

Data Science in  
Bioinformatics  
ws 22/23

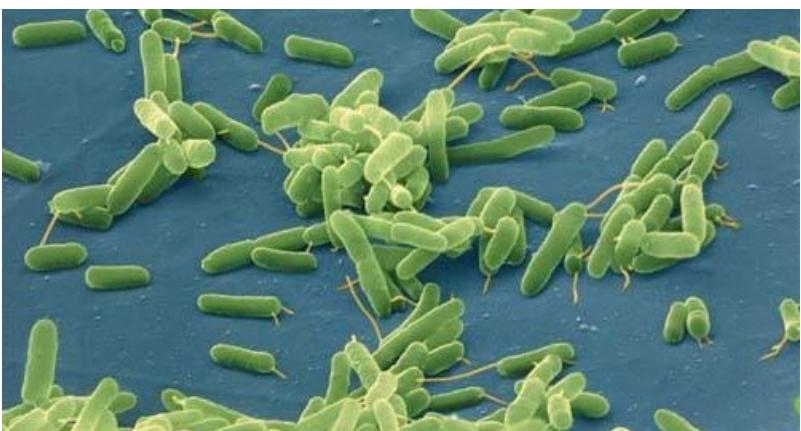
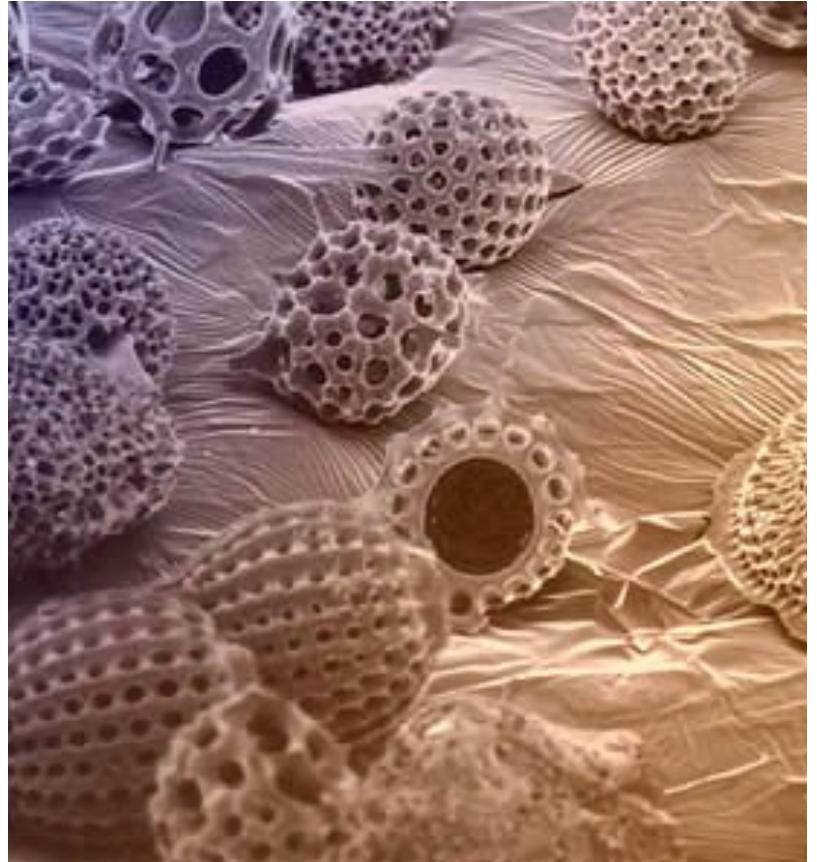
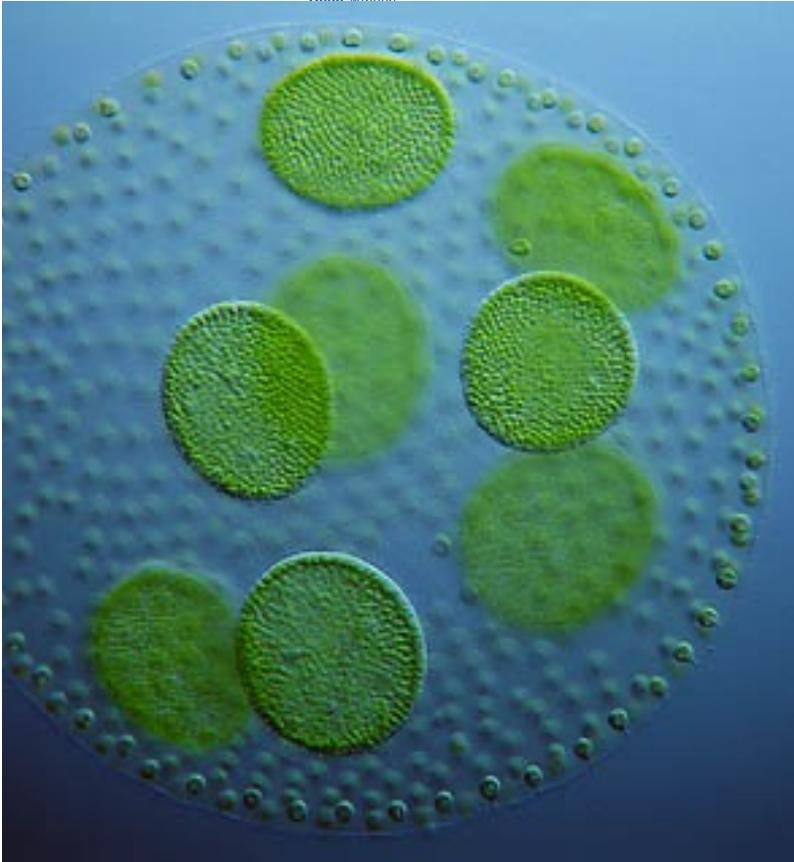
# From strands to strings

## Life, DNA and sequences: biological basics for bioinformaticians

Lecturer: Dr. Simon Magin  
Course Date: 17.10.2022

# Todays Topics

- **Life on earth: Diversity on common ground**
- **Chemistry of life: Biological Macromolecules (DNA, RNA, Proteins) and their structure**
- **Fundamental cellular processes: Replication, Transcription, Translation**
- **Bioinformatics: Fasta format and excercise**
- **Classification: Cellular Blueprints and Domains of Life**
- **Finally: What about Viruses?**





