Proteinstrukturen: Faltung, Vergleich und Vorhersage

Annalisa Marsico

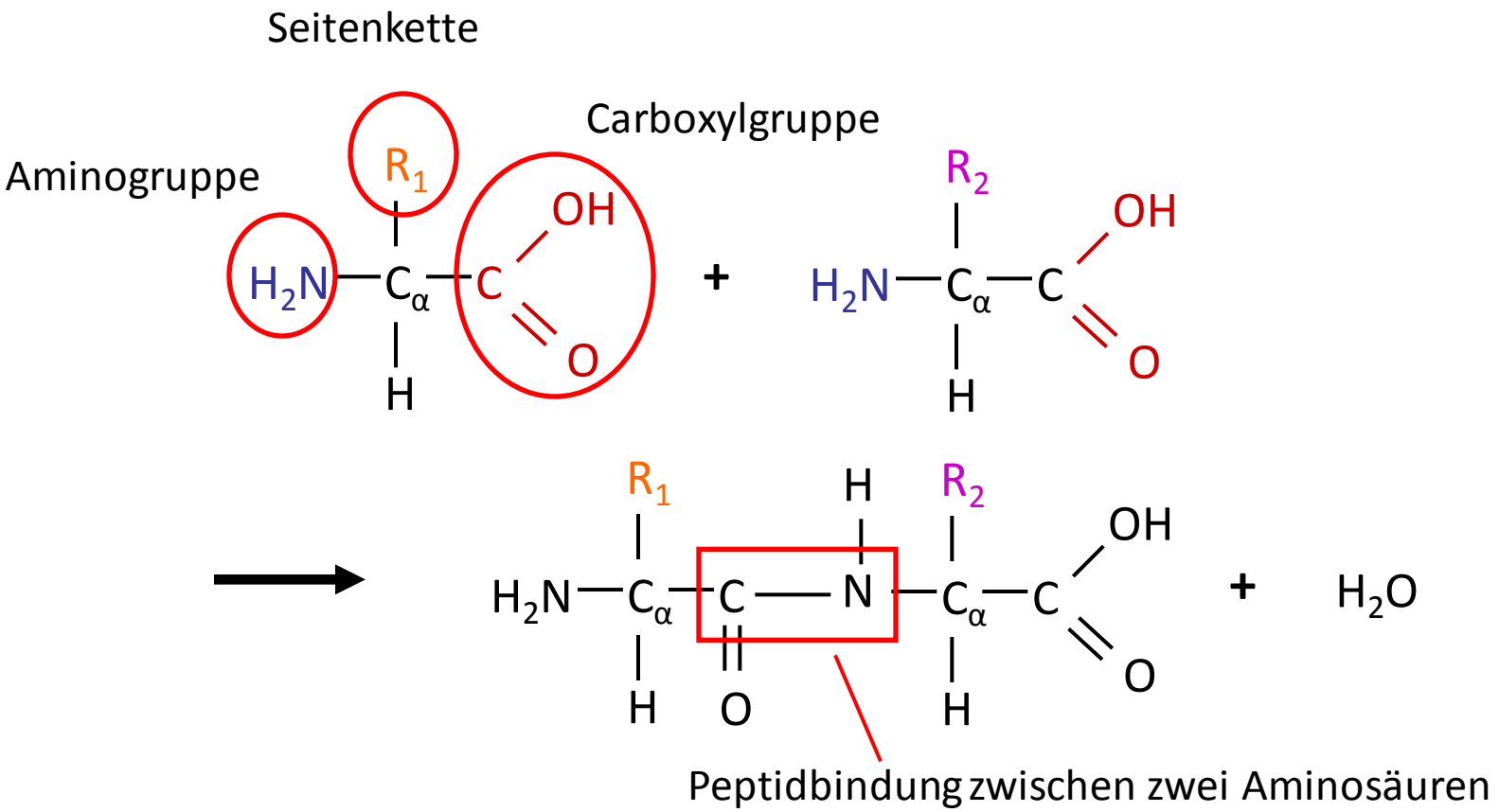
OWL RNA Bionformatics group

Max Planck Institute for Molecular Genetics

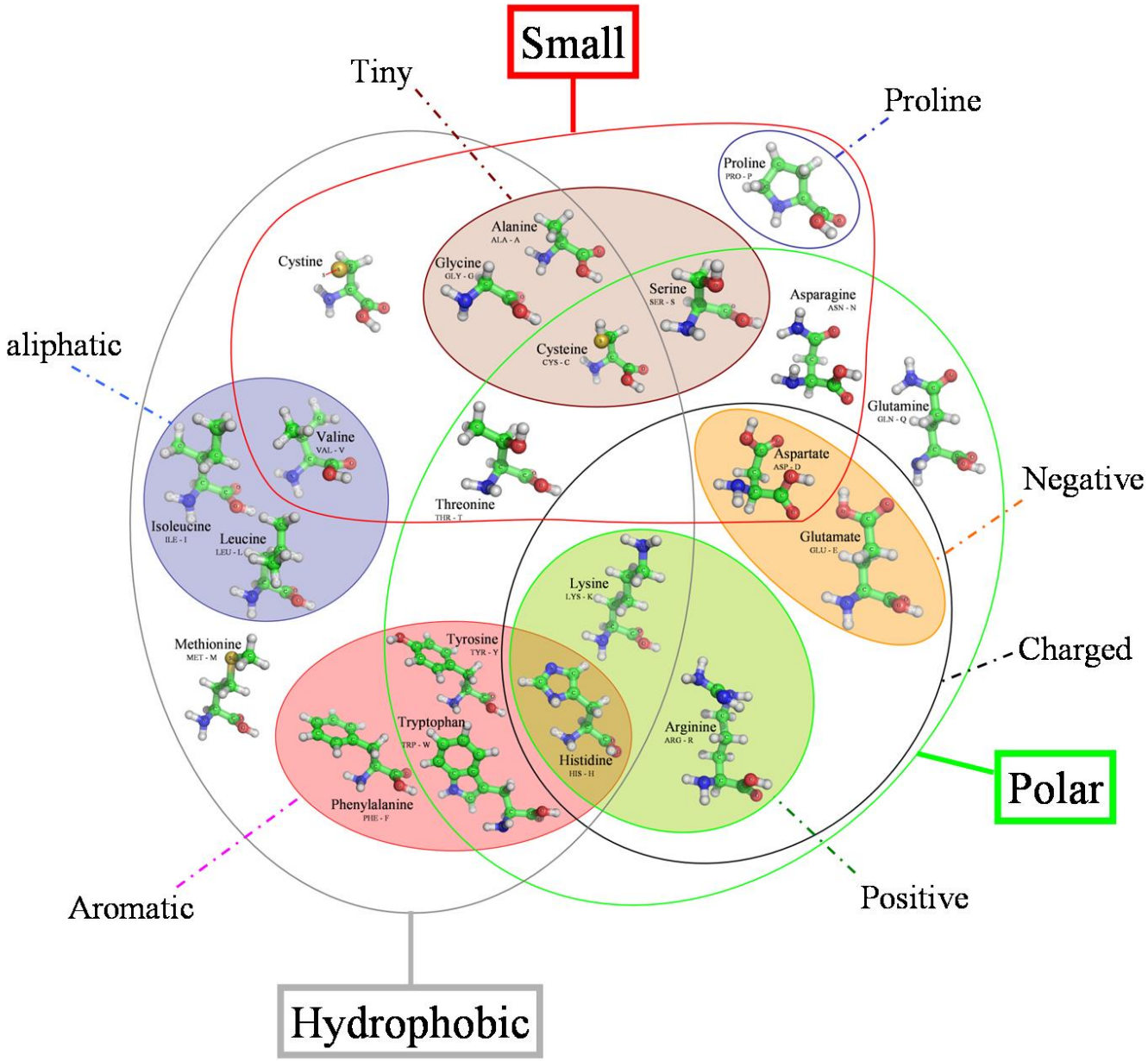
Freie Universität Berlin

19/01/2015

Proteinstrukturen: Einführung

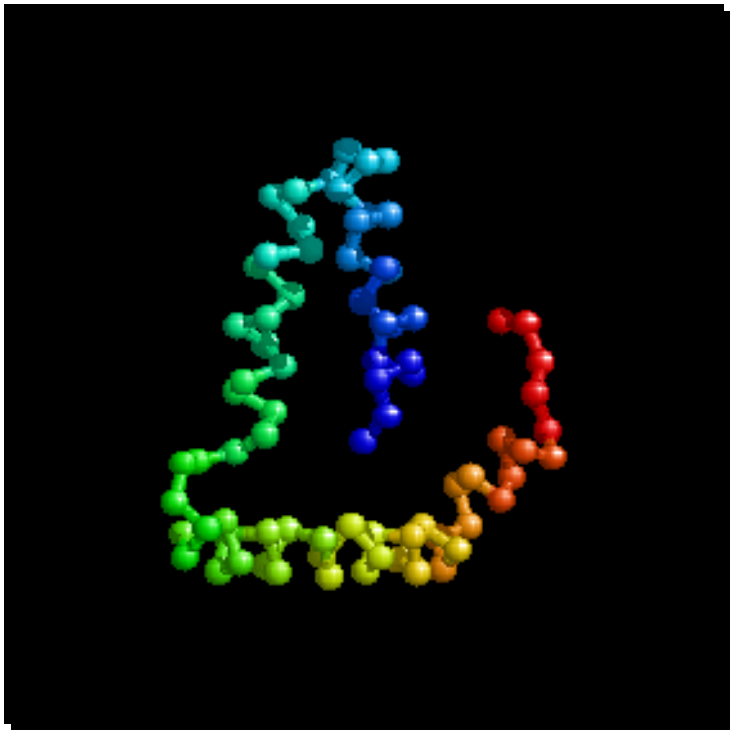


Klassifizierung von Aminosäuren



Proteinfaltung

- Proteine sind lineare Polymere mit unterschiedlichen Aminosäureseitenketten
- 3D-Struktur: Anordnung der Atome im 3D Raum
- Proteine falten sich spontan in einen Zustand **minimaler Energie**
- Seiten- und Hauptketten interagieren miteinander und mit dem Lösungsmittel
- Proteinstrukturen sind dynamische Objekte
- **Die Anzahl von Wasserstoffbrückenbindungen** wird **maximiert** während des Faltens



Proteinfaltung

Einen grossen Einfluss auf die Faltung haben folgende Eigenschaften

- Wasserstoffbrückenbindungen zwischen **polaren Seitenketten** und Wassermolekülen (hydrophile Aminosäuren)
- Interaktionen zwischen hydrophoben Aminosäuren
- Wasserstoffbrückenbindungen zwischen polaren Atomen im Rückgrat (Sekundärstruktur Elemente, alpha-Helices und beta-Stränge)
- Andere Faktoren (schwache Interaktionen)

Warum Strukturen, wenn wir bereits Sequenzen haben?

- In evolutionär-verwandten Proteinen sind Strukturen stärker konserviert als die Sequenzen
- Strukturmotive können ähnliche Funktionen vorhersagen
- Ein Einblick in Proteinfaltung ermöglicht Fehlfaltungen zu verstehen (ist die Ursache vieler Krankheiten)

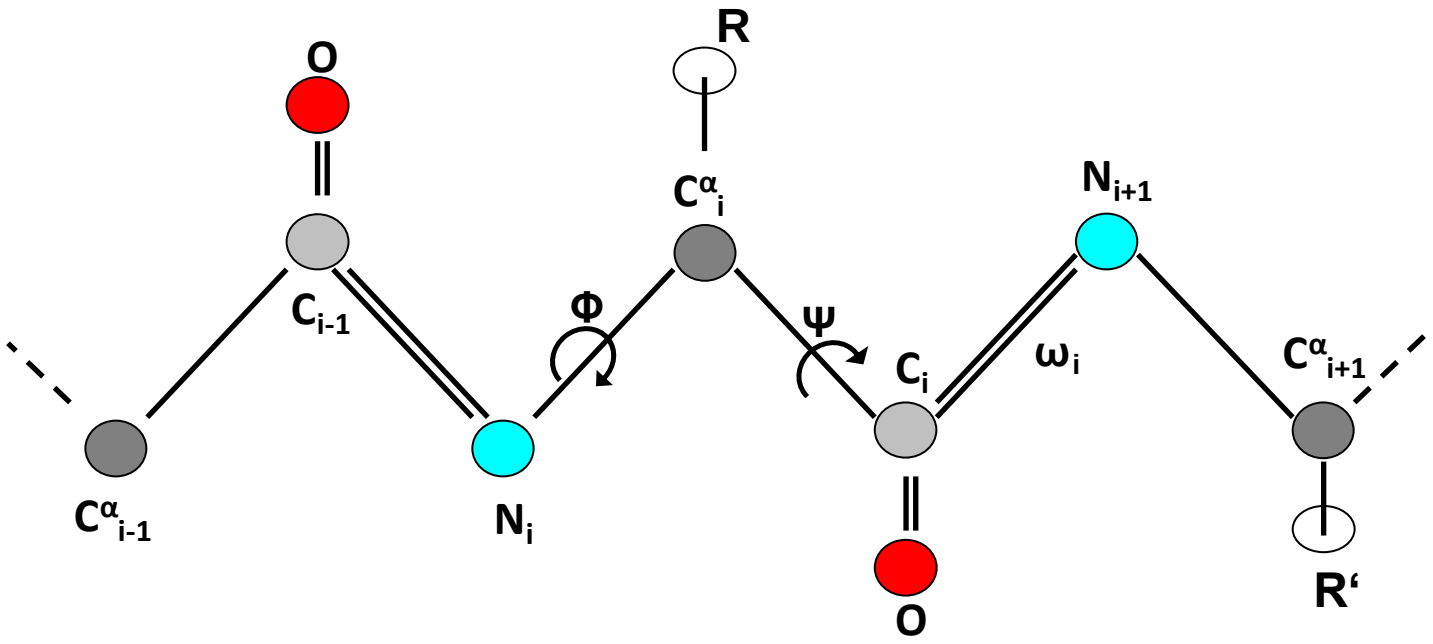
Warum brauchen wir automatisierte Algorithmen?

- Die steigende Zahl von großen Strukturdatenbanken erlaubt keine manuelle (visuelle) Analyse mehr und erfordern effiziente 3D-Suche und Klassifikationsverfahren .

- Structural Genomics Aufwand

Proteingerüst: Torsionswinkel

Geometrisch ist das Rückgrat eines Protein eine Folge von Atomen in Raum
.....$C_{i-1} = N_i - C_i^{\alpha} - C_i = N_{i+1} - C_{i+1}^{\alpha} - C_{i+1}$,.....



Φ Winkel um die Bindung $N_i - C^{\alpha}_i$
Ψ Winkel um die Bindung $C^{\alpha}_i - C_i$
ω Winkel um die Bindung $C_i - N_{i+1}$. Die Peptidbindung ist planar und in einem der beiden Zuständen:
trans **ω≈180° (meistens)** und *cis,* **ω≈0°** (selten, Proline)

Ramachandran Plot

- **Linien = energetisch bevorzugt (rot referenziert spezielle sekundäre Strukturelemente)**
- **Ausserhalb der Linie = nicht erlaubt**
- Die meisten Aminosäuren fallen in α_R Regionen (*right-handed alpha helix*) oder β Regionen (*beta-strand*)
- Glycine hat zusätzliche Konformationen (e.g. left-handed alpha helix = α_L Region) und Konformationen im rechten unteren Panel

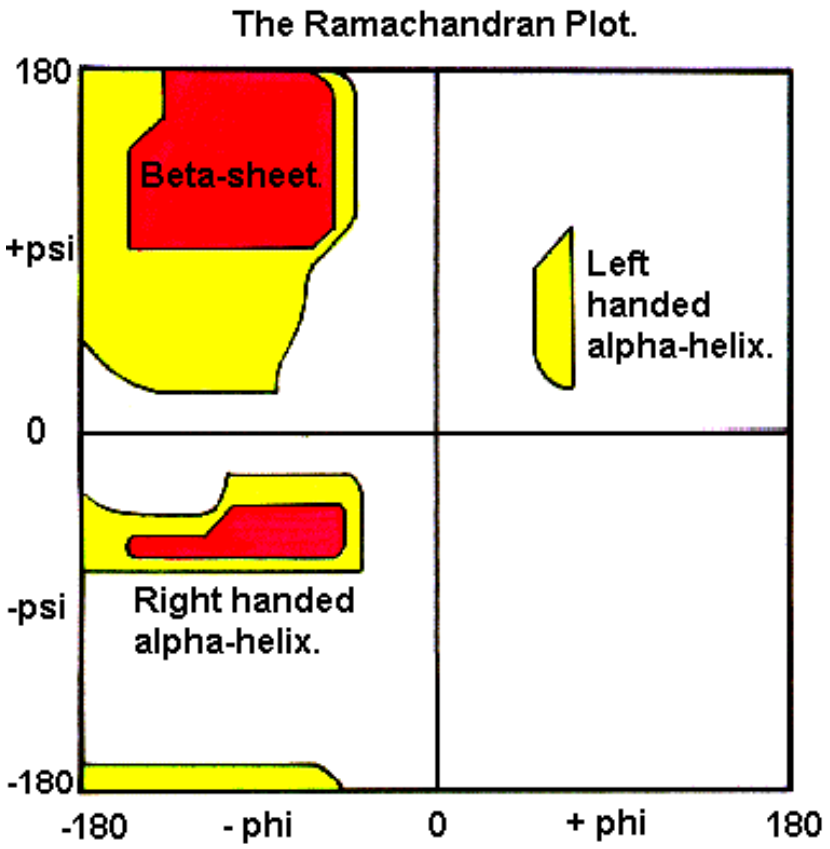


Image taken from www.expasy.org/swissmod/course

Ramachandran Plot

Plot eines Proteins
mit überwiegend Beta-Faltblätter

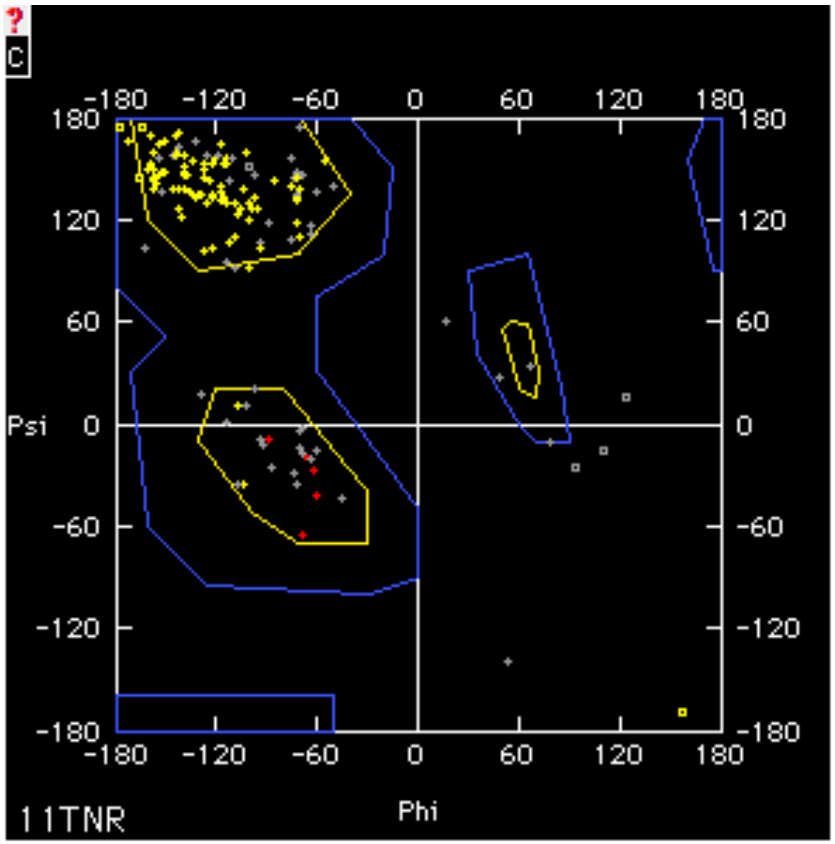
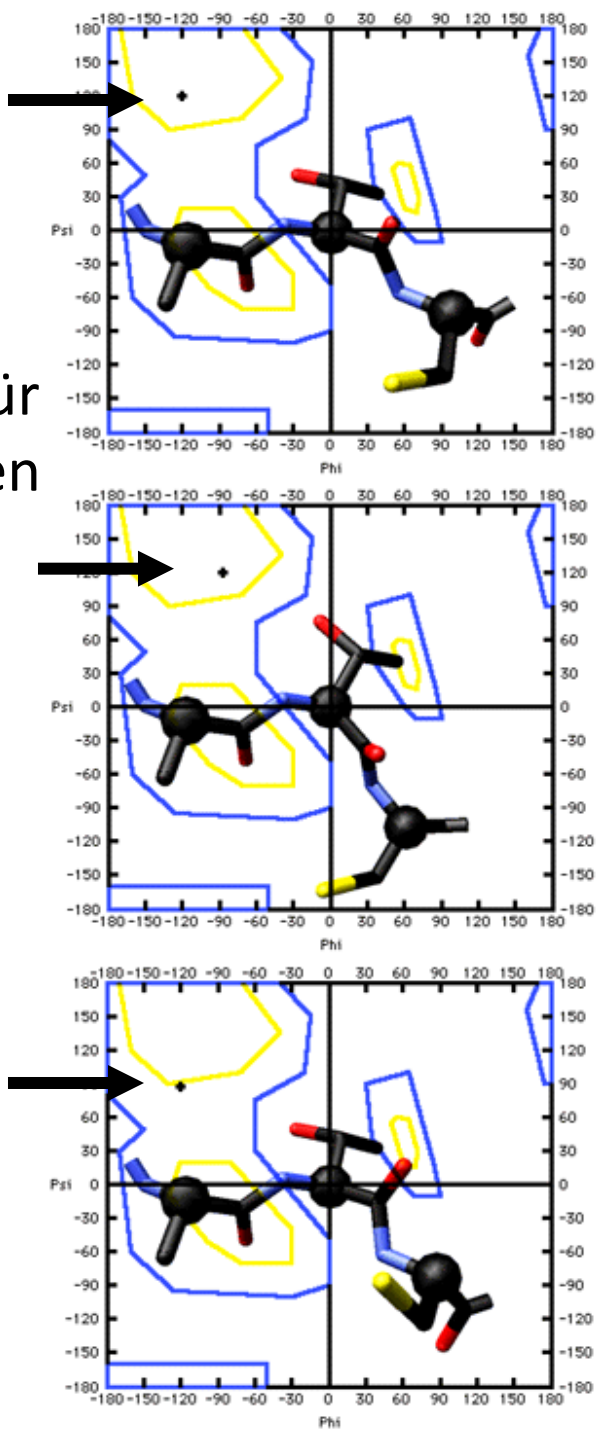


Image taken from www.expasy.org/swissmod/course

Beispiele für
Konformationen

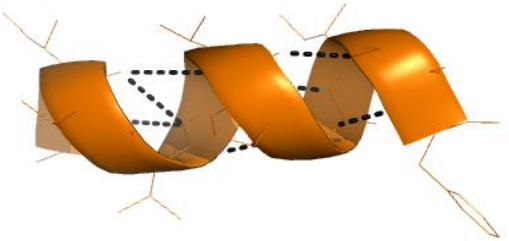


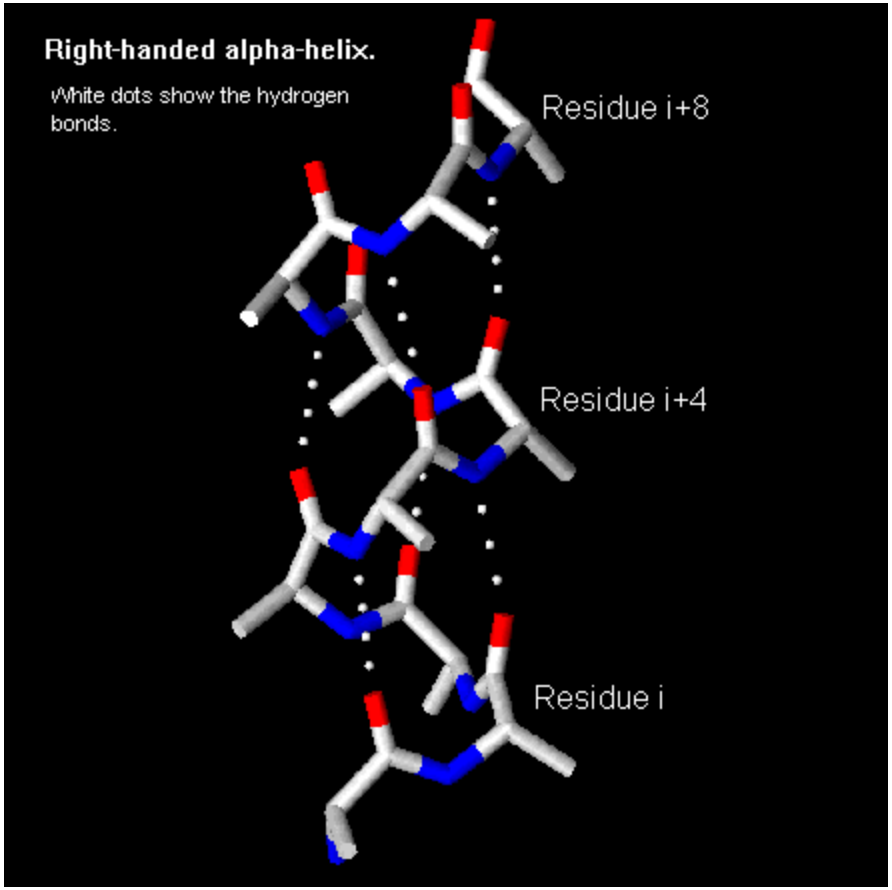
Sekundärstrukturelemente (SSEs)

Helices und Stränge

- Aufeinanderfolgende Reste in alpha- oder beta-Konformation erzeugen **alpha-Helices und beta-Stränge**
- Solche Sekundärstrukturelemente werden durch schwache **Wasserstoffbrücken** stabilisiert
- Sie werden durch Windungen (Turns) und Schleifen verbunden (Regionen, in denen die Kette die Richtung ändert)
- Wendungen sind oft an der Oberfläche ausgesetzt und enthalten polare Reste

Alpha Helix



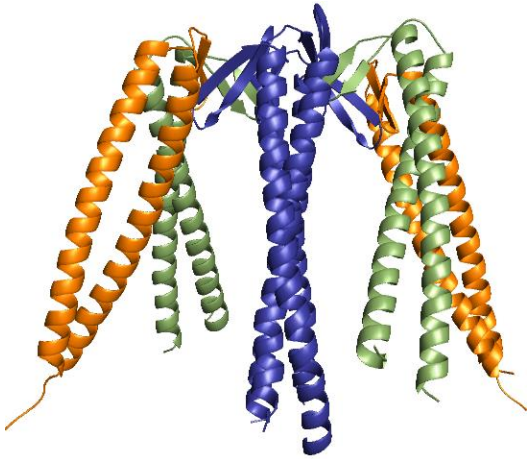


- Rest *i* bildet eine Wasserstoffbrücke mit Rest *i+4*
- 3.6 Reste pro Drehung
- 1.5 Å Anstieg pro Runde
- Ideal Torsionswinkel:

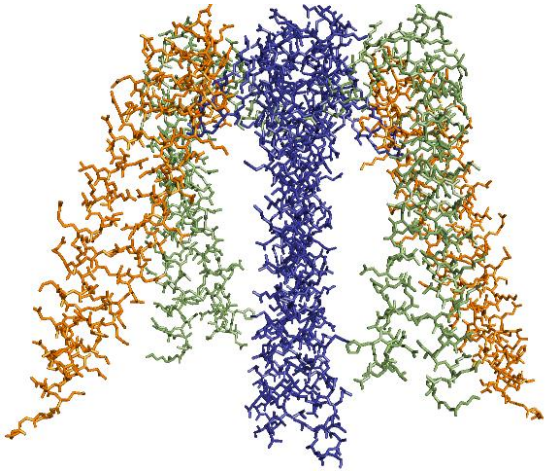
ϕ =~ -60 °

ψ =~ -45 °

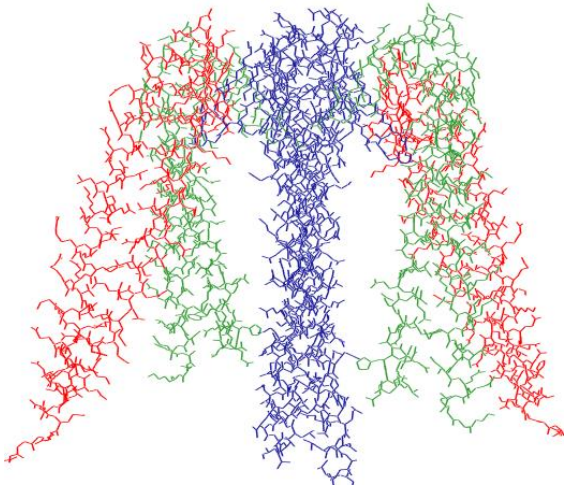
Beispiel 1- Proteine hauptsächlich aus alpha-Helices geformt



