Lecture 1

1. To understand the flow of biochemical information from DNA to RNA to proteins and ultimately to biochemical and cellular function and dysfunction (disease)

Central Dogma:

DNA→RNA→Protein→Phenotype

Nucleotide→gene→DNA→chromosome→genome

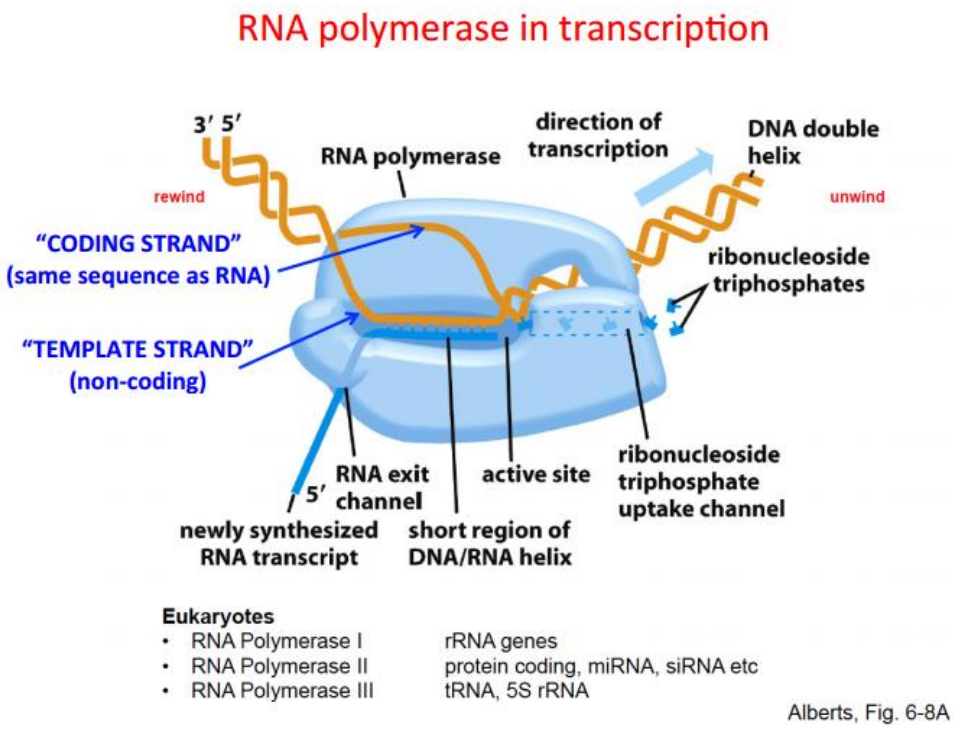
- In reality it is much more complex

DNA sequence → RNA sequence → Protein sequence → Protein structure → Protein function

→ RNA structure →RNA function

Know:

- Structure of DNA/RNA and proteins
- Intermolecular interactions:
  - Hydrogen bonds
  - Disulfur bonds
  - Ion bonds
  - Polar bonds
- Main aspects of the central dogma
  - Translation (at the ribosome)
  - Transcription
  - Genetic code



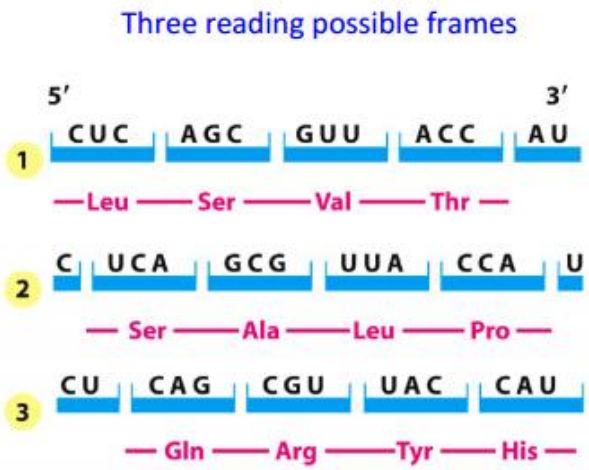
DNA structure:

- Sugar-phosphate backbone
- 4 bases
  - A/G = purines
  - T/C = pyrimidines
  - A-T and G-C
- 5′-3′ antiparallel (always read 5′→3′)
  - Addition occurs at the 3′ end
  - 5′ position to commence reading depends on promotor
  - Reading frame depends on promotor
- Coding and template strand:
  - Depends on the position of the promotor

RNA structure:

- U replaces T
- Self-complementarity = annealing of strand to itself
- tRNA/mRNA/pre-RNA
- Spliced (exons remain) to form mature RNA

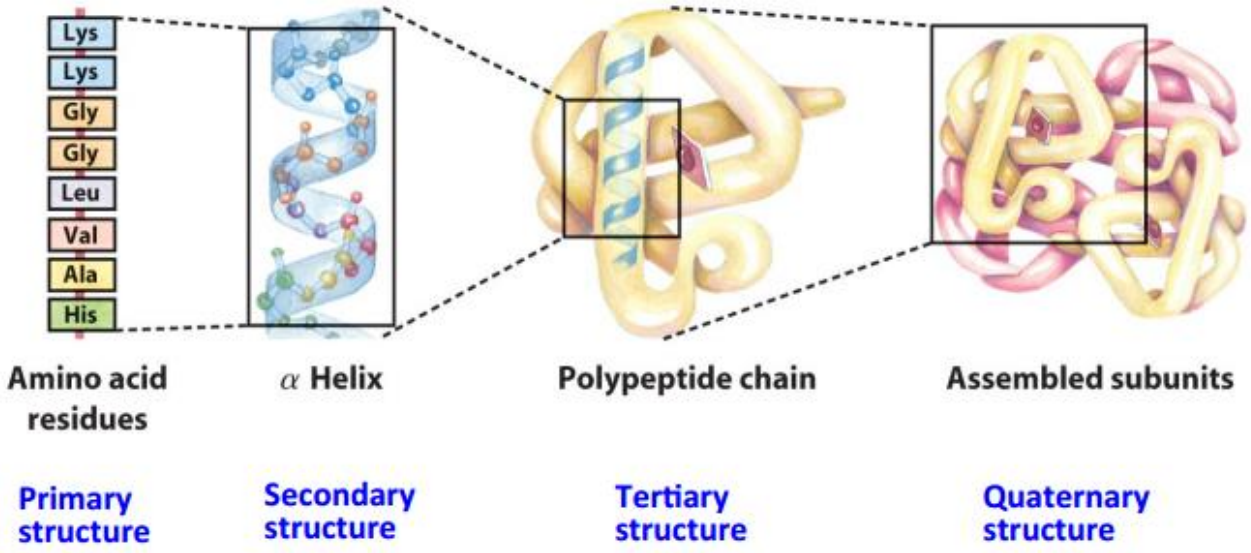
How many possible amino acid sequences could be encoded by this dsDNA?



Therefore 6:

- 3 from the codon from the top 5’ end, 3 from the codon from the bottom 5’ end.

Protein Structure:

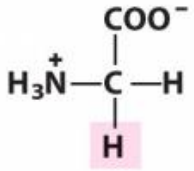


<u>Chemical Properties</u>

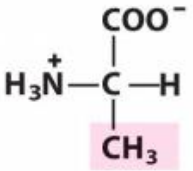
Depends on N or C-terminus, peptide bonds and side chains

- Non-polar aliphatic
- Polar but uncharged
- Aromatic
- Positively charged
- Negatively charged

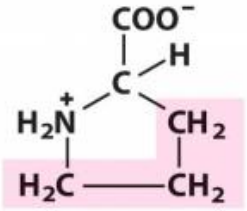
Nonpolar, aliphatic R groups



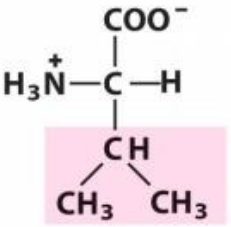
Glycine



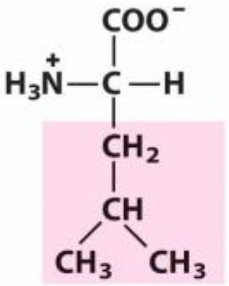
Alanine



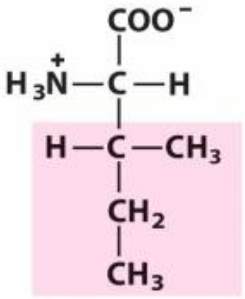
Proline



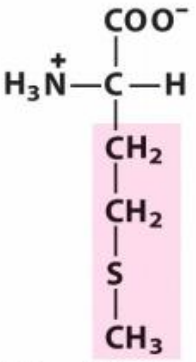
Valine



Leucine

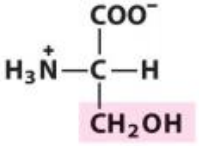


Isoleucine

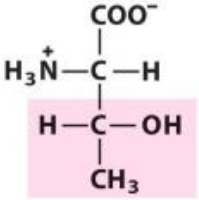


Methionine

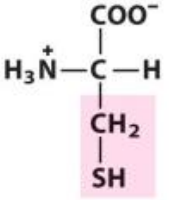
Polar, uncharged R groups



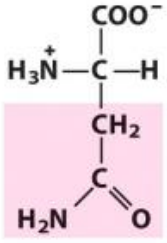
Serine



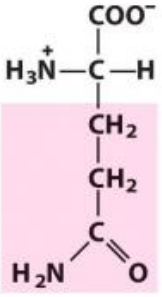
Threonine



Cysteine

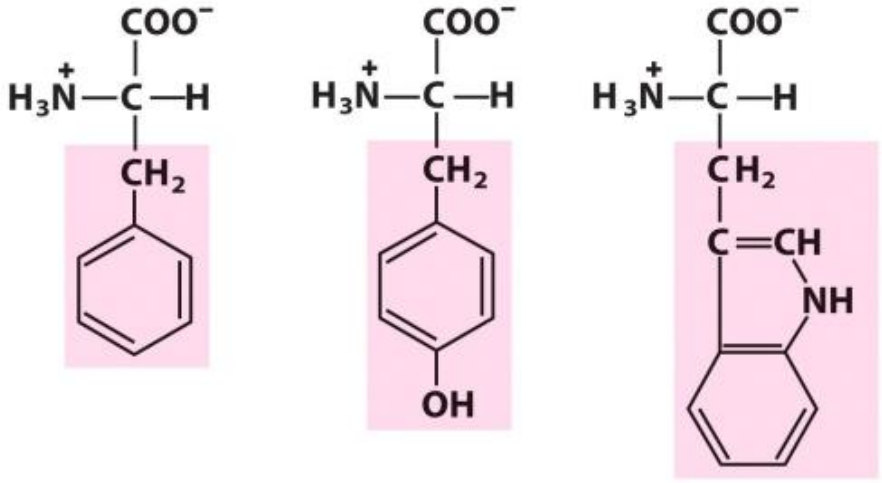


Asparagine



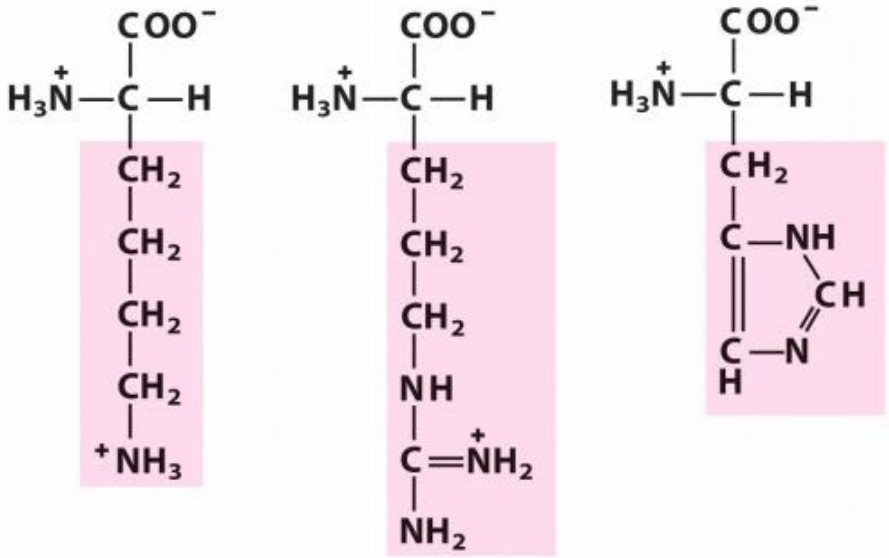
Glutamine

Aromatic R groups



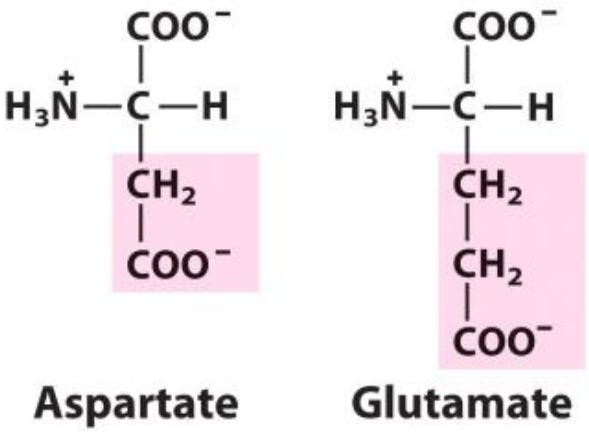
Phenylalanine Tyrosine Tryptophan

Positively charged R groups



Lysine Arginine Histidine

Negatively charged R groups



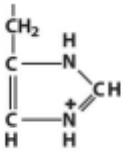
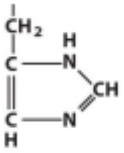
Aspartate **Glutamate**

- Nonpolar, aliphatic
- Aromatic
- Polar, uncharged
- Positively charged
- Negatively charged

•Which type are you likely to find in the hydrophobic core of a protein molecule?

•Which type are likely to be on the surface of the protein molecule, exposed to water?

Amino Acid Side Chain Ionisation

Amino Acid(s)	pK_a	Low pH Form*	High pH Form
C-terminus	~2	-COOH	$-COO^-$
Asp, Glu	~4	-COOH	$-COO^-$
His	~6		
Cys (reduced form)	~8	-SH	$-S^-$
Lys, N-terminus	~10.5	$-NH_3^+$	$-NH_2$
Arg	~12.5	$-NH-C(=NH_2^+)-NH_2$	$-NH-C(=NH)-NH_2$

with respect to pK

pK_a = pH at which the protein has a charge of zero.

Alpha-helices: side chains point sideways

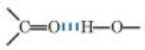
Beta-helices:

- Parallel and anti-parallel to produce alternate the direction the side chains point
- In reality, there is a combination of parallel and anti-parallel side chains
- Different bonds between NH and CO groups in each direction of side chain

Intermolecular Interactions

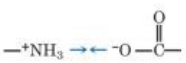
TABLE 2-5 Four Types of Noncovalent ("Weak") Interactions among Biomolecules in Aqueous Solvent

Hydrogen bonds

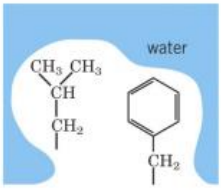
Between neutral groups 

Between peptide bonds 

Ionic interactions

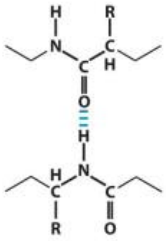
Attraction 

Repulsion 

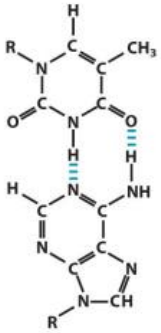
Hydrophobic interactions 

van der Waals interactions — Any two atoms in close proximity

Between peptide groups in polypeptides



Between complementary bases of DNA



* Can you identify the H-bond donor and acceptors??