10/17/22



Basic Biology - Data Science in Bioinformatics WS 22/23



4

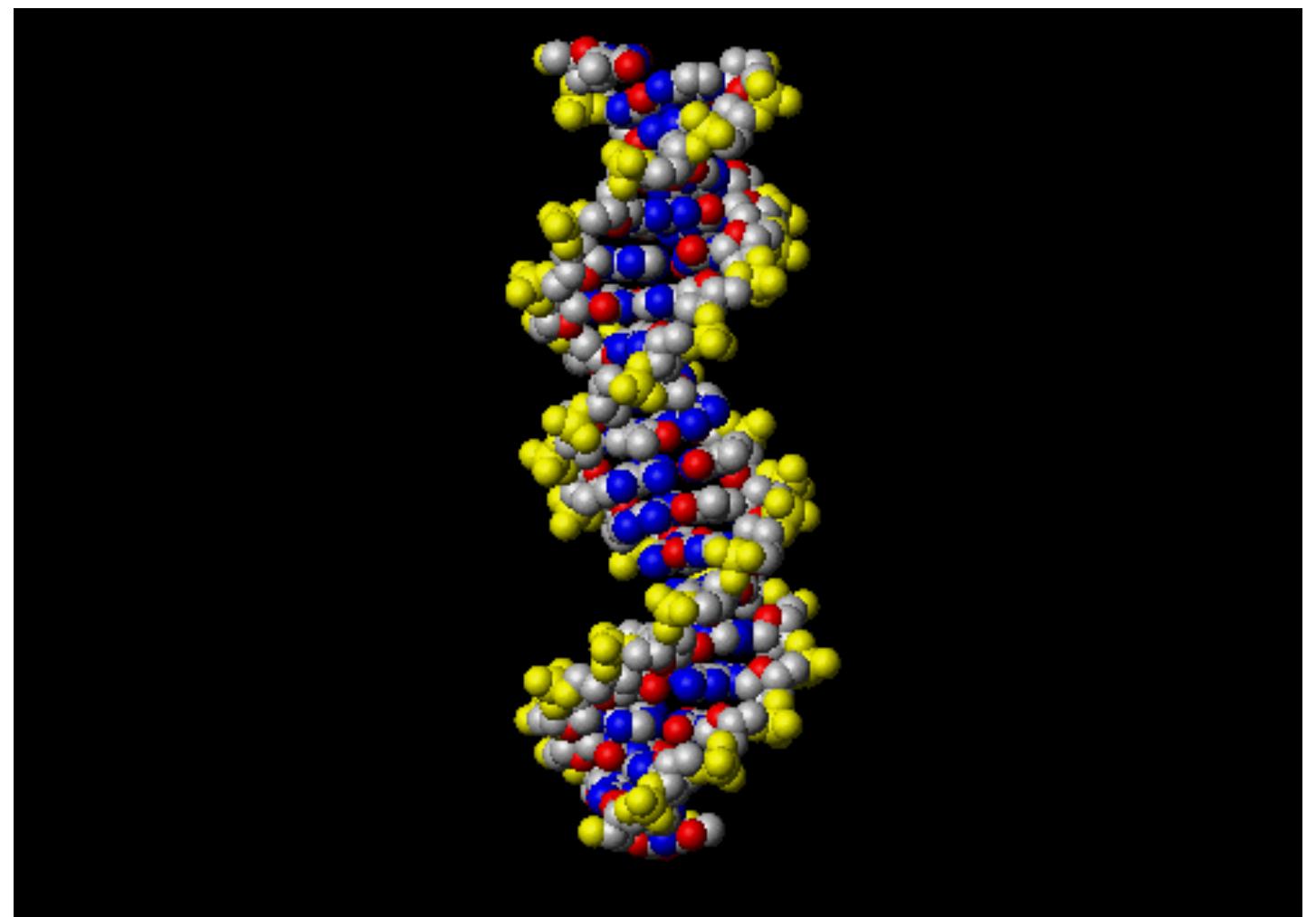


10/17/22

Basic Biology - Data Science in Bioinformatics WS 22/23

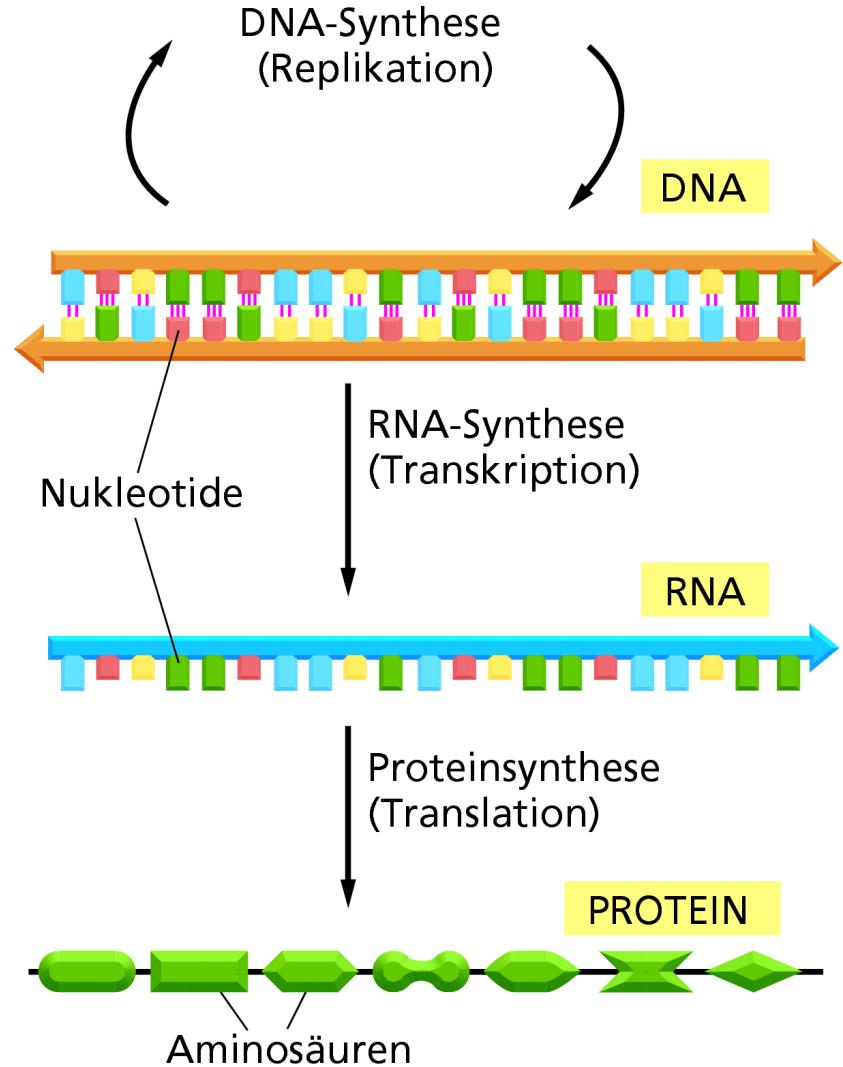


Double-stranded DNA  
(deoxyribonucleic acid)



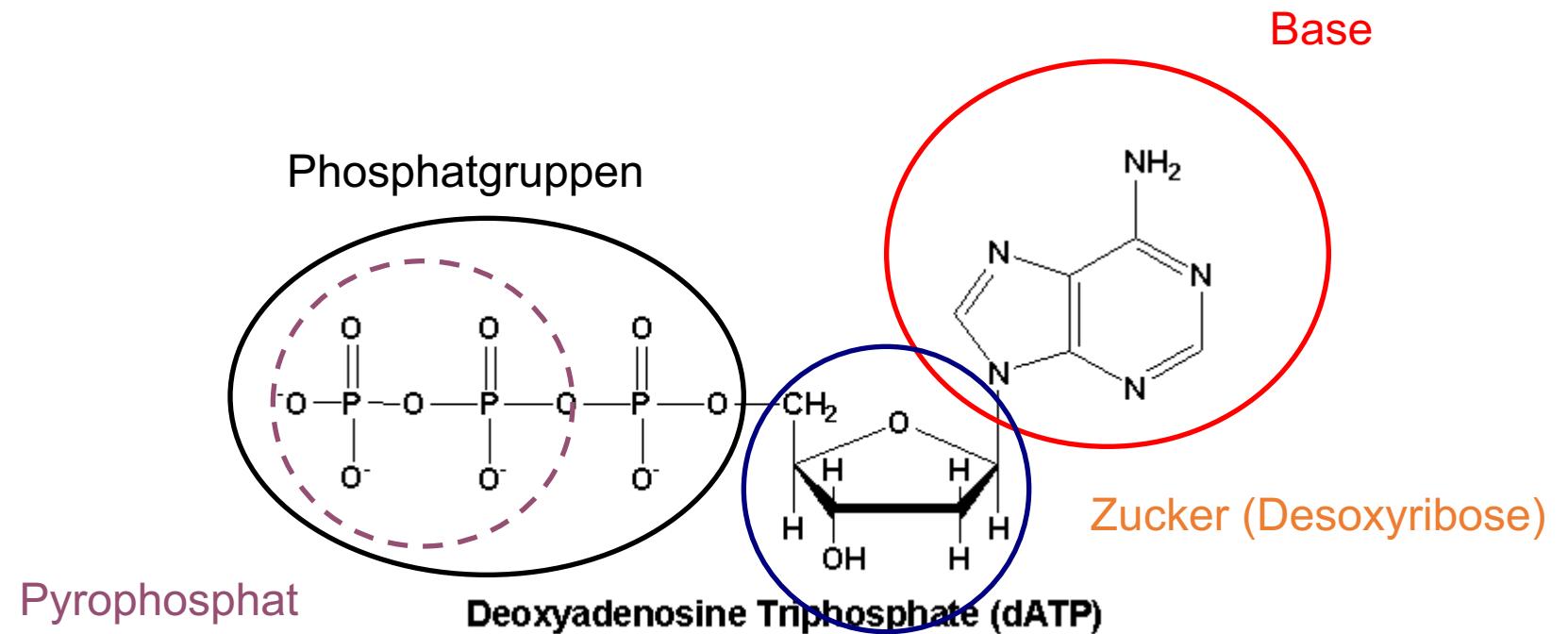
# Chemistry of Life

All the macromolecules which are central for the workings of a cell, be it for storage of information or to carry out work, are polymers (strings) of basic building blocks.



© 2012 Wiley-VCH, Weinheim  
Alberts - Lehrbuch der Molekularen Zellbiologie  
ISBN: 978-3-527-32824-6 Fig. 01-002

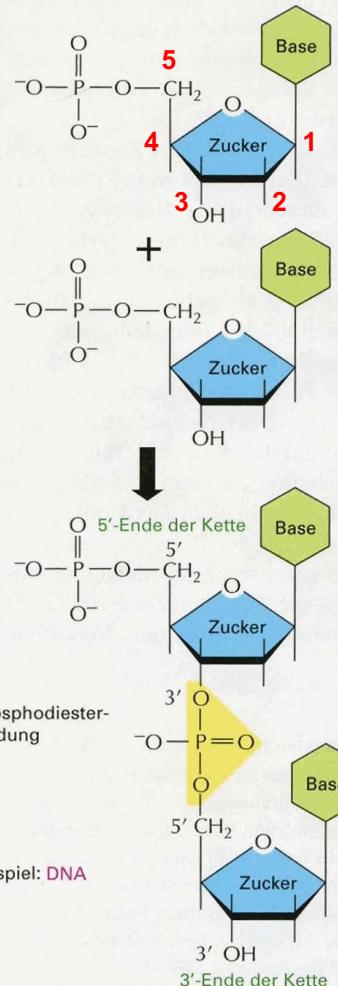
# deoxynucleotide triphosphates dNTPs



## NUCLEINSÄUREN

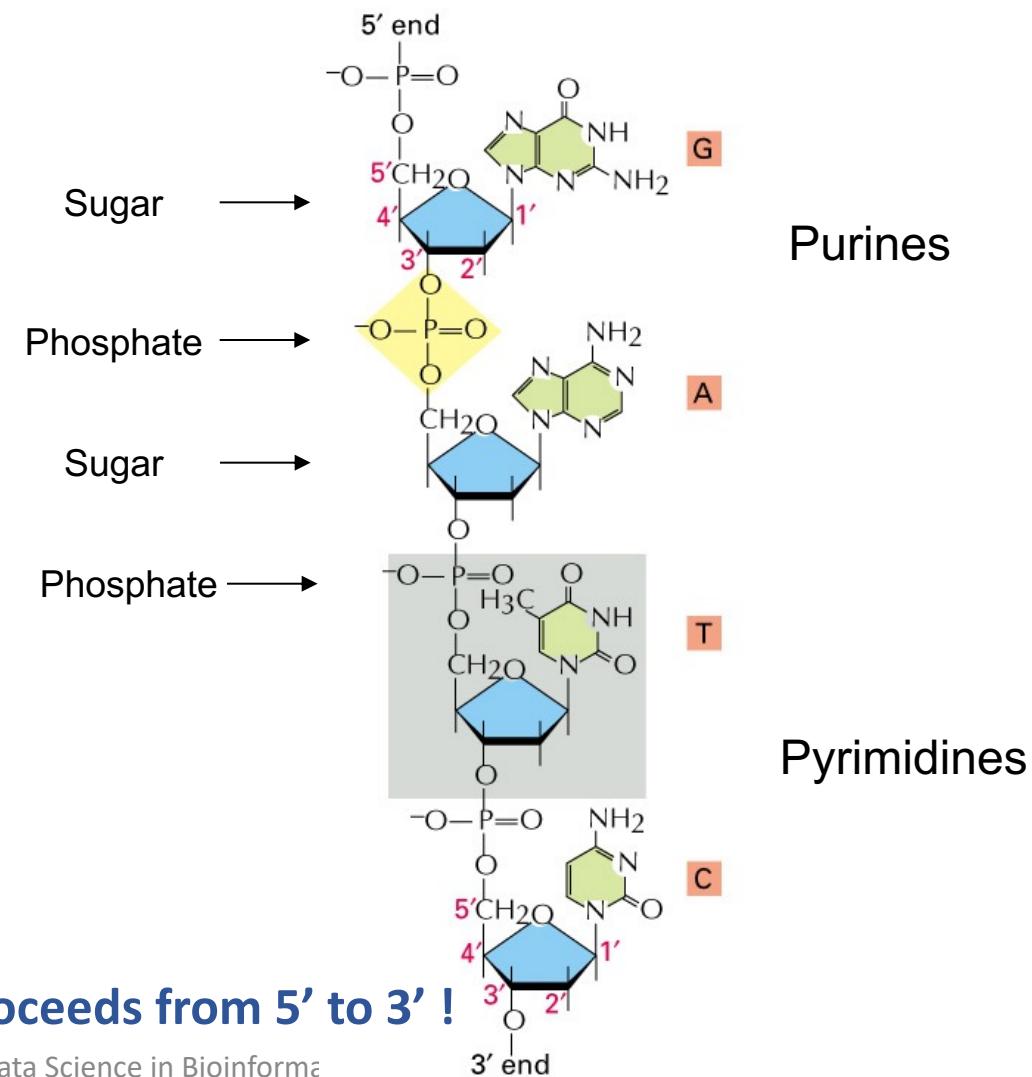
Nucleotide werden durch einen einzigen Typus einer Phosphodiesterbindung zwischen den 5'- und 3'-Kohlenstoffatomen zu Nucleinsäuren verknüpft.

