Proteine: Struktur, Modellierung, Dynamik

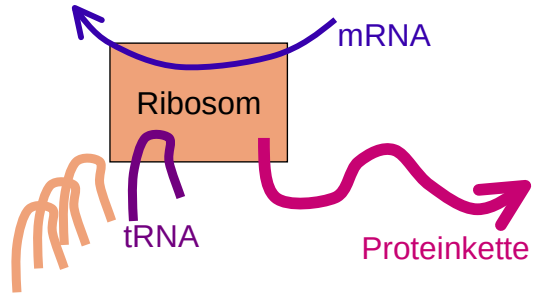
Algorithmische Bioinformatik
WS 2002

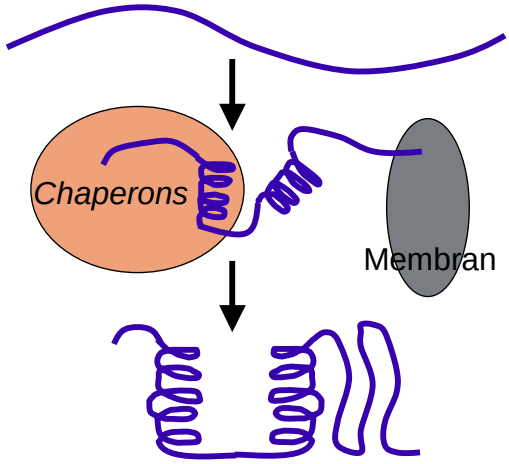
Thomas Steinke
Zuse Institute Berlin (ZIB) <www.zib.de>
Berlin Center for Genom Based Bioinformatics (BCB) <www.bcbio.de>
steinke@zib.de

Einführung

Sequenz → Struktur

- Sekundärstrukturvorhersage
  - 70-75%
- Faltungserkennung
  - PDB: ~ 20 000 (1/2003)
  - SwissProt: ~ 121 000 (1/2003)
- Strukturvorhersage
- Aktivität, Prozesse
- Metabolismus
- Drug-Design



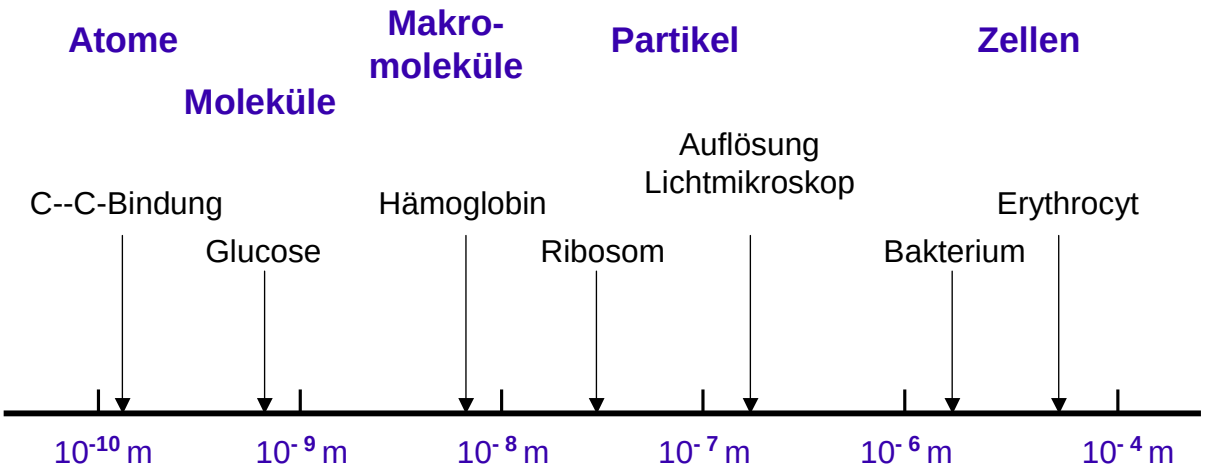


Sequenz-Struktur-Projekte

- Protein Struktur Fabrik (PSF), Berlin
  www.proteinstrukturfabrik.de

- Protein Structure Initiative (PSI)
  www.structuralgenomics.org
  proteome.bnl.gov

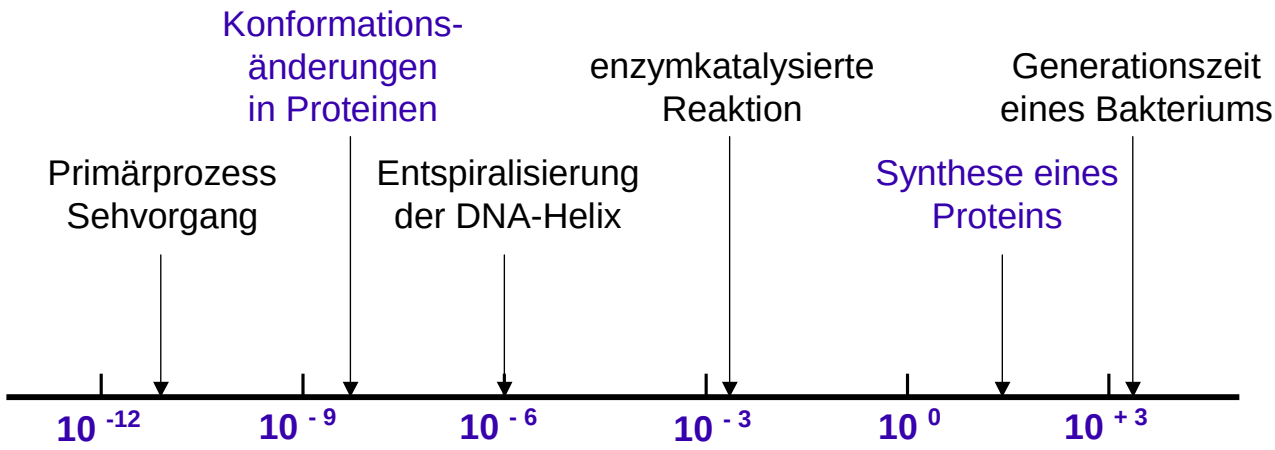
Dimensionen ...



Länge

steinke@zib.de

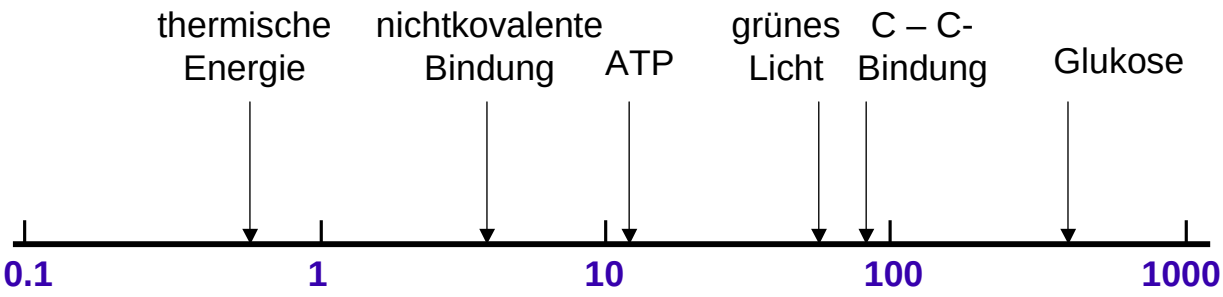
Typische Zeiten ...



Zeit [s]

steinke@zib.de

Energien



Energie [kcal/mol]

Darstellung molekularer Strukturen

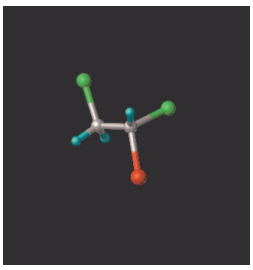
Topologie, Konfiguration, Konformation

- Topologie
  - 2D-Darstellung chemischer Bindungsverhältnisse
- Konfiguration
  - räumliche lokale Bindungsverhältnisse
- Konformation
  - gesamte räumliche atomare Anordnung

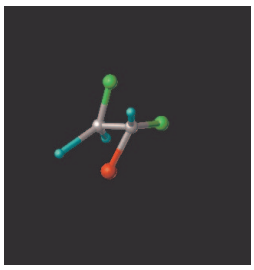
Konformation: eine Definition*

- Chemische Strukturvorstellungen

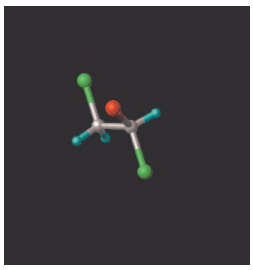
Gestalt



Topologie



Flexibilität



Energiegehalt

- Konformationsisomere
  - Moleküle mit gleicher Konstitution und gleicher Topologie
  - unterscheiden sich nur in räumlicher Anordnung von Atomen oder –gruppen
  - Anordnungen kommen durch Rotation um Bindungen oder Inversion pyramidaler Gruppen zustande
  - sind stabil

** /Bernd Kallies, 2002/*

Wasserstoffbrückenbindung

- X – H ---- Y
- X, Y: elektronegative Atome wie O, N, Halogene, C

- Abstände:

-O—H ... O^---	2.63 A
-O—H ... O--	2.70 A
-O—H ... N--	2.88 A
-N—H ... O--	3.04 A
-N—H ... N--	3.10 A

- Bindungsenergie: $\Delta_B E$ ~ -2 ... -13 kcal/mol
  - Wasser: -13 kcal/mol

Schwache molekulare Wechselwirkungen

- Dispersions-WW
  - immer ΔE<0, Atom 1,2:

$$\Delta E_{12} \approx \frac{\alpha_1 \alpha_2}{r_{12}^6}$$

- Lennard-Jones 6-12 Potential Teilchen i,j:
  - Repulsion + Attraktion (Dispersion)

$$E_{ij} = E_0 \cdot \left[-\left(\frac{R_0}{r_{ij}}\right)^{12} + 2\left(\frac{R_0}{r_{ij}}\right)^{6} \right] = \frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}}$$

van-der-Waals :=

(schwache) Elektrostatik + Dispersion + Repulsion

Hydrophobizität (I)

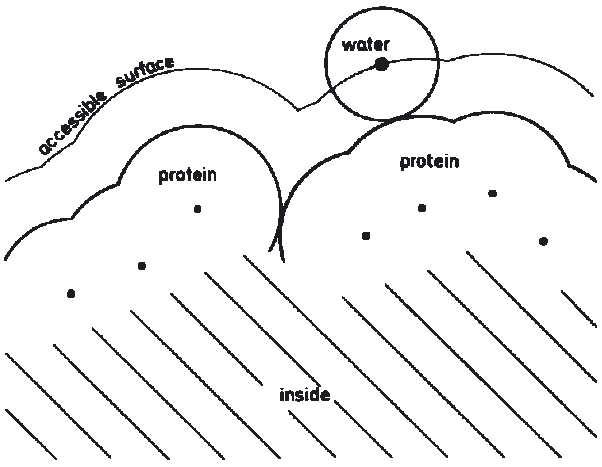
- entropisches (thermodynamisches) Phänomen
- Wechselwirkung Seitenketten mit Wasser
  - → Proteinfaltung
  - Kohlenwasserstoffketten: Leu, Phe

- Thermodynamik:
  - Energie (Enthalpie), Entropie, freie Enthalpie

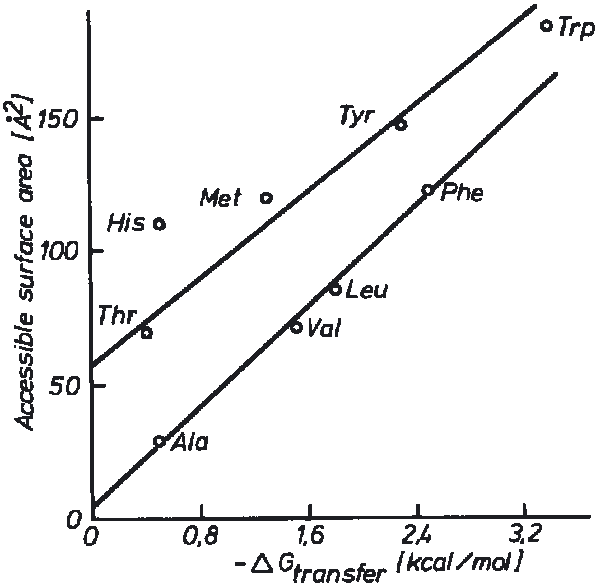
$$\Delta G = -RT \ln K = \Delta H - T\Delta S$$

  - Solvatation:
    - Hohlraum (*cavity*): Zerstörung d. Wasserstruktur
    - Solvenshülle: Umorientierung d. Wassermoleküle
    - Wechselwirkung: Protein -- Wasser

Hydrophobizität (II)



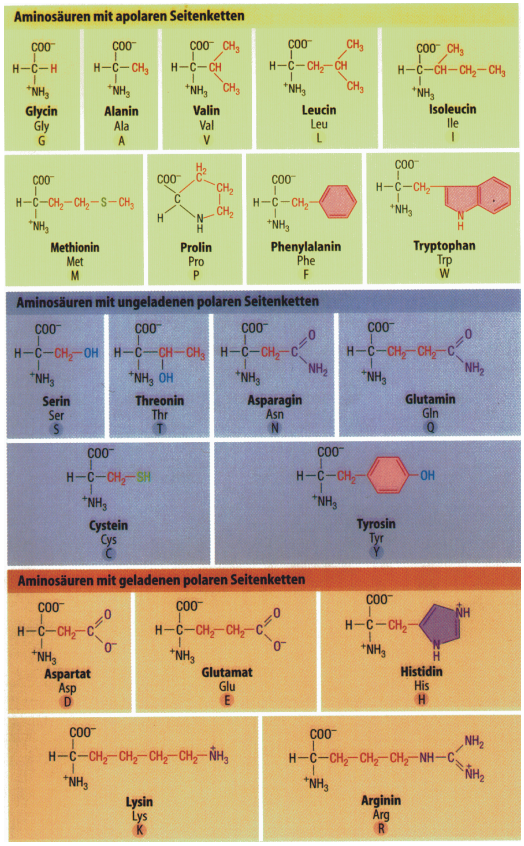
Definition der Solvens Accessible Surface (SAS)



Korrelation zwischen der dem Wasser zugänglichen Moleküloberfläche und der freien Enthalpieänderung bei Übergang von Wasser zu Ethanol/Dioxan

G.E. Schulz, R.H. Schirmer: Principles of Protein Structure, Springer, S. 13

Aminosäure: "Hydrophobizität"



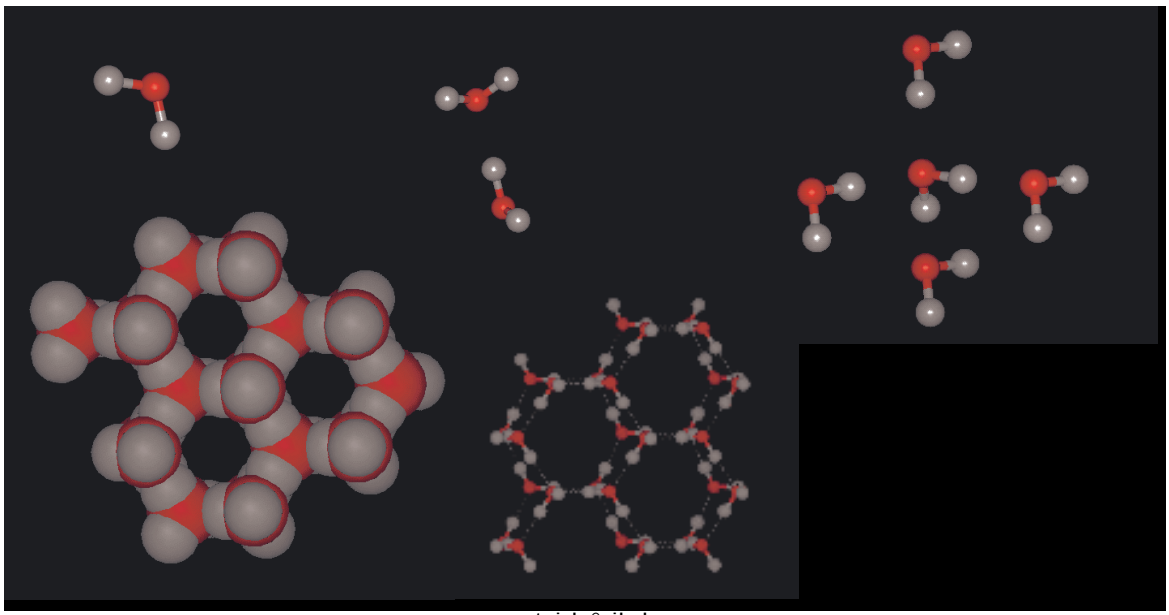
Hydrophobizitäts-Skala /Lesk/

Trp	2.25
Ile	1.80
Phe	1.79
Leu	1.70
Cys	1.54
Met	1.23
Val	1.22
Tyr	0.96
Pro	0.72
Ala	0.31
Thr	0.26
His	0.13
Gly	0.00
Ser	-0.04
Gln	-0.22
Asn	-0.60
Glu	-0.64
Asp	-0.77
Lys	-0.99
Arg	-1.01

steinke@zib.de

Wasser: ein besonderes Lösungsmittel

- besondere physiko-chem. Eigenschaften:
  - Dampfdruck, Siedetemperatur, Verdunstungswärme
  - elektr. Leitvermögen
- Netzwerkstruktur des Wassers



