnatural amino acid. However, the amino acids tryptophan, tyrosine and cysteine absorb light in the UV wavelength:

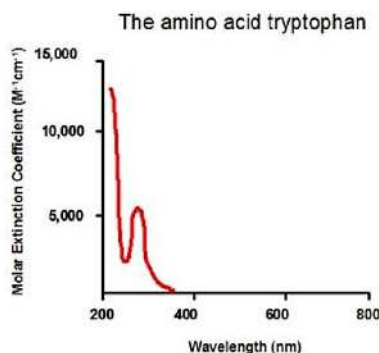


Figure 5.3.1: Tryptophan absorption

- Tryptophan has a peak of absorption at 280nm in the UV range
- This is a useful wavelength to quantitate the absorption of tryptophan
- Since the absorption is proportional to concentration, this is a useful way to quantitates protein concentration (for proteins containing Trp)

Nucleic acids

The aromatic rings in the bases of nucleic acids also absorb in the UV range:

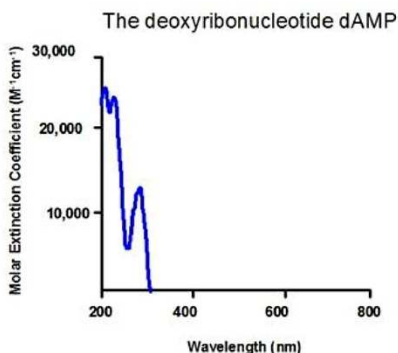


Figure 5.3.2: dAMP Absorption

- Each DNA and RNA base has a slightly different absorption spectrum
- 260 or 280nm is a typically useful wavelength to monitor concentration of nucleic acids

Note that samples of nucleic acids and proteins can both absorb at 280nm, therefore, samples of biological molecules should be **pure** in order to quantitate using UV absorption spectroscopy (any contaminating nucleic acids in a protein sample will increase the apparent absorbance, likewise for contaminating proteins in a nucleic acid sample).

Important aspects of quantification of proteins using UV absorbance

- If a protein contains Trp, Tyr or Cys residues it will absorb in the UV. If it does not contain these amino acids, it will not absorb UV light, and we cannot quantify it using this method
- Multiple Trp, Tyr or Cys residues will contribute to the Extinction coefficient for the protein. Thus, we need to know how many of these residues are present in the protein to know the correct extinction coefficient
- Nucleic acids (DNA, RNA) contaminant will also absorb UV light, as will other proteins with Trp, Tyr and Cys residues. Thus, the sample must be PURE to use UV absorption to quantify a protein

Molar extinction coefficients of Trp, Tyr and Cys amino acids:

Amino Acid	$E_{280\text{nm}}$ ($\text{M}^{-1} \text{cm}^{-1}$)
Trp	5690
Tyr	1280
Cys	120

Example 5.3.1: Bovine insulin

Bovine insulin contains 4 Tyr residues, 6 Cys residues and 0 Trp residues. We can determine the expected molar extinction coefficient at 280nm, $E_{280\text{nm}}$, by the following calculation:

$$E_{280\text{nm}} = (0)(5690) + (4)(1280) + (6)(120)$$

$$E_{280\text{nm}} = 5840 \text{ M}^{-1} \text{ cm}^{-1}$$

Thus, a 1.0M solution of pure bovine insulin would give an absorbance of 5,840 at 280nm (obviously, it would have to be diluted considerably to be read accurately).

A useful expression relating the parameters of E, concentration (C) and A are derived from the Beer-Lambert law (assuming 1cm path length):

$$A/E = C$$

For example, if a sample of bovine insulin was observed to give an absorbance at 280nm of 0.745 we could calculate the concentration to be:

$$0.745/5840 \text{ M}^{-1} \text{ cm}^{-1} = C$$

$$C = 1.28 \times 10^{-4} \text{ M (note: cm}^{-1} \text{ drops out with 1cm pathlength)}$$

It should be noted that a deuterium lamp is required for UV spectrophotometry, as well as quartz cuvettes (since glass absorbs UV light)

Colorimetric (chromogenic) methods of protein concentration determination

Chromo means color and genesis mean creation, so chromogenic means the "creation of color". Color means color (duh...) and metric means to measure, so colorimetric is to "measure color". Both terms refer to the same sort of thing in the present case - we can modify the protein sample with appropriate reagents so as to produce a color reaction (in visible spectrum) and measure protein concentration using a VIS spectrophotometer. Advantages are:

- Cheap lamp! (tungsten light bulb versus deuterium for UV)
- Cheap cuvette! (cheap glass or plastic versus quartz)
- Not contaminating absorbance from proteins or nucleic acids! (no absorption in VIS spectrum)

We will consider three methods: The **Biuret**, **Lowry** and **Bradford** methods of colorimetric determination of proteins.

Biuret

Under high pH (alkaline) conditions the copper II ion (Cu^{2+}) is believed to form a complex with peptide nitrogens of proteins:

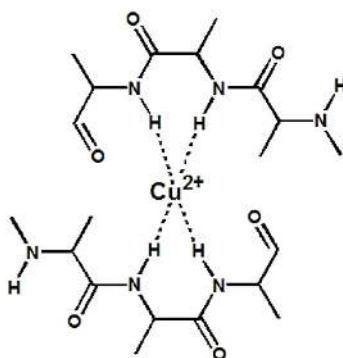


Figure 5.3.3: Cu^{2+} complex

This complex absorbs light at 550nm and has the following useful properties:

- It is dependent upon at least a dipeptide structure (see above), thus, contaminating amino acids will not contribute to the 550nm absorption
- The binding depends upon the peptide backbone nitrogen and not the side chain functional group. Thus, the binding is independent of sequence
- Nucleic acids do not interfere since they don't share the peptide backbone structure
- However, ammonia and certain amines can interfere. Thus, ammonium sulfate salts, and the common biological buffer TRIS (tris hydroxymethyl amino ethane) will provide a false positive and cannot be present in the sample
- Also, the absorption is relatively weak, thus, the method is somewhat insensitive and requires a relatively high concentration of protein

Lowry

The Lowry method is a modification of the Biuret method. After treatment with Copper II, the protein is treated with phosphomolybdotungstate mixed acids (acidic compounds of molybdenum and tungsten ions). These acids are known as the **Folin-Ciocalteu (or just Folin) reagent**. The Folin reagent is added under alkaline conditions, and the Folin reagent is subsequently reduced by the the Copper ions as well as Tyr, Trp and polar amino acid side chains. The product of this reaction is **heteropolymolybdenum blue**, which absorbs strongly at 750nm. This assay has the following properties:

- More sensitive than the Biuret assay (can detect lower concentrations of protein)
- Somewhat dependent upon amino acid composition (i.e. relative concentrations of Tyr, Trp and polar amino acids)
- Absorption reaction is linearly dependent upon protein concentration, but only at low concentrations of protein (i.e. the standard curve and assay must be performed at a low concentration regime).
- Sensitive to contaminants as with the Biuret method, as well as others related to the Folin reagent and redox reactions.
- More critical to timing and precision of person doing the assay

Bradford

A dye known as Coomassie Brilliant Blue was developed by the textile industry. It was noticed to stain skin as well as the textiles. Thus, this dye (which normally absorbs at 465nm) was known to bind to proteins and to absorb strongly at 595nm. The dye forms a wide variety of strong, but non-covalent, interactions including hydrogen bonding donor and acceptor interactions as well as hydrophobic (non-polar) interactions.

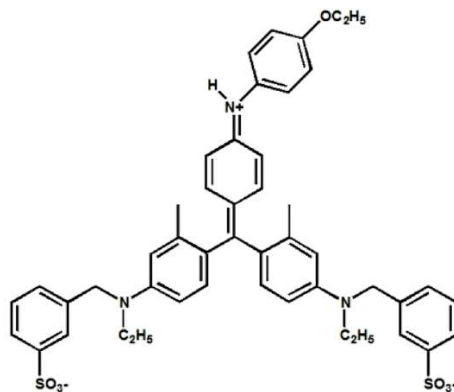


Figure 5.3.4: Coomassie Brilliant Blue binding

The method is quite simple: a single step in which the dye is added to the protein solution under acidic conditions, and then the absorbance is read at 595nm. However, the method has the following properties:

- The response is generally independent of the amino acid composition
- The assay is sensitive, but somewhat non-linear. Thus, a standard curve must always be performed (using known concentrations of pure protein)
- The dye stains pretty much everything, including cuvettes, floors, countertops. It can be a real mess if spilled (I know this from personal experience). And since it is a textile dye, if you get it on your clothes, you will need to learn to like blue polka dots.

This page titled [5.3: Quantification of Protein Concentration](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

5.4: Chromatography

"Chroma" refers to color and "graphy" refers to writing.

Chromatography involves the **physical separation of a mixture of compounds**, where historically the identification of the individual compounds is by their unique color. Chromatography can be used as a purification method (preparative scale) and also for the identification of compounds based on their chromatographic behavior (analytical scale).

There are many types of chromatography, but all involve dissolving the compounds to be separated in a **liquid solution** (the "mobile" phase) and then passing this solution across (or through) some solid support (or "matrix") known as the "stationary" phase. The stationary phase is typically a slurry of solid beads of silica or polysaccharide compounds, or a solid surface of these compounds. As the mobile phase travels through or across the solid phase the dissolved solutes interact to varying degrees with the stationary phase (i.e non-covalent **attractive** forces). **The stronger the attraction, the more the molecules are retarded** in comparison to the mobile phase. If there is no attraction for the solid support, then the solute molecules move with the mobile phase. The different rates of movement for the different solutes results in their separation from each other.

An example you can do at home involves a coffee filter (stationary phase) and water (mobile phase) in the separation and analysis of the dye in a water soluble marker pen (the mixture of compounds). If you draw a line with a purple marker on a strip of the filter paper, and place the bottom in a dish of water, the water will wick up the paper, and the differential affinity of the dye inks for the solid support (cellulose) results in differential migration. In the example below, the red ink has some affinity for the paper, but the blue ink does not (and migrates with the front edge of the mobile phase).



Figure 5.4.1: Chromatography of purple pen

Chromatography in biochemistry typically utilizes not paper, but beads of polysaccharide (often chemically derivatized) packed into a column, as the solid support. The solid support is often called the chromatography "resin". Some common types of chromatographic resins include:

- Ion exchange
- Affinity
- Hydrophobic
- Gel filtration

Often, the solutes to be separated are proteins, and so the discussion will focus upon protein chromatography (but the principles are the same for any solute)

Ion exchange

Ion exchange resins contain charged groups.

- These may be acidic in nature (in which case the resin is a cation exchanger)
- or basic (in which case it is an anion exchanger).

- Cation and anion exchangers may be broken down further into weak and strong exchangers (reflecting binding affinity).

Type of exchanger	Functional group	Common name
Weak cation exchanger	carboxymethyl (-)	CM cellulose/sephadex
Strong cation exchanger	sulfopropyl (-)	SP sephadex
Weak anion exchanger	diethylaminoethyl (+)	DE cellulose/sephadex
Strong anion exchanger	quaternary amine (+)	QAE sephadex

Usually, samples are **loaded under low ionic strength** conditions (which promotes electrostatic interactions) and bound material is eluted using either a step or gradient elution of buffer with higher ionic strength.

- Generally speaking, a protein will **bind to a cation exchange resin if the buffer pH is lower than the isoelectric point (pI) of the protein, and will bind to an anion exchange resin if the pH is higher than the pI.**
- Knowledge of the pI of the protein is therefore helpful in designing a purification protocol using ion exchange resins (however, you can always simply try different resins to see which works best).

Elution of proteins from ion exchange resins

Proteins bound to ion exchange resins are bound via non-covalent ionic (salt-bridge) interactions. We can compete for these ionic binding sites on the resin with other ionic groups, namely, salts

- There are two general types of methods when eluting with a salt solution: **1. Gradient elution and 2. Step elution**
- A gradient elution refers to a smooth transition of salt concentration (from low to high) in the elution buffer. Weakly binding proteins elute first, and stronger binding proteins elute last (i.e. they require higher salt concentrations in the buffer to compete them off the column)
- A gradient salt concentration can be made using a gradient maker. In its simplest form, this consists of two containers (must be the same shape) connected by a siphon (or tube at the bottom). One container contains the low salt buffer, and the other contains high salt buffer. The buffer is withdrawn from the low salt container:

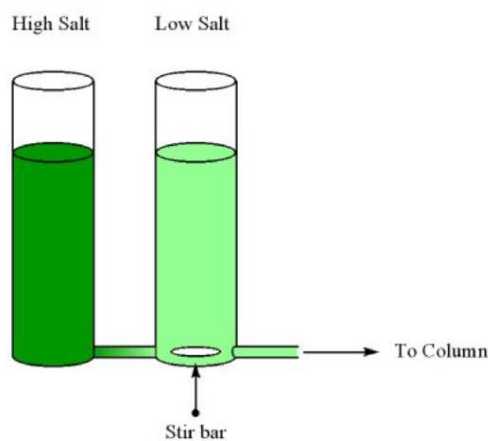


Figure 5.4.2: Gradient maker

- This will produce a linear gradient from low to high salt concentrations over the total volume of the gradient
- If we know the concentration range of salt over which a protein of interest will elute we can simply elute with a buffer containing that concentration of salt. This is known as a step elution.
- Step elutions are generally faster to run, and elute the protein in a smaller overall volume than with gradient elutions. They generally work best when contaminants elute at a significantly different salt concentration than the protein of interest

Note that after ion exchange chromatography the protein of interest will be in a buffer with a potentially high salt concentration. This must be taken into account before proceeding with the next step in the purification scheme

Affinity chromatography

Affinity chromatography is a general term which applies to a wide range of chromatographic media. It can be basically thought of as some inert resin to which has been attached some compound which has a *specific affinity* for your protein of interest.

- Thus, a specific antibody attached to an inert resin would be a type of affinity chromatography.
- Other examples might include: a protease inhibitor attached to some matrix, designed to bind a specific protease
- a cofactor bound to some matrix, designed to bind to a particular enzyme
- a metal ion bound to a matrix, designed to chelate a protein with a metal binding site, and so on.

In each case, the type of resins used and the method of attachment may vary, as will the method of elution. One generalization regarding method of elution is that the **bound ligand can be competed off of the column's functional group by including in the elution buffer a high concentration of the free functional group**. For example, if the functional group of the column is a cofactor, then the bound protein can be competed off the column by passing a buffer containing a high concentration of cofactor (or cofactor analog) through the column.

Other methods of elution include changing the buffer conditions such that the protein is no longer in the native state (since it is the native state which confers the structure required for the specific binding interaction). This can be achieved by changing pH or by adding denaturing agents such as urea or guanidine.

With affinity chromatography, typically the purification achieved in a single step can be **dramatic** - on the order of several thousand fold. Single step purifications with specific affinity columns are not unheard - in fact it is an ideal goal of purification - *a matrix which recognizes only the protein of interest and none other*.

Hydrophobic resins

Hydrophobic resins contain a non-polar functional group, such as an alkane or aromatic group.

- Many proteins are able to sequester such groups on their surface and this exclusion from solvent provides the basis of the binding energy (i.e. the "hydrophobic effect").
- This interaction is enhanced by **increasing ionic strength**, such that proteins may **bind under high salt conditions and elute under low salt conditions**.
- As such these columns may be used to not only provide purification, but to *desalt* samples (for example after an initial ammonium sulfate precipitation).
- It is usually not possible to predict in advance which particular resin will bind a given protein, this is usually determined empirically. However, the longer the alkane, or the larger the aromatic compound, the stronger the binding typically will be.

Due to the nature of hydrophobic interactions and ionic strength, hydrophobic chromatography and ion exchange chromatography can be conveniently used sequentially. For example, after ion exchange the protein is in high salt conditions, thus it can be loaded directly onto a hydrophobic column. Conversely, a hydrophobic column is eluted in low salt, which is a requirement for binding to an ion exchange resin.

A distinction should be noted between hydrophobic interaction chromatography and reverse phase chromatography.

- Hydrophobic interaction chromatography is performed in aqueous solvent conditions and changes in ionic strength are used to elute the column. The protein typically binds in the native state via hydrophobic groups located on the surface of the protein. The native state is retained during the elution conditions
- Reverse phase chromatography utilizes a hydrophobic solvent (typically acetonitrile) and the binding of a ligand is a function of the phase partition between the hydrophobic nature of the solvent and column functional group. Proteins are typically denatured in such solvents and bind due to the hydrophobic nature of the entire polypeptide sequence. Since the majority of hydrophobic groups are located in the core of globular proteins, the binding is related to the denaturation of the protein and the accessibility of these groups to the column functional groups. Proteins can be purified using reverse phase chromatography, but usually must be refolded in some way to regain functionality (i.e. the native state)

Gel filtration

Gel filtration does not rely on any chemical interaction with the protein, rather it is based on a **physical** property of the protein - that being the *effective molecular radius* (which relates to mass for most typical globular proteins).

- Gel filtration resin can be thought of as beads which contain pores of a defined size range.
- Large proteins which cannot enter these pores pass around the *outside* of the beads. Therefore, **the volume of the column appears smaller to a large molecule.**
- Smaller proteins which can enter the pores of the beads have a larger volume that they can explore, **thus the volume of the column appear larger to a small molecule.**
- Both large and small molecules experience **the same flow rate of mobile phase** (i.e. L/min). Thus, a sample of proteins passing through a gel filtration column will *separate based on molecular size: the big ones will elute first and the smallest ones will elute last* (and "middle" sized proteins will elute in the middle).

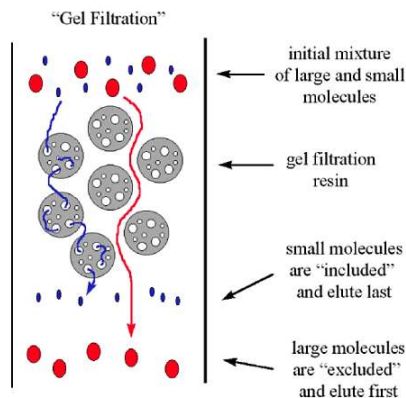


Figure 5.4.3: Gel filtration

- If your protein is unusually "small" or "large" in comparison to contaminating proteins then gel filtration may work quite well.

Where will a protein elute in a gel filtration experiment?

- There are two extremes in the separation profile of a gel filtration column.
- There is a critical molecular mass (large mass) which will be **completely excluded** from the gel filtration beads. All solutes in the sample which are equal to, or larger, than this critical size will behave identically: they will all eluted in the excluded volume of the column
- There is a critical molecular mass (small mass) which will be **completely included** within the pores of the gel filtration beads. All solutes in the sample which are equal to, or smaller, than this critical size will behave identically: they will all eluted in the included volume of the column
- Solutes between these two ranges of molecular mass will elute between the excluded and included volumes

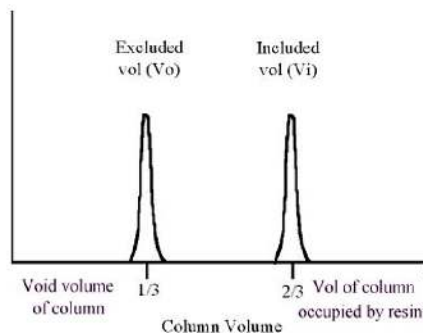


Figure 5.4.4: Protein elution in gel filtration

As a general rule of thumb, the excluded volume (V_o) is approximately equal to one third of the column volume, the included volume is approximately equal to two thirds of the column volume (the "missing" third is taken up by the volume of the resin material).

- In gel filtration **the resolution is a function of column length** (the longer the better)
- However, one drawback is related to the maximum sample volume which can be loaded. The larger the volume of sample loaded, the more the overlap between separated peaks. Generally speaking, the sample size one can load is limited to about **3-5% of the total column volume**.
- Thus, gel filtration is best saved for *the end stages of a purification*, when the sample can be readily concentrated to a small volume.
- Gel filtration can also be used to remove salts from the sample, due to its ability to separate "small" from "large" components.
- Finally, gel filtration can be among the most "gentle" purification methods due to the lack of chemical interaction with the resin.

This page titled [5.4: Chromatography](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

5.5: Gel Electrophoresis of Proteins

Gel electrophoresis is used to characterize one of the most basic properties - **molecular mass** - of both polynucleotides and polypeptides. Here we will focus exclusively on gel electrophoresis of proteins

Gel electrophoresis can be used to determine:

- the *purity* of a protein sample
- *heterogeneity* and extent of degradation of a protein sample
- *subunit composition* of a protein sample

How does it work?

The underlying principle of electrophoresis is the **migration property of charged species within an electric field**. Thus, it is the simple behavior of opposite charges attracting. An electric field is established across the electrodes of a power supply, and charged ions move in this electric field. Note that in such an apparatus, the word "ANODE" refers to the positively charged electrode (the ANODE attracts ANIONS) and the word "CATHODE" refers to the negatively charged electrode (the CATHODE attracts CATIONS). The cathode and anode terms are consistent with their redox reaction definitions in that reduction is subsequently occurring at the cathode and oxidation is occurring at the anode:

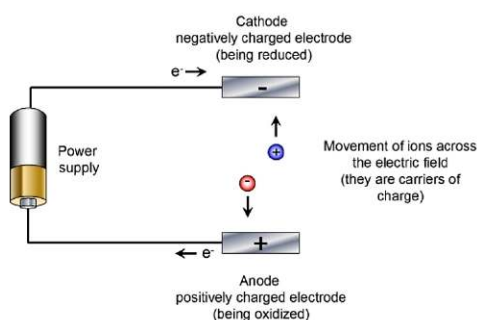


Figure 5.5.1: Anode and cathode

Warning: read the following note only if you are curious about batteries and labeling of electrodes, otherwise, don't bother (it may only be confusing).

Note

Redox chemistry within the external power supply is driving the redox reaction at the electrodes. Note that the electrons from the anode are going to the "+" terminal of the battery. Reduction at this electrode of the battery must be occurring and is driving the oxidation at the anode of the external electrodes. If this terminal of the battery is being reduced, then it must be the cathode in the redox reaction within the battery. Thus, batteries have their cathode labeled as "+" and anode labeled as "-". Took me hours to figure this out.

Proteins are comprised of the 20 common amino acids, which include both negatively charged (i.e. acidic) side chains (e.g. aspartic acid, glutamic acid) and positively charged (i.e. basic) side chains (e.g. histidine, lysine and arginine). Thus, proteins can be charged, and will migrate in an electric field.

At this point there are a couple of things to consider:

1) Any such separation is a non-equilibrium process. By this, we mean that if we let the process continue on until some equilibrium condition is met, all the anions will be on one electrode and all the cations will be on the other. It would be better to halt the separation process at some intermediate time point to permit achieve separation:

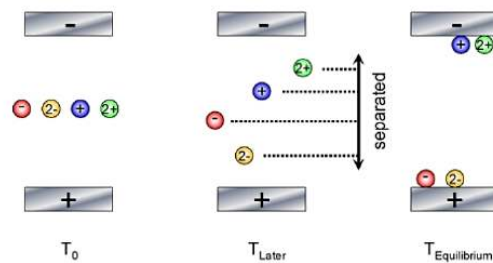


Figure 5.5.2: Separation process

2) The other problem is that once the electric field is switched off, diffusion will cause the separated ions to move around (i.e. we will lose the separation we have tried to achieve). To solve this problem, the separation is not performed in solution, but within a matrix (i.e. a molecular mesh or network). The matrix provides a **frictional component** that resists diffusion. Furthermore, the friction of the matrix is an important factor in the rate of migration of the ions.

3) Note also that separation is achieved by initial application of the sample within a narrow zone (band). If the sample is initially dispersed, although the ions will move, they won't be neatly separated.

Factors that influence the rate of electrophoresis migration (R_f)

Three factors affect the rate of migration:

- Strength of the electric field, E (directly proportional to migration rate)
- Charge on the ionic species, q (directly proportional to migration rate)
- Frictional coefficient of the support matrix, f (inversely proportional to migration rate)

$$R_f \propto qE/f$$

These factors can be varied in the following way:

- The strength of the field is a function of the **voltage** of the power supply. Thus, we can vary the voltage directly. In a related issue, the voltage is proportional to the resistance across the electrodes. Current comes into play here also, but in short, it is difficult to achieve high voltage across the electrodes if the resistance is low. The resistance of a solution is **inversely proportional to the ionic strength** (i.e. concentration of ions). Thus, with high salt concentrations, the resistance is low, it is difficult to achieve high voltage, and the migration rate will decrease.
- The charge on an ionic molecule is a function of the **pI of the molecule and the pH of the solution**

If $pI > pH$ the molecule is cationic
(migrates towards cathode)

If $pI < pH$ the molecule is anionic
(migrates towards anode)

If $pI = pH$ the molecule is neutral
(no migration in electric field)

We can therefore alter the migration rate (and possibly the direction of migration) by altering the pH of the solution

- The frictional coefficient of the matrix can potentially be altered. The matrix is typically a polymer network (see below), referred to as a **gel**, and the frictional coefficient can be increased by increasing the polymer concentration

The gel matrix for gel electrophoresis of proteins

Gel electrophoresis of proteins almost exclusively utilizes **polyacrylamide**. This is a polymer comprised of two covalently-linked components:

- acrylamide
- bis acrylamide.

The bis acrylamide is essentially a **cross-linking** component of the acrylamide polymer. A typical value for the acrylamide:bisratio is 19:1 and the **total** acrylamide concentration in the gel affects the migration of proteins through the matrix (i.e. determines the frictional coefficient).

- **High molecular mass proteins** are separated using low frictional coefficient (i.e. **low concentrations**) of polyacrylamide.
- **Low molecular mass proteins** are separated using high frictional coefficient (i.e. **high concentrations**) of polyacrylamide.
- Although we now have everything in place to perform a gel electrophoresis experiment to separate proteins, there is another consideration that often is undesirable (although sometimes useful). This situation is the fact that different proteins have different charges for a given value of pH. Thus, proteins will migrate as a function of **both** their **mass** (big ones move slowly) and their **overall net charge**.
- If there were some way to cause each protein to have an **identical** charge to mass ratio, we could separate a mixture of proteins **based only upon mass effects**

The role of sodium dodecyl sulfate detergent in polyacrylamide gel electrophoresis

Sodium dodecyl sulfate (**SDS**; also known as "laurel sulfate") is an ionic detergent with the following structure:

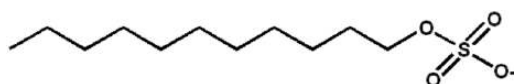


Figure 5.5.3: Sodium laurel sulfate structure

- The SDS binds, via hydrophobic interactions, to the proteins in a stoichiometry **approximately proportional to the size of the protein** (i.e. a small protein will bind a few molecules, and a large protein will bind a lot of molecules of SDS)
- Due to the charged nature of the SDS molecule the proteins thus will have an approximate **constant charge to mass ratio** due to the charge provided by the SDS, and will migrate through the gel at a rate proportional to their molecular mass
- The proteins migrate towards the **anode** since the charge on the SDS is negative under all pH conditions except highly acidic.

Protein gels are usually performed under **denaturing conditions**, meaning that the sample preparation involves heating the protein in the presence of SDS to fully unfold the protein and permit binding of SDS throughout the length of the polypeptide. Once SDS has been bound, the characteristic pI values of the proteins is no longer relevant; the protein takes on a negative charge, and each protein has essentially the same charge to mass ratio.

Migration rate, protein mass, and the % acrylamide in the gel

The greater the percent acrylamide in the gel support, the greater the frictional coefficient, and the slower the migration rate. If the proteins to be separated are of a high molecular mass, and if the % gel is high, the proteins may not even enter the gel (due to overwhelming friction). Thus, it is **essential to match the % gel to the mass of the proteins being separated**. The following table provides a general guideline:

Acrylamide	Range of separation of Polypeptides (length in amino acids)
8%	25-200 kDa (225-1800 a.a.)
10%	15-100 kDa (135-900 a.a.)
12.5%	10-70 kDa (90-630 a.a.)
15%	6-60 kDa (55-550 a.a.)
20%	4-40 kDa (36-360)

In setting up the SDS PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) experiment we **need to know when to stop the experiment** (since it is not an equilibrium process). This is somewhat difficult to determine since proteins (even with SDS bound) do not absorb in the visible spectrum (i.e. we cannot simply look at the gel to determine when the proteins have been separated). Therefore, it is common to include in the protein sample a **small anionic dye molecule (e.g. bromophenol blue)**:

- The size of the dye molecule is chosen to be very small so that there is essentially **no friction** coefficient with the gel
- The dye is chosen to be **anionic** so that it will migrate in the same direction as the protein/SDS complexes (i.e. towards the anode)
- Since the dye is anionic, and small, it will **migrate the fastest** of any component in the mixture to be separated
- The dye molecule is also chosen so as to absorb in the **visible** spectrum (and, hence, be visually detectable while the gel is running)
- The protein/SDS/dye mixture is loaded on the top of the gel (i.e. cathode side) and when the dye molecule (the "dye front") **reaches the bottom of the gel, the power is turned off and the experiment halted**

Visualization of the separated proteins

Although the gel support provides some friction to molecular motions, as soon as the power is turned off the separated protein bands will begin to diffuse (they are freely soluble in aqueous solution). To prevent this, the gel is treated with an acetic acid and methanol solution which causes almost all proteins to precipitate (become insoluble). This is called "fixing" the gel. Now the separated proteins will not diffuse.

The fixed proteins are, however, still invisible and must be visualized by staining. A common protein stain is Coomassie Brilliant Blue R-250 (related to the dye used previously in the Bradford assay). The fixed gel is incubated in a solution of "Coomassiestain" and then the stain is washed out of the gel by incubation in a weak solution of acetic acid and methanol. The stain will not bind to the acrylamide, and will wash out (leaving a clear gel). However, it remains strongly bound to the proteins in the gel, and these take on a deep blue color.

Determination of molecular mass

With SDS treatment, the proteins will migrate as a function of their molecular mass. The approximate molecular mass of the separated proteins is therefore a function of their **migration distance**. If a series of proteins with different and known molecular masses is analyzed on an identical gel, under identical conditions, then a standard curve can be established that can be used to determine the molecular mass of an unknown protein. A typical "molecular weight standard" includes the following mixture of proteins:

Protein	Molecular Mass (kDa)
Phosphorylase B	94
Bovine Serum Albumin	67
Ovalbumin	43
Carbonic Anhydrase	30
Soybean Trypsin Inhibitor	20.1
α -Lactalbumin	14.4

The molecular mass is quantified as follows:

- Measure (in cm) the migration distance of the **dye front**
- Measure (in cm) the migration distance of the **protein** (will always be less than the dye front)
- Divide the protein migration distance by the dye front distance to get the **relative mobility** value (always < 1)
- Plot the relative mobility value (along x-axis) versus the **log** of the molecular mass (along y-axis)
- The relationship between the relative mobility and the log of the molecular mass should be a linear function - thus, providing a standard curve against which the molecular mass of an unknown protein can be determined from its relative mobility

Here is an example of a molecular mass standard, and two unknown samples, analyzed on a 10% SDS PAGE stained with Coomassie:

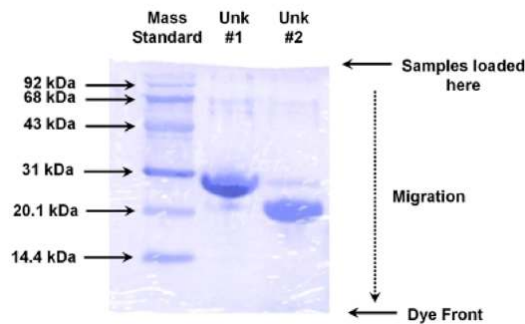


Figure 5.5.4: Example gel electrophoresis

The calculation of migration distances, relative mobility and relationship to Log of the molecular mass yields the following:

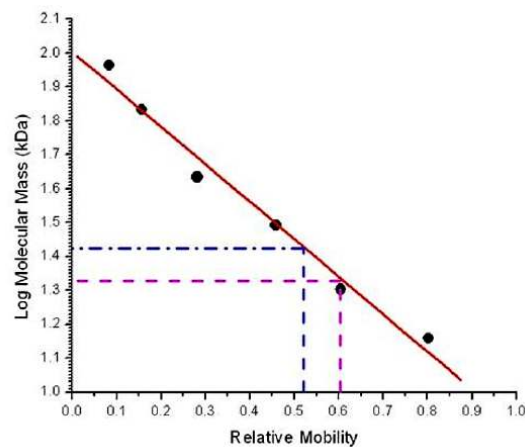


Figure 5.5.5: Relative mobility and molecular mass

- Unknown #1 has a value of 1.43 for the log of the molecular mass, or a mass of ~26.9 kDa
- Unknown #2 has a value of 1.33 for the log of the molecular mass, or a mass of ~21.4 kDa
- A visual check of the unknowns versus the standards in the above gel indicate the calculations appear to be correct

There are other variations of SDS PAGE including:

- Discontinuous gels (i.e. a composite of one gel on top of another) to achieve "focusing" of protein bands (i.e. sharper bands, yielding greater sensitivity and resolution)
- Gradient gels (instead of pouring a single % polyacrylamide gel, pour a gradient with a high concentration at the bottom, and a lower concentration towards the top. This permits resolution of a wider range of molecular mass samples)
- Silver staining (a more sensitive type of stain in comparison to Coomassie - allowing detection of lower concentrations of proteins)
- Transfer of resolved protein bands to a secondary support (e.g. nitrocellulose) for probing with other reagents (i.e. antibodies)
- PAGE (no SDS) of native proteins (to allow detection of non-covalent complexes of two or more proteins - since SDS will disrupt such complexes)
- SDS PAGE with the addition of reducing agents (e.g. β -mercaptoethanol, dithiothreitol, etc.). These reagents will reduce disulfide bonds and separate polypeptide chains that are connected by such bonds

However, the above information covers the basics.

This page titled [5.5: Gel Electrophoresis of Proteins](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

5.6: Introduction to Protein Purification

A successful protein purification procedure can be nothing short of amazing. Whether you are starting off with a recombinant protein which is produced in *E. coli*, or trying to isolate a protein from some animal or plant tissue, you are typically starting with a complex mixture of proteins, nucleic acids, polysaccharides, lipids, etc. from which you may have to extract milligram (or microgram!) quantities of the desired protein, often with high purity, and hopefully with high yield.

Protein purification can be thought of as a series of **fractionation steps** designed so that:

- The protein of interest is found almost exclusively in one fraction (and with good yield)
- A significant amount of the contaminants can be found in a different fraction

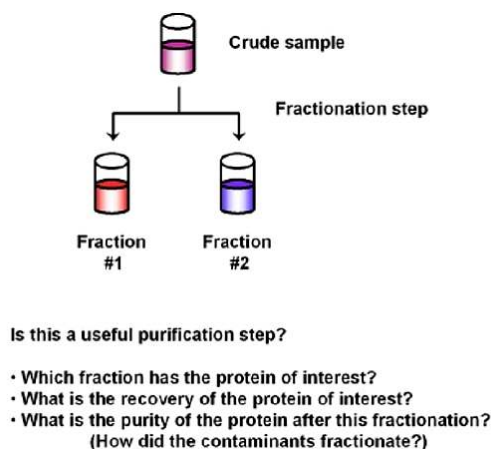


Figure 5.6.1: Fractionation

The first step in any purification is the development of an assay for the protein of interest

The assay can be based upon some **unique characteristic** of the protein of interest, possibly involving:

- Enzymatic activity
- Immunological activity
- Physical characteristics (e.g. molecular mass, spectroscopic properties, etc.)
- Biological activity
- Ideally, an assay should be
 - **specific** (you don't want a false positive or false negative)
 - **rapid** (you don't want to wait a week for the results)
 - **sensitive** (you don't want to consume all your sample in order to assay it)
 - **quantitative** (you need an accurate way to measure the quantity of your protein at each step in the purification)

During purification you will **need to keep track of several parameters**, including:

1. *The total volume of your sample*
2. *The total amount of protein in your sample*

• can be estimated by using $E^{1\%}_{280\text{nm}} = 14.5$

• In other words, measure absorbance at 280nm, divide by 1.4, and you will have the approximate protein concentration in mg/ml

3. *The total amount of the protein of interest (the one you are trying to purify). This information will be determined from your quantitative assay*

This basic information will allow you to keep track of the following parameters during each step of purification:

1. *% yield* for each purification step
2. *Specific activity* of the desired protein ("units of activity" of desired protein/mg total protein)

3. Purification enhancement of each step (e.g. "3.5x purification)

In designing a purification scheme you typically have to balance *purification* with *yield*.

- For example, it may be relatively straightforward to obtain 90% pure material with good yield.
- However, it may be difficult to improve that purity by an additional few percentile and still maintain a good yield.
- **The planned application of the purified protein determines the target purity.**
- If the protein is to be used to determine amino acid sequence information, maybe 90% is acceptable. However, if the material is to be used in clinical trials, 99.999+% may be the target purity.

Initial steps in purification

- Typically (but not always) the protein of interest is contained within the cytoplasm of a cell (sometimes, however, it might be secreted by the cell into the extracellular environment). Thus, cells must be broken open to release the cytoplasm, a process known as cell lysis.
- It is extremely helpful to have some information not only on the general physical and chemical characteristics of the protein you are trying to purify, but also on the *contaminating* components.
- For example, many *E. coli* proteins are generally *low molecular weight* (<50,000 Da) and somewhat *acidic* in isoelectric point

A typical general outline of a protein purification procedure might look like this:

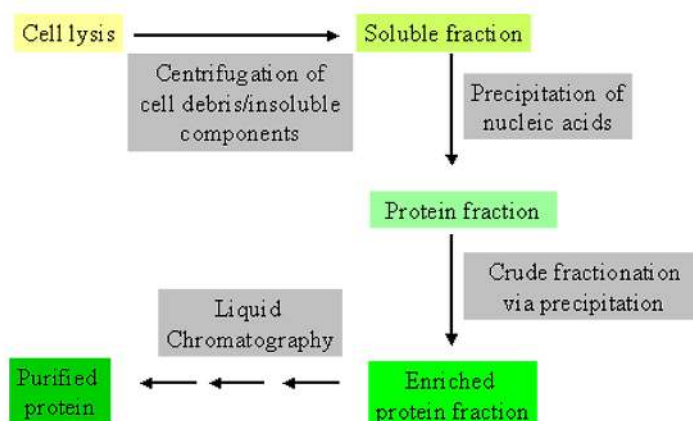


Figure 5.6.2: Protein purification steps

Usually the initial steps in purification make use of general physical and/or chemical differences between soluble proteins and other cell components.

- For example, soluble proteins can be separated from general cellular debris, and intact cells, by *centrifugation*.
- Thus, cells are physically disrupted (via homogenization or a cell press) to allow release of cell contents. This is then followed by centrifugation to separate generally soluble components from those which are insoluble.
- It is at this point that data collection begins in order to monitor the purification.

Nucleic acids can sometimes be readily removed from the sample by the addition of large cationic compounds such as *polyethylene imine*, or *streptomycin sulfate*.

- The nucleic acids bind to these compounds via electrostatic interactions and the complex precipitates and can be removed via centrifugation.
- The same general result can be obtained by mixing in ion exchange resins which are *anion exchangers* (i.e. the resins contain cationic groups) and then filtering or centrifuging to remove. As with either method, it should be confirmed that the desired protein is *not* bound as well.

Crude fractionations of proteins can be achieved by adding various quantities of precipitants such as *ammonium sulfate*, or *polyethylene glycol* (PEG).

- For this type of purification step an initial experiment is performed to monitor the fraction of overall protein, as well as desired protein, remaining in solution (and pellet) as a function of precipitant concentration.

Ammonium Sulfate (% saturated)	0	1	2	3	4	5	6	7	8	90
Sample A ₂₈₀ (in solution)	1000	900	600	200	100	75	50	40	25	20
Activity assay(units) (in solution)	200	200	200	190	170	150	300	500	0	0

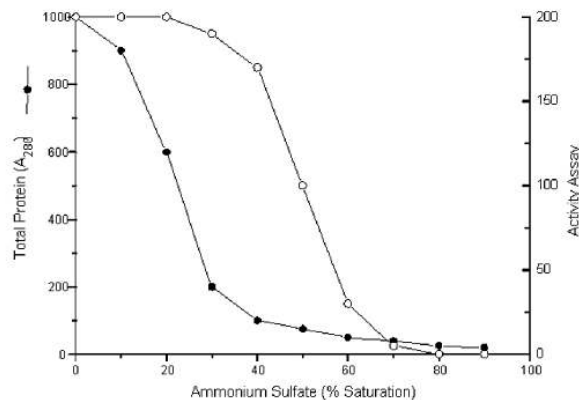


Figure 5.6.3: Protein concentration and precipitant concentration

How can we use this information to utilize ammonium sulfate precipitation as a useful purification step in our overall purification procedure?

Let's begin by calculating the **specific activity** (a measure of purity) of the **starting sample**:

Specific activity = total units of desired protein / mg of total protein

To calculate these parameters we need to know the total sample volume (in addition to the information in the above table). In this case, the total sample volume is 1 L. Knowing this, and the absorbance, we can determine the total protein in the sample:

Sample A₂₈₀/1.4 » mg/ml of sample

1000/1.4 = 714 mg/ml

714 mg/ml * 1.0L(1000ml/L) = **714,000 mg total protein**

The protein of interest was assayed at 200 units present in the starting sample. Therefore, the specific activity of the starting sample is:

Specific activity = 200 units / 714,000 mg = **2.80 x 10⁻⁴ units/mg**

Note

Note: the "units" of activity depend upon the protein of interest. It could be a measure of enzymatic activity - i.e. 1.0 unit of enzymatic activity results in 1mmol of substrate being consumed in 1 min time.

We can now expand upon the above table to include the specific activity of the sample

Ammonium Sulfate (% saturated)	0	10	20	30	40	50	60	70	80	90
Sample A ₂₈₀ (in solution)	1000	900	600	200	100	75	50	40	25	20
Total mg protein (in solution)	714,000	643,000	429,000	143,000	71,400	53,600	35,700	28,600	17,900	14,300
Activity assay(units) (in solution)	200	200	200	190	170	100	30	5	0	0
Specific activity (units/mg)	2.80×10^{-4}	3.11×10^{-4}	4.66×10^{-4}	13.3×10^{-4}	23.9×10^{-4}	18.7×10^{-4}	8.40×10^{-4}	1.75×10^{-4}	0	0

We can plot the specific activity of the sample remaining in solution for the different concentrations of added ammonium sulfate:

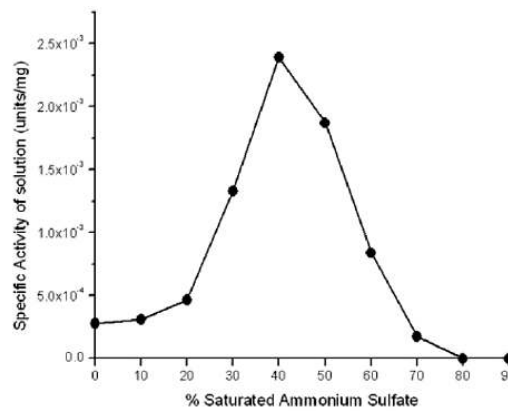


Figure 5.6.4: Specific activity and precipitant concentration

- The highest specific activity is observed in solution when ammonium sulfate is added to 40% saturation
- What is the increase in purity for the 40% solution sample?

Purification = final specific activity / initial specific activity

Purification = $(23.9 \times 10^{-4} \text{ units/mg}) / (2.80 \times 10^{-4} \text{ units/mg}) = 8.5\text{-fold purification}$

However, purification is not the entire story. We also need to consider yield:

Yield (%) = (final total units / initial total units) x 100

And we can add this information our purification statistics:

Ammonium Sulfate (% saturated)	0	10	20	30	40	50	60	70	80	90
Sample A_{280} (in solution)	1000	900	600	200	100	75	50	40	25	20
Total mg protein (in solution)	714,000	643,000	429,000	143,000	71,400	53,600	35,700	28,600	17,900	14,300
Activity assay(units) (in solution)	200	200	200	190	170	100	30	5	0	0
Specific activity (units/mg)	2.80×10^{-4}	3.11×10^{-4}	4.66×10^{-4}	13.3×10^{-4}	23.9×10^{-4}	18.7×10^{-4}	8.40×10^{-4}	1.75×10^{-4}	0	0
Purification	N/A	1.1	1.7	4.8	8.5	6.7	3.0	0.6	-	-
Yield (%)	100	100	100	95	85	50	15	2.5	0	0

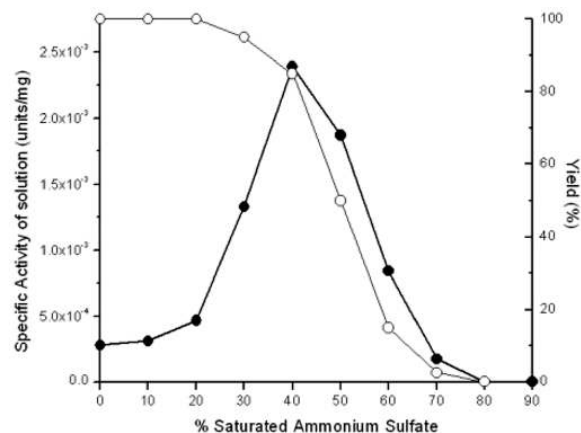


Figure 5.6.5: Percent yield and precipitant concentration

Information on the yield shows that our protein of interest begins to precipitate around 30% ammonium sulfate. Thus, while the addition of ammonium sulfate to 40% provides the greatest purification, it is associated with a 15% loss of our protein of interest. Another option is to add ammonium sulfate to 30%. In this case, we would realize a purification of 4.8-fold, and suffer losses of only 5%.

At each step of a purification scheme we are typically faced with a decision regarding a tradeoff between purity and yield

- If the starting material is cheap, plentiful, or easy to obtain, **we may choose our fractionation steps so as to maximize purity at the expense of yield**
- If the starting material is expensive, limited, or difficult to get, **we may choose to maximize yield at the expense of purity**
- Again, these decisions are also determined by the intended purpose of the final material, and therefore, the final purity desired.

Thus, in the above experiment, we would choose to add ammonium sulfate to a concentration between 20-30% if we wanted to maximize yield, or perhaps 40% (or slightly higher) if we wanted to maximize the purification.

Practical matters with ammonium sulfate fractionation steps

We would probably like to remove the ammonium sulfate from our sample (perhaps the next purification procedure cannot tolerate high ionic strength buffers). Notice that from the data in the above table it looks like all our protein of interest is precipitated if ammonium sulfate is added to a concentration of 80%. Thus, this we can use this precipitation property to remove our protein from most of the ammonium sulfate:

1. Add ammonium sulfate to our sample to a concentration of 20-40% saturation (the exact percentage being determined by whether we wish to maximize purity or yield)
2. Centrifuge and **discard the pellet**
3. Add ammonium sulfate to 80% saturation to the retained solution
4. Centrifuge and **keep the pellet**. Resuspend the pellet in buffer to solubilize the protein.

This page titled [5.6: Introduction to Protein Purification](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

CHAPTER OVERVIEW

6. Lab Notes Part 2

- 6.1: Genetic Transformation (using bacteria and the pGLO plasmid)
- 6.2: Enzyme kinetics
- 6.3: Ligand binding
- 6.4: Restriction Mapping
- 6.5: Polymerase Chain Reaction (PCR)
- 6.6: Use of PC and internet for biochemical research

This page titled [6. Lab Notes Part 2](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

6.1: Genetic Transformation (using bacteria and the pGLO plasmid)

Genetic transformation is the process by which an organism acquires and expresses a new gene. Genetic engineering is the directed transfer of a gene, or piece of DNA, into a cell (typically a bacteria). Typically the intent is to force the cell to express (produce) the protein that the newly introduced piece of DNA codes for (known as heterologous expression). The organism commonly used for genetic transformation and heterologous expression of human genes/proteins is the single celled bacteria known as *Escherichia coli* (E. coli). This organism has several traits of importance in the laboratory:

- Single cell organism
- Doubling time is **20 minutes** (in rich media) to 1 hour (minimal media)
- Naturally lives in human gut (part of normal intestinal flora). Thus, **normal growth temperature is 37°C** (human body temperature)
- *E. coli* has a **single chromosome** that is a **circular DNA molecule**
- *E. coli* is **gram negative** - has a inner cytosolic membrane, periplasmic space, and outer cell membrane

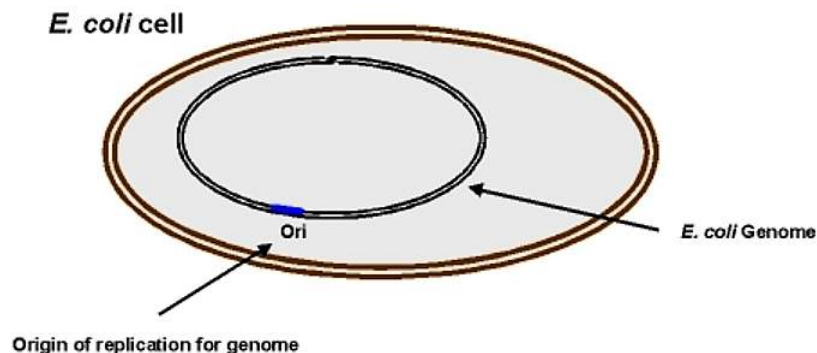


Figure 6.1.1: *E. Coli* genome

Foreign DNA can be introduced into the *E. coli* as a second "chromosome", or as another circular DNA molecule that can **replicate autonomously** due to having its **own origin of replication**. Such "autonomously replicating elements" are referred to as "**plasmids**" or "**vectors**".

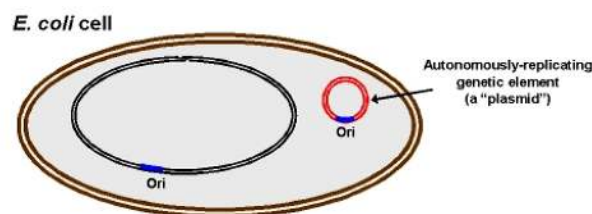


Figure 6.1.2: Plasmid in *E. Coli*

In order to "**stably retain**" the plasmid, there needs to be some type of metabolic reason for the *E. coli* to keep the plasmid around. If the plasmid contains a gene that codes for a protein that protects against antibiotics, then, only cells that have the plasmid will survive in the presence of that antibiotic. Drug resistance can therefore form the basis of a "**selectable marker**" for the presence of the plasmid in a sample of *E. coli*. Ampicillin is an antibiotic for gram negative cells such as *E. coli* (it is part of the penicillin family of antibiotics):

- Production of the outer cell membrane/proteoglycan structure is inhibited in the presence of ampicillin (a lethal event for the bacteria)
- Ampicillin, as a molecule, contains a **b-lactam ring** structure. This ring can be cleaved (and the ampicillin destroyed) by the enzyme b-lactamase
- b-lactamase enzyme is the product of the *bla* gene (genes are typically lower case and italicized)
- If a plasmid contains the *bla* gene, it will confer resistance to ampicillin to the host *E. coli*

- Such *E. coli* grown in the presence of ampicillin will be **selected for**. Under these conditions, any wild type *E. coli* in the sample will be **selected against**)
- Thus, ampicillin resistance is a **selectable marker** for the **plasmid**

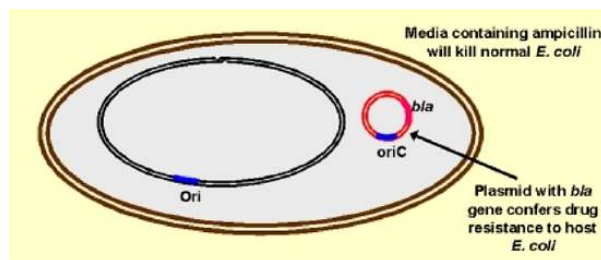


Figure 6.1.3: Ampicillin resistance plasmid

Other genes (that express other proteins) can now be introduced into the plasmid, and the host *E. coli* forced to express the protein of interest

pGLO

Plasmids are typically abbreviated with an acronym that begins with the lower case "p", and the name can provide some information regarding the person that designed the plasmid, or the contents of the plasmid. The pGLO plasmid contains an origin or replication, a selectable marker, and the gene for **Green Fluorescent Protein (GFP)**. The plasmid also contains a gene for the arabinose C protein, which is a protein that regulates expression from the arabinose BAD promoter (P_{BAD}). Promoters are usually indicated with an acronym that begins with an upper case "P".

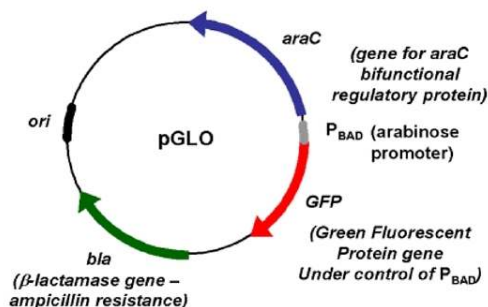


Figure 6.1.4: pGLO Plasmid. The **ori** is the origin of replication for the pGLO plasmid, **bla** is the gene that codes for β -lactamase, and is the selectable drug-resistant marker for the plasmid, **GFP** is the GFP gene and **araC** is the gene coding for the arabinose C protein

A "**promoter**" is a region of DNA that signals RNA polymerase to initiate **transcription** (for production of mRNA). Promoters are typically located at the **start** (5' end) of a gene that codes for a protein (since mRNA production proceeds 5'→3'). The *bla* gene includes a promoter at the 5' end of the gene. This is a weak constitutive promoter (always "on" at a low-level). It will instruct RNA polymerase to continually make a low-level of mRNA for this gene. The mRNA will be translated to produce low-levels of the β -lactamase protein. The GFP is transcribed by the **arabinose P_{BAD} promoter**. This is a **strong** promoter and will cause RNA polymerase to make a large number of copies of mRNA from this gene (and therefore, a lot of GFP protein). However, the arabinose P_{BAD} promoter is **regulated** by the protein coded for by the *araC* gene (which has its own promoter, much like the *bla* gene):

- In the **absence** of the sugar **arabinose**, the *araC* protein binds to the P_{BAD} promoter and **prevents transcription** (of *GFP*)
- In the **presence** of **arabinose** in solution, the *araC* protein binds the arabinose, and this results in a conformational change to the *araC* protein, the result of which is that it now instructs RNA polymerase to make **many copies** of the *GFP* mRNA (and thus, a lot of GFP protein)

How is the pGLO plasmid introduced into the *E. coli* cell?

The general process by which foreign DNA is introduced into a cell is called **transformation**. There are several ways to transform DNA into an *E. coli* cell, but the most common way is by making the cells **competent**. "Competent" cells have the ability to take up DNA molecules from the environment. Normal *E. coli* are **not** competent, however, if they are treated with a solution of **calcium chloride** their cell membranes become competent. Actually, only a small fraction of the cells treated with CaCl_2 are able to take up foreign DNA, however, since the number of cells in a sample is large, the low efficiency of transformation is not much of a problem. The selectable markers ensure that only the cells that did take up the foreign DNA (i.e. plasmid) will survive and grow.

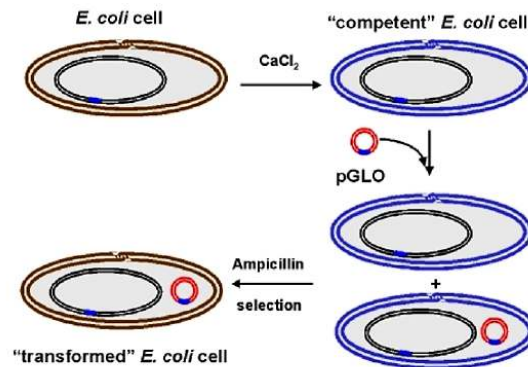


Figure 6.1.5: Transformation of *E. Coli*

Expression of GFP

Addition of arabinose sugar to the growth media will cause RNA polymerase to start transcribing the GFP gene (i.e. making mRNA molecules). The cellular machinery (e.g. ribosomes) will translate this mRNA into corresponding GFP protein.

- GFP will fluoresce green under UV light
- The presence of green light from *E. coli* cells indicates that the transformation (and drug resistance selection) has been successful
- GFP is a gene from a jelly fish and is the reason that some jelly fish glow green. It is not a normal gene for *E. coli*, however, if introduced into *E. coli*, it will make the GFP protein (and fluoresce green)

This page titled [6.1: Genetic Transformation \(using bacteria and the pGLO plasmid\)](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

6.2: Enzyme kinetics

Enzymes are **protein catalysts**, they influence the **kinetics** but **not the thermodynamics** of a reaction

- Increase the rate of a chemical reaction
- Do not alter the equilibrium

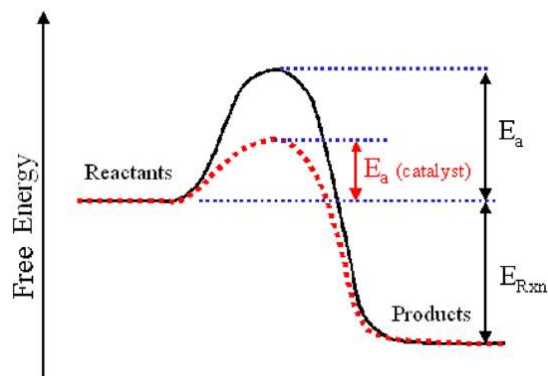
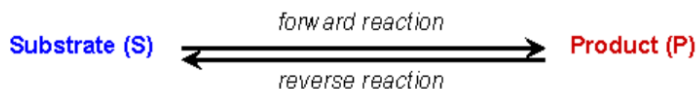


Figure 6.2.1: Catalyst activity

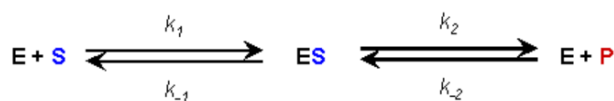
- They increase the rate by stabilizing the transition state (i.e. lowering the energy barrier to forming the transition state (they do not affect the energetics of either the reactant(s) or product(s))

Michaelis-Menten derivation for simple steady-state kinetics

The Michaelis-Menten equation is a **mathematical model** that is **used to analyze simple kinetic data**. The model has certain **assumptions**, and as long as these assumptions are correct, it will accurately model your experimental data. The derivation of the model will highlight these assumptions.

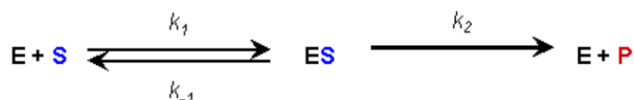


In an enzyme catalyzed reaction the substrate initially forms a reversible complex with the enzyme (i.e. the **enzyme and substrate have to interact** for the enzyme to be able to perform its catalytic function). The standard expression to show this is the following:



ASSUMPTION #1:

- There is no product present at the start of the kinetic analysis
- Therefore, as long as we monitor **initial reaction rates** we can **ignore** the reverse reaction of **E+P going to ES**



ASSUMPTION #2:

- During the reaction an **equilibrium** condition is established for the binding and dissociation of the Enzyme and Substrate (Briggs-Haldane assumption)
- Thus, the **rate of formation** of the ES complex is equal to the **rate of dissociation plus breakdown**

ASSUMPTION #3:

- $[E] \ll [S]$
- The enzyme is a catalyst, it is not destroyed and can be recycled, thus, only small amounts are required
- The amount of S bound to E at any given moment is **small** compared to the amount of free S
- It follows that $[ES] \ll [S]$ and therefore **[S] is constant** during the course of the analysis (NOTE: this assumption requires that the reaction is monitored for a **short period**, so that not much S is consumed and [S] does not effectively change - see next assumption)

ASSUMPTION #4:

- Only the **initial velocity** of the reaction is measured
- $[P] = 0$ (reverse E + P reaction can be ignored)
- $[S] \gg [S]_{\text{initial}}$

ASSUMPTION #5:

- The enzyme is either present as free enzyme or as the ES complex
- $[E]_{\text{total}} = [E] + [ES]$

Michaelis-Menten derivation using above assumptions:

$$\text{Rate of ES formation} = k_1[E][S] + k_{-2}[E][P]$$

Assumption #1 says we can ignore the k_{-2} reaction, therefore:

$$\text{Rate of ES formation} = k_1[E][S]$$

Assumption #5 says $[E] = [E]_{\text{total}} - [ES]$, therefore:

$$\text{Rate of ES formation} = k_1([E]_{\text{total}} - [ES])[S]$$

The rate of ES breakdown is a combination of the dissociation and the conversion to product:

$$\text{Rate of ES breakdown} = k_{-1}[ES] + k_2[ES]$$

$$\text{Rate of ES breakdown} = (k_{-1} + k_2)[ES]$$

Assumption #2 says the rate of ES formation equals the rate of breakdown:

$$k_1([E]_{\text{total}} - [ES])[S] = (k_{-1} + k_2)[ES]$$

Rearrange to define in terms of rate constants:

$$([E]_{\text{total}} - [ES])[S] / [ES] = (k_{-1} + k_2) / k_1$$

$$([E]_{\text{total}} [S] / [ES]) - [S] = (k_{-1} + k_2) / k_1$$

Define a new constant, $K_m = (k_{-1} + k_2) / k_1$

$$([E]_{\text{total}} [S] / [ES]) - [S] = K_m$$

Solve for the [ES] term (for reasons that will be given in the next step):

$$[ES] = [E]_{\text{total}} [S] / (K_m + [S])$$

The actual reaction velocity measured at any given moment is given by:

$$V = k_2[ES]$$

Multiple both sides of the above equation by k_2 :

$$k_2[ES] = k_2[E]_{\text{total}} [S] / (K_m + [S])$$

thus

$$V = k_2[E]_{\text{total}} [S] / (K_m + [S])$$

The maximum possible velocity ($V_{\max}$) occurs when all the enzyme molecules are bound with substrate $[ES] = [E]_{\text{total}}$, thus:

$$V_{\max} = k_2[E]_{\text{total}}$$

Substituting this into the prior expression gives:

$$V = V_{\max} [S] / (K_m + [S])$$

This is the mathematical expression that is used to model your experimental kinetic data

It is known as the Michaelis-Menten equation

Experimental approach

The general approach is to add a **known concentration of substrate** to the enzyme and to determine the **initial reaction rate** for that concentration of substrate

- Reaction rates are typically given as moles (or micromole) of product produced per unit of time (sec or min) per mole (or micromole) of enzyme
- The experiment is repeated for a wide range of substrate concentrations
- A table of $[S]$ versus V datapoints are collected
- These datapoints are plotted (V versus S) and should fit a curve that agrees with the Michaelis-Menten equation

The $V_{\max}$ and K_m terms are **intrinsic** properties of the particular enzyme/substrate combination that you are studying

- They will be determined from the features of the V versus S plot

$V_{\max}$

There are a limited number of enzyme molecules and they can only perform a single reaction at a time. Thus, at **high $[S]$** the enzymes can be **saturated**

- Under saturating conditions the reaction is going as fast as it can, and additional increases in $[S]$ do not increase the reaction rate.
- The maximum observable rate is $V_{\max}$ and the data will **asymptote to this value** at high $[S]$
- At low $[S]$ the reaction rate is generally linearly proportional to the $[S]$ (i.e. at low $[S]$ if you double $[S]$ the V will double)

K_m

$$K_m = (k_{-1} + k_2) / k_1 = (\text{rate of breakdown of ES}) / (\text{rate of formation of ES})$$

- K_m is **similar**, but not exactly equal to, a dissociation constant (K_d) for the ES complex
- If $k_{-1} \gg k_2$, then $K_m \gg K_d$
- Due to this similarity to the expression for K_d , a **low value of K_m** is often interpreted as a **high affinity of the enzyme for the substrate**, and a large value for K_m is often interpreted as a weak affinity of the enzyme for the substrate
- K_m has units of molar concentration (just like the units for $[S]$)

There is a mathematical treatment that allows for the determination of K_m from the experimental V versus $[S]$ data

- Consider the situation when the $[S]$ being evaluated results in a value of V that is exactly $1/2$ of the maximum reaction velocity:

$$V = V_{\max} [S] / (K_m + [S])$$

$$1/2 V_{\max} = V_{\max} [S] / (K_m + [S])$$

$$1/2 = [S] / (K_m + [S])$$

$$K_m + [S] = 2[S]$$

$$K_m = [S]$$

Thus, K_m equals the substrate concentration that results in exactly one half the maximum possible reaction velocity

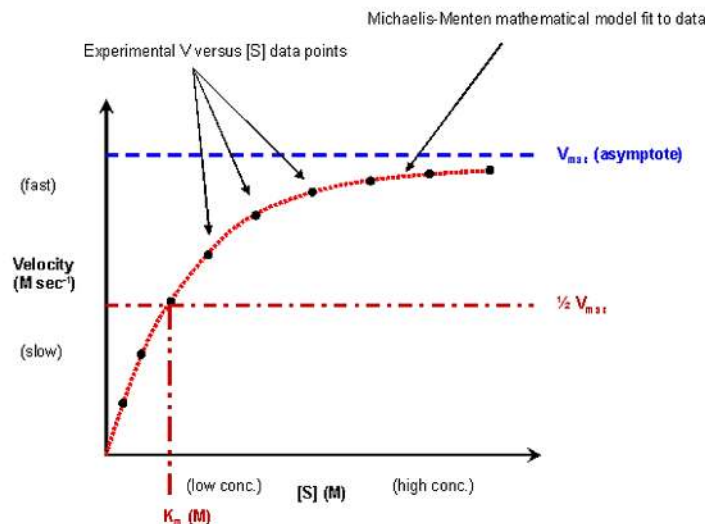


Figure 6.2.2: K_m

Lineweaver-Burke (the "double reciprocal" plot)

- The Michaelis-Menten equation can be rearranged by taking the reciprocal, to yield:

$$\frac{1}{V} = \frac{K_m}{V_{\max} [S]} + \frac{1}{V_{\max}}$$

- If $X = 1/[S]$ and $Y = 1/V$ then this is a linear equation with a slope of $K_m/V_{\max}$ and a Y intercept of $1/V_{\max}$

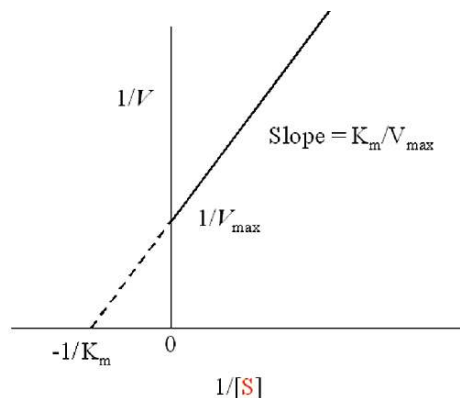


Figure 6.2.3: $1/S$ and $1/V$

- Since the plot of $1/[S]$ versus $1/v$ data should be a straight line, it is easier to fit a linear function to the data in this form, and $V_{\max}$ and K_m can be readily determined from the plot

Reversible Inhibition

There are two major categories of reversible inhibitors: **competitive** reversible inhibitors, and **noncompetitive** reversible inhibitors:

Competitive inhibitors

The inhibitor (**I**) competes with the substrate (S) for the enzyme active site (also known as the **S-binding site**). Binding of either of these molecules in the active site is a **mutually exclusive** event

- The substrate and inhibitor share a high degree of structural similarity. However, the inhibitor cannot proceed through the reaction to produce product.
- Increasing the concentration of substrate will **outcompete** the inhibitor for binding to the enzyme active site
- A competitive reversible inhibitor can be identified by its characteristic effects upon kinetic data

The expression for the Michaelis-Menten expression in the presence of a reversible competitive inhibitor is:

$$V = V_{\max} [S] / (K_m(1+[I]/K_i) + [S])$$

Where K_i is the actual EI complex dissociation constant

The effects of the reversible competitive inhibitor on the kinetics are as follows:

- If no inhibitor is present (i.e. if $[I] = 0$) then the equations are the same
- As inhibitor is added, the effect is to modify the apparent value of K_m . In particular, **the apparent K_m will be increased** by a value equal to $(1 + [I]/K_i)$. If K_m is increased, **the reaction velocity v will decrease**.
- Note that as $[S]$ gets very large the value of the denominator is essentially equal to $[S]$ and $v @ v_{\max}$. Thus, the reaction velocity can be driven to $v_{\max}$ with a high enough substrate concentration

The diagnostic criteria for reversible competitive inhibition is that while the *apparent* K_m is affected by addition of the inhibitor, the value of $v_{\max}$ does not change

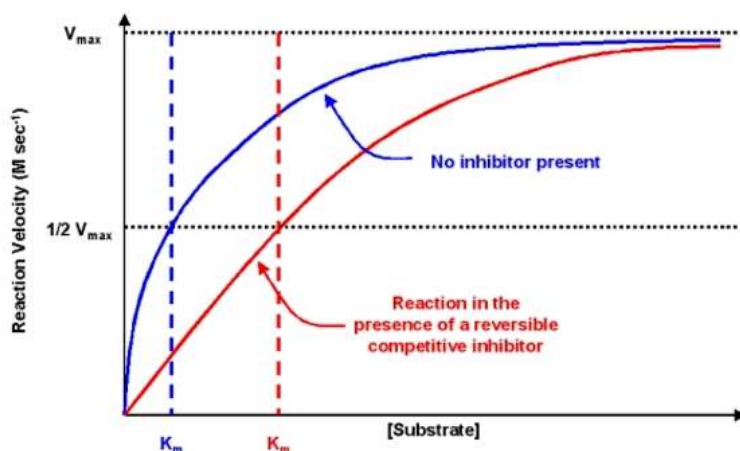


Figure 6.2.4: Effect of reversible competitive inhibitor

How is the Lineweaver-Burke double reciprocal plot affected by the presence of a reversible competitive inhibitor?

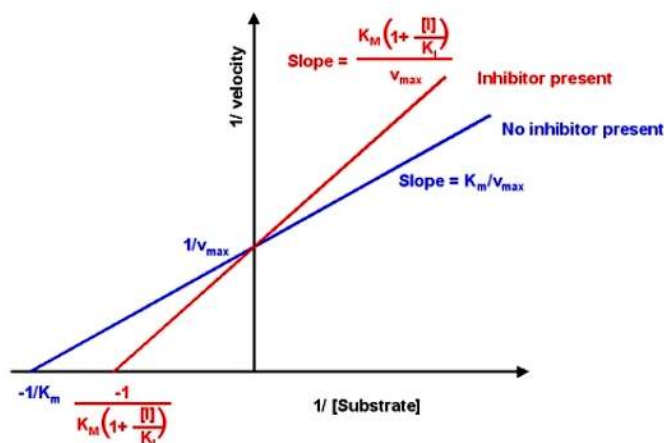


Figure 6.2.5: Double reciprocal plot with reversible competitive inhibitor

Noncompetitive Inhibitors

Noncompetitive inhibitors react with both E **and** ES (this is because the noncompetitive inhibitor does not bind at the same site in the enzyme as the substrate)

- Inhibition **cannot** be overcome by increasing the concentration of S
- The effect on kinetics is as if the enzyme were less active ($v_{\max}$ **is reduced**), but that the affinity for substrate is unaffected (K_m **remains the same**) since the substrate binding site is not occupied by the noncompetitive inhibitor.

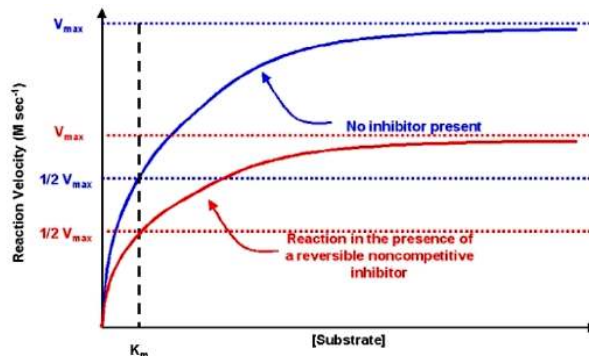


Figure 6.2.6: Effect of reversible noncompetitive inhibitor

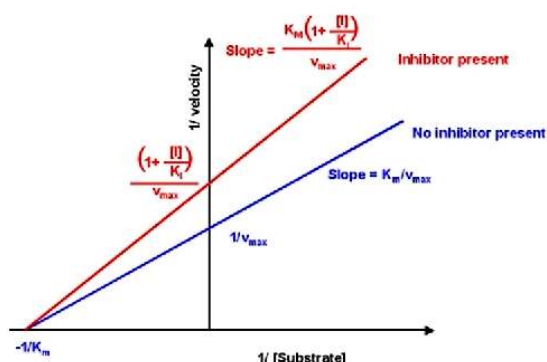


Figure 6.2.7: Double reciprocal plot with noncompetitive inhibitor

This page titled 6.2: Enzyme kinetics is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

6.3: Ligand binding

Molecular recognition is a key aspect of virtually all biochemistry, and involves:

- DNA replication
- Gene regulation
- Transcription of DNA into mRNA
- Translation of mRNA into proteins
- Enzyme function
- Molecular assembly
- Signal transduction
- Cell-cell communication
- Immunology

Intermolecular interactions, in the simplest example, involves **two molecules** recognizing and binding to each other:

- The **ligand**
- The **receptor**

These are broad terms, but the general properties of receptor and ligand are as follows:

- The ligand is typically, a small molecule, and it diffuses throughout the environment until it binds to a specific receptor
- The receptor is typically a large, relatively stationary molecule that contains a specific binding site for the ligand. Upon binding ligand, the structural and/or chemical properties of the receptor are typically altered

The interactions between receptor and ligand are typically non-covalent

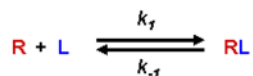
- There is a general molecular complementarity between the receptor and ligand
- Often, but not always, there is an entropic cost associated with binding of the ligand to receptor, but this is more than offset by a favorable enthalpy (i.e. the molecular complementarity)
- The binding energy is typically utilized to cause a structural change associated with a molecular communication event

There are two general methods to study receptor/ligand interactions:

1. **Equilibrium thermodynamics**, and
2. **Association and dissociation kinetics**

Equilibrium ligand/receptor binding analysis

The two possible states of a ligand/receptor interaction, and the rate constants associated with their formation, are given as:



The **equilibrium association (formation) constant**, K_f , is given as:

$$K_f = \frac{[\text{RL}]}{[\text{R}][\text{L}]}$$

- $[\text{RL}]$ is the concentration of the complex, $[\text{R}]$ is the concentration of free R at equilibrium, and $[\text{L}]$ is the concentration of free L at equilibrium
- K_f has units of M^{-1}
- A **large value for K_f** means that the equilibrium favors the complex - there is a **high affinity of the receptor for the ligand**
- A **small value for K_f** means that the equilibrium favors the separation of the receptor and ligand, and the receptor has a **low affinity for the ligand**

Just to make life difficult, the typical frame of reference when describing receptor affinity for ligand is to reference the **dissociation constant**, K_d :

$$K_d = \frac{1}{K_f} = \frac{[R][L]}{[RL]}$$

- K_d has units of M
- A **small value for K_d** means that the equilibrium favors the complex - there is a **high affinity of the receptor for the ligand**
- A **large value for K_d** means that the equilibrium favors the separation of the receptor and ligand, and the receptor has a **low affinity for the ligand**

At **equilibrium**, the **rate of formation** of the RL complex is **equal** to the **rate of dissociation**:

$$\text{Rate of formation of RL} = k_f[R][L]$$

$$\text{Rate of dissociation of RL} = k_d[RL]$$

$$\text{At equilibrium: } k_f[R][L] = k_d[RL]$$

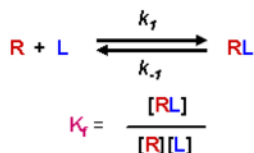
$$\frac{[R][L]}{[RL]} = \frac{k_d}{k_f}$$

The left-hand expression is the definition of the dissociation constant, K_d , therefore:

$$K_d = \frac{[R][L]}{[RL]} = \frac{k_d}{k_f}$$

The free energy expression and K_f and K_d

From the original expression of the complex formation:



The standard state free energy change, ΔG^0 , for the process (i.e. starting with 1M everything) would be given as:

$$\Delta G^0 = -RT \ln K_f$$

Since K_d is the inverse of K_f :

$$\Delta G^0 = -RT \ln K_f = RT \ln K_d$$

In this case, a negative value for ΔG^0 (indicating spontaneity) will occur when $K_d < 1.0$. This will happen when there is high affinity and the equilibrium favors the complex formation.

Practical relationships between $[R]$, $[L]$, K_d and $[RL]$

For the same concentration of added ligand and receptor, more complex formation will occur with a smaller value of K_d

Example 6.3.1

Receptor and ligand are added to a total concentration of $1.0 \times 10^{-4}M$. The K_d for the RL complex is $1.0 \times 10^{-4}M$. What is the concentration of $[R]$, $[L]$ and $[RL]$ at equilibrium?

$$K_d = \frac{[R][L]}{[RL]} = 1.0 \times 10^{-4}$$

The added ligand, L, will be partitioned between free and bound forms:

$$L_{\text{total}} = L + RL = 1.0 \times 10^{-4}M$$

$$L = 1.0 \times 10^{-4} M - RL$$

Similarly, the added receptor, R, will be partitioned between free and bound forms:

$$R_{\text{total}} = R + RL = 1.0 \times 10^{-4} M$$

$$R = 1.0 \times 10^{-4} M - RL$$

Substituting these values of L and R into the expression for K_d yields:

$$\frac{(1.0 \times 10^{-4} - RL)(1.0 \times 10^{-4} - RL)}{[RL]} = 1.0 \times 10^{-4}$$

$$(1.0 \times 10^{-4} - RL)(1.0 \times 10^{-4} - RL) = 1.0 \times 10^{-4} [RL]$$

$$1.0 \times 10^{-8} - 2.0 \times 10^{-4} RL + RL^2 = 1.0 \times 10^{-4} [RL]$$

$$RL^2 - 3.0 \times 10^{-4} RL + 1.0 \times 10^{-8} = 0$$

This is a quadratic with values $a = 1$, $b = -3.0 \times 10^{-4}$ and $c = 1.0 \times 10^{-8}$. This yields two possible values for RL:

$$RL = 2.62 \times 10^{-4} M \text{ or}$$

$$RL = 3.82 \times 10^{-5}$$

Since the maximum possible value for [RL] is 1.0×10^{-4} (i.e. given the starting concentrations of R and L, this is the most amount of RL that can be formed), the first result is not possible and **[RL] = $3.82 \times 10^{-5} M$**

Thus, 38.2% of the added receptor is in the complex formation (i.e. $3.82 \times 10^{-5} M / 1.0 \times 10^{-4}$) $\times 100\%$

Example 6.3.2

Same amount of added R and L, but this time the K_d for the RL complex is $1.0 \times 10^{-6} M$. What is the concentration of [R], [L] and [RL] at equilibrium?

$$\frac{(1.0 \times 10^{-4} - RL)(1.0 \times 10^{-4} - RL)}{[RL]} = 1.0 \times 10^{-6}$$

$$(1.0 \times 10^{-4} - RL)(1.0 \times 10^{-4} - RL) = 1.0 \times 10^{-6} [RL]$$

$$1.0 \times 10^{-8} - 2.0 \times 10^{-4} RL + RL^2 = 1.0 \times 10^{-6} [RL]$$

$$RL^2 - 2.01 \times 10^{-4} RL + 1.0 \times 10^{-8} = 0$$

his is a quadratic with values $a = 1$, $b = -2.01 \times 10^{-4}$ and $c = 1.0 \times 10^{-8}$. This yields two possible values for RL:

$$RL = 1.11 \times 10^{-4} M \text{ or}$$

$$RL = 9.05 \times 10^{-5}$$

Since the maximum possible value for [RL] is 1.0×10^{-4} (i.e. given the starting concentrations of R and L, this is the most amount of RL that can be formed), the first result is not possible and **[RL] = $9.05 \times 10^{-5} M$**

Thus, 90.5% of the added receptor is in the complex formation (i.e. $9.05 \times 10^{-5} M / 1.0 \times 10^{-4}$) $\times 100\%$

For the same concentration of added ligand and receptor, more complex formation will occur with a smaller value of K_d

Determination of dissociation (K_d) or binding (K_f) constants

The relationship between K_d , [R], [L] and [RL] is given by:

$$K_d = \frac{[R][L]}{[RL]}$$

- This is an equation with four variables, if you know three of them, you can solve for the fourth

- In principle, you can **solve for K_d** if you know the concentration of **free receptor** [R], **free ligand** [L] and **receptor/ligand complex** [RL] at **equilibrium**

To be able to solve for K_d , you need an assay that will provide the **molar concentrations** for each component. However, this is typically difficult to achieve. It is much more common to have an assay that will provide you with information regarding the **fraction** of added receptor (or ligand) that is **present in the free or complex** form.

