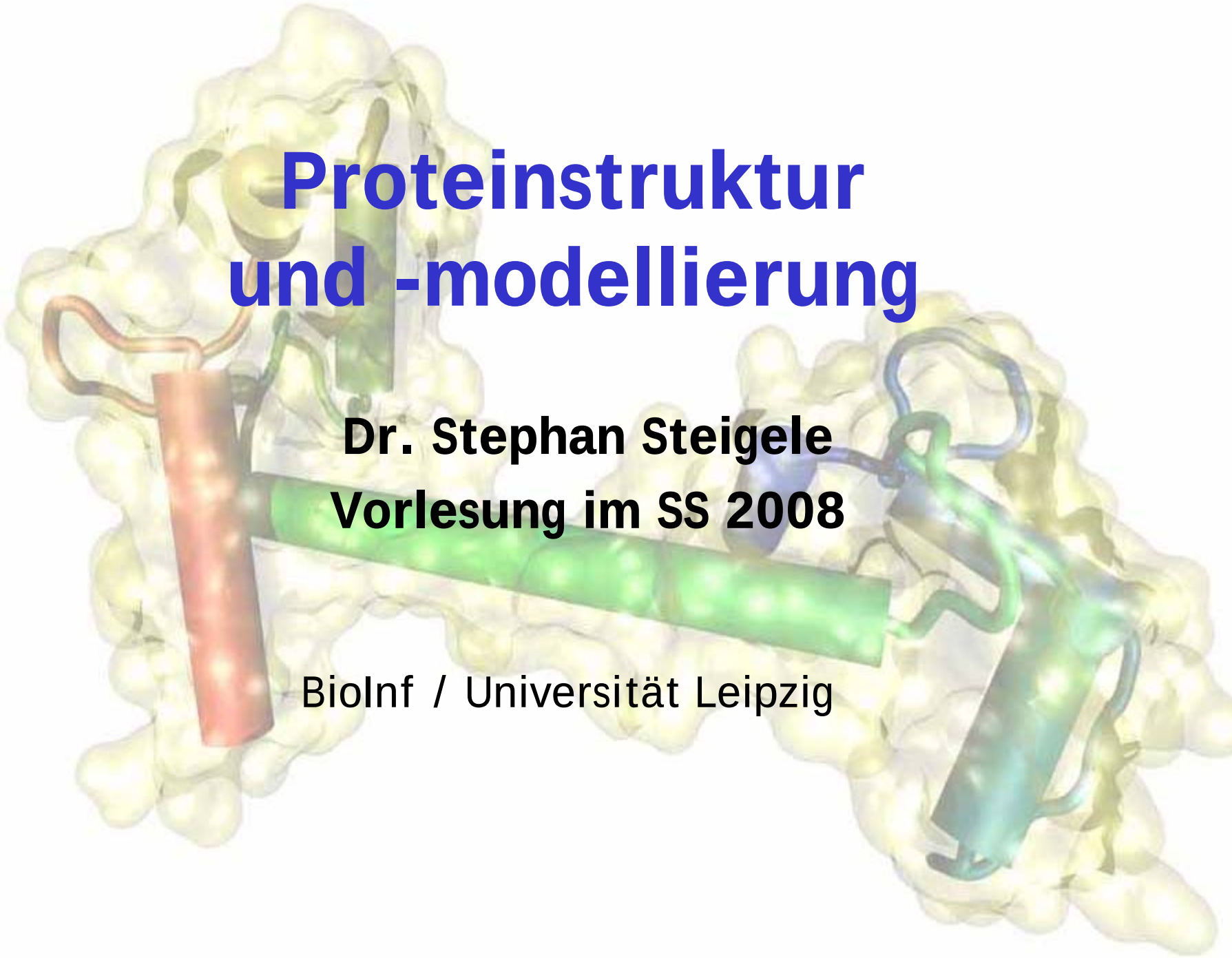


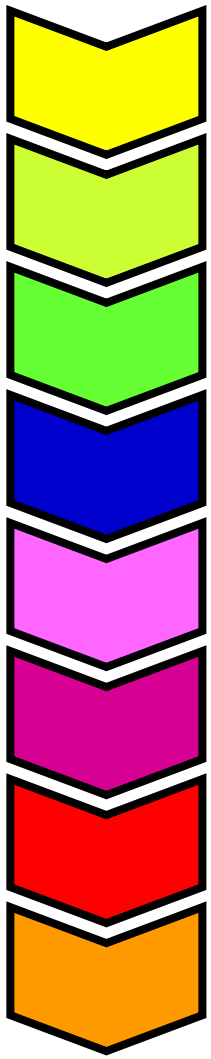
Proteinstruktur und -modellierung

Dr. Stephan Steigele

Vorlesung im SS 2008

BioInf / Universität Leipzig

Remark



Viele Folien dieser Vorlesung wurden

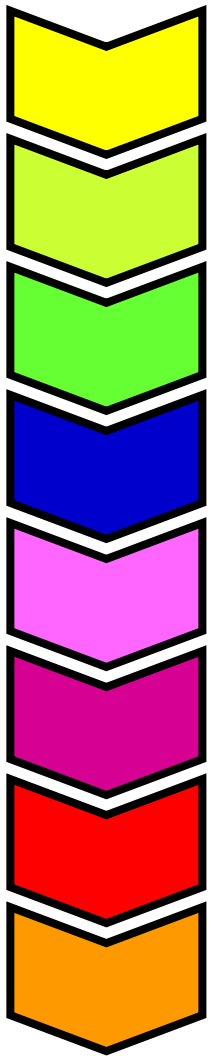
von **Prof. Oliver Kohlbacher** vom **Lehrstuhl zu Simulation Dynamischer Systeme**

in Tübingen entwickelt.

Die folgende Vorlesung wurde allerdings deutlich **adaptiert**, so dass

jegliche vorhandenen Fehler ausschliesslich die Schuldigkeit des vorlesenden treffen .. ;)

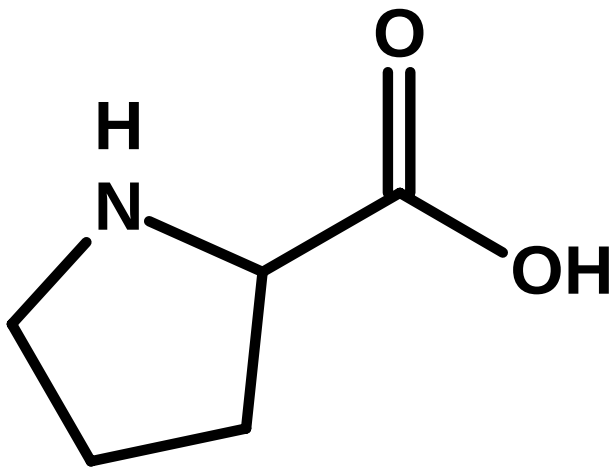
Überblick über die Vorlesung

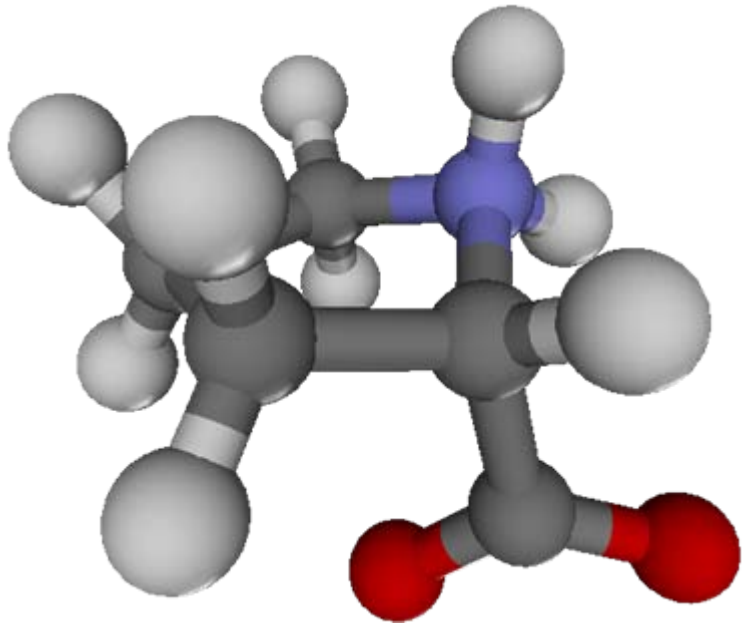


1. Einleitung und Überblick
2. Aufbau und Struktur der Proteine
3. Modellierung von Proteinstrukturen
4. Proteinfaltung
5. Proteinstruktur-Vorhersage

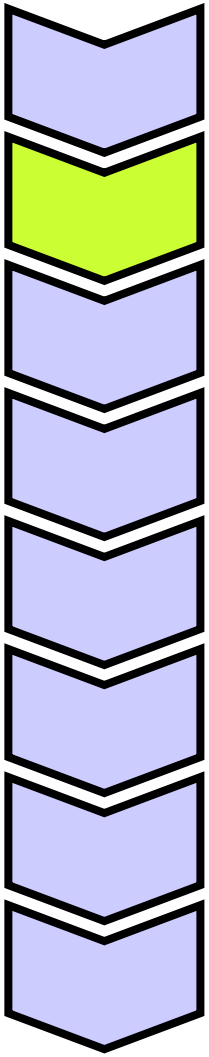
Aminosäuren

- Detaillierte Eigenschaften der Aminosäuren
- Relevant für
  - Sequenz-Struktur-Beziehungen
  - Wechselwirkungsmöglichkeiten
  - Verständnis der Strukturen





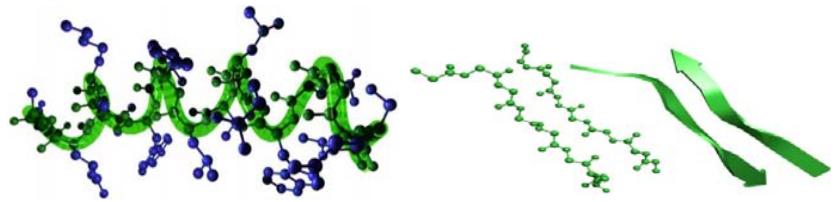
Proteinstruktur – Überblick



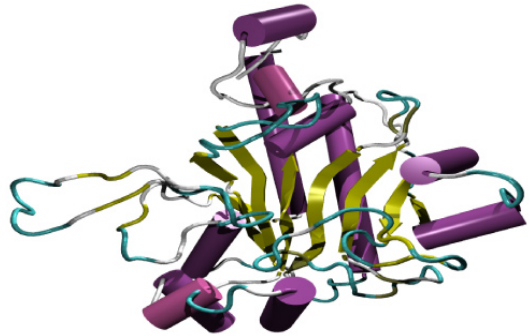
Primärstruktur

...LGFCYWS...

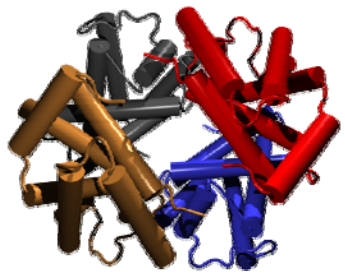
Sekundärstruktur



Tertiärstruktur

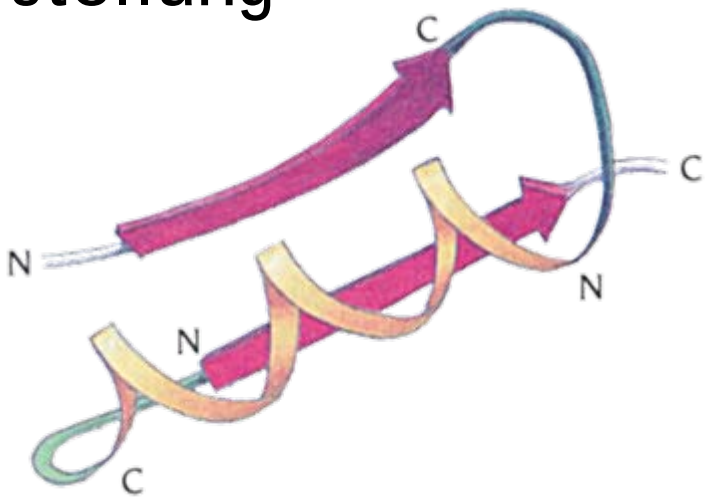


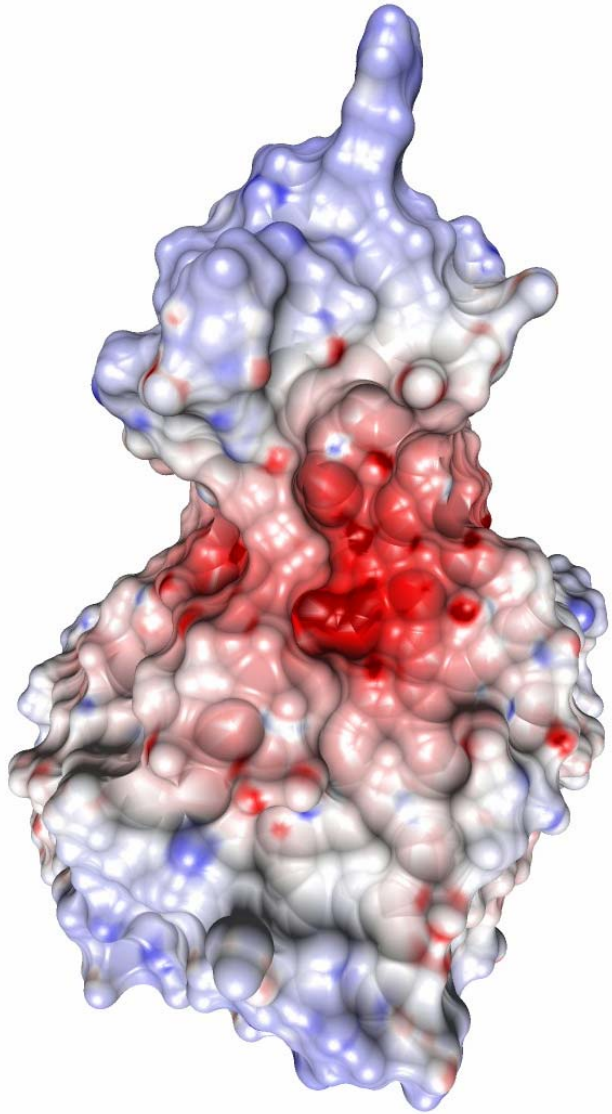
Quartärstruktur



Visualisierung

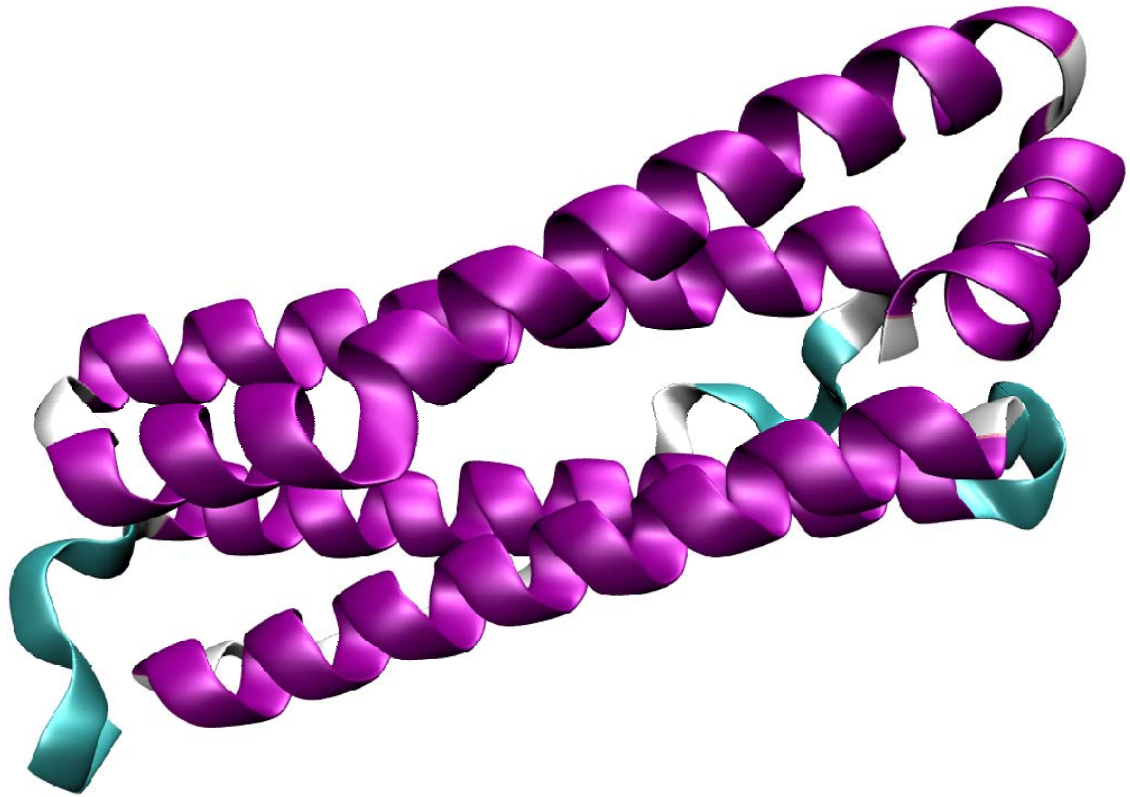
- Arten der Darstellung von Proteinstrukturen
- Interpretation der Abbildungen
- Software zur Darstellung



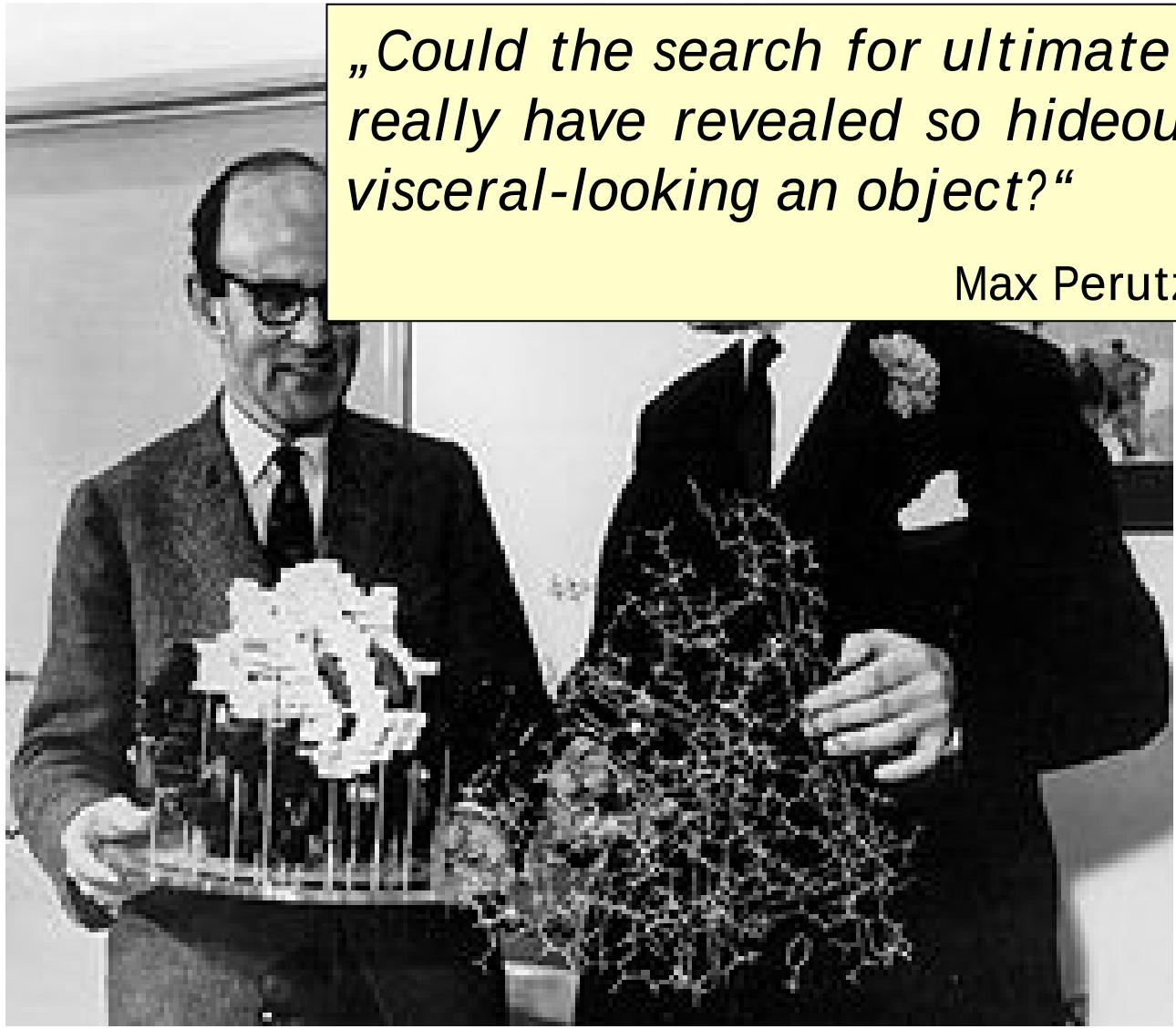


Strukturfamilien

- Proteine bilden eine Vielfalt unterschiedlicher Strukturen, die sich in Familien gliedern lassen
- Struktur ist viel besser konserviert als Sequenz



Wie sehen Proteine aus?



„Could the search for ultimate truth really have revealed so hideous and visceral-looking an object?“

Max Perutz, 1964

Kraftfelder

$$E = E_{stretch} + E_{bend} + E_{tors} + E_{vdW} + E_{ES}$$

$$= \sum_{bonds\ (ij)} \frac{k^{(ij)}}{2}\left(r_{ij} - r_0{}^{(ij)}\right)^2 + \sum_{angles\ (ijk)} \frac{k^{(ijk)}}{2}\left(\phi_{ij} - \phi_0{}^{(ijk)}\right)^2$$

$$+ \sum_{torsions\ (ijkl)} \frac{k^{(ijkl)}}{2}\left(1 + \cos(n^{(ijkl)}\tau - \tau_0{}^{(ijkl)}\right)^2$$

$$+ \sum_{pairs\ (ij)} \left(\frac{A(ij)}{rij} - \frac{B(ij)}{rij}\right) + \frac{1}{4\pi\varepsilon\varepsilon_0} \sum_{pairs\ (ij)} \frac{q_i q_j}{r_{ij}}$$

Kraftfelder

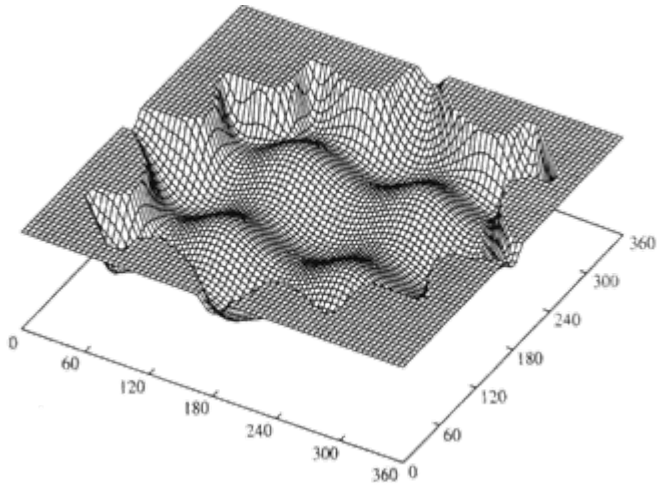
Kraftfeld	# Atom-Typen	vdW	ES	HB	stretch	bend	cross	Domäne
ECEPP/2	21	6-12	MP	10-12	-	-	-	Proteine
TRIPOS	31	6-12	MP	-	P2	P2	-	Allgemein
AMBER	54	6-12	MP	-	P2	P2	-	Proteine, DNA
CHARMM	29	6-12	MP	-	P2	P2	-	Proteine, DNA
MM2	71	Exp-6	DP	-	P3	P6	sb	Allgemein
MM3	155	Exp-6	DP	-	P4	P6	sb, bb, st	Allgemein
MMFF94	99	7-14	MP	7-14	P4	P3	sb	Allgemein

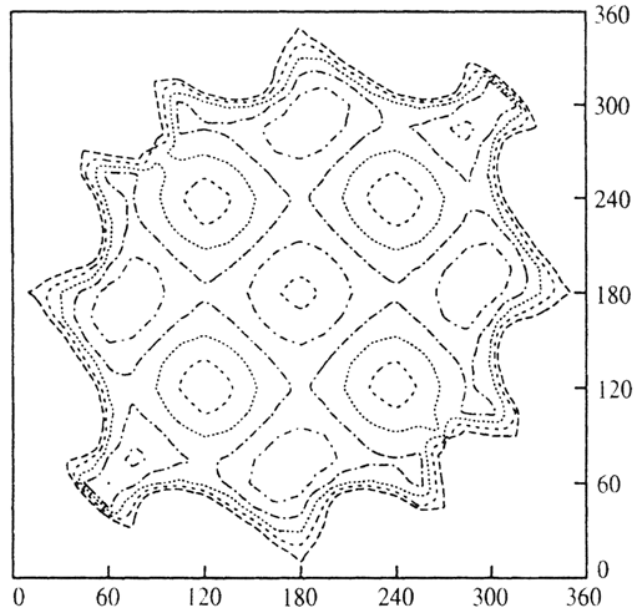
DP – Dipol-Dipol-WW, MP – Monopol (Coulomb),
sb – stretch/bend, bb – bend/bend, st – stretch/torsion

Energiehyperflächen von Proteinen

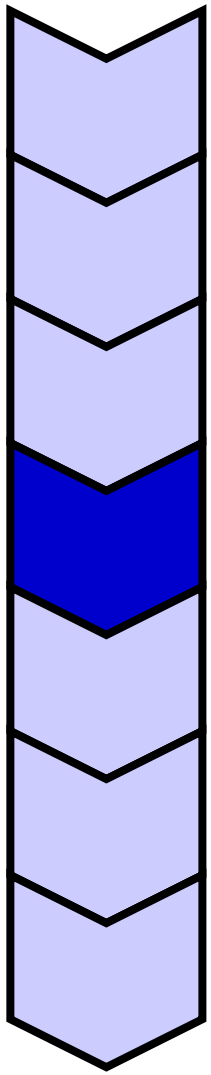
- **Minima**
  - Entsprechen günstigen Konformationen (Konformeren)
  - Häufig **lokale** Minima!
- **Globales** Minimum?

  (Bsp.: Proteinfaltung!)
- Kann man die Oberfläche systematisch durchmustern?
- Muss man das?



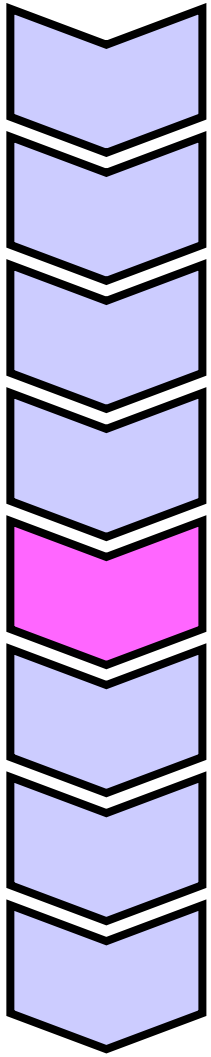


Modellierung von Proteinstrukturen



- Konformationsraum
- Wechselwirkungen
- Kraftfelder
  - Definition
  - Beispiele: AMBER, CHARMM, ...
- Algorithmen zur Energieminimierung
- Durchmustern des Konformationsraums
  - Systematische Suche
  - Stochastische Methoden

Proteinfaltung



- Wie falten sich Proteine?
- Warum falten sich Proteine?
- Welche Modelle beschreiben die Proteinfaltung?
- Kann man die Faltung vorhersagen?

Proteinstruktur – Proteinfaltung

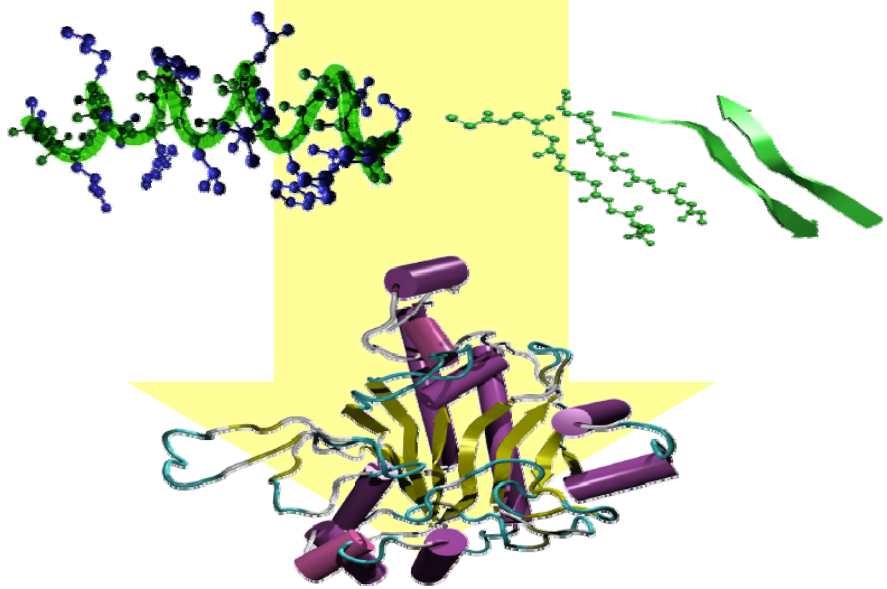
Primärstruktur

Sekundärstruktur

Tertiärstruktur

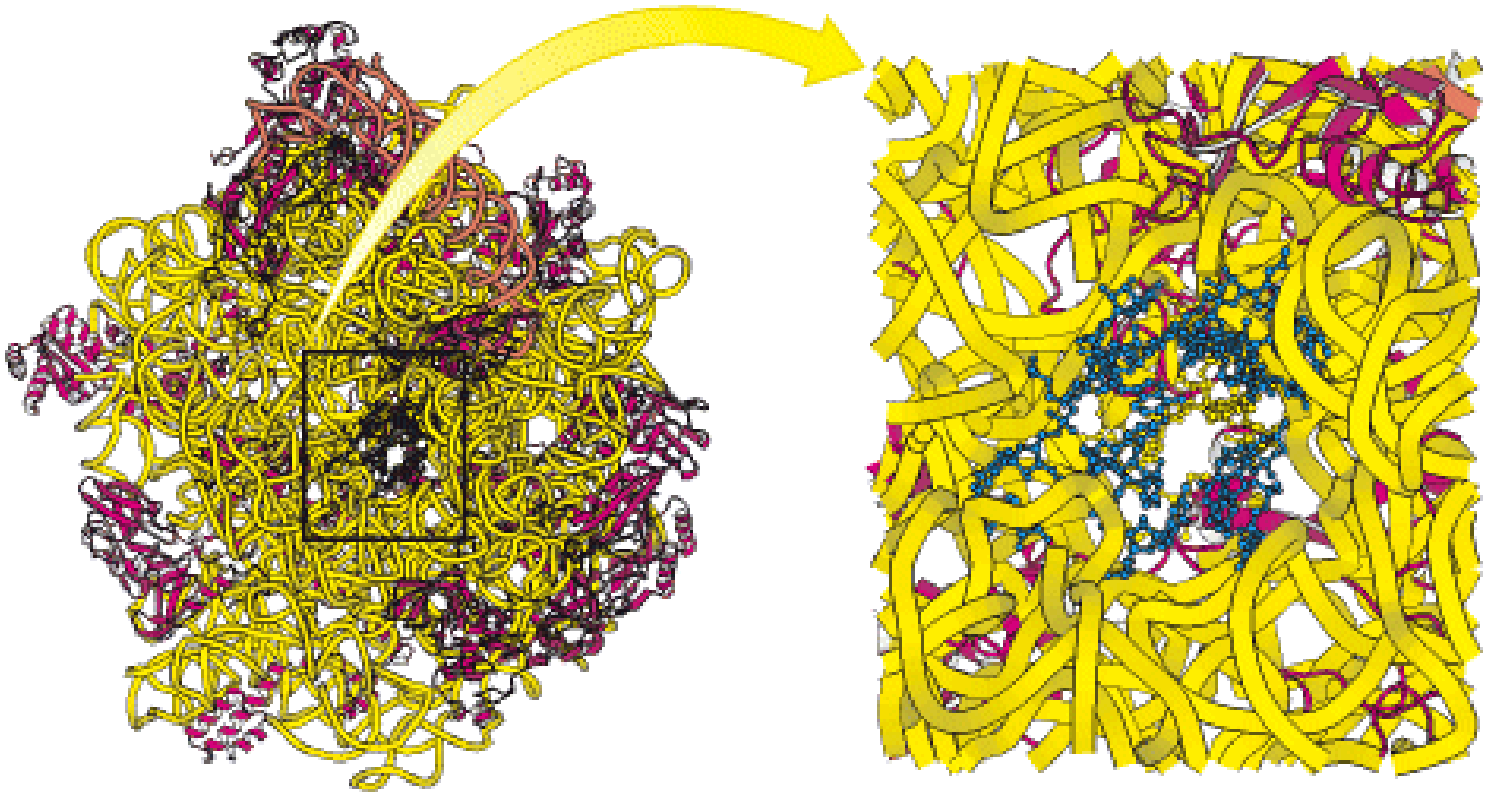
Qu

. . . LGFCYWS . . .



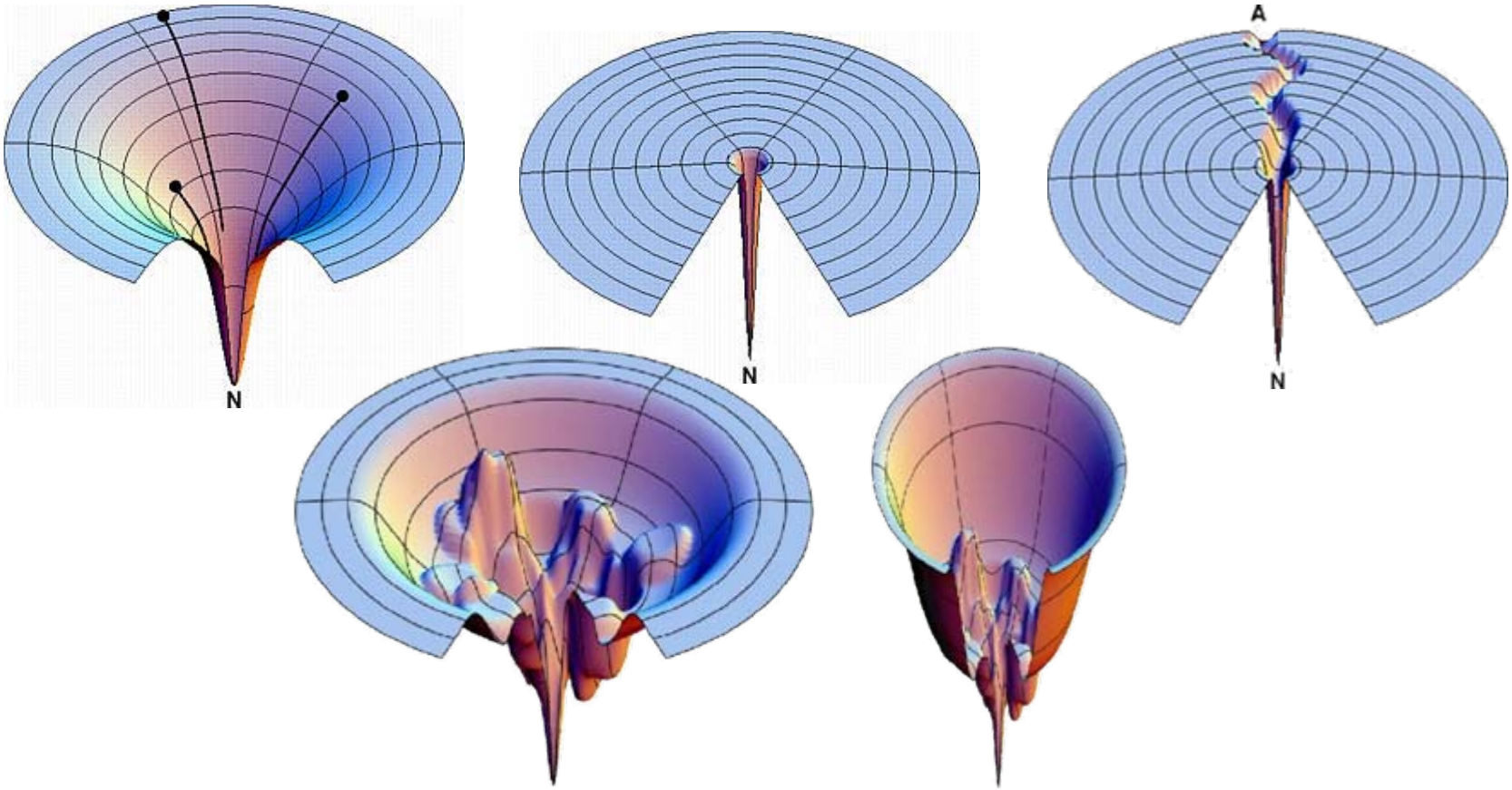
Woher weiß das Protein, wie es sich zu falten hat?

Protein-Biosynthese



Austrittspfad des neu synthetisierten Proteins aus dem Ribosom

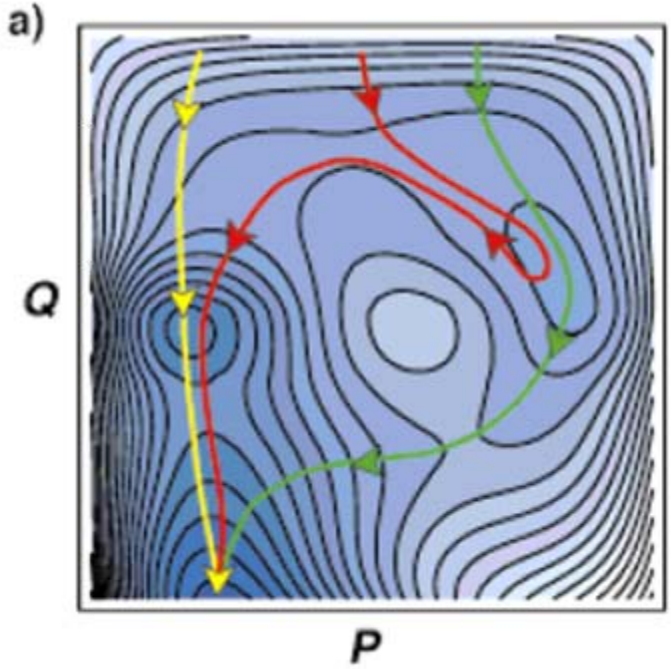
Faltungstrichter

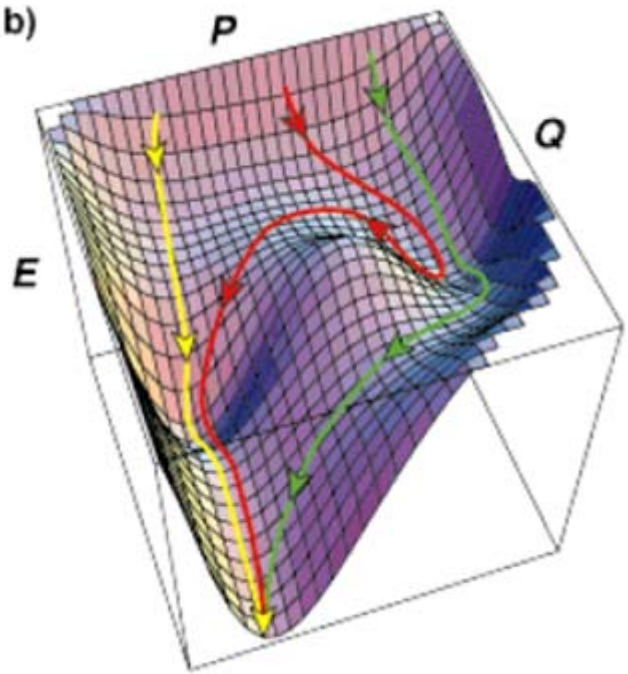


Bei der Faltung folgen die Proteine Pfaden durch eine sehr komplexe Energielandschaft

Dobson, Šali, Karplus, Angew. Chemie (1998), 110(7), 908

Faltungspfade

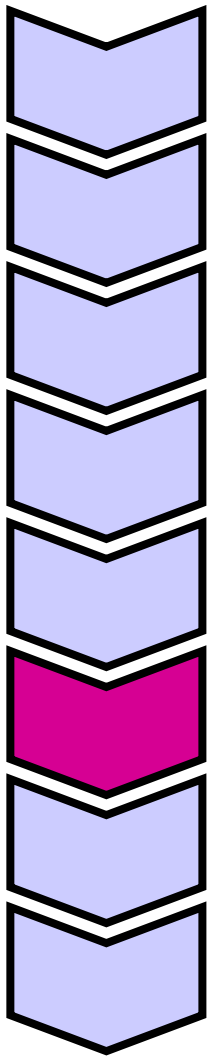




- Viele Proteine kennen mehrere Faltungspfade
- Zwischenzustände in denen je eine der beiden Domänen gefaltet ist entsprechen Minima der Energiehyperfläche
- Faltungsgeschwindigkeiten der Domänen unterschiedlich: eine überwiegend
- Pfade unterschiedlich schnell (gelb schnell, grün + rot langsam)

Dobson, Šali, Karplus, Angew. Chemie (1998), 110(7), 908

Proteinstruktur-Vorhersage



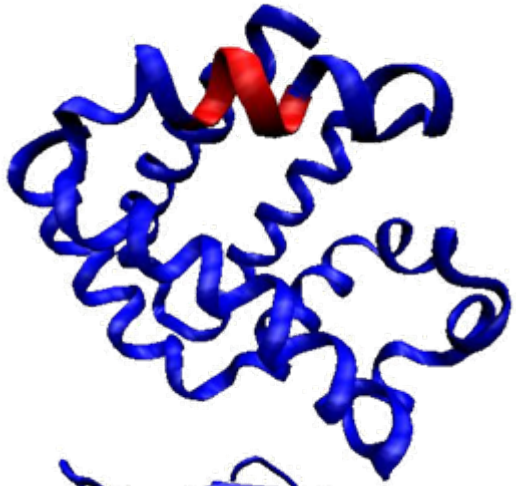
Proteinstruktur-Vorhersage

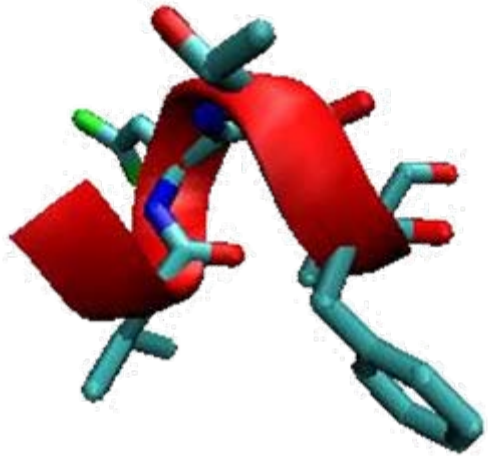
- Wie kann man Sekundärstrukturelemente vorhersagen?
- Kann man die gesamte 3D-Struktur vorhersagen,
  - ausgehend von einem ähnlichen Protein?
  - ohne Kenntnis ähnlicher Strukturen?
- Welche Algorithmen und Werkzeuge gibt es dazu?
- Wie gut sind die Vorhersagen?

Proteinstruktur ist nichtlokal

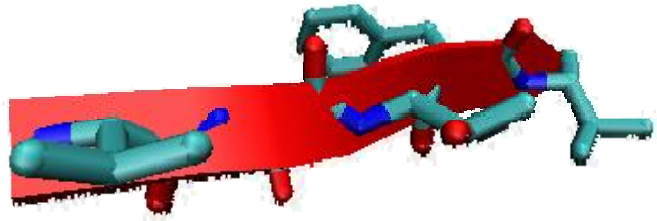
Identische Sequenzen bilden unterschiedliche Sekundärstrukturen aus, je nach **Umgebung**!

1ECN

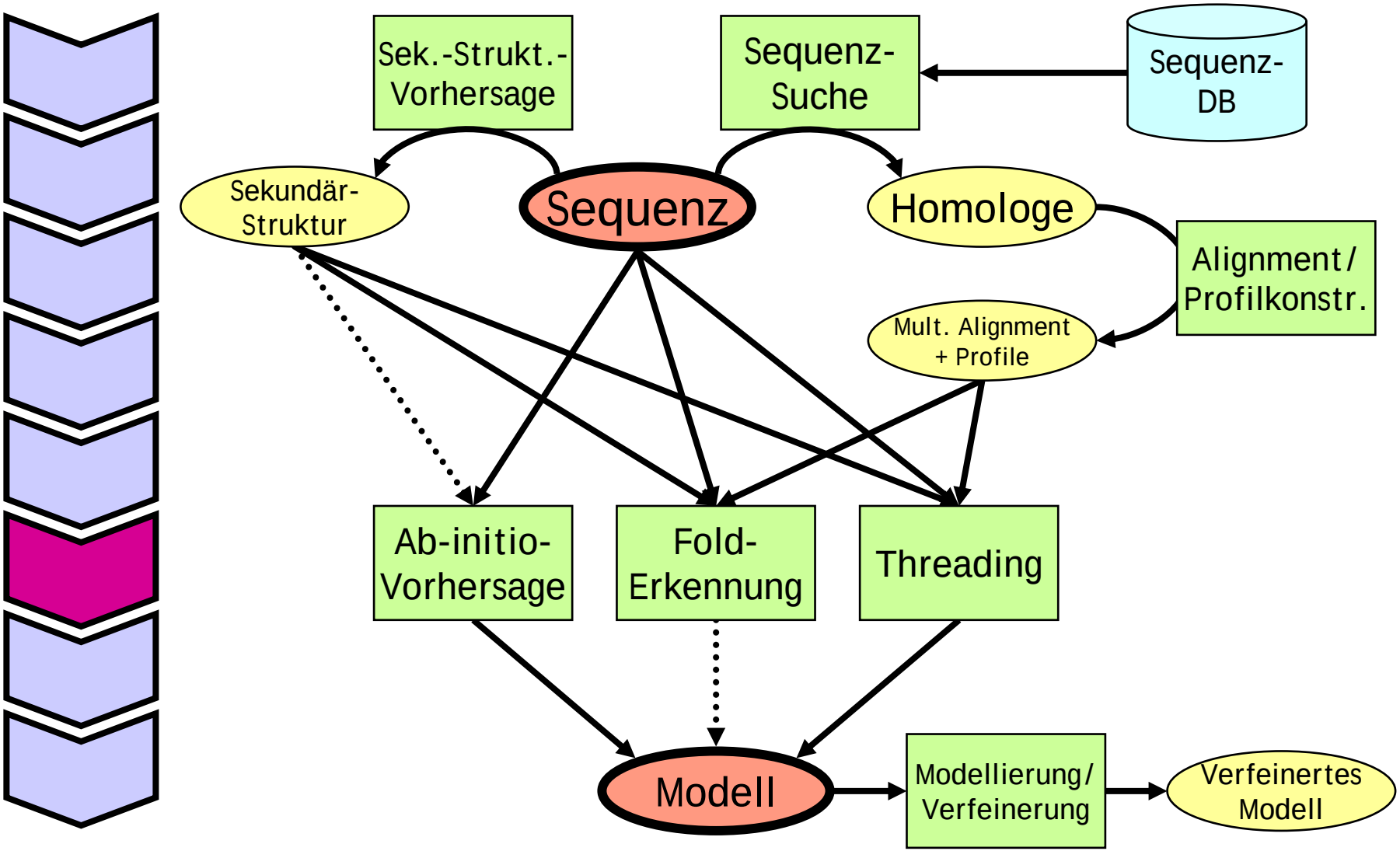




9RSA



Methoden der Strukturvorhersage

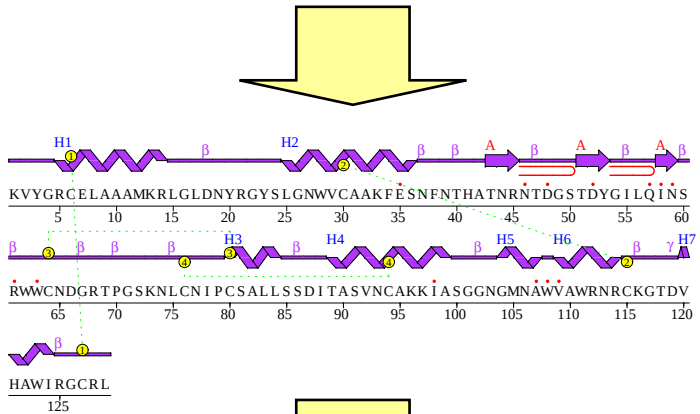


Nach: Zimmer, Lengauer: Bioinformatics – From Genomes to Drugs, Wiley VCH, 2001

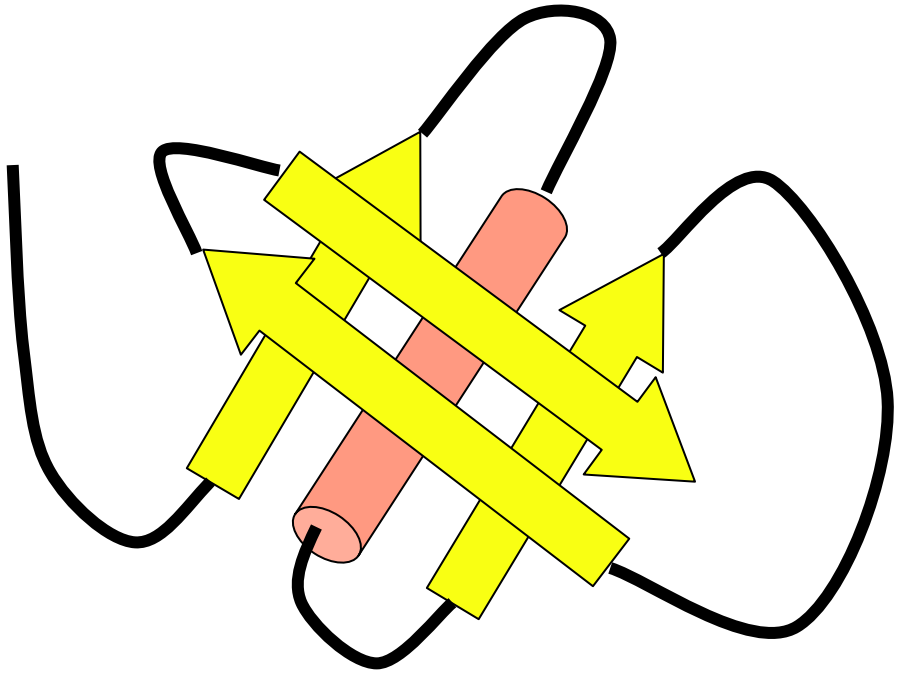
Sekundärstruktur-Vorhersage

- Sekundärstruktur-Zuordnung definiert Topologie des Proteins
- Packung der Sekundärstrukturen im Raum definiert Faltungsklasse
- Wichtiger Anhaltspunkt für Tertiärstruktur

```
KVYGRCELAAAMKRLGLDNYRGYSLGNWVC
AAKFESNFNTHATNRNTDGSTDYGILQINS
RWWCNDGRTPGSKNLCNIPCSALLSSDITA
SVNCAKKIASGGNGMNAWVAWRNRCKGTDV
HAWIRGCRL
```



Threading

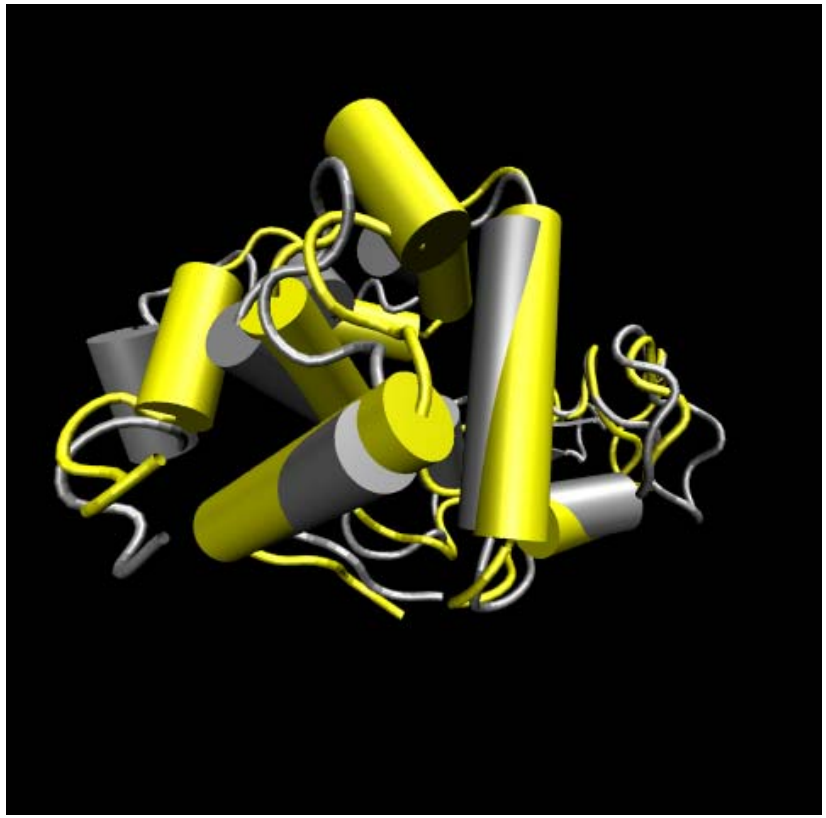


```
...LGFCYWS...
...ILVGCIL...
```

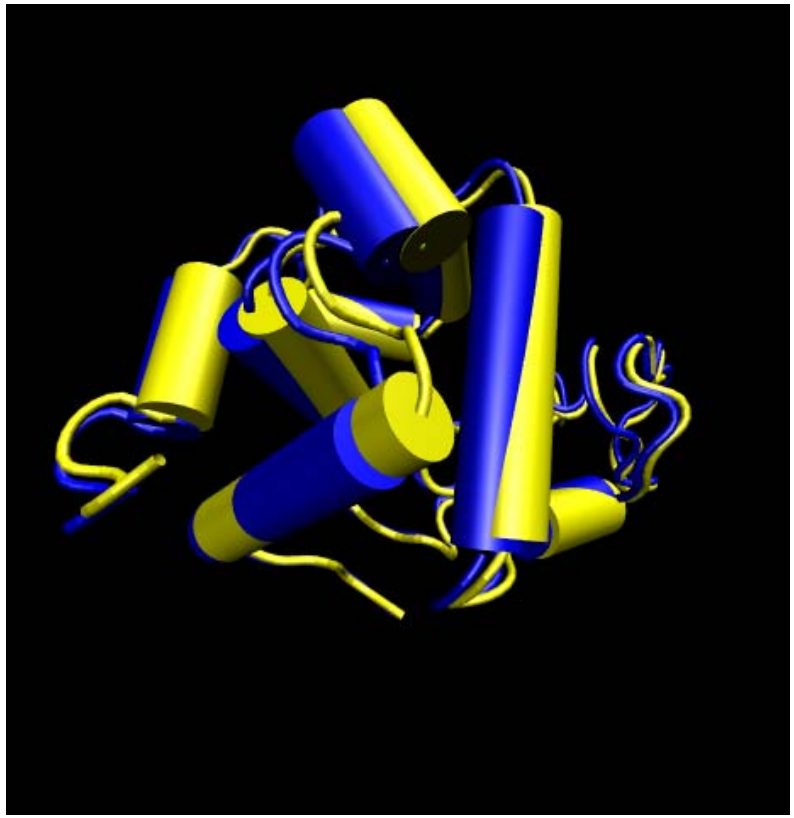
Gegeben

- eine (oder mehrere) Struktur(en) (Schablonen)
- Eine Zielsequenz

Beispiel

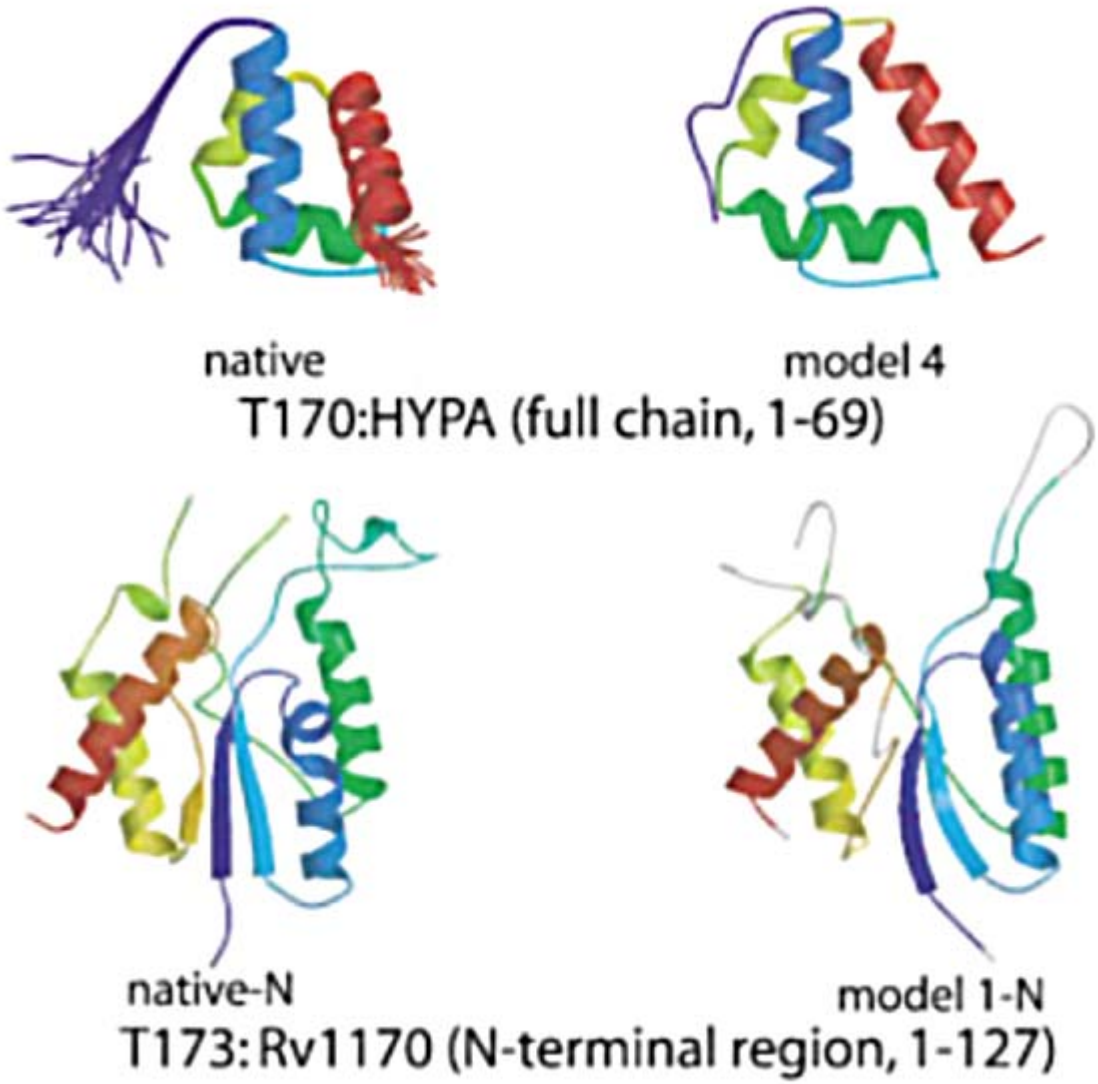


Grau: 1IVM
Gelb: 1IVM gethreaded auf 1LZY

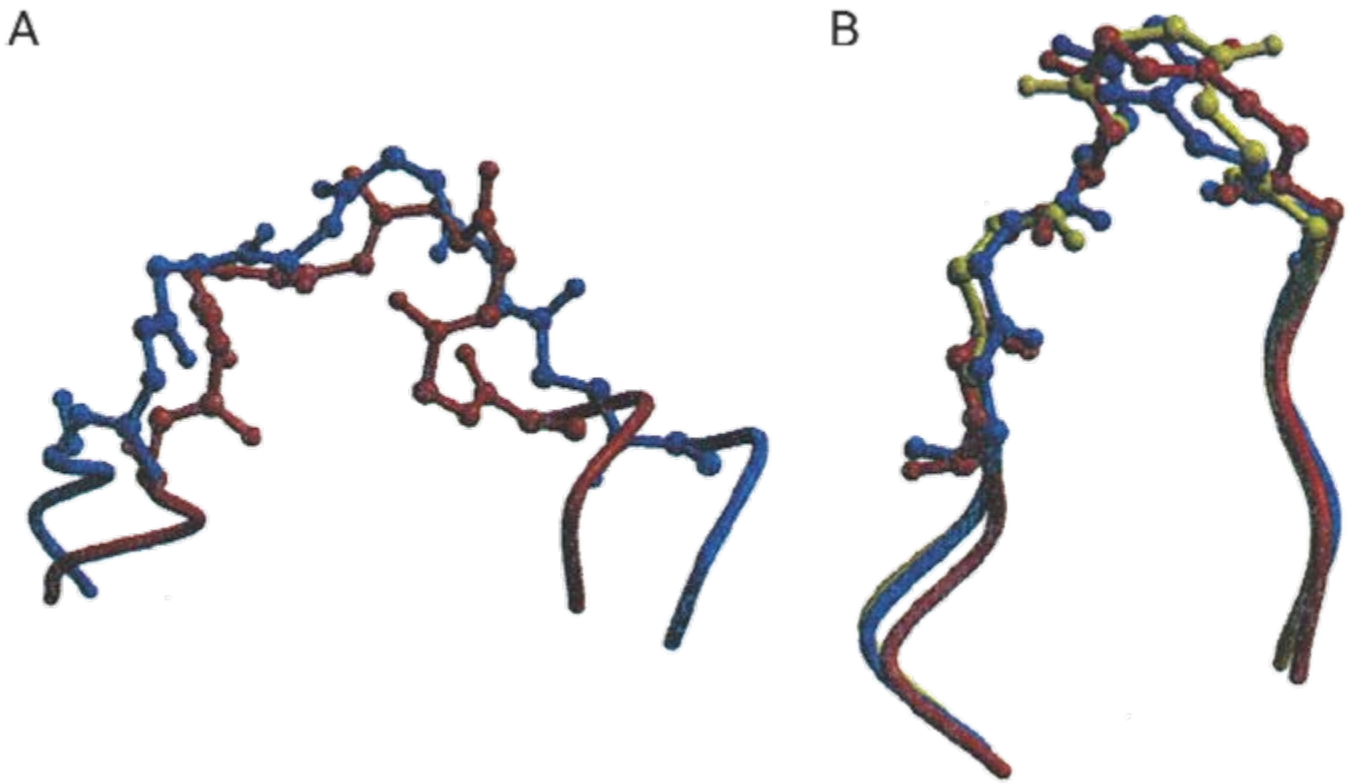


Blau: 1LZY
Gelb: 1IVM gethreaded auf 1LZY

ROSETTA – Ergebnisse CASP5

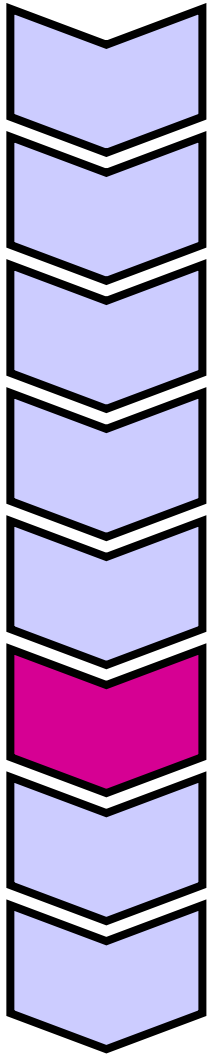


Modellierung von Schleifen



Nativ: blau, modellierte: rot,
ähnlichste Struktur: gelb

Überblick über die Vorlesung



Proteinstruktur-Vorhersage

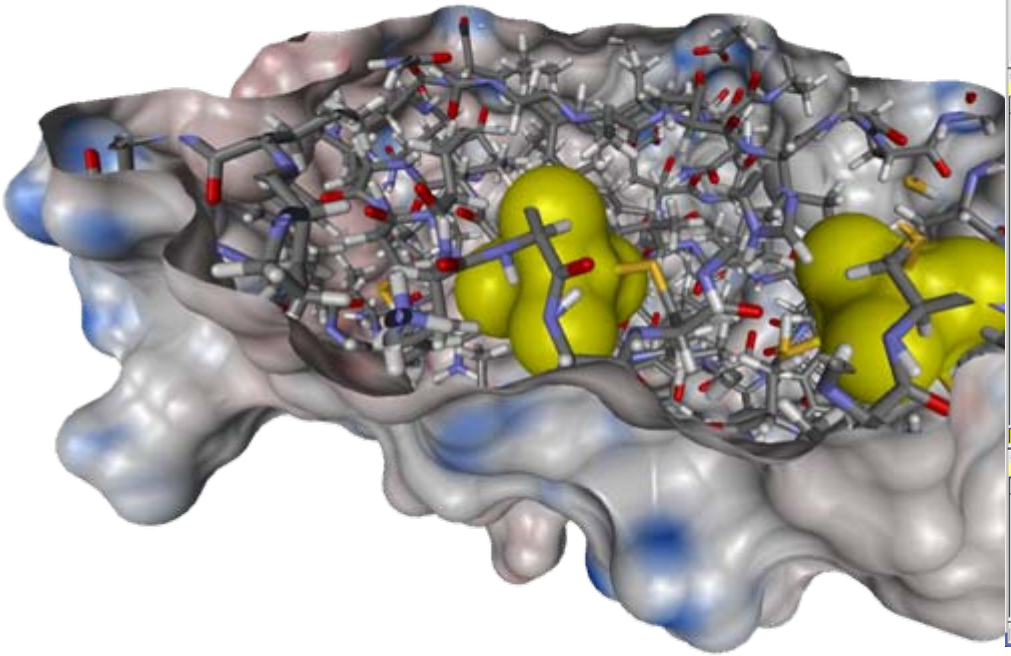
- Sekundärstrukturvorhersage
- Tertiärstrukturvorhersage
  - Fold Recognition, Threading
  - Ab-initio-Ansätze
- Homologiemodellierung
  - Schleifenmodellierung
  - Seitenkettenplatzierung
  - Optimierung, Validierung
  - Werkzeuge

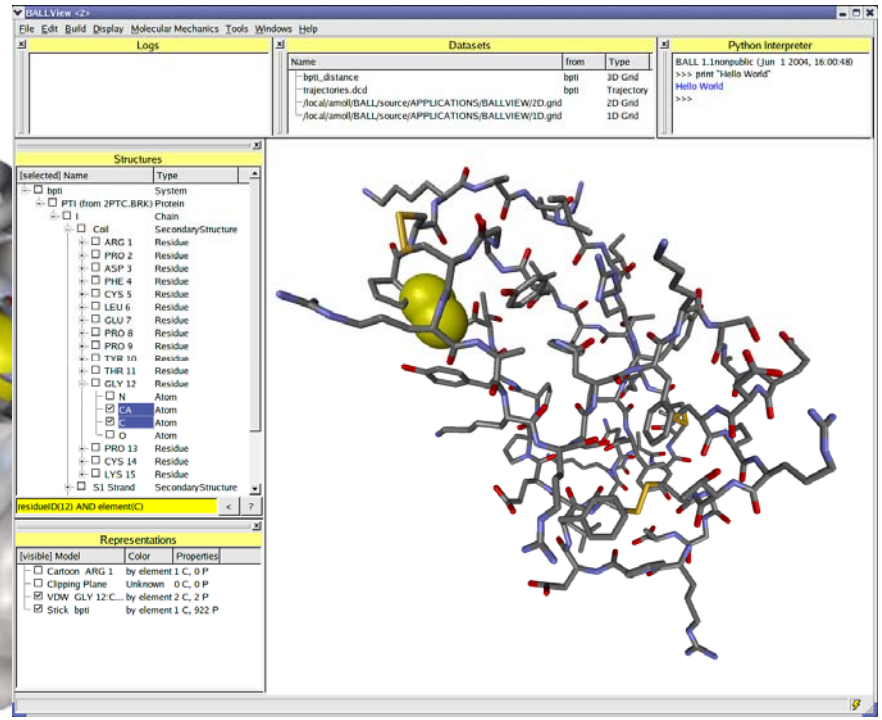
Empfohlene Software

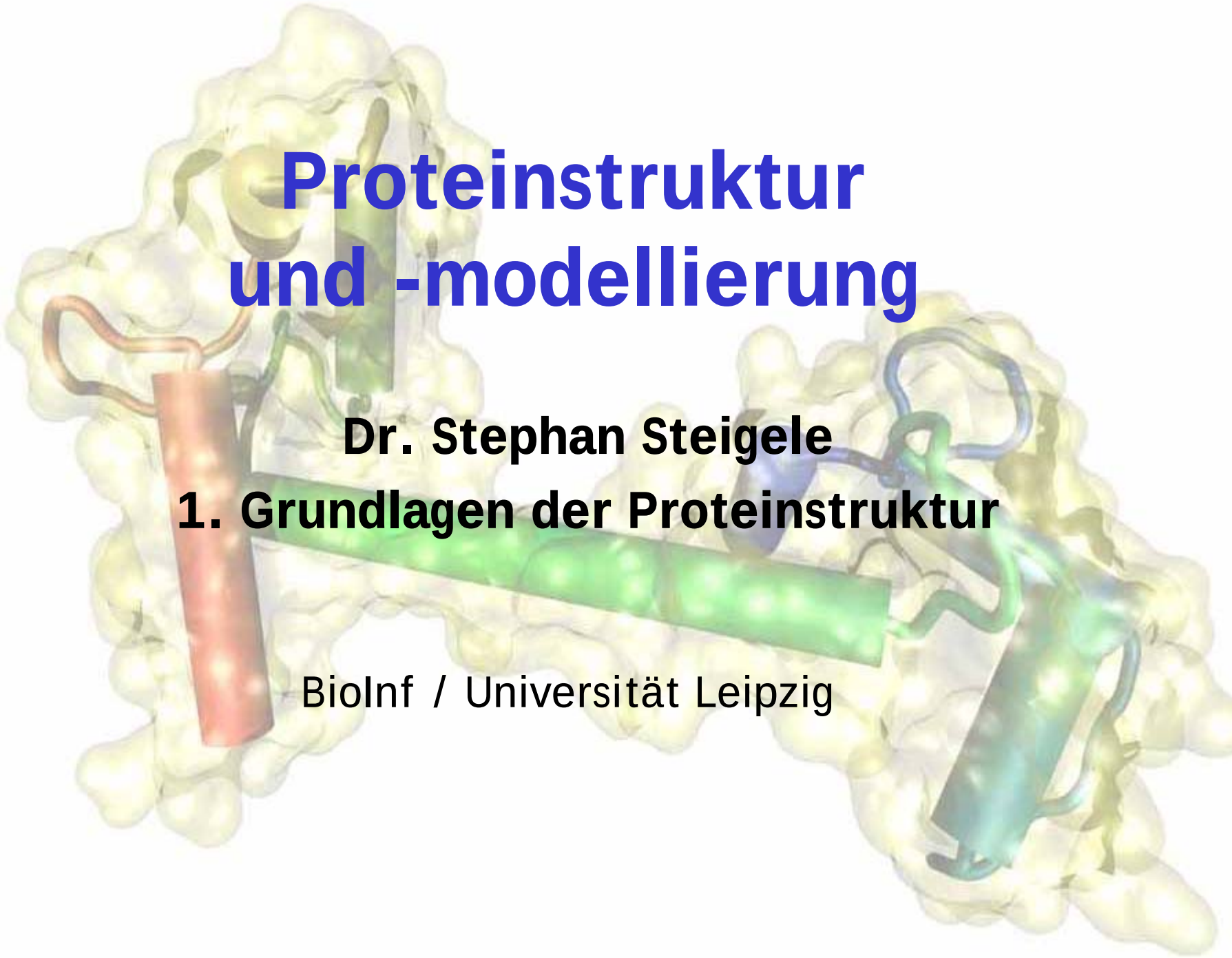
BALLView

Ein Werkzeug zur Visualisierung und Modellierung von Proteinstrukturen.

www.ballview.org







Proteinstruktur und -modellierung

Dr. Stephan Steigele

1. Grundlagen der Proteinstruktur

BioInf / Universität Leipzig

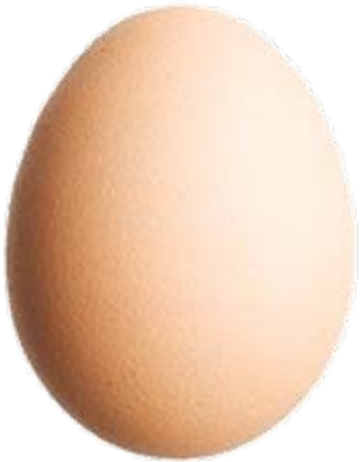
Überblick

- Aminosäuren
  - Struktur
  - Eigenschaften
- Proteine
  - Struktur und Visualisierung
  - Strukturhierarchie
  - Freiheitsgrade
  - Sekundärstrukturelemente
  - Tertiärstrukturen
  - Quartärstruktur

Einstieg: Warum Proteine?

Proteine:

- 50% der Trockenmasse tierischer Zellen
- Hauptfunktionen
  - Strukturelle Funktion, Bewegung…
  - Katalyse, Transport, Signalübertragung…
- Modellierung: Protein-Design, Wirkstoff-Entwurf



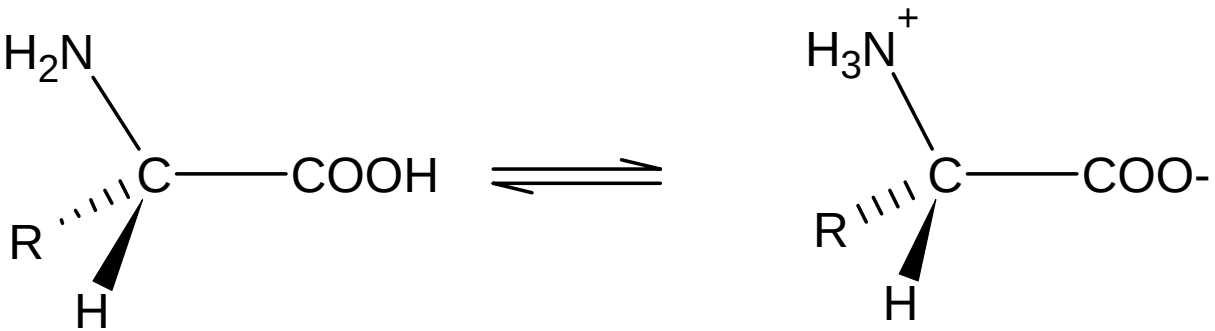
Zentrales Dogma

Ein Gen = ein Protein

- Sequenz der DNA bestimmt eindeutig mRNA, mRNA bestimmt eindeutig Sequenz des Proteins, Sequenz des Proteins bestimmt eindeutig die Struktur
- Bekannte Ausnahmen
  - Retroviren: kehren Richtung der Transkription um!
  - Prionen kennen mehr als eine stabile Struktur
  - Spleißvarianten des selben Gens

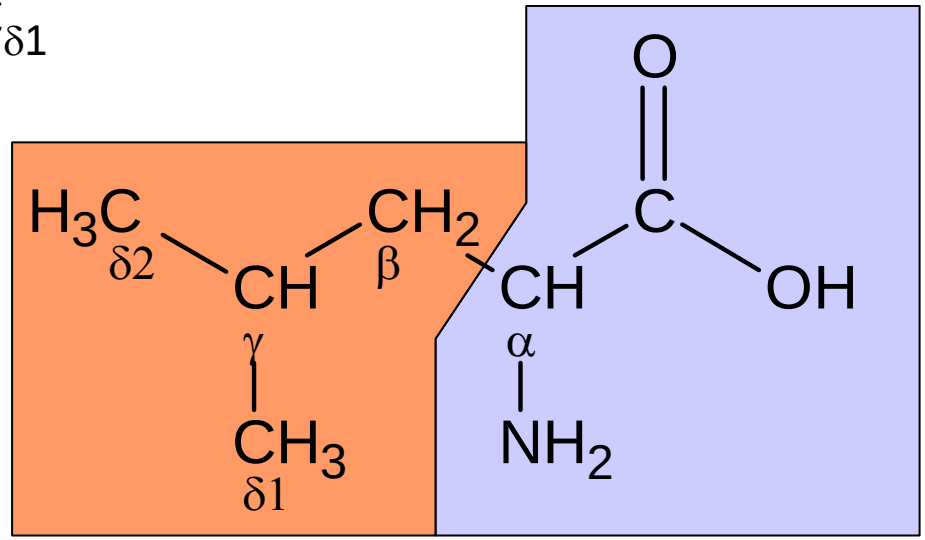
Aminosäuren I

- Proteine bestehen aus **α-Aminokarbonsäuren**
- Natürliche Aminosäuren (AS, aa)
  - Besitzen eine **Karbonsäurefunktion** –COOH
  - Besitzen eine **primäre Aminofunktion** –NH_2
  - Liegen gewöhnlich als Zwitterionen vor (– NH_3^+, –COO^-)
  - Meistens **chiral**: L-Aminosäuren (in S-Konfiguration)
- 20 **proteinogene** Aminosäuren
- Unterschiede liegen in den Seitenketten



Aminosäuren II

- Rückgrat
- Seitenkette
- Atome des Rückgrats: C, O, N, H, C_α, H_α
- Nummerierung der Seitenkettenatome
  - „Entfernung“ vom C_α: β, γ, δ, ε, η
  - Atome auf gleicher Ebene mit arabischen Ziffern, z.B. $C_{\delta 1}$



Aminosäuren III

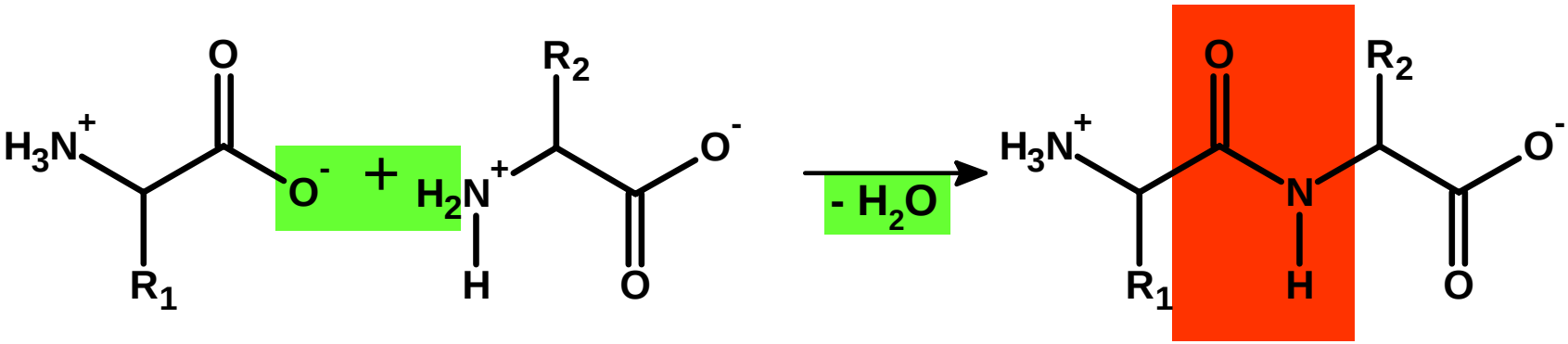
- Die 20 proteinogenen AS unterscheiden sich in ihren Seitenketten
- Benennnung üblicherweise mit **Ein- oder Drei-Buchstaben-Kürzeln** (*one letter code*, 1LC, *three letter code*, 3LC)

Name	3LC	1LC
Alanin	Ala	A
Cystein	Cys	C
Asparaginsäure	Asp	D
Glutaminsäure	Glu	E
Phenylalanin	Phe	F
Glycin	Gly	G
Histidin	His	H
Isoleucin	Ile	I
Lysin	Lys	K
Leucin	Leu	L

Name	3LC	1LC
Methionin	Met	M
Asparagin	Asn	N
Prolin	Pro	P
Glutamin	Gln	Q
Arginin	Arg	R
Serin	Ser	S
Threonin	Thr	T
Valin	Val	V
Tryptophan	Trp	W
Tyrosin	Tyr	Y

Peptidbindung I

- Amino- und Karbonsäurefunktion können unter **Kondensation** verknüpft werden
- Es entsteht eine **Peptidbindung**:



- Das entstandene **Dipeptid** kann mit weiteren AS verknüpft werden

Definitionen und Begriffe

(Oligo-)Peptid: 2 – 10 AS

Polypeptid: 10 – 100 AS

Protein: > 100 AS

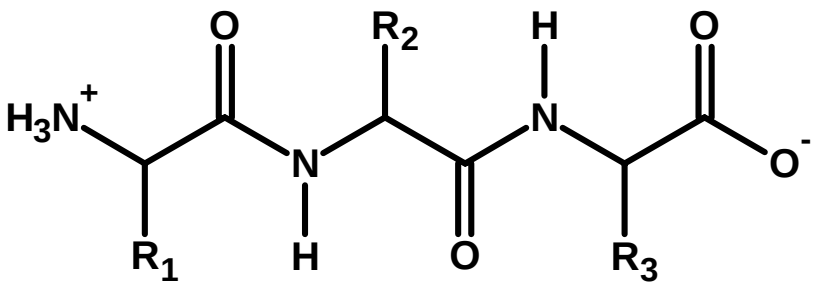
N-Terminus: das Ende mit der freien Aminofunktion

C-Terminus: das Ende mit der freien Carboxylfunktion

Sequenz: Abfolge der AS, gelesen vom N-Terminus zum C-Terminus

Bsp.:

N-Terminus



C-Terminus

Sequenz: $R_1R_2R_3$

Aminosäuren V - Klassifikation

- Es gibt viele Arten AS nach ihrem Rest zu klassifizieren
- Gängig sind z.B. Merkmale wie
  - Polar/unpolar
  - Geladen/ungeladen
  - Sauer/Basisch
  - Aliphatisch/aromatisch
  - Groß/klein
  - Schwefelhaltig/nicht schwefelhaltig
- In der Folge werden wir folgende Einteilung verwenden:
  - **Ungeladen** Ala, Gly, Phe, Ile, Met, Leu, Pro, Val
  - **Geladen** Asp, Glu, Lys, Arg
  - **Polar** Ser, Thr, Tyr, His, Cys, Asn, Gln, Trp

Aminosäuren VI - Ungeladene

Ala	R = $-CH_3$	
Gly	R = –H	(kleinste AS!)
Phe	R = $-CH_2-C_6H_5$	(aromatisch!)
Ile	R = $-CH(CH_3)-CH_2-CH_3$	
Met	R = $-CH_2-CH_2-$ **S** $-CH_3$	(schwefelhaltig!)
Leu	R = $-CH_2-CH(CH_3)-CH_3$	
Pro	R = $-CH_2-CH_2-CH_2-$ **N** –	(Iminosäure!)
Val	R = $-CH(CH_3)-CH_3$	

Aminosäuren VII - Geladene

Asp	R = – CH_2 – **COO^-**
Glu	R = – CH_2 – CH_2 – **COO^-**
Lys	R = – CH_2 –CH_2 –CH_2 –CH_2 –**NH_3^+**
Arg	R = – CH_2 –CH_2 –CH_2 –NH – C(=**NH_2^+**) – NH_2

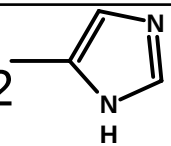
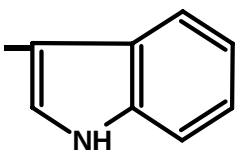
Asp, Glu: saure AS, zusätzliche Karbonsäurefunktion
Lys, Arg: basische AS, zusätzliche Aminofunktion

Die geladenen AS liegen bei physiologischen Bedingungen in der Regel **deprotoniert** (sauer) bzw. **protoniert** (basisch) vor

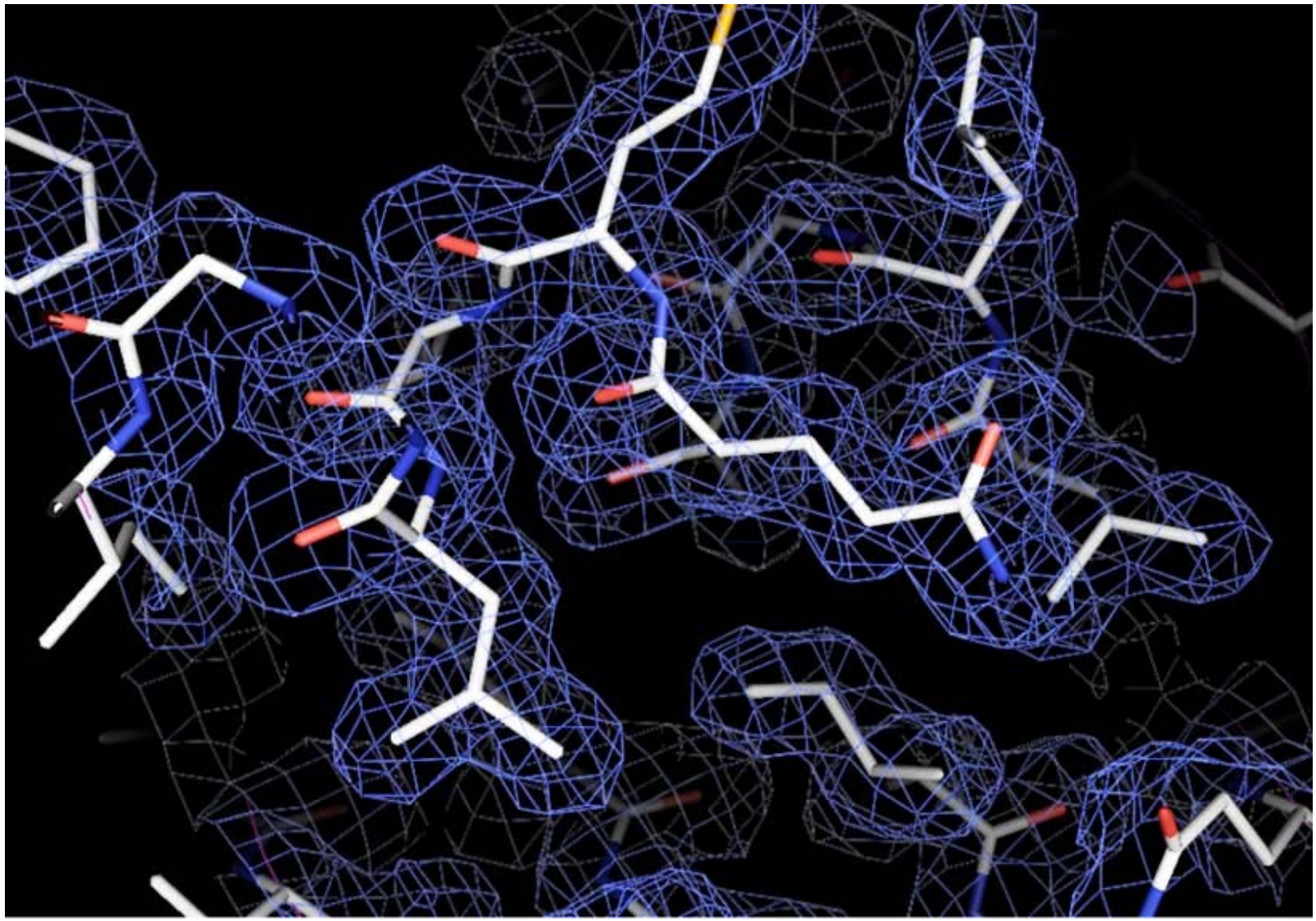
⇒ saure AS: negativ geladen

⇒ basische AS: positiv geladen

Aminosäuren VIII - Polare

Ser	R = – CH_2 –OH	
Thr	R = – CH(OH) –CH_3	
Tyr	R = – CH_2-C_6H_4 –OH	(aromatisch!)
His	R = – CH_2 	
Cys	R = –CH_2 – SH	(Thiol!)
Asn	R = –CH_2 –$CONH_2$	
Gln	R = –CH_2 –CH_2-$CONH_2$	
Trp	R = – CH_2 	(aromatisch!)

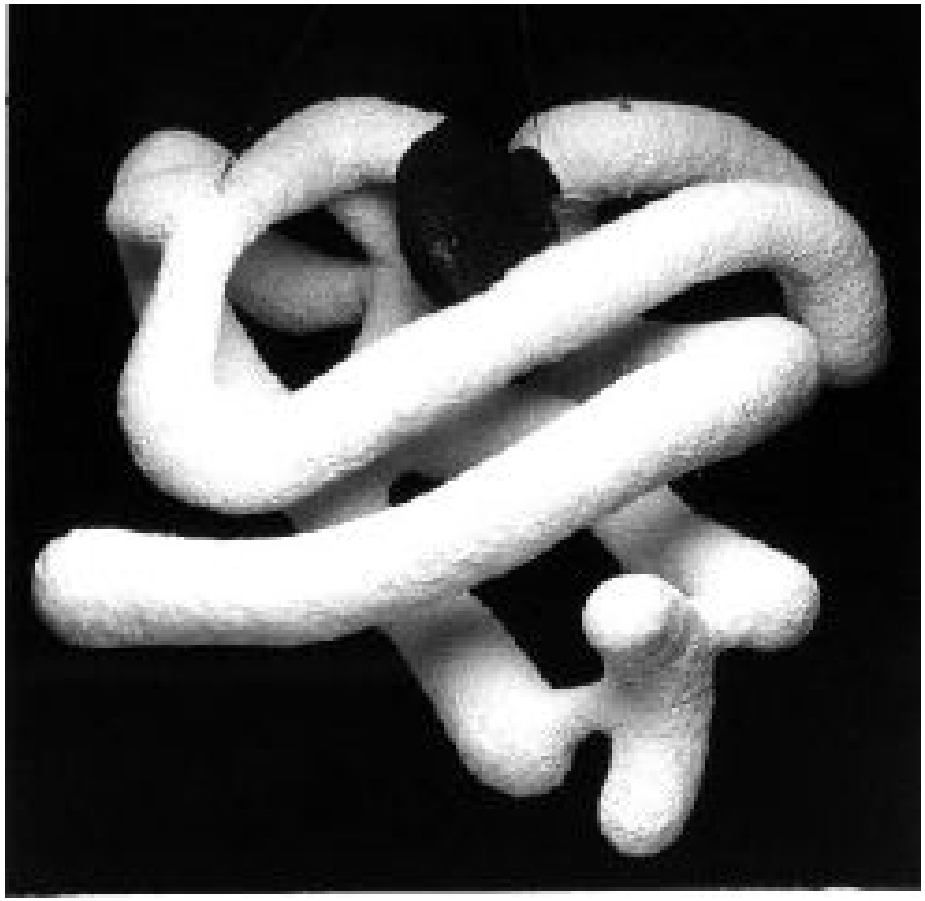
Wie sehen Proteine aus?

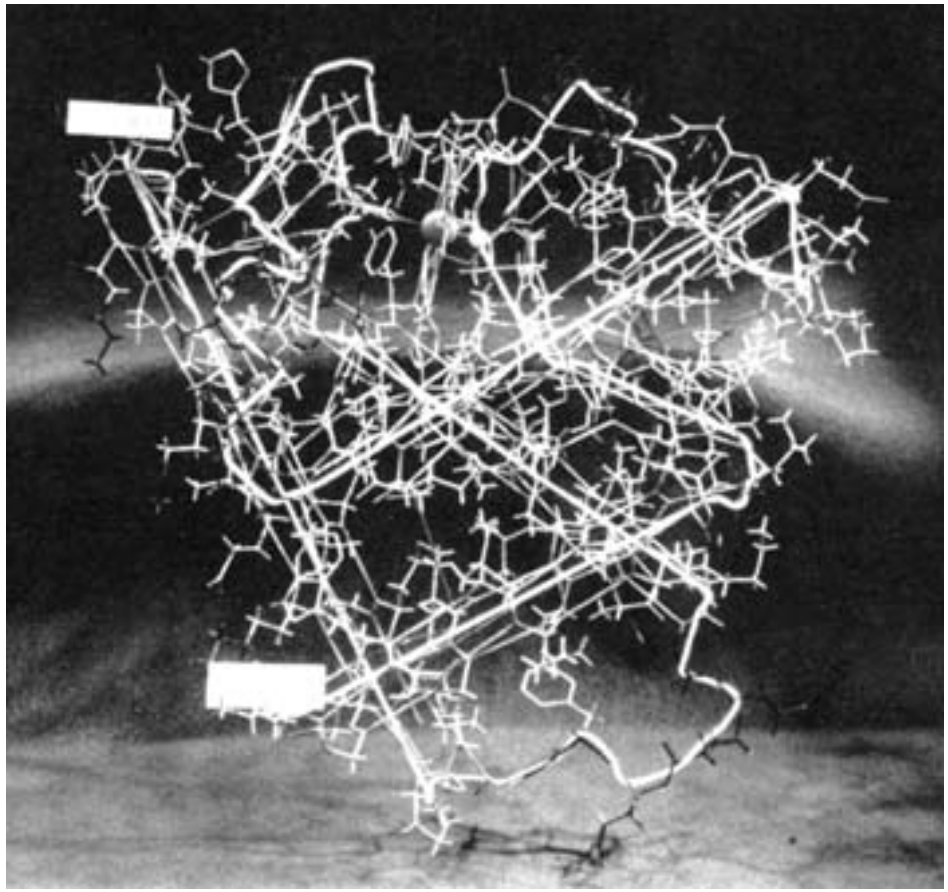


Wie sehen Proteine aus?

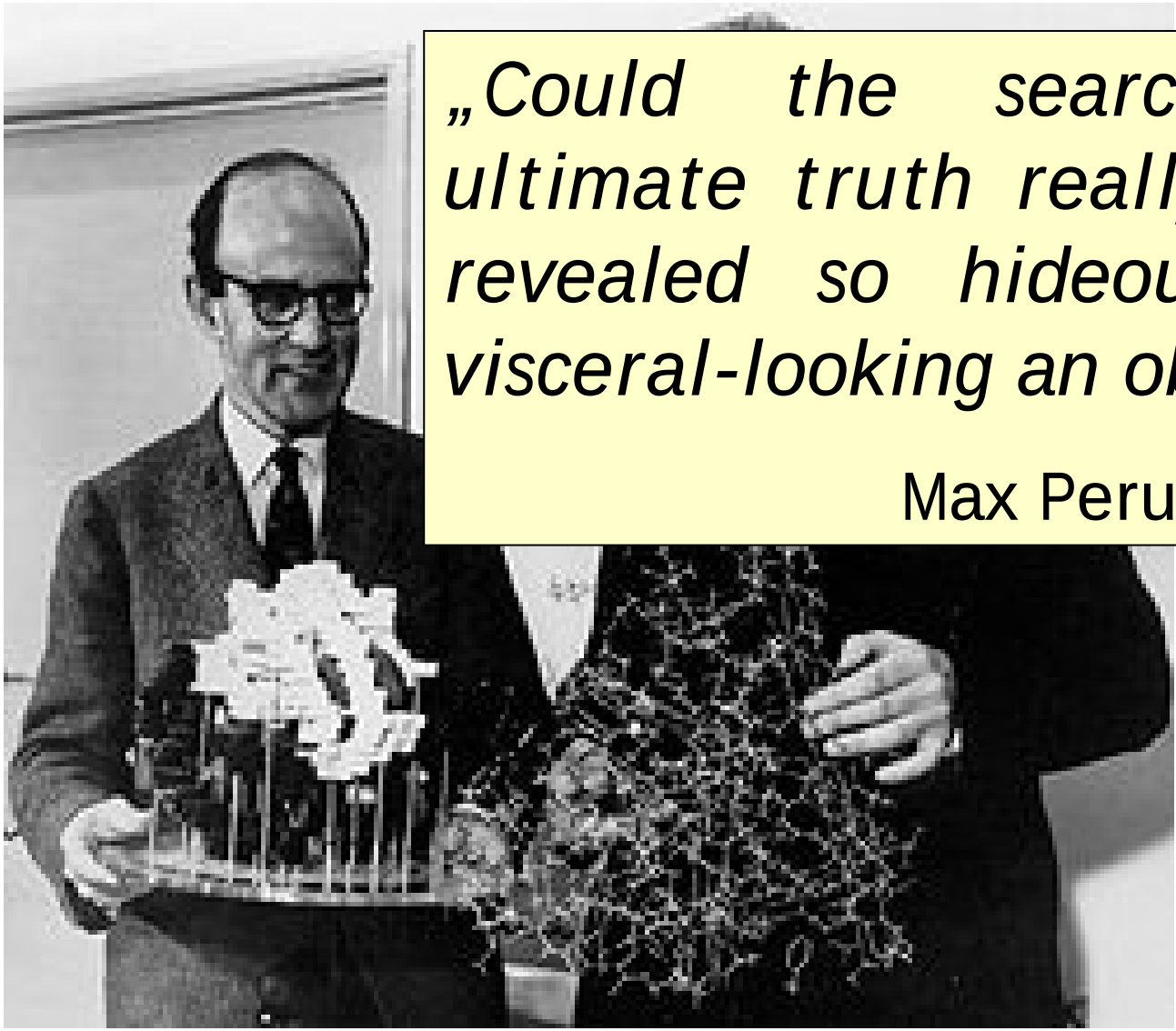
Wie sehen Proteine aus?

Wie sehen Proteine aus?





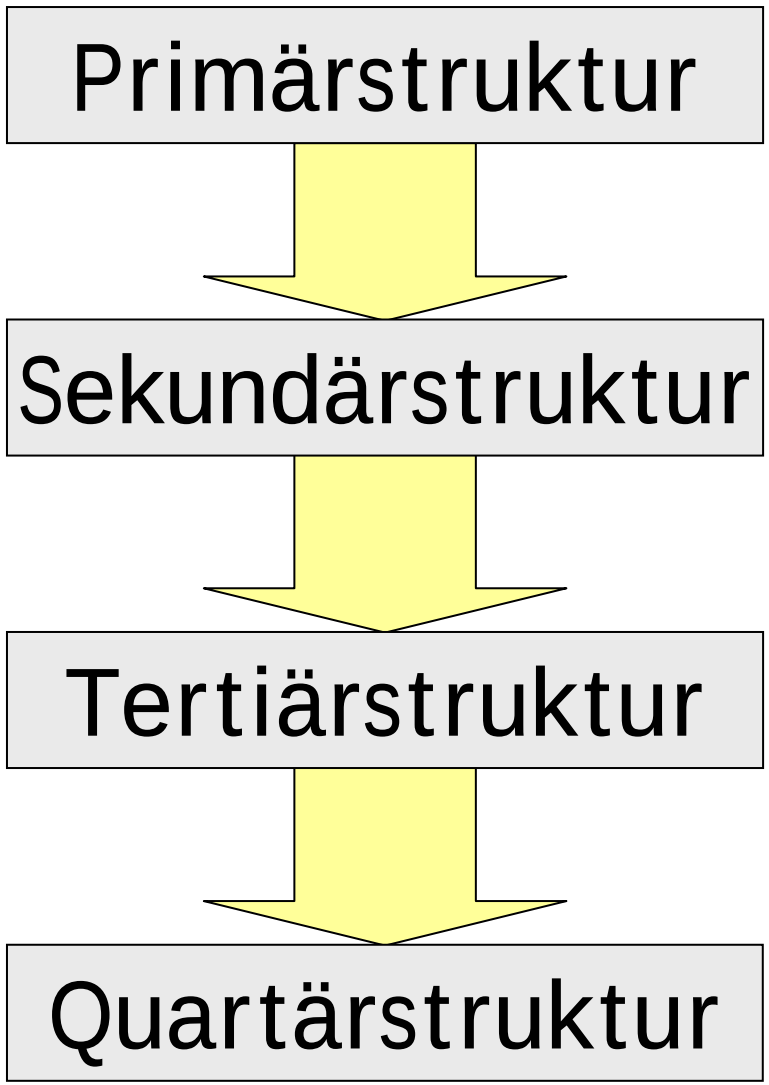
Wie sehen Proteine aus?



„Could the search for ultimate truth really have revealed so hideous and visceral-looking an object?“

Max Perutz, 1964

Proteinstruktur – Überblick



Sequenz: . . . **LGFCYWS** . . .

Relative Anordnung der AS zueinander,
Regelmäßige *Sekundärstrukturelemente*

Anordnung der Sekundärstrukturelemente
im Raum (*Faltung*)

Anordnung der einzelnen Proteine in
größeren *Komplexen*

Proteinstruktur – Überblick

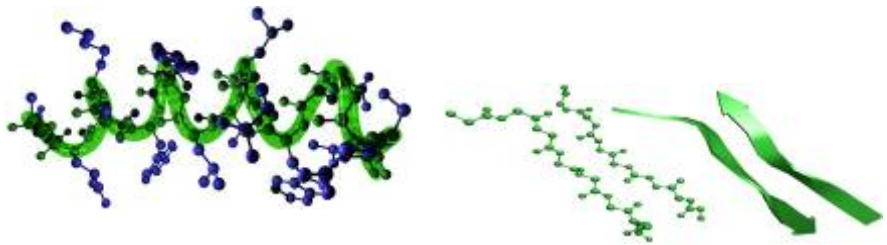
Primärstruktur

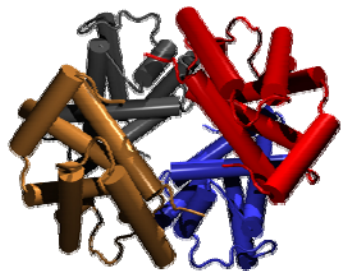
Sekundärstruktur

Tertiärstruktur

Quartärstruktur

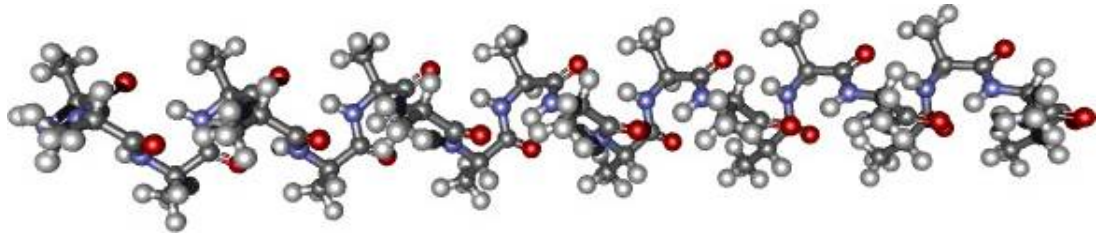
Sequenz: . . . **LGFCYWS** . . .





Sekundärstrukturelemente

- **Sekundärstrukturen** sind repetitive oder nicht repetitive Teilstrukturen, die durch eine **ausgezeichnete Geometrie** definiert sind
- Diese Geometrien sind **energetisch** besonders **günstig**
- Stabilisiert werden die Sekundärstrukturen durch intramolekulare **Wasserstoffbrückenbindungen**
- **Repetitive**
  - α-Helix
  - β-Faltblatt
- **Nichtrepetitive**
  - β-Schleifen

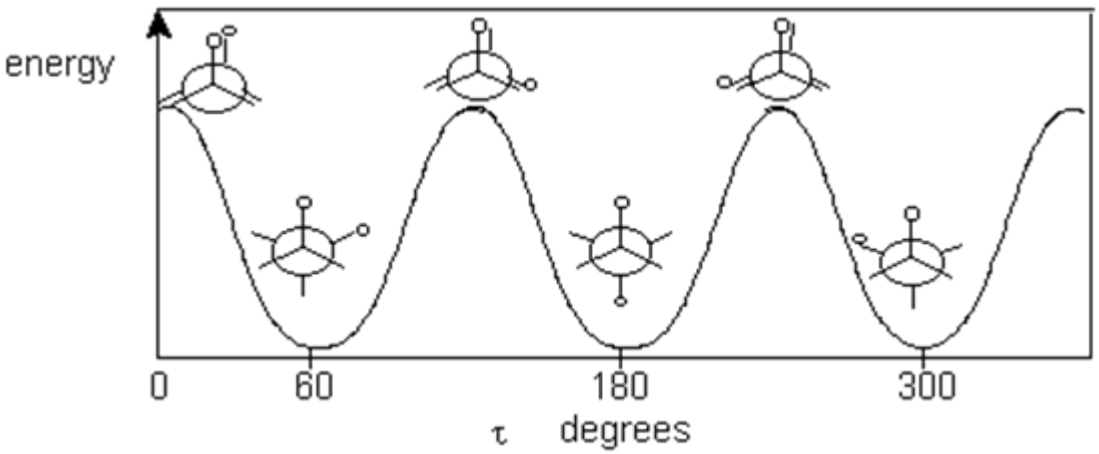


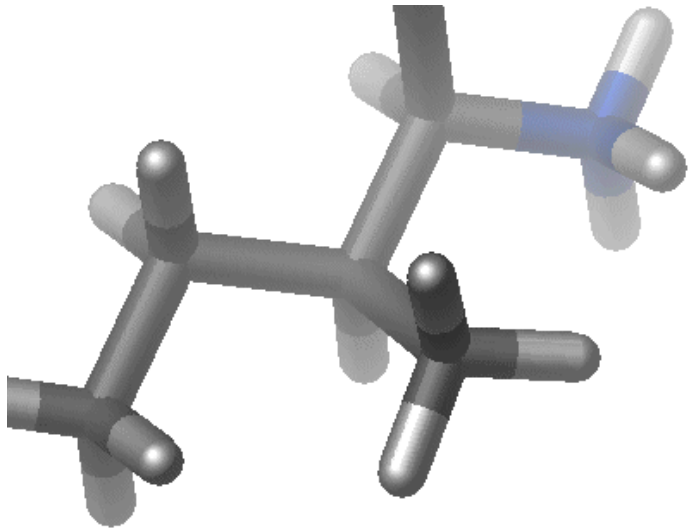
3_{10}-Helix

- Repetitive Sekundärstrukturelemente wiederholen die gleiche Geometrie mehrmals hintereinander

Torsionswinkel

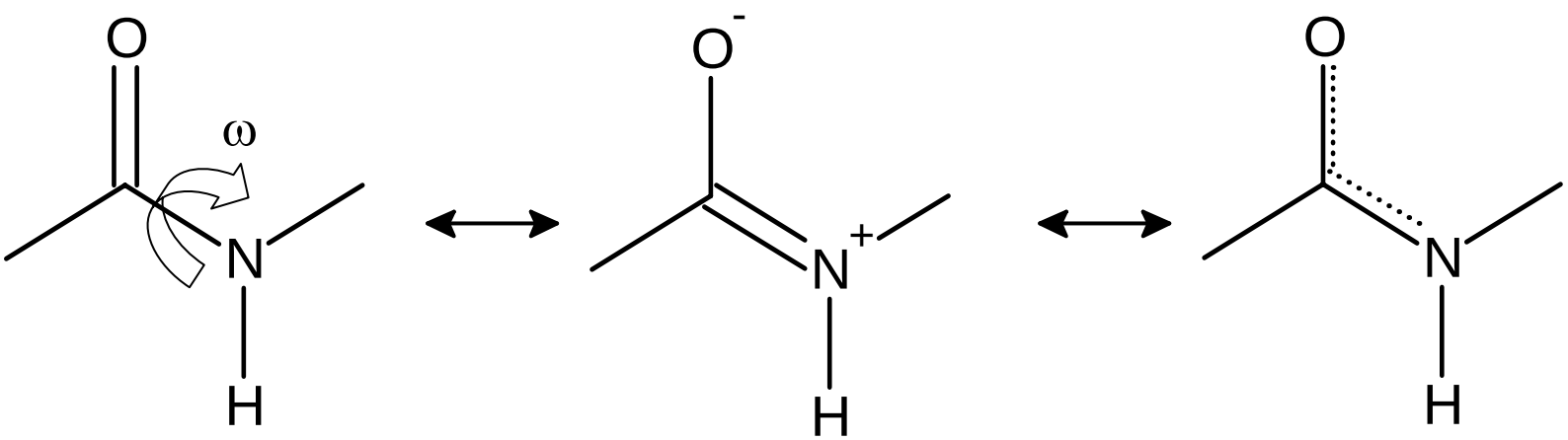
- Unterschiedliche Geometrien der Sekundärstrukturelemente entstehen durch **Rotation um die Einfachbindungen des Rückgrats**
- Rotation um Bindungen werden durch **Torsionswinkel** beschrieben
- Deformation bezüglich Bindungslängen und –winkel erfordert höhere Energien als Änderung der Torsionswinkel
- Torsionsbarrieren für Seitenketten liegen bei etwa 20 kJ/mol





Peptidbindung II – Geometrie

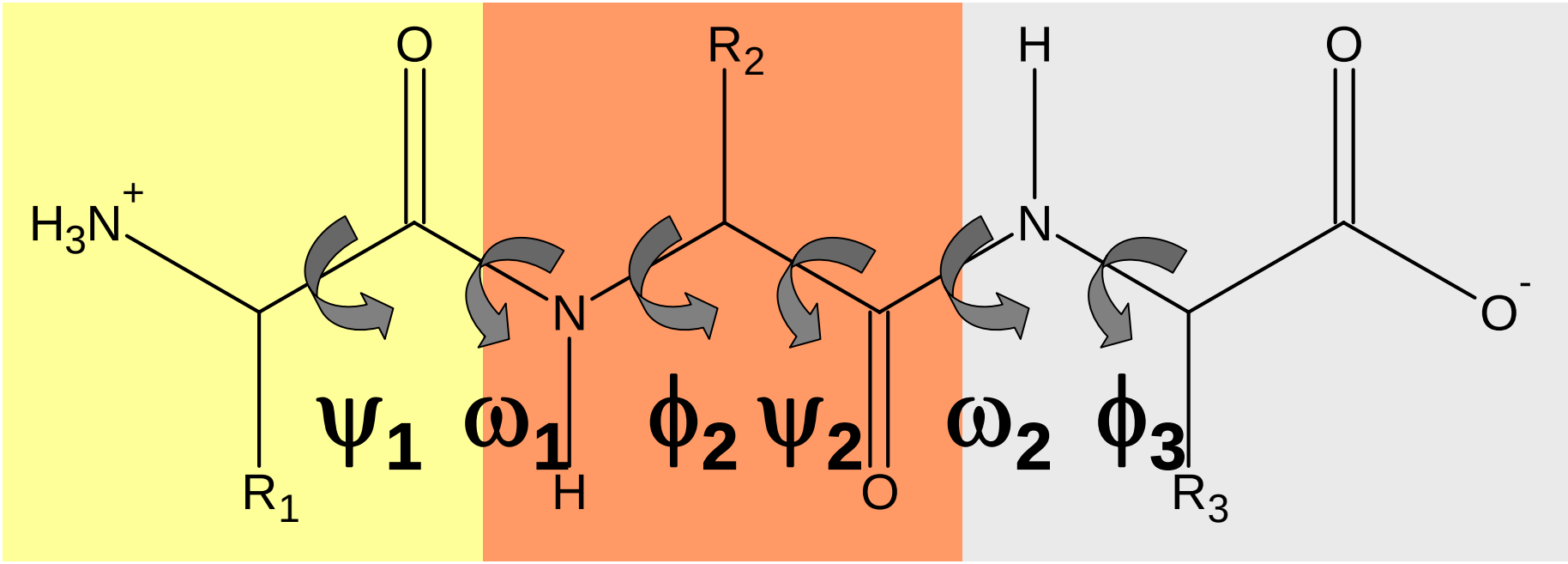
- C-N-Bindung hat ca. 40% Doppelbindungscharakter
- Ausgelöst durch Resonanz
- Frei Drehbarkeit behindert, planare Konformation bevorzugt



- Zwei Konformere: trans ($\omega = 180°$) und cis ($\omega = 0°$)
- Trans-Konformer um ca. 8 kJ/mol stabiler als cis-Konformer, Torsionsbarriere mit 160 kJ/mol recht hoch

Peptidbindung III – Torsionen

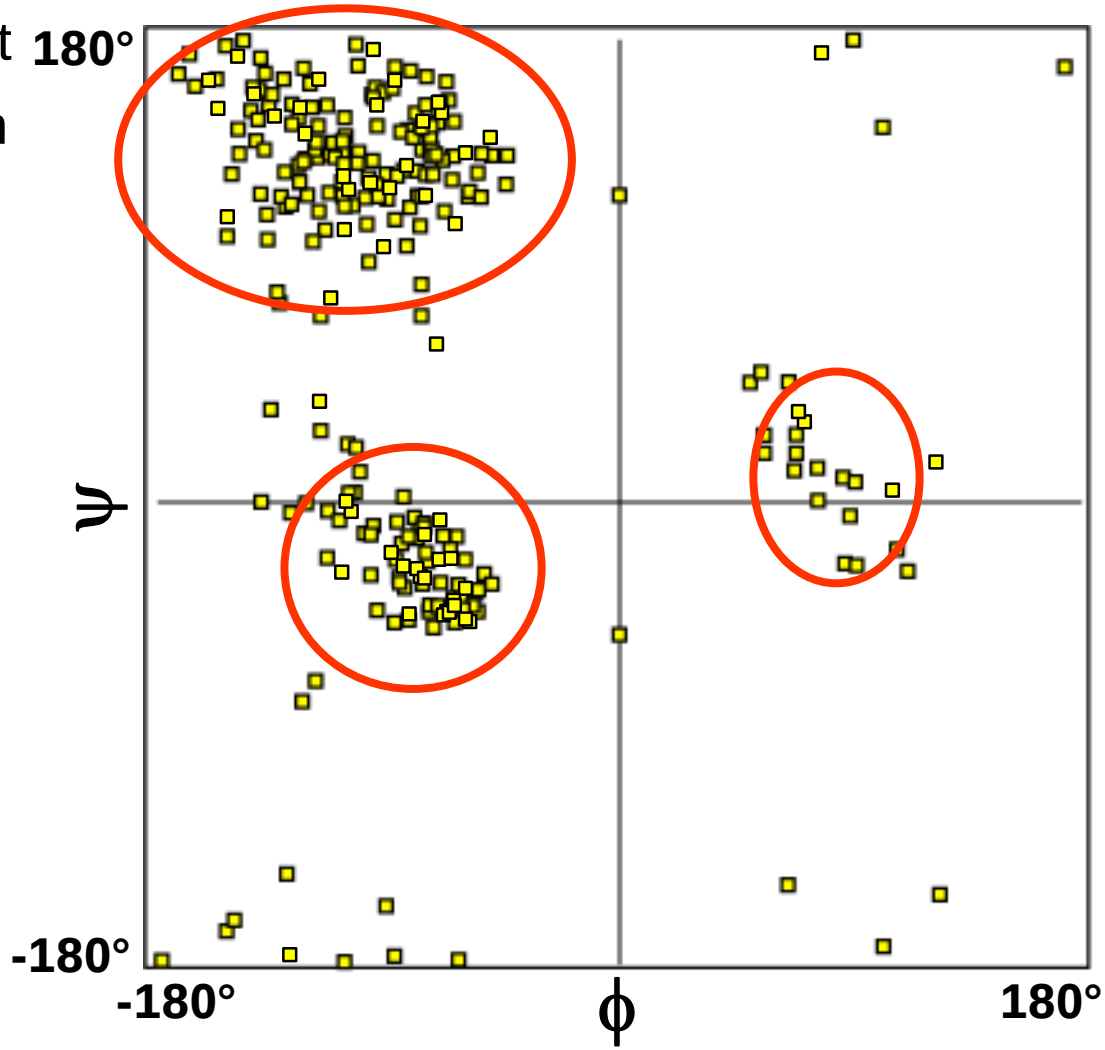
- Drei Torsionswinkel pro AS
  - ϕ entlang der Bindung zwischen N-C_α
  - ψ entlang der Bindung zwischen C_α-C
  - ω entlang der Peptidbindung
- Am N-Terminus entfällt ϕ, am C-Terminus ψ



Torsionswinkel – Ramachandran-Plot

- Im **Ramachandran-Plot** stellt man jeweils Paare (ϕ, ψ) von Torsionswinkeln einer AS dar
- Bestimmte Torsionswinkel-kombinationen sind energetisch bevorzugt, bestimmte sterisch ausgeschlossen
- **Beispiel**

  der Ramachandran-Plot des Proteinkomplexes Trypsin/BPTI (2PTC)

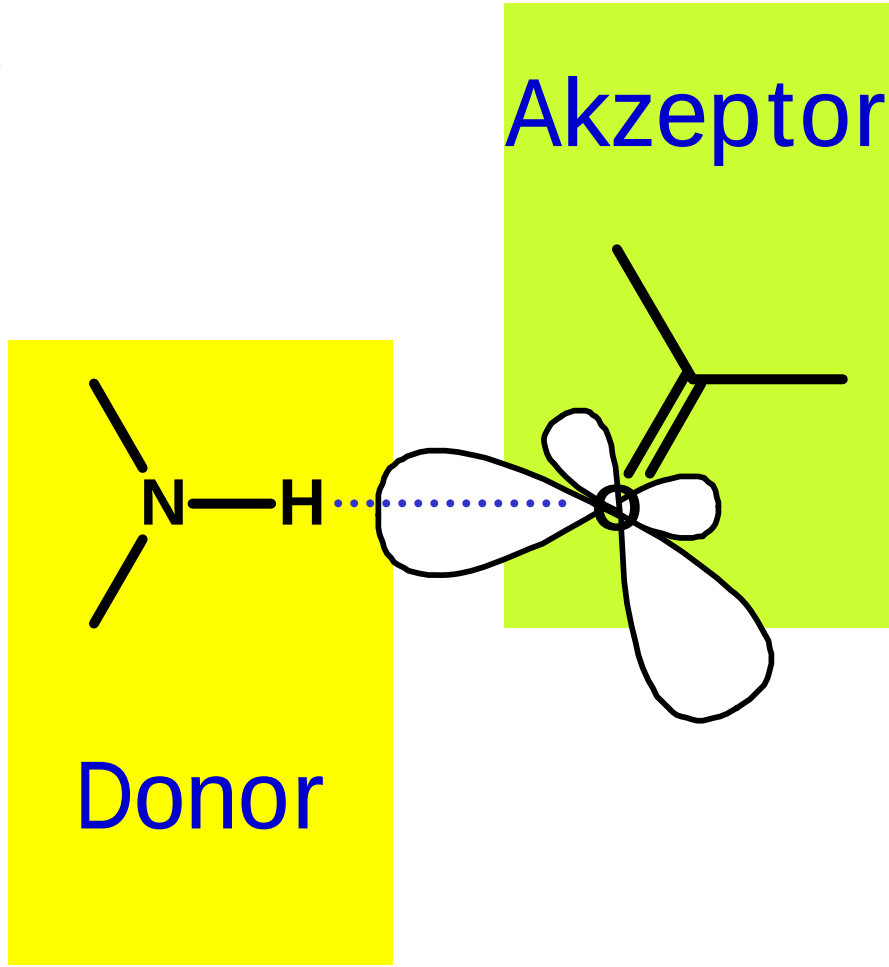


Wasserstoffbrücken

- Wasserstoffe gebunden an stark elektronegative Partner sind polar (z.B. an N, O, F)
- **Polare H** an Donor D wechselwirken mit freien Elektronenpaaren an Akzeptor A

  D—H⋯⋯A

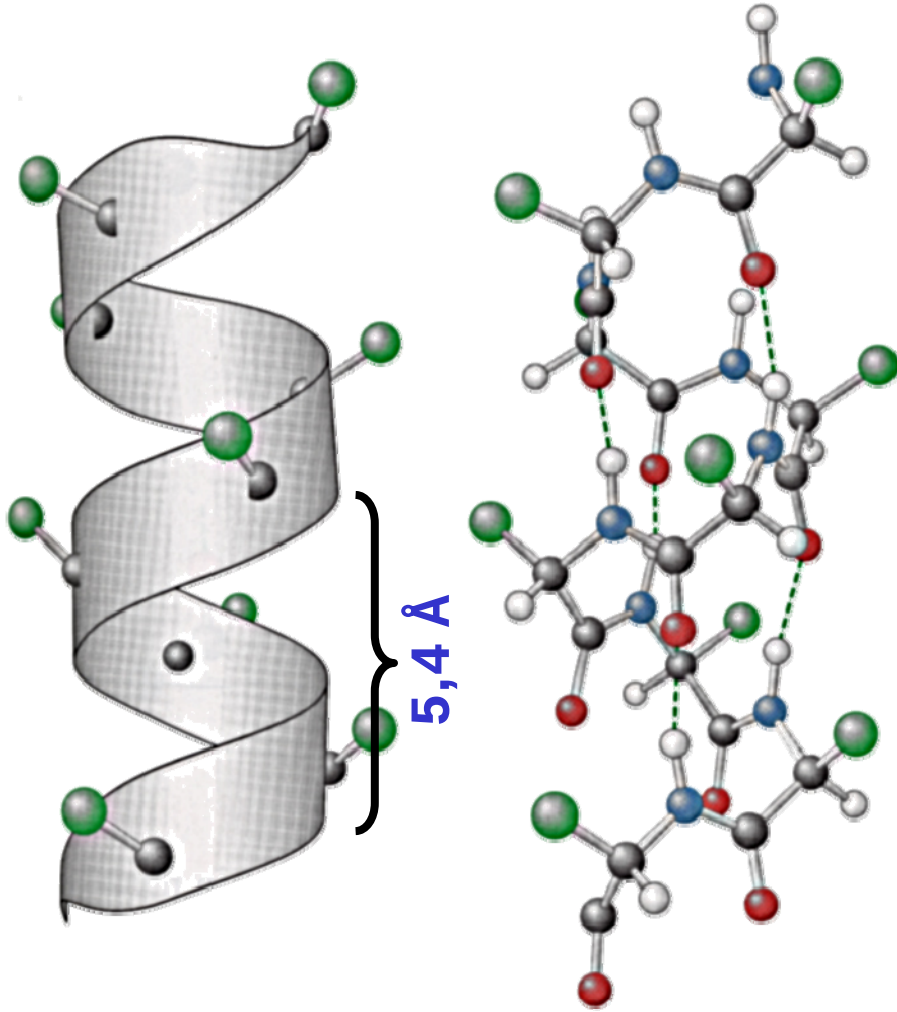
- Grenzfall zwischen kovalenter Bindung und intermolekularer WW



Sekundärstruktur – α-Helices

- α-Helix: meist rechtsgängige Helix
- pro Windung
  - 3,6 AS
  - 5,4 Å
- Stabilisiert durch regelmäßige H-Brücken (i → i + 4)
- Torsionswinkel

  $(\phi, \psi) = (-60°, -50°)$



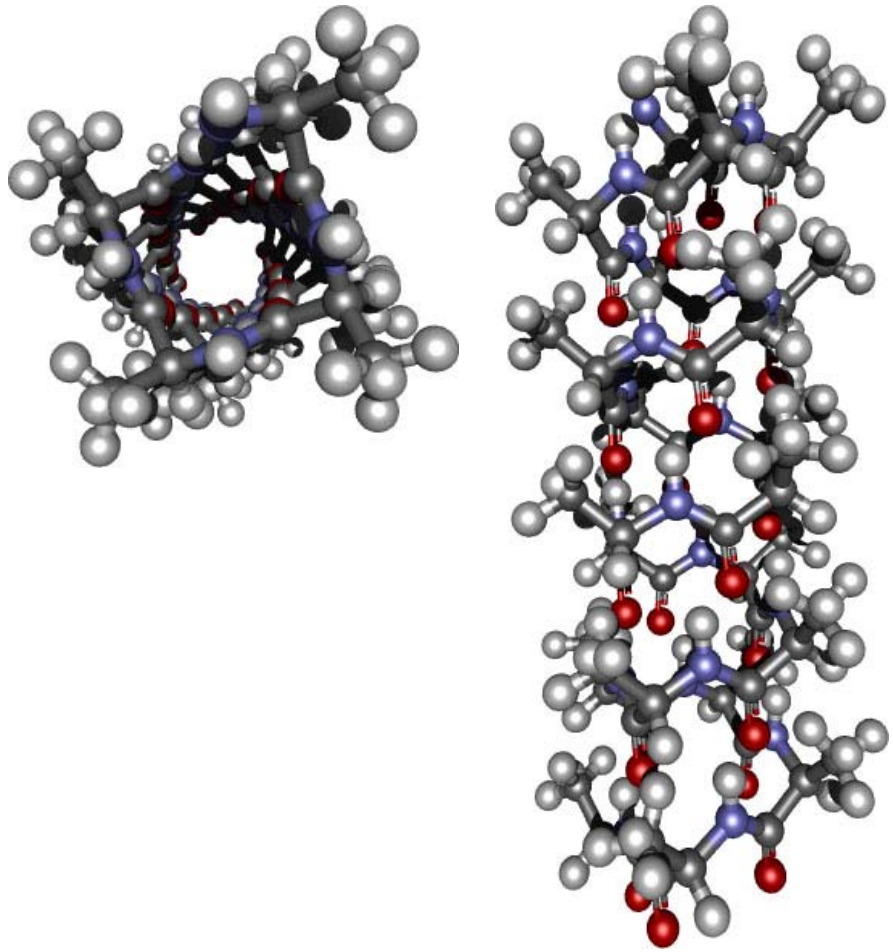
π–Helix

- π-Helices sind weniger eng gepackt als α-Helices
- H-Brücke zwischen den AS

$i = i + 5$

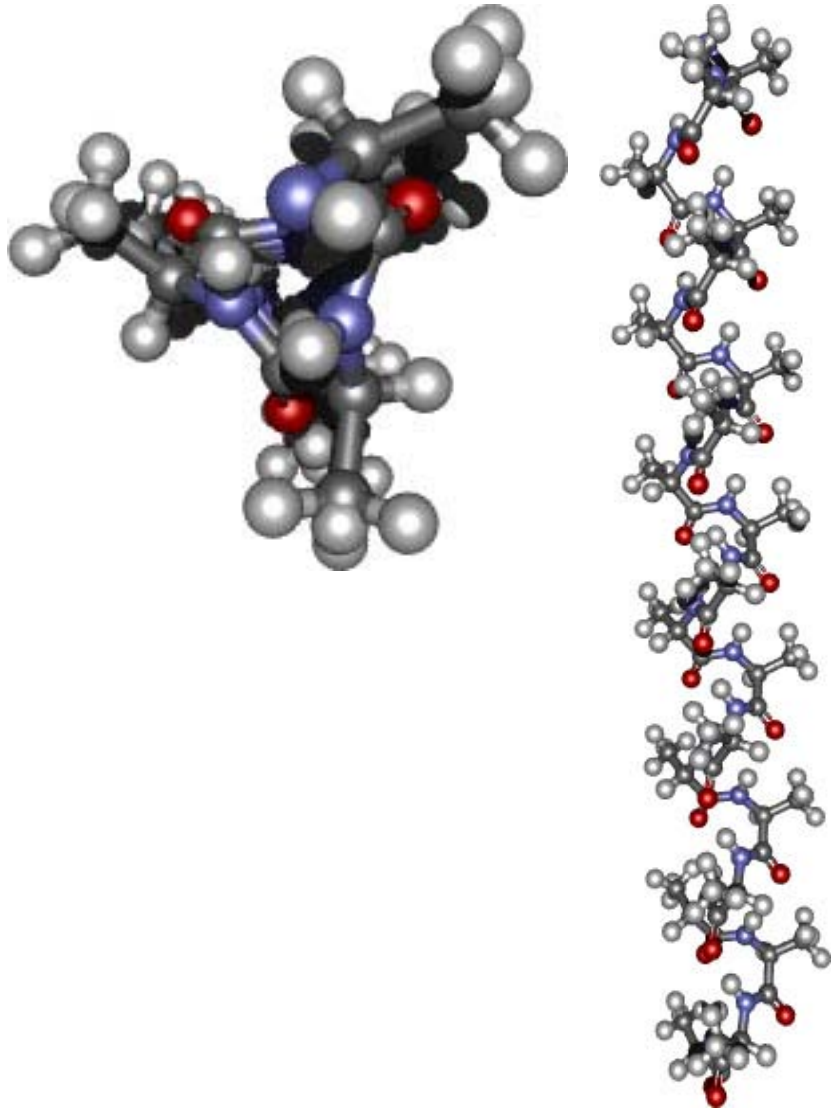
- Loser gepackt, als α-Helix
  - Atome berühren sich nicht an der Helixachse

  - Loch in der Mitte
- π-Helices sind sehr selten und immer kurz (meist weniger als eine Windung)



3_{10}-Helix

- 3_{10}-Helices sind enger gepackt als α-Helices
- 3 AS pro Windung
- H-Brücke zwischen AS *i* ! *i* + 3
- Sehr selten, immer kurz (meist weniger als eine Windung)
- Helixachse geschlossen

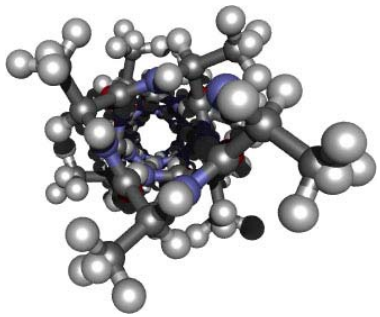


Helices – Überblick

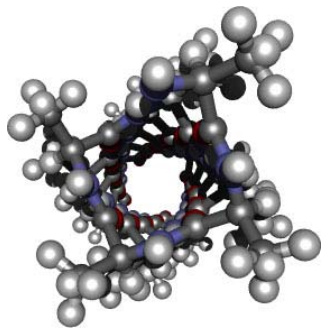
- Rechtsgängige α-Helices sind die mit Abstand häufigste Form von Helices
- Rechtsgängige α-Helices können sehr lang werden (~40 AS), im Durchschnitt sind sie etwa 10 AS
- Sehr selten sind linksgängige α-Helices: diese sind sterisch ungünstig und daher nur 3-5 AS lange Stücke
- π- und 3_{10}-Helices sind ebenfalls selten
  - In der Regel nur als kurze Helixstücke (eine Windung)
  - Oft am Ende von α-Helices
  - Energetisch ungünstiger als α-Helices, da Rückgrat zu lose/dicht gepackt

Helices – Überblick

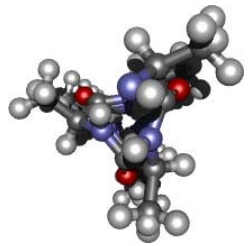
Helixart	Abstand H-Brücken	# AS pro Windung	ϕ/ψ (optimal)	Packung
α	$i \rightarrow i + 4$	3,6	-58° / -47°	optimal
π	$i \rightarrow i + 5$	4,4	-57° / -70°	lose
3_{10}	$i \rightarrow i + 3$	3	-74° / -4°	eng



α



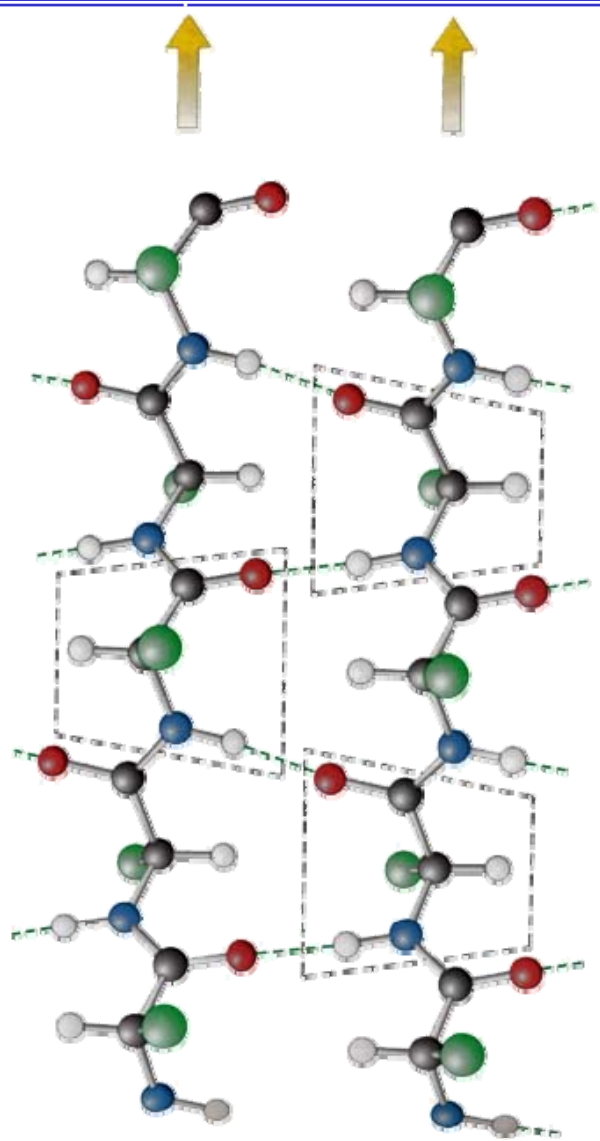
π



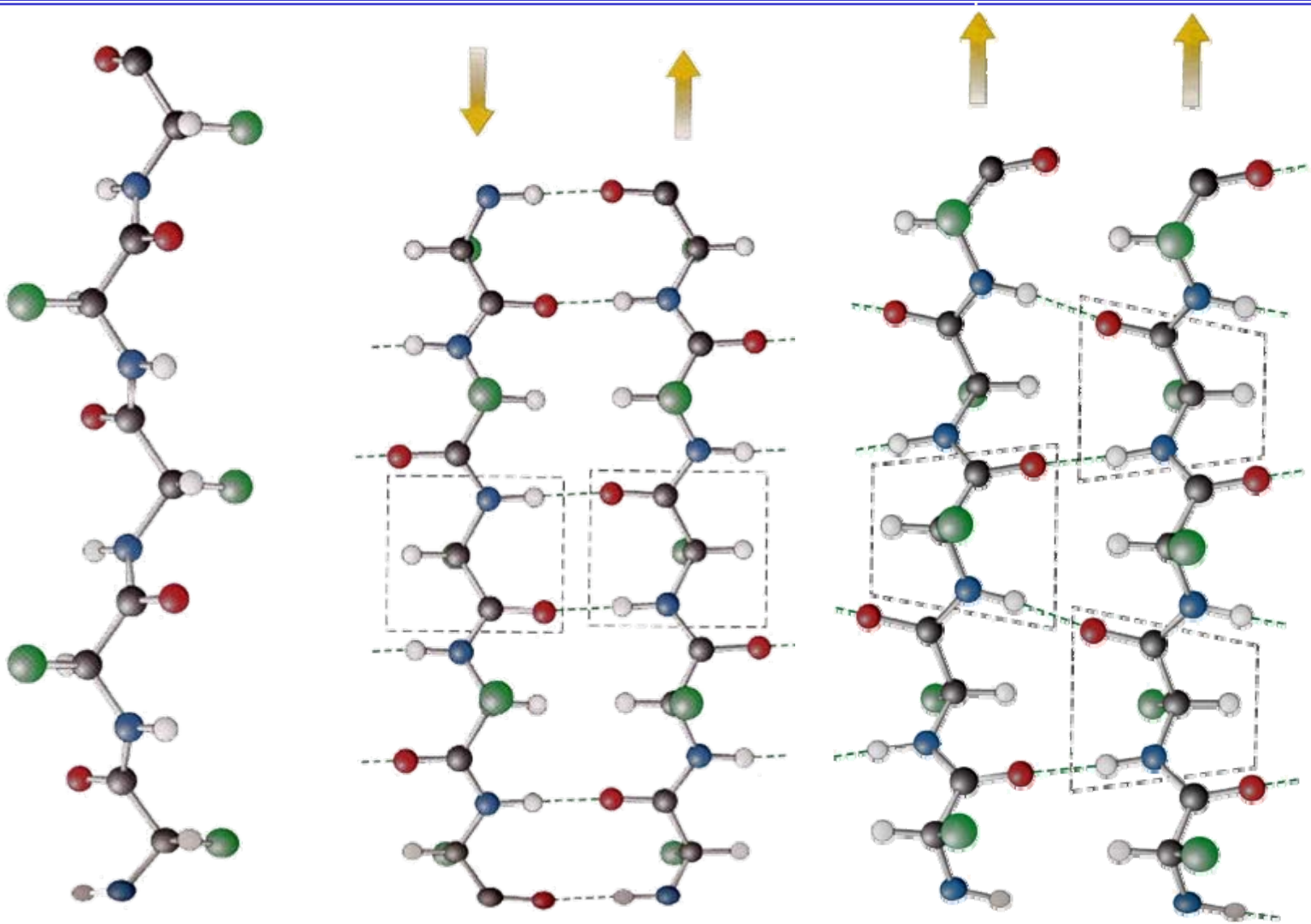
3_{10}

β-Faltblätter

- **Faltblätter** (*sheets*) bestehen aus mehreren parallelen oder antiparallelen **Strängen** (*strands*)
- Verbunden durch H-Brücken des Rückgrats (C=O → H-N)
- Abstand zwischen Strängen ~3.5 Å
- Torsionswinkel (ϕ, ψ)
  - **Parallel** (-120°, 115°)
  - **Antiparallel** (-140°, 135°)



Sekundärstruktur – β-Faltblätter

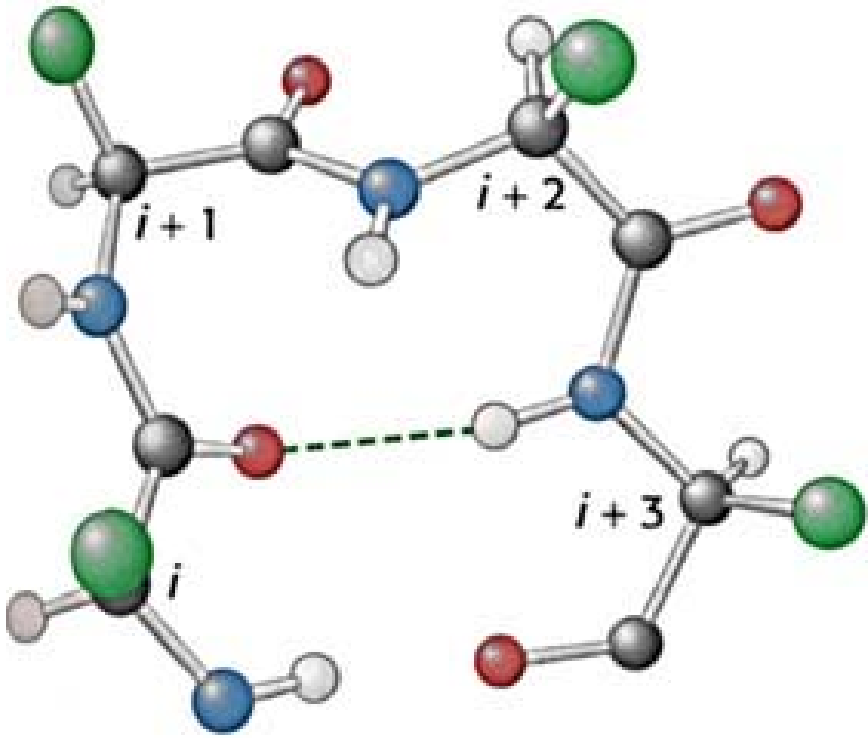


Berg, Tymoczko, Stryer: Biochemistry, 5th ed.

