V6 Homologie-basierte Proteinmodellierung

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

  Twighlight-Zone

  Sequenzprofil

- **Lernziele:**

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

  (2) wie gut (genau) sind Homologiemodelle?

Letzte Woche haben wir uns Methoden angeschaut, mit denen man in einer Proteinsequenz die Sekundärstrukturelemente vorhersagen kann.

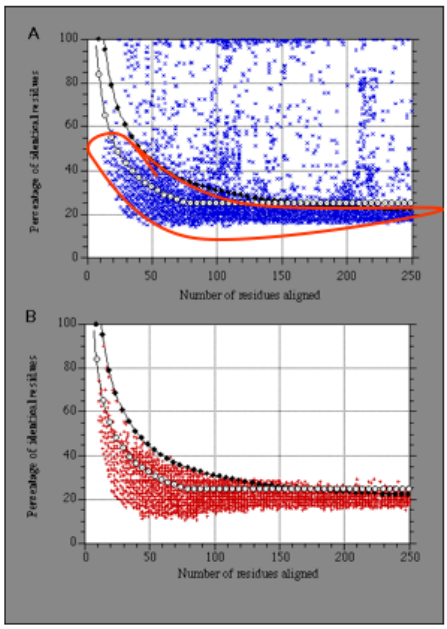
In dieser sechsten Vorlesung beschäftigen wir uns heute mit der Vorhersage der dreidimensionalen Struktur von Proteinen, von denen wir nur die Sequenz kennen.

Zwei wichtige Methode dazu ist die sogenannte Threading-Methode und die Homologie-Modellierung.

1 Twilight Zone

Die schwarzen Diamant-Symbole kennzeichnen eine Kurve. Die rot umrandete Region unter dieser Kurve wird als „**Twilight Zone**“ bezeichnet.

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)

Diese Folie legt die Grundlage für die Homologie-Modellierung, hat aber auch Relevanz für Threading. Es geht um den Vergleich zweier Proteinsequenzen A und B, deren 3D-Strukturen bekannt sind. Die Frage ist, bei welchem Prozentanteil Sequenzidentität zwischen A und B die Strukturen von A und B ebenfalls ähnlich zueinander sind. In beiden Abbildungen ist auf der x-Achse die Länge des Sequenzalignments aufgetragen, d.h. die Anzahl an Residuen, die aufeinander abgebildet werden kann. Auf der y-Achse ist die Sequenzidentität zwischen A und B aufgetragen.

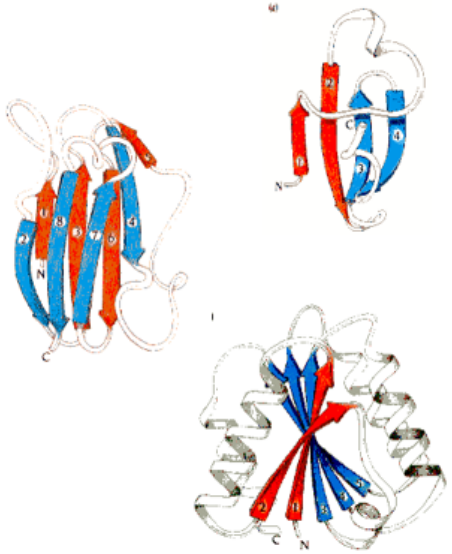
Die obere Abbildung (blaue Kreuze) zeigt eine Statistik von Proteinpaaren A und B mit ähnlicher Struktur, die untere Abbildung (rote Kreuze) enthält nur Proteinpaare mit unterschiedlichen Strukturen. Die untere Abbildung zeigt, dass Proteinpaare mit unterschiedlichen Strukturen nie eine Sequenzidentität oberhalb der schwarzen Linie aufweisen.

Die obere Abbildung zeigt, dass Proteinpaare mit ähnlicher Struktur sowohl hohe wie auch niedrige Sequenzidentität aufweisen können. Die Region unterhalb der schwarzen Linie bezeichnet man als „twilight zone“ (Zwielicht), über die man nichts aussagen kann. Sobald ein Proteinpaar eine höhere

Sequenzidentität besitzt, haben die beiden Proteine mit äußerst hoher Wahrscheinlichkeit eine ähnliche Struktur. Nur für solche Proteinpaare werden wir die Homologiemodellierung anwenden. Man sieht weiterhin, dass das Mindestmass an Sequenzähnlichkeit für kurze Alignments höher ist (ca. 40% bei 50 alignierten Positionen) als bei langen Alignments (ca. 25% bei 250 Positionen).

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.



6. Vorlesung WS 2021/22 Softwarewerkzeuge 3

Threading ist eine alternative Methode zur Homologie-Modellierung, die auch schon bei geringerer Sequenzidentität ganz brauchbare Ergebnisse liefern kann. Der Ausdruck „threading" bedeutet z.B. einen Faden durch ein Loch zu fädeln. Hier werden eine gesamte Proteinsequenz durch eine Strukturvorlage („fold") hindurchfädeln. Genau genommen werden wir die Sequenz durch eine repräsentative Menge aller etwa 2000 bekannten folds (= Proteinarchitekturen) hindurchfädeln. Auf der rechten Seite sind beispielhaft 3 davon gezeigt. Wenn man die Sequenz hindurchfädelt, muss man außerdem theoretisch noch all deren Aminosäuren als mögliche Startpositionen in Betracht ziehen. Jede dieser Möglichkeiten wird dann mit einer energetischen Bewertungsfunktion bewertet, ob das resultierende Proteinstrukturmodell typische Eigenschaften eines Proteins besitzt, d.h. ob die innenliegenden Bereiche vorwiegend hydrophob sind und die außenliegenden Bereiche eher hydrophob sind und was für Aminosäurekontakte im Inneren des Proteins existieren.

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*:

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

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)

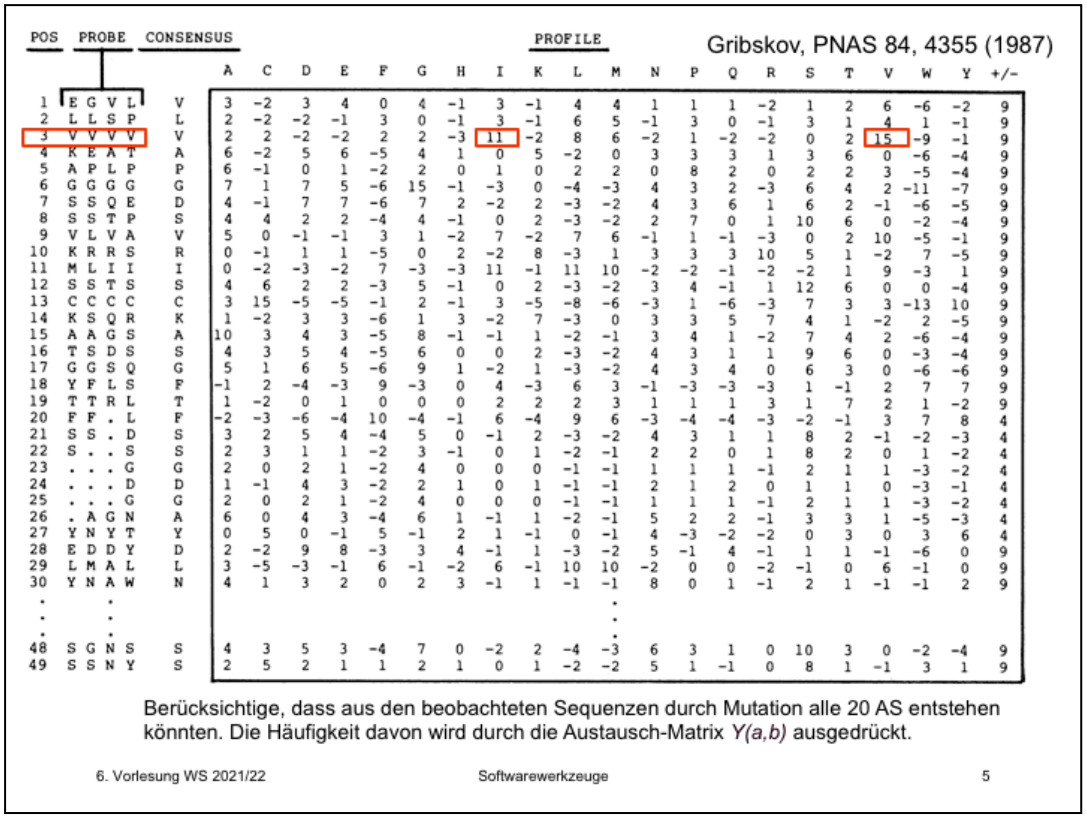
$$M(p,a) = \Sigma_{b=1}^{20} W(p,b) \times Y(a,b),$$

