Bio Data Management

Kapitel 1

Motivation und Grundlagen

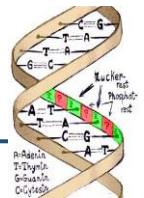
Wintersemester 2014/15

Anika Groß
Universität Leipzig, Institut für Informatik, Abteilung Datenbanken
http://dbs.uni-leipzig.de

UNIVERSITÄT LEIPZIG

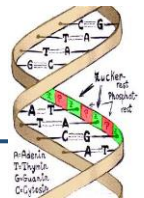
Vorläufiges Inhaltsverzeichnis

1. Motivation und Grundlagen
2. Bio-Datenbanken
3. Datenmodelle und Anfragesprachen
4. Modellierung von Bio-Datenbanken
5. Sequenzierung und Alignments
6. Genexpressionsanalyse
7. Annotationen
8. Ontologiematching in den Lebenswissenschaften
9. Datenintegration: Ansätze und Systeme
10. Versionierung von Datenbeständen
11. Neue Ansätze



Lernziele

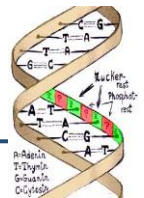
- Überblick zu den wichtigsten biomedizinischen Begriffen und deren Zusammenhang
  - Entitäten: Spezies, Gen, DNA, RNA, Protein
  - Prozesse: Transkription, Translation
- Entstehende Daten und deren Verwendung in der Bioinformatik



Bio-/Lebenswissenschaften

„Erkenntnisgewinn“ über Prozesse oder Strukturen von/in Lebewesen

- Molekularbiologie
- Systembiologie (Bio-)Medizin
- Biophysik, Biochemie
- Bioinformatik
- Spezies, Artenvielfalt
- Translationale Medizin
- …

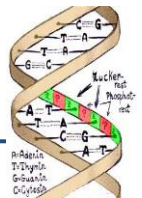


Bioinformatik

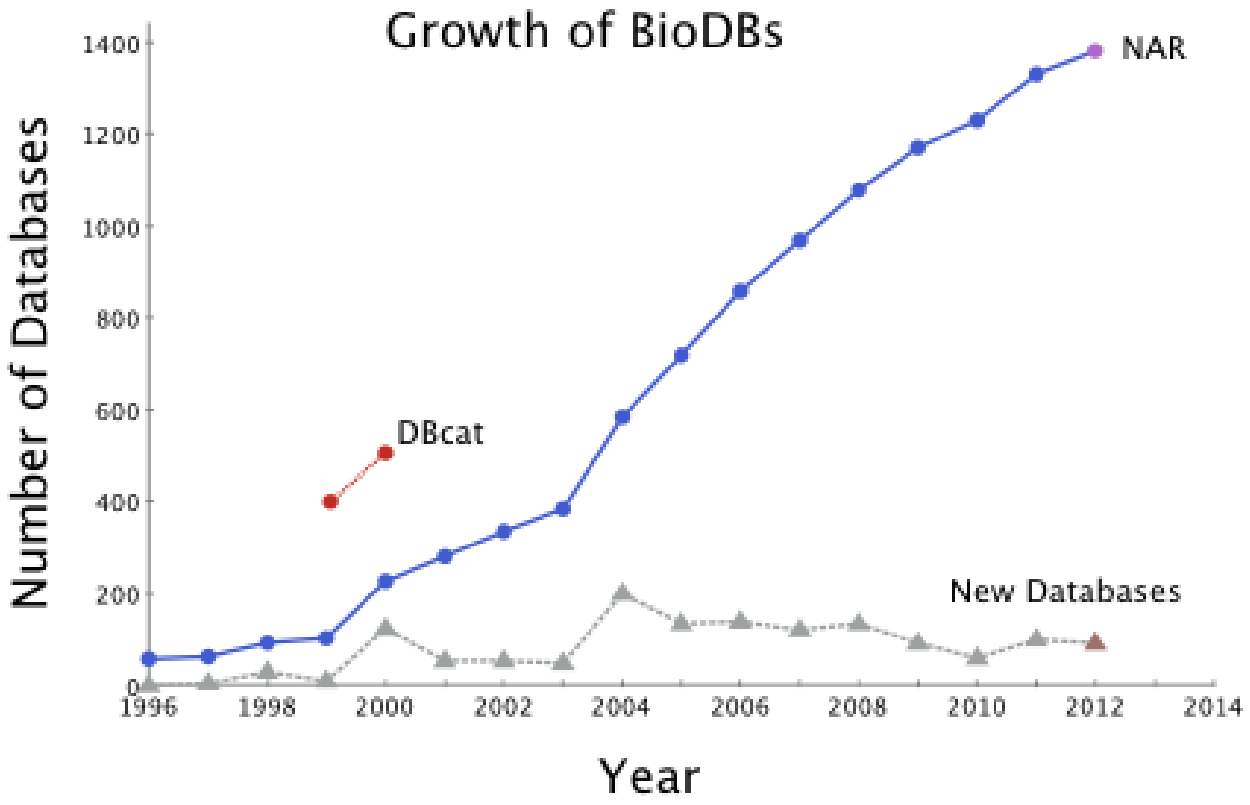
- Sequenzanalyse
- Strukturbioinformatik
- Datenverwaltung

„Die rasch wachsende Menge biologischer Daten, insbesondere DNA- und Proteinsequenzen, deren Kommentierung (die Annotation), 3D-Proteinstrukturen, Interaktionen biologischer Moleküle und Hochdurchsatzdaten von beispielsweise Microarrays stellt besondere Anforderungen an die Handhabung dieser Daten. Ein wichtiges Problem der Bioinformatik besteht daher in der Datenaufbereitung und Speicherung in geeignet indizierten und verknüpften ***biologischen Datenbanken****.*
Die Vorteile liegen dabei in der einheitlichen Struktur, der leichteren Durchsuchbarkeit und der Automatisierbarkeit von Analysen durch Software.“

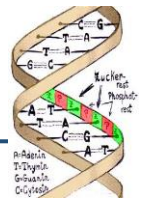
http://de.wikipedia.org/wiki/Bioinformatik



Starkes Wachstum von Bio-Datenbanken

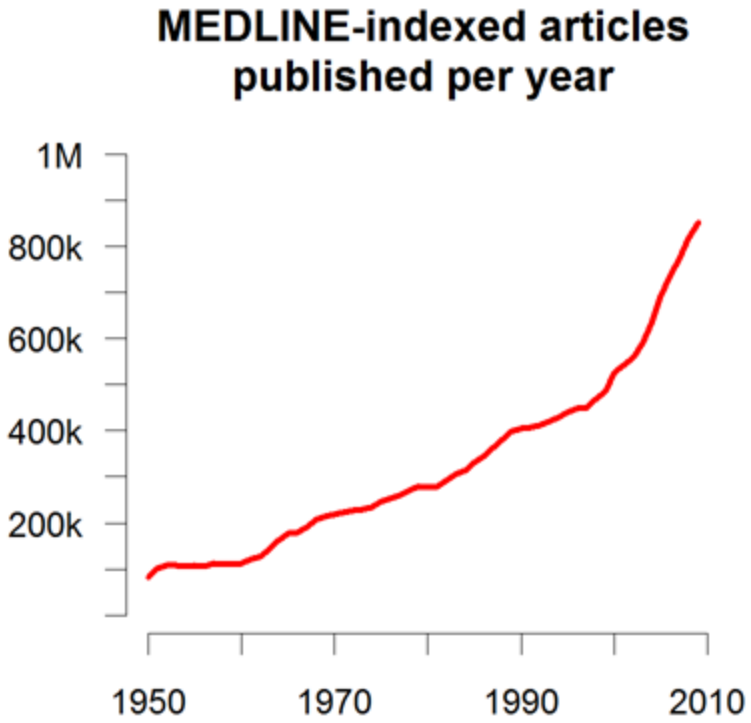


Number of existing (circles) and new databases (triangles) are plotted from 1996 to 2011. New databases are difference between the number of existing databases for each year. DBcat (red) is shown with NAR (blue) counts.