Schleife, Turn

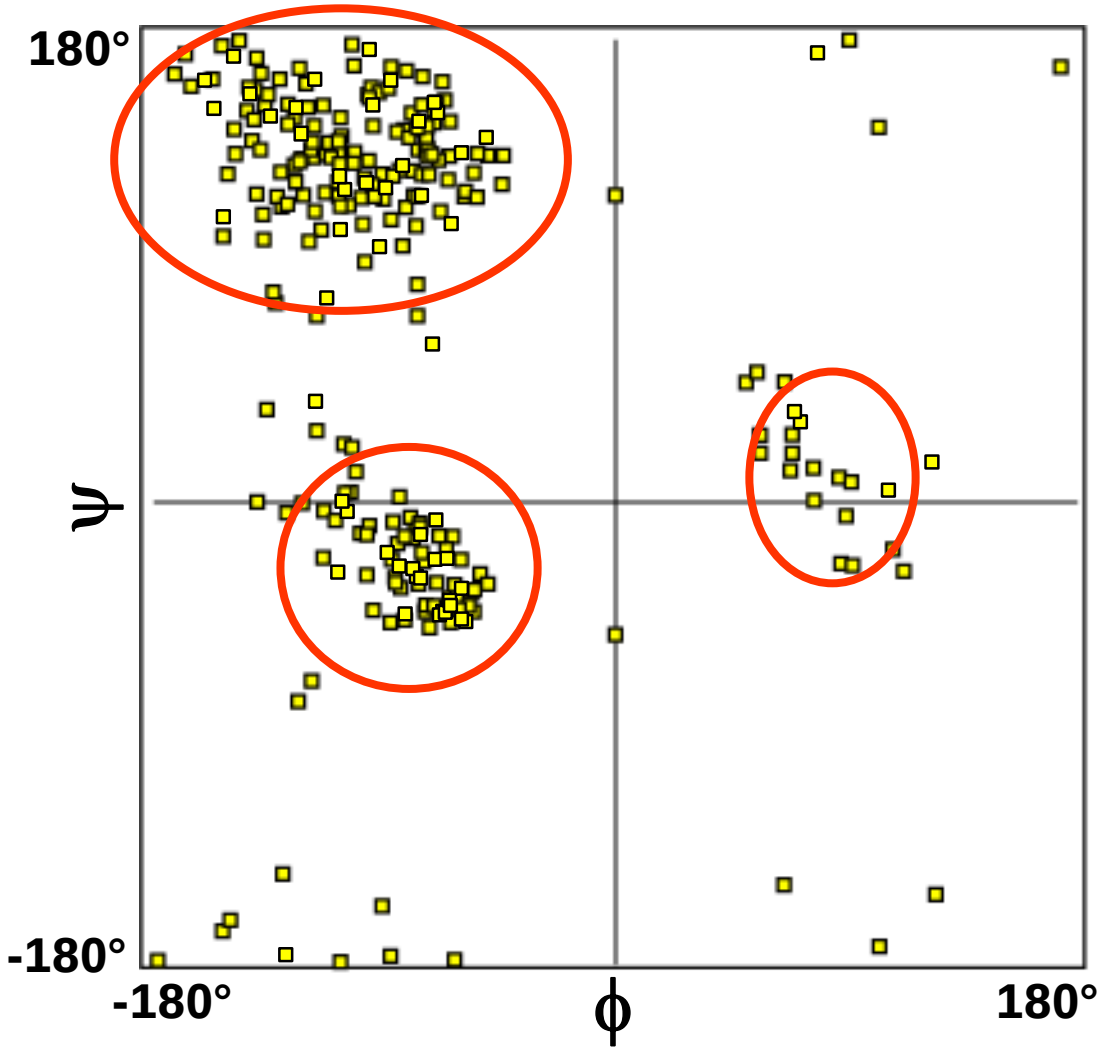
- Repetitive Strukturelemente (Helices, Faltblätter) machen etwa 50% eines Proteins aus
- Nichtrepetitive Abschnitte sind die **Schleifen** (*loops*)
- *Turns* sind enge 180°-Schleifen aus mindestens drei AS
- *Turn* wird durch H-Brücke

  i → i + 3 stabilisiert
- Oft an Wechselwirkungen mit anderen Proteinen oder am aktiven Zentrum beteiligt



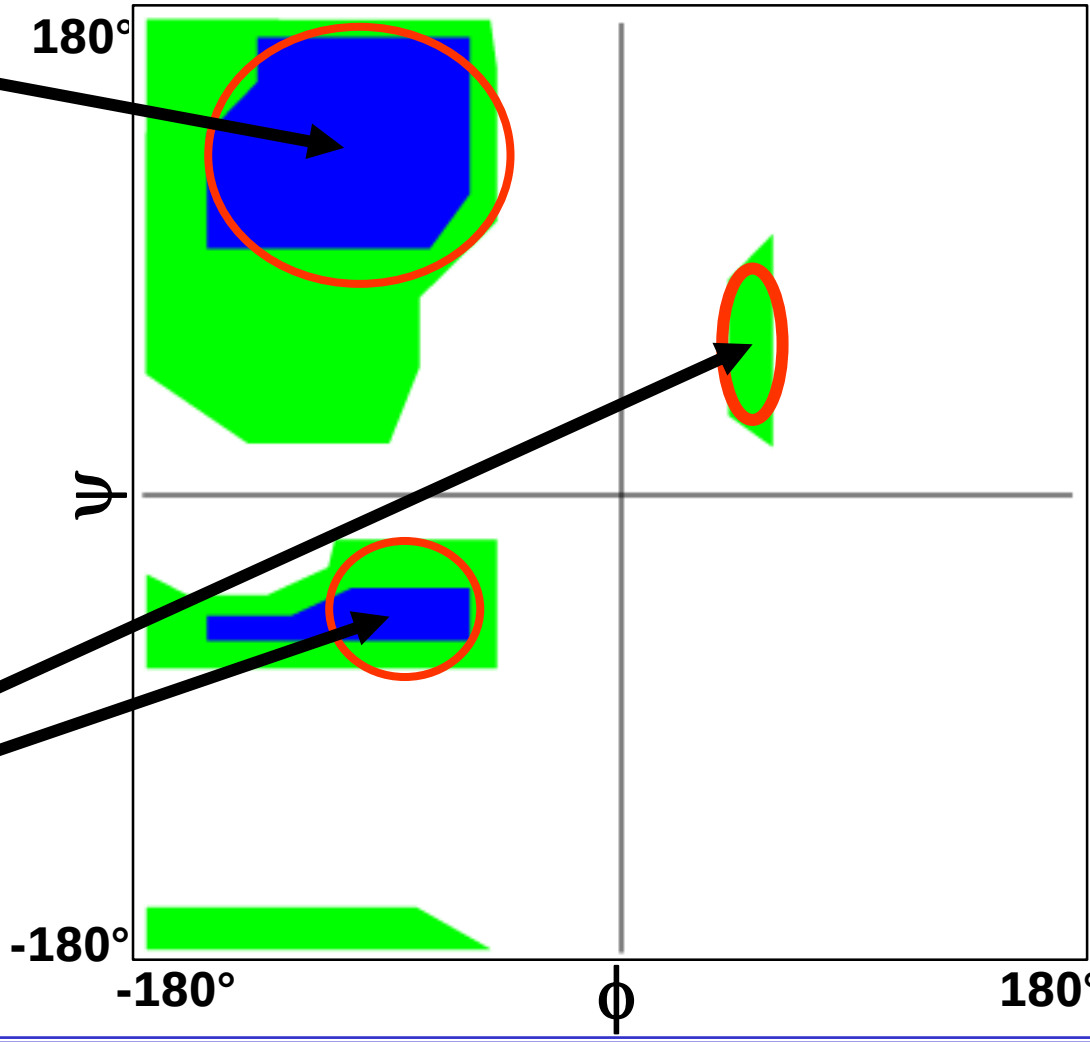
Torsionswinkel – Ramachandran-Plot

- Sekundärstruktur-elemente besitzen wohldefinierte Winkelkombinationen
- Ramachandran-Plot zeigt diese Regionen deutlich



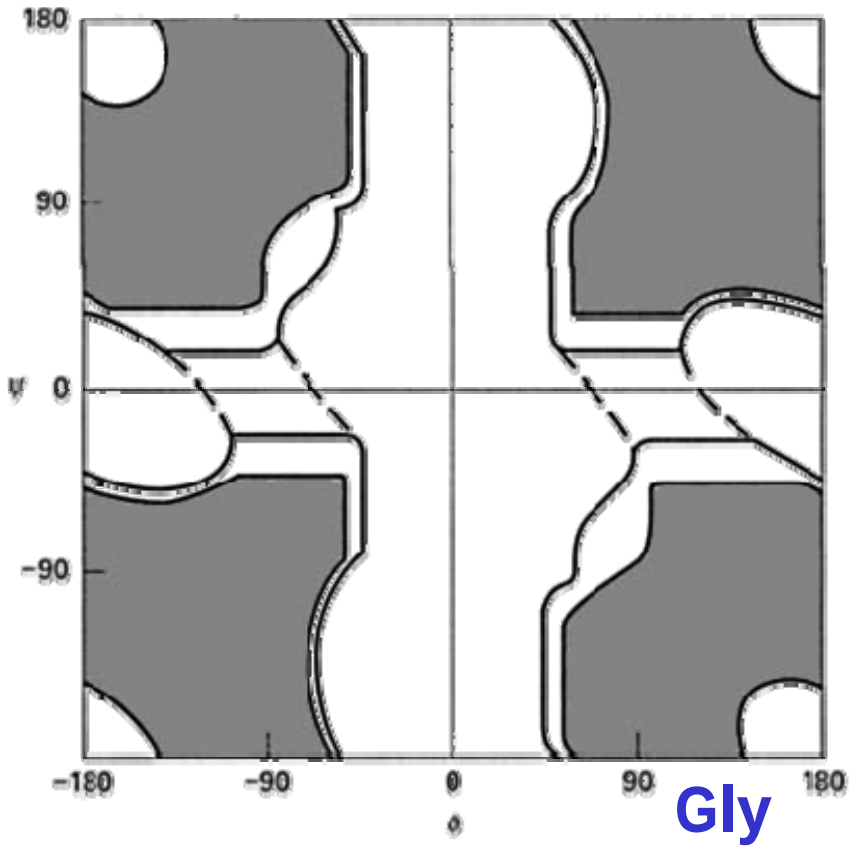
Torsionswinkel – Bevorzugte Bereiche

- β-Faltblätter
- α-Helices
  - Linksgängig (selten)
  - Rechtsgängig

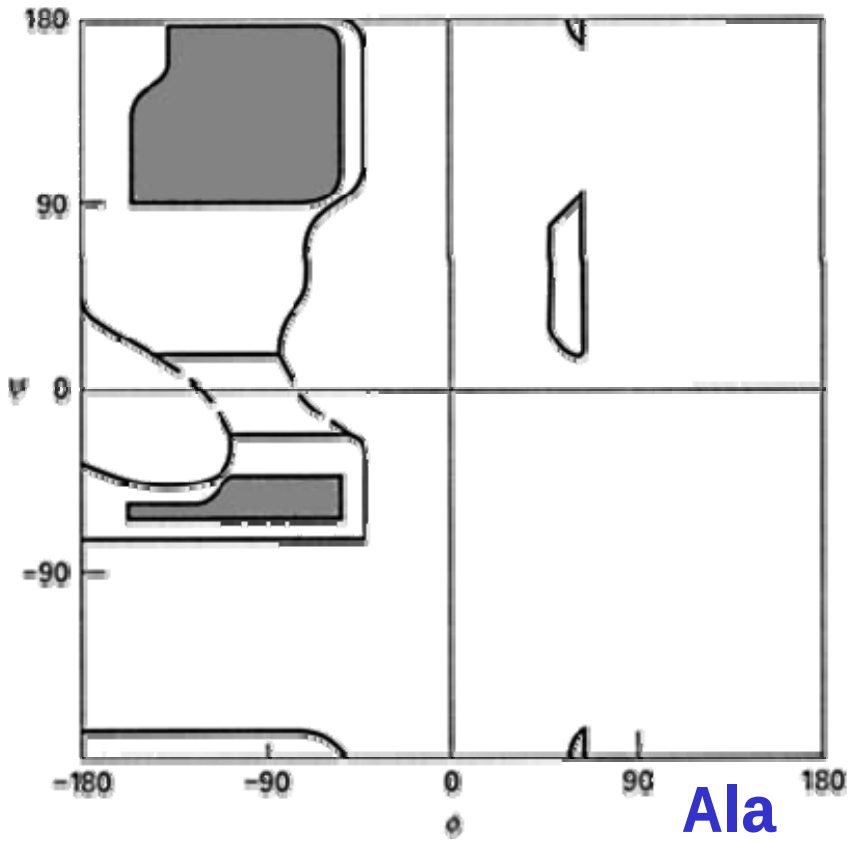


Torsionswinkel - Gly

- Gly hat kein C_β, daher große Freiheit in den Torsionswinkeln
- Ramachandran-Plot von Gly ist symmetrisch (Gly nicht chiral!)
- Bereits Ala mit seiner kleinen Seitenkette ist sterisch eingeschränkt



Gly



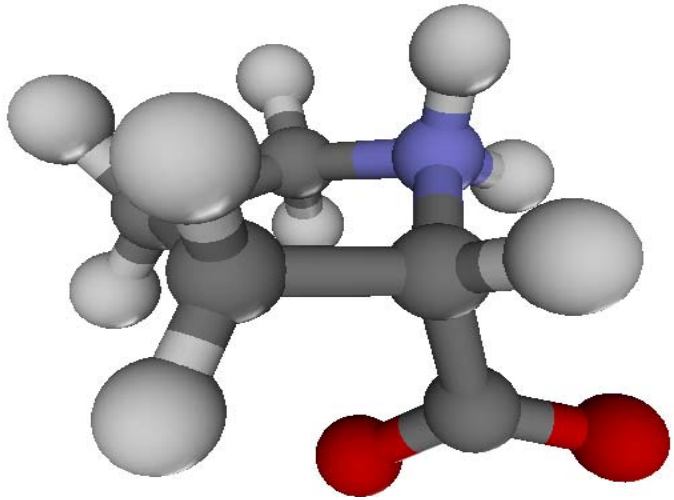
Ala

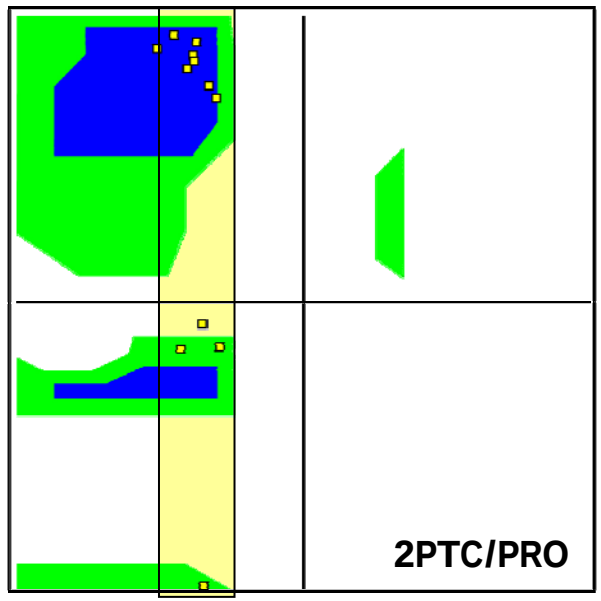
Torsionswinkel - Pro

- Pro ist **Iminosäure**

  $\Rightarrow$ keine H-Brücke mit NH des Rückgrats

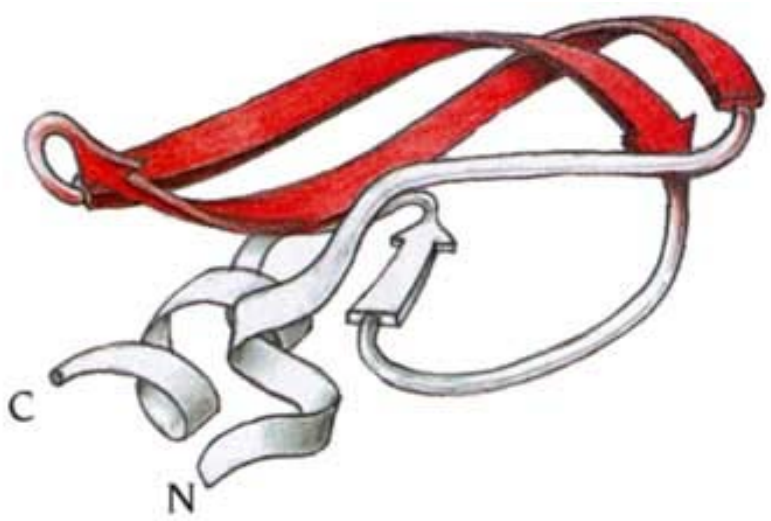
- ϕ ist durch Ring auf -60° festgelegt
- „**Helixbrecher**“: Geometrie nicht mit α-Helix kompatibel
- Pro passt aber sehr gut am N-terminalen Ende eine α-Helix



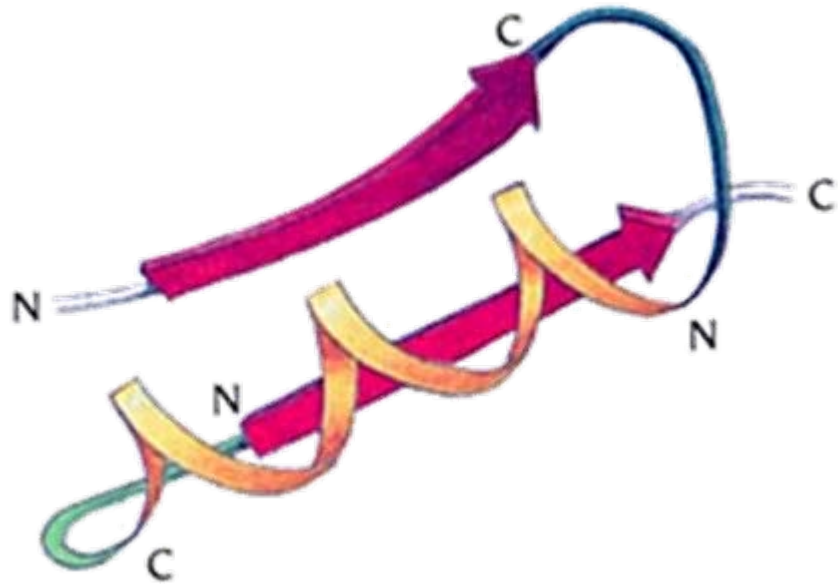


Supersekundärstrukturen

- Sekundärstrukturelemente formen häufig einfache Motive (**Supersekundärstrukturen**)
- Häufig wiederkehrende Motive sind z.B.
  - **Haarnadel-Motiv** (*hairpin*)
  - **β-α-β-Motiv**



Haarnadel



β-α-β

Tertiärstruktur – Visualisierung

- **BALLView**

  Ein Werkzeug zur Visualisierung und Modellierung von Proteinstrukturen.

  Download von www.ballview.org

  Binaries Windows, MacOSX, Source für Linux

- **VMD**

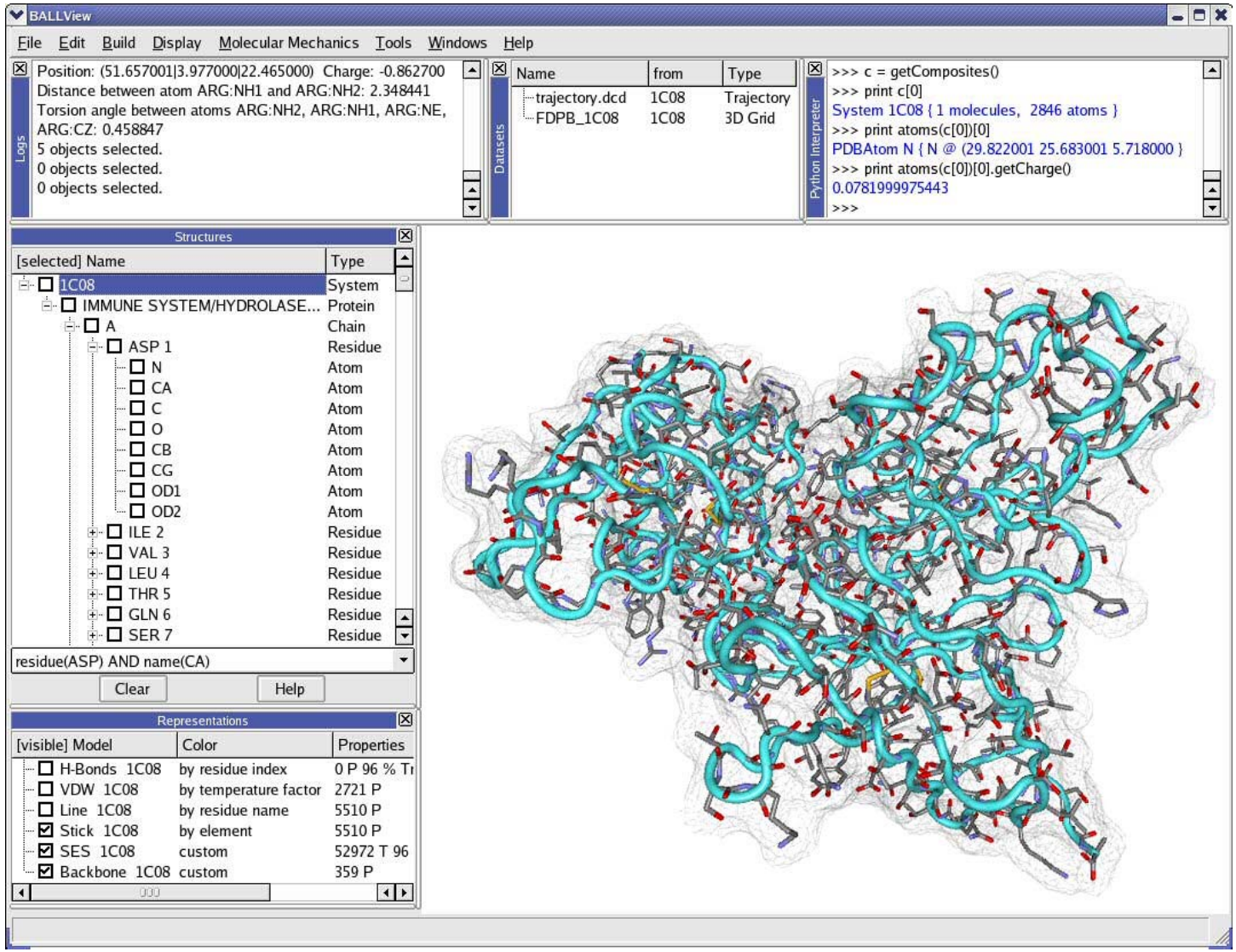
  Ein Werkzeug zur Proteinvisualisierung.

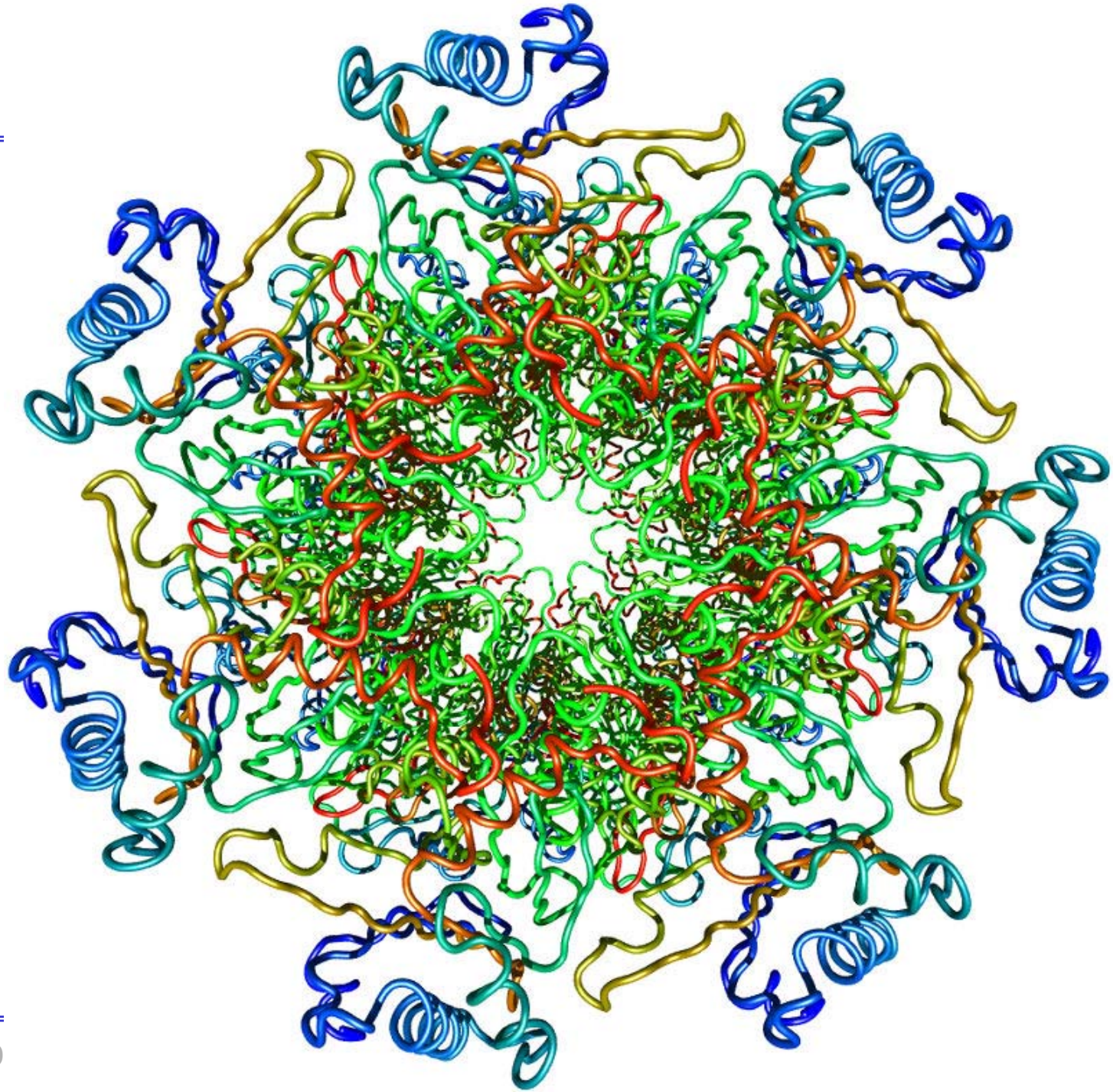
  Download von

  http://www.ks.uiuc.edu/Research/vmd/

  für verschiedene Plattformen.

BALLView

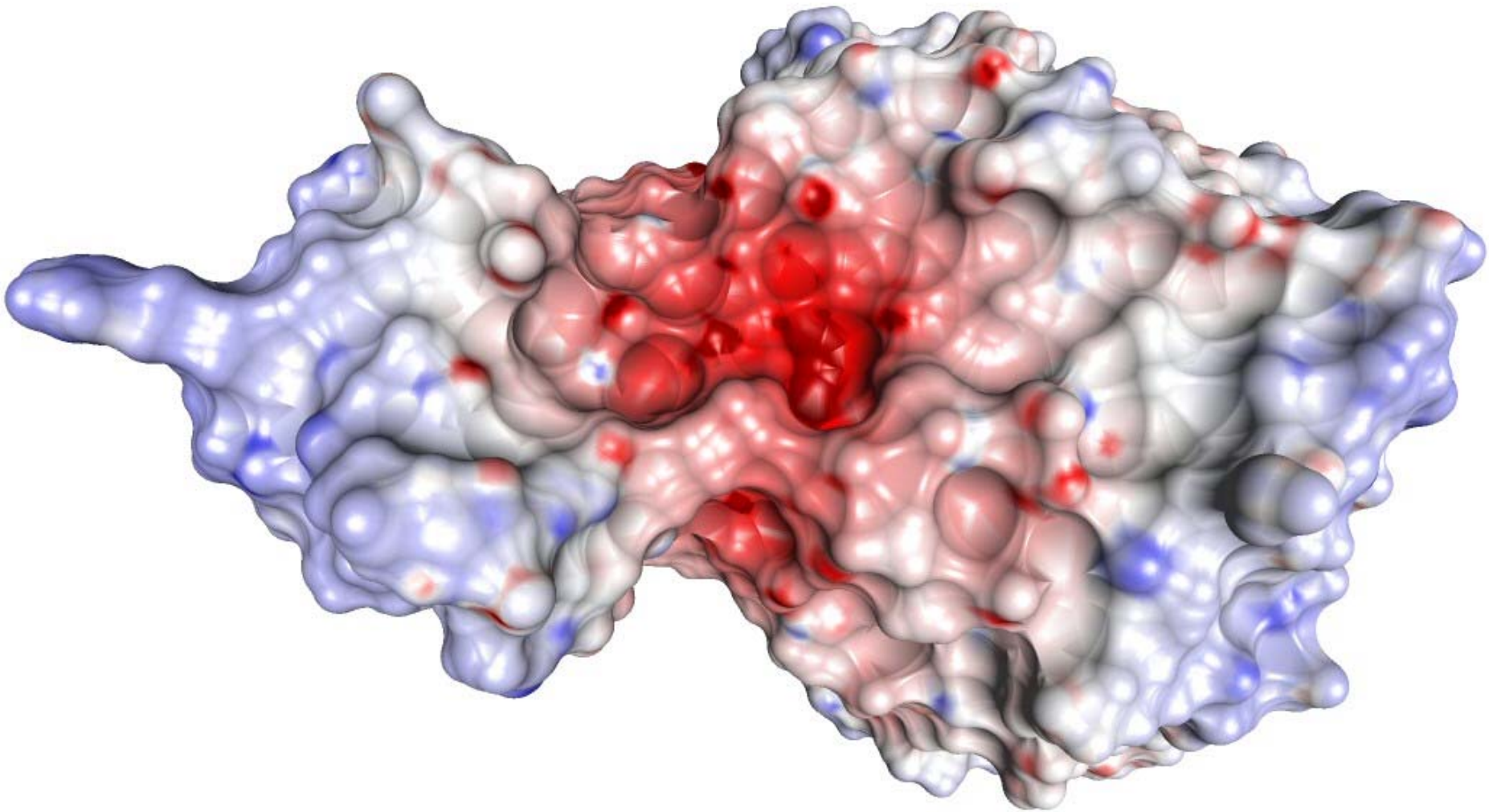

BALLView
File Edit Build Display Molecular Mechanics Tools Windows Help
Logs
Position: (51.657001|3.977000|22.465000) Charge: -0.862700
Distance between atom ARG:NH1 and ARG:NH2: 2.348441
Torsion angle between atoms ARG:NH2, ARG:NH1, ARG:NE, ARG:CZ: 0.458847
5 objects selected.
0 objects selected.
0 objects selected.
Datasets
Name from Type
trajectory.dcd 1C08 Trajectory
FDPB_1C08 1C08 3D Grid
Python Interpreter
>>> c = getComposites()
>>> print c[0]
System 1C08 { 1 molecules, 2846 atoms }
>>> print atoms(c[0])[0]
PDBAtom N { N @ (29.822001 25.683001 5.718000 }
>>> print atoms(c[0])[0].getCharge()
0.0781999975443
>>>
Structures
[selected] Name Type
1C08 System
IMMUNE SYSTEM/HYDROLASE... Protein
A Chain
ASP 1 Residue
N Atom
CA Atom
C Atom
O Atom
CB Atom
CG Atom
OD1 Atom
OD2 Atom
ILE 2 Residue
VAL 3 Residue
LEU 4 Residue
THR 5 Residue
GLN 6 Residue
SER 7 Residue
residue(ASP) AND name(CA)
Clear
Help
Representations
[visible] Model Color Properties
H-Bonds 1C08 by residue index 0 P 96 % Tr
VDW 1C08 by temperature factor 2721 P
Line 1C08 by residue name 5510 P
Stick 1C08 by element 5510 P
SES 1C08 custom 52972 T 96
Backbone 1C08 custom 359 P



Step

nstructure + Modelling

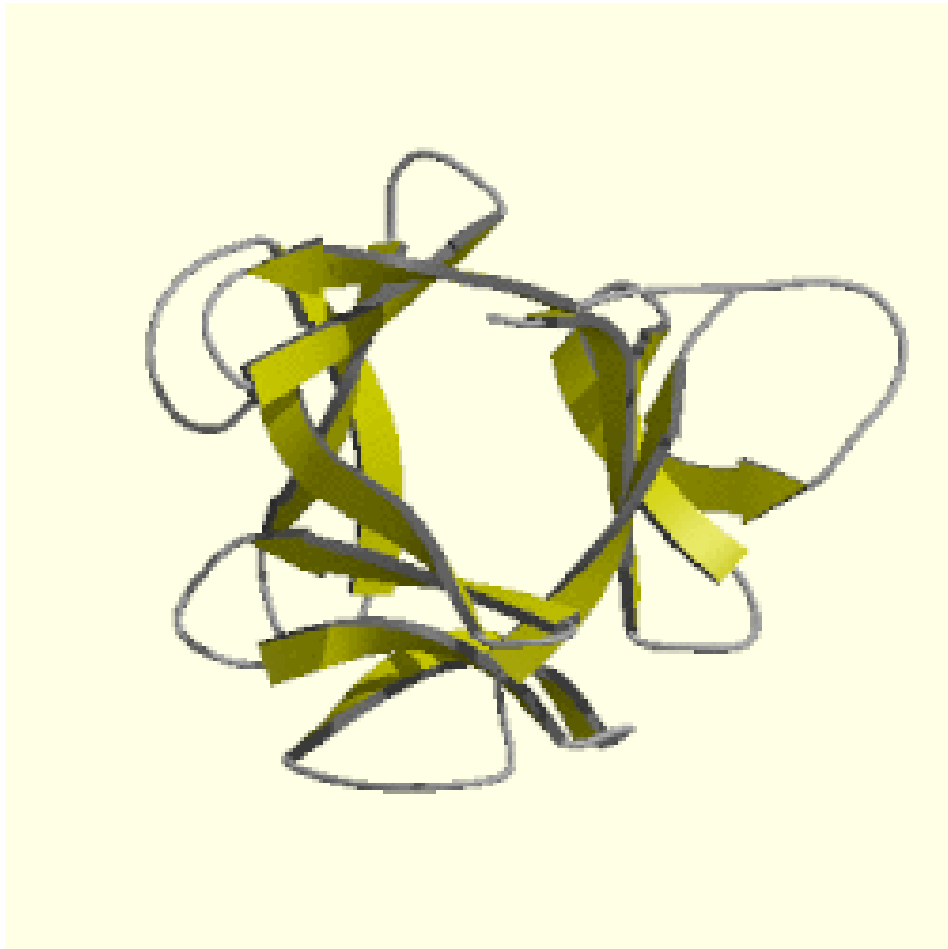
1PMA



Exkurs: Faltungsklassen

α: nur Helices

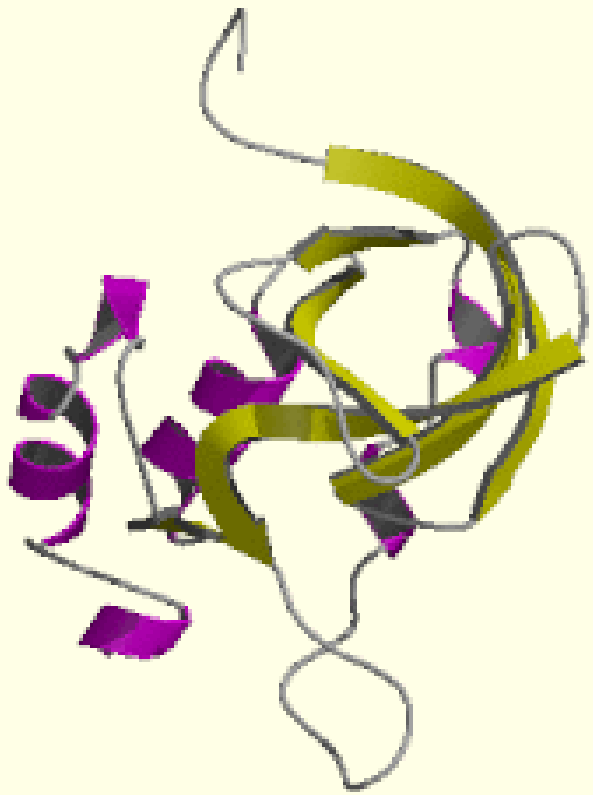
β: nur Faltblätter



Exkurs: Faltungsklassen

$\alpha+\beta$: Helices und Faltblätter in der Sequenz getrennt, Faltblätter meist durch *Turns* verbunden

Ubichinon-konjugierendes Enzym (1UBQ)



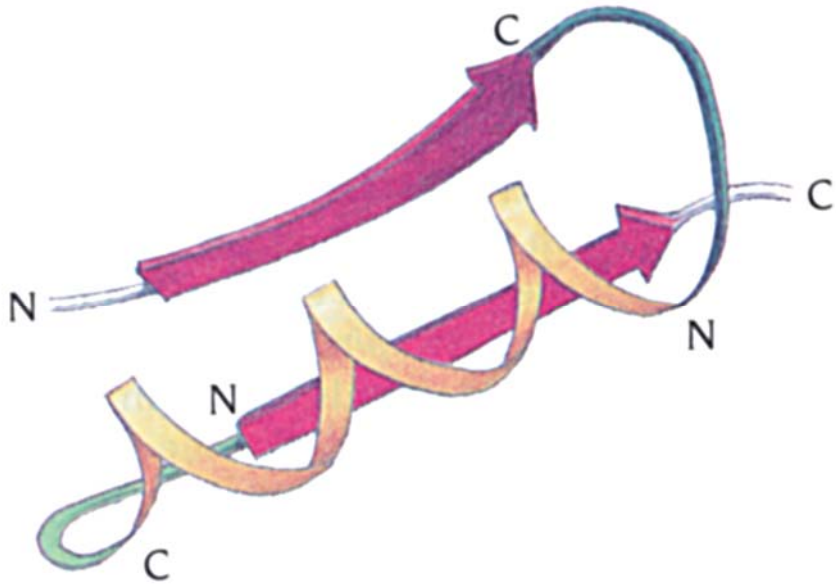
Staphylokokken-Nuklease (2SNS)

Exkurs: Faltungsklassen

α/β: Faltblatt mit verbindenden Helices (basierend auf dem β-α-β-Motiv)

TIM barrel

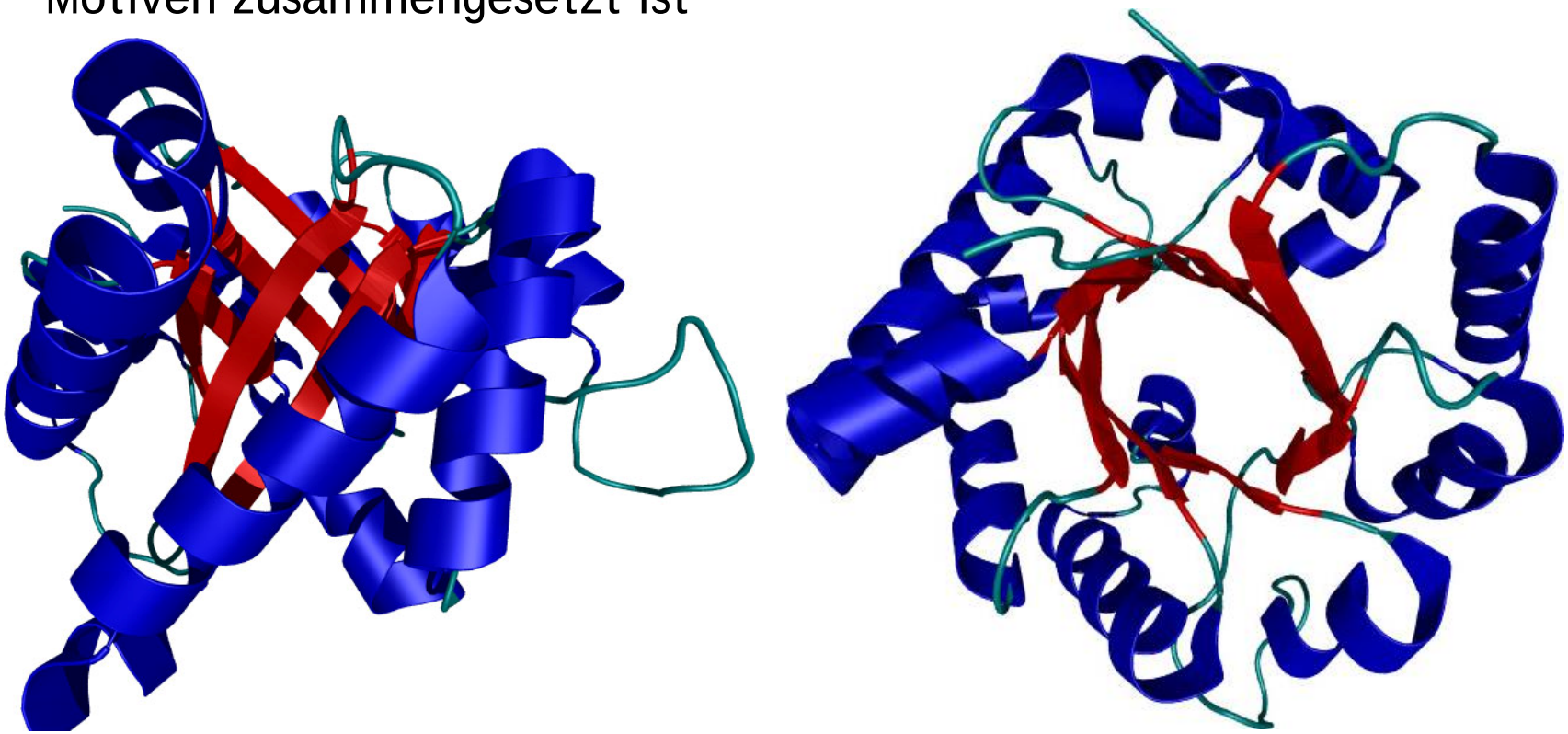
(TIM = **T**riosephosphat**i**so**m**erase)



β–α–β-Motiv

Exkurs: Faltungsklassen

- Es gibt eine ganze Hierarchie von typischen Faltungsmustern
- Eine sehr bekanntes Fold ist z.B. das TIM-Barrel (Triosephosphatisomerase)
- Eine Reihe von Proteinen nimmt dieses Fold an, das aus β-α-β-Motiven zusammengesetzt ist



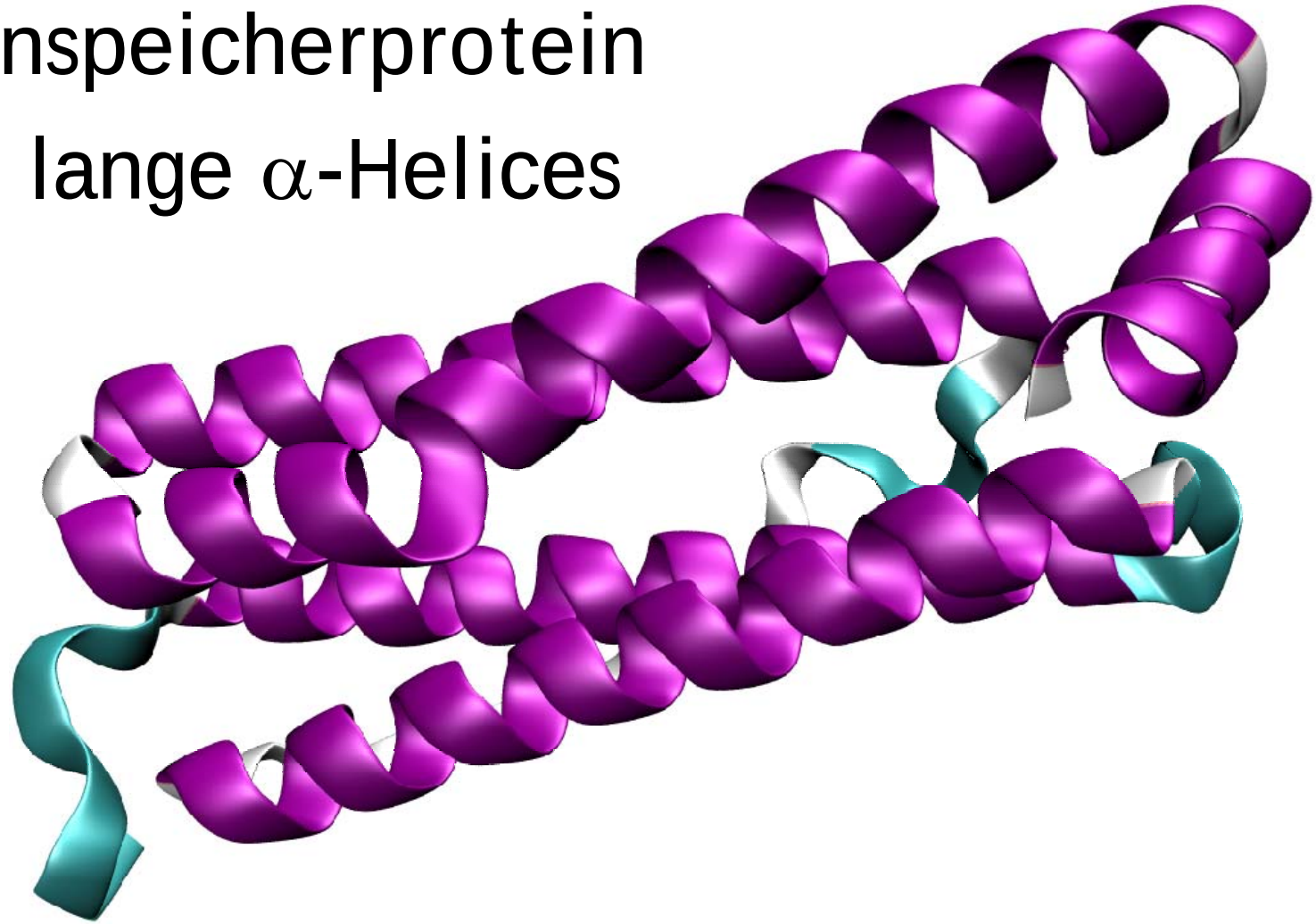
Datenbanken

- Die Datenbanken SCOP und CATH enthalten eine hierarchische Klassifikation von Proteindomänen nach ihren Strukturen
- ~ 27.000 Strukturen in der PDB
- ~ 600 Faltungsklassen
- Anzahl Faltungsklassen in SCOP (03/2001)

138	α
93	β
97	α/β
184	$\alpha + \beta$
23	Multidomänen-Proteine
11	Membranproteine/Zelloberflächenproteine
54	Kleine Proteine
$\sum$ 605	

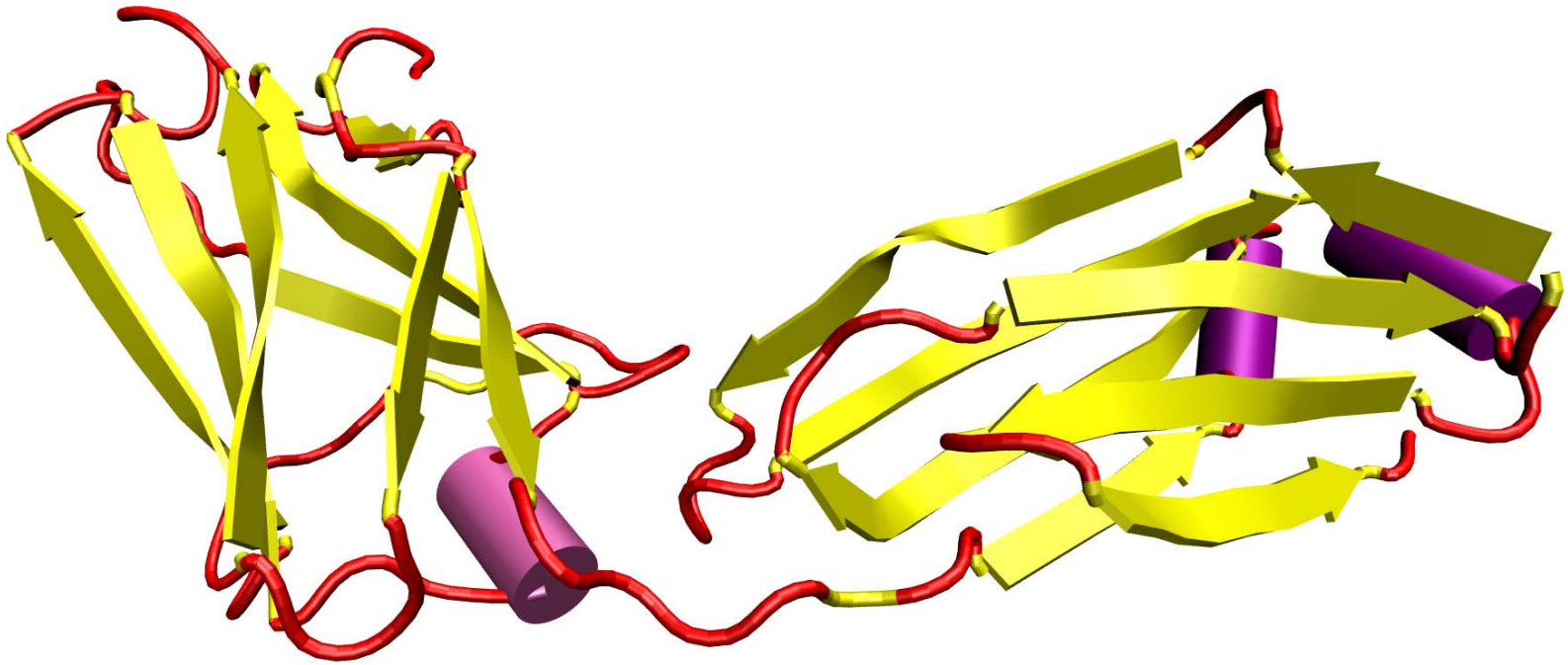
Ferritin – ein α-helikales Protein

- Eisenspeicherprotein
- Vier lange α-Helices



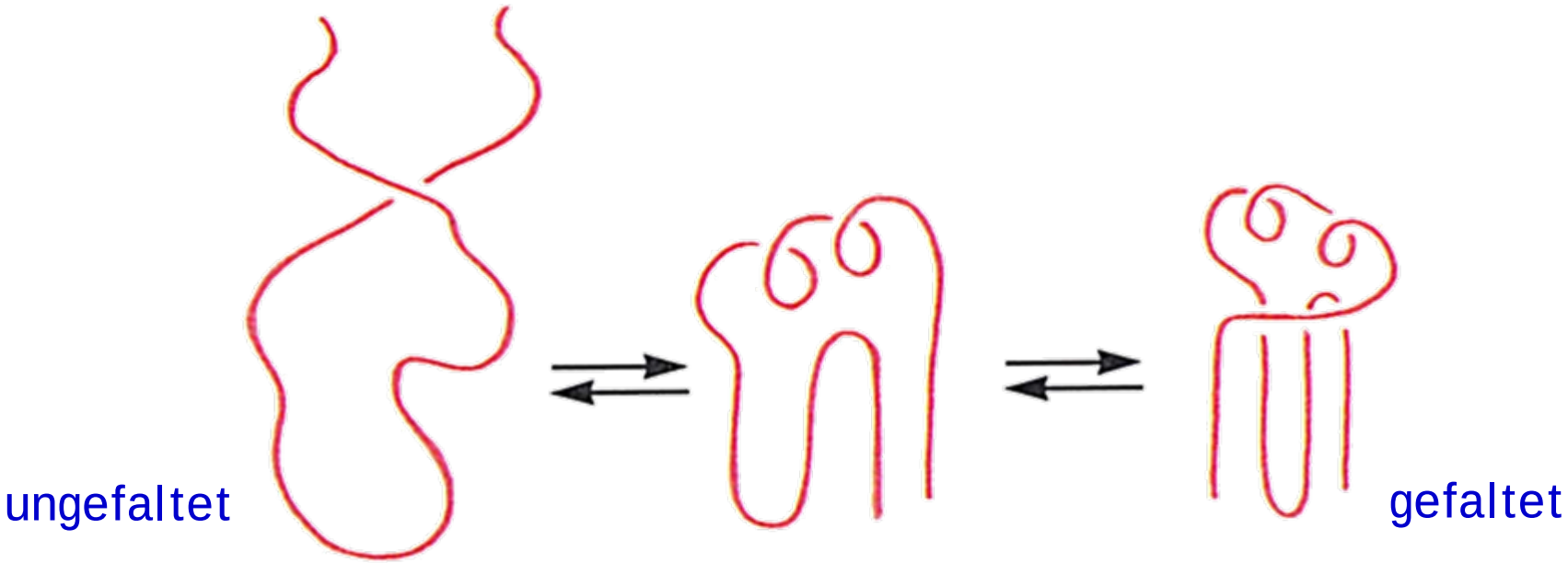
L-Kette eines Antikörpers

- Antikörper bestehen aus einer leichten und einer schweren Kette
- Leichte Kette besteht fast ausschließlich aus β-Faltblättern



Faltung

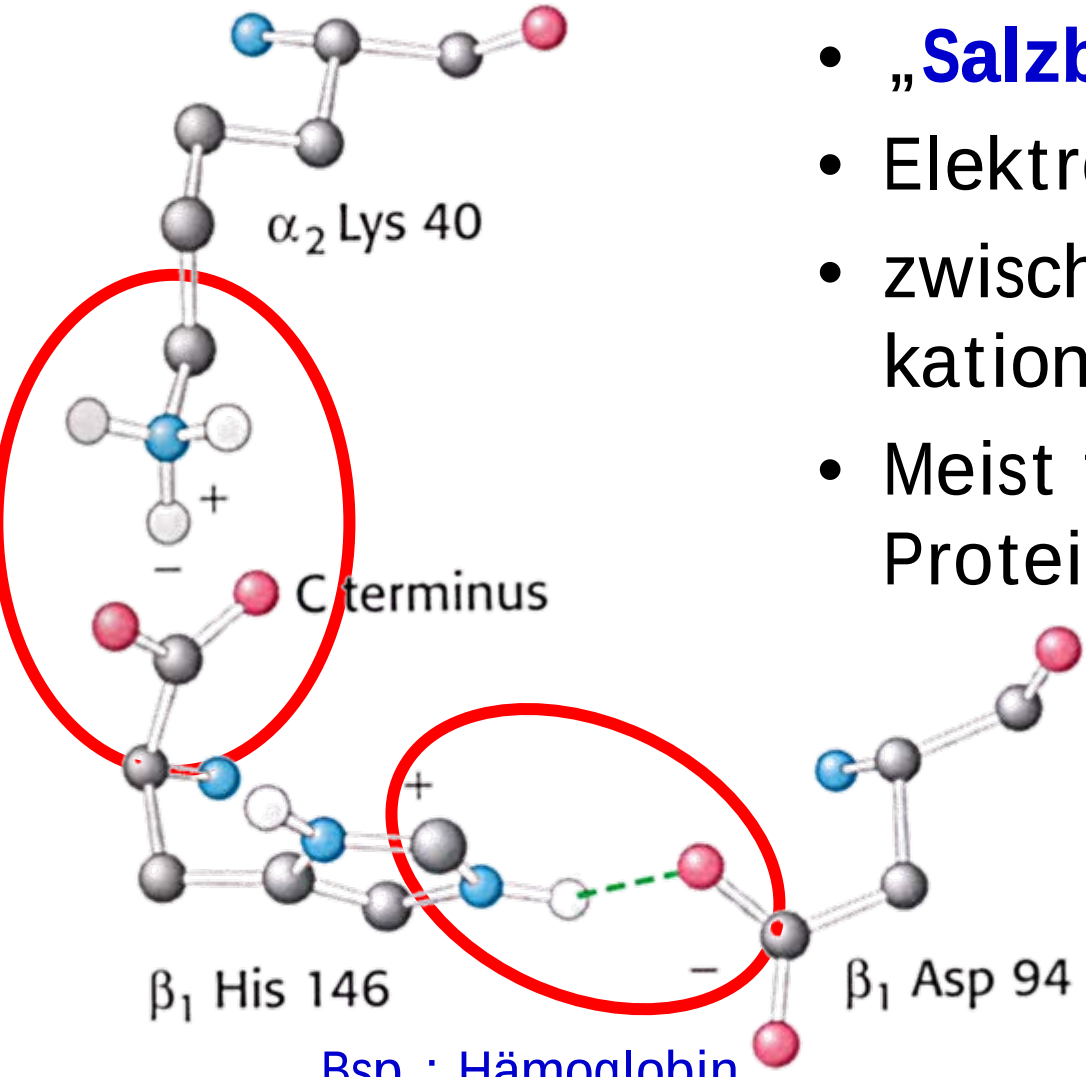
- Die am Ribosom synthetisierte Polypeptidkette ist zunächst noch ungeordnet
- Es bilden sich in zunächst lokal Sekundärstrukturelemente aus
- Kette faltet sich im Raum und bildet die **Tertiärstruktur** aus
- Der umgekehrte Prozeß zur **Faltung** ist die **Entfaltung** oder **Denaturierung**



Tertiärstruktur

- Tertiärstruktur beschreibt die räumliche Anordnung der Sekundärstrukturelemente
- Ca. 50% der Proteinstruktur werden von repetitiven Elementen gebildet (Helix, Strang)
- Sekundärstrukturelemente werden durch eine Reihe von Wechselwirkungen stabilisiert
- gefaltete Struktur stabiler als entfaltete.
- Wichtigste stabilisierende Faktoren sind
  - Hydrophobe Wechselwirkung
  - Ionische Wechselwirkung
  - Schwefelbrücken
  - Wasserstoffbrücken

Tertiärstruktur – Ionische WW



- „**Salzbrücken**“
- Elektrostatische WW
- zwischen anionischen und kationischen Resten
- Meist tief im hydrophoben Proteinkern (*buried*)

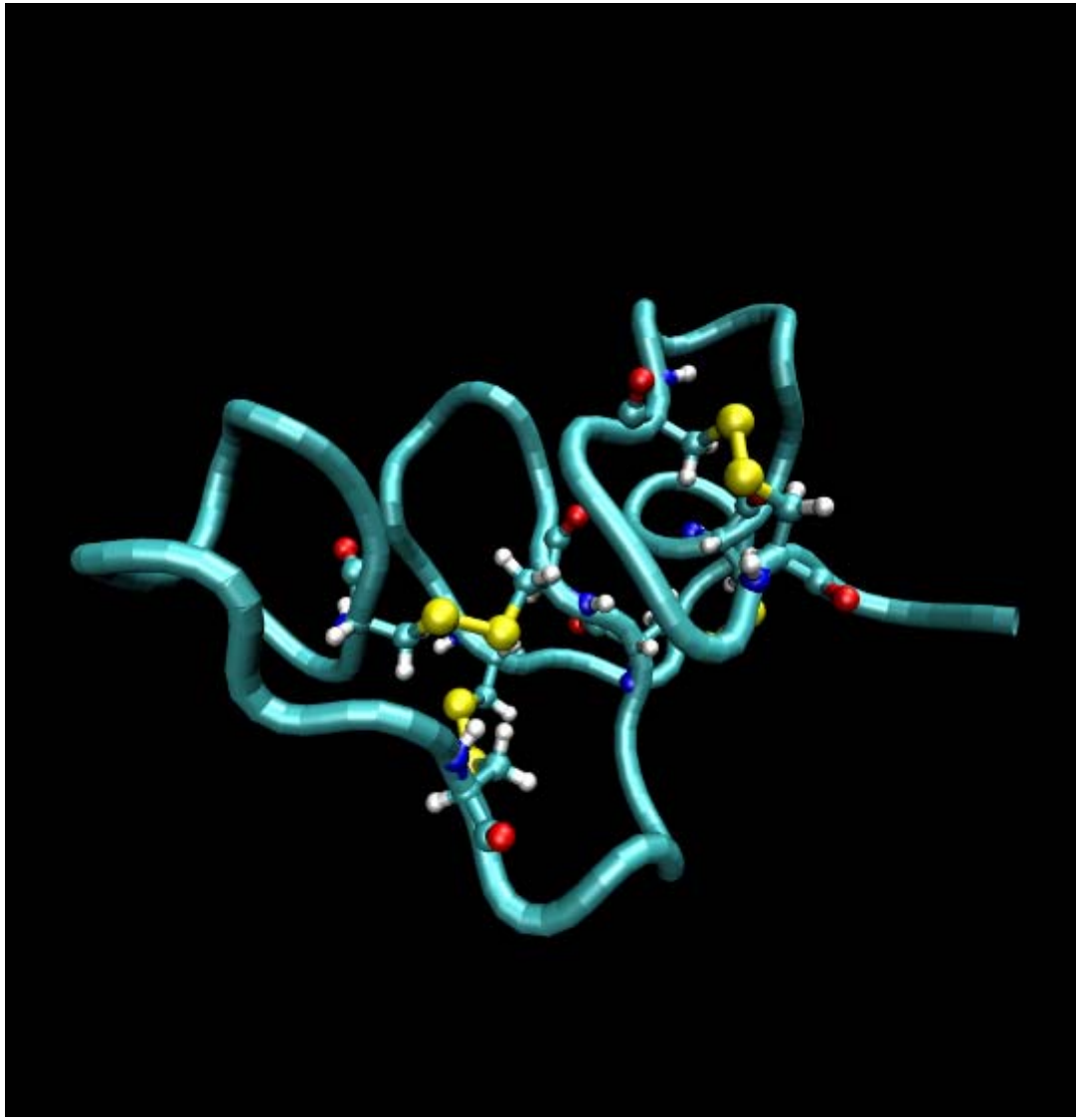
Bsp.: Hämoglobin

Tertiärstruktur – Hydrophobe WW

- Hydrophobe Seitenketten versuchen Wasser zu vermeiden
- Analogie: Öl in Wasser
- Bildung eines **hydrophoben Kerns** im Protein (*hydrophobic core*)
- **Entropischer Effekt** – Eigentlich keine Wechselwirkung innerhalb des Proteins, sondern des Proteins mit dem umgebenden Wasser

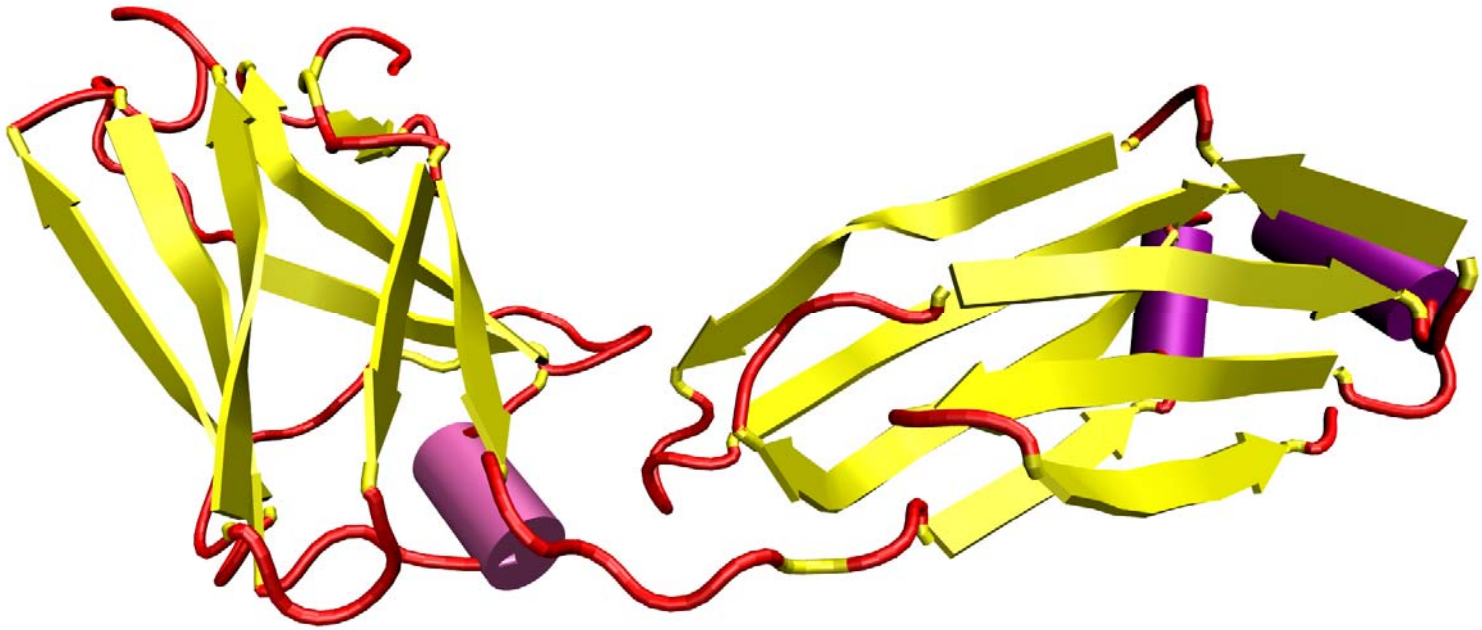
Tertiärstruktur – Schwefelbrücken

- Sekundärstrukturelemente oder Ketten können kovalent durch **Disulfidbrücken** (Schwefelbrücken) miteinander verbunden
- Schwefelbrücken können reduktiv geöffnet werden und oxidativ wieder geschlossen werden (Dauerwelle!)
- S-Brücken ungleich stärker als hydrophobe WW und H-Brücken

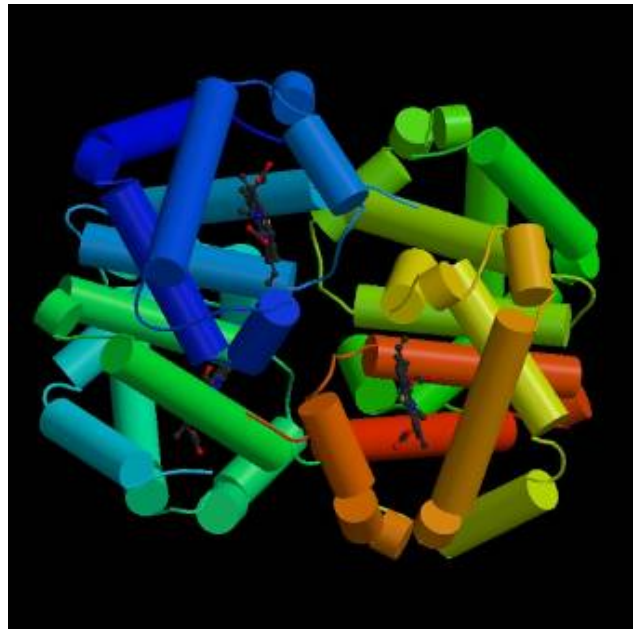


Domänen

- Häufig falten sich Proteine in mehrere kompakte **Untereinheiten** (Domänen) die durch **flexible Polypeptidketten** verbunden sind
- **Domänen**
  - haben häufig spezifische Aufgaben innerhalb des Proteins (z.B. katalytische Aktivität)
  - sind meist 100 – 400 AS lang
  - werden oft von einzelnen Exons kodiert



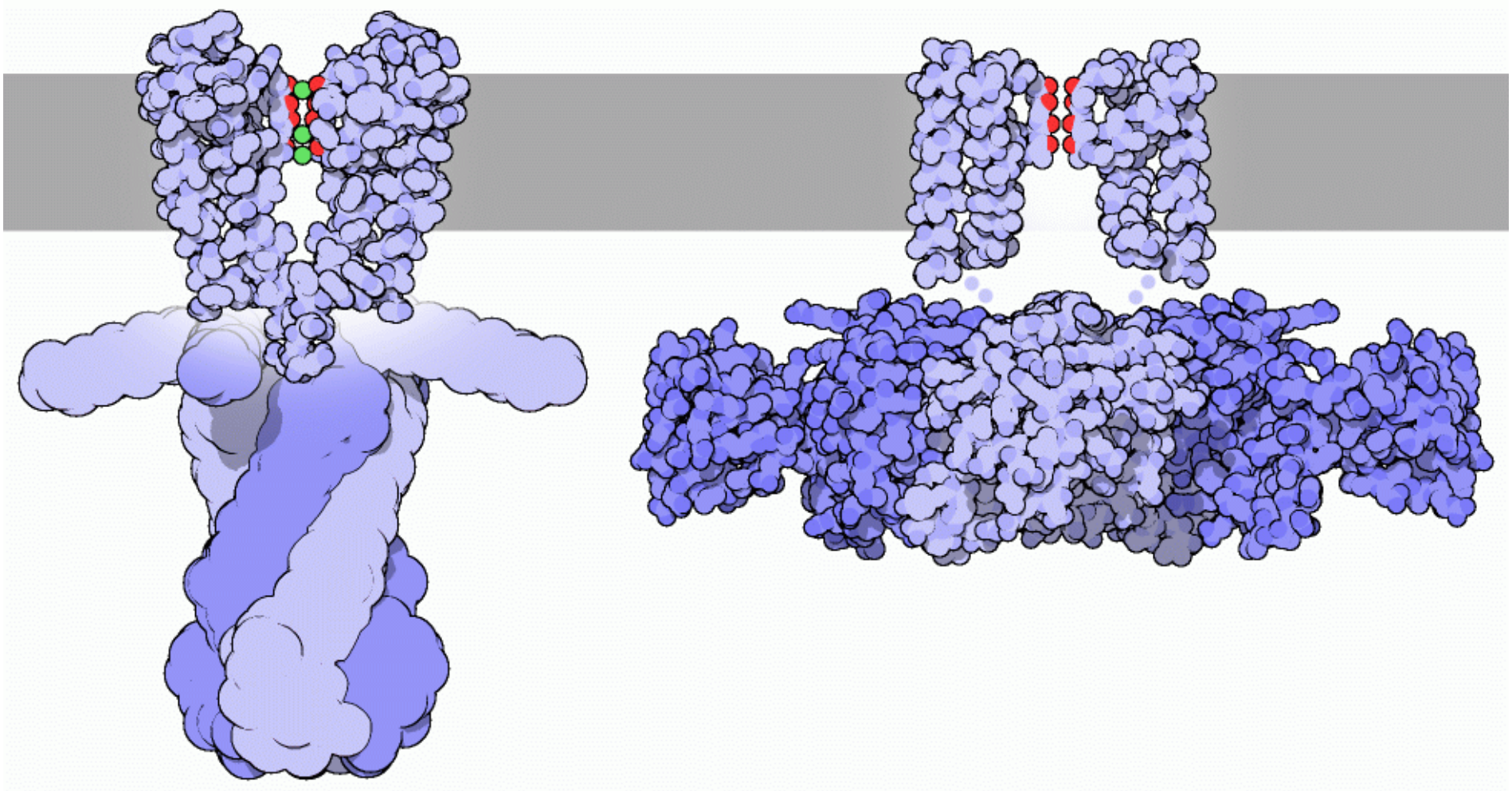
Quartärstruktur



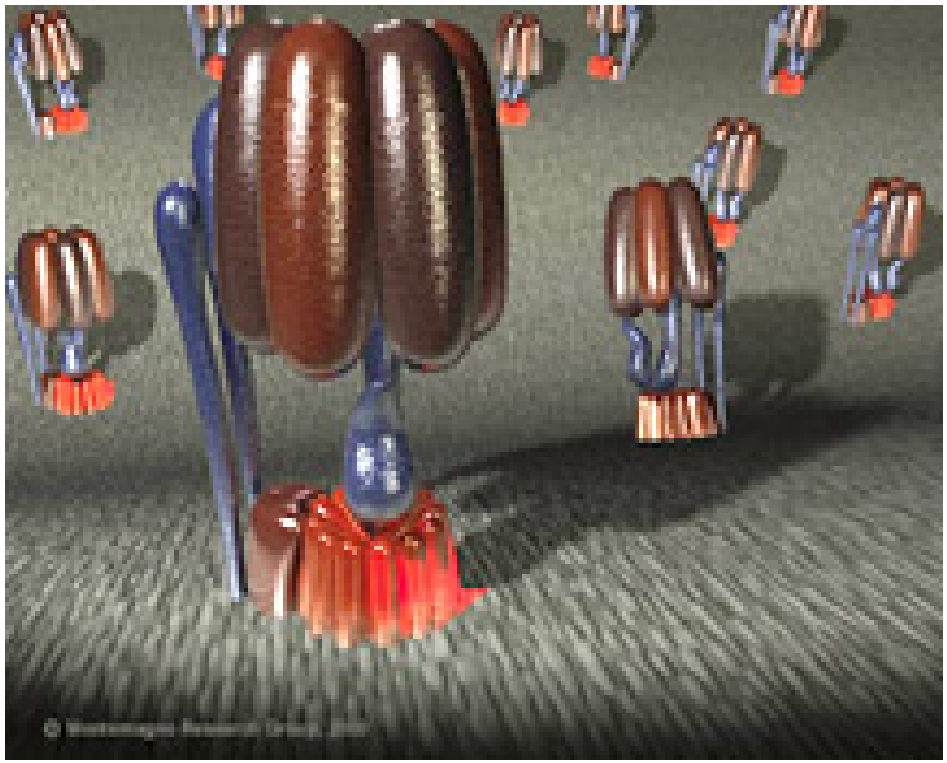
- Proteine liegen häufig als Aggregate mehrerer Untereinheiten (Ketten) vor
- **Dimer** = zwei Untereinheiten
- **Trimer** = drei Untereinheiten, …
- **Homodimer**: zwei identische Untereinheiten (AA)
- **Heterodimer**: unterschiedliche Untereinheiten (AB)
- Funktion oft an Quartärstruktur gebunden

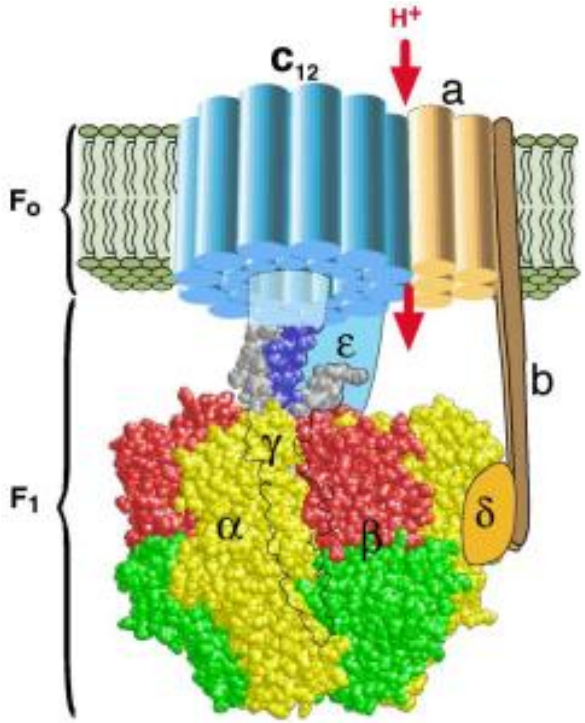
  (Beispiel: Bindungsstelle an der Grenze zweier Domänen)
- Bsp.: **Hämoglobin**
  - Sauerstofftransporter im Blut
  - α-β-α-β-Tetramer
  - Alternativ: Homodimer aus zwei α-β-Untereinheiten

Komplexe - Ionenknäle



Quartärstruktur – Komplexe

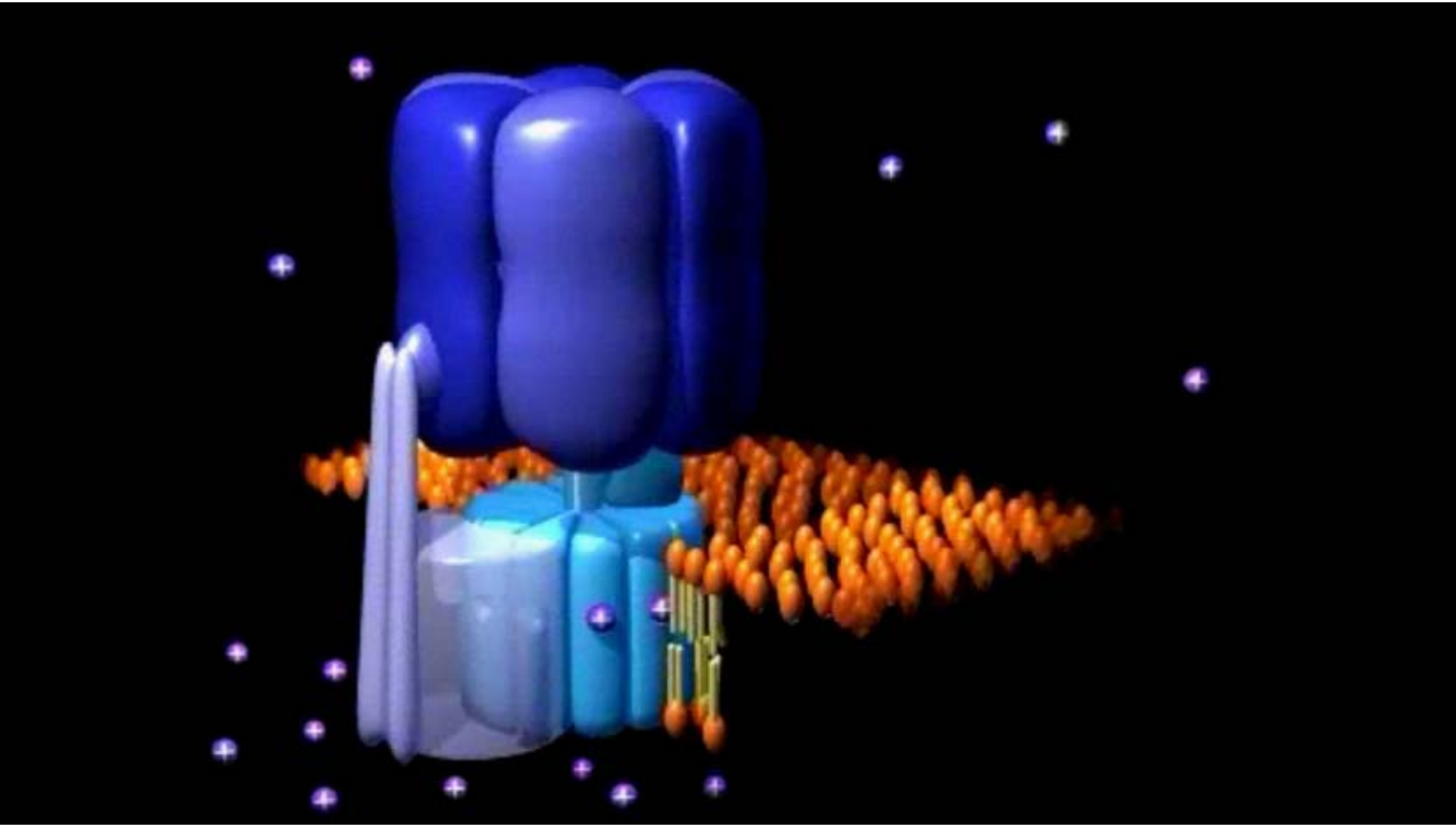




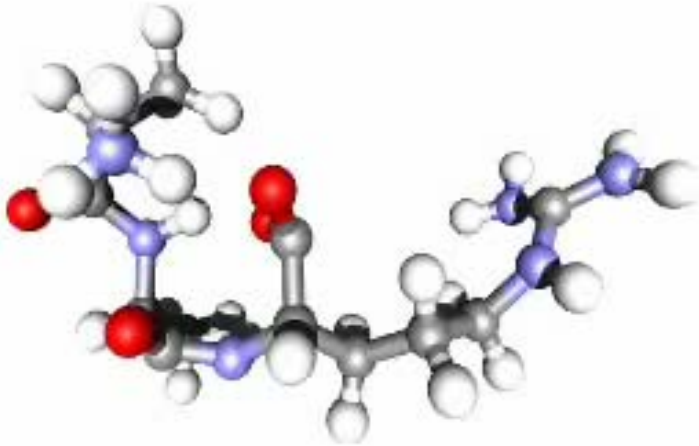
H. Wang and G. Oster (1998). Nature 396:279-282.

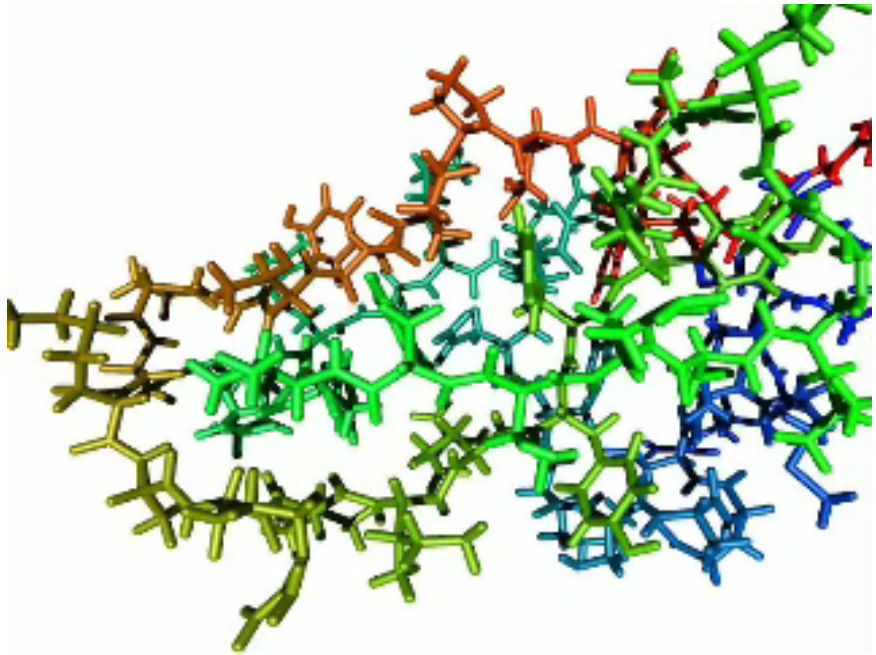
- F_0-F_1-ATPase sitzt in der Zellmembran
- Erzeugt ATP (aus ADP und P_i) unter Ausnutzung des H^+-Gradienten entlang der Membran
- Dabei erzeugt der F_0-Motor aus dem H^+-Gradienten eine Rotation, die der F_1-Teil zur Synthese von ATP aus ADP einsetzt

Quartärstruktur – Komplexe



Dynamik von Proteinen





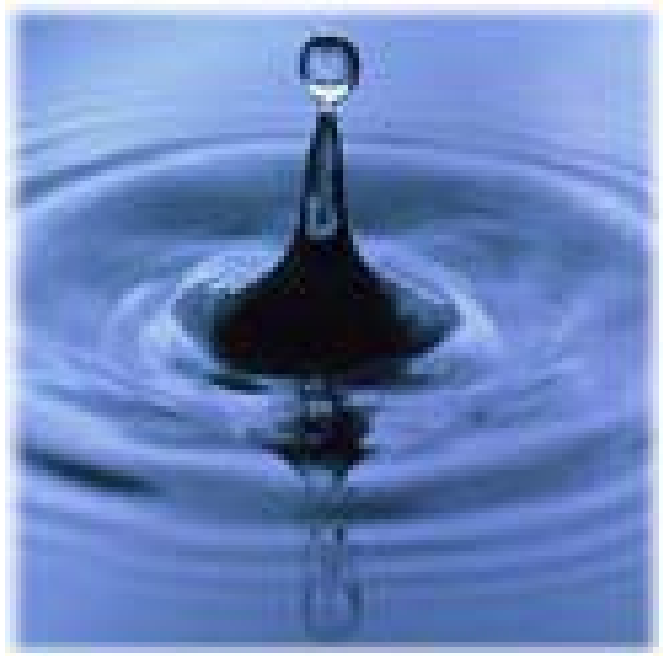
- Überwiegend Rotationen um Torsionswinkel
- Auch komplexere Bewegungen möglich, z.B. Bewegungen ganzer Domänen um flexible „Scharnier“-Bereiche
- Rückgrat meist mehr oder minder starr
- Seitenketten an der Oberfläche sind sehr flexibel

Verteilung der AA

- **Oberfläche**
  - Überwiegend polare, geladene AS
  - Ermöglicht WW mit Wasser
- **Kern**
  - Überwiegend unpolare, hydrophobe AS
  - WW zwischen unpolaren Seitenketten günstiger als zwischen unpolarer SK und Wasser
  - Ausnahmen: Salzbrücken

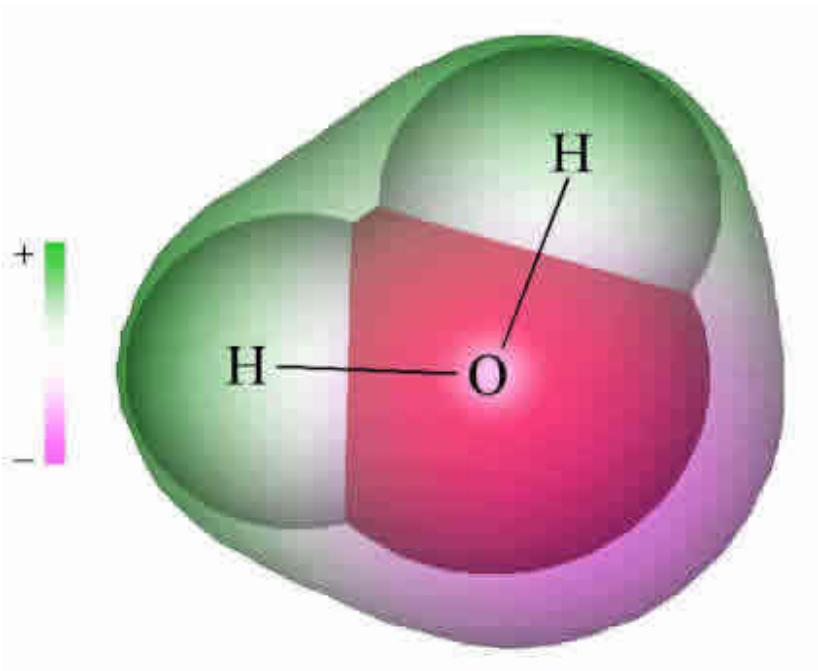
Wasser

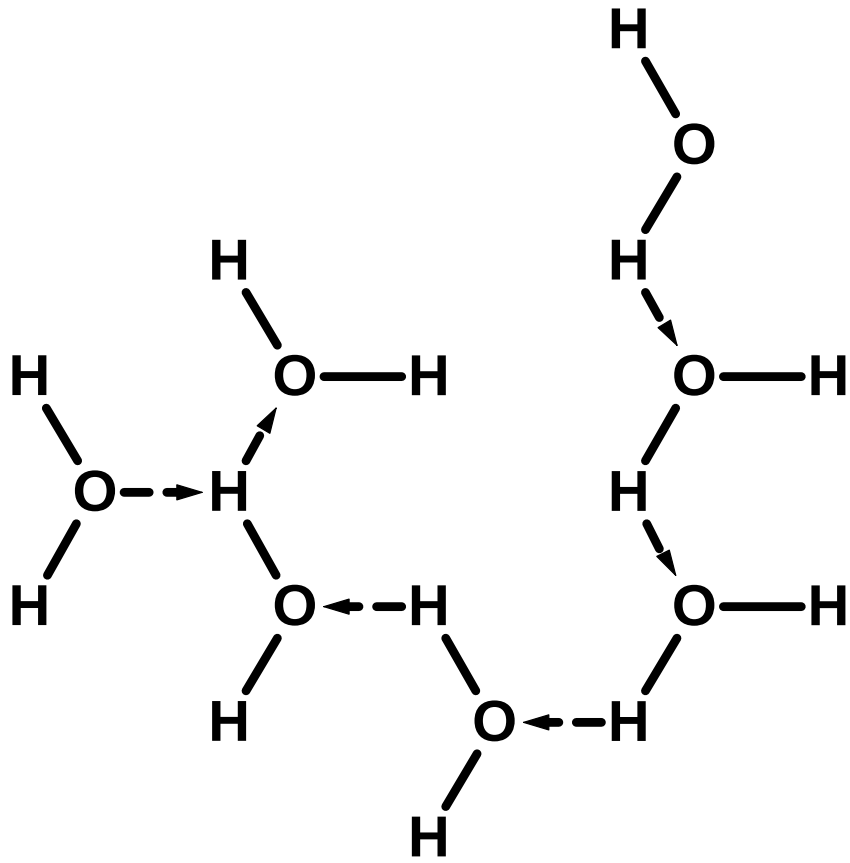
- Biochemie findet in Wasser statt
- Wasser hat **ungewöhnliche Eigenschaften**
  - Wasserstoffbrücken
  - Polarität
  - Hohe Dielektrizitätskonstante



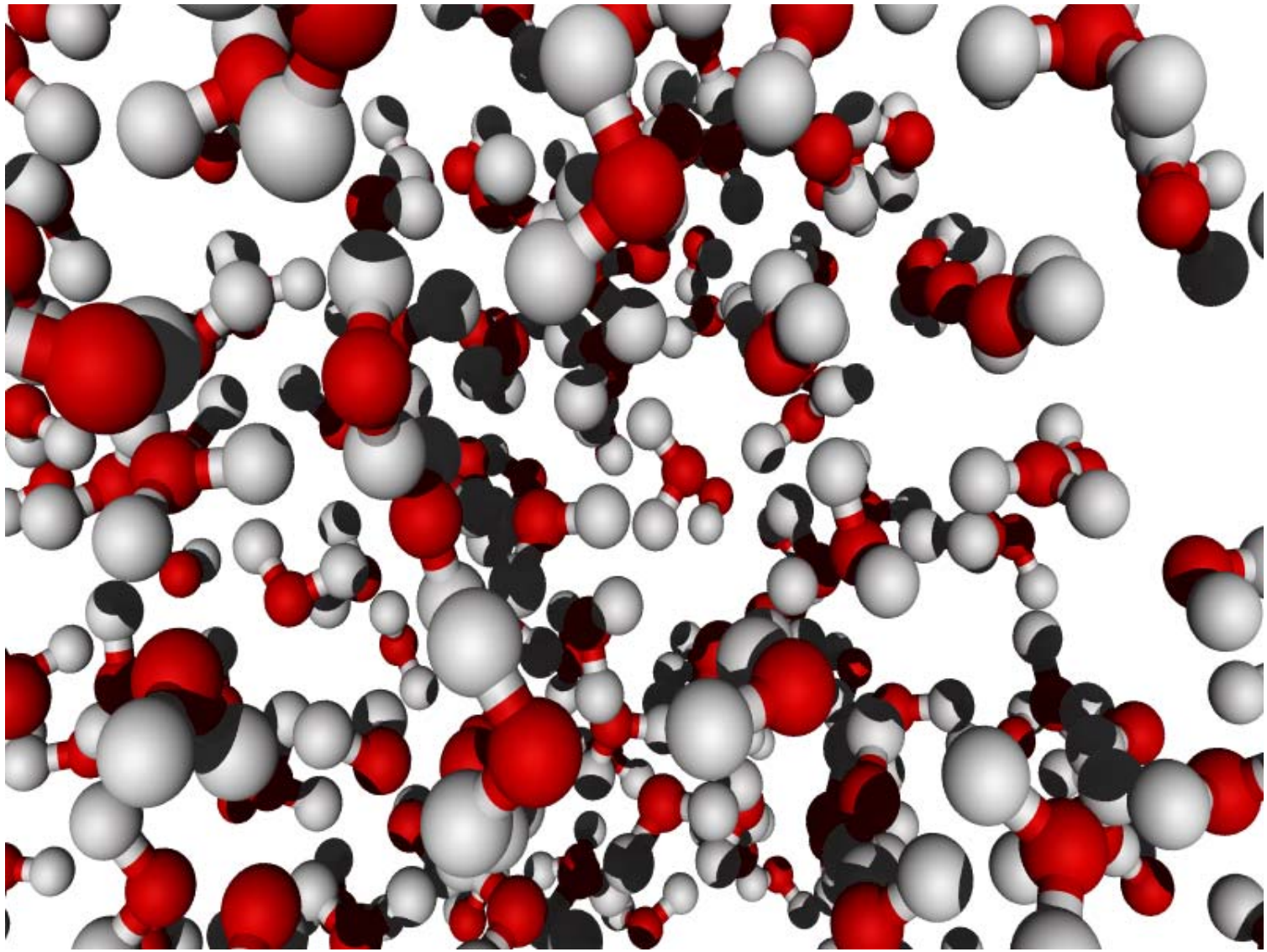
- Ideal für die Biochemie
- Fürchterlich für die Bioinformatik!

Wasser – Aus der Nähe



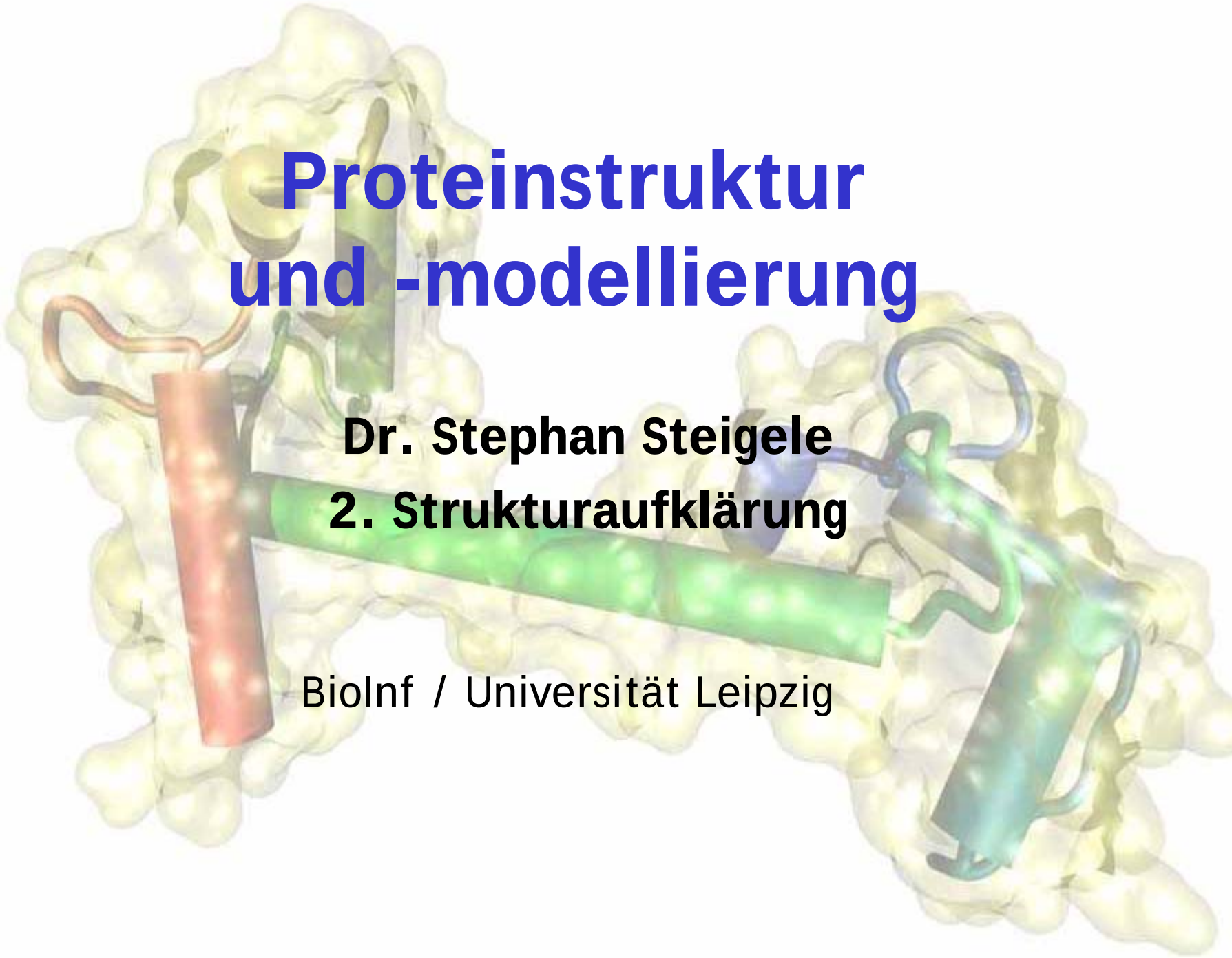


Wasser – Aus der Nähe…



Probleme mit Wasser

- Durch Ausbildung des H-Brücken-Netzwerks sind Lösemitteleffekte nicht lokal
- Modelle müssen
  - die **Struktur von Wasser** berücksichtigen
  - die **Wechselwirkungen im Wasser** modellieren
- Modelle zur Beschreibung von Wasser sind daher häufig sehr aufwändig
- Modellierung von Solvatation, Elektrostatik, Entropie ist generell schwierig!



Proteinstruktur und -modellierung

Dr. Stephan Steigele

2. Strukturaufklärung

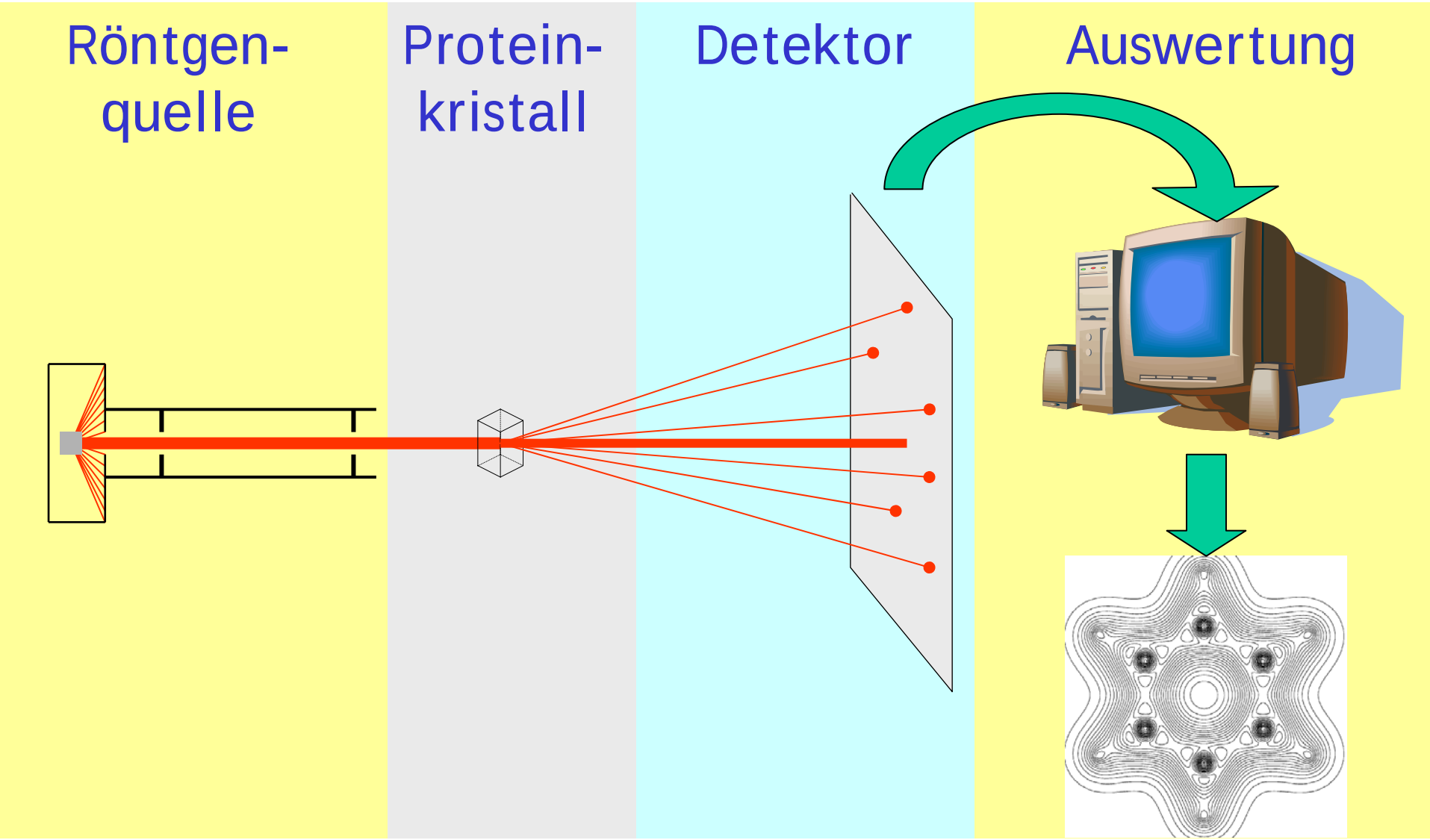
BioInf / Universität Leipzig

Übersicht

Experimentelle Methoden

- **Röntgenkristallografie (XRD)**
- **Neutronenstreuung**
- **Elektronenmikroskopie (TEM)**
- **Kernmagnetische Resonanz (NMR)**

Röntgen-Kristallografie

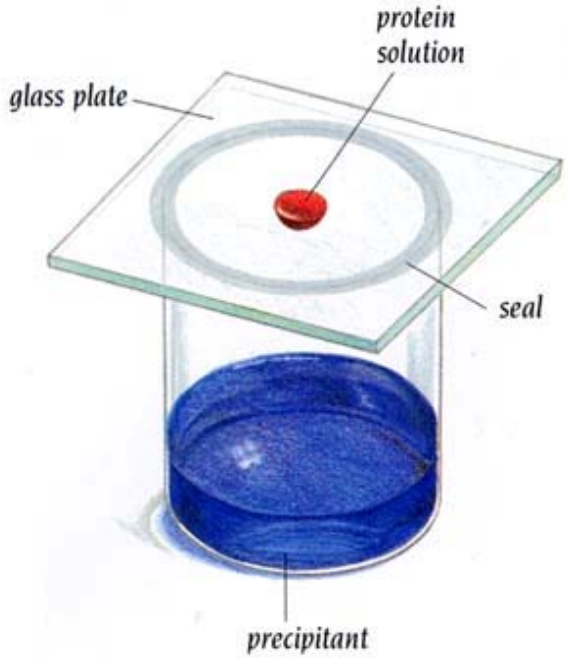


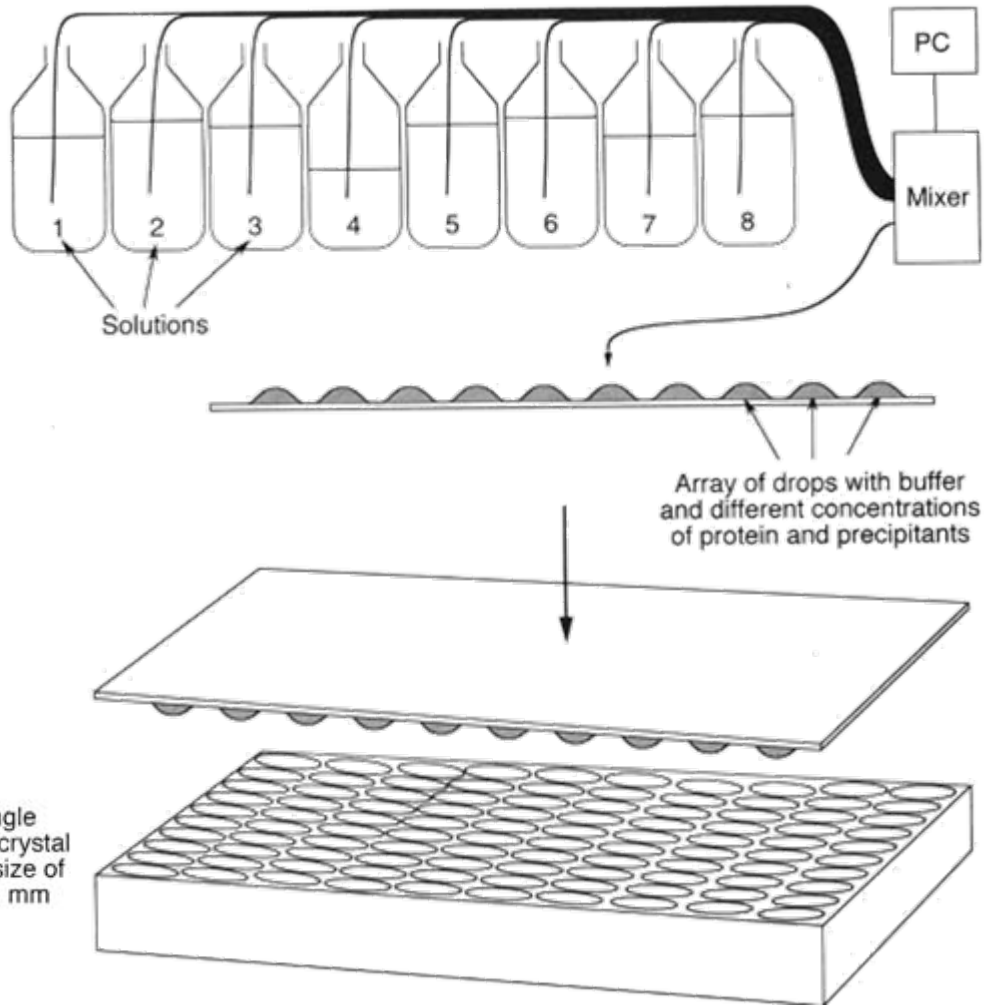
Protein-Kristalle

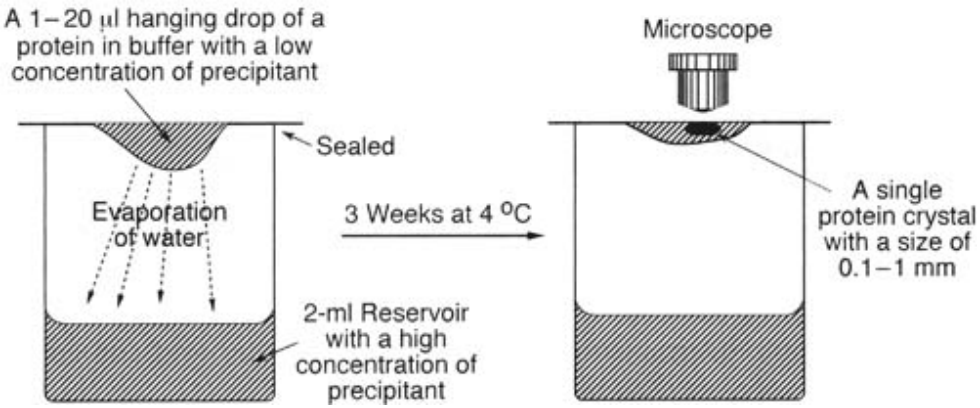
Proteine sind **schwierig zu kristallisieren**

- Unregelmäßige Struktur
  ⇒ Große „Löcher“ im Kristall
- Große Kristalle erforderlich
  (0.1 – 0.5 mm)
- Ausreichend Protein erforderlich
- Große Reinheit erforderlich
- Wachstum sehr langsam (teilweise mehrere Monate)
- Teilweise nur Einzeldomänen kristallisierbar

Kristallisation – „Hängender Tropfen“





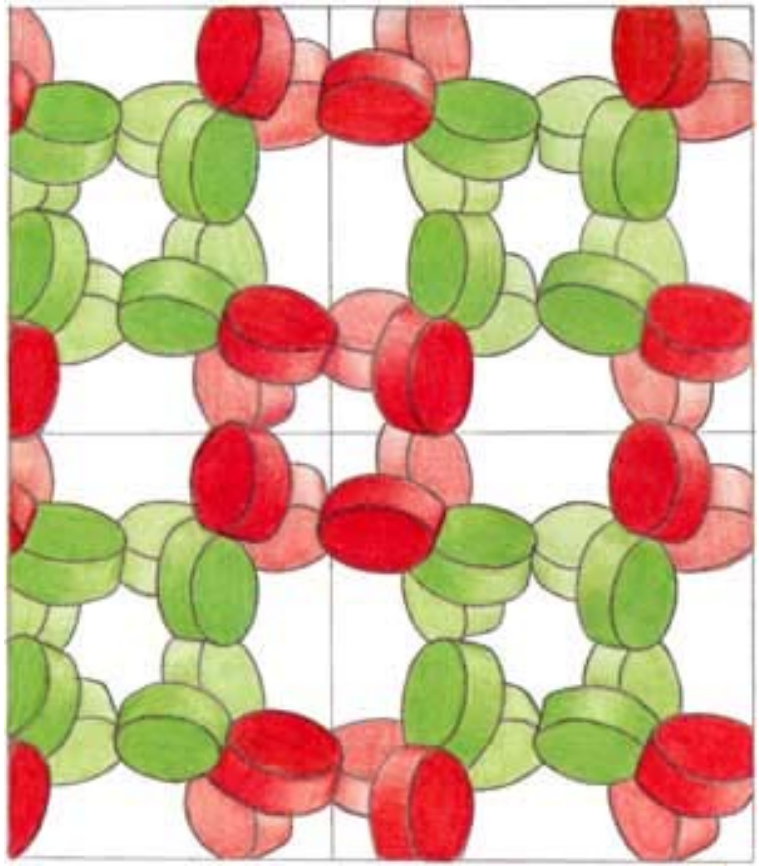


Nölting, S. 70,
Branden/Tooze, S. 376
Proteinstructure + Modelling

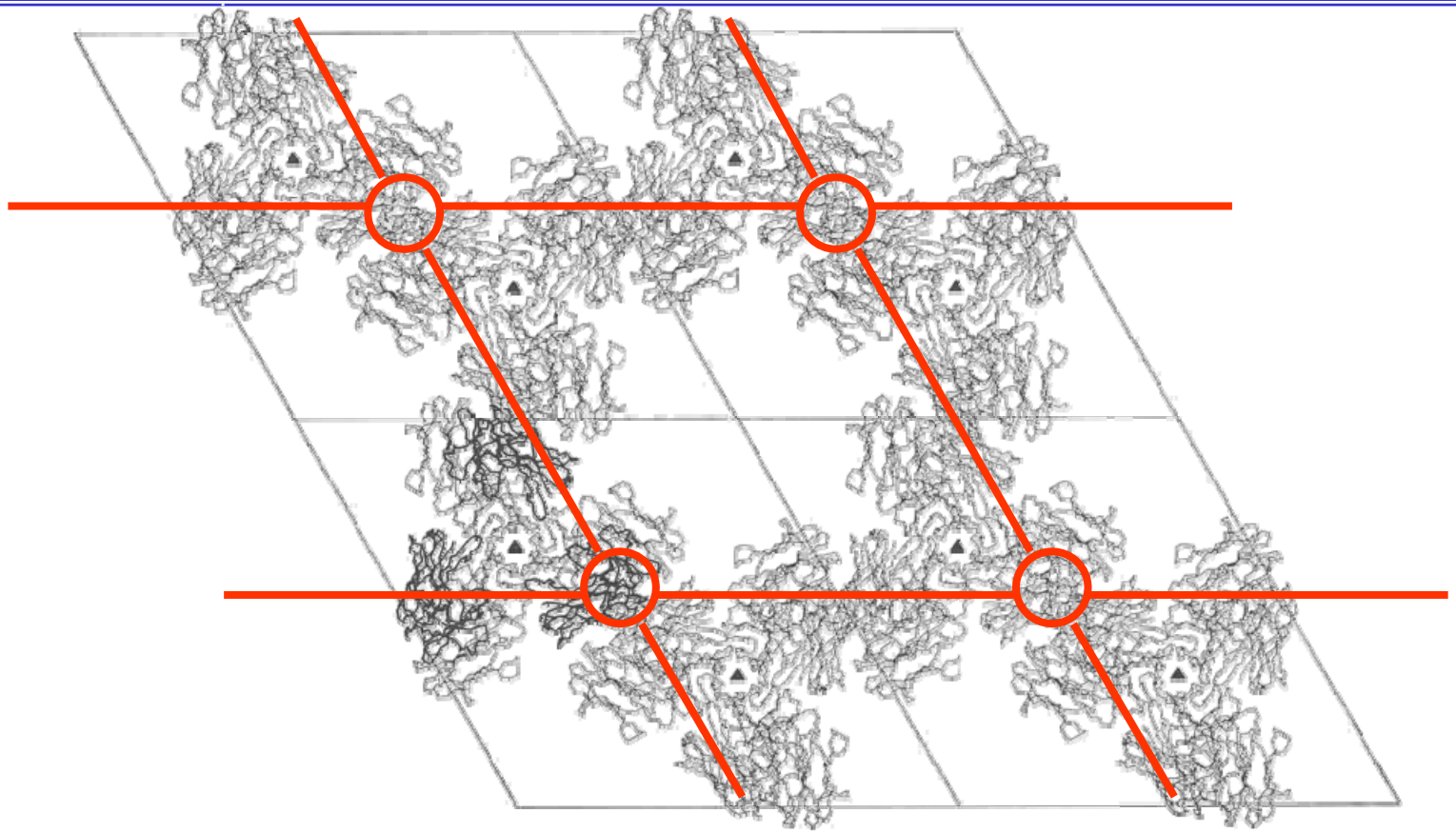
Protein-Kristalle

- Regelmäßige Anordnung einzelner Protein-Moleküle in einem Gitter
- Unregelmäßige Form der Proteine bedingt „Löcher“ im Kristall

  ⇒ **hoher Wassergehalt** (20 – 90%)
- **Einheitszelle:** die kleinste Untereinheit im Kristall, die allein durch Translation den gesamten Kristall erzeugen kann
- **Einheitszelle** enthält meist mehrere Protein-Einheiten



Protein-Kristalle



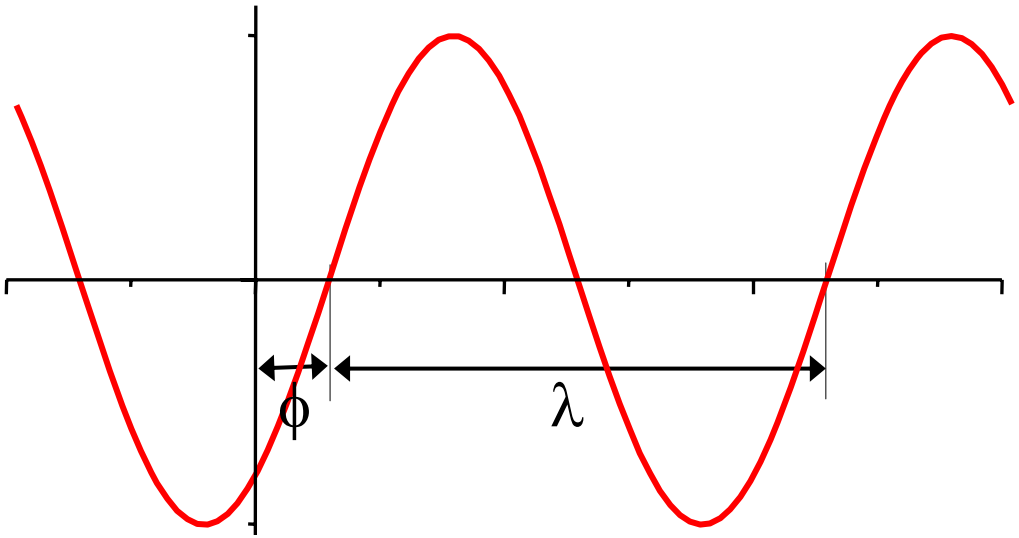
- Bsp.: Fab – Einheitszelle enthält zwei Kopien von Fab
- Kristall entsteht durch **Translation** dieser Einheitszelle entlang eines **regelmäßigen Gitters**

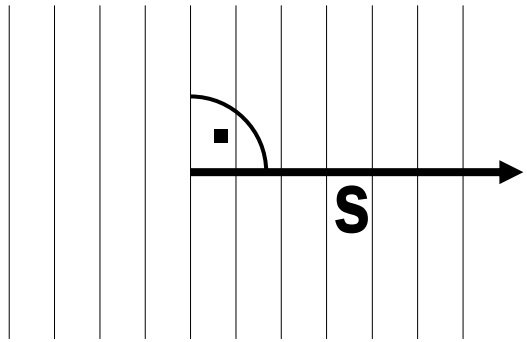
Röntgenbeugung an Proteinen

- **Bernal und Crowfoot** beobachteten bereits **1934**, dass Pepsin-Kristalle wohl definierte Beugungsmuster erzeugen
- Dennoch waren fast drei Jahrzehnte und die Erfindung des Computers notwendig bis **Kendrew und Perutz 1960** die ersten Strukturen (Myoglobin, Hämoglobin) aufklären konnten

Max Perutz, John Kendrew

Wellengleichungen





Welle zum Zeitpunkt *t* in **r** wird beschrieben durch:

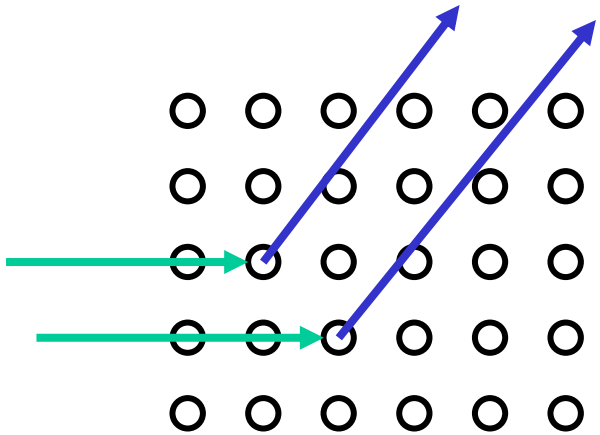
$$E(\mathbf{r},t) = E_0\, e^{-2\pi i\left(\frac{\mathbf{sr}}{\lambda}-\omega t+\phi\right)}$$

mit dem Einheitsvektor **s** in Richtung der Wellenfront,

der Frequenz ω, der Wellenlänge λ, der Phase ϕ und $i^2 = -1$

Beugung am Gitter

- Röntgenstrahlen haben Wellenlängen in der Größenordnung von Atomen
- Sie wechselwirken mit der Elektronenhülle der Atome
- Röntgenstrahlen werden an Atomen **gestreut**
- Gestreute Röntgenstrahlen **interferieren** miteinander

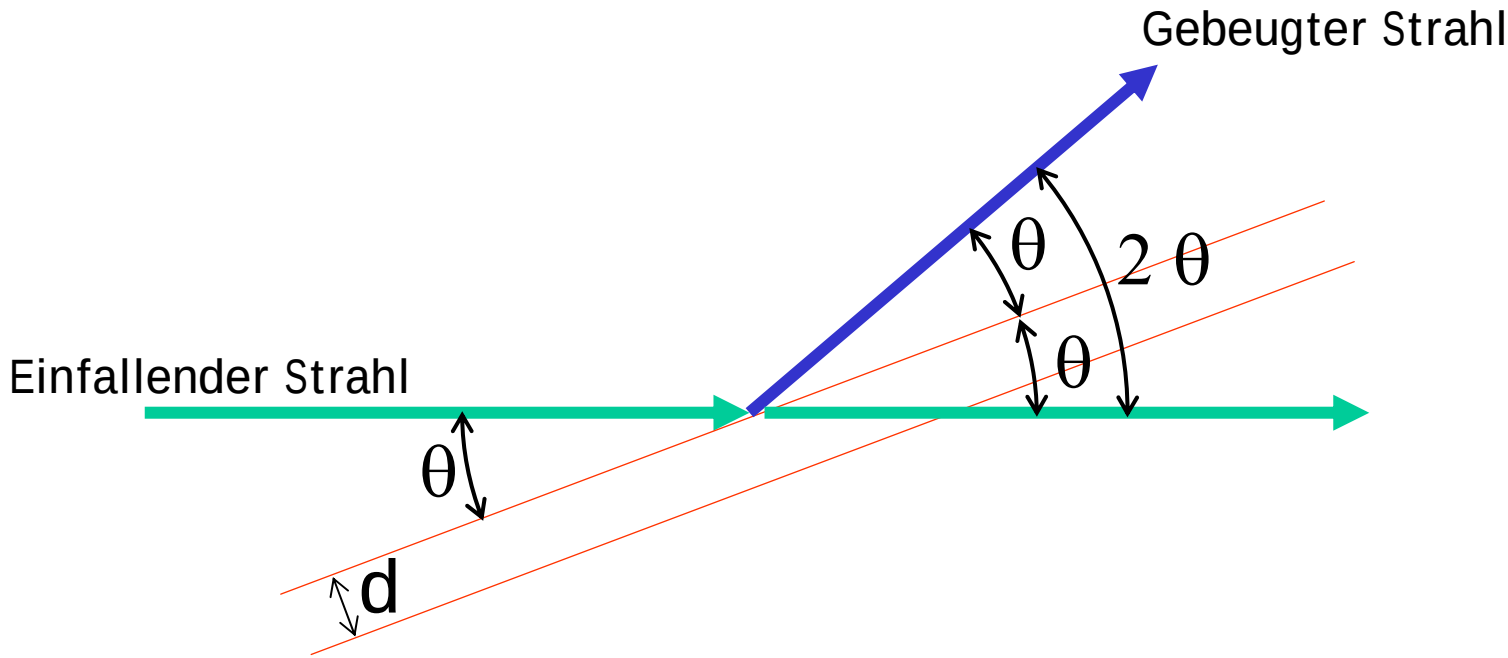


Beugung am Gitter

- **Braggsches Gesetz**

$$2d \sin \theta = \lambda$$

- Konstruktive Interferenz tritt auf unter Winkeln, die der Reflektion an Netzebenen im Kristall entsprechen

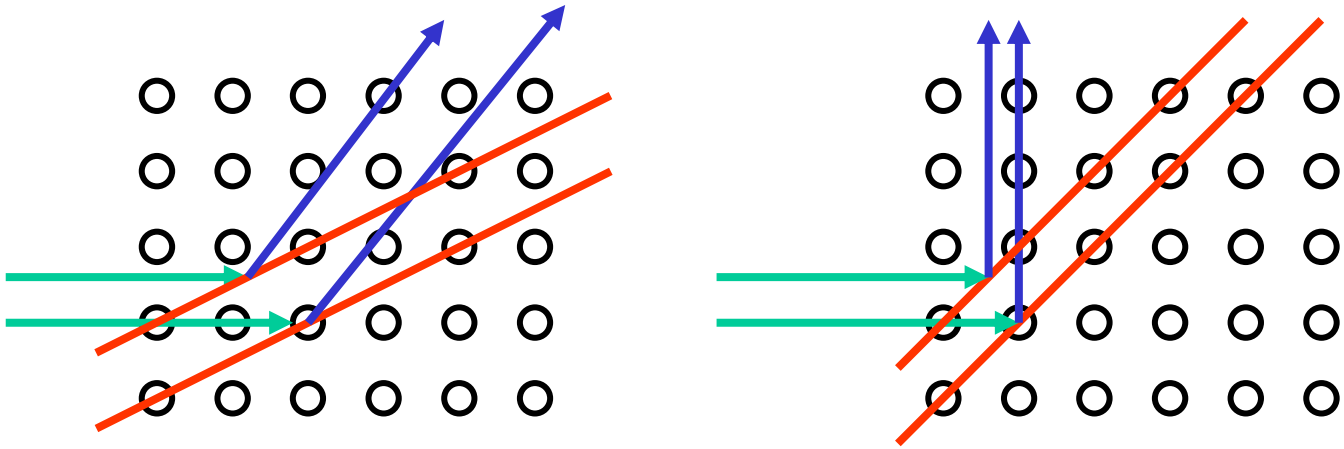


Beugung am Gitter

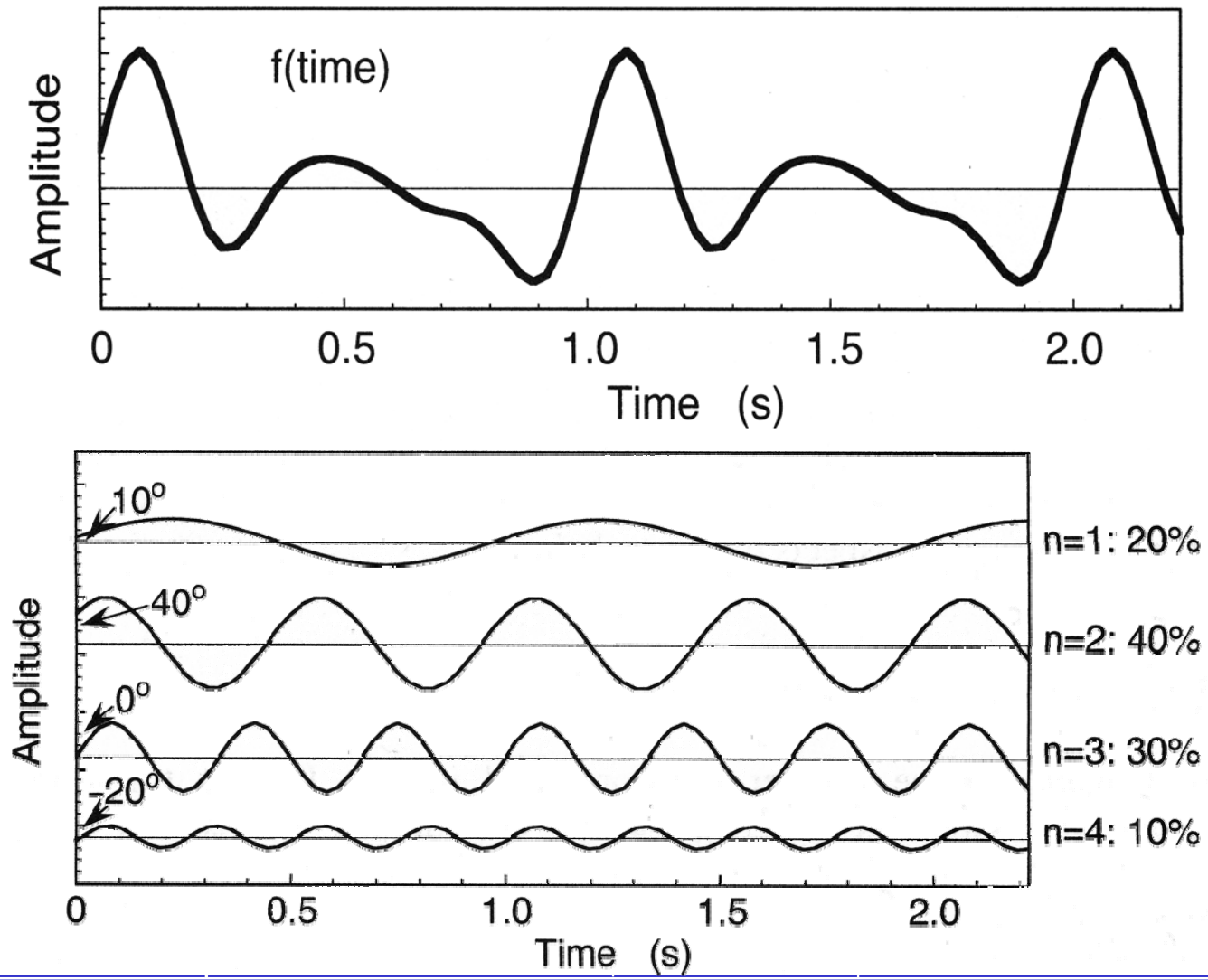
- **Braggsches Gesetz**

$$2d \sin \theta = \lambda$$

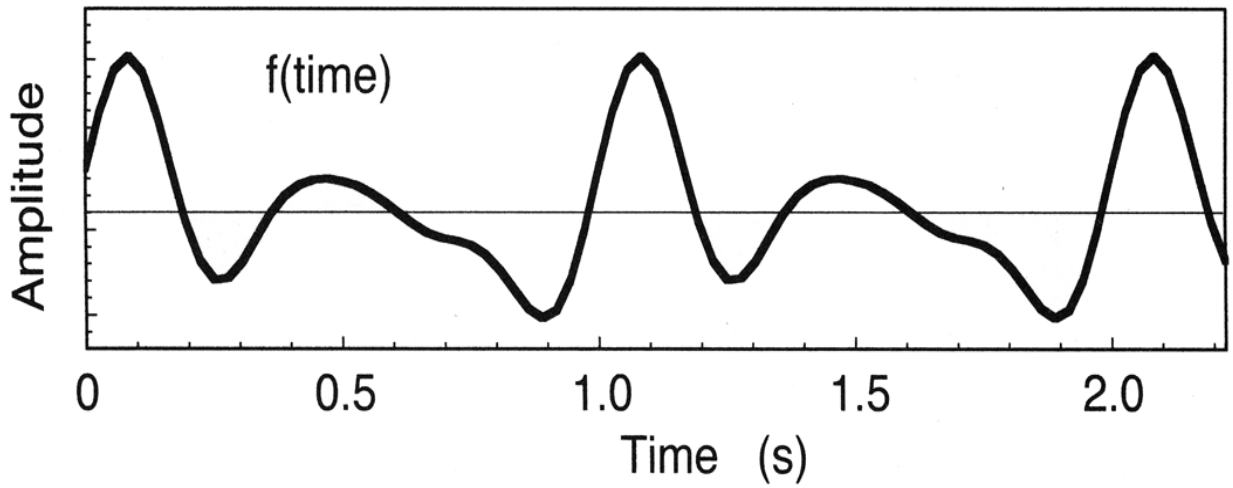
- Konstruktive Interferenz tritt auf unter Winkeln, die der Reflektion an **Netzebenen** im Kristall entsprechen

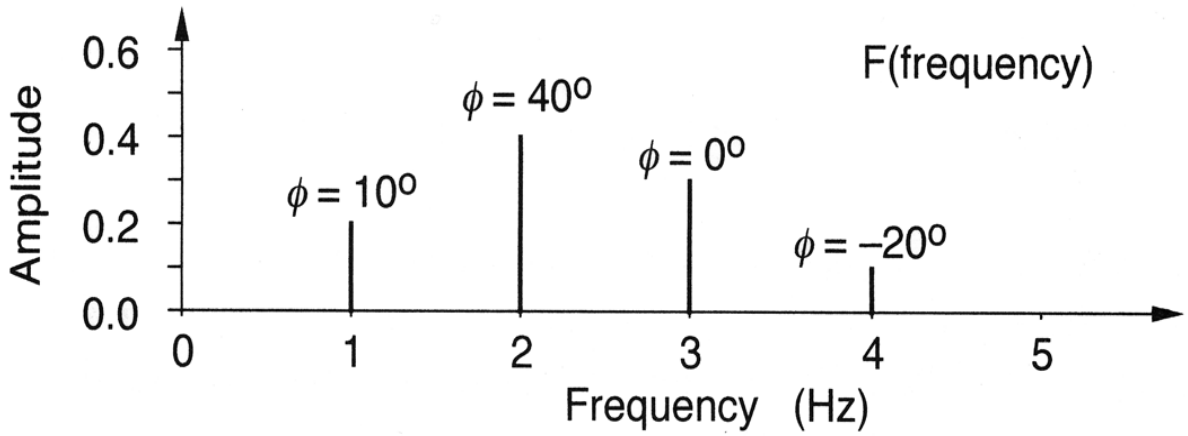


Fourier-Analyse

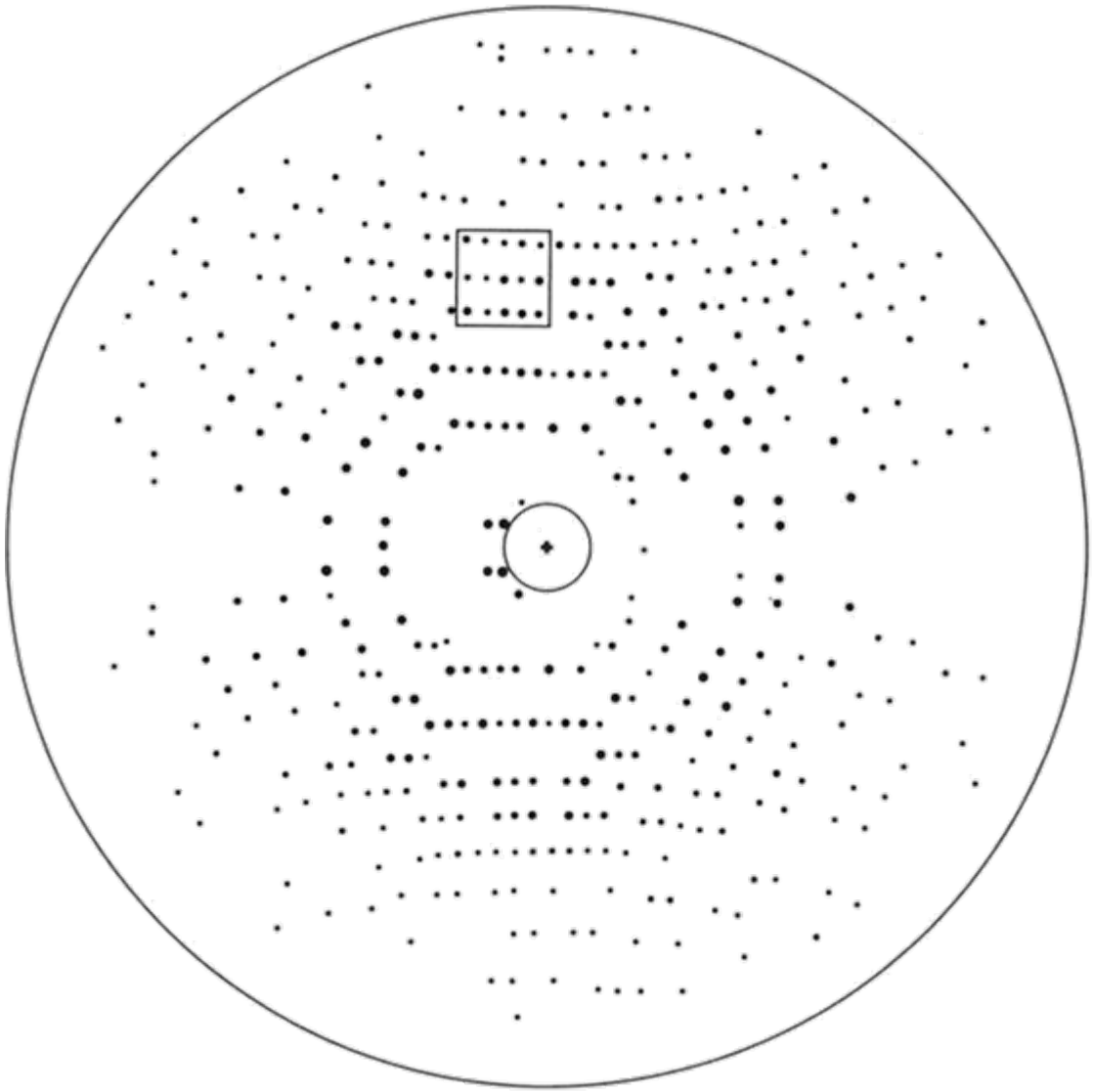


Fourier-Analyse

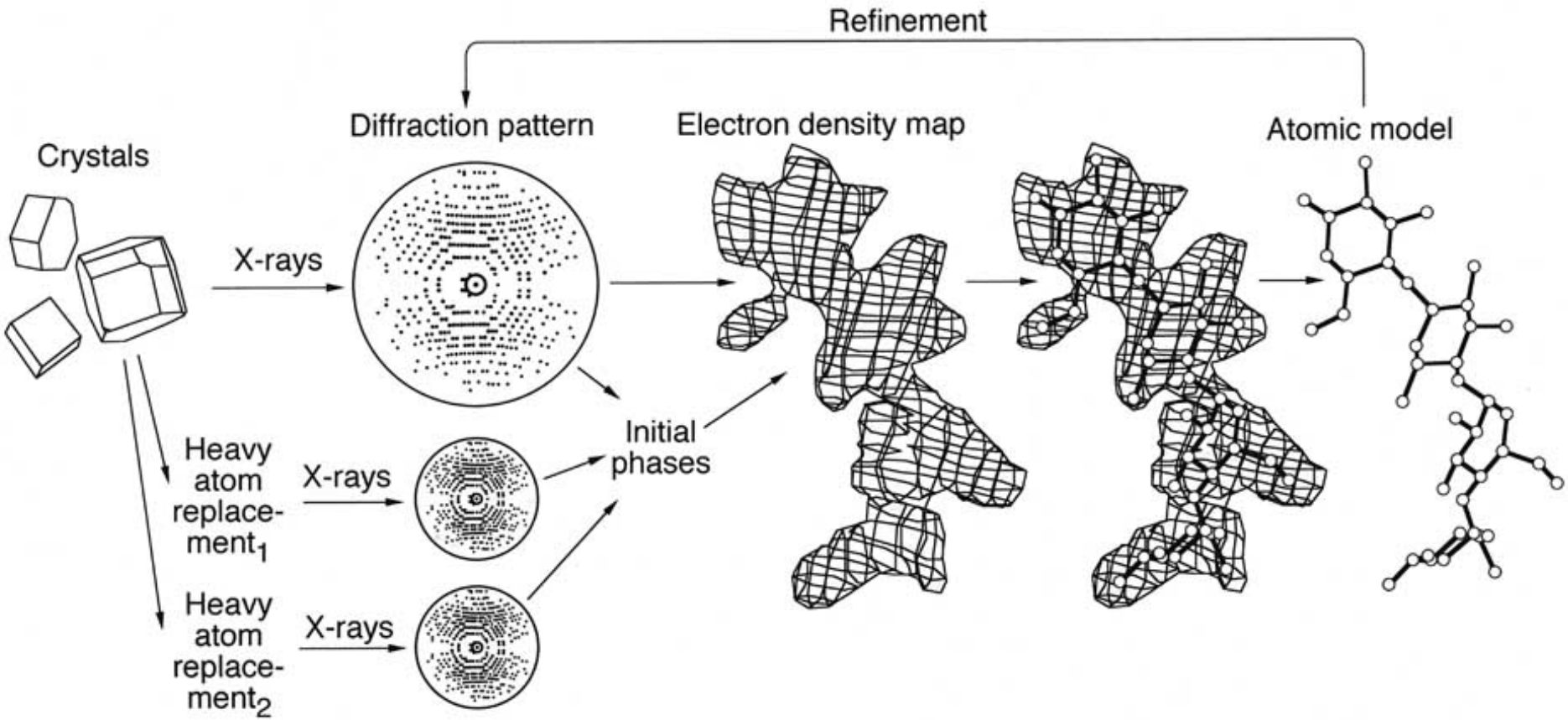




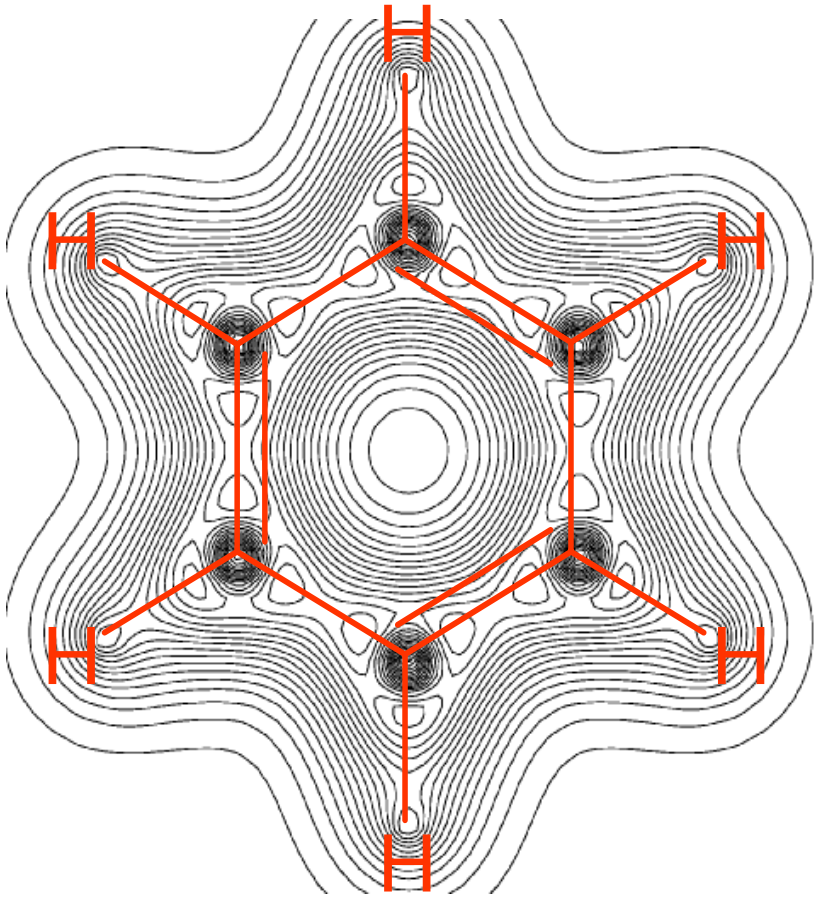
Diffraktionsmuster eines Proteins



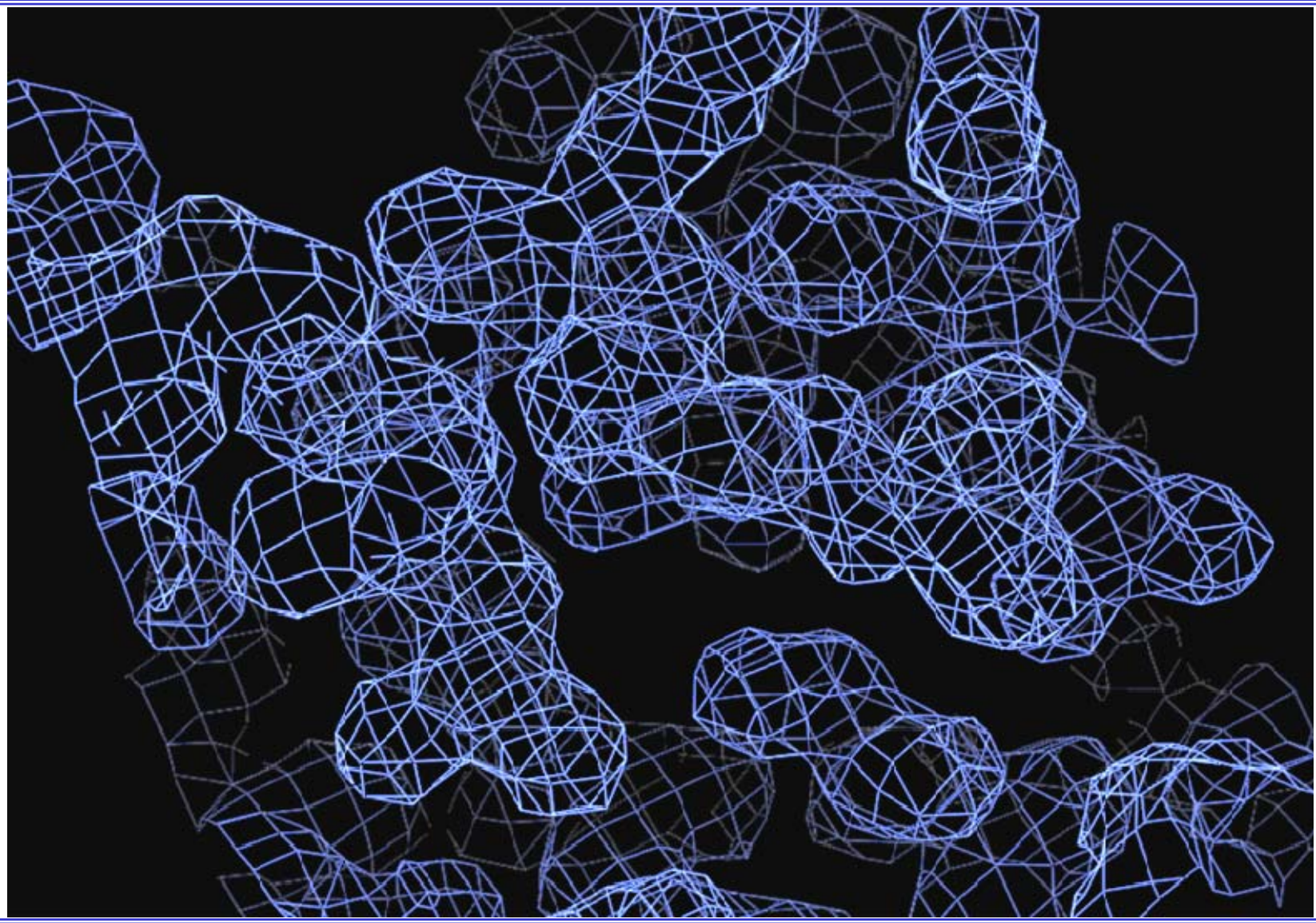
Überblick Röntgenbeugung



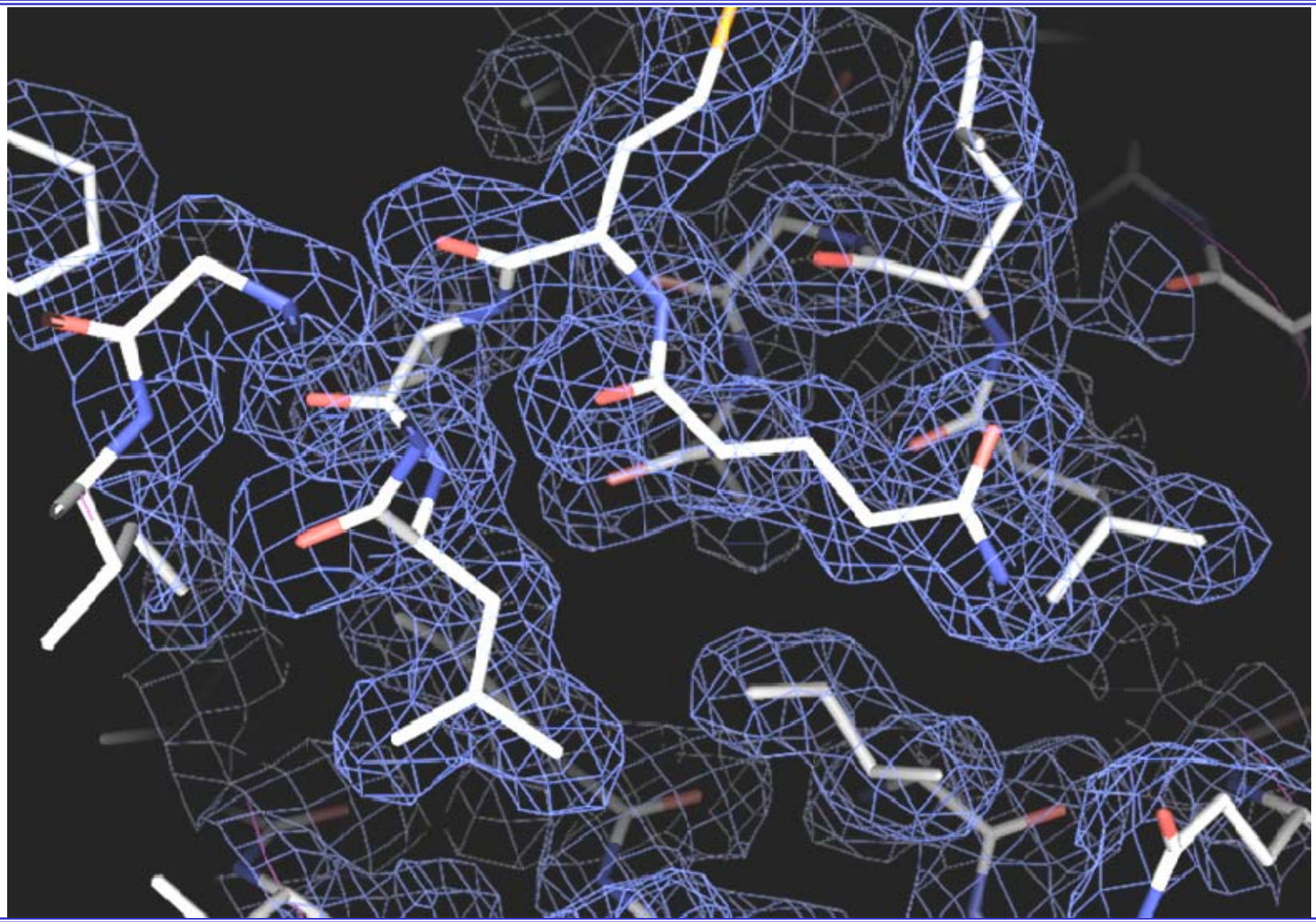
Elektronendichte-Karte



Elektronendichte-Karte

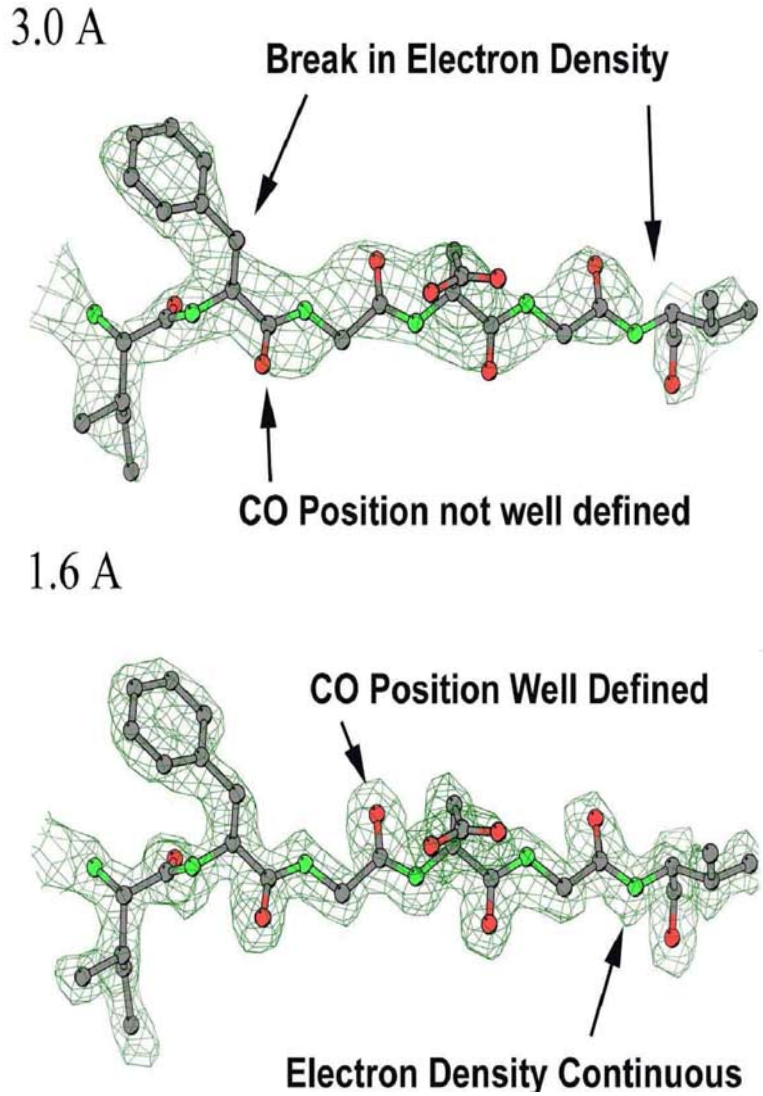


Elektronendichte-Karte



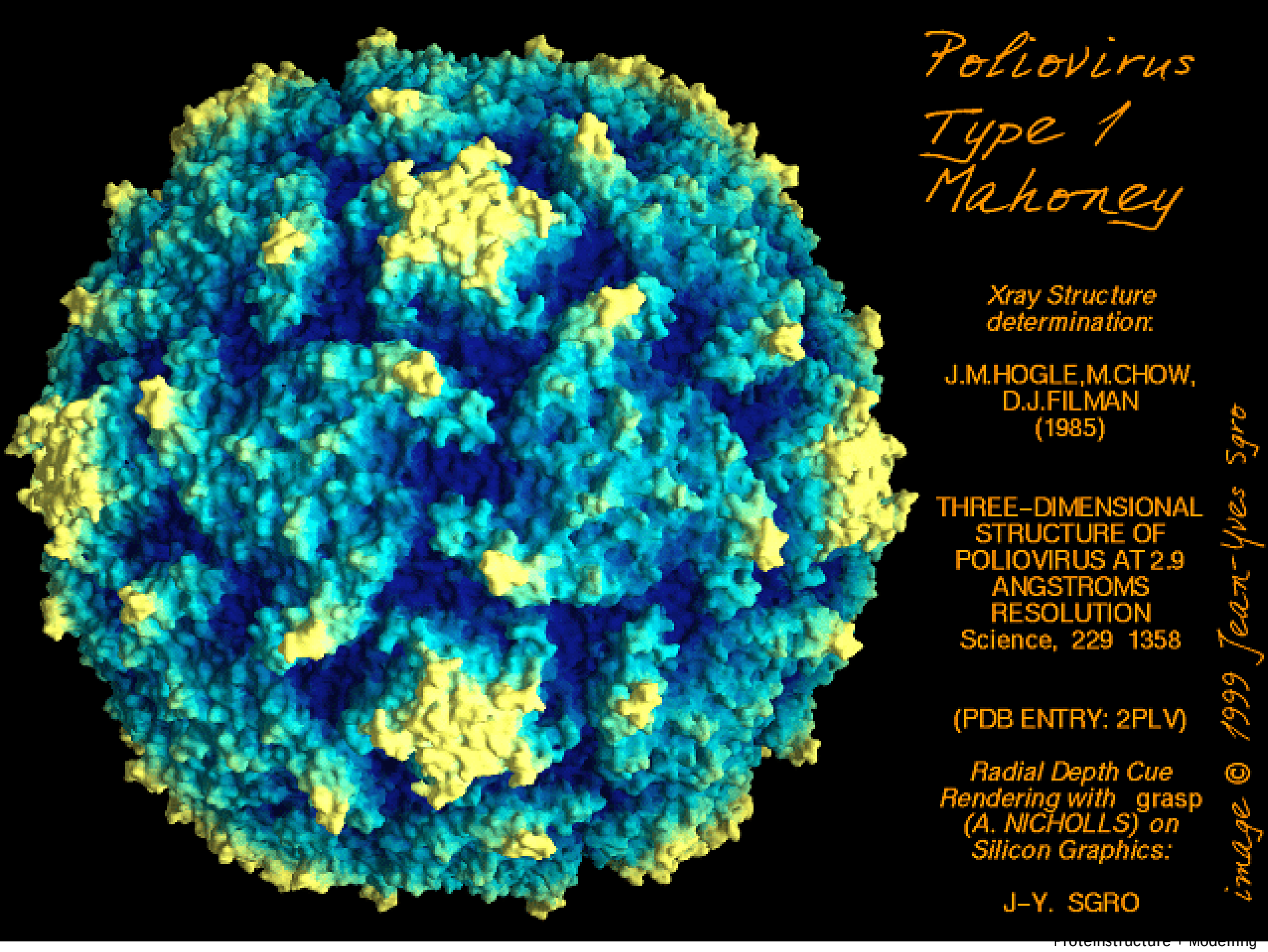
Auflösung

- Auflösung bestimmt die atomaren Details die erkennbar sind
- Schlechte Auflösung (hoher Wert) lässt atomare Details verschwimmen
- Auflösung wird angegeben in Ångstrom
- Auflösung von 2 Å bedeutet, dass Reflexe gesehen werden, die durch Streuung an parallelen Netzebenen im Abstand von 2 Å entstehen noch sichtbar sind
- Es bedeutet *nicht*, dass die Atomkoordinaten um 2 Å unbestimmt sind
- Fehler in der Atomkoordinaten liegen in diesem Fall bei ca. 0,3 Å

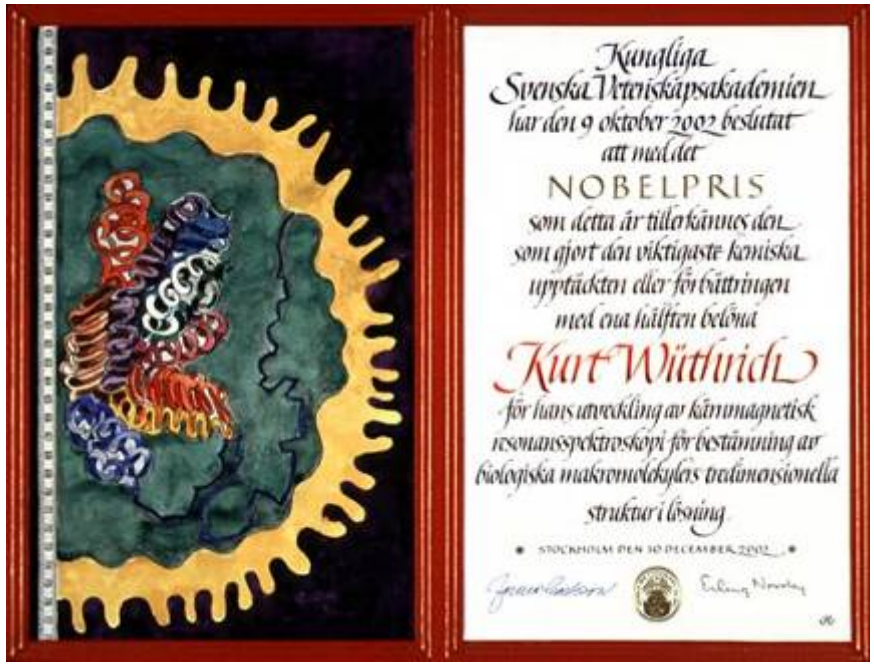


Auflösung

Auflösung [Å]	Enthaltene Information	
4.0	Faltungsklasse, einige Sekundärstrukturen	**schlecht**
3.5	Helices von Faltblättern unterscheidbar	
3.0	Die meisten Seitenketten erkennbar	**typisch**
2.5	Alle Seitenketten wohl definiert, ϕ und ψ im Rückgrat teilweise wohl definiert, Wasser erkennbar	
1.5	Torsionen im Rückgrat wohl definiert, erste H-Atome erkennbar	**sehr gut**
1.0	H-Atome sichtbar	**möglich**


Poliovirus
Type 1
Mahoney
Xray Structure determination:
J.M.HOGLE,M.CHOW, D.J.FILMAN (1985)
THREE-DIMENSIONAL STRUCTURE OF POLIOVIRUS AT 2.9 ANGSTROMS RESOLUTION
Science, 229 1358
(PDB ENTRY: 2PLV)
Radial Depth Cue Rendering with grasp (A. NICHOLLS) on Silicon Graphics:
J-Y. SGRO
image © 1999 Jean-Yves Sgro

NMR an Biomolekülen



Nobelpreis 2002 in Chemie für Kurt Wüthrich

„for his development of nuclear magnetic resonance spectroscopy for determining the three-dimensional structure of biological macromolecules in solution.“

Datenbanken – Swiss-Prot

Swiss-Prot

- Protein-Sequenz-Datenbank
  - 162781 Sequenzen (Release 44.7)
  - 314833 Literaturzitate
- Querverlinkt mit vielen anderen Datenbanken
- Homologie- und Ähnlichkeitssuche
- URL: **http://www.ebi.ac.uk/swissprot/index.html**

Protein-Datenbanken

Sequenzdaten

- **Swiss-Prot** – Protein-Sequenzen
  http://www.ebi.ac.uk/swissprot/index.html

Strukturdaten

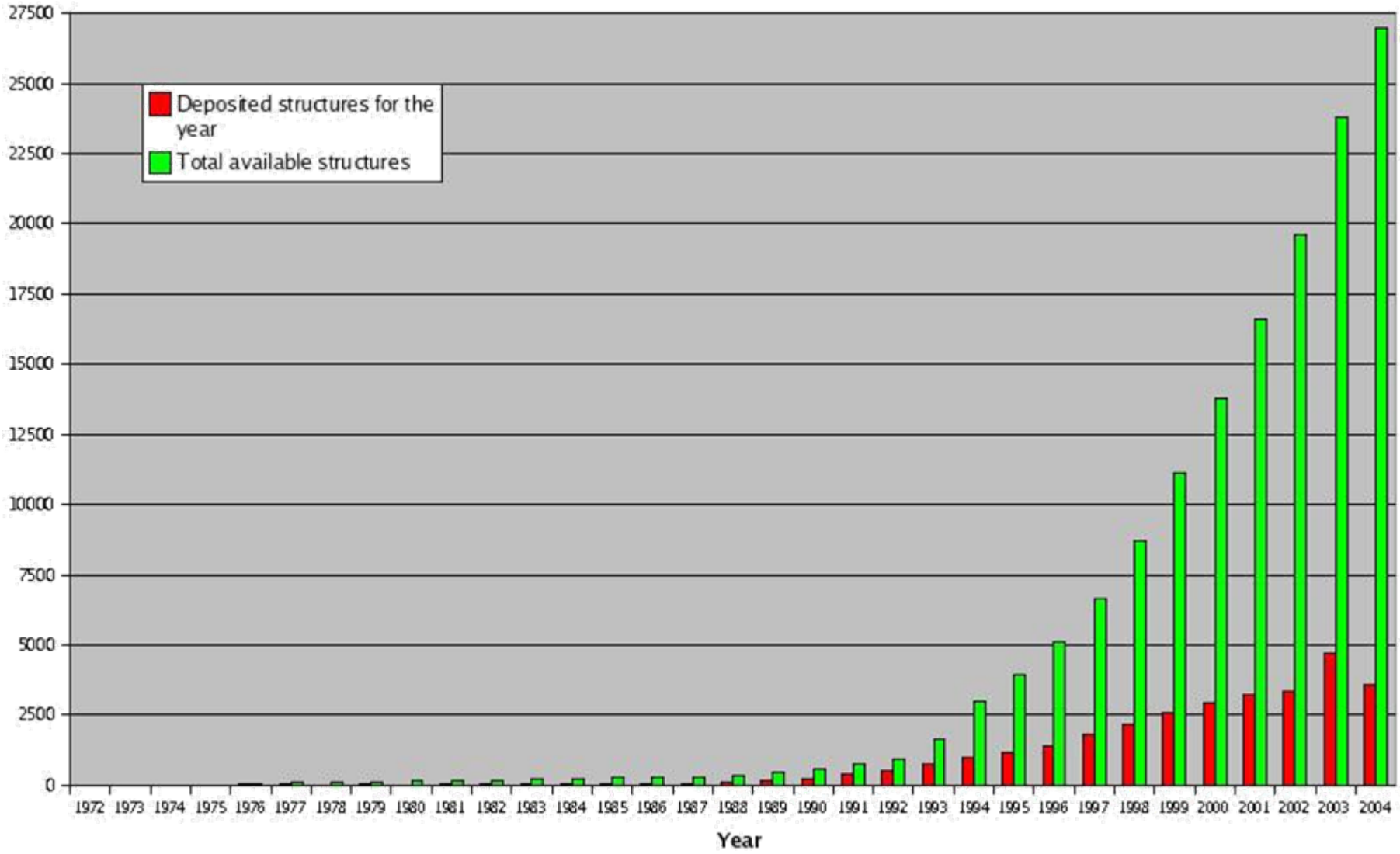
- **PDB** – 3D-Strukturen
  http://www.rcsb.org
- **BMRB** – NMR-Daten
  http://www.bmrb.wisc.edu
- **CATH –** Domänenklassifizierung
  http://www.biochem.ucl.ac.uk/bsm/cath/
- **SCOP** – Faltungsklassen
  http://scop.mrc-lmb.cam.ac.uk/scop/

Datenbanken – PDB

PDB (Protein Data Bank) – http://www.rcsb.org

- Strukturdaten von Biomolekülen
- Geführt von RCSB *(Research Collaboratory for Structural Bioinformatics)*
- Ablegen von Strukturen in der PDB heute Voraussetzung für strukturbiologische Publikation
- Alle Strukturen werden mit eindeutiger ID versehen
  - 4 Zeichen
  - 1. Zeichen – Version
  - 2. – 4. Zeichen – Struktur ID
  - Bsp.:
    - 2PTI, 3PTI, 4PTI sind drei Strukturen des Proteins BPTI
    - 2PTI: 1973, 3PTI: 1976, 4PTI: 1983

PDB – Wachstum



PDB – Statistik

	Proteine Peptide Viren	Protein-NA-Komplexe	Nuklein-säuren	Zucker	**Gesamt**
XRD, ND, TEM	21606	1067	748	14	23435
NMR	3276	103	610	4	3993
Gesamt	24882	1170	1358	18	**27428**

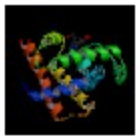
PDB – Dateiformate

- PDB
  - Uralt
  - Fortran-style spaltenbasiert)
  - Immer noch Standardformat!
  - „Lochkarten“ (records = cards!)
- mmCIF
  - Star-basiert (Strukturiert, Keywords in Dictionaries)
  - Auch alt (Star aus den 1970ern)
  - „Tabellen“
- XML
  - XML-Schema-basierend
  - Derzeit Beta-Test
  - „Baum“

PDB – Der erste Eintrag!