Gribskov, PNAS 84, 4355 (1987)

Für die Threading-Methode werden wir Sequenzprofile verwenden. Ein solches Profil haben wir bereits bei PSIBLAST und PSIPRED kennengelernt. Ein Profil entspricht einer Matrix mit den Dimensionen 21 (für die 20 Aminosäuren plus Gap) mal der Länge der Sequenz. Für jede Position drücken die Einträge in dieser Reihe/Spalte die Wahrscheinlichkeit aus, mit der diese Aminosäure an dieser Position auftreten kann. Falls wir diese Statistik aus „ganz vielen“ verwandten Sequenzen berechnen könnten, bekämen wir daraus eine gute Abschätzung über diese Wahrscheinlichkeiten. Manchmal gibt es jedoch nicht viele verwandten Sequenzen. Dann bekämen wir nur für die wenigen darin auftretenden Aminosäuren eine Aussage über deren Häufigkeit an dieser Position, für die anderen Aminosäuren wäre die Häufigkeit = 0. Davon einen Logarithums (observed/expected) zu nehmen, ist nicht definiert.

Welche Aminosäuren würden in weiteren verwandten Proteinsequenzen auftreten, die es evtl. noch geben könnte, die aber noch nicht sequenziert wurden? Dazu können wir einfach eine Aminosäure-Austauschmatrix verwenden. Damit kann man abschätzen, dass die tatsächlich an Position p beobachteten Aminosäuren

(deren Anzahl ist n(p,b) bzw. die normalisierte Wahrscheinlichkeit W(p,b) mit den Austauschwahrscheinlichkeiten Y(a,b) in andere Aminosäuren a mutieren würden. Genau dies wird durch die untere Formel ausgedrückt.



PROFILE — Gribskov, PNAS 84, 4355 (1987)

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

6. Vorlesung WS 2021/22 Softwarewerkzeuge 5

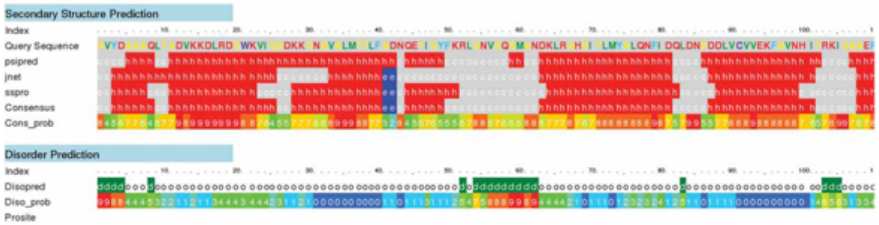
Dies ist ein Sequenzprofil, das aus 4 beobachteten, verwandten Sequenzen (unterhalb von „PROBE“) berechnet wurde. Man könnte aus den 4 Sequenzen auch einen Konsensus-Sequenz generieren. Wie bei dem Beispiel der Transkriptionsfaktorbindemotive in V4 würde man dadurch aber sehr viel Information verlieren.

Schauen wir uns mal die 3. Position an. Dort kommen in den 4 Sequenzen nur Valine vor. Entsprechend erhält V im Profil den höchsten Eintrag (15, rot markiert). Die verwandte Aminosäuren Isoleucin erhält die zweithöchste Bewertung (11), da sie relativ häufig durch Mutation erzeugt werden könnte.

4 Methode zur Fold-Erkennung: Phyre2 webserver

- Webserver verwendet repräsentative Bibliothek für bekannte folds
- Lese Eingabesequenz mit unbekannter Struktur ein
- 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)

Der Webserver Phyre2 stammt aus der bekannten Bioinformatik-Gruppe von Prof. Michael Sternberg am Imperial College London.

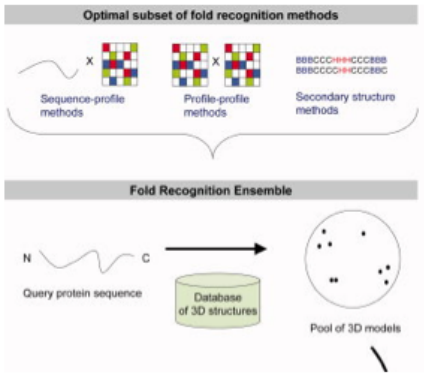
Für eine Eingabesequenz wird mit PsiBlast nach verwandten Sequenzen gesucht. Aus diesen Sequenzen wird dann ein Sequenzprofil für die Eingabesequenz berechnet.

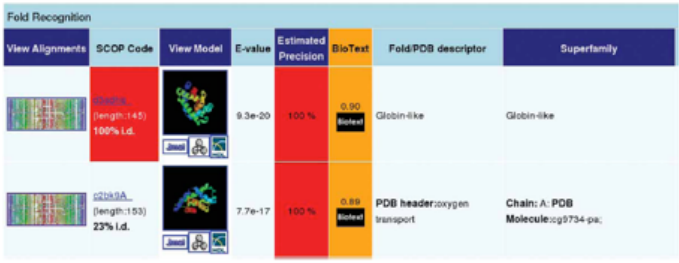
Ausserdem wird mit verschiedenen Methoden jeweils eine Sekundärstrukturvorhersage gemacht und daraus ein Konsensus gebildet, siehe Abbildung oben.

Mit 2 Tools werden ungeordnete Regionen vorhergesagt.

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)

6. Vorlesung WS 2021/22 Softwarewerkzeuge 7

Für eine bekannte 3D-Struktur ist natürlich bekannt, wo deren Sekundärstrukturelemente liegen. Somit kann man für jede Strukturvorlage („fold“) die Passung der vorhergesagten Sekundärstrukturelemente für die Eingabesequenz auf die Elemente der Vorlage berechnen. Ausserdem erzeugt man ebenfalls für die 3D-Vorlage ein Sequenzprofil. Dann bildet man das Sequenzprofil der Eingabesequenz auf das Sequenzprofil der 3D-Vorlage ab. Dies bezeichnet man als Profil-Profil-Alignment.

