

Introduction to Molecular Biology

INTRODUCTION TO MOLECULAR BIOLOGY

For use in BIOL-L211 Indiana
University Bloomington.

SAPNA MEHTA



Introduction to Molecular Biology by Sapna Mehta is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/), except where otherwise noted.

This book was written, compiled and curated from ***Creative Commons with Attribution Licenses***. Chapter level licenses are provided. Many of the resources are under the CC-BY-NC-SA and therefore reproduction of any of the material in this book should take into consideration the original source.

This book was produced with Pressbooks (<https://pressbooks.com>) and rendered with Prince.

CONTENTS

Preface	xv
<i>About this Book</i>	xv
<i>To the Instructor</i>	xvi
Course Context	1
<i>“Why is the course taught this way, and what will I get out of this class?”</i>	1
Learning from this book	3

Part I. [Molecular Biology: From DNA to RNA to Protein](#)

Themes from Intro Bio	9
<i>Introduction</i>	9
<i>Gene Expression</i>	15
<i>Protein Synthesis Overview</i>	17
<i>Attributions</i>	20

1. Protein Structure and Function	22
1.1 <i>Introduction: From Computer to the Clinic and Beyond</i>	24
1.2 <i>Building Blocks of Proteins</i>	28
1.3 <i>Levels (Orders) of Structure</i>	33
1.4 <i>Protein Structure Deep Dive</i>	50
1.5 <i>Domains in Protein Structure.</i>	57
1.6 <i>Protein Modifications can affect structure and function</i>	59
1.7 <i>Intrinsically Disordered Proteins</i>	62
1.7 <i>X-Ray Crystallography: Art Marries Science</i>	65
<i>References and Attributions</i>	70
2. Methods to Study Proteins	72
2.1 <i>Introduction: Molecules to Medicine</i>	72
2.2 <i>Biochemical Assays</i>	76
2.3 <i>Isolating Proteins from Complex Mixtures</i>	81
2.4 <i>Column Chromatography</i>	85
2.5 <i>SDS-PAGE and Western Blotting</i>	99
<i>References and Attributions</i>	113

3. Nucleic Acids, Identity of DNA as molecule of inheritance	115
3.1 <i>Introduction: The Stuff of Genes</i>	115
3.2 <i>Chemistry of Nucleic Acids</i>	119
3.3 <i>Identity of DNA as Genetic Material</i>	125
3.4 <i>DNA Structure: The Double Helix</i>	136
3.5 <i>Analysis of Nucleic Acids</i>	144
3.6 <i>From Genes to Genomes</i>	150
<i>References and Attributions</i>	159
4. DNA Packaging in Eukaryotes	162
4.1 <i>Introduction: “Inner Life of the Genome”</i>	162
4.2 <i>Types of Chromatin in Eukaryotic Cells</i>	166
4.3 <i>Chromatin Organization</i>	168
4.4 <i>Regulation of Chromatin Structure</i>	181
4.5 <i>Links to Medicine</i>	186
<i>References and Attributions</i>	188

5. DNA Replication	191
5.1 Introduction	191
5.2 The Basic Rules of Replication	195
5.3 Structure of DNA Polymerase	215
5.4 Process of Replication	219
5.4.1 Stages of Replication	221
5.5 Eukaryotic Replication	241
References and Attributions	255
6. Transcription in Prokaryotes	257
6.1 Introduction	257
6.2 Basics of Transcription	263
6.3 Rules of Transcription and Terminology	267
6.4 Genes are Transcription Units	273
6.5 RNA Polymerase Enzymes	278
6.6 Steps in Transcription	286
References and Attributions	296

7. Regulation of Gene Expression - Prokaryotes	297
	297
7.1 Introduction	297
7.2 Overview of Regulation of Gene Expression	302
7.3 Gene Regulation in Prokaryotes	310
7.4 The lac Operon- An example regulation of bacterial gene expression	320
References and Attributions	338

8. Eukaryotic Transcription and Regulation	340
	340
<i>Learning Objectives</i>	340
<i>8.1 Introduction</i>	342
<i>8.2 Eukaryotic Cells Have Three Types of RNA Polymerase</i>	343
<i>8.3 Overview of Gene Expression (From DNA to Protein)</i>	345
<i>8.4 Details of Eukaryotic Transcription Initiation</i>	348
<i>8.5 How Transcription factors Work</i>	362
<i>8.5 Bringing it all together</i>	364
<i>8.6 Relevant Biological Concepts</i>	365
<i>8.7 Transcription Elongation and Termination- mRNA Processing</i>	369
<i>References and Attributions</i>	374

9. mRNA Splicing and Alternative Splicing	376
9.1 Introduction	378
9.2 The Split Gene	379
9.3 A Detailed Look at mRNA Splicing	383
9.4 Alternative Splicing	393
9.5 Clinical Insight	401
	405
Before you continue	405
References and Attributions	406
10. Genetic Code and Translation	408
10.1 Overview of Translation	408
10.1.2 Overview of Genetic Code	412
10.2 tRNA's: The Interpreter of the Code	425
10.3 Ribosome Structure	433
10.4 Details of Translation	441
10.5 Elongation and Termination	452
References and Attributions	456

Part II. Tools and Techniques of Molecular Biology

Introduction	461
<i>Introduction</i>	461
Methods of Molecular Genetic Analysis based on DNA replication process	463
1. <i>Methods of Molecular Genetic Analysis based on the DNA replication process</i>	463
1.1 <i>Polymerase Chain Reaction</i>	465
1.2 <i>Variations of PCR</i>	474
1.4 <i>DNA Sequencing</i>	486
<i>References and Attributions</i>	492

Recombinant DNA Technology	494
<i>Introduction</i>	496
<i>Molecular Cloning and Recombinant DNA Technology</i>	497
<i>References and Attributions</i>	520

PREFACE

About this Book

Introduction to Molecular Biology is a textbook curated from open-source materials for BIOL-L211- a core course required for students majoring in biology at Indiana University Bloomington. Students enrolled in this course have completed college-level introductory biology and chemistry. The text is tailored for a flipped/hybrid format and serves as the primary source of learning material before a class meeting focused on problem-solving. The book closely reflects the structure of the course with Molecular Biology in the News interwoven with content.

Following the principles of Universal Design for Learning, multiple means of representation are provided for students to engage with the content. Links to outside content for further exploration, instructor-endorsed videos, and animations have been paired with the text.

This book contains material from several Open educational resources (OER) with CC-BY, CC-BY-SA or CC-BY-NC-SA licensed content.

Material has been curated, modified, reworded, re-ordered,

and combined to create a text uniquely aligned with the course and for my students.

I would like to thank all of the authors of the open-source texts that allowed for derivatives to be created. Chapter-level licenses and attributions have been provided for those wishing to reuse this text.

This book would not have been possible without Indiana University [Course Material Transformation Fellowship Program](#) OER Award Program spearheaded by Sarah Hare Scholarly Communication Librarian.

Undoubtedly the in-class experience of using it in my teaching and feedback from some of the many hundreds of students that take this course will be critical, as it was for creating this resource in the first place.

To the Instructor

This document is intended to be a living text and will be updated organically as time permits. Instructors may notice some of the topics typically discussed within a molecular course are **not** included- such as DNA Damage and Repair, the role of regulatory RNA, and Genome Editing. Future editions will include this material, additional interactive content, and suggestions for case study teaching using molecular biology in the news examples.

In keeping with the core focus of the course on principles

over facts, there isn't a huge emphasis on definitions. Therefore a glossary has not been created at this time. In addition, there is an intentional simplification of specific topics and a deliberate omission of gene/protein names for some processes.

A list of key terms is provided to the students separately and terms or words that students should get familiar with are highlighted in the text.

As noted above, most pages have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) provided at Chapter Level before doing so.

My goal in releasing it as an open access resource is to make it easier for colleagues at other colleges and universities to create versions of their own as well as provide suggestions and feedback.

Offers for collaboration , comments, critiques, and requests for access to problems/ learning activities used in class are welcome! Please reach out at sapmehta@iu.edu.

COURSE CONTEXT

“Why is the course taught this way, and what will I get out of this class?”

This course is going to be very different from other biology classes you might have taken. It is geared towards engaging creative thinking and problem solving to give students a feeling of how biology research takes place.

I have designed the course to maximize interaction with the material, each other and to promote metacognitive tools to learn better.

The teaching team of graduate associate instructors and undergraduate teaching associates are here to assist you throughout the semester.

In addition, we will be using learning groups to facilitate collaborative thinking. You learn best from one another!

It is critical for students in biology or related fields to develop a strong conceptual foundation and to demonstrate their ability to use it in contexts that may be novel to them.

We expect you to go beyond “fact-learning” and use what we teach you to solve problems that require critical thought.

- Students in L211 will be asked to begin developing the ability to identify and articulate the key scientific and biological questions that are at the core of the course content.
- Students will be expected to learn and use correct technical vocabulary in their discussions of course content.
- Students will be expected to begin conceptualizing course content from a question-driven and problem-solving perspective.

When you have time to learn and practice the “facts” outside of class, we can use class time covering the complex ways of thinking about molecular biology and practice the higher-order levels of understanding.

Successful completion of this class will equip you with **the set of tools** necessary to master upper-level courses – not only in biology.

LEARNING FROM THIS BOOK

This book was made especially for you, the student. It arose out of observations and feedback from several semesters of teaching, and more recently during the online semesters of 2020 -2021.

1. Some students liked to have a reference, but textbooks are prohibitively expensive, dense, and not always aligned with what is taught in class.
2. Many students prefer to watch quick videos (often Khan Academy) and often seek out resources online.

Unfortunately, many of them (*and yes even some Khan Academy ones!*) are inaccurate, propagate misconceptions, or are not appropriate for the level of the class.

So I did the work for you.

As you read you will see embedded videos, animations, and additional links to learning. These deliver the information and content as it pertains to this class. You will also have access to lecture videos for some topics.

I hope this digital text will be a happy medium between the two options.

Make sure you have exhausted all the options (Links to Learning, watching animations, and other videos provided). If after doing all of that you still find yourself looking for alternative explanations- please seek out help before looking online for other resources!

Avoid:

⊗ Only watching the videos and not reading the text.

⊗ Only reading the text and not watching any of the linked resources.

Because

3. Cognitive science and studies of how our brain process information has shown *that the more ways* you look at information the better it is for long-term retention.

Hopefully by having multiple options *within one text* will provide for a seamless experience.

How is this book organized?

In each chapter, you will see the following features to guide you:

- **Learning Objectives** guide you through what you can expect to learn by reading and completing the chapter exercises. ALL exam and quiz questions are paired with these learning objectives! **You should take these very seriously.**
- **Exercises** allow you an immediate opportunity to put new information into practice.
- **Links to Learning** sections provide you with opportunities to continue exploring the concepts you are learning through connections to other helpful information.

My recommendation

During your established study time set aside for each ‘Lecture’ as per CANVAS consider doing the following:

1. Print out the active learning notes. Read the learning objectives, glance at the problem set questions.
2. Begin reading the assigned sections of the book – pause and watch the lecture videos *if instructed within the text*. If you think you have a good grasp you can watch them at increased speed!
 - Print out (or have available a digital form) ‘Lecture

Video Slides' when made available and keep them nearby to make any notes or look at the figures.

3. Fill in the active learning notes as you make your way through the material (try to do it **without looking** at the text first!)
4. Use the List of Terms to make concept maps connecting as many as you can.
5. Complete any associated assessments or practice in your LMS (Learning management system- CANVAS).
 - Treat the assessment as a **QUIZ** to **test your understanding** of the factual knowledge.
 - Take note of where you are guessing, what you got wrong so you can review.

This book is will continue to change. Your experience *using* the book and suggestions will be valuable as I continue to modify chapters!

PART I

MOLECULAR BIOLOGY: FROM DNA TO RNA TO PROTEIN

THEMES FROM INTRO BIO

Introduction

“Consider just three of Earth’s inhabitants: a bright yellow daffodil that greets the spring, the single-celled creature called *Thermococcus* that lives in boiling hot springs, and you. Even a science-fiction writer inventing a story set on a distant planet could hardly imagine three more different forms of life.

Yet you, *Thermococcus* and the daffodil are related! Indeed, all of the Earth’s billions of living things are kin to each other!”

Underlying this cellular diversity is biochemical unity. All cells comprise of molecules that encode information and can be copied—the nucleic acids DNA and its simpler relative, RNA.

Proteins—workhorse molecules that perform important tasks. And encapsulating them all, there’s a membrane made from fatty acids.

From the Genetic Science Learning Center,

University of Utah, <http://learn.genetics.utah.edu>
take a look at video highlighting the unity of life:
[Shared Functions, Shared Genes](#)

Below is a refresher into some themes that you encountered in earlier introductory classes. Throughout the semester we will explore these in greater detail as well as some unfamiliar but exciting new topics.

What is a Gene?

The gene is the basic physical unit of inheritance. Genes are passed from parents to offspring and contain the information needed to specify traits. Genes are arranged, one after another, on structures called chromosomes. A chromosome contains a single, long DNA molecule- only a portion of which corresponds to a single gene- as well as the structural proteins (called histones) that the DNA molecule wraps around. Humans have approximately 20,000 genes arranged on their chromosomes.

Watch the following brief video for an animated view of the relationship between chromosomes and genes.



An interactive H5P element has been excluded from this version of the text. You

can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1098#h5p-34>

We will learn more about replication in Chapter 5

DNA and RNA

The two main types of nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). As described earlier in this chapter, DNA is the genetic material in all living organisms, ranging from single-celled bacteria to multicellular mammals.

It is in the nucleus of eukaryotes and in the organelles mitochondria and chloroplasts. In prokaryotes, the DNA is not enclosed in a membranous envelope.

The ***cell's entire genetic content is its genome***, and the study of genomes is genomics. In eukaryotic cells but not in prokaryotes, a DNA molecule may contain tens of thousands

of genes. Many genes contain information to make protein products (e.g., mRNA). Other genes code for RNA products.

DNA controls all of the cellular activities by turning the genes “on” or “off.”

The other type of nucleic acid, RNA, is mostly involved in protein synthesis. ***The DNA molecules never leave the nucleus*** but instead use an intermediary molecule to communicate with the rest of the cell.

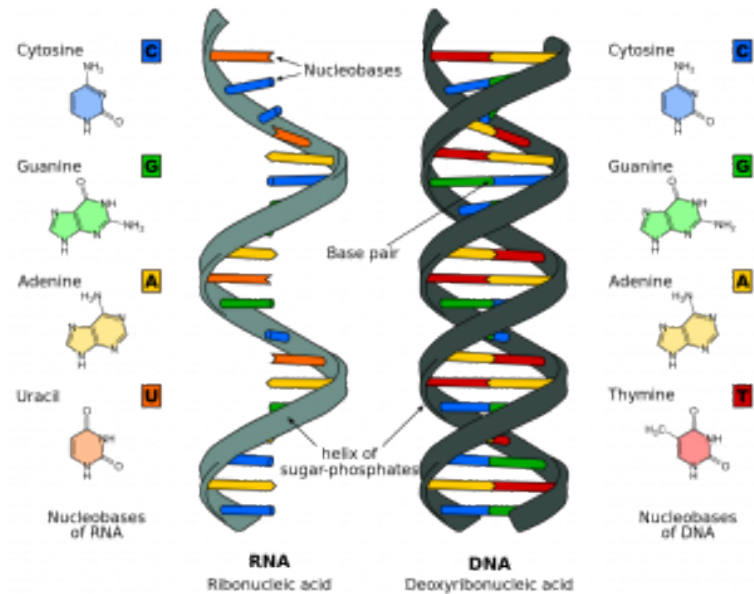
This ***intermediary is the messenger RNA (mRNA)***.

Other types of RNA—like rRNA, tRNA, and microRNA—are involved in protein synthesis and its regulation but do not carry code for proteins (they are non-coding RNAs)

DNA and RNA are comprised of ***monomers*** that scientists call ***nucleotides***. The nucleotides combine with each other to form a ***polynucleotide***, DNA or RNA.

Three components comprise each nucleotide: a nitrogenous base, a pentose (five-carbon) sugar, and a phosphate group. Each nitrogenous base in a nucleotide is attached to a sugar molecule, which is attached to one or more phosphate groups.

Therefore, although the terms “base” and “nucleotide” are sometimes used interchangeably, a nucleotide contains a base as well as part of the sugar-phosphate backbone.



Comparison of the molecular structure of RNA and DNA.

Comparison of RNA (left molecule) and DNA (right molecule). The color of the bases in RNA and DNA aligns with the colored boxes next to each base molecule.

Exercises

Examine the image above and then answer the following questions:



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1098#h5p-35>



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1098#h5p-36>



An interactive H5P element has been excluded from this version of the text. You can view it online here:

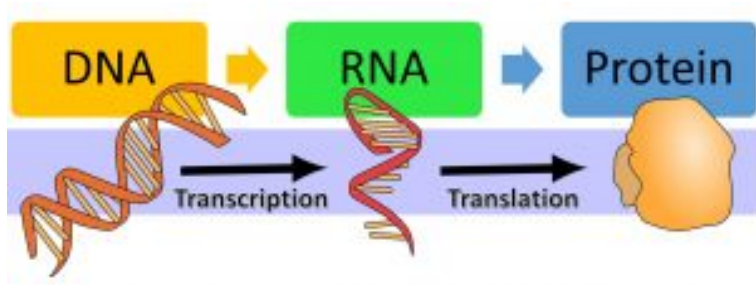
[https://iu.pressbooks.pub/
iul211smehta/?p=1098#h5p-37](https://iu.pressbooks.pub/iul211smehta/?p=1098#h5p-37)

Gene Expression

The mechanism for how information coded in DNA is ‘brought to life’ or ‘expressed’ into proteins is often referred to as the Central Dogma of life.

It involves 2 processes: Transcription and Translation.

Because proteins are coded by genes, the term “***gene expression***” refers to protein synthesis (i.e., making proteins), including the ***regulation*** of that synthesis.



The central dogma states that DNA is used to make RNA via transcription, which is used to make protein via translation.

Transcription the first step in gene expression results in production of RNA (functional RNA that do not code for proteins) and mRNA (code for proteins) from DNA.

Note that DNA never “becomes” RNA; rather, the DNA is “read” to make an RNA copy.

In eukaryotic cells, the mRNA leaves the nucleus and then, through the process of **translation**, the mRNA is read to create an amino acid sequence that folds into a protein.

Consider what the terms “transcribe” and “translate” mean in relation to language. To “transcribe” something means to rewrite text again in the same language while to “translate” something means to rewrite the text in a different language.

Similar to these meanings, in biology, **DNA is transcribed into RNA: both DNA and RNA are made of nucleic acid** (i.e., the same “language”). With the assistance of proteins, DNA is “read” and transcribed into an mRNA sequence.

To read RNA and create protein, though, we refer to it

as being translated: *RNA is made of nucleic acid, and protein is made of amino acids (i.e., different “languages”)*. Therefore, DNA is transcribed to create an mRNA sequence, and then the mRNA sequence is translated to make a protein.

Protein Synthesis Overview

The following is an overview of each of these processes.

Each process will be described in more detail in future chapters.

Transcription

A gene is complex: it contains not only the code for the resulting protein but also several regulatory factors that determine if and when the region that codes for a protein are read to create protein.

What follows is a diagram of the components of a gene that are used in transcription.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1098#h5p-38>



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1098#h5p-39>

Exercise

Given a specific DNA strand, what is the sequence of the resulting mRNA molecule? We will learn about how mRNA is created in a later chapter.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1098#h5p-40>

Translation

Translation involves different types of RNA, and we will explain them in more detail in later chapters: rRNA, tRNA, mRNA, and microRNA.

After an mRNA is created, it leaves the nucleus and is attracted to or attracts a ribosome, which is a molecule made of rRNA and polypeptides. Then, in the ribosome, and with the assistance of tRNAs, the mRNA is read and an amino acid sequence is created.

DNA and mRNA create sequences with just four types of bases; yet, these bases code for 20 unique amino acids (the makeup of protein). How is this possible? Watch the following video to find out!



An interactive H5P element has been excluded from this version of the text. You

can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1098#h5p-41>

Regulation of Gene Expression

As we shall see throughout the semester, our cells resort to a number of ways and resources in controlling gene expression in space and time- what genes are expressed, when they are expressed and how robust is the expression. Regulation may occur at any point in the expression of a gene, from the start of the transcription phase of protein synthesis to the processing of a protein after synthesis occurs.

This tightly controlled regulation is key to development as well as the smooth functioning of an already developed organism.

Attributions

This chapter is a modified derivative of the following articles:

“[Gene](#)” by National Human Genome Research Institute,

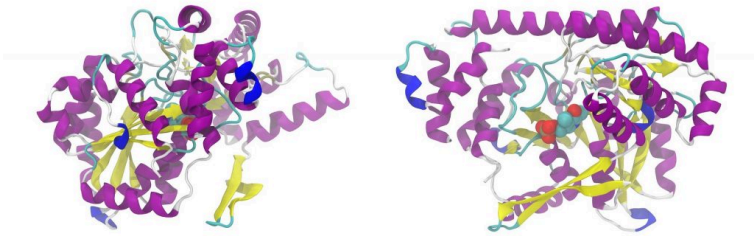
National Institutes of Health, *Talking Glossary of Genetic Terms*.

“Nucleic Acids” by OpenStax College, [Biology 2e](#), CC BY 4.0. Download the original article at <https://openstax.org/books/biology-2e/pages/3-5-nucleic-acids>

Genetic Science Learning Center. (2017, August 1) Shared Functions, Shared Genes. Retrieved December 16, 2021, from <https://learn.genetics.utah.edu/content/evolution/sharedfunctions>

1.

PROTEIN STRUCTURE AND FUNCTION



Learning Objectives

When you have mastered the information in this chapter, you should be able to:

Level 1 and 2 (Knowledge and Comprehension)

- Draw and label the chemical structure of single amino acid, a dipeptide, a tripeptide and identify the various components and bonds (especially peptide bonds)
- Classify the amino acid side chains as polar and charged, polar and uncharged, hydrophobic, or special.
- Predict the types of non-covalent interactions possible between any two amino acids.
- Define and distinguish between the orders of protein structure.
- Differentiate between α -sheet, α -helix, and 'random coil' structures based on the atomic interactions involved on each.
- Describe how globular proteins arise from the hydrophobic and hydrophilic interactions that drive protein folding.



Level Up (Application, Analysis, Synthesis)

- Formulate a hypothesis to explain why the amino acid glycine is a disruptor of alpha-helical polypeptide structure.
- Identify the role of disulfide bonds in protein folding, articulate why secreted proteins often contain disulfide bonds.
- Propose a likely function or property for a protein or protein domain based upon a given three-dimensional structure.
- Within a given set of parameters, predict *modified or substituted amino acids affect* the structure of a protein and if that it will alter the function of that protein.

Before you begin make sure you are thinking about answers to the learning objectives. Level 1 and Level 2 form the foundation. Level Up will be the target for assessments.

1.1 Introduction: From

Computer to the Clinic and Beyond

“Imagine that you are a scientist probing the secrets of living systems not with a scalpel or microscope, but much deeper — at the level of single molecules, the building blocks of life.

You’ll focus on the detailed, three-dimensional structure of biological molecules. You’ll create intricate models of these molecules using sophisticated computer graphics. You may be the first person to see the shape protein offers clues about the role it plays in the body. It may also hold the key to developing new medicines, materials, or diagnostic procedures. You are part of the growing field of structural biology.” (1)

The molecules whose shapes most tantalize structural biologists are **proteins**. Virtually everything that goes on inside of cells happens as a result of the actions of proteins. Nature has programmed proteins to do **nearly every job in the body**:

- protein enzymes catalyze the vast majority of cellular reactions,
- proteins mediate signaling,
- give structure both to cells and to multicellular organisms and
- proteins exert control over the expression of genes.

Life, as we know it, would not exist if there were no proteins.

Therefore, it is not surprising that diseases at the molecular level are often due to the malfunction of these proteins.

Like many everyday objects, proteins are shaped to get their job done. The shape or structure of a protein offers clues about the role it plays in the body. It also holds the key to developing new medicines, materials, or diagnostic procedures. As a result, efforts to solve, predict and modify protein structures have been central to therapeutic science and the development of lifesaving and life-altering medicines for often debilitating symptoms.

Some examples include injectable insulin, antiretroviral therapy for AIDs, Celebrex a drug to treat arthritis to name just a few.

There are a variety of experimental techniques for solving protein structures- nuclear magnetic resonance, X-ray crystallography, and cryo-electron microscopy. However, to date, scientists only know the structure of a tiny fraction of all the known proteins. For long scientists have grappled with 2 research problems.

1. The “protein folding problem” – the general problem of predicting protein structure directly from the amino acid sequence.
2. The Protein Design problem- starting with the desired shape – something brand new and finding the amino acid sequence that can adopt that shape.

In recent years there has been a revolution in computational methods for both predicting how proteins fold based on knowledge of their amino acid sequence **AND** creating **new designer proteins from scratch!**

Watch the following TED talk by Dr. David Baker, named as one of the world's most influential scientists talks about his work, and why it's important to create new proteins.

Molecular Biology in the News: “Artisanal Proteins”

WATCH: TED Talk by Dr. David Baker.

Link here: <https://youtu.be/PJLT0cAPNfs> or directly embedded below.



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul21smehta/?p=43#oembed-1>

READ: [The biochemist engineering proteins from scratch](#)

NOTE: Your CANVAS course site includes .pdfs of this article above.

In order to understand how scientists can design proteins OR predict the structure of proteins from a given amino acid sequence, we need to understand the rules of protein folding.

In this chapter, we first dive into how a protein adopts its shape. We then look at X-ray crystallography one of the methods for obtaining atomic level information about protein shape. **In Chapter 2** you will learn how one can exploit the differences in proteins to separate a given protein from a complex mixture inside cells.

First, let's review what proteins are made of.

1.2 Building Blocks of Proteins

The building blocks of all proteins are amino acids. The sequence of amino acids in individual proteins is encoded in the DNA of the cell. All amino acids have the same basic structure, which is shown in Figure 1.1

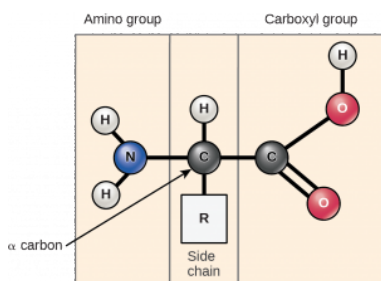


Figure 1.1 Amino acid structure.

At the “center” of each amino acid is a carbon called the α -carbon and attached to it are four groups – hydrogen, an α -carboxyl group, an α -amine group, and an R-group, often referred to as a **side chain**.

The α -carbon, carboxyl, and amino groups are common to all amino acids, so the R-group is the only unique feature in each amino acid. (A minor exception to this structure is that of proline, in which the end of the R-group is attached to the α -amine.)

With the exception of glycine, which has an R-group consisting of a hydrogen atom, all of the amino acids in proteins have four different groups attached to them and consequently can exist in two mirror-image forms, L and D. With only very minor exceptions, every amino acid found in cells and in proteins is in the L configuration.

Cells use only 20 amino acids to make polypeptides and proteins (these are specified by the genetic code), although they do use a few additional amino acids for other purposes.

The 20 amino acids have diverse properties which are determined by the chemistry of the side chains. If you compare groupings of amino acids in different textbooks, you will see different names for the categories and (sometimes) the same amino acid being categorized differently by different authors.

One such grouping based on polarity and charge is shown in Figure 1.2 below.

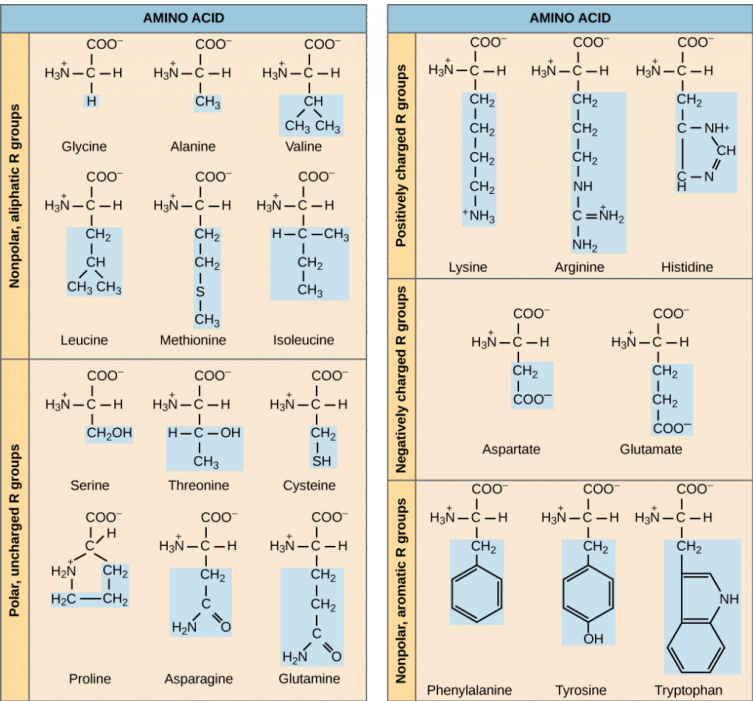


Figure 1.2. Chemical characteristics of the 20 amino acids found in the proteins of cells.

Note that some amino acids have ionizable side chains in addition to the They can give up a proton or take a proton. This ability is dependent on the pH! The amino acids in the figure are shown at physiological pH 7.0.

See here for a refresher on Acid, Base, pH, and pKa. pKa **may be** a new term. For this course, you must understand the relationship between the pH of a solution and its influence on the charge on an amino acid. **We will refer to pKa in Chapter 2.**



An interactive H5P element has been excluded from this version of the text.

You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=43#h5p-2>

Classifying amino acids by polarity is important to learn how to do- because polarity affects which non-covalent interactions amino acids can form- which in turn determines protein shape.

TRY THIS

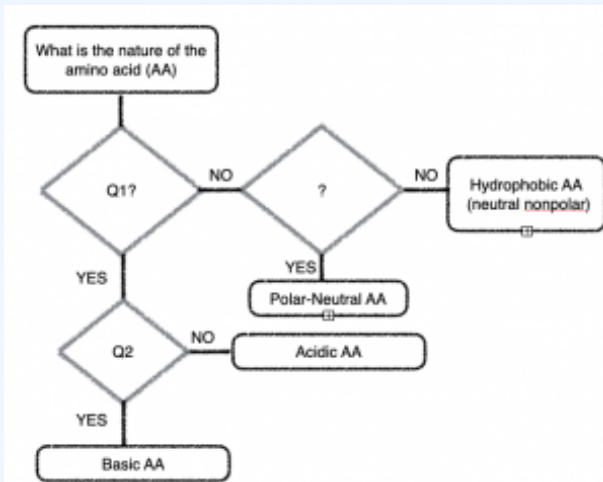
You should be able to infer the properties of the side chain from the 2D chemical diagram and the 3D structure. For example, which amino acids have polar sidechains? Which have planar aromatic groups?

You can do this without memorizing!

- When given a structural formula for an amino acid, you can determine its type by **asking three simple questions!**

First: Study the different groups of amino acids shown in Figure 1-2. We want to classify them as Acidic, Basic, Polar Neutral, and Non-polar Neutral. What features did you look for?

Come up with the 3 simple “yes or no” questions you could ask based on your analysis to plug into the flowchart.



Hint: Think about the chemistry of acids and bases (what do the acidic side chains look like?), Think about what makes a covalent bond a 'polar covalent bond'

Ask me in class, or ask your classmates to see if you came up with similar questions.

1.3 Levels (Orders) of Structure

Now that we are familiar with the building blocks of proteins,

we dive into how strings of amino acids twist and buckle, folding in upon themselves, the knobs of some amino acids nestling into grooves in others.

We shall examine protein structure at four distinct levels (Figure 1.3)

- 1) how the sequence of the amino acids in a protein (primary structure) gives identity and characteristics to a protein;
- 2) how local interactions between one part of the polypeptide backbone and another affect protein shape (secondary structure);
- 3) how the polypeptide chain of a protein can fold to allow amino acids to interact with each other that are not close in primary structure (tertiary structure); and
- 4) how different polypeptide chains interact with each other within a multi-subunit protein (quaternary structure)

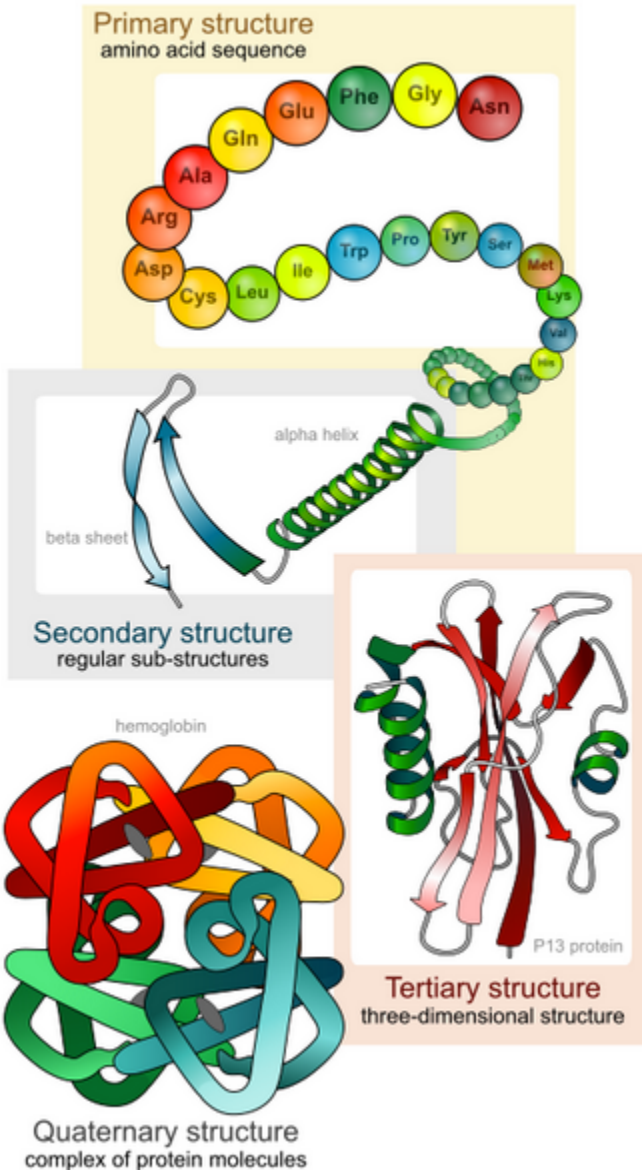


Figure 1.3. The four orders (levels) of protein structure. Primary, secondary and tertiary structures describe polypeptides; quaternary structure applies to proteins

composed of 2 or more polypeptides. (“Main protein structure levels en” by LadyofHats is in the Public Domain)

It is important to note that for each polypeptide the final structure (**native fold**) is the tertiary structure. Quaternary structure refers to associations *of two or more polypeptides, creating higher-order protein structures*. Superimposed on these basic levels are other features of protein structure. These are created by the specific amino acid configurations in the mature, biologically active protein.

Some clarifications of confusing terminology you may encounter!

We use the term **polypeptide** to refer to a single polymer of a long stretch of amino acids—the translation product of a gene. It may or may not have folded into its final, functional form.

The term **protein** is sometimes used *interchangeably* with **polypeptide**, as in “protein synthesis”. It is generally used, however, to refer to a folded, functional molecule. As we

shall see a protein may be made up of MORE than ONE polypeptide. For example, Hemoglobin is a **protein**– but it is made up of 4 separate polypeptides that come together to adopt a final shape.

1.3.1 Primary Structure

The specific order of amino acids in a protein is known as its ***primary structure***. Inside a cell synthesis of proteins occurs in the ribosomes and proceeds by joining the carboxyl terminus of the first amino acid to the amino terminus of the next one (Figure 1.4).

Note the chemistry of the reaction! Individual amino acids are joined together by the attack of the nitrogen of an amino group of one amino acid on the carbonyl carbon of the carboxyl group of another to create a covalent peptide bond and yield a molecule of water.

This is a condensation reaction or dehydration reaction that is common in the making of all biological polymers.

Amino acids that are part of a protein are referred to as **'amino acid residue(s)'**. You will read and hear this term being utilized often.

Because the synthesis takes place from the alpha-amino group of one amino acid to the carboxyl group of another amino acid, the result is that there will always be a free amino group on one end of the growing polymer (the amino or N-terminus) and a free carboxyl group on the other end (the carboxyl or C-terminus).

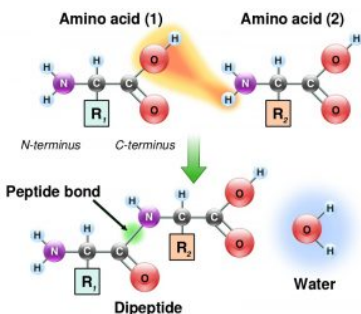


Figure 1.4. Linking of amino acids through peptide bond formation. (" Peptide bond formation by YassineMrabetTal. This W3C-unspecified vector image was created with Inkscape ., Public domain, via [Wikimedia Commons](#))

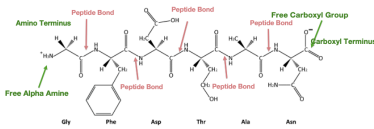


Figure 1.5. Simple view of polypeptide. Image credit: Ahern K, Rajagopal I and Tan T. (2013). *Biochemistry Free for All* (Version 1.3). Licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).

Since proteins are synthesized starting with the amino terminus and ending at the carboxyl terminus, therefore by convention amino acid sequences are written left to right from amino to carboxy-terminus. The name of the N-terminal residue is always the first amino acid. The name of each amino acid then follows.

Reversing the directionality indicates a very different protein sequence!

It is the order of amino acids that dictates the 3-D conformation (shape) the folded protein will have. This conformation, in turn, will determine the function of the protein.

Practice Exercises



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=43#h5p-3>

1.3.2 Secondary Structure

As protein synthesis progresses, interactions between amino acids close to each other begin to occur, giving rise to local patterns called secondary structures. Two common types of secondary structures in proteins are alpha (α) helices and beta (β) strand/sheets. (Figure 1.6)

Secondary structure conformations occur due to the spontaneous formation of hydrogen bonds between amino groups and oxygens along the polypeptide backbone.

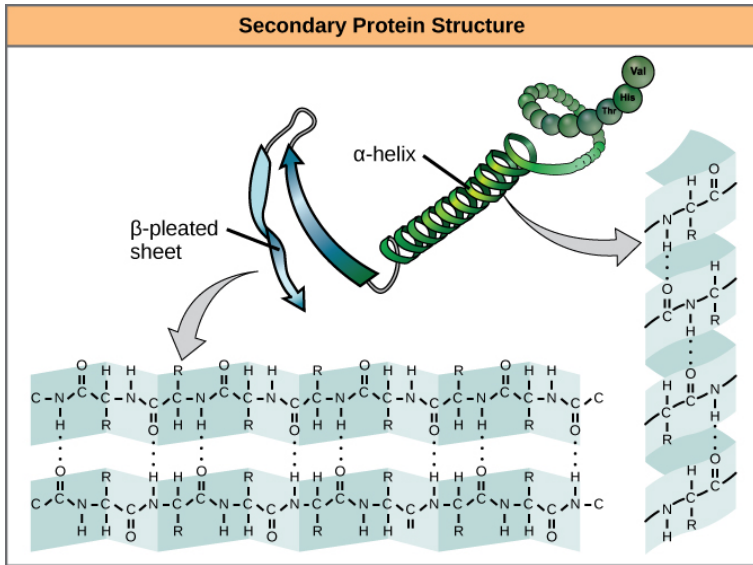


Figure 1.6. In the secondary structure of a polypeptide, more organized alpha-helical and beta-pleated sheet structures are separated by less organized, random coil stretches of amino acids.

α -helix

In the α -helix, hydrogen bonds form between $\text{C}=\text{O}$ groups and $\text{N}-\text{H}$ groups within the polypeptide backbone that is four amino acids distant. These hydrogen bonds are the primary forces stabilizing the α -helix.

The cartoon representation of a protein's structure represents α -helices as coils.

β strand versus β -sheet

A flattened form of the helix in two dimensions is a common description for a β -strand. Rather than coils, β -strands have bends and these are sometimes referred to as pleats, like the pleats in a curtain.

Stretches of β -strands embedded within a single polypeptide chain form a β -sheet. Within a β -sheet, hydrogen bonds form between the backbone atoms of *separate* β -strands. When multiple strands from different regions of a polypeptide interact in this way a relatively flat, sheet-like surface is created, the β -sheet.

The cartoon representation of a protein's structure represents β -sheets as flat arrows.

Segments of the peptide backbone that do not form secondary structures are called loops. Loops are linkers that connect regions of secondary structure. Loops are shown in cartoon representations as lines connecting regions of secondary structure.

1.3.3 Tertiary Structure

For all proteins, the unique final three-dimensional structure adopted by the polypeptide is its tertiary structure.

This structure is determined by all types of non-covalent interactions that involve amino acid side chains (side-chain and backbone interactions, or side chain-side chain interactions).

These non-covalent bonds can be

Ionic interactions (strongest): between pairs of charged amino acids also known as **salt bridges**.

Hydrogen bonds: between polar groups- one of these polar groups is acting as a hydrogen donor the other as a hydrogen acceptor.

van-der Waals interactions (weakest): act only over short distances although they are present between any pair of atoms in close proximity.

Further, proteins fold in an aqueous environment, and that environment is critical to how the proteins adopt their native conformation.

Hydrophobic molecules do not form strong bonds with water. In aqueous solutions, hydrophobic molecules are driven together to the exclusion of water. As a protein folds to its final three-dimensional structure, the hydrophobic parts of the protein are forced together and away from the aqueous environment of the cell.

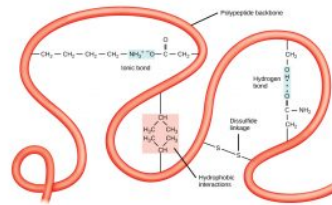


Figure. 1.7. Tertiary structure is created by non-covalent hydrophobic amino acid interactions as well as H-bonding in the interior of a polypeptide, leaving charged (hydrophilic) amino acid side chains to interact with water on the exterior of a typical “globular” protein. Stable covalent disulfide bonds between cysteine amino acids help stabilize tertiary structures.

Amino acid residues involved in these interactions can come from distant parts of the polypeptide chain bringing the chain into a more compact shape.

Disulfide bonds stabilize tertiary structure

Based on non-covalent bonds, tertiary structures are nonetheless strong simply because of the large numbers of otherwise weak interactions that form them.

However, one covalent bond is possible and is formed when two cysteine amino acids (sulfhydryl-containing side chain) are close together as a result of tertiary structure formation.

The sulfhydryl group is highly reactive and will covalently bond with another sulfhydryl group to form a covalent disulfide bond (the disulfide ($-S-S-$) bond).

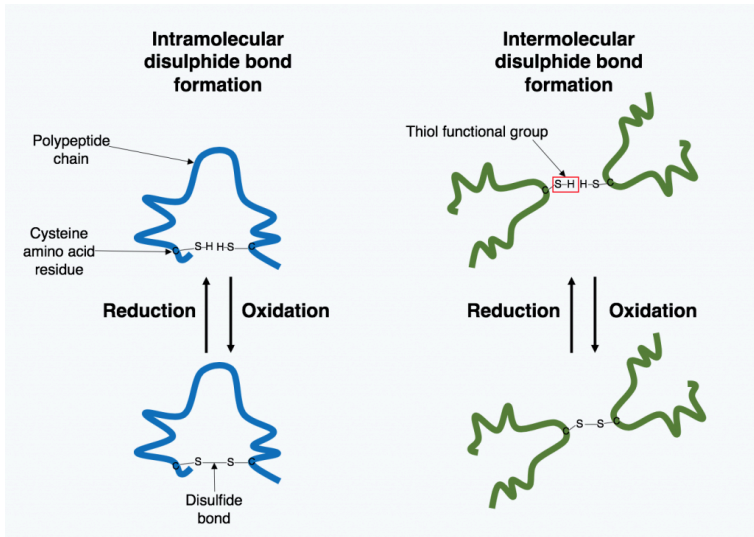


Figure 1.8. Shown is the reaction for the formation of a disulfide bond between two cysteine side chains. These can occur between distant regions of a polypeptide chain (blue) or between 2 different polypeptide chains (green). [Image is by Kep17, licensed under CC BY-SA 4.0, via Wikimedia.](#)

The formation of this bond depends on the environment. Oxidizing environment favors bond formation. Reducing environments break disulfide bonds.

Did you know?

Your hair provides a familiar example of disulfide

bonds. Hair features lots of these bonds, as they are important for its strength. Beauty salons take advantage of the disulfide bonds in your hair. Want to go from curly to straight? Add reducing agents to your hair. This breaks the disulfide bonds and chemically transforms them back into free cysteines. Want to make your hair curly? Get a perm? Curled hair around rollers which places different sulfhydryl groups in close proximity, and treat with an oxidizing agent, usually hydrogen peroxide, to form new disulfide bonds. Now instead of straight hair, you have curled hair.

1.3.4 Quaternary Structure

The fourth level of **protein** structure is that of quaternary structure. It refers to structures that arise as a result of interactions between **multiple polypeptides**, often referred to as **subunits**.

The subunits can be identical to each other or can be different polypeptide chains. The stabilizing interactions that hold the multiple polypeptides together are the same non-covalent interaction interactions (hydrogen bonding, ionic bonding, and hydrophobic interactions) and covalent disulfide

bonds, that stabilize the tertiary structure. Except, that they occur among between the polypeptide chains.

There are many examples of proteins that require more than one polypeptide to be functional and we will encounter several molecular machines (DNA polymerase, RNA polymerase) that are comprised of more than 10 different polypeptides.

The nomenclature used for such multi-subunit proteins reflects the number of polypeptides *and* the similarity between them.

For example, Estrogen (the hormone) receptor consists of **two identical** peptide chains coming together to form the functional protein. This is called a **homodimer**. [homo (similar) di (two) mer]

Nomenclature: Mono- Single; Di = Two, Tri = Three and so on. Homo = similar/same ; Hetero = Different

Adult hemoglobin is composed of **two identical chains** called α and **two identical chains** called β . Proteins with multiple polypeptides where at least one of the polypeptides is different from the other are Hetero-mers. Hemoglobin would be a **hetero-tetra-** mer.

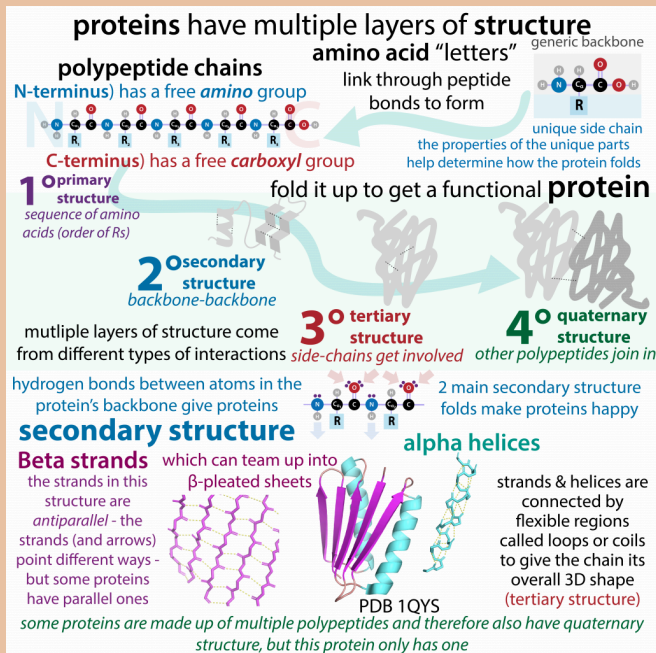
Key Takeaways: Via Sketchnotes!

A great learning tool is to create concept maps that highlight the key terms and connect them to one another in meaningful ways.

Included is an example of a free-to-use sketch note or graphic organizer uploaded by a young scientist who blogs and writes about Biochemistry. **[You can find her delightful work here: “the bumbling biochemist”:<https://thebumblingbiochemist.com/>].**

Making one for yourself makes for better learning than memorizing one. You don’t have to be an artist to do this!

Image attribution: [Biochemlife, CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0)
<<https://creativecommons.org/licenses/by-sa/4.0>>,
via [Wikimedia Commons](#)



Before you continue

1. Watch the Lecture video that covers the material above.
Linked here: [Dr. Mehta Lecture Video: Protein Structure Review](#)
2. Complete the associated Lecture Quickcheck.

1.4 Protein Structure Deep Dive

In the earlier part of the chapter, you were introduced to the protein folding and protein design problem. Numerous different 3-dimensional shapes can be adopted by polypeptides, only one is the final native fold. Yet, we see that only two secondary structures are commonly observed in proteins, the alpha helix, and beta-structures.

Why is this the case? To answer this we need to revisit the peptide bond as the chemical properties of the peptide bond are important for determining the shape of the polypeptide chain.

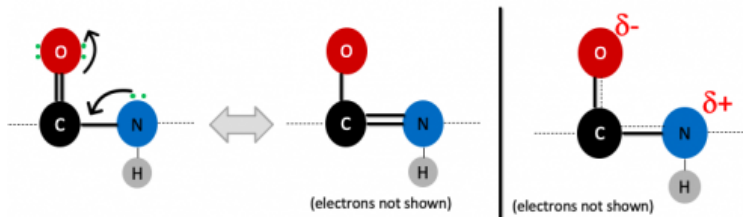
1.4.1 Peptide bond is polar, rigid, and planar (flat)

When two amino acids are joined a peptide bond is formed. Even though we draw the peptide bond with a single line it turns out, however, that the peptide bond does not behave like a single bond and has some double-bond character.

The carbonyl electrons are partially shared by the carbon-nitrogen peptide bond giving the carbonyl carbon/amide nitrogen bond a slight double bond character.

Because of the charge separation between oxygen and nitrogen, the peptide bond is also **polar**.

Because of this partial double bond character, **rotation around the C-N peptide bond is prevented and the peptide bond is planar.**



Both the C-O and C-N have a double bond character

The peptide bond is a partial double bond. Image created by Sapna Mehta for this book (CC-BY)

Planar means that – the six atoms: the carbonyl carbon, the carbonyl oxygen, the alpha carbon attached to the carbonyl carbon, the amide nitrogen (of the second amino acid), the hydrogen attached to the amide nitrogen, and the alpha carbon of the second amino acid are **all confined to a single plane**. These planes can be seen in the shaded blue in the diagram below. (Figure 1.8)

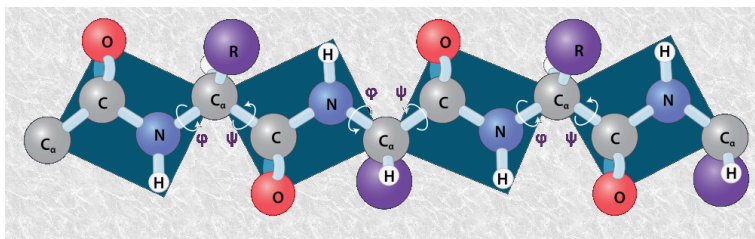


Figure 1.9. Diagram of a generic polypeptide chain. Blue rectangles indicate sets of six atoms that are coplanar due to the double-bond character of the peptide bond. (Image is adapted from Alejandro Porto obtained at <https://commons.wikimedia.org/wiki/File:Enlace-peptidico-caracter-planar.png> and licensed CC-BY-SA.)

Since there is **no rotation** around the C-N bond of the peptide bond, the only possible freedom of rotation in an amino acid residue is the carbonyl carbon-alpha carbon single bond, which is denoted as psi(Ψ), and the amide nitrogen-alpha carbon single bond, which is called phi(Φ).

Each blue rectangle is free to rotate relative to the other, which puts constraints on the possible conformations it can take, thereby influencing how the overall protein will fold.

The preference is to avoid clashes between R-groups. Given the bulkiness of R-groups, the phenomenon of steric hindrance, and the tendency of close side chains to interact with each other, one might expect that only certain values of ϕ and ψ , and hence conformations of the peptide, are permitted whilst others are not.

Turns out the conformations that are most stable are the alpha helices and beta strands!

The preference to avoid clashes between R-groups also results in the arrangement of **sequential R-groups alternating on either side of the polypeptide chain or 'trans' configuration.**

Thus the chemistry and nature of the peptide bond directly contribute to the observation that all proteins have mainly 2 types of secondary structures!

To watch a simulation of the above go to:
<https://youtu.be/O1ftYq13XKk> you can start the video at time stamp **3:13 and end at 5:58**

1.4.3 Features of Alpha-Helices and Beta Sheets

Shown below is a diagram for the alpha-helix highlighting the key features of the helix. (Figure 1.10)

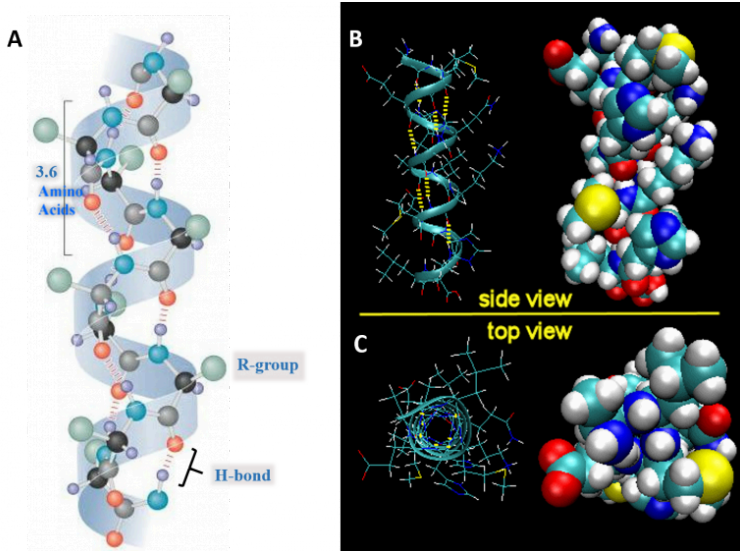


Figure 1.10 (A) Ball and Stick Model Side View. A total of 3.6 amino acids are required to form one turn of an α -helix. Hydrogen bonding between the carbonyl oxygen and the nitrogen of the 4th amino acid stabilizes the helical structure. On the structure shown, the black atoms are the α carbon, grey are carbonyl carbons, red are oxygen, blue are nitrogen, green are R-groups, and light purple are hydrogen atoms. (B) Expanded Side View Linear Structure and Space-Filling Model (C) Expanded Top View Linear Structure and Space-Filling Model. (Image credit: from <https://wou.edu/chemistry/courses/online-chemistry-textbooks/ch450-and-ch451-biochemistry-defining-life-at-the-molecular-level/chapter-2-protein-structure/>. Original images: Image A modified from: [Maksim](#) Image B and C from [Henry Jakubowski](#)

1. Each full turn of the helix (360°) is 3.6 amino acids in length.

[Note that the H-bond forms between C=O groups and N-H groups in the polypeptide backbone that are four amino acids distant]

2. Helices are predominantly right-handed.
3. The hydrogen bonds are parallel to the axis of the helix, located inside the helix and are in a regular arrangement. [All C=O bonds point in one direction, all N-H bonds point in opposite direction].
4. R-groups extend away from the helix.

The stability of an alpha helix is affected by different factors.

- Amino acids whose R-groups are too large (tryptophan, tyrosine) or too small (glycine) **destabilize** α -helices.
 - Glycine has a lot of conformational flexibility given its side chain is simply an H atom! Glycine is found in the more flexible regions of proteins which are the loops.
- Certain combinations of adjacent amino acids. For example, a run of positively charged or negatively charged side chains will repel one another.
- Presence of Proline- the 'helix breaker'.
 - Take a look again at the **side chain of the proline in Figure 1.2**.
 - Notice that the hydrogen connected to the N of the amine group is not available to hydrogen bond! In

fact, proline disrupts both types of secondary structures. However, proline **is** commonly found in turns or loops between the beta-strands, connecting secondary structure elements and occasionally in the **first helical turn** where the side chain geometry does not create a problem.

Parallel and Antiparallel β -Sheets

In a β -sheet, two or more sections of the polypeptide run alongside each other and are linked in a regular manner by hydrogen bonds between the main chain C=O and N-H groups. In parallel beta-sheets, the strands all run in one direction, whereas in antiparallel adjacent sheets run in opposite directions.

Practice Exercises



An interactive HSP element has been



excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=43#h5p-5>

1.5 Domains in Protein Structure.

An important concept in protein structure is that of the **protein domain**. In many cases, a single polypeptide can be seen to contain two or more physically distinct substructures, known as domains. Often linked by a flexible hinge region these domains are compact, stable, and fold independently of the rest of the protein chain. Again, since the shape is important for function, domains have a shape that is best suited for a particular cellular function and are named accordingly!

For example; Nucleotide-binding domain, Calcium-binding domain. Occasionally they are named after their discoverers like Pleckstrin Homology (PH) domain. Figure 1.

10 shows structures of two different proteins both of which have the PH domain.

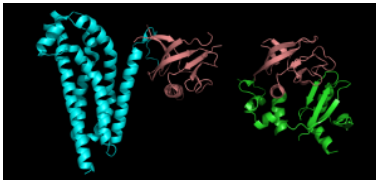


Figure 1.10 Protein Domains. Example of local structural homology. The two shown proteins, Dbs (left) and Grb10 right, share a common PH domain (maroon), which binds phosphatidyl-inositol triphosphate. [Image credit: Fdardel, CC BY-SA 3.0 via Wikimedia Commons](#)

Proteins that have this domain, can bind a molecule of phosphatidylinositol triphosphate that is generated as part of a common cell-signaling pathway.

Proteins sharing ***more than a few*** common domains are encoded by members of evolutionarily related genes comprising **gene families**. Genes for

proteins that share only one or a few domains may belong to gene superfamilies. Superfamily members can have *one function* in common (as predicted by the domain), but the rest of their sequences are otherwise unrelated.

Out of thousands of structures that have now been solved we can see similar structures among proteins with very different sequences. This suggests that there are a limited number of stable folds and almost all novelty in protein structure comes from the way these single domains are arranged. Unlike the number of novel single domains, the number of multidomain families being added to the public databases is still rapidly increasing.

There are several advantages conferred by multidomain protein architecture:

1. *Creation of catalytic or substrate-binding sites:* These sites are often formed at the interface between two domains, typically a cleft. The movement of the domains relative to each other allows the substrate to bind.
2. *Segregation of function.*
3. *Multifunctional proteins* A multidomain protein may have more than one function, often related, and each function is performed by a distinct domain. For example, *E. coli* DNA polymerase a multi-subunit protein we will encounter has both polymerase activity, and two kinds of exonuclease activity, all of which are required for DNA replication and all of which reside on distinct domains of this protein.

1.6 Protein Modifications can affect structure and function

Proteins in the cell are often modified post-translationally by the addition of functional groups via covalent bonds to the side chains of amino acids. These groups are added by enzymes. Think of these modifications as an additional makeover or accessorizing the protein. We have spent time revealing how the nature of the amino acid is crucial for

protein folding, and the final shape is important for function, changing key amino acids chemically can impact the structure and the function. The changes in shape are often subtle but can have a dramatic impact on function, activity, stability, localization, and/or interacting partner molecules.

Two modifications that we will encounter include:

1. Addition of phosphate groups (*phosphorylation*)
2. Addition of acetyl and methyl groups (*acetylation, methylation*)

Other types of modifications are shown in the diagram below (Figure 1.11).

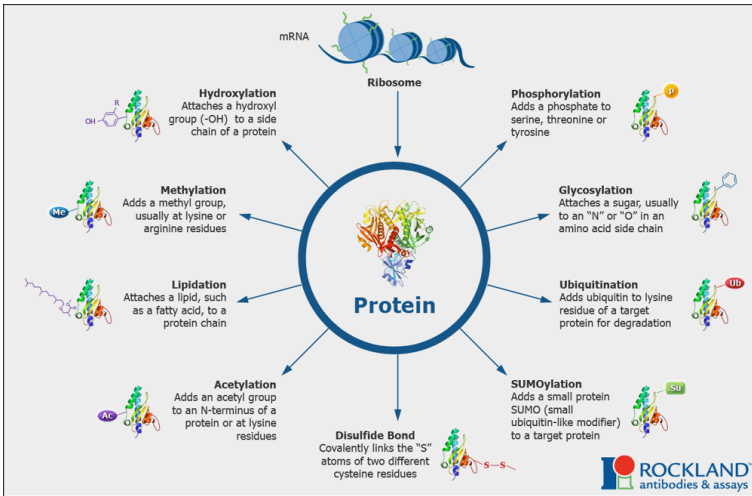


Figure 1.11 Types of post-translational modifications. (Image: Copyright Rockland Antibodies and Assays @ <https://rockland-inc.com/post-translational-modification-antibodies.aspx>. with permission)

Collectively these modifications account for and enhance the molecular and functional diversity of proteins within and across species. Much of the work of proteins inside cells involved signaling or communication.

Often these modifications are reversible! Meaning that just as there are enzymes that add these groups, there are enzymes that can remove them. Having reversible modifications allows many proteins to function as signaling nodes: turning on (when modified) or off (when modifications are removed) or vice versa!

