

Biologische Grundlagen

Jens Quedenfeld

Jan Stricker

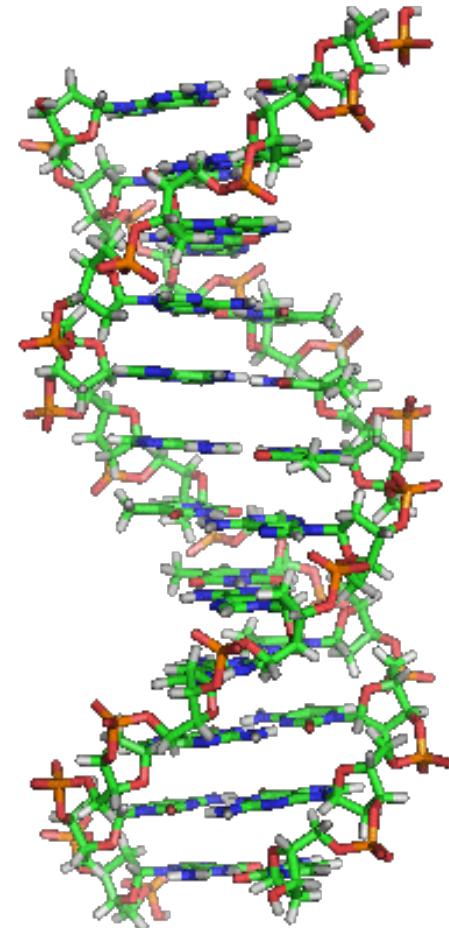
03. April 2014

Inhalt

- DNA und Chromosomen
- Zellteilung, DNA-Replikation
- Proteinbiosynthese
- Vererbung
- Mutationen und Varianten
- Sequenzierung
- Dateiformate

DNA

- DNA = **Deoxyribonucleic acid**
(Desoxyribonukleinsäure)
- Träger der **Erbinformation**
- Information für den Bau von
Proteinen
- Kommt in allen Lebewesen vor

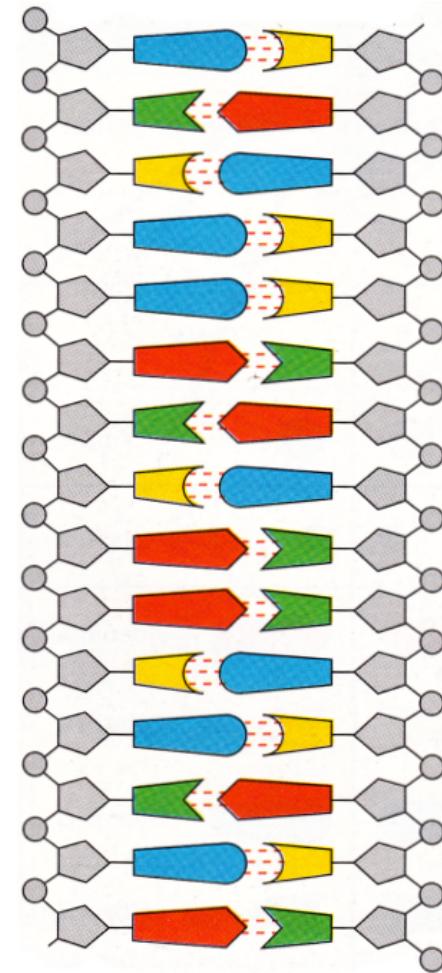


Quelle: <http://de.wikipedia.org/wiki/DNA>

Aufbau der DNA

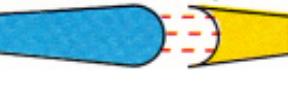
- Doppelstrang aus
- Nucleotid {
- Zucker (Desoxyribose) 
 - Phosphatrest 
 - Nuleobase
 - Adenin 
 - Guanin 
 - Cytosin 
 - Thymin 

DNA-Moleköl

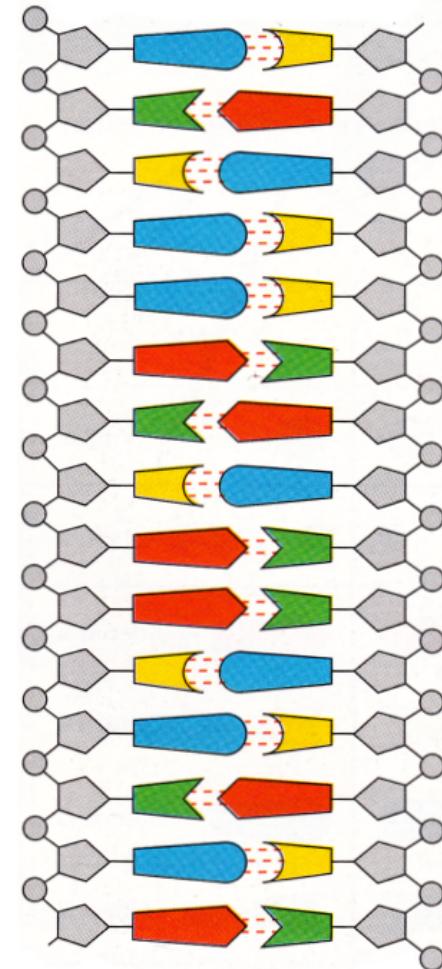


Quelle: Knodel, Linder Biologie, S. 345/346

Aufbau der DNA

- Basensequenzen sind **komplementär**
 - Adenin (A)  (T) Thymin
 - Guanin (G)  (C) Cytosin

DNA-Moleköl

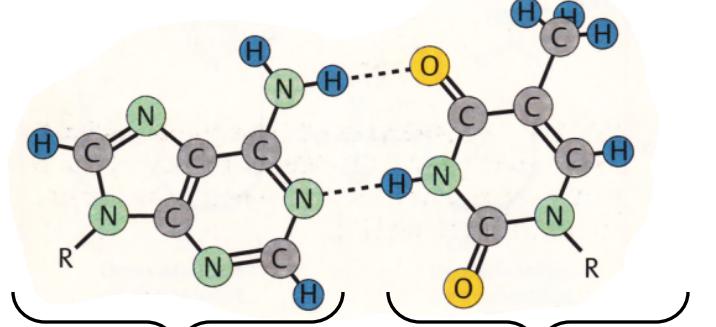


Quelle: Knodel, Linder Biologie, S. 345/346

Chemische Struktur der DNA

- Basensequenzen sind **komplementär**
 - Adenin (A) (T) Thymin

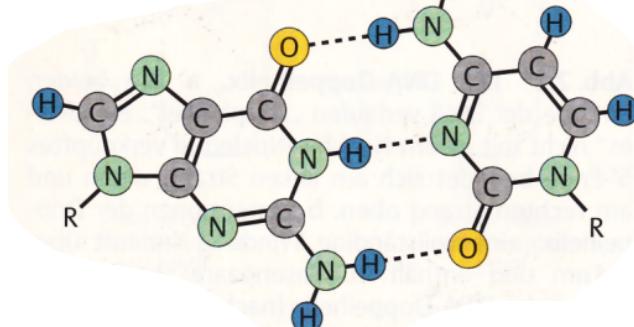
2 Wasserstoff-Brücken-bindungen



Purin-Basen

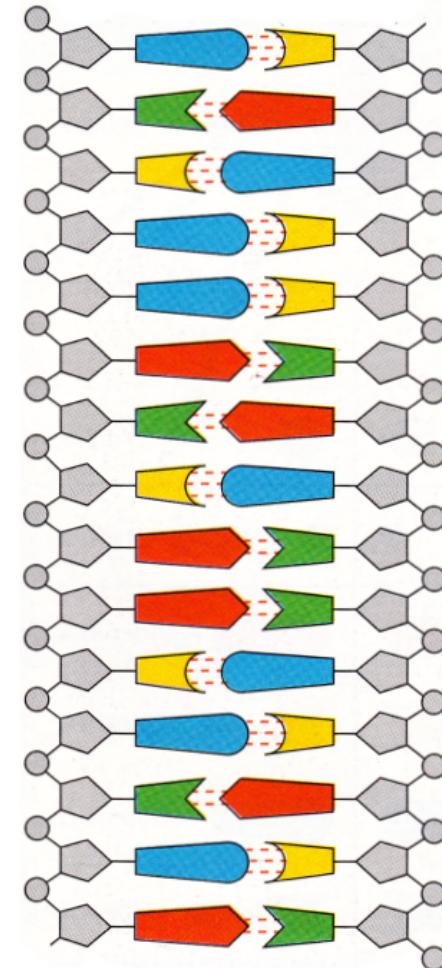
Pyrimidin-Basen

3 Wasserstoff-Brücken-bindungen



- Guanin (G) (C) Cytosin

DNA-Moleköl

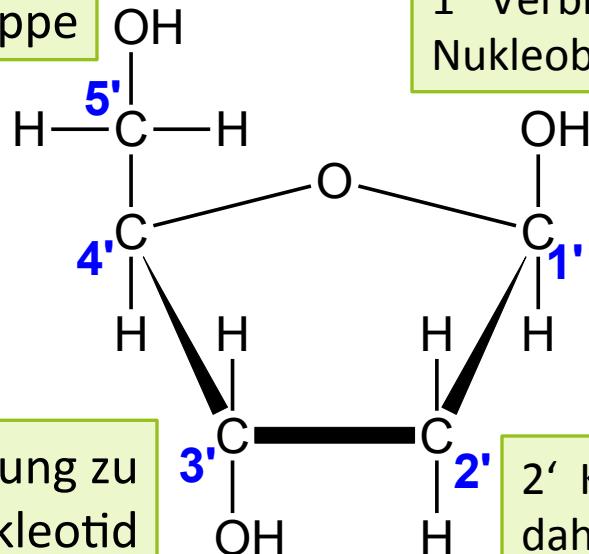


Quelle: Knodel, Linder Biologie, S. 345/346

Chemische Struktur der DNA

- Nummerierung der C-Atome der **Desoxyribose** (Zucker): 

5' Verbindung zur Phosphatgruppe

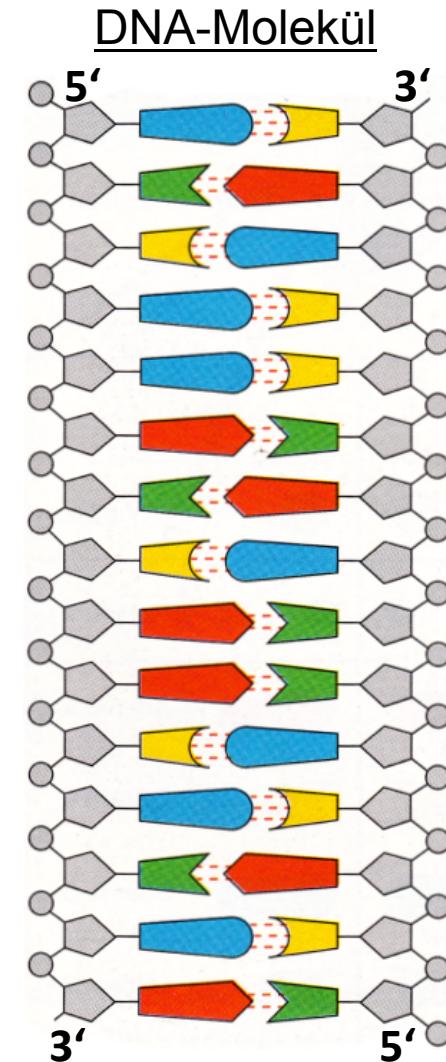


1' Verbindung mit der Nukleobase (A, C, G, T)

3' Verbindung zu nächstem Nukleotid

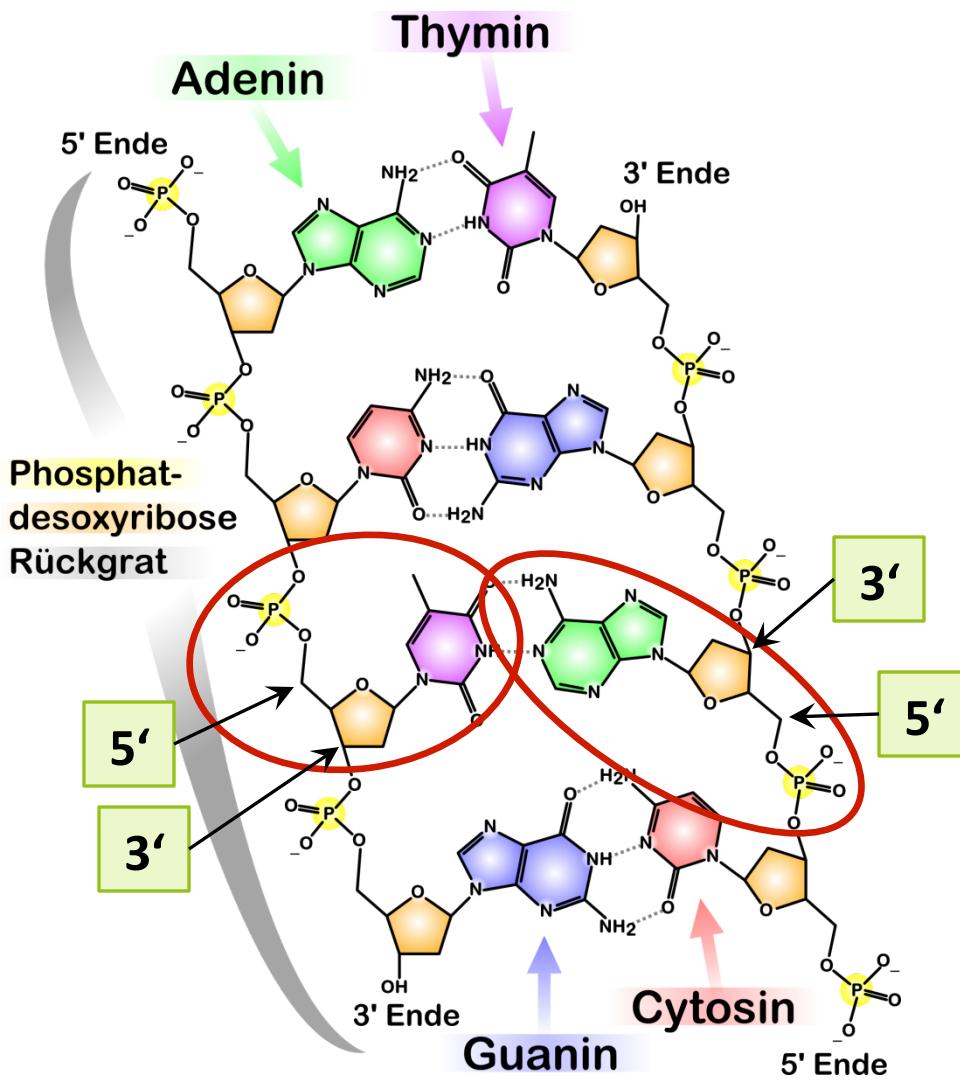
2' Keine OH-Gruppe, daher „Desoxy“

→ DNA-Stränge haben eine **Richtung**
3'-5' oder 5'-3'

