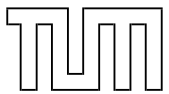
Algorithmische Bioinformatik 1

Dr. Hanjo Täubig

Lehrstuhl für Effiziente Algorithmen
(Prof. Dr. Ernst W. Mayr)
Institut für Informatik
Technische Universität München

Sommersemester 2009



Übersicht

Vorlesungsdaten

- Modul: IN2179
- Bereich:
  Informatik III (Theoretische Informatik)
- Semesterwochenstunden:
  4 SWS Vorlesung + 2 SWS Übung
- ECTS: 8 Punkte
- Vorlesungszeiten:

  Dienstag 15:15 – 16:45 Uhr (MI Hs 2)
  Freitag 13:15 – 14:45 Uhr (MI Hs 3)

- Web:
  http://www14.in.tum.de/lehre/2009SS/cb/

Ausfall am 26.05.2009

Am

Dienstag, dem 26.05.2009

ist der Hörsaal aufgrund einer Klausur besetzt.
An diesem Tag entfällt die Vorlesung.

Dozent

- Dr. Hanjo Täubig
  (Lehrstuhl für Effiziente Algorithmen / Prof. Mayr)
- eMail:
  taeubig@in.tum.de
- Web:
  http://www14.in.tum.de/personen/taeubig/
- Telefon: 089 / 289-17740
- Raum: MI 03.09.039

- Sprechstunde: Mittwoch 13-14 Uhr
  (oder nach Vereinbarung)

Übung

- Übungsleiter: Johannes Krugel (krugel@in.tum.de)
- Zeit ??? ??:?? – ??:?? Uhr
- Raum: MI 03.11.018 (?)
- Web:
  http://www14.in.tum.de/lehre/2009SS/cb/uebung/
- Umfang: ca. 10 Übungsblätter mit jeweils 3 bis 4 Aufgaben
- Ablauf:
  - Jede Woche wird auf der Webseite ein Übungsblatt bereitgestellt, das dann zu bearbeiten und (freiwillig) abzugeben ist.
  - In der Übung nach dem Abgabetermin werden die Lösungen besprochen und eventuelle Fragen zum Vorlesungsstoff diskutiert.

Voraussetzungen

- Voraussetzungen:
  Stoff des Informatik Grundstudiums:
  - Einführung in die Informatik
  - Grundlagen: Algorithmen und Datenstrukturen
  - Diskrete Strukturen
  - Einführung in die Theoretische Informatik

- vorteilhaft, aber nicht unbedingt notwendig:
  Effiziente Algorithmen und Datenstrukturen I/II

Inhalt

- Molekularbiologische Grundlagen
  - Vererbung, Nukleinsäuren, Proteine
- Algorithmen zur Textsuche
  - KMP-, AC-, BM-Algorithmus, Suffix-Bäume
- Paarweises Sequenzen-Alignment
  - Distanz/Ähnlichkeit, globale/lokale/semiglobale Alignments
- Mehrfaches Sequenzen-Alignment
  - Maße, Dynamische Programmierung, Divide-and-Conquer, Center-Star-Approximation, Konsensus-/Steiner-Strings
- Fragment Assembly
  - Shotgun Sequencing, Overlap Detection, Fragment Layout

Literatur

- Michael S. Waterman:
  Introduction to Computational Biology – Maps, Sequences and Genomes
  Chapmann and Hall, 1995

- João C. Setubal, João Meidanis:
  Introduction to Computational Molecular Biology
  PWS, 1997

- Dan Gusfield:
  Algorithms on Strings, Trees, and Sequences – Computer Science and Molecular Biology
  Cambridge University Press, 1997

Literatur (Forts.)

- Pavel A. Pevzner:
  Computational Molecular Biology – An Algorithmic Approach
  MIT Press, 2000

- Peter Clote, Rolf Backofen:
  Computational Molecular Biology – An Introduction
  Wiley, 2000

- Neil C. Jones, Pavel A. Pevzner:
  An Introduction to Bioinformatics Algorithms
  MIT Press, 2004

Vererbung

- Gregor Mendel: Untersuchung der Verteilung von verschiedenen Eigenschaften nach der Vererbung bei Erbsen (1860)
- Beobachtung:
  - Kreuzung von (reinerbigen) Erbsen mit glatter und runzliger Oberfläche lieferte nur glatte Kinder,
  - Kinder der zweiten Generation waren glatt und runzlig, aber im Verhältnis 3:1
- Annahme: jede Erbse bekommt für ein bestimmtes Merkmal (z.B. die Art der Oberfläche: runzlig oder glatt) je eine Erbinformation von beiden Eltern

Vererbung

- **Uniformitätsregel**:
  Aus zwei jeweils homozygoten Eltern, die sich in einem einzigen Merkmal unterscheiden, entsteht eine Kindgeneration mit gleichartigen Individuen.
- **Spaltungsregel**:
  Werden zwei Mischlinge der ersten Kindgeneration gekreuzt, so spalten sich die Merkmale in der zweiten Kindgeneration im Verhältnis 1:3 (bei dominant/rezessiven Merkmalen) bzw. im Verhältnis 1:2:1 (bei intermediären Merkmalen) auf.
- **Unabhängigkeitsregel**:
  Unterschiedliche Merkmale (z.B. Samenfarbe und Samenform) werden unabhängig voneinander vererbt.

  (Gilt nur, wenn sich beide Gene auf unterschiedlichen Chromosomen befinden bzw. wenn sie innerhalb eines Chromosoms einen genügend großen Abstand haben.)

Vererbung

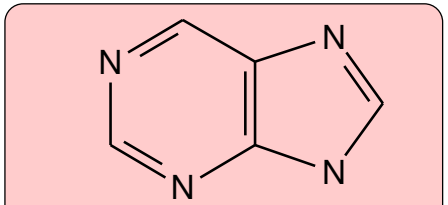
- Solche vererbten Informationen für Merkmale nennen wir heute Gene.
- Unterscheidung:
  Genotyp (vererbte Information) / Phänotyp (äußeres Erscheinungsbild)
- Sind beide Informationen (Allele) gleich, nennt man den Organismus homozygot, ansonsten heterozygot (bezüglich eines Gens)
- Speichermedium der Gene ist die DNA.
- Gesamtheit der Gene eines Organismus: Genom
- Chromosomen sind Träger der Gene

Polymere

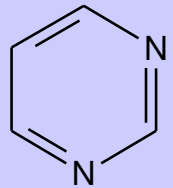
- Polymere:
  Stoffe, die aus vielen gleichen oder ähnlichen Teilen bestehen

- bedeutende Stoffklassen in Organismen sind (Bio-)Polymere:
  - DNA
  - RNA
  - Proteine
  - Kohlenhydrate
  - Lipide

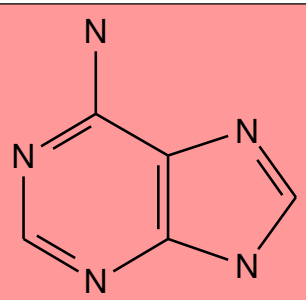
Nukleinbasen



Purin

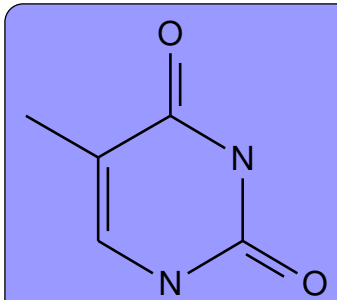


Pyrimidin

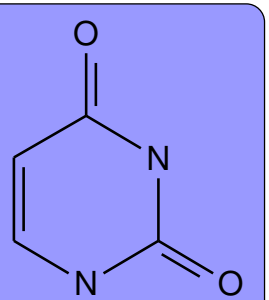


Adenin (A)