1.7 Intrinsically Disordered Proteins

We have thus far focused on how proteins can adopt a very specific three-dimensional shape. However, we now know that entire classes of proteins termed Intrinsically Disordered Proteins lack any well-defined secondary or tertiary structure! Some proteins exhibit regions that remain unfolded (IDP regions) even as the rest of the polypeptide folds into a structured form- provided the protein “hinges” that can move a protein domain in a controlled way, loops that have an open or closed conformation.

Intrinsically disordered proteins and disordered regions within proteins have, in fact, been known for many years, but were initially regarded as an anomaly.

With the realization that IDPs and IDP regions are widespread among eukaryotic proteins, that it has been recognized that the observed disorder is a **“feature, not a bug”**.

Comparison of IDPs shows that they share sequence characteristics that appear to favor their disordered state. That is, just as some amino acid sequences may favor the folding of a polypeptide into a particular structure, the amino acid sequences of IDPs favor their remaining unfolded. IDP regions are seen to be low in hydrophobic residues and unusually rich in polar residues and proline.

The presence of a large number of charged amino acids in the IDPs can inhibit folding through charge repulsion, while the lack of hydrophobic residues makes it difficult to form a stable hydrophobic core, and proline discourages the formation of helical structures. The observed differences between amino acid sequences in IDPs and structured proteins have been used to design algorithms to predict whether a given amino acid sequence will be disordered.

What is the significance of intrinsically disordered proteins or regions? The fact that this property is encoded in their amino acid sequences suggests that their disorder may be linked to their function. The flexible, mobile nature of some IDP regions may play a crucial role in their function, permitting a transition to a folded structure upon binding a protein partner or undergoing post-translational modification.

Studies on several well-known proteins with IDP regions suggest some answers. IDP regions may enhance the ability of proteins like the lac repressor to translocate along the DNA to search for specific binding sites. The flexibility of IDPs can also be an asset in protein-protein interactions, especially for proteins that are known to interact with many different protein partners.

Concepts in Context (Molecular Biology in the News): “Jumping Insects”

LISTEN to the Science Friday episode (link below).

<http://www.sciencefriday.com/segments/flexible-insect-protein-inspires-super-rubber/> (Links to an external site.)

Resilin, as described in the podcast, has many desirable mechanical properties. Some of these include rubber-like elasticity, high extensibility, efficient energy storage (think bouncy ball or springs). Resilin is **a disordered protein**; however, its segments may take on secondary structures under different conditions. Glycine and proline are two particular residues that often appear in the sequence of disordered proteins and in the case of resilin glycine is the most substantial proportion (30–40%) of the total residues.

COMPLETE: Don't forget to complete the associated concepts in context assignment.

For a video review of the material above click on the JoVE (Journal of Visualized Experiments) Quick Review link provided on your CANVAS page.

1.7 X-Ray Crystallography: Art Marries Science

How would you examine the shape of some thing too small to see in even the most powerful microscope?

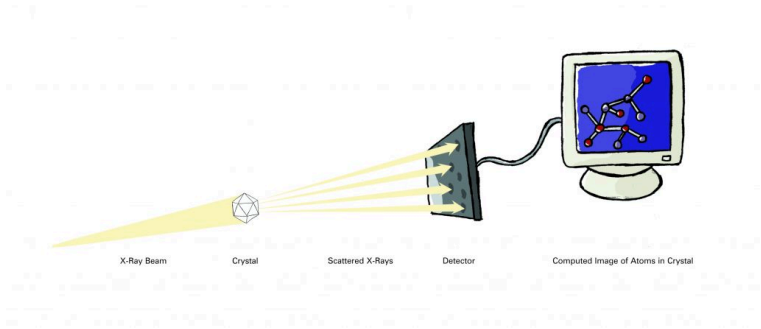
Scientists trying to visualize the complex arrangement of atoms within molecules have exactly that problem, so they solve it indirectly. By using a large collection of identical molecules — often proteins — along with specialized equipment and computer modeling techniques, scientists are able to calculate what an isolated molecule would look like.

The two most common methods (prior to computational methods of protein prediction) used to investigate molecular structures are X-ray crystallography (also called X-ray diffraction) and nuclear magnetic resonance (NMR) spectroscopy. Researchers using X-ray crystallography grow solid crystals of the molecules they study. Those using NMR study molecules in solution. Each technique has advantages

and disadvantages. Together, they provide researchers with a precious glimpse into the structures of life.

More than 85 percent of the protein structures that are known have been determined using X-ray crystallography. In essence, crystallographers aim high-powered X-rays at a tiny crystal containing trillions of identical molecules. The crystal scatters the X-rays onto an electronic detector like a disco ball spraying light across a dance floor. The electronic detector is the same type used to capture images in a digital camera. After each blast of X-rays, lasting from a few seconds to several hours, the researchers precisely rotate the crystal by entering its desired orientation into the computer that controls the X-ray apparatus. This enables the scientists to capture in three dimensions how the crystal scatters, or diffracts, X-rays. The intensity of each diffracted ray is fed into a computer, which uses a mathematical equation called a Fourier transform to calculate the position of every atom in the crystallized molecule.

The result — the researchers' masterpiece — is a three-dimensional digital image of the molecule. This image represents the physical and chemical properties of the substance and can be studied in intimate, atom-by-atom detail using sophisticated computer graphics software.



Crystal Cookery

An essential step in X-ray crystallography is growing high-quality crystals. The best crystals are pure, perfectly symmetrical, three-dimensional repeating arrays of precisely packed molecules. They can be different shapes, from perfect cubes to long needles. Most crystals used for these studies are barely visible (less than 1 millimeter on a side). But the larger the crystal, the more accurate the data and the more easily scientists can solve the structure.

Crystallographers grow their tiny crystals in plastic dishes. They usually start with a highly concentrated solution containing the molecule. They then mix this solution with a variety of specially prepared liquids to form tiny droplets (1-10 microliters). Each droplet is kept in a separate plastic dish or well. As the liquid evaporates, the molecules in the solution become progressively more concentrated. During this process, the molecules arrange into a precise, three-dimensional pattern and eventually into a crystal — if the researcher is lucky.

Sometimes, crystals require months or even years to grow. The conditions — temperature, pH (acidity or alkalinity), and concentration — must be perfect. And each type of molecule is different, requiring scientists to tease out new crystallization conditions for every new sample. Even then, some molecules just won't cooperate. They may have floppy sections that wriggle around too much to be arranged neatly into a crystal. Or, particularly in the case of proteins that are normally embedded in oily cell membranes, the molecule may fail to completely dissolve in the solution.

Some crystallographers keep their growing crystals in air-locked chambers, to prevent any misdirected breath from disrupting the tiny crystals. Others insist on an environment free of vibrations — in at least one case, from rock-and-roll music. Still, others joke about the phases of the moon and supernatural phenomena. As the jesting suggests, growing crystals remains one of the most difficult and least predictable parts of X-ray crystallography.

It's what blends art with science.

Watch the lecture videos in this playlist:

[Understanding X-ray crystallography](#)

Remember to:

1. Watch the videos that cover the material above. This will help reinforce certain concepts if they were unclear.
2. Complete the associated Lecture Quick checks
3. Complete any Molecular Biology in the News Assignments
4. Begin work on Weekly Problem Set

SOME KEY TERMS

disulfide bonds, C-terminus, N-terminus, polypeptide, peptide bond, primary, secondary, Quaternary, alpha helix, beta -sheet and beta-strand, Ramachandran plot, phi and psi angles, phosphorylation, acetylation, methylation, protein domain.

References and Attributions

This chapter is curated from and contains material from the following **CC-licensed content**. Changes include rewording, replacing and removing paragraphs with original material.

- Section 1.1 and Section 1.8 from Structure of Life (2007). Booklet, retrieved from National Institute of General Medical Sciences <https://www.nigms.nih.gov/education/Booklets/The-Structures-of-Life/Pages/Home.aspx>; (United States Government Work, in the public domain)
- Ahern K, Rajagopal I and Tan T. (2013). Biochemistry Free for All (Version 1.3). Licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](#). The entire textbook is available for free from the authors at <http://biochem.science.oregonstate.edu/content/biochemistry-free-and-easy>.
- Details of Protein Structure” by Gerald Bergstrom, LibreTexts which is licensed under CC BY. The chapter can be found online at <https://bio.libretexts.org/@go/page/16428>

Images

Unless otherwise noted within the text, images on this page are licensed under CC-BY 4.0 by OpenStax. **Located**

at: <https://openstax.org/books/biology/pages/3-4-proteins>. Access for free at <https://openstax.org/books/biology/pages/1-introduction>



Introduction to Molecular Biology by Sapna Mehta is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/), except where otherwise noted.

2.

METHODS TO STUDY PROTEINS



A. Iletin (Insulin, Lilly) Display 15-B sales kit, vials demonstrating steps of insulin extraction and purification
Image credit: National Museum of American History, Smithsonian Record ID: edanmdm:nmah_1071412



B. Recombinant insulin vials, pen and syringes. Image credit: Blausen.com staff (2014). "Medical gallery of Blausen Medical 2014". Wikijournal of Medicine 1 (2); CC BY 3.0 , via Wikimedia Commons

2.1 Introduction: Molecules to Medicine

On Jan. 11, 1922, Leonard Thompson a 14-year-old boy received an experimental injection of insulin for the treatment of Diabetes at the Toronto General Hospital. Just days earlier he had been admitted for 'end-stage diabetes mellitus. At the time diabetes was a life-shortening disease with a life expectancy of generally less than a year from diagnosis. No one expected the boy to live. The results however were dubbed

as miraculous. History documents the day as the turning point for patients all over the world suffering from diabetes.

Just a short 11 months earlier, Frederick Banting and medical student Charles Best had begun the classic experiments to show that pancreatic extracts could restore the ability of diabetic dogs to manage blood sugar.

With the help of biochemist, James Collip the team began to work on improving the purity of insulin made starting with cow pancreas and using various extraction techniques.

To see if insulin was present in each solution, and in what amount they developed an activity assay which included monitoring blood sugar levels of rabbits following injection of each solution. Collip developed a measure of activity based on the ability of the extract to lower blood sugar in the rabbit.

The Nobel Prize in Physiology or Medicine 1923 was awarded jointly to Frederick Grant Banting and John James Rickard Macleod “for the discovery of insulin”. (1)

Links to Learning

See here [Celebration 100](#) for an interactive on the history of insulin discovery. It includes footage of the lab, original records, and images.

2021 is the 100th anniversary of insulin's discovery – the first life-saving treatment for diabetes. The story of insulin discovery and subsequent production is a great example of serendipity, persistence and the power of pursuing questions of consequence. It also highlights nicely the role of biochemistry and the dawn of medicinal molecular biology.

In this chapter, you will learn the various protein methods to isolate proteins from a complex mixture in cells. In subsequent chapters, we will learn the tools of recombinant technology that allow us to generate human insulin using bacteria or yeast as production factories.

Learning Objectives

Levels 1 and 2 are factual information, knowledge-based. Level up indicated by the “Target” symbol is the goal.

When you have mastered the information in this chapter, you should be able to:

Level 1 and 2 (Knowledge and

Comprehension)

- Compare and contrast Gel Exclusion, Ion Exchange, and Affinity chromatography methods.
- Describe SDS-PAGE. Explain the function of SDS, DTT/Beta Mercaptoethanol in SDS-PAGE.
- List the steps in performing a Western Blot.
- Explain when the technique of western blot is applied, and the kinds of questions one can answer by performing a western blot

⊕Level Up (Application, Analysis, Synthesis)

- Develop a purification scheme (Steps, type of chromatography, pH of buffers) for a protein within a given set of parameters.
- Interpret co-immunoprecipitation data.
- Interpret a western blot image/data from a scientific article.
- Identify or troubleshoot problems with a protein purification scheme based on the image of a stained SDS-page gel.

A special note for students: Lab methods can be difficult to learn by just 'reading' instead of 'doing'.

I have provided videos and animations here to help you visualize the process. I hope those will help and I highly recommend watching them.

*At the same time understanding the 'why's of the methodology is critical for troubleshooting, and troubleshooting is an important problem-solving skill. Therefore, **do not skip** reading the chapter as the theory is better laid out here.*

In short, do both- Read and watch/listen!

2.2 Biochemical Assays

A successful protein purification procedure can be nothing short of amazing. What are the scenarios that require the isolation of a protein? A few are listed below

- 1) You want to study the function of a known protein.
- 2) You want to recapitulate a cellular process in a test tube and need to assemble the protein components.

3) You want to isolate a protein in pure form for structural biology (X-ray crystallography)

4) You are trying to **identify a novel protein** responsible for a function. For example, we will soon learn about DNA replication, an elaborate process of making identical copies of our genomes.

The proteins and enzymes involved in replication were identified by separating the cells into fractions and isolating those that demonstrated the desired activity (like the ability to add nucleotides to a growing strand).

Whether you are starting off with a recombinant protein 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, polysaccharides, etc. from which you may have to extract milligram (or microgram!) quantities of desired protein at high purity, and hopefully with high yield.

As illustrated in the story of insulin, the first step in any purification is the development of a **specific assay** for the protein of interest.

An assay is a test. Some way to measure and quantify both the **presence** of your protein inside the sample and its activity/function.

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)

Examples: Developing Assays

Deriving an assay is one of the most creative parts of doing biochemistry! In this video below Dr. Ron Vale describes the experiment to reconstitute axonal transport of membrane organelles in a test tube and the unexpected results that were encountered. The results of this experiment led to the discovery of a new motor protein, kinesin. **(If you are short on**

time, you can watch time stamps: Beginning to 3:42, 11:24 to 16:00. approx 8 minutes)



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://iu.pressbooks.pub/iul211smehta/?p=168#oembed-1>

[Reconstituting Motility from the Squid Giant Axon](#) from [XBio Media](#) on [Vimeo](#). For more information, see Ronald Vale's Key Experiment on Motility in a Test Tube in The Explorer's Guide to Biology (explorebiology.org/collections/cell-biology/motility-in-a-test-tube).

During purification, you will **need to keep track of several parameters**, but two are critical

1. ***The total amount of all proteins present in your sample*** (estimates obtained by measure absorbance at 280nm often suffice)
2. ***The total amount of the protein of interest*** (the one you are trying to purify). This information will be

determined from your **quantitative assay**.

From the 2 measurements, you can obtain **Specific Activity** of the desired protein (“units of activity” of desired protein/mg total protein) and **% yield** (the fraction or percentage of the final desired protein divided by the amount of material in the original mixture)

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 a good yield.
- However, it may be difficult to improve that purity by an additional few percentiles 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.

Did I get this?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=168#h5p-6>

2.3 Isolating Proteins from Complex Mixtures

With an assay in hand, the next step in protein purification is to break open the cell/ source material. Early insulin preparations utilized cow pancreas!

2.3.1 Cell Disruption

There are several ways to break open cells. Lysis methods include lowering the ionic strength of the medium cells are kept in. This can cause cells to swell and burst. Mild detergents may be used to enhance the efficiency of lysis.

Depending on your starting material (bacteria? plants?) you

may need additional assistance. Most bacteria, yeast, and plant cells have a cell wall and are resistant to such osmotic shocks. Here, enzymes may be useful in helping to degrade the cell walls.

Lysozyme, for example, is very useful for breaking down bacterial walls and commonly added to lysis buffers. Other enzymes commonly employed include cellulase (plants), glycanases, proteases, mannases, and others.

Mechanical agitation may be employed in the form of tiny beads that are shaken with a suspension of cells. As the beads bombard the cells at high speed, they break them open.

Sonication (20-50 kHz sound waves) provides an alternative method for lysing cells. The method is noisy, however, and generates heat that can be problematic for heat-sensitive compounds.

Pressure Disruption: In this method, cells are placed under very high pressure (up to 25,000 psi). When the pressure is released, the rapid pressure change causes dissolved gases in cells to be released as bubbles which, in turn, break open the cells.

Cryopulverization: is often employed for samples having a tough extracellular matrix, such as connective tissue or seeds. In this technique, tissues are flash-frozen using liquid nitrogen and then ground to a fine powder before extraction of cell contents with a buffer.

Whatever method is employed, the crude lysates obtained contain all of the molecules in the cell, and thus, must be

further processed to separate the molecules into smaller subsets, or fractions. This soupy mess of everything inside the cells is called **Crude Extract/ Homogenate/Cell Lysate**.

2.3.2 Fractionation

Fractionation of crude extract starts with centrifugation. Using a centrifuge, one can remove cell debris, and fractionate organelles (when working with eukaryotic cells), and cytoplasm. For example, nuclei, being relatively large, can be spun down at fairly low speeds. Once nuclei have been sedimented, the remaining solution, or supernatant, can be centrifuged at higher speeds to obtain the smaller organelles, like mitochondria.

This stepwise process of fractionating cellular organelles is called **Differential Centrifugation**.

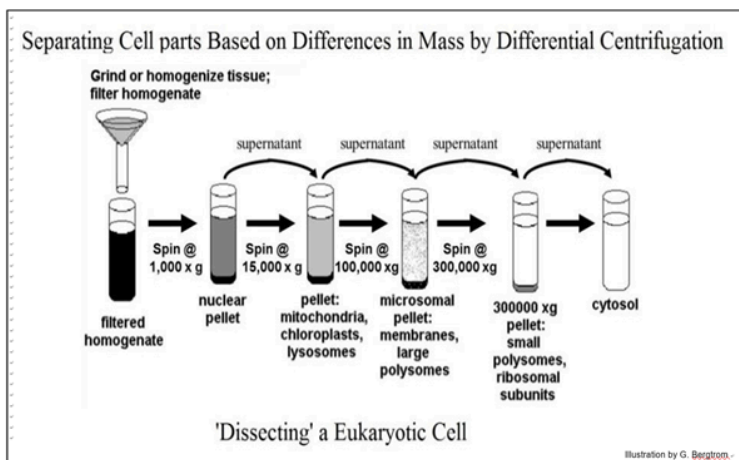


Figure 2.2. Subcellular fractionation by differential centrifugation. *Image Credit: Gerald Bergtrom from "Book: Basic Cell and Molecular Biology (Bergtrom)" licensed under CC BY.*

Each of these fractions will contain a subset of the molecules in the cell. Although every subset contains fewer molecules than does the crude lysate, there are still many hundreds of molecules in each.

Separating the molecule of interest from the others is where chromatography comes into play.

Dr. Mehta's Lecture Video: [How to Study Proteins- Cell Lysis and Centrifugation](#)

(lecture slides have been provided on CANVAS to follow along)

2.4 Column Chromatography

Proteins differ in size, charge, shape, and affinity for other proteins or molecules. These differences in properties can be exploited to begin to separate a protein of choice away from others. The method most commonly used is Chromatography.

During chromatography, the mobile phase (buffer or other solvents) moves through the stationary phase (usually a solid matrix) carrying the components of the mixture. Separation of the components is achieved, because the different components move at different rates, for reasons that vary, depending on the type of chromatography used.

In most cases chromatography is performed in long glass tubes filled with a matrix or resin of beads, hence the term **Column Chromatography**.

The type of bead/resin is chosen based on the property that will be exploited for separation. We will discuss the following three:

- Ion exchange chromatography- separation based on charge
- Gel exclusion chromatography- separation based on size
- Affinity chromatography- separation based on binding affinities of sample molecules (typically proteins) for molecules covalently linked to the support beads.

Regardless of the type of chromatography, there are some common **terms and a workflow** associated with all types of chromatography.

First: the column is filled with the beads and **equilibrated** with an appropriate buffer.

Second: The 'lysate' which is a mixture of proteins is added

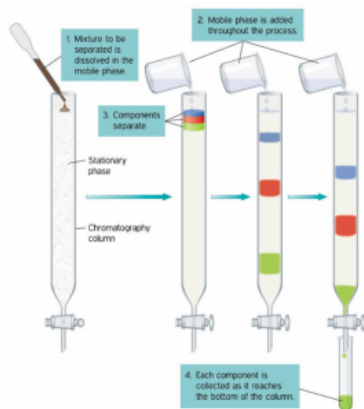


Figure 2.2 Principle of separation of a complex mixture of compounds by chromatography. Image credit: EjupPod, CC BY-SA 3.0 via Wikipedia commons]

to the top of the column and the same buffer is allowed to flow through the column. As the buffer **flows through** the column the mixture of proteins is drawn down the column and interacts with the matrix or resin, proteins not interacting with resin flow through and come out of the column.

Third: Material that is collected from the bottom of the column in small volumes called '**fractions**'.

Fourth: The desired protein will eventually drip out or *elute*, from the column. This is often done by changing the buffer conditions. The step is referred to as **elution** and the buffers are referred to as '**Elution Buffers**'.

Usually, the bulk of the fractions are kept and assayed for protein content and activity. Tubes with the highest specific activity are pooled and taken on to the next step.

The process of collecting fractions is often automated. (See **video at end of chapter or a protein purification process at IU from start to finish**)

2.4.1 Ion exchange chromatography

The names of chromatography are clues to the principle of separation! In ion-exchange chromatography, the support consists of tiny beads to which are attached chemicals possessing a charge (**ions**). Before use, the beads are equilibrated in a solution containing an appropriate counter-ion to the charged molecule on the bead.

Thus, in a **cation-exchange** column, the chemicals

attached to the beads are **negatively charged** and the counter-ions are positively charged. When positively charged compounds present in a mixture are passed through the column they will exchange with the counter-ions and “stick” to the negatively charged groups on the beads. Molecules in the sample that are neutral or negatively charged will pass quickly through the column.

On the other hand, in **anion-exchange** chromatography, the chemicals attached to the beads are **positively charged** and the counterions are negatively charged (chloride, for example). **Negatively charged molecules** in the cell lysate will “stick” and other molecules will pass through quickly. To remove the molecules “stuck” to a column, one simply needs to add a high concentration of counter-ions to release them.

Consequently, in order to elute the ‘stuck’ proteins, it is usually necessary to alter the affinity of the protein for the column by either adding salt to interfere with the electrostatics or changing the charge on the protein by altering the pH.

The selection of pH of sample buffer and the column is key in this kind of separation as it determines the charge of the protein.

Recall in Chapter 1 we discussed the relationship between the charge on the ionizable side chains of proteins and the pH of the environment.

Another term we need to know is the *isoelectric pH /point / pI* of a protein. This refers to the pH value at which the total (net) charge on the protein is zero.

At this pH value, the negative and positive charges of the protein are equal, and the protein is at neutral charge.

Since the charge on side chains relies on the pH of the environment (or buffer in the case of experiments), the net protein charge can be altered by *changing* pH or by adding salt to neutralize the charge.

REMEMBER THIS

If the pH is less than the pI, then the protein will be positively charged.

If the pH is greater than the pI, the protein will be negatively charged.

How does this help? In Cation exchange for example ‘stuck’ proteins can be eluted by raising the pH to above the pI of the protein!

Therefore, knowing the pI of the protein helps experimenters to determine the appropriate pH for the buffers and design a purification protocol using ion exchange resins

For proteins with a known amino acid sequence, there are handy computer programs that generate an approximate pI.

Watch the videos below: (link out by clicking titles or

directly play the embedded video on your device if you are able)

[Introduction to Ion-Exchange](#)



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=168#oembed-2>

[Principle of Ion Exchange Chromatography](#)



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=168#oembed-3>

2.4.2 Size Exclusion Chromatography

Size exclusion chromatography (also called **molecular exclusion** chromatography, **gel exclusion** chromatography,

or **gel filtration** chromatography) is a separation method that is based on a **physical** property of the protein – the *effective molecular radius* (which relates to **mass** for most typical globular proteins).

The column resin consists of beads with tiny “tunnels” in them with openings of a precise size.

The size of the opening is referred to as an “exclusion limit,” which means that molecules above that limit (a certain molecular weight) **will not be able** to pass through the tunnels.

They can pass through the column relatively quickly by making their way between the beads.

Smaller proteins that 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).**

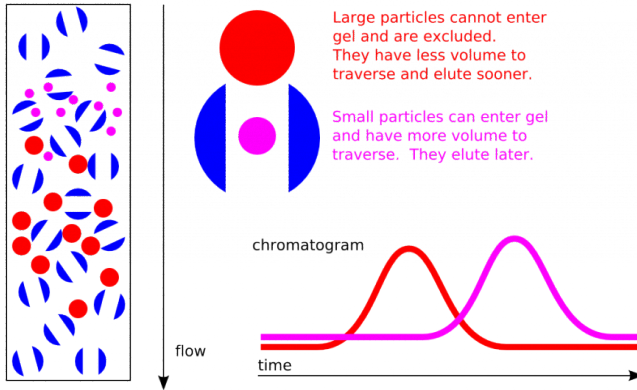


Figure 2.4: Size exclusion chromatography. Image credit: The original uploader was [Takometer at English Wikipedia](#)., CC BY 2.5 via [Wikimedia Commons](#).



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=168#oembed-4>

For most general lab methods the **size of protein is equated with its mass or molecular weight.**

The unit of molecular weight for proteins is **Daltons** (Da- atomic mass unit) or more commonly in **kilodaltons or (kDa– 1000 Da)**. Typically, the protein size or molecular weight is an **approximation of the sum of the molecular weights of all its amino acids.**

There are online programs that will tell you the size of a protein of a known sequence.

For a ***really rough*** estimate scientists can use the average molecular weight of an amino acid – 110Da and multiply by the # of amino acids in the protein.

For example, A protein with 100 amino acids is ~ 11.0 Kda.

Practice- SOLVED PROBLEM

Learning Objective – Development of a Purification Scheme.

If characteristics are known about the target protein, such as size and charge, it is possible to devise a purification scheme to obtain the target protein. Additional proteins may co-purify with the target protein, so it may be necessary to modify or add additional steps to the purification scheme.

Question: Devise a purification scheme to purify protein C from a mixture of A, B, C, and D. The properties of the proteins are given below. You can use gel filtration, anion, or cation exchange columns in your purification scheme.

Protein	# of Amino Acid Residues (Mol Weight)	pI (Iso electric point)
A	120 (13,200 Da)	7.5
B	120 (13,200 Da)	7.5
C	120 (13,200 Da)	6.5
D	240 (26,400 Da)	8.0

Think about what do you need to know to solve this problem?

1. What are the differences between the proteins in size, charge that can be used to separate the proteins?

Molecular Weight: We can separate A, B, and C from D based on differences in molecular weight.

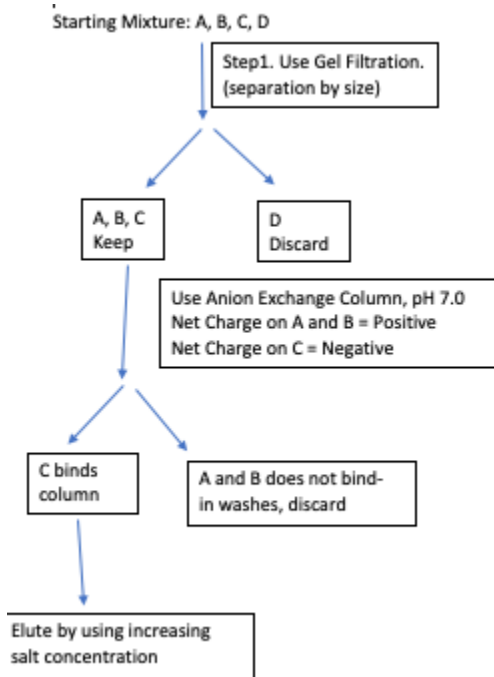
Charge: We can separate C from A and B based on the difference in isoelectric point.

2. What is the relationship between pH and pI?

Buffer pH will determine the net charge on proteins. If the pH of the buffer is the same as the pI of the protein, then the protein will carry zero net charge. If the protein is suspended in a buffer whose pH value is less than the protein's pI, the protein will pick up protons from the buffer and hence have a net positive charge. Conversely, if the protein is suspended in a buffer more alkaline than the protein's pI, the protein will lose protons to the buffer and become net negatively charged.

3. Types of chromatography.

Solution:



Dr. Mehta's Lecture Video's Playlist:

[Size Exclusion and Ion-Exchange Chromatography](#)

Before you continue you should

1. Watch the Lecture videos that cover the material above.
 2. Complete the associated Lecture Quickcheck.
-

2.4.3 Affinity chromatography

Affinity chromatography is a very powerful and selective technique that exploits the binding *affinities* of sample molecules (typically proteins) for molecules covalently linked to the support beads. In contrast to ion-exchange chromatography, where all molecules of a given charge would bind to the column, affinity chromatography exploits the specific binding of a protein or proteins to a ligand that is immobilized on the beads in the column.

For example, if one wanted to separate all of the proteins in a cell lysate that binds to ATP from proteins that do not bind ATP, one could use a column that has ATP attached to the support beads and pass the sample through the column. All proteins that bind ATP will “stick” to the column, whereas

those that do not bind ATP will pass quickly through it. The bound proteins may then be released from the column by adding a solution of ATP that will displace the bound proteins by competing, for the proteins, with the ATP attached to the column matrix.

A common form of affinity chromatography involved the use of a specific antibody attached to an inert resin.

When done on a small scale this is also termed **Immunoprecipitation** or Immuno-Affinity chromatography.

Antibodies are powerful tools for protein biochemistry, commonly used in research, and will feature again in this chapter.

Watch the video below for why antibodies are useful and commonly used in research. Parts between time stamps 1:50 and 3:01 can be skipped.

Bonus: As you listen to the description of antibody protein structure pay attention to terms you encountered in Chapter 1!



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=168#oembed-5>

Going back to the general workflow for chromatography. In affinity chromatography after binding, washing and elution of bound protein.

One generalization regarding the method of elution is that the bound ligand (protein) can be competed off from 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 that recognizes only the protein of interest and none other.*

2.5 SDS-PAGE and Western Blotting

Most proteins are colorless! During and at the end of a protein purification scheme how can you tell if you have finally obtained a pure protein?

Assays will establish a protein is present, but cannot tell you if there are other contaminants. What if there was some way to ‘visualize’ the components of that colorless liquid in the fractions?

A commonly used technique for detecting and identifying the presence of proteins inside a sample is SDS-PAGE and western blotting. Since western blotting includes SDS-PAGE as its first step we begin with SDS-PAGE.

2.5.1 SDS-PAGE – Sodium Do-decyl Sulfate Poly Acrylamide Gel Electrophoresis

Let’s break the name apart:

Gel Electrophoresis is a method that uses an **electric** field across a **gel** matrix to **separate** (phoresis) large molecules. Samples are loaded into the wells of the gel matrix that can separate molecules by size and an electrical field is applied across the gel. This field causes negatively-charged molecules to move towards the positive electrode. The gel matrix, itself, acts

as a **sieve**, through which the **smallest** molecules pass rapidly, while **longer** molecules are slower-moving.

Polyacrylamide: when the gel-like matrix is made utilizes a chemical **polyacrylamide**. This is a polymer comprised of two covalently-linked components:

- acrylamide
- bis acrylamide.

When mixed together they form a mesh that functions as a sieve.

As the proteins undergo electrophoresis, they are separated according to their molecular weight because of the sieving properties of the polymer strand in the gel.

- Smaller proteins have a much easier time moving and therefore migrate a larger distance.
- Larger proteins have a more difficult time and therefore migrate a smaller distance.

NOTE: Most common SDS-PAGE gels ***will not be able to resolve proteins that differ by a few amino acids in size.*** A different method is needed to further separate 2 proteins that are close in size.

[Casting A Gel:](#)



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://iu.pressbooks.pub/iul211smehta/?p=168#oembed-6>

That covers the PAGE part of SDS- PAGE.

We now have everything in place to perform a gel electrophoresis experiment to separate proteins, but there is another consideration related to proteins we need to account for. Remember, 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**.

Role of SDS

Sodium dodecyl sulfate (**SDS**; also known as “laurel sulfate”) is an ionic detergent. It has an anionic head group and a **lipophilic** tail. 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)

SDS accomplishes a couple of useful things:

1) SDS causes proteins to denature and disassociate from each other (excluding covalent cross-linking) and essentially unravel into linear molecules.

2) Due to the charged nature of the SDS molecule the proteins all proteins will have a negative charge!

This is useful to us! Now all proteins in the sample will have an approximate ***constant charge to mass ratio*** and will migrate through the gel at a rate **proportional to their molecular mass alone**.

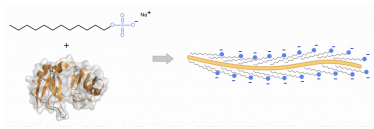


Figure 2.5 Denaturation of a protein (bottom-left) by sodium dodecyl sulphate (top-left). Detergent molecules coat the unfolded protein chain (right), which adopts a rod-like shape as a consequence of electrostatic repulsion. Negative charges contributed by SDS dominate the total charge of the aggregate and are used as the driving force in SDS gel electrophoresis.

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.

In addition, recall that many proteins can be stabilized by disulfide bonds which are covalent. The samples that are prepared for SDS-PAGE are also often treated with **reducing agents** like **b-mercaptoethanol**, **dithiothreitol** (DTT). These reagents will reduce disulfide bonds and separate polypeptide chains that are connected by such bonds.

In setting up the SDS PAGE experiment we need to **know when to stop the experiment**. Again, this is somewhat difficult since proteins (even with SDS bound) do not absorb in the visible spectrum or have a color! i.e. we cannot simply look at the gel to determine when the proteins have been separated or reached the bottom of the gel.

Therefore, it is common to include in the protein sample a small anionic dye molecule (e.g. bromophenol blue).

Typically, all these reagents (SDS, reducing agents, dye) are pooled in a 'Sample Buffer'. 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

Once the proteins are separated within the gel they are still invisible and must be visualized by staining. A common

protein stain is Coomassie Brilliant Blue R-250. The fixed gel is incubated in a solution of “Coomassie stain” 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.

Since 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 (Molecular Weight Marker, Standard, or Ladder) is included alongside the samples in the same gel then a crude estimation can be made about the protein size.

[Note: Accurate estimations can be made by measuring the distance a protein migrated in a gel and establishing a standard curve.]

For our purposes, remember we started with wanting a way to ‘know’ if a protein(s) is/are present in the sample. If we started with knowing what the molecular weight was then SDS-PAGE provides a good way of visualizing the progress of a protein purification scheme.

Below is an example from a scientific paper that illustrates the latter.

Here the authors are isolating (purifying) a particular

enzyme in Snake venom to study the mechanism of muscle damage.

Problem Solving Exercise

Study the figure below.

- Look at the axes and labels.
- Try to correlate the graphical presentation with your understanding of chromatographic techniques.
 - (A) (C) and (E) are chromatograms.
What does the Y-axis represent? What are the peaks?
- B, D and F are SDS-PAGE gels.
 - What do the numbers in left represent?
What are the fractions? What do the bands show?
- Fractions 62–67 were *pooled* and subject to gel filtration chromatography.
 - Notice the difference in the SDS-PAGE

gel image between D and F ? What do you think happened during gel filtration?

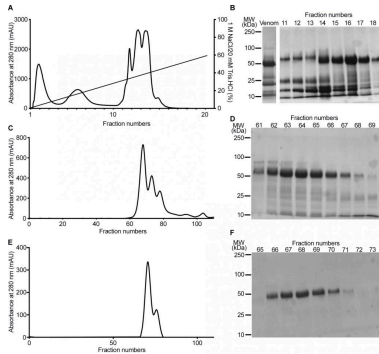


Figure 2. 6. Image from <https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0007041#> (© 2019 Williams et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution License](#) (A) chromatogram demonstrates the purification profile of 10mg of whole *C. atrox* venom by anion exchange chromatography. **B**, a Coomassie-stained gel displays the protein profile of whole *C. atrox* venom and fractions 11–18 of anion exchange chromatography. **A** chromatogram (**C**) and Coomassie-stained gel (**D**) show the purification profile of fractions 14–18 of anion exchange chromatography by gel filtration. **E**, a chromatogram of the second step of gel filtration for

2.5.2 Western Blotting

While for many purposes SDS-PAGE followed by staining is sufficient to identify the presence of a protein by molecular weight, it does not tell you the **identity** of the protein beyond that! Western blotting is a technique that allows for the visualization of a **specific** protein in a gel.

In a western blot procedure, proteins are first separated on an SDS-PAGE gel and then transferred to a membrane. This membrane replica is treated with antibodies that specifically recognize a protein or epitope of interest. Additional processing steps generate a signal at the

fractions 62–67 from the previous step, and (F) a Coomassie-stained gel shows the purified protein at approximately 50kDa.

position of the bound antibody. Between the steps, various washes are done to increase the signal-to-noise ratio on the final, developed

blot.

Steps of Western Blot after SDS page include:

- Electrophoretic transfer of proteins from an SDS-PAGE gel to a membrane
- Blocking of nonspecific protein binding sites on transfer membranes
- Incubation of the membrane with a primary antibody specific for the epitope of interest
- Incubation with a secondary antibody that recognizes primary antibodies
- Visualization of bound antibodies

The major steps in a typical western blot are diagrammed below and relevant ones greater detail in sections that follow.

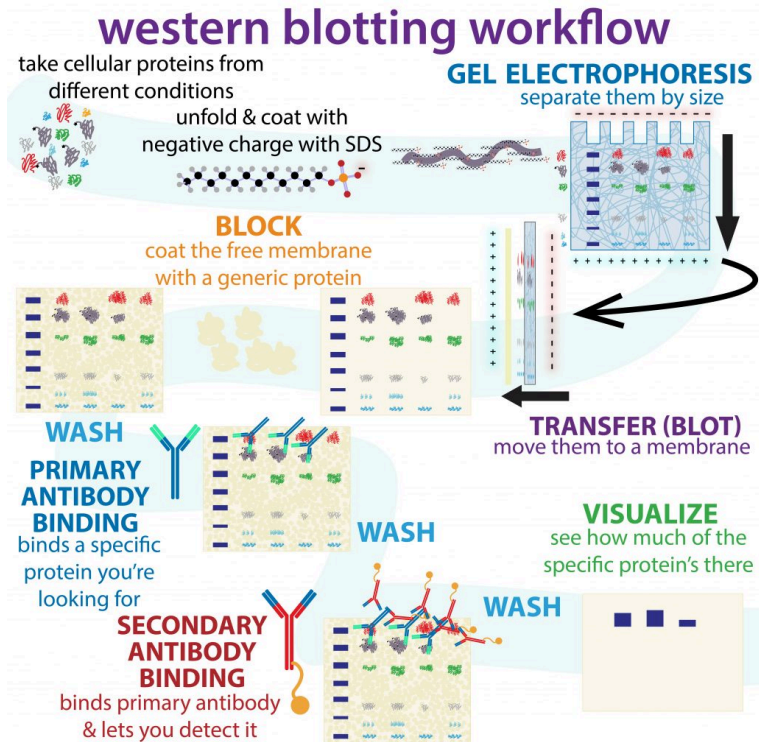


Figure 2.7 An infographic overview of the workflow for performing a western blot. Image credit: Bumbling Biochemist. Graphic made available under Biochemlife, CC BY-SA 4.0, via Wikimedia Commons. Artist URL here: <https://thebumblingbiochemist.com/graphics/>

Electrophoretic transfer of proteins from an SDS-PAGE gel to a membrane

A replica of the SDS-PAGE gel is generated by **transferring**

proteins electrophoretically to a synthetic membrane with a high protein binding capacity.

During the transfer process, the gel and membrane are placed directly against each other within a “sandwich” of pre-wet filter papers and foam pads (see the video). During the electrophoretic transfer, the current should flow evenly across the entire surface area of the gel. After the electrophoretic transfer, the membrane replica with the transferred proteins can be allowed to dry out and stored for later visualization with antibodies.

Blocking of non-specific protein binding sites on membranes

The transfer membranes used in western blots bind proteins nonspecifically. Before the membranes are incubated with specific (and expensive) antibodies, they must be pretreated with blocking solutions that contain high concentrations of abundant (and cheap) proteins to saturate non-specific binding sites. Think of this step as analogous to an artist priming a canvas with a lower quality paint before the more expensive media is applied. If the transfer membranes are not adequately blocked before the antibody is applied, the nonspecific sites on the membranes will absorb some of the antibodies, reducing the amount of antibody available to bind the target proteins. I

Primary, Secondary antibody, and Visualization of proteins.

Protein detection usually employs two antibodies, the first of which is not labeled.

The primary antibody: specifically binds to the protein of interest on the blot. Increasingly, researchers are using epitope-tagged proteins in their experiments, because antibodies against naturally- occurring proteins are expensive and time-consuming to prepare. In addition, an antibody directed against an epitope can be used to detect many different proteins carrying that same epitope.

The secondary antibodies: Are designed to bind the FC fragments of primary antibodies and also carries an enzyme or reagent which can cause a reaction to produce a color upon further treatment.

Thus, the primary antibody binds the protein on the membrane, the secondary antibody binds the primary antibody and in the end, if the molecule of interest is in the original mixture, it will “light” up and reveal itself on the blot.

This sandwich methods allow for better signal and is cost-effective since secondary antibodies can be used with a number of different types of primary antibodies.

Co-immunoprecipitation (Co-IP)

A commonly used technique to study protein-protein

interactions inside cells is an extension of IP called Co-IP for short if antibodies against the proteins of interest are available.

The principle is similar to that of a fishing line with bait. The goal is to use the potential of IP reactions to capture and purify the primary target (i.e., the antigen) but also any other protein that may naturally be associated with it inside the cells. The presence of the second protein (binding partners) is visualized during the western blotting stage by using antibodies to the different components.

In other words- you fish with one antibody. If interactions inside the cell are real then you should pull out binding partners by association.

[Dr. Mehta's Lecture Video Playlist: Affinity Chromatography, Western Blot, Co-ip](#)

For a complete protein purification video from start to finish see below. This video was made by graduate teaching assistants for this course.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://iu.pressbooks.pub/iul211smehta/?p=168>

Remember to:

1. Watch the Lecture videos that cover the material above.
This will help to clarify or reinforce certain concepts if they were unclear.
2. Complete the associated Lecture Quickchecks.
3. Begin work on Problem Set

References and Attributions

1. The Nobel Prize in Physiology or Medicine 1923.
NobelPrize.org. Nobel Prize Outreach AB 2021. Sun. 25 Jul 2021. <<https://www.nobelprize.org/prizes/medicine/1923/summary/>>
2. Williams, Harry F., Ben A. Mellows, Robert Mitchell, Peggy Sfyri, Harry J. Layfield, Maryam Salamah, Rajendran Vaiyapuri, et al.. 2019. Purification of CAMP from the venom of *C. atrox*. (version 1). PLOS Neglected Tropical Diseases. <https://doi.org/10.1371/journal.pntd.0007041.g001>.

This chapter is curated, adapted, and modified from the following **CC-licensed sources**:

- **“Fractionation and Chromatography Techniques” by Kevin Ahern, Indira Rajagopal,**

& Taralyn Tan, LibreTexts is licensed under CC BY-NC-SA. The entire textbook is available for free from the authors at

<http://biochem.science.oregonstate.edu/content/biochemistry-free-and-easy>.

- “Protein Purification” by Michael Blaber, Licensed under CC-BY-NC-SA. Available at: <https://bio.libretexts.org/@go/page/18140>]
- **Western blot: ” Western blots involve many steps” by Clare M. O’Connor, LibreTexts is licensed under CC BY-NC-SA.** Retrieved July 24, 2021, from <https://bio.libretexts.org/@go/page/17590>.

3.

NUCLEIC ACIDS, IDENTITY OF DNA AS MOLECULE OF INHERITANCE

3.1 Introduction: The Stuff of Genes

You have learned that all of Earth's billions of living things are kin to each other. Every living thing does one thing the same way: To make more of itself, it first copies its molecular instruction manual—its genes—and then passes this information on to its offspring. This cycle has been repeated for three and a half billion years.

But how did we and our very distant relatives come to look so different and develop so many different ways of getting along in the world? A century ago, researchers began to answer that question with the help of a science called genetics.

When genetics first started, scientists didn't have the tools they have today. They could only look at one gene, or a few

genes, at a time. Now, researchers can examine all of the genes in a living organism—its genome—at once. They are doing this for organisms on every branch of the tree of life and finding that the genomes of mice, frogs, fish, and a slew of other creatures have many genes similar to our own.” (1)

It’s likely that when you think of heredity you think first of DNA, but in the past few years, researchers have made surprising findings of another molecular actor that plays a starring role- RNA.

DNA and RNA are one of the four biological macromolecules that you began learning about in BIOL-112 (or through AP Bio). Recall that nucleic acids are made up of monomers called nucleotides joined together by strong covalent bonds.

Here we begin with a **quick review** of components of nucleic acids (which were identified long before it was known that DNA is the stuff of genes). We then look at classic experiments that led to our understanding that genes are composed of DNA.

Learning Objectives

Level 1 and 2 are factual information, knowledge-based. Level up indicated by the “Target” symbol is the goal.

When you have mastered the information in this chapter, you should be able to:

Level 1 and 2 (Knowledge and Comprehension)

- Be able to identify the sugar, phosphate, and nitrogenous base portions of a nucleotide.
- Be able to identify a major structural feature that distinguishes a purine nucleotide from a pyrimidine and to justify the specific pairings of these types of nucleotides in the double-stranded structure of DNA.
- Be able to distinguish a ribonucleotide from a deoxyribonucleotide
- Be able to identify the 5' and 3' ends of a nucleic acid strand and know how the two DNA strands are oriented in the double helix.
- Describe the various features of the Watson-Crick double helix model of DNA.
- Label a diagram of a dsDNA molecule to show its features.
- Identify errors in a diagram showing a dsDNA structure or base-pairing of nucleotides.

⊕ Level Up (Application, Analysis, Synthesis)

- Explain and understand the experimental

work leading to the conclusions that DNA is heredity material. ***Predict what the conclusions of experiments discussed would be given a different result.***

- Predict alternative conclusions for structure of DNA when provided different data from Chargaff.
- Justify/hypothesize why the chemical structure of DNA is better suited to long-term information storage than that of RNA.
- Be able to explain why A/T-rich DNA strands associate more weakly than G/C-rich DNA strands.
- Be able to give examples of data that contributed to Watson and Crick's proposal of the double-helical, base-paired structure of DNA.
- Explain the chemical basis of molecular hybridization.
- Explain why T_m is related to base composition.

3.2 Chemistry of Nucleic Acids

Our current understanding of DNA began with the discovery of nucleic acids followed by the development of the double-helix model. In the 1860s, Friedrich Miescher, a physician by profession, isolated phosphate-rich chemicals from white blood cells (leukocytes). He named these chemicals (which would eventually be known as DNA) *nuclein* because they were isolated from the nuclei of the cells.

3.2.1 The building blocks of nucleic acids are nucleotides.

The term nucleotide refers to the building blocks of both DNA (deoxyribonucleoside triphosphates, dNTPs) and RNA (ribonucleoside triphosphates, NTPs). In order to discuss this important group of molecules, it is necessary to define some terms.

Nucleotides contain three

Link to Learning

To see Miescher conduct his experiment that led to his discovery of DNA and associated proteins in the nucleus, click

through [this review](#).

primary structural components. These are a nitrogenous base, a pentose sugar, and at least one phosphate.

Molecules that contain only sugar and a nitrogenous base (no phosphate) are called nucleosides.

The nitrogenous bases found in nucleic acids include adenine and guanine (called purines) and cytosine, uracil, or thymine (called pyrimidines). There are two sugars found in nucleotides – deoxyribose and ribose (Figure 3.1).

By convention, the carbons on these sugars are labeled 1' to 5'. (This is to distinguish the carbons on the sugars from those on the bases, which have their carbons simply labeled as 1, 2, 3, etc.)

Deoxyribose differs from ribose at the 2' position, with ribose having an OH group, whereas deoxyribose has H.

Nucleotides containing deoxyribose are called deoxyribonucleotides and are the forms found in DNA.

Nucleotides containing ribose are called ribonucleotides and are found in RNA. Both DNA and RNA contain nucleotides with adenine, guanine, and cytosine, but with very minor exceptions, RNA contains uracil nucleotides, whereas DNA contains thymine nucleotides.

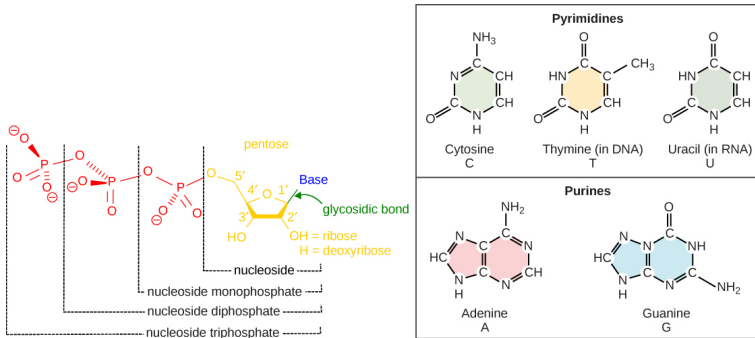


Figure 3.1. Schematic showing the structure of nucleoside triphosphates. This figure also shows the five common nitrogenous bases found in DNA and RNA on the right. Image credit: Public domain, via Wikimedia Commons. [File URL: https://upload.wikimedia.org/wikipedia/commons/b/b9/Nucleotides.png](https://upload.wikimedia.org/wikipedia/commons/b/b9/Nucleotides.png)

VISUAL CONNECTION

The images above illustrate the five bases of DNA and RNA.

Examine the images and explain why these are called “nitrogenous bases.”

How are the purines different from the pyrimidines? How is one purine or pyrimidine

different from another, e.g., adenine from guanine?

How is a nucleos**ide** different from a nucleot**ide**?

The purines have a double ring structure with a six-membered ring fused to a five-membered ring. Pyrimidines are smaller in size; they have a single six-membered ring structure.

Scientists classify adenine and guanine as purines, and cytosine, thymine, and uracil as pyrimidines. (See Mnemonic to help you remember).

Useful Mnemonic

Purines
are
Adenine
and

In molecular biology shorthand, we know the nitrogenous bases by their symbols A, T, G, C, and U.

DNA contains A, T, G, and C; whereas, RNA contains A, U, G, and C.

Building Nucleic Acid strands

The substrates for making DNA or RNA polymers are dNTPS (de-oxyribonucleoside triphosphates- DNA) or NTPs (for RNA).

Each DNA strand is built from dNTPs by the formation of a phosphodiester bond, catalyzed by DNA polymerase, between the 3'OH of one nucleotide and the 5' phosphate of the next.

The result of this directional growth of the strand is that one end of the strand has a free 5' phosphate and the other a free 3' hydroxyl group (Figure 3.2). These are designated as the 5' and 3' ends of the strand.

The three phosphates are named using greek letters α (alpha), β (beta), and γ (gamma) with alpha being the phosphate linked to the 5' carbon.

During the formation of a nucleic acid polymer, the incoming nucleotides are added to a growing chain. During the reaction, the two outer phosphate groups (beta and gamma) from the incoming dNTP are released. These two outer phosphates are called **pyrophosphate** after they are released.

Guani
ne or
PURe
As
Gold!

The remaining
(T, C and U are
Pyrimidines)

Remember This: In-vitro (in a test tube) assays to study the formation of DNA (DNA replication!) or RNA (Transcription) utilize radiolabeled dNTP's. Since the **alpha phosphate becomes part of the polymer**, it is the one that is in the form of a radioactive isotope of Phosphorous (^{32}P). The incorporation of the label into the growing DNA or RNA strand can then be quantified.

RNA: Diverse Roles

We should take a moment to talk about RNA. While it's true that DNA is the basic ingredient of our genes and, as such, it often steals the limelight from RNA, the other form of genetic material inside our cells.

“But, while they are both types of genetic material, RNA and DNA are rather different.

The chemical units of RNA are like those of DNA, except that RNA has the nucleotide uracil (U) instead of thymine (T). Unlike double-stranded DNA, RNA usually comes as only a single strand. And the nucleotides in RNA contain ribose sugar molecules in place of deoxyribose.

RNA is quite flexible—unlike DNA, which is a rigid, spiral-staircase molecule that is very stable. RNA can twist itself into

a variety of complicated, three-dimensional shapes. RNA is also unstable in that cells constantly break it down and must continually make it fresh, while DNA is not broken down often. RNA's instability lets cells change their patterns of protein synthesis very quickly in response to what's going on around them.

Many textbooks still portray RNA as a passive molecule, simply a “middle step” in the cell's gene-reading activities. But that view is no longer accurate.” (1)

Each year, researchers unlock new secrets about RNA. While we will leave RNA behind for the remainder of this chapter, we will come back to this amazing molecule and its role in the regulation of gene expression and genetic medicine in later chapters.

Study Tip: Watch Dr. Mehta Lecture Video before continuing: [L211 Nucleic Acids \(Review\)](#)

3.3 Identity of DNA as Genetic Material

That eukaryotic cells contain a nucleus was understood by the late 19th century. By then, histological studies had shown

that nuclei contained largely proteins and DNA. At around the same time, the notion that the nucleus contains genetic information was gaining traction.

However, DNA was thought to be a monotonous, repetitive string of nucleotides, which would not be useful for information storage. The prevailing hypothesis by Phoebus Levene was the “**tetranucleotide hypothesis**,” which postulated that DNA’s four bases are present in equal amounts and repeat over and over again along the chromosome in a fixed pattern. This idea that a simple molecule made up of only 4 nucleotides couldn’t possibly account for the inheritance of so many different physical traits.

The recognition that enzyme activities were inherited in the same way as morphological characteristics led to the ***one-gene-one enzyme*** hypothesis that earned G. W. Beadle, E. L. Tatum, and J. Lederberg the 1958 Nobel Prize for Physiology and Medicine.

When enzymes were later shown to be proteins, the hypothesis became ***one-gene-one protein***. When proteins were shown to be composed of one or more polypeptides, the final hypothesis became ***one-gene-one-polypeptide***. However, this relationship between genes and polypeptides failed to shed any light on how DNA might be the genetic material.

In fact, quite the contrary! As chains of up to 20 different amino acids, polypeptides and proteins had the potential for enough structural diversity to account for the growing

number of heritable traits in a given organism. Thus, proteins seemed more likely candidates for the molecules of inheritance.

The experiments you will read about here began around the start of World War I and lasted until just after World War 2. During this time, we learned that DNA was no mere tetramer, but was in fact a long polymer.

This led to some very clever experiments that eventually forced the scientific community to the conclusion that DNA, not protein, was the genetic molecule, despite being composed of just four monomeric units.

3.3.1 Transforming Principle – Griffiths Experiments

in 1928, British bacteriologist Frederick Griffith reported the first demonstration of bacterial transformation—a process in which external DNA is taken up by a cell, thereby changing its morphology and physiology.

He had discovered three immunologically different strains of *Streptococcus pneumonia* (Types I, II and III). The virulent strain (Type III) was responsible for much of the mortality during the **Spanish Flu** (influenza) *pandemic* of 1918-1920. This pandemic killed between 20 and 100 million people, many because the influenza viral infection weakened the immune system of infected individuals, making them susceptible to bacterial infection by *Streptococcus pneumonia*.

In the 1920s, Frederick Griffith was working with virulent

wild type (*Type III*) and **benign** (*Type II*) strains of *S. pneumonia*. The two strains were easy to tell apart in petri dishes because the virulent strain grew as morphologically *smooth colonies*, while the benign strain formed *rough colonies*. For this reason, the two bacterial strains were called **S** (smooth) and **R (rough)**, respectively.

When Griffith injected the living S strain into mice, they died from pneumonia. In contrast, when Griffith injected the live R strain into mice, they survived. In another experiment, when he injected mice with the heat-killed S strain, they also survived. This experiment showed that the capsule alone was not the cause of death.

In a third set of experiments, a mixture of live R strain and heat-killed S strain were injected into mice, and—to his surprise—the mice died.

Upon isolating the live bacteria from the dead mouse, only the S strain of bacteria was recovered. When this isolated S strain was injected into fresh mice, the mice died. Griffith concluded that something had passed from the heat-killed S strain into the live R strain and transformed it into the pathogenic S strain. He called this the *transforming principle* (Figure). These experiments are now known as Griffith's transformation experiments.

TIME TO THINK.

What is the significance of Griffiths's experiments as it pertains to biotechnology?

Answer: This was the first experimental demonstration of gene /nucleic acid transfer into cells! A method that powers today's biotechnology industry! Diabetic patients, for example, are treated with human insulin made by bacteria **transformed** with the human insulin gene.

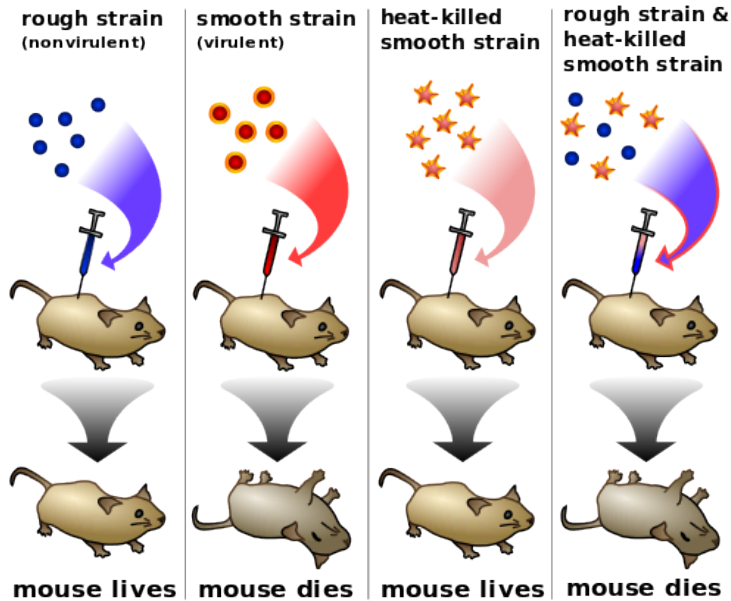


Figure 3.3 The experiments of F. Griffith demonstrating the existence of a chemical that could transfer a genetic trait (i.e., virulence) between bacteria (see text for details). [Image credit: No machine-readable author provided. Madprime assumed \(based on copyright claims\), CC0, via Wikimedia Commons.](#)

3.3.2 The Avery-MacLeod-McCarty and the Hershey-Chase experiments.

While Griffith didn't know the chemical identity of his

transforming principle, his experiments led to studies that proved DNA to be ***the stuff of genes***. With improved molecular purification techniques developed in the 1930s, O. Avery, C. MacLeod, and M. McCarty transformed R cells *in vitro* (that is, without the help of a mouse!), in other words, ‘a **biochemical assay**’ for transformation!

They isolated the S strain from the dead mice and isolated the proteins and nucleic acids (RNA and DNA) as these were possible candidates for the molecule of heredity. They used enzymes that specifically degraded each component and then used each mixture separately to transform the R strain. They found that when DNA was degraded, the resulting mixture was no longer able to transform the bacteria, whereas all of the other combinations were able to transform the bacteria.

This led them to conclude that DNA was the transforming principle.

WATCH: Whiteboard video



One or more interactive elements has
been excluded from this version of the
text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=428#oembed-1>

Attribution: This video is from The Explorer's Guide to Biology (XBio) and is offered under a CC-BY-NC licensing agreement. For more information, see Ronald Vale's Narrative on DNA Structure in The Explorer's Guide to Biology (explorebiology.org/collections/genetics/dna-structure).

Although the experiments of Avery, McCarty, and McLeod had demonstrated that DNA was the informational component transferred during transformation, DNA was still considered to be too simple a molecule to carry biological information. Since this result stood against the dogma of the time, it was not readily accepted.

The decisive experiment, conducted by Martha Chase and Alfred Hershey in 1952, provided confirmatory evidence that DNA was indeed the genetic material and not proteins.

Chase and Hershey were studying a bacteriophage—a virus that infects bacteria. Viruses typically have a simple structure: a protein coat called the capsid, and a nucleic acid core that contains the genetic material (either DNA or RNA). The bacteriophage infects the host bacterial cell by attaching to its surface, and then it injects its nucleic acids inside the cell. The phage DNA makes multiple copies of itself using the host machinery, and eventually the host cell bursts, releasing a large number of bacteriophages. Hershey and Chase selected radioactive elements that would specifically distinguish the protein from the DNA in infected cells.

They labeled one batch of phage with radioactive sulfur, ^{35}S , to label the protein coat. Another batch of phage was labeled with radioactive phosphorus, ^{32}P .

Because phosphorous is found in DNA, but not protein, the DNA and not the protein would be tagged with radioactive phosphorus. Likewise, sulfur is absent from DNA but is present in several amino acids such as methionine and cysteine.

Each batch of phage was allowed to infect the cells separately. After infection, the phage bacterial suspension was put in a blender, which caused the phage coat to detach from the host cell. Cells exposed long enough for infection to occur were then examined to see which of the two radioactive molecules had entered the cell. The phage and bacterial suspension were spun down in a centrifuge. The heavier bacterial cells settled down and formed a pellet, whereas the

lighter phage particles stayed in the supernatant. In the tube that contained phage labeled with ^{35}S , the supernatant contained the radioactively labeled phage, whereas no radioactivity was detected in the pellet.

