Intro to Molecular Biology: Basic Concepts and Applications in Research

Objectives: To understand the basic process by which cells access the information encoded in the genome.
To understand the concepts and applications of basic experimental methods used in cell and molecular biology to study genes, cellular structure and function.

When completed, the student should be able to understand and describe/explain/discuss:

1. Basic components and structure of nucleic acids.
2. Structural and functional differences between RNA and DNA
3. Role of histones in higher order structure of DNA and in modulating transcription of genes
4. Components and structural organization of a typical eukaryotic gene including promoter, exons, introns, untranslated regions, coding regions
5. Basic steps by which the cell generates proteins, starting from the DNA. (transcription, RNA processing, translation, protein processing)
6. Codons and how they relate to protein synthesis
7. Major types of RNA in the cell and their functions
8. Differences between primary, secondary and tertiary protein structure.
9. To understand the basic principles and applications of molecular biology in research and diagnostics, including PCR, restriction enzymes, cloning, western blot analysis, DNA sequencing, gene chips.

Recommended Reading (on reserve in the library)
Introduction to Genetic Analysis; Griffiths et al. 11th edition
Chapters 1, 7, 8, 9, 10, 11, 12 cover material presented in class.

NOTE: The book presents this material in more detail than we can cover in the time allowed. You are not responsible for material that is not included in the handout or in lecture. However, it may be useful to read the text for a better understanding any material that is unfamiliar to you.

General class policies:

- Office Hours by appointment: e-mail: dotteson@central.uh.edu
- Students may not audio or video record lectures without prior permission from Dr. Otteson.
- It is considered a violation of the Academic Honesty Policy for students to discuss the content of the examination with any of their classmates who have yet to complete the exam. Students are expected to comply with the UH Academic Honesty Policy with regards to all aspects of the course. The Academic Honesty Policy can be found and read in its entirety at: http://www.uh.edu/provost/academic-affairs/policy-guidelines/honesty-policy/
- In accordance with 504/ADA guidelines, we will attempt to provide reasonable academic accommodations to students who request and require them. Please notify your instructor if you qualify.
I. Background

The success of the human genome project is beginning to change medical practice by providing new tools for diagnosis and treatment of diseases. Among the goals of genome based medicine is to identify ‘biomarkers’ in the genome that will allow us to predict who will benefit from a particular therapy, who will have an adverse effect from a drug, who is likely to develop a particular disease/condition. In addition, by understanding the patho-physiology and genetic basis of diseases, it is possible to develop gene-based therapies for treatment. Molecular biology provides the tools needed for success in these approaches.

Human Genome project:

1990 – 2003 Original Human Genome project
  time: 13 years  cost: $3 billion

2004 The National Institutes of Health has issued a challenge to produce a sequencing method that costs less than $10,000 per genome by 2009, and a method for $1,000 or less by 2014.

2006 Archon X Prize for Genomics “build a device and use it to sequence 100 human genomes within 10 days or less, with an accuracy of no more than one error in every 100,000 bases sequenced, with sequences accurately covering at least 98% of the genome, and at a recurring cost of no more than $10,000 per genome.”

2007 Nobel laureate James Watson – co-discoverer of the DNA double helix and father of the Human Genome Project –the first human to receive the data that encompass his personal genome sequence.  time: 2 months  cost: $2 million

2009 Illumina launches personal genome sequencing $48,000

2010 Illumina reduces “personal genome sequencing” pricing
  $19,500 for individuals
  $9500 for physicians ordering sequencing for clinically relevant studies

2012 $50,000 benchtop sequencing instrument available from Life Technologies
  Estimate costs for sequencing @ $1000 per genome in 2 hours

2013 Archon cancels X prize competition claiming that there was insufficient interest and that technological advances have outpaced expectations, making the prize irrelevant; the Harvard team disagreed.

2015 Next big challenges
  Reduce genome sequencing to $100
  Sequence more genomes from more individuals to compare
  How to interpret data???

from: National Human Genome Research Institute, NIH
A. Applications of Molecular Biology and Genetics in Biomedical Research and Medicine

- Understand processes regulate normal and abnormal function in cells and tissues
- Identify genetic factors that cause disease
- Identify mutated genes and how these result in disease
- Understanding normal function of genes
- Predict susceptible individuals/prevention
- Identify biomarkers
- Develop gene replacement/corrective therapies
- Identify new drug/treatment targets
- Identify patients who will benefit from different therapies (different causes = different therapies)

B. Problems that must be overcome for success in gene-based and regenerative medicine

a) How to treat the underlying cause of disease?
   - Stop pathological processes
   - Eliminate malfunctioning proteins
   - Replace missing metabolites/proteins/genes
   - Add genes/cells that will make therapeutic peptide factors

b) To do this we need to:
   - Understand the pathological processes
   - Identify mutated genes
   - Understanding normal function of genes
   - Find methods to deliver the therapy into the eye/into specific cells
   - Specificity
     - Target to affected cells
     - Reduce unwanted side-effects
     - Regulated expression level
   - May need to also eliminate mutant protein and/or RNA to eliminate toxic effects
   - Sustained delivery
     - Eliminate need for frequent injections/drops/pills
     - Minimize need for repeated surgical interventions

c) Timing: early diagnosis is critical to be able to stop/reverse the disease process before too much damage has occurred.