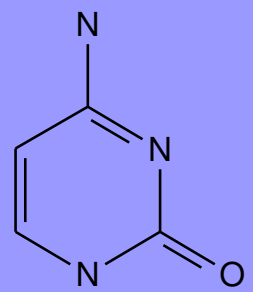
Guanin (G)



Thymin (T)

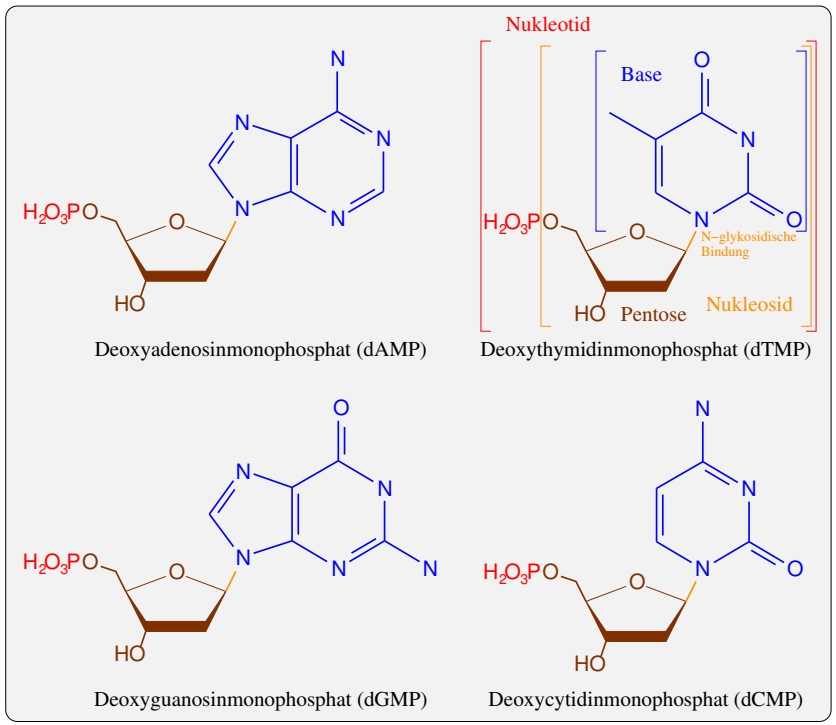


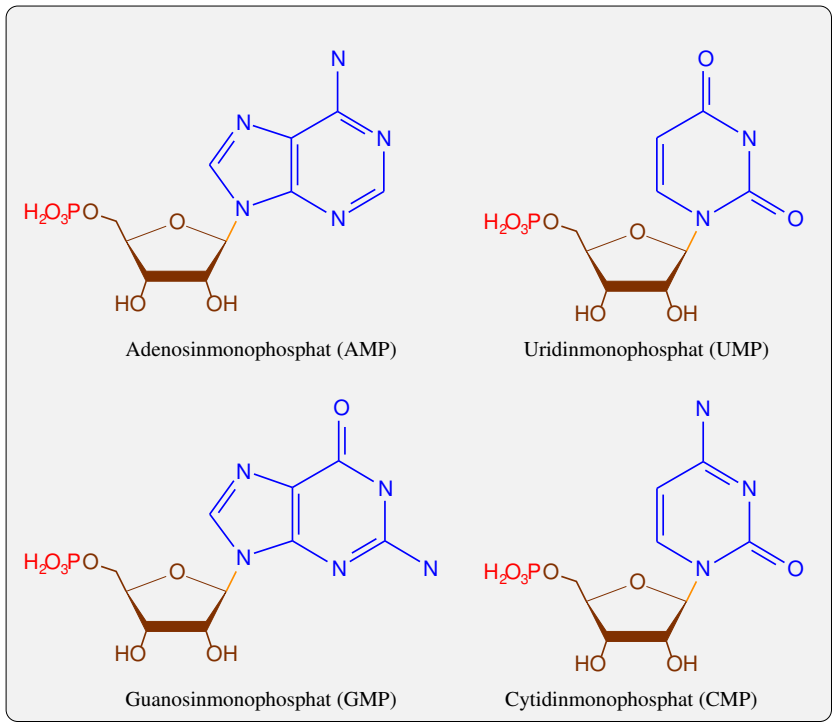
Uracil (U)



Cytosin (C)

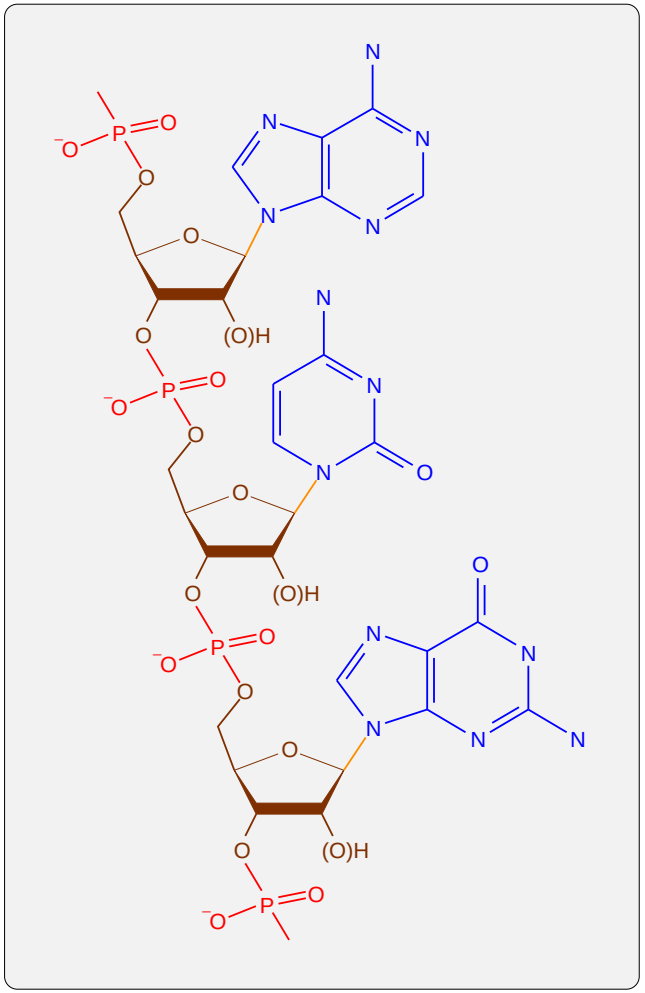
Nukleoside und Nukleotide





- Pentose = Ribose / Deoxyribose
- Nukleosid = Nukleobase + Pentose
- Nukleotid = Nukleobase + Pentose + Phosphatrest

DNA/RNA-Strang



DNA/RNA-Struktur

- Bildung eines Doppelstrangs mit komplementären Basen (Publikation Watson/Crick, 1953, basierend auf Arbeiten von Maurice Wilkins und Rosalind Franklin)
- Basenpaarung: Ausbildung von Wasserstoffbrückenbindungen $A = T \qquad C \equiv G$
- Die Basenfolge eines Strangs besitzt eine Orientierung.
- Der haploide Chromosomensatz des Menschen besteht aus mehr als 3 Milliarden Basenpaaren.
- Davon kodieren aber nur wenige Prozent für die ca. 20.000-25.000 Gene.

