V5: Proteinstruktur: Sekundärstruktur

INHALT

- Hierarchischer Aufbau der Proteinstruktur
- **Ramachandran-Plot**
- Vorhersage von Sekundärstrukturelementen aus der Sequenz
- Membranproteine
- 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

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

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.

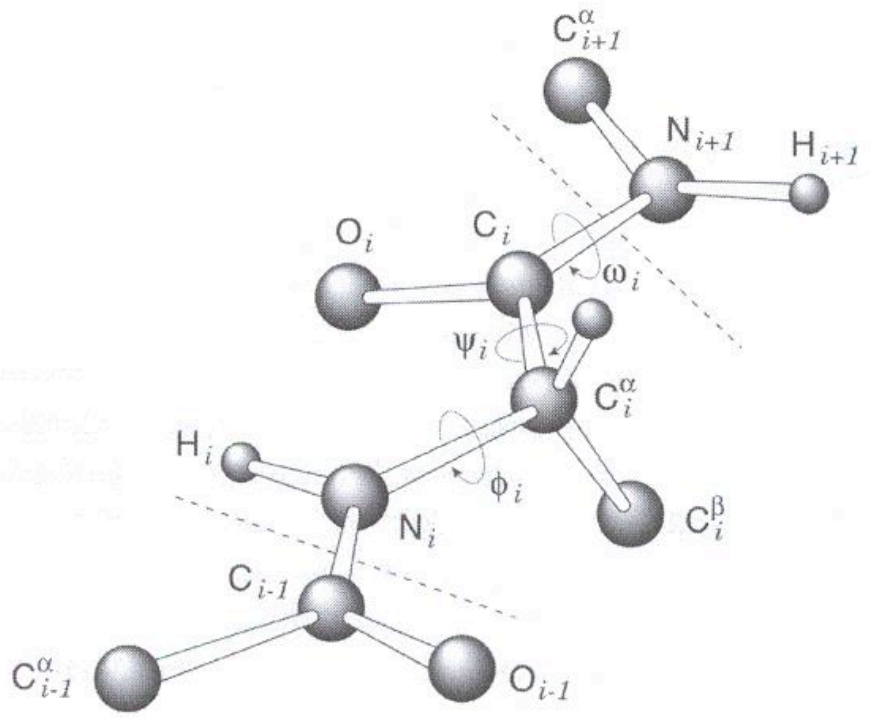
Membranproteine: sind im **Transmembranbereich** außen hydrophober als innen.

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

Diederwinkel des Proteinrückgrats

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

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



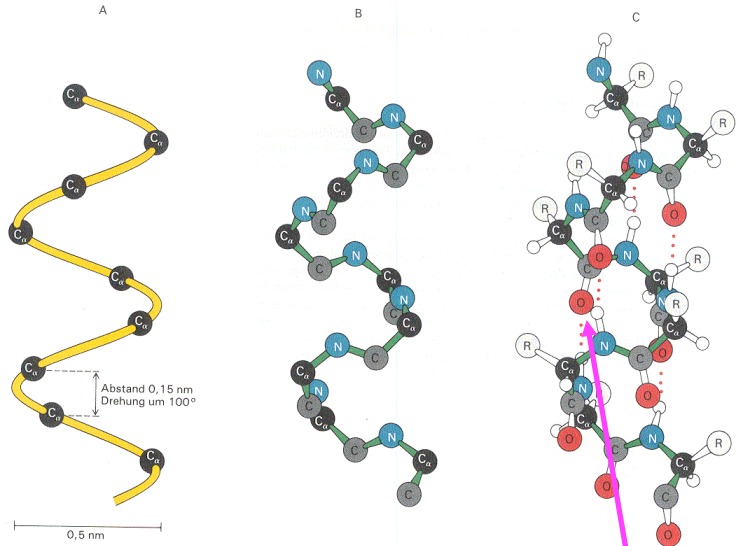
Definition der Konformationswinkel im Polypeptidrückgrat.

Lesk-Buch

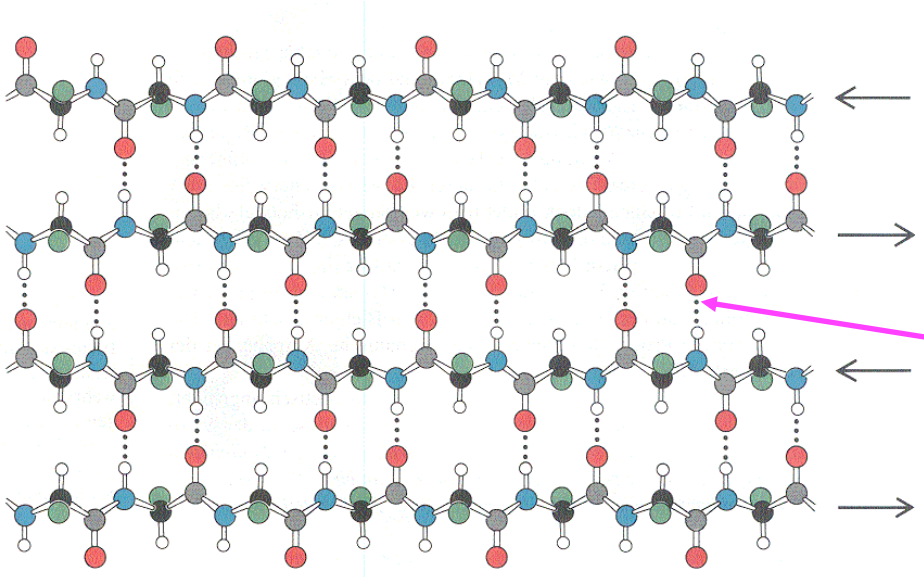
Sekundärstrukturelemente

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.

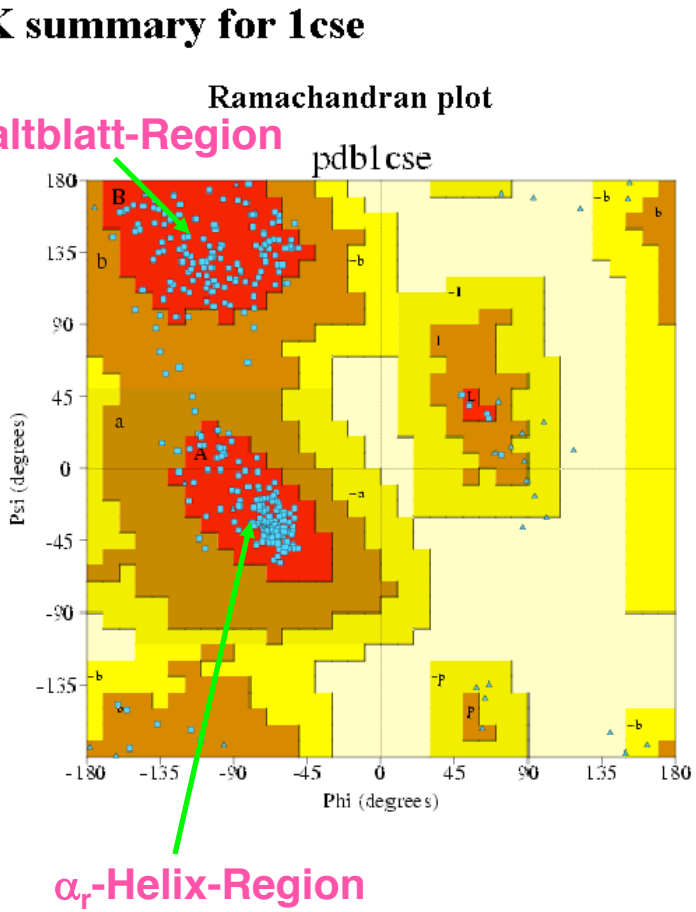
Stabilität und Faltung von Proteinen

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.

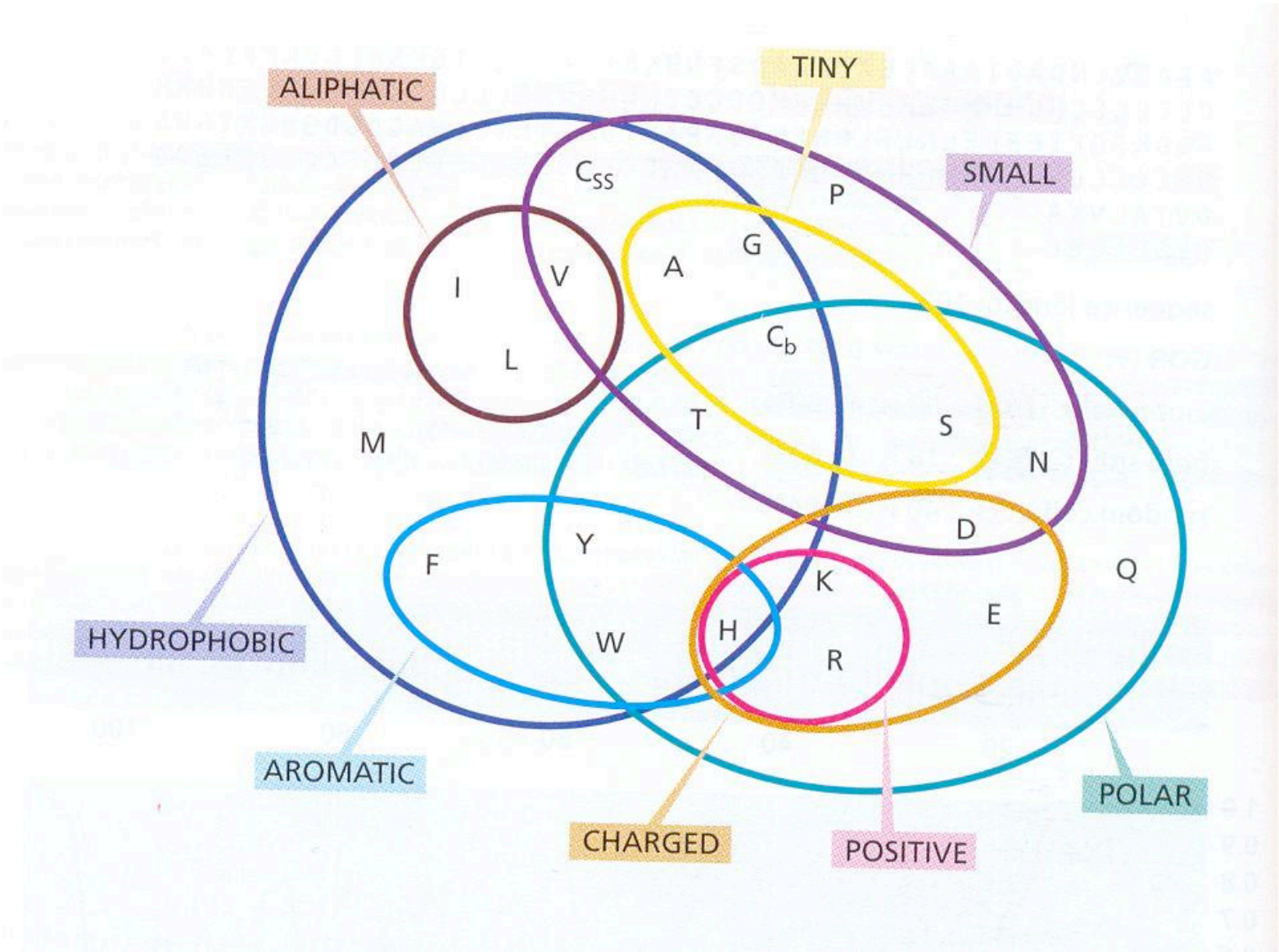
Die einzige Residue, die außerhalb der erlaubten Bereich liegt, also alle möglichen Torsionswinkel annehmen kann, ist **Glycin**.

Grund: es hat keine Seitenkette.

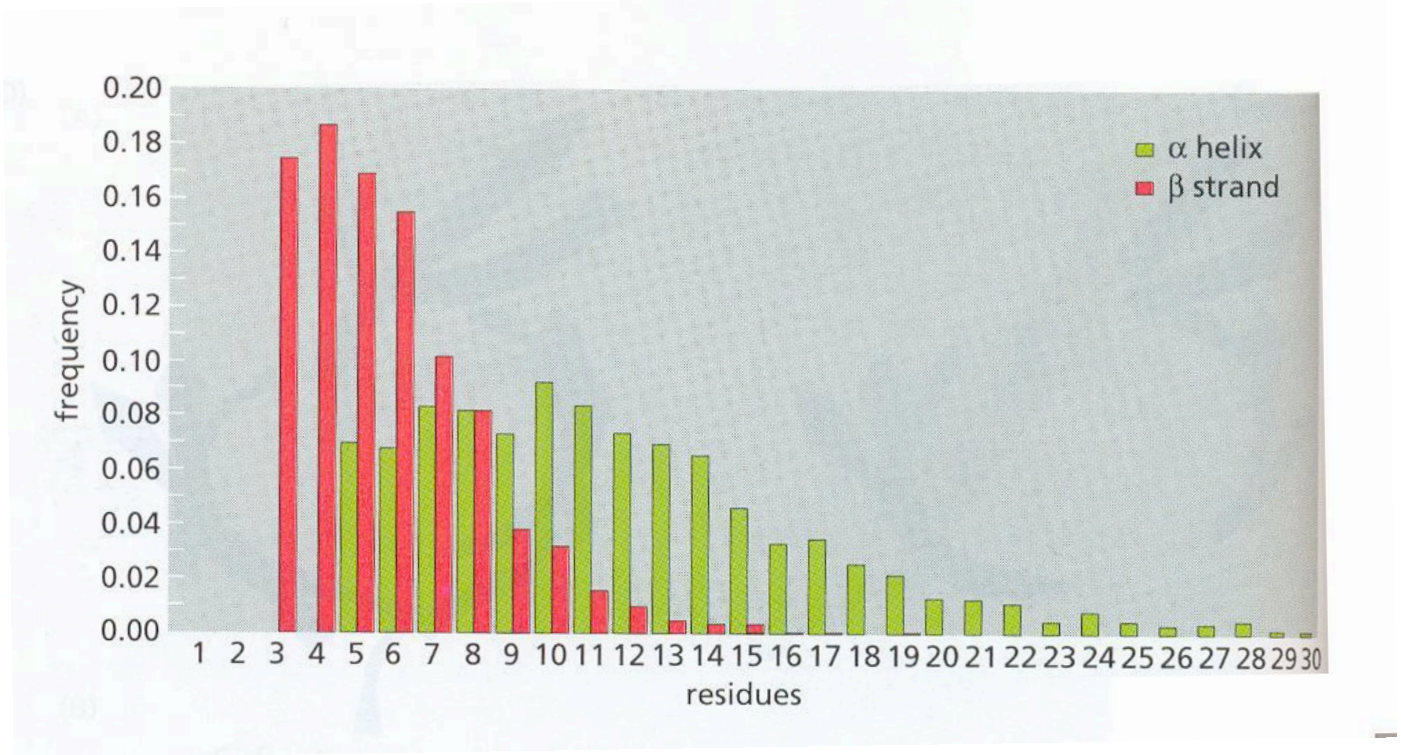


(rechtsgängige Helix)

Die 20 natürlichen Aminosäuren



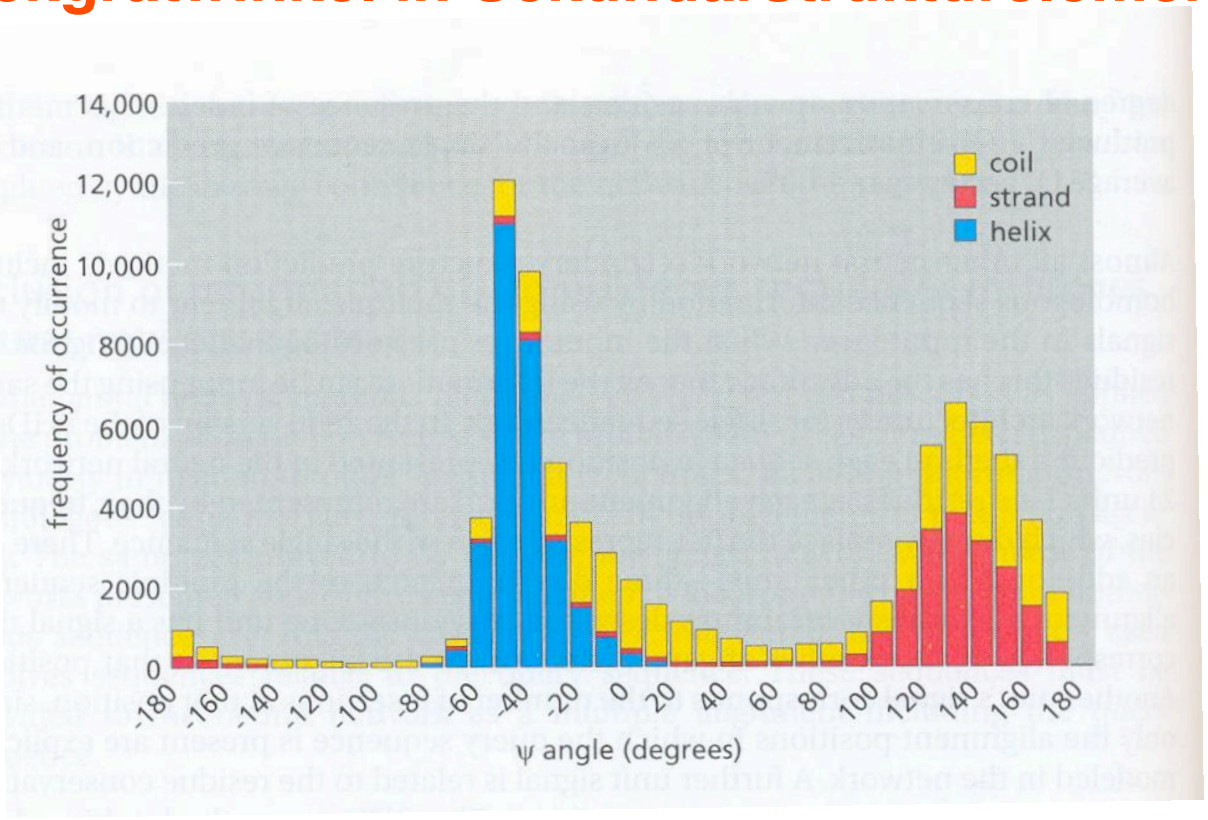
Sekundärstruktur-Auftreten in löslichen Proteinen



