V5: Proteinstruktur: Sekundärstruktur

INHALT

- Hierarchischer Aufbau der Proteinstruktur
- **Ramachandran-Plot**
- Vorhersage von Sekundärstrukturelementen aus der Sequenz
- Membranproteine
- Distanzmatrix, Strukturvergleich (DALI)

LERNZIELE

- lerne Prinzipien der Proteinstruktur kennen
- stelle Proteinstrukturen graphisch dar (Übung)

WOZU IST DAS GUT?

- Verständnis der dreidimensionalen Proteinstruktur macht erst deutlich, was die **Funktion** vieler Proteine ist.
- viele interessante **Strukturmotive** können bereits aus der Sequenz mit Bioinformatik-Methoden vorhergesagt werden

Funktion von Proteinen

Strukturproteine (Hüllenproteine von Viren, Cytoskelett)

Enzyme, die chemische Reaktionen katalysieren

Transportproteine und **Speicherproteine** (Hämoglobin)

Regulatoren wie Hormone und **Rezeptoren/Signalübertragungsproteine**

Proteine, die die Transkription kontrollieren
oder an Erkennungsvorgängen beteiligt sind:

Zelladhäsionsproteine, Antikörper

Warum sind Proteine so groß?

Proteine sind große Moleküle.

Ihre **Funktion** ist oft in einem kleinen Teil der Struktur,
dem **aktiven Zentrum**, lokalisiert.

Der Rest?
- Korrekte **Orientierung** der Aminosäuren des aktiven Zentrums
- **Bindungsstellen** für Interaktionspartner
- Konformationelle **Dynamik**

Evolution der Proteine: Veränderungen der Struktur, die durch Mutationen in ihrer Aminosäuresequenz hervorgerufen werden.

Hierarchischer Aufbau

Primärstruktur – Sekundärstruktur – Tertiärstruktur – Quartärnere Struktur – Komplexe

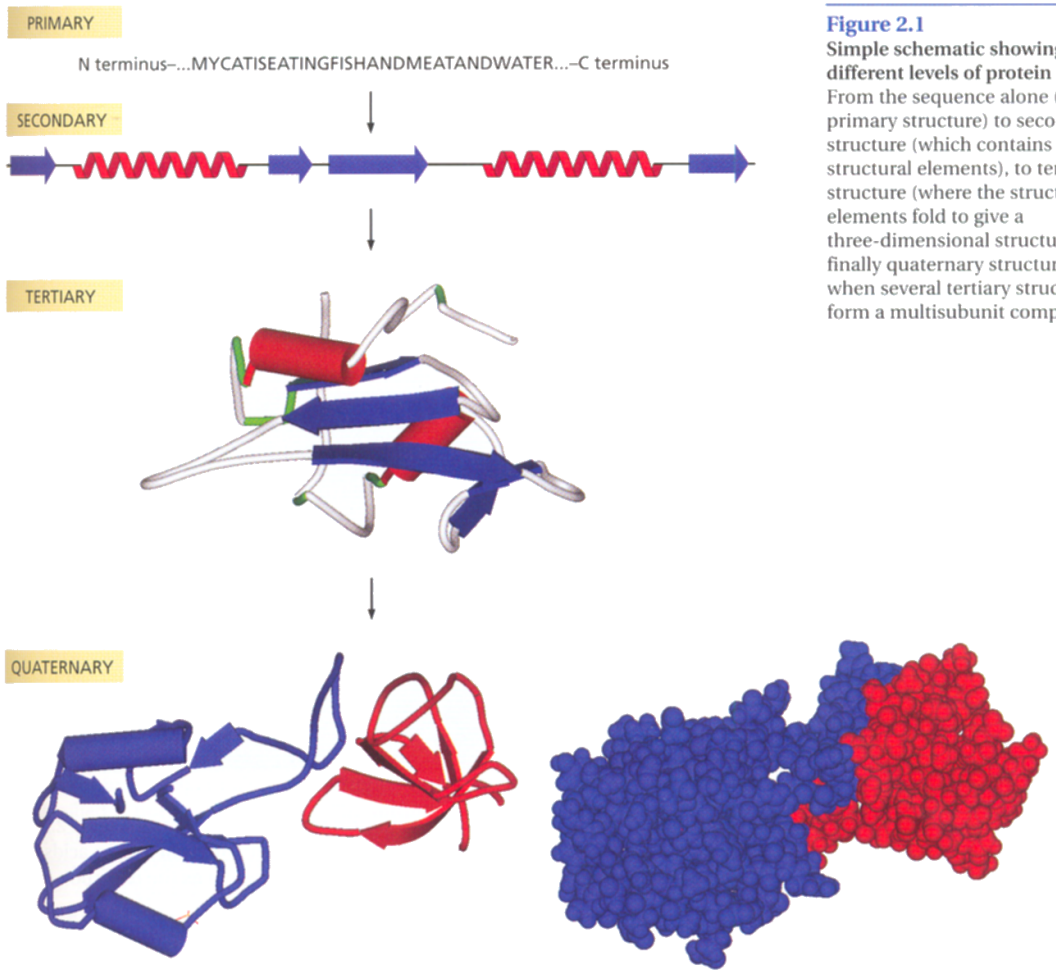


Figure 2.1
Simple schematic showing the different levels of protein structure. From the sequence alone (the primary structure) to secondary structure (which contains local structural elements), to tertiary structure (where the structural elements fold to give a three-dimensional structure), to finally quaternary structure found when several tertiary structures form a multisubunit complex.

Hierarchischer Aufbau

Welche „Kräfte“ sind für die Ausbildung der verschiedenen „Strukturen“ wichtig?

Lösliche Proteine: wichtigstes Prinzip ist der **hydrophobe Effekt**.

Der Beitrag hydrophober WW zur Freien Enthalpie bei der Proteinfaltung und der Protein-Liganden-Wechselwirkung kann als proportional zur Grösse der während dieser Prozesse vergrabenen hydrophoben Oberfläche angesehen werden.

Membranproteine: sind im **Transmembranbereich** außen hydrophober als innen.

Die wasserlöslichen Bereiche von Membranproteinen ähneln in ihrer Zusammensetzung den löslichen Proteinen.

Hydrophober Effekt

Beobachtung, dass die Überführung einer unpolaren Substanz/Oberflächenbereichs aus einem organischen bzw. Unpolaren Lösungsmittel nach Wasser

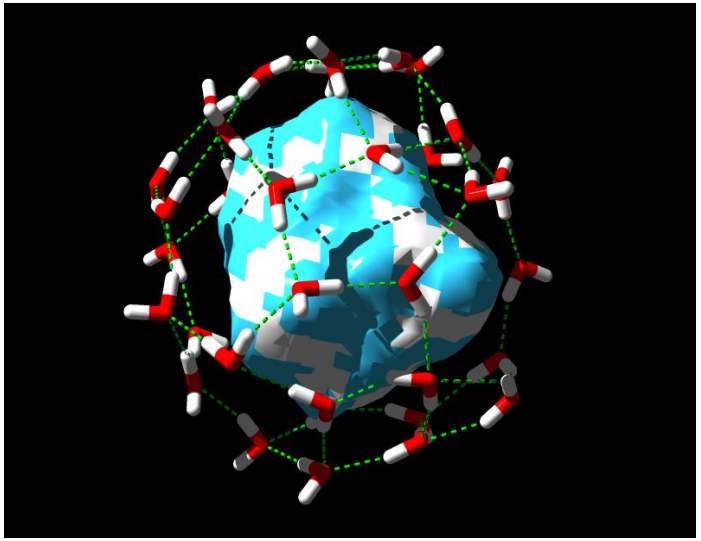
(a) energetisch stark ungünstig ist

(b) bei Raumtemperatur zu einer Abnahme der Entropie führt

(c) zu einer Zunahme der Wärmekapazität führt.

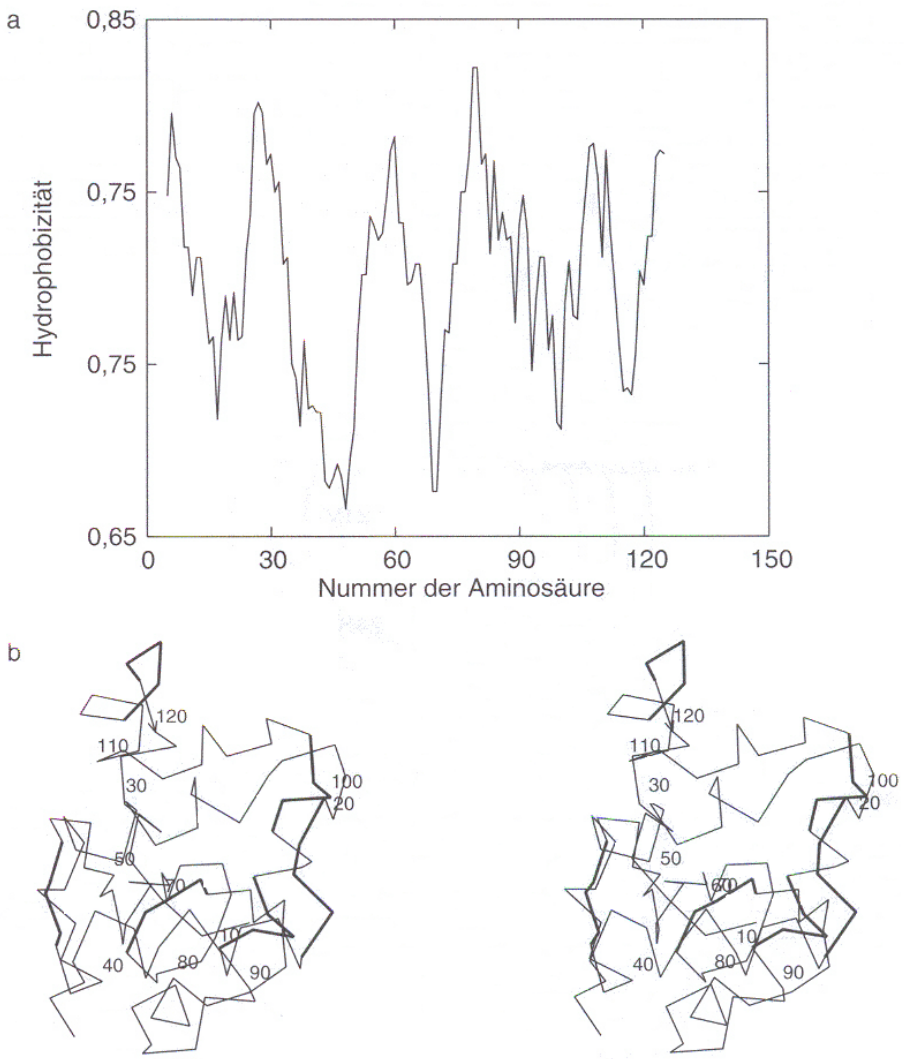
Eisberg-Modell

W. Kauzman 1959



Wassermoleküle an einer hydrophoben Oberfläche sind in ihren möglichen Orientierungen stark eingeschränkt -> dies ist entropisch ungünstig.

Anwendungen der Hydrophobizität



Lesk-Buch

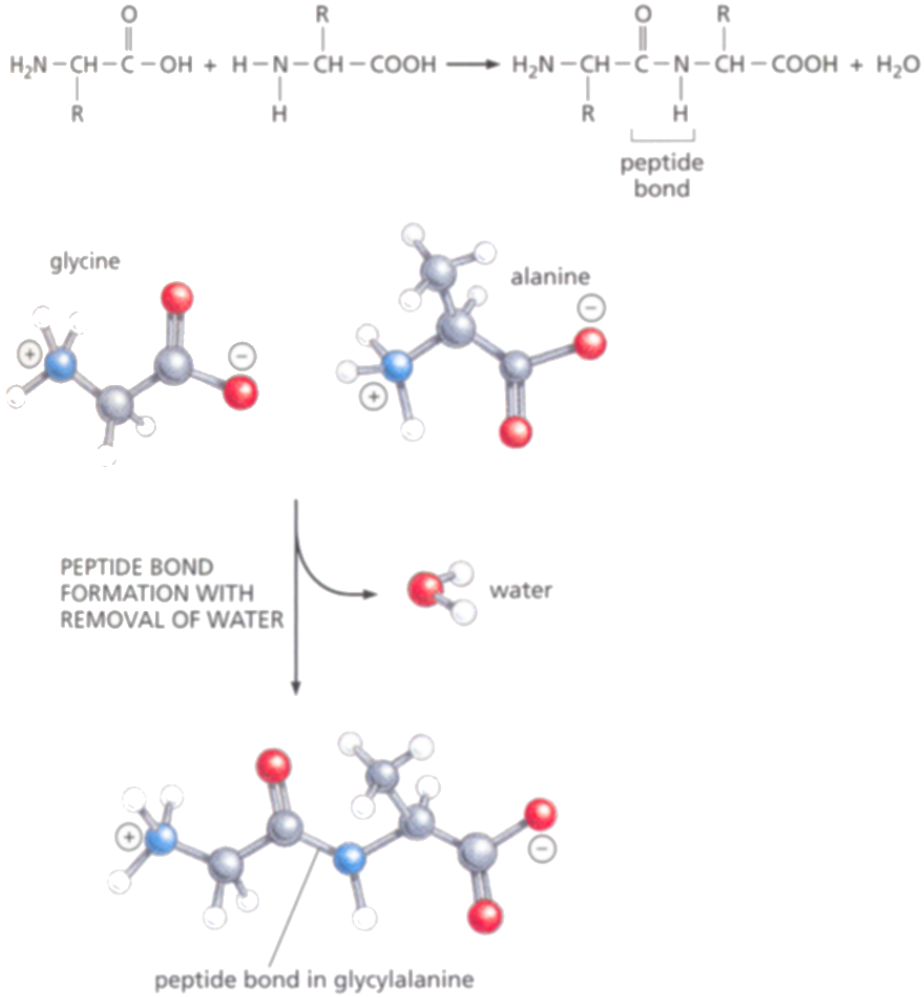
5.4 a) Hydrophobizitätsprofil des Lysozyms aus Hühnereiweiß (erzeugt mithilfe der „Primary Structure Analysis"-Werkzeuge unter `http://www.expasy.ch`). b) Struktur des gleichen Enzyms. Abschnitte, die den Minima im Hydrophobizitätsprofil entsprechen, sind durch etwas dickere Linien gekennzeichnet.

Peptidbindung

In Peptiden und Proteinen sind die Aminosäuren miteinander als lange Ketten verknüpft.
Ein Paar ist jeweils über eine „**Peptidbindung**“ verknüpft.

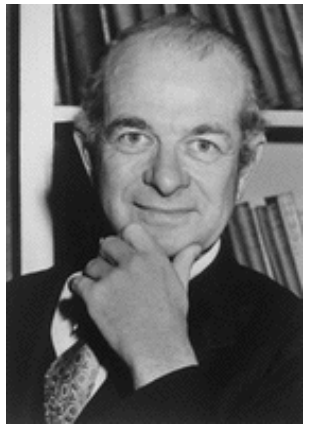
Die Aminosäuresequenz eines Proteins bestimmt seinen „**genetischen code**“.

Die Kenntnis der Sequenz eines Proteins allein verrät noch nicht viel über seine Funktion.
Entscheidend ist seine **drei-dimensionale Struktur.**



Eigenschaften der Peptidbindung

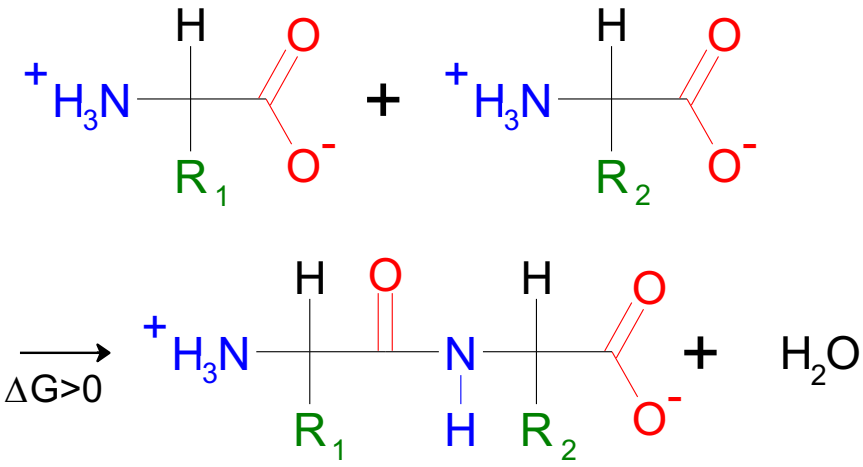
E.J. Corey und **Linus Pauling** studierten die Petidbindung in den 1940‘ern und 1950‘ern.



Linus Pauling
Nobelpreise für
Chemie 1954 und
Frieden 1963

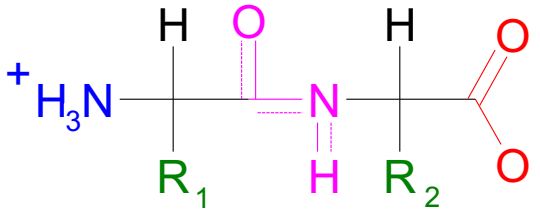
Sie fanden: die C-N Länge ist 1.33 Å.
Sie liegt damit zwischen 1.52 Å und 1.25 Å,
was die Werte für eine Einfach- bzw.
Doppelbindung sind.

Die benachbarte C=O Bindung hat eine Länge
Von 1.24 Å, was etwas länger als eine typische
Carbonyl- C=O Doppelbindung ist (1.215 Å).



→ die Peptidbindung hat einen teilweise konjugierten Charakter und ist nicht frei drehbar.

Es bleiben damit pro Residue 2 frei drehbare Diederwinkel des Proteinrückgrats übrig.

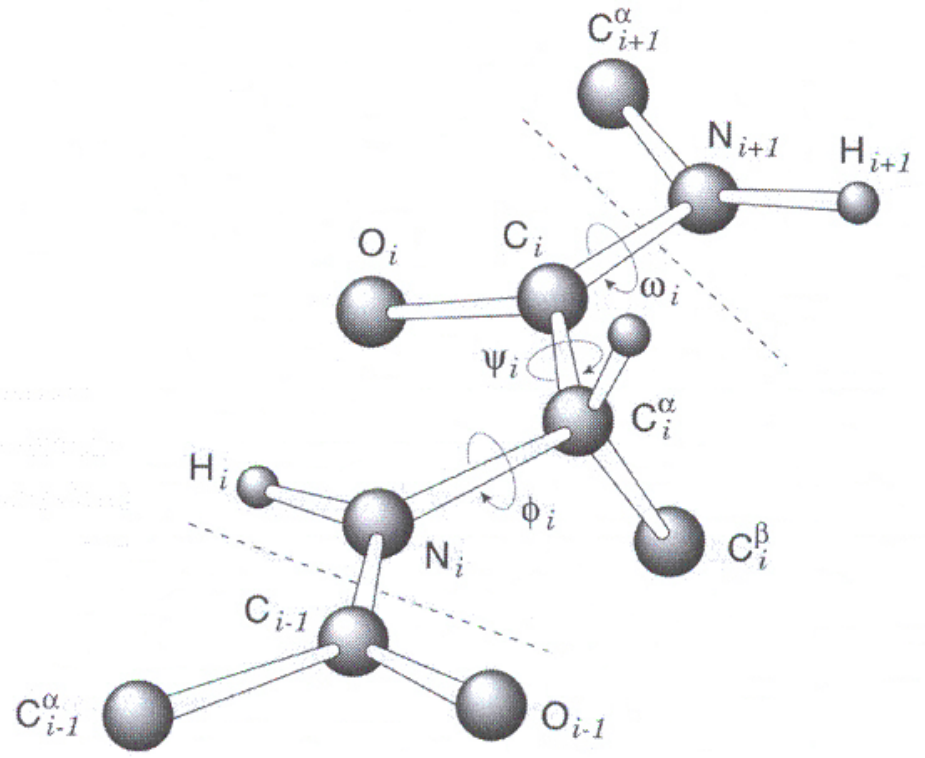


peptide bond

Diederwinkel des Proteinrückgrats

Die dreidimensionale Faltung des Proteins wird vor allem durch die **Diederwinkel** bzw. Dihedralwinkel des Proteinrückgrats bestimmt.

Pro Residue gibt es 2 frei drehbare Diederwinkel, die als Φ und Ψ bezeichnet werden.



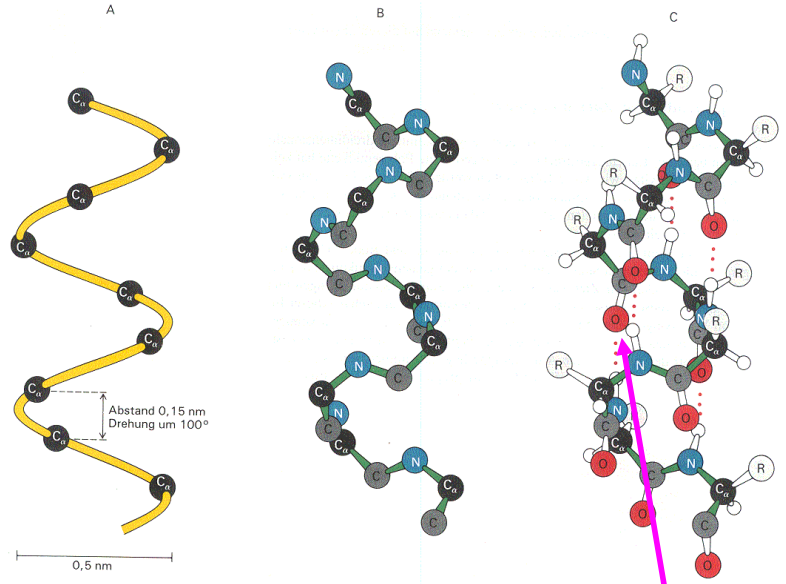
Definition der Konformationswinkel im Polypeptidrückgrat.

Lesk-Buch

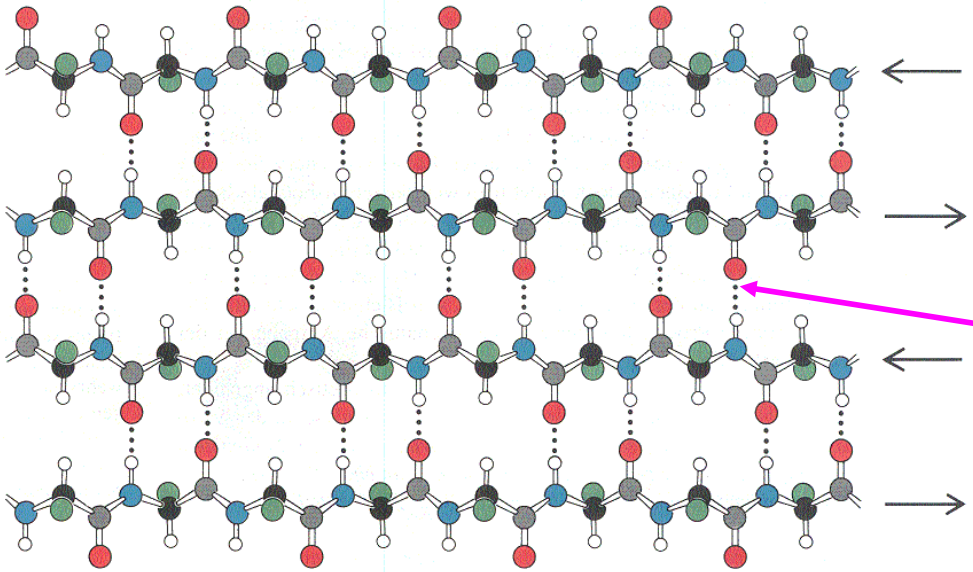
Sekundärstrukturelemente

Wie seit den 1950‘er Jahren bekannt, können Aminosäure-Stränge **Sekundärstrukturelemente** bilden:

(aus Stryer, Biochemistry)



α-Helices



und β-Stränge.

In diesen Konformationen bilden sich jeweils **Wasserstoffbrückenbindungen** zwischen den C=O und N-H Atomen des Rückgrats. Daher sind diese Einheiten strukturell stabil.

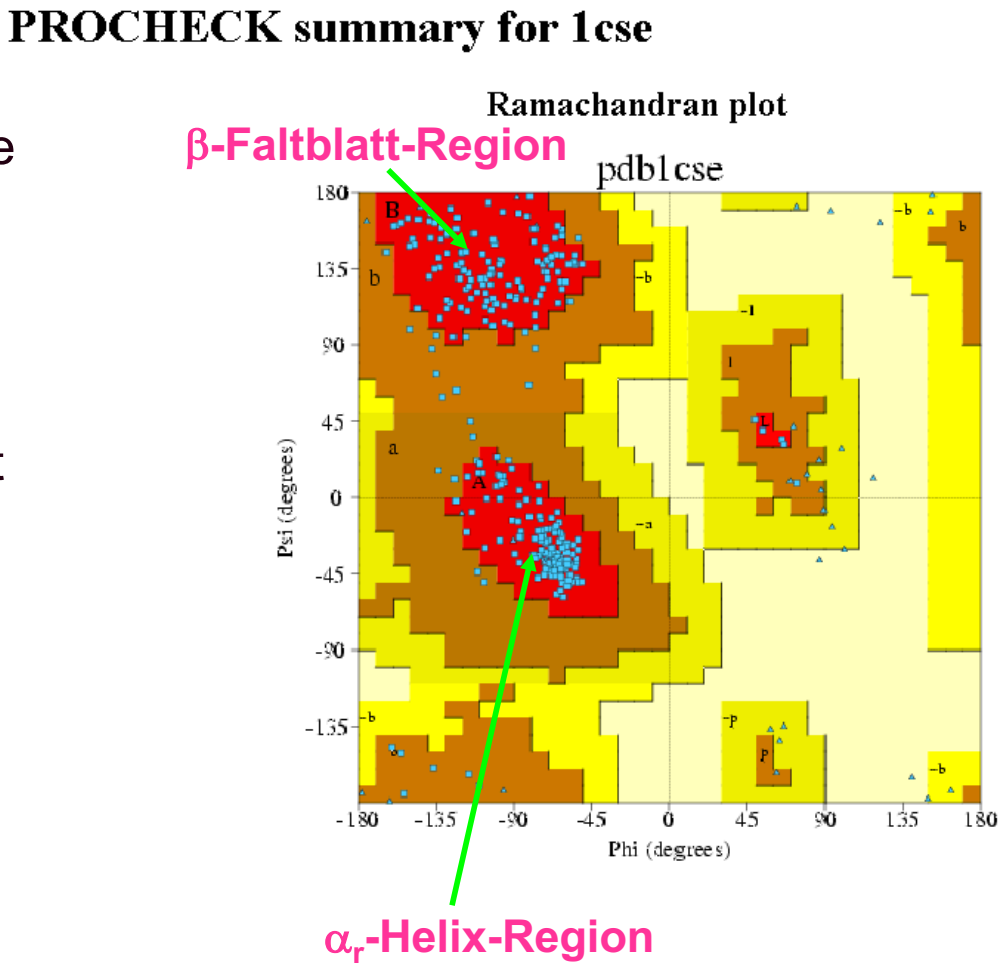
Stabilität und Faltung von Proteinen

Die gefaltete Struktur eines Proteins ist die Konformation, die die günstigste freie Enthalpie ΔG für diese Aminosäuresequenz besitzt.

Der **Ramachandran-Plot** charakterisiert die energetisch günstigen Bereiche des Aminosäurerückgrats.

Die einzige Residue, die außerhalb der erlaubten Bereich liegt, also alle möglichen Torsionswinkel annehmen kann, ist **Glycin**.

Grund: es hat keine Seitenkette.



(rechtsgängige Helix)

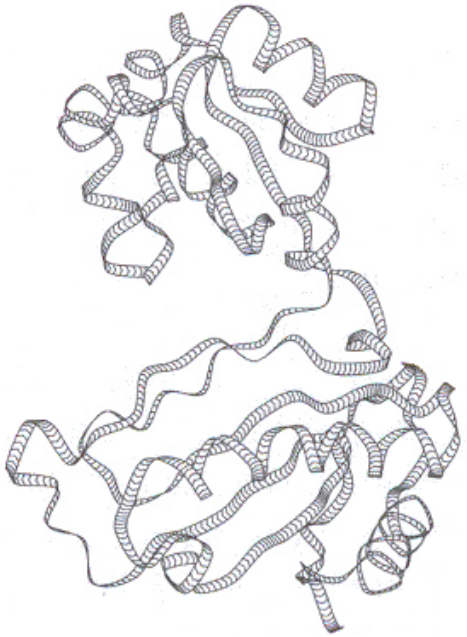
Domänen

Kompakter Bereich im Faltungsmuster einer Molekülkette, der den Anschein hat,

“er könnte auch unabhängig von den anderen stabil sein”.

cAMP-abhängige Proteinkinase

SERCA Calcium-Pumpe



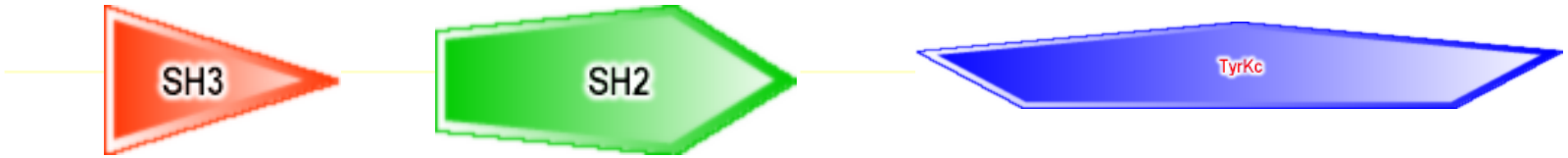
Lesk-Buch

Modular aufgebaute Proteine

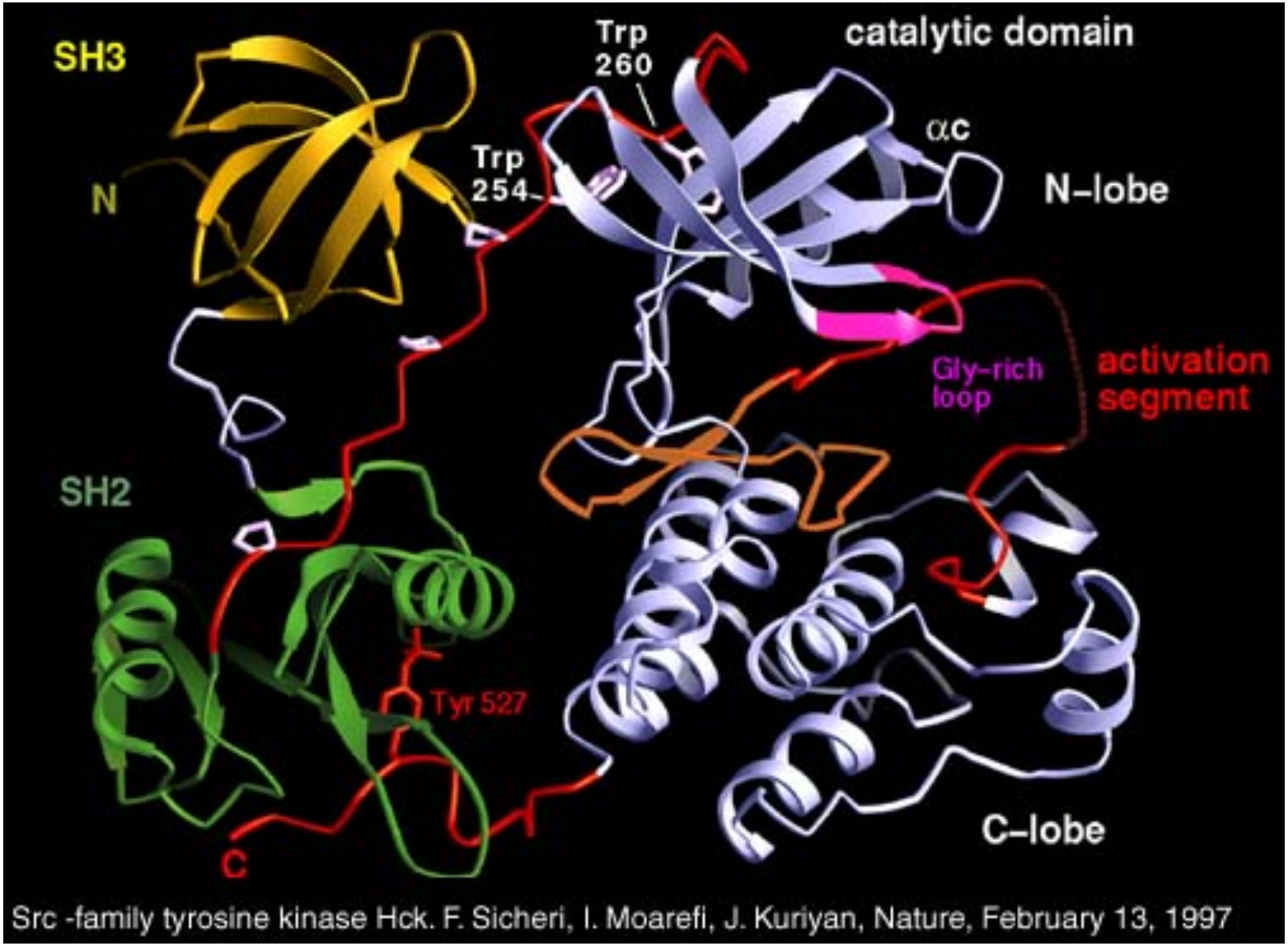
Modular aufgebaute Proteine bestehen aus mehreren Domänen.

Anwendung von SMART (www.smart.embl-heidelberg.de) für die Src-Kinase HcK ergibt

```
Sequenz:
MGGRSSCEDP GCPRDEERAP RMGCMKSKFL QVGGNTFSKT ETSASPHCPV
YVPDPTSTIK PGPNSHNSNT PGIREAGSED IIVVALYDYE AIHHEDLSFQ
KGDQMVVLEE SGEWWKARSL ATRKEGYIPS NYVARVDSLE TEEWFFKGIS
RKDAERQLLA PGNMLGSFMI RDSETTKGSY SLSVRDYDPR QGDTVKHYKI
RTLDNGGFYI SPRSTFSTLQ ELVDHYKKGN DGLCQKLSVP CMSSKPQKPW
EKDAWEIPRE SLKLEKKLGA GQFGEVWMAT YNKHTKVAVK TMKPGSMSVE
AFLAEANVMK TLQHDKLVKL HAVVTKEPIY IITEFMAKGS LLDFLKSDEG
SKQPLPKLID FSAQIAEGMA FIEQRNYIHR DLRAANILVS ASLVCKIADF
GLARVIEDNE YTAREGAKFP IKWTAPEAIN FGSFTIKSDV WSFGILLMEI
VTYGRIPYPG MSNPEVIRAL ERGYRMPRPE NCPEELYNIM MRCWKNRPEE
RPTFEYIQSV LDDFYTATES QYQQQP
```



Beispiel: Src-Kinase HcK



http://jkweb.berkeley.edu/

Klassifikation von Proteinen

Die Klassifikation von Proteinstrukturen nimmt in der Bioinformatik eine Schlüsselposition ein, weil sie das Bindeglied zwischen Sequenz und Funktion darstellt.

Scop Classification Statistics

SCOP: Structural Classification of Proteins. **1.69** release
25973 PDB Entries (1 Oct 2004). 70859 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	218	376	608
All beta proteins	144	290	560
Alpha and beta proteins (a/b)	136	222	629
Alpha and beta proteins (a+b)	279	409	717
Multi-domain proteins	46	46	61
Membrane and cell surface proteins	47	88	99
Small proteins	75	108	171
Total	945	1539	2845

Die allgemeinste Einteilung in Familien von Proteinstrukturen stützt sich auf die Sekundär- und Tertiärstrukturen:

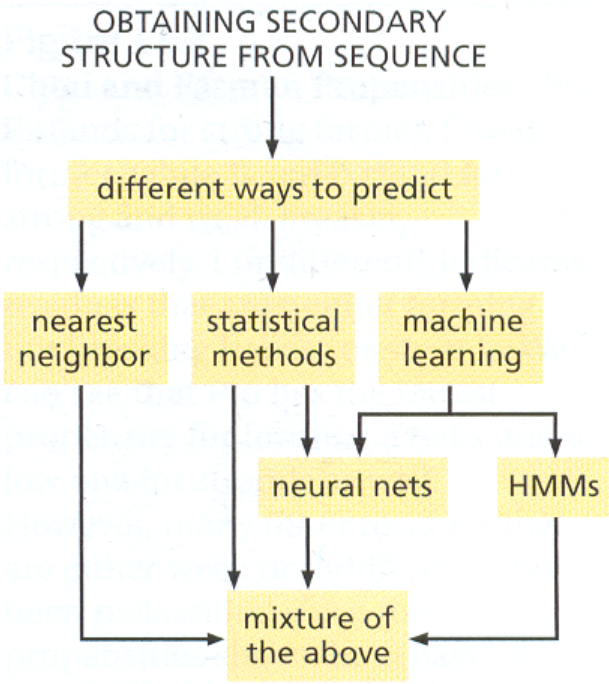
Klasse	Merkmale
α-helikal	Sekundärstruktur ausschließlich oder fast ausschließlich α-Helix
β-Faltblatt	Sekundärstruktur ausschließlich oder fast ausschließlich β-Faltblatt
$\alpha + \beta$	α-Helix und β-Faltblatt getrennt in verschiedenen Molekülteilen; kein β-α-β als Supersekundärstruktur
α/β	Helices und Faltblätter aus β-α-β-Einheiten zusammengesetzt
– α/β-linear	Mittellinie von Strängen der Faltblätter ungefähr linear
– α/β-Tonnen (*barrels*)	Mittellinie von Strängen der Faltblätter ungefähr kreisförmig
wenig oder gar keine Sekundärstruktur	

Lesk-Buch

Sekundärstruktur-Vorhersage

- Sekundärstrukturvorhersage für lösliche Proteine

- Sekundärstrukturvorhersage für Membranproteine



Flow Diagram 11.1
The key concept introduced in this section is that many different approaches have been taken in deriving methods for predicting protein secondary structure.

Literatur:
Kapitel 11 und 12 in
Understanding Bioinformatics
Zvelebil & Baum

am häufigsten auftretende Sekundärstrukturen

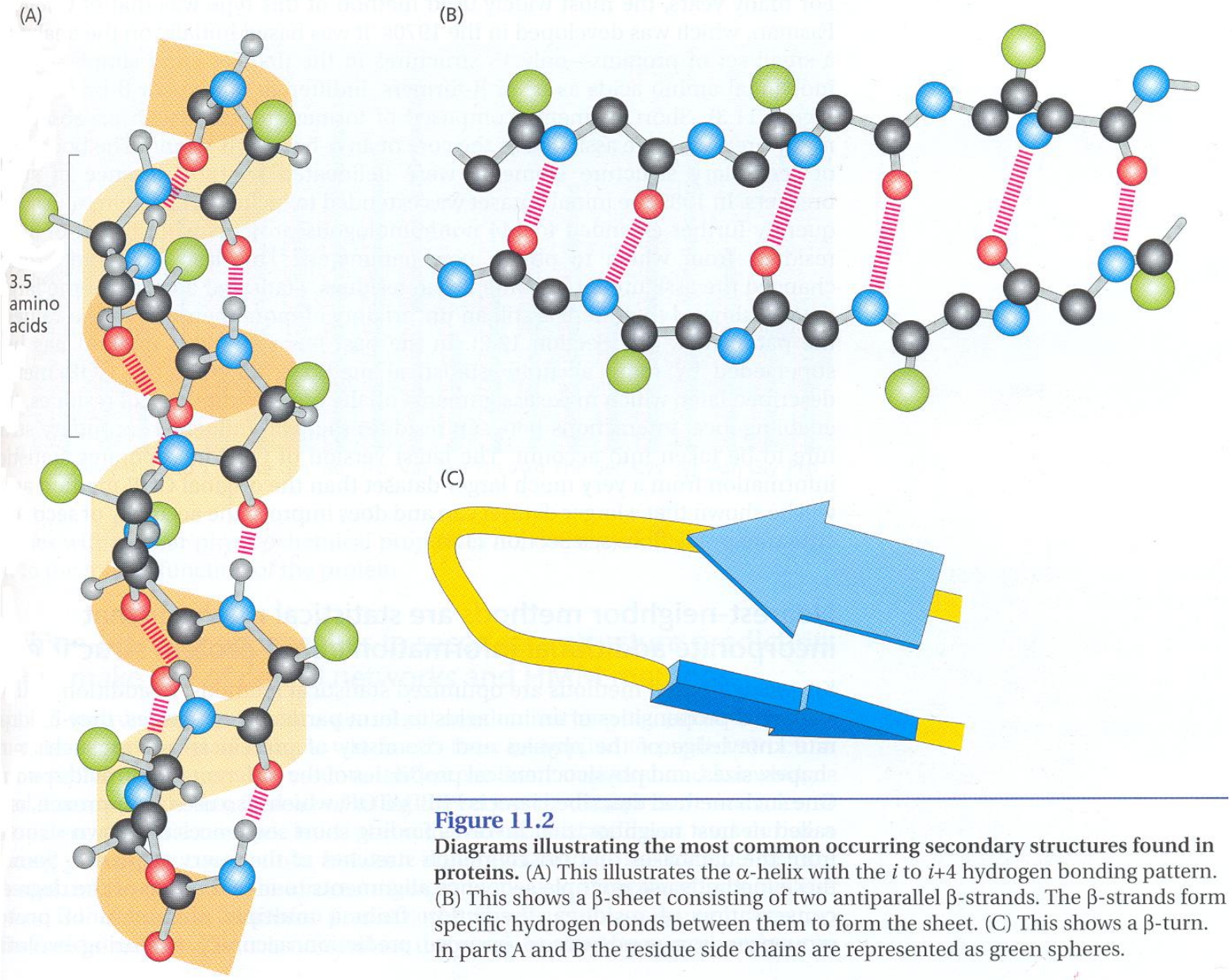
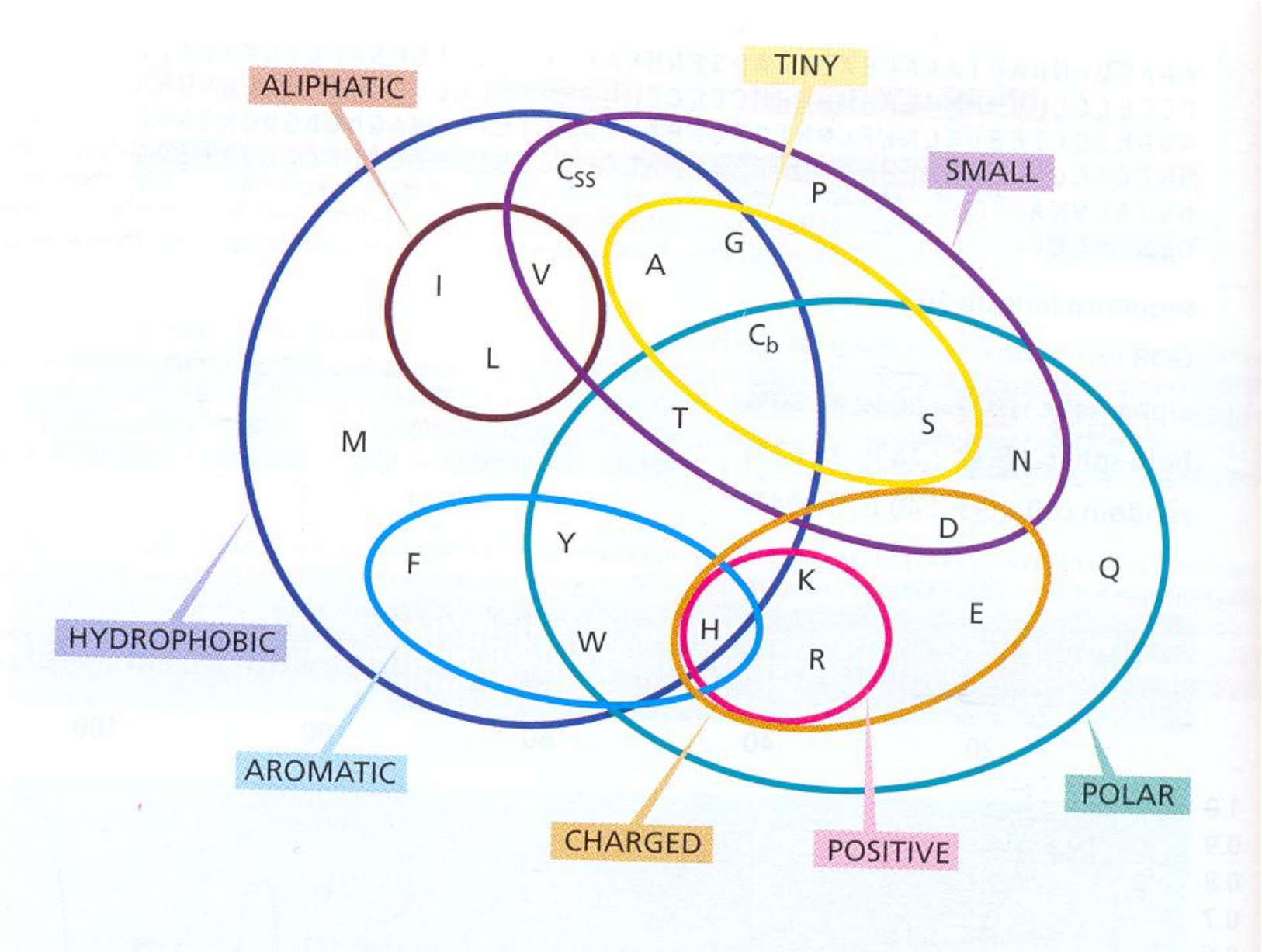
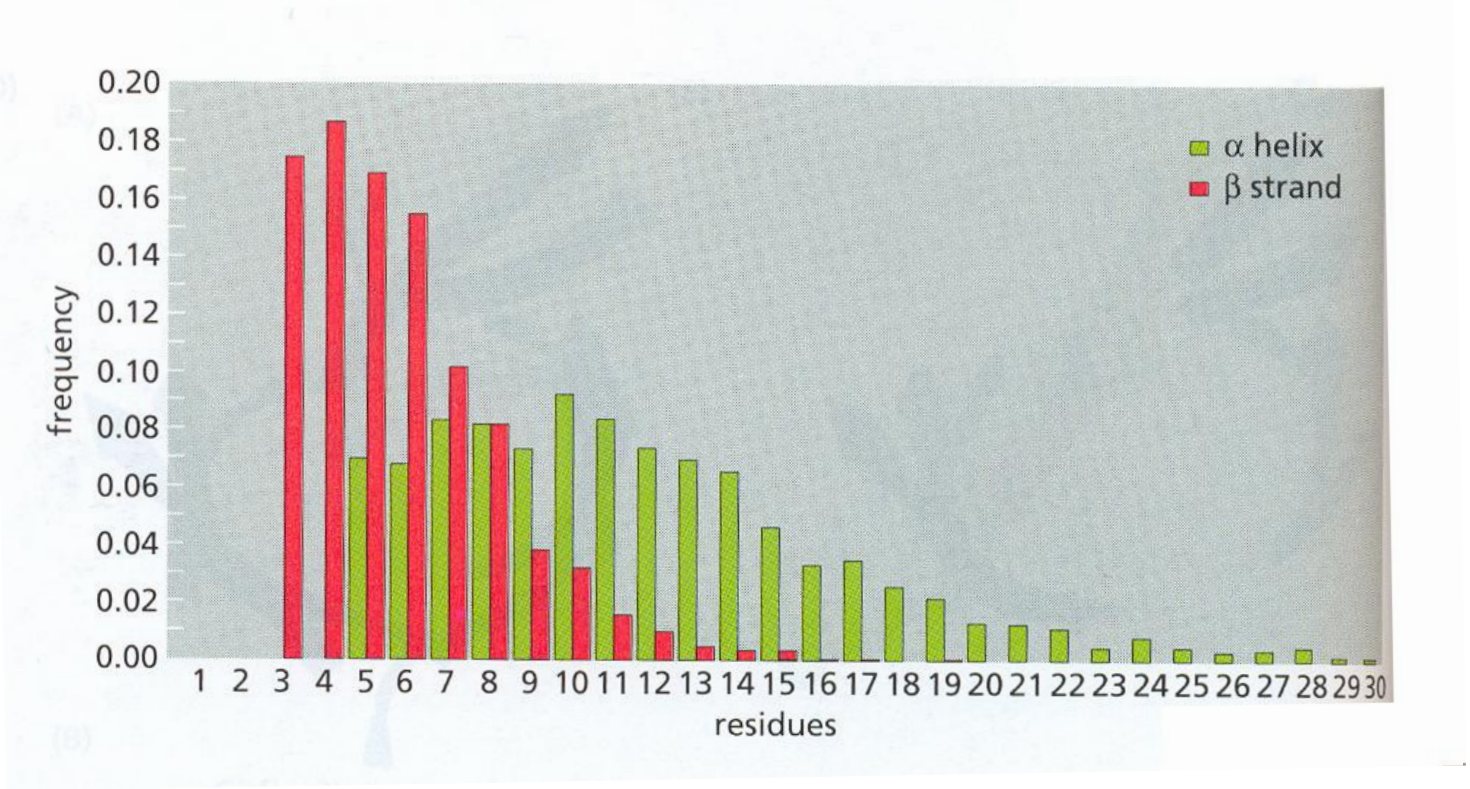


Figure 11.2
Diagrams illustrating the most common occurring secondary structures found in proteins. (A) This illustrates the α-helix with the *i* to *i*+4 hydrogen bonding pattern. (B) This shows a β-sheet consisting of two antiparallel β-strands. The β-strands form specific hydrogen bonds between them to form the sheet. (C) This shows a β-turn. In parts A and B the residue side chains are represented as green spheres.

Die 20 natürlichen Aminosäuren



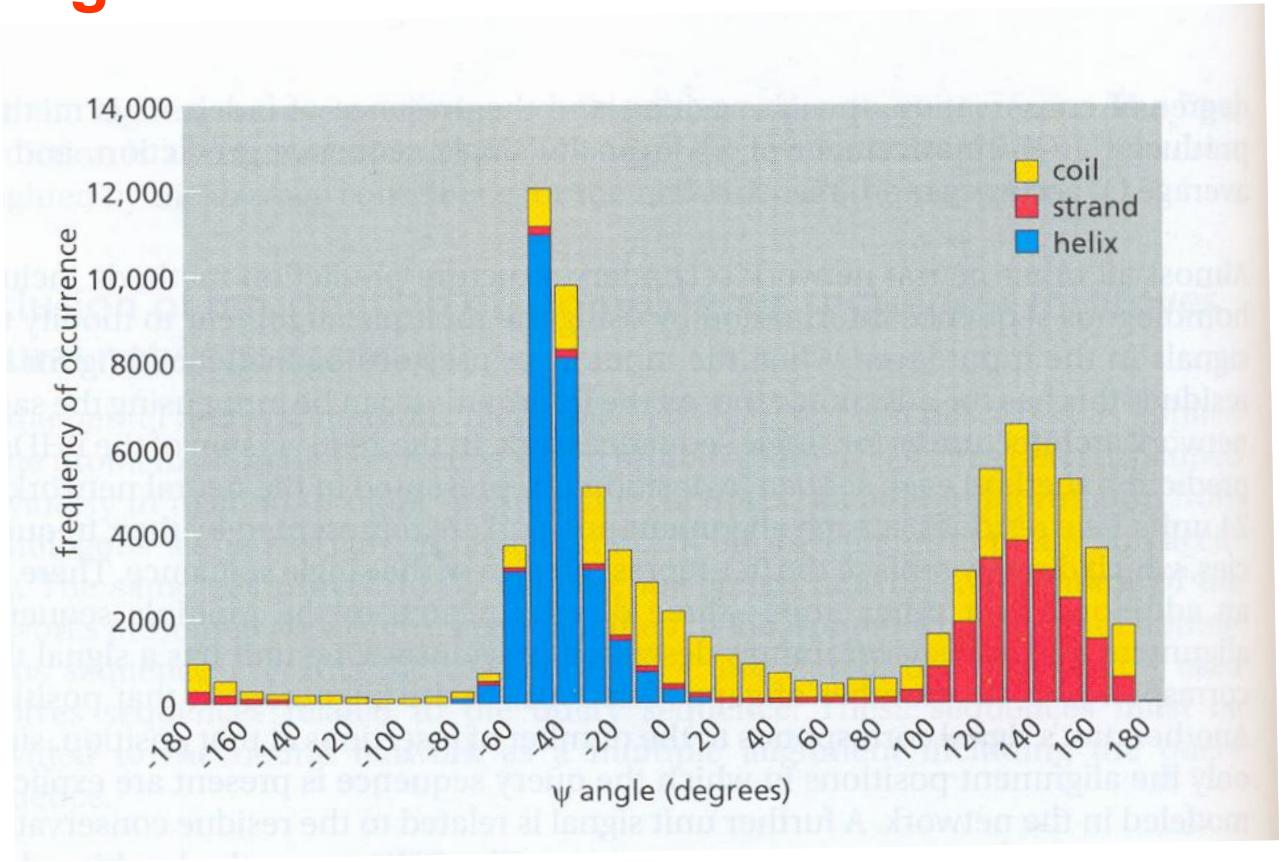
Sekundärstruktur-Auftreten in löslichen Proteinen



Längenverteilung von Sekundärstrukturelementen.

Statistische Daten für eine große Menge an Proteinen mit bekannter Struktur.

Rückgratwinkel in Sekundärstrukturelementen



Chou & Fasman Propensities

Amino Acid	helix			
	Designation	*P*	Designation	*P*
Ala	F	1.42	b	0.83
Cys	I	0.70	f	1.19
Asp	I	1.01	B	0.54
Glu	F	1.51	B	0.37
Phe	f	1.13	f	1.38
Gly	B	0.61	b	0.75
His	f	1.00	f	0.87
Ile	f	1.08	F	1.60
Lys	f	1.16	b	0.74
Leu	F	1.21	f	1.30
Met	F	1.45	f	1.05
Asn	b	0.67	b	0.89
Pro	**B**	**0.57**	**B**	**0.55**
Gln	*f*	*1.11*	*h*	*1.10*
Arg	I	0.98	I	0.93
Ser	I	0.77	b	0.75
Thr	I	0.83	f	1.19
Val	f	1.06	F	1.70
Trp	f	1.08	f	1.37
Tyr	b	0.69	F	1.4