Sequenz-Hypothese & Zentrales Dogma

The Sequence Hypothesis

. . . assumes that the specificity of a piece of nucleic acid is expressed solely by the sequence of its bases, and that this sequence is a (simple) code for the amino acid sequence of a particular protein.

The Central Dogma

. . . states that once 'information' has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein.

Standardpfad biologischer Informationsübertragung

$$\text{DNA} \xrightarrow{\text{Transkription}} \text{RNA} \xrightarrow{\text{Translation}} \text{Protein}$$

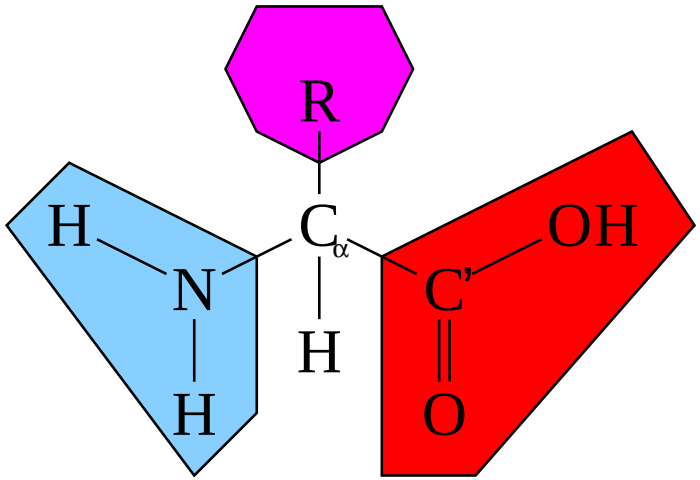
Genetischer Code: universell / fehlertolerant

1. Base	2. Base U	2. Base C	2. Base A	2. Base G	3. Base
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

Proteinbiosynthese

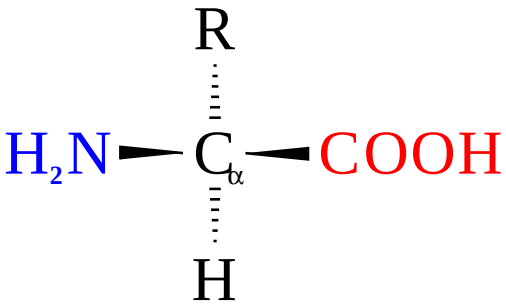
1. Transkription
   1. Ablesen eines Gens auf der DNA und Umschreibung in mRNA (messenger-/Boten-RNA)
   2. Enzym RNA-Polymerase erzeugt mit ATP, UTP, CTP und GTP einen zum DNA-Strang komplementären RNA-Strang
   3. findet bei Prokaryoten im Cytoplasma und bei Eukaryoten im Zellkern statt, so dass bei letzteren eine Überführung der mRNA ins Cytoplasma notwendig ist
2. Translation
   1. Übersetzung der mRNA-Basenfolge (interpretiert als Tripletts) in die entsprechende Aminosäuresequenz
   2. erfolgt an den Ribosomen
   3. tRNA (Tranfer-RNA) transportieren Aminosäuren spezifisch zum Ribosom und erkennen mit ihrem Anticodon das entsprechende Anticodon der mRNA

Aminosäuren

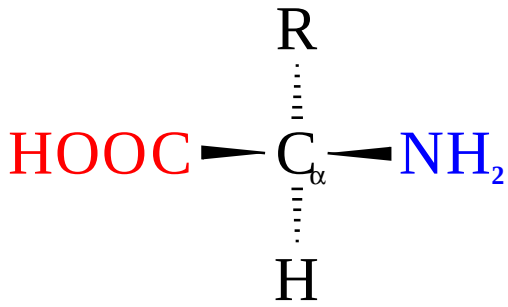


- (α-)Aminosäuren bestehen aus: dem zentralen Kohlenstoffatom C_α, der Amino-Gruppe NH_2, der Carboxylgruppe $COOH$, dem Rest R und dem Wasserstoffatom H

Chiralität der Aminosäuren



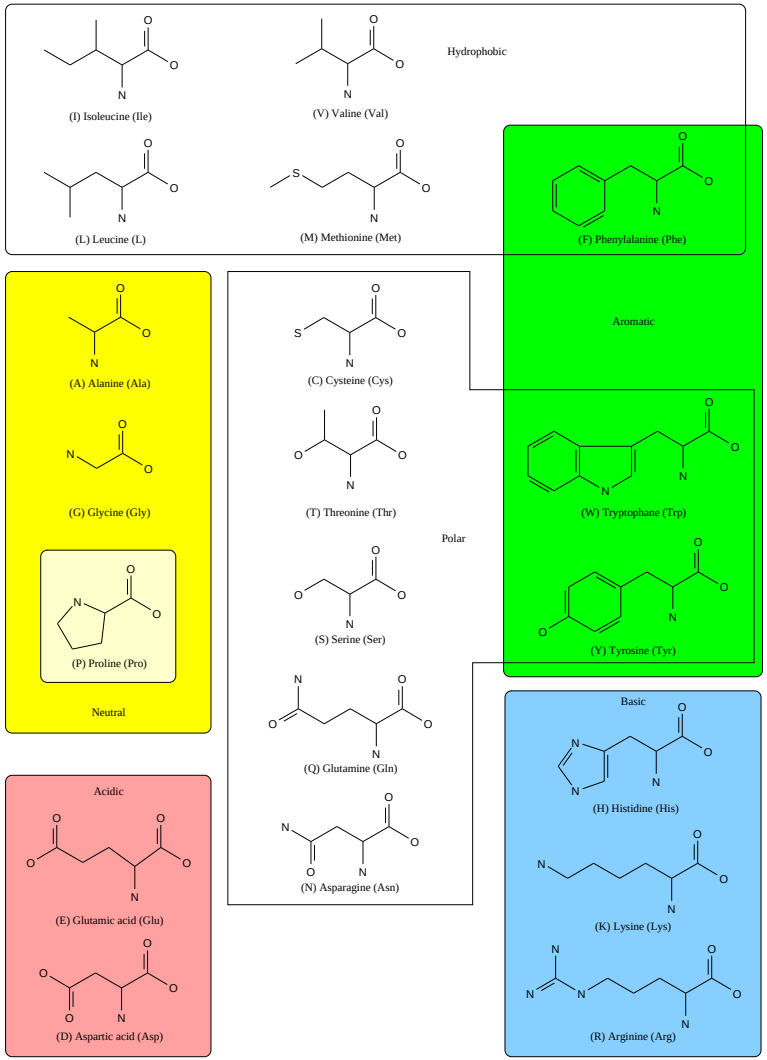
L-α-Aminosäure



D-α-Aminosäure

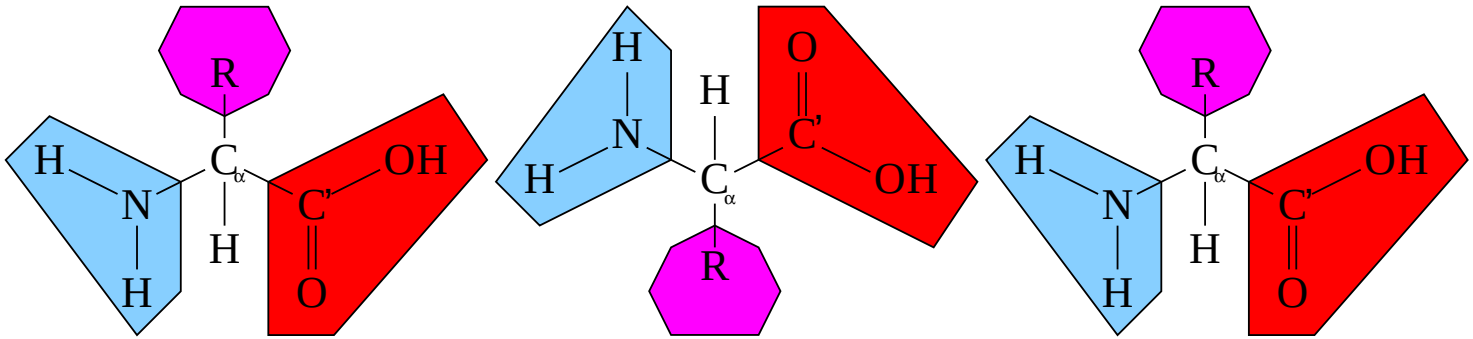
- α-Aminosäuren sind chiral (außer Glycin)
- Proteine bestehen aus L-α-Aminosäuren