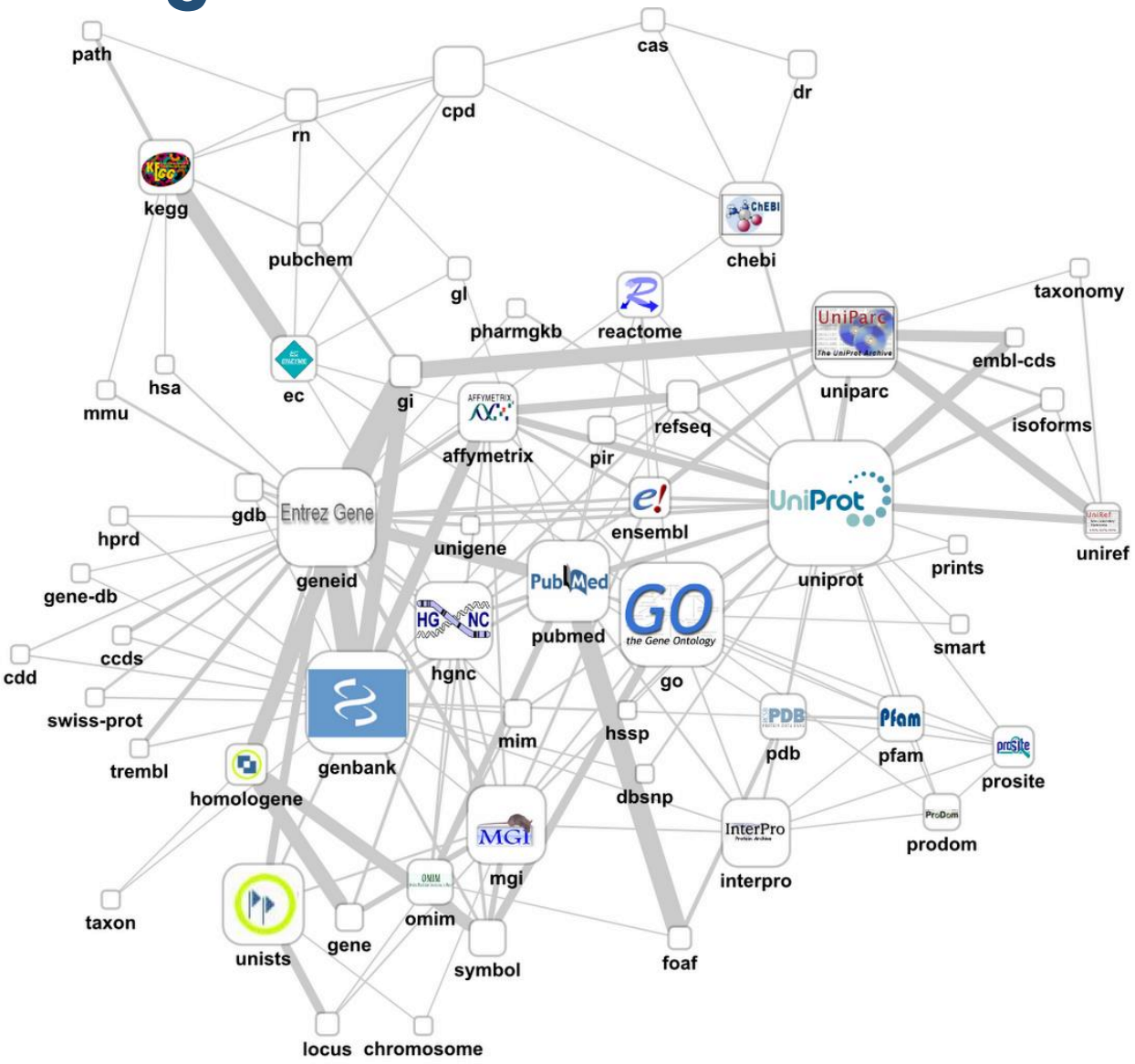
Wachstum von Publikationen

- MEDLINE literature growth chart



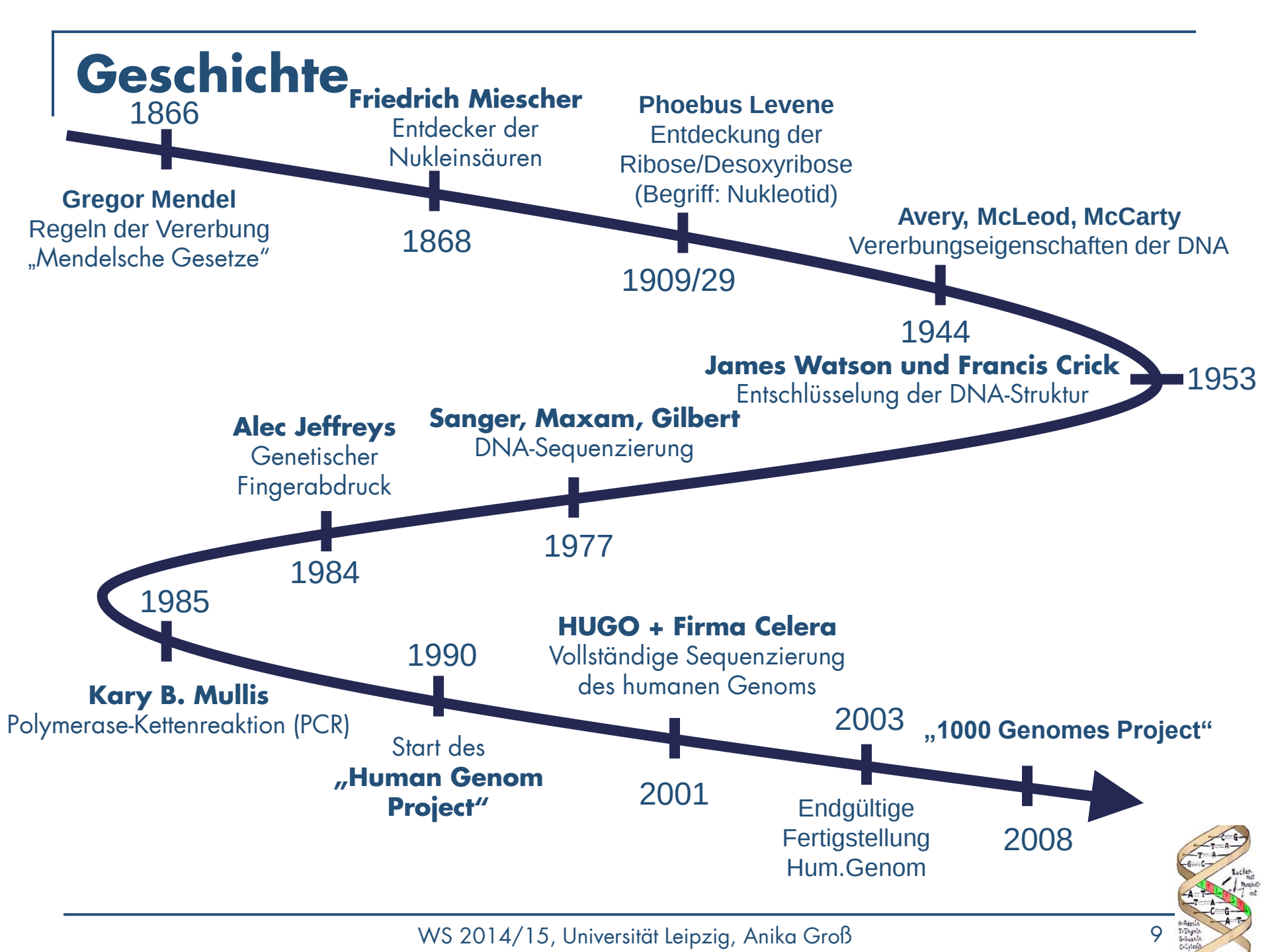
http://jasonpriem.org/2010/10/medline-literature-growth-chart/

Vernetzung von Bio-Datenbanken



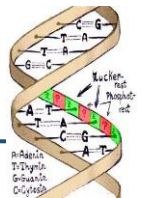
http://www.mquter.qut.edu.au/bio/bio2rdf.jpg
Originalquelle: ttp://bio2rdf.wiki.sourceforge.net/

Geschichte


1866
Gregor Mendel
Regeln der Vererbung
„Mendelsche Gesetze“
1868
Friedrich Miescher
Entdecker der
Nukleinsäuren
1909/29
Phoebus Levene
Entdeckung der
Ribose/Desoxyribose
(Begriff: Nukleotid)
1944
Avery, McLeod, McCarty
Vererbungseigenschaften der DNA
1953
James Watson und Francis Crick
Entschlüsselung der DNA-Struktur
1977
Sanger, Maxam, Gilbert
DNA-Sequenzierung
1984
Alec Jeffreys
Genetischer
Fingerabdruck
1985
Kary B. Mullis
Polymerase-Kettenreaktion (PCR)
1990
Start des
„Human Genom
Project“
2001
HUGO + Firma Celera
Vollständige Sequenzierung
des humanen Genoms
2003
Endgültige
Fertigstellung
Hum.Genom
2008
„1000 Genomes Project“

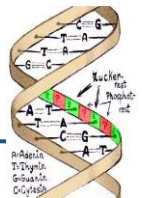
Fragestellungen

- Welche Sequenz hat „mein Gen“? Wo liegt das Gen im Genom?
- Welche Funktionen hat das Gen?
  An welchen Prozessen ist es beteiligt?
- Gibt es homologe Sequenzen im Genom der betrachteten Spezies oder in anderen Spezies?
- Ist das Gen konserviert?
- Wie sieht die Struktur des Genprodukts (Proteins) aus?
- Welche Interaktionen hat das Genprodukt (Protein) mit anderen Proteinen?
- Ist das Gen an einer Krankheit beteiligt?
- In welchen Publikationen finde ich Informationen zu dem Gen?
- …



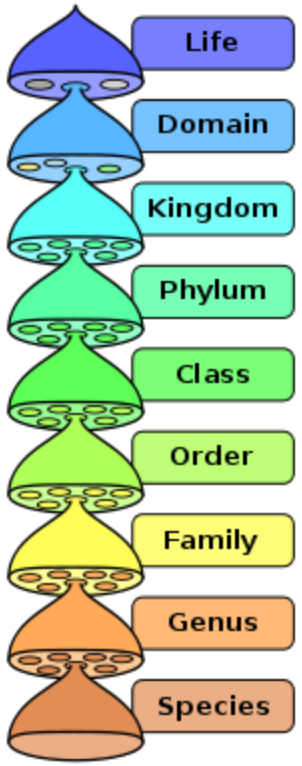
Gliederung

- Spezies und Organismus
- Genom und molekular-biologische Grundlagen
- Proteine
- Transkription und Translation
- Stoffwechsel

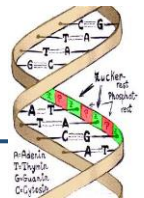


Taxonomie der Spezies

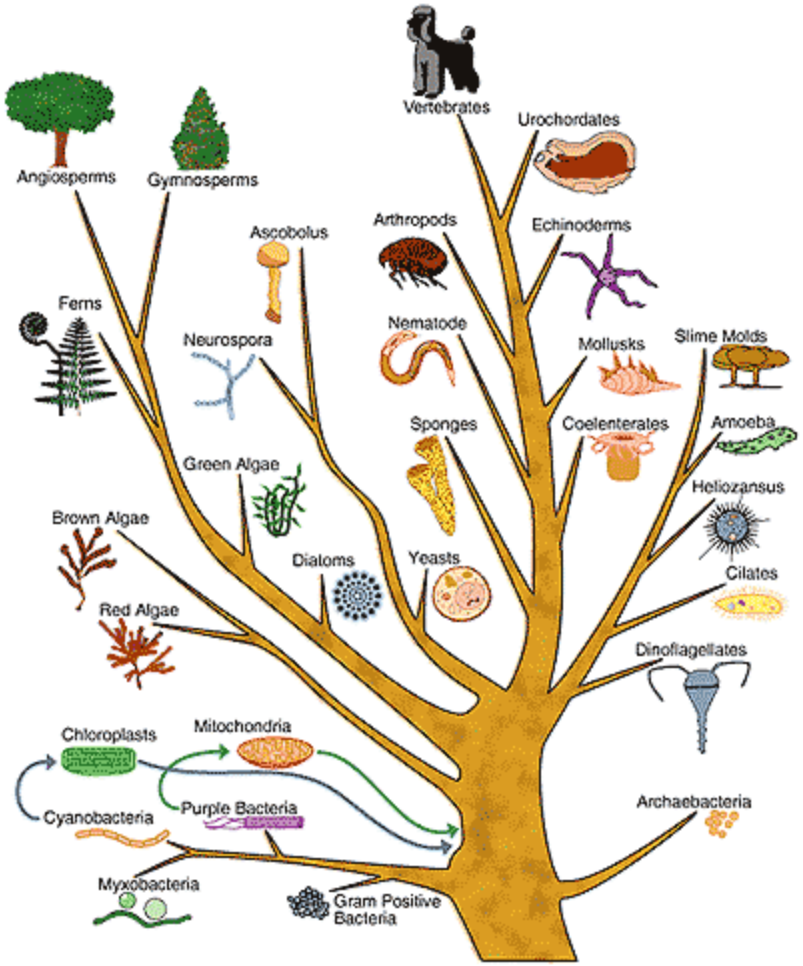
- Verschiedene Definitionen
  - Spezies: Klasse von Organismen, die einen einheitlichen "Genpool" besitzen
  - Fuzzy: "einheitlich"!
  - Exakte Abgrenzung unter kontroverser Diskussion
- Hierarchische Ordnung von Spezies
- Ebenen der Ordnung, spezifischen Begriffen
- Entdeckung und Einordnung
  - Erst Beschreibung, dann Namenszuordnung
  - Heute: Überprüfung anhand genetischer Merkmale → Neuordnung
- Übergroße Anzahl an Spezies
  - 7 – 100 Mio (identifiziert+unidentifiziert)



Bildquelle: Wikipedia

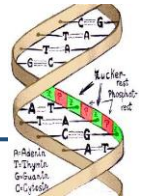


Taxonomie der Spezies



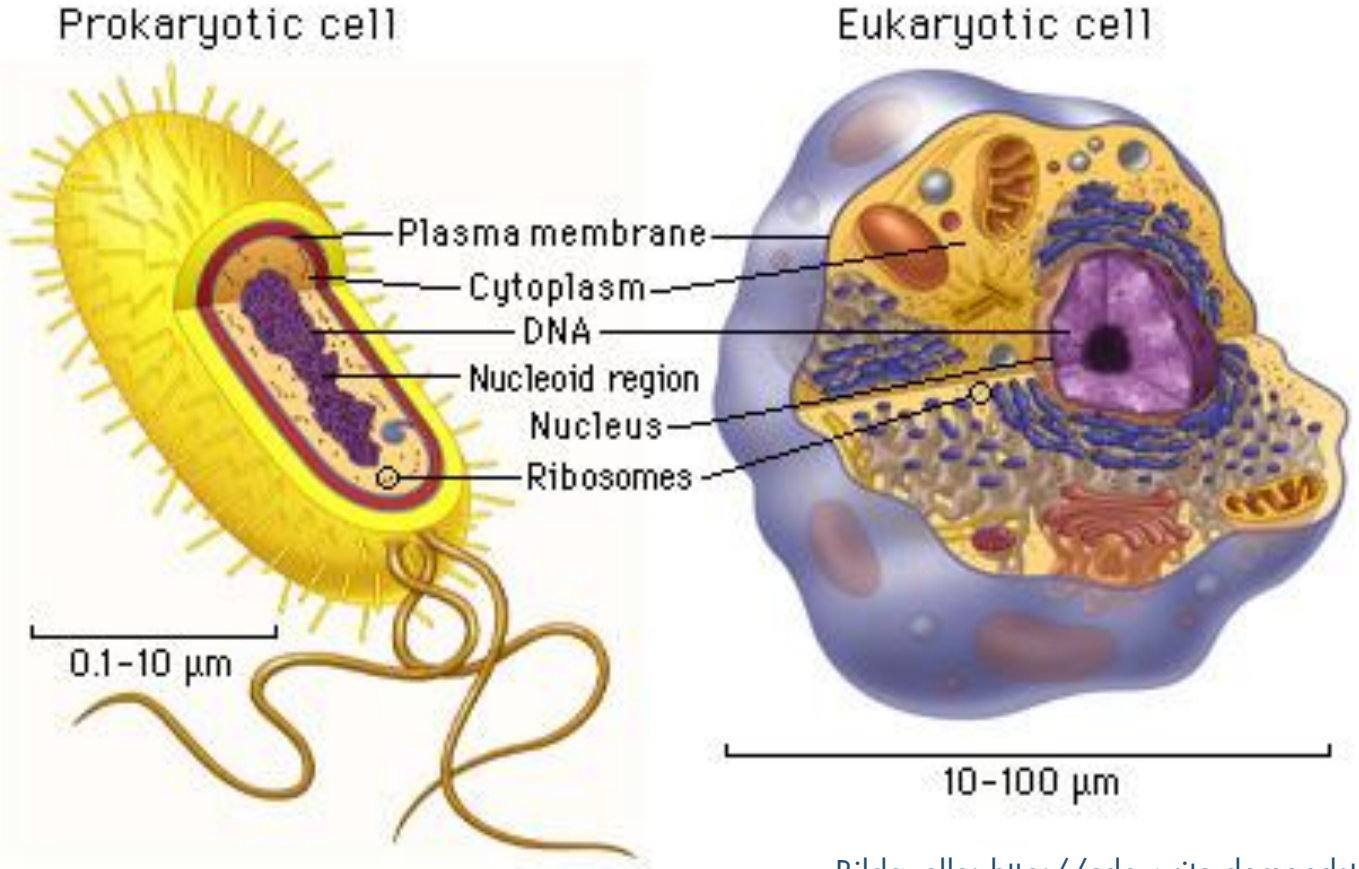
- Systematische Ordnung anhand von Pfaden im Baum
- Innere Knoten = zeitliche Auftrennung einer Spezies in Unterarten
- Länge der Äste = evolutionäre Distanzen

Bildquelle: http://creationwiki.org/pool/images/thumb/0/04/Evolution_tree_of_life.png/300px-Evolution_tree_of_life.png



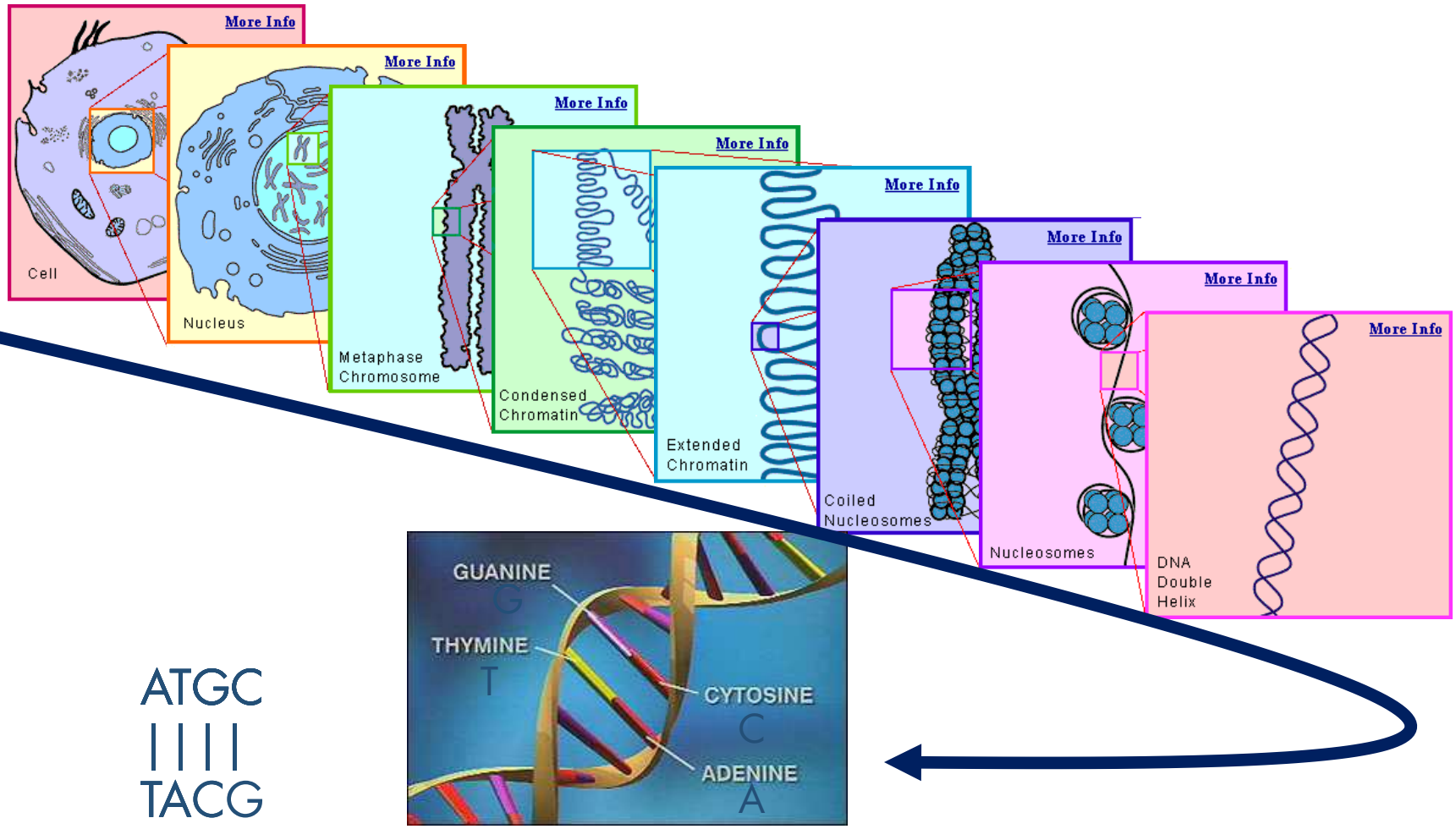
Zellen als kleinste funktionelle Einheit