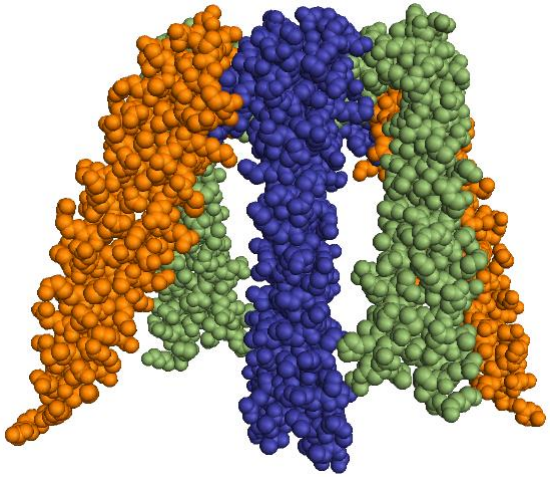
‚Cartoon‘ Darstellung



‚Stick‘ Darstellung

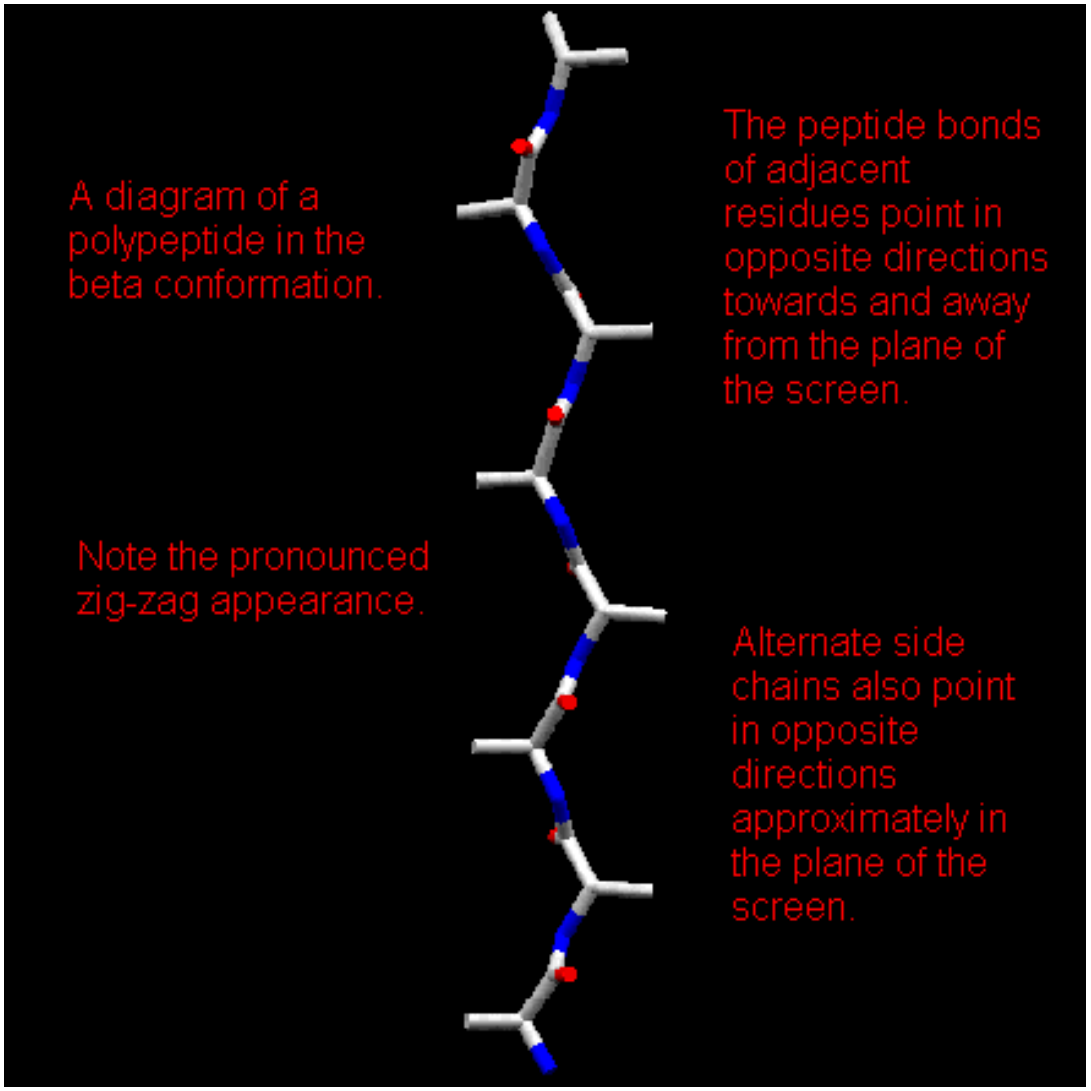


‚Wires‘ Darstellung



‚Balls‘ Darstellung

Beta Strang

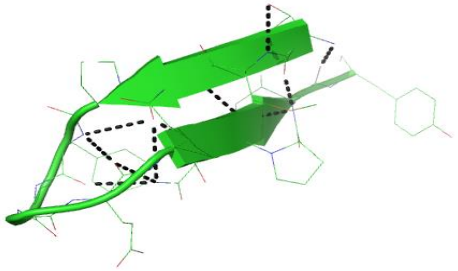


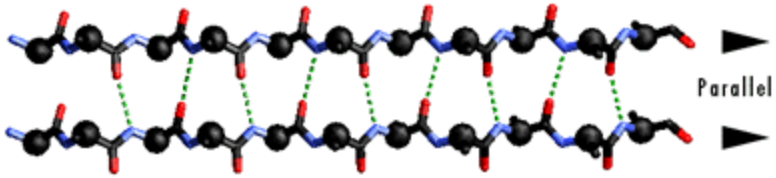
ϕ = -140 °,

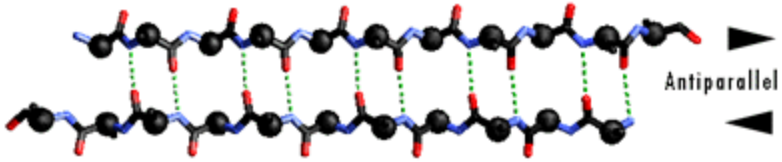
ψ = 130 °

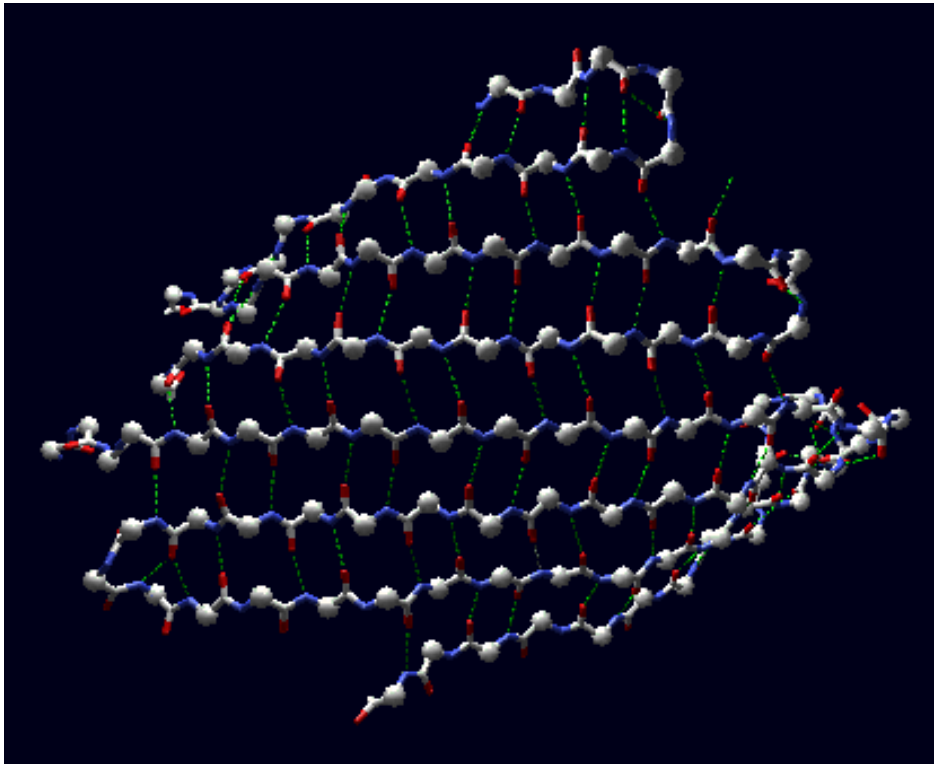
Image taken from www.expasy.org/swissmod/course

Beta Blätter

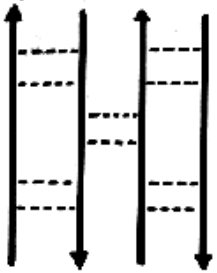




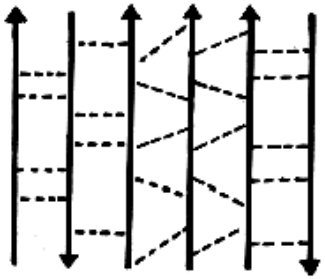




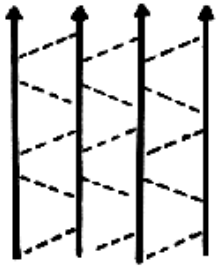
Antiparallel beta-sheet



The different types of beta-sheet. Dashed lines indicate main chain hydrogen bonds.

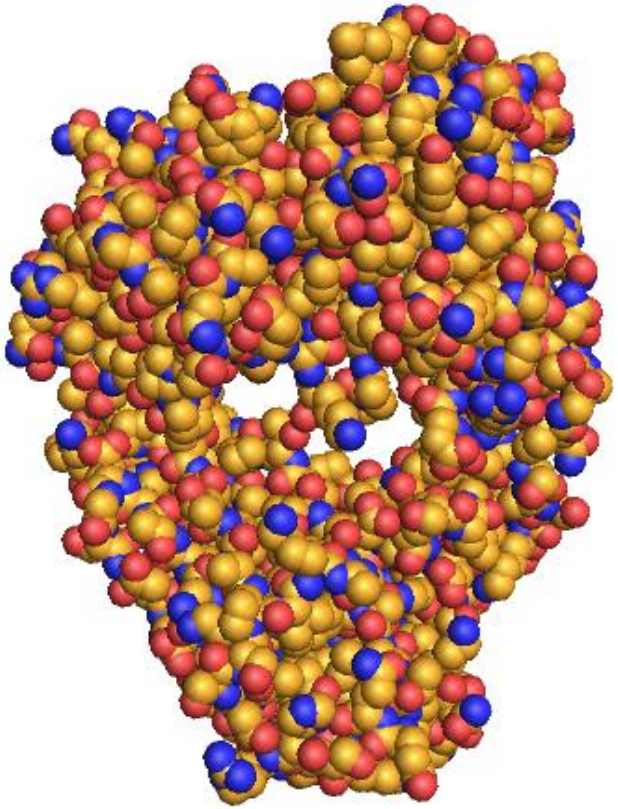


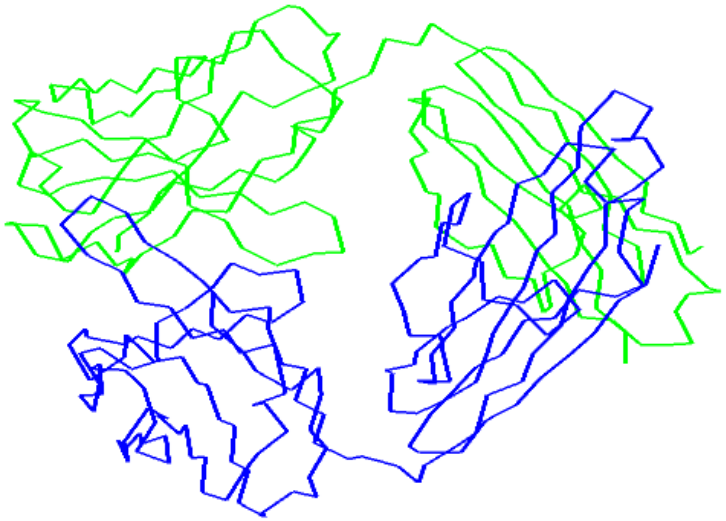
Mixed beta-sheet

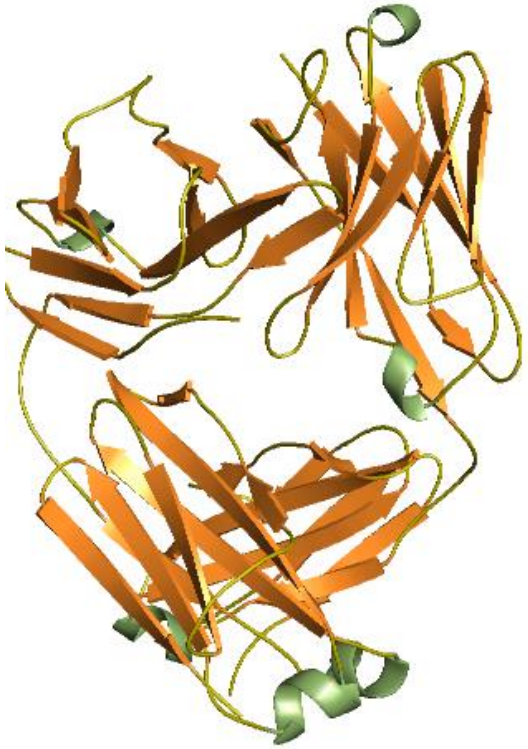


Parallel beta-sheet

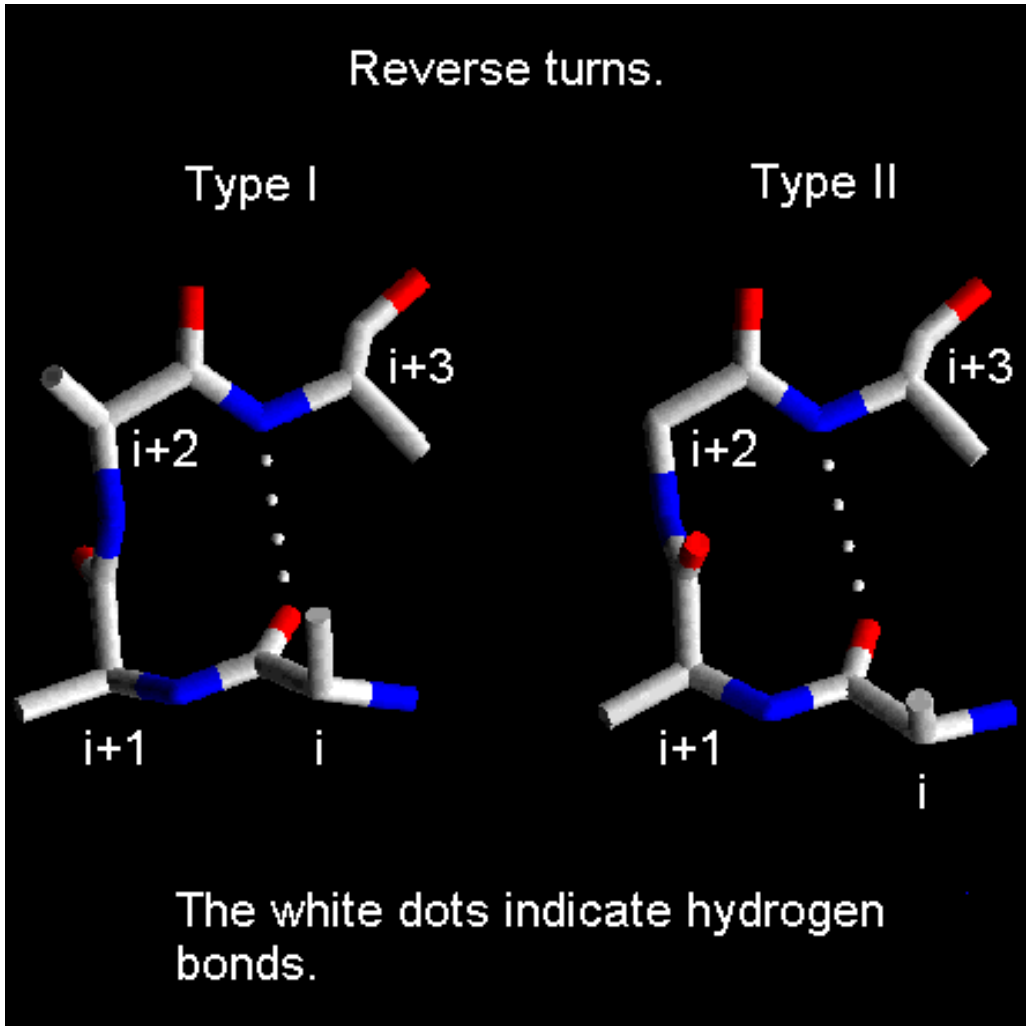
Beispiel 2- Proteine hauptsächlich aus beta-Faltblätter geformt







Turns



- Rest *j* ist um Rest *j+3 gebunden*
- Typ II: oft proline and glycine

Formale Definition von alpha-Helices

- Sei Hbond (i, j) eine Wasserstoffbindung zwischen der CO-Gruppe des Restes i und der NH-Gruppe des Restes j.

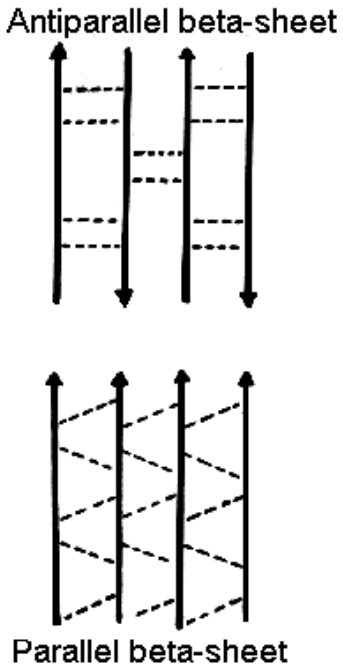
- Helices werden zwischen benachbarten Resten gebildet
- Drei verschiedene Arten von Helices :
  - **α-helix**: Hbond(i,i+4), Hbond(i+1,i+5),...

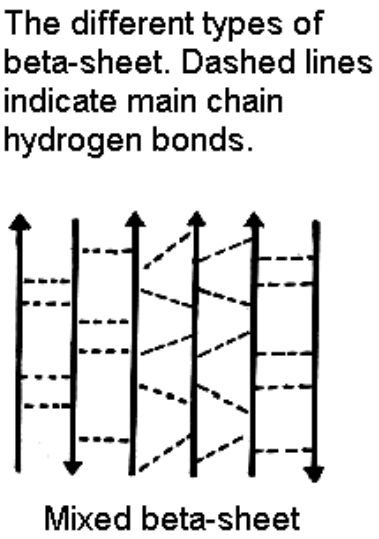
    Durchschnittliche Länge: 10 Reste (häufigste Art)
  - **3_{10}-helix**: Hbond(i,i+3), Hbond(i+1,i+4),..
  - **π-helix**: Hbond(i,i+5), Hbond(i+1,i+6),...

Formale Definition der Beta-Faltblätter

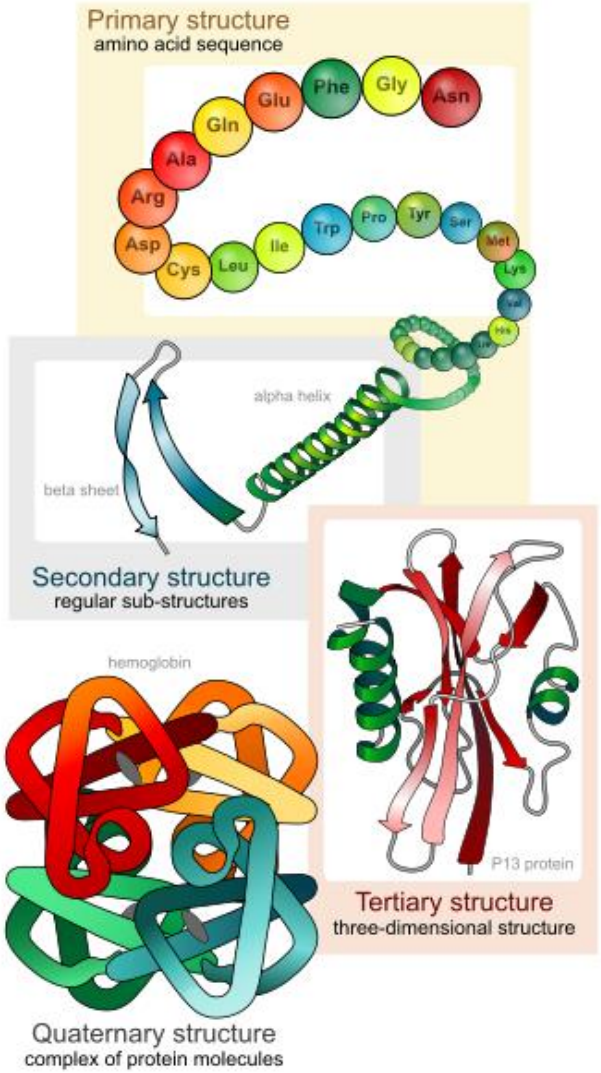
- Stränge und Blätter werden durch sukzessive Wasserstoffbindungen zwischen weitentfernten Resten in der Sequenz gebildet

- Sei Rest j auf einem Strang
  und Rest i auf einem anderen Strang:
  - **Parallele Bindung**:
    [Hbond(i,j), Hbond(j,i+2)],
    [Hbond(i+2,j+2), Hbond(j+2,i+4)],...
  - **Antiparallele Bindung**:
    [Hbond(j,i), Hbond(i,j)],
    [Hbond(j+2,i+2), Hbond(i+2,j+2)],..





Von der Primärstruktur zur gefalteten Struktur



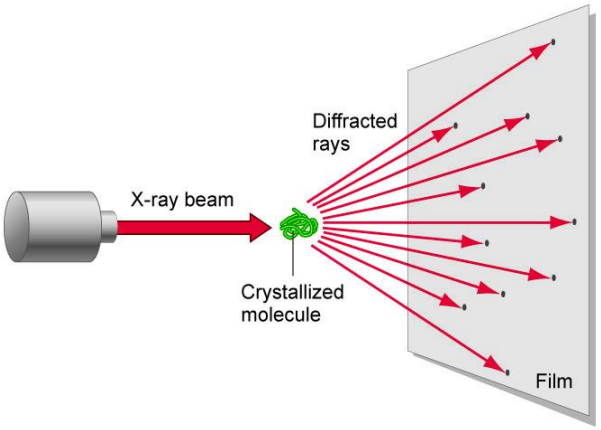
Monomer = Protein aus nur einer Kette
Multimer = Protein aus mehr als einer Kette

Dimer = Protein aus zwei Ketten
Homo-dimer = Protein aus zwei identischen Ketten
Hetero-dimer = Protein aus zwei unterschiedlichen Ketten

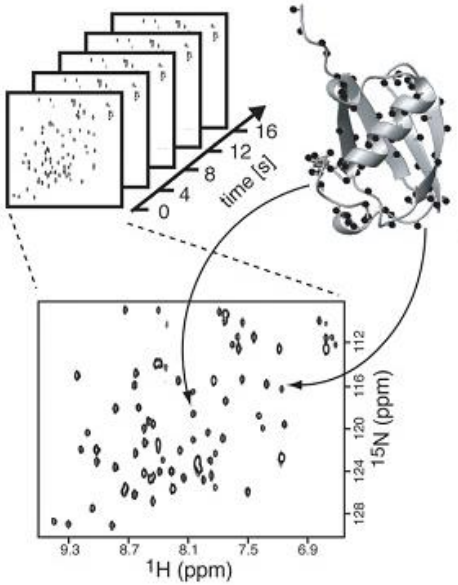
Image taken from Wikipedia

Experimentelle Methoden zur Strukturbestimmung

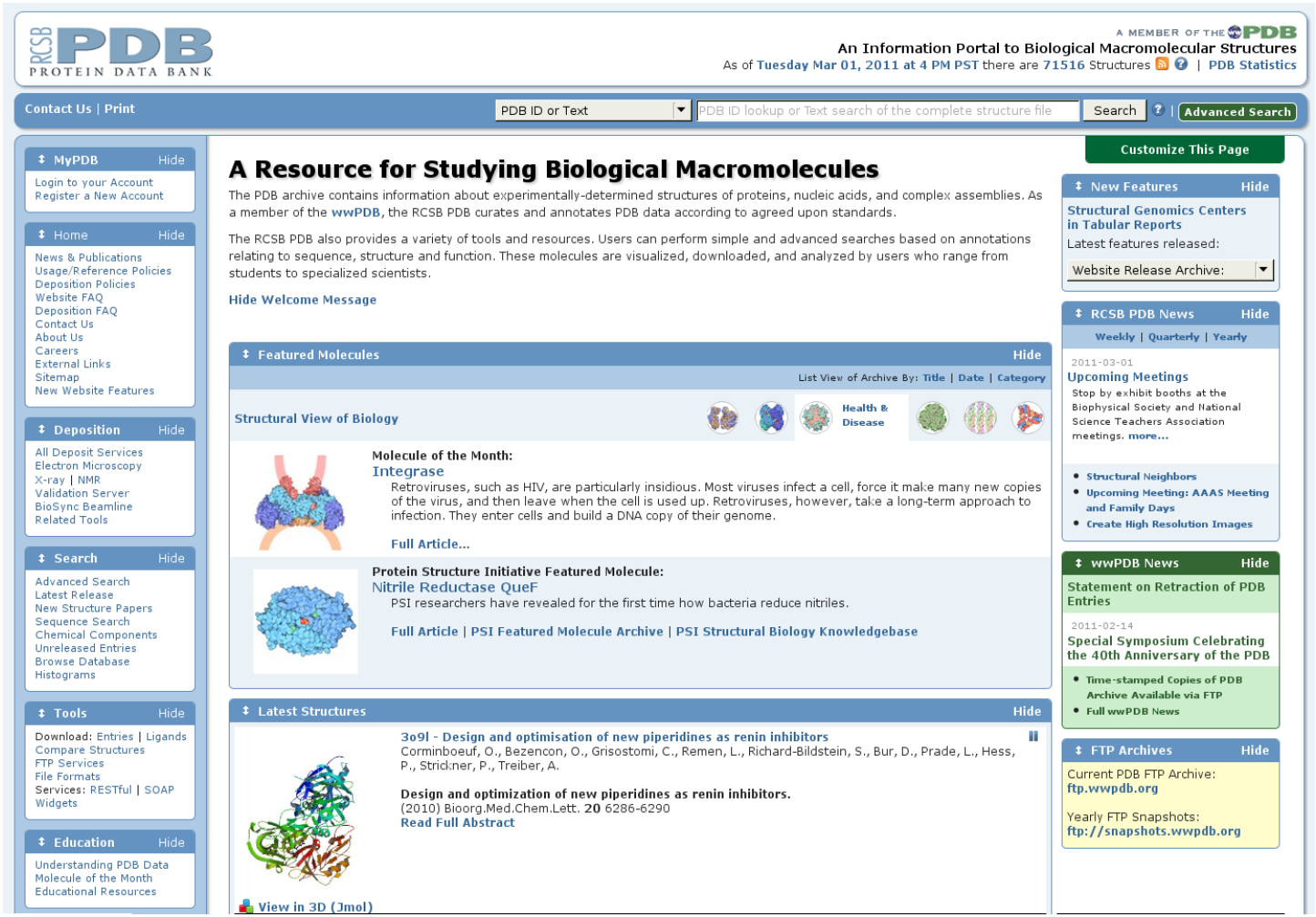
- X-Ray Crystallographie



- Nucleic magnetic resonance (NMR)



Protein Data Bank (PDB)



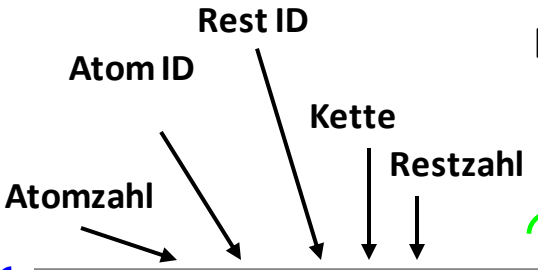
http://www.pdb.org/

Die PDB Datei - Text Format

```
HEADER    TRANSFERASE                             17-JUN-02   1M17
TITLE     EPIDERMAL GROWTH FACTOR RECEPTOR TYROSINE KINASE DOMAIN
TITLE    2 WITH 4-ANILINOQUINAZOLINE INHIBITOR ERLOTINIB
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: TYROSINE KINASE DOMAIN (RESIDUES 671-998);
COMPND   5 SYNONYM: RECEPTOR PROTEIN-TYROSINE KINASE ERBB-1;
COMPND   6 EC: 2.7.1.112;
COMPND   7 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 GENE: EGFR;
SOURCE   5 EXPRESSION_SYSTEM: SPODOPTERA FRUGIPERDA;
SOURCE   6 EXPRESSION_SYSTEM_COMMON: FALL ARMYWORM;
SOURCE   7 EXPRESSION_SYSTEM_STRAIN: AUTOGRAPHICA
SOURCE   8 CALIFORNICA/T.NICOPLUSIA;
SOURCE   9 EXPRESSION_SYSTEM_CELL_LINE: SF9;
SOURCE  10 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE  11 EXPRESSION_SYSTEM_PLASMID: PVL1392
KEYWDS    TRANSFERASE, TYROSINE KINASE DOMAIN
EXPDTA    X-RAY DIFFRACTION
AUTHOR    J.STAMOS,M.X.SLIWKOWSKI,C.EIGENBROT
REVDAT   2   25-FEB-03 1M17    1       JRNL
REVDAT   1   04-SEP-02 1M17    0
JRNL        AUTH   J.STAMOS,M.X.SLIWKOWSKI,C.EIGENBROT
JRNL        TITL   STRUCTURE OF THE EPIDERMAL GROWTH FACTOR RECEPTOR
JRNL        TITL 2 KINASE DOMAIN ALONE AND IN COMPLEX WITH A
JRNL        TITL 3 4-ANILINOQUINAZOLINE INHIBITOR.
JRNL        REF    J.BIOL.CHEM.                  V. 277 46265 2002
JRNL        REFN   ASTM JBCHA3  US ISSN 0021-9258
REMARK   1
REMARK   2
REMARK   2 RESOLUTION.  2.60 ANGSTROMS.
```

Minimaler Abstand zwischen Atomen, der aufgelöst werden konnte

Die PDB Datei - Text Format



ATOM:

Normalerweise Protein oder DNA

```
ATOM      2  CA  GLY A 672      54.168   8.340  69.707  1.00104.94           C
ATOM      3  C   GLY A 672      52.692   8.194  69.380  1.00105.46           C
ATOM      4  O   GLY A 672      51.877   9.045  69.750  1.00108.67           O
ATOM      5  N   GLU A 673      52.359   7.101  68.691  1.00102.41           N
ATOM      6  CA  GLU A 673      50.994   6.785  68.274  1.00 89.17           C
ATOM      7  C   GLU A 673      50.624   5.325  68.585  1.00 81.77           C
ATOM      8  O   GLU A 673      51.438   4.411  68.405  1.00 81.88           O
ATOM      9  CB  GLU A 673      50.850   7.050  66.777  1.00 96.53           C
ATOM     10  CG  GLU A 673      50.252   8.399  66.438  1.00 99.19           C
ATOM     11  CD  GLU A 673      48.788   8.486  66.827  1.00115.45           C
ATOM     12  OE1 GLU A 673      48.062   7.477  66.681  1.00116.71           O
ATOM     13  OE2 GLU A 673      48.356   9.561  67.286  1.00113.58           O
ATOM     14  N   ALA A 674      49.387   5.109  69.023  1.00 67.27           N
ATOM     15  CA  ALA A 674      48.912   3.768  69.370  1.00 63.11           C
ATOM     16  C   ALA A 674      48.702   2.826  68.174  1.00 58.54           C
ATOM     17  O   ALA A 674      48.064   3.183  67.186  1.00 62.02           O
ATOM     18  CB  ALA A 674      47.616   3.866  70.189  1.00 47.04           C
ATOM     19  N   PRO A 675      49.260   1.612  68.240  1.00 55.66           N
ATOM     20  CA  PRO A 675      49.087   0.665  67.134  1.00 52.95           C
ATOM     21  C   PRO A 675      47.629   0.261  66.997  1.00 48.19           C
```

HETATM:

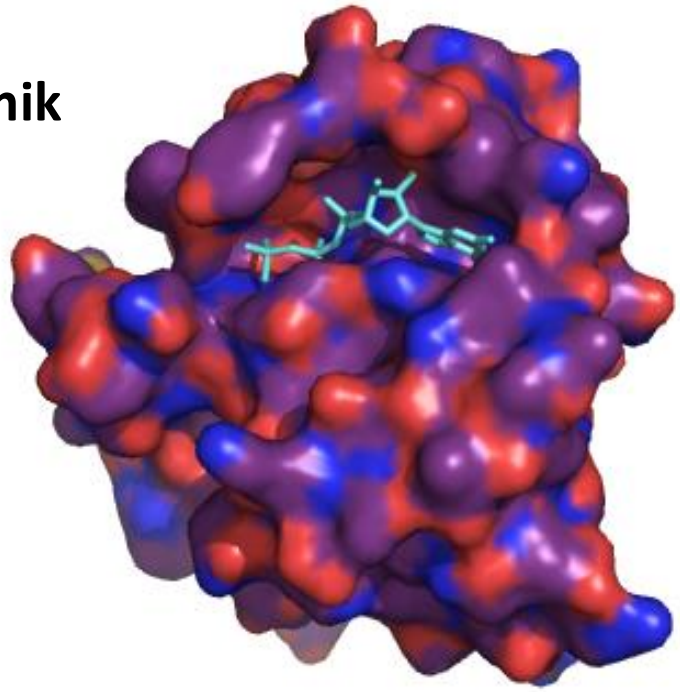
Normalerwise Ligand, Ion, Wasser

```
HETATM 2517  C5  AQ4   774      25.725   0.972  53.258  1.00 75.72           C
HETATM 2518  N1  AQ4   774      24.289   0.712  53.215  1.00 63.33           N
HETATM 2519  C6  AQ4   774      23.410  -0.217  53.908  1.00 56.92           C
HETATM 2520  C7  AQ4   774      22.037  -0.309  53.572  1.00 52.23           C
HETATM 2521  C8  AQ4   774      21.501   0.476  52.546  1.00 48.18           C
HETATM 2522  C9  AQ4   774      20.143   0.376  52.218  1.00 52.11           C
HETATM 2523  O1  AQ4   774      19.589   1.220  51.120  1.00 82.48           O
HETATM 2524  C10 AQ4   774      20.550   1.362  50.041  1.00 83.98           C
HETATM 2525  C11 AQ4   774      20.235   2.645  49.262  1.00 91.80           C
```

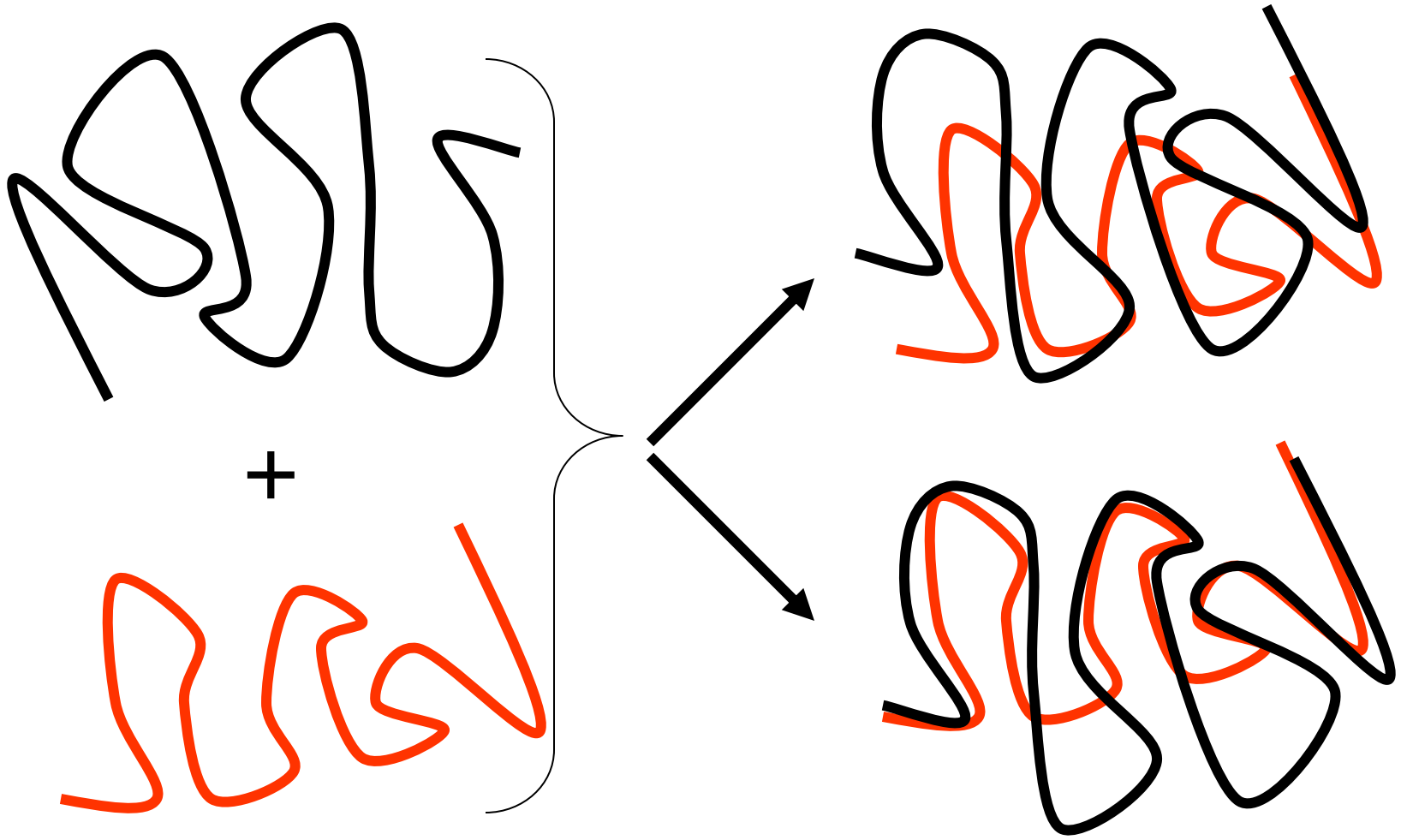
Software für Struktur-Repräsentation

Programme für Molekular-Grafik:

- **RasMol**
- **JMol**
- **Pymol**
- **VMD -> Molekulare Dynamik**
- **Insight II -> Drug Design**
- **.....**



Struktur Alignment



Struktur Alignment

- Motivation
- Grundlagen
- Ein einfacher Algorithmus
- Ein etwas verfeinerter Algoruthmus basierend auf DP
- Doppelte Dynamische Programmierung

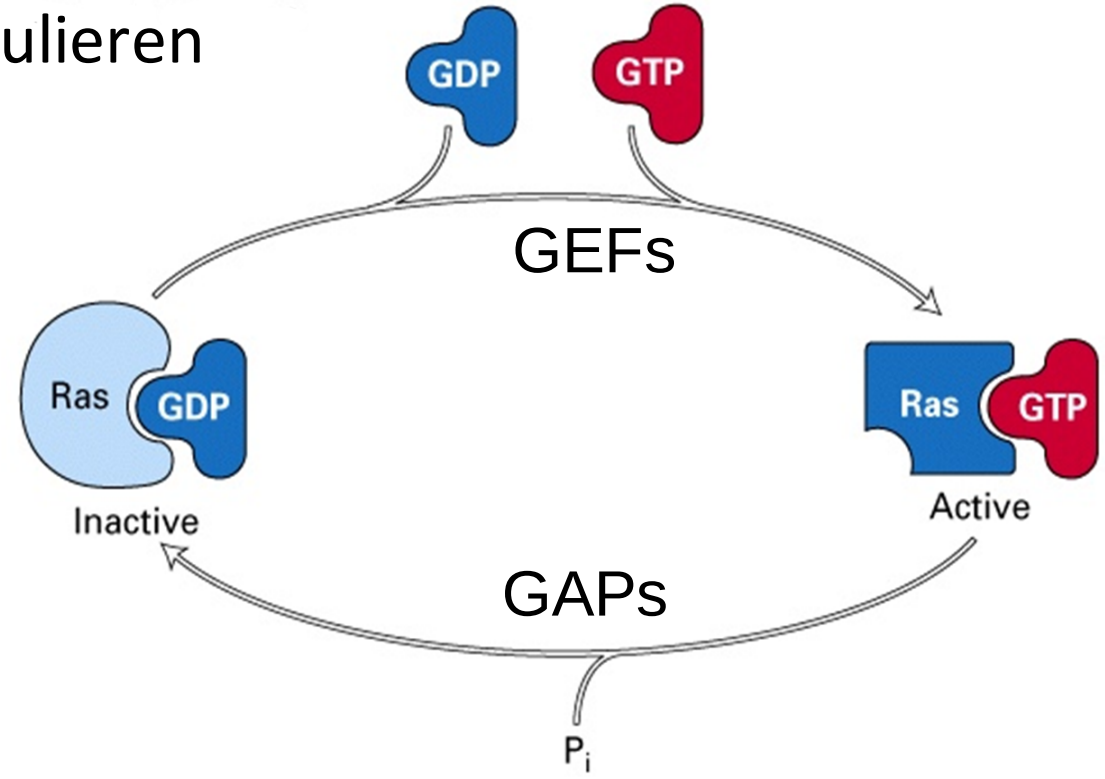
Warum brauchen wir strukturelles Alignment?

- Biomolekulare Erkennung
- Strukturell ähnliche Rezeptoren können ähnliche Medikamente binden
- Konformationsänderungen
  - Die Struktur kann sich nach der ligand Bindung ändern
  - Strukturelles Alignment kann diese Änderungen zeigen
- Erkennung entfernter Verwandter (Proteine)

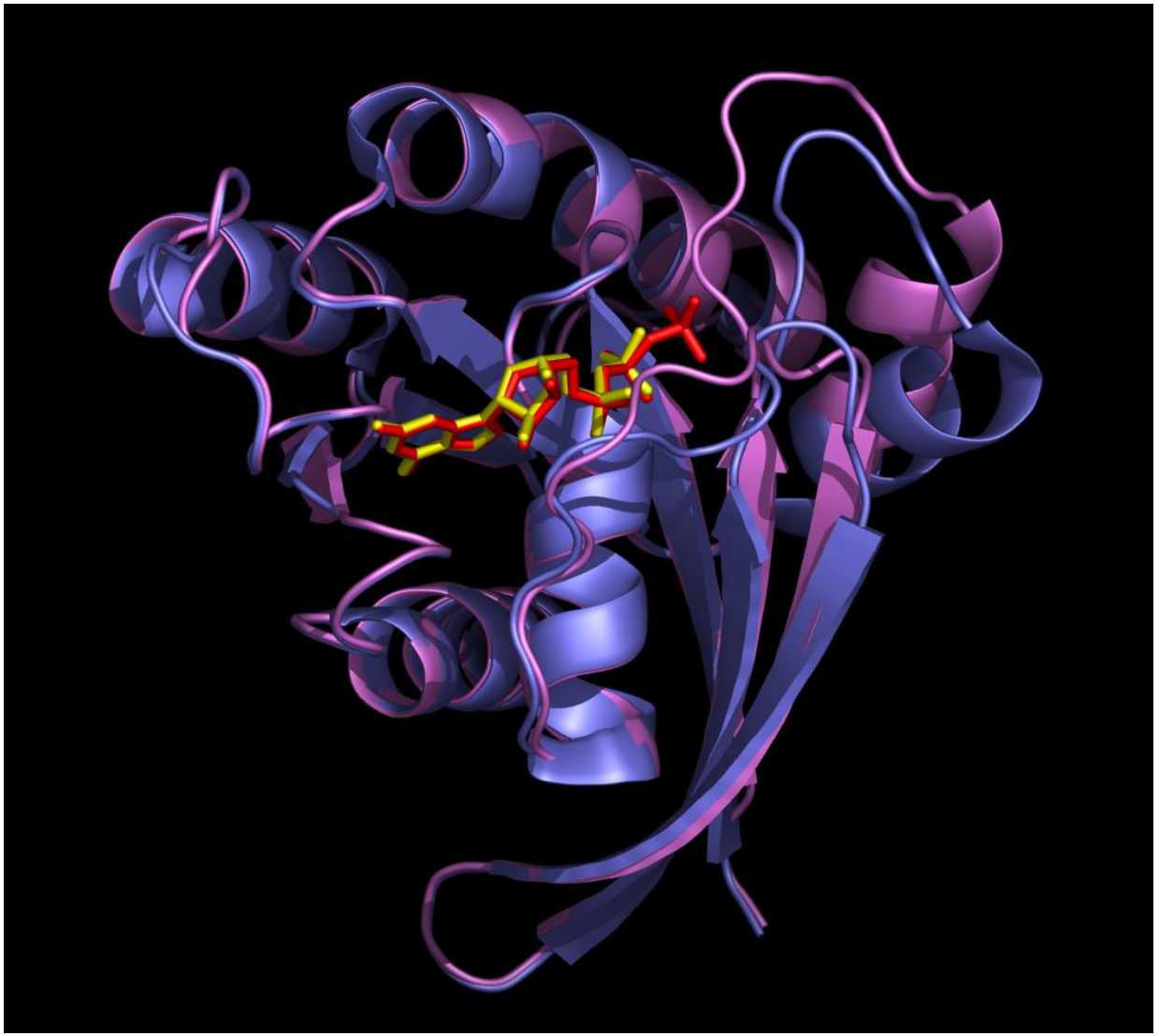
Motivation: Konformationsänderungen

- Kleine GTPases agieren als molekulare Schalter, die wichtige Funktionen und Wege in der Zelle kontrollieren und regulieren

- Aktiviert durch *guanine nucleotide exchange factors* (GEF)
- Deaktiviert durch *GTPase activating proteins* (GAP)

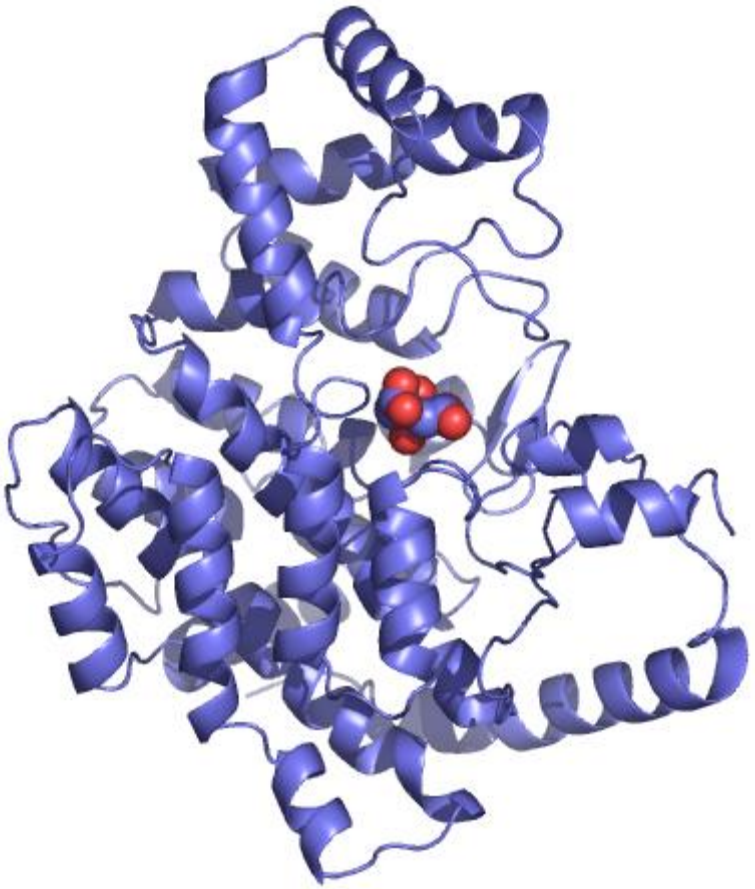


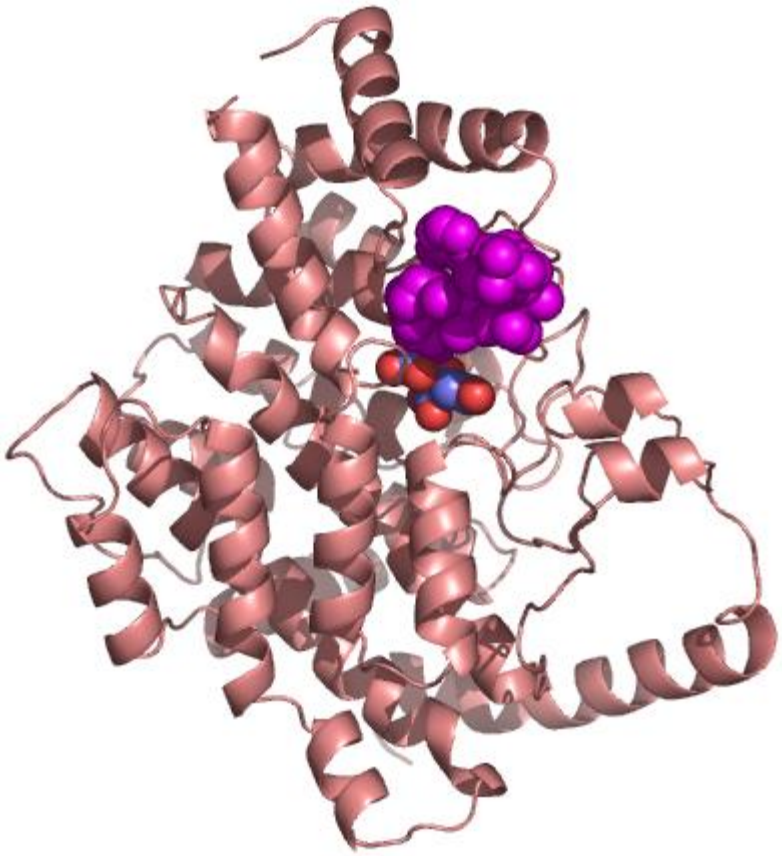
G Proteins: Konformationsänderungen in den GTP und GDP Bindungszuständen

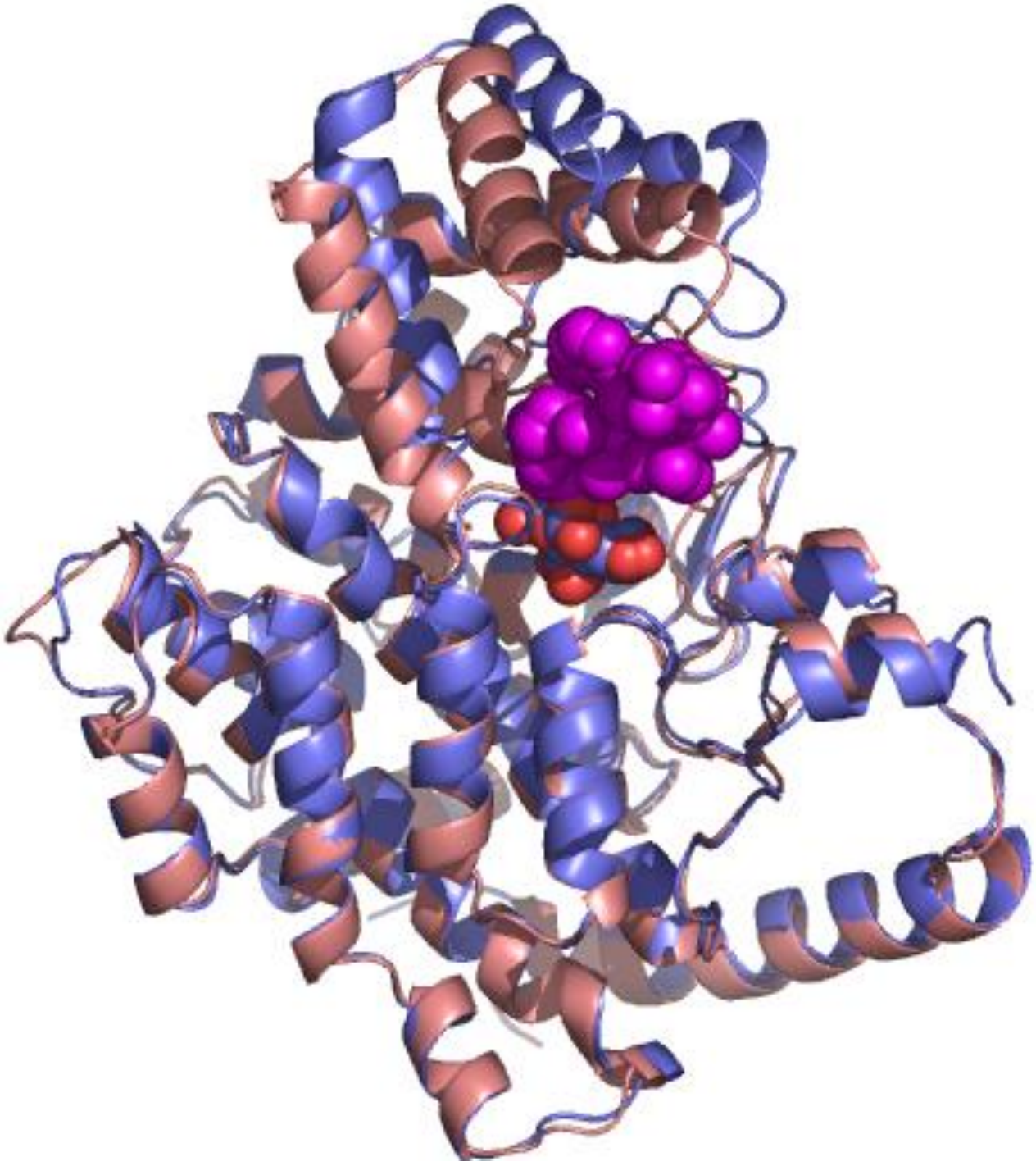


Geöffnete und geschlossene Konformation von *cytrate synthase* (1cts,5cts)

- Offen: oxalacetate, Geschlossen: oxalacetate und co-enzyme A
- Schleife zwischen zwei Helices bewegt sich um 6Å und dreht sich um 28º; einige Atome bewegen sich um 10Å

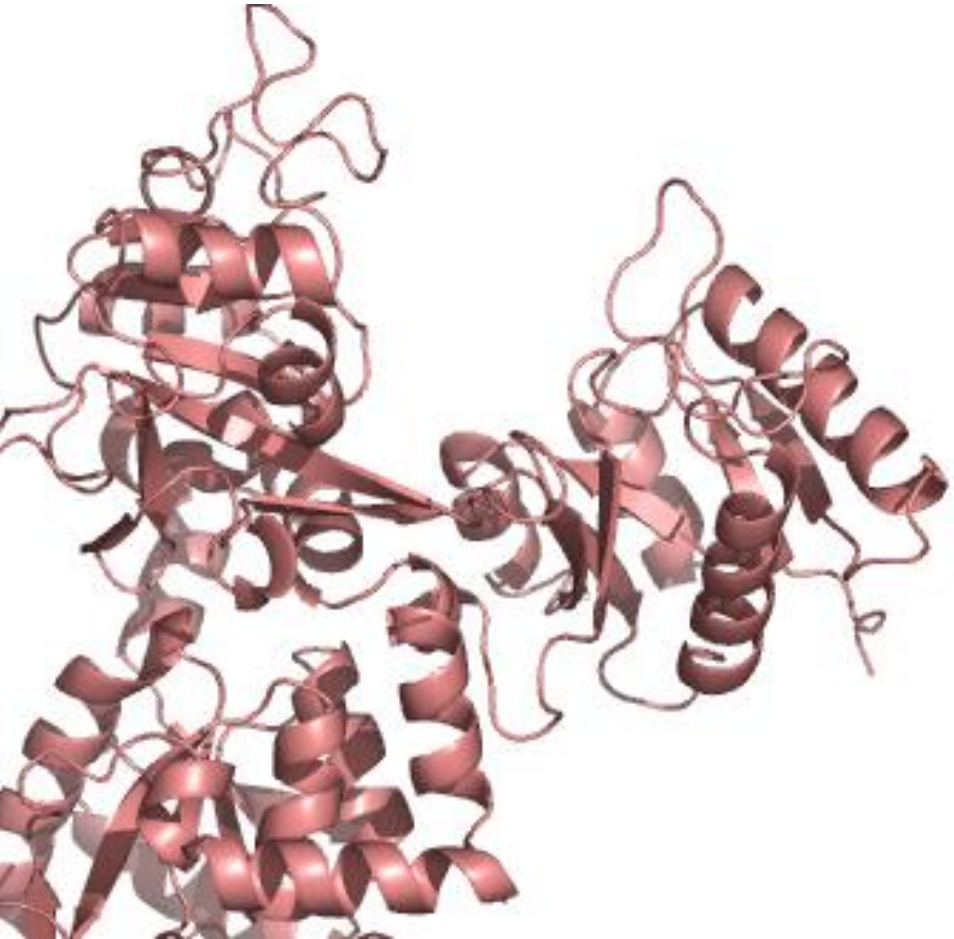


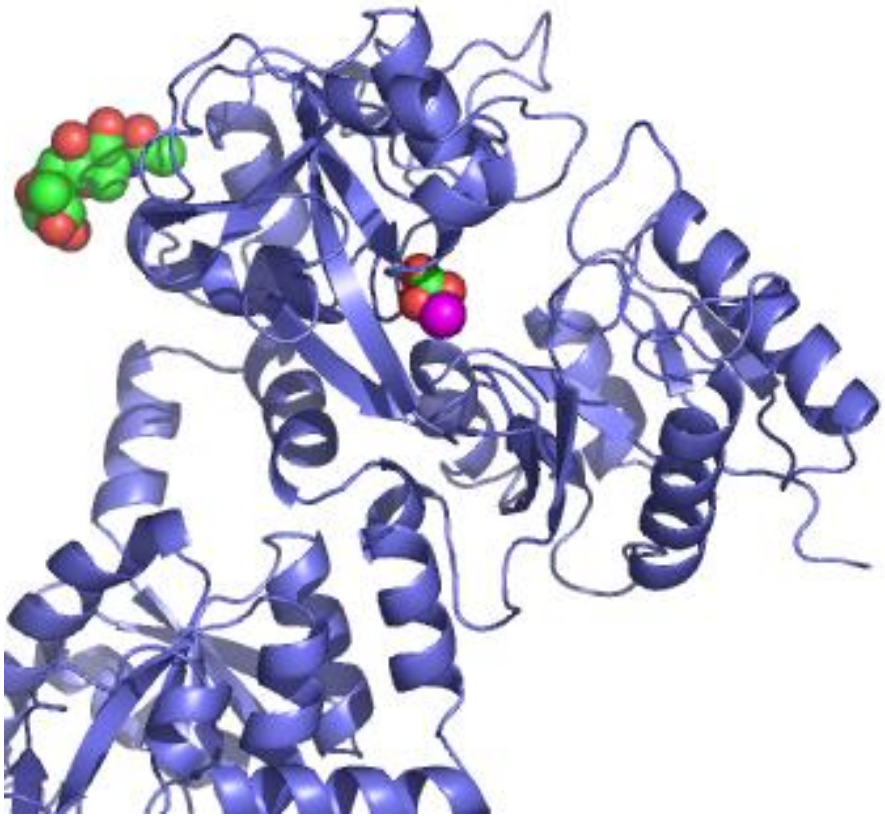




Hinge Motion in Lactoferrin (1lfh, 1lfg)

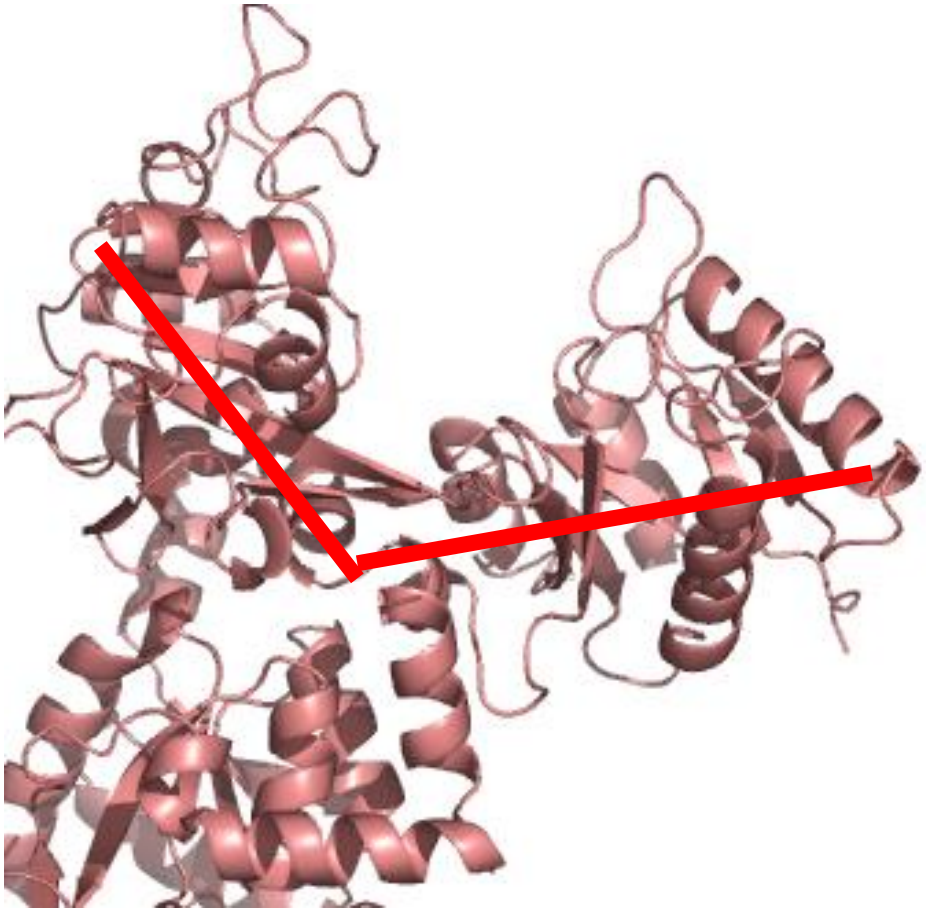
- Lactoferrin ist ein Eisen-bindendes Protein, welches in Sekreten wie Milch oder Tränken gefunden werden kann
- Drehung um 54º nach Eisen-Binding

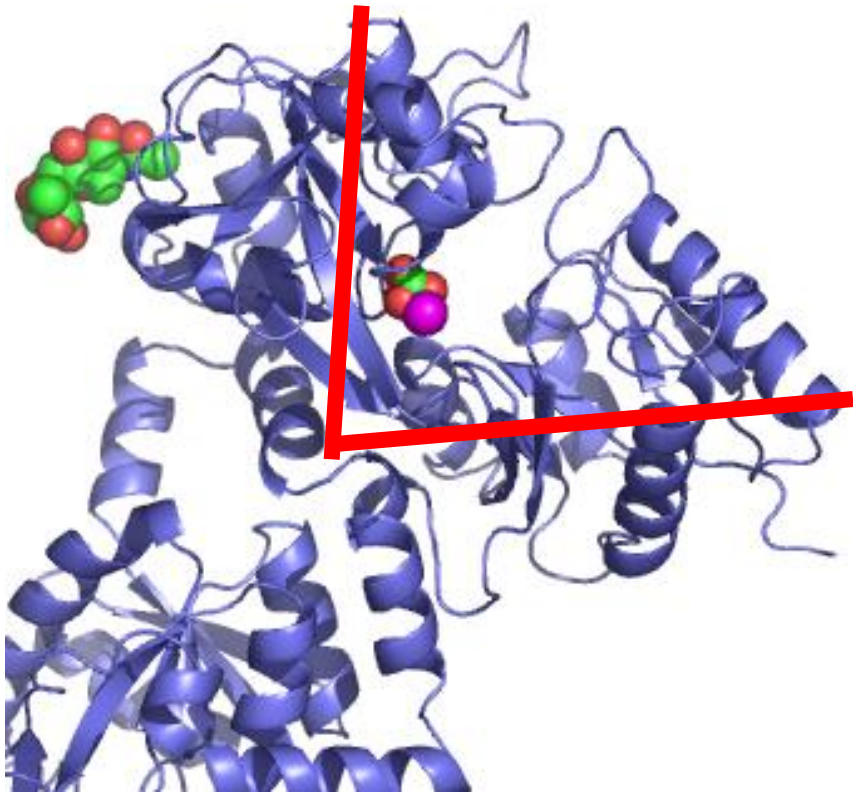


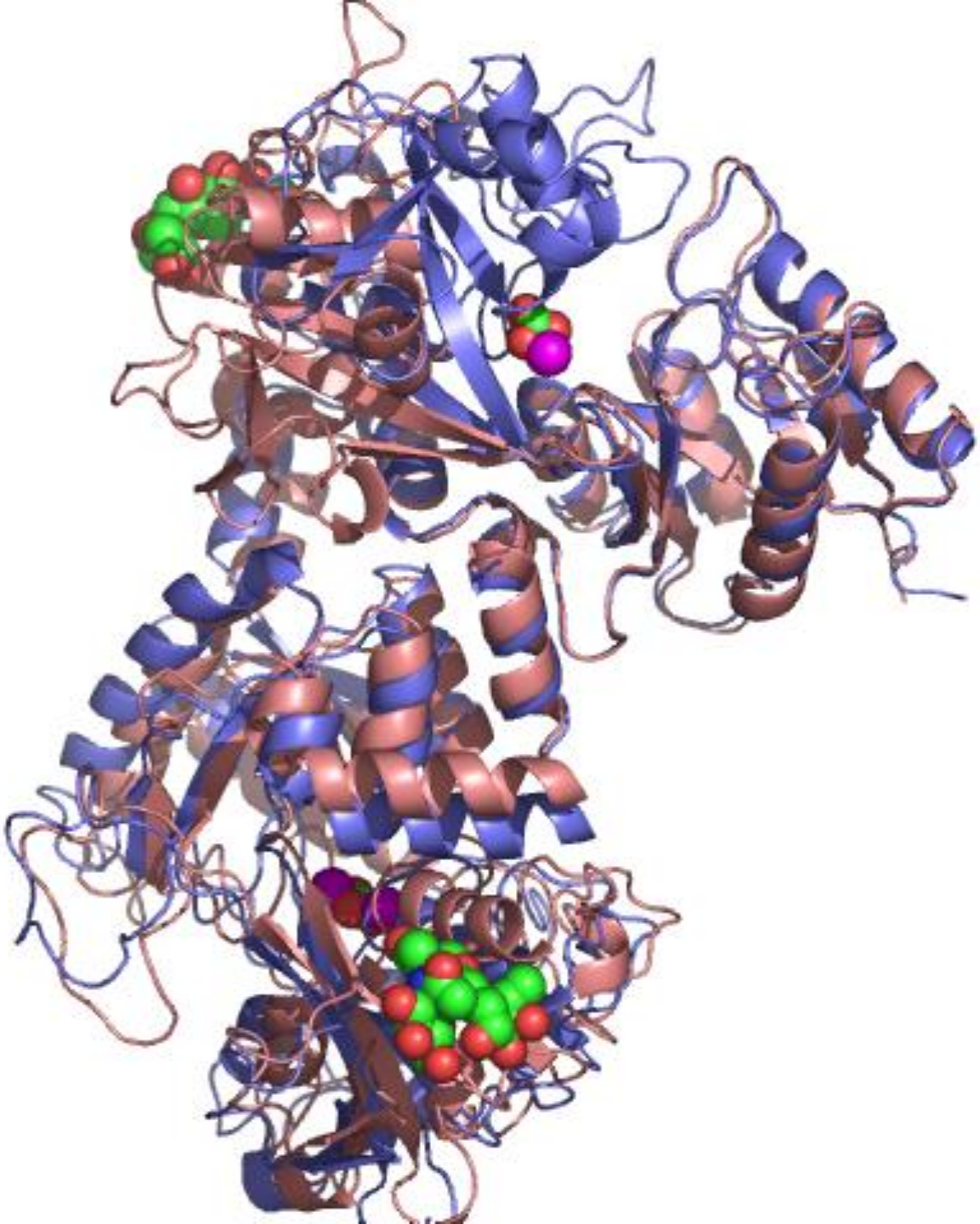


Hinge motion in Lactoferrin (1lfh, 1lfg)

- Lactoferrin ist ein Eisen-bindendes Protein, welches in Sekreten wie Milch oder Tränken gefunden werden kann
- Drehung um 54º nach Eisen-Binding

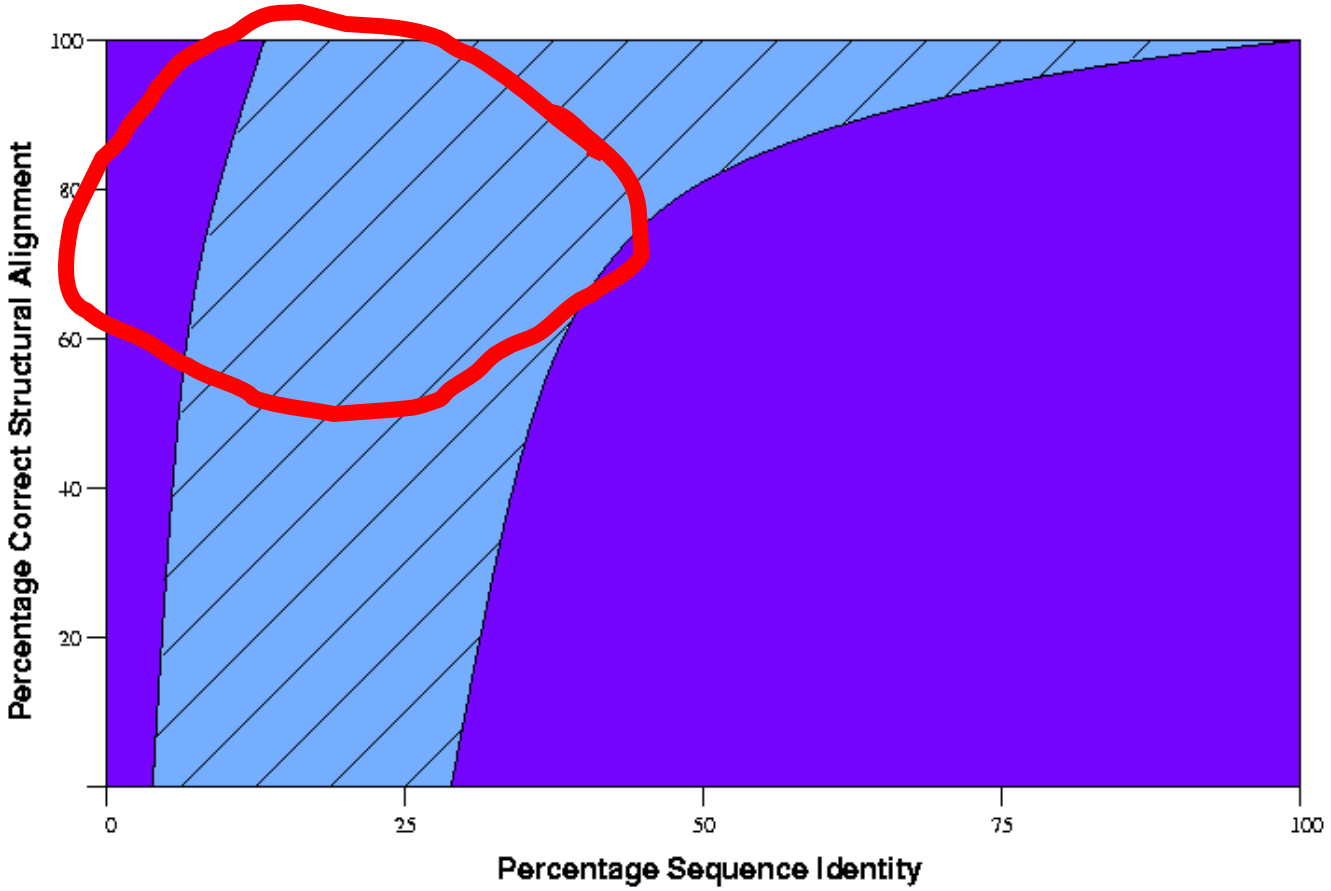






Motivation: (Entfernte) Verwandte

- Sequenzähnlichkeit mag gering sein, aber strukturelle Ähnlichkeit kann immer noch hoch sein

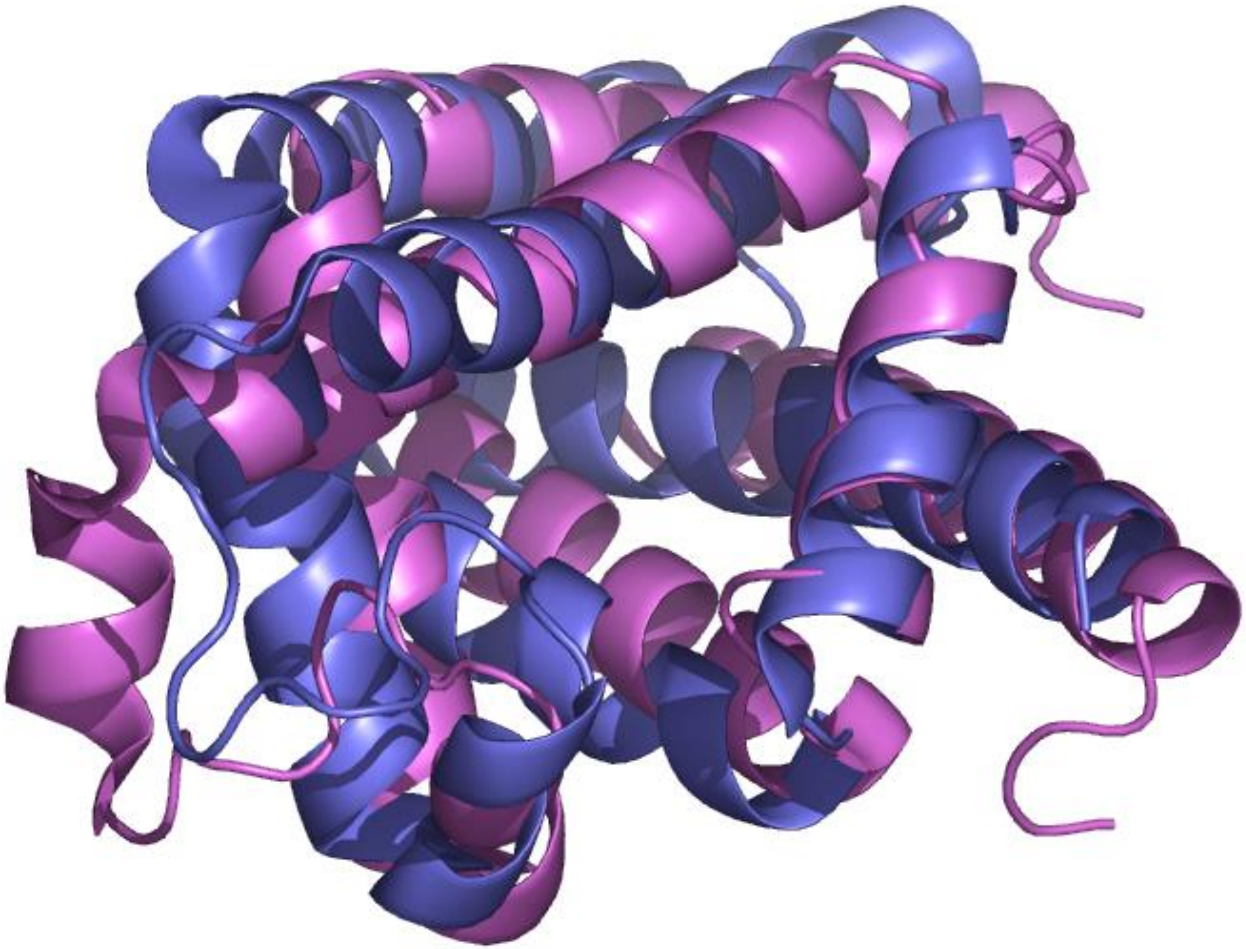


Picture from www.jenner.ac.uk/YBF/DanielleTalbot.ppt

Entfernte Verwandte (Beispiel)

- Globins occur widely
- Primary function: binding oxygen
- Assembly of helices surrounding haem group

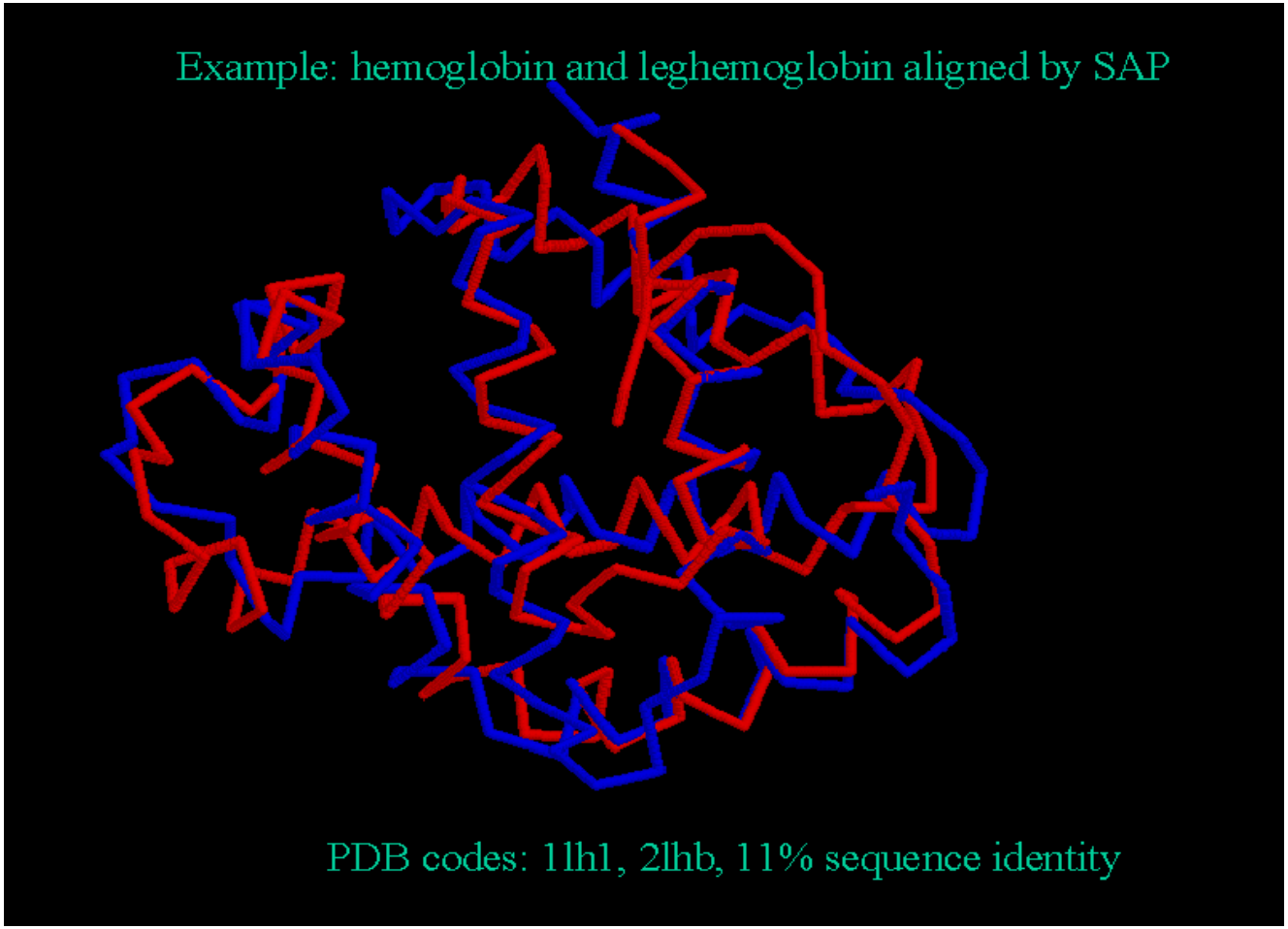
Verwandte



Wal-Sperma myoglobin (2lh7) und Lupin leghaemoglobin (1mbd),

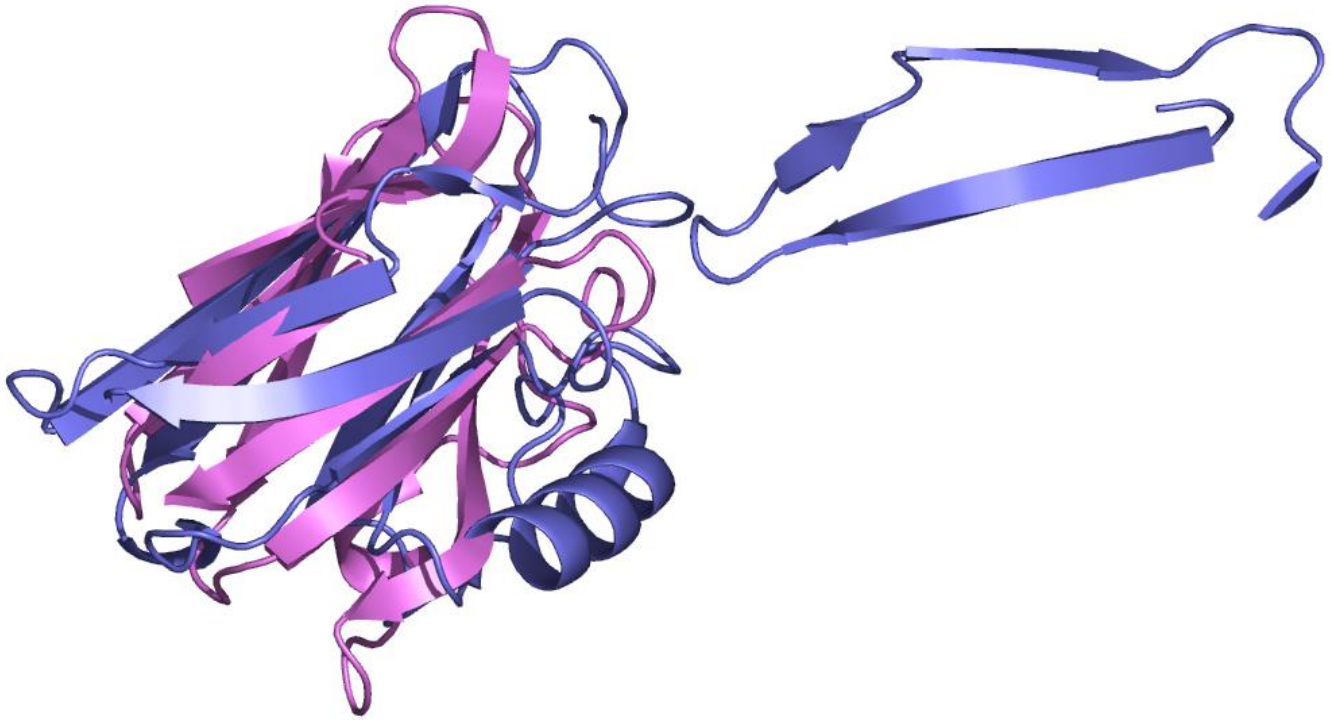
22% Sequenz Identität

Entfernte Verwandte



Verwandte

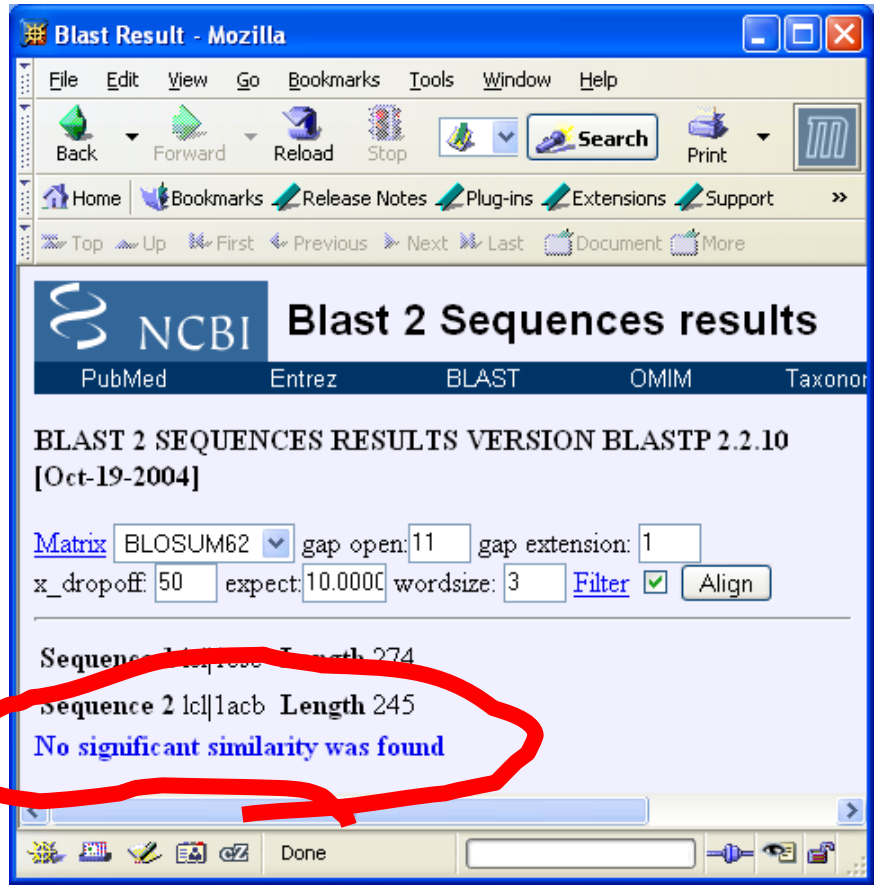
- Plastocyanin (5pcy) und azurin (2aza), 24% Sequenz Identität
- Kern der Struktur ist konserviert



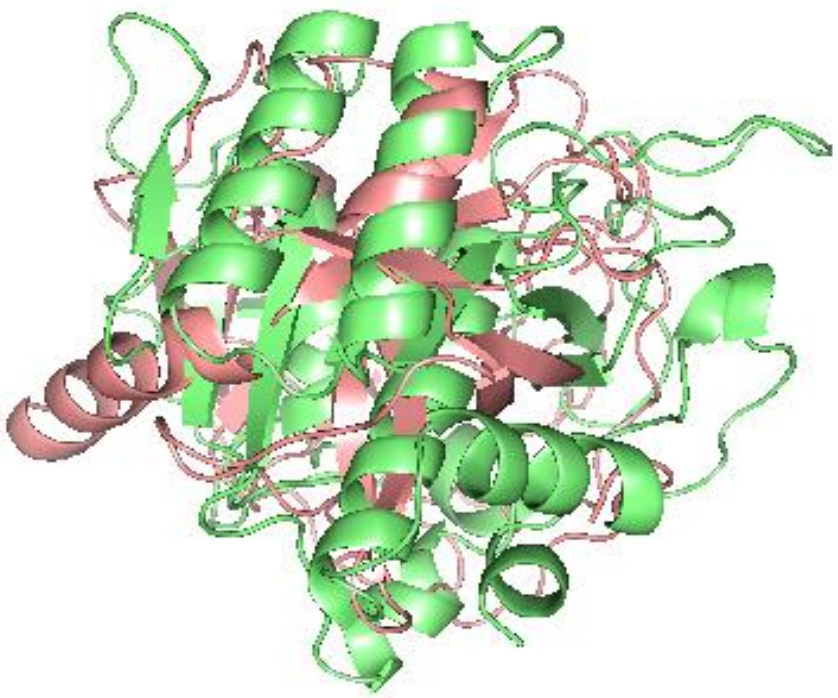
Motivation: Konvergente Evolution

Sehr niedrige Sequenzähnlichkeit..

```
>1cse Subtilisin
AQTVPYGIPLIKADKVQAQGFKGANVKVAVLDTGIQA
SHPDLNVVGGASFVAGEAYNTDGNGHGTHVAGTVAAL
DNTTGVLGVAPSVSLYAVKVLNSSGSGSYSGIVSGIE
WATTNGMDVINMSLGGASGSTAMKQAVDNAYARGVVV
VAAAGNSGNSGSTNTIGYPAKYDSVIAVGAVDSNSNR
ASFSSVGAELEVMAPGAGVYSTYPTNTYATLNGTSMA
SPHVAGAAALILSKHPNLSASQVRNRLSSTATYLGSS
FYYGKGLINVEAAAQ
>1acb Chymotrypsin
CGVPAIQPVLSGLSRIVNGEEAVPGSWPWQVSLQDKT
GFHFCGGSLINENWVVTAAHCGVTTSDVVVAGEFDQG
SSSEKIQKLKIAKVFKNSKYNSLTINNDITLLKLSTA
ASFSQTVSAVCLPSASDDFAAGTTCVTTGWGLTRYTN
ANTPDRLQQASLPLLSNTNCKKYWGTKIKDAMICAGA
SGVSSCMGDSGGPLVCKKNGAWTLVGIVSWGSSTCST
STPGVYARVTALVNWVQQTLAAN
```



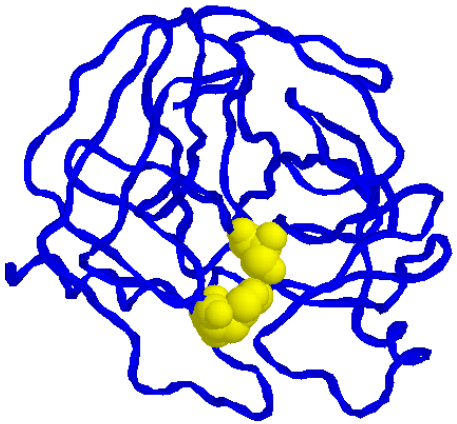
Strukturelle Ähnlichkeit: Niedrig



1CSE:E, 1ACB:E

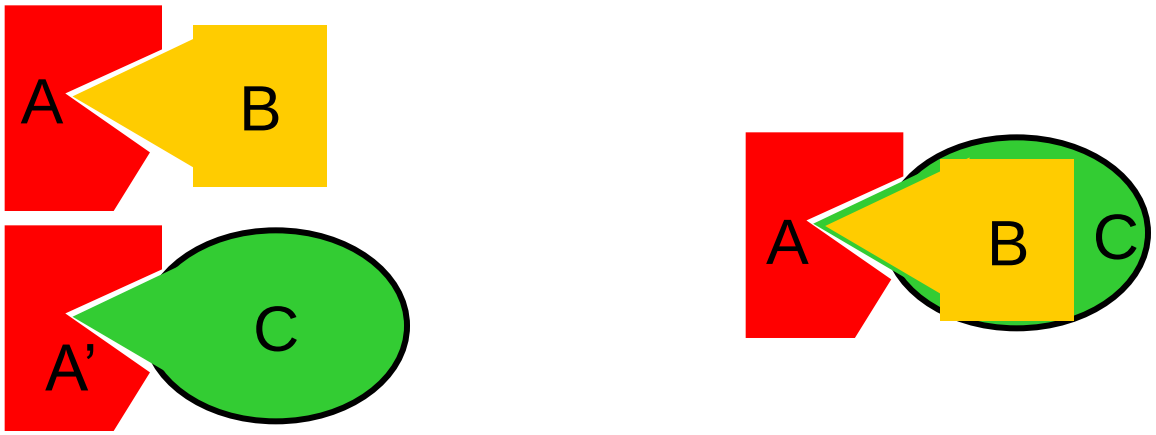
Drei Reste machen die Arbeit!

>1cse Subtilisin
AQTVPYGIPLIKADKVQAQGFKGANVKVAVL**D**TGIQA
SHPDLNVVGGASFVAGEAYNTDGNG**H**GTHVAGTVAAL
DNTTGVLGVAPSVSLYAVKVLNSSGSGSYSGIVSGIE
WATTNGMDVINMSLGGASGSTAMKQAVDNAYARGVVV
VAAAGNSGNSGSTNTIGYPAKYDSVIAVGAVDSNSNR
ASFSSVGAELEVMAPGAGVYSTYPTNTYATLNGT**S**MA
SPHVAGAAALILSKHPNLSASQVRNRLSSTATYLGSS
FYYGKGLINVEAAAQ
>1acb Chymotrypsin
CGVPAIQPVLSGLSRIVNGEEAVPGSWPWQVSLQDKT
GFHFCGGSLINENWVVTAA**H**CGVTTSDVVVAGEFDQG
SSSEKIQKLKIAKVFKNSKYNSLTINN**D**ITLLKLSTA
ASFSQTVSAVCLPSASDDFAAGTTCVTTGWGLTRYTN
ANTPDRLQQASLPLLSNTNCKKYWGTKIKDAMICAGA
SGVSSCMGD**S**GGPLVCKKNGAWTLVGIVSWGSSTCST
STPGVYARVTALVNWVQQTLAAN



Konvergente Evolution

Strukturelle Alignment kann die gemeinsame Interaktionsschnittstelle in ‘nicht verwandten’ Proteinen zeigen und Fälle von konvergenter Evolution erklären
Die Mechanismen werden durch Viren ausgenutzt



A und B sind nativ, C ist viral

Henschel et al., Bioinformatics 2006

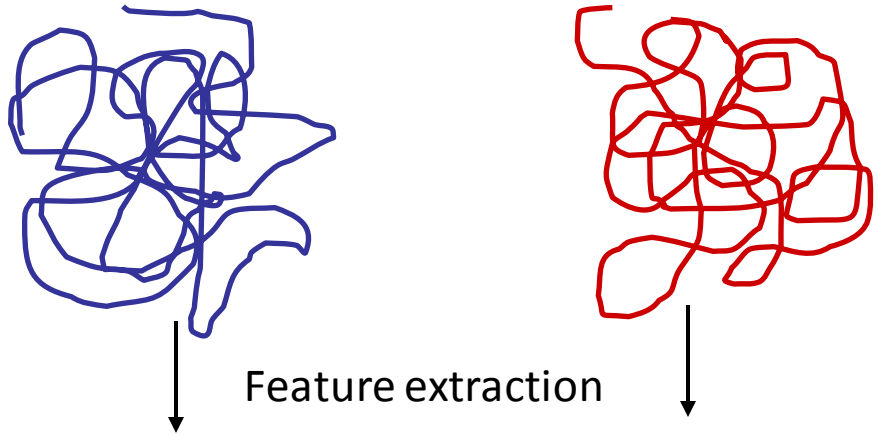
Basis für Strukturelles Alignment: Was brauchen wir?

- **Repräsentation der Proteinform**
  - durch diskrete 3D **Features**
  - *Invariant* nach Translation und Rotation (Proteine sind ‚rigid bodies')
  - *Robust:* die Beschreibung ändert sich nicht nach kleinen Änderungen der Struktur (Fehler)
  - *Verschidene Strukture* haben vershiedene Beschreibungen

- **Algorithmen für den Vergleich**
  - Äquivalente Reste finden
  - Hauptoperationen
  - Bewertungschema
  - (Bedingungen)

- Evaluation des Vergleichs basierend auf einigen Kriterien

Framework für paarweisen Strukturvergleich



Feature extraction

Struktur Beschreibung 1

Struktur Beschreibung 2

Scoring

Vergleichalgorithmus

Bedingungen

Äquivalenz

Evaluation

Framework für paarweisen Strukturvergleich - Äquivalenz

Wir müssen die **Korrespondenzen** zwischen den Elementen
der zwei Strukturen finden
Seien A und B zwei Objekte, mit den Elementen A_1, ...A_m und B_1, ...B_n

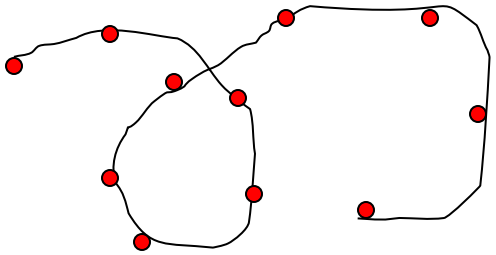
Eine **Äquivalenz** ist definiert als eine Menge von Paaren
$E(A,B)=\{(A_{i1}, B_{j1}), (A_{i2}, B_{j2}), ...(A_{ir}, B_{jr})\}$

Die Äquivalenz wird Alignment genannt wenn:

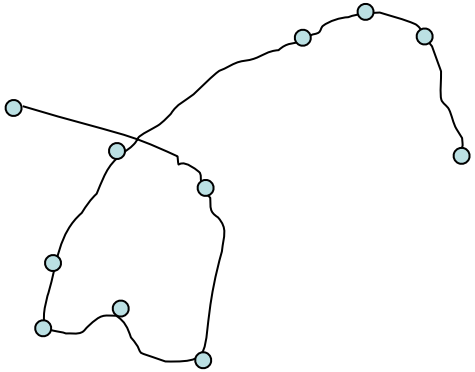
- **A und B geordnet sind und**
- **die Paaren in E kollinear sind**:
  $i_1 < i_2 < ....< i_r$ und $j_1 < j_2 <....j_r$ (bedeutet das die Äquivalenz die Ordnung erhält)

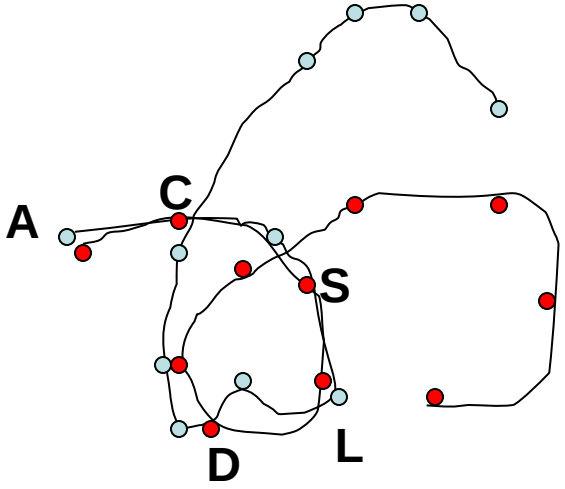
Warum kann dynamische Programmierung nicht angewendet werden?

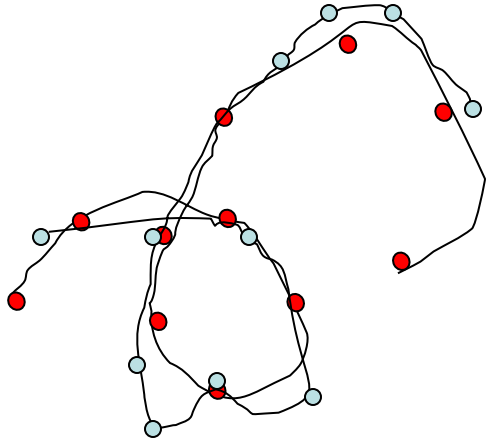
A : ACSLDRTSIRV



B : ATLREKSSLIR







Sequenz Alignment basierend auf DP

```
ACSL-DRTS-IRV
A-TLREKSSLIR-
```

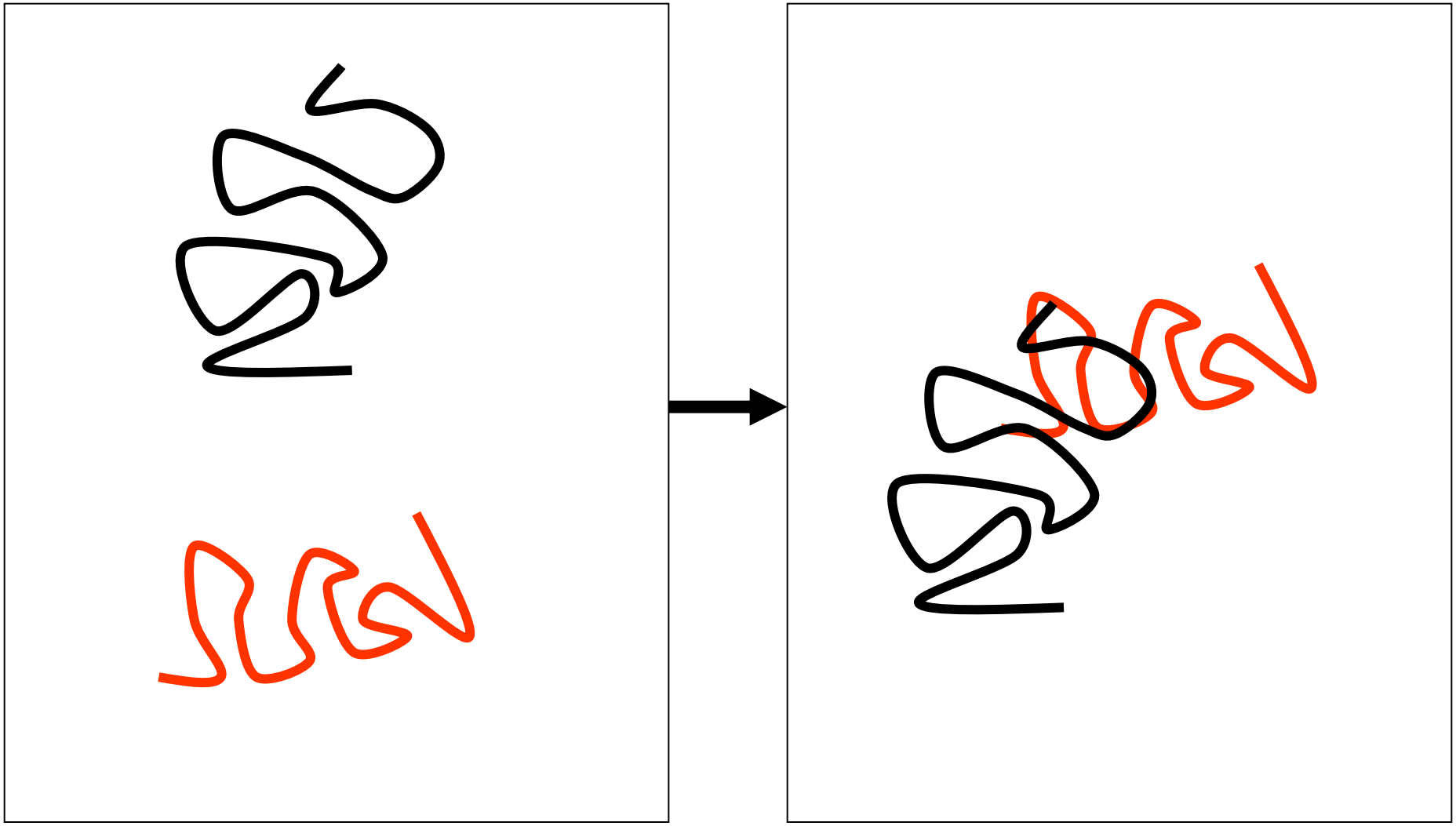
Basis für Strukturelles Alignment: Was brauchen wir?

- **Repräsentation der Proteinform**
  - durch diskrete 3D **Features**
  - *Invariant* nach Translation und Rotation (Proteine sind ‚rigid bodies'
  - *Robust:* die Beschreibung ändert sich nicht nach kleinen Änderunge der Struktur (Fehler)
  - *Verschidene Strukture* haben vershiedene Beschreibungen

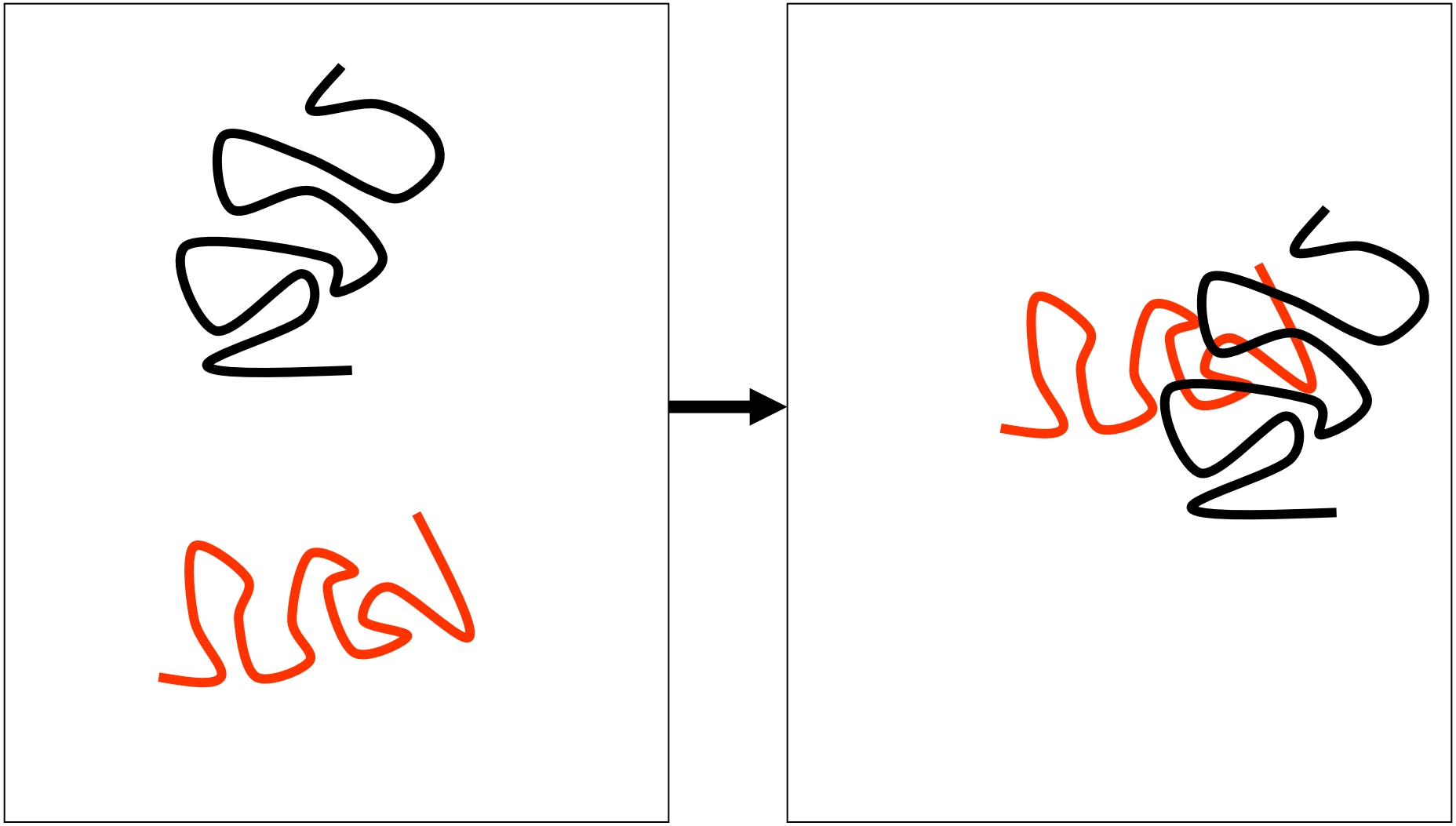
- **Algorithmen für den Vergleich**
  - **Äquivalente Reste finden** → **Schwierige Aufgabe**
  - Hauptoperationen → **Translation und Rotation!**
  - Bewertungschema
  - (Bedingungen)

- Evaluation des Vergleichs basierend auf einigen Kriterien

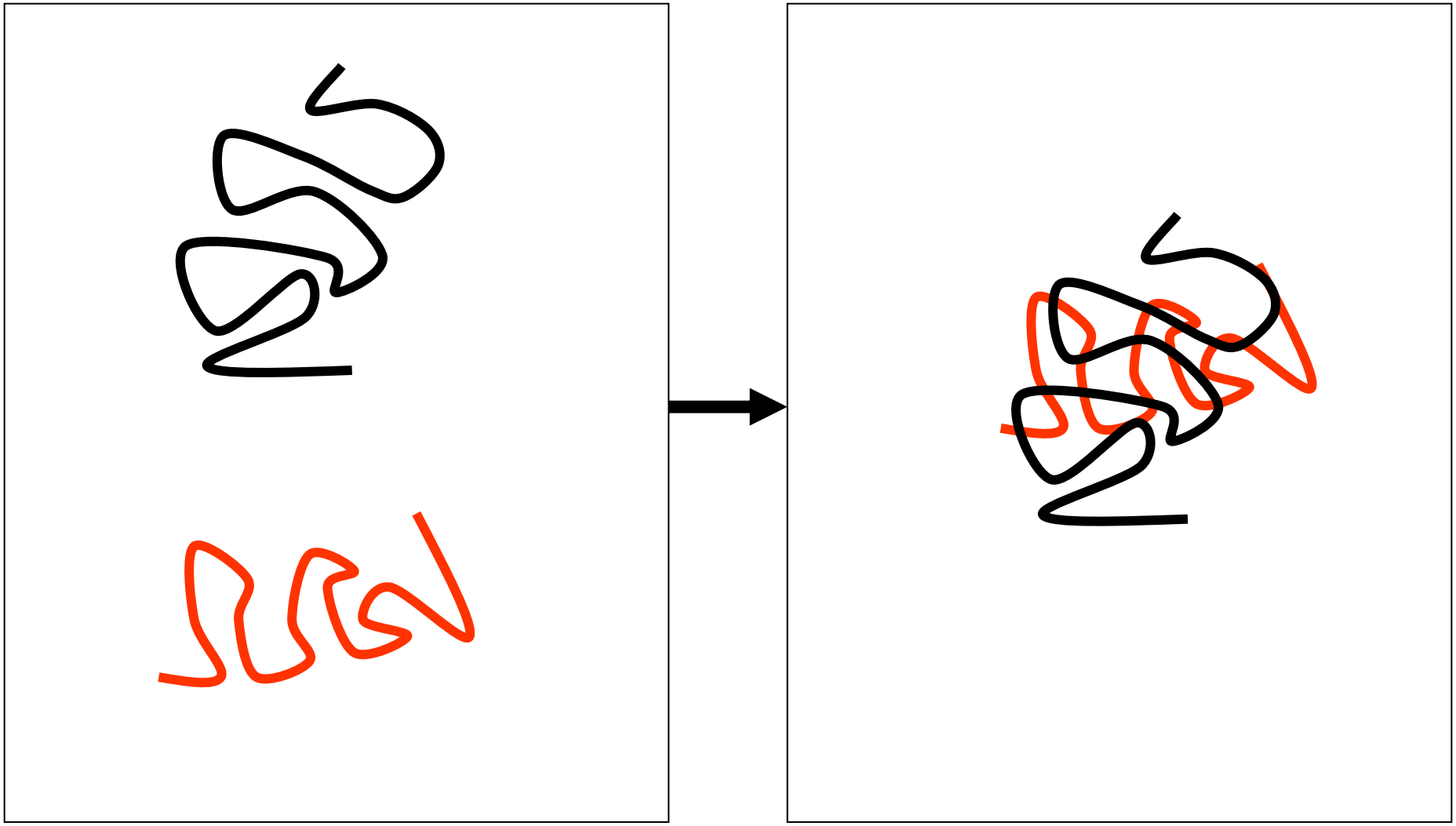
Basis Operationen: Translation



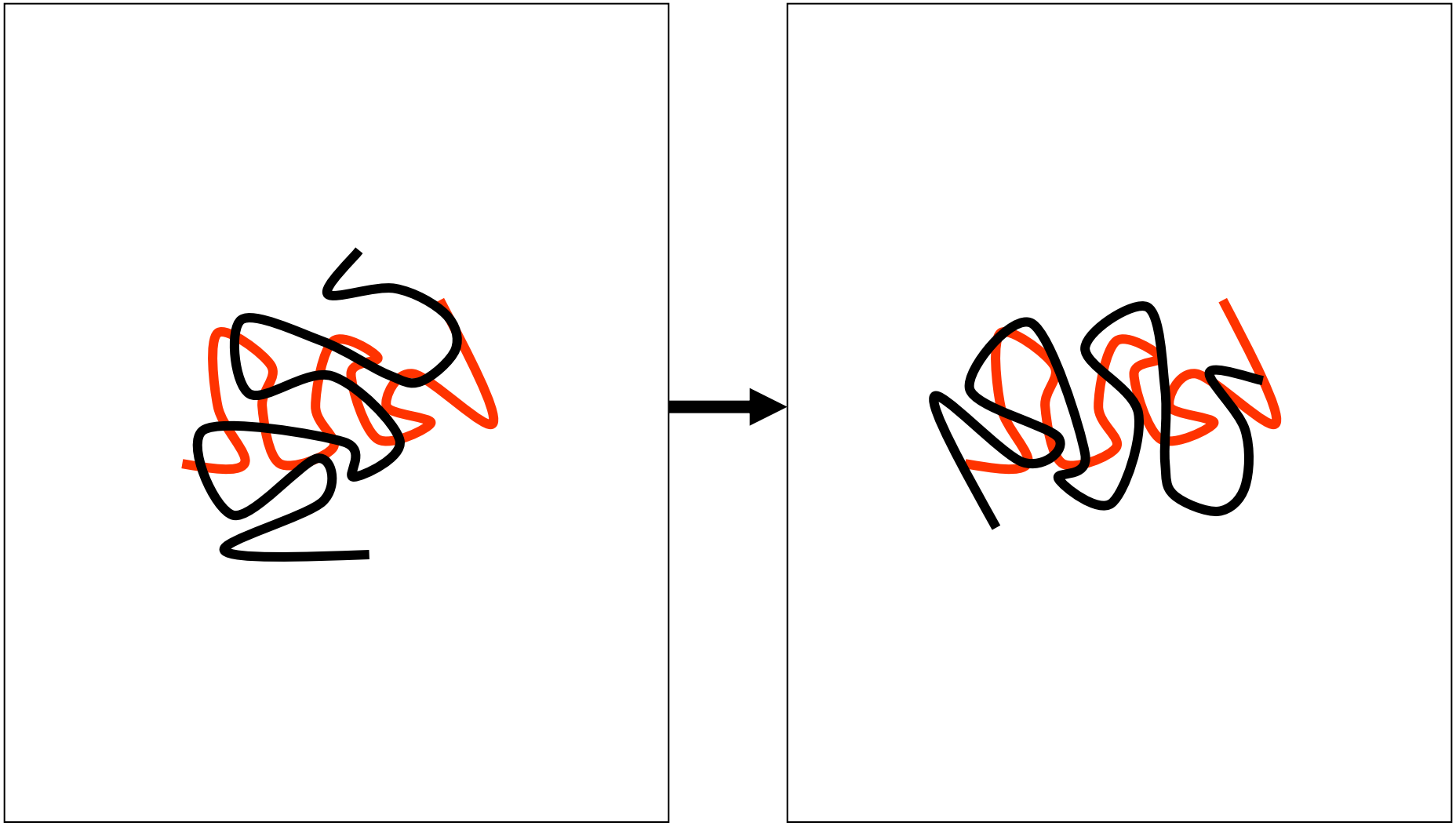
Basis Operationen: Translation



Basis Operationen: Translation



Basic Operationen: Rotation



Basis für Strukturelles Alignment: Was brauchen wir?

- **Repräsentation der Proteinform**
  - durch diskrete 3D **Features**
  - *Invariant* nach Translation und Rotation (Proteine sind ‚rigid bodies')
  - *Robust:* die Beschreibung ändert sich nicht nach kleinen Änderungen der Struktur (Fehler)
  - *Verschidene Strukture* haben vershiedene Beschreibungen

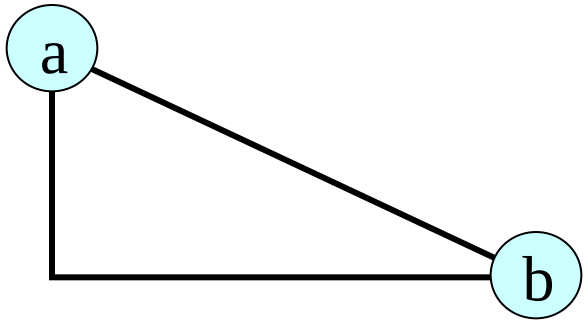
- **Algorithmen für den Vergleich**
  - **Äquivalente Reste finden** → **Schwierige Aufgabe**
  - Hauptoperationen → **Translation und Rotation!**
  - Bewertungschema → **Finde die beste Überlagerung zwischen Korrespondierenden Resten- RMSD**
  - (Bedingungen)

- Evaluation des Vergleichs basierend auf einigen Kriterien

Bewertung:
Root Mean Square Deviation (RMDS)

- Was ist die Abstand zwischen zwei Punkten *a* mit Koordinaten x_a und y_a und *b* mit Koordinaten x_b und y_b?
  - Euclidean Abstand:

$$d(a,b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$



  - Und in 3D?

Root Mean Square Deviation (RMDS)

- In einem Struktur Alignment, der RMSD misst wie weit die ausgerichteten Atome durchschnittlich voneinander entfernt sind
- Gegeben sind die Distanzen d_i zwischen n ausgerichteten Atomen, der *root mean square deviation* ist definiert durch:

$$\text{rmsd} = \sqrt{1/n \sum d_i^2}$$

Quälität des Alignments

- Die Einheit des RMSD => z.B. Ångstroms
  - Identische Strukturen => *RMSD* = “0”
  - Gleiche Strukturen => *RMSD* ist klein (1 – 3 Å)
  - Entfernte Strukturen => *RMSD* > 3 Å

$RMSD_C$: *Coordinate RMSD*

- Sei (α_1, β_1), ...(α_r, β_r) die Koordinatenmenge von äquivalenten Elementen (E) (α_i aus A und β_i aus B)
- Jede Koordinatenmenge besteht aus drei Werten (3D)
- Aufgabe: Finde eine Transformation für A die den RMDS minimiert

$$RMSD_C(E) = \min_T \sqrt{\frac{1}{\sum_{i=1}^{r} \omega_i} \sum_{i=1}^{r} \omega_i (T\alpha_i - \beta_i)^2}$$

$$\downarrow$$

$$\sum_{i=1}^{r} (R\alpha_i + t - \beta_i)^2$$

$$\downarrow$$

$$\sum_{i=1}^{r} (R\alpha_i - \beta_i)^2$$

Nach der Translation zum gemeinsamen Ursprung

Ein sehr einfacher Algorithmus…

- …um identische Strukturen auszurichten (mit Konformationsänderungen)
  - Ein Sequenz **Alignment** bilden (nicht notwendig, wenn beide Sequenze 100% identisch sind)
  - Die Massezentren von beiden Strukturen bestimmen
  - Die beide Strukturen so bewegen, dass die Massezentren im Ursprung sind
  - Die Winkel zwischen allen ausgerichteten Resten berechnen
  - Struktur A um den Median von allen Winkeln rotieren

Ein sehr einfacher Algorithmus…

- …um identische Strukturen auszurichten (mit Konformationsänderungen)
  - Ein Sequenz **Alignment** bilden (nicht notwendig, wenn beide Sequenze 100% identisch sind)
  - Die Massezentren von beiden Strukturen bestimmen
  - Die beide Strukturen so bewegen, dass die Massezentren im Ursprung sind
  - Die Winkel zwischen allen a… berechnen
  - Struktur A um den Median … n

Frage: Wie?
Gegeben *n* Atome
(x_1,y_1,z_1) zu (x_n,y_n,z_n)
(für eine Struktur)

Ein sehr einfacher Algorithmus…

- …u
  Kor

Frage: Wie? Gegeben *n* Atome(x_1,y_1,z_1) zu $(x_n,y_n,z_{n:})$
Massezentrum $(x_{CoM},y_{CoM},z_{CoM})$ =
$(1/n \sum^{n}_{i=1} x_i \ , \ 1/n \sum^{n}_{i=1} y_i \ 1/n \sum^{n}_{i=1} z_i)$

  - Ein Sequenz **Alig**ilden (nicht notwendig, wenn beide Sequenze 100% identisch sind)
  - Die Massezentren von beiden Strukturen bestimmen
  - Die beide Strukturen so bewegen, dass die Massezentren im Ursprung si
  - Die Winkel zwischen allen aus berechnen
  - Struktur A um den Median von allen Winkeln rotieren

Frage: Wie?

Ein sehr einfacher Algorithmus…

- …um identische Strukturen auszurichten (mit Konformationsänderungen)
  - Ein Sequenz **Alignment** bilden (nicht notwendig, wenn beide Sequenze 100% identisch sind)
  - Die Massezentren von beiden Strukturen bestimmen
  - Die beide Strukturen so bewegen, dass die Massezentren im Ursprung sind
  - Die Winkel zwi[illegible]nen allen ausgerichteten Resten berechne[illegible]
  - Str[illegible]Median von allen Winkeln rotieren

For all *i*: do $x_i := x_i - x_{CoM}$, $y_i := y_i - y_{CoM}$, $y_i := y_i - y_{CoM}$,

Ein sehr einfacher Algorithmus…

- …um identische Strukturen auszurichten (mit Konformationsänderungen)
  - Ein Sequenz **Alignment** bilden (nicht notwendig, wenn beide Sequenze 100% identisch sind)
  - Die Massezentren von beiden Strukturen bestimmen
  - Die beide Strukturen so bewegen, dass die Massezentren im Ursprung sind
  - Die Winkel zwischen allen ausgerichteten Resten berechnen
  - Struktur A um den Median von allen Winkeln rotieren

Warum Median und nicht der Mittelwert?

Fazit
(Was haben wir heute gelernt?)

- Biochemie von Proteinstukturen
- Sekundär Strukturelemente
- Bedingungen für Konformationen: Tortionswinkel
- Strukturelles Alignment
  - Einige Fälle zur Motivation
  - Basis Algorithmus basierend auf Translation und Rotation
  - Wie bewerten wir strukturelle Alignments: RMSD Maßnahme

Strukturelle Ausrichtung: Ist das so einfach?

Ein verfeinertes Alignment: Abwechslung von Alignment und Überlagerung

- 1. P = Anfangsalignment
  (e.g. basierend auf einem Sequenzalignment)
- 2. Überlagern von Strukturen A und B auf der P-Basis
- 3. Berechnen einer Scoringmatrix R, basierend auf den Entfernungen der Reste
- 4. Dynamische Programmierung verwenden zur Ausrichtung A und B mit Scoring matrix R
- 5. P‘ = Neualignment, abgeleitet aus dem dynamischen Programmierschritt
- 6. Wenn sich P ' von P unterscheidet, dann wieder zu Schritt 2 zurückspringen

Entfernungsbasierte Scoringmatrix

- Sei $d(a_i, b_j)$ die euklidische Distanz zwischen a_i *und* b_j
- Sei t die obere Abstandsgrenze für die Reste a_i *und* b_j

- Die Scoring-Matrix R wird wie folgt definiert :

$$R(a_i, b_j) = 1 / d(a_i, b_j) - 1 / t$$

if $R(A_i, B_j) >$ max. Score, dann $R(A_i, B_j) =$ max. score

- Die Gap / Mismatch-Strafe wird auf 0 gesetzt

Iteriert DP Algorithmus - Pseudocode

Maximale Anzahl von Zyklen

Initiale Äquivalenz

Wert in der R Zelle

```
Const pmax, Einit
Var   p,R,Ep,T
proc
dist(ai,bj) distance between residues
Score(d)    calculate score from a distance
begin
        E0:=Einit; p:=0
        repeat
                T:=transformaiton for RMSD
                A*:=T(A)  superimpose A to B giving A*
                # calculate the scoring matrix R
                forall pairs(i,j) do Rij:=score(dist(ai*,bj)) end
                (s,P):=AR(A,B)using R (find path P with score s)
                 p:=p+1
                Ep:={(ai1,bj1), ...(air, bjr)}
        while Ep!=Ep-1 and p < pmax
end
```

Doppelte Dynamische Programmierung

- **Ziel**: gleichzeitiges alignieren und überlagern der Strukturen
- Doppelte Dynamische Programmierung* ist eine Heuristik, die dieses Ziel zu erreichen versucht
- Implementiert als Teil von SSAP (z. B. durch CATH verwendet)

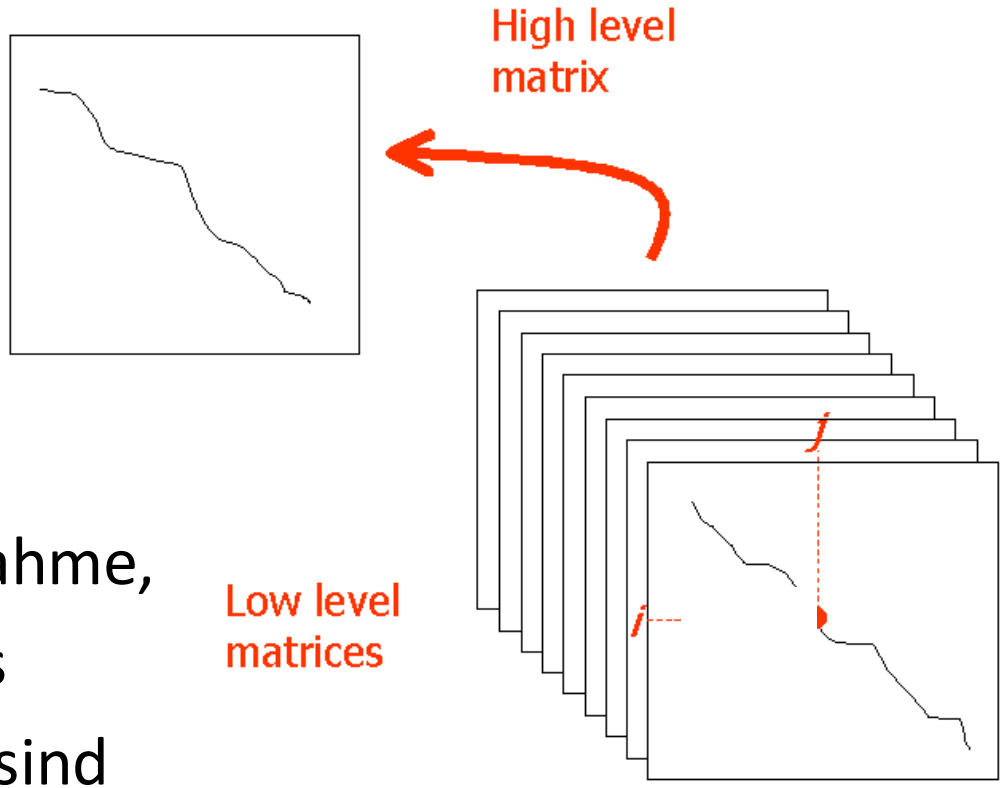
* Wird verwendet, wenn die Sequenzähnlichkeit zwischen A und B sehr gering ist

Doppelte Dynamische Programmierung

- Die bisheringen Methoden sind schwer zu verwenden, wenn wir kein gutes Alignment haben
- Die Idee bestehet darin, zwei Ebenen von DP zu verwenden
  - **Erste Stufe (low level)**
    - die zwei Strukturen auf viele verschieden Weisen überlagern (DP)
    - Wie? Angenommen jedes Paar von Resten (ai und bj) matcht, dann versuche di anderen Reste zu alignieren
    - Finde das optimale Alignment unter der Bedingung, dass a_i and b_j aligniert sind
    - Berechne die *low level scoring Matrizen* für alle Kombinationen von a_i und b_j
  - **Zweite stufe (high level):** kombiniere die Ergebnisse aus low level DP
    - Aufsummieren in der *high-level Matrix*
    - Wie? Fülle den Score nur wenn die Zellen auf dem optimalen Weg liegen (Summe über alle möglichen Low-Level Matrizen)
    - Der Score wird von den Low-level zu High-Level Matrizen übertragen

Idee der DDP (mit anderen Worten)

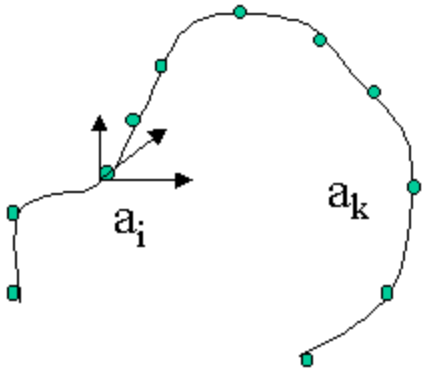
- Verwende zwei Ebenen von Dynamischer Programmierung:
  - *High level*, die low level DP aufsummiert
  - *Low level*, die Alignments werden erzeugt unter der Annahme, dass a_i und b_j Teil eines optimalen Alignments sind



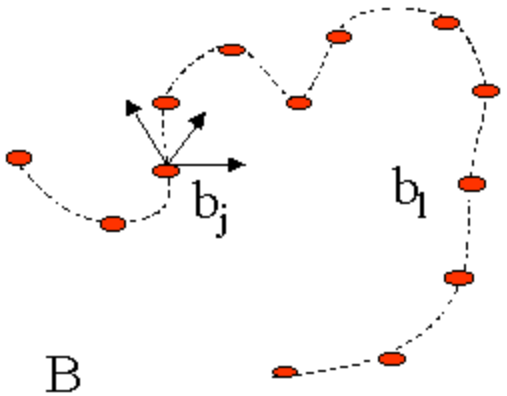
Low level Matrix

- ${}^{ij}R$ ist die low level scoring Matrix angenommen, dass a_i and b_j aligniert sind
- ${}^{ij}R_{kl}$ ist der Score, der zeigt wie gut a_k auf b_l passt unter der Bedingung, dass a_i and b_j aligniert sind
- Ausführen der Dynamischen Programmierung für alle Matrizen (${}^{ij}R$ benutzen) unter der Bedingung dass das optimale Alignment (i.j) enthält

Example

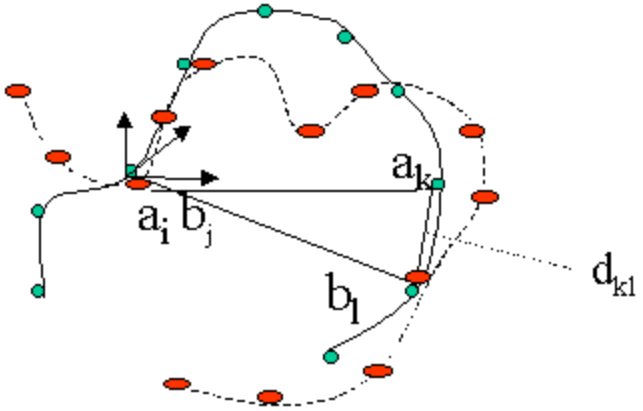


Structure A



B

Koordinatensystem
mit Ursprung an ai und bj



Einfacher Score: eine Funktion des Abstandes zwischen ak und bl

Low level Matrix - Beispiel

	H	S	E	R	R	H	V	F
G	33	37	22	13				
Q	27	31	26	17				
V	22	26	30	21				
G	13	17	21	**25**	21	17	13	9
M				21	34	30	26	22
A				17	30	34	30	36
C				13	26	31	34	32

Alignment:

H-SERRHVF-

GQV-GM--AC

Wie ist der maximale Score hier?
Und die gap Strafe?

	H	S	E	R	R	H	V	F
G	25	17						
Q	29	21						
V	21	**25**	21	17	13	9	5	1
G		21	17	28	24	20	16	12
M		17	22	24	23	28	24	20
A		13	18	23	24	24	33	29
C		9	14	19	20	24	29	43

Alignment:

HSE-RRHVF-

-GQVG-M-AC

Rauschen

- Wie können die low-level Matrizen zur High-level Matrix kombiniert werden? Alle Werte in den Zellen sammeln? Zu viel Rauschen..

- Die high-level Matrix R ist die Summe aus n*m low-level Matrizen. Die meisten von Ihnen enthalten Paare, die nicht auf dem optimalen Alignment liegen (sie fügen Rauschen hinzu..)

- Nur Paare, die auf dem optimalen Alignment liegen, werden berücksichtig

High level Matrix - Beispiel

	H	S	E	R	R	H	V	F
G	58	37	0	0	0	0	0	0
Q	29	0	26	0	0	0	0	0
V	0	25	51	0	0	0	0	0
G	0	0	0	**53**	24	0	0	0
M	0	0	0	0	34	58	26	0
A	0	0	0	0	0	0	33	36
C	0	0	0	0	0	0	0	75

High level
Scoring Matrix

	H	S	E	R	R	H	V	F
G	58	37						
Q	29	58	63	63	63	63	63	63
V		58	109	109	109	109	109	109
G		58	109	162	162	162	162	162
M		58	109	162	196	220	220	220
A		58	109	162	196	220	253	256
C		58	109	162	196	220	253	328

High level
Dynamische Programmierung

Algorithmus für DDP

Erste Stufe DP

Zweite Stufe DP

$R := \{0\}$ *High level scoring matrix 0*

for each pair (a_i, b_j) **do**

compute the low level scoring matrix ${}^{ij}R$

$(s,P) := \boldsymbol{DP}^{*}_{{}^{ij}R}(A,B)$ *Low level DP forced through* (a_i, b_j)

P is the optimal path, s the score

forall (a_p, b_q) in P **do** $R_{pq} := R_{pq} + {}^{ij}R_{pq}$ *Accumulate into R*

end

$(s,P) := \boldsymbol{DP}_R(A,B)$ *High level DP using R*

Multiple structural alignment

Warum?

- Ähnlichkeiten in eine Menge von homologen Strukturen zu finden
- Gemeinsame Kerne / Strukturmotive und lokale Patterns zu finden
- Protein Klassifizierung in Falten und Familien druchzuführen

Wie?

- Wir brauchen eine multiple Äquivalenz zwischen den Resten
- Wähle eine Struktur als Basis ({A1,A2,....Am} Strukture, A1 basis)
  - Überlagern alle Strukturen auf der Basis
  - Zuerst überlagern die Massenzentren
  - Dann minimire:

$$\sum_{j=2}^{m} \sum_{i=1}^{r} \omega_{ij} \left(R_j A_i^j - A_i^1 \right)^2$$

Finde die beste Rotationmatrix für jede Struktur

r = Anzahl von Atomen; m = Anzahl von Strukturen; w_{ij}= Gewichte von Resteskoordinaten

Proteinklassifizierung

Warum Strukturen klassifizieren?

- Die strukturelle Ähnlichkeit ist eine guter Indikator für Homologie (in allgemeinen die Struktur mehr conserviert als Sequenz ist)
- Klassifizierung ist auf vershiedenen Ebenen durchgeführt
  - **Ähnliche Falten** (Strukture nicht unbedingt verwandte)
  - Nidrige Sequenzähnlichkeit, aber **ähliche Struktur und Funktion** implizieren sehr wahrscheinlich **Homologie**)
  - **Hohe Sequenzähnlichkeit** beteutet, ähnliche Strukturen und Homologie
- Klassifizierung kann helfen, um evolutionäre Beziehungen zu untersuchen und ähnliche Funktione herauszufinden

Strukturelle Klassifizierung

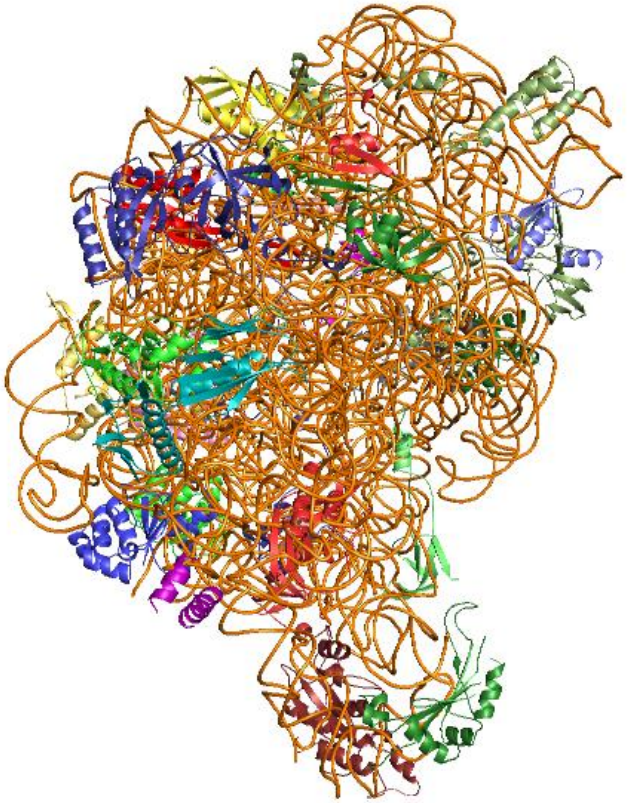
- **SCOP**: Structural Classification of Proteins
  - **Hand kuratiert** mit einigen Automatisierungen (Alexei Murzin, Cambridge)
- **CATH**: Class, Architecture, Topology, Homology
  - Automatisiert, wenn möglich. Einige Prüfungen werden von Hand gemacht (Orengo, UCL, London)
- **FSSP**: Fold basierend auf Structure-Structure Alignment von Proteinen
  - vollautomatisch

Ziel der Klassifizierung

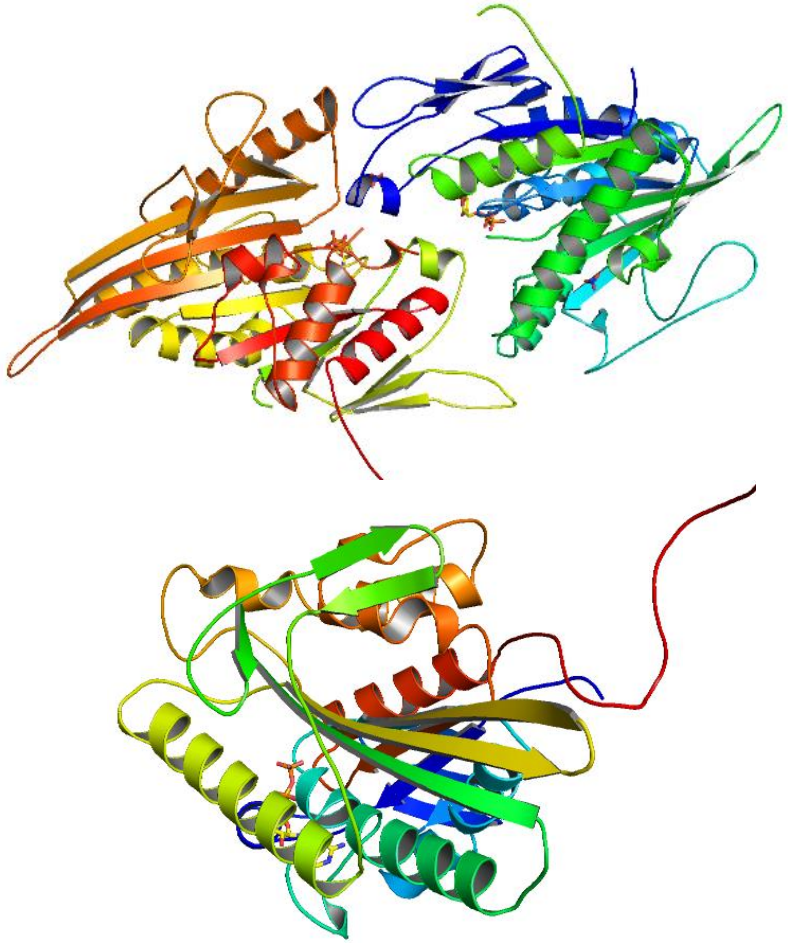
Evolutionär verwandten Strukturen in Clustern verpacken, die ähnliche Falte oder ähnliche Sekundärstruktur haben

Was ist eine Proteindomäne?

“Eine Proteindomäne ist ein Bereich eines Proteins mit stabiler, meist kompakter Faltungsstruktur, der funktional und strukturell (quasi-)unabhängig von benachbarten Abschnitten ist”
by Wikipedia



Multi-Domäne Protein



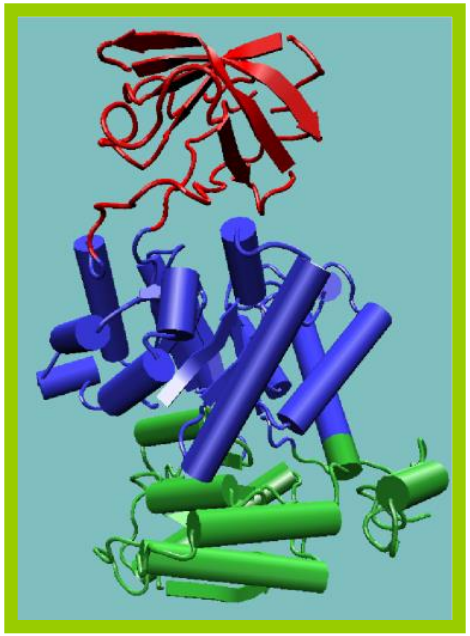
Einzeldomäne Protein

Was ist eine Proteindomäne?

- **Functionelle**: eine Domain ist eine unabhängige **Funktionseinheit**, die in mehr als ein Protein gefunden ist.
- **Chemie**: Domäne haben eine **hydrophoben Kern**
- **Topologie**: **Intra-Domain** Abstände von Atomen sind **minimal**, Interdomain Abstände maximal

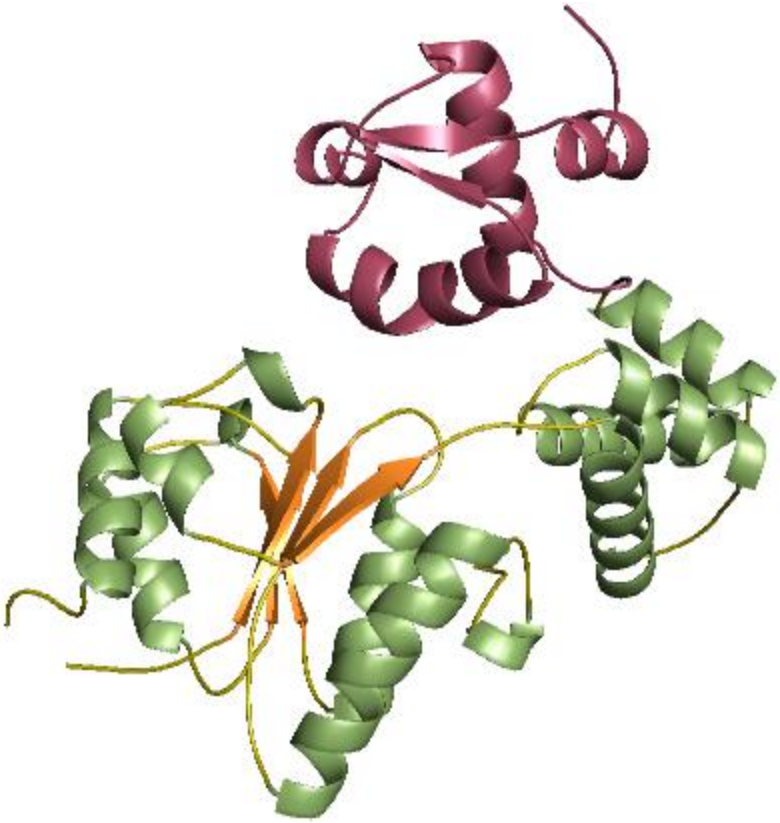
Es ist schwierig, genau zu Domains Grenzen zustimmen

Proteindomäne

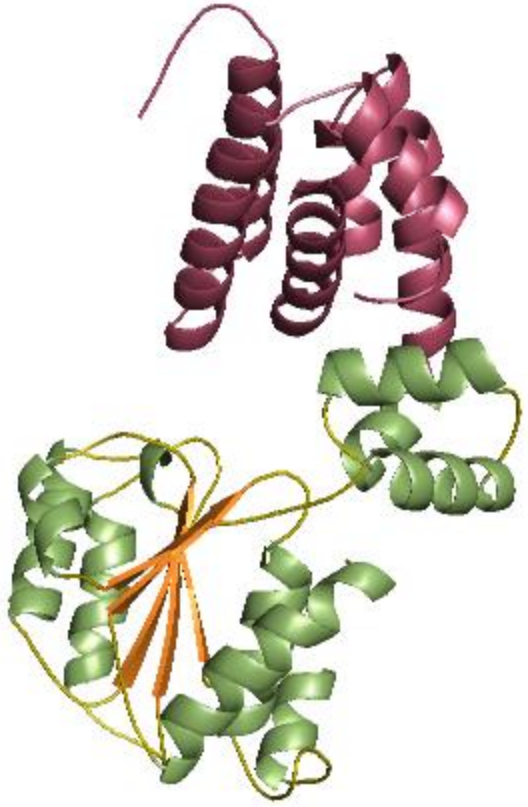


SKSHSEAGSAFIQTQQLHAAMADTFLEHMCRLDIDSAPITARNTG
IICTIGPASRSVETLKEMIKSGMNVARMNFSHGTHEYHAETIKNV
RTATESFASDPILYRPVAVALDTKGPEIRTGLIKGSGTAEVELKK
GATLKITLDNAYMAACDENILWLDYKNICKVVEVGSKVYVDDGLI
SLQVKQKGPDFLVTEVENGGFLGSKKGVNLPGAAVDLPAVSEKDI
QDLKFGVDEDVDMVFASFIRKAADVHEVRKILGEKGKNIKIISKI
ENHEGVRRFDEILEASDGIMVARGDLGIEIPAEKVFLAQKMIIGR
CNRAGKPVICATQMLESMIKKPRPTRAEGSDVANAVLDGADCIML
SGETAKGDYPLEAVRMQHLIAREAEAAMFHRKLFEELARSSSHST
DLMEAMAMGSVEASYKCLAAALIVLTESGRSAHQVARYRPRAPII
AVTRNHQTARQAHLYRGIFPVVCKDPVQEAWAEDVDLRVNLAMNV
GKAAGFFKKGDVVIVLTGWRPGSGFTNTMRVVPVP

Domäne in verschiedenen Proteinen



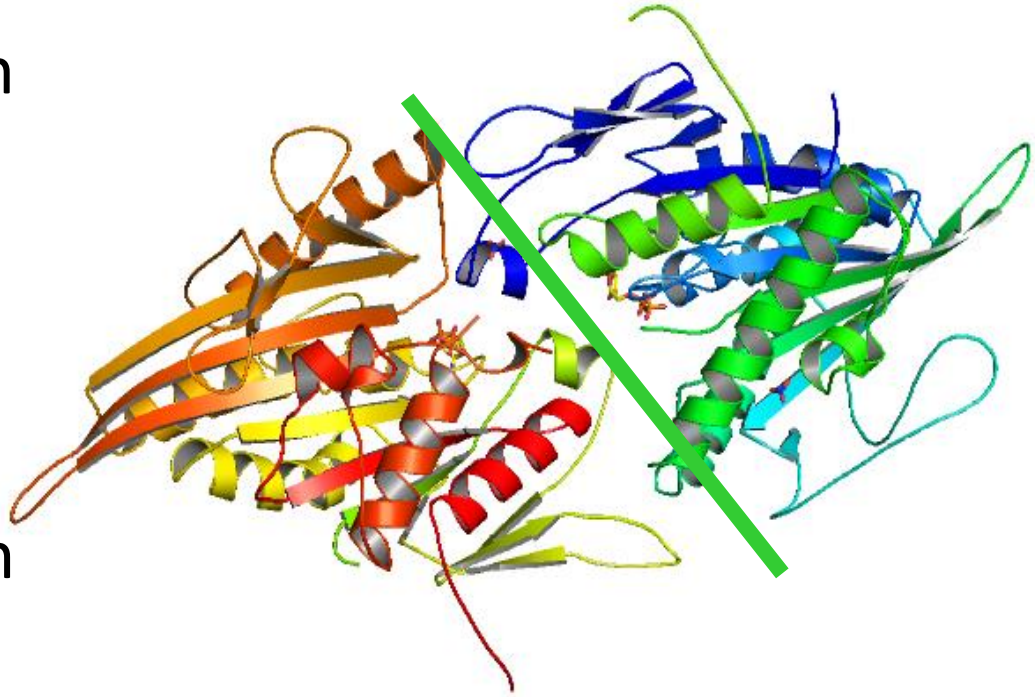
1in5: interaction of P-loop domain (green & orange) and winged helix DNA binding domain



1a5t: interaction of P-loop domain (green & orange) and DNA polymerase III domain

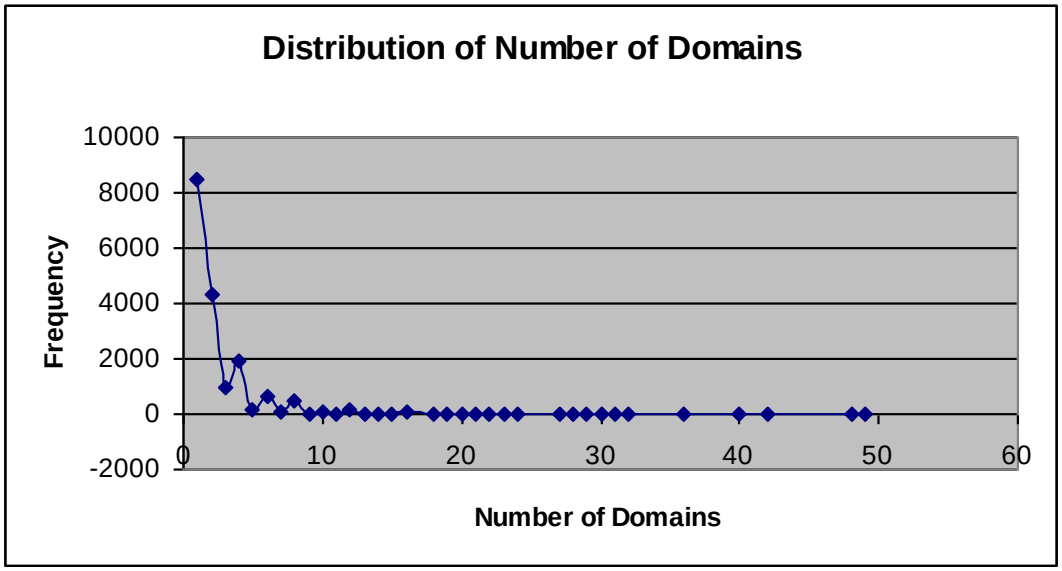
Intradomain Abstände minimal

- Entfernungen zwischen den Atomen innerhalb der Domäne sind minimal
- Entfernungen zwischen den Atomen von zwei verschiedenen Domänen sind Maximal