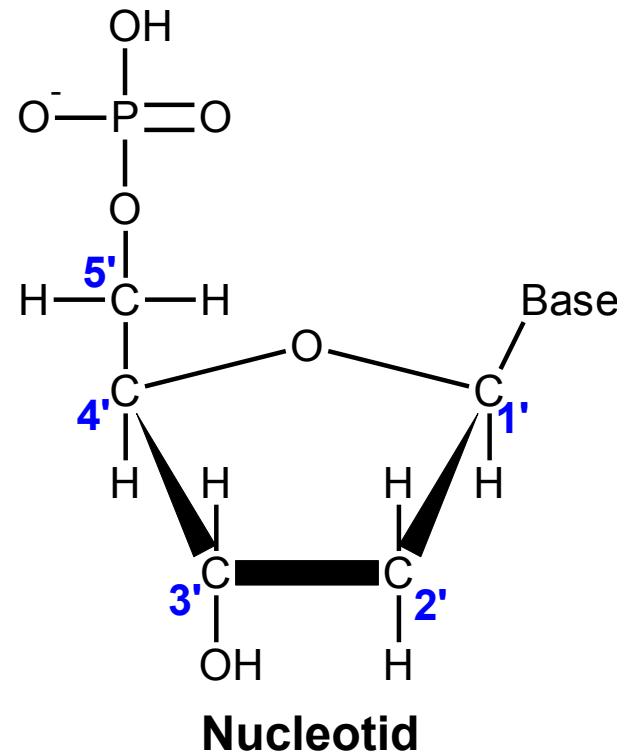


Quelle: Knodel, Linder Biologie, S. 345/346

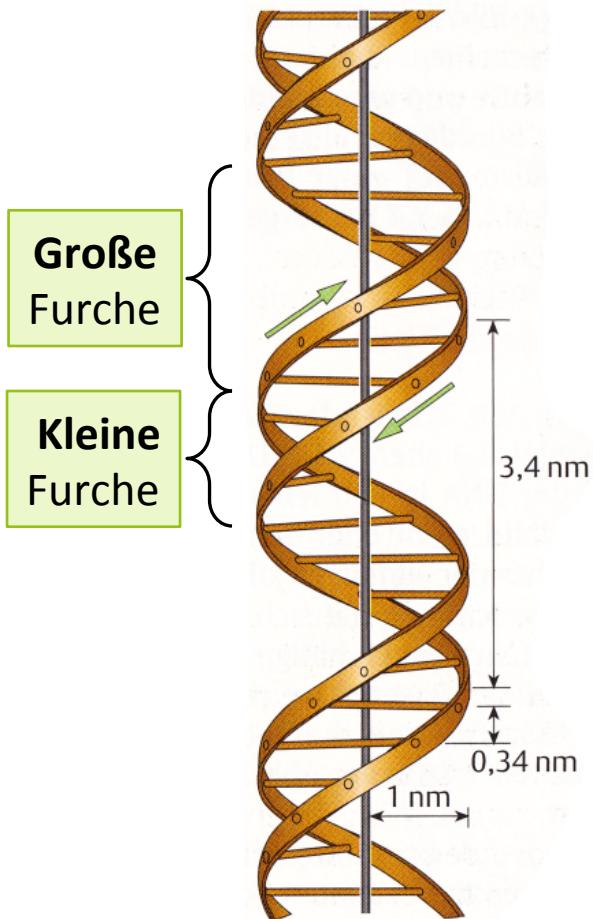
Chemische Struktur der DNA



- DNA-Stränge sind **antiparallel**

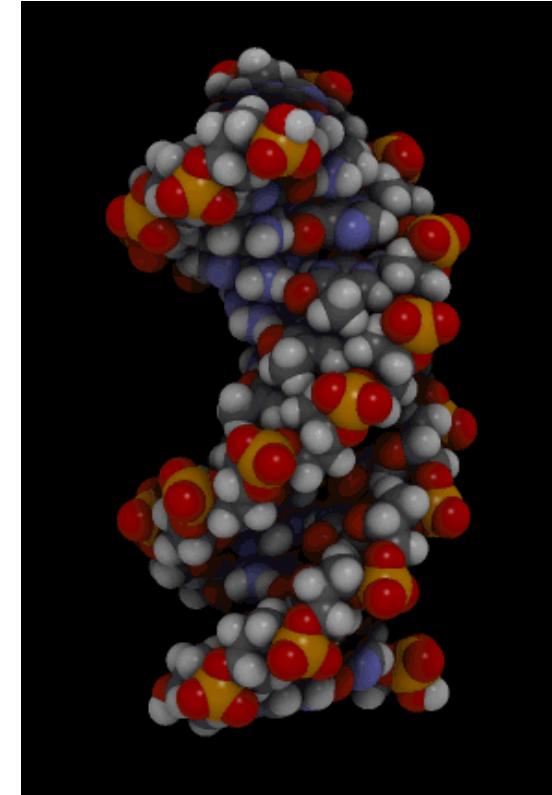


Räumliche Struktur



Quelle: Knippers, Molekulare Genetik, S.11

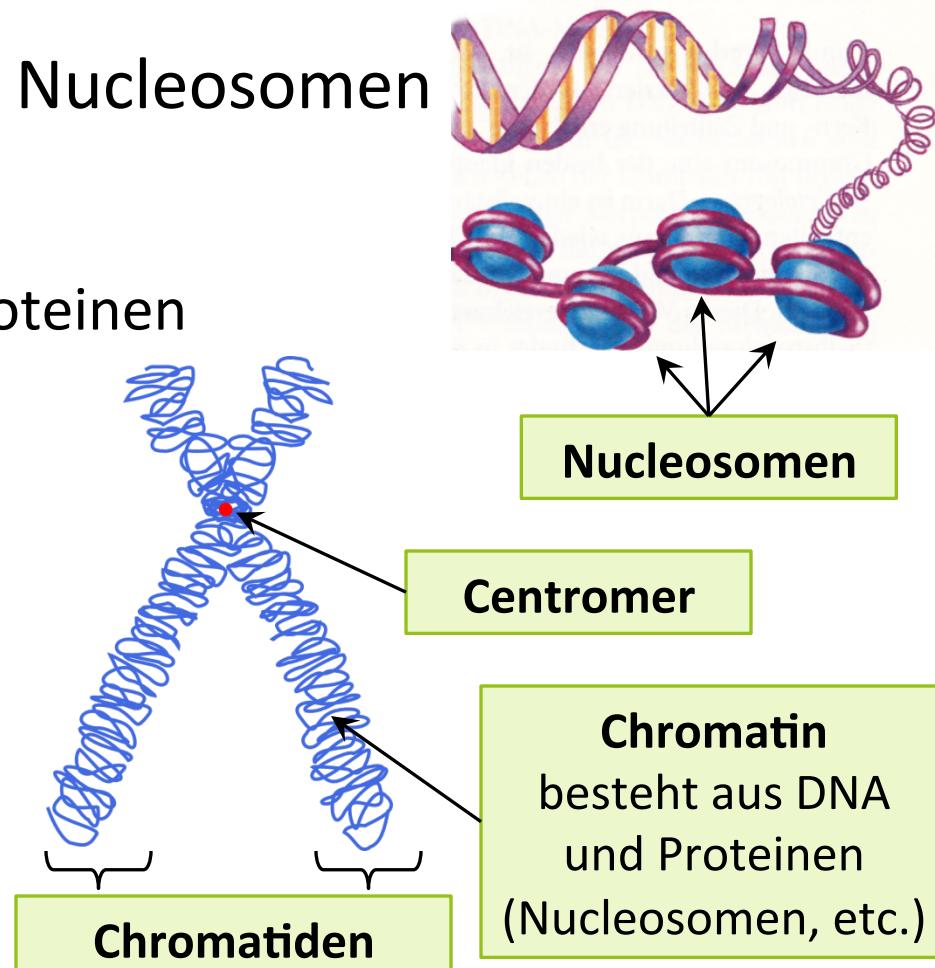
- Doppelhelix
- rechtsläufig
- Etwa 10 Basenpaare pro Windung



Quelle: http://commons.wikimedia.org/wiki/File:Bdna_cropped.gif

Chromosomen

- DNA-Moleküle können sehr lang werden
 - Bis zu 263 Millionen Basenpaare (Mbp) beim Menschen
- Aufwicklung der DNA um Nucleosomen
- Chromosomen
 - Bestehen aus DNA und Proteinen
 - **Vor** Zellteilung:
2 Chromatiden
 - **Nach** Zellteilung:
1 Chromatid
 - **Zwischen** Zellteilungen:
freies Chromatin



Quelle: <http://de.wikipedia.org/wiki/Chromosom>

Chromosomen

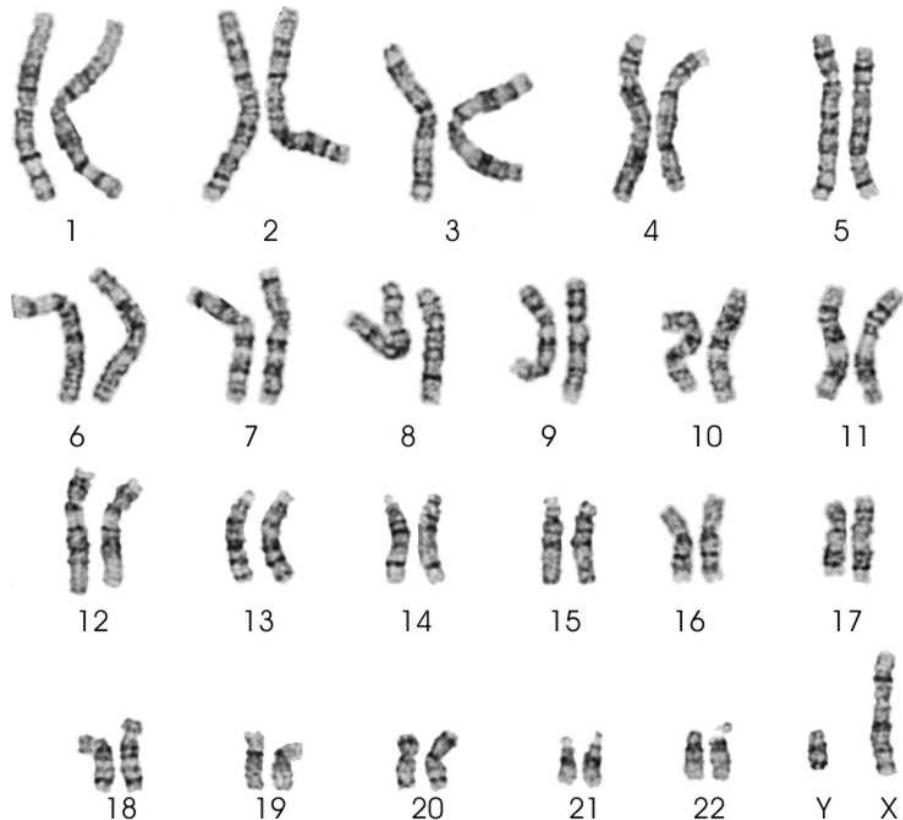
Beim Menschen:

- 46 Chromosomen
- 23 Paare
 - Jeweils eines von Vater und Mutter geerbt
- 2 Geschlechts-Chromosomen
 - Frauen = XX
 - Männer = XY
- 3 Milliarden

Basenpaare (3 Gbp)

- Informationsgehalt: 750 MB
- Länge: 1 Meter

Karyogramm des Menschen



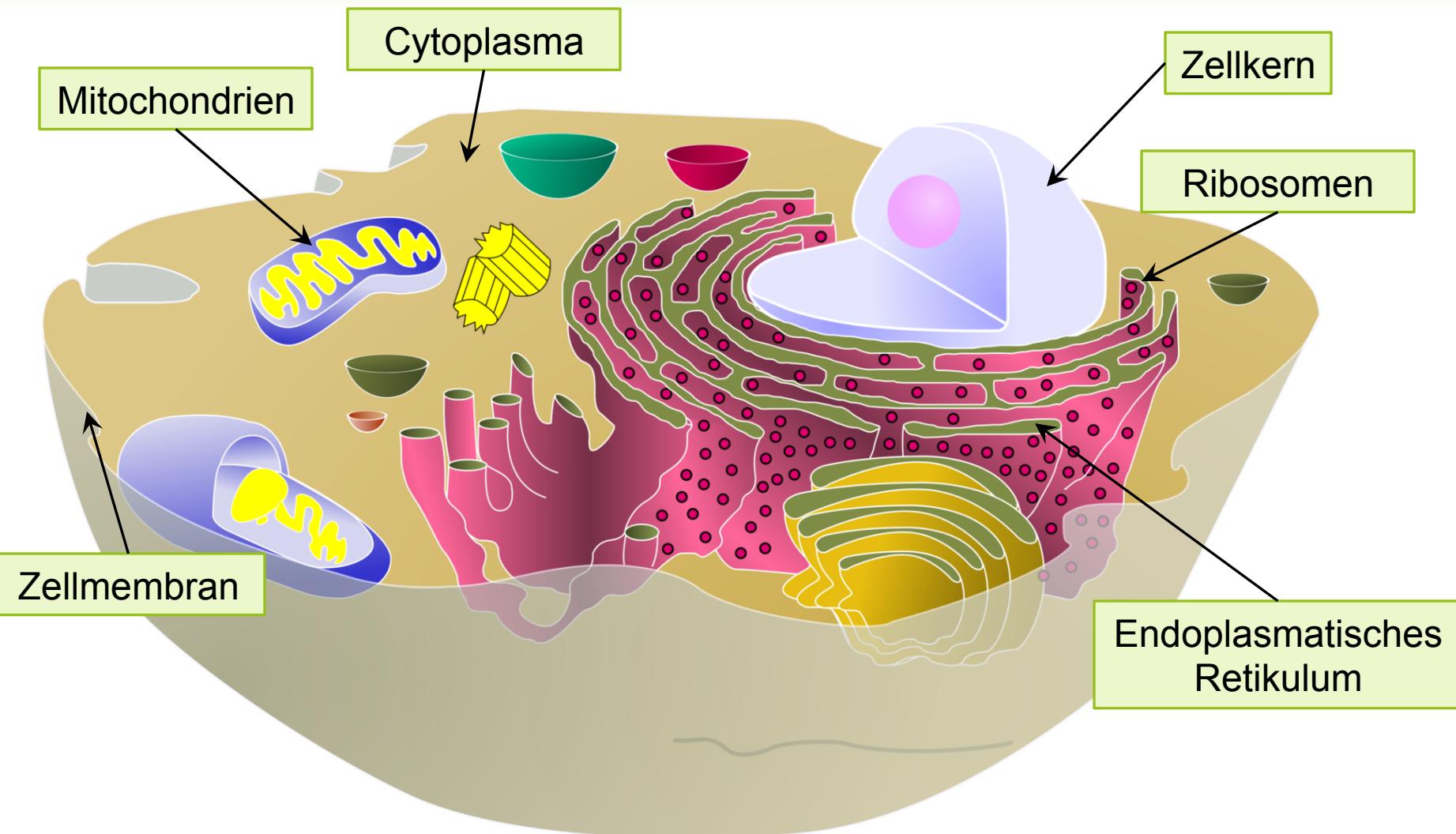
Quelle: <http://www.laborlexikon.de/images/Karyogramm-813.JPG>

Prokaryoten / Eukaryoten

Unterscheidung zwischen

- **Eukaryoten** (= Menschen, Tiere, Pflanzen, Pilze)
 - Chromosomen
 - DNA im Zellkern
- **Prokaryoten** (= Bakterien, Archaebakterien)
 - Kein Zellkern
 - Ringförmiges, in sich geschlossenes DNA-Molekül

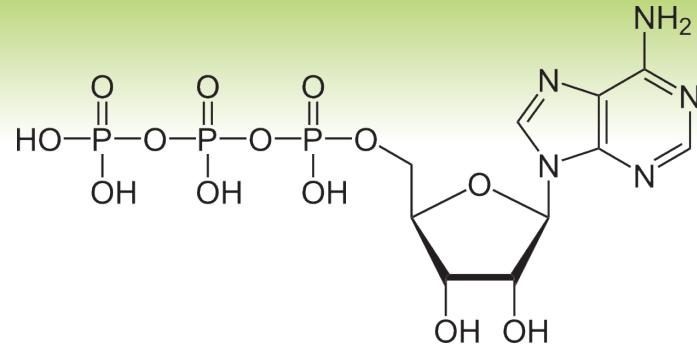
Aufbau einer eukaryotischen Zelle



Quelle: http://commons.wikimedia.org/wiki/File:Biological_cell.svg

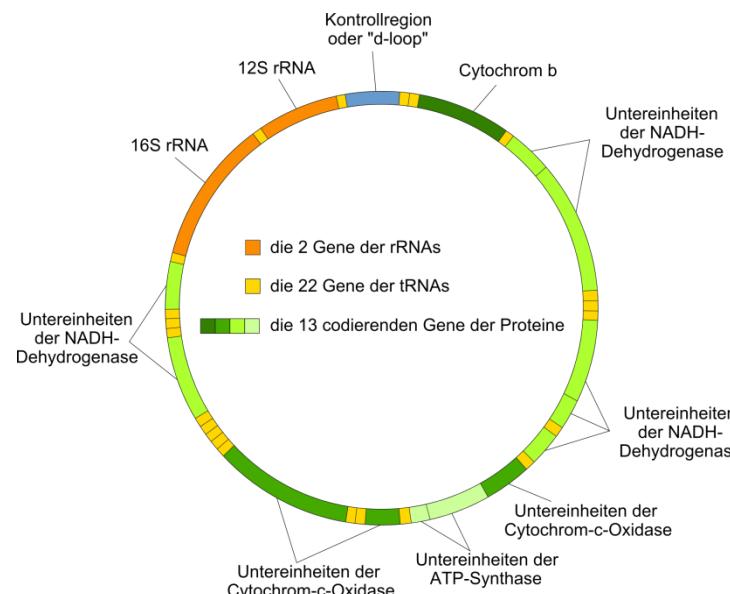
Mitochondrien

- In fast allen Zellen vorhanden
- „Energiekraftwerke“
- ATP-Herstellung
- Eigene, **ringförmige DNA**
- 16.569 Basenpaare
- Mitochondrien werden nur von der Mutter weitervererbt