F : starke Tendenz

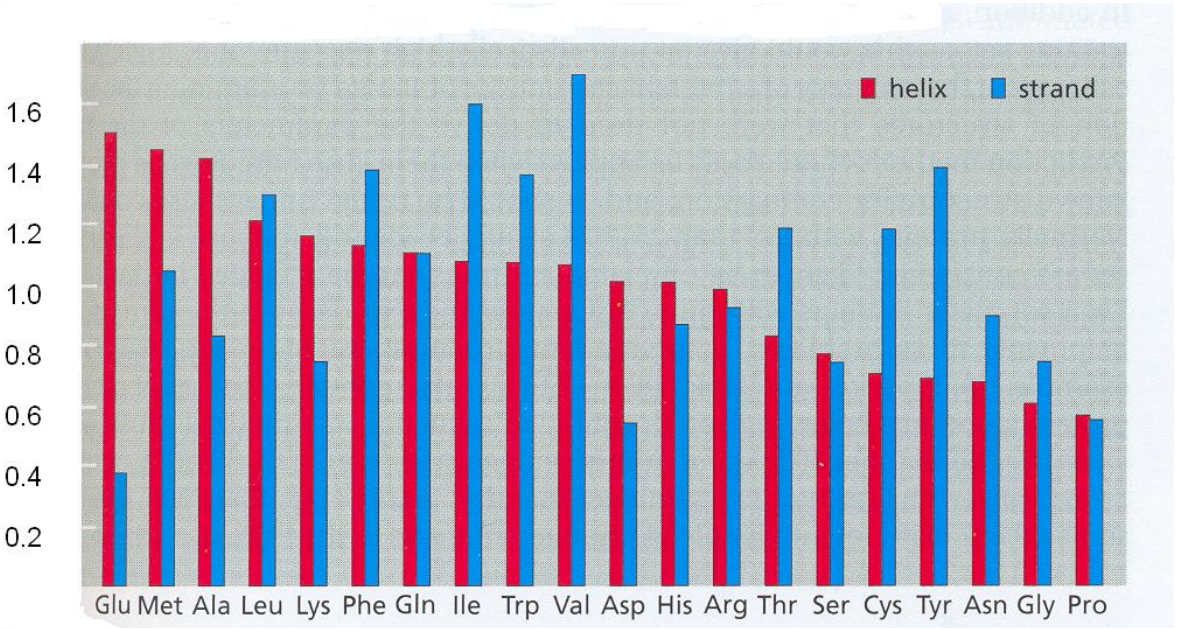
f : schwache Tendenz

B : starker (Unter-) Brecher

b : schwacher (Unter-) Brecher

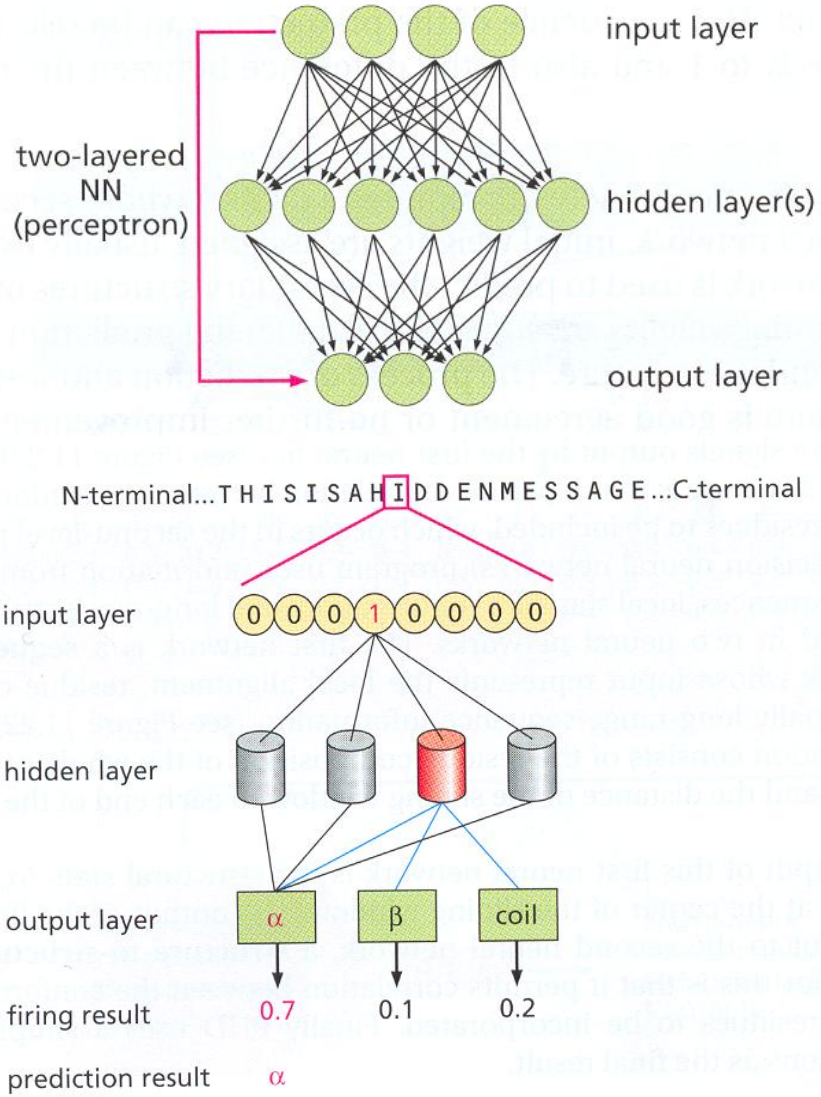
I : indifferent

Prolin: stärkster Helixbrecher sowie für Betastränge



Vorhersage mit Neuronalen Netzwerken

zweilagiges Neuronales Netzwerk

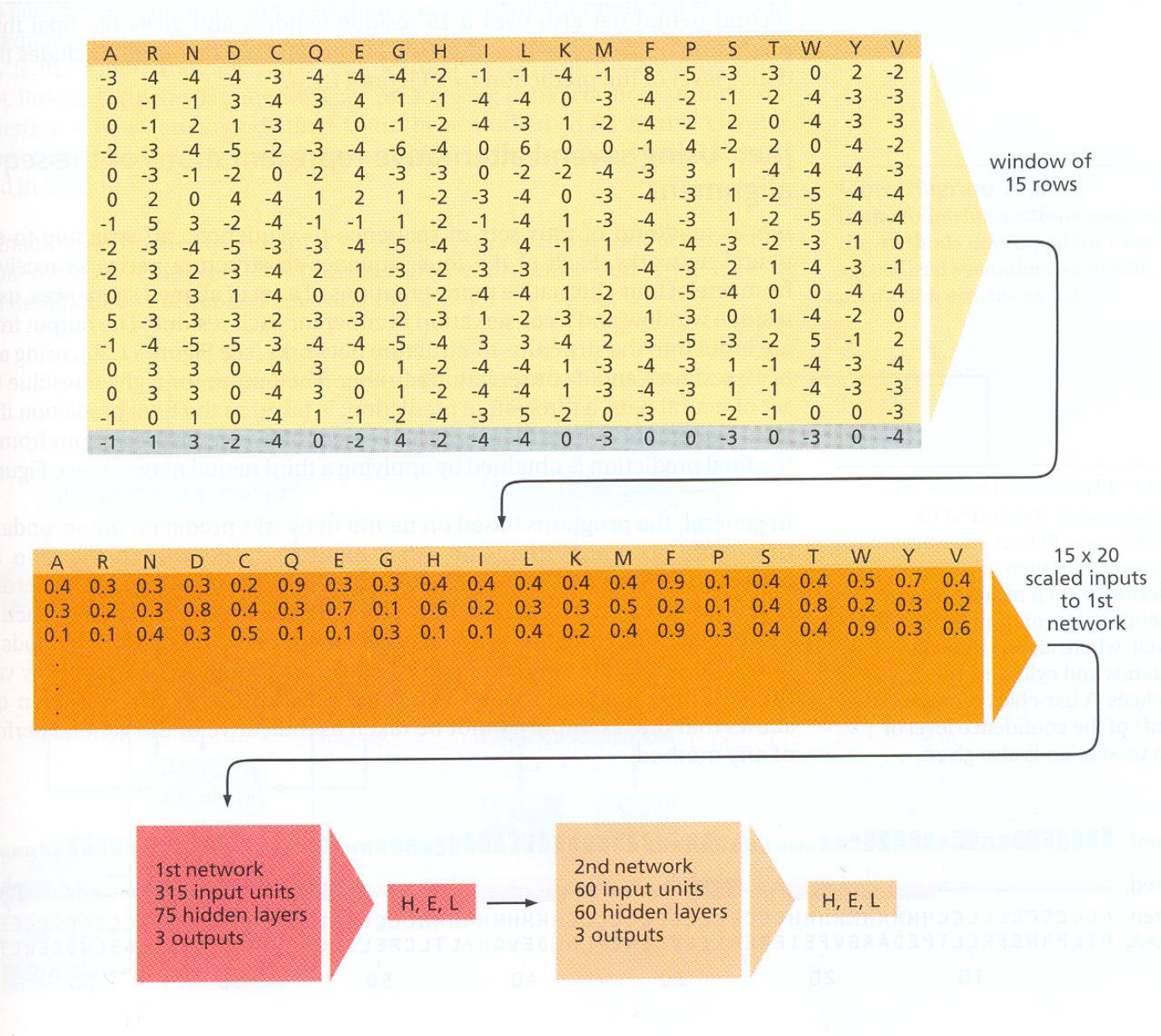


Feed-forward NN zur Vorhersage von Sekundärstrukturen

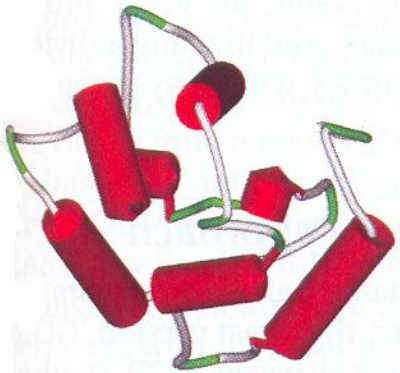
PSIPRED

Benutze Profil aus PSIBLAST.

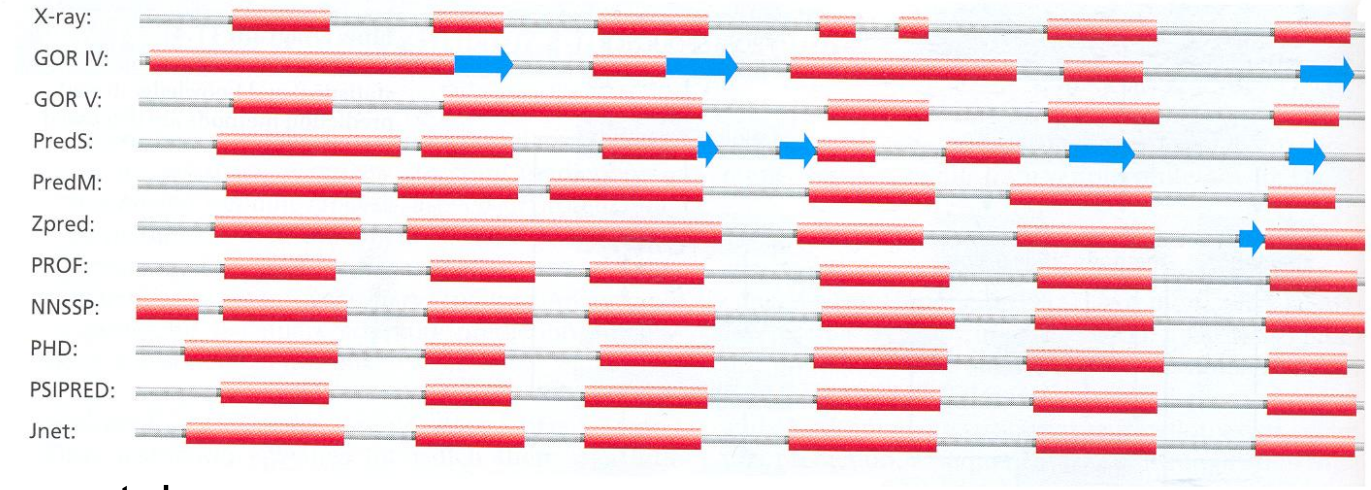
Skaliere Werte auf Intervall [0.0;1.0].



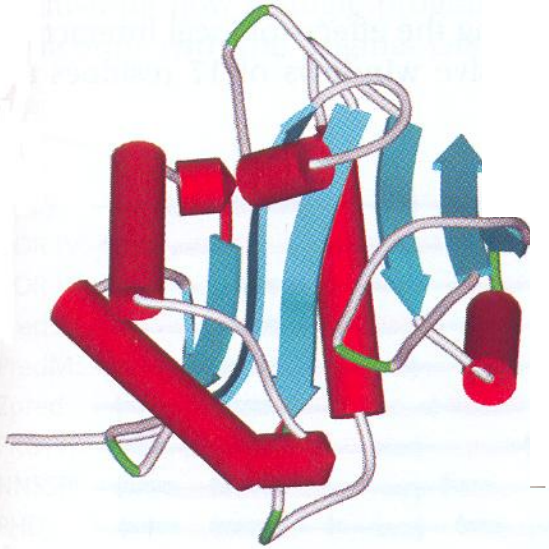
Qualität der Sekundärstruktur-Vorhersagen



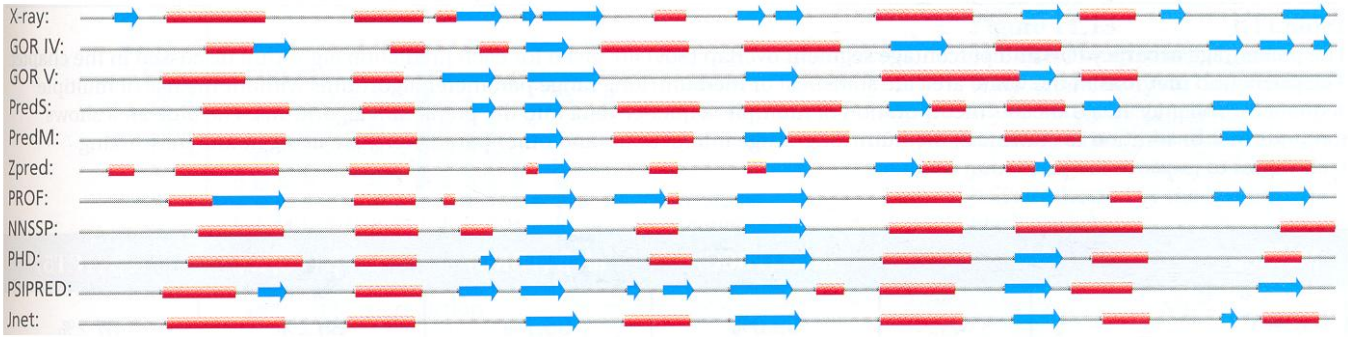
1B8C



all α protein



1CJW

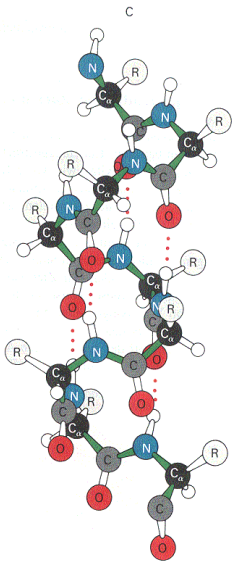


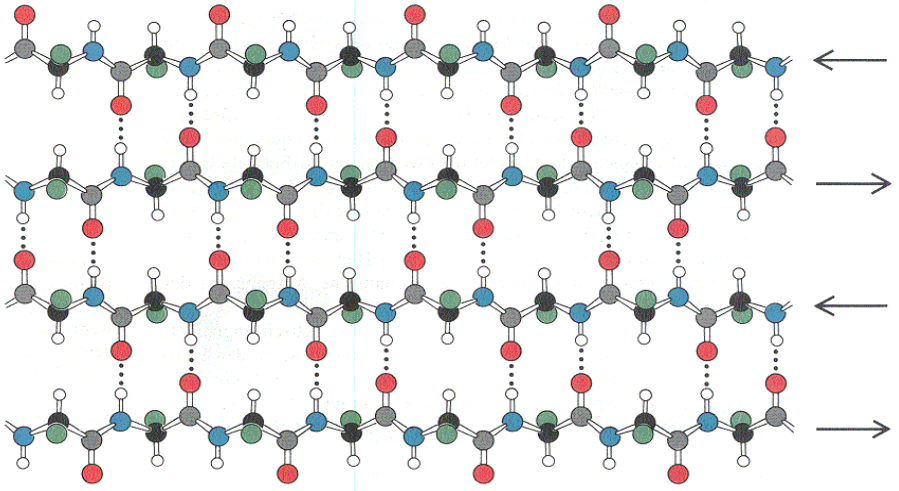
$\alpha + \beta$ protein

etwa 75% Genauigkeit

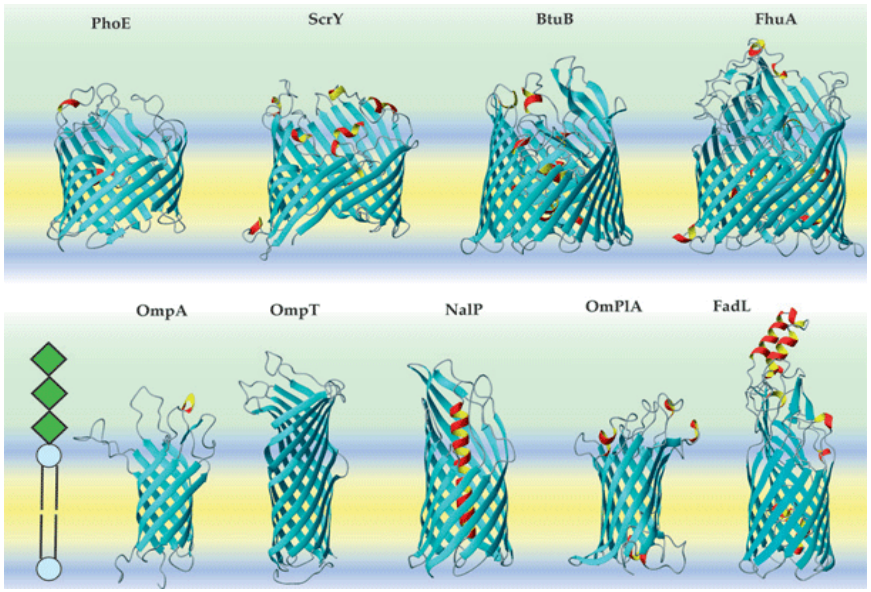
Topologie von Membranproteinen

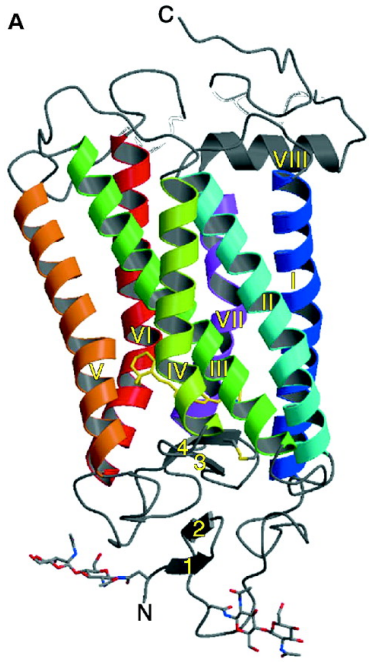
Im Inneren der Lipidschicht kann das Proteinrückgrat keine Wasserstoffbrücken-Bindungen mit den Lipiden ausbilden →
die Atome des Rückgrats müssen miteinander Wasserstoffbrückenbindungen ausbilden,
sie müssen entweder helikale oder β-Faltblattkonformation annehmen.





Topologie von Membranproteinen





Die hydrophobe Umgebung erzwingt, dass (zumindest die bisher bekannten) Strukturen von Transmembranproteinen entweder reine β-Barrels (links) oder reine α-helikale Bündel (rechts) sind.

http://www.biologie.uni-konstanz.de/folding/Structure%20gallery%201.html

Vorhersage von Transmembranhelices

Einfaches Kriterium: Hydrophobizitäts-Skalen

TMHs können aus der Abfolge von hydrophoben und polaren Regionen in der Sequenz vorhergesagt werden (siehe helikales Rad).

Man beobacht folgende immer wiederkehrende Motive:
1. TMHs sind meistens apolar und 12-35 Residuen lang,

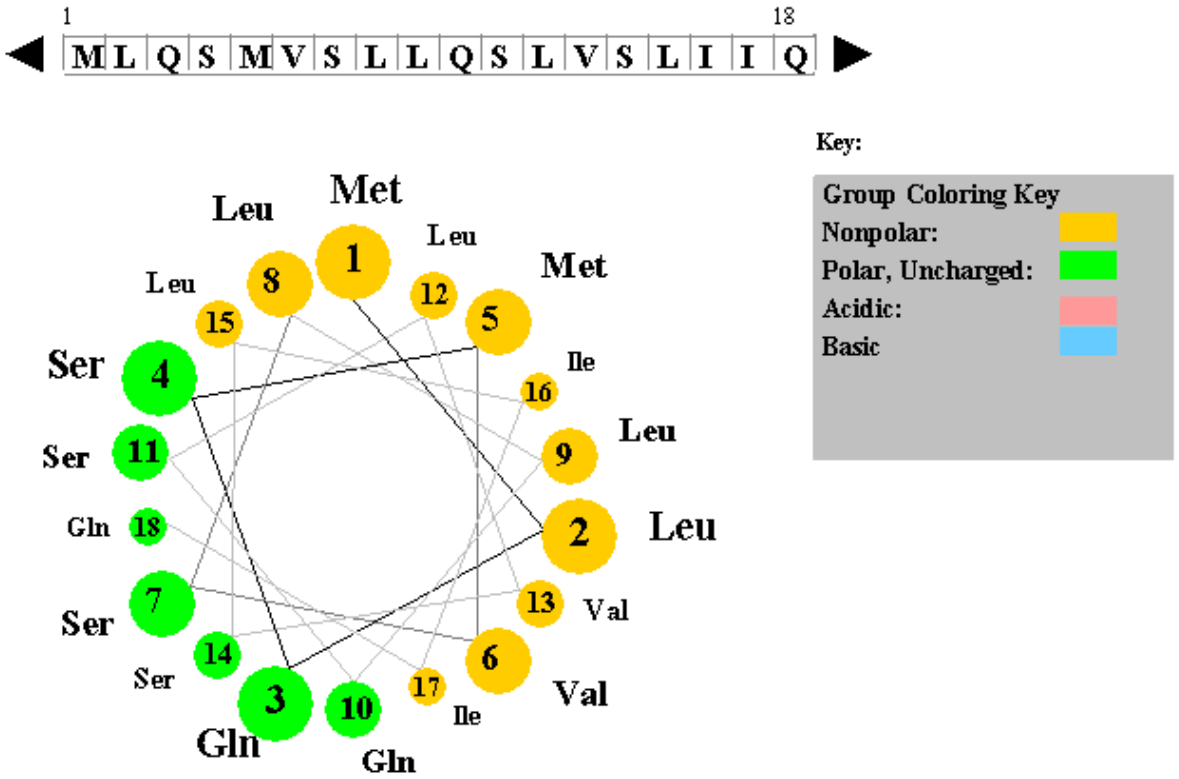
2. Globuläre (d.h. kompakte oder kugelförmige) Regionen zwischen den TMHs sind kürzer als 60 Residuen,

3. die meisten TMH Proteine haben eine spezifische Verteilung der positiv geladenen Aminosäuren Arginin und Lysin, ”**positive-inside-rule**“ (Gunnar von Heijne), die ”loop“ Regionen innen haben mehr positiv geladene Aminosäuren als außen.

Gunnar von Heijne

4. Lange globuläre Regionen (> 60 Residuen) unterscheiden sich in ihrer Anordnung von den Regionen, die der ”Innen-Außen-Regel“ unterliegen.

Helikale Räder



Helikale Räder dienen zur Darstellung von Helices.

Man kann so leicht erkennen, welche Seite der Helix dem Solvens zugewandt ist und welche ins Proteininnere zeigt.

http://cti.itc.Virginia.EDU/~cmg/Demo/wheel/wheelApp.html

Kyte-Doolittle Hydrophobizitätsskala (1982)

Jede Aminosäure erhält Hydrophobizitäts-wert zugeordnet.

Um TM-Helices zu finden, addiere alle Werte in einem **Sequenzfenster** der Länge *w.*

Alle Fenster oberhalb einer Schranke *T* werden als TM-Helix vorhergesagt.

Beobachtung:
Gute Parameter sind $w = 19$ und $T > 1.6$.

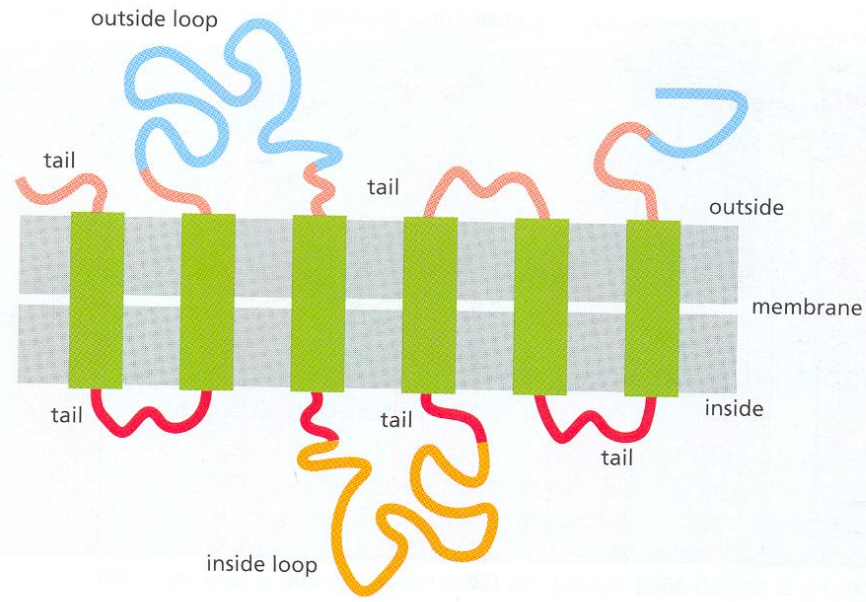
Hydrophobicity Scales

	Kyte-Doolittle	Hopp-Woods
Alanine	1.8	-0.5
Arginine	-4.5	
Asparagine	-3.5	
Aspartic acid	-3.5	
Cysteine	2.5	
Glutamine	-3.5	
Glutamic acid	-3.5	
Glycine	-0.4	
Histidine	-3.2	
Isoleucine	4.5	
Leucine	3.8	
Lysine	-3.9	
Methionine	1.9	
Phenylalanine	2.8	
Proline	-1.6	
Serine	-0.8	
Threonine	-0.7	
Tryptophan	-0.9	
Tyrosine	-1.3	
Valine	4.2	

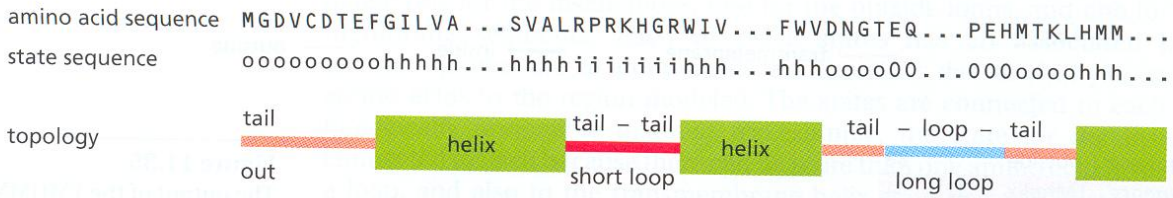
TM-Vorhersage mit Hidden Markov Modellen

HMMTOP: verwendet ein Hidden Markov-Modell um 5 strukturelle Zustände zu unterscheiden:

- Nicht-Membran Region innen
- TMH-Ende innen
- Membranehelix
- TMH-Ende außen
- Nicht-Membran Region außen



HMMTOP Vorhersage



Wie kann man 2 Proteinstrukturen vergleichen?

Paarweise Sequenzvergleiche

Paarweise Strukturvergleiche?

DALI (Distance-matrix Alignment)

L. Holm & C. Sander

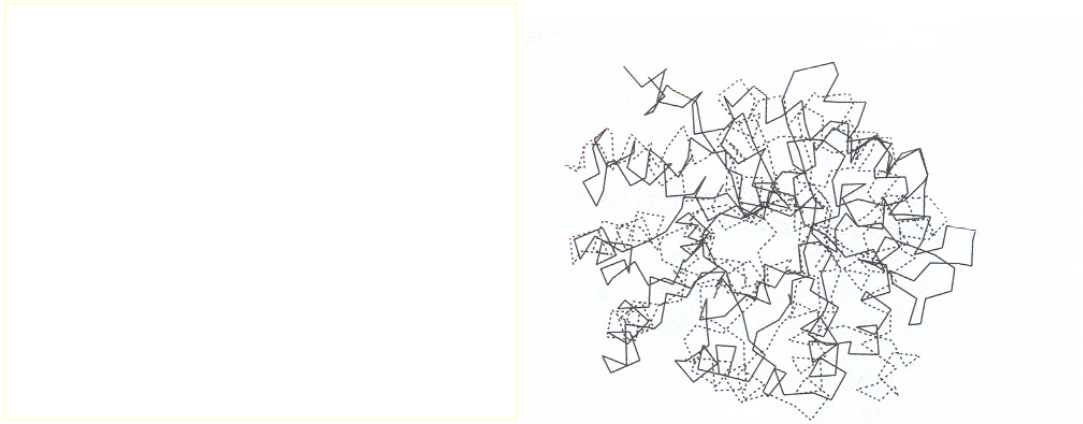
Während der Evolution eines Proteins verändert sich seine Sequenz.

Was häufig erhalten bleibt, ist die Verteilung der Kontakte zwischen den Aminosäuren.

→ Konstruiere Kontaktmatrizen für beide Proteine (leicht)

→ finde maximal übereinstimmende Untermatrizen der Kontaktmatrizen (schwierig)

http://www.ebi.ac.uk/dali



5.7 Abschnitte mit gemeinsamen Faltungsmustern, ermittelt mit dem Programm DALI von L. Holm und C. Sander. Es handelt sich um zwei Proteine mit TIM-*barrels*, die Adenosindesaminase der Maus [1FKX] (durchgezogene Linien) und die Phosphotriesterase aus *Pseudomonas diminuta* [1PTA] (gestrichelte Linien). Nach dem hier gezeigten Alignment stimmen die Ketten nur in 13 Prozent ihrer Aminosäuren überein – ein Wert, der eher im mitternächtlichen Dunkel denn in der Grauzone liegt.

Bedeutung von struktureller Äquivalenz

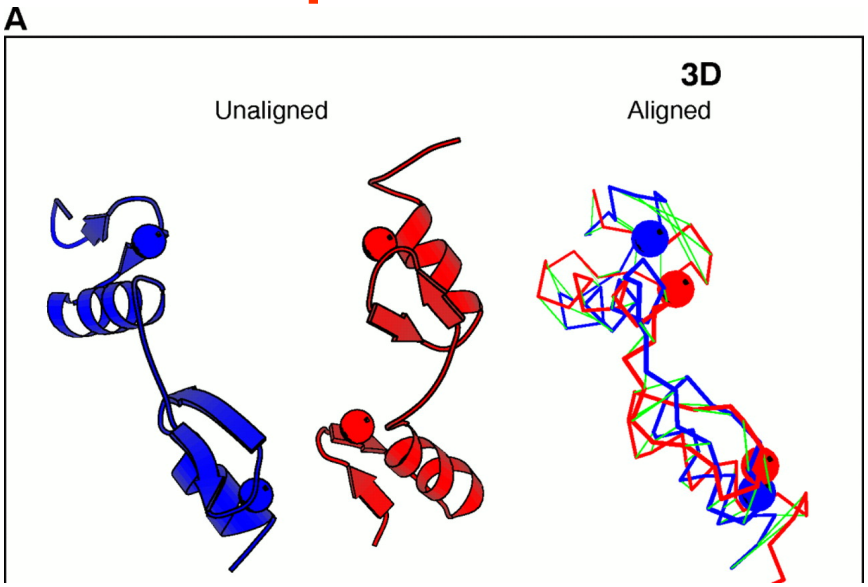
Beim Strukturvergleich sollen äquivalente Strukturblöcke zweier Proteine einander zugeordnet werden.

Darstellung

- in 3D als Überlagerung (superimposition) starrer Körper
- in 2D als ähnliche Muster in Distanz-Matrizen
- in 1D als Sequenzalignment

Rechts: Strukturvergleich von zwei **Zinkfinger-Proteinen**, tramtrack und MBP-1 [1bbo].

Holm, Sander Science 273, 5275 (1996)



3D-Überlagerung: finde Translation und Rotation eines Moleküls (rot: 1bbo), so dass es optimal auf das andere Molekül passt (blau: 2drpA).

Das Problem ist hier, dass die zwei Domänen der beiden Proteine unterschiedlich gegeneinander verdreht sind (vgl. parallele Lage der beiden roten Helices bzw. senkrechte Lage der beiden blauen Helices).

Ähnlichkeit zwischen zwei Proteinen

A

strong	medium	twilight zone	
100 % — 70 %		30 % — 0 %	sequence identity
point mutations only, essentially identical structures	insertions/ deletions mainly in loops	may or may not imply structural similarity	structural implication

B

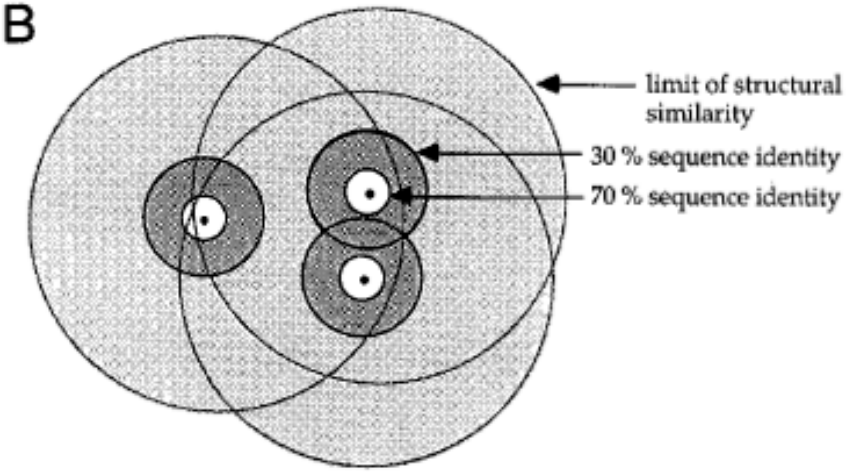


Fig. 1. Ranges of similarity between proteins. **A:** The structural implication of sequence similarity. Very little structural variation is observed in the 70–100% range of sequence identity. **B:** Protein families and relatives included in the database. There is one data set for each of 154 representative families. The central member of each family, i.e., the search structure, is shown by a dot in the center of concentric circles corresponding to strong sequence similarity (white), medium sequence similarity (shaded), and twilight zone (lightly shaded) with structural similarity but without significant sequence similarity. The central member of each family has less than 30% sequence identity with any other central member. For reasons of economy, the database of three-dimensional (3D) alignments only includes relatives in the medium and twilight zone range, but not in the strong range. In the medium range, the database contains all available 3D structures; in the twilight zone range, only representatives of each protein family. By definition, families can overlap, especially in the twilight zone; only a few proteins in the representative set share relatives in the medium zone (pair on the right). The large radius of convergence of the structure comparison search precludes the construction of nonoverlapping families but is of great advantage in discovering remote structural similarities.

Holm et al. Prot Sci 1, 1691 (1992)

Überraschende Ähnlichkeit zwischen papD und CD4 T-Zellrezeptor

Fig. 2. Structure alignment of papD and CD4. Structural alignment and optimal superposition of papD protein (3DPA) and CD4 T-cell receptor (1CD4). The C^{α}-rmsd after optimal superposition is 1.5 Å for 41 aligned residues. PDB residue numbers of the aligned fragments follow. For 3DPA: 16–38, 85–94, 105–112. For 1CD4: 111–133, 154–163, 166–173. The alignment was generated using the Suppos algorithm.

Holm et al. Prot Sci 1, 1691 (1992)

Überraschende Ähnlichkeit zwischen Flavodoxin und Malat-Dehydrogenase

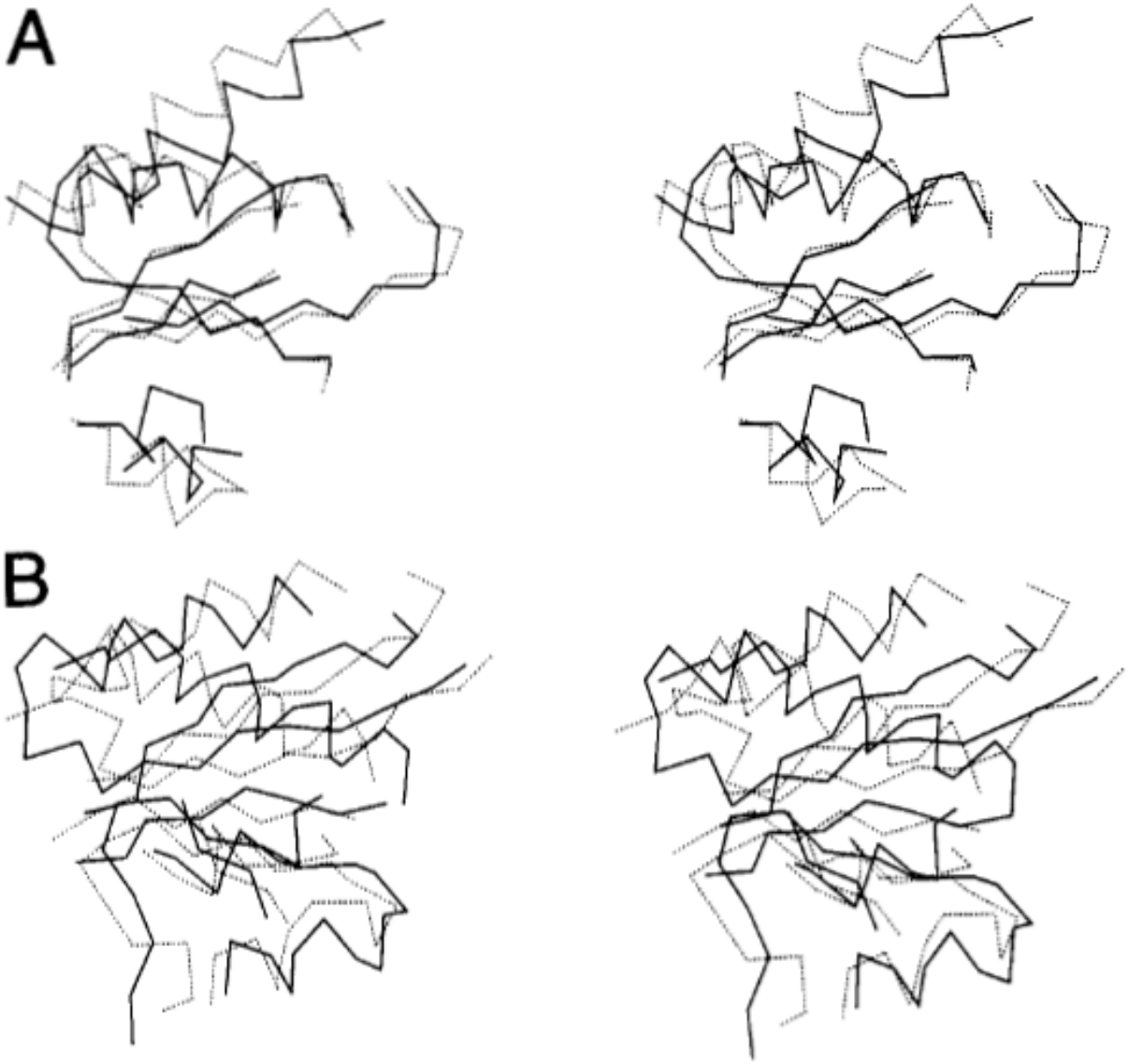


Fig. 3. Structure alignment of arabinose binding protein (ABP) with flavodoxin (FXN) and malate dehydrogenase (MDH). **A:** Structural alignment and optimal superposition of 1ABP and 4FXN. The C^{α}-rmsd after optimal superposition is 2.7 Å for 78 aligned residues. PDB residue numbers of the aligned fragments follow. For 1ABP: 5–15, 17–40, 49–54, 59–64, 73–79, 83–89, 253–269. For 4FXN: 1–35, 39–44, 47–52, 72–78, 80–86, 122–138. **B:** Structural alignment and optimal superposition of 1ABP and 4MDH. The C^{α}-rmsd after optimal superposition is 3.4 Å for 97 aligned residues. Residue numbers of the aligned fragments follow. For 1ABP: 2–27, 31–41, 45–53, 59–66, 73–94, 103–106, 254–270. For 4MDH (chain A): 1–26, 31–41, 72–88, 113–130, 135–138, 151–154, 240–256. The alignments were generated using the Comp3D algorithm.

Holm et al. Prot Sci 1, 1691 (1992)

Überraschende Ähnlichkeit zwischen Tryptophansynthase und Flavocytochrom b2

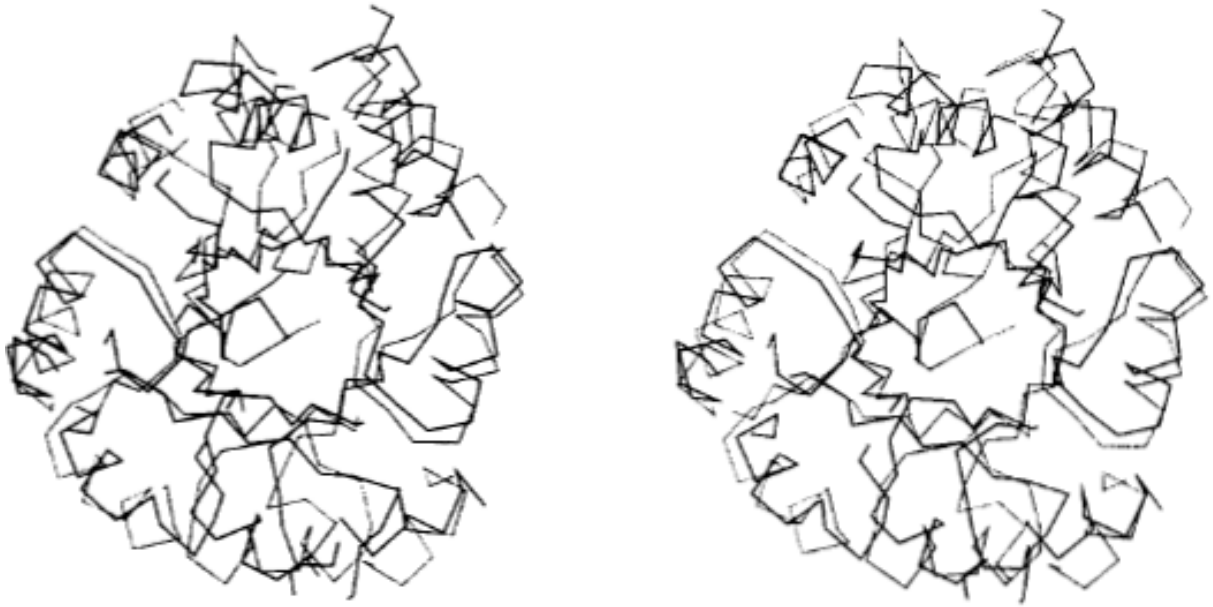


Fig. 4. Structure alignment of $(\alpha\beta)_8$ barrels. Structural alignment and optimal superposition of tryptophan synthase (1WSY) and flavocytochrome b2 (1FCB). The C^{α}-rmsd after optimal superposition is 3.1 Å for 198 aligned residues. PDB residue numbers of the aligned fragments follow. For 1WSY (chain A): 3–8, 17–28, 29–43, 44–53, 84–92, 93–105, 108–130, 131–144, 148–159, 162–177, 192–201, 202–243, 250–265. For 1FCB (chain A): 182–187, 190–201, 205–219, 221–230, 234–242, 245–257, 259–281, 330–343, 344–355, 356–371, 387–396, 400–441, 442–457. The alignment was generated using the Dali algorithm.

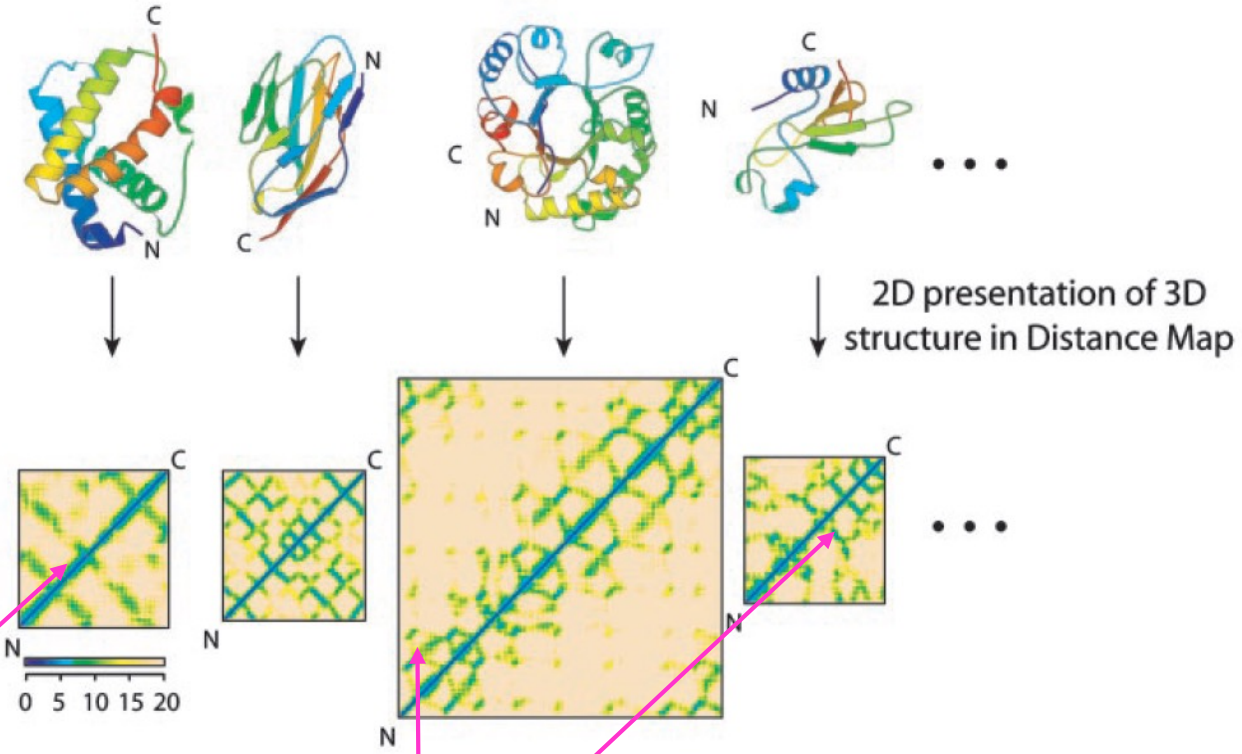
Holm et al. Prot Sci 1, 1691 (1992)

Distanzmatrix für Proteinstrukturen

Distanzmatrix:

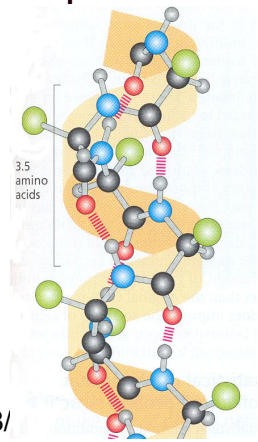
Auf beiden Achsen wird jeweils die Proteinsequenz aufgetragen.

Die Einträge der Matrix enthalten die Abstände zwischen den C_α – Atomen der Aminosäuren *i* und *j* dieses Proteins in der 3D-Struktur.



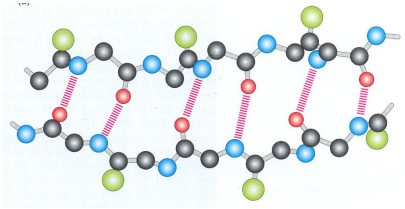
In einer **α-Helix** liegt Aminosäure *i* jeweils nah bei AS *i* + 4

→ In der Distanzmatrix ergeben diese Kontakte eine um 4 verschobene Linie parallel zur Diagonalen.



Parallele β-Stränge : ihre Kontakte ergeben ebenfalls eine verschobene Linie parallel zur Diagonalen.

Antiparallele β-Stränge : ihre Kontakte ergeben um 90° gekippte Linien.



Choi et al. PNAS 101, 3797 (2004)

Distanzmatrix bzw. Kontaktmatrix

(**B**) Distanzmatrix: schwarze Punkte markieren Paare von Residuen in 1bbo (unten) und 2drpA (oben) mit Abstand unter 12 Å.

Links: ohne Alignierung, schlechte Übereinstimmung der Kontakte.

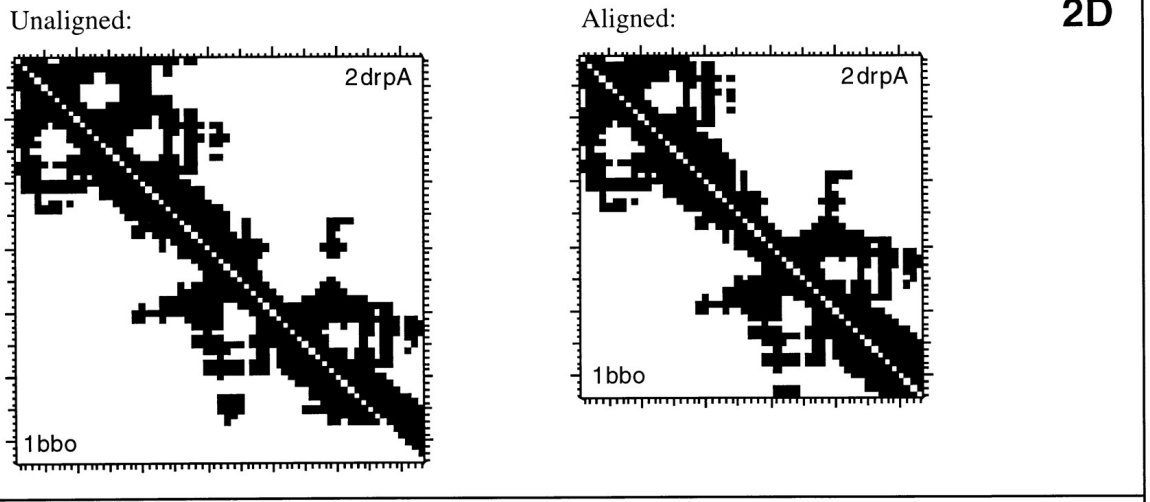
Rechts: nach Alignierung, wenn nur die Spalten und Reihen für sich strukturell entsprechende Residuen behalten werden.

(**C**) 1D Sequenzalignment.

Die die Zinkatome koordinierenden Histidin-Residuen werden aligniert.

Unterstrichen: Sekundärstruktur-elemente.

B **2D**



C **1D**

Unaligned:

```
1bbo     1 KYICEECGIRXKKPSMLKKHIRTHTDVRPYHCTYCNFSFKTKGNLTKHMKSKAHSKK  57

2drpA  103 FTKEGEHTYRCKVCSRVYTHISNFCRHYVTSHKRNVKVYPCPFCFKEFTRKDNMTAHVKIIHK 165
```

Aligned:

```
1bbo     1 .......KYICEECGIRXKKPSMLKKHIRTHt..DVRPYHCTYCNFSFKTKGNLTKHMKSKAHskk  57
                   | |  |       |    | |     | | | |   |  | | | | |
2drpA  103 ftkegehTYRCKVCSRVYTHISNFCRHYVTShkrNVKVYPCPFCFKEFTRKDNMTAHVKIIHK... 165
```

Holm, Sander Science 273, 5275 (1996)

DALI verwendet einen branch-and-bound Algorithmus

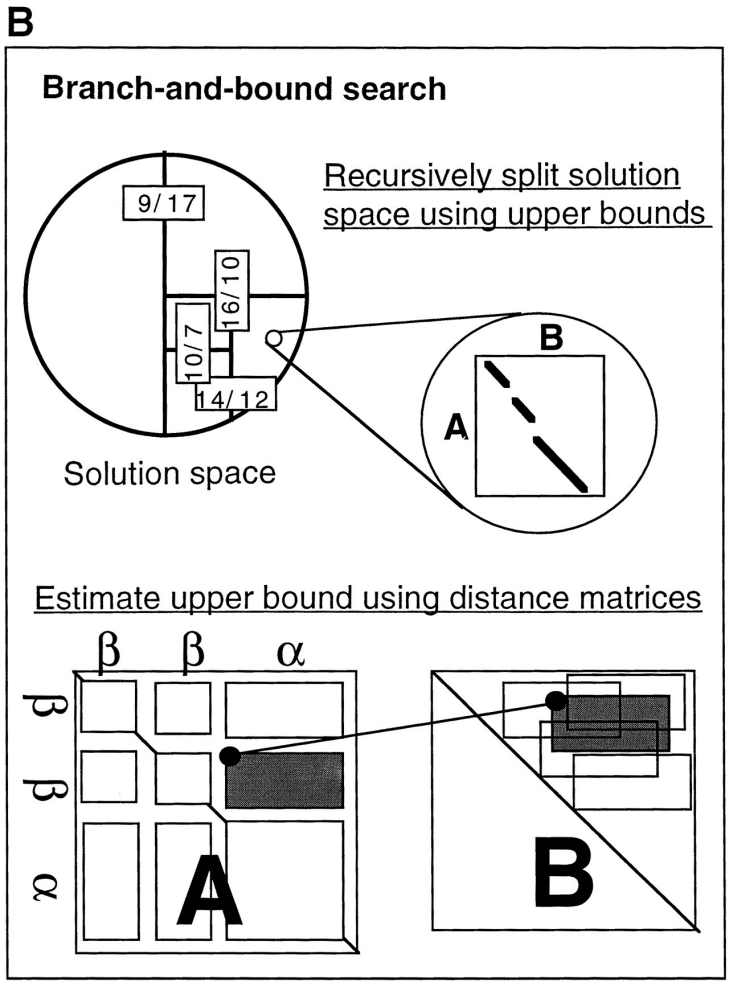
(**B**) A branch-and-bound algorithm is guaranteed to yield the global optimum but may, in the worst case, need an exponential number of steps to do so.

First, protein structures A and B are represented by distance matrices (bottom left and right; each point in a matrix is a residue-residue distance; an internal square is a set of contacts made by two segments; the secondary structure segments are β,β, and α).

The problem of shape comparison becomes one of finding a best subset of residues in each matrix (subsets of rows and columns) such that the set of residues in protein A has a similar pattern of intramolecular distances as the set in protein B.

A single solution to the problem is given in terms of the two sets of equivalent residues (an alignment). The solution space consists of all possible placements of residues in protein B relative to the segments of residues of protein A. The key algorithmic idea is to recursively split the solution subspace (schematically shown as a circle at upper left, in which each point is a solution to the problem and the lines divide subsets of solutions) that yields the highest upper bound until there is a single alignment trace left:

start with the entire circle; calculate the upper bound for the left (9) and right (17) half; choose the right half and split it into top (upper bound 10) and bottom (upper bound 16) quarters; choose the bottom part and split it (left: 14; right: 12); choose the right part; and so on until the area of solution space has shrunk to a single solution (shown as the residue-residue alignment matrix enlarged at right). The upper bound for each part of the solution space is estimated in terms of a simplified subproblem that asks for the best match of residues in protein B onto a predefined set of residues in protein A (the match is illustrated by the circle-ended line connecting the single square in matrix A with a set of candidate squares in matrix B). The best match is the one with the maximal pair score (sum of similarities of distances between the square in A and the square in B). The predefined set corresponds to residues in secondary structure elements. The upper bound for each of the segment-segment submatrices of matrix A is found by calculating the similarity scores between the submatrix in A and all accessible submatrices in B. An upper bound of the total similarity score (sum over all segment-segment submatrices in A) for one set of solutions is given by the sum of separately calculated upper bounds for each segment-segment pair of matrix A.



Holm, Sander Science 273, 5275 (1996)

Folie nicht klausurrelevant

Zusammenfassung

- Proteinstrukturen sind hierarchisch aufgebaut

- Die Kenntnis der 3D-Struktur erlaubt es, die Proteinfunktion mechanistisch zu verstehen, z.B. von Enzymen katalysierte chemische Umwandlungsschritte.

- die strukturelle Bioinformatik beschäftigt sich u.a. mit der Vorhersage von 2D- und 3D-Struktur aus der 1D-Struktur (Sequenz)

- Vorhersagen von 2D-Strukturelementen sind ca. 80% genau

- Die Aminosäurezusammensetzung der Membranregionen von Membranproteinen ist sehr verschieden von der löslicher Proteine.

- Dadurch kann man Transmembranregionen recht zuverlässig identifizieren

- Der Vergleich mehrerer Proteinstrukturen ist nicht trivial.