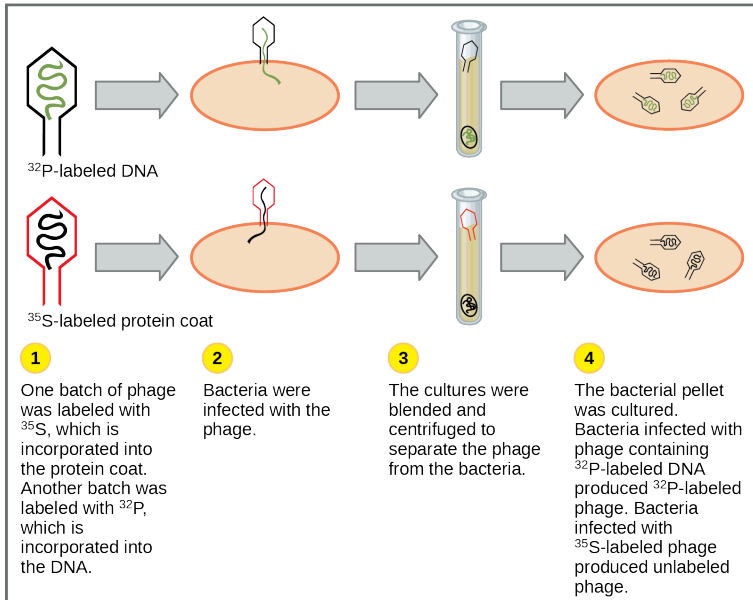


Figure 3.4 Hershey and Chase's experiments demonstrate that DNA is the chemical that makes up genes. Image credit: <https://openstax.org/books/biology-2e/pages/1-introduction>

3.3.3 Chargaff's Data

Around this same time, Austrian biochemist Erwin Chargaff examined the content of DNA in different species and found that the amounts of adenine, thymine, guanine, and cytosine

were not found in equal quantities and that relative concentrations of the four nucleotide bases varied from species to species, but not within tissues of the same individual or between individuals of the same species. He also discovered something unexpected: That the amount of adenine equaled the amount of thymine, and the amount of cytosine equaled the amount of guanine (that is, $A = T$ and $G = C$). Different species had equal amounts of *purines* ($A+G$) and *pyrimidines* ($T + C$), but different ratios of $A+T$ to $G+C$. These observations became known as Chargaff's rules.

Chargaff's findings proved immensely useful when Watson and Crick were getting ready to propose their DNA double helix model!

You can see after reading the past few paragraphs how science builds upon previous discoveries, sometimes in a slow and laborious process.

Key Takeaways

1. DNA was first isolated from white blood cells by Friedrich Miescher, who called it nuclein because it was isolated from nuclei.

2. Frederick Griffith's experiments with strains of *Streptococcus pneumoniae* provided the first hint that DNA may be the transforming principle.
3. Avery, MacLeod, and McCarty showed that DNA is required for the transformation of bacteria.
4. Later experiments by Hershey and Chase using bacteriophage T2 proved that DNA is the genetic material.
5. Chargaff found that the ratio of A = T and C = G, and that the percentage content of A, T, G, and C is different for different species.

3.4 DNA Structure: The Double Helix

When DNA was accepted as the *stuff of genes*, the next questions were

- What did DNA look like?
- How did its structure account for its ability to encode and reproduce life?

While the nature of DNA, its composition of 4 nucleotides had been known for some time, it became mandatory to explain how such a “simple molecule” could inform the thousands of proteins necessary for life.

The answer to this question was to lie at least in part in an understanding of the physical structure of DNA, made possible by the advent of *X-Ray Crystallography*.

3.4.1 Wilkins, Franklin, Watson & Crick – DNA Structure Revealed.

William Astbury demonstrated that high molecular weight DNA had just such a regular structure. His *crystallographs* suggested DNA was a linear polymer of stacked bases (nucleotides), each nucleotide separated from the next by 0.34 nm.

Maurice Wilkins, an English biochemist, was the first to isolate highly pure, high molecular weight DNA. Working in Wilkins laboratory, Rosalind Franklin was able to crystalize this DNA and produce very high-resolution X-Ray diffraction images of the DNA crystals. Franklin’s most famous (and definitive) crystallograph was “Photo 51” (Fig. 3.5 -b).

This image confirmed Astbury’s **0.34 nm** repeat dimension and revealed two more numbers, **3.4 nm**, and **2 nm**, reflecting additional repeat structures in the DNA crystal. When James Watson and Francis Crick got hold of these numbers, they

used them along with other data to build DNA models out of nuts, bolts, and plumbing.

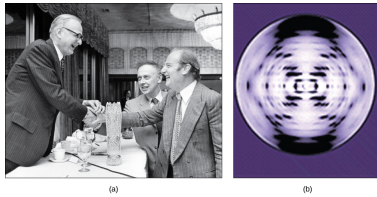


Figure 3.5. The work of pioneering scientists (a) James Watson, Francis Crick, and Maclyn McCarty led to our present-day understanding of DNA. Scientist Rosalind Franklin discovered (b) the X-ray diffraction pattern of DNA, which helped to elucidate its double-helix structure. (credit a: modification of work by Marjorie McCarty, Public Library of Science)

Watson and Crick proposed that DNA is made up of two strands that are twisted around each other to form a right-handed helix. Base pairing takes place between a purine and pyrimidine on opposite strands, so that A pairs with T, and G pairs with C (suggested by Chargaff's Rules).

Thus, adenine and thymine are complementary base pairs, and cytosine and guanine are also

complementary base pairs.

The base pairs are stabilized by hydrogen bonds: *adenine and thymine form two hydrogen bonds and cytosine and guanine form three hydrogen bonds.*

The two strands are anti-parallel in nature; that is, the 3' end of one strand faces the 5' end of the other strand. The

sugar and phosphate of the nucleotides form the backbone of the structure, whereas the nitrogenous bases are stacked inside, like the rungs of a ladder.

Each base pair is separated from the next base pair by a distance of 0.34 nm, and each turn of the helix measures 3.4 nm. Therefore, 10 base pairs are present per turn of the helix. The diameter of the DNA double-helix is 2 nm, and it is uniform throughout.

Only the pairing between a purine and pyrimidine and the antiparallel orientation of the two DNA strands can explain the uniform diameter.

The twisting of the two strands around each other results in the formation of uniformly spaced major and minor grooves or indentations on the sides of the helix (Figure 3.6).

Protein-DNA interactions occur in these spaces. The grooves expose the edges of the bases to the external environment, making them accessible for protein binding.

The nature of the atoms that are exposed in the grooves forms a type of code or pattern that proteins that bind DNA use to recognize sequences without needing to pry open the DNA helix! Discrimination of the different sequences must be made by having access to the bases inside the structure since the backbone structure is common to all sequences of DNA.

[See video below on DNA Structure and Lecture Videos in the playlist for further explanation]

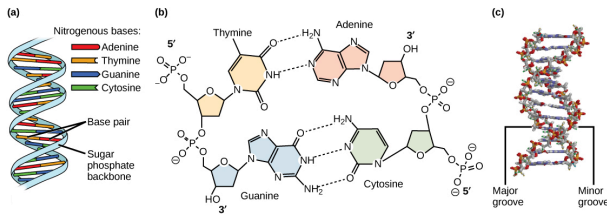


Figure 3.6 DNA has (a) a double helix structure and (b) phosphodiester bonds; the dotted lines between Thymine and Adenine and Guanine and Cytosine represent hydrogen bonds. The (c) major and minor grooves are binding sites for DNA binding proteins during processes

such as
transcription (the
copying of RNA
from DNA)
and
replication. Image
credit: From
<https://openstax.org/books/biology-2e/pages/1-introduction>

TIME to THINK:

To function as the molecule of inheritance the molecule should be stable, store information (vast amount to account for all the variability), capacity to change (to account for evolution and diversity of life), and faithfully replicate (to account for the continuity of life and perpetuation of species). It is not surprising that Proteins were initially favored over DNA as being the molecule of inheritance given the tall order.

However, once Watson and Crick proposed and published their model of DNA as a double helix the answer was obvious. The structure perfectly suited its function!

Think about how the simple elegant structure and features of the double helix can fulfill the requirements needed of a molecule of inheritance. Jot down your thoughts.

Dr. Mehta Lecture Videos: [L211 Identity of DNA as Genetic Material](#)

Watch this video on DNA structure:



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=428#oembed-2>

Molecular Biology in the News: “New Life -Forms, No DNA Required”

READ: “New Life -Forms, No DNA Required” found here [Scientific American Article Link](#)

NOTE: Your CANVAS course site includes .pdfs of this article.

The DNA in our cells is not naked. All of our DNA is always complexed with proteins and the double helix is coiled and supercoiled.

We will learn more about this in the upcoming chapters.

Before you continue you should

1. Watch the Lecture videos that cover the material above.
 2. Complete the associated Lecture Quickcheck.
 3. READ the Molecular Biology in the News article.
-

3.5 Analysis of Nucleic Acids

When Watson and Crick submitted their historic one-page paper on the model DNA structure they included a provocative message.

“It has not escaped our notice that the specific pairing that we have postulated immediately suggests a possible copying mechanism for the genetic material.”

While the exact mechanism of replication was eventually experimentally proven, fundamental to the process was the separation of DNA strands. Experimentally scientists had shown that when DNA molecule was heated or treated with chemicals (like Urea) the viscosity of the DNA solution dropped.

The forces holding duplexes together include hydrogen bonds between the bases of each strand that, like the hydrogen bonds in proteins, can be broken with heat or urea. (Another important stabilizing force for DNA arises from the stacking interactions between the bases in a strand.)

The separation of strands of the double helix into single strands is **Denaturation**.

Single strands absorb light at 260 nm more strongly than double strands. This is known as the **hyperchromic effect** (Figure 3.) and is a consequence of the disruption of interactions among the stacked bases. The changes in absorbance allow one to easily follow the course of DNA denaturation.

Denatured duplexes can readily **renature** when the temperature is lowered below the “**melting temperature**” or **T_m**, the temperature at which half of the DNA strands are in duplex form. Under such conditions, the two strands can reform hydrogen bonds between the complementary sequences, returning the duplex to its original state.

In most organisms, the T_m of the chromosomal DNA ranges from 85-100 degrees C. It is possible to determine the composition of the DNA experimentally from its T_m because the T_m of DNA is directly proportional to the GC content of the DNA.

For DNA, this principle of strand separation and renaturation (also called hybridization or annealing) are important for many techniques- and most importantly will feature again when we learn about polymerase chain reaction (PCR).

Historically, the separation of DNA and subsequent renaturation was used to assess the complexity of the genome, % GC content, and similarity between sequences.

Did I get this?



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=428#h5p-8>

Agarose Gel Electrophoresis

DNA molecules can be separated by size and visualized using electrophoresis. The principle is similar to that discussed for proteins with a few exceptions.

1. Since DNA is naturally negatively charged there is no need to use detergents like SDS. DNA molecule migrates towards the positively charged electrode (cathode).
2. The polymer used for DNA electrophoresis is Agarose (hence the name).

3. Agarose gels are usually poured and run horizontally.

The rate of migration is directly dependent on the ability of each DNA molecule to worm or wiggle its way through the sieving gel. The agarose matrix provides openings for macromolecules to move through. The largest macromolecules have the most difficult time navigating through the gel, whereas the smallest macromolecules slip through it the fastest.

4. Visualization of DNA occurs by using a fluorescent dye **ethidium bromide**.

This compound contains a planar group that **intercalates** between the stacked bases of DNA. The dye is usually incorporated into the gel and running buffer, the stain is visualized by irradiating with a UV light source (i.e. using a transilluminator)

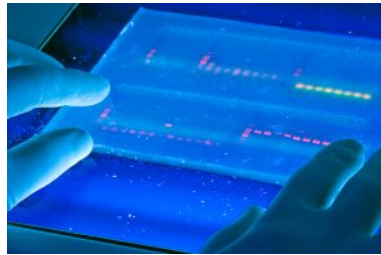


Figure 3.7. Agarose gel with UV illumination – Ethidium bromide-stained DNA glows orange (close-up). [Image credit: School of Natural Resources from Ann Arbor, CC BY 2.0 , via Wikimedia Commons.](#)

and photographing with polaroid film. (Figure 3.7)

Ethidium being a DNA intercalating agent is a powerful mutagen! There are several newer DNA stains by companies that are advertised as being safer and less mutagenic.

5. Reference DNAs of known sizes are alongside the samples (DNA ladders). This allows ones to determine the sizes of the DNA fragments in the sample.

It is useful to note that, by convention, DNA fragments are not described by their molecular weights (unlike proteins), but by their **length in base-pairs**

(bp) or kilobases (kb).

NOTE: Like for proteins the principle of separation relies on DNA molecules being linear. However, as we shall soon see genomes are often circular (bacteria) or really long (millions of bps!). Typically DNA for electrophoresis is linearized using special enzymes (restriction endonucleases) that will cut DNA at specific sequences OR shorter pieces of DNA are analyzed.

Below is a Virtual Lab Simulation of Agarose Gel Electrophoresis:

explorebiology.org/activities/agarose-gel-electrophoresis

This virtual lab simulation video takes you through the steps of agarose gel electrophoresis, a method used in biology and biotechnology to separate different-sized DNA molecules on the basis of their movement in an electric field.

This video also describes how this method can be used to analyze the outcomes of an experiment with DNA.

If you want to simulate running your own gel, see the Gel Electrophoresis Activity by Shawn Douglas in The Explorer's Guide to Biology (explorebiology.org/activities/agarose-gel-electrophoresis).



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=428#oembed-3>

Dr. Mehta Lecture Video: [L211 Analysis of Nucleic Acids](#)

Remember to:

1. Watch the Lecture videos that cover the material above. This will help to clarify or reinforce certain concepts if they were unclear.
 2. Complete the associated Lecture Quickchecks.
-

3.6 From Genes to Genomes

We use the word “**genome**” to describe all of the genetic material of the cell. That is, a genome is an entire sequence of nucleotides in the DNA that is in all of the chromosomes of a cell. When we use the term genome without further qualification, we are generally referring to the chromosomes in the nucleus of a eukaryotic cell.

As you know, eukaryotic cells have organelles like mitochondria and chloroplasts that have their own DNA. These are referred to as the mitochondrial or chloroplast genomes to distinguish them from the nuclear genome.

Starting in the 1980s, scientists began to determine the complete sequence of the genomes of many organisms, in the hope of better understanding how the DNA sequence specifies cellular functions.

The Human Genome Project (HGP) was one of the great feats of exploration in history. Rather than an outward exploration of the planet or the cosmos, the HGP was an inward voyage of discovery led by an international team of researchers looking to sequence and map all of the genes — together known as the genome — of members of our species, *Homo sapiens*.

Beginning on October 1, 1990 and completed in April 2003, the HGP gave us the ability, for the first time, to read nature’s complete genetic blueprint for building a human being.

The technological advances in the ability to sequence DNA that were a direct result of this colossal endeavor resulted in a new field of study: Genomics; the comprehensive study of whole sets of genes and their interactions.

The results are deposited into publicly available databases. You can find many of them at the [National Center for Biotechnology Information](#).

The number of available, completely sequenced genomes numbers in the tens of thousands—over 2,000 eukaryotic genomes, over 600 archaeal genomes, and nearly 12,000 bacterial genomes. Tens of thousands of more genome sequencing projects are in progress.

As the sequence databases compile ever more information, the fields of computational biology and bioinformatics have arisen, to analyze and organize the data in a way that helps biologists understand what the information in DNA means in the cellular context.

With this many genome sequences available— scientists have been asking many questions about what we see in these genomes. What patterns are common to all genomes? How many genes are encoded in genomes? How are these organized? How many different types of features can we find? What do the features that we find do? How different are the genomes from one another?

3.6.1 Diversity of genomes

Diversity of sizes, number of genes, and chromosomes

Let's start by examining the range of genome sizes. In the table below, we see a sampling of genomes from the database. We can see that the genomes of free-living organisms range tremendously in size. The smallest known genome is encoded in 580,000 base pairs while the largest is 150 billion base pairs—for reference, recall that the human genome is 3.2 billion base pairs. That's a huge range of sizes.

NUCLEIC ACIDS, IDENTITY OF DNA AS MOLECULE OF INHERITANCE | 153

organism	genome size (base pairs)	protein coding genes	number of chromosomes
model organisms			
model bacteria <i>E. coli</i>	4.6 Mbp	4,300	1
budding yeast <i>S. cerevisiae</i>	12 Mbp	6,600	16
fission yeast <i>S. pombe</i>	13 Mbp	4,800	3
amoeba <i>D. discoideum</i>	34 Mbp	13,000	6
nematode <i>C. elegans</i>	100 Mbp	20,000	12 (2n)
fruit fly <i>D. melanogaster</i>	140 Mbp	14,000	8 (2n)
model plant <i>A. thaliana</i>	140 Mbp	27,000	10 (2n)
moss <i>P. patens</i>	510 Mbp	28,000	27
mouse <i>M. musculus</i>	2.8 Gbp	20,000	40 (2n)
human <i>H. sapiens</i>	3.2 Gbp	21,000	46 (2n)
viruses			
hepatitis D virus (smallest known animal RNA virus)	1.7 Kb	1	ssRNA
HIV-1	9.7 kbp	9	2 ssRNA (2n)
influenza A	14 kbp	11	8 ssRNA
bacteriophage λ	49 kbp	66	1 dsDNA
Pandoravirus salinus (largest known viral genome)	2.8 Mbp	2500	1 dsDNA
organelles			
mitochondria - <i>H. sapiens</i>	16.8 kbp	13 (+22 tRNA +2 rRNA)	1
mitochondria - <i>S. cerevisiae</i>	86 kbp	8	1
chloroplast - <i>A. thaliana</i>	150 kbp	100	1
bacteria			
<i>C. rudii</i> (smallest genome of an endosymbiont bacteria)	160 kbp	182	1
<i>M. genitalium</i> (smallest genome of a free living bacteria)	580 kbp	470	1
<i>H. pylori</i>	1.7 Mbp	1,600	1
Cyanobacteria <i>S. elongatus</i>	2.7 Mbp	3,000	1
methicillin-resistant <i>S. aureus</i> (MRSA)	2.9 Mbp	2,700	1
<i>B. subtilis</i>	4.3 Mbp	4,100	1
<i>S. cellulosum</i> (largest known bacterial genome)	13 Mbp	9,400	1
archaea			
Nanoarchaeum equitans (smallest parasitic archaeal genome)	490 kbp	550	1
Thermoplasma acidophilum (flourishes in pH<1)	1.6 Mbp	1,500	1
Methanocaldococcus (Methanococcus) jannaschii (from ocean bottom hydrothermal vents; pressure >200 atm)	1.7 Mbp	1,700	1
Pyrococcus furiosus (optimal temp 100°C)	1.9 Mbp	2,000	1
eukaryotes - multicellular			
pufferfish <i>Fugu rubripes</i> (smallest known vertebrate genome)	400 Mbp	19,000	22
poplar <i>P. trichocarpa</i> (first tree genome sequenced)	500 Mbp	46,000	19
corn <i>Z. mays</i>	2.3 Gbp	33,000	20 (2n)
dog <i>C. familiaris</i>	2.4 Gbp	19,000	40
chimpanzee <i>P. troglodytes</i>	3.3 Gbp	19,000	48 (2n)
wheat <i>T. aestivum</i> (hexaploid)	16.8 Gbp	95,000	42 (2n=6x)
marbled lungfish <i>P. aethiopicus</i> (largest known animal genome)	130 Gbp	unknown	34 (2n)
herb plant <i>Paris japonica</i> (largest known genome)	150 Gbp	unknown	40 (2n)

This table shows some genome data for various organisms. 2n = diploid number. Source: <http://book.bionumbers.org/how-big-are-genomes/>

Notice the genome sizes of Prokaryotes in base pairs and compare them with those of multicellular organisms. At first glance, we see that the overall genomes of viruses, archaea, bacteria (and some unicellular eukaryotes) are smaller.

A common-sense assumption about genomes would be that if genes specify proteins, then the more proteins an organism made, the more genes it would need to have, and thus, the larger its genome would be.

However, when we begin to see how much of the genome is devoted to protein-coding genes a different story emerges that indicates that in fact there is NO direct relationship between the complexity of an organism and the size of its genome.

Compare the pufferfish genome to the chimpanzee genome, we note that they encode roughly the same number of genes (19,000), but they do so on dramatically differently sized genomes—400 million base pairs versus 3.3 billion base pairs, respectively.

That implies that the pufferfish genome **must have much less space between its genes** than what might be expected to be found in the chimpanzee genome. Indeed, this is the case, and the difference in **gene density** is not unique to these two genomes.

In fact, we can broadly categorize genomes as being either: Small, compact genomes like those of viruses, archaea, and bacteria or large and expanded where the bulk of the genome is non-coding.

To understand how this could be true, it is necessary to

recognize that while genes are made up of DNA, all DNA does not consist of genes (for purposes of our discussion, we define a gene as a section of DNA that encodes an RNA or protein product).

One of the surprising findings of the human genome data was that less than 2% of the total DNA seems to be the sort of coding sequence that directs the synthesis of proteins. For many years, non-coding DNA in genomes was believed to be useless and was described as “junk DNA” although it was perplexing that there seemed to be so much “useless” sequence. Recent discoveries have, however, demonstrated that much of this so-called junk DNA plays important roles in evolution, as well as in the regulation of gene expression

What is all the “extra stuff” in the eukaryotic genomes?

Introns

We know that even coding regions in our DNA are interrupted by non-coding sequences called introns. This is true of most eukaryotic genomes. An examination of genes in eukaryotes shows that non-coding intron sequences can be much longer than the coding sections of the gene, or exons. Most exons are relatively small, and code for fewer than a hundred amino acids, while introns can vary in size from several hundred base pairs to many kilobase pairs (thousands of base pairs) in length. For many genes in humans, there is much more of intron

sequence than coding (a.k.a. exon) sequence. Intron sequences account for roughly a quarter of the genome in humans.

Introns get removed (spliced) after transcription, something you will learn in upcoming chapters.

Regulatory Sequences

What other kinds of non-coding sequences are there? One function for some DNA sequences that do not encode RNA or proteins is in specifying when and to what extent a gene is used, or expressed. Such regions of DNA are called regulatory regions and each gene has one or more regulatory sequences (**promoters and enhancer** sequences) that control its expression. However, regulatory sequences do not account for all the rest of the DNA in our genomes, either.

Repetitive DNA

More than 50% of the genome and the ‘intergenic’ regions consist of highly repetitive sequences. Some of these are found in regions of the DNA that go on to form structural markers of chromosomes like centromeres and telomeres.

Genome-Wide or Interspersed Repeats

Many of the repetitive sequences are known to be transposable elements (transposons), sections of DNA that can move around within the genome. Sometimes referred to as “jumping genes” these transposable elements can move from one

chromosomal location to another, either through a simple “cut and paste” mechanism that cuts the sequence out of one region of the DNA and inserts it into another location, or through a process called retrotransposition involving an RNA intermediate. These are dispersed throughout the genome. These repetitive DNA elements are further subdivided into two categories based on their length.

Tandem Repeats

In contrast to the repeats that are dispersed, tandem repeats are placed next to each other in an array. Amongst these are **short tandem repeats or STRs**. Each repeat is a unique DNA sequence that ranges from 2 to 6 bp repeated over and over, eg GACA GACA GACA.

The number of repeats is variable in different individuals and since these are inherited they are the basis of forensic genetic analysis to generate a DNA profile of an individual.

Interestingly the fact that our genomes consisted of highly repetitive DNA sequences was known as early as the '60s by denaturation and renaturation experiments!

See below for an example of the kind of scientific application connection with concepts discussed earlier.

[Link to Learning](#)

Go to: <http://www.dnafb.org/31/index.html>

Click through the Animation tab to learn about early experiments revealing the presence of repetitive DNA in genomes.

ENCODE

The sequencing phase of the Human Genome Project provided a massive data set of ordered bases but did not show where genes begin or end or what they do.

The project ENCODE or the Encyclopedia of DNA Elements (ENCODE) set out to identify all functional elements in the human and mouse genome sequences – this includes protein-coding genes, non-protein-coding genes, transcriptional regulatory elements, and sequences that mediate chromosome structure and dynamics. The ENCODE Project started in 2003 with the [ENCODE Pilot Project](#), which focused on 1% of the human genome and is now in its fourth phase.

Dr. Mehta Lecture Video: [L211 Dr. Mehta Genes to Genomes](#)

Before you continue you should

1. Complete the associated Lecture Quickcheck.
2. Complete the Concepts in Context ” 1000 Genome Project”

References and Attributions

This chapter contains material taken from the following CC-licensed content. Changes include rewording, removing paragraphs, and replacing them with original material.

- (1) *Introduction and RNA- Diverse Roles From the New Genetics* is available online at:
<http://publications.nigms.nih.gov/thenewgenetics>.
NIH Publication No. 10-662. Revised April 2010. (US Government Work)

- Ahern K, Rajagopal I and Tan T. (2013). Biochemistry Free for All (Version 1.3). Licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](#). The entire textbook is available for free from the authors at <http://biochem.science.oregonstate.edu/content/biochemistry-free-and-easy>.
- "DNA, Chromosomes and Chromatin" by Gerald Bergstrom, LibreTexts is licensed under [CC BY](#). The chapter can be found online at <https://bio.libretexts.org/@go/page/16463>. The entire textbook Basic Cell and Molecular Biology: What We Know & How We Found Out – 4e can be found at <https://open.umn.edu/opentextbooks/textbooks/cell-and-molecular-biology-2e-what-we-know-how-we-found-out>.

Section 3.6 From Genes to Genomes:

- Genomes: a Brief Introduction. (2019, June 2). <https://bio.libretexts.org/@go/page/9388>
- "Genes and Genomes" by Kevin Ahern, Indira Rajagopal, & Taralyn Tan, LibreTexts is licensed under CC BY-NC-SA. The entire textbook is available free from the authors at <http://biochem.science.oregonstate.edu/content/biochemistry-free-and-easy>

Images

Figure images without any attribution in the caption are licensed under CC-BY 4.0 by OpenStax. Located at: <https://openstax.org/books/biology-2e/pages/14-1-historical-basis-of-modern-understanding>. Access for free at <https://openstax.org/books/biology-2e/pages/1-introduction>

4.

DNA PACKAGING IN EUKARYOTES

4.1 Introduction: “Inner Life of the Genome”

To review differences between prokaryotic cells and eukaryotic cells go here: [Khan Academy Video](#)

The size of the genome in one of the most well-studied prokaryotes, *E. coli*, is 4.6 million base pairs (approximately 1.1 mm, if cut and stretched out).

Let's do a rough calculation about how much DNA a **eukaryotic** cell has.

Assume

1. A somatic cell has 2 full copies of every chromosome. The total

number of base pairs for ALL chromosomes combined is 6×10^9 bp.

2. We know the length of each base-pair is 0.34nm (see DNA structure chapter!)

If you were to take the DNA of all the chromosomes, stretch it out, and lay them end to end then **what will the total length of ALL the DNA in the cell be?**

Hopefully, you got to a number of 2.0 meters! For reference that is approximately the height of Lebron James! Now consider the size of the eukaryotic nucleus which ranges from 2- 10 microns! For a good frame of reference, a SINGLE grain of salt is 100 microns which is still 10 times larger than the upper range for nucleus size!

This results in an engineering problem that cells need to solve.

1) How to does all of the DNA fit inside a bacterial cell and even more amazingly inside the nucleus of a eukaryotic cell?

At the same time consider all the different cells in the body. The human body contains approximately many different cell types, but each cell type shares the same genomic sequence. In spite of having the same genetic code, cells not only develop into distinct types from this same sequence but also maintain the same cell type over time and across divisions.

They are also specialized in their function- liver cells for example produce enzymes that help with detoxification but they do not produce antibodies, that is the job of a different cell type.

These cells have different expression patterns due to the

temporal and spatial regulation of genes. (See video below for a good analogy).

This leads to the second problem to contend with- once the genome has been folded and all packaged **up how can we open small sections of it so the machinery to read and express the information coded with the genes can gain access?**

The epigenome (“epi” means above in Greek, so epigenome means above genome) is the set of chemical modifications or marks that influence gene expression and are transferred across cell divisions and, in some limited cases, across generations of organisms.



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=498#oembed-1>

In this chapter, you will learn about the solution to the ‘Packaging Problem’ and how cells regulate **access to genes.**

The solution involves interactions of DNA with specific proteins, leading to the formation of a nucleoprotein complex called **CHROMATIN.**

We will focus exclusively on Eukaryotic chromatin,

although bacterial chromosomes (which are circular) also coil and supercoil with the help of proteins.

Learning Objectives

Levels 1 and 2 are factual information, knowledge-based. Level up indicated by the “Target” symbol is the goal.

When you have mastered the information in this chapter, **you should be able to:**

Level 1 and 2 (Knowledge and Comprehension)

- Draw/Label nucleosomes in a 10nm fiber to identify core histones, linker DNA, core DNA.
- Outline the steps by which a nucleosome is assembled.
- List the five major types of histone proteins, and describe what role each of them plays in the nucleosome
- Explain what form of chromatin is present during interphase.
- Define the roles of modifying amino acid side chains in altering nucleosome structure and

how the structure of chromatin can be used for regulating gene expression.

- What modifications would lead to activation of gene expression?

⊕ **Level Up (Application, Analysis, Synthesis)**

- Analyze/Predict/Interpret experimental data connected with DNA organization into chromatin.
- Analyze/Predict/Interpret experimental data connecting chromatin structure with gene-expression.
- Predict how mutations in key histones can alter chromatin structure /gene expression.
- Outline an experiment to purify histones from chromatin.

4.2 Types of Chromatin in Eukaryotic Cells

When we visualize chromosomes we usually think of the characteristic highly condensed structures (X shaped).

This is the form the DNA takes only during a brief part of

its life cycle (during Mitosis) where it is maximally condensed. These discrete packets are ideal for the job at hand which is to accurately separate and segregate to the opposite poles. However, the majority of the time a cell spends in **interphase**.

During this stage [which can be further subdivided into Gap 1-G1, S-phase, and G2] the cell is active. DNA, RNA, proteins are synthesized, sometimes on demand and depending on the needs of the cell.

During *interphase* (Figure 4.1 below) chromatin exists in various states of condensation called ***heterochromatin*** and ***euchromatin*** respectively.

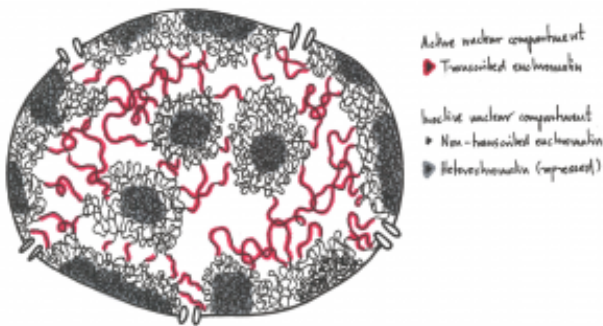


Figure 4.1 A hand-drawn sketch illustrating the spatial organization of different types of chromatin as typically seen in the nucleus of biological cells. Image credit: Lennart Hilbert, CC BY-SA 4.0, via Wikimedia Commons

The transition between these chromatin forms involves changes in the amounts and types of proteins bound to the

chromatin that can occur during gene regulation, i.e. when genes are turned on or off. Experiments to be described later showed that active genes tend to be in the more dispersed **euchromatin** where enzymes of replication and transcription have easier access to the DNA. Transcriptionally *inactive* genes are **heterochromatic**, obscured by additional chromatin proteins present in heterochromatin.

STUDY TIP: Pause here and watch Dr. Mehta's lecture **Video 1 from playlist: [L211 Dr. Mehta Chromatin Structure](#)**

4.3 Chromatin Organization

We can define three levels of chromatin organization in general terms:

1. DNA wrapped around histone proteins (*nucleosomes*) like “beads on a string”.
2. Multiple nucleosomes coiled (condensed) into 30 nm fiber (solenoid) structures.
3. Higher-order packing of the 30 nm fiber into the eventual familiar metaphase chromosome.

4.3.1 Histones

Before we can discuss how these aspects were determined we should introduce the proteins that play a central role in the folding of DNA- the histones.

There are 5 major classes of histones: H1, H2A, H2B, H3, and H4. Histones are basic proteins containing many *lysine* and *arginine* amino acids. Their positively charged side chains enable these amino acids to bind to the acidic, negatively charged *phosphodiester backbone* of double-helical DNA. About a gram of histones is associated with each gram of DNA.

Histones are among the most highly conserved proteins. Very few amino acid differences distinguish a human histone from histones in a mouse, sea urchin, or yeast cell. For instance, the H4 from cows differs from H4 in peas by only 2 amino acids. This eludes to the fundamental aspect of the role and function of these proteins.

Note that only eukaryotes (i.e., organisms with a nucleus and nuclear envelope) have histones. Prokaryotes, such as bacteria, do not.

4.3.2 First level of organization- The Nucleosome

Aspects of chromatin structure were determined by gentle

disruption of the nuclear envelope of nuclei, followed by salt extraction of extracted chromatin. Salt extraction dissociates most of the proteins from the chromatin. The results of a low [salt] extraction are shown in Fig.4.2 (below).

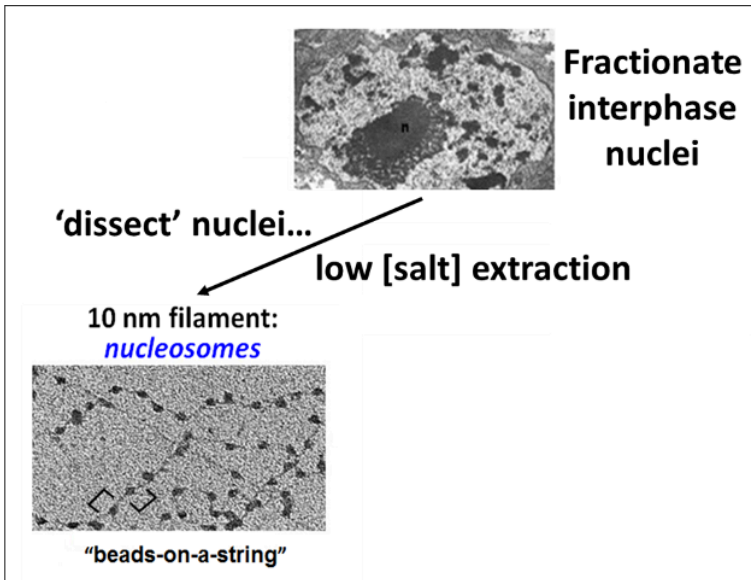


Figure 4. 2 Low salt fractionation of interphase nuclei yields 10nm nucleosome beads on a string.

When the low salt extract is centrifuged and the pellet resuspended, the remaining chromatin looks like *beads on a string*. DNA-wrapped *nucleosomes* are the beads, which are in turn linked by uniform lengths of the metaphorical DNA 'string'.

Roger Kornberg, one son of Nobel Laureate Arthur

Kornberg (discoverer of the first DNA polymerase enzyme of replication-see the next chapter), participated in the discovery and characterization of nucleosomes while still a post-doc!

The techniques he used are some of the ones we discussed in earlier chapters! Two types of experiments were used, one involved partial digestion of chromatin with micrococcal nuclease (an enzyme that degrades DNA).

When he carried out electrophoresis of DNA extracted from digests of nucleosome beads-on-a-string preparations, he found that each it generated DNA fragments about 200 base pairs of DNA and that *nucleosomes* are separated by a **“linker” DNA** stretch of about 80 base pairs.

In contrast, digestion of naked DNA (not associated with proteins) yielded a continuous smear of randomly sized fragments. These results suggested that the binding of proteins to DNA in chromatin protects regions of the DNA from nuclease digestion so that the enzyme can attack DNA only at sites separated by approximately 200 base pairs.

This led him to propose a model that the basic unit of chromatin- a Nucleosome (the beads) consists of ~ 200bp of DNA wound around a core, some space (linker DNA), and another nucleosome.

The length of the DNA extracted from the nucleosomes varies between organisms, ranging from 170-240 but the variation results entirely from the linker DNA length between the nucleosomes.

The length of DNA that is wrapped around the histone

proteins that make up the core is always about 147 bp base pairs long.

The nucleosome core particle consists of an octameric protein complex (two copies each of H2A, H2B, H3, and H4) with the 147 bp DNA wound around it.

The identity of the proteins that make up the **core of the nucleosome** came next and included familiar techniques of separating proteins and running them on SDS-PAGE gels.

Assembly of the Nucleosome

Today, researchers know that nucleosomes have a common structure comprising of: Two of the histones H2A, H2B, H3, and H4 that come together to form a histone octamer, which binds and wraps approximately 1.7 turns of DNA, or about 146 base pairs.

All core histones share a conserved structure known as the histone fold, which consists of three alpha-helices connected by 2 short loops.

The N-terminus of the proteins that do not participate in the fold are referred to as **Histone Tails** and play a crucial role in the regulation of chromatin structure.

The assembly of the nucleosome occurs in a step-wise process with specific associations:

2 subcomplexes are first formed. H3 interacting with H4 to form a heterodimer, H2A and H2B interacting to form H2A.H2B heterodimers. The interaction between histones utilized histone folds in what is termed the handshake fold.

H3.H4 heterodimers then interact to form a tetramer and the DNA begins to wrap around it, the H2A. H2B dimers then cap the complex at the top and bottom. (Figure 4. 3)

In the figure below the core histones are color-coded (yellow = H2A, red = H2B, blue = H3, green = H4) and cylinders represent helices.

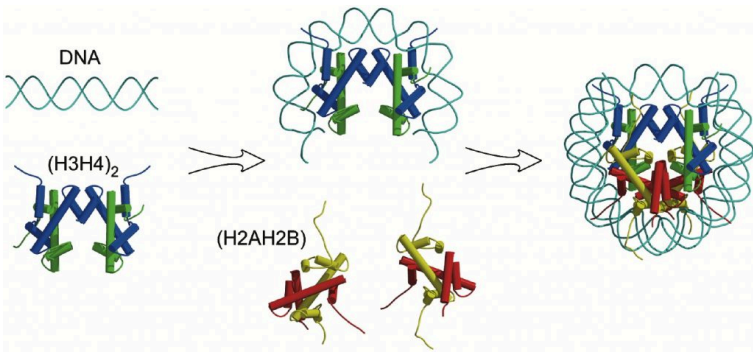


*Figure 4.3. Schematic Representation of Stepwise Assembly of the Nucleosome Core Particle, Based on Published NCP Structures (Luger, 2003). Image credit: Structure Volume 12 Issue 12 Pages 2098-2100 (December 2004)
DOI: 10.1016/j.str.2004.11.004*

STUDY TIP: Pause here and watch Dr. Mehta

VIDEO 2 within the playlist: [L211 Dr. Mehta Chromatin Structure](#)

Before you continue you should

1. Watch the Lecture videos that cover the material above.
(if you haven't already)
2. Complete the associated Lecture Quickcheck.

Histone-DNA interactions:

The three-dimensional structure of the nucleosome using X-ray crystallography was solved by Dr. Karolin Luger which provided a deeper understanding of chromatin organization.

1. Electrostatic interactions of negatively charged DNA (phosphodiester backbone) with positively charged amino acids of histones.
2. Additional interactions occur via hydrogen bonds with bases of the DNA.

This association is consistent with the need to package any type of DNA without regard to sequence.

However, there are areas of the genome where nucleosomes position themselves preferentially. These include regions with A-T rich sequences that are more accommodating of the ‘bend’ or ‘compression’ that occurs in the minor groove as the DNA wraps around the histone octamer.

4.3.2 30-nm Fibres and Higher-Order Chromatin

The packaging of DNA into nucleosomes shortens the fiber length about sevenfold, not enough to fit in the nucleus just yet. Therefore, chromatin is further coiled into an even shorter, thicker fiber, termed the “30-nanometer fiber,” because it is approximately 30 nanometers in diameter (Figure 4.4).

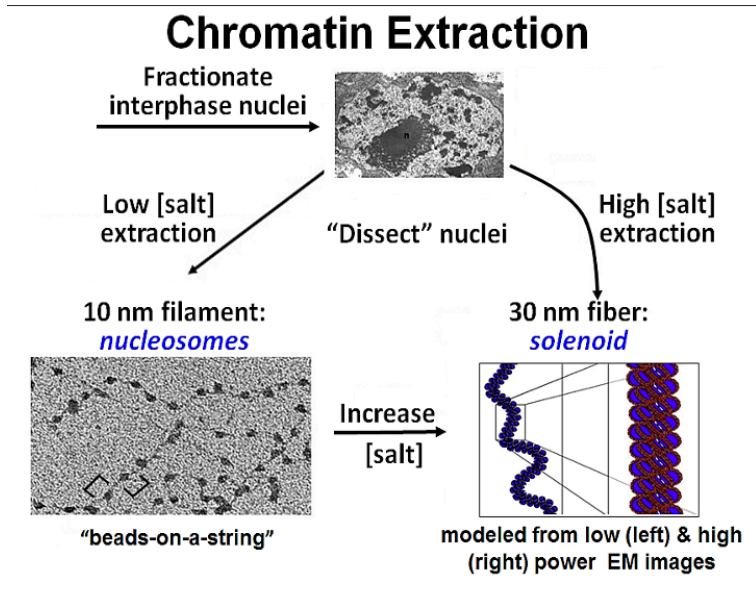


Figure 4.4 High salt chromatin extraction from nuclei, or high salt treatment of 10 nm filaments yields 30 nm solenoid structures, essentially coils of 10 nm filaments.

Experimentally this structure can be observed after a high salt chromatin extraction. As shown in the illustration, increasing the salt concentration of an already extracted nucleosome preparation will cause the ‘necklace’ to fold into the 30nm solenoid structure.

In recent years histone H1 also known as **‘Linker histone’** has been shown to play a role in establishing or stabilizing the 30nm structure.

In experiments when H1 is present the 30-nanometer fibers form readily. H1 binds DNA where the DNA joins and leaves

the histone octamer and helps lock the DNA into place, acting as a clamp around the nucleosome.

Additionally, the tails of the core histones have been shown to be important for the formation of 30nm fibers.

Non-Histone Proteins

In fact, there are at least five levels (*orders*) of chromatin structure (Fig. 4.5).

The first 2 (#1 and #2 in figure 4.5) we discussed above but other extraction protocols revealed other aspects of chromatin structure shown in #s 3 and 4 (in Figure 4.5). We know the most about the 10 nm fiber and have yet to fully elucidate how the chains of nucleosomes fold into the final condensed form.

Non-Histone proteins like **condensins** and

scaffold proteins and cohesins play a role to further coil and compact the DNA until it resembles the metaphase chromosome.

Remember that the final metaphase chromosome shape is established and made only in preparation for metaphase. The bulk of the time the cell is in interphase and chromosomes do

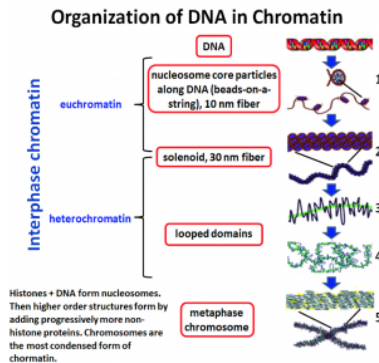


Figure 4.5 Five different levels (orders) of chromatin structure (see text for details). CC-BY-SA 3.0; Adapted From <https://en.wikipedia.org/wiki/Chromatin>

not resemble those discrete units we are used to thinking of when we think of chromosomes!

See animation below showing the folding process: [Animation: How DNA is packaged](#)



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://iu.pressbooks.pub/iul211smehta/?p=498#oembed-2>

STUDY TIP: Pause here and watch [BIOL-L211 Dr. Mehta Lecture Video: Higher-Order Chromatin Structure](#)

Concepts in Context: Shape of the Genome

Watch:



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=498#oembed-3>

In this video, you learn how new tools have allowed scientists to probe chromatin organization by finding points of contact between different regions.

COMPLETE: Don't forget to complete the assignments associated with Mol Bio in the News in CANVAS.

Key Takeaways

- The primary structure of chromatin is a 10 nm fiber “beads on a string” structure
- The ‘bead’ is a nucleosome, which includes about 200 bp of DNA wrapped around a histone core consisting of 2 copies each of core histones (H2A, H2B, H3, and H4)
- The DNA closest to the Histone core- Core DNA is 147 bp.
- The next level of chromatin is 30 nm fibers- formed by interactions between neighboring nucleosomes.
- Histone H1 is associated with linker DNA and assist in the formation of 30 nm fibers.
- 30nm fibers are folded into higher-order structures (loops upon a proteinaceous scaffold, and then loops are further condensed to form metaphase chromosomes)
- Non-Histone proteins are used for higher-order structures.

4.4 Regulation of Chromatin Structure

STUDY TIP: Watch the Lecture Video first. In this part of the chapter, I will highlight just some key points of how chromatin structure is regulated from the lecture videos.

Link here: [Dr. Mehta: Regulation of Chromatin Structure](#)

To recap we have just seen that chromosomal DNA associates with histones, forming an organized complex known as chromatin. This accomplished the goal of allowing DNA to fit into a smaller volume so within a eukaryotic cell's nucleus.

However, compaction of the DNA also limits the accessibility of the DNA to proteins (transcription factors- which we will learn about soon) that will transcribe the gene [Gene Expression].

In order for gene expression to occur, the chromatin structure needs to move back and forth between condensed and decondensed forms.

Changes to chromatin structure can be brought about in

2 ways and they are not mutually exclusive adding layers of complexity to regulation.

1. **Via Histone modifications**
2. **Via Chromatin remodeling complexes.**

4.4.1 Histone Modifications

The amino acid residues of histone tails are subject to many post-translational modifications and the list is ever-growing. In particular, histones contain a large percentage of **lysine and arginine (basic amino acids)** residues and serines and threonines.

The side chains of these residues are modified by the action of enzymes.

The most studied modifications include **acetylation (adding acetyl groups)** of lysines, **phosphorylation (adding phosphate groups)** of serines, **methylation (adding methyl groups)** of lysines and arginines.

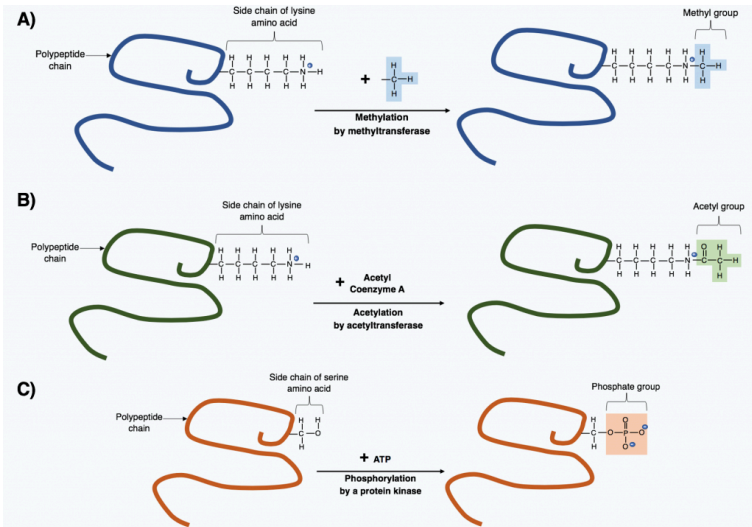


Figure 4.6 The post-translational modification of proteins by methylation, acetylation, and phosphorylation. Kep17, CC BY-SA 4.0 via Wikimedia Commons.

Connecting Concepts



An interactive HSP element has been excluded from this version of the text. You

can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=498#h5p-9>

The addition of the groups is facilitated by enzymes (collectively known as epigenetic “writers”) whilst de-modifying enzymes (or “erasers”) remove these marks. (see table below)

Histone Acetyl transferases (HATs)	Adds acetyl groups to histones
Histone Deacetylases (HDACS)	Removes acetyl groups from histones
Histone Methyltrasnderases	Adds methyl groups
Histone Demethylases	Remove methyl groups

These opposing activities enable a highly dynamic regulation of gene expression as modifications can be added or removed depending on whether a particular writer or eraser is recruited to a specific location of the genome.

Modifications can occur at different amino acids on

different histones and can create more than 100 unique potential changes in the histones.

Modifications not only regulate chromatin structure by merely being there, but they also recruit proteins that recognize particular modified amino acid residues.

Proteins with BROMODOMAINS- recognize Acetylated Lysine residues.

Proteins with CHROMODOMAINS- recognize methylated residues.

To add to the complexity, these proteins and modifying enzymes are contained within a larger multiprotein complex and can recruit chromatin remodeling complexes that further reposition nucleosomes (discussed below).

4.4.2 Chromatin Remodeling Complexes

- As the name suggests, these are protein complexes meaning they contain multiple different proteins working together.
- They ‘remodel’ chromatin- alter chromatin structure by repositioning nucleosomes. The mechanisms include
 - sliding nucleosomes along DNA
 - evicting nucleosome components (such as H2A–H2B dimers)
 - ejecting full nucleosomes (creating nucleosome-free regions)

- replacing with variant histone subunits.

There are many families of chromatin remodeling complexes with a variety of different names, and while they are diverse in protein composition they all have at least one enzymatic ‘ATPase subunit’ which allows them to utilize the energy released from ATP hydrolysis to reposition nucleosomes. (1)

Additionally, as mentioned above they are often recruited to the chromatin by ‘reading’ the ‘marks’ left by histone modifiers OR contain proteins with histone-modifying activities and histone recognition activities.

4.5 Links to Medicine

Given the intimate role chromatin structure plays in regulating gene expression, it is not surprising that changes in acetylation signaling resulting from misregulated HATs or HDACs can cause abnormal gene expression patterns and have been identified in numerous cancers. (2)

Some examples are

- Genetic alterations in HATs in hematological and solid cancers (3)
- Hyper-active HDACs in many cancers (4)
- Components of several chromatin-remodeling complexes are highly mutated in cancer.

Not surprisingly there are a wide variety of small-molecule inhibitors targeting acetylation signaling pathways in development for use as anti-cancer drugs (see table below).

Family	Target	Inhibitors	Phase	Tested cancer type	Effects on the DDR	References
HAT	CBP/p300	Curcumin	Clinical	multiple	PARP1, CPT and HJ sensitivity; (-) BRCA1 mRNA; (-) ATR activity; (-) Ku70/Ku80 recruitment	[172]
	p300	C646	Preclinical	melanoma	Cisplatin sensitivity; (-) DNA repair genes mRNA i.e., Rad51; (-) γH2AX	[176]
				lung	IR sensitivity; (-) IR-p-CHK1	[199]
HDAC	Class I/II	Vorinostat (SAHA)	FDA approved in 2006	multiple	IR sensitivity; (-) HR & NHEJ proteins, i.e., Rad50, Ku70; (+) γH2AX IRIF; (-) IR-induced HR & NHEJ proteins, i.e., Rad51, Ku80	[165–167,200,201]
	Class I	Romidepsin (FK228)	FDA approved in 2009	ovarian	(+) γH2AX/Rad51/53BP1 foci with cisplatin c.t., and in s.c.	[202]
				lung	IR sensitivity; (+) IR-γH2AX	[203]
				thyroid	(+)γH2AX/ROS; (-) Ku70/80 and Rad51; (+) γH2AX in s.c.	[168]
				renal cell	(+) γH2AX/ROS with 5-FU c.t. and in s.c.	[204]
	Class I/II	Panobinostat (LBH589)	FDA approved in 2015	lung, bladder leukemia	IR sensitivity; (+) IR-γH2AX; (-) Mre11/Nbs1/ Rad51 protein (+) γH2AX; (-) DNA repair protein/signaling, i.e., Chk1/p-Chk1 with TOP1II c.t.	[205,206] [207,208]

Modified table from Article: Acetylation Reader Proteins: Linking Acetylation Signaling to Genome Maintenance and Cancer. Gong F, Chiu LY, Miller KM (2016) Acetylation Reader Proteins: Linking Acetylation Signaling to Genome Maintenance and Cancer. PLOS Genetics 12(9): e1006272. <https://doi.org/10.1371/journal.pgen.1006272>. Works published by PLOS are licensed under the Creative Commons Attribution (CC-BY) license.

Remember to:

1. Watch the Lecture videos that cover the material above. This will help to clarify or reinforce certain concepts if they were unclear.
2. Complete any associated Lecture Quickchecks.

3. Start work on Problem Set

References and Attributions

This chapter contains material taken from the following CC-licensed content and Public Domain content. Changes include rewording, removing paragraphs and replacing with original material, and combining material.

Images

Images in this chapter unless otherwise noted are from the textbook Basic Cell and Molecular Biology: What We Know & How We Found Out – 4e by Gerald Bergtrom licensed CC-BY. Available here: <https://open.umn.edu/opentextbooks/textbooks/cell-and-molecular-biology-2e-what-we-know-how-we-found-out>.

Attributions provided within the original text for the diagrams are included below

Fig. 4.2: Low salt fractionation of interphase nuclei yields 10nm nucleosome *beads on a string*.

- Upper; From Bergtrom et al., (1977) J. Ultrastr. Res. 60:395-406: Research by G. Bergtrom;
- Lower left; CC-BY-SA 3.0; Adapted from: <https://commons.wikimedia.org/wiki/>

[File:Chromatin_nucleofilaments_%28detail%29.png](#)

Fig. 4.3: High salt chromatin extraction from nuclei or high salt treatment of 10 nm filaments yields 30 nm *solenoid* structures, essentially coils of 10 nm filaments.

- Electron micrograph of nucleus, From Bergtrom et al., (1977) J. Ultrastr. Res. 60:395-406: Research by G. Bergtrom
- CC-BY-SA 3.0; Adapted from:
https://commons.wikimedia.org/wiki/File:Chromatin_nucleofilaments_%28detail%29.png
- CC BY-SA 4.0; Alt: Adapted from Richard Wheeler
<https://en.wikipedia.org/w/index.php?curid=53563761>

Fig. 4.5: Five different levels (orders) of chromatin structure. CC-BY-SA 3.0; Adapted From <https://en.wikipedia.org/wiki/Chromatin>.

References

- (1) Clapier, C., Iwasa, J., Cairns, B. *et al.* Mechanisms of action and regulation of ATP-dependent chromatin-remodeling complexes. *Nat Rev Mol Cell Biol* **18**, 407–422 (2017). <https://doi.org/10.1038/nrm.2017.26>
- (2) Gong F, Chiu LY, Miller KM (2016) Acetylation Reader Proteins: Linking Acetylation Signaling to Genome Maintenance and Cancer. *PLOS Genetics* 12(9): e1006272. <https://doi.org/10.1371/journal.pgen.1006272>

(3) Di Cerbo V, Schneider R. Cancers with wrong HATs: the impact of acetylation. *Briefings in functional genomics*. 2013;12(3):231–43. pmid:23325510

(4) Gui CY, Ngo L, Xu WS, Richon VM, Marks PA. Histone deacetylase (HDAC) inhibitor activation of p21WAF1 involves changes in promoter-associated proteins, including HDAC1. *Proceedings of the National Academy of Sciences of the United States of America*. 2004;101(5):1241–6. pmid:14734806

5.

DNA REPLICATION

5.1 Introduction

A fundamental property of living organisms is their ability to reproduce. Bacteria and fungi can divide to produce daughter cells that are identical to the parental cells. Sexually reproducing organisms produce offspring that are similar to themselves.

On a cellular level, this reproduction occurs by mitosis, the process by which a single parental cell divides to produce two identical daughter cells.

Meiosis is the process in which cells with a diploid genome produce four germ cells (haploid cells). Regardless of process Mitosis or Meiosis, an essential step is the duplication of all of the genetic material such that each daughter cell receives a full complement of genetic material.

In this chapter, we look at the details of replication as well as differences in detail between prokaryotic and eukaryotic replication that arise because of differences in DNA packing.

We will begin with a discussion on the general features of replication common to the replication of ‘naked’ prokaryotic

DNA and of chromatin-encased eukaryotic DNA, arisen early in the evolution of replication biochemistry.

You will read about experiments that explored or revealed answers to some basic questions like:

1. Where does replication begin? Is it one region on the chromosome or several?
2. Is the start random or at a specific location?
3. Once replication begins in which direction does replication occur?

Answers to many of these questions arose from experiments carried out in *E. coli*, which has a circular genome.

We will then introduce all the proteins involved in carrying out this essential process to develop a complete picture of how DNA replication actually occurs, ending with differences between prokaryotic and eukaryotic replication that arise because of differences in DNA packing.

Learning Objectives

Levels 1 and 2 are factual information, knowledge-

based. Level up indicated by the “Target” symbol is the goal.

When you have mastered the information in this chapter, **you should be able to:**

Level 1 and 2 (Knowledge and Comprehension)

- Be able to explain why DNA replication is semiconservative
- Be able to explain the data obtained by Meselson-Stahl to prove that replication is semi-conservative.
- Outline the general feature of replication that are common amongst prokaryotes and eukaryotes
- Outline the general feature of replication that are common amongst prokaryotes and eukaryotes
- Accurately label within a replication fork (a) the polarity of the newly synthesized strands (5'-3'), (b) leading and lagging strands.
- Identify individual proteins and their roles in DNA replication.
- Why do eukaryotes need telomeres, but prokaryotes do not?
- What are the ways in which bacterial and

eukaryotic replication differ?

- What are the ways in which bacterial and eukaryotic replication are similar?
- Explain how eukaryotic cells distribute their histones during replication and what impact it has on chromatin state inheritance.

⊕ Level Up (Application, Analysis, Synthesis)

- What are the requirements for in vitro synthesis of DNA under the direction of DNA polymerase I?
- Apply the concepts of directionality of DNA synthesis to **new modes of replication**.
- Draw/Interpret modes of replication when given fictitious Meselsohn and Stahl data.
- Predict the effect on DNA replication with mutations in components of DNA replication machinery.
 - Predict which component of DNA replication machinery is mutated, when given descriptions of phenotypes of mutants.

5.2 The Basic Rules of Replication

DNA Replication is essential for life and as a process, it must be

1. Accurate (make few mistakes)
2. Fast
3. Complete (although we will see one exception when it comes to the ends of linear chromosomes.)

It takes an army of proteins and some specialized DNA sequences that function together to accomplish this!

While the individual components are important, it is important to not lose sight of some general principles of replication that are common to all of life.

5.2.1 DNA Replication is always semiconservative: Meselson and Stahl Experiments

Watson and Crick's double helix model suggested a replication mechanism. In presenting the complementary pairing of bases in the double helix, Watson and Crick immediately realized that the base sequence of one strand of DNA can be used as a template to make a new complementary strand.

This was the semi-conservative model of replication. However, how DNA can separate (denature) was not understood and there were alternative models of replication (Figure 5.1)

Conservative Model: Perhaps there is some DNA synthesizing machine in the cell that can take dsDNA and make a copy of it ?

Dispersive Model: Perhaps the process of replication could break the parental DNA into pieces and use them to seed the synthesis of new DNA?

It was far from obvious what the mechanism would be. However, these three models make different predictions about the *behavior* of the two strands of the parental DNA during replication.

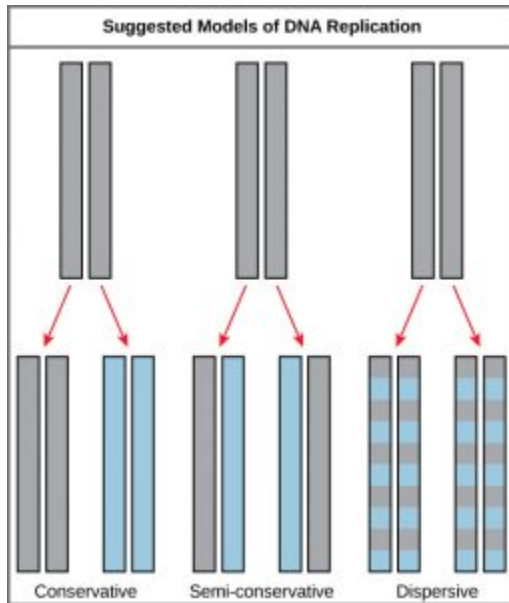


Figure 5. 1 Three suggested models of DNA replication. Grey indicates parental DNA, blue indicates newly synthesized DNA. Image CNX OpenStax, CC BY 4.0, via Wikimedia Commons.

Meselson and Stahl's set out to test the three possible models of replication and confirmed that DNA replication is in fact semi-conservative. *(Due to the simplicity and indisputable results of the experiment, the Meselson-Stahl experiment has been called "the most beautiful experiment in biology." see Link to Learning)*

In their experiment, *E. coli* cells were grown in a medium containing ^{15}N , a 'heavy' nitrogen isotope. After many

generations, all of the DNA in the cells had become labeled with the heavy isotope. At that point, the ^{15}N -tagged cells were placed back in a medium containing the more common, 'light' ^{14}N isotope and allowed to grow for exactly one generation.

Fig. 5.2 (below) shows Meselson and Stahl's predictions for their experiment. Meselson and Stahl knew that ^{14}N -labeled and ^{15}N -labeled DNA would **form separate** bands after centrifugation on *CsCl chloride density gradients*.