d) Ethics: is likely to be a problem with stem-cell therapies for regeneration of damaged tissues
C. Applications of DNA technologies in research, diagnosis and therapies
   a) Identify mutations in genes that cause the disease
      • Genetics can be used to localize disease phenotypes to particular regions of
        chromosome and can be identified in the region of interest.
      • Candidate genes with a known or predicted function that makes sense with
        the disease of interest and/or that are expressed (turned on) in the affected
        tissues can be identified and studied
   b) GWAS: genome-wide association studies: Look for sequence variants in the genomic
        DNA that are differentially present in affected and unaffected individuals
      • Many of these variants are SNP: single nucleotide polymorphism.
      • Single nucleotides within the genome that vary between individuals
      • Used in mapping gene/disease locations
      • Often found outside of gene
      • A specific SNP may or may not have direct health effects (depends on
        location and molecular change)
   c) Looking for Linkage: Traits (SNP, genes, phenotypes) that are inherited together
      more frequently than predicted by chance alone are said to be linked
      • Two traits (e.g disease and SNP) that are physically close within a
        chromosome are more likely to be inherited together. So they are ‘linked’
   d) Approach: Sequence DNA from affected and unaffected individuals (can be in a single
      family or in larger groups of individuals) and identify DNA changes that associate with
      disease state
      • Evaluate changes
      • Look for “candidate” genes in region of SNP to see if it could be related to
        disease process? Expressed in relevant tissues? Involved in relevant cellular
        processes?
      • Are there DNA sequence changes in a gene? Could they alter the protein?
      • Clone the gene to test function in vivo and in vitro
      • Determine pattern of expression
      • Test effects of changes/mutations in animal models (make
        transgenics/knockouts)
      • Use sequence/mutation information to generate diagnostic tests for mutation
        and for gene replacement therapies.
   e) Once you found a mutation, you have to prove that it causes the disease! How?
      • Clone normal and “mutant” genes
      • Analyze RNA and Protein expression (normal vs. mutant)
      • Determine function
• Compare function of normal and mutant proteins
• Generate animal model (gene knock-out/gene replacement with mutated sequence)
• Other.........?

To understand these approaches, you have to have some basic knowledge of cell and molecular biology and the molecular basis of genetics.

II. General Cell and Molecular Biology

What is molecular biology? A powerful approach for studying (a) cellular/biological/biochemical processes at the molecular level and within biological systems

**Molecular Biology tends to be “gene” based.** All multicellular organisms grow from a single egg. Since each egg contains the DNA in the nucleus that codes the genetic ‘blueprint’ or code that tells the cell how to make all of the proteins necessary for building the mature organism. This information is encoded in the DNA

**Central dogma of Molecular Biology** is that the genes encoded in the DNA are transcribed to RNA, which is translated into Proteins. Understanding the cellular mechanisms underlying these processes is the basic for molecular genetics and allows us to manipulate genetic materials and study cellular processes, gene and protein function and associated disease processes.
A. DNA (deoxyribonucleic acid)

a) DNA is organized into chromosomes
- The entire collection of DNA (chromosomes) in the cell is called the **GENOME**
- All cells in the body contain the entire genome.
- Because we have two copies of every chromosome, our genomes are **DIPOID**
- Germ cells (egg and sperm) have only one set of chromosomes and therefore are **HAPLOID**.
- Humans have 46 chromosomes in total: 23 from each parent; these chromosomes are classified as autosomes (22 x 2) and sex chromosomes (X and Y)
- Different organisms have different numbers of chromosomes

b) The functional units of the chromosomes are called **genes**
- Human genome contains between 20,000 and 30,000 genes
- Most genes code for proteins (more on this later)
- Mitochondria also contain DNA that codes for some of the mitochondrial proteins

c) DNA structure, function and analysis

**Chemical Composition:** Components of DNA:
- 4 Nitrogen bases: two types of bases: purines (two rings) and pyrimidines (one ring)
- Deoxyribose sugar
- Phosphate groups
- Ribose sugars connected by phosphodiester bonds

**d) Nitrogen bases**

<table>
<thead>
<tr>
<th>Purines</th>
<th>(first synthesized in laboratory: ~1888)</th>
<th><strong>Pyrimidines</strong> (first synthesized in laboratory: ~1879)</th>
</tr>
</thead>
<tbody>
<tr>
<td>abbreviation</td>
<td>base</td>
<td>nucleoside*</td>
</tr>
<tr>
<td>A</td>
<td>adenine</td>
<td>adenosine</td>
</tr>
<tr>
<td>G</td>
<td>guanine</td>
<td>guanosine</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>cytosine</td>
<td>cytidine</td>
</tr>
<tr>
<td>D</td>
<td>thymine</td>
<td>thymidine</td>
</tr>
</tbody>
</table>

*Base + deoxyribose = nucleoside

**Base + phospho-deoxyribose = nucleotide (= nucleoside with phosphate)**

e) **Deoxyribose sugar:**
• Each base is attached to a deoxyribose (5 carbons) sugar. For DNA, the ribose lacks an oxygen group from one of the carbons (= de-oxy)
• **Carbons are numbered 1 to 5; the** base is attached at carbon C1
• **5’ and 3’ carbons** indicate orientation of DNA strand (phosphate attached to 5’ carbon)

f) **Base pairing (“Watson-Crick Pairing”)**
- Purine always pairs with pyrimidine
- A pairs with T (2 Hydrogen bonds)
- G pairs with C (3 Hydrogen bonds)
- The pairs are held together by hydrogen bonds
- Therefore, in any piece of double stranded DNA:

  \[
  \% A = \% T \quad \text{and} \quad \% C = \% G
  \]

  But A does not have to equal C

Each species has unique DNA composition in terms of the overall \%G + \%C in its genomic DNA


g) **DNA structure determined by Watson and Crick (published 1953), using some of the data acquired from Franklin and Wilkins.**
- DNA is a polymer consists of two polymers that are oriented in opposite directions (anti-parallel)
- These polymers are twisted to form a double helix
- The two polymers are made up of nucleotides
- Nucleotides have three main components: a base; a ribose sugar and a phosphate groups
- DNA = Double-stranded alpha-helix with *anti-parallel orientation of 2 strands*

```
5’−CAGTGATTACA−3’
| | | | | | | | |
3’−GTCACTAATGT−5’
```

h) **DNA replication**
Before a cell can divide, it must copy the DNA so that each daughter cell receives the full set of chromosomes.
- Base pairing is the basis of DNA replication: the existing DNA is the template for its copy. Each copy is the reverse complement of the original strand.
• To do this, the DNA helix is unwound and the hydrogen bonds between the bases are broken.
• **DNA Polymerase enzymes** make copies of each strand of the DNA.
• The two resulting DNA helices will contain one strand that is the original DNA and one that is the newly synthesized DNA.
• This type of replication is called semi-conservative.
• **A KEY FEATURE OF DNA** is the ability of single stranded DNA to base pair with a second strand of DNA with the ‘reverse complement’ sequence.
  o This is basis for PCR, gene chip diagnostics, sequencing and more molecular technologies

i) **DNA: higher order structure**
• The overall length of the DNA in nucleus of a single cell is about 2 meters
• When condensed for mitosis, chromosomes are about 2 micrometers long.
• How can the cell pack all that DNA into the nucleus of the cell?
• DNA is organized into chromosomes which consist of DNA and proteins
• The DNA in the nucleus of eukaryotic cells is wrapped around proteins called histones.
  o Histones are octameric proteins consisting of 8 subunits.
  o The DNA wraps around the histone ~2 times to form a nucleosome.
  o Modifications of histone tails allow DNA to be in an open configuration (available for transcription) or to condensed (silenced)
    ▪ In general: increased acetylation of histone tails creates open structure, is found in transcribed genes
    ▪ Increased methylation creates closed/condensed structure, genes are not transcribed. These areas of silenced DNA also tend to have increased DNA methylation.

B. **RNA (ribonucleic acid)**
   a) **Genes are transcribed (copied) from the DNA to make RNA**
      • DNA in nucleus contains the genetic information for the cell
      • DNA is the template for making RNA
      • RNA copies of the genes are called transcripts
   b) **The entire collection of all transcripts in a cell are the TRANSCRIPTOME**
      • Different cells will have different transcripts
      • This is because only some genes that are needed in a specific type of cell, therefore a cell only transcribes the genes that it needs.
   c) **RNA composition**
RNA is a single stranded polymer
  ▪ consists of ribose sugar, purine or pyrimidine base and phosphate
  • Bases of RNA are: purines [Adenine, Guanine] and pyrimidines
   [Uracil (instead of Thymine), Cytosine]

Ribonucleotides are joined by phosphodiester bonds between ribose sugars

RNA is single stranded (but can fold on itself and base pair with itself)

d) There are different kinds of RNA
  • mRNA (messenger RNA)= protein coding RNA is the template for making
   proteins
  • functional RNAs = non-protein coding RNA is the functional end product
  types of functional RNA: rRNA = ribosomal RNA (Protein synthesis)
   tRNA = transfer RNA (Protein synthesis)
   snRNA = small nuclear RNA (RNA processing/splicing)
   microRNAs=regulatory RNAs: used by cell to silence gene
   cRNA = cyclic ribonucleotides (e.g. cAMP, cGMP) used as
   second-messengers in intracellular signaling (e.g. opening/closing gated channel proteins, phototransduction, etc.)

e) RNA and Transcription: RNA synthesis
  • Genes in the DNA encode the information needed by the cell. To access this
   information, the DNA must be transcribed to make RNA.
  • Orientation of genes on chromosome vary
    • Orientation is designated based on direction of transcription: 5’ end of gene ~ 5’ end of transcribed RNA.
    • Strand of DNA with the same sequence as the RNA (but recall that DNA has T, RNA as U), is called the coding strand or sense strand (because if you ‘read’ the sequence you can directly translate it to match the protein sequence – see protein synthesis)
    • The strand of DNA that is used to make the RNA copy is the template strand and
   it is the REVERSE COMPLEMENT of the RNA.

Organization of a eukaryotic gene (protein coding)
  • Eukaryotic genes are organized with a regulatory region (promoter, enhancers) that
   determine when a gene is transcribed and how much RNA is made
  • Promoter is located at the 5’ end of the gene (More details below).
  • The body of the gene contains exons and introns.
  • The entire gene (both exons and introns) are transcribed to make hnRNA
  • Exons contain information needed to code for protein (coding region)
  • Exons also contain untranslated regions (UTRs) that do not code for protein
REGULATORY REGIONS of genes

- **Promoter**: contains regulatory sequences that bind to proteins required for making the RNA and are important for instructing the cell which genes to activate or shut down.
  - Promoters contain specific DNA sequences that are recognized by transcription factors. Transcription factors control transcription.
  - There are general transcription factors that are used in all cells and are required for transcription of all genes. There are also tissue-specific transcription factors that control which cells will turn on which gene.
  - **Minimal or Basal promoter**: the smallest promoter region needed for transcription. Contains binding sites for RNA polymerase and some general transcription factors. Usually not sufficient for high level expression. Typically does not contain sites for transcription factors that regulate cell-specificity – can drive low level expression in most cell types.
  - **Proximal Promoter Region**: Located near transcription start site typically extends ~200 bp from the transcription start site and includes the basal promoter and adjacent promoter elements. Proximal promoter region contains sequences (elements) that are recognized by a variety of general and tissue-specific transcription factors needed for appropriate cell-specific transcription. Because proximal promoters are the site for assembly of the transcription machinery of the cell, it must be positioned in the right place or the gene will not be transcribed. Therefore, it is not surprising that they are position and orientation specific. If we alter the location or orientation of the proximal promoter relative to the gene, transcription will not proceed properly.

- **Enhancers**: Located much further from the transcription start site. Enhancers contain binding sites for additional regulatory proteins and transcription factors. These are often involved in regulating the overall activity of the promoter and play a role in cellular specificity of transcriptional activation. Because enhancers are located at a distance from the proximal promoter, the transcription factors that bind the enhancer can bend the DNA to bring it close to the proximal promoter region. Because of the flexibility of the DNA, enhancers are generally position and orientation independent. If you move them, the DNA just bends a bit more to bring them to the right place.

Transcription = the generation of RNA from a DNA template

- Transcription of RNA occurs in the nucleus of eukaryotic cells
- NOTE: prokaryotes and mitochondria have no nuclei and so transcription occurs in the cytoplasm! (not discussed here)
h) Regulation of transcription:
  - Each cell controls which genes are transcribed
    - Different cells transcribe different genes
  - Stress or other environmental signals can later gene expression patterns.
  - Regulation requires that transcription factors (proteins) and RNA polymerase (a multiprotein complex) bind to the promoter region and access the DNA template so the RNA polymerase can make the RNA copy.
  - Chromatin Structure
    - Because the DNA is organized on histones, genes cannot be transcribed until the histones are moved to expose the DNA.
    - To regulate access of transcription factors, DNA modifying enzymes and RNA polymerase to the promoter and enhancer regions
    - In eukaryotes, the default state of chromatin is condensed and tightly coiled. Thus, transcription is blocked and genes are ‘turned off’.
      - Closed chromatin/gene silencing
        - Densely packed nucleosomes
        - Histone modifications stabilize structure (deacetylation, methylation)
        - DNA methylation at CpG sequences can permanently shut down expression
        - Transcriptional regulatory are proteins excluded from heterochromatin
      - Open chromatin/gene transcription
        - Opening DNA structure is a dynamic process and is required for transcription
        - Open chromatin associated with increased acetylation of histone tails.
        - Histone modifications open the DNA structure (acetylation)
        - For transcription, histones are moved aside (sliding and/or disassembly)
        - This allows transcriptional regulatory proteins to bind to DNA
        - Transcription factors can recruit histone modifying enzyme, RNA POLYMERASE and other components of the transcription machinery
          - Modifications of histone tails by acetylation/deacetylation and methylation/demethylation alter the ‘tightness’ of chromatin condensation
            - Acetylation: open, loose structure, available for transcription
            - Methylation/deacetylation: closed, condensed, tight structure, DNA not transcribed
  i) Transcription:
    - Open DNA is bound by a variety of proteins to activate or repress transcription
      - These include general transcription factors (typically present in all cells)
- RNA polymerase complex
- Cell and tissue specific transcription factors (DNA binding proteins required for transcription.
- Many transcription factors are positive regulators (activators) of transcription and may recruit histone acetyltransferase proteins that further open DNA structure. Others bend the DNA to open it further for binding of additional transcription factors.
- Other transcription factors are negative regulators (repressors) and can recruit chromatin modifying proteins (histone deacetylases, histone methyltransferases, DNA methyltransferases) to close the chromatin structure and block transcription.
- The combination of transcription factors present in a cell determine which genes are transcribed or silenced.
  - Once chromatin is open and appropriate transcription factors are recruited, RNA polymerase transcribes the RNA

j) **THE TRANSCRIBED REGIONS of a GENE** = the portion of the gene that is ‘downstream’ from the promoter
  - For protein coding genes:
    RNA polymerase generates a full length RNA copy of the template strand of the entire transcribed region = Heterogeneous nuclear RNA (hnRNA):
    - hnRNA is the collection of primary transcripts in the nucleus
    - contains RNA copies of both exons and introns
    - Exons are the part of the gene that will end up in the final mature RNA. The regions coding for the protein are in the Exons. Exons also contain regions that do not code for proteins (untranslated region = UTR).
    - Introns are the intervening sequences located between the exons.
  - Various lengths of hnRNA for a specific gene can exist at the same time in a nucleus of a cell, since transcription is ongoing/dynamic process
  - hnRNA molecules for mRNAs are processed in nucleus during and following transcription

k) **PROCESSING of mRNA (eukaryotes)**
  - During transcription, protein coding hnRNAs are processed to become mature mRNA
    - As the gene is transcribed, the 5’ end of the hnRNA extends away from the RNA polymerase and is subject to damage and degradation. To protect the RNA and to generate a mature mRNA requires RNA PROCESSING.
- **NOTE:** Although we describe these as separate and sequential, in reality, RNA processing begins while the gene is being transcribed. As the hnRNA is synthesized,

  - **Capping:**
    - Addition of a modified nucleotide in the reverse orientation (methylated G) at the 5’ end of the transcript
    - Capped RNAs are stable and protected from nuclease degradation
    - The cap is also used as a recognition site for the ribosomes during protein synthesis.

  - **Splicing**
    - Removes introns from the hnRNA and joins of remaining exons in the RNA
    - This is done by a complex of snRNAs and proteins called the *splicosome*.
    - **Alternative splicing** results in removal/inclusion of different subsets of exons in a mRNA.
      - Alternative splicing allows multiple mRNAs and multiple proteins to be made from a single gene.
      - Alternative transcripts can yield multiple related, but structurally unique, proteins from a single gene.
      - Contributes to tissue specific regulation (different splice variant may be present in different cell types)

  - **Poly-adenylation:** Addition of a string of A bases to the 3’ end of the RNA; polyA tail further stabilizes the RNA and functions as the signal for nuclear export
    - When the RNA polymerase reaches the end of the gene, the RNA is cut (cleavage) and a poly-A tail is added (polyadenylation).
    - The polyA tail consists of 150-200 adenosine nucleotides
    - Attached to the 3’ end of the RNA (the opposite end from the cap)
    - Marks the end of the RNA.
    - The poly-A tail also increases RNA stability
    - Signals that the mRNA is mature
    - Signals for nuclear export of mRNA to the cytoplasm for translation
    - Only mRNA is poly adenylated (has a poly-A tail). Functional RNAs (rRNA, tRNA) do not have poly-A tails.

- **mRNA** = messenger RNA = mature RNA = protein coding RNA. This is fully processed protein coding RNA after the introns have been removed (splicing) and the rest of RNA processing (capping and polyadenylation) is complete. The mRNA is exported from the nucleus to the cytoplasm where it functions as the template for protein synthesis.
After RNA is transcribed from the DNA and processed in the nucleus, the mature mRNA (messenger RNA) is exported to the cytoplasm and used as a template for protein synthesis.

l) Transcription and processing of other RNAs prior to nuclear export:
- tRNAs are processed: cleavage, splicing, modification of bases, but are not polyadenylated
- rRNA are not spliced or polyadenylated

C. Proteins (polypeptides)

a) General:
- mRNA serves as the template for making proteins
- in cytoplasm, ribosomes bind the mRNA to synthesize the protein
- Translation of mRNA into protein uses the genetic code
- Free ribosomes make cytoplasmic proteins
- Ribosomes attached to the endoplasmic reticulum (ER) synthesize membrane bound and secreted proteins
- The collection of all proteins made by a cell is called the PROTEOME

b) Protein Structure
- Proteins are made of amino acids that are joined together to form a polypeptide.
- There are 20 amino acids
  - classified by their structure and chemical properties [polar, charged (positive or negative), uncharged, aromatic (ring structure)].
  - These different properties will contribute to the properties and shape of the final mature protein.
  - Substitution of one amino acid for a different amino acid (e.g. as a result of mutation) can have different effects on the protein depending on whether or not the new amino acid is in the same class as the original.
- Primary Structure of a protein: the linear sequence of amino acids, as coded in the mRNA transcript
- Secondary Structure of a protein: In the cell, most proteins are not configured as a straight strand, but are folded.
  - A single protein may have multiple domains that are folded into different shapes.
  - These domains constitute the secondary structure of the protein.
  - Two common secondary structures are a helix or coil and pleated sheet.
  - These are stabilized by hydrogen bonds between the strands.
• **Tertiary structure** is the 3-dimensional structure of the entire polypeptide (as opposed to just the structure of individual domains/regions within the protein = secondary structure).
  - Folding into final tertiary structure is controlled by proteins in the cell called **chaperones**, which bind to the newly generated proteins and help them to achieve their final correct structure.
  - (Interestingly, the crystallin proteins in the human lens are structurally very similar to chaperones).
  - The complex tertiary structure is stabilized by hydrogen bonds and other chemical bonds.
  - As a consequence of folding, amino acids that are located at different (non-adjacent) locations along the protein sequence can be brought into close association and often will create critical functional sites (e.g. binding sites for proteins, channel pores) that would not be predicted based on the primary structure.

• **Quaternary structure** refers to the formation of multi-protein complexes. Many mature proteins are composed of multiple subunits. These complex proteins are called multimers
  - each component is called a monomer
  - dimer = 2 subunits; trimer = 3 subunits; tetramer = 4 subunits, etc;
  - The subunits may be separate proteins that are encoded by different genes.
  - In some cases, the subunits are part of the same original protein that has been cut by protease enzymes to form 2 or more subunits that recombine to form the active protein.

c) **Genetic Code**
- The RNA ‘encodes’ the amino acid sequence of the protein in base triplets called **codons**.
  - Since there are 4 bases, there are $4 \times 4 \times 4 = 64$ possible combinations.
  - Some amino acids are coded for by more than one codon. These typically vary at the third position in the codon.
  - In addition there are three codons called stop codons that signal the end of the protein = **stop codons**.
  - The codons are read in sequence and do not overlap.
  - The sequence of codons for the protein starting with the start codon and ending with the stop codon is called the **open reading frame**.
  - If you **mutate** one base and change the codon, it will only change one amino acid in the protein
  - If you **delete a base or insert an extra base**, it will cause a ‘frame shift’ that can alter all of the downstream codons that follow the deleted/inserted base.
d) Translating RNA sequence into proteins

- A **reading frame** is the continuous sequence of adjacent codons within the RNA.
- Since the codons consist of three bases, there are three possible reading frames in each mRNA, depending on which base is chosen to be the first base of the first codon.
- In most genes there is only one correct reading frame that will code for the protein and this is called the **open reading frame**.
- It is typically the longest continuous sequence of codons that is bounded at each end by a start codon and a stop codon.
- The other reading frames typically will have many stop codons that would result in formation of very small proteins if they were mistakenly translated.

e) Protein synthesis

- **Machinery of TRANSLATION**
  - mRNA: The processed mRNA (after capping, splicing, poly-adenylation) is exported from the nucleus into the cytoplasm
  - Amino acids
  - tRNA: transfer RNA: this is a small RNA molecule that carries an amino acid to the ribosome. There are multiple tRNAs with unique structures: each tRNA is linked to the appropriate amino acid by an enzyme based on structure. Each tRNA has a sequence called an **anti-codon** that is the reverse complement of a codon. This allows it to bind to the codon in the RNA sequence that codes for the amino acid carried by the tRNA.
  - Ribosomes: These are the cellular machines that translate an mRNA sequence into a protein. They make the proteins. The ribosomes consist of 2 subunits (small and large) and each of the subunits is composed of both ribosomal proteins and ribosomal RNA.

- **Steps of Protein Synthesis**
  - The small ribosomal subunit binds to the 5’ cap on the mRNA and scans along the mRNA looking for a start codon.
  - A start codon is **nearly always** AUG and codes for methionine.
  - The ribosome recognizes the correct start codon by the presence of a Kozak sequence immediately upstream of the AUG.
  - It binds to the DNA and recruits the tRNA carrying Methionine (Met) and the rest of the ribosome assembles.
  - There are three sites within the ribosome where the action of protein synthesis occurs. They are called A (acceptor site), P (peptidyl site) and E (exit site).
- Methionine tRNA positioned at the P site (first amino acid almost always methionine)
- Second tRNA recruited to A site
- Amino acid on tRNA in P site joined to amino acid at A site
- Ribosome translocates, moving tRNA with growing peptide chain from A to P site
- This puts the empty tRNA into E (exit site) and it is released
- The next tRNA is recruited to A site
- Repeat until ribosome reaches stop
- There is no tRNA for the stop codon. When the ribosome reaches a stop codon, a termination factor binds, the ribosome dissociates and releases peptide chain

  - There are additional modifications that can occur during and after protein synthesis.
    - If the protein is a cytosolic protein, it is translated on the free ribosomes.
    - If the protein is a transmembrane protein or a secreted protein, a signal sequence within the growing polypeptide chain will cause the ribosome to associate with the endoplasmic reticulum of the cell and the growing polypeptide chain is put through the membrane into the lumen of the endoplasmic reticulum. This creates what is called the rough endoplasmic reticulum because the surface of this part of the endoplasmic reticulum is studded with ribosomes generating proteins.
    - Proteins can be further modified by addition of various chemical groups. Most common are:
      - Proteolytic cleavage
      - Glycosylation: addition of sugar groups
      - Methylation: addition of methyl groups
      - Acetylation: addition of acetyl groups
      - Ubiquitination: addition of ubiquitin (targets protein for destruction in the lysozomes)
    - In addition, proteins must be sent to their proper destination.
      - There are specific sequences in the protein that signal its destination.
      - In addition to the signal sequence (above) there are nuclear localization signals.
      - These are typically on transcription factors and target the mature protein to the nucleus.
III. What can we do with Molecular Biology?

1. Clone and sequence DNA and RNA
   Why?
   - Identify all of the genes that are present at a particular developmental or disease state
   - Determine what genes are necessary for structure/function of specific cell types (e.g. what are the genes expressed in a photoreceptor? How do these differ from those expressed by ganglion cells? or corneal keratocytes?)
   - Which genes are expressed in early vs. late development? healthy vs. diseased tissues?
   - Identify mutations associated with disease.

First need to identify the affected gene
- Genotype patients, mice
  - Sequence the whole genome?
  - Sequence the whole gene? Which parts?
    - What is the difference between an exon and an intron?
    - What is the coding region?
    - How do you get the pieces of DNA to sequence?

- Sequence exons containing coding regions of candidate genes in patients
  - How do you sequence DNA?
  - How do you identify “mutations” (sequence differences predicted to alter protein function) in patients with disease that are not present in normal individuals

- Study function of specific genes/proteins.
  - Generate large amounts of proteins in vitro to study structure, function.
  - Put new genes into cells to assess their function
  - Mutate genes to model genetic diseases; generate transgenic mice

- Develop gene therapy
  - replace missing genes in patients with disease
  - add in survival factors to slow progression.

- In genetic studies, patient DNA is isolated and analyzed for mutations.
  - After comparison to the genome databases, sequence changes are identified.
  - Often only a few genes of interest are analyzed (if you already know something about the mutations/disease)
  - Advanced sequencing technologies now allow analysis of the entire ‘exome’ (all exons in all known genes). More data, harder to find the change. Need high throughput computational resources.
  - Once mutation is identified, need to study effects on structure and function in vitro and in vivo.

- Bioinformatics: some useful genome and gene databases
  - National Library of Medicine/NIH
  - University of California Santa Cruz http://genome.ucsc.edu/
• Ensembl  http://www.ensembl.org
• There are many others, too numerous to list here

2. What are some of the methods used to look for mutations and analyze expression and function?

• To study a gene’s function or the effects of mutation, we have to isolate the DNA that codes for the gene and clone it. Once in a clone, we can generate the protein in vitro or in cells to study its biochemistry, overexpress normal or mutated proteins in cells to look at function, make transgenic/knockout animals to look at effects/phenotype.

