

## 2. Genes and proteins

- The sequence as a model of DNA
- Genes: from Mendel to molecular biology
- **The genetic code**
- A translation algorithm
- Implementing the genetic code
- Algorithms + data structures = programs
- The algorithm design trade-off
- DNA sequencing
- Whole genome sequencing
- How to find genes?

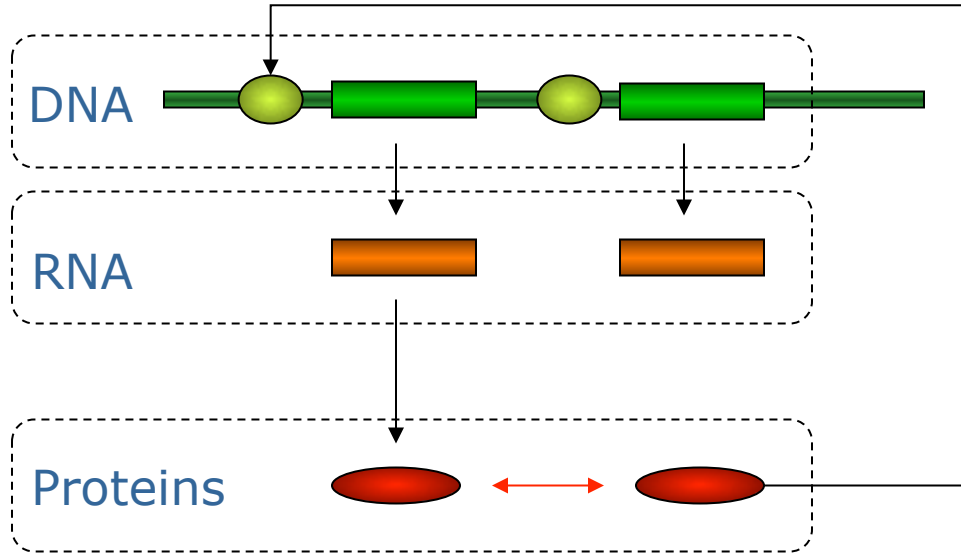
# The genetic code

# Proteins

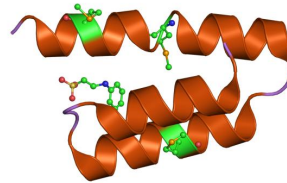
- Genes are DNA regions coding for proteins
- Proteins are made up of amino acids (AA)
- 20 different amino acids in protein sequences
- one-letter, 3-letter and full names

Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Transcription



Translation



# Translation

- Gene RNA sequences are translated into AA sequences
- What is the code?
  - i.e. the correspondance between  
DNA/RNA sequences (4-letter alphabet)  
and AA sequences (20-letter alphabet)
- 1 nucleotide / letter → only 4 possible AAs
- 2 nucleotides / letters → only 16 possible AAs  
AA AC AG AT CA CG CC CT GA GC GG GT TA TC TG TT
- 3 nucleotides / letters → 64 ( $4 \times 4 \times 4$ ) possible AAs

# The genetic code as a 3-entry table

1st base	2nd base								3rd base	
base	U		C		A		G		base	
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine	U	
	UUC		UCC		UAC		UGC		C	
	UUA				UCA	UAA	Stop (Ochre)	UGA	Stop (Opal)	A
	UUG				UCG	UAG	Stop (Amber)	UGG	(Trp/W) Tryptophan	G
C	CUU	(Leu/L) Leucine	CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine	U	
	CUC		CCC		CAC		CGC		C	
	CUA		CCA		CAA	(Gln/Q) Glutamine	CGA		A	
	CUG		CCG		CAG		CGG		G	
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine	U	
	AUC		ACC		AAC		AGC		C	
	AUA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A	
	AUG <sup>[A]</sup>	(Met/M) Methionine	ACG		AAG		AGG		G	
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine	U	
	GUC		GCC		GAC		GGC		C	
	GUA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A	
	GUG		GCG		GAG		GGG		G	

[A]: start codon

# The genetic code is redundant

- It is said to be “redundant”
  - Several different triplets code for the same amino acid
  - Example: CCU, CCC, CCA and CCG code all for proline / P
  - Coding triplets of nucleotides are called codons
- The triplet ATG (or AUG) codes for methionine  
but may also be the triplet where translation begins (start codon)
- UAA, UAG and UGA are stop triplets where translation ends

# The genetic code as an array

- The genetic code can be represented as an array of 64 rows and 2 columns
- Here, the first 12 rows (over 64)

TTT	F
TTC	F
TTA	L
TTG	L
TCT	S
TCC	S
TCA	S
TCG	S
TAT	Y
TAC	Y
TAA	Stop
TAG	Stop



# Pictures & movies : material licensing

p. 4 : Public domain, <http://www.ebi.ac.uk/>