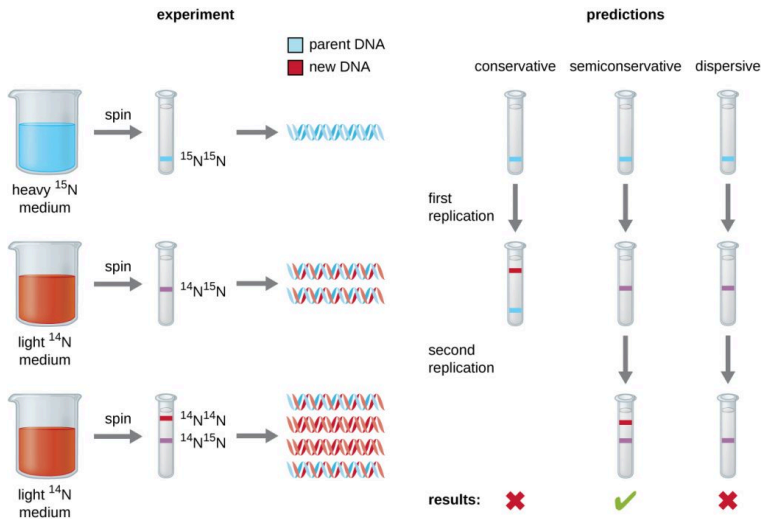


Figure 5. 2. Meselson and Stahl experimented with *E. coli* grown first in heavy nitrogen (^{15}N) then in ^{14}N . DNA grown in ^{15}N (blue band) was heavier than DNA grown in ^{14}N (red band) and sedimented to a lower level on ultracentrifugation. After one round of replication, the DNA sedimented halfway between the ^{15}N and ^{14}N levels (purple band), ruling out the conservative model of replication. After the second round of replication, the dispersive model of replication was ruled out. These data supported the semiconservative replication model. Image credit: CNX OpenStax, CC BY 4.0, via Wikimedia Commons.

They tested their predictions by purifying and centrifuging the DNA from the ^{15}N -labeled cells grown in ^{14}N medium for one generation.

They found that this DNA formed a single band with a density between that of ^{15}N -labeled DNA and ^{14}N -labeled

DNA, eliminating the *conservative model* of DNA replication (possibility #1).

That left two possibilities: replication was either semiconservative (possibility #2) or dispersive (possibility #3).

The dispersive model was *eliminated* when DNA isolated from cells grown for a 2nd generation on ^{14}N were shown to contain two bands of DNA on the CsCl density gradients.

Study Tip: Pause here and watch Dr. Mehta

Lecture Video 1 [DNA Replication- Meselson and Stahl Expt](#)

Links to Learning: Hear from Meselson and Stahl themselves!

[iBiology Video: The most beautiful experiment in science](#)

Talk Overview

Matt Meselson and Frank Stahl were in their mid-20s when they performed what is now recognized as one of the most beautiful experiments in modern biology. In this short film, Matt and Frank share how they devised the groundbreaking experiment that proved semiconservative DNA replication, what it was like to see the results for the first time, and how it felt to be at the forefront of molecular biology research in the 1950s. This film celebrates a lifelong friendship, a shared love of science, and the serendipity that can lead to foundational discoveries about the living world.

5.2.2 DNA Replication Begins at Specific Chromosomal Sites

Where does replication begin? Does the DNA unwind at one region on the genome or multiple places? Is it random or are there special sequences on the DNA where the process of unwinding begins?

The answers to some of the fundamental questions came from the direct visualization of bacterial DNA using a technique called **autoradiography**.

In fact, this technique showed that the genome of E-coli is in fact circular!

In 1963, John Cairns cultured *E. coli* cells for long periods on ^3H -thymidine ($^3\text{H-T}$) to make all of their cellular DNA radioactive. He then extracted the DNA (during replication) and allowed it to adhere to membranes.

A sensitive film was placed over the membrane and time was allowed for the radiation to expose the film, which was later developed ***to generate the autoradiographs***.

This produced another famous image in biology (pictured in Figure 5.3), the dark lines (tracks of silver grains in the autoradiographs) that revealed the pattern of replicating DNA molecules.

Cairns called these replicating chromosomes ***theta images*** because they resembled the Greek letter theta (θ).

From his many autoradiographs, he arranged a sequence of his images to illustrate his inference that replication starts at

a single **origin of replication** on the bacterial chromosome, proceeding around the circle to completion.

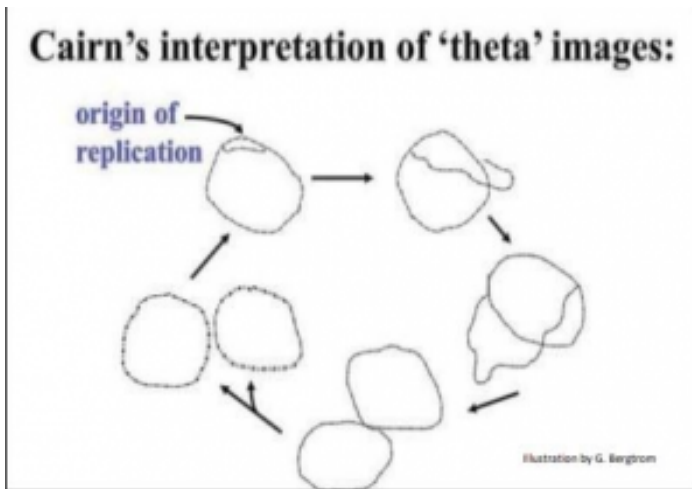


Figure 5.3 An ordering of Cairns' autoradiograph images to suggest the progress of replication of the E. coli circular chromosome. Illustration by G. Bergtrom. From: Bergtrom, G. (2020) Cell and Molecular Biology: What We Know & How We Found Out [CMB4e] (http://dc.uwm.edu/biosci_facbooks_bergtrom/)

Terminology: DNA replication involves the formation of a **replication bubble** (and prokaryotic replication involves a **single origin**

of replication). The term 'replication fork' also comes from the image above. Note the Y-shaped junctions, similar to the classic fork in the road. DNA at these forks is being separated and replicated.

5.2.3 Most DNA Replication Is Bidirectional

The theta images can be explained by both **uni-directional** replication- where a new DNA is being synthesized in one direction (DNA continues to unwind at ONE replication fork) and moves around the circular DNA until complete.

It can also be explained by **bi-directional** replication, two growing points originating (DNA continues to unwind and new DNA synthesized at BOTH replication forks) until they meet at the opposite end.

David Prescott demonstrated ***bidirectional replication***.

- Cells were labeled with ^3H -thymidine with low specific activity to **lightly label the replication bubble**.
- Then the cells were labeled with a much **stronger radioactive isotope** for a short time.
- Any newly synthesized DNA would be labeled with the

stronger label and appear darker.

- Visualization of chromosome revealed darker segments on BOTH ends of the replication bubble.

Conclusion: Replication indeed begins at an origin of replication, but that double helix then unwinds in *opposite directions*, replicating DNA *both* ways away from the origin from two ***replication forks***.

STUDY TIP: Pause here and Lecture Video: [L211](#)

[DNA Replication: General Features](#)

5.2.4 DNA Polymerase Catalyzes Phosphodiester-Linkage Formation

Before we consider what happens at replication forks in detail, let's focus our attention on **DNA POLYMERASES**- a class of enzymes that catalyze the step-wise addition of nucleotides to a DNA strand.

DNA Polymerase enzymes have some unique properties:

1. DNA polymerase is a **template-directed enzyme** that

synthesizes a product with a base sequence complementary to that of the template.

BUT

1. All DNA polymerases discovered to date can only ***elongate a preexisting DNA or RNA strand***, they cannot **initiate chains**!

How would DNA synthesis begin then?

Solution: RNA polymerases do not require a preexisting base-paired 3' end to initiate synthesis.

In the cell, an RNA polymerase enzyme, called **Primase** assembles a short stretch of RNA – the **PRIMER** base-paired to the parental DNA template. Given the similarity between RNA and DNA, DNA polymerases can extend the free 3' OH group provided by RNA!

The PRIMER-TEMPLATE junction is the cellular substrate recognized by DNA polymerase.

Primase is a slow (relatively!) and error-prone polymerase. The error-prone nature of its activity is not a problem, because the cell will remove all the RNA primers later in the process of DNA synthesis and replace them with DNA nucleotides- and we'll revisit this later in this reading.

Recall: The nucleotides used in DNA synthesis are deoxyribonucleoside triphosphates or dNTPs. As can be inferred from their name, such nucleotides have a deoxyribose sugar and three phosphates, in addition to one of the four DNA bases, A, T, C or G.

Chemistry of DNA Synthesis

Synthesis of DNA involves adding nucleotides, one by one, in the exact order specified by the original (template) strand.

DNA polymerase catalyzes the reaction by which an incoming deoxyribonucleotide, complementary to the template, is added onto the 3' end of the previous nucleotide. The importance of the 3'OH group lies in the nature of the reaction that builds a chain of nucleotides.

The reaction catalyzed by the DNA polymerase occurs through the nucleophilic attack by the 3'OH group of the last nucleotide of growing strand on the **α phosphate** of the **incoming dNTP** (Figure 5.4). The 5' phosphate of the new nucleotide binds to the 3' OH group of the nucleotide to make a phosphodiester bond.

The immediate hydrolysis of the pyrophosphate that is cleaved off the incoming dNTP drives the reaction forward.

Each added nucleotide provides a new 3'OH, allowing the chain to be extended for as long as the DNA polymerase continues to synthesize the new strand.

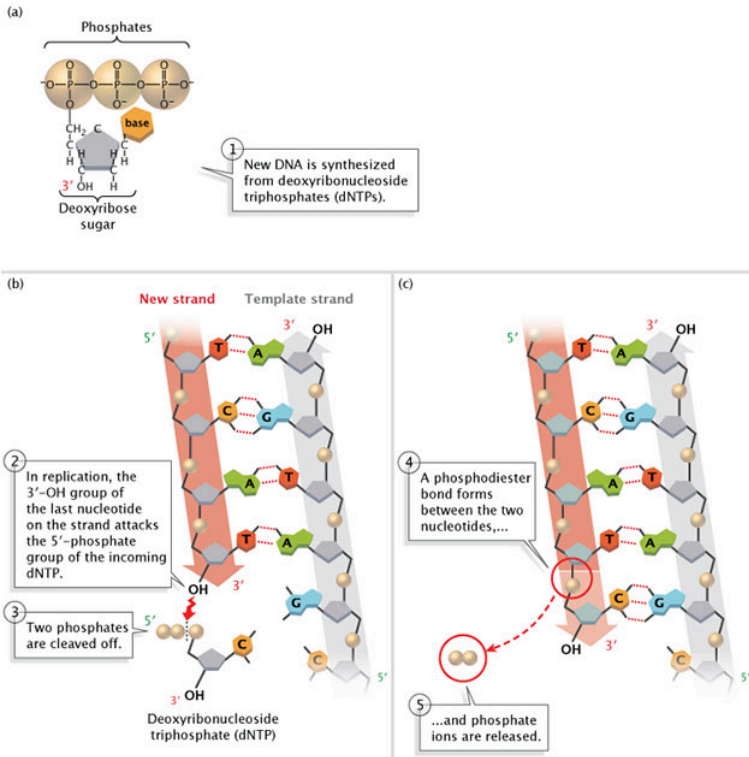


Figure 5.4. (A) A deoxyribonucleoside triphosphate (dNTP). (B) During DNA replication, the 3'-OH group of the last nucleotide on the new strand attacks the 5'-phosphate group of the incoming dNTP. Two phosphates are cleaved off. (C) A phosphodiester bond forms between the two nucleotides and phosphate ions are released. © 2014 Nature Education Adapted from Pierce, Benjamin. *Genetics: A Conceptual Approach*, 2nd ed. All rights reserved. Image can be viewed here: <https://www.nature.com/scitable/topicpage/major-molecular-events-of-dna-replication-413/>

Thus DNA polymerases can only extend a strand in the

5' to 3' direction. Scientists have yet to identify a polymerase that can add bases to the 5' ends of DNA strands.

Now consider one replication fork. Here the DNA strands of separated, the DNA ahead of the fork remains wound in a double helix and the separated strands serve as templates for new daughter strand synthesis.

Remember this fork is moving- the downstream DNA is continuously being unwound and upstream DNA is continuously being replicated into **hybrid helices of old and new DNA strands. But this DNA has to be anti-parallel to the parental DNA.**

Time to Think



An interactive H5P element has been excluded from this version of the text.

You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=576#h5p-10>

So if no 3' to 5' synthesizing activity can be found, how

is the new strand oriented 3' to 5' in the direction of the replication fork synthesize **without violating one of the two fundamental rules of nucleic acid chemistry?**

5.2.5 DNA synthesis at the replication fork semi-discontinuous

This conundrum is solved, by copying DNA **continuously** on one template strand and **discontinuously** on the other.

One template strand is oriented such that the 5'-to-3' direction in which the daughter strand synthesis is taking place is in the direction of fork movement; this strand can be synthesized in a CONTINUOUS manner and is known as **LEADING STRAND synthesis.**

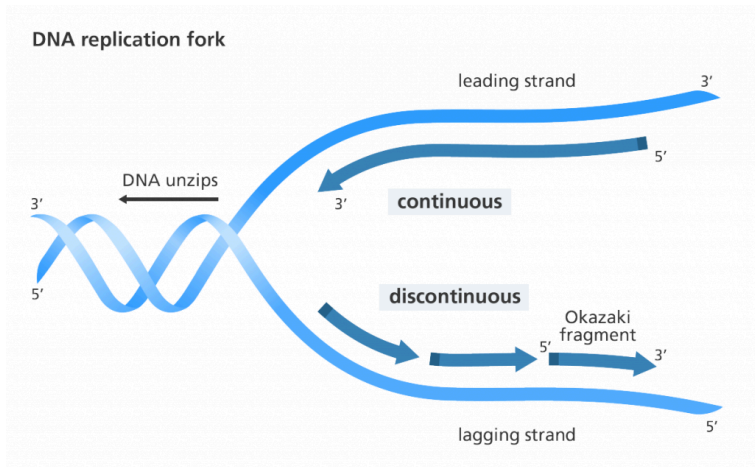


Figure 5.5 An illustration to show replication of the leading and lagging strands of DNA. Image credit: Genome Research Limited

DNA copied from the other template strand, on the other hand, is made in short segments **away from the fork or discontinuously**.

This is known as **LAGGING STRAND synthesis**. Lagging strand synthesis takes place in a 5'-to-3' direction and no rules of nucleic chemistry are violated.

Eventually, these short segments of DNA will have to be stitched together by an enzyme called **DNA LIGASE**.

The short stretches of DNA generated during lagging strand synthesis are commonly known as Okazaki fragments after their discoverer Reiji Okazaki. In bacteria, Okazaki fragments are 1,000-2,000 nucleotides in length,

whereas in the cells of higher organisms they are typically only 100-200 nucleotides long.

See here for an animation of this process: [Animation of Bidirectional Replication \(YouTube\)](#)

Did I Get This?



An interactive HSP element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=576#h5p-11>



An interactive HSP element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=576#h5p-12>

Key Takeaways

- Replication is initiated at origin, creating two replication forks that move bidirectionally (in opposite directions)
- All DNA polymerases require a 3'-OH end to initiate DNA synthesis.
- The DNA polymerase advances continuously when it synthesizes the leading strand, but synthesizes the lagging strand in short fragments (Okazaki fragments)

Study Tip. Pause here and watch the Lecture Video: [Chemistry of DNA Synthesis](#)

Prof Mehta's Study Note: An Important Distinction:

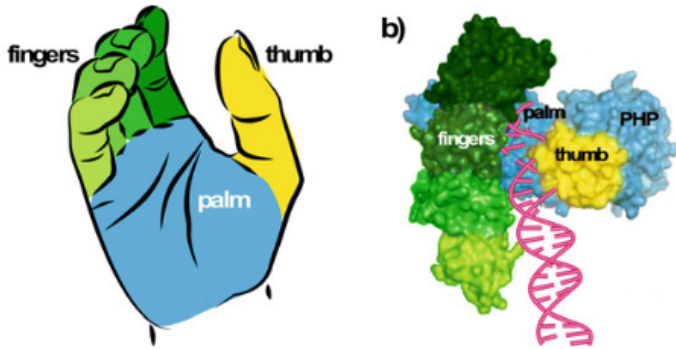
Bidirectional replication means that there are **2 replications forks moving in opposite**

directions. It does NOT refer to the fact that the 2 DNA strands (leading and lagging strands) are made in opposite directions. That is called semi-discontinuous synthesis. At every fork DNA replication is semi-discontinuous.

5.3 Structure of DNA Polymerase

The three-dimensional structures of a number of DNA polymerase enzymes are known. The shape has been compared to that of a hand (or baseball mitt) with the protein domains that are referred to as the fingers, the thumb, and the palm.

Each subdomain carries a specific function. The fingers domain binds the incoming dNTP, the thumb domain helps grip duplex DNA, and the palm domain contains the amino acids that are part of the active site of the enzyme. The catalytic palm domain is the most conserved of the domains,



Crystal structure analysis of various domains (b) shows that they all have the shape of a right hand, with subdomains referred to as palm, fingers, and thumb. (a). Image from the following Nina Y. Yao & Mike E. O'Donnell (2016) Evolution of replication machines, *Critical Reviews in Biochemistry and Molecular Biology*, 51:3, 135-149, DOI: 10.3109/10409238.2015.1125845

5.3.1 Accuracy of Replication

How accurate is the copying of the information by DNA polymerase?

As you are aware, changes in DNA sequence (mutations) can change the amino acid sequence of the encoded proteins, and that this is often, though not always, deleterious to the functioning of the organism.

When billions of bases in DNA are copied during replication, how do cells ensure that the newly synthesized

DNA is a faithful copy of the original information? DNA synthesis is extremely high fidelity making only one error per 10^9 bases!

How is this remarkable accuracy achieved?

3 Mechanisms assure the accuracy of replication

1. **The change in the structure of the enzyme when it binds the correct nucleotide.** The DNA polymerase clasps the ds DNA tightly before binding the new dNTP. Within the active site of a DNA polymerase, the incoming dNTP is tested. A perfect fit triggers a conformational change: the finger domain rotates to form a tight pocket into which only a properly shaped base pair will readily fit.

Recall that the complementary base pairing (A with T and G with C) results in just the correct width for DNA!

The proper base is favored by the formation of a base pair, which is stabilized by specific hydrogen bonds. The binding of a non-complementary base is less likely because the interactions are energetically weaker.

2. 3'-5' Exonucleolytic Proofreading: DNA polymerases are their own editors! They can correct mistakes in DNA by removing mismatched nucleotides. This happens with a separate 3'-5' EXONUCLEASE activity that allows it to snip out the incorrect base and replace it with the correct base and resume replicating the template strand.

Some polymerases like **DNA polymerase I** can also remove RNA primers in the 5' to 3' direction, though that is not a common activity of polymerases. You will encounter this shortly.

3. Strand Directed Mismatch Repair: A special mechanism inside cells to correct mismatches that already got incorporated.

We will learn about mismatch repair in the DNA Repair unit.

STUDY TIP Pause here and watch Lecture Video:
[L211 Structure of DNA Polymerase](#)

Animation: <https://youtu.be/6hck-4S68U>



*One or more interactive elements has
been excluded from this version of the*

text. You can view them online here:

[https://iu.pressbooks.pub/
iu/211smehta/?p=576#oembed-1](https://iu.pressbooks.pub/iu/211smehta/?p=576#oembed-1)

Before you continue you should

Complete the associated Lecture Quickcheck on CANVAS

5.4 Process of Replication

Our understanding of the process of DNA replication is derived from studies using bacteria, yeast, and other systems. These investigations have revealed that DNA replication is carried out by the action of a large number of proteins that

act together **as a complex protein machine**. (Table 5.1) This complex machine is known as **REPLISOME**.

Although the specific proteins involved are different in bacteria and eukaryotes, the basic mechanisms and principles are relevant **in all cells**.

E. coli Gene	Enzyme/Protein Function	Description
<i>dnaA</i>	Initiator Protein	Melts DNA at <i>oriC</i> , exposing two template ssDNA strands
<i>dnaB</i>	Helicase	Unwinds the DNA helix at the front end of each replication fork during replication
<i>dnaC</i>	Helicase Loader	Loads the DnaB Helicase onto the ssDNA template strands
<i>dnaG</i>	Primase	Synthesizes RNA primers used to initiate DNA synthesis
<i>dnaE</i>	α -Catalytic Subunit of DNA Polymerase III	Catalytic subunit of the main replicative polymerase during DNA replication
<i>dnaQ</i>	ϵ -Editing Subunit of DNA Polymerase III	Editing subunit of the main replicative polymerase during DNA replication
<i>dnaN</i>	β -clamp subunit of DNA Polymerase III	Clamping subunit of the main replicative polymerase during DNA replication
<i>polA</i>	DNA Polymerase I	Processes Okazaki fragments and also fills in gaps during DNA repair processes
<i>polB</i>	DNA Polymerase II	Proofreading and editing, especially on lagging strand synthesis and some involvement in DNA repair
<i>ssb</i>	Single Stranded Binding Proteins (SSB)	Bind with single-stranded regions of DNA in the replication fork and prevent the strands from rejoining
A dimer encoded by <i>gyrA</i> and <i>gyrB</i>	DNA Gyrase	Type II Topoisomerase involved in relieving positive supercoiling tension caused by the action of Helicase
A dimer encoded by <i>parC</i> and <i>parE</i>	Topoisomerase IV	Type II Topoisomerase involved in decatenation of daughter chromosomes during DNA replication
<i>ligA</i>	DNA Ligase	Fixes nicks in the DNA backbone during DNA replication, DNA damage, and DNA repair processes
Note: Only the genes involved in the formation of the catalytic domain of DNA polymerase III are listed		

Table 5.1 Enzymes Involved in DNA Replication in the prokaryote, *E. coli*. From: Flatt, P.M. (2019) Biochemistry – Defining Life at the Molecular Level. Published by Western Oregon University, Monmouth, OR (CC BY-NC-SA). Available at: https://wou.edu/chemistry/courses/online-chemistry-textbooks/ch450-and-ch451-biochemistry-defining-life-at-the-molecular-level/?preview_id=4919&preview_nonce=cca8f0ce36&preview=true”

We will primarily describe the process as discovered in Prokaryotes and point out differences with Eukaryotes towards the end.

The identification and roles of many of the proteins in replication were elucidated with genetic experiments, using bacterial mutants. Table 5.1 lists the gene names that code for the proteins! **You are not expected to know the gene names!** I will only refer to the protein name (or sometimes *the class of enzymes* that the protein belongs to). **It is the FUNCTION of the protein that is most relevant.**

5.4.1 Stages of Replication

As with many processes in molecular biology, we conventionally look at replication as being made up of three phases – initiation, elongation, and termination.

Stage	What happens	Goal
INITIATION: Getting Started	DNA first unwound at the origins of replication.	Make space!! Open up the ds DNA to allow the machinery to load!! Assembly of the replication apparatus on the origin of replication.
ELONGATION: Making DNA	DNA is made! Leading strand is synthesized continuously Lagging strand is synthesized discontinuously.	Make DNA, avoid errors, be efficient (fast!)
TERMINATION: Stop	2 replication forks meet about halfway around the bacterial chromosome.	Avoid re-replication. Ensure completion of replication. Disengage apparatus and proteins.

5.4.1 Stage 1: Initiation of Replication

As we have seen, DNA synthesis starts at one or **more origins or replication**.

How does the replication machinery *know* where to begin?
What are these origins?

Origins are **specific nucleotide sequences** where replication begins. In *E. coli* (as do most prokaryotes), there is a single origin of replication (ori-C) on its one chromosome. This region is depicted in Figure 5.6

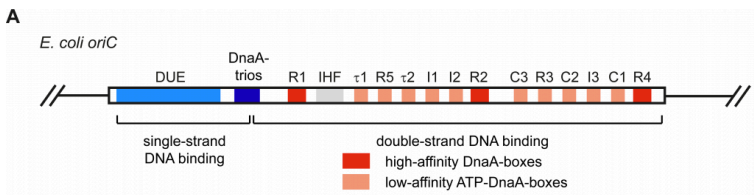


Figure 5.6. Schematic of the architecture of *E. coli* origin oriC. Franziska Bleichert, CC BY 4.0 via Wikimedia Commons. Original from Ekundayo B, Bleichert F (2019) Origins of DNA replication. *PLOS Genetics* 15(9): e1008320. <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008320>

The region includes:

- Several conserved sequences that are binding sites for a **protein called DnaA [DnaA boxes]**
- An **AT-rich** region also called the **DNA Unwinding Element**
- GATC Methylation sites. (not shown in Figure above)

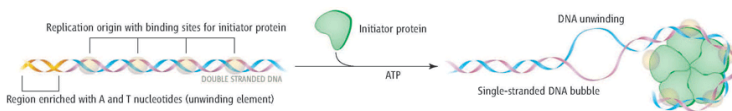
You may remember, A-T base pairs, which have two hydrogen

bonds between them are more readily disrupted than G-C base pairs which have three hydrogen bonds and therefore can be easily 'melted' or 'open up'.

Initiation of replication begins when **INITIATOR PROTEINS** (DnaA for *E. coli*) will bind to the origins of replication.

For *E. coli*

- DnaA proteins bind the DnaA- Boxes (orange above)- these are the recognition sequences for the binding of the DnaA initiator protein.
- The binding of these proteins promotes the recruitment of multiple DnaA subunits that form a **helical oligomer**
- Bending of the DNA and torsional stress promotes the melting of the adjacent region (AT)



*The *E. coli* origin is illustrated as an example. Initiator binding and hydrolysis of ATP results in oligomerization of initiator proteins and the formation of a single-stranded DNA bubble, into which the DNA replication machinery will assemble.*

Image credit: Roxana E. Georgescu, and Mike O'Donnell
Science 2007;317:1181-1182. <https://science.sciencemag.org/content/317/5842/1181>

Unwinding

Once a small region of the DNA is opened up at each origin of replication, the DNA helix must be unwound to allow replication to proceed.

How are the strands of the parental DNA double helix separated?

The unwinding of the DNA helix requires the action of a *class of enzymes* called **HELICASES**.

Helicases are motor proteins that move (at speeds up to that of a jet engine!) directionally along a nucleic acid phosphodiester backbone, separating two annealed nucleic acid strands such as DNA and RNA, using energy from ATP hydrolysis.

Helicase loads onto the lagging strand and is a ring-shaped protein.

Note that a replication bubble is made up of two replication forks that “move” or open up, in opposite directions. At each replication fork, the parental DNA strands must be unwound to expose new sections of the single-stranded template.

[In E.coli the loading of the helicase is actually a two-step process but let's not complicate things!]

Two problems arise as a result of unwinding DNA

1. The separated single strands of DNA must be kept from coming back together so the new strands can continue to be synthesized.

Solution: To ensure that unwound regions of the parental DNA remain single-stranded and available for copying, the separated strands of the parental DNA are quickly coated by many molecules of a protein called **single-strand DNA binding protein. (SSB)**

2. A knotty problem: What is the effect of unwinding one region of the double helix? The local unwinding of the double helix causes over-winding (increased positive supercoiling) ahead of the unwound region. The DNA ahead of the replication fork has to rotate, or it will get twisted on itself and halt replication. This is a major problem, not only for circular bacterial chromosomes but also for linear eukaryotic chromosomes, which, in principle, could rotate to relieve the stress caused by the increased supercoiling.

Solution: Topoisomerases

Topoisomerases are special enzymes that can relieve the topological stress caused by local “unwinding” of the extra winds of the double helix. They do this by cutting one or both strands of the DNA and allowing the strands to swivel around each other to release the tension before rejoining the ends.

In *E. coli*, the topoisomerase that performs this function is called **GYRASE**.

DNA gyrase moves ahead of the replication fork and is essential for the process of replication.

Watch this animation to better understand the concept of supercoiling



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://iu.pressbooks.pub/iu121smehta/?p=576#oembed-2>

Loading Primase

Okay, so we have opened up DNA. What's next?

Remember DNA polymerases **cannot** begin synthesis of a new DNA strand de novo and require a free 3' OH to which they can add DNA nucleotides.

The helicase recruits ***primase*** which synthesizes RNA primers used to initiate DNA synthesis.

Did I get this?



An interactive HSP element has been



excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=576#h5p-17>

NOTE: A lot of textbooks and diagrams show the primase and helicase as separate. In reality, the 2 enzymes interact with one another. The N-terminal domain (NTD) the helicase interacts with the carboxy-terminal domain (CTD) of the primase and forms a functional ***primosome***. Within the primosome, the helicase acts to unwind the double-stranded helix and the primase synthesizes RNA primers on both the leading and lagging DNA strands.

Study Tip. Pause here and watch Dr. Mehta
Lecture Video: [L211 Initiation of DNA](#)
[Replication](#)

Key Takeaways

- Origin of replication in E.coli (*oriC*) contains conserved sequences called DnaA boxes and AT-rich regions.
- Initiator proteins (DnaA) bind to DnaA boxes, forming a complex that leads to the melting of DNA.
- DNA helicase forms at the replication fork.
- SSB – single-stranded binding proteins and DNA gyrase are also needed at the origin,
- Primase is bound to helicase and when activated synthesizes the RNA primer.

5.4.2 DNA Polymerases

E. coli has a total of five **DNA polymerases**. Three of these enzymes are involved in DNA replication (DNA polymerases I, II, and III). **DNA polymerase III is the main polymerase involved in both leading strand biosynthesis and the synthesis of the Okazaki Fragments during DNA replication.**

The DNA polymerase III consists of 10 different proteins organized into three **functionally distinct**, but **physically interconnected** assemblies and referred to as **DNA Pol-III Holoenzyme**. (Figure 5.7)

A cartoon of the arrangement is shown on the right in the diagram below.

There are copies [the 3 glove-like structures (in black) in the cartoon] of DNA Pol-III catalytic core or enzyme. This is the protein responsible for synthesizing DNA. These are connected via flexible linker proteins to another sub-structure (green and red) that consists of several proteins that together make up the **Sliding Clamp loader** and **Sliding clamp**.

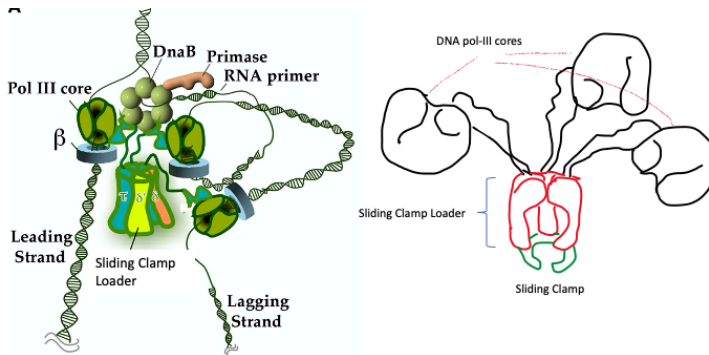


Figure 5.7. On left is a diagram showing the DNA pol-III holoenzyme. Three Pol-III cores are shown. Two of the Pol III cores are proposed to function on the lagging strand. The τ subunits (blue) of the clamp loader are shown with flexible linker proteins that connect the clamp loading device to the helicase and pol-III core. A cartoon representation of just the DNA Pol III holoenzyme is on right. Image credit: Left image is from [Peter McInerney, Aaron Johnson, Francine Katz, Mike O'Donnell \(2007\) Molecular Cell Volume 27, Issue 4 Pages 527-538.](#)

On right is a cartoon diagram of the DNA Pol III holoenzyme-illustration by S. Mehta.

Role of Sliding Clamp and Clamp Loader

Consider that the chromosome of *E. coli* consists of almost five million base pairs. Because DNA replication in *E. coli* takes place simultaneously from two replication forks, the overall rate of DNA synthesis is 1,600 nucleotides per second. Thus, *E. coli* is capable of duplicating its entire chromosome in as little as 40 minutes, and as we will see below, it does so with considerable accuracy.

This property of remaining attached to the DNA through many rounds of nucleotide addition is referred to as **processivity**.

Processivity maximizes the speed of DNA synthesis.

If the polymerase frequently fell off the DNA and had to re-bind in order to resume nucleotide incorporation, the rate of DNA synthesis would be much slower.

This is where the sliding clamp and clamp loader comes in.

The shape of the clamp – a ring-shaped structure provides a geometrically elegant and simple solution.

DNA polymerase is tethered to the DNA by a **sliding clamp**, which (being ring-shaped) fully encircles the DNA helix.

Sliding clamps cannot load onto DNA spontaneously because they are closed circles. The job of the **clamp loader** is to ‘load’ the clamp onto the DNA.

Adenosine triphosphate (ATP)–dependent **clamp loaders** open the sliding clamps and load them onto primer-template junctions. Hydrolysis of ATP to ADP then changes the affinity of the clamp loader for the

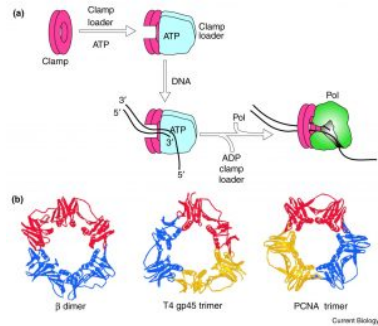


*Figure 5.8. Top: Clamp-loading reaction. The clamp loader has a low affinity for both clamp and primer-template DNA in the absence of ATP. Upon binding ATP, the clamp loader can bind the clamp and open it. The binding of primer-template DNA activates ATP hydrolysis, leading to the ejection of the clamp loader. Bottom: Sliding clamps have similar architecture. From left to right are structures of *E.coli*, Bacteriophage, and Eukaryotic sliding clamps. Figure from: Current Biology Volume 11 Issue 22 Pages R935-R946 (November 2001). Open Access DOI: [https://doi.org/10.1016/S0960-9822\(01\)00559-0](https://doi.org/10.1016/S0960-9822(01)00559-0)*

clamp and leaves it behind. The clamp spontaneously closes encircling the DNA helix. (Figure 5.8).

The sliding clamp is then joined by the DNA Polymerase. In the presence of the sliding clamp, DNA polymerases are much more processive, making replication faster and more efficient.

Note that the sliding clamp is loaded at the **Primer-Template junction**. Once the sliding clamp and DNA pol III cores are engaged the process of elongation begins!

Before you continue you should list/order the events that have taken place in Initiation thus far.

Let's Review



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=576#h5p-16>

5.4.3 Stage 2: Elongation of Replication

During elongation, the core polymerase (glove part of DNA pol III) adds DNA nucleotides to the 3' end of the newly synthesized polynucleotide strand. The template strand specifies which of the four DNA nucleotides (A, T, C, or G) is added at each position along the new chain. Only the nucleotide complementary to the template nucleotide at that position is added to the new strand.

1. The leading and lagging strand polymerases are both tethered to the clamp loader. Recall, however, that the lagging strand (chemistry) necessitates synthesizing it in a direction opposite to the growing replication fork.

How can DNA be **simultaneously** synthesized on both the leading and lagging stands **if the polymerases are bound to each other?**

2. Note that there are 3 core polymerases. **What is the purpose of the third polymerase?**

In an updated model for replication known as the **TROMBONE model of replication** (see animation below). The lagging template strand is looped out so that it passes through the polymerase site in one subunit of polymerase III in the 3' to 5' direction.

After adding about 1000 nucleotides, DNA polymerase III lets go of the lagging-strand template by releasing the sliding clamp.

This mode of replication has been termed the trombone model *because the size of the loop lengthens and shortens like the slide on a trombone!*

This looping allows the primase and the Pol III active site can accomplish the discontinuous synthesis of the lagging template strand even though the general direction of the Pol III complex is opposite to the required direction of DNA synthesis.

Thus the lagging strand requires **2- Core polymerases** (in E.coli) to allow for replication to be completed in a timely fashion. One core-polymerase engages with the looped Okazaki fragment. It begins extension and the DNA is released as helicase proceeds forward.

As the core polymerase continues to extend the Okazaki fragment a newly primed section is formed at the replication fork, which is then captured by the third polymerase.

DNA Replication Animation below highlights the whole process:



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=576#oembed-3>

Dr. Mehta Lecture Video: [Trombone Model of Replication](#)

Completing Elongation

As each newly formed segment of the lagging strand approaches the 5' end of the adjacent Okazaki fragment (the one just completed).

How are these RNA/DNA fragments converted into one long continuous DNA strand?

The RNA could be removed by a polymerase that has 5'→3' exonuclease activity, however, ***Pol III lacks this activity.***

Unlike polymerase III, DNA polymerase I has a 5' → 3' exonuclease activity.

In E.coli,

- RNAse H first removes all the ribonucleotides in the primer except for the rNTP linked to the DNA end.
- **DNA Pol I use nick-translation-** utilizing its polymerizing ability to extend the 3' end of the newly formed Okazaki fragment and the 5'-3' exonuclease activity to remove the remaining ribonucleotide from the neighboring Okazaki fragment.

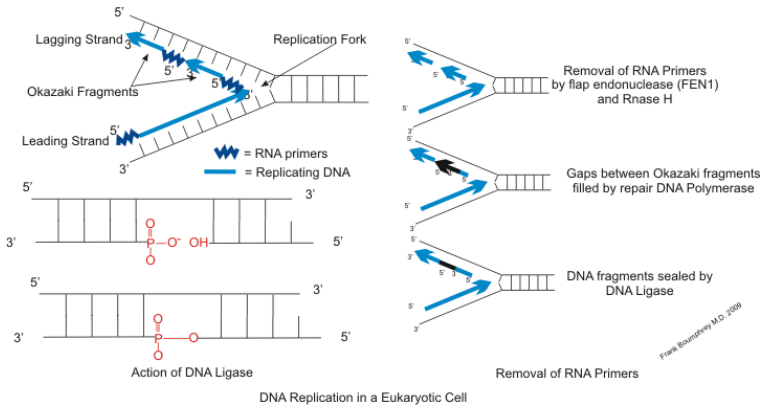


Figure 5.9. Processing of Okazaki Fragments. RNA primers are removed by RNAase H, the ends of Okazaki fragments (light blue) are extended by DNA Pol- δ . DNA ligase seals the gap. Image attribution: Boumphreyfr, CC BY-SA 3.0 <<https://creativecommons.org/licenses/by-sa/3.0/>>, via Wikimedia Commons.

Finally, another critical enzyme, **DNA ligase**, joins adjacent completed fragments, making the final phospho-diester bond in the ‘nicks’ left behind between Okazaki fragments. (Figure 5.9)

5.4.4 Stage 3: Termination of Replication

Proper termination of DNA replication is important for genome stability. *E. coli* replication terminates in the region

opposite *oriC*.

The two replication forks meet at a termination region, setting off a series of events that leads to the eventual completion of replication and subsequent chromosome separation.

The termination region contains special conserved sequences called **(Ter) sites**. These sequences are recognized by **Tus proteins**.

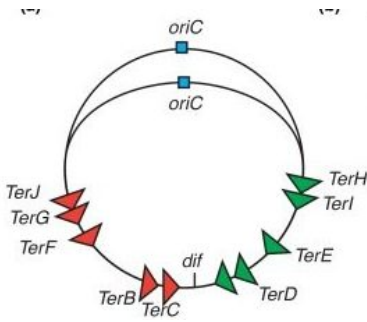


Figure 5.10. Termination of replication. Ten terminator sites flank the terminus region. Clockwise replication is arrested when the replication fork meets Tus bound Ter C,B,F,G J site.

Counterclockwise replication by the other set of terminators. Image from: Xu, Z-Q. and Dixon, N.E. (2018) Curr Op Struct Biol 53:159-168

The 10 Ter sites are split and organized as two ***oppositely orientated groups of five***. The association of the Tus proteins to the Ter sites is also polar, such that there is a permissive face that allows the replisome to pass unhindered and a non-permissive face that can block the replisome (Figure 5.9).

This configuration allows the replisome to pass the first Tus-Ter complex unhindered and be blocked at the second. When the 2 replisomes meet

the apparatus disassembles.

We now have 2 circular chromosomes that are linked to

one another. Here the activity of a different enzyme from the Topoisomerase family becomes useful in separating or un-linking the 2 chromosomes.

Dr. Mehta Lecture Video: [Replication Elongation and Termination](#)

Practice: Terminology Crossword



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=576#h5p-19>

Before you continue you should

1. Watch the Lecture videos that cover the material above (if you haven't)
 2. **Complete the associated Lecture Quickcheck on CANVAS**
-

5.5 Eukaryotic Replication

Replication in eukaryotes is mechanistically similar to replication in bacteria but is more challenging in the following ways.

1) Sheer size: *E. coli* must replicate 4.6 million base pairs, whereas a human diploid cell must replicate 6 billion base pairs.

2) The second challenge is the fact that while the genetic information for *E. coli* is contained on one chromosome, human beings have 23 pairs of chromosomes that must be replicated.

3) While the *E. coli* chromosome is circular, human chromosomes are linear. The third challenge arises because of the nature of DNA synthesis on the lagging strand. Linear chromosomes are subject to shortening with each round of replication unless countermeasures are taken.

4) The events of replication have an additional twist in eukaryotes. Recall that DNA is found in eukaryotic cells as chromatin, a complex of the DNA with proteins. The nucleosome structure must be disrupted to make DNA available for replication and restored after replication is completed.

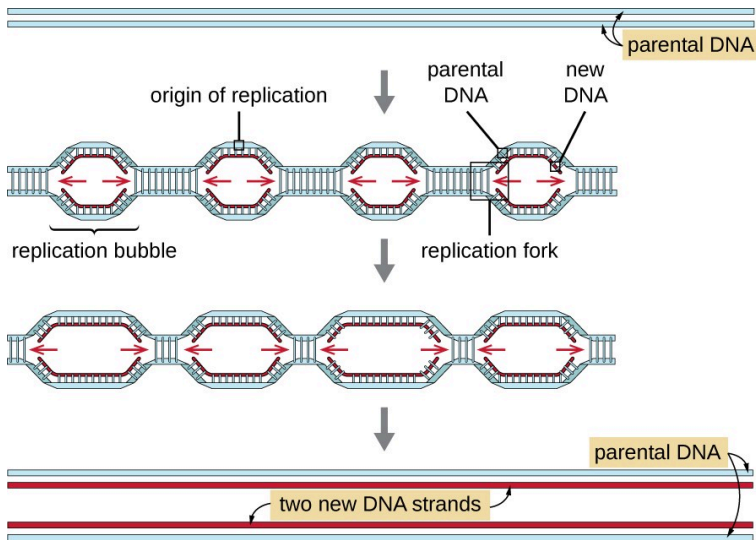


Figure 5.11 Eukaryotic chromosomes are typically linear, and each contains multiple origins of replication. Image credit: From Nina Parker, Mark Schneegurt, Anh-Hue Thi Tu, Philip Lister, Brian M. Forster. Microbiology (Open Stax) Access for free at: <https://openstax.org/books/microbiology/pages/11-2-dna-replication>

The solution to Challenge 1 and 2 :

Origin organization, specification, and activation in

eukaryotes are more complex than in bacterial or archaeal kingdoms and significantly deviate from the paradigm established for prokaryotic replication initiation.

To account for replicating large genome sizes – eukaryotic chromosomes have **multiple origins of replication**, which are located between 30 and 300 kilobase pairs (kbp) apart.

In human beings, replication of the entire genome requires about 30,000 origins of replication, with each chromosome containing several hundred.

Each origin of replication represents a replication unit or replicon, and at each origin, replication is semi-discontinuous and bi-directional.

The essential steps of replication are the same as in prokaryotes:

Before replication can start, the DNA has to be made available as a template.

- A helicase using the energy from ATP hydrolysis opens up the DNA helix.

- Replication forks are formed at each replication origin as the DNA unwinds.

- The opening of the double helix causes over-winding, or supercoiling, in the DNA ahead of the replication fork. These are resolved with the action of topoisomerases.

- Primers are formed.

The differences often are in the number of proteins needed for these functions.

For example:

1. The protein needed to bind the sequences that define the origins in eukaryotic genomes is often part of a large complex of proteins- Origin Recognition Complex.

2. The number of DNA polymerases in eukaryotes is much more than prokaryotes: 14 are known, of which five are known to have major roles during replication and have been well studied. They are known as pol α , pol β , pol γ , pol δ , and pol ϵ .

At the eukaryotic replication fork, three distinct replicative polymerase complexes contribute to canonical DNA replication: α , δ , and ϵ .

The ‘Primase’ for eukaryotes is DNA polymerase α (Pol α) – This protein has 2 parts- one which synthesizes DNA (not unlike all the DNA polymerases) the other has RNA priming activity.

Thus this complex accomplishes the priming task by synthesizing a primer that contains a **short ~10-nucleotide RNA stretch followed by 10 to 20 DNA bases.**

Importantly, this priming action occurs at origins to begin leading-strand synthesis and also at the 5’ end of each Okazaki fragment on the lagging strand.

However Pol α is not able to continue DNA replication- it is not processive!

3. After priming, synthesis is “handed off” to 2 other polymerases to continue elongation. The **polymerase switching** requires clamp loaders. [The function of which is exactly the same as in prokaryotes). The leading strand is ‘handed off’ to the enzyme **pol δ** , the lagging strand to **pol ϵ**

5.5.1 Dealing with Histones

As seen in Chapter 4 chromosomes are packaged by wrapping ~147 nucleotides (at intervals averaging 200 nucleotides) around an octamer of histone proteins, forming the ***nucleosome***. The histone octamer includes two copies each of histone H2A, H2B, H3, and H4.

It was highlighted that histone proteins are subject to a variety of post-translational modifications, including phosphorylation, acetylation, methylation, and ubiquitination that represent vital epigenetic marks.

The nucleosome structure must be disrupted to make DNA available for replication **and** restored after replication is completed.

Furthermore, it is important to transmit the epigenetic information found on the parental nucleosomes to the daughter nucleosomes, in order to preserve the same chromatin state.

In other words, ***the same histone modifications should be present on the daughter nucleosomes as were on the parental nucleosomes***. This must all be done while doubling the amount of chromatin, which requires the incorporation of **newly synthesized histone proteins**.

This process is accomplished by ***histone chaperones*** and ***chromatin remodeling complexes***.

Ahead of the replication fork, the chromatin structure is **disassembled** by ATP-dependent chromatin remodeling

complexes, allowing access to the DNA template. The loss of the histone octamer from the parental DNA during DNA replication is accompanied by the dissociation of H3/H4 tetramers and H2A/H2B dimers.

Special histone protein chaperones recruit histone H3-H4 dimers to the replication fork- helping load both newly synthesized (light purple) histones to establish chromatin behind the fork. Previously loaded histones (dark purple) are also deposited on both daughter DNA strands.

Thus in the reassembled chromatin, half of the histones are recycled from the parental chromatin while the other half are newly-synthesized.

(Figure 5.12)

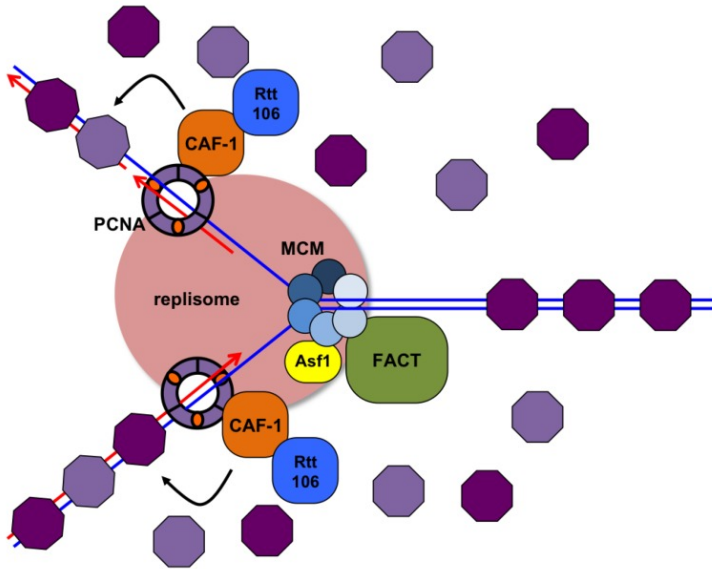


Figure 5. 12 Nucleosome displacement and deposition during DNA replication. Figure from: [Lemanm A.R. and Noguchi, E. \(2013\) Genes 4\(1\):1-32](#)

The presence of some recycled histones allows for one possible mechanism of how the epigenetic state is re-established. Presumably, the old histones would still be carrying the modifications (acetyl or methyl groups for example) which can then facilitate the spread of modifications using Bromo and Chromodomain-containing complexes that have histone-modifying activity.

For example, old histone bearing an acetyl group is recognized by a bromodomain-containing Histone acetyl Transferase, which will then acetylate adjacent histones.

Scientists are still uncovering how this important aspect of ‘somatic’ cell identity is maintained.

5.5.1 End Replication Problem-Replicating Telomeres

As you’ve learned, the enzyme DNA polymerase can add nucleotides only in the 5’ to 3’ direction. Leading strand synthesis continues until the end of the chromosome is reached. On the lagging strand, DNA is synthesized in short stretches, each of which is initiated by a separate primer.

When the replication fork reaches the end of the linear chromosome, there is no way to replace the primer on the 5’ end of the lagging strand leaving a structure called a 3’-overhang or a 5’-gap (**Fig 5.13**).

Interestingly, this overhang is critically important in forming a ‘cap-like structure at the end of the chromosome called the **telomere**.

The telomere has a highly repetitive noncoding sequence that is recognized by proteins that bind the DNA sequence and ‘hide’ the linear end of the chromosome by forming a loop. This structure is important because it is telling the cell “Hey, I’m the normal end of the chromosome-leave me alone; don’t try and fix me!”

Because the cell needs this 3’-overhang to create the correct end structure, the leading strand is resected (or chewed up) by an exonuclease to create a 3’-overhang.

After each round of replication, the leading strands are continually shortened (the lagging strands also gradually shorten due to the removal of the last RNA primer, but this shortening is actually less detrimental compared to the degradation on the leading strands). If the shortening at the ends of the chromosomes was not fixed, eventually the highly repetitive sequence of the telomere would be lost and the cap structure at the ends of the chromosomes would no longer form, causing the cell to signal that the DNA was damaged.

To combat the loss of DNA sequence in the telomere and prevent a DNA damage response, an enzyme called telomerase (Figure 5.13), adds nucleotides to the ends of chromosomes.

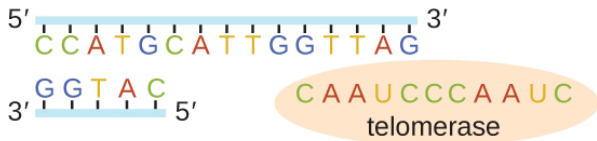
Telomerase is highly expressed in early development and gametic cells; it is moderately expressed in many adult stem cells, but not expressed in most adult somatic cells.

Telomerase is an enzyme that contains a catalytic part and a built-in RNA template. Because it uses an RNA template to build DNA (transcription in reverse), it is called **reverse transcriptase**.

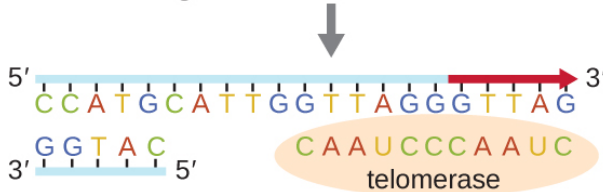
Telomerase attaches to the 3'-overhang (location of a 3'OH group) of the chromosome and adds DNA nucleotides that are **complementary to the RNA template**.

Once the 3' end of the template is sufficiently elongated, DNA polymerase can add a new primer and fill in the nucleotides complementary to the ends of the chromosomes (Figure 5.13).

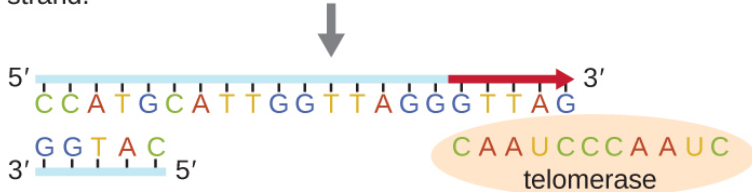
The final RNA primer will be removed, re-establishing the 3'-overhang and allowing the cap structure to be formed once again.



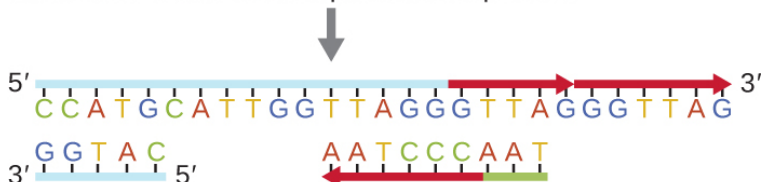
Telomerase has an associated RNA that complements the 3' overhang at the end of the chromosome.



The RNA template is used to synthesize the complementary strand.



Telomerase shifts, and the process is repeated.



Primase and DNA polymerase synthesize the complementary strand.

Figure 5.13. Telomerase. The ends of linear chromosomes are maintained by the action of telomerase. Image from Nina Parker et al Microbiology from Open Stax, Access for free at: <https://openstax.org/books/microbiology/pages/11-2-dna-replication>

In many cancer cell types, the expression of the telomerase

activity has been reactivated, allowing cancer cells to lengthen telomeres and become immortal.

Because of this action, there is great interest in better understanding how telomerase and the telomere are regulated in cells and if there are ways to exogenously control the enzyme.

Dr. Mehta Lecture Video: [L211 Eukaryotic Replication Playlist](#)

Link to Learning

This [You-Tube Video](#) compares the process of prokaryotic and eukaryotic replication

Concepts in Context: Telomeres and Ageing.

Cells that undergo cell division continue to have their telomeres shortened because most somatic cells do not make telomerase. This essentially means that telomere shortening is associated with aging. With the advent of modern medicine, preventative health care, and healthier lifestyles, the human life span has increased, and there is an increasing demand for people to look younger and have a better quality of life as they grow older.

For her discovery of telomerase and its action, Elizabeth Blackburn (1948–) received the Nobel Prize for Medicine or Physiology in 2009.

WATCH: TED Talk from Elizabeth Blackburn, where she highlights the discovery and data on the relationship between **stress and telomerase**.



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=576#oembed-4>

COMPLETE: Don't forget to complete the assignments associated with Mol Bio in the News in CANVAS.

Check your understanding





An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://iu.pressbooks.pub/iul211smehta/?p=576#h5p-18>

References and Attributions

This chapter contains material taken from the following CC-licensed content. Changes include rewording, removing paragraphs and replacing with original material, and combining material from the sources.

1. Bergtrom, Gerald, “Cell and Molecular Biology 4e: What We Know and How We Found Out” (2020). *Cell and Molecular Biology 4e: What We Know and How We Found Out – All Versions*. 13.
https://dc.uwm.edu/biosci_facbooks_bergtrom/13
2. Works contributed to LibreTexts by Kevin Ahern and Indira Rajagopal. LibreTexts content is licensed by CC BY-NC-SA 3.0

3. Fundamentals of Cell Biology by Shoshana D. Katzman, et al., Georgia Gwinnet College, licensed by Attribution-NonCommercial-ShareAlike 4.0 (available <https://alg.manifoldapp.org/projects/fundamentals-of-cell-biology>)

Figure image attributions are provided within the text.

D. Samson, L. H. John F. Cairns (1922–2018). *Nat Struct Mol Biol* **26**, 149–150 (2019). <https://doi.org/10.1038/s41594-019-0194-1>. See here: <https://rdcu.be/crdJs>

6.

TRANSCRIPTION IN PROKARYOTES

6.1 Introduction

In the preceding sections, we have discussed the replication of the cell's DNA and the mechanisms by which the integrity of the genetic information is carefully maintained.

What do cells do with this information? How does the sequence in DNA control what happens in a cell?

If DNA is a giant instruction book containing all of the cell's "knowledge" that is copied and passed down from generation to generation, what are the instructions for? And how do cells use these instructions to make what they need?

Genes must be expressed

In earlier chapters and from previous classes you have learned that all living organisms have genes, these genes carry information in the form of a code (temporary instructions or mRNA) that is used to make proteins. The language of

the genome is universal. This is what allows bacteria to make human insulin!

This description of the flow of information from DNA to RNA to protein is often called the central dogma of molecular biology and is a good starting point for an examination of how cells use the information in DNA.

Recall that proteins are polymers, or chains, of many amino acid building blocks. The sequence of bases in a gene (that is, its sequence of A, T, C, G nucleotides) translates to an amino acid sequence.

A **triplet** is a section of three DNA bases in a row that codes for a specific amino acid.

Similar to the way in which the three-letter code *d-o-g* signals the image of a dog, the three-letter DNA base code signals the use of particular amino acid.

For example, the DNA triplet CAC (cytosine, adenine, and cytosine) specifies the amino acid valine. Therefore, a gene, which is composed of multiple triplets in a unique sequence, provides the code to build an entire protein, with multiple amino acids in the proper sequence (**Figure 6.1**).

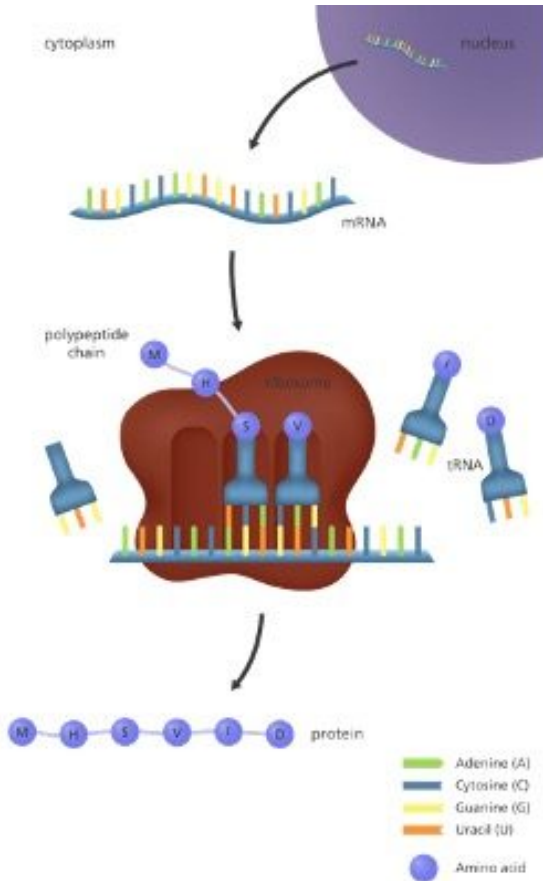


Figure 6.1. An illustration showing the process of translation. Image credit: Genome Research Limited.
<https://www.yourgenome.org/facts/what-is-gene-expression>

Because proteins are coded by genes, the term “*gene expression*” refers to protein synthesis (i.e., making proteins), including the *regulation* of that synthesis.

Gene Expression Is Regulated

Consider that all of the cells in a multicellular organism have arisen by division from a single fertilized egg and therefore, all have the same DNA. Division of that original fertilized egg produces, in the case of humans, over a trillion cells, by the time a baby is produced from that egg (that's a lot of DNA replication!).

Yet, we also know that a baby is not a giant ball of a trillion identical cells, but has the many different kinds of cells that make up tissues like skin and muscle and bone and nerves.

How did cells that have identical DNA turn out so different? The answer lies in the **REGULATION of GENE EXPRESSION** which is the process by which the information in DNA is used. Although all the cells in a baby have the same DNA, each different cell type uses a different subset of the genes in that DNA to direct the synthesis of a distinctive set of RNAs and proteins.

We began to see glimpses of this idea when we discussed chromatin regulation.

In this section and through the rest of the sections we will unravel all the ways in which our GENOME is actually put to work.

We begin by examining RNA synthesis in the simple prokaryote- *E. Coli*. Like many of the discussions prior, studies in this organism gave insight into the basic biochemical

processes that hold true in Eukaryotes as well – just with added complexity.

Learning Objectives

Level 1 and Level 2 (Knowledge and Comprehension)

- **Explain** what is meant by functional RNA
- Draw, describe or identify key features within a transcription unit.
- Explain the role of Sigma factor in bacterial transcription.
- Describe the three steps of transcription initiation that occur before the elongation phase begins.
- Explain the molecular mechanism behind transition from closed-open complex during initiation—(isomerization)
- Understand the meaning of promoter consensus sequences.

- Understand the different associations between RNA polymerase and DNA during transcription stages.
- Explain how termination of bacterial transcription occurs
- Distinguish between the 2 types of terminators sequences
- Explain the differences and similarities between RNA polymerase and DNA polymerase.

⊕ Level Up (Application, Analysis, Synthesis)

- When given an illustration that shows: a portion of a gene undergoing transcription, the template and coding strands labeled, and a DNA sequence you should be able to-
 - Indicate the direction in which RNA polymerase moves as it transcribes this gene.
 - Write the polarity and sequence of the RNA transcript from the DNA sequence given.
 - Identify the location of the promoter for

the gene.

- Identify elements of a gene, when given a gene sequence from a database.
- **Predict how mutations in promoter sequences, genes coding for sigma factors would affect transcription.**

6.2 Basics of Transcription

Transcription is the first step of the “central dogma” of information transfer from DNA to protein in which genetic information in genes is transcribed into RNA.

You can think of mRNA as being a copy of one book/ page within the book with the information needed for a particular assignment. While your genome is the entire library/master copy of the entire book that is in the restricted section of the library.

You can't take that book out of the restricted section, so you need to make a copy of the information.



Are all the RNA's in a cell messenger RNA (mRNA) molecules?

No! While we focus mainly on mRNA or 'protein coding' RNA when we think about transcription, less than 5% of total RNA in a cell is actually messenger RNA.