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

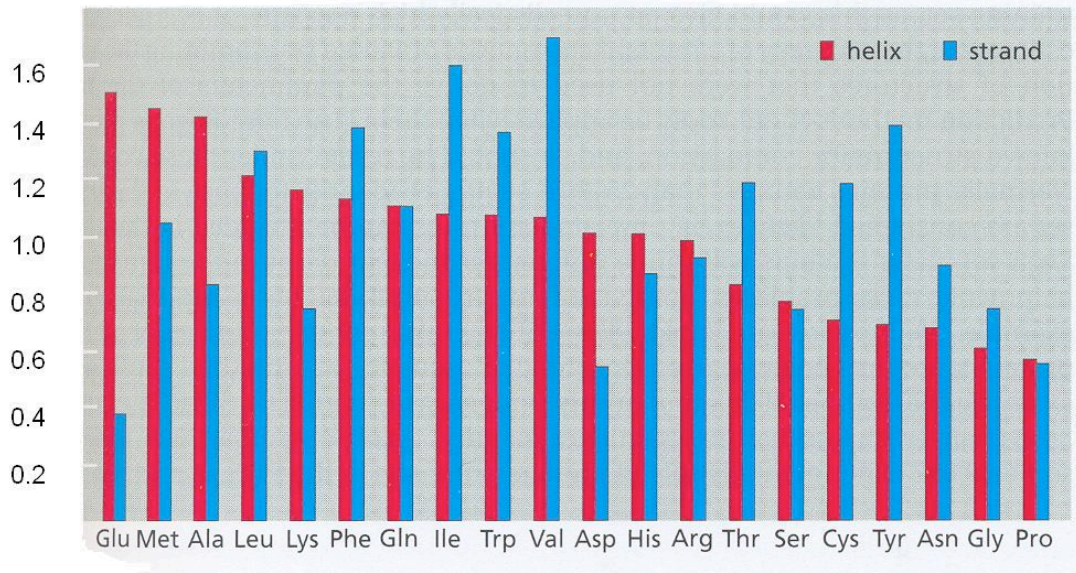
f : schwache Tendenz

B : starker (Unter-) Brecher

b : schwacher (Unter-) Brecher

I : indifferent

Prolin: stärkster Helixbrecher sowie für Betastränge

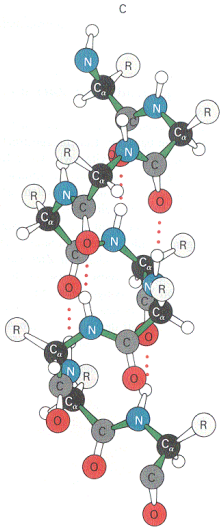


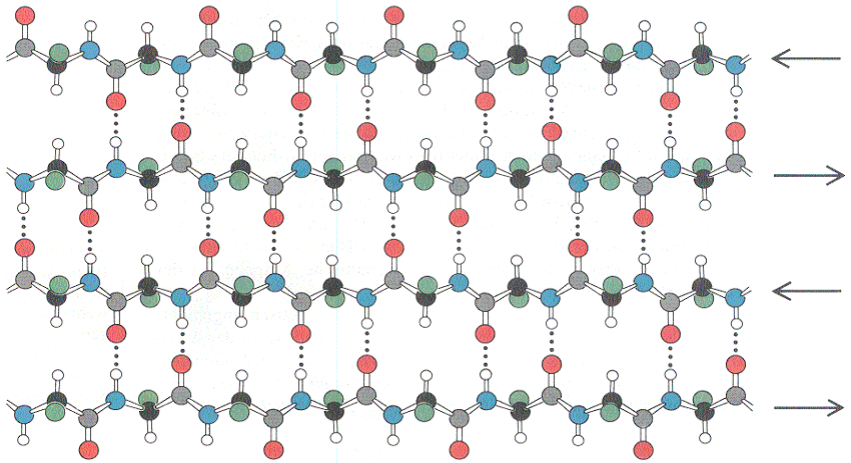
Topologie von Membranproteinen

Im Inneren der Lipidschicht kann das Proteinrückgrat keine Wasserstoffbrücken-Bindungen mit den Lipiden ausbilden →

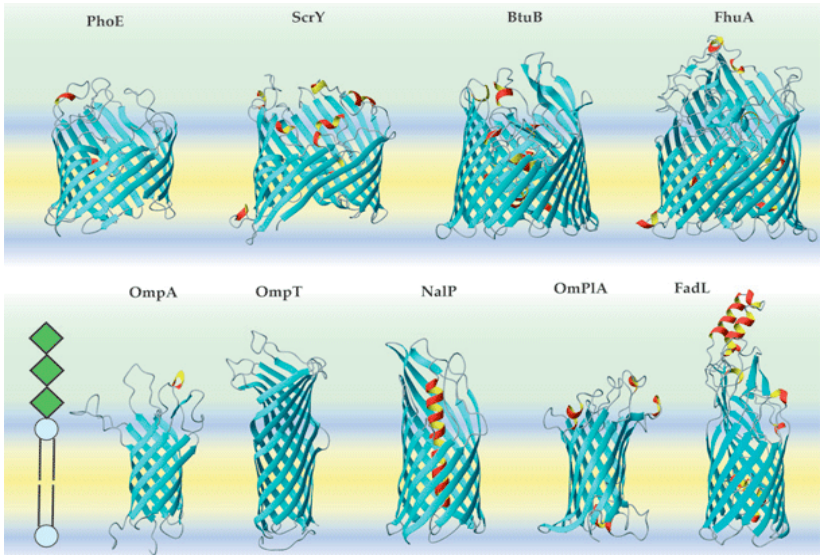
die Atome des Rückgrats müssen miteinander Wasserstoffbrückenbindungen ausbilden,

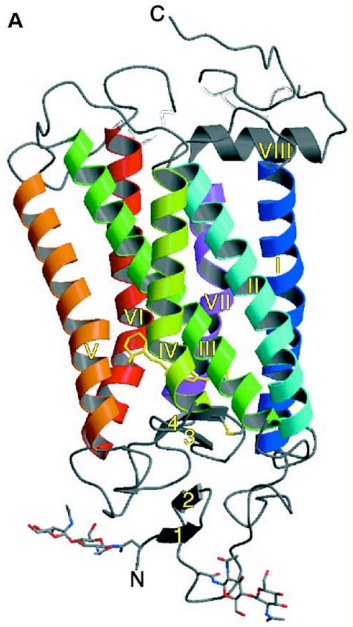
sie müssen entweder helikale oder β-Faltblattkonformation annehmen.





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,

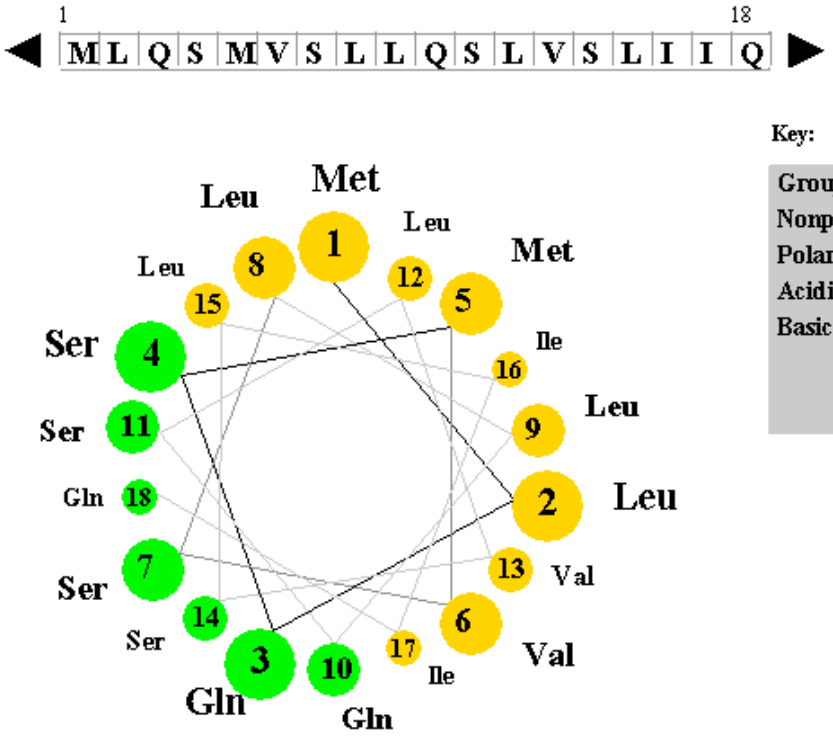
2. Globuläre (d.h. kompakte oder kugelförmige) Regionen zwischen den TMHs sind kürzer als 60 Residuen,

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	

DALI (Distance-matrix Alignment)

L. Holm & C. Sander

Während der Evolution eines Proteins verändert sich seine Struktur.

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

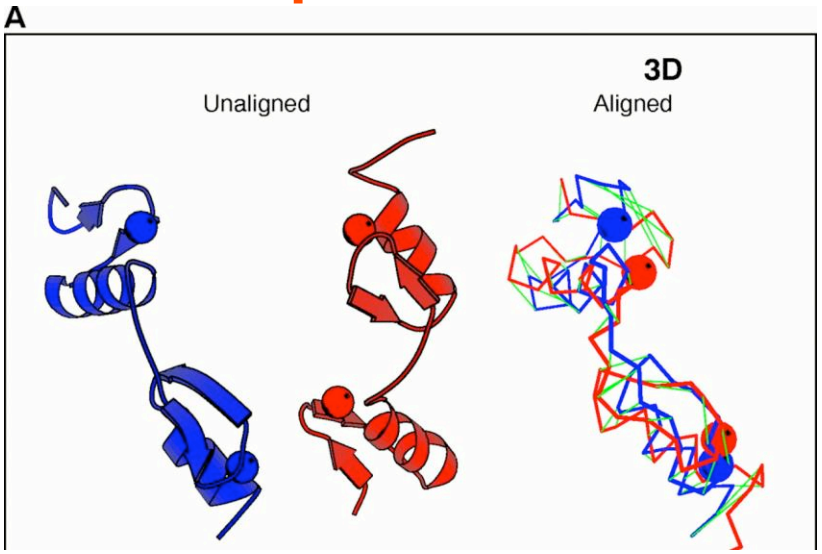
Beim Strukturvergleich sollen äquivalente Strukturblöcke zweier Proteine einander zugeordnet werden.

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).

Distanzmatrix bzw. Kontaktmatrix

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

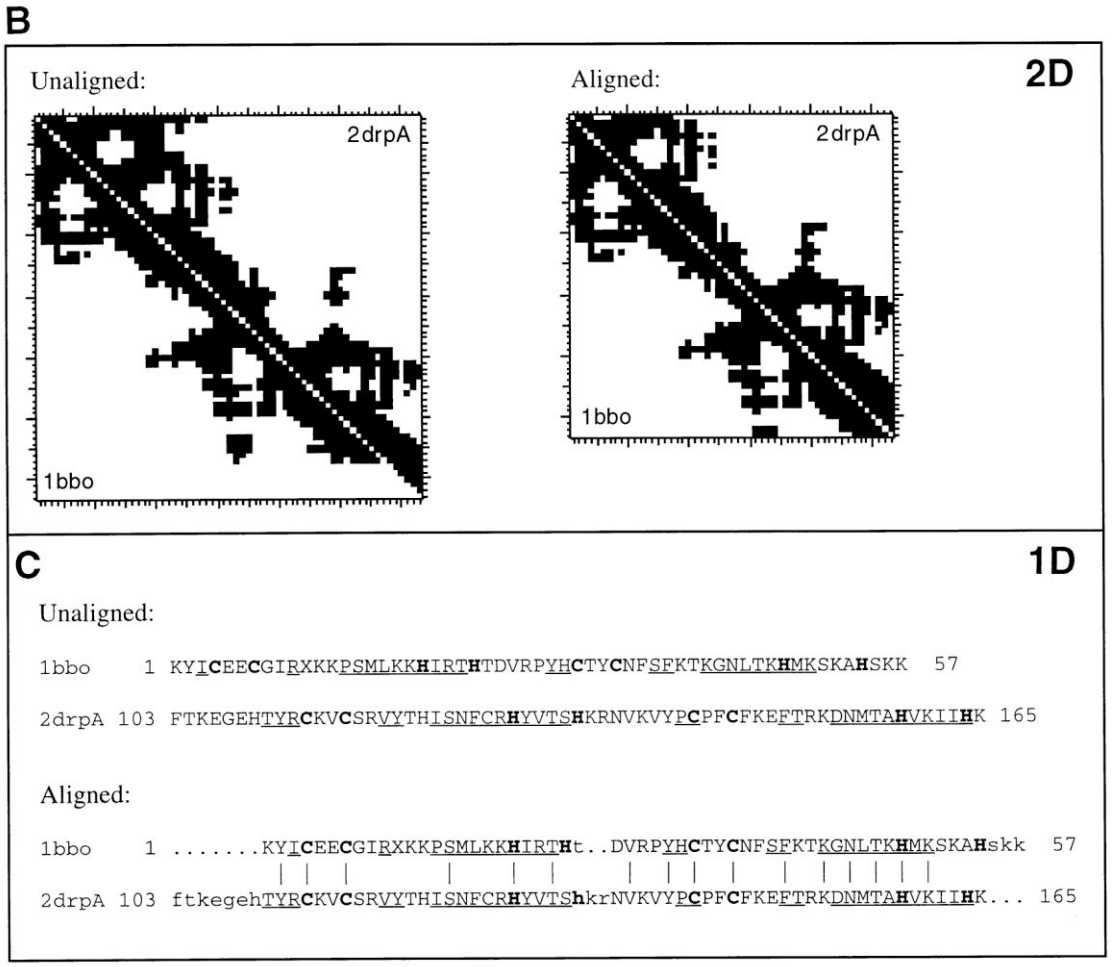
Links: ohne Alignierung, schlechte Übereinstimmung der Kontakte.

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)

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.

V6 Homologie-basierte Proteinmodellierung

- **Idee:** Sequenzähnlichkeit führt oft zur Ähnlichkeit der 3D-Struktur

  Twighlight-Zone

- **Lernziele:**

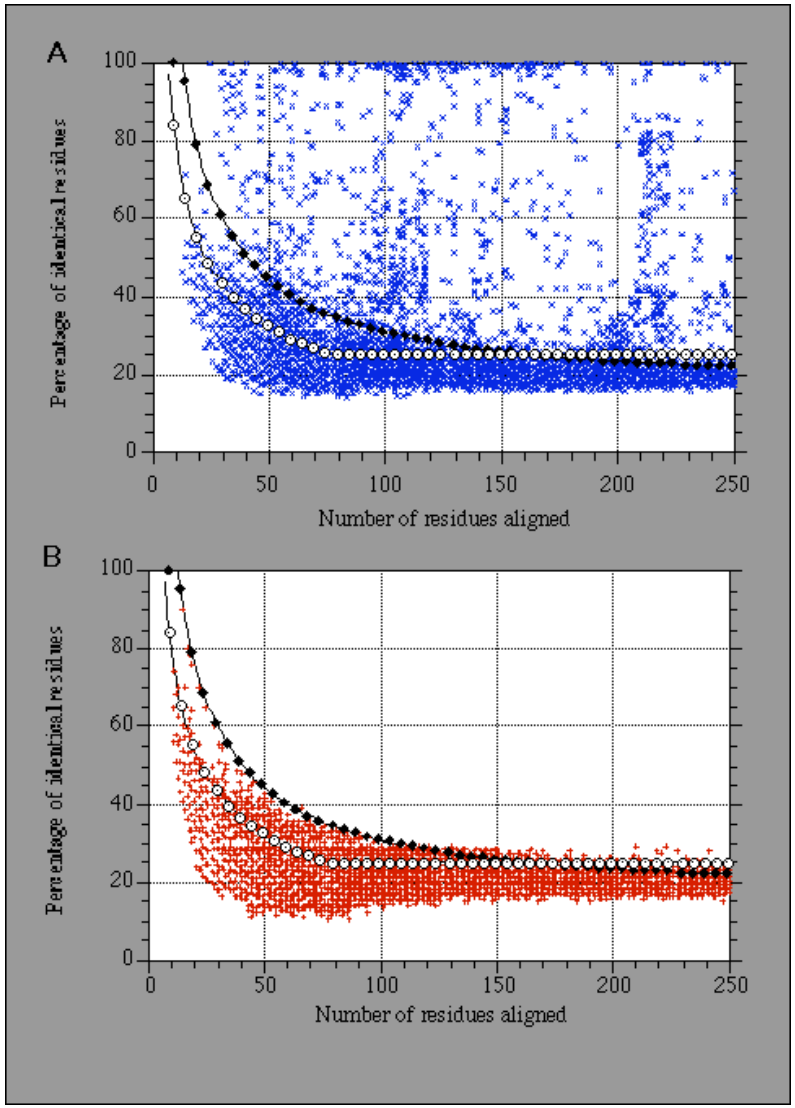
  (1) verstehe, wie Threading- und Homologiemodelle konstruiert werden

  (2) wie gut (genau) sind Homologiemodelle?

1 Twilight Zone

Die schwarzen Diamant-Symbole kennzeichnen eine Kurve, die als „**Twilight Zone**“ bezeichnet wird.

Paare von Proteinsequenzen mit größerer Sequenz-Identität als die Kurve haben mit Sicherheit eine ähnliche Struktur.



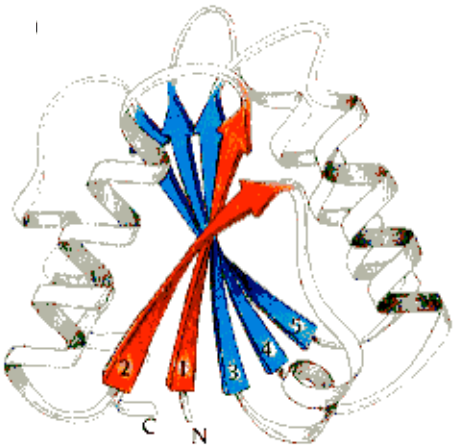
A „true positives“: Proteinpaare mit ähnlicher Struktur liegen sowohl oberhalb und unterhalb der Kurve, können also hohe oder niedrige Sequenzidentität haben.