Die lineare Sequenz der Nucleotide in einer Nucleinsäurekette wird im Allgemeinen durch einen Einbuchstaben-Code abgekürzt, A-G-C-T-T-A-C-A, wobei das 5'-Ende linkerhand steht.



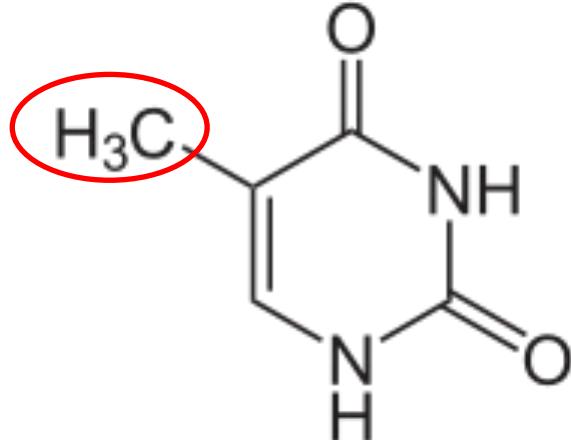
# DNA synthesis

Chaining nucleotides into a DNA strand through polymerization

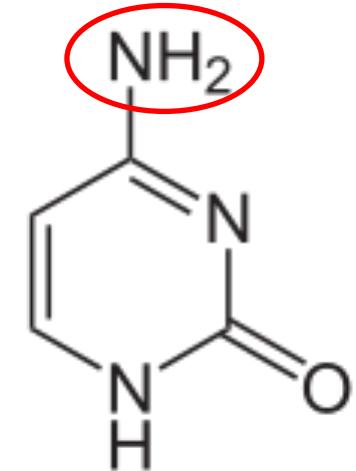


**DNA synthesis always proceeds from 5' to 3' !**

Pyrimidines

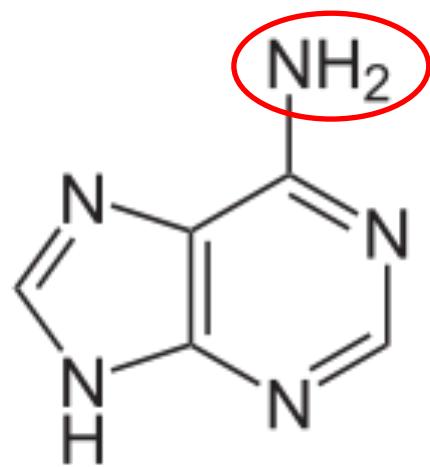


Thymine

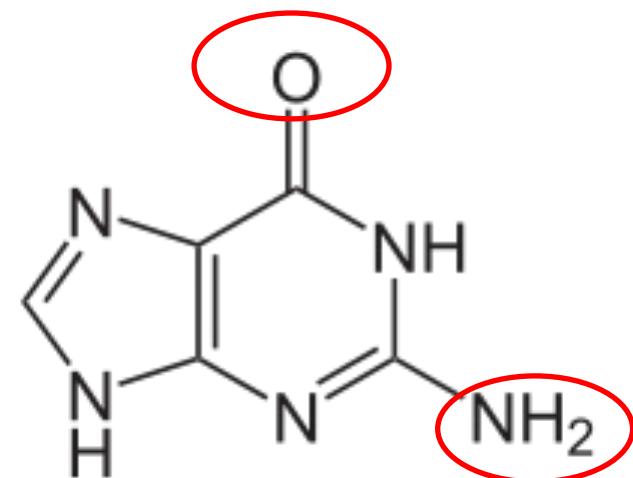


Cytosine

Purines

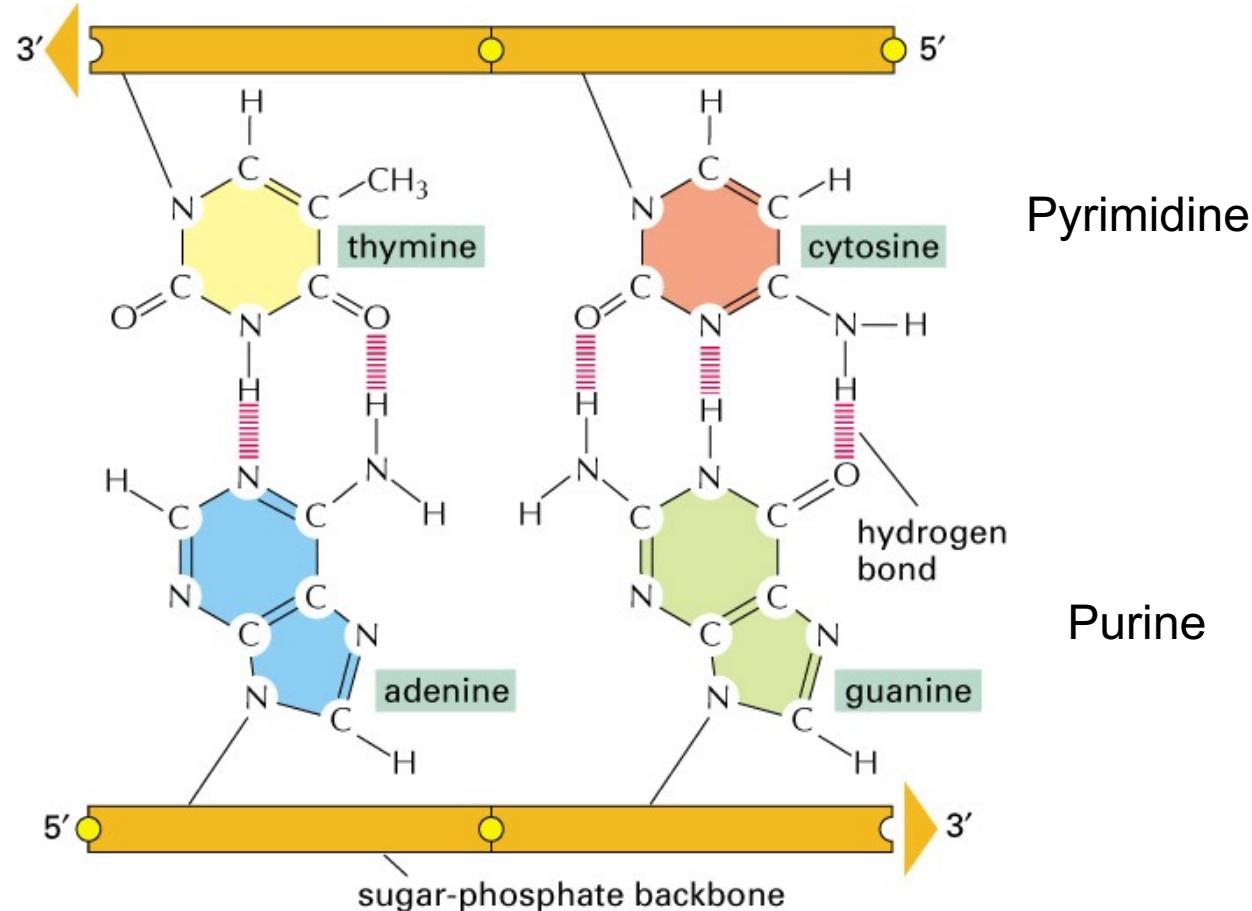


Adenine



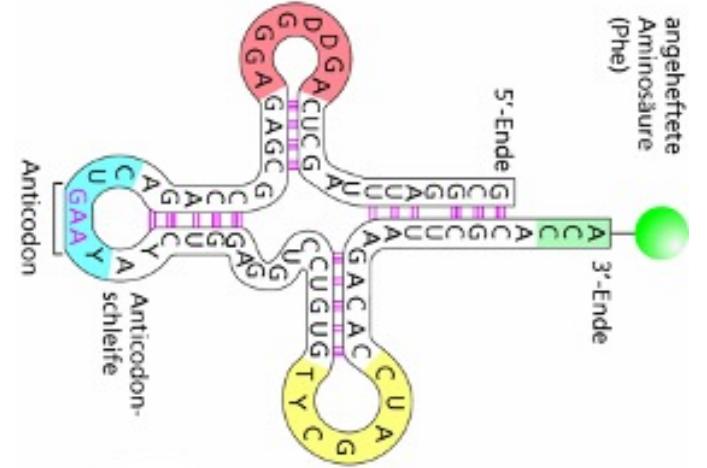
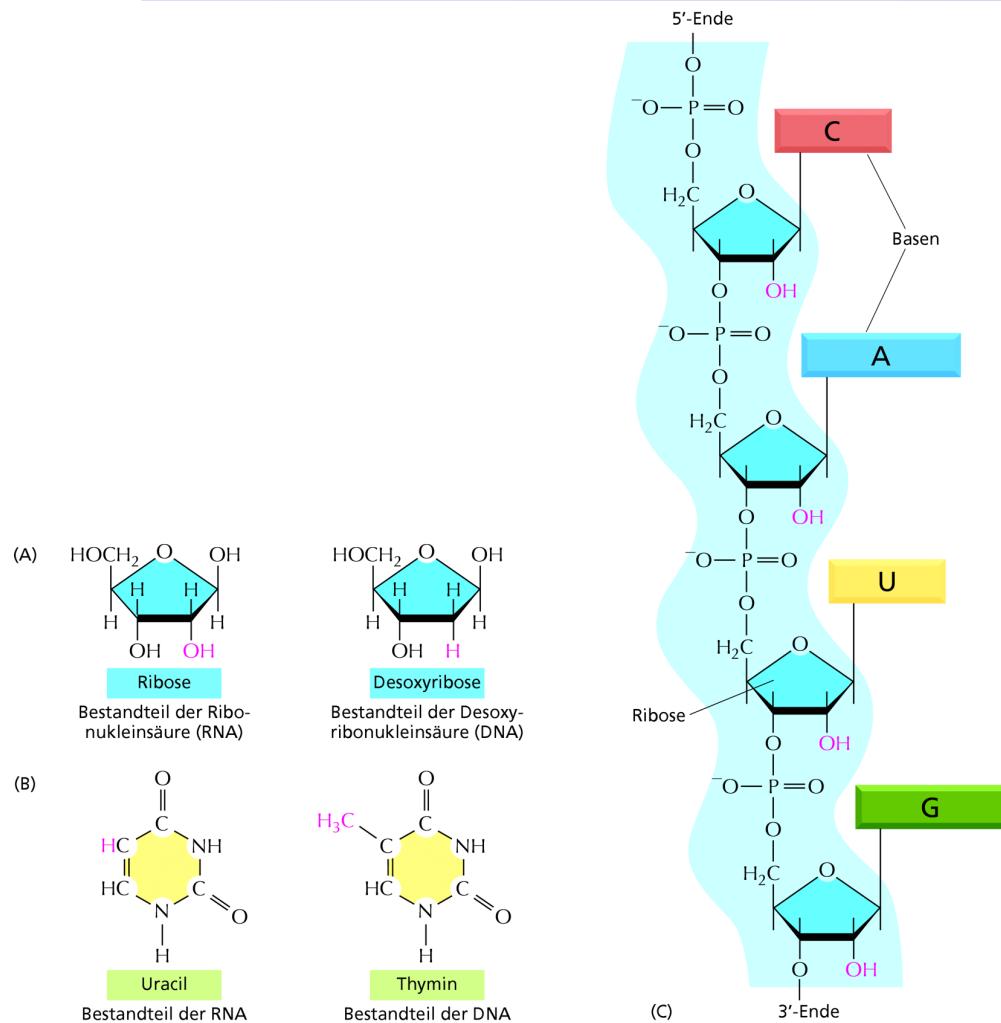
Guanine

# Complementary base pairing in the DNA double helix



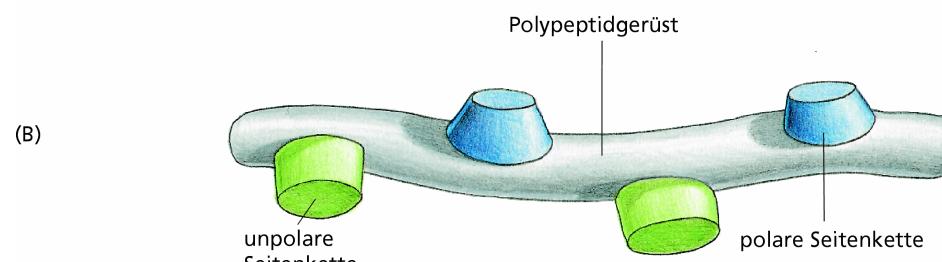
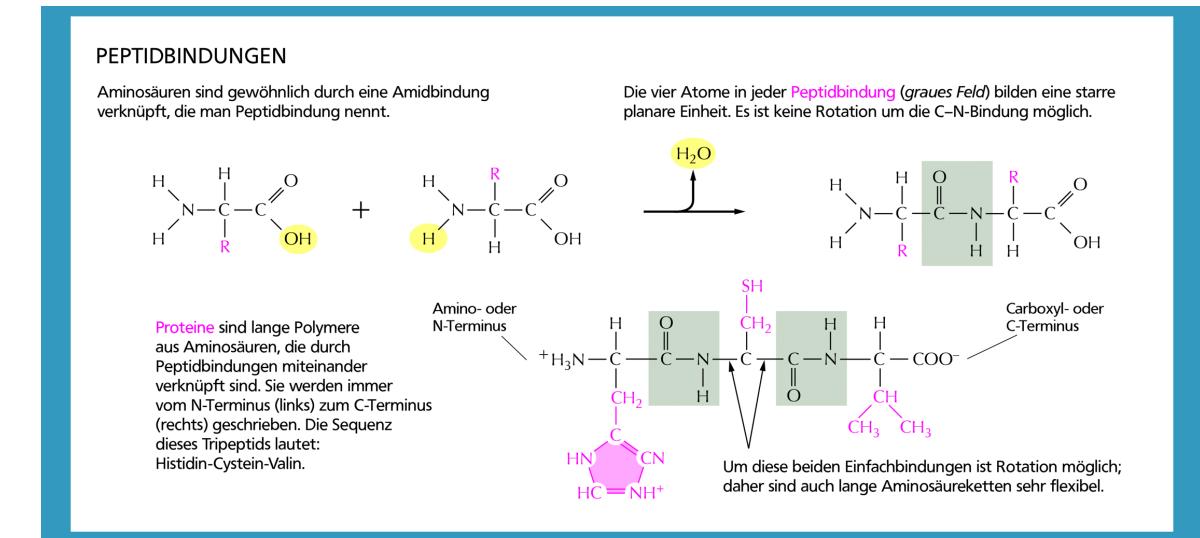
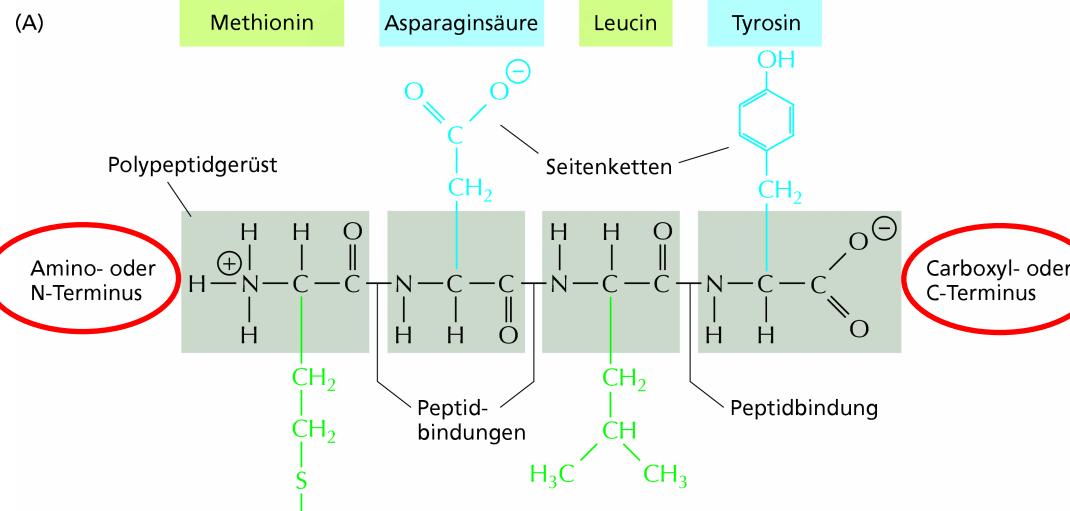
- Two strands of DNA form a **double helix** through **base-pairing**
- Stability increases with the number of **hydrogen bonds**
- The **5'-3' directionality of the strands is opposing**
- Genes can be located on each of the two complementary, antiparallel strands
- Genes can be overlapping and interleaved

# RNA



- Sugar is Ribose (as opposed to Deoxyribose)
- In RNA the base Uracil is used in place of Thymine
- Responsible for transport and translation of information
- Also structural and enzymatic functions (rRNA, ribozymes)
- Typically single stranded (except secondary structures)

# Proteins



# Proteins

There are 20 Amino acids of which proteins are composed

AMINOSÄURE		SEITENKETTE	
Asparaginsäure	Asp	D	negativ
Glutaminsäure	Glu	E	negativ
Arginin	Arg	R	positiv
Lysin	Lys	K	positiv
Histidin	His	H	positiv
Asparagin	Asn	N	ungeladen polar
Glutamin	Gln	Q	ungeladen polar
Serin	Ser	S	ungeladen polar
Threonin	Thr	T	ungeladen polar
Tyrosin	Tyr	Y	ungeladen polar

polare Aminosäuren (hydrophil)

© 2012 Wiley-VCH, Weinheim  
Alberts - Lehrbuch der Molekularen Zellbiologie  
ISBN: 978-3-527-32824-6 Fig. 04-003

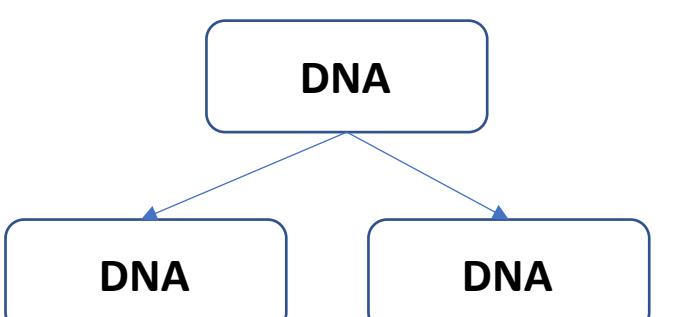
AMINOSÄURE		SEITENKETTE	
Alanin	Ala	A	unpolar
Glycin	Gly	G	unpolar
Valin	Val	V	unpolar
Leucin	Leu	L	unpolar
Isoleucin	Ile	I	unpolar
Prolin	Pro	P	unpolar
Phenylalanin	Phe	F	unpolar
Methionin	Met	M	unpolar
Tryptophan	Trp	W	unpolar
Cystein	Cys	C	unpolar

unpolare Aminosäuren (hydrophob)

# Three fundamental cellular processes to remember

## DNA replication

DNA Polymerase

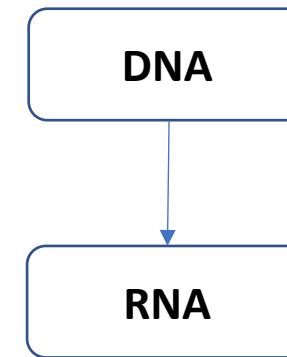


Prerequisite for cell division

Duplicates nucleic acid

## Transcription

RNA Polymerases

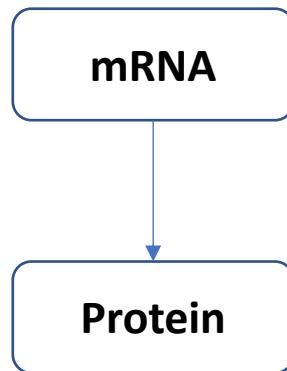


generation of mRNA, rRNA, tRNA .....

Transcribes one type of nucleic acid into another

## Translation

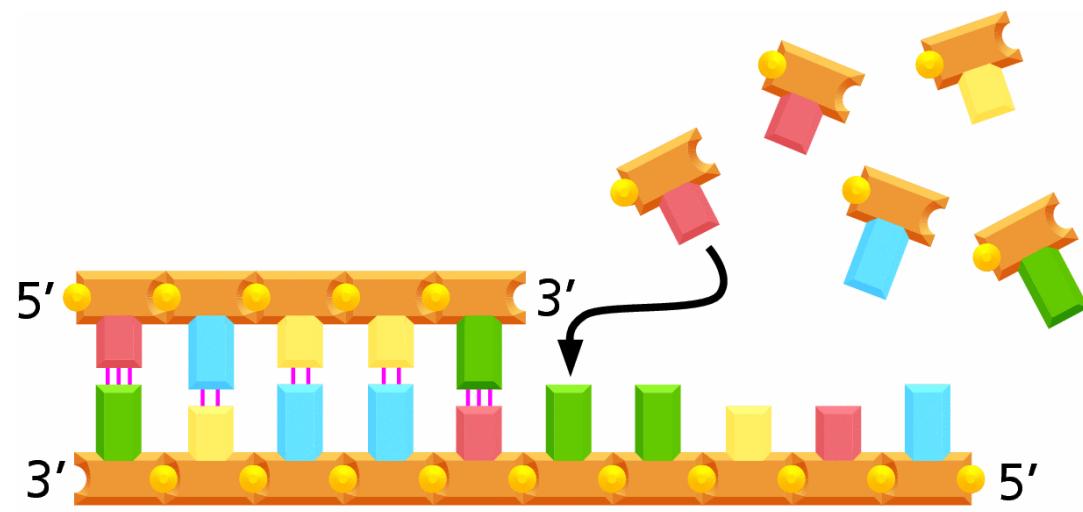
Ribosomes, tRNA



Proteinbiosynthesis

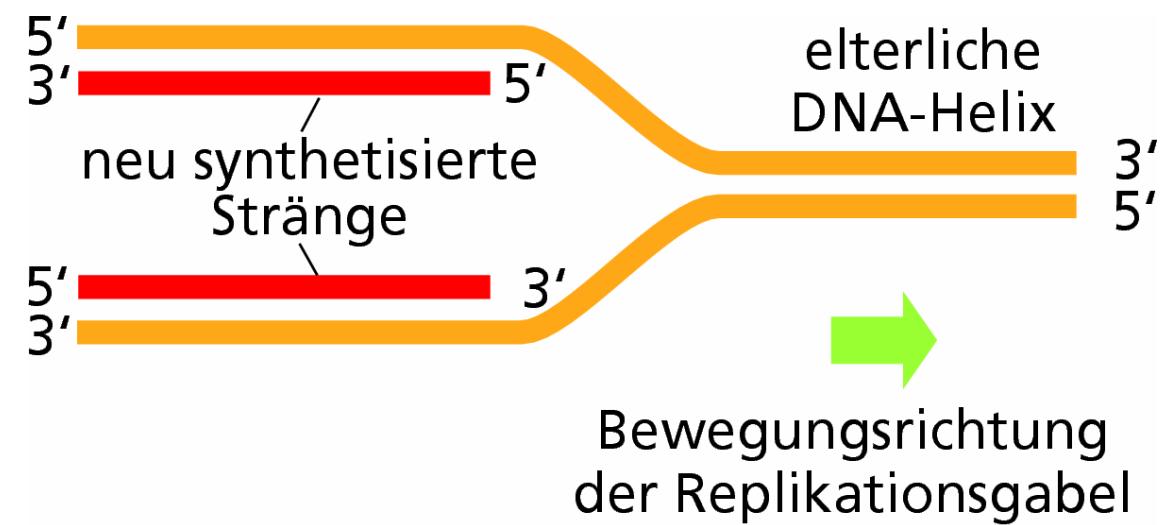
Translates nucleic acid to amino acid sequence

# DNA Replication



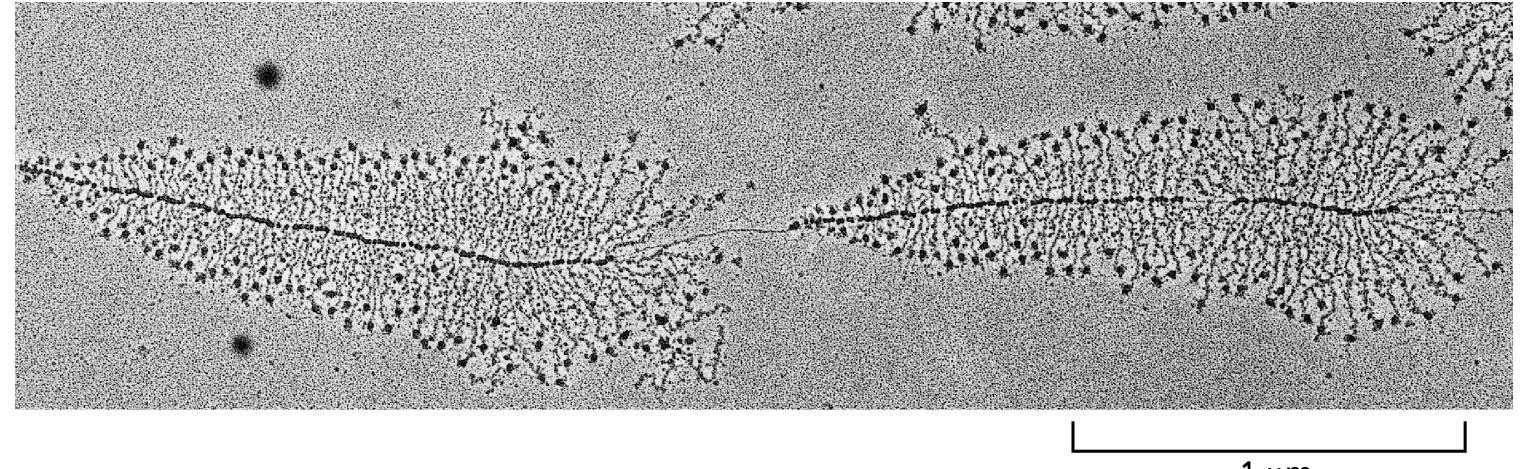
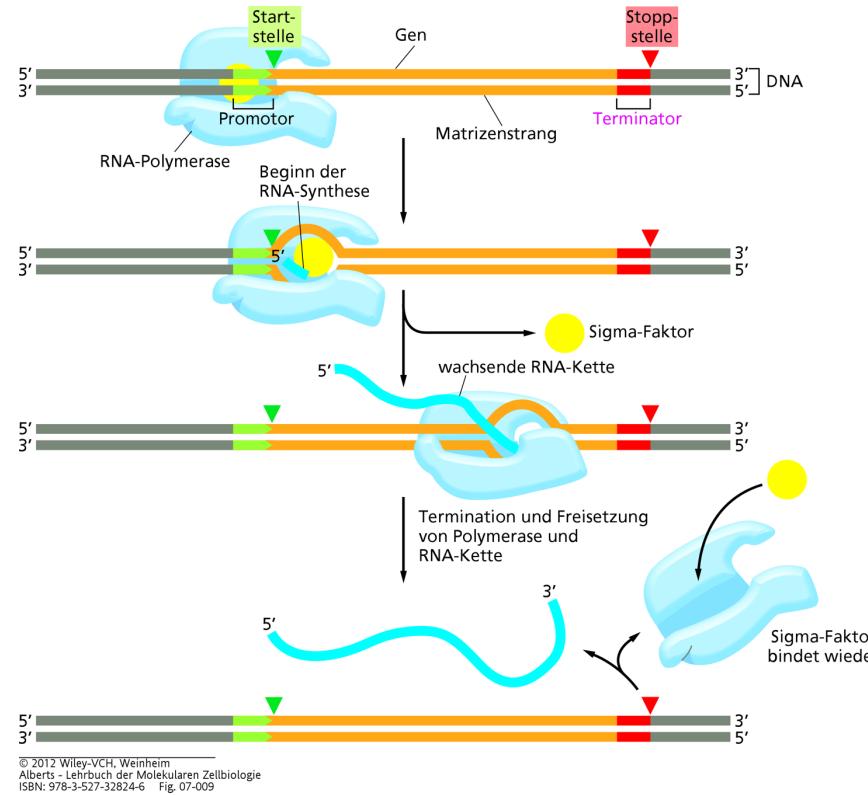
© 2012 Wiley-VCH, Weinheim  
Alberts - Lehrbuch der Molekularen Zellbiologie  
ISBN: 978-3-527-32824-6 Fig. 06-002

- Principle of complementarity underlies DNA replication
- replication mode is semi-conservative
- DNA de-novo synthesis has to go from 5' -> 3' direction on the leading and lagging strand