There are many other types of RNAs in the cell. These RNAs are referred to as functional or **non-coding RNA (ncRNA)**. (Figure 6.2)

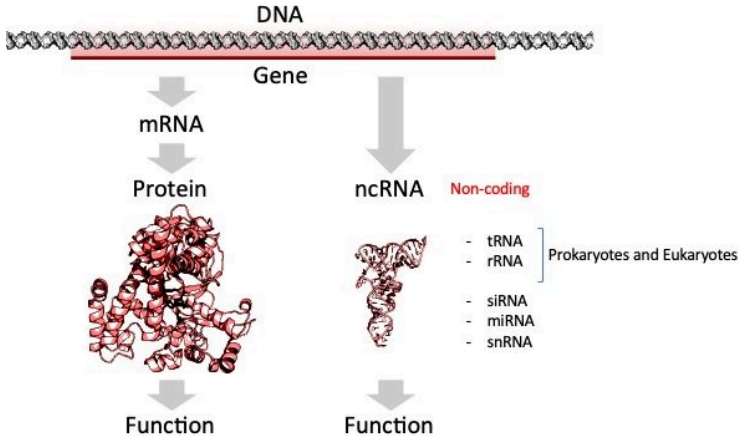


Figure 6.2. Transcripts may code for a protein or may be functional as RNAs. Image Credit: [Thomas Schaffee](https://commons.wikimedia.org/)
<https://commons.wikimedia.org/>

The RNA *itself fulfills* an important function, either enzymatic, structural, or as we shall see later in the semester in control of gene expression.

All cells (prokaryotes and eukaryotes) make three main kinds of RNA: **ribosomal RNA (rRNA)**, **transfer RNA (tRNA)**, and **messenger RNA (mRNA)**.

tRNA carries the appropriate amino acids into the ribosome for inclusion in the new protein. Meanwhile, the ribosomes themselves consist largely of ribosomal RNA (rRNA) molecules. **Quantitatively, rRNAs are by far the most abundant RNAs in the cell.**

Other small ncRNAs and long ncRNA molecules play a role in the regulation of transcriptional and translational processes and we will learn about those in later chapters.

The diagram below shows the central dogma in prokaryotes and eukaryotes. See if you can spot the differences between them as it relates to transcription by comparing the diagrams below!

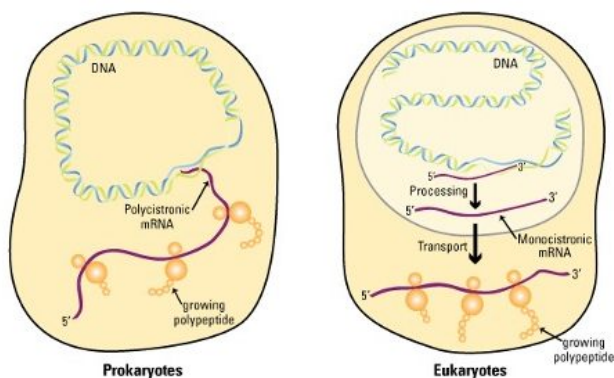


Figure 6.3 Comparison of transcription and translation in prokaryotes vs. eukaryotes. Image credit: Overview of Protein Expression. <https://www.thermofisher.com/us/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-protein-expression-systems.html>

Location

Most RNA transcripts in prokaryotes emerge from transcription ready to use! In fact, transcription and translation are coupled, with the association of ribosomes with mRNA and the translation of a polypeptide beginning even before the transcript is finished. This is because these cells have no nucleus.

Eukaryotic transcripts must exit the nucleus before they encounter the ribosomes in the cytoplasm.

mRNA type :

Eukaryotic transcripts have to undergo additional processing by trimming, splicing, or both!

In contrast to eukaryotes, some bacterial genes are part of operons whose mRNAs encode multiple polypeptides.

Chromatin:

DNA in bacteria is virtually ‘naked’ in the cytoplasm while eukaryotic DNA is wrapped up in chromatin proteins in a nucleus.

6.3 Rules of Transcription and Terminology

Since we are coming straight off a unit of Replication, it is useful in thinking about these 2 processes by comparing the 2 polymers (DNA v/s RNA) – we can use that fundamental information to get a basic outline of transcription and the rules of transcription like we did for replication.

Building an RNA strand is very similar to building a DNA strand. This is not surprising, knowing that DNA and RNA are very similar molecules.

Transcription is catalyzed by the enzyme **RNA Polymerase**. “RNA polymerase” is a general term for an enzyme that makes RNA. There are several different kinds of

RNA polymerases in eukaryotic cells, while in prokaryotes, a single type of RNA polymerase is responsible for all transcription.

Unlike DNA replication, however, only short sections of the genome (the genes) are transcribed. Different genes may be copied into RNA at different times in the cell's life cycle.

Further, our cells need billions of copies of proteins which will be made using these instructions. Therefore unlike DNA replication, many copies of RNA are made during transcription.

RNA synthesis: The substrates

Like DNA polymerases, RNA polymerases synthesize new strands **only in the 5' to 3' direction**, but because they are making RNA, they use ribonucleotides (i.e., RNA nucleotides) rather than deoxyribonucleotides. **(Figure 6.4)**

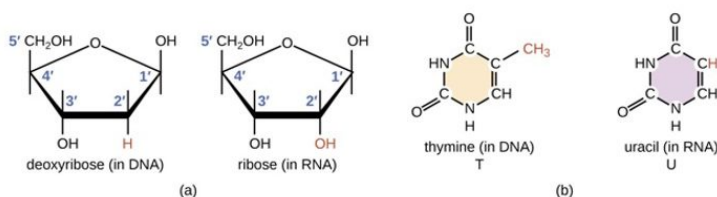


Figure 6. 4 RNA Structural Elements (a) Ribonucleotides contain the pentose sugar ribose instead of the deoxyribose found in deoxyribonucleotides. (b) RNA contains the pyrimidine uracil in place of thymine found in DNA. Figure from:Parker, et al (2019) Microbiology from Openstax

Ribonucleotides are joined in exactly the same way as deoxyribonucleotides, i.e., the 3'OH of the last nucleotide on the growing chain is joined to the 5' phosphate on the incoming nucleotide to make a phosphodiester bond.

One important difference between DNA polymerases and RNA polymerases is that the latter **does not require a primer** to start making RNA.

Once RNA polymerases are in the right place to start copying DNA, they just begin making RNA by joining together RNA nucleotides complementary to the DNA template.

RNA synthesis: The Template

A gene (DNA) is double-stranded, but **only one is transcribed into RNA!**

Within each section of a 'gene', ONE of the TWO strands carries the code/information to make the protein and is referred to as the sense strand or coding strand. (Figure 6.5)

The goal of transcription is to make a copy of that code accurately such that the 5' end of the code is also represented at the 5' end of the RNA!

Therefore during transcription, the opposite strand (antisense) or **non-coding strand** is used as a **template**.

RNA that is synthesized is complementary and antiparallel to the DNA template strand.

We can differentiate the two strands of DNA on the basis of their relation to the RNA product.

The sequence of the **template strand of DNA** is the **complement of that of the RNA transcript**. It is also referred to as the **non-coding strand or anti-sense (-) strand**.

In contrast, the **coding strand of DNA** has the **same sequence** as that of the **RNA transcript** except for thymine (T) in place of uracil (U). The coding strand is also known as the **sense (+) strand**.

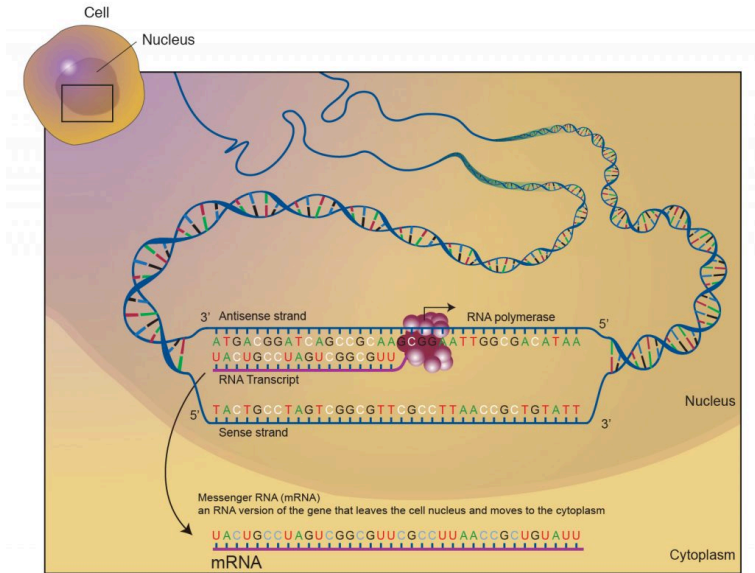


Figure 6.5 Illustration showing sense and antisense strand.
 Image credit: <https://www.genome.gov/genetics-glossary/messenger-rna>

Either of the 2 polynucleotide strands may contain a gene, and hence the determination of sense and antisense is gene-specific!

(This is very important to understand!)

The lecture video will walk you through some of the important terms. Make a similar template like the slides to write out information as you listen. Dr. Mehta Lecture Video: [L211 Introduction to Transcription \(1\)](#)

Did I Get This?



An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://iu.pressbooks.pub/iul211smehta/?p=833#h5p-22>

6.4 Genes are Transcription Units

Unlike the situation in replication, where every nucleotide of the parental DNA must eventually be copied, transcription, as we have already noted, only copies selected portions of the DNA into RNA at any given time.

Consider the challenge here: in a human cell, there are approximately 6 billion base pairs of DNA. Much of this is non-coding DNA, meaning that it will not need to be transcribed. The small percentage of the genome that is made up of coding sequences still amounts to between 20,000 and 30,000 genes in each cell. Of these genes, only a small number will need to be expressed at any given time.

This imposes a fundamental problem on the cell: how to recognize individual genes and transcribe them at the proper time and place. How do RNA polymerases know which DNA strand to read and where to start and stop?

That information is carried within the DNA sequence itself. There are patterns that indicate where RNA polymerase should start and end transcription.

The signatures within the DNA sequence also help a scientist (you!) take a sequence of letters and identify the **beginning, middle, and end of the gene.**

These **sequences** are recognized by the RNA polymerase

or by proteins that help RNA polymerase determine where it should bind the DNA to start transcription.

Gene Structure:

The stretch of DNA that encodes an RNA molecule and includes all the sequences necessary for its transcription is the transcription unit.

Much of the gene structure is *broadly similar* between eukaryotes and prokaryotes. These common elements largely result from the shared ancestry of cellular life in organisms with roughly 3.8 billion years of evolution.

Key differences in gene structure between eukaryotes and prokaryotes reflect their divergent transcription and translation machinery. Understanding gene structure is the foundation of understanding gene annotation, expression, and function.

Promoters:

Special DNA sequences, called **promoters**, direct the RNA polymerase to the proper site for the initiation of transcription.

A promoter is described as being situated upstream of the gene that it controls (Figure 7.57). What this means is that on the DNA strand that the gene is on, the promoter sequence is “before” the gene, or to put it differently, it is on the side of the gene opposite to the direction of transcription.

In this manner, the promoter actually indicates which of

the two DNA strands is to be read as the template and the direction of transcription.

Also, notice that the promoter is said to “control” the gene it is associated with.

This is because the expression of the gene is dependent on the binding of RNA polymerase to the promoter sequence to begin transcription. If the RNA polymerase and its helper proteins do not bind at the promoter, the gene cannot be transcribed and it will, therefore, not be expressed. Promoters also control the frequency of transcription.

Coding Region (Open Reading Frame):

The first nucleotide that will be transcribed into RNA is the **transcription start site** and given the number +1. (Note: the promoter is NOT part of the mRNA!)

All nucleotides added after form the open reading frame or RNA -coding region.

*Recall from the paragraphs earlier: only one of the strands encodes information that the RNA polymerase reads to produce protein-coding mRNA or non-coding RNA. This ‘sense’ or ‘coding’ strand, runs in the 5' to 3' direction where the numbers refer to the carbon atoms of the backbone’s ribose sugar. The **open reading frame** (ORF) of a gene is therefore usually*

represented as an arrow indicating the direction in which the sense strand is read

Terminator Region:

The DNA sequence that indicates the endpoint of transcription, where the RNA polymerase should stop adding nucleotides and dissociate from the template is known as a terminator sequence.

Terminators are usually part of the RNA-coding sequence; transcription stops only after the terminator has been copied into RNA.

Upstream and Downstream:

When DNA sequences are written out, often the sequence of only one of the two strands is listed. Molecular biologists typically write the sequence of the **nontemplate strand (coding strand)** because it will be the same as the sequence of the RNA transcribed from the template strand.

Upstream refers to sequences in the opposite direction from expression and are assigned **negative numbers**.

Nucleotides **downstream** of the start site are **assigned positive numbers**. There is no nucleotide numbered 0.



Figure 6.6 Transcription Unit showing essential elements of a gene. Image: Illustration by S. Mehta

Pause and watch: [L211 Introduction to Transcription \(2\)- Terminology](#)

Key Takeaways

- During Transcription, only certain sections of the DNA are transcribed at any one time.
- RNA is transcribed from a single strand of DNA. Within a gene, only one of the two DNA strands—the template strand—is usually

copied into RNA.

- Ribonucleoside triphosphates are used as the substrates in RNA synthesis.
- The transcribed RNA molecule is antiparallel and complementary to the DNA template strand. **Transcription is always in the 5'→3' direction, meaning that the RNA molecule grows at the 3' end.**
- The transcription unit contains all of the sequences that are necessary for both making the RNA as well as regulating it.
- Upstream of the start of the gene are Promoter sequences that are crucial for the binding of RNA polymerase to DNA.
- Terminator sequences signal the end of the gene, these are transcribed and found in RNA.

6.5 RNA Polymerase Enzymes

RNA Polymerase Enzymes (RNAPs) are required to carry out the process of transcription and are found in all cells ranging

from bacteria to humans. All RNAPs are multi-subunit assemblies (an example of a quaternary structure) and carry out the same reaction.

Bacterial RNAPs are the simplest form of RNA polymerases and provide an excellent system to study how they control transcription.

The **RNAP catalytic core** within bacteria contains five major subunits ($\alpha_2\beta\beta'\omega$ - 2 copies of alpha, beta, beta prime and omega) (Fig 10.7B). To position this catalytic core onto the **correct promoter** requires the association of a sixth subunit called the sigma factor (σ).

Together, the σ subunit and core polymerase make up what is termed the **RNA polymerase holoenzyme**.

The core polymerase is the part of the RNA polymerase that is responsible for the actual synthesis of the RNA, while the σ subunit is necessary for binding the enzyme at **promoters** to initiate transcription. The sigma (σ) subunit helps to find a site where transcription begins, participates in the initiation of RNA synthesis, and then dissociates from the rest of the enzyme.

The first sigma factor to be identified was sigma70 (σ^{70}) because it has a mass of 70kDa.

σ^{70} is the housekeeping sigma factor that is responsible for transcribing most genes in growing cells. It keeps essential genes and pathways operating.

How We Know

Biochemical assays combined with protein purification techniques of the kind you saw in Chapter 2 led to the identification of Bacterial RNA polymerase subunits.

It was found that RNA polymerase activity was associated with two protein species. A core polymerase (with subunit structure $\alpha 2\beta\beta'$) can transcribe DNA into RNA inefficiently and nonspecifically. When the sigma subunit, σ^{70} , is added, it can bind to core forming a holoenzyme ($\alpha 2\beta\beta'\sigma$) that is capable of specific engagement with duplex DNA at the beginning of genes (promoters) as well as efficient initiation of transcription. (1)

WATCH Lecture Video 3 for the experiments that led to the identification and role of the sigma factor: [L211 Introduction to Transcription \(3\): RNA polymerase](#)

Before you continue you should

1. Watch the Lecture videos that cover the material above.
 2. Complete the associated Lecture Quickcheck.
-

The role of the sigma factor is to help RNA polymerase find ‘authentic’ genes. Those with promoter sequences. What are those sequences?

Bacterial Promoters

Because the same RNA polymerase has to bind to many different promoters, it would be predicted that promoters would have some similarities in their sequences. Scientists examined many genes and their surrounding sequences and as expected, common sequence patterns were seen to be present in many promoters.

The locations of these nucleotides relative to the transcription start site are also similar in most promoters.

This common sequence pattern is called a **CONSENSUS SEQUENCE**. It is important to understand that each nucleotide in a consensus sequence is simply the one that appeared at that position in the majority (*consensus*) of promoters examined, and does not mean that the entire consensus sequence is found in all promoters.

The most common sequences found in bacteria are

A -10 sequence (Pribnow box): this is a 6 bp region centered about 10 bp upstream of the start site. The consensus sequence at this position is TATAAT. In other words, if you count back from the transcription start site, the sequence found at roughly -10 in the majority of promoters studied is TATAAT.

A -35 sequence: this is a 6 bp sequence at about 35 base pairs upstream from the start of transcription. The consensus sequence at this position is TTGACA.

The distance between these conserved sequences is 17-19bp. While the exact sequence of the spacer is not important that length with a separation of 17 nucleotides optimal.

Notice that there is a high proportion of (A)denines and (T)hymines in the σ^{70} promoter sequences. This is true for many promoters in both prokaryotic and eukaryotic genes. As you probably suspected, this is advantageous because there are only two H-bonds between A-T pairs (as opposed to 3 H-bonds between G-C pairs), which means that it is 33% easier to unzip.

The **figure** below shows the most common prokaryotic

promoter: the σ^{70} promoter, so-called because it is recognized and bound by the σ^{70} transcription factor.

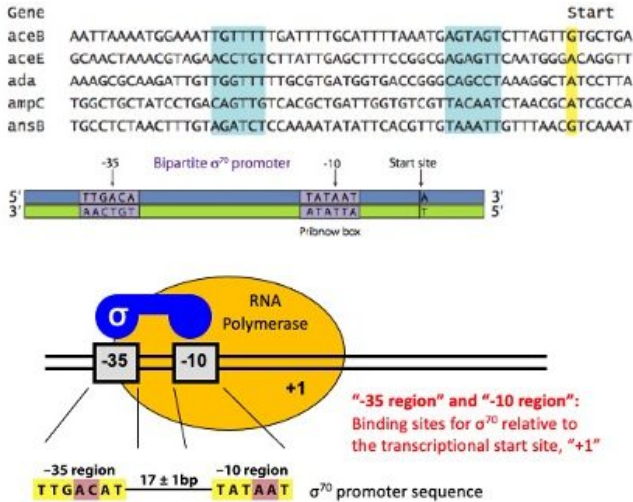


Figure 6.6. The image on top shows promoter sequences of various *E. coli* genes. Below is the consensus sequence for σ^{70} promoters. The sigma factor recognizes and makes contact with these sequences.

Notice the relationship between the various individual promoters and the consensus sequence. In general, those promoters with more matches to the consensus sequence are **stronger promoters**.

What does it mean to be a stronger (or weaker) promoter? First, keep in mind that the expression of any given gene is

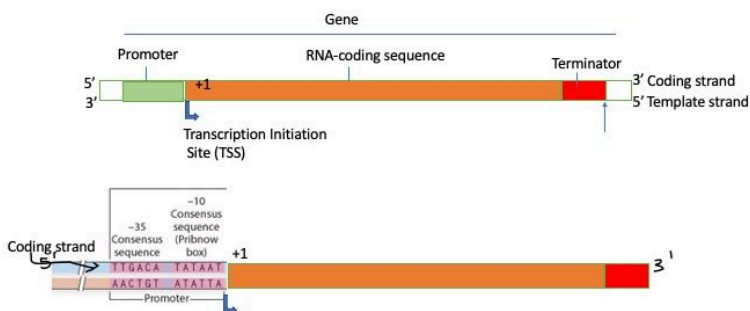
not automatic, or 100%. At any point in time, many of a cell's genes will be near 0% or shut off.

However, even genes that are turned on are transcribed at different rates. One of the governing factors is the recognition of the promoter site.

Genes with strong promoters are transcribed frequently—as often as every 2 seconds in *E. coli*. The RNA polymerase holoenzyme (with sigma) is more likely to recognize the site, dock properly, open up the double helix, and begin transcribing.

In contrast, genes with very weak promoters are transcribed about once in 10 minutes. RNA polymerase can potentially recognize weaker promoters, but it is less likely to do so, instead of passing it by as just another unimportant stretch of DNA.

We can now add more detail to our transcription unit as shown below. Note that in the gene shown below the promoter is on the left of the start site.



Alternative Sigma Factors

Within bacteria, there are multiple different sigma factors that can associate with the catalytic core of RNAP that help to direct the catalytic core to the correct DNA locations where RNAP can then initiate transcription.

E. coli σ^{70} is the housekeeping sigma factor that is responsible for transcribing most genes in growing cells. It keeps essential genes and pathways operating. Other sigma factors are activated during certain environmental situations, such as σ^{38} which is activated during starvation or when cells reach the stationary phase.

Alternative sigma factors give RNA polymerase holoenzyme specificity for different sets of promoters – as these sigma factors recognize different sets of consensus sequences at the -10 and -35 positions.

Therefore all genes (gene products) that are needed in response to assist bacteria during periods of starvation will share the same -10 and -35 consensus and will be coordinately expressed when needed.

This is our first look at ONE way in which prokaryotic cells can **regulate gene expression** of entire sets of genes!

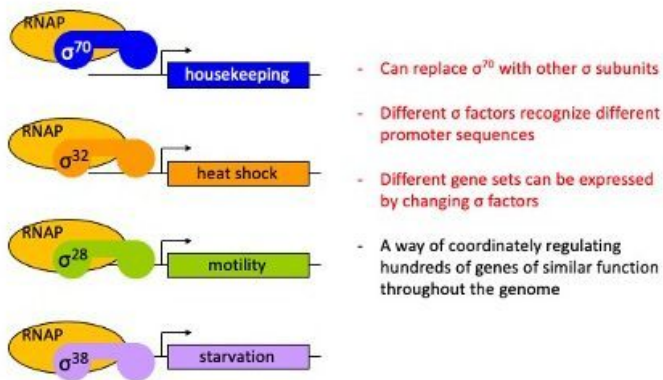


Figure 6.7 Alternative sigma factors can associate with RNA polymerase core enzyme, directing the complex to transcribe subsets of genes.

Pause and watch the Lecture video: [L211 Bacterial Transcription \(1\)](#)

6.6 Steps in Transcription

Like DNA Replication, RNA polymerase synthesizes RNA in three distinct phases that are also conceptually similar

Initiation: ‘Setting up’. In this phase RNA polymerase

holoenzyme locates and binds to promoter DNA to begin the synthesis of RNA>

Elongation: ‘The actual synthesis of RNA’. Here **RNA polymerase is moving** down the DNA template, unwinding the DNA ahead of it, and adding nucleotides one at a time (extending the 3’ OH)

As already mentioned, an RNA chain, complementary to the DNA template, is built by the RNA polymerase by the joining of the 5’ phosphate of an incoming ribonucleotide to the 3’OH on the last nucleotide of the growing RNA strand. Behind the RNA polymerase, the DNA template is rewound, displacing the newly made RNA from its template strand.

Termination: ‘Ending transcription’. Here the terminator sequences are recognized, the separation of the RNA molecule from the DNA template occurs and the enzyme ‘falls off’ as it is no longer needed.

Not unlike DNA Replication, the most elaborate step in transcription is Initiation which we will discuss first

6.6.1 Transcriptional Initiation

The events that comprise transcriptional initiation can be conceptually broken down into

1) Recognition of promoter 2) Unwinding of dsDNA and creation of bubble and 3) Initial synthesis and **escape of RNA polymerase** from the promoter.

Below is a summary of the events

a. RNA polymerase holoenzyme binds to the promoter to form a **closed complex**; at this stage, there is no unwinding of DNA.

b. Closed complex is then converted to an “open” complex by the separation of the DNA strands to create a transcription bubble about 12-14 base-pairs long. The conversion of the closed complex to the open complex also requires the presence of the σ subunit. The section of promoter DNA that is within it is known as a ‘**transcription bubble**’. **The transcription bubble is about 12-14 base-pairs long.**

Role of sigma factor

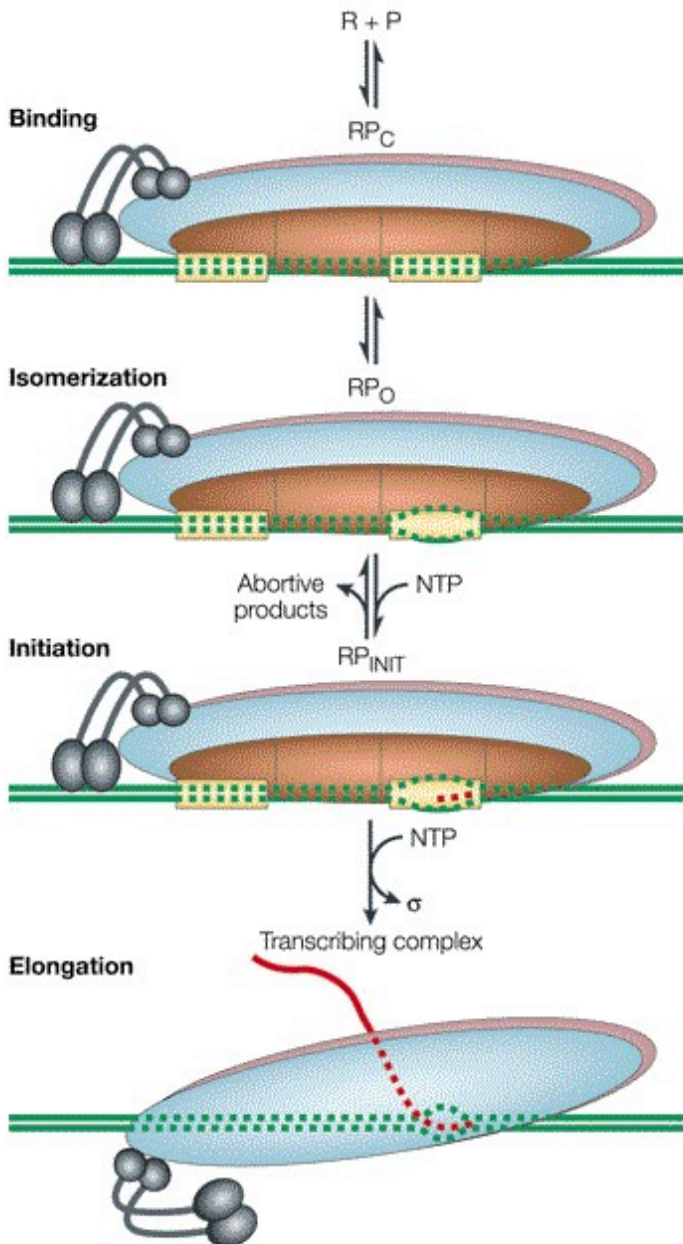
- Initial specific binding to the promoter by sigma factors of the holoenzyme sets in motion conformational changes that result in the separation of the two strands of DNA and expose a portion of the template strand.
- The sigma protein consists of different subunits. Some portions of sigma interact with the -35 element. The duplex DNA just upstream of the -10 element (-17 to -13) interacts with other parts of the protein σ^{2A} .
- The structure of the sigma unit is such that there is a recognition pocket for a nucleotide– the $A_{-11}(\text{nt})$ base from the duplex DNA. This base gets flipped into its recognition pocket in σ^{2A} . σ^A_2 is thought to be the key event in the initiation of promoter melting and the formation of the transcription bubble.
- Once the transcription bubble has formed and

transcription initiates.

c. **Abortive initiation:** Once the open complex has formed, the DNA template can begin to be copied, and the core polymerase adds nucleotides complementary to one strand of the DNA. The polymerase adds several nucleotides while still bound to the promoter, and without moving along the DNA template. Initially, **short pieces of RNA a few nucleotides long** may be made and released, **without the polymerase leaving the promoter**.

Part of it is due to the contacts the sigma factor still has with the promoter.

d. **Promoter Escape:** After several abortive initiation attempts, the polymerase synthesizes an RNA molecule from 9 to 12 nucleotides in length, which allows the polymerase to transition to the elongation stage. The σ subunit also dissociates from the core enzyme which ‘breaks free’ or ‘escapes’ into the gene.



The pathway of transcription initiation at bacterial promoters. The RNA polymerase (R) interacts with promoter DNA (P) to form the closed complex (RPC). Dashed lines show the promoter DNA that is bound by the RNA polymerase holoenzyme. The duplex DNA around the transcript start site is unwound (represented by a 'bubble' in the DNA that is bound by the RNA polymerase holoenzyme) to form the open complex (RPO). The initiating complex (RP_{INIT}) is formed and synthesis of the DNA-template-directed RNA chain (shown as a dashed red line) begins with formation of a phosphodiester bond between the initiating and adjacent phosphodiester nucleoside triphosphates (NTPs). Elongation is the final stage, and the RNA chain length increases, shown as a solid red line. Image from : February 2004
[Nature Reviews Microbiology](#) 2(1):57-65
DOI:[10.1038/nrmicro787](https://doi.org/10.1038/nrmicro787)

Pause and watch Lecture Video: [L211 Bacterial Transcription – Initiation](#)

Exercises



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=833#h5p-21>

6.6.2 Transcriptional Elongation

The core polymerase can move along the template, unwinding the DNA ahead of it to maintain a transcription bubble of 12-15 base pairs and synthesizing RNA complementary to one of the strands of the DNA.

RNA polymerases are by themselves processive (with no need for additional apparatus) adding hundreds or thousands of bases to the growing RNA at about 20-50 nucleotides per second.

Unlike DNA polymerases there is no elaborate mechanism for proof-reading. Although research has shown that RNA polymerase is capable of a type of proofreading in the course of transcription.

When RNA polymerase incorporates a nucleotide that does not match the DNA template, **it backtracks** and cleaves the last two nucleotides.

Transcription Termination

As mentioned earlier, a sequence of nucleotides called the terminator is the signal to the RNA polymerase to stop transcription and dissociate from the template.

Some terminator sequences, known as **intrinsic terminators**, allow termination by RNA polymerase *without* the help of any additional factors, while others, called **Rho-dependent terminators**, require the assistance of a protein factor called rho (ρ).

How does the sequence of the terminator cause the RNA polymerase to stop adding nucleotides and release the transcript?

To understand this, it is useful to know that the terminator sequence precedes the last nucleotide of the transcript. *In other words, the terminator is part of the end of the sequence that is transcribed.*

Intrinsic terminators

These makeup about 50% of all terminators in prokaryotes and have two common features.

First, they contain **inverted repeats**, which are sequences

of nucleotides on the same strand that are **inverted and complementary**. This sequence when transcribed into RNA can base-pair with each other to form a hairpin structure that contains a GC-rich run in the “stem” of the hairpin.

Second, adjacent to the inverted repeat is a stretch of 7- 9 Adenines. These when transcribed get converted to ‘U’s.

As RNA polymerase reaches and transcribes the terminator region the RNA will have a stem-loop structure. The secondary structure formed by the folding of the end of the RNA into the hairpin causes the RNA polymerase to pause!

Meanwhile, the run of U’s at the end of the hairpin permits the RNA-DNA hybrid in this region to come apart, because the base-pairing between A’s in the DNA template and the U’s in the RNA is relatively weak.

This allows the transcript to be released from the DNA template and from the RNA polymerase.

Rho-dependent termination

Transcription termination factor Rho is an essential protein in *E. coli* first identified for its role in transcription termination at Rho-dependent terminators, and is estimated to terminate ~20% of *E. coli* transcripts. The *rho* gene is highly conserved and nearly ubiquitous in bacteria.

Rho is a helicase and consists of a hexamer of six identical monomers arranged in an open circle. The protein can separate the transcript from the template it is paired

As in intrinsic termination, rho-dependent termination requires the formation of a hairpin structure in the RNA that causes pausing of the RNA polymerase.

Meanwhile, rho binds to a region of the **transcript called the rho utilization site (rut)**, a ~ cytidine-rich and poorly structured RNA sequence, and moves along the RNA till it reaches the paused RNA polymerase.

It then acts on the RNA-DNA hybrid, releasing the transcript from the template.

Lecture Videos: [L211 Bacterial Transcription – Elongation and Termination](#)

Remember to:

1. Watch the Lecture videos that cover the material above. This will help to clarify or reinforce certain concepts if they were unclear.
 2. Complete the associated Lecture Quick checks
 3. Begin work on Problem Set.
-

References and Attributions

This chapter contains material taken from the following CC-licensed content. Changes include rewording, removing paragraphs and replacing with original material, and combining material from the sources.

1. Bergtrom, Gerald, “Cell and Molecular Biology 4e: What We Know and How We Found Out” (2020). *Cell and Molecular Biology 4e: What We Know and How We Found Out – All Versions*. 13.

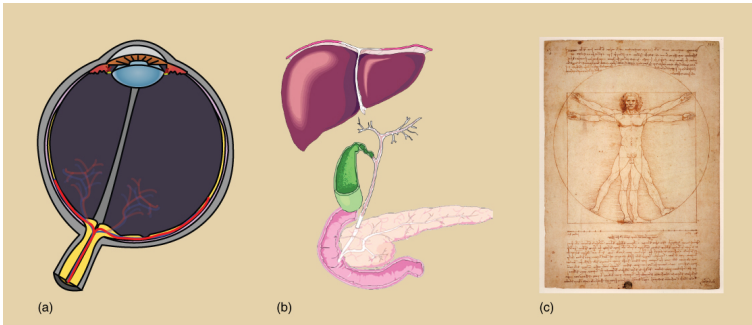
https://dc.uwm.edu/biosci_facbooks_bergtrom/13

2. Works contributed to LibreTexts by Kevin Ahern and Indira Rajagopal. LibreTexts content is licensed by CC BY-NC-SA 3.0. The entire textbook is available for free from the authors at <http://biochem.science.oregonstate.edu/content/biochemistry-free-and-easy>

3. Flatt, P.M. (2019) Biochemistry – Defining Life at the Molecular Level. Published by Western Oregon University, Monmouth, OR (CC BY-NC-SA). Available at: https://wou.edu/chemistry/courses/online-chemistry-textbooks/ch450-and-ch451-biochemistry-defining-life-at-the-molecular-level/?preview_id=4919&preview_nonce=cca8f0ce36&preview=true

7.

REGULATION OF GENE EXPRESSION - PROKARYOTES



7.1 Introduction

Each somatic cell in the body generally contains the same DNA. A few exceptions include red blood cells, which contain no DNA in their mature state, and some immune system cells that rearrange their DNA while producing antibodies. In general, however, the genes that determine whether you have green eyes, brown hair, and how fast you

metabolize food are the same in the cells in your eyes and your liver, even though these organs function quite differently. If each cell has the same DNA, how is it that cells or organs are different? Why do cells in the eye differ so dramatically from cells in the liver?

Similarly, all cells in two pure bacterial cultures inoculated from the same starting colony contain the same DNA, with the exception of changes that arise from spontaneous mutations. How is it that the same bacterial cells within two pure cultures exposed to different environmental conditions can exhibit different phenotypes?

Gene regulation is how a cell controls which genes, out of the many genes in its genome, are “turned on” (expressed).

Before you begin make sure you look at learning objectives.

Learning Objectives

Level 1 and 2 (Knowledge and Comprehension)

1. Draw a picture illustrating the general structure of an operon, and identify its parts.
2. Know the difference between positive and negative control? What is the difference between inducible and repressible operons?
3. Briefly describe the lac operon and how it controls the metabolism of lactose.
4. What is catabolite repression? How does it allow a bacterial cell to use glucose in preference to other sugars?

Level Up (Application, Analysis, Synthesis)

1. Predict for the following types of transcriptional control whether the protein produced by the

regulator gene will be synthesized **initially as** an active repressor or as an inactive repressor.

- Negative control in a repressible operon
- Negative control in an inducible operon

2. Predict for the following types of transcriptional control whether the protein produced by the regulator gene will be synthesized **initially as** an active form or inactive form.

- Positive control in a repressible operon
- Positive control in an inducible operon

NOTE: The mechanism of prokaryotic regulation was deciphered with the use of bacterial mutants. (See Link to learning).

3. Predict the effect of mutations in the following elements on the transcription of an operon

- Promoter
- Mutation at the operator prevents the regulator protein from binding, if regulator protein is a repressor AND operon is repressible operon.
- Mutation at the operator prevents the regulator protein from binding, if regulator protein is a repressor AND

operon is inducible operon

4. Use genetic data (phenotypes of mutant strains) for a fictitious operon to determine

- Type of operon (inducible, repressible)
- Which sequences are the promoter sequences.
- Which sequences correspond to regulatory gene.
- Which sequences are the structural genes.

5. Identify the level of transcription of a lac operon under given cellular conditions

6. Predict the effect of mutations within the following elements on the transcription of the Lac operon under different conditions.

- CAP (such that it can no longer bind the CAP site)
- Operator sequences
- Lac-I gene (repressor protein lac-I non-functional)
- Promoter of lac-operon

7.2 Overview of Regulation of Gene Expression

Define the term regulation as it applies to genes

For a cell to function properly, necessary proteins must be synthesized at the proper time.

In a given cell type, not all genes encoded in the DNA are transcribed into RNA or translated into protein because specific cells in our body have specific functions. Specialized proteins that make up the eye (iris, lens, and cornea) are only expressed in the eye, whereas the specialized proteins in the heart (pacemaker cells, heart muscle, and valves) are only expressed in the heart. At any given time, only a subset of all of the genes encoded by our DNA is expressed and translated into proteins. The expression of specific genes is a highly regulated process with many levels and stages of control. This complexity ensures the proper expression in the proper cell at the proper time.

The process of turning on a gene to produce RNA and protein is called **gene expression**. Whether in a simple unicellular organism or a complex multi-cellular organism, each cell controls *when* and *how* its genes are expressed.

Genomic DNA contains both *structural genes*, which encode products that serve as cellular structures or enzymes, and *regulatory genes*, which encode products that regulate gene expression. The expression of a gene is a highly regulated process. Whereas regulating gene expression in multicellular organisms allows for cellular differentiation, in single-celled organisms like prokaryotes, it primarily ensures that a cell's

resources are not wasted making proteins that the cell does not need at that time.

Elucidating the mechanisms controlling gene expression is important to the understanding of human health. Malfunctions in this process in humans lead to the development of cancer and other diseases. Understanding the interaction between the gene expression of a pathogen and that of its human host is important for the understanding of a particular infectious disease. Gene regulation involves a complex web of interactions within a given cell among signals from the cell's environment, signaling molecules within the cell, and the cell's DNA. These interactions lead to the expression of some genes and the suppression of others, depending on circumstances.

Prokaryotic versus Eukaryotic Gene Expression

Prokaryotic organisms are single-celled organisms that lack a cell nucleus, and their DNA, therefore, floats freely in the cell cytoplasm. To synthesize a protein, the processes of transcription and translation occur almost simultaneously. When the resulting protein is no longer needed, transcription stops. As a result, the primary method to control what type of protein and how much of each protein is expressed in a prokaryotic cell is the regulation of DNA transcription. All of the subsequent steps occur automatically. When more protein

is required, more transcription occurs. Therefore, in **prokaryotic cells**, the control of gene expression is mostly at the **transcriptional level**.

Recall from Chapter 6 that in eukaryotic cells, things are more complex. For one the DNA is contained inside the cell's nucleus and there it is transcribed into RNA. The newly synthesized RNA is then transported out of the nucleus into the cytoplasm, where ribosomes translate the RNA into protein. The processes of transcription and translation are physically separated by the nuclear membrane; transcription occurs only within the nucleus, and translation occurs only outside the nucleus in the cytoplasm. Thus regulation of gene expression in eukaryotes occurs at all stages of the process. Regulation may occur when the DNA is uncoiled and loosened from nucleosomes to bind transcription factors (epigenetic level) when the RNA is transcribed (transcriptional level) when the RNA is processed and exported to the cytoplasm after it is transcribed (post-transcriptional level) when the RNA is translated into protein (translational level), or after the protein has been made (post-translational level).

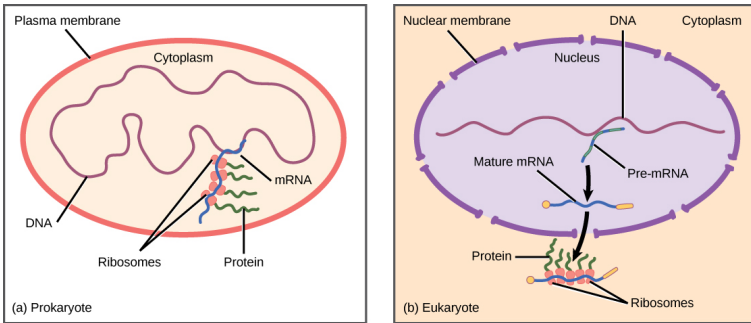


Figure 7.1. Prokaryotic transcription and translation occur simultaneously in the cytoplasm, and regulation occurs at the transcriptional level. Eukaryotic gene expression is regulated during transcription and RNA processing, which take place in the nucleus, and during protein translation, which takes place in the cytoplasm. Further regulation may occur through post-translational modifications of proteins.

The differences in the regulation of gene expression between prokaryotes and eukaryotes are summarized in the table below.

Differences in the Regulation of Gene Expression of Prokaryotic and Eukaryotic Organisms	
Prokaryotic organisms	Eukaryotic organisms
Lack a membrane-bound nucleus	Contain nucleus
DNA is found in the cytoplasm	DNA is confined to the nuclear compartment
RNA transcription and protein formation occur almost simultaneously	RNA transcription occurs prior to protein formation, and it takes place in the nucleus. Translation of RNA to protein occurs in the cytoplasm.
Gene expression is regulated primarily at the transcriptional level	Gene expression is regulated at many levels (epigenetic, transcriptional, nuclear shuttling, post-transcriptional, translational, and post-translational)

The regulation of gene expression is discussed in detail in subsequent modules. In this chapter, we consider systems of gene regulation in bacteria. Before we do let's look at examples of why understanding the regulation of gene expression in bacteria is relevant.

Clinical and Biological Relevance

It is becoming increasingly clear bacterial cells live in communities, interacting with other cells of their own species and of different species. They also exhibit community behaviors such as coordinated expression of genes within cells! For example, a type of community behavior is Quorum Sensing.

Here bacteria 'count' the presence of others and when there is an appropriate cell density reached (quorum!) then turn on the synthesis of genes together. This behavior is medically important. For example, some microbial species, such as *Staphylococcus aureus*, can encase their community within a self-produced matrix of hydrated extracellular polymeric substances that include polysaccharides, proteins, nucleic acids, and lipid molecules. These encasements are known as **biofilms**. Organisms within the biofilm are more than 1000-fold **more resistant to antibiotics** and more able to evade the host immune response than are free-living bacteria. Patients implanted with prosthetic devices, including simple bladder catheters, are especially at risk for biofilm formation.

WATCH LECTURE VIDEO: [Regulation of Transcription- Why it Matters](#)

Review



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=918#h5p-23>



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=918#h5p-25>



An interactive H5P element has been excluded from this version of the text. You

can view it online here:

[https://iu.pressbooks.pub/
iul211smehta/?p=918#h5p-26](https://iu.pressbooks.pub/iul211smehta/?p=918#h5p-26)

7.3 Gene Regulation in Prokaryotes

Operons

Bacterial genes with related functions—such as the genes that encode the enzymes that catalyze the many steps in a single biochemical pathway— are regulated together and found next to each other on the DNA. This cluster of genes share **ONE** promoter and regulatory sequences (explained below) that control the transcription of the unit.

The organization of genes in this manner is called an **OPERON**.

Transcription of the **OPERON** forms a **polycistronic mRNA** (Figure 7.2) – one mRNA that contains the

information to **make more than one protein**. The promoter then has simultaneous control over the regulation of the transcription of these structural genes because they will either **all be needed at the same time, or none will be needed**.

Grouping related genes under a common control mechanism allows bacteria to rapidly adapt to changes in the environment.

The organization of an operon is illustrated below in Figure 7.2

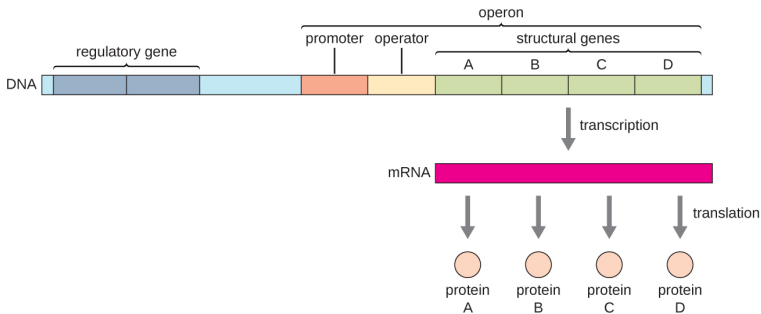


Figure 7.2 Schematic Representation of an Operon. In prokaryotes, structural genes of related function are often organized together on the genome and transcribed together under the control of a single promoter. The operon's regulatory region includes both the promoter and the operator. If a repressor binds to the operator, then the structural genes will not be transcribed. Alternatively, activators may bind to the regulatory region, enhancing transcription. Figure from: Parker, N., et. al. (2019) Microbiology. Openstax

The genes that encode these proteins used in metabolism or biosynthesis or that play a structural role in the cell are called **STRUCTURAL GENES**

Each operon includes DNA sequences that influence its own transcription; these are located in a region called the regulatory region.

The regulatory region includes the promoter and the region surrounding the promoter, to which **transcription factors**, proteins encoded by regulatory genes, can bind. Transcription factors influence the binding of **RNA polymerase** to the promoter and allow its progression to transcribe structural genes.

A **repressor** is a transcription factor that suppresses transcription of a gene in response to an external stimulus by binding to a DNA sequence within the regulatory region called the **operator**, which is located between the RNA polymerase binding site of the promoter and the transcriptional start site of the first structural gene. Repressor binding physically blocks RNA polymerase from transcribing structural genes.

Conversely, an **activator** is a transcription factor that increases the transcription of a gene in response to an external stimulus by facilitating RNA polymerase binding to the promoter.

An **inducer**, a third type of regulatory molecule, is a small molecule that either activates or represses transcription by interacting with a repressor or an activator.

Other genes in prokaryotic cells are needed all the time. These gene products will be **constitutively expressed**, or turned on continually. Most constitutively expressed genes are “housekeeping” genes responsible for overall maintenance of a cell.

The genes that code *for* these regulatory proteins (Activators and Repressors) are called **Regulatory Genes**. Transcription of these genes is under the control of its own promoter but often found next to the operon on the same DNA.

How does one transcript code for multiple proteins?

Just like DNA replication and transcription have different start and stop signals, translation also its own start and stop signals.

DNA replication starts at *origins* (this is on DNA), transcription starts at *promoters* (also on DNA) and translation begins on mRNA. The coding information for protein is buried within the mRNA and does not start at the transcriptional start site.

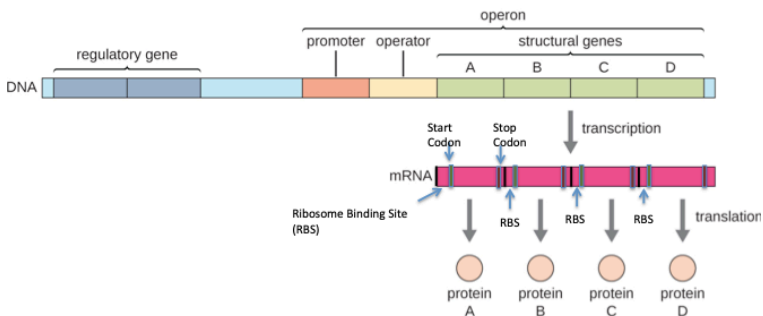
Just like DNA has extra sequences like the promoter that enable the RNA polymerase to bind and signals were mRNA transcription to begin, the **mRNA** as at its 5' end a leader sequence (untranslated region on the 5' end) that carries a ribosome binding site.

Translation of the protein begins at the translational **START Codon** (we will revisit this when we learn about translation in upcoming modules) and ends at the translation

STOP codon. The region between the 2 is the open reading frame.

Thus a polycistronic transcript carries many such open reading frames, each beginning with a translation initiation codon and consisting of a linear sequence of codons that specifies the protein.

A more accurate representation of an operon is shown below. Note that while the transcription start site is ONE (it is one long message- ONE promoter), within the message are present **multiple start** and stop codons and the region between represent the information to make the protein (Polycistronic).



Check your understanding

Use this quiz to check your understanding and decide whether to (1) study the previous section further or (2) move on to the next section.

<https://assessments.lumenlearning.co...sessments/6905>

cis- and *trans* regulators

Another term for DNA sequences that regulated transcription is *cis*-elements because they must be located on the same piece of DNA as the genes they regulate.

Binding sites for proteins involved in transcriptional regulation of the operon- **promoters, operators, and activator binding sites** are called ***cis*-elements**

On the other hand, the proteins that bind to these *cis*-elements are called ***trans*-regulators** because (as diffusible molecules) they do not necessarily need to be encoded on the same piece of DNA as the genes they regulate.

Cis and trans acting elements is a concept that will be relevant when predicting the effect of mutations. See Link to Learning.

Recall that regulation of gene expression or operons is occurring *in response* to some environmental signal. Therefore operons can be

Inducible: Where transcription is **normally off** (not taking place); something must happen to **induce** transcription, or turn it on. Usually, this is in response to a metabolite (a small molecule undergoing metabolism) that regulates the operon.

Repressible: Where transcription is **normally on (mRNA made, proteins made)**; something must happen to **repress** transcription, or turn it off.

The type of control (how these operons can be induced or repressed) is defined by the *mechanism it uses*.

For operons under negative control: The regulatory protein used is a repressor. Genes are expressed unless they are turned OFF by a **repressor that** binds to DNA and inhibiting transcription. Thus the operon will be turned OFF when the repressor is present, but ON when the repressor is absent or somehow inactivated.

For operons under positive control: The genes are expressed only when an active regulator protein, e.g. an

activator, is present. Thus the operon will be turned off when the positive regulatory protein is absent or inactivated.

SOLVED PROBLEM:

Let's take the basic parts and see if we can build the logic of gene regulatory circuitry.

We have a **NEGATIVE INDUCIBLE OPERON**.

Inducible = the operon is normally OFF
(inhibited) = no genes are expressed.

Negative Control = Repressor protein is used to control the expression of genes within this operon.

How it would work:

Operon normally off because repressor protein is bound to operator preventing RNA polymerase to bind.

A small molecule called an inducer accumulates and binds to the repressor protein. The binding of the inducer alters the shape of the repressor, preventing it from binding to DNA and thus turning ON (inducing transcription)! **The repressor is inactivated.**

Concepts in Context

WATCH:

Bonnie Bassler discovered that bacteria “talk” to each other, using a chemical language that lets them coordinate defense and mount attacks. The find has stunning implications for medicine, industry — and our understanding of ourselves.



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=918#oembed-1>

COMPLETE: Don't forget to complete the associate assignment on CANVAS

Check your understanding

Use this quiz to check your understanding and decide whether to (1) study the previous section further or (2) move on to the next section.

Figure out the logic of control for Positive Inducible, Negative Repressible and Positive Repressible operons.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=918#h5p-24>

WATCH Lecture Video: [Regulation of Prokaryotic Transcription-Terminology](#)

7.4 The lac Operon- An example regulation of bacterial gene expression

One of the best-understood examples of gene regulation is the *lac* operon of *Escherichia coli*. Regulation of the *lac* operon was first described by François Jacob and Jacques Monod (**See Link to Learning**). Their discoveries gave rise to a large subdiscipline within molecular biology devoted to the understanding of genetic regulation.

The preferred carbon and energy source for *E. coli* is glucose, but *E. coli* will instead metabolize lactose **if no glucose** is present in the growth medium.

Lactose is a disaccharide composed of the sugars galactose and glucose. β -galactosidase cleaves the glycosidic bond (a β -glycosidic bond that links the 1 position of galactose to the 4 positions of glucose) that connects galactose and glucose, thereby releasing free glucose and free galactose, which another cellular enzyme converts into glucose. (Figure 7.3)

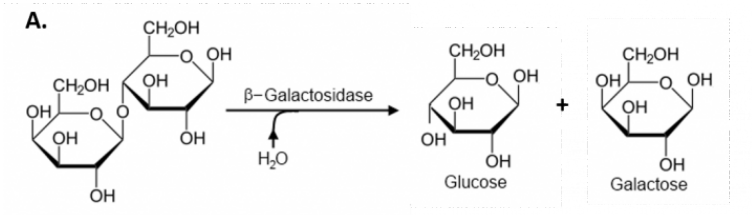


Figure 7.3. Reaction Controlled by the Expression of the Lac Operon. (A) Expression of the β -galactosidase enzyme enables the breakdown of lactose into the simple sugars, glucose and galactose for *E. coli* to use as a food resource. Figure modified from: [Andreas Piehler](#).

2 conditions must be met in order for the *lac*-operon to be expressed

- 1) Glucose must be absent AND 2) lactose must be present.

The operon consists of *lacZ*, *lacY*, and *lacA* genes that were called **structural genes**. By definition, structural genes encode proteins that participate in cell structure and metabolic function. As already noted, the *lac operon* is transcribed into an mRNA encoding the Z, Y and A proteins.

Let's take a closer look at the structure of the *lac* operon and the function of the Y, Z and A proteins (See Figure 7.4 A).

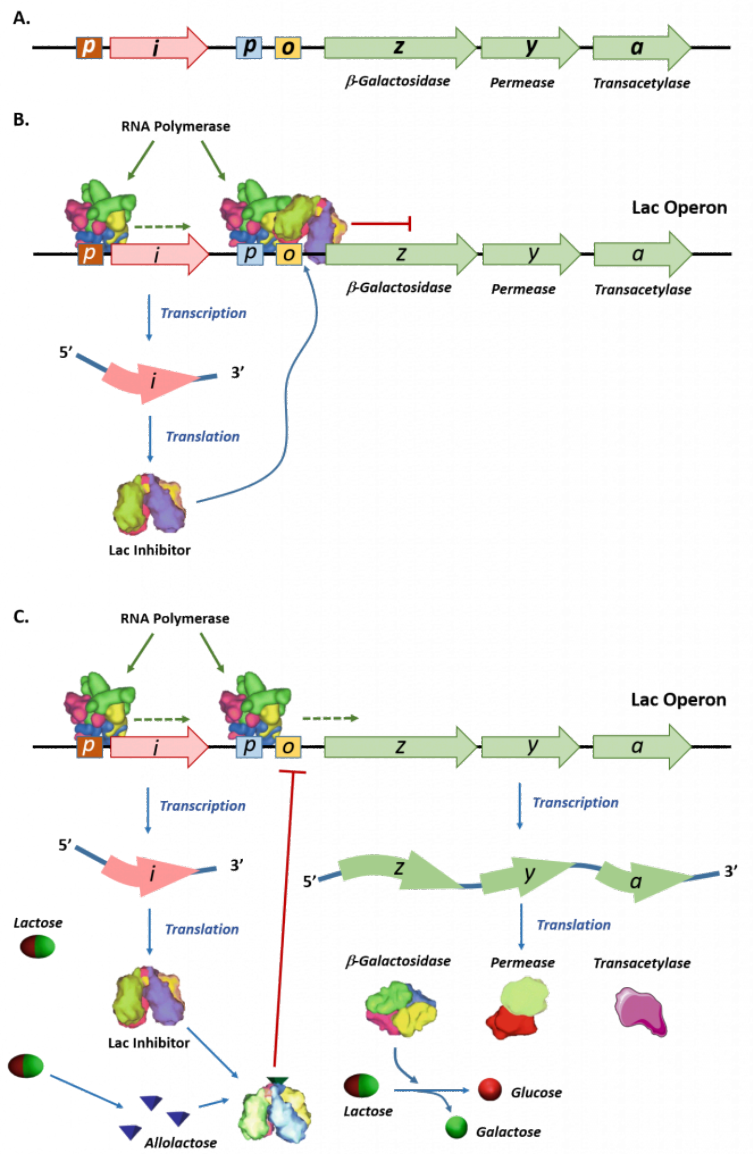


Figure 7.4 (A) Schematic representation of the lac operon in *E. coli*. The lac operon has three structural genes, *lacZ*, *lacY*, and *lacA* that encode for β -galactosidase, permease, and

galactoside acetyltransferase, respectively. The promoter (p) and operator (o) sequences that control the expression of the operon are shown. Upstream of the lac operon is the lac repressor gene, lacI, controlled by the lacI promoter (p). (B) Shows the lac repressor inhibition of the lac operon gene expression in the absence of lactose. The lac repressor binds with the operator sequence of the operon and prevents the RNA polymerase enzyme which is bound to the promoter (p) from initiating transcription. (C) In the presence of lactose, some of the lactose is converted into allolactose, which binds and inhibits the activity of the lac repressor. The lac repressor-allolactose complex cannot bind with the operator region of the operon, freeing the RNA polymerase and causing the initiation of transcription. Expression of the lac operon genes enables the breakdown and utilization of lactose as a food source within the organism. Figure modified from: [Esmaeili, A., et. al. \(2015\) BMC Bioinformatics 16:311](#)

The lacZ gene encodes **β -galactosidase**, the enzyme that breaks lactose (a disaccharide) into *galactose* and glucose.

The lacY gene encodes lactose **permease**, a membrane protein that facilitates lactose entry into the cells.

The role of the lacA gene (a **transacetylase**) in lactose energy metabolism is not well understood.

The **I gene** to the left of the lac Z gene is a **regulatory gene** (to distinguish it from structural genes).

The **operator** sequence separating the I and Z genes is a transcription regulatory DNA sequence.

7.4.1 Negative Regulation of the

Lac Operon by Lactose

How does **lactose** turn ON transcription of the lacoperon?

The regulatory protein is known as the lactose operon repressor or LacI. The gene for LacI is located just upstream of the lacoperon and is transcribed from its own separate promoter.

Lac I is always made and present in *E. coli* cells!

In the absence of lactose in the growth medium, the repressor protein binds tightly to the **operator DNA**.

Since the operator partially overlaps with the promoter, the presence of LacI blocks RNA polymerase from accessing the promoter and hence blocks transcription. Under these conditions, little or no transcript is made. **(Figure 7.4 B)**

Let's look more closely at how the repressor prevents RNA polymerase from binding to the promoter. When RNA polymerase binds to the promoter, it physically contacts a stretch of DNA that extends upstream to roughly position -40 relative to the start site of transcription (***recall that the sigma factor contacts the -35 and -10 sequences***) and downstream to roughly position +20.

Meanwhile, the stretch of DNA contacted by the repressor, the operator, overlaps with the downstream region of the promoter, covering the transcription start site and extending past the end of the promoter **(Figure 7.5)**. Thus, when the repressor binds to the operator, it physically occludes RNA polymerase.

LacI is therefore a classic example of negative regulation in which the binding of a regulatory protein to DNA represses transcription. (We will come to positive regulation presently.)

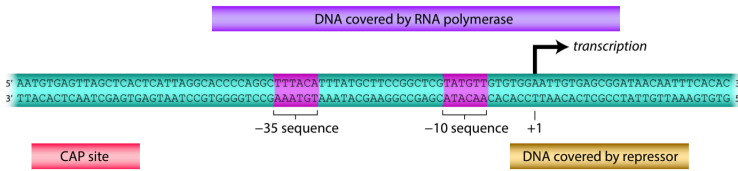
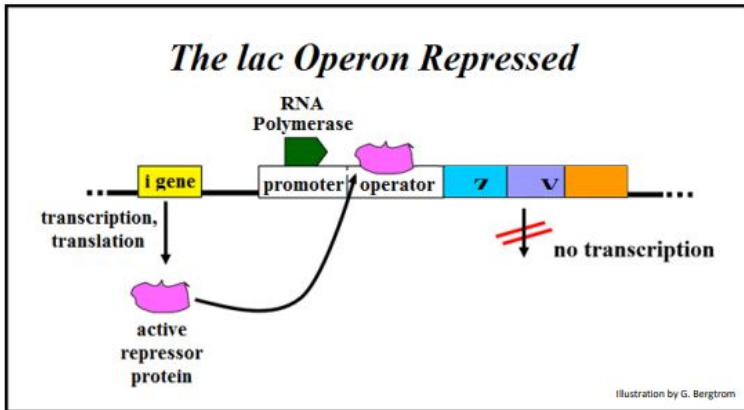


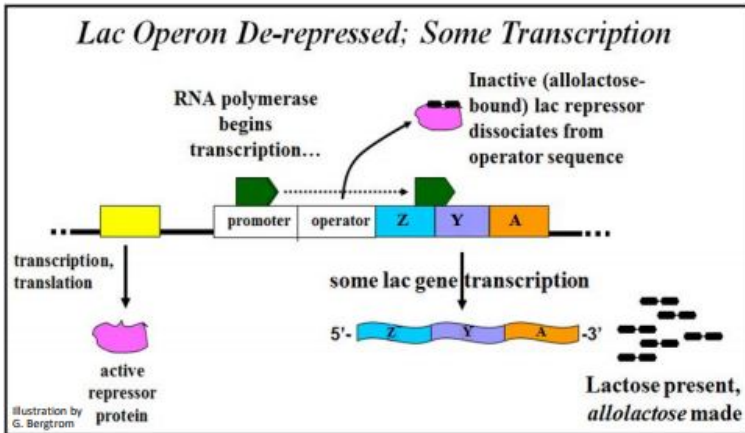
Figure 7.5 Binding of the repressor to the operator occludes RNA polymerase. Shown are the DNA binding sites for RNA polymerase, the repressor, and CAP, which is introduced below. Image from : <https://www.labxchange.org/library/items/lb:LabXchange:a17ca615.html:1>. The content within the pages is from <https://projects.iq.harvard.edu/lifesciences1abookv1>



How does the lacoperon escape repression to turn on the synthesis of β -galactosidase when lactose is present in the growth medium instead of glucose?