Adenosintriphosphat

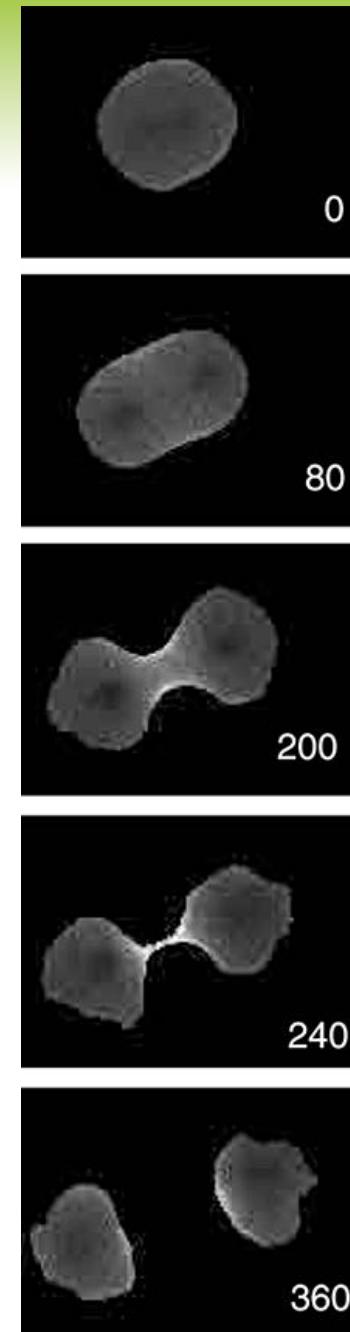
Bei Abspaltung einer Phosphatgruppe wird Energie frei



Mitochondriale DNA (mtDNA)

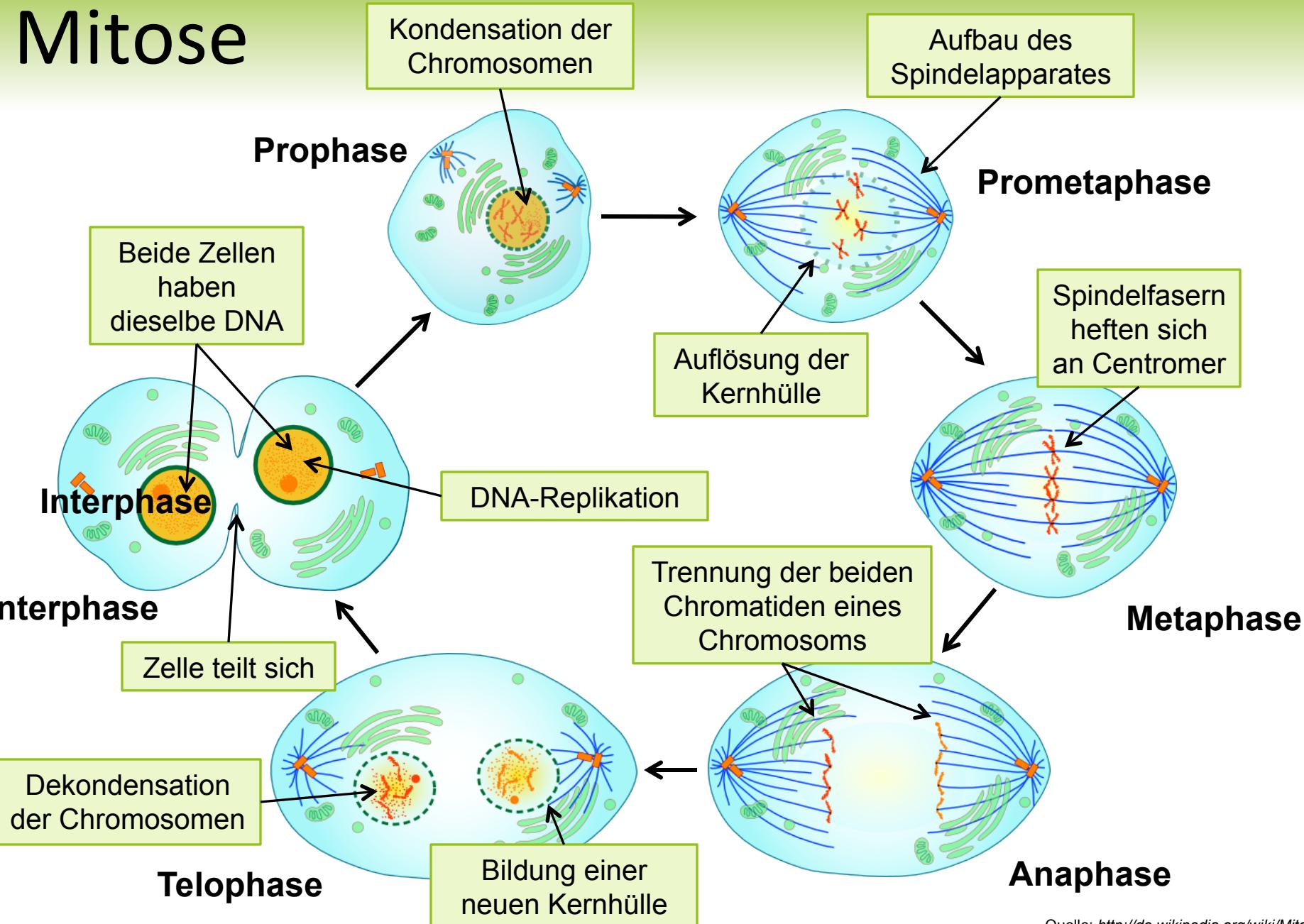
Zellteilung

- Verdoppelung der DNA
→ DNA-Replikation
- Aufteilung der DNA auf je eine Tochterzelle
→ Mitose



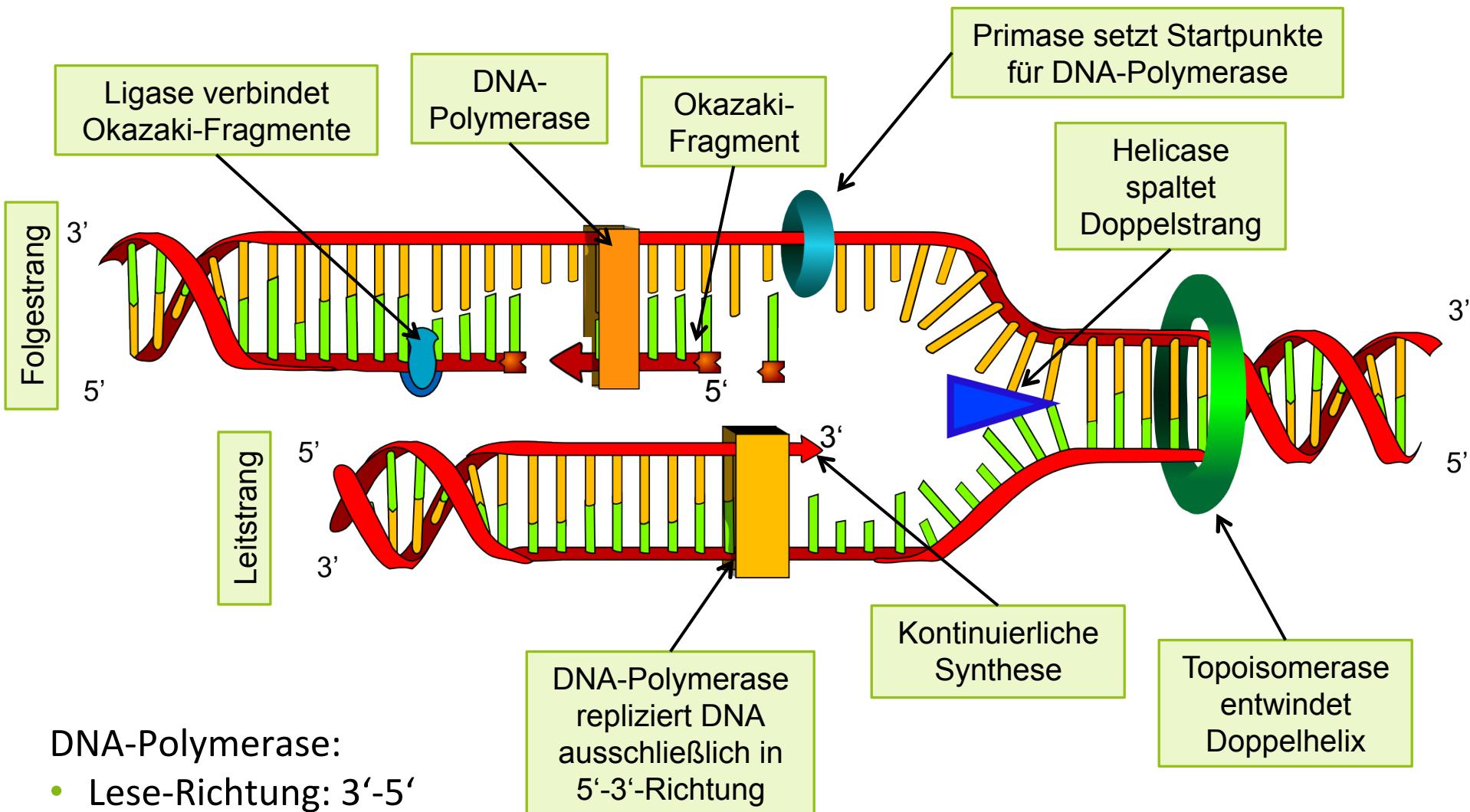
Quelle: <http://de.wikipedia.org/wiki/Zellteilung>

Mitose



Quelle: <http://de.wikipedia.org/wiki/Mitose>

DNA-Replikation



DNA-Polymerase:

- Lese-Richtung: 3'-5'
- Synthese-Richtung: 5'-3'

Quelle: <http://de.wikipedia.org/wiki/DNA-Replikation>

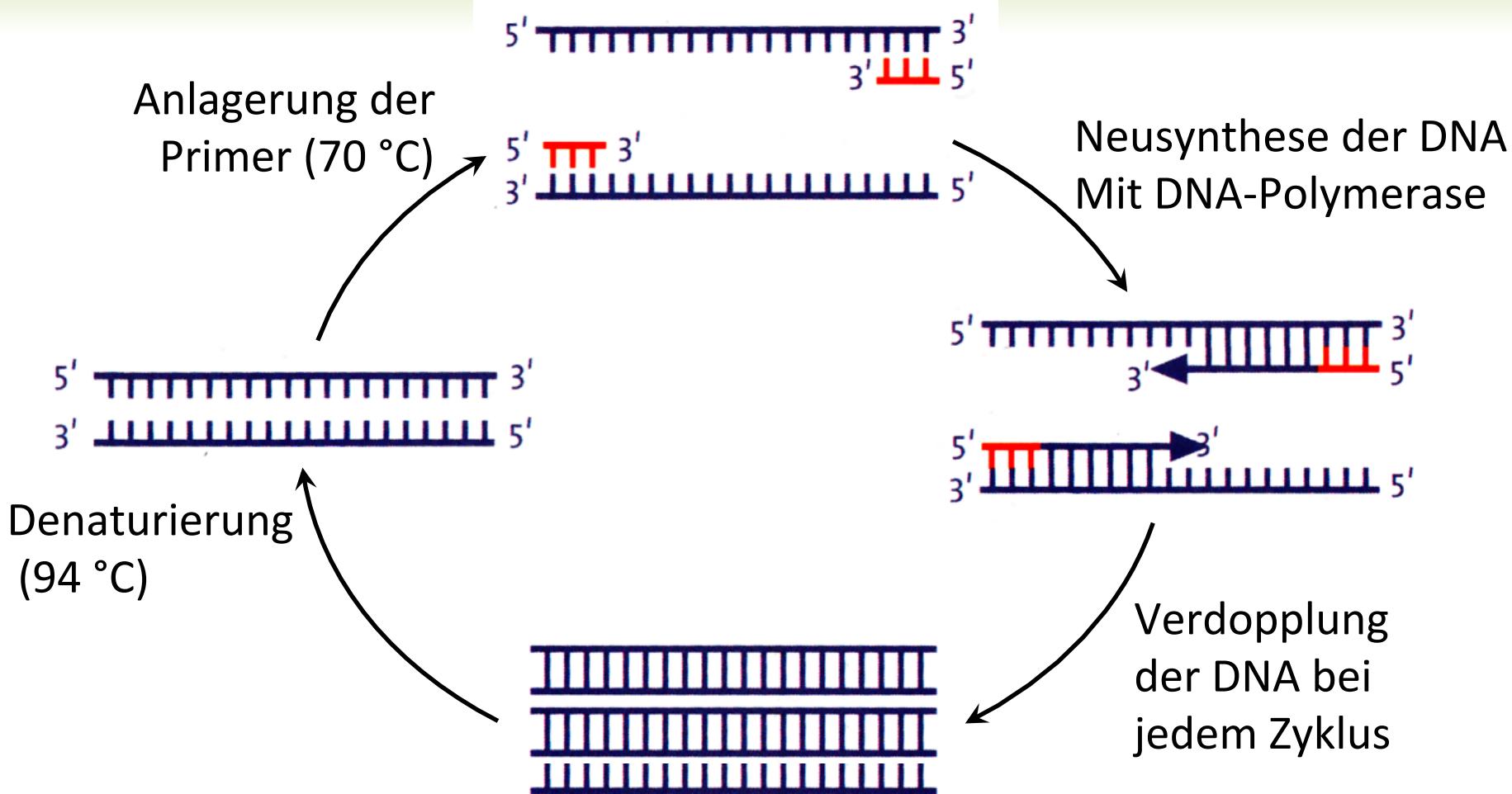
Künstliche DNA-Replikation - PCR

- PCR = Polymerase Chain Reaction
- Vervielfältigung von geringsten DNA-Mengen
- Anwendungen
 - genetischer Fingerabdruck
 - Vaterschaftstest
 - Sequenzierung
 - Analyse fossiler DNA
 - Erbkrankheiten nachweisen
- Wie funktioniert die PCR?



Quelle: http://www.celliculturetechnology.de/fileadmin/user_upload/Geraete/PCR-Thermocycler.jpg

Künstliche DNA-Replikation - PCR



- Problem** normale DNA-Polymerase wird bei 94° C zerstört
Lösung Taq-Polymerase von thermophilen Bakterien

Quelle: Knodel, Linder Biologie, S. 350

Protein-Biosynthese

Protein-Biosynthese = Übersetzung eines DNA-Abschnitts (Gens) in ein Protein

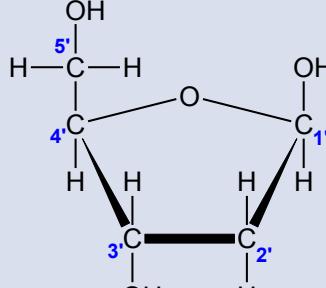
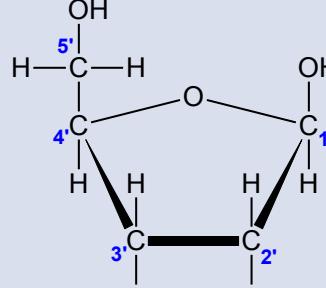
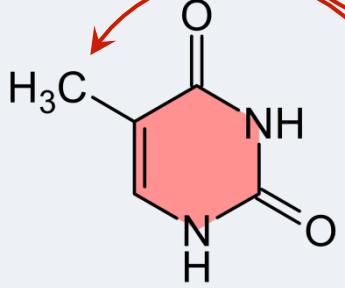
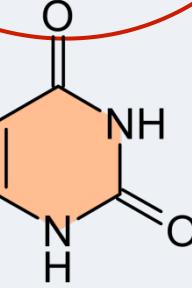
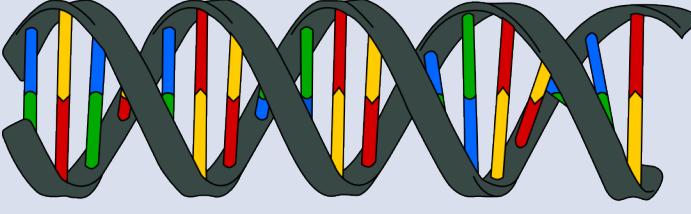
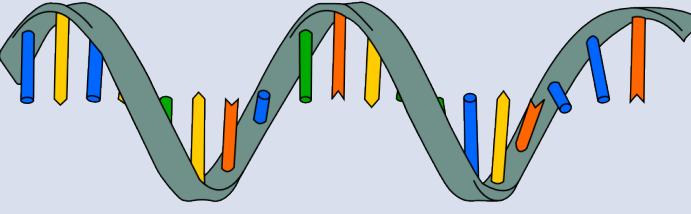
1. Transkription

- Ablesen eines Gens aus der DNA
- Erstellen einer RNA-Kopie dieses Gens (messenger-RNA = mRNA)

2. Translation

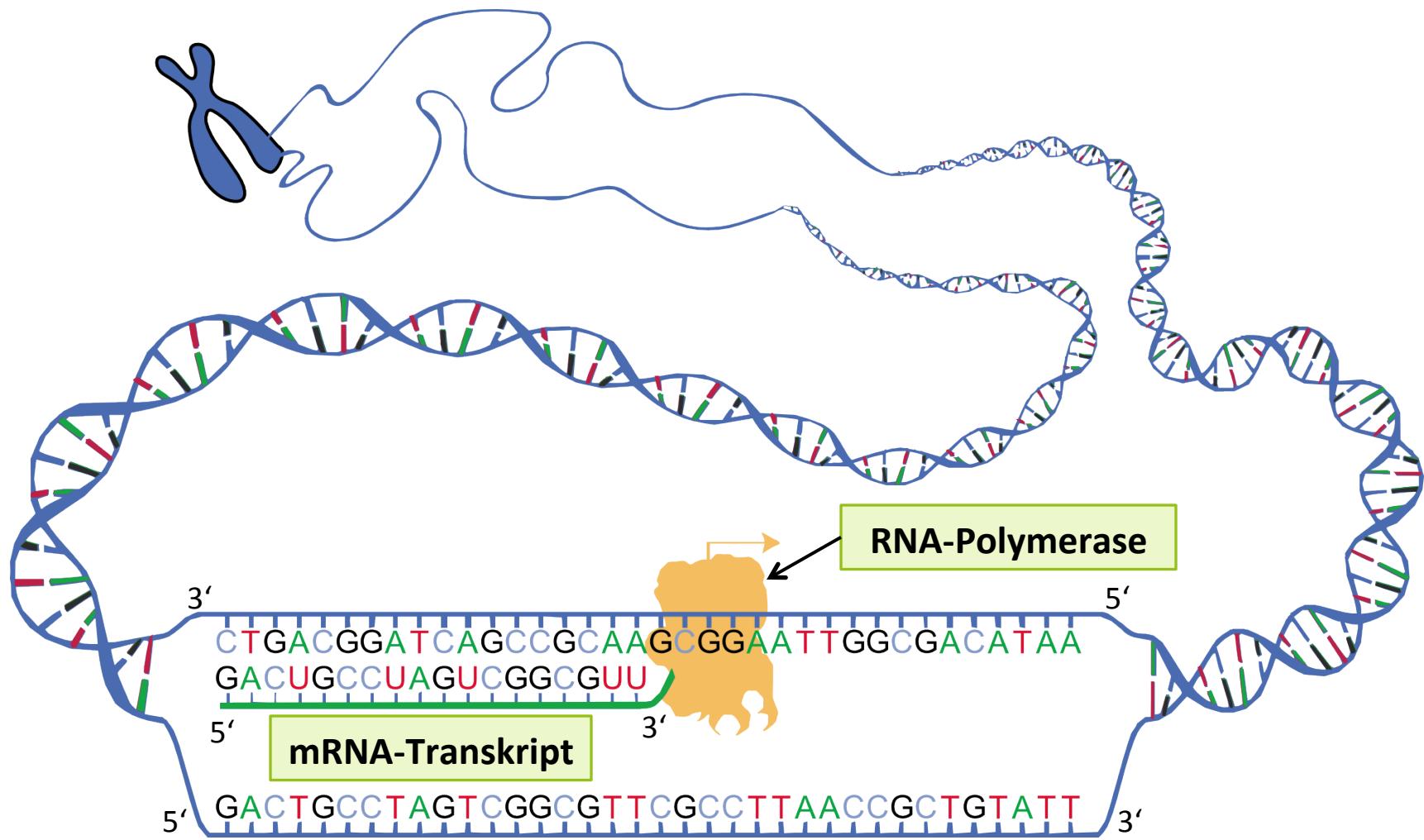
- Übersetzung der RNA in eine Aminosäure-Sequenz (= Vorstufe eines Proteins)

Unterschiede DNA und RNA

DNA		mRNA
Desoxyribose		
Thymin (T)		(U) Uracil 
Doppelstrang		Einzelstrang 

Quelle: http://commons.wikimedia.org/wiki/File:Difference_DNA_RNA-DE.svg

Transkription



Quelle: http://commons.wikimedia.org/wiki/File:DNA_transcription.svg

Transkription

Initiation

- RNA-Polymerase setzt sich an **Promotor**
 - DNA-Sequenz mit Information, wann und in welchem Zelltyp Gen transkribiert werden soll
 - Codiert kein Protein, sondern reguliert Genexpression

Elongation

- RNA-Polymerase liest DNA in 3'-5'-Richtung
- Erstellt komplementäre mRNA-Kopie

Termination

- Bei bestimmter DNA-Sequenz oder mit Hilfe von Proteinen

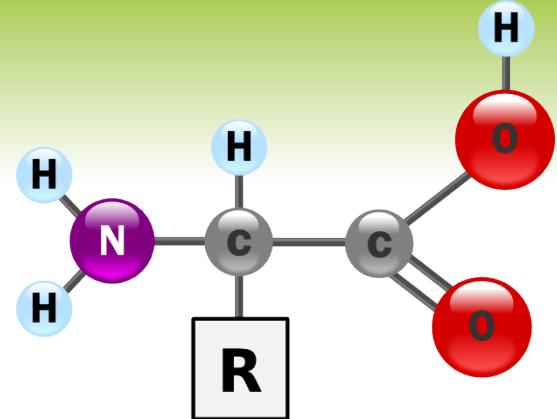
Translation

Übersetzung der mRNA in ein Protein

- **Prokaryoten**
 - Translation noch während der Transkription
- **Eukaryoten**
 - **Prozessierung** → *gleich*
 - Reife mRNA verlässt durch Kernpore den Zellkern
 - Danach Translation → *jetzt*

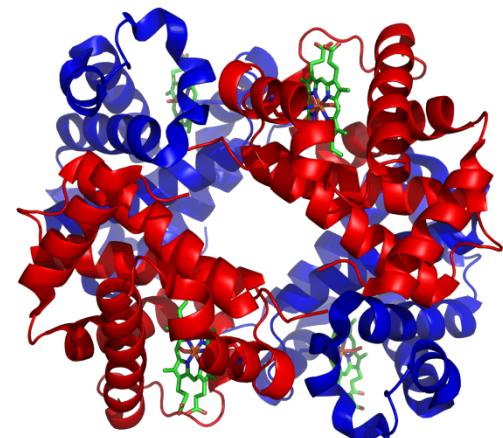
Was ist ein Protein?

- Proteine bestehen aus **Aminosäuren**
- Es gibt **20** natürliche Aminosäuren
- mRNA codiert eine Aminosäure-Sequenz (= Polypeptid)
- Proteine können aus mehreren **Aminosäuren-Sequenzen** bestehen
- **Faltung** für Funktion entscheidend
- Funktionen im Organismus
 - Stofftransport
 - Katalysator für chemische Reaktionen (Enzyme)
 - Ionen-Kanäle
 - etc.



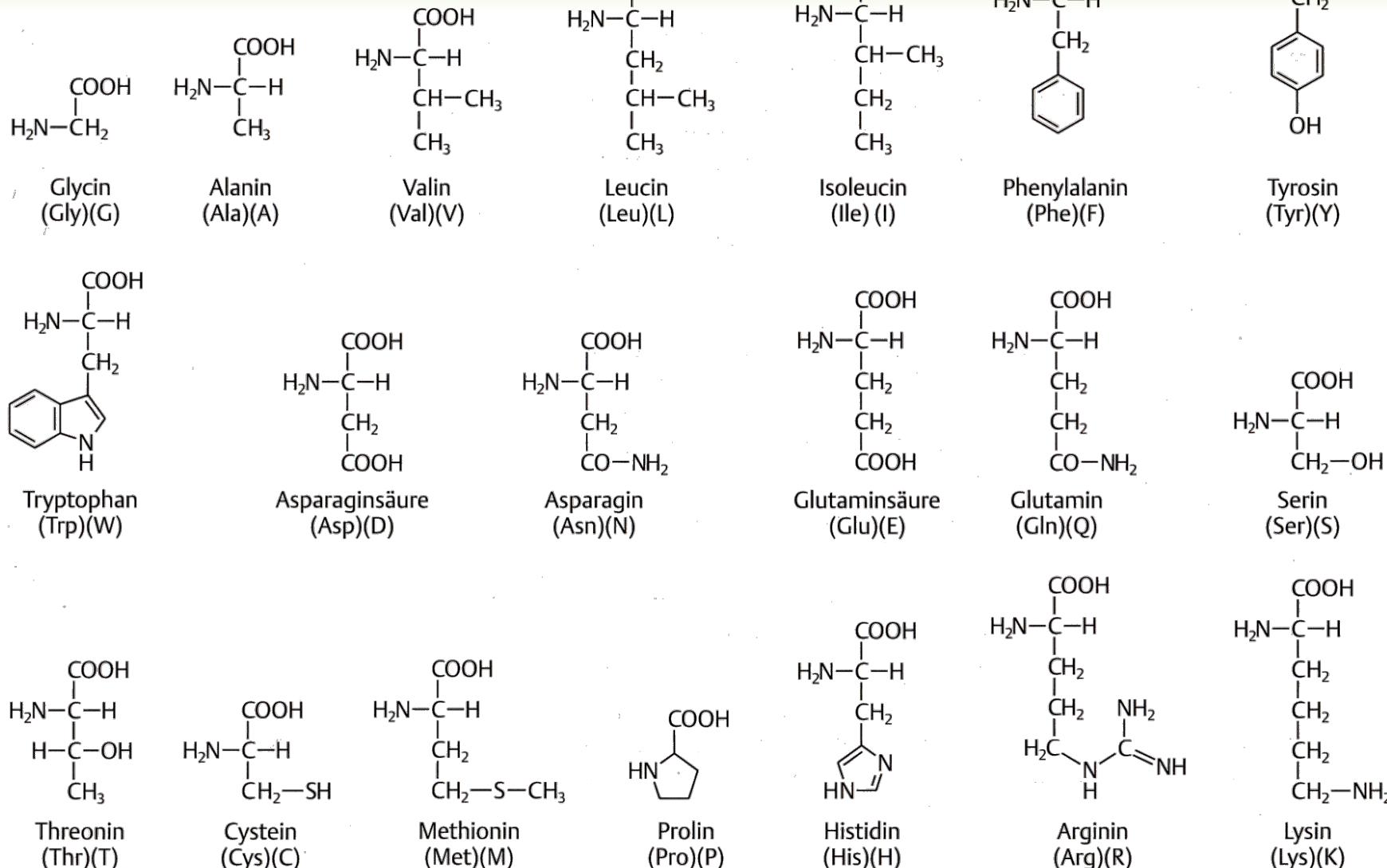
Allgemeine Struktur einer Aminosäure

Die 20 Aminosäuren unterscheiden sich im Rest R



Hämoglobin A
Sauerstoff-Transport

Aminosäuren



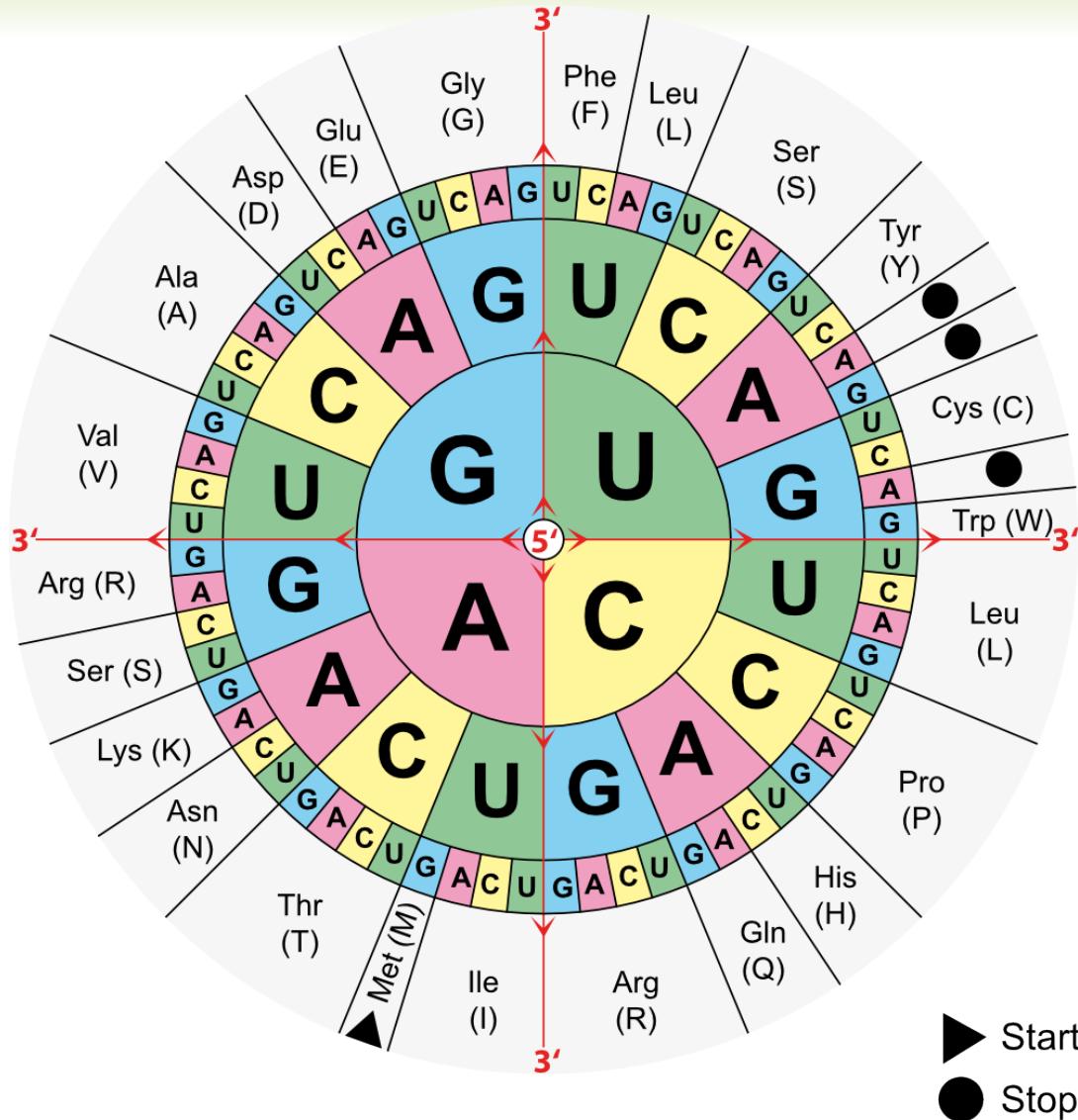
Quelle: Knippers, Molekulare Genetik, S.38

Translation

- mRNA-Strang codiert eine Aminosäure-Sequenz
 - **Problem** Es gibt 20 Aminosäuren,
aber nur 4 verschiedene Basen (A,C,G,U)
 - **Drei** Basen bilden ein **Codon**
und codieren eine Aminosäure
- Genetischer Code

Genetischer Code

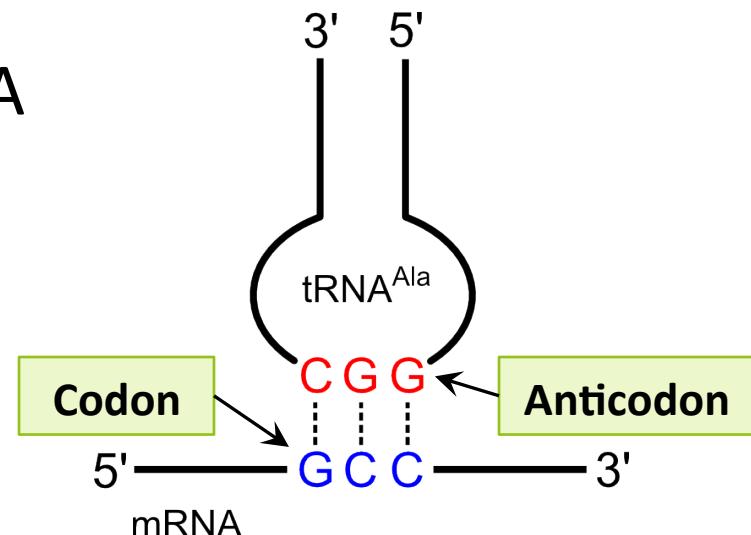
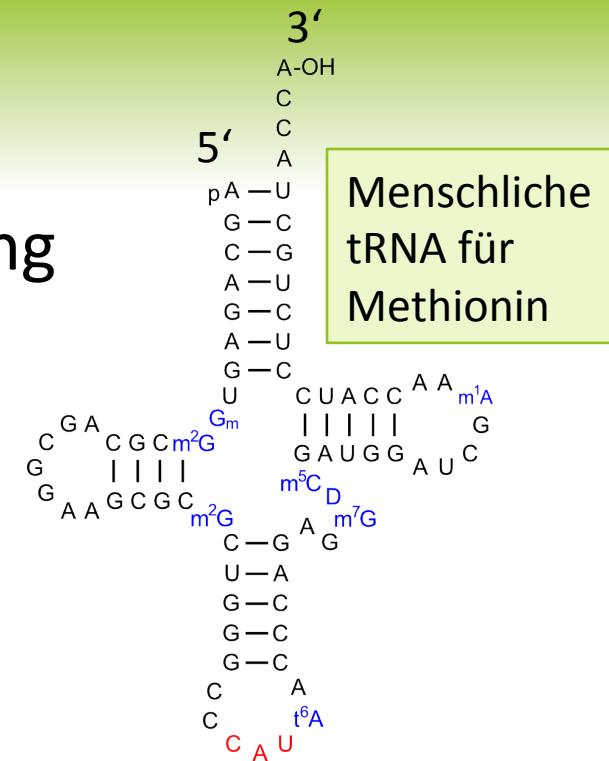
- Drei Basen bilden ein **Codon** und codieren eine Aminosäure
- Ablesen der mRNA in 5'-3'-Richtung
- Startcodon: AUG
 - Begin der Translation
- Stopcodons
 - „Abwurf“ der fertigen Aminosäure-Sequenz



Quelle: http://de.wikipedia.org/wiki/Genetischer_Code

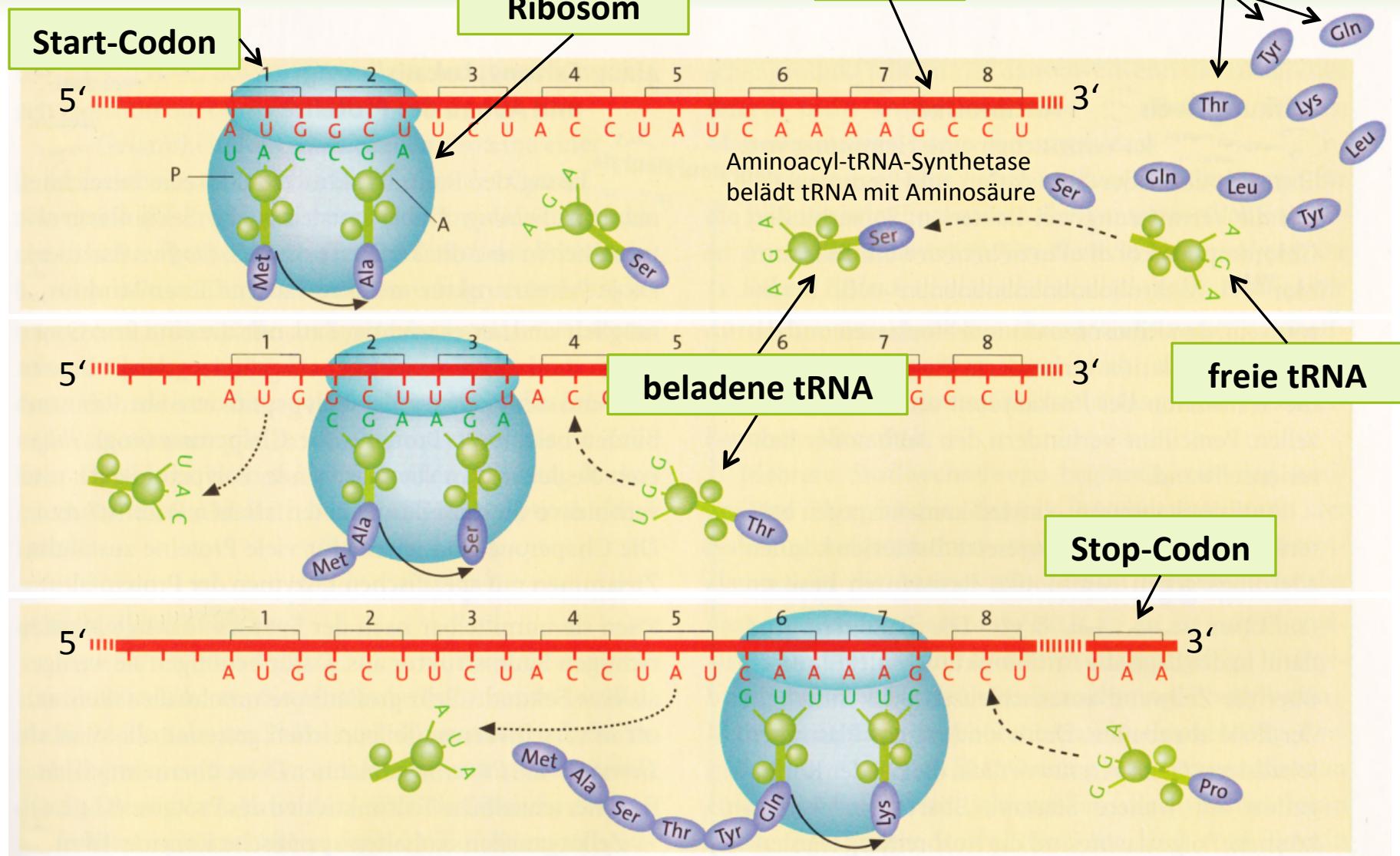
Translation

- Ribosom liest mRNA in 5'-3'-Richtung
 - tRNA (= transfer-RNA) transportiert Aminosäure
 - besitzt zu Codon komplementäres **Anticodon**
 - für jede Aminosäure existiert (mindestens) eine tRNA
 - Ribosom bringt mRNA und tRNA zusammen und verbindet Aminosäuren
 - Typische Länge:
100-800 Aminosäuren



Quelle: <http://de.wikipedia.org/wiki/TRNA>

Translation



Quelle: Knodel, Linder Biologie, S. 359

RNA-Prozessierung

Zwischen Transkription und Translation

- Bei Eukaryoten
- prä-mRNA wird zu „reifer“ mRNA

- **Capping**

- Zusätzliches, modiziertes Guanin-Nukleotid am 5'-Ende

- **Polyadenylierung**

- Verlängerung des 3'-Endes mit Adenin-Nukleotiden

- **RNA-Editing** → *jetzt*

- **Splicing** → *gleich*

Verhinderung des vorzeitigen Abbaus der prä-mRNA durch Enzyme

Veränderung der codierenden mRNA-Bereiche

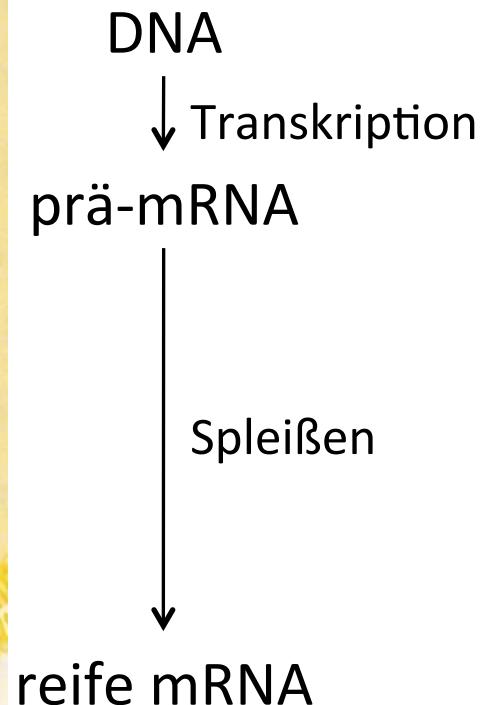
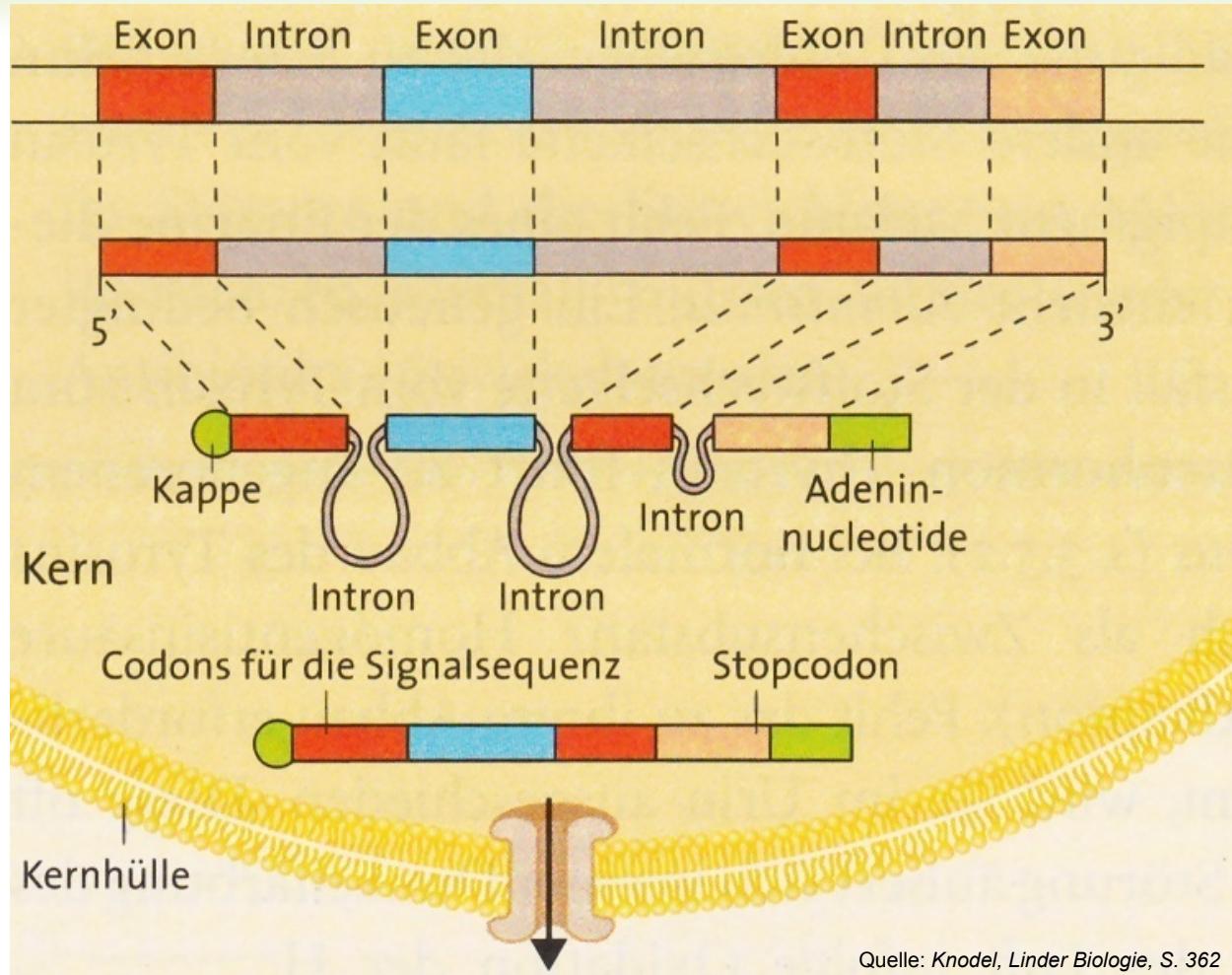
RNA-Editing

- Änderung der Nukleotid-Sequenz nach der Transkription

Beispiel: Apolipoprotein B

- Zwei Isoformen
 - Leberzellen Apo B100 mit 4536 Aminosäuren
 - Dünndarmzellen Apo B48 mit 2153 Aminosäuren
- Beide Isoformen entstehen durch die **gleiche** prä-mRNA mit etwa 14.000 Nukleotiden
- Ursache: RNA-Editing
 - Umwandlung von Cytosin in Uracil an spezifischer Stelle
 - Dadurch entsteht ein **Stopp-Codon**
- RNA-Editing sehr häufig bei Eukaryoten
- Teilweise neue Basen (\neq A,C,G,U), wie Inosin
 - kann mit allen Basen eine Verbindung eingehen
 - kommt z.B. als Anticodon in tRNA vor

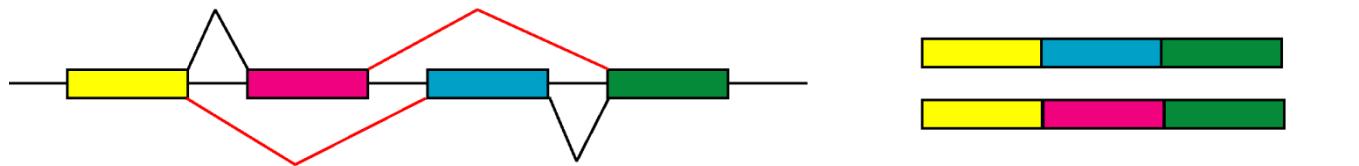
Splicing



- **Exons** Protein-codierende Sequenzen
- **Introns** werden beim Spleißen herausgeschnitten

Alternatives Splicing

- prä-mRNA kann auf verschiedene Weisen gespleißt werden
 - Überspringen von Exons



- Alternative Spleißstellen



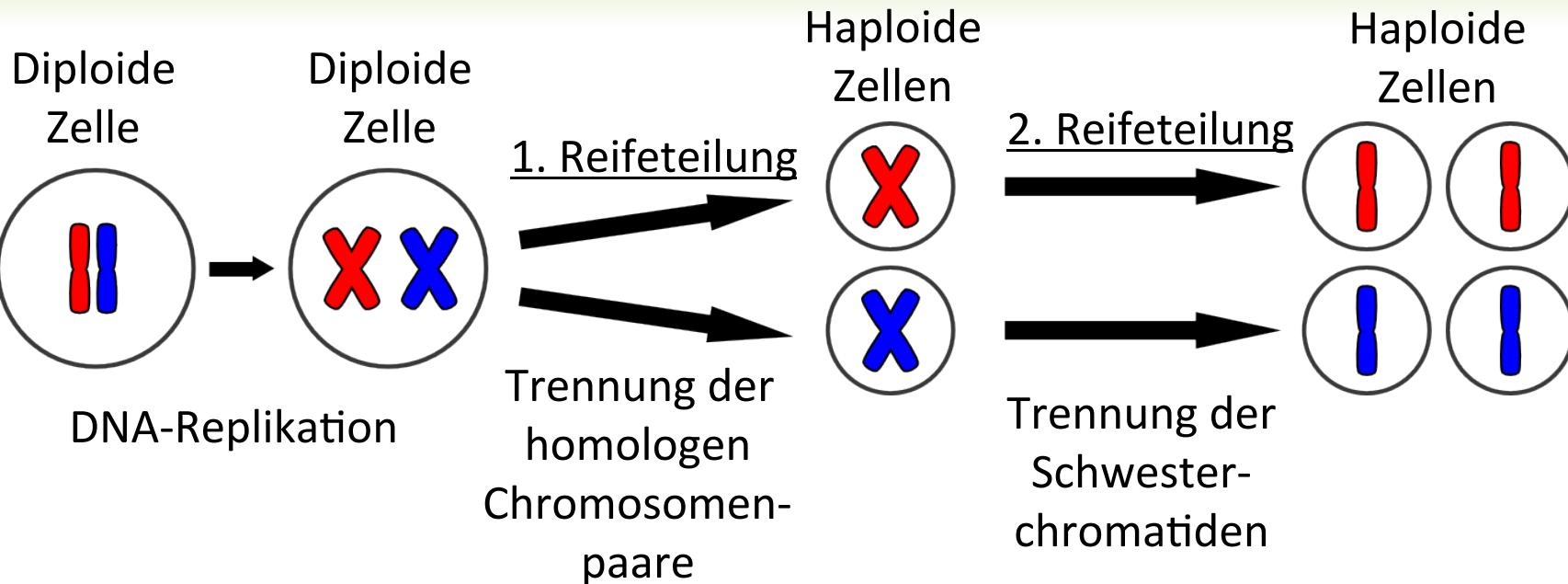
- Ein Gen (eine prä-mRNA) kann somit verschiedene Proteine codieren

Quelle: http://de.wikipedia.org/wiki/Alternatives_Spleißen

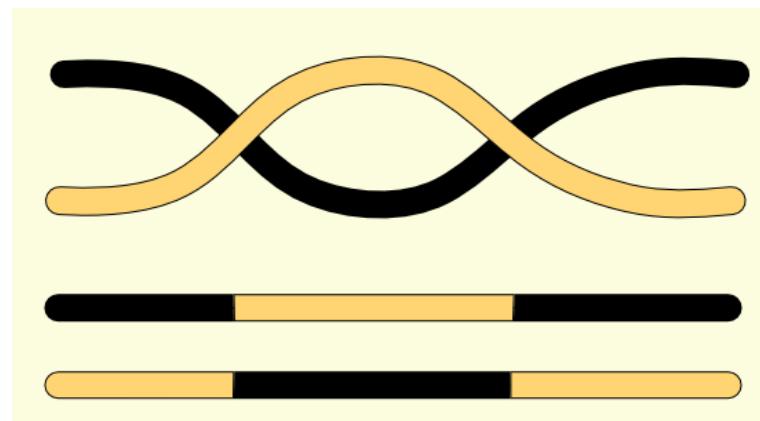
Vererbung

- DNA wird an Kinder vererbt
- Problem:
 - Vater vererbt 46 Chromosomen
 - Mutter vererbt 46 Chromosomen
- Das Kind hätte 92 Chromosomen
- Eizellen und Spermien haben nur 23 Chromosomen
 - **Haploider Chromosomensatz**
 - alle anderen Zellen sind diploid (46 Chromosomen)
- Halbierung der Chromosomenzahl → **Meiose**

Meiose



- Zusätzlich **Crossing-Over**
 - Austausch von DNA-Sequenzen zwischen den homologen Chromosomenpaaren
 - Während der 1. Reifeteilung



Quelle: <http://de.wikipedia.org/wiki/Crossing-over>

Vererbung

- **Gen** DNA-Abschnitt, der die Information zur Herstellung eines Proteins trägt
- **Allel** Zwei oder mehr unterschiedliche Ausbildungsformen eines Gens
- **Genotyp** Gesamtheit der Gene
- **Phänotyp** äußere Erscheinungsbild
- **Dominantes Allele** wirken bei Ausbildung des Phänotyps bestimmend und unterdrücken das **rezessive Allel** in seiner Wirkung

Blutgruppe (Beispiel)

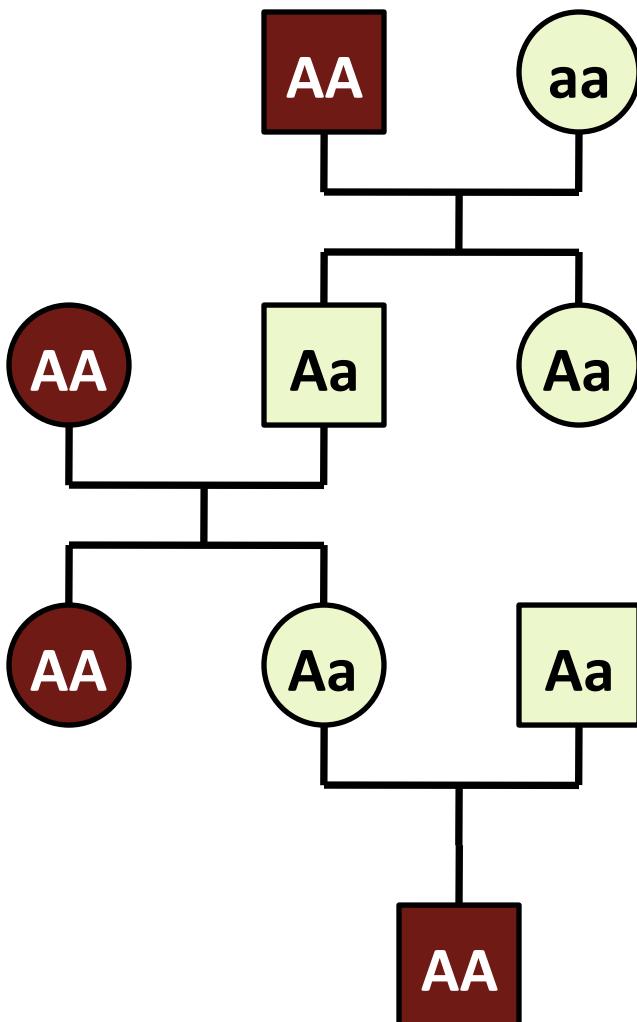
Drei Allele i , i^A und i^B

Blutgruppe	Phänotyp	Genotyp
A	Blutgruppensubstanz A	$i^A i^A$ oder $i^A i$
B	Blutgruppensubstanz B	$i^B i^B$ oder $i^B i$
AB	Blutgruppensubstanz A und Blutgruppensubstanz B	$i^A i^B$
0	Kein Blutgruppensubstanz	ii

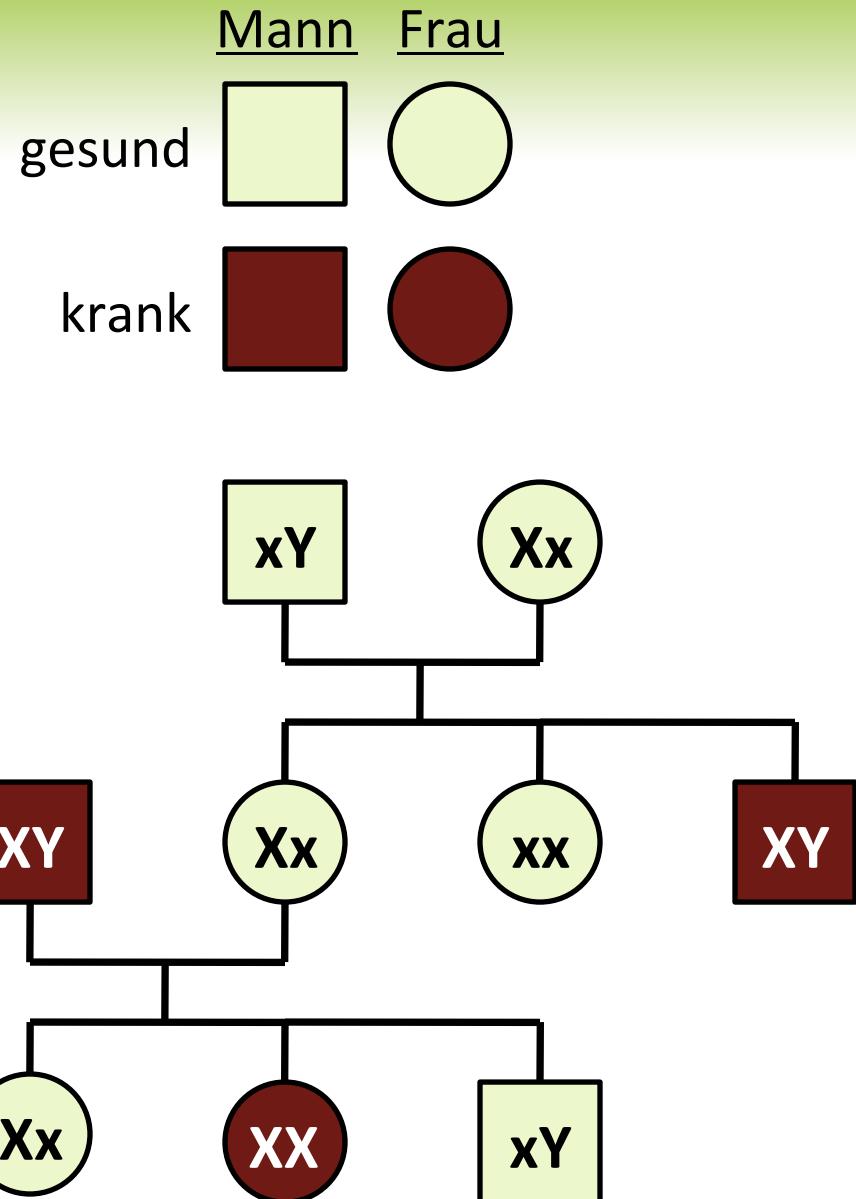
Homozygot = reinerbig $i^A i^A$ $i^B i^B$ ii

Heterozygot = mischerbig $i^A i$ $i^B i$ $i^A i^B$

Stammbäume



rezessiv-autosomal



rezessiv-genosomal

Mutationen

- Genom-Mutationen
 - Veränderung der Anzahl der Chromosomen
 - Mögliche Ursache: Ungleiche Aufteilung der Chromosomen bei der Meiose
 - Beispiel: Trisomie 21 (Down-Syndrom)
- Chromosomen-Mutationen
 - Strukturelle Veränderung eines Chromosoms
 - Mögliche Ursache: Ungleiches Crossing-Over (Meiose)
 - Beispiel: Katzenschrei-Syndrom
- Gen-Mutationen
 - Veränderung eines Gens

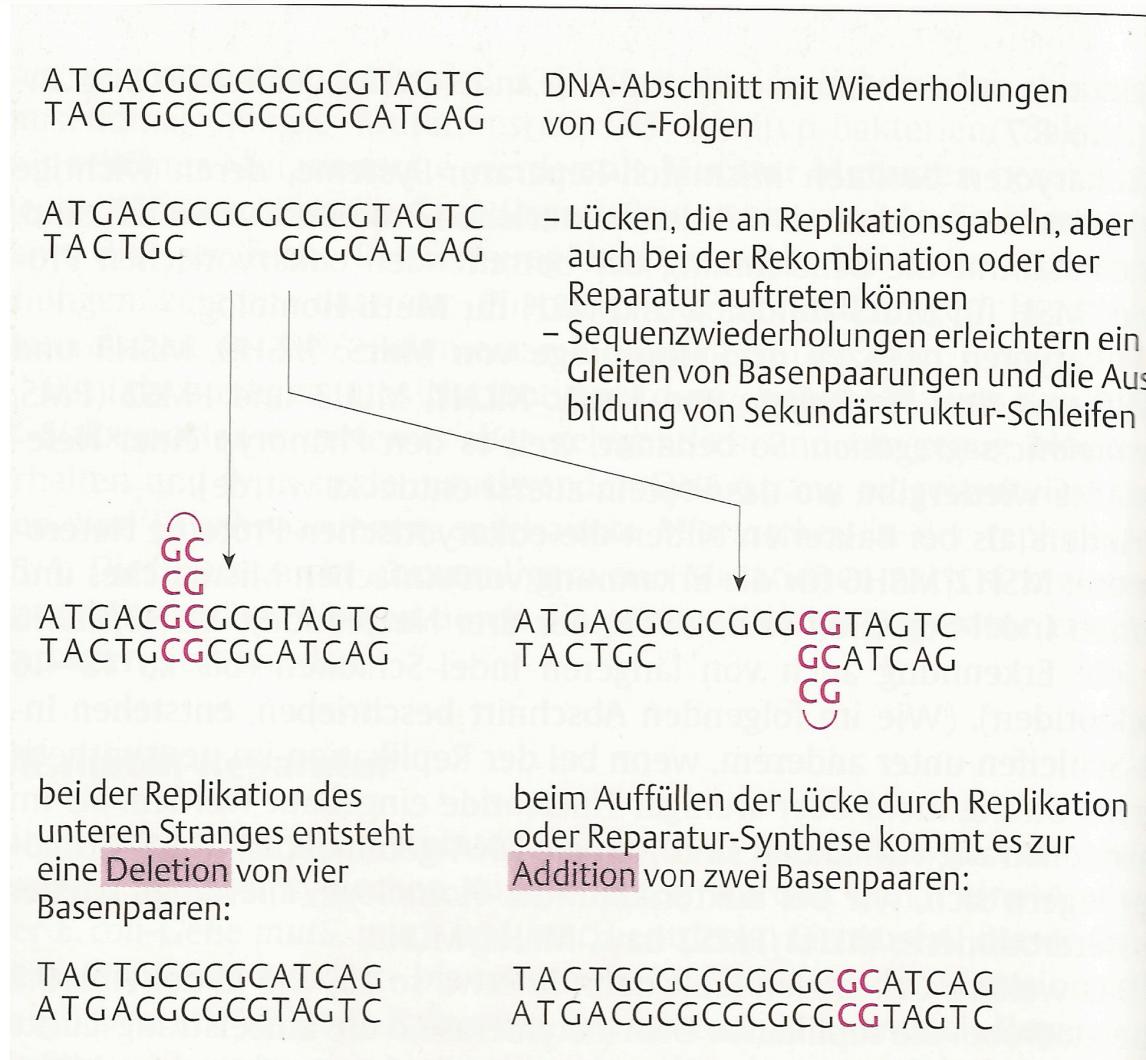
Genvarianten

- Entstehen durch Mutationen
- Verschiedene Arten
 - Indels
 - SNPs
 - Mikrosatelliten-Polymorphismus
 - Wiederholungsfolgen der DNA
 - Hunderttausendfaches Auftreten über das ganze Genom
 - Genetischer Fingerabdruck
 - Entstehen durch Verrutschungen während der Replikation

Genmutation

- Veränderung der Basenpaarsequenz innerhalb eines Gens
 - Nukelotid-Austausch
 - Indels (Insertion / Deletion)
- Änderung der Kodierung von Proteinen
 - Neutral
 - Missense
 - Andere Aminosäure wird kodiert
 - Nonsense
 - Erzeugung eines „Stop-Codon“
 - Terminierung der Synthese
 - Funktionsverlust des Proteins

Beispiel - Leserastermutation



Bildquelle Knippers: Molekulare Genetik 9. Auflage, Thieme 2006, S. 260

Beispiele der Mutationsarten

Ausgangsnukleotid

CTT	AGT	GAC	TAC	CGG	TAA	A	DNA				
GAAT	CACT	GTG	ATG	GCC	CAT	TTT					
CUU	AGU	GAC	UAU	CGG	GUU	AAA	mRNA				
Leu	·	Ser	·	Asp	·	Tyr	·	Gly	·	Lys	Protein

Neutrale Mutation

CTT	AGC	GAC	TAC	CGG	TAA	G	DNA				
GAAT	TCG	CTG	GAT	GCC	CAT	T					
CUU	AGC	GAC	UAC	CGG	GUU	A	mRNA				
Leu	·	Ser	·	Asp	·	Tyr	·	Gly	·	Lys	Protein

Missense-Mutation

CC	T	AGT	GA	A	T	A	CGG	T	A	AA	A	DNA	
GG	A	T	CA	CT	T	A	T	G	CC	C	AT	TT	
CC	U	A	G	UGA	A	U	A	C	GG	U	AA	A	mRNA
Pro	·	Ser	·	Glu	·	Tyr	·	Gly	·	Lys	Protein		

Nonsense-Mutation

CTT	AGT	GAC	TAC	G	GG	T	A	AA	A	DNA	
GAAT	CACT	GTG	ATG	C	CC	C	A	TT	T		
CUU	AGU	GAC	CUA	A	GUU					mRNA	
Stop-Codon											
Leu	·	Ser	·	Asp							Protein

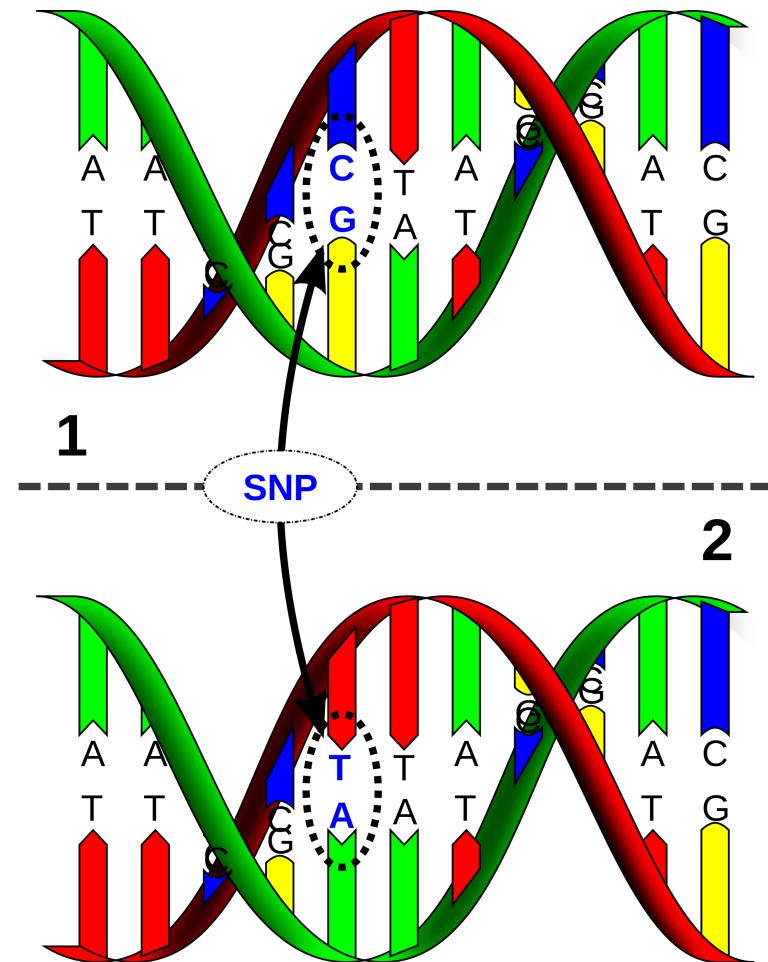
Bildquelle Knippers: Molekulare Genetik 9. Auflage, Thieme 2006, S. 249

Mutationen

- Luria-Delbrück-Experiment
 - Mutationen treten zufällig, spontan und ungerichtet auf
 - Begünstigung von Mutanten durch Umweltbedingungen
- Mutationen in Körperzellen
 - Funktionsverluste oder Tod der Zelle
- Mutationen in Keimzellen
 - keine direkten Konsequenzen für betroffenen Organismus
 - Veränderungen werden bei Nachkommen sichtbar

SNP – Single Nucleotide Polymorphism

- Variation einzelner Basenpaare
- Treten durchschnittlich an jedem 1000. Basenpaar auf
- Hotspots
 - Regionen, an denen SNPs häufiger auftreten
- Geschätzte Gesamtzahl:
10 – 15 Millionen
 - 1999 waren 7000 bekannt