Sample Question 1.

Q. Which of the following amino acid sequences is encoded by the second reading frame (beginning with the underlined nucleotide) in a coding strand of DNA with the sequence 5´-AAGTTGCGCTCGAC-3´?

A) Gln-Leu-Ala-Leu

B) Lys-Leu-Arg-Ser

C) Ser-Cys-Ala-Arg

D) Glu-Val-Arg-Ala

E) Val-Ala-Leu-Asp

Template strand = mRNA

Coding strand =AA = ribosome

Sample Question 2

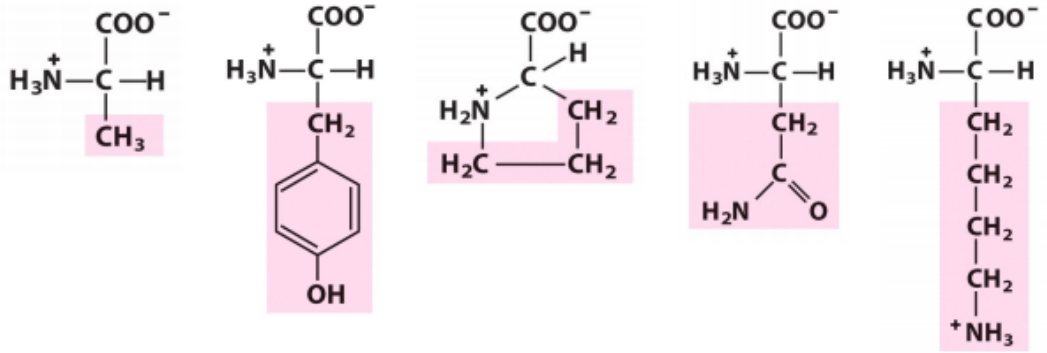
Example EMQ

Consider a peptide with the following amino acid sequence:

Ala-Val-Ala-Ser-Lys-Asn-Ala-Thr-Asp-Gln-Ser-Thr-Pro

Q. Which of the following amino acids is not present in this peptide?

(a) (b) (c) (d) (e)



Sample Question 3

Consider a peptide with the following amino acid sequence:

Ala-Val-Ala-Ser-Lys-Asn-Ala-Thr-Asp-Gln-Ser-Thr-Pro

NP NP Po + Po – Po NP

neutral

Q. How many positively charged and negatively charged side chain groups are present in this peptide at pH=7?

(a) None

(b) One positive and one negative

(c) One positive and two negative

(d) Two positive and one negative

(e) Two positive and two negative

Lecture 2

1. Using experimentally-determined rules we can use bioinformatics to analyze DNA sequences to identify genes, structural and regulatory elements, and to infer protein sequences
2. DNA sequence repeats and palindromes are often associated with structural and regulatory elements
3. Promoter sequences are an example of DNA motifs occurring in many genes, which can be identified through bioinformatics analysis. A consensus sequence is built using the most commonly found bases at each position in a motif.
4. Gene identification – look for ORFs, promoters, terminators
5. Protein sequence - can be deduced from gene sequence
6. Importance and function of a gene - can be inferred from mutation

Remembering DNA has two strands:

1. Translate the code

2. Identify potential binding sites

3. Compare wild-type, allelic and mutant sequences

Burden of Disease

- Australia suffers most from complex/chronic disease

Other types:

- Infectious
- Respiratory
- Congenital
- Neurological

Complex disease: multifactorial

- Nearly all conditions and diseases have a genetic component
- Single gene mutation rarely cause disease
- Commonly, disease is due to:
    - Multiple genes
    - Environmental factors
    - Lifestyle factors
- Often hereditary
- No clear-cut pattern
- Researchers continue to search for major contributing genes for most common complex disorders

Genomics: structure/function/evolution and mapping of the genome

- Driven by technology
- Results in data explosion
    - Increase and overwhelming amounts of information
    - Creates an info analysis gap
    - Cannot analyse the data fast enough

1. Gene analysis finds:
    a. Transcripts encoded
    b. Proteins encoded
    c. Evolution
    d. Building sites for machinery
2. Research
    a. Required to confirm or extend analysis
3. Experimentation

The lectures will explore:

- How to discover (eukaryotic) genes:
    - Recognise protein-encoding genes
    - Recognise and map gene variants
    - Recognise non-coding entities
- How to work out what gene products do
    - RNA
    - Protein
- How we can use this information in biomedical science and medicine:
    - Assess disease risk
    - Disease management

Gene: basic unit of linear arrays

Prokaryote: haploid (single chromosome/no nucleus)

Eukaryote: diploid (multiple chromosomes)

Unaltered gene: wildtype

Altered gene: mutation

DNA replication is semi-conservative

Read: 5'→3' (conventionally written 5'→3' left to right)

Addition: 3'→5'

Double helix due to polarity:

p-5'→3'-OH

N-5'→3'-C

Chargaff's rules:

1. Base composition of DNA varies from species to species, but is constant within a species
2. DNA isolated from different tissues of the same animal has the same composition
3. The base composition in a given species does not change with age, nutritional state or changing environment
4. In all cellular DNAs, the number of adenines equals the number of thymidines, and the number of guanosines equals the number of cytidine residues

DNA properties:

- Flexible

- Majority of helix is in beta-form
- Histones form chromatin
- Can be modified via methylation
    - Cytosine is commonly methylated
    - Prokaryotes self-methylate and identify/destroy foreign DNA by recognising different methylation patterns
- Carries information:
    - Replication
    - Repair
    - Recombination
    - Packaging
    - RNA and protein synthesis

Sequence annotation:

1. Identify gene via open reading frame and codons
2. Identify proteins associated with each codon
3. Identify structural elements (introns/exons)
4. Identify markers (cleavage points for restriction enzymes)
5. Identify alleles and mutations

Initiation codon: ATG

- Not every ATG is used as some proteins may need sequence motifs for promoters or a ribosome binding site

Stop codon: TAG/TAA/TGA

61 codons code for 20 amino acids producing redundant codes

Each protein has a preferential sequence (eg. Arg is encoded by 6 codons. AGA occurs 48% of the time, the rest occur 10% of the time)

- Preferential coding allows scientists to identify if the sequence they are experimenting on is a real gene

Mutation:

- Substitution via point mutation
- Frame shift via deletion of insertion
- Stop codon via substitution

Open Reading Frame:

- Not all ORFs are coding sequences
- All sequences are ORFs
- Only sequences starting with Met are coding sequences
- There is only one starting point though a coding sequence may have multiple Met proteins
- Computers can be used to identify ORFs

Experimental work is used to define motifs

- DNA/RNA binding factors can bind to more than one sequence
    - Related sequences are compared to find a consensus (best-fit) sequence
        - Eg. Palindromes
        - Repeats
            - CTCTCT
            - AAT AAT AAT
            - ATGTCAXXXXXATGTCA (directional repeat)
            - AAGCTT
              TTCGAA (inverse repeats has an axis of symmetry)

        - Repeats provide binding sites/indicate mobile genetic elements/contribute to single-strand secondary structure
    - Promoter/enhancer transcription initiator sites
    - Splicing
    - Ribosome binding
    - Termination
- Bioinformatics is used to determine the aforementioned things

Consensus sequences

- Prokaryote promoters:
    - RNA polymerase recognizes and binds to a DNA motif at the beginning of a gene (promoter sequence)
    - Terminator sequence tells the RNA polymerase to discontinue RNA synthesis

Consensus sequence shows the most commonly found bases at each promoter region indicated optimal sequence for RNA polymerase binding