Here lactose itself serves as the inducer! If cells are grown in the presence of lactose, some of the lactose entering the cells is converted to **allolactose**. (*conversion to allolactose occurs by β -galactosidase!*)

Allolactose binds to the repressor sitting on the operator DNA to form a 2-part complex. The allosterically altered repressor dissociates from the operator and RNA polymerase can transcribe the *lac* operon genes as illustrated in **Figure 7.4 C and image below**.



WATCH Lecture Video: [Lac-Operon -Negative Regulation](#)

7.4.2 Positive Regulation of the Lac Operon; Induction by Catabolite Activation

Bacteria typically have the ability to use a variety of substrates as carbon sources. However, because glucose is usually preferable to other substrates, bacteria have mechanisms to ensure that alternative substrates *are only used* when glucose has been depleted.

Recall that 2 conditions must be met in order for the *lac*-operon to be expressed.

1) Glucose must be absent **AND** 2) lactose must be present.

How does the bacterial cell sense the availability of glucose?

In addition to being subject to negative control by repressor binding to the operator at the downstream end of the promoter, the lacoperon is subject to positive control by an activator called CAP (cAMP-bound **catabolite activator protein** or cAMP receptor protein).

CAP binds to a site (CAP Binding Site) just upstream of the promoter such that both CAP and RNA polymerase can sit side-by-side on the DNA. This is in contrast to the repressor, whose binding site overlaps with the binding site for RNA polymerase.

Why does RNA polymerase require the assistance of CAP to bind to the promoter in the presence of an inducer?

If the inducer is present, then, as we have seen, the LacI repressor is **not** bound to the operator and hence RNA polymerase *should be* able to bind to the promoter and initiate transcription?

The answer is that the lac promoter is a poor match to the -35 and -10 consensus sequences. As you will recall, the ideal -35 and -10 sequences are 5'-TTGACA-3' and 5'-TATAAT-3', respectively. The promoter for the *lac* operon differs from these ideal sequences at three positions. Hence,

the *lac* promoter is an ***intrinsically weak promoter*** to which RNA polymerase only weakly binds.

This is the basis for positive control; an activator compensates for the promoter's poor match to the consensus sequence by helping to facilitate the binding of RNA polymerase.

How does CAP facilitate the binding of RNA polymerase? It does so by directly contacting the RNA polymerase, and the favorable free energy from this protein-protein interaction helps to stabilize the binding of RNA polymerase to the otherwise weak promoter. Situations such as these in which an activator stabilizes the binding of RNA polymerase to DNA are often referred to as **recruiting** RNA polymerase.

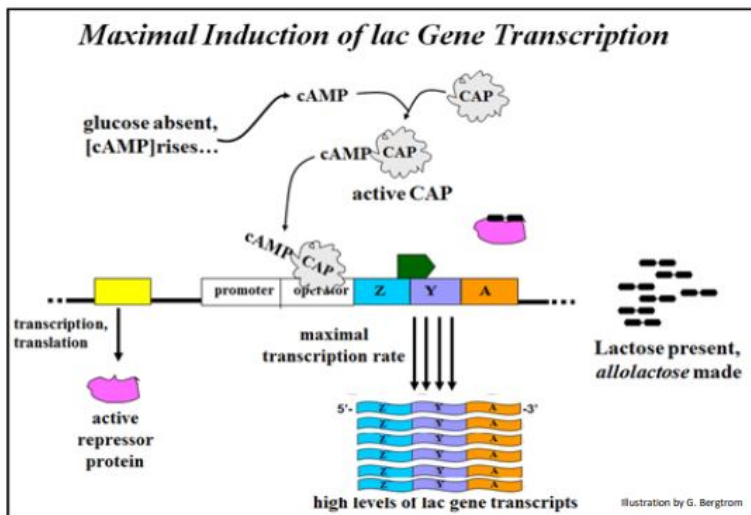
Just as the affinity of the Lac-I repressor for DNA is governed by a small molecule, the inducer allolactose, the ability of CAP to adhere to its binding site is strongly influenced by a small molecule, 3',5'-cyclic adenosine monophosphate (cyclic-AMP- cAMP)

When glucose is available, cellular levels of cAMP are low in the cells and CAP is in an inactive conformation. When glucose is scarce, the accumulating cAMP binds to **catabolite activator protein (CAP)**. The complex binds to the promoter region of the *lac* operon. The binding of the CAP-cAMP complex to this site increases the binding ability of RNA polymerase to the promoter region to initiate the transcription of the structural genes.

Thus, in the case of the *lac* operon, for transcription to

occur, lactose must be present (removing the lac repressor protein) and glucose levels must be depleted (allowing binding of an activating protein). The result is the synthesis of **higher levels of lac enzymes** that facilitate efficient cellular use of lactose as an alternative to glucose as an energy source.

Maximal *activation* of the lac operon in high lactose and low glucose is shown below.



When glucose levels are high, there is catabolite repression of operons encoding enzymes for the metabolism of alternative substrates. Because of low cAMP levels under these conditions, there is an insufficient amount of the CAP-cAMP complex to activate transcription of these operons.

WATCH Lecture Video: [L211 Lac-Operon -Positive Regulation](#)

Did You Get It?

-  An interactive H5P element has been excluded from this version of the text. You can view it online here:
<https://iu.pressbooks.pub/iul211smehta/?p=918#h5p-27>

Let's look at some of the classic experiments that led to our understanding of E. coli gene regulation in general, and of the lac operon in particular.

LINK TO LEARNING: How we know

LO: Predict the effect of mutations in the following elements on the transcription of an operon.

In molecular biology, one of the most common methods for figuring out a gene's function is to mutate it and measure the resulting effects on its organism's phenotype.

François Jacob and Jacques Monod first described the “operon model” for the genetic control of lactose metabolism in *E. coli* in 1961. Jacob and Monod deduced the structure of the operon *genetically* by analyzing the interactions of mutations that interfered with the normal regulation of lactose metabolism.

They knew that wild-type *E. coli* would **not** make the β -galactosidase, β -galactoside permease, or β -galactoside transacetylase proteins when grown **on glucose**.

Of course, they also knew that the cells would switch to lactose for growth and reproduction if they were deprived of glucose! They then searched for and isolated different *E. coli* mutants that could not grow on lactose, even when there was no glucose in the growth medium.

Jacob and Monod deduced the structure and various regulatory elements using genetics.

You already know how regulation of lac operon occurs and therefore should be able to predict the effect of mutations on various components of the lac operon.

GO TO: <http://www.dnafb.org/33/index.html>

Follow the tabs: Concept, Animation, and Problem.

Crucial to the experiments described in the video was the creation of **partial diploid strains** of *E. coli*, in which 2 copies of the operon were present: one on the chromosome and one on a plasmid.

These partially diploid prokaryotes **merodiploids** (“mero-” comes from the Greek word for “part”, or “partial”). Merodiploids can be produced in a lab setting,

Merodiploids of *E. coli* is a fantastic research tool. They allow us to examine how wild-type and

mutated alleles interact within a living organism. Genetic tests using such diploids distinguished between mutations in the genes coding for trans-acting elements or within the regulatory sequences.

Some terms:

UN-INDUCIBLE mutants: Mutations in the regulatory circuit that abolish expression of the operon

CONSTITUTIVE mutants: A mutant in which a protein is produced at a constant level, as if continuously induced; a bacterial regulatory mutant in which an operon is transcribed in the absence of inducer; a mutant in which a regulated enzyme is in a continuously active form.

For example Mutations in Lac I gene (that codes for the repressor) and the Operator sequences would be constitutive.

The partial diploid strains would be useful in distinguishing between the options.

Cis-acting mutations: Only affect those genes on the contiguous stretch of DNA. Mutations in promoter sequences, regulatory sequences (operator) were identified as cis-acting mutations.

Cis-dominant: A site or mutation that affects the

properties of its own molecule of DNA, often indicating that it does not encode a diffusible product.

Trans-acting mutations: Repressors and activators are trans-acting; that is, they affect the expression of their regulated genes no matter on which DNA molecule in the cell these are located.

For In-Class Activity/Problem

You will be looking at a variety of mutations that can occur in lac operon genes and discussing the effects of those mutations on *E. coli*. To do this, we'll be using the following symbols to represent the individual components of the lac operon:

I P O Z Y A

Since the function of *lac A* is not well defined, we'll be leaving it out of this model more often than not.

When all the sequences are, the lac operon functions normally. We'll represent this using the following notation

$$I^+ P^+ O^+ Z^+ Y^+ A^+$$

If a given gene is mutated, we'll change the superscript above that gene. Listed below are the specific mutations

Null mutation: Denoted by X^- (where X can be any genetic element on the operon), DNA sequences with this mutation have completely lost their normal activity. In protein-coding genes, this means no protein is produced. In regulatory genes, this means that regular binding sites are non-functional

Constitutive activity: Denoted by O^C , this mutation is specific to the operator region. Constitutively active operator regions always block the binding of repressor protein to the operator region. This results in transcription of the operon whether or not lactose is present, because the repressor is unable to block RNAPol from binding to the promoter.

When Merodiploids are used the following notation is used:

$$I^+ P^+ O^+ Z^+ Y^+ A^+ / I^+ P^+ O^+ Z^+ Y^+ A^+$$

In this notation, we show a chromosomal lac operon and the plasmid lac operon side by side. Again, we've included the *lacA* gene here for completeness but will be leaving it out of our exercises.

Because merodiploids have two copies of a given set of genes, mutations affect them differently. For example, if a single copy of a protein-coding gene is inactivated, the second copy may still continue to produce viable protein, effectively masking the mutation. Here is where 'Trans-acting' and 'Cis-acting' becomes relevant!

For extra practice:

Try this [simulation](#) (System requirements: this requires JaVA but should operate on most computers)

Remember to:

1. Watch the Lecture videos that cover the material above.
This will help to clarify or reinforce certain concepts if they were unclear.
 2. Complete the associated Lecture Quick checks
 3. Begin work on Problem Set.
-

References and Attributions

This chapter contains material taken from the following CC-licensed content. Changes include rewording, removing paragraphs and replacing with original material, and combining material from the sources.

1. Bergtrom, Gerald, “Cell and Molecular Biology 4e: What We Know and How We Found Out” (2020). *Cell and Molecular Biology 4e: What We Know and How We Found Out – All Versions*. 13.

https://dc.uwm.edu/biosci_facbooks_bergtrom/13

2. Flatt, P.M. (2019) Biochemistry – Defining Life at the Molecular Level. Published by Western Oregon University, Monmouth, OR (CC BY-NC-SA). Available at: <https://wou.edu/chemistry/courses/online-chemistry-textbooks/ch450-and-ch451-biochemistry-defining-life-at->

[the-molecular-level/?preview_id=4919&preview_nonce=cca8f0ce36&preview=true](https://openstax.org/r/the-molecular-level/?preview_id=4919&preview_nonce=cca8f0ce36&preview=true)

3. Biology 2e. **Provided by:** OpenStax. **Located at:** <http://cnx.org/contents/185cbf87-c72e-48f5-b51e-f14f21b5eabd@10.8>. **License:** *CC BY: Attribution*. **License Terms:** Access for free at <https://openstax.org/books/biology-2e/pages/1-introduction>

4. Positive Regulation of the Lac Operon; Induction by Catabolite Activation modified from following. **License Terms:** Access for free at <https://www.labxchange.org/library/items/lb:LabXchange:f4fa7330:html:1>

8.

EUKARYOTIC TRANSCRIPTION AND REGULATION

Learning Objectives

1. What polymerases transcribe eukaryotic genes? (Name and type of gene it transcribes).
2. Describe the three processes that commonly modify eukaryotic pre-mRNA.
3. Answer these questions concerning promoters: •What role do promoters play in transcription? • Explain why eukaryotic promoters are more variable than bacterial promoters. •What elements constitute a 'core promoter'.
4. Bacterial and eukaryotic gene transcripts can differ, in the transcripts themselves, in whether the

transcripts are modified before translation, and in how the transcripts are modified. For each of these three areas of contrast, describe what the differences are.

5. What are Enhancer sequences and how are they different from core promoter sequences?
6. What does it mean that enhancers are position- and orientation-independent?
7. What is combinatorial control?
8. What role does the mediator play in transcription
9. How do transcriptional repressors work? (In lecture videos)
10. What purposes do capping and poly-A tail addition serve for eukaryotic mRNAs?
 - Show the pathway for cap formation
 - Answer at what stage of mRNA formation is the cap added to the RNA molecule.
11. Describe the basic assembly of PIC as deduced from in vitro experiments.
12. What steps in the eukaryotic transcription cycle are stimulated by phosphorylation of the carboxyl-terminal (CTD) of the large subunit of RNA polymerase II?

LEVEL UP (combines molecular bio methods introduced thus far, and others introduced later)

1. Interpret reporter gene assay data for the identification of regulatory elements (all types

- including enhancers) in eukaryotic genes.
2. Interpret data that leads to the identification of general transcription factors.
 3. Explain how mutations in regulatory regions of genes differ from mutations in the coding region.

8.1 Introduction

In Chapter 7 Introduction to Transcription, you learned about the basics of transcription (which is the same for prokaryotes and eukaryotes) as well as some differences between transcription in prokaryotes and eukaryotes. As a reminder, some differences are listed again below

- *E. coli* uses a **single RNA polymerase** enzyme to transcribe all kinds of RNAs while eukaryotic cells use **different RNA polymerases** to catalyze the syntheses of **ribosomal RNA** (*rRNA*), **transfer RNA** (*tRNA*), and **messenger RNA** (*mRNA*).
- In contrast to eukaryotes, some bacterial genes are part of **operons** whose mRNAs encode multiple polypeptides. Eukaryotic genes are not part of operons.
- Most RNA transcripts in prokaryotes emerge from transcription ready to use for translation! In eukaryotes, transcription and translation occur in different compartments (nucleus – cytoplasm)
- Eukaryotic transcripts synthesized as longer precursors

undergo *processing* by *trimming*, *splicing*, or both!

Importantly eukaryotic DNA is wrapped up in chromatin proteins in a nucleus. Therefore the default state of transcription in eukaryotes is 'off'!

The implication of this is that to begin transcription of any eukaryotic gene, there must be mechanisms to 'activate it'. Eukaryotic transcription initiation cannot be separated from regulation!

**Begin by watching the Lecture Videos
within CANVAS**

This chapter consists of descriptions of key concepts and terms introduced in the lecture videos.

Research methodology and specific biomedical or relevant examples are highlighted in detail within the lecture videos.

8.2 Eukaryotic Cells Have

Three Types of RNA Polymerase

RNA Polymerase I (Pol I) is responsible for the synthesis of the majority of rRNA transcripts, whereas RNA Polymerase III (Pol III) produces short, structured RNAs such as tRNAs and 5S rRNA. RNA Polymerase II (Pol II) produces all mRNAs and most regulatory and untranslated RNAs.

Did You Know?

The death cap mushroom produces a toxin α -Amanatin. The lethal effect of this toxin is due to its effect on RNA polymerases. The poison binds very tightly to RNA polymerase II and effectively prevents transcription.

The chemistry of RNA polymerization is identical in all types of organisms, and the three eukaryotic RNA polymerases are structurally related to *E. coli* RNA Polymerase; consist of homologs of 5 prokaryotic **core subunits** that form the same characteristic crab-claw shape in addition to other subunits.

In addition to homologs of the core subunits, there are many more polypeptides that make up the eukaryotic RNA polymerases.

One of the subunits of RNA Polymerase II possesses a unique CTD (carboxy-terminal domain) consisting of multiple repeats of a special heptameric (Hepta-7) amino acid sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser that repeats itself.

In mammals, this domain consists of 52 repeats of the amino acid sequences. Serines in each repeat unit can be modified by the addition of a phosphate group, causing a substantial change in the properties of the polymerase.

The phosphorylation of the CTD of RNA polymerase plays an important role in transcription and mRNA processing.

8.3 Overview of Gene Expression (From DNA to Protein)

We focus on initiation by RNA pol II, the polymerase that produces all mRNAs and most regulatory and untranslated RNAs. Below (Figure 8.1) is a diagram of elements of the eukaryotic gene- that include all the sequences necessary to regulate transcription in addition to the protein-coding sections.

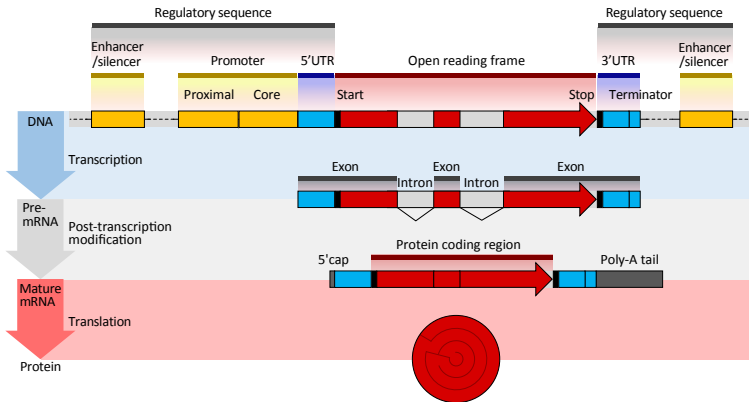


Figure 8.1 The structure of a eukaryotic protein-coding gene. Regulatory sequence controls when and where expression occurs for the protein coding region (red). Promoter and enhancer regions (yellow) regulate the transcription of the gene into a pre-mRNA which is modified to add a 5' cap and poly-A tail (grey) and remove introns. The mRNA 5' and 3' untranslated regions (blue) regulate translation into the final protein product. Image Attribution: [Thomas Shafee, CC BY 4.0 <https://creativecommons.org/licenses/by/4.0>, via Wikimedia Commons](https://creativecommons.org/licenses/by/4.0/)

The structure of eukaryotic genes includes features not found in prokaryotes (Figure 8.1).

Eukaryotic genes typically have more regulatory elements to control gene expression compared to prokaryotes.

An additional layer of regulation occurs for protein-coding genes after the mRNA has been processed to prepare it for translation to protein.

Only the region between the start and stop codons encodes

the final protein product. The flanking untranslated regions (UTRs) contain further regulatory sequences. The 3' UTR contains a terminator sequence, which marks the endpoint for transcription and releases the RNA polymerase, and also contains sequences that regulate mRNA stability.

The 5' UTR contained sequences that serve as landing pads for translational machinery (ribosome and other factors). In the case of genes for non-coding RNAs, the RNA is not translated but instead folds to be directly functional.

The most striking difference is the extent to which eukaryotic mRNA (pre-mRNA) is modified **to produce mature mRNA ready for translation into protein.**

These include:

- addition of a 5' CAP at 5' end of mRNA produced.
- splicing, the removal of the intron regions, and joining together of exons (the protein-coding portions)
- addition of a Poly A tail (polyadenylation) that is an inherent part of the termination mechanism.

Importantly most processing occurs while mRNA is being synthesized (co-transcriptional) and some soon after transcription (post-transcriptional). For example, the cap is added as soon as transcription has been initiated, splicing and editing begin while the transcript is still being made.

However to deal with all of these events together would be confusing, with too many different things being described at once.

We will therefore postpone mRNA processing until

after we have talked about the Initiation of transcription. We will consider splicing completely separately after we discuss capping, elongation, and polyadenylation.

8.4 Details of Eukaryotic Transcription Initiation

As depicted in Figure 8.1 transcription starts downstream of the promoter and creates a transcript that begins with a 5' untranslated region (5'UTR) followed by the coding region **which may include multiple introns** and ending in a 3' untranslated region or 3'UTR.

RNA Pol II gene transcription in eukaryotes is tightly regulated, controlled by a highly complex multicomponent machinery. A plethora of proteins, *more than a hundred in humans*, are organized in often very large multiprotein assemblies.

8.4.1 Eukaryotic Promoters

Eukaryotic promoters are more complex. They include all the sequences that are necessary for both initiation of transcription of a gene as well as regulatory sequences.

The promoter is located at the 5' end of the gene and can be divided into the

CORE PROMOTER– which represents a **minimal set of sequences** necessary for assembly of the transcription machinery and transcription initiation. This allows for **BASAL LEVELS of TRANSCRIPTION**.

‘PROXIMAL PROMOTER ELEMENTS’ – regulatory sequences next to the core promoter sequences.

While the assembly of the initiation complex can occur on core promoter sequences, almost all genes have additional proteins called Transcriptional Activators that bind to Proximal Promoter elements and ‘promote’ gene expression.

An alternative term used for these are “upstream promoter elements” or “upstream regulatory elements”

“ENHANCERS/SILENCER- DISTAL”– these sequences are located many thousands of base pairs away. The binding of different transcription factors, therefore, regulates the rate of transcription initiation at different times and in different cells.

Enhancer and Silencer sequences dictate **when (developmental stage) and where (what tissue) a gene gets expressed**.

Core Promoter Sequences:

The core promoter is a region encompassing approx 50 base pairs (bp) upstream and aprox 50 bp downstream of the TSS.

It includes or encompasses the TSS!

First, it is important to note that not all eukaryotic genes look alike! There is no exact set sequence or a minimum

number of core promoter sequences. Most genes have some combinations of elements, and scientists are still identifying core promoter consensus sequences.

Below are some “typical core promoter consensus” sequences.

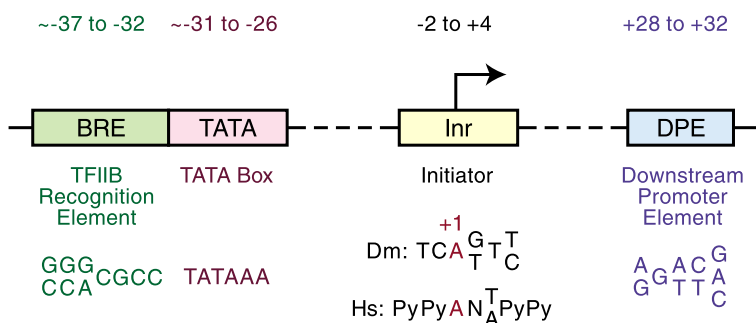


Figure 8.2 Overview of the four core promoter elements, B recognition element (BRE), TATA box, initiator element (Inr), and downstream promoter element (DPE), showing their respective consensus sequences and their distance from the transcription start site. The Inr consensus sequence is shown for the model organism *Drosophila melanogaster* (Dm) as well as for humans (Hs).

A TATA box (consensus 5'-TATAAA-3') –about 25-35 base pairs upstream of the start of transcription (+1). (Note this is **not the same** as the sequence found in prokaryotes but is similarly A-T-rich to facilitate the opening of DNA and formation of transcription bubble)

Initiator (Inr) sequence located around nucleotide +1 (the TSS)

DPE (downstream promoter element): is a common component of RNA polymerase II promoters that do not contain a TATA box (TATA-less promoters).

8.4.2 Role of General Transcription Factors

A key difference between the initiation of transcription in *E. coli* and eukaryotes is that **eukaryotic polymerases do not directly recognize their core promoter sequences.**

The core promoter sequences described above are recognized by a set of proteins called **general (or basal) transcription factors. These proteins are not part of the RNA polymerase II complex.**

These general transcription factors are found in all eukaryotes, suggesting that the fundamentals of transcription are conserved in higher organisms.

These proteins are identified as TF_NX, where N is a roman numeral I, II, or III (signifying the polymerase) and X is a letter. Therefore TF-II – means Basal Transcription Factor for RNA Pol II.

Note:

The bulk of the work identifying general transcription factors as well as the order of assembly was done using genes

with promoters that have the TATA box and deduced **by in vitro experiments**.

Establishing the Pre-Initiation Complex

In order for transcription to occur RNA polymerase II needs to be recruited to the appropriate location- around the transcription start site, on the core promoter. The first steps in eukaryotic transcription involve the regulated assembly of the **general transcription factors (GTFs)**.

These proteins serve as a platform for RNA polymerase II recruitment.

The GTFs include the factors TFIIA, TFIIB, TFIID, TFIIIE, TFIIF, TFIIH, RNA polymerase (RNA pol II).

We will only focus on the functions of 2 GTFs:

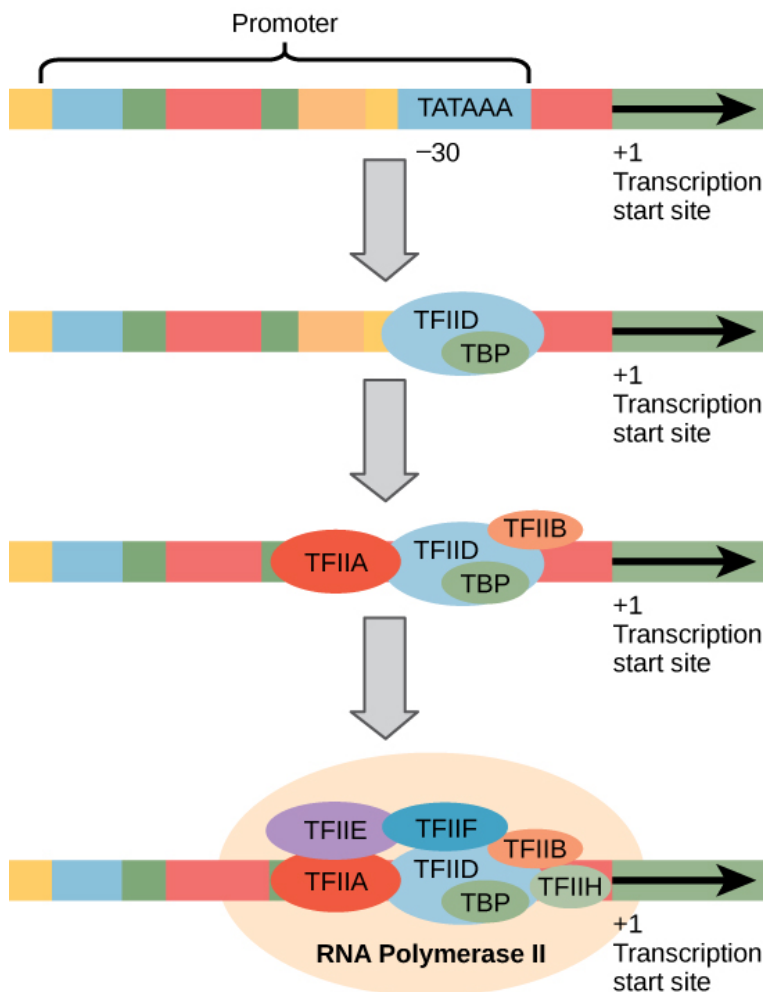
As the name “TATA-binding protein” suggests: TBP is a sequence-specific protein that binds to the TATA box. X-ray crystallography studies of TBP show that it has a saddle-like shape that wraps partially around the double helix. (Chasman DI, et al)

2. Binding of TFIID to the core promoter is followed by the recruitment of further GTFs and eventually RNA pol II.

The combination of all the GTFs along with RNA Pol II is the Pre-initiation Complex (PIC)

PIC first adopts an inactive state, the “closed” complex, which

is incompetent to initiate transcription. This complex is 'poised for transcription'.



Eukaryotic Transcription Initiation: A generalized promoter of a gene transcribed by RNA polymerase II is shown. Transcription factors recognize the promoter, RNA polymerase II then binds and forms the transcription initiation complex. Image Attribution: OpenStax College, Eukaryotic Transcription October 16, 2013. Provided by: OpenStax CNX. Located at: <https://cnx.org/resources/cf10220587e828cab2594cc6f5a229a6b66b92e2/>

[Figure_15_03_01.jpg](#) License: CC BY: Attribution

Abortive Initiation, Promoter Clearance, and Elongation

TFII-H plays a key role in the transition from 'closed to open complex'.

This protein has **2 activities**:

1) ATP-dependent helicase type activity- that opens up about 11 to 15 base pairs around the transcription start site leading to forming the transcriptional bubble.

Abortive transcription- once the RNA polymerase binds, it can begin to assemble a short stretch of RNA. This must be followed by promoter clearance, in order to move down the template and elongate the transcript.

2) TF II H- Kinase activity: adds phosphates onto the C-terminal domain (CTD) of the RNA polymerase II. (specifically certain amino acids within the CTD (C-terminal domain) of RNA polymerase II get phosphorylated)

[**Terminology alert**: Kinases are the name given to a class of enzymes that catalyze the transfer of a phosphate group from ATP to proteins. Commonly modified amino acids include Serines, Threonines, Tyrosines]

This phosphorylation appears to be the signal that releases the RNA polymerase from the basal transcription complex and allows it to move forward on the template, building the new RNA as it goes

After the departure of the polymerase, at least some of the GTFs detach from the core promoter,

8.4.3 Other Regulatory Sequences

Fundamentally, a key difference with bacterial transcription is that the pre-initiation complexes do not assemble efficiently and the basal rate of transcription initiation is therefore very low, regardless of how ‘strong’ the promoter is.

As was discussed in the outlining the structure of eukaryotic promoter, in most eukaryotic genes to achieve effective initiation, the formation of the complex must be **activated by additional proteins**.

Any protein that stimulates transcription initiation is called a **Transcriptional Activator**. These proteins bind either the Promoter Proximal Elements or Enhancer/Silencer sequences. This binding is sequence-specific.

Promoter Proximal Elements

Are several different consensus sequences to which *different regulatory* transcription factors can bind.

In different promoters, **transcription factor binding sites** are mixed and matched in different combinations. Each

promoter is regulated by a unique combination of transcription factors.

The binding of transcription factors to the consensus sequences in the regulatory promoter affects the assembly or stability of the basal transcription apparatus at the core promoter.

Example: Red blood cell development

An example of the former is the upstream element AACCAAT and its associated transcription factor, CP1. Another transcription factor, Sp1, is similarly common and binds to a consensus sequence of ACGCCC.

Both are used in the control of the beta-globin gene, along with more specific transcription factors, such as GATA-1, which binds a consensus AAGTATCACT and is primarily produced in blood cells.

CP1 is found in many types of cells. GATA-1 is present in only a few types of cells including red blood cells; therefore are thought to contribute to the cell-type specificity of β -globin gene expression

This illustrates another option found in eukaryotic control that is not found in prokaryotes: tissue-specific gene expression.

Response Elements

In addition, many genes have common regulatory elements called 'RESPONSE ELEMENTS'. These response elements

are binding sites for Transcriptional Activators and enable transcription initiation to respond to general signals from outside of the cell:

Examples:

- the cyclic AMP response module CRE (consensus 5'-WCGTCA-3'), recognized by the CREB activator
- heat-shock module (HSE) (consensus 5'-CTNGAATNTTCTAGA-3'), recognized by HSP70 and other activator
- steroid- hormone response element [Glucocorticoid Response Element, Estrogen Receptor Element]

For Biological examples of how this relates to Hormone Signaling please watch Lecture Videos associated with this module.

Enhancers and Silencers

Enhancers are regulatory elements that stimulate the transcription of distant genes. Silencers inhibit transcription.

Both regulate transcription over long distances in a position- and orientation-independent manner.

Enhancers are transcription activator binding sites grouped in units. *Multiple enhancers enable a gene to respond differently to different combinations of activators.*

This arrangement gives cells, in a developing organism, exquisitely fine control over their genes in different tissues or at different times!

The ‘looping of DNA’ between the enhancer sites and the core promoter region helps proteins bound to the enhancer to interact with the transcriptional apparatus.

Research Technique: Reporter Gene Assays are instrumental in identifying regulatory sequences in promoters of genes.

Concepts in Context

Watch the short film *The Making of the Fittest: Evolving Switches, Evolving Bodies*. Pay close

attention to how the switches regulate the expression of the *Pitx1* gene in stickleback embryos as well as the ‘Reporter Gene Assay’ used



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1216#oembed-1>

REMEMBER: Don't forget to complete the associated assignment in CANVAS.

8.4.4 Mediator Complex

Experimental studies of transcription in vitro showed that in addition to the GTFs, another multisubunit complex **mediate's** communication between activating TFs (at enhancer and upstream activator sequences) and the GTFs and RNA pol II, **hence the name “Mediator” for this complex.** ([Ref](#))

According to current gene activation models, the *Mediator complex* forms a **physical bridge** between proteins bound to distant regulatory regions and promoters, and transcription machinery at the core promoter.

The mediator is a huge complex of 25 to 30 subunits with a mass of more than 1-MDa.

A picture of transcription initiation that includes all the elements is shown in Figure 8.4 below

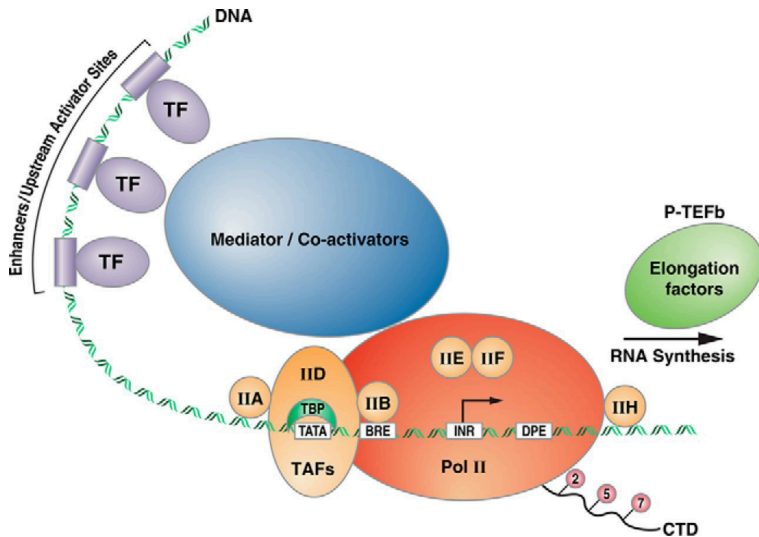


Figure 8. 4 Upstream activator sites and enhancers are bound by a variety of transcription factors, composed of DNA-binding domains (shown as cylinders on the DNA) and activation domains (shown as circles). These proteins serve to recruit co-activators, which can act on chromatin to facilitate transcription complex assembly (see below), or mediator, a large multisubunit complex that communicates with and is part of the core transcription machinery. Image from: [https://www.jbc.org/article/S0021-9258\(20\)36478-4/fulltext](https://www.jbc.org/article/S0021-9258(20)36478-4/fulltext) which is an Open Access article distributed under the terms of the Creative Commons CC-BY license.

8.5 How Transcription factors Work

Experiments using recombinant proteins showed that transcriptional activators are modular containing 2 domains.

Details of the experimental approach are presented in the lecture video associated with this module

[Recall the function of a protein domain from Chapter 1]

DNA binding domain: which contacts the regulatory sequences

The DNA binding domains fall into one of four representative families that are distinct structurally.

Activation domain: responsible for 'activation' or recruitment of transcriptional machinery.

Transcription factors often work as dimers.

In general, most regulatory transcription factors do not bind directly to the RNA polymerase

Ways in which Transcriptional Activators influence transcription include

1. Influence the PIC at the promoter directly via TF-II D or indirectly via the mediator
2. Influence the chromatin structure!

The main way in which this can be achieved as was discussed in Chapter 4 is

1. Covalent modification of histones
2. ATP-dependent chromatin remodeling.

Both of these activities are present in many transcription factors!

8.5 Bringing it all together

Overall the picture of transcription initiation then is less of an ON or OFF but that of fine-tuning.

Transcription initiation *in vivo* requires the presence of transcriptional activator proteins (coded by gene-specific transcription factors). These proteins bind to specific short sequences in DNA (enhancers).

A typical eucaryotic gene has many activator proteins, which together determine its rate and pattern of transcription. Sometimes acting from a distance of several thousand nucleotide pairs, these gene regulatory proteins help RNA polymerase, the general factors, and the mediator all assemble at the promoter.

In addition, activators attract ATP-dependent chromatin-remodeling complexes and histone acetylases.

Each individual gene transcription can be adjusted in amount based on the tissue type, developmental stage, biochemical condition. Factors like the number of sequences, types of enhancers, presence or absence of transcriptional

factors, co-activator, or repressors all dictate the final outcome of transcription initiation.

8.6 Relevant Biological Concepts

The mixing and match of these regulatory sequences is the principle behind 2 important features of eukaryotic transcriptional regulation

Coordinated Gene Regulation and Combinatorial Control

The presence of the **same response element** in different genes allows a **single stimulus** to activate multiple genes by virtue of binding to a single transcriptional regulator.

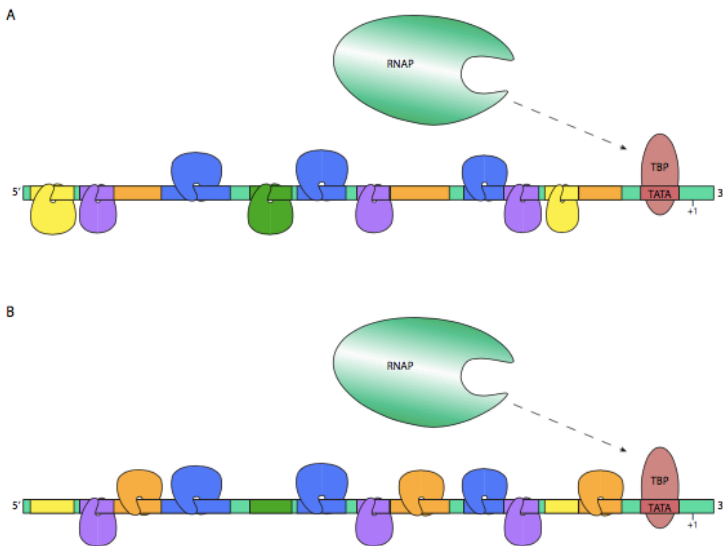
This phenomenon is also behind the succession of gene expression patterns during development, which results in establishing cell fate.

Similarly, the Transcription factors and other proteins that bind to regulatory sites on DNA provide RNA polymerase, access to specific genes. Therefore a given regulatory protein can have different effects, depending on **what other proteins are present in the same cell. This phenomenon is combinatorial control.**

An illustrative example is shown below. In the figure, the

transcription factors hanging downward are representative of inhibitory TFs, while those riding upright on the DNA are considered enhancers. Thus, the RNA polymerase in (A) has a lower probability of transcribing this gene, while the RNAP in (B) is more likely to, perhaps because the TF nearest the promoter interacts with the RNAP to stabilize its interactions with TFIID.

In this way, the same gene may be expressed in very different amounts and at different times depending on the transcription factors expressed in a particular cell type.



Molecular mechanisms that create and maintain specialized cell types depend on --Combinatorial gene control. Combinations of master transcription regulators specify cell types by controlling the expression of many genes.

This discovery is the basis behind **induced pluripotent stem (iPS) cells- the ability to take** specialized cells (like skin or fibroblasts) and reprogram them to become immature cells. The addition of four genes, encoding transcription factors can induce these cells to become **pluripotent stem cells, i.e. immature cells that are able to develop into all types of cells in the body.**

The Nobel Prize in Physiology or Medicine 2012 was awarded for this discovery and has implications broader biomedical implications- including creating organs within a lab for organ donation.

REMEMBER: To complete the reading associated with this within your module.

Learning Objectives: You should be able to:

1. Describe the three processes that commonly modify eukaryotic pre-mRNA.
2. Bacterial and eukaryotic gene transcripts can differ, in the transcripts themselves, in whether the transcripts are modified before translation, and in how the transcripts are modified. For each of these three areas of contrast, describe what the differences are.
3. What purposes do capping and poly-A tail addition serve for eukaryotic mRNAs?
4. Show the pathway for cap formation
5. Answer at what stage of mRNA formation is the cap added to the RNA molecule
6. Explain how Polyadenylation and transcription termination are linked.

8.7 Transcription Elongation and Termination- mRNA Processing

After initiation, the mechanics of transcription elongation are similar to that in Prokaryote, however, a big difference is the modification of the mRNA as it emerges from the RNA pol II enzyme.

The first modification occurs at the 5' end

5' end-capping.

Once the 5' end of a nascent RNA extends free of the RNAP II approximately 20-30 nt, it is ready to be capped by a 7-methylguanosine structure.

It consists of a guanine nucleotide connected to mRNA via **an unusual 5' to 5' triphosphate linkage** (Figure below). This guanosine is methylated on the 7 position directly after capping *in vivo* by a methyltransferase.

It is referred to as a **7-methylguanylate cap, abbreviated m⁷G.**

The process involves three steps.

First, RNA triphosphatase removes the 5'-terminal triphosphate group.

Second, Guanylation by GTP is catalyzed by a capping enzyme, forming an unusual 5'-5' "backward" bond between the new guanine and the first nucleotide of the RNA transcript.

Finally, guanine-7-methyltransferase methylates the newly attached guanine.

This 5' "cap" has many functions.

1. serves as a recognition site for transport of the completed mRNA out of the nucleus and into the cytoplasm
2. Prevention of degradation by exonucleases
3. Promotion of translation (see ribosome and translation)

Once the CAP is made it is recognized and bound by a complex of proteins (CAP Binding Protein -CBP) that remain associated with the cap till the mRNA has been transported into the cytoplasm.

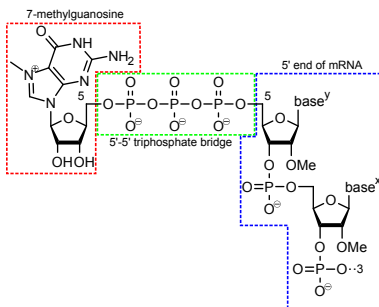


Figure 8.5 5' CAP Structure.
Image attribution: [Naturwiki, CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/)
<<https://creativecommons.org/licenses/by-sa/4.0/>>, via [Wikimedia Commons](https://creativecommons.org/licenses/by-sa/4.0/)

3' end Polyadenylation and termination

The 3' end **of the gene (within the 3'UTR)** is the signature sequence for signaling the end of transcription and polyadenylation.

It consists of a Poly A site flanked by a polyadenylation signal (AATAAA) and a downstream element that is GT-rich.

As the transcriptional machinery marches along the gene it will eventually transcribe this region generating the consensus AAUAAA sequence and the downstream element which will be a GU-rich sequence.

A protein complex called CPSF (the cleavage and polyadenylation specificity factor, CPSF) recognizes the poly-A signal. It has endonuclease activity and cuts the pre-mRNA between an AAUAAA consensus sequence and a GU-rich sequence, leaving the AAUAAA sequence on the pre-mRNA and a free 3' OH.

Note: This effectively releases the mRNA from the transcribing machinery!

An enzyme called **poly-A polymerase** then adds a string of approximately 200 Adenine residues, called the **poly-A tail**.

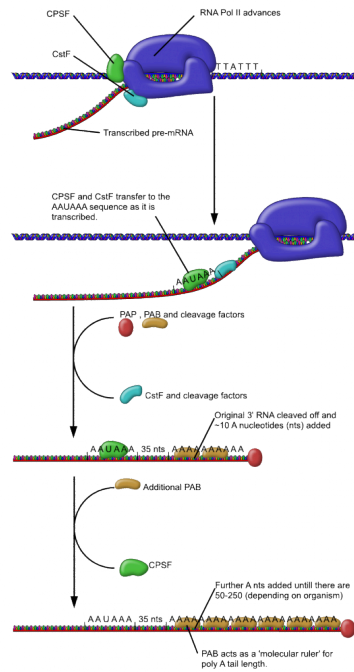


Figure 8. 6 Process of Polyadenylation. Image Attribution: Zephyris (en Wikipedia user), CC BY-SA 3.0 <<http://creativecommons.org/licenses/by-sa/3.0/>>, via Wikimedia Commons

Thus the Poly A tail is NOT a part of coded information of the gene but added post-transcriptionally.

Evidence suggests that the Poly A tail influences the efficiency of translation. The poly-A tail also has an effect on the stability of transcripts in the cytoplasm.

Splicing

The third and most complicated modification to newly-transcribed eukaryotic RNA is *splicing*. Splicing is the process by which the non-coding regions, known as *introns*, are removed, and the coding regions, known as *exons*, are connected together. We will be discussing the mechanism of splicing separately, although it is useful to introduce it here because splicing is also occurring during transcription!

Role of CTD of RNA Polymerase in mRNA processing

That transcription and mRNA processing are coupled is highlighted by the fact that proteins utilized for capping, splicing, and polyadenylations **are recruited to the CTD of RNA pol II!**

Recall that the CTD consists of multiple repeats of a special heptameric (Hepta-7) amino acid sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser. In particular, the Serines may be phosphorylated

in the various repeats. This occurs in a sequential manner and creates a signature (like a code) for many of the processing proteins to bind to the tail!

The RNA Pol-II enzyme physically carries the processing enzymes with it- and they get deployed as needed!

References and Attributions

This chapter contains material taken from the following CC-licensed content. Changes include rewording, removing paragraphs and replacing with original material, and combining material from the sources.

1. Bergtrom, Gerald, “Cell and Molecular Biology 4e: What We Know and How We Found Out” (2020). *Cell and Molecular Biology 4e: What We Know and How We Found Out – All Versions*. 13.

https://dc.uwm.edu/biosci_facbooks_bergtrom/13

2. Works contributed to LibreTexts by Kevin Ahern and Indira Rajagopal. LibreTexts content is licensed by CC BY-NC-SA 3.0. The entire textbook is available for free from the authors at <http://biochem.science.oregonstate.edu/content/biochemistry-free-and-easy>

3. Flatt, P.M. (2019) Biochemistry – Defining Life at the Molecular Level. Published by Western Oregon University, Monmouth, OR (CC BY-NC-SA). Available at:

https://wou.edu/chemistry/courses/online-chemistry-textbooks/ch450-and-ch451-biochemistry-defining-life-at-the-molecular-level/?preview_id=4919&preview_nonce=cca8f0ce36&preview=true

4. “Eukaryotic Transcriptional Regulation” by E. V. Wong, [LibreTexts](#) is licensed under [CC BY-NC-SA](#)

Other References

Chasman DI, Flaherty KM, Sharp PA, Kornberg RD. Crystal structure of yeast TATA-binding protein and model for interaction with DNA. *Proc. Natl Acad. Sci. USA.* (1993);90:8174–8178. [[PMC free article](#)]

9.

MRNA SPLICING AND ALTERNATIVE SPLICING

Learning Objectives

- Know the similarities and differences that exist between pre-mRNA and mRNA.
- Know/draw/label the steps involved in spliceosome formation.
- Describe the roles of snRNPs in splicing
- Explain what are the benefits of RNA processing.
- What is transesterification, and what transesterification reactions are needed to splice introns?
- Explain why fidelity of splicing is important?
- Explain what controls the fidelity of splicing.
- Predict how mutations at the 5' splice site, 3' splice site, and branch point might disrupt splicing and alter

the phenotype.

- Describe the different forms/patterns of alternate splicing.
 - What does constitutive exon mean?
 - Draw splicing diagrams to show alternative splicing patterns.
 - Identify when given a description/diagram of an alternately spliced gene which type it is.
- How is alternative splicing regulated?
- What are exonic and intronic splicing silencers or enhancer sequences? (ESS, ISS, ESE, and ISE)
- Using examples explain how anti-sense technology is used to correct for splicing defects.

CONNECTING CONCEPTS

Chapters 8 and 9 introduced you to different components of eukaryotic gene structure and RNA molecules transcribed. These include promoters, 5' and 3' untranslated regions (UTR), coding sequences (exons), introns, 5' caps, Poly A signal, and poly(A) tails.

You should be able to draw /identify/annotate when given a gene sequence the elements of the gene above.

You should be able to draw the pre-mRNA and mRNA derived from it.

9.1 Introduction

The immediate product of RNA polymerase II is sometimes referred to as **pre-mRNA** or the **primary transcript**.

As we saw in Chapter 8, the initial products of transcription are further processed acquiring a cap at their 5' end and poly-A tail at their 3' end. Most importantly nearly ALL mRNA precursors are **spliced**

Splicing is the process by which the **non-coding regions**, known as *introns*, are removed, and the coding regions, known as *exons*, are connected together.

Introns were initially thought to be entirely a feature of the eukaryotic genome. In recent years the presence of intron-containing genes has been documented in archaea, bacteriophages, and even some bacteria. They are also present within mitochondrial and chloroplast genes. **These introns are called Group I and II introns and are *self-splicing*!**

That means for those RNAs, splicing happens autonomously, with part of the RNA acting as an enzymatic catalyst for the process. This requires that the RNA have a specific secondary and tertiary structure, bringing the two exons close together while looping out the intron. It was the study of this phenomenon that led to the discovery of ribozymes, which are enzymes made of RNA.

Until the discovery of ribozymes, it had been assumed that

enzymes could only be generated with the diversity of structures possible with the amino acids in proteins.

The splicing process we will study however is **carried out within the nucleus on mRNA** using a multi-subunit protein complex known as the **Spliceosome**.

Spliceosome-mediated splicing IS a unique feature of eukaryotes.

9.2 The Split Gene

Discovery of Interrupted Genes

The discovery of eukaryotic split genes with introns and exons came as quite a surprise.

Go to the [DNA Learning Center website](#) and click on the **Interactive Animation** that outlines the experiments that led to the discovery that eukaryotic genes have non-coding regions.

For their discovery of split genes, Richard J. Roberts and Phillip A. Sharp shared the Nobel Prize for Physiology in 1993.

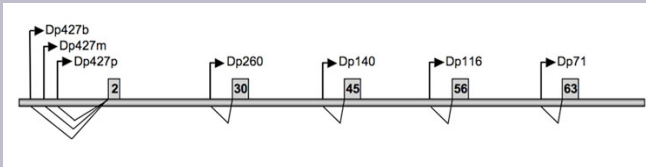
In fact, all but a few eukaryotic genes are split, and some have one, two (**or more than 30-50!**) **introns** separating bits of coding DNA, the **exons**. The size of the intron (# of base pairs) and number of introns may vary.

An extreme example of splicing and of medical relevance is the human dystrophin gene.

Human Dystrophin Gene- Muscular Dystrophy

The *DMD* gene is one of the largest known gene in humans, **spanning 2.6 million base pairs (bp)** consisting of almost 0.1% of the human genome or about **1.5% of the entire X chromosome**. This very large gene is highly fragmented into 79 exons and introns of variable size ranging from 107 bp (intron 14) to >200kb (intron 44).

It is transcribed in a 14 kb mRNA, and the **11kb cDNA** encodes a 3685 amino acid protein of 427 kDa called Dystrophin.



Dystrophin Gene Structure. From “[TREAT-NMD DMD Global database](#)”

Dystrophin is located primarily in muscles used for movement (skeletal muscles) and in heart (cardiac) muscle. Small amounts of dystrophin are present in nerve cells in the brain.

Many different types of mutations have been described for DMD including large deletions and duplications, point mutations, and small rearrangements that underlie various forms of Muscular Dystrophy.

9.2.1 So, Why Splicing?

While the dystrophin gene is an extreme example, overall in humans

- a) Median Gene is about 23, 000 bp, 7 introns
- b) Median Intron is about 1800 bp in size and

c) Median Exon is only about 123 bp in size!

We can see that 90 -95% of the RNA transcript that gets synthesized is essentially thrown out! So why do higher organisms have split genes in the first place?

While the following discussion can apply to all splicing, it will reference mainly **spliceosomal introns**. Here are some answers to the question “Why splicing?”

1) Introns in nuclear genes are typically longer (often much longer!) than exons. Since they are non-coding, they are large targets for mutation. In effect, noncoding DNA, including **introns can buffer the ill effects of random mutations**.

2) Gene duplication on one chromosome (and loss of a copy from its homolog) arise from unequal recombination (non-homologous crossing over). It occurs when similar DNA sequences align during synapsis of meiosis. Unequal recombination can also occur between similar sequences (e.g., in introns) in the same or different genes, resulting in a sharing of exons between genes. After unequal recombination between introns flanking an exon, one gene will acquire another exon while the other will lose it. The gene with the extra exon may produce the same protein, but one with a new structural domain and function. Like a complete duplicate gene, one with a new exon that adds a new function to an old gene has been entered in the pool of selectable DNA. Thus, this phenomenon of **exon shuffling** increases species diversity!

Exon shuffling has occurred, creating proteins with different

overall functions that nonetheless share at least one domain and one common function.

3) Presences of introns allow for increasing protein diversity. As shown in Figure 9.1 one gene with 5 exons can produce at least 3 different isoforms as illustrated, depending on which exons are joined together.

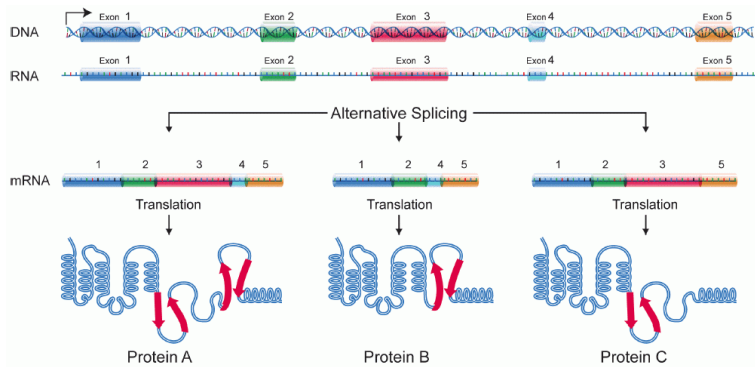


Figure 9.1. In this diagram a gene with 5 exons produce three protein isoforms. Protein A includes all of the exons, whereas Proteins B and C do not.

9.3 A Detailed Look at mRNA Splicing

Splicing must be exquisitely sensitive because the nucleotide information is converted into protein information in three non-overlapping nucleotide sequences called codons ([See Themes from Inro Bio](#) and Chapter on Translation). Thus, a

one-nucleotide shift in the course of splicing would alter the nucleotide information on the processed mRNA resulting in inaccurate coding sequence and protein.

Therefore the splicing machinery **must be able to recognize splice junctions** (i.e., where each exon ends and its associated intron begins) in order to correctly cut out the introns and join the exons to make the mature, spliced mRNA.

9.3.1 Finding the Intron/Exon junctions

Whether it is self-spliced or using the spliceosome, the junctions between exons and introns are indicated by specific base sequences and guide the splicing process (Figure 8.4.8" > 9.2) called SPLICE SITES.

They are named for their positions relative to the intron. These include

1. A **5' splice site** with the consensus sequence AG|GURAGU.

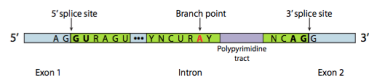


Figure 9.2 Consensus sequences for splicing.

In this sequence, the intron starts with the second G (R stands for any purine)

2. A **3' splice site** that starts with an 11-nucleotide polypyrimidine tract followed by NCAG|G.

...and somewhere in between the two,

3. A branchpoint **adenine**, typically within a YNCURAY sequence (Y is a pyrimidine, N is any nucleotide, R is a purine)

The branch point Adenine is ‘invariant’- meaning it is always needed and present in introns. The importance of this site will be seen when we consider the steps of splicing.

Notice that the INTRON is defined by a **GU** at its 5' end and an **AG** at its 3' end. Splicing occurs utilizing sets of GU-AG marking the boundary of the intron at either end where the 'cut/ splice' needs to occur.

This is often referred to as the GU-AG rule: (originally called the GT-AG rule in terms of DNA sequence) describing the requirement for these constant dinucleotides at the first two and last two positions of introns in pre-mRNAs.

[Lecture Video: Discovery of Split Gene, Why Splicing, Finding Introns/Exon Junctions](#)

9.3.2 The Chemistry of Splicing: Transesterification

There are two main steps in splicing.

STEP 1: The nucleophilic attack by the **2'OH of the**

branch point A on the 5' splice site (the junction of the 5' exon and the intron), releasing the 5' exon with a free 3' hydroxyl group. **A looped lariat-shaped molecule composed of the 5' end of the intron connected to the branchpoint via a 2',5'-phosphodiester bond.**

What is
Transesterification
?

Transesterification is a chemical reaction of an **alcohol** with an ester to form a different alcohol and a different ester.

We can now see why the **BRANCH point A** is crucial for the splicing to occur.

As a result of a transesterification reaction, the 5' exon is released, **and a lariat-shaped molecule composed of the 3' exon and the intron sequence is generated (Figure 9.3).** The lariat 5' end of the intron is connected to the branchpoint **via a 2',5'-phosphodiester bond.**

STEP 2: The 3' OH of the 5' exon which was released in step 1, attacks the phosphate at the 3' splice junction, connecting exons 1 and 2. The lariat intermediate is released.

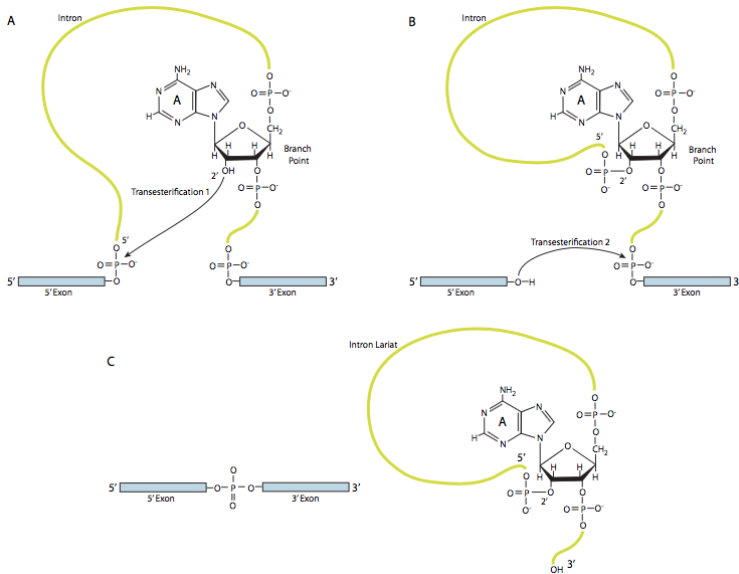


Figure 9.3 (A) Transesterification reaction 1 connects the 2'-OH of the branch- point ribose to the 5'-phos- phosphate of the 3' end of exon 1. (B) Transesterification 2 is an attack by the remaining -OH on the 3' end of exon 1 on the 5' phosphate of exon 2. (C) This simultaneously releases the lariat and connects the exons. Image from: ["Post-Transcriptional Processing of RNA" by E. V. Wong, LibreTexts is licensed under CC BY-NC-SA.](#)

Lecture Video: Chemistry of Splicing

9.3.3 The Spliceosome

More than 300 proteins and a group of special RNAs come assemble on the pre-mRNA to form the machine called the ‘Spliceosome’ that controls mRNA splicing.

The components of this machine include **small nuclear ribonucleoprotein** or **snRNPs** (pronounced “snurps”) for short.

As the name suggests, snRNPs contain **proteins and a small nuclear RNA (snRNA) component**. The snRNA’s are designated as **U1, U2, U4, U5, and U6**. These are **100-200 nucleotides long and adopt elaborate tertiary structures**.

They associate with many proteins and together form the snRNP. The snRNPs are named after the RNA component, for example, the U1 snRNP consists of many proteins along with a U1 snRNA component.

There are therefore the U1 snRNP, U2 snRNP, U4 snRNP, U5 snRNP, and U6 snRNP. The snRNPs recognize the conserved sequences within introns and quickly bind these sequences once the pre-mRNA is made and initiate splicing.

It is the RNA component that base pairs with the consensus site and brings the snRNP to the accurate location.

Although many details remain to be worked out, it appears that components of the splicing machinery associate with the

CTD of the RNA polymerase and that this association is important for efficient splicing.

Other proteins that play a role include U2AF (U2-associated factor, which binds to the polypyrimidine tract, and branch-point protein BPP, which binds to consensus sequence near the branchpoint.

There are also a variety of other less-studied splicing factors from the SR protein family (C-terminal Serine-Arginine binding motif) and the hnRNP (heterogeneous nuclear ribonucleoprotein) families that act to recruit the primary members of the spliceosome to their proper locations (see Alternative splicing regulation)

The spliceosome is not ‘pre-assembled’ but builds on the mRNA in a step-wise fashion as the mRNA emerges from the RNA polymerase.

Key Takeaways

Nuclear pre-mRNA introns contain three consensus sequences critical to splicing: a 5' splice site, a 3' splice site, and a branch point. The splicing of pre-mRNA takes place within a large complex called the spliceosome, which consists of snRNAs and proteins.

Practice Questions



An interactive HSP element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1292#h5p-48>

9.3.4 Steps in splicing

We will illustrate the assembly of the spliceosome using an example transcript. In this example, the pre-mRNA contains two exons and one intron. (**Figure 9. 4**)

First, the **U1 snRNP** binds the 5' splice site. Additional proteins such as U2AF (AF = associated factor) are also loaded onto the pre-mRNA near the branch site.

Next, the U2 snRNP binds to the **consensus site around the branchpoint**, but importantly, there is no base-pairing to

branchpoint A itself. Instead, due to base pairing of U2 with the surrounding sequence, the branch point A is forced to bulge out from the rest of the RNA in that region.

Next, a complex of the U4/U6 and U5 snRNPs is recruited to the spliceosome to generate a pre-catalytic complex. This com-

plex undergoes rearrangements that alter RNA-RNA and protein-RNA interactions, resulting in the displacement of the U4 and U1 snRNPs and the formation of the **catalytically active spliceosome**.

This complex is then responsible for executing the 2 transesterification steps: First, the 5' end of the intron is cut. The 5' GU end of the intron is then connected to the A branch site, which creates a lariat structure. This releases the 5' exon (and the whole 5' half of the RNA for that matter), but it is kept in close proximity to the 3' exon (and the rest of the RNA) by U5, which attaches to both exons.