- Organismus besteht aus Zelle(n)
- Unterteilung in Prokaryoten & Eukaryoten



Bildquelle: http://cdn-write.demandstudios.com

Von der Zelle zur DNA



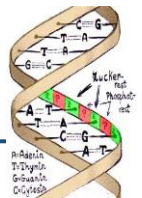
ATGC
||||
TACG

- DNA (***d**eoxyribo**n**ucleic **a**cid*), DNS (**D**esoxyribo**n**uklein**s**äure)
  - Träger der Erbinformation (Gene)
  - Kommt in allen Lebewesen und DNA-Viren vor

Genom

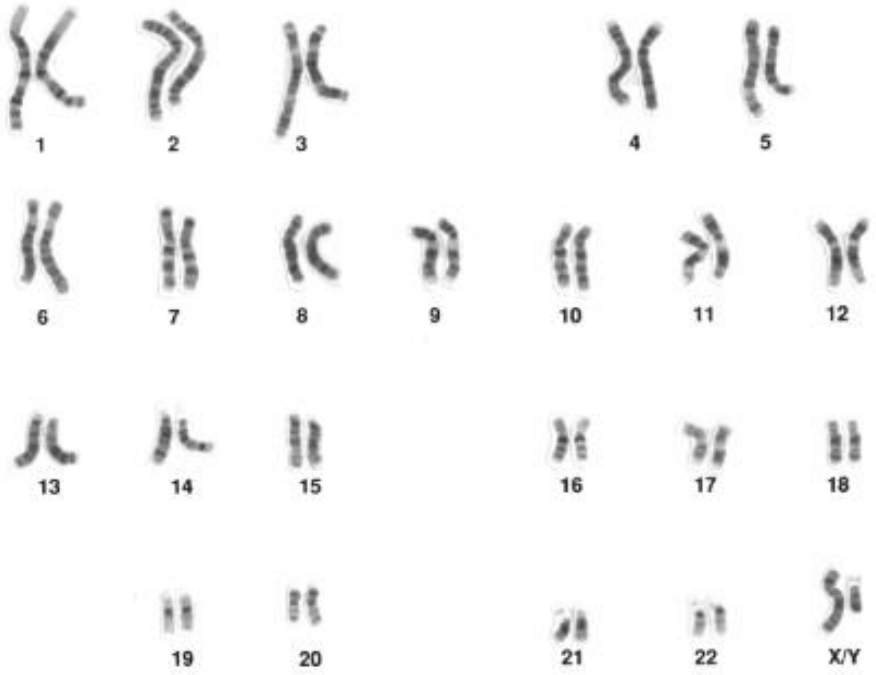
- Erbgut eines Lebewesens oder eines Virus
- Gesamtheit der vererbbaren Information einer Zelle bzw. eines Viruspartikels
- Gesamtheit der Chromosomen, der Gene oder der DNA

- Humanes Genom
  - Ca. drei Milliarden Basenpaare
  - 46 Chromosomen (23 Paare)

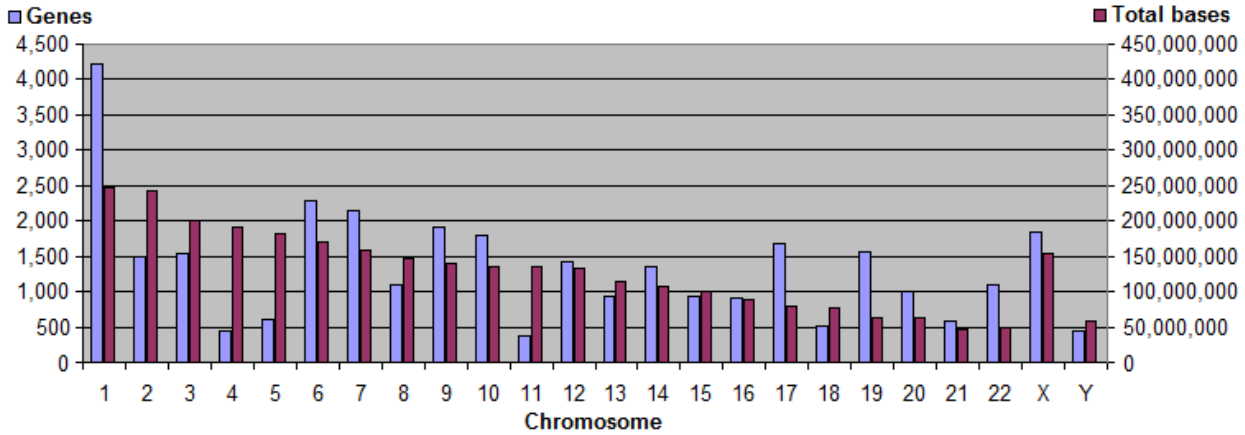


Chromosomen

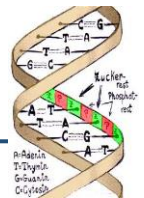
- Strukturen, welche Gene, regulatorische Elemente und andere Nukleotidsequenzen enthalten
- Bestehen aus DNA und Proteinen



Verteilung von Genen und Basenanzahl pro Chromosom (Mensch)

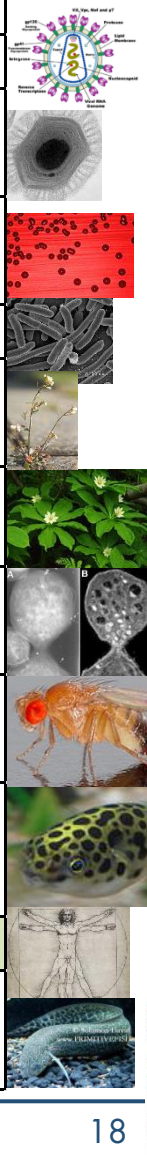


Bildquelle: http://en.wikipedia.org/wiki/Genome



Genome verschiedener Spezies

Organismus Typ	Organismus	Genomgröße in Basenpaaren		Notiz
Virus	HIV	9.749	9,7 kbp	
Virus	Megavirus	1.259.197	1,3 Mbp	Größtes bekanntes virales Genom
Bakterium	Haemophilus influenzae	1.830.000	1,8 Mbp	1.Genom eines lebenden Organismus, Juli 1995
Bakterium	Escherichia coli	4.600.000	4,6 Mbp	
Pflanze	Arabidopsis thaliana	157.000.000	157 Mbp	1. Pflanzen Genom, sequ.im Dezember 2000
Pflanze	Paris japonica (Japanese-native pale-petal)	150.000.000.000	150 Gbp	Größtes bekanntes Pflanzengenom
Hefe	Saccharomyces cerevisiae (Bierhefe, Bäckerhefe)	12.100.000	12,1 Mbp	1. eukaryotisches Genom, sequ.im 1996
Insekt	Drosophila melanogaster (Fruchtfliege)	130.000.000	130 Mbp	
Fisch	Tetraodon nigroviridis (Grüner Kugelfisch)	385.000.000	390 Mbp	Kleinstes Vertebratengenom
Säugetier	**Homo sapiens**	**3.200.000.000**	**3,2 Gbp**	
Fisch	Protopterus aethiopicus (Äthiopischer Lungenfisch)	130.000.000.000	130 Gbp	Größtes bekanntes Vertebratengenom

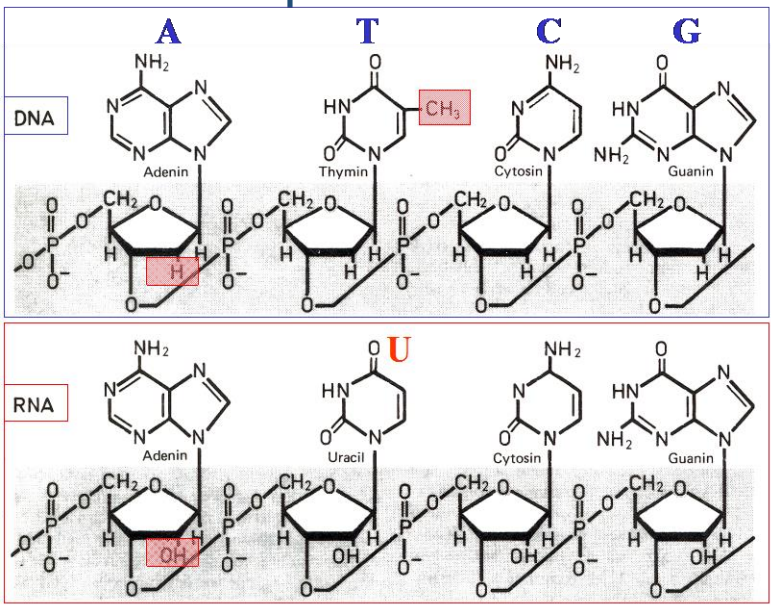


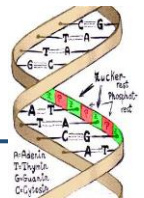
http://en.wikipedia.org/wiki/Genome

Nukleinsäuren (DNA, RNA)