• Since transcribed regions lack the introns present in genomic DNA, to study the gene function, we typically clone a DNA copy of the RNA called cDNA (complementary DNA, because it is complementary to the mRNA).

a) Isolate genomic DNA (gDNA): can be used for genotyping, to clone regulatory regions of DNA-gene promoters, to study splicing mechanisms, DNA/protein interactions, etc)
- Sources of DNA
  - Isolate DNA from patient/subject
  - Blood
  - Cheek swab
  - Biopsy (e.g. cancer)
  - DNA in all cells should be the same **

b) Isolate RNA and make DNA copy = cDNA (can be used to overexpress genes, synthesize proteins in vitro, generate gene replacement therapies, generate mutated versions of genes to model disease, etc)

1. When you purify total RNA from sample, it contains mixture of ribosomal RNA, tRNA and mRNA (messenger RNA= protein coding RNA)
- RNA is unstable
- single stranded
- easily degraded: tricky to keep intact
- May need to purify mRNA from total RNA
- mRNA has poly-A tail
- Poly-A tail can be used to selectively enrich poly-A RNA fraction

2. For easier manipulation of RNA: first convert to cDNA
- cDNA = complementary DNA
- Use Reverse transcriptase to generate a reverse complementary strand of DNA from the RNA template
- Can use DNA polymerase to generate 2nd strand of DNA to make double stranded DNA for cloning
3. **Reverse transcription reaction: convert RNA to cDNA**
   - for mRNA: use Oligo-dT primer anneals to polyA tail
   - for total RNA: can use random primers; will reverse transcribe all RNA (rRNA, tRNA, mRNA etc)
   - Reverse Transcriptase (from retrovirus) synthesizes the first strand DNA
   - DNA polymerase synthesizes 2nd strand of DNA
   - cDNA library contains DNA copies of only RNA present in starting tissue (Transcriptome)
   - No introns or regulatory sequences,
   - No non-coding sequences
   - Expressed (transcribed) genes only
   - Content variable: depends on RNA source

c) **Basic PCR amplification target sequence**

1. **Components needed:**
   - DNA template (gDNA or cDNA)
   - Primers (short pieces of single stranded DNA designed to be complementary to the two strands of the DNA)
   - dNTP (nucleotides for copying DNA = dATP, dCTP, dGTP, dTTP)
   - Taq DNA polymerase: heat stable, isolated from Thermus aquaticus
   - Some salts needed to make Taq polymerase happy
   - Thermocycler (allows controlled changes in temperature for the reactions)

2. **Steps for PCR reaction:**
   - **denature** (heat to melt DNA, separates strands of DNA)
   - **anneal** (cool; allows primers to bind to target sites on DNA)
   - **extend** (Taq polymerase synthesizes DNA beginning at end of primers)
   - **Repeat:** 35-45 cycles: amplifies target

3. **Standard PCR**
   - end point assay: target is present/absent
   - are there alternative splice forms of the RNA present
   - qualitative, not quantitative

d) **RT-qPCR = quantitative RT-PCR**
   - Sometimes we want to know how much of the target is present in our sample
   - Is the gene of interest expressed in a particular tissue?
   - Does expression change in disease state or with manipulation of other factors/signaling pathways?
   - How much of each different splice forms is present?
   - uses fluorescent stains/labels to quantify PCR product synthesis
   - reads fluorescence level at every cycle
e) **Clone DNA and PCR products**

1. Bacterial *restriction enzymes* can cut DNA at specific sites to generate smaller pieces that can be used to isolate regions of interest
   - The DNA inserts are cloned into plasmid or viral vectors cut with same restriction enzymes
   - This allows propagation of the plasmids or virus in bacteria or eukaryotic cells in the lab. Can generate large amounts of DNA and proteins for study.
   - Relies on antibiotic resistance genes in plasmids that allow selection of cells with the clone/plasmid/vector by killing cells that lack the plasmid

2. **Vectors used in Cloning:**
   a) **Plasmid Vectors**
      - Small, circular extrachromosomal DNA in bacteria
      - Replicate independently of host genome
      - Many copies/cell
      - Carry antibiotic resistance genes
      - Small inserts: <5,000 bp
      - Useful for cDNA, small genomic, PCR products
   b) **Virus Vectors (bacteriophage)**
      - Can be grown in bacterial cells
      - Many copies per cell
      - Larger inserts: ~15,000 bp
      - Useful for cloning genomic DNA
   a) **Cosmids**
      - Similar to plasmids, but hold huge DNA inserts
      - 35,000-45,000 bp; few copies per cell
   b) **YAC = yeast artificial chromosome**
      - 100,000 – 1,000,000 bp insert
   c) **BAC = bacterial artificial chromosome**
      - 100,000 – 200,000 bp insert
      - Maintained in host cell as independent chromosome
      - Passed to daughter cells
      - One copy per cell

3. **Growing plasmid clones in bacteria: Overview**
   - Add plasmid DNA to permeabilized bacterial cells
   - Plate cells on media with antibiotics
   - Antibiotic resistance gene in plasmid allows cells containing plasmid to grow in presence of antibiotic
• Once colonies identified, they can be expanded to generate many cells, frozen for storage and future use, purification of the plasmid DNA

f) Working with DNA

Once you have a PCR product or plasmid DNA how can you know if it is what you wanted?

1. Gel electrophoresis: Is it the correct (predicted) size?
   • Isolate DNA
   • Cut with restriction enzyme
   • Uses fluorescent stains to detect DNA or RNA in gel

2. Basic DNA Sequencing:
   • Sequencing: modern sequencing uses PCR based method but includes (1) only one primer and (2) modified nucleotides (di-deoxyribonucleotides) that terminate DNA synthesis when incorporated
   • Each reaction contains 4 deoxyribonucleotides (dATP, dCTP, dTTP and dGTP)
   • Also includes di-deoxyribonucleotides (ddATP, ddCTP, ddTTP or ddGTP). If using radioactive di-deoxynucleotides, need 4 reactions each containing only one of the 4 di-deoxynucleotides.
   • All reactions use a single gene-specific primer; copies only one strand of the DNA target. Only get linear amplification.
   • In each PCR cycle, you end up with a mixture of different sized PCR products.
   • Each product will end with the specific di-deoxyribonucleotide in the reaction.
   • After many cycles, each reaction will contain a mixture of products that will end at each of the bases for the dideoxyribonucleotide in the mix:
     • ddATP reaction, contains all possible fragments that end in A.
     • Mixtures of products are separated on a gel. Each lane contains one reaction.
     • Short fragments migrate quickly; long fragments migrate slowly. Over the 4 reactions, there will be one fragment ending on every nucleotide within the sequenced fragment.
     • Starting at the bottom of the gel, you can read the sequence of the products based on the end ddNTP incorporated into each reaction.
     • Modern sequencing uses fluorescent ddNTPs and they are all mixed in a single reaction. The products are separated in a capillary gel so that each fragment (band) travels past a detector that ‘reads’ the fluorescence of the terminal ddNTP