B: „false positives“: Strukturen, die keine bzw. wenig Übereinstimmung aufweisen, liegen stets unter der Kurve.

Rost, Prot. Eng. 12, 85 (1999)

2 Methode zur Fold-Erkennung: Threading

- Gegeben:
  - Sequenz: IVACIVSTEYDVMKAAR…
  - Ein Datenbank von möglichen Proteinarchitekturen (“**folds**”)



- Naive Idee: Bilde die Sequenz auf jeden fold ab

- Starte dabei bei jeder möglichen Position

- Bestimme anhand einer energetischen Bewertungsfunktion, welcher Fold am besten zu dieser Sequenz passt.

3 Sequenz-Profil

Profil: Sequenzpositionsspezifische Bewertungsmatrix *M(p,a)* mit 21 Spalten und *N* Reihen.

- Reihe *p* entspricht einer bestimmten Position in den N_R alignierten Inputsequenzen.

- Die ersten 20 Spalten enthalten jeweils die Bewertung dafür, an dieser Position eine der 20 Aminosäuren zu finden.

Eine Extraspalte enthält einen Bestrafungsterm für Insertionen oder Deletionen.

Frequenz *W(p,b)* für das Auftreten der Aminosäure *b* an Position *p*:

$$W(p,b) = c \log (n(b,p) / N_R) \text{ oder } n(b,p) / N_R$$

n(b,p) : beobachtete Häufigkeit der Aminosäure *b* an Position *p* in den N_R Inputsequenzen;

setze außerdem *n(b,p)* = 1 für jede Aminosäure, die nie in *p* auftritt.

Berechne *M(p,a)* aus der Frequenz *W(p,b)* und einer Austauschmatrix *Y(a,b)* (PAM/BLOSUM)

$$M(p,a) = \Sigma_{b=1}^{20} W(p,b) \times Y(a,b),$$

Gribskov, PNAS 84, 4355 (1987)

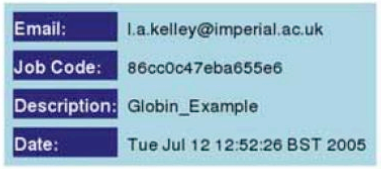
Gribskov, PNAS 84, 4355 (1987)

POS	PROBE	CONSENSUS	PROFILE	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
1	E G V L	V		3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
2	L L S P	L		2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
3	V V V V	V		2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9
4	K E A T	A		6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4	9
5	A P L P	P		6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9
6	G G G G	G		7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9
7	S S Q E	D		4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9
8	S S T P	S		4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9
9	V L V A	V		5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9
10	K R R S	R		0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	3	10	5	1	-2	7	-5	9
11	M L I I	I		0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9
12	S S T S	S		4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9
13	C C C C	C		3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9
14	K S Q R	K		1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	4	1	-2	2	-5	9
15	A A G S	A		10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9
16	T S D S	S		4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9
17	G G S Q	G		5	1	6	5	-6	9	1	-2	1	-3	-2	4	3	4	0	6	3	0	-6	-6	9
18	Y F L S	F		-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9
19	T T R L	T		1	-2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	-2	9
20	F F . L	F		-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4
21	S S . D	S		3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	-1	-2	-3	4
22	S . . S	S		2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	0	1	8	2	0	1	-2	4
23	. . . G	G		2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
24	. . . D	D		1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4
25	. . . G	G		2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
26	. A G N	A		6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4
27	Y N Y T	Y		0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4
28	E D D Y	D		2	-2	9	8	-3	3	4	-1	1	-3	-2	5	-1	4	-1	1	1	-1	-6	0	9
29	L M A L	L		3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9
30	Y N A W	N		4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9
.	.													.										
.	.													.										
.	.													.										
48	S G N S	S		4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9
49	S S N Y	S		2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9

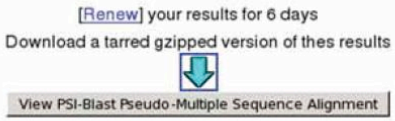
Berücksichtige, dass aus den beobachteten Sequenzen durch Mutation alle 20 AS entstehen könnten. Die Häufigkeit davon wird durch die Austausch-Matrix *Y(a,b)* ausgedrückt.

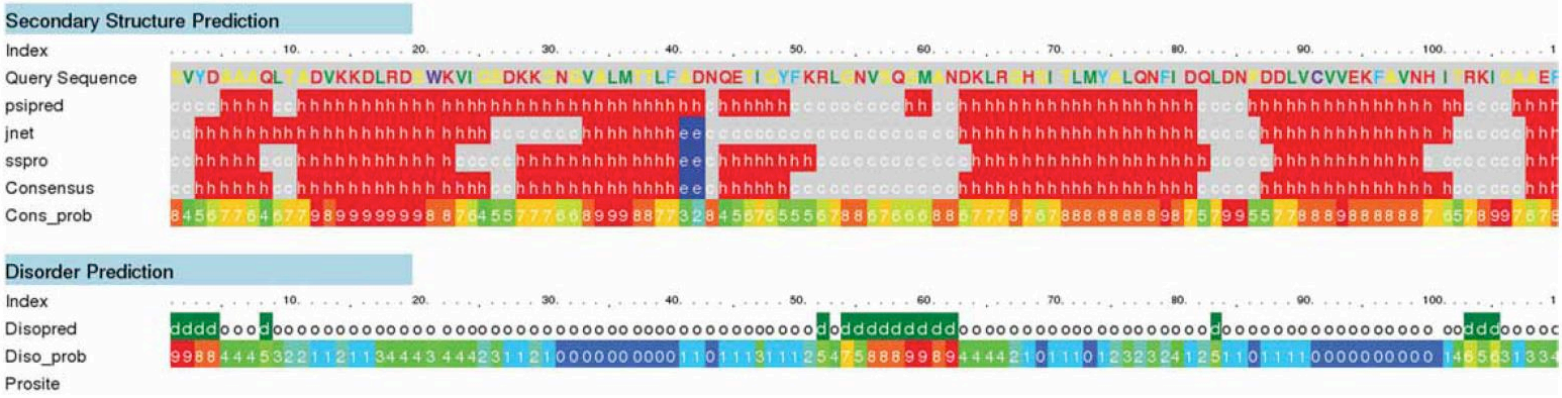
4 Methode zur Fold-Erkennung: Phyre2 webserver

- Webserver verwendet repräsentative Bibliothek für bekannte folds
- Lese Eingabesequenz mit unbekannter Struktur
- 5 Iterationen mit PsiBlast; finde nah und fern verwandte Sequenzen (richtiges MSA zu aufwändig)
- Berechne “Profil” aus den Sequenzen



- Sekundärstrukturvorhersage mit Psi-Pred, SSPro, Jnet, bilde Konsensus



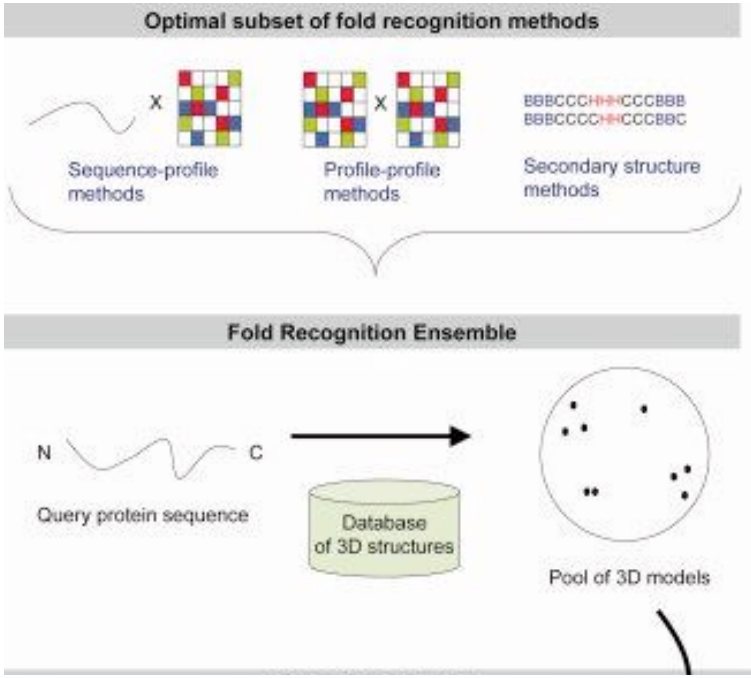


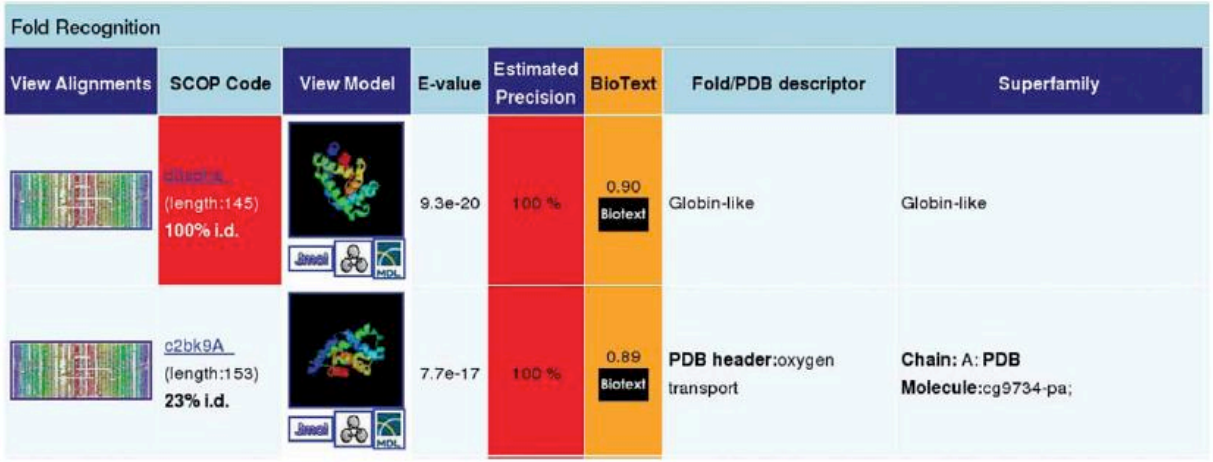
+ Vorhersage ungeordneter Regionen

Kelley, Nature Protocols 4, 363 (2009)

Methode zur Fold-Erkennung: Phyre2 webserver

- Profile-Profile Alignment zwischen Profil für Eingabesequenz und Profilen für Strukturfolds
- Berücksichtige auch, wie gut die vorhergesagte Sekundärstruktur zu jeder 3D-Strukturvorlage passt
- Berechne Scores für Passung zu allen 3D-Strukturen in der “fold library”
- Konstruiere komplette Strukturen für die 10 besten Scores
- Ergibt manchmal sehr gute Strukturmodelle bei 15-25% Sequenz-Identität.





Bennet-Lovsey, Proteins 70, 611 (2008)

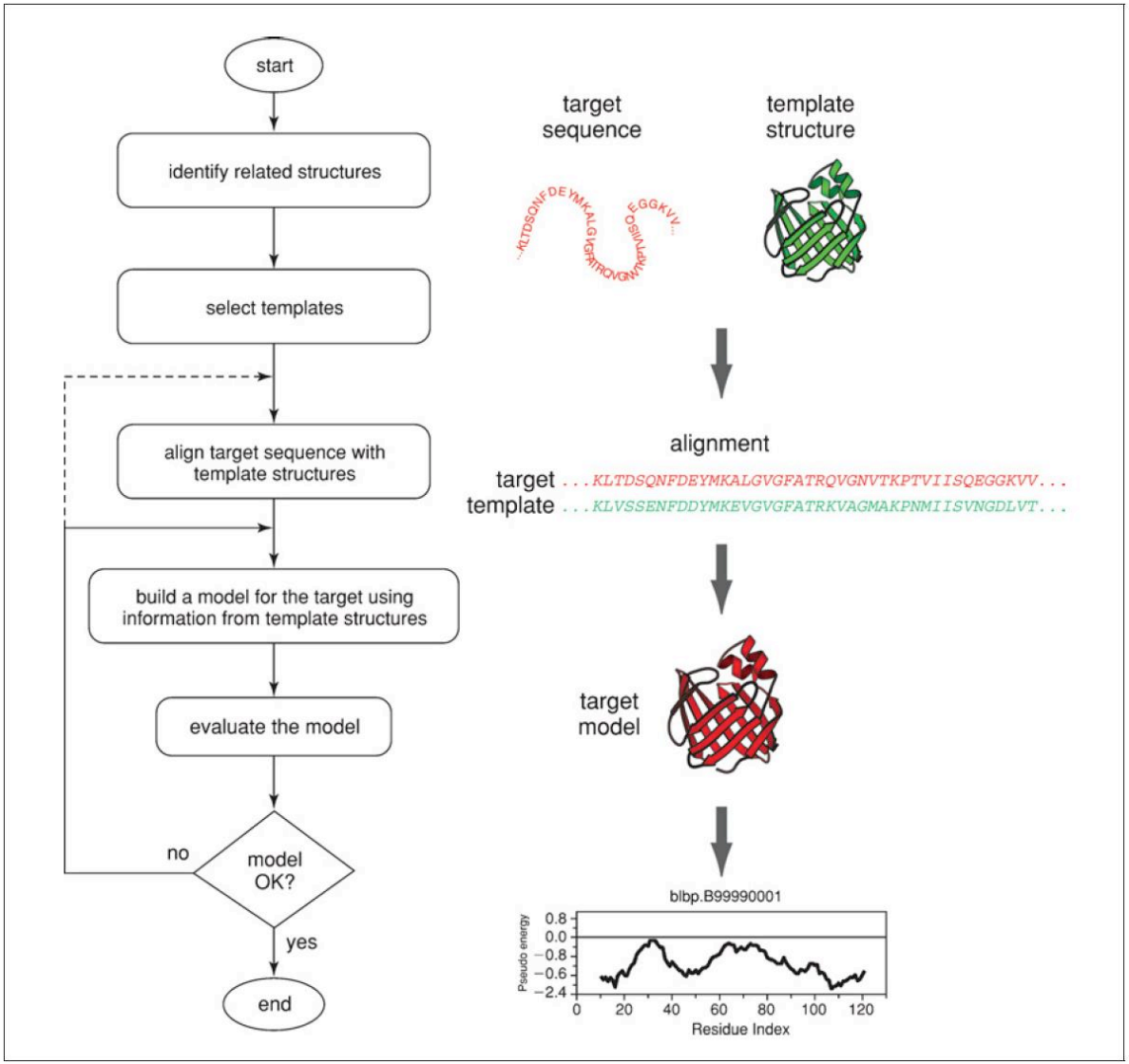
5 Homologie-basierte Proteinmodellierung (SwissModel)

- **Methode:** Ebenfalls wissensbasierter Ansatz.
- **Erfordernis: Mindestens** 1 bekannte 3D-Struktur eines verwandten Proteins,
- **Prozedur:**
- finde Proteine bekannter Struktur, die zu Inputsequenz verwandt sind.
- Erzeugung eines multiplen Sequenzalignments mit der Zielsequenz.
- Generierung eines **Frameworks** für die neue Sequenz.
- Konstruiere fehlende **Loops**.
- Vervollständige und korrigiere das **Proteinrückgrat**.
- Korrigiere die **Seitenketten**.
- Überprüfe die **Qualität** der modellierten Struktur und deren Packung.
- Strukturverfeinerung durch Energieminimierung und Moleküldynamik.

Homologie-basierte Proteinmodellierung (Modeller)

Andrej Sali, UCSF

http://salilab.org



Eswar, Curr. Protocols in Bioinf. (2006)

Konstruktion fehlender Loops

Konformationen für strukturell abweichende Loops zu konstruieren, ist ein ernstes Problem bei der vergleichende Modellierung. Seine Lösung ist (noch) offen.

Dies gilt nicht nur für lange Loops, in denen zahlreiche Mutationen auftraten, sondern auch für kurze Loops im Fall von Insertionen und Deletionen.

Sobald das Alignment von Zielsequenz und der Vorlagesequenz vorliegt, sollte man überprüfen, ob die eingefügten Gaps außerhalb von Sekundärstrukturelementen in der 3D-Struktur der Vorlage liegen.

Ein paar Regeln:

- bei sehr kurzen Loops können wir Daten über beta-turns verwenden

Konstruktion fehlender Loops

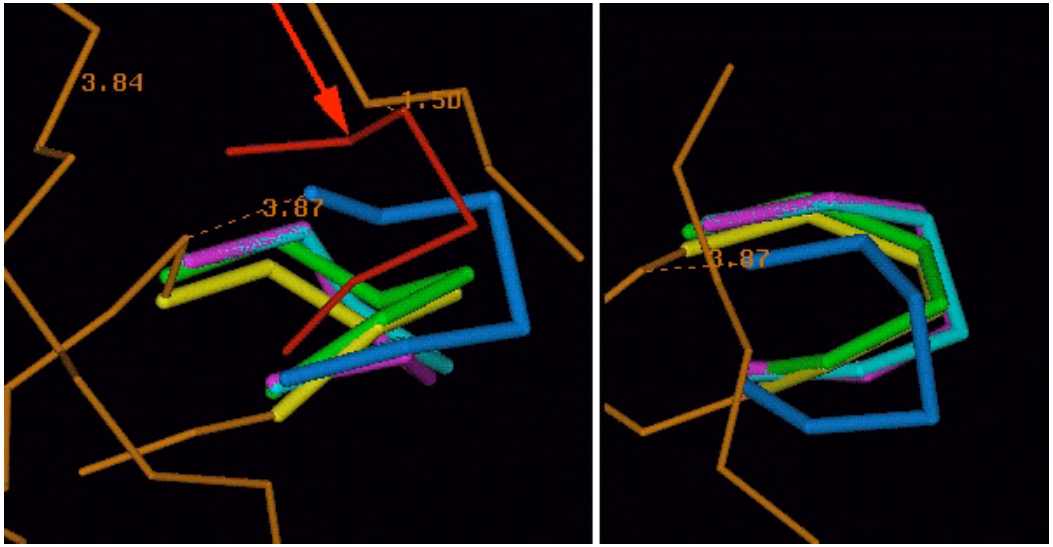
Basierend auf den Verankerungen der Loops

(a) wird entweder eine Datenbank bekannter Loopfragmente in der PDB-Datenbank durchsucht.