Diese Schritte werden für alle 3D-Vorlagen in der Datenbank durchgeführt. Die 10 besten Passungen werden ausgegeben.

5 Homologie-basierte Proteinmodellierung (SwissModel)

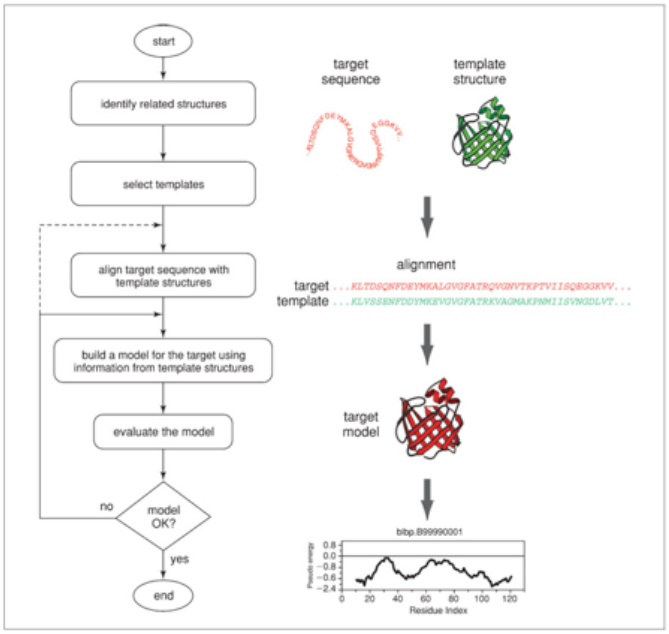
- **Methode:** Ebenfalls wissensbasierter Ansatz.
- **Erfordernis: Mindestens** 1 bekannte 3D-Struktur eines verwandten Proteins,
- **Prozedur:**
- finde Proteine bekannter Struktur, die zu Inputsequenz verwandt sind.
- Erzeugung eines multiplen Sequenzalignments mit der Zielsequenz.
- Generierung eines **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.

Nun besprechen wir im Rest dieser Vorlesung die Methode der Homologiemodellierung. Diese wird auch Inhalt des zweiten Projekts sein. Gegeben sei eine Proteinsequenz A mit unbekannter Struktur. Man sucht nun nach ähnlichen Sequenzen A‘, A‘‘, etc, deren 3D-Strukturen bekannt sind. Wenn die Sequenzidentität hoch genug ist (siehe Folie 2 zur Twilight-Zone) kann man mit hoher Sicherheit davon ausgehen, dass die Struktur von A sehr ähnlich zu den bekannten Strukturen A‘ und A‘‘ sein wird. Die Homologiemodellierungsmethode geht schrittweise vor, gewissermaßen von grob nach fein. Wir werden die Schritte nun im Einzelnen besprechen.

Homologie-basierte Proteinmodellierung (Modeller)

Andrej Sali, UCSF

http://salilab.org



Eswar, Curr. Protocols in Bioinf. (2006)

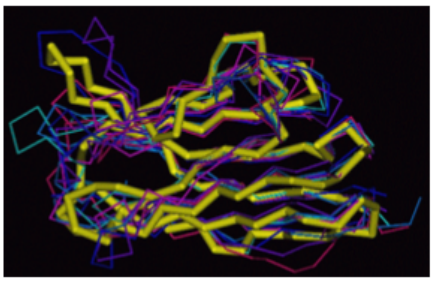
Die Gruppe von Andrej Sali an der UCSF entwickelt das verbreitete Tool Modeller. Der abgebildete Workflow ist sehr ähnlich zu dem des Tools Swissmodel, das wir im Folgenden besprechen werden und das das Arbeitsinstrument im zweiten Projekt sein wird.

3D Framework für die neue Sequenz

(a) Zunächst wird mit Sequenzalignments eine optimale Vorlage-Struktur ausgewählt. Meist diejenige mit größter Sequenzidentität. Allerdings spielt auch eine Rolle, ob ein Ligand und Ko-Faktoren gebunden sind und in welchem Zustand man das Eingabeprotein modellieren möchte.

(b) Für alle im Alignment konservierten Aminosäuren, kann man die Atomkoordinaten kopieren. Von mutierten Aminosäuren nur die Atome des Rückgrats. Falls man mehrere Vorlagen verwendet, kann man mittlere Positionen erzeugen.

(c) Seitenketten mit völlig inkorrekter Geometrie (die nicht passen) werden entfernt.



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

Die Auswahl der Vorlage und die Bestimmung des optimalen Alignments zwischen Vorlage und Eingabesequenz sind die wichtigsten Schritte des ganzen Prozesses. Alle Fehler, die man hier einbaut, lassen sich später nicht mehr korrigieren.

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

Die meisten Abweichungen (Insertionen, Deletionen, Mutationen) treten in Loops auf. Dies ist generell zu erwarten, da hier der geringste Selektionsdruck auf die Proteinstruktur herrscht. Die Modellierung von Loopstrukturen ist leider im Allgemeinen recht schwierig.

Beta-Turns

Eine Aminosäurekette kann ihre Richtung dadurch umkehren, dass ein „reverse turn“ durch Bildung einer H-Bindung zwischen C=O und H-N gebildet wird.

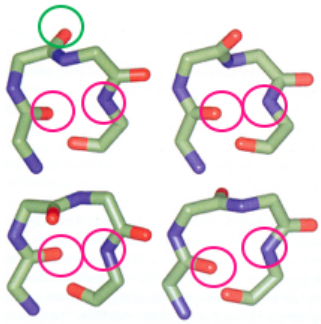


Figure 1.9 The four types of beta turn described in Table 1.1, types I and I′ are shown on the top, types II and II′ on the bottom.

Wenn dies zwischen zwei antiparallelen beta-Strängen geschieht, nennt man diesen eine beta-Haarnadel (hairpin). Es ergeben sich folgende Diederwinkel:

Table 1.1 Turns are regions of the protein chain that enable the chain to invert its direction. The ϕ and ψ angles of some commonly occurring turns are listed.

Turn type	ϕ_1	ψ_1	ϕ_2	ψ_2
I	−60	−30	−90	0
I′	60	30	90	0
II	−60	120	80	0
II′	60	−120	−80	0

[Tramontano book]

6. Vorlesung WS 2021/22 Softwarewerkzeuge 12

In einem Beta-Turn macht die Aminosäureketten innerhalb von 4 Aminosäuren eine enge Wendung um 180 Grad. Kommt Ihnen das bekannt vor? Was ist mit dem saarländischen Symbol, der Saarschleife?

Wie in der Abbildung gezeigt, gibt es 4 mögliche Anordnung der Rückgratatome. Aufgrund der alternierenden H-Bindungen von links nach rechts und von rechts nach links und der relativ planaren Konformation bezeichnet man dieses Strukturelement als Beta-turn. In den beiden linken Strukturen zeigen die pink umkreisten Sauerstoff- und Stickstoffatome „nach hinten“. Der Unterschied zwischen oben und unten ist dann die relative Position des grün umkreisten Sauerstoffatoms, das entweder ebenfalls nach hinten zeigt (oben), oder nach vorne (unten).

Die beiden rechten Konformationen sind analog, wobei dabei die pinken Atome nach vorne zeigen.