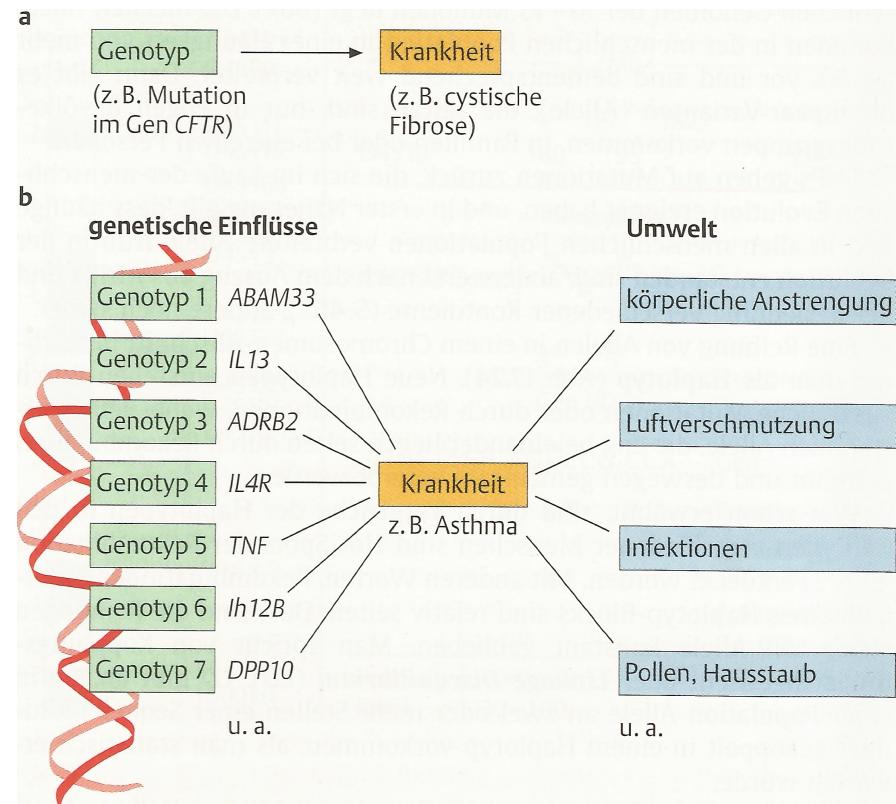
Bildquelle: <http://commons.wikimedia.org/wiki/File:Dna-SNP.svg>

SNP – Single Nucleotide Polymorphism

- sehr stabil
 - 10^{-8} Änderungen pro Nucleotid pro Generation
 - Großteil entstanden als sich Menschen entwickelten
 - Homo sapiens teilt sich keine SNPs mit Primaten
 - 85% aller SNPs können in jeder menschlichen Population vorkommen
- Ursache für Unterschiede zwischen Menschen
 - Haut- und Haarfarbe
 - Körpergröße und –form
 - Verhaltensweisen
- Ursache für Empfänglichkeit für Krankheiten

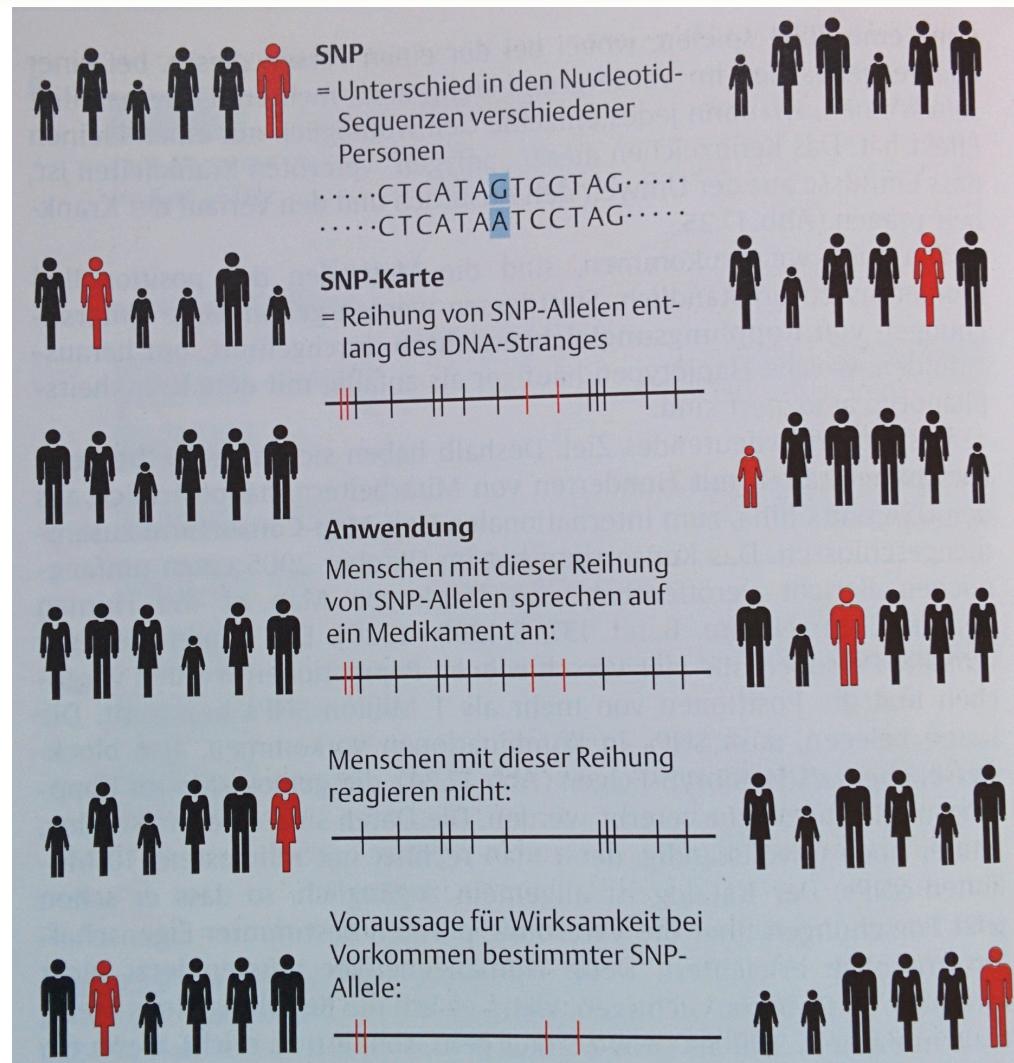
SNP – Single Nucleotide Polymorphism

- Kein direkter Auslöser von Krankheiten
 - Eher Kombinationen von SNPs in Schlüsselgenen und Umweltfaktoren
 - Jedoch existieren monogenetische Krankheiten
- Assoziationsstudien
 - Zusammenhänge zwischen Varianten/SNPs und verschiedenen Krankheiten



Bildquelle Knippers: Molekulare Genetik
9. Auflage, Thieme 2006

SNP in der Pharmakogenetik



Bildquelle Knippers: Molekulare Genetik 9. Auflage, Thieme 2006, S. 249

Abstraktion von DNA -> String

- DNA besteht aus einer Kette von Nukleotiden
- Zeichenkette auf festem Eingabealphabet
 - DNA $\Sigma = \{A, C, G, T\}$
 - RNA $\Sigma = \{A, C, G, U\}$
 - Protein $\Sigma = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$
- Komplementäre Basenpaare
 - A – T
 - G – C
- Beispielsequenz

A C C G T A G C
T G G C A T C G

Sequenzierung

- Bestimmung der Nukleotidfolge in DNA
- Analyse von Organismen und genetisch bedingten Erkrankungen
- Klassische Methoden der 70er Jahre waren aufwendig und konnten nur kleine Abschnitte erfassen
- Moderne Ansätze ab 2000 sind parallel und hocheffizient durchführbar

Methoden

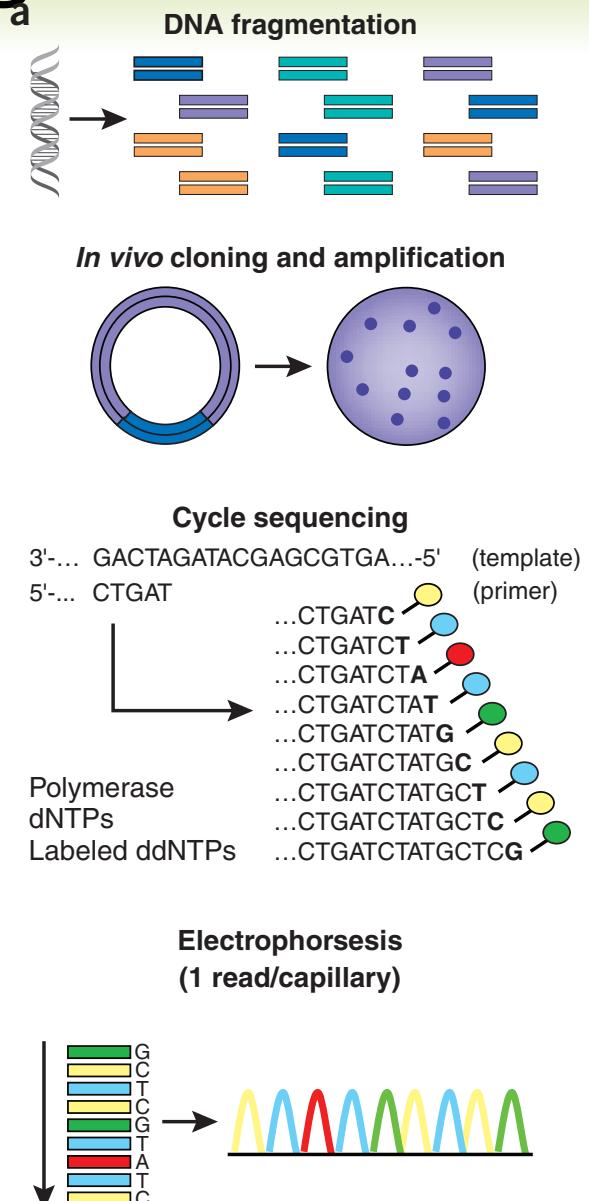
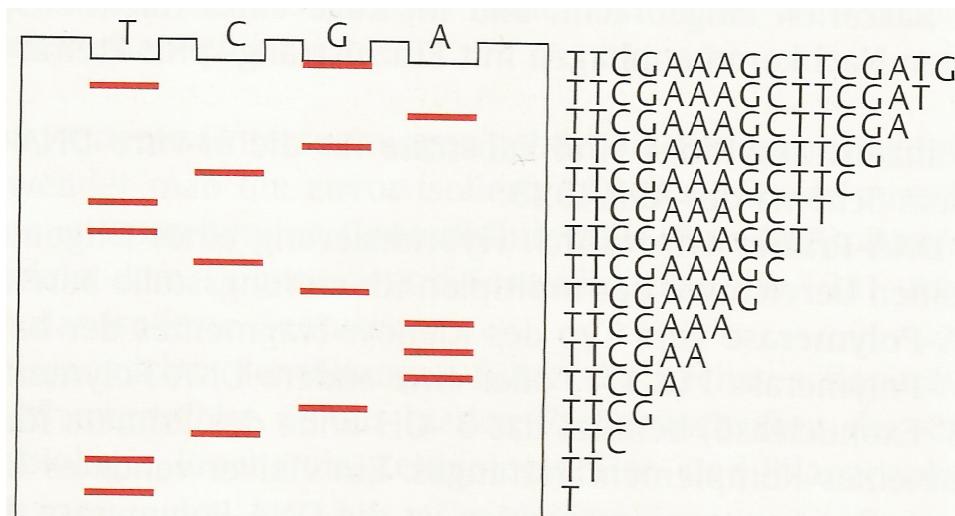
- Klassische Methoden
 - Sanger-Sequenzierung
 - Maxam-Gilbert-Sequenzierung
- Next-Generation Sequencing
 - Sequenzierung durch Hybridisierung
 - Pyrosequenzierung
 - Cyclic Reversible Terminator (CRT)
 - Sequenzierung durch Beobachtung der Bindung
 - SBL und SOLID
 - Echtzeit-Sequenzierung

Sequenzierung nach Sanger

- „Kettenabbruchmethode“
- DNA wird kloniert um einzelstrangigen DNA-Ring zu erhalten
- In jeder Sequenzierungsrounde (cycle sequencing) wird ein radioaktiv markiertes ddNTP hinzugefügt
 - Sorgt für Anhalten der Synthese
 - Nucleoid wird an der Stelle getrennt

Sequenzierung nach Sanger

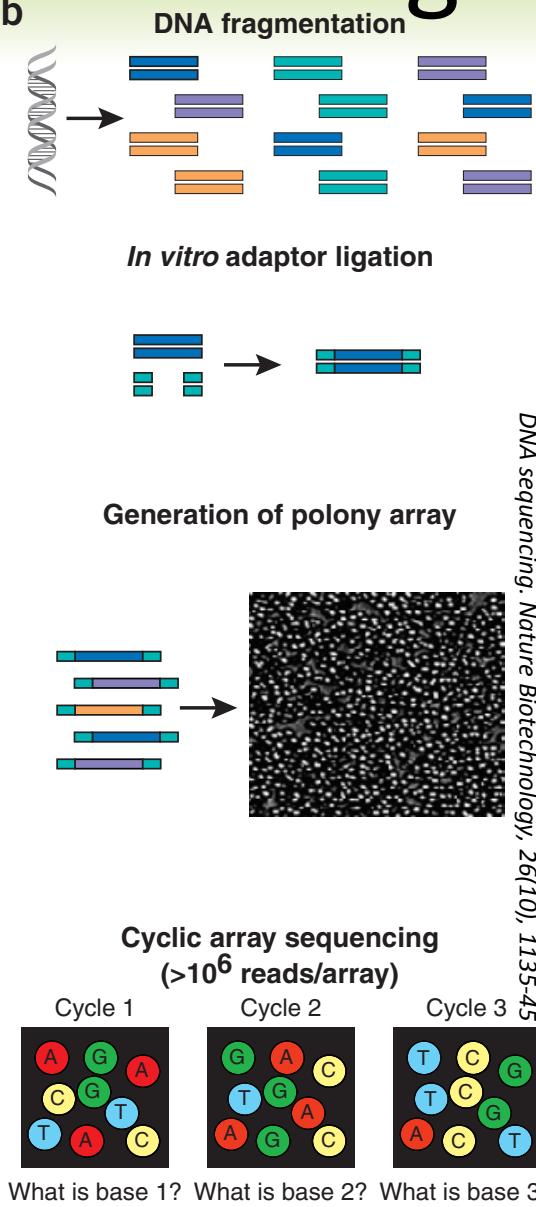
- Sequenz ergibt sich durch Untersuchung der Markierungen
 - Kürzeste & am weitesten gewanderte Nucleoid ist das erste der Sequenz



Bildquelle Knippers: Molekulare Genetik 9. Auflage, Thieme 2006

Sequenzierung durch Hybridisierung

- Konzeptioneller Ablauf:
 - Zufällige Aufteilung der DNA
 - Klonierung durch bspw. PCR
 - DNA wird unbeweglich auf einer Oberfläche befestigt und untersucht
- Vorteile:
 - Parallel Bearbeitung
 - Gleichzeitige enzymatische Behandlung
- Nachteile:
 - Sehr kurze Read-Längen
 - Geringe Genauigkeit