Für den neuen Loop verwendet man dann entweder das am besten passende Fragment oder ein Framework aus den 5 besten Fragmenten.

(b) oder es wird der Torsionsraum der Loopresiduen durchsucht

- 7 erlaubte Kombinationen der Φ-Ψ Winkel
- benötigter Raum für den gesamten Loop



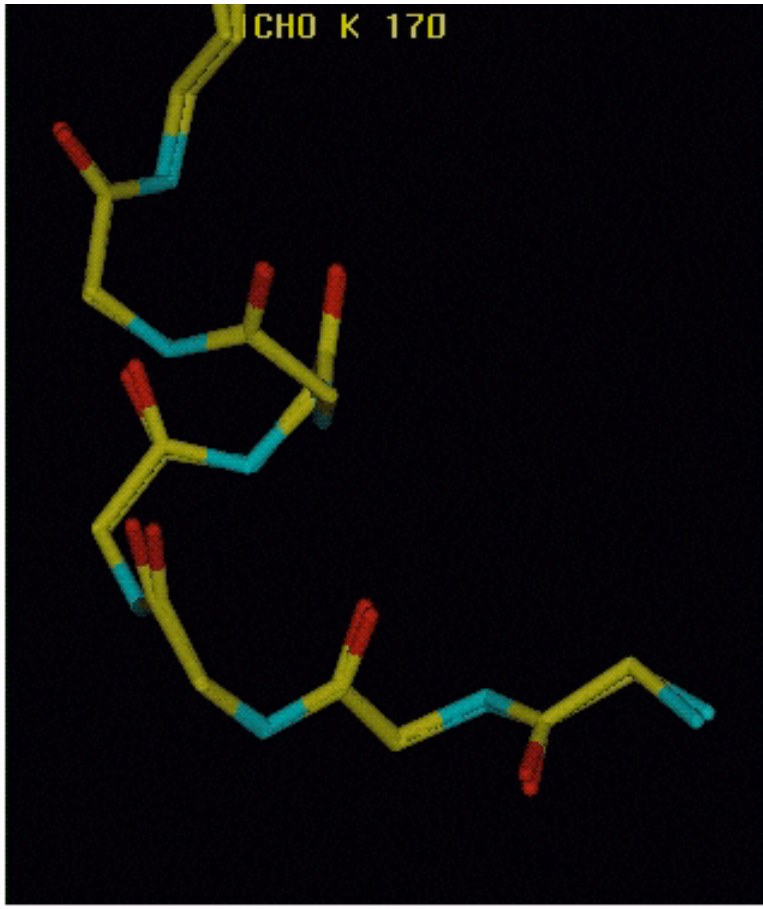
www.expasy.org/swissmodel/
SWISS-MODEL.html

Rekonstruktion von fehlendem Proteinrückgrat

Das Rückgrat wird auf der Grundlage von C_α -Positionen konstruiert.

- 7 Kombinationen der Φ-Ψ Winkel sind erlaubt.
- Durchsuche Datenbank für Backbone-Fragmente mit Fenster aus 5 Residuen,

Verwende die Koordinaten der 3 zentralen Residuen des am besten passenden Fragments.



www.expasy.org/swissmodel/SWISS-MODEL.html

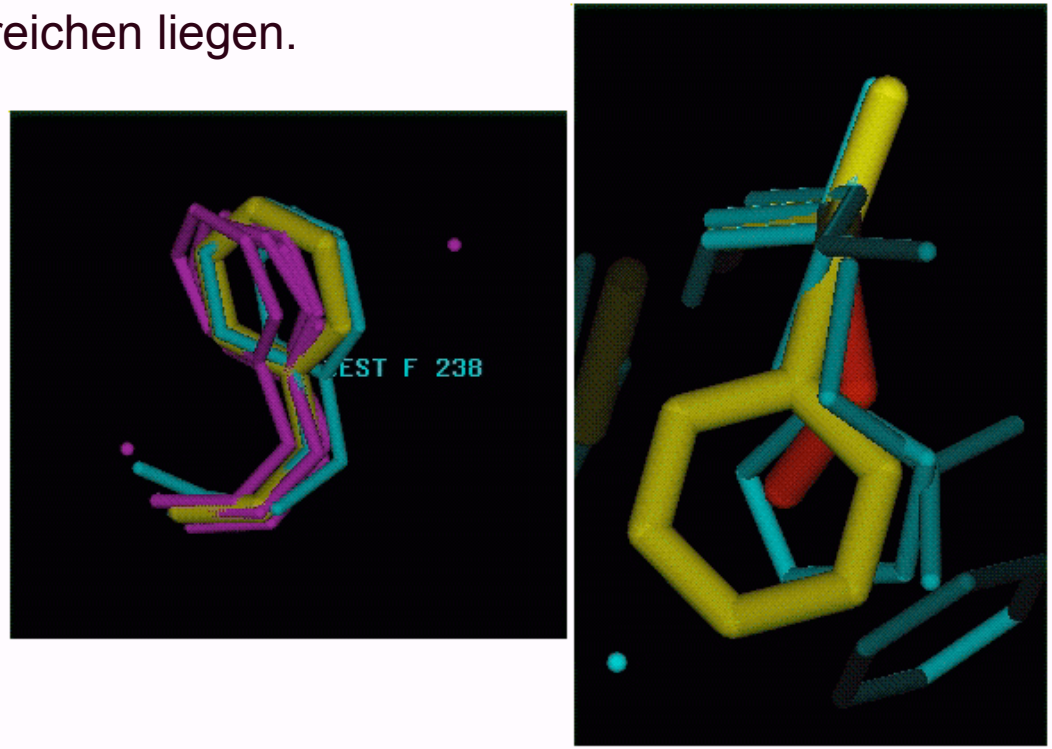
Konstruktion unvollständiger/fehlender Seitenketten

Ponder & Richards (1987): einige Aminosäuren bevorzugen bestimmte Winkelbereiche für ihre Seitenkettenwinkel → Rotamerbibliotheken.

Verwende Bibliothek erlaubter Seitenketten-Rotamere geordnet nach der Häufigkeit des Auftretens in der PDB-Datenbank.

- Erst werden verdrehte (aber komplette) Seitenketten korrigiert.
- fehlende Seitenketten werden aus der Rotamer-Bibliothek ergänzt.

  Teste dabei, ob van-der-Waals Überlapps auftreten und ob die Torsisonswinkel in erlaubten Bereichen liegen.



www.expasy.org/swissmodel/SWISS-MODEL.html

Rotamer-Bibliotheken: günstige Diederwinkel

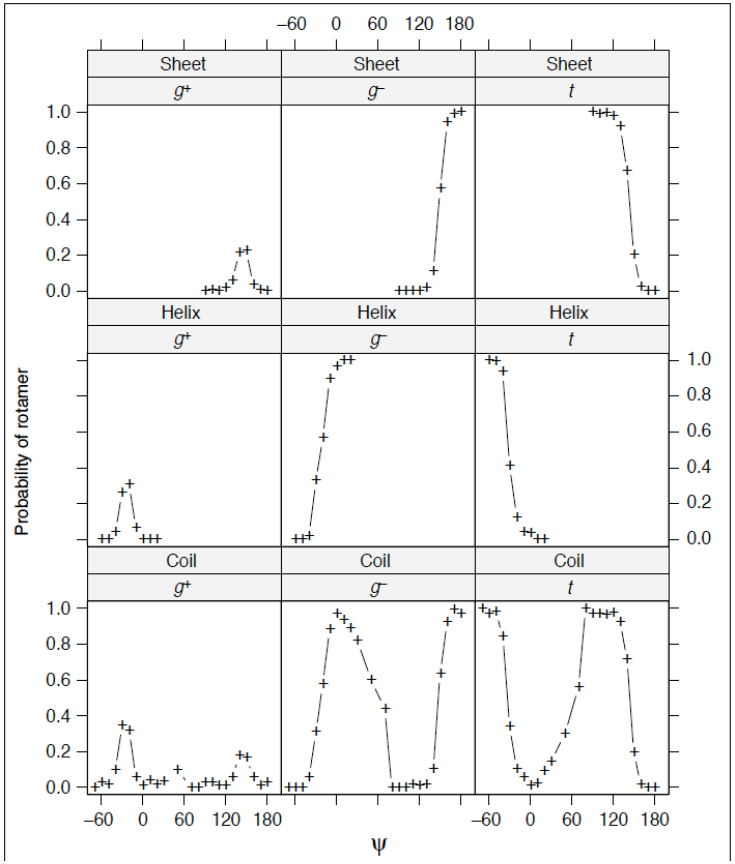
Günstige χ_1-Drehwinkel der Valin-Seitenkette:
beobachtete Häufigkeit der Rotamere

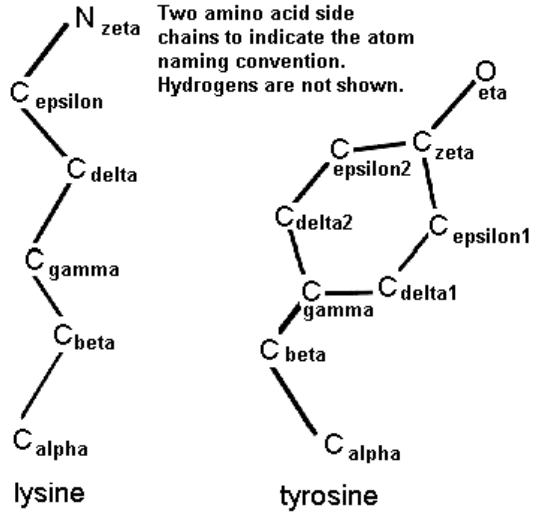
gauche$^+$ ($\chi_1 \sim +60°$)

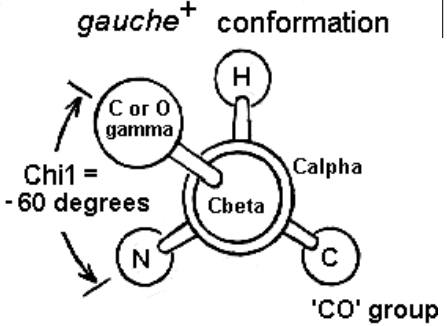
gauche$^-$ ($\chi_1 \sim -60°$)

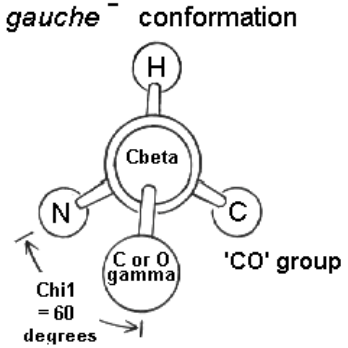
trans ($\chi_1 \sim 180°$)

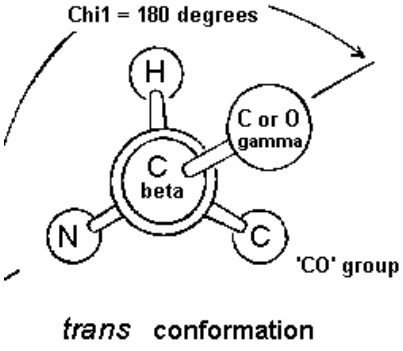
in verschiedenen Sekundärstrukturen
als Funktion des Rückgratsdiederwinkels Ψ.











R. Dunbrack (2002) Curr.Opin.Struct.Biol. 12, 431
http://swissmodel.expasy.org/course/text/chapter3.htm

Paarungs-Präferenz von Aminosäuren

Bei der Orientierung der Seitenketten wird üblicherweise jede für sich betrachtet (Rotamer-Bibliothek berüchsichtigt zwar die Konformation des Rückgrats, aber nicht die Umgebung).

Aminosäuren nehmen jedoch je nach Umgebung unterschiedliche Konformationen ein.

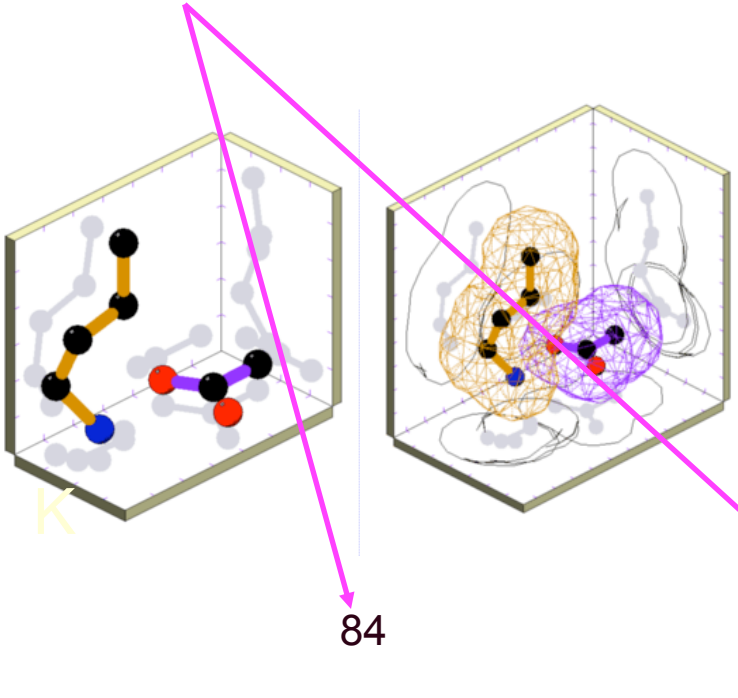
Diese “Packungseffekte” können ebenfalls für komparative Modellierung berücksichtigt werden.

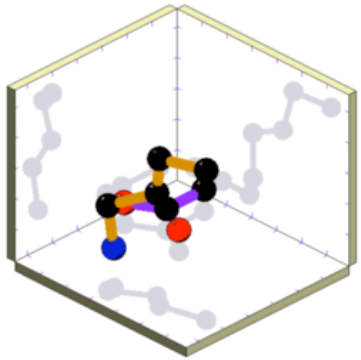
Salzbrücken

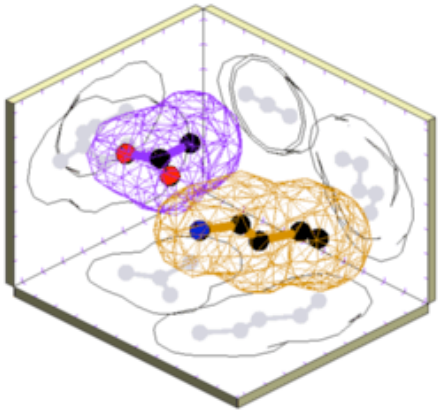
Datenbank für statistische Präferenz für die Orientierung von Aminosäure-Seitenketten in der PDB-Datenbank:

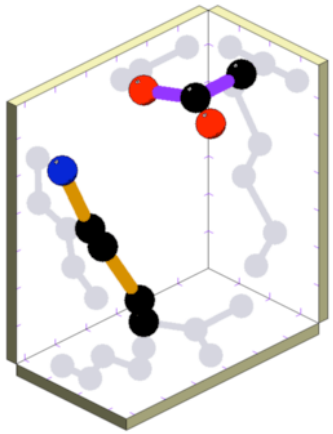
http://www.biochem.ucl.ac.uk/bsm/sidechains/index.html#

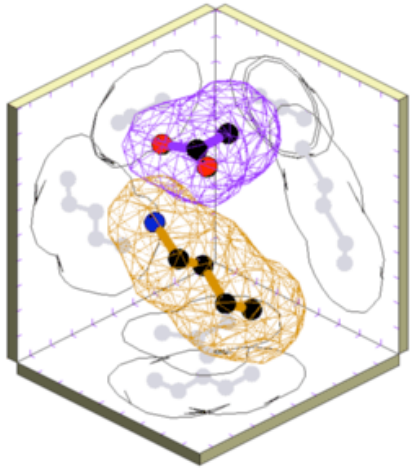
Häufigkeit von 1845 Asp-Lys-Kontakte in PDB-Datenbank




66


34


60

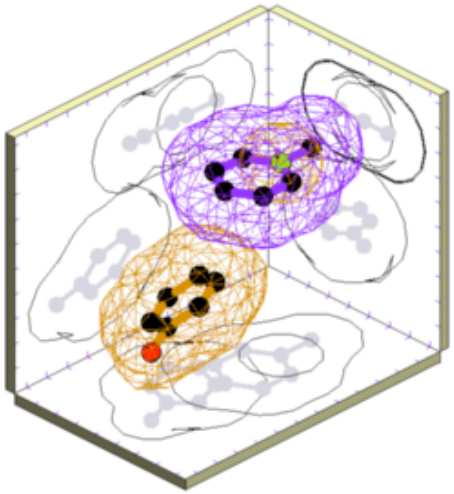

31

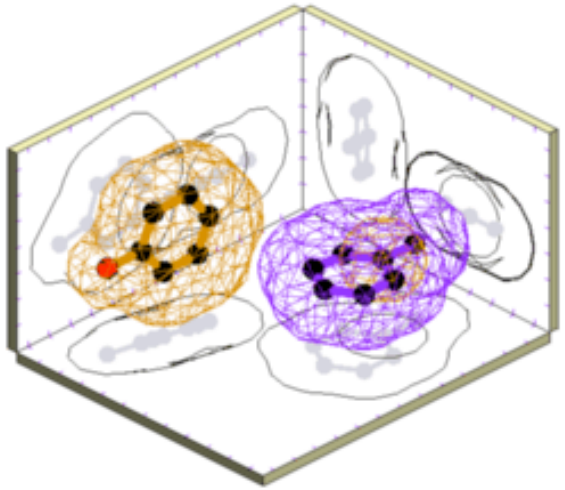
Π-stacking von aromatischen Ringen

Aromatische Ringe (z.B. Phenol, Benzol, Seitenketten von Tyrosin, Phenylalanin, Tryptophan, Histidin …) besitzen delokalisiertes Elektronensystem ausserhalb der Ringebene.

Mehrere dieser Ringe “packen” gerne aufeinander bzw. senkrecht zueinander.