3. “Next Generation” Sequencing (massively parallel sequencing) for genome and transcriptome analysis
   • Allows sequencing of entire genome or transcriptome at one time.
   • Fractionate the DNA or cDNA to ~ 200 bp
   • Ligate on adaptors with ‘bar codes’-unique sequence for each library; allows multiplexing
• Attach to solid surface (flow cell or bead) and amplify each fragment in place to make cluster
• Clusters are sequences simultaneously

g) Analysis of RNA and Gene function
• Useful for looking at changes in splicing, tissue distribution of gene expression
• Must isolate RNA from affected tissue
• Widely used in animal studies
• Problematic for many eye diseases in human patients unless you can biopsy (enucleation, cancer) or isolate tissue sample (e.g. tears) for analysis

1. Analysis of mRNA Expression:
• RT-PCR and RT-qPCR: quantitative, uses RNA from whole tissues, isolated cells, etc.
• in situ Hybridization
  • Purpose: Determine the cellular expression of RNA within a tissue
  • Tissue: frozen or paraffin sections
  • Probe: anti-sense RNA (complementary to endogenous RNA)
    o make by in vitro transcription
    o labeled with digoxigenin protein-tagged ribonucleotides
  • Hybridize probe to RNA in tissue
  • Antibody detection of protein-tag; colorimetric detection of antibodies

2. Analysis of Protein Expression
• Immuno-histochemistry / Immuno-fluorescence
  • Purpose: Determine the cellular expression of protein within a tissue
  • Tissue: frozen or paraffin sections
  • Probe: Antibodies that recognize specific protein (primary antibody)
  • Detection: Tagged antibodies that recognize primary antibody
    o Immunohistochemistry: secondary tagged with enzyme; apply substrate for colorimetric detection. (e.g. Alkaline phosphatase or horseradish peroxidase)
    o Immunofluorescence: secondary antibody tagged with fluorescent molecule (Cy3, Cy5, Alexafluor, TRITC, FITC/Fluorescein)

3. Western Blot
• Purpose: look at global protein expression, size, post-translational modifications of proteins
• Tissue: isolated tissue, homogenized for proteins; can do sub-cellular fractionation (e.g. nuclear vs. cytoplasmic vs. membrane bound)
• Probe: Antibodies that recognize specific protein (primary antibody)
  • Separated by Poly-Acrylamide Gel Electrophoresis (PAGE)
  • Transferred to nitrocellulose filter
• Incubated with antibody against protein of interest (e.g. KLF15-transcription factor)
• Antibodies detected using enzymatic reaction that emits light for imaging
• Bands show protein size and amount
• Different isoforms of the protein may be present in different tissues’

4. **Transgenic Technology**
   - **Transgenic**
     - General term for genetically modified animal; typically refers to random gene insertion
   - **Knock-out**
     - Specific gene is deleted/ inactivated in the genome
   - **Knock-in**
     - Knocks out target gene and replaces with foreign gene

• **Why Create Transgenic Animals?**
  - To produce therapeutic proteins e.g. Transgenic sheep and goats to produce human proteins in their milk (insulin production)
  - Research tools
  - Develop models for studying disease
    - e.g. Lumican knockouts: study corneal structure
    - MMP-9 knockouts to study dry eye disease
    - Rhodopsin knock-ins to model *Retinitis pigmentosa*
  - Replace normal genes with mutated genes to determine consequences of mutation

• **Approach: Transgenic/random insertion**
  - Generate DNA construct in plasmid for overexpression (transgenic, random insertion) construct consists of promoter to drive expression, cDNA of interest (normal or mutated gene, reporter such as green fluorescent protein, etc).
  - May be modified to enhance integration using viral sequences
  - Introduce DNA into cell or oocyte (electroporated, microinject, virus)
  - Allow random insertion; grow up animal and study directly or breed to generate more; study cells in a dish.

• **Approach: targeted insertion/deletion/replacement**
  - Generate DNA construct
  - Construct consists of genomic DNA of ~3-5,000 bp on either side of targeted sequence; mutated sequence and/or selection genes for isolating targeted cells
  - Embryonic stem cells
  - Constructs introduced into cells
    - Homologous recombination (gene targeting) is a rare event, so need to be able to eliminate cells that did not have the new gene correctly targeted.
    - Uses positive selection: insert carries drug resistance factor, cells without new transgene are killed by antibiotics—eliminates cells with no insert
• Also uses negative selection: ends of transgene construct carry toxic gene; if insert is randomly integrated, the entire sequence will be present, so the toxic gene kills the cells
• After selection, only cells with targeted insertion remain alive.
• After verifying insert by PCR, inject ES cells into mouse blastocyst
• Recover chimeric mice carrying normal and targeted cells
• Breed to recover germline transmission and establish mice to study

5. **Applications of molecular biology:** Limited only by your imagination... MORE on this in the *Eye Development and Pathophysiology Advanced Module* in Spring 2016
   • Development of gene replacement therapies
   • Identifying targets for drug development
   • Generating animal models of disease
   • Understanding gene/protein function
   • Determining changes associated with disease