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

CS234

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

January 10, 2017

Roadmap

- Proteins
- DNA
- RNA
- Central Dogma
- Transcription
- Translation
- Genomics and Epigenetics
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 \((\text{COOH})\) and a side chain

• The side chain is what distinguishes one amino acid from another

Amino acids
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
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: $\alpha$-helices and $\beta$-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≈135 and psi≈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 holds 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 100s of millions of base pairs
DNA

• Each base in one strand is paired to its complement on the other strand: A↔T  C↔G (Watson-Crick pairs)

• Reverse-complementation operation
  e.g. y = ATTGCGGAT
       ŷ = ATCCGCAAT

• 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 (~3Gbps) 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)
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

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' ATGCCGTTAGACCGTTAGCGGACCTGAC 3'  
3' TACGGCAATCTGGCAATCGGCTGAGCTG  

RNA synthesis

5' AUGCCGUJAGACCGUJAGCCGACCUGAC 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

RNA

Trailer

3' UTR

5' N

3' C

Protein defines coding region

Length of RNA defines region of gene
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>

*Figure 2.10* Interrupted genes are expressed via a precursor RNA. Introns are removed when the exons are spliced together. The mRNA has only the sequences of the exons.
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”

- If A and B then D
- If B then NOT D
- If C then D
- If D then B

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
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 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 RNA 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 7.1 All the triplet codons have meaning: 61 represent amino acids, and 3 cause termination (STOP).

**GENETIC CODE**

<table>
<thead>
<tr>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU</td>
<td>UCU</td>
<td>UUA</td>
<td>UUG</td>
</tr>
<tr>
<td>UUC</td>
<td>UCC</td>
<td>UCA</td>
<td>UCG</td>
</tr>
<tr>
<td>UAU</td>
<td>UAC</td>
<td>UAA</td>
<td>UAG</td>
</tr>
<tr>
<td>UGU</td>
<td>UGC</td>
<td>UGA</td>
<td>UGG</td>
</tr>
<tr>
<td>CUC</td>
<td>CCA</td>
<td>CUG</td>
<td></td>
</tr>
<tr>
<td>CCA</td>
<td>CGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAG</td>
<td>CAU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CGU</td>
<td>CAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CGG</td>
<td>CGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CGC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CGA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GUG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GUC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GUA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GUG</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 7.2 The number of codons for each amino acid does not correlate closely with its frequency of use in proteins.
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
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><em>Mycoplasma genitalium</em></td>
<td>0.58</td>
<td>470</td>
<td></td>
</tr>
<tr>
<td><em>Rickettsia prowazekii</em></td>
<td>1.11</td>
<td>834</td>
<td></td>
</tr>
<tr>
<td><em>Haemophilus influenzae</em></td>
<td>1.83</td>
<td>1,743</td>
<td></td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>13.5</td>
<td>6,034</td>
<td>3,600</td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td>165</td>
<td>12,000</td>
<td>3,100</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>97</td>
<td>10,069</td>
<td></td>
</tr>
<tr>
<td><em>H. sapiens</em></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 dispersed distribution extending to very large sizes.

**Figure 3.1** DNA content of the haploid genome is related to the morphological complexity of lower eukaryotes, but varies extensively among the higher eukaryotes. The range of DNA values within a phylum is indicated by the shaded areas.
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

- *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
Detection of structural variations

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 ~147bps 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