© 2012 Wiley-VCH, Weinheim  
Alberts - Lehrbuch der Molekularen Zellbiologie  
ISBN: 978-3-527-32824-6 Fig. 06-011

# Transcription

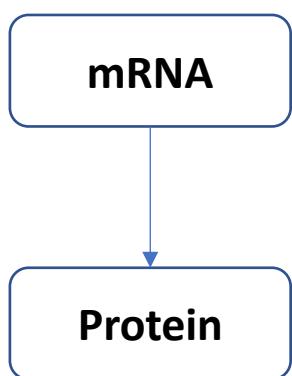


© 2012 Wiley-VCH, Weinheim  
Alberts - Lehrbuch der Molekularen Zellbiologie  
ISBN: 978-3-527-32824-6 Fig. 07-008

- Transcription is initiated at regulatory elements in the DNA called promoters
- Many RNA copies can be produced simultaneously
- The DNA Template remains unchanged
- Amount of RNA copies in the cell depends on rate of transcription (gene regulation, strength of Promoter) and half-live of RNA

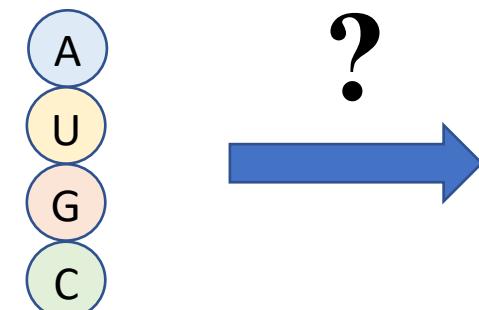
# Translation

**Translation**  
Ribosomes, tRNA



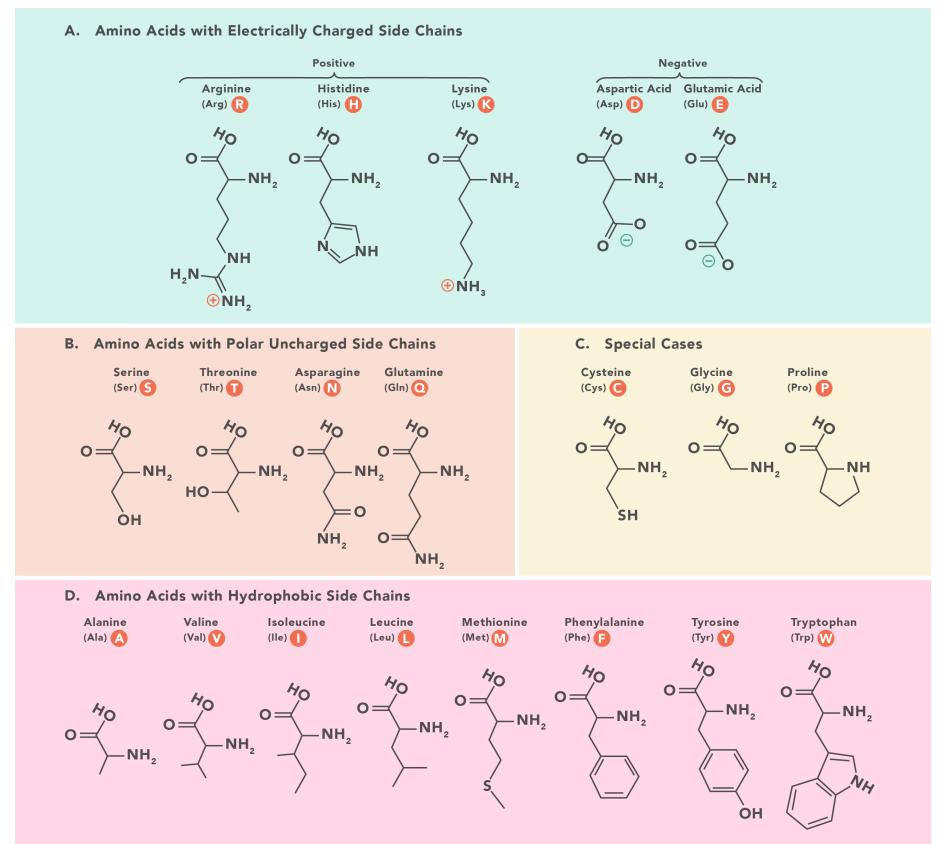
Proteinbiosynthesis

Translates nucleic acid  
to amino acid sequence



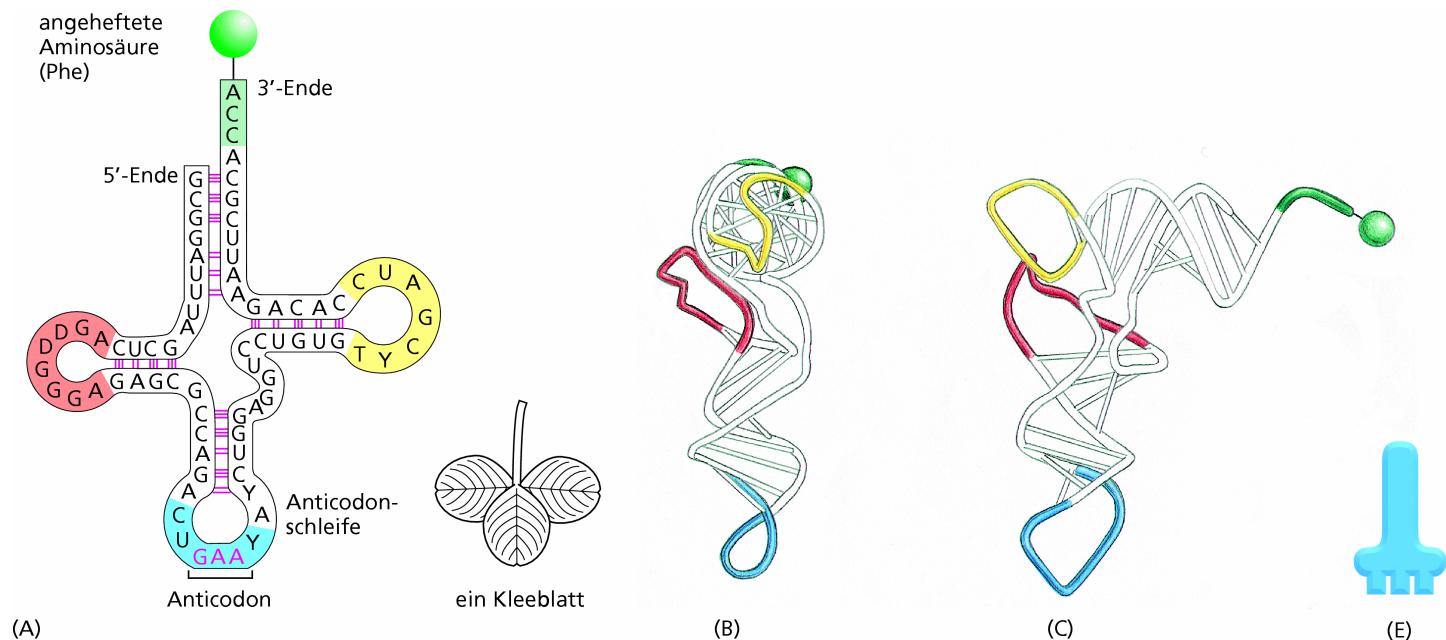
Four-letter RNA "alphabet"  
(4 bases)

Twenty-letter Protein alphabet  
(20 amino acids)



# Translation

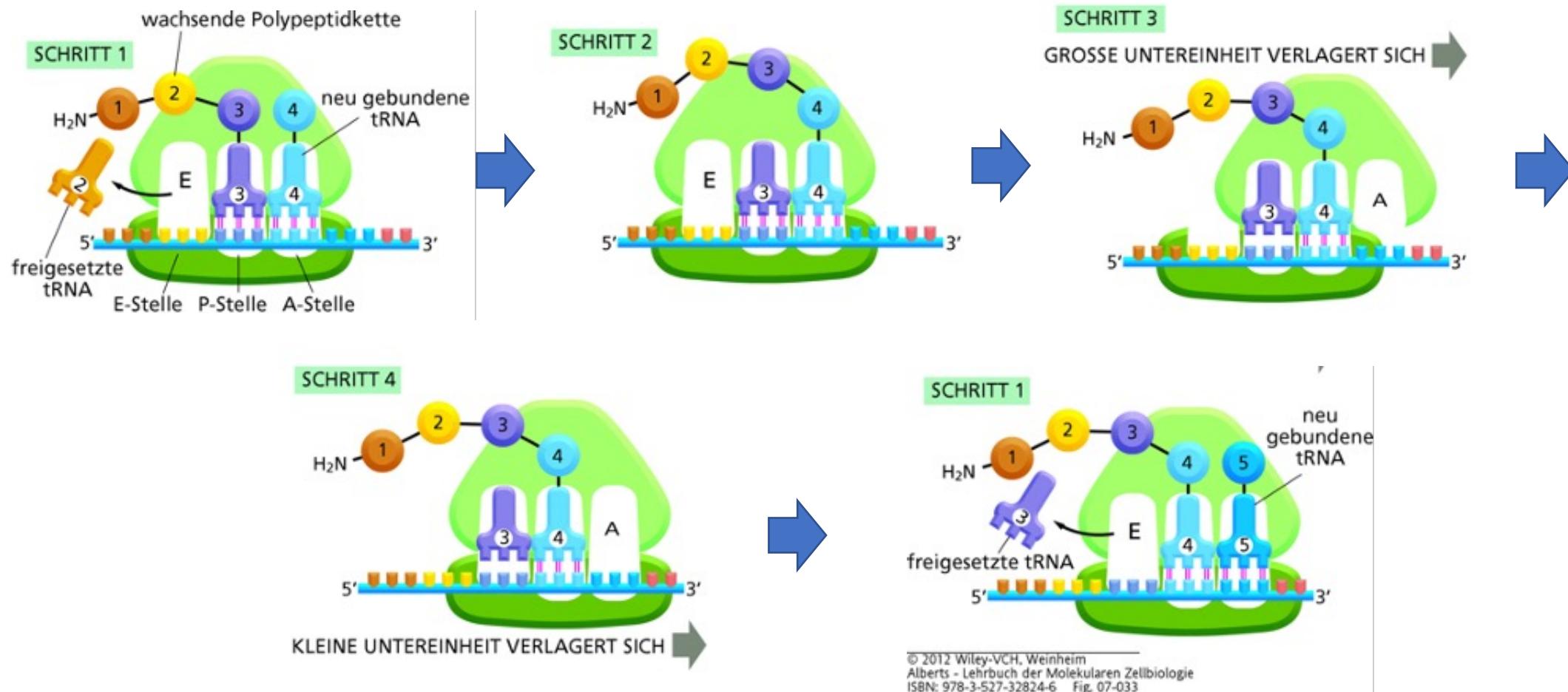
tRNAs act as adaptors to connect nucleic-acid sequences to amino-acids



