## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Molecular Biology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Properties of Nucleic Acids</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Proteins and Amino Acids</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Replication of Nucleic Acids</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Making an RNA Copy of DNA</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>The Genetic Code for Protein Synthesis</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Making a Genetic Probe</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>The Control Regions of Genes</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>Cloning Genes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overview</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Cutting DNA</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Joining DNA--DNA Ligase</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Host Strains</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Cloning Vectors for <em>E. coli</em></td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Plasmid Cloning Vectors for <em>E. coli</em></td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Transforming Cells with Plasmids</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Selecting Transformed Cells</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Replica Plating for Colony Screening</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Nucleic Acid Probes</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Purifying Plasmid DNA</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Colored Indicators</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Let's Clone!</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Stomatostatin</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Insulin</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>More Advanced Cloning Techniques</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bacteriophage I Vectors</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>Cosmid Vectors</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Finding Genes: DNA Libraries</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>cDNA Libraries</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Genomic Libraries</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>More Cloning</td>
<td>59</td>
</tr>
<tr>
<td>5</td>
<td>Polymerase Chain Reaction (PCR)</td>
<td>64</td>
</tr>
<tr>
<td>6</td>
<td>Insulin and Diabetes</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Glucose Flux</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>Diabetes</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>References and Readings</td>
<td>75</td>
</tr>
</tbody>
</table>
Chapter 1 Introduction

This section of the bioengineering course will introduce you to the molecular biology portion of the course. That is, we will examine the properties of the nucleic acids that constitute genes and then proteins which are the end product of gene action and the main focus of genetic engineering. We will investigate our model, the hormone insulin, its structure and the general role it plays in regulating the blood level of glucose. Individuals who suffer from diabetes either are unable to produce insulin, or have some defect in responding to the hormone signals.

There are many analogies between general engineering principles and the manner in which the body regulates glucose levels:

<table>
<thead>
<tr>
<th>Fluid system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glucose sensor</td>
</tr>
<tr>
<td>-- the beta-cell of the pancreas</td>
</tr>
<tr>
<td>Valves to regulate glucose level</td>
</tr>
<tr>
<td>-- glucose transporters under control by the hormone insulin</td>
</tr>
<tr>
<td>-- glucose release from liver under control by the hormone glucagon</td>
</tr>
<tr>
<td>Steady state glucose blood levels</td>
</tr>
<tr>
<td>-- analogous to a continuous feed bioreactor</td>
</tr>
<tr>
<td>Substrate (glucose) reservoir</td>
</tr>
<tr>
<td>-- the liver is the glucose buffer system of the body</td>
</tr>
</tbody>
</table>
Chapter 2 Molecular Biology

This section is intended to provide a brief overview of the properties of the components of genes (nucleic acids and their chemical constituents) and the end products of gene action (proteins and their amino acid sequences). Additional details will be introduced as needed in subsequent sections. The student should supplement the material given here with the suggested readings.

2.1 Properties of nucleic acids

Nucleic acids (deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are extremely long polymers made up of phosphate-sugar-nitrogenous base (nucleotide) units. The bases found in DNA are adenine and guanine (both purines) and cytidine and thymine (both pyrimidines). Thymine is replaced by uracil in RNA. The nucleotides are linked by 3’ to 5’ phosphodiester bonds. That is, a phosphate group on the 5’ position of the sugar residue becomes linked to the 3’ hydroxyl group of the preceding sugar group on the chain as the long polymer is synthesized.
FIGURE 1. The structure of nucleic acids showing the bases found in DNA and how nucleotides are linked together

The double stranded structure of DNA was worked out in 1953 by Watson and Crick. It is absolutely essential to understand the key concepts behind this structure since all cloning strategies are based upon them.

The double stranded (double helix) molecule consists of two strands wound around a central axis with the bases stacked inside. The order of the strands are in opposite directions (5’ to 3’ in one, 3’ to 5’ in the other). The bases stack together in the center of the helix because they interact with one another via weak hydrogen bonds.
Hydrogen bonds are much weaker than covalent bonds and are continually forming and disassociating. They form between two electronegative atoms, one of which has a covalent hydrogen. The special properties of water are due to hydrogen bonds of the type:

\[
H-O-H \quad \text{Hydrogen bond}
\]

The hydrogen bonds found in nucleic acids are between N and O:

\[
N-H \quad \text{and} \quad O-H
\]

In double stranded nucleic acids, adenine will only form hydrogen bonds with thymine (or uracil)
and cytidine will only form hydrogen bonds with guanine. Thus, the amount of adenine always equals the amount of thymine and the amount of cytidine always equals that of guanine for the DNA of a given species. The per cent G-C (and thus the per cent A-T) in DNA is surprisingly variable from species to species.

![Base pairing between A:T and G:C.](image)

**FIGURE 3.** Base pairing between A:T and G:C.

The base pairs form a flat plane in the interior of the helix with adenine forming two hydrogen bonds with thymidine and cytidine forming three hydrogen bonds with guanine. It immediately occurred to Watson and Crick that if you zipped open the two strands and began copying them so that everywhere that there was an adenine in the original strand, the new strand had a thymidine, (a guanine would be matched to a cytidine etc.), then soon you would exactly duplicate both strands using only the concept of base pairing and whatever enzymes and substrates that might be necessary to carry out the synthesis.

At the completion of replication, each original strand (parental strand) is paired with a new (daughter) strand. The replication is said to be semiconservative. Double stranded DNA may be denatured by alkaline conditions or by heat. For example, the two strands will unwind when a solution of DNA is heated to $90^\circ$ and then will reassociate if the temperature is lowered gradually. Longer DNA molecules take longer to reform their double-stranded structure in perfect alignment.

Even small stretches of DNA (polynucleotides) will anneal with large single-stranded DNA molecules if the base sequence matches a sequence somewhere on the single stranded DNA (by base pairing). A sequence of 10 or more matching base pairs usually is needed to form enough hydrogen bonds for the complex to be stable. Many of techniques used to manipulate DNA make use of this observation.
2.2 Proteins and amino acids

Proteins are made up of long sequences of amino acids. There are 20 common amino acids used for synthesis of the many thousands of proteins found in living organisms. The properties of the proteins depend upon the exact sequence of their amino acids which in turn is determined by the genetic code discussed below. A major goal of genetic engineering is to be able to make useful quantities of peptides and proteins, some of which are as small as a dozen amino acids on up to proteins containing hundreds of amino acids. The proteins we want to make (perhaps using bacteria to produce them) might be human hormones, antibodies and other rare proteins that are difficult or impossible to obtain by other means.
Many proteins found in cells are enzymes that catalyze metabolic reactions. These proteins must have a three dimensional structure that encourages binding of substrates and one or more active sites where the reactions take place. Peptide hormones must be of precise structure in order to bind to their receptors on cells.

**FIGURE 4. Structure of the common amino acids**
The exact amino acid sequence and the precise folding of the protein structure is essential for fulfilling their biological roles. Proteins fold into characteristic structures determined by (guess what) hydrogen bonding, usually into helical form (a-helix) or sheet-like structures (b-sheets). A single protein often has several of each of these folding patterns. Also, during protein folding, the amino acids with aliphatic and aromatic side chains tend to locate in the interior of the protein away from water, with charged amino acids at the surface.

_The naming of amino acids is trivial (asparagine was isolated from asparagus)_

The usual structure is a carboxyl and an amino group covalently bound to the same carbon atom (the a-carbon):

\[
\text{NH}_2--\text{CH}--\text{COO}--
\]

\[\text{R}\]

Various side chains are attached to the alpha-carbon (the R-group; see table). There are several amino acids that have only carbon and hydrogen on the R group (alanine, valine, leucine, isoleucine), several that have a positive charge at neutral pH (histidine, lysine, arginine), several that are acidic at neutral pH (aspartic acid, glutamic acid), and several that have aromatic rings (phenyl alanine, tryrosine, tryptophan). The amino acid cysteine (HS-CH$_2$ = R) is of particular interest because disulfide bonds between two cysteine residues are often found that cross-link between two peptide chains or between two parts of a single peptide (to give R-Ch$_2$-S-S-CH$_2$-R’). The amino acids often are represented by three letter abbreviations (ala, val, leu, iso, etc.) and even one letter abbreviations (A=alanine, R=arginine, D=aspartic acid, etc) to save space when writing out sequences.
Proteins are made up of combinations of helical and sheet structures.

**α-Helix**
only the N-Cα-backbone is represented. The vertical line is the helix axis.

**β-Strand**
Note that the amide planes are perpendicular to the page.

"Shorthand" α-helix

"Shorthand" β-strand

FIGURE 5. Proteins are made up of combinations of helical and sheet structures.
During synthesis of proteins, amino acids are added one at a time according to the genetic code using complex structures called ribosomes. Each amino acid is added to the preceding one by a splitting out of water between the carboxyl group of one amino acid and the amino group of the next amino acid. The resulting bond between the two is known as the peptide bond:

\[
\text{--CH--CO--NH--CH--COO--}
\]

It follows that the first amino acid of a protein usually will have a free amino group (the amino end) and the last amino acid will usually have a carboxyl group (the carboxy end). I have said “usually” this is true, because sometimes the ends of proteins are modified after synthesis.

2.3 Replication of Nucleic Acids

Polymerase enzymes catalyze the duplication of nucleic acids. All of the DNA polymerases can only catalyze the addition of additional nucleotides to an existing piece of DNA that has a free 3’-hydroxyl group (they are unable to start a new strand without a primer attached). This limitation imposes some special conditions on the way new DNA is made in cells. The reaction itself uses deoxynucleotide triphosphates as substrates. The triphosphate supplies the free energy to drive the reaction (figure). Note that the enzyme only will add the next nucleotide base onto a primer, and the base to be added must be able to hydrogen bond with the base exposed on the strand to be copied (the template strand). During the reaction, inorganic pyrophosphate is liberated and immediately hydrolyzed to inorganic phosphate, insuring that the reaction will be irreversible.

The reaction only proceeds in the 5’ to 3’ direction (new bases are added to the 3’ hydroxyl of the preceding sugar.)
What are some of the requirements for DNA replication starting with intact DNA? First of all, replication starts at special sequences of bases that define the origin of replication (indicated as ori in our models). Then we must begin to unwind the DNA to provide access for DNA polymerase and other needed enzymes and proteins. As the DNA unwinds, we see two fork shaped areas that are indeed the replication forks. They proceed in both directions down the double stranded DNA, and each must have its own set of enzymes. Remember that the two strands of DNA are running in opposite directions. One strand is ready to serve as a template for the 5’ to 3’ synthesis, but the other is in the wrong orientation. To solve this problem, RNA polymerase is used to synthesize a small stretch of RNA (Okazaki fragment, named for the scientist who discovered this mechanism) to provide a primer so that DNA polymerase may continue the chain. The RNA stretches are removed, filled in with DNA and sealed together with an enzyme called ligase. A more detailed representation of the replication fork complex is shown in Figure 7. Notice that proteins are needed to help unwind the DNA, which is tightly coiled, before the synthesis may continue. By looping one strand around (the one that needs the RNA start) we may visualize the replication complex as all occurring as one large aggregate at the replication fork.
2.4 Making an RNA copy of DNA

You may already know that it is RNA that carries the genetic code to ribosomes with instructions for protein synthesis. Our master DNA tape stays safely away from the action and sends messenger RNA out with the code.

One strand of DNA is copied into RNA using RNA polymerase. In order to copy the genetic code from DNA molecules, we need to copy one of the strands of the DNA into RNA using RNA polymerase and the same base pairing principles that we used for DNA synthesis. Again we have the problem of gaining access to a tightly wound coil of DNA. The DNA must be unwound slightly to allow the enzyme to bind and begin copying the sequence.

Control regions are found upstream of the coding region of the gene. The binding only occurs at a region of the gene called the **promoter** region. This base sequence occurs before the 5’ end of the gene (upstream) along with many other sequences that serve to control reading of the gene. Synthesis proceeds with the required base paring (remember adenine now pairs with uracil) using the ribonucleotide triphosphates (containing the sugar ribose) as substrates. The next figure shows a more complete model of the polymerization with an indication of how the unwinding of DNA coils may occur.
There are some differences between bacterial messenger RNA (mRNA) and that of higher organisms that have great importance if we are cloning the gene. Bacterial mRNA is used directly without modification. In fact, ribosomes often attach to mRNA of bacteria as it is being synthesized and immediately use the code for new protein synthesis.

The RNA of higher organisms must be processed before it becomes messenger RNA. In higher organisms, there are stretches of RNA that are cleaved out and discarded (called introns). The ends of the remaining RNA that contain the actual code (called exons) are reattached (spliced) to each other to give a shorter mRNA that is now ready to serve as a coding strand for protein synthesis. The mRNA of higher organisms also usually has polyadenylate tail added to the 3’ end of each mRNA.

Scientists puzzled for some time on how four nitrogenous bases could provide the code for the 20 amino acids used in protein synthesis. The puzzle was solved once it was realized that three bases were used to code for one amino acid. Now the number of possible combinations could be calculated as \(4^3 = 64\), or more than enough!
In general, each amino acid can be coded for by several different three-base codons (the code is redundant; see table). For example, leucine is coded for by CUU, CUC, CUA, CUG, UUA, and UUG. Notice that in the first four examples, only the last base is different. Three codons, TAA, TAG, and TGA are used as stop codons to tell ribosomes to stop reading the code and stop making protein. Methionine has only one codon (AUG) and it is used as the first amino acid to start a new protein (it sometimes is removed later, as all proteins do not have an N-terminal methionine).

**FIGURE 10. The genetic code**

We now may think back concerning the necessary sequence of events. One strand of the DNA must be copied into an RNA sequence beginning at the promoter region. Either DNA strand may be used, depending on the location of the promoter region. Each three-base code of the RNA
specifies a specific amino acid that will be polymerized together on ribosomes to give the final amino acid sequence which, in turn, will determine the properties of the new protein.

We will not go into the details of protein synthesis, but the general features are shown in the figure. Ribosomes assemble on the mRNA in the region containing a ribosome binding site and the initiation codon (AUG for methionine). Amino acids are brought to the proper site on the ribosome in the form of complexes with special small RNA molecules known as transfer RNAs. Each transfer RNA is specific for its own amino acid and has an “anti-codon” on one end of the loop shaped molecule and the amino acid at the other. In the case of methionine, the anti-codon would be UAC which will associate with AUG of the bound RNA. After a second transfer RNA with its amino acid enters the site, a peptide bond is synthesized between the two amino acids. The ribosome then shuttles three bases along the mRNA so that the next transfer RNA and its amino acid may enter and undergo polymerization. The process continues until a stop codon is reached and the ribosome falls off the mRNA.
Some proteins have short stretches of amino acids that serve to direct the protein into specific cell compartments as it is being synthesized. These “signal” sequences are then removed by specific enzymes. In this way, proteins may be captured into vesicles for excretion out of the cell (insulin) or, in bacteria, may be directed to the space between the inner and outer membranes or excreted.
2.6 Making a Genetic Probe

Often we may know the amino acid sequence (or part of it) of a protein of interest, and frequently we will wish to synthesize a short nucleic acid sequence that might correspond to the gene sequence used by the cell. Such polynucleotides may be used as molecular probes to help locate the gene and as primers for polymerase reactions.

The redundancy of the genetic code complicates our problem because we have no real way of guessing which codon the cell may be using for a particular amino acid (Figure 12), although a given organism often uses some redundant codons more efficiently than others. In this case, we would choose the sequence starting with glu which uses two codons and proceed through a region where there is a minimal number of possible codons. We still need to consider ten possible sequences in this stretch of 15 nucleotides.

<table>
<thead>
<tr>
<th>AMINO ACID SEQUENCE</th>
<th>leu-glu-asn-tyr-cys-asn</th>
</tr>
</thead>
<tbody>
<tr>
<td>POSSIBLE CODONS:</td>
<td>UUA GAA AAU UAU UGU</td>
</tr>
<tr>
<td>AAC</td>
<td>UUG GAG AAC UAC UGC</td>
</tr>
<tr>
<td>AAC</td>
<td>CUU</td>
</tr>
<tr>
<td>AAC</td>
<td>CUC</td>
</tr>
<tr>
<td>AAC</td>
<td>CUG</td>
</tr>
<tr>
<td>SYNTHEtIC PROBES:</td>
<td>GA(A)-AA(U)-UA(U)—</td>
</tr>
<tr>
<td>UG(U)-AA(U)</td>
<td>(G)</td>
</tr>
<tr>
<td>(C)</td>
<td>(C)</td>
</tr>
</tbody>
</table>

**FIGURE 12.** A nucleic acid probe should be selected around the least redundant sequence of possible codons. This sequence of amino acids is taken from the insulin molecule. (see FIG 14)

2.7 The Control Regions of Genes

It is essential that the cell be able to regulate the relative amounts of proteins and enzymes that are synthesized according to the needs of cell in its environment. The process is very well studied in bacteria and an example of one such regulatory region and how it works was worked out by Jacob and Monod and is shown in the figure. *E. coli* makes the proteins and enzymes that are needed to metabolize lactose only if lactose is present in the environment, and actually prefers to use glucose if it is available. Whether or not these genes are transcribed is under control of a repressor protein and several other factors, including cyclic AMP. The gene for the
repressor protein lies upstream from the lactose control region and, in the absence of lactose, the repressor (as a tetramer) binds to a gene segment called the operator and prevents RNA polymerase from binding at the adjacent promoter site. If lactose is present, it binds to repressor and lowers the binding affinity of the repressor for the operator site. An enhanced reading of the lactose genes occurs upon binding of CAP protein (catabolite activator protein) and cyclic AMP leading to an active transcription of the genes and enabling the cell to metabolize lactose. The level of cyclic AMP in the cell is regulated by glucose. Thus, when glucose concentration is high, cyclic AMP concentration declines, and the lactose genes will not be read. We will use parts of this same system later in cloning experiments. The lacZ gene is the structural gene for β-galactosidase, the enzyme that hydrolyzes lactose to glucose plus galactose. It can accumulate to equal as much as 10% of total soluble, cellular protein following induction.

FIGURE 13. Control of transcription of genes involved in lactose metabolism

2.8 Insulin and its Genes

We now may examine the amino acid sequence of insulin and how its gene is organized. Insulin is a small protein made up of two peptide chains held together by disulfide bonds
between cysteine residues. The A chain has 21 amino acids and the B chain 30. The amino acid sequence was determined by Sanger in 1955.

Insulin is produced only in the beta cells of the pancreas. Of course, the gene for insulin is in all cells, but insulin is not produced in other tissues because of special control regions that lie in front of the gene and prevent it from being expressed. Proinsulin is the precursor of insulin and contains an extra peptide chain (the C chain). The C chain that links chains A and B allows the entire molecule to be synthesized in the b-cell as one protein. The C chain is split out after proinsulin is made to give the mature insulin structure of 51 amino acids with a molecular weight of 5,700.

Insulin is actually produced in the b-cell as **preproinsulin** with an extra 24 amino acids which serves to guide the new protein to vesicles where it is stored. This "signal sequence" is removed shortly after protein synthesis begins.
Insulin is a member of a large family of related proteins and over 80 of these have been sequenced. Even insects have insulin-like proteins. The disulfide bonds are conserved, and the molecules form three helices in their native form. The three-dimensional shape of the insulin molecule is shown in the figure. Other animal insulins have almost identical amino acid sequences to that of human insulin, which means that we may sometimes use them in place of human insulin.

Pork insulin has only one amino acid change; in chain B at position 30, alanine is found in place of threonine
Beef insulin has two additional changes; in chain A at position 8, alanine is found in place of threonine and at position 10, valine is found in place of isoleucine.
Chapter 3 Cloning Genes

3.1 Overview

Now we are ready to examine the technical procedures (and some of the theories behind them) that allow molecular biologists to detect, study and clone individual genes—and even produce the proteins coded for by these genes in commercial amounts. These developments have sparked a revolution in the biological sciences and will continue to have far-reaching implications for every area of biology and medicine in the years ahead.

The practice of genetic engineering, or gene cloning, had to await the development of suitable methods in the early 1970s for cleaving, joining and manipulating nucleic acids. The basic ideas behind the procedures are easy to state: genetic information is encoded in genes which are made up of long polymers of nucleotides in the form of double stranded DNA molecules.

Cleave genomic DNA into large fragments using restriction endonucleases.
Restriction endonucleases are used to cleave genomic DNA into large fragments at defined nucleotide sequences.

Insert the DNA into plasmids. These pieces of DNA are inserted into plasmids, which are self-replicating, extrachromosomal genetic elements originally isolated from the bacterium, Escherichia coli. The circular plasmid DNA is opened using the same endonuclease that was used to cleave the genomic DNA.

Join the ends of DNA with the enzyme, DNA ligase. The inserted DNA is joined to the plasmid DNA using another enzyme, DNA ligase, to give a recombinant DNA molecule. The new plasmid vector contains the original genetic information for replication of the plasmid in E. coli plus the inserted DNA which may contain, for example, the human gene for insulin.

Introduce the new vector into E. coli. The new vector is inserted back into E. coli where many copies of the genetic sequence are made as the bacteria grow and divide with the replicating vector inside.

Isolate the newly-synthesized DNA, or the protein coded for by the inserted gene. The E. coli may even transcribe and translate the gene and obligingly produce insulin, if
that was the gene incorporated into the vector. Alternatively, many copies of the DNA
gene itself may be isolated for sequencing the nucleic acid or for other biochemical
studies.

Many variations on the basic methodology have been developed. We will consider only
three: the use of vectors made from bacterial viruses (or bacteriophages); cosmid vectors;
and the polymerase chain reaction (PCR). Additional methodologies may be found in the
reference list.
FIGURE 16. The general scheme for genetic engineering
3.2 Cutting DNA

The most important enzymes for gene cloning are the restriction endonucleases. These nucleases make internal cuts in the nucleic acid molecule only where certain sequences of nucleotides are present. Thus, they tend to cleave DNA into rather large pieces rather than small fragments.

These enzymes are found in various bacterial species. They are believed to protect the organisms from any invading foreign DNA, such as might be introduced by bacterial viruses. The DNA of the host is protected by methylation of the DNA bases that are a part of the recognition sequence.

Example Escherichia coli K has a restriction endonuclease EcoK which is coded for by the gene hsdR. However, EcoK does not cleave the host DNA because E. coli K has a methylase coded for by the gene hsdM that methylates the bases in the region where the endonuclease would cleave unprotected DNA. Both of the enzymes require the presence of gene hsdS which is needed for DNA recognition.

Over 800 restriction endonucleases are known. The most useful for cloning work are known as Type II which generally make a staggered cut within the recognition site and across the double stranded DNA.

The enzymes are named by a three letter code indicating the organism from which they were isolated. For example, EcoRI was isolated from E. coli and BamHI from Bacillus amyloliquefaciens.

The action of EcoRI is as follows:

```
5' - - G AATT C - - -
3' - - C TTAA G - - -
```

```
5' - - G-3'OH
3' - - CTTAA-
```

```
5' - - G AATT C - - -
3' - - C TTAA G - - -
```

```
5' - - G-3'OH
3' - - CTTAA-
```

```
AATTC - -
3'OH-G - -
```
Notice that the enzyme recognizes a six-base pair sequence in the DNA, and the cleavage results in single-stranded ends (sticky ends) that would readily reform hydrogen bonded base pairs again with each other.

Most restriction endonucleases recognize four, five or six base pair sequences. It can be calculated that if the base sequence of DNA were random, then the expected frequency for a given sequence would be 4 (for the number of different bases to the nth power:

\[
\text{Expected frequency of cleavage} = 4^n \quad \text{where } n \text{ is the length of the recognized sequence}
\]

Thus, the *EcoRI* site might be expected to occur about every 4096 base pairs. Enzymes that recognize a four base pair sequence (such as *HaeIII*) would be expected to produce shorter fragments (around 256 base pairs in length). Other restriction endonucleases are shown in the table. After cleaving with restriction endonucleases, plasmid DNA and the DNA to be inserted are allowed to anneal *via* their sticky ends.
FIGURE 17. Some commonly used type II restriction endonucleases

*SceI*  
(5') ATCGAT (3')  
TAGCTA  
*↓*  

*EcoRI*  
(5') GAATTC (3')  
CTTAAG  
*↓*  

*HaeIII*  
(5') GGGC (3')  
CCGC  
*↓*  

*HindIII*  
(5') AAGCTT (3')  
TTCGAA  
*↓*  

*NcoI*  
(5') GGGGCCC (3')  
GGCGGGCG  
*↓*  

*PstI*  
(5') CTGCAG (3')  
GACGTC  
*↓*  

*PvuII*  
(5') CAGCTG (3')  
GTGACGAC  
*↓*  

*SmaI*  
(5') CCCG (3')  
GGGGG  
*↓*  

*Tth111I*  
(5') GACNNNNGTC (3')  
CTGNNNCAG  
*↓*
FIGURE 18. The three-dimentional structure of a restriction endonuclease showing the binding of DNA

3.3 Joining DNA: DNA Ligase

The second key enzyme for cloning is DNA ligase. This enzyme will rejoin DNA strands that are close together and that have a free -3’-OH facing an adjacent 5’-phosphate. This will come about naturally if the sticky ends of DNA molecules that have been cleaved by the same nuclease are allowed to anneal with one another. The most commonly used DNA ligase is one coded for by bacteriophage T4. It may be purified from *E. coli* cells that are infected by T4.
DNA ligase is activated by ATP which adds an adenylate group to the enzyme. The adenylate group is transferred to the 5’ terminal phosphate of the DNA:

\[
\begin{align*}
\text{3’OH} & \quad \text{Adenosine Monophosphate} \\
\text{---------G pApApTpTpC-------} \\
\text{---------CpTpTpApAp G------} \\
\text{OH} \quad \text{AMP} \\
\text{3’} & \quad \text{5’} \\
\end{align*}
\]

The 3’ OH now can perform a nucleophilic attack on the phosphate linkages and form the phosphodiester bond to seal the stands:

\[
\begin{align*}
5’ & \quad \text{GpApApTpC-------} \\
3’ & \quad \text{CpTpTpApApG------} \\
\end{align*}
\]

\[+ 2 \text{ AMP}\]

The enzymatic reaction is run at 5-15° (which is below its temperature optimum of 37°) so that the hydrogen bonds which hold the sticky ends together will not dissociate. Sometimes the ligase will simply reseal the original plasmid back together, and often it may seal together the two ends of the DNA gene that we wish to insert instead of inserting it into the plasmid. We will see how to deal with these problems when we screen for transformants. Since there is a binding site for both strands of the DNA on the ligase enzyme, it is possible to bind and ligate blunt-ended DNA molecules, but with much lower efficiency.

### 3.4 Host Strains

The most commonly used host strains have been derived from \textit{E. coli} K-12, although other bacteria, yeast cells, plant cells and mammalian cells are used according to the needs of the experiment. For example, if you are producing a human protein that has specific carbohydrate residues needed for activity, it may be essential to produce that
protein in mammalian cells. Also, eukaryotic genes usually have introns and exons that cannot be directly transcribed into useful mRNA by prokaryotes, such as *E. coli* (see molecular biology and cDNA libraries sections). Sometimes a bacterial host may be used for initial experiments and gene isolation, and then the vector is moved to a mammalian host for expression.

Our examples use *E. coli* K-12 strains that have the following properties:
1. These bacteria are easy to handle in the laboratory, and may be used with many different vectors. The fundamentals of genetic engineering were developed using these organisms and are based upon an exhaustive knowledge of their molecular genetics.
2. The strains are selected to contain mutations in host restriction and usually other mutations involving nutritional requirements so that they will not grow outside the laboratory.
3. They are easily grown on minimal medium containing salts and an energy source such as glucose. They also grow on a rich nutrient broth with a doubling time of 20 min.

*E. coli* is a Gram-negative short rod. The genome consists of a large, compactly folded circle containing $4 \times 10^6$ base pairs. *E. coli* proteins are made on ribosomes in the cytoplasm and either remain there or are extruded into the periplasmic space, or even excreted into the medium by special mechanisms. Recombinant proteins often accumulate as insoluble inclusions in the cytoplasm.
Figure 19. E. coli is a commonly used host cell.

3.5 Cloning Vectors for *E. coli* are of three main types
1. Plasmids--circular DNA vectors that are self-replicating (separate from the host DNA). Plasmids are usually 5-40 kilobase pairs (Kbp) in size, and DNA fragments up to 15 Kbp can be inserted into the plasmids.

2. Bacteriophages--include modified bacteriophage lambda, which has 48.5 Kbp. About one-third of the genome is not essential for DNA replication and may be replaced with up to 23 Kbp of inserted DNA. The total length of the DNA must be of the correct size to be packaged into virus particles within the infected cells.

3. Cosmids--a combination of plasmid and bacteriophage. Small (5-7 Kbp) circular DNA containing an origin for DNA replication (ori), markers and restriction sites plus a sequence from lambda needed for packaging the DNA (cos site). Cosmids may be used to clone large DNA molecules of up to 45 Kbp.

3.6 Plasmid Cloning Vectors for *E. coli*

There are hundreds of different plasmids that today may be used for cloning experiments with *E. coli*. We will concentrate on describing the plasmid known as pBR322 which was developed by Bolivar and Boyer.

The plasmid known as pBR322 was developed by Bolivar and Boyer in 1977. This plasmid has 4,362 base pairs that includes a special sequence of DNA that is needed as an origin (the ori site) for DNA replication and that allows for efficient replication. Usually 10--20 copies of the circular DNA are produced in each bacterial cell that harbors the plasmid.
FIGURE 20. The pBR322 plasmid

Two genes of pBR322 confer resistance to antibiotics to any cell that contains the plasmid. *AmpR* confers resistance to ampicillin and *tetR* confers resistance to tetracycline to cells containing the plasmids. *AmpR* is a gene that codes for the periplasmic enzyme beta-lactamase that cleaves the ring structure found in ampicillin, which is a penicillin antibiotic. *TetR* is a gene that codes for a protein that modifies the bacterial cell wall and prevents tetracycline from entering the cell.

Multiple restriction endonuclease sites are present where foreign DNA fragments may be inserted.
“Relaxed” plasmid DNA replication continues in the presence of chloramphenicol. An interesting feature of this plasmid is that “relaxed” plasmid DNA replication continues even in the presence of an inhibitor of protein synthesis such as chloramphenicol. This feature allows increased yields of plasmid/cell of up to 100-fold.

3.7 Transforming Cells with Plasmids

As you might imagine, it is not easy to force large DNA molecules into bacterial cells. *E. coli* is protected by two membranes, the outer membrane and an inner plasma membrane, and large, charged molecules do not ordinarily pass through phospholipid membranes. The organism also has protective enzymes; the restriction endonucleases we already have learned about and other nucleases that are ready to attack foreign DNA as it entered the cell. In 1974, Lederberg and Cohen developed a method to partially overcome these natural barriers.

Intact DNA does not ordinarily enter bacterial cells. First, treat *E coli* cells at 0° with a dilute solution of CaCl2 to weaken the cell wall. Next, added DNA forms hydroxy-calcium-phosphate bonds and the complex adsorbs to the cells. The calcium-phosphate complex not only adsorbs to cells, but the DNA is protected inside the complex.

Next, Raise the temperature to 37-40° to “heat shock” the cells and force DNA inside. During the heat shock, a few of the cells take up complete molecules of DNA. Usually about 1 in 1000 DNA molecules yield a transformant. This mean that 10^5 - 10^7 transformants may be obtained for each microgram of pBR322 DNA. Other methods may be used. For example, DNA may be driven into cells by an electric pulse which creates a temporary pore in the membranes, a technique known as electroporation.

Special host strains of *E. coli* are used that lack the nucleases that would otherwise attack and degrade the inserted plasmid DNA (see host strain section)

3.8 Selecting Transformed Cells

Remember that only a small percentage of the bacterial cells may have taken up the plasmid DNA and not all of the plasmids will have the gene we wished to insert. Some transformed cells may contain original plasmid that closed upon itself without
incorporating any of the new DNA--and not all of the DNA that might be inserted in the plasmid may be intact and functional. We will see later that when we are making a gene library, the one gene that interests us may be only one of thousands present in the DNA sample that we have cloned into the plasmids. We clearly need very sensitive and accurate methods for screening for just the transformed cells that we want.

In even the simplest cases, the transformation mixture will contain large numbers of untransformed cells, usually some cells that are transformed with only the original plasmid, and a few cells that are transformed with the complete vector containing the gene of interest. There are several procedures that may be used to create very efficient selection techniques:

**Include in the plasmid a gene that is required for growth of the transformed host cell--usually a gene for resistance to an antibiotic.** If the plasmid brings in a gene for antibiotic resistance, then normal host cells will die if the antibiotic is added to the culture and only transformed cells will survive.

**Include in the plasmid a gene that will yield colored colonies when a suitable substrate is added.** Only those colonies that are transformed by plasmids that have a new gene for an enzyme that will convert a substrate to a colored produce will be colored. Normal cell colonies will appear white and colorless.

**A radioactive nucleic acid probe (such as $^{32}$P-mRNA) may be used to detect those colonies that have the new DNA gene.** For this procedure, we must isolate the mRNA of the gene or chemically synthesize in radioactive form a small section of the DNA sequence which we know to be a part of the gene. Of course, we do not always have this information.

**Antibodies (often labeled with radioactivity) that react with the protein product of the gene may be used to determine which colonies contain the new gene.** For this procedure to work, the transformed cell must make mRNA copies of the new gene and then translate these messages into new protein products. For example, *E. coli* containing the complete gene for proinsulin will make large quantities of human insulin. These cells could be detected using anti-insulin antibody. The general techniques for these procedures will be described under “replica plating” and other refinements are described in the Nicholl text.
3.9 Replica Plating for Colony Screening

We will now examine how to select for transformed cells that harbor the plasmid of interest if that plasmid contains genes for antibiotic resistance. Think back to the plasmid pBR322. The original plasmid contains genes for resistance to ampicillin and tetracycline. Cells transformed with the original plasmid will be resistant to both of these antibiotics. If we cut the plasmid with *Bam*H1, and insert our new gene at this site we will interrupt the gene for tetracycline resistance. If we transform cells with this new construct, the cells will not be resistant to tetracycline, but they will still be resistant to ampicillin.

**FIGURE 21.** Replica platting

Spread about $10^7$ bacteria from the transformation reaction mixture on the surface of a plate containing nutrient agar. Incubate the plates overnight at 37°C to allow the cells to grow as colonies.

Press a piece of velvet against the surface to pick up some of the cells.
Transfer to new plates containing normal media or medium containing antibiotics and grow overnight.

Expected Result: The colonies of cells pressed onto normal medium will grow vigorously as before and they represent the total number of cells that we plated. In our example, cells plated on medium containing ampicillin will not grow unless they were transformed with either the original plasmid or the plasmid containing the inserted gene. The number of colonies tells us about the efficiency of transformation.

Select colonies of cells containing the desired transformation. Cells growing on medium containing both ampicillin and tetracycline must be those cells that were only transformed with the original plasmid and not the plasmid containing the gene we want. We now must go back to the ampicillin plate and select and save those few colonies that were unable to grow in the presence of tetracycline. These are the cells that we will save and grow and which contain the desired plasmid.

3.10 Nucleic Acid Probes

A similar replica plating procedure may be used if we have a radioactive nucleic acid probe that will aneal with our new gene. The procedure is as follows:
Grow the cells on the master plate as before, but substitute a nitrocellulose filter for the velvet cloth to blot the colonies.

Press a nitrocellulose filter onto the colonies; some cells will adhere. When the attached cells are treated with alkali, the cells will break open and the DNA will attach to the nitrocellulose filter. A radioactive probe derived from the new gene may then be incubated with the filters.

Remove the filter with attached cells and treat with alkali.
Incubate with a radioactive nucleic acid probe (RNA or DNA) under annealing conditions.
Detect where the probe bound by autoradiography using X-Ray film.

The probe binds to those colonies that contain the new gene. We then may go back to the master plate and select the desired colonies for further growth

3.11 Purifying Plasmid DNA

Any of the commonly used techniques for purifying nucleic acids from bacterial cells may be used to isolate and purify plasmid DNA from *E. coli*.

Grow bacteria containing the plasmid in several hundred ml of culture medium. It is customary to start from a single bacterial colony and grow the bacteria in a rich liquid medium containing the same antibiotic that was used in the original selection procedure. This prevents the growth of any non-transformed cells that would grow faster than the cells containing the plasmid.
Harvest the cells by centrifugation and treat with lysozyme followed by detergent. The enzyme lysozyme is used to weaken the outer cell membrane, and the detergent lyses the cells by interaction with the plasma membrane.
Isolate the DNA by gel electrophoresis or density-gradient centrifugation
Minipreps may be prepared by lysis at pH12

![FIGURE 23. Separation of DNA fragments by gel electrophoresis](image-url)
If small amounts of DNA are enough for the experimental purpose, cells may be lysed directly using alkaline conditions. The host DNA will be denatured, but the supercoiled plasmid DNA will not. The two forms of DNA may be separated easily by gel electrophoresis or density-gradient centrifugation and recovered for use.

3.12 Colored indicators

Color reactions may be used in several ways in cloning. Usually an enzyme such as beta-galactosidase is detected by using a substrate that gives a colored reaction product after being attacked by the enzyme. *E. coli* will synthesize b-galactosidase after it runs out of glucose if lactose is present in the medium. An analog of lactose known as IPTG (isopropyl-thiogalactoside) will induce gene transcription leading to beta-galactosidase synthesis, but the enzyme is unable to hydrolyze the compound.

Another compound known as X-gal (5-bromo-4-chloro-3-indolyl-b-D-galactopyranoside) is a substrate for the enzyme and a blue colored product results after galactose is hydrolyzed off and air oxidation of the indoxyl product takes place. X-gal is not an inducer of the enzyme. The situation actually is a little more complicated that this because two gene products are needed to give an active enzyme and the system is under control of a repressor (see Molecular Biology section). If we wish to use the beta-galactosidase system in a plasmid, then the host cells must contain mutations in one or more of the so-called lacZ sequences of the host genes.
FIGURE 24. Detection with colored indicators

The beta-galactosidase system is especially useful because it is inducable. That is the promoter is only activated (and the gene transcribed) in the presence of the inducer (which as we have noted could be lactose or IPGT). Thus, if we want to produce large amounts of a new protein in *E. coli*, a straight-forward procedure is to incorporate the very effective inducable promoter and *lacZ* gene into the plasmid directly ahead of our gene of interest. A “fusion” protein consisting of the first part of β-galactosidase and our new gene product will be produced after we have grown a dense culture of transformed *E. coli* and only after we add IPGT. Examples of this approach are given in the next section.

The reaction is equally useful in detecting successful transformations. If the plasmid has an expressible promoter and *lacZ* gene region (such as the pUC series of plasmids), any transformed cells will give blue plaques with incubated with IPGT (inducer) and X-gal (chromogenic substrate). If we insert our new gene into the *lacZ* sequence (interrupting it) then the plaques will be colorless. The pUC plasmid series also have an ampicillin resistance gene for additional selection and a short sequence containing multiple restriction sites to make them highly useful vectors.

3.13 Let’s Clone!
We now may use the various procedures that we have learned to examine making a plasmid vector, transforming cells and producing human polypeptides in *E. coli*.

**Somatostatin (Itakura, et al., Milestones readings, p. 84)**

Somatostatin is a very small peptide hormone and provides a nice example of the major principles to be considered when cloning a gene (see figure 25). The polypeptide amino acid sequence is known, is only 14 amino acids long, and contains one internal disulfide bond. The hormone is produced in the pancreas and other tissues and inhibits the secretion of insulin and glucagon. It also reverses glucagon effects in the liver.

![GENETIC CODE](image)

**FIGURE 25.** Cloning somatostatin

The DNA gene was synthesized chemically after deciding likely codons for each amino acid. Notice that deoxynucleotides A through H were prepared so that they would anneal with one another in an overlapping sequence that could be sealed by ligase. Also notice
the *EcoRI* site synthesized at one end of the molecule and a *BamHI* at the other. When we open pBR322 using these two restriction endonucleases, a segment of the plasmid is cleaved out and discarded and our new gene will enter the opened plasmid in only one orientation. Notice also that the *lac* promoter and control region and part of the β-galactosidase gene have been inserted into the plasmid. This means that the protein will not be produced until an inducer has been added to the medium and that the protein product will consist of part of the β-galactosidase protein linked to the polypeptide that we want.

**FIGURE 26.** Strategy for the chemical synthesis of the gene for somatostatin

Another trick that was used was to insert the codon for methionine as the first amino acid in front of the code for somatostatin. Somatostatin itself does not contain methionine. Cyanogen bromide may be used to chemically break the polypeptide bond at all methionine residues, thus liberating the polypeptide from the β-galactosidase-somatostatin protein product. Two stop signals were inserted into the gene at the end of the somatostatin gene sequence to ensure that protein synthesis would not continue beyond this point. Recombinant *E. coli* colonies were selected as usual using antibiotic resistance markers, and clones yielding the highest amounts of somatostatin were further screened with a radioimmune assay. The protein product that contained somatostatin was found to reach over 3% of the total cellular protein in some of the clones. The peptide hormone was successfully purified from the transformed cell.
Almost the same procedure as just described was used for the first cloning of insulin. In this case, synthetic DNA sequences were prepared for both the A (77 base pairs) and B (104 base pairs) polypeptide chains, and these were inserted separately into plasmids containing lac sequences. The individual vectors could be used to transform separate cultures of E. coli. After replica plating and selection, clones were found that produced high yields of either the A- or the B-chain. The insulin polypeptide chains also do not contain methionine, so that cyanogen bromide could be used to cleave out the mature polypeptides from proteins produced in transformed cells. The A- and B-chains were purified away from E. coli cellular material and beta-galactosidase fragments, and conditions were worked out so that when the two purified chains were incubated together under reducing conditions, mature insulin was formed with high yield and with the disulfide bonds in the correct position.
FIGURE 27. Cloning insulin
Chapter 4 More Advanced Cloning Techniques

4.1 Bacteriophage Lambda Vectors

Interestingly enough, even bacteria are plagued by viruses! You may know that when viruses infect cells they manage to transfer their nucleic acid genetic material to the inside of the host cell. There it replicates to produce up to several hundred new viruses. If we could replace one or more of the viral genes with a gene we are interested in amplifying, then we would have another type of vector in addition to the plasmids already discussed. Furthermore, the host cell may obligingly package up the replicated DNA containing our gene into new virus particles that are liberated from the cell.

The lambda (λ) genome is 48.5 kbases in length and contains about 46 genes. The bacteriophage lambda is frequently used as a vector. It has a head structure that contains the viral DNA and a long tail which is used in attachment of the virus to \textit{E. coli}. The original virus has a set of genes that allows the viral DNA to insert itself into the bacterial chromosome. These genes may be cut out and discarded to provide a location for our new gene to be inserted.

![Diagram of the λ genome](image)

**FIGURE 28.** The λ genome

Following attachment of the virus to \textit{E. coli}, the DNA in the head structure travels down the hollow tail and enters the host cell. The linear, viral DNA has naturally “sticky ends” consisting of a complementary sequence of 12 bases on each end of the molecule. The
sticky ends of the linear DNA (the \textit{cos} site) associate to form circular DNA. The virus codes for its own DNA polymerase which leads to the synthesis of several hundred copies of viral DNA.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{phage_life_cycle.png}
\caption{The phage life cycle}
\end{figure}

The DNA is synthesized in very long repeated units which are cleaved at the \textit{cos} sites to give unit length viral DNA for insertion into new particles. The bacteriophage DNA also codes for all of the proteins needed to put together new virus particles which then are assembled inside the host cell.

The new DNA is trapped inside the newly-formed viral coat proteins and the host cell bursts, liberating the new particles. A most remarkable property of the proteins and the viral DNA is that complete particles will assemble in the test tube when all of the reactants are mixed together.
The infection spreads to adjacent cells and the cycle is repeated. If the infection is carried out with *E. coli* cells that are spread over an agar surface, the infected areas will appear as clear dots (or plaques) surrounded by live cells. Bacteriophage may be recovered from the clear areas. Each plaque originates from a single infected cell.

![Phage plaques on a lawn of bacteria](image)

**FIGURE 30.** Phage plaques on a lawn of bacteria

Virus particles are separated from cell debris and the DNA isolated. The process may be scaled up by infecting cells growing in liquid culture. Over 100 different lambda vectors have been prepared for use in cloning. The DNA of the cloning vector must be the correct size for it to be packaged into virus particles (between 38 and 51kb).

The vector lgt10 (43.8 kbases) has an *EcoRI* site inserted within the cI gene (lambda repressor) and so could accept a DNA fragment of 7.6 kbases before becoming too large to be packaged into virus particles.
FIGURE 31. The vectors lgt10 and Charon 16A

Other lambda vectors can incorporate DNA fragments of up to 22 kbases (e.g., the Charon series).

To use lgt10, we would cut the vector DNA and the DNA to be inserted with EcoRI, incubate the two together, seal with ligase and incorporate the recombinant molecule into particles in the test tube:

*Open lgt10 with EcoRI and insert double stranded DNA also cleaved with EcoRI. Seal with ligase.*

*Package the DNA into particles with a mix of head and tail precursors.*
*Infect E. coli with the reconstituted virus particles containing the new gene and recover the modified bacteriophage from infected cells as described above.*

4.2 Cosmid vectors

All of the lambda DNA sequence may be removed, except for the cos sites to allow larger pieces of DNA to be inserted. A site for the origin of replication (ori) and perhaps an antibiotic resistance gene also are retained:

*Insert DNA from 33 to 47 kbases in length into the cos site fragments using an appropriate restriction endonuclease.*
*Seal with ligase.*
Package the DNA into particles with the in vitro packaging mixture.

Since no viral genes are included in the vector, they replicate as plasmids in *E. coli*. The amplified DNA or its gene product may be isolated from the infected cells.

**FIGURE 32.** Assembly of λ phage in the test tube

*Infect cells, allow the cosmid to replicate, and isolate DNA.* This method allows insertion of larger pieces of DNA and also has the advantage of efficient infection of cells by the bacteriophage particles. Recall that transformation of cells by plasmids alone is relatively inefficient (perhaps 1 in 1000 molecules) and is usually limited to inserted DNA fragments of up to 15 kb.

**FIGURE 33.** Incorporation of DNA into λ--derived particles
4.3 Finding genes: DNA Libraries

We now have all the tools at hand that we need to actually clone a gene. Each new cloning problem is different from the previous one and our experimental approach must be planned carefully. The first major decision that must be made is whether to clone from the genomic DNA or to start with the messenger RNA that has the final code for a protein of interest.

Why do the two approaches differ? You may have to review this part of the molecular biology section before proceeding. Recall that the DNA of both bacteria and higher organisms has information at the 5’ end of a gene that is related to control of that gene—turning it on or off. You may or may not want to include this information in your cloned gene.

Most important, the typical DNA gene from higher organisms usually contains introns and exons and the sequences that are a part of introns must be cleaved out of the mRNA copy before it may be used to code for a protein. Thus, if you want to read a gene from a higher organism (such a human) in a bacterium (such as E. coli), to produce a protein you will want to start with the sequences found in mature mRNA. We then need to make a DNA copy (called cDNA) of the mRNA so that we may use the cloning techniques we have just learned. If the host cell is to be an eukaryotic one, you most likely will want to isolate the complete genomic DNA.

cDNA libraries

In our example, we will devise an approach to isolate a set of mRNA molecules from cells that are actively expressing the gene of interest, make DNA copies of the RNA and clone these DNA molecules into E. coli.
Isolate mRNA from cells that are actively expressing the gene of interest. Most cells express hundreds of genes at the same time, but often we can take advantage of some specialized tissue that is producing much more of the protein we are interested in than do other cells or tissues. It then follows that these cells will be making increased amounts of the specific mRNA that we want. Examples include liver cells producing albumin, lymphocytes making antibodies, and our favorite: beta-cells of the pancreas producing insulin (the only cells of the body to do so). If we isolate mRNA from these special
tissues, we increase our chances of being able to capture the correct mRNA in our cDNA library.

**Use reverse transcriptase to make DNA copies of the RNA.** Reverse transcriptase is an enzyme that is found in some viruses, where it makes DNA copies of RNA viral genes. We also can make use of another trick by remembering that most eukaryotic mRNAs have a string of adenine nucleotides at their 3’ end so that we can isolate the whole mRNA on a column containing poly dT bound to beads. The poly A tail will aneal to the poly dT and be captured on the column. The column may be washed to remove other nucleic acids and protein and then the mRNA may be eluted. Remember also that polymerases need a primer; now we can use poly dT for this purpose. We might also save only the mRNAs that are of the approximate size that we suspect would be needed to code for our protein.

**Use terminal transferase + dCTP to add a poly dC tail which will serve as a primer for second-strand synthesis.** We need a primer to start the second strand synthesis, just as we needed one to start synthesis of the first strand. A convenient way of providing this is use an enzyme known as terminal transferase to add cytidine residues to the 3’ end of the newly synthesized DNA using dCTP as the substrate.

If we separate the strands in alkali, the original RNA will be hydrolyzed and only our new cDNA copy will be left. We now may add an oligo dG primer and synthesize the second strand using DNA polymerase. We now have created a full length double-stranded DNA copy of the original mRNA. We may also add restriction endonuclease sites to the ends of the molecules (see below).

**Insert the DNA copies into a vector (a plasmid or bacteriophage l) after creating “sticky ends” for annealing and to provide sites for restriction endonucleases.** One of several methods may be used at this point. For example, we could add dC tails to each end of the cDNA using terminal transferase. Then we might cut the plasmid with *Pst*I and add dG tails to the exposed sticky ends, again using terminal transferase. The DNA now inserts readily into the vector.
The gaps may be filled in with DNA polymerase after annealing. Note that the $PstI$ site is preserved in the recombinant DNA and will provide an easy way to clip out the new DNA after amplification and isolation. Another method would be to use blunt end ligation (which occurs with low efficiency) to ligate “linker” oligonucleotides to each end of the cDNA. The linkers would contain one or more restriction sites.

**Isolate many different clones that contain the vector with inserted DNA sequences to create a cDNA library.** Presumably we now have a complex mixture of cDNAs inserted into vectors. The cDNAs represent sequences of all the major mRNA that were present in the cells. We might wish to save all those colonies that contain plasmids with functional inserts (detected by antibiotic resistance tests, etc. as described before). We even could go back to this library at a latter date and screen for other genes of interest that might be represented among the cDNAs.

**Screen for the desired clone using any of the techniques discussed in the previous sections. Antibodies specific for the new protein or radioactive nucleic acid probes are particularly useful.** We now need to screen our library more specifically for the gene of interest. We need quite a specific test in order to select the correctly transformed cell colonies. If we are looking for the insulin gene, we might test with antibodies for insulin (radioactive antibodies will bind to those colonies producing insulin). Since we know the amino acid sequence for insulin, we could synthesize a radioactive nucleic acid probe that will anneal to the cDNA and perform this test (as described previously).
FIGURE 37. A library from mRNA
Genomic Libraries

The goal of creating genomic libraries is to fragment the entire genome of an organism into a series of overlapping fragments and incorporate them all into suitable vectors. The library (or clone bank) may be stored as transformed bacteria (in the case of plasmid vectors) or infected bacteria or bacteriophages (in the case of lambda vectors).

**Estimate the number of clones needed to produce a library that will include all of genes of an organism.** This estimated number will depend heavily on the size of the genome and will be much lower for the gene of a bacterium than for a human. \( N = \frac{\ln(1-P)}{\ln(1-a/b)} \) helps to estimate the number of clones needed where \( P \) is the probability that the desired sequence will be represented in the library (usually 0.95-0.99), \( a \) is the average size of the DNA fragments, and \( b \) is the size of the genome. For the human genome, we might have to screen about one half million clones of average size of 20 kbases pairs in order to find a gene present as a single copy, because the human genome has a total of \( 3 \times 10^9 \) base pairs!

**Cut genomic DNA randomly into fragments suitable for cloning.** DNA may be sheared (simply be passing DNA solutions through a fine hypodermic needle) but no restriction sites will be generated. We noted that restriction endonucleases that recognize six base pairs give fragments of about 4096 base pairs which is on the small side if we hope to capture complete genes. Also, we know nothing about the actual distribution of restriction sites (such as for *EcoRI*) within a given gene. For these reasons, the usual procedure is to use a restriction enzyme that cuts frequently, on average, (such as *Sau3A*; once every 256 bases) but only allow a partial digestion of the DNA. Under these conditions, cutting is nearly random and restriction endonuclease sticky ends are produced. Either the time of the reaction or the enzyme concentration may be adjusted to obtain the desired size range of fragments.
FIGURE 38. Selecting sizes of nucleic acid to clone. Genomic DNA was cleaved with increasing amounts of Sau3A giving progressively smaller fragments. The desired size may be isolated from the gel and used for cloning.

DNA of suitable size for cloning may be purified by gel electrophoresis. Bacteriophage lambda or cosmids usually are the vectors of choice for DNA libraries because 20 Kb fragments and above can be inserted and cloned.

The vector is opened with a suitable restriction endonuclease, mixed with the DNA fragments, annealed and ligated. If Sau3A was used in the fragmentation of DNA, the vector may be opened with BamHI to produce the correct cohesive ends. If lambda is used, the recombinant molecules may be incorporated into particles in vitro and the library stored in this way.

Bacteriophage lambda libraries may be amplified by one passage through E. coli, stored and used to screen for many different genes. Indeed, libraries based on lambda are available commercially and represent an easy starting point for screening for human genes.

Specific screening techniques (such as antibodies or nucleic acid probes) are used to select transformed clones of interest.
4.4 More Cloning

Several variations on the techniques just presented will illustrate practical cloning solutions (see FIG. 37). In the first, a cDNA library is constructed. Instead of adding a tail of dC residues to provide a binding site for second strand primer, the investigators relied on the ability of reverse transcriptase to make hairpin turn at the 3’ end of the new strand. This short hairpin now provides a primer to complete the second strand. The hairpin may be cleaved by S1 nuclease that makes cuts in single stranded regions, such as the extreme end of the hairpin would be. Another feature of this cloning is the attachment of short adapters molecules to each end of the new double stranded DNA using DNA ligase (blunt-ended ligation). These adapters contain one or more restriction sites such as the one shown for HinDIII. The plasmid also was opened with HinDIII for insertion of the DNA.

The next example (FIG. 39) illustrates the construction of a genomic library. The genomic DNA was digested with limiting amounts of HaeIII and AluI to give DNA fragments in the 10--50 Kbp range. The DNA was fractionated on agarose gels and fragments of about 20 Kbp were selected for the cloning. The fragments were methylated at sites that would protect against EcoRI action (can you explain why?), and then an adapter was added to each end with ligase. The completed molecule was digested with EcoRI to create sticky ended sites. A Charon 4A was selected and annealed to bind together the cos sites, and then the circular molecule was digested with EcoRI. The large, central fragment was isolated (containing the cos site and EcoRI sticky ends) and ligated with the DNA fragments to obtain long, linear molecules that could be packaged in vitro into new phage particles. The phage particles constitute the library, or the library may be expanded by growth in E. coli.
At this point we could prepare a cDNA library starting from pancreatic beta-cells that should contain the mRNA for proinsulin. The series of steps would be very similar to those described in an earlier section, except we would use reverse transcriptase to create the cDNA copy instead of chemically synthesizing the DNA genes for the separate A and B chains.
chains. We again would add the code for methionine so that we may split out the finished proinsulin from the beta-galactosidase fusion product using cyanogen bromide. After the proinsulin is isolated, we need to split out and discard the C peptide using protease enzymes. We also could use another approach and isolate the insulin gene from a genomic library as well.
FIGURE 40. Cloning the complete gene for insulin
There are special enzymes used to cleave out the C peptide that are found in pancreatic beta-cells. These special peptidases are used by the cell to specifically produce mature insulin without accumulation of unwanted side products.
Chapter 5 Polymerase Chain Reaction (PCR)

5.1 Introduction

Polymerase Chain Reaction (or PCR) was invented by Kary Mullis as a method to amplify a specific sequence of DNA millions of times in the test tube. Thus, we may think of PCR as an *in vitro* technique for gene cloning. Instead of plasmids replicating inside of bacteria, a DNA template, suitable primers and substrates together with a special DNA polymerase carry out the process in a test tube. Often the DNA from a single cell--such as a single hair or sperm is sufficient to carry out the procedure.

FIGURE 41. The first step in PCR

5.2 The Reaction

The reaction mixture is heated to denature the DNA and separate the double strands. It is not necessary to have a pure DNA to start the procedure. It is necessary to
know some of the base sequence at each end of the DNA that you wish to amplify and to prepare single-stranded primers containing the correct base sequences.

The temperature is lowered to allow the primers (added in excess) to bind to the 3’ regions of the DNA strands.

DNA polymerase then synthesizes the complementary strand on each DNA. The DNA polymerase was isolated from an organism that lives at high temperatures so that the enzyme is stable to heating to over 90° and is enzymatically active at 70°.

The heating and cooling cycle is repeated many times. Each time the concentration of the newly-amplified DNA should double. Notice that the first round of synthesis produces new strands of indeterminate length, but the products become more uniform as the reaction proceeds. This happens because the products are soon defined by the primers at each end of the desired sequence. Commercial thermal cyclers are available that will precisely automate the many heating and cooling cycles.

A particular gene often may be amplified to microgram quantities in as little as four hours.
FIGURE 42. Multicycle PCR
5.3 Reaction Particulars

Reaction conditions
Specific conditions for maximal yields of a given DNA sequence usually have to be worked out. As a starting point, the general reaction mixture (in 0.1 ml) may contain:

- About 100 ng of complex DNA (~$10^4$ molecules)
- Buffer of 50 mM KCl, 10 mM Tris.HCl (pH 8.4), 1 mg/ml gelatin
- 1.5 mM MgCl2
- 0.2 mM of each deoxynucleotide triphosphate (dATP, dCTP, dGTP and dTTP)
- 50 mM of each primer (one for each end of the sequence)
- 2.5 Units of Taq DNA polymerase
- A drop of mineral oil to cover the reaction mixture and prevent evaporation

The concentration of Mg ions is quite critical, and may have to be determined by testing various concentrations.

Typically, the thermal cycler will be set to denature at 94°C for 20 sec, anneal the primers to the DNA at 55°C for 20 sec, and allow the polymerase to synthesize DNA at 72°C for 30 sec. The machines may heat at about 0.3°C per sec and cool at about 1°C per sec, so that a complete cycle takes 3.75 min. The automated machine might run this program for 20 to 30 cycles to amplify the desired sequence.


Primer Selection
Select primers (if possible) that do not have repeat stretches of one base, or primers that might fold over on themselves to form internal hairpin structures, or that might hydrogen bond with each other. Primers should be between 20 and 30 nucleotides in length and be of typical G plus C content. If only shorter primers are available (12 to 15 nucleotides) the polymerase reaction temperature will have to be lowered to perhaps 50-60°C so that the primers do not dissociate from the template DNA before polymerase can act. Unrelated nucleotide sequences may be added onto the 5’ end of primers to provide other functions in the product. Different sequences could be placed at each end. An example might be the addition of an endonuclease restriction site]
The DNA polymerase
The PCR method first used DNA polymerase I from *E. coli* (the Klenow fragment). This enzyme has an ordinary temperature stability, and so fresh enzyme had to be added after each denaturation step. The *Taq* polymerase was isolated from *Thermus aquaticus* which is was found in a hot spring in Yellowstone National Park. The microorganism grows optimally at 70-75° and the isolated enzyme catalyzes DNA synthesis optimally at 75-80°. It is quite stable at still higher temperatures and retains about 50% of its activity after one hr at 94°. Thus, the enzyme needs to be added only once at the beginning of the procedure. In contrast to the *E. coli* enzyme, *Taq* polymerase has no “editing” function, and may make more errors in synthesizing DNA. Usually this is not a serious problem, but if the product is used to determine the base sequence of a gene, several cloned PCR products should be sequenced to be sure no errors have crept in.

The DNA sample
Often, whole cells may be added to the reaction mixture and the high temperature and alkaline conditions of the denaturation step will lyse the cells and liberate DNA. This procedure works with isolated white blood cells, but not with whole blood. Hematin and other porphyrin breakdown products of hemoglobin were found to inhibit PCR. In some cases, it may be preferable to isolate and quantitate the DNA before using it in PCR.

A procedure for plucked hairs is as follows:
Cut off about 0.5 cm at the root end (containing the cell body). Place in PCR buffer with nonionic detergents (0.45 % NP40 and 0.45% Tween 20) and Proteinase K (0.01 mg in 0.1 ml). Incubate at 50° for one hour to digest cell proteins. Incubate at 95° for 10 min to inactivate Proteinase K. Proceed with PCR or store frozen.

The product
Usually the approximate size of the gene being isolated is known or suspected. Gel electrophoresis should be used to verify that PCR yielded predominantly one product in the correct size range. If impurities are detected, often the DNA in the band of the correct size may be eluted from the gel and used. The DNA may be precipitated with ethanol, redissolved and used in DNA sequencing reactions (see above reference for details). PCR also may be used to produce mostly single-stranded product by using one primer in 100-fold excess over the other.
**Reaction plateau**

Amplification usually slows or stops after 20-30 cycles. The reason for this is believed to be that the Taq polymerase cannot keep up with the many strands to be copied after their concentration reaches high levels. Another cause might be that the new strands will anneal with one another before the primers can bind and be extended. If more product is needed, it is simplest to run as many duplicate tubes as required.]
Chapter 6 Insulin and Diabetes

6.1 Insulin

The normal level of glucose in blood is between 80-90 mg/100ml or 4-6 mMolar. When the level of glucose goes up (after eating a meal), the beta (b) cells of the pancreas secrete insulin. When the level of glucose begins to fall, the alpha cells of the pancreas secrete glucagon which acts mostly on the liver to liberate glucose from stored glycogen. Thus, the effects of the two hormones counterbalance each other.

**FIGURE 43.** Relationships between insulin and glucagon action

For insulin to act, it must bind to insulin receptors. These receptors are membrane proteins found on many tissues (see diagram). The action of insulin is complex and changes with time.
FIGURE 44. The binding of insulin to its receptor leads to complex intercellular reactions.

**Short term:**

Within seconds after insulin binds to a receptor complex, the receptor proteins become phosphorylated and a receptor tyrosine kinase is activated which may lead to other cellular proteins becoming phosphorylated.

Within minutes there is a large increase in transport of glucose due to a large migration of glucose transporter proteins from the interior of fat and muscle cells to the cell membrane. Liver always has a very large capacity to take up glucose. The excess glucose is converted into glycogen for storage. A change in gene regulation occurs, leading to turning on of new genes. Some insulin receptor in taken into the cell.
**Longer term:**

There is an increase in glycogen synthesis in liver, fat synthesis increases in fat cells, and growth-promoting factors are synthesized (insulin-like growth factor). As glucose levels fall in blood, insulin secretion also declines. Soon glucagon is released which stimulates glycogen breakdown in liver with release of glucose into the circulation and activation of a hormone-sensitive lipase in fat cells which stimulates release of fatty acids to be used as alternate energy sources to glucose.

**Glucose Flux:**

Humans use about 10 gm/hr of glucose when resting and 40 gm/hr during exercise. After a meal, about 30 gm/hr may pour into the blood stream. This glucose must be rapidly removed by the liver and muscle and stored as glycogen.
FIGURE 45. Glucose flux in the resting state (A) and after eating (B). There are rapid increases in glucose transporters in muscle (squares) and fat cells (squares) in respond to insulin action.
6.2 Diabetes
About 2% of the USA population suffer from diabetes. There are two main types.

**Insulin-dependent diabetes mellitus (IDDM).** The pancreas stops making insulin, which then must be supplied (usually by injection). IDDM accounts for 10-25% of cases. Patients do not make enough insulin because of autoimmune destruction of the beta cells. Insulin must be given by injection on a carefully controlled basis. Lack of insulin results in high blood glucose, ketoacidosis and even coma and death.

**Non-insulin-dependent diabetes mellitus (NIDDM).** Is very common. The beta-cells seem normal, but fail to increase their secretion of insulin following a rise in blood glucose. NIDDM accounts for 75-90% of cases. The beta cells appear normal, but do not respond to glucose by producing insulin. Common in older people, especially if overweight. Often may be controlled by diet and weight loss.
References

Text:


References for Engineering Biotechnology:


Recommended articles from Primrose: *Milestones in Biotechnology*:

Itakura, K., et al., Expression in Echerichia coli of a chemically synthesized gene for the hormone somatosatatin (1977) p 84

Southern, E., Detection of specific sequences among DNA fragments separated by gel electrophoresis (1975) p 122

Suggs, S. et al., Use of synthetic oligonucleotides as hybridization probes: Isolation of cloned cDNA sequences for human b2-microglobulin (1981) p 140

Ullrich, S., et al., Rat insulin genes: Construction of plasmids containing the coding sequences (1977) p 243

Innis, M., et al., DNA sequencing with Thermus aquaticus DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA (1988) p 7