PDB
PROTEIN DATA BANK

Structure Explorer - 1MBN



Summary Information

Summary Information
View Structure
Download/Display File
Structural Neighbors
Geometry
Other Sources
Sequence Details

Explore
SearchLite SearchFields

Title: The Stereochemistry of the Protein Myoglobin
Compound: Myoglobin (Ferric Iron - Metmyoglobin)
Authors: H. C. Watson, J. C. Kendrew
Exp. Method: X-ray Diffraction
Classification: Oxygen Storage
Source: Physeter catodon
Primary Citation: Watson, H. C.: The Stereochemistry of the Protein Myoglobin *Prog.Stereochem.* 4 *pp.* 299 (1969)
[PubMed search]

Deposition Date: 05-Apr-1973 *Release Date:* 19-May-1976

Resolution [Å]: 2.00 *R-Value:* not available
Space Group: $P\,2_1$
Unit Cell: *dim [A]:* *a* 64.50 *b* 30.90 *c* 34.70
angles [°]: *alpha* 90.00 *beta* 106.00 *gamma* 90.00

PDB – Der erste Eintrag!

```
HEADER    OXYGEN STORAGE                          05-APR-73   1MBN      1MBNH  1
COMPND    MYOGLOBIN  (FERRIC IRON - METMYOGLOBIN)                        1MBN   4
SOURCE    SPERM WHALE (PHYSETER CATODON)                               1MBNM  1
AUTHOR    H.C.WATSON,J.C.KENDREW                                       1MBNG  1
[…]
REVDAT  20   27-OCT-83 1MBNS   1       REMARK                          1MBNS  1
JRNL        AUTH   H.C.WATSON                                         1MBNG  2
JRNL        TITL   THE STEREOCHEMISTRY OF THE PROTEIN MYOGLOBIN       1MBNG  3
JRNL        REF    PROG.STEREOCHEM.             V.   4   299 1969    1MBNG  4
JRNL        REFN   ASTM PRSTAP  US ISSN 0079-6808                419  1MBNG  5
[…]
SEQRES   1    153  VAL LEU SER GLU GLY GLU TRP GLN LEU VAL LEU HIS VAL  1MBN  39
[…]
HET    HEM      1      44     PROTOPORPHYRIN IX WITH FE(OH), FERRIC     1MBND 10
FORMUL   2  HEM    C34 H32 N4 O4 FE1 +++ .                             1MBNG 25
FORMUL   2  HEM    H1 O1                                               1MBNG 26
HELIX    1   A SER      3  GLU     18  1 N=3.63,PHI=1.73,H=1.50        1MBN  52
[…]
TURN     1 CD1 PHE    43  PHE    46     BETW C/D HELICES IMM PREC CD2   1MBN  60
[…]
ATOM      1  N   VAL     1      -2.900  17.600  15.500  1.00  0.00   2  1MBN  72
ATOM      2  CA  VAL     1      -3.600  16.400  15.300  1.00  0.00   2  1MBN  73
ATOM      3  C   VAL     1      -3.000  15.300  16.200  1.00  0.00   2  1MBN  74
ATOM      4  O   VAL     1      -3.700  14.700  17.000  1.00  0.00   2  1MBN  75
ATOM      5  CB  VAL     1      -3.500  16.000  13.800  1.00  0.00   2  1MBN  76
ATOM      6  CG1 VAL     1      -2.100  15.700  13.300  1.00  0.00   2  1MBNP  4
ATOM      7  CG2 VAL     1      -4.600  14.900  13.400  1.00  0.00   2  1MBNL  8
ATOM      8  N   LEU     2      -1.700  15.100  16.000  1.00  0.00   1  1MBN  79
ATOM      9  CA  LEU     2       -.900  14.100  16.700  1.00  0.00      1MBN  80
ATOM     10  C   LEU     2      -1.000  13.900  18.300  1.00  0.00      1MBN  81
ATOM     11  O   LEU     2       -.900  14.900  19.000  1.00  0.00      1MBN  82
ATOM     12  CB  LEU     2        .600  14.200  16.500  1.00  0.00      1MBN  83
ATOM     13  CG  LEU     2       1.100  14.300  15.100  1.00  0.00   1  1MBN  84
ATOM     14  CD1 LEU     2        .400  15.500  14.400  1.00  0.00   1  1MBNL  9
[…]
```

Datenbanken – BMRB

BMRB *(BioMagResBank)* – **http://www.bmrb.wisc.edu/**

- Enthält NMR-Daten von Biomolekülen
- University of Wisconsin – Madison
- Daten (u.a.)
  - Chemische Verschiebungen
  - NOEs
  - Kopplungskonstanten
  - Experimentelle Bedingungen
- Querverlinkt mit PDB, d.h. zusätzliche Annotationen für NMR-Strukturen der PDB
- Format: NMRStar, identifiziert durch BMRB Accession Number

Datenbanken – BMRB

- Derzeit (10/03) 2832 Einträge
  - 2711 Proteine, Peptide
  - 100 DNA
  - 48 RNA
- Daten für chemische Verschiebungen

  ~720,000 ^{1}H

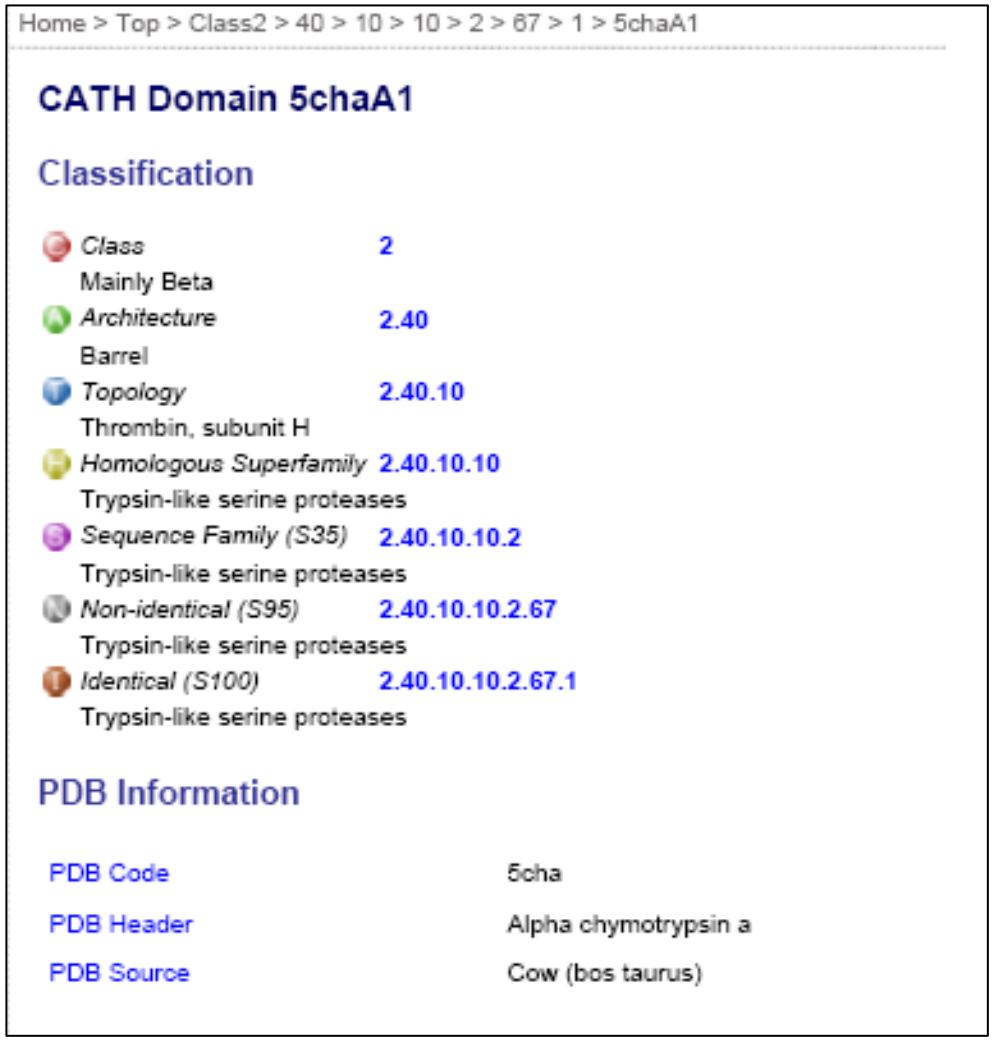
  ~350,000 ^{13}C

  ~120,000 ^{15}N

  ~500 ^{31}P (nur DNA, RNA)

Datenbanken – CATH

- Hierarchische Klassifizierung von Proteindomänen nach 3D-Struktur



Datenbanken – SCOP

SCOP – *Structural Classification of Proteins*

- Hierarchische Gruppierung aufgrund der 3D-Struktur

Protein: Ribonuclease inhibitor from Pig (*Sus scrofa*)

Lineage:

1. Root: scop
2. Class: Alpha and beta proteins (a/b)
   Mainly parallel beta sheets (beta-alpha-beta units)
3. Fold: Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)
   2 curved layers, a/b; parallel beta-sheet; order 1234...N
4. Superfamily: RNI-like
   regular structure consisting of similar repeats
5. Family: 28-residue LRR
6. Protein: Ribonuclease inhibitor
   duplication: consists of 16 repeats
7. Species: Pig (*Sus scrofa*)

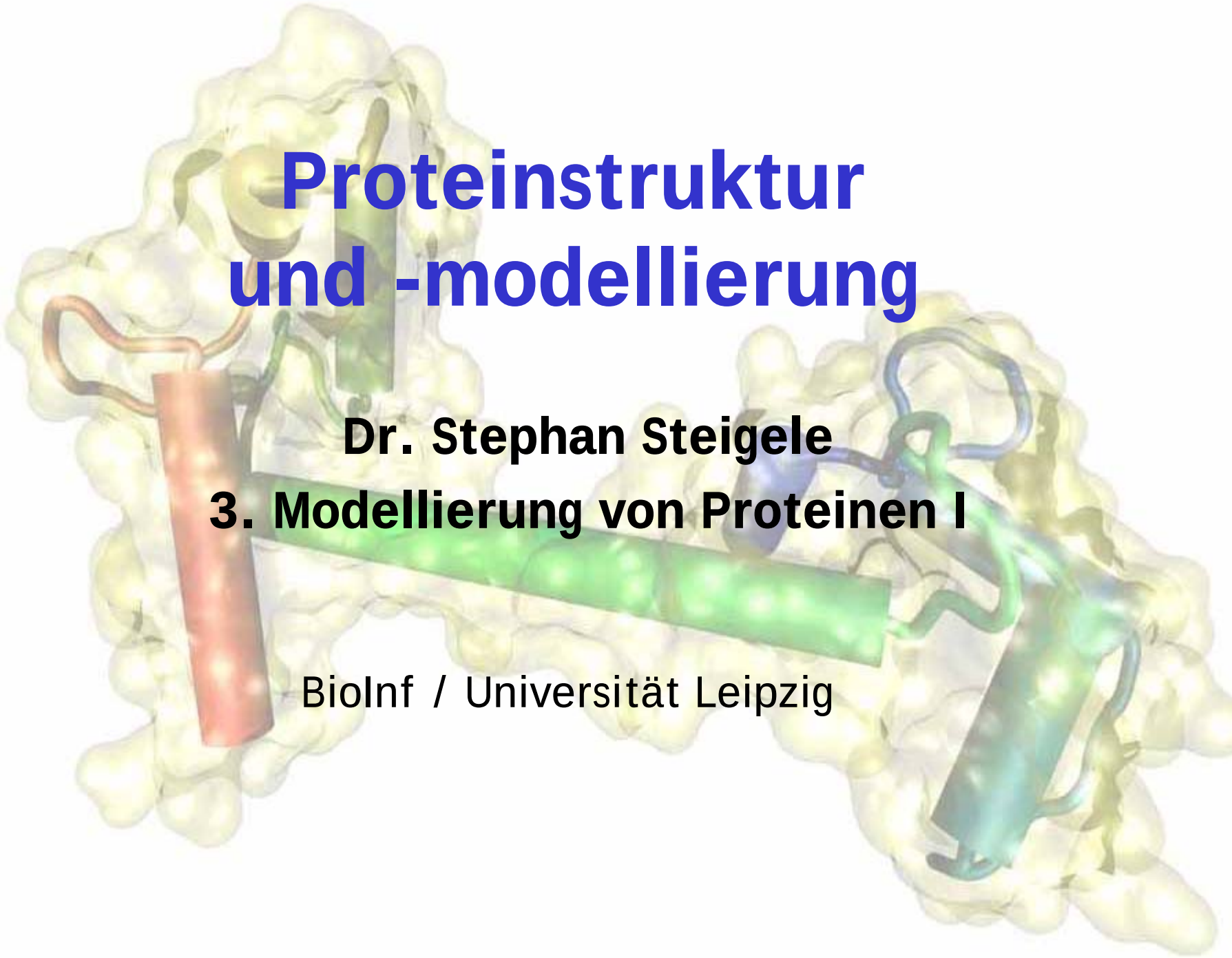
PDB Entry Domains:

1. 2bnh
   complexed with ace
2. 1dfj
   complexed with ace, so4
   1. chain i

Literatur

NMR

- **H. Günther, NMR-Spektroskopie,** Thieme, Stuttgart
- **H. Friebolin, Basic One- and Two-Dimensional NMR Spectroscopy**, VCH, Weinheim
- **Kurt Wüthrich, NMR of Proteins and Nucleic Acids**. John Wiley and Sons, 1986
- **J. Cavenagh, W. J. Fairbrother, A. G. Palmer, and N. J. Skelton, Protein NMR Spectroscopy: Principles and Practice**, Academic Press Inc., San Diego, 1996.



Proteinstruktur und -modellierung

Dr. Stephan Steigele

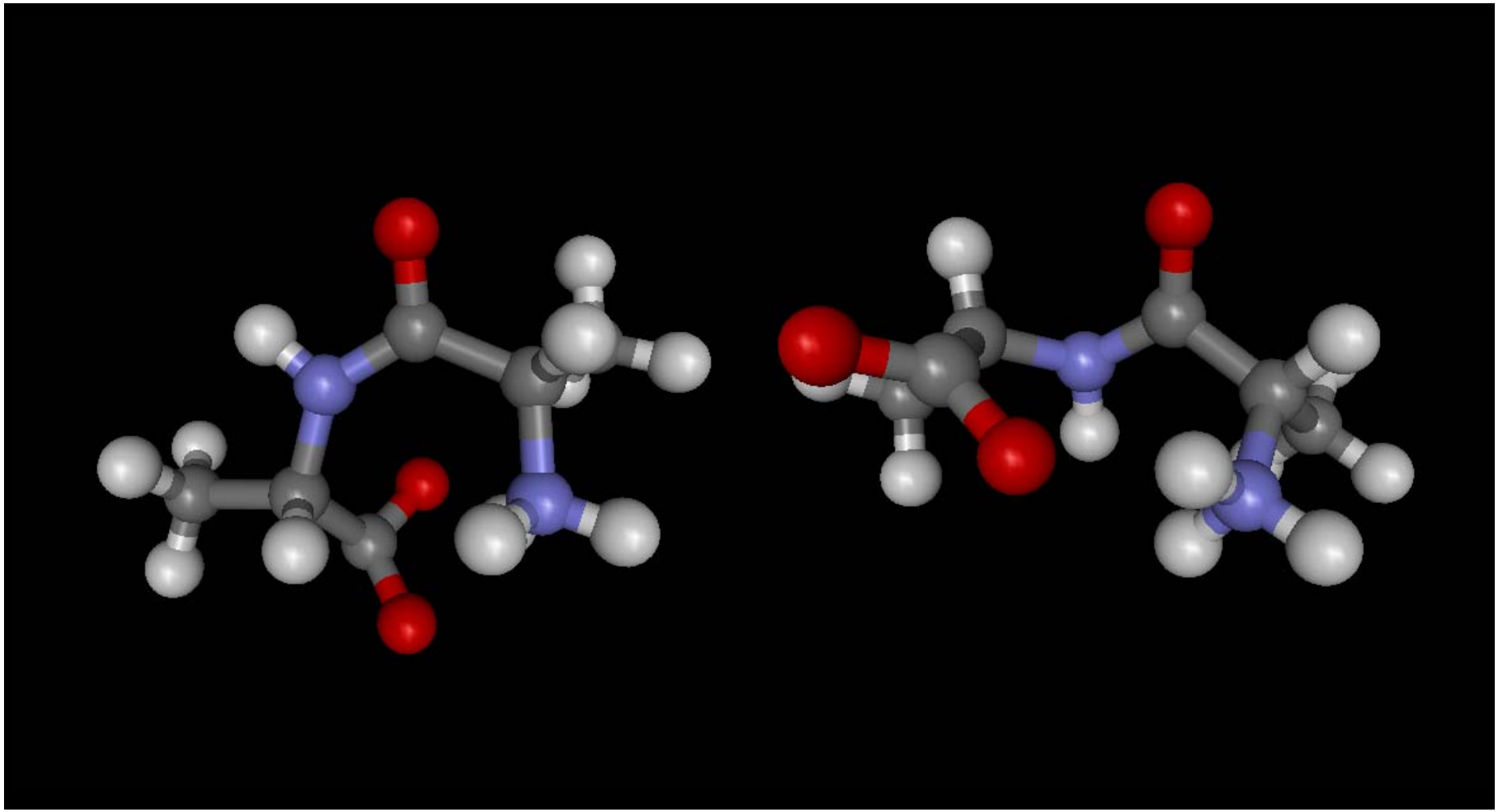
3. Modellierung von Proteinen I

BioInf / Universität Leipzig

Überblick

- Modellierung von Molekülen
- Molekülmechanik
  - Quantenmechanische Näherungen
  - Grundlagen der Molekülmechanik
    - Arten der Wechselwirkungen
    - Physikalische Grundlagen
    - Mathematische Modellierung
  - Kraftfelder
    - Definition
    - Einteilung
    - Beispiele

Motivation: Konformationsenergien



Welche Konformation ist günstiger? Um welchen Energiebetrag?

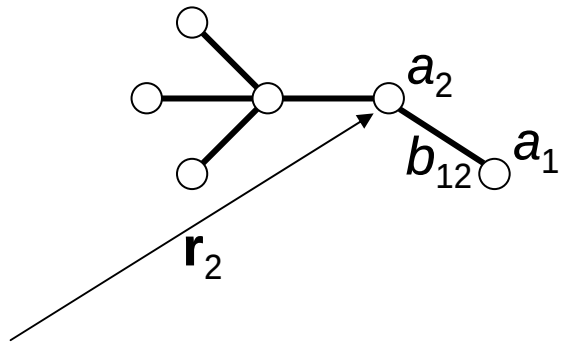
Einfaches Molekülmodell

Vereinfachtes Modell eines Moleküls

- Molekül mit N Atomen $A = \{a_i\}$ hat N **Koordinaten** $\mathbf{r}_i = (r_i^x, r_i^y, r_i^z)$, $R = \{\mathbf{r}_i\}$
- Kovalente Bindungen zwischen zwei Atomen a_i und a_j werden durch **Bindungen** $B = \{b_{ij} = (a_i, a_j)\}$ dargestellt
- **Molekül** *M* ist definiert durch einen Satz von Atomen A mit Koordinaten R und Bindungen B

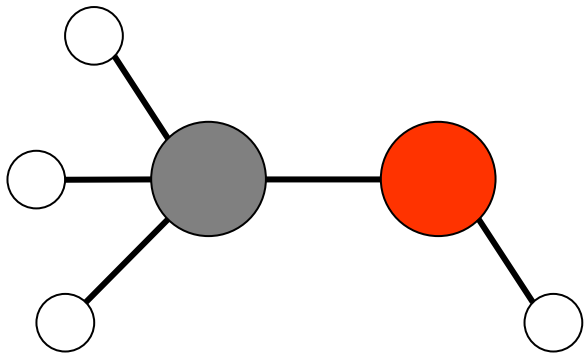
$$M = (A, R, B)$$

- Da alle Atome eines Moleküls „zusammenhängen“, besitzt der durch die Atome a_i und Bindungen b_{ij} definierte Graph exakt eine **Zusammenhangskomponente**



Einfaches Molekülmodell

- In diesem Modell werden keine Bindungen erzeugt oder gebrochen, d.h. die **Topologie ist unveränderlich** (*B* = const.).
- Koordinaten sind variabel und beschreiben Konformation des Moleküls
- Koordinaten $\mathbf{r}_i$ spannen den **Konformationsraum** des Moleküls auf
- Diese Beschreibung lässt sich auch auf Systeme mehrerer Moleküle anwenden) Graphen mit mehreren Zusammenhangskomponenten.
- Knoten des Graphs können beliebige **Beschriftungen** tragen, z.B. das Element oder Ladung



Modellierung

- Beim Modellieren von Proteinen ist die Topologie meist gegeben (Sequenz!). Je nach Problem sind Konformationen gegeben oder gesucht.
- Gesucht sind meist
  - **Optimale Konformationen**

    In der Natur kommt in der Regel die energieärmste Konformation am häufigsten vor. Eine gute Energiefunktion, kann diese Konformation identifizieren.
  - **Energiedifferenzen**

    Hat man mehrere unterschiedliche Konformation, will man die energetischen Unterschiede quantifizieren (z.B. ist cis oder trans günstiger)
  - **Absolute Energien**

    Insbesondere für intermolekulare Energien ist man daran interessiert, absolute Energien zu bestimmen (z.B. Vorhersage von Bindungsenergien von Wirkstoffen)
  - **Dynamisches Verhalten**

    Proteine sind flexibel. Die Energiefunktion läßt auch die Beschreibung des dynamischen Verhaltens (zeitliche Entwicklung) zu.

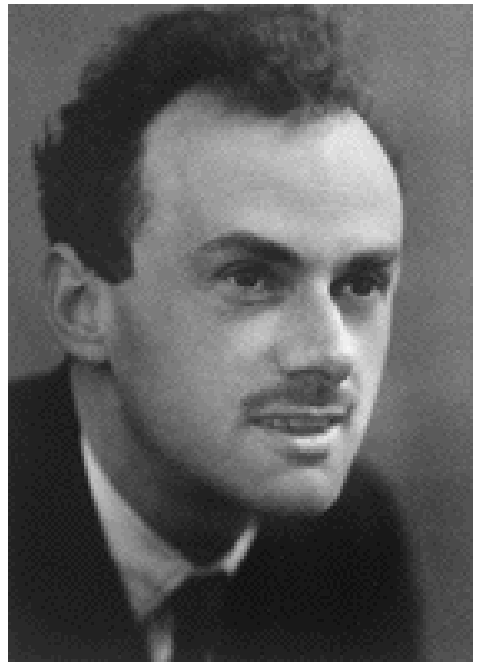
Modellierung

- Zur Modellierung von Proteinen (oder Molekülen allgemein) muss man die physikalischen Wechselwirkungen zwischen Atomen und Molekülen beschreiben
  - *Intra***molekulare Wechselwirkungen** (innerhalb)
  - *Inter***molekulare Wechselwirkungen** (zwischen)
- Diese Wechselwirkungen lassen sich in der Regel als Energien beschreiben, d.h. es existiert eine **Energiefunktion** *E(M)* die jeder Konformation *M* eine Energie zuordnet
- Die Energiefunktion ist in der Regel von der **Konformation** und von der **Topologie** des Moleküls abhängig

Wechselwirkungen

Paul Dirac (1929):

"The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation."



Paul A. M. Dirac

P.A.M. Dirac, Quantum Mechanics of Many-Electron Systems, Proc. R. Soc. London, A 123, 714 (1929)
http://en.wikipedia.org/wiki/Image:Dirac.gif

Quantenmechanik

- Diracs Aussage bezog sich auch die Quantenmechanik (QM)
- Im Zentrum der QM steht die Schrödinger-Gleichung

$$\mathbf{H}\Psi = \mathrm{E}\Psi$$

Dabei ist **H** der **Hamilton-Operator**, der das System (d.h. Elektronen und Kerne) beschreibt und Ψ die **Wellenfunktion**, deren Quadrat die Aufenthaltswahrscheinlichkeit der Elektronen beschreibt

- Um zu einer Energiefunktion zu kommen, muss man die Topologie und Geometrie in den Hamilton-Operator verpacken und obige Gleichung nach E auflösen.

Quantenmechanik

- Für hinreichend kleine Systeme ist das kann man quantenmechanisch rechnen
- Unter der **Born-Oppenheimer-Approximation**, trennt man die Bewegungen der Elektronen von der (langsameren) Bewegung der Kerne
- **H** enthält dann nur noch die elektrostatische Wechselwirkung der Elektronen untereinander und mit den Kernen
- Topologie entfällt, da die Positionen der Kerne für eine quantenmechanische Beschreibung ausreichend sind

Quantenmechanik

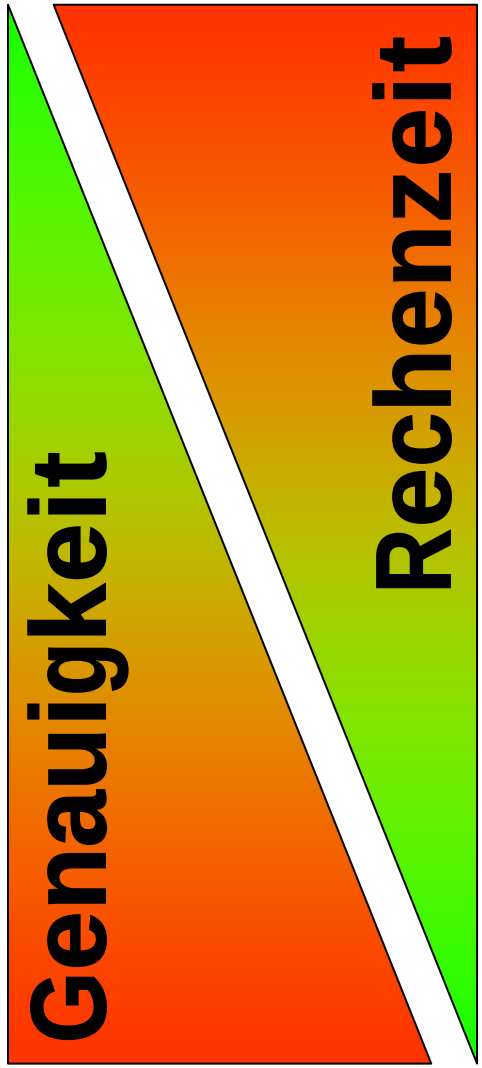
- Berechnet man die Wellenfunktion Ψ, so erhält man Elektronendichte und Energie des Systems
- Berechnungen sind extrem aufwändig
- Gute Approximationen skalieren wie $O(N^4)$ – $O(N^8)$ mit der Anzahl N der Elektronen(!) im Molekül
- Peptid mit 60 AS = 400 Schweratome, 400 H

  $\Rightarrow$ ~ 3000 e^-
- Selbst heute nicht möglich komplett quantenmechanisch zu rechnen

  $\Rightarrow$ **weitere Näherungen notwendig**

Hierarchie der Näherungen

- **Quantenmechanische Methoden**
  - **Ab initio** – volle QM-Rechnung ohne empirische Näherungen
  - **Semiempirisch** – QM-Rechnung, bei der einige Teile durch (empirische) Näherungen ersetzt werden
- **Molekülmechanische Methoden**
  - Klassisch-mechanische Modelle, an empirische Daten angepasst



Molekülmechanik

- Beschreibung der WW durch **klassisch mechanische Modelle**
- Keine explizite Betrachtung von Elektronen
- Effiziente Berechnung aufgrund einfacher (klassischer) Terme
- Schwierigkeit
  - Viele **Parameter**, die an experimentelle Daten angepasst werden müssen
  - **Genauigkeit und Übertragbarkeit** müssen verifiziert werden

Wechselwirkungen

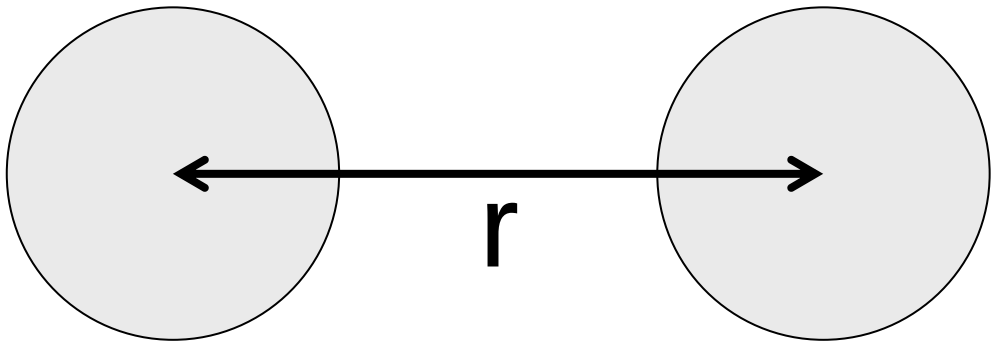
Einfluss auf die Energie haben

- **Topologie und Atomeigenschaften**
  - Art der Atome (Element)
  - Ladungen
  - Bindungen
    - Art (einfach, doppelt, aromatisch)
    - Länge
- **Geometrie**
  - Position der Atome
    - Torsionswinkel
    - Bindungswinkel
    - Bindungslängen
    - ….

Edelgasatome

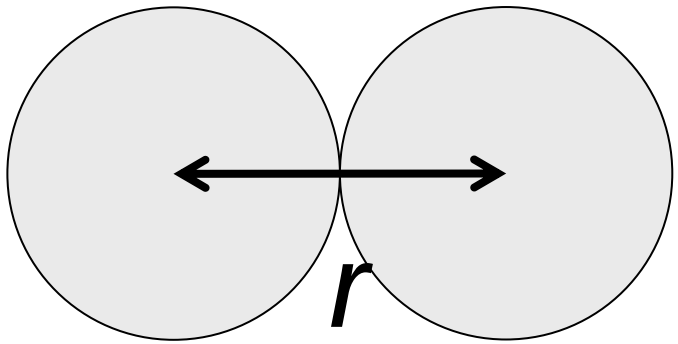
Beispiel:

- Welche Energiefunktion beschreibt z.B. zwei Edelgasatome (Ar – Ar)
- Keine Bindungen (Edelgas!)
- Eindimensionales System: Abstand *r*(Ar—Ar) ist ausreichend zur Beschreibung



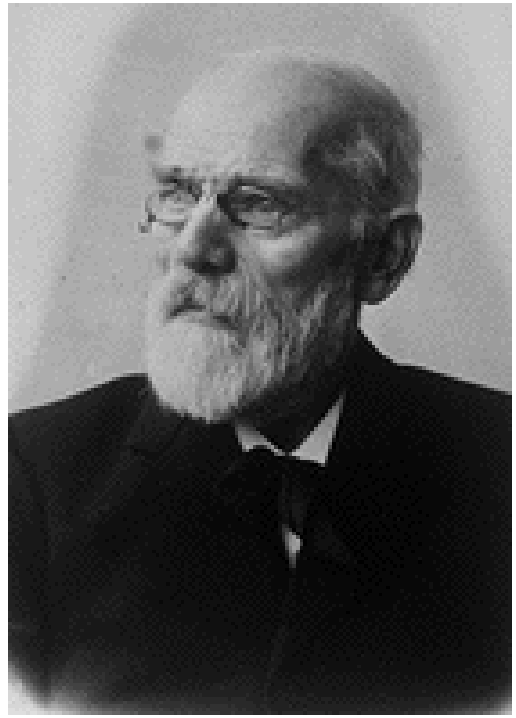
Edelgasatome

- Energiefunktion $E(r)$ ordnet jeder Anordnung der beiden Atome eine Energie zu
- $E(r)$ muss für sehr kleine Werte von r positiv sein (E positiv = energetisch ungünstig), da Ar nicht beliebig kompressibel ist (Atome können sich nicht durchdringen)
- E ist also **repulsiv** für kleine Abstände



Reale und ideale Gase

- **Johannes van der Waals** postulierte 1873 bei der Untersuchung von Gasen, dass es Kräfte zwischen den Atomen/Molekülen geben müsse, die die Abweichungen zwischen realen und idealen Gasen erklären
- Diese Kräfte müssen von der Natur her anziehende (attraktive) Kräfte sein, da die realen Gase beim gleichen Volumen einen niedrigeren Druck besitzen als ideale Gase.



J. D. van der Waals

menten dus gelijk nul zal zijn. Bij gevolg wordt onze bewegingsvergelijking, zoo p den uitwendigen druk, v het volume, b eenige malen het molekulair-volume, en a de specifieke attractie voorstelt.

$$\left(p + \frac{a}{v^2}\right)(v - b) = \Sigma \frac{1}{3} m V^2 \quad (\gamma)$$

Daar $\Sigma \frac{1}{3} m V^2$ met wat men gewoonlijk door temperatuur verstaat toeneemt, zullen wij daarvoor in de plaats schrijven

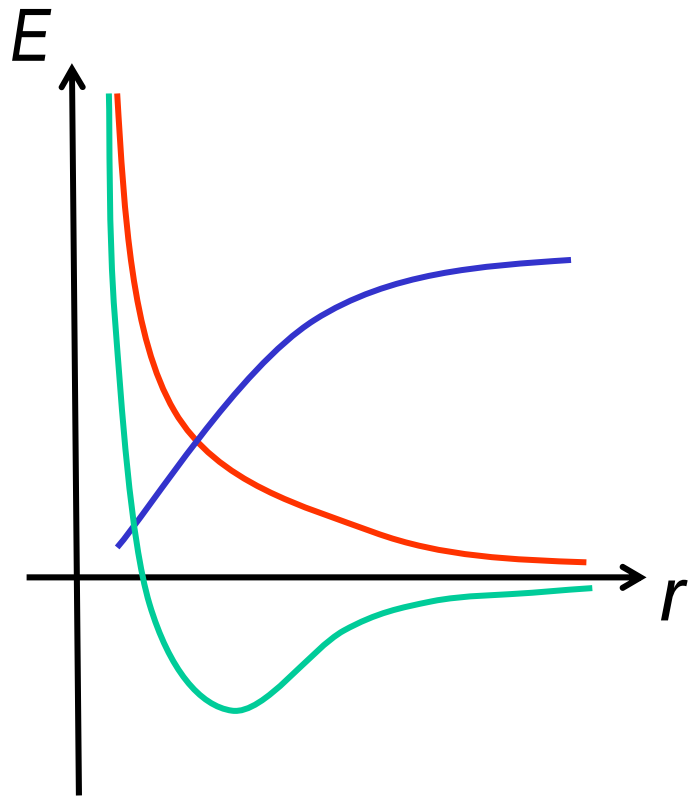
J. D. van der Waals: Over de Continuiteit van den Gas- en Vloeeistoftoestand, Dissertation, Leiden, 1873
http://nobelprize.org/physics/laureates/1910/waals.gif

Edelgasatome

- Zwei Beiträge
  - **Attraktiv**
  - **Repulsiv**
- Wechselwirkungen verschwinden im Unendlichen

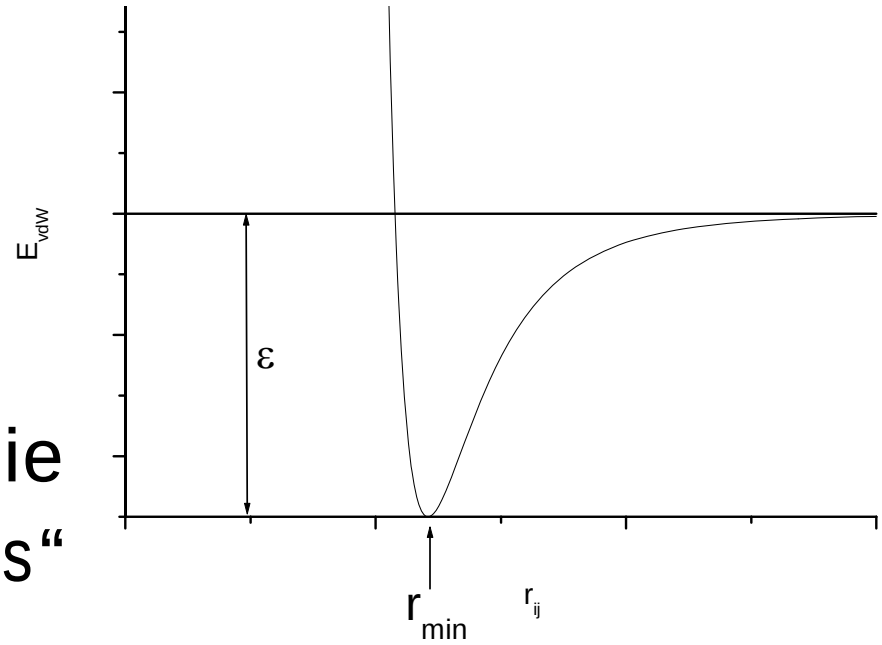
$$\lim_{r \to \infty} E(r) = 0$$

- Gesamtenergie: Summe aus attraktivem und repulsivem Beitrag



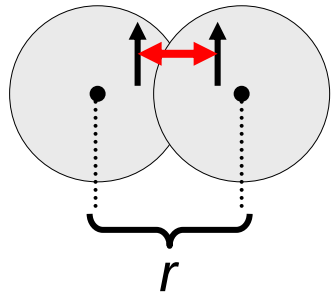
Van-der-Waals-Wechselwirkung

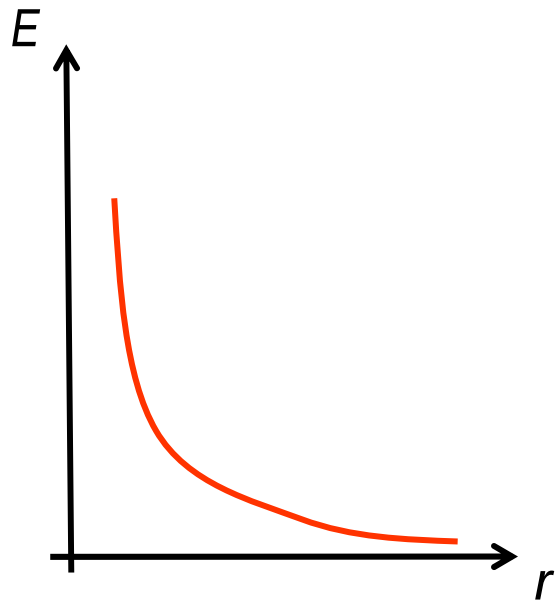
- Man bezeichnet diese Art der Energiefunktion (oder Potential) allgemein als van-der-Waals-Potential
- Beschrieben wird das Potential über die Lage r_{min} des Minimums und die Tiefe ε des „Energietopfs“
- Analytische Funktionen zur Beschreibung variieren



Repulsionsenergie

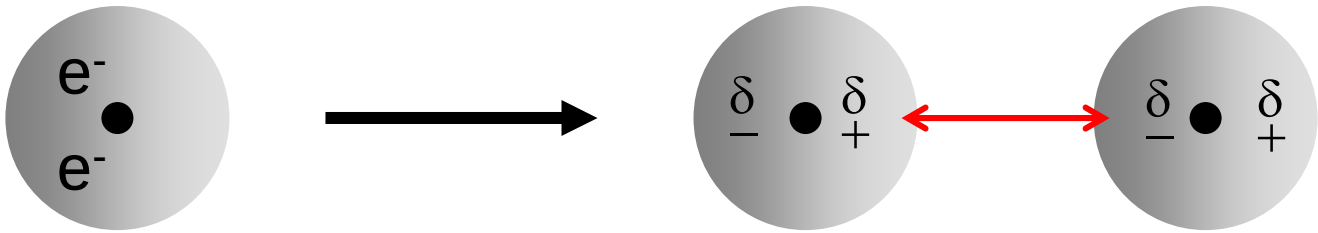
- Abstoßender Teil der vdW-Energie stammt aus der Abstoßung der überlappenden nicht bindenden Orbitale (**Austausch-Wechselwirkung**, Pauli-Verbot)
- Stärke der Abstoßung daher von den Elektronenkonfigurationen der beiden beteiligten Atome abhängig
- **Quantitative Herleitung** ist **quantenmechanisch** möglich
- Abstoßung wächst sehr schnell („**hartes Potential**“)





Dispersionsenergie

- Attraktiver Teil des Potentials verursacht durch **Dispersions-Wechselwirkung**
- Elektrostatischer Effekt
  - Dipol-Dipol-WW
  - Induzierte Dipole der polarisierbaren Elektronenhüllen



Spontane asymmetrische Elektronenverteilung

Spontaner Dipol

Induzierter Dipol in benachbartem Atom) anziehende Kraft

Dispersionsenergie

- **Fritz London** konnte 1930 quantenmechanisch eine Näherung für die Dispersionsenergie (oder seither auch *London-Energie*) herleiten
- Näherung beruht auf einem einfachen Modell zweier wechselwirkender Dipole
- Ergebnis:

$$E_L = -\frac{3\alpha^4 \hbar\omega}{4(4\pi\varepsilon_0)^2}\frac{1}{r^6}$$

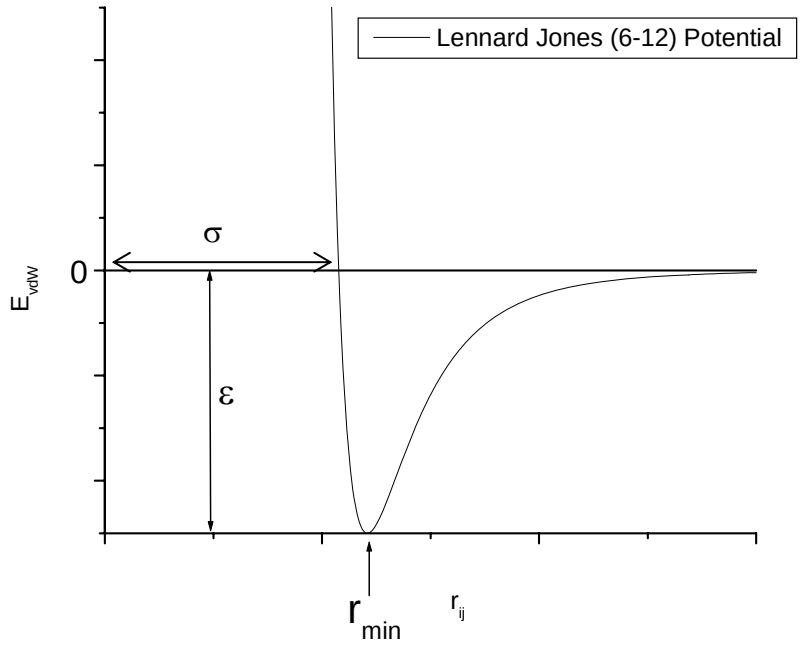
(wichtig hier nur: $E \sim r^{-6}$)

Van-der-Waals-Potentiale

- Verschiedene analytische Beschreibungen für die vdW-WW möglich
- Wichtige Formen sind
  - **Lennard-Jones-Potential**
  - **Born-Mayer-Potential**
- Unterschiede liegen in
  - Form und Breite des Potentialtopfs
  - Steilheit des asymptotischen Anstiegs/Abfalls

Lennard-Jones-Potential

- Nach **Sir John Lennard-Jones** (um 1920)
- Auch **6-12-Potential** genannt
- Repulsionsteil theoretisch nicht fundiert
- Dispersion: London-Potential



$$E_{\mathrm{vdW}} = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right]$$

Born-Mayer-Potential

- Um 1932 vorgeschlagen
- Auch Exp-6-Potential genannt
- Exponentieller Verlauf des Repulsionsterms trifft physikalische Realität besser als r^{-12}-Term

$$E_{BM} = Ae^{-Br} - \frac{C}{r^6}$$

- Probleme
  - Exponentialfunktion rechenaufwändig
  - $\lim_{r \to 0} E_{BM} = -\infty$

Energiefunktion für Argon

- Für Edelgase reines vdW-Potential ausreichend
- Beispiel Ar:

  $\varepsilon = 1.65 \cdot 10^{-21}$ J, $\sigma = 3.405 \cdot 10^{-10}$ m

  Damit lässt sich für zwei Ar-Atome mit beliebigem Abstand die Energie berechnen.
- Wie sieht die Energiefunktion aus für 20 Ar-Atome?
- vdW-Potential ist ein **paarweises, additives Potential**, d.h. für mehr als zwei Atome ist die Summe aller Paare *(i, j)* mit Abständen r_{ij} zu betrachten:

$$E_{\mathrm{vdW}} = \sum_i \sum_{j<i} 4\varepsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right]$$

Energiefunktion für Argon

$$E_{\mathrm{vdW}} = \sum_i \sum_{j<i} 4\varepsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right]$$

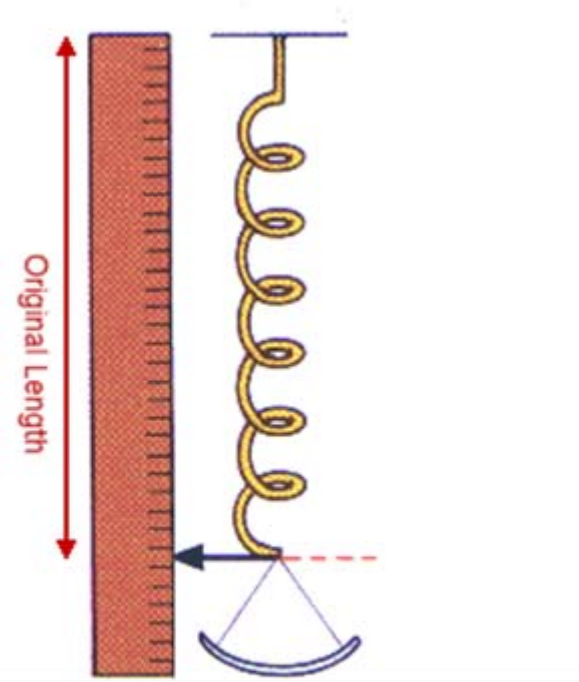
- Für Argon ist diese Beschreibung vollständig, da keine anderen WW zu betrachten sind
- Für jedes System aus „Molekülen“ (hier: Ar-Atomen) liefert uns diese Funktion eine Energie
- Parameter:
  - ε und σ sind vom Element abhängig
  - Üblicherweise rechnet man in molaren Einheiten, also $[\varepsilon]$ = 1 kJ/mol
  - Man zieht 4ε und σ häufig zu zwei Parametern *A* und *B* zusammen

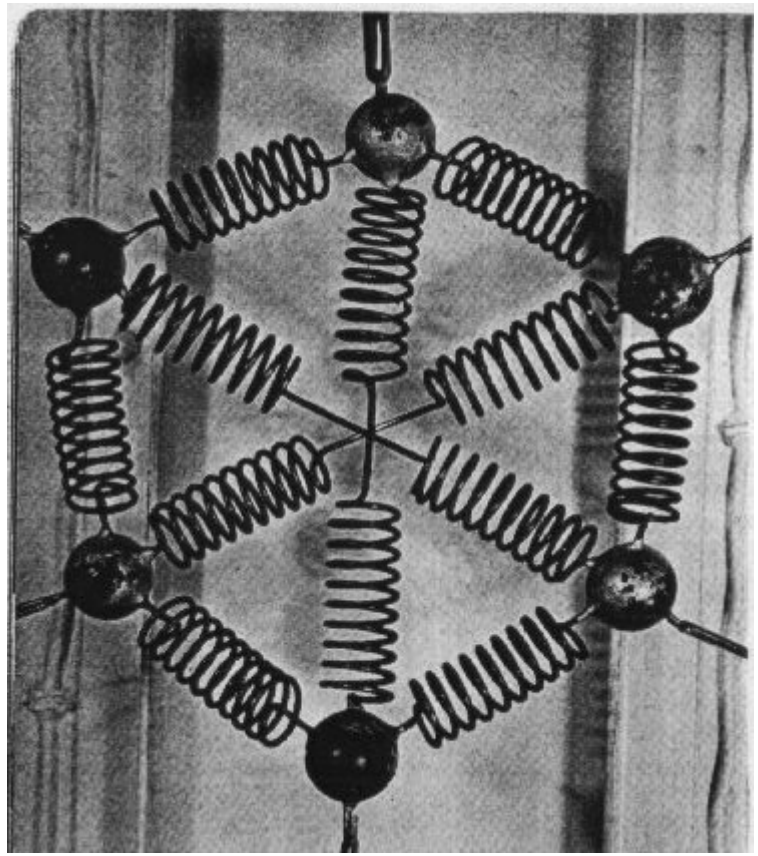
$$4\varepsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] = \frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}}$$

Bindungslänge *(stretches)*

- H – H, Abstand r = Bindungslänge
- Experimentelle Befunde zur Bindungslänge liegen vor
- Für einzelnes H_2-Molekül ist Bindungslänge r die einzige Koordinate $\Rightarrow E = f(r)$
- vdW spielt innerhalb des Moleküls keine Rolle
  - Repulsion ist größtenteils außer Kraft durch Überlappen der Bindungsorbitale
  - Bindungsenergie viel höher als die vdW-Energie

Klassische Mechanik



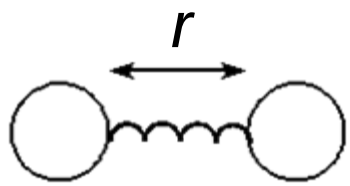


Hookesches Gesetz: $F = k\,\Delta s$

$\Rightarrow E = \frac{1}{2}\,k\,(\Delta s)^2$

Bindungslänge

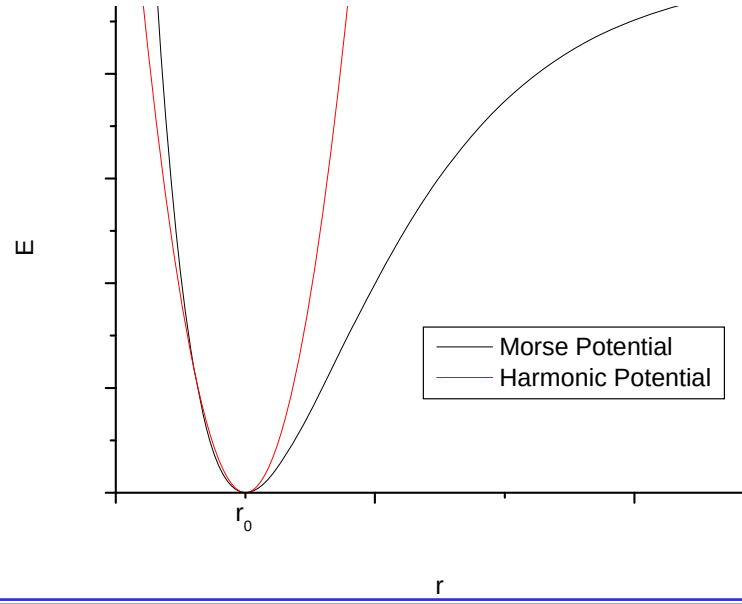
- Energie in Abhängigkeit von der Bindungslänge wird gut durch ein **Morse-Potential** beschrieben



$$E_{\text{Morse}} = A(1 - e^{-B(r-r_0)})^2$$

- Bindungslängen weichen nur sehr wenig vom Minimumsabstand r_0 ab
- Daher oft als **harmonisches Potential** angenähert

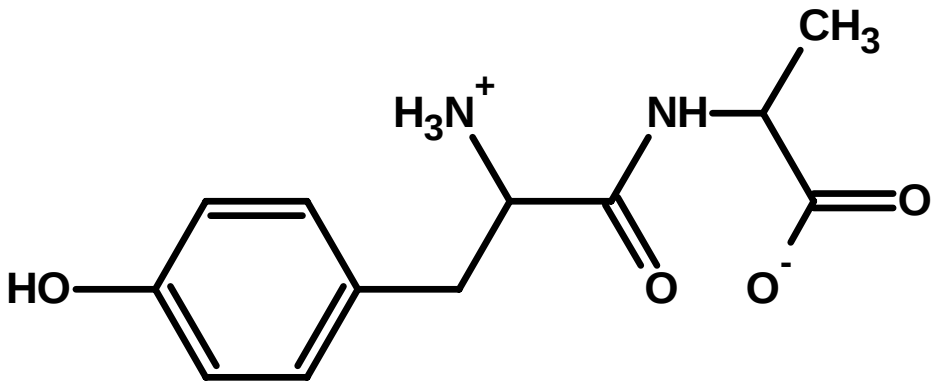
$$E_H = k_{IJ}(r_{ij} - r_0)^2$$



Arten von Wechselwirkungen

- WW werden häufig klassifiziert in
  - **Bindungsvermittelte** WW *(bonded interactions)*
  - **Nicht bindungsvermittelte** WW *(nonbonded)*
- vdW: nicht bindungsvermittelt
- Bindungslänge: bindungsvermittelt
- In System mit *N* Atomen gibt es in der Regel
  - O(*N*) Bindungen und bindungsvermittelte WW
  - O(N^2) nicht bindungsvermittelte WW

Weitere Arten von WW



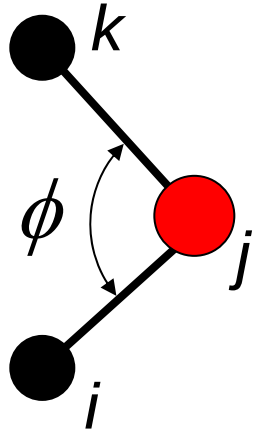
Für die Modellierung von Proteinen sind noch eine Reihe weiterer Wechselwirkungen unerlässlich

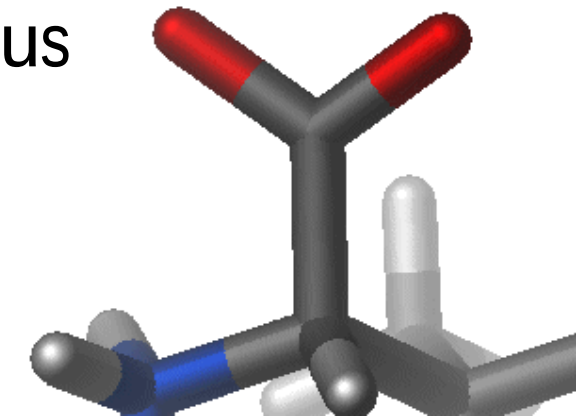
- Elektrostatik (Ladungen) *(nicht bindungsvermittelt)*
- Torsionen *(bindungsvermittelt)*
- Bindungswinkel *(bindungsvermittelt)*
- ....

Bindungswinkel *(bends)*

- Winkel ϕ zwischen zwei benachbarten Bindungen
- Wieder als **harmonisches Potential**:

  $E = k_{ijk} (\phi - \phi_0^{ijk})^2$
- Zwingt die Bindungen auf „natürliche“ Geometrien wie sie aus der Orbitaltheorie folgen
- Weniger steif als Bindungslänge

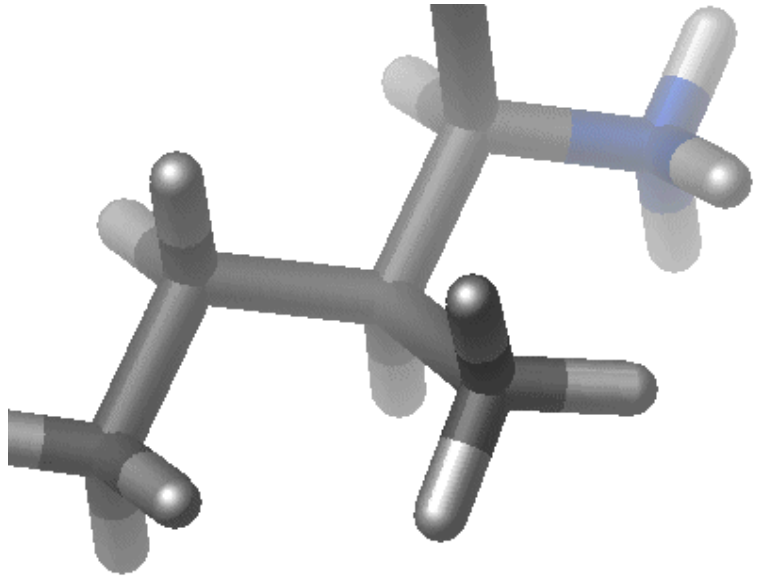


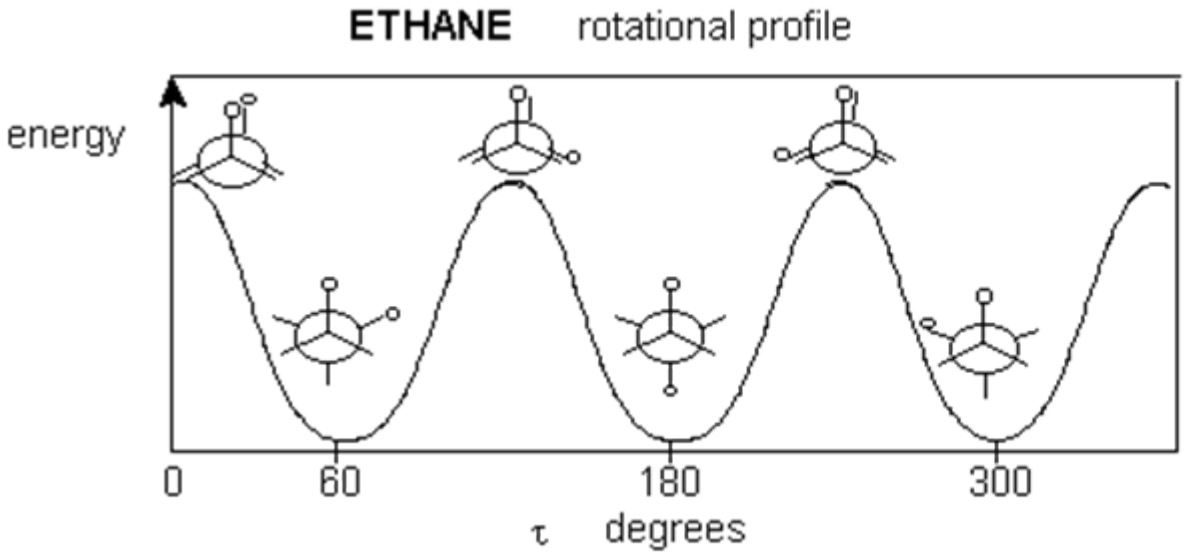


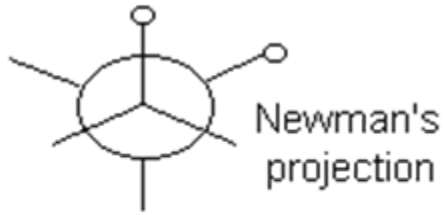
Torsionen *(dihedrals)*

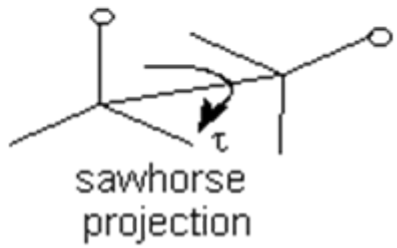
Torsionsenergie

- Beschreibt sowohl Einfach- als auch Mehrfachbindungen
- periodisch









Torsionen

Torsionsenergie

- Beschrieben durch
  - Torsionswinkel τ
  - Multiplizität *n*
- Energetisch weniger aufwändig als Bindungslänge und –winkel

  („weicher“ $\Rightarrow$ Flexibilität in Proteinen!)
- Mögliches Potential

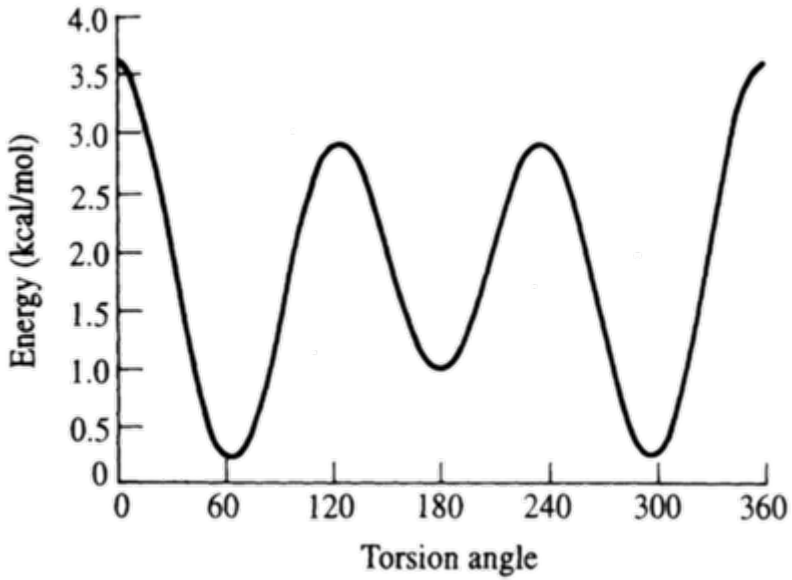
$$E_{\text{tors}} = k(1 + cos(n\tau - \tau_0))$$

- n, τ, k sind Konstanten und abhängig von den Atomtypen I, J, K , L

Torsionen

- Häufig müssen mehrere Torsionen zu einer kombiniert werden um die Energie korrekt zu beschreiben

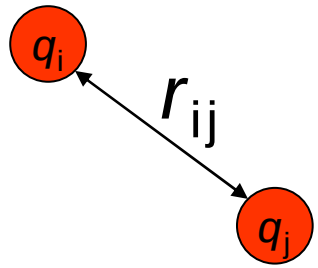
$$E_{\text{tors}} = \sum_{n=0}^{N} \frac{k_n}{2}(1 + cos(n\tau - \tau_0))$$



Elektrostatik – Monopol-Ansatz

- Klassischer Ansatz: **Coulombsches Gesetz**

$$E_{ES} = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$



(Mit der Permittivität des Vakuums ε_0)

- Ladungen liegen an Atompositionen (Punktladungen)
- Für ein System vieler Ladungen ergibt sich die Energie wieder als Summe der Potentiale

$$E_{ES} = \frac{1}{4\pi\varepsilon_0} \sum_i \sum_{j<i} \frac{q_i q_j}{r_{ij}}$$

Kraftfelder I

- Als **Kraftfeld** bezeichnet man eine konsistente Zusammenstellung mehrerer Wechselwirkungsterme
- Es gibt eine große Vielfalt von Kraftfeldern
  - Für unterschiedliche **Systeme**

    Wasser, kleine Moleküle, Proteine, DNA, Zucker, …
  - Zur Vorhersage unterschiedlicher **Eigenschaften**
    - Optimale Geometrien
    - Interne Energien
    - Wechselwirkungsenergien

Kraftfelder II

Ein Kraftfeld umfasst

- Eine **analytische Form**
- **Parametersätze** für die einzelnen WW
- **Regeln**, um die Parameter auf die passenden Atome eines Moleküls anzuwenden

Bsp.: Argon

Unser einfaches Kraftfeld bestand aus

- LJ-Potential
- Parametern für Ar (σ/ε)
- (triviale) Regel: verwende Parameter für Ar

ECEPP/2

Die ECEPP-Familie beruht auf starren Bindungslängen und –winkeln. Torsionen sind für die Flexibilität allein verantwortlich.

- Keine Bindungslängen- und –winkel-Terme
- vdW-WW: Lennard-Jones-Potential
- Elektrostatik: Coulomb
- Torsionen: Kosinus-Term
- Wasserstoffbrücken: 10-12-Potential

ECEPP/2

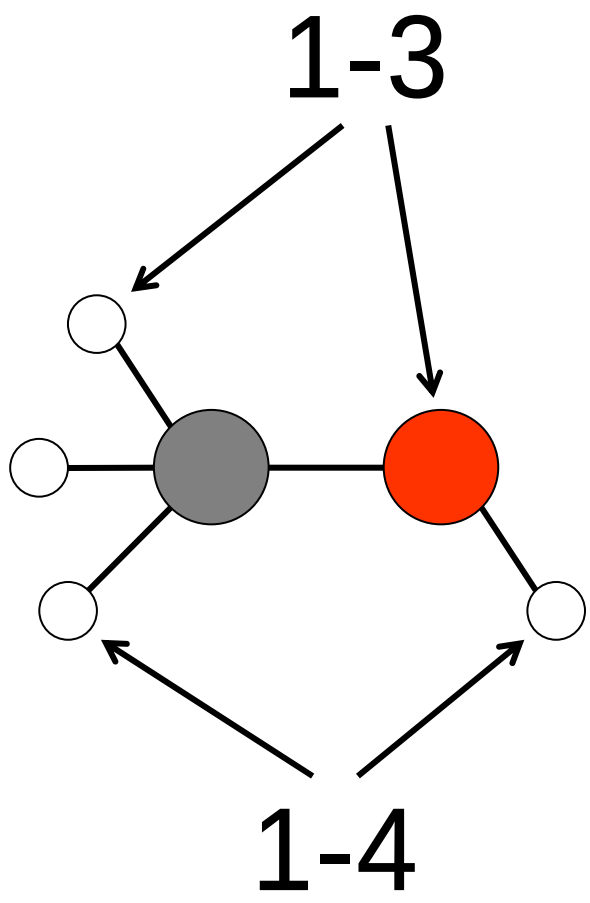
$$
\begin{aligned}
E &= E_{\mathrm{ES}} + E_{\mathrm{LJ}} + E_{\mathrm{HB}} + E_{\mathrm{tors}} \\
&= k_{\mathrm{ES}} \sum_{i,j<i} \frac{q_i q_j}{r_{ij}} + \sum_{i,j<i} F \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}} \\
&+ \sum_{(i,j)\in \mathrm{HB}} \frac{A'_{ij}}{r_{ij}^{12}} - \frac{B'_{ij}}{r_{ij}^{10}} \\
&+ \sum_{(i,j,k,l)\in tors} A_{ijkl}(1 \pm \cos n\tau_{ijkl})
\end{aligned}
$$

ECEPP/2 - vdW

$$E_{\mathrm{LJ}} = \sum_{i,j<i} F \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}}$$

- Gewöhnliches LJ-Potential
- *F* ist ein Faktor, der die Repulsion abschwächt

$$F = \begin{cases} 0.0 \text{ für 1-2-, 1-3-WW} \\ 0.5 \text{ für 1-4-WW} \\ 1.0 \text{ ansonsten} \end{cases}$$



ECEPP/2 – Wasserstoffbrücken

$$E_{\mathrm{HB}} = \sum_{(i,j)\in \mathrm{HB}} \frac{A'_{ij}}{r_{ij}^{12}} - \frac{B'_{ij}}{r_{ij}^{10}}$$

- H-Brücken werden erfasst durch
  - Elektrostatik
  - vdW
  - Zusatzpotential, das Differenzen zu Experiment ausgleicht
- 10-12-Potential ist für H-Brücken häufig verwendet, jedoch nicht fundiert
- Summe läuft über alle Paare aus H-Atome an Donoren und Akzeptoren

AMBER

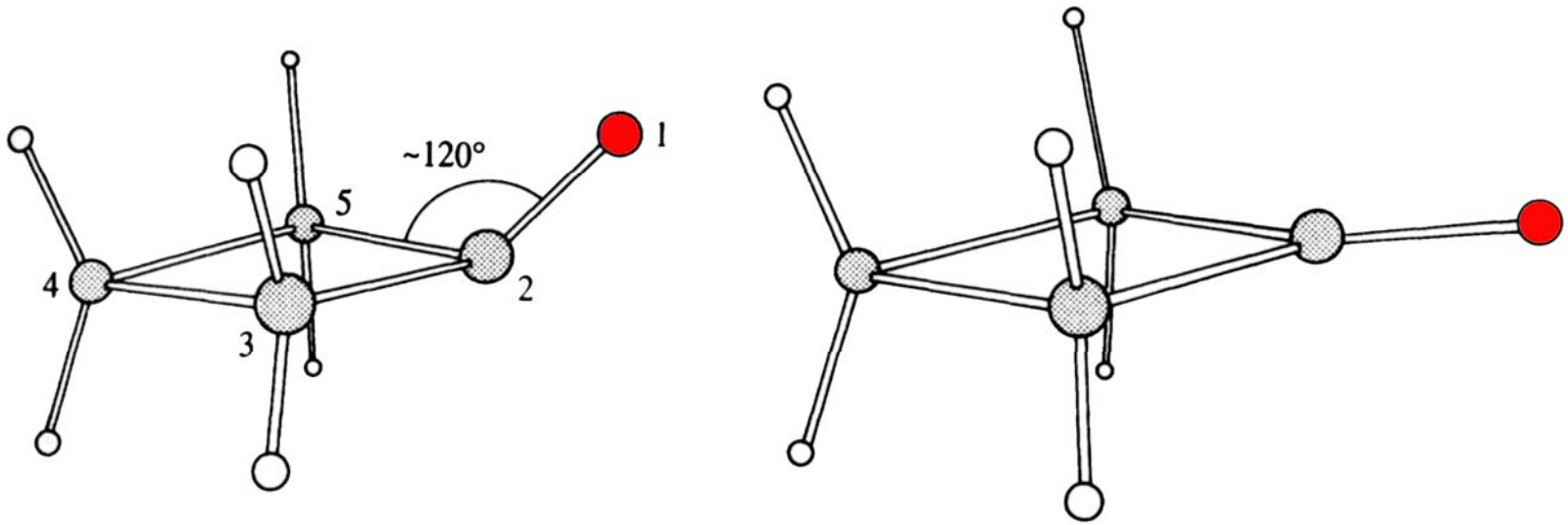
- **AMBER** – **A**ssisted **M**odel **B**uilding with **E**nergy **R**efinement
- An der UCSF in den 80ern entwickelt
- Fünf Beiträge
  - Bindungslänge
  - Bindungswinkel
  - Torsionen
  - Van-der-Waals
  - Elektrostatik
- Geeignet für Proteine und DNA
- Entwickelt zur Konstruktion von Modellen für XRD, später auch NMR
- Es gibt nicht *ein* AMBER-Kraftfeld, sondern eine ganze Familie, die üblicherweise mit dem Jahr der Publikation benannt werden (etwa AMBER89, AMBER94, AMBER99)

AMBER

$$
\begin{aligned}
E &= E_{\text{stretch}} + E_{\text{bend}} + E_{\text{tors}} + E_{\text{vdW}} + E_{\text{ES}} \\
&= \frac{1}{2} \sum_{(i,j)\in \text{B}} k_b^{IJ} (r_{ij} - r_0^{IJ})^2 \\
&+ \frac{1}{2} \sum_{(ijk)\in A} k_a^{IJK} (\phi_{ijk} - \phi_0^{IJK})^2 \\
&+ \frac{1}{2} \sum_{(ijkl)\in T} k_t^{IJKL} \left(1 + \cos(n^{IJKL}\tau - \tau_0^{IJKL})\right) \\
&+ \sum_{i<j} \left(\frac{A^{IJ}}{r_{ij}^{12}} - \frac{C^{IJ}}{r_{ij}^{6}} \right) + \frac{1}{4\pi\varepsilon_0} \sum_{i<j} \frac{q_i q_j}{r_{ij}}
\end{aligned}
$$

AMBER – Uneigentliche Torsionen

- AMBER enthält zusätzlich (im Torsionsterm versteckt) eine weitere Wechselwirkung: **uneigentliche Torsionen** *(improper torsions)*
- Zweck:
  - Atome in Ringebene halten (z.B. aromatische Systeme)



AMBER – Parameter

- AMBER 94 enthält
  - 83 Bindungslängen-Parameter
  - 191 Bindungswinkel-Parameter
  - 81 Torsions-Parameter
  - 31 uneigentliche Torsions-Parameter
  - 34 van-der-Waals-Parameter
  - 54 Atomtypen

AMBER – Atomtypen

- Zuordnung zu den Parametersätzen wird durch **Atomtypen** ermöglicht
- Jedem Atom wird ein Typ zugeordnet, der von seiner **chemischen Umgebung** abhängt (Hybridisierung, Bindungsnachbarn etc.)
- Parametersätze referenzieren nur noch Atomtypen
- AMBER96 kennt z.B. **54 Atomtypen**, davon 13 für Kohlenstoff und 12 für Wasserstoff
  - BR – Brom
  - C – sp^2-Kohlenstoff in Carbonylgruppe
  - CT – aliphatischer sp^3-Kohlenstoff
  - CM – sp^2-Kohlenstoff an Position 5 & 6 in Pyrimidinen
  - ....

AMBER – Parameter

Parameter-Dateien referenzieren nur noch Atomtypen

Beispiel: Bindungswinkel

```
CT-C -N     70.0       116.60    AA general
N -C -O     80.0       122.90    AA general
O -C -O     80.0       126.00    AA COO- terminal res.
O2-C -O2    80.0       126.00    AA GLU
[...]
```

Für gegebenen Bindungswinkel aus drei Atomen *(i,j,k)*

- Bestimme zugehörige Atomtypen *(I,J,K)*
- Lies aus obiger Tabelle Parameter k^{IJK} und $\phi_0{}^{IJK}$ aus

Atomtypen – Zuweisung

- Zuweisung der Atomtypen hängt von der Implementierung ab
- Zwei verbreitete Ansätze
  - **Vorgefertigte Tabellen**
    - Einfach bei Proteinen, DNA
    - Sehr schnell
    - Für alle anderen Moleküle nicht anwendbar
  - **Regelbasiert**
    - Aufwändiger
    - Für alle parametrisierten Strukturen anwendbar
    - Gefahr der Fehlzuweisung

Atomtyp-Zuweisung – Tabellen

Beispiel: Zuweisung der Atomtypen (und Ladungen) für Ala aus der AMBER-Implementierung von BALL

```
[ChargesAndTypeNames]
ver:version key:name value:q value:type
@unit_q=e0
   1.0 ALA:N      -0.41570 N
   1.0 ALA:H       0.27190 H
   1.0 ALA:CA      0.03370 CT
   1.0 ALA:HA      0.08230 H1
   1.0 ALA:CB     -0.18250 CT
   1.0 ALA:1HB     0.06030 HC
   1.0 ALA:2HB     0.06030 HC
   1.0 ALA:3HB     0.06030 HC
   1.0 ALA:C       0.59730 C
   1.0 ALA:O      -0.56790 O
```

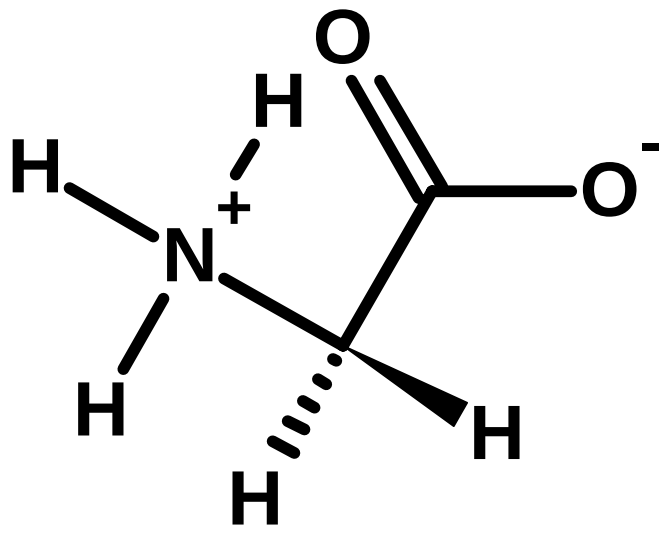
Atomtyp-Zuweisung – Regeln

Beispiel:

Regelsatz für die Zuweisung einiger AMBER-Typen aus der AMBER-Implementierung von BALL

```
; any sp3 (four explicit substituents)
CT = sp3Hybridized()
; any sp2 carbonyl carbon
C  = sp2Hybridized() AND connectedTo(=O)
; sp2 aromatic carbon in a five-membered
; ring next to two carbons (e.g.
; C_gamma in tryptophan)
C* = sp2Hybridized() AND inRing(5) AND
     connectedTo((~C)(~C))
; any sp2 aromatic carbon...
CA = sp2Hybridized() AND connectedTo((~*)(~*))
```

Beispiel: Glycin mit AMBER



Zuweisung der Atomtypen

O2	O in Carboxylgruppen
C	C in Carbonyl- und Carboxylgruppen
N3	N in Ammoniumgruppen
H	H in Aminen, Amiden, Ammonium
CT	aliphatische sp^3-C
H1	aliphatische H

Beispiel: Glycin mit AMBER

Torsionen

15 Torsionswinkel

- 4 x H1-CT-C-O2
- 2 x N3-CT-C-O2
- 6 x H-N3-CT-H1
- 3 x H-N3-CT-C

Bindungslänge

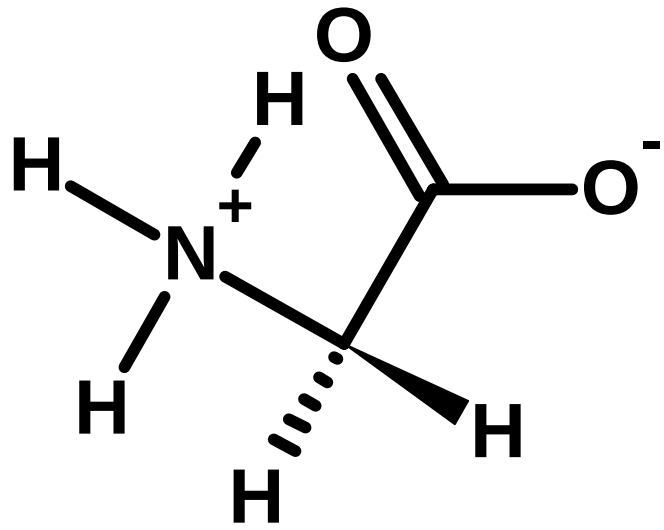
9 Bindungen

- 2 x C-O2
- 1 x C-CT
- 2 x CT-H1
- 1 x CT-N3
- 3 x N3-H

Bindungswinkel

14 Winkel

- 1 x O2-C-O2
- 2 x CT-C-O2
- 2 x H1-CT-C
- 2 x H1-CT-N3
- 1 x C-CT-N3
- 3 x H-N3-CT
- 3 x H-N3-H



Beispiel: Glycin mit AMBER

vdW

16 1-4-Paare

- 4 x H1-O2
- 3 x N3-C
- 6 x H-H1
- 3 x H-C

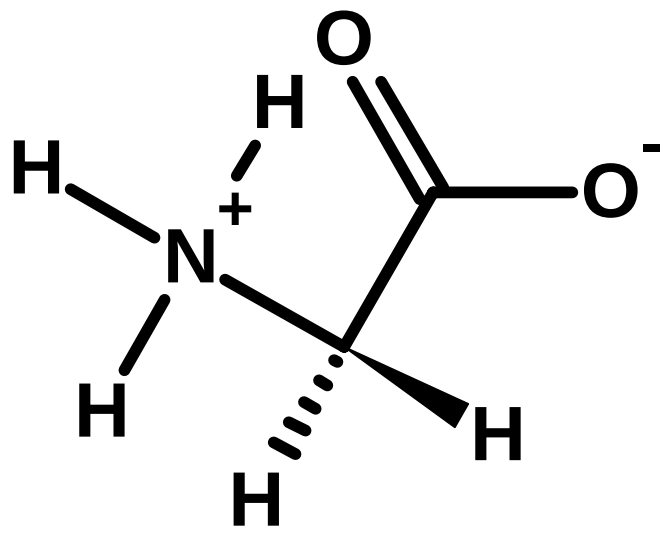
6 Gewöhnliche Paare

- 6 x H-O2

Elektrostatik

Wie vdW…

Ladungen aus QM-Rechnung oder Tabelle



Beispiel: Glycin mit AMBER

- Typ- und Parameterzuweisung ist programmtechnisch recht aufwändig und oft unterschätzt
- Alleine für Glycin müssen zugewiesen werden
  - 6 verschiedene Atomtypen für 10 Atome
  - 10 verschiedene Stretch-Parameter für 9 Bindungen
  - 14 verschiedene Bend-Parameter für 14 Winkel
  - 4 verschiedene Parametersätze für 15 Torsionen
  - 10 verschiedene LJ-Parameter für 22 NB-Paare

Kraftfeldklassen

Klassifizierung nach bindungsvermittelten WW

- **Klasse I** (AMBER, CHARMM, GROMOS)
  - Nur harmonische Terme, keine Kreuzterme
  - Reproduzieren optimale Geometrien recht gut
- **Klasse II** (MM2, MM3, MMFF94)
  - Terme dritten und höheren Grades, Kreuzterme
  - Vorhersage von spektroskopischen Eigenschaften
- **Klasse III**
  - Zusätzlich Integration chemischer Eigenschaften und Effekte (Elektronegativität, Hyperkonjugation, …)

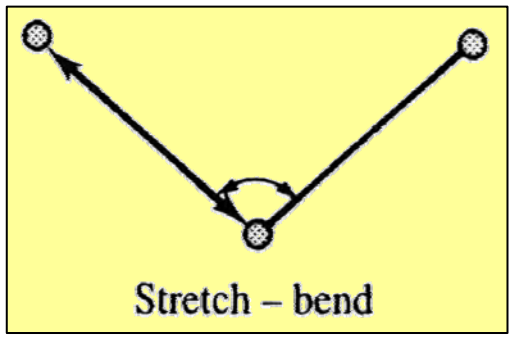
Klasse-II-Kraftfelder

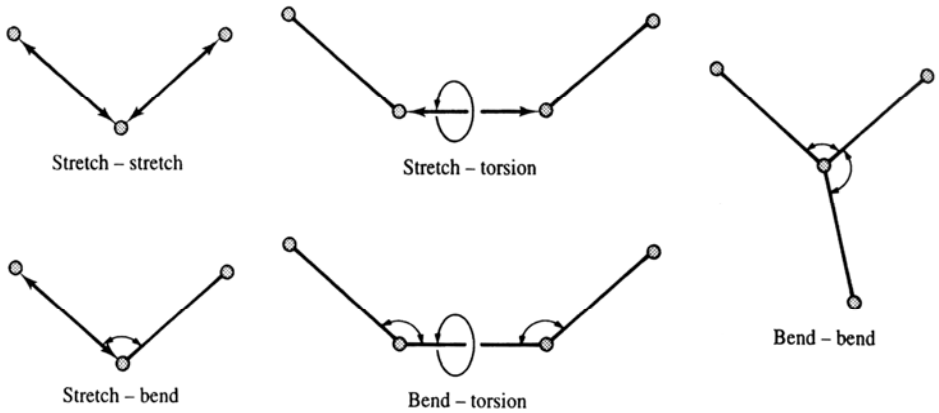
- **Anharmonische Potentiale** beschreiben die Bindungsstreckungen und –biegungen wesentlich besser als harmonische (vgl. Morse-Potential)
- **Klasse-II-Kraftfelder** verwenden dazu häufig direkt das Morse-Potential oder Polynome höheren Grades (3., 4.)
- Anharmonische Potentiale geben die Vibrationen, die man z.B. mit spektroskopischen Methoden untersucht, wesentlich besser wieder

$$
\begin{aligned}
E = {} & S_b\{\sum_b [{}^2K_b(b-b_0)^2 + {}^3K_b(b-b_0)^3 + {}^4K_b(b-b_0)^4]\} \\
& + S_\theta\{\sum_\theta [{}^2K_\theta(\theta-\theta_0)^2 + {}^3K_\theta(\theta-\theta_0)^3 + {}^4K_\theta(\theta-\theta_0)^4]\} \\
& + S_\phi\{\sum_\phi [{}^1K_\phi(1-\cos\phi) + {}^2K_\phi(1-\cos 2\phi) + \\
& \quad {}^3K_\phi(1-\cos 3\phi)]\} \\
& + S_\chi\{\sum_\chi K_\chi\chi^2\} + \sum_{i>j}\frac{q_iq_j}{r_{ij}} + \sum_{i>j}\epsilon\left[2\left(\frac{r^*}{r_{ij}}\right)^9 - 3\left(\frac{r^*}{r_{ij}}\right)^6\right] \\
& + S_c\{\sum_b\sum_{b'} K_{bb'}(b-b_0)(b'-b'_0) + \sum_\theta\sum_{\theta'} K_{\theta\theta'} \times \\
& \quad (\theta-\theta_0)(\theta'-\theta'_0) \\
& + \sum_b\sum_\theta K_{b\theta}(b-b_0)(\theta-\theta_0) \\
& + \sum_\phi\sum_b (b-b_0)[{}^1K_{\phi b}\cos\phi + {}^2K_{\phi b}\cos 2\phi + {}^3K_{\phi b}\cos 3\phi] \\
& + \sum_\phi\sum_{b'} (b'-b'_0)[{}^1K_{\phi b'}\cos\phi + {}^2K_{\phi b'}\cos 2\phi + \\
& \quad {}^3K_{\phi b}\cos 3\phi] \\
& + \sum_\phi\sum_\theta (\theta-\theta_0)[{}^1K_{\phi\theta}\cos\phi + {}^2K_{\phi\theta}\cos 2\phi + {}^3K_{\phi\theta}\cos 3\phi] \\
& + \sum_\phi\sum_\theta\sum_{\theta'} K_{\phi\theta\theta'}(\theta-\theta_0)(\theta'-\theta'_0)\cos\phi\} \qquad (2)
\end{aligned}
$$

Klasse-II-Kraftfelder

- **Kreuzterme** beschreiben die Kopplung zweier Terme, z.B. der Bindungslänge an die Bindungswinkel

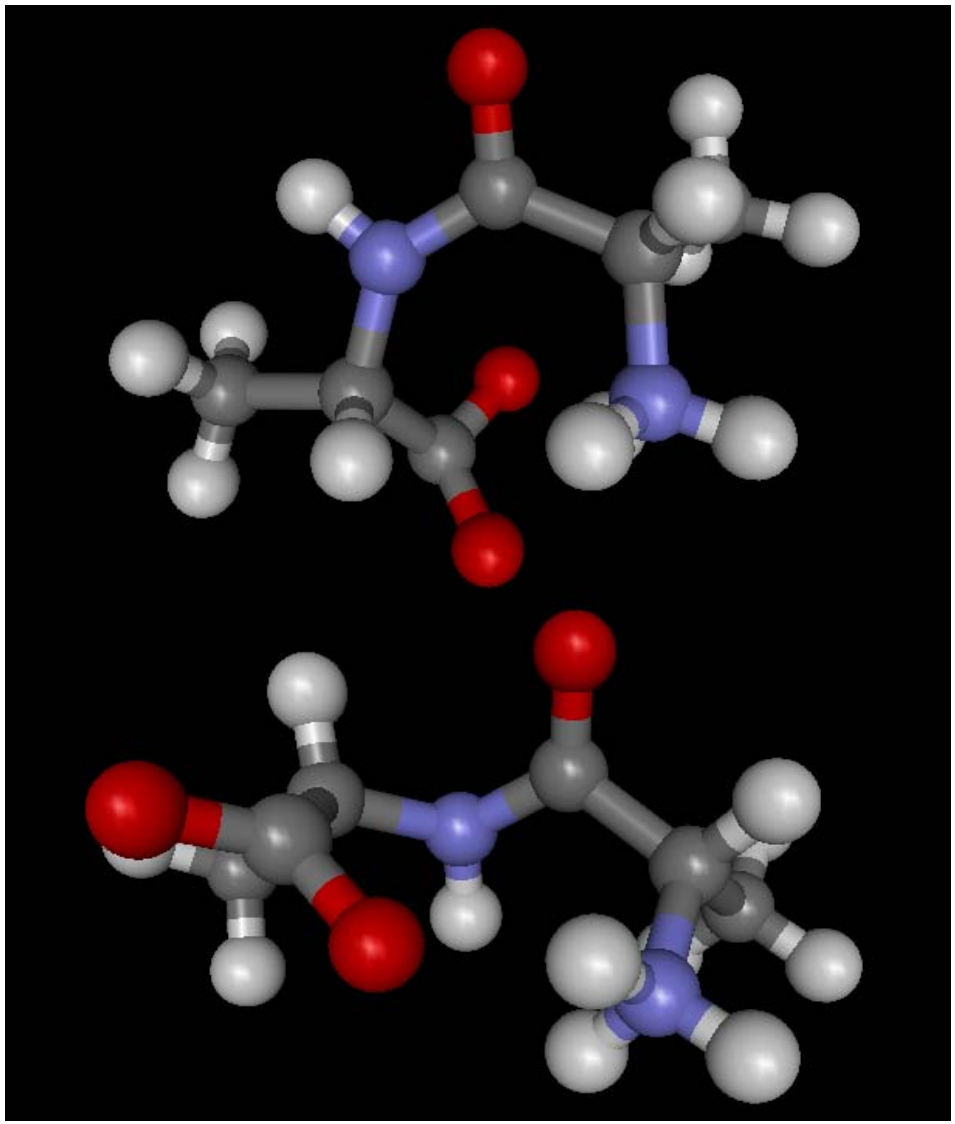




$$
\begin{aligned}
E = {} & S_b\{\sum_b [{}^2K_b(b-b_0)^2 + {}^3K_b(b-b_0)^3 + {}^4K_b(b-b_0)^4]\} \\
& + S_\theta\{\sum_\theta [{}^2K_\theta(\theta-\theta_0)^2 + {}^3K_\theta(\theta-\theta_0)^3 + {}^4K_\theta(\theta-\theta_0)^4]\} \\
& + S_\phi\{\sum_\phi [{}^1K_\phi(1-\cos\phi) + {}^2K_\phi(1-\cos 2\phi) + \\
& \quad {}^3K_\phi(1-\cos 3\phi)]\} \\
& + S_\chi\{\sum_\chi K_\chi\chi^2\} + \sum_{i>j}\frac{q_iq_j}{r_{ij}} + \sum_{i>j}\epsilon\left[2\left(\frac{r^*}{r_{ij}}\right)^9 - 3\left(\frac{r^*}{r_{ij}}\right)^6\right] \\
& + S_c\{\sum_b\sum_{b'} K_{bb'}(b-b_0)(b'-b'_0) + \sum_\theta\sum_{\theta'} K_{\theta\theta'} \times \\
& \quad (\theta-\theta_0)(\theta'-\theta'_0) \\
& + \sum_b\sum_\theta K_{b\theta}(b-b_0)(\theta-\theta_0) \\
& + \sum_\phi\sum_b (b-b_0)[{}^1K_{\phi b}\cos\phi + {}^2K_{\phi b}\cos 2\phi + {}^3K_{\phi b}\cos 3\phi] \\
& + \sum_\phi\sum_{b'} (b'-b'_0)[{}^1K_{\phi b'}\cos\phi + {}^2K_{\phi b'}\cos 2\phi + \\
& \quad {}^3K_{\phi b}\cos 3\phi] \\
& + \sum_\phi\sum_\theta (\theta-\theta_0)[{}^1K_{\phi\theta}\cos\phi + {}^2K_{\phi\theta}\cos 2\phi + {}^3K_{\phi\theta}\cos 3\phi] \\
& + \sum_\phi\sum_\theta\sum_{\theta'} K_{\phi\theta\theta'}(\theta-\theta_0)(\theta'-\theta'_0)\cos\phi\} \qquad (2)
\end{aligned}
$$

(4)

Konformationsenergien



Energien mit AMBER96

	cis [kJ/mol]	trans [kJ/mol]
Bindungen	3.4	1.8
Winkel	62.1	31.5
Torsionen	27.6	44.7
ES	-488.2	-488.2
vdW	37.58	31.8
Gesamt	**-357.5**	**-379.0**

Überblick Kraftfelder

Kraftfeld	# Atom-Typen	vdW	ES	HB	stretch	bend	cross	Domäne
ECEPP/2	21	6-12	MP	10-12	-	-	-	Proteine
TRIPOS	31	6-12	MP	-	P2	P2	-	Allgemein
AMBER	54	6-12	MP	-	P2	P2	-	Proteine, DNA
CHARMM	29	6-12	MP	-	P2	P2	-	Proteine, DNA
MM2	71	Exp-6	DP	-	P3	P6	sb	Allgemein
MM3	155	Exp-6	DP	-	P4	P6	sb, bb, st	Allgemein
MMFF94	99	7-14	MP	7-14	P4	P3	sb	Allgemein

DP – Dipol-Dipol-WW, MP – Monopol (Coulomb),
sb – stretch/bend, bb – bend/bend, st – stretch/torsion

Parametrisierung

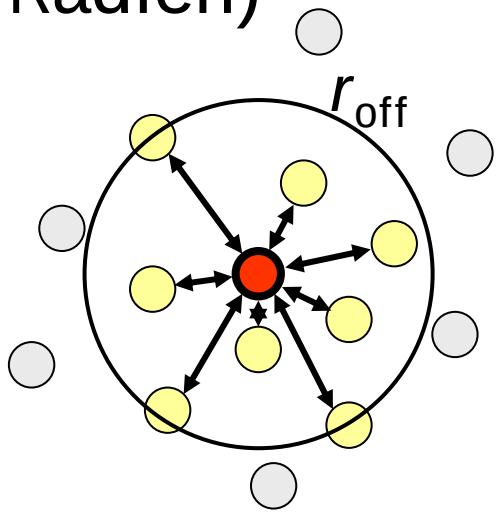
- **Kalibrierung an QM-Rechnungen**
  - Auch exotische Geometrien zugänglich
  - Resultat hängt von Qualität der QM-Methode ab
  - Kritisch: Wahl von Modellverbindungen
  - Relativ unaufwändig
- **Kalibrierung an experimentellen Daten**
  - Gute experimentelle Daten rar
  - Nur natürlich vorliegende Geometrien zugänglich
  - Geeignet
    - Geometrien aus XRD, Neutronenstreuung
    - Spektroskopisch ermittelte Kraftkonstanten

Implementierung

- Häufig sehr viele Berechnungen nötig (z.B. Simulationen mit > 10^8 Schritte)
- Geschwindigkeit essenziell
- Komplexität (*N* = Anzahl Atome)
  - Bindungsvermittelte WW generell O(N)
  - vdW, Elektrostatik generell O(N^2)
- Beschleunigung z.B. durch
  - Effiziente Implementierung
    - Assembler-Code
    - Vorhalten der Energiefunktionen in Tabellen
    - ….
  - Geschickte Näherungen
    - **Abschneideradien**
    - (Multipol-Methoden)
    - …

Abschneideradien

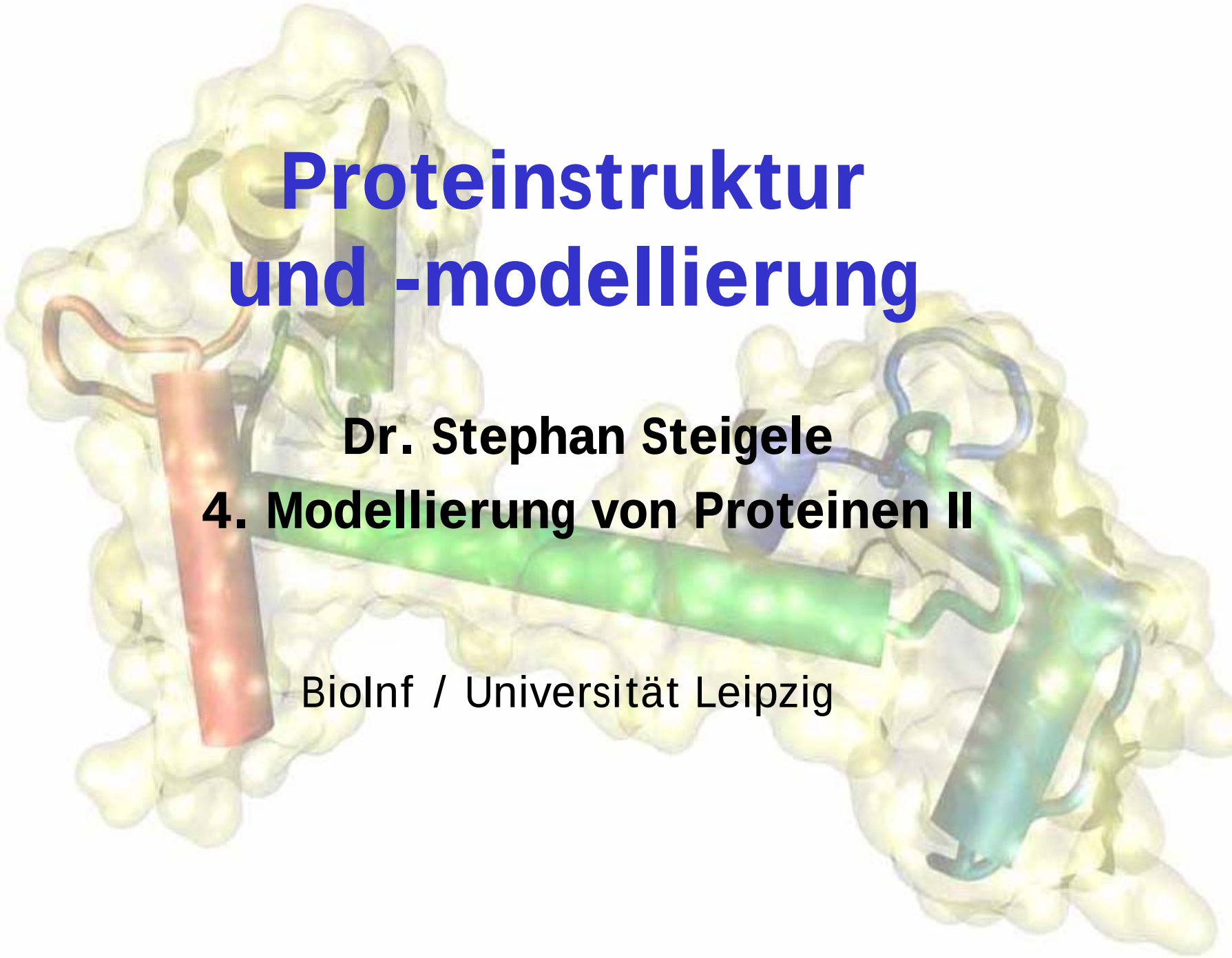
- O(N^2) viele Paare für vdW und ES benötigen meist die Hauptrechenzeit
- Viele Paare tragen dabei nichts zur Energie bei, da Beiträge für $r > 8 - 9$ Å nahe Null
- Lösung: **Abschneideradien** (cut-off-Radien)
  - Berechne nur Paare für $r < r_{off}$
  - Es gibt nur O(N) solcher Paare
  - Halte die Paare in Paarliste
  - Berechne die Paarliste regelmäßig neu



Literatur

Molekülmechanik

- **Andrew R. Leach, Molecular Modelling – Principles and Applications**, Prentice Hall, 2001
- **Anthony J. Stone, The Theory of Intermolecular Forces**, Clarendon Press, 1996
- **Daan Frenkel, Berend Smit, Understanding Molecular Simulation**, Academic Press, 1996
- **Martin J. Field, A practical introduction to the simulation of molecular systems**, Cambridge University Press, 1999
- **Tamar Schlick, Molecular Modeling and Simulation**, Springer, 2003
- **Ulrich Burkert, Norman L. Allinger, Molecular Mechanics**, American Chemical Society, 1982



Proteinstruktur und -modellierung

Dr. Stephan Steigele

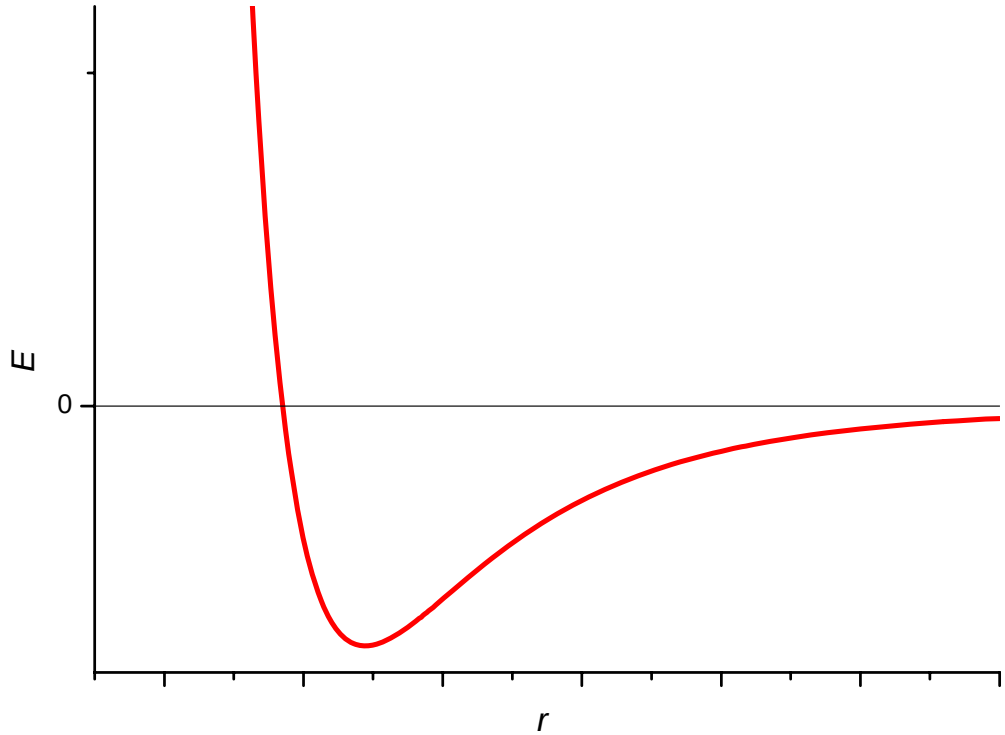
4. Modellierung von Proteinen II

BioInf / Universität Leipzig

Gliederung

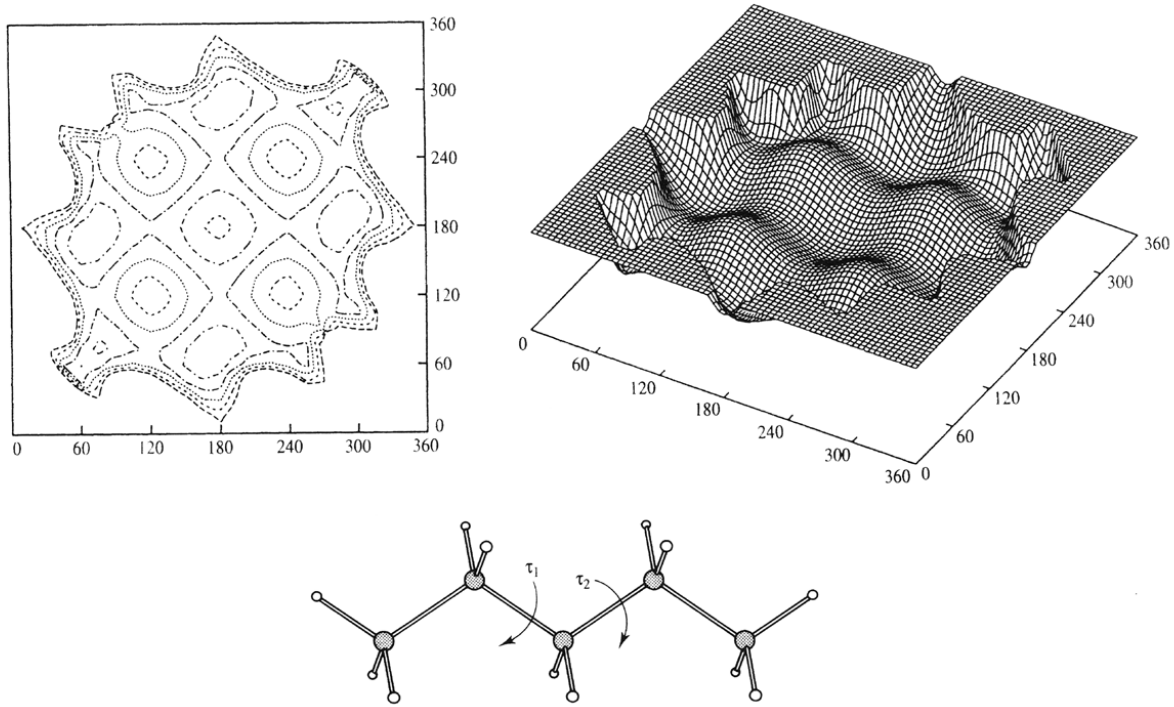
- Energieminimierung
  - Energiehyperflächen
  - Algorithmen
    - Grundlagen der Geometrieoptimierung
    - Liniensuche
    - Steilster Abstieg
    - Konjugierte Gradienten
  - Anwendungen

Energiehyperflächen



- Jedem Punkt des 3*N*-dimensionalen Konformationsraum ist eine Energie zugeordnet
- Energie beschreibt **Hyperfläche** in diesem Raum
- 1D (Ar—Ar): Kurve in der Ebene

Energiehyperflächen

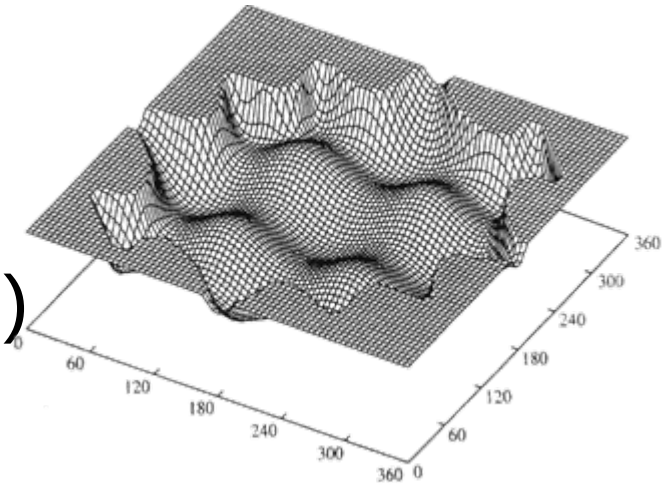


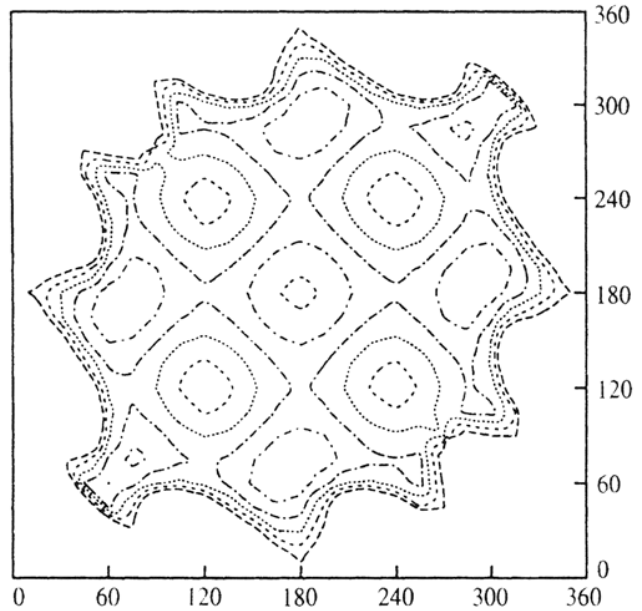
- 2D (2 Torsionswinkel) – Fläche in 3D
- Energie beschreibt eine Hyperfläche im (3N+1)-dimensionalen Raum
- **Energiehyperfläche = PES** (*Potential Energy Surface*)

Wofür ist das alles interessant?

- **Minima**
  - Entsprechen günstigen Konformationen (Konformeren)
  - Häufig **lokale** Minima!
- **Globales** Minimum?

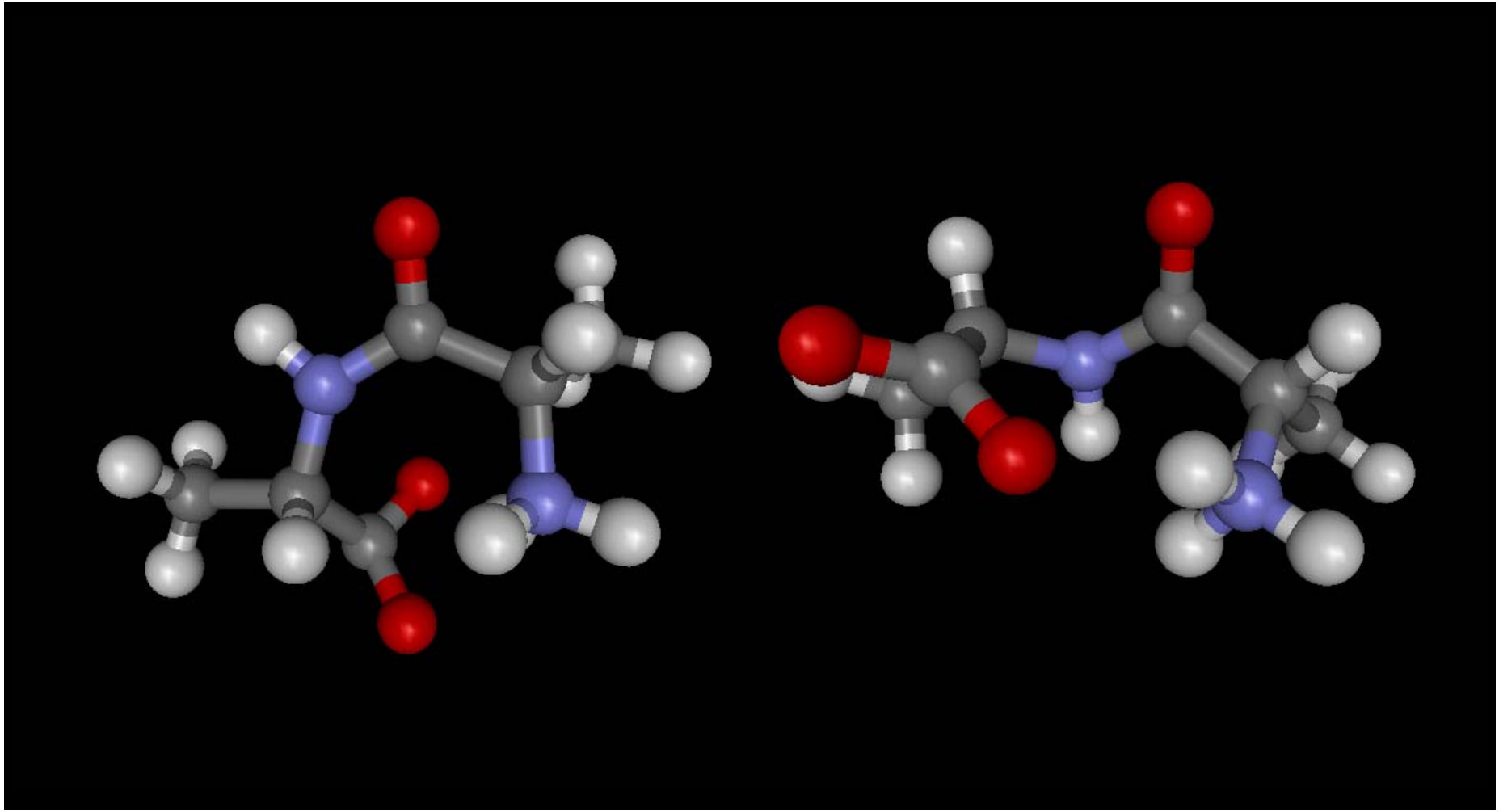
  (Bsp.: Proteinfaltung!)
- Kann man die Oberfläche systematisch durchmustern?
- Mittelung über Ensembles





Leach, Molecular Modelling, S. 254

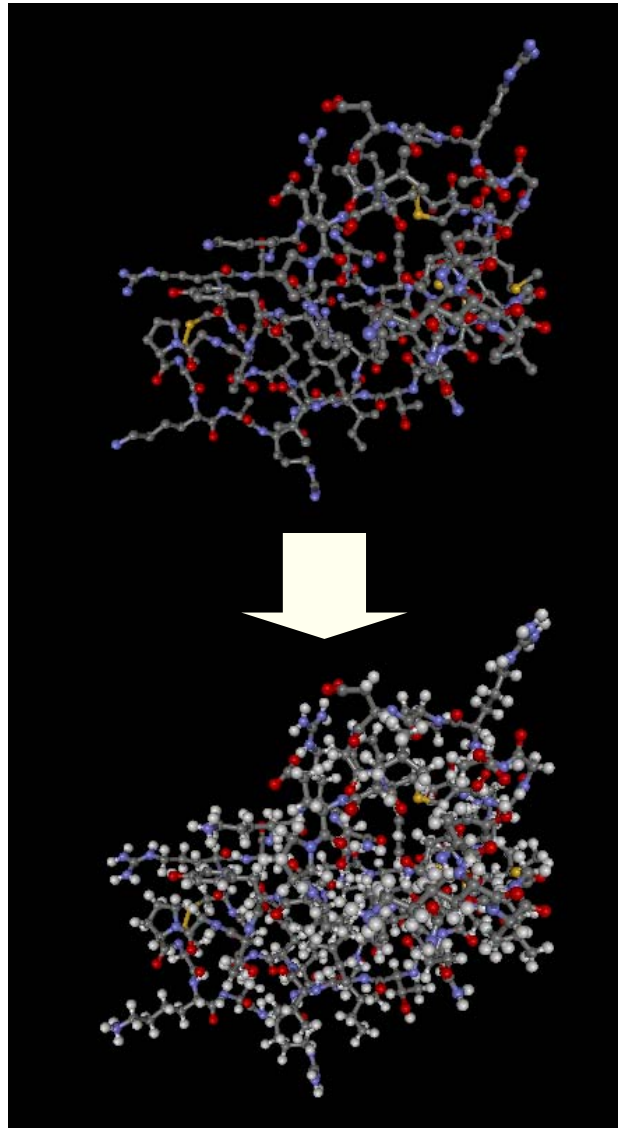
Suche nach Minima



Bsp. Ala_2: Gibt es günstigere Konformationen?

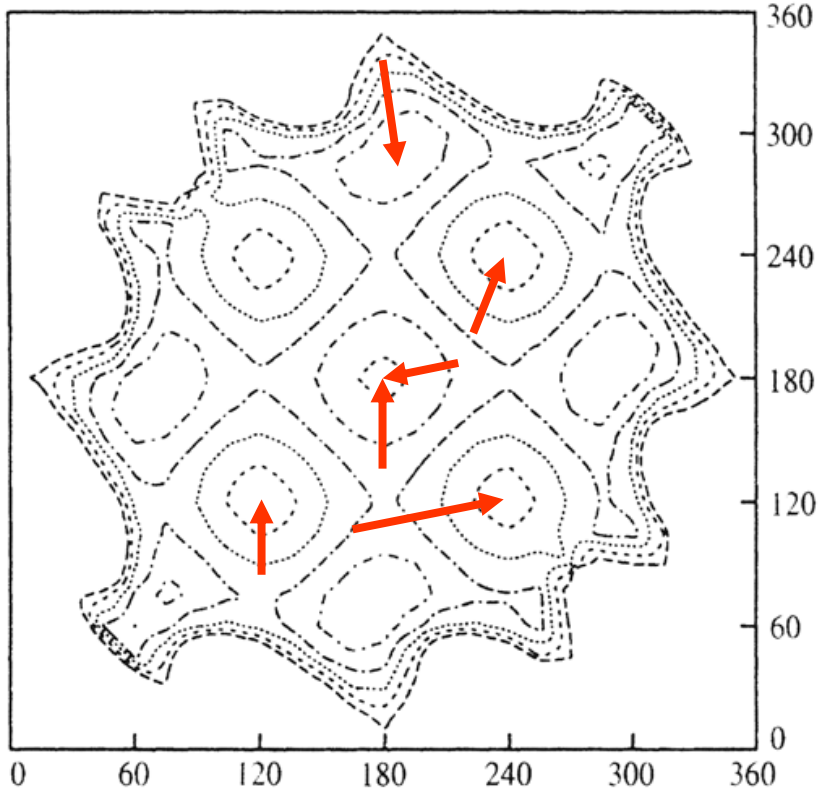
Strukturverfeinerung

- XRD zeigt keine H-Atome
- Hinzufügen der H-Atome liefert Positionen in denen Atome oft überlappen
- Diese Überlappungen sind physikalisch nicht möglich
- Kann man die H-Atome so positionieren, dass die Energie minimal ist?