- If $[R_t]$ is the **total** concentration of receptor in the sample, we know that it will be divided into two populations - bound and free:

$$[R_t] = [R] + [RL]$$

- The fraction of bound receptor sites is given by the ratio:

$$\text{Fraction of bound receptor molecules} = \frac{[RL]}{[R_t]}$$

- The fraction of bound receptors is known by the variable **a**, and **a** has a value between 0 and 1

$$\alpha = \frac{[RL]}{[R_t]}$$

- Substituting the value for $[R_t]$ in terms of [R] and [RL] given above yields:

$$\alpha = \frac{[RL]}{[R] + [RL]}$$

- The relationship between [RL] and K_d is given by the expression for K_d :

$$K_d = \frac{[R][L]}{[RL]}$$

$$[RL] = \frac{[R][L]}{K_d}$$

- This expression for [RL] can be substituted into the expression for α to yield:

$$\alpha = \frac{\frac{[R][L]}{K_d}}{[R] + \frac{[R][L]}{K_d}}$$

- Simplify by first dividing through by [R]:

$$\alpha = \frac{\frac{[L]}{K_d}}{1 + \frac{[L]}{K_d}}$$

- Then multiply by K_d :

$$\alpha = \frac{[L]}{K_d + [L]}$$

This equation has three variables:

- The concentration of free ligand at equilibrium [L]

2. The dissociation constant, K_d , for the RL complex
3. The fraction of receptor that is in the bound form at equilibrium, a

If you know two of these variables, you can solve for the third (i.e. if the experimental assay quantifies $[L]$ and a , you can solve for K_d)

The above equation can be manipulated further by substitution of the value of a in terms of $[RL]$ and $[R_t]$:

$$\frac{[RL]}{[R_t]} = \frac{[L]}{K_d + [L]}$$

$$[RL] = \frac{[R_t][L]}{K_d + [L]}$$

This equation has four variables:

1. The concentration of free ligand at equilibrium $[L]$
2. The dissociation constant, K_d , for the RL complex
3. The concentration of receptor that is in the bound form at equilibrium $[RL]$
4. The total concentration of receptor in the sample

Usually, the ligand-binding experiment is setup so that $[R_t]$ is held constant and $[RL]$ is monitored as a function of $[L]$

This equation looks a lot like the Michaelis-Menten equation:

$$V = \frac{V_{\max} [S]}{K_m + [S]}$$

Plotting V versus $[S]$ in the M-M equation yields a rectangular hyperbola. Thus, **plotting $[RL]$ versus $[L]$** will likewise yield a **rectangular hyperbola**

- The asymptote of the curve will yield the value of $[R_t]$
- The free ligand concentration $[L]$ that results in one-half the maximum value of $[RL]$ is equal to the K_d value

Yet another modification is to realize that if a equals the fraction of receptor bound in a complex (with a value $0 \leq a \leq 1.0$), then $(1 - a)$ is the **fraction of free receptor** (i.e. **fraction of unbound receptor**)

$$a = \frac{[RL]}{[R_t]}$$

$$1 - a = 1 - \frac{[RL]}{[R_t]}$$

$$1 - a = \frac{[R_t]}{[R_t]} - \frac{[RL]}{[R_t]}$$

$$1 - a = \frac{[R_t] - [RL]}{[R_t]}$$

$$\text{if } [R_t] = [R] + [RL]$$

$$\text{then } [R_t] - [RL] = [R]$$

$$1 - a = \frac{[R]}{[R_t]}$$

This value can be substituted into the three variable expressions above:

$$\alpha = \frac{[L]}{K_d + [L]}$$

$$1 - \alpha = 1 - \frac{[L]}{K_d + [L]}$$

$$\frac{[R]}{[R_t]} = 1 - \frac{[L]}{K_d + [L]}$$

$$[R] = [R_t] - \frac{[R_t][L]}{K_d + [L]}$$

$$[R] = \frac{[R_t] K_d + [R_t][L]}{K_d + [L]} - \frac{[R_t][L]}{K_d + [L]}$$

$$[R] = \frac{[R_t] K_d}{K_d + [L]}$$

This equation has four variables:

1. The concentration of free ligand at equilibrium [L]
2. The dissociation constant, K_d , for the RL complex
3. The concentration of receptor that is in the free form at equilibrium [R]
4. The total concentration of receptor in the sample $[R_t]$

Usually, the ligand-binding experiment is setup so that $[R_t]$ is held constant and $[R]$ is monitored as a function of $[L]$

(Note: Although this looks like a M-M equation, K_d is a constant.)

The above derivations are called "**binding isotherms**" in reference to their ability to solve for K_d under equilibrium conditions (and the original equilibrium thermodynamic equation utilized constant temperature)

"Binding Isotherms"

$$[RL] = \frac{[R_t][L]}{K_d + [L]} \quad \text{Determine } K_d \text{ from knowing } [RL], [R_t] \text{ and } [L]$$

$$[R] = \frac{[R_t] K_d}{K_d + [L]} \quad \text{Determine } K_d \text{ from knowing } [R], [R_t] \text{ and } [L]$$

Plotting data for various values of [L], $[R_t]$, and either [RL] or [R], and fitting to the appropriate function above, allows a more accurate determination of K_d (although a single experiment will provide an answer).

Due to the similarity of the [RL] versus [L] plot with the M-M plot, it is often the more desirable experimental setup. In particular, a **double-reciprocal plot** can be used (much like the double-reciprocal M-M plot) to more accurately determine K_d

$$\frac{1}{[RL]} = \frac{K_d + [L]}{[R_t] [L]}$$

$$\frac{1}{[RL]} = \frac{K_d}{[R_t] [L]} + \frac{1}{[R_t]}$$

$$\frac{1}{[RL]} = \frac{K_d}{[R_t]} \frac{1}{[L]} + \frac{1}{[R_t]}$$

$$Y = m \cdot X + B$$

- If you plot $1/[RL]$ versus $1/[L]$ you will get a straight line whose slope is equal to $K_d/[R_t]$
- Y intercept will be equal to $1/[R_t]$. Since this value is fixed and is known, the value of the y intercept can serve as a check to make sure the experiment fits the model

An equilibrium dialysis experiment

A semi-permeable membrane can be used in a dialysis-based equilibrium experiment to determine K_d for a ligand-receptor pair.

- The membrane should be freely permeable as far as the ligand is concerned (usually it is a small molecule), but present a barrier to the receptor
- The experiment is established with a known concentration of receptor on one side, and a known concentration of ligand on the other
- The system is allowed to come to equilibrium
- The experiment requires an assay for the ligand

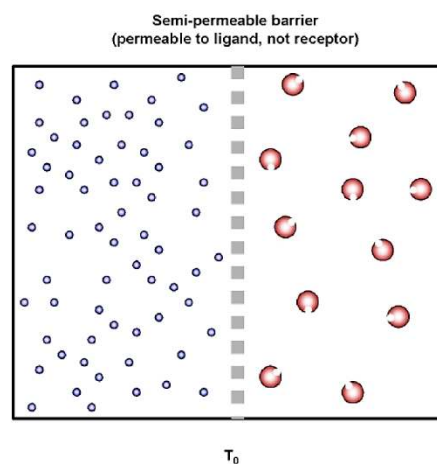


Figure 6.3.1: Semipermeable barrier

After equilibrium is reached the ligand concentration on both sides of the semi-permeable membrane is assayed:

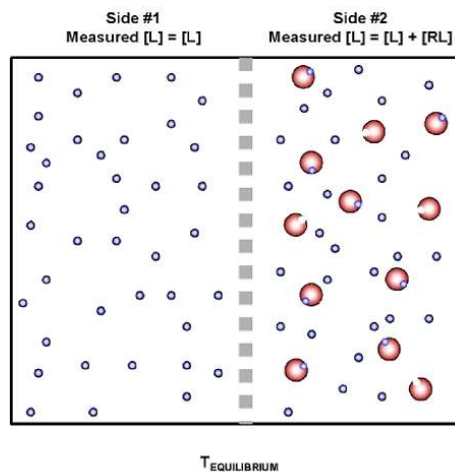


Figure 6.3.2: Equilibrium concentration

$$[R_t] = \text{known}$$

$$[L] = \text{Side\#1}[L]$$

$$[RL] = (\text{Side\#2}[L] - \text{Side\#1}[L])$$

$$[R] = [R_t] - [RL]$$

$$[RL] = \frac{[R_t][L]}{K_d + [L]}$$

$$[R] = \frac{[R_t]K_d}{K_d + [L]}$$

- Since the ligand is unaffected by the membrane the ligand concentration on the side with no receptor is equal to the free ligand concentration $[L]$ everywhere in the sample (i.e. on both sides of the membrane)
- The ligand concentration will be equal to the free ligand concentration $[L]$ plus the ligand present in the receptor complex $[RL]$ (in other words, there should be a higher concentration of ligand on the side with the receptor if the receptor has affinity for the ligand)
- The concentration of the receptor/ligand complex $[RL]$ can be determined from the difference between these two values
- The total receptor concentration is a known value (experimenter determined at the start of the experiment, and does not change)
- Either binding isotherm can be used to solve for K_d (since $[R]$ can also be determined)

This page titled [6.3: Ligand binding](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

Since different bacterial strains and species have potentially different R/M systems, their characterization has made available **hundreds** of endonucleases with different sequence specific cleavage sites.

- They are one of the primary tools in modern molecular biology for the manipulation and identification of DNA sequences.

- Restriction endonucleases are commonly named after the bacterium from which it was isolated.

Name	Source	Recognition Sequence	Comments
<i>Alu I</i>	<i>Arthrobacter luteus</i>	5' ... A G C T ... 3' 3' ... T C G A ... 5'	"Four cutter". Leaves blunt ends to the DNA.
<i>Bfa I</i>	<i>Bacteroides fragilis</i>	5' ... C T A G ... 3' 3' ... G A T C ... 5'	"Four cutter". Leaves 5' overhang.
<i>Nci I</i>	<i>Neisseria cinerea</i>	C 5' ... C C G G G ... 3' 3' ... G G C C C ... 5' G	"Five cutter". Middle base can be either cytosine or guanine. Leaves 5' overhang. <u>Different recognition sites may have non-complementary sequences.</u>
<i>Eco RI</i>	<i>Escherichia coli</i>	5' ... G A A T T C ... 3' 3' ... C T T A A G ... 5'	"Six cutter". Leaves 5' overhang. Behaves like a "four cutter" ('star' activity) in high salt buffer. \$44 for 10,000 units.

<i>Hae II</i>	<i>Haemophilus aegyptius</i>	5'... Pu G C G C Py ... 3' 3'... Py C G C G Pu ... 5'	"Six cutter". Pu is any purine, Py is any pyrimidine. Leaves 3' overhang.
<i>EcoO109I</i>	<i>Escherichia coli</i>	5'... Pu G G N C C Py ... 3' 3'... Py C C N G G Pu ... 5'	"Seven cutter". Pu is any purine, Py is any pyrimidine, N is any base. Leaves 5' overhang. <u>Different recognition sites may have non-complementary sequences.</u>
<i>Bgl I</i>	<i>Bacillus globigii</i>	5'... GCCN NNNNGGC ... 3' 3'... CGGNNNN NCCG ... 5'	"Six cutter with interrupted palindrome". Leaves 5' overhang. <u>Different recognition sites may have non-complementary sequences.</u>
<i>Bsa HI</i>	<i>Bacillus stearothermophilus</i>	5'... G Pu C G Py C ... 3' 3'... C Py G C Pu G ... 5'	"Six cutter". Different recognition sites <u>will</u> be complementary.
<i>Aat II</i>	<i>Acetobacter aceti</i>	5'... G A C G T C ... 3' 3'... C T G C A G ... 5'	"Six cutter" with 3' overhang. Same recognition sequence as Bsa HI, but different cleavage position.

<i>Bpm I</i>	<i>Bacillus pumilus</i>	5'... C T G G A G N₁₆ ... 3' 3'... G A C C T C N₁₄ ... 5'	Non-palindrome, distal cleavage. Leaves 3' overhang. \$50 for 50 units.
<i>Not I</i>	<i>Nocardia otitidis cavarius</i>	5'... G C G G C C G C ... 3' 3'... C G C C G G C G ... 5'	"Eight cutter". Leaves 5' overhang.
<i>Bsm I</i>	<i>Bacillus stearothermophilus</i>	5'... G A A T G C N ... 3' 3'... C T T A C G N ... 5'	"Weird". Leaves 3' overhang.

- The utility of restriction endonucleases lies in their specificity and the frequency with which their recognition sites occur within any given DNA sample.
- If there is a 25% probability for a specific base at any given site, then the frequency with which different restriction endonuclease sites will occur can be easily calculated (0.25ⁿ):**

Specificity	Example	Frequency of Occurrence
Four base sequence	<i>Alu I</i>	1 Alu site in every 256 bases (0.25 Kb)
Five base sequence	<i>Nci I</i>	1 Nci I site in every 1024 bases (1.0 Kb)
Six base sequence	<i>EcoR I</i>	1 EcoR1 site in every 4,096 bases (4.1 Kb)
Seven base sequence	<i>EcoO109I</i>	1 EcoO109I site in every 16,384 bases (16.4 Kb)
Eight base sequence	<i>Not I</i>	1 Not I site in every 65,536 bases (65.5 Kb)

Thus, on average, any given DNA will contain an Alu I site every 0.25 kilobases, whereas a Not I site occurs once about every 65.5 kilobases.

- Not I is therefore a very useful enzyme for isolating large regions of DNA, typically in research involving genomic DNA manipulations.
- Alu I would be expected to digest a DNA sample into lots of little pieces.