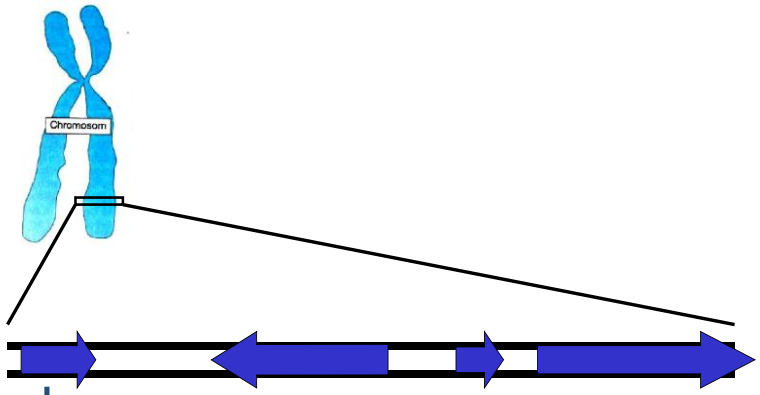
- DNA (DNS): Desoxyribonucleinacid (… säure)
- RNA (RNS): Ribonucleinacid (… säure)
- Endgültige Strukturaufschlüsselung der DNA durch Watson & Crick 1953 (nach Vorarbeiten von Chargaff und Wilkins & Franklin), 1962 Nobelpreis

- Doppelhelix aus Zucker-phosphatrückgrat und Basen (**A**denin, **G**uanin, **C**ytosin, **T**hymin, **U**racil)
- Feste Basenpaare
  - DNA: A-T, G-C
  - RNA: A-U, G-C



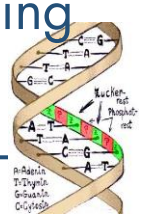


Gen



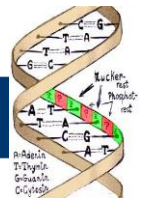
- Keine abschließend geklärte Definition für „Gen“!

- Allgemein: Gene sind Einheiten in der DNA, die Vererbungsinformation tragen
- "locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions“
  [Pearson, 2006, Nature]
- Typische Charakteristik
  - ca. 2.000 - 100.000 Basenpaare lang
  - ca. 50.000 Gene im humanen Genom
  - nur ca. 28% des Genoms beinhalten Gene (also sogenannte Coding Sequence(s) - CDS)



What is a Gene (4)? [GBR+07]

- The same gene?
  - Genes may generate different assemblies (differential splicing)
  - Genes may have interspersed genes
  - Gene duplications in a genome
  - The „same" gene in another organism
  - Mutation of a gene
  - Genes with a different start site
  - Post-translational modifications
- A gene?
  - Pseudo genes (never transcribed, yet highly similar)
  - Non-coding genes
  - miRNA (25 bases!)
- Gene definitions change(d) over centuries, decades, and last years



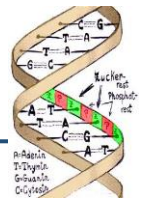
Größe des humanen Genoms

- 23/46 Chromosomen, ca. 20000 Gene,
  ca. 3,2 *10^9 ≈ 3 Milliarden Basenpaare ≈ 3 Gigabasen (Gb)
- Wenn 1 Byte pro Base: 3 Gigabyte
- *Achtung: haploides vs. diploides Genom, binär vs. ASCII (Faktor 8), Informationsgehalt vs. Speicherplatz, Genom + Annotationen deutlich größer, …*
- **Bit** (**b**inary d**i**g**it**): Maßeinheit für den Informationsgehalt
- 1 Bit: 2 Zustände (0/1), 8 Bit=1 Byte
- 1 Base: 4 Zustände (A/T/C/G)

CD ROM	**Menschliches Genom**
700 MB	1359 MB *
3,6 x 10^5 Datenblöcke	2,3 x 10^1 Datenblöcke
Brennen: 6,5 x 10^7 Bits pro Sekunde	DNA-Replikation: 2 x 10^2 Bits pro Sekunde
Block Error Rate: 1 Fehler/3 x 10^5 Bit	Fehlerrate der DNA-Replikation: 1 Fehler/2 x 10^{10} Bits

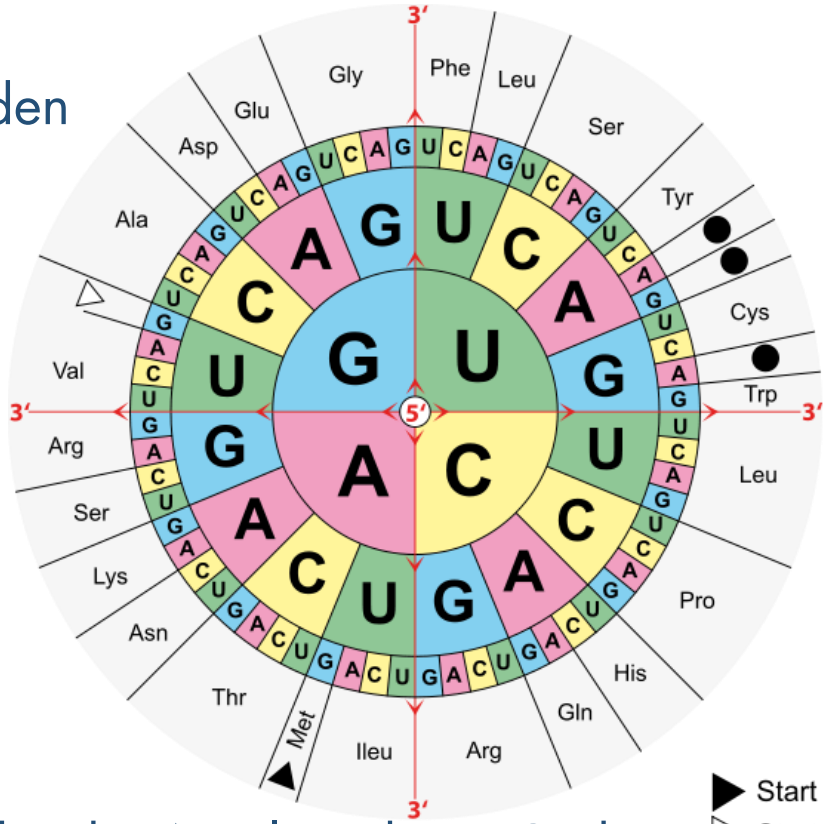
* Reiner Informationsgehalt

Quelle: Joe Dramiga, 17.4.2010, SciLogs: http://www.scilogs.de/wblogs/blog/die-sankore-schriften/medizin/2010-04-17/vergleich-zwischen-einer-cd-rom-und-dem-menschlichen-genom



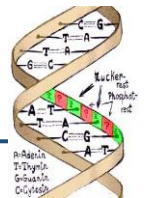
Genetischer Code

- 20 Aminosäuren
- Aminosäuren: Grundlegende Struktureinheiten von Proteinen
- Bildung aus RNA in Translation
- Codon: Abfolge von 3 Nucleotiden (innen→außen)
- Start und Stop-Codons
- Eigenschaften
  - Triplett-Code
  - Universell
  - Degeneriert
  - „Kommalos“
  - Nicht überlappend



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

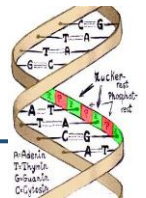
- Anzahl Kombinationen: $4^3=64$
- Stärkere Gewichtung der ersten beiden Nucleotide im Codon
- Unterschiedliche Häufigkeit von Codons pro Aminosäure



AS: Ein- oder Dreibuchstabencode

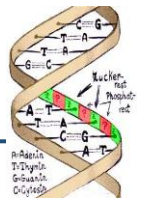
Aminosäure	Dreibuchstabencode	Einbuchstabencode
Alanin	Ala	A
Arginin	Arg	R
Asparagin	Asn	N
Asparaginsäure	Asp	D
Cystein	Cys	C
Glutamin	Gln	Q
Glutaminsäure	Glu	E
Glycin	Gly	G
Histidin	His	H
Isoleucin	Ile	I
Leucin	Leu	L
Lysin	Lys	K
Methionin	Met	M
Phenylalanin	Phe	F
Prolin	Pro	P
Serin	Ser	S
Threonin	Thr	T
Tryptophan	Trp	W
Tyrosin	Tyr	Y
Valin	Val	V

- Verwendung in Aminosäure-sequenzen



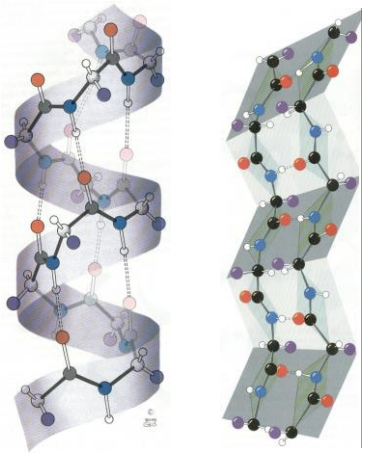
Proteine

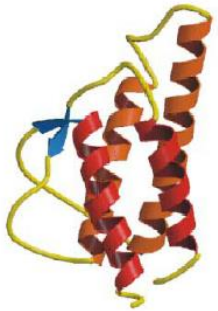
- Zentrale Elemente des Stoffwechsels (als Enzyme)
- Produkt eines oder mehrerer Gene nach Transkription und Translation
- Lineare Abfolge von Aminosäuren
  - Sequenzierung eines Proteins am Stück schwierig (bereits Länge von 20 Aminosäuren nicht-trivial), daher oft Sequenzierung des zugehörigen Gens

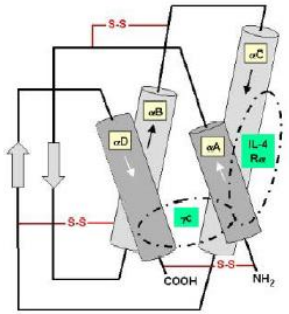


Proteine: Struktur

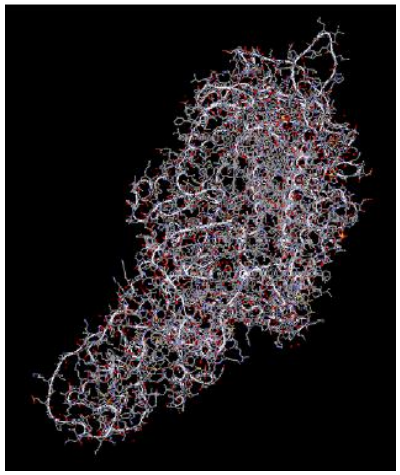
- Verschiedene Strukturebenen: Primär-… Quartärstruktur
  - Primärstruktur: Aminosäuresequenz (1D)
  - Sekundärstruktur: Faltung und Helixbildung in 2D
  - Tertiärstruktur: räumliche Anordnung der Sekundärstruktur
  - Quartärstruktur: Bindung zwischen Proteinen
    → Proteinkomplexen



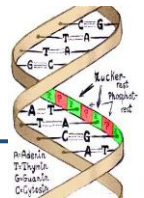




Reinemer/Sebald/Duschl: Angew. Chemie Int. Ed. 39:2834 (2000)

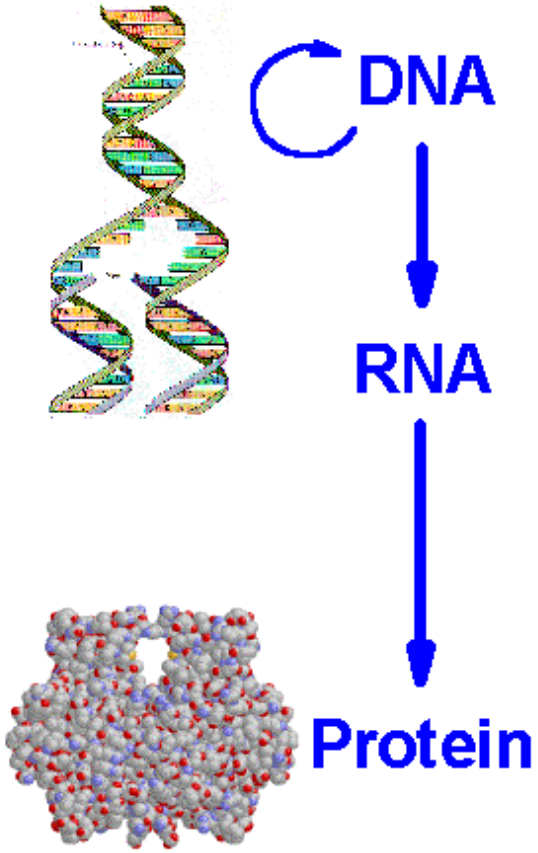


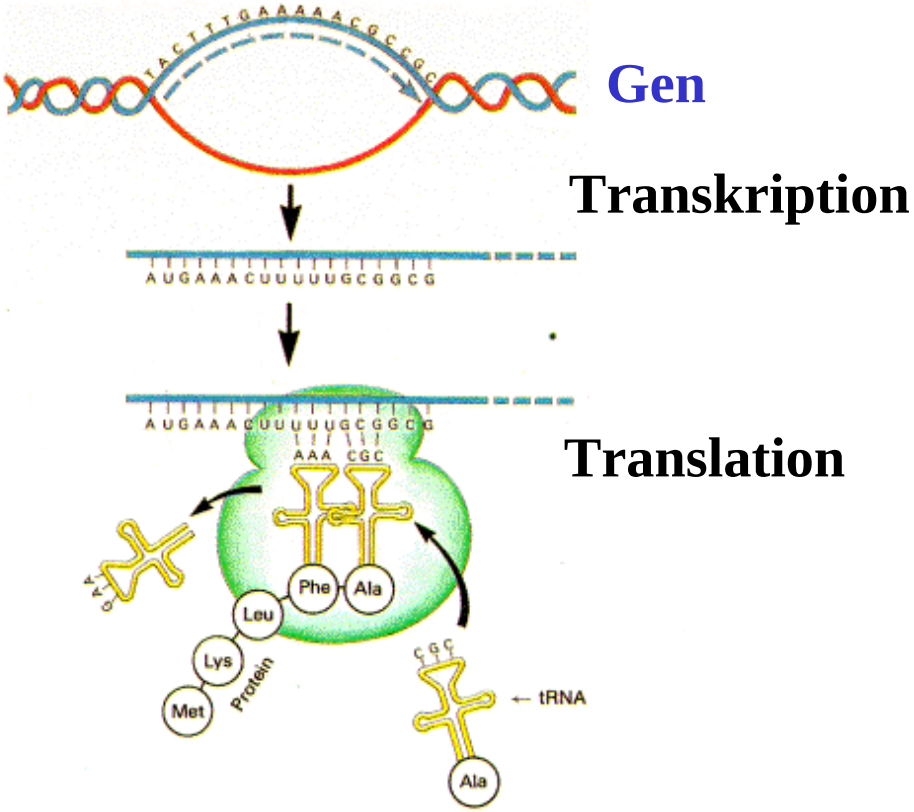
© J. Deisenhofer, O. Epp., K. Miki, R. Huber, H. Michel



Zentrales Dogma der Molekularbiologie

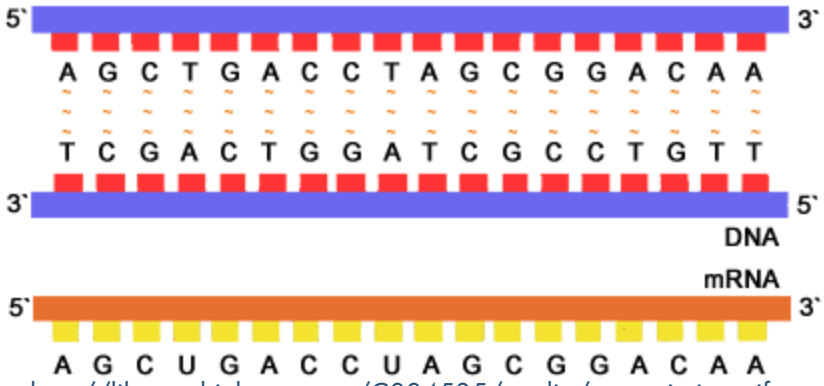
- Proteinbiosynthese → Expression der Gene



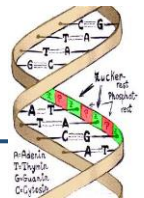


Transkription

- Übertragung („Abschreibung“) eines DNA Abschnitts (Gen) in mRNA (Messenger-RNA)
  - Vorgang im Zellkern
  - Aufspaltung der Doppelhelix
  - Erstellung eines komplementären Strangs
    - Nucleotide A, C, G werden unverändert überschrieben
    - Nutzung des Nucleotids U (Uracil) statt T (Thymin)
    - Zucker: Ribose statt Desoxyribose
  - Durch Enzym RNA-Polymerase (und andere Proteine) katalysiert

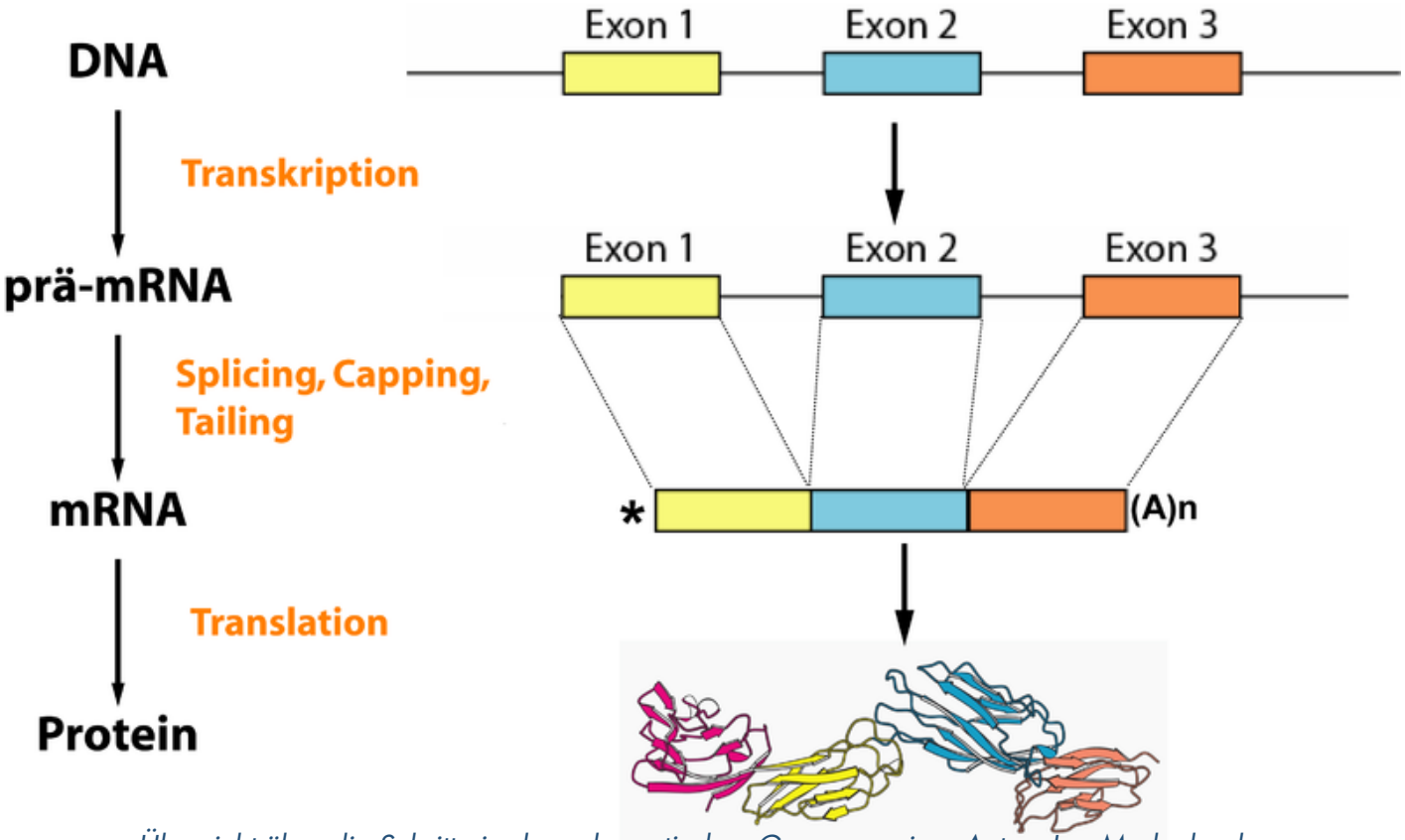


http://library.thinkquest.org/C004535/media/transcription.gif



Proteinbiosynthese - etwas genauer ...

- prä-mRNA (in Transkription gebildet) enthält noch Introns und Exons
- **Splicing**: aus prä-mRNA entsteht die reife mRNA

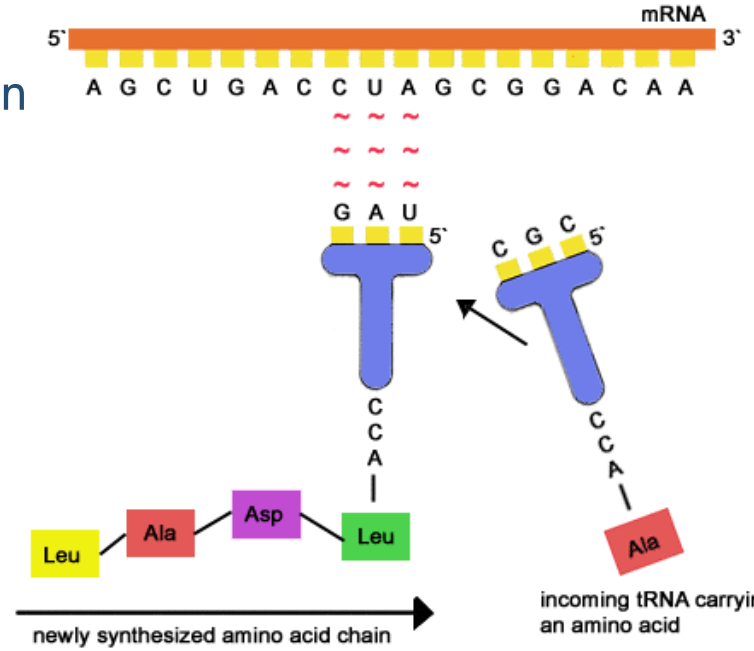


Übersicht über die Schritte in der eukaryotischen Genexpression, Autor: Jan Medenbach
http://commons.wikimedia.org/wiki/File:Eukariotische_Genexpression.png

Translation

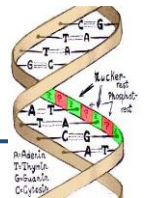
- Übersetzung der Basensequenz der mRNA in die Aminosäuresequenz des Proteins
- Transport vom Zellkern zu Ribosomen

- Drei aufeinander folgende Basen bilden ein Codon (Basentriplett)
- Jedes Codon codiert für eine Aminosäure (genetischer Code)
- Sequenzielle Translation der AS entsprechend der Abfolge der Codons
- Verwendung von Aminosäuren-"Transportern": tRNAs (Transfer-RNAs)



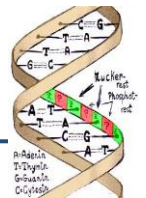
http://library.thinkquest.org/C004535/media/translation.gif

- Co- und Posttranslationale Modifikationen



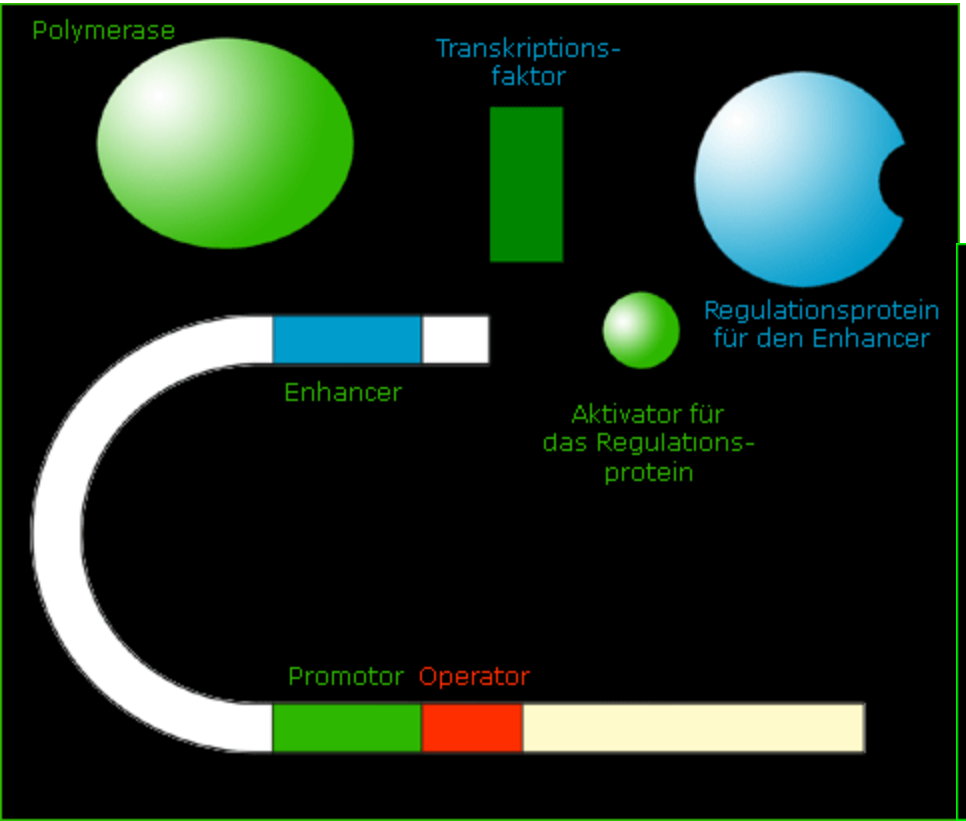
Genregulation

- Genregulation bezeichnet die Steuerung der Aktivität von Genen, d.h. die Steuerung der Genexpression
- Begriff Genexpression (auch Expression oder Exprimierung)
  - Gesamter Prozess des Umsetzens der im Gen enthaltenen Information in das entsprechende Genprodukt
  - Exprimiertes Gen: ausgeprägt, aktiv *(als Protein Teil des Stoffwechsels)*
  - Nicht exprimiertes Gen: abgeschaltet, stummgeschaltet
  - z.B. sind in unterschiedlichen Geweben unterschiedliche Gene exprimiert
- Regulatorische Faktoren können in jedem Schritt der Genexpression wirken
  - Initiation der Transkription
  - Termination der Transkription
  - Capping, Polyadenylierung, Spleißen (bei Eukaryoten)
  - Transport ins Cytoplasma (bei Eukaryoten),
    Stabilität der mRNA im Cytoplasma
  - Initiation der Translation
  - Posttranslationale Modifikationen an synthetisierten Proteinen

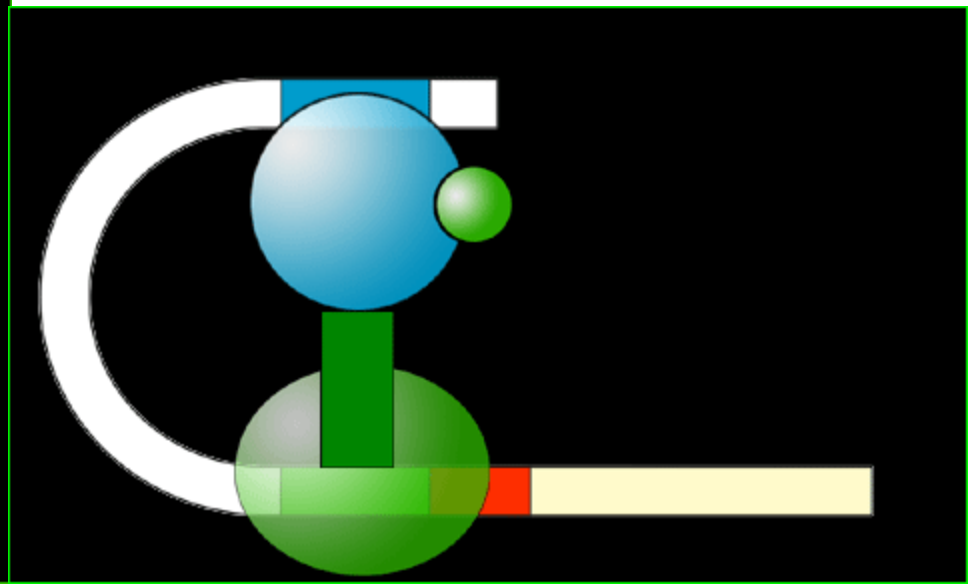


Promotor

- Nukleotid-Sequenz auf der DNA, welche die regulierte Expression eines Gens ermöglicht; Bestandteil eines Gens
- Wechselwirkung mit Transkriptionsfaktoren (oft DNA bindend)
- Transkriptionsfaktoren vermitteln den Start der Transkription des Gens
- Enhancer/Silencer

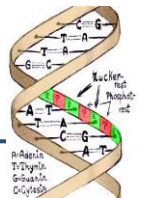


http://www.u-helmich.de/bio/gen/reihe2/25/25-3-s.html



Mutationen

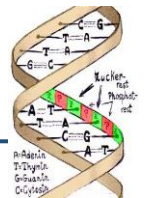
- Unterschiedlichkeit der Organismen innerhalb einer Spezies, z.B. Hauttyp, Haarfarbe
- Einfluss auf die Transkription & Translation und damit auf die Proteinbildung → erblich bedingte Krankheitsmerkmale
- Unterschiedliche Typen: Substitution, Deletion, Insertion
  - Veränderung des Leserahmens bei I & D
- Punktmutation vs. Intervallmutation
  - Single Nucleotide Polymorphism (SNP)
  - Mutation ganzer Bereiche (z.B. Chromosomen-, Genommutation)
- Beispiele
  - Rot-Grün-Blindheit
  - Trisomie 21 (Down-Syndrom)
  - Monosomie X (Turner Syndrom, X0)



Phänotyp vs. Genotyp

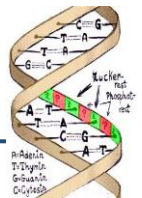
- Stammbaumanalyse: Analyse der auf Verwandtschaft beruhenden genetischen Beziehungen zwischen Individuen
- Phänotyp: beobachtetes Merkmal für einen Organismus (Haarfarbe, Blutgruppe, …)
- Genotyp: einem Phänotyp zugrunde liegende genetische Information
- Allele: unterschiedliche Formen der genetischen Info
  - Diploide Organismen: je Gen 2 Allele
  - Verschiedene Ausprägungsformen, z.B. Blutgruppen
  - heterozygot, homozygot
  - dominant, rezessiv

- Beispiel Sichelzellanämie: homozygot vs. heterozygot



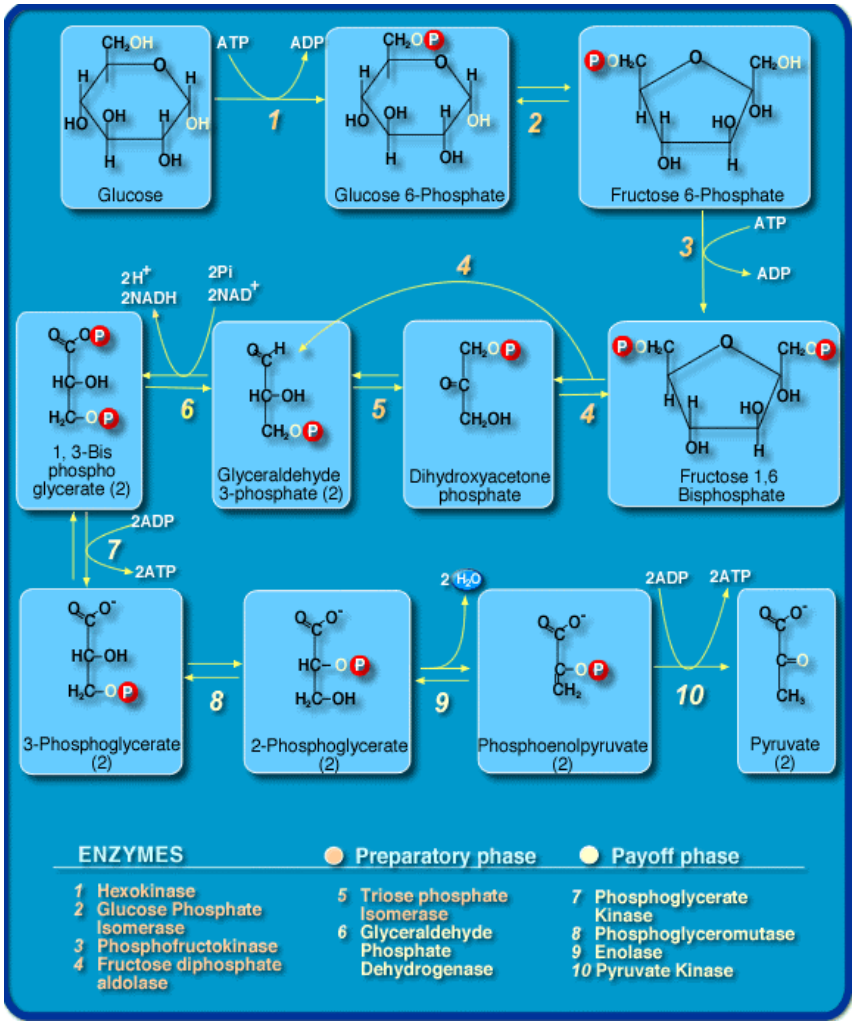
Stoffwechsel

- Gesamtheit aller für einen Organismus notwendigen biochemischen Umwandlungsprozesse
- Hauptsteuerung durch als Enzyme (Katalysatoren) agierende Proteine
- Pathway: Folge von biochemischen Reaktionen (meist einer oder mehreren Funktion(en) im Organismus zugeordnet)
- Grobeinteilung der Pathways in
  - Stoffwechselwege (metabolic pathways)
  - Regulatorische Pfade (regulatory pathways)

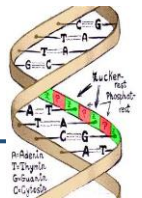


Stoffwechsel: Metabolische Netzwerke

- Metabolismus: Gesamtheit aller lebensnotwendigen biochemischen Vorgänge beim Aufbau, Abbau und Umbau eines Organismus sowie seinem Austausch mit der Umwelt
- 2 grundlegende Stoffwechselvorgänge
  - Assimilation/Anabolismus (z.B. Photosynthese)
  - Dissimilation/Katabolismus (z.B. Zellatmung, Gärung)

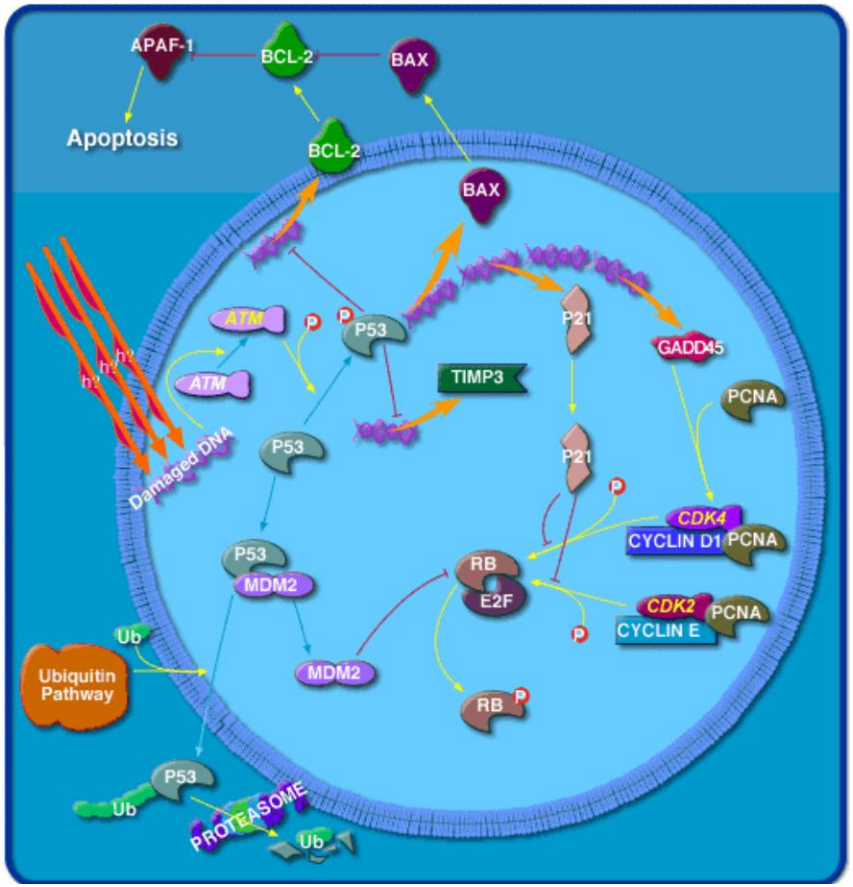


Beispiel Glykolyse



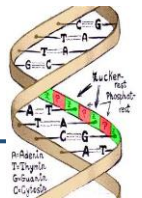
Stoffwechsel: Regulatorische Netzwerke

- Regulation der Genexpression (genetic networks, genetic regulatory pathways)
- Signalwege (signalling pathways, signal-transduction cascades)
- Beispiel: p53-Signalweg
  - Funktion: Terminieren des Zellzyklus im Falle von beschädigter DNA; Apoptose
  - p53 mutiert in fast allen Tumoren vorhanden



Zusammenfassung

- Spezies und Organismus
- Genom, DNA, RNA, Gene, Proteine
- Transkription und Translation
- Stoffwechsel, Pathways



Fragen ?