Lokale und globale Minima

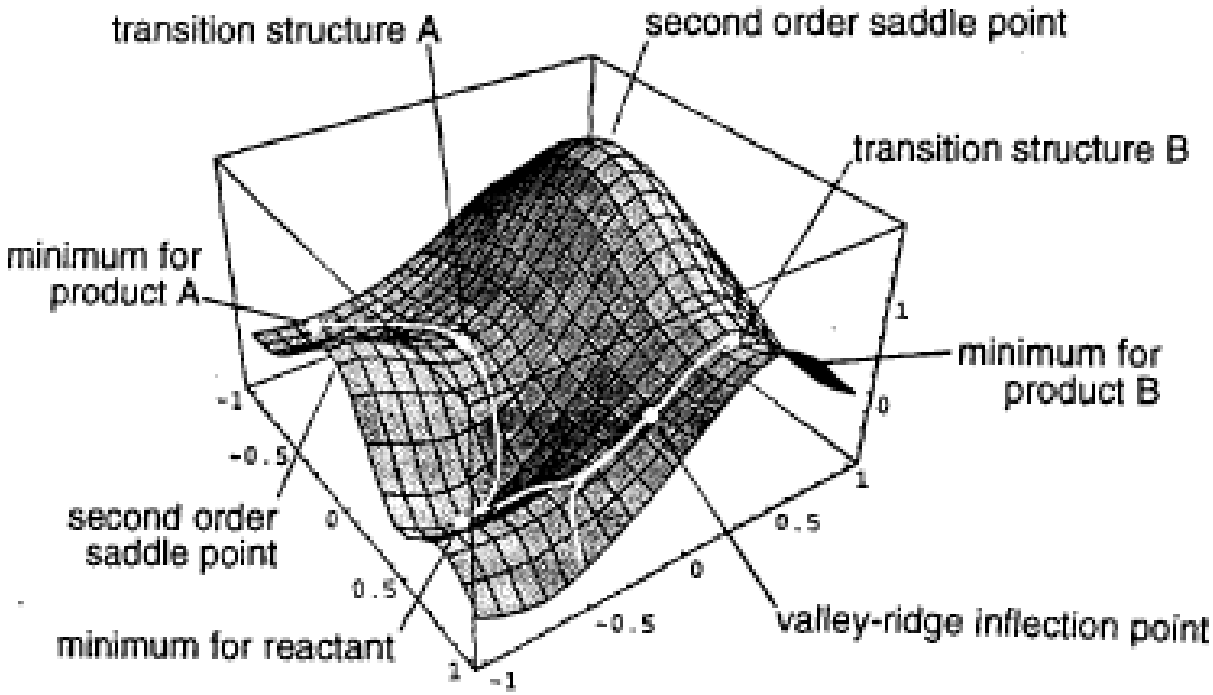
- Lokale Minima finden ist einfacher als das globale
- Viele Algorithmen finden das nächste lokale Minimum
- Für viele Fragen sind lokale Minima ausreichend
- Verschiedene Startpositionen können zu verschiedenen Minima führen



PES und ihre Ableitungen

Gradient $\mathrm{r} = (\partial/\partial x_0, \partial/\partial x_1, \ldots)$

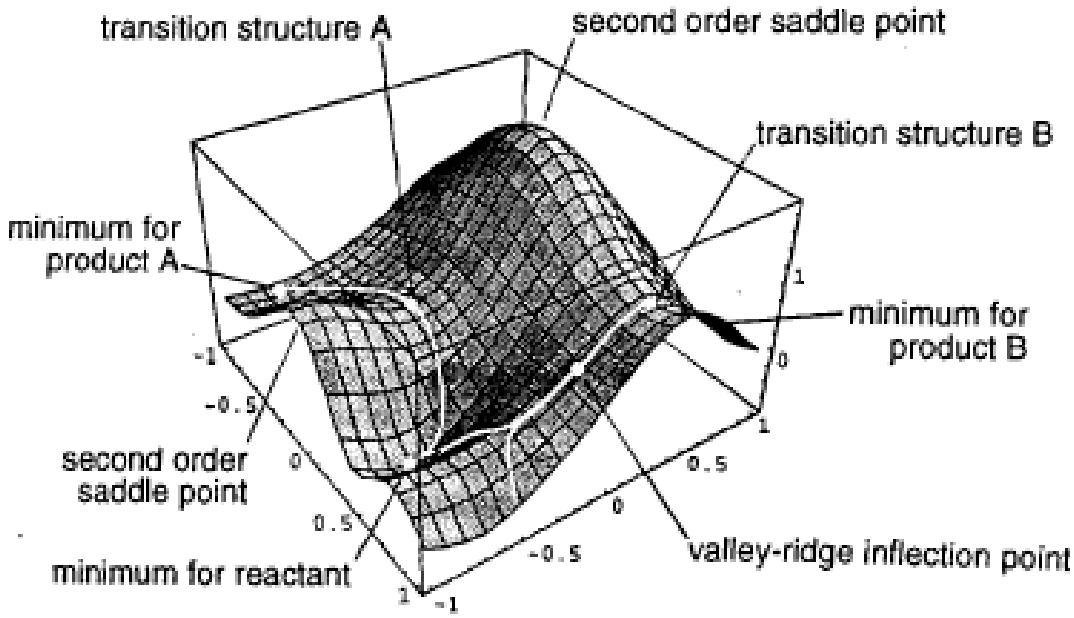
- rf ist die Richtung in der f am stärksten ansteigt
- rf = 0 an Minima, Maxima, Sattelpunkten



PES und ihre Ableitungen

Hesse-Matrix **H**(**x**), Matrix der zweiten partiellen Ableitungen: $\mathbf{H}_{ij} = \partial^2 / (\partial x_i \, \partial x_j)$

- **H**(**x**) beschreibt Krümmung von *f* in **x**
  - konvex, falls **H** positiv definit
  - konkav, falls **H** negativ definit



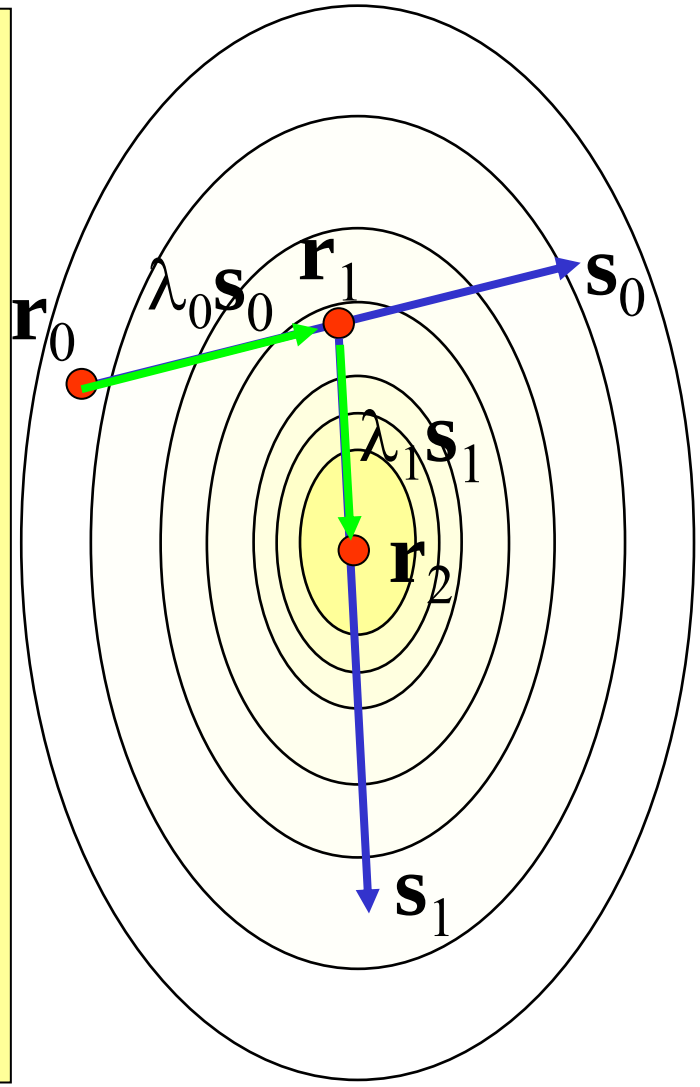
Minimierung – Basisalgorithmus

Für k = 0 bis Konvergenz

- Wähle Suchrichtung $\mathbf{s}_k$
- Wähle Schritt λ_k entlang $\mathbf{s}_k$, der die Energie verringert
- Aktualisiere Koordinaten:

$$\mathbf{r}_{k+1} = \mathbf{r}_k + \lambda_k \mathbf{s}_k$$

- $k = k + 1$, nächster Schritt



Minimierung – Grundsätze

- **s** sei ein Einheitsvektor, **r** die Atompositionen
- Gradient bezeichnen wir mit $\mathbf{g} \equiv \nabla E(\mathbf{r})$
- Gradient zeigt in Richtung des stärksten Anstiegs
- Suchrichtung sollte **„bergab“** führen

) entgegen des Gradienten

Satz:

Für ein beliebiges **s** mit negativer Komponente entlang **g** existiert ein positives λ mit

$$E(\mathbf{r} + \lambda\mathbf{s}) < E(\mathbf{r}).$$

Konvergenz

- Zur Abschätzung ob man im Minimum angelangt ist, berechnet man den **RMS-Gradienten** (RMS = *root mean square*)

$$RMS(\mathbf{g}) = \frac{1}{\sqrt{3N}}|\mathbf{g}| = \frac{1}{\sqrt{3N}}\sqrt{\sum g_i^2}$$

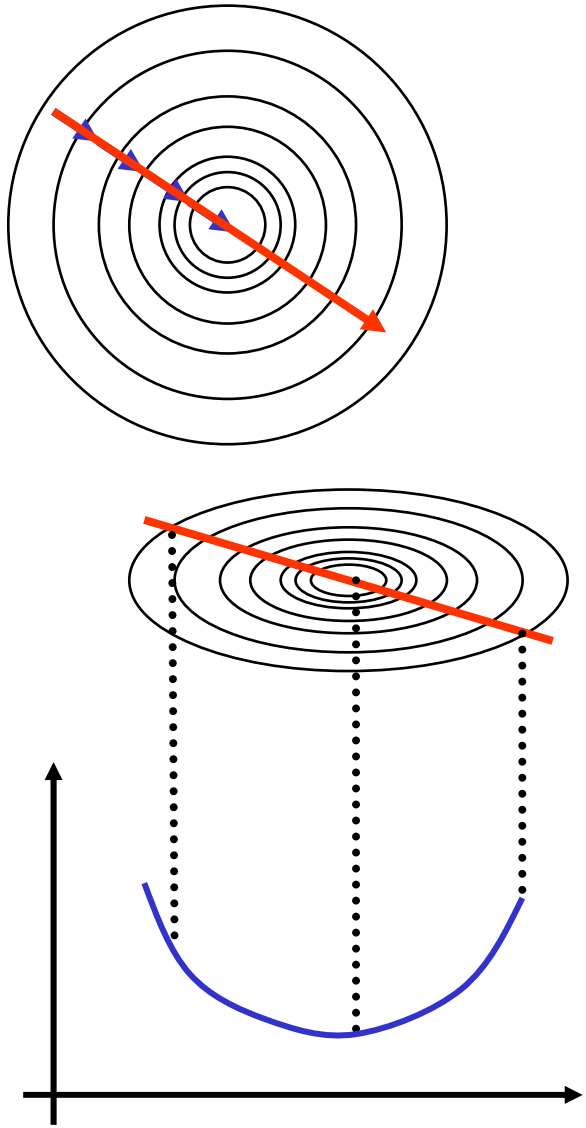
- Erreicht die Suche ein Minimum, wird $\mathbf{g} = 0$
- Üblicherweise gibt man einen Schwellenwert RMS_{max} an
- Konvergenz ist erreicht bei $RMS(\mathbf{g}_k) \cdot RMS_{min}$
- Je nach Anwendung wählt man RMS_{min} unterschiedlich streng

Steilster Abstieg

Steepest Descent (SD)

Idee: Gieriger Algorithmus

- Nimm Schritt in Richtung des stärksten Gefälles
- Nächster Schritt, bis es nicht mehr tiefer geht
- Stärkstes Gefälle entlang negativem Gradient $-\nabla E$
- Schrittweite?



Liniensuche

Gegeben:

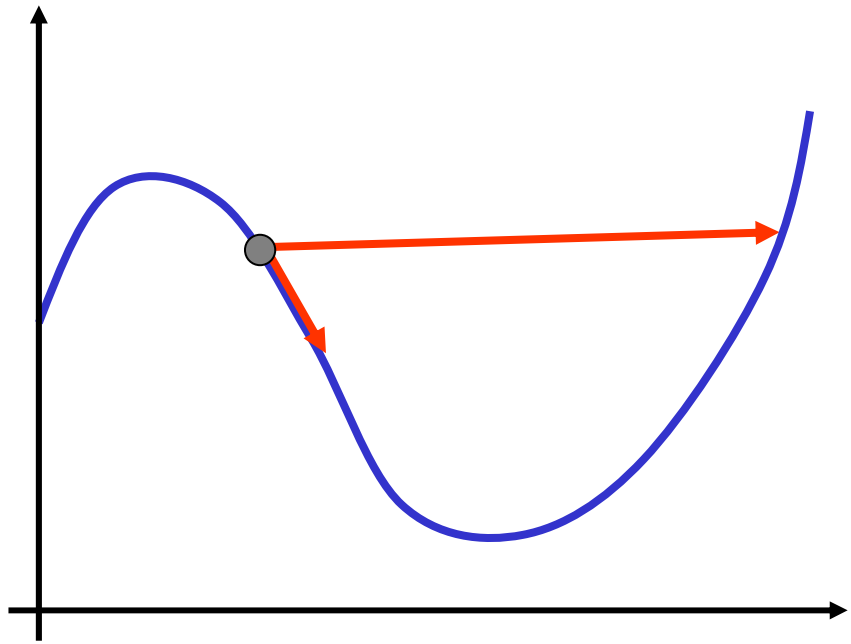
Suchrichtung **s**

(3N-dimensionaler Einheitsvektor)

Aufgabe:

Finde sinnvollen Schritt in Suchrichtung

- Zu kurz: ineffizient
- Zu lang: verpasst Minimum



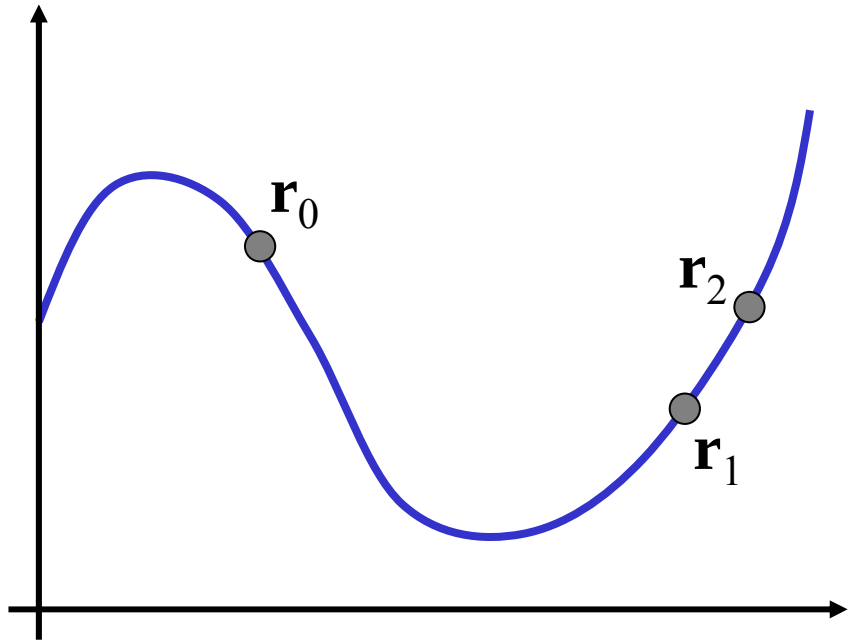
Liniensuche

- Sehr viele mögliche Varianten des Algorithmus
- Häufig wird folgende Idee benutzt
  - Minimum eingrenzen
  - Minimumsposition interpolieren
- Eingrenzen findet zur Startposition $\mathbf{r}_0$ zwei Punkte $\mathbf{r}_1$ und $\mathbf{r}_2$ mit

  $\mathbf{r}_1 = \mathbf{r}_0 + \lambda_1\mathbf{s}$ Æ $\mathbf{r}_2 = \mathbf{r}_0 + \lambda_2\mathbf{s}$

  für $\lambda_1, \lambda_2 > 0$ für die gilt

  $E(\mathbf{r}_1) < E(\mathbf{r}_0)$ Æ $E(\mathbf{r}_1) < E(\mathbf{r}_2)$



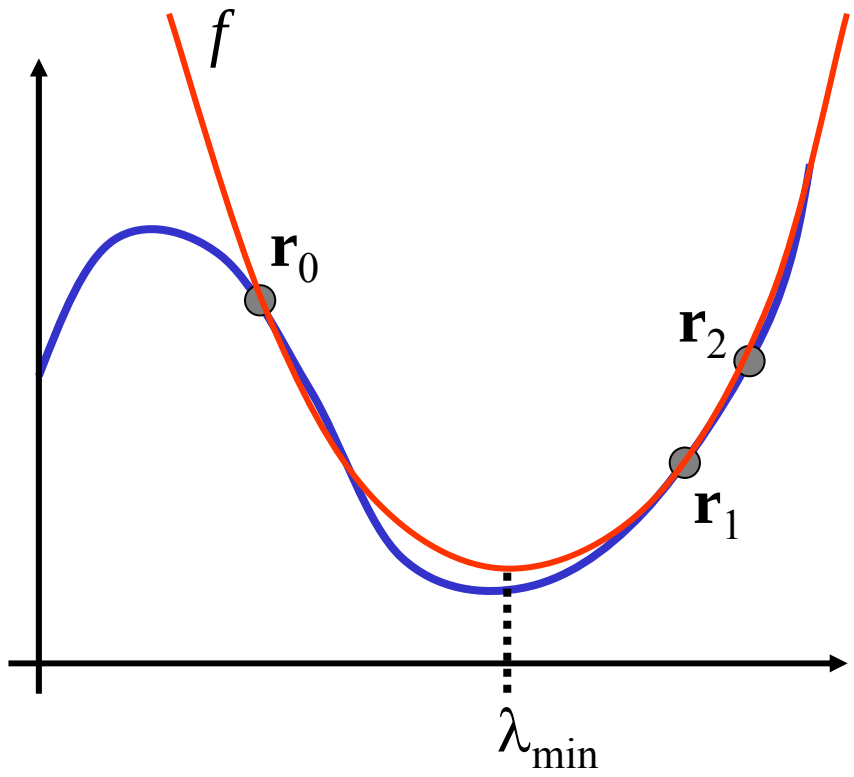
Liniensuche

- **Eingrenzen** findet zur Startposition $\mathbf{r}_0$ zwei Punkte $\mathbf{r}_1$ und $\mathbf{r}_2$ mit

  $\mathbf{r}_1 = \mathbf{r}_0 + \lambda_1\mathbf{s}$ Æ $\mathbf{r}_2 = \mathbf{r}_0 + \lambda_2\mathbf{s}$

  für $\lambda_1, \lambda_2 > 0$ für die gilt

  $E(\mathbf{r}_1) < E(\mathbf{r}_0)$ Æ $E(\mathbf{r}_1) < E(\mathbf{r}_2)$
- **Interpolation** schätzt die Lage des Minimums λ_{min} durch eine Interpolationsfunktion $f(\lambda)$ (meist quadratisch oder kubisch) ab
- f wird so gewählt, dass $f(\mathbf{x}) = E(\mathbf{x})$ für $\mathbf{x} = \mathbf{r}_0, \mathbf{r}_1, \mathbf{r}_2$

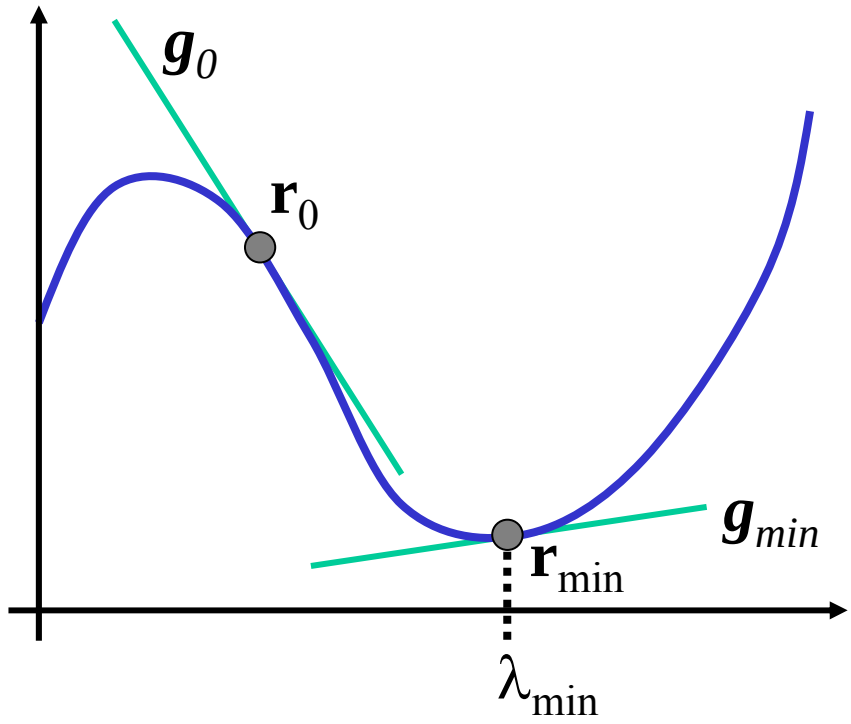


Liniensuche

- Schritt nach $\mathbf{r}_{min} = \mathbf{r}_0 + \lambda_{min}\mathbf{s}$ wird akzeptiert, falls folgende **Kriterien** erfüllt sind
  - $E(\mathbf{r}_0 + \lambda_{min}\mathbf{s}) \cdot E(\mathbf{r}_0) + \alpha\, \lambda_{min}\, \mathbf{g}_0\mathbf{s}$
    (ausreichende Energieabnahme)
  - $|\mathbf{g}_{min}\, \mathbf{s}| \cdot \beta\, |\mathbf{g}_0\, \mathbf{s}|$
    (ausreichende Abnahme des Gradienten)
- Wahl von α und β ist etwas heikel. Bewährt hat sich

  $\alpha = 0.0001$ und $\beta = 0.9$
- Bei exakter Liniensuche gilt:

  $\mathbf{s}\, \mathbf{g}_{min} = 0$



Steilster Abstieg

STEILSTER_ABSTIEG(r):

Für k = 0 bis Konvergenz

- Berechne Gradient $\mathbf{g}_k$
- $\mathbf{s}_k = -g_k / |g_k|$
- λ_k = **LINIENSUCHE**($\mathbf{r}_k$, $\mathbf{s}_k$)
- Aktualisiere Koordinaten mit

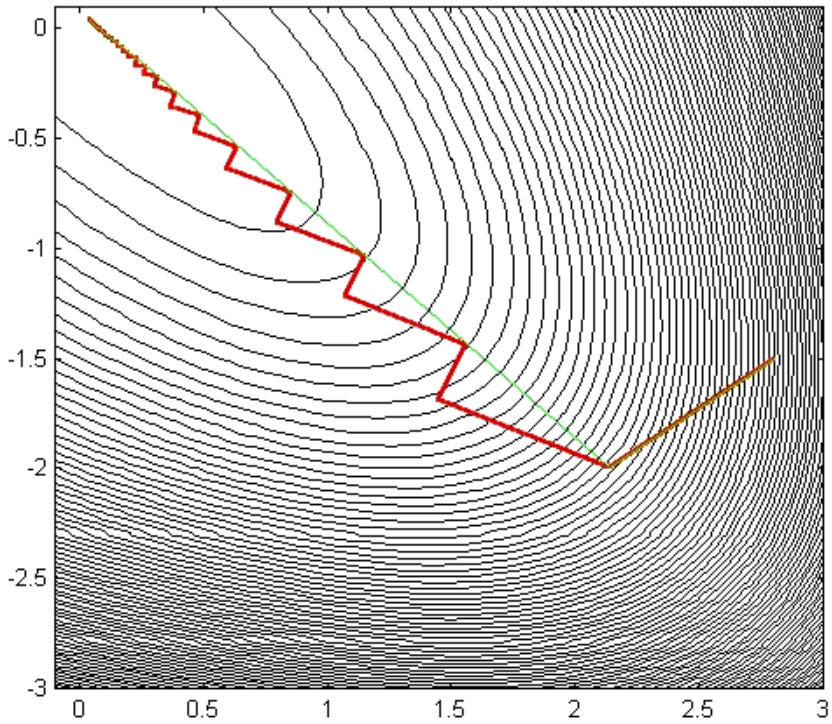
$$\mathbf{r}_{k+1} = \mathbf{r}_k + \lambda_k \mathbf{s}_k$$

- Abbruch falls Konvergenz
- $k = k + 1$, nächster Schritt

Steilster Abstieg

- **Vorteile**
  - Sehr einfach
  - Sehr schnell für sehr steile Gradienten
- **Nachteile**
  - Langsame Konvergenz
  - $\mathbf{s}_k$ und $\mathbf{s}_{k+1}$ sind zueinander orthogonal!

) Zick-Zack-Verhalten in schmalen Tälern!



Newton-Verfahren

- Besser als CG wird man, wenn man zusätzlich die 2. Ableitungen **H** berücksichtigt
- Entwicklung von E(r) um r_0 (mit $\Delta\mathbf{r} = \mathbf{r} - \mathbf{r_0}$)

$$E(\mathbf{r}) = E(\mathbf{r}_0) + \mathbf{g}(\mathbf{r_0})\Delta\mathbf{r} + \tfrac{1}{2}\,\Delta\mathbf{r}^{\mathbf{T}}\mathbf{H}(\mathbf{r_0})\Delta\mathbf{r}$$

- Es gibt eine Reihe von Techniken die **H** nutzen (Newton, Truncated Newton, Quasi-Newton, …)
  - Vorteil: schnellere Konvergenz
  - Nachteile
    - Sehr speicherintensiv (**H** wächst wie $O(N^2)$)
    - Sehr rechenintensiv
  - Für Proteine selten verwandt

Anmerkungen

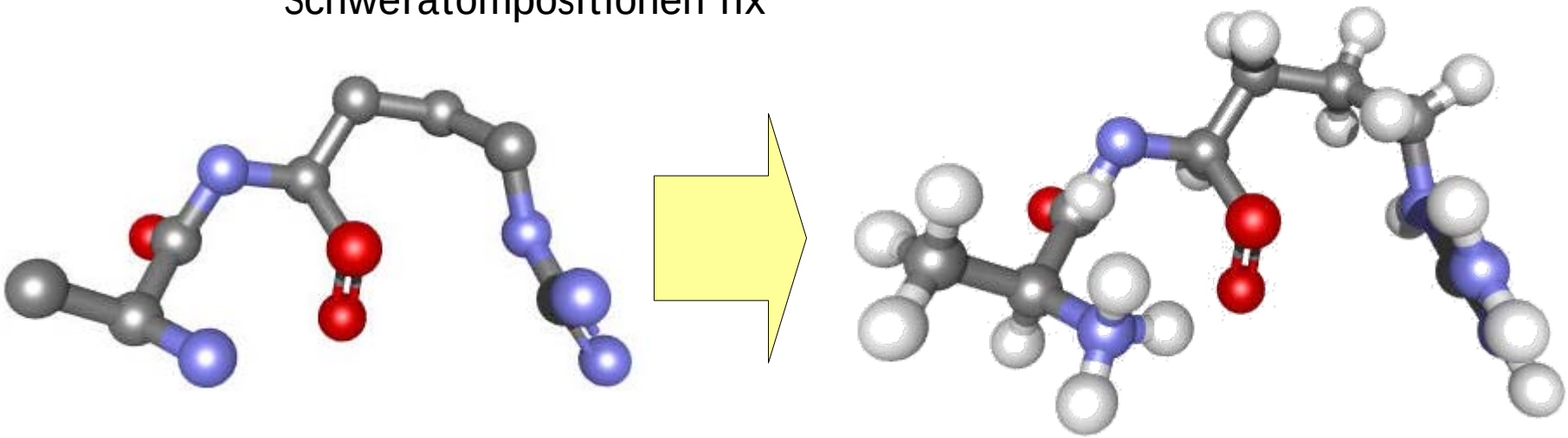
- Energieminimierung ist unerlässlich für die Konstruktion neuer Strukturen
- In vielen der später in der Vorlesung behandelten Methoden taucht Energieminimierung als ein Teilschritt auf
- Für große Proteine können Minimierungen sehr lange dauern
- In vielen Fällen kann man die Rechenzeit drastisch reduzieren, indem man die Konvergenzkriterien geschickt wählt (ein Gradient von Null ist of nicht notwendig)

Anmerkungen

- Achtung! Alle beschriebenen Algorithmen sind Heuristiken, die lokale Minima liefern
- Endergebnis hängt ab von
  - Startposition
  - Kraftfeld, Implementierung
  - Minimierungsalgorithmus
  - Wahl der Parameter (z.B. Konvergenzkriterium)

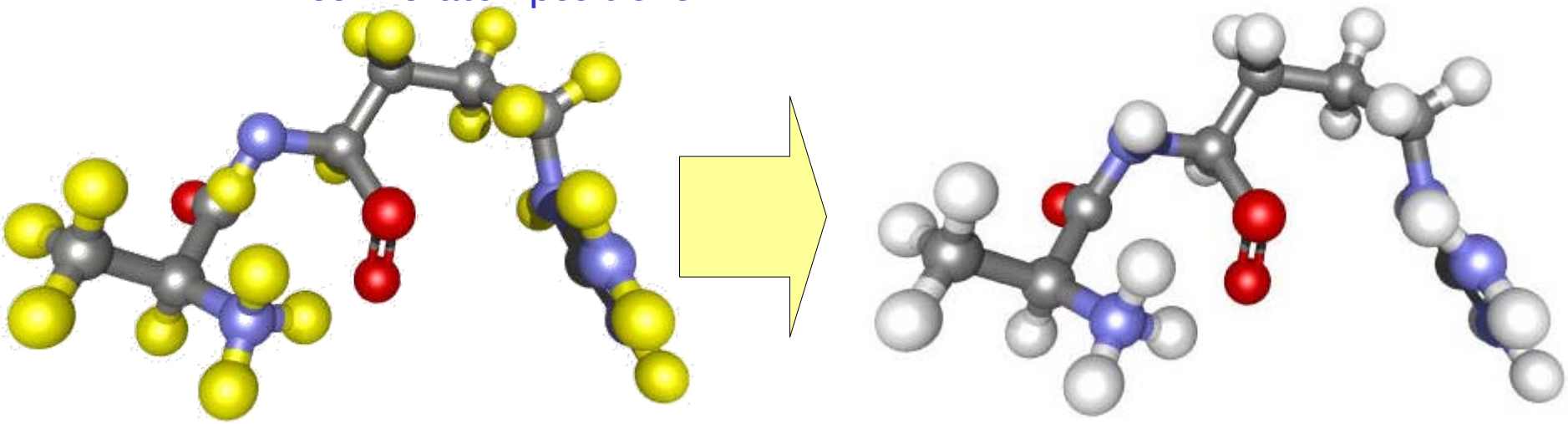
Beispiel

- Vervollständigung von Kristallstrukturen
  - Hinzufügen fehlender (Wasserstoff-)Atome
  - Bei Wasserstoffatomen sind die generierten Koordinaten sehr akkurat (nur eine Bindung ! wenig Abhängigkeiten)
  - Vorgehen:
    - Konstruiere Wasserstoffe in “Standardkonformation”, d.h. sinnvoller Abstand, annähernd sinnvolle Bindungswinkel
    - Minimiere Position der Wasserstoffe, aber halte Schweratompositionen fix



Beispiel

- Vervollständigung von Kristallstrukturen
  - Hinzufügen fehlender (Wasserstoff-)Atome
  - Bei Wasserstoffatomen sind die generierten Koordinaten sehr akkurat (nur eine Bindung ! wenig Abhängigkeiten)
  - Vorgehen:
    - Konstruiere Wasserstoffe in “Standardkonformation”, d.h. sinnvoller Abstand, annähernd sinnvolle Bindungswinkel
    - Minimiere Position der Wasserstoffe, aber halte Schweratompositionen fix



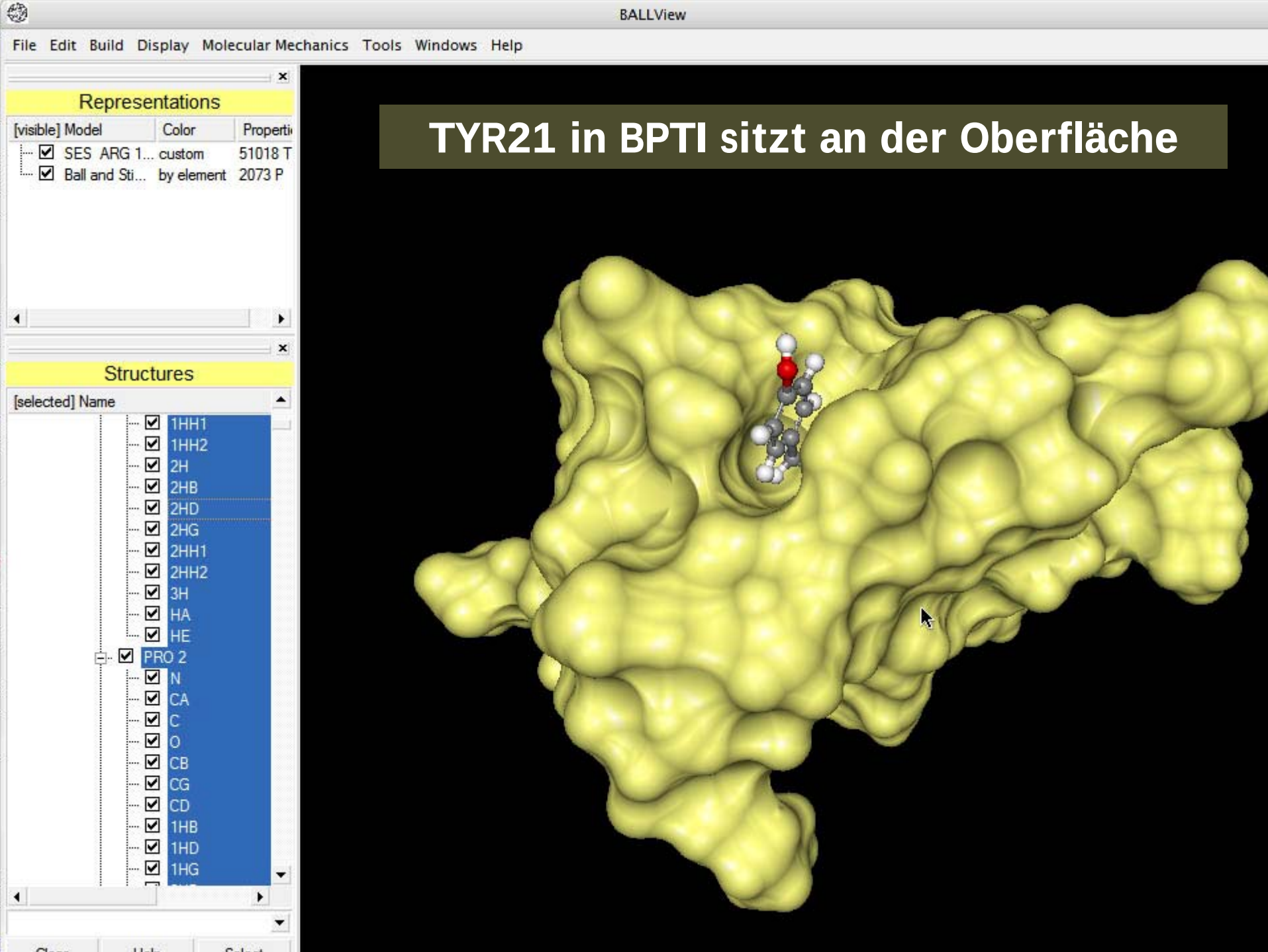
Beispiel

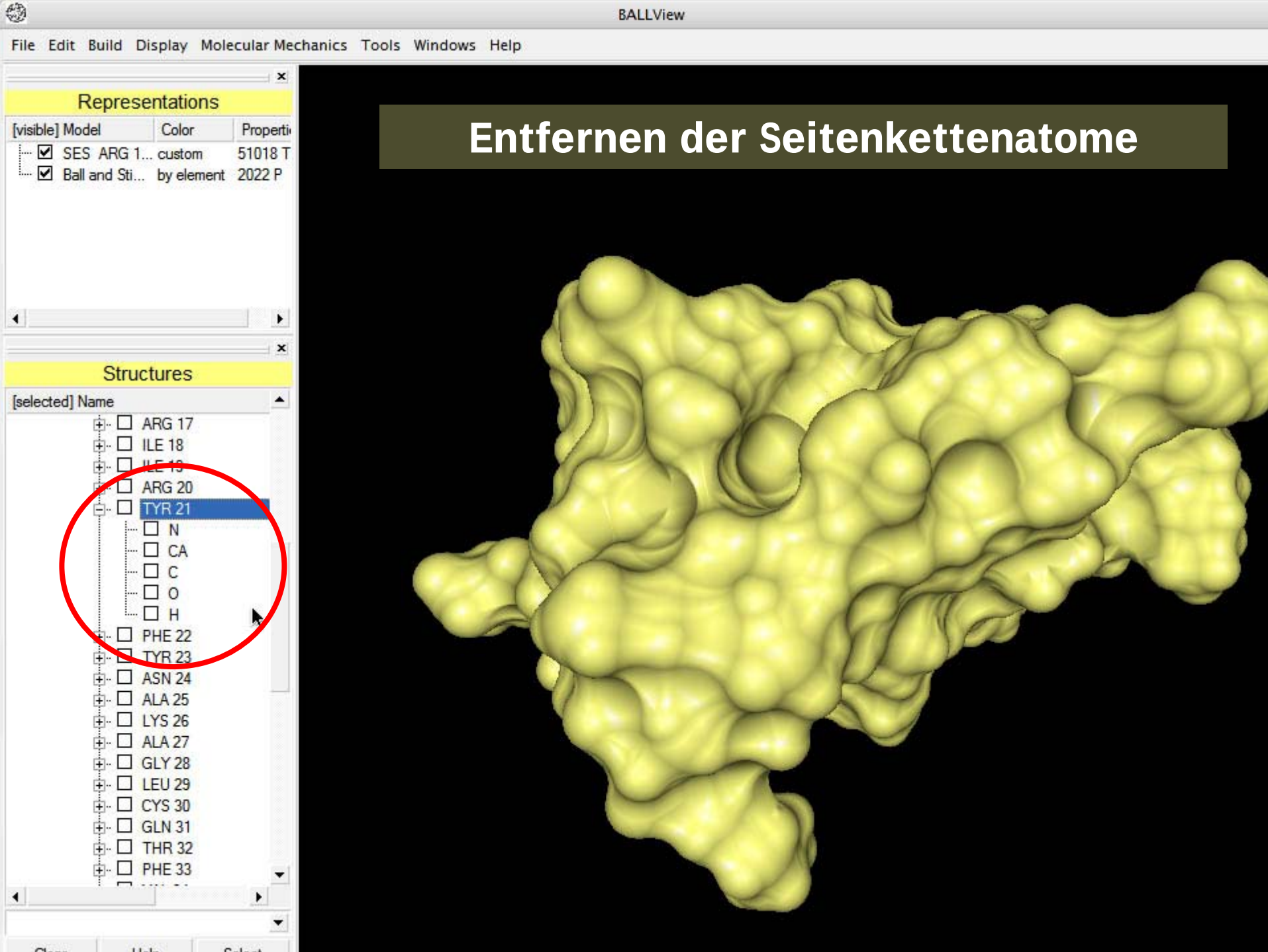
- Einführen von Punktmutationen
  - Auf triviale Weise (wir werden später sehen, wie man dies besser macht…) durch Ersetzen der Seitenkette
  - **Beispiel**

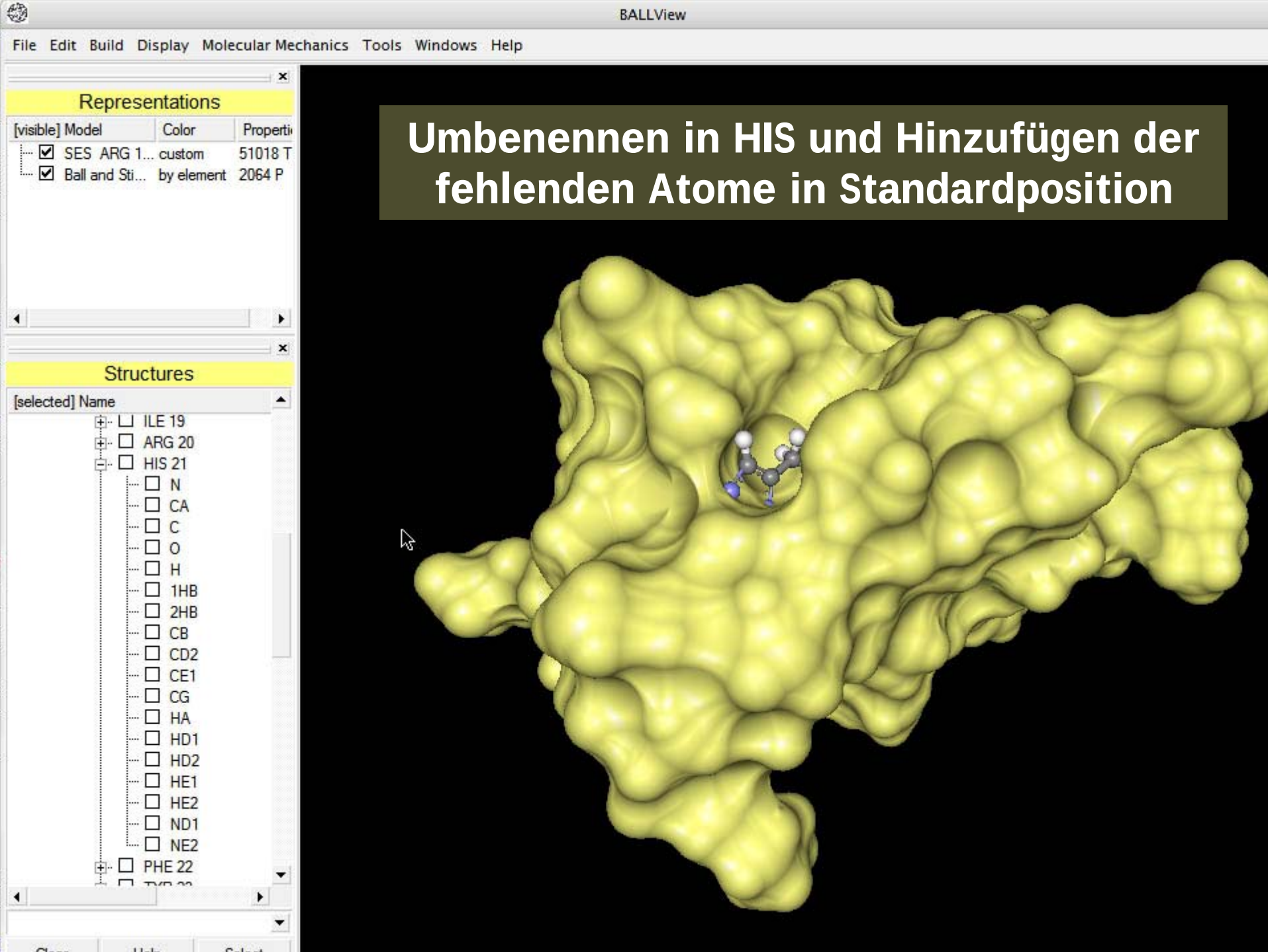
    BPTI, TYR 21 ! HIS

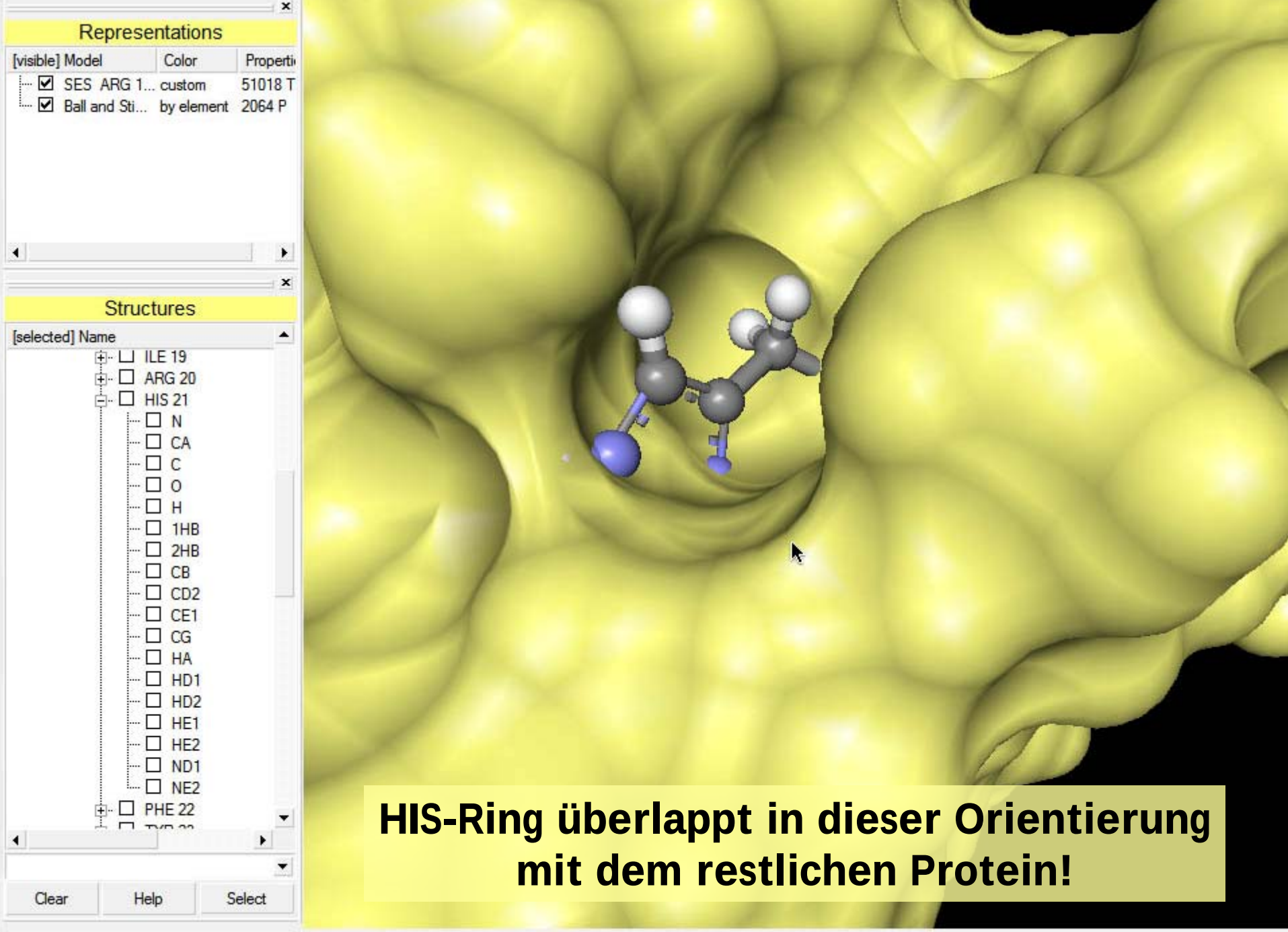
    Vorgehensweise:

    - Alte Seitenkette löschen
    - Neue Seitenkette in Standardorientierung einfügen
    - Überlappende Atome werden durch Geometrieoptimierung zu sinnvoller Struktur


BALLView
File Edit Build Display Molecular Mechanics Tools Windows Help
Representations
[visible] Model Color Properti
SES ARG 1... custom 51018 T
Ball and Sti... by element 2073 P
Structures
[selected] Name
1HH1
1HH2
2H
2HB
2HD
2HG
2HH1
2HH2
3H
HA
HE
PRO 2
N
CA
C
O
CB
CG
CD
1HB
1HD
1HG
TYR21 in BPTI sitzt an der Oberfläche


BALLView
File Edit Build Display Molecular Mechanics Tools Windows Help
Representations
[visible] Model Color Properti
SES ARG 1... custom 51018 T
Ball and Sti... by element 2022 P
Structures
[selected] Name
ARG 17
ILE 18
ARG 20
TYR 21
N
CA
C
O
H
PHE 22
TYR 23
ASN 24
ALA 25
LYS 26
ALA 27
GLY 28
LEU 29
CYS 30
GLN 31
THR 32
PHE 33
Entfernen der Seitenkettenatome


BALLView
File Edit Build Display Molecular Mechanics Tools Windows Help
Representations
[visible] Model Color Properti
SES ARG 1... custom 51018 T
Ball and Sti... by element 2064 P
Structures
[selected] Name
ILE 19
ARG 20
HIS 21
N
CA
C
O
H
1HB
2HB
CB
CD2
CE1
CG
HA
HD1
HD2
HE1
HE2
ND1
NE2
PHE 22
Umbenennen in HIS und Hinzufügen der fehlenden Atome in Standardposition

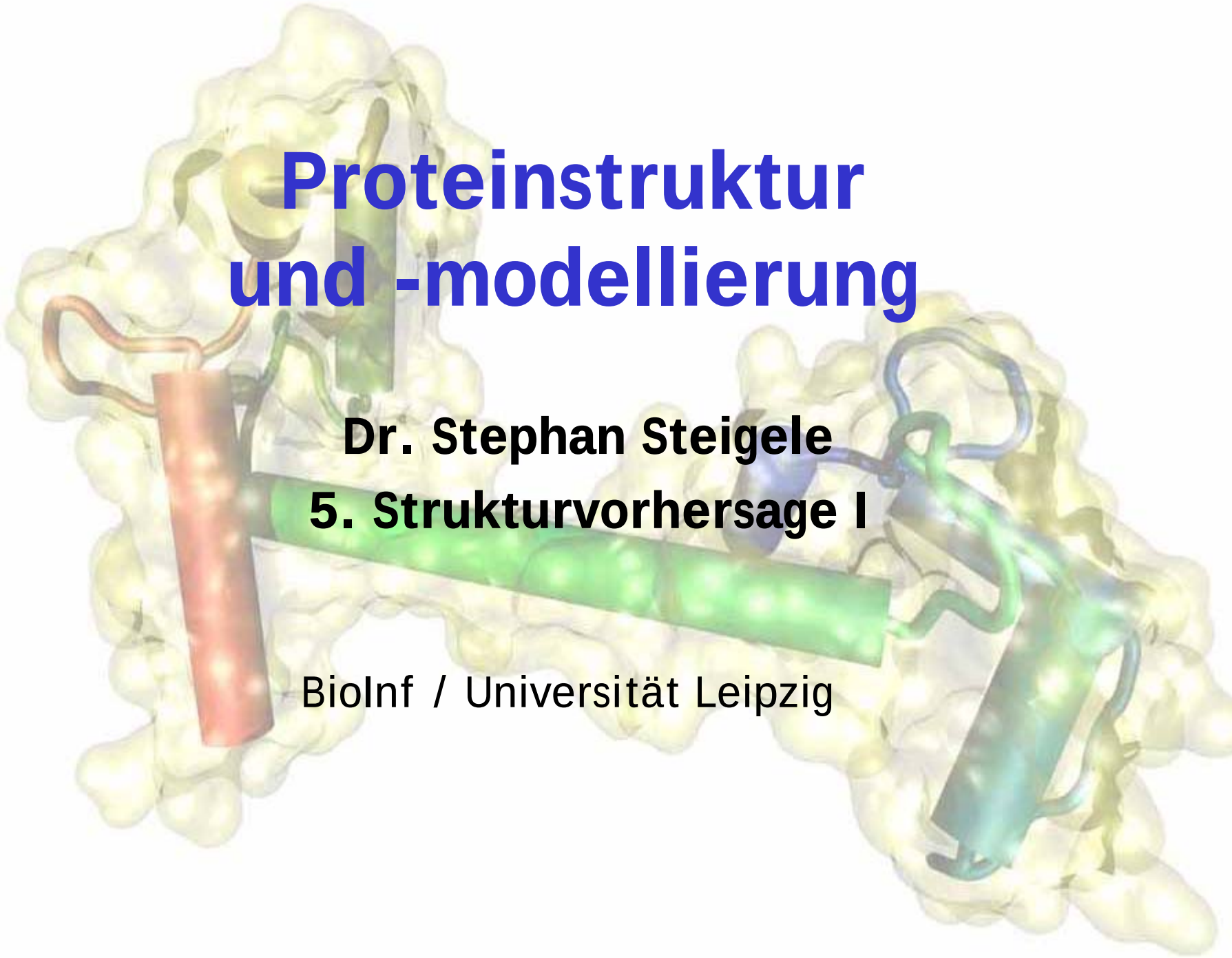

Representations
[visible] Model
Color
Properti
SES ARG 1...
custom
51018 T
Ball and Sti...
by element
2064 P
Structures
[selected] Name
ILE 19
ARG 20
HIS 21
N
CA
C
O
H
1HB
2HB
CB
CD2
CE1
CG
HA
HD1
HD2
HE1
HE2
ND1
NE2
PHE 22
Clear
Help
Select
HIS-Ring überlappt in dieser Orientierung
mit dem restlichen Protein!


BALLView
File Edit Build Display Molecular Mechanics Tools Windows Help
Representations
[visible] Model Color Properti
SES ARG 1... custom 51045 T
Ball and Sti... by element 2064 P
Structures
[selected] Name
4pti
PROTEINASE INHIBITOR (
ARG 1
PRO 2
ASP 3
PHE 4
CYS 5
LEU 6
GLU 7
PRO 8
PRO 9
TYR 10
THR 11
GLY 12
PRO 13
CYS 14
LYS 15
ALA 16
ARG 17
ILE 18
ILE 19
Geometrieoptimierung entfernt Überlapp
und führt zu sinnvoller Struktur

Literatur

Molekülmechanik

- **Andrew R. Leach, Molecular Modelling – Principles and Applications**, Prentice Hall, 2001
- **Daan Frenkel, Berend Smit, Understanding Molecular Simulation**, Academic Press, 1996
- **Martin J. Field, A practical introduction to the simulation of molecular systems**, Cambridge University Press, 1999
- **Tamar Schlick, Molecular Modeling and Simulation**, Springer, 2003
- **Ulrich Burkert, Norman L. Allinger, Molecular Mechanics**, American Chemical Society, 1982



Proteinstruktur und -modellierung

Dr. Stephan Steigele

5. Strukturvorhersage I

BioInf / Universität Leipzig

Strukturvorhersage – Übersicht

- Problemdefinition / klassifizierung
- Sekundärstrukturvorhersage
- Fold-Recognition -> behandeln wir nicht
- Threading
- ab-initio-Vorhersage
- CASP/CAFASP

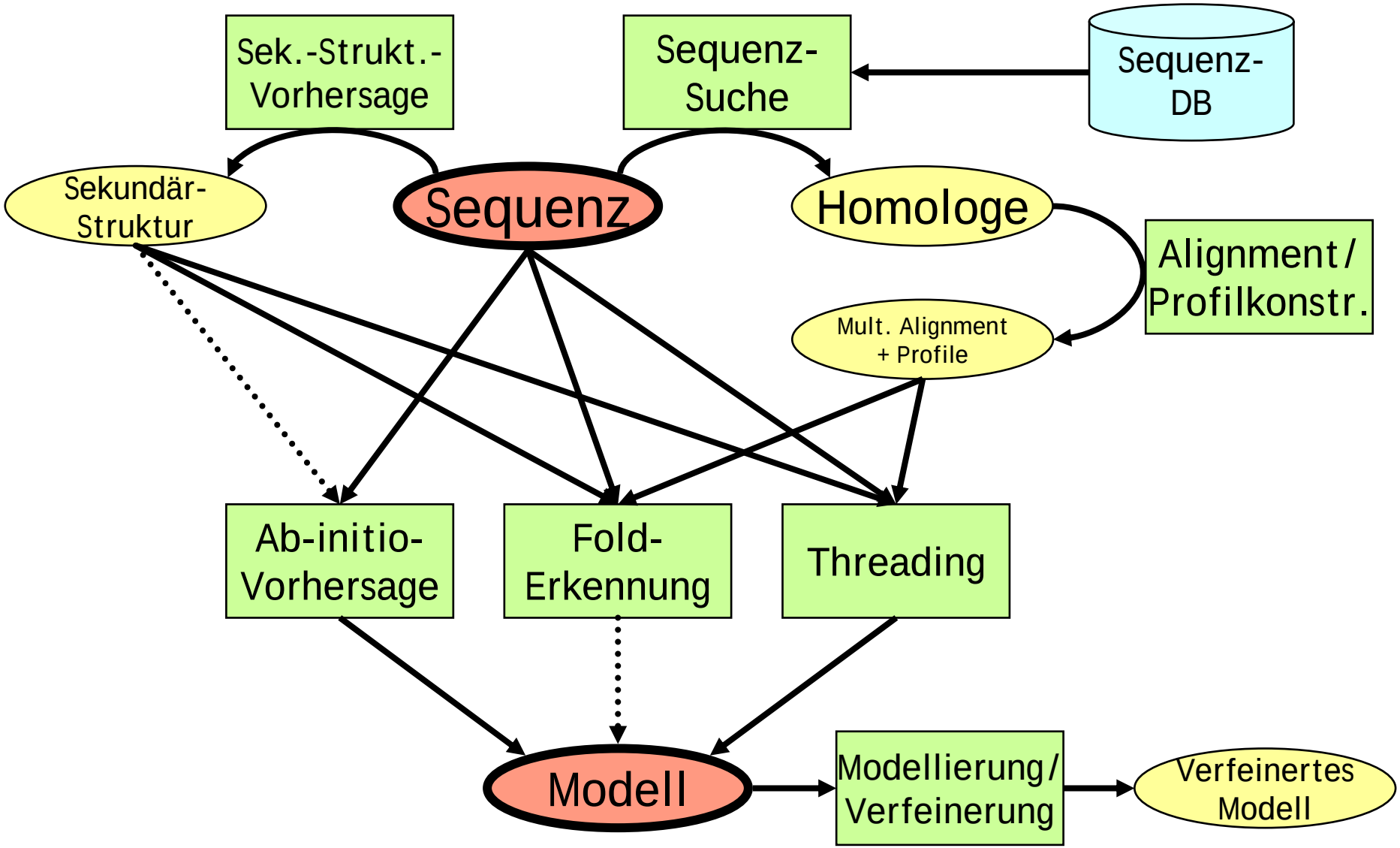
Protein-Strukturvorhersage

- **Grundproblem:**

  Gegeben Sequenz, finde Struktur

- Geeignete Methode abhängig von
  - Verfügbarkeit homologer Strukturen
  - Verfügbarkeit experimenteller Daten
  - Qualität/Auflösung des gewünschten Modells
- Vorhersage nur von Backbone-Positionen!
  - Seitenketten werden getrennt modelliert
  - Entsprechende Techniken werden wir später besprechen

Methoden

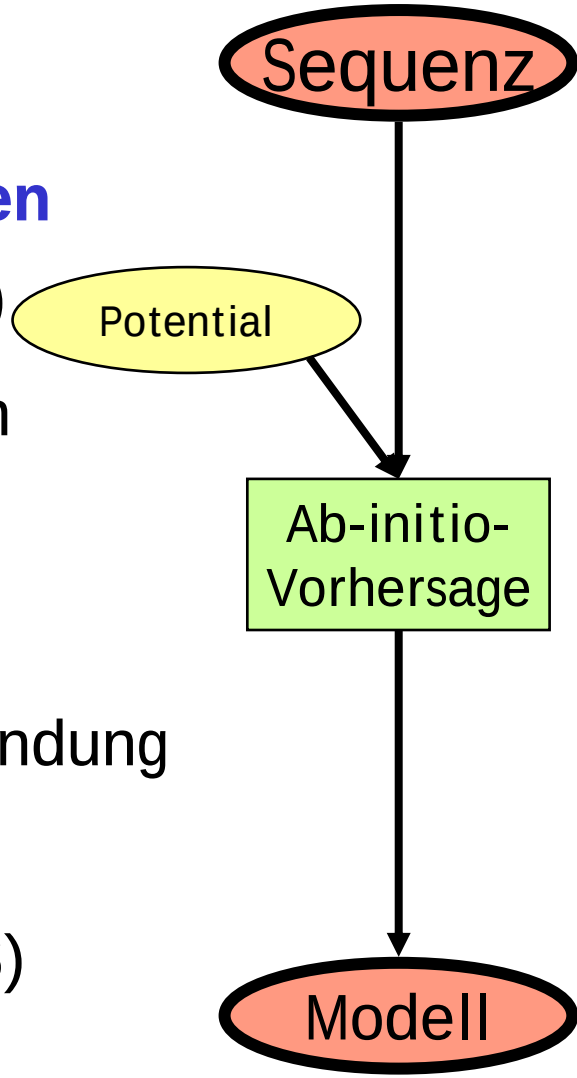


Nach: Zimmer, Lengauer: Bioinformatics – From Genomes to Drugs, Wiley VCH, 2001

Definition ab-initio-Vorhersage

Ab-initio-Vorhersage

- Vorhersage ausgehend von **physikalischen Modellen** (*ab initio* = „erste Prinzipien“)
- Keine Verwendung homologer Strukturen
- **Vorhersage neuer Folds** möglich
- Fragment-Assemblierung (z.B. ROSETTA)
  streng genommen nicht *ab initio* (Verwendung von bekannten Strukturfragmenten)
- Anwendbar für **kleine Proteine** (<100 AS)



Definition Threading

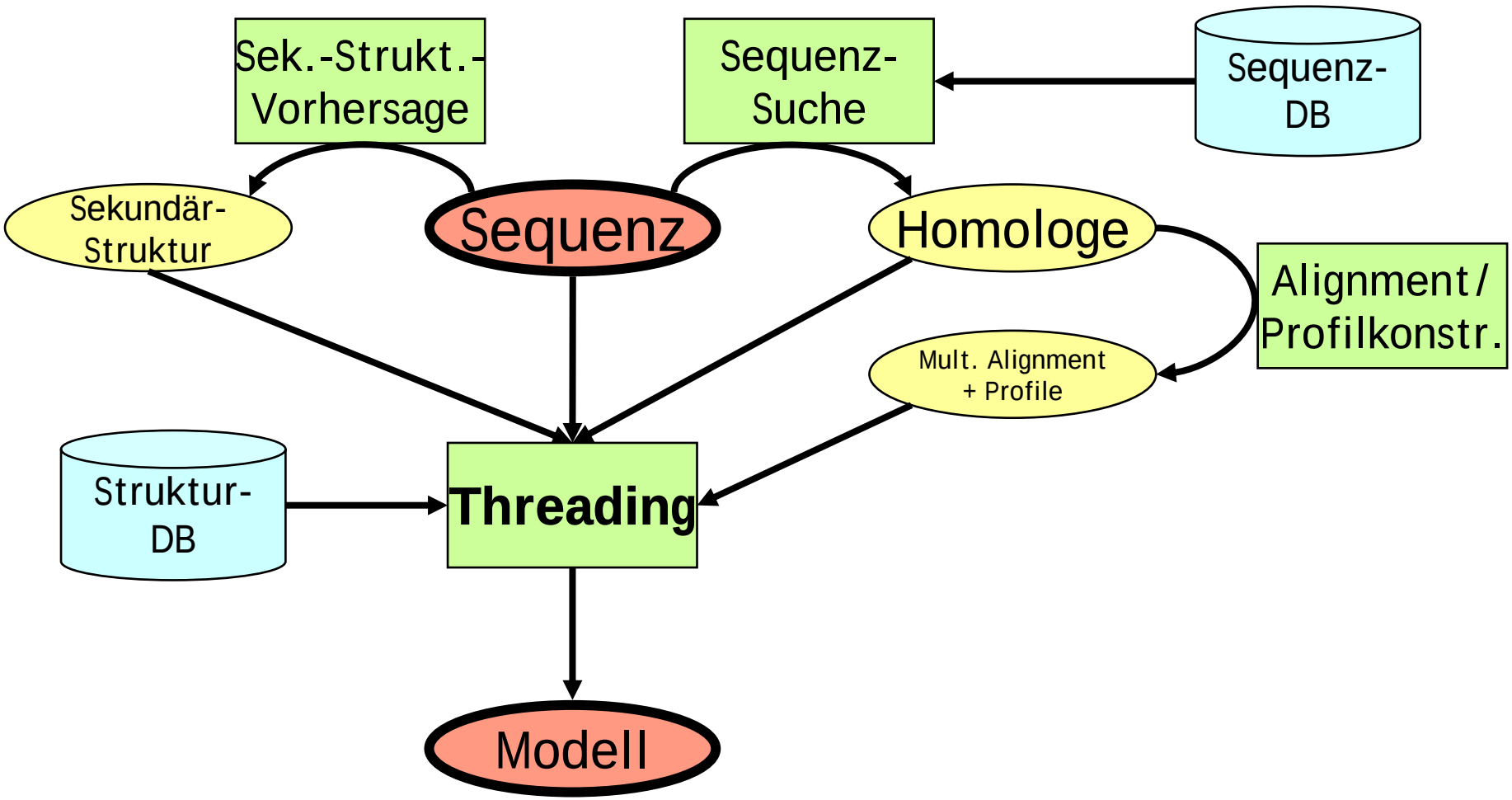
Threading („Auffädeln“)

- Modellierung der **Ziel-Sequenz** (*target*) auf homologe Struktur (**Schablone**, *template*)
- Nur Modellierung **bekannter Faltungsklassen**
- Abbildung passender Sequenzfragmente auf die Schablonen-Struktur

Fold-Erkennung

- Vereinfachte Version des Threading-Problems
- **Identifiziere Faltungsklasse** der Ziel-Sequenz

Definition Threading



Sekundärstruktur-Vorhersage

Gegeben: Sequenz

```
KVYGRCELAAAMKRLGLDNYRGYSLGNWVC
AAKFESNFNTHATNRNTDGSTDYGILQINS
RWWCNDGRTPGSKNLCNIPCSALLSSDITA
SVNCAKKIASGGNGMNAWVAWRNRCKGTDV
HAWIRGCRL
```

Gesucht:

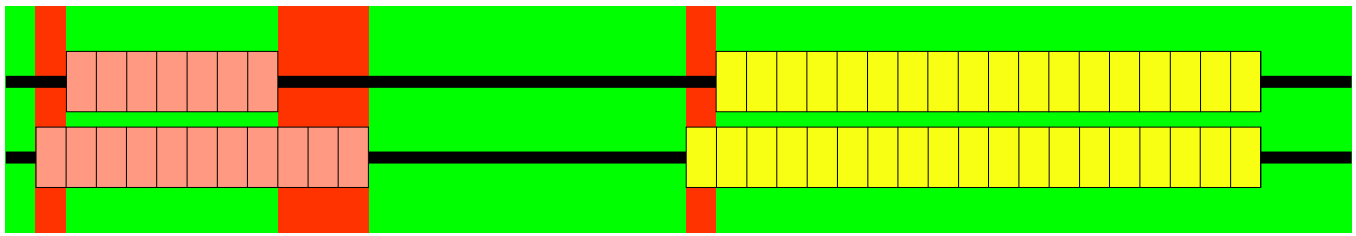
Sekundärstruktur-Zuordnung mit den Klassen E (*extended*, Faltblatt), H (Helix), C/– (*coil*, Schleife) zu jeder Aminosäure

```
KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNTHATNRNTD
-----HHHHHHHHH-------------EEEEE----------------

GSTDYGILQINSRWWCNDGRTPGSKNLCNIPCSALLSSDITASVNCAK
----EEEEEE-------------------------------HHHHHH

KIASGGNGMNAWVAWRNRCKGTDVHAWIRGCRL
HHH-------EEE--------------------
```

Qualitätsmaße

- Drei-Zustands-Klassifikation (C/H/E – Coil/Helix/Extended)
- **Q_3-Score**: Prozentsatz an korrekt zugewiesenen AS in der Sequenz
- Maximal 100% = alles korrekt vorhergesagt
- Da insbesondere die Enden der Sekundärstrukturelemente

  nicht einwandfrei klassifizierbar sind sind > 80% sehr gut



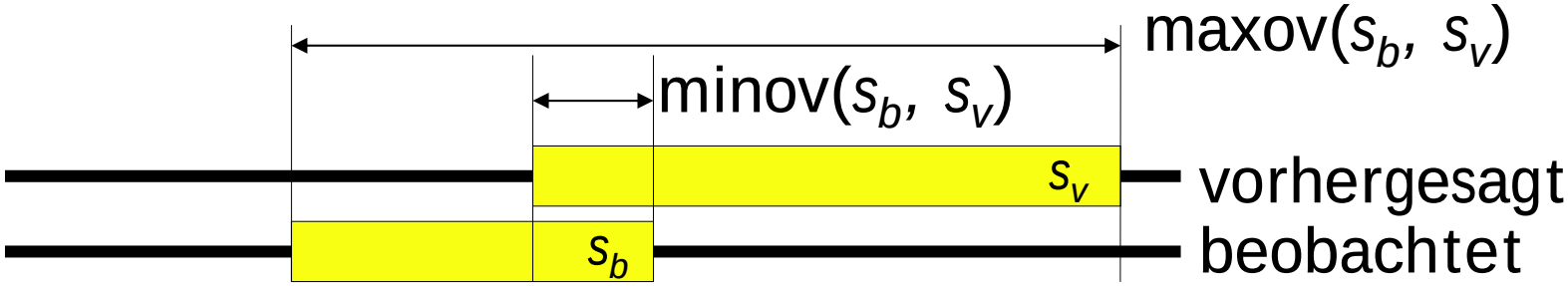
vorhergesagt
beobachtet

Qualitätsmaße

- Gelegentlich auch **Acht-Zustands-Klassifikation** (C/H/E/G/I/T/B/S/L)
  - 3_{10}-Helix (G)
  - α-Helix (H)
  - π-helix (I)
  - Helix-Turn (T)
  - Faltblatt (E)
  - β-Brücke (B)
  - Bend (S)
  - Andere/Loop (L)
- **Q_8-Score**: Prozentsatz an korrekt zugewiesenen AS in der Sequenz
- Acht Klassen lassen sich auf drei abbilden:
  - HELIX = 3_{10}-Helix + α-Helix + π-Helix
  - EXTENDED = Faltblatt + β-Brücke
  - LOOP = Loop + Bend + Helix-Turn
- Q_8-Score geringer als Q_3-Score

Segment OVerlap – SOV

- Maß für den Überlapp in Vorhersage und wirklicher Zuordnung
- Vergleicht beobachtete (s_b) und vorhergesagte (s_v) Segmente gleichen Typs (Typ: H, C oder E)
- 100% für korrekte Zuordnung
- minov(s, t): Länge des „Schnitts“ von s und t
- maxov(s, t): Länge der „Vereinigung“ von s und t



Sekundärstruktur-Vorhersage

Mehrere Generationen von Algorithmen

1. Generation

Nur Eigenschaften einzelner AS (Q_3 > 50 – 60%)

2. Generation

Einbeziehung lokaler Umgebung (Q_3 > 65%)

3. Generation

Einbeziehung homologer Sequenzen (Q_3 > 70%)

4. Generation

Konsensus-Methoden die Ergebnisse mehrerer Methoden der 2. + 3. Generation kombiniert (Q_3 > 75-80%)

Testdaten

- Vergleich üblicherweise mit **bekannten Strukturen**, für die die Sekundärstrukturzuordnung automatisch mit **DSSP** bestimmt wird
- DSSP betrachtet dabei für die Kristallstruktur die Backbone-Torsionswinkel, H-Brückenmuster, Lösemittelexponiertheit und andere Parameter, die für verschiedene Sekundärstrukturen charakteristisch sind
- Jeder AS wird dabei eine von drei bzw. acht Sekundärstrukturklassen zugewiesen
- Probleme ergeben sich für
  - NMR-Strukturen
  - Strukturen mit Auflösung schlechter als 2 Å
- Testsatz aus Strukturen niedriger Homologie

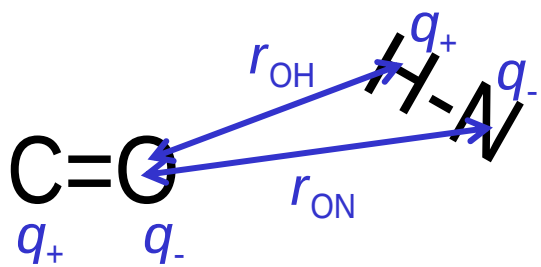
  (< 25% Sequenzidentität)

DSSP

- Kern des DSSP-Algorithmus ist eine Funktion zur Erkennung von H-Brücken des Protein-Rückgrats
- Ob eine H-Brücke zwischen zwei AS existiert, wird über die elektrostatische Energie (Coulomb) entschieden:

$$E = \frac{q_+ q_-}{4\pi\varepsilon_0}\left(\frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}}\right)$$

- Dem Modell liegt die Annahme zugrunde, dass die beiden Bindungen C=O und H-N polarisiert sind und daher Partialladungen q_+ und q_- tragen:



$q_- = -0.20\ e_0$

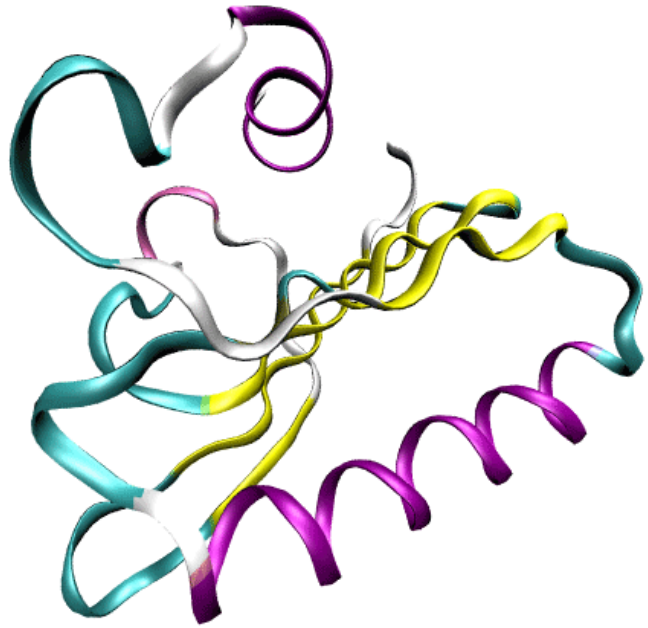
$q_+ = +0.42\ e_0$

DSSP

- Wasserstoffpositionen für NH werden dabei aus Standardgeometrien konstruiert (nicht in Kristallstrukturen enthalten!)
- Für alle Paare *(i,j)* von Aminosäuren wir eine H-Brücke angenommen, wenn E_{ij} kleiner (negativer) als ein Threshold t = -2.4 kJ/mol ist
- Liegen H-Brücken für (*i*,*i*+3), (*i*,*i*+4) oder (*i*,*i*+5) vor, wird eine 3-, 4-, oder **5-Turn** angenommen
- Mehrere benachbarte Turns gleichen Typs entsprechen 3_{10}-, α- und π-**Helices**
- Eine β**-Brücke** liegt vor, wenn H-Brücken existieren für
  - *(i-1,j)* und *(j,i+1)* [parallel]
  - *(i,j)* und *(j,i)* [antiparallel]
- Benachbarte β-Brücken vom gleichen Typ entsprechen parallelen/antiparallelen **Faltblättern**

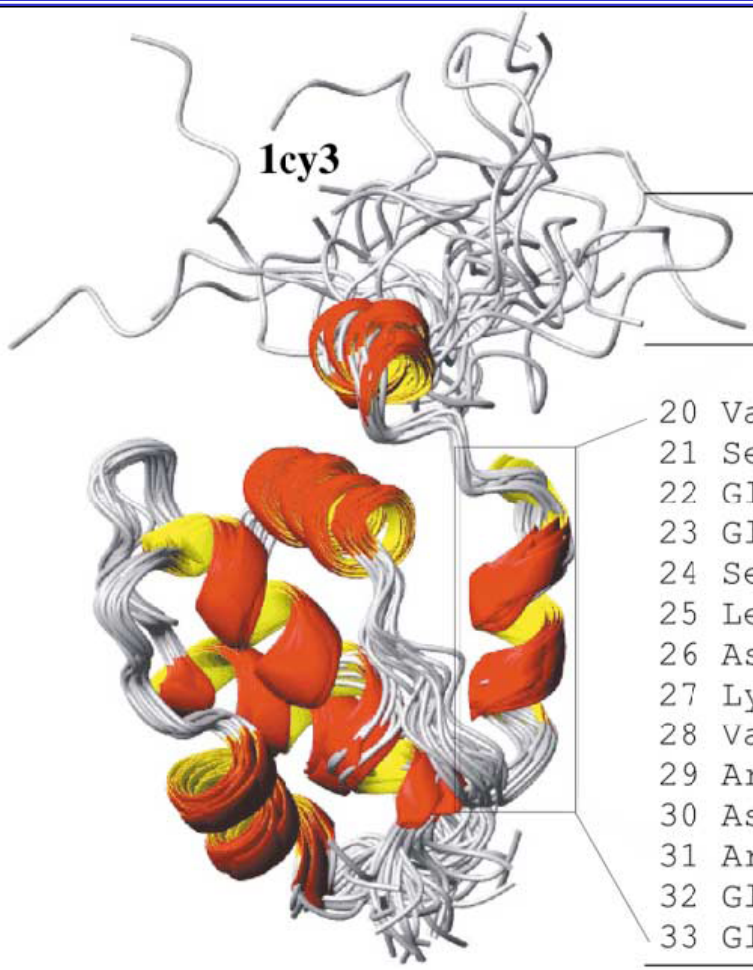
DSSPcont

- Zuweisung einer Sekundärstruktur nicht unbedingt eindeutig
  - Strukturen sind flexibel
  - Teile der Struktur fluktuieren zwischer mehreren Zuständen
- Beispiel: schwache (in der Nähe des Thresholds von DSSP) H-Brücken am Ende einer Helix
- **DSSPcont**: statt fester Zuordnung, Wahrscheinlichkeit mit der eine AS eine Sekundärstruktur annimmt



Andersen et al., Structure (2002), 10, 175

DSSPcont



A DSSP assignment for models 1-23

		1 5 10 15 20 23
20	Val	LLLLLLLLLLLLLLLLLLLLLLL
21	Ser	TLLHTLLTLLLLHLLLHLLTLLT
22	Glu	TTTHTTTTHTSSHTSTHSSTSTT
23	Glu	STTHHTTHHTSSHTHTHSSSHTH
24	Ser	HHHHHHHHHHHHHHHHHHHHHHH
25	Leu	HHHHHHHHHHHHHHHHHHHHHHH
26	Asn	HHHHHHHHHHHHHHHHHHHHHHH
27	Lys	HHHHHHHHHHHHHHHHHHHHHHH
28	Val	HTHTHHTHHTHHHTHHHHHTHHH
29	Arg	TTTTTTTTHTTTTTTTTTTTTTT
30	Asn	SSTSSSSSSTSSSSSTSTSSSTS
31	Arg	SSTSTSSSSTTSSTTTTTTSSTS
32	Glu	SSTSTLSLSTTSSTTTTTTLSLL
33	Glu	LLLLLLLLLLLLLLLLLLLLLLL

B DSSPcont assignment for model 1

G	H	I	T	E	B	S	L
0	0	0	32	0	0	0	68
0	0	0	100	0	0	0	0
0	0	0	100	0	0	0	0
0	32	0	21	0	0	47	0
0	99	0	1	0	0	0	0
0	99	0	1	0	0	0	0
0	99	0	1	0	0	0	0
1	99	0	0	0	0	0	0
1	87	0	12	0	0	0	0
1	53	0	46	0	0	0	0
0	0	0	10	0	0	90	0
0	0	0	32	0	0	88	0
0	0	0	32	0	0	88	0
0	0	0	0	0	0	0	100

- Variabilität der Sekundärstruktur (gerade an den Grenzen der Helix) in den 23 Modellen von 1CY3 wird durch DSSPcont korrekt wiedergegeben
- In diesen Bereichen ist auch Sekundärstrukturvorhersagen schwierig

Andersen et al., Structure (2002), 10, 175

Chou-Fasman-Algorithmus

- **Idee:** statistische Unterschiede in der „Neigung“ der AS zur Ausbildung von Sekundärstrukturen
- **Analyse von Strukturdatenbanken**: wie oft welche AS in welcher Sekundärstruktur
- n_j sei die Anzahl der Vorkommen von AS j in allen Proteinen der Strukturdatenbank
- **Wahrscheinlichkeit** p_j die AS j in einem Protein zu finden ist dann

$$p_j = n_j \; / \; \sum_j n_j$$

- Analog definiert man die Wahrscheinlichkeit AS j in Sekundärstruktur k (mit k = {C, H, E}) zu finden als

$$p_{j,k} = n_{j,k} \; / \; \sum_j n_{j,k}$$

Chou, Fasman, Biochemistry (1974), 13, 211

Chou-Fasman-Algorithmus

- Analog die Wahrscheinlichkeit $f_{j,k}$ die AS j in Sekundärstruktur k zu finden:

$$f_{j,k} = n_{j,k} / n_j$$

- Die mittlere Häufigkeit eine beliebige der 20 AS in der Sekundärstruktur k zu finden kann man damit schreiben als

$$\langle f_k \rangle = \sum_j f_{j,k} / 20 = \sum_j n_{j,k} / \sum_j n_j$$

- Die relative Häufigkeit, dass für AS j in Sekundärstruktur k auftritt ist somit:

$$P_{j,k} = f_{j,k} / \langle f_k \rangle$$

- Diese **relativen Häufigkeiten** beschreiben die Präferenzen einer jeden AS für eine gewisse Sekundärstruktur und bilden die Grundlage des Chou-Fasman-Algorithmus

Chou-Fasman-Algorithmus

- Einteilung der 20 AS in Klassen nach P_α^i
  - Starke Helixbildner H_α (Glu, Ala, Leu)
  - Helixbildner h_α (His, Met, Gln, Trp, Val, Phe)
  - Schwache Helixbildner I_α (Lys, Ile)
  - Indifferente i_α (Asp, Thr, Ser, Arg, Cys)
  - Schwache Helixbrecher b_α (Asn, Tyr)
  - Starke Helixbrecher B_α (Pro, Gly)
- Analog für β-Faltblätter
  - H_β, h_β, i_β, b_β, B_β

Chou, Fasman, Biochemistry (1974), 13, 211

Chou-Fasman-Parameter

AS	P_α	Klasse	AS	P_β	Klasse
Glu	1.53	H_α	Met	1.67	H_β
Ala	1.45		Val	1.65	
Leu	1.34		Ile	1.60	
His	1.24	h_α	Cys	1.30	h_β
Met	1.20		Tyr	1.29	
Gln	1.17		Phe	1.28	
Trp	1.14		Gln	1.23	
Val	1.14		Leu	1.22	
Phe	1.12		Thr	1.20	
Lys	1.07	I_α	Trp	1.19	

AS	P_α	Klasse	AS	P_β	Klasse
Ile	1.00	I_α	Ala	0.93	I_β
Asp	0.98	i_α	Arg	0.90	i_β
Thr	0.82		Gly	0.81	
Ser	0.79		Asp	0.80	
Arg	0.79		Lys	0.74	b_β
Cys	0.77		Ser	0.72	
Asn	0.73	b_α	His	0.71	
Tyr	0.61		Asn	0.65	
Pro	0.59	B_α	Pro	0.62	
Gly	0.53		Glu	0.26	B_β

Chou, Fasman, Biochemistry (1974), 13, 222

Chou-Fasman-Algorithmus II

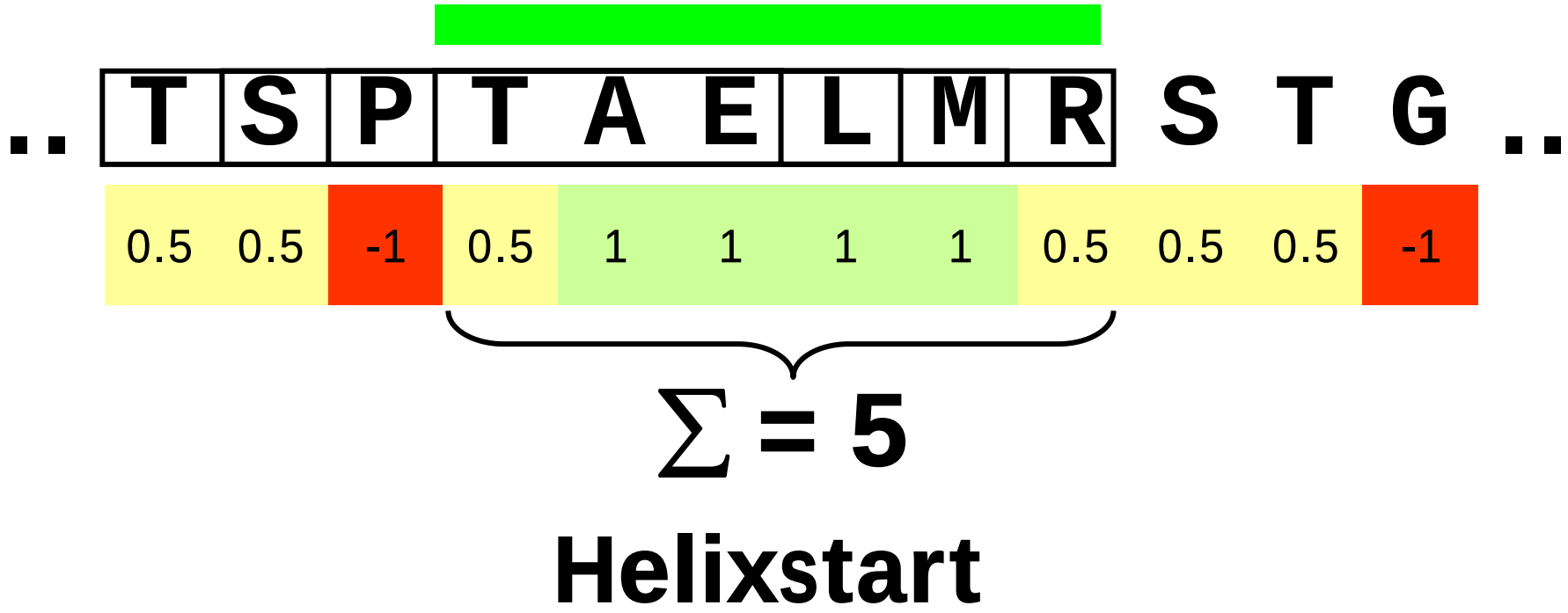
Beispiel:

.. T S P T A E L M R S T G ..

i_α	i_α	B_α	i_α	H_α	H_α	h_α	H_α	i_α	i_α	i_α	B_α
0.5	0.5	-1	0.5	1	1	1	1	0.5	0.5	0.5	-1

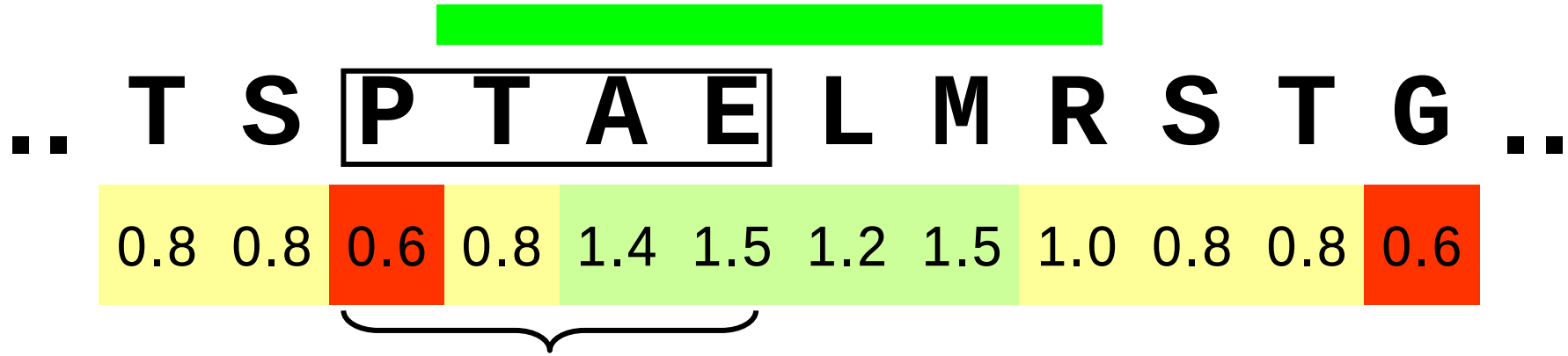
Chou-Fasman-Algorithmus II

Beispiel:



Chou-Fasman-Algorithmus II

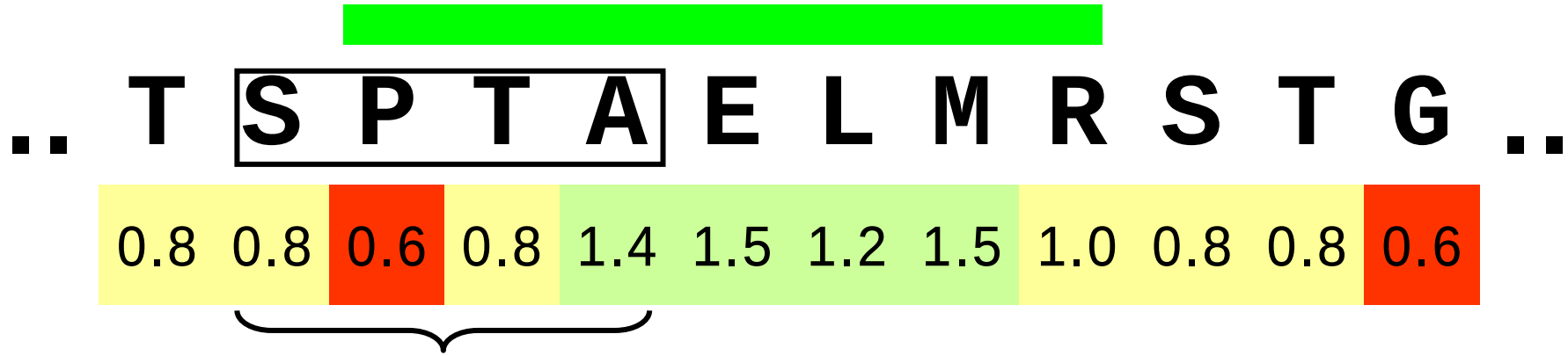
Beispiel:



Ausdehnen nach links mit 4er-Fenster
(auf den P_α-Werten!)

Chou-Fasman-Algorithmus II

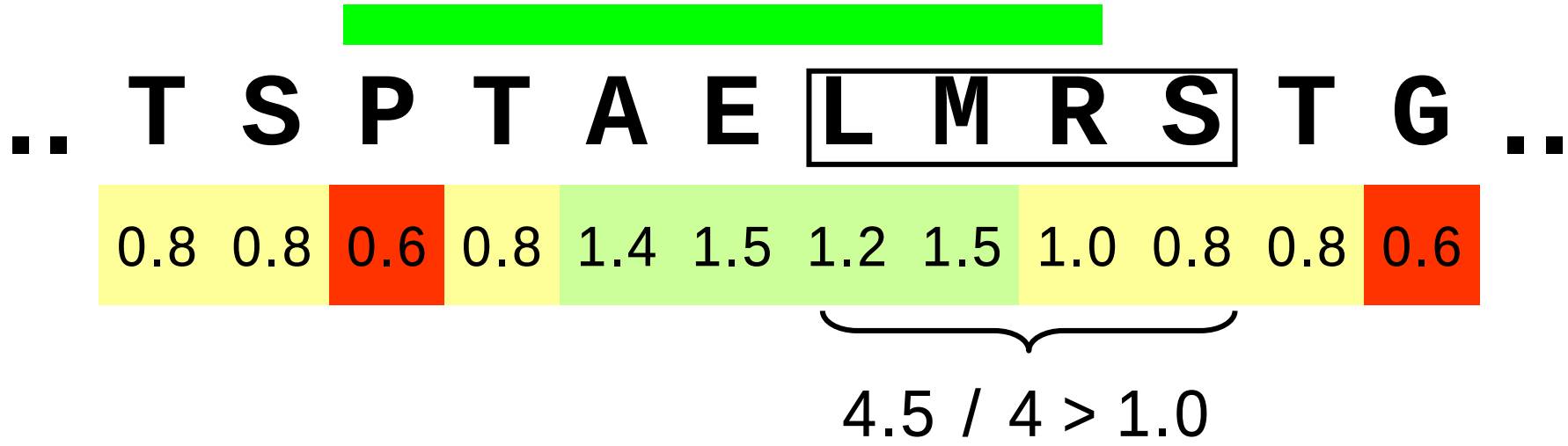
Beispiel:



Ausdehnen nach links mit 4er-Fenster
(auf den P_α-Werten!)

Chou-Fasman-Algorithmus II

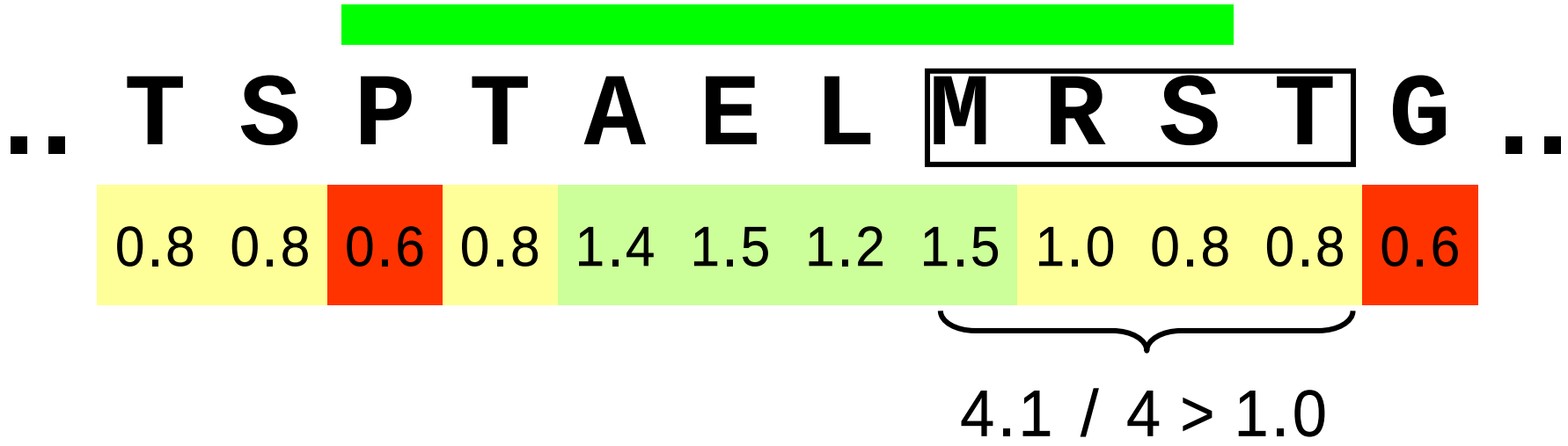
Beispiel:



Ausdehnen nach rechts mit 4er-Fenster
(auf den P_α-Werten!)

Chou-Fasman-Algorithmus II

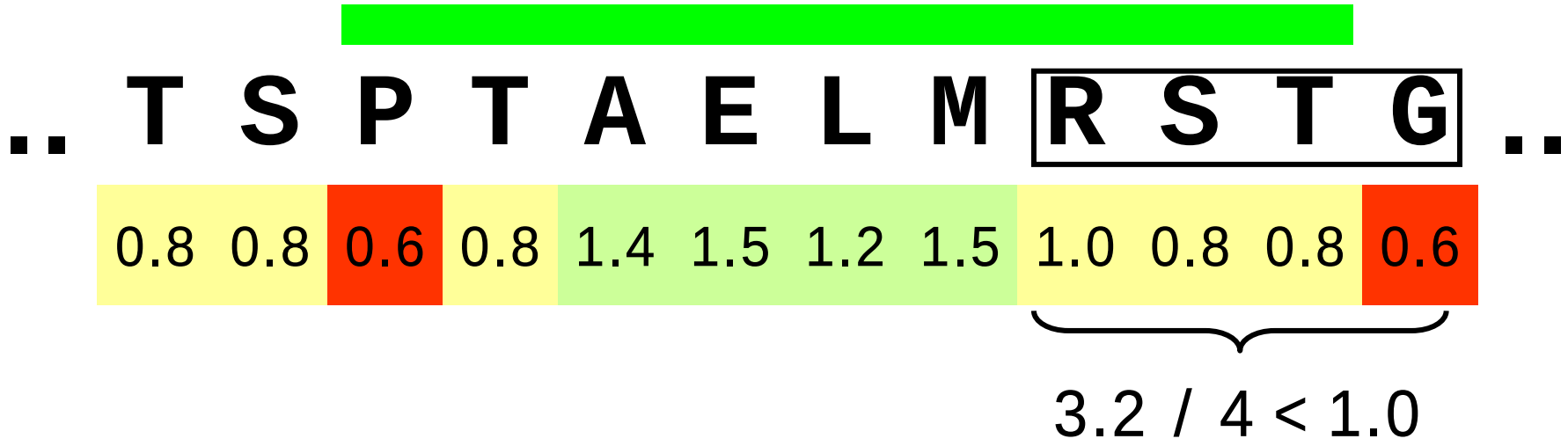
Beispiel:



Ausdehnen nach rechts mit 4er-Fenster
(auf den P_α-Werten!)

Chou-Fasman-Algorithmus II

Beispiel:



Ausdehnen nach rechts mit 4er-Fenster
(auf den P_α-Werten!)

Chou-Fasman-Algorithmus II

Beispiel:

Anschließend analog für Faltblätter

Chou-Fasman-Algorithmus I

Algorithmus (vereinfacht!)

- Ordne jeder AS der Sequenz $S = s_1 s_2 \ldots s_k$ α/β-Klassen zu

A: HELICES

- Weise jeder AS Gewicht w_i zu mit $w(H_\alpha) = w(h_\alpha) = 1$, $w(I_\alpha) = 0.5$, $w(b_\alpha) = w(B_\alpha) = -1$
- Finde Helix-Kerne
  - Fenster der Länge 6 mit $\sum w_i > 4$
- Erweitere Kerne nach links oder rechts
  - Fenster der Länge 4
  - Links oder rechts schieben bis $\sum P_\alpha^{s_i} < 4$
  - Kompatible AS des abbrechenden Peptids sind Teil der Helix

Chou-Fasman-Algorithmus II

Algorithmus (vereinfacht!)

B: STRANDS

- Weise jeder AS Gewicht w_i zu mit $w(H_\beta) = w(h_\beta) = 1$, $w(I_\alpha) = 0.5$, $w(b_\alpha) = w(B_\alpha) = -1$
- Finde Strand-Kerne
  - Fenster der Länge 5 mit
    - Drei oder mehr H_β oder h_β
    - Höchstens ein B_β oder b_β
- Erweitere Kerne nach links oder rechts
  - Fenster der Länge 4
  - Links oder rechts schieben bis $\sum P_\beta^{s_i} < 4$

Chou-Fasman-Algorithmus III

Algorithmus (vereinfacht!)

C: KONFLIKTE

- Für Bereiche die α und β markiert sind:
  - Berechne Mittelwerte P_α^{avg} und P_β^{avg}
  - Helix, falls $P_\alpha^{avg} > P_\beta^{avg}$
  - Faltblatt, falls $P_\alpha^{avg} < P_\beta^{avg}$

- Vollständiger „Algorithmus“ enthält noch weitere zusätzliche Regeln zur Zuweisung von Enden und zur Beseitigung von Konflikten

Chou, Fasman, Biochemistry (1974), 13, 222

Chou-Fasman-Algorithmus

- Online Vorhersage:

  `http://fasta.bioch.virginia.edu/fasta_www/chofas.htm`

- Vorhersagegenauigkeit sehr gering (50-60%)
- Es existieren eine Reihe verbesserter Varianten
  - Vorhersage von Turns
  - Bessere Statistiken (Chou, Fasman: 15 Proteine!)
  - Eine Variante ist z.B. **SSP**

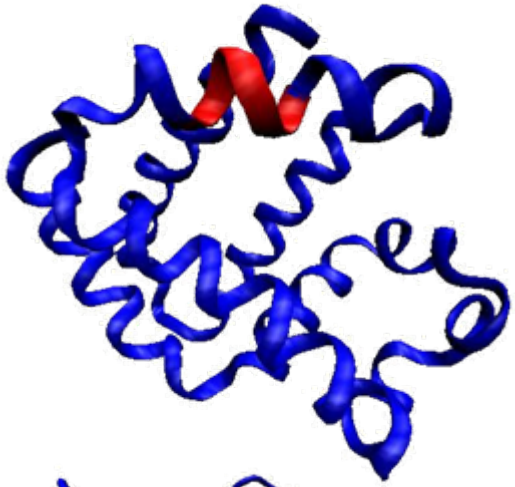
    (Solovyev, Salamov, 1991, ! *Alg. in Bioinformatics*)

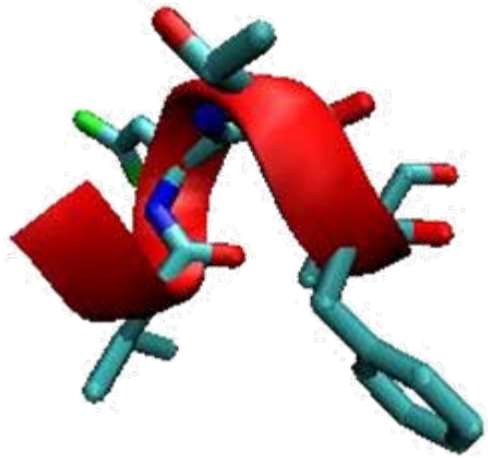
Nichtlokalität

Selbe Sequenz bildet unterschiedliche Sekundärstrukturen

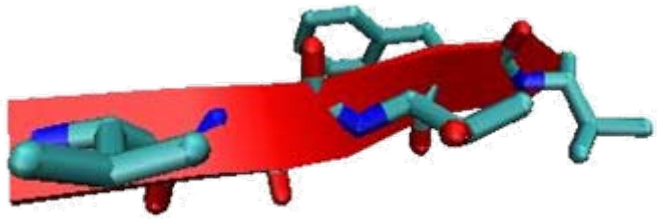
aus: Val-Asn-Thr-Phe-Val in 1ECN (80-84) und 9RSA (43-47)

1ECN



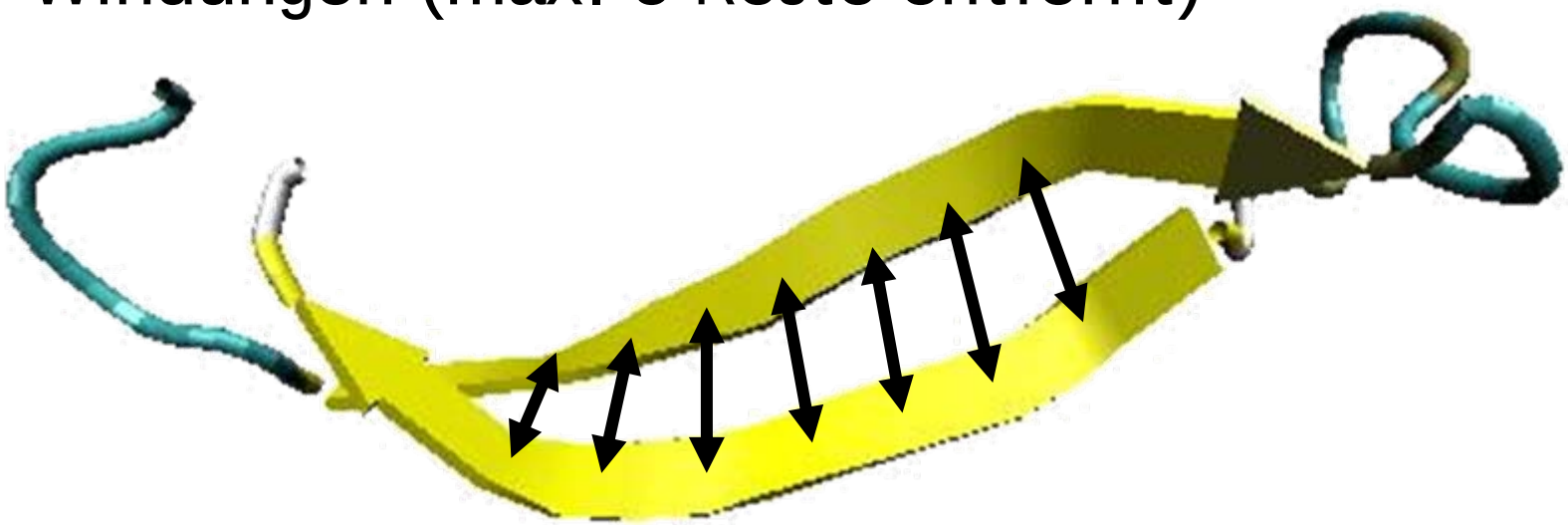


9RSA



Nichtlokalität

- Faltblätter zeigen stärkere **Nichtlokalität** als Helices: WW zwischen entfernten Sequenzbereichen notwendig um Faltblätter zu stabilisieren
- Helices: WW zwischen benachbarten Windungen (max. 5 Reste entfernt)



Methoden der 2. Generation

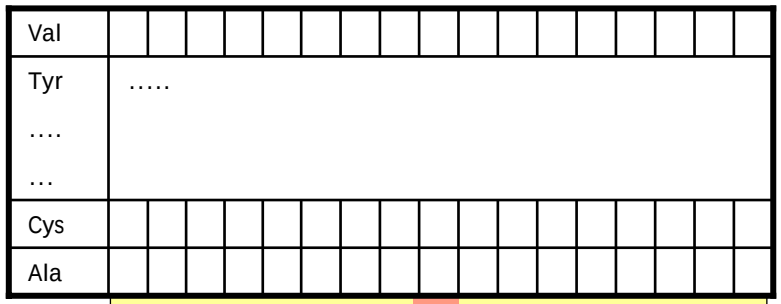
- Einbeziehung benachbarter Reste
- Verbessert Vorhersage für Helices deutlich
- Faltblätter immer noch schwierig
- Vielzahl von Methoden basierend auf
  - Künstlichen neuronalen Netzen
  - LDFs (*Linear Discrimant Function*)
  - Nächster-Nachbar-Klassifizierer
  - Support-Vektor-Maschinen
  - Hidden-Markov-Modellen

GOR-Methode

- Garnier-Osguthorpe-Robson-Methode
  - Verschiedene Varianten (GOR I – GOR IV)
- Hier: GOR IV als Beispiel einer Methode der 2. Generation
- Bezieht umliegende Sequenz mit ein (Fenster)
- Fensterlänge:
  - GOR III: 17 AS
  - Helices ca. 5-40 AS
  - Strands ca. 4-10 AS

GOR IV

- Statt P_i^j gibt es nun drei 17x20-Matrizen (PSSM)
- Je eine für die drei Klassen H, C, E
- Wert in der Matrix entspricht der Wahrscheinlichkeit einen bestimmten Rest in dieser Umgebung in der jeweiligen Sekundärstruktur zu finden



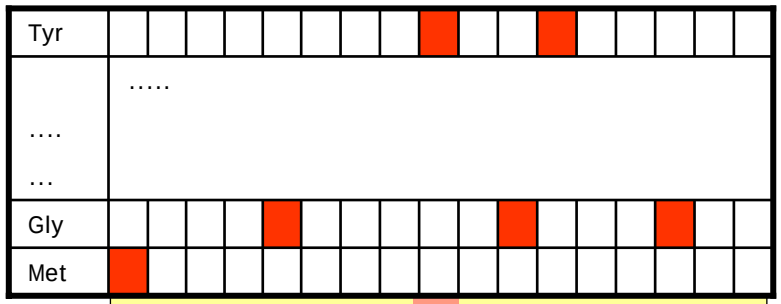
KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNF

GOR IV

- Matrix-Elemente S_α^{ij} berechnen sich zu

$$S_\alpha^{ij} = ln \frac{P(\alpha_i | aa_{i+j})}{P(\alpha_i)} \quad \text{mit} \quad -8 \leq j \leq 8$$

- Score für Position *i* ergibt sich durch Summation über das Fenster



KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNF

GOR IV

- Benötigt große statistische Basis um alle Matrixelemente mit hinreichender Genauigkeit zu berechnen
- Mehrdeutigkeiten, insbesondere an den Enden der Sekundärstrukturelemente
- Qualität der Vorhersagen: Q_3 ~ 64%
- Online verfügbar unter

  `http://abs.cit.nih.gov/`

- Mittlerweile existierte eine verbesserte Version (GOR V)

Methoden der dritten Generation

- Nur etwa 65% der Information sind lokaler Natur

  => Methoden der 1. + 2. Generation können nicht viel besser werden
- **Beobachtung**

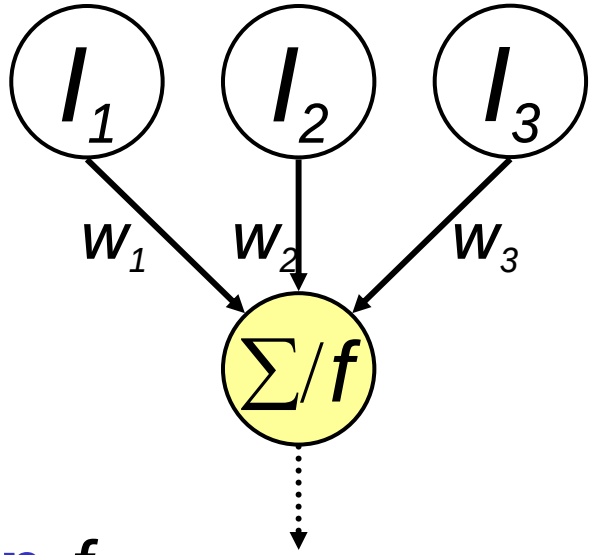
  etwa 67% der Reste einer Sequenz kann man austauschen ohne die Sekundärstruktur zu ändern
- Im Laufe der Evolution wurden viele dieser neutralen Mutationen durchprobiert
- Evolutionär verwandte (homologe) Sequenzen enthalten diese Information
  - Treten an einer Position Helixbrecher in homologen Sequenzen häufig auf, ist es unwahrscheinlich, dass dort eine Helix liegt
  - Diese Art von Information lässt sich auf einfache Art in Form von Profilen einbringen

PHD

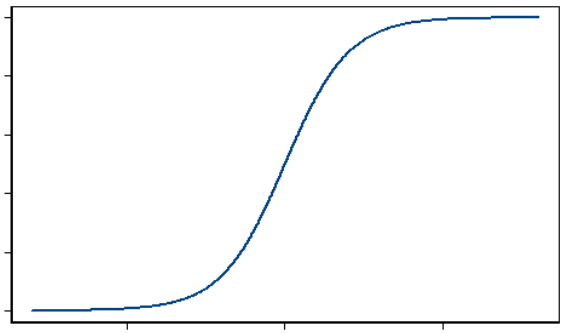
- Kurzer Rückblick auf *Alg. in Bioinformatics*
- PHD verwendet
  - künstliche neuronale Netze
  - Profile von homologen Sequenzen
- Dreischichtiges künstliches neuronales Netz (ANN)
- 1. + 2. Schicht: Mapping der Sequenz (bzw. des Profils) auf die Strukturklassen
- 3. Schicht: Mehrheitsentscheid

Wdh.: ANNs

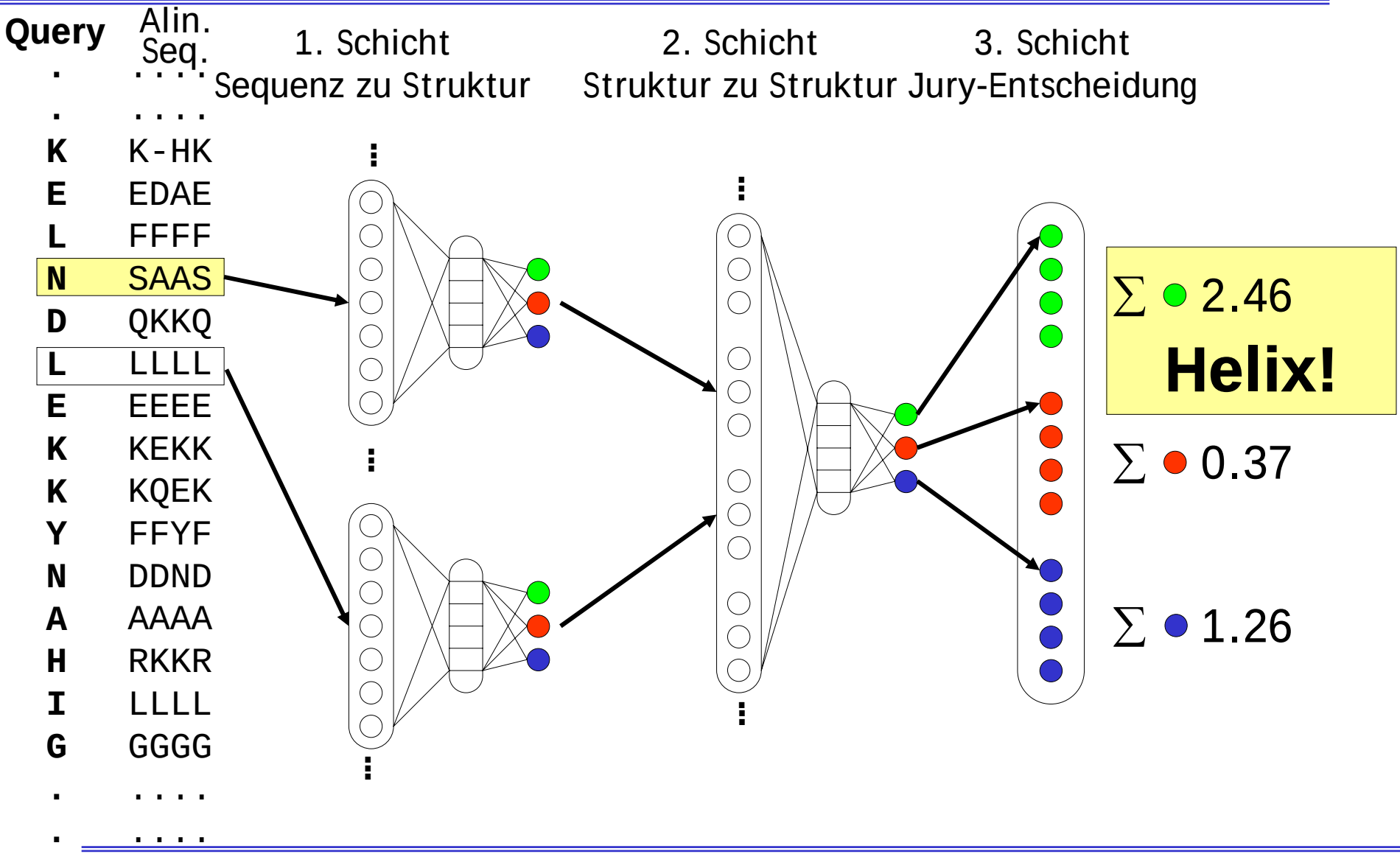
- Graph definiert die **Topologie**
- Meist in Schichten angeordnet
- Kanten sind Gewichte zugeordnet
- Eingangssignale werden **gewichtet summiert**
- (Nichtlineare) **Aktivierungsfunktion** *f*
- Häufig verwendet: *f* = Logistikfunktion



$$O = \frac{1}{1+\exp(-\sum_i w_i I_j)}$$



PHD – Struktur des ANN



PHD

- Nachbearbeitungsschritt entfernt Helices mit Länge < 3
- Training auf Kristallstrukturen/DSSP

Ergebnisse:

- Verwendung von Profilen verbessert Q_3 um etwa 6% gegenüber Einzelsequenz, Mehrheitsentscheid um ca. 2%
- Verbesserte Version PHD3 steigert Q_3 auf etwa 75%

PSIPRED

- Dreistufiger Algorithmus
  - Erzeugung eines **Profils**
  - Vorhersage mit zweistufigem **ANN**
  - **Filtern** der Vorhersagen
- Profilerzeugung
  - PSI-BLAST-Lauf (drei Iterationen) der Sequenz gegen große, nicht-redundante Proteinsequenz-Datenbank
  - PSI-BLAST-Profil (Scoring-Matrix) ist die Eingabe für die erste Schicht des ANN

Konsensus-Methoden – JPRED

- **Meta-Server:** verwendet sechs unabhängige Methoden parallel
  - NNSSP (Variante von SSP)
  - PHD
  - MULPRED (Mehrfach-Vorhersage inkl. GOR, Chou & Fasman)
  - ZPRED
  - PREDATOR
  - DSC
- Ergebnis durch Mehrheitsentscheid pro AS
- Bei Unentschieden: verwende PHD-Ergebnis!
- Genauigkeit: 73% (1% besser als PHD)

Vergleich der Algorithmen

- Regelmäßig findet CASP (*Critical Assessment of Structure Prediction*) statt
- Offener Wettbewerb zur Vorhersage unbekannter Strukturen aus der Sequenz
- Strukturbiologen stellen ihre unveröffentlichten Strukturen zur Verfügung
- Teams erhalten die Sequenz, geben innerhalb der vorgeschriebenen Zeit Modelle ab
- Im Rahmen einer Konferenz werden Ergebnisse diskutiert

Vergleich der Algorithmen

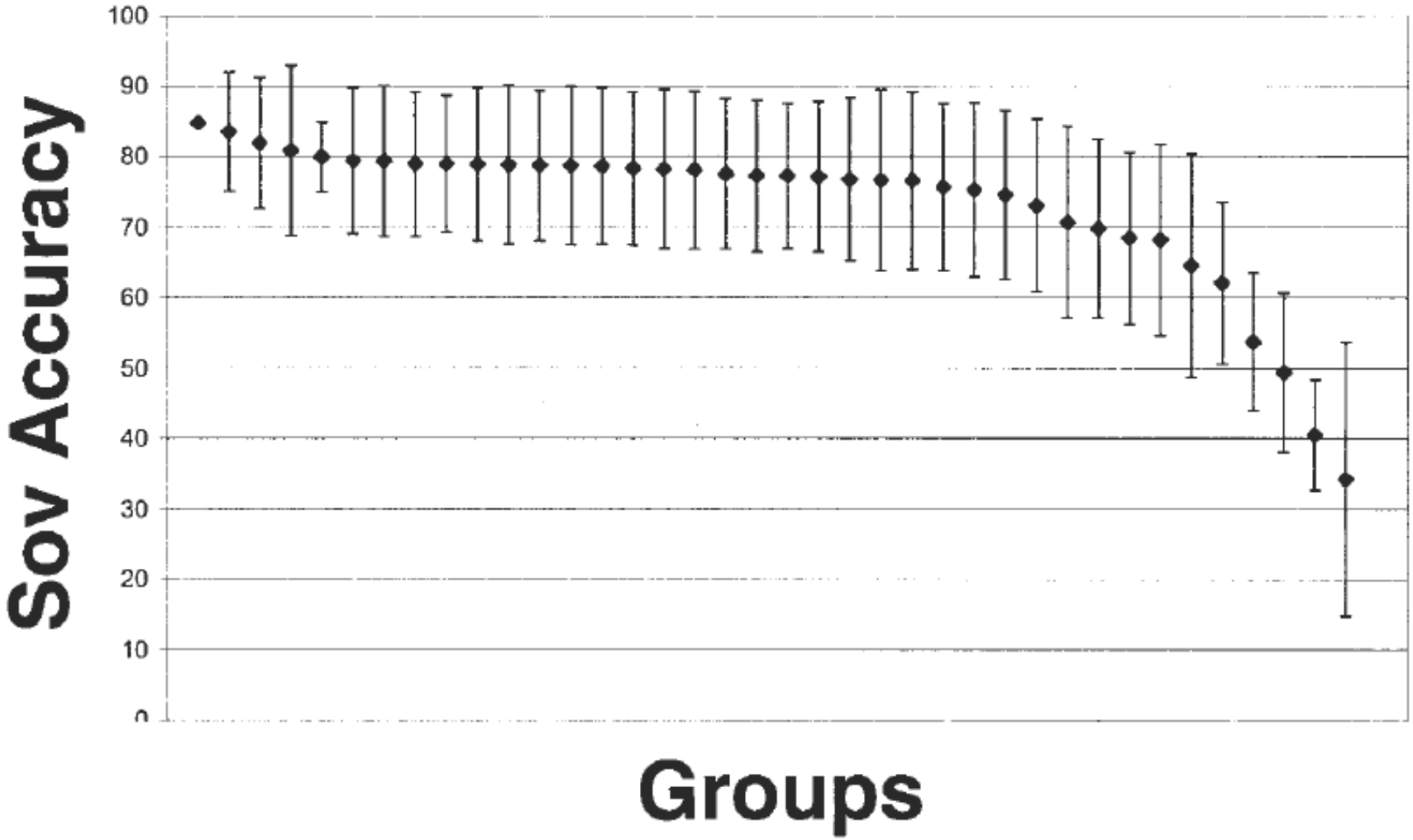
- CASP5 fand im Dezember 2002 statt
- CASP5
  - Verschiedene Disziplinen
    - Threading
    - Ab-initio-Vorhersage
    - Sekundärstrukturvorhersage
  - Sekundärstrukturvorhersage
    - 54 Zielstrukturen mit 78 Domänen
    - 38 Gruppen lieferten 2626 vorhergesagte Sekundärstruktur-Zuordnungen ab

CASP5 – Ergebnisse

- An der Spitze liegen Meta-Server
- TOP 10 hat SOV von 80%

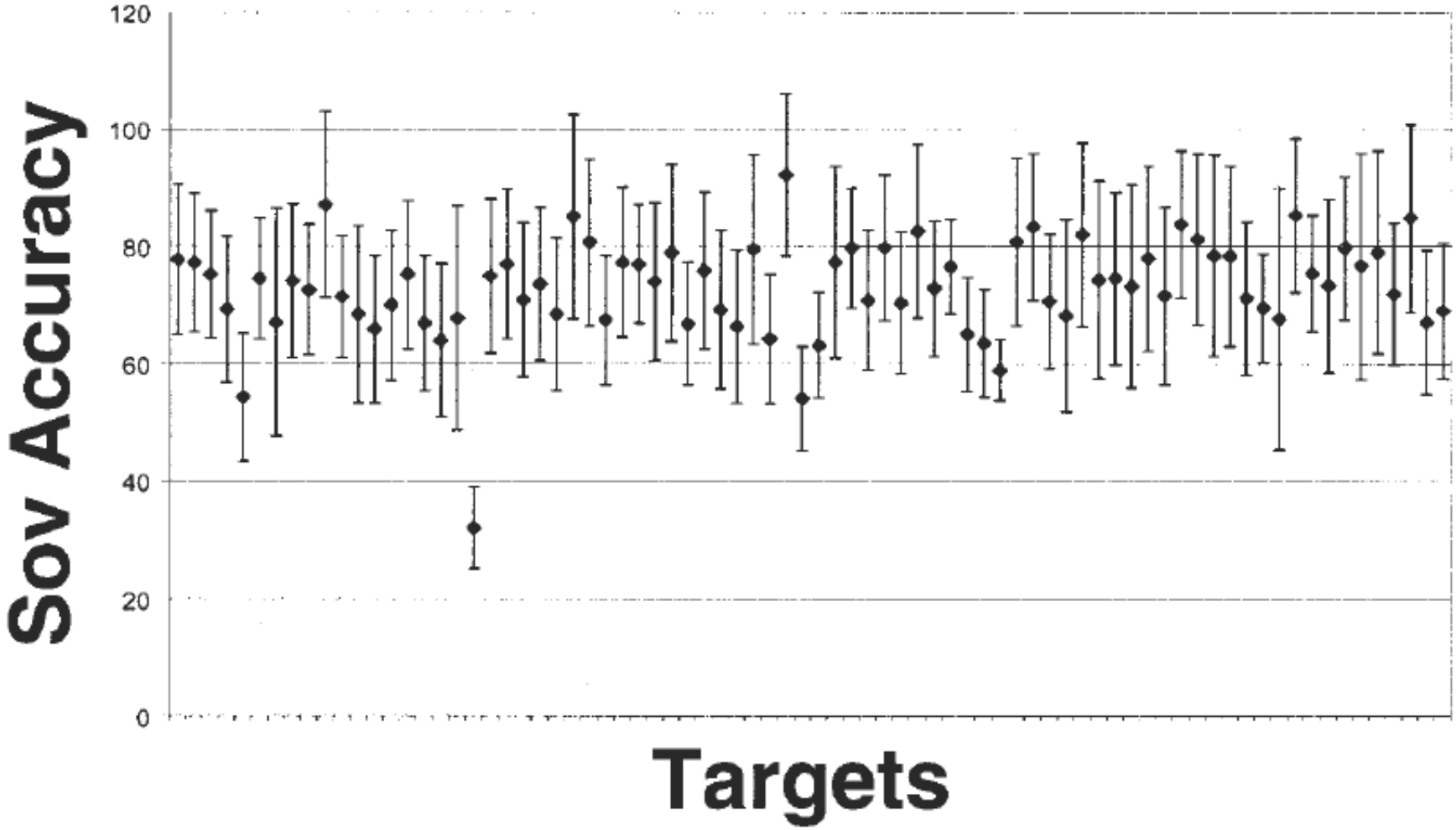
  (CASP4, 2000: 76%)
- Erfolgreiche Meta-Server basieren auf **ssPRO**, **PSIPRED** und/oder **SAM-T02** (HMM-Ansatz)
- Helices immer noch ca. 10% besser vorhergesagt als Faltblätter

CASP5 Sekundärstruktur



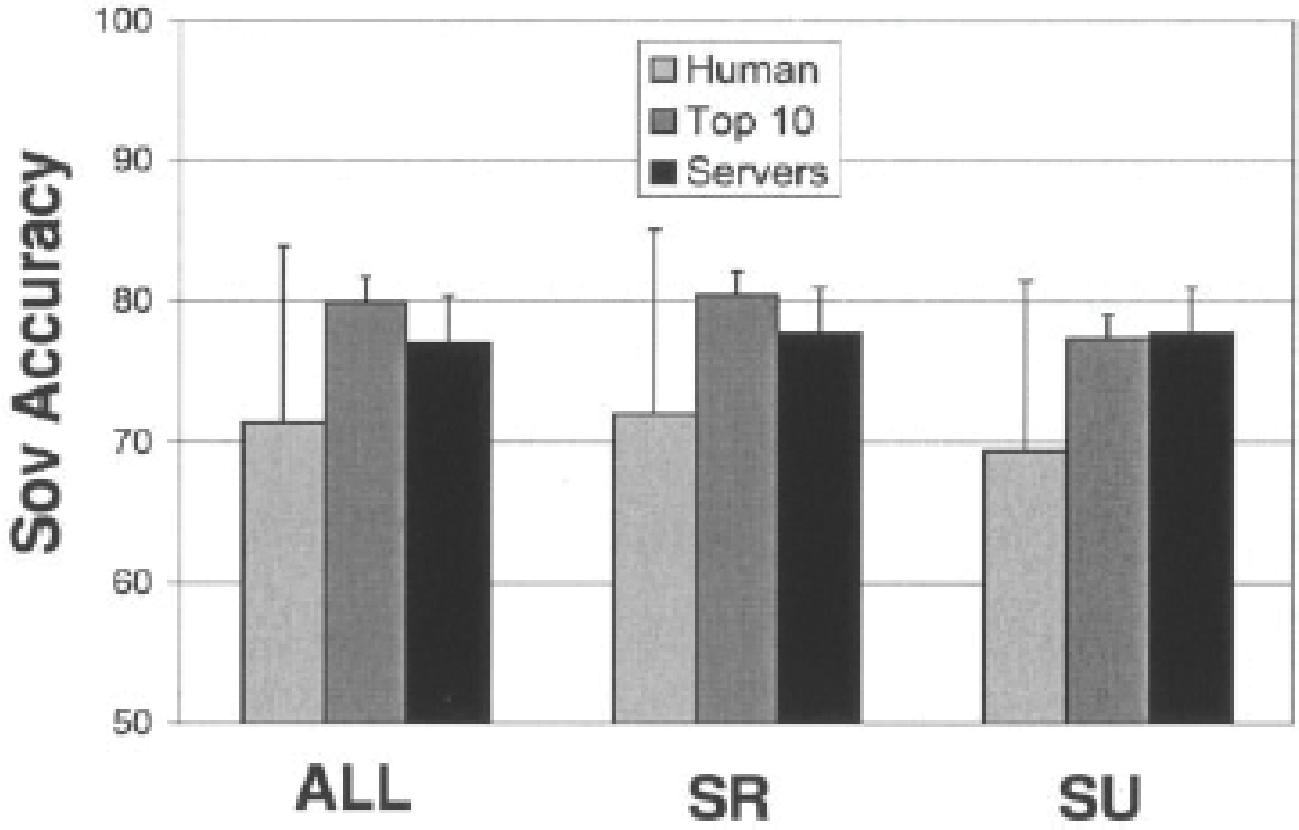
Aloy et al., Proteins: Structure, Function, Genetics (2003), 53, 436

CASP5 Sekundärstruktur



Aloy et al., Proteins: Structure, Function, Genetics (2003), 53, 436

CASP5 Sekundärstruktur



SR = Targets mit Ähnlichkeit zu bekannten Strukturen,
SU = keine Ähnlichkeit, ALL = SU + SR

Aloy et al., Proteins: Structure, Function, Genetics (2003), 53, 436

Zusammenfassung

- Sekundärstrukturvorhersage ist ein erster Schritt in der Vorhersage der Tertiärstruktur
- Gute Methoden betrachten große Sequenzabschnitte und beziehen evolutionäre Information mit ein
- Meta-Server sind leicht besser als einzelne Algorithmen
- Man kann Vorhersagegenauigkeiten (Q_3) von 75-80% erwarten

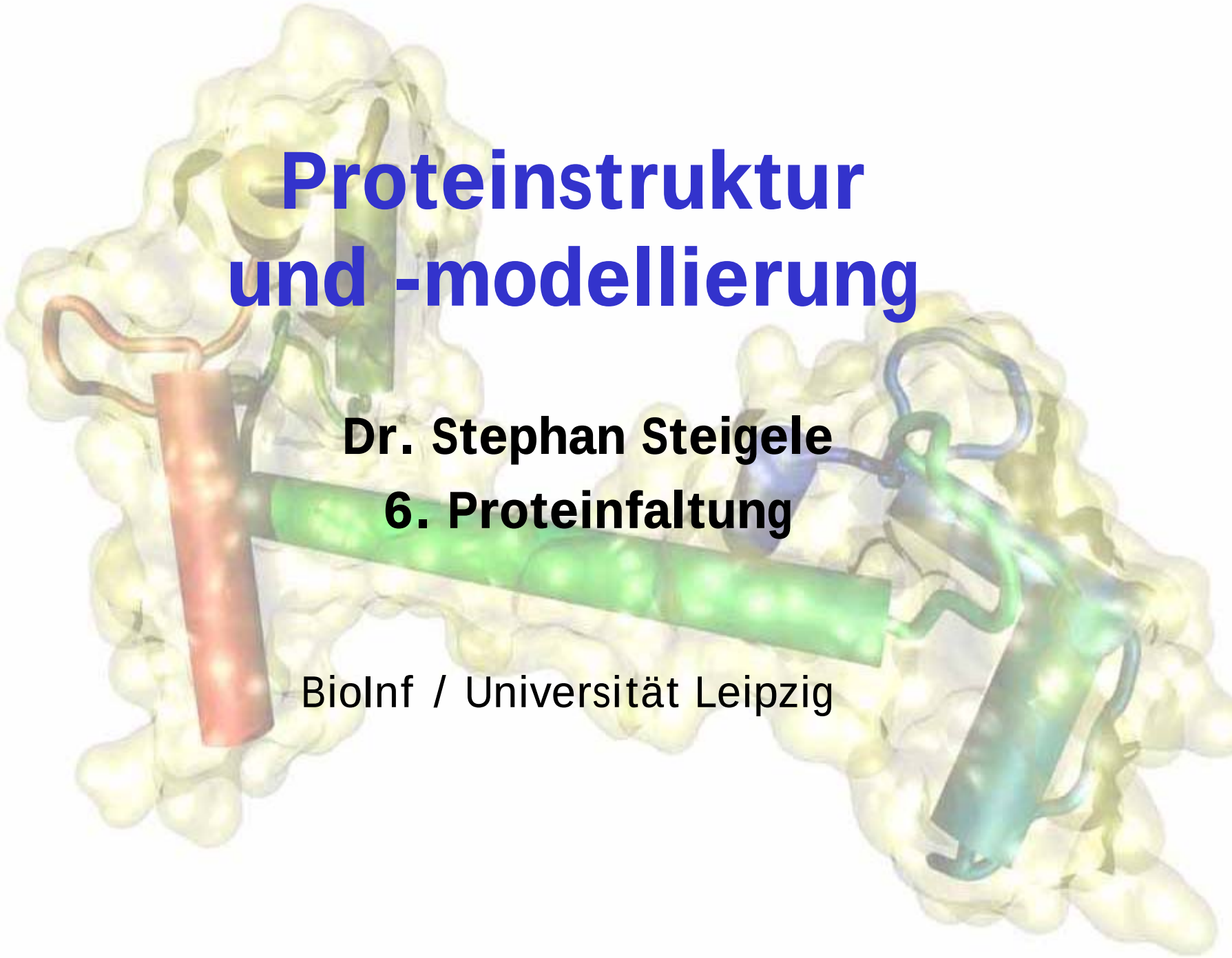
Literatur

Sekundärstrukturvorhersage:

- Burkhard Rost: Prediction in 1D, In: Structural Bioinformatics (Hrsg.: P. E. Bourne, H. Weissig), Wiley, 2003

- Ralf Zimmer, Thomas Lengauer: Structure Prediction, Chapter 5 in T. Lengauer (Hrsg.): Bioinformatics: From Genomes to Drugs, Wiley, 2002

- Publikationen zu den einzelnen Methoden: siehe Website zur Vorlesung



Proteinstruktur und -modellierung

Dr. Stephan Steigele

6. Proteinfaltung

BioInf / Universität Leipzig

Gliederung

- Problemdefinition
- **Biochemie**
  - Protein-Biosynthese
  - Proteinfaltung: beteiligte Enzyme, Chaperone
- **Biophysik**
  - Thermodynamik
  - Kinetik
- **Modelle** für die Faltung
  - Gittermodelle (Komplexität)
  - Molekulardynamik

Proteinstruktur – Proteinfaltung

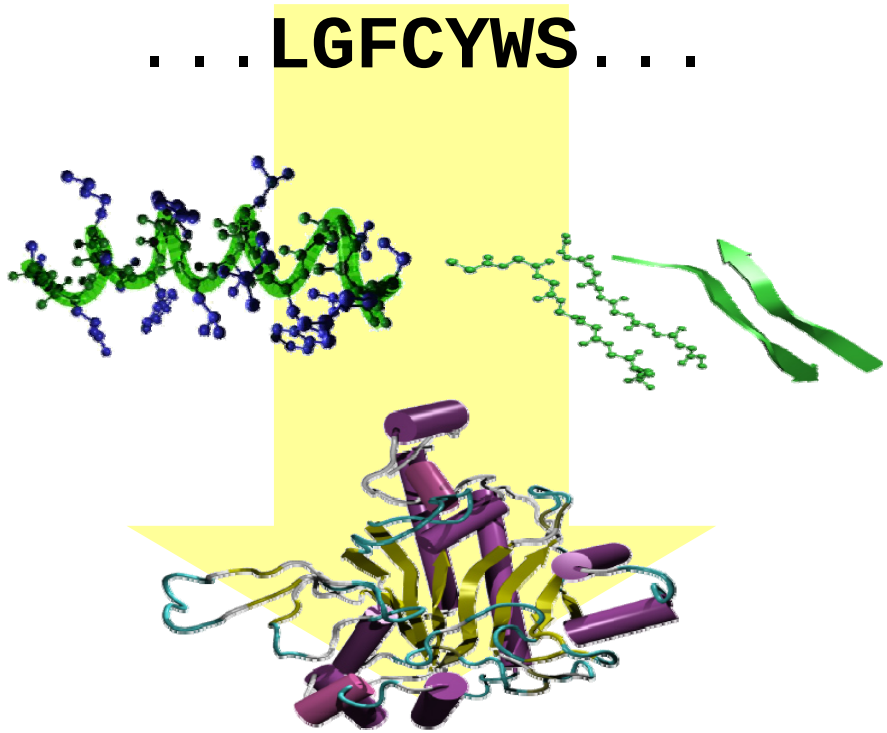
Primärstruktur

Sekundärstruktur

Tertiärstruktur

Q

. . . LGFCYWS . . .



Woher weiß das Protein, wie es sich zu falten hat?

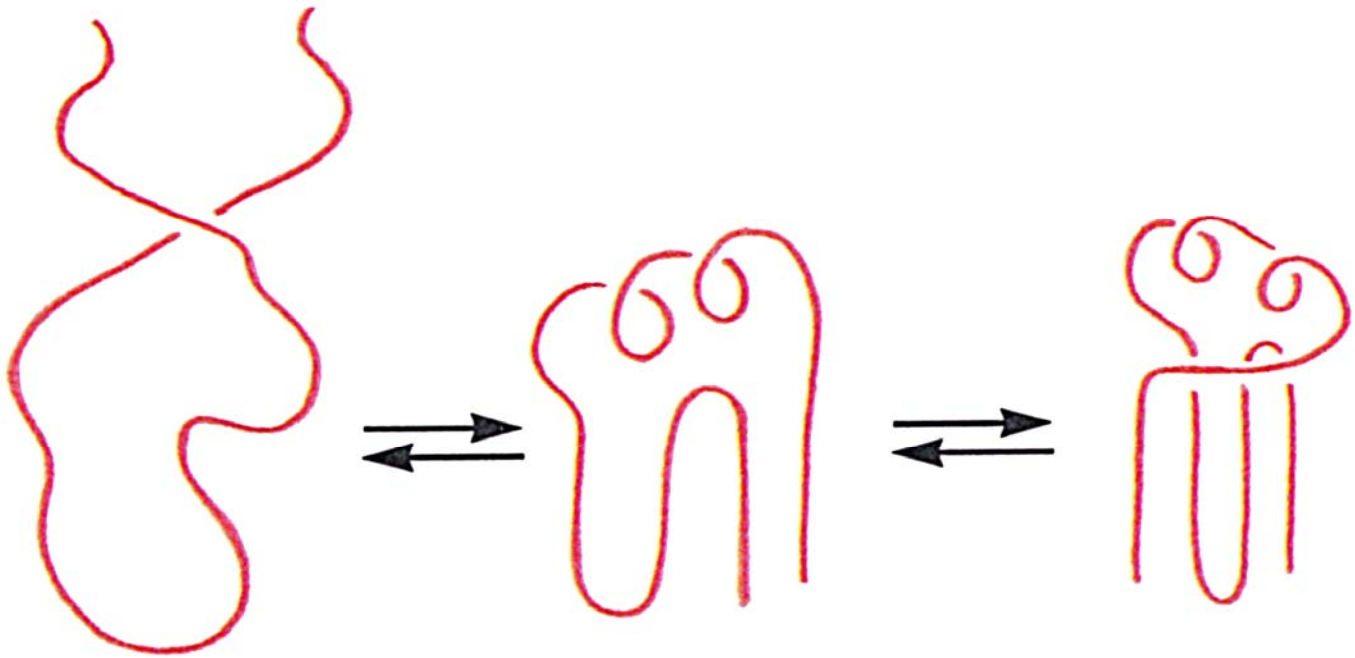
Faltungsproblem

Schlüsselfragen:

- Wie kommt das Protein von seiner Sequenz zu seiner Struktur?
- Wie können wir aus der Sequenz die Struktur vorhersagen?
- **Anwendungen**:
  - Vorhersage der Struktur
  - Vorhersage der Funktion

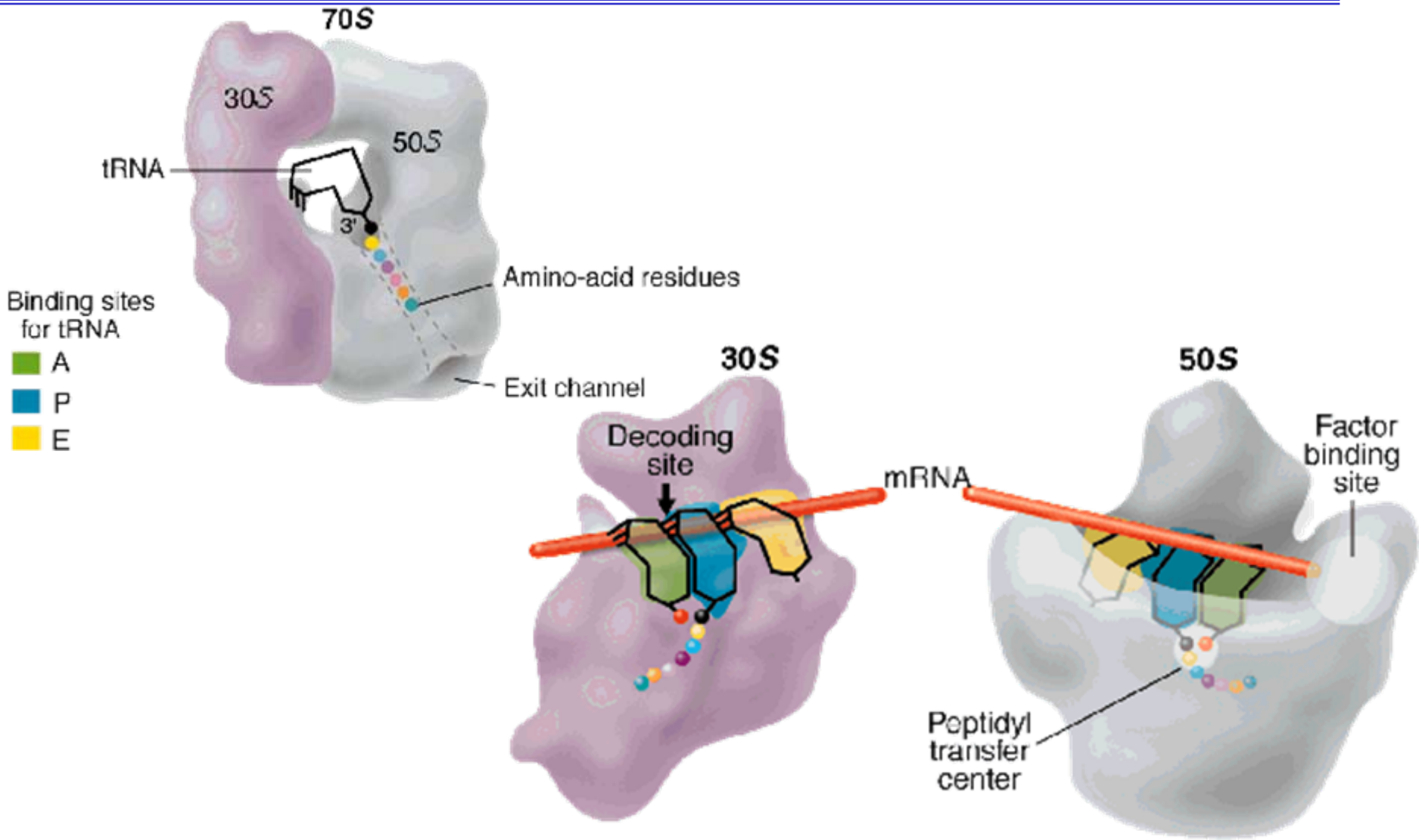
Proteinfaltung

Definition: Übergang vom ungeordneten (entfalteten) Zustand zum wohldefinierten (nativen, gefalteten) Zustand



Faltung ist reversibel (z.B. Denaturierung durch Erhitzen!)

