COSC 348:
Computing for Bioinformatics

Lecture 2: Introduction contd.

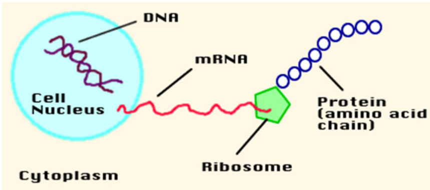
Lubica Benuskova

http://www.cs.otago.ac.nz/cosc348/

1

Central dogma of molecular biology

Crick 1958: "DNA makes RNA, which makes proteins, which make us".

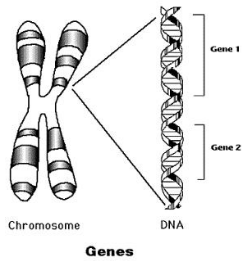


DNA remains in the nucleus, but in order to get its instructions translated into proteins, it must send its message to the ribosomes, where proteins are made. The molecule that carries this message is **Messenger RNA** (mRNA)

2

DNA to RNA transcription: a GENE

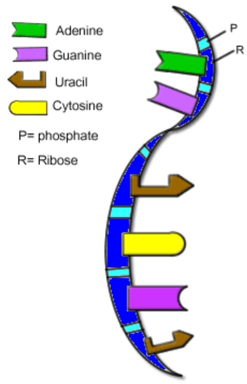
- It is not the whole DNA that is transcribed into mRNA, but only a small portion called a gene.
- Gene: Region of DNA that codes for a protein or for an RNA that has a function in the organism.
- A gene is associated with regulatory regions, transcribed regions, and other functional sequence regions.



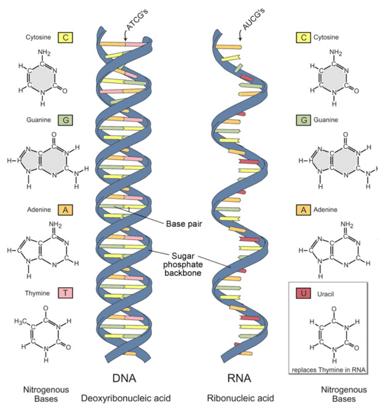
3

RNA - <u>**R**</u>ibo<u>**N**</u>ucleic <u>**A**</u>cid

- One strand instead of two and has ribose instead of deoxyribose
- 4 bases: Adenine (A), Uracil (U), Guanine (G), Cytosine (C)
- mRNA has the job of taking the message from the DNA from the nucleus to the ribosomes
- (There are also other types of RNA with different functions (e.g. regulatory functions) that are also created from DNA)



DNA versus RNA



Pyrimidine: 1 ring

Purine: 2 rings

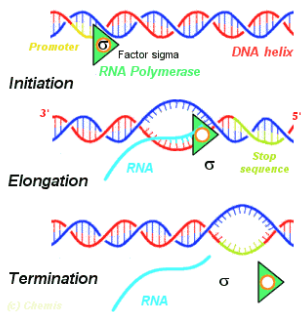
Pyrimidine: 1 ring

5

Image adapted from: National Human Genome Research Institute.

Transcription: RNA is made from DNA in 3 steps

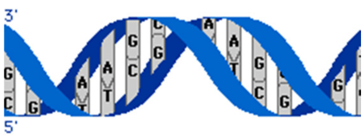
- **Initiation:** RNA polymerase + factor s bind to a special subsequence of DNA called ***promoter*** (for instance TATA, CAAT, etc.)
- **Elongation:** involves successive addition of RNA bases
- **Termination:** when Stop signals are encountered (GC-rich *palindrome* followed by *oligo*-A region). RNA and polymerase with s fall off.



Note: A **palindrome** is a sequence of units that has the property of reading the same in either direction (e.g., GCCG). **Oligo** means "a few" in Greek.

6

Transcription animation: RNA is made from DNA



Transcription is the process in which DNA is converted into RNA. Enzyme called RNA polymerase unzips DNA double helix, recruits RNA bases and matches them by base pairing, to the DNA sequence.

The pairing rule (DNA → RNA) is: G → C, T → A, A → U, and C → G

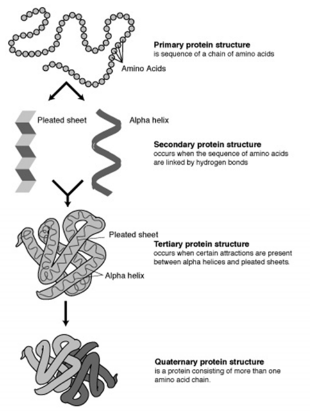
Protein synthesis

- Synthesis of proteins from genes is a 2-stage process.
  - The first stage, copying the instruction stored in DNA into RNA, is called ***transcription***.
  - The second stage, the actual protein building is called ***translation***.

- What are the rules of translation from mRNA to proteins?

Protein structure

- Proteins are the major functional and structural constituents of cells, we are made of proteins, and they make everything we need
- Proteins are formed by polypeptid chains of basic units called **amino acids**
- There are 20 amino acids, and the *order or primary structure* of a protein determines its structure and properties



Amino acid	3-letter code	1-letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cystein	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F

Amino acid	3-letter code	1-letter code
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V
No acid (gap)	---	-
Any acid	Xaa	X
Asn or Asp		B
Ile or L	Xle	J
Gln or Glu		Z
Pyrrolysine	Pyl	O
Selenocysteine	Sec	U

Protein sequence

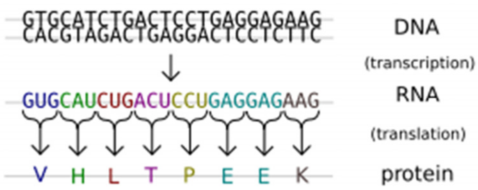
- An example of a protein sequence with 550 amino AA:

```
---------+---------+---------+---------+---------+
MKLLQRGVALALLTTFTLASETALAYEQDKTYKITVLHTNDHHGHFWRNE 50
YGEYGLAAQKTLVDGIRKEVAAEGGSVLLLSGGDINTGVPESDLQDAEPD 100
FRGMNLVGYDAMAIGNHEFDNPLTVLRQQEKWAKFPLLSANIYQKSTGER 150
LFKPWALFKRQDLKIAVIGLTTDDTAKIGNPEYFTDIEFRKPADEAKLVI 200
QELQQTEKPDIIIAATHMGHYDNGEHGSNAPGDVEMARALPAGSLAMIVG 250
GHSQDPVCMAAENKKQVDYVPGTPCKPDQQNGIWIVQAHEWGKYVGRADF 300
EFRNGEMKMVNYQLIPVNLKKKVTWEDGKSERVLYTPEIAENQQMISLLS 350
PFQNKGKAQLEVKIGETNGRLEGDRDKVRFVQTNMGRLILAAQMDRTGAD 400
FAVMSGGGIRDSIEAGDISYKNVLKVQPFGNVVVYADMTGKEVIDYLTAV 450
AQMKPDSGAYPQFANVSFVAKDGKLNDLKIKGEPVDPAKTYRMATLNFNA 500
TGGDGYPRLDNKPGYVNTGFIDAEVLKAYIQKSSPLDVSVYEPKGEVSWQ 550
```

- The size of a synthesized protein can be measured by the number of amino acids it contains and by its total molecular mass, which is normally reported in units of daltons, Da (synonymous with atomic mass units). Proteins can have from tens to thousands of amino acids.

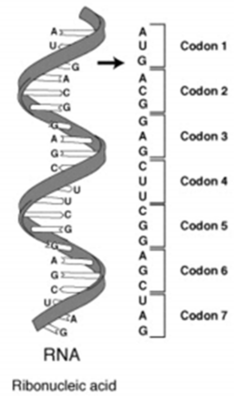
The Genetic Code

- The genetic code is a set of rules, by which information encoded in DNA and RNA sequences is translated into an amino acid (AA) sequence.
- The genetic code defines a mapping between nucleotide sequences and amino acid sequence.



The Genetic Code

- Every **triplet of bases, called a codon,** in a DNA / RNA sequence specifies a single amino acid.
- Different triplets of nucleotide bases code different amino acids.
- Each code word is a unique combination of three letters that codes a single amino acid in a polypeptide chain of a protein



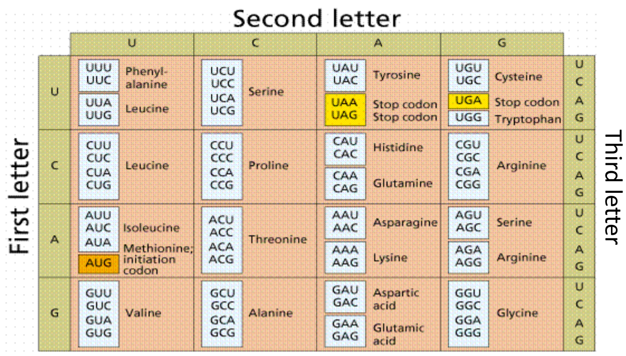
Directionality of DNA

- From which end of RNA we start to read the code?
- Directionality: end-to-end chemical orientation of a single strand of DNA/RNA due to its chemical properties.
- The chemical convention of numbering carbon atoms in the (deoxy)ribose gives rise to the so-called 5'-end and a 3'-end of DNA/RNA (pronounced as "five prime end" and "three prime end").
- DNA/RNA can only be assembled in a 5'- to 3'-direction. By convention, single strands of DNA and RNA sequences are written in 5'- to 3'-direction.

Coding and template DNA strands

- Relationship:
  - (5' → 3') **ATGGAATTCTCGCTC** (Coding, sense strand)
  - (3' → 5') **TACCTTAAGAGCGAG** (Template, antisense strand)
  - (5' → 3') **AUGGAAUUCUCGCUC** (mRNA made from template)
- One strand of DNA, with the same sequence as mRNA, is called the *coding* strand or sense strand.
- The complementary DNA strand, from which the mRNA is actually copied, is called *the template* strand or antisense strand.
- Since *mRNA is made from the template* strand, it has the same information as the coding strand.

Canonical or standard Genetic Code



First letter \ Second letter	U	C	A	G	Third letter
U	UUU, UUC Phenyl-alanine UUA, UUG Leucine	UCU, UCC, UCA, UCG Serine	UAU, UAC Tyrosine UAA Stop codon UAG Stop codon	UGU, UGC Cysteine UGA Stop codon UGG Tryptophan	U C A G
C	CUU, CUC, CUA, CUG Leucine	CCU, CCC, CCA, CCG Proline	CAU, CAC Histidine CAA, CAG Glutamine	CGU, CGC, CGA, CGG Arginine	U C A G
A	AUU, AUC, AUA Isoleucine AUG Methionine; initiation codon	ACU, ACC, ACA, ACG Threonine	AAU, AAC Asparagine AAA, AAG Lysine	AGU, AGC Serine AGA, AGG Arginine	U C A G
G	GUU, GUC, GUA, GUG Valine	GCU, GCC, GCA, GCG Alanine	GAU, GAC Aspartic acid GAA, GAG Glutamic acid	GGU, GGC, GGA, GGG Glycine	U C A G

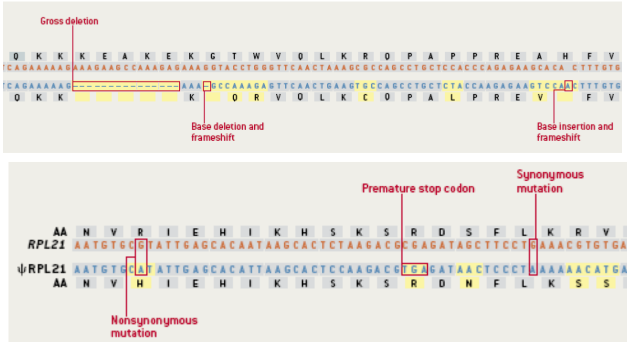
Properties of genetic code: universality

- Because the vast majority of proteins in all organisms on earth are encoded with exactly the same code, this particular code is called the canonical or standard genetic code.
- There are only slight variations to this code
  - *Mycoplasma* (one type of bacteria) translates the codon UGA as tryptophan;
  - in bacteria and archaea, GUG and UUG are common start codons;
  - in rare cases, certain specific proteins may use alternative start codons;
  - in certain proteins, non-standard amino acids (U or O) are substituted for standard stop codons), etc.
- To conclude: the **genetic code is almost universal**.

Properties of genetic code: redundancy

- **Redundancy** or degeneracy (> 1 codon per AA) but **no ambiguity**.
- We need to code for 20 amino acids and a stop codon, i.e. 21 unique codes is required. If there were 2 bases per 1 codon, then only 16 amino acids could be coded for (4^2=16). A three-letter code gives 4^3 = 64 possible codons, meaning 43 codons are redundant.
- But redundancy makes a huge advantage: it makes the genetic code more **fault-tolerant** for **point mutations** (exchange of one base for another by mistake), which can have a fatal consequence.

Types of DNA mutations



All these mutations (except synonymous) cause alteration(s) in the AA sequence.

Causes of mutations: mutagens

- Radiation: solar radiation, radiation from radioactive sources (Uranium), X-rays, radiation from universe, UV light.
- Viruses: viruses insert their genetic material into the host cell DNA and this may lead to errors. About 5000 viruses known.
- Chemical pollution:
  - e.g., tar – present in cigarettes.
  - e.g., benzene, an industrial solvent and precursor in the production of drugs, plastics, synthetic rubber and dyes.
- Transposons (“jumping genes”), mobile genetic elements that move from one site to another by “copy and paste” operation. Role – unknown (gene duplication, change of gene expression regulation, etc.)

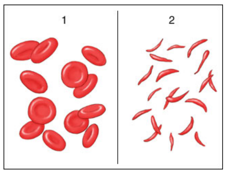
Harmful and good mutations

- Mutations are more frequent than we think and most mutations are fatal – for instance, estimated 30% of all embrya are spontaneously aborted during the first month of pregnancy.
- Many mutations cause cancer:
  - E.g., UV light causes mutation in the skin cells that turn malignant
  - Radiation from nuclear plant failure causes all kinds of cancer
  - Mutation caused by tar in cigarettes causes lung cancer, etc.
- Inherited gene mutations are the cause of genetic diseases.
  - Huntington disease, Alzheimer dementia, multiple sclerosis, etc.
- Mutations that result in an improved trait, drive evolution.
  - E.g., bigger claws, better eyesight, opposing thumb, bigger brains, etc.

Example of genetic disease: sickle-cell anaemia

- Sickle-cell anaemia is a genetic life-long blood disorder that is due to the mutation of a single nucleotide in the haemoglobin gene, from a GAG to GTG codon mutation.
- This results in replacing glutamic acid with valine in the haemoglobin protein and malformation of red blood cells under the condition of low oxygen.

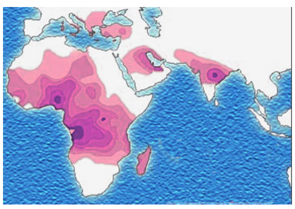
Normal red blood cells



Sickled red blood cells

Spread of sickle-cell anaemia

- Malformation of red blood cells deprives the tissues of oxygen which causes gradual organ damage.
  - an average life expectancy of 42 in males and 48 in females.
- People with SCA are resistant to malaria. In areas where malaria is common, there is a survival value in carrying a single sickle-cell gene. 30% of population in areas with malaria have the mutation.



Inheritance modes of genetic disease

- A person that receives the defective gene from both father and mother develops the disease. (Gene forms are called alleles.)
- A person that receives one defective and one healthy gene remains healthy, but can pass on the disease and is known as a carrier.
- Carriers produce a few sickled red blood cells, not enough to cause symptoms, but enough to give resistance to malaria.
- If two parents who are carriers have a child, there is a 1-in-4 chance of their child's developing the disease and a 1-in-2 chance of their child's being just a carrier.