Intro to Molecular Biology
(all you need to know … in about 90 slides)

CS234

Chapters 1-5, 9, Lewin
Chapter 1&7, Krane & Raymer

December 29, 2015

Roadmap

• Proteins
• DNA
• RNA
• Central dogma
• Transcription
• Translation
• Genome and epigenome
Proteins

- A *protein* is a chain of molecules, called *amino acids*
- Every amino acid has a central carbon atom, known as *alpha carbon* ($C_\alpha$), an *amino group* ($\text{NH}_2$), a *carboxyl group* (COOH) and a *side chain*
- The side chain is what distinguishes one amino acid from another

### Amino acids

- **Alanine** $\text{CH}_2\text{CH}(\text{NH}_2)\text{COOH}$
  - Amino group
  - Carboxyl group
  - Side chain

- **Glycine** $\text{NH}_2\text{CH}_2\text{COOH}$
  - Amino group
  - Carboxyl group

- **Cysteine** $\text{SH}\cdot\text{CH}_2\text{CH}(\text{NH}_2)\text{COOH}$
  - Amino group
  - Carboxyl group
  - Side chain

- **Tyrosine** $\text{C}_6\text{H}_5\text{OH}\cdot\text{CH}_2\text{(NH}_2)\text{COOH}$
  - Amino group
  - Carboxyl group
  - Side chain
Proteins

• Amino acids are linked by **peptide bonds** between the **carboxyl group** and the **amino group**

• A typical protein is composed by 300 amino acids, but there are proteins with as few as 100 or with as many as 5,000 amino acids

Formation of a peptide bond

![Diagram showing the formation of a peptide bond](image)
Amino acids

• We (usually) find 20 amino acids
  – Alanine (A), Cysteine (C), Aspartic Acid (D), Glutamic Acid (E), Phenylanine (F), Glycine (G), Histidine (H), Isoleucine (I), Lysine (K), Leucine (L), Methionine (M), Asparagine (N), Proline (P), Glutamine (Q), Arginine (R), Serine (S), Threonine (T), Valine (V), Tryptophan (W), Tyrosine (Y)

Amino acid classification

One of many classifications that are possible, but is probably that which most people would agree covers the most protein contexts. C-SS and C-SH denote the two oxidation states of Cysteines. C-SS denotes those cysteines that are involved in disulphide bonds (i.e. connected to each other).
Proteins

- **Primary** structure: the linear sequence of amino acids, ordered from the N-terminal (amino group) to C-terminal (carboxy group)

- **Secondary** structure: α-helices and β-sheets

- **Tertiary** structure: the 3D conformation (folding) in space
• Most of the backbone is rigid
• The chemistry of a protein forces most of the backbone to remain planar
• The chemical bonds to the alpha carbons can rotate
• The angle of rotation for each alpha carbon bonds are called $\phi$ and $\psi$
• Phi and psi are the degree of freedom of the protein

Ramachandran plot

• Not all (phi, psi) combinations are possible
• The Ramachandran plot shows the value of phi and psi that are physically realizable
• Glycine can achieve additional configuration
Alpha helix

- Exactly 3.6 residues per turn
- Hydrogen bonds
- Two types
  - Right-handed
  - Left-handed

Beta sheet

- Regions of extended (nearly linear) backbone conformation with $\phi \approx 135$ and $\psi \approx 135$
- Hydrogen bonds
- Two types
  - Parallel
  - Anti-parallel
Some common tertiary structures

- Four helical bundle
- Alpha-beta barrel
- Open twisted beta sheet
- Open twisted beta sheet

**Hemoglobin** is a total of four proteins which assemble around a **Fe** atom that binds and hold on the oxygen in the blood.
Protein structure

- The *function* of a protein is determined by its tertiary structure
- Structure is much more conserved than sequence
- Predicting the folding from the primary sequence is very hard (see CASP competition)
- *Binding*: the interaction between two or more proteins (or protein-DNA) which have a “compatible” 3D structure (*docking*)
(A) The folding of the polypeptide chain typically creates a crevice or cavity on the protein surface. This crevice contains a set of amino acid side chains disposed in such a way that they can make bonds only with certain ligands. (B) Close-up view of an actual binding site showing the hydrogen bonds and ionic interactions formed between a protein and its ligand (in this example, cyclic AMP is the bound ligand).
DNA

- DNA is a *double stranded* chain of sugar molecules and phosphate residues
- Each sugar molecule contains five carbon atoms (labeled 1’ through 5’)
- Backbone bonds are between the 3’ carbon and the 5’ carbon
- Orientation of DNA is by convention 5’ to 3’
DNA

- Attached to the 1’ we can have one of four possible bases: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T)
- A,G are purines
- C,T are pyrimidines
- *Nucleotide* = sugar + phosphate + base
- DNA can reach in the thousands of millions of base pairs
DNA

- Each base in one strand is paired to its *complement* on the other strand: \( A \leftrightarrow T \quad C \leftrightarrow G \) (Watson-Crick pairs)
- **Reverse-complementation** operation
  e.g. \( y = ATTGC\text{GGA}T \)
  \( \tilde{y} = AT\text{CCGCAAT} \)
- **Replication**: the process of DNA to make an exact copy of itself
DNA replication is carried out by DNA polymerase. DNA polymerase is fast: it can replicate the entire human genome (~3 Gb) in about one hour.

DNA polymerase synthesizes in the 5'→3' direction and reads in the 3'→5' direction.

DNA polymerase has an error detection mechanism: the actual error rate of DNA polymerase is less than 1 error in a billion nucleotides.
DNA

- *Denaturation:* the process of strand separation (usually obtained by raising the temperature)
- *Renaturation:* the process of two separated complementary strands to reform in a double helix
- *Hybridization:* the process of two separated strands to reform in a duplex structure (the extent of hybridization is determined by their complementary)

![Diagram of DNA denaturation, renaturation, and hybridization.]
RNA

- Single stranded
- Uracil (U) instead of thymine (T)
- Different types of RNA
  - mRNA (messenger RNA)
  - tRNA (transfer RNA)
  - rRNA (ribosomal RNA)
  
  ... and recently discovered ncRNA in the “RNAi world”: miRNA, siRNA, snoRNA, stRNA, snRNA

- RNA is much less stable than DNA
Central Dogma

- Replication
- Reverse transcription
- Transcription
- Translation

Figure 1.34 The central dogma states that information in nucleic acid can be perpetuated or transferred, but the transfer of information into protein is irreversible.
Genes

- **Gene**: a segment of DNA which encodes for at least one polypeptide chain (usually mRNA)
- It includes regions preceding and following the coding region (UTR) and intervening sequences (*introns*)
- Genes usually lie in non-repetitive DNA

Transcription

- The synthesis of mRNA on a DNA template
- *RNA polymerase* is the enzyme that catalyzes this process (*pol II* in eukaryotes)
- RNA polymerase transcribes 1Kbps/sec
- **Start point** (+1) the first base pair transcribed (TSS)
- **Upstream** is the sequence prior to the start point (-1, -2, …)
- **Downstream** …
**Figure 1.28** RNA is synthesized by using one strand of DNA as a template for complementary base pairing.

DNA consists of two base-paired strands

5' ATGCGTTAGACCGTTAGCGGACCTGAC 
3' TACGGCAATCTGGCAATCGCCTGGACTG 

RNA synthesis

5' AUGCCGUUAGACCGUUGACCGGACCUGAC 
3' 

RNA has same sequence as DNA top strand; is complementary to DNA bottom strand

---

**Figure 1.29** The gene may be longer than the sequence coding for protein.

DNA

Leader

5' UTR

5' UTR

Protein

Protein defines coding region

Length of RNA defines region of gene

RNA

Trailer

3' UTR
Prokaryotes and Eukaryotes

According to the most recent evidence, there are three main branches to the tree of life

- Prokaryotes include Archaea ("ancient ones") and bacteria
- Eukaryotes include plants, animals, fungi and certain algae

**Prokaryotes**: organisms lacking nuclear membrane
- *E. coli, B. subtilis, H. influenzae, H. pylori, ...*

**Eukaryotes**: organisms whose DNA is inside the nucleus
- *A. thaliana, C. elegans, D. melanogaster, H. sapiens, M. musculus, S. cerevisiae (models)*

Eukaryotic genes may have **splicing sites**
# Prokaryotes and Eukaryotes

<table>
<thead>
<tr>
<th>Prokaryotes</th>
<th>Eukaryotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single cell</td>
<td>Single or multi cell</td>
</tr>
<tr>
<td>No nucleus</td>
<td>Nucleus</td>
</tr>
<tr>
<td>No organelles</td>
<td>Organelles</td>
</tr>
<tr>
<td>One piece of circular DNA</td>
<td>Chromosomes</td>
</tr>
<tr>
<td>No mRNA post transcriptional modification</td>
<td>Exons/Introns splicing</td>
</tr>
</tbody>
</table>
Alternative splicing & genetic diseases

Muscular dystrophy (MD) refers to a group of genetic disorders whose major symptom is muscle wasting. There are two major forms of MD, differing in severity and age of onset. In Duchenne muscular dystrophy, symptoms are noticeable in early childhood and quickly become debilitating. Becker muscular dystrophy, on the other hand, is of later onset and less severe. Both forms of MD are caused by mutations in the dystrophin gene, a large (2.6Mb) gene comprised of 97 exons.
Transcription

- **Promoter**: a region of DNA involved in binding of RNA polymerase to initiate transcription

- **Enhancer**: a region of DNA that increases the utilization of (some) promoters (it can function in either orientations and any location relative to the promoter)

- **Repressor**: a region of DNA that decreases the utilization of (some) promoters
Transcription control

- Different factors are involved in the transcription machinery
  - binding of transcription factors to DNA
  - ability of DNA to bend
  - relative location of the binding sites
  - interaction between transcription factors
  - DNA methylation, nucleosomes (epigenetics)
  - presence CpG islands ("p" is for phosphate)
  - ...

Example: Bacterial Promoter
Genetic “circuits”

Slide by Serafim Batzoglou, Stanford U.
Example: A Human heat shock protein

- TATA box: positioning transcription start
- TATA, CCAAT: constitutive transcription
- GRE: glucocorticoid response el.
- MRE: metal response element
- HSE: heat shock element

Reverse Transcription

- The process of transcription can be reversed: RNA viruses (retroviruses) can reverse-transcribe their RNA genomes into DNA, which is then integrated into the host genome and replicated along with it.
- *Reverse transcription*: the synthesis of double-stranded complementary DNA (cDNA) from single-stranded mRNA.
- For mRNA, the process starts from the polyA tail and moves backwards to its head.
Reverse Transcription

Translation

- The synthesis of a protein on the mRNA template
- Takes place inside **ribosomes**
- Ribosomes are made of rRNA
- Ribosomes translate about 60 bases/sec (<0.0001% error rate)
- mRNA is translated into the corresponding amino acids by ribosomes + tRNA
**Figure 1.30** Gene expression is a multistage process.

**Figure 5.8** A polyribosome consists of an mRNA being translated simultaneously by several ribosomes moving in the direction from 5' to 3'. Each ribosome has two tRNA molecules: one carrying the nascent protein, the second carrying the next amino acid to be added.
Figure 5.13 Transcription units can be visualized in bacteria. Photograph kindly provided by Oscar Miller.
Figure 7.2: The number of codons for each amino acid does not correlate closely with its frequency of use in proteins.

Figure 7.3: Third bases have the least influence on codon meanings. Boxes indicate groups of codons within which third-base degeneracy ensures that the meaning is the same.

<table>
<thead>
<tr>
<th>Third base relationship</th>
<th>Third bases with same meaning</th>
<th>Number of codons</th>
</tr>
</thead>
<tbody>
<tr>
<td>third base irrelevant</td>
<td>U, C, A, G</td>
<td>32</td>
</tr>
<tr>
<td>purines differ</td>
<td>U or C</td>
<td>14</td>
</tr>
<tr>
<td>from pyrimidines</td>
<td>A or G</td>
<td>10</td>
</tr>
<tr>
<td>unique</td>
<td>U, C, A</td>
<td>3</td>
</tr>
<tr>
<td>definitions</td>
<td>G only</td>
<td>2</td>
</tr>
</tbody>
</table>
The RNAi world

- dsRNA: double stranded RNA, typically longer than 30 nt
- miRNA: microRNA, 21-25 bases
  - Encoded by endogenous (‘within’) genes
  - Hairpin precursors
  - Recognize multiple targets
- siRNA: short-interfering RNA, 21-25 bases
  - Mostly exogenous origin
  - dsRNA precursors
  - May be target specific

RNAi: Two Phase Process

- Initiation
  - Generation of mature siRNA/miRNA

- Execution
  - Silencing of target gene
  - Degradation/inhibition of translation
A Bestiary of RNA

Once thought to be simply a messenger between the DNA and the protein-building processes of a cell, RNA now plays many roles. RNA can perform its own gene function. New types of drugs help to make use of some of the many types of RNA to halt or reduce the production of specific proteins.

RNA Transcribed

Human cells contain thousands of genes needed for protein synthesis, each performing a specific function. To build a protein, a section of DNA unwinds, exposing the nucleotide bases. The strand that contains the protein codons is a copy made by RNA polymerase. Once the protein strand is formed, it leaves the nucleus.

A Primer of RNA

RNA acts as a template for DNA replication. It is also involved in the translation of proteins and in the regulation of gene expression. RNA is transcribed from DNA in the nucleus, then exported to the cytoplasm. RNA degradation is essential for cell function, ensuring that RNA is not overproduced.

Chromosome

Chromosomal DNA is stored in 23 pairs of chromosomes within the nucleus of each cell.

An Acting RNA

RNA Polymerase

RNA polymerase is a large enzyme that synthesizes RNA from DNA templates. It replicates the genetic material in the nucleus and transcribes it into mRNA, which is then translated into proteins in the cytoplasm.

A Constraint RNA

Ribosomal RNA

Ribosomal RNA (rRNA) is a component of ribosomes, the molecular machines that translate RNA into proteins. RRNA is synthesized in the nucleolus of the cell nucleus and then exported to the cytoplasm, where it assembles into ribosomes.

A Cloaked RNA

Transfer RNA (tRNA)

Transfer RNA (tRNA) is a type of RNA that transfers amino acids to the ribosome during protein synthesis. It recognizes specific codons on the mRNA and delivers the corresponding amino acid to the ribosome.

A Circulating RNA

MicroRNA (miRNA)

MicroRNA (miRNA) is a type of small RNA that regulates gene expression by binding to complementary sequences in the target mRNA. miRNAs play a crucial role in various cellular processes, including development, differentiation, and disease.

A Confounding RNA

Long Non-coding RNA (lncRNA)

Long non-coding RNA (lncRNA) is a type of RNA that is not translated into protein. LncRNAs have various functions, including regulating gene expression, RNA processing, and epigenetic modifications.

A Converging RNA

Double-stranded RNA (dsRNA)

Double-stranded RNA is a form of RNA that is formed when RNA molecules are cleaved by RNAse or by viral replication. dsRNA can trigger the RNA interference (RNAi) pathway, leading to the degradation of complementary messenger RNA (mRNA).

A Collaborating RNA

Small Interfering RNA (siRNA)

Small interfering RNA (siRNA) is a type of RNA that is synthesized by enzymes called Dicer from double-stranded RNA. siRNA molecules guide RNAi enzymes to specific sites on the target mRNA, leading to the degradation of the mRNA and inhibition of gene expression.

A Critical RNA

miRNA Interference: Blocking translation

For more details, please refer to the original source.
Genome

- In higher organisms, DNA is contained in *chromosomes*
- Number of chromosomes in characteristic of the specie (e.g., *H. sapiens* has 46, *S. cerevisiae* has 32, *C. elegans* has 12, *D. melanogaster* has 8)
- Eukaryotic chromosomes usually appear in pairs
- Corresponding genes in *homologous* chromosomes may differ (*alleles*)

Haploid vs Diploid

- Chromosomes
  - *Sex chromosomes*
  - *Autosomes*
- Humans have 22 (pairs of) autosomes and 2 sex chromosomes (female XX, male XY)
- *Haploid set* of chromosomes contains one copy of each autosome and one sex chromosome
- *Diploid set* of chromosomes contains two copies of each autosome and two sex chromosomes
- *Polyploid set* …
Genome

Figure 3.10 Genome sizes, gene numbers and lethal loci.

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome (Mb)</th>
<th>Genes</th>
<th>Lethal loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mycoplasma genitalium</td>
<td>0.58</td>
<td>470</td>
<td></td>
</tr>
<tr>
<td>Rickettsia prowazekii</td>
<td>1.11</td>
<td>834</td>
<td></td>
</tr>
<tr>
<td>Haemophilus influenzae</td>
<td>1.83</td>
<td>1,743</td>
<td></td>
</tr>
<tr>
<td>Methanococcus jannaschii</td>
<td>1.66</td>
<td>1,738</td>
<td></td>
</tr>
<tr>
<td>B. subtilis</td>
<td>4.2</td>
<td>4,100</td>
<td></td>
</tr>
<tr>
<td>E. coli</td>
<td>4.6</td>
<td>4,283</td>
<td>1,800</td>
</tr>
<tr>
<td>S. cerevisae</td>
<td>13.5</td>
<td>6,034</td>
<td>3,600</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>165</td>
<td>12,000</td>
<td>3,100</td>
</tr>
<tr>
<td>C. elegans</td>
<td>97</td>
<td>19,099</td>
<td></td>
</tr>
<tr>
<td>H. sapiens</td>
<td>3,000</td>
<td>100,000</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.23 Most genes are uninterrupted in yeast, but most genes are interrupted in flies and mammals. (Uninterrupted genes have only 1 exon, and are totalled in the leftmost column.)

Figure 2.24 Yeast genes are small, but genes in flies and mammals have a disparate distribution extending to very large sizes.
Human Genome by numbers

- About 3.1 billion nucleotide bases
- Less than 1.5% of the genome codes for proteins (~26% of the genome is introns)
- ENCODE project: 80% of the entire human genome is either transcribed, binds to regulatory proteins, or is associated with some other biochemical activity
- Repeated sequences make up at least 50% of the human genome (transposons, tandem repeats, etc.)
- The number of protein-coding genes is estimated ~20,000 (pseudogenes ~13,000, ncRNA genes ~18,400)
Human Genome by numbers

- The average gene consists of 3,000 bases, but sizes vary greatly: the largest is *dystrophin* at 2.2Mbp; smallest encodes histone H1a at 781 bp; Titin has the longest coding sequence (80,780 bp), the largest number of exons (364), and the longest single exon (17,106 bp)

- The median size of an exon is 122 bp ($\mu = 145$ bp), the median number of exons is 7 ($\mu = 8.8$), and the median coding sequence encodes 367aa ($\mu = 447aa$)

Human Genome by numbers

- *Single nucleotide polymorphism* (SNP) occur on average between every 1 in 100 and 1 in 300 bps
- Large-scale structural variations range from a few thousand to a few million bps: these variations include differences in the number of copies individuals have of a particular gene, deletions, translocations and inversions (*copy number variations* or CNV)
- A high proportion of the genome (currently estimated at up to 12%) is subject to CNV
- SNPs and CNVs may either be inherited or caused by *de novo* mutation
Human Genome

- The human genome gene-dense regions are predominantly composed of G/C
- Genes appear to be clustered along the genome
- Stretches of up to 30,000 CpG bases repeating over and over often occur adjacent to gene-rich areas, forming a barrier between the genes and the “junk DNA”
- These CpG islands are believed to help regulate gene activity (DNA methylation)

Genome (chromosome) 3D structure

- Conformation of chromosomes in nuclei is important to many cellular processes such as gene regulation, DNA replication, maintenance of genome stability
Epigenetics: DNA methylation

- Cytosines (Cs) can be methylated, which turns them in a “fifth” type of nucleotide (other types of methylation have also been observed)
- In mammals 60%-90% of all CpGs are methylated
- DNA methylation is associated with gene silencing and plays a role in the development of nearly all types of cancer, and imprinting
- It can be inherited through cell division
- It is a dynamic process, still poorly understood
Epigenetics: Nucleosomes/Histones

- Eight core histones (2 x H2A, 2 x H2B, 2 x H3, and 2 x H4) and ~147 bps of DNA wrapped around form the 'nucleosome'
- Nucleosomes are thought to carry epigenetically inherited information in the form of covalent modifications of their core histones (histone code)
- Nucleosome “slide” (?) and their position regulate gene expression

Nucleosomes organization
Epigenetics: Nucleosome positioning

Further reading


Lewin, Genes XI, Jones & Bartlett, 2012