The assortment of DNA fragments would represent a specific "**fingerprint**" of the particular DNA being digested. Different DNA would not yield the same collection of fragment sizes. Thus, DNA from different sources can be either matched or distinguished based on the assembly of fragments after restriction endonuclease treatment. **These are termed "Restriction Fragment Length Polymorphisms", or RFLP's.** This simple analysis is used in various aspects of molecular biology as well as a law enforcement and genealogy. For example, genetic variations that distinguish individuals also may result in **fewer or additional restriction endonuclease recognition sites.**

Gel Electrophoresis of DNA

The most common gel electrophoresis solid support matrix for DNA molecules is

- **agarose** and
- **acrylamide.**

DNA agarose gels

The electrophoretic migration rate of DNA through agarose gels is dependent upon **four main parameters**:

1. The **molecular size** of the DNA. Molecules of linear duplex DNA travel through agarose gels at a rate which is inversely proportional to the log of their molecular weight.

$$M_r \propto 1/\log (M_w)$$

Example: Compare molecular mass vs. expected migration rate:

Molecular Mass (Da)	log (Molec. Mass)	1/log (Molec. Mass) <i>i.e. relative Mr</i>
100,000	5.0	0.20
50,000	4.7	0.21
10,000	4.0	0.25
5,000	3.7	0.27
1,000	3.0	0.33

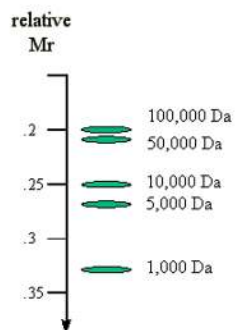


Figure 6.4.4: Molecular mass and migration rate

2. The **agarose concentration**. There is an inverse linear relationship between the logarithm of the electrophoretic mobility and gel concentration.

Agarose (%)	Range of separation of linear DNA (in kilobases)
0.3	60 - 5
0.6	20 - 1
0.7	10 - 0.8
0.9	7 - 0.5
1.2	6 - 0.4
1.5	4 - 0.2
2.0	3 - 0.1

3. The conformation of the DNA.

- closed circular DNA (**form-I**) - typically supercoiled (compact)
- nicked circular (**form-II**) - nick relaxes any supercoiling
- linear DNA (**form-III**)

These different forms of the *same* DNA migrate at different rates through an agarose gel.

- Almost always the **linear form** (form-III) migrates at the **slowest** rate of the three forms
- Supercoiled DNA (form-I) usually migrates the **fastest**

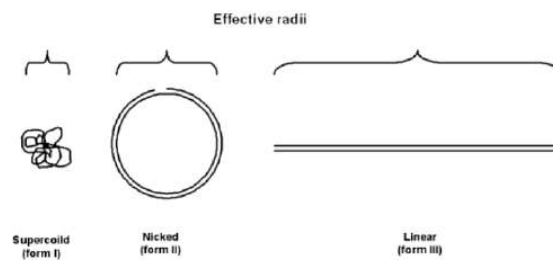


Figure 6.4.5: Forms of DNA

4. The applied **voltage**.

Other details:

- Typical value for running an agarose gel is **5 volts per cm** (length of gel).
- Agarose gels are usually poured and run horizontally
- Finally, the DNA being an acidic molecule, migrates towards the positively charged electrode (cathode). DNA naturally has a constant charge to mass ratio, so no detergents need to be added (as with proteins)

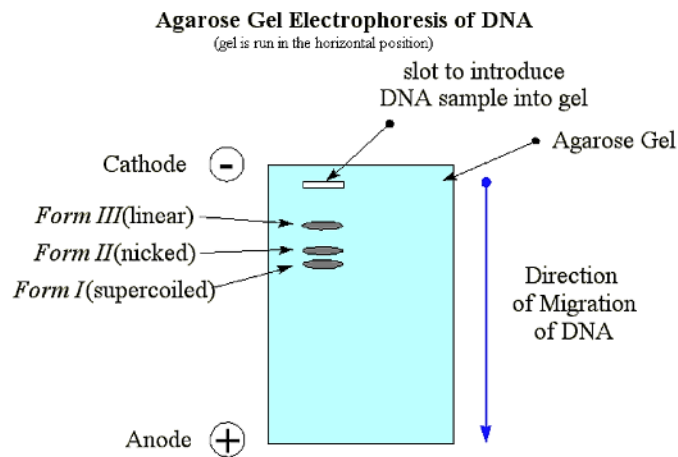


Figure 6.4.6: Gel electrophoresis

DNA acrylamide gels

- Acrylamide gels are useful for separation of **small DNA fragments**
 - typically oligonucleotides <100 base pairs.
 - These gels are usually of a *low acrylamide concentration* ($\leq 6\%$) and contain the non-ionic *denaturing agent* Urea (6M).
 - The denaturing agent prevents secondary structure formation in oligonucleotides and allows a relatively accurate determination of molecular mass.

Staining of DNA

- The most convenient method to visualize DNA in gel electrophoresis is staining with the fluorescent dye **ethidium bromide**.

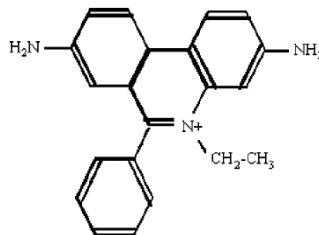


Figure 6.4.7: Ethidium bromide

- This compound contains a planar group that **intercalates** between the stacked bases of DNA.
- The orientation and proximity of ethidium with the stacked bases causes the dye to display an increased fluorescence compared to free dye (in solution).
 - U.V. radiation at 254 nm is absorbed by the DNA and transmitted to the bound dye.
 - The energy is re-emitted at 590 nm in the **red-orange** region of the spectrum.
- Ethidium bromide is usually prepared as a stock solution of 10 mg/ml in water, stored at room temp and protected from light.
- The dye is usually incorporated into the gel and running buffer, or conversely, the gel is stained after running by soaking in a solution of ethidium bromide (0.5 ug/ml for 30 min).
- The stain is visualized by irradiating with a UV light source (i.e. using a transilluminator) and photographing with polaroid film.
- The usual sensitivity of detection is better than 0.1 ug of DNA.

Because ethidium is a DNA intercalating agent, it is a powerful mutagen. Incorporation of ethidium in the DNA of living organisms (i.e. you and I) can cause (unwanted) mutations.

Combining restriction endonuclease digestion with gel electrophoresis of DNA: Restriction mapping

A given sequence of DNA (e.g. a gene) will have a specific sequence, and therefore, specific restriction endonuclease sites

- The number and location of such sites is a unique and predictable property for a given DNA molecule
- The fragmentation pattern (i.e. number and size of fragments after restriction endonuclease digestion) can be characterized by gel electrophoresis as a type of "DNA fingerprint"
- Any changes in the DNA sequence of a gene can result in the elimination of particular restriction sites, and conversely, create new ones:

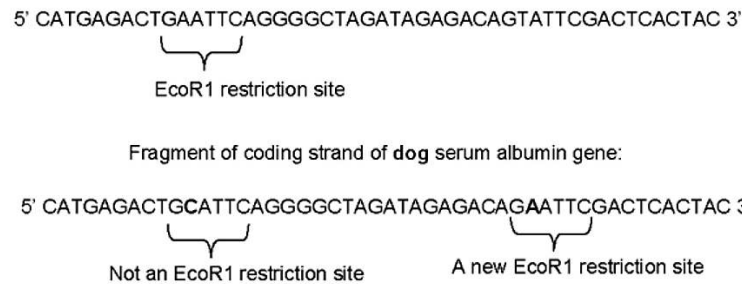


Figure 6.4.8: Change in restriction sites

- The gel electrophoresis fragment pattern for the human and dog albumin gene will be characteristically different from each other
- A sample of blood can potentially be identified as either human or dog by observing the **restriction fragment length polymorphism**
- Genetic differences between individuals can also be identified using restriction fragment length polymorphism analysis

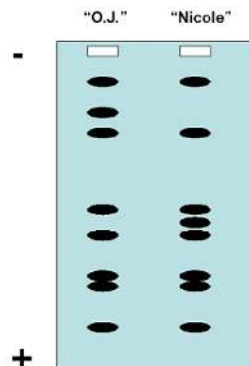


Figure 6.4.9: Using RFLP to see genetic differences

This page titled 6.4: Restriction Mapping is shared under a not declared license and was authored, remixed, and/or curated by Michael Blaber.

6.5: Polymerase Chain Reaction (PCR)

Review

DNA replication is accomplished by the enzyme DNA polymerase which has the following characteristics:

1. DNA polymerase catalyzes extension of an oligonucleotide in the **5' → 3' direction**. It cannot go in the opposite direction.
2. A **DNA template** (i.e. single strand oligonucleotide) is required. This is the information that is replicated (via Watson-Crick base pairing). If you want to replicate the "sense" strand of a DNA duplex, then the template is the anti-sense strand. Since DNA is a duplex, held together by non-covalent forces, we can separate the strands of the duplex (and obtain an appropriate template) by heat (or alkali) denaturation.
3. DNA polymerase can only extend an **existing oligonucleotide**, it cannot initiate replication. Thus, in addition to a template, DNA polymerase requires a primer oligonucleotide. DNA polymerase extends the primer in the 5' → 3' direction.
4. Nucleotides incorporated into the nascent oligonucleotide must be **nucleoside triphosphates** (i.e. dNTP's). The energy released in the hydrolysis of the phosphate anhydride bonds is used in the formation of covalent linkages (incorporation of the α-phosphate into the phosphodiester backbone of the growing oligonucleotide)
5. Although DNA polymerases catalyze extension in the 5' → 3' direction, some have the ability to "**proof-read**". "Proof reading" is the ability to recognize a misincorporation of a base, and the ability to backup and excise the incorrect base, and then continue on.

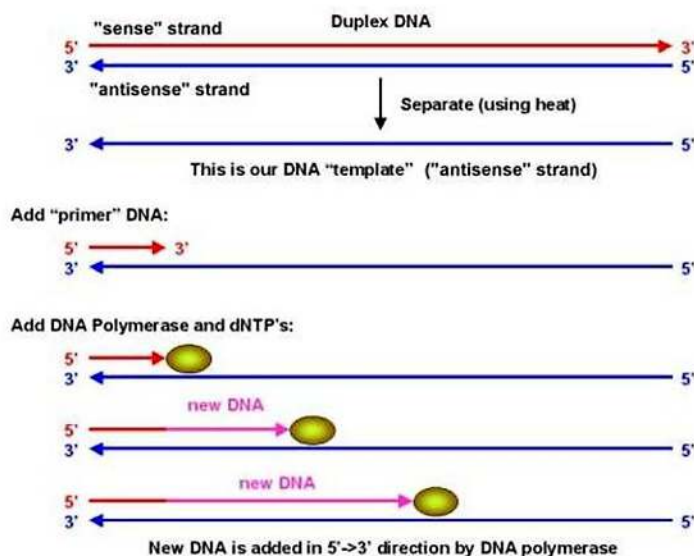


Figure 6.5.1: DNA synthesis

Polymerase Chain Reaction (PCR)

- PCR is an *in vitro* technique for the amplification of a region of DNA which lies **between two regions of known sequence**.
- PCR amplification is achieved by using oligonucleotide primers. These are typically short, single stranded oligonucleotides which are **complementary to the outer regions** of known sequence.

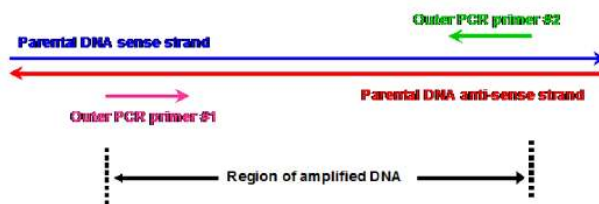


Figure 6.5.2: PCR amplification

The oligonucleotides serve as **primers** for DNA polymerase and each of the denatured strands of the parental DNA duplex serves as the **template**.

This results in the synthesis of new DNA strands which are complementary to the parent template strands.

The steps of:

1. *Template denaturation*
2. *Primer annealing*
3. *Primer extension*

comprise a single "cycle" in the PCR amplification methodology.

- After each cycle the newly synthesized DNA strands can **serve as templates** in the next cycle (the PCR primers are typically added in substantial molar excess to the template DNA)

Summary of products at the end of each PCR cycle:

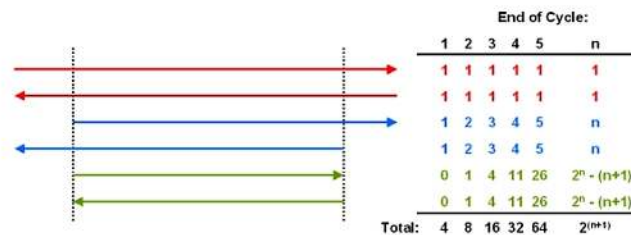


Figure 6.5.3: PCR products

The desired PCR product will be a **duplex of the defined length fragment**. The question is: how many will be produced?

• The expected **amplification** of the desired defined length product with respect to the original template concentration 'x' can thus be represented by the formula:

$$[(2^n - (n + 1)) - (n + 1)] x$$

or

$$(2^n - 2(n + 1))$$

(this is often abbreviated to a simple rule of thumb for the **amplification**: $(2^n - 2n) x$)

• The interpretation of this formula is that

- For a given number of cycles 'n' we make ' $2^n x$ ' **total possible duplexes**
- For a given number of cycles there will be ' $2(n+1)$ ' (or $2n$ in our approximation) x ' duplexes which are formed from either the original template, or a fragment of indeterminate length, along with a fragment of defined length (and represent an undesired product)
- Thus, the total concentration of desired product (duplexes with a length defined by the PCR primers) will be $(2^n - 2(n+1)) x$ (where x is the concentration of the original duplex)

The theoretical amplification value is never achieved in practice. Several factors prevent this from occurring, including:

1. Competition of complementary daughter strands with primers for reannealing (i.e. two daughter strands reannealing results in no amplification).
2. Loss of enzyme activity due to thermal denaturation, especially in the later cycles
3. Even without thermal denaturation, the amount of enzyme becomes limiting due to molar target excess in later cycles (i.e. after 25 - 30 cycles too many primers need extending)
4. Possible second site primer annealing and non-productive priming

Thermal cycling parameters

The thermal cycling parameters are critical to a successful PCR experiment. The important steps in each cycles of PCR include:

1. denaturation of template (typically performed at highest temp - 100°C)
2. annealing of primers (temperature is chosen based upon melting temperature of primer)
3. extension of the primers (performed at optimum for the polymerase being used)

A representative temperature profile for each cycle might look like the following:

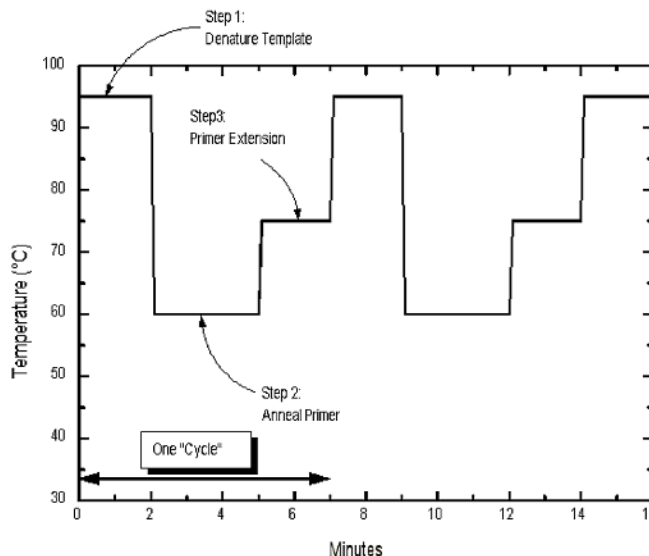


Figure 6.5.4: Thermal cycling

Buffers and MgCl₂ in PCR reactions

A typical reaction buffer for PCR would something like:

- 10 mM Tris, pH 8.3
- 50 mM KCl
- 1.5 mM MgCl₂
- 0.01% gelatin
- The MgCl₂ concentration in the final reaction mixture is usually between 0.5 to 5.0 mM, and the optimum concentration is determined empirically (typically between 1.0 - 1.5 mM). **Mg²⁺ ions:**
 - form a soluble complex with dNTP's which is essential for dNTP incorporation
 - stimulate polymerase activity
 - increase the T_m (melting temperature) of primer/template interaction (i.e. it serves to stabilize the duplex interaction)

Generally,

- **low** Mg²⁺ leads to low yields (or no yield) and
- **high** Mg²⁺ leads to accumulation of nonspecific products (mispriming).