Aminosäuregruppen



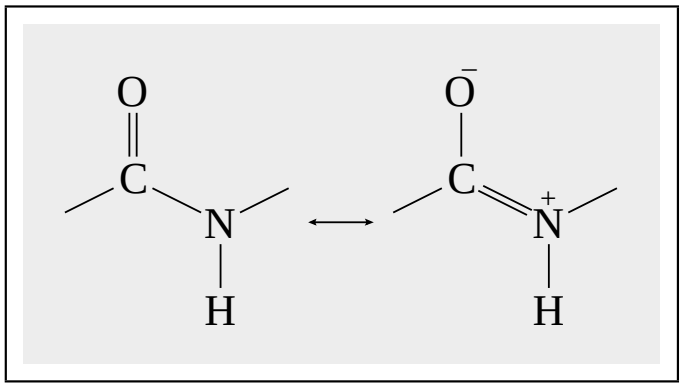
Polypeptide

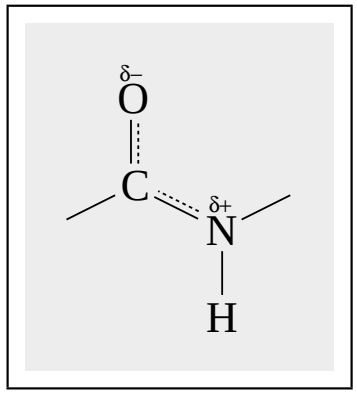
- Polypeptide sind Ketten aus Aminosäuren



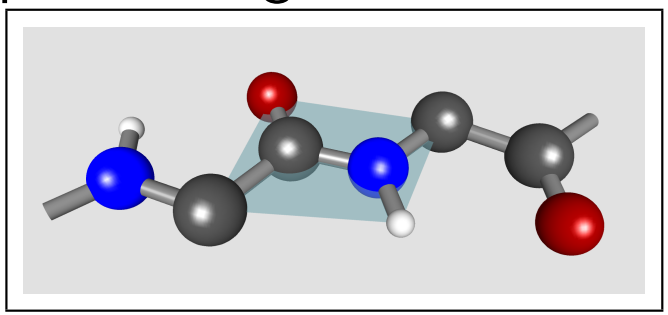
Peptidbindung

- Es entstehen Peptidbindungen:

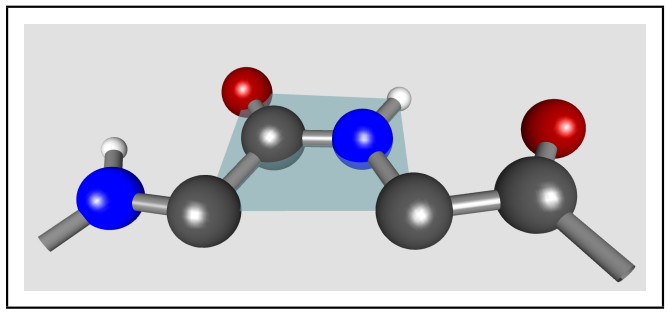




- Peptidbindungen können 2 Formen haben:



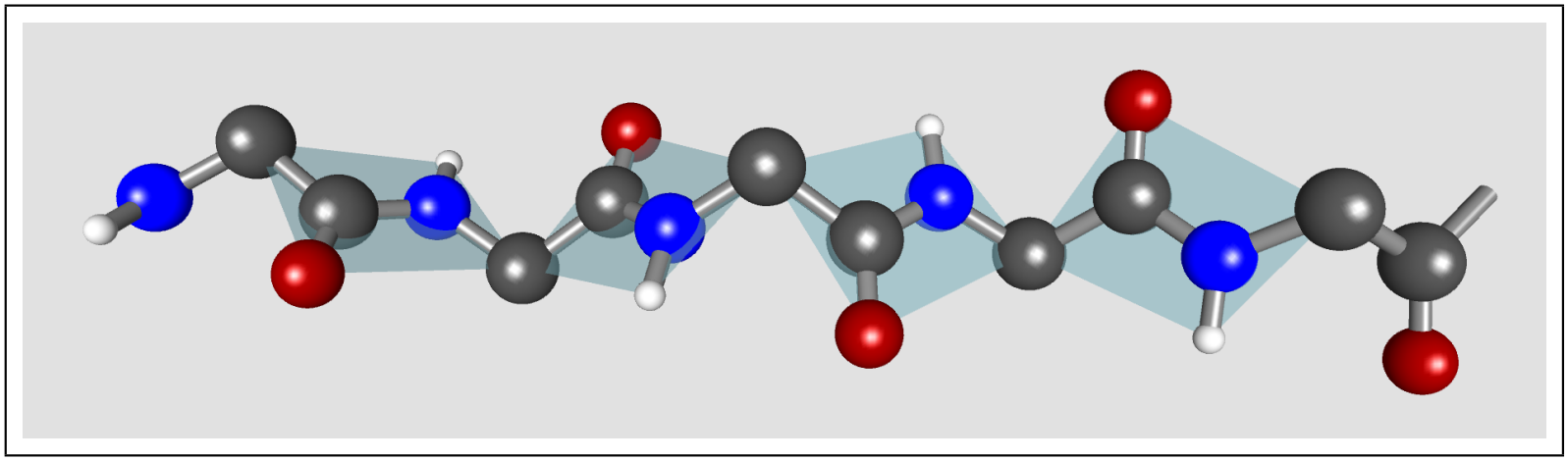
trans



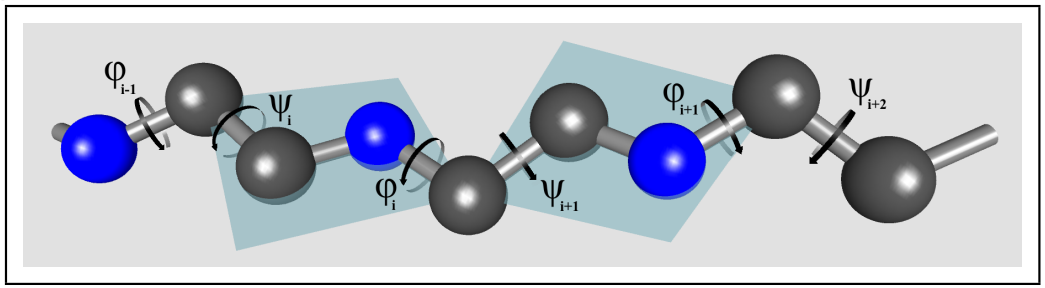
cis

Rückgrat (Backbone)

- Es bildet sich ein Rückgrat:



- Das Rückgrat ist 'verdrehbar':



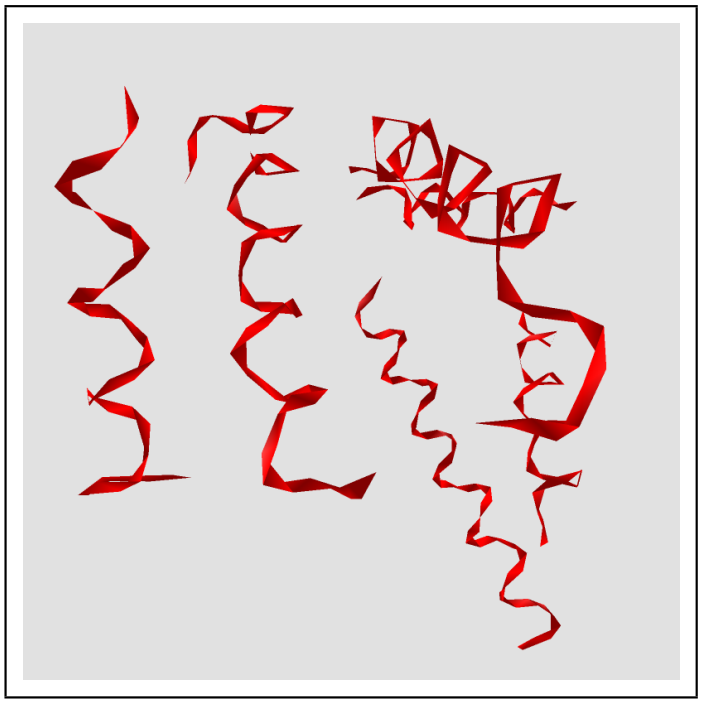
Faltung

- Die 20 proteinogenen Aminosäuren unterscheiden sich in ihren physikalischen / chemischen Eigenschaften
  - Größe,
  - Polarität / Ladung und
  - Aromatizität
- Die verschiedenen Aminosäurereste (Seitenketten) eines Polypeptids bestimmen durch ihre Eigenschaften die Faltung der Kette.
- bei globulären Proteinen (in wässriger Umgebung): hydrophober Kollaps
  - polare Reste an die Oberfläche und
  - hydrophobe Reste ins Innere des Proteins

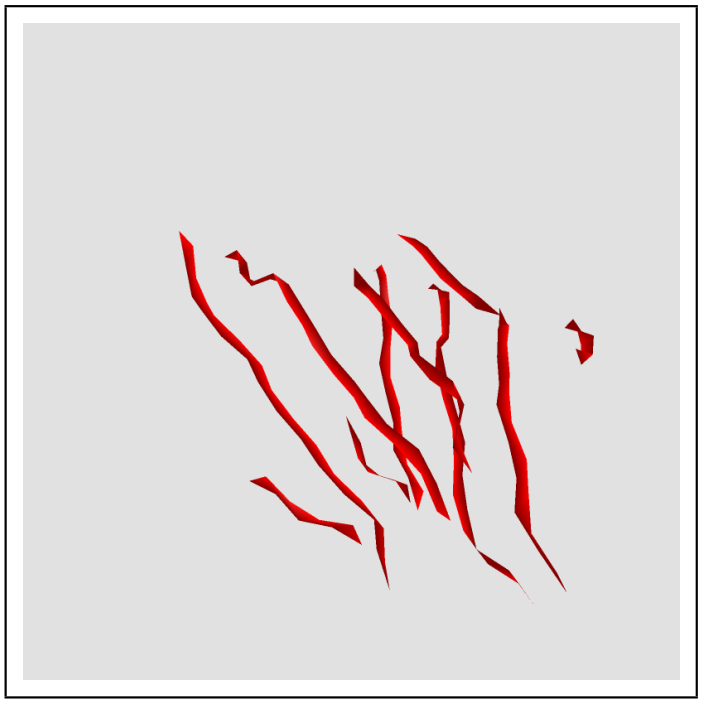
Primär-/Sekundärstruktur

- Primärstruktur: Aminosäuresequenz
- Sekundärstruktur

α-Helix

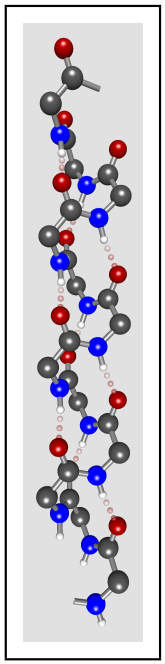


β-Faltblatt

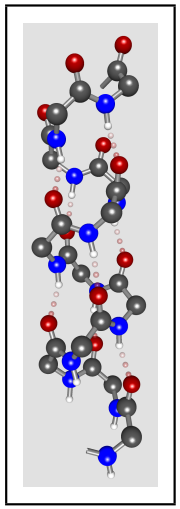


Helices und Faltblätter

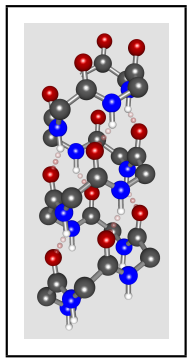
- Unterschiedliche Helix-Typen:



3_{10}-Helix



α-Helix



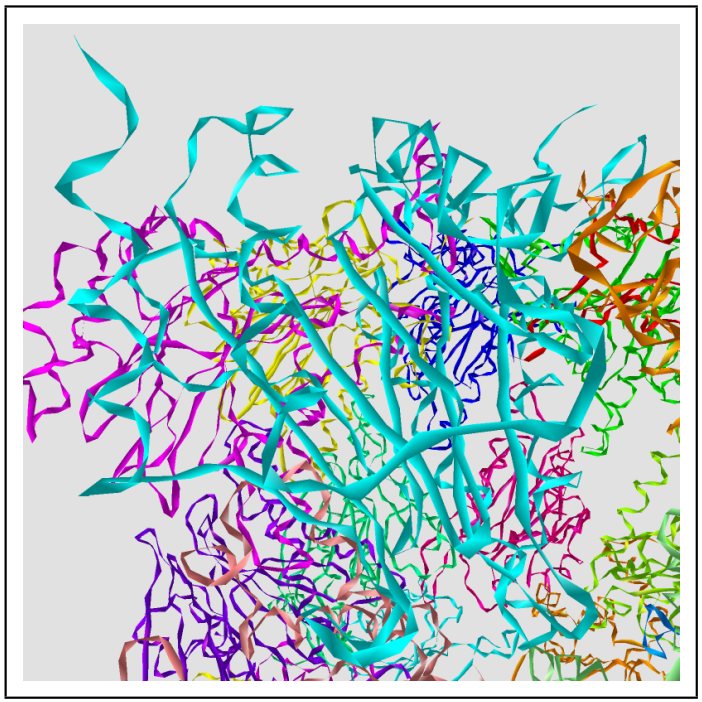
π-Helix

- Faltblätter (Sheets) können aus parallelen oder antiparallelen (oder gemischten) Strands zusammengesetzt sein.

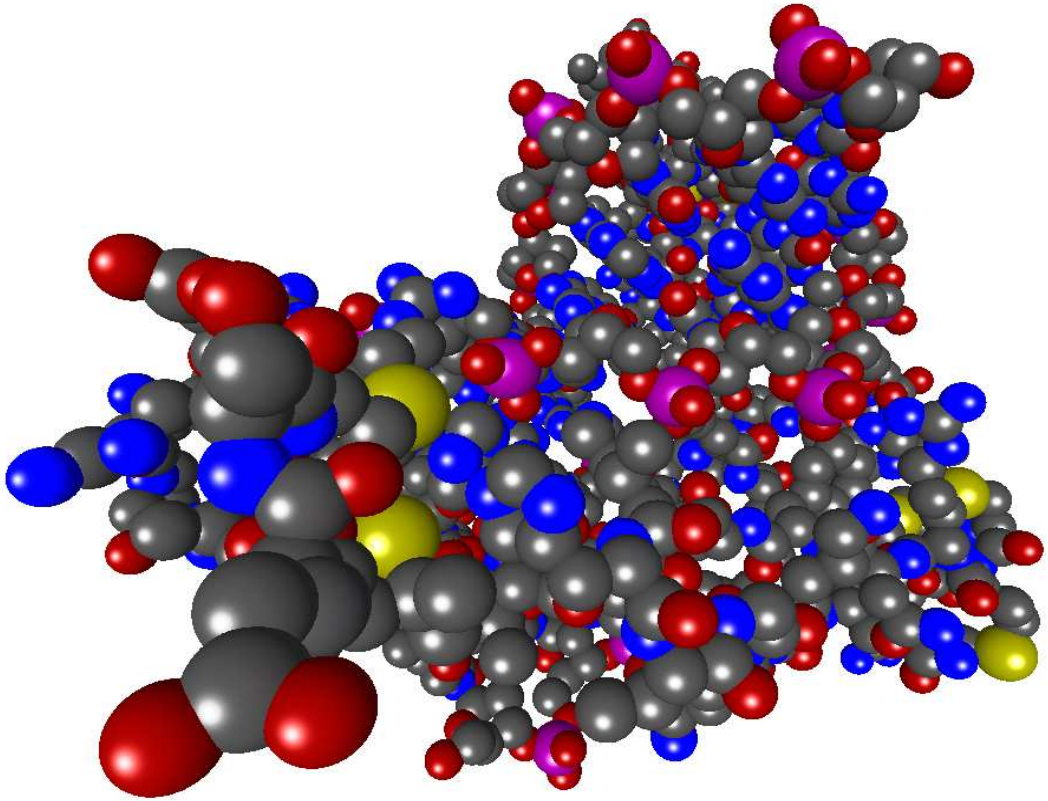
Tertiär-/Quartärstruktur

- Tertiärstruktur (ganze Kette)

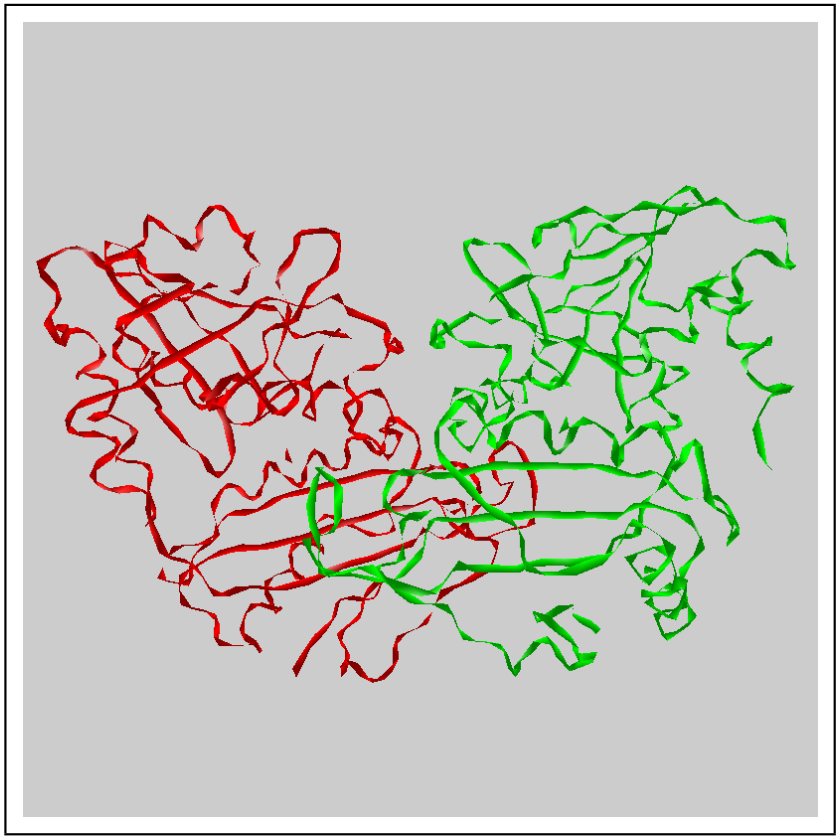
Quartärstruktur (mehrere Ketten)

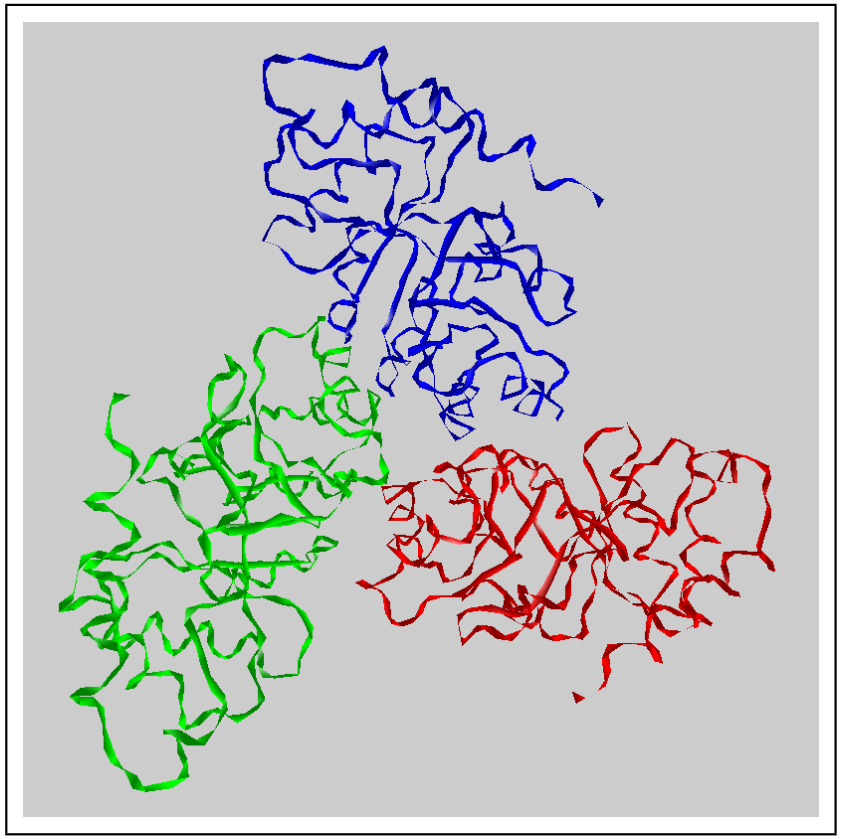


Tertiärstruktur: Beispiel

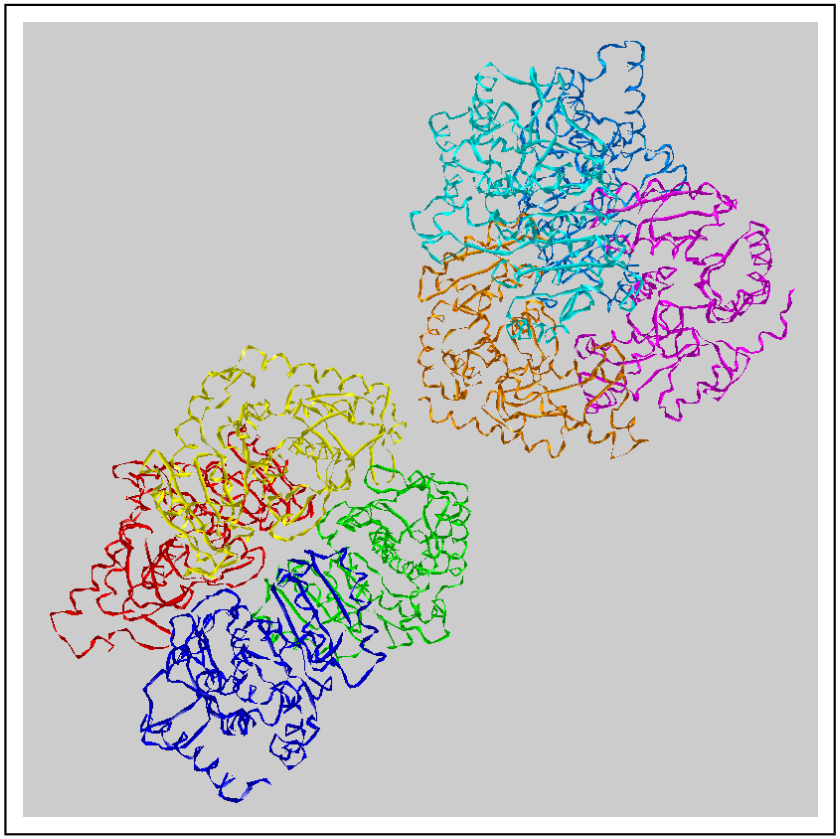


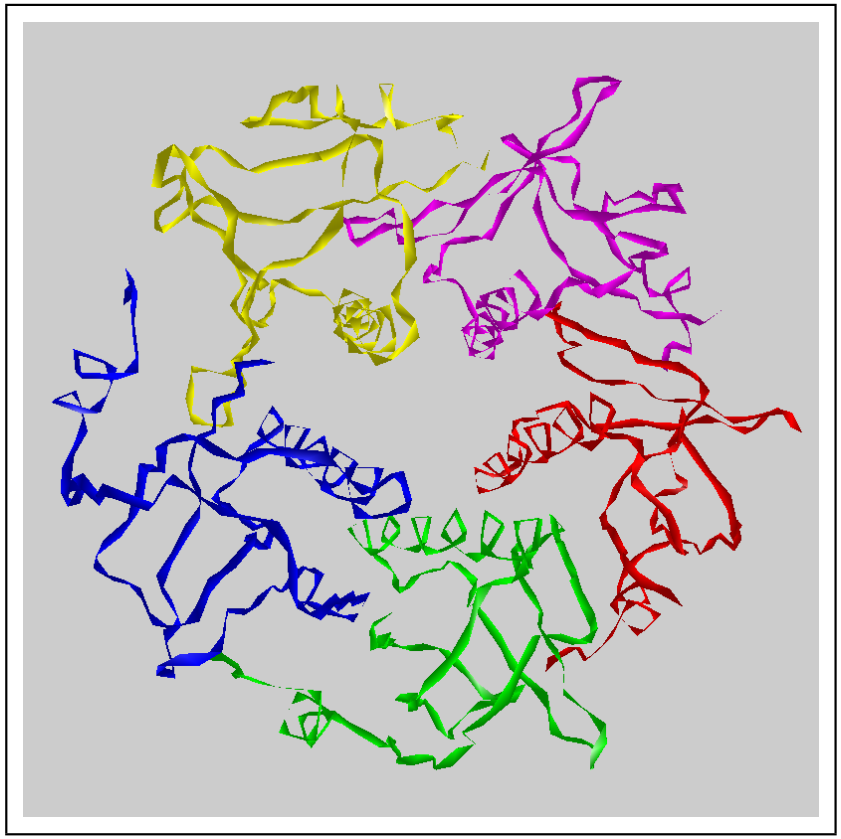
Quartärstruktur: Beispiele



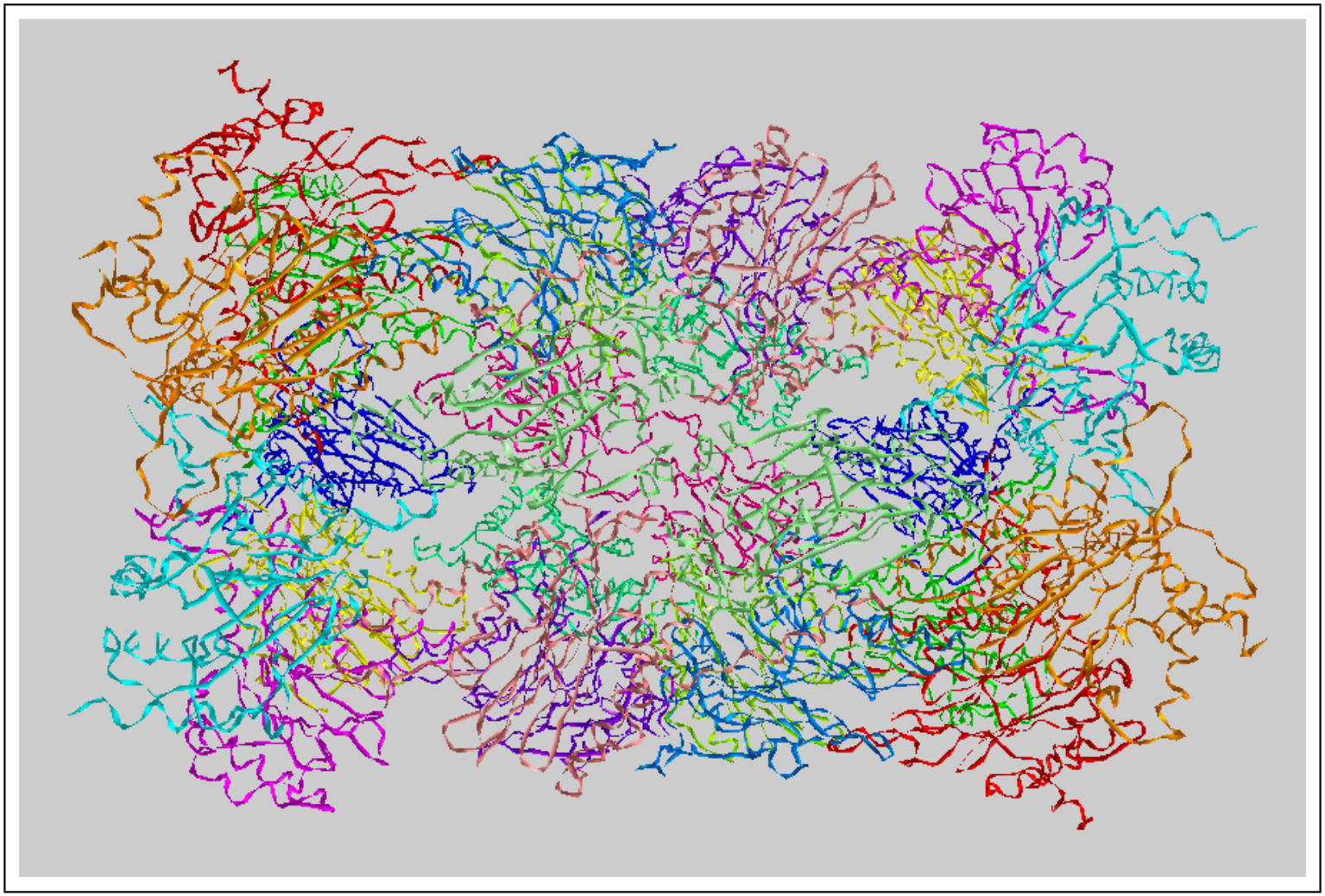


Quartärstruktur: Beispiele

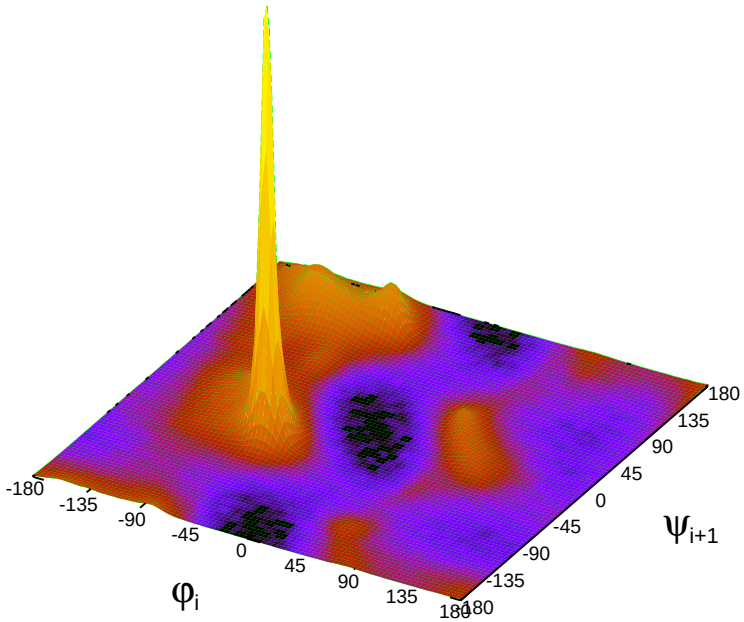


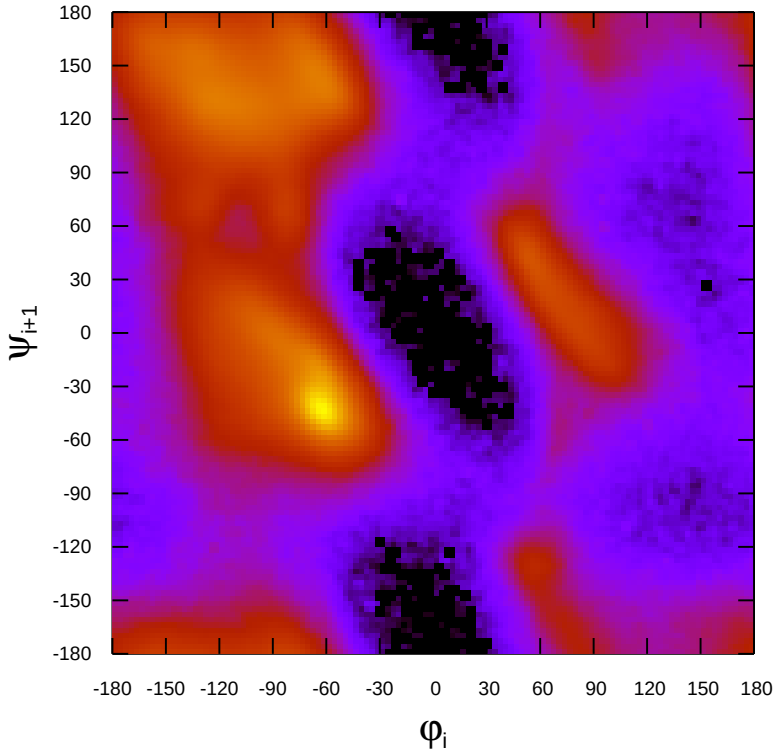


Quartärstruktur: Beispiele