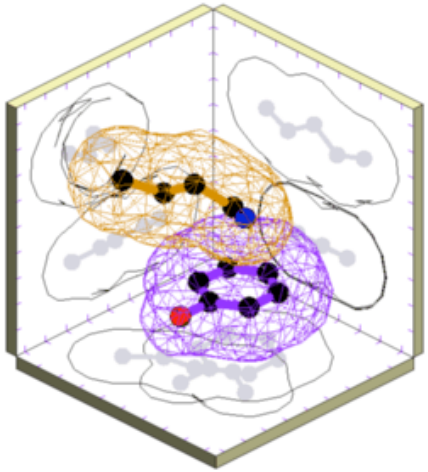
Cluster Phe-Tyr 1 4





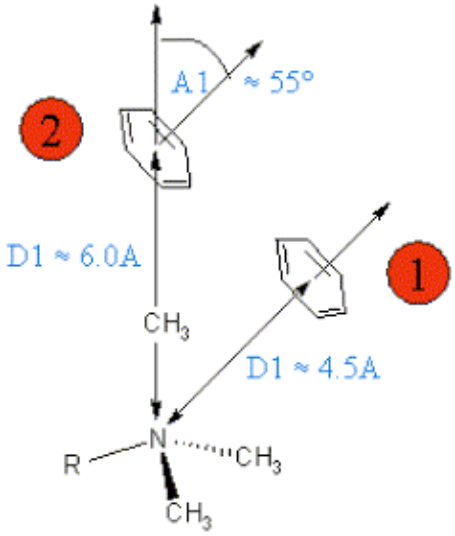
Kationen-Π-Wechselwirkung

Die gleichen **aromatischen Ringe** wechselwirken gerne senkrecht zur Ringebene mit positiv geladenen Gruppen.



Tyr-Lys Cluster 6

Beispiele: Acetylcholin in Bindungstasche von Acetylcholinesterase



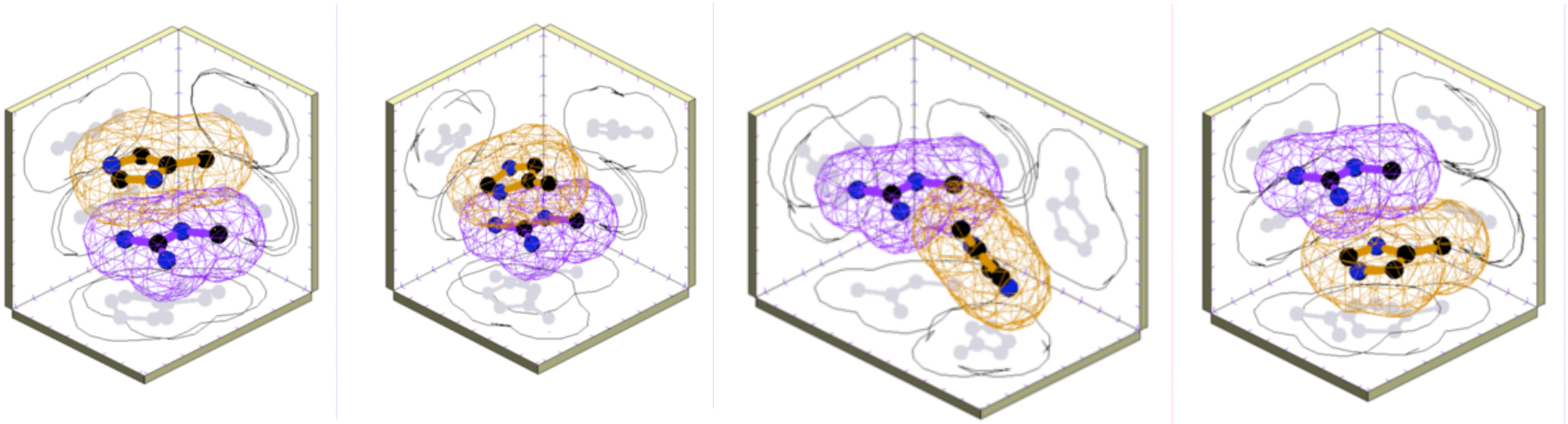
Bevorzugte Geometrien für die Wechselwirkung von Trimethyl-Ammoniumgruppen mit Phenyl-Ringen
Gohlke & Klebe, JMB 2000

Kationen-Π-Wechselwirkung

Wechselwirkung der positiv geladenen Guanidinium-Gruppe von Arg mit dem Π-Elektronensystem von His.

Fast immer planare Packung. Nur in Cluster 3 Ausbildung einer Wasserstoffbrücke N-H … N

1 2 3 4

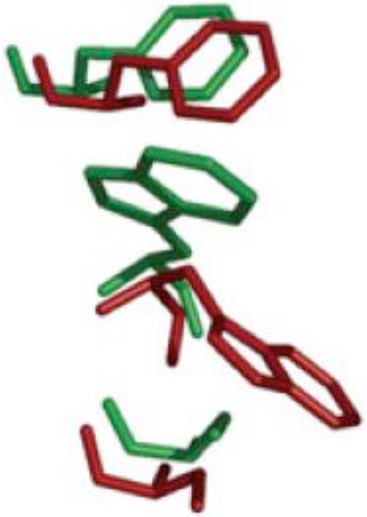


Typische Fehler bei Homologie-Modellierung (I)

(1) Fehlerhafte Packung der Seitenketten.

In rot gezeigt ist die Kristallstruktur des cellular retinoic acid binding protein I (CRAB1) aus Maus.

Die modellierte Struktur der Tryptophan Residue 109 (Mitte) ist in grün gezeigt.

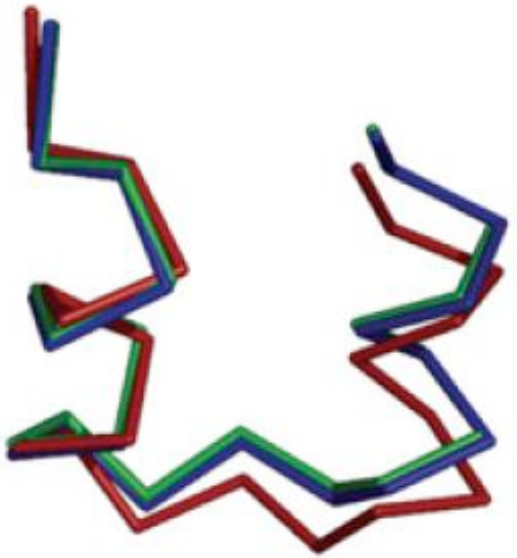


Eswar, Curr. Protocols in Bioinf. (2006)

Typische Fehler bei Homologie-Modellierung (II)

(B) **Verschiebungen in korrekt alignierten Regionen**.

Hier ergeben sich leichte Abweichungen des Modells des CRAB1 Proteins (grün) von der Kristallstruktur des CRAB1 (rot) entsprechend der Kristallstruktur des fatty acid binding protein (blau), das als Vorlage benutzt wurde.



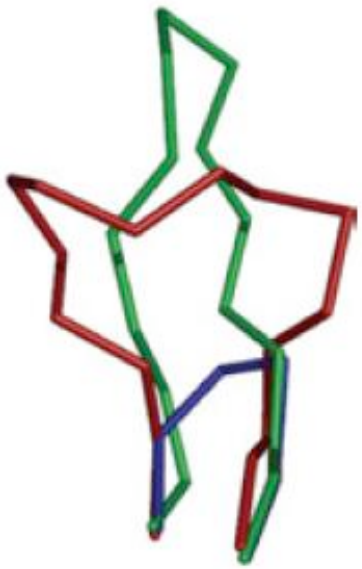
Eswar, Curr. Protocols in Bioinf. (2006)

Typische Fehler bei Homologie-Modellierung (III)

(C) **Fehler in Regionen ohne Vorlage**.

Gezeigt ist die Verbindung zwischen den Cα-Atomen der Schleife 112–117 für

- die Kristallstruktur des menschlichen eosinophil neurotoxin (rot),
- dessen Modell (grün),und
- die Vorlagestruktur Ribonuclease A (blau).



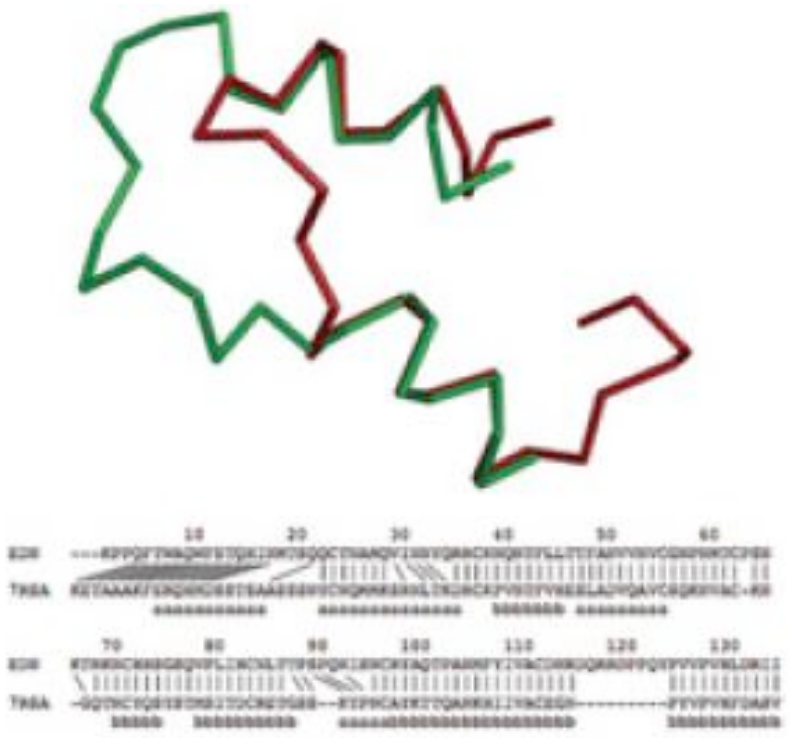
Eswar, Curr. Protocols in Bioinf. (2006)

Typische Fehler bei Homologie-Modellierung (IV)

(D) **Fehler durch Misalignment**.

N-terminale Region der Kristallstruktur von menschlichem eosinophil neurotoxin (rot) im Vergleich mit dem Modell (grün).

Der Fehler resultiert aus dem ungünstigen Alignment mit der Vorlage Ribonuclease A (unten).



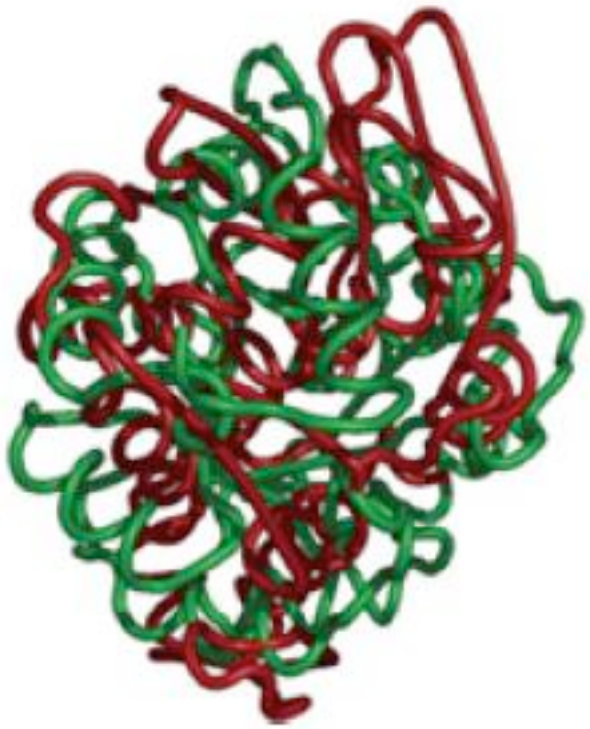
Eswar, Curr. Protocols in Bioinf. (2006)

Typische Fehler bei Homologie-Modellierung (V)

(E) **Fehler durch inkorrekte Vorlage**.

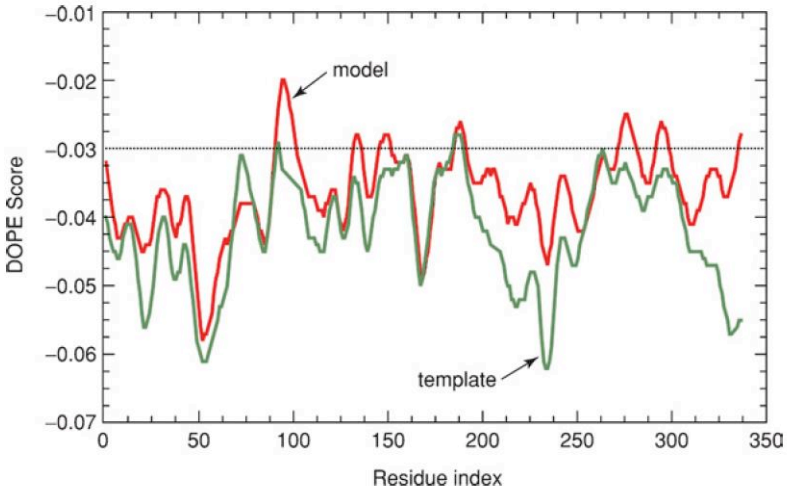
Vergleich der Kristallstruktur für

α-trichosanthin (rot) mit dem Modell (grün), das mit Indol-3-Glycerophosphat-Synthase als Vorlage erzeugt wurde..



Eswar, Curr. Protocols in Bioinf. (2006)

Bewertung von Strukturmodellen (Modeller)



Modeller verwendet das DOPE-Potential (Discrete Optimized Protein Energy) zru Bewertung von Strukturmodellen.

Niedrigere Energien sind besser.

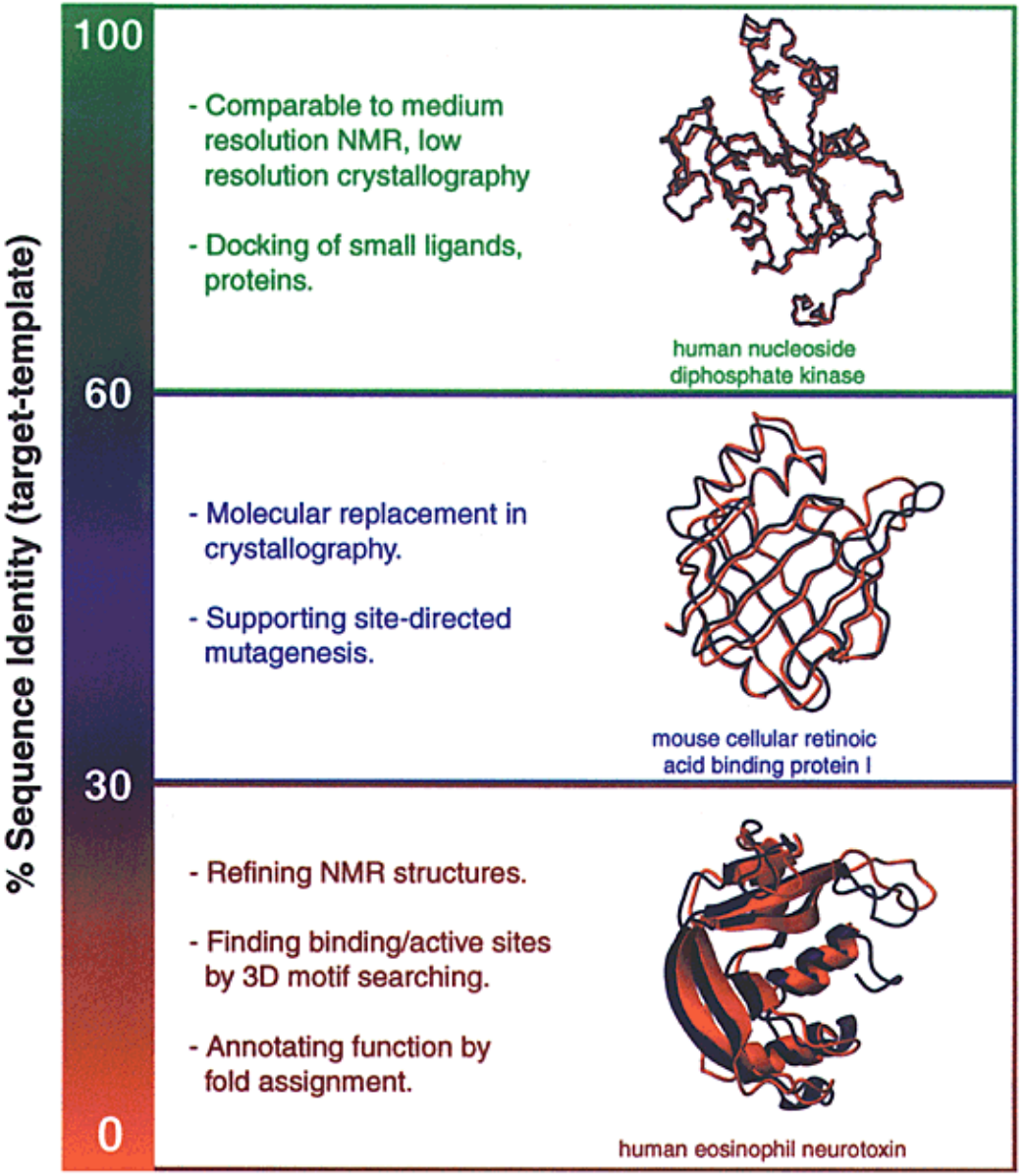
DOPE ist ein statistisches Potential für die Wahrscheinlichkeiten, wie häufig bei einem bestimmten Abstand das Atompaar *i* – *j* in den bekannten Proteinstrukturen auftritt.

Eswar, Curr. Protocols in Bioinf. (2006)

Homologie/Komperative Modellierung

Qualität der Modellierung hängt von Sequenzidentität mit Vorlage ab.

Man sollte stets beachten, dass die Vorlage nicht aus der Twilight Zone stammt.



Protein structure modeling for structural genomics.
R. Sánchez *et al.* Nat. Struct. Biol. 7, 986 - 990 (2000)

Bewertung der Qualität eines Homologiemodells - Allgemeine Gesichtspunkte

- Ein Modell wird als **falsch** angesehen, wenn mindestens eines seiner strukturellen Elemente gegenüber dem Rest des Modells falsch angeordnet ist. Dies kann durch ein falsches Sequenzalignment entstehen.

  Das Modell kann dennoch korrekte Stereochemie besitzen.

- Man kann ein Modell als **ungenau** ansehen wenn seine atomare Koordinaten mehr als 0.5 Å von einer experimentellen Kontrollstruktur abweichen.

- Ungenauigkeiten können auch in der Stereochemie (Bindungslängen und –winkel auftreten). Dies kann leicht mit **WhatCheck** überprüft werden.

- **Statistische Paarpotentiale** für die Verteilung von Aminosäuren in bekannten Proteinen erlauben manchmal die Aufspürung von fehlerhaften Modellen.

www.expasy.org/swissmodel/SWISS-MODEL.html

Proteinkern und Loops

Fast jedes Proteinmodell enthält nicht-konservierte Loops, die als die am wenigsten zuverlässigen Teile des Proteinmodells angesehen werden können.

Andererseits sind diese Bereiche der Struktur oft auch am flexibelsten –

hohe Temperaturfaktoren in Kristallstrukturen oder hohe Unterschiede zwischen verschiedenen (gleichsam gültigen) NMR-Strukturen.

Die Residuen im Proteinkern werden gewöhnlich fast in der identischen Orientierung wie in experimentellen Kontrollstrukturen modelliert.

Residuen an der Proteinoberfläche zeigen grössere Abweichungen.

www.expasy.org/swissmodel/SWISS-MODEL.html

Vergleich zweier Strukturen: RMSD

Root mean square deviation:

$$RMSD_{1,2} = \sqrt{\frac{\sum_{i=1}^{n} \left(x_{1,i} - x_{2,i}\right)^2}{n}}$$

Man vergleicht zwei Proteinstrukturen 1 und 2 durch die Berechnung des mittleren quadratischen Abstands der Koordinaten der *n* sich entsprechenden Atome.

Dann nimmt man noch die Wurzel daraus.

Werte unterhalb von 0.2 nm oder 2 Å kennzeichnen eine hohe strukturelle Ähnlichkeit.

Zum Vergleich: die Länge einer C-C Bindung beträgt 0.15 nm.

Die Distanzen aller Atome weichen also höchstens etwa um eine Bindungslänge voneinander ab.

Test für die Zuverlässigkeit von SwissModell

3DCrunch-Projekt von Expasy zusammen mit SGI.

Idee: Generiere „Homologie-Modelle“ für Proteine mit bekannter 3D-Struktur um zu überprüfen, wie genau die mit Homologie-Modellierung erzeugten Strukturmodelle sind.

Die Vorlagen besaßen 25 – 95 % Sequenzidentität mit dem Zielprotein.

1200 Kontrolle-Modelle wurden erstellt.

Grad der Identität [%]	Modell innerhalb von x Å RMSD zur Vorlage					
	< 1	< 2	< 3	< 4	< 5	> 5
25-29	0	10	30	46	67	33
30-39	0	18	45	66	77	23
40-49	9	44	63	78	91	9
50-59	18	55	79	86	91	9
60-69	38	72	85	91	92	8
70-79	42	71	82	85	88	12
80-89	45	79	86	94	95	5
90-95	59	78	83	86	91	9

www.expasy.org/swissmodel/SWISS-MODEL.html

Ligandendocking in Homologiemodelle ??

- Homologiemodelle können zwar recht gut sein, aber nicht immer für Ligandendocking geeignet sein
- Grund: falsche Seitenkettenrotamere in Bindungstasche

- Ansatz1: verwende flexibles Docking, wo auch Teile des Proteins flexibel sind

- Ansatz2: verwende zusätzliches experimentelles Wissen, verlangt manuelles Vorgehen

- Ansatz3: erstelle Homologiemodell in Anwesenheit eines modellierten Liganden, dessen Position z.B. aus Modell-Vorlage stammt

Zusammenfassung – Homologiemodellierung

- Gemeinsamer Kern von Proteinen mit 50% Sequenzidentität besitzt ca. 1 Å RMSD
- Dies gilt sogar für absolut identische Sequenzen.
- Der zuverlässigste Teil eines Proteinmodells ist der Sequenzabschnitt, den es mit der Vorlage gemeinsam hat. Die größten Abweichungen liegen in den konstruierten Schleifen.
- Die Wahl der Modellvorlage ist entscheidend!

  Die An- oder Abwesenheit von Ko-faktoren, anderen Untereinheiten oder Substraten kann Proteinkonformation sehr beeinflussen und somit alle Modelle, die von ihnen abgeleitet werden.
- Jeder Fehler im Alignment produziert falsche Modelle!

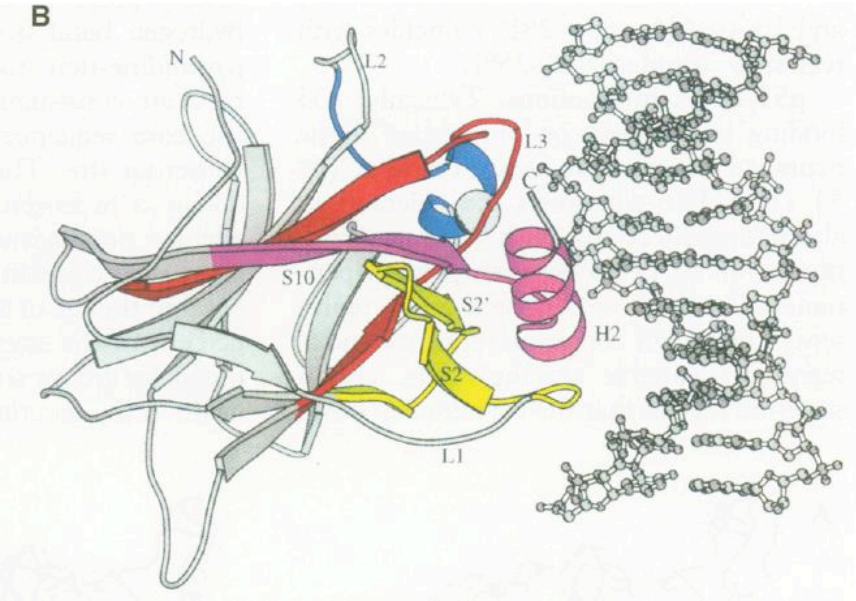
  Solche Alignment-Fehler treten bei Sequenzidentität unter 40% auf.

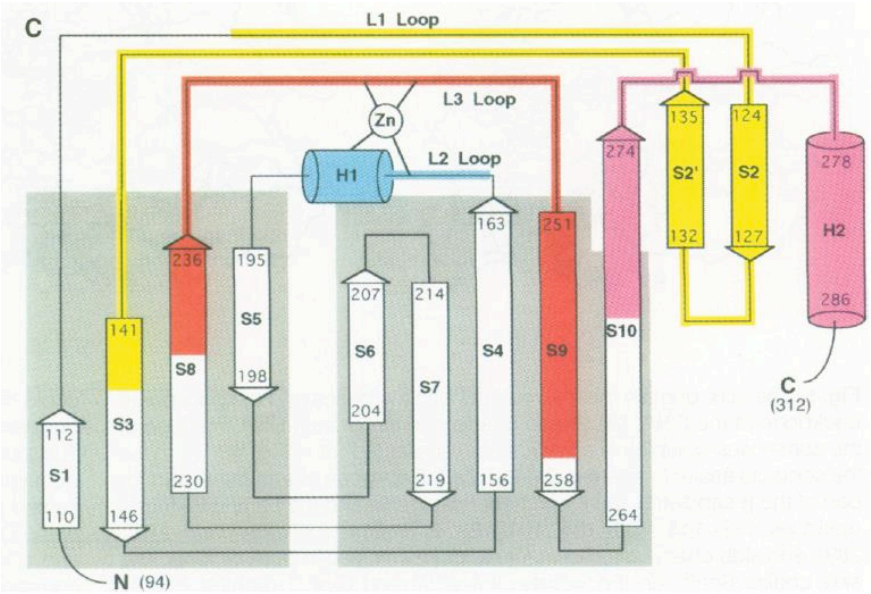
V7 Modellierung von biomolekularen Komplexen

- **Protein-Protein-Docking**

- **Protein-DNA-Komplexe**

Konservierte Bereiche sind am Interface angereichert

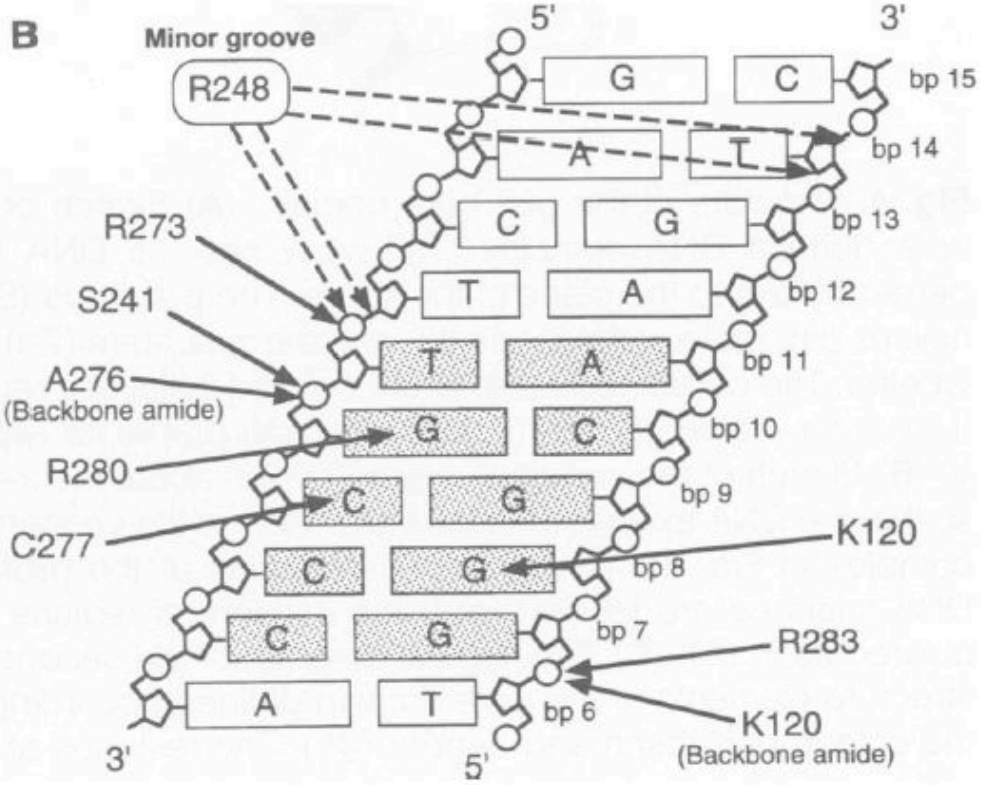


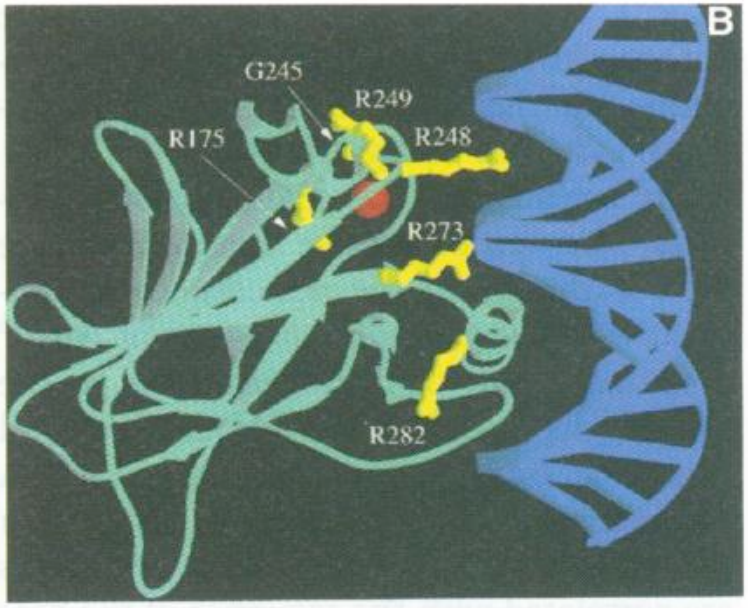


colored as in (A). (**C**) Topological diagram of the secondary structure elements of the core domain defined according to the criteria of Kabsch and Sander (*45*). The residues at the start and the end of each secondary structure element are indicated. The conserved regions are colored according to (A), and the boundaries of the two β sheets that make up the β sandwich are shaded.

Science 265, 346-355 (1994)

Kontakt-Residuen





Links: Protein – DNA-Kontakte enthalten viele Arginin- und Lysin-Reste

Rechts: die 6 am häufigsten bei **Krebs** mutierten Aminosäuren (gelb)
in p53 liegen alle am Interface!

Science 265, 346-355 (1994)

Modellierung von Protein-DNA-Komplexen

Eine besondere Herausforderung bei der Modellierung von Protein-DNA-Kontakten ist, dass beide Bindungspartner flexibel sind.

Wichtigstes Prinzip:

- Elektrostatische Komplementarität – DNA ist stark negativ geladen, Proteinoberfläche muss entsprechend positiv geladen sein

Science 265, 346-355 (1994)

Eigenschaften von Protein-Protein-Interfaces

Table 2. *Properties of protein–protein interfaces*

Parameter	Protein–protein complexes[a]	Homodimers[b]		Weak dimers[c]	Crystal packing[d]
		Bahadur	Dey		
Number in dataset	70	122	276	19	188
BSA (Å^2)	1910	3900	3700	1620	570, 1510
(S.D.)	(760)	(2200)	(2160)	(670)	(520)
Amino acids per interface	57	104	100	50	48
BSA (Å^2) per amino acid	34	38	37	32	32
Composition (BSA %)					
Non-polar	58	65	65	62	58
Neutral polar	28	23	22	25	25
Charged	14	12	13	13	17
Atomic packing					
f_{bu} (buried atoms %)	34	36	35	28	21
L_D packing index	42	45	43	34	32
S_c complementarity score	0·69	0·70			0·63
R_p propensity score[e]	0·9	4·3	2·1	0·5	−1·1
Chain segments[f]	5·6	3·4	3·2	5·8	6·3
H bonds					
n_{HB} (number per interface)	10	19	18	7	5
BSA per bond (Å^2)	190	210	209	230	280
Water molecules[g]					
Number per interface	20	44			23
Number per 1000 Å^2	10	11			15
Bridging H bonds	6	13			6
Residue conservation[h]					
% in core	55	60			40
s in core and rim	0·65 and 0·80	0·63 and 0·77			0·98 and 0·99

Homodimere sind meist permanente Komplexe
-> sehr große Schnittstellen, großer Anteil an unpolaren Aminosäuren

BSA: buried surface area

Artifizielle Kristallkontakte enthalten einen wesentlich kleineren Anteil an konservierten Residuen (40% gegenüber 55-60%).

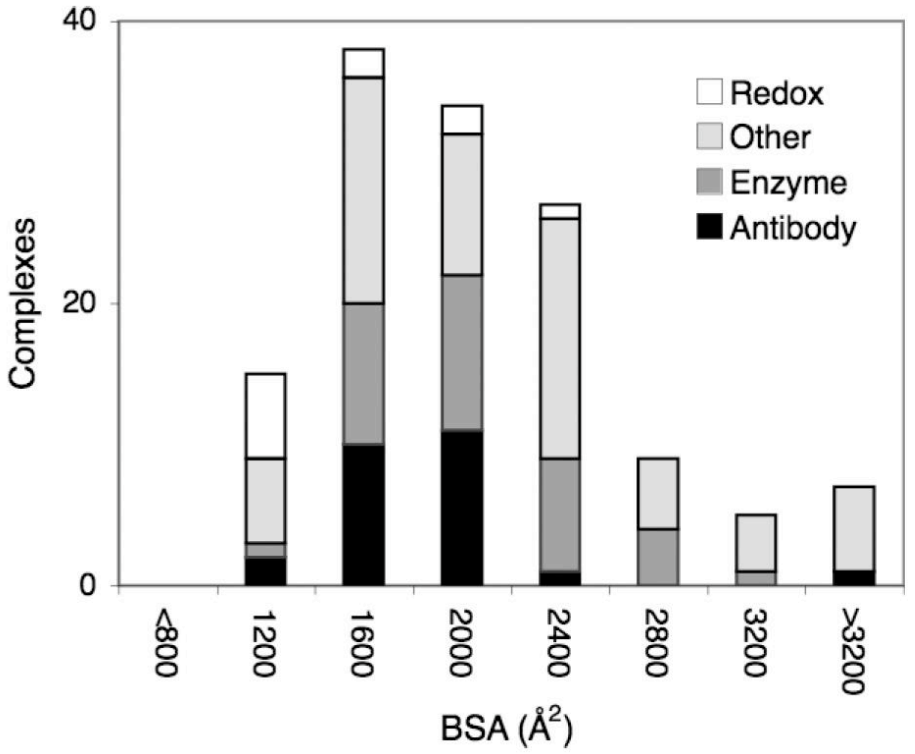
Janin et al. Quart Rev Biophys 41, 133-180 (2008)

Fläche von Protein-Protein-Interfaces

Redox-Komplexe dienen z.B. zur Übertragung von Elektronen (Cytochrom c).
Beobachtung: Sie haben recht kleine Schnittstellen.

Ihre Funktion benötigt nur eine transiente = kurzlebige Bindung.

Dies wird durch die kleine Größe der Schnittstelle unterstützt.



Janin et al. Quart Rev Biophys 41, 133-180 (2008)

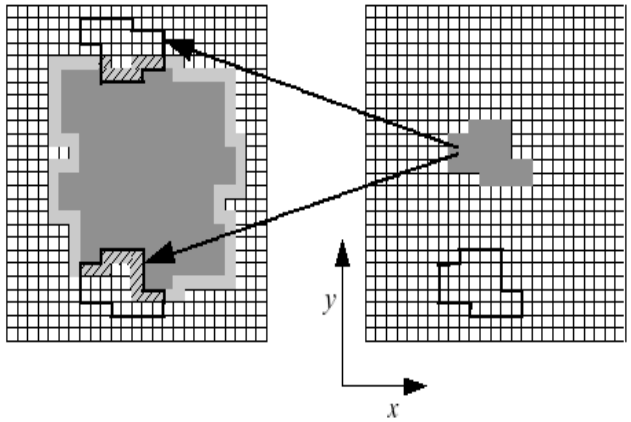
Protein-Protein-Komplexe

Die bekannten Strukturen von Protein-Protein-Interaktionen des Menschen (bekannte experimentelle Strukturen plus Homologie-Modelle) decken nur etwa 4% der geschätzten Anzahl von etwa 300.000 Protein-Protein-Interaktionen zwischen menschlichen Proteinen ab.

Quelle: Proteins, 81, 2192–2200 (2013)

Idee: versuche, die Strukturen von PP-Komplexen durch **Docking** zu modellieren.

Berechne die Komplementarität der Oberflächen zwischen beiden starren Proteinen in allen möglichen Orientierungen (alle erlaubten Translationen + Rotationen), Suche in 6 Freiheitsgraden



Proteins, 81, 2192–2200 (2013)

Workflow für Protein-Protein-Docking

(1) Docking von starren Proteinen
mit Katchalski-Kazir-Algorithmus
z.B. FTDock-Program, verwendet FFT,
Optimale Lösung hat maximales **a** x **b**

Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 2195–2199, March 1992
Biophysics

Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques

(protein–protein interaction/surface complementarity/macromolecular complex prediction/molecular docking)

EPHRAIM KATCHALSKI-KATZIR[†‡], ISAAC SHARIV[§], MIRIAM EISENSTEIN[¶], ASHER A. FRIESEM[§], CLAUDE AFLALO[‖], AND ILYA A. VAKSER[†]

Departments of [†]Membrane Research and Biophysics, [§]Electronics, [¶]Structural Biology, and [‖]Biochemistry, Weizmann Institute of Science, Rehovot 76100, Israel

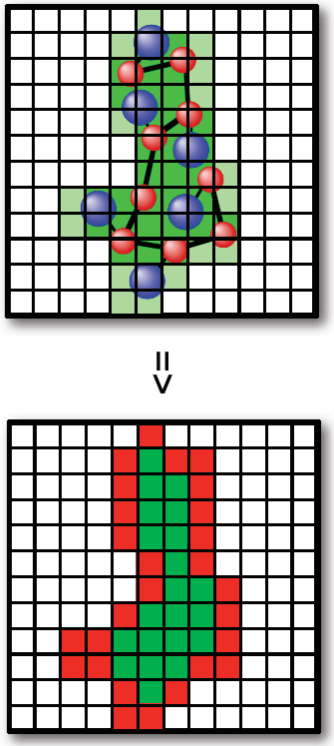
Next, to distinguish between the surface and the interior of each molecule, we retain the value of 1 for the grid points along a thin surface layer only and assign other values to the internal grid points. The resulting functions thus become

$$\bar{a}_{l,m,n} = \begin{cases} 1 & \text{on the surface of the molecule} \\ \rho & \text{inside the molecule} \\ 0 & \text{outside the molecule,} \end{cases} \quad [2a]$$

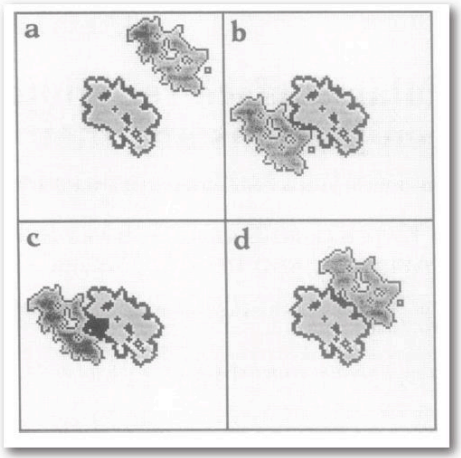
and

$$\bar{b}_{l,m,n} = \begin{cases} 1 & \text{on the surface of the molecule} \\ \delta & \text{inside the molecule} \\ 0 & \text{outside the molecule,} \end{cases} \quad [2b]$$

where the surface is defined here as a boundary layer of finite width between the inside and the outside of the molecule. The parameters ρ and δ describe the value of the points inside the molecules, and all points outside are set to zero. Two-



2D cross sections at l = 46 (N = 90)



a) no contact
b) limited contact
c) overlap (black area)
d) good geometric match

Typical values: ρ = –15, δ = 1
=> penalty for overlap of volumes

Workflow für Protein-Protein-Docking

(2) Wende Zdock-Scoring-Funktion auf die FTDock-Ergebnisse an

-> wähle 1000-2000 Kandidaten-Modelle mit dem besten Score aus.

Zdock Scoring-Funktion: statistisches Paar-Potential zwischen Kontaktresiduen, Oberflächenkomplementarität und Elektrostatik

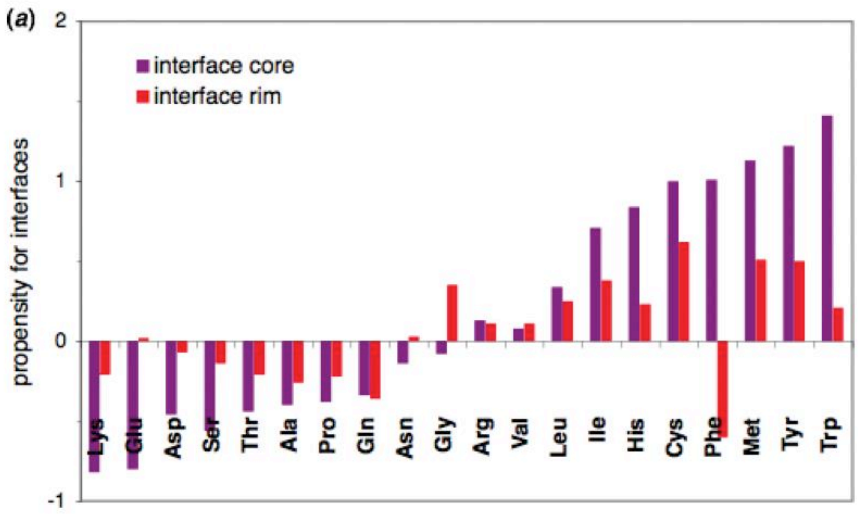
(3) Bewerte Zdock-Lösungen noch einmal mit pyDOCK-Scoring-Funktion:

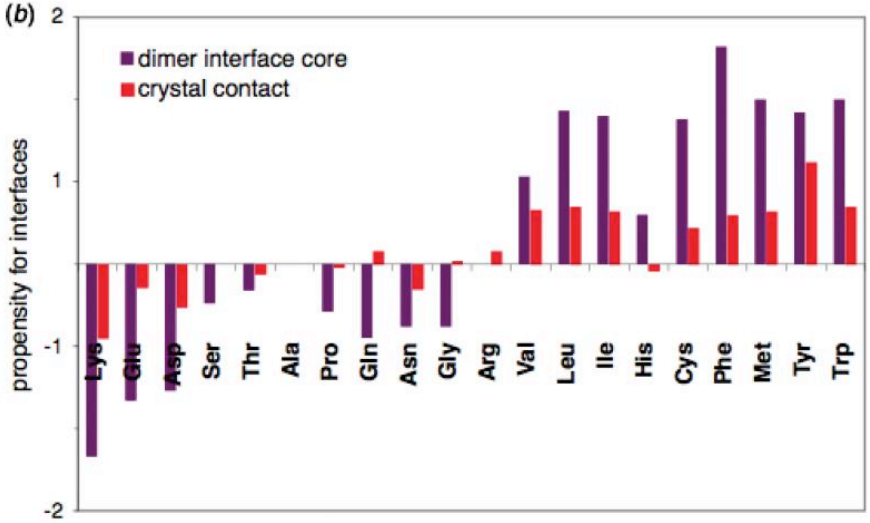
- **Desolvationsenergie** proportional zur Abnahme der gesamten Solvent-accessible surface area (-> Hydrophober Effekt in V5)

- **Coulomb-Elektrostatik** und **van-der-Waals-Wechselwirkungen** zwischen Kontakt-Residuen

Proteins, 81, 2192–2200 (2013)

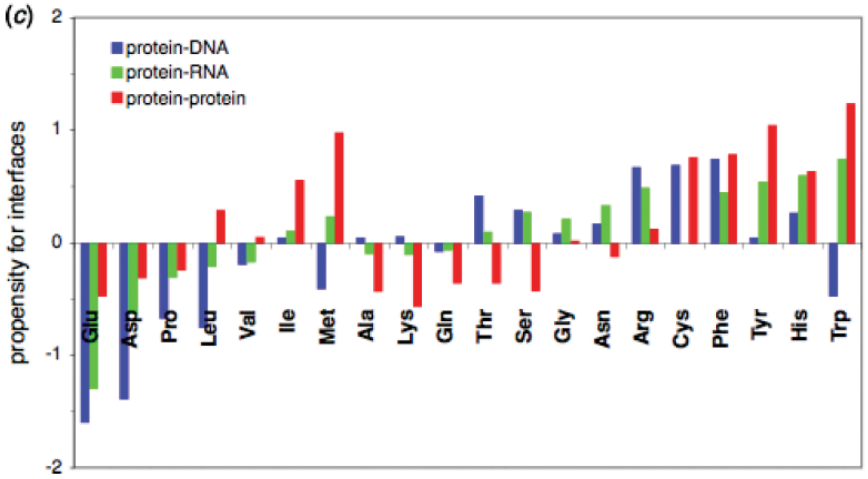
Komposition von Bindungsschnittstellen





(b) Artifizielle Kristallkontakte haben zu wenig hydrophobe Aminosäuren am Interface.

(c) An Interfaces mit negativ geladener DNA gibt es viel weniger negative Glu- und Asp-Residuen als bei PP-Kontakten. Dafür mehr positive Arg-Residuen.



Janin et al. Quart Rev Biophys 41, 133-180 (2008)

Protein-Nukleinsäure-Komplexe

Table 3. *Properties of protein–nucleic acid interfaces*

Average value	Protein/RNA[a]	Protein/DNA[b]	Protein/protein[c]
Number of complexes	81	75	70
BSA (Å^2)			
Mean	2530	3100	1910
S.D.	(1210)	(1050)	(760)
Protein/nucleic acid	1210/1320	1540/1560	–
Number of amino acids/nucleotides	43/18	48/18	57
BSA (Å^2) per amino acid/nucleotide	28/75	33/72	34
Composition (protein/nucleic acid, BSA %)			
Non-polar	55/33	52/41	58
Neutral polar	21/41	24/16	28
Charged (negative)	4/26	2/43	5
Charged (positive)	20/0	23/0	9
f_{bu} (% buried atoms, protein/nucleic acid)	29/29	24/28	34
L_D (packing index, protein/nucleic acid)	37/43	39/46	42
H bonds			
n_{HB} (number per interface)	20	22	10
BSA per bond (Å^2)	125	145	190
Water molecules			
Number per interface	32	21	20
Number per 1000 Å^2	13	7	10
Bridging H bonds	11		6

Eher größere Kontaktflächen.

Hoher Anteil an positiv geladenen Aminosäuren.

Janin et al. Quart Rev Biophys 41, 133-180 (2008)

V8 Genexpression - Microarrays

- **Idee:** analysiere Ko-Expression von mehreren Genen um auf funktionelle Ähnlichkeiten zu schließen

- **wichtige Fragen:**

  (1) wie wird Genexpression reguliert?

  (2) was wird mit MicroArray-Chips gemessen?

  (3) wie analysiert man Daten aus MicroArray-Experimenten?

  (4) was bedeutet Ko-Expression funktionell?

- **Inhalt V8:**

  (1) Hintergrund zu Transkription und Genregulationsnetzwerken

  (2) Micro-Arrays

  (3) Übung: analysiere selbst Daten aus einem MicroArray-Experiment

Was mißt man mit Microarrays?

Häufig verwendet werden **Zweifarben-MicroAssays**:

Sample A: rot

Sample B: grün

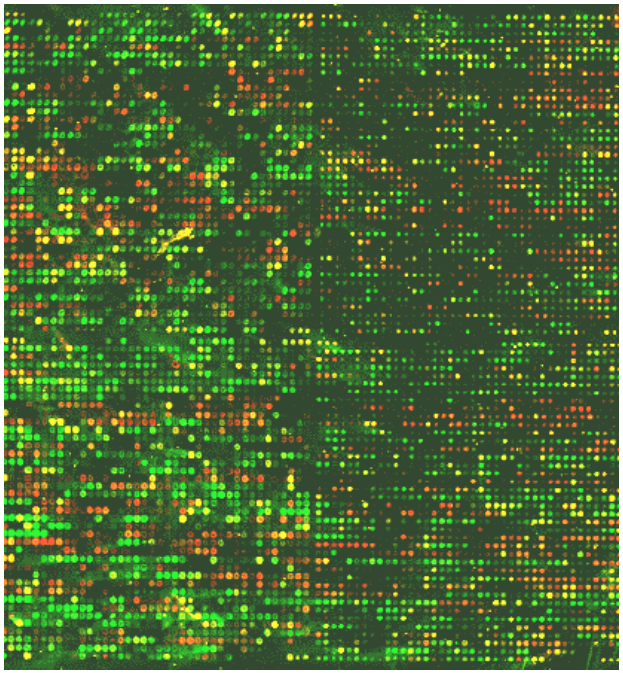
Ziel: bestimme das Verhältnis rot/grün

dunkel: Gen weder in A noch B exprimiert

rot: Gen nur in A exprimiert (bzw. viel stärker)

grün: Gen nur in B exprimiert

gelb: Gen in A und in B exprimiert.



Das Licht wird von zwei Farbstoffen (roter Cy5 und grüner Cy3) erzeugt, die an die cDNA angeheftet wurden (die cDNA wurde „gelabelt“) und die unter Laserlicht fluoreszieren.

pgrc.ipk-gatersleben.de

Experimentelles Vorgehen

Isolierung einer Zelle im Zustand X

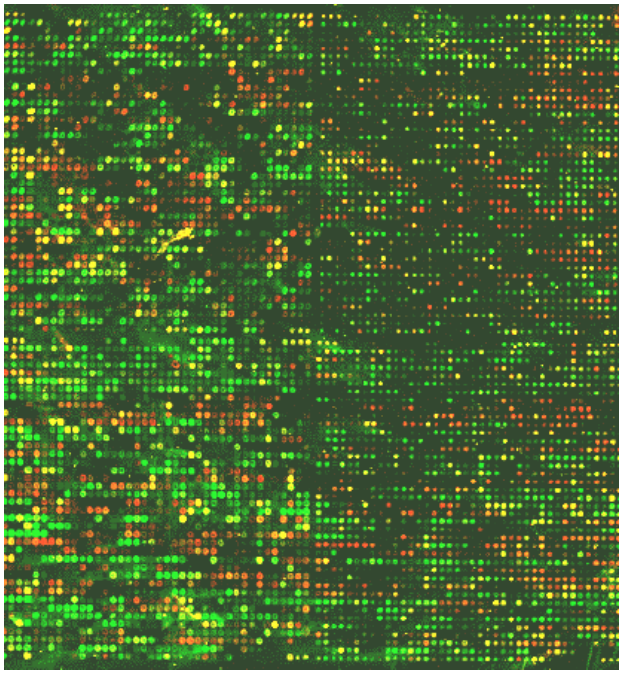
Extraktion aller RNA

Umwandlung in cDNA

Markierung mit Farbstoff (rot oder grün)

Pipette enthält markierte cDNA aller in der Zelle exprimierten Gene.

Man bringt nacheinander die cDNA aus zwei verschiedenen Zellpräparationen auf, die unterschiedlich (rot/grün) gelabelt wurden.

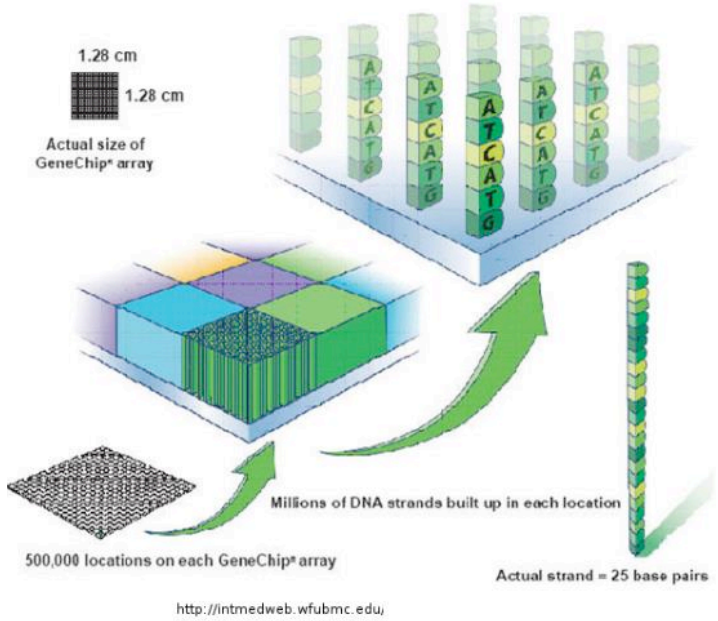


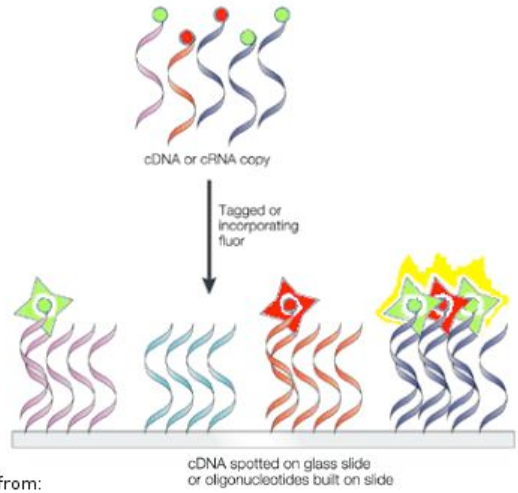
pgrc.ipk-gatersleben.de

Experimentelles Vorgehen

Aufbringen des zellulären cDNA-Gemischs auf die einzelnen Zellen des Arrays.

Jede Zelle enthält an die Oberfläche funktionalisiert einen cDNA-Klon aus einer cDNA-Bibliothek.





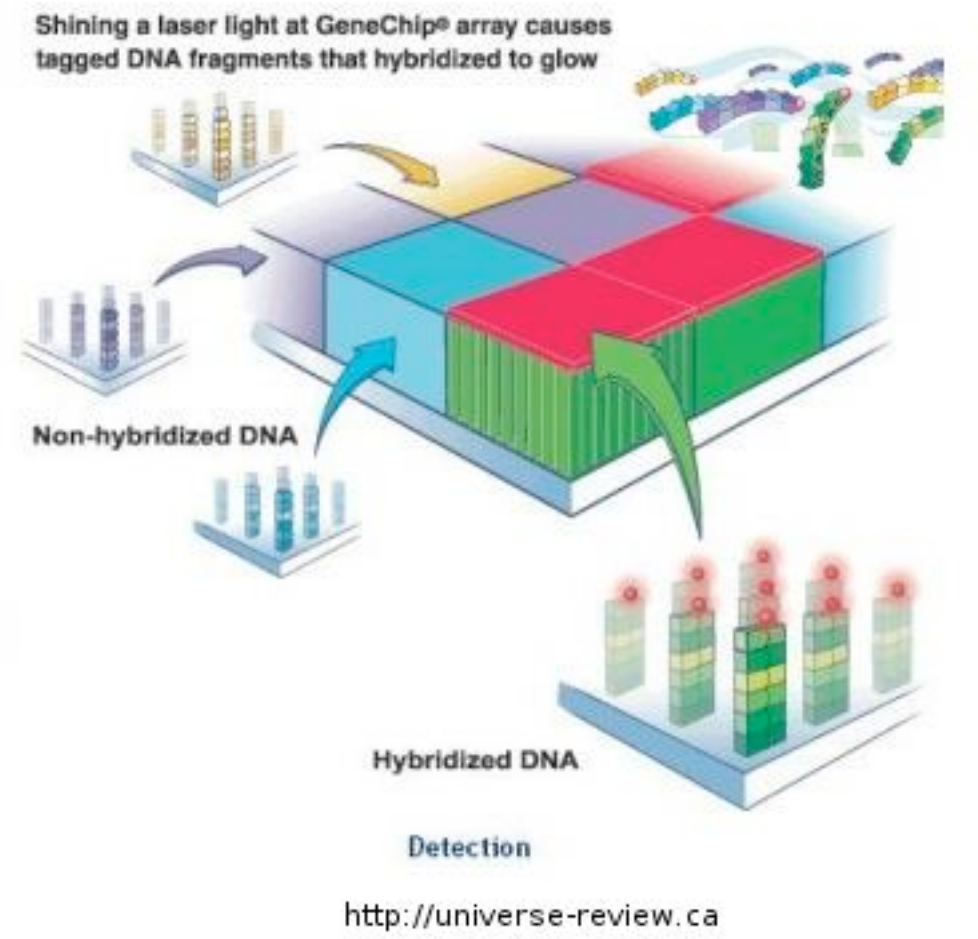
changed from:
A. Butte, Nature Reviews Drug Discovery 1, 951-960, 2002

Jede **Zelle** misst daher die Expression eines **einzelnen Gens**.

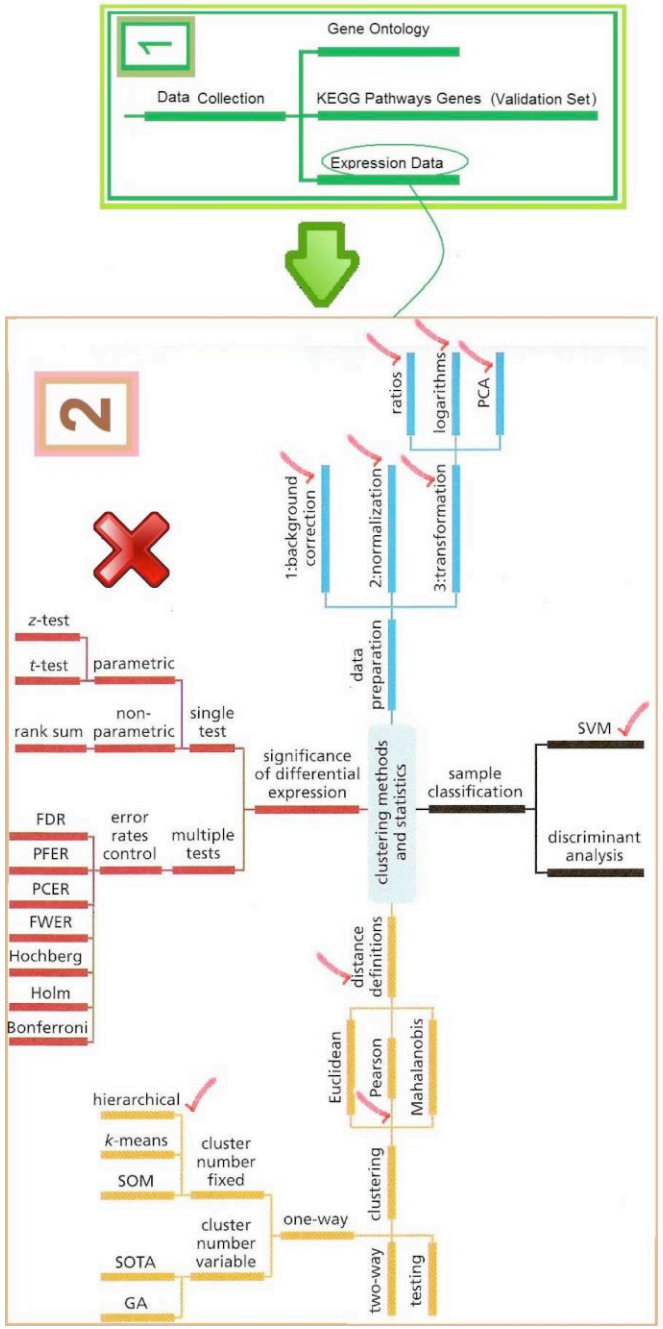
pgrc.ipk-gatersleben.de

Auslesen der Probe: Laserlicht

Man stimuliert sowohl die Fluoreszenz bei der roten als auch bei der grünen Wellenlänge.



Auswertung von Microarray-Experimenten



Korrekte Reihenfolge bei Prozessierung der Daten:

0. Löschen fehlerhafter Daten
1. Background-correction

(Details werden hier nicht behandelt)

1. Normalisierung
2. Transformation (z.B. Log)

Dann

Clustering

bzw.

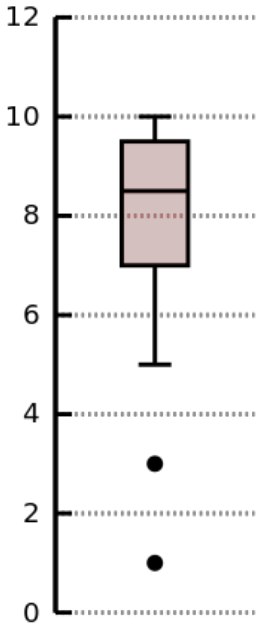
Analyse der Signifikanz

Boxplot

Die Boxplot-Darstellung erlaubt es, schnell einen Überblick über die Werteverteilung in einem Datensatz zu erhalten. Beispiel:

Datenpunkt	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Wert (unsortiert)	9	6	7	7	3	9	10	1	8	7	9	9	8	10	5	10	10	9	10	8
Wert (sortiert)	1	3	5	6	7	7	7	8	8	8	9	9	9	9	9	10	10	10	10	10

Kennwert	**Beschreibung**	**Lage im Boxplot**
Minimum	Kleinster Datenwert des Datensatzes	Ende eines Whiskers oder entferntester Ausreißer
Unteres Quartil	Die kleinsten 25% der Datenwerte sind kleiner oder gleich diesem Wert	Beginn der Box
Median	Die kleinsten 50% der Datenwerte sind kleiner oder gleich diesem Kennwert	Strich innerhalb dieser Box
Oberes Quartil	Die kleinsten 75% der Datenwerte sind kleiner oder gleich diesem Kennwert	Ende der Box
Maximum	Größter Datenwert des Datensatzes	Ende eines Whiskers oder entferntester Ausreißer



www.wikipedia.de

1. Normalisierung von Arrays

Wie alle anderen biologischen Experimente zeigen auch Microarrays **zufällige** und **systematische Abweichungen**.

Zufällige Schwankungen treten auf

- in der absoluten Menge an mRNA, die eingesetzt wird,
- in der Hybridisierungs-Technik und
- in Waschschritten.

Systematische Unterschiede gibt es z.B. bei den physikalischen Fluoreszenz-eigenschaften der beiden Farbstoffmoleküle bzw. durch inhomogen hergestellte Microarray-Chips.

Um diese systematischen Abweichungen der Genexpressionslevel zwischen zwei Proben zu unterdrücken, verwendet man **Normalisierungsmethoden**.

Normalisierung

1. Schritt einer Normalisierung: wähle einen Satz von Genen, für die ein Expressionsverhältnis von 1 erwartet wird (z.B. House keeping-Gene).

2. Berechne eine Normierungsfaktor aus der beobachteten Variabilität der Expression für diese Gene.

3. Wende diesen Normierungsfaktor auf die Expressionswerte der anderen Gene an.

Wichtig: durch diese Normierung werden die Daten verändert.

Normalisierung

Man nimmt grundsätzlich an, dass die absolute Menge an RNA in beiden Messungen (bzw. Proben) dieselbe ist und dass die selbe Anzahl an RNA-Molekülen in beiden Messungen mit dem Microarray hybridisieren.

Dann sollten auch die Hybridisierungsintensitäten der beiden Gen-Mengen gleich sein.

Berechne Normierungsfaktor

$$N_{total} = \frac{\sum_{k=1}^{N_{gene-set}} R_k}{\sum_{k=1}^{N_{gene-set}} G_k}$$

Reskaliere die Intensitäten, so dass

$$G_k^{'} = G_k \times N_{total}$$

und

$$R_k^{'} = R_k$$

D.h. nur die Grünwerte werte skaliert, die Rotwerte bleiben unverändert.

Expressionsverhältnis

Der relative Expressions-Wert eines Gens kann als Menge an rotem oder grünen Licht gemessen werden, die nach Anregung ausgestrahlt wird.

Man drückt diese Information meist als **Expressionsverhältnis** T_k aus:

$$T_k = \frac{R_k}{G_k}$$

Für jedes Gen k auf dem Array ist hier R_k der Wert für die Spot-Intensität für die Test-Probe und G_k ist die Spot-Intensität für die Referenz-Probe.

Man kann entweder absolute Intensitätswerte verwenden, oder solche, die um den mittleren Hintergrund (Median) korrigiert wurden.

In letzterem Fall lautet das Expressionsverhältnis für einen Spot:

$$T_{median} = \frac{R_{median}^{spot} - R_{median}^{background}}{G_{median}^{spot} - G_{median}^{background}}$$

M. Madan Babu, An Introduction to Microarray Data Analysis

Bereich der Expressionsverhältnisse

Das Expressionsverhältnis stellt auf intuitive Art die Änderung von Expressions-Werten dar. Gene, für die sich nichts ändert, erhalten den Wert 1.

Allerdings ist die Darstellung von Hoch- und Runterregulation nicht balanciert.

Wenn ein Gen um den Faktor 4 hochreguliert ist, ergibt sich ein Verhältnis von 4.

$$R/G = 4G/G = 4$$

Wenn ein Gen jedoch um den Faktor 4 runterreguliert ist, ist das Verhältnis 0.25.

$$R/G = R/4R = 1/4.$$

D.h. Hochregulation wird aufgebläht und nimmt Werte zwischen 1 und unendlich an, während die Runterregulation komprimiert wird und lediglich Werte zwischen 0 und 1 annimmt.

M. Madan Babu, An Introduction to Microarray Data Analysis

Logarithmische Transformation

Eine bessere Methode zur Transformation ist, den Logarithmus zur Basis 2 zu verwenden.

d.h. $\log_2$(Expressionsverhältnis)

Dies hat den großen Vorteil, dass Hochregulation und Runterregulation gleich behandelt werden und auf ein kontinuierliches Intervall abgebildet werden.

Für ein Expressionsverhältnis von 1 ist $\log_2(1) = 0$, das keine Änderung bedeutet.

Für ein Expressionsverhältnis von 4 ist $\log_2(4) = 2$,

für ein Expressionsverhältnis von 1/4 ist $\log_2(1/4) = -2$.

Für die **logarithmierten Daten** ähneln die Expressionsraten dann oft einer **Normalverteilung** (Glockenkurve).

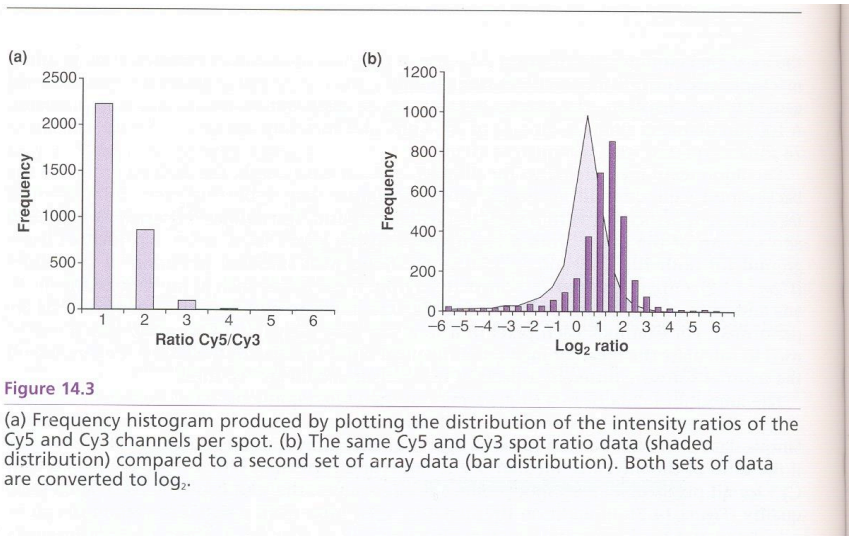


Figure 14.3

(a) Frequency histogram produced by plotting the distribution of the intensity ratios of the Cy5 and Cy3 channels per spot. (b) The same Cy5 and Cy3 spot ratio data (shaded distribution) compared to a second set of array data (bar distribution). Both sets of data are converted to $\log_2$.

M. Madan Babu, An Introduction to Microarray Data Analysis

Orengo-Buch

Daten-Interpretation von Expressionsdaten

Annahme:

Funktionell zusammenhängende Gene sind oft ko-exprimiert.

Z.B. sind in den 3 Situationen

X ➔	Y	(Transkriptionsfaktor X aktiviert Gen Y)
Y ➔	X	(Transkriptionsfaktor Y aktiviert Gen X)
Z ➔	X, Y	(Transkriptionsfaktor Z aktiviert Gene X und Y)

die Gene X und Y ko-exprimiert.

Durch Analyse der Ko-Expression (beide Gene an bzw. beide Gene aus) kann man also funktionelle Zusammenhänge im zellulären Netzwerk entschlüsseln.

Allerdings nicht die kausalen Zusammenhänge, welches Gen das andere reguliert.

Hierarchisches Clustering zur Analyse von Ko-Expression

Man unterscheidet beim Clustering zwischen anhäufenden Verfahren (**agglomerative clustering**) und teilenden Verfahren (**divisive clustering**).

Bei den anhäufenden Verfahren, die in der Praxis häufiger eingesetzt werden, werden schrittweise einzelne Objekte zu Clustern und diese zu größeren Gruppen zusammengefasst, während bei den teilenden Verfahren größere Gruppen schrittweise immer feiner unterteilt werden.

Beim Anhäufen der Cluster wird zunächst jedes Objekt als ein eigener Cluster mit einem Element aufgefasst.

Nun werden in jedem Schritt die jeweils einander nächsten Cluster zu einem Cluster zusammengefasst.

Das Verfahren kann beendet werden, wenn alle Cluster eine bestimmte Distanz zueinander überschreiten oder wenn eine genügend kleine Zahl von Clustern ermittelt worden ist.

k-means Clustern

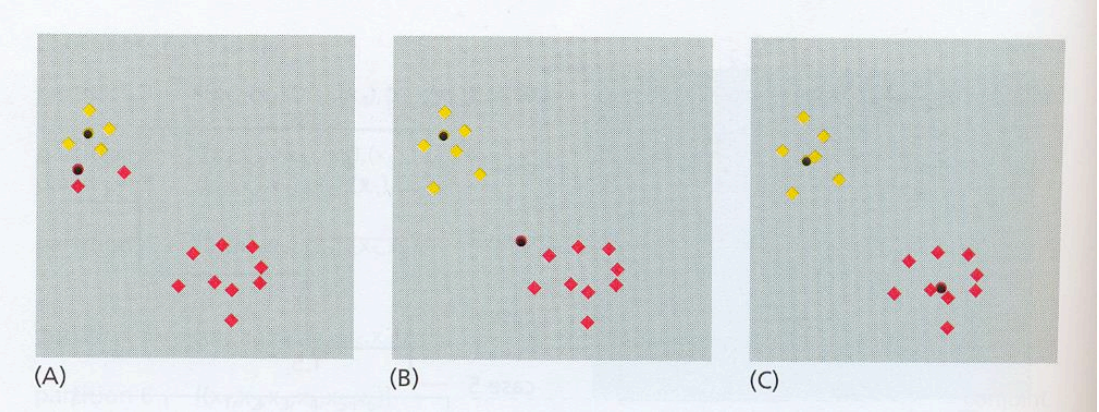
Ein Durchlauf der *k*-means Clustering Methode erzeugt eine Auftrennung der Datenpunkte in *k* Cluster. Gewöhnlich wird der Wert von *k* vorgegeben.

Zu Beginn wählt der Algorithmus *k* Datenpunkte als Centroide der *k* Cluster. Anschließend wird jeder weitere Datenpunkt dem nächsten Cluster zugeordnet.

Nachdem alle Datenpunkte eingeteilt wurden, wird für jedes Cluster das Centroid als Schwerpunkt der in ihm enthaltenen Punkte neu berechnet.

Diese Prozedur (Auswahl der Centroide - Datenpunkte zuordnen) wird so lange wiederholt bis die Mitgliedschaft aller Cluster stabil bleibt.

Dann stoppt der Algorithmus.



Zusammenfassung

Die Methode der Microarrays erlaubt es, die Expression aller möglichen kodierenden DNA-Abschnitte eines Genoms experimentell zu testen.

Die **Zwei-Farben-Methode** ist weit verbreitet um differentielle Expression zu untersuchen.

Aufgrund der natürlichen biologischen Schwankungen müssen die Rohdaten **prozessiert** und *normalisiert* werden.

Durch **Clustering** von Experimenten unter verschiedenen Bedingungen erhält man Gruppen von **ko-exprimierten Genen**.

Diese haben vermutlich **funktionell** miteinander zu tun.