This allows the second transesterification to take place, in which the 3'-OH of the first exon attacks the 5' phosphate at the beginning of the second exon, thus simultaneously breaking the bond between the intron and the second exon, and also connecting the two exons via a conventional 3',5'-phosphodiester bond.

After the second catalytic step, the intron in the form of a lariat is released along with U2, U5, and U6 snRNPs.

The intron will be degraded and the snRNPs are used again

to splice other pre-mRNAs. The mature mRNA transcript is now ready to be exported to the cytoplasm for translation.

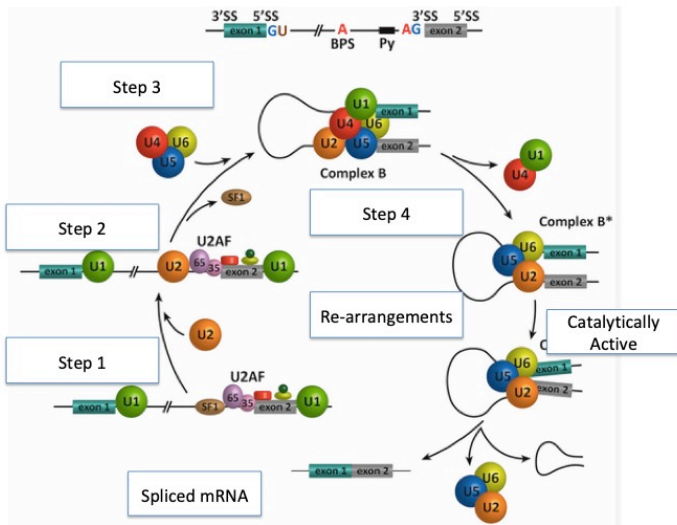


Figure 9. 4 Schematic representation of the spliceosome assembly and pre-mRNA splicing. In the first step of the splicing process, the 5' splice site (GU, 5' SS) is bound by the U1 snRNP, and the splicing factors SF1/BBP and U2AF cooperatively recognize the branch point sequence (BPS), the polypyrimidine (Py) tract, and the 3' splice site (AG, 3' SS). The binding of the U2 snRNP to the BPS results in the pre-spliceosomal complex A. Subsequent steps lead to the binding of the U4/U5–U6 tri-snRNP and the formation of catalytically active complex, which is responsible for the two transesterification reactions at the SS. Additional rearrangements result in the excision of the intron, which is removed as a lariat RNA, and ligation of the exons. The U2, U5, and U6 snRNPs are then released from the complex and recycled for subsequent rounds of splicing. *Figure modified from: Suñé-Pou, M., et. al. (2017) Genes 8(3):87 (Open Access)*

Once splicing is complete proteins called **Exon Junction Complexes (EJC)** are deposited marking the previous boundaries.

Along with the CAP-Binding Protein at the 5' end, the poly A tail binding protein at the 3' end the presence of these protein markers indicates a processed and mature mRNA transcript that is ready for export out of the nucleus into the cytoplasm where it will be translated into protein.

[**See a Step-By-Step Animation of how Introns are removed at this website**](#)

[Lecture Video: Mechanism of Splicing](#)

9.4 Alternative Splicing

Splicing is an efficient (with respect to genome size) way to generate protein diversity. In alternative splicing, some potential introns may be spliced out under certain

circumstances but remain as coding sequence under other circumstances.

Alternative Splicing (AS) thus offers an additional mechanism for regulating protein production and function.

Recall that the splice sites are recognized by base-pairing and therefore, there can be stronger and weaker splice sites depending on how close they are to the consensus and the complementary sequence on the snRNPs. Therefore, a gene with several potential introns may have all introns spliced out 80% of the time, but the other 20% of the time, perhaps only one or two introns are spliced out.

Adding variability, there are splicing factors that may bind near splice sites and can either make them more easily recognizable, or nearly hidden (see **Alternative Splicing Regulation below.**)

A classic example of alternative splicing is the gene encoding α -tropomyosin.8.4.11">By splicing in/out different combinations of exons, a single gene can generate seven different proteins, depending on the tissue type (**Figure 9.5**)

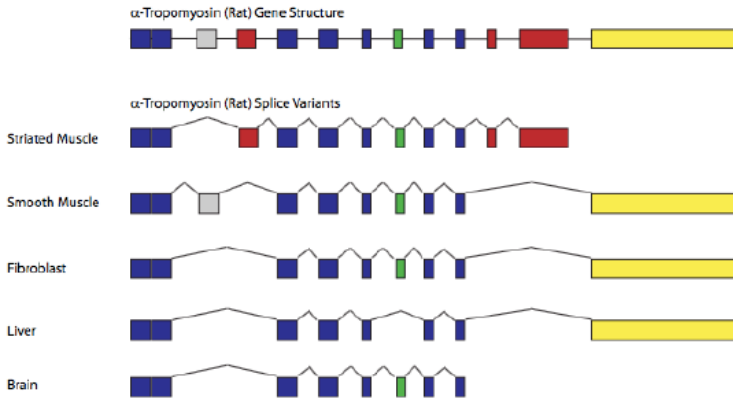


Figure 9.5 Alternative splicing of the α -tropomyosin gene leads to different forms of the mRNA and protein in different cell types.

9.4.1 Alternative Splicing Terminology

Splice forms /Splice variants/ Isoforms: Refers to all the different ways in which a pre-mRNA can be spliced to generate different forms of mature mRNA

CONSTITUTIVE EXON: Those exons that are always included in all splice forms

REGULATED EXONS: Those that are sometimes included, sometimes not.

***** In splicing diagrams the exons are drawn as boxes or cylinders, the introns are represented as lines between them. The splice patterns are indicated using lines (solid or dotted) showing which exons are connected *****

9.4.2 Alternative Splicing Patterns

Exon skipping: which is the major AS event in higher eukaryotes. In this type of event, the exon and the intron flanking on either side is removed from the pre-mRNA (Fig. 9.6 a).

Alternative 3' and 5' SS selection (Fig. 9.6 b & c): Occurs when the spliceosome recognizes *two or more (an alternative) splice sites at one end of an exon.*

In the diagram below starting from the left moving into the mRNA, the spliceosome uses the GU-AG rule. In 3' SS selection the spliceosome pairs the 5' SS with an *alternative 3'SS*. Conversely, an alternative 5' SS is used to pair with a normal 3'SS.

Use of alternative 3'SS and 5' SS also results in EXON EXTENSION.

Intron retention (Fig. 9.6 d), in which an intron remains

in the mature mRNA transcript. This AS event is much more common in plants, fungi, and protozoa than vertebrates.

Other events that affect the transcript isoform outcome include **mutually exclusive exons** (Fig. 10.26 e), **alternative promoter usage** (Fig. 9.6 f), and **alternative polyadenylation** (Fig. 9.6 g).

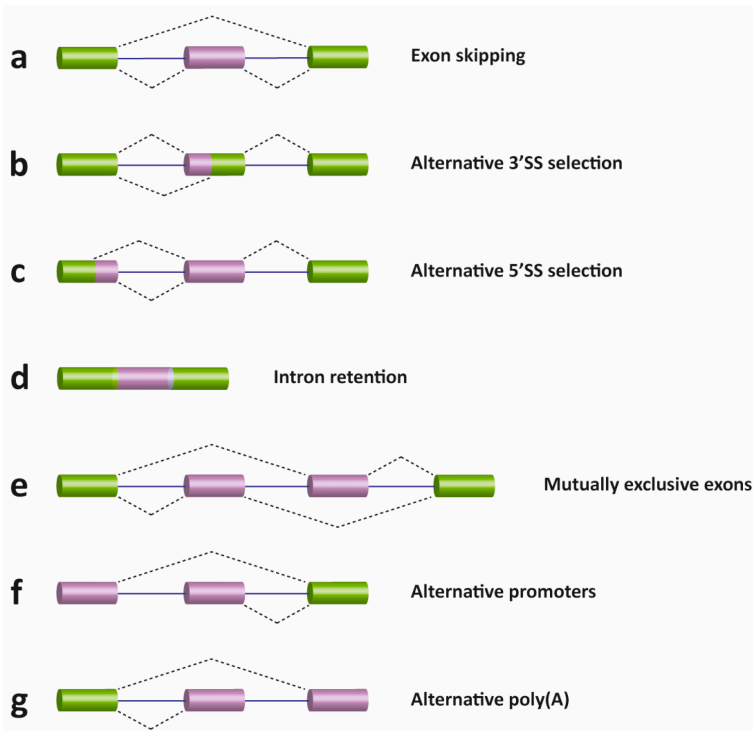


Figure 9.6 Schematic representation of different types of alternative transcriptional or splicing events, with exons (boxes) and introns (lines). Constitutive exons are shown in green and alternatively spliced exons in purple.

Dashed lines indicate the AS event. Exon skipping (**a**); alternative 3' (**b**) and 5' SS selection (**c**); intron retention (**d**); mutually exclusive exons (**e**); alternative promoter usage (**f**); and alternative polyadenylation (**g**) events are shown. Like alternative splicing (AS), usage of alternative promoter and polyadenylation sites allow a single gene to encode multiple mRNA transcripts.

Figure from: [Suñé-Pou, M., et. al. \(2017\) Genes 8\(3\):87](#)

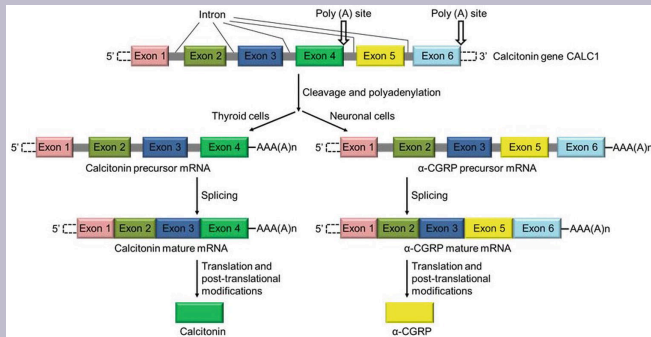
Examples: Alternative Polyadenylation

The CT/GRP gene is an example of alternative splicing using alternative poly(A) site. The gene produces the same pre-mRNA transcript in many cells, including thyroid cells and neuronal cells. The transcript contains six exons and five introns and includes two alternative polyadenylation sites, one in exon 4 and the other following exon 6.

In thyroid C cells, the gene product contains exon 4. The inclusion of Exon 4 facilitates utilization of the Poly(A) site and cleavage of mRNA transcript and termination of transcription. Translation of this

mRNA forms calcitonin, a hormone that regulates calcium in the body.

In sensory neuronal cells the same pre-mRNA is spliced to include exons 1,2,3 5 and 6, skipping exon 4. Therefore Poly adenylation takes place at the site in Exon 6. The mature mRNA when translated forms α -CGRP a regulatory neuropeptide.



Schematic diagram showing the synthesis of calcitonin and α -CGRP from a common gene CALC-I. Calcitonin gene CALC-I undergoes alternative splicing and forms protein calcitonin in thyroid C cells, and α -CGRP in the sensory neurons. Figure from: [Kumar A, Potts JD and DiPette DJ \(2019\) Protective Role of \$\alpha\$ -Calcitonin Gene-Related Peptide in Cardiovascular Diseases. *Front. Physiol.* 10:821. doi: 10.3389/fphys.2019.00821.](#) (An open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#).)

9.4.3 Alternative Splicing Regulation

Alternative splicing is not a random process. Instead, it's typically controlled by regulatory proteins. The proteins bind to specific sites on the pre-mRNA and “tell” the splicing factors which exons should be used.

These sites are sequence elements within the mRNA, known as *exonic* and *intronic* splicing silencers or enhancers (ESS, ISS, ESE, and ISE, respectively), participate in the regulation of alternative splicing.

Specific RNA-binding proteins, including heterogeneous nuclear ribonucleoproteins (hnRNPs) and **serine/arginine-rich** (SR) proteins, recognize these sequences to positively or negatively regulate alternative splicing (**Figure 9.7**).

In general, SR proteins bound to enhancers facilitate exon definition, and hnRNPs inhibit this process. These *trans-acting* elements are **expressed differentially within different locations or under different environmental stimuli to regulate alternative splicing**.

These regulators, together with an ever-increasing number of additional auxiliary factors, provide the basis for the specificity of this pre-mRNA processing event in different cellular locations within the body.

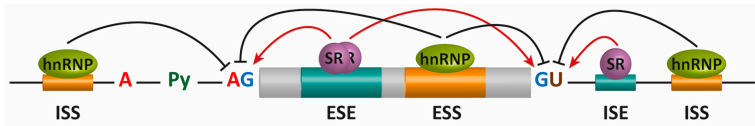


Figure 9.7 Alternative splicing (AS) regulation by *cis* mRNA elements and *trans-acting* factors. The core *cis* sequence elements that define the exon/intron boundaries (5' and 3' splice sites (SS), GU-AG, polypyrimidine (Py) tract, and branch point sequence (BPS)) are poorly conserved. Additional enhancer and silencer elements in exons and in introns (ESE: exonic splicing enhancers; ESI: exonic splicing silencers; ISE: intronic splicing enhancers; ISI: intronic splicing silencers) contribute to the specificity of AS regulation. *Figure from: Suñé-Pou, M., et. al. (2017) Genes 8(3):87*

[Lecture Video: Alternative Splicing](#)

9.5 Clinical Insight

Many human genetic diseases arise from mutations that affect pre-mRNA splicing; **indeed, about 15% of single-base substitutions that result in human genetic diseases alter pre-mRNA splicing.** Some of these mutations interfere with recognition of the normal 5' and 3' splice sites. Others arise

from mutations in splicing factors that can also cause defective mRNA splicing.

Listed below are some common diseases and the associated splicing defect.

Disease	Mutation	Splicing Effect
Familial dysautonomia (FD)	T > C mutation at position 6 of intron 20 of the IKBKAP gene	Exon skipping; introduction of a premature termination codon (PTC)
Spinal muscular atrophy (SMA)	C > T mutation at position 6 of exon 7 of the SMN2 gene	Alteration of a putative ESE
Hutchinson-Gilford progeria syndrome (HGPS)	c1824C > T mutation in exon 11 of LMNA gene	Activation of a cryptic splice site
Duchenne muscular dystrophy (DMD)	T > A mutation in exon 31 of the Dystrophin gene	Creation of a PTC and introduction of ESS
Amyotrophic lateral sclerosis (ALS)	Mutations in TDP-43	Altered gene splicing
Autosomal dominant retinitis pigmentosa (RP)	Mutations in genes of the core spliceosome (PRPF31, PRPF8, PRPF3, RP9)	Disruption of basal spliceosome function

9.5.1 Antisense Oligonucleotide Therapy

Concepts in Context: Molecular Biology in the News

In December 2016 the FDA approved Spinraza, the first FDA-approved therapy for all ages and types of SMA- Spinal Muscular Atrophy. SMA affects approximately 1 in 11,000 births in the U.S., and about 1 in every 50 Americans is a genetic carrier. SMA can affect any race or gender.

This discovery came after years of basic research to understand splicing and the mechanisms behind it. It was conceived and tested over several years in mouse models of SMA by Professor [Adrian Krainer](#), Ph.D., and his colleagues at Cold Spring Harbor Laboratory (CSHL)

Watch this video on the mechanism of action of Spinraza

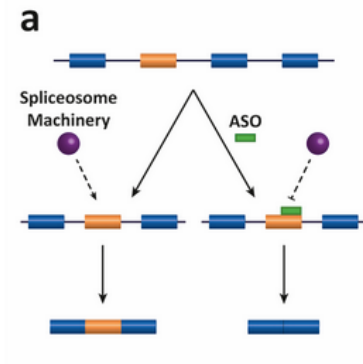


One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1292#oembed-1>

Spinraza is one of a growing list of gene therapies using **an antisense oligonucleotide (ASO) for modifying and fixing errors in the splicing process.**

Antisense drugs are small snippets of synthetic genetic material that bind to ribonucleic acid (RNA), so they can be used to fix the splicing of genes like *SMN2*.



The diagram depicts the antisense oligonucleotide (ASO)-based strategy for altering splicing patterns. Here the antisense oligos target an alternatively spliced exon (in orange). In

the absence of the ASO, the spliceosome is assembled and the exon is included in the mRNA; in the presence of the ASO, the spliceosome is sterically blocked and the exon is skipped and not included in the mRNA.

Before you continue

- Complete the Lecture Quick checks associated with this unit.
 - Complete the associated Concepts in Context Assignment.
-

References and Attributions

This chapter contains material taken from the following CC-licensed content. Changes include rewording, removing paragraphs and replacing with original material, and combining material from the sources.

1. Bergtrom, Gerald, “Cell and Molecular Biology 4e: What We Know and How We Found Out” (2020). *Cell and Molecular Biology 4e: What We Know and How We Found Out – All Versions*. 13.

https://dc.uwm.edu/biosci_facbooks_bergtrom/13

2. Works contributed to LibreTexts by Kevin Ahern and Indira Rajagopal. LibreTexts content is licensed by CC BY-NC-SA 3.0. The entire textbook is available for free from the authors at <http://biochem.science.oregonstate.edu/content/biochemistry-free-and-easy>

3. Flatt, P.M. (2019) Biochemistry – Defining Life at the Molecular Level. Published by Western Oregon University, Monmouth, OR (CC BY-NC-SA). Available at: https://wou.edu/chemistry/courses/online-chemistry-textbooks/ch450-and-ch451-biochemistry-defining-life-at-the-molecular-level/?preview_id=4919&preview_nonce=cca8f0ce36&preview=true

Other References

Suñé-Pou, M., Prieto-Sánchez, Boyero-Corral, S., Moreno-Castro, C., El Yousfi, Y., Suñé-Negre, J.M., Hernández-Munain, C., and Suñé, C. (2017) Targeting splicing in the

treatment of human disease. *Genes* 8(3):87. Available at:
<https://www.mdpi.com/2073-4425/8/3/87/htm>

10.

GENETIC CODE AND TRANSLATION

10.1 Overview of Translation

Within this chapter, we will cover the details of prokaryotic and eukaryotic translation. Translation is the process of converting the information housed in mRNA into the protein sequence. Essentially, you are translating the language of nucleotides into the language of amino acids.

Amino acids are linearly strung together via covalent bonds (called peptide bonds) between amino and carboxyl termini of adjacent amino acids. The sequential polymerization of amino acids, in a strict order determined by the sequence of an mRNA, is catalyzed by a ribonucleoprotein complex called the **ribosome** working with decoding “keys” termed **charged tRNAs**.

Recall that prokaryotic and eukaryotic transcription and translation systems differ in large part due to the compartmentalization of larger eukaryotic cells. Due to this compartmentalization, transcription and translation are

separated spatially and temporally within the cell. Transcription occurs within the nucleus of eukaryotes and translation occurs within the cytoplasm (Fig. 10.1 B). Prokaryotes do not have compartmentalization and have, thus, evolved a coupled transcription/translation system where both process occur simultaneously (Fig. 10.1 A).

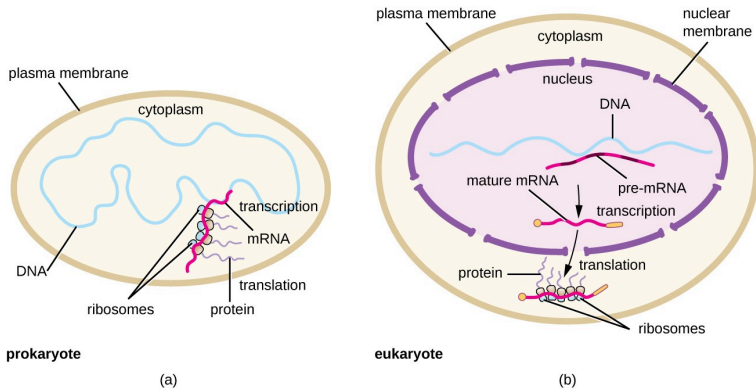


Figure 10.1 Cellular Location of Transcription and Translation in Prokaryotes and Eukaryotes. (a) Prokaryotes lack cellular compartmentalization and show coupled transcription-translation processing, whereas (b) eukaryotes have a high degree of compartmentalization and separate the processes of transcription, which is in the nucleus of the cell, from the processes of translation, which is localized in the

cytoplasm. Figure from: [Baccei, A., and Rice, M. Lumen Learning](#)

Recall that peptide formation is a dehydration reaction that combines the carboxylic acid of the upstream amino acid with the amine functional group of the downstream amino acid to form an amide linkage (Fig. 10.2). Water is the by-product. The ribosome (a large complex of peptides and rRNA molecules) serves as the enzyme that mediates this reaction. It requires a mature mRNA to serve as the template, and performs peptide bond synthesis in a directional fashion from the N to the C-terminal of the growing peptide/protein.

This is known as *N- to C-synthesis*.

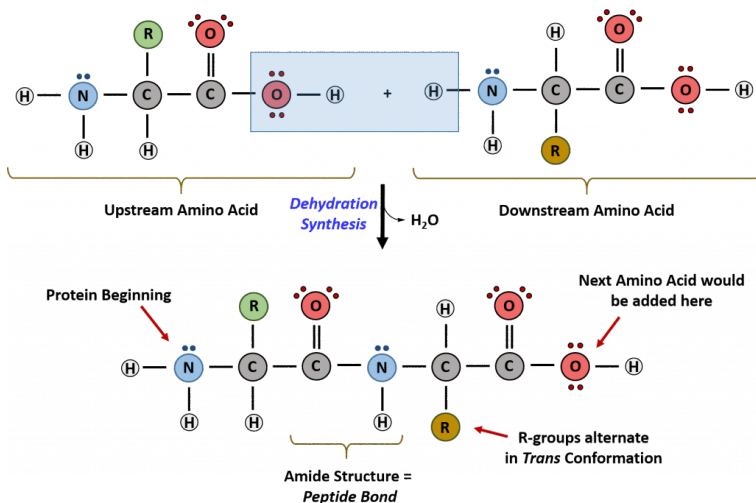


Figure 10.2 Formation of the Peptide Bond. The addition of two amino acids to form a peptide requires dehydration synthesis. The carboxylic acid of the upstream amino acid is joined with the amine functional group of the downstream amino acid to form the amide linkage. Within the ribosome, this reaction is highly directional and only occurs in the N to C orientation. Figure from: Flatt, P.M. (2019) Biochemistry – Defining Life at the Molecular Level. Published by Western Oregon University, Monmouth, OR (CC BY-NC-SA). Available at: <https://wou.edu/chemistry/courses/online-chemistry-textbooks/ch450-and-ch451-biochemistry-defining-life-at-the-molecular-level/chapter-11-translation/>

Learning Objectives

- Explain what the terms nonoverlapping, commaless (without punctuation), degenerate, and unambiguous mean with respect to the genetic code.
- Define the following terms as they apply to the genetic code: Reading frame, Initiation codon , Termination codon , Sense Codon
- What is the wobble hypothesis and how does it fit

with the fact that the genetic code is degenerate?

- Be familiar with the genetic code and be able to use it to deduce the primary structure of a polypeptide from an mRNA sequence

10.1.2 Overview of Genetic Code

We speak of genes (i.e., DNA) coding for proteins and the central dogma, which states that DNA makes RNA makes protein.

What does this actually mean? A code can be thought of as a system for storing or communicating information.

Analogy: A familiar example is the use of letters to represent the names of airports (e.g., PDX for Portland, Oregon and ORD for Chicago's O'Hare). When a tag on your luggage shows IND as the destination, it conveys information that your bag should be sent to Indianapolis, Indiana. To function well, such a set-up must have unique identifiers for each airport and people who can decode the identifiers correctly. That is, IND must stand only for Indianapolis, Indiana and no other airport. Also, luggage handlers must be able to correctly recognize what IND stands for so that your luggage doesn't land in Iowa, instead.

How does this relate to genes and the proteins they encode?

Genes are first transcribed into mRNA, as we have already discussed. The sequence of an mRNA, copied from a gene, directly specifies the sequence of amino acids in the protein it encodes.

The genetic code is the information for linking amino acids into polypeptides in an order based on the base sequence of **3-base codewords (codons) in a gene and its messenger RNA (mRNA)**.

For example, the amino acid tryptophan is encoded by the sequence UGG on an mRNA. All of the twenty amino acids used to build proteins have, likewise, 3-base sequences that encode them.

Concept Note: The emphasis on understanding the polarity of DNA, RNA (coding versus non-coding) should be even clearer now. The code is always read in a fixed direction, i.e., in the 5' → 3' direction! If the code is read in opposite direction (i.e., 3' → 5'), it would specify 2 different proteins since the codon would have reversed base sequence.

Fig. 10.3A (below) shows representations of the genetic code in the ‘language’ of RNA.

The left-hand vertical column indicates the first (5') position in a codon, the horizontal bar across the top indicates the second position, and the right-hand vertical column indicates the third (3') position

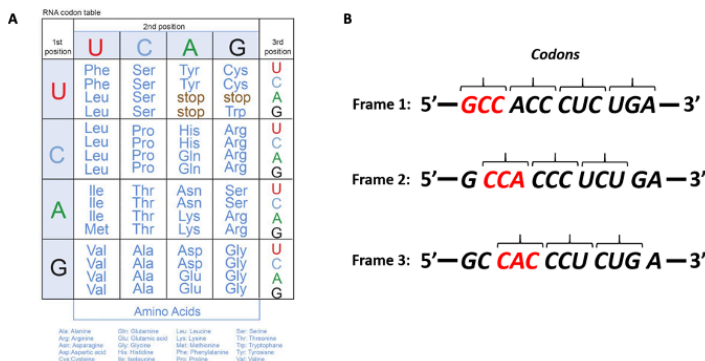


Figure 10.3 A. This figure shows the genetic code for translating each nucleotide triplet in mRNA into an amino acid or a termination signal in a nascent protein. Figure from: <https://www.genome.gov/genetics-glossary/Genetic-Code>, in the Public Domain. **(B)** Possible reading frames

10.1.3 Features of the Genetic Code

Three nucleotides encode an amino

acid

Template mRNA is read by the ribosome in groups of three nucleotides, called a **codon** (Fig. 10.3 B). Simple calculations hypothesized (a minimum of 3 bases would be needed to code for 20 amino acids) and genetic experiments ultimately proved this to be the case.

Non-overlapping and Unambiguous

The template is non-overlapping and read in discrete groups of three. This is known as the **reading frame** of the **mRNA**, and it is always read from the 5' to 3' direction.

Thus, for **each mRNA**, there are **three potential reading frames** (Fig. 10.3A). Only one reading frame will be the correct one for protein synthesis.

The ribosome must recognize and align the correct reading frame of the mRNA such that the correct codon sequences can be read.

Look at the codon chart in Figure 10.3A. We can see that each codon is **specific for a single amino acid**.

For example UUU is always coding for Phenylalanine and not any other amino acid.

There is **very little ambiguity** within the code.

There is no punctuation

The sequence of bases is read continuously without stopping or skipping nucleotides.

Degeneracy and Redundancy

Given that there are 4 bases in RNA, the number of different 3-base combinations that are possible is 4³, or 64. There are, however, only 20 amino acids that are used in building proteins in cells.

This discrepancy in the number of possible codons and the actual number of amino acids they specify is explained by the fact that **the same amino acid may be specified by more than one codon.**

In fact, with the exception of the amino acids **methionine and tryptophan**, all the other amino acids are encoded by multiple codons.

Codons for the same amino acid are often related, with the first two bases the same and the third being variable.

An example would be the codons for alanine: GCU, GCA, GCC, and GCG all stand for **alanine**.

This sort of redundancy in the genetic code is termed **degeneracy**.

Rules of translation: Stop and start codons

All codons that code for an amino acid are also referred to as SENSE Codons.

Whereas 61 of the 64 possible triplets code for amino acids, three of the 64 codons **do not code for an amino acid**; they terminate protein synthesis, releasing the polypeptide from the translation machinery.

These are called stop codons or nonsense codons. The three stop codons in the Standard Genetic Code ‘tell’ ribosomes the location of the last amino acid to add to a polypeptide.

The three stop codons are UAA, UGA and UAG. The three stop codons also have colloquial names: UAA (ochre), UAG (amber), UGA (opal), with UAA being the most common in prokaryotic genes.

In contrast, evolution has selected the codon for methionine, **AUG**, as the **start** codon for all polypeptides (regardless of their function) and for the insertion of methionine within a polypeptide.

Thus, all polypeptides begin life with a methionine at their amino-terminal end!

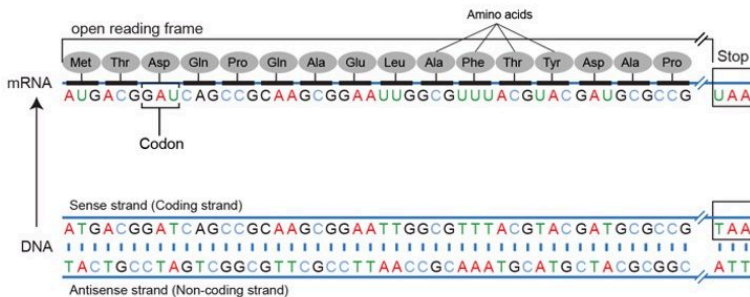
Analogy: This ingenious system is used to direct the assembly of a protein in the same way that you might string together colored beads in a particular order using instructions

that used symbols like UGG for a red bead, followed by UUU for a green bead, CAC for yellow, and so on, till you came to UGA, indicating that you should stop stringing beads.

Open Reading Frame

As mentioned above when the genetic code is read on mRNA there are three potential reading frames.

The frame is set by the AUG start codon near the 5' end of the mRNA. Each set of three nucleotides following this start codon is a codon in the mRNA message until the termination codon is the reading frame.



Open Reading Frame. Image from <https://www.genome.gov/genetics-glossary/Open-Reading-Frame> (in Public Domain)

Using the same diagram as above we can see that of the potential reading frames only one of them makes an intelligible protein.

The first terminates immediately, the second runs into the end and still no termination codon.

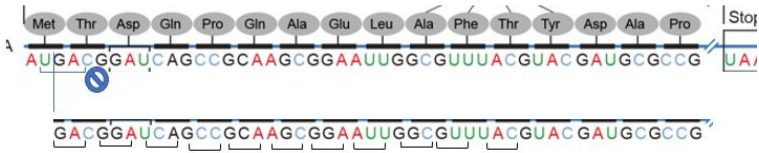


Figure 10.4 Other frames of reading do not code for accurate a protein, and thus are likely not the correct reading frame. Image modified from: <https://www.genome.gov/genetics-glossary/Open-Reading-Frame>

In practice usually, the one with the longest stretch of codons is typically indicative of an open reading frame.

Did I get this? Concept of Open Reading Frames

When scanning a genome for genes that may encode proteins, scientists use bioinformatics programs like **ORF Finder** to look for start codons, stop codons, and stretches of DNA in between the two that code for proteins at least 50 to 300 amino acids long.

Other clues include presence of promoter sequences ahead of the start codon.

These open reading frames can then be analyzed further, using bioinformatics tools like BLAST searches and phylogenetic analyses to determine whether these areas are similar to other known genes from other organisms, which may then warrant further study in the lab.

Gene sequences are largely conserved – so if an ORF sequence is present in multiple genomes, it likely represents a gene!

Q. For a **double-stranded DNA molecule** how many reading frames are possible?



An interactive H5P element has been excluded from this version of the text. You

can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1347#h5p-50>

(Nearly) Universal

With a few exceptions and with minor changes (some prokaryotes, mitochondria, chloroplasts, ciliated protozoa), the genetic code is the same in all organisms from viruses and bacteria to humans, providing support for a single origin of life.

Exceptions include some protozoans using UAA and UAG as codons for amino acids rather than as stop signals; UGA is their sole termination signal. Mitochondrial DNA encodes for a distinct set of mitochondrial tRNA;s which can recognize alternative codons. Thus the genetic codes in *nearly* universal.

Practice: How to Read a Codon Chart



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1347#oembed-1>

[Click here for Dr. Mehta's Lecture Video for Genetic Code \(13 minutes\)](#)

Molecular Biology in the News: Concepts in Cotenxt

[A new universe of mini proteins is upending cell biology and genetics! Tiny proteins help power muscles and provide the toxic punch to many venoms.](#)

Read the associated article (PDF files on CANVAS if the link above is paywalled).

As you read think about:

The criteria that scientists used for identifying genes and why these mini proteins were missed and how the method to determine

which proteins cells are making that as developed helped them find these proteins.

Learning Objectives

1. What is the adaptor hypothesis and how does the tRNA fit into this hypothesis?
2. Be able to predict how mutations in anticodon section of a tRNA will affect 'translation' of mRNA.
3. Describe the roles and relationships between: tRNA synthetases and tRNA molecules, tRNA anticodon sequences and mRNA codon sequences. How are tRNAs linked to their corresponding amino acids?
4. Bacteria have two different kinds of tRNA^{met}. What roles do these tRNA play in polypeptide synthesis?
5. How does translation initiation in eukaryotes differ from that in prokaryotes?
6. Describe or illustrate or be able to label the initiation stage in bacterial polypeptide synthesis. Discuss the

role of each protein factors that participate in this process.

7. Give the elongation factors used in bacterial translation and explain the role played by each factor in translation.
8. What is the role of codons UAA, UGA, and UAG in translation? What events occur when one of these codons appears at the A site of the ribosome?
9. Compare and contrast the process of protein synthesis in bacterial and eukaryotic cells, giving similarities and differences in the process of translation in these two types of cells.
10. In a diagram of translation be able to label:

5' and 3' ends of the mRNA ; A, P, and E sites ; Start codon; Stop codon ; Amino and carboxyl ends of the newly synthesized polypeptide chain; Approximate location of the next peptide bond that will be formed; Place on the ribosome where release factor 1 will bind

LEVEL UP

1. Explain what the potential effect of a mutation in the part of the tRNA gene that encodes: (a) the acceptor stem; (b) the anticodon
2. Explain in a cell-free protein-synthesizing system (in a test tube) the effects of omitting various factors in translation.(What, if any, type of protein would be produced? Explain your reasoning.)
3. Explain how some antibiotics work by affecting the

process of protein synthesis.

10.2 tRNA's: The Interpreter of the Code

While the ribosomes are the factories that join amino acids together using the instructions in mRNAs, another class of RNA molecules, the transfer RNAs (tRNAs) are also needed for translation.

In terms of the bead analogy above, someone or something has to be able to bring a red bead in when the instructions indicate UGG, and a green bead when the instructions say UUU. This, then, is the function of the tRNAs.

They act as ADAPTORS or interpreters of the code. They act as adaptors by binding to the codon on one end and carrying the amino acid on the other end.

There is at least one tRNA for each amino acid.

All transfer RNAs share common features and are structurally similar. These include

1. Transfer RNAs are small single-stranded RNA molecules, about 75-90 nucleotides long.
2. Transfer RNAs are extensively modified post-transcriptionally and contain a large number of **unusual** bases and **modified** bases like Inosine.
3. Mature tRNAs take on a three-dimensional structure

where the single-stranded tRNA folds on itself and base-pairs to form what is sometimes described as a ***stem-loop, or cloverleaf*** structure.

This structure is crucial to the function of the tRNA, providing both the sites for attachment of the appropriate amino acid and for recognition of codons in the mRNA.

The cloverleaf consists of five parts: the acceptor stem (containing the tRNA's 5'- and 3'-ends), the D-arm, the anticodon arm, the variable loop and the TYC-arm (T-arm).

4. All tRNA's contain at the 3'-terminus (acceptor stem) the nucleotides CCA. These are added to the tRNA post-transcriptionally by CCA-adding enzymes.

5. For all tRNA amino acid is attached to a hydroxyl group of the A (of the CCA sequence).

6. At the other end of acceptor's arm is the ***anticodon loop***.

Every tRNA has a sequence of 3 bases, the **anticodon**, that is complementary to the codon for the amino acid it is carrying. When the tRNA encounters the codon for its amino acid on the messenger RNA, the anticodon will base-pair with the codon.

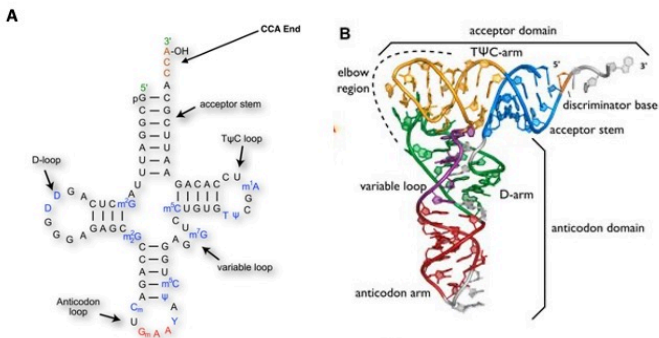


Figure 10.5 . (A) 2 D representations of tRNA showing various features (B) The L-shaped tertiary structure of the cytosolic tRNA^{Phe} from *S. cerevisiae*. Protein Data Bank entry (PDB): 1EHZ. The acceptor domain is composed of a stacked T-arm and acceptor stem, whereas D- and anticodon arm form the anticodon domain. Figure (B) From [Lorenz, C., et. al. \(2017\) Biomolecules 7\(2\):35](#)

Note that the pairing of anticodon with codon within the message like all other forms of nucleic acid interactions is ‘antiparallel’.

Also note that the sequences are both written, by convention, in the 5’ to 3’ direction as we have seen earlier for all written forms.

For the tryptophan tRNA this is what it would look like:

The sequence of tryptophan codon in mRNA: 5’ -UGG- 3’

The codon-anticodon basepair in the *antiparallel* orientation then would be:

5' – UGG -3'
 3'- ACC- 5'

Wobble Base Pairing

The degeneracy of the genetic code – where many tRNA molecules can recognize **more than one codon using a single anticodon** is due to a feature known as ‘Wobble Base Pairing’.

Wobble base pair is a pairing between two nucleotides in RNA molecules **that does not follow Watson-Crick base pair rules.**

The four main **wobble base pairs** are guanine-uracil (G-U), hypoxanthine-uracil (I-U), hypoxanthine-adenine (I-A), and hypoxanthine-cytosine (I-C).

The ***wobble base*** position is usually the first position of the anticodon (read in the 5' – 3' direction), which aligns with the 3rd position of the mRNA codon.

In order to maintain consistency of nucleic acid nomenclature, “I” is used for hypoxanthine because hypoxanthine is the nucleobase of the **inosine nucleotide—one of the modified bases on tRNA's**

The thermodynamic stability of a wobble base pair is comparable to that of a Watson-Crick base pair.

Wobble base pairs are fundamental in RNA secondary structure and are critical for the proper translation of the genetic code.

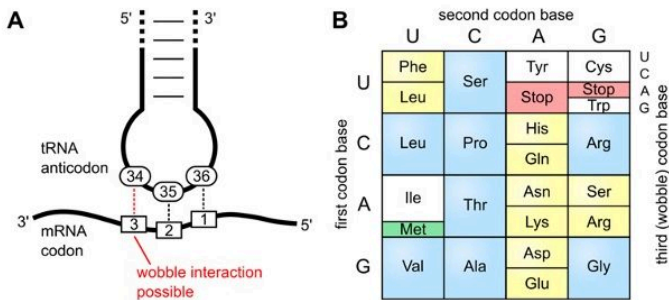


Figure 10. 6 Anticodon Loop Structure and Codon Degeneracy. (A) The interaction of the anticodon bases (34–36) of a tRNA with the corresponding bases of the mRNA codons (3, 2, 1). A wobble interaction is possible between codon base 3 and anticodon base 34. The latter is frequently modified and directs the wobble interactions with the third codon base; (B) The standard genetic code is illustrated as a simple decoding table, 2-fold degenerate codon boxes are colored yellow, 4-fold degenerate boxes are blue. Start and stop codons are colored green and red, respectively. *Figure from: Lorenz, C., et. al. (2017) Biomolecules 7(2):35*

Watch Dr. Mehta Lecture Video on [tRNA and Wobble Base Pairing](#) (includes a time to think exercise) (10 minutes)

Did I get this?

A series of tRNAs have the following anticodons. Consider the wobble rules listed in Table 10.6, and give all possible codons with which **each tRNA can pair with**

A) 5'-GGC-3'

B) 5'-AAG-3'

Key Takeaways

Complete this exercise to summarize the key takeaways thus far.



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/>

iul211smehta/?p=1347#h5p-51

10.2.1 Charging of tRNA's"- Amino Acyl tRNA Synthetases

The fidelity (accuracy) of protein synthesis is maintained by the ribosome's ability to match the code from the template mRNA strand with the appropriate amino acid.

However, it is the tRNA that forms a physical link between the mRNA and the amino acid the codon represents.

Therefore, the accuracy of translation hinges upon the very important process of ensuring that the correct amino acid is added to the appropriate tRNA!

Before the tRNA is brought to the ribosome, amino acids are attached to the tRNA! The attachment of amino acids to the tRNA is known as 'Charging of tRNA'. A pool of charged tRNAs is necessary to carry out protein synthesis.

The attachment of amino acids to tRNA's is the job of an important class of enzymes called: *Aminoacyl tRNA synthetases*.

There are 20 aminoacyl-tRNA synthetases- ONE for each of the 20 amino acids!

They are named after the aminoacyl-tRNA product generated, as such, methionyl-tRNA synthetase (abbreviated as MetRS) charges tRNA^{Met} with **methionine**.

tRNAs bearing an aminoacyl linkage (amino acid) are said

to be charged. Charged tRNA's are often indicated in written form as with the amino acid in superscript. **Example: met-tRNA^{Met}**

The amino acid gets attached to the tRNAs in a 2 step process that involves – recognizing both a specific tRNA and a specific amino acid, binding an ATP for energy, and then joining them together.

Depending on the class of synthetase, the amino acid attaches to the **2'-OH of the terminal A (class I)** or to the **3'-OH of the terminal A (class II)** of the tRNA.

In order to ensure the faithful translation of the genetic message, synthetases must identify and pair particular tRNAs with their corresponding amino acid which relies on the proper recognition of both substrates.

This can prove extremely challenging for the synthetases. They need to discriminate the correct tRNA amongst a set of other tRNAs very similar in structure and chemical composition, but also be able to select the correct amino acid amidst an extremely large pool of similar amino acids.

The evolutionary pressure to maintain fidelity has driven aaRSs to develop an elevated specificity for their substrates, both the tRNA and the amino acid. There is also a built-in pre-attachment proofreading mechanism in that tRNA molecules that fit the synthetase well (i.e. the correct ones) maintain contact longer and allow the reaction to proceed whereas ill-fitting and incorrect tRNA molecules are likely to

disassociate from the synthetase before it tries to attach the amino acid.

Watch Dr. Mehta's Lecture Video on [Aminoacyl tRNA synthetases \(6 minutes\)](#)

10.3 Ribosome Structure

The ribosome is a highly conserved molecular machine.

In all organisms, it is composed of two unequal subunits, called LARGE and SMALL subunits. Each consists of a distinct set of ribosomal RNA (rRNA) and ribosomal proteins (RPs) that combine to form a large nucleoprotein complex.

Prokaryotic ribosomes: have a mass of about 2500 kDa and size of **70S** (or Svedberg units: A Svedberg unit is a measure of the sedimentation rate in a centrifuge and thus is representative of size).

A complete ribosome (70S) can be dissociated into large subunit (50S) and a small subunit (30S)

Eukaryotic ribosomes are larger than their prokaryotic counterparts at approximately **80S** (although there is some modest variation between eukaryotic species). Human

cytosolic ribosomes are composed of a **large subunit (60S)** and a **small subunit (40S)**.

The ribosome structures in all living organisms regardless of size however function similarly and carry out three important tasks.

- 1) Bind the mRNA and find the start codon, where translation will begin

- 2) Facilitate (provide a place) for the tRNA to come in and 'decode'. Molecularly- facilitate the complementary base pairing of mRNA codons and tRNA anticodons that determines amino acid order in the polypeptide.

- 3) Catalyze the peptide bond formation between the amino acids.

Structurally they harbor three different tRNA binding sites that helps the ribosomes carry out these tasks.

The A-site, where decoding occurs and the correct aminoacyl-tRNA (aa-tRNA) is selected on the basis of the mRNA codon displayed.

During protein synthesis tRNA's charged with amino acid are entering here.

The P-site, which holds the peptidyl-tRNA, (the growing polypeptide)

The E-site, binds exclusively to deacetylated tRNAs (uncharged tRNAs) that are exiting the ribosome.

Thus, during translation the tRNA moves from the A-site through the P- and E-site, where it leaves the ribosome (Fig. 10.7).

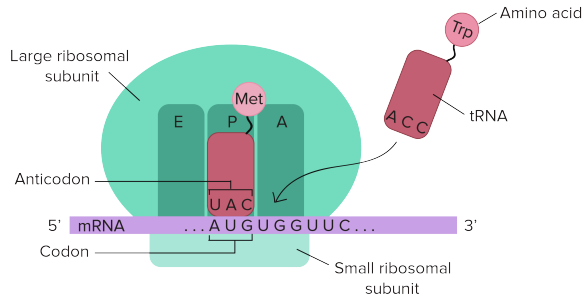


Figure 10.7 Schematic Structure of an Active Ribosome. The mRNA (shown in purple) is assembled between the small subunit and the large subunit of the ribosome (shown in green). tRNA molecules (shown in red) that are loaded with their cognate amino acid (shown in pink) are transitioned through the A-P-E sites of the ribosome during the elongation phase of translation. The movement of the tRNA molecules also shifts the position of the mRNA causing the next three codon bases to line up in the A-site of the ribosome.

Figure from: [The Khan Academy where it was modified from Openstax College Biology](#)

10.3.1 Overall Steps in Translation

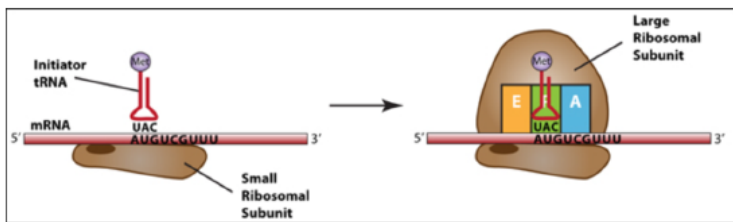
Translation occurs in three phases and involves cycling in and of tRNAs through the ribosome sites mentioned above.

Initiation

Finding the correct AUG sets the open reading frame. (Needs initiator tRNA's and initiation factors).

At the end of initiation, the start codon (AUG) is positioned to base pair with the tRNA in the **P-site (peptidyl site)**.

This is the only time tRNA charged with amino acids occupies the P-site.



Elongation: joining of adjacent amino acids –carried by the tRNA successively.

A tRNA bound to its amino acid (known as an aminoacyl-tRNA) that is able to base pair with the next codon on the mRNA arrives at the A site.

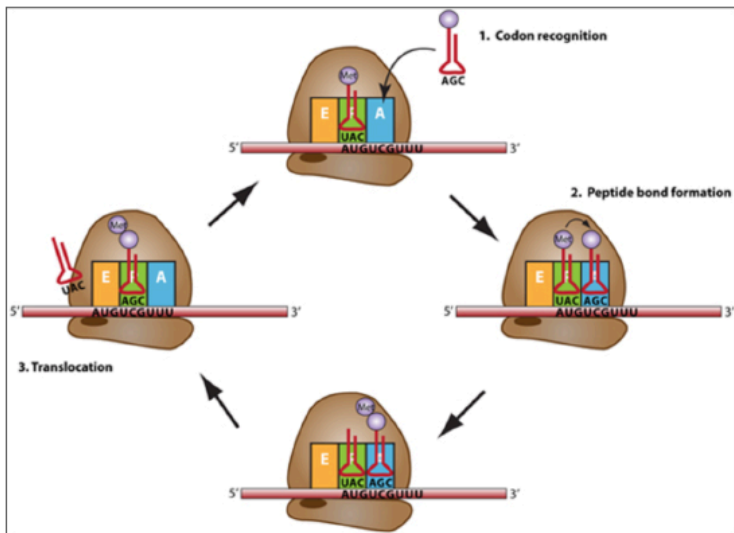
The preceding amino acid (Met at the start of translation) is covalently linked to the incoming amino acid with a **peptide bond**.

The bond between the amino acid and the tRNA in the P-site is broken and the dipeptide is joined to the tRNA on the A-site.

The initiator tRNA moves to the E site and the ribosome

moves one codon downstream. This shifts the most recent tRNA from the A site to the P site, opening up the A site for the arrival of a new aminoacyl-tRNA.

This cycle continues, with the ribosome moving on the mRNA one codon at a time, until the stop codon reaches the A-site.



Termination: Termination codons are recognized by release factors. Completed polypeptide chain is released.

The ribosome then dissociates into the small and large subunits, once more.

[Link to Learning](#)

Watch this NDSU Virtual Cell Animations “Translation” for an **overview of Translation**.

Note: This video uses **Eukaryotic mRNA** as an example.



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1347#oembed-2>

Question. What information about the mRNA described lets us know this video is talking about eukaryotic mRNA?

Answer at end.

10.3.2 Polysomes

Each mRNA molecule is simultaneously translated by many ribosomes, all synthesizing protein in the same direction: reading the mRNA from 5' to 3' and synthesizing the polypeptide from the N terminus to the C terminus.

The complete structure containing an mRNA with multiple associated ribosomes is called a polyribosome (or polysome). In bacteria, before transcriptional termination occurs, each protein-encoding transcript is already being used to begin the synthesis of numerous copies of the encoded polypeptide (s) because the processes of transcription and translation can occur concurrently, forming polyribosomes (Figure 11.4.2">10.8).

This allows a prokaryotic cell to respond to an environmental signal requiring new proteins very quickly.

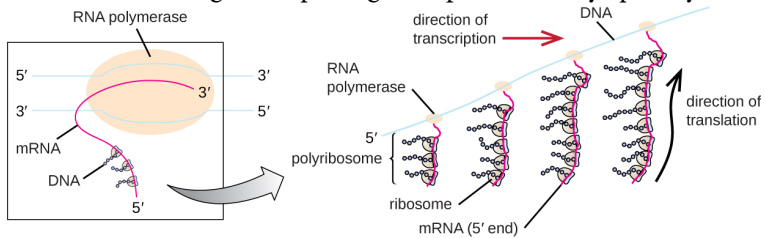


Figure 11.4.2">10.8

In prokaryotes, multiple RNA polymerases can transcribe a single bacterial gene while numerous ribosomes concurrently translate the mRNA transcripts into polypeptides. In this way, a specific protein can rapidly reach a high concentration in the bacterial cell. Figure from: [“Protein Synthesis \(Translation\)” by OpenStax, LibreTexts is licensed under CC BY.](#)

Watch Dr. Mehta's Lecture Video on [Ribosomes](#).

[Polysomes, and Overview of Translation](#) (14 minutes)

Concept Check

Concept: In both prokaryotic and eukaryotic cells, multiple ribosomes may translate a single mRNA molecule simultaneously, generating a structure called a **polyribosome**.

In a polyribosome, the polypeptides associated with which ribosomes will be the longest?

- a) Those at the 5' end of mRNA
- b) Those at the 3' end of mRNA
- c) Those in the middle of mRNA
- d) All polypeptides will be the same length.

Answer at end.

10.4 Details of Translation

Having considered the steps of translation in broader terms, we can now look at them in greater detail.

As in all the processes we have learned, the first step Initiation is where most of the differences occur between Prokaryotic and Eukaryotic Translation.

10.4.1 Initiation of Protein Translation

There are three steps to translation initiation (achieved differently in eukaryotes and in bacteria) that conceptually involve

1. Identifying the start codon (involves the small ribosome subunit)
2. Positioning the initiator tRNA in P-site
3. Forming the active complex by joining of large ribosome subunit.

Protein factors called **initiation factors** facilitate these steps, ensure speed and accuracy to the overall process.

First, watch the lecture video on [Prokaryotic Initiation](#)

Messenger RNAs have non-coding sequences both at their 5' and 3' ends, with the actual protein-coding region sandwiched in between these untranslated regions (called the 5' UTR and 3' UTR, respectively).

The ribosome must be able to recognize the 5' end of the mRNA and bind to it, then determine where the start codon is located.

10.4.2 Prokaryotic Initiation Key Features

Initiator tRNA

Initiation also requires the binding of the **first tRNA** to the ribosome. As we have noted earlier, the initiation or start codon is usually AUG, which codes for the amino acid **methionine**.

Thus, the initiator tRNA is one that carries methionine and is designated as tRNA^{met} or methionyl tRNA^{met}.

In prokaryotes, the methionine on the initiator tRNA is modified by the addition of a **formyl group** and is designated **tRNA^{fmet}**.

The initiator tRNA carrying methionine to the AUG is different from the tRNAs that carry methionine intended for other positions in proteins. As such, the initiator tRNA is sometimes referred to as **tRNA^{i-met}**

fMet is only used for the initiation of protein synthesis

and is thus found only at the N-terminus of the protein. Unmodified methionine is used during the rest translation. Once protein synthesis is completed, the formyl group on methionine may be removed and on occasion, the entire methionine residue can be further removed by special enzymes.

Shine-Dalgarno sequence

In prokaryotes, the 5' end of the mRNA is the only free end available, as transcription is tightly coupled to translation and the entire mRNA is not transcribed before translation begins.

Nevertheless, the ribosome must be correctly positioned at the 5' end of the messenger RNA in order to initiate translation.

How does the ribosome “know” exactly where to bind in the 5'UTR of the mRNA?

Examination of the sequences upstream of the start codon in prokaryotic mRNAs reveals that there is a short purine-rich sequence ahead of the start codon that is crucial to recognition and binding by the small ribosomal subunit.

This sequence, called the **Shine-Dalgarno sequence**, is **complementary to a stretch of pyrimidines at the 3' end of the 16S rRNA component of the small ribosomal subunit**

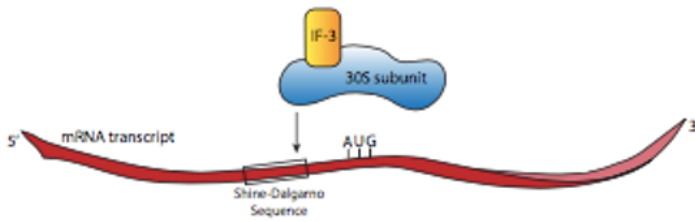


Figure from: “Prokaryotic Translation” by E. V. Wong, LibreTexts is licensed under CC BY-NC-SA .

Base-pairing between these complementary sequences positions the small ribosomal subunit at the right spot on the mRNA, with the AUG start codon at the P-site.

Initiation factors

Think about answers to Level-Up learning objective #2 as you review

The binding of the small ribosomal subunit to the mRNA requires the assistance of three protein factors called **Initiation Factors 1, 2, and 3 (IF1, IF2, IF3)**.

These proteins, which are associated with the small ribosomal subunit, are necessary for its binding to mRNA, but dissociate from it when the 50S ribosomal subunit binds.

IF3= Initiation Factor 3

1. An antiassociation factor; prevents association between the large and small ribosomal subunits.
2. It also must be associated with the small subunit for it to form an initiation complex, i.e. for the small subunit to correctly bind mRNA and fmet-tRNA^f.
3. It dissociates prior to binding of the large subunit

IF1 = Initiation Factor 1

1. Prevents premature association of amino-acyl tRNAs with the small ribosome subunit

IF2

1. Brings **fmet-tRNA^f** to the partial P site on the small subunit.
2. IF2 activates a GTPase activity in the small subunit. The resulting change in conformation may allow the large subunit to bind.

Once the small ribosomal subunit is bound to the mRNA and the initiator tRNA is positioned at the P-site, the large ribosomal subunit is recruited and the initiation complex is formed.

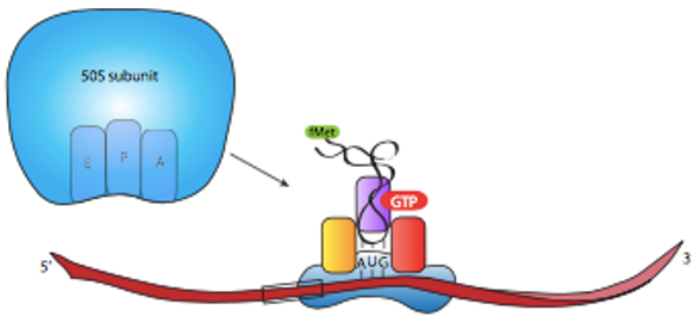


Figure from: “Prokaryotic Translation” by E. V. Wong, LibreTexts is licensed under CC BY-NC-SA .

The binding of the 50S ribosomal subunit is accompanied by the dissociation of all three initiation factors.

The removal of IF1 from the A-site on the ribosome frees up the site for the binding of the charged tRNA corresponding to the second codon.

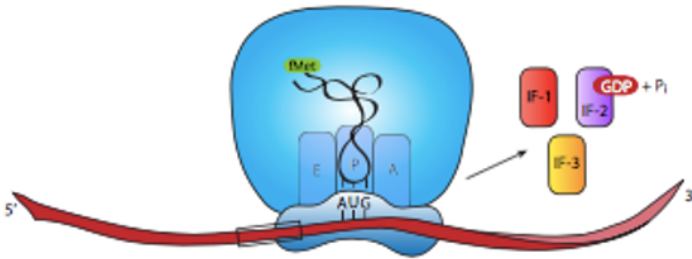


Figure from: “Prokaryotic Translation” by E. V. Wong, LibreTexts is licensed under CC BY-NC-SA .

Before you continue you should

1. Watch the videos (as was instructed in the chapter)
 2. Test yourself with the Lecture Quickcheck (as a quiz)
 3. Attempt Weekly Problem Questions
-

10.4.3 Eukaryotic Initiation

First, watch the lecture video here: [Eukaryotic](#)

[Translation Initiation \(17 minutes\)](#)

Key Points

The initiation process is slightly more complicated, but the elongation and termination processes are the same, but with eukaryotic homologs of the appropriate elongation and release factors.

Eukaryotic initiation factors are written as – eIFs where the ‘e’ stands for eukaryotic and IF for initiation factor.

Eukaryotes have a **large number of IFs** involved in the binding of the initiator tRNA to the small subunit, as well as in association of the small subunit with mRNA and subsequent attachment of the large subunit.

We will not cover the action of **all the eIFs** in detail but rather focus on a few key steps.

Table: Comparison of Prokaryotic and Eukaryotic Translation Initiation Factors

Prokaryotic Factors	Function	GTP Hydrolase	Eukaryotic Factors	Function'	GTP Hydrolase'
IF1	Associates with 30S subunit at A-site. Prevents premature assembly of large subunit	No	eIF1, eIF1A	induce open conformation of mRNA binding channel of 40S subunit	No
IF2	Recruits fMet-tRNA to P-site	Yes	eIF2, eIF5B	delivers tRNAi to 40S subunit	Yes
IF3	Required for correct association of mRNA with small subunit rRNA	No	eIF3	binds to 40S subunit and helps recruit mRNA through interaction with other eIFs	No
			eIF4E	Bind to 5' CAP of mRNA	No
			eIF4G	Binds PABP	No
			eIF4A	RNA helicase important for resolving secondary structure	No
			eIF5	GTPase activating protein that helps eIF2 in recruiting the large subunit	No
			eIF6	Binds large subunit of ribosome and helps with eIF3 to link with small subunit	No
			PABP	Binds to polyA tail of mRNA and links it to 5' end	No

Actively Translating eukaryotic mRNA's are circular!

In eukaryotes the processed mRNA contains additional modifications:

A) A CAP at the 5' end – bound by CAP binding protein

B) A Poly-A tail at the 3' end- bound by Poly A Binding Proteins.

This processed mRNA exits the nucleus, and in the cytoplasm **eukaryotic Initiation factors eIF, replace the CAP binding protein one of which is eIF4E.**

Another protein **eIF4G** connects with the eukaryotic initiation factors assembled at the 5' end with the 3' end poly A- binding proteins to create a circular structure. (See **Figure 10.9 below**)

The binding of the mRNA cap by eIF4E is often considered the rate-limiting step of ***cap-dependent initiation***, and the concentration of **eIF4E is a regulatory nexus of translational control.**

Link to Learning: Concepts in Context

Certain viruses cleave a portion of eIF4G that binds eIF4E, thus preventing cap-dependent translation to hijack the host machinery in favor of the viral (cap-independent) message!

Ribosome assembly and key difference with prokaryotic initiation:

Unlike prokaryotes -the assembly of the translation machinery in eukaryotes **begins with the binding of the initiator tRNA to the 40S (small) subunit BEFORE the subunit binds the mRNA.**

This step requires the assistance of **eIF2** and other factors. The complex of the small ribosome accompanied by eukaryotic initiation factors and Met-tRNA_i is known as the **ternary complex.**

Next, the **small complex** with the initiator tRNA binds to the **7-methyl G cap on the 5'end of the mRNA.**

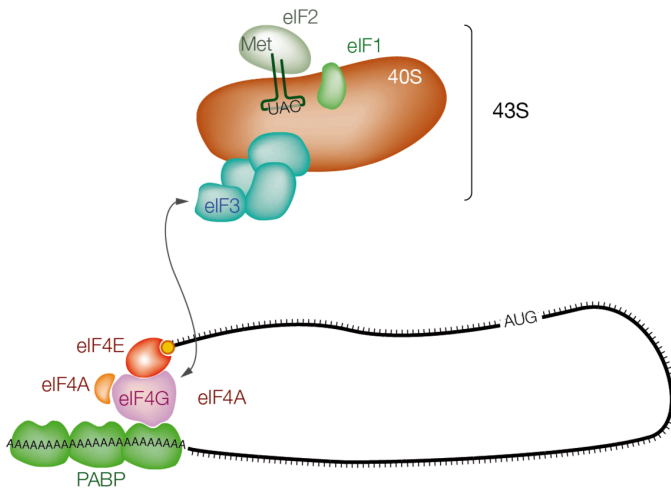


Figure 10. 9 Eukaryotic Translation Initiation. This is a simplified diagram of eukaryotic translation initiation detailing some of the eIFs involved in the process. eIF2 is critical for recruiting the initiation tRNA_i to the 40S subunit. The 43S pre-initiation complex through the interaction of the eIF4 factors and causes the scanning of the pre-initiation complex down the mRNA to locate the start codon (usually AUG). Poly A Binding Proteins (PABPs) bind with the polyA tail sequence of the mRNA and also interact with the eIF4 factors causing the circularization of the mRNA. *Figure from: [Eukaryotic Translation, Wikiwand](#)*

This 43S preinitiation complex accompanied by the protein factors moves along the mRNA chain toward its 3'-end, in a process known as 'scanning', to reach the start codon (typically AUG).