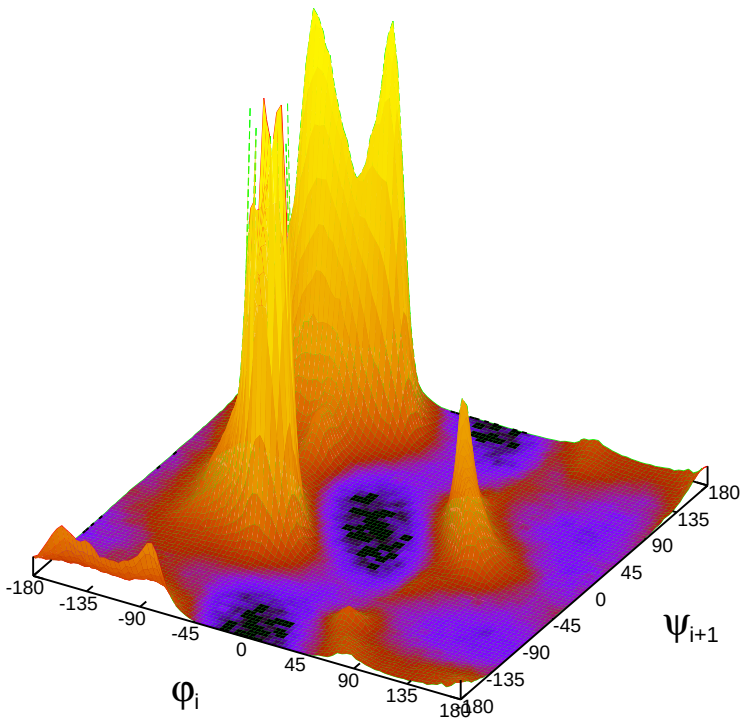


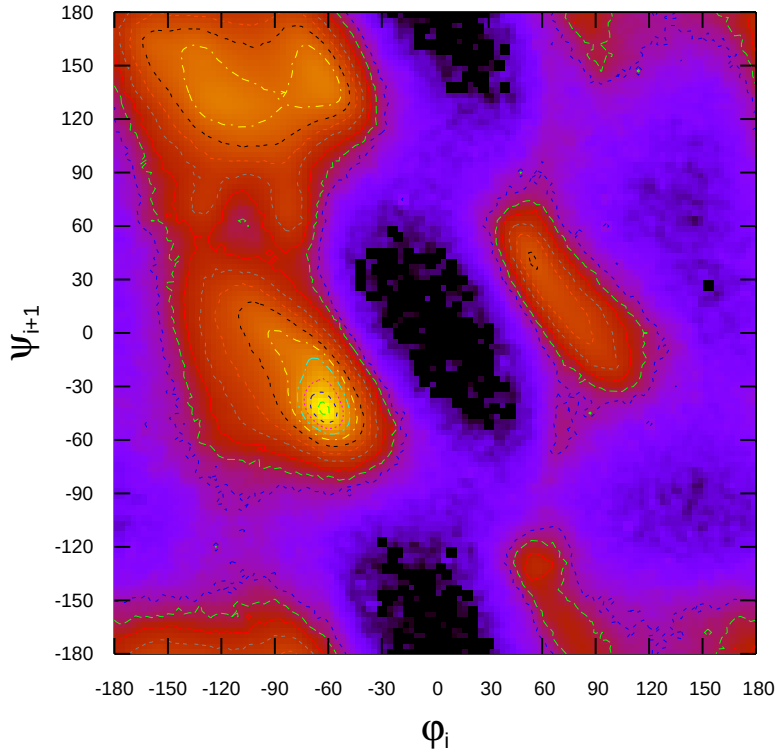
Ramachandran Plot





Ramachandran Plot





Levinthal-Paradoxon

- Obwohl die Berechnung der dreidimensionalen Struktur aus der puren Aminosäuresequenz ein schweres Problem darstellt, faltet sich ein normales Protein innerhalb von Sekunden in seine natürliche Form.