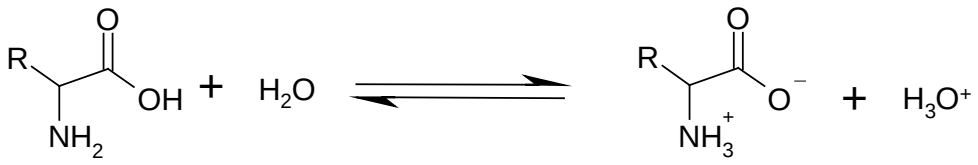
steinke@zib.de

Aminosäuren im Wasser

- Säure-Base-Gleichgewicht im Wasser
  - ionisierbare funktionelle Gruppen:

R-COOH + H_2O ⇆ R-COO^- + H_3O^+ Säure

R-NH_2 + H_2O ⇆ R-NH_3^+ + OH^- Base



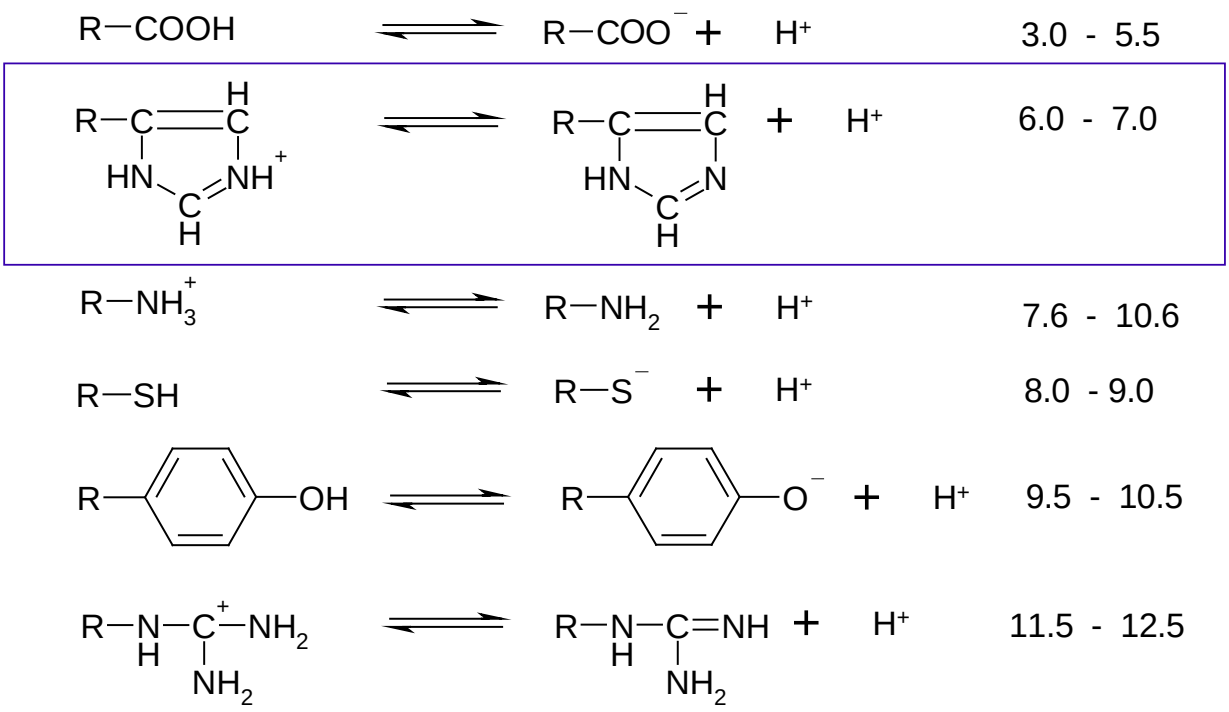
- pK_a – Wert:
  - Dissoziationsgrad (Protolyse)

$$pK_a = -\log K_a$$

$$K_a = \frac{[H_3O^+][B^-]}{[A]}$$

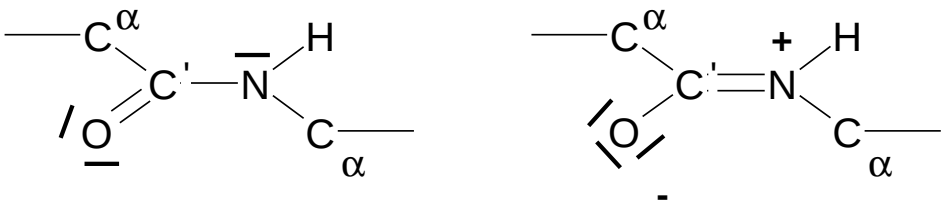
$$pH = pK + \log \frac{[B^-]}{[A]}$$

Charakteristische pK_a-Werte



Peptidbindung

- „Elektronenstruktur“ der Peptidbindung

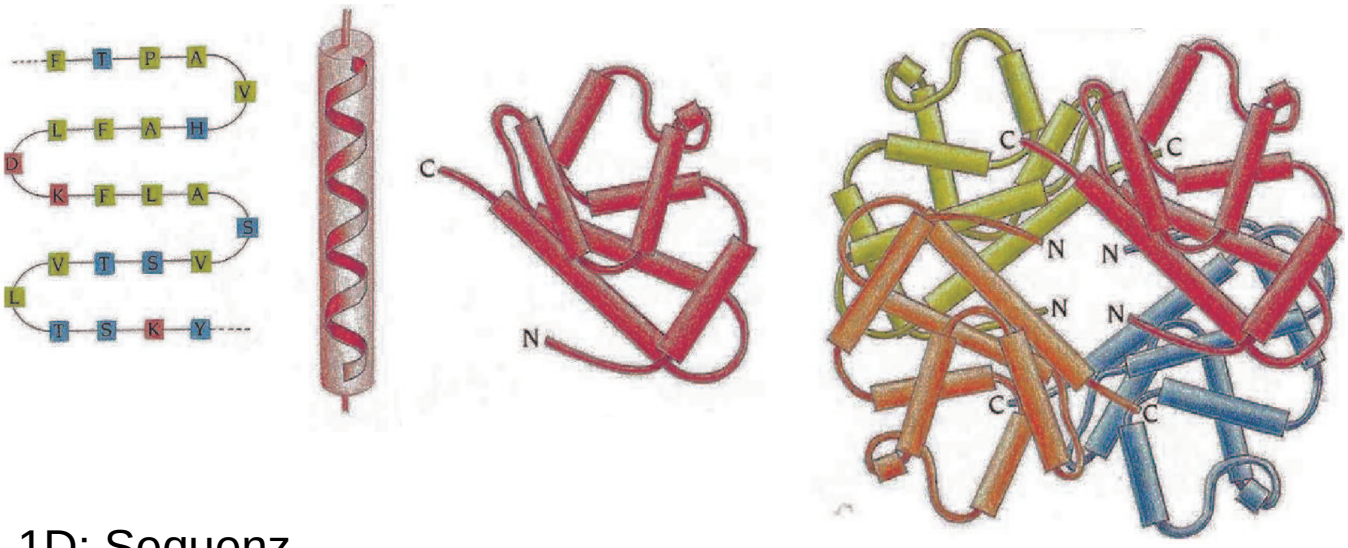


steinke@zib.de

Strukturelemente: 2D

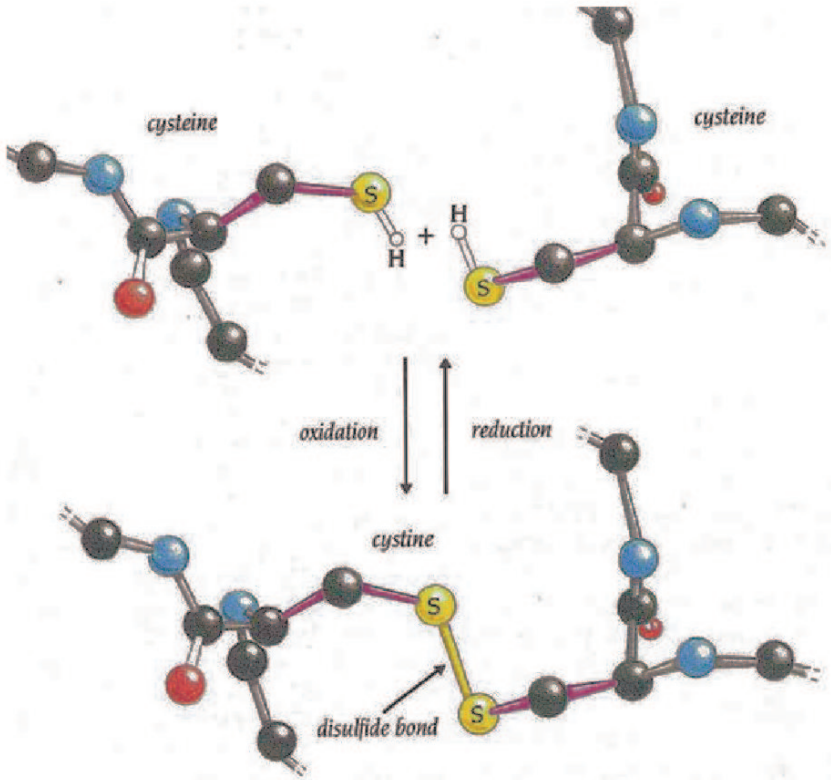
steinke@zib.de

Primär-, Sekundärstruktur: Strukturmotive

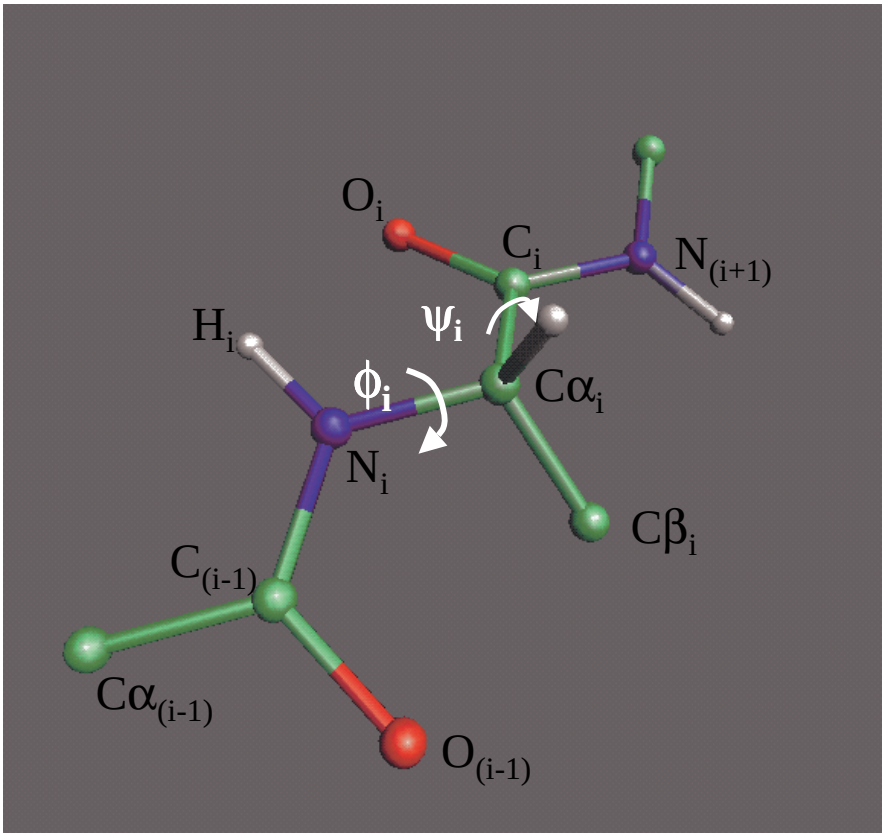


1D: Sequenz
2D: lokale reguläre Struktur
3D: kompakte (globuläre) Einheiten
4D: Molekülaggregate

Cysteine: Disulfidbrücken



Freiheitsgrade in Peptide/Proteine

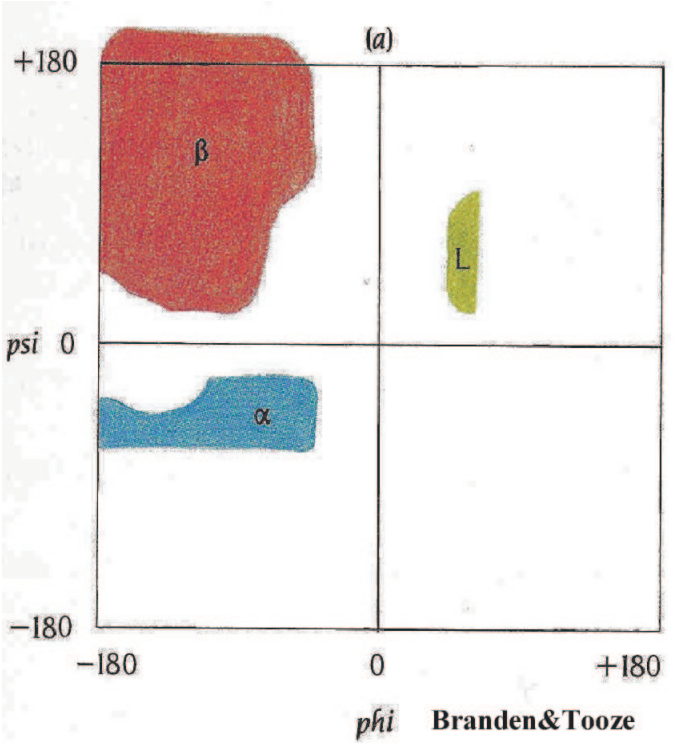


Primärstruktur

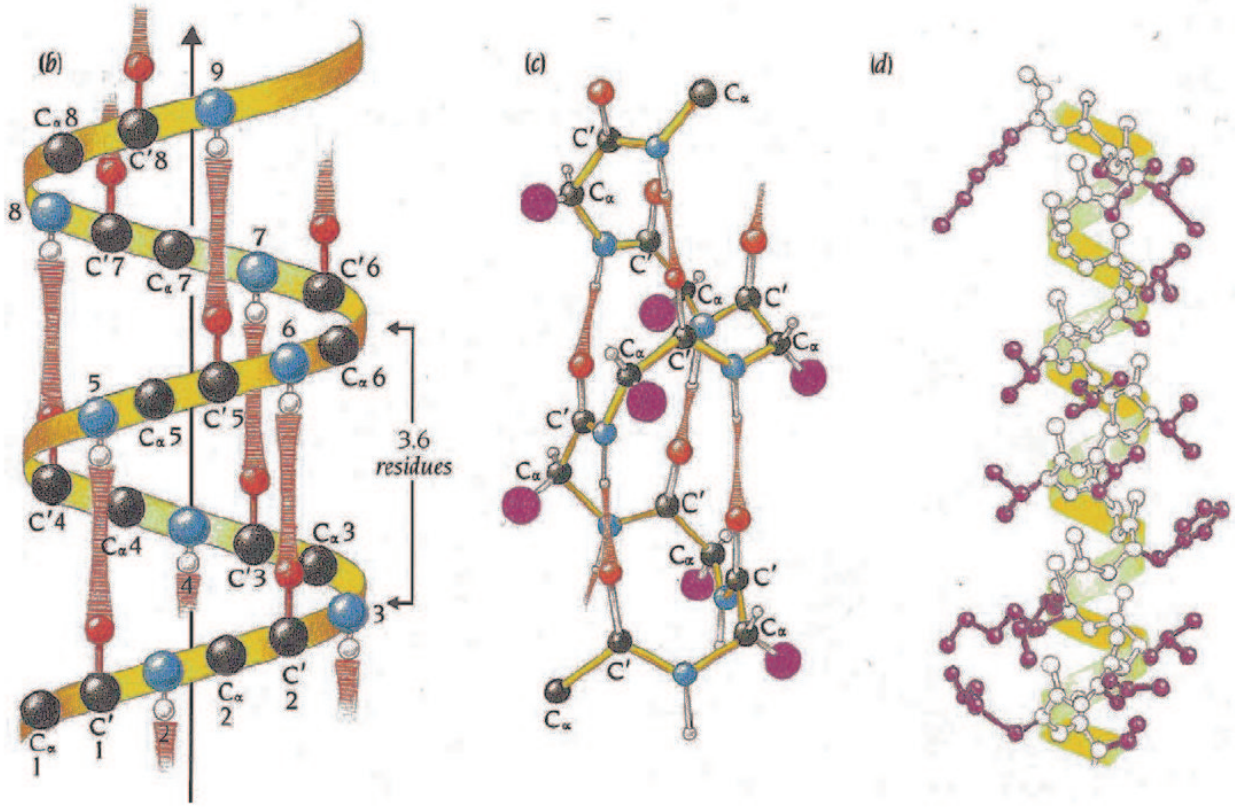
- Peptideinheit
  - planar
  - trans-Konformation (Winkel {O=C-N}, {C-N-H} = 180 Grad)
    - -> Proline?
- Kettenrückrat (*backbone, mainchain*): N-C_α-C
- Bewegungsfreiheit des Peptidkette: Φ, Ψ (Ramachandran)
- Glycin: grösste Bewegungsfreiheit

Sekundärstruktur: Ramachandran-Plot

- Aufteilung in sterisch bevorzugte Φ-Ψ-Kombinationen
- sterisch "erlaubte" Bereiche:
  - α: rechts-drehende α-Helix
  - β: β-Faltblatt
  - L: links-drehende Helix



α-Helix

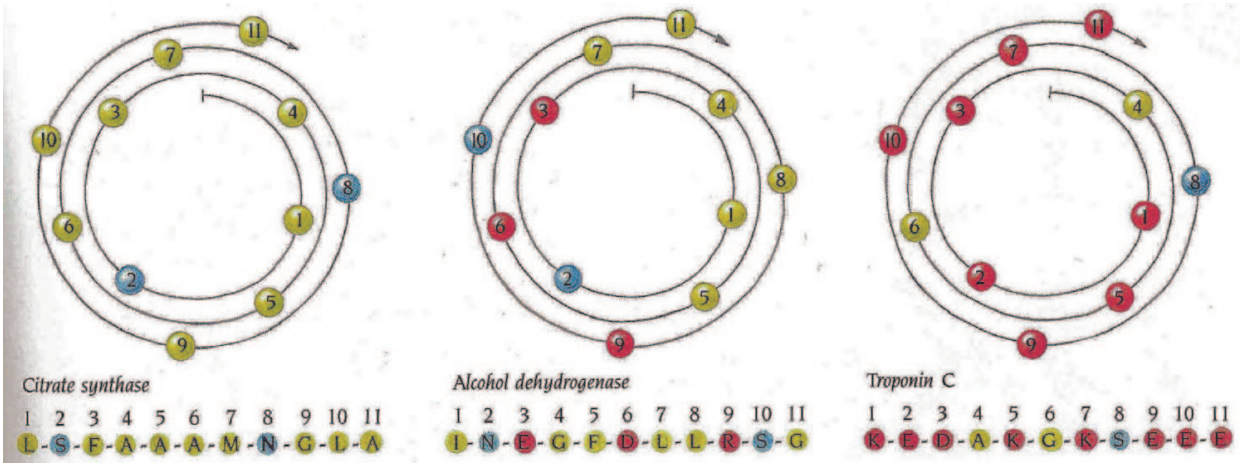


Branden & Tooze

Helix-Typen

- α-Helix
  - HBB: $N_{(i)}$-H ... O=$C_{(i+4)}$
  - Rotation zw. Residuuen ~ 100º , + 1.5 A / turn
- 3_{10}-Helix
  - $N_{(i)}$-H ... O=$C_{(i+3)}$
  - engere Windung
- π-Helix
  - $N_{(i)}$-H ... O=$C_{(i+5)}$
  - lockere Windung, seltener

"Helixrad"

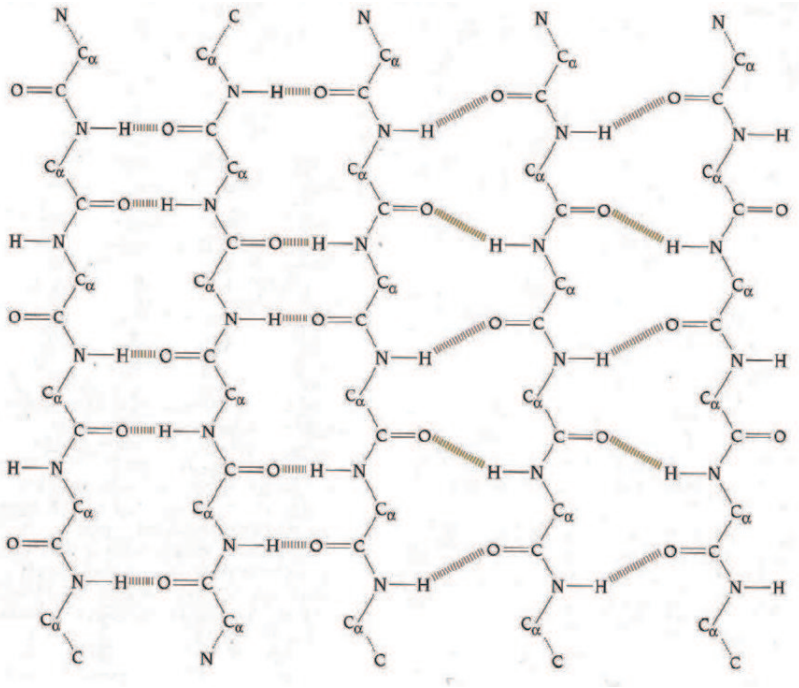


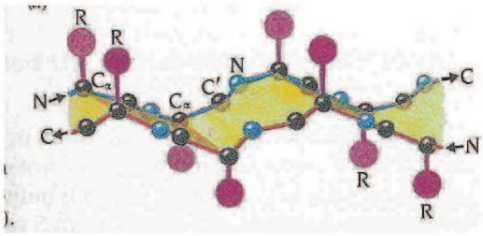
hydrophob
polar
geladen

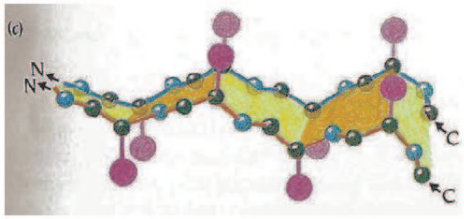
- 100º-Verschiebung des Residuen in verschiedenen α-Helices
- hydrophile/hydrophobe Verteilung je nach Umgebung

Branden & Tooze

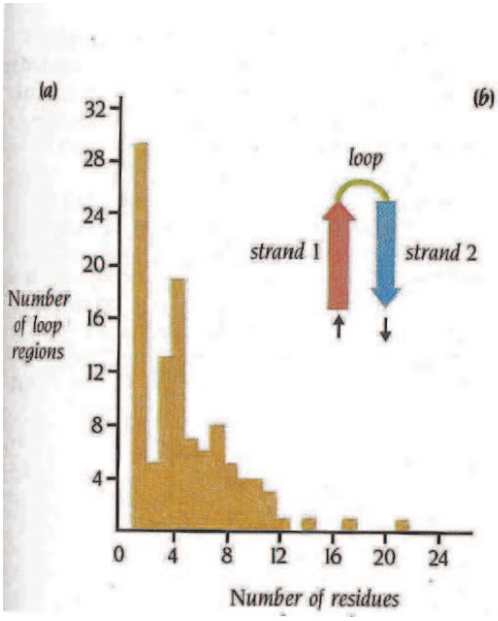
β-Faltblatt

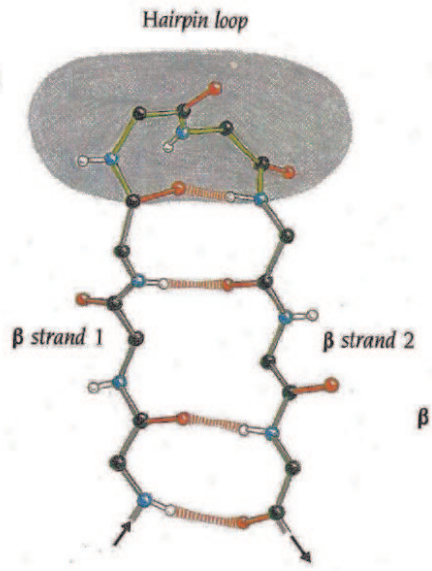


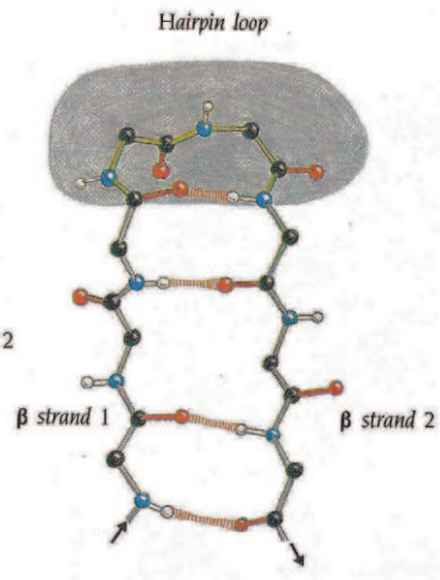




Haarnadel (*hairpin*)

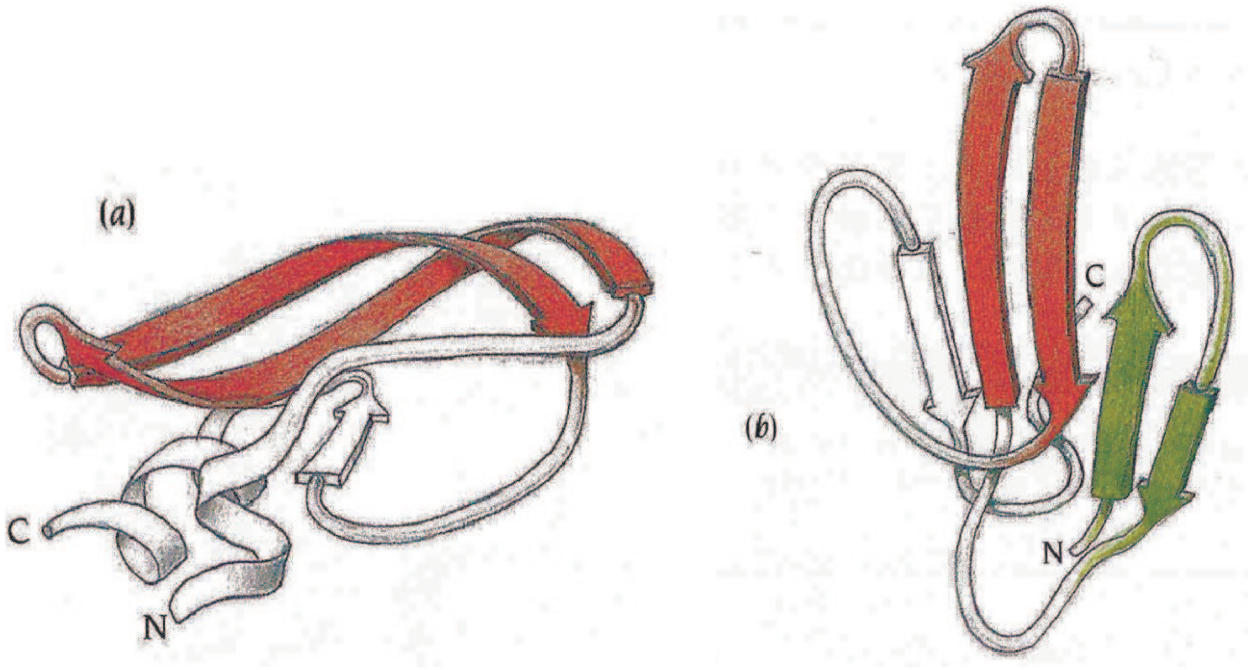






Strukturmotive

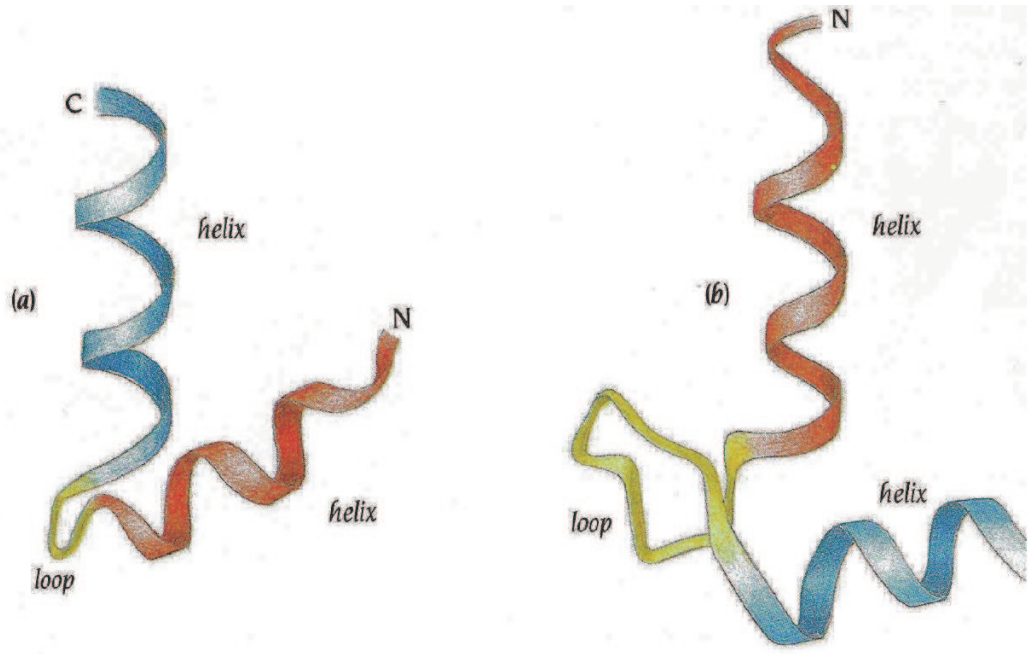
Haarnadelmotiv



keine spezifische Funktion

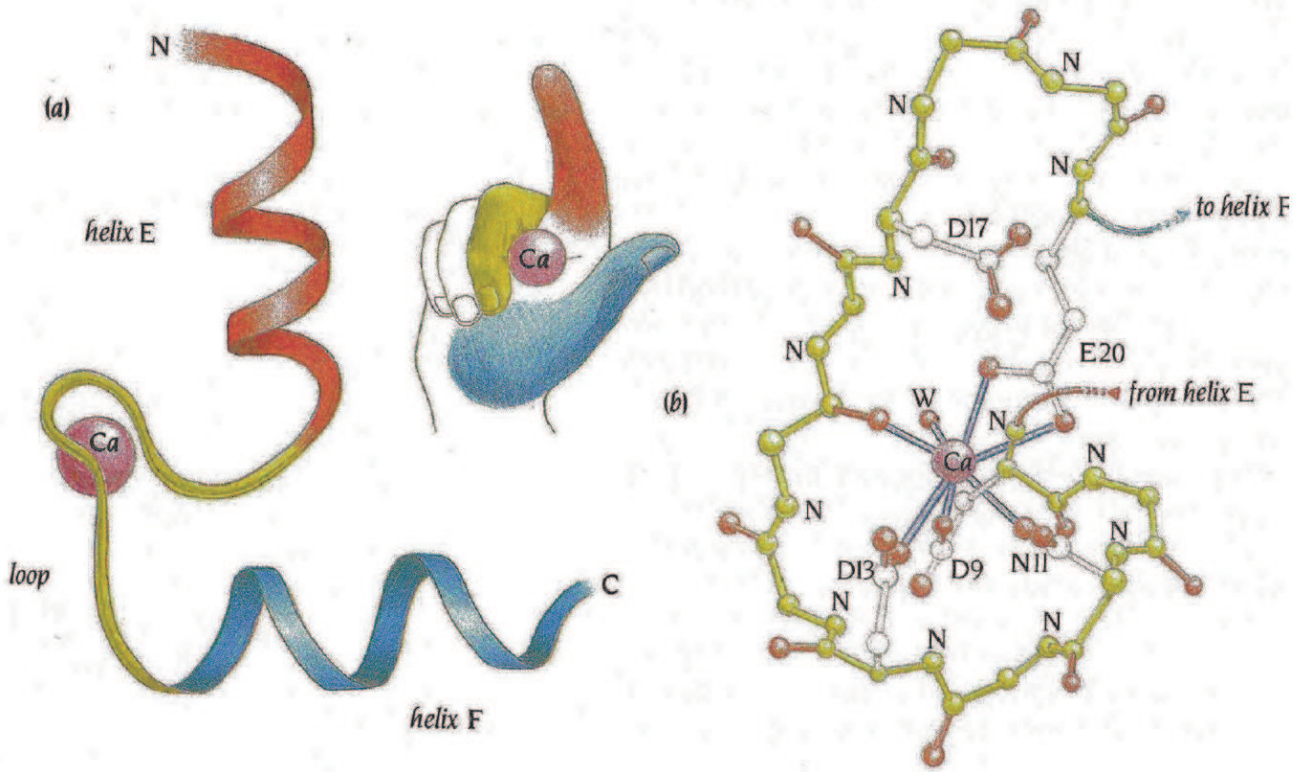
Branden & Tooze

Strukturmotive: Helix-Turn-Helix

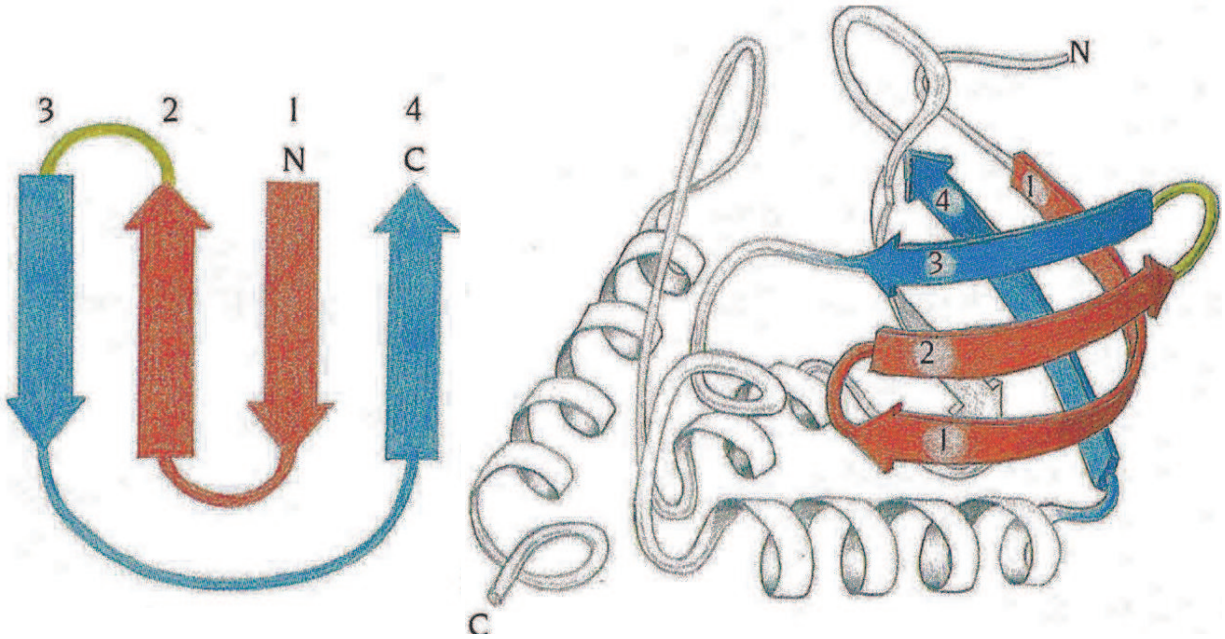


Funktion: z.B. DNA-Bindung

Ca-Bindungsmotiv



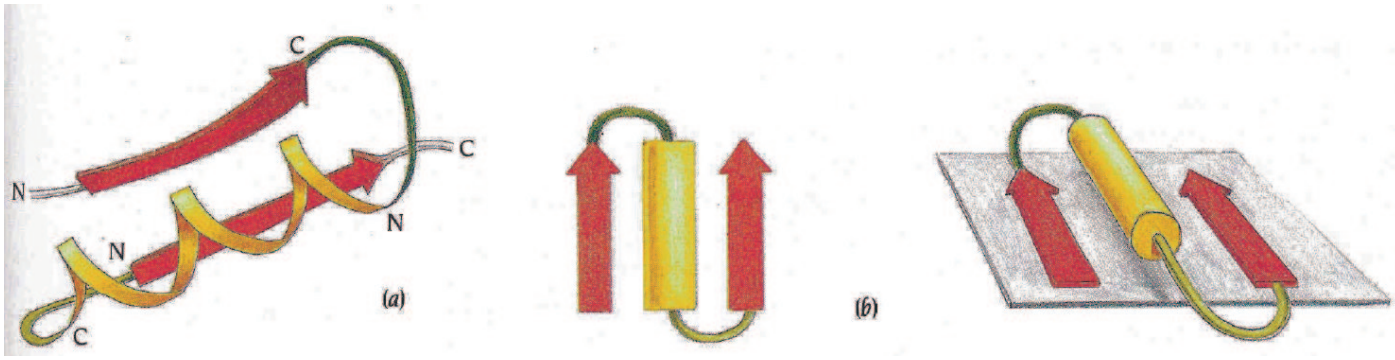
"Greek Key"-Motiv



Staphylococcus Nuklease

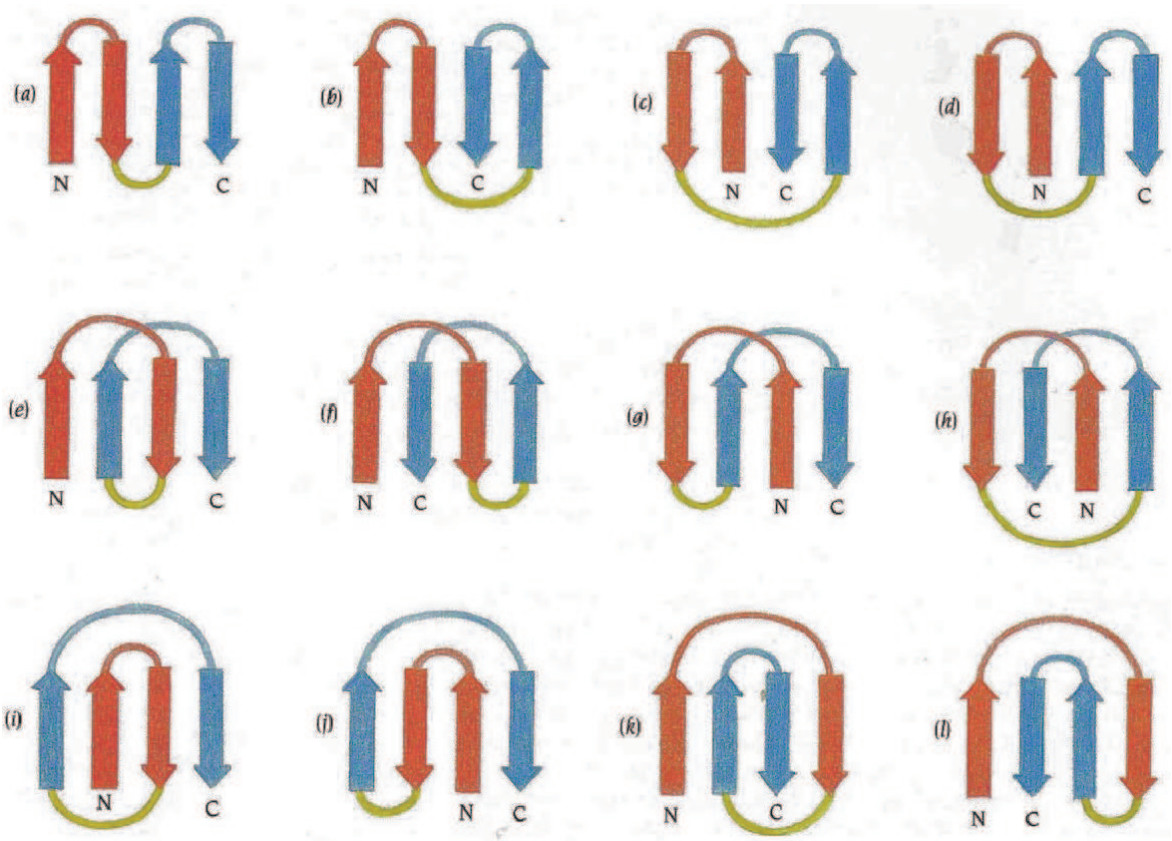
keine spezifische Funktion

βαβ-Motiv



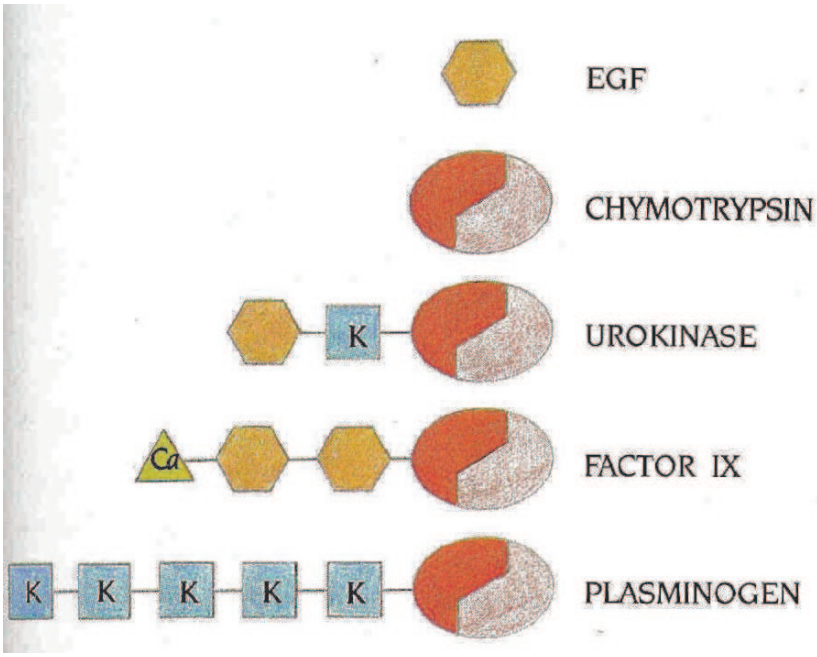
- oberhalb der Faltblattebene: viele Enzyme
- unterhalb der Faltblattebene: bisher nur *Subtilisin*

β-Faltblatt-Topologien



Branden & Tooze

Domänen



homologe Domäne zum epidermalen Wachstumsfaktor

Serinprotease-Domäne

"Kringle"-Domäne mit 3 internen Disulfidbrücken

Kalzium-Bindungsdomäne

Branden & Tooze

Strukturmotive
Klassifikation der Proteine
Beispiele

Hierarchie von Proteinstrukturen (II)

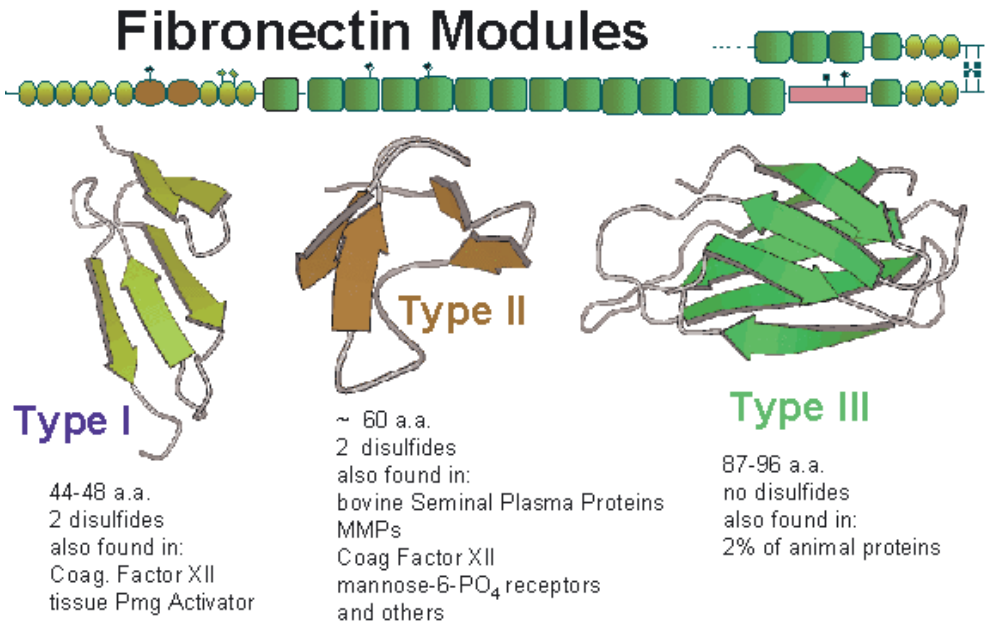
- Supersekundärstruktur: α/β-Bausteine, konsekutiv in 1D
  - β-α-β Einheit
- Domain: zwischen supersekundär u. 3D
  - kompakte Einheit, hydrophob. Kern + polare Oberfläche
  - „unabhängiges“ Faltungsbild einer Sequenz
- modulare Proteine
  - mehrere, z.T. eng-verwandte Domainen
  - Fibronectin: 29 Domainen - $(F1)_6\ (F2)_2\ (F1)_3\ (F3)_{15}\ (F1)_3$

www.gwumc.edu/biochem/ingham/fnpage.htm

Hierarchie von Proteinstrukturen (III)

- Fibronectin: 29 Domainen - $(F1)_6\ (F2)_2\ (F1)_3\ (F3)_{15}\ (F1)_3$

Hierarchie von Proteinstrukturen (III)



- Fibronectin: 29 Domainen - $(F1)_6\ (F2)_2\ (F1)_3\ (F3)_{15}\ (F1)_3$

41
steinke@zib.de

Helices und Faltblätter

- Strukturbildung
  1. möglichst minimale Energie pro Monomer
  2. Stabilisierung durch Ausbildung von HBB-Netzwerke
  3. Ausbildung kompakter, optimal gepackter Gesamtstrukturen

- Ramachandran-Plot → α_R bzw. β-Konformation f. Hauptkette
- helikale (α) / langkettige (β) Strukturen
- Helix-/Faltblatt-Motive mit Loops (*turns*)

Protein-Klassifikation: Helices

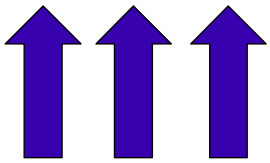
α-Helix := bevorzugter Helixtyp Hydrolase [107L]	**3_{10}-Helix** C-Helix im Myoglobin des Pottwal [1MBO]
π-Helix *photoactive yellow protein* [2PHY]	

β-Faltblätter

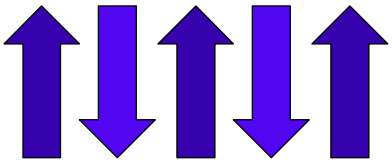
- β-Faltblatt
  - aus mehreren Strängen
  - nicht-konsektiv auf Sequenzebene (Gegensatz zur α-Helix)
- relative Stranganordnung:
  - parallel / anti-parallel / gemischt
- Verknüpfung:
  - anti-parallele Faltblätter: kurzer Loop → β-Hairpin
  - parallele Faltblätter: raumfüllendes Element, z.B. α-Helix
    - β-α-β Einheit

Protein-Klassifikation: Faltblätter

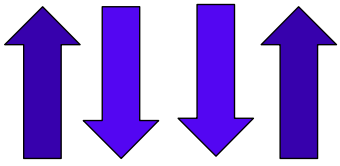
paralleles Faltblatt
[1BRS]



antiparalleles Faltblatt
[1IGM]



antiparallel-paralleles Faltblatt
[1PGA]



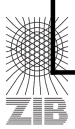
β-Ausbeulung (*bulge*)
[1CHG]

Loops

- Verknüpfung Helix/Faltblatt-Elemente
  - Ausbildung supersekundäre Strukturen
- 1/3 in globulären Proteinen
- funktionelle Aminosäuren
- flexible, daher einfache Konformationsänderung (z.B. Triose-Phosphat-Isomerase)
  - (nicht ausreichend bei allosterische strukturellen Änderungen)

Protein-Klassifikation: Supersekundär

β-*hairpin* [1BHC, 6PTI]	**β–α–β** [8TIM]
helix-turn-helix [1LBM]	***β-barrel*** [1EMA]



Protein-Klassifikation *

Klassen	Charakteristika	Beispiele
α-helikal	(fast) ausschließlich α-helikal	Myoglobin, Cytochrom C, Zitratsynthase
β-Faltblatt	(fast) ausschließlich β-Faltblatt	Chymotrypsin, Immunoglobin-Domäne
α+β	α und β Anteile getrennt keine β–α–β Motive	Papain, Staphylococcen-Nuklease
α/β	β–α–β Motive	
α/β linear	Linien durch Zentren der β-Stränge annähernd linear	Alkoholdehydrogenase, Flavodoxin
α/β-Fass (*barrel*)	Linien durch Zentren der β-Stränge annähernd zirkulär	Triosephosphat-Isomerase, Glycolatoxidase
wenig oder keine Sekundärstruktur		Ferredoxin, Weizenkeim-Agglutinin

* Levitt, Chotia

Katalogisierung von Proteinstrukturen

- SCOP: *Structural Classification of Proteins*
- CATH: *Class, Architecture, Topology, Homologous superfamily*
- FSSP: *Fold classification based on Structure-Structure alignment of Proteins*
- DALI: *Domain Dictionary*

SCOP

- hierarchische Organisation
- evolutionärer Ursprung + strukturelle Ähnlichkeit
- Domainen-basiert, ← PDB
- Domainen → Familien von Homologen
  - Struktur, Funktion, Sequenz
- Superfamilien:
  - ähnlich in Struktur, Funktion
  - keine Sequenzähnlichkeit
- Faltung: Superfamilien ähnlicher Faltungstopologie
- Klassen: Faltungstypen
  - α, β, α+β, α/β, kleine Proteine

http://scop.mrc-lmb.cam.ac.uk/scop/

SCOP: Statistik (Jan/2000)

Level	Anzahl Einträge
Klassen	7
Faltungen	520
Superfamilien	771
Familien	1212
Domainen	21529

SCOP: Flavodoxin

Protein: Flavodoxin from *Escherichia coli*

Lineage:

1. **Root: scop**
2. **Class: Alpha and beta proteins (a/b)**
   Mainly parallel beta sheets (beta-alpha-beta units)
3. **Fold: Flavodoxin-like**
   3 layers, a/b/a; parallel beta-sheet of 5 strand, order 21345
4. **Superfamily: Flavoproteins**
5. **Family: Flavodoxin-related**
   binds FMN
6. **Protein: Flavodoxin**
7. **Species: *Escherichia coli***

CATH

- Hierarchie-Level: Klassen, Architektur, Topologie, homologe Superfamilien
- ähnl. in Struktur, Sequenz + Funktion → Sequenzfamilie
- homologe Superfamilie: gemeinsame Abstammung
- Topologie/Faltungsfamilie: homolog. SF mit gleicher räuml. Anordnung *und* Abfolge/Konnektivität der 2D-Elemente
- Architektur: ähnl. Anordnung d. 2D-Elemente, *aber* versch. Konnektivität
- Klassen: α, β, α+β, α/β, Domaine m. wenig 2D

http://www.biochem.ucl.ac.uk/bsm/cath/

DALI + FSSP (I)

- DALI [Holm, Sander]:
  - Vergl. von Proteinstrukturen (Struktur-Struktur-Alignment)
    - gemeinsame Substruktur(en), Alignment gemeinsamer Residuen
  - Erkennung entfernter Ähnlichkeiten
  - schnell, kompletter PDB-Scan möglich

- DALI → PDB liefert:
  - FSSP := Fold classification based on Structure-Structure alignment of Proteins)
  - DALI Domain Dictionary

DALI + FSSP (II)

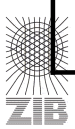
- FSSP
  - Clustering von Ketten > 30 Residuen mit 25% Sequenzähnlichkeit
  - Auswahl eines Repräsentanten pro Cluster
  - DALI klassifiziert Ähnlichkeiten zwischen Repräsentanten
  - FSSP-Eintrag:
    - dessen Alignment mit Strukturhomologen
    - struktur-äquivalente Residuen
- DALI *Domain Dictionary*
  - automat. extrahierte, häufige Domainen

Protein-Klassifikation & Evolution

- Die Evolution selektiert über die Funktion von Proteinen.
- Die Klassifikationsschemata sind eng an geometrische Ähnlichkeitskriterien geknüpft. Diese müssen nicht im Zusammenhang mit Funktion stehen.

α-helikale Globine: Aggregate

Monomer: Glycera-Hämoglobin [1HBG]	**Dimer:** Bluthemmerprotein (*ark clam globin*) (*Scapharca inaequivalvis*) [4SDH]
Tetramer: Humanes Hämoglobin [4HHB]	

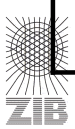


Protein-Klassifikation: α-helikale Globine

- SCOP enthält 70 Faltungen für α-helikale Proteine
- PDB enthält über 200 Globin-Strukturen von 30 Spezies, die α-helikal sind z.B. Hämoglobin, Myoglobin:
  - liegen als Monomere, Dimere, Tetramere und höhere Aggregate vor
  - Monomere binden an Häm-Gruppe
  - Quartärstruktur höherer Aggregate ermöglicht allosterische Eigenschaften
- Cytochrom C bindet Häm-Gruppe, ist α-helikal, aber kein Globin
- α-helikale Proteine müssen nicht klein sein (s. Ras-GTPase)

α-helikale Proteine (Beispiele)

Cytochrom C v. Thunfisch [5CYT]	Cytochrom C v. Reis [1CCR]
GCN4-bZIP (*leucin zipper*) [2DGC]	Bacterio-Rhodopsin Membranprotein mit parallelen Helices [1BRD]



Strukturprinzipien: α-helikale Proteine

F Wieviele Möglichkeiten gibt es, 4 paarweise antiparallele Helices anzuordnen?

in := in die Ebene
out := aus der Ebene
die Zahlen nummerieren die Helices

2-in 1-out 3-out 4-in **A**	2-in 1-out 4-in 3-out **B**	3-out 1-out 2-in 4-in **C**
4-in 1-out 3-out 2-in **D**	3-out 1-out 4-in 2-in **E**	4-in 1-out 2-in 3-out **F**

steinke@zib.de

Strukturprinzipien: α-helikale Proteine

F Wieviele Möglichkeiten gibt es, 4 paarweise antiparallele Helices anzuordnen?

- Fast alle α-helikale Typen besitzen Topologie A oder D!
- Beobachtung: A und D haben keine Diagonalverbindungen zwischen Helices. Alle nächst-benachbarten Helices sind antiparallel.

in := in die Ebene
out := aus der Ebene
die Zahlen nummerieren die Helices

2-in 1-out 3-out 4-in **A**	2-in 1-out 4-in 3-out **B**	3-out 1-out 2-in 4-in **C**
4-in 1-out 3-out 2-in **D**	3-out 1-out 4-in 2-in **E**	4-in 1-out 2-in 3-out **F**

60
steinke@zib.de

Strukturprinzipien: α+β - Proteine

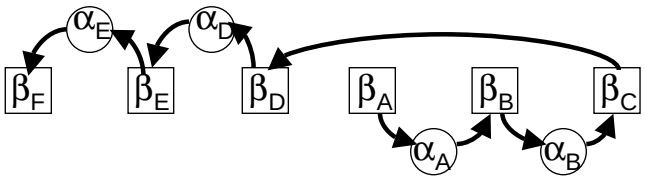
Staphylococcen-Nuclease [2SNS]	Humanes Ubiquitine-konjugierte Enzym [1U9B]

Strukturprinzipien: α/β - Proteine

Lineare / offene β–α–β Proteine

NAD-Bindungsdomäne der Pferde-leber-Alkoholdehydrogenase
[6ADH]

- Nukleotid bindende Proteine: enthalten Domäne mit bis 6 β-α-Einheiten folgender Topologie:



geschlossenes β–α–β Fass (barrel)

Glycolate Oxidase [1GOX]

- 8 β-α Einheiten → zylindrische Topologie, Helices außen
- entdeckt 1975 an **T**riosephosphat-**I**so**m**erase des Huhns (TIM) → Name *TIM barrel*
- heute 40 Enzyme bekannt
- aktives Zentrum am Fassende im Bereich der C-Termini der Stränge
- gemeinsamer Vorfahr nicht nachweisbar

Strukturprinzipien: (α/β), TIM-Barrel

- Faltblatt aus 8 parallelen β-Strängen
- Stränge bilden eine 36^{o} Winkel zur Helixachse
- Helices sind annähernd parallel zu Strängen
- aus Sicht des aktiven Zentrums geht die Kette gegen den Uhrzeigersinn um das *barrel* nach dem Muster:
  - Strang *auf* - Helix *ab*
  - Ausnahmen: fehlende Helix oder invertierter Strang [Enolase])
- Seitenketten der Stränge zeigen abwechselnd mit den Residuen in das Innere oder nach außen, die Inneren sind in drei Ebenen angeordnet

β-Barrel: Geometrieparameter (TIM-Barrel)

a = C_α-C_α Abstand	3.3 A
b = Abstand zweier Stränge	4.4 A
n = Anzahl der Stränge	8
S = Scherzahl (shear)	8
R = Radius = $[(Sa)^2 + (nb)^2]^{1/2} / [2n \sin(\pi/n)]$	6.5-7.5 A
α = Winkel Stränge-Barrel-Achse; tan α = Sa/(nb)	36^{o}

[1GOX]

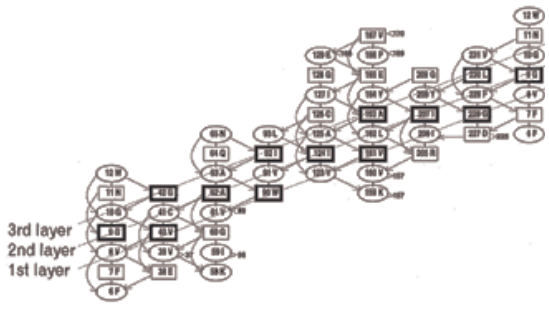
$S = n = 8$

1tim chain A

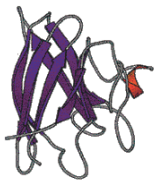
str. No. 1 2 3 4 5 6 7 8 1'

$S = n = 8$

1tim chain A



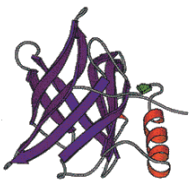
str. No. 1 2 3 4 5 6 7 8 1'



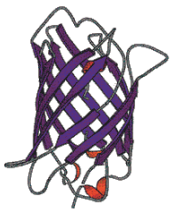
3sod
(n,S)=(8,6)
1 family

1timA
(n,S)=(8,8)
16 families

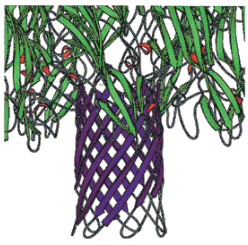
1avdA
(n,S)=(8,10)
3 families



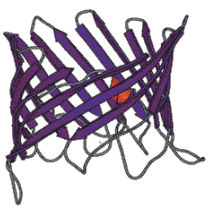
1brp
(n,S)=(8,12)
1 family



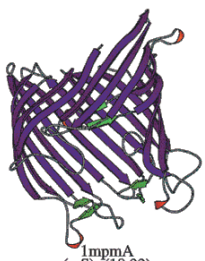
1ema
(n,S)=(11,14)
1 family



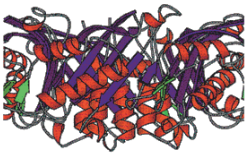
7ahl
(n,S)=(14,14)
1 family



1prn
(n,S)=(16,20)
1 family



1mpmA
(n,S)=(18,22)
1 family



1gtpABCDE domain 2
(n,S)=(20,20)
1 family

65
steinke@zib.de

Strukturprinzipien: irreguläreStrukturen

Stabilisierung durch **Disulfidbrücken** Agglutinin [9WGA] (Weizenkeim)	Stabilisierung durch **Eisen-Schwefel-Cluster** Ferrodoxin [6FD1]
Stabilisierung durch **Disulfidbrücken + kurze 2-strängige Faltblätter** „Kringle“ Domäne in Prothrombin [2PF1]	

Referenzen

- C.-I. Branden, J. Tooze: Introduction to Protein Structure
- Higgins, Taylor (ed.): Bioinformatics, Oxford Uni Press, 2000

Proteine: Struktur, Modellierung, Dynamik

Algorithmische Bioinformatik
WS 2002

Thomas Steinke
Zuse Institute Berlin (ZIB) <www.zib.de>
Berlin Center for Genom Based Bioinformatics (BCB) <www.bcbio.de>
steinke@zib.de

Strukturvorhersage
- Einführung -

Struktur: Warum so wichtig?

- Struktur bestimmt Funktion
- Struktur besser konserviert als Sequenz
- Struktur →
  - Verständnis biomolekularer Wechselwirkungen
  - Rationales Drug-Design
  - Protein-Engineering

Experiment vs. „Vorhersage“

Experimentell:

- Röntgen-Kristallographie (X-Ray)
- Kernmagnetische Resonanzspektroskopie (NMR)
- Neutronenbeugung
- Kryoelektronenmikroskopie (Cryo-EM)
- Zirkulardichroismus (CD)

Theoretisch:

- 2D-Strukturvorhersage
- *Threading*
- *Homology Modeling*
- *Ab-initio* Strukturvorhersage
- Docking: Protein-Protein, Protein-DNA, Protein-Drug

Wege zur (Protein-) Struktur

X-ray	nD-NMR	*n*-Beugung
✓ Auflösung	✓ in Lösung	✓ H-Atome
✗ Zeitaufwand (Kristallzüchtung)	✓ Dynamik	✗ Kristallgrösse
	✗ Sequenzlänge	

Simulation

✓ allg. Ansatz

✗ konkurrenzfähig?

Vor- & Nachteile

- Kristallographie
  - hohe Genauigkeit
  - Kristall notwendig (ca. 20 mg Material) → Zeit
- NMR
  - schnelle Ermittlung flexibler & starrer Regionen
  - beschränkt auf ca. 120 Residuen
  - Protein muss löslich sein (ca. 30 mg/ml)
- Risiken bei exp. Strukturbestimmung (IBM, BlueGene Info)
  - Kosten: ca. $100K
  - Personal: > 1 Post-doc Jahr
  - Erfolgsaussichten: beschränkt

Faltungstruktur & Funktion

- Beobachtung I: Reversibilität (Beispiel Ribonuklease)
  - Dauer d. Rückfaltung: ~ Sekunden

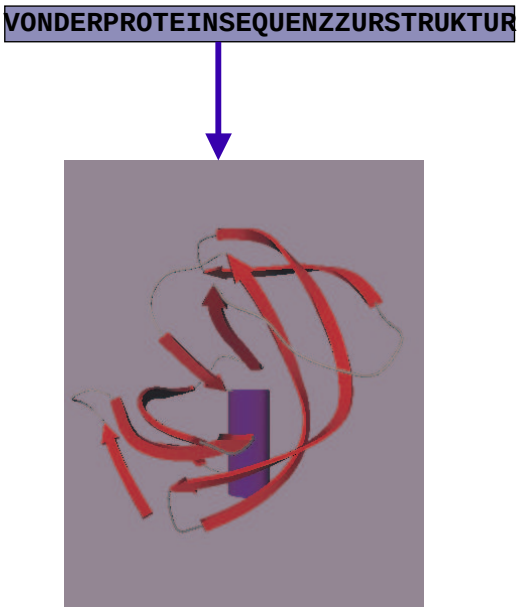
native RN	Hitze, Reduktion →	entfaltete RN	Oxidation →	native RN
voll aktiv definierte 3D-Struktur		inaktiv zufällige Knäulung		voll aktiv definierte 3D-Struktur

→ **Theorem: Struktur ist eindeutig durch AA-Sequenz definiert**

- Beobachtung II: einige Proteine benötigen Hilfsmoleküle (*chaperons*)
- Beobachtung III: Faltung in andere Konformation (Prione)

Das Faltungsproblem

(1D) Sequenz → (3D) Struktur



- intensive Forschung seit 1950
- Vielzahl von Verfahren/Modifikationen
- manuell vs. automatisch
- freie Enthalpie ΔG
  - "das richtige unversielle" Potential für ΔG
  - Minimum = native Struktur?
- Dimensionsproblem
  - 100 Residuen
  - Torsionswinkel Φ, Ψ: je 3 Werte
  - → 3^{198} Konfigurationen ($\approx 3 \cdot 10^{94}$)

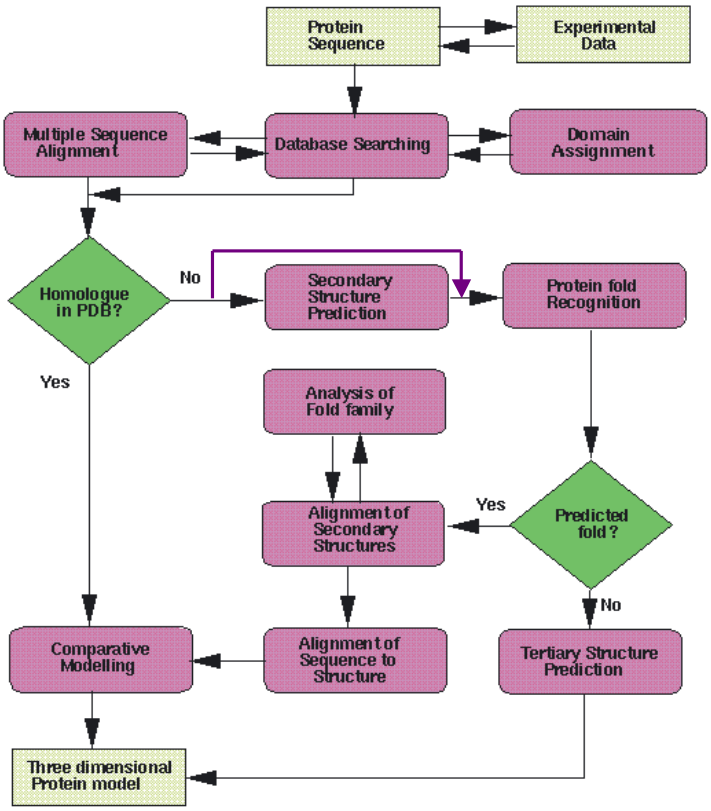
Faltungsproblem: Status

- Min (ΔG) ≡ native Struktur unklar
  - korrektes Potential, Rechenaufwand

→ keine eindeutige Abbildungsvorschrift bekannt

- allg. "Energieterme": Kontakt, hydrophob/hydrophil
- heuristische Methoden
  - Neuronale Netze (2D-Vorhersage)
  - Mustererkennung aus statistischer Analyse bekannter 3D's

Strukturvorhersage: Vorgehen



http://www.bmm.icnet.uk/people/rob/CCP11BBS/flowchart2.html

2D-Vorhersage

steinke@zib.de

Vorhersage der Sekundärstruktur

- Proteine: 3 Zustände:
  - α-Helix
  - β-Faltblatt
  - Loop
- Beispiel:

```
          MKFIIAFFVATLAVMTVSGEDKKHDYQNEF
          *HHHHHHHHHHHHHHHHLLLLLLLHHHHHH
     Helix1                Helix2
```

- Treffgenauigkeit: <= 80%

steinke@zib.de

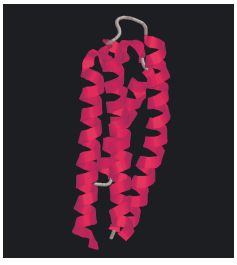
Sequenzähnlichkeit: Begriffe

- **Ähnlichkeit**: Grad der Übereinstimmung in bestimmten gemeinsamen Komponenten in zwei Objekten
- **Homologe**: Ähnlichkeit mit gemeinsamen Urspung
  - **Orthologe**: Homologe mit konservierter Funktion
  - **Paraloge**: Homologe mit divergierender Funktion
- **Analoge**: Ähnlichkeit aufgrund konvergierter evolutionärer Entwicklung

- ***Sequenzähnlichkeit***: zwei Sequenzen enthalten identische oder in Beziehung stehende Zeichen in entsprechenden Sequenzpositionen

Primäranalyse der Proteinsequenz

- Sequenz aus Gen-Vorhersage: → multiple Domainen (?)
- Seqment-Analyse:
  - Transmembran-Protein:
    - Transmembran-Segmente separieren extra- u. intra-zelluläre Domainen
    - TMHMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/)
  - Mehrfach-Helix-Domainen (*coiled coils*)
    - COIL-Srv
      http://www.ch.embnet.org/software/COILS_form.html)



Erkennen von Domainen

- Homologie zu bekannten Sequenzen nur in Teilbereiche
  - SMART (Simple Modular Architecture Research Tool)
    http://smart.embl-heidelberg.de/)
  - PFAM (Protein families database of alignments and HMMs)
    www.sanger.ac.uk/Software/Pfam/
- Domain-Separatoren: n* Pro/Glu/Ser/Thr
- Sekundärstrukturvorhersage: → Zugehörigkeit von 2D-Elementen zu bestimmten Domainen

Sequenzsuche in Datenbanken (I)

- → Sequenzhomologe
  - BLAST (Basic Local Alignment Search Tool)
    @NCBI: www.ncbi.nlm.nih.gov/BLAST/
    @EBI: www.ebi.ac.uk/blast2/
  - gapped BLAST, PSI-BLAST (position specific iterated BLAST)
    - empfindlicher (weniger false positive)
    - finden entfernter Homologie durch iterative Konstruktion von Profilen und erneuter DB-Suche nach neuen Homologen
  - FASTA
    http://alpha10.bioch.virginia.edu/fasta/

Sequenzsuche in Datenbanken (II)

- Verwendung von Informationen anhand mehrfacher Sequenzen-Vergleiche
  - Steigerung der Empfindlichkeit
- Konstruktion von Profilen aus multiplen Sequenz-Alignment
- Profil: Bewertung jeder Aminosäure in jeder Position
  - PSI-BLAST
    www.ncbi.nlm.nih.gov/BLAST/
  - HMMER: Profile Hidden Markov Models
    http://hmmer.wustl.edu/

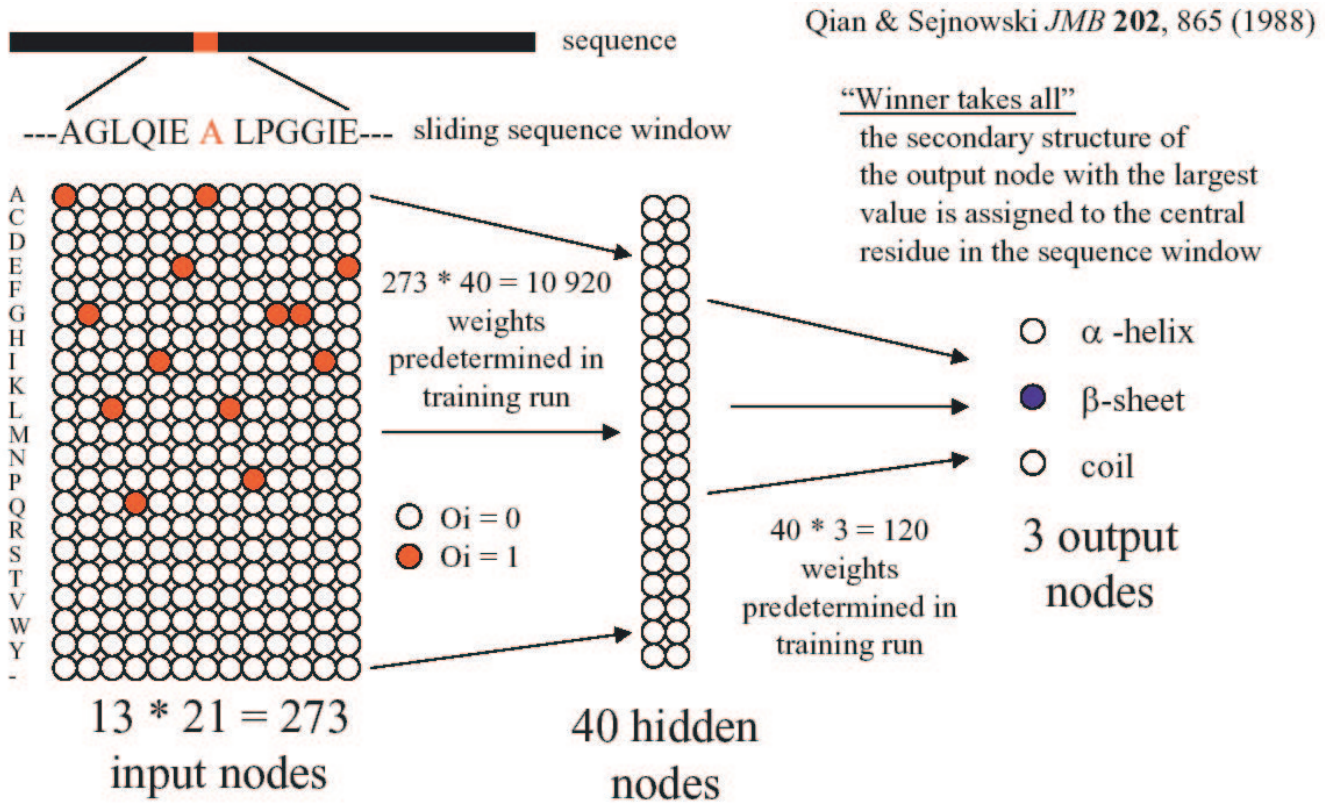
Nutzung von Sequenz-Motiven

- nur die invarianten Positionen im Alignment → konservierte Sequenzen (Familie) [Signature]
- Beispiel:
  `H-[FW]-x-[LIVM]-x-G-x(5)-[LV]-H-x(3)-[DE]`
  Familie von DNA-bindenen Proteinen
  - PROSITE: www.expasy.ch/tools/scanprosite/

2D-Vorhersage für einzelne Sequenzen

- nur einzelne Sequenzen, keine Familien-Daten
  - Chou-Fasman
  - Garnier, Osguthorbe & Robson (GOR)
  - 56-60% Genauigkeit
- Daten von homologen Sequenzen + neue Analyse Techniken (Neuronale Netze)
  - 70 ... 80%%
  - genaue Vorhersage der Core-Regionen
- automatisierte Methoden
  - PHD, ZPred, NNSSP, ...

Beispiel (Qian, Sejnowski)



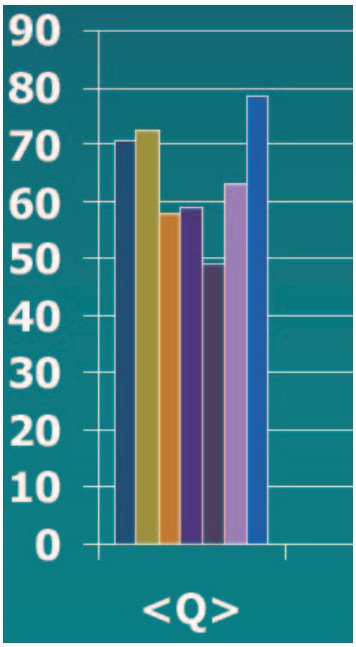
PHD3: Multiples Alignment + NN

- Rost, Sander 1994
- multiples Sequenz-Alignment
- 12 NNs mit versch. Gewichten
- aus 12 versch. Trainigsverfahren

```
Sequence1 AGLQIE A LPGGIE
Sequence2 GGVNLE A IPAPID
Sequence3 AGINID A GAGPID
Sequence4 AGWQIT A CGAPIE
```

A	75	0	0	0	0	0	100	0	25	50	0	0	0
C	0	0	0	0	0	0	0	0	25	0	0	0	0
D	0	0	0	0	0	25	0	0	0	0	0	0	50
E	0	0	0	0	0	50	0	0	0	0	0	0	50
F	0	0	0	0	0	0	0	0	0	0	0	0	0
G	25	100	0	0	0	0	0	0	25	25	50	25	0	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	25	0	75	0	0	25	0	0	0	100	0
K	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	25	0	25	0	0	25	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	50	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	50	0	75	0	0
Q	0	0	0	50	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	25	0	0	0	0	0	0	0
V	0	0	25	0	0	0	0	0	0	0	0	0	0
W	0	0	25	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0
-	0	0	0	0	0	0	0	0	0	0	0	0	0

Genauigkeit von 2D-Vorhersagemethoden



- PhD (Rost, Sander)
- PhD3 (Rost, Sander)
- COMBINE
- GORIII (Garnier, Osguthorpe, Robson)
- Chou / Fassman
- PrISM (Yang)
- PSIPRED (Jones)

/Cheng/

Struktur-Sequenz-Alignment (Threading)

- ähnliche Sequenz → ähnliche Struktur **?**
  - wenn ja: 1D(A) ≈ 1D(B) → 3D(A) ≈ 3D(B)
  - leider nicht immer!

- Beobachtung: ähnliche Domäne aber unähnliche Sequenz
  ← evolutionäre Selektion nach Funktion

→ **Inverses Faltungsproblem:** Fit von bekannter 3D auf 1D

Faltungserkennung (Threading)

- Ziel: Erkennung *des* Faltungstyps aus 500+ (-700) bekannten Faltung (PDB: 20 000 Strukturen)
- paarweise Sequenz-Ähnlichkeit zu bekannten Strukturen: <= 30%

steinke@zib.de

Faltungserkennung

- „Threading“ 1992 (Jones)
- 10 Superfolds in 50% ähnlicher Superfamilien
- → Suche nach dem Faltungstyp in bekannten DBs (Faltungserkennung)
- Entdeckung von Strukturähnlichkeiten ohne Sequenzähnlichkeit

- bislang mit Threading vorgeschlagene Strukturen zu ungenau für atomistische Simulationen (Drug Design)

steinke@zib.de

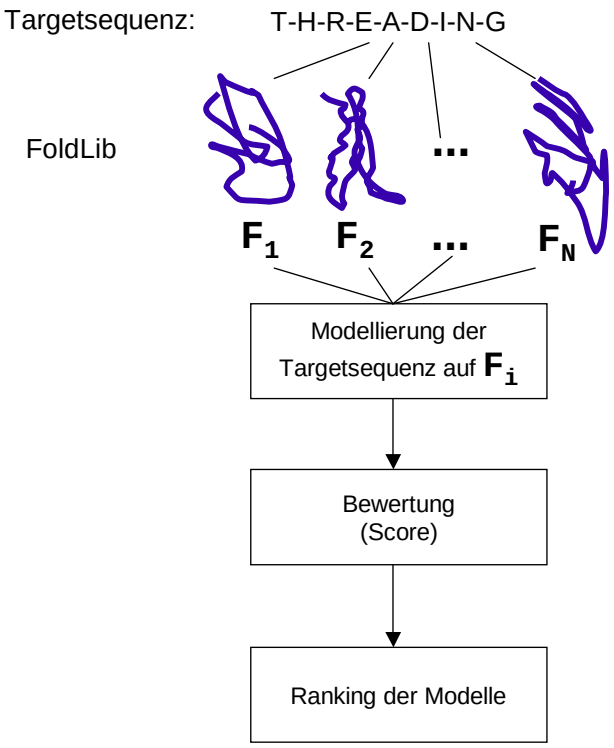
Threading

1. Bibliothek (FoldLib)
   - repräsentative Strukturen: Ketten, Domänen, konservierte Proteinkerne
   - Erstellung des Strukturmodells

2. Mappen/Alignen der Targetsequenz auf jeden Repräsentanten der FoldLib (Sequenz-Struktur-Alignment)
   - InDel in Loops
     - Dynamische Programmierung
     - Gibbs-Sampling
     - Branch&Bound
   - → Bewertung

3. Bewertungsfunktion
   - Pseudo-Energie

4. Auswahl nach Ranking



steinke@zib.de

Threading: Faktoren für Güte

- Auswahl der Faltungsrepräsentanten F_i
- Art der Bewertungsfunktion
- Alignment der Sequenz auf die Struktur der F_i, Optimierung
- Auswahl des „besten Fits“

steinke@zib.de

1D-3D Profile (Bowie et.al., 1991)

- Faltung charakterisiert durch Umgebung:
  - lokale 2D-Elemente: α, β, Loop
  - Lösungsmittelzugänglichkeit: verdeckt, teilw. verdeckt, offen
  - Anteil polarer Atome im verdeckten Core
- Annahme: Umgebung eines Residuums mehr konserviert als der Residuentyp selbst
- 3D übersetzt in 1D Zeichenkette
  - Alignment mit DP

\+ Ähnlichkeiten bei entfernten Sequenzen

-- Ähnlichkeit strukturell divergenter Sequenzen?

-- konvergierte Strukturen

Threading (Jones et.al.,1992)

- Faltung: modelliert als „Netzwerk“ von paarweisen atomaren Wechselwirkungstermen
- Spezifität eines Residuums erhalten (anstelle mittlere Umgebung)
- → keine einfache DP mit Zeichenketten
- Alignment einer Sequenz mit Backbone-Struktur

  → doppelt dynam. Programmierung (DDP) [Jones et.al.]

1D – 3D – Alignment (Jones et.al.)

- Problem: Alignment einer Targetsequenz mit Backbone-Koordinaten eines Templates; mit paarweisen WW
- Umgebungspotential e. Residuums
- v_{ij}: aus DB-Daten statistische Potentiale (Sippl):

$$V_i = \sum_{j \neq i} v_{ij}$$

  - Sequenzabstand k <= 10 / 11<= k <= 30, k > 30
    - 2D-Matching / Super-2D Motife / Tertiärpackung
  - Atompaare: C_β-C_β, C_β-N, C_β-O, N-C_β, N-O, O-C_β, O-N
  - v_{ij} ← Atompaar *i,j* ; k ; Abstand R_{ij}

    → „Energie“: ~ Wahrscheinlichkeit der betrachteten Einbettung / dieser Wechselwirkung im native Protein
- \+ Solvatationsterm: Häufigkeit e. Residuums mit Anteil an der SAS

Threading (Lathrop, Smith) (I)

Strukturmodell :=

- in 3D jedes Residuum wird durch Platzhalter ersetzt
- Platzhalter behält ausgewählte physikochemische Eigenschaften des betr. Residuums
- Eigenschaften:
  - Zugehörigkeit zu 2D-Element (α, β, Loop)
  - räumliche Umgebung, Abstände zu anderen Residuen
  - Anteil an SAS oder Core-Region (innen/aussen)

Sequenz → Alignment auf Strukturmodell

- kombinatorisches Optimierungsproblem, NP hart

Threading (Lathrop, Smith) (II)

- Eingabedaten für Threading
  1. Proteinsequenz A mit n Residuen a_i
  2. Strukturmodell mit m Segmenten (cores) C_j mit
     a. Länge L_j jedes Segmentes
     b. Segmente C_j u. C_{j+1} sind durch Loop-Region λ_k mit maximaler bzw. minimaler Länge l_k^{max} bzw. l_k^{min} verbunden
     c. lokale strukturelle Umgebung jedes Residuums (chemische Eigenschaften, räumliche Einschränkungen usw.)
  3. Bewertungsfunktion (score function) zur Bewertung eines Threading

- Ausgabedaten: optimales Threading $\mathbf{T} = \{t_j\}$; $j=1,...,m$
  - Wert t_j zeigt auf Residuum aus $\{a_i\}$, das die erste Position vom Segment C_j besetzt

Threading (Lathrop, Smith) (III)

- Bewertungsfunktion → Schlüsselfunktion
  - Richtigkeit + Robustheit der Ergebnisse
  - wesentliche Wechselwirkungen, die zur Strukturausbildung beitragen
  - Heuristiken ... physikalische Formulierungen

- Ansatz: 1-Zentren u. 2-Zentren-WW zwischen Segmenten
  - → NP hart: Lösung mit Branch-and-Bound

Threading (Lathrop, Smith) (IV)

Branch-and-Bound basiertes Threading

Einschränkung des Lösungsraums:

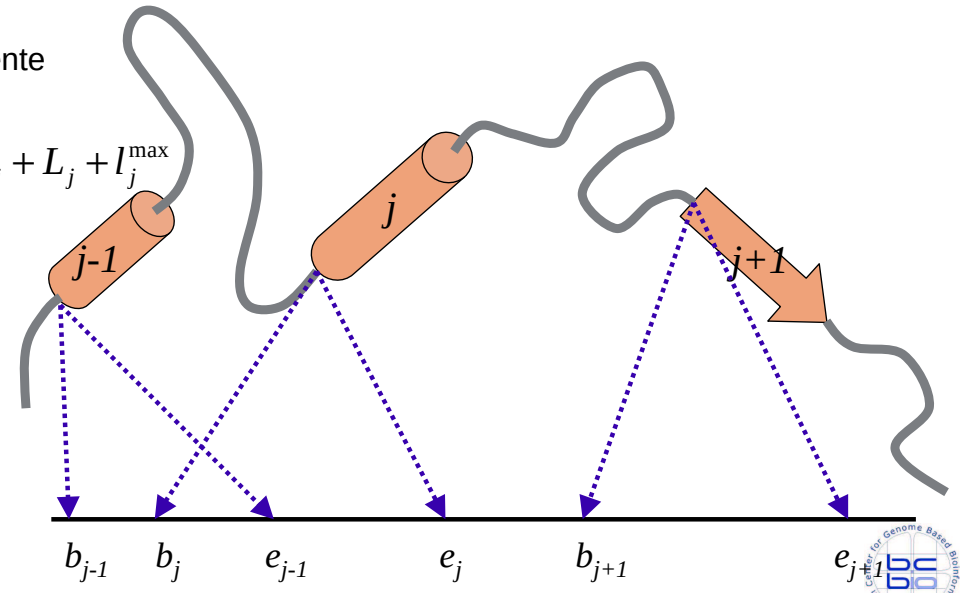
- räumliche Einschränkung

$$1+\sum_{j<i}(L_j+l_j^{\min}) \le t_i \le n+1-\sum_{j\ge i}(L_j+l_j^{\min})$$

- Reihenfolge der Segmente (kein Überspringen)

$$t_j + L_j + l_j^{\min} \le t_{j+1} \le t_j + L_j + l_j^{\max}$$

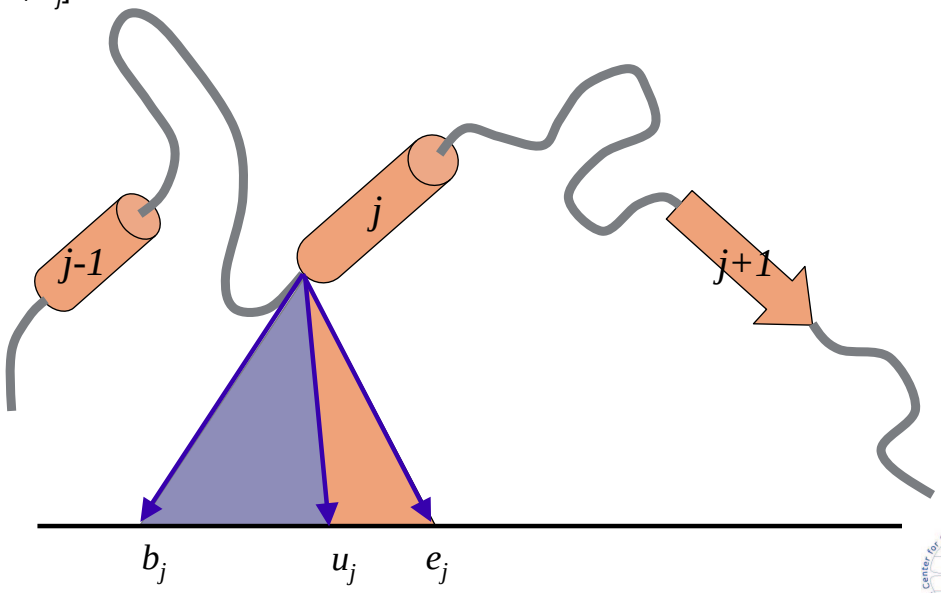
$$(b_j \le t_j \le e_j)$$



steinke@zib.de

Threading (Lathrop, Smith) (V)

- "Branch"-Algorithmus
  - wähle Segment *j* und Position in [b_j, e_j] aus
  - teile den Lösungsraum in 3 Teile:
    1. t_j in [b_j, u_j -1]
    2. t_j in [u_j +1, e_j]
    3. $t_j = u_j$



steinke@zib.de

Threading (Lathrop, Smith) (VI)

- Untere Grenze ("Bound"-Algorithmus)
- Ansatz für Bewertungsfunktion *f* für Threading **T**:

$$f(\mathbf{T}) = \sum_j g_1(j, t_j) + \sum_j \sum_{k>j} g_2(j, k, t_j, t_k)$$

  - WW über Segmente, nicht individuelle Residuen
- einf. Berechnung der unteren Grenze:

$$\min_{\mathbf{T}\in\Theta} f(\mathbf{T}) = \min_{\mathbf{T}\in\Theta} \sum_j \left[g_1(j, t_j) + \sum_{k>j} g_2(j, k, t_j, t_k) \right]$$

$$\geq \sum_j \left[\min_{b_j \leq x \leq e_j} g_1(j, x) + \sum_{k>j} \min_{\substack{b_j \leq y \leq e_j \\ b_k \leq z \leq e_k}} g_2(j, k, y, z) \right]$$

  - untere Schranken jedes Terms einzeln

B&B Protein Threading (Lathrop, Smith) (VII)

read {a_i}, Strukturmodell, C_j, f(**T**), f_{lb}

Θ ← alle möglichen Threadings
lb ← f_{lb} (Θ)
! Threadings in Prio-Queue Q
! Q_i Satz von Threadings u. lb
Insert (Q, Θ, lb)
do while .TRUE.
    ! holen der besten Kandidatenmenge aus Q
    Θ_c ← RemoveMin (Q)
    if |Θ_c|=1 **then**
        ! übriggebliebenes Threading ist die Lösung
        write T = Θ_c ; **return**
    else
        Split(Θ_c)
        do für jede neue Untermenge Θ_i aus Θ_c
            lb_i ← *flb*(Θ_i)
            Insert (Q, Θ_i lb_i)
        end do
    end if
end do

Genauigkeit

- abhängig von Kategorie des als ähnliche gefundenen Strukturniveaus
- Familie:
  - klare Sequenzähnlichkeit
  - Genauigkeit 1.0-3.0 Å
- Superfamilie:
  - gemeinsamer Ursprung v. Target- u. Template-Sequenz
  - Modelle teilweise korrekt (gut in Active Site)
  - RMSD 3.0-6.0 Å
- Analoge:
  - kein gemeinsamer Ursprung
  - Model niedriger Qualität
  - höchstens toologisch korrekt: richtige 2D-Alignments

steinke@zib.de

Probleme beim *Threading*

- Richtigkeit:
  - ~ 70% in Top10 enthalten korrekte Faltung
  - ~ 30% keine der Top-Hits korrekt
- Erhöhung der Spezifität: ← mehr Informationen
  - Funktion/Struktur-Relationen
  - Strukturmotive
  - ob u. welche Position von Kontakt-Residuen
- Qualität der Vorhersage
  - menschliche Expertise
  - Informationen von anderen Methoden

steinke@zib.de

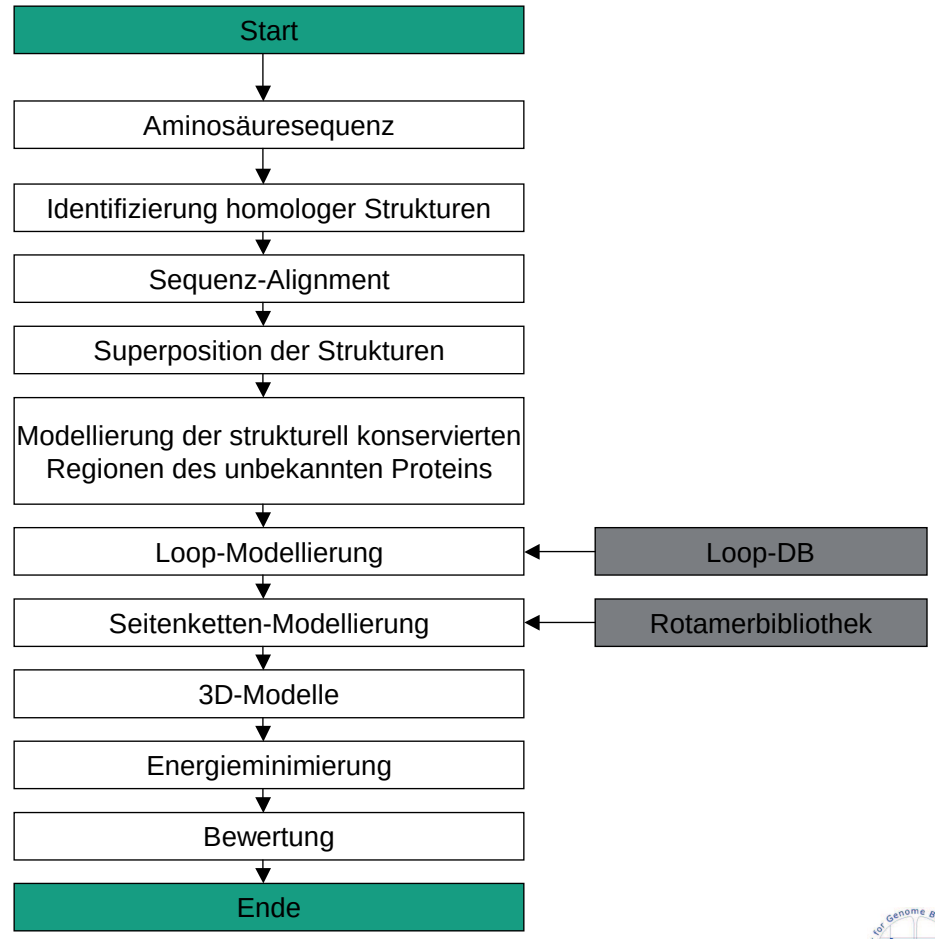
Homology Modeling

steinke@zib.de

Homology-Modeling

- Struktur ist konserviert → für homologe Sequenzen vermutlich Struktur ähnlich
- paarweise Sequenz-Ähnlichkeit: > 40%
- Koordinaten des Protein-*Backbones* homologer Strukturen als *Template* für neues Strukturmodell
- Güte der Modellierung:
  - >= 70%: sehr gute Modelle, Substitution v. Seitenketten
  - 40%-65%: moderat gute Modelle, signifikante Fehler im Backbone, speziell Loops
- Automatische Methoden:
  - GeneMine (UCSD)
  - Modellor (Rockefeller)

steinke@zib.de

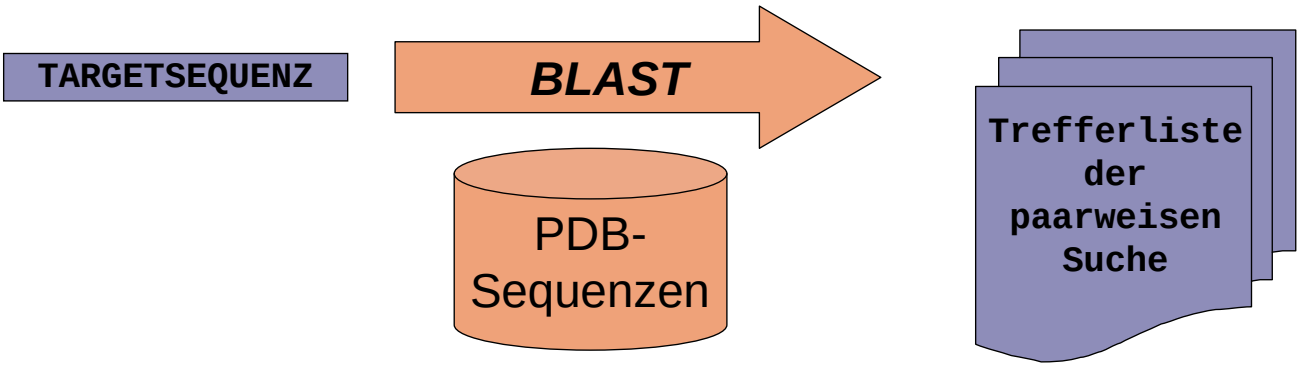


Anmerkungen

- Sequenz-Alignment
  - InDel in Loop-Regionen
  - Hauptkette → Basis f 3D-Konstruktion, < 1 Å RMSD
- Loop-Modellierung
  - Ramachandran-Statistik
- Seitenketten-Modellierung
  - Diederwinkel so identisch wie möglich
- Inspektion des Modells:
  - Atomkollisionen, visuell, Software
- Modellverbesserung durch eingeschränkte Minimierung
  - Seitenketten („kosmetische Korrektur“)

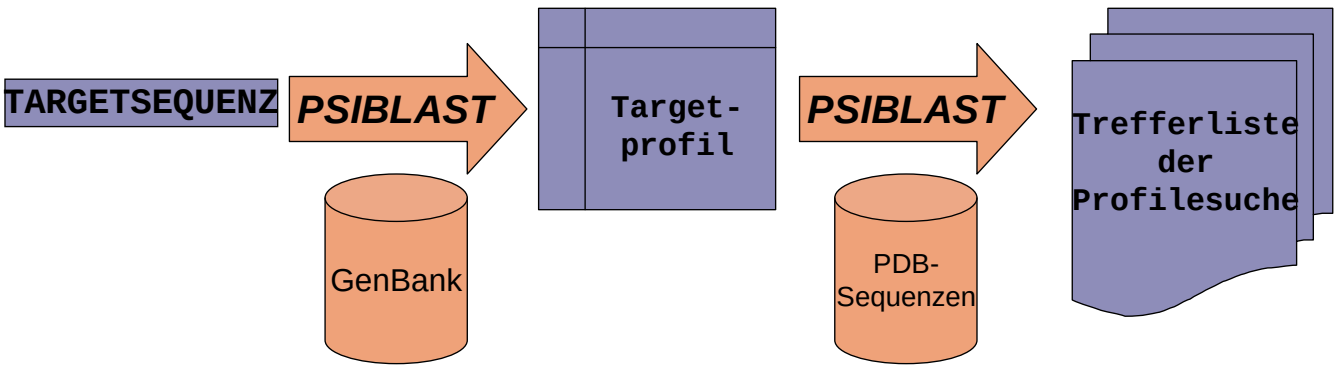
HM-Schritte (I)

- Identifizierung von Sequenzhomologen aus 3D-DB ...
  a) durch **paarweise** Sequenzsuche in PDB



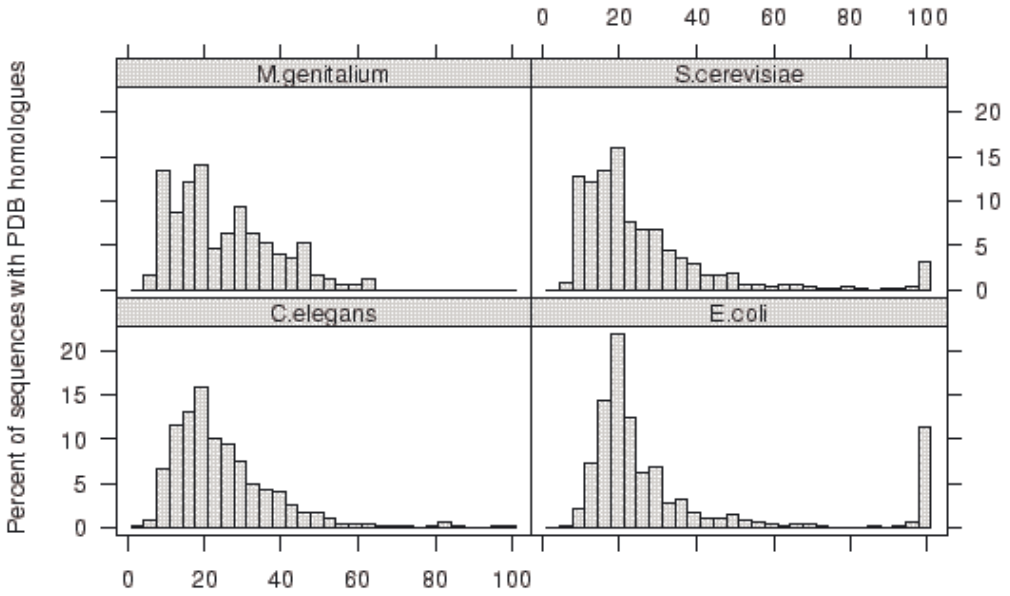
HM-Schritte (II)

- Identifizierung von Sequenzhomologen aus 3D-DB ...
  b) durch **profil-basierte** Sequenzsuche in PDB



Homologiesuche: Beobachtung

- Die meisten Proteine sind nur entfernt ähnlich zu Proteinen bekannter Struktur
  - Verbesserung mit wachsender Zahl von PDB-Einträgen



R.L. Dunbrack, Jr.

steinke@zib.de

HM-Schritte (III)

- Schritt 3: Loop Modellierung

Verschiebung Entfernung Einschub

R.L. Dunbrack, Jr.

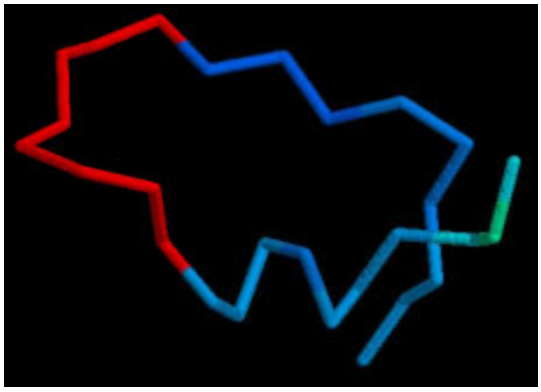
steinke@zib.de

Loop-Konstruktion (Auswahl)

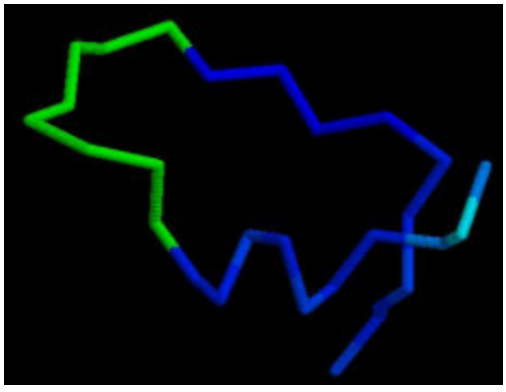
- Von Scratch: verbinden der Endpunkte von 2Ds
- Nutzung von Ramachandran-Plots: Suche nach möglichen Konformationen, + MD-Minimierung
- Random search (Monte Carlo / MD)

Loops: Gute Vorhersage

Vorhersage



Röntgenstruktur

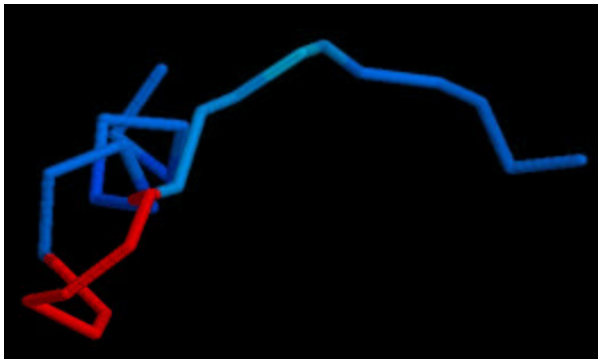


1JG1: Residuen 165-172. RMS=0.25 Å

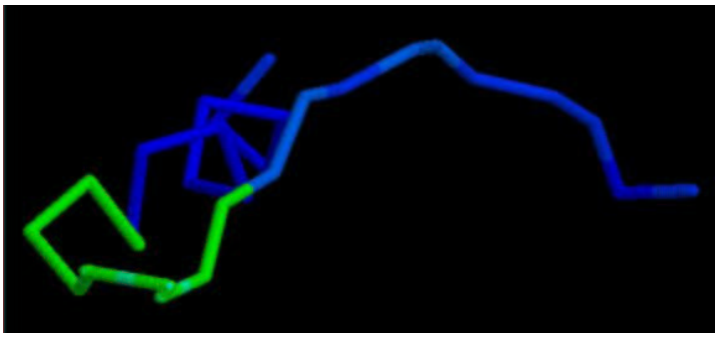
R.L. Dunbrack, Jr.

Loops: schlechte Vorhersage

Vorhersage



Röntgenstruktur



1I40: Residuen 142-149. RMS=2.29 Å

R.L. Dunbrack, Jr.

steinke@zib.de

SWISS-Model

- Comparative Modeling via Internet
  http://www.expasy.ch/swissmod/SWISS-MODEL.html

Vergleich der Strukturen mit dem Experiment

Sequenz Identität(%)	Anzahl Modell	max. RMS (Å)					
		1	2	3	4	5	>5
25-29	125	0	10	30	46	67	33
40-49	156	9	44	63	78	91	9
60-69	145	38	72	85	91	92	8
90-95	88	59	78	83	86	91	9

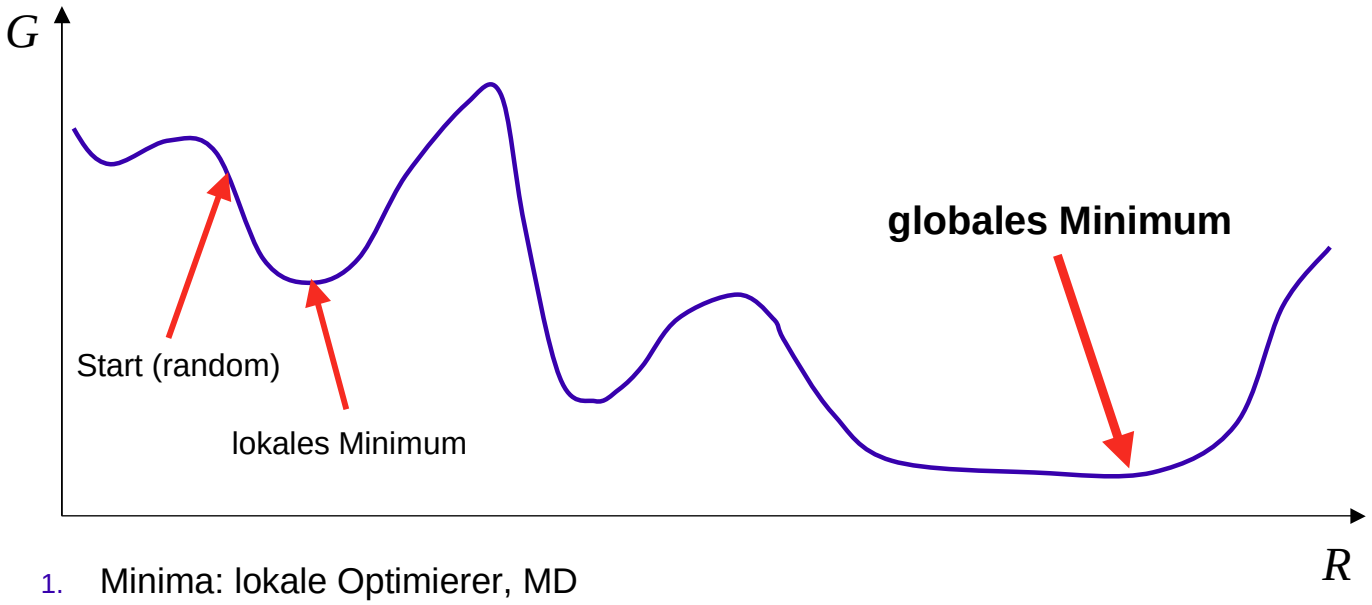
steinke@zib.de

Ab-initio Strukturvorhersage

- keine weiteren Informationen ausser Sequenz
- Modellierung der physikalischer Wechselwirkungen u. Prozesse der Faltung in die native Konformation
- thermodynamische Stabilität:
  - Proteinstruktur in nativer Konformation ist globales Minimum auf der Hyperfläche der Freien Enthalpie

    (Anfinsen‘s Hypothese); Faltung führt unabhängig in stabile Konformation
  - nicht bei allen Proteine gegeben (→ Chaperones)

Energiehyperfläche (*Free Energy Landscape*)



1. Minima: lokale Optimierer, MD
2. Barrieren / lokale Minima: MC, SA

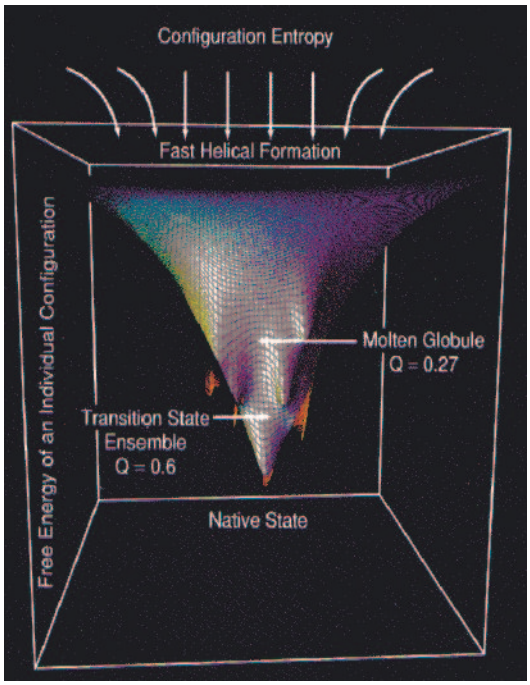
→ Multi-Minima-Problem!!!

Vorhersage: Methodischer Überblick

	Homologie-Modellierung	Faltungserkennung	Ab-initio
Methode	1. Suche nach Templates (Sequenzhomologe) 2. Modell durch Sequenzalignment 3. Modellierung v. Bereichen niedriger Ähnlichkeit	1. Klassifikation d Faltung 2. Threading 3. 3D-Profile 4. Verbesserung	1. Kraftfeld 2. Optimierung 3. globales Minimum
Nachteile	1. > 25% Sequenzähnlichkeit 2. Modellierung v. Loops u. Seitenketten kritisch	1. ausreichende Zahl v. Proteinen pro Faltungstyp erforderlich 2. Bewertungsfunktion kritisch	1. Rechenaufwand 2. physikalischen Modellierung
Auflösung	< 3 A	3 – 7 A	> 5 A
Zeitbedarf	< Tag	~ Tag	>> Tage

MD Simulation

- Simulation der Proteinfaltung: Modellierung der atomaren Wechselwirkungen
- globales Optimum: Trichter in Landschaft von ΔG := nativer Zustand
- Ziel: in-silico-Faltung



Computerentwicklung

- Strukturvorhersage := Grand Challenge Problem
  - PFlop/s Dauerleistung

  → Anreiz für Entwicklungen neuer Computerarchitekturen

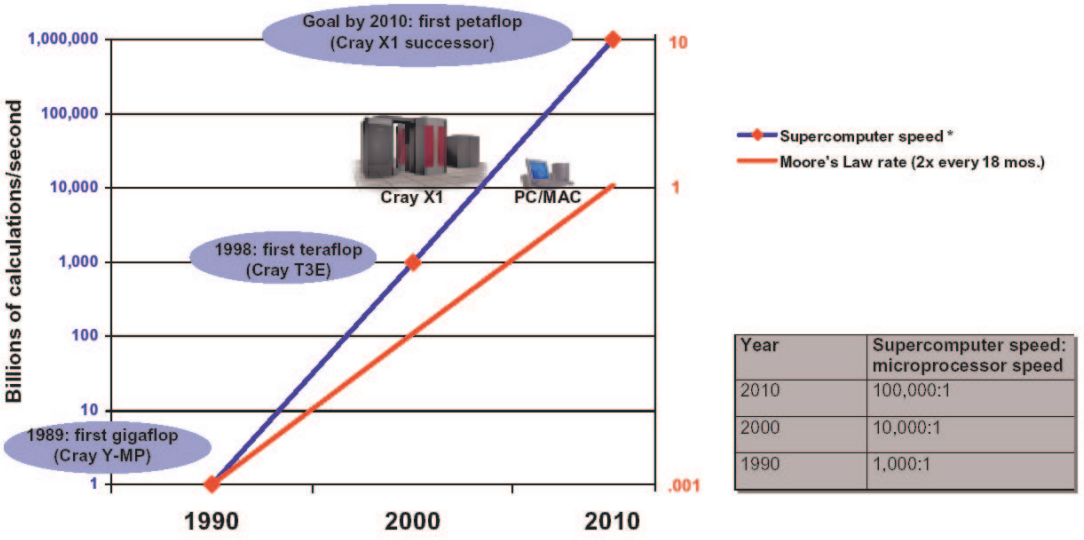
- aktuelle Projekte
  - CRAY Inc.: bis 2010 PFlop/s **Anwenderleistung**
  - IBM: BlueGene (Spezialrechner?)

Ab-initio Proteinfaltung: Warum 1 Petaflop/s?

Beschreibung	**Anzahl**	**Kommentar**
Atome	~ 32 000	300 Aminosäuren + Wasser
Kräfte/Zeitschritt	10^9	paarweise atomare WW
FLOPs / Kraftberechnung	150	
FLOPs / Zeitschritt	$1.5\ 10^{11}$	
Zeitschritt	~ 10^{-15} s	1 – 5 fs
Simulationszeit	10^{-3} s	Faltung in ms-Bereich
gesamte Zeitschritte	$2\ 10^{11}$	
FLOPs / SImulation	$3\ 10^{22}$	FLOP/s für Faltung
Ausführungszeit	$3\ 10^7$ s	1 Jahr
erforderliche FLOP/s	~ $1\ 10^{15}$	1 PFlop/s

www.research.ibm.com/bluegene/

1990-2010: Supercomputer speeds increases will outpace Moore's Law progress by 100x



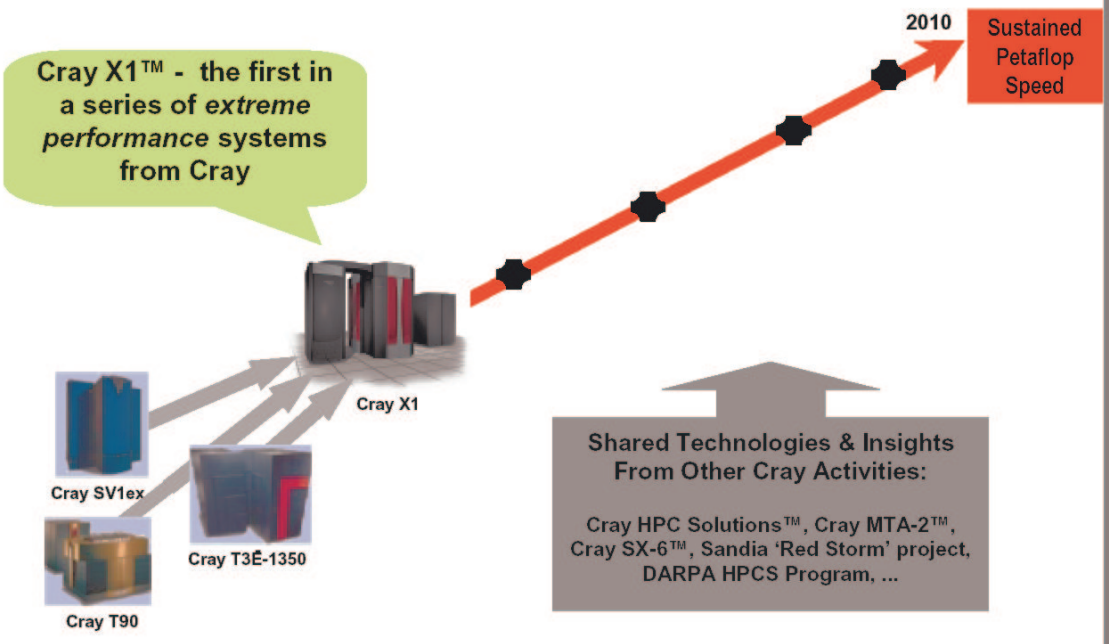
Year	Supercomputer speed: microprocessor speed
2010	100,000:1
2000	10,000:1
1990	1,000:1

*Supercomputer speeds shown are actual, sustained performance on full 64-bit applications. Microprocessor speeds represented are theoretical maximums (MIPS ratings) and are higher than actual speeds. Gigaflop: 1 billion (10^9) calculations/second. Teraflop: 1 trillion (10^{12}) calculations/second. Petaflop: 1,000 trillion (10^{15}) calculations/second.

SLIDE 3 | Cray X1 Supercomputer and Petaflop Speed
12/6/2002 | Cray Inc.

www.cray.com/products/systems/x1/

By 2010, Cray will deliver a system capable of sustained petaflop speed on a variety of challenging applications



SLIDE 6 | Cray X1 Supercomputer and Petaflop Speed
12/6/2002 | Cray Inc.

www.cray.com/products/systems/x1/

Blue Gene-Projekt

IBM Announces $100 Million Research Initiative to build World's Fastest Supercomputer

"Blue Gene" to Tackle Protein Folding Grand Challenge

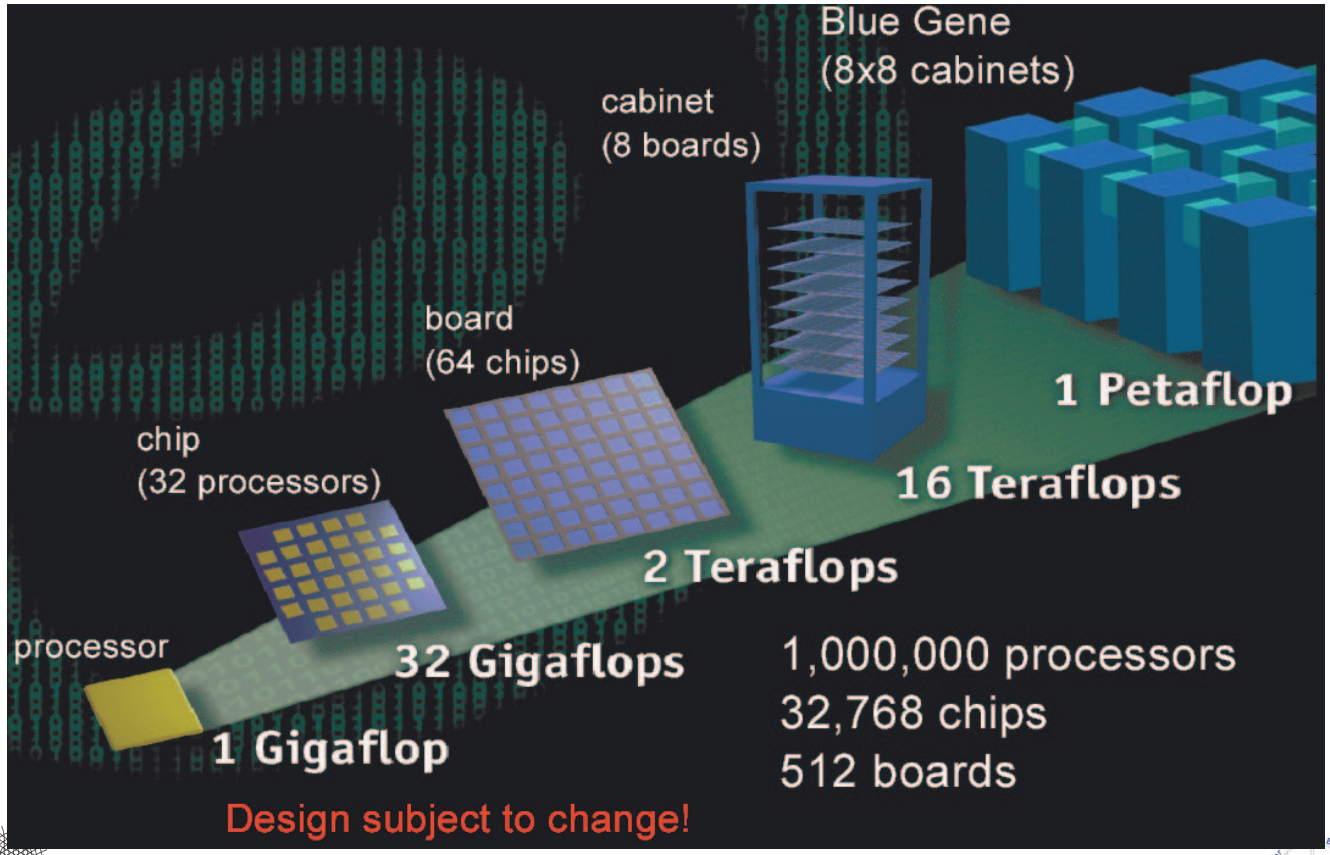
YORKTOWN HEIGHTS, NY, December 6, 1999 -- IBM today announced a new $100 million exploratory research initiative to build a supercomputer 500 times more powerful than the world's fastest computers today.

The new computer -- nicknamed "Blue Gene" by IBM researchers -- will be capable of more than one quadrillion operations per second (one petaflop). This level of performance will make Blue Gene 1,000 times more powerful than the Deep Blue machine that beat world chess champion Garry Kasparov in 1997, and about 2 million times more powerful than today's top desktop PCs.

Blue Gene's massive computing power will initially be used to model the folding of human proteins, making this fundamental study of biology the company's first computing "grand challenge" since the Deep Blue experiment. Learning more about how proteins fold is expected to give medical researchers better understanding of diseases, as well as potential cures.

http://www.research.ibm.com/bluegene/press_release.html

Blue Gene: Architektur



www.research.ibm.com/bluegene/

Epilog

Remember to keep all of the available data in mind when doing predictive work. Always ask yourself whether a prediction agrees with the results of experiments. If not, then it may be necessary to modify what you've done.

Robert B. Russell

- S-S-Bindungen: starke Randbedingung für Positionen von Cysteinen im Raum
- 2D-Strukturinfo (NMR)
- exp. Mutationsstudien
  → welche Residuen sind an aktiven und/oder bindenen Orten
- Kenntnisse über Abbaureaktionen, post-translationale Modifikationen (Phosphorylierung, Glycosylierung)
  → welche Reste sind zugängig

Referenzen

- B. Cheng: Protein Structure prediction - An Overview http://s-star.org/downloads/lecture7/notes/prediction.pdf
- P. Mittl: Protein structure prediction, Vorlesung, Uni Zürich
- Robert B. Russell: A Guide to Structure Prediction (v. 2) http://www.bmm.icnet.uk/people/rob/CCP11BBS/
- Higgins, Taylor (ed.): Bioinformatics, Oxford Uni Press, 2000
- Setubal, Meidanis: Introduction to Computational Molecular Biology, PWS Publ. Comp., 1997
- R.L. Dunbrack, Jr.: BMB612, Lecture 2 Protein Structure Prediction http://www.fccc.edu/research/labs/dunbrack/
- CRAY Inc., http://www.cray.com/products/systems/x1/
- W. Pulleyblank: *Protein Folding and the Blue Gene Petaflops Computer*, Supercomputing, Mannheim, 2000
- IBM: http://www.research.ibm.com/bluegene/

steinke@zib.de