Choice of Polymerases for PCR

- One of the important advances which allowed development of PCR was the availability of thermostable polymerases.
- This allowed initially added enzyme to survive temperature cycles approaching 100 °C.
- Properties of DNA polymerases used in PCR

	<i>Taq</i> /Amplitaq®	Vent™	Deep Vent™	<i>Pfu</i>	<i>Tth</i>	ULTma™
95 °C half-life	40 min	400 min	1380 min	>120 min	20 min	>50 min
Extension rate (nt/sec)	75	>80	?	60	>33	?
Resulting ends	3' A	>95% blunt	>95% blunt	blunt	3' A	blunt
Strand displacement		+	+			
5'→3' exo	+				+	
3'→5' exo		+	+	+		+

Primers

Primer design

- Generally, primers used are 20 - 30 mer in length. This provides for practical annealing temperatures (of the high temperature regimen where the thermostable polymerase is most active).
- Primers should avoid stretches of polybase sequences (e.g. poly dG) or repeating motifs - these can hybridize with inappropriate register on the template.
- Inverted repeat sequences should be avoided so as to prevent formation of secondary structure in the primer, which would prevent hybridization to template
- Sequences complementary to other primers used in the PCR should be avoid so as to prevent hybridization between primers (particularly important for the **3' end** of the primer)
- If possible the 3' end of the primer should be rich in G, C bases to enhance annealing of the end which will be extended
- The distance between primers should be less than 10 Kb in length. Typically, substantial reduction in yield is observed when the primers extend from each other beyond ~3 Kb.

Melting temperature (T_m) of primers

- The T_m of primer hybridization can be calculated using various formulas. The most commonly used formula is:

$$(1) T_m = [(\text{number of A+T residues}) \times 2 ^\circ\text{C}] + [(\text{number of G+C residues}) \times 4 ^\circ\text{C}]$$

- Examples of T_m calculations

	G	A	T	C	T_m
15 mer	3	5	2	5	46
20 mer	6	5	4	5	62
30 mer	8	6	8	8	92

If the annealing temperature is too high, the primers will not anneal. If it is too low, mispriming can occur at sites of similar DNA sequence to the intended primer binding site

Analysis of the PCR experiment

- 25-30 cycles are common
- Major product should be a duplex DNA molecule whose 5' and 3' ends are defined by the two PCR primers
- Other products can, however, be produced. These will often represent products of secondary priming sites (mispriming) or misincorporation errors of the polymerase
- Thus, PCR product can be heterogeneous and must be separated and analyzed
- **Agarose gel electrophoresis**, in combination with ethidium bromide staining, is the most common method to separate and analyze PCR products

This page titled [6.5: Polymerase Chain Reaction \(PCR\)](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

6.6: Use of PC and internet for biochemical research

There are a variety of resources on the web related to biochemical research. These resources typically include general categories of:

- Databases of information
- Software for analysis of data
- Software for graphing and representing data
- Published reports and reviews (literature)

Databases of information

There are numerous repositories of information related to biological molecules and biochemistry. A few of these (along with some sample links) include:

- Databases of nucleic acid sequences, including commonly used plasmids.

[The Entrez Nucleotide Database](#)

[New England Biolabs™ Database of vectors and plasmids](#)

- Protein sequence databases

[Protein sequence database \(Protein Information Resource\)](#)

[Swiss-Prot knowledge database](#)

- Protein structure database

[The Protein Databank \(PDB\)](#)

- Nucleic acid structure database

[The Nucleic Acid Database](#)

- Protein stability and thermodynamic database

[Protherm Database](#)

- Protein 2-D gel electrophoresis databases (proteomics)

[EXPASY 2-D PAGE links](#)

Software for analysis of data

Software can either be run on a local machine (i.e. your PC on your desktop), or on a server over the web. Sometimes, it is a combination of the two (i.e. a Java™ applet that is served on a host machine, but requires Java software installed on your PC).

In general terms, if your data is expected to follow a particular mathematical model (e.g. Michaelis-Menten kinetics) then you need general-purpose curve-fitting software. Links to non-linear least-square fit software include:

[Datafit™ from Oakdale Engineering](#)

[LSM Least Squares Method curve fitting](#)

Programs such as these allow you to simply write an equation, input your experimental values, and then the parameters of the equation (i.e. model) will be refined to give the best fit to the data. Thus, a single program can be used for a variety of applications (i.e. any situation where an equation is available that describes the behavior of the data)

Special purpose programs are available for specific applications.

Software for graphing and presenting data

In distinction to software for analysis (although sometimes combined in the same package) is software for graphing and representing data. The subtle point here is that after the data has been analyzed, you may need to prepare a report to communicate the results. In this case, there are other issues of data presentation - i.e. colors, line thicknesses, arrow styles, axis-breaks, fonts, legends, etc. that may need to be fussed with. Basic fitting software often does not include extensive presentation tools, and

similarly, presentation software often does not have robust and general purpose fitting tools. Examples of graphing and presentation software for scientific data are:

[Origin™ from MicroCal™](#)

[KaleidaGraph from Synergy Software™](#)

[SigmaPlot™ from SPSS, Inc.](#)

In addition to graphic representation of experimental data, some forms of data require three-dimensional representations. For example, biomolecular structures (e.g. RNA, DNA and protein molecules) are stored in databases (e.g. PDB) as a series of atomic coordinates, however, the data only really makes sense if you can convert such data into a representation of a three-dimensional molecule. Furthermore, understanding the structural information is possible by manipulation of such representations (e.g. rotations, translations, zooming in/out, etc.). Software for such manipulations includes:

[Swiss PDB viewer](#)

[DS Viewer™ from Accelrys](#)

[Pymol](#)

[Q](#)

Furthermore, publication of figures of molecular structures often necessitate artistic embellishments - such as raytracing, shading, etc. The above programs can often output their figures in POV format, which allows such modifications using the following:

[POV-RAY™](#)

Literature databases

Scientific reports are published in various journals, meeting proceedings, and various books. Search engines are available to search electronic repositories of these materials. PubMed, a service of the National Library of Medicine, includes over 14 million citations for biomedical articles back to the 1950's. These citations are from MEDLINE and additional life science journals. PubMed includes links to many sites providing full text articles and other related resources.

[Here is a link to PubMed™](#)

Additionally, it is also useful to be able to easily download, maintain, and format such references in papers or manuscripts that you have to prepare. Bibliography software allows you to do this quite easily. Examples of such software include ProCite, Reference Manager and Endnote (all from ISI ResearchSoft:

[Endnote™ from ISI ResearchSoft™](#)

Understanding DNA sequence information files

When requesting DNA sequence information there is a lot of additional information that is included in the file - in addition to the actual sequence. Here is an example of part of the file for the sequence of the plasmid pUC19:

Plasmid pUC19
Update 6/5/02
Features:
469- 146 lacZ alpha CDS (start 469, complementary strand)
519- 514 Plac promoter -10 sequence (TATGTT)

543- 538	Plac promoter -35 sequence (TTTACA)
575- 563	CAP protein binding site
396- 452	multiple cloning site (EcoRI-HindIII)
1455- 867	origin of replication (counterclockwise)
	(RNAII -35 to RNA/DNA switch point):
1273-1278	RNAI transcript promoter -35 sequence (TTGAAG)
1295-1300	RNAI transcript promoter -10 sequence (GCTACA)
1309-1416	RNAI transcript
1419- 867	RNAII transcript (complementary strand)
1434-1429	RNAII transcript promoter -10 sequence (CGTAAT)
1455-1450	RNAII transcript promoter -35 sequence (TTGAGA)
2486-1626	beta-lactamase (bla; amp-r) CDS
	(start 2486, complementary strand)
2486-2418	beta-lactamase signal peptide CDS
	(start 2486, complementary strand)
2521	bla RNA transcript start (complementary strand)
2535-2530	bla promoter -10 sequence (GAGACA)
2556-2551	bla promoter -35 sequence (TTCAAA)
puc19.seq Length: 2686 June 5, 2002 13:46 Type: N Check: 4141 ..	
1	TCGCGCGTTT CGGTGATGAC GGTGAAAACC TCTGACACAT GCAGCTCCCG

51 GAGACGGTCA CAGCTTGTCT GTAAGCGGAT GCCGGGAGCA GACAAGCCCCG

101 TCAGGGCGCG TCAGCGGGTG TTGGCGGGTG TCGGGGCTGG CTTAACTATG

151 CGGCATCAGA GCAGATTGTA CTGAGAGTGC ACCATATGCG GTGTGAAATA

The actual DNA sequence starts towards the bottom of this file (it is truncated because it goes on for about 3000 bases). The start of the file is called the "header" and provides notes about features of the DNA sequence. For example, nucleotide bases 867 - 1455 comprise an origin of replication. It is listed as 1455 - 867 to communicate that the orientation of this "ori" is actually counterclockwise around the plasmid (plasmids are circular). There also is a gene for b-lactamase (bla; which provides for ampicillin resistance, or amp^R selectable marker) that runs from basepair 1626 to 2486. It is also counterclockwise. Furthermore, the promoter for transcribing this gene is located from basepairs 2551 - 2521 (i.e. just "upstream" of the bla gene). The vector looks like this:

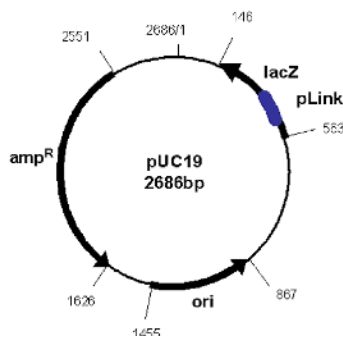


Figure 6.6.1: pUC Vector

The "start" of the sequence is basepair number 1, and is oriented at 12:00 (straight up). The sequence then runs clockwise around the plasmid. The plasmid diagram will use arrows to indicate the locations of the various features of the DNA, and the basepair numbers associated with the start and end locations of such features is also typically given on such figures.

This page titled [6.6: Use of PC and internet for biochemical research](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Michael Blaber](#).

Index

C

central dogma

[2.4: Central Dogma and the Genetic Code](#)

E

E. coli

[6.1: Genetic Transformation \(using bacteria and the pGLO plasmid\)](#)

enzyme induction

[2.5: Gene and Operon](#)

enzyme repression

[2.5: Gene and Operon](#)

ethidium bromide

[3.1: Gel Electrophoresis](#)

G

galactose containing media

[4.5: Protein-Protein Recognition Probed Using a Yeast Transcriptional Activator System](#)

gel electrophoresis

[3.1: Gel Electrophoresis](#)

gene

[2.5: Gene and Operon](#)

genetic code

[2.4: Central Dogma and the Genetic Code](#)

I

introns

[2.5: Gene and Operon](#)

K

kinase

[1.4: DNA Modifying Enzymes](#)

L

Ligase

[1.4: DNA Modifying Enzymes](#)

M

M13 Phage

[4.2: M13 Phage](#)

Methylases

[1.4: DNA Modifying Enzymes](#)

methylation

[1.4: DNA Modifying Enzymes](#)

Molecular Imprinting

[4.6: Molecular Imprinting](#)

N

Nucleases

[1.4: DNA Modifying Enzymes](#)

O

operons

[2.5: Gene and Operon](#)

P

PCR

[3.3: Polymerase Chain Reaction \(PCR\) and Cloning of PCR Products](#)

pGLO plasmid

[6.1: Genetic Transformation \(using bacteria and the pGLO plasmid\)](#)

Phage Display Libraries

[4.3: M13 Phage Display Libraries](#)

Phosphatase

[1.4: DNA Modifying Enzymes](#)

polymerase chain reaction (PCR)

[6.5: Polymerase Chain Reaction \(PCR\)](#)

Polymerases

[1.4: DNA Modifying Enzymes](#)

R

Restriction Mapping

[6.4: Restriction Mapping](#)

S

SELEX

[4.4: SELEX \(Selective Evolution of Ligands by Exponential Enrichment\)](#)

T

transcriptional activator protein

[4.5: Protein-Protein Recognition Probed Using a Yeast Transcriptional Activator System](#)

transformation

[6.1: Genetic Transformation \(using bacteria and the pGLO plasmid\)](#)

translation

[2.8: mRNA Translation](#)

U

Upstream Activator Sequences (UAS's)

[4.5: Protein-Protein Recognition Probed Using a Yeast Transcriptional Activator System](#)

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Supplemental Modules (Biochemistry)

Webpages: 54

All licenses found:

- **Undeclared:** 100% (54 pages)

By Page

- Supplemental Modules (Biochemistry) - *Undeclared*
 - Front Matter - *Undeclared*
 - TitlePage - *Undeclared*
 - InfoPage - *Undeclared*
 - Table of Contents - *Undeclared*
 - Licensing - *Undeclared*
 - 1: DNA - *Undeclared*
 - 1.1: DNA as Genetic Material - *Undeclared*
 - 1.2: Structure of DNA and RNA - *Undeclared*
 - 1.3: Bacterial Restriction/Modification system - *Undeclared*
 - 1.4: DNA Modifying Enzymes - *Undeclared*
 - 1.5: DNA Replication - Introduction to Prokaryotic replication - *Undeclared*
 - 1.6: DNA Supercoiling and Topoisomerases - *Undeclared*
 - 2: Bacteria - *Undeclared*
 - 2.1: The Bacteria - An Introduction - *Undeclared*
 - 2.2: Bacterial Growth and Reproduction - *Undeclared*
 - 2.3: Extrachromosomal elements, plasmids, selectable markers - *Undeclared*
 - 2.4: Central Dogma and the Genetic Code - *Undeclared*
 - 2.5: Gene and Operon - *Undeclared*
 - 2.6: The lac Operon; CAP site; DNA footprinting - *Undeclared*
 - 2.7: Transcriptional Regulation: Transcription termination, the trp operon - *Undeclared*
 - 2.8: mRNA Translation - *Undeclared*
 - 3. Biotechnology 1 - *Undeclared*
 - 3.1: Gel Electrophoresis - *Undeclared*
 - 3.2: DNA Sequence Analysis - *Undeclared*
 - 3.3: Polymerase Chain Reaction (PCR) and Cloning of PCR Products - *Undeclared*
 - 3.4: Prokaryotic Expression Vectors - *Undeclared*
 - 3.5: Protein Sequencing, Peptide Mapping, Synthetic Genes - *Undeclared*
 - 3.6: cDNA and Genomic Libraries - *Undeclared*
 - 4. Biotechnology 2 - *Undeclared*
 - 4.1: Protein Purification - *Undeclared*
 - 4.2: M13 Phage - *Undeclared*
 - 4.3: M13 Phage Display Libraries - *Undeclared*
 - 4.4: SELEX (Selective Evolution of Ligands by Exponential Enrichment) - *Undeclared*
 - 4.5: Protein-Protein Recognition Probed Using a Yeast Transcriptional Activator System - *Undeclared*
 - 4.6: Molecular Imprinting - *Undeclared*
 - 5. Lab Notes Part 1 - *Undeclared*
 - 5.1: Basic practices and techniques, Laboratory notebooks, Presentation of data, Sections of a scientific report - *Undeclared*
 - 5.2: Spectrophotometry - *Undeclared*
 - 5.3: Quantification of Protein Concentration - *Undeclared*
 - 5.4: Chromatography - *Undeclared*
 - 5.5: Gel Electrophoresis of Proteins - *Undeclared*
 - 5.6: Introduction to Protein Purification - *Undeclared*
 - 6. Lab Notes Part 2 - *Undeclared*
 - 6.1: Genetic Transformation (using bacteria and the pGLO plasmid) - *Undeclared*
 - 6.2: Enzyme kinetics - *Undeclared*
 - 6.3: Ligand binding - *Undeclared*
 - 6.4: Restriction Mapping - *Undeclared*
 - 6.5: Polymerase Chain Reaction (PCR) - *Undeclared*
 - 6.6: Use of PC and internet for biochemical research - *Undeclared*
 - Back Matter - *Undeclared*
 - Index - *Undeclared*
 - Glossary - *Undeclared*
 - Detailed Licensing - *Undeclared*