Kapitel 1: Einführung und biologische Grundlagen

Ziele der Vorlesung

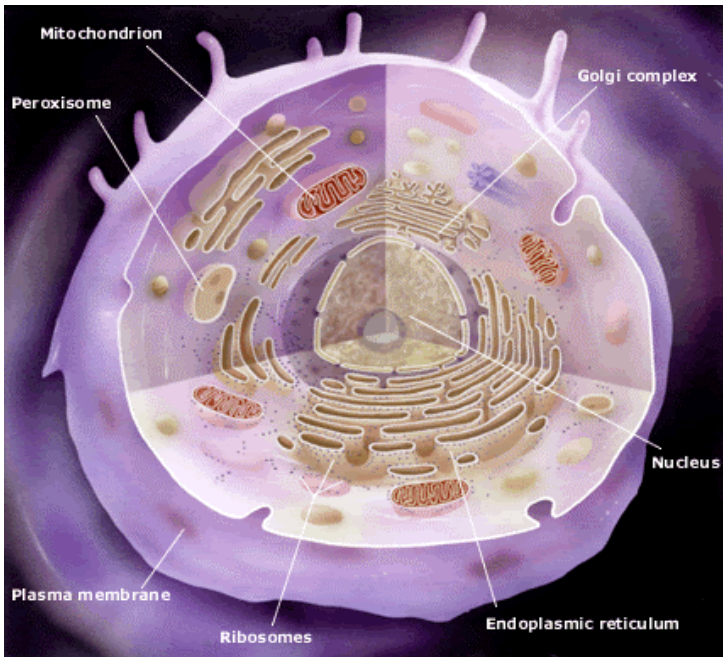
- Grundverständnis wichtiger Verfahren zur Datengewinnung
  - Sequenzierung, Microarrayanalyse, ...
- Klassifizierung von Bio-Datenbanken, Kenntnis typischer Bio-Datenbanken
  - Mapping-, Sequenz-, Protein-, Stoffwechsel-, Publikations-Datenbanken
  - Semantik und Qualität der Daten, Modelle, Zugriffsmethoden, Verwendung
- Kenntnis wichtiger Datenbank-Technologien und ihrer Anwendung auf Bio-Datenbanken
  - Datenmodellierung, Datenbankintegration in der Bioinformatik
  - Datenretrieval, Datenverarbeitung, Data Mining

Literatur und verwendete Materialien

Literatur			
Autoren	**Titel**	**Verlag**	**Jahr**
St. I. Letovsky	Bioinformatics - Database and Systems	Kluwer	2001
Z. Lacroix, T. Critchlow	Bioinformatics: Managing Scientific Data	Morgan Kaufmann	2003
David W. Mount	Bioinformatics: Sequence and Genome Analysis	Cold Spring Harbor Laba-tory Press	2001
Pavel A. Pevzner	Computational Molecular Biology: An Algorithmic Approach	MIT Press	2000
Michael S. Waterman	Introduction to Computational Biology: Maps, Sequences and Geno-mes	CRC Press	1995

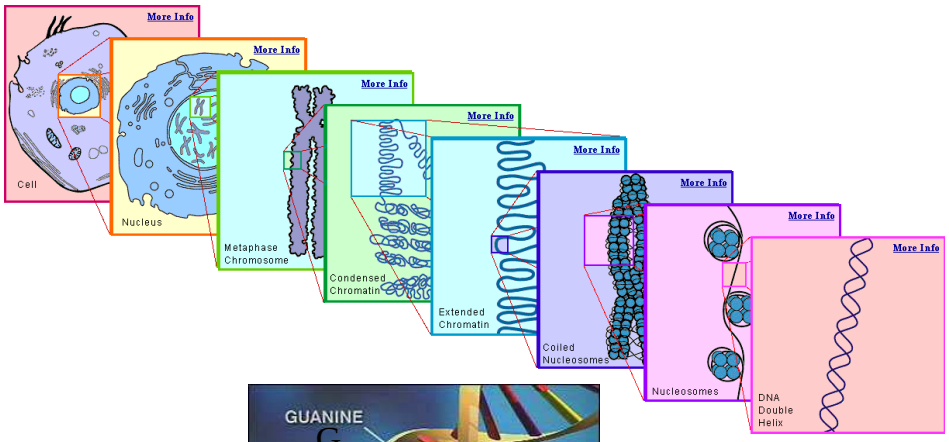
Verwendete Vorlesungsmaterialien u.a.	
Autoren	**Titel / Webadresse**
Prof. Ulf Leser (HU Berlin)	Molekularbiologische Datenbanken (http://www.informatik.hu-berlin.de/wbi/teaching/sose03/mdb/index.html)
Prof. Johann Chr. Freytag (HU Berlin)	Bioinformatik (http://www.dbis.informatik.hu-berlin.de/%7Edbis/lehre/WS0203/BioInformatik/index.html)

Zellaufbau (Eukaryonten)

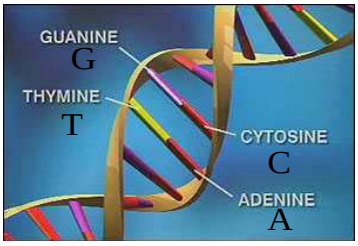


- Prokaryonten (z.B. Bakterien): Kein Zell-kern

Genom

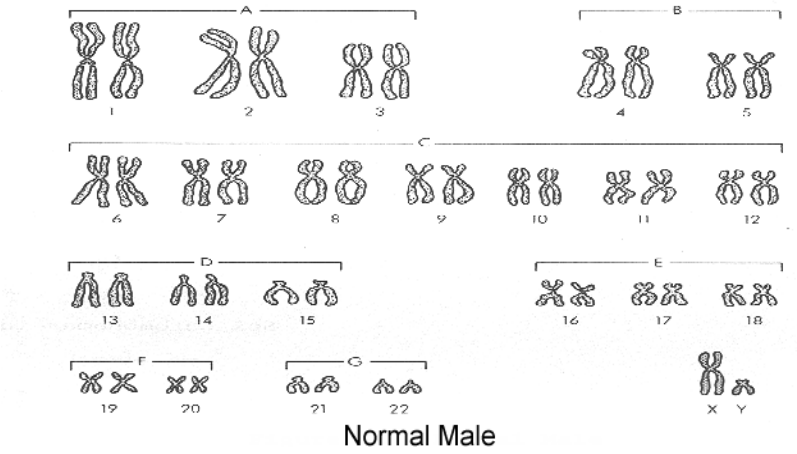


ATGC
||||
TACG



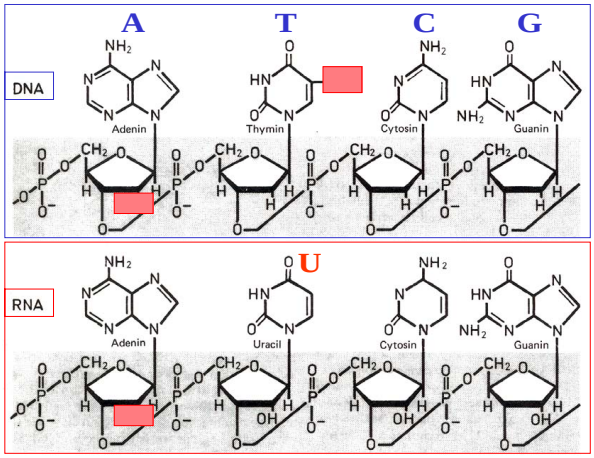
Genom: Chromosomen

- 46 menschliche Chromosomen
- Zusammen circa 3 Milliarden Basenpaare



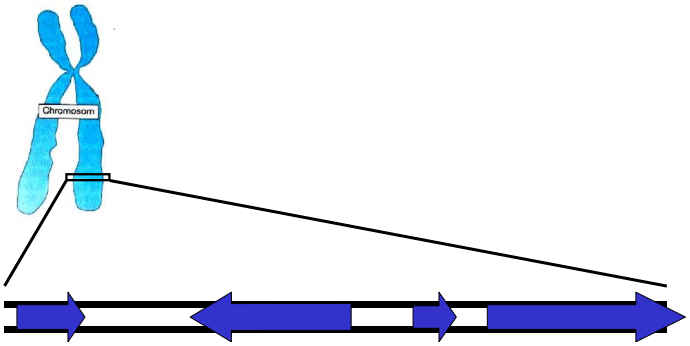
Genom: Nukleinsäuren (DNA, RNA)

- 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
- Feste Basenpaare:
  - DNA: A-T, G-C
  - RNA: A-U (U= Uracil), G-C
- Universaler Codierungs-Mechanismus in allen Spezies



Gen

- Gene sind die Funktionseinheiten in der DNA
- Gen: Ein Abschnitt der DNA, der für ein Protein kodiert
  - 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)



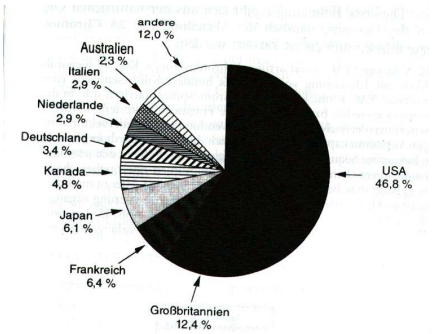
(C) Prof. R. Müller, Prof. E. Rahm

Genom: Sequenzierung

- Sequenzierung: Bestimmung der Reihenfolge der Basen in den Doppelsträngen der DNA-Moleküle

- Wegen Basenkomplementarität genügt es *einen* der beiden komplementären Stränge (Texte) zu bestimmen
- 2 Sequenzierungsprojekte
  - Human Genom Projekt (HGP, Hugo; öffentlich gefördert; multinational), Abschluss 2003 (www.genome.gov)
  - Celera Genomics (kommerziell), http://www.celera.com

HGP: Beteiligte Staaten



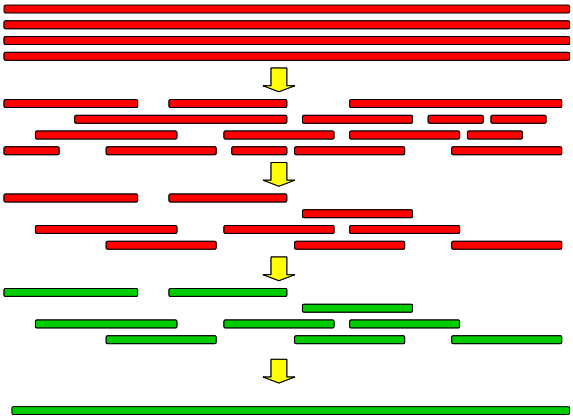
Nutzen und Problematik der Genomsequenzierung

- Verbesserung der Krankheits-Diagnostik
- Frühere Erkennung von Prädispositionen für Krankheiten
- Medikamenten-Design
- Gentherapie
- Organersatz (Eignung des Spenders, in vitro Herstellung)
- Ethische und rechtliche Problematik
  - Gentests zur Krankheitsdiagnose, z.B.: Soll/darf ein Gentest durchgeführt werden, wenn noch keine Therapie verfügbar ist? Wer hat Zugang zu den Testergebnissen? Wie verlässlich sind die Gentests?
  - Kommerzialisierung: Darf ein Gen patentiert werden? (Derzeitige Rechtslage: Nein). Wer hat Zugang zu den Datenbanken?

Sequenzierungsverfahren

??

??



DNA Zielmolekül

(1) Kopieren

(2) Zerkleinern

(3) Auswählen

(4) Sequenzieren

(5) Assemblieren

AACCTTACTACTGGGGTTTTATGCATGCATGCCCCGGGA
TTGGAATGATGACCCCAAAATACGTACGTACGGGGCCCT

Celera hatte 300 ABI 3700 DNA Sequenzierer im Einsatz

Sequenzfragestellungen

- Kartierungsproblematik
  - Auf welchem Chromosom befindet sich welches Gen (welche Sequenz) an welcher Stelle
- Codierung
  - Welche Teilsequenzen codieren (d.h. sind CDS), welche nicht?
- Datenbanksuche nach ähnlichen Sequenzen (Texten) (z.B. für Verwandschaftsbeziehungen)
  - Gegeben ein Pattern P und eine Menge von Texten (Sequenzen) T = $\{t_1, t_2, ... , t_s\}$: Suche alle Sequenzen t_i, die P lokal oder global ähneln
  - Gegeben ein Pattern P und ein großer Text T: Suche alle Teilsequenzen von T, die dem Pattern P oder Teilsequenzen des Pattern ähneln
- Berechnung von Sequenzalignments
- Sequenz-Assemblierungs-Problem (Sequence Assembly Problem):
  - Gegeben die Überlappungsinformationen und Alignments von Fragmenten einer "unbekannten" Sequenz. Man bestimme die Reihenfolge der Buchstaben (Basen) der "unbekannten" Sequenz

Editierdistanz in der Bioinformatik*

- Bestimmung eines *Alignments* zweier Sequenzen s1 und s2:
  - Übereinanderstellen von s1 und s2 und durch Einfügen von Gap-Zeichen Sequenzen auf dieselbe Länge bringen: Jedes Zeichenpaar repräsentiert zugehörige Editier-Operation
  - Kosten des Alignment: Summe der Kosten der Editier-Operationen
  - *optimales Alignment*: Alignment mit minimalen Kosten (= Editierdistanz)
  - Komplexität: O (n*m) mit n, m Länge der beiden Sequenzen
- Details zu Alignments in Kap 4. der Vorlesung Algorithmen und Datenstrukturen 2 (Prof. Rahm)
  - http://dbs.uni-leipzig.de/de/lehre/db-lernmaterial-vorl.html

* www.techfak.uni-bielefeld.de/bcd/Curric/PrwAli/node2.html

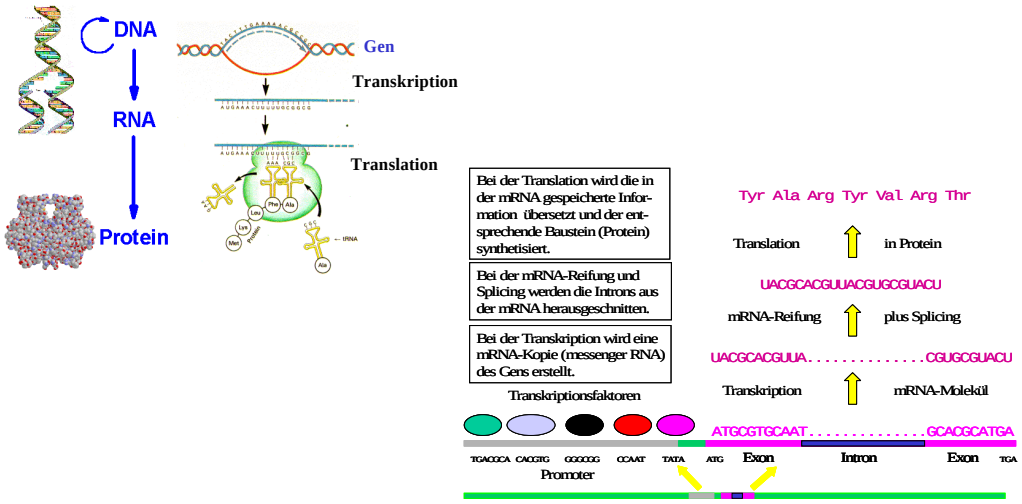
Alignment: Beispiel

Genome verschiedener Spezies

	Organismus	Genomgröße [kb]	Anzahl der Gene	bekannte Genomsequenz [%]
	Bakterium *Escherichia coli*	$4{,}80 \times 10^6$		100 %
	Hefe *Saccharomyces cerevisiae*	$1{,}44 \times 10^7$	~ 6.000	100 %
	Blütenpflanze *Arabidopsis thaliana*	$1{,}00 \times 10^8$	~ 25.000	> 95 %
	Fadenwurm *Caenorhabditis elegans*	$1{,}00 \times 10^8$	~ 20.000	100 %
	Fliege *Drosophila melanogaster*	$1{,}65 \times 10^8$	~ 25.000	100 %
	Maus *Mus musculus*	$3{,}00 \times 10^9$	~ 50.000	> 80 %
	Mensch *Homo sapiens*	$3{,}00 \times 10^9$	~ 50.000	100 %

Transkription und Translation

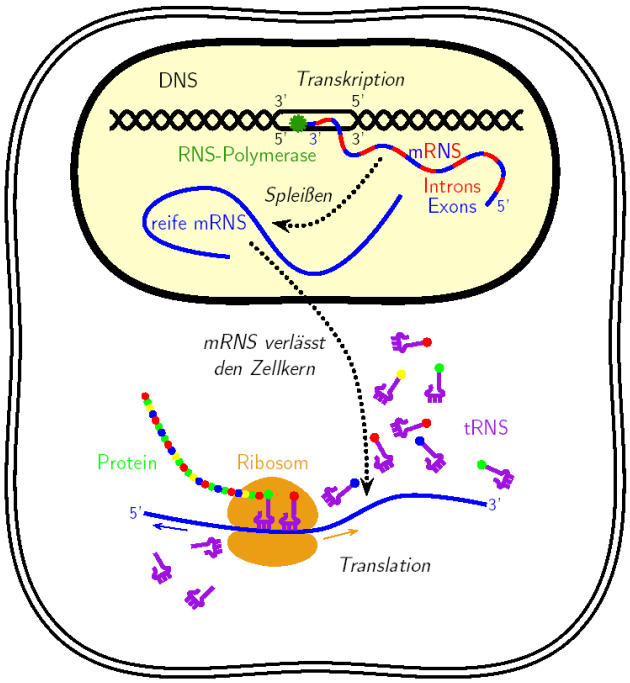
- Gene kodieren die Baupläne für den Aufbau der Proteine, die wiederum (als Enzyme) alle weiteren biomolekularen Vorgänge steuern



  - nur ca. 28% des Genoms werden transkribiert
  - nur ca. 2% der DNA kodiert für Proteine

Splicing

- Splicing: Entfernen (*Spleißen*) von Stücken, die keine Erbinformation tragen (*Introns*), aus der Boten-RNS (mRNS)
- Zusammensetzung der codierenden Teile (*Exons*) zu sogenannter *reifer Boten-RNS* (*mature messenger RNA*)



Genetischer Code

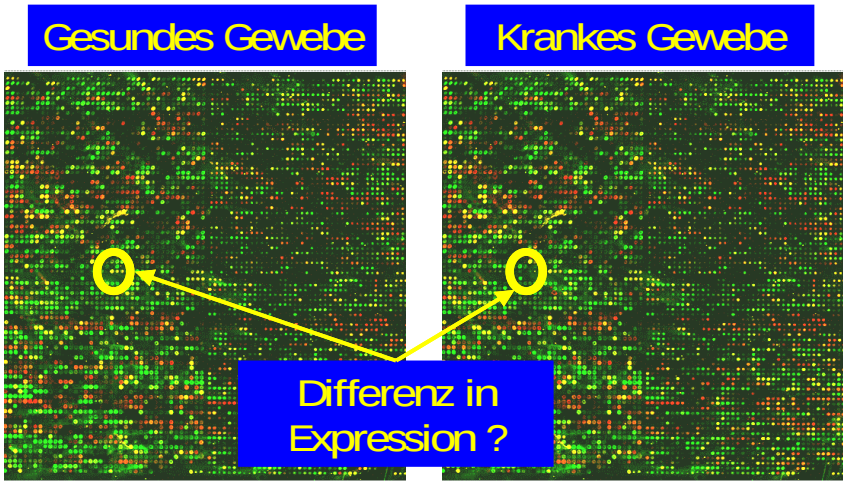
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Alanin	ala
Arginin	arg
Asparagin	asn
Asparaginsäure	asp
Cystein	cys
Glutamin	gln
Glutaminsäure	glu
Glycin	gly
Histidin	his
Isoleucin	ile
Leucin	leu
Lysin	lys
Methionin	met
Phenylalanin	phe
Prolin	pro
Serin	ser
Threonin	thr
Tryptophan	trp
Tyrosin	tyr
Valin	val

(C) Prof. R. Müller, Prof. E. Rahm

Genexpression

- Zielsetzung: Messen der "Expressionsniveaus" aller Gene einer bestimmten Zelle zu einem bestimmten Zeitpunkt
- Microarray-Verfahren
  - Unterschiedliche Expressionsniveaus erzeugen unterschiedliche Farbniveaus
  - Einsatz von Methoden der Bildverarbeitung
- Dazu mehr in Kapitel 5 (Genexpressions-Datenbanken)



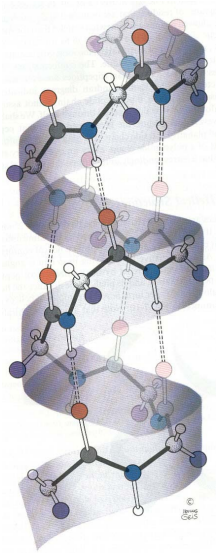
Proteine

- Zentrale Elemente des Stoffwechsels (als Enzyme)
- Besitzen Primär-, Sekundär-, Tertiär- und ggf. Quartärstrukturen
- Primärstruktur
  - Aminosäuresequenz (lineare Abfolge)
  - Sequenzierung eines Proteins am Stück schwierig (bereits Länge von 20 Aminosäuren nicht-trivial), daher oft Sequenzierung des zugehörigen Gens

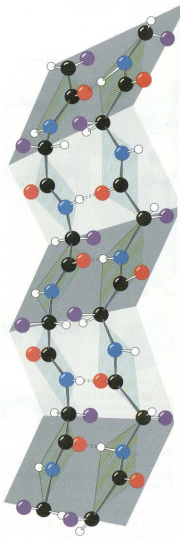
Proteine (2)

- Sekundärstruktur
  - 2-dim. Anordnung in der Ebene
  - Typen: α-Helix (Hohlstruktur, Pauling & Corey 1951, am häufigsten), β-Helix (Faltblatt, Pauling & Corey 1951), Kollagenhelix, random coil (coil = Windung, ohne erkennbares 2-dim. Muster)
  - Oft lagern sich zwei oder drei Sekundärstrukturelemente zu sogenannten *Motifs* zusammen, z.B. zu *coiled coils* aus zwei verdrillten α-Helices (spielen wichtige Rolle in Faser-Proteinen)

α-Helix



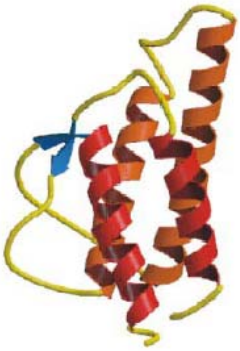
β-Helix

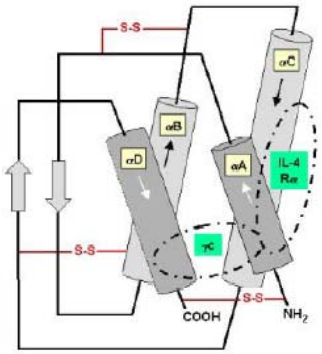


Proteine (3)

- Tertiärstruktur
  - dreidimensionale Raumstruktur eines Proteins
  - Determiniert Fuktion eines Proteins
  - Beispiel: Struktur von Interleukin-4, einem Protein mit immunregulierenden Aufgaben: 4 α-Helices (rot), zwei sehr kurze, einsträngige ß-Faltblätter (blau) und verbindende loops in random coil Struktur (gelb)
  - Wichtiges Ziel der Biologie: Vorhersage der Funktion aufgrund der Primärstruktur (→ Protein-Design)

Interleukin-4

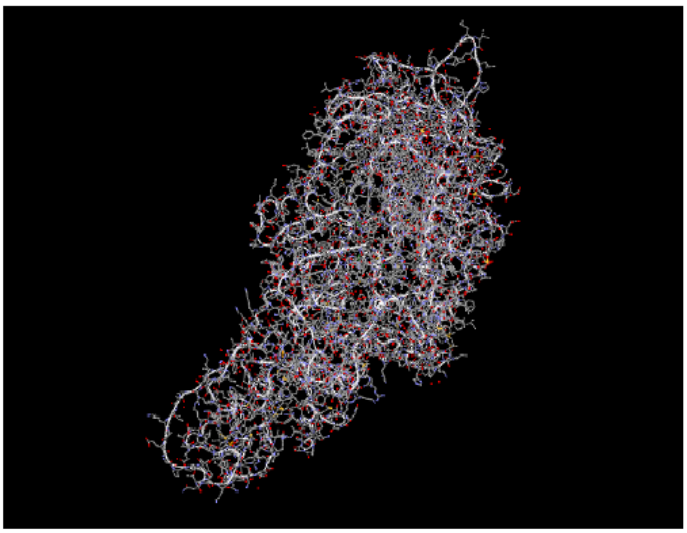




Reinemer/Sebald/Duschl: Angew. Chemie Int. Ed. 39:2834 (2000)

Proteine (4)

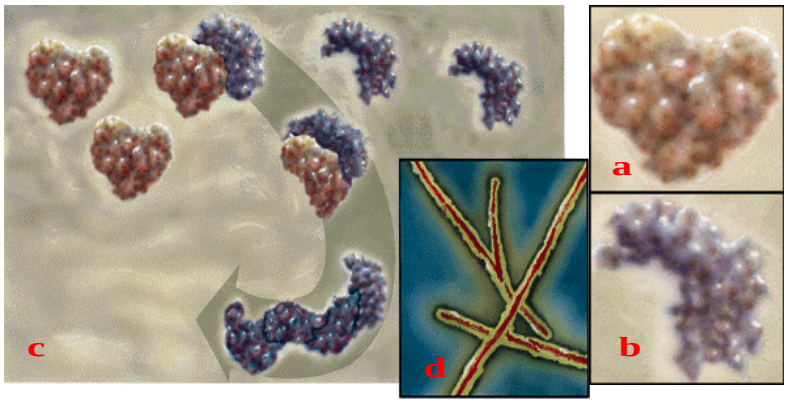
- Quartärstruktur
  - entsteht durch Assoziation mehrerer separater Proteine, die durch nicht-kovalente Wechselwirkungen zusammengehalten werden
  - Nicht alle Proteine besitzen Quartärstruktur
  - Im Bild: Das photosynthetische Reaktionszentrum von Rhodopseudomonas viridis, ein grosser Komplex aus mehreren Proteinen (Nobelpreis für Chemie für die Aufklärung dieser Struktur; 1988 Michel, Deisenhofer & Huber)



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

Prione

- Proteine, die die Struktur von anderen Proteinen verändern können
- Ursache von BSE, Creutzfeldt-Jakob-Krankheit ...
- Prion-Hypothese
  - 2 Prion-Formen: Das normale unschädliche Prion-Protein (PrP^{C}, α-Helix, a) kann zur pathogenen Isoform (PrP^{Sc}, β-Helix, b) umgewandelt werden. Diese Konversion schreitet in Form einer Kettenreaktion (c) fort. Dabei bilden sich lange filamentäre Aggregate (d), die schrittweise neuronales Gewebe zerstören



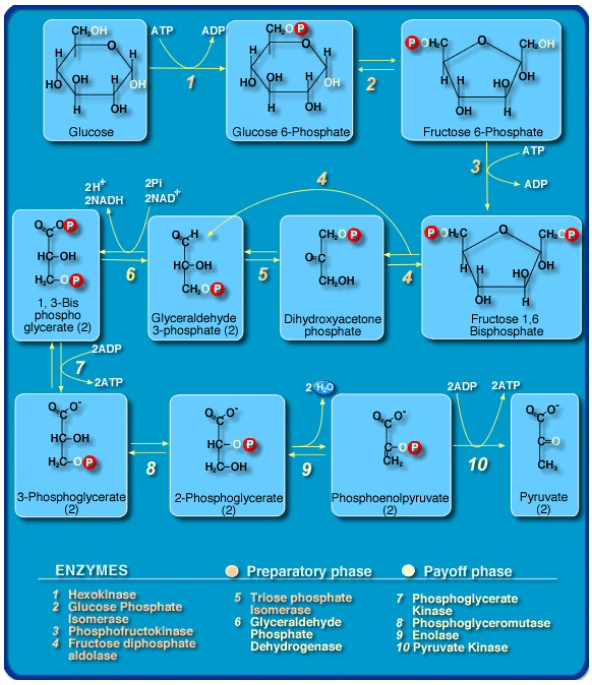
- Entdeckung der Prione führte zu Dogmenrelativierung Ende der neunziger Jahre, denn:
  - Gewisse Proteine (eben die Prione) können sowohl Helix als auch Faltblattstruktur annehmen (→ nicht alle Proteine sind durch Basensequenz determiniert)
  - Proteine allein (als Prione) können schon Krankheiten übertragen (ohne Viren, Bakterien etc.)

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: Metabolic Pathways

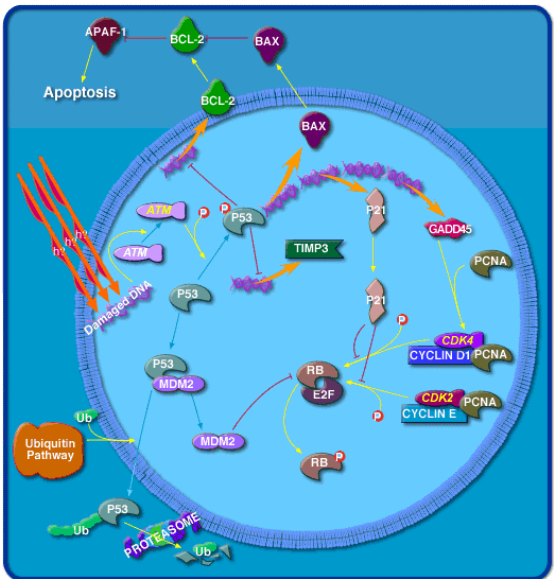
- Metabolismus: Gesamtheit aller lebensnotwendigen biochemischen Vorgänge beim Aufbau, Abbau und Umbau eines Organismus sowie seinem Austausch mit der Umwelt
- 2 grundlegende Stoffwechselvorgänge
  - Assimilation/Anabolismus (z.B. Photosynthese)
  - Dissimilation/Katabolismus (z.B. Atmung, Gärung)



Beispiel Glykolyse

Stoffwechsel: Regulatory Pathways

- 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
  - p53 mutiert in fast allen Tumoren vorhanden



Zusammenfassung

- Genom
- Proteine und Prione
- Translation und Transkription
- Stoffwechsel