© 2012 Wiley-VCH, Weinheim  
Alberts - Lehrbuch der Molekularen Zellbiologie  
ISBN: 978-3-527-32824-6 Fig. 07-026

# Translation

## takes place at Ribosomes



© 2012 Wiley-VCH, Weinheim  
Alberts - Lehrbuch der Molekularen Zellbiologie  
ISBN: 978-3-527-32824-6 Fig. 07-033

# Translation

## The genetic code

AGA										UUA						AGC				
AGG										UUG						AGU				
GCA	CGA									CUA						CCA	UCA			
GCC	CGC									CUC						CCC	UCC			
GCG	CGG	GAC	AAC	UGC	GAA	CAA	GGA	GGC	CAC	AUA	AAA	UUC	UUC	UUA	AGC	ACA				GUA
GCU	CGU	GAU	AAU	UGU	GAG	CAG	GGG	GGU	CAU	AUC	AAG	UUU	CCG	CCU	AGU					GUC
										CUG	CUU	AAA	CCG	UCG	ACG					UAA
										AUU	AAG	AUG	CCG	UCG	ACG					UAG
												UUU	CCU	UCU	ACU					UGA
													UCU	UGG	UAC					UAG
														UGG	UAU					UAA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

© 2012 Wiley-VCH, Weinheim  
Alberts - Lehrbuch der Molekularen Zellbiologie  
ISBN: 978-3-527-32824-6 Fig. 07-024

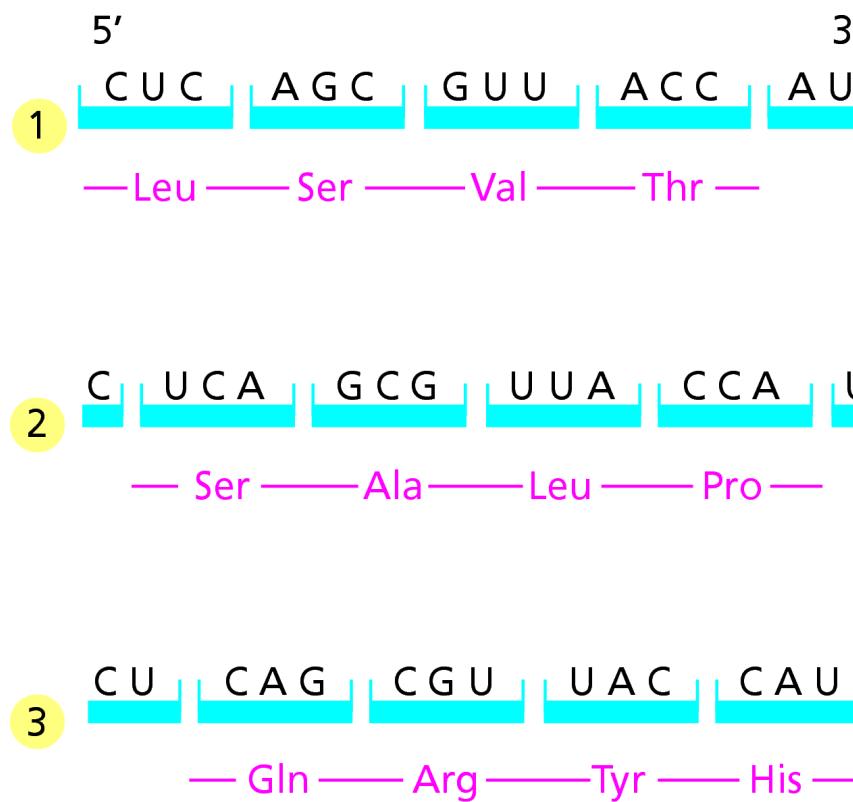
Start

AUG

- The Genetic code is like a dictionary used to translate nucleic acid sequence into amino acid sequence
- The RNA is read in **triplets** of bases (corresponding to the anticodons of the tRNAs) called codons
- Since RNA is a linear Polymer consisting of 4 different bases, there are  $4 \times 4 \times 4 = 64$  possible combinations of bases
- Since there are only 20 different amino acids, the code has evolved to be redundant (**degenerate code**)
- There are two special types of codons signaling the start and stop of a protein: **Start & Stop codons**
- This genetic code is used by all life forms on earth

# Translation

## Open Reading Frame (ORFs)



- The partitioning of the genetic code into triplets gives rise to 3 potential “reading frames” on any given stretch of DNA/RNA
- The presence of a Start codon not only signals the onset of translation, but at the same time defines the reading frame
- An ORF is opened by a start-codon and closed by a subsequent stop-codon lying within the same reading frame
- Deletions or insertion within a coding sequence shift the reading frame and therefore disrupt the information that follows them

# From strands to strings

- **Fasta (.fasta, .fa, .fna)** is a text-based format for biological sequences
- Headers/deflines are preceded by ">" and contain a name or description of the following sequence
- The monomers (nucleotides or amino acids) are represented by single letters in the sequence
- Nucleic acid sequence always depicted as single strand and from 5'-3' direction
- Protein sequence typically starting from the N-Terminal end
- Sequence can either be contained in one line (**sequential**) or in multiple lines (**interleaved**)

```
1 >Example_DNA_seq_1
2 ATGCGTGGGTGTCCATGCGTATGTCGTATGCTATGTGATGCTTGTAGTTTTCCACTCTATCTTAAGCGGACTTGCCTGATGTCGTATGCTATGTGATGGC
3 >Example_DNA_seq_2
4 GCCATCACATAGCATACGACATCACGCAAGTCCGCTTAAGATAGAGTGGAAAAAACTACGACAAGCATCACATAGCATACGACATCACGATGGACACCCACGCAT
5
```

```
1 >Example_Protein_seq_1
2 MFVFLVLLPLVSSQCVNLTTTQLPPAYTNSFRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFD
3 NPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKCEFQFCNDPFLGVYYHKNNSWMESEFRVY
4 GWIFGTTLDSKTQSLLIVNNATNVVIKCEFQFCNDPFLGVYYHKNNSWMESEFRVYSSANNCTFEYVSQPFLMDLEGK
5 >Example_Protein_seq_2
6 SSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQT
7 LLALHRSYLTPEGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTTDAVDCALDPLSETKCTLKSFTVEKGIIYQTSNFRV
8 QPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGSPTKLNDLCFTNVYADSF
9
```



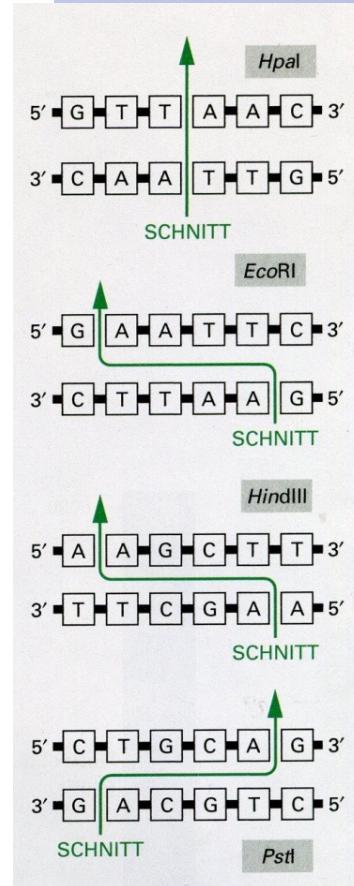
# Excercise

**Write a python script that:**

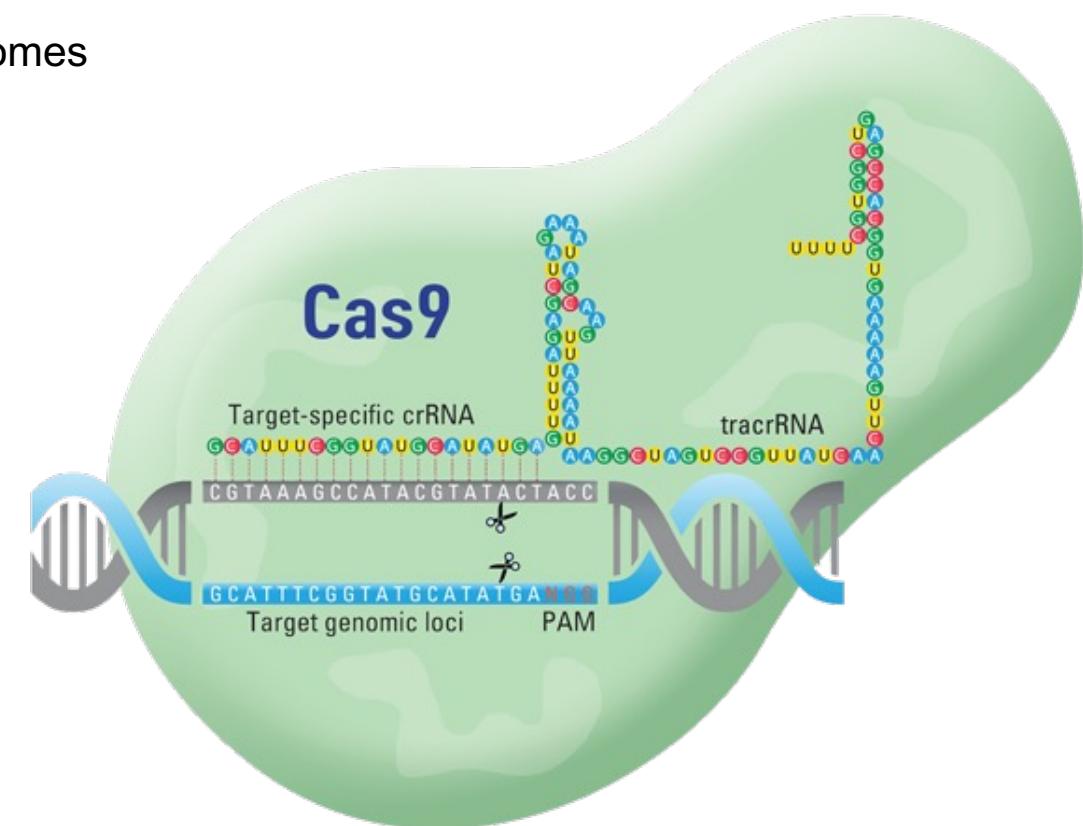
- 1. finds all ORFs in the provided DNA sequence**
- 2. Selects the ORFs that code for proteins >40 amino acids**
- 3. Selects the ORFs that do not lie within a larger ORF**
- 4. translates the selected ORFs into an amino acid sequence**

# Tools for molecular biology

## Restriction enzymes and other endonucleases



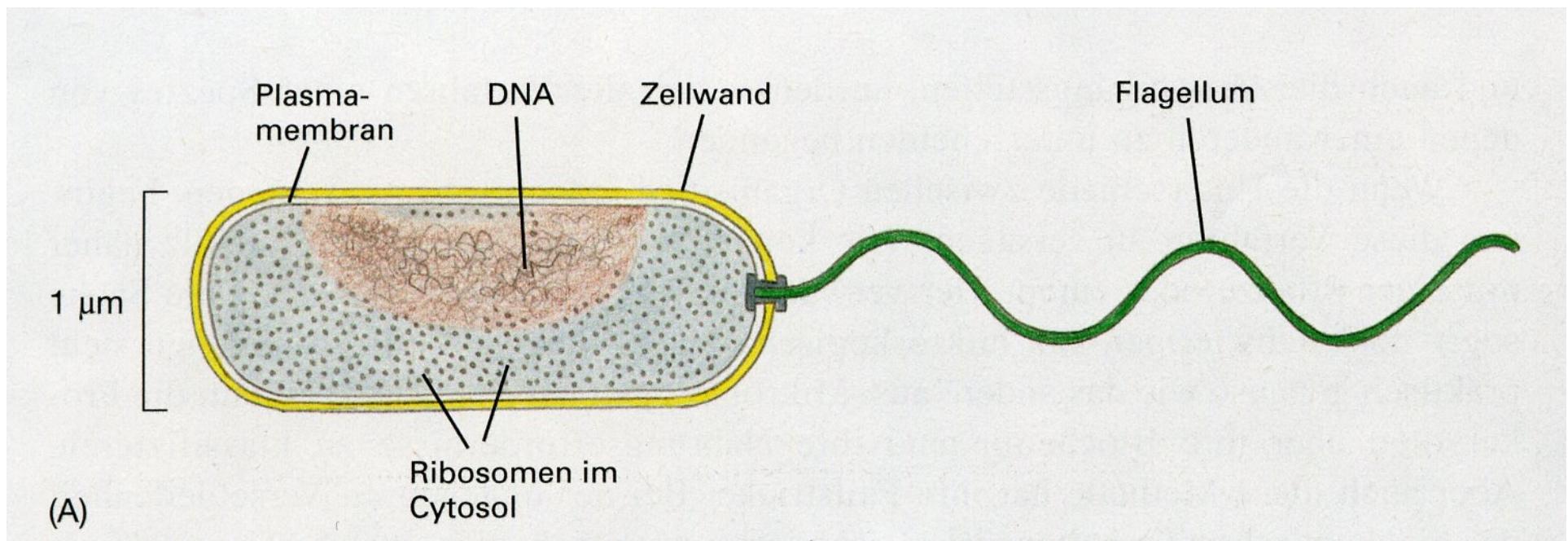
- **Restriction enzymes** recognize specific short DNA motifs and cleave the DNA at these positions.
  - Restriction sites are typically sequence palindromes
- 
- **CRISPR/Cas9 system**
  - Cas9 is a RNA-guided endonuclease
  - Short, tailored RNA molecules can specify the cutting site
  - All genomic sites become available to manipulation



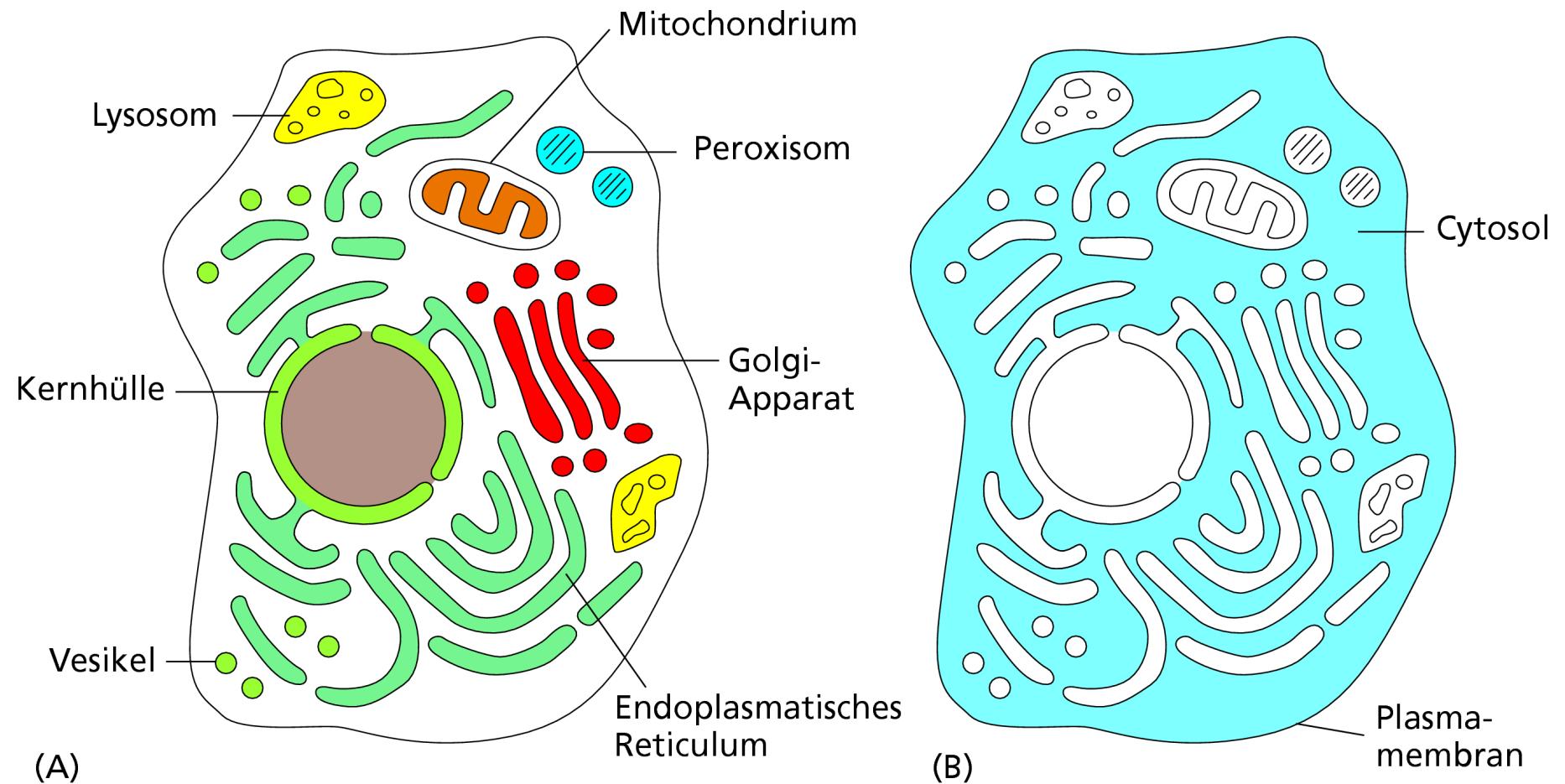
# Classification of life

- How are those fundamental processes connected and organized in order to actually bring forth life?
- The most elementary unit of live is a cell
- A cell is an enclosed space in which the reactions of life can take place and that contains the hereditary material through which the information necessary to maintain those reactions can be passed on
- Basically there are two blueprints for cells

# Prokaryotic cells



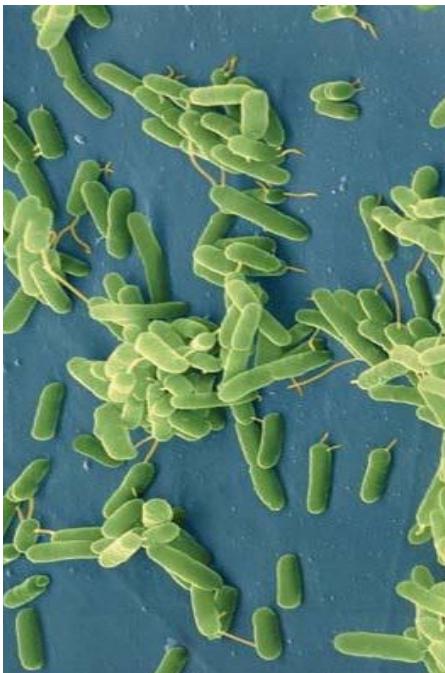
# Eukaryotic cells



# Domains

## based on molecular criteria (rDNA sequence)

Bacteria



Archaea



Eukaryota



# What about Viruses?

- **Viruses are evolutionary very old (may date back to the last universal common ancestor)**
- **Viruses infecting all types of life exist: from bacteria over yeast to humans**
- **Through their ubiquitous interactions with life forms on all branches of the evolutionary tree they have been affecting and shaping life on this planet throughout its existence**
- **However: Virus aren't capable of replicating themselves independently of their hosts**
- **Next: A spotlight on human pathogenic viruses**

Data Science in  
Bioinformatics  
WS 22/23

# Next: Viruses



Lecturer: PD Dr. Ricarda Schmithausen  
Course Date: 17.10.2022