After recognition of the start codon, the large ribosomal subunit (60S) assembles to form the 80S initiation complex,

Kozak sequences

Specific sequences surrounding the AUG, called Kozak sequences for the scientist who defined them, have been shown to be necessary for the binding of the 40S subunit, with the bases at -4 and +1 relative to the AUG being especially important.

Once the small subunit is properly positioned, the large ribosomal subunit (60S) binds, forming the initiation complex.

10.5 Elongation and Termination

First, watch the lecture video here: [Translation Elongation and Termination \(19 minutes\)](#)

Note: The lecture video has additional information on Antibiotics and Translation to help with applied questions.

Key Points

We only discuss elongation and termination in a prokaryotic system, due to the similarity between the processes between organisms.

Elongation

After the ribosome is assembled with the initiator tRNA positioned at the AUG in the P-site, the addition of further amino acids can begin.

In both prokaryotes and eukaryotes, the elongation of the polypeptide chain requires the assistance of elongation factors.

In bacteria, the binding of the second charged tRNA at the A-site requires the elongation factor EF-Tu complexed with GTP.

When the charged tRNA has been loaded at the A-site, EF-Tu hydrolyzes the GTP to GDP and dissociates from the ribosome.

The free EF-Tu can then work with another charged tRNA to help position it at the A-site, after exchanging its GDP for a new GTP.

The reaction that joins the amino acids occurs in the

ribosomal peptidyl transferase center, which is part of the large ribosomal subunit. This reaction is catalyzed by rRNA components of the large subunit, making the formation of peptide bonds an example of the activity of RNA enzymes, or ribozymes.

IMPORTANT: A common and understandable misconception is that the new amino acid brought to the ribosome is added *onto* the growing polypeptide chain. In fact, the mechanism is exactly the opposite: the polypeptide is added **to the** new amino acid. This begins with the second amino acid to be added to a new protein. The first amino acid, a methionine, you should recall, **came in along with IF-2 and the initiator tRNA.**

The result of the peptidyl transferase activity is that the tRNA in the A-site now has two amino acids attached to it, while the tRNA at the P-site has none. This “empty” or deacylated tRNA is moved to the E-site on the ribosome, from which it can exit.

The tRNA in the A-site then moves to occupy the vacated

P-site, leaving the A-site open for the next incoming charged tRNA.

Yet another elongation factor, **EF-G complexed with GTP**, is required for the translocation of the ribosome along the mRNA in bacteria.

Repeated cycles of these steps result in the elongation of the polypeptide by one amino acid per cycle, until a termination, or stop codon is in the A-site.

Termination

When a stop codon is in the A-site, proteins called release factors (RFs) are needed to recognize the stop codon and cleave and release the newly made polypeptide.

In bacteria, RF1 is a release factor that can recognize the stop codon UAG, while RF2 recognizes UGA. Both RF1 and RF2 can recognize UAA. A third release factor, RF3, works with RF1 and RF2 to hydrolyze the linkage between the polypeptide and the final tRNA, to release the newly synthesized protein.

This is followed by the dissociation of the ribosomal subunits from the mRNA, ending the process of translation.

Before you continue you should

1. Watch the videos (as was instructed in the chapter)
2. Test yourself with the Lecture Quickcheck (as a quiz)
3. Attempt Weekly Problem Questions

Answers to Problems in text:

[modifications of mRNA ; (c)]

References and Attributions

This chapter contains material taken from the following CC-licensed content. Changes include rewording, removing paragraphs and replacing with original material, and combining material from the sources.

1. Bergtrom, Gerald, “Cell and Molecular Biology 4e: What We Know and How We Found Out” (2020). *Cell and Molecular Biology 4e: What We Know and How We Found Out – All Versions*. 13.

https://dc.uwm.edu/biosci_facbooks_bergtrom/13

2. Works contributed to LibreTexts by Kevin Ahern and Indira Rajagopal. LibreTexts content is licensed by CC BY-NC-SA 3.0. The entire textbook is available for free from the authors at <http://biochem.science.oregonstate.edu/content/biochemistry-free-and-easy>

3. Flatt, P.M. (2019) Biochemistry – Defining Life at the

Molecular Level. Published by Western Oregon University, Monmouth, OR (CC BY-NC-SA). Available at: https://wou.edu/chemistry/courses/online-chemistry-textbooks/ch450-and-ch451-biochemistry-defining-life-at-the-molecular-level/?preview_id=4919&preview_nonce=cca8f0ce36&preview=true

4. “Translation” by Katherine Harris, LibreTexts is licensed under CC BY-NC-SA .

5. “Protein Synthesis (Translation)” by OpenStax, LibreTexts is licensed under CC BY .

PART II

TOOLS AND TECHNIQUES OF MOLECULAR BIOLOGY

INTRODUCTION

Introduction

Molecular biology as a discipline was defined with the convergence of biochemical and genetic techniques. Technologies that enabled manipulation of genetic material and introduce it into cells- *genetic engineering and recombinant DNA*. This is such seminal technology that just realizing it could be done and then doing it in a test tube for the first time earned Paul berg a half-share in the 1980 Nobel Prize in Chemistry (the other half was shared by Walter Gilbert and Frederick Sanger for studies that enabled efficient *DNA sequencing*). Molecular Biology is an experimental science, and a central element to the understanding of molecular biology is an appreciation of the approaches taken to yield the information from which concepts and principles are deduced. All aspects of biological research that investigate cellular mechanisms will employ or utilize various molecular biology tools.

In this unit, we will explore some commonly used techniques, learn the basics of recombinant DNA technology, commonly referred to as DNA cloning. We will also dive into newer genome-wide technologies that allow scientists to study

gene expression on a larger scale and ways to edit the genome with unprecedented precision using CRISPR.

METHODS OF MOLECULAR GENETIC ANALYSIS BASED ON DNA REPLICATION PROCESS

1. Methods of Molecular Genetic Analysis based on the DNA replication process

Learning Objectives

LEVEL 1 and 2

- List the 5 chemical components of a PCR reaction and describe their roles.

- List the functions of the 3 temperature cycles which are repeated during a PCR reaction.
- Describe the process of observing results and interpreting the results of a PCR experiment.
- List possible uses of PCR in genetic testing and in research.
- Explain the Quantitative or Real Time PCR method and differences with regular PCR

⊗ **LEVEL-UP**

- Design appropriate forward and reverse primer pairs when given a gene sequence.
- Design a PCR based diagnostic test.
- Interpret electrophoresis results by distinguishing DNA fragments by length and determining whether individuals are homozygous or heterozygous at different STR loci.
- Interpret qPCR data to evaluate differences in gene expression.

The Polymerase Chain Reaction (PCR) an in-vitro method for amplifying DNA and the chemistry to sequence DNA; two cornerstone techniques that are widely used in molecular and recombinant DNA technology were developed directly

from an understanding of enzymes used in the process of DNA replication.

1.1 Polymerase Chain Reaction

Developed in 1983 by Kary Mullis, PCR is a common technique used in medical and biological research labs for a variety of applications including the following:

- DNA cloning for sequencing; DNA-based phylogeny, or functional analysis of genes
- The diagnosis of hereditary diseases
- The identification of genetic fingerprints (used in forensic sciences and paternity testing)
- The detection and diagnosis of infectious diseases

In Vitro vs. In Vivo Replication

PCR is an *In Vitro* process that mimics cellular DNA replication in the test tube, repeatedly copying the target DNA over and over, to produce large quantities of the desired DNA. This laboratory technique is modeled after the *In vivo* process. However, unlike all the various components needed to replicate DNA discussed in this chapter, PCR works by using just one of these enzymes- DNA polymerase.

Mullis imagined a chemical reagent and a temperature change step in the method that could perform the work of the initiation and eliminate the need for helicases. It should be noted that because Dr. Kerry Mullis had learned about the details of *in vivo* DNA replication, he could create this science-changing *in vitro* method.

1.1.1 Components of PCR

A basic PCR setup requires the following components and reagents:

1. **The DNA template** is a sample of DNA that contains the DNA region (target) to be amplified.

2. **Two primers.**

Primers are short **oligonucleotides of DNA**, usually around 20 base pairs in length. Because the purpose of PCR is to amplify a specific section of DNA in the genome, such as a known gene, primers of specific sequences must be used. The experimenter will design a **forward primer** to bind to one strand and a **reverse primer** that complements and binds to the other strand.

The primer design process to select forward and reverse primers is important in designing the PCR and example exercises are provided later in this reading.

3. **Thermostable DNA polymerase** (to carry out the synthesis). The enzyme must have a good activity rate around 75°C. Second, the enzyme should be able to withstand

temperatures of 95-100°C without denaturing and losing activity.

One of the best-known thermostable DNA polymerases was first isolated from bacteria that grow in hot springs, such as those found in Yellowstone National Park, *Thermus aquaticus*. Named after the species, Taq polymerase is DNA polymerase I that retains polymerising activity even at the high temperatures needed for melting the templates, and it is active at a temperature between the melting and annealing temperature.

Since then many better-performing DNA polymerases have been isolated with varying levels of efficiencies from other bacteria and archaea.

4. dNTPs (DNA nucleotides to build the new DNA strands). DNA polymerase will add each **complementary** base to the new growing DNA strand according to the original strand's sequence following normal A-T and C-G pairings.

5. Finally, a reaction buffer. Buffer solution provides a suitable chemical environment for optimum activity and stability of the DNA polymerase.

1.1.2 The Three Steps

A key insight to the success of PCR as an in vitro DNA replication process which generates millions of specific sequence copies was a three-temperature cycle that accomplishes three parts of DNA replication.

Denaturation of the double-stranded template, **annealing** of the primers to the single strands, and **extension** of new strand synthesis by DNA pol III.

Before reading the description of PCR steps and how the components are used, watch the video below to help you visualize the importance of each step.

LINK to LEARNING

This [video animation](#) on LabExchange animation illustrates the process of PCR.

Denaturation step: This step is the first regular cycling event and consists of heating the reaction to 94-98°C, near-boiling! At this temperature, all double-stranded DNA is “melted” into single strands

[Step 1 in Figure 1.1]

Annealing step: The reaction temperature is lowered to 50-65°C (range varies) This allows the primers bind to **complementary** sequences in the single-stranded DNA template the region of interest

The two primers are called the **forward** and the **reverse**

primer and are designed because their sequences will target the desired segment of the DNA template for replication.

[Step 2 in Figure 1.1]

When planning the PCR analysis the experimenter “designs” chooses the appropriate sequences for the forward and reverse primers *and then buys them from a vendor who can synthesize single-stranded DNA that has a specific sequence and length.*

The two most important criteria for primer design are the following.

1. One primer must have a sequence that complements one of the template strands and the other primer must be complementary to the other strand. **BOTH strands need to be primed for PCR!**
2. The primers must bind so that their 3' ends are ‘pointing’ in the direction of the other primer. primer. This ensures that the sequence between the primers is replicated in the PCR cycles.

Designing primers for a PCR reaction is an important step because the way they bind to the template DNA dictates much of the success of the PCR reaction. The interactive link below presents the rules for designing precise nucleotide sequences called PCR primers.

LINK to LEARNING: Primer Design

Complete [this interactive on Lab-Exchange](#) to learn about primer design.

Primers are present in millions of fold excess over the template. This is important because each newly made DNA strand starts from a primer. Since the primers are present in great excess, the complementary sequences they target are readily found and base-paired to the primers.

Polymerase chain reaction - PCR

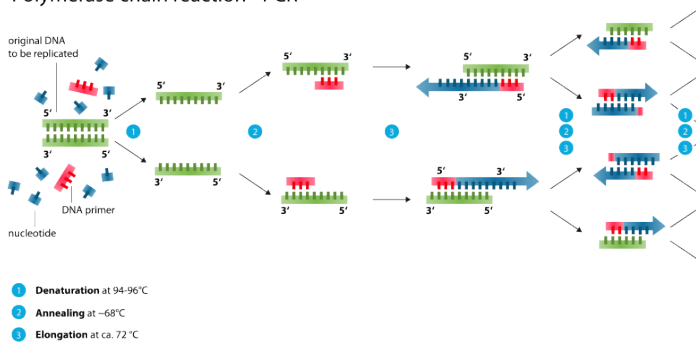


Figure 1.1 Polymerase Chain Reaction Steps. Image from [Wikimedia commons, CC-BY-SA](https://commons.wikimedia.org/wiki/File:PCR_Steps.png)

Extension: The final PCR step is when the **DNA polymerase enzyme** reads the template and connects new nucleotides to the primer's 3' end, extending a new complementary strand of DNA. The test tube is heated to around 72-75°C, the optimal temperature for the polymerase to operate. DNA pol. III activity and the newly synthesized DNA strand is extended as the template strand is read by DNA pol. III.

[Step 3 in Figure 1.1]

There are now twice as many copies of your gene of interest as when you started!

Because thousands of copies of the forward and reverse primer are added at the start of PCR, all the single strand templates, both the original, the copies in cycle 3 and beyond,

and the copies of the copies made from previous cycles will be primed for the extension step of the cycles.

Thermocycling in a typical PCR amplification is illustrated below.

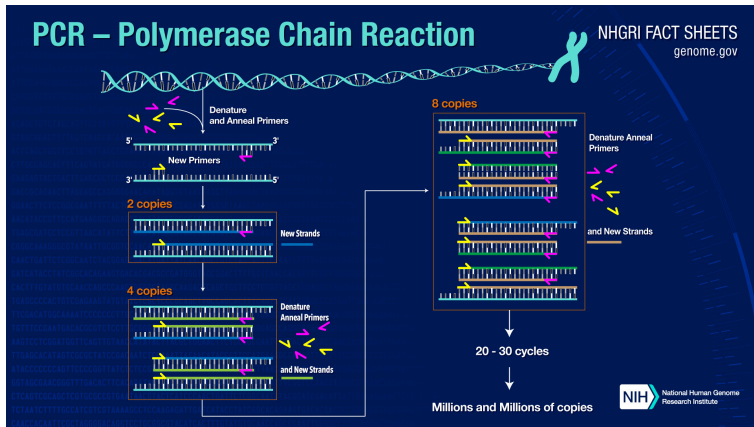
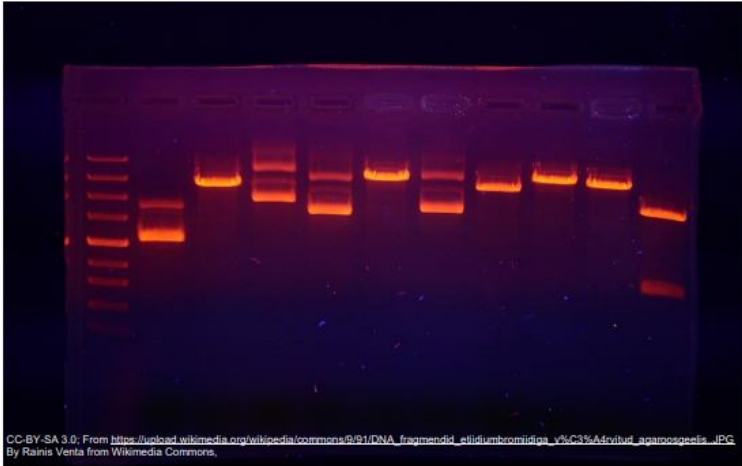


Figure 1.2 This image shows how a single piece of DNA is replicated to create many copies through a process called the polymerase chain reaction (PCR). This image is from [National Human Genome Research Institute](https://www.genome.gov) which is in the Public Domain.

You can see from the illustration that the second cycle of PCR has generated the two DNA strands that will be templates for doubling and re-doubling the desired product after each subsequent cycle. A typical PCR reaction might involve 30 PCR cycles, resulting in a nearly exponential amplification of the desired sequence

Visualizing the Results with Electrophoresis

Once a PCR reaction has been completed, we need to be able to see the results. To do this, a sample of the **PCR** mixture is loaded into an agarose gel for electrophoresis.



DNA fragments in agarose gel stained with ethidium bromide. DNA and ethidium bromide form a complex that emits orange light if placed under UV. Image Attribution: [Rainis Venta, CC BY-SA 3.0](https://creativecommons.org/licenses/by-sa/3.0/) <<https://creativecommons.org/licenses/by-sa/3.0/>>, via Wikimedia Commons

Did I Get This?

Apply concepts of primer design by completing this activity



An interactive H5P element has been excluded from this version of the text. You can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1155#h5p-44>

1.2 Variations of PCR

1.2.1 Quantitative PCR (qPCR) or Real-time PCR

qPCR or Real-Time PCR follows the same principle of amplification as regular PCR (exponential amplification), however instead of needing gels to visualize the end product the process is monitored in “real-time”, as suggested by the name.

A machine “watches” the reaction occur with a camera or detector. There are many techniques to allow monitoring of

PCR in real-time but they all have one thing in common: linking the amplification of DNA to the generation of fluorescence. Thus as there is more DNA, there should be more fluorescence.

The ability to monitor in real-time and measure accumulation of fluorescence means unlike regular PCR, this method can provide truly quantitative analysis of gene expression. Hence the other name for the same technique- Quantitative PCR (qPCR)

Methods used for generating fluorescence:

1) Use of an intercalating agent that binds only to double-stranded DNA.

A common one is Sybr Green. The principle is simple:

- the dye itself is fluorescent
- it can bind dsDNA and when it does it alters the structure of the dye making it fluoresce MORE.
- So very simply as the PCR creates more DNA, more dye binds more fluorescence is observed.

2) TaqMan (style) Probes.

More commonly, qPCR is performed using TaqMan [Based on the name of the company and enzyme] technology.

With TaqMan, a **third primer (TaqMan probe)** is designed in the middle of the area to be amplified. This middle primer is designed with a hairpin self-complementarity so that the 5' and 3' ends are in close proximity.

At one end, a **fluorescent reporter** is attached while the other terminus has a **quencher** that absorbs any fluorescence signal. Under normal circumstances, measurements of fluorescence will be very low since the quencher is physically close to the reporter.

Taq Polymerase used in these reactions retains its 5'-3' exonuclease activity. The probe is cleaved by the enzyme during the reaction separating the quencher and reporter. The quencher no longer has its effect on the reporter and the level of fluorescence increases.

This means that with every cycle of PCR more probe is cleaved and more fluorescence is generated.

The name TaqMan is a play on words since it is imagined that the polymerase is chewing up the probe like Pacman. With increased distance between quencher/reporter, the fluorescence signal from this probe can now be measured.

This method is much more specific than Sybr Green. However, the use of specific probes increases the cost considerably.

Watch this video below explaining this technology:



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1155#oembed-1>

Nomenclature Used in Real Time PCR and Data Interpretation.

First take a look at representation of real time data.

Model of real time quantitative PCR plot

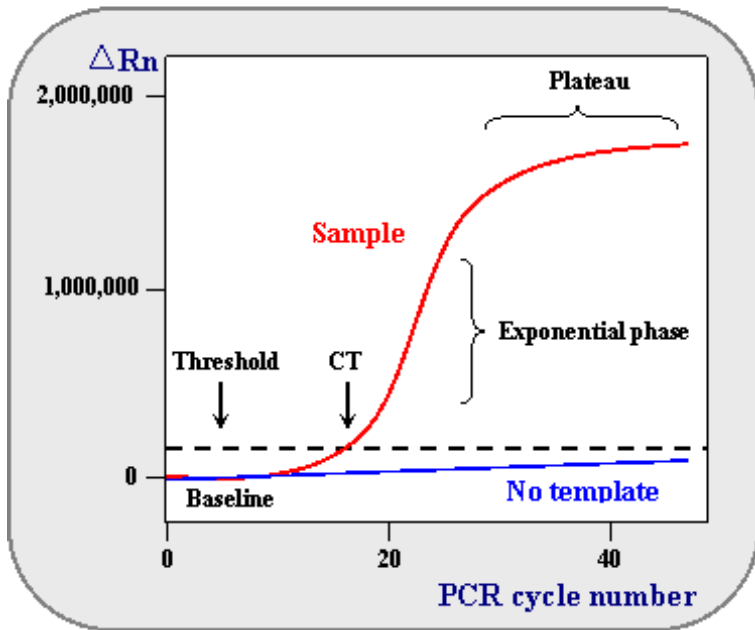


Figure 1.3 Example RT-PCR Plot. Image in Public Domain from <https://www.ncbi.nlm.nih.gov/probe/docs/techqpcr/>

The horizontal axis shows the PCR cycle number, meaning how many cycles of melting, annealing, and primer extension have occurred.

Baseline: Is defined as PCR cycles in which a reporter **fluorescent signal is accumulating** but is beneath the limits of detection of the instrument. Fluorescence measurement early during the PCR process will be very low due to the small number of dsDNA molecules (Sybr Green) or most TaqMan

primers being quenched. At this point remember that PCR cycles are still occurring.

During this exponential DNA production, a threshold will be reached in which the fluorescence will linearly increase.

ΔR_n : Note that values obtained do not have absolute units associated with them. What is represented is an increment of fluorescent signal at each time point. The ΔR_n values are plotted versus the cycle number.

Threshold: The dotted line is an arbitrary level of fluorescence chosen on the basis of the baseline variability. A signal that is detected above the threshold is considered a real signal that can be used to define the threshold cycle (C_t) for a sample. Threshold can be adjusted for each experiment so that it is in the region of exponential amplification across all plots.

C_t value: is defined as the fractional PCR cycle number at which the reporter fluorescence is greater than the threshold. **The C_t is a basic principle of real time PCR and is an essential component in producing accurate and reproducible data.**

The C_t value is instructive when comparing samples or determining relative amount of starting material.

Watch the 2 videos below explaining how to interpret qPCR data.

Link [here](#) or see below.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://iu.pressbooks.pub/iul211smehta/?p=1155#oembed-2>

What are Ct Values?



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://iu.pressbooks.pub/iul211smehta/?p=1155#oembed-3>

1.2.2 Quantitative Reverse Transcription PCR

Reverse transcription–PCR is a sensitive technique for the detection of mRNA. It is routinely used as a diagnostic test for the presence of RNA viruses, such as the agents causing AIDS, or SARsCoV2.

The difference between the regular PCR and Reverse-Transcription PCR is in the first step. Since the starting material is mRNA we need to first convert it into DNA using a special enzyme called reverse transcriptase. We learn more about the details of how the conversion is done later in this unit.

Application in Diagnostics

Quantitative PCR can be used to detect the presence of genetic material from pathogens, like coronaviruses in human cells.

Coupling of **reverse transcription** followed by quantitative real-time PCR [qRT-PCR] and it becomes a powerful quantitative measurement to compare **levels of gene expression** (how much RNA was there to begin with in the sample?).

This concept is also the basis behind using RT- quantitative PCR diagnostic for establishing viral load and/or infectivity for viruses with RNA genomes- like SARS-CoV-2!

1.3 Example Applications of PCR

1.3.1 Forensics- DNA Profiling

You learned in Chapter 3 that the difference in nucleotide sequences between humans lies is minimal. People are greater

than 99% similar. But when you look at your classmates around the room, you can see that that small difference amounts to quite a bit of variation within our species. The bulk of these differences aren't even within the coding sequences of genes but lie outside in regulatory regions that change the expression of those genes.

One such area of variation are the **Short Tandem Repeats (STR)**. Recall that STRs are comprised of units of bases, typically two to five bases long (ex. TAATTAATTAAT). that repeat multiple times. The repeat units are found at different locations, or loci, throughout the genome.

If you were to read a repetitive set of sequences and count the repetition, you would make mistakes and lose count. Likewise, DNA polymerase will make errors or stutter in areas of repetitiveness and produce changes – expanding or contracting the number of repeats.

Any variation of a locus is referred to as an **allele**. In the case of STRs, these alleles are simply a difference in the **number of repeats** surrounded by nonvariable segments of DNA known as flanking regions.. This means the length of DNA within this locus is either longer or shorter based on the number of repeats. This variability within the STR regions can be used to distinguish one DNA profile from another.

For example the TH01 is a locus on chromosome 11 that has a repeating sequence of TCAT. There are reported to be between **3-14 repeats** in this locus.

In the example below the repeat unit [TCAT] repeats 6

times, and along with the flanking region it will have set length in basepairs.

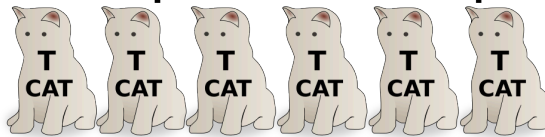
A different allele of this same STR would have a different number of TCAT repeat units but the **same flanking regions**.

Flanking regions are important because knowing their sequences enables geneticists to isolate the STR using polymerase chain reaction, or PCR, amplification.

The TH01 locus contains repeats of TCAT.

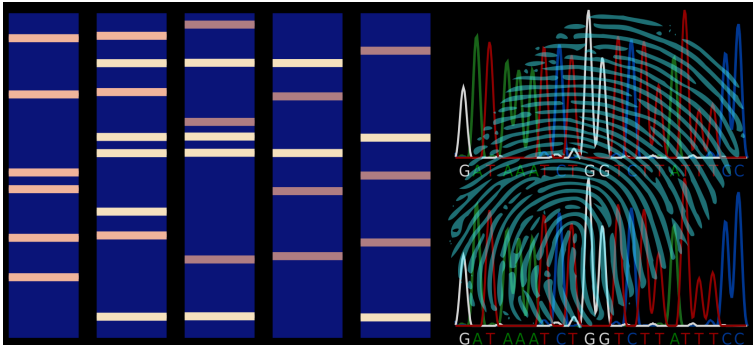
CCC **TCAT TCAT TCAT TCAT TCAT TCAT** AAA

This example has 6 TCAT repeats.



TH01 STR: Outside of the STR, there are flanking areas of known sequence. The primers that amplify TH01 in PCR recognize these flanking sequences to amplify the TCAT repeats.
Credit: Jeremy Seto (CC-BY-NC-SA)

CoDIS



DNA fingerprinting. Credit: [Helixitta](#), [The Photographer](#) and [Jeremy Soto](#) (CC-BY-SA 3.0)

The FBI and local law enforcement agencies have developed a database called the Combined DNA Index System (**CoDIS**) that gathers data on a number of STRs.

By establishing the number of repeats of a given locus, law enforcement officials can differentiate individuals based on the repeat length of these alleles.

CoDIS uses a set of 20 loci that are tested together. As you would imagine, people are bound to have the same alleles of certain loci, especially if they were related. The use of 20 different loci makes it statistically improbable that 2 different people could be confused with each other.

Think about this in terms of physical traits. As you increase the number of physical traits used to describe someone, you are less likely to confuse that person with someone else based on those combinations of traits.

Using the CoDIS loci increases the stringency since there are many alleles for each locus.



One or more interactive elements has been excluded from this version of the text. You can view them online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1155#oembed-4>

How DNA Profiling is Done

At a crime scene, scant evidence in the form of a few cells found within bodily fluids or stray hairs can be enough to use as DNA evidence. DNA is extracted from these few cells and amplified by PCR using the specific primers that **flank the STRs used in CoDIS**.

Amplified DNA will be separated by gel electrophoresis and analyzed. Size reference standards and samples from the crime scene and the putative suspects would be analyzed together.

In a paternity test, samples from the mother, the child, and the suspected father would be analyzed in the same manner. A simple cheek swab will supply enough cells for this test.

Typically as mentioned above many STRs are analyzed at once during genetic profiling.

Watch this lecture video to learn about the how this is carried out.

[DNA Profiling Lecture Video](#)

This method while most well known for use in crime scenes has a wide-variety of applications beyond that- from establishing paternity, matching organ donors to recipients, identify victims of catastrophes (Tsunami and earthquakes), identify protected and endangered species to aid in conservation biology.

1.4 DNA Sequencing

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 developed by Fred Sanger in the 1970s.:

- the use of **di-deoxy base** incorporation in a **polymerization** reaction
- leads to termination of primer extension

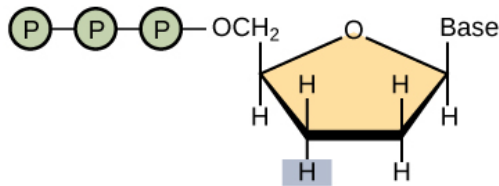
Principle of Sanger Sequencing

It helps to understand the principle of sequencing by learning how it used to be carried out when first discovered.

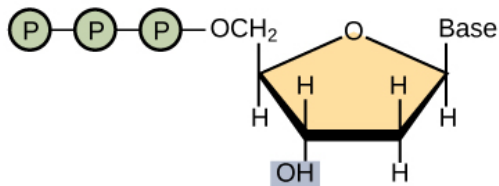
The basic steps involved

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.

The ddNTPs are monomers that are missing a hydroxyl group (–OH) at the site at which another nucleotide usually attaches to form a chain. When a dideoxynucleotide is incorporated into a DNA strand, DNA synthesis stops.



Dideoxynucleotide (ddNTP)



Deoxynucleotide (dNTP)

Figure 1.4 A dideoxynucleotide is similar in structure to a deoxynucleotide, but is missing the 3' hydroxyl group (indicated by the box). Image from: <https://openstax.org/details/books/biology-2e>

Watch [this visual](#) demonstrating chain termination

4. Four PCR mixes are set up, each containing stocks of normal nucleotides plus one dideoxynucleotide (ddA, ddT, ddC or ddG) that was **made radioactive**.

5. As a typical PCR will generate over 1 billion DNA molecules, each PCR mix should generate all the possible terminating fragments for that particular base, especially if only a small amount of ddNTP is used.

This results in multiple short strands of replicated DNA that are each terminated at a different point during replication.

6. When the fragments are separated using gel electrophoresis, the multiple newly replicated DNA strands form a ladder because of the differing sizes. Because the **ddNTPs are labeled**, each band on the gel reflects the size of the DNA strand **and** the ddNTP that terminated the reaction. The base sequence can be determined by ordering fragments according to length and ‘reading’ the gel or ladder from bottom up.

Click through the interactive at the link below that explains the process of Sanger Sequencing

[Sanger Sequencing DNA Learning Center](#)

This process has since been automated eliminating the need for high amounts of radioactivity or dealing with cumbersome gels. Instead the ddNTPs are fluorescently labeled in different colors.

Template DNA could be sequenced **in a single tube**, containing all the required components, including *all* four dideoxynucleotides! That’s because the fluorescence detector

in the sequencing machine separately sees all the short ddNTP-terminated fragments as they move through the electrophoretic gel.

After the sequencing reactions, the reaction products are electrophoresed on an ‘automated DNA sequencer’. UV light excites the migrating dye-terminated DNA fragments as they pass through a detector. The color of their fluorescence is detected, processed and sent to a computer, generating color-coded graph known as **Chromatogram** like the one below, showing the order (and therefore length) of fragments passing the detector and thus, the sequence of the strand

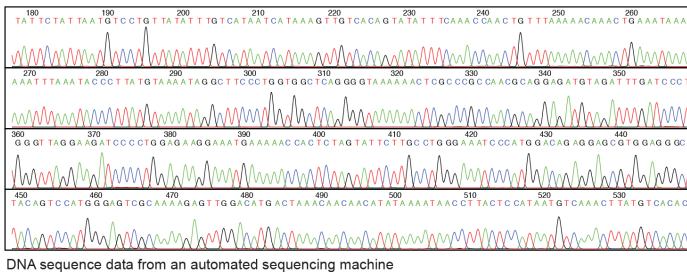


Image from <https://www.genome.gov/genetics-glossary/DNA-Sequencing> in Public Domain.

Second Generation and Next-generation Sequencing

The Sanger method of sequencing DNA is often called “first-

generation” sequencing because they were the first to be developed. In the late 1990s, new methods, called second-generation sequencing methods, that were faster and cheaper, began to be developed.

We will learn about these newer techniques later in the semester as they have become useful tools for analyzing genome wide gene expression and studies into regulation of gene expression.

DNA Sequencing Lecture Videos: [Dr. Mehta](#) or [Sanger Sequencing with Simulation](#)

Key Takeaways

- The lack of the second deoxy group on an dNTP making it ddNTP, stops the incorporation of further nucleotides, this

termination creates DNA lengths stopped at every nucleotide, this is central to further identifying each nucleotide.

- Using fluorescent labels, dideoxy sequencing can be automated allowing high-throughput methods which have been utilized to sequence entire genomes.

Terms:

- **chromatogram:** The visual output from a chromatograph. Usually a graphical display or histogram.
- **dideoxynucleotide:** Any nucleotide formed from a deoxynucleotide by loss of an a second hydroxy group from the deoxyribose group

References and Attributions

This chapter contains material taken from the following CC-licensed content. Changes include rewording, removing

paragraphs and replacing with original material, and combining material from the sources.

1. Bergtrom, Gerald, “Cell and Molecular Biology 4e: What We Know and How We Found Out” (2020). *Cell and Molecular Biology 4e: What We Know and How We Found Out – All Versions*. 13.

https://dc.uwm.edu/biosci_facbooks_bergtrom/13

2. Works contributed to LibreTexts by Kevin Ahern and Indira Rajagopal. LibreTexts content is licensed by CC BY-NC-SA 3.0. The entire textbook is available for free from the authors at <http://biochem.science.oregonstate.edu/content/biochemistry-free-and-easy>

3. “Amplifying DNA – The Polymerase Chain Reaction” by LibreTexts is licensed under CC BY-SA.

4. “Analysis of STRs” by LibreTexts is licensed [CC BY-NC-SA](https://creativecommons.org/licenses/by-nc-sa/3.0/). <https://bio.libretexts.org/@go/page/72191>

RECOMBINANT DNA TECHNOLOGY

Learning Objectives

When you have mastered the information in this chapter, you should be able to:

1. Describe the characteristics of plasmids/ cloning vectors
2. List in words or indicate in a drawing the important features of a plasmid vector that are required to clone a gene. Explain the purpose of each feature.
3. Explain how vectors are used in cloning a gene.
4. Explain what are expression vectors and how they are used.

5. Explain how to use restriction enzymes to create a recombinant plasmid.
6. What properties of the DNA restriction fragments enable ligation of these fragments?
7. Explain what role do restriction enzymes have in nature?
8. Explain how DNA ligase is used to create a recombinant plasmid
9. Describe possible recombinant plasmids that form when ligating a restriction digest
10. Design a cloning experiment/strategy when given vector sequences and insert sequences
11. Predict or troubleshoot the reasons for a 'failed' cloning experiment.
12. Due to a mishap in the lab, bacteria carrying a plasmid with a kanamycin-resistant gene and bacteria carrying a plasmid with an ampicillin-resistance gene were accidentally mixed together. How would you design an experiment allowing you to sort out the two kinds of bacteria?
13. Explain the use of reporter gene assays.
14. Outline an experiment using reporter gene assays to identify regulatory elements in promoter sequences of genes (prokaryotic

and eukaryotic)

Introduction

It is likely you know somebody who has diabetes (a disease that occurs when a person's blood glucose [sugar] is too high). In certain types of diabetes, the body is unable to manufacture or produce insulin, the protein hormone needed to manage blood glucose levels. In [Chapter 1](#) and [Chapter 2](#) you learned about the breakthrough research that led to the isolation of insulin, first from dogs and subsequently from pancreases of cows and subsequent demonstration that purified insulin when injected in humans was able to treat the symptoms.

By 1923, insulin had become widely available in mass production, and Banting and Macleod were awarded the Nobel Prize in medicine. Diabetes was no longer a life-shortening disease. While these methods were effective, using animal-produced proteins sometimes resulted in adverse reactions, and making large enough quantities was difficult. This all changed with the advent of **recombinant DNA technology**.

This technology allowed scientists to take human DNA (for

example insulin gene) and introduce it into bacterial DNA, allowing the bacteria to produce a human protein.

The ability to make **recombinant DNA** is such a seminal technology that just realizing it could be done and then doing it in a test tube for the first time earned Paul berg a half-share in the 1980 Nobel Prize in Chemistry (the other half was shared by Walter Gilbert and Frederick Sanger for studies that enabled efficient **DNA sequencing**).

Molecular Cloning and Recombinant DNA Technology

Joining together of DNA fragments from different sources creates **recombinant DNA**. The ability to cut and paste DNA might seem like purely a technical feat, but one key application that arose out of this is molecular cloning.

Typically cloning is referring to ‘molecular cloning’, the process of making multiple copies of a particular sequence of DNA, **expression of genes, and study of specific genes**.

In molecular cloning a gene of interest can be inserted into a vector, usually a **plasmid**, by cutting both the vector and the gene (called the insert) with the same enzyme to generate sticky ends and joining the two pieces together to generate a recombinant.

A plasmid is a type of autonomously replicating,

extrachromosomal DNA. It is quite simple to extract plasmids from the cells, engineer them to contain the gene of interest and re-introduce the recombinant plasmid into the bacteria. The idea was that when the plasmid DNA was replicated, the extra inserted gene would also be copied. Thus, by growing up a lot of the bacteria carrying the plasmid, many copies of the gene of interest could be obtained, to provide sufficient amounts of the gene to use in experiments.

How do you begin to 'clone' a gene or make a recombinant DNA molecule?

First, let's think about what 'cloning' requires conceptually.

1. We need to isolate the gene/ DNA sequence (for example, a gene for a human therapeutic protein) away from the rest of the genome. This would involve cutting DNA in a reproducible manner and procedures for viewing the DNA fragments.
2. We need a way to 'paste' this DNA into a vector (a different DNA molecule) to create a recombinant molecule
3. This vector should be able to replicate inside a host cell.
4. We need procedures for *introducing the recombinant DNA molecule* into the appropriate cell (eukaryotic or

prokaryotic)

Key developments that allowed for the revolution in recombinant DNA technology came from studying prokaryotic biology.

These include the identification of DNA cutting enzymes “restriction endonucleases”, isolation of small extrachromosomal DNA found in bacteria called ‘Plasmids’ and knowledge that bacteria can take up DNA from exogenous sources- ‘transformation’.

Digestion of DNA with restriction endonucleases, pasting with ligases

The creation of recombinant DNA molecules is possible due to the use of naturally occurring **restriction endonucleases** (**restriction enzymes**), bacterial enzymes produce as a protection mechanism to cut and destroy foreign cytoplasmic DNA that is most commonly a result of bacteriophage infection. Stewart Linn and Werner Arber discovered restriction enzymes in their 1960s studies of how *E. coli* limits bacteriophage replication or infection (**Read Link to learning**).

Today, we use restriction enzymes extensively for cutting DNA fragments that can then be pasted with another DNA molecule to form recombinant molecules. These enzymes are purified from various bacterial species and can be purchased

commercially. They are named after the bacterium from which they were first isolated. For example, ***EcoRI*** and *EcoRV* are both enzymes from *E. coli*. HindIII came from the bacteria *Haemophilus influenzae*.

Restriction enzymes serve as molecular scissors recognizing a specific sequence in the DNA called restriction sites. The sites are usually palindromic, typically between four to six base pairs in length.

Palindromic sequences are the same sequence forwards and backward. Some examples of palindromes: RACE CAR, CIVIC, A MAN A PLAN A CANAL PANAMA.

With respect to DNA, there are 2 strands that run antiparallel to each other. Therefore, the reverse complement of one strand is identical to the other. This means that the sequence of the recognition site when *read in the 5' to 3' direction* for the top strand is exactly the same as that of the bottom strand. Upon binding this sequence, the enzyme cuts both strands of the DNA generating a double-stranded break.

However, based on where the enzyme cuts different kinds of ends are generated. Some enzymes cut asymmetrically leaving a overhang at the 5' end of one strand of the duplex; some leave an overhang at the 3' end. These ends are referred to as “**sticky ends**”. (Figure 2B)

Some enzymes cut symmetrically leaving no overhanging sequence overhangs and generate “**blunt ends**”(Figure 2A)

This distinction in cutting is important because “sticky ends” are thus called because they are self-complementary!

Molecules with **complementary** or **compatible** sticky ends can

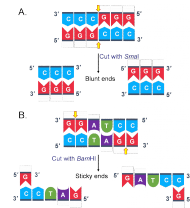


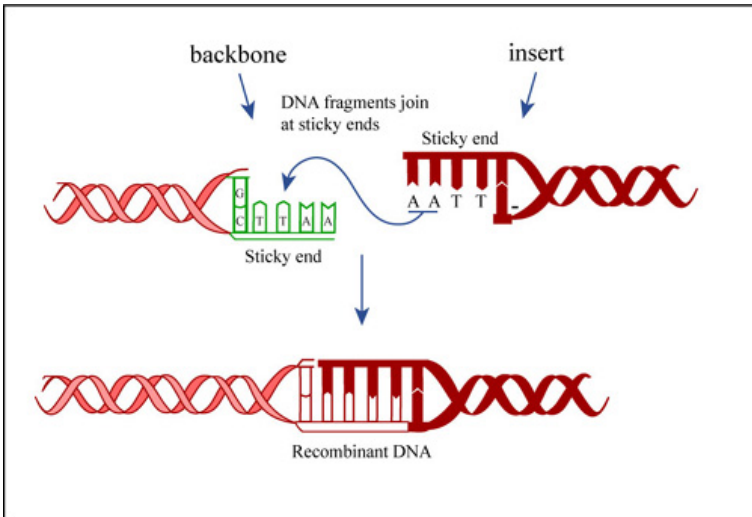
Figure 2.1
Example of how restriction enzymes cut DNA. (A) Treating the DNA with SmaI results in fragments with blunt ends. (B) Whereas treatment with BamHI produces fragments with “cohesive” or “sticky” ends. Image by Walter Suza. From: Genetics, Agriculture, and Biotechnology by Walter Suza; Donald Lee; Marjorie

easily **anneal (or stick to one another)** forming hydrogen bonds between complementary bases, at their overhangs.

With the help of an enzyme we are already familiar with, **DNA Ligase** the 2 pieces of DNA are joined together. through covalent bonding, making the molecule a continuous double strand.

With a combination of RE and DNA ligase, any two fragments regardless of their origin (animal, plant, fungal, bacterial) can be joined in vitro to form recombinant molecules, especially if they have compatible sticky ends.

Hanneman;
and Patricia
Hain is
licensed under
a Creative
Commons
Attribution-No
nCommercial-S
hareAlike 4.0
International
License



Lecture Activities

Some of the links below will be helpful in understanding how Restriction Enzymes Work.

1. [Watch this DNA Learning Center Animation on 'Cutting and Pasting DNA'](#). **Note the 3' OH and 5' Phosphates of the overhangs created.**
2. Watch **this video up to time stamp 1:24** to learn how to distinguish between 3' and 5' overhangs after restriction enzyme cutting.



One or more interactive elements has been excluded from this version of the text. You can view them online here: <https://iu.pressbooks.pub/iul211smehta/?p=1187#oembed-1>

3. Go through [this scrollable interactive on LabExchange](#). It demonstrates how restriction enzymes work and why they can be used as tools to analyze DNA and create recombinant plasmids.

TRY THIS ACTIVITY:



An interactive H5P element has been excluded from this version of the text. You can view it online here: <https://iu.pressbooks.pub/iul211smehta/?p=1187#h5p-45>

Link to Learning: Discovery of Restriction Endonucleases.

Why would bacteria have enzymes that could cut up DNA?

The discovery of this special class of enzymes came from early work in the 1950s on certain strains of bacteria. Scientists observed that certain strains of *E. coli*, a common bacterium found in the human gut, were resistant to infection by **bacteriophage**—a virus that infects bacteria by injecting its DNA into the cell and commandeering the host cell's molecular processes to make more bacteriophage. This peculiar (at the time) phenomenon was called 'Host-Restriction'. 1960s, Werner Arber was investigating the possible mechanisms and observed a dramatic change in the **bacteriophage DNA** after it invaded these resistant strains of bacteria: the bacteriophage DNA was degraded and cut into pieces! To explain the resistance of certain bacterial strains to bacteriophage infection, Arber hypothesized that bacteriophage-resistant bacterial cells might express a specific enzyme that degrades *only invading bacteriophage DNA, but not their own DNA*. Further investigation of this primitive bacterial "immune system" led to the discovery of restriction enzymes, proteins that recognize cut DNA at specific sequences called **restriction sites**. Thus, they restrict the growth of bacteriophages by recognizing and destroying the phage DNA. Read the following article to learn how bacteria protect their own DNA from being chewed up.

(<https://www.nature.com/scitable/spotlight/restriction-enzymes-18458113/>)

Plasmids

Bacteria carry a single circular genome (chromosome) that carries all the information needed by the organism to survive and reproduce. However, in addition, many bacteria also carry small circular DNA molecules, ranging in size from 1,000 to 200,000 *base pairs* present in multiple copies separate from the chromosomal DNA called Plasmids.

Plasmids carry some useful features:

1. They have their own **origin of replication**, which allows the plasmid to produce many copies (replicate) inside of a bacterial cell. Additionally, as mentioned above they replicate at a higher frequency than the bacterial chromosome, which means there could be thousands of copies of plasmid within a single cell.
2. They have a promoter for transcription of an inserted gene if so desired.
3. They carry a gene for antibiotic resistance, that provides for a **mechanism to select** for a small percentage of bacterial cells that have taken up the plasmid

Scientists have taken advantage of plasmids to use them as

tools to clone, transfer, and manipulate genes. Plasmids that are used experimentally have been repurposed to make them suitable for cloning and called **vectors**.

Often the terms plasmid and vector are used interchangeably.

Vectors (plasmids) used by scientists today come in many sizes and vary broadly in their functionality. The greatest variety of cloning vectors has been developed for use in the **bacterial host *E. coli***.

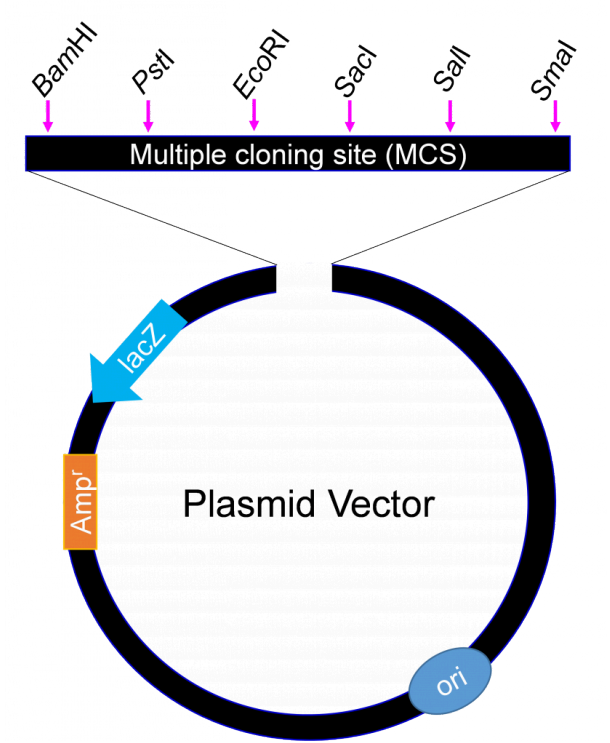
They all however must have *at least* three features

- 1) An origin of replication
- 2) A selectable marker (antibiotic resistance gene)

Cells that take up the plasmid will be able to grow in the presence of the antibiotic. If bacterial cells to which the plasmid has been added are plated on agar containing the antibiotic, the cells which took up the plasmid will be able to grow, while the others will not.

3) A cluster of restriction enzyme sites which called a **Multiple cloning site (MCS)**.

The MCS enables researchers to introduce any piece of DNA into the vector by restriction enzymes digestion and ligation. As a general rule, the restriction sites in the MCS are unique and not located elsewhere in the plasmid backbone, which is why they can be used for cloning by restriction enzyme digestion but when designing a cloning experiment a detailed 'Restriction Map' can be obtained".



**Watch this video
embedded below or go to
[this Youtube Link from
Addgene](#)**

Figure 2.2 Basic map of a plasmid vector. The MCS contains several restriction enzymes sites and few are shown here as examples. Image by Walter Suza. From: [Genetics, Agriculture, and Biotechnology](#) by Walter Suza; Donald Lee; Marjorie Hanneman; and Patricia Hain is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](#),



*One or more
interactive
elements has*

*been excluded from this version of the text. You can
view them online here: [https://iu.pressbooks.pub/
iul211smehta/?p=1187#oembed-2](https://iu.pressbooks.pub/iul211smehta/?p=1187#oembed-2)*

Building a recombinant plasmid: The workflow

Below is the general workflow of how molecular cloning takes place.

Step 1. Isolate DNA of interest (called insert in a cloning experiment)

Step 2. Identify cloning vector/plasmid

Step 3. Set up restriction digests for your insert and plasmid

– once the target and vector DNA have been identified, both types of DNA are cut using **restriction endonucleases**.

Step 4. Purify the cut fragments: Unfortunately you can't just throw the digestion mixtures together. You need to isolate your insert and backbone from the enzymes used to digest them as well as any pieces cut out or not needed!

An easy way to do this using a technique you are already familiar with- Agarose Gel Electrophoresis. You simply run the cut vector, insert and appropriate controls on an Agarose Gel, isolate the desired fragment and purify the DNA using special procedures that remove the 'gel' material and leave only the nucleic acid to carry forward to the next step

[Note: Go back to Chapter 3 and watch The Visual Simulation of Gel Electrophoresis that shows restriction digest analysis!]

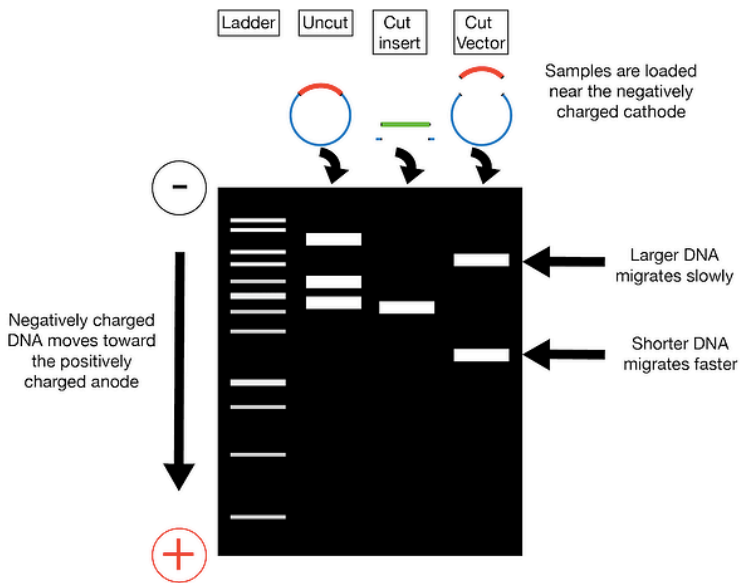


Image from <https://blog.addgene.org/plasmids-101-restriction-cloning>.

Step 4. Combine purified target and vector DNA – the two types of DNA are combined together in a single tube with the addition of DNA ligase and appropriate buffer. This results in the creation of **recombinant DNA**.

Step 5. Introduce recombined molecule into host cell

Once the target DNA has been stably combined with vector DNA, the recombinant DNA must be introduced into a host cell, in order for the genes to be replicated or expressed. There are different methods for introducing the recombinant DNA, largely depending upon the complexity of the host organism.

In the case of bacteria, **transformation** is often the easiest

method, using competent cells to pick up the recombinant DNA molecules. Alternatively, **electroporation** can be used, where the cells are exposed to a brief pulse of high-voltage electricity causing the plasma membrane to become temporarily permeable to DNA passage.

Step 6. Grow the bacteria and plate on agar plates that contain the antibiotic. Bacterial cells carrying the plasmid will grow, multiply and form visible colonies.

The steps involved in molecular cloning using bacterial transformation are outlined in this graphic flowchart.

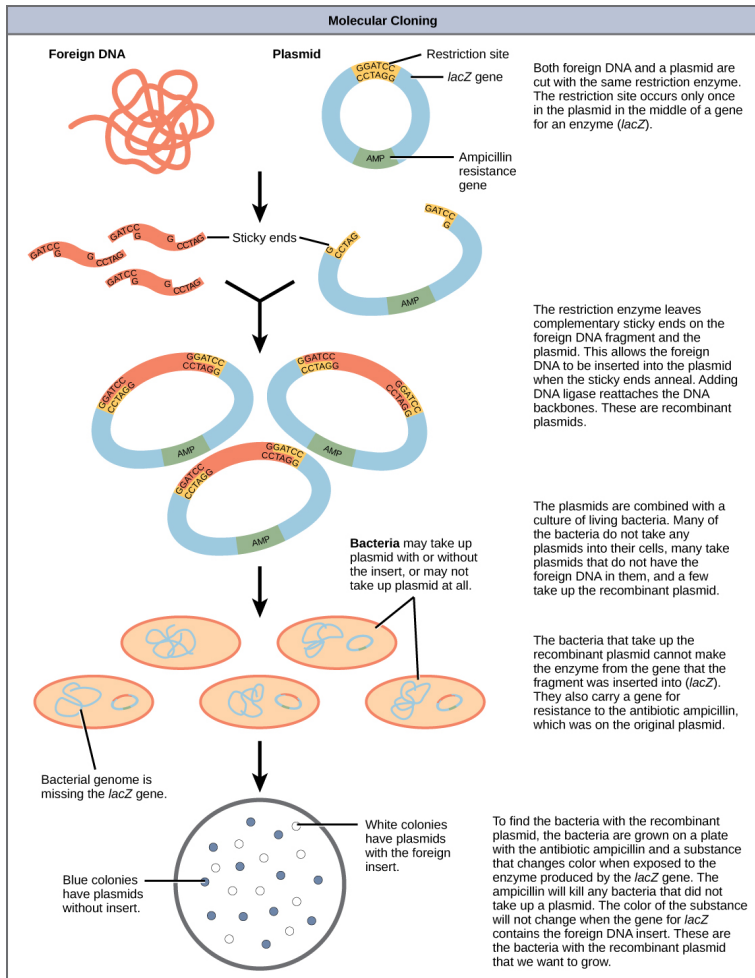


Image from: Concepts of Biology; Access for free Book URL: <https://openstax.org/books/concepts-biology/pages/1-introduction>.

Cloning with Restriction Enzymes-Some Considerations

When designing a cloning experiment where you want to cut out an insert from one source such as genomic DNA and insert it into a second source such as a plasmid (vector), there are a few considerations.

- 1) The restriction sites should occur in the desired location in your recipient plasmid (usually in the Multiple Cloning Site (MCS) but not elsewhere in the vector.

- 2) They should be on either side of your insert, but do not cut within your insert!

- 3) Using 2 different enzymes makes cloning more efficient preventing the plasmid from reforming a circle without the inserted gene. Importantly it allows the inserted gene to be oriented in the correct position for transcribing the “sense” strand of DNA (the strand that codes for the protein)!

As we learned in Introduction to Transcription, the strand of DNA that has the coding sequence of the gene (5'-3' direction) is called the Sense strand. Insert being in the correct orientation in the recipient plasmid is necessary – you don't want to express the antisense version of your gene!

Here is an example where I walk through the logic of picking restriction enzymes to use.



An interactive H5P element has been excluded from this version of the text. You

can view it online here:

<https://iu.pressbooks.pub/iul211smehta/?p=1187#h5p-46>

Cloning by PCR

While we mostly focused on restriction enzyme cloning. Another convenient way to isolate DNA sequence of interest is Polymerase chain reaction (PCR)! The basic steps in PCR reaction were discussed [earlier](#).

This by far the most common method of cloning. In this method when designing primers for your region of interest researchers introduce restriction site to the 5' end of the primers.

Then upon amplification the PCR product will have a restriction site added to either end that can be digested with the appropriate restriction enzyme. 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**.

Click here for [Dr. Mehta Lecture Videos on Recombinant DNA Technology](#)

Types of Plasmids/Vectors

There are many different types of plasmids or vectors for use in molecular biology experiments. The choice of vector depends on the biological question or ultimate use of the cloned DNA. Ones you will encounter in this course or are

General Purpose Vectors: Used to facilitate the cloning of DNA fragments. Cloning vectors tend to be very simple, often containing only a bacterial resistance gene, origin of replication, and MCS. They are small and optimized to help in the initial cloning of a DNA fragment.

Genome Engineering Plasmids: Used for editing and modifying genes using CRISPR, which we shall see in Genome Editing chapter.

Expression Vectors: Used for gene expression (for the purposes of gene study). Expression vectors must contain a promoter sequence, a transcription terminator sequence, and the inserted gene. The promoter region is required for the generation of RNA from the insert DNA via transcription.

The terminator sequence on the newly synthesized RNA signals for the transcription process to stop.

Many sophisticated variations on such vectors have been created that have made it easy to produce and purify large amounts of any protein of interest for which the gene has been cloned.

For example: A handy feature in some expression vectors is a sequence encoding an affinity tag either up- or downstream of the gene being expressed. This sequence allows a short affinity tag (such as a run of histidine residues) to be fused onto the encoded protein. The tag can be used to readily purify the protein, as described in the section on affinity chromatography using special beads that bind to His-tag.

Another handy feature is fusing the coding sequence of fluorescent proteins like GFP to the coding sequence of gene of interest and create a recombinant fusion protein. This allows researchers to view the presence of the protein inside a living cell by microscopy. The two genes are under the same promoter element, transcribed as though it is one gene into a single messenger RNA molecule. The mRNA is then translated into protein.

To connect back to protein folding, in these cases, it is important **that both proteins be able to properly fold into their active conformations when fused together** and interact with their substrates despite being fused. GFP protein is small and is a single domain protein. It has been shown to adopt its three- dimensional shape independent of the

addition of extra amino acids in many cases. Creating fluorescent fusion proteins is a very important tool used by cell biologist (see side bar link to learning for stunning live cell microscopy of a 10 day tumor metastasizing)

Link to Learning

To see a 10-day time-lapse of an engineered human micro-tumor forming and metastasizing click here.

More at: <https://www.nikonsmallworld.com/galleries/2021-small-world-in-motion-competition>

In all examples of Expression Cloning, where the goal is to transcribe mRNA and/or also have the host organism (Bacterial or Eukaryotic) make protein cloning; directionality matters!

Another special type of expression vector is a reporter plasmid which we discuss in detail below.

Reporter Plasmids and Reporter Gene Assays: Tool to Study Gene Expression

A common tool for studying the regulation of gene expression by cis- or trans-acting factors, are reporter gene assays.

In these assays, a ‘reporter gene’ acts as a surrogate (reports) for the coding region of the gene under study. Commonly used reporter genes are those that have visually identifiable characteristics usually involve fluorescent and luminescent proteins.

Examples include the gene that encodes jellyfish green fluorescent protein (GFP), which causes cells that express it to glow green under blue light, the enzyme luciferase, which catalyzes a reaction with luciferin to produce light, a

Other genes that serve as good reporters are types of enzymes that themselves may not be visible when made, but upon addition of a substrate generate a color that can be measured.

A common example of such a gene is Lac Z from E.coli which encodes *β -galactosidase*. This enzyme causes bacteria expressing the gene to appear blue when grown on a medium that contains the substrate analog X-gal or cells that express

the gene to stain blue. The choice of reporter gene will often depend on biological question.

Using recombinant DNA technology one or more gene regulatory elements being analyzed are introduced (cloned) ***upstream*** of the coding sequence of the reporter gene. Following introduction of the reporter construct into cells and experimental treatment, expression levels of the reporter gene are monitored by quantifying the reporter protein enzymatic activity.

Dr. Mehta Lecture Video: [L211 Reporter Gene Assays](#)

References and Attributions

This chapter contains material taken from the following CC-licensed content. Changes include rewording, removing paragraphs and replacing with original material, and combining material from the sources.

1. Bergtrom, Gerald, “Cell and Molecular Biology 4e: What We Know and How We Found Out” (2020). *Cell and Molecular Biology 4e: What We Know and How We Found Out – All Versions*. 13.

https://dc.uwm.edu/biosci_facbooks_bergtrom/13

2. Works contributed to LibreTexts by Kevin Ahern and Indira Rajagopal. LibreTexts content is licensed by CC BY-NC-SA 3.0. The entire textbook is available for free from the authors at <http://biochem.science.oregonstate.edu/content/biochemistry-free-and-easy>

3. [Genetics, Agriculture, and Biotechnology](#) by Walter Suza; Donald Lee; Marjorie Hanneman; and Patricia Hain is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](#)

4. OpenStax Microbiology. **Provided by:** OpenStax CNX.
Located at: <http://cnx.org/contents/e42bd376-624b-4c0f-972f-e0c57998e765@4.2>. **License:** [CC BY: Attribution](#). **License Terms:** Download for free at <http://cnx.org/contents/e42bd376-624b-4c0f-972f-e0c57998e765@4.2>