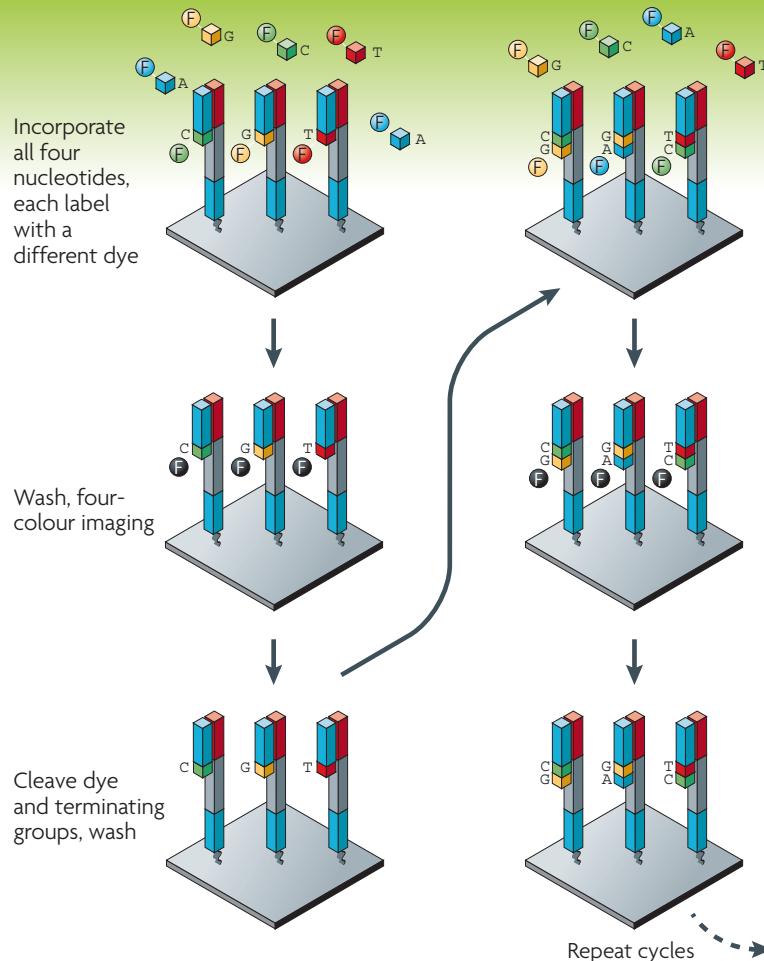
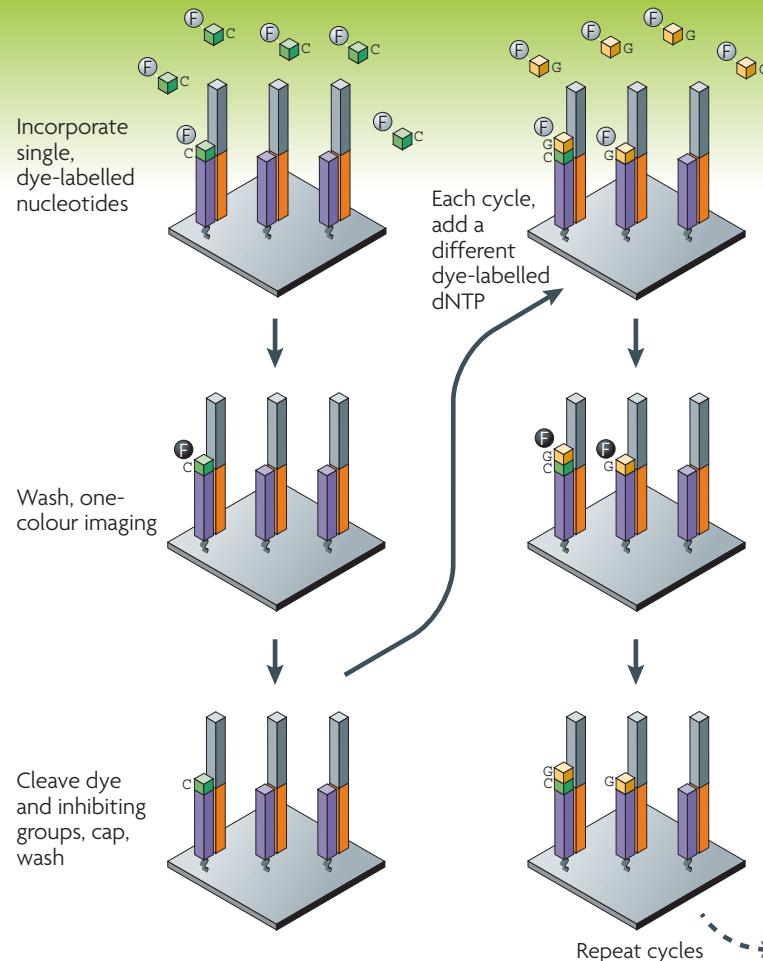
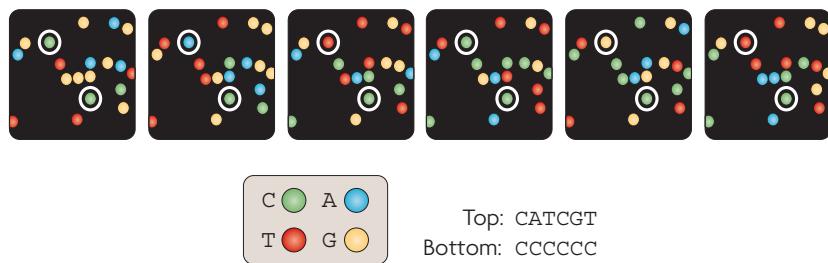


Pyrosequenzierung

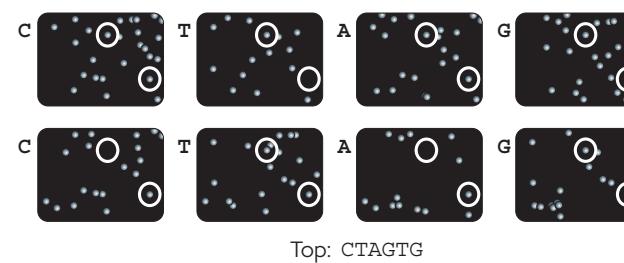
- Analog zu Sanger: Nutzung von DNA-Polymerase
 - Abspaltung von Pyrophosphaten der dNTPs
 - Umwandlung mittels Sulfurylasen zu ATP
- Hinzufügen von Luziferase
 - Nachweis von ATP durch Lichtblitze
 - Messung der Lichtintensität
 - bei zwei Nukleotiden erfolgt ein Lichtblitz mit doppelter Intensität
- Vorteile:
 - Schnelle Laufzeit
- Nachteile:
 - Hohe Fehlerraten

Cyclic reversible termination (CRT)

- Grundidee: Einsatz von reversiblen Terminatoren im Vergleich zur Sanger-Methode
- In jedem Schritt wird ein einzelnes, fluoreszierendes Nukleotid hinzugefügt
 - DNA-Synthese stoppt
 - Restliche Nukleotide werden weggewaschen
 - Visuelle Aufnahme des eingebrachten Nukleotids
 - 4-farbig oder 1-farbig
 - Terminator wird entfernt

a Illumina/Solexa — Reversible terminators**c Helicos BioSciences — Reversible terminators****b**

Top: CATCGT
Bottom: CCCCCC

d

Top: CTAGTG
Bottom: CAGCTA

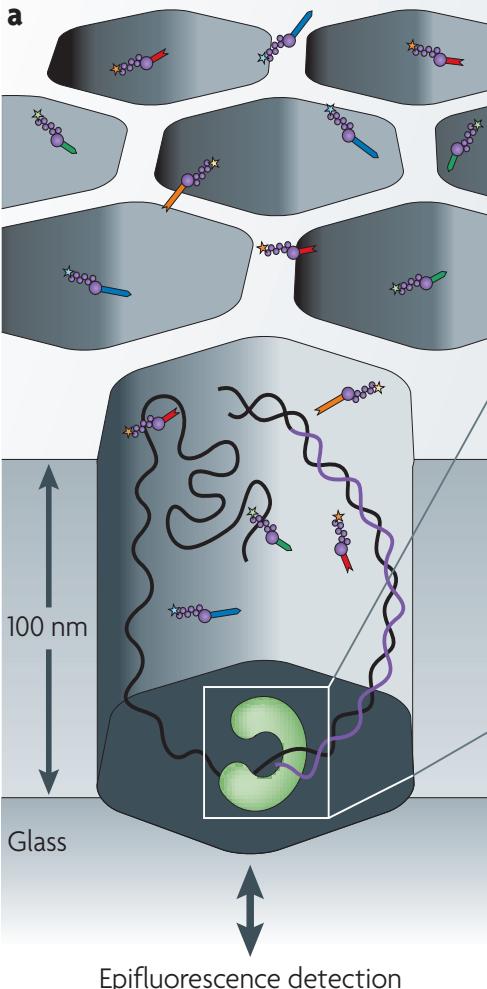
Bildquelle: Metzker, M.L. (2010). Sequencing technologies – the next generation. *Nature Reviews Genetics*, 11(1), 31-46

Echtzeit-Sequenzierung

- Beobachtung der Polymerase
 - Freiwerdende und diffundierende Pyrophosphate senden ein fluoreszierendes Signal
- Aufwendige Vorbehandlung entfällt
- kostengünstig
- Aber sehr fehleranfällig
 - Genauigkeit bei einem Read liegt bei etwa 83%
 - Mehrfaches Sequenzieren führt zu einer Genauigkeit von > 99,9%

Echtzeit-Sequenzierung

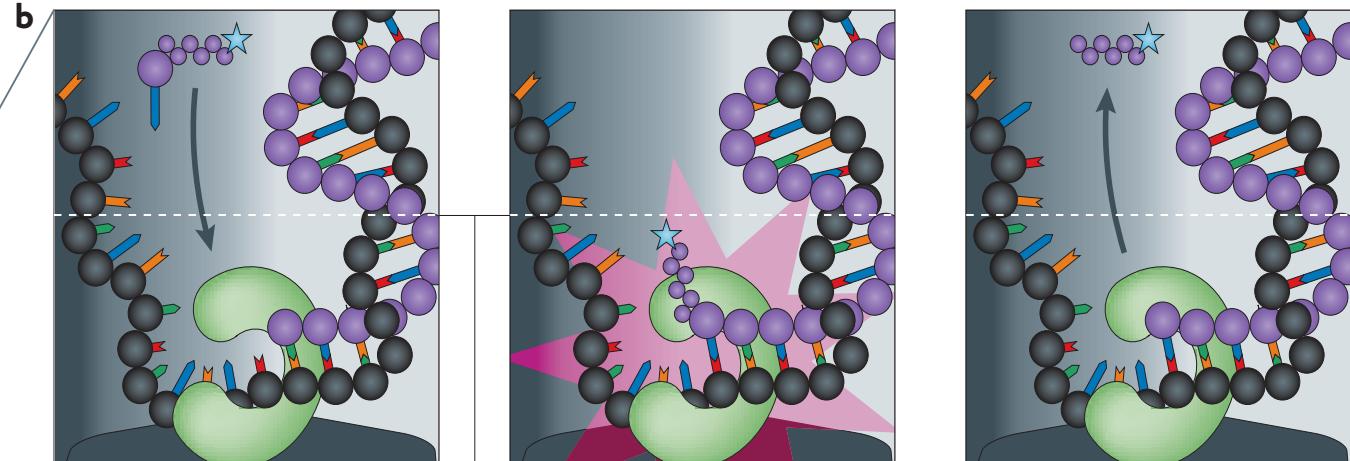
Pacific Biosciences — Real-time sequencing



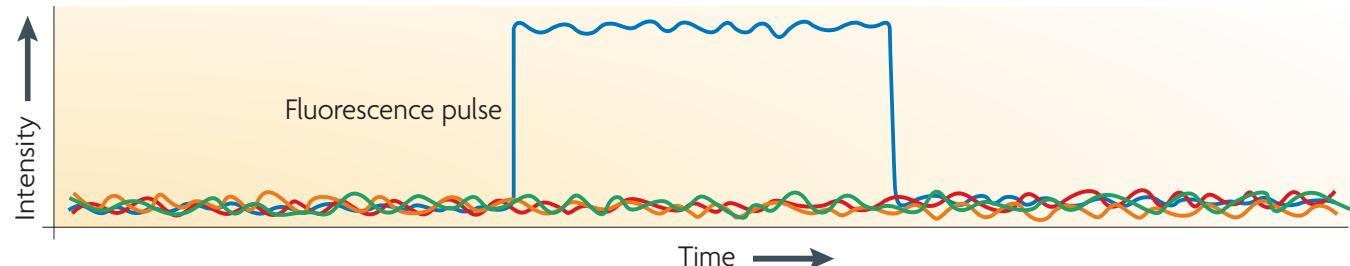
Phospholinked hexaphosphate nucleotides



b



Limit of detection zone



Bildquelle: Metzker, M.L. (2010). Sequencing technologies – the next generation. *Nature Reviews Genetics*, 11(1), 31-46

Dateiformate

- FASTA
- FASTQ
- SAM

FASTA

- Einfaches Dateiformat aus dem FASTA-Projekt
- 1985 entwickelt
- Einfache Struktur
 - Kopfzeile mit ID und Sequenzname
 - Nukleotidsequenz ohne Leerzeichen
 - In 5' -> 3' Richtung

Kodierung

Nukleinbase	Bedeutung	Nukleinbase	Bedeutung
A	Adenin	S (G ∨ C)	Starke Wechselwirkung
C	Cytosin	W (A ∨ T)	Weiche Wechselwirkung
G	Guanin	B (G ∨ T ∨ C)	Nicht A
T	Thymin	D (A ∨ G ∨ T)	Nicht C
U	Uracil	H (A ∨ C ∨ T)	Nicht G
R (G ∨ A)	Purine	V (A ∨ C ∨ G)	Nicht T
Y (T ∨ C)	Pyrimidine	N	Wildcard
K (G ∨ T)	Ketone		
M (A ∨ C)	Aminogruppe		

IUB Nucleotide Codes nach <http://biocorp.ca/IUB.php>

Beispiel des Bakteriums E.coli

```
>gi|556503834|ref|NC_000913.3| Escherichia coli str. K-12 substr.  
MG1655, complete genome  
AGCTTTCACTCTGACTGCAACGGCAATTATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC  
TTCTGAACGGTTACCTGCCGTGAGTAAATTAAAATTATTGACTTAGGTCACTAAATACTTTAACCAA  
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC  
ATTACCACCACCATCACCATTACCAACAGGTAAACGGTGCAGGCTGACGCGTACAGGAAACACAGAAAAAG  
CCCGCACCTGACAGTGCAGGCTTTTCGACCAAAGGTAACGAGGTAACAACCATTGCGAGTGAA  
GTTCGCGGGTACATCAGTGGCAAATGCAGAACGTTCTGCGTGTGCCGATATTCTGGAAAGCAATGCC  
AGGCAGGGGCAGGTGCCACCGTCCTCTGCCCGCAAAATACCAACCACCTGGTGGCGATGATTG  
AAAAAACCATTAGCGGCCAGGATGCTTACCCAATATCAGCGATGCCAACGTATTGCGAACATT  
GACGGGACTCGCCGCCAGCCAGCCGGGTTCCCGCTGGCGCAATTGAAAACCTTCGTCATCAGGAATT  
GCCCAAATAAAACATGTCCTGCATGGCATTAGTTGTTGGGCAGTGCCCGGATAGCATAACGCTGCGC  
TGATTGCGTGGCGAGAAAATGTCGATGCCATTATGCCGGTATTAGAACGCGCGGTACAACGT  
TACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGCATTACCTCGAATCTACCGTCGATATTGCT  
GAGTCCACCCGGCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTTCACCG  
(...)  
TAGAGCAACGAGACACGGCAATGTTGCACCGTTGCTGCATGATATTGAAAAAAATATCACCAAATAAAA  
AACGCCTTAGTAAGTATTTC
```

Quelle: <http://www.ncbi.nlm.nih.gov/nuccore/556503834?report=fasta>

FASTQ

- Erweiterung von FASTA
- Angabe von Erfassungsqualität
 - 1. Zeile enthält ID
 - 2. Zeile enthält Sequenzierung
 - 3. Zeile ist optional und enthält Kommentare
 - 4. Zeile enthält Erfassungsqualität nach ASCII

Beispiel

@EECRH8001A0WUU
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGTAATGCCGTCGCCCGCCTGT
CCGGTGACGATTCCAGCGGCCATGCCACAGGCAATCAGCAGTGGCGCAA
CAGAAATCACGCTCCCCGGCTGTGCTTGCTGGCATGAGGATGAACACGCGA
CGACCAGACGGTGAATTCTGATTGCCAACATAGCTGAAGGCACCCGGCCAC
GGATCGGCAACGGCACGTACCATGTTGTGCAG
+
/ (\$"!!!!!!"; D==<C===
9FB0=C==<<@7A9<<=<9<EA . D==< :
8 ; C=< : : =A9 : <>6 ; >5=<<<= ; ; @8<<=C<= ; =EA / : : =<<<=FB5 & C <
; 5= ; =8FB / ; <<C= ; <=<4C<= ; <@7<< ; == ; ; : <9B : : =79D= ; 8B ; E @ -
=<< : B ; <<3A9= : = ; <=<=A9B ; : =C ?
* D=<3==A : ==8C=<C<7C=<<99=8C<< ; <A9<==<9<

http://orione.crs4.it/library_common/ldda_info?

library_id=9cb534f738e19216&show_deleted=False&cntrller=library&folder_id=9b4cd4a5204a4b3e&use_panels=False&id=23fc845439f64453

Angabe der Erfassungsqualität

- W'keit der korrekten Erfassung der Base
- Codiert durch ASCII

!"#\$%&'()*+,-./0123456789:;<=>?

@ABCDEFGHIJKLMNPQRSTUVWXYZ[\]^_`abcdefghijklmno
pqrstuvwxyz{|}~

- Erfassungsqualität und deren Angabe schwanken je nach Sequenziermethode und -maschine

SAM

- Entwickelt 2009 im SAMTools-Projekt
- Container für erfasste Sequenz und der jeweiligen Reads
- Tab-delimited

header section

- Headerzeilen beginnen mit @
- Allgemeine Informationen über das Template

Zeile	Tags	Beschreibung
@HD	VN	Version des Dateiformats
@SQ	SN	Name der referenzierten Sequenz. Verwendung in RNAME & PNEXT
	LN	Länge der referenzierten Sequenz
@RG	ID	Identifikation der Read-Group
@PG	ID	Identifikation des Programms

alignment section

- Jede Zeile besteht auf 11 Segmenten

Feld	Beschreibung
QNAME	Name des Query Template
FLAG	Bitweise Beschreibung des Reads
RNAME	Name des referenzierten Genoms
POS	Position des Reads im Genoms
MAPQ	Angabe der Qualität der Positionseinordnung
CIGAR	CIGAR-String
RNEXT	Angabe des nächstes Reads ("=" falls RNEXT = RNAME)
PNEXT	Position des nächsten reads
TLEN	Länge des Reads
SEQ	Eigentliche Sequenzierung der BASEN
QUAL	ASCII-Informationen über die Erfassungsqualität

Beispiel

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATAT *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA *
SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC *
SA:Z:ref,9,+,5S6M,30,1;
r001 83 ref 37 30 9M = 7 -39 CAGCGGCAT *
NM:i:1
```