PDB, Proteine und Domäne

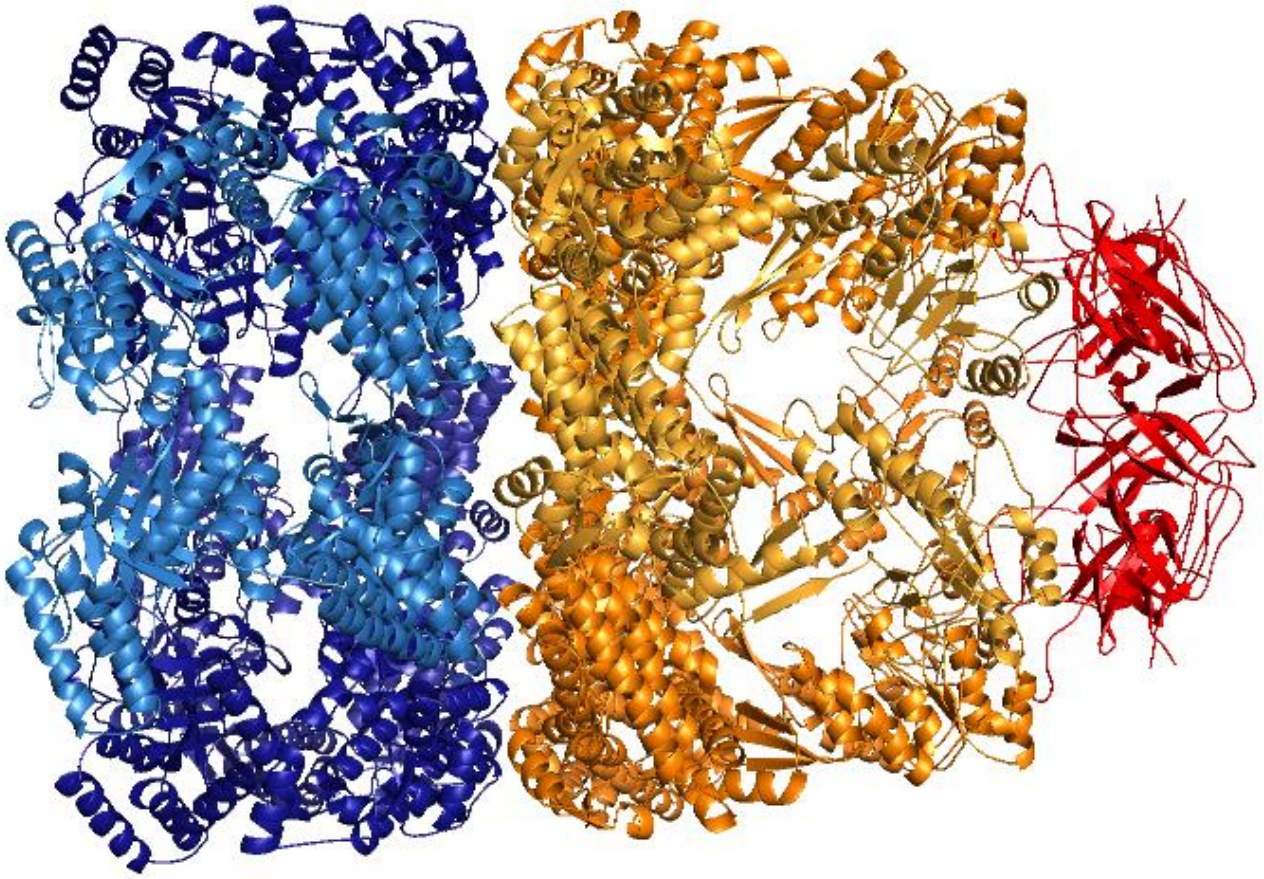
- Ca. 30.000 Strukturen in PDB
  - 50% einzige Domäne
  - 50% mehrere Domäne
  - 90% haben weniger als 5 Domäne



Eine Struktur mit 49 Domäne

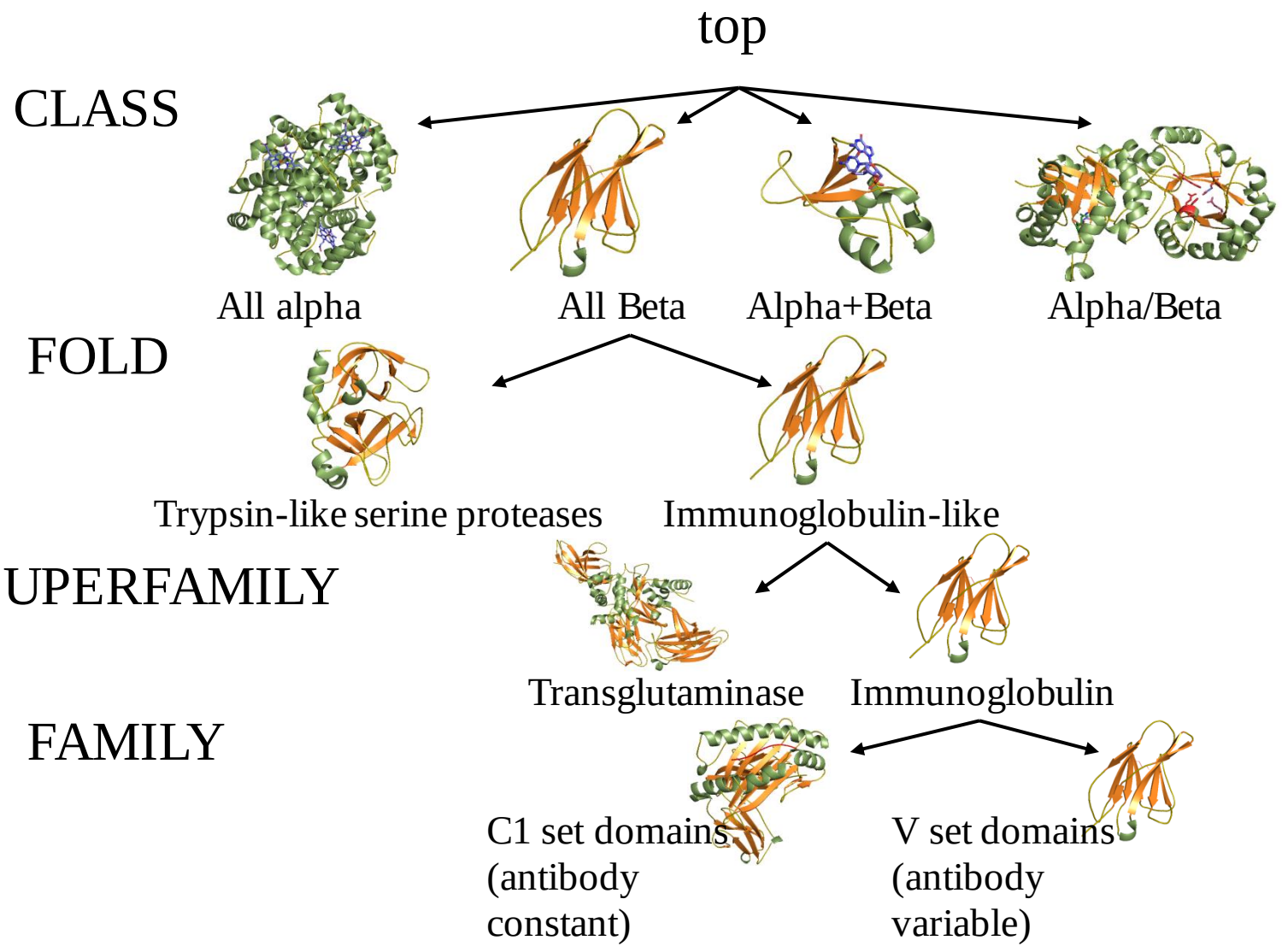
1AON, Asymmetrisch Chaperonin Complex

Groel/Groes/(ADP)7



SCOP: Structural Classification of Proteins

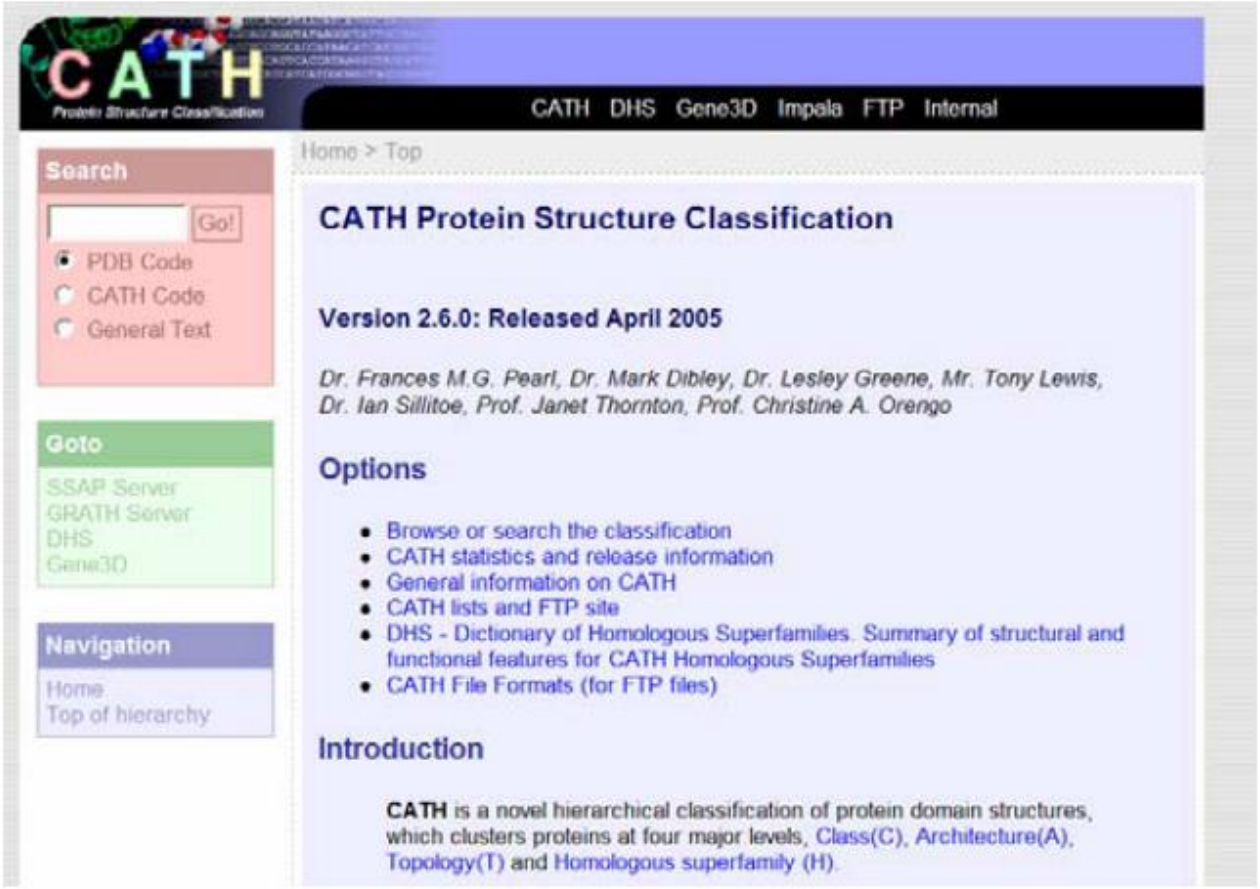
http://scop.mrc-lmb.cam.ac.uk/scop/



Falte, Superfamily, Family

- **Falte**
  - **gemeinsame Kernstruktur**
    - i.e. die selbe Sekundärstrukturelementen in der gleichen Anordnung mit dem gleichen topologischen Struktur
- **Superfamily**
  - Sehr ähnliche Struktur und Funktion
- **Family**
  - Sequenzidentität (>30%) und sehr ähnliche Struktur und Funktion

Klassifizierung von Proteinstrukturen mit CATH

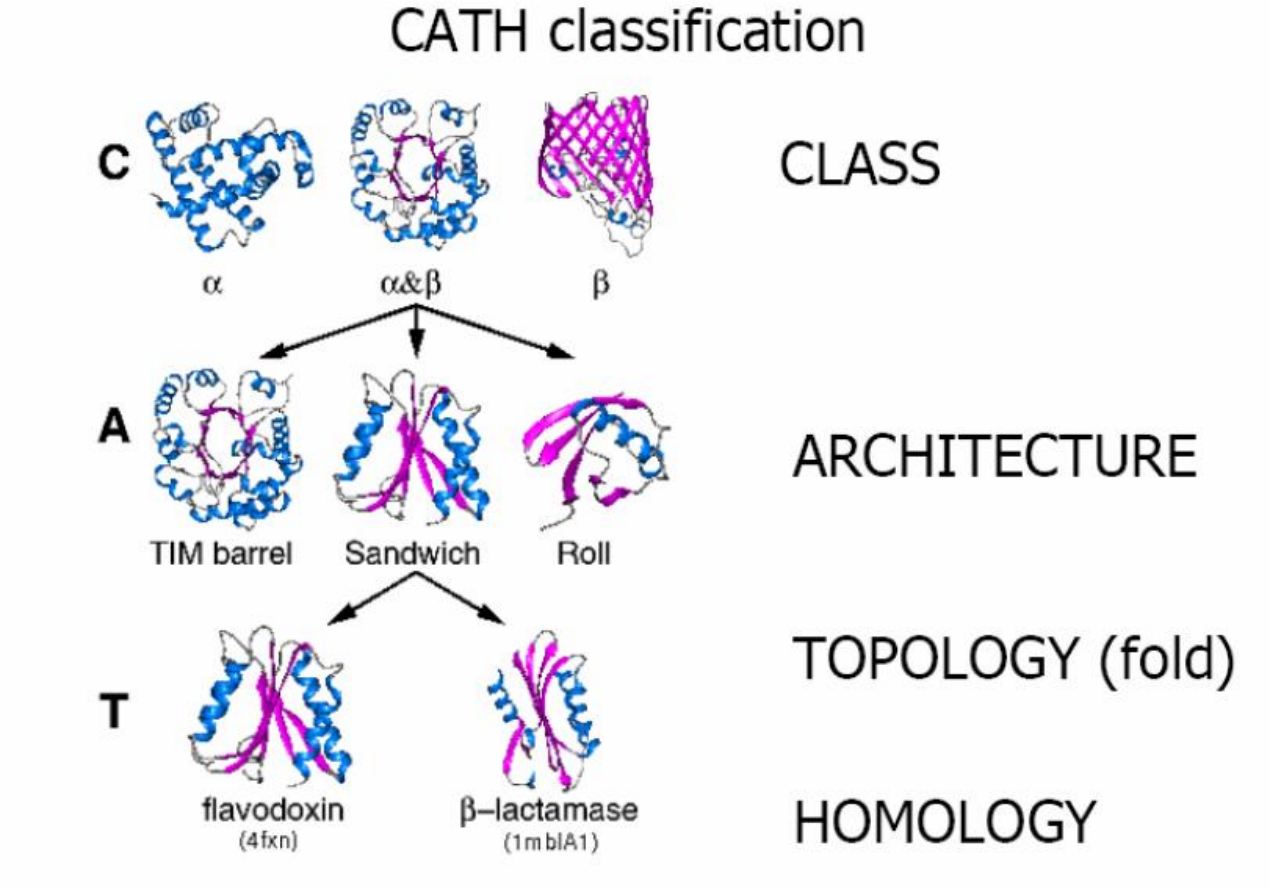


http://www.biochem.ucl.ac.uk/bsm/cath

CATH Hierarchie

Schwerpunkt: Clustering von Proteinen mit ähnlichen Falten.

Klassifizierung wird von einzelnen Proteindomänen durchgeführt



CATH

- **Klassen**:

  überwiegend alpha, überwiegend beta, gemischt alpha/beta, kaum regelmäßige Sekundärstruktur

- **Architektur**:

  Anordnung und Orientierung von Sekundärstrukturelementen, unabhängig von der Topologie

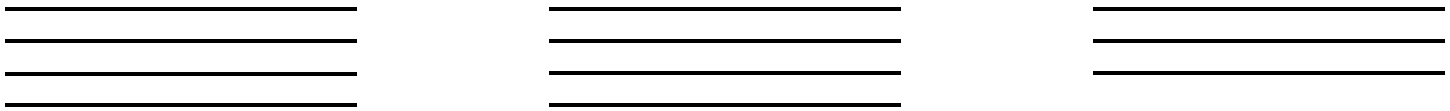
- **Topologie** (Falten Familien)

  Faltengruppe = Form und Konnektivität

- **Homologie**: evolutionäre Verwandte

CATH generieren

1) Gruppiere Ketten von PDB in Sequenzfamilien mit wenigstens 35% Sequenzidentität (Homologe)



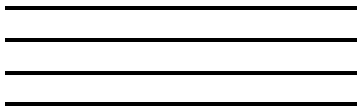
Familie 1 > 35% Identität

Familie 2 > 35% Identität

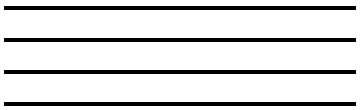
Familie 3 > 35% Identität

CATH generieren

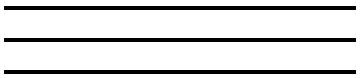
1) Gruppiere Ketten von PDB in Sequenzfamilien mit wenigstens 35% Sequenzidentität (Homologe)



Familie 1 > 36% Identität

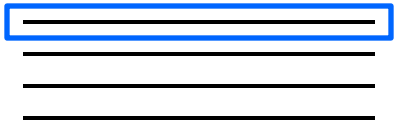


Familie 2 > 36% Identität



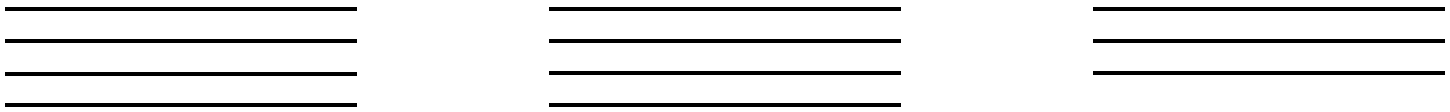
Familie 3 > 36% Identität

2) Wähle einen Vertreter pro Familie aus und finde die Domänen der Kette des Vertreters



CATH generieren

1) Gruppiere Ketten von PDB in Sequenzfamilien mit wenigstens 35% Sequenzidentität (Homologe)

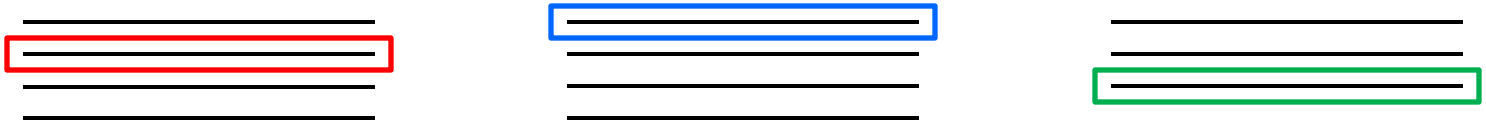


Familie 1 > 36% Identität

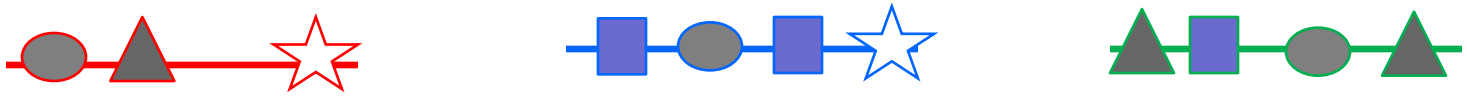
Familie 2 > 36% Identität

Familie 3 > 36% Identität

2) Wähle einen Vertreter pro Familie aus und finde die Domänen der Kette des Vertreters

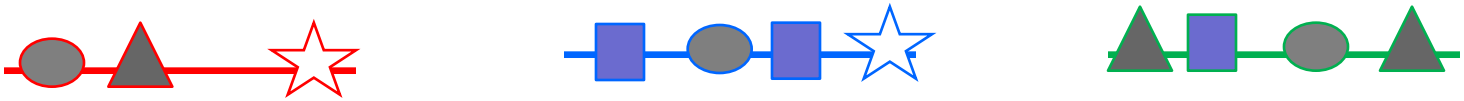


3) Finde die Domäne für jeden Vertreter jeder Familie

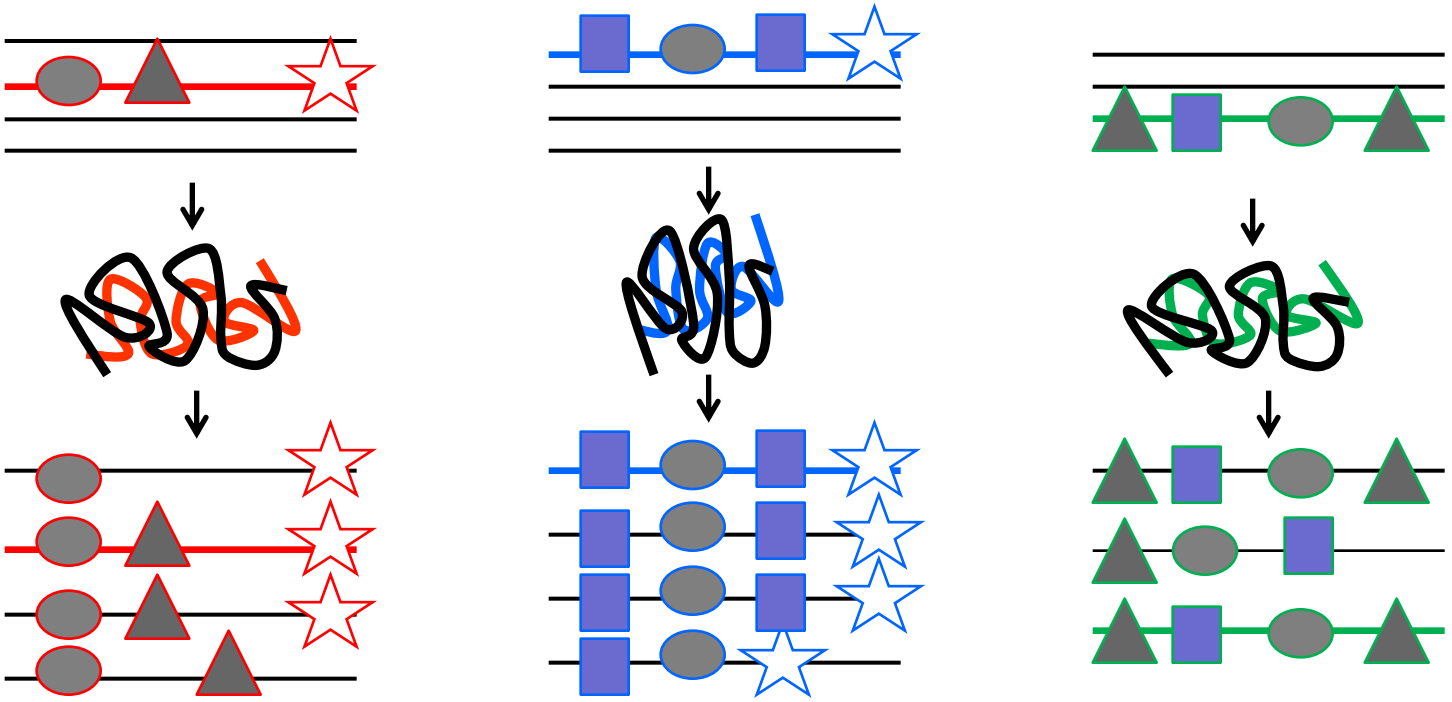


CATH generieren

3) Finde die Domäne für jeden Vertreter jeder Familie

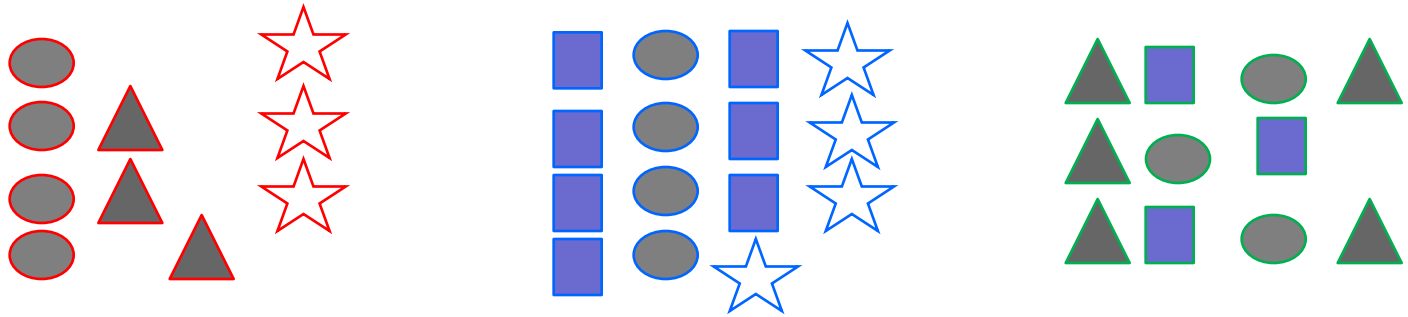


4) Suche Domänen jeder Kette – Berechne das strukturelle Alignment zwischen jeder Kette und dem Vertreter (mit SAP, DDP) – lass die Kette die Domänedefinition beim Vertreter erben

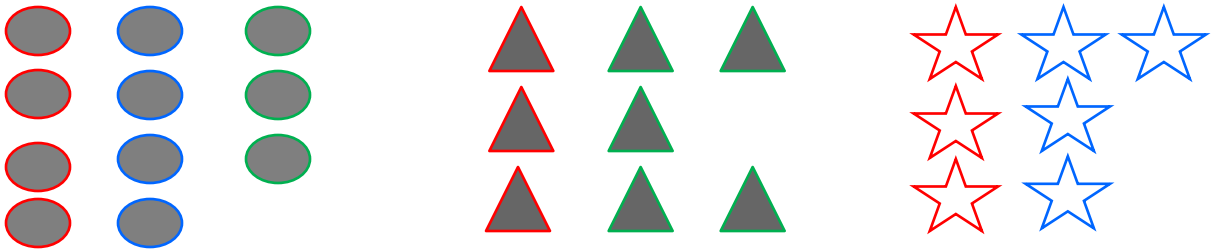


CATH generieren

5) Teile die Ketten von Domänen durch die oben gefundenen Domänengrenzen

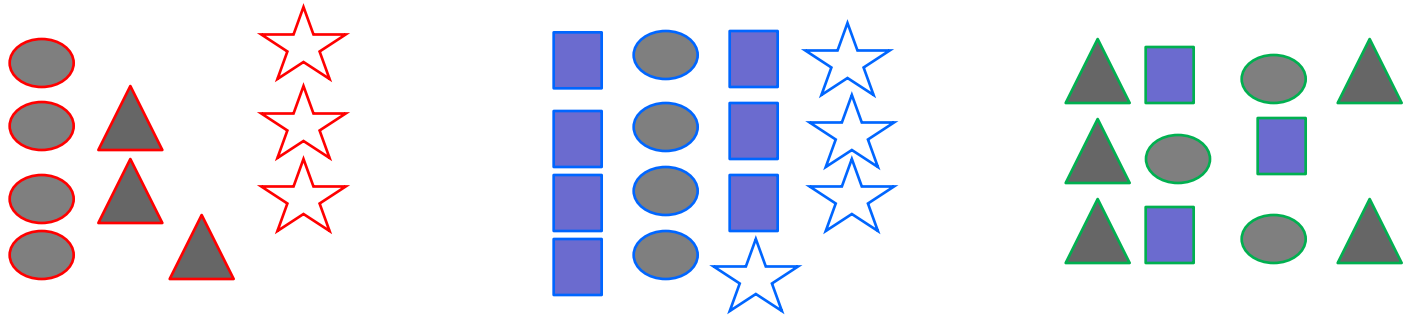


6) Gruppiere Domänen in Sequenzfamilien (35% Sequenzidentität)

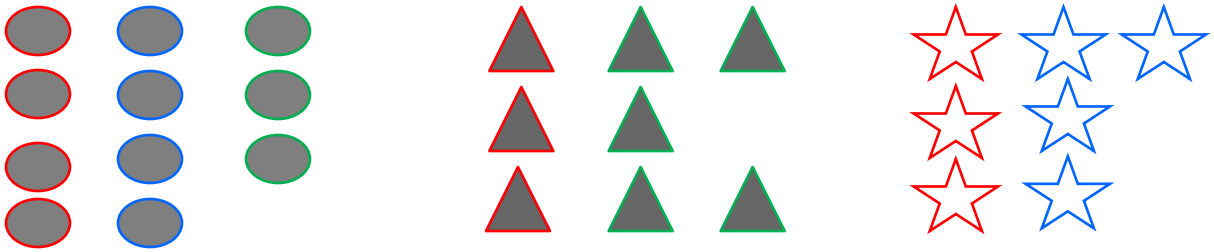


CATH generieren

5) Teile die Ketten von Domänen durch die oben gefundenen Domänengrenzen



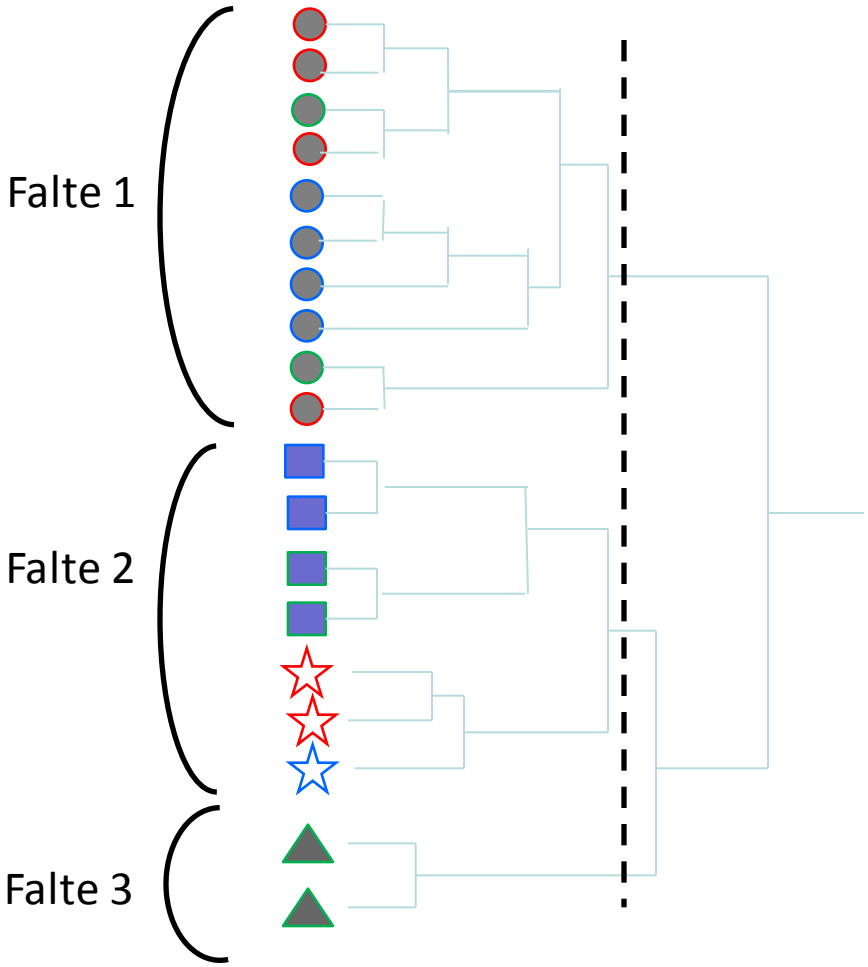
6) Gruppiere Domänen in Sequenzfamilien (35% Sequenzidentität)



7) Weise eine Klasse zu jeder Domäne zu

CATH generieren

8) Cluster in Superfamilien und Falten basierend auf Strukturvergleich (SSAP, DDP)



9) Manuelle Zuweisung der Architektur

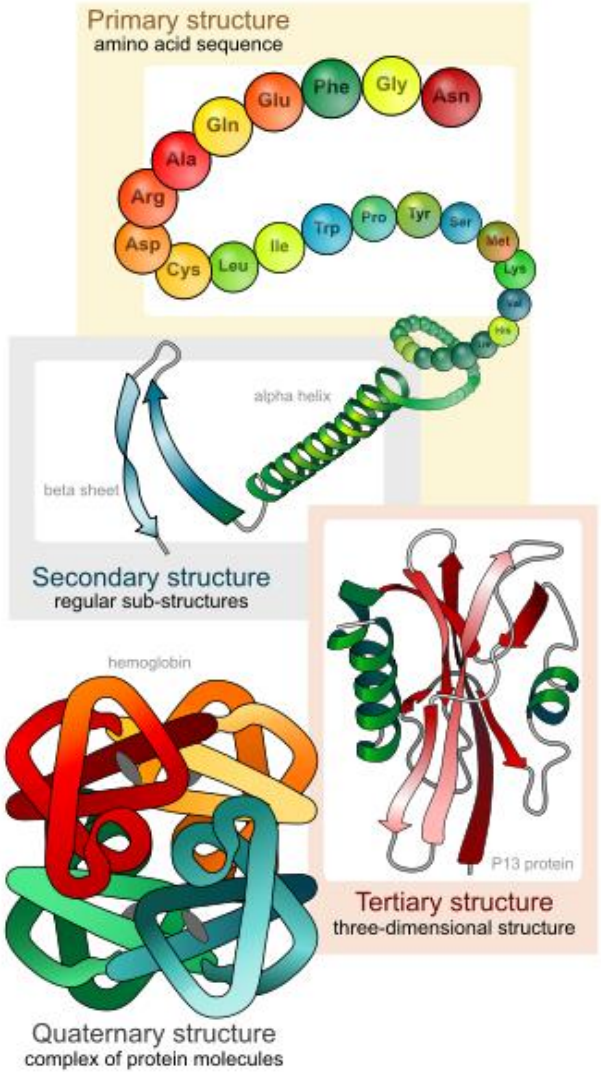
Proteinfaltung

Proteinfaltung - Strukturvorhersage

Ziel: Die Strukturvorhersage versucht, Modelle von 3D-Strukturen von Proteinen zu bauen, die nützlich für das Verständnis von Struktur-Funktions-Beziehungen sind.

- Wenn die Struktur verfügbar ist, ist es leichter, die Position der aktiven Stellen zu vermuten

Von der Primärstruktur zur gefalteten Struktur



Monomer = Protein aus nur einer Kette
Multimer = Protein aus mehr als einer Kette

Dimer = Protein aus zwei Ketten
Homo-dimer = Protein aus zwei identischen Ketten
Hetero-dimer = Protein aus zwei unterschiedlichen Ketten

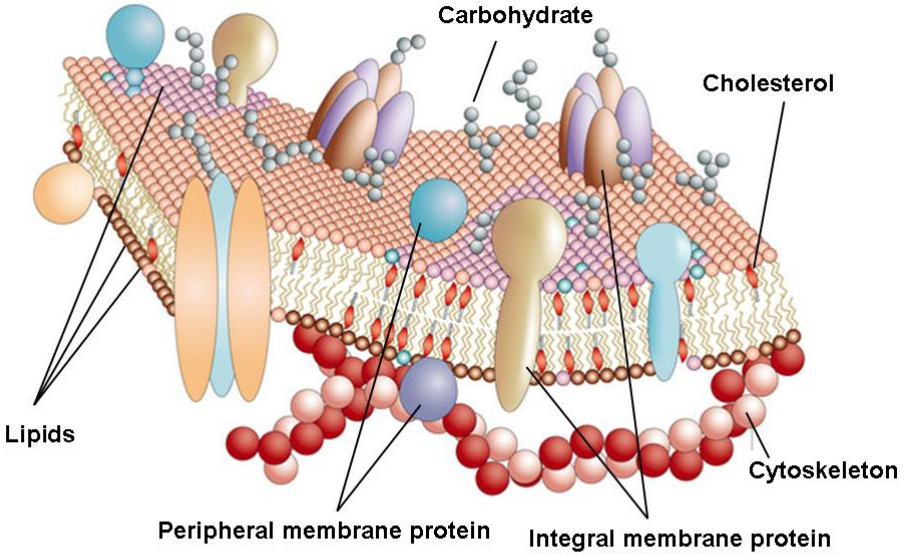
Image taken from Wikipedia

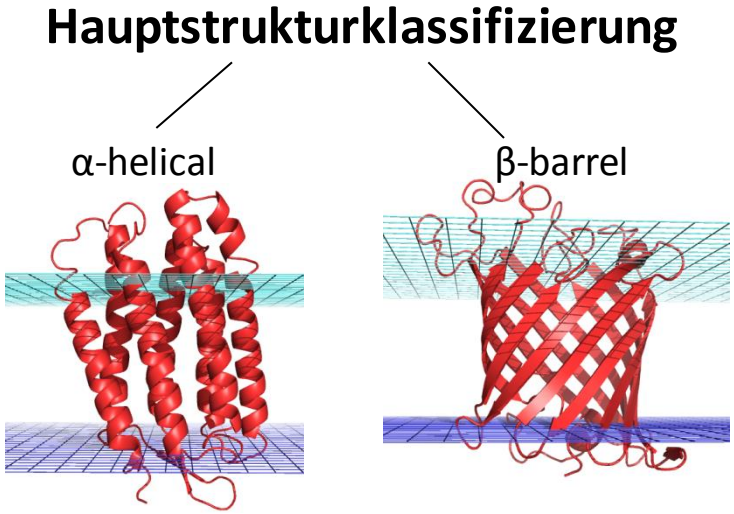
Wie wird eine Struktur gefaltet?

- **Alle Reste** müssen **erlaubte Konformationen** haben
- **Verschüttete polare Atome mussen in Wasserstoffbrücken gebunden sein**
  - Wasserstoffbrücken mit dem Lösungsmittel sind auch energetisch günstiger
- **Genug hydrophobe Oberflächen muss begraben und das Innere des Proteins muss dicht gepackt werden**
- Es gibt Anzeichen dafür, dass Faltung hierarchisch auftritt: zu erst Sekundärstrukturelemente, dann supersekundär, ...
- Dies rechtfertigt die Verwendung von hierarchischen Methoden bei der Simulation von Falten

Membranproteine

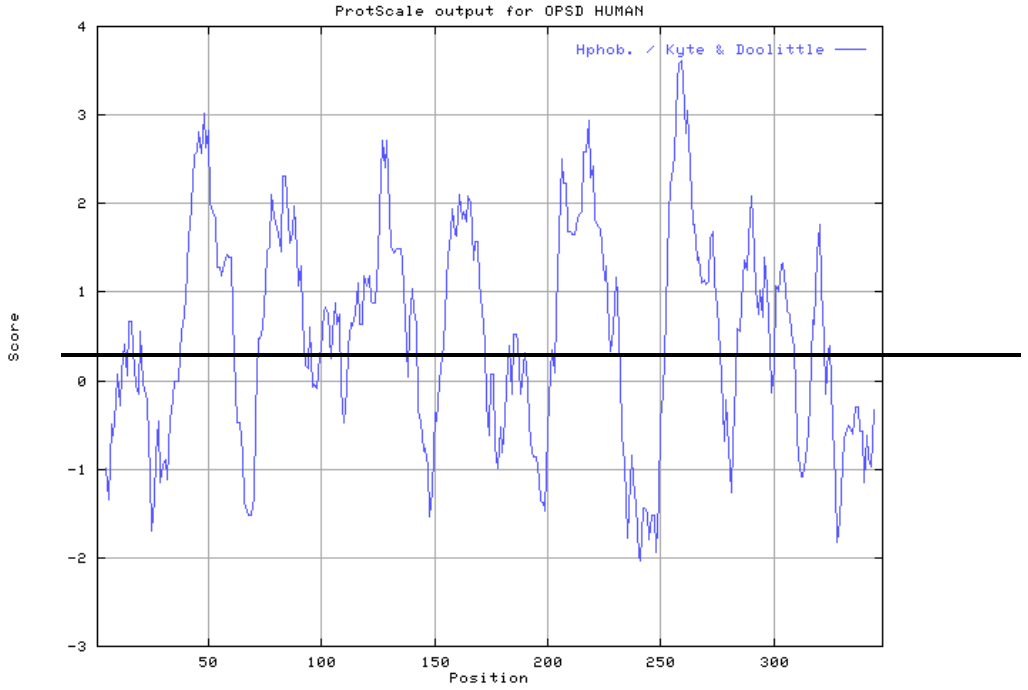
- Sie bilden 20-30% der Gene in einem typischen Genom
- Sie werden in vielen zellulären Prozessen wie Transport und Signalisierung beteiligt
- Mutationen führen zu schweren Erkrankungen, beispielsweise zystische Fibrose, Nachtblindheit - die den Großteil der Medikamenten-Ziele darstellen
- Problem: wenige bekannte Strukturen (< 2% in Protein Data Bank)





Hydrophobie plot: Kyte und Doolittle Ausmaß (JMB 1982)

SOAP, ProtScale: Programm, um den Hydropathie-Charakter einer Proteinsequenz zu plotten.
Basierend auf die Kyte und Doolittle Ausmaß
Hydropathie-Werte werden auf der Basis verschiedener experimenteller Parameter berechnet



Hintergrund Hydrophobie

Hydrophobizitätsplot des menschlichen Rhodopsin durch Expasy ProtScale erstellt.
Fenstergrösse = 9, Kyte und Doolittle Ausmaß

Verfahren der Proteinmodellierung

- ***Ab initio* Methode**

  Lösung eines Proteinfaltungsproblems
  Suche in Konformationsraum
  - Energieminimierung

- ***Knowledge-based* Methode:**
  Homologiemodellierung
  Faltenerkennung (Threading)

Homologiemodellierung

Beim Sequenzidentität von mehr als 30%

Wie wird es gemacht: Hauptschritte

- Identifiziere eine/mehrere Vorlage(n) - anfängliches Alignment
- Das Alignment (manuell) verbesseren
- Rückgrat-Erzeugung
- Modellierung von Schleifen
- Seitenkettenmodellierung
- Verfeinerung
- Validierung

Vorlagenidentifikation

- Suche mit der Sequenz
  - Blast
  - Psi-Blast

→ **Choose one with highest similariy Where structure is known, where structure Is good**

- Benutze biologische Information
- Benutze funktionale Annotation in Datenbanken
- Aktive Stelle/Motive

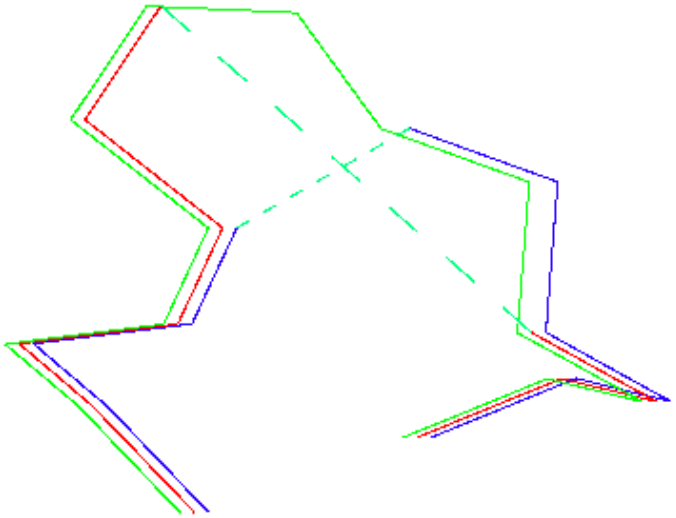
Vorlagenidentifikation – Initiales Alignment*

FNICRLPGSADAVC	Original Sequenz
FNVCRTP---DAIC	Homolgous Sequenz- Alignment 1
FNVCR---TPDAIC	Homolgous Sequenz- Alignment 2 Verbessertes Alignment

* Manchmal kann man mehr als eine Vorlage nutzen

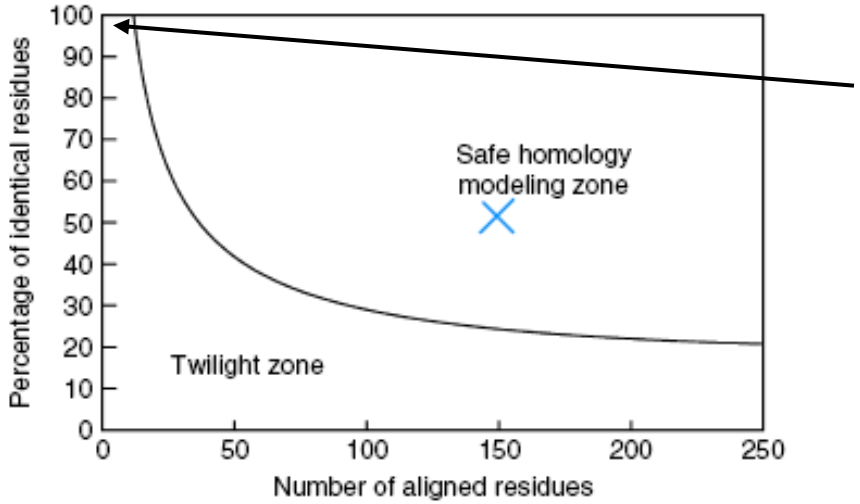
Vorlagenidentifikation – Initiales Alignment*

FNICRLPGSADAVC	Original Sequenz
FNVCRTP---DAIC	Homolgous Sequenz- Alignment 1
FNVCR---TPDAIC	Homolgous Sequenz- Alignment 2 Verbessertes Alignment



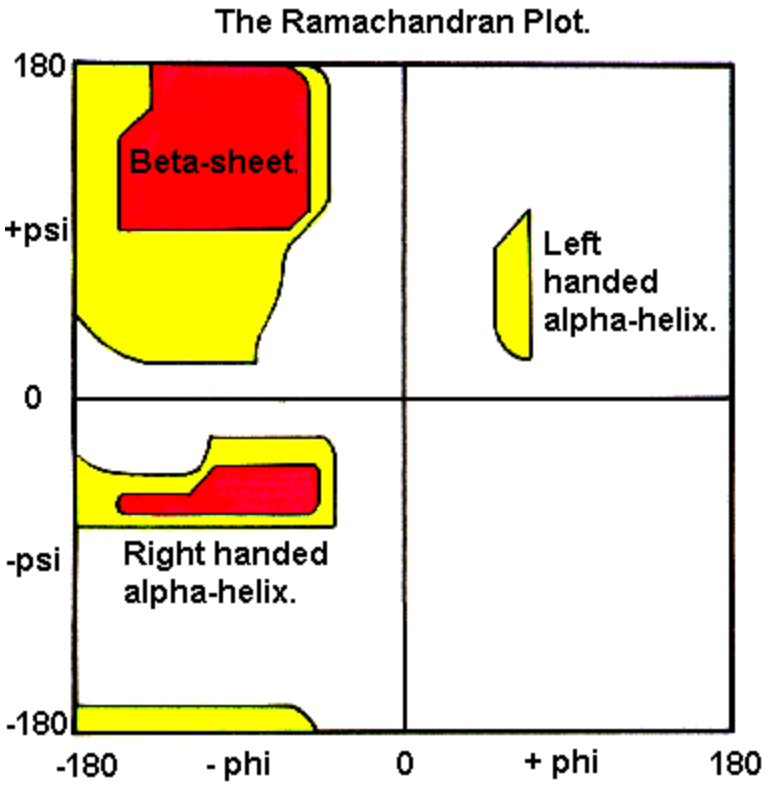
Qualität der Vorlagen

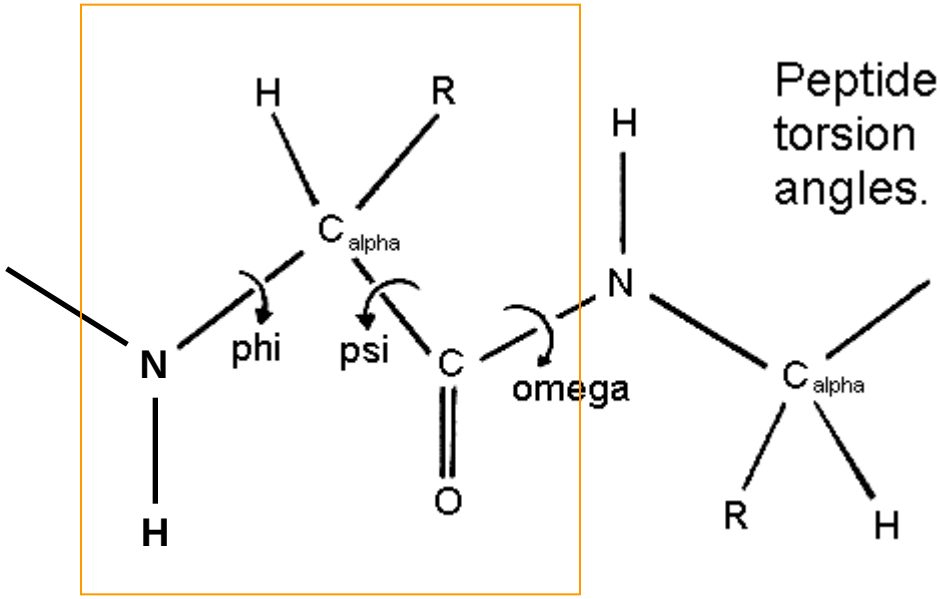
- Die auswahl des besten Vorlage ist entscheidend!
- Die beste Vorlage kann auch die sein, die nicht die höchste id % hat (bester p-value…)
  - Vorlage 1: 93% id, 3.5 Å Auflösung ☹
  - Vorlage 2: 90% id, 1.5 Å Auflösung ☺



Qualität der Vorlagen – Ramachandran Plot

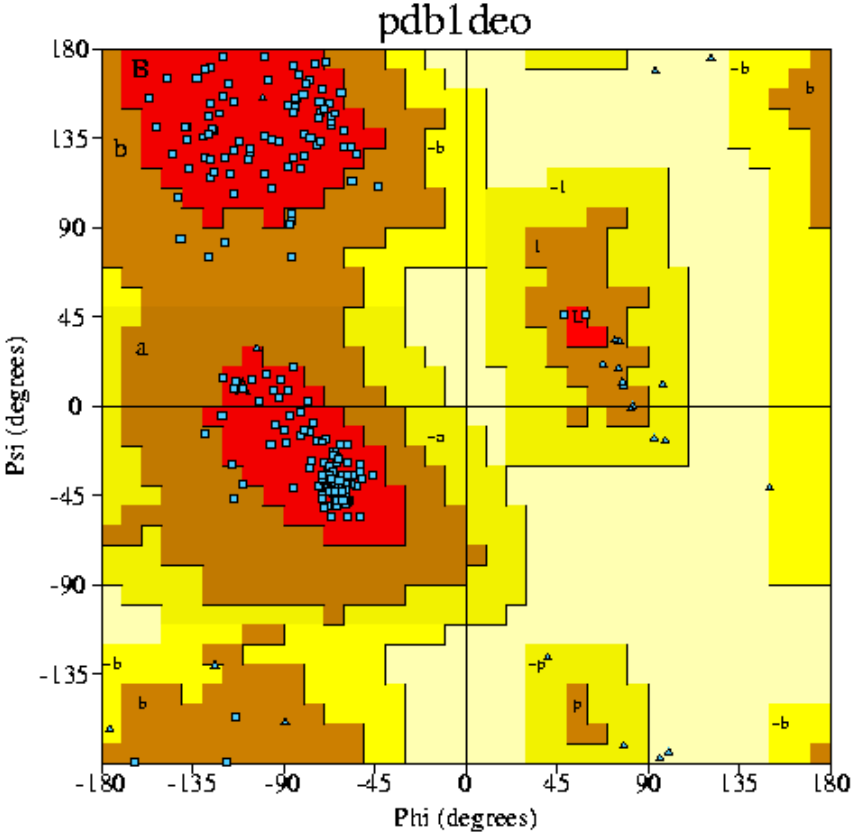
- **erlaubte Rückgrat Torsionswinkel in Proteinen**



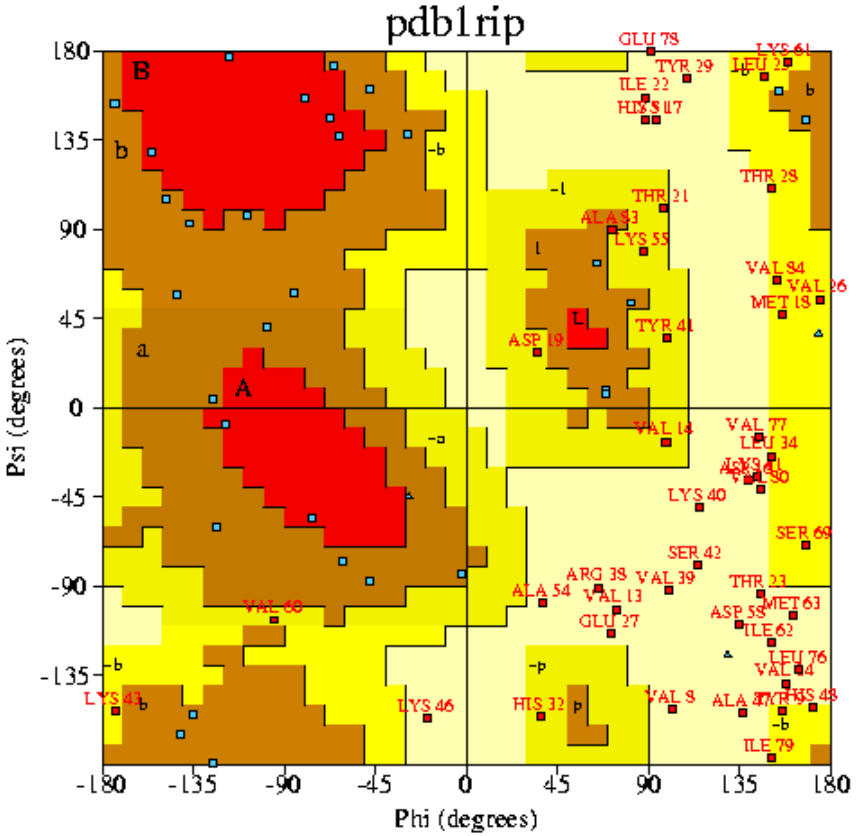


Aminosäure Reste

Qualität der Vorlagen – Ramachandran Plot



X-ray Struktur – Gute Daten.



NMR Struktur – Niedrige Qualität der Daten…

Erzeugung der Rückgratstruktur

- Generiere die Koordinaten der Rückgratatome aus der Vorlage (für die alignirenden Regionen). http://salilab.org/modeller/modeller.html

- Schleifen-Modellierung: Verwende eine Energiefunktion, um die Qualität der Schleife zu bewerten und minimiere diese Funktion durch Monte-Carlo (Sampling)

Vorhersage von Seitenkettenkonformationen

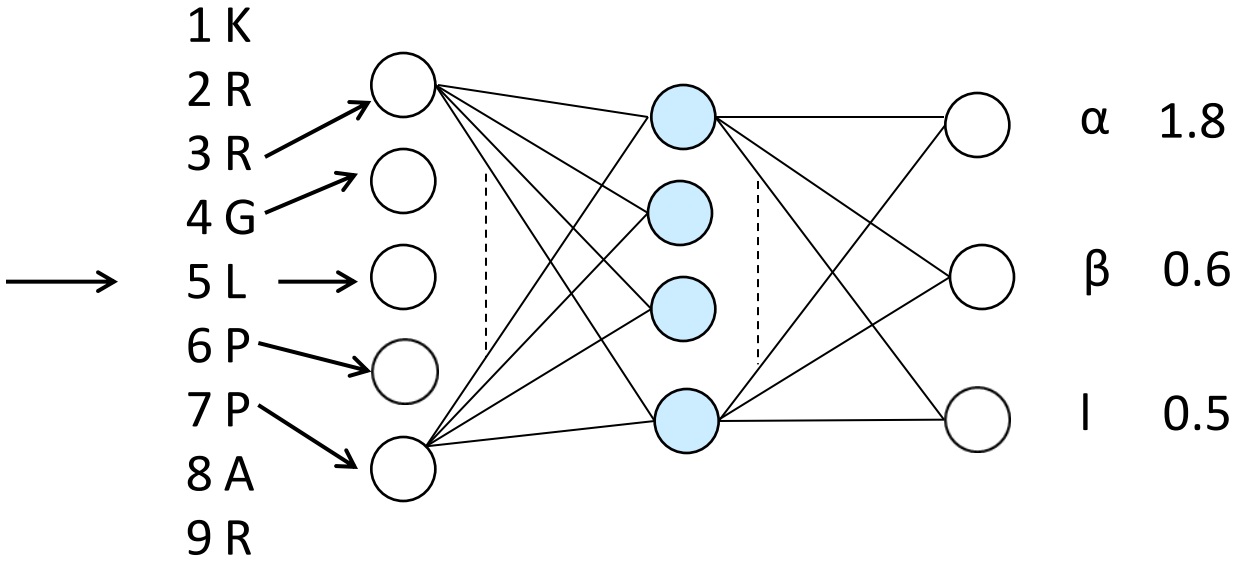
- Seitenketten Rotamere sind abhängig von den Rückgratkonformationen.
- Die erfolgreichste Methode war SCWRL von Dunbrack *et al.*:
  - Methode, auf Graph-Theorie basierend, um das kombinatorische Problem der Seitenketten Modellierung zu lösen.

Protein Sekundärstruktur Vorhersage – PHD Programm

Basierend auf ANNs (Artificial Neural Networks)

Als Input dient die Proteinsequenz; der Output ist die vorhergesagte Sekundärstruktur

Der wichtigste Faktor für die Sekundärstruktur eines Restes ist ihre lokale Sequenz



Threading

- Ziel: **Vergleiche die 1-dimensionale Sequenz mit dreidimensionalen Strukturen (1D-3D Verfahren**). Codiere die Struktur als String und dann vergleiche es, als ob es eine Proteinsequenz wäre
- Die Struktur wird als eine Reihe von unterschiedlichen strukturellen Zuständen kodiert, die mit einer Proteinsequenz mit dem Sequenzabgleichverfahren verglichen werden kann

Threading – Definition von Strukturzuständen

- Sekundärstruktur
  - α-Helix
  - β-Blätter
  - keines von beiden
- Für Lösungsmittel zugängliche Oberflächen
  - verbogen
  - ausgesetzt

Threading – The 3D-1D matching Methode

Abfragesequenz

STAYDILEYNTGA

ATGTSHFDEIYGA
A-GTHHIEDIYTA
STGHSYLEELR-A
ATSTSYFDEILGA

Multiples Sequenz-Alignment
/ Multiples Struktur -Alignment

6 Zustände für jeden Rest

- α-Helix & exposed
- α-Helix & buried
- β-sheets & exposed
- B-sheet & buried
- neither & exposed
- Neither & buried

$S_{aj} = \log(P_{aj}/P_a)$

P_{aj} = Wahrscheinlichkeit Aminosäure a in Umgebung j zu finden

P_a = Wahrscheinlich Aminosäure a irgendwo zu finden