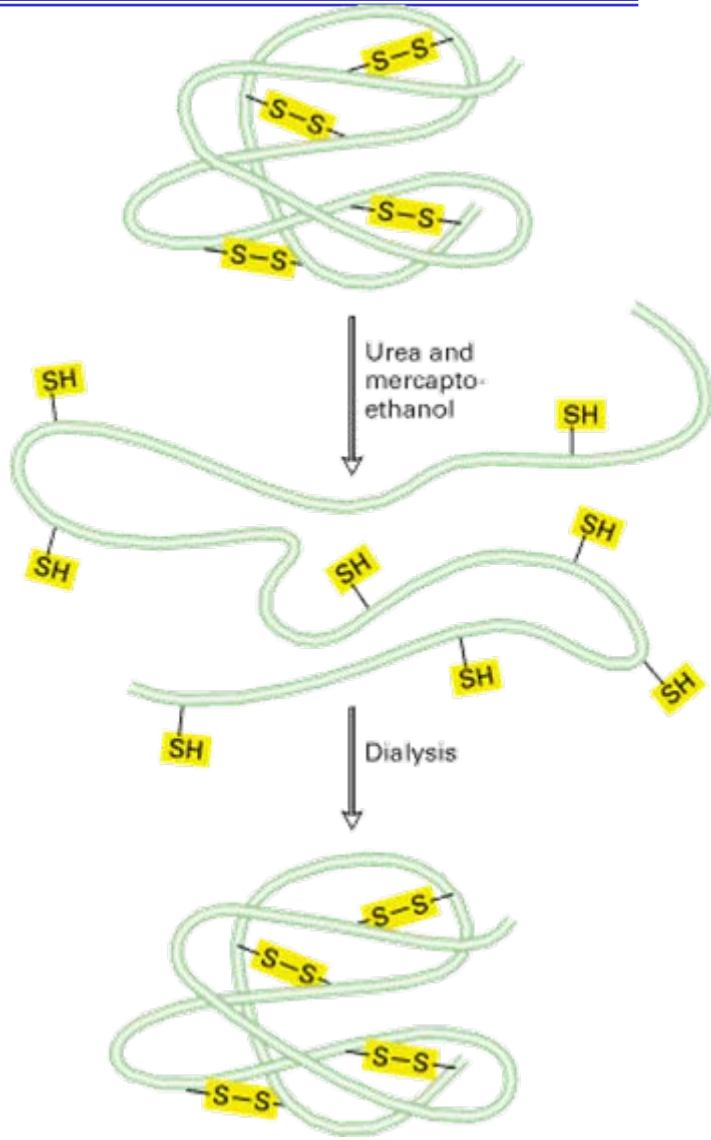
Protein-Biosynthese



Reversibilität – Anfinsen

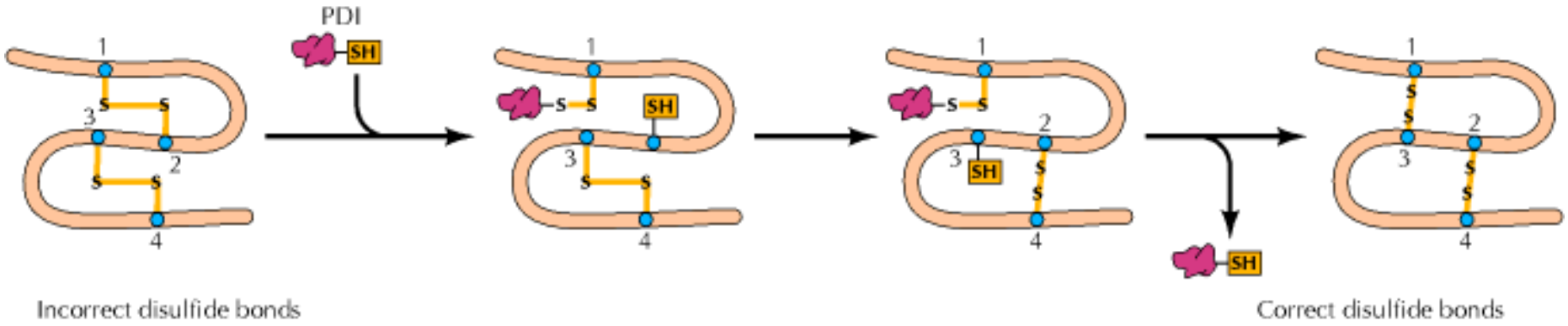
- Christian B. Anfinsen führte in den 60ern wichtige Experimente zur Faltung durch
- **Mercaptoethanol** (EtSH) vermag die Schwefelbrücken reduktiv zu lösen
- **Harnstoff** bei niedrigem pH schwächt die Wechselwirkungen innerhalb des Proteins

  => Entfaltung
- Entfernen der beiden Reagenzien (Dialyse) führt meist zum nativen Zustand

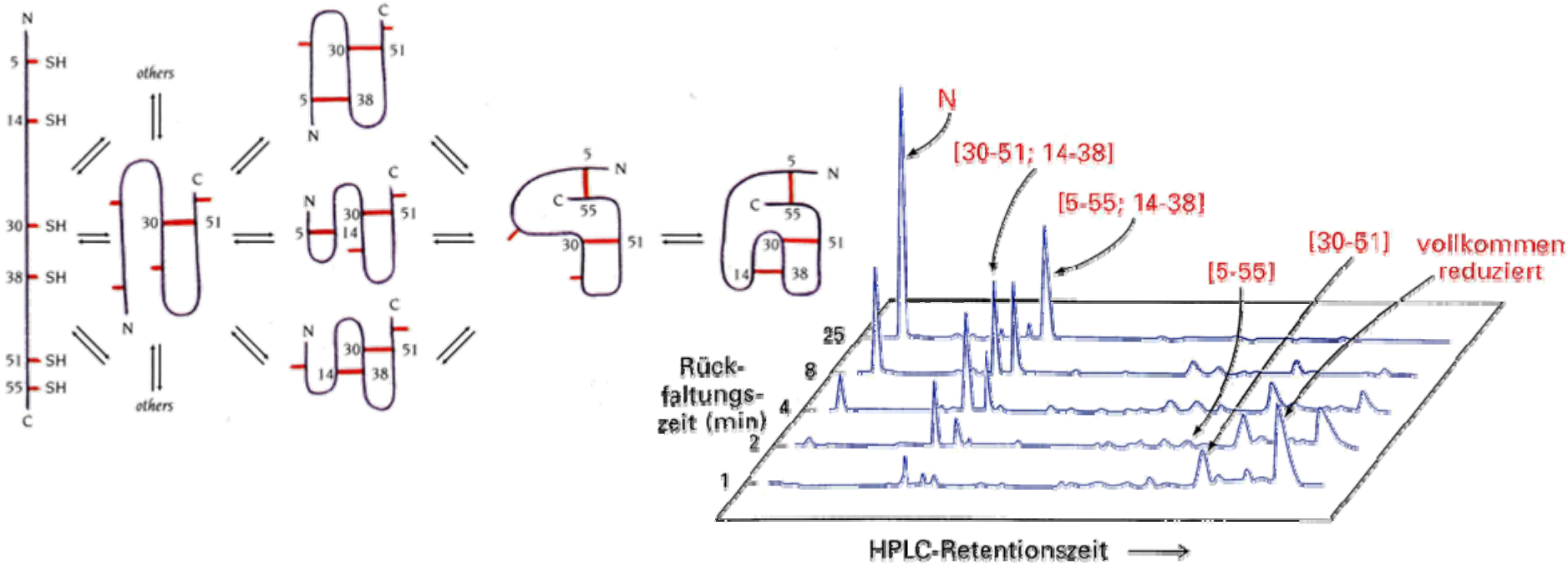


Schwefelbrücken

- Ausbildung von S-Brücken wird enzymatisch katalysiert
- In Bakterien erfolgt die Ausbildung im periplasmatischen Raum durch eine Familie von Enzymen (Dsb)
- In Eukaryota erfolgt die Ausbildung der S-Brücken im ER durch die Protein-Disulfid-Isomerase (PDI)



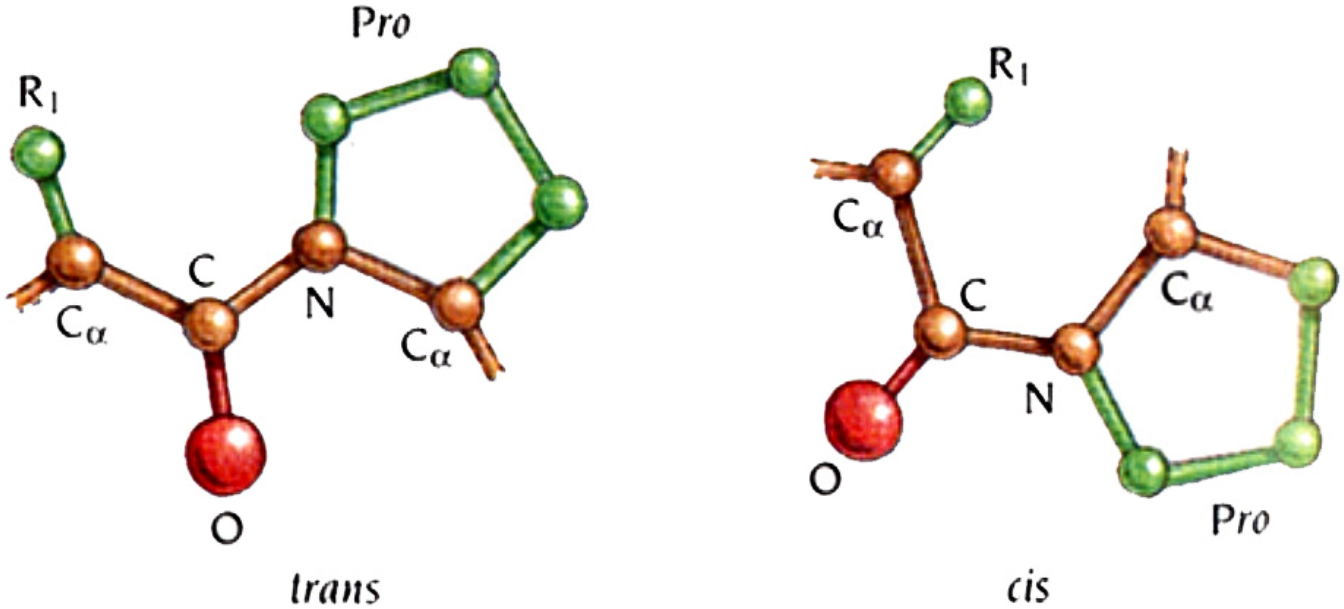
Ausbildung von S-Brücken in BPTI



- BPTI bildet im korrekt gefalteten Zustand drei SS-Brücken aus
- Intermediär werden Zustände mit 1-2 S-Brücken ausgebildet
- Der korrekt gefaltete Zustand hat die niedrigste Energie

Darby, Morin, Talbo, Creighton, J Mol Biol. (1995), 249, 463-77
Branden, Tooze: Introduction to Protein Structure, S. 96

Prolin-Isomerisierung



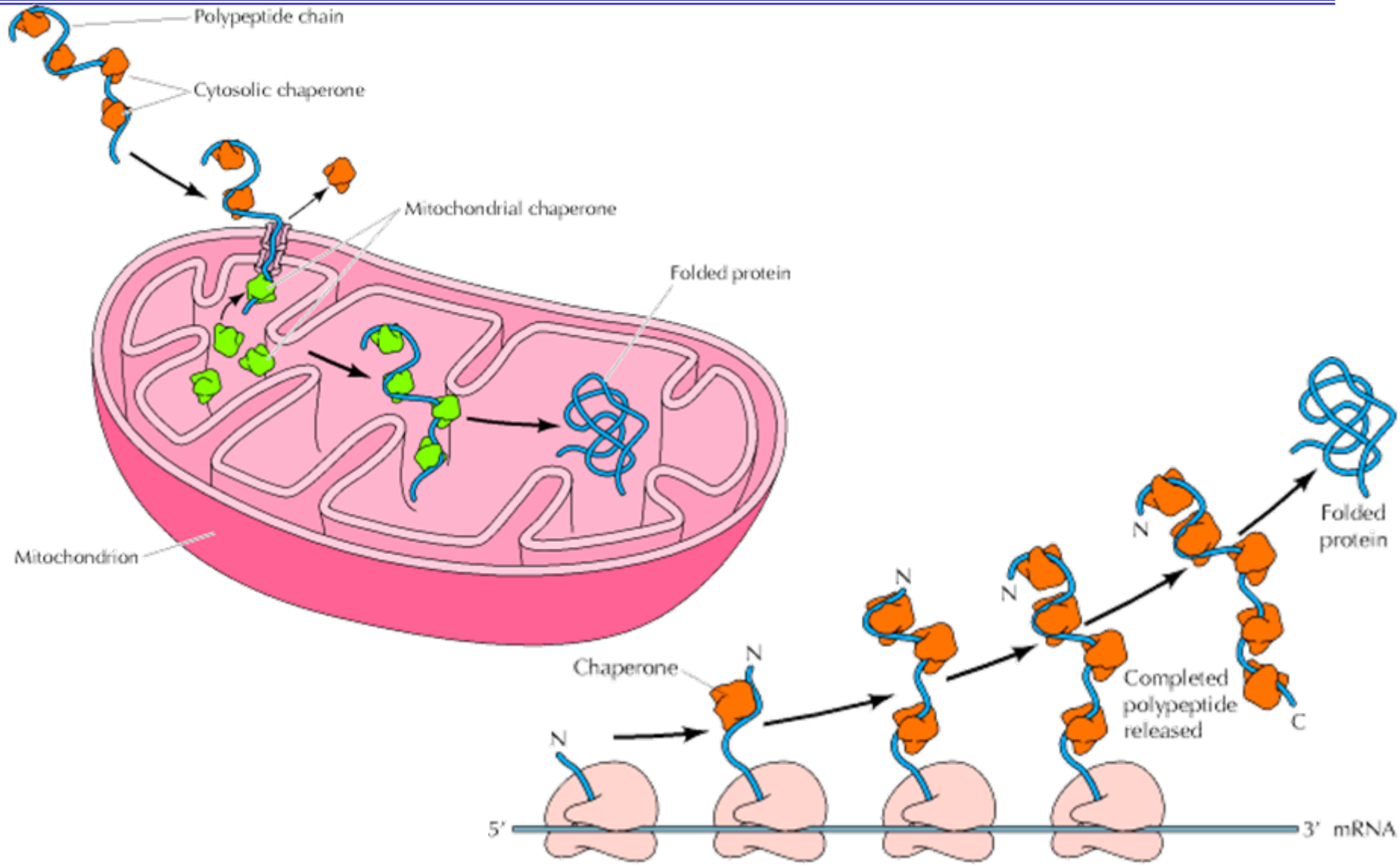
- *cis-trans*-**Isomerisierung** von Pro kann als langsamster Schritt die Faltung dominieren
- **Peptidyl-Prolyl-Isomerasen** beschleunigen die Isomerisierung dramatisch

Chaperone

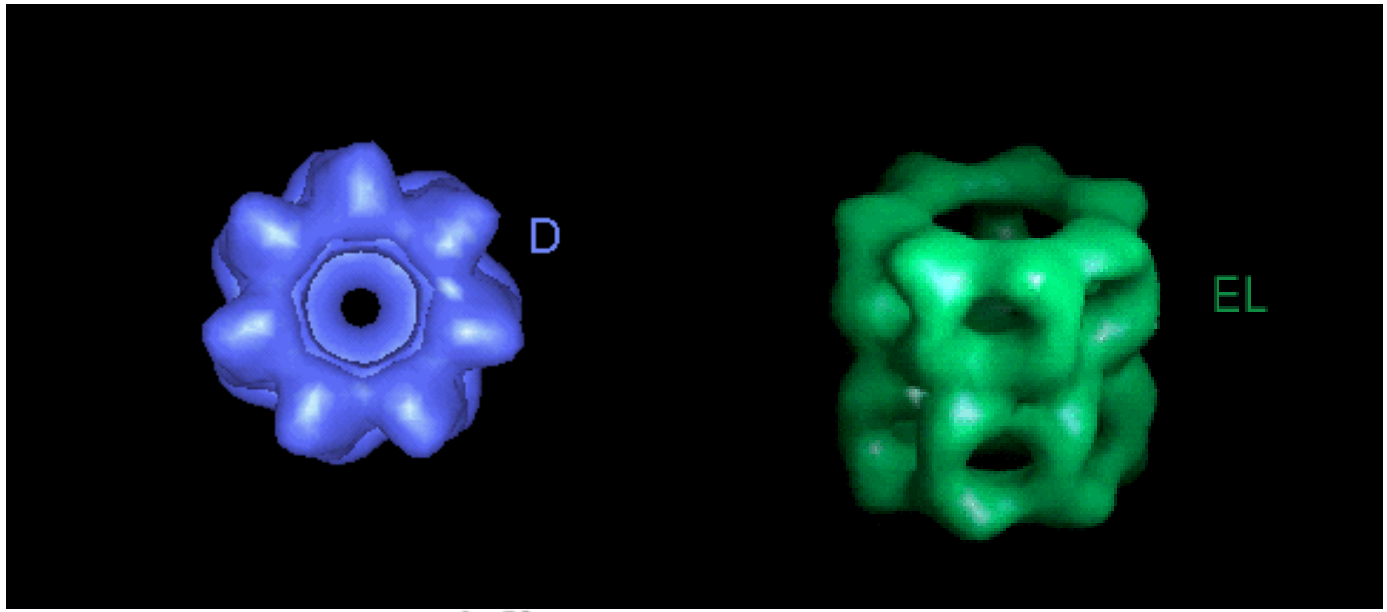
- **Chaperone** sind Proteine, die bei der Faltung assistieren
- Sie werden insbesondere bei thermischem Stress stärker exprimiert um teilweise denaturierte Proteine wieder zurückzufalten

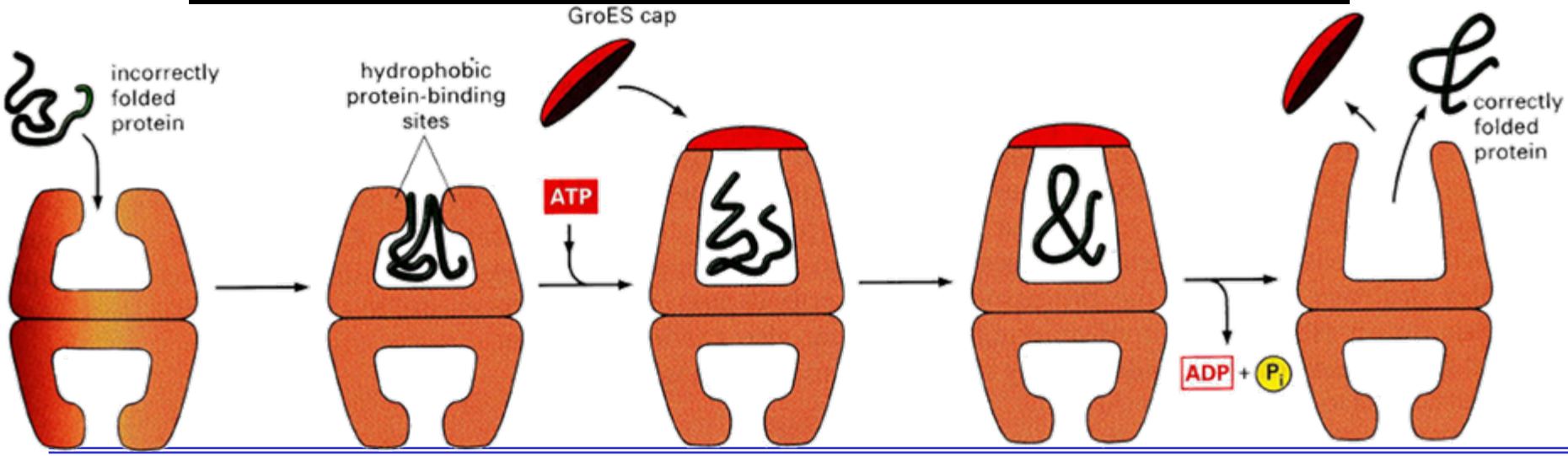
  => **Hitzeschockproteine**
- Stabilisierung und Faltung erfolgt überwiegend durch hydrophobe Wechselwirkung des Chaperon-Innenraums mit den Proteinen

Chaperone


Polypeptide chain
Cytosolic chaperone
Mitochondrial chaperone
Folded protein
Mitochondrion
Chaperone
N
Completed polypeptide released
C
5'
3' mRNA
Folded protein

Chaperone – GroEL/GroES





Einfache Proteine

- Aufgrund der Kompliziertheit der Vorgänge wollen wir in der Folge die Effekte von
  - Peptidyl-Prolyl-Isomerasen
  - Disulfid-Isomerasen
  - Chaperonen

  außer Acht lassen

=> Beschränkung auf kleine Eindomänen-Proteine

Proteinfaltung – Übersicht

- Problemdefinition
- Biochemie
  - Protein-Biosynthese
  - Proteinfaltung: beteiligte Enzyme, Chaperone
- **Biophysik**
  - **Thermodynamik**
  - **Kinetik**
- Modelle für die Faltung
  - Gittermodelle (Komplexität)
  - Molekulardynamik

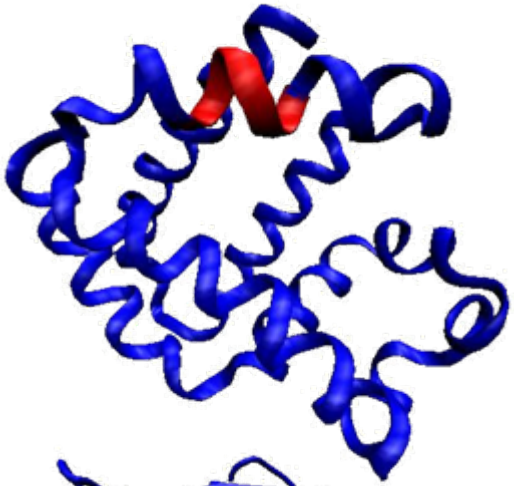
Biophysik der Proteinfaltung

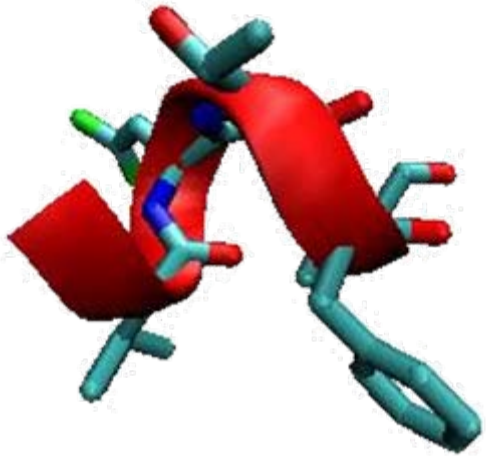
- **Eine Sequenz = eine Tertiärstruktur**
- Aber
  - gleiche Teilsequenzen kommen in unterschiedlichen Strukturen vor
  - Unterschiedliche Sequenzen haben die gleiche Tertiärstruktur (z.B. Punktmutation)
  - Kleine Änderungen in der Sequenz können komplett unterschiedliche Tertiärstruktur bewirken
- Faltungsproblem ist **nichtlokal**!

Proteinstruktur ist nichtlokal!

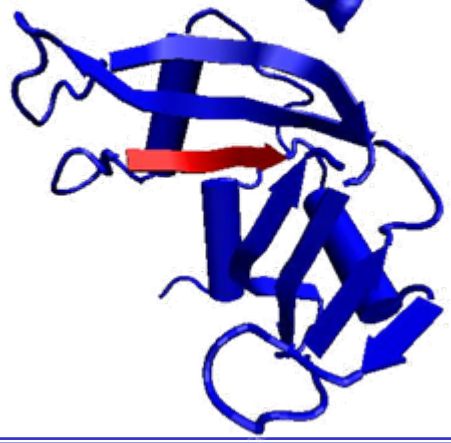
Identische Sequenzen bilden unterschiedliche Sekundärstrukturen aus, je nach **Umgebung**!

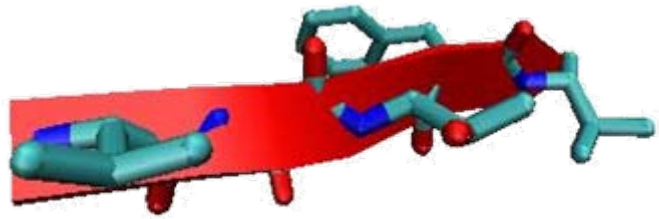
1ECN





9RSA



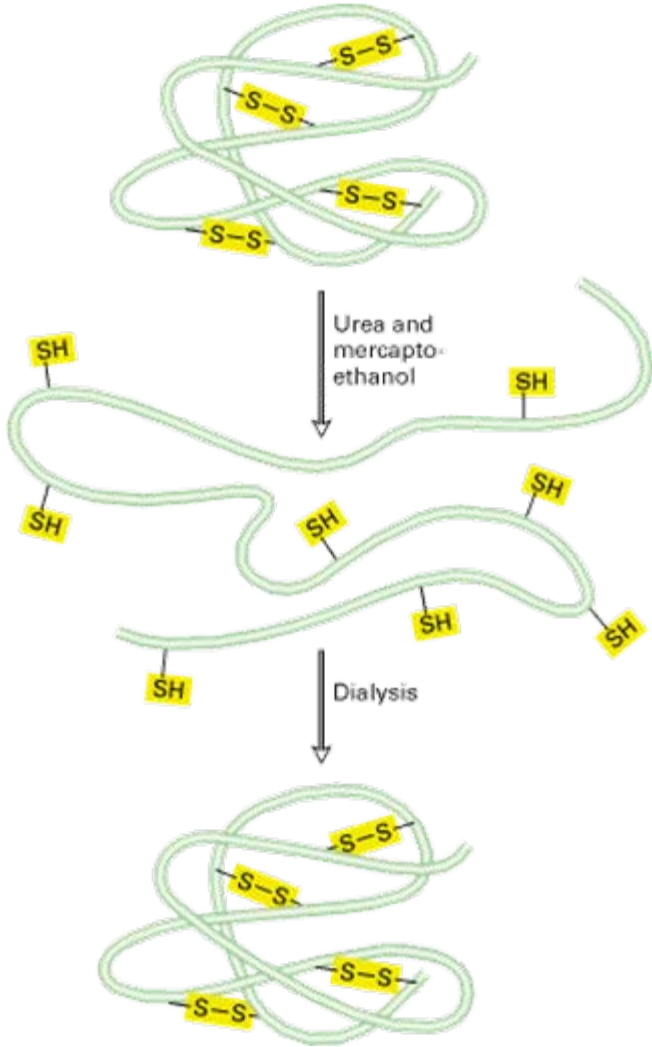


Thermodynamische Hypothese

Anfinsens Experiment führte zur Formulierung der

Thermodynamischen Hypothese:

Faltung erfolgt zum thermodynamisch günstigsten Zustand



Übergangszustände sind energetisch ungünstiger, müssen während der Faltung aber durchlaufen werden

Thermodynamik der Faltung

Thermodynamische Hypothese: ΔG bestimmend

$$\Delta G = \Delta H - T\Delta S$$

- Gefalteter Zustand muss energetisch günstiger sein als alle entfalteten Zustände ($\Delta G < 0$)
- **Globales Minimum** sollte deutlich von weiteren Minima getrennt sein
- **Entropie** muss eine Rolle spielen (thermische Denaturierung!)

Wechselwirkungen

WW	Effekt	Wichtigkeit
Elektrostatik	Destabilisierend	Gering
Salzbrücken	Stabilisierend	Mittel
Konf.-Entropie	**Destabilisierend**	**Groß**
H-Brücken	Stabilisierend	Mittel
van der Waals	Stabilisierend	Gering
Hydrophobe WW	**Stabilisierend**	**Groß**

Konformationsentropie

- **Entfalteter Zustand**
  - Viele mögliche Konformationen („*random coil*“)
  - Seitenketten können praktisch alle möglichen Zustände annehmen
  - Viele zugängliche Freiheitsgrade

  => hohe Entropie
- **Gefalteter Zustand**
  - Eine einzige Konformation des Rückgrats
  - Viele der Seitenketten-Torsionen sind „eingefroren“

  => niedrige Entropie

Konformationsentropie

- Faltung entspricht **Verlust an Freiheitsgraden**
- $\Delta S_{F\ vs.\ u} = \Delta S_F - \Delta S_U < 0$
- Konsistent mit thermischer Denaturierung

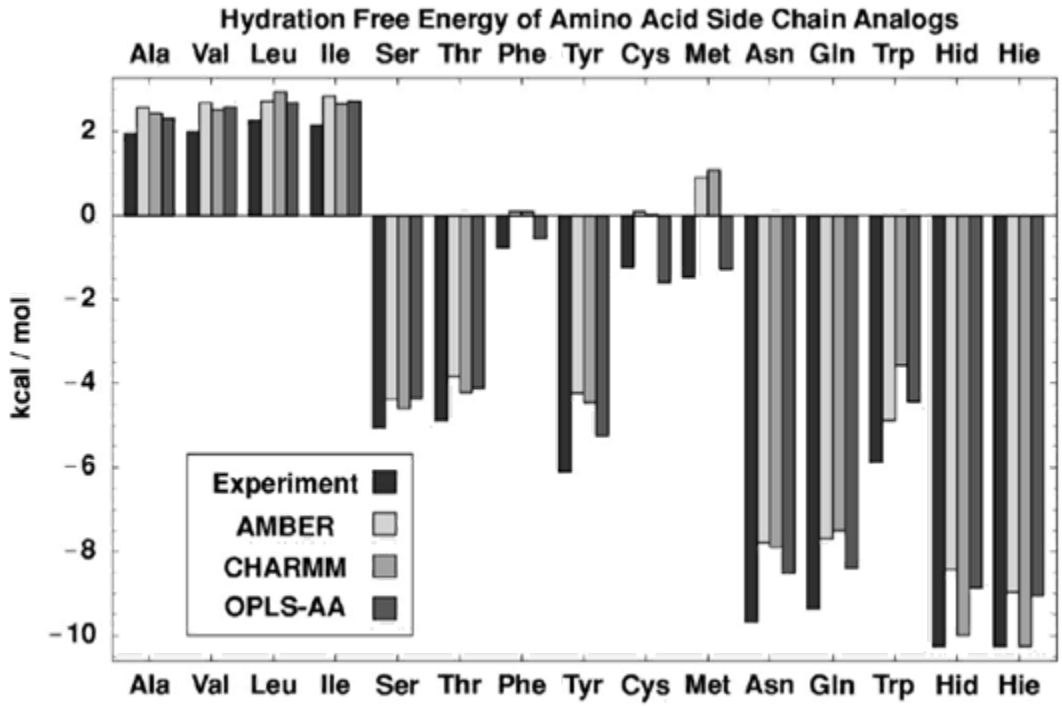
  => $-T\Delta S_{F\ vs.\ u}$ positiv, daher für hohe T Entfaltung begünstigt
- **Einfache Modellrechnung:**
  - Lysozym (164 aa)
  - pro AS ~ 10 Konformationen
    - $N_U = 10^{164}$ mögliche Konformationen (ungefaltet)
    - $N_F = 1$ (eine einzige native Konformation)

  $\Delta S_{F\ vs.\ u} = R \ln (N_F/N_U) \sim -490\ R$ (KORREKT?)

  => $\Delta G_{conf} = -T\Delta S_{conf} \sim 1200$ kJ/mol (exp.: -2188 kJ/mol)

Hydrophober Effekt

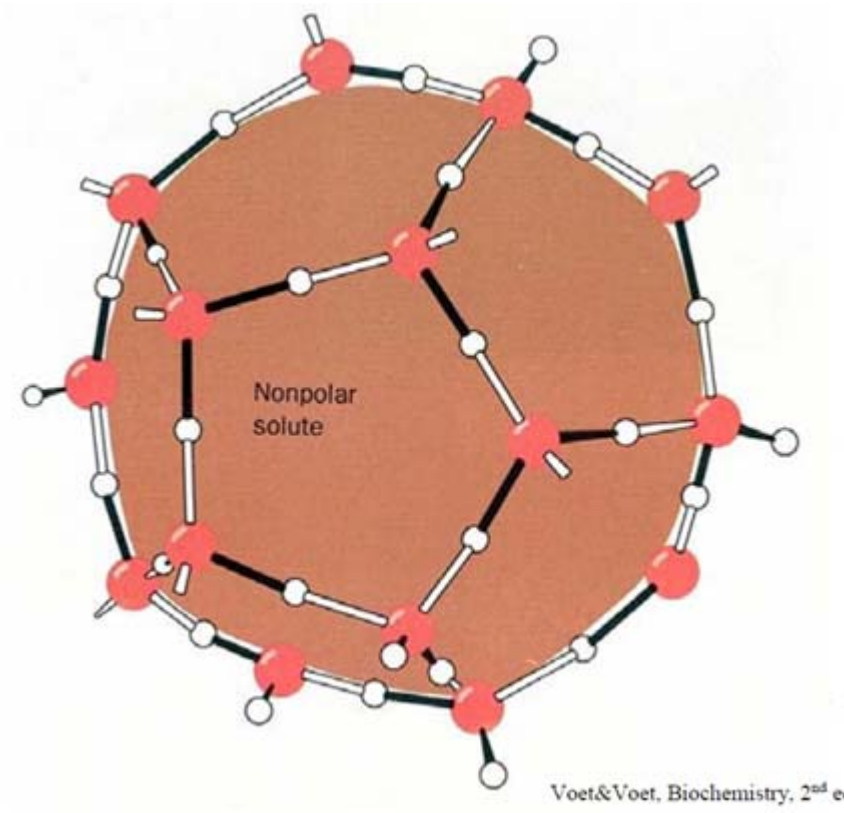
- **Experiment**
  - Transfer von Seitenkette von Vakuum in Wasser
  - Freie Hydratationsenergien
- Negativ für geladene polare Aminosäuren
- Positiv für unpolare Aminosäuren



Hydrophober Effekt

- Entropie stammt aus dem Lösemittel
- Unpolare Seitenketten vermögen keine H-Brücken zu bilden
- Wasser nimmt geordnete „**Käfigstrukturen**“ an
- **Verlust an „Unordnung“**

  => $\Delta S < 0$



Voet&Voet, Biochemistry, 2nd ed.

Gesamtbilanz

Beispiel:

Lysozym bei 25°C

$\Delta H = -2245$ kJ/mol

- 1881 kJ/mol von den nichtpolaren Gruppen
- 364 kJ/mol von den polaren Gruppen

$-T\Delta S = -2186$ kJ/mol

$\Rightarrow \Delta G = -59$ kJ/mol

~ 0.4 kJ/mol pro Aminosäure!

Gefalteter Zustand nur unwesentlich gegenüber den entfalteten bevorzugt!

Zwischenbilanz

- Nativer Zustand thermodynamisch am günstigsten
- $|\Delta G_{U \text{ vs. } F}|$ recht gering
- Wie wir der native Zustand erreicht?
  - Zufällige Suche im Konformationsraum?
  - Spontaner Kollaps in die native Struktur?
  - Gibt es mehrere Pfade zur nativen Struktur?
- Auf welchen Zeitskalen passiert die Faltung?

Kinetik der Proteinfaltung

- Auf welchen **Zeitskalen** läuft die Faltung ab?
- Welche **Zwischenzustände** werden angenommen?
- Welche Schritte sind **geschwindigkeitsbestimmend**?
- Bilden sich Sekundär- und Tertiärstruktur gleichzeitig aus?

Levinthal-Paradoxon

Cyrus Levinthal (1968), Donald Wetlaufer (1973)

trast, is only 0.04 sec. To make this calculation, we assume a polymer chain of n links, each of which has two rotatable bonds with three energetically equal structures for each bond. Then the total number of structures is 3^{2n}, which can be approximated for convenience to 10^n. Assume that each of the internal rotations occurs independently at a frequency of 10^{13} sec^{-1}, then 2×10^{13} n structures are sampled per second. The time t required to sample each conformation m times is $t = (m\ 10^n\ \text{sec}/2 \times 10^{13} n)$. This argument was first pointed out to me by Prof. V. Bloomfield (personal communication, 1968); a similar argument was suggested by Levinthal in a generally unavailable publication (7).

D. Wetlaufer, Proc. Natl. Acad. Sci. USA (1973), 70, 697

Levinthal-Paradoxon

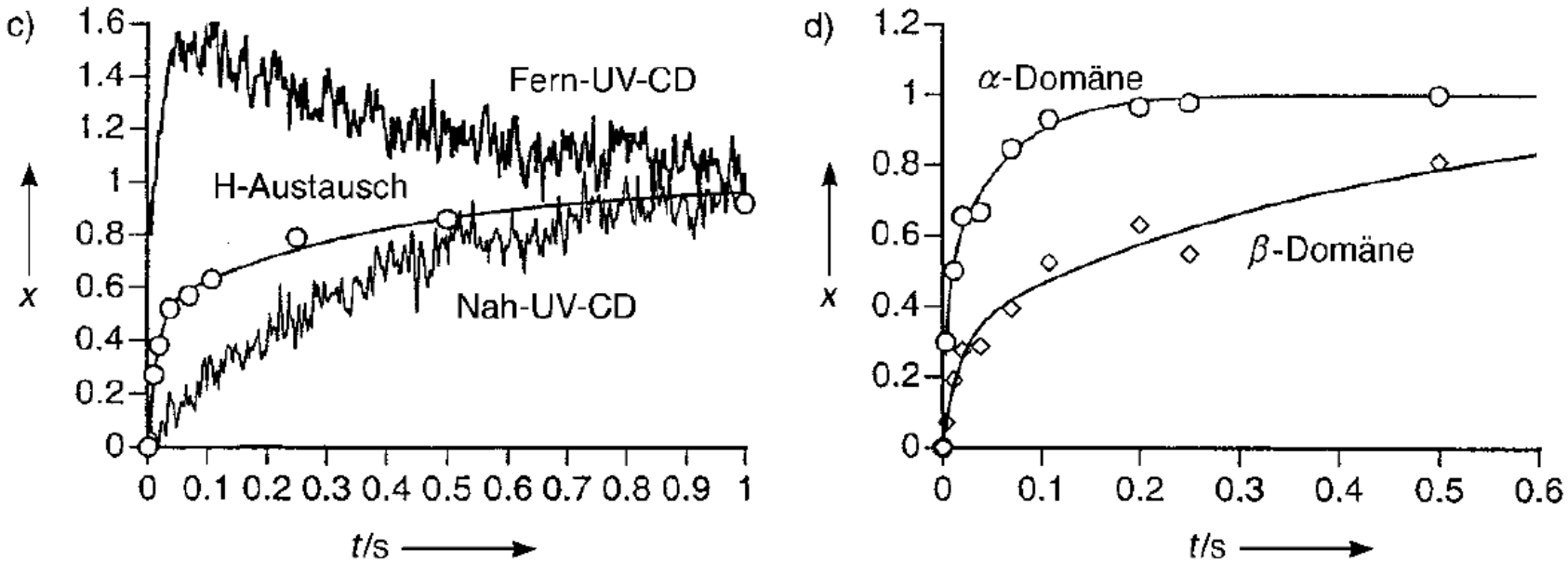
Levinthal-Paradoxon

Wenn ein Protein seinen nativen Zustand durch zufällige Suche erreichen soll, braucht diese Suche viel zu lange.

Auflösung

Faltung erfolgt nicht durch zufällige Suche, sondern ist auf irgendeine Art und Weise gerichtet.

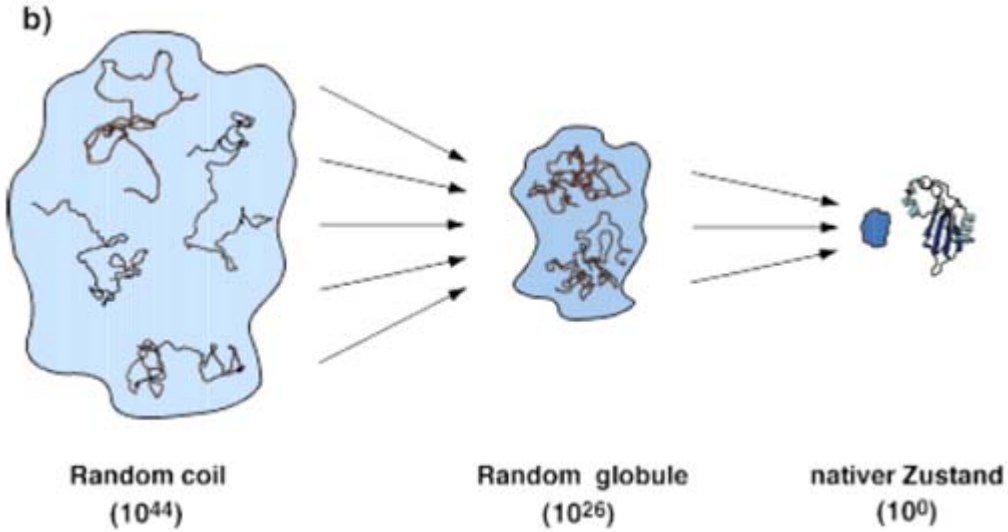
Kinetik der Proteinfaltung



- Experimentelle Daten für die Faltung von Lysozym (Dobson et al., Angew. Chemie (1998), 110, 922)
- Nah-UV-Signal: Maß für Ausbildung der Tertiärstruktur, Fern-UV-Signal: Maß für Ausbildung der Sekundärstruktur
- Schutz vor H-Austausch: Maß für „Kompaktheit“ der Struktur (im Protein verborgene Wasserstoffe werden nicht mehr mit dem Lösemittel ausgetauscht)
- Faltung dauert typischerweise zwischen 0.1 und 1000 s

Geschmolzene Kügelchen

- Zunächst Ausbildung teilweise geordnete Übergangszustände (schnell, einige ms)
- Kompakter als der ungeordnete Zustand
- Lokale Ordnung: viele der nativen Sekundärstrukturelemente ausgebildet
- Dieser Zustand wird *Molten Globule* („Geschmolzene Kügelchen“) oder *Random Globule* genannt

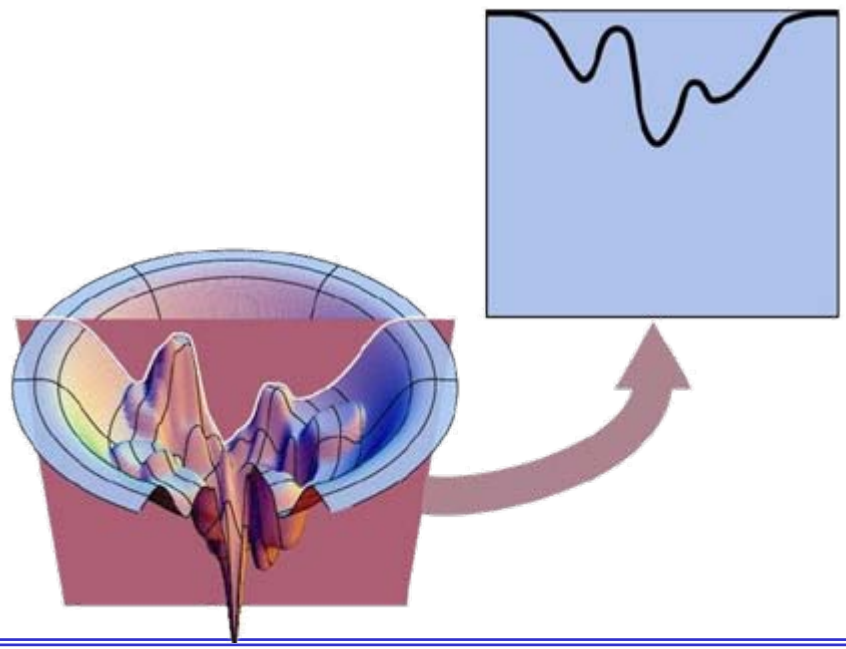


Geschmolzene Kügelchen

- Molten-Globule-Zustände
  - Weniger kompakt als der native Zustand
  - Besitzen noch nicht alle Kontakte im Inneren
  - Innere Seitenketten noch flexibel („flüssig“)
  - Schleifen und Strukturen an der Oberfläche überwiegend ungefaltet
  - Kein wohl definierter Zustand, sondern ein Ensemble ähnlicher Strukturen
- Wie kommt es zur Bildung dieser Übergangszustände?
- Wie geht es von da aus weiter?

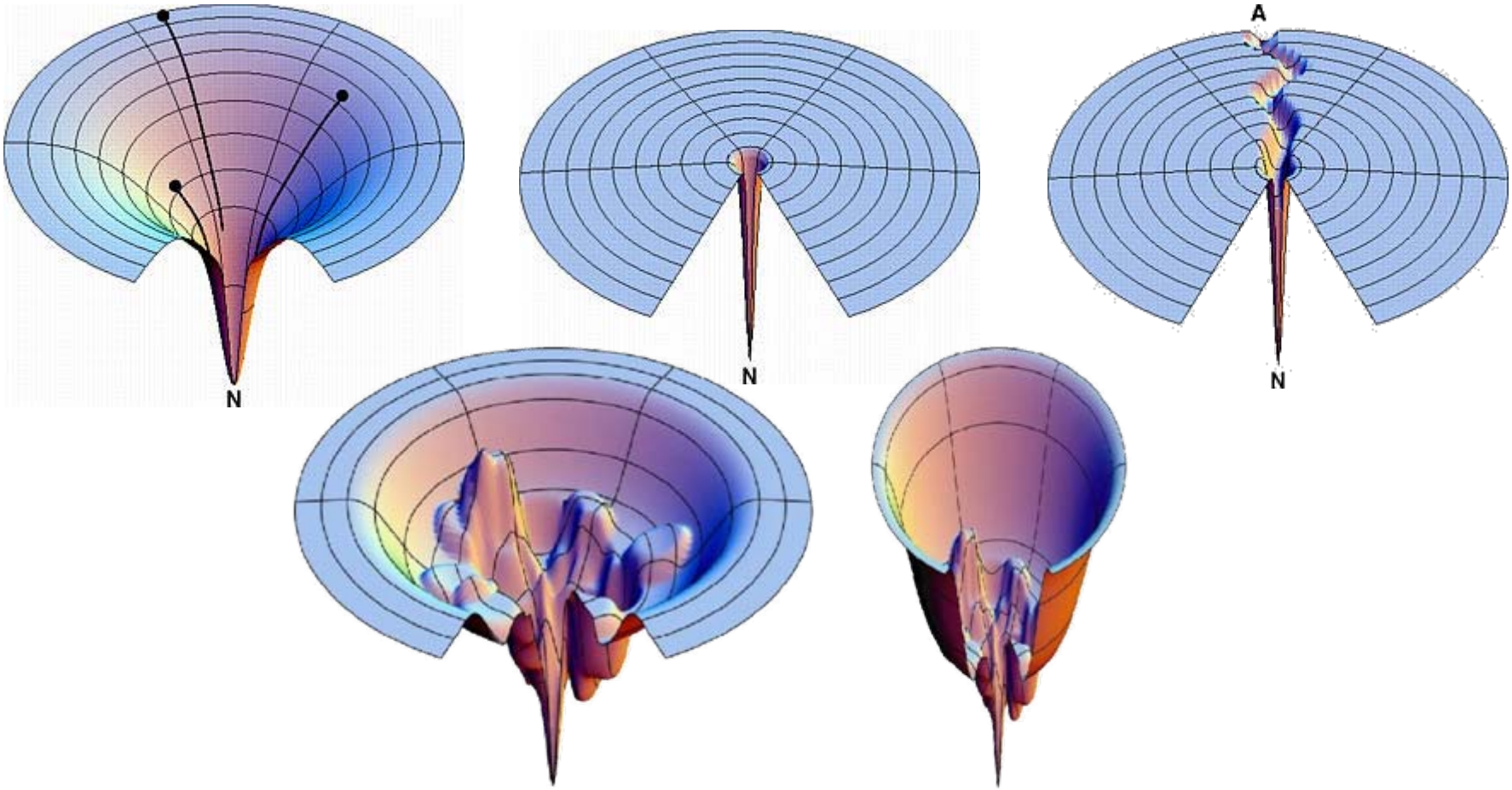
Faltungspfade

- Faltungsprozess entspricht Trajektorie im Konformationsraum der zu globalem Minimum läuft
- Unterschiedliche Startkonformationen führen offensichtlich zum selben Minimum
- Gibt es ausgezeichnete solche Trajektorien?
- Form und „Rauheit“ der Energiehyperfläche nahe des Minimums?



Dobson, Šali, Karplus, Angew. Chemie (1998), 110(7), 908

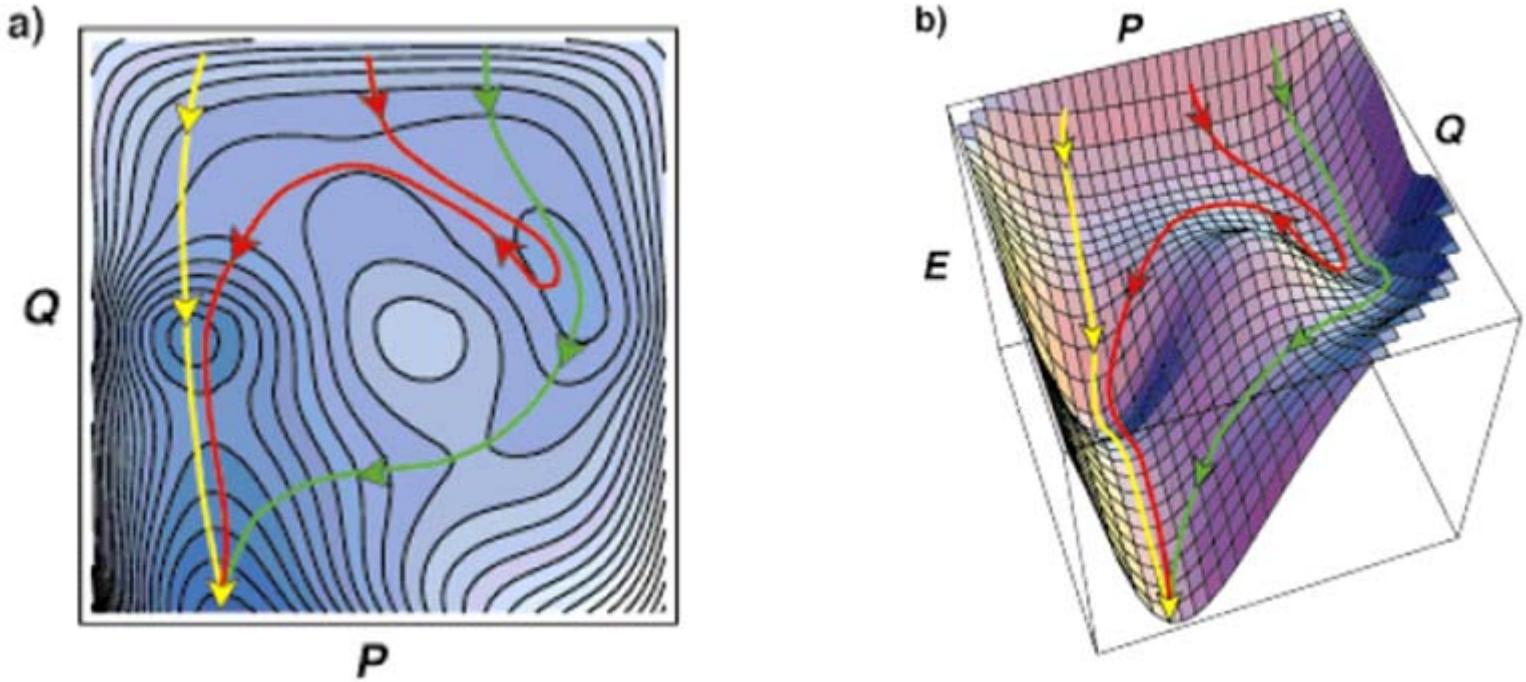
Faltungstrichter



Ken Dill: Begriff des **Faltungstrichters** (*folding funnel*)
Pfade konvergieren durch Trichter zum Minimum

Dobson, Šali, Karplus, Angew. Chemie (1998), 110(7), 908

Faltungspfade



- Lysozym kennt mehrere Faltungspfade
- Zwischenzustände in denen je eine der beiden Domänen gefaltet ist
- Faltungsgeschwindigkeiten der Domänen unterschiedlich: eine überwiegend α, die andere überwiegend β
- Pfade unterschiedlich schnell (gelb schnell, grün + rot langsam)

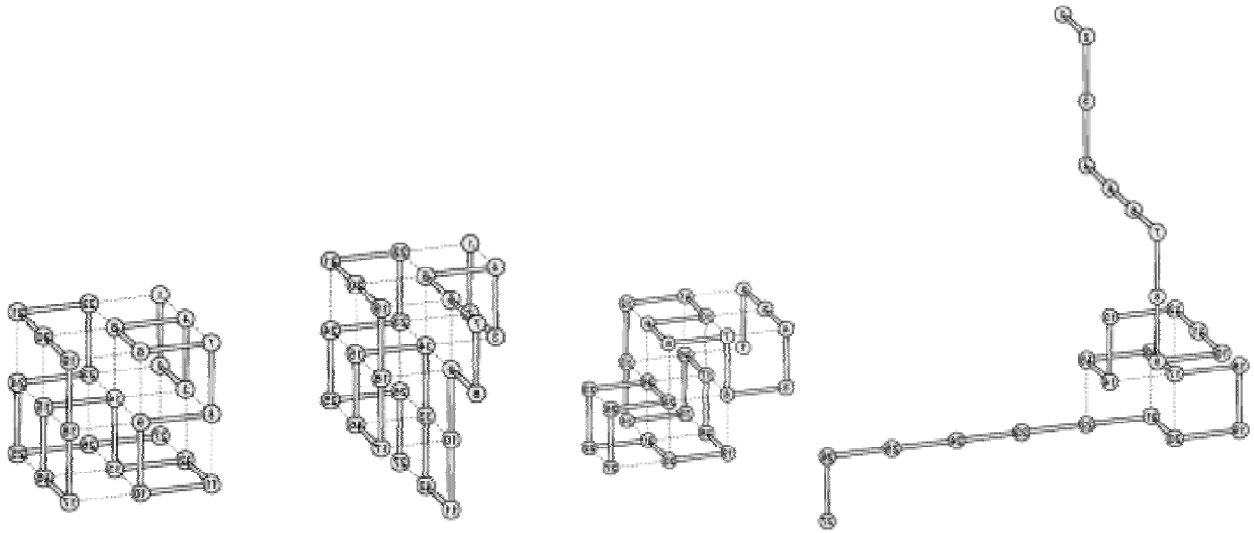
Dobson, Šali, Karplus, Angew. Chemie (1998), 110(7), 908

Proteinfaltung – Übersicht

- Problemdefinition
- Biochemie
  - Protein-Biosynthese
  - Proteinfaltung: beteiligte Enzyme, Chaperone
- Biophysik
  - Thermodynamik
  - Kinetik
- **Modelle für die Faltung**
  - **Gittermodelle (Komplexität)**
  - **Molekulardynamik (machen wir nicht)**

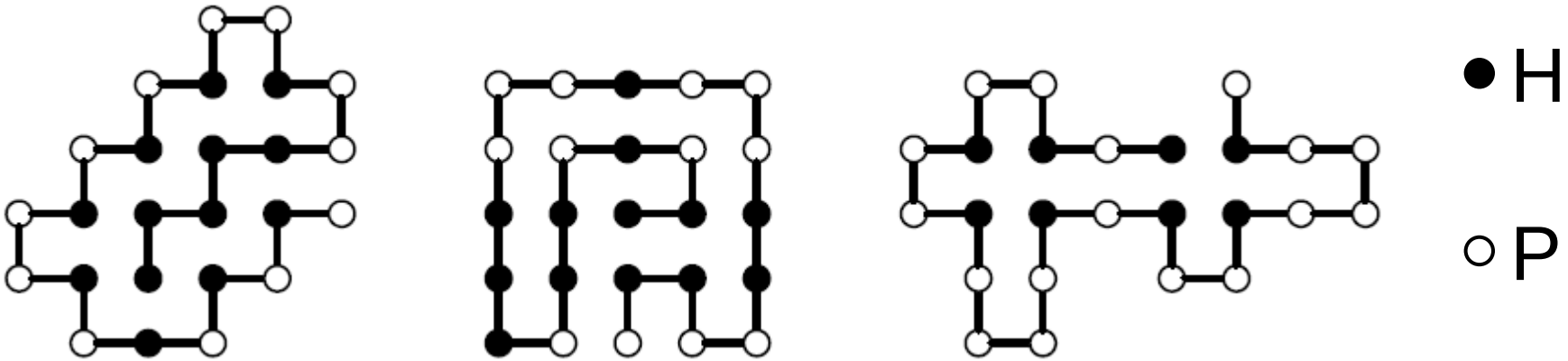
Gittermodelle

- Um Proteinfaltung im Detail zu studieren wurden sehr stark vereinfachte Gittermodelle eingeführt
- Protein entspricht Kette von Kugeln (AS)
- Kugeln können nur auf Gitterpositionen liegen
- Benachbart



Das HP-Modell

- Ken Dill führte das **HP-Modell** (**H**ydrophob-**P**olar) ein
- Sequenz reduziert auf Sequenz **hydrophober** (H) und **hydrophiler** (polarer, P) Reste
- *Self avoiding walk* auf 2D oder 3D Gitter
- Wechselwirkung nur zwischen benachbarten, nicht direkt verbundenen Resten



Das HP-Modell

- Trotz grober Vereinfachung erhält das HP-Modell viele Eigenschaften des Faltungsproblems
  - Levinthal-Paradoxon gilt weiterhin
  - Größenordnung der Faltungsgeschwindigkeit abschätzbar
- Aufgrund seiner Einfachheit sind aber interessante Aussagen möglich über
  - **Komplexität des Problems**
  - **Struktur der Energielandschaften**

Faltungsproblem im HP-Modell

Gegeben

- Sequenz $S = \{H|P\}^n$
- Ein Gitter
- Eine Energiefunktion E

  (z.B. E(H,H) = -1, E(H,P) = 0, E(P,P) = 0)

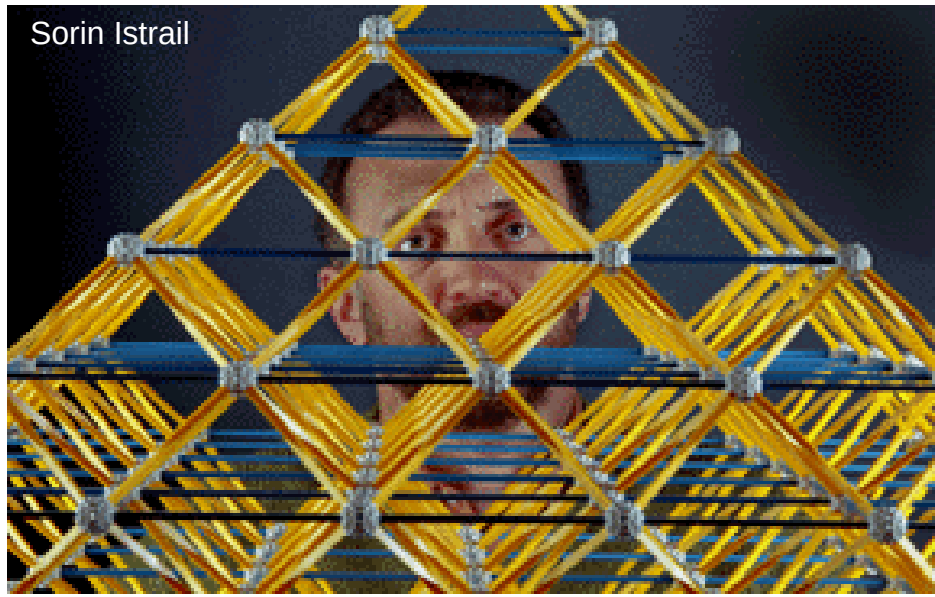
Gesucht

Die Anordnung der Sequenz auf dem Gitter mit minimaler Energie

Alle möglichen Anordnungen sind nur für sehr kurze Sequenzen (~30 AS) enumerierbar.

Komplexität der Faltung

- W. Hart und S. Istrail zeigten, dass Proteinfaltung im HP-Modell NP-hart ist
  - Für alle sinnvollen Gittertypen
  - Für Energiefunktionen ähnlich zu van-der-Waals
- Approximation in Linearzeit möglich (auf 86% der minimalen Energie)
- Für realistischere Modelle gibt es keine entsprechenden Beweise



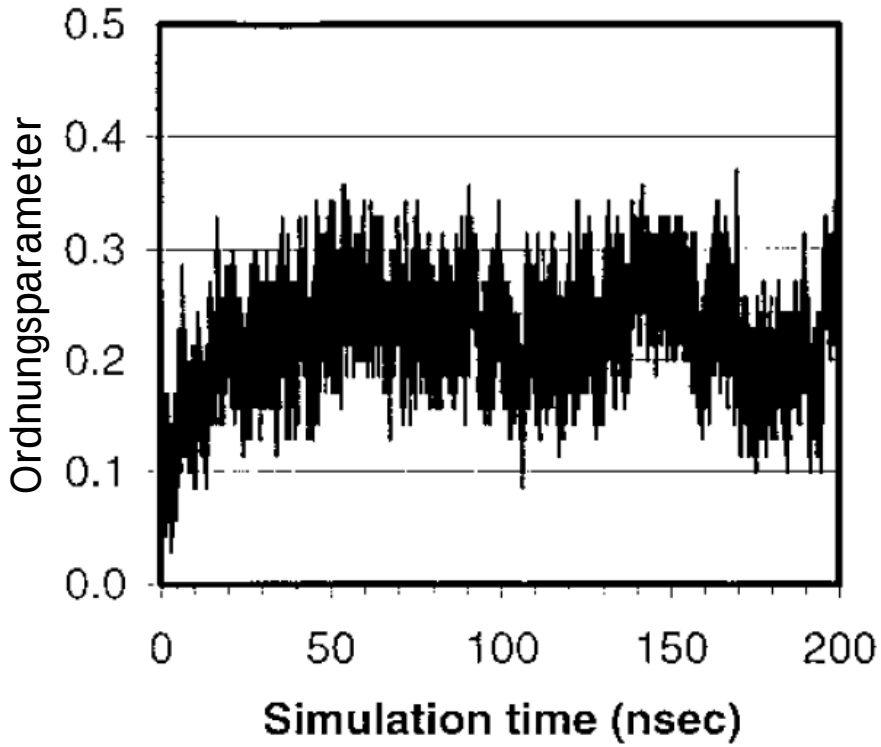
Faltung durch MDS

- Faltung als Prozess sollte von der Energie her, als auch von der Dynamik her mit MD-Simulation vorhersagbar sein
- Probleme
  - **Präzision** der Energiefunktion – das globale Minimum muss erhalten bleiben
  - **Aufwand**
    - Faltung kleiner Protein im Bereich von µs

) 10^8-10^{10} Iterationen!
    - Berücksichtigung von Wasser vergrößert Rechenzeit

Duan & Kollman

- MD-Simulation eines kleinen Peptids (36 AS)
- AMBER auf CRAY T3D, 256 CPUs
- Simulationszeit: 200 ns (10^8 Iterationen)
- Konstante Temperatur, konstanter Druck
- Mehrere Monate Rechenzeit
- Nativer Zustand nicht erreicht



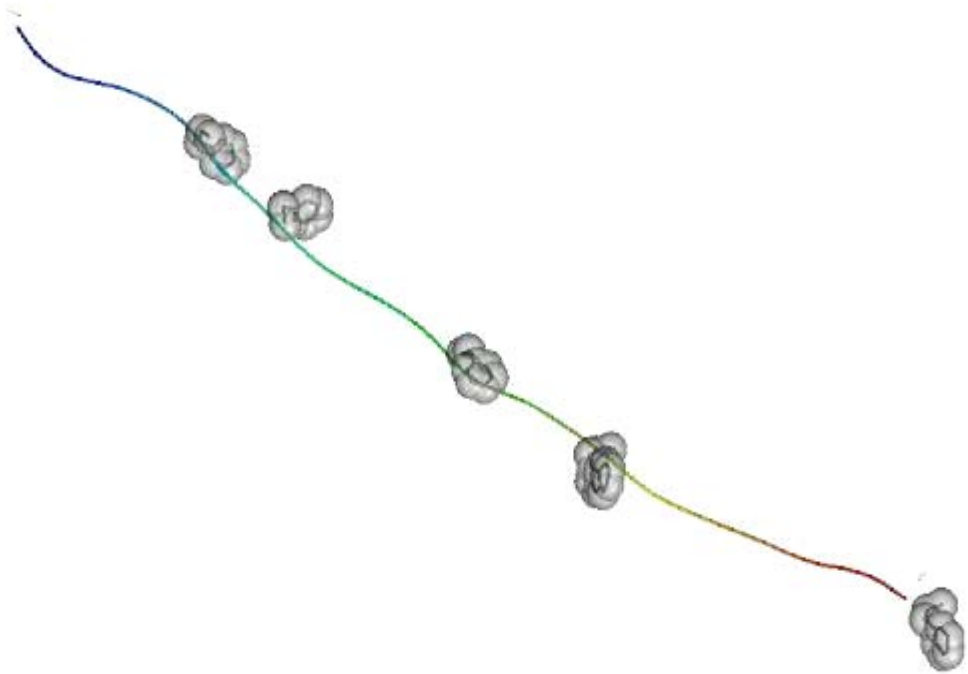
Duan, Wang, Kollman, Proc. Natl. Acad. Sci. USA (1998), 95, 9897

Folding@home

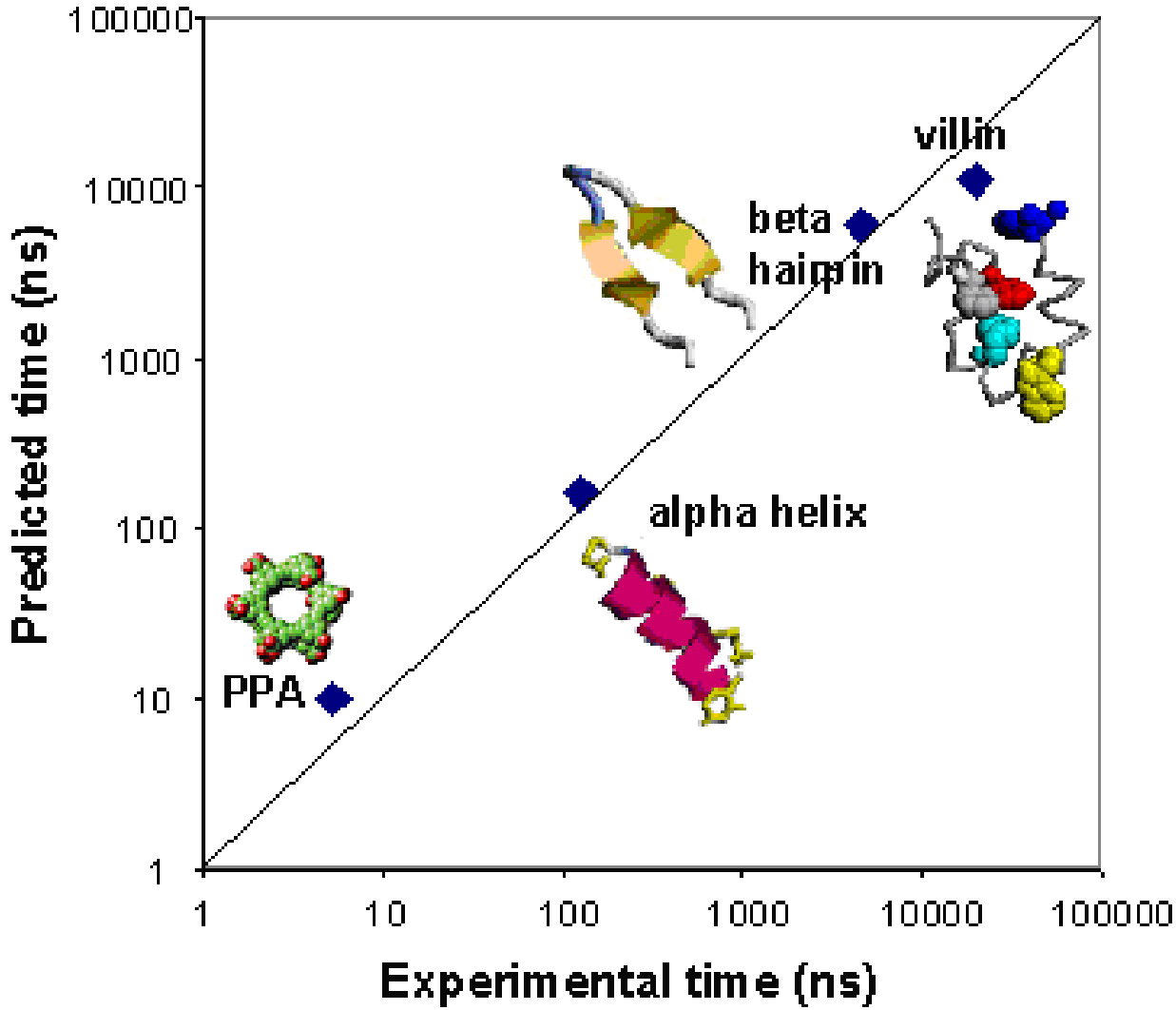
- Weiltweit verteilte Simulation („Bildschirmschoner“)
- „Supercluster“ viel leistungsstärker als größte Supercomputer
- Simulationszeiten im ms-Bereich machbar

Number CPUs	Number Active CPUs	Number Users	Number Teams	Last Update
581974	118572	275910	28335	2003-12-09 12:22:45

folding@home



folding@home



IBMs Blue-Gene-Projekt

- Rechner für Molekulardynamik (Speziell auch Proteinfaltung)
- 65536 CPUs
- 360 TFLOPS Peak
- Aufbau bis 2005

Zusammenfassung

Vorhersage der Proteinfaltung ist **schwierig**…

…für Simulationen – wegen der **Zeitskalen**

…von der Komplexitätstheorie – sogar **NP-hart!**

…von der Biophysik – da die **Energiedifferenzen** sehr gering sind

…von der Biochemie – aufgrund der **komplexen Vorgänge** (Chaperone, Enzyme, …)

Links, Literatur

Protein-Biosynthese

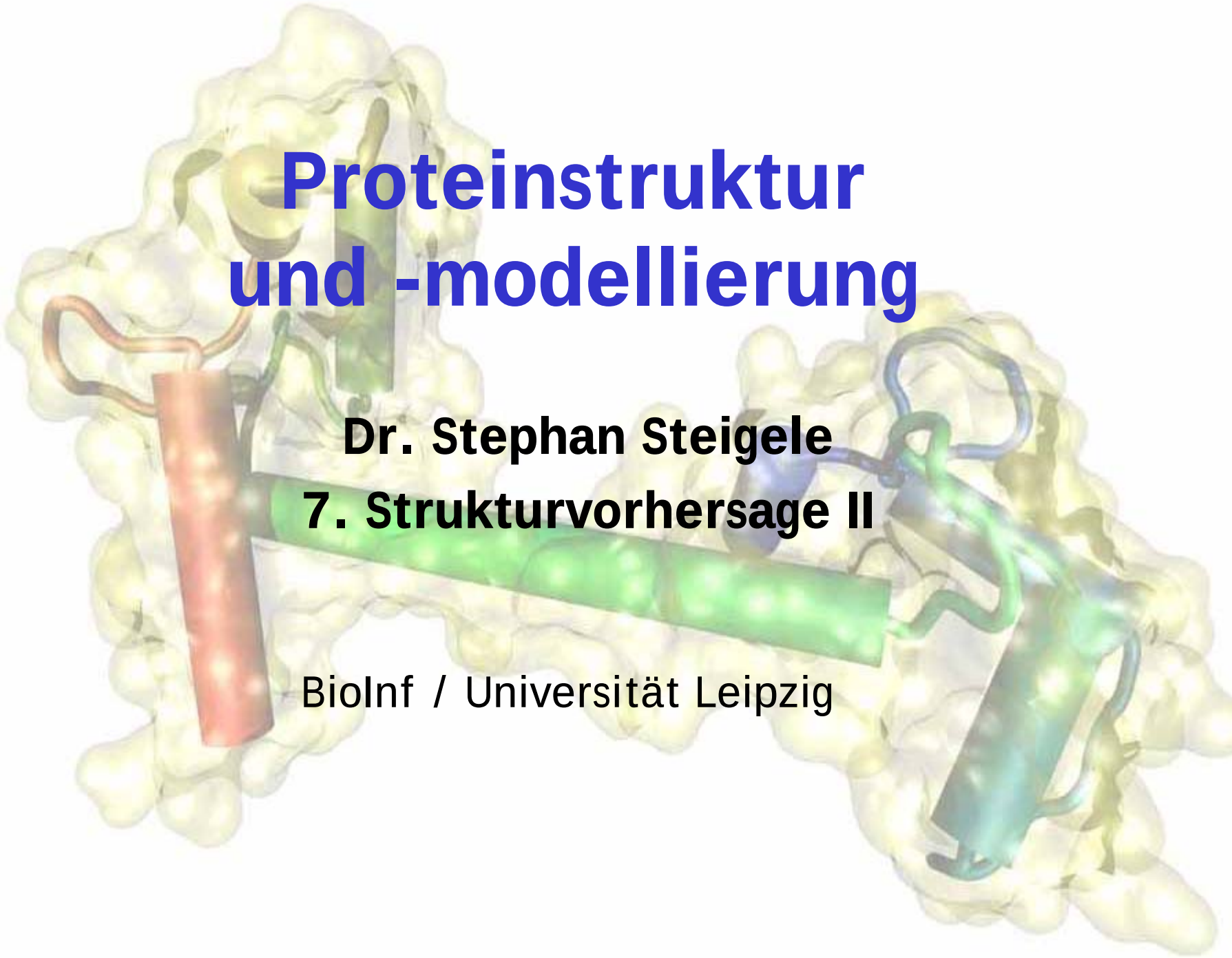
- Animation der Biosynthese:
  `http://www.ncc.gmu.edu/dna/ANIMPROT.htm`
- Alberts et al., Molecular Biology of the Cell, Garland Science

Protein-Faltung

- Dobson, Sali, Karplus, Angew. Chemie (1998), 110, 908
- Branden, Tooze, Introduction to Protein Structure

Modelle und Simulationen

- Dill, Biochemistry (1985), 24, 1501
- Duan, Wang, Kollman, Proc. Natl. Acad. Sci. USA (1999), 95, 9897
- folding@home: **`http://folding.stanford.edu/`**



Proteinstruktur und -modellierung

Dr. Stephan Steigele

7. Strukturvorhersage II

BioInf / Universität Leipzig

Strukturvorhersage – Übersicht

- Problemdefinition/-klassifizierung
- Sekundärstrukturvorhersage
- Fold-Recognition
- Threading
- ab-initio-Vorhersage
- CASP/CAFASP

Fold-Recognition und Threading

- Begriffe und Definitionen
  - Faltungsklassen
  - Fold Recognition
  - Threading
- Threading
  - Threading als Alignment-Problem
  - Potentiale zum Threading
  - Komplexität des Threading-Problems
  - Algorithmen
  - Qualität der Resultate

Strukturvorhersage - Nomenklatur

Schrittweises Vorgehen:

- Template-Auswahl
- Threading der Zielsequenz auf Template
- Modellierung divergenter Regionen (Loops)
- Seitenkettenmodellierung

Fold Recognition

Threading

Homologie-Modellierung

Strukturvorhersage - Nomenklatur

Fold Recognition:

Gegeben Sequenz,

finde Faltungsklasse

Threading:

Gegeben Sequenz,

erzeuge grobes 3D-Modell

Homologiemodellierung:

Gegeben Sequenz + Template,
konstruiere vollständiges Modell

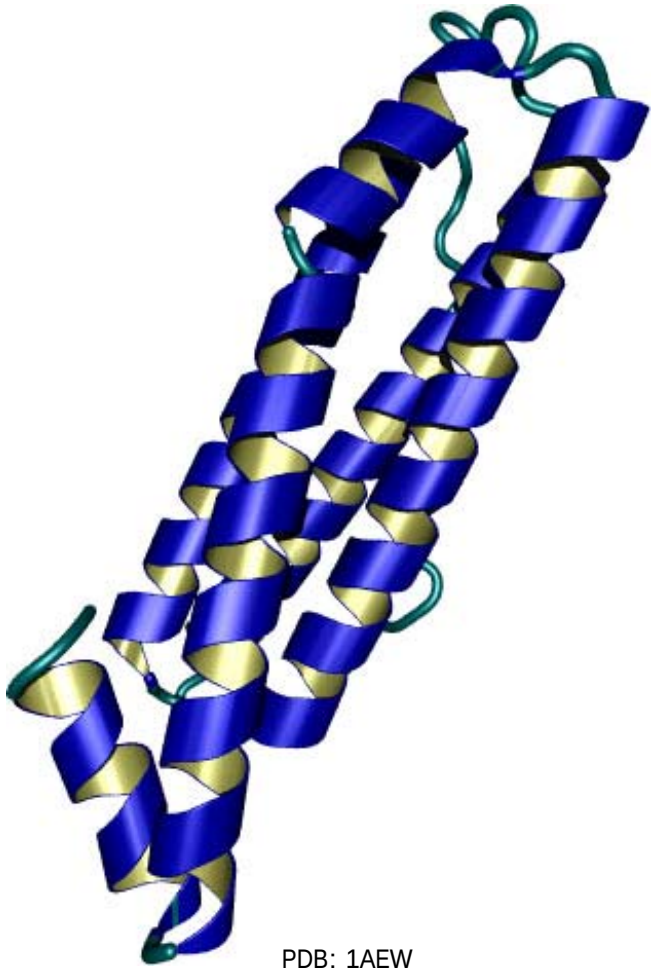
Faltungsklassen

SCOP (03/2001)

- ~ 13.000 Strukturen aus der PDB
- ~ 31.500 Domänen
- ~ 600 Faltungsklassen
- α: nur Helices
- β: nur Faltblätter
- α/β: Faltblatt mit verbindenden Helices (β-α-β)
- $\alpha+\beta$

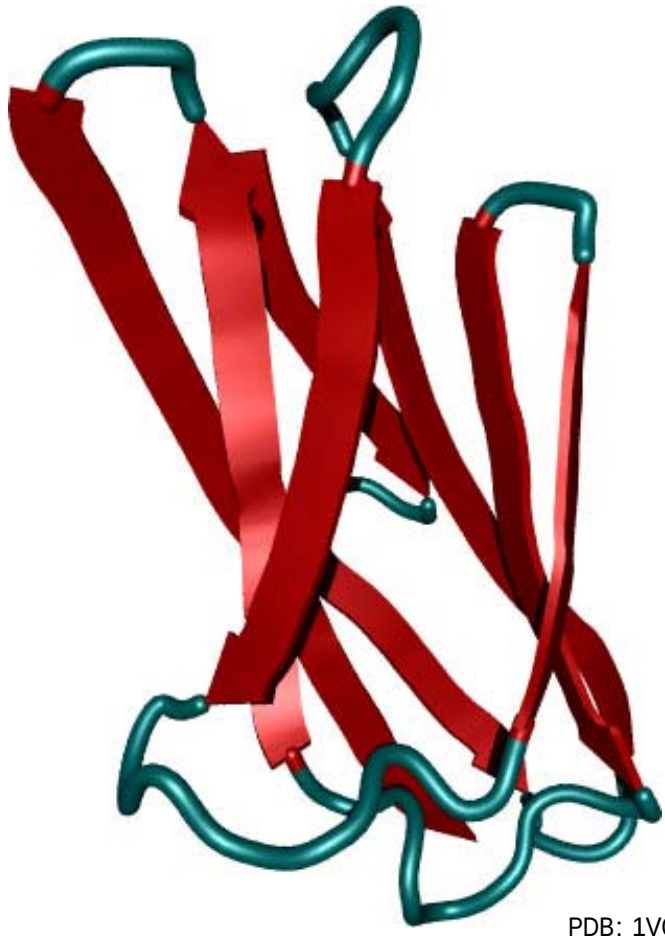
Exkurs: Faltungsklassen

α: nur Helices



PDB: 1AEW

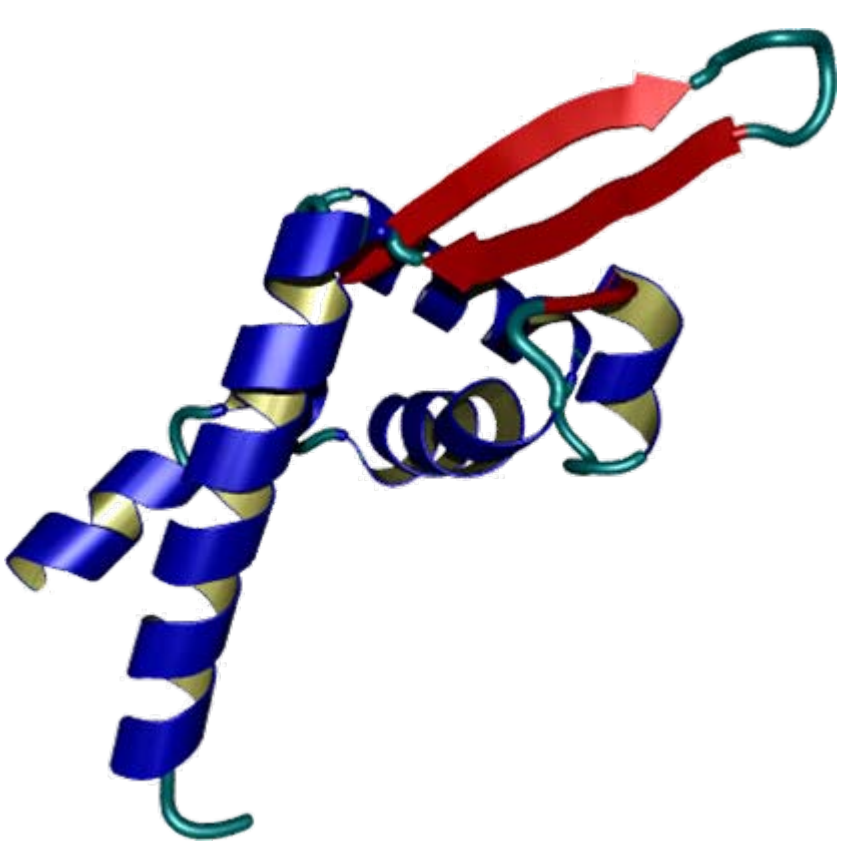
β: nur Faltblätter



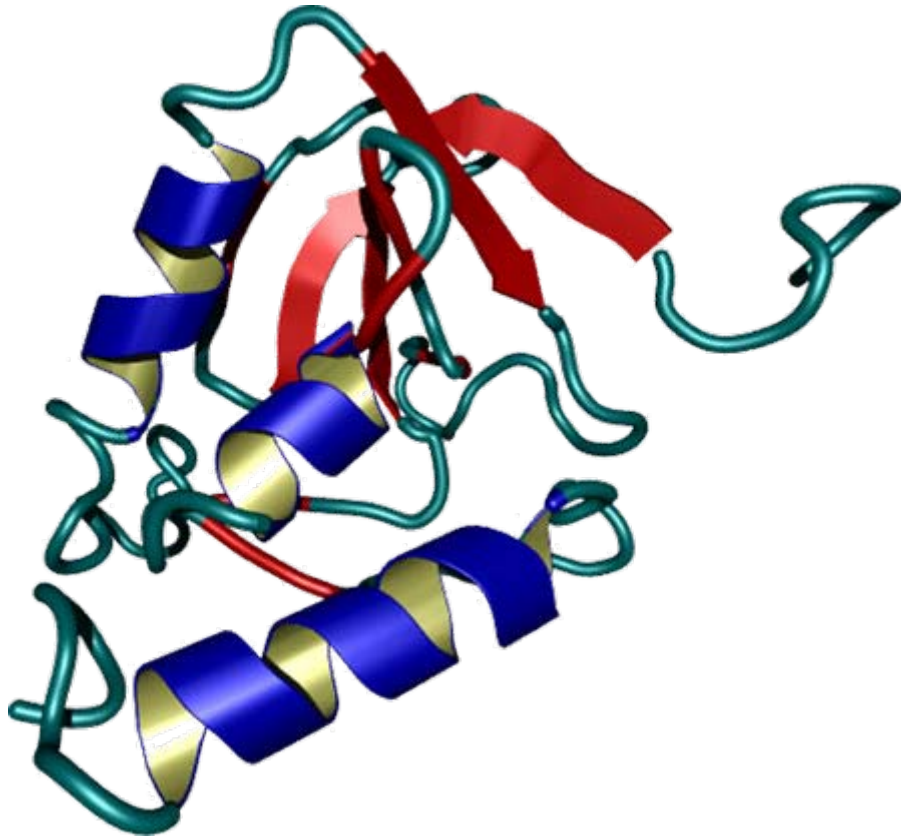
PDB: 1VCA

Exkurs: Faltungsklassen

α+β: Helices und Faltblätter in der Sequenz getrennt, Faltblätter meist durch *Turns* verbunden



Ubichinon-konjugierendes Enzym (1UB9)



Staphylokokken-Nuklease (2SNS)

Exkurs: Faltungsklassen, SCOP

- ~ 20.000 Strukturen in der PDB
- ~ 600 Faltungsklassen
- Anzahl Faltungsklassen in SCOP (03/2001):

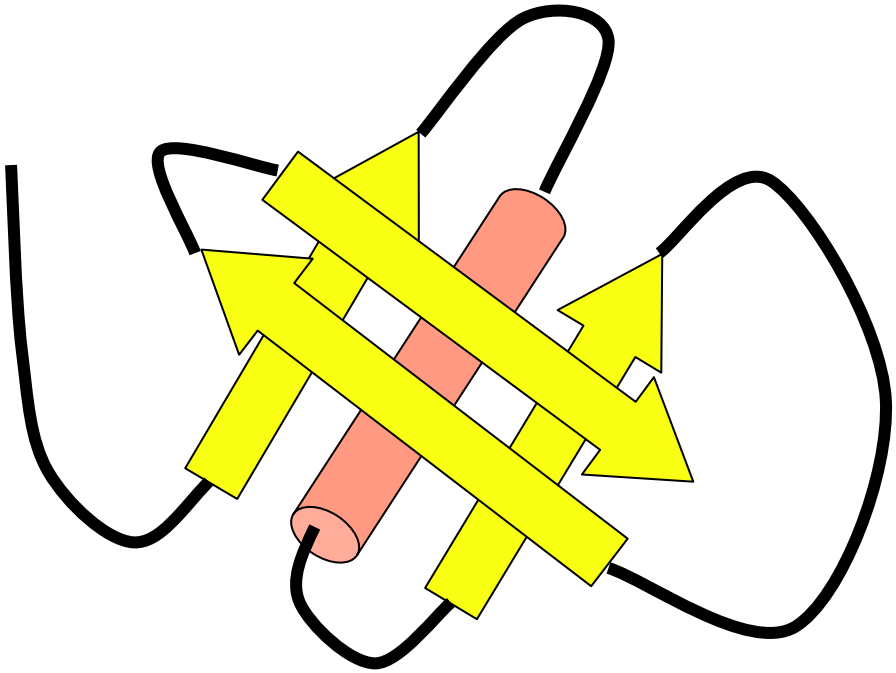
138	α
93	β
97	α/β
184	$\alpha + \beta$
23	Multidomänen-Proteine
11	Membranproteine/Zelloberflächenproteine
54	Kleine Proteine
$\sum$ 605	

Zwei Proteine, eine Faltungsklasse

- **Divergente Evolution**
  - Proteine sind evolutionär verwandt
  - Häufig sehr geringe Sequenzähnlichkeit
- **Konvergente Evolution**
  - Funktionelle Gemeinsamkeiten führen zu ähnlichen strukturellen Lösungen
  - Sehr wenige Beispiele, häufig nur Ähnlichkeiten in kleinen Fragmenten (aktives Zentrum o.ä.)
- **Begrenzte Anzahl Folds**
  - Anzahl der möglichen Faltungsklassen ist beschränkt (1000?)
  - Proteine landen „zufällig“ in der selben Klasse
- **Fehlklassifizierung**
  - Scheinbare strukturelle Ähnlichkeit basiert auf falscher Klassifizierung
  - SCOP, CATH und FSSP weisen z.B. nur in 60% der Fälle Identität in der Klassifizierungen auf!

Fold Recognition Methods, In: Structural Bioinformatics (Hrsg.: P. E. Bourne, H. Weissig), Wiley, 2003

Threading

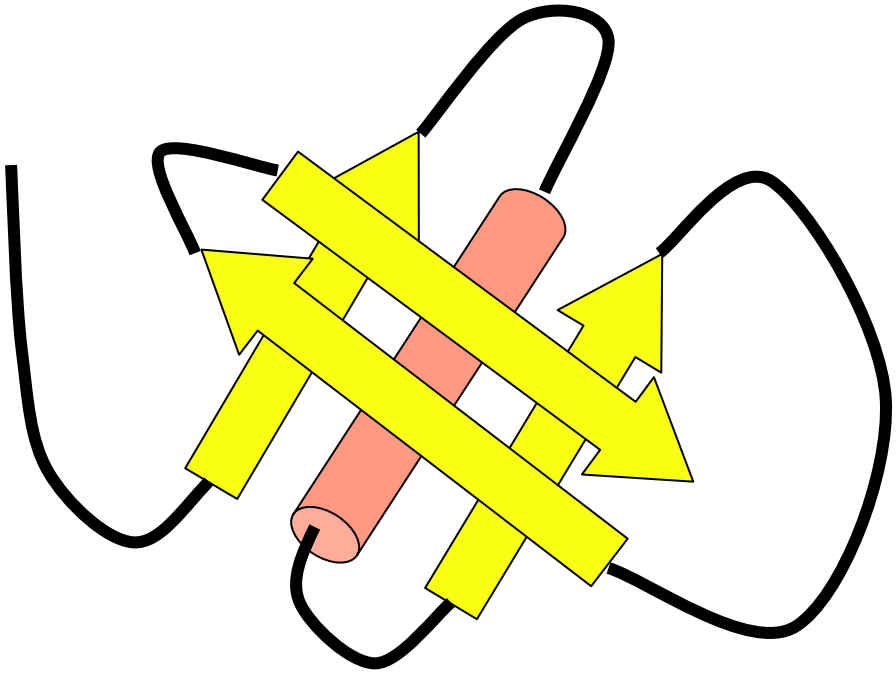


```
...LGFCYWS...
...ILVGCIL...
```

Gegeben

- Eine (oder mehrere) Struktur(en) (Schablonen)
- Eine Zielsequenz

Threading



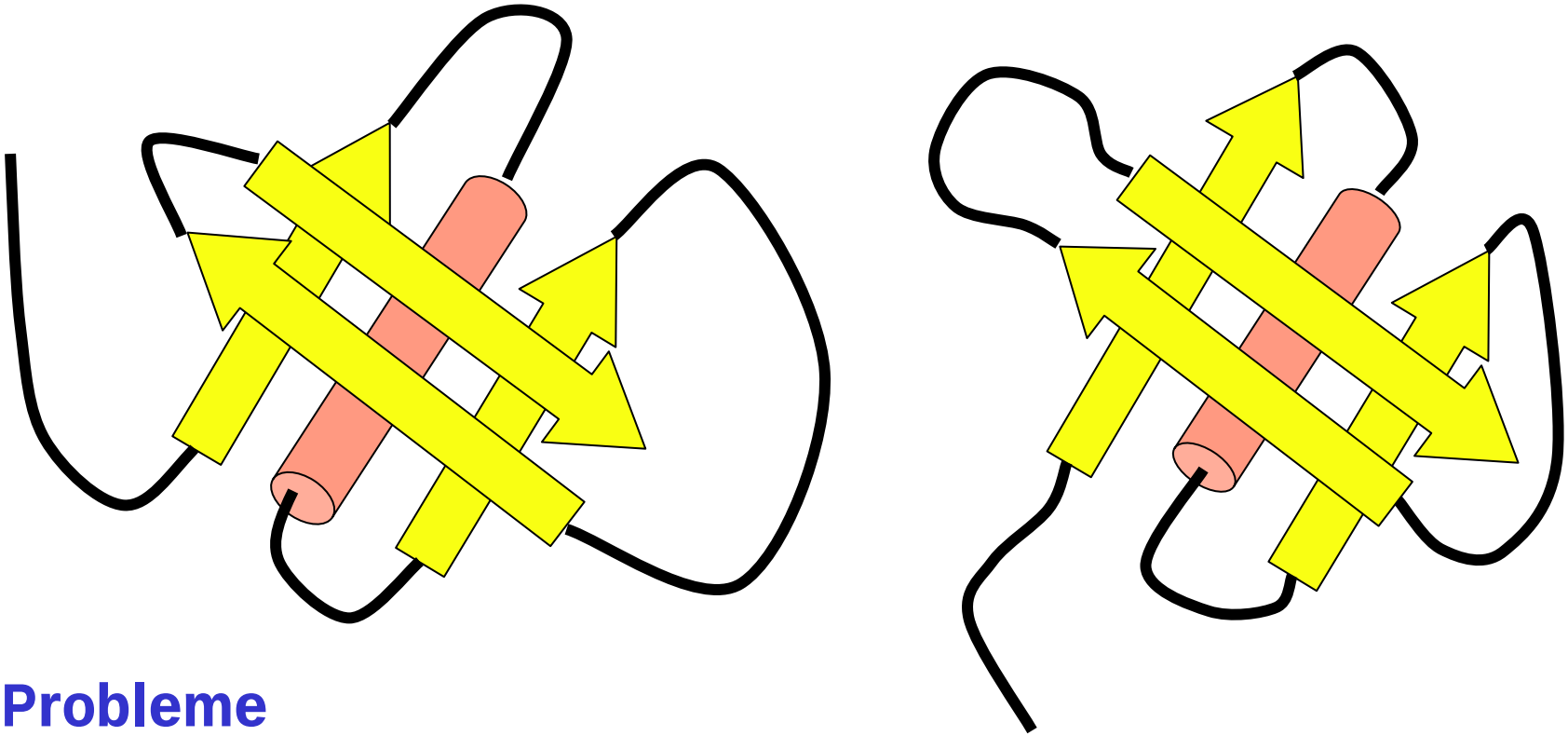
```
...LGFCYWS...
...ILVGCIL...
```

Gesucht

Zuordnung der Zielsequenz zu Positionen in der Schablonenstruktur

=> **Sequenz-Struktur-Alignment**

Threading



Probleme

- Proteine einer Faltungsklasse differieren in **Schleifenregionen** (abgesehen von funktional wichtigen Loops)
- Meist nur **Kernelement der Struktur** konserviert
- Länge und Beginn der Strukturelemente kann differieren

=> Alignment mit Gaps!

Threading als Alignment

- **Gegeben**
  - Zielsequenz *x*
  - Schablone *y*
  - Energiefunktion *E*
- **Gesucht**

  Bezüglich *E* global optimales **Sequenz-Struktur-Alignment** von *x* auf *y*

- Optimierungsproblem
  - Suche nach dem Alignment minimaler Energie
  - Welche Art von Energiefunktion?

Potentialfunktionen

Probleme

- Struktur auf atomarer Auflösung unbekannt (z.B. Seitenkettenpositionen unklar)
- Strukturen unvollständig (Lücken im Alignment)

=> Energiefunktion mit reduzierter Auflösung (AS-Ebene)

Wissensbasierte Energiefunktionen

- *„Knowledge-based potentials“*
- Abgeleitet aus Analyse von Strukturdatenbanken
- **Inverse Boltzmannstatistik**
  - Häufigkeit des Auftretens in DB nähert Ensemblehäufigkeit
  - Boltzmann-Verteilung liefert aus Häufigkeiten Energien

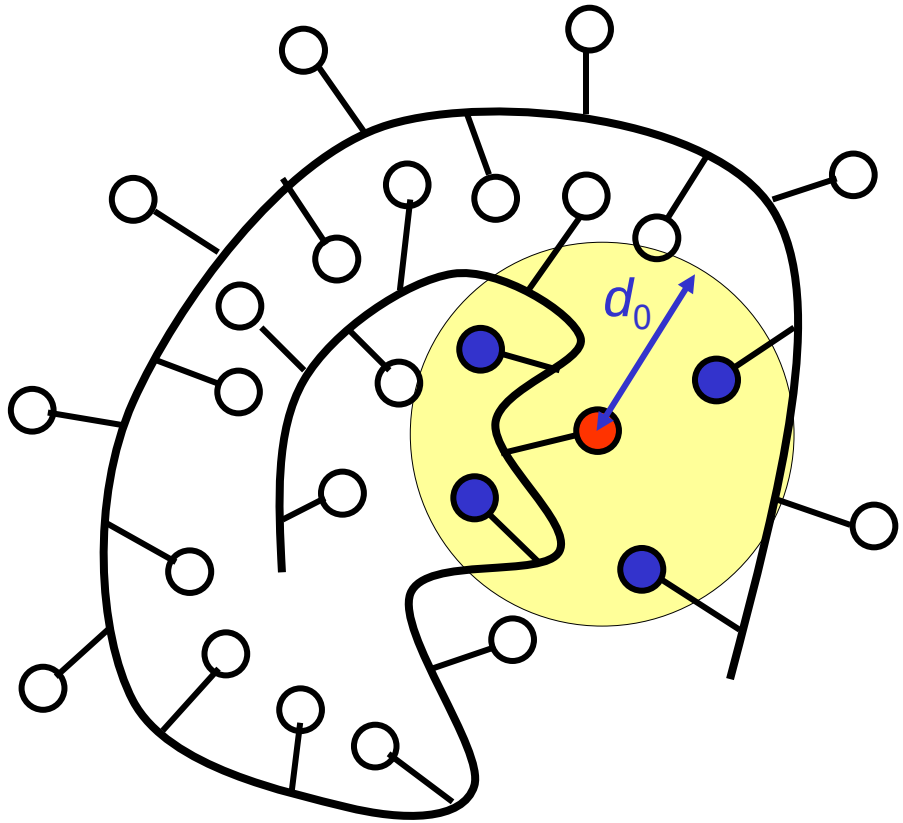
=> Pseudo-Energien!

=> Keine physikalischen Modelle notwendig!

Kontakte in Proteinen

Definition Paarkontakt

- Reduktion des Proteins auf C_α- oder C_β-Positionen
- AS *i*, *j* in Kontakt, falls Abstand $d(C_\beta^i, C_\beta^j) < d_0$ (z.B. d_0 = 7 Å)
- Hilfskonstruktion für Gly (kein C_β!)



Sippl-Potential I

- WW-Energie $E_{i,j}$ zwischen zwei AS *i, j* als Funktion des C_β-C_β-Abstands *d(i, j)*, der AS-Typen und des Sequenzabstands *k*
- Analyse einer Datenbank mit ~100 Proteinstrukturen
- Ableitung von Energien aus Häufigkeiten
- **Boltzmann-Verteilung**

$$\rho = \frac{N_i}{N} = e^{-\frac{E_i}{k_B T}} / \sum_i e^{-\frac{E_i}{k_B T}} = \frac{1}{Q}\, e^{-\frac{E_i}{k_B T}}$$

  mit der Wahrscheinlichkeitsdichte ρ
- Annahme: In einem Ensemble von Strukturen sind die Interaktionsenergien Boltzmann-verteilt

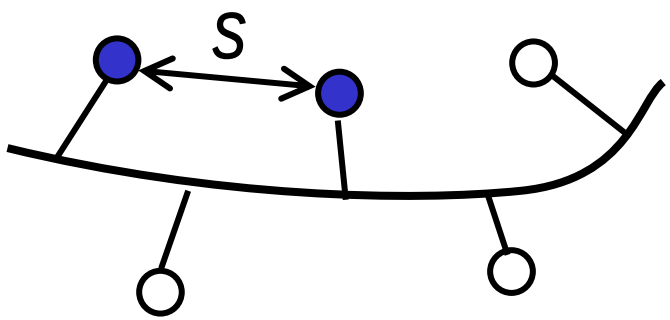
M. Sippl, J. Mol. Biol. (1990), 213, 859, M. Hendlich et al., J. Mol. Biol. (1990), 216, 167

Sippl-Potential II

- „Inverser“ Boltzmann-Ansatz

$$E_i = -k_B T(\ln \rho_i + \ln Q)$$

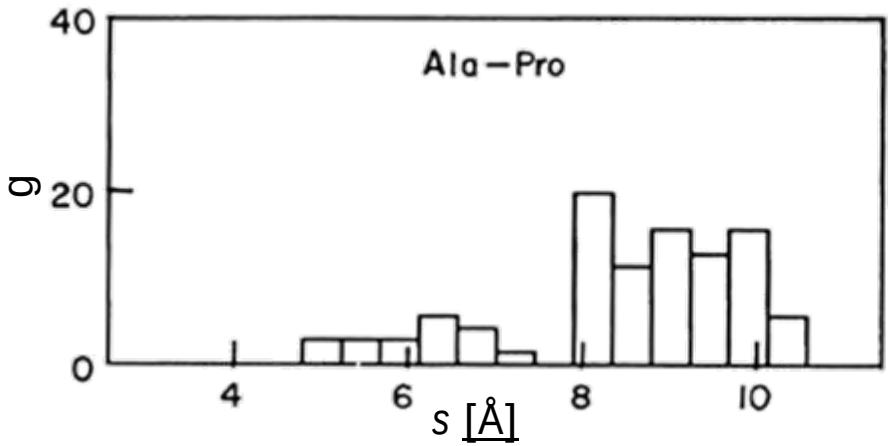
- Approximation der Wahrscheinlichkeitsdichten aus beobachteten Häufigkeiten *g*
- Bestimmung von $E_{a,b}^{k}(s)$ für alle AS-Kombinationen *(a, b)*, Sequenzabstände *k* und (diskretisierte) Abstände *s*

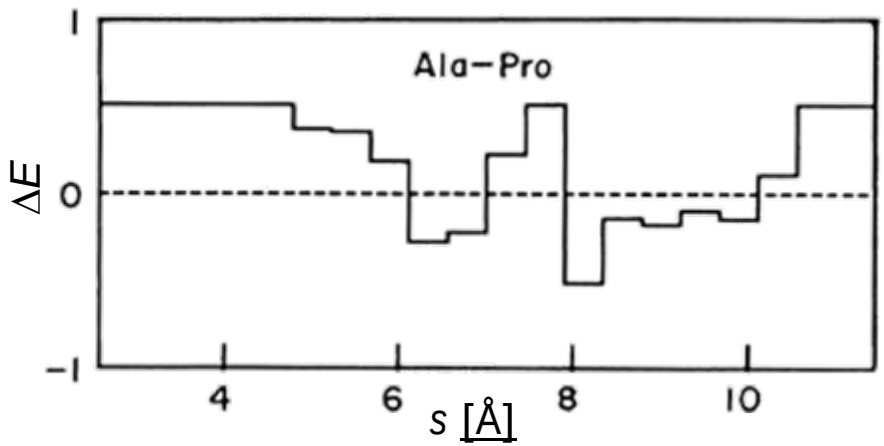


$k = 2$

Sippl-Potential III

$$\Delta E_k^{ab}(s) = -k_B T \ln \frac{g_k^{ab}(s)}{g_k(s)} - k_B T \ln \frac{Q_k^{ab}}{Q_k}$$





M. Sippl, J. Mol. Biol. (1990), 213, 859, M. Hendlich et al., J. Mol. Biol. (1990), 216, 167

Paar-Interaktions-Potentiale

Vorteile

- Kein physikalisches Modell notwendig
- Rechnerisch sehr einfach (Nachsehen eines Tabelleneintrags)
- Keine Modellierung der Seitenketten notwendig

Nachteile

- Sehr grobe Näherungen, da Aminosäure als Einheit betrachtet wird

  => viele spezifische, seltener auftretende Effekte nicht erfassbar
- Abhängig von guten Strukturdatenbanken zur Parametrisierung

Solvatationspotentiale

- **Solvatation** wichtig für Faltung
- Jones *et al.* leiten Solvatationspotential wie Sippl-Potential her
- Abhängig von der **Lösemittelzugänglichkeit** *r*
- *r* beschreibt, wie stark der Rest für Wasser zugänglich ist (*r* = Anteil der Oberfläche der AS der Kontakt mit Wasser hat)

  r = 0: im Inneren (*buried*)

  r = 1: komplett zugänglich

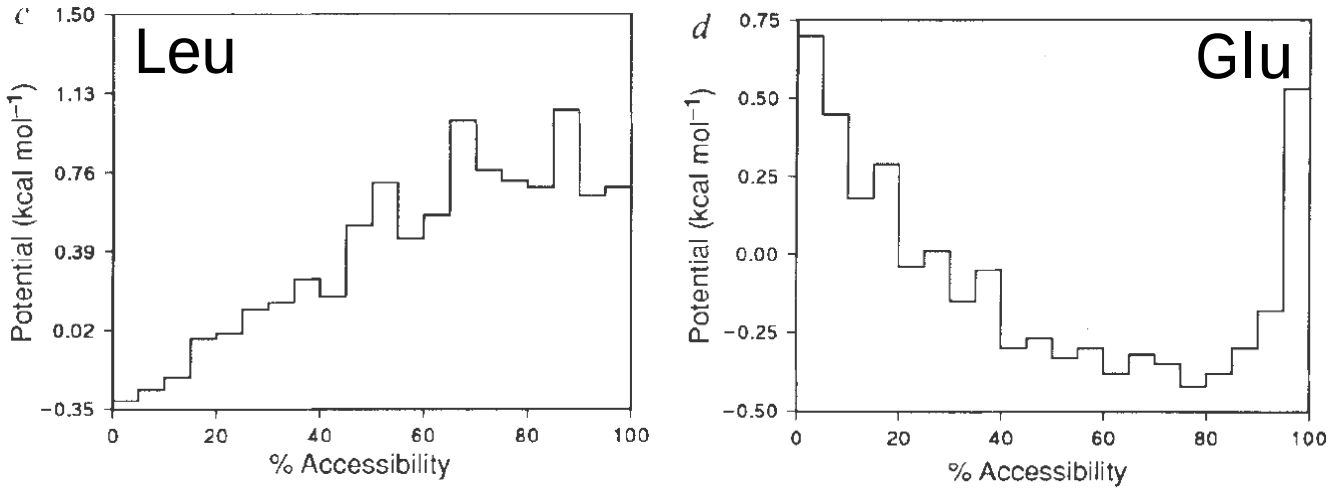
Solvatationspotentiale

Definition analog zu Sippl-Potential

$$\Delta E^{a}_{solv}(r) = -k_B T \ln \frac{f^{a}(r)}{f(r)}$$

$f^{a}(r)$ Häufigkeit von AS *a* mit Zugänglichkeit *r*

$f(r)$ Gesamthäufigkeit von AS mit Zugänglichkeit *r*



Kontaktkapazitätspotentiale

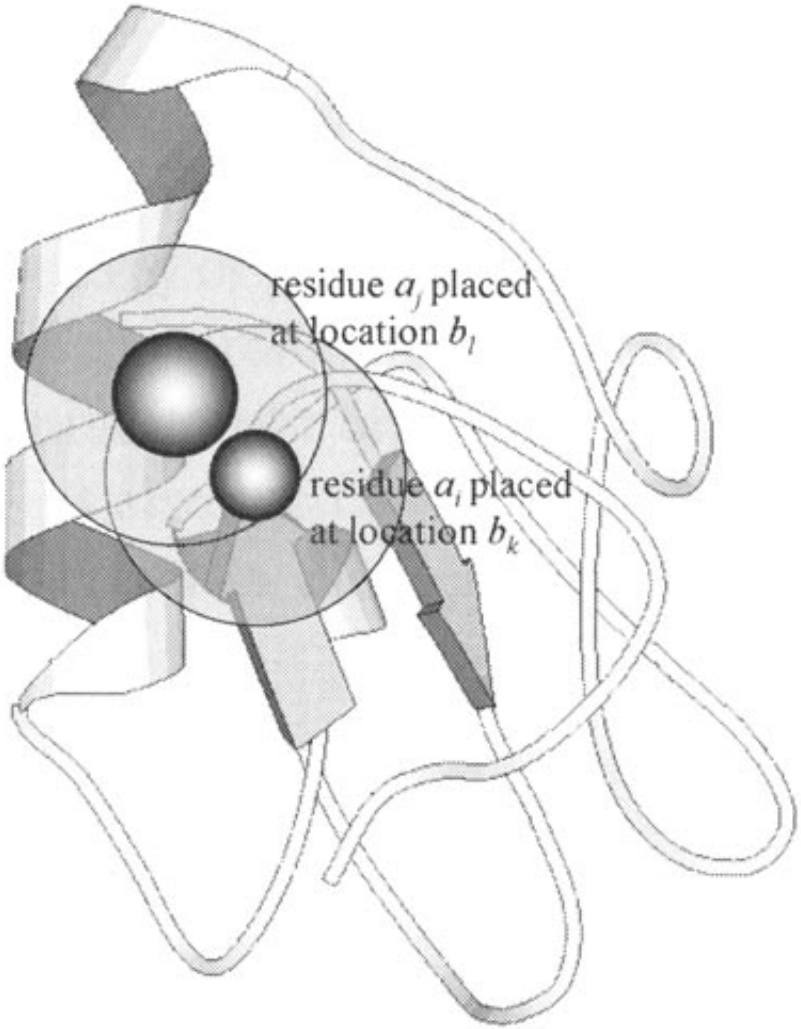
CCP (*Contact Capacity Potential*)

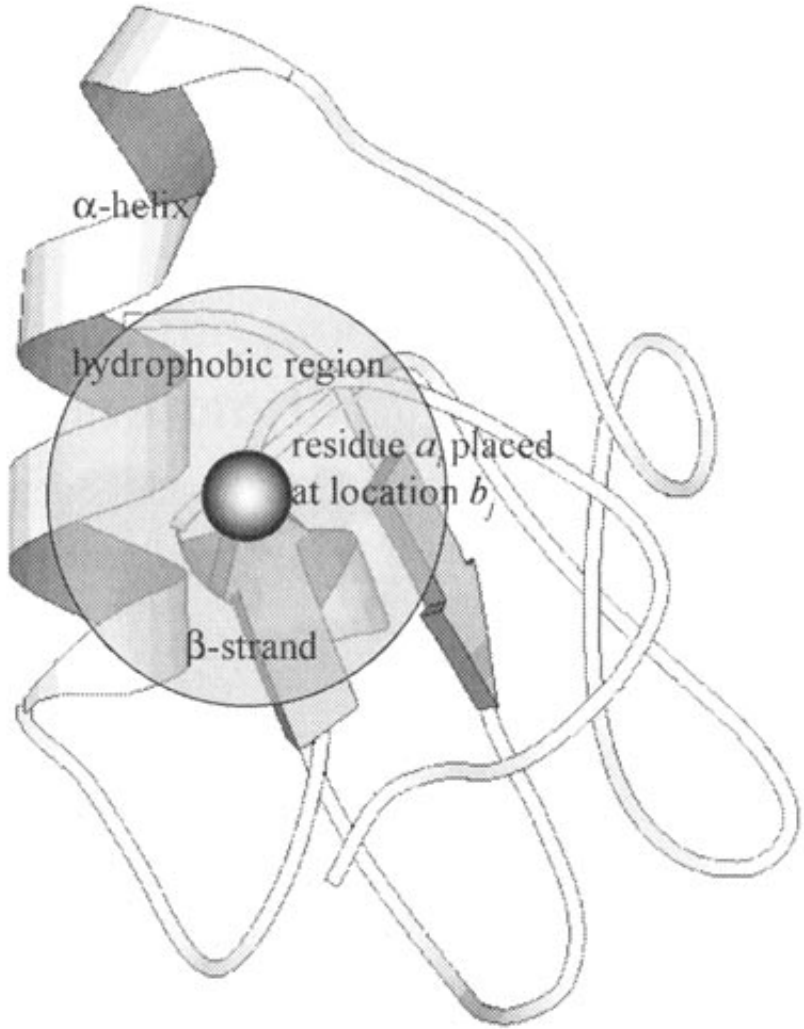
Idee

Zielstruktur und Schablone haben identisches Fold, damit auch ähnliche „Chemie“ (z.B. Hydrophobizität)

- Hydrophobe Bereiche in der Schablone sollten von Hydrophoben AS in der Zielsequenz ausgefüllt werden
- **Hydrophobe Reste haben mehr Kontakte** („im Inneren“)
- Herleitung ähnlich wie Kontaktpotentiale
  - Zählen der Reste in der Nähe einer Position der Schablone
  - Unabhängig vom Typ des Rests an dieser Position in der Schablone
- Prinzipiell ein aus der Schablone abgeleitetes Profil für die AS-Präferenz an einer bestimmten Position
- **Profil nicht abhängig von der zu threadenden Sequenz!**

CCP vs. CP

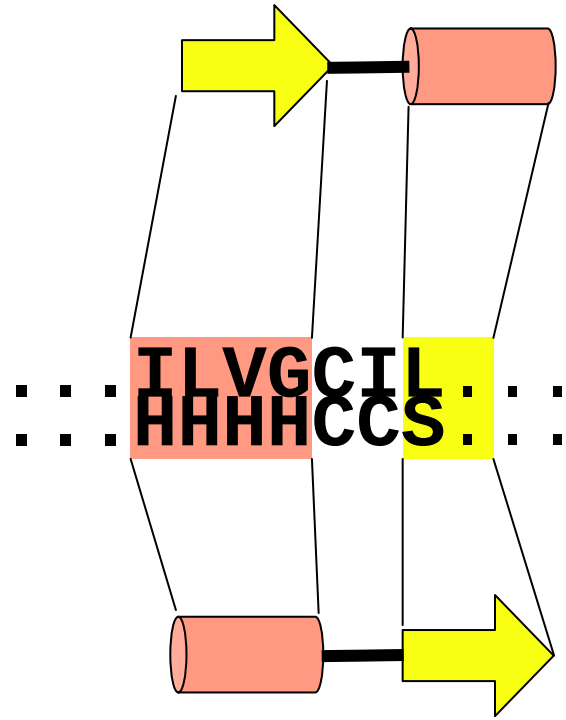




Lengauer, Zimmer, In: Bioinformatics: From Genomes to Drugs, T. Lengauer (Hrsg.), Wiley, 2002

Sekundärstruktur-Präferenzen

- Gewöhnliche Sekundärstruktur-Vorhersagen
- Vorhersage der Sekundärstruktur-Elemente für Zielsequenz
- Bewertungsfunktion beschreibt wie groß die Präferenz der AS x_i für die Sekundärstruktur an der Zielposition ist
  - Bewerte korrekte Sekundärstrukturen im Alignment günstig, unpassende ungünstig
  - Gewichte Alignment mit den Präferenzen der jeweiligen AS für die Sekundärstruktur in Schablone



Komplexität I

Satz:

Threading mit

- paarweisen Potentialen und
- beliebigen Gaps

ist **NP-hart.**

Beweis:

- 1994 durch Lathrop durch **Reduktion auf ONE-IN-THREE 3SAT**
- ONE-IN-THREE 3SAT: Variante von 3SAT, wobei *exakt* ein Literal zur Erfüllung jeder Klausel notwendig ist
- Wie 3SAT ist **ONE-IN-THREE 3SAT NP-vollständig**

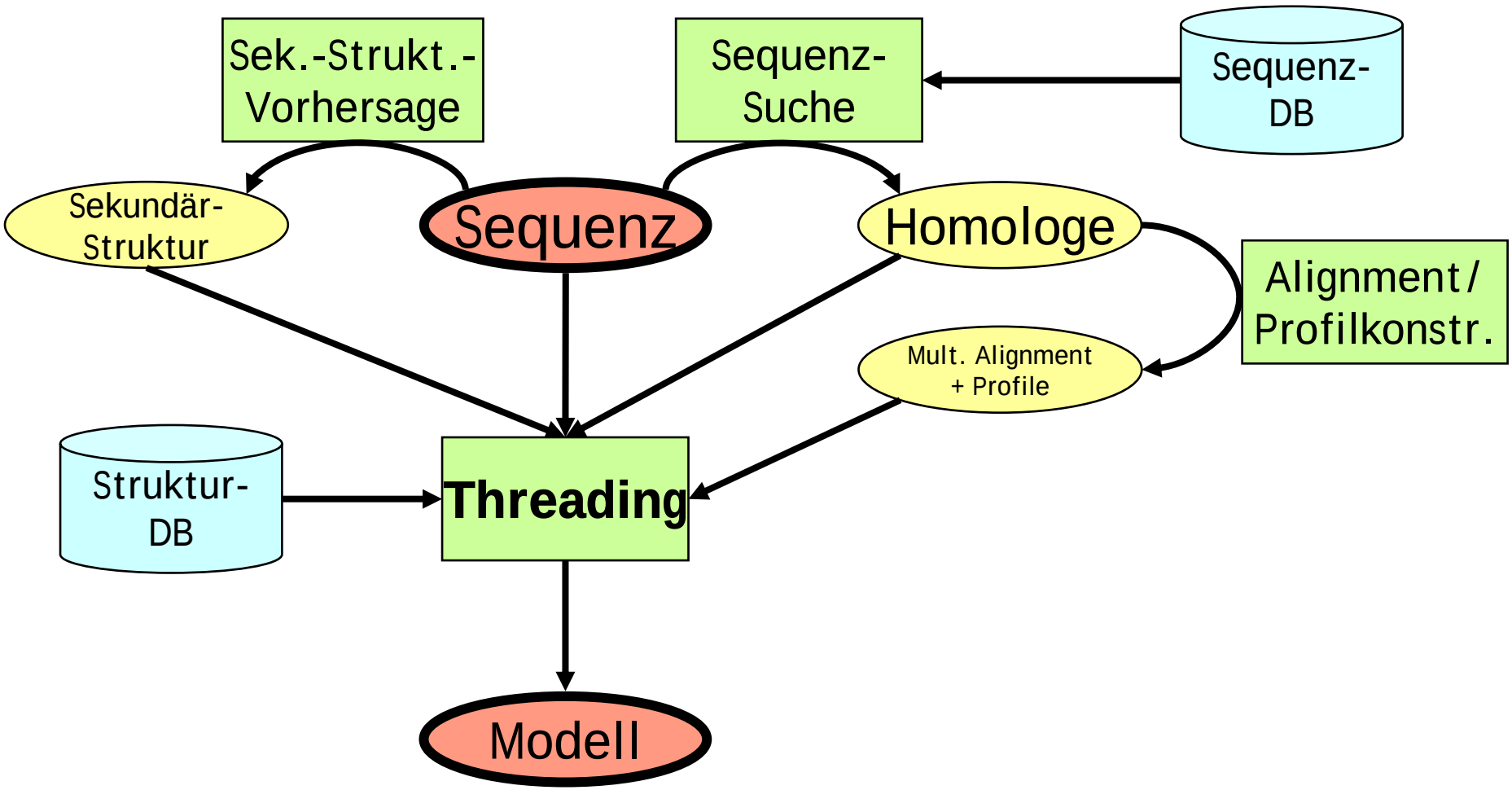
Lathrop, Protein Engin. (1994), 7, 1059

Komplexität II

- Dieses Ergebnis würde man intuitiv analog zum Faltungsproblem erwarten: **Threading = inverses Faltungsproblem**
- Verursacht durch
  - exponentielle Zahl möglicher Alignments
  - Nichtlokalität der Potentialfunktion
    (Wechselwirkung in der Sequenz beliebig
    angeordneter AA miteinander)

  => paarweise Potentiale für Threading ungünstig!
- Alignments ohne Gap resultieren in einer linearen Anzahl Alignments
- Einfachere Potentialfunktionen (CCP) lösen die Nichtlokalität auf
  - Alignmentproblem wird zu gewöhnlichem Profilalignment
  - Mit DP in polynomieller Zeit lösbar ($O(N^2)$/$O(N^3)$, Gotoh- oder Needleman-Wunsch-Algorithmus)

Threading – Schema



Algorithmen (Auswahl)

- Lathrop & Smith (1996)

  Optimales globales Alignment mit paarweisem Potential über Branch&Bound-Ansatz

- 123D (Alexandrov et al., 1996)

  Alignment (DP) mit CCP, Sekundärstrukturpräferenz

- SAM-T02 (Karplus et al., 2003)

  HMM/ANN-basierter Ansatz

- 3D-PSSM (Kelley et al., 2000)

  Sekundärstrukturpräferenzen, Solvationspotentiale, DP

- .....

Lathrop & Smith 1996

- Verwendung von Paarpotentialen
- Suchraum typischerweise 10^{30} mögliche Threadings pro Schablonenstruktur
- Rechnerisch sehr aufwändig, in der Regel nicht praktikabel
- Liefert beweisbar **global optimale Lösung**
- **Ansatz**
  - Branch & Bound
  - Reduktion des Problems („Core Threading“)

Richard H. Lathrop, Temple F. Smith, J. Mol. Biol. (1996), 255, 641

Lathrop & Smith – Formalismus

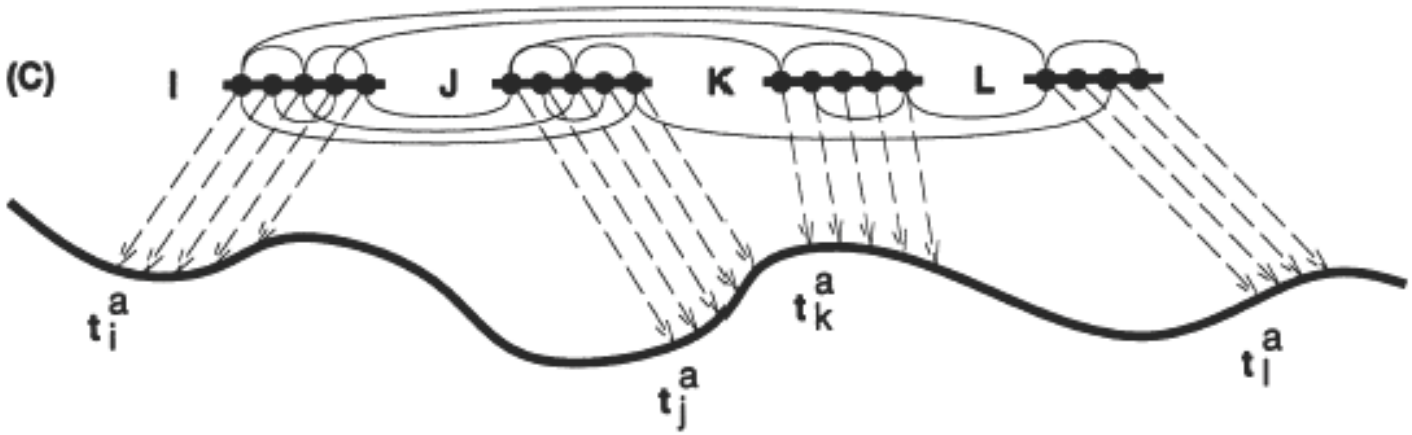
Sequenz *a* bestehend aus *n* AS a_i

Kern-Modell *C* aus *m* Segmenten

C_i mit Länge c_i und Positionen $C_{i,j}$

Threading $\mathbf{t}^a$ von *a* auf *C* bildet jeweils einen Rest t^a_i auf die erste Position eines Segments $C_{i,1}$ ab

$$\mathbf{t}^a = (t^a_1, t^a_2, \ldots, t^a_m)$$



Richard H. Lathrop, Temple F. Smith, J. Mol. Biol. (1996), 255, 641

Lathrop & Smith – Potential

Generalisiertes Paarpotential $f(\mathbf{t})$

$$f(\mathbf{t}) = \sum_i g_1(i, t_i) + \sum_i \sum_{i<j} g_2(i, j, t_i, t_j) + \ldots.$$

$g_1(i,\ t_i)$:

Energie für Reste in Segment *i* beginnend ab Sequenz-Position t_i

$g_2(i,\ j,\ t_i,\ t_j)$

Paarweise Segment-Segment-WW der Reste in Segmenten *i* und *j*

$g_3 \ldots g_m$

analog definierte Terme höherer Ordnung, in der Regel vernachlässigbar

Lathrop & Smith – Potential

Generalisiertes Paarpotential $f(\mathbf{t})$

$$f(\mathbf{t}) = \sum_i g_1(i, t_i) + \sum_i \sum_{i<j} g_2(i, j, t_i, t_j) + \ldots.$$

g_1 enthält **lokale Beiträge** (z.B. Sekundärstruktur-Präferenzen), Paarweise WW innerhalb des Segments.

Im Falle des Sippl-Potentials wird $g_1(i, t_i)$ also zur Summation über Energien der Paare (k, l) innerhalb des Segments C_i

g_2 enthält alle **paarweisen Beiträge** von Reste aus unterschiedlichen Segmenten. g_2 ist für den NP-Charakter des Problems verantwortlich. Beschränkung auf g_1 oder triviale g_2 macht das Threading in polynomieller Zeit lösbar.

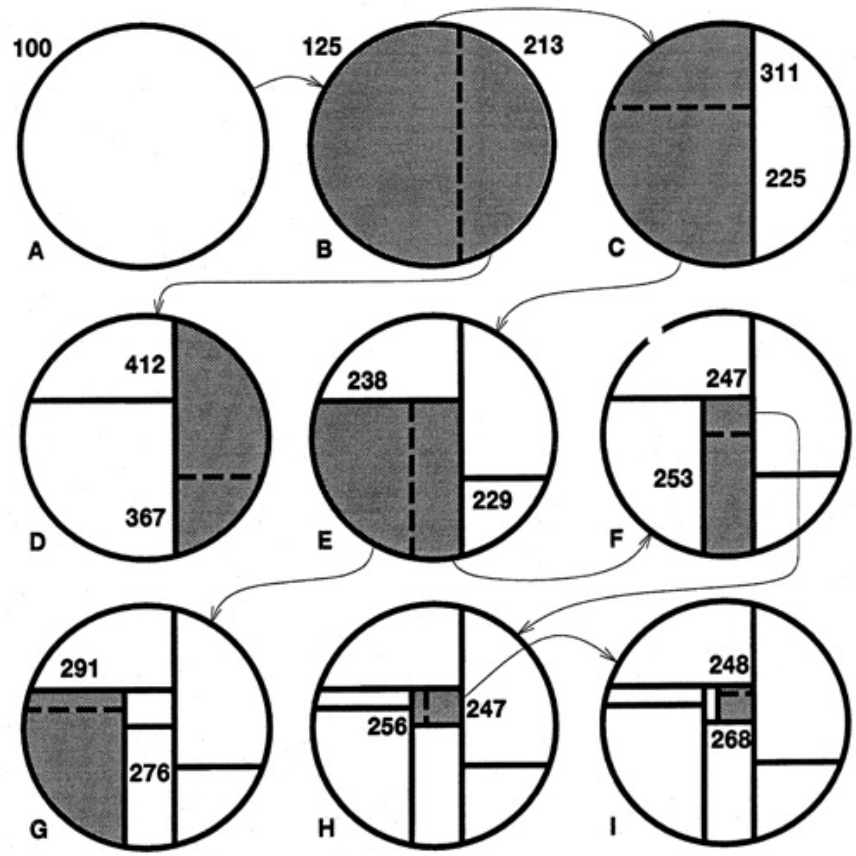
Branch & Bound

Grundannahmen

- Gesamtproblem *P* in Teilprobleme p_i zerlegbar
- untere Schranke $u(p_i)$, die schnell berechenbar ist

Algorithmus

- Calc **u(P)**
- priority queue $q = \{(u(P), P)\}$
- Iteration:
  - $p = q$.insert$(u(P), P)$
  - Wenn Anzahl Threadings in $p = 1$:
    - Abbruch, fertig!
  - Zerlege p in Teilprobleme p_i
  - p_i: q.insert$((u(p_i), p_i)$



Richard H. Lathrop, Temple F. Smith, J. Mol. Biol. (1996), 255, 641

Aufteilung in Teilprobleme

- **Menge von Threadings T** = $\{\mathbf{t} \mid b_i > t_i > d_i\}$
- **Intervalle** $[b_i, d_i]$ definieren Sequenzindices in denen die Segmentgrenze C_i liegen darf
- $\mathbf{b} = (b_1, b_2, \ldots, b_m)$ und $\mathbf{d} = (d_1, d_2, \ldots, d_m)$ spannen m-dimensionales Rechteck auf
- Alle möglichen Threadings sind Punkte darin
- Nicht alle dieser Threadings sind legal
  - **Reihenfolge** muss beachtet werden
  - **Abstands-Constraints**
  - **Illegale Threadings** werden ignoriert ($f(\mathbf{t}^{ill}) = +1$)

Aufteilung in Teilprobleme

- Splitten der Probleme erfolgt durch Auswahl eines Segments C_i und eines zugehörigen Punkts t_i^{split}
- $[b_i, d_i]$ wird dann in drei Intervalle aufgeteilt
  - $[b_i, t_i^{split} - 1]$
  - $[t_i^{split}, t_i^{split}]$
  - $[t_i^{split} + 1, d_i]$
- Jedes dieser Teilintervalle definiert drei unabhängige Mengen von Threadings mit eigener unterer Schranke
- Wahl des Splitpunkts durch Heuristik

Lathrop & Smith, 1996

- Leistung von B&B hängt sehr stark von der Qualität der Schranke ab
- Lathrop & Smith definieren mehrere untere Schranken
- Einfachste Version: Summation über die Minima der Einzelterme

$$\min_{\mathbf{t}\in\mathcal{T}} \; f(\mathbf{t}) = \min_{\mathbf{t}\in\mathcal{T}} \sum_i \left[g_1(i,t_i) + \sum_{i<j} g_2(i,j,t_i,t_j) \right]$$

$$\geq \sum_i \left[\min_{b_i \leq x \leq d_i} g_1(i,x) + \sum_{j>i} \min_{\substack{b_i \leq y \leq d_i \\ b_j \leq z \leq d_j}} g_2(i,j,y,z) \right]$$

Lathrop & Smith, 1996

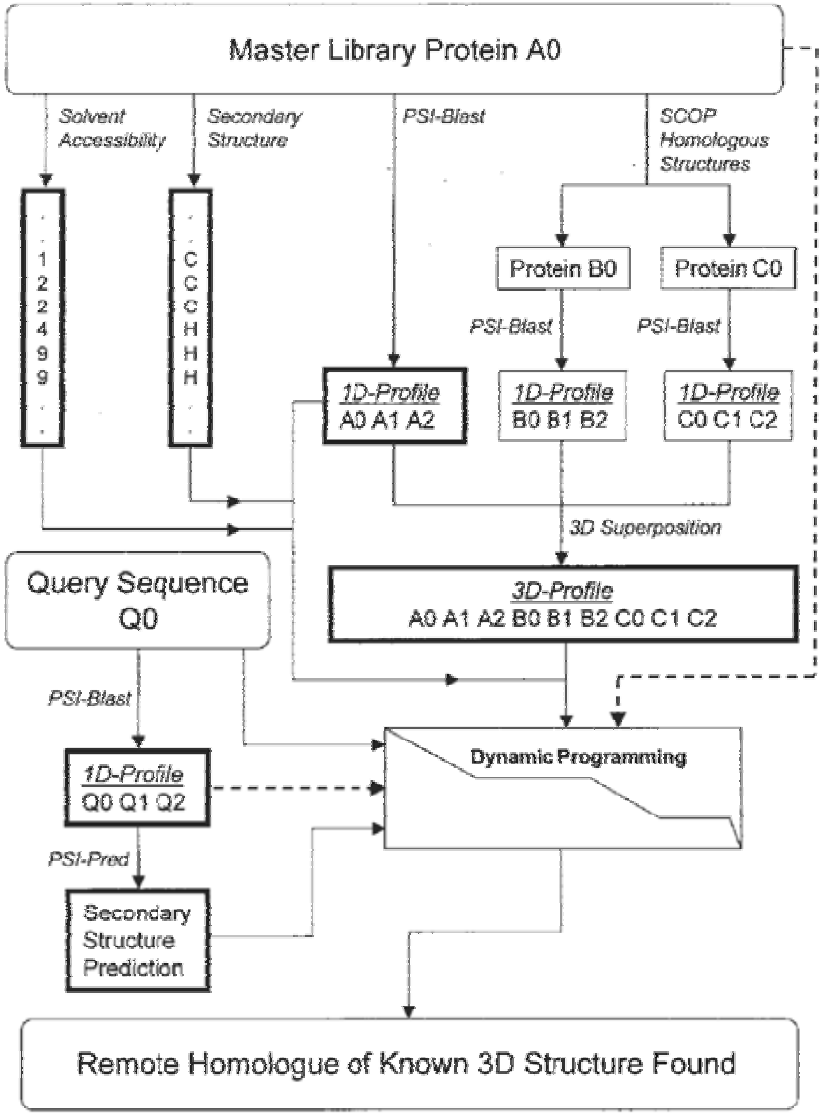
- Diese Schranke erlaubt das effiziente Durchmustern von Suchräumen bis ca. 10^{12}
- Bessere Schranken ermöglichen das Durchmustern von Suchräumen bis zu 10^{35}
- Schwierig sind effiziente Implementierung der Schranke und die Sicherstellung der Legalität der Threadings
- Details finden sich im Anhang von Lathrop & Smith

3D-PSSM

- *3D-PSSM = 3D position-specific scoring matrix*
- Kelley, McCallum, Sternberg (1999)
- Berechnung eines Profils (= *scoring matrix*) aus
  - Solvatationspotential (Lösemittelzugänglichkeit)
  - Nahen Homologen (PSIBLAST, iterativ)
- DP-Alignment der Zielsequenz auf das Profil
- Einbeziehung von Sekundärstrukturvorhersage (PSIPRED)

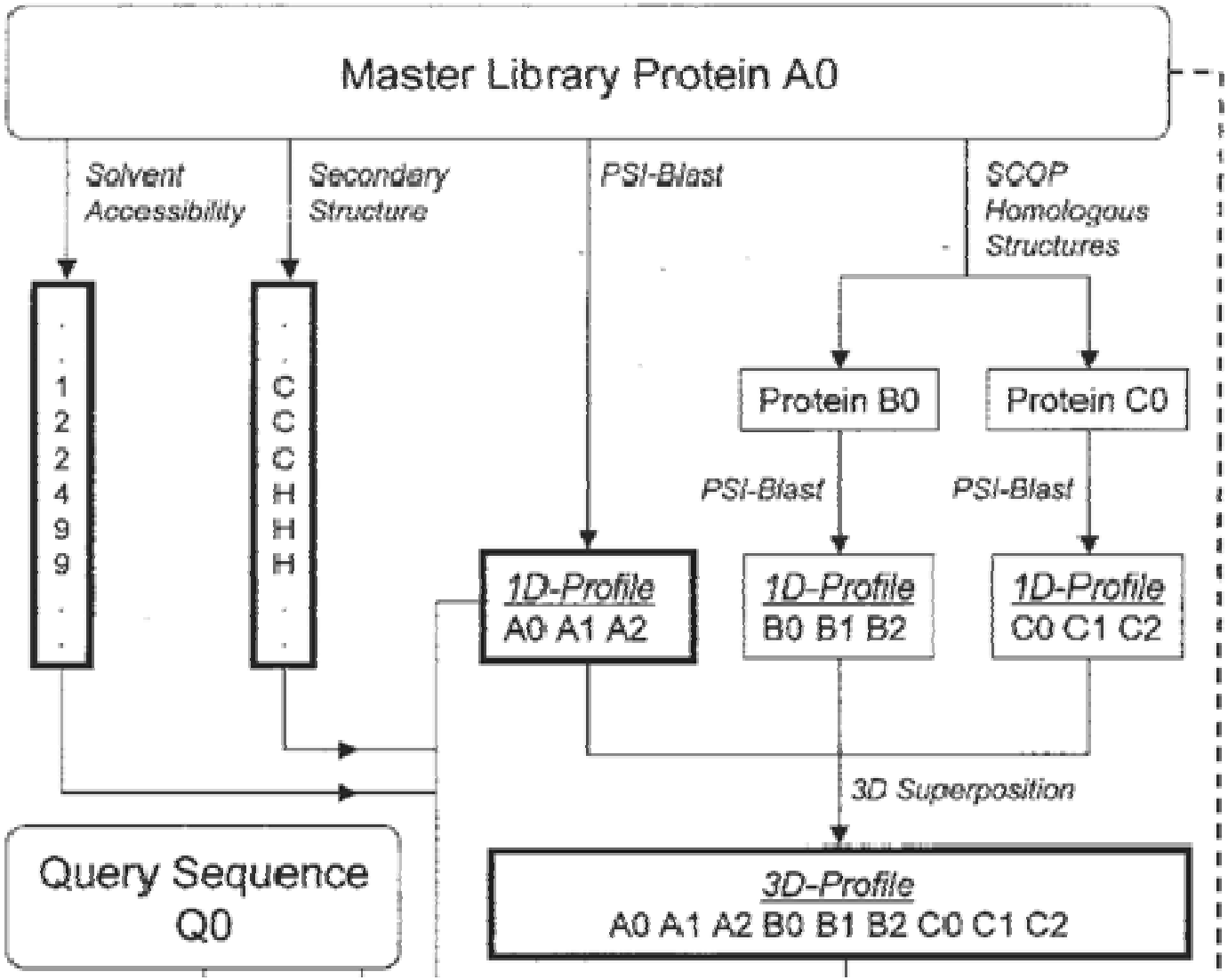
L.A. Kelley, R. Maccallum, M.J.E. Sternberg, RECOMB 99, 218-225

3D-PSSM



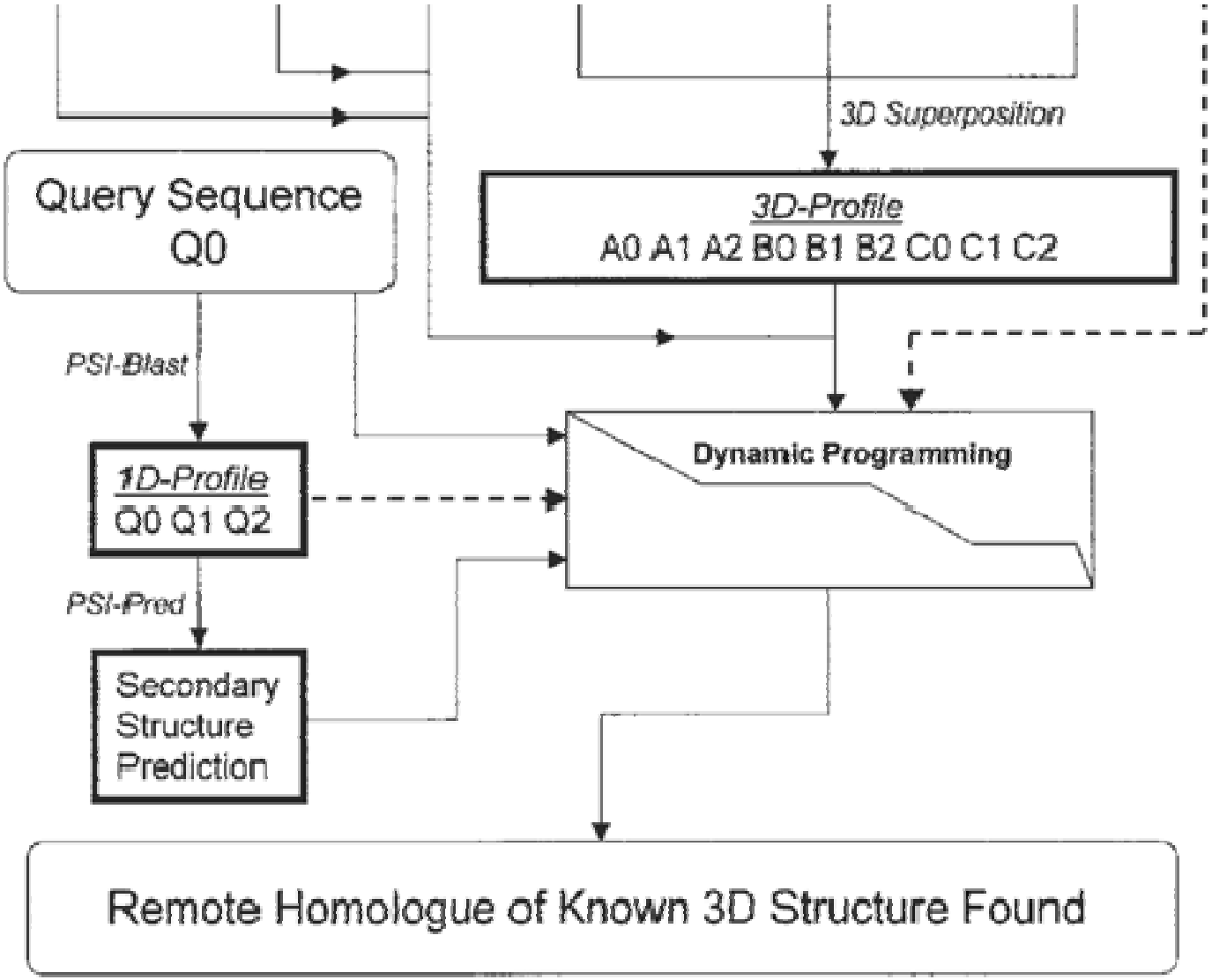
Lengauer, Zimmer, In: Bioinformatics: From genomes to drugs, Wiley, 2002

3D-PSSM



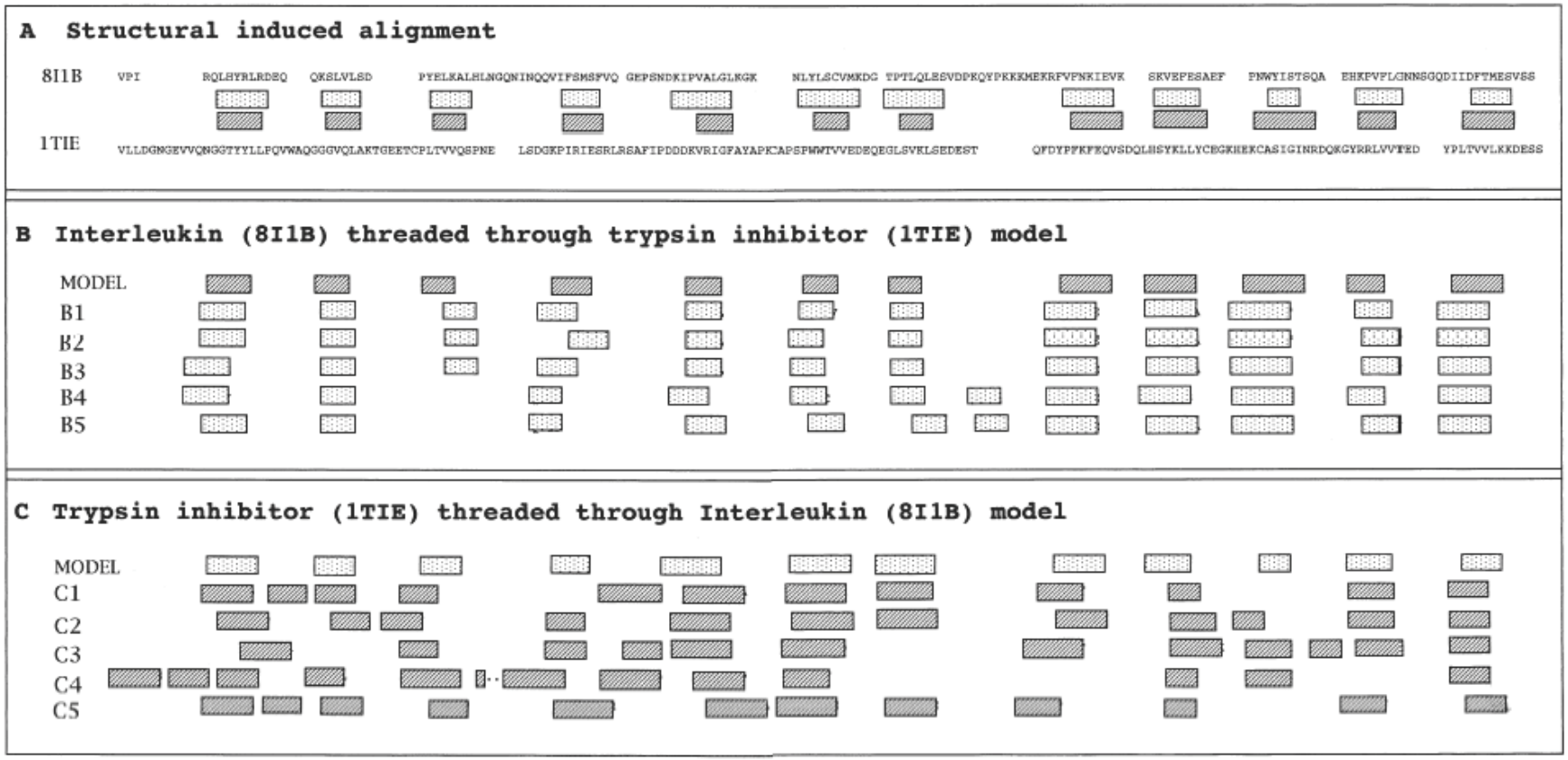
Lengauer, Zimmer, In: Bioinformatics: From genomes to drugs, Wiley, 2002

3D-PSSM



Lengauer, Zimmer, In: Bioinformatics: From genomes to drugs, Wiley, 2002

Abhängigkeit vom Potential



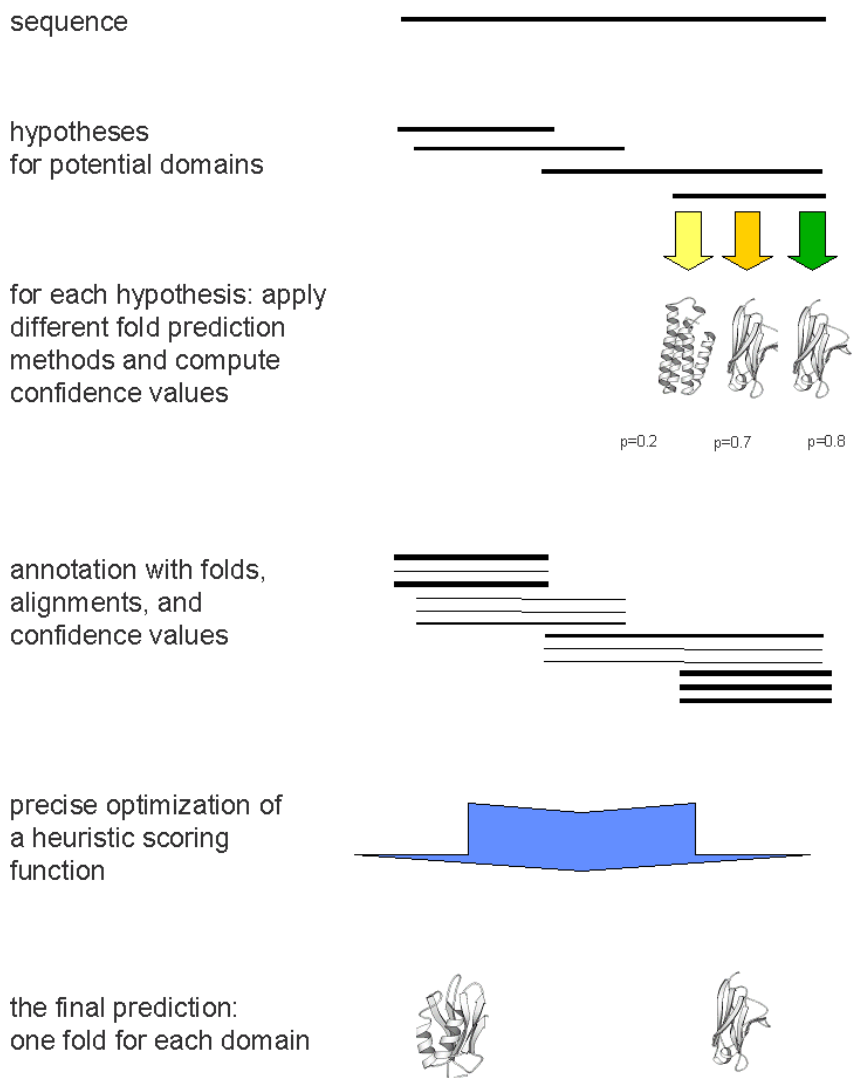
- Optimalen Threadings von 1TIE und 8I1B auf die jeweils andere Struktur
- Vergleich der Potentiale von
  (1) Bryant & Lawrence (1993), (2) Maiorov & Crippen (1992),
  (3) Miyazawa & Jernigan (1985), (4) Sippl (1990, 1993), (5) White et al. (1994)

Optimale vs. suboptimale Lösungen

- Alle der eben gezeigten Lösungen sind optimal!
- **Optimalität nur bezüglich des jeweiligen Potentials**
- Potentiale nicht exakt sondern nur approximativ
- Biologisch relevant sind meist Lösungen die bezüglich der Potentiale suboptimal sind
- **Häufig ist die echte Lösung** zwar nahe des Optimums, aber selbst **suboptimal**
- Suboptimale Lösungen genauso wichtig wie optimale Lösungen

Domänenidentifizierung

- **Threading nur für Einzeldomänen** sinnvoll
- Vorverarbeitungsschritt (z.B. in Arby) ist die **Aufspaltung in wahrscheinliche Domänen** in Loop-Regionen
- Threading aller Domänenhypothesen, dann Auswahl des wahrscheinlichsten Ergebnisses



Konstruktion des initialen Modells

- Trivialstes Vorgehen
  - Verwerfe divergente Regionen (Loops)
  - Übernehme Backbone-Koordinaten der Schablone
- Loop-Regionen können übernommen werden soweit sie konserviert sind
- Resultat ist ein rohes Modell ohne Seitenketten
- Modell enthält in der Regel Lücken (Gaps im Alignment)

Homologiemodellierung

- Homologiemodellierung konstruiert ein vollständiges Modell eines Proteins auf eine Schablonenstruktur
- Grundsätzlich folgende Schritte
  - Rohes Modell (Backbone)
  - Hinzufügen der Seitenketten in den Kernregionen
  - Modellierung der Loops
  - Modellierung und Optimierung aller Seitenketten

Beispiel

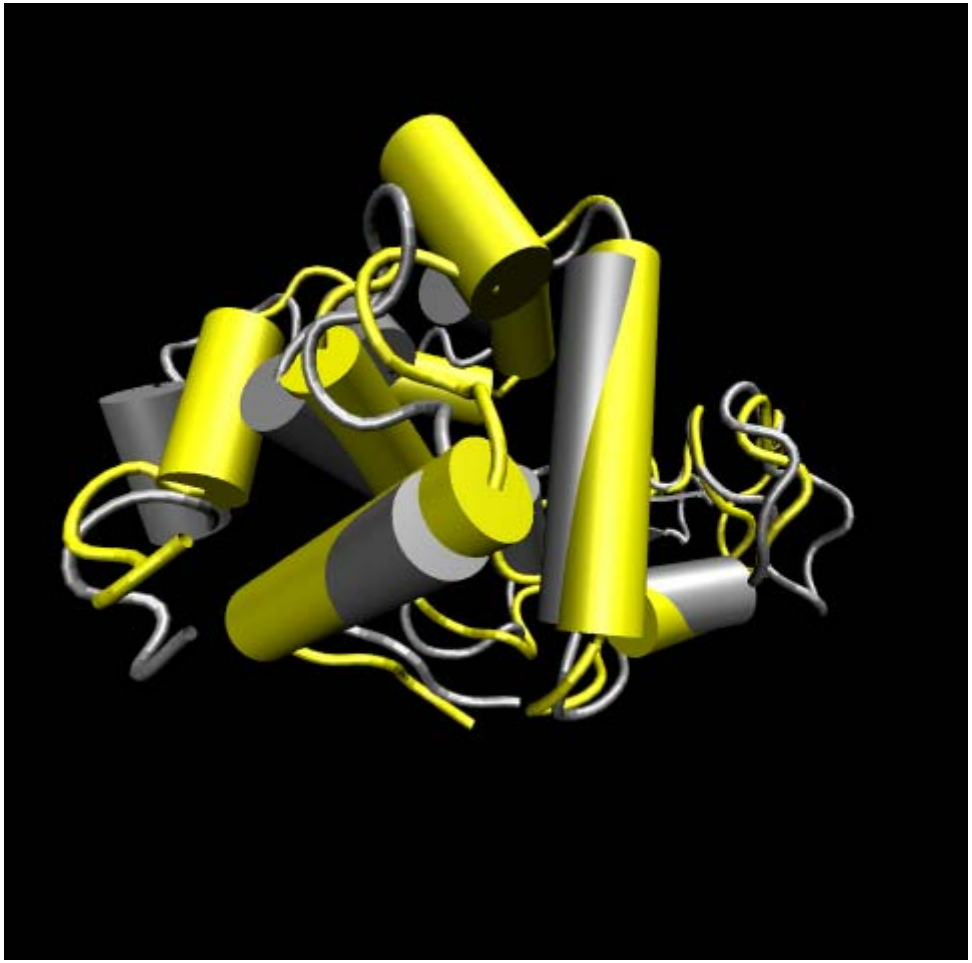
- Lysozyme sind Zucker spaltende Enzyme
- Spielen eine Rolle bei der mikrobiellen Verteidigung (Auflösen der Zellwand)
- Triviales Beispiel für Threading
  - 1LZY – Lysozym (Truthahn)
  - 1IVM – Lysoyzm (Maus)
- 57% Sequenzidentität
  => sehr einfaches Problem

SCOP:

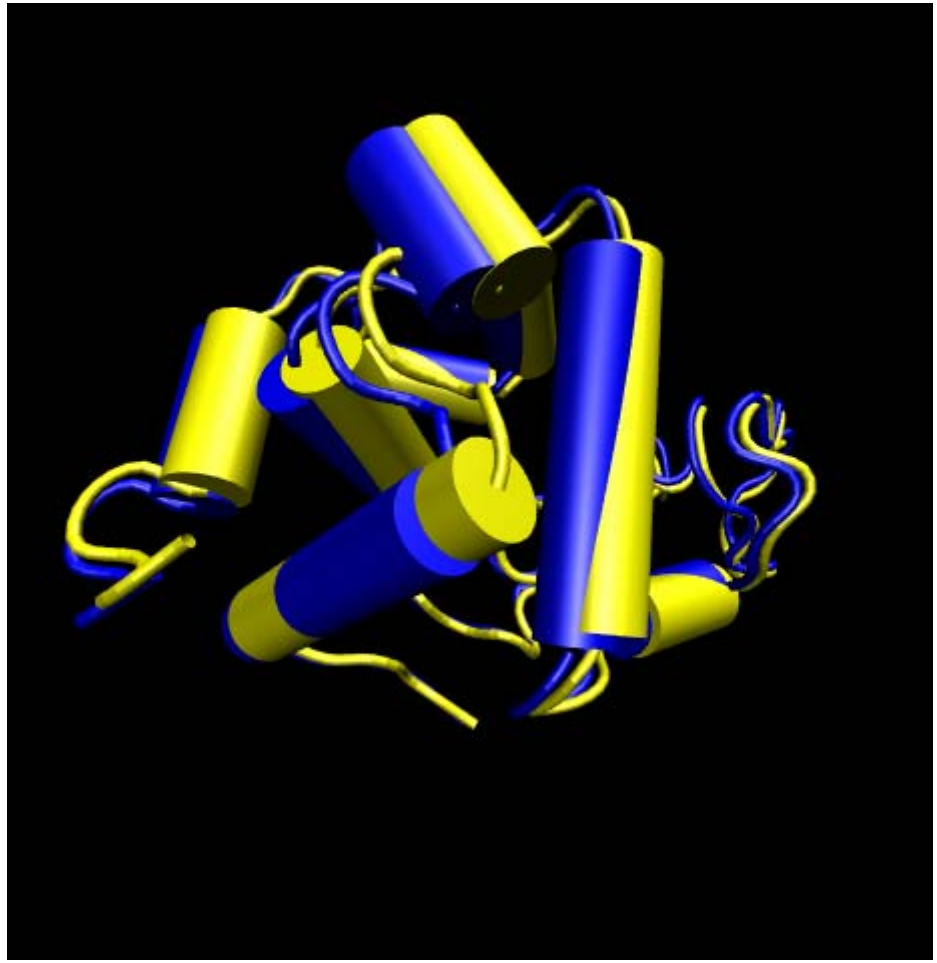
Protein:
Lysozyme from Turkey
(*Meleagris gallopavo***)**

Lineage:
Class: Alpha and beta proteins (a+b)
Mainly antiparallel beta sheets (segregated alpha and beta regions)
Fold: Lysozyme-like
common alpha+beta motif for the active site region
Superfamily: Lysozyme-like
Family: C-type lysozyme
Protein: Lysozyme
ubiquitous in a variety of tissues and secretions
Species: Turkey (*Meleagris gallopavo***)**

Beispiel



Grau: 1IVM
Gelb: 1IVM gethreaded auf 1LZY



Blau: 1LZY
Gelb: 1IVM gethreaded auf 1LZY

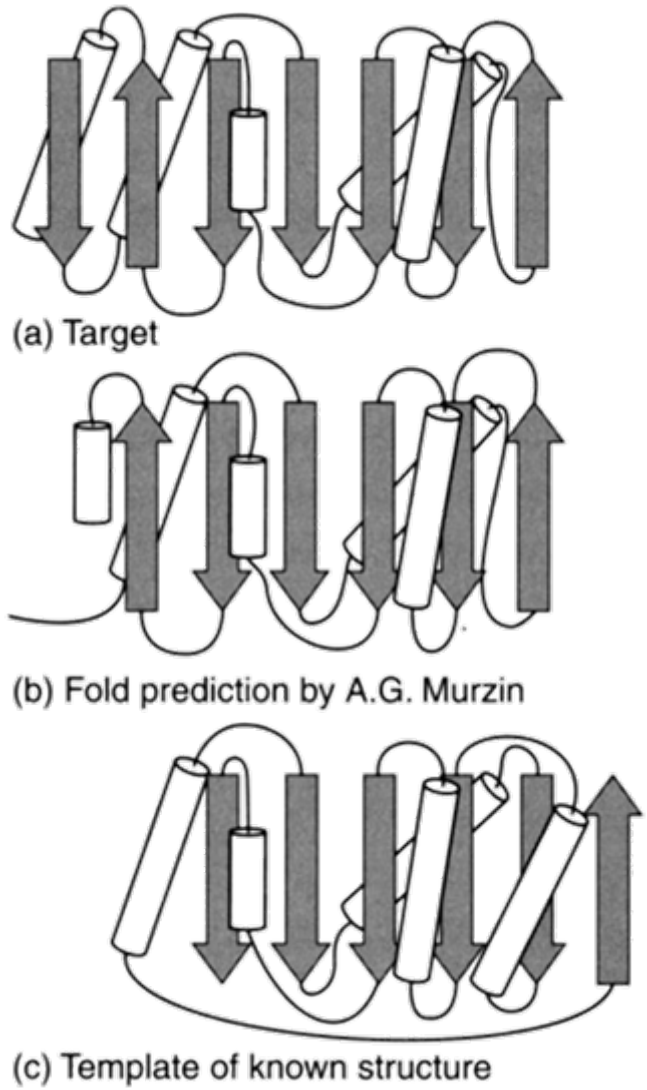
Beispiel

- Einige Sekundärstrukturelement zu sehr aus der Schablone übernommen
- Genauer Anfang/Ende von Helices/Faltblättern generell schwierig (wie bei Sekundärstrukturvorhersage)
- Einige sehr gut erhaltene Loops
  - Loops in Lysozym im aktiven Zentrum beteiligt
  - Diese Loops sind hochkonserviert

Qualität

Beispiel aus CASP4:

- Unbekanntes Protein aus *H. Influenzae*
- Die Topologie der von A. Murzin vorhergesagten Struktur (b) stimmt bis auf die Termini sehr gut
- Die Vorhersage ist näher am Target als die nächste bekannte Struktur (c)



A. Lesk, Introduction to Bioinformatics, Oxford University Press, 2002

Qualität – CASP5

- Qualitätskriterien
  - RMSD der C_α-Atome der modellierten Struktur mit der Kristallstruktur
  - Prozentsatz korrekt alinierter Reste (richtige Sekundärstruktur)
  - GDT_TS-Wert (GDT: *global distance test*)
    - Abbildung von Strukturfragmenten aufeinander
    - Maximale RMSD von *d* Angstrom
    - *Pd* = Prozentsatz der mit RMSD < *d* abbildbaren Reste

    GDT_TS = (*P1* + *P2* + *P4* + *P8*) / 4
- GDT_TS gilt als zuverlässigstes Maß

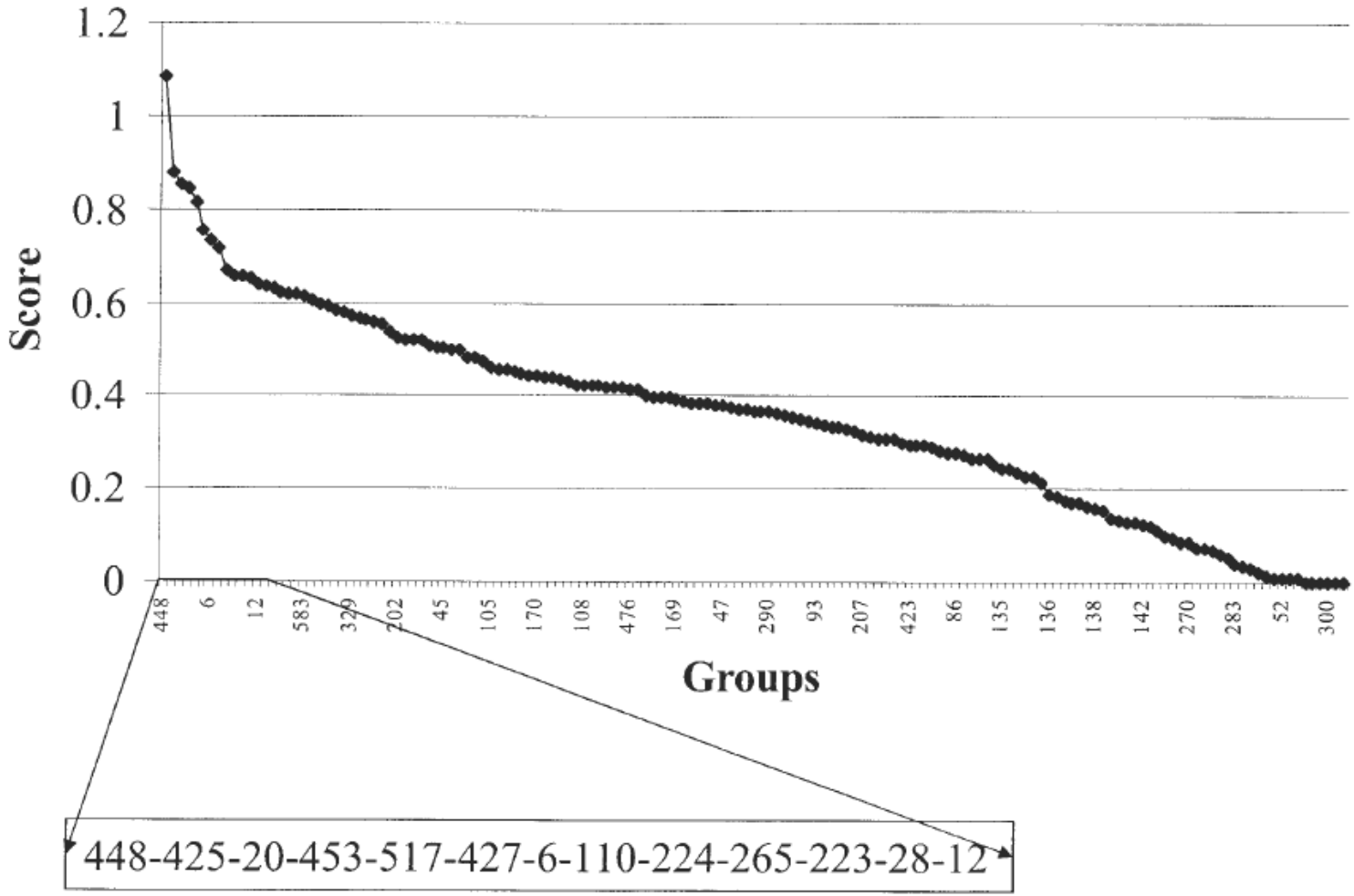
Adam Zemla, Nucl. Acid. Res. 2003, 31, 3370

Qualität – CASP5

- Ergebnisse generell recht gut, wenn > 25% Sequenzidentität zwischen Ziel und Schablone
- Es zeichnet sich ab, dass die sequenzielle Vorgehensweise (Fold Recognition, Threading, Modellierung der Schleifen, Seitenkettenmodellierung) zu restriktiv ist
- Erfolgreiche Methoden verwenden meist mehrere Schablonen-Strukturen auf die parallel modelliert wird
- Auswahl des besten Modells erst im letzen Schritt

Anna Tramontano and Veronica Morea, Proteins (2003), 53, 352-368

Qualität – CASP5



448: Alexey Murzin!

Literatur

Faltungsklassen

- **Arthur M. Lesk:** Introduction to Protein Architecture, Oxford University Press, 2001

Fold Recognition, Threading, Potentiale

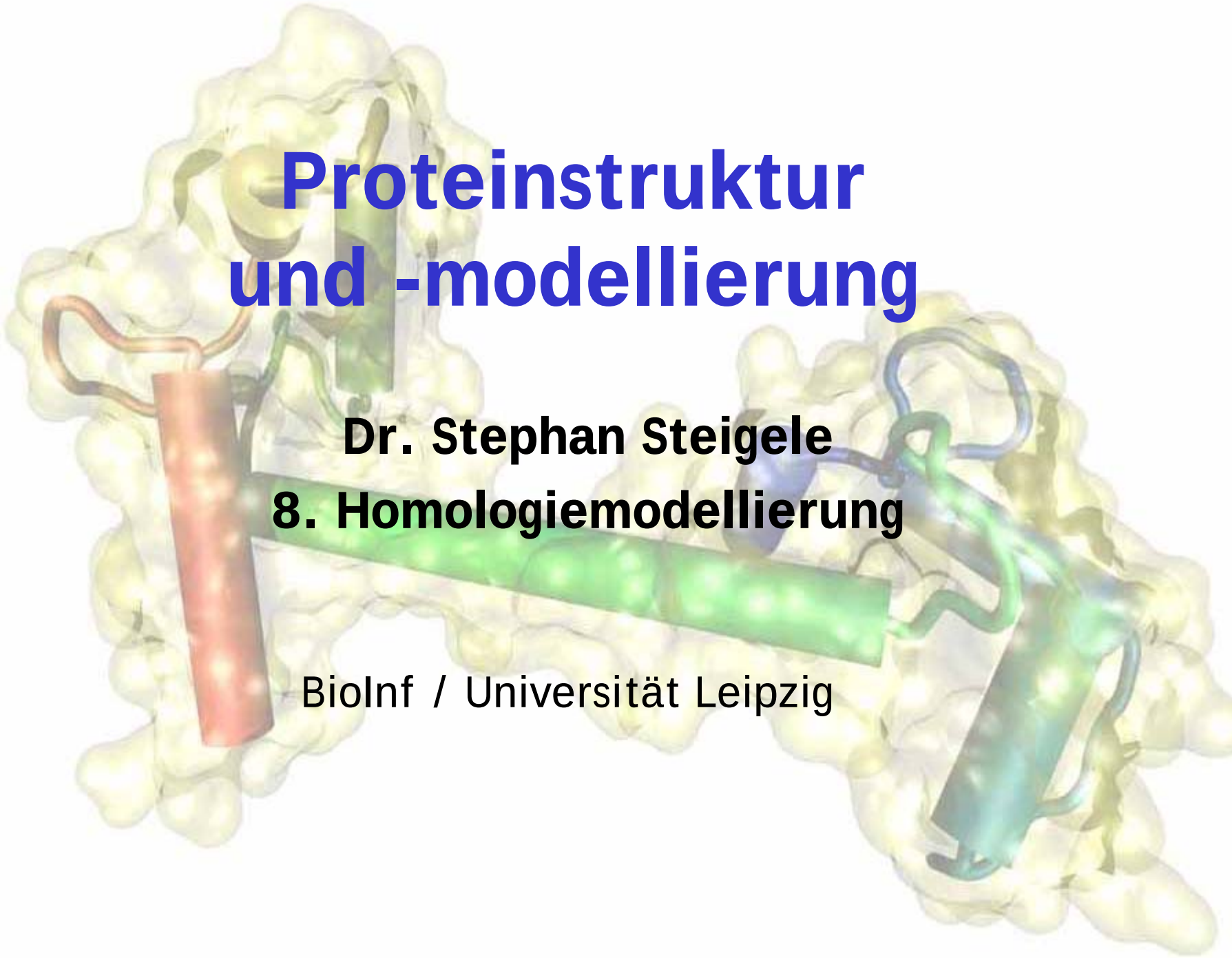
- **Adam Godzik**: *Fold Recognition Methods*, In Phillip E. Bourne, Helge Weissig (Hrsg.), Structural Bioinformatics, Wiley (2003)

- **Ralf Zimmer, Thomas Lengauer**: *Structure Prediction*, Chapter 6 in T. Lengauer (Hrsg.): Bioinformatics: From Genomes to Drugs, Wiley, 2002

- **Manfred Sippl**, Calculation of Conformational Ensembles from Potentials of Mean Force, J. Mol. Biology (1990), **213**, 859-883

- **Richard H. Lathrop, Temple F. Smith**, Global Optimum Protein Threading with Gapped Alignment and Empirical Pair Score Functions, J. Mol. Biology (1996), 255, 641-665

- **Nick Alexandrov, Ruth Nussinov, Ralf Zimmer**, *Fast protein fold recognition via sequence to structure alignment and contact potentials*. In Lawrence Hunter and Teri E. Klein (Hrsg.), Pacific Symposium on Biocomputing 1996, p. 53-72 (1996)



Proteinstruktur und -modellierung

Dr. Stephan Steigele

8. Homologiemodellierung

BioInf / Universität Leipzig

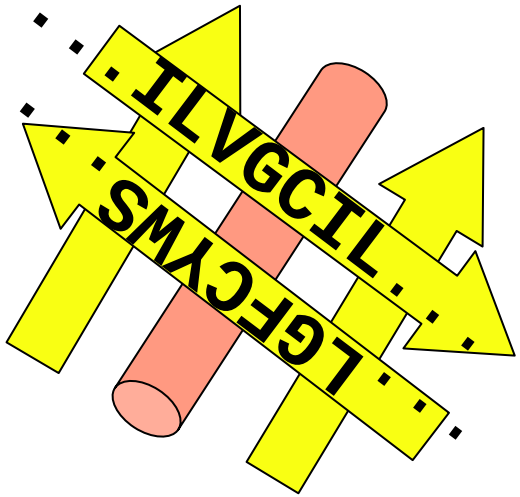
Gliederung

- Übersicht, Begriffe
- Schleifenmodellierung
  - Datenbanken
  - Algorithmen
- Seitenkettenmodellierung
  - Rotamerbibliotheken
  - Algorithmen
- Optimierung und Verifikation
- Programmpakete
  - SWISS-MODEL
  - MODELLER

Begriffe

- **Homologiemodellierung**
  - Modellierung eines **hochauflösenden Modells** (alle Atome)
  - Ausgangspunkt: **Sequenz-Struktur-Alignment**
- Wir lassen das Sequenz-Struktur-Alignment hier außer Acht (siehe Threading)
- Als gegeben nehmen wir einen partiell aufgefädelten Backbone an (C_α-Koordinaten)
- Gesucht sind dann die Koordinaten der fehlenden Atome

Homologiemodellierung



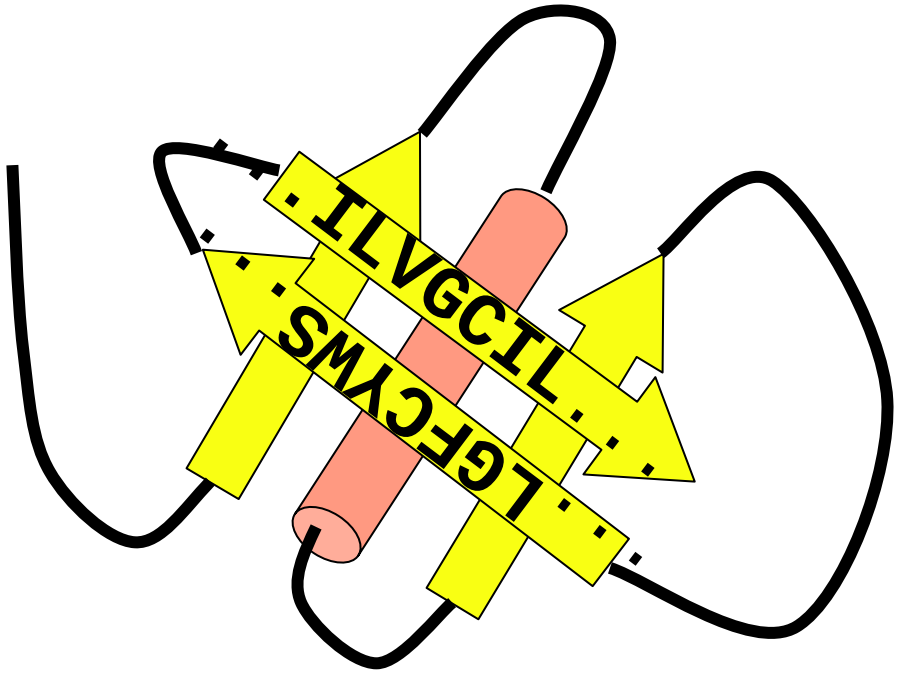
Gegeben

Sequenz-Struktur-Alignment

Gesucht

Positionen aller Atome

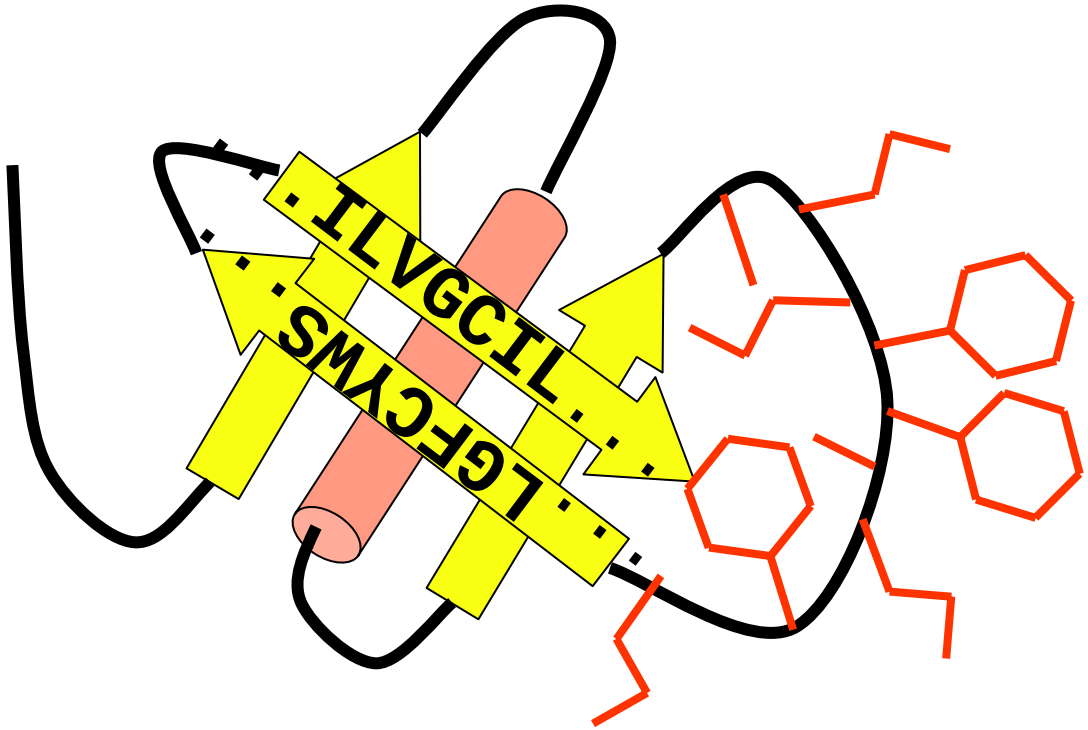
Homologiemodellierung



1. Schritt: Hinzufügen der fehlenden Loops

- Schließen der Lücken in der Struktur
- Erzeugen eines durchgängigen, sinnvollen Rückgrats

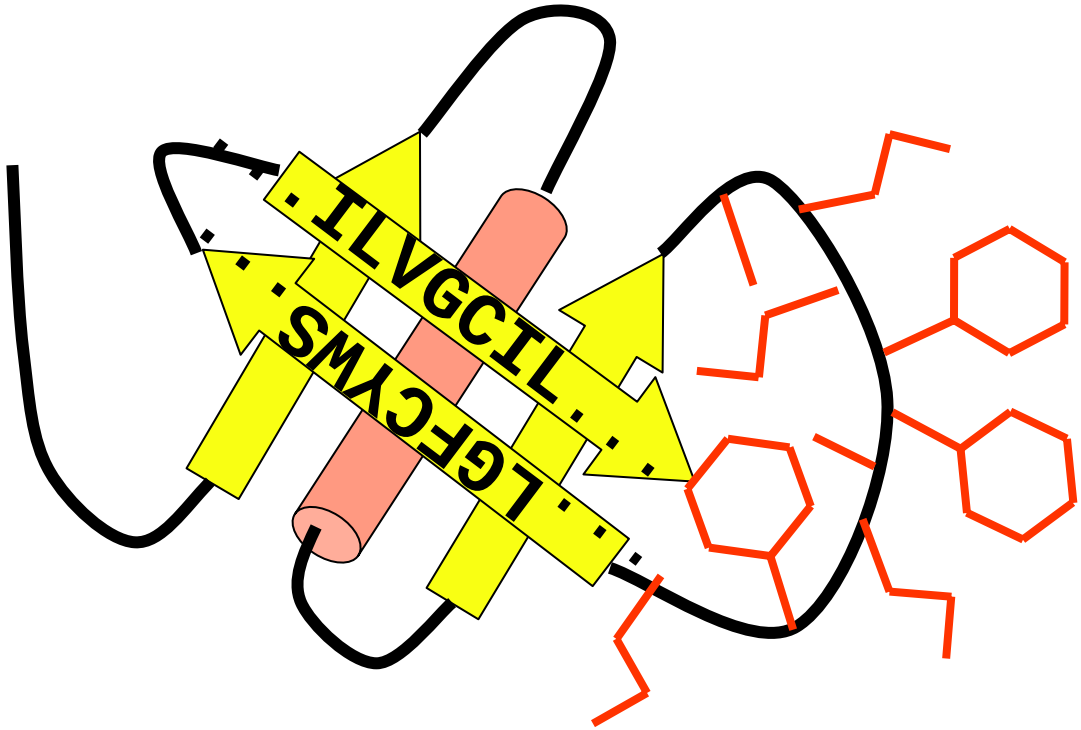
Homologiemodellierung



2. Schritt: Hinzufügen der Seitenketten

- Ergibt vollständige, physikalisch sinnvolle Struktur
- Positionierung in sinnvollen Positionen schwierig

Homologiemodellierung



3. Schritt: Optimierung der Gesamtstruktur

- Lokale Nachoptimierung, „Entspannung“ der Struktur
- Entfernt Artefakte der Konstruktionsalgorithmen

Homologiemodellierung – Überblick

- Vorgehen üblicherweise sequenziell
  - Konstruktion eines groben **Backbone-Modells**
  - Hinzufügen der divergenten **Schleifenregionen**
  - Positionierung der **Seitenketten**
  - **Optimierung** und Validierung der Gesamtstruktur
- Einige Schritte lassen sich auch zusammenfassen oder parallel durchführen, so z.B. Modellierung der Schleifen und Seitenketten

Anwendungen

- Zur Konstruktion eines Modells nach dem Threading

  => z.B. **Wirkstoffentwurf**

  - Konstruiere Modell zur Auswahl von Wirkstoffen
  - Verständnis von Mechanismen anhand der Struktur

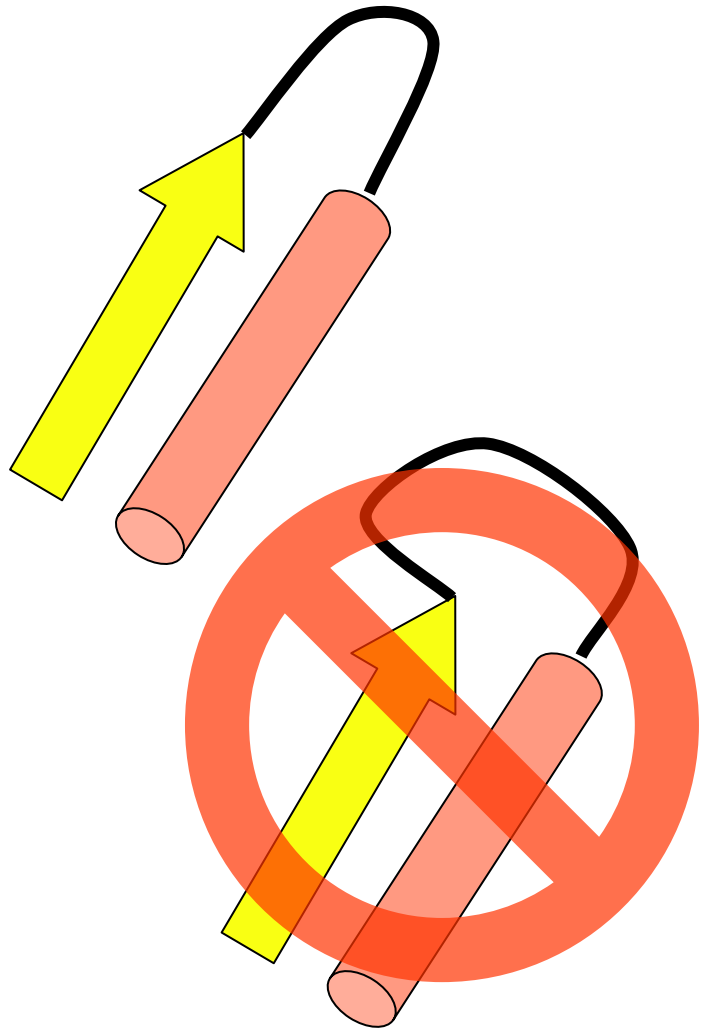
- Zur Analyse von Modellen von veränderten Proteinen (z.B. Punktmutationen)

  => z.B. **Protein-Entwurf und -Optimierung**

  - Vergleiche Strukturen und Aktivitäten verschiedener Mutanten
  - Berechne aus Struktur Stabilität der Mutanten

Loop-Modellierung

- Was beschreibt Loops?
  - Sequenz
  - Torsionswinkel (Backbone)
- Randbedingungen
  - Ende-Ende-Abstand
  - „Anschluss“ an Sekundärstrukturelemente
  - Keine Überschneidungen mit Backbone



Loop-Modellierung

Zwei grundlegende Ansätze

- **Datenbank/Fragment-basierte Ansätze**
  - Verwenden Loop-Regionen aus bekannten Proteinstrukturen
  - Chancen bei längeren Loops sehr schlecht
- **Ab-initio**
  - Suche ohne Strukturfragmente
  - Energiefunktion meist basierend auf Molekülmechanik
  - Suchraum recht groß

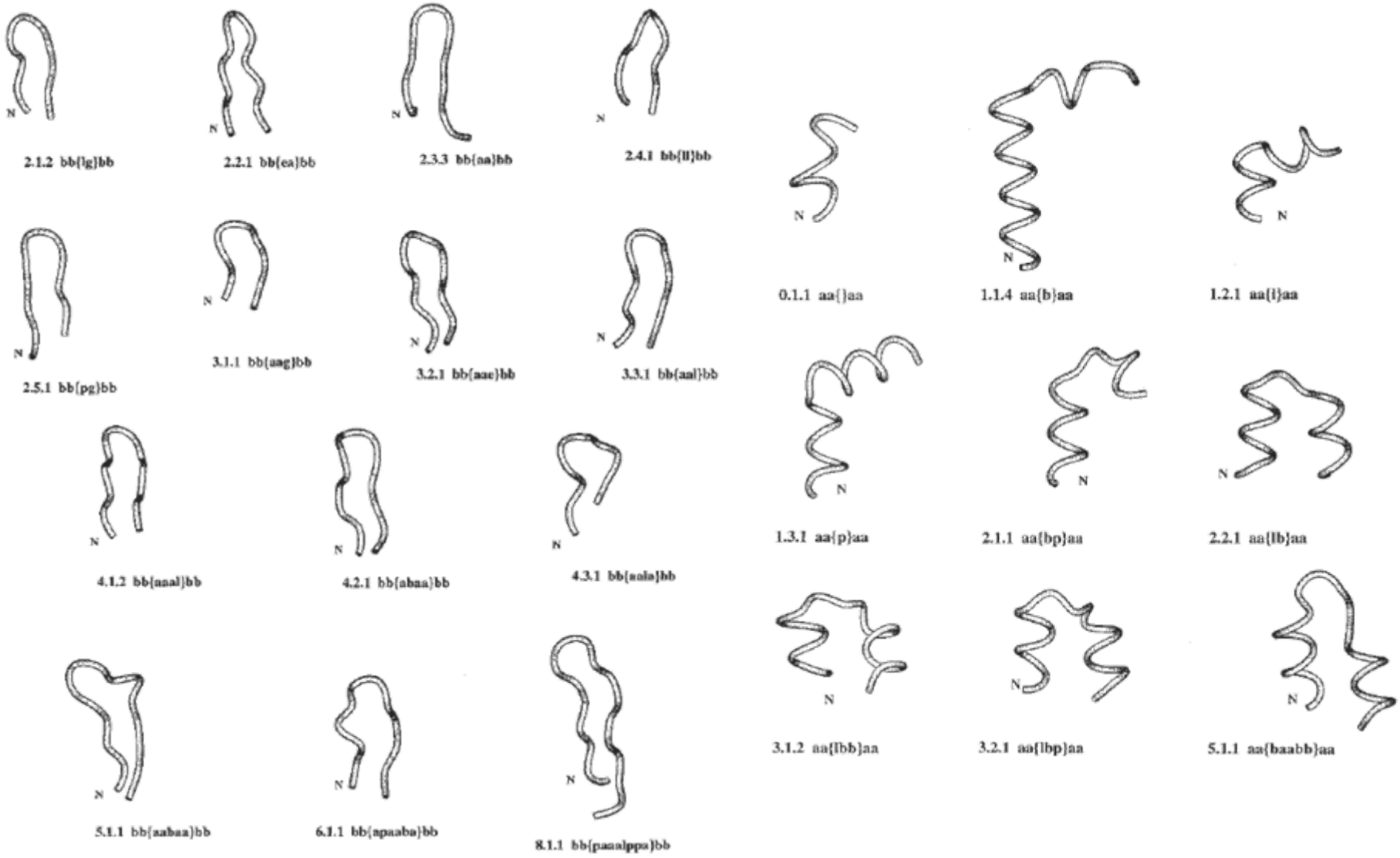
    => Meist stochastische Suchstrategien

Loop-Datenbanken

- Oliva et al. Stellten 1997 eine Loop-Datenbank vor
- Clustering der Loop-Regionen von 233 Proteinstrukturen führt zu 56 Loop-Klassen
- Fünf große Loop-Klassen nach den Sekundärstrukturen die sie verbinden
  - α-α
  - β-β
  - β-β-Haarnadeln
  - α-β
  - β-α

Oliva et al., J. Mol. Biol. (1997), 266, 814

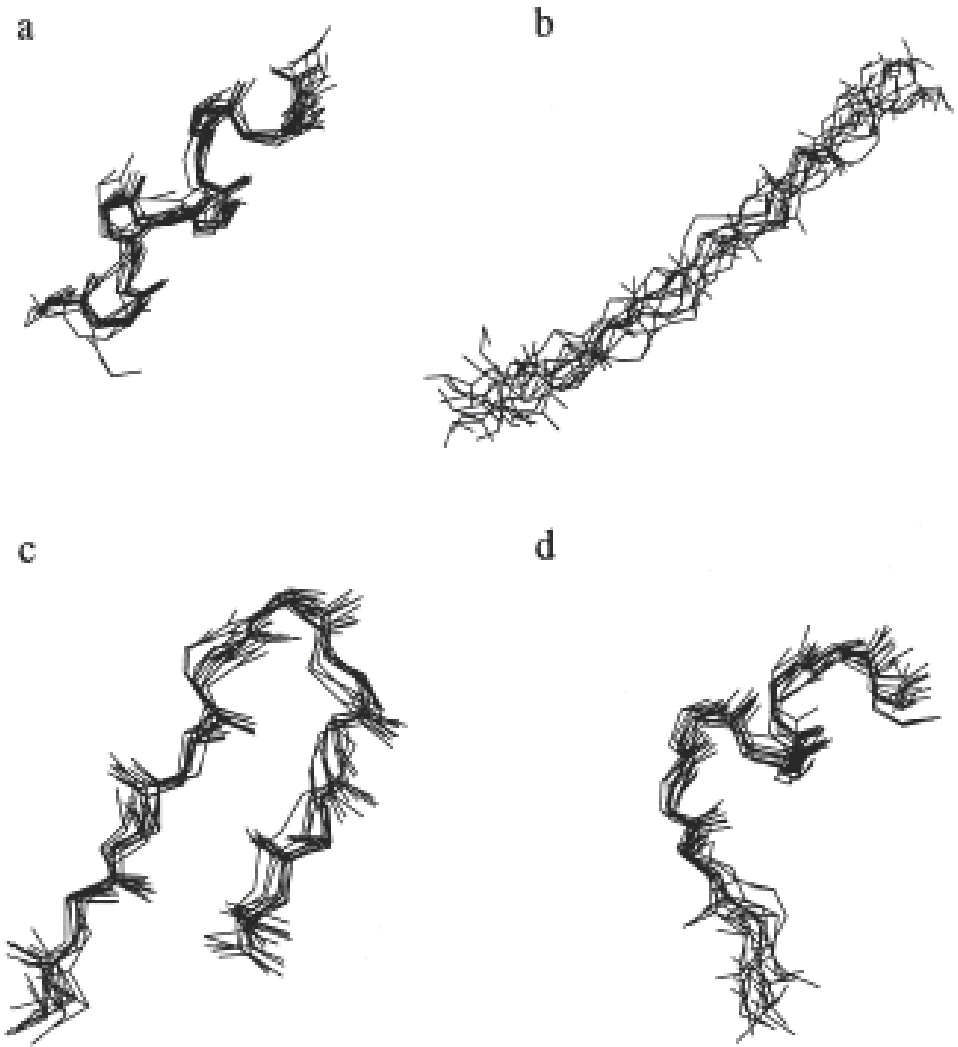
Loop-Datenbanken



Oliva et al., J. Mol. Biol. (1997), 266, 814

Loop-Datenbanken

- **Clustering** liefert große Zahl sehr ähnlicher Fragmente
- Cluster werden üblicherweise auf einzelne **Repräsentanten** reduziert
- Methoden
  - Hierarchisches Clustering
  - Nächste Nachbarn



Fechteler, Dengler, Schomburg, J. Mol. Biol. (1995), 253, 114-131

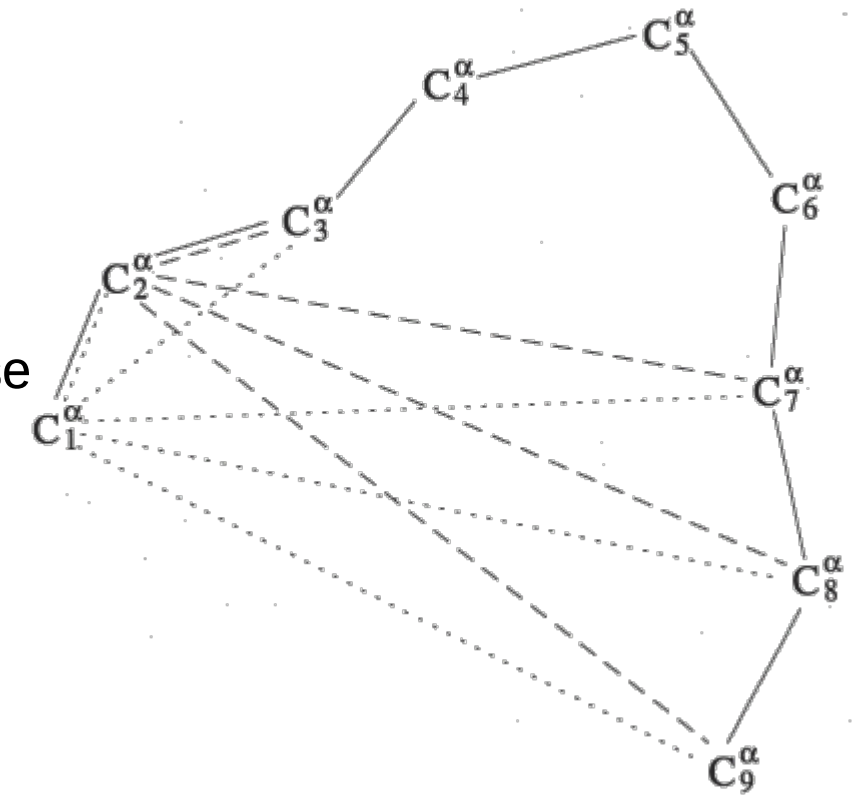
Loop-Datenbanken

- Für kurze Loops ist eine Datenbanksuche eine geeignete Möglichkeit
- Für längere Loops ist die Wahrscheinlichkeit ein geeignetes Fragment in der Datenbank zu finden zu gering
- Charakteristisch für die Schleifenregion
  - Sequenz
  - Abstand der Enden
  - Torsionswinkel („glatter“ Anschluß)
- Ausreichender Überlapp des Fragments mit modelliertem Backbone (Anker-Regionen)

DB-basierte Methoden – BRAGI

BRAGI

- Basierend auf Loop-DB die durch eine Kombination aus hierarchischem und NN-Clustering konstruiert wurde
- Alle Loops der DB werden auf die „Anker-Reste“ platziert (üblicherweise drei AS links und rechts)
- Auswahl der Fragment durch Ähnlichkeit in den Distanzen zu Ankerresten (1-3, 7-9)
- Günstigstes Fragment ist das mit den wenigsten Überlappungen mit dem Protein-Backbone

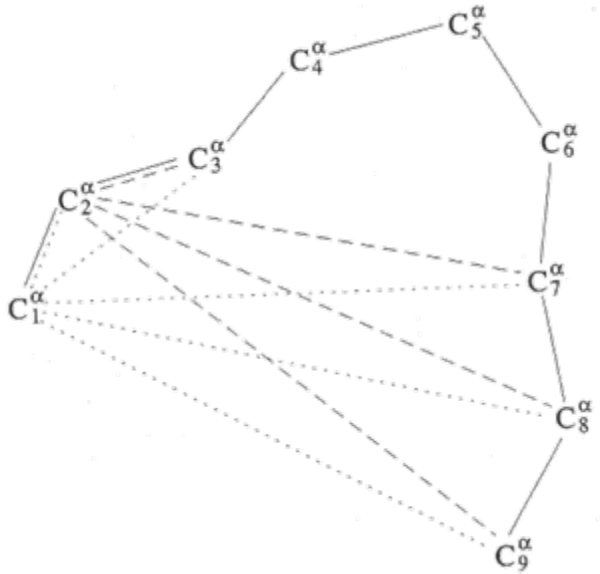


Fechteler, Dengler, Schomburg, J. Mol. Biol. (1995), 253, 114-131

DB-basierte Methoden – BRAGI

Algorithmus

- Berechnung der Matrizen mit allen C_α-Distanzen für Template und alle entsprechenden Fragmente
- Vergleich der Distanzmatrizen
- Wähle 100 beste Fragmente (geringste Abweichung der Matrixelemente = Distanzen)
- Bilde Fragment auf Ankerreste ab
- Berechne Überlappungen zwischen Fragment und Protein-Backbone
  - Radius für O, N: 0.8 Å
  - Radius für C: 0.9 Å
- Ordne Fragmentliste nach Anzahl Kollisionen
- Baue bestes Fragment ein

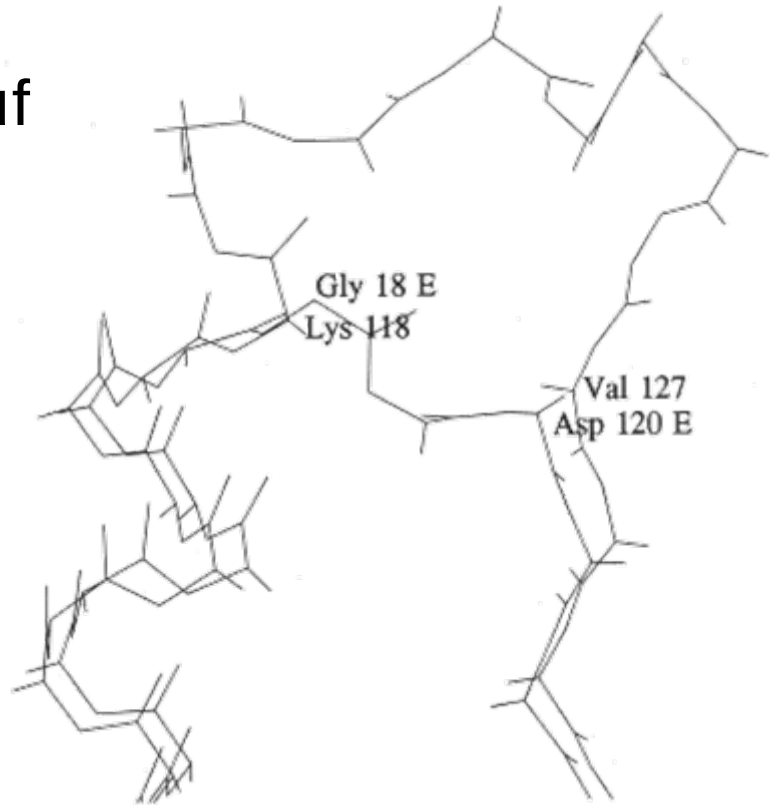


Fechteler, Dengler, Schomburg, J. Mol. Biol. (1995), 253, 114-131

DB-basierte Methoden – BRAGI

Beispiel

Einfügen von sieben zusätzlichen AS in Proteinase K (2PRK) wenn auf Subtilisin Carlsberg (2CSE) modelliert



...ASDKNNRNCPKGVASLS...
...TTNGM-------VINMS...

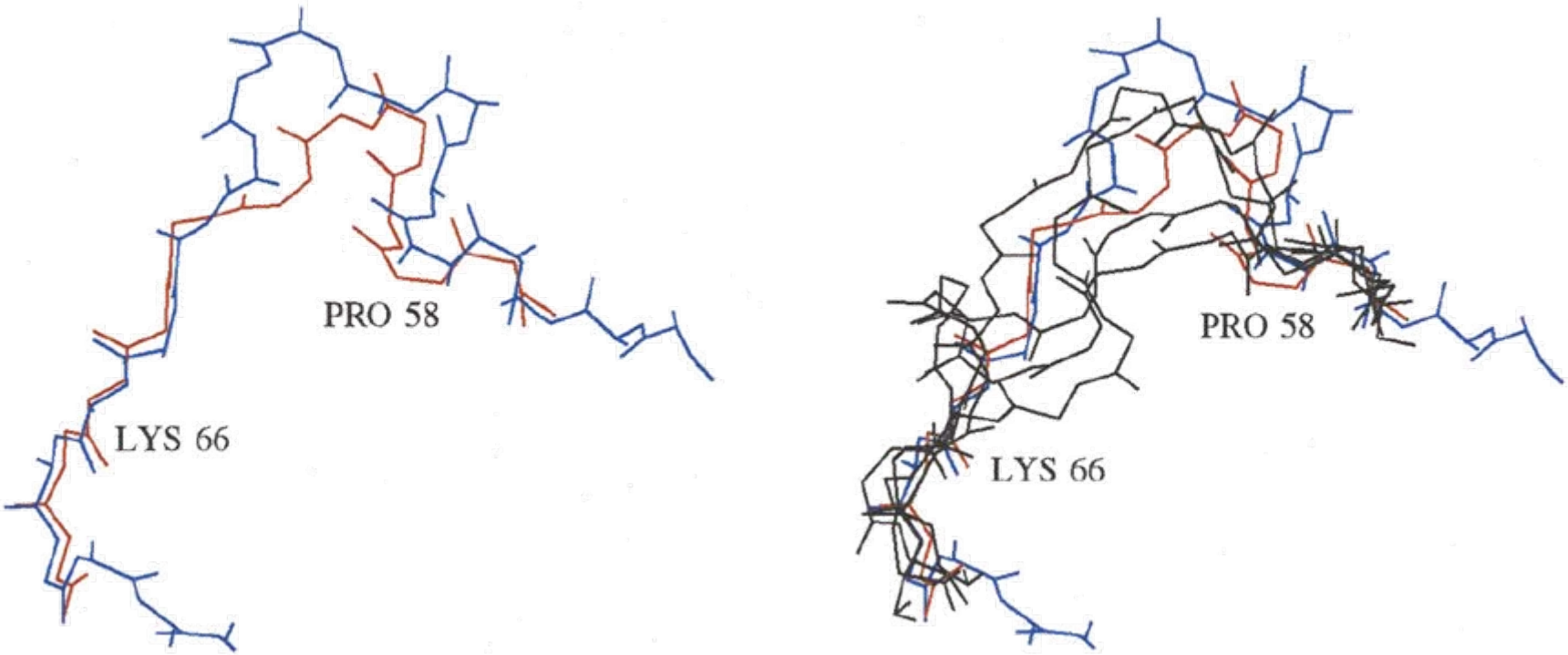
DB-basierte Methoden

Beispiel

Entfernung zweier AS aus Endothiapepsin (4APE) mit BRAGI

Links: native Struktur (blau), bestes Fragment (rot)

Rechts: zusätzlich die nächstbesten drei Fragmente



Fechteler, Dengler, Schomburg, J. Mol. Biol. (1995), 253, 114-131

Ab-initio-Konstruktion

- Moult & James haben einen der ersten Algorithmen zur **ab-initio-Konstruktion** vorgestellt
- Konstruktion **ohne Datenbank**
- Systematische Suche
  - Sinnvoll bis zu Längen von 6 AS
  - Suche eingeschränkt durch Betrachtung sinnvoller Backbone-Torsions-Paare
  - Konstruktion parallel von beiden Enden
  - Abbruch, falls verbleibende AS die Lücke nicht mehr füllen können
  - Überlappungen in den Seitenketten führen zum Verwerfen der Konformation

Moult, James, Proteins (1986), 1, 146

Ab-initio-Konstruktion

CONGEN

- Verwendet Raster von Backbone-Winkeln (15° oder 30°) in zugänglichen Bereichen des Ramachandran-Plots
- Sonderbehandlung für Gly, Pro
- Energien werden mit CHARMM berechnet
- Seitenketten werden parallel genauso (30°-Inkremente) konstruiert

Bruccoleri, Karplus, Biopolymers (1987), 26, 137

Gliederung

- Übersicht, Begriffe
- Schleifenmodellierung
  - Datenbanken
  - Algorithmen
- Seitenkettenmodellierung
  - Rotamerbibliotheken
  - Algorithmen
- Optimierung und Verifikation
- Programmpakete
  - SWISS-MODEL
  - MODELLER

Seitenkettenmodellierung

- Seitenketten-Modellierung: Optimierungsproblem
- **Modell**
  - Torsionswinkelraum der Seitenketten (SK)
  - Alle SK-Atome oder -Schweratome
- **Gegeben**
  - Backbone-Koordinaten (fix)
  - Startkoordinaten der Seitenkettenatome
- **Gesucht**

  Anordnung der Seitenketten minimaler Energie

  (**GMEC** = *global minimum energy conformation*)

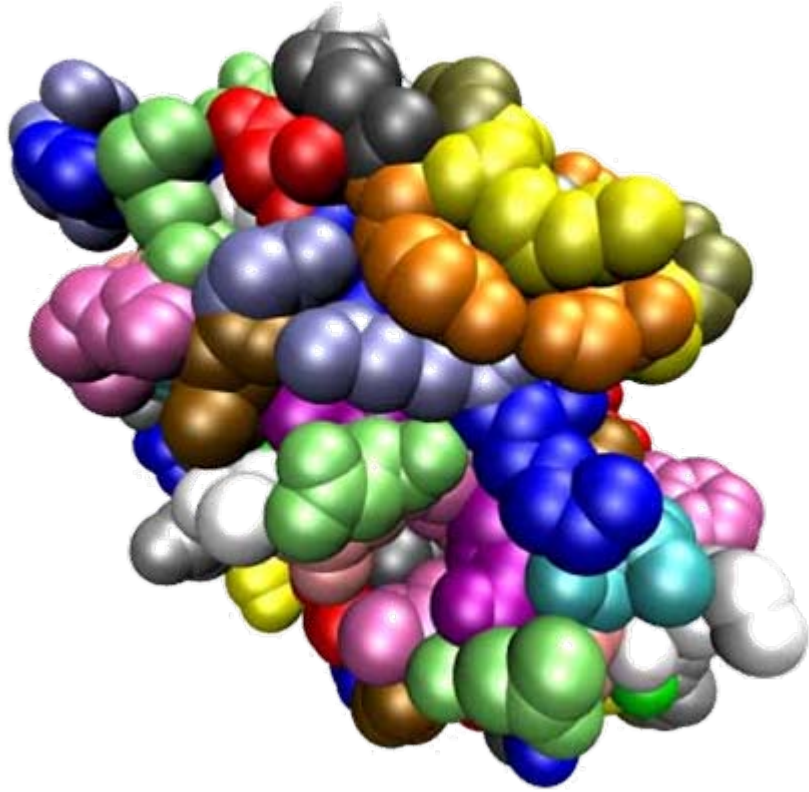
Seitenkettenmodellierung

Optimierungsproblem nichttrivial, da

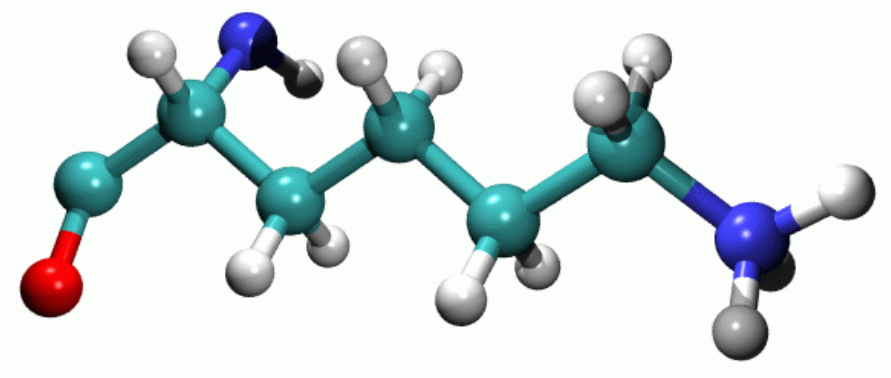
- alle Seitenketten miteinander wechselwirken
- Seitenketten sind eng gepackt

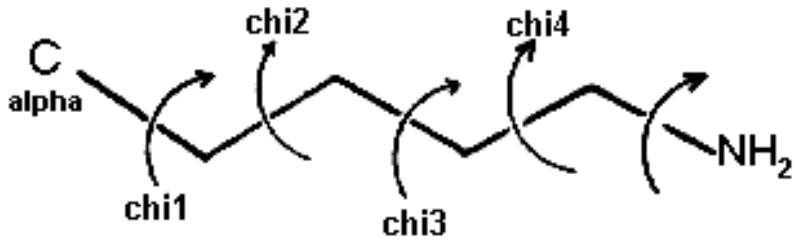
  => Überlappungen unzulässig

  => Verdrehen einer SK verschiebt benachbarte Seitenketten mit



SK-Flexibilität



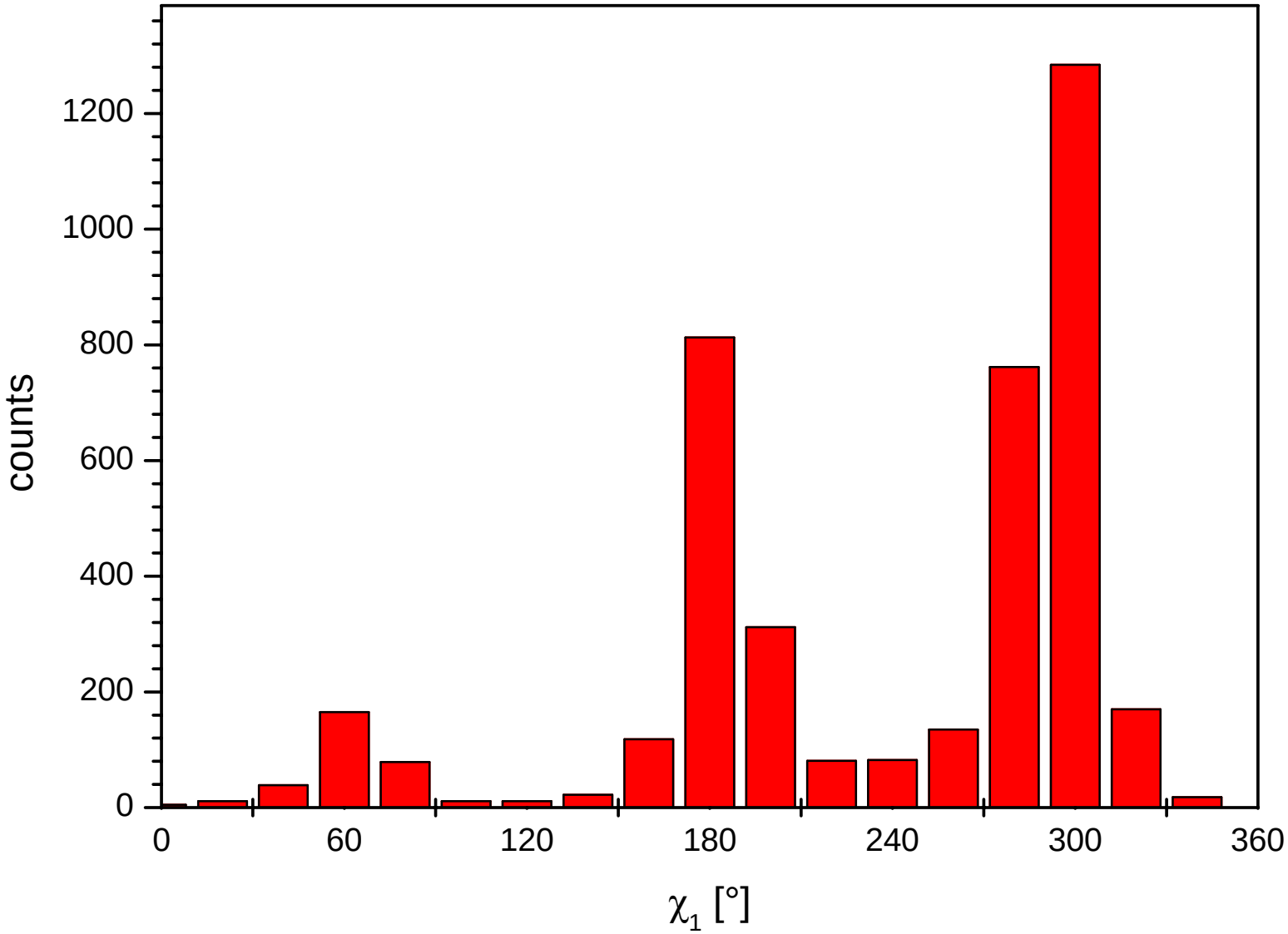


- Reduktion auf Torsionswinkelraum
- SK-Torsionswinkel: χ_1, χ_2, ...
- Seitenketten ohne Torsionswinkel:

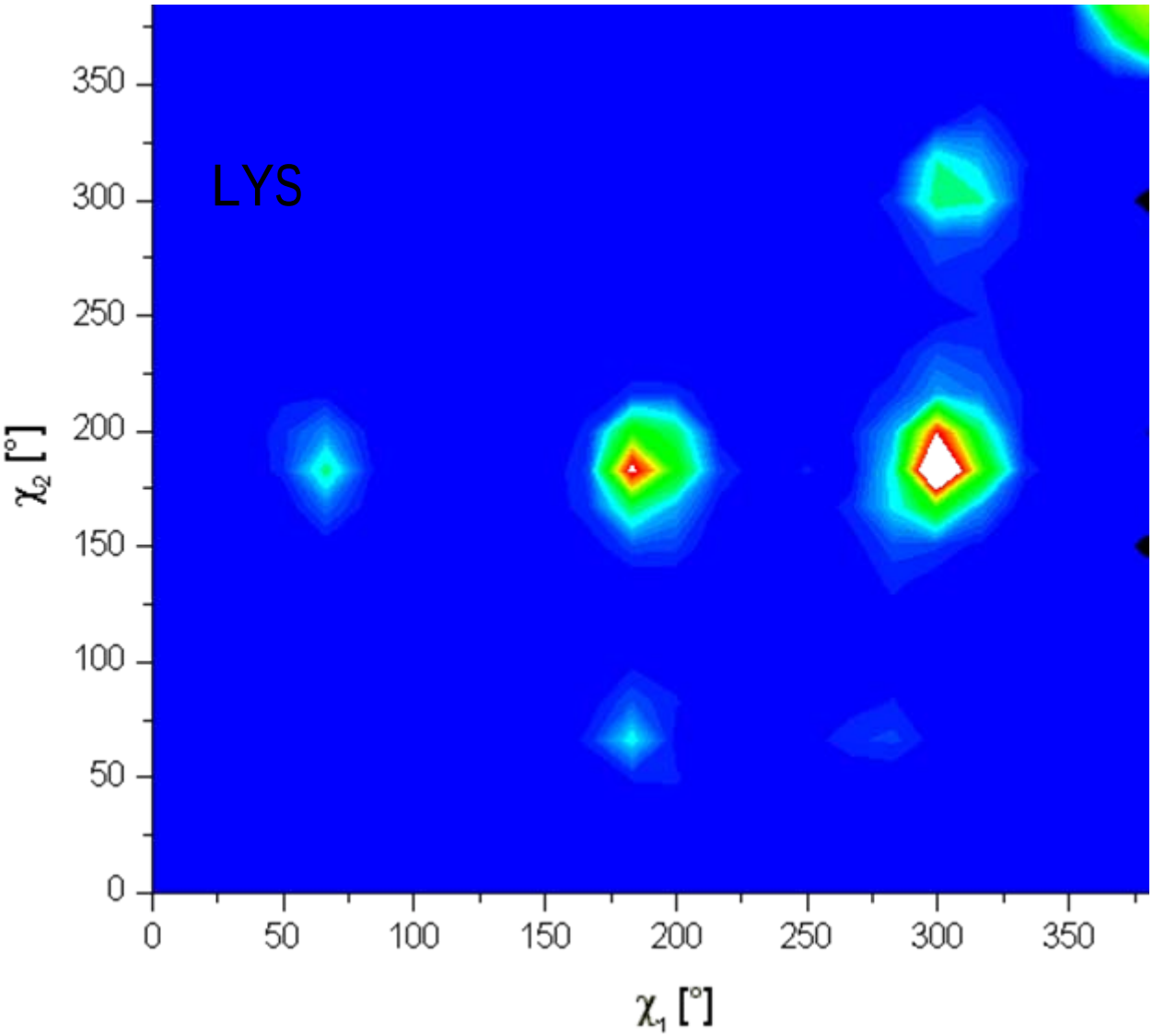
  Gly (Ala, falls nur Schweratome betrachtet werden)
- Eingeschränkte Torsionen:

  Pro (zwei Verkippungen des Rings)

Torsionswinkelverteilung – LYS

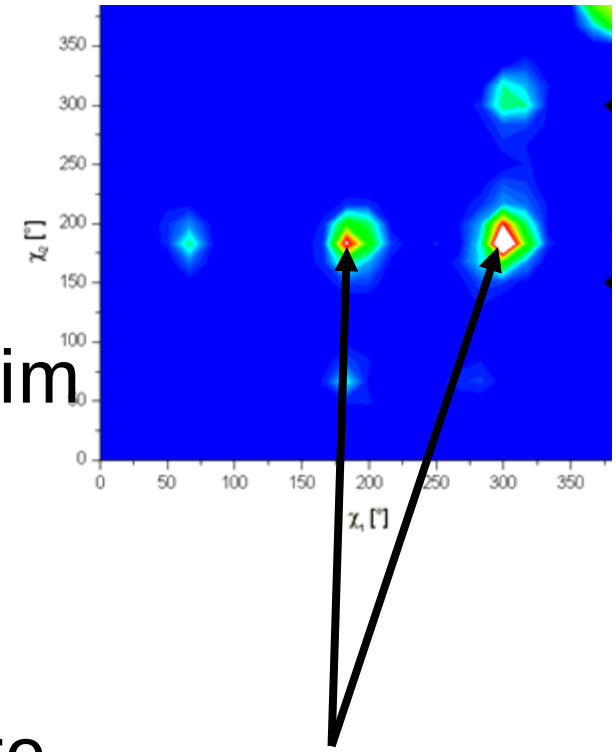


Torsionswinkelverteilung – LYS



Rotamere

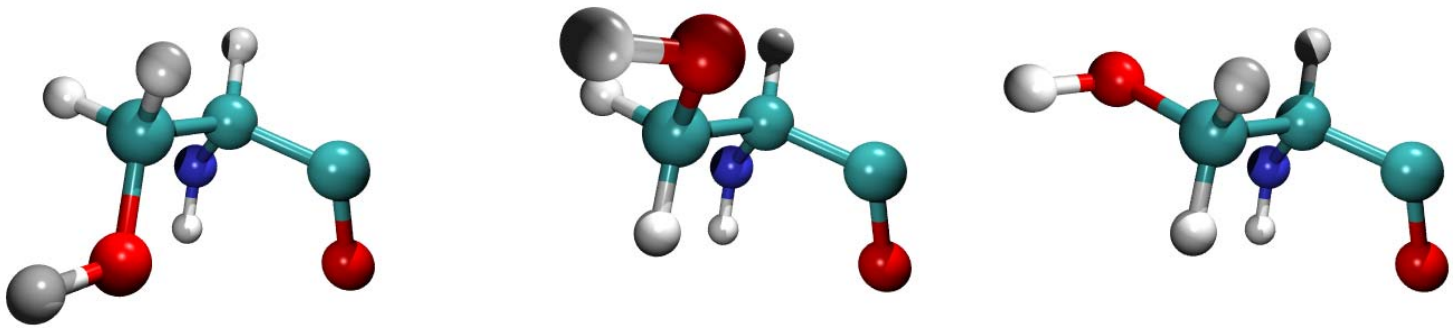
- **Torsionswinkel** werden **nicht unabhängig voneinander** angenommen
- Es existieren **ausgezeichnete Winkelbereiche** die Konformeren im Torsionsraum entsprechen
- Da diese **Konformere** durch Rotation um Torsionswinkel entstehen, nennt man sie Rotamere
- **Rotamere:** Seitenketten-konformationen minimaler Energie (**Rota**tionskonfor**mere**)



Rotamere von LYS

Rotamerbibliotheken

- Bereits um 1970 wurde vermutet, dass die Seitenketten der AS nur einige wenige günstige Konformationen (Rotamere) annehmen.
- Seitdem stellten viele Autoren Rotamerbibliotheken aus der PDB zusammen, die diese tabellieren.



SER

Sasisekharan, Ponnuswamy, Biopolymers (1970), 9, 1249
Ponder, Richards, J. Mol. Biol. (1987), 193, 775

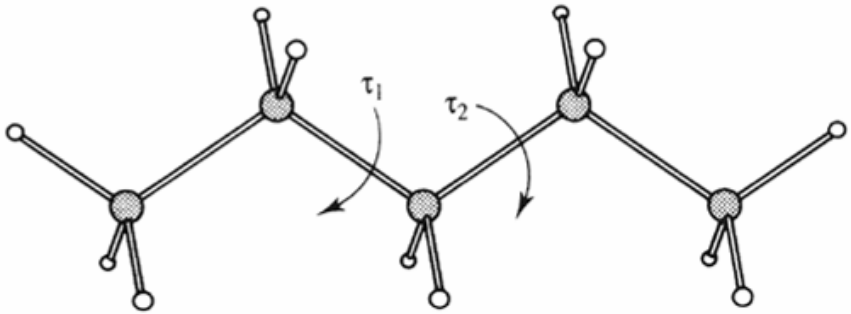
Dunbrack-Rotamerbibliothek

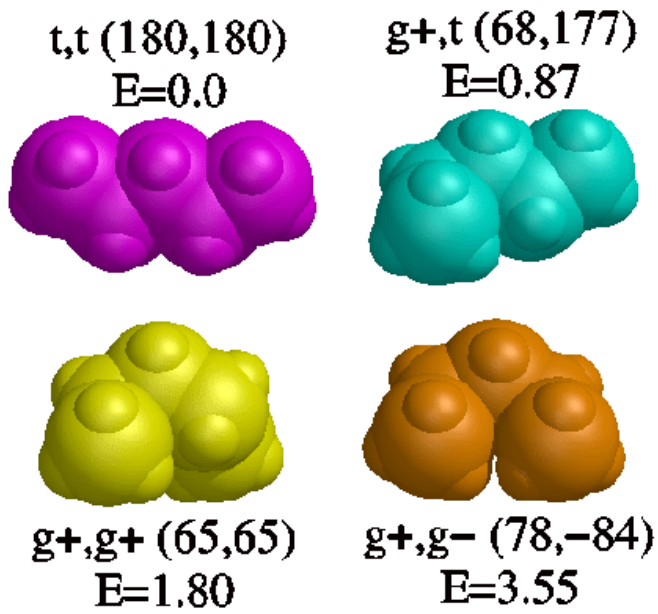
- Eine der wichtigsten und besten Rotamerbibliotheken wird von Roland Dunbracks Gruppe gepflegt
- Zwei Varianten
  - **Backbone-independent**

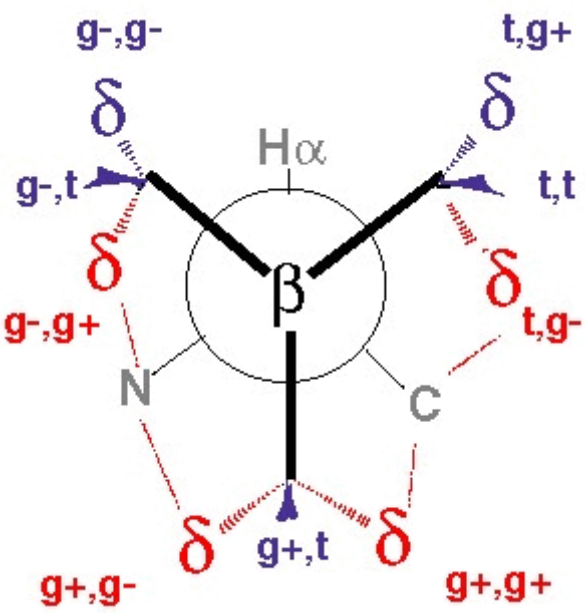
    Rotamer sind unabhängig von Backbone-Torsionswinkeln ϕ, ψ
  - **Backbone-dependent**

    Zu jedem Paar von Werten (ϕ, ψ) werden die Rotamere getrennt tabelliert
- Für jedes Rotamer werden angegeben
  - Häufigkeiten $n(\chi_1,\ldots)$ für Rotamer in DB
  - Torsionswinkel χ_1, ...
  - (bedingte) Wahrscheinlichkeiten für das Auftreten des Rotamers ($P(\chi_1, \chi_2)$, $P(\chi_2|\chi_1)$, ...)
  - Standardabweichungen σ

Torsionswinkel – Nomenklatur



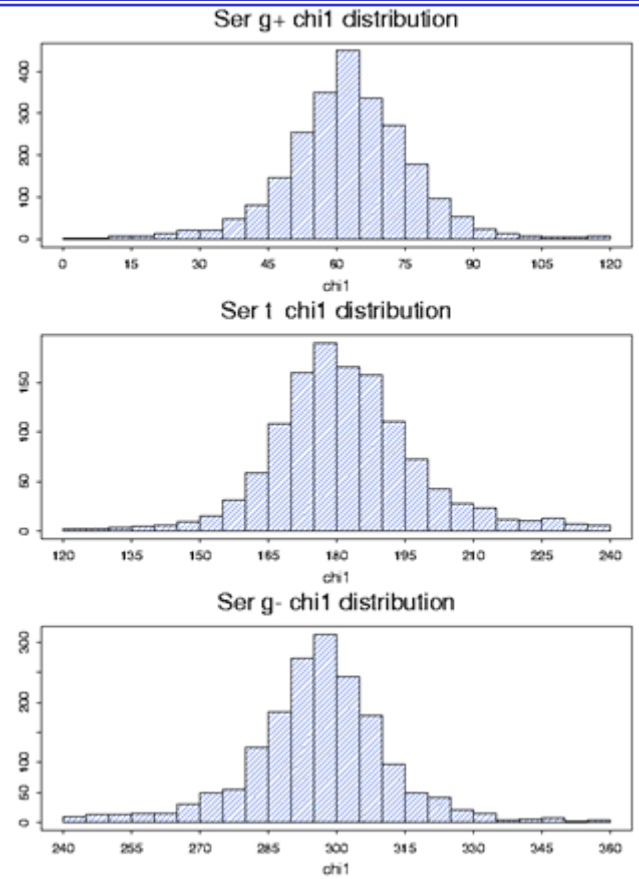


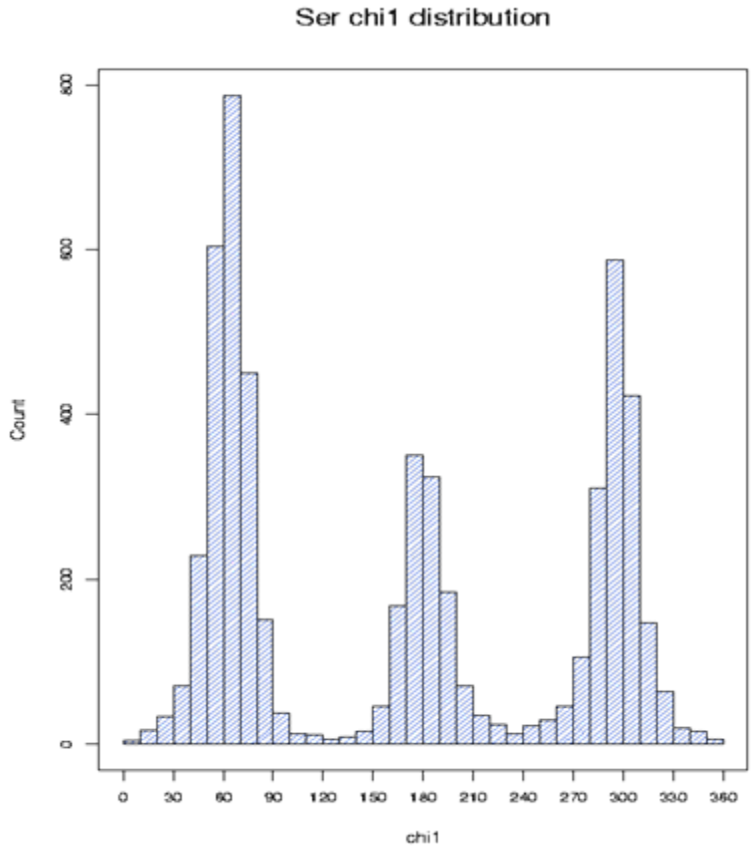


Nomenklatur:

t	trans	$120° \cdot \chi_1 < 240°$
g-	gauche (-)	$-120° \cdot \chi_1 < 0$
g+	gauche (+)	$0° \cdot \chi_1 < 120°$

Dunbrack-Rotamerbibliothek (bbind)

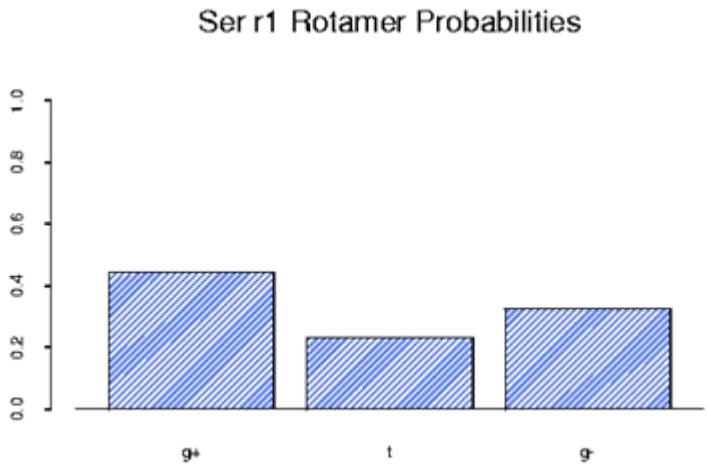


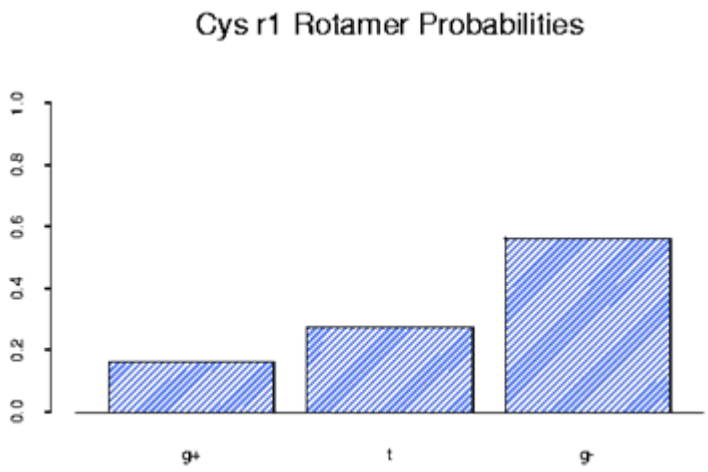


- Basiert auf Analyse von Häufigkeiten der einzelnen Winkel
- Betrachtet wird nichtredundante Teilmenge der PDB

Dunbrack-Rotamerbibliothek (bbind)

- Es werden nur Seitenketten betrachtet (*backbone independent*)
- Es werden nur Schweratome betrachtet
- Torsionswinkel werden zu **Clustern** zusammengefasst, die durch ein Rotamer repräsentiert werden
- **Bibliothek** enthält für jedes Rotamer
  - Torsionswinkel
  - Häufigkeiten in DB
  - a-priori-Wahrscheinlichkeiten
  - bedingte Wahrscheinlichkeiten (Annahme eines fixen χ_1)

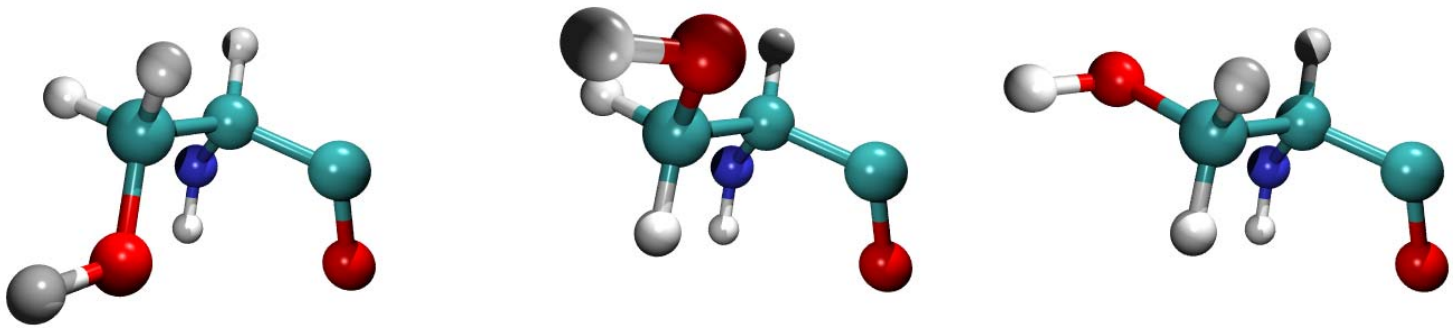




Dunbrack-Rotamerbibliothek (bbind)

$n(\chi_1)$ $n(\chi_1,\chi_2)$ $P(\chi_1,\chi_2)$ σ $P(\chi_2|\chi_1)$ σ χ_1 σ

```
...
SER 1 0 0 0    4931    4931    47.07    0.40  100.00    0.00    65.3 10.5
SER 2 0 0 0    2466    2466    23.54    0.34  100.00    0.00   179.0 11.6
SER 3 0 0 0    3078    3078    29.38    0.36  100.00    0.00   -63.9 10.5
...
```



SER

Dunbrack-Rotamerbibliothek (bbind)

AS	# Rotamere	AS	# Rotamere
Ala	0	Leu	9
Arg	81	Lys	81
Asn	18	Met	27
Asp	9	Phe	6
Cys	3 / 0	Pro	2
Gln	36	Ser	3
Glu	27	Thr	3
Gly	0	Trp	9
His	9	Tyr	6
Ile	9	Val	3

Gly, Ala Keine Torsionen (nur Schweratome!)

Pro: zwei Rotamere – verkippte Ringe

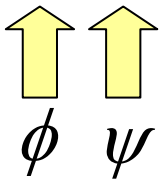
Cys: keine Rotamere wenn in Schwefelbrücke

Dunbrack-Rotamerbibliothek (bbdep)

- Seitenketten- und Backbone-Atome kollidieren für manche Rotamere

  => die möglichen Rotamere hängen auch von ϕ/ψ ab
- Die Backbone-abhängige Bibliothek (bbdep) tabelliert die Rotamere für jede ϕ/ψ-Kombination (in 10°-Schritten)

```
...
SER  -180 -180    54    1 0 0 0  0.854351    69.6     0.0     0.0     0.0     12.3     0.0     0.0     0.0
SER  -180 -180    54    2 0 0 0  0.144673  -166.0     0.0     0.0     0.0     18.2     0.0     0.0     0.0
SER  -180 -180    54    3 0 0 0  0.000976   -63.9     0.0     0.0     0.0      9.6     0.0     0.0     0.0
SER  -180 -170    22    1 0 0 0  0.756981    70.6     0.0     0.0     0.0     15.2     0.0     0.0     0.0
SER  -180 -170    22    2 0 0 0  0.241662  -169.7     0.0     0.0     0.0     12.5     0.0     0.0     0.0
SER  -180 -170    22    3 0 0 0  0.001357   -63.8     0.0     0.0     0.0     10.3     0.0     0.0     0.0
SER  -180 -160     2    1 0 0 0  0.664463    65.6     0.0     0.0     0.0     11.2     0.0     0.0     0.0
...
```



Größe des Suchraums

- Dunbracks Rotamerbibliothek enthält bis zu 81 Rotamere pro AS
- Bereits für ~50 AS gibt es 10^{55}-10^{60} mögliche Rotamer-Kombinationen
- Für große Sequenzbereiche, ganze Proteine **kaum mehr systematisch durchsuchbar**
- Interaktionen der Seitenketten untereinander führen dazu, dass das Finden des globalen Maximums erschwert wird

Komplexität

- Anzahl der Konformationen ist klar exponentiell:

  $(\#\ \text{Rotamere/AS})^{\#\text{AS}}$

- Zur energetischen Bewertung werden paarweise Potentiale verwendet
- Analog zum Faltungsproblem:

  SK-Optimierung mit Rotameren und paarweisen Potentialen ist NP-hart

- Zusätzlich kann man zeigen, dass das Problem **nicht approximierbar** ist

N. A. Pierce and E. Winfree, *Prot. Eng.*, 15(10):779–782, 2002.
B. Chazelle, C. Kingsford, M. Singh, Proc. ACM FCRC'2003, Paris Kanellakis Memorial Workshop (2003), 86-94.

Algorithmen zur SK-Optimierung

- **Monte-Carlo-Ansätze**
  - Holm, Sander, Proteins (1992), 14, 213
  - Levitt, J. Mol. Biol. (1992), 226, 507
- **A*-Algorithmus**
  - Leach, Lemon, *Proteins (1998)*, 33, 227
- **Branch&Bound und Verwandte**
  - **Desmet, De Maeyer, Hazes, Lasters, *Nature* (1992), 356, 539**
  - Bower, Cohen, Dunbrack, *J. Mol. Biol.* (1997), 267, 1268
- **Ganzzahlige Lineare Programmierung (ILP)**
  - Althaus, Kohlbacher, Lenhof, Müller, *J. Comput. Biol.* (2002), 9, 597
- **Semidefinite Programmierung**
  - Chazelle, Kingsford, Singh, *Proc. ACM FCRC* 2003, 86

Energiefunktionen

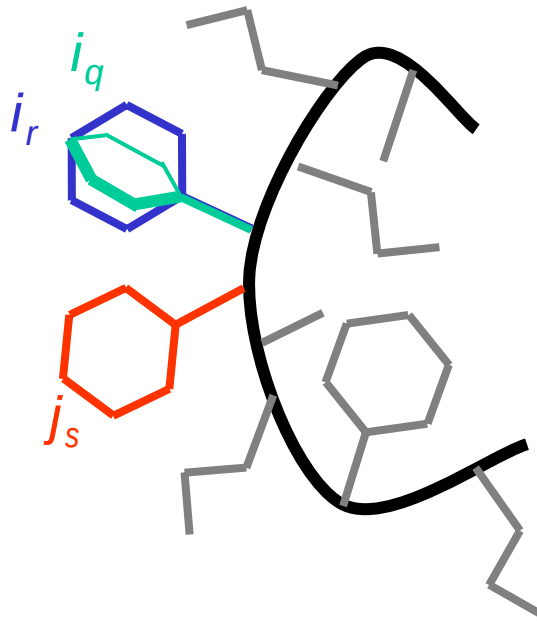
- Energiefunktion muss **atomare Auflösung** haben
- Größe des Suchraums

  => keine aufwändigen Energiefunktionen möglich
- Häufig verwendet
  - molekülmechanische Kraftfelder

    (AMBER, CHARMM)
  - Sonstige Kombinationen aus Torsionen, Elektrostatik, vdW
- **Energien** immer **zerlegbar** in
  - Interne Energie der Seitenkette (Torsionen)
  - Paarweise WW der SK untereinander und mit dem Backbone

Energiefunktionen

- Energie des Backbones: E^{bb}
- Energie für Rotamer *r* der Seitenkette *i*: $E^{bb}_{i_r}$
  - Interne Energie des Rotamers (Torsionen)
  - WW des Rotamers mit dem Backbone
- Paarweise WW der Rotamere *r*, *s* für Seitenketten *i*, *j* miteinander: $E^{pw}_{i_r,j_s}$
- **Gesamtenergie**



$$E_{ges} = \underbrace{E^{bb}}_{\text{const.}} + \underbrace{\sum_i E^{bb}_{i_r}}_{\text{const. für Rotamer}} + \underbrace{\sum_i \sum_{j<i} E^{pw}_{i_r,j_s}}_{\text{Paarweise WW}}$$

Dead End Elimination

- Desmet *et al.* schlagen einen Branch&Bound-ähnlichen Algorithmus vor, den sie **Dead End Elimination (DEE)** nennen
- Dazu werden alle Rotamerenergien und alle Rotamer-Rotamer-WW berechnet
- Das **DEE-Theorem** besagt nun:

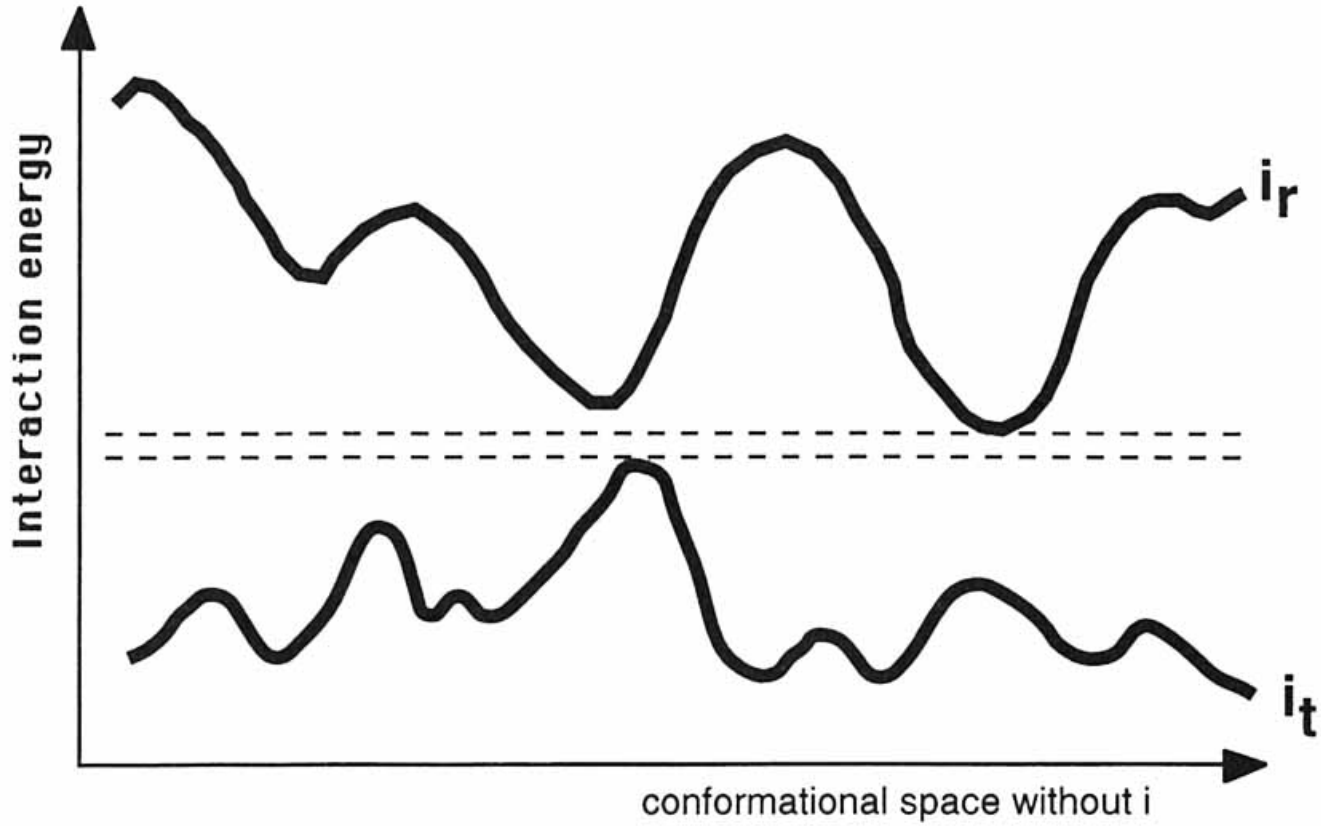
  Falls für zwei Rotamere i_r, j_s gilt

$$E_{i_r} + \sum_j \min_s E_{i_r,j_s} > E_{i_t} + \sum_j \max_s E_{i_t,j_s}$$

  dann ist i_r nicht Teil der optimalen Lösung.

  i_r ist also ein „totes Ende“, wenn die beste Energie (min) für i_r immer schlechter ist als die schlechteste Energie (max) für ein anderes Rotamer i_t.

Dead End Elimination



$$E_{i_r} + \sum_j \min_s E_{i_r,j_s} > E_{i_t} + \sum_j \max_s E_{i_t,j_s}$$

De Maeyer et al., In: Protein Structure Prediction (D. Webster, Hrsg.), Humana Press (2000), S. 296

Dead End Elimination

- Das **DEE-Theorem** definiert also eine **untere Schranke** für die Qualität der Lösung die i_r enthält
- Im DEE-Algorithmus werden iterativ Rotamere entfernt, die das DEE-Theorem verletzen

Algorithmus

- Berechne Wechselwirkungsenergien
- Für alle Seitenketten i:
  - Für alle Rotamer-Paare (i_r, i_t):
    - Prüfe DEE-Kriterium
    - Entferne i_r, wenn erfüllt
- Abbruch, wenn kein Paar mehr gefunden

Dead End Elimination

- DEE bricht ab, wenn keine Paare mehr gefunden werden
- In der Regel bleibt dann ein großer Restsuchraum übrig

  (z.B. für 50 AS wird der Suchraum oft von 10^{50} – 10^{60} auf 10^{30} – 10^{40} reduziert)
- Es existieren weitere untere Schranken, z.B. von Goldstein oder DEE für Rotamerpaare, die eine weitere Einschränkung des Suchraums ermöglichen
- Der verbliebene Suchraum kann schließlich durch Enumeration auf das Minimum geprüft werden

Dead End Elimination

- **Goldsteins Kriterium** besagt:

  Rotamer i_r kann entfernt werden, wenn gilt:

$$E_{i_r} - E_{i_s} + \sum_{i \neq j} \min_{j_s} (E_{i_r,j_s} - E_{i_t,j_s}) > 0$$

  d.h., i_r kann eliminiert werden, falls ein anderes Rotamer i_s von i stets eine günstigere WW-Energie mit allen anderen Seitenketten besitzt als i_r.

- Dieses Kriterium ist stärker als das ursprüngliche DEE Theorem.
- Weitere Varianten für Paare von Rotameren existieren, die eine weitere Reduktion ermöglichen.

Goldstein, Biophys J. (1994), 66, 1335-40

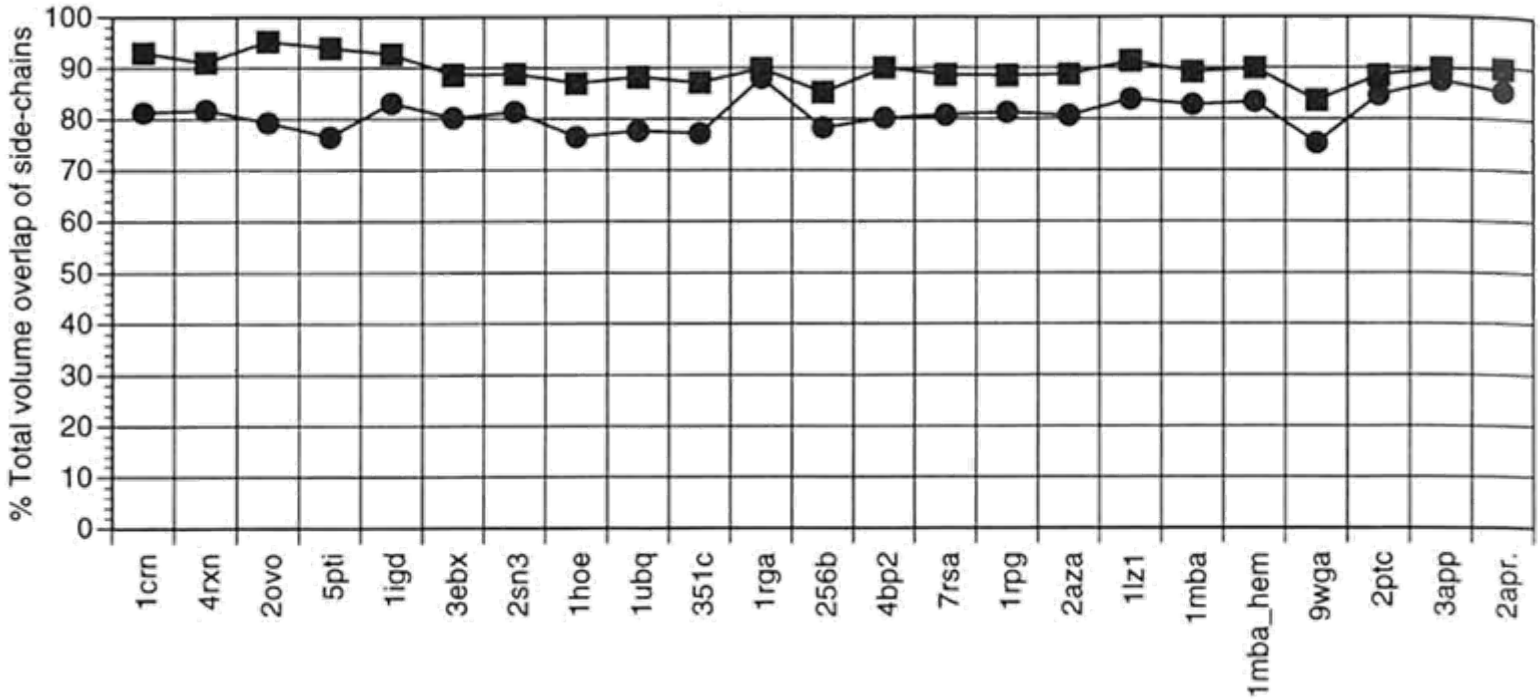
Qualität der Resultate

- Abhängig von Suchalgorithmus und von Qualität der Rotamerbibliothek
- Maß für Qualität:
  - Prozentsatz korrekte χ_1-Zuordnungen
  - Prozentsatz korrekte χ_1+χ_2-Zuordnungen
- χ_1-Zuordnung generell einfacher
- Äußere Torsionswinkel häufig nicht wohl definiert
  - Geladene AS haben viele Torsionswinkel (Lys, Arg!)
  - Geladene Reste sitzen üblicherweise an der Oberfläche

    => nicht dicht gepackt
  - Geladene Reste ragen ins Wasser und sind meist in der Lage das geladene Ende im Wasser zu drehen

Qualität der Resultate

- Erreichbare Werte für Seitenketten im Proteinkern
  - χ_1: ~90% korrekt
  - χ_{12}: ~80% korrekt



De Maeyer et al., In: Protein Structure Prediction (D. Webster, Hrsg.), Humana Press (2000), S. 296

Gliederung

- Übersicht, Begriffe
- Schleifenmodellierung
  - Datenbanken
  - Algorithmen
- Seitenkettenmodellierung
  - Rotamerbibliotheken
  - Algorithmen
- Optimierung und Verifikation
- Programmpakete
  - SWISS-MODEL
  - MODELLER

Optimierung

- Platzierung der Seitenketten nur **bezüglich der Diskretisierung** auf Rotamere **optimal**!
- Abweichungen von den idealen Rotamerwinkeln sind die Regel

  => Entspannen der Struktur
- Führt nicht *per se* zu genereller Verbesserung des Modells, sondern nur zur **Auflösung lokaler Probleme** in der Struktur
- In der Regel durch
  - MD-Simulation
  - Energieminimierung
- Kraftfelder sind üblicherweise AMBER oder CHARMM

Homologiemodellierung: Fehlerquellen

- Wahl der falschen Schablone
- Inkorrektes Alignment

} Threading-Probleme

- Fehler in Regionen ohne Schablone
- Verzerrungen in korrekt alinierten Regionen
- Fehler in der Seitenkettenplatzierung

Sánchez, Šali, In: Protein Structure Prediction, (Webster, Hrsg.), Humana Press (2000), 121

Validierung

Um die Qualität der Struktur abzuschätzen, gibt es Kontrollmöglichkeiten

- **Experimentelle Daten**

  Idealerweise besitzt man experimentelle Daten, die zwar nicht für eine vollständige Strukturaufklärung ausreichen, aber genügen um einzelne Aspekte des Modells zu verifizieren (z.B. NOEs)

- **Güte der Struktur**

  Gemessen an Packung, Einhaltung der Standardgeometrien (Winkel, Bindungslängen), niedrige Energien

Validierung

Güte der Struktur wird beurteilt nach

- **Geometrie**
  - Überlapp, unübliche Winkel, Abstände
  - Unübliche ϕ/ψ-Kombinationen im Ramachandran-Plot
- **Energie** gemäß Paarpotentialen (z.B. Sippl)

- Eine gewisse Anzahl Abweichungen sind üblich (auch für exp. bestimmte Strukturen), treten die Fehler aber gehäuft an bestimmten Stellen der Sequenz auf, ist dies ein Indiz für einen Fehler in der Homologiemodellierung.

WHAT_CHECK

- Teil des Programmpakets **WHAT IF** von Gert Vriend
- Entwickelt zur Überprüfung von Proteinstrukturen für Kristallographie und NMR
- WHAT_CHECK testet anhand von Standardwerten (aus der PDB), ob die Geometrie ungewöhnliche Abweichungen aufweist bezüglich
  - Bindungswinkel
  - Bindungslängen
  - Torsionswinkel
  - H-Brücken-Donoren und –Akzeptoren im Kern ohne Partner
  - Überlappungen von Atomen
  - Ungewöhnliche Packung von Resten (zu viele/wenige Kontakte)
  - …
- Viele dieser Abweichungen werden erst in einer Minimierung/MDS eingeführt, sind aber Indiz dafür, dass die Struktur inkorrekt ist (Struktur weicht dem Problem lokal durch Deformation aus)

Hooft, Vriend, Sander, Abola, Nature (1996), 381, 272

Gliederung

- Übersicht, Begriffe
- Schleifenmodellierung
  - Datenbanken
  - Algorithmen
- Seitenkettenmodellierung
  - Rotamerbibliotheken
  - Algorithmen
- Optimierung und Verifikation
- Programmpakete
  - SWISS-MODEL
  - MODELLER

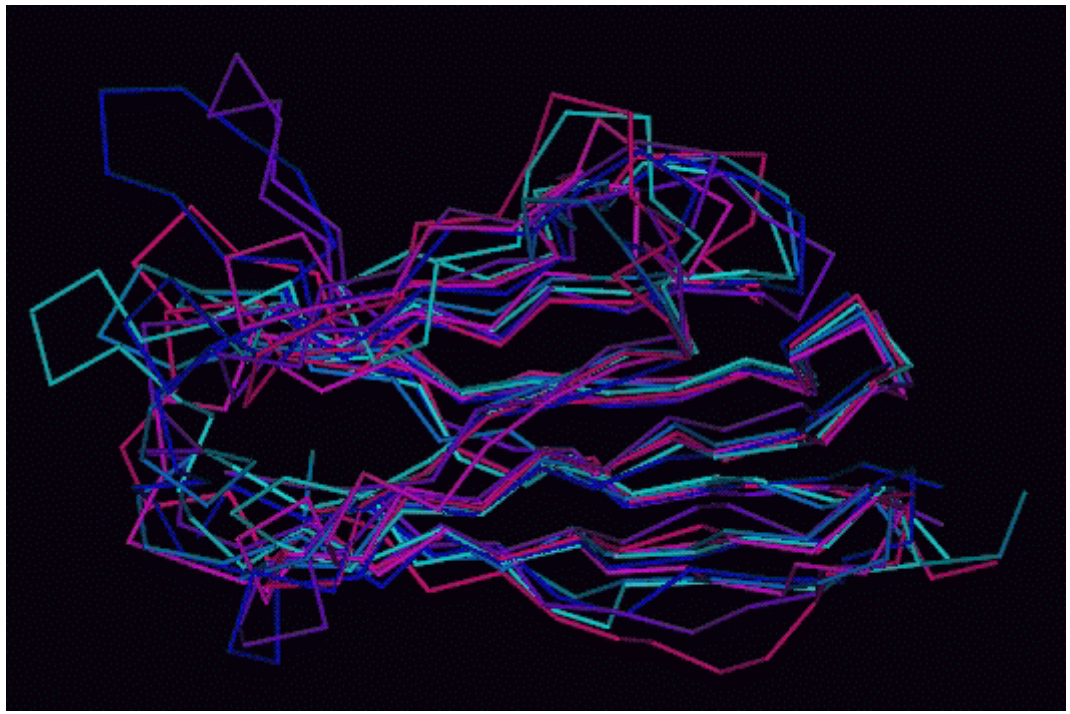
SWISS-MODEL

- Web-Dienst: **`http://www.expasy.org/swissmod/SWISS-MODEL.html`**
- Standard-Vorgehensweise
    1. Identifizierung der Schablone(n)
    2. Alignment
    3. Konstruktion des Kern-Backbones
    4. Loop-Modellierung
    5. Seitenkettenmodellierung
    6. Energieminimierung
- Kern bildet **ProMod II**, das eine Sequenz auf einen Satz von Strukturen modelliert (2-6)

ProModII

1. Schritt von ProModII:

Überlagerung aller Strukturen basierend auf Alignment der Sequenzen und 3D-Überlagerung alinierter Regionen der Strukturen

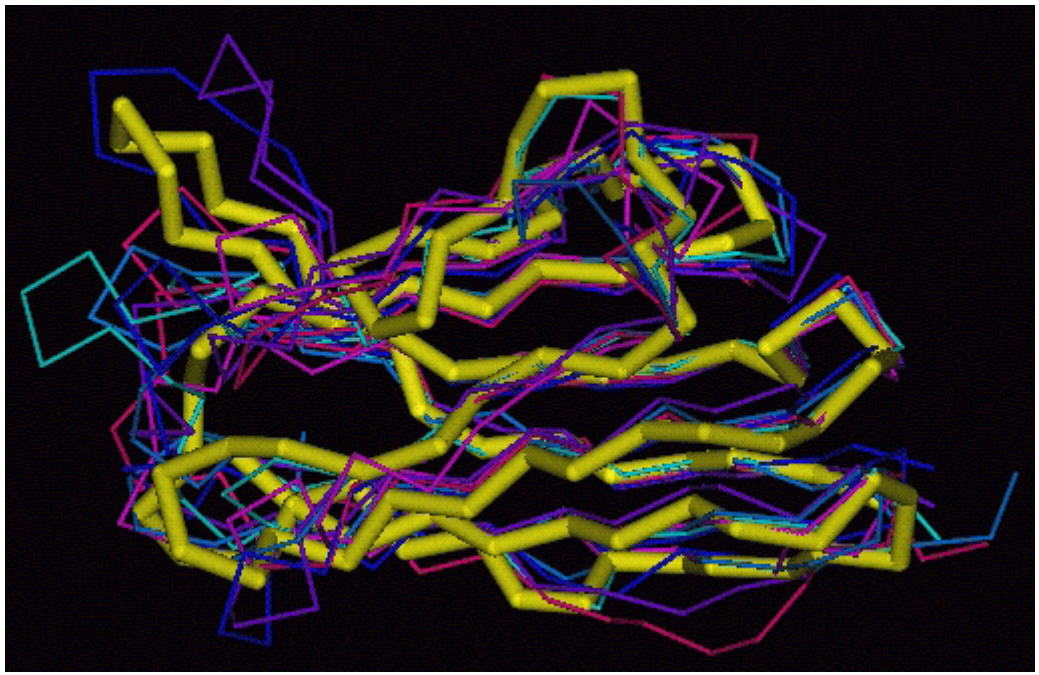


Guex, Peitsch, Electrophoresis (1997), 18, 2714-2723

ProModII

2. Schritt:

Berechnung von gemittelten Backbone-Koordinaten ausgehend von Sequenzähnlichkeit im multiplen Alignment

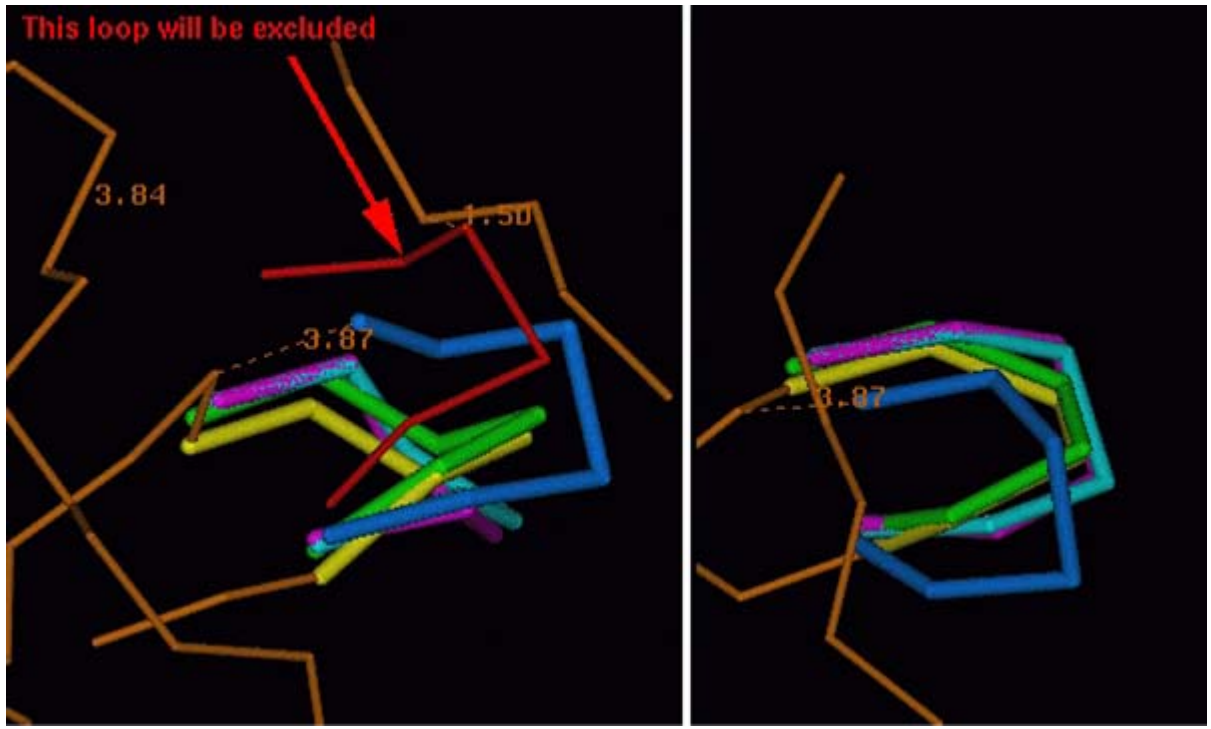


Guex, Peitsch, Electrophoresis (1997), 18, 2714-2723

ProModII

3. Schritt:

Konstruktion der nicht konservierten Loops mit einer Kombination aus DB-Suche und ab-initio-Konstruktion.



Guex, Peitsch, Electrophoresis (1997), 18, 2714-2723

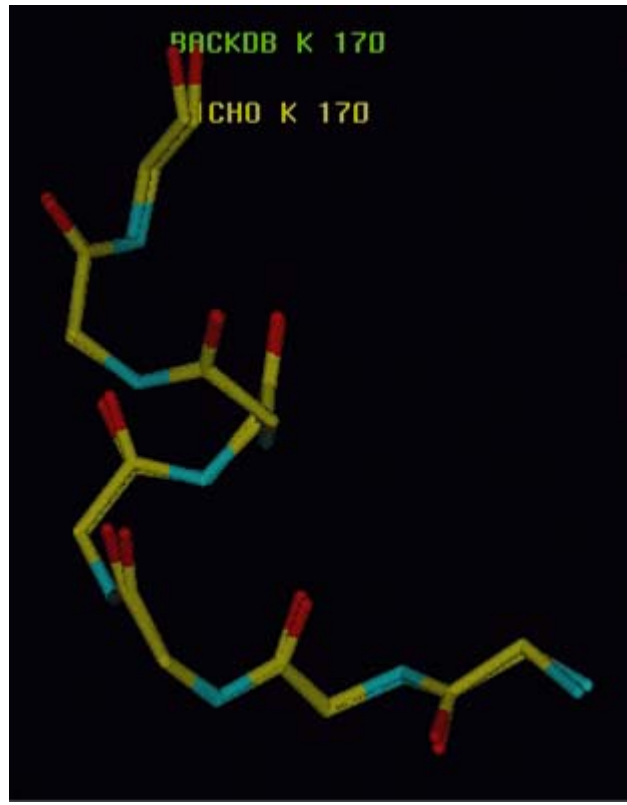
ProModII

4. Schritt:

Vervollständigung des Backbones aus den C_α-Koordinaten.

Koordinaten werden dabei aus Strukturfragmenten konstruiert:

- Suche Fragment der Länge 5, dessen Koordinaten (C_α) am besten auf den jeweiligen Backbone-Abschnitt passen
- Übernehme Koordinaten der übrigen Backbone-Atome (C, N, O) der mittleren drei AS



Guex, Peitsch, Electrophoresis (1997), 18, 2714-2723

ProModII

5. Schritt:

Platzierung der Seitenketten (probabilistisches Modell, Backbone-abhängige Rotamerbibliothek)

6. Schritt:

Energieminimierung zur Entfernung lokaler Überlappungen (GROMOS96-Kraftfeld, Steilster Abstieg)

MODELLER – Loop-Konstruktion

- Ab-initio-Loop-Konstruktion inkl. Seitenketten
- **Modell**

  Alle Schweratome der Loop inkl. Seitenketten
- **Energiefunktion**

  Basierend auf CHARMM, ergänzt um statistische Präferenzen für bestimmte Torsionswinkel
- **Algorithmus**

  Randomisierte Suche kombiniert mit MD-Simulation und konj. Gradientenverfahren

Fiser, Do, Šali, Protein Sci. (2000), 9, 1753

MODELLER

$$E_{\text{CHARMM}} = \sum_{\text{bonds}} k_b(r_{ij} - r_0)^2 + \sum_{\text{angles}} k_\phi(\phi_{ijk} - \phi_0)^2 + \sum_{\text{tors}} |k_\psi| - b_\psi \cos(n\psi_{ijkl} + \delta) + \sum_{\text{improp}} k_\theta(\theta_{ijkl} - \theta_0)$$

- CHARMM22-Kraftfeld ohne ES, vdW
- Ergänzt um folgende Komponenten
  - Atombasiertes, interpoliertes Sippl-Potential
  - Harmonische Repulsion für Überlappungen
  - Wahrscheinlichkeitsbasierte Constraints für Torsionen

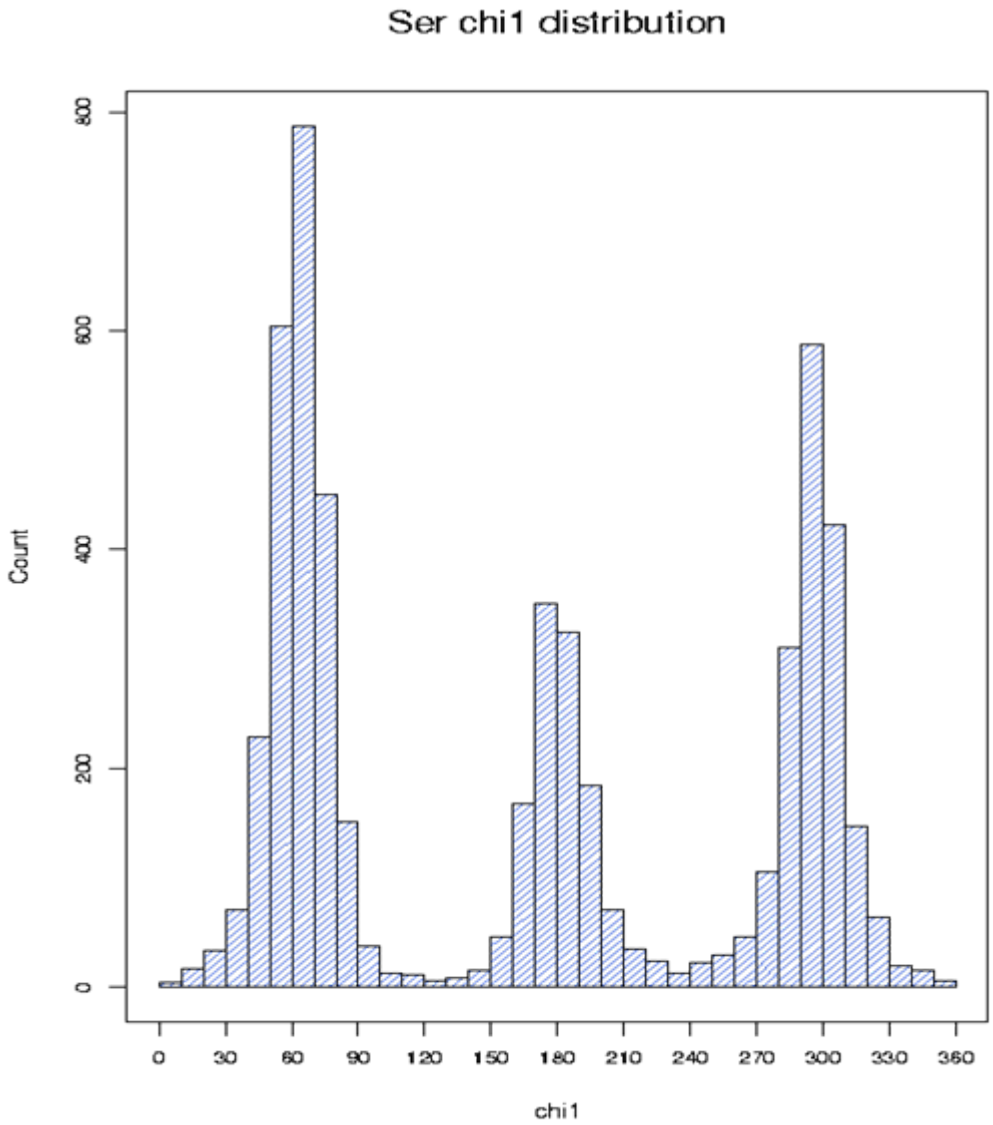
$$E = E_{\text{CHARMM}} - \sum_{\text{sc tors}} \ln p_s(\chi) - \sum_{\text{residues}} \ln p_\omega(\omega) - \sum_{\text{residues}} \ln p_{\phi\psi}(\phi, \psi) + \sum_{i,j \in \text{nb}} \xi[E_{\text{pw}}(i,j) + S(i,j)]$$

MODELLER

$$E = E_{\text{CHARMM}} - \sum_{\text{sc tors}} \ln p_s(\chi) - \sum_{\text{residues}} \ln p_\omega(\omega)$$
$$- \sum_{\text{residues}} \ln p_{\phi\psi}(\phi,\psi) + \sum_{i,j\in\text{nb}} \xi[E_{\text{pw}}(i,j) + S(i,j)]$$

- Umwandlung der Wahrscheinlichkeiten für die Annahme bestimmter Torsionswinkel in Pseudoenergien (durch Logarithmieren)
- $p_s(\chi)$ abhängig von
  - Typ der AS
  - Torsion innerhalb der AS (χ_1, χ_2,...)
- Ableitbar aus Strukturdatenbanken

MODELLER – Energiefunktion



MODELLER – Energiefunktion

$$E = E_{\text{CHARMM}} - \sum_{\text{sc tors}} \ln p_s(\chi) - \sum_{\text{residues}} \ln p_\omega(\omega) - \sum_{\text{residues}} \ln p_{\phi\psi}(\phi,\psi) + \sum_{i,j\in\text{nb}} \xi[E_{\text{pw}}(i,j) + S(i,j)]$$

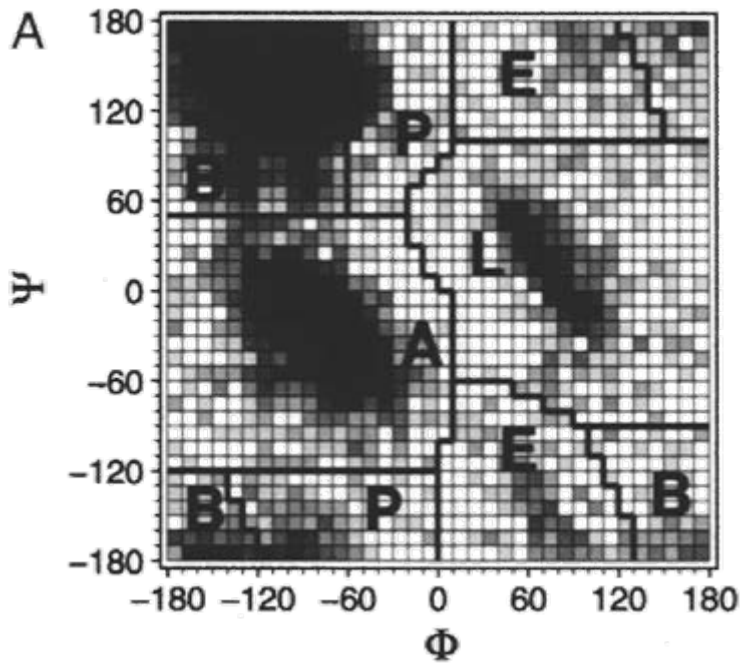
- Analog werden die Beiträge für die Torsionen der Peptidbindungen p_ω modelliert
- Es gibt hierbei nur eine Winkelklasse
- Sollwinkel liegt bei 180° bei einer Standardabweichung von 5°
- Komplexer sind die Beträge für ϕ und ψ, da hier im wesentlichen der Ramachandran-Plot für alle AS interpoliert werden muss

MODELLER – Energiefunktion

- Plot wird in fünf Klassen eingeteilt
- Gesamtwahrscheinlichkeit wird wieder als gewichtete Summe der Klassenwahrscheinlichkeiten dargestellt:

$$p_\phi\psi(\phi,\psi) = \sum_{i=1}^{5} \omega_i p_i$$

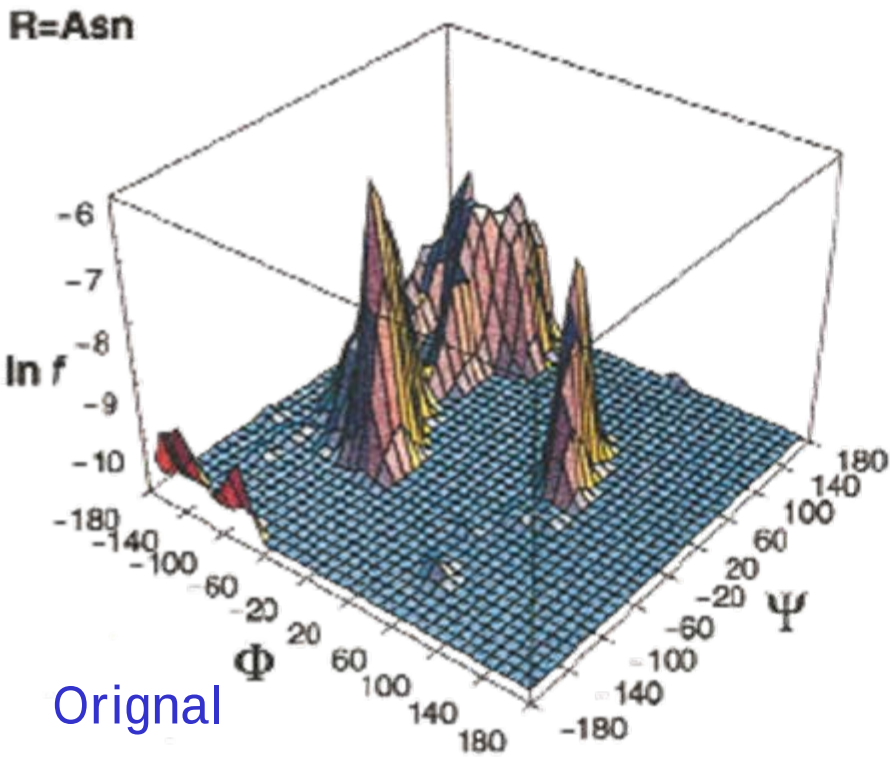
- An die rechts gezeigte Analyse von ca. 218.000 ϕ/ψ-Paaren lässt sich auch anfitten
- Gauß-Funktion für Ramachandran-Plot ungeeignet



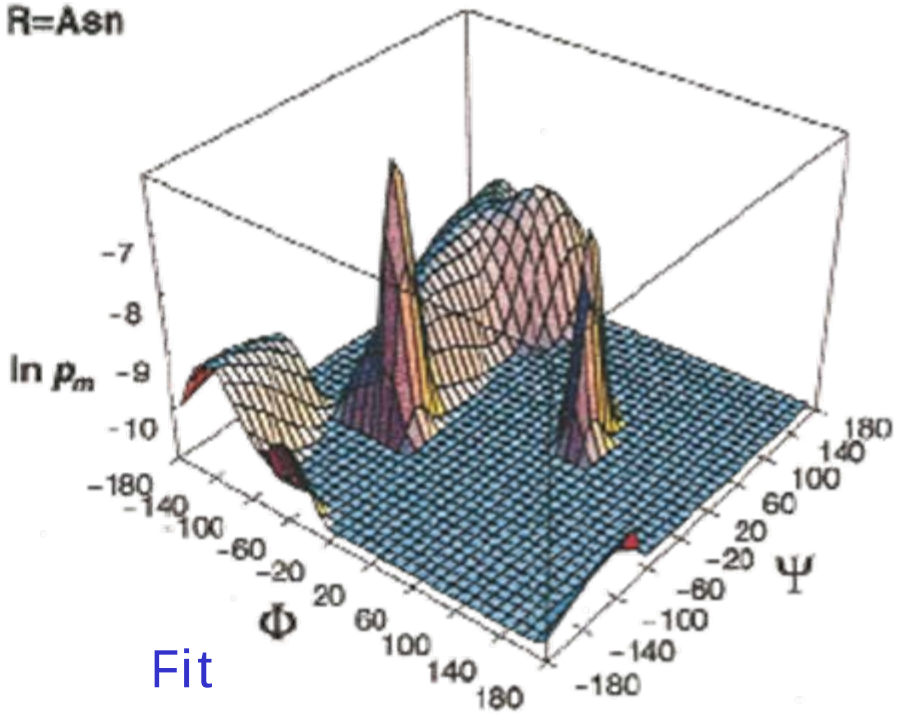
MODELLER – Energiefunktion

Ramachandran-Plot lässt sich durch folgenden Ausdruck **anfitten**:

$$p_i = \frac{1}{2\pi\sigma_{\phi,i}\sigma_{\psi,i}\sqrt{1-\rho_i^2}} \exp\left[\frac{1}{1-\rho_i^2}\left(\frac{1-\cos(\phi-\phi_i^0)}{\sigma_{\phi,i}^2} - \rho_i\frac{\sin(\phi-\phi_i^0)}{\sigma_{\phi,i}}\frac{\sin(\psi-\psi_i^0)}{\sigma_{\psi_i}} + \frac{1-\cos(\psi-\psi_i^0)}{\sigma_{\psi,i}^2}\right)\right]$$



Orignal



Fit

MODELLER – Energiefunktion

$$E = E_{\text{CHARMM}} - \sum_{\text{sc tors}} \ln p_s(\chi) - \sum_{\text{residues}} \ln p_\omega(\omega)$$
$$- \sum_{\text{residues}} \ln p_{\phi\psi}(\phi,\psi) + \sum_{i,j\in\text{nb}} \xi[E_{\text{pw}}(i,j) + S(i,j)]$$

- Schließlich erfasst der letzte Term WW zwischen nichtgebundenen (nb) Atomen
- E_{pw} entspricht dabei einem Sippl-Potential (aber für alle Atome, statt für gesamte AS)
- S ist ein zusätzliches abstoßendes harmonisches Potential, dass die Ungenauigkeiten des empirischen Potentials für kleine Abstände unterdrückt
- ξ ist ein Skalierungsfaktor, der das An- und Ausschalten der paarweisen WW erlaubt

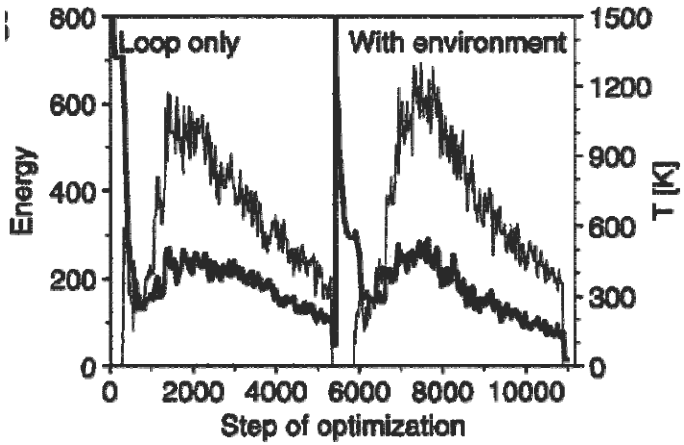
MODELLER – Konstruktion

- Für jede Schleife werden die Seitenketten in gerader Linie zwischen die Ankerreste platziert
- Dann wird eine **zufällige Startkonformation** erzeugt, indem alle Atomkoordinaten zufällig durch Addition einer Zufallszahl um ±5 Å im Raum verschoben werden
- Diese zufällige Startstruktur wird dann einer **Optimierung** unterzogen
- 50-500 solcher Startkonformationen werden unabhängig erzeugt und jeweils optimiert
- Aus diesen optimierten Strukturen wird die energetisch günstigste Loop-Struktur ausgewählt

MODELLER – Optimierung

Jede Startkonformation wird folgendermaßen optimiert:

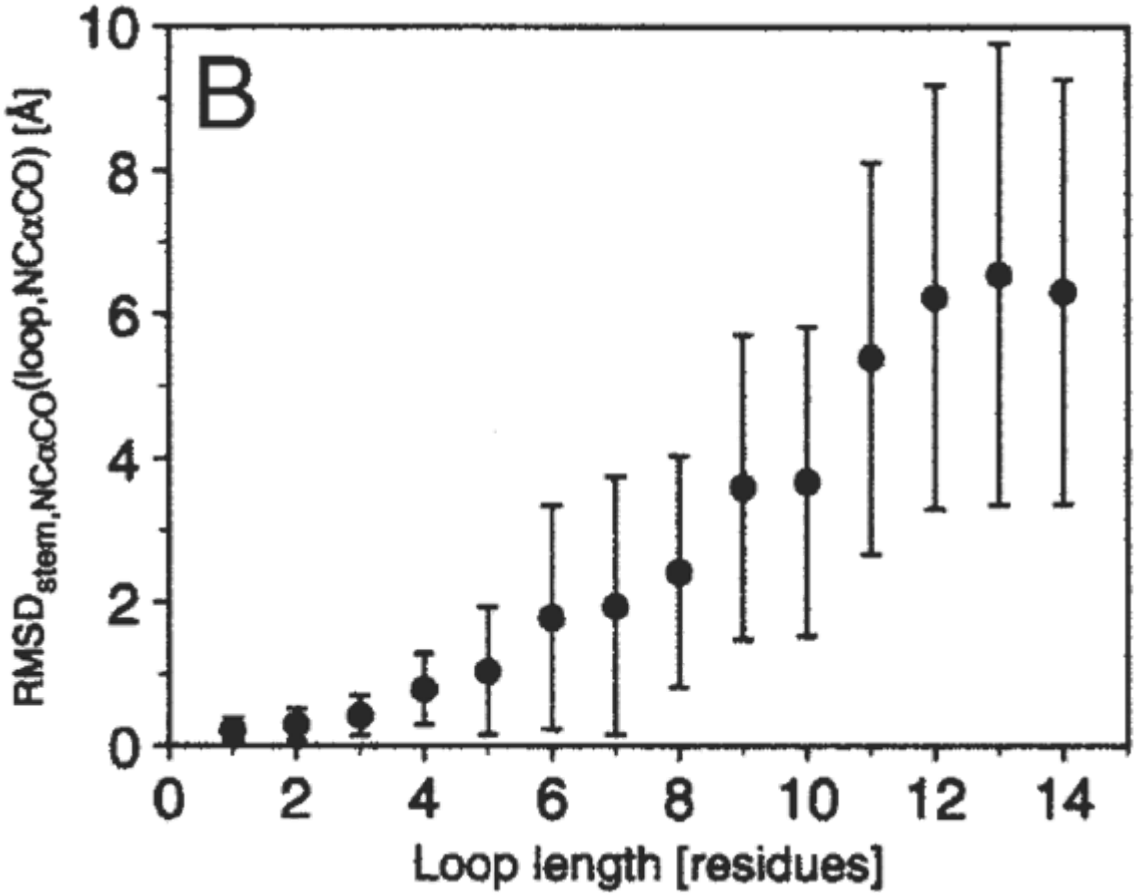
- Fünf **Minimierungen** mit konjugiertem Gradient
  (200 Schritte jeweils, mit ξ = 0/0.01/0.1/0.5/1.0)
  Dabei werden durch zunehmendes ξ die abstoßenden Paarpotentiale Stück für Stück eingeschaltet
- **Aufheizen mit MD-Simulationen** (je 4 fs) bei 150/250/400/700/1000 K
- **Abkühlen mit MD-Simulationen** (je 4 fs) bei 1000/800/600/500/400/300 K
- **Minimierung** mit konjugiertem Gradient



MODELLER – Laufzeiten

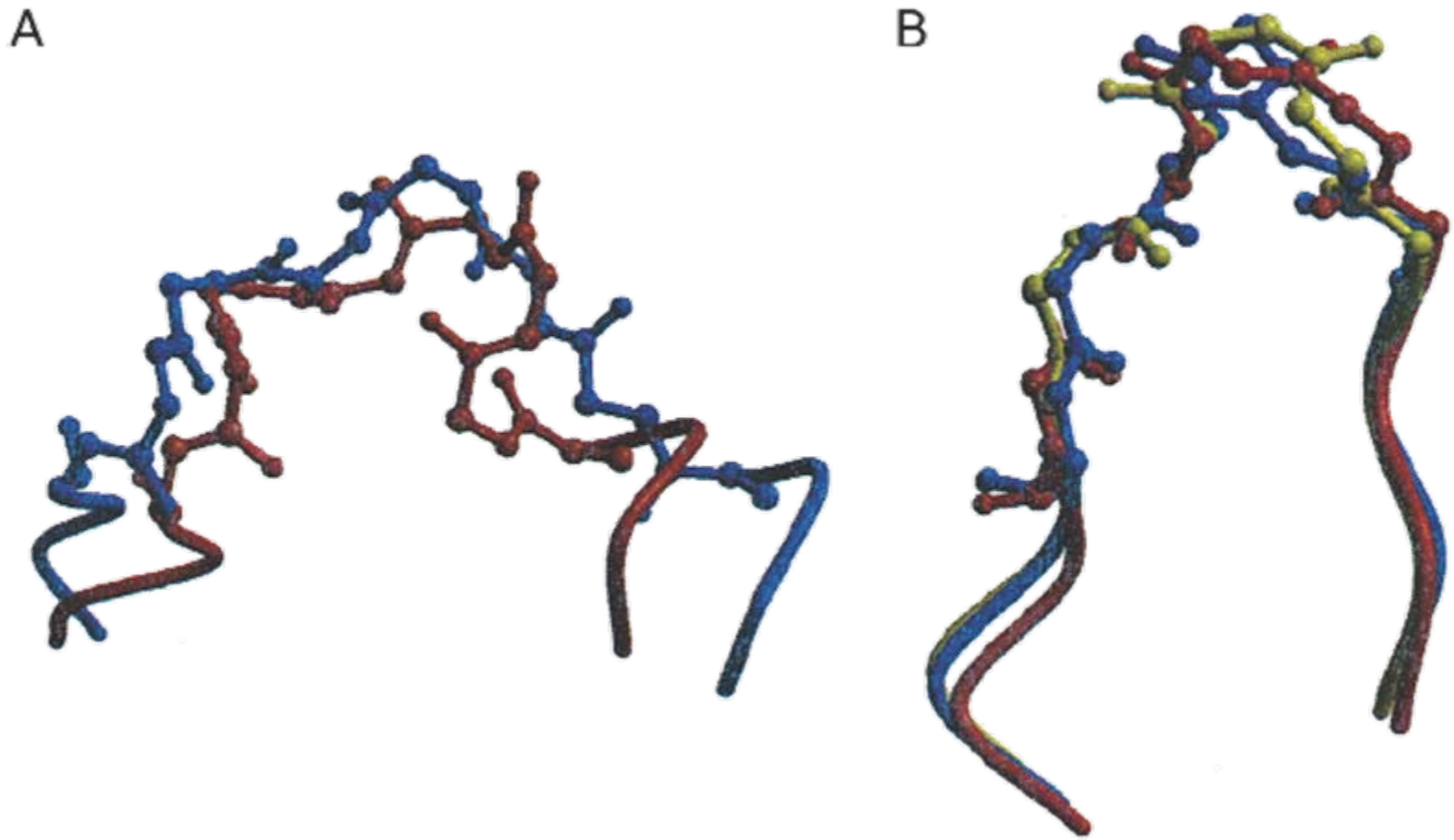
- Dieser Zyklus wird zunächst für die Loop alleine
- durchgeführt, dann für die Loop in der Umgebung des
- Templates wiederholt
- Optimierungszyklus ist extrem rechenaufwändig
- Für Loops der Länge 8 werden etwa 8-30 CPU h benötigt
- Die erhaltenen Strukturen sind jedoch qualitativ recht gut und scheinen die Laufzeit zu rechtfertigen
- Die Simulationen lassen sich problemlos verteilen und parallel ausführen (Cluster)

MODELLER – Ergebnisse



- RMSD für Loops unterschiedlicher Länge
- Bis zu Längen von acht Resten sehr gute Genauigkeit

MODELLER – Ergebnisse



Zwei Beispiele aus CASP3:
Nativ: blau, modellierte: rot, nächste Struktur: gelb

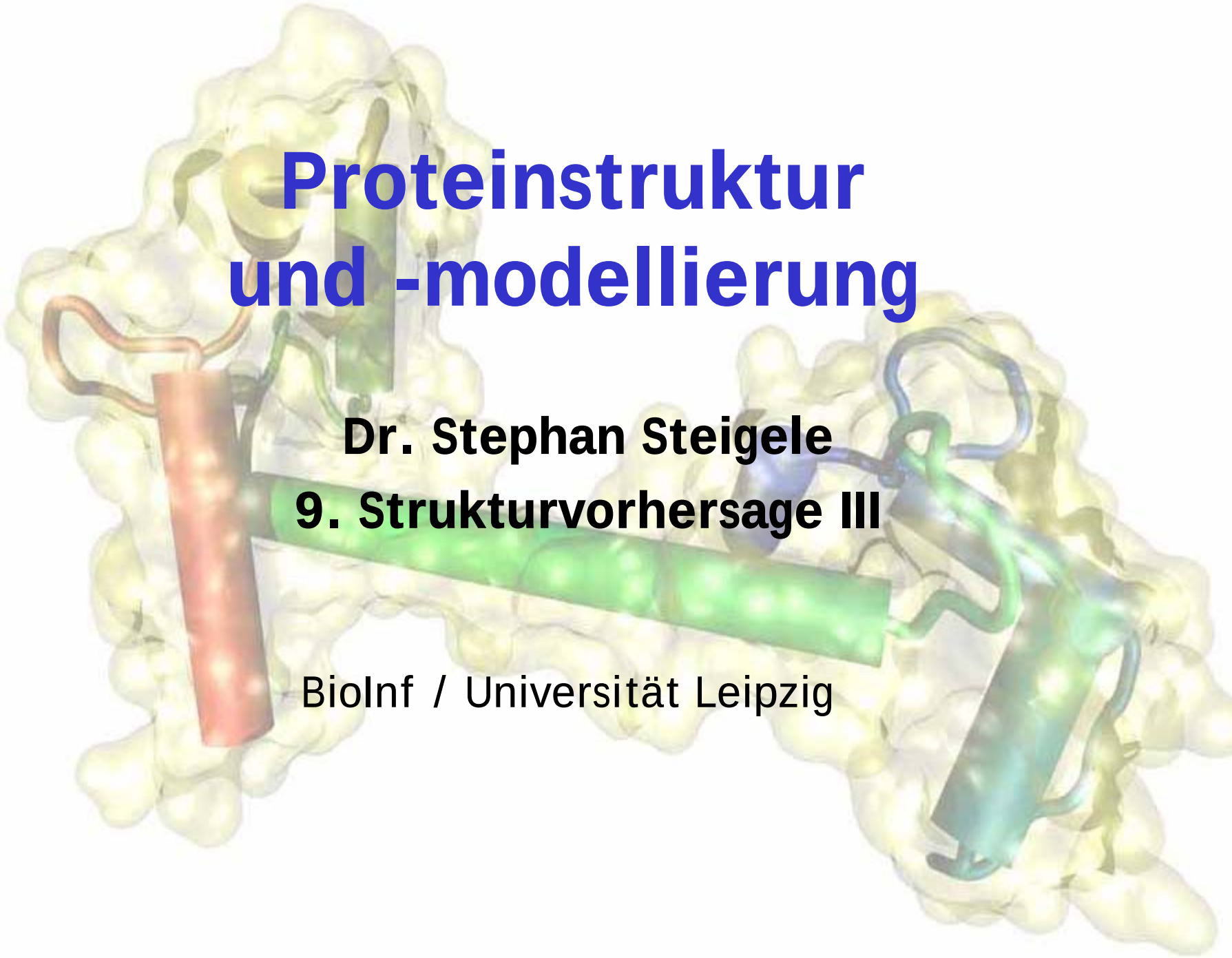
Literatur

Homologiemodellierung

- **Roland Dunbrack**: Homology Modeling in Biology and Medicine, Chapter 5 in T. Lengauer (Hrsg.): Bioinformatics: From Genomes to Drugs, Wiley, 2002

- **Roberto Sánchez, Andrej Šali**: Comparative Protein Structure Modeling, in D. Webster (Hrsg.): Protein Structure Prediction: Methods and Protocols, Humana Press, New Jersey, 2000
-
- **Patric Koehl, Michael Levitt**: A brighter future for protein structure prediction, Nat. Struct. Biol. (1999), 6, 108

Publikationen zu den einzelnen Methoden: siehe Website



Proteinstruktur und -modellierung

Dr. Stephan Steigele

9. Strukturvorhersage III

BioInf / Universität Leipzig

Strukturvorhersage – Übersicht

- Problemdefinition/-klassifizierung
- Sekundärstrukturvorhersage
- Fold-Recognition
- Threading
- ab-initio-Vorhersage

Gliederung

- Definition, Begriffe
- Übersicht: Modelle und Algorithmen
- Exemplarische Methoden
  - Grobe Modelle – Packen von Helices (Nanias et al.)
  - MM-definierte Potentiale (Eyrich et al.)
  - Fragment-Assembly: ROSETTA (Simons et al.)
- Stand der Technik: CASP5

Definition

Ab-initio-Vorhersage (de-novo-Vorhersage)

- Ausgehend von „ersten Prinzipien“
- Ohne Einbeziehung einer Struktur mit gleichem Fold
- Erzeugt eine Struktur mit **neuem Fold**
- Kann auch zur reinen Fold Recognition verwendet werden

Nachteile

- **Weniger genau** als Threading
- Problem ist schwieriger als Threading (größerer Suchraum)

Vorteil

- Anders als Threading **für neue Folds geeignet**
- Nicht durch Templates (Bibliothek) beeinflusst

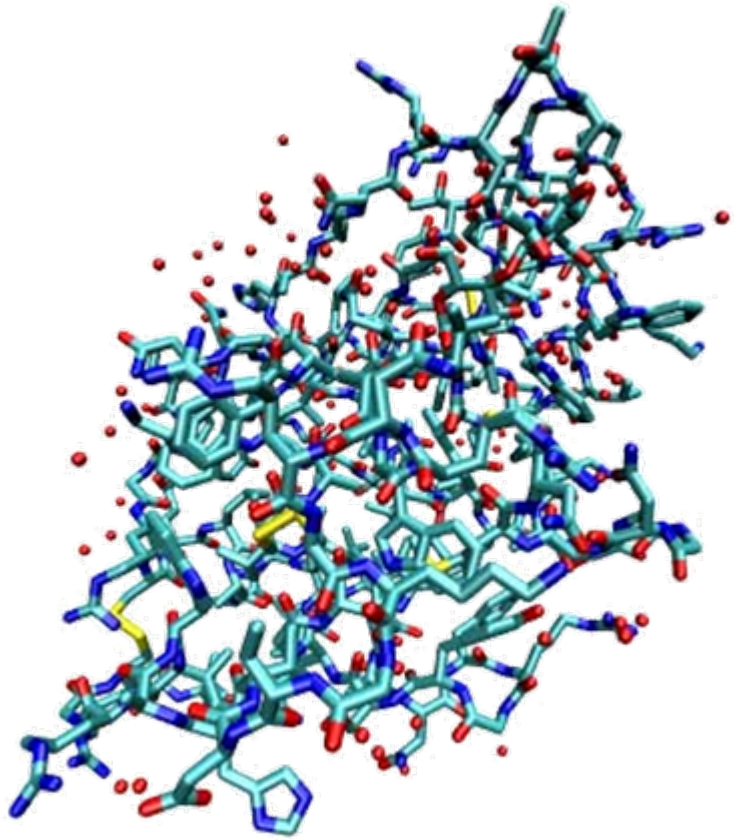
Methoden

- Methoden unterscheiden sich hinsichtlich
  - Potentialen
  - Auflösung der Modelle
  - Suchstrategie
- Allgemeine Problemformulierung
  - Gegeben eine Sequenz, finde Struktur
  - Ohne Verwendung bekannter Strukturen
  - Struktur sei optimal bezüglich Energiefunktion
  - Freiheitsgrade gegeben durch Modellauflösung

Modellauflösungen

Atomare Auflösung

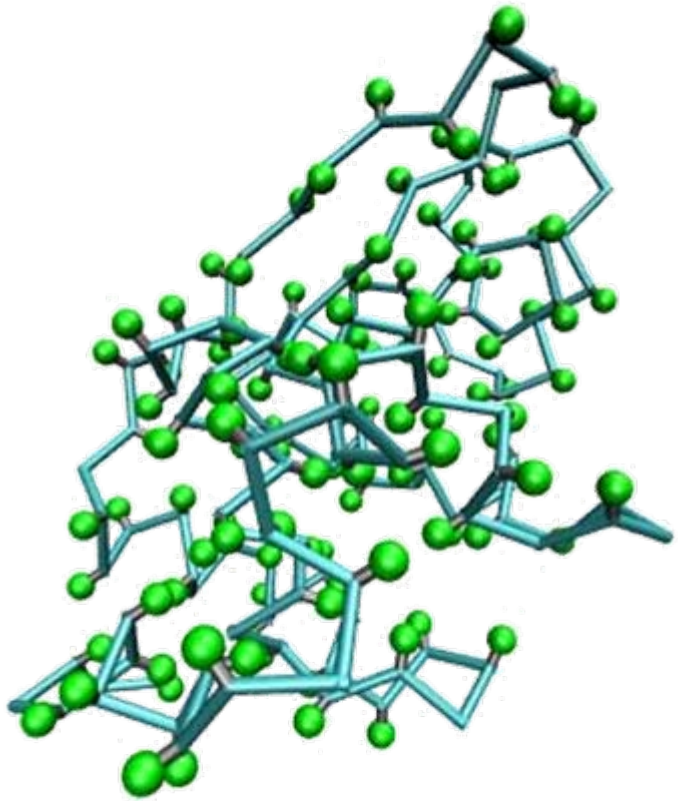
- Notwendig für MM-Kraftfelder
- $>10^3$ Freiheitsgrade
- Nur für kleine Peptide machbar
- Methoden
  - MD-Simulation
  - MC-Simulation
- Weitere Reduktion der FG durch Wechsel in Torsionsraum



Modellauflösungen

Reduktion der Seitenketten

- Seitenketten werde durch C_β oder Schwerpunkt ausgedrückt
- # FG: 2 ¢ #AS (Torsionen ϕ, ψ)
- Potentiale beschränkt auf empirische Kontaktpotentiale
- Immer noch schwieriger als Threading mit paarweisen Potentialen (keine Schablone!)
- Ähnlich: Gittermodelle



Modellauflösungen

Reduktion auf Sekundärstrukturen

- Auch Reduktion auf Helices allein
- Sekundärstrukturelemente aus entsprechenden Vorhersagen
- Empirische Kontaktpotentiale
- Sehr geringe Anzahl an FG (Positionen, Orientierungen)
- Nur zur Fold Recognition geeignet

Modellauflösungen

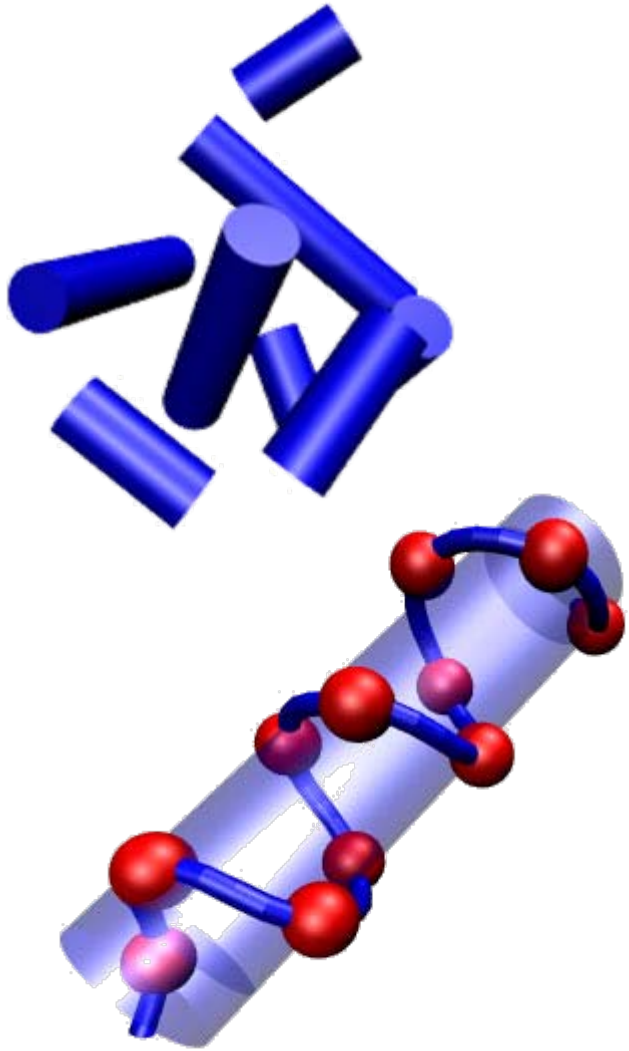
- Es existieren viele weitere **Stufen der Modellauflösung** zwischen den drei genannten
- Potentiale und Suchstrategien müssen spezifisch an Modellauflösung angepasst werden
- Modellauflösung bestimmt Ergebnisse: von reiner Fold Recognition bis zu gut aufgelösten Modellen
- Je höher die Modellauflösung (und damit die Anzahl der Freiheitsgrade), desto geringer die Erfolgsaussichten

Algorithmen

- **Monte-Carlo-Varianten**
  - Nanias et al., PNAS (2003), 100, 1706
  - Simons et al, J. Mol. Biol. (1997), 268, 209 (ROSETTA)
- **Molekulardynamik**
  - Lee et al., J. Mol. Biol. (2001), 313, 417
- **Evolutionäre Algorithmen**
  - Pedersen, Moult, J. Mol. Biol. (1997), 269, 240
- **Branch & Bound-Ansätze**
  - Eyrich et al., Proteins (1999), 35, 41
- ...

Nanias et al. (2003)

- Optimales Packen von Helices (nur α-helikale Proteine)
- **Modell**
  - Reduktion auf **starre Helices** aus C_α
  - Helices durch **Schwerpunktlage und Richtung** beschrieben
  - Lage der AS in der Helix vorgegeben durch **idealisierte Helixgeometrie**
    - 3,6 AS pro Windung
    - 1,5 Å Höhe pro AS
    - 3,8 Å Abstand zwischen benachbarten C_α



Nanias, Chinchio, Pillardy, Ripoll, Scheraga, PNAS (2003), 100, 1706

Miyazawa & Jernigan

- Miyazawa & Jernigan stellten ein Kontaktpotential vor, das auf einem einfachen Gittermodell basiert und Kontakte mit anderen Aminosäuren bzw. Lösemittel (leere Gitterplätze) zählt
- Seitdem wurden viele verbesserte Versionen dieses Potentials vorgestellt
- Problematisch:
  - Wahl des **Referenzzustandes**
    - Was ist der „interaktionslose Zustand“ eines AS-Paares?
    - Interaktion mit Wasser? Interaktion mit einer Referenzaminosäure?
  - Potential basiert auf Paarkontakten, dadurch ist die Energiefunktion **nicht stetig**, was für Optimierungen ungünstig ist

Miyazawa, Jernigan, Macromolecules (1985), 18, 534-552
Miyazawa, Jernigan, J. Mol. Biol. (1996), 256, 623

Miyazawa & Jernigan

Table 3. Contact energies in RT units; e_{ij} for upper half and diagonal and e'_{ij} for lower half

		Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Asn	Gln	Asp	Glu	His	Arg	Lys	Pro	
	Cys	–5.44	–4.99	–5.80	–5.50	–5.83	–4.96	–4.95	–4.16	–3.57	–3.16	–3.11	–2.86	–2.59	–2.85	–2.41	–2.27	–3.60	–2.57	–1.95	–3.07	Cys
	Met	0.46	–5.46	–6.56	–6.02	–6.41	–5.32	–5.55	–4.91	–3.94	–3.39	–3.51	–3.03	–2.95	–3.30	–2.57	–2.89	–3.98	–3.12	–2.48	–3.45	Met
	Phe	0.54	–0.20	–7.26	–6.84	–7.28	–6.29	–6.16	–5.66	–4.81	–4.13	–4.28	–4.02	–3.75	–4.10	–3.48	–3.56	–4.77	–3.98	–3.36	–4.25	Phe
	Ile	0.49	–0.01	0.06	–6.54	–7.04	–6.05	–5.78	–5.25	–4.58	–3.78	–4.03	–3.52	–3.24	–3.67	–3.17	–3.27	–4.14	–3.63	–3.01	–3.76	Ile
	Leu	0.57	0.01	0.03	–0.08	–7.37	–6.48	–6.14	–5.67	–4.91	–4.16	–4.34	–3.92	–3.74	–4.04	–3.40	–3.59	–4.54	–4.03	–3.37	–4.20	Leu
	Val	0.52	0.18	0.10	–0.01	–0.04	–5.52	–5.18	–4.62	–4.04	–3.38	–3.46	–3.05	–2.83	–3.07	–2.48	–2.67	–3.58	–3.07	–2.49	–3.32	Val
	Trp	0.30	–0.29	0.00	0.02	0.08	0.11	–5.06	–4.66	–3.82	–3.42	–3.22	–2.99	–3.07	–3.11	–2.84	–2.99	–3.98	–3.41	–2.69	–3.73	Trp
	Tyr	0.64	–0.10	0.05	0.11	0.10	0.23	–0.04	–4.17	–3.36	–3.01	–3.01	–2.78	–2.76	–2.97	–2.76	–2.79	–3.52	–3.16	–2.60	–3.19	Tyr
	Ala	0.51	0.15	0.17	0.05	0.13	0.08	0.07	0.09	–2.72	–2.31	–2.32	–2.01	–1.84	–1.89	–1.70	–1.51	–2.41	–1.83	–1.31	–2.03	Ala
	Gly	0.68	0.46	0.62	0.62	0.65	0.51	0.24	0.20	0.18	–2.24	–2.08	–1.82	–1.74	–1.66	–1.59	–1.22	–2.15	–1.72	–1.15	–1.87	Gly
	Thr	0.67	0.28	0.41	0.30	0.40	0.36	0.37	0.13	0.10	0.10	–2.12	–1.96	–1.88	–1.90	–1.80	–1.74	–2.42	–1.90	–1.31	–1.90	Thr
	Ser	0.69	0.53	0.44	0.59	0.60	0.55	0.38	0.14	0.18	0.14	–0.06	–1.67	–1.58	–1.49	–1.63	–1.48	–2.11	–1.62	–1.05	–1.57	Ser
	Asn	0.97	0.62	0.72	0.87	0.79	0.77	0.30	0.17	0.36	0.22	0.02	0.10	–1.68	–1.71	–1.68	–1.51	–2.08	–1.64	–1.21	–1.53	Asn
	Gln	0.64	0.20	0.30	0.37	0.42	0.46	0.19	–0.12	0.24	0.24	–0.08	0.11	–0.10	–1.54	–1.46	–1.42	–1.98	–1.80	–1.29	–1.73	Gln
	Asp	0.91	0.77	0.75	0.71	0.89	0.89	0.30	–0.07	0.26	0.13	–0.14	–0.19	–0.24	–0.09	–1.21	–1.02	–2.32	–2.29	–1.68	–1.33	Asp
	Glu	0.91	0.30	0.52	0.46	0.55	0.55	0.00	–0.25	0.30	0.36	–0.22	–0.19	–0.21	–0.19	0.05	–0.91	–2.15	–2.27	–1.80	–1.26	Glu
	His	0.65	0.28	0.39	0.66	0.67	0.70	0.08	0.09	0.47	0.50	0.16	0.26	0.29	0.31	–0.19	–0.16	–3.05	–2.16	–1.35	–2.25	His
	Arg	0.93	0.38	0.42	0.41	0.43	0.47	–0.11	–0.30	0.30	0.18	–0.07	–0.01	–0.02	–0.26	–0.91	–1.04	0.14	–1.55	–0.59	–1.70	Arg
	Lys	0.83	0.31	0.33	0.32	0.37	0.33	–0.10	–0.46	0.11	0.03	–0.19	–0.15	–0.30	–0.46	–1.01	–1.28	0.23	0.24	–0.12	–0.97	Lys
	Pro	0.53	0.16	0.25	0.39	0.35	0.31	–0.33	–0.23	0.20	0.13	0.04	0.14	0.18	–0.08	0.14	0.07	0.15	–0.05	–0.04	–1.75	Pro
$e_{ir} - 2.55$	e_{ir}	–3.57	–3.92	–4.76	–4.42	–4.81	–3.89	–3.81	–3.41	–2.57	–2.19	–2.29	–1.98	–1.92	–2.00	–1.84	–1.79	–2.56	–2.11	–1.52	–2.09	
$e_r - 3.60$	e	–4.29	–4.73	–5.57	–5.29	–5.71	–4.72	–4.41	–3.87	–3.17	–2.53	–2.63	–2.27	–2.14	–2.35	–2.02	–2.07	–2.94	–2.43	–1.82	–2.53	
$f_r - 3.60$	f_i	–5.58	–6.14	–7.39	–7.09	–7.88	–6.15	–5.34	–4.60	–3.24	–2.22	–2.48	–1.92	–1.74	–1.93	–1.54	–1.49	–2.91	–2.07	–1.17	–1.97	
N_{ir}/N_i	2.096	2.723	2.722	2.780	2.811	2.893	2.728	2.537	2.493	2.143	1.840	1.973	1.771	1.699	1.720	1.598	1.508	2.075	1.787	1.343	1.629	
q_i 7.162	6.281	6.646	6.137	5.870	6.042	6.087	6.155	5.793	6.037	6.334	6.284	6.486	6.582	6.574	6.469	6.487	6.235	6.241	6.318	6.569	5.858	

- Diese Tabelle wurde aus 1168 Proteinstrukturen mit ca. 113.000 Paarkontakten abgeleitet
- Betrachtet wurden Kontakte von C_α-Atomen mit einem Maximalabstand von 6.5 Å
- Kontaktenergie berechnet sich nach dem inversen Boltzmann-Ansatz

Miyazawa, Jernigan, J. Mol. Biol. (1996), 256, 623

Nanias et al. (2003)

- Potentialfunktion basierend auf dem Kontaktpotential von Miyazawa und Jernigan
- Lennard-Jones-ähnliche Glättung des Kontaktpotentials in kontinuierliche Funktion (Reste *i, j* vom Typ *a, b):*

$$E_{ij}(r_{ij}) = \frac{E^0_{ab}}{q \pm p}\left[q\left(\frac{r_0}{r_{ij}}\right)^p \pm p\left(\frac{r_0}{r_{ij}}\right)^q\right]$$

r_{ij}	C_α-Abstand
E^0_{ab}	Kontaktenergie im Abstand r_0 für AS-Typen *a, b* (aus Miyazawa-Jernigan)
p, q, r_0	Wählbare Parameter Optimal: p = 15, q = 14, r_0 = 7,5 Å

Miyazawa, Jernigan, Macromolecules (1983), 18, 534-552
Nanias, Chinchio, Pillardy, Ripoll, Scheraga, PNAS (2003), 100, 1706

Nanias et al. (2003)

- Gesamtenergie: Summation über alle Paare *zwischen* Helices
- WW *innerhalb* der selben Helix konstant (starre Helix)
- Keine Wechselwirkungen mit Loops Zusätzliche Einschränkung
  - Helixenden dürfen nicht weiter als Schleifenlänge voneinander entfernt sein
  - Maximale Länge: #AS ¢ 3,8 Å (C_α-Abstand)

Nanias, Chinchio, Pillardy, Ripoll, Scheraga, PNAS (2003), 100, 1706

Nanias et al. (2003)

Optimierung: modifizierter MC-Algorithmus

- Arbeitet auf **Familien** von Strukturen statt auf einer Einzelstruktur (CFMC – *Configuration Family MC*)
- Zwei Klassen von MC-Schritten
  - **Global**
    - Translationen/Rotationen einer beliebigen Anzahl Helices ändern
    - Translationen bis zu 15 Å, Rotationen um 360°

) **Erzeugen drastisch unterschiedlicher Strukturen**
  - **Lokal**
    - Translation um bis zu 4 Å, Rotation um bis zu 50° ändern
    - Rotation um Helixachse (bis zu 180°)
    - Verschiebung entlang Helixachse (bis zu 3 Å)

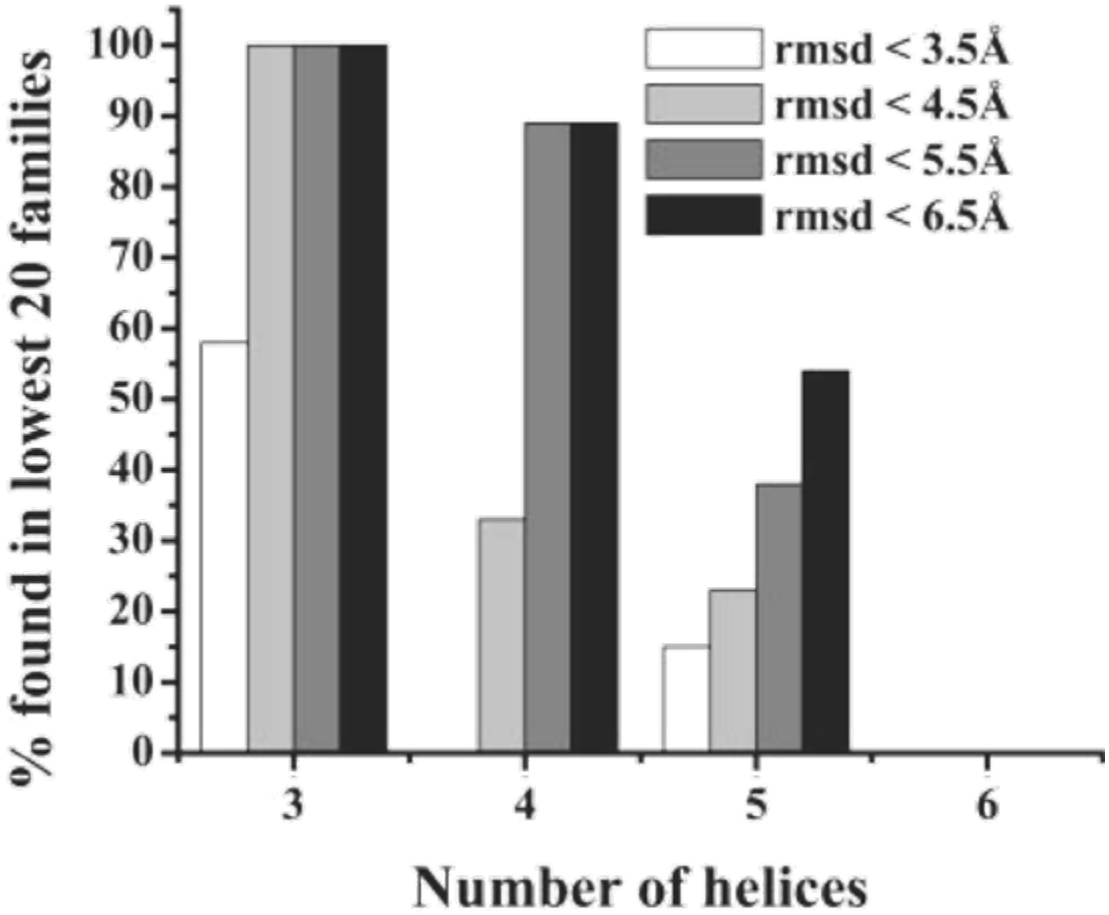
) **Kleine Bewegungen zur lokalen Optimierung**
- 75% der Schritte global, 25% lokal

) Suche nach globalem Optimum

Nanias et al. (2003)

Testsatz

- 42 Proteine
- bis zu 188 AS
- bis zu 8 Helices
- Helices aus **DSSP!**
- Helices aus JPRED ergaben jedoch ähnliche Resultate



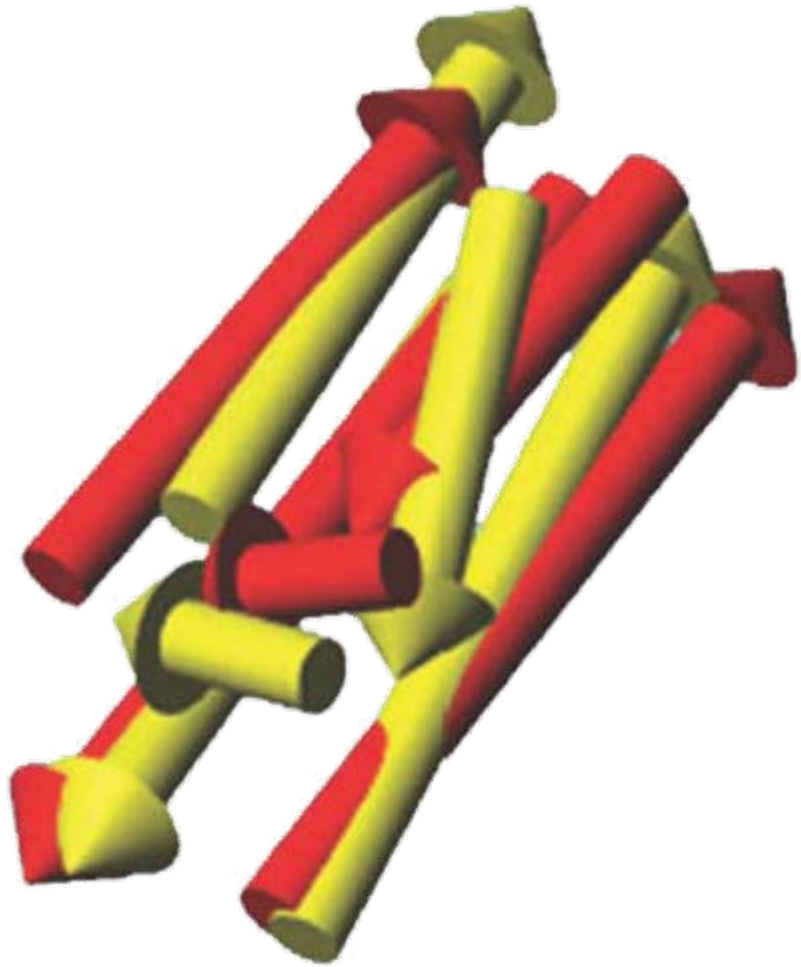
Nanias, Chinchio, Pillardy, Ripoll, Scheraga, PNAS (2003), 100, 1706

Nanias et al. (2003)

Zusammenfassung

- Sehr trivialer Ansatz
- Ab-initio-Fold-Erkennung für α-helikale Proteine mit geringem Aufwand
- Nicht ohne weiteres auf Faltblätter generalisierbar
- Modellqualität teilweise sehr gut (RMSD ~ 3 Å)

) selbst grobe Modelle recht nützlich: Membranproteine?



Nanias, Chinchio, Pillardy, Ripoll, Scheraga, PNAS (2003), 100, 1706

Potentialfunktionen

- Weit verbreitete Potentialtypen sind
  - Empirische Kontaktpotentiale
  - Solvatationspotentiale
  - Molekülmechanische Kraftfelder
- Potentiale müssen der Modellauflösung angepasst sein
- Verschmelzung von wissensbasierten Potentialen und Molekülmechanik möglich

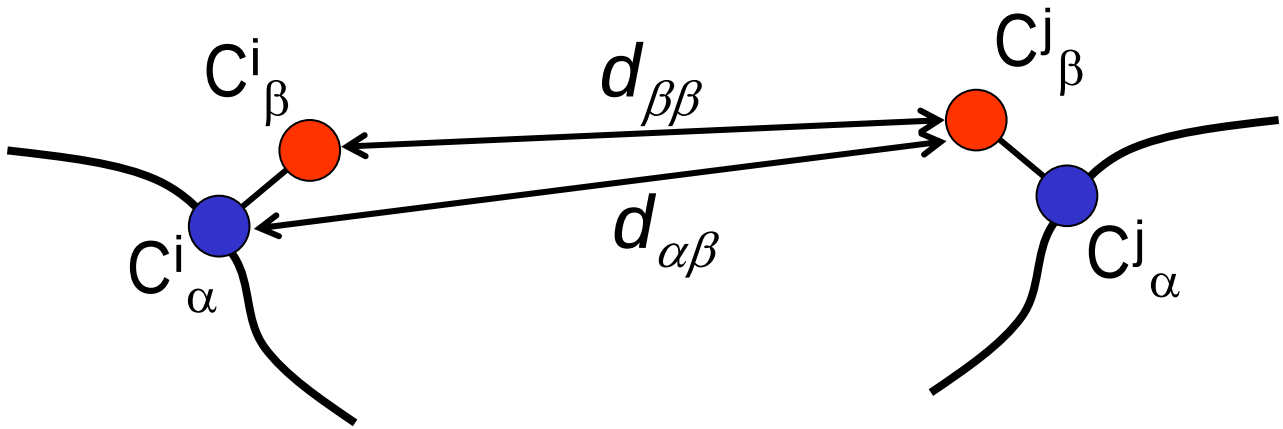
SK-Potentiale mit AMBER

- Eyrich et al. repräsentieren Seitenkette (SK) durch C_β
- **Idee**
  - Proteinstrukturen enthalten sinnvolle Seitenketten-konformationen
  - Statt Pseudoenergien aus der Häufigkeit zu berechnen, berechne gemittelte WW mit AMBER
  - Interpoliere die so bestimmten Potentiale analytisch
- **Vorteile**
  - Weniger Artefakte aus Strukturdatenbanken
  - Höhere Genauigkeit
  - Analytische, differenzierbare Funktion für Potential

Eyrich, Standley, Felts, Friesner, Proteins (1999), 35, 41

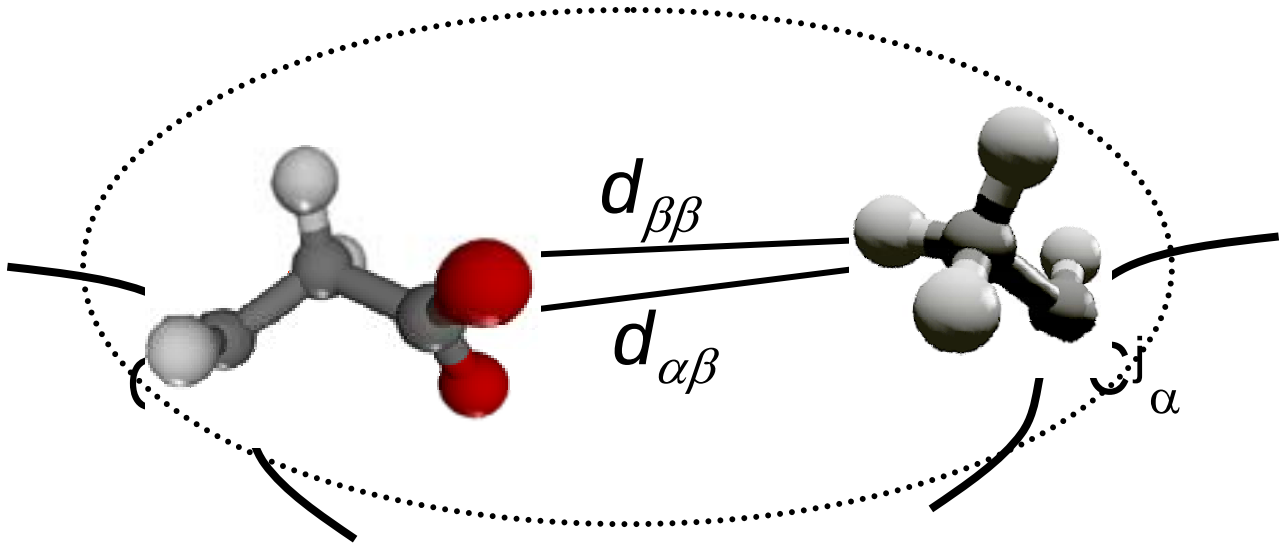
Eyrich et al. (1999)

- WW zwischen zwei AS *i*, *j* ist Funktion von
  - AS-Typen
  - C_α/C_β-Abstände ($d_{\alpha\alpha}$, $d_{\alpha\beta}$, $d_{\beta\alpha}$, $d_{\beta\beta}$)
- $E_{ab}(d_{\alpha\alpha}, d_{\alpha\beta}, d_{\beta\alpha}, d_{\beta\beta})$ wird tabelliert für 30 verschiedene Abstandsbereiche für die Distanzen (0 – 28 Å)



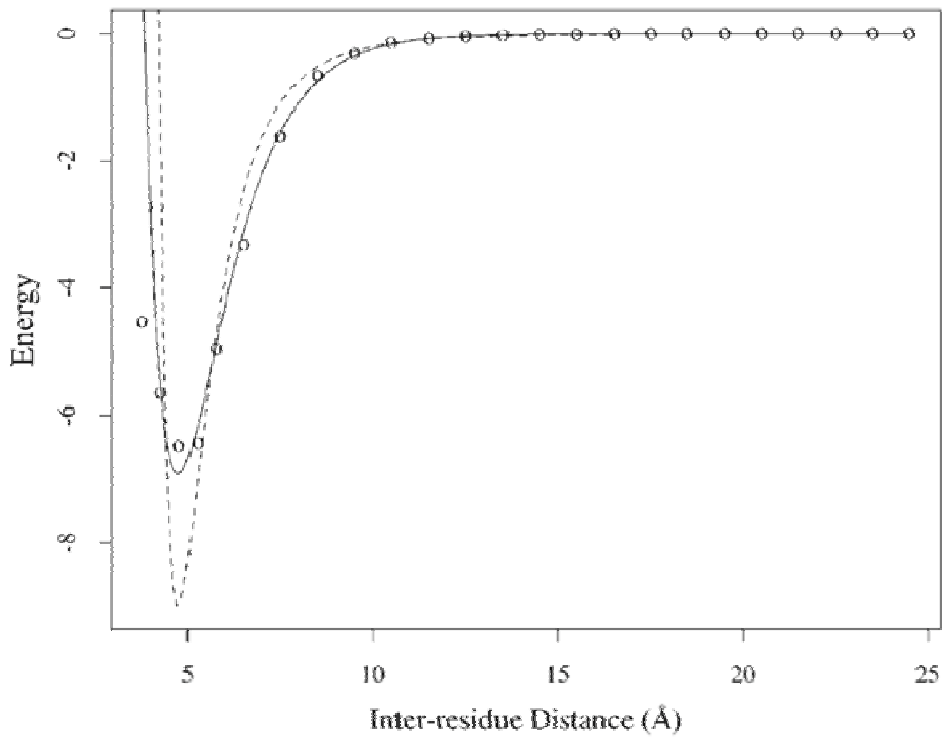
Eyrich et al. (1999)

- Für jedes Paar der Datenbank
  - Extrahiere Koordinaten der beiden Reste
  - Berechne vdW-Energie mit AMBER89
  - Speichere Energie für die jeweiligen d_{ij}
- Mittele die Energien für jeden Tabelleneintrag



Interpolation des Potentials

- Morse-Potential E_M (——)
  nähert die WW besser an als
  Lennard-Jones-Potential (- - - -)
- Morse-Potential ist „weicher“ als LJ-Potential
- Seitenketten sind flexibel
- Potential ist gemittelt über viele Konformationen
- Seitenketten können einander ausweichen



$$E_M(r) = \varepsilon \left(1 - e^{-\beta(r-\sigma)}\right)^2 - \varepsilon$$

E_M wird an die Tabellenwerte für jeden der vier Indices interpoliert.

Gesamtes vdW-Potential

- Um die Stärke des repulsiven Teils getrennt variieren zu können, wird dieser Teil als Polynom modelliert
- vdW-Potential also abschnittsweise definiert

$$E_{vdW}(r) = \begin{cases} ar^4 + b & \text{falls} \quad r \leq r_0 \\ E_M(r) & \text{falls} \quad r > r_0 \end{cases}$$

mit $a = -\varepsilon\beta / r_0^3$, $b = \varepsilon\beta r_0$

- Potential zwischen zwei Resten wird für alle vier Distanzen ($d_{\alpha\alpha}$, $d_{\beta\beta}$, $d_{\alpha\beta}$ und $d_{\beta\alpha}$) aufsummiert
- Potential wird zwischen allen nichtbenachbarten Resten summiert

Gesamte Potentialfunktion

- Neben der vdW-Interaktionsenergie werden noch folgende Terme betrachtet
  - E_{hyd} – Hydrophobe WW: Casari & Sippl, JMB (1992) (empirisches Kontaktpotential)
  - E_{ov} – Überlappungsterm: Morsepotential das Überlappungen der Seitenketten zusätzlich bestraft

$$E_{ov} = \exp\left(-\left(\frac{r_{ij}}{r_0}\right)^{10}\right)$$

- Gesamtpotential

$$E_{ges} = c_{vdW}E_{vdW} + c_{hyd}E_{hyd} + c_{ov}E_{ov}$$

- Da die Exponentialfunktion in den Morsepotentialen rechnerisch aufwändig ist, werden die Potentiale für feste Abstände tabelliert und dazwischen schnell interpoliert

Algorithmus

- Eyrich *et al.* verwenden einen **Branch & Bound-Algorithmus**
- Suche erfolgt dabei im **Torsionswinkelraum**, jedem Winkel ist dabei ein Intervall zugeordnet
- Es wird ein Torsionswinkel ausgewählt
- Intervall dieses Winkels geteilt, sodass zwei unabhängige Teilprobleme entstehen
- Für jedes Teilproblem wird eine untere Schranke durch eine **quadratische Interpolationsfunktion** berechnet
- Teilproblem mit dem niedrigsten Wert für die untere Schranke wird dann weiter unterteilt

Ergebnisse

- Vorhersage der Struktur für acht Proteine
- Für sechs der acht wurde eine korrekte Struktur (RMSD < 6 Å) als optimale Lösung gefunden

Nachteile

- Rechenzeiten sind sehr hoch: bis zu 120 CPU h pro Protein

Vorteile

- Algorithmus ist deterministisch
- Resultate sind global optimal (bezüglich *E*)

Fragment Assembly

- Fragment-Assembly-Methoden wie ROSETTA verwenden

  Fragmente bekannter Strukturen

) nicht „ab initio“ im engeren Sinn

  (auch „Mini-Threading“ genannt)

- Fragmente haben eine Reihe von Vorteilen
  - Fragmente **reduzieren und diskretisieren den Suchraum**
  - Fragmente sind nahe an **optimalen Teilkonformationen**: jedes Fragment kommt aus einer Struktur die selbst optimal ist

ROSETTA

Kernideen

- Betrachtung des Konformationsraums für Teilsequenzen
- Teilsequenzen nehmen nur geringe Anzahl energetisch günstiger Konformationen an
- Diese Konformationen werden durch einen Satz Fragmente äquivalent repräsentiert
- Konformationen der Fragmente überwiegend durch lokale WW bestimmt

Simons et al., J. Mol. Biol. (1997), 268, 209
Simons et al., Proteins (1999), 34, 82

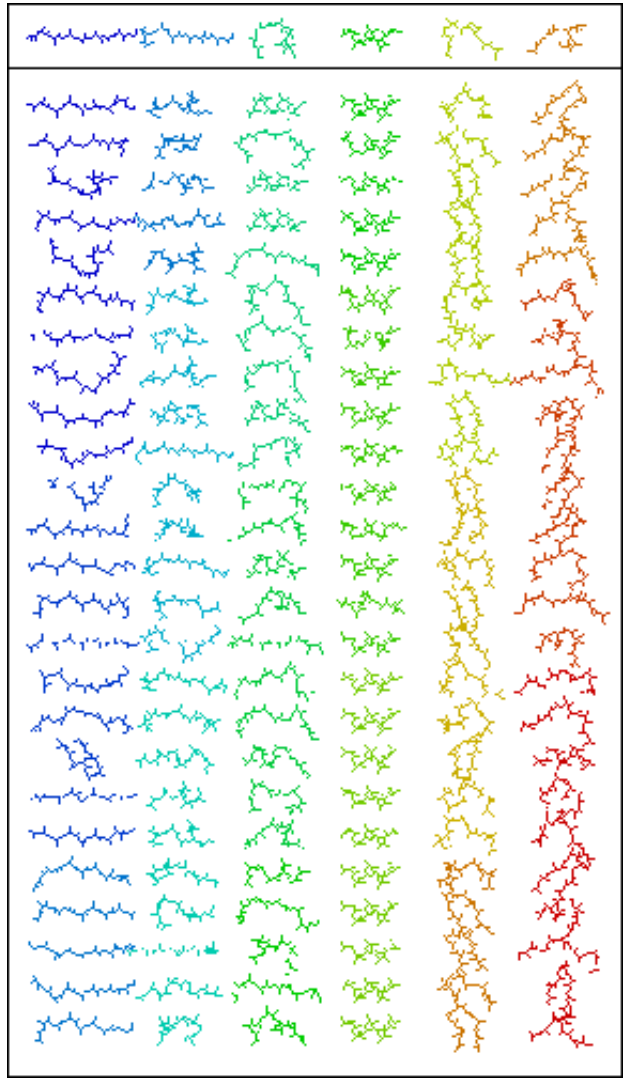
ROSETTA

- **Modell**
  - Torsionswinkelraum, reduziert auf Fragmente
  - Seitenketten auf C_β reduziert
- **Potentialfunktion**
  - Wahrscheinlichkeitsbasiert (Bayes-Ansatz)
- **Algorithmus**
  - Simulated Annealing:

    MMC mit linear sinkender Temperatur
  - Feste Anzahl Schritte (10000)

Fragmentbibliothek

- Abgeleitet aus nicht-redundantem Teilsatz der PDB
- Aus den Strukturen werden alle 9-mere und 3-mere gesammelt
- Zu jeder Teilsequenz der Zielsequenz werden daraus die 25 nächsten Fragmente ausgewählt
- ROSETTA 97 verwendet Fragmente der Länge 9
- ROSETTA 99 verwendet Fragmente der Längen 3 und 9



Bewertungsfunktion I

Idee

Bestimme die **wahrscheinlichste** Struktur der Sequenz ausgehend von Wissen aus bekannten Strukturen.

- Berechne also die Wahrscheinlichkeit *P(y|x)* dafür, dass die Sequenz *x* die Struktur *y* annimmt. Der negative log dieses Werts dient als Bewertungsfunktion.

- Mit dem **Satz von Bayes** ergibt sich für *P(y|x)*:

$$P(y|x) = P(y)\frac{P(x|y)}{P(x)}$$

- Vergleicht man verschiedene Strukturen für die selbe Sequenz, so ist *P(x)* **= const.**

$$P(y|x) \propto P(y)P(x|y)$$

Bewertungsfunktion II

- Die Bewertungfunktion von ROSETTA, d.h. die Definition von *P(y)* und *P(y|x)*, hat sich mehrmals geändert. Wir betrachten zwei Varianten:
  ROSETTA 97 (Simons et al., J. Mol. Biol. (1997), 268, 209) und
  ROSETTA 99 (Simons et al., Proteins (1999), 35, 43)

ROSETTA 97

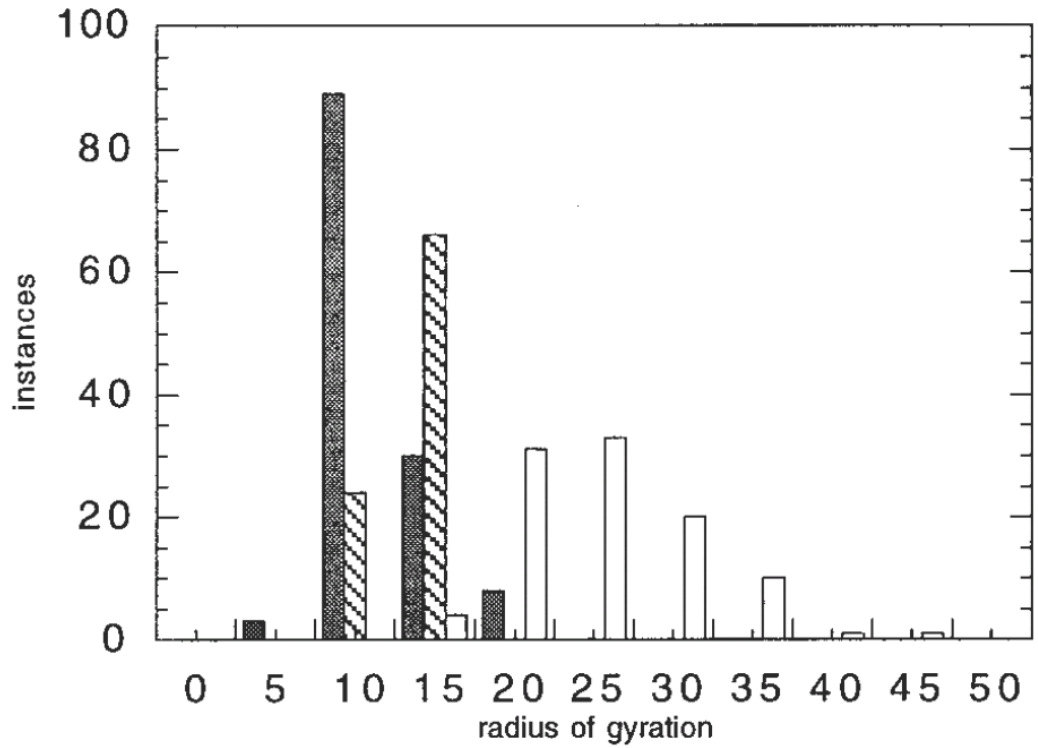
P(y) : a-priori-Wahrscheinlichkeit für die Struktur *y*.
„Wahrscheinlichkeit, dass *y* eine Proteinstruktur ist."

- *P(y)* = 0, falls *y* überlappende Atome enthält (chemisch unsinnig)
- Sinnvolle Proteinstrukturen sollten **kompakt** sein.
) Nähere *P(y)* über den **Trägheitsradius** r_g (*radius of gyration*) an:

$$P(y) \propto e^{-r_g^2}$$

Bewertungsfunktion III

- r_g für die Bewertung zu nehmen ist eine triviale, aber wirkungsvolle Möglichkeit
- Abb. zeigt die Häufigkeit bestimmter Werte von r_g
  - Kleine Proteine aus der PDB
  - Zufällig aus Fragmenten generierte Strukturen
    - Ohne Bewertungsfunktion
    - Mit $P(y) = \exp(r_g^2)$
- Mit Bewertung werden realistischerer Werte für r_g erzielt



schwarz: Strukturen aus der PDB (50-150 AS)
weiß: zufällig generierte Strukturen
schraffiert:zufällig generiert mit Bewertungsfunktion $P(y) = \exp(r_g^2)$

Bewertungsfunktion IV

ROSETTA 97

P(x|y) : wie gut „passt“ die Sequenz *x* in die Struktur *y*

Annahmen:

- Paare von AS sind **unabhängig**

) Gesamtwahrscheinlichkeit als Produkt über Paare

- Wahrscheinlichkeit für Paar ist **Funktion des Abstands**

$$P(x|y) = \prod_{i<j} P(x_i, x_j | r_{ij})$$

Dann gilt:

$$P(x_i, x_j | r_{ij}) = P(x_i, x_j) \frac{P(r_{ij}|x_i, x_j)}{P(r_{ij})}$$

Unabhängig von Struktur

Sippl-Potential!

Bewertungsfunktion V

ROSETTA 99 – *P(x|y)*

Idee: Reihenentwicklung von P(x|y)

$$
\begin{aligned}
P(x|y) &= P(x_1, x_2, \ldots, x_n|y) \\
&= \prod_i P(x_i|y) \prod_{i<j} \frac{P(x_i, x_j|y)}{P(x_i|y)P(x_j|y)} \\
&\quad \prod_{i<j<k} \frac{P(x_i, x_j, x_k|y)P(x_i|y)P(x_j|y)P(x_k|y)}{P(x_i, x_j|y)P(x_j, x_k|y)P(x_j, x_k|y)} \cdots
\end{aligned}
$$

- Die ersten beiden Terme scheinen relevant
- Erster Term: **lokale Umgebung**
- Zweiter Term: **paarweise WW**

Bewertungsfunktion VI

- Annäherung der ersten Terme:

$$P(x|y) = P_{\text{lok}} P_{\text{pw}}$$

- Erster Term: **lokale Umgebung**

$$P_{\text{lok}} = \prod_i P(x_i|E_i^{\text{lok}})$$

$P(x_i|E_i^{lok})$: Wahrscheinlichkeit, AS *i* in Umgebung E_i^{lok} zu finden

Lokale Umgebung E_i^{lok} definiert über Anzahl Kontakte) **Kontaktkapazitätspotential**

Bewertungsfunktion VII

Zweiter Term: **paarweise WW**

$$P_{\mathrm{pw}} = \prod_{i<j} \frac{P(x_i,x_j|E_i,E_j,r_{ij})}{P(x_i|E_i^{\mathbf{pw}},r_{ij})P(x_j|E_j^{\mathbf{pw}},r_{ij})}$$

- Potential abgeleitet aus Strukturdatenbank
- r_{ij}: grobe Distanzklassen (0-7, 7-10, 10-12, >12 Å)
- E_i^{pw} in zwei Klassen eingeteilt
  - *buried*: mehr als 16 weitere Reste im Abstand < 10 Å
  - *not buried*: 16 oder weniger Reste in Kontakt

) **hydrophobes Kontaktpotential**

Bewertungsfunktion VIII

ROSETTA 99

P(y): abgewandelter Term, der folgende Beiträge enthält:

vdW – eine van-der-Waals-ähnliche WW

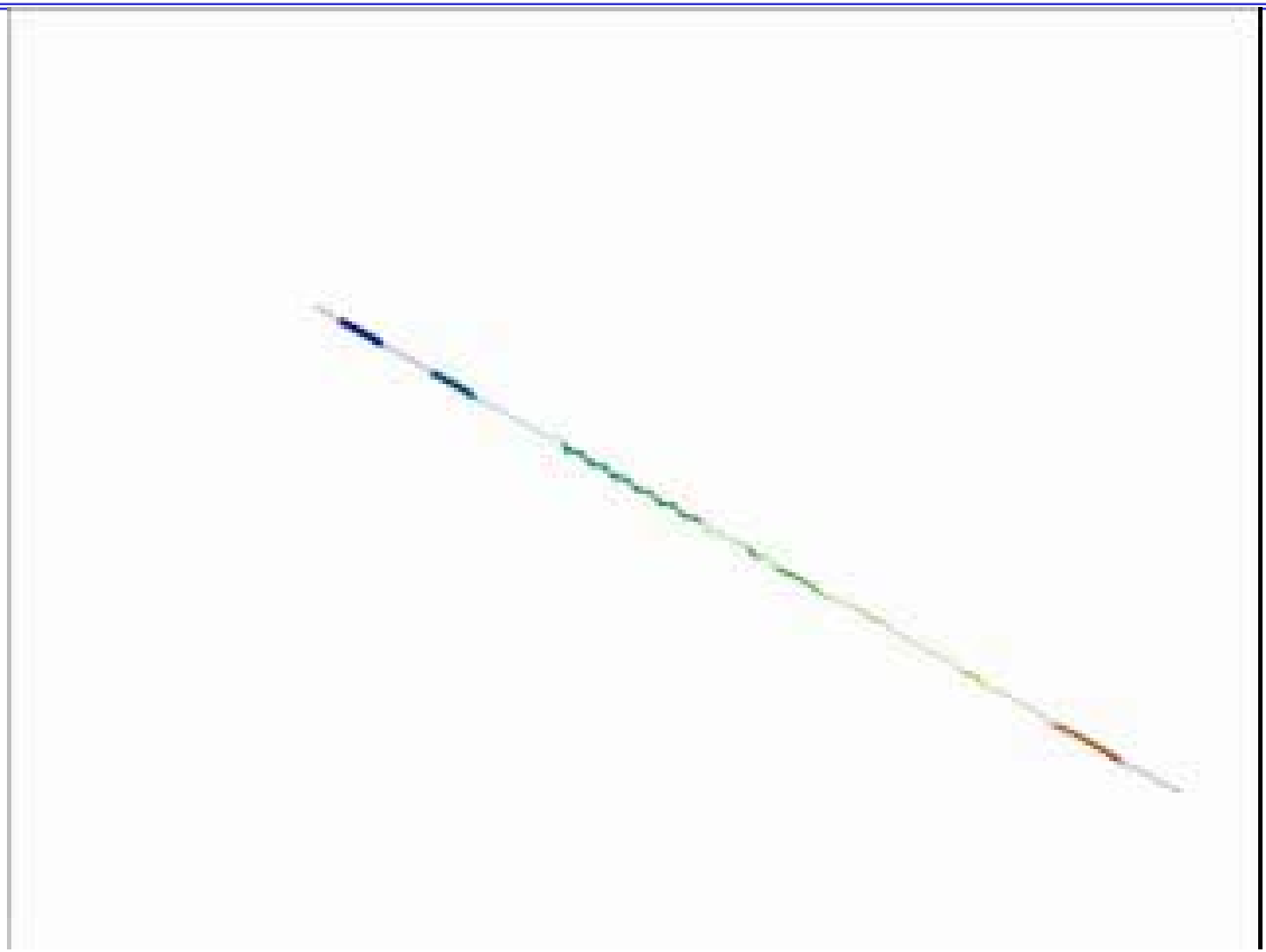
$P(r, \phi, \theta, \sigma, hb|Sep)$

Berücksichtigt Packung zweier Sekundärstruktur-elemente (Helix-Helix, Helix-Faltblatt, Faltblatt-Faltblatt) im Abstand *r*, mit dem Sequenzabstand *Sep*, mit den Verkippungswinkeln ϕ, θ, σ und einem Beitrag zu den Wasserstoffbrücken von *hb*

Algorithmus

- Bestimme 25 nächste Nachbarn für jede Teilsequenz
- Starte mit gestreckter Struktur
- Für 10000 Iterationen:
  - Wähle zufällig eine Teilsequenz *x´* aus *x*
  - Wähle zufällig *x´´* aus den Fragmenten für *x´*
  - Ersetze die Torsionswinkel in *x´* mit denen aus *x´´*
  - Wenn dadurch Atome überlappen, verwerfe Zug
  - Berechne Score
  - Akzeptiere oder verwerfe Zug gemäß MMC-Kriterium
  - Passe Temperatur an

ROSETTA



ROSETTA-Trajektorie von 1UBI

Verwendung von ROSETTA

- ROSETTA erzeugt eine ganze Anzahl von Strukturen
- Zur Vorbereitung muss ein Fragmentbibliothek mit rosettaFRAGMENT gebaut werden
- Diese Strukturen werden mit Hilfe der Scoring-Funktion bewertet: bester (negativster) Score = beste Struktur
- Je mehr Durchläufe (d.h. unabhängige Simulationen) durchgeführt werden, desto höher die Chance eine sehr gute Struktur zu finden
- Laufzeiten liegen im Minuten- bis Stundenbereich

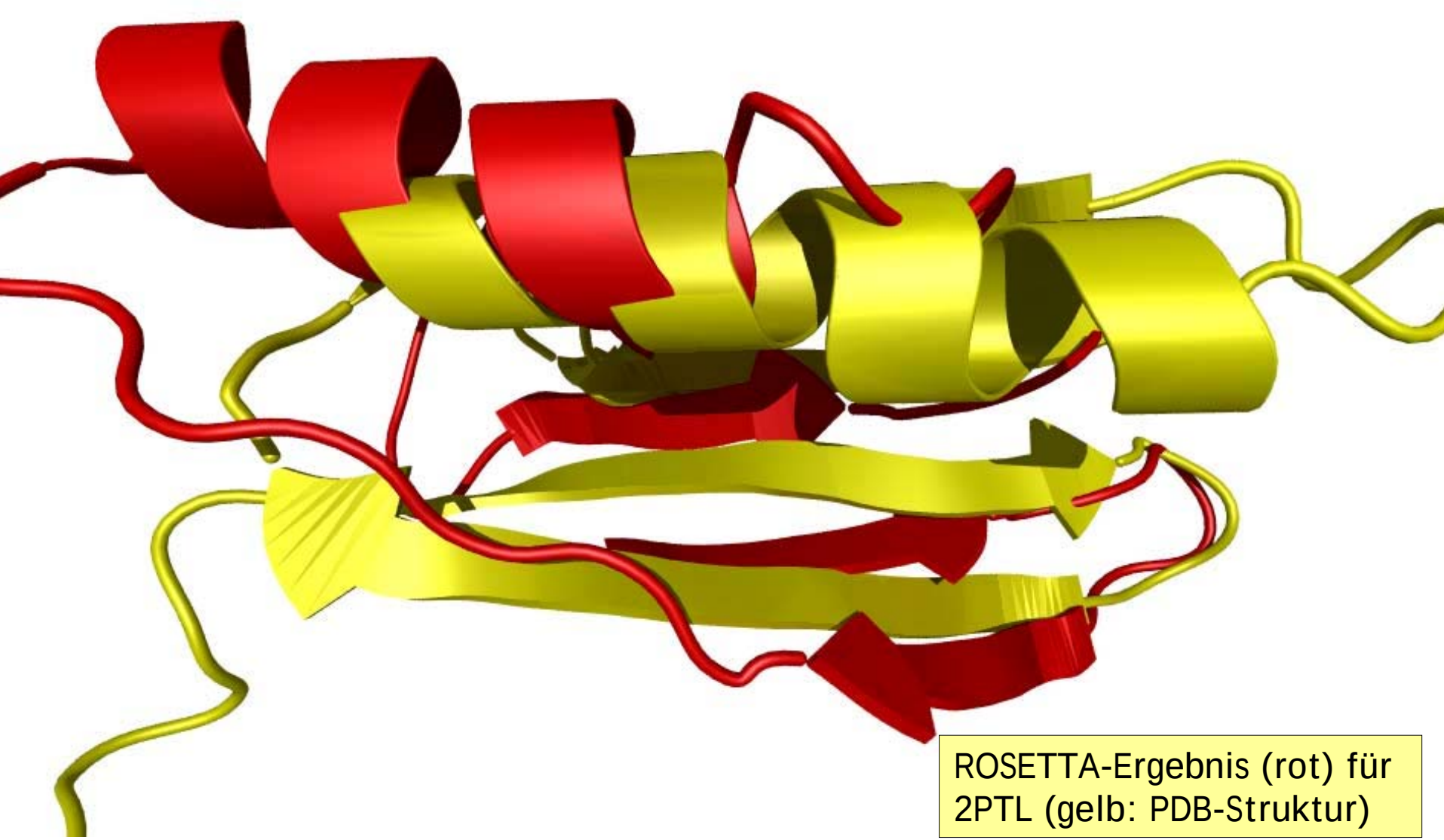
filename	score	env	pair	vdw	hs	ss
no_pdbfile_fail	-71.86	-27.02	-18.88	1.99	-1.69	-22.96
aa2PTL0001.pdb	-67.85	-24.50	-12.15	0.58	-4.43	-21.45
no_pdbfile_fail	-64.30	-27.23			-2.72	-17.13
aa2PTL0002.pdb	-67.77	-18.90	-9.64	2.06	2.63	-33.28
no_pdbfile_fail	-69.70	-31.18	-12.69	1.47	-1.76	-21.73
aa2PTL0003.pdb	-91.94	-18.14	-13.14	2.01	-8.71	-40.83
no_pdbfile_fail	-43.20	-11		83	-7.51	-11.75

Lauf fehlgeschlagen

Beste Energie

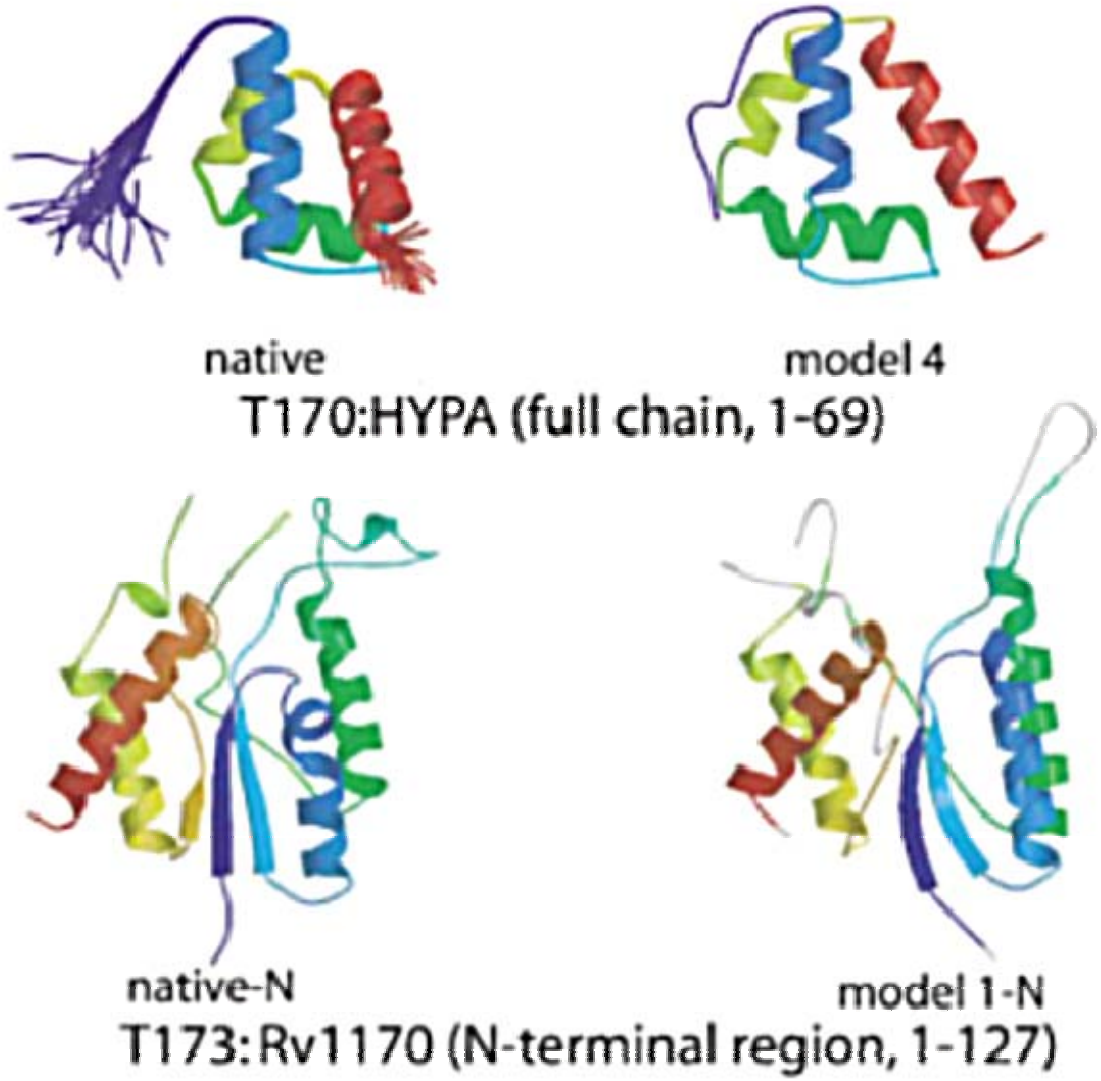
Beste Struktur

ROSETTA – Ergebnis



ROSETTA-Ergebnis (rot) für 2PTL (gelb: PDB-Struktur)

ROSETTA – Ergebnisse CASP5



MDS und ROSETTA

- Volle MD-Simulationen der Faltung von Proteinen sind *per se* möglich (folding@home, Blue Gene), aber bisher noch weitgehend erfolglos
- Lee *et al.* schlugen 2001 vor MDS und ROSETTA zu kombinieren
  - Fragmente zwingen Strukturen auf diskrete Torsionen, Strukturen sind nicht optimal
  - ROSETTAs Energiefunktion ist schnell, aber simpel
  - Methoden wie MM-PBSA (AMBER mit Berücksichtigung der Protein-Wasser-WW) sind aufwändig, aber liefern gute Energien

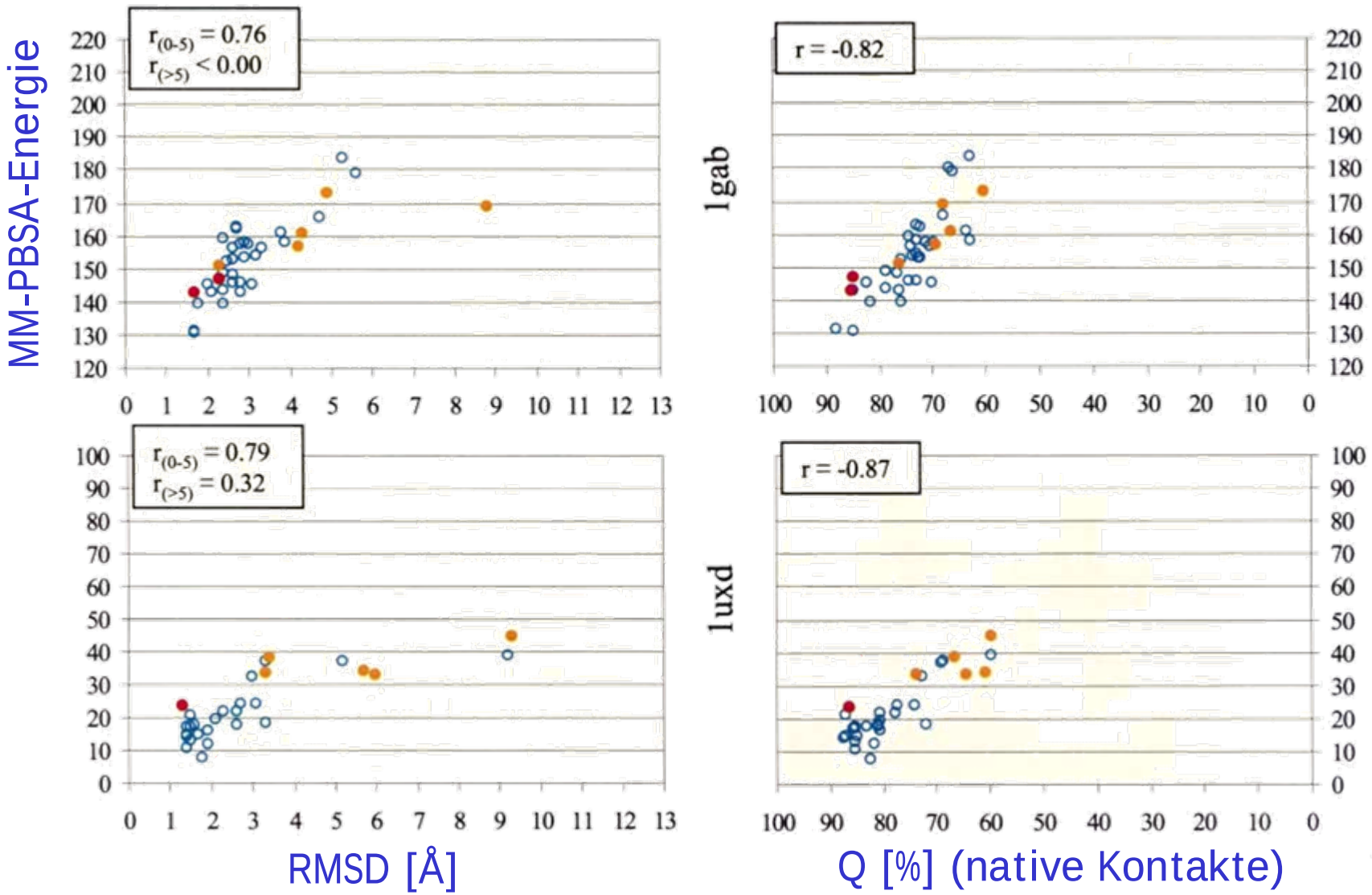
) ROSETTA zur Erzeugung guter **Startstrukturen** der MDS

) **MD-Simulation entspannt Strukturen**

) Verbesserte Energiefunktion liefert **besseres Ranking**

Lee et al., J. Mol. Biol. (2001), 313, 417

MDS und ROSETTA



(orange: Top5-ROSETTA-Strukturen, rot: nativer Zustand)

Lee et al., J. Mol. Biol. (2001), 313, 417

MDS und ROSETTA

Vorteile

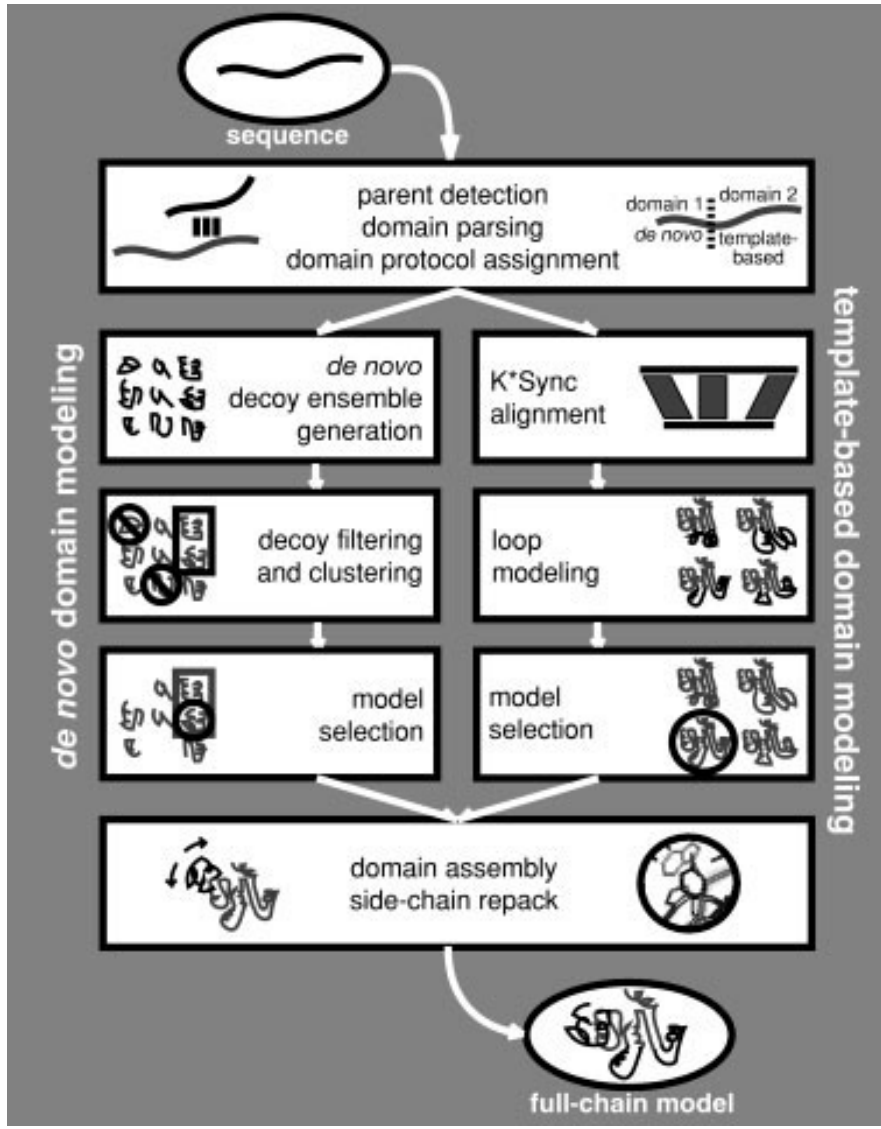
- MD-Simulationen verbessert die Qualität der ROSETTA-erzeugten Strukturen deutlich
- Energien aus MM-PBSA sind für das Ranking besser geeignet als ROSETTA-Score
- Kombination der Methoden verbessert Zuverlässigkeit der Vorhersage

Nachteile

- Immer noch extreme Rechenzeiten (Tage) allein für die Nachbearbeitung der ROSETTA-Strukturen

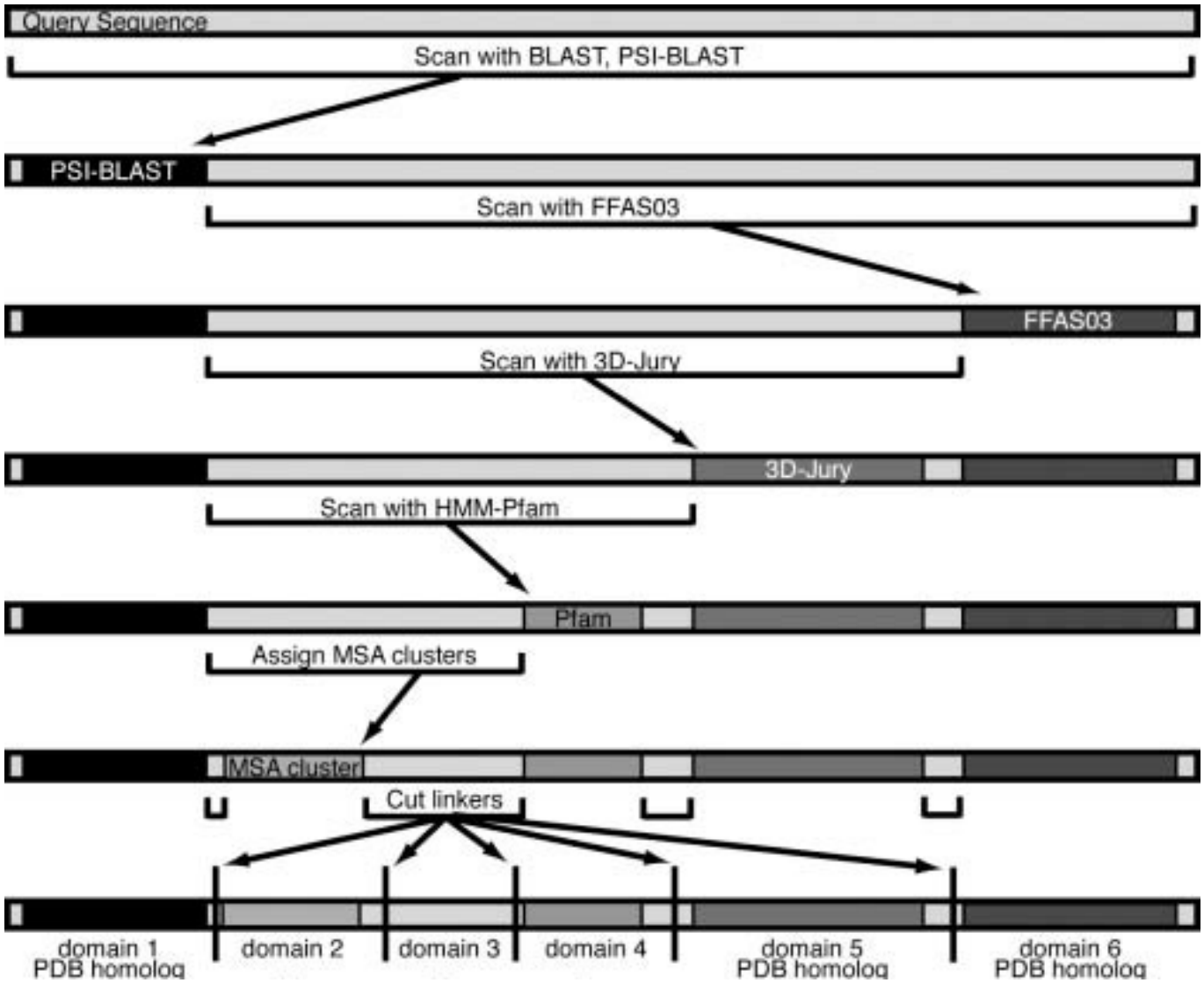
ROBETTA

- ROBETTA ist ein vollautomatischer Online-Server zur Proteinstruktur-Vorhersage
- ROBETTA kam erstmals bei CASP5 und dessen Äquivalent für vollautomatische Server, CAFASP3, zum Einsatz
- ROBETTA kombiniert
  - Domänenzerlegung
  - Threading für Domänen mit bekanntem Fold
  - Ab-initio-Vorhersage für Domänen ohne homologe Struktur



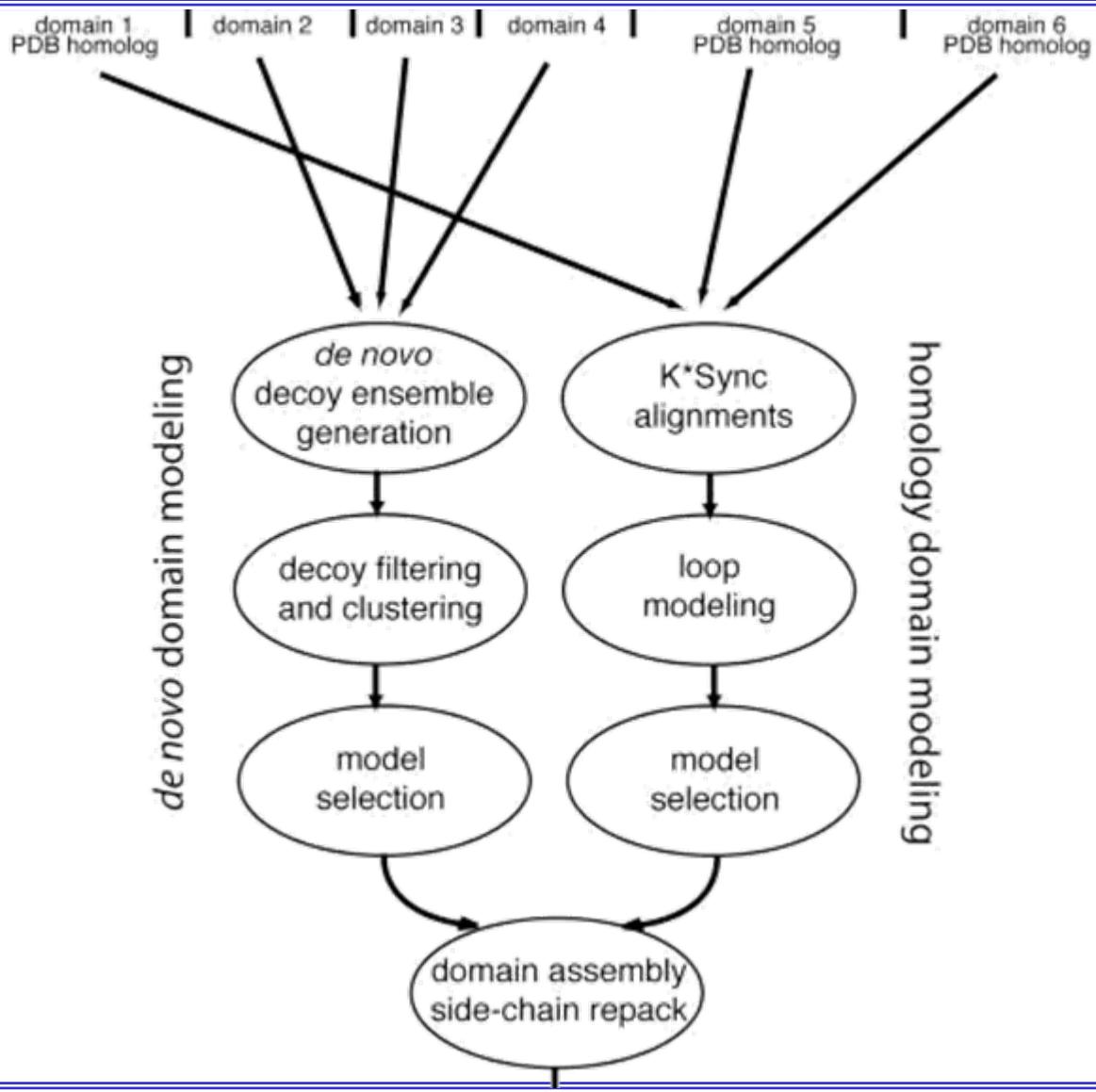
Chvian et al., Proteins (2003), 53, 524

ROBETTA – Überblick



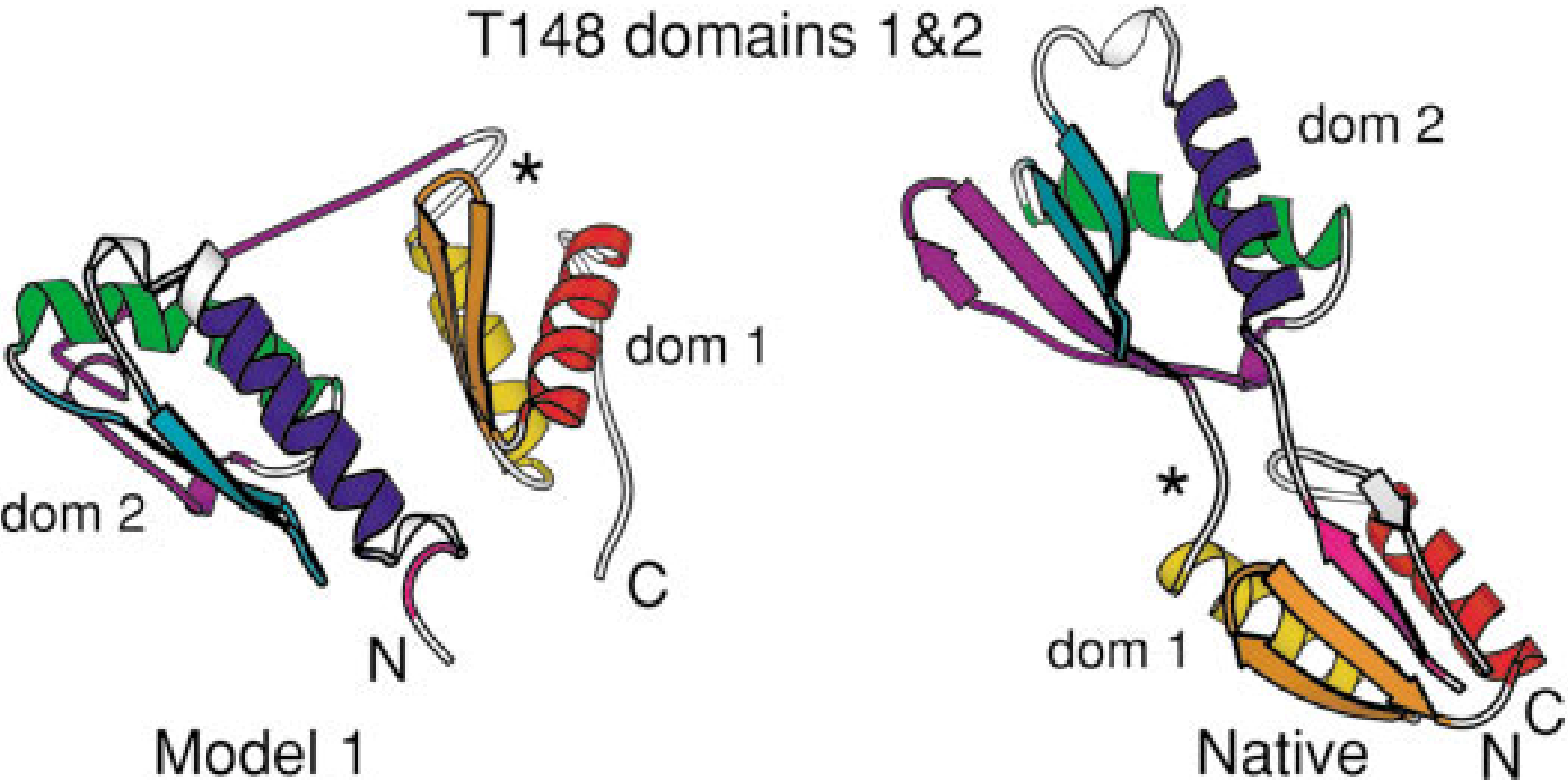
Kim et al., Nucleic Acids Res. (2004), 32, W526

ROBETTA – Überblick



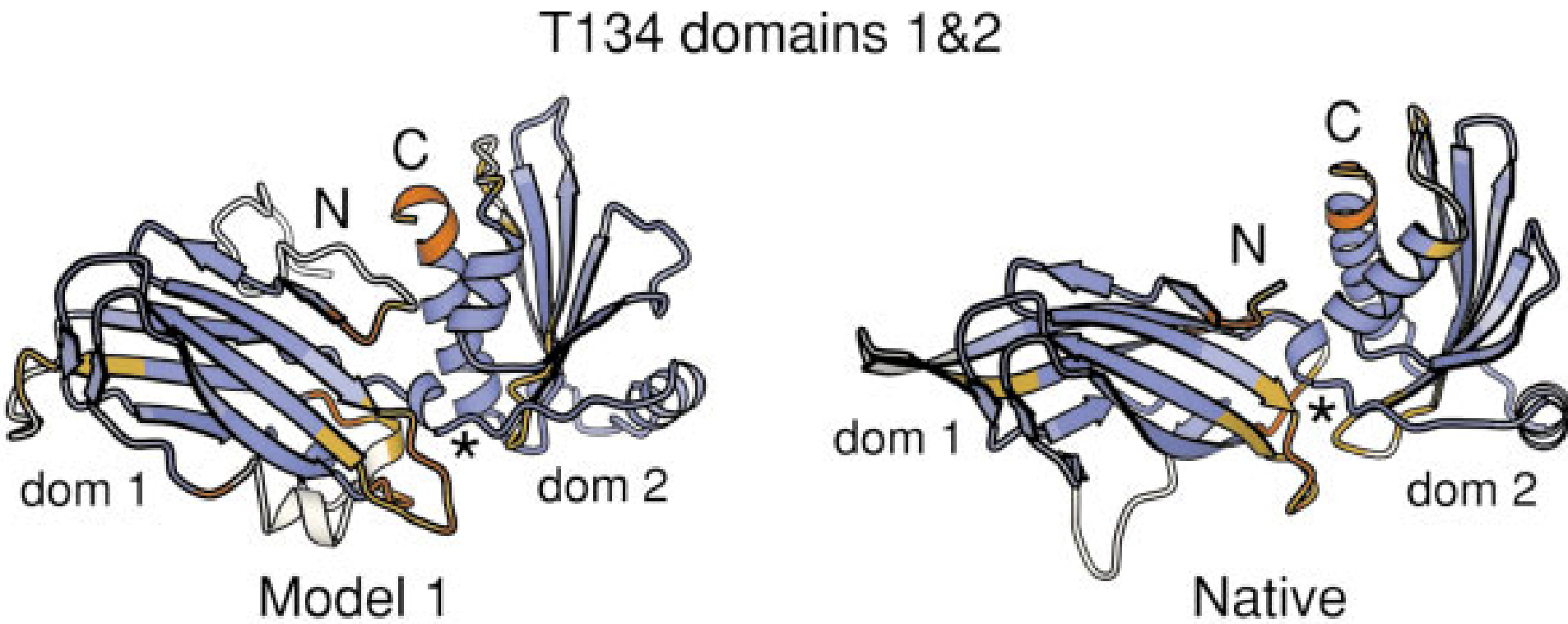
Kim et al., Nucleic Acids Res. (2004), 32, W526

ROBETTA – CASP5-Ergebnisse



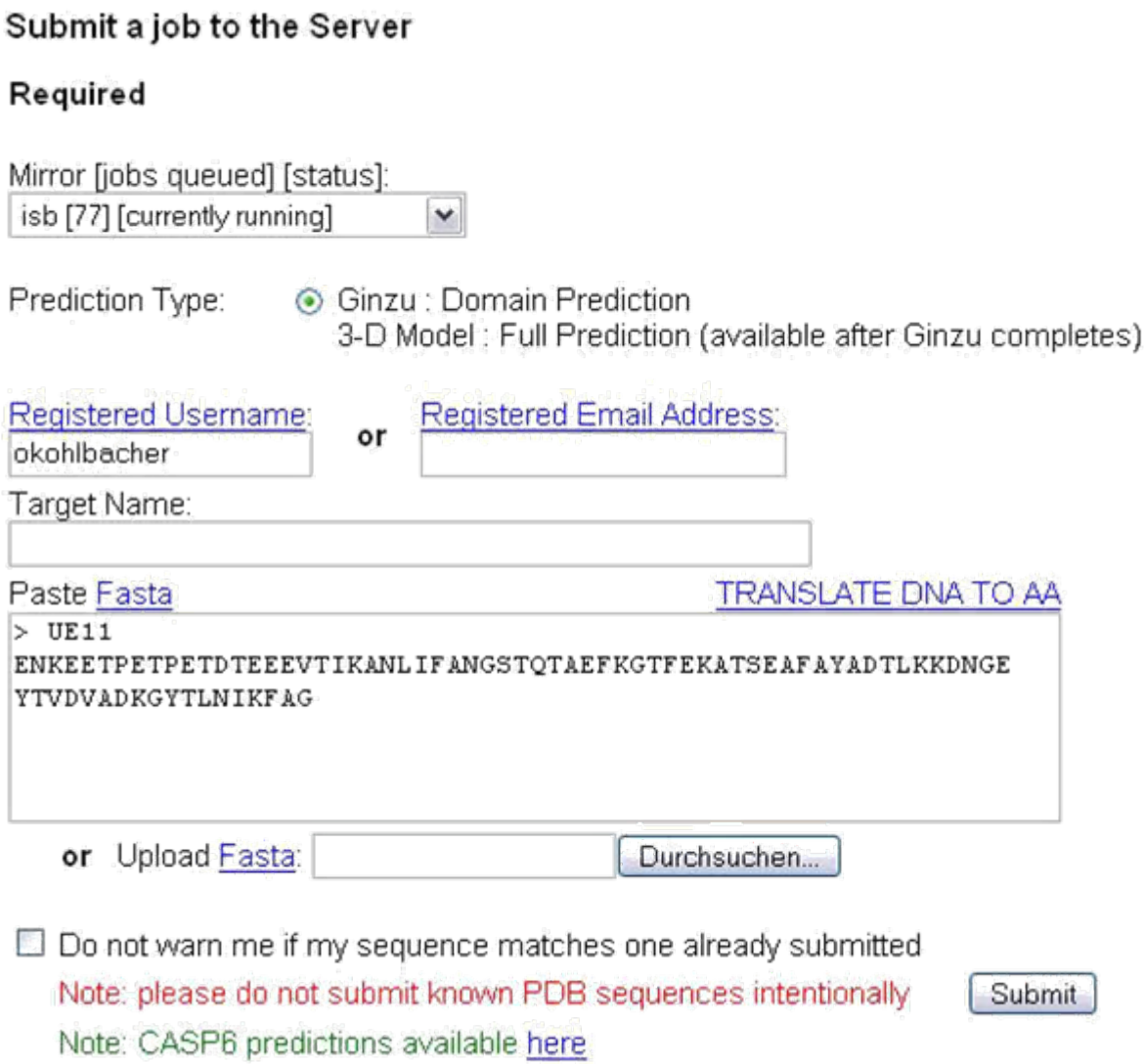
Chvian et al., Proteins (2003), 53, 524

ROBETTA – CASP5-Ergebnisse



Chvian et al., Proteins (2003), 53, 524

ROBETTA – Submission



ROBETTA – Status

Page 1 of 3

To conserve disk space, please remove your jobs after they complete by following these instructions.
*ETC is the Estimated Time of Completion in days. These values may increase when higher priority users submit jobs.

1 2 3

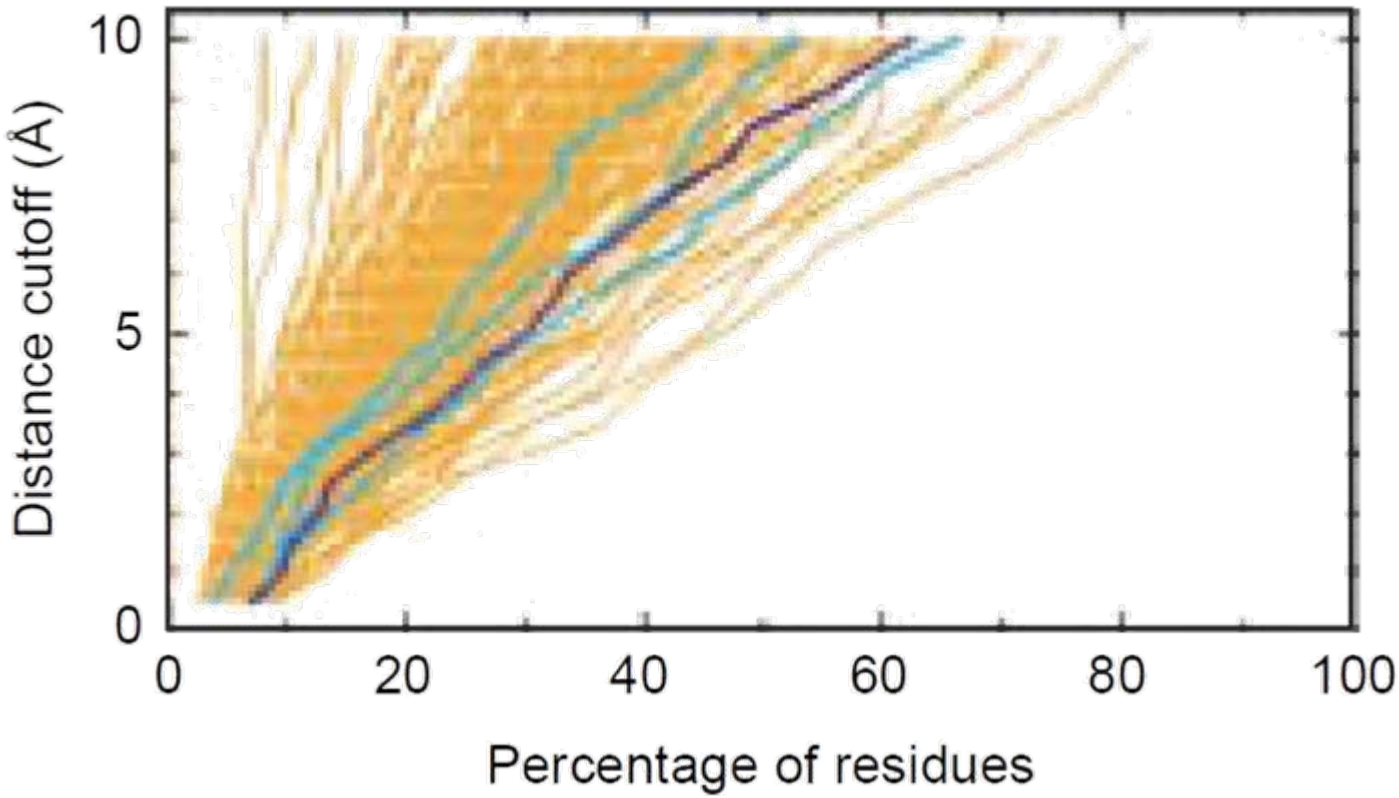
X	ID	Mirror	Status (*ETC)	Method	Username	Target	Length	Domain Prediction	Host
☐	2886	isb	Queued (26)	Ginzu	okhlbacher	Uebung11	78	--	p213.54.154.166.tisdip.tiscali.de
	2885	isb	Queued (25)	Ginzu	tomekj	a4.Mval	454	--	deimos.iu-bremen.de
	2884	lanl	Queued (3)	Ginzu	dekim	y2h_binary_seqs - Tcru004375AAA	258	--	dbfw.bchem.washington.edu
	2883	lanl	Queued (2)	Ginzu	dekim	y2h_binary_seqs - Tcru002553AAA	712	--	dbfw.bchem.washington.edu
	2882	lanl	Queued (2)	Ginzu	dekim	y2h_binary_seqs - Pfal008570AAA	314	--	dbfw.bchem.washington.edu
	2881	lanl	Queued (1)	Ginzu	dekim	y2h_binary_seqs - Pfal008464AAA	292	--	dbfw.bchem.washington.edu
	2880	lanl	Queued (1)	Ginzu	dekim	y2h_binary_seqs - Pfal006851AAA	484	--	dbfw.bchem.washington.edu
	2878	isb	Queued (3)	Ginzu	dhalphen	PsCOX2B&Ainv04	257	--	132.248.16.182
	2877	isb	Queued (24)	Ginzu	Torsten	Domain1	75	--	dilbert.pz.unibas.ch
	2876	lanl	Queued (5)	Ginzu	JARMENGAUD	SSOmod	166	--	194.254.179.136
	2875	isb	Queued (23)	Ginzu	rebecca	period2 protein	1000	--	61.3.96.230
	2874	isb	Queued (22)	Ginzu	vhalttun	NCAP_UUK	254	--	virusmac77.hi.helsinki.fi
	2873	isb	Queued (20)	Ginzu	subrama2	hcd47xDwosigpep	124	--	pcp01342817pcs.wilog501.pa.comcast.net
	2872	isb	Queued (19)	Ginzu	manmuell	PLQ1_277	388	--	vpn-global-dhcp3-163.ethz.ch
	2870	isb	Queued (18)	Ginzu	sudhar	mtkasb	438	--	ip24-250-177-222.bc.dl.cox.net

CASP – New Folds

- Topologie der Folds generell vorhersagbar
- Gesamtstruktur meist noch stark fehlerbehaftet
- Qualität der Methoden schwankt sehr stark
- Es gibt immer noch Folds die sich jeder Vorhersage entziehen
- CASP6 wird schwieriger: die Anzahl neuer Folds nimmt stetig ab (begrenzte Anzahl?)

Aloy, Stark, Hadley, Russell, Proteins, 2003, 53, 436

CASP – New Folds



- „Hubbard Plot“ gibt wieder wie viel Prozent der Reste für welchen Cutoff (RMSD) korrekt vorhergesagt wurden
- Einzelne Kurven entsprechen unterschiedlichen Methoden
- Für diese Zielstruktur aus CASP4 konnte keine der Methoden mehr als 60% für vernünftige Abweichungen korrekt vorhersagen

CASP – New Folds

- Drei Gruppen heben sich in CASP5 ab
  - **David Baker (ROSETTA)**
  - David Jones (FRAGFOLD)
  - David Shortle
- Erfolgreiche Algorithmen verwenden
  - Sekundärstrukturvorhersage (zu 80% PSIPRED)
  - Meistens Fragment-basierte Ansätze
  - ~ 50% der Ansätze verlangen manuelle Hilfe

Aloy, Stark, Hadley, Russell, Proteins, 2003, 53, 436

Literatur

Ab-Initio-Vorhersage

Dylan Chivian, Timothy Robertson, Richard Bonneau, David Baker, Ab initio Methods, In: Structural Bioinformatics, P. Bourne, H. Weissig (Hrsg.), Wiley, 2003

Patrick Aloy, Alexander Stark, Caroline Hadley, Robert Russell, Predictions Without Templates: New Folds, Secondary Structure, and Contacts, Proteins (2003), **53**, 436

Corey Hardin, Taras Pogorelov, Zaida Luthey-Schulten, Ab initio protein structure prediction, Curr. Opin. Struct. Biolo. (2002), 12:176

Publikationen zu den einzelnen Methoden: siehe Website

Proteinstruktur und -modellierung

Dr. Stephan Steigele

10. Modellierung von Proteinen III

BioInf / Universität Leipzig

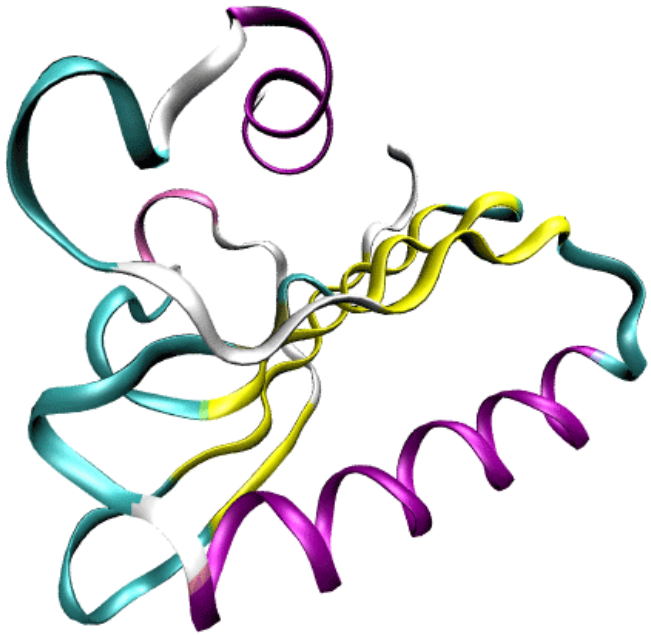
Überblick

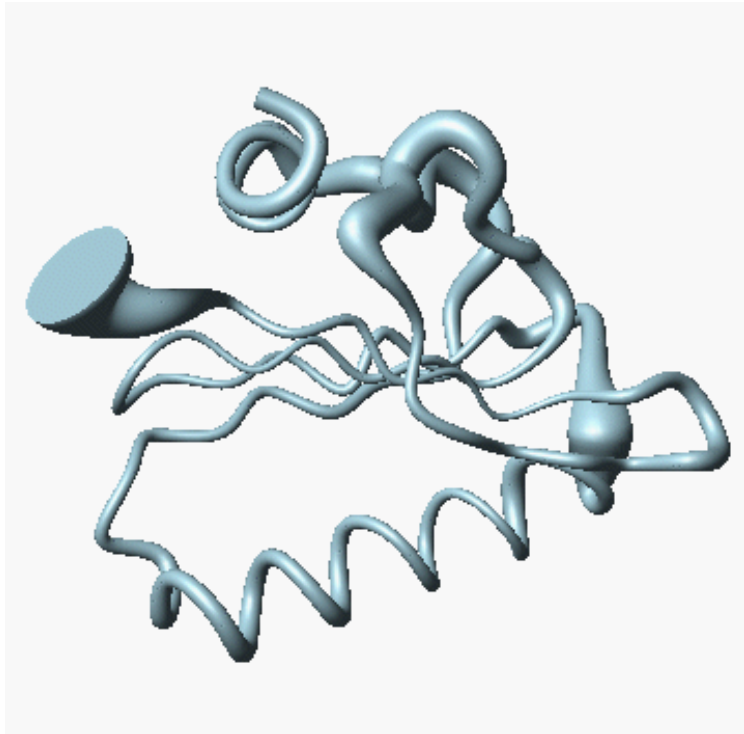
Molekulardynamik

- Bewegungsgleichungen
- Ein einfaches Modellsystem
- Algorithmen zur Integration der Bewegungsgleichungen
  - Verlet
  - Velocity-Verlet
  - Leap-Frog
- Ensembles und Thermodynamik
- Simulationen mit periodischen Randbedingungen
- Beispiele und Anwendungen

Dynamik von Proteinen

- Proteine sind flexibel
- Flexibilität ist essenziell für die Funktion

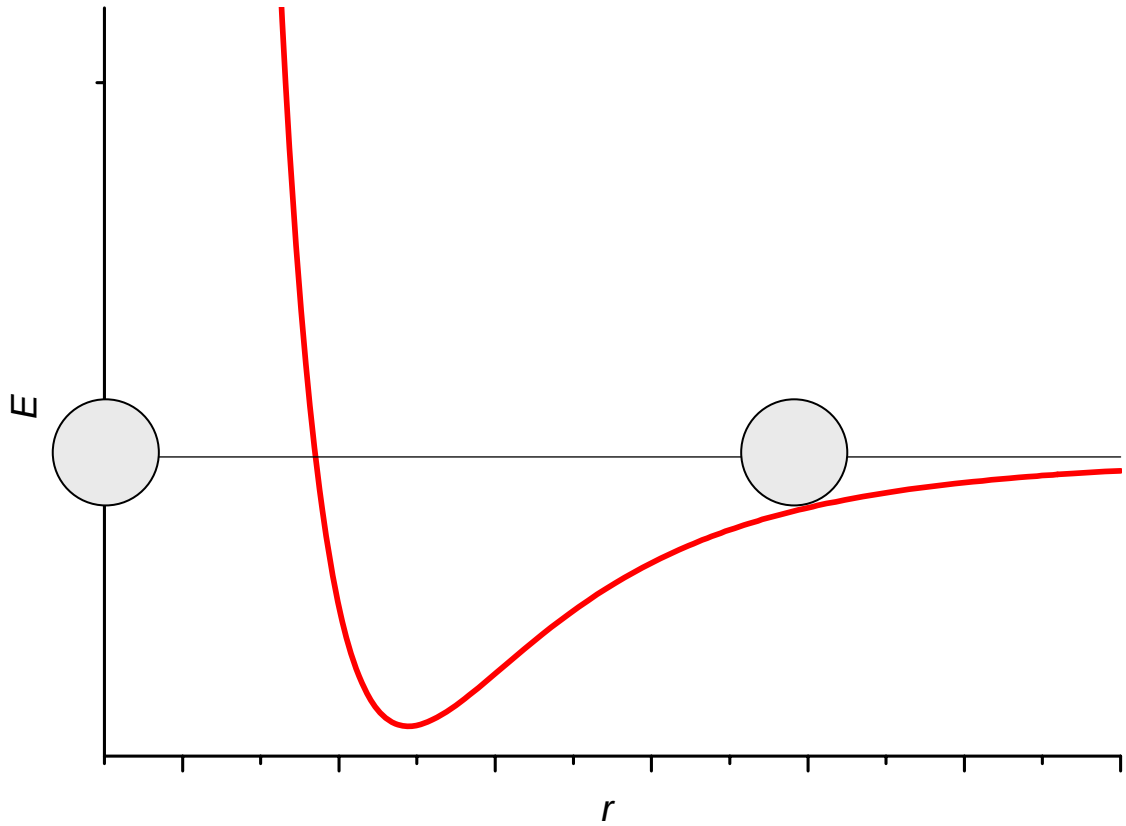




Wie simuliert man das dynamische Verhalten?

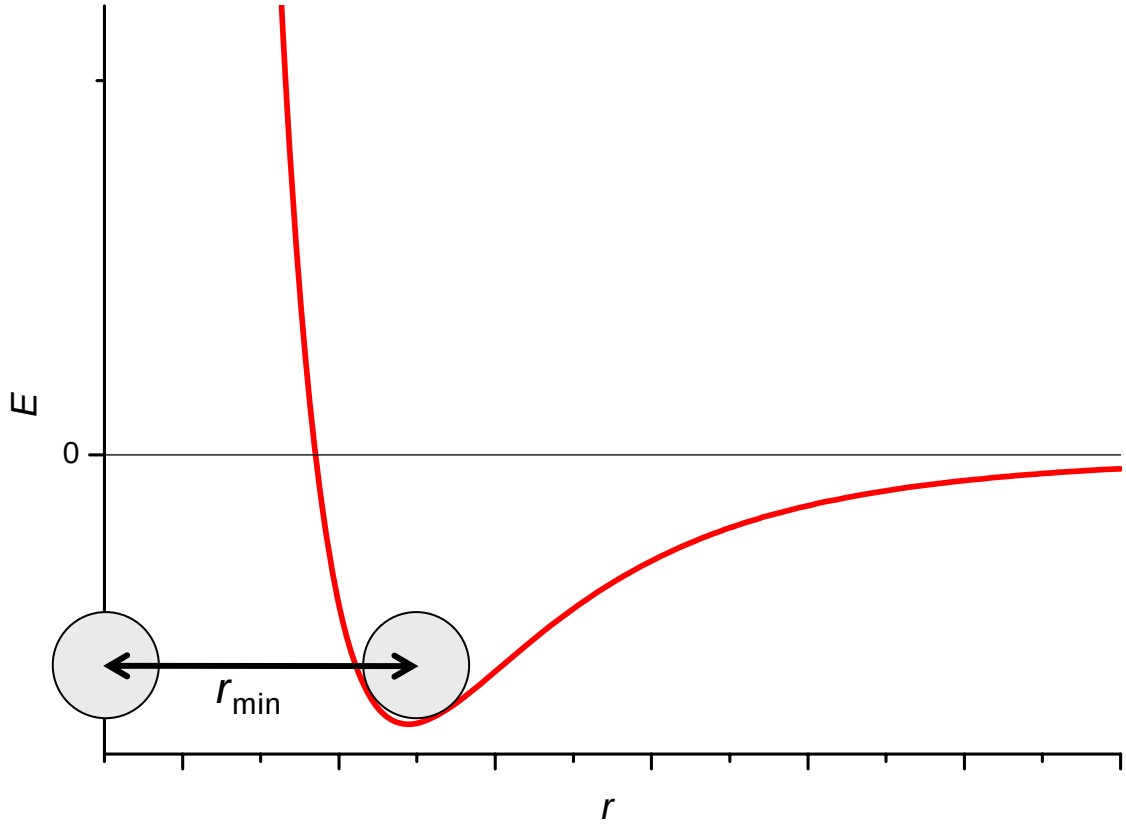
Modellsystem

- Zwei Argon-Atome an Positionen $\mathbf{r}_1$ und $\mathbf{r}_2$
- Wir kennen die Energiefunktion $E(R)$



Modellsystem

- System versucht minimale Energie anzunehmen
- **Attraktive Wechselwirkung** = anziehende Kraft!



Grundgrößen

Bewegung eines Teilchens wird beschrieben durch

- Ort **r**
- Geschwindigkeit **v**
- Beschleunigung **a**

in Abhängigkeit von der Zeit *t*.

Dabei gilt:

$$\mathbf{v} = \frac{\partial}{\partial t}\mathbf{r} = \dot{\mathbf{r}}$$

$$\mathbf{a} = \frac{\partial}{\partial t}\mathbf{v} = \frac{\partial^2}{\partial t^2}\mathbf{r} = \ddot{\mathbf{r}}$$

Newtonsche Axiome

1. **Newtonsches Axiom (N1) – Trägheitsgesetz**

   Jeder Körper verharrt im Zustand der Ruhe oder der gleichförmigen, geradlinigen Bewegung, solange er nicht durch äußere Kräfte gezwungen wird, seinen Bewegungszustand zu ändern.

2. **Newtonsches Axiom (N2) – Dynam. Grundgesetz**

   Die Bewegung eines Körpers ändert sich proportional zur einwirkenden Kraft, wobei die Masse der Proportionalitätsfaktor ist: $\mathbf{F} = m\,\mathbf{a}$

3. **Newtonsches Axiom (N3) – Reaktionsgesetz**

   Actio aequat reactionem. Die von zwei Körpern aufeinander ausgeübten Kräfte sind stets gleich groß und entgegengesetzt.

Kraft und Beschleunigung

- **N1:** ein Körper wird aus der Ruhe gebracht (beschleunigt), wenn auf ihn eine Kraft wirkt
- Kraft bewirkt Beschleunigung, also Änderung der Geschwindigkeit (**N2**).
- Jedes Teilchen *i* hat eine Masse m_i
- Mit **N2** gilt dann für seine Beschleunigung

$$\mathbf{a}_i = \mathbf{F}_i \; / \; m_i$$

- **Kenntnis der Kräfte $\mathbf{F}_i$ ermöglicht** also die **Berechnung der Beschleunigungen.**

Vom Kraftfeld zur Kraft

- Wechselwirkungen zwischen Atomen resultieren in der Regel in Wirkungen, d.h. in Kräften zwischen den Atomen
- **Kraft ist** dabei gerade der **negative Gradient der Energie**

$$\mathbf{F}(\mathbf{r}) = -\nabla E(\mathbf{r}) = -\,\mathrm{grad}\, E(\mathbf{r})$$

- Für dreidimensionale kartesische Koordinaten ist der Gradientenoperator **Nabla** definiert als

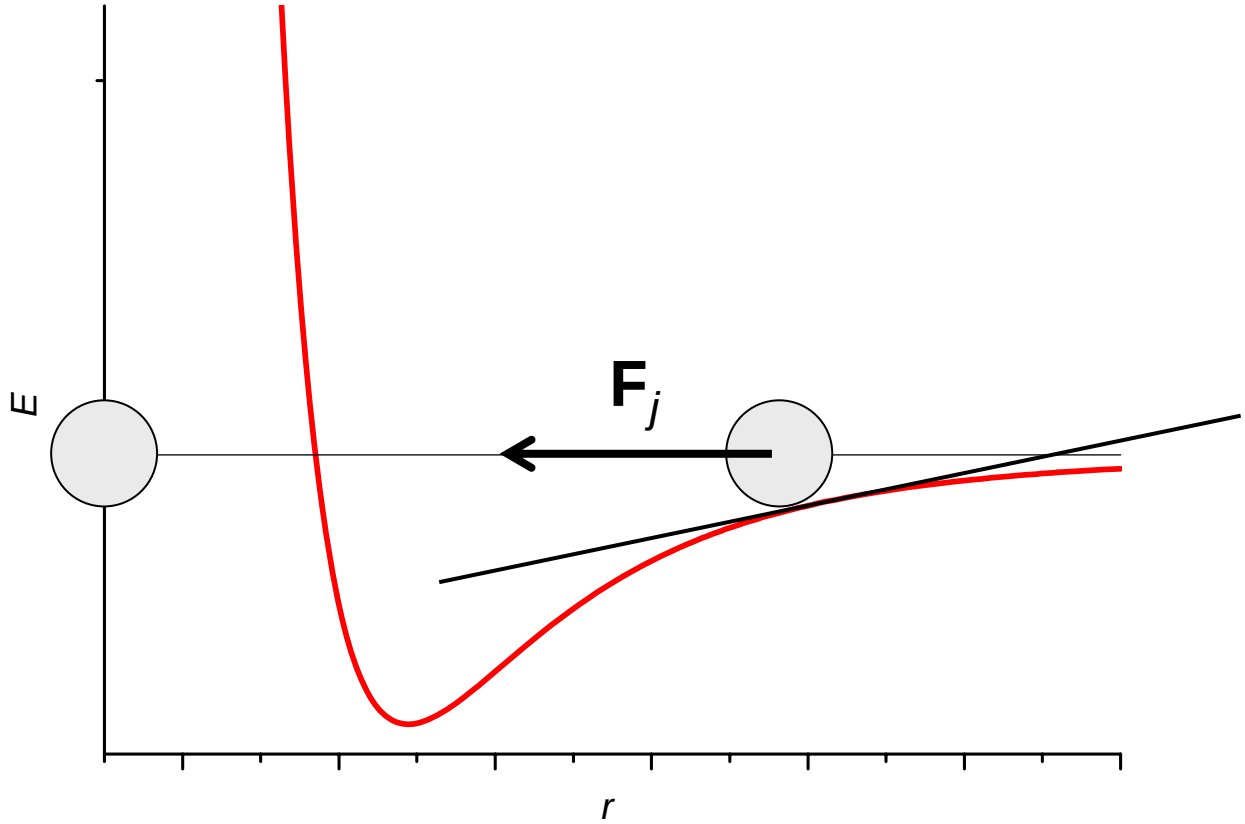
$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right) = \frac{\partial}{\partial \mathbf{r}}$$

- Damit kann man aus jeder **differenzierbaren Energiefunktion** *E* die auf jedes Atom *i* wirkende **Kraft** $\mathbf{F}_i$ berechnen.

Modellsystem

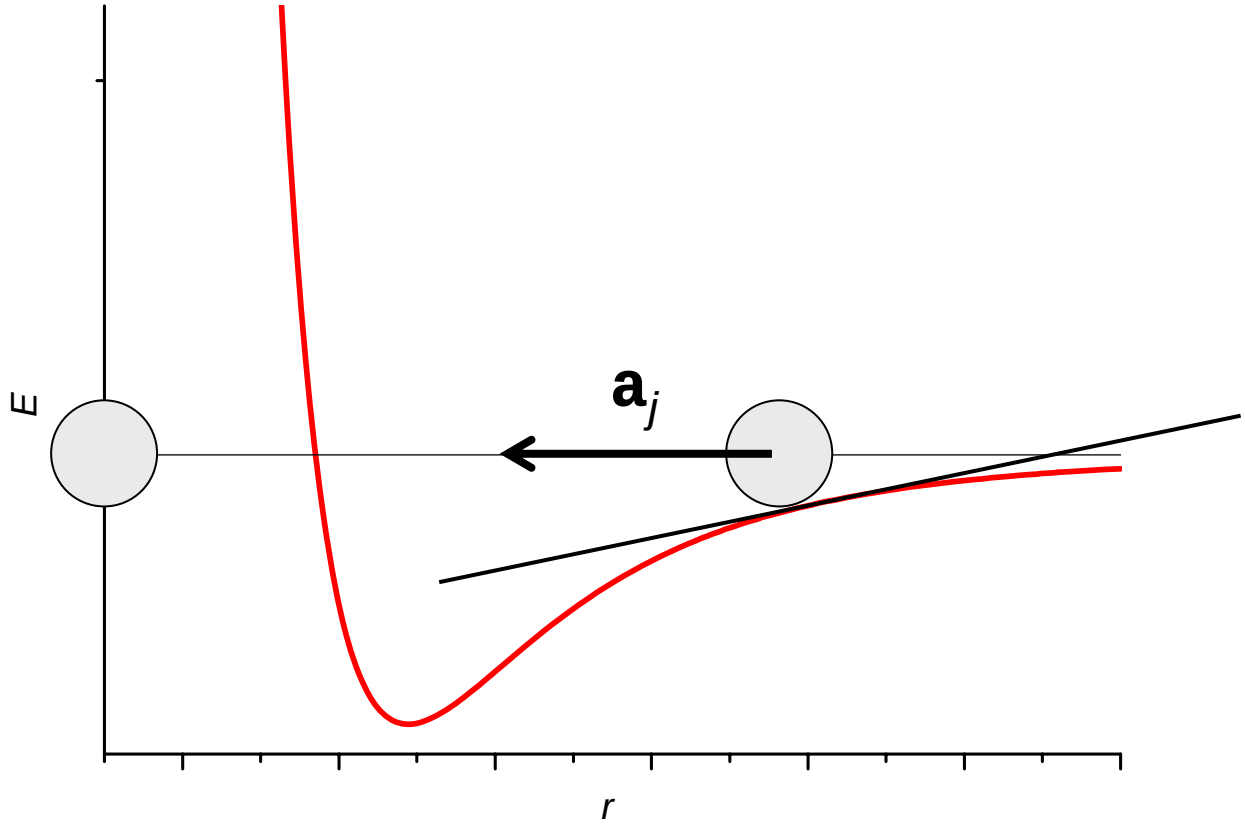
- Kraft entspricht der Steigung von *E*

$$F_j = -\mathrm{r}E(\mathrm{r}) = \partial / \partial\, x_j\, E(r)$$



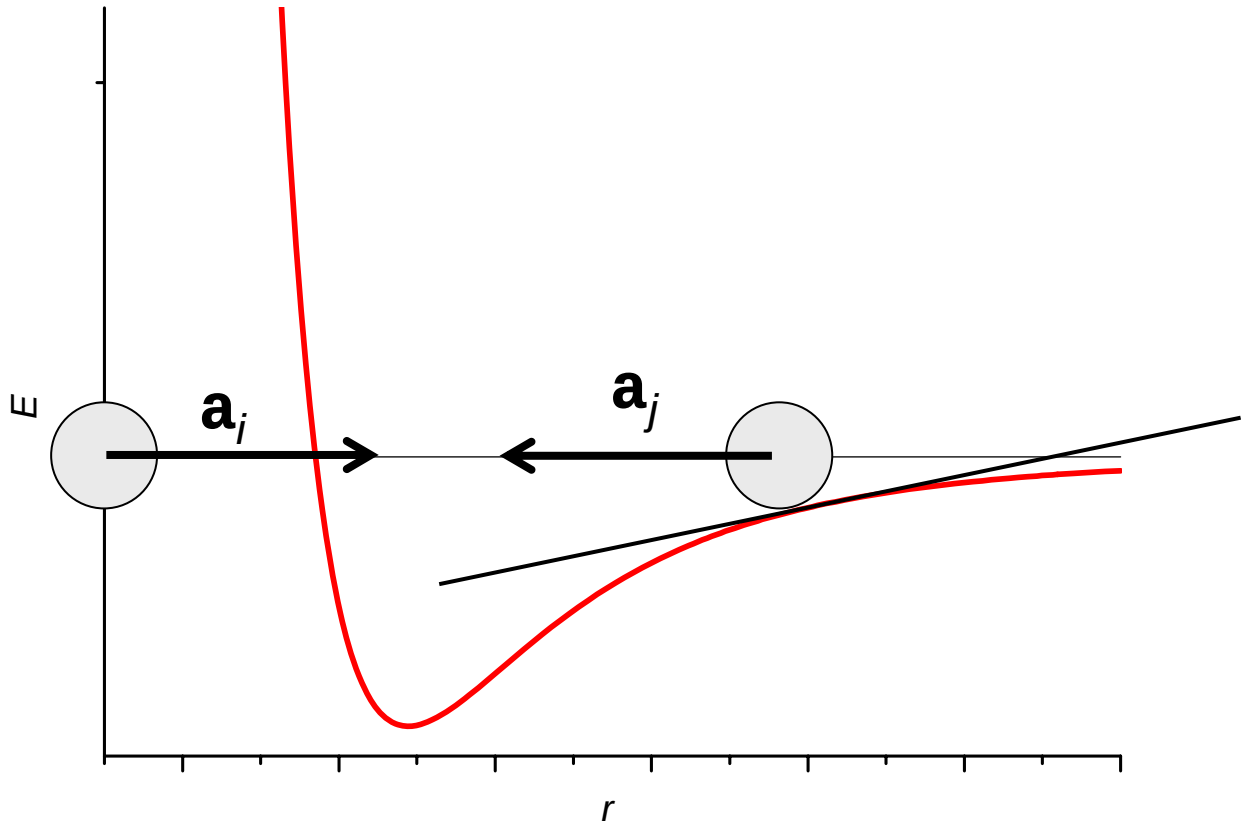
Modellsystem

- Liegt ein Teilchen also an einer Stelle mit r$E \neq 0$, wird es gemäß **N2** beschleunigt: $\mathbf{a}_j = \mathbf{F}_j \,/\, m_j$



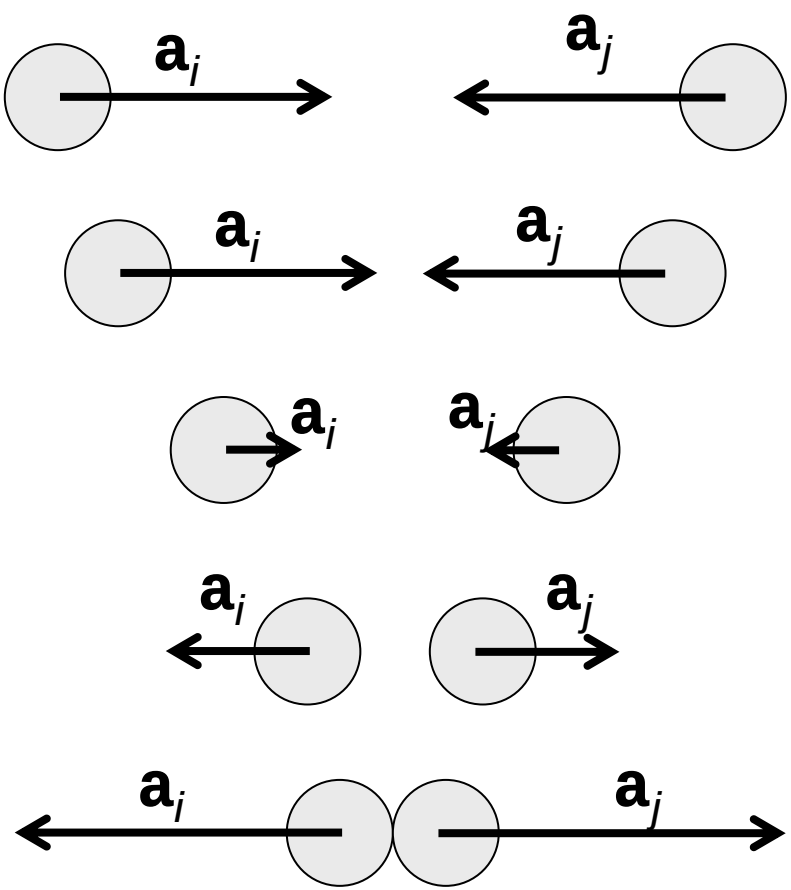
Modellsystem

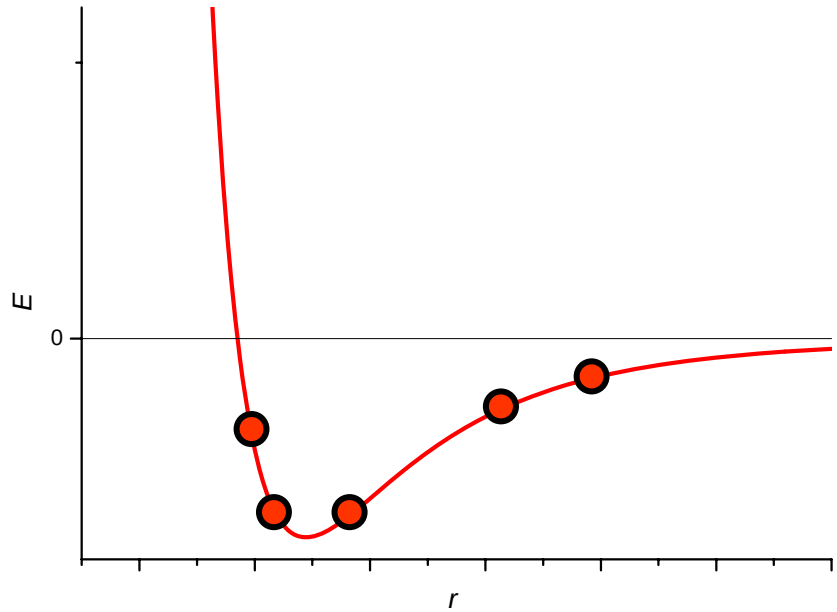
- Beschleunigung führt zu einer Bewegung der Teilchen *i* und *j* aufeinander zu.



Modellsystem

- Teilchen bewegen sich aufeinander zu, bis über das Minimum von *E* hinweg (Trägheit), dann wieder zurück.





Modellsystem

- In unserem einfachen Modellsystem erhält man für die Beschleunigungen:

$$\mathbf{a}_i = \frac{\mathbf{F}_i(\mathbf{r}_i)}{m_i} = \frac{\partial}{\partial \mathbf{r}_i}\mathbf{E}_i(\mathbf{r}_i)/m_i = \frac{\partial^2}{\partial t^2}\mathbf{r}_i = \ddot{\mathbf{r}}_i$$

- Dieses **System von Differenzialgleichungen** beschreibt die Abhängigkeit der Atompositionen $\mathbf{r}_i$ von der Zeit t.
- Lösen dieser **Bewegungsgleichungen** liefert dann die zeitliche Variation der Position, die gesuchte Dynamik!

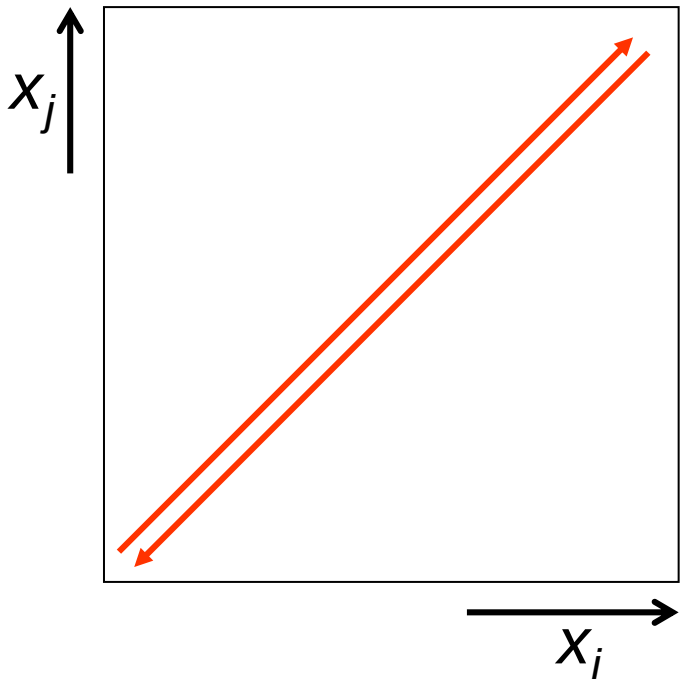
Lösung der Bewegungsgleichungen

- Lösen der Bewegungsgleichungen erfolgt durch Integration nach der Zeit.
- **Beispiel:** konstante Beschleunigung

$$\begin{aligned}
\mathbf{v}(t) &= \int \mathbf{a} dt = \mathbf{a} \int dt = \mathbf{a}t \\
\mathbf{r}(t) &= \int \mathbf{v}(t) dt \\
&= \int \left(\int \mathbf{a}(t) dt \right) dt = \frac{1}{2} \mathbf{a} t^2
\end{aligned}$$

Von der Kraft zur Dynamik

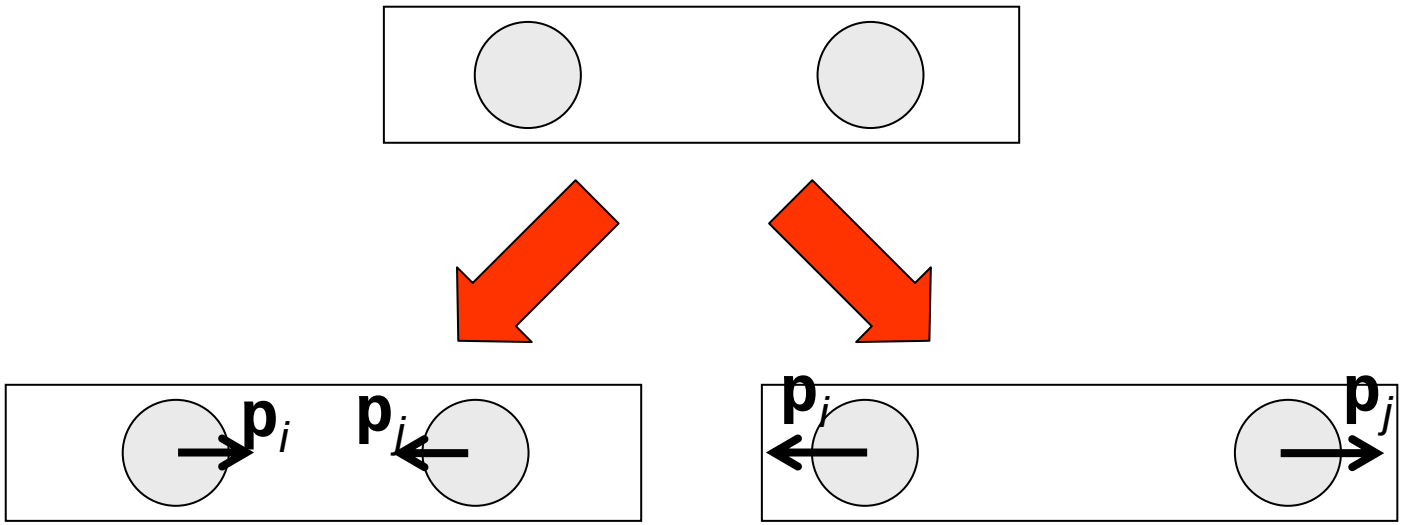
- Jede Konformation eines Systems mit *N* Teilchen entspricht gerade einem Punkt im 3*N*-dimensionalen Konformationsraum des Systems
- Lösen der Bewegungsgleichungen resultiert in den **Koordinaten als Funktion der Zeit**
- Entspricht **Kurve durch den Konformationsraum**
- Eine solche Kurve nennt man **Trajektorie**



Trajektorie für unser zweiatomiges Modellsystem

Phasenraum

- Ein Punkt im Konformationsraum ist nicht eindeutig um den Zustand eines Systems zu beschreiben
- Vollständig beschrieben ist das System, wenn man zusätzlich für jedes Teilchen die Geschwindigkeit **v** oder alternativ den Impuls **p** = *m***v** kennt
- Orts- und Impulskoordinaten spannen den (6N-dimensionalen) **Phasenraum** auf



MD-Simulation

- Simulation der Dynamik eines molekularen Systems basierend auf einem Kraftfeld und der Lösung der Bewegungsgleichungen nennt man

  Molekulardynamik-Simulation (MDS)

- Resultat der MDS ist eine Trajektorie und zugehörige Energien
- Trajektorie beschreibt die Bewegung des Systems in Abhängigkeit von der Zeit
- Durch Simulation entsprechend langer Zeiträume können auch langwierige Prozesse (z.B. Faltung) simuliert werden

MDS mit AMBER

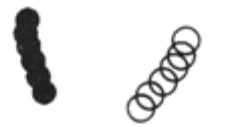
- AMBER ist ein differenzierbares Kraftfeld für Molekülmechanik
- Molekulardynamik-Simulation ist damit also möglich
- Erforderlich:
  - Ableitungen der einzelnen Terme

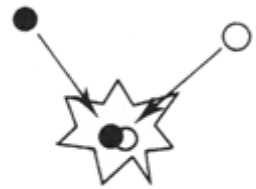
    → **Bewegungsgleichungen**
  - Effiziente Lösung der Bewegungsgleichungen

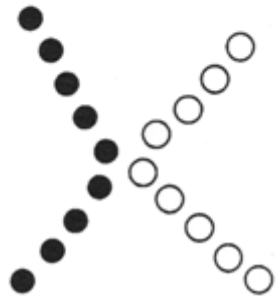
    → **Trajektorie**

Wahl des Zeitschritts

- Wahl der Größe des Zeitschritts kritisch
- Zu klein: Simulation dauert zu lange
- Zu groß: **Simulation „explodiert“**: Energien steigen extrem stark an, Teilchen werden extrem beschleunigt
- Größenordnung: 10^{-14} bis 10^{-16} s

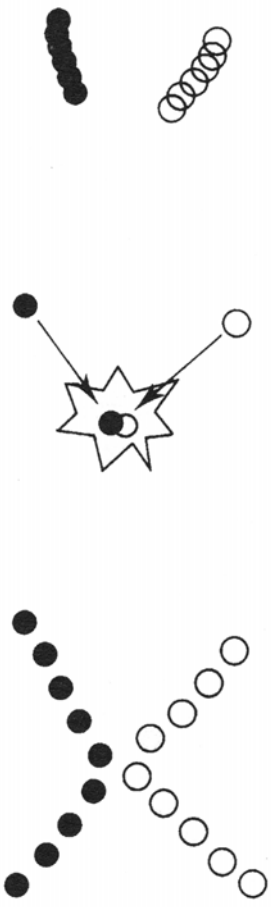




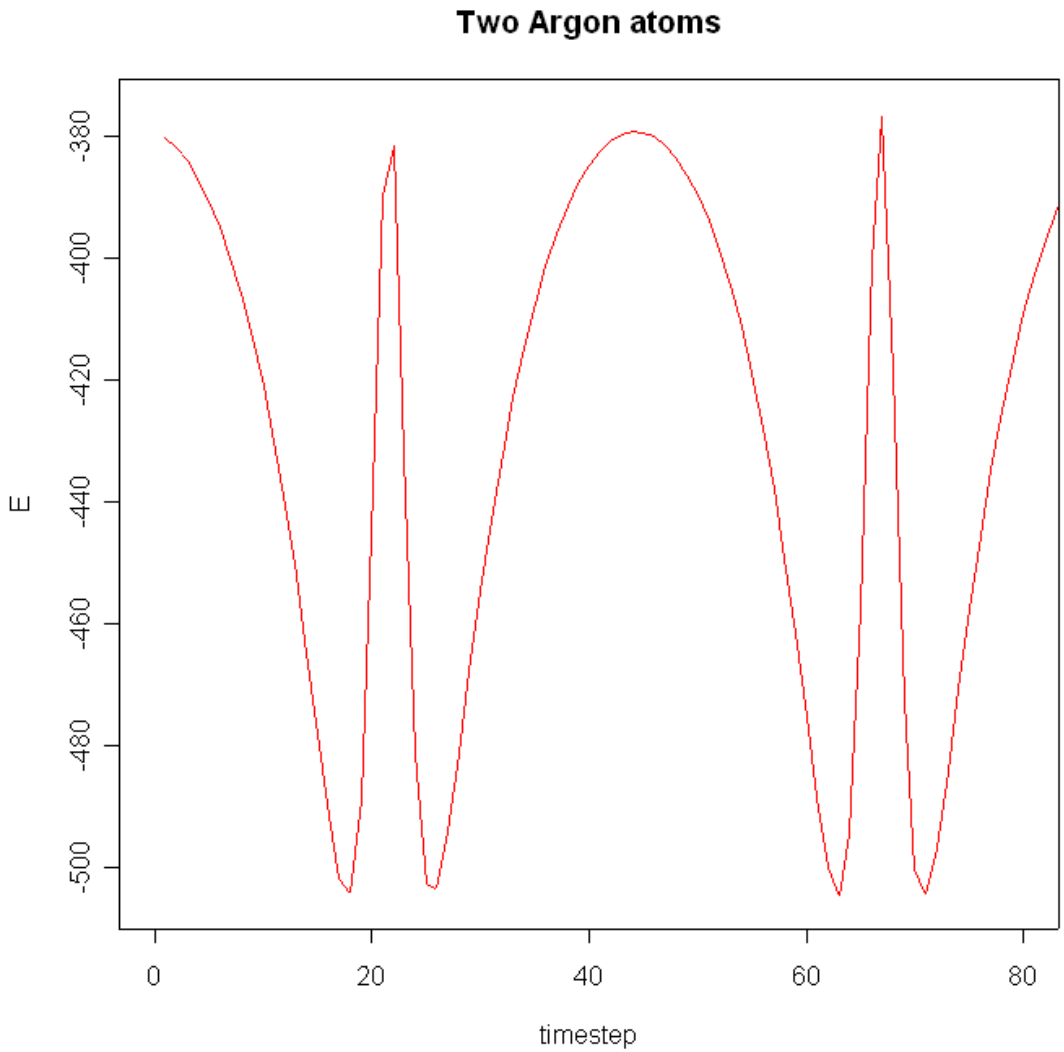


„Explosionen“

- Atombewegungen durchbrechen „Energiebarriere“
- Grund: Zu großer Zeitschritt nicht vereinbar mit linearer Approximation
- Resultierende Bewegungen pro Zeitschritt zu groß) stark überlappende Atome
- Hartes repulsives Potential führt zu extremen Beschleunigungen der Teilchen) Explosion



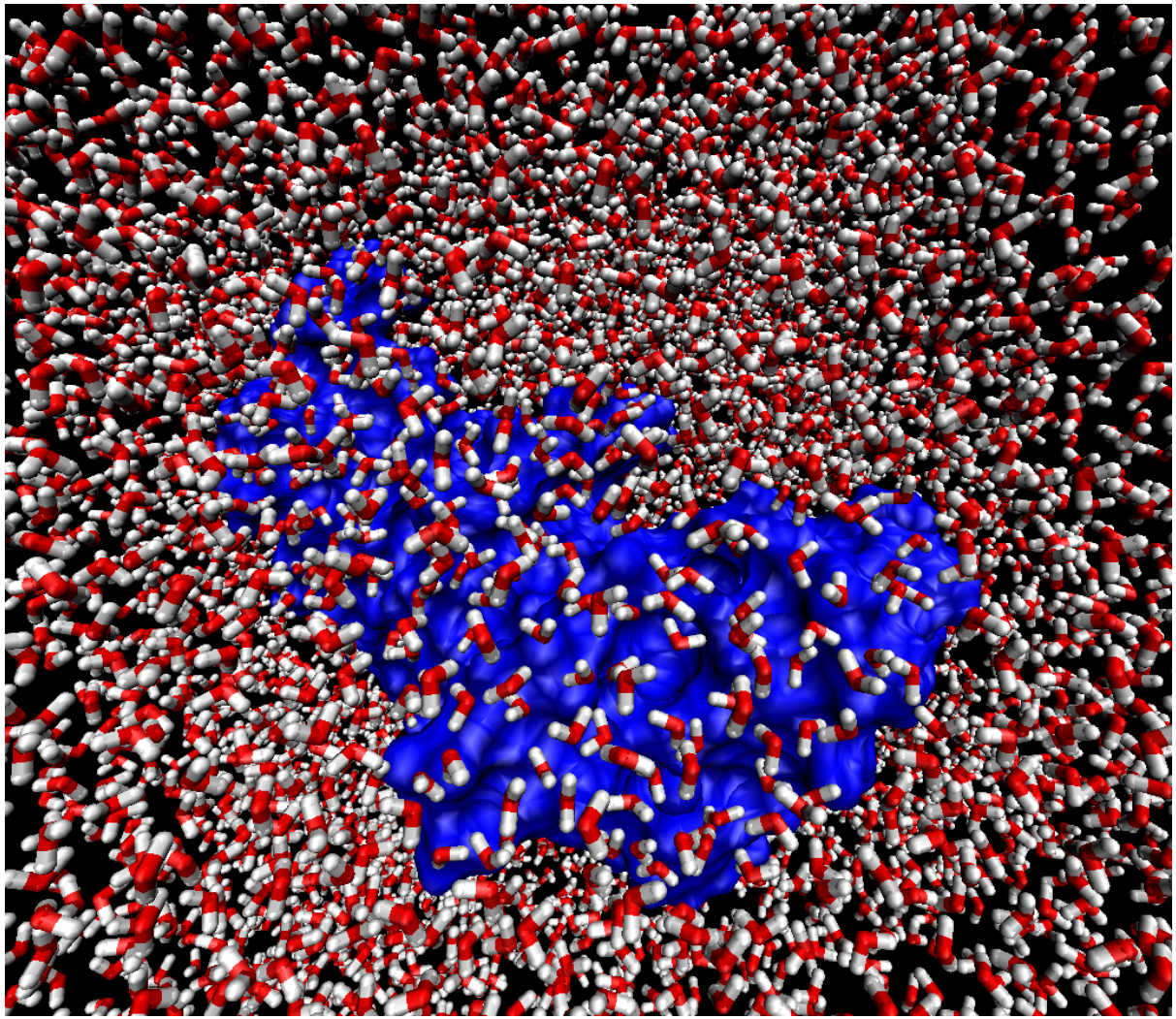
Typischer Energieverlauf



Explizites Wasser

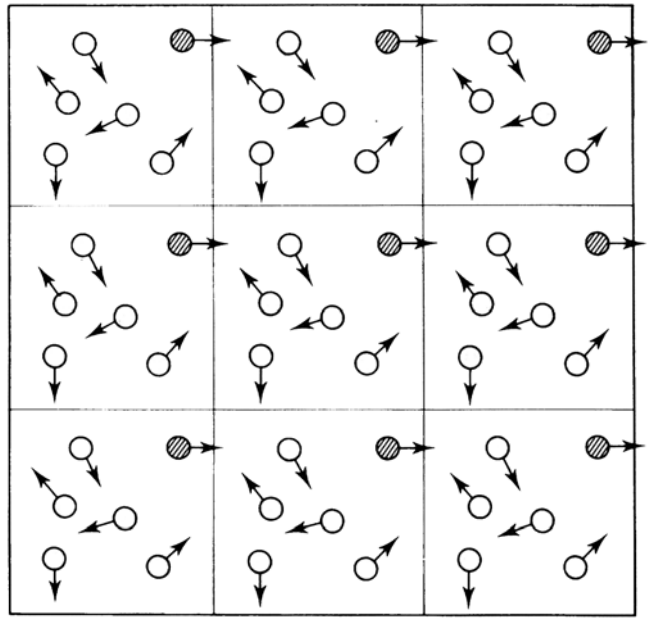
- Biologische Prozesse finden im Wasser statt
- Simulation von Protein ohne umgebendes Wasser reproduziert natürliches Verhalten schlecht
- Wasser genauso mit Hilfe der Molkülmechanik modellierbar
- Notwendig: Einbindung **explizit modellierter Wassermoleküle**

Explizites Wasser

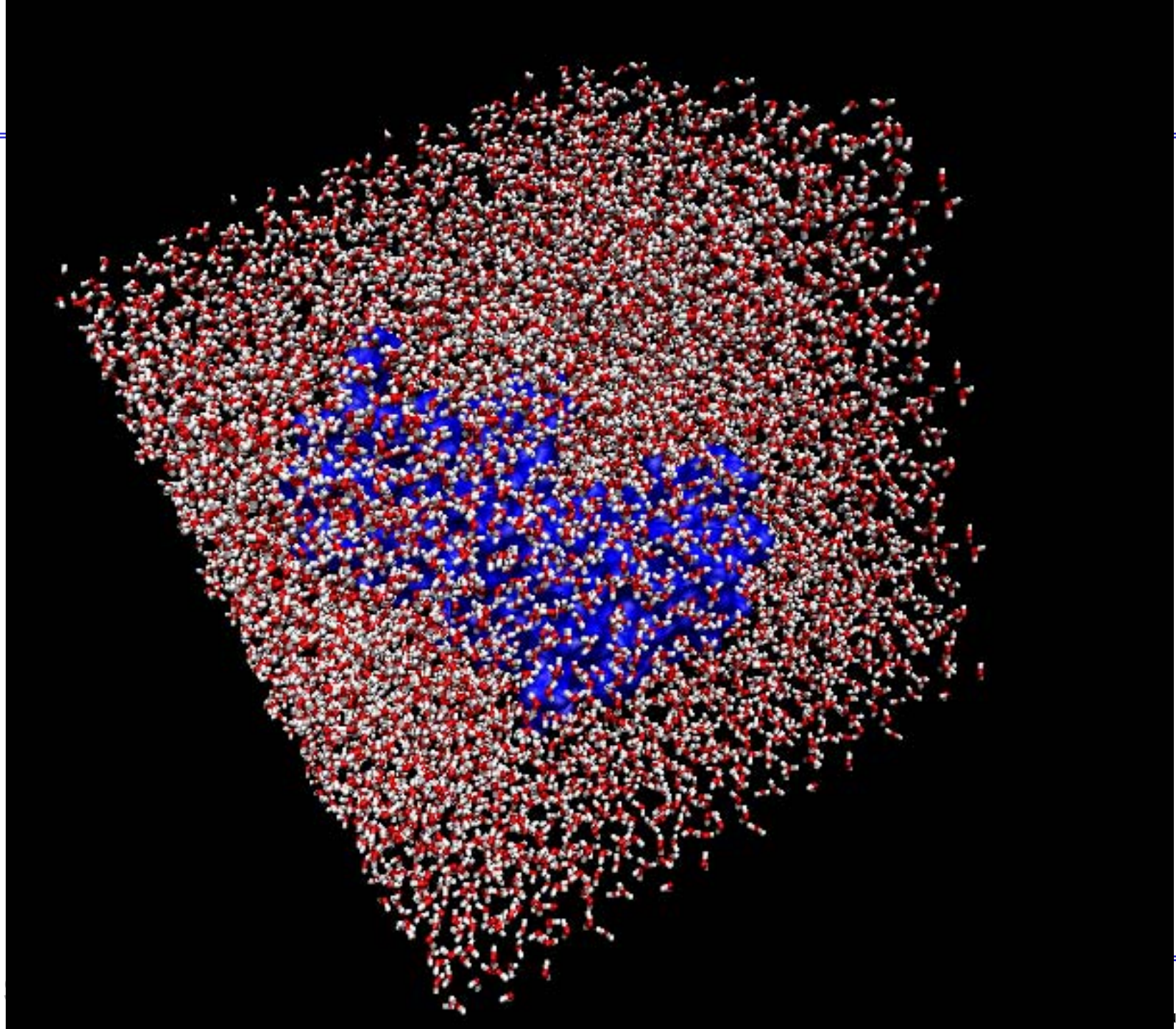


Periodische Randbedingungen

- Aussagekräftige Simulationen erfordern große Wasserboxen
- Wasser-Vakuum-Grenze verursacht unerwünschte Effekte
- Viele explizite Wassermoleküle verlangsamen Simulation

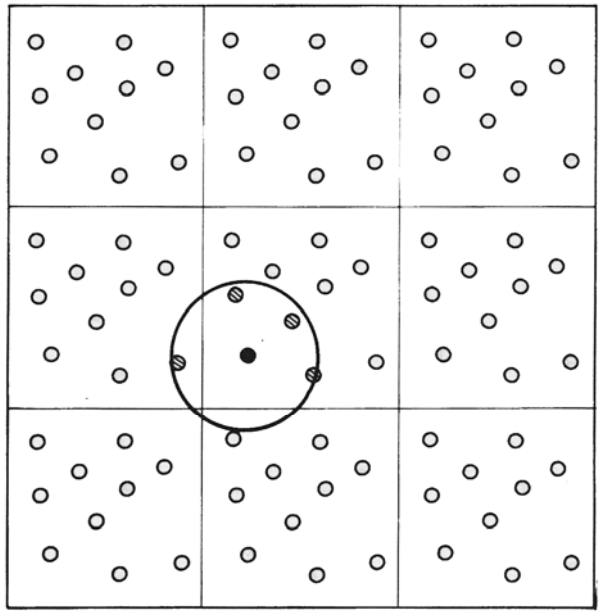


- Trick: **periodische Randbedingungen**
- Teilchen werden in Nachbarboxen „gespiegelt“
- Berechnungen benutzen diese Phantom-Teilchen für langreichweitige Wechselwirkungen



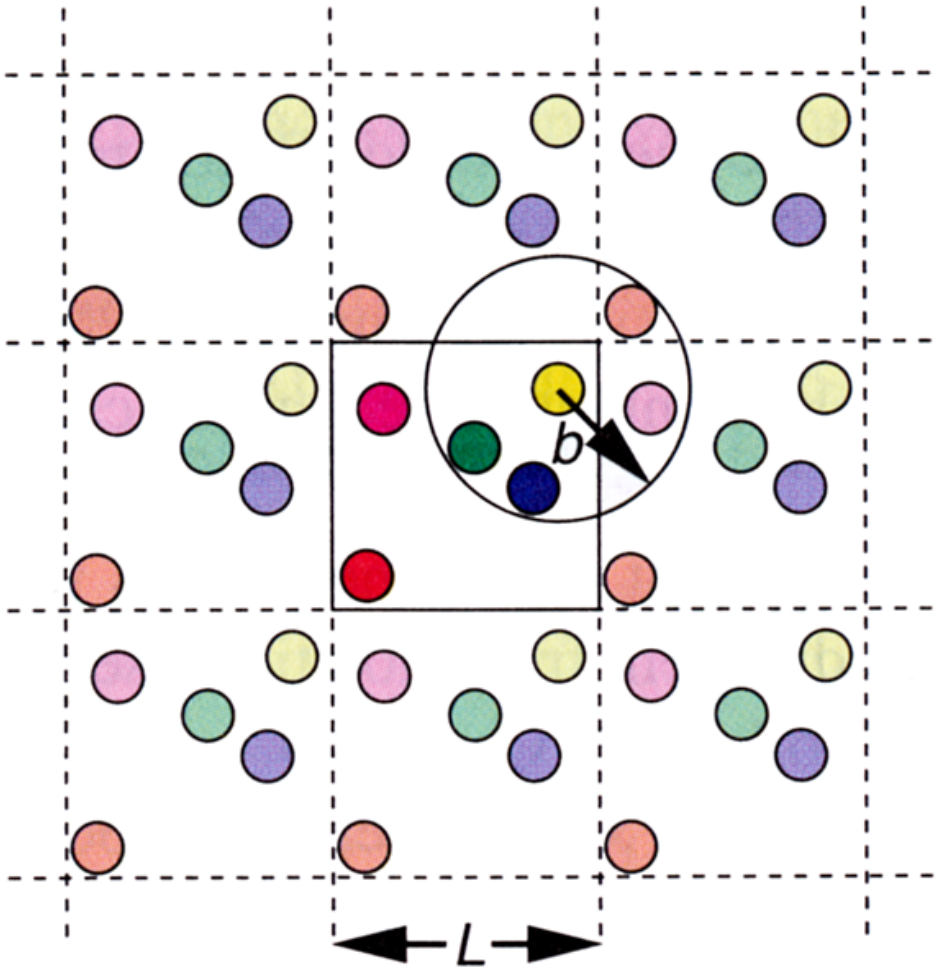
Minimum-Image-Konvention

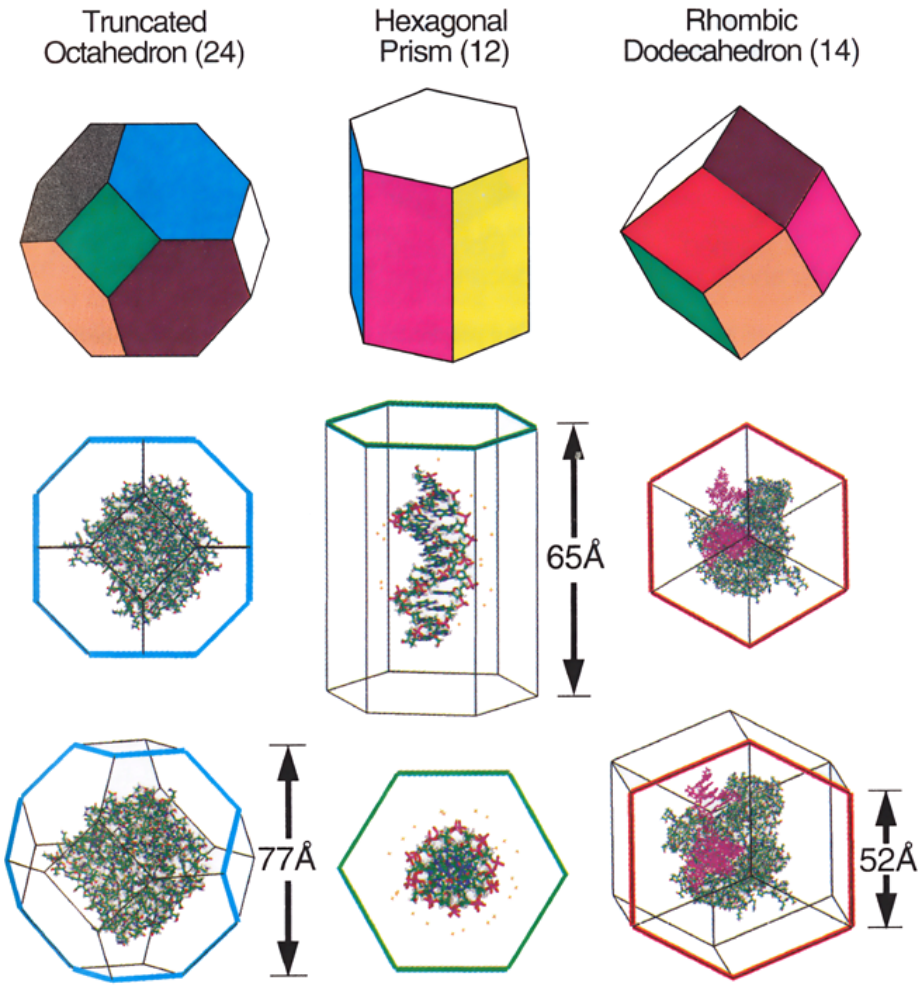
- Zeitaufwändig: Non-bonded Wechselwirkungen
- Aber: Starker Abfall der Kurven (vdW: proportional zu $1/r^6$)
- Wechselwirkung schon über kurze Distanzen sehr schwach



- **Minimum-Image-Konvention:** jedes Atom sieht nur das räumlich nächste Bild eines anderen Atoms
- Zusätzlicher Cut-Off beschleunigt Berechnung weiter

Periodische Randbedingungen





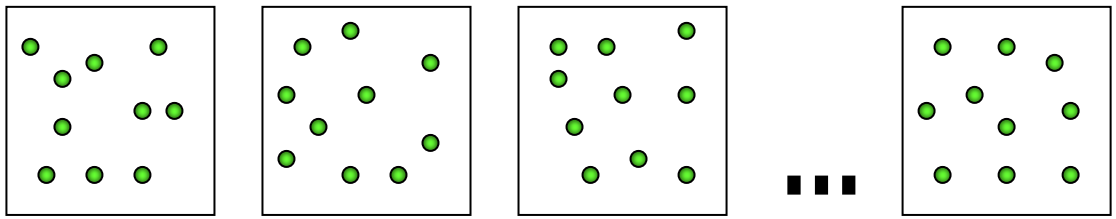
Tamar Schlick: Molecular Modeling and Simulation, p. 272, 273

Thermodynamische Eigenschaften

- Man kann viele thermodynamische Eigenschaften eines Systems aus der ausreichend langen Trajektorien berechnen
- Thermodynamische **Zustandsvariablen**
  - Temperatur T
  - Druck p
  - Volumen V
  - Teilchenzahl N
  - Innere Energie E

Ensembles

- Ensemble: Menge von Systemen mit **verschiedenen mikroskopischen Zuständen**, aber **identischem makroskopischem Zustand**



- Makroskopischer Zustand definiert durch thermodynamische Kenngrößen
- Verschiedene Arten von Ensembles
  - **Kanonisch** (NVT)
  - **Mikrokanonisch** (NVE)
  - **Isobar/isothermal** (NPT)

Ergodentheorem

- Das **Ergodentheorem** besagt, dass in einem abgeschlossenen System der **zeitliche Mittelwert** einer thermodynamischen Größe **gleich ihrem Ensemblemittel** ist.

) Zwei äquivalente Möglichkeiten, um thermodynamische Größen zu berechnen:

  - Aus einem **Ensemble** (vielen mikroskopisch unterschiedlichen Systemen)
  - Aus der Simulation eines einzigen Systems über **lange Zeiträume**

Simulationen bei T = const.

- Im kanonischen Ensemble sind konstant:
  - Teilchenzahl N
  - Volumen V
  - Temperatur T
- Variabel sind Druck und Energie
- In der Regel wird die Temperatur vorgegeben und auch geändert, so z.B. um die Entfaltung eines Proteins durch Erhöhen der Temperatur zu simulieren
- Dazu benötigt man eine Möglichkeit diese Temperatur einzustellen
- Entsprechende Algorithmen sind auch unter dem Begriff „**Thermostaten**“ bekannt

Thermostat

- Einfacher Thermostat: Skalierung der Geschwindigkeiten (*velocity rescaling*)
- Bilde Differenz ΔT zwischen Soll- und Ist-Wert und berechne daraus den Skalierungsfaktor λ

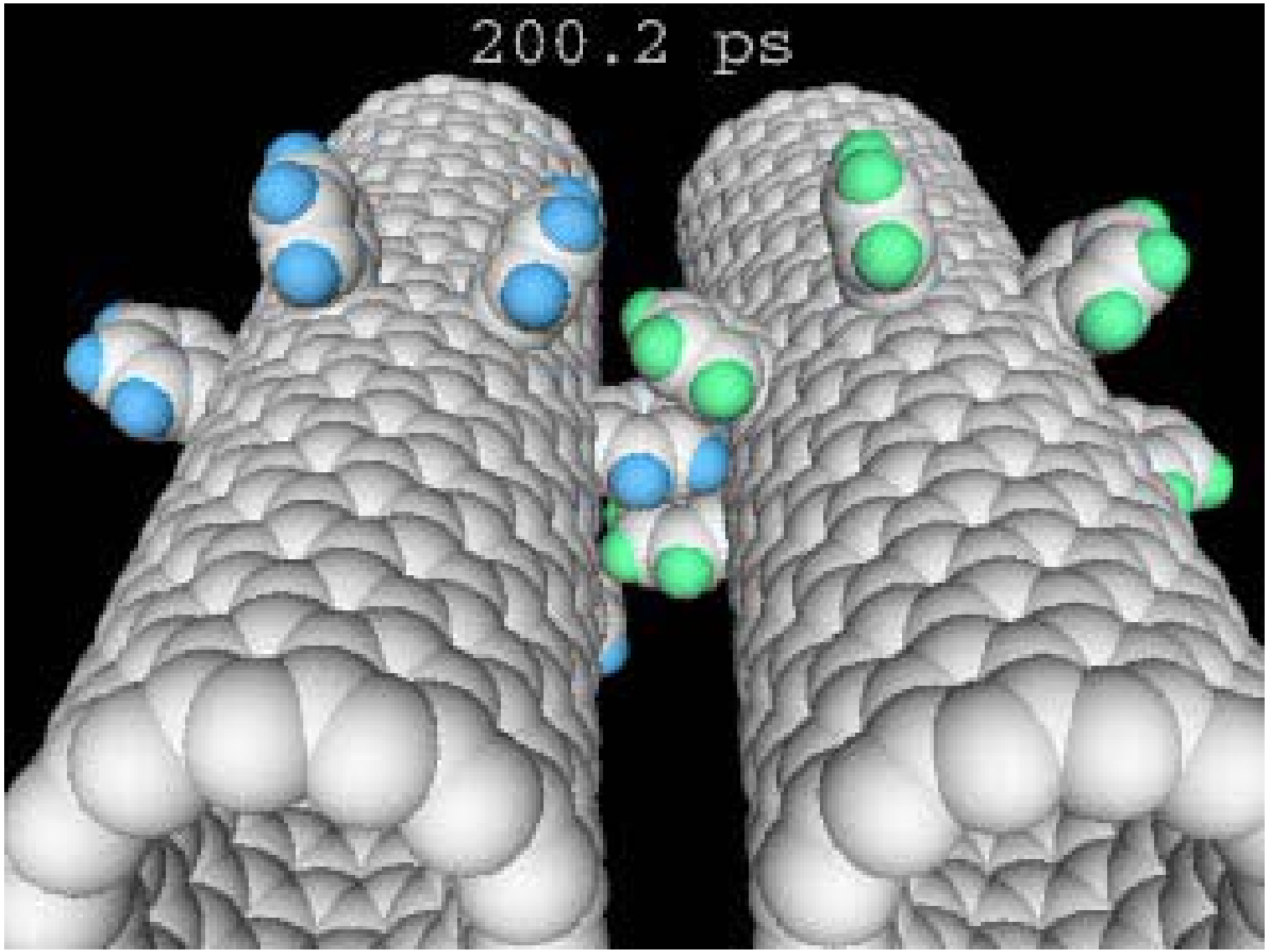
$$\begin{aligned} \Delta T &= \frac{1}{3Nk_B}\sum(\lambda v_i)^2 - \frac{1}{3Nk_B}\sum v_i^2 \\ &= (\lambda^2 - 1)T(t) \\ \Rightarrow \lambda &= \sqrt{\frac{T_{\text{neu}}}{T(t)}} \end{aligned}$$

- Berechne in jedem Zeitschritt die aktuelle Temperatur und λ und skaliere die $\mathbf{v}_i$

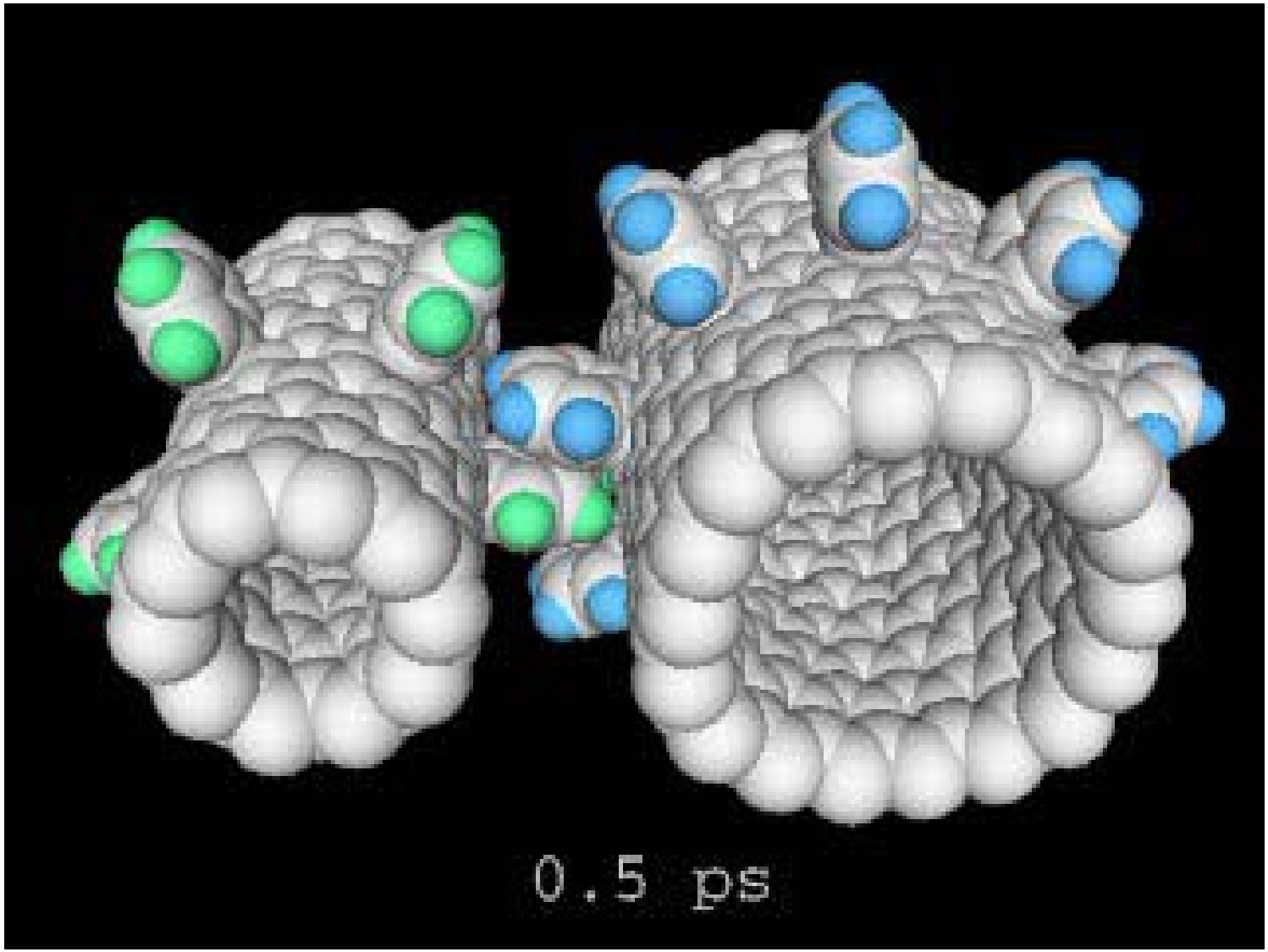
Simulationen bei konstantem Druck

- Isobares Ensemble
- Geeignet zur Simulation druckinduzierte Effekte (Phasenübergänge)
- Makroskopische Systeme halten Druck konstant durch Änderung des Volumens
- Einführung eines **Mannostaten**
- Skalierung des Volumens analog zum Thermostat

Nanogetriebe



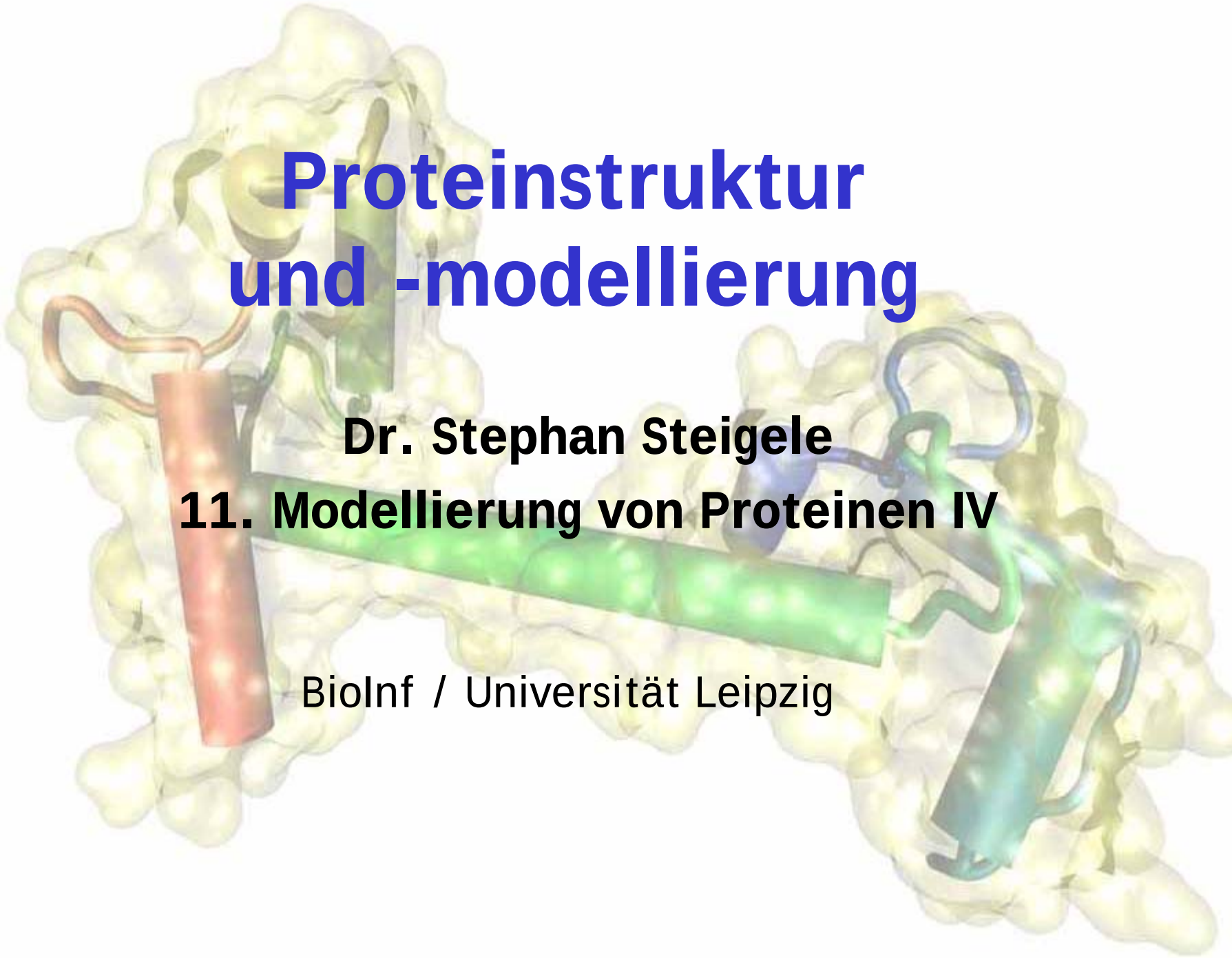
Nanogetriebe II



Literatur

Molekülmechanik

- **Andrew R. Leach, Molecular Modelling – Principles and Applications**, Prentice Hall, 2001
- **Anthony J. Stone, The Theory of Intermolecular Forces**, Clarendon Press, 1996
- **Daan Frenkel, Berend Smit, Understanding Molecular Simulation**, Academic Press, 1996
- **Martin J. Field, A practical introduction to the simulation of molecular systems**, Cambridge University Press, 1999
- **Tamar Schlick, Molecular Modeling and Simulation**, Springer, 2003
- **Ulrich Burkert, Norman L. Allinger, Molecular Mechanics**, American Chemical Society, 1982



Proteinstruktur und -modellierung

Dr. Stephan Steigele

11. Modellierung von Proteinen IV

BioInf / Universität Leipzig

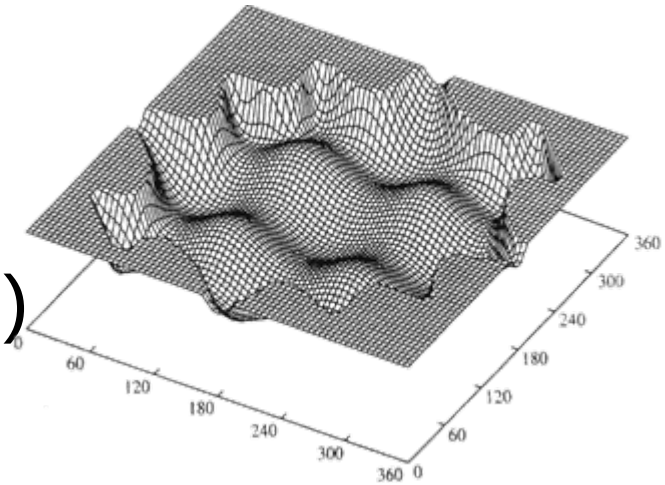
Gliederung

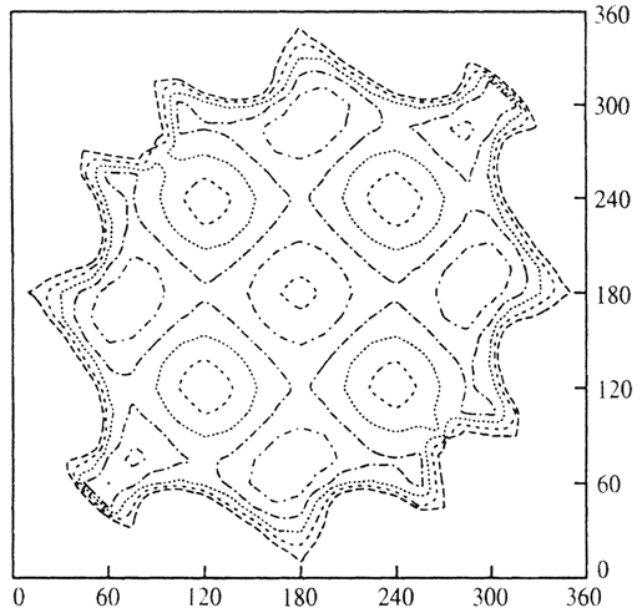
- Durchmustern des Konformationsraums
  - Motivation
  - Systematische Suche
  - Molekulardynamik
  - Monte Carlo
    - Theorie
    - Algorithmen
    - Berechnung von Ensemblemitteln
  - Vergleich MD/MC
- Molekülmechanik-Implementierungen
- Übersicht und Zusammenfassung Molekülmechanik

Suche im Konformationsraum

- Minima
  - Entsprechen günstigen Konformationen (Konformeren)
  - Häufig lokale Minima!
- Globales Minimum?

  (Bsp.: Proteinfaltung!)
- **Kann man die Oberfläche systematisch durchmustern**
- **Mittelung über Ensembles**





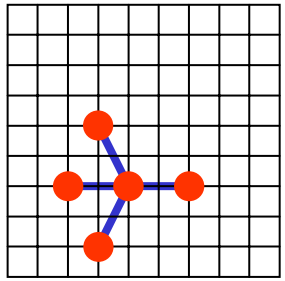
Systematische Suche

- Geht nur für kleine Anzahl Freiheitsgrade (kombinatorische Explosion)
- **Beispiel**
  - Protein mit 1000 Atomen
  - Koordinaten in Würfel von 20 Å Seitenlänge
  - Diskretisierung mit 0.2 Å Abstand

) 100 mögliche Werte für jede der 3000 Koordinaten

) 10^{6000} mögliche Energien

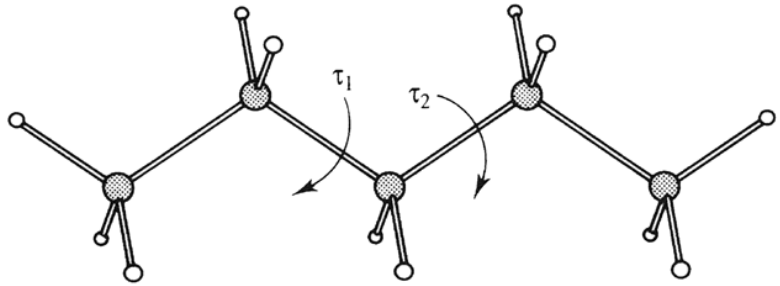
  (sichtbares Universum enthält ca. 10^{80} Teilchen!)

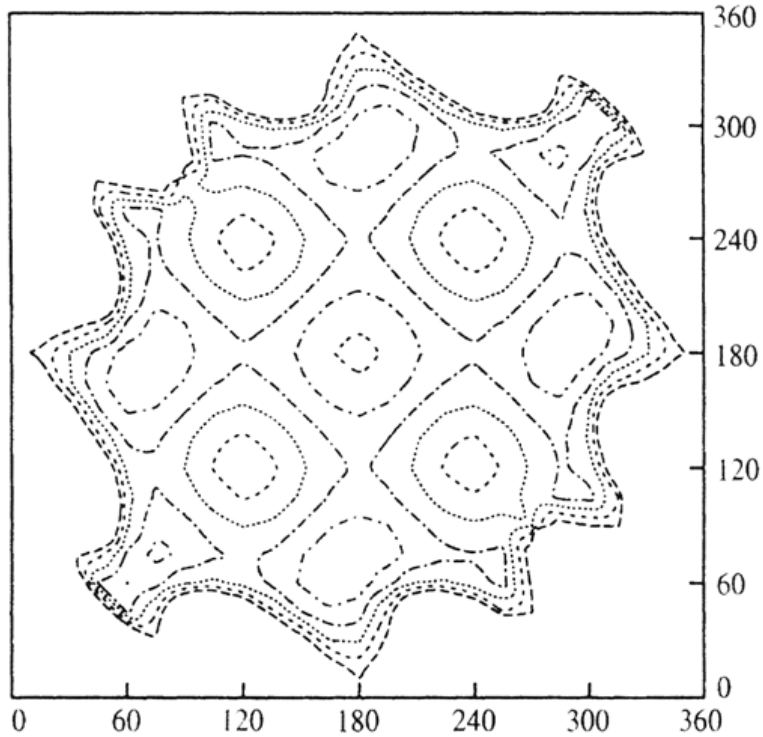


Systematische Suche

- Unabhängige Betrachtung der Koordinaten ist naiv
- Flexibilität wird überwiegend durch Torsionen bestimmt
- Wesentlich geringere Anzahl Freiheitsgrade (ca. 2-7 pro AS)
- Wenige Minima in den Torsionen

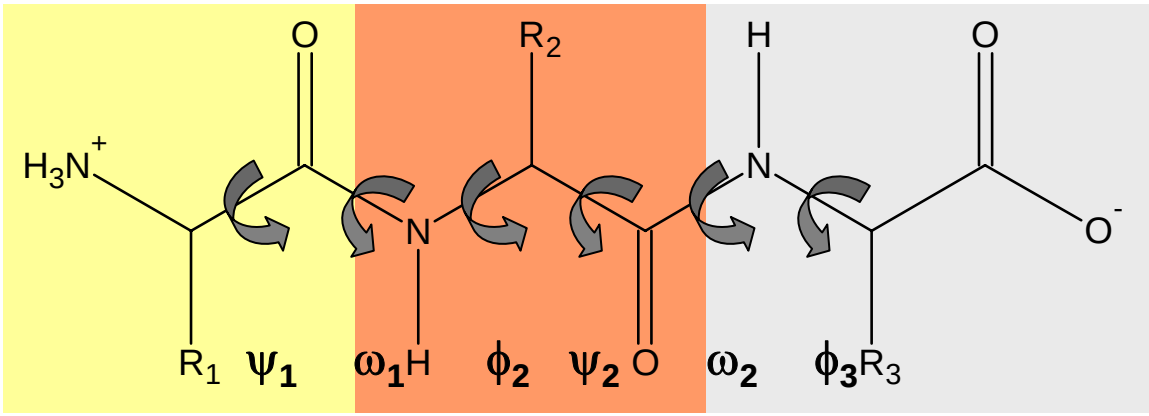
) grobe Rasterung (0/120/240°)
- Vermeidet Betrachtung physikal. unsinniger Konformationen





Baumsuche

- Systematische Suche durch den Torsionsraums eines Proteins lässt sich durch Aufzählungsbaum beschreiben
- Dabei entsprechen Knoten möglichen (minimalen) Konformationen von Backbone/Seitenketten



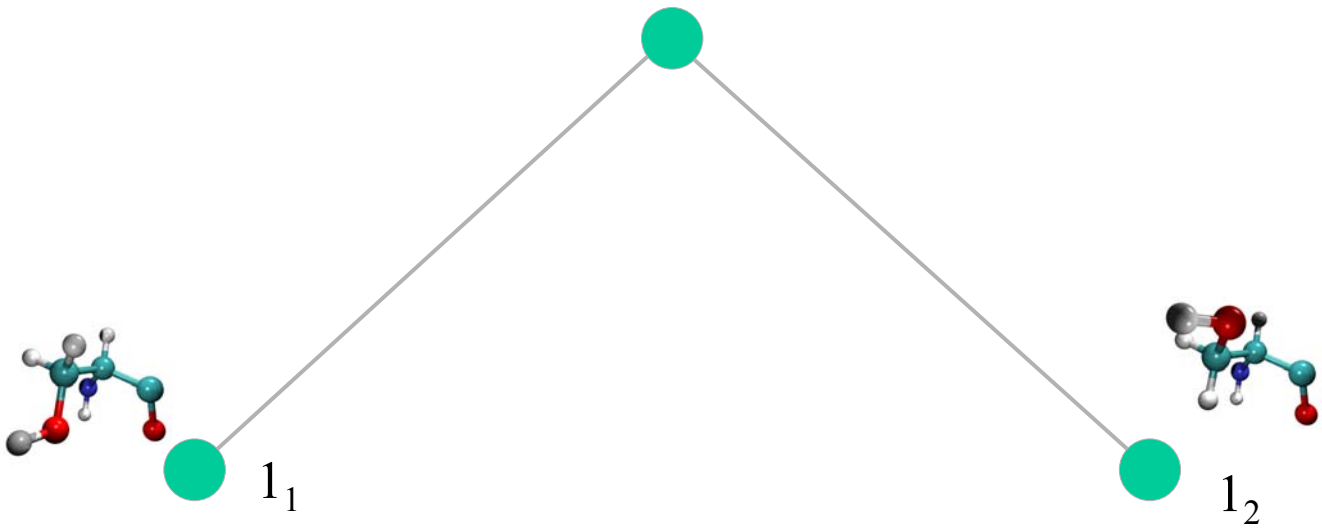
Baumsuche

- Kombinatorische Explosion durch exponentiellen Anstieg mit der Anzahl der Torsionen/AS
- Sucht man nur energetisch günstige Konformationen, kann man eine simple (Multi-Greedy-)Heuristik nutzen
  - Beschränke die Anzahl der Knoten je Schicht des Aufzählungsbaums
  - Lineares Wachstum des Baums!

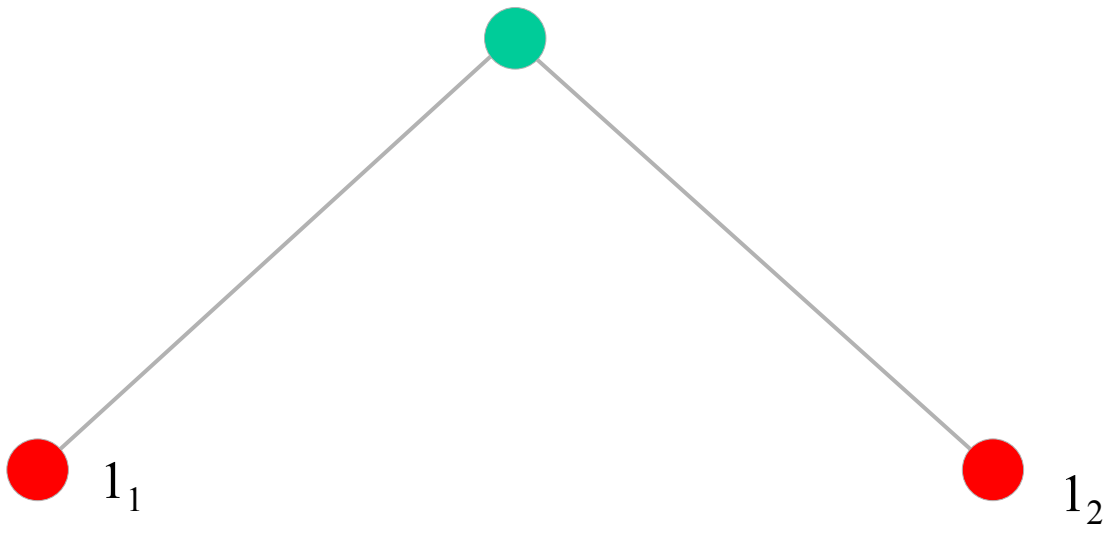
Multi Greedy Method

Startknoten

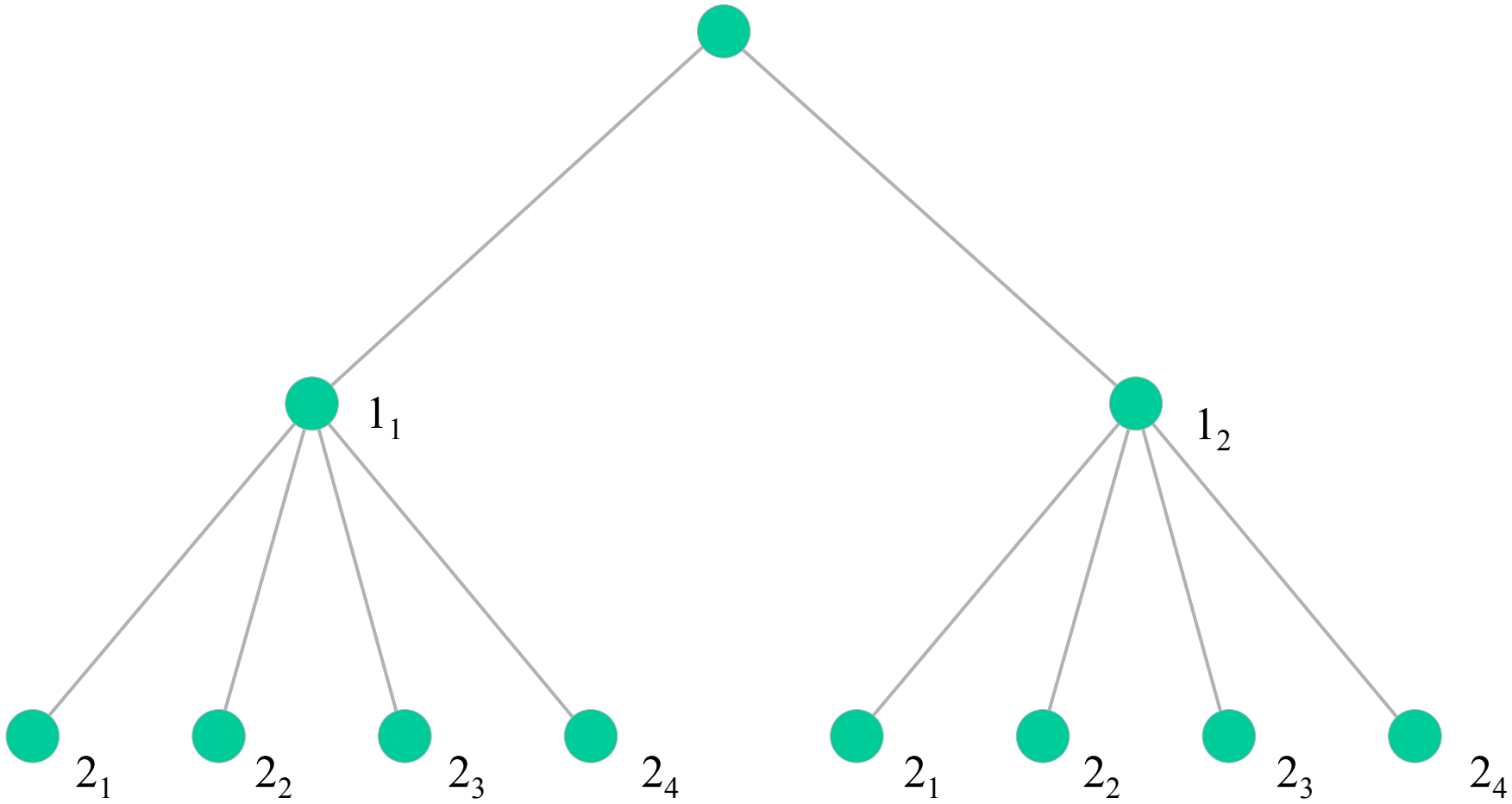
Baumsuche



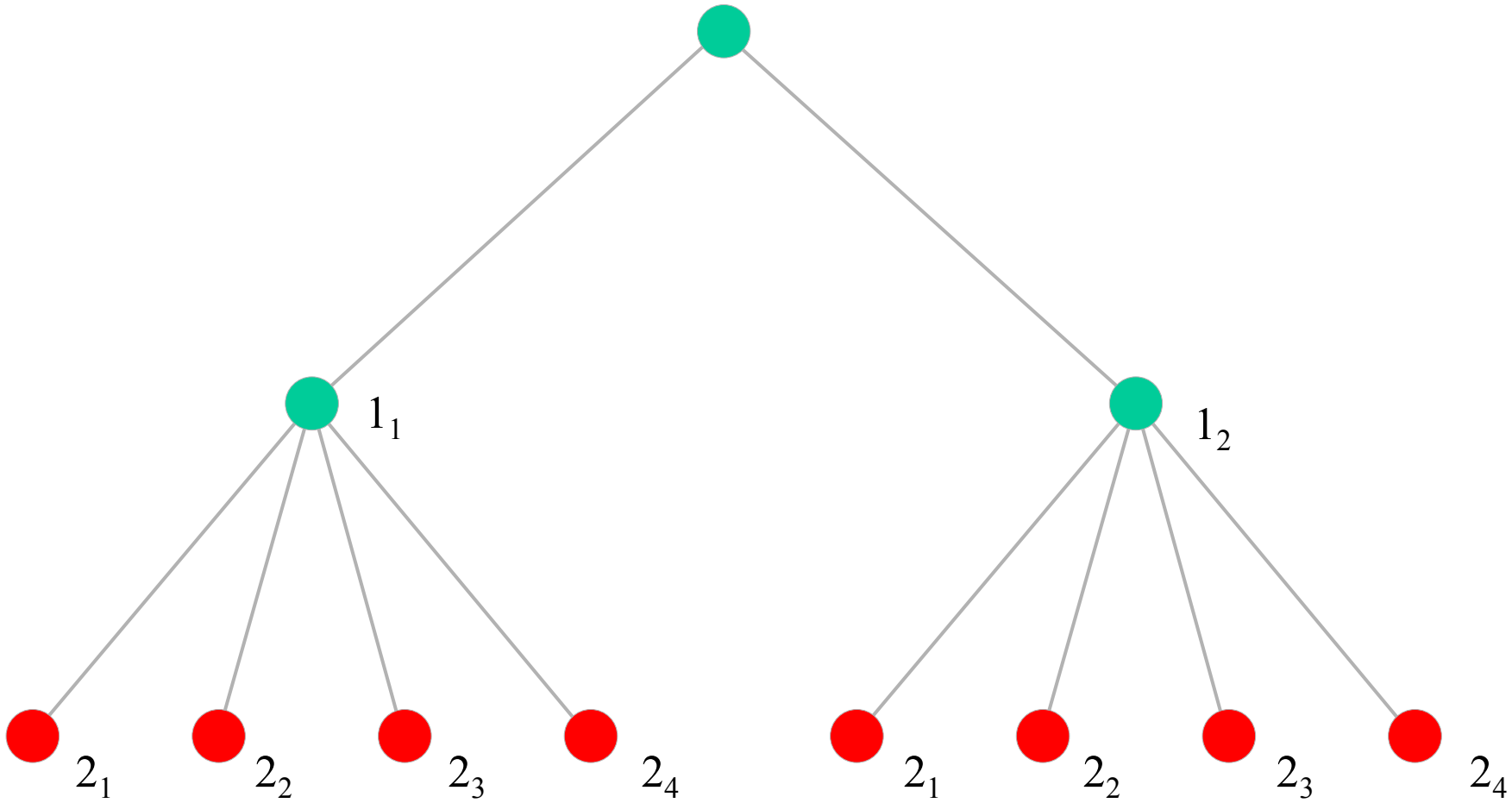
Baumsuche



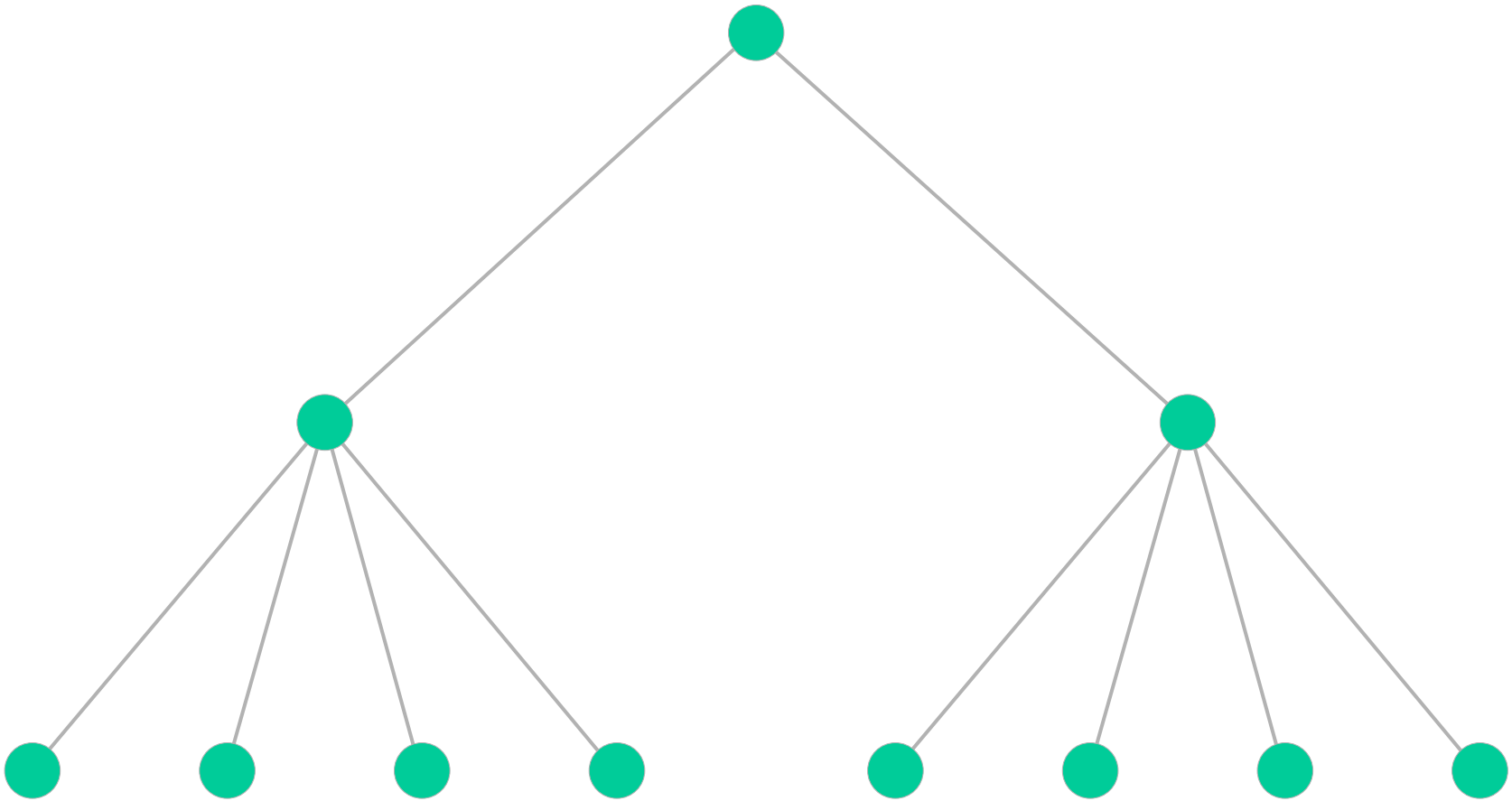
Baumsuche



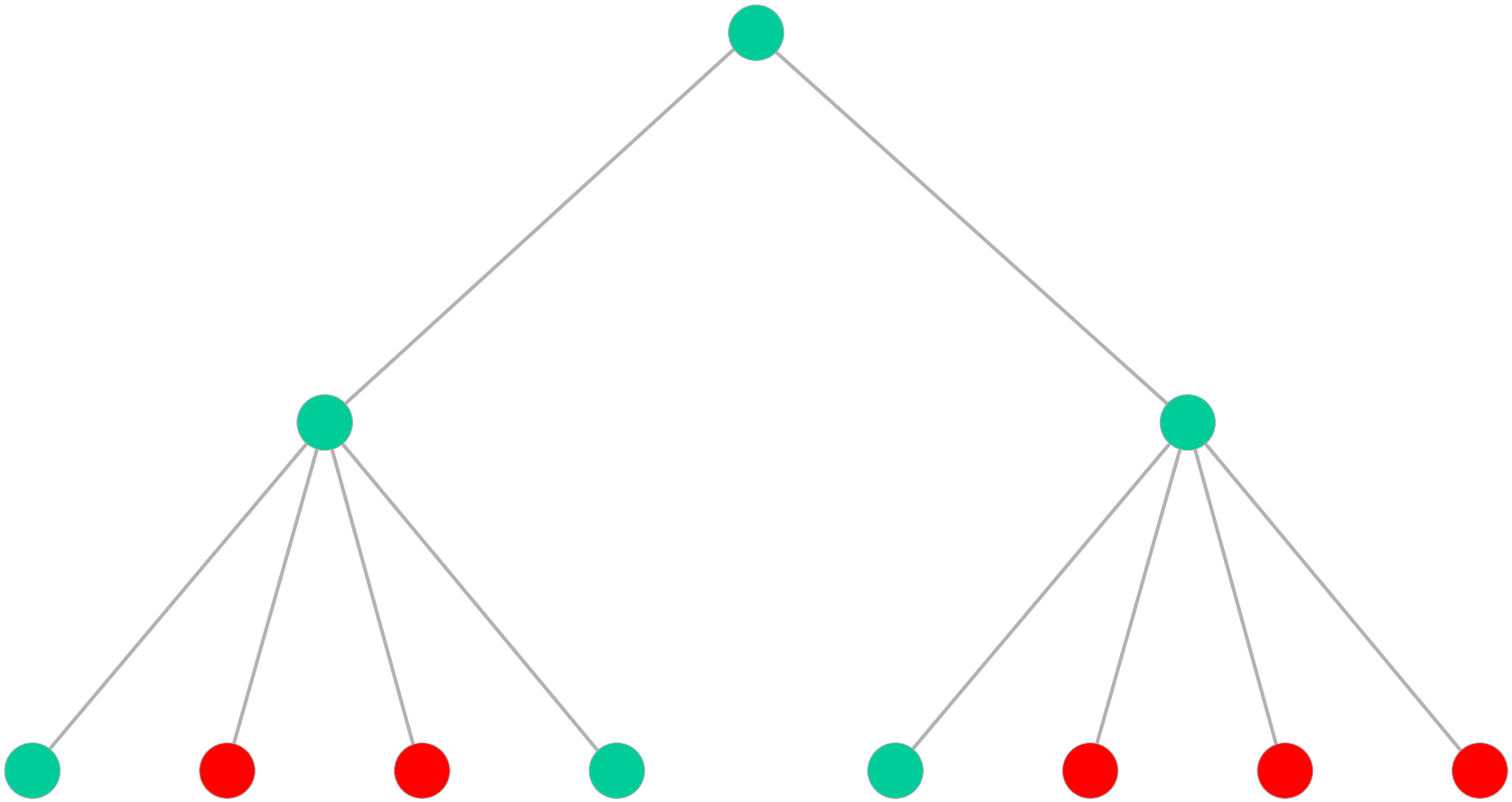
Baumsuche



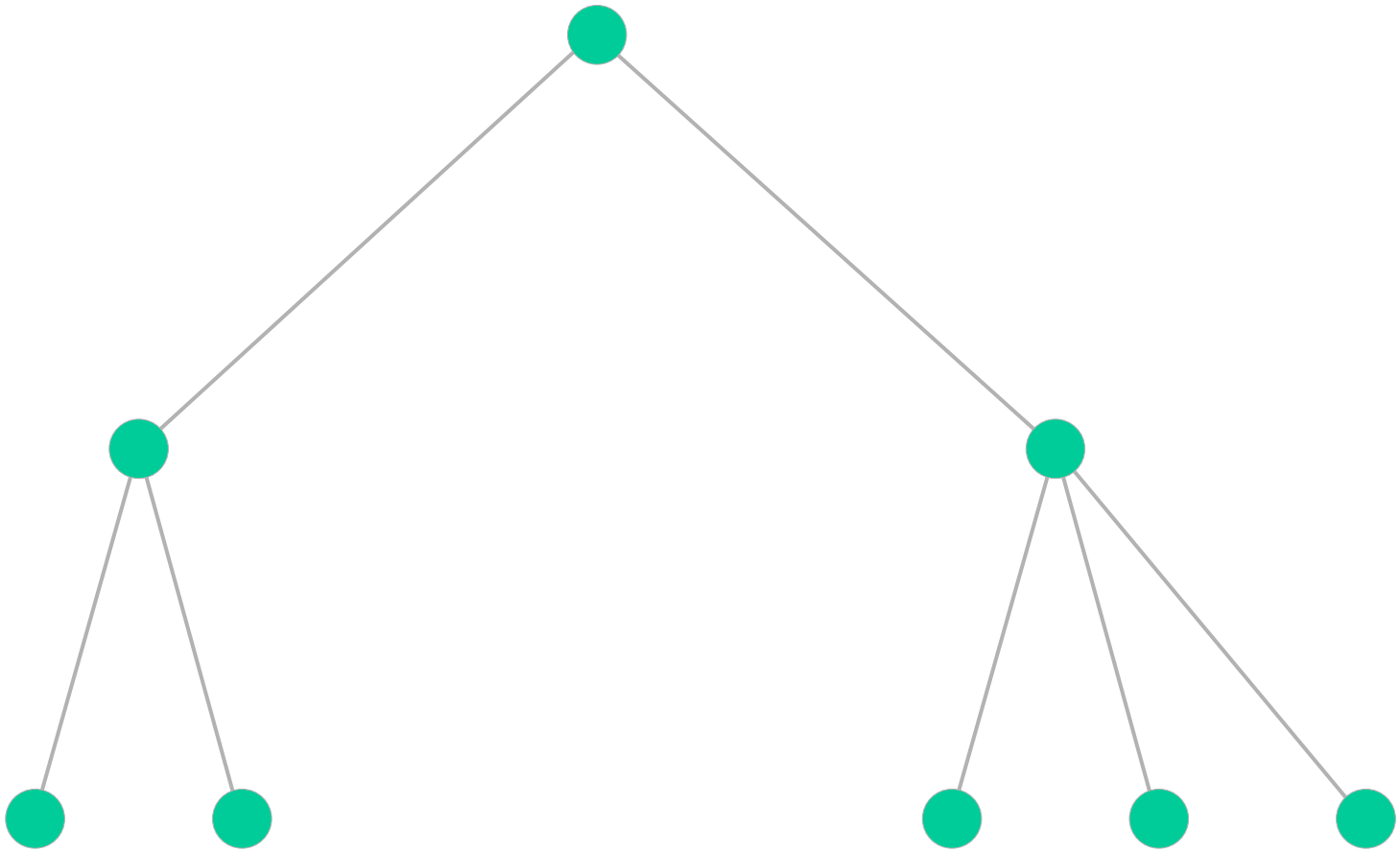
Baumsuche



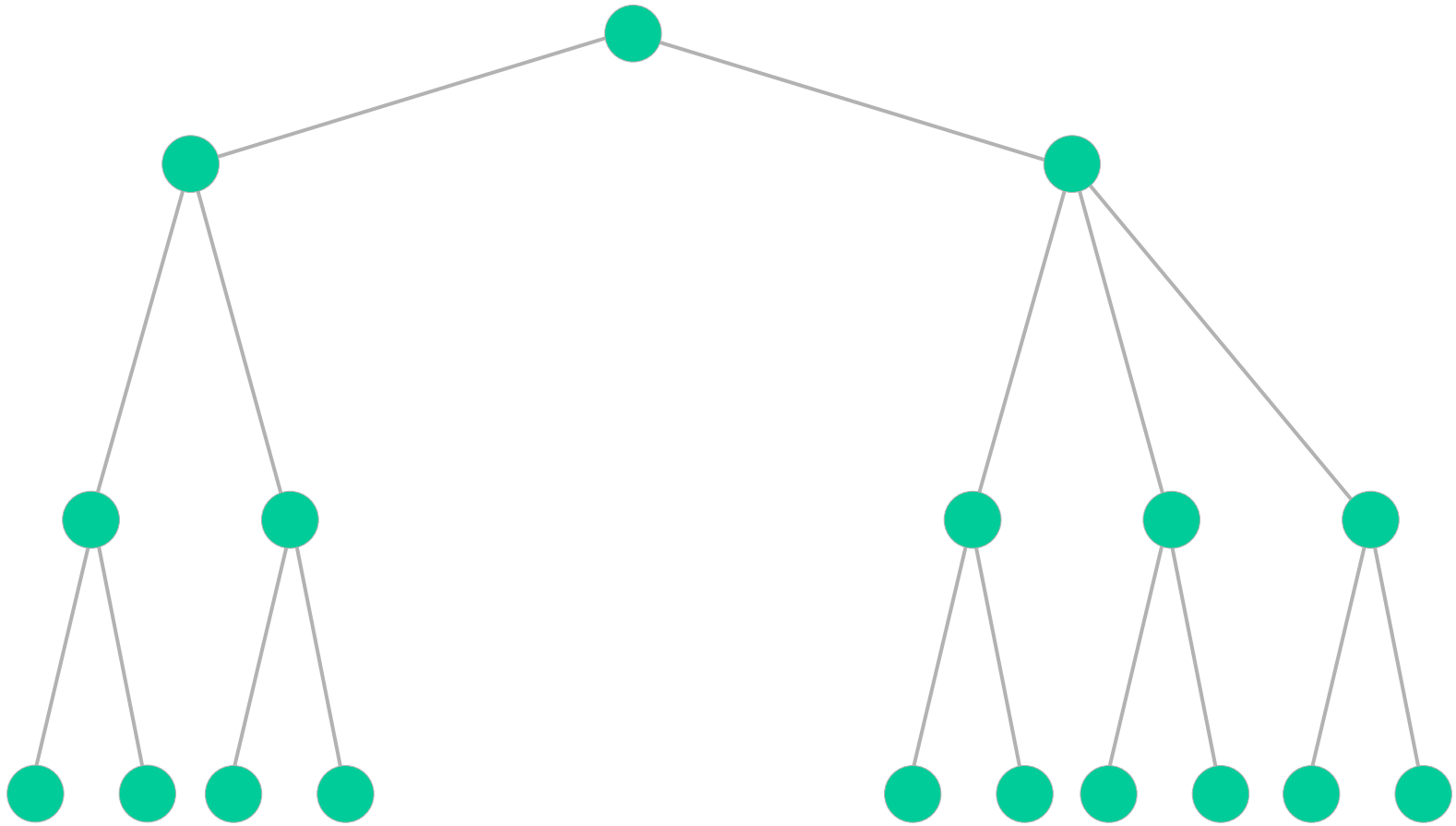
Baumsuche



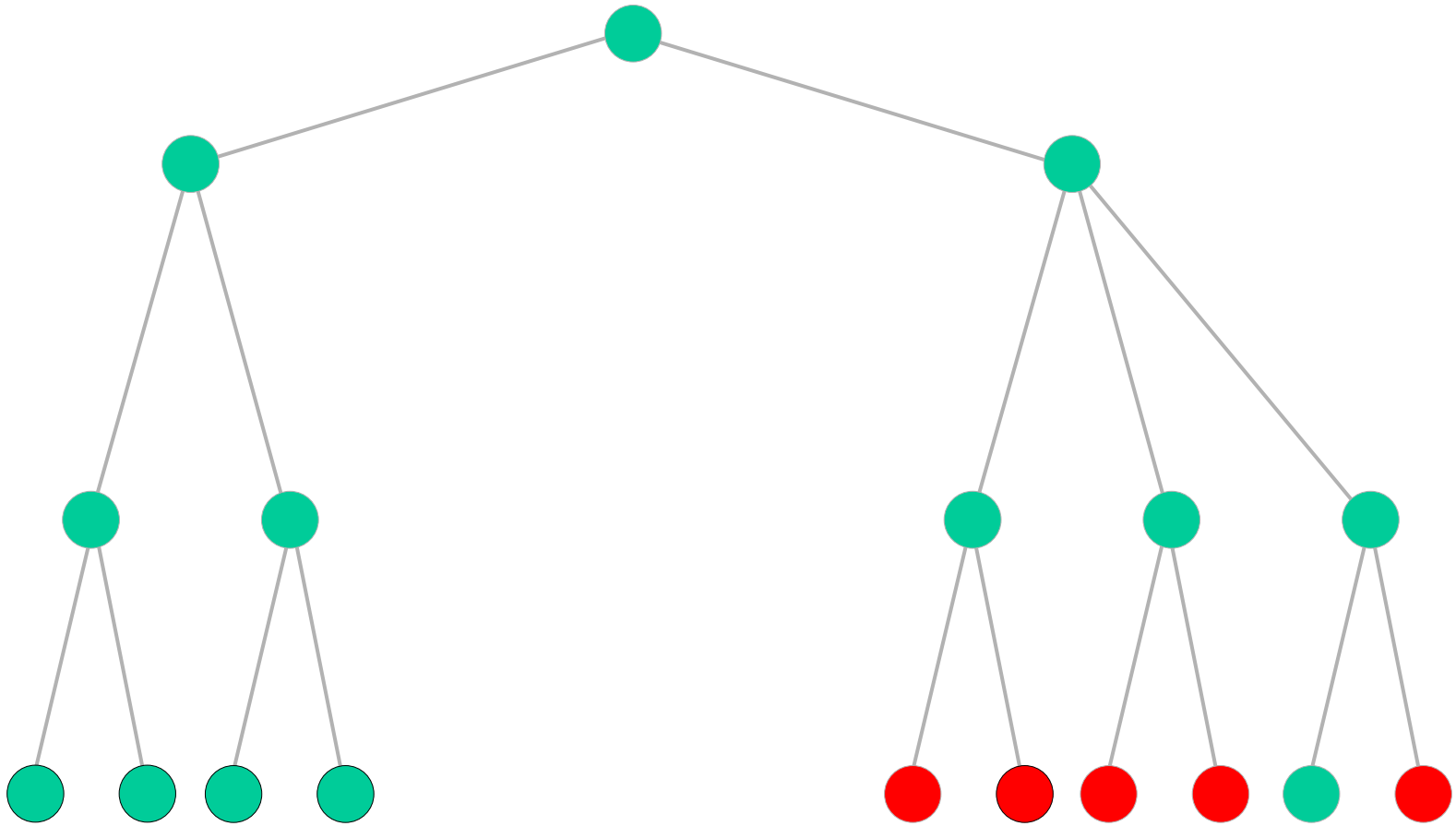
Baumsuche



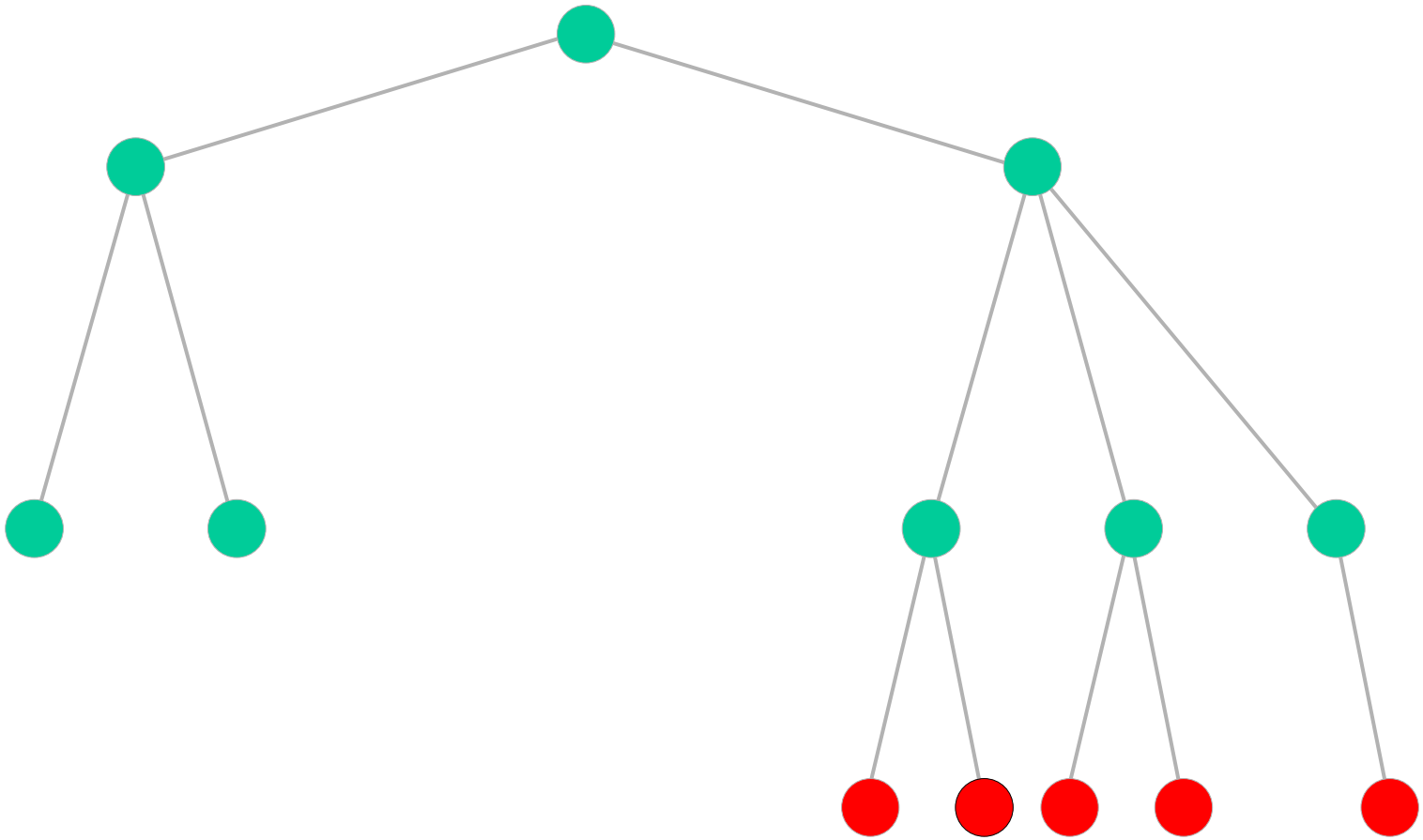
Baumsuche



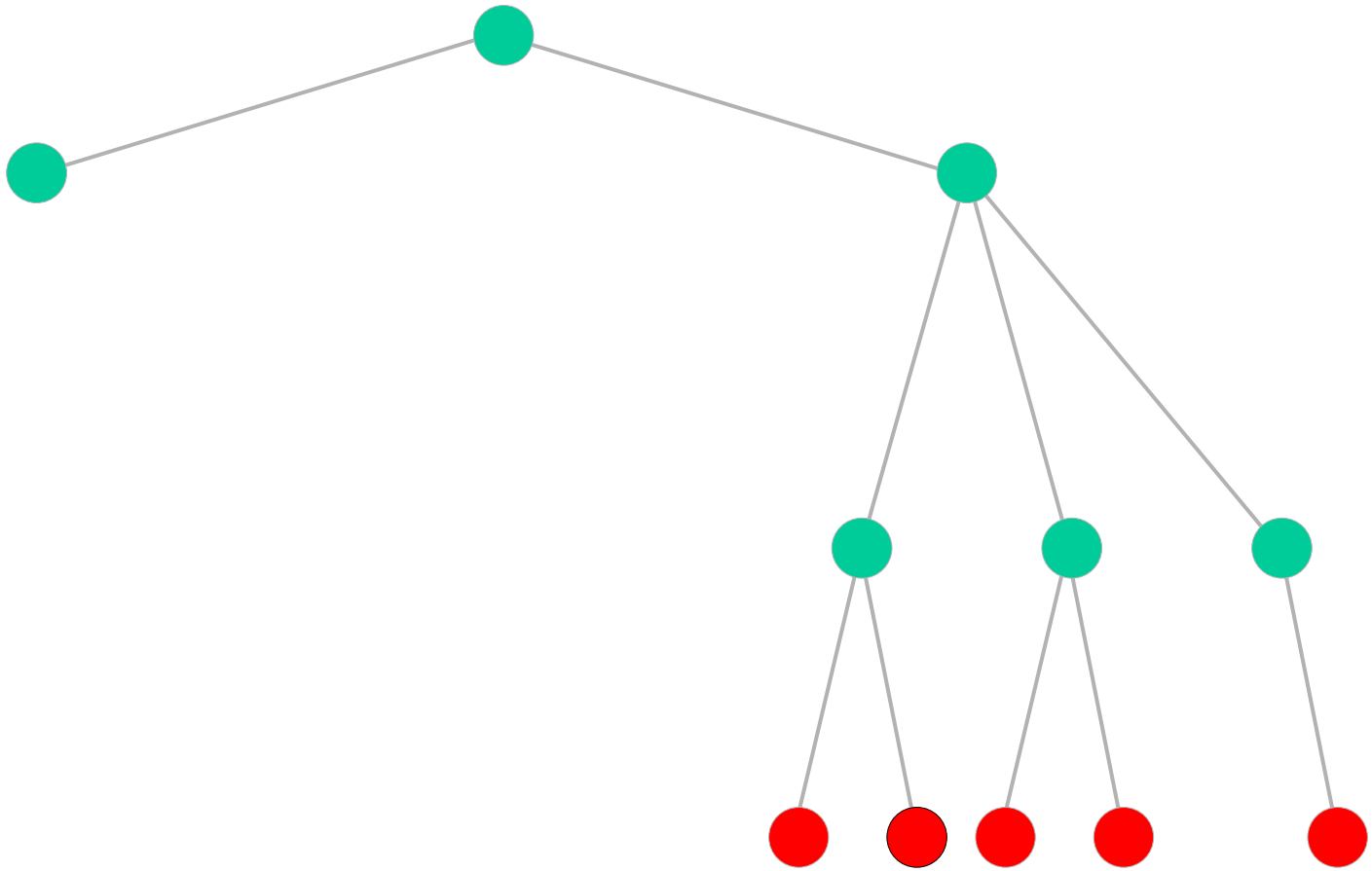
Baumsuche



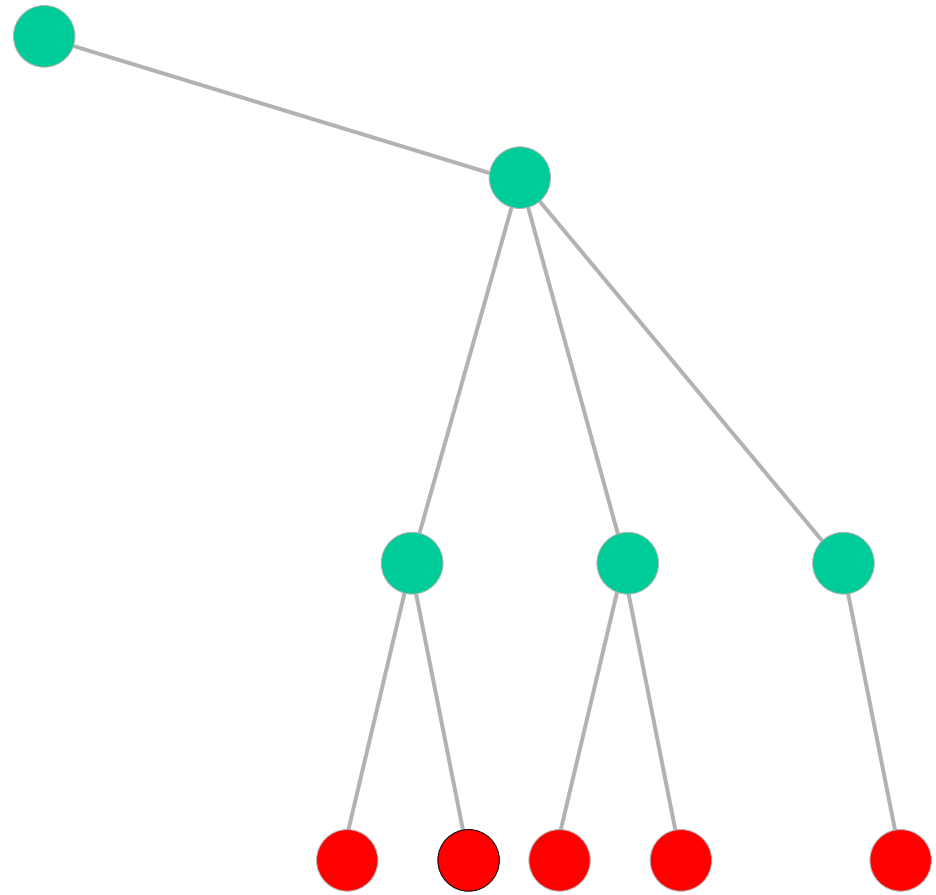
Baumsuche



Baumsuche



Baumsuche

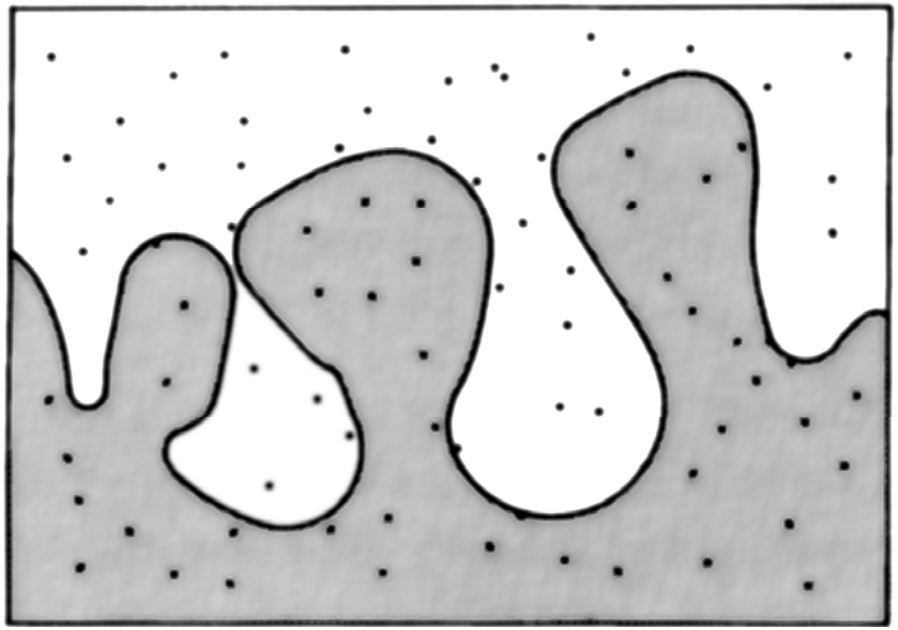


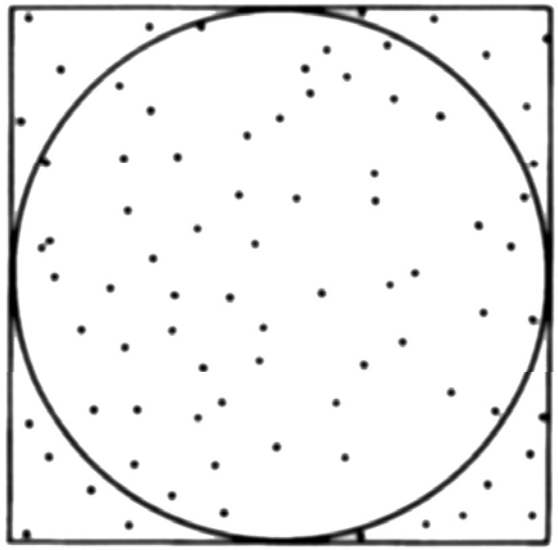
Systematische Suche

- Große Suchräume ($<10^{80}$) können in vertretbarer Zeit nach dem Minimum durchmustert werden
- Geschicktere Algorithmen werden wir beim Proteindesign kennen lernen
- **Nachteile**
  - Minima können zwischen das Raster fallen
  - Keine sinnvolle Mittelung möglich
  - Die meisten durchmusterten Konformationen sind nicht interessant
- Alternative Idee: stochastische Methoden

Stochastische Methoden

- Ensemblemittel entspricht einem Phasenraumintegral
- Integral kann durch zufällige Stichprobe geschätzt werden

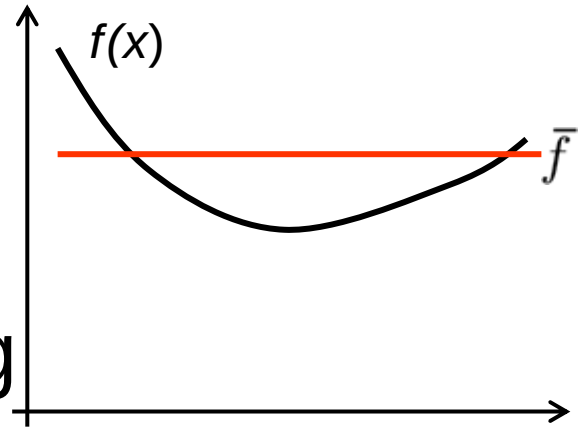




Leach: Molecular Modelling, S. 413

Monte-Carlo-Integration

- **Gegeben:** Funktion $f(x)$
- **Gesucht:** $I = \int_a^b f(x)\,dx$
- **Idee:** Mittelwertberechnung



$$\bar{f} = \frac{\int_a^b f(x)dx}{b-a} \Rightarrow I = (b-a)\bar{f}$$

- Schätze den Mittelwert aus einer Stichprobe an zufälligen Stellen $x_1, \ldots x_N$

Monte-Carlo-Integration

- Mittelwertbildung wird dann zu Mittelung über Funktionswerte an $x_0, \ldots, x_N$

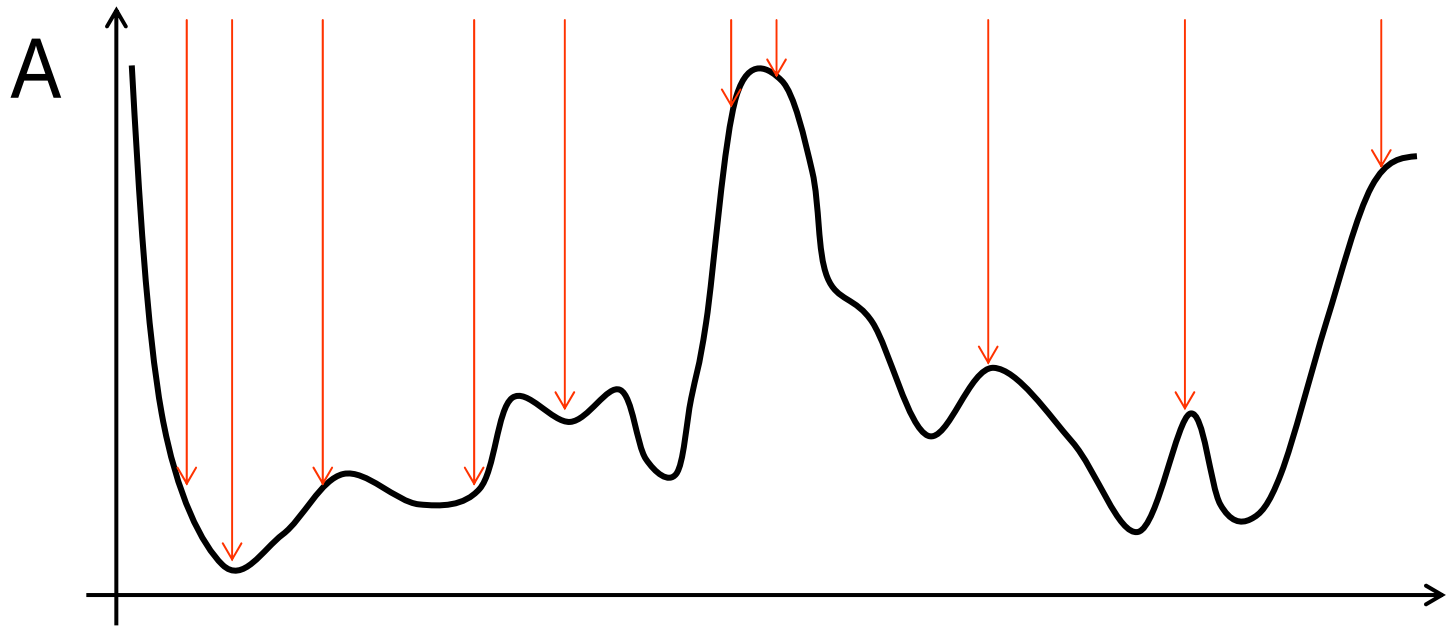
$$\bar{f} \approx \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

- Damit ergibt sich das Integral als

$$\int_a^b f(x)dx \approx \frac{b-a}{N} \sum_{i}^{N} f(x_i)$$

- Je größer die zufällige Stichprobe, desto genauer die Integration

Stochastische Methoden

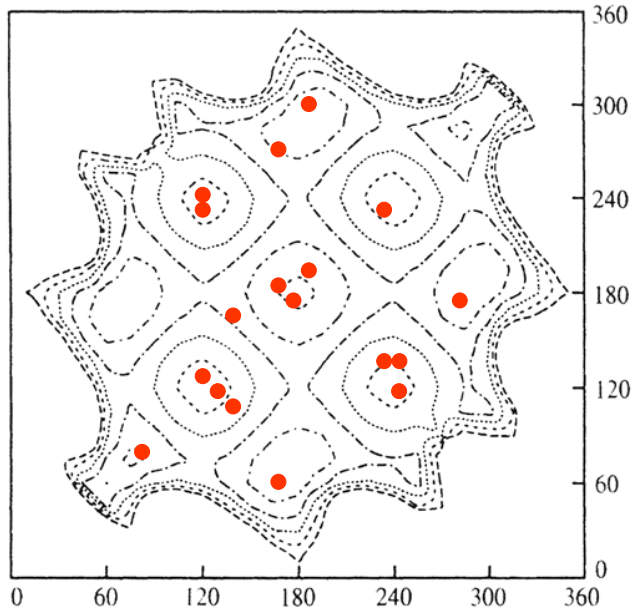


- Zufälliges Abtasten des Phasenraums
- **Problem**
  - Energetisch ungünstige Punkte werden in der Natur seltener angenommen

) Übergewichtung energetisch ungünstiger Punkte

Abtastung des Konformationsraums

- Protein kann *per se* beliebige Punkte im Konformationsraum annehmen
- Fast alle sind energetisch sehr ungünstig
- In der Realität liegt ein **Ensemble** von Molekülen vor
- Einzelne Konformationen treten mit ihrer Energie gewichtet auf



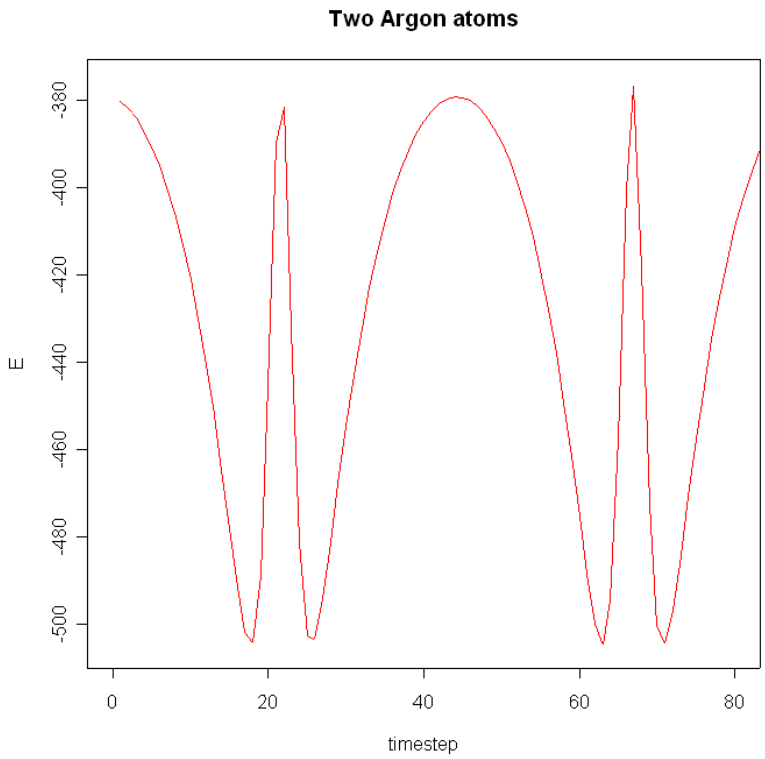
Ensemblemittel

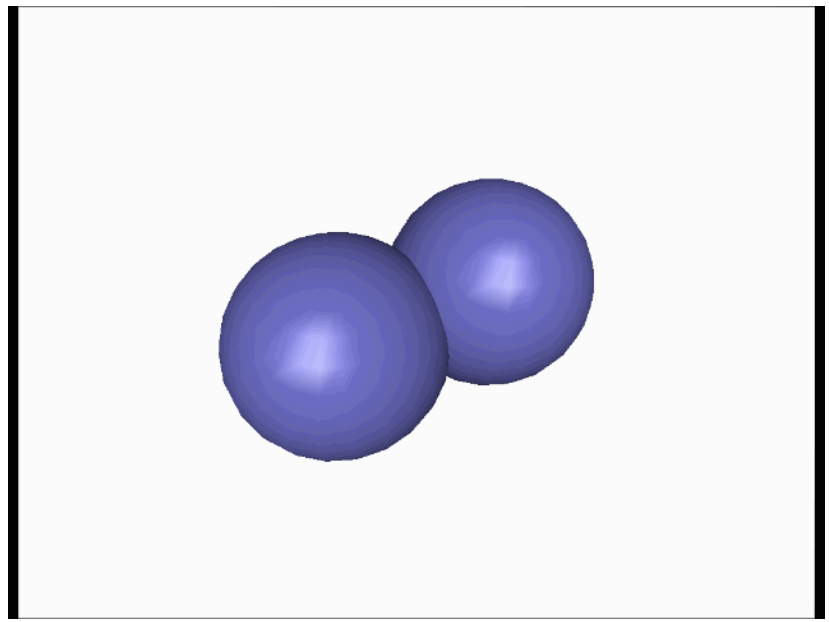
- Ensemblemittel liefert den Wert der gesuchten Größe
- Ensemblemittel gewichtet nach „Bedeutung“
  - „Seltene“ Zustände (energetisch ungünstig) tragen wenig bei
  - „Häufige“ Zustände (energetisch günstig) tragen stark bei
- Ergodenhypothese umgedreht:

  Was MD vermag, kann auch ein Ensemble aus nicht zeitlich korrelierten Zuständen
- **Beispiele**
  - **Flexibilität**: Mittel der Abweichung von Atompositionen
  - **Diffusionskoeffizient**: wie schnell wandert ein Protein durchs Wasser

Ensemblemittel

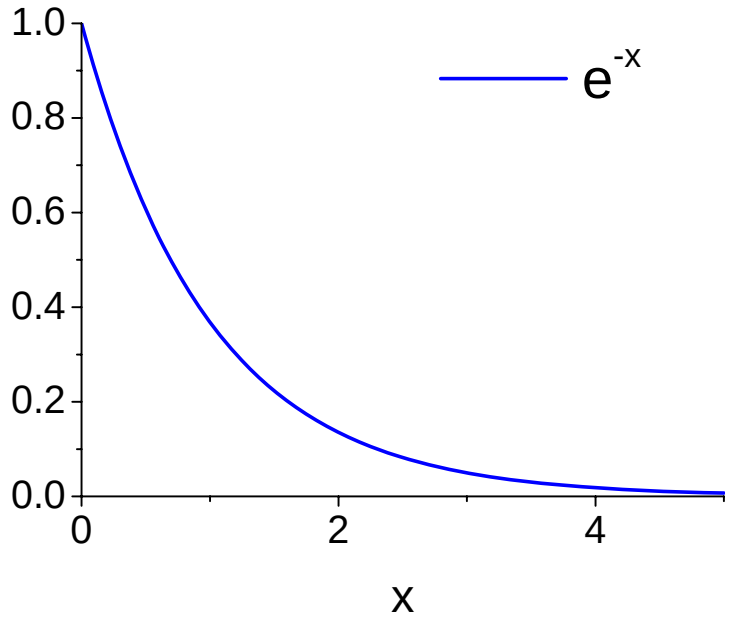
- Einzelner Punkt im Konformationsraum bedeutungslos!
- Physikalische **Observablen** sind meist **Ensemblemittel**





Boltzmann-Statistik

- Gegeben ein System
  - N Teilchen
  - Konstante Gesamtenergie
  - Zustände $E_0 \ldots E_k$ mit $E_0 < E_1 < E_2 \ldots$
  - N_i Teilchen sind in E_i
  - Gesamtzahl $\sum N_i = N$
- Im **Gleichgewicht** verteilen sich die Teilchen auf die Zustände gemäß einer Boltzmann-Verteilung



$$\frac{N_i}{N} = \frac{e^{-\frac{E_i}{k_B T}}}{\sum_i e^{-\frac{E_i}{k_B T}}}$$

Wahrscheinlichkeitsdichte

- Boltzmann-Verteilung entspricht der **Wahrscheinlichkeitsdichte** ρ im NVT-Ensemble

$$\rho(\mathbf{r},\mathbf{p}) = \frac{1}{Q} e^{-\frac{E(\mathbf{r},\mathbf{p})}{k_b T}}$$

mit der **Zustandssumme Q**

$$Q = \frac{1}{N! h^{3N}} \int\int e^{-\frac{E(\mathbf{r},\mathbf{p})}{k_b T}} d\mathbf{r} d\mathbf{p}$$

- $\rho(\mathbf{r}, \mathbf{p})$ = Wahrscheinlichkeit ein Teilchen des Ensembles im Zustand ($\mathbf{r}$, $\mathbf{p}$) zu finden.

Ensemblemittel

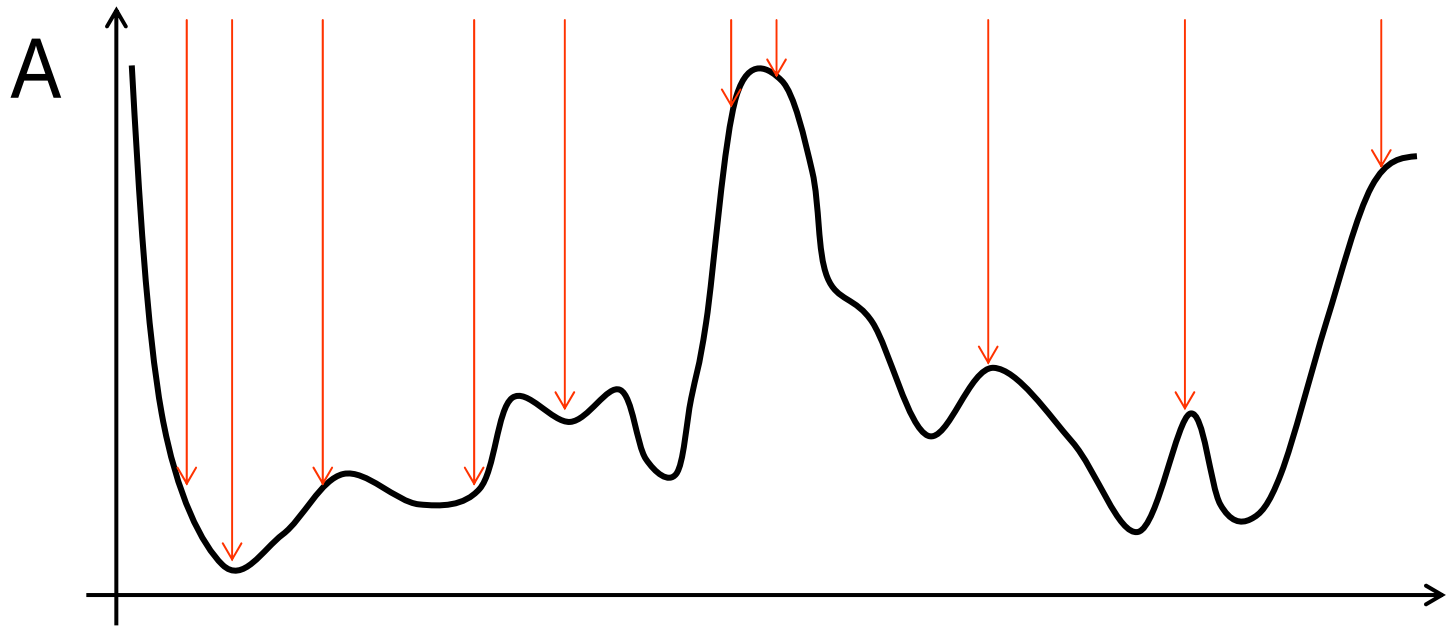
Erwartungswert <*A*> einer Größe *A* entspricht dem Ensemblemittel

$$< A >= \int \int A(\mathbf{r}, \mathbf{p}) \rho(\mathbf{r}, \mathbf{p}) d\mathbf{p} d\mathbf{r}$$

<A> entspricht auch dem Zeitmittel (nach Ergodenhypothese)

$$< A >= \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A(\mathbf{r}(t), \mathbf{p}(t)) dt$$

Stochastische Methoden



- Ensemblemittel:

$$< A >= \int \int A(\mathbf{r},\mathbf{p})\rho(\mathbf{r},\mathbf{p})d\mathbf{p}d\mathbf{r}$$

- Gewichte die Stichproben mit ρ!

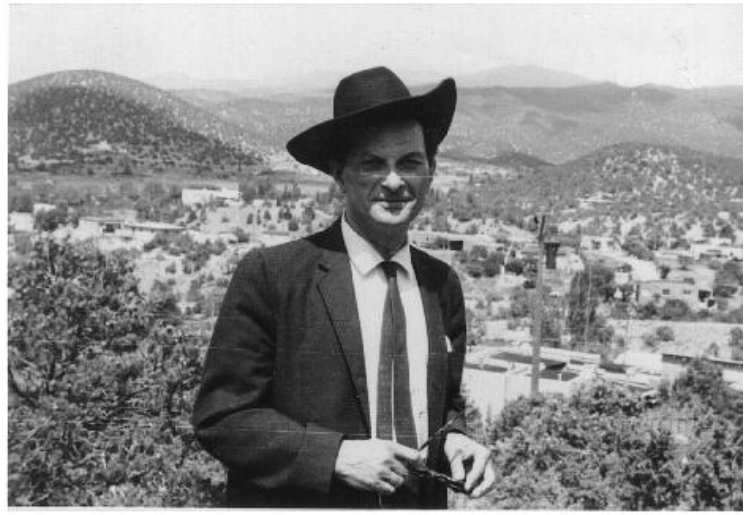
Monte-Carlo-Methode

- Die Monte-Carlo-Methode hat ihren Namen aus der Verwendung von Zufallszahlen
- 1949: **Metropolis und Ulam** verwenden den Begriff zum ersten Mal
- 1953 Metropolis-Algorithmus
- 1970 und 1995 von Hastings und Green zur **Metropolis-Hastings-Green-Methode** generalisiert

Metropolis, Ulam. The Monte Carlo method, *J. Am. Statist. Assoc. (1949)*, 44 (247), 335-341

Monte-Carlo-Methode

Stanislaw Ulam



„The first thoughts and attempts I made to practice [the Monte Carlo Method] were suggested by a question which occurred to me in 1946 as I was convalescing from an illness and playing solitaires. The question was what are the chances that a Canfield solitaire laid out with 52 cards will come out successfully? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether a more practical method than "abstract thinking" might not be to lay it out say one hundred times and simply observe and count the number of successful plays. This was already possible to envisage with the beginning of the new era of fast computers, and I immediately thought of problems of neutron diffusion and other questions of mathematical physics, and more generally how to change processes described by certain differential equations into an equivalent form interpretable as a succession of random operations. Later … [in 1946, I] described the idea to John von Neumann, and we began to plan actual calculations."

Ensemblemittel mit MC

$$\rho(\mathrm{r},\mathrm{p}) = \frac{1}{Q} e^{-\frac{E(\mathrm{r},\mathrm{p})}{k_b T}}$$

$$Q = \frac{1}{N! h^{3N}} \int\int e^{-\frac{E(\mathrm{r},\mathrm{p})}{k_b T}} d\mathrm{r} d\mathrm{p}$$

- Für *k* Iterationen
  - Wähle zufälligen Punkt ($\mathbf{r}_k$, $\mathbf{p}_k$)
  - Bestimme zu mittelnde Größe $A_k(\mathbf{r}_k, \mathbf{p}_k)$
  - Bestimme Energie $E_k(\mathbf{r})$
  - Berechne Boltzmann-Faktor $\exp(-E_k(\mathbf{r}_k)/(k_B T))$
  - Addiere gewichtete Mittel von *A* auf
  - Addiere Boltzmann-Faktoren auf (zur Berechnung von *Q*)
- Berechne Ensemblemittel als

$$< A >= \sum_k A_k(\mathrm{r},\mathrm{p}) \rho_k(\mathrm{p},\mathrm{r})$$

Problem: zu viele sinnlose Konformationen ($\rho \sim 0$)!

Metropolis-Monte-Carlo

- Zustandssumme (und damit ρ_k) ist aufwändig zu berechnen
- Leicht dagegen: ρ_i / ρ_j

$$\frac{\rho_i}{\rho_j} = \frac{Qe^{-\frac{E_i}{k_B T}}}{Qe^{-\frac{E_j}{k_B T}}} = e^{-\frac{E_i - E_j}{k_B T}}$$

- **Idee**

  Erzeuge **Markov-Kette** von Zuständen, so dass Häufigkeit des Zustands H_i ! ρ_i

Markov-Kette

- Eine **Markov-Kette** ist ein stochastischer Prozess $X(t) = \{X_t,\ t\ 2\ N_0\}$, für den gilt
- $P\{X_{t+1} = j \mid X_t, X_{t-1}, \ldots X_0\} = P\{X_{t+1} = j \mid X_t\}$
- Aktueller Zustand also nur vom vorhergehenden Zustand abhängig (gedächtnislos)
- Übergangswahrscheinlichkeiten

$$\pi_{i!\,j}(t,\ t+1) := \pi_{i\,!\,j} = P\{X_{t+1} = j \mid X_t = i\}$$

Metropolis-Monte-Carlo

- Für eine Markov-Kette im Gleichgewicht gilt

  $H_i \, \pi_{i!j} = H_j \, \pi_{j!i}$ (*detailed balance*)

- Die Häufigkeiten H_i nähern ρ_i an, falls

$$\frac{\pi_{i \to j}}{\pi_{j \to i}} = \frac{H_j}{H_i} = \frac{\rho_j}{\rho_i} = e^{-\frac{E_j - E_i}{k_B T}}$$

- Metropolis-Algorithmus konstruiert eine entsprechende Markov-Kette
- Daher auch der Name **MCMC (*Markov Chain Monte Carlo*)** für die Methode

Gewöhnliches MC und MMC

- Häufig wird unter „Monte Carlo“ Metropolis-Monte-Carlo verstanden
- **MC**
  - Erzeugt gleichverteilte Zustände
  - Gewichtet diese mit Boltzmann-Faktor
- **MMC**
  - Erzeugt Zustände mit Boltzmann-Wahrscheinlichkeit
  - Zählt alle Zustände gleich
- Von hier an sei MC := MMC

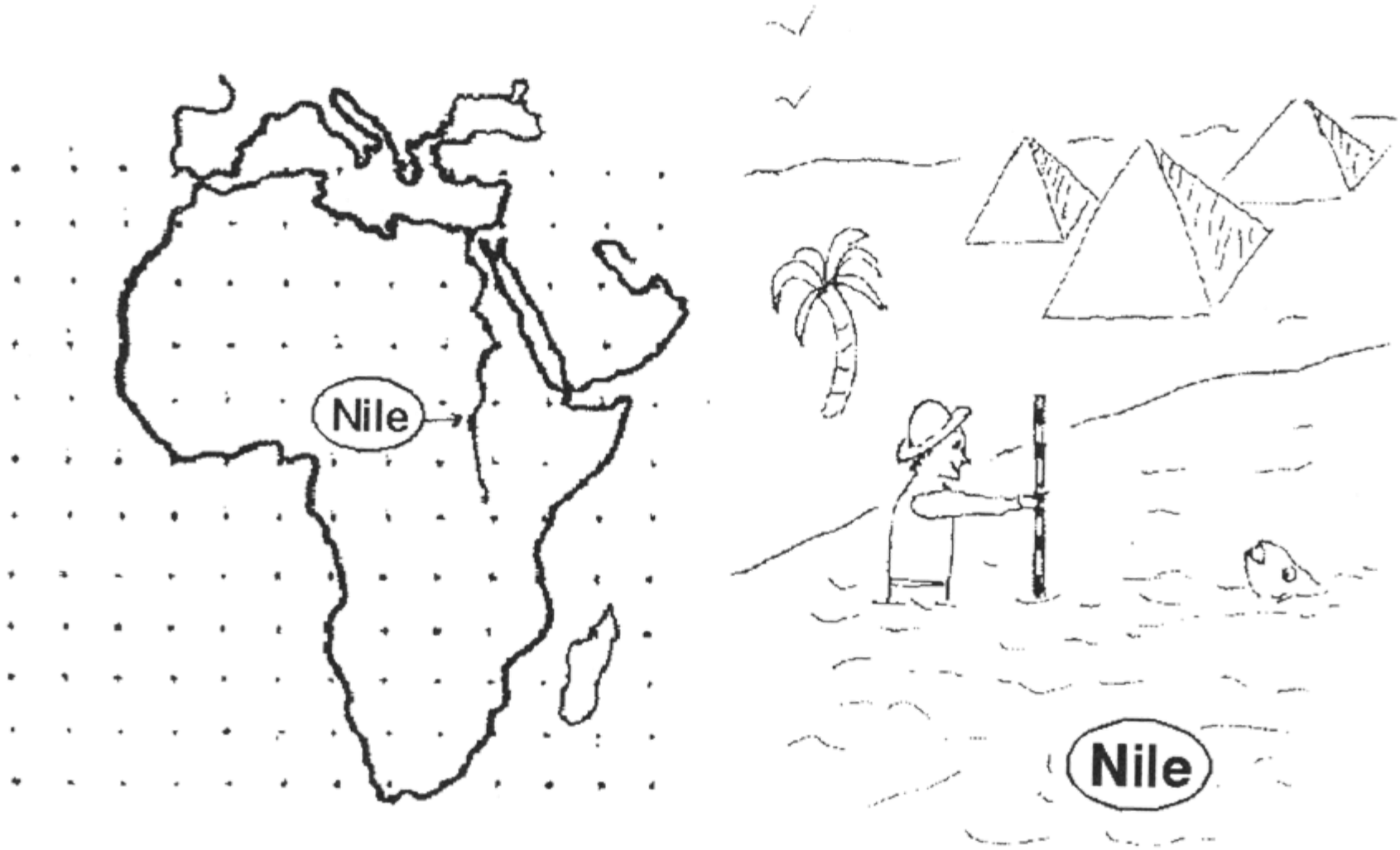
MMC-Algorithmus

- Für k Iterationen
  - Wähle zufälligen Punkt $\mathbf{r}_k$ im Konformationsraum
  - Falls $E_k \cdot E_{k\text{-}1}$
    - Akzeptiere neuen Punkt
  - Falls $E_k > E_{k\text{-}1}$
    - Wähle gleichverteilte Zufallszahl x 2 [0, 1]
    - Falls $x < \exp(-(E_k - E_{k\text{-}1}) / (k_B T)))$
      - Akzeptiere neuen Punkt
    - Andernfalls
      - Bleibe bei alter Konformation

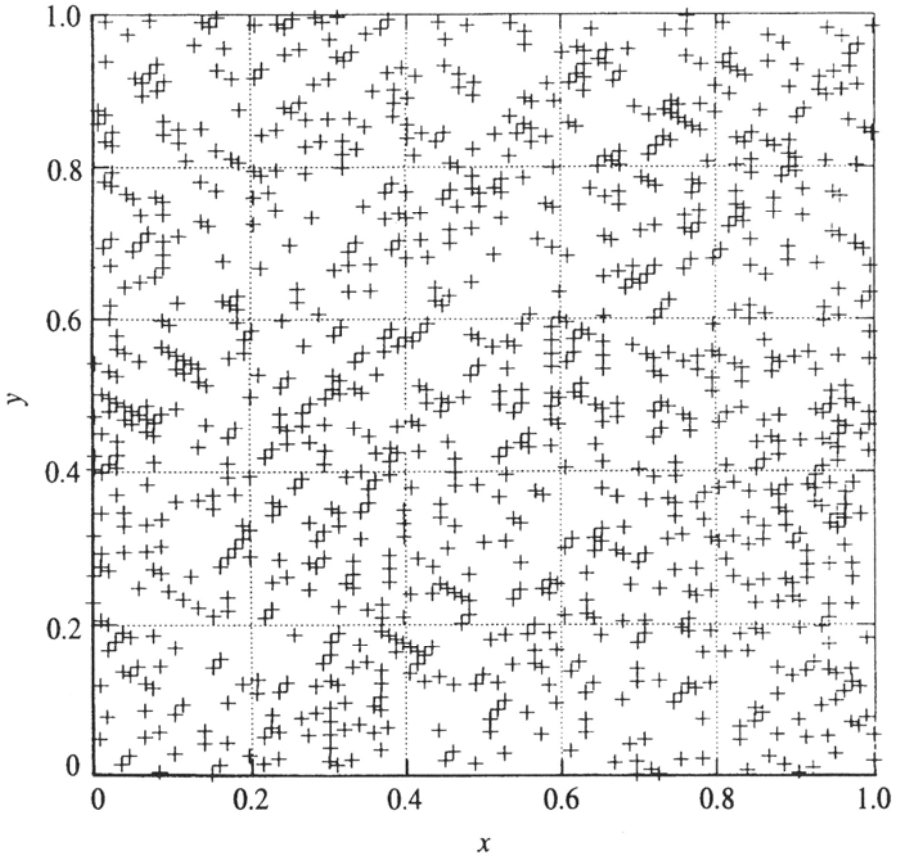
MMC-Algorithmus

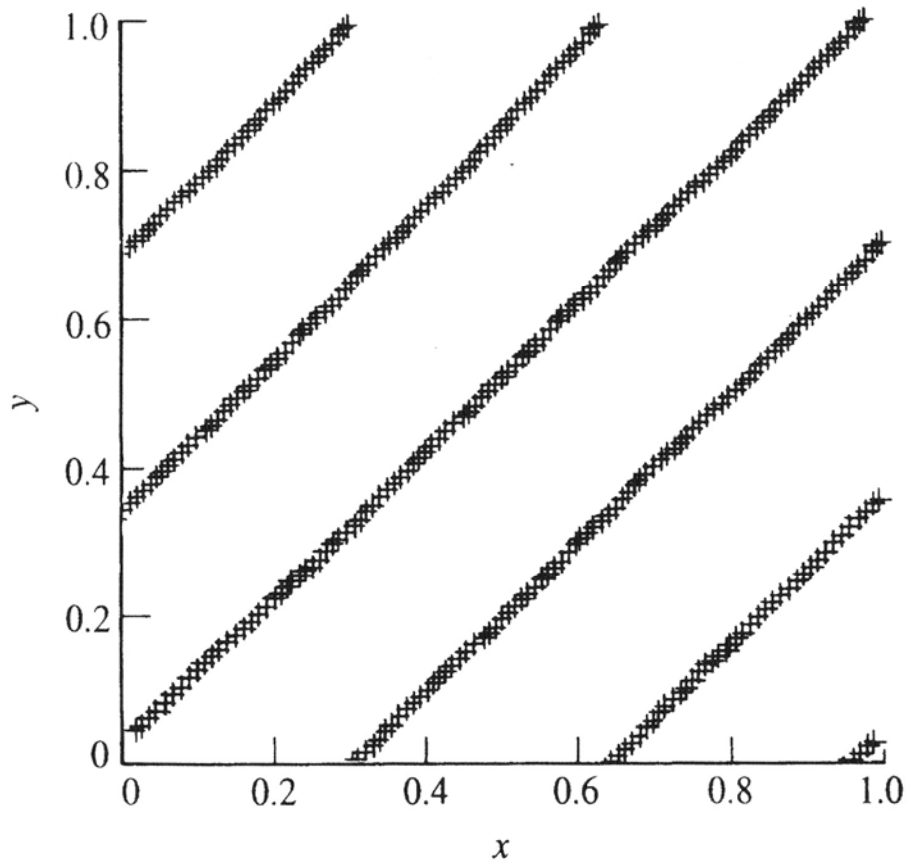
- Resultat einer MMC-Simulation ist eine Trajektorie $\mathbf{r}_k$ und zugehörige Energie E_k
- Im Gegensatz zu einer MDS besteht **kein zeitlicher Zusammenhang** zwischen aufeinander folgenden Punkten der Trajektorie
- Neben der Bestimmung von Mittelwerten kann man selbstverständlich auch Minima damit bestimmen

Systematische Suche vs. MC



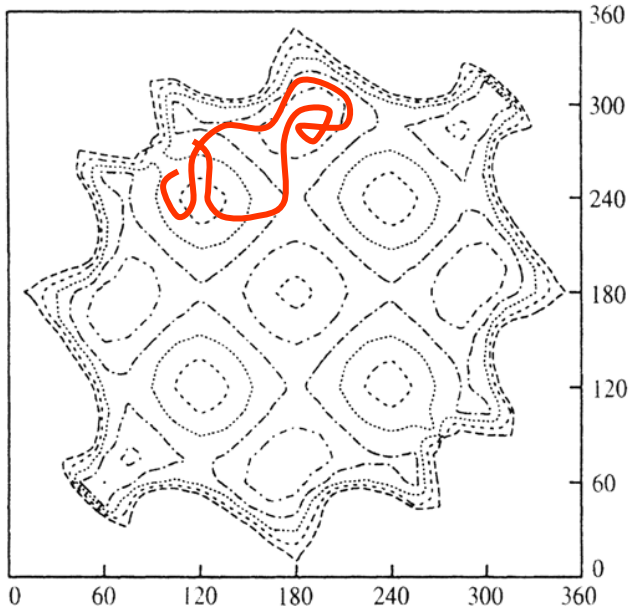
Zufallszahlengeneratoren

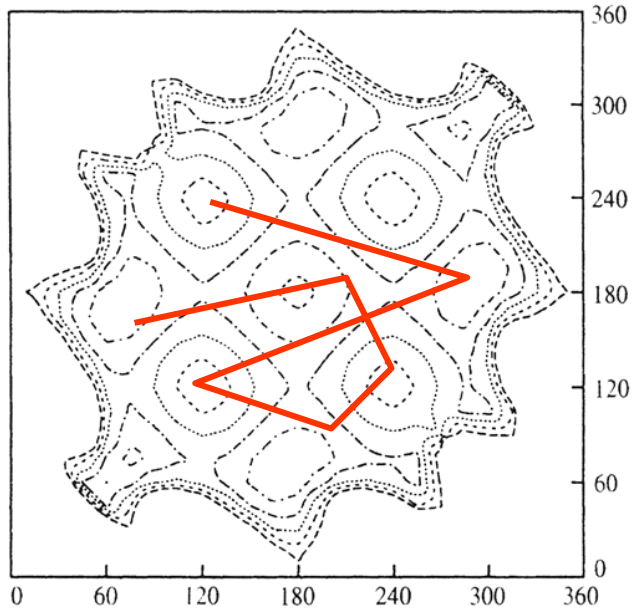




- Viele „Zufallszahlen“-Generatoren weisen für lange MC-Simulationen (und auch MD-Simulationen mit zufälligen Stößen) keine ausreichende Qualität auf
- Auswahl des Generator kritisch – immer überprüfen!

Vergleich MD/MC





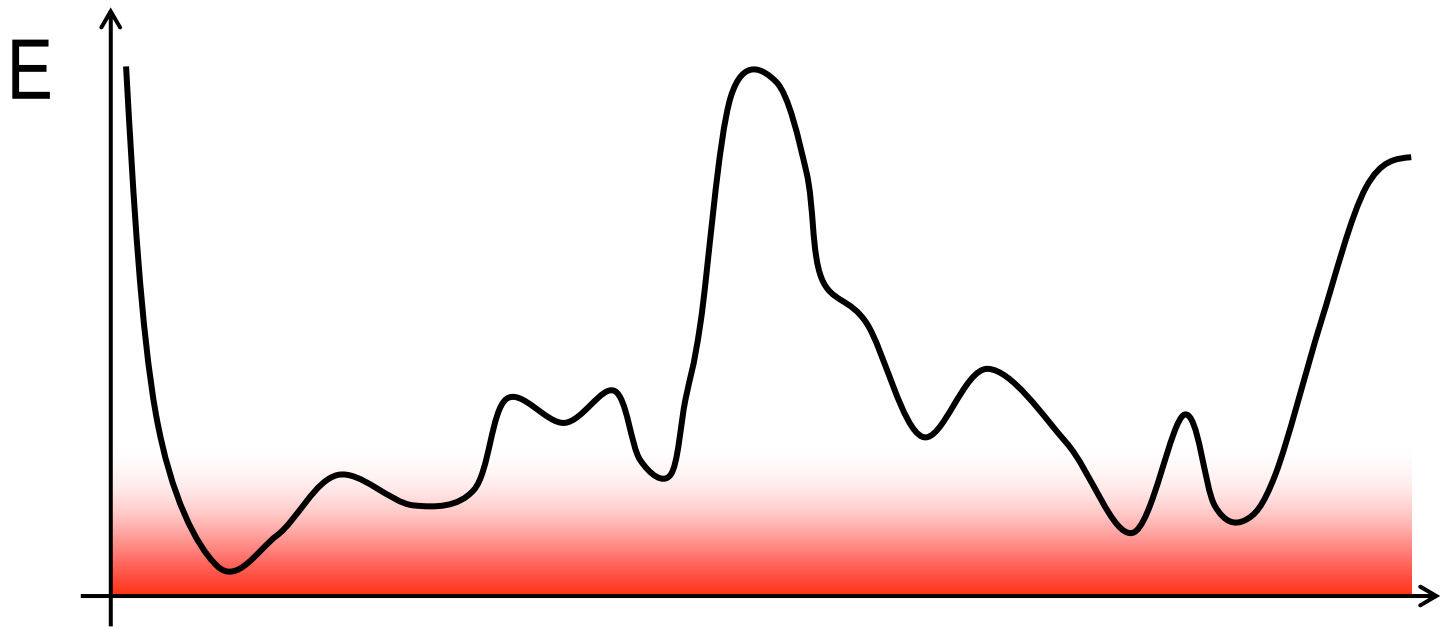
Im Konformationsraum:

- MD: glatte Kurven
- MC: Sprünge

Vergleich MD/MC

- MD gut zu Berechnung zeitabhängiger Größen (z.B. Diffusionskoeffizienten)
- MC zeigt häufig schneller Konvergenz bei der Simulation thermodynamischer Größen
- MD erforscht besser lokale Umgebung
- MC erforscht globalen Konformationsraum besser
- MD und MC komplementär
- Methoden die beide Ansätze kombinieren sind auch in Gebrauch

Einfluss der Temperatur

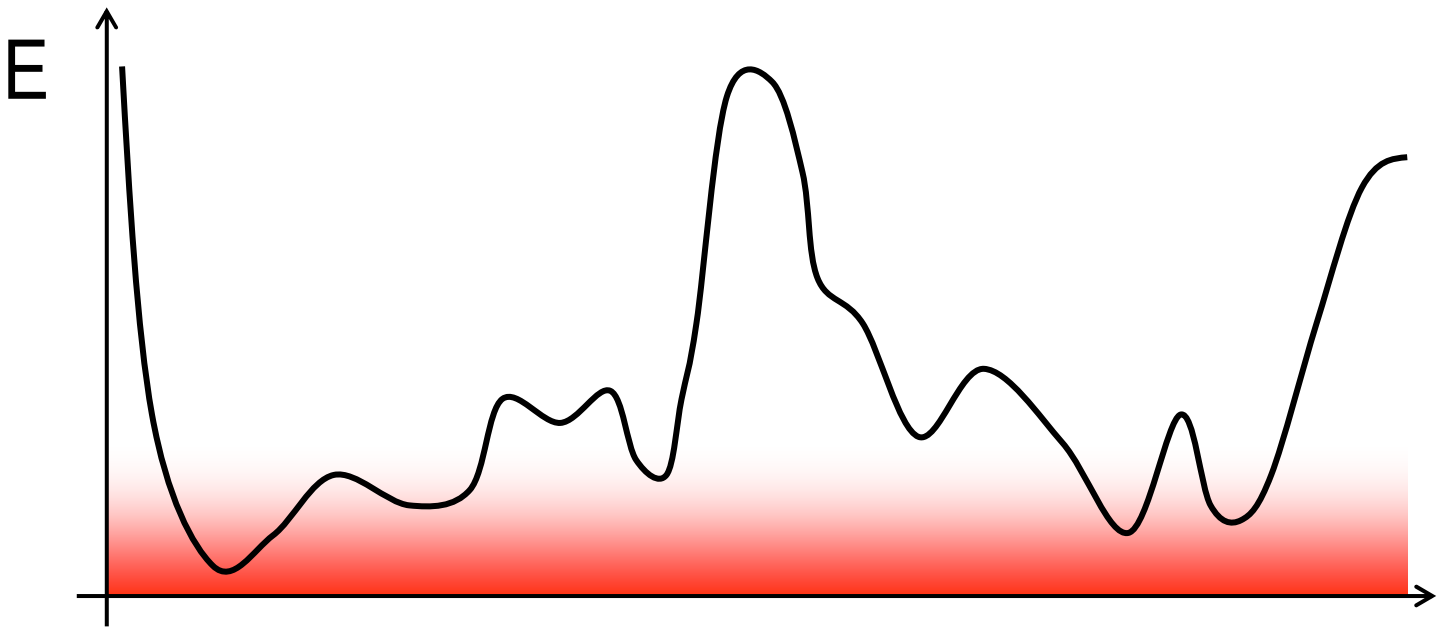


- Temperatur entspricht Energie:

  $E_{kin} = 3/2\ RT$

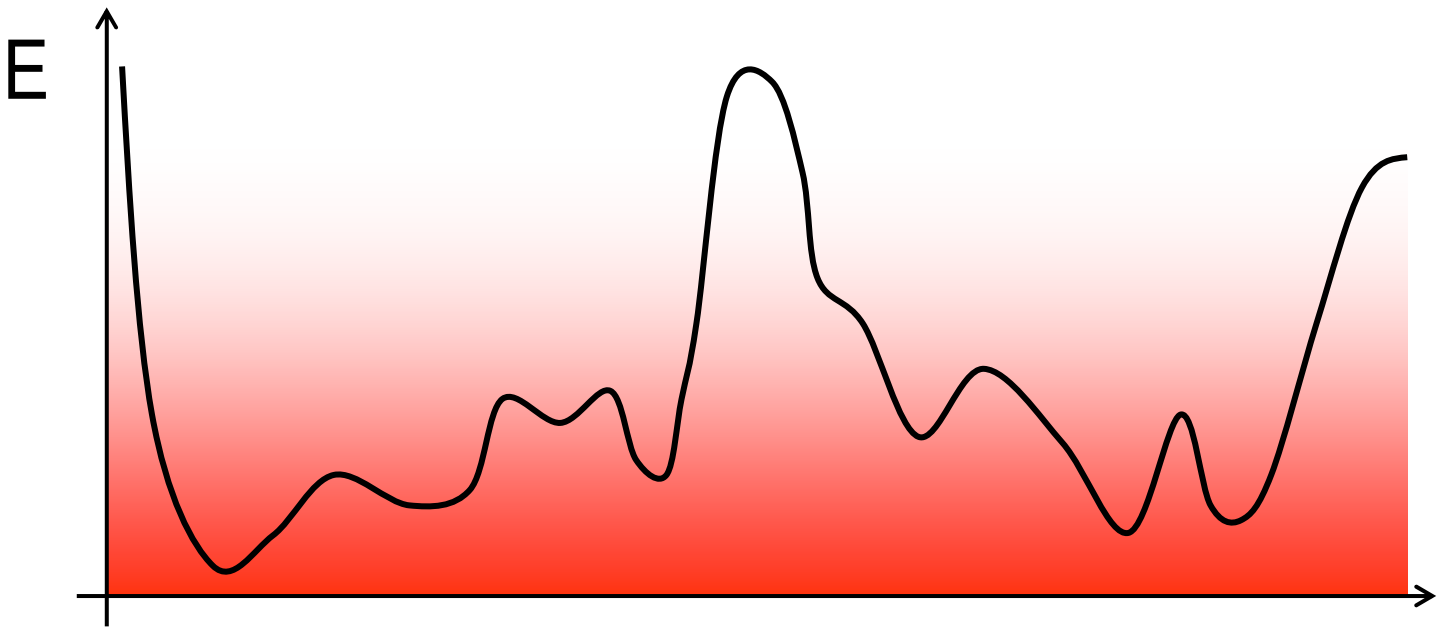
- Mit der Gaskonstante $R = 8.314\ J/(K\ mol)$ ergibt sich für Raumtemperatur (298 K): $E_{kin} = 3.7\ kJ/mol$

Einfluss der Temperatur



- Maxima die tiefer oder auf Höhe der Temperatur liegen, stellen für die Simulation kein Problem dar.
- Höhere Maxima werden mit einer gewissen Wahrscheinlichkeit erst überschritten, wenn die Temperatur entsprechend hoch ist.

Einfluss der Temperatur

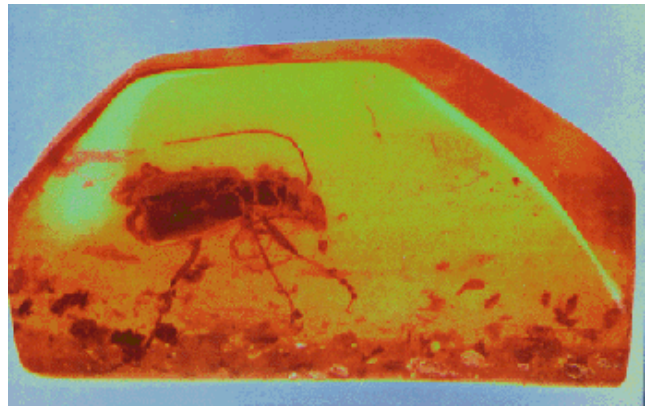


- Maxima die tiefer oder auf Höhe der Temperatur liegen, stellen für die Simulation kein Problem dar.
- Höhere Maxima werden mit einer gewissen Wahrscheinlichkeit erst überschritten, wenn die Temperatur entsprechend hoch ist.

Existierende Implementierungen

- **AMBER** – Der Klassiker von Kollman
- **CHARMM** – Der Klassiker aus dem Hause Karplus
- **GROMACS** – Superschnell!
- **MMTK** – Für Python-Fans und sonstige Bastler
- **BALL** – Für C++-Fans

AMBER



- **A**utomated **M**odel **B**uilding and **R**efinement
- Original-Implementierung des klassischen AMBER-Kraftfelds
- Ursprünglich in der Gruppe von Peter Kollman, UCSF, jetzt: Scripps, La Jolla
- In Fortran, mittlerweile mit Teilen in C
- Recht schnell, parallelisiert
- Grausame Formate
- MD-Simulation, Minimierung, Konformationssuche, Trajektorienanalyse

`http://amber.scripps.edu`

CHARMM

- **C**hemistry at **HAR**vard **M**olecular **M**echanics
- Entwickelt in der Gruppe von Martin Karplus (Harvard)
- Kommerziell vertrieben
- MD-Simulation, Energieminimierung
- Fortran-Code

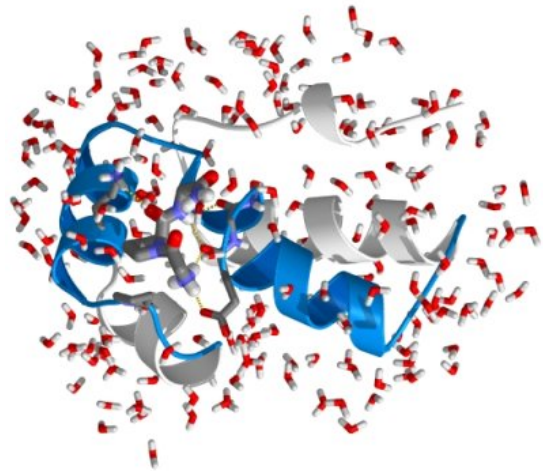
http://www.accelrys.com/support/life/charmm/
http://www.scripps.edu/brooks/c29docs/Charmm29.Html

GROMACS

- MD-Simulation, Minimierung
- GROMOS-Kraftfeld

  (*GROMOS: GROningen MOlecular Simulation software*)
- Schnellster MD-Code
  - Handoptimierter Assembler-Code
  - Sehr effiziente Algorithmen
  - Implementierungstricks
- Sehr effizient parallelisiert
- Open Source (GPL)
- 3-10x schneller als anderer Code

`http://www.gromacs.org`



MMTK

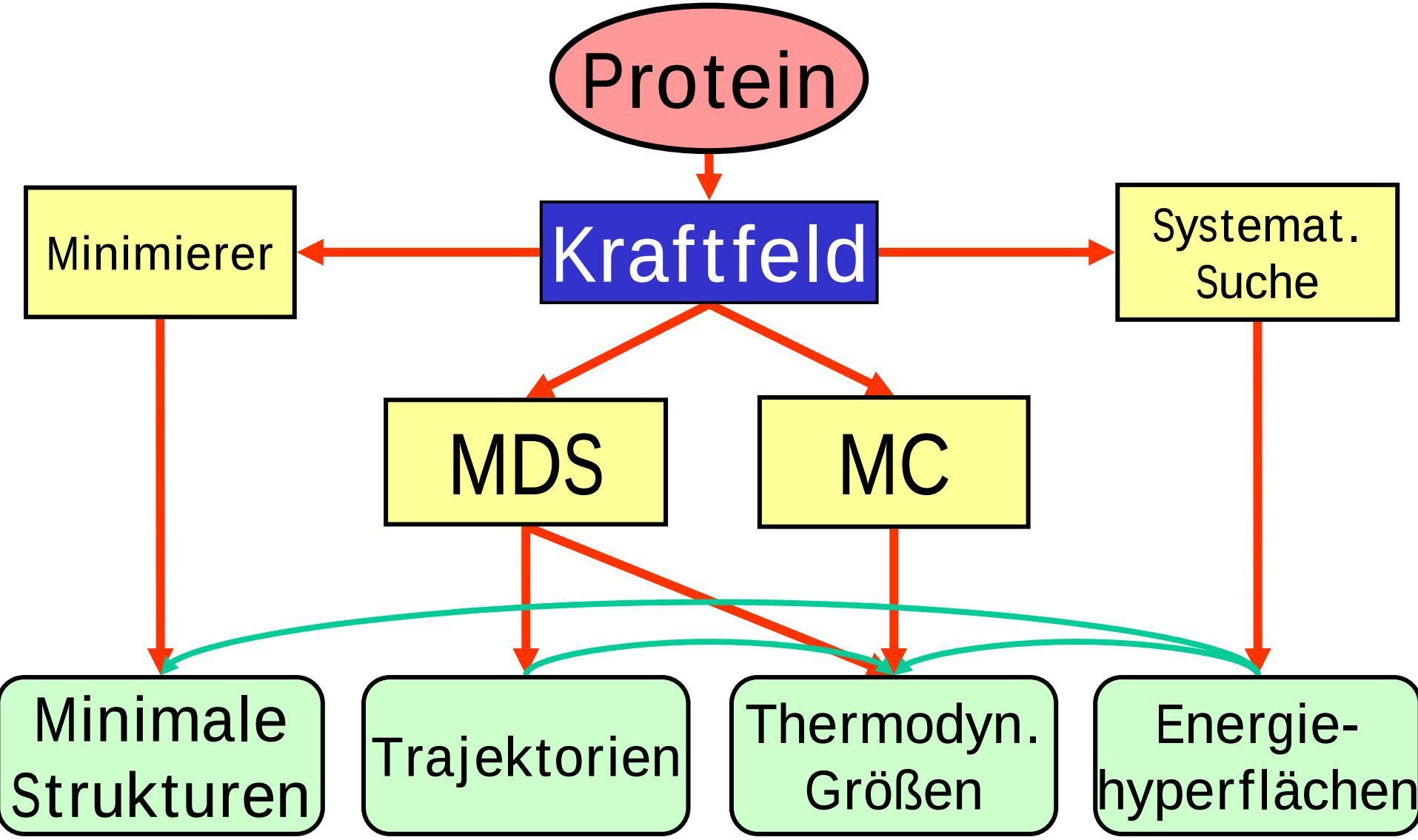
- Molecular Modeling Tool Kit
  (Konrad Hinsen, CNRS, Orlèans)
- Sehr beschränkte Funktionalität
  - AMBER
- Open Source
- Implementiert in Python, C
- Sehr einfach zu bedienen
- AMBER94-Kraftfeld, nicht sehr effizient
- Sehr gut zum Lernen und für einfache Experimente

`http://starship.python.net/crew/hinsen/MMTK/`

Zusammenfassung MM

- Molekülmechanik erlaubt **Vorhersage/Simulation** von
  - Geometrien
  - Energien, thermodynamischen Größen
  - Dynamischen Eigenschaften
- Ein geeignetes **Kraftfeld** liefert
  - Energien
  - Kräfte (Gradienten)
- Kraftfelder unterscheiden sich bezüglich
  - Güte (Klasse I/II/III)
  - Anwendbarkeit (Proteine, DNA, kleine Moleküle)

Übersicht Molekülmechanik



Literatur

Molekülmechanik

- **Andrew R. Leach, Molecular Modelling – Principles and Applications**, Prentice Hall, 2001
- **Daan Frenkel, Berend Smit, Understanding Molecular Simulation**, Academic Press, 1996
- **Martin J. Field, A practical introduction to the simulation of molecular systems**, Cambridge University Press, 1999
- **Tamar Schlick, Molecular Modeling and Simulation**, Springer, 2003
- **Ulrich Burkert, Norman L. Allinger, Molecular Mechanics**, American Chemical Society, 1982