Konstruktion fehlender Loops

Ein paar Regeln:

- falls mittellange Loops kompakte Substrukturen bilden, spielt die Ausbildung von Wasserstoffbrücken-bindungen mit den Atomen des Rückgrats die wichtigste Rolle für ihre Konformation.

- falls mittellange Loops ausgedehnte Konformationen haben, ist für ihre Stabilisierung meistens eine hydrophobe Seitenkette verantwortlich, die ins Proteininnere zeigt und zwischen die Sekundärstrukturelemente gepackt ist, zwischen denen der Loop liegt.

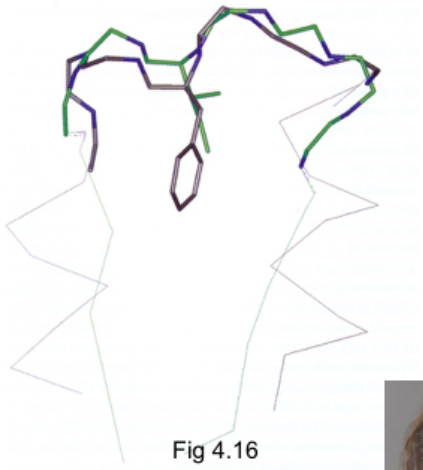


Fig 4.16

Figure 4.16 The figure shows two loops with similar conformations stabilized by the packing of a central hydrophobic amino acid. Note that one of the loops connects two alpha helices and the other two beta strands.

[Tramontano book]

Anna Tramontano (1957-2017)

https://www.nature.com/articles/nsmb.3410

Auch für Loops mittlerer Länge (z.B. 8 – 12 Aminosäuren) gibt es ein paar Erfahrungswerte. In ihrem Buch beschreibt Anna Tramontano, dass solche Loops manchmal eine hydrophobe Aminosäure in der Mitte besitzen. Im Bild sind zwei solche Loops gezeigt. Einer enthält ein Phenylalanin, der andere ein Isoleucin. Die anschließenden Elemente des Loops links und rechts davon sind entweder alpha-helikal oder extended.

Anna Tramontano war eine sehr bekannte Bioinformatikerin, die z.B. sehr aktiv in der Organisation des CASP-Wettbewerbs und in der ISCB war.

Konstruktion fehlender Loops

Sehr ähnliche Konformation dreier Loops mit unterschiedlicher Sequenz.

Zwei Loops enthalten ein *cis*-Prolin.

Die stabilisierenden H-Bindungen werden mit sehr unterschiedlichen Proteingruppen ausgebildet.

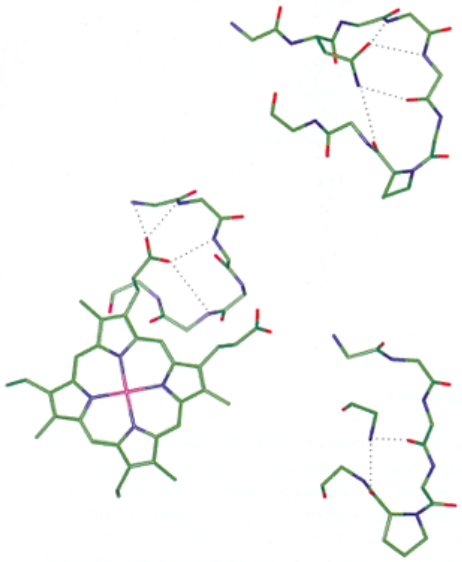


Figure 4.17 The three loops shown in the figure are very similar and stabilized by hydrogen-bonds, however the partners of these interactions are different in the three different proteins (an immunoglobulin, a viral protein, and a cytochrome).

[Tramontano book]

In diesem Fall bilden äußere Loops ähnliche H-Bindungsmuster entweder mit sich selbst (oben rechts), oder mit dazwischenliegenden Gruppen des restlichen Proteins, unten links um eine Hämgruppe herum.

Diese Beobachtung deutet darauf hin, dass es eine gute Strategie ist, eine Bibliothek von in der Protein Datenbank beobachteten Loopkonformationen aufzubauen und diese als Vorlage für Loop-Modelling zu benutzen.

Konstruktion fehlender Loops

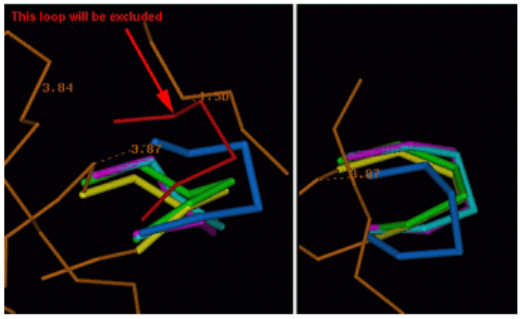
Basierend auf den Verankerungen der Loops

(a) entweder wird 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

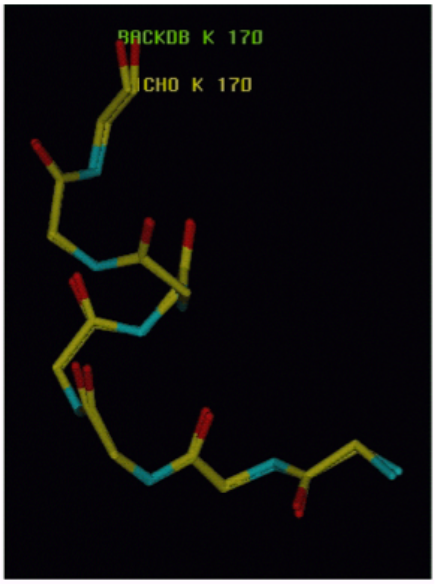
Das Homologie-Modellierungstool Swissmodel verwendet entweder Loop-Vorlagen (siehe (a)), oder konstruiert kurze Loops „de novo“ durch einen kombinatorischen Ansatz (siehe (b)). Letzteres ist jedoch nur für kurze Loops möglich. Wichtig ist natürlich, dass die Enden des Loops jeweils zu den Fortsetzungen in der Proteinstruktur passen. In der Abbildung passen mehrere mögliche Modelle ganz gut, bloß die rote Vorlage passt nicht zu den vorgegebenen Enden.

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

6. Vorlesung WS 2021/22 Softwarewerkzeuge 16

Durch das Alignment von Vorlage und Eingabesequenz kann es vorkommen, dass in dem Rückgrat der Vorlage extra-Residuen eingefügt werden müssen, bzw. herausgeschnitten werden müssen.

In diesem Fall wendet man eine ähnliche Strategie wie bei der Loop-Modellierung an. Entweder verwendet man eine passende Vorlage aus der Proteindatenbank (d.h. dort gibt ein passendes Stück), oder der Abschnitt wird “de novo” durch kombinatorische Suche konstruiert.

Konstruktion unvollständiger/fehlender Seitenketten

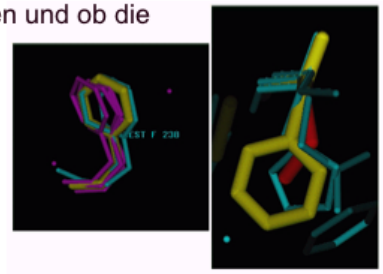
Rotamere: Seitenkettenkonformationen mit niedriger (günstiger) Energie.

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

Swissmodel verwendet 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 Torsionswinkel in erlaubten Bereichen liegen.



