INHALT

V5: Proteinstruktur: Sekundärstruktur

- 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

In zweiten Teil dieser Vorlesung werden wir uns 3 Doppelstunden lang mit Aspekten der Proteinstruktur beschäftigten. Heute führen wir zunächst ein paar Begriffe ein, die Ihnen wohlbekannt sein dürften wie etwa der Ramachandran-Plot oder die Sekundärstrukturelemente Alpha-Helix und beta-Strang. Am Ende der heutigen Vorlesung werden wir uns dann mit einer Datenstruktur beschäftigen, der sogenannten Distanzmatrix, die sehr gut geeignet ist um zwei Proteinstrukturen miteinander zu vergleichen. In Analogie zum paarweisen Alignment zweier Sequenzen geht es dabei als um das paarweise Alignment zweier Strukturen.

Im Tutorial werden Sie damit vertraut gemacht, dreidimensionale Proteinstrukturen zu visualisieren. Dies ist ein wichtiger Arbeitsschritt z.B. im virtual drug design.

Man kann sich natürlich fragen, ob es wirklich notwendig ist, die räumliche Struktur von Proteinen mit aufwändigen experimentellen Methoden aufzuklären. Ja, ist die eindeutige Antwort. Erst durch die Kenntnis der Proteinstruktur kann man wirklich die Funktion von Proteinen aufklären. Allerdings zeigen sich viele interessante strukturelle Eigenschaften bereits in den Sequenzen einer Proteinfamilie. Dies hatten wir bereits in V3 am Beispiel der Thioredoxine kennengelernt.

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

Proteine übernehmen in einer biologischen Zelle eine Vielfalt an unterschiedlichen Aufgaben. Vermutlich wird dies auch Auswirkungen auf ihre Struktur und Dynamik haben. So haben Strukturproteine naturgemäß eher steife Konformationen, wohingegen sich Enzyme oft durch eine gewisse Beweglichkeit auszeichnen.

Warum sind Proteine so groß?

Proteine sind große Moleküle.

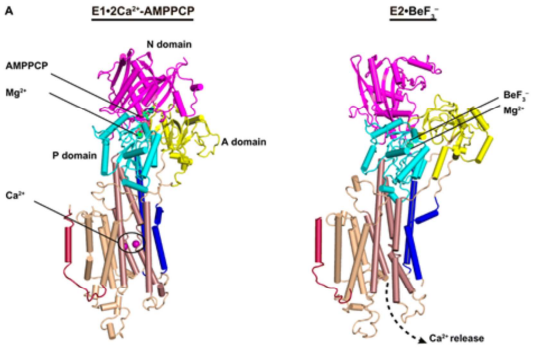
Ihre **Funktion** ist oft in einem kleinen Teil der Struktur,
dem **aktiven Zentrum**, lokalisiert.

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

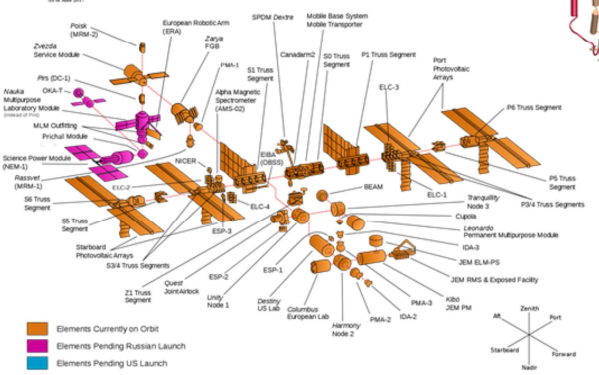
Manche Proteine sind „riesengroß“, obwohl ihre wirkliche Funktion in einem relativ kleinen „aktiven Zentrum“ konzentriert ist. Als ein aktives Zentrum bezeichnet man die Aminosäuren eines Enzyms, aus denen die Bindungstasche gebildet wird, in der dann das Substrat der enzymatisch katalysierten Reaktion bindet. Allerdings können diese Aminosäuren ja nicht einfach „in der Luft aufgehängt“ werden. Stattdessen müssen sie im Allgemeinen an Sekundärstrukturelementen des Proteins verankert sein, damit sie eine feste Verankerung besitzen um z.B. chemische Verbindungen der Substrate aufbrechen bzw. auseinanderziehen zu können. Diese umliegenden Sekundärstrukturelemente müssen wiederum in ein bestimmte relative Orientierung zueinander gebracht werden, was einfach eine bestimmte minimale Größe des Proteins mit sich bringt. Auf der Oberfläche des Proteins liegen außerdem die Bindungs-Schnittstellen für andere Proteine, DNA oder Membranen. Auch diese müssen eine bestimmte Fläche besitzen um stabile Interaktionen ausbilden zu können. Manche Proteine müssen zudem große Konformationsänderungen durchführen. Auch dazu müssen sie groß genug sein, z.B. aus 2 oder mehr Domänen bestehen.

Warum sind Proteine so groß?

Evolution der Proteine:
Veränderungen der Struktur, die durch Mutationen in ihrer Aminosäure-sequenz hervorgerufen werden.



ISS Configuration



https://de.wikipedia.org/wiki/Internationale_Raumstation
Zhang et al. *Science Advances* (2020) 6, eabb0147

5. Vorlesung WS 2020/21 Softwarewerkzeuge 4

Ein wichtiger Grund für die Größe von Proteinen ist allerdings ihre Entstehung. Oft werden komplizierte Funktionen nämlich dadurch realisiert, dass sich im Laufe der Evolution mehrere Proteindomänen hintereinander miteinander verknüpften, wobei jede Domäne eine bestimmte Funktion hat. Ein Ingenieur hätte diese Kombination an Funktionen vielleicht auch mit einem anderen Design erreichen können. In der Natur war die Aneinanderheftung bewährter Elemente oft die einfachste und naheliegendste Lösung. Und danach gilt das bewährte Prinzp „never change a winning system“, bzw. die Konservierung von bewährten Bauelemente.

Man kann diese Entstehung mit dem Aufbau der links gezeigten internationalen Raumstation ISS vergleichen, die ja ebenfalls aus vielen verschiedenen Modulen besteht, die nacheinander aneinander angeheftet wurden und werden.

Rechts sind zwei Cryo-EM-Struktur der SERCA-Pumpe in verschiedenen Zuständen gezeigt. Dies ist ein integrales Membranprotein, das Calcium durch die Membran in das endoplasmatische Retikulum pumpt. Der untere Teil besteht aus 10 Transmembranhelices, die die Membran (die man sich waagerecht hinzudenken kann) durchqueren. Der obere Teil enthält eine Aktivierungsdomäne, eine Nukleotid-bindende Domäne und eine Phosphorylierungsdomäne. Dieser externe Bereich des Proteins führt eine sehr große Konformationsänderung aus um in dem daran gekoppelten Membranbereich ein Calcium-Ion entgegen dem Konzentrationsgradienten von Calcium durch die Membran ins ER

hineinpumpen zu können. Die dafür benötigte Energie stammt aus der Bindung des Nukleotids. Diese komplizierte Funktion bedingt einfach eine bestimmte Mindestgröße und Komplexität der Proteinstruktur.

Hierarchischer Aufbau

Primärstruktur – Sekundärstruktur – Tertiärstruktur – Quartärnere Struktur – Komplexe

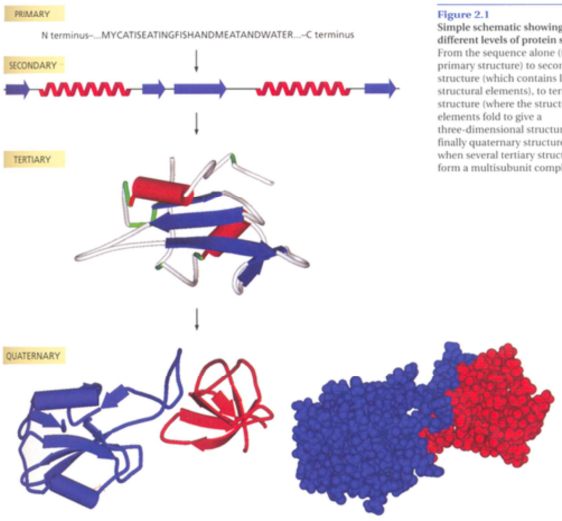


Figure 2.1
Simple schematic showing the different levels of protein structure. From the sequence alone (the primary structure) to secondary structure (which contains local structural elements), to tertiary structure (where the structural elements fold to give a three-dimensional structure), to finally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

Für die Proteinstruktur gilt der wohlbekannte hierarchische Aufbau: Primärstruktur (Sequenz) – Sekundärstrukturelemente (dazu kommen wir gleich) – gefaltete Tertiärstruktur eínes Proteins – Quartärstruktur (Aneinanderlagerung mehrerer einzelner Proteine).

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.
Man bezeichnet sie daher auch als „inside out“ Proteine.
Die wasserlöslichen Bereiche von Membranproteinen ähneln in ihrer Zusammensetzung den löslichen Proteinen.

5. Vorlesung WS 2020/21 Softwarewerkzeuge 6

Eine wichtige Überlegung für das Verständnis von Proteinstrukturen ist, welche Wechselwirkungen die Faltung von Proteinen begünstigen.

Entropisch gesehen (nach dem 2. Hauptsatz der Thermodynamik wird der Zustand „maximaler Unordnung“ angestrebt) ist die Faltung einer langen Proteinsequenz in eine kompakte Struktur aus Sicht der langen Peptidkette sehr ungünstig. Allerdings kommt es bei der energetischen Betrachtung nicht nur auf die Peptidkette an, sondern auch auf das umgebende Lösungsmittel. Dieses ist meist Wasser. Wassermoleküle finden hydrophobe Moleküle „schrecklich“, da sie mit ihnen keine Wasserstoffbrückenbindungen ausbilden können. Jedes Wassermolekül bildet im flüssigen Zustand 3.7 Wasserstoffbrückenbindungen aus. Aus Sicht des Wassers wäre es für das Gesamtsystem die optimale Lösung, wenn sich die Peptidkette auf ein möglichst kleines Volumen „zurückzieht“ und dabei möglichst alle hydrophoben Aminosäurereste in seinem Inneren begräbt und auf der Oberfläche des Proteinklumpens sich möglichst nur polare oder geladene Reste befinden. Genau dies ist die Basis des **hydrophoben Effekts**.

Bei Membranproteinen gilt dasselbe Prinzip für alle Bereiche außerhalb der Membran. Innerhalb der Membran ist der Fall jedoch umgekehrt. Umgeben wird das Protein dort von den extrem hydrophoben Lipidketten der Phospholipidmembran. In ihrem Inneren enthalten Transporter oder Kanäle oft eine relative polare Pore, durch die der Austausch von polaren Substanzen von einer Seite der Membran auf die andere stattfindet. Deshalb haben Transmembranbereiche eine „inside out“

Zusammensetzung, also außen hydrophob und innen relativ polar.

Hydrophober Effekt

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

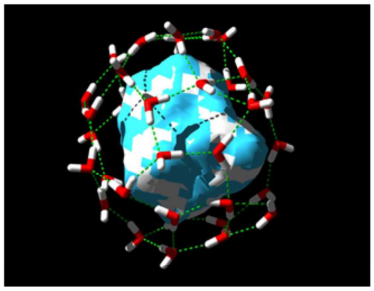
(a) energetisch stark ungünstig ist

(b) bei Raumtemperatur zu einer Abnahme der Entropie führt

(c) zu einer Zunahme der Wärmekapazität führt.

Eisberg-Modell

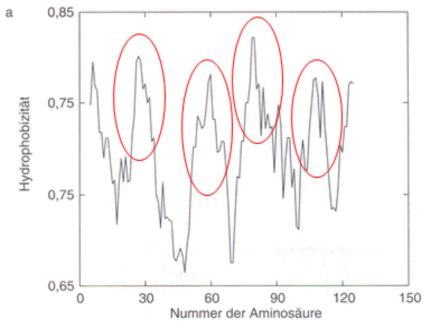
W. Kauzman 1959

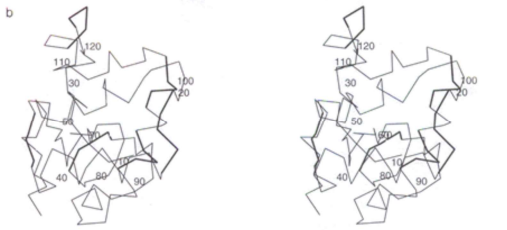


Wassermoleküle an einer hydrophoben Oberfläche sind in ihren möglichen Orientierungen stark eingeschränkt -> dies ist entropisch ungünstig.

Nach dem sogenannten Eisberg-Modell, das auf Walter Kauzman zurückgeht, „frieren“ Wassermoleküle um eine hydrophobe Substanz herum gewissermaßen wie im Eiszustand ein. Dies führt zu einer ungünstigen Abnahme von deren Entropie und ist außerdem enthalpisch ungünstig. Dieses Modell ist zwar physikalisch mittlerweile überholt, allerdings behält es eine didaktische Einfachheit und Symbolik, so dass wir es hier zur Illustration des hydrophoben Effekts weiterhin verwenden werden.

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.

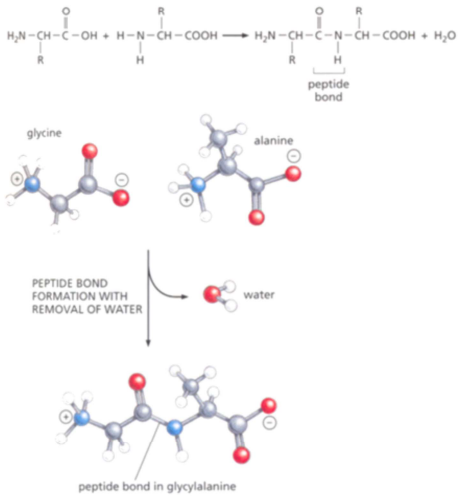
Man kann nun einfach einmal die Hydrophobizität der einzelnen Aminosäuren entlang einer Proteinsequenz auftragen. Dies ist im oberen Beispiel für das Enzym Lysozym aus Hühnereiweiß gezeigt. Die Minima dieses Hydrophobizitätsprofils liegen in der unteren Proteinstruktur von Lysozym auf der Oberfläche des Proteins (dickere Linien). Daher kann man ableiten, dass die Minima des Hydrophobizitätsprofils hydrophile Abschnitte sind und die Maxima des Hydrophobizitätsprofils im Proteininneren liegen. Das obere Profil legt nahe, dass die Aminosäurekette das Proteininnere etwa viermal von einer Seite zur anderen durchquert.

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**“.

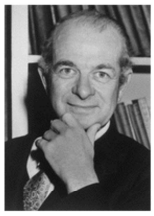
Die Kenntnis der Sequenz eines Proteins allein verrät noch nicht viel über seine Funktion. Entscheidend ist seine **drei-dimensionale Struktur.**



Jeweils 2 Aminosäuren bilden unter Wasserabspaltung eine Peptidbindung aus. Die gesamte Sequenz des Proteins entspricht einer linearen Abfolge von Peptidbindungen.

Eigenschaften der Peptidbindung

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



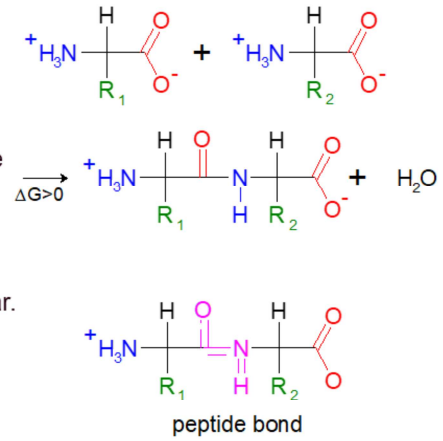
Linus Pauling
Nobelpreise für
Chemie 1954 und
Frieden 1963

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.

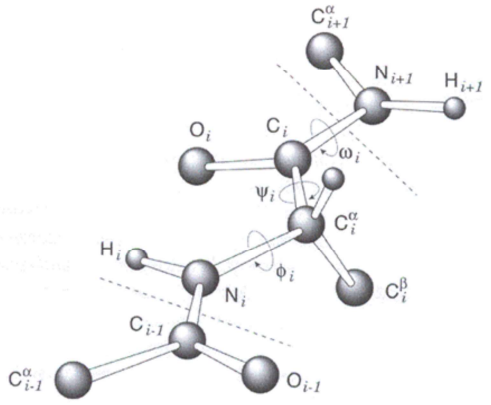


Jede Aminosäure steuert 3 Bindungen zum Peptidrückgrat bei. Die zentrale C-N-Bindung hat einen partiellen Doppelbindungscharakter, d.h. eine Sigma-Bindung und die Hybridisierung der Pi-Elektronenorbitale. Deshalb ist diese Bindung nicht frei drehbar, die beiden anderen jedoch schon.

Diederwinkel des Proteinrückgrats

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

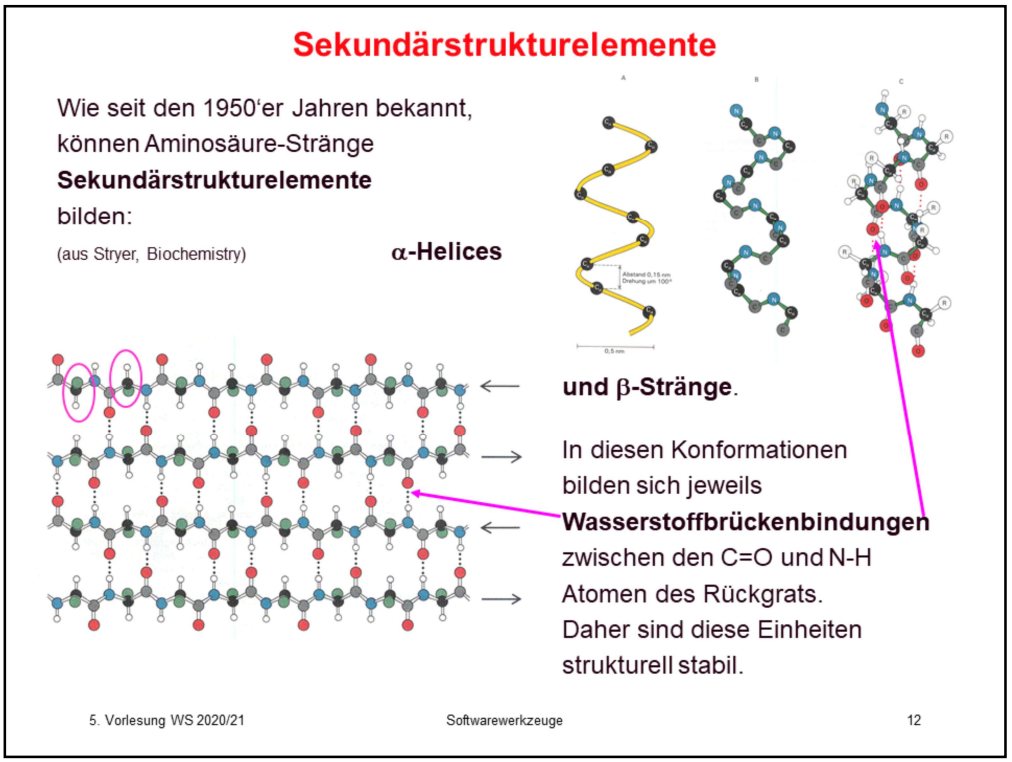
Pro Residue gibt es 2 frei drehbare Diederwinkel, die als Φ und Ψ bezeichnet werden.



Definition der Konformationswinkel im Polypeptidrückgrat.

Lesk-Buch

Ein Diederwinkel beschreibt den Winkel zwischen vier Atomen, d.h. wie die Ebenen der Atome 1-2-3 und der Atome 2-3-4 gegeneinander gekippt sind.



In der rechts gezeigten alpha-Helix zeigen alle N-H-Bindungen des Rückgrats „nach oben“ und alle C=O-Bindungen „nach unten“. Zwischen den N-Hs der i-ten Aminosäure und den C=Os der i+4-ten Aminosäure bildet sich dabei jeweils eine Wasserstoffbrückenbindung aus. Dies hat zwei Konsequenzen:

(1) N-H-Gruppen sind partiell elektrisch positiv geladen, C=O-Gruppen partiell elektrisch negativ. Jedes dieser Paare bildet einen kleinen elektrischen Dipol. Da nun alle diese Dipole in dieselbe Richtung zeigen, bilden alpha-Helices einen Nettodipol. An ihrem positiven Ende kann sich ein Anion günstig anlagern und umgekehrt ein Kation am negativen Ende.

(2) Falls in einer Helix ein Prolin liegen würde, besitzt dieses keine freiliegende N-H-Gruppe im Rückgrat, da dort die Seitenkette kovalent an das Stickstoffatom gebunden ist. Daher kann das Prolin keine H-Bindung ausbilden und die alpha-Helix bekommt zwangsläufig einen Knick an dieser Stelle, da sich die darüberliegende C=O-Gruppe eine andere Möglichkeit suchen muss um eine H-Bindung auszubilden.

Unten ist ein antiparalleles beta-Faltblatt gezeigt, das aus 4 beta-Strängen besteht, die abwechselnd von rechts nach links und von links nach rechts laufen. Genau wie bei der alpha-Helix bilden sich H-Bindungen zwischen N-H-Gruppen und C=O-Gruppen aus. Interessanterweise zeigen die Aminosäurereste abwechselnd aus der Tafelebene hinaus und hinein (lila Kreise). Dies hatten wir bereits in V3 bei der Thioredoxin-Familie

besprochen.

DSSP

Der DSSP-Algorithmus geht zurück auf Wolfgang Kabsch & Chris Sander (1983).

DSSP steht für *Define Secondary Structure of Proteins*.

DSSP benutzt eine elektrostatische Energiefunktion um H-Bindungen zwischen Atomen des Proteinrückgrats zu identifizieren.

Man unterscheidet dann 3 **helikale Konformationen**:

3_{10} Helix (DSSP-Symbol **G**) – mehrere H-Bindungen zwischen Residuen *i* und *i+3*

α Helix (**H**) – mehrere H-Bindungen zwischen Residuen *i* und *i+4*

π Helix (**I**) – mehrere H-Bindungen zwischen Residuen *i* und *i+5*

2 Typen von **Beta-Faltblatt-Strukturen**:

beta Brücken (**B**) bzw. längere Abfolgen von H-Bindungen (**E**)

Turns (**T**)

S : sehr gekrümmte Abschnitte

C: sonstige – meist Loops (Schleifen) an der Proteinoberfläche

DSSP ist eine sehr verbreitete strukturelle Klassifizierung von Sekundärstrukturelementen. Es gibt neben der kanonischen alpha-Helix zwei weitere Arten von Helices.

Einen beta-Strang bezeichnet man mit E. Einen sehr kurzen beta-Abschnitt als beta-Brücke.

Bei einem Turn vollzieht die Peptidkette eine sehr abrupte 180 Grad-Umkehr, vergleichbar mit der **Saarschleife**.

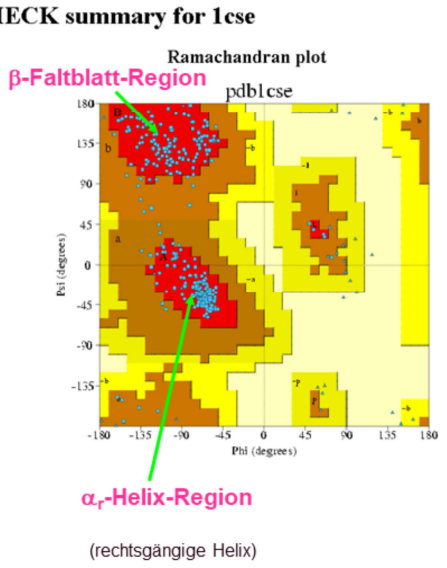
Stabilität und Faltung von Proteinen

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.



Dies ist der bekannte Ramachandran-Plot, bei dem wir die Winkelkombination für das Rückgrat jeder einzelnen Aminosäure eines Proteins eintragen. Hier z.B. für die PDB-Struktur 1cse.

Auf der x-Achse liegt der Winkel Phi, auf der y-Achse der Winkel Psi. Es gibt 2 Bereiche, die sehr viele Aminosäuren enthalten. Dies sind die Regionen der Sekundärstrukturelemente.

Der Einfachheit halber kann man sich merken, dass beta-Faltblätter fast planar sind, d.h. die beiden Winkel des Rückgrats nahe bei 180 Grad liegen.

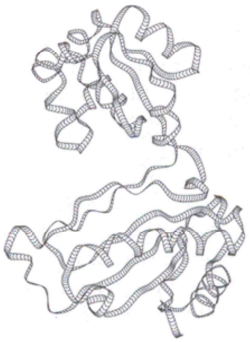
Domänen

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

5. Vorlesung WS 2020/21 Softwarewerkzeuge 15

Das nächste Strukturelement oberhalb der Sekundärstrukturelemente sind die Domänen von Proteinen. Als Domäne bezeichnet man eine räumliche kompakte Einheit.

Die linke Abbildung zeigt die katalytische Untereinheit der cAMP-abhängigen Proteinkinase. Die katalytischen Untereinheiten von Kinasen bestehen aus 2 Domänen, einer großen alpha-helikalen Domäne (unten) und einer kleinen beta-Faltblatt-Domäne (oben). Dazwischen liegt die Bindungstasche für ATP. In der Abbildung ist ATP als stick model angedeutet. Die beiden Domänen können sich wie ein PacMan-Männchen relativ zueinander bewegen und dadurch bzgl. der ATP-Bindungstasche entweder eine offene oder eine geschlossene Konformation einnehmen.

Die Abbildung in der Mitte zeigt wiederum die SERCA-Pumpe, die wir bereits kennengelernt haben.

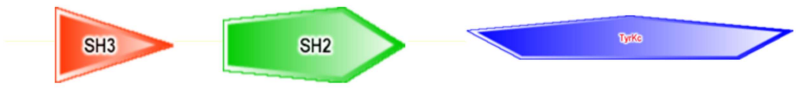
Das rechte Beispiel ist wiederum ein Protein, das aus 2 Domänen besteht.

Modular aufgebaute Proteine

Modular aufgebaute Proteine bestehen aus mehreren Domänen.

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

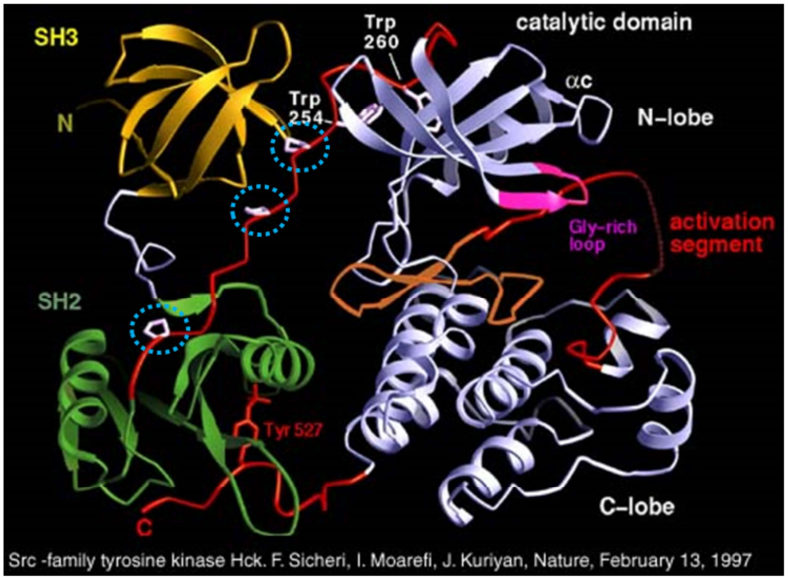
```
Sequenz:
MGGRSSCEDP GCPRDEERAP RMGCMKSKFL QVGGNTFSKT ETSASPHCPV
YVPDPTSTIK PGPNSHNSNT PGIREAGSED IIVVALYDYE AIHHEDLSFQ
KGDQMVVLEE SGEWWKARSL ATRKEGYIPS NYVARVDSLE TEEWFFKGIS
RKDAERQLLA PGNMLGSFMI RDSETTKGSY SLSVRDYDPR QGDTVKHYKI
RTLDNGGFYI SPRSTFSTLQ ELVDHYKKGN DGLCQKLSVP CMSSKPQKPW
EKDAWEIPRE SLKLEKKLGA 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
```



SMART identifiziert Domänen durch deren HMM-Signatur (für > 1300 verschiedene Domänen).

Der Webserver SMART identifiziert Domänen in Proteinsequenzen mit Hilfe von einem Hidden Markov-Modellen für die unterschiedlichen Domänen (vgl. V4). In dem gezeigten Beispiel identifiziert SMART in der Proteinkinase HcK zunächst eine SH3-Domäne (die bekanntermaßen Prolin-reiche Peptide binden kann), eine SH2-Domäne (die Peptide binden kann, welche ein phosphoryliertes Tyrosin enthalten) und eine katalytische Tyrosin-Kinase-Domäne. Wie könnten diese 3 Domänen nun strukturell angeordnet sein?

Beispiel: Src-Kinase HcK



http://jkweb.berkeley.edu/

Diese Abbildung zeigt die Kristallstruktur der Proteinkinase HcK. Rechts ist weiß gefärbt die katalytische TyrK gezeigt. Sie besteht wiederum aus einer großen alpha-helikalen Domäne (unten, mit C-lobe beschriftet) und der kleinen beta-Faltblatt-Domäne (oben, mit N-lobe beschriftet). In der Sequenz davor liegt die SH2-Domäne (grün, links unten). Der Linker zwischen Kinase-Domäne und SH2 (rot gefärbt) zieht sich diagonal von Mitte oben nach links unten und enthält die Tryptophane 260 und Trp254 und 3 Proline (durch gestrichelte, blaue Kreise umrandet). In dieser Konformation bindet die SH2-Domäne an das phosphorylierte Tyr527 im C-Terminus der Kinase-Einheit. Somit wird ein enger struktureller Kontakt mit der Kinase-Domäne hergestellt. Vor der SH2-Domäne liegt in der Sequenz die SH3-Domäne (gelb, links oben). Sie ist durch einen weißen Linker mit der SH2-Domäne verbunden. Zudem macht sie Kontakte mit 2 Prolinen in dem roten Linker zwischen SH3-Domäne und Kinase-Domäne. Dadurch „krallt“ sich die SH3-Domäne an den beiden anderen Domänen fest.

Strukturelle 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

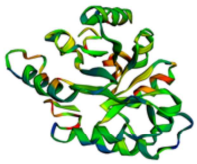
SCOP: Structural Classification of Proteins. **1.69** release
25973 PDB Entries (1 Oct 2004). 70859 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

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

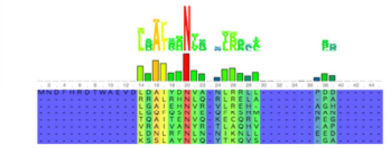
Die allgemeinste Einteilung in Familien von Proteinstrukturen stützt sich auf die Sekundär- und Tertiärstrukturen:

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	

Representative CATH Domain



ce alignment



Lesk-Buch

https://academic.oup.com/nar/article/47/D1/D280/5162467

5. Vorlesung WS 2020/21 Softwarewerkzeuge 18

Man unterscheidet generell alpha-helikale und beta-Faltblatt-Proteine, sowie Mischformen (siehe Tabelle links unten).

Es gibt zwei wichtige automatische strukturelle Klassifikations-Schemata für Proteine:

(1) die sogenannte SCOP-Klassifikation (http://scop.mrc-lmb.cam.ac.uk/) . SCOP wurde von Cyrus Chothia am berühmten MRC Laboratory for Molecular Biology (LMB) in Cambridge initiiert. Am LMB wurde die Sanger-Sequenzierung von Proteinsequenzen entwickelt, die ersten Proteinstrukturen bestimmt etc. Insgesamt wurden bisher 12 Nobelpreise an Mitglieder des LMB verliehen. Dies ist die Originalpublikation von SCOP: https://pubmed.ncbi.nlm.nih.gov/7723011/

Dies ist der Link zu der aktuellen Publikation zu SCOP: https://academic.oup.com/nar/article/48/D1/D376/5625529

Die strukturelle Klassifikation von Faltungsmustern (folds) in alpha/beta-Proteine wurde manuell vorgenommen. Der aktuelle Release enthält 1388 Folds.

Diesen zugeordnet sind 5060 Proteinfamilien und 2455 Superfamilien, die wieder über Hidden Markov Modelle klassifiziert werden.

(2) Eine zweite, weit verbreitete Klassifikation CATH wird am University College London in der Gruppe von Christine Orengo gepflegt (https://www.cathdb.info/).

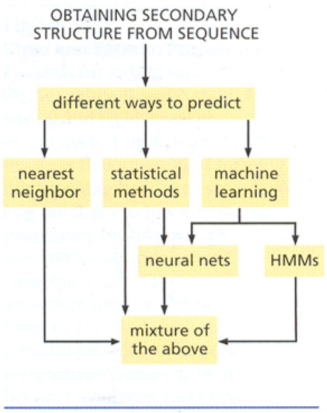
Die aktuelle Publikation zu CATH ist
https://academic.oup.com/nar/article/47/D1/D280/5162467

Unten rechts ein Beispiel gezeigt, wie eine Eingabesequenz QUERY wiederum aufgrund eines HMM-Sequenzmotifs einer Proteinfamilie zugeordnet wird. Die gezeigte Kristallstruktur gehört zu der Sequenz 2rjgAO2 und zeigt die repräsentative CATH-Domäne dieser Proteinfamilie.

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

Im zweiten Teil der heutigen Vorlesung beschäftigen wir uns mit Algorithmen zur Vorhersage von Sekundärstrukturelementen in Proteinsequenzen. Dies ist ein klassisches Feld der strukturellen Bioinformatik. Eine der ersten Methoden stammt von Peter Y. Chou und Gerald D. Fasman und wurde 1974 publiziert (https://pubmed.ncbi.nlm.nih.gov/4358940/).

Heute werden verschiedene Methode des statistischen/maschinellen Lernens für diese Aufgabe eingesetzt, siehe Abbildung.

Aufgrund ihrer Verschiedenheit müssen wir das Problem separat für lösliche Proteine und Membranproteine angehen.

am häufigsten auftretende Sekundärstrukturen

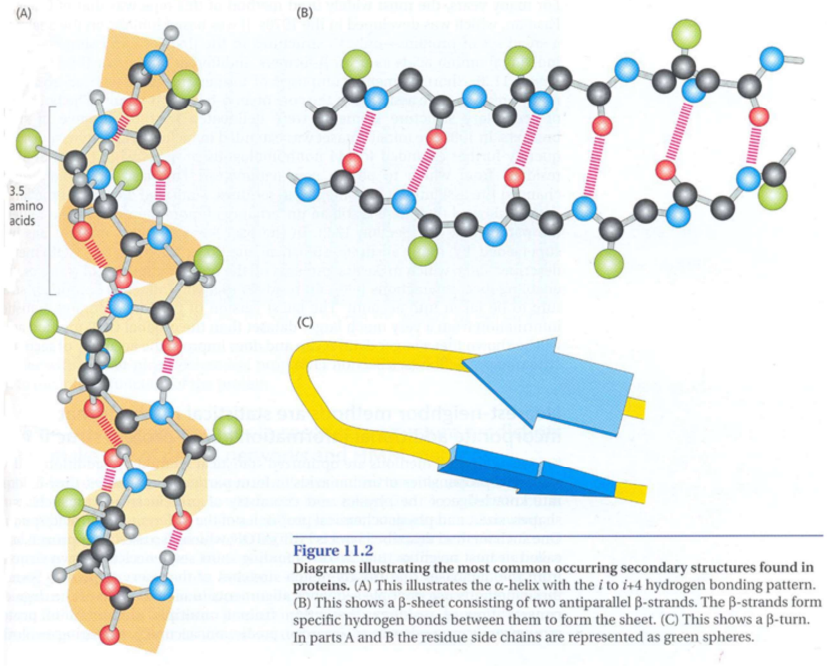
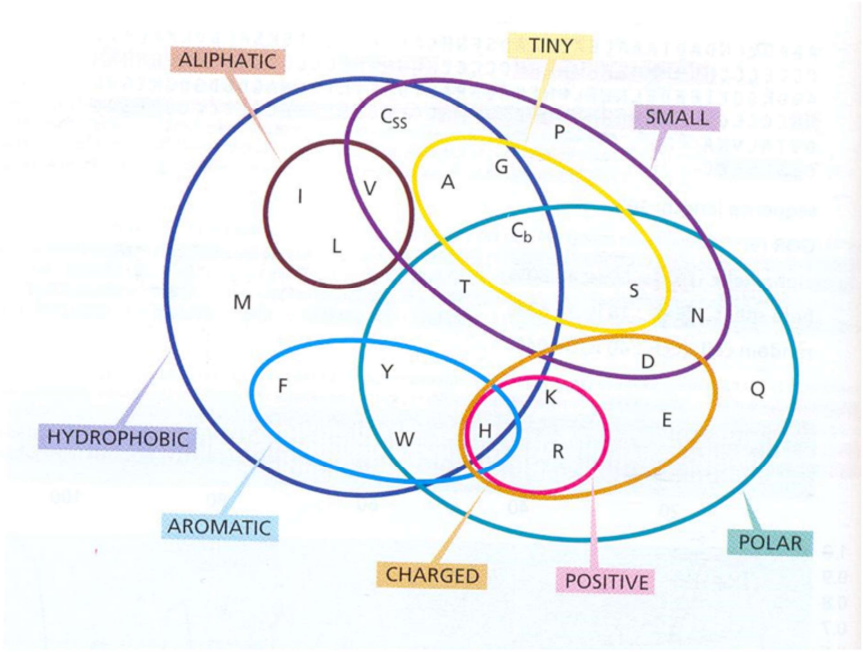


Figure 11.2
Diagrams illustrating the most common occurring secondary structures found in proteins. (A) This illustrates the α-helix with the *i* to *i*+4 hydrogen bonding pattern. (B) This shows a β-sheet consisting of two antiparallel β-strands. The β-strands form specific hydrogen bonds between them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

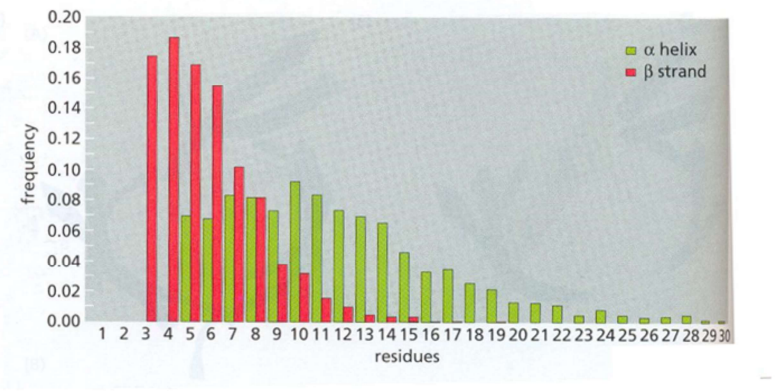
Diese Folie erinnert noch einmal an die zwei grundlegenden Arten von Sekundärstrukturelementen (alpha und beta), die wir heute bereits besprochen haben.

Die 20 natürlichen Aminosäuren



Diese Folie erinnert noch einmal an die unterschiedlichen Eigenschaften der 20 natürlich in Proteinen vorkommenden Aminosäuren (siehe V1).

Sekundärstruktur-Auftreten in löslichen Proteinen

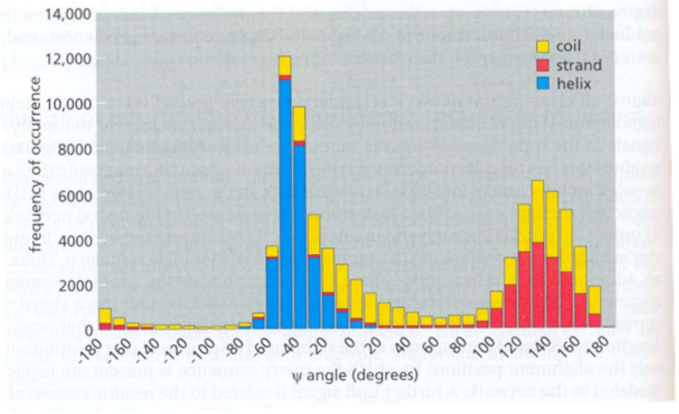


Längenverteilung von Sekundärstrukturelementen.

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

Hier gezeigt ist eine Statistik über die Längenverteilung von alpha-Helices und beta-Faltblättern in Proteinen mit bekannter Struktur. Alpha-Helices (grün) sind deutlich länger im Durchschnitt. Wir benötigen diese Statistik um geeignete Algorithmen entwickeln zu können. Wir müssen ja schließlich wissen wonach wir suchen wollen.

Rückgratwinkel in Sekundärstrukturelementen



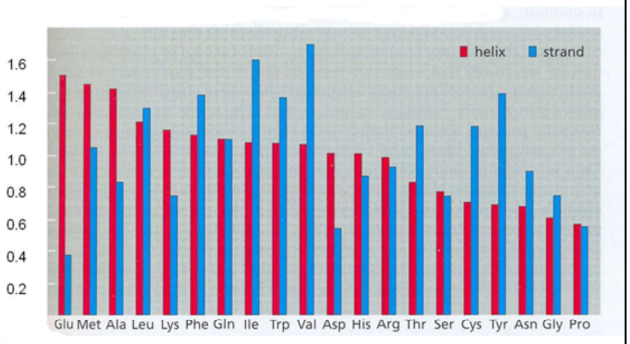
Diese Folie zeigt die Verteilung des Psi-Winkels in alpha-Helices (blau), beta-Strängen (rot) und in ungeordneten Schleifen/coils (gelb). Diese Statistik entspricht der Projektion des Ramachandran-Plots auf die Psi-Achse. Beta-Stränge haben Psi-Winkel nahe 180 Grad.

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

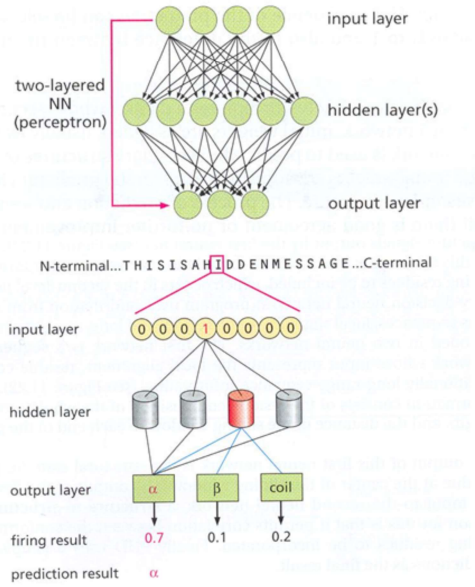


5. Vorlesung WS 2020/21 Softwarewerkzeuge 24

Chou und Fasman ordneten jeder Aminosäure eine Präferenz für alpha-Helix und beta-Strang zu. Wenn mehrere aufeinanderfolgende Aminosäuren (4 aus 6) einen Schrankenwert für alpha-Helix überschreiten, wird diese Region als alpha-helikal vorhergesagt bzw. als beta-Strang (wenn 3 aus 5 dessen Schrankenwert überschreiten).

In der Abbildung sind die Aminosäuren gemäß absteigender helikaler Präferenz angeordnet. Prolin hat für alpha und beta die niedrigste Präferenz. Überraschend finde ich die hohe helikale Präferenz von Glutamat. Alanin ist als Helix-Former bekannt. Die höchsten beta-Präferenzen haben Val und Ile. Diese liegen oft im Proteininneren.

Vorhersage mit Neuronalen Netzwerken



zweilagiges Neuronales Netzwerk

Feed-forward NN zur Vorhersage von Sekundärstrukturen

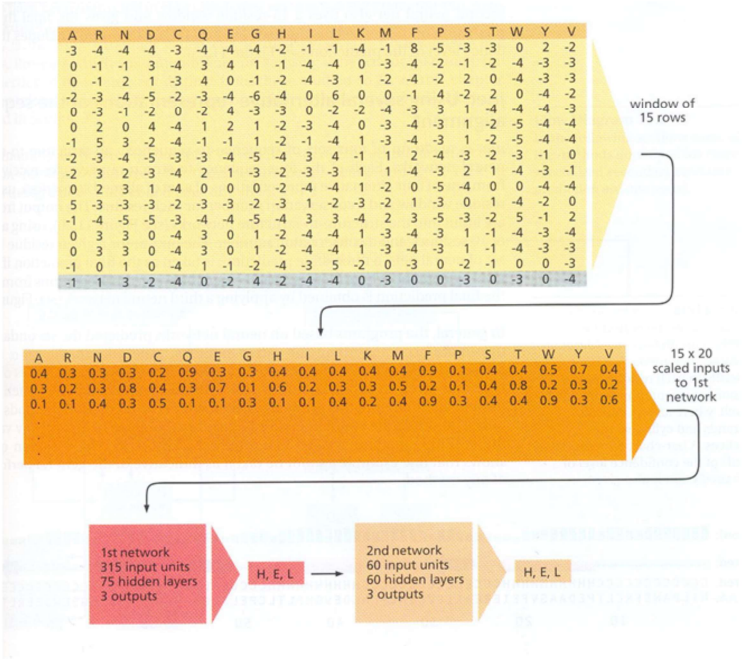
5. Vorlesung WS 2020/21 Softwarewerkzeuge 25

Wir werden heute die Methode PSIPRED vorstellen, die sogenannte neuronale Netzwerke verwendet. Hier ist zunächst einmal in der oberen Abbildung die Topologie eines einfachen Netzwerks gezeigt, das nur eine Schicht von mittleren Knoten enthält. Ein neuronales Netzwerk bildet Eingabedaten in Ausgabedaten (Vorhersagen) ab. Die Pfeile entsprechen gewichteten mathematischen Operationen. Die Gewichte werden in der Trainingsphase (wie beim HMM) trainiert.

Die untere Abbildung illustriert, wie man eine Buchstaben-Sequenz numerisch kodieren kann. Für jede der 20 Aminosäuren gibt es ein Element in einem Eingabevektor. Die Ausgabe enthält hier 3 Zustände für alpha-helikal (70% Zuversicht für die rot umkreiste Aminosäure Ile), beta-Strang (30% Zuversicht) und coil/Schleife.

PSIPRED

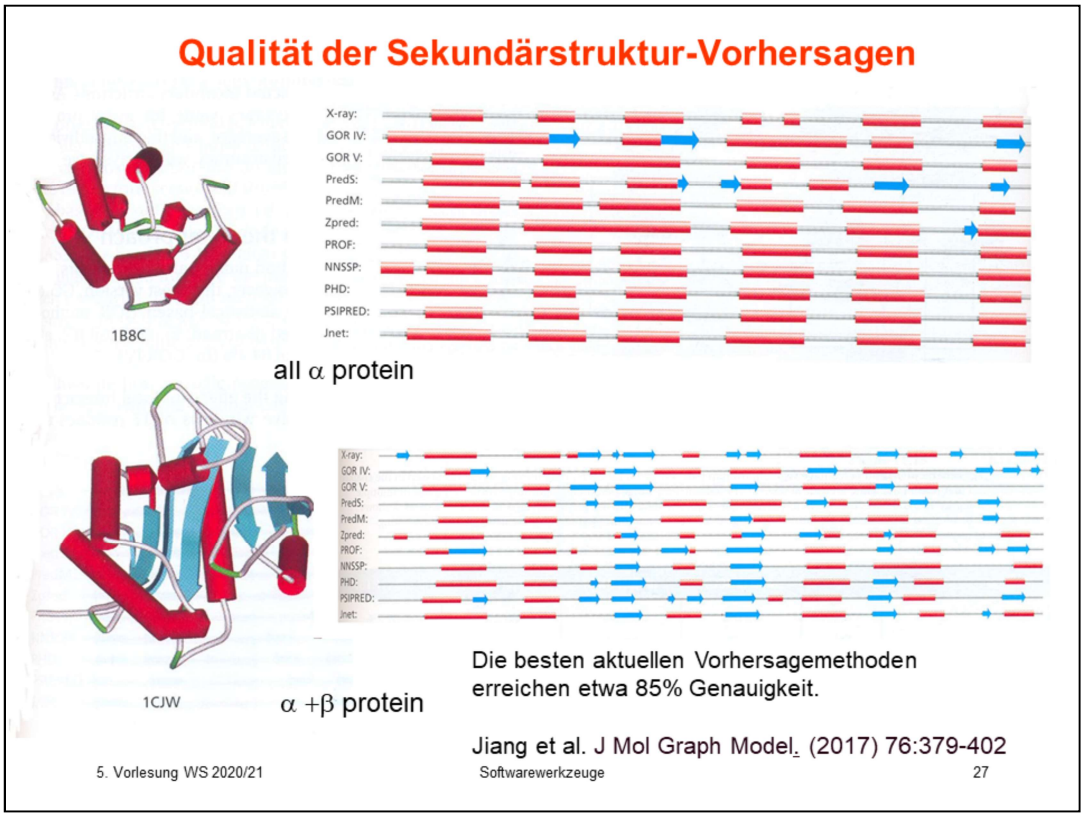
Benutze Profil aus PSIBLAST.
Skaliere Werte auf Intervall [0.0;1.0].



PSIPRED wurde 1999 von David Jones entwickelt: https://pubmed.ncbi.nlm.nih.gov/10493868/

Man verwendet ein Sequenzprofil von PSIBLAST (siehe V2) mit den log-Häufigkeiten aller 20 AS in einem Sequenzfenster der Länge 15. Diese Werte werden zunächst einmal auf das Intervall (0,1) skaliert (mittlere Tabelle). Diese Daten werden dann in ein erstes NN gefüttert. Die Ausgaben H,E,L stehen für Helix, Strand und Loop.

Daran schliesst sich ein zweites Netzwerk an, das wiederum die Vorhersagen des ersten Netzwerks für ein Sequenzfenster von nun 20 Positionen aggregiert (3 Zustände H/E/L für 20 Positionen).

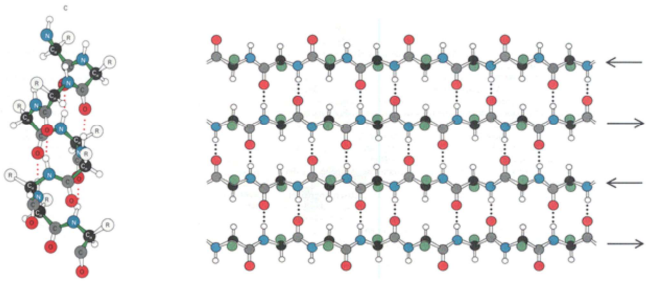


Dies sind Vorhersagen verschiedener Tools für 2 Beispiele unterschiedlicher Komplexität. Die oberste Zeile ist jeweils die Annotation aus der entsprechenden Kristallstruktur. Helices werden eigentlich von allen Methoden gut erkannt. GOR ist eine sehr alte Methode. Im unteren Beispiel ergeben sich stärkere Unterschiede zwischen verschiedenen Methoden.

Mittlerweile erreichen moderne deep learning Methoden etwa 85% Genauigkeit.

Topologie von Membranproteinen

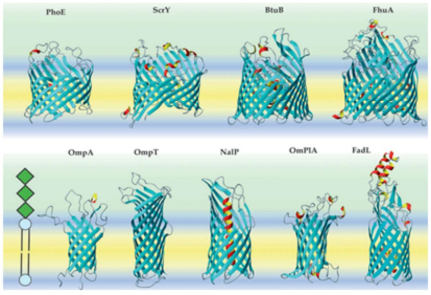
Im Inneren der Lipidschicht kann das Proteinrückgrat keine Wasserstoffbrücken-Bindungen mit den Lipiden ausbilden →
die Atome des Rückgrats müssen miteinander Wasserstoffbrückenbindungen ausbilden,
sie müssen entweder helikale oder β-Faltblattkonformation annehmen.

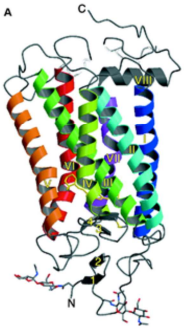


5. Vorlesung WS 2020/21 Softwarewerkzeuge 28

Transmembranproteine durchspannen die hydrophobe Phospholipid-Doppelschicht. In deren Innerem ist das Protein von hydrophoben Fettsäureketten umgeben, mit denen die Gruppen des Proteins keine H-Bindungen ausbilden können. Daher liegen auf der Proteinoberfläche dann vorwiegend hydrophobe Aminosäuren. Allerdings haben auch diese ein polares Rückgrat. Aus energetischen Gründen ist es erforderlich, dass die N-H-Gruppe und die C=O-Gruppe des Rückgrats H-Bindungen ausbilden können. Die einzige Möglichkeit dafür ist, dass sich in der Membranschicht Sekundärstrukturelemente (alpha oder beta) bilden, sodass die Gruppen des Rückgrats durch H-Bindungen abgesättigt sind. Man findet in der Membran keine Schleifen/coils, höchstens im Inneren eines Transmembranproteins.

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

Daher haben Transmembranproteine entweder eine komplette alpha-helikale Struktur wie das rechts gezeigte Rhodopsin oder eine beta-Barrel-Struktur wie die links gezeigten Proteine, die vor allem in den äußeren Membranen von grampositiven Bakterien vorkommen.

Vorhersage von Transmembranhelices

Einfaches Kriterium: Hydrophobizitäts-Skalen wie die von Kyte & Doolittle
TMHs sind meistens apolar und 12-35 Residuen lang,

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.

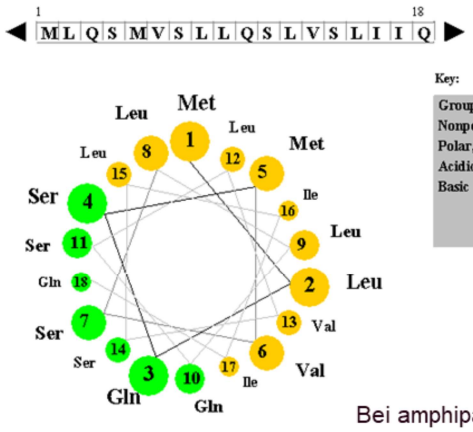
Beobachtung:
Gute Parameter sind $w = 19$ und $T > 1.6$.

Hydrophobicity Scales

	Kyte-Doolittle
Alanine	1.8
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

Ein einfaches Kriterium um die Position von Transmembranhelices vorherzusagen sind empirische Skalen wie die Kyte-Doolittle-Skala (https://pubmed.ncbi.nlm.nih.gov/7108955/).

Helikale Räder



Helikale Räder dienen zur Darstellung von Helices.

Man kann so leicht erkennen, welche Seite der Helix einen polaren bzw. hydrophoben Charakter hat.

Bei amphipathischen Helices (die flach auf der Membranoberfläche liegen) zeigt die hydrophobe Seite in die Lipidschicht der Membran und die polare Seite ins Wasser.

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

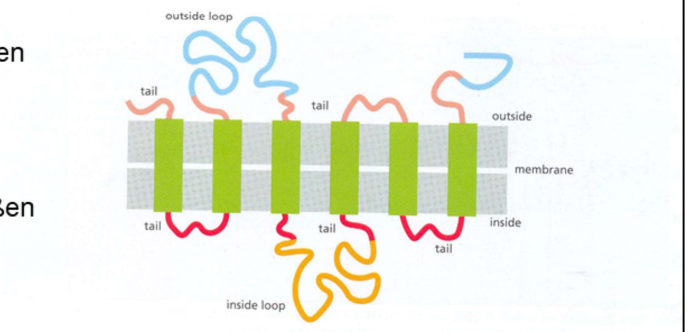
5. Vorlesung WS 2020/21 Softwarewerkzeuge 31

Membranproteine können ebenfalls sogenannte **amphipathische Helices** enthalten. Diese kann man gut durch die Darstellung auf einem **helical wheel** erkennen. Man fädelt eine Sequenz entlang einer alpha-Helix und schaut dann quasi „von vorne“ oder „von hinten“ durch die Helix hindurch.

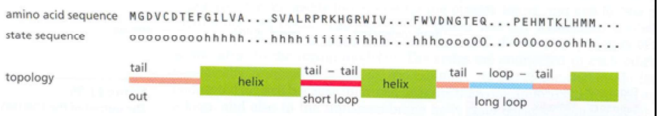
TM-Vorhersage mit Hidden Markov Modellen

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

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



HMMTOP Vorhersage



Moderne Methoden wie HMMTOP verwenden ebenfalls Hidden Markov Modelle um Transmembranhelices zu erkennen. HMMTOP verwendet die 5 aufgelisteten Zustände. Für die Sequenz des oben schematisch gezeigten Proteins mit (in Wahrheit) 6 Transmembranhelices erstellt HMMTOP die unten gezeigte Vorhersage (nur die erste Hälfte der Sequenz ist gezeigt), so dass man erkennen kann, welche Loops außen und innen von der Membran liegen.

Gunnar von Heijne postulierte die empirische „positive inside“ Regel, nach der die innenliegenden Loops mehr positive Aminosäuren enthalten als außenliegende Loops. Diese Regel funktioniert in der Praxis erstaunlich gut – die physikochemischen Hintergründe dafür sind bis heute jedoch nicht vollkommen klar.

Die Annahme von HMMTOP und ähnlichen Tools ist, dass eine Membranhelix die Membran stets komplett durchquert. Dies ist bei Membrantransportern allerdings manchmal nicht der Fall. Dort gibt es Strukturen, bei denen eine Halbhelix bis ins Innere der Membran reicht, dort eine kurze Schleife macht und dann eine kurze Halbhelix in dieselbe Richtung wie zum Beginn wieder zurückführt. HMMTOP würde solch einen Fall als eine komplette Helix erkennen, so dass die weitere Annotation des Proteins dann jeweils „auf der falschen Seite“ liegen würde.

Vergleich von 2 Proteinstrukturen: DALI (Distance-matrix Alignment)

L. Holm & C. Sander

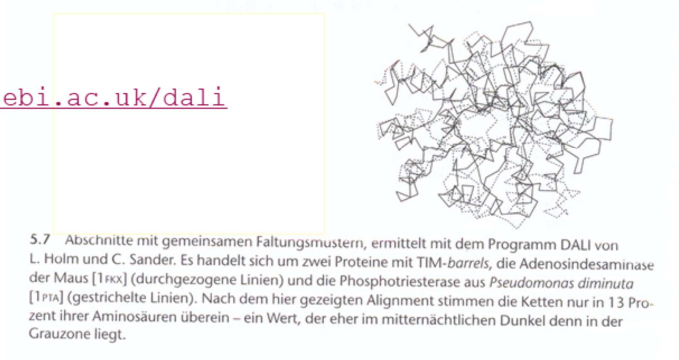
Während der Evolution eines Proteins verändert sich seine Sequenz.

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.

Im letzten Teil der heutigen Vorlesung beschäftigen wir uns mit dem Vergleich zweier Proteinstrukturen. Es kann nämlich durchaus eine strukturelle Ähnlichkeit zwischen entfernt miteinander verwandten Proteinen bestehen, zwischen denen mittlerweile keine Sequenzähnlichkeit mehr besteht.

Bedeutung von struktureller Äquivalenz

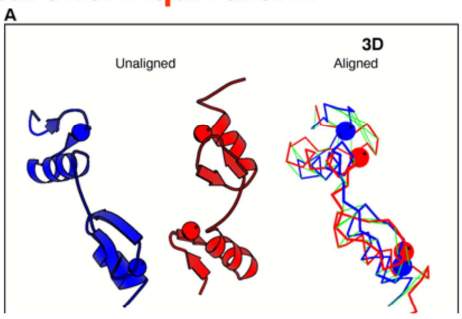
Beim Strukturvergleich sollen äquivalente Strukturblöcke zweier Proteine einander zugeordnet werden.

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).

Gezeigt sind hier die Kristallstrukturen von zwei Zinkfinger-Transkriptionsfaktoren tramtrack und MBP-1. Beide Proteine enthalten jeweils zwei Kopie einer Domäne (1 Helix und 2 Beta-Stränge). Man würde erwarten, dass die beiden Proteine strukturell perfekt aufeinander passen. Allerdings sind die beiden Domänen im roten und blauen Fall unterschiedlich gegeneinander verdreht, so dass sie bei einer Überlagerung der starren Körper (rechts, „aligned“) eine hohe strukturelle Abweichung (RMSD) aufweisen.

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)

Dies ist ein Beispiel zweier sehr ähnlicher Proteinstrukturen, die zu zwei Proteinen mit komplett unterschiedlicher Struktur gehören. papD ist ein Transportprotein in Bakterien, das andere ein Protein des Immunsystems.

Überraschende Ähnlichkeit zwischen Flavodoxin und Malat-Dehydrogenase

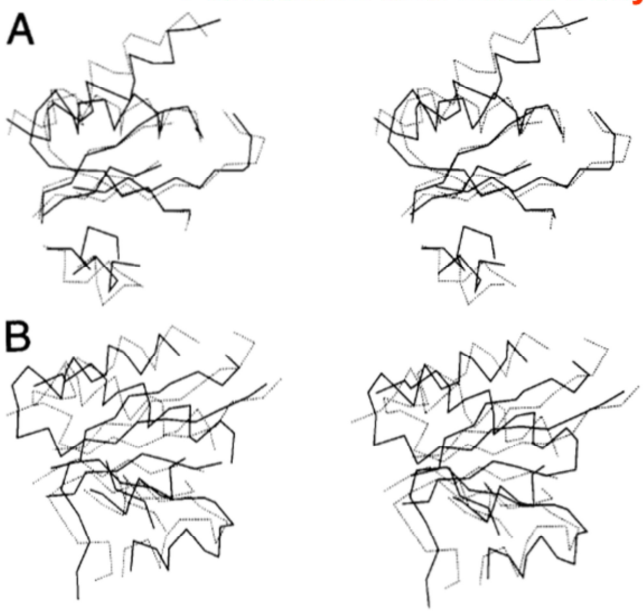


Fig. 3. Structure alignment of arabinose binding protein (ABP) with flavodoxin (FXN) and malate dehydrogenase (MDH). **A:** Structural alignment and optimal superposition of 1ABP and 4FXN. The C^{α}-rmsd after optimal superposition is 2.7 Å for 78 aligned residues. PDB residue numbers of the aligned 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)

Dies ist noch ein Beispiel großer struktureller Ähnlichkeit.

Überraschende Ähnlichkeit zwischen Tryptophansynthase und Flavocytochrom b2

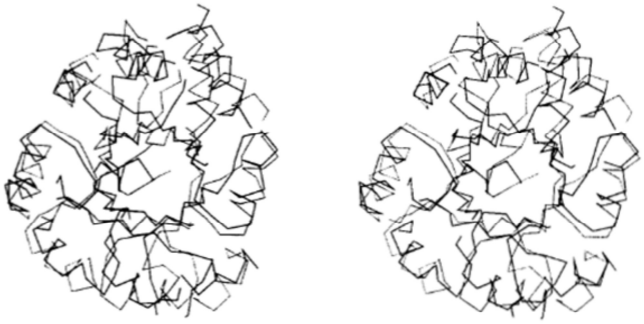


Fig. 4. Structure alignment of $(\alpha\beta)_8$ barrels. Structural alignment and optimal superposition of tryptophan synthase (1WSY) and flavocytochrome b2 (1FCB). The C^{α}-rmsd after optimal superposition is 3.1 Å for 198 aligned residues. PDB residue numbers of the aligned fragments follow. For 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)

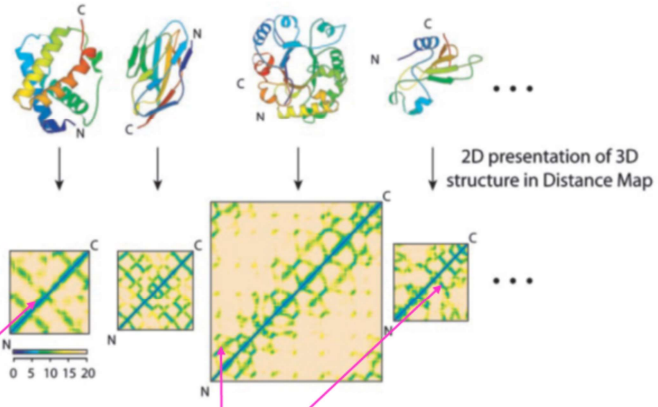
Dies ist noch ein Beispiel großer struktureller Ähnlichkeit.

Distanzmatrix für Proteinstrukturen

Distanzmatrix:

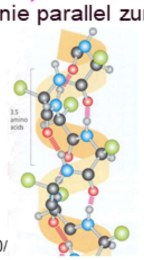
Auf beiden Achsen wird jeweils die Proteinsequenz aufgetragen.

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



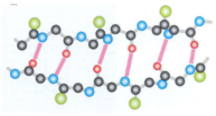
In einer **α-Helix** liegt Aminosäure *i* jeweils nah bei AS *i* + 4

→ 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)

Eine geeignete Datenstruktur für Strukturvergleiche ist die sogenannte Distanzmatrix. Dort wird auf beiden Achsen die Proteinsequenz aufgetragen. Aminosäurepaare, die voneinander einen Abstand unterhalb eines Schrankenwerts aufweisen, werden gekennzeichnet. Oft verwendet man einen Schrankenwert von 8 Angstroem.

Distanzmatrix bzw. Kontaktmatrix

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

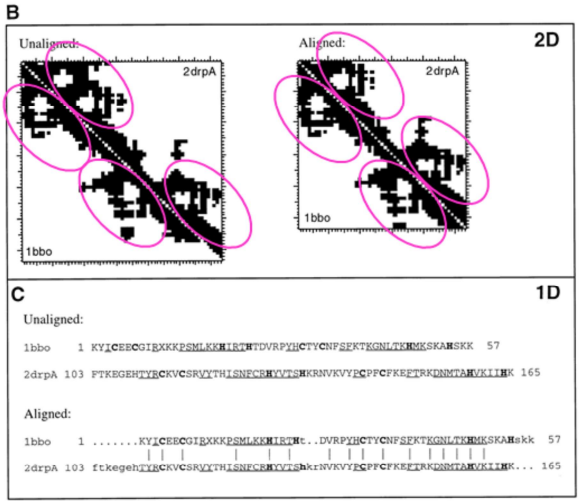
Links: ohne Alignierung, schlechte Übereinstimmung der Kontakte.

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.



Holm, Sander Science 273, 5275 (1996)

Hier sind die Kontaktmatrizen für das Beispiel der beiden Zinkfingerproteine gezeigt. In der linken Hälfte von Panel B enthält jede Diagonalmatrix entweder die Kontaktmatrix für 1bbo oder 2drpA. Mit lila Kreisen sind die kompakten Domänen umkreist. Man sieht sofort, dass die beiden Proteine nicht gut aufeinander passen. Allerdings bekommt man aus dem unten gezeigten Sequenzalignment die Idee, dass die Sequenz von 2drpA einen längeren Linker als in 1bbo enthält. Wenn man diesen auf Länge von 1bbo kürzt, passen die beiden Proteine wie rechts gezeigt nahezu perfekt aufeinander. Genau dies ist die Idee des DALI-Algorithmus.

DALI verwendet einen branch-and-bound Algorithmus

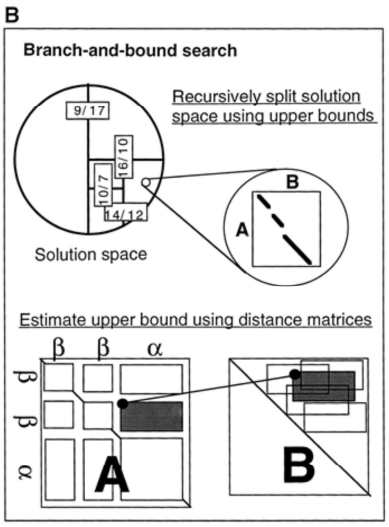
(**B**) A branch-and-bound algorithm is guaranteed to yield the global optimum but may, in the worst case, need an exponential number of steps to do so.

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

Diese Folie (Abb. und Legende) stammt aus der Originalpublikation von DALI. Die Idee ist, mit Hilfe eines branch-and-bound-Algorithmus einen jeweils optimalen Fit für kurze lokale Abschnitte der Strukturen zu identifizieren (unterer Teil der Abbildung). Wir werden die Details des Algorithmus nicht in der Vorlesung behandeln.

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.

In dieser ersten Vorlesung aus dem Bereich Proteinstruktur haben wir den prinzipiellen Aufbau von Proteinen wiederholt und dann Algorithmen zur Identifizierung von Sekundärstrukturelementen und zum Vergleich von zwei Proteinstrukturen vorgestellt. In V6 werden wir uns mit der Methode der Homologiemodellierung von Proteinstrukturen beschäftigen.