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

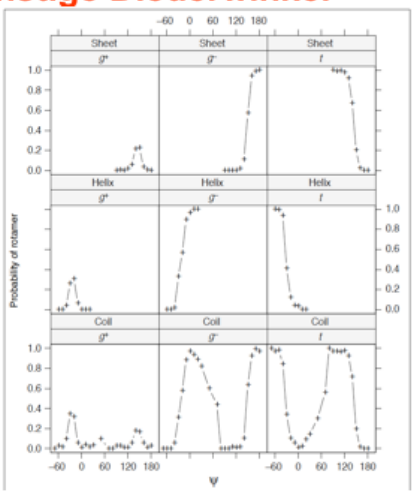
Die Konstruktion der fehlenden Seitenketten, bzw. die Korrektur von konservierten, aber nicht passenden Seitenketten klingt zunächst schwierig, ist aber vermutlich der einfachste Schritt in dem Workflow. Durch statistische Analysen stellte man fest, dass die Winkel der Seitenketten wenige bevorzugte Orientierungen (sogenannte Rotamere) einnehmen. Man braucht also „bloß“ die verschiedenen Möglichkeit kombinatorisch durchzuprobieren. Dieses Problem kann man mittels des Dead End Elimination-Algorithmus (siehe https://www.nature.com/articles/356539a0) sogar auf optimale Weise lösen.

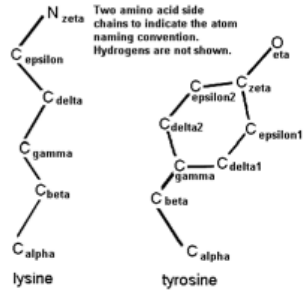
Rotamer-Bibliotheken: günstige Diederwinkel

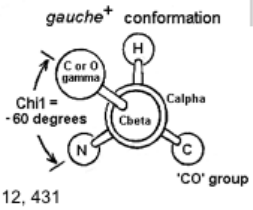
Abbildung rechts und rechts unten:
Günstige χ_1-Drehwinkel der Valin-Seitenkette:
beobachtete Häufigkeit der Rotamere

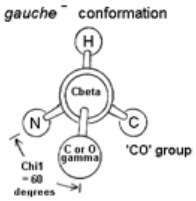
gauche$^+$ ($\chi_1 \sim +60^\circ$)
gauche$^-$ ($\chi_1 \sim -60^\circ$)
trans ($\chi_1 \sim 180^\circ$)

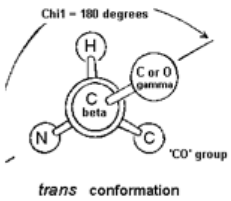
in verschiedenen Sekundärstrukturen
als Funktion des Rückgratsdiederwinkels Ψ.











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

Die Abbildungen links unten zeigen, wie die Atome der Seitenketten von Lysin und Tyrosin benannt werden. Der erste Winkel der Seitenkette beschreibt, welche Orientierung das C_gamma-Atom einnimmt, wenn man durch die Verbindungslinie (Bindung) zwischen C_beta und C_alpha hindurchschaut. C_gamma liegt also vor der Tafelebene, die Atome des Rückgrats (N und CO-Gruppe) dahinter.

C_gamma muss aus energetischen Überlegungen (möglichst geringer sterischer Überlapp) „zwischen“ den 3 Gruppen H-Atom / N-Atom / C-Atom liegen. Es gibt also 3 Einstellungen dafür. Dies gilt eigentlich für jede Aminosäure. Man bezeichnet die Einstellungen als gauche+, gauche- und trans. In der trans-Konformation liegt C_gamma gegenüber von dem Stickstoffatoms des Rückgrats.

Die Abbildung rechts oben illustriert, dass die drei möglichen Einstellungen mit unterschiedlicher Häufigkeit angenommen werden, je nach der Konformation des Rückgrats (Sheet, Helix und Coil von oben nach unten).

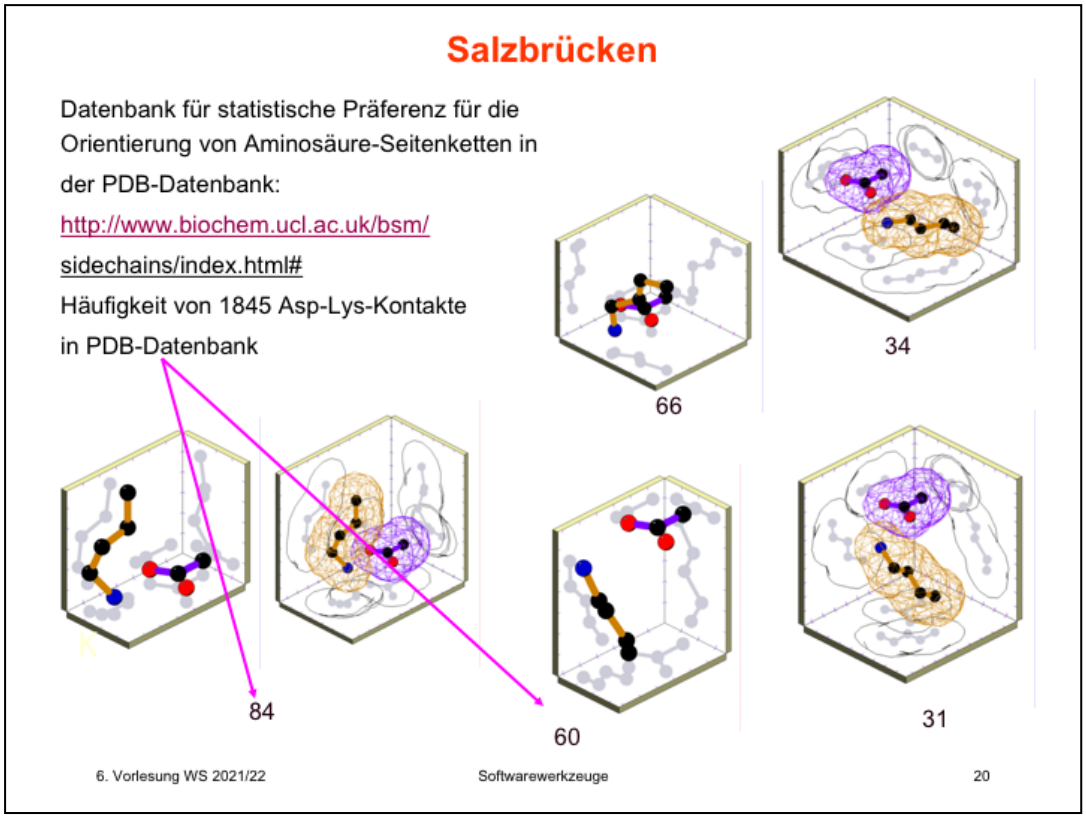
Paarungs-Präferenz von Aminosäuren

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.

Ein wichtiger Punkt bei der Orientierung der Seitenketten ist aber noch, in welch unterschiedlicher Umgebung sind landen.



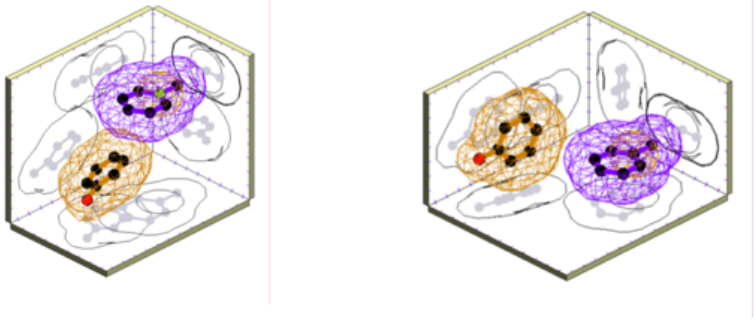
Die Gruppe von Prof. Janet Thornton am EBI hat einen schönen Atlas von Seitenketten-Paarungen erstellt. Hier ist eine Statistik über die relative Orientierung von negativ geladenen Aspartaten (Asp) und positiv geladenen Lysinen (Lys) gezeigt. Das (blaue) Stickstoffatom des Lysins liegt eigentlich immer direkt „vor“ den beiden Carboxyl-Sauerstoffatomen des Aspartats. Unterschiede gibt es lediglich in der Orientierung der Lysin-Seitenkette. Lediglich in der Abbildung links oben liegt Lysin seitlich neben Aspartat.

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.

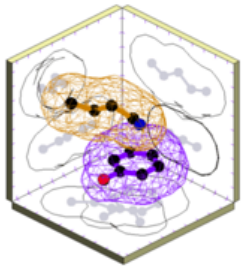
Cluster Phe-Tyr 1 4



Zwei aromatische Ringe (hier von Phenylalanin und Tyrosin) können entweder seitlich versetzt übereinander packen (links) oder in einer T-Konformation (rechts). Dieselben energetisch günstigen Konformationen erhält man auch in quantenchemischen Rechnungen.

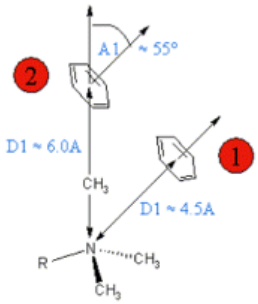
Kationen-Π-Wechselwirkung

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

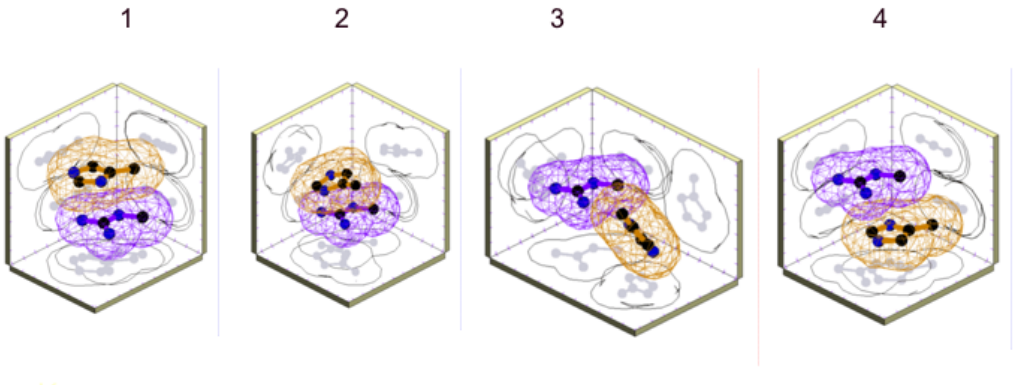
Das positiv geladene Lysin positioniert das Stickstoffatom senkrecht über dem aromatischen Ring des Tyrosins. Dort kann der formal positive Stickstoff gut mit der negativen Pi-Elektronenwolke des aromatischen Rings wechselwirken.

Im unteren Bild ist eine ähnliche Wechselwirkung für die Bindung des Neurotransmitter-Moleküls Acetylcholin (CH3)3-N-R in der Bindungstasche des Enzyms Acetylcholinesterase gezeigt. Der formal positive Stickstoff von Acetylcholin wird durch zwei aromatische Ringe des Proteins koordiniert. Da am Stickstoff noch 3 Methylgruppen hängen, sind die aromatischen Ringe etwas weiter entfernt (4.5 A und 6.0 A).

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



Diese Beispiele zeigen Anordnungen der positiv geladenen Kopfgruppe von Arginin bzgl. des partiell aromatischen Rings von Histidin.

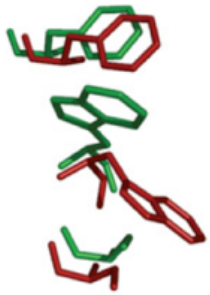
Diese Betrachtungen der Seitenketten-Paarungen werden bei der Homologie-Modellierung zunächst nicht berücksichtigt, sondern erst im Nachhinein bei der energetischen Bewertung der erzeugten Modelle. Wir kommen darauf in Kürze bei dem DOPE-Potential zurück.

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.



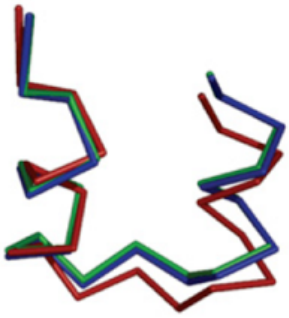
Eswar, Curr. Protocols in Bioinf. (2006)

Welche Fehler können nun bei der Homologie-Modellierung auftreten. Falls es genügend „Platz“ gibt, kann eine Seitenkette (wie hier das grüne Tryptophan) in eine unterschiedliche Orientierung modelliert werden als in der roten Kristallstruktur gefunden wird. An solchen Beispielen (wenn beide Kristallstrukturen von Vorlage und für die Eingabesequenz bekannt sind) kann die Korrektheit der Modellierung überprüfen.

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)

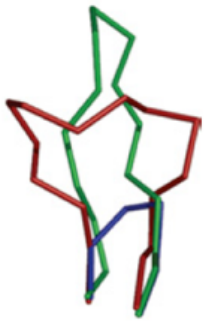
Dies ist eine leicht unterschiedliche Positionierung eines Loops. So etwas ist kein „Fehler“, sondern eine leichte Ungenauigkeit.

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)

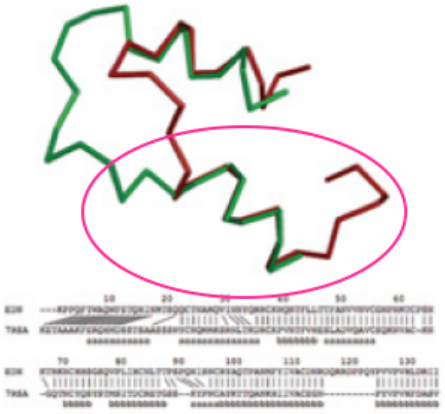
Auch so etwas ist kein Fehler. Die Modellierung musste in die blaue Vorlage aus einem kurzen Loop einen längeren Loop bauen. Daraus wurde das kompakte grüne Modell erstellt. Vielleicht „kannte“ das Tool die Regeln von Anna Tramontano (siehe Folie 13). In der Kristallstruktur ist der Loop jedoch stärker ausgedehnt. Mit solchen Abweichungen muss man stets rechnen.

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).



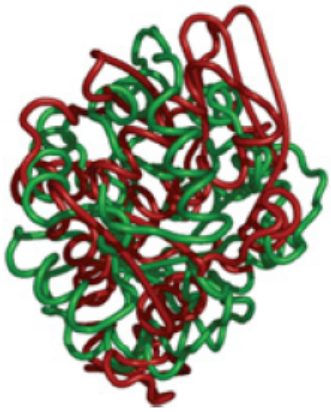
Eswar, Curr. Protocols in Bioinf. (2006)

Dieses ist jedoch eine deutliche Abweichung. Durch ein ungünstiges Sequenzalignment (unten) ergibt sich eine deutliche Verschiebung der pink umrandeten Helix.
Dies hat dann nicht nur Auswirkungen auf diesen Teil der Struktur, sondern kann sich in die anschließenden Bereiche fortsetzen.

Typische Fehler bei Homologie-Modellierung (V)

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

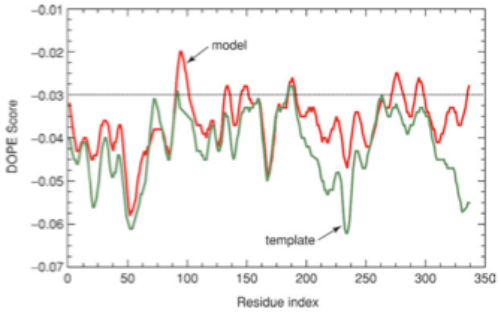
Vergleich der Kristallstruktur für α-trichosanthin (rot) mit dem Modell (grün), das mit Indol-3-Glycerophosphat-Synthase als Vorlage erzeugt wurde..



Eswar, Curr. Protocols in Bioinf. (2006)

Dies ist ein Beispiel für eine Modellierungs-Katastrophe. Falls man eine nicht passende Vorlage gewählt hat (z.B. mit zu geringer Sequenzidentität), kann es passieren, dass Modell und Vorlage überhaupt nichts miteinander gemeinsam haben.

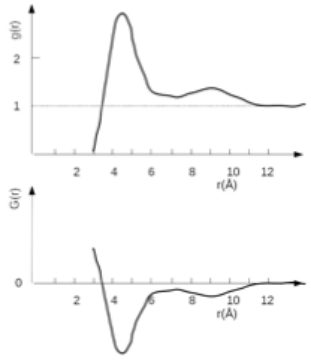
Bewertung von Strukturmodellen (Modeller)



Modeller verwendet das statistische DOPE-Potential (Discrete Optimized Protein Energy) zur 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.



Statistisches Potential:

Aus den in Proteinstrukturen beobachteten radialen Häufigkeiten *g(r)* für Aminosäure-Paare (-> *bei welchem Abstand gibt es mehr als erwartet?*), berechnet man durch „Boltzmann-Inversion“ deren effektive Wechselwirkungsstärke.

$$\frac{p_1}{p_2} = e^{-\frac{E_1 - E_2}{kT}} \longrightarrow G(r) = -k_B T \ln p(r)$$

Eswar, Curr. Protocols in Bioinf. (2006)

6. Vorlesung WS 2021/22 Softwarewerkzeuge 29

Nachdem das Modell erzeugt wurde, möchte man gerne bewerten, ob das Modell wie eine typische Proteinstruktur aussieht, bzw. welche Teile gute Eigenschaften haben. Der oben gezeigte DOPE score basiert auf statistischen Potentialen für alle Aminosäure-Paare. Je günstiger (negativer) die DOPE-Bewertung, desto günstiger ist die Umgebung einer bestimmten Aminosäure. In diesem Beispiel liegt das grüne Profil der Vorlage fast überall unter dem roten Profil des Modells. Das ist nicht schlimm. Ein Bereich des Modells liegt oberhalb eines Schrankenwerts von -0.03, was darauf hindeutet, dass dieser Bereich vielleicht nicht optimal modelliert wurde.

Die mittlere Abbildung zeigt die radiale Verteilungsfunktion zweier Aminosäuren in den bekannten Proteinstrukturen. Diese beiden Aminosäuren haben einen deutlichen Peak bei etwa 4.5 Angstroem Abstand, also in direktem Kontakt. Vermutlich sind dies zwei hydrophobe Aminosäuren, die häufig im Proteinkern in kurzen Abständen zueinander auftreten.

Durch Invertierung der bekannten Boltzmann-Formel kann man aus der Verteilungsfunktion p(r) (d.h. der Häufigkeitsfunktion) eine freie Enthalpie bei verschiedenen Abständen definieren (untere Abbildung).

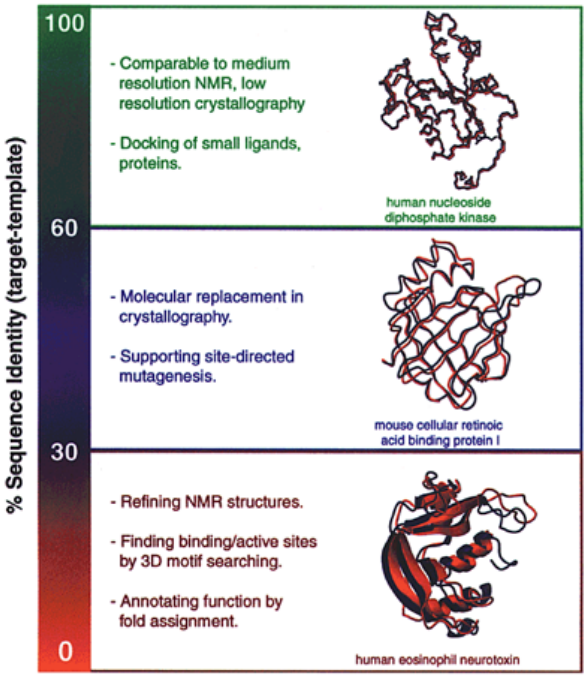
Der Peak in der Häufigkeit wird zu einem Minimum der freien Enthalpie und umgekehrt.

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)



Diese Folie gibt einen Überblick, bei welchem Grad an Sequenzidentität man Homologiemodelle sinnvoll einsetzen kann.

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

Kein Kommentar.

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

Kein Kommentar.

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.

Die Passung zweier Strukturen kann man mit dem RMSD-Wert bewerten. Dies funktioniert dann, wenn beide Proteine exakt gleich viele Atome (bzw. genau gleich viele C-alpha-Atome) enthalten. Ansonsten kann man eben nur die Teile der Strukturen miteinander vergleichen, die in beiden Proteinen vorkommen.

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

In diesem bereits älteren Benchmark wurde die Genauigkeit von Homologiemodellen überprüft. Man erstellt also Homologiemodelle für Proteine, deren Strukturen man kennt und kann dann die Abweichungen zwischen Modell und tatsächlicher Struktur als RMSD messen.

CAMEO: Continuous Automated Model Evaluation

Wettbewerb: Für die innerhalb von 5 Wochen neu bei der PDB eingereichten Kristallstrukturen von Proteinen werden aufgrund von deren Sequenzen mit automatisierten Homology-Modeling-Webservern Strukturmodelle erzeugt und mit den experimentellen Strukturen verglichen.

CAMEO verwendet für die Bewertung keine RMSD-Abweichungen, da deren Werte stark durch Rotationen zwischen Proteindomänen beeinflusst werden.

Stattdessen werden lDDT-scores (local distance difference test) berechnet:

- Betrachtet werden alle paarweisen Distanzen zwischen Atomen (< 15 Å).
- Getestet wird, wie viele Distanzen um weniger als 0.5 Å, 1 Å, 2 Å und 4 Å abweichen.
- Daraus wird der durchschnittliche Prozentanteil gebildet.

Haas et al. Proteins 86, 387-398 (2017)

Heutzutage gibt es eine kontinuierliche, automatische Qualitätskontrolle von Homologie-Modellierungs-Webservern. Wenn eine neue Kristallstruktur bei der PDB eingereicht wird, wird die Sequenz an verschiedene Webserver geschickt um mit diesen automatisch ein Modell für dasselbe Protein zu erstellen.

Man ist davon abgekommen, Strukturen per RMSD miteinander zu vergleichen, da dies bei großen Proteinen nicht gut funktioniert, wenn diese aus mehreren Domänen bestehen. Wenn diese zwischen Modell und richtiger Struktur nur leicht gegeneinander verdreht sind, erhält man gleich recht hohe RMSD-Unterschiede, obwohl die einzelnen Domänen evtl. sehr gut übereinstimmen.

Stattdessen werden Abständen zwischen Atompaaren gemessen und ein mittlerer Anteil berechnet, welche Anteile des Modells um wieviel von der korrekten Struktur abweichen.

CAMEO: Continuous Automated Model Evaluation

TABLE 2 CAMEO IDDT based ranking for the time frame 2016-05-01 to 2016-07-30[a]

Server Name	IDDT (all Targets)
Robetta	65.3 ± 16.45
RaptorX	63.8 ± 16.57
IntFOLD3-TS[c]	62.2 ± 17.53
IntFOLD4-TS	62.0 ± 16.32
SWISS-MODEL	56.5 ± 22.49
SPARKS-X	56.3 ± 18.25
Princeton_template	55.6 ± 15.68
IntFOLD2-TS[c]	55.1 ± 17.47
HHpredB[b]	47.6 ± 17.41
M4T[c]	45.1 ± 16.74
Phyre2[c]	44.5 ± 23.135
NaiveBLAST[c]	43.3 ± 25.59
RBO Aleph[b]	38.6 ± 16.28

TABLE 4 Average response time from submission of the sequence to reception of structure prediction by CAMEO[a]

Server name	Avg. response time (hh:mm:ss)
HHpredB[b]	00:21:38
SWISS-MODEL	00:27:35
SPARKS-X	01:29:17
Phyre2[c]	02:14:08
Princeton_template	03:22:15
NaiveBLAST[c]	04:07:31
M4T[c]	08:20:48
RaptorX	13:53:17
RBO Aleph[b]	16:12:39
IntFOLD2-TS[c]	28:09:28
IntFOLD3-TS[c]	28:23:14
Robetta	29:03:28
IntFOLD4-TS	36:42:55

Robetta macht die besten Modelle, benötigt dafür jedoch aufgrund von aufwändigem Konformationssampling sehr lange.

Haas et al. Proteins 86, 387-398 (2017)

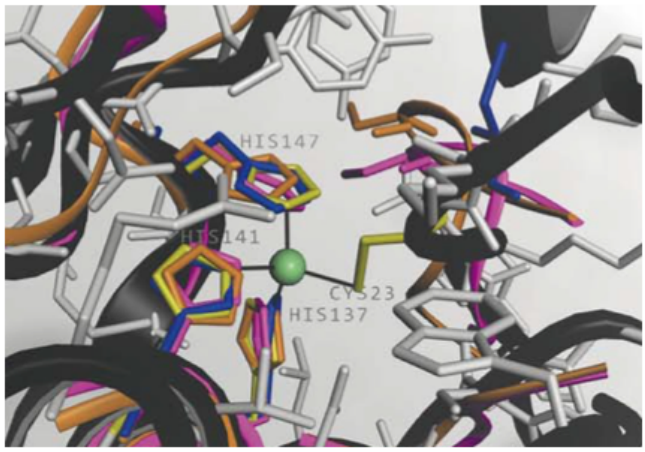
Der Server Robetta macht gemäß des IDDT-Scores die besten Modelle (65,3), braucht dafür aber im Mittel 29 Stunden. Demgegenüber steht Swiss-Modell (IDDT 56,5), das nur 27 Minuten pro Struktur benötigt.

CAMEO: Continuous Automated Model Evaluation

Kristallstruktur für Myroilysin (PDB ID 5CZW, schwarz).

Homologiemodelle mit SWISS-MODEL (orange), IntFOLD4-TS (blau), Sparks-X (magenta).

In der Mitte liegt ein Zink-Ion (grüne Kugel), das von 4 Residuen (3 His, 1 Cys) koordiniert wird (die gelben Residuen stammen aus der Kristallstruktur).



Alle Tools machen gute Vorhersagen für die drei Histidin-Residuen 137, 141 und 147, die das Zink-Ion koordinieren, weichen jedoch für Cystein 23 voneinander ab.

Haas et al. Proteins 86, 387-398 (2017)

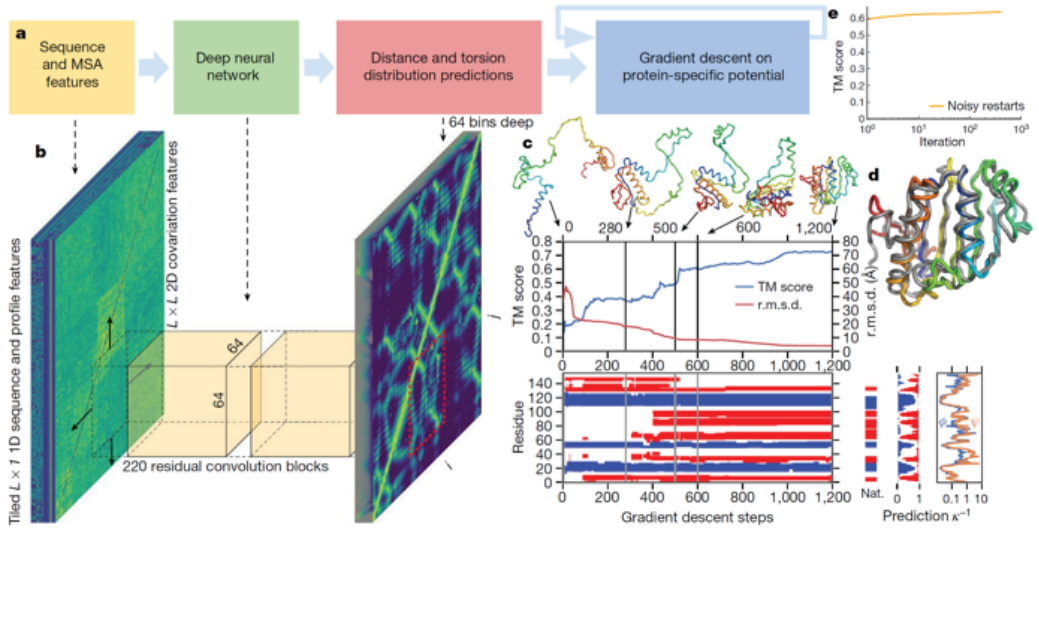
Dieses Beispiel vergleicht die Vorhersagen von verschiedenen Homologiemodellierungs-Servern für die Koordinierung des grünen Zinkatoms. His137, 141 und 147 liegen sehr gut aufeinander, für das rechts liegende Cys23 gibt es jedoch deutliche Unterschiede. Nur die gelbe Kristallstruktur zeigt, dass das Cystein der vierte Ligand des Zinkatoms ist.

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

Oft möchte man Homologiemodelle verwenden, um damit Liganden-Docking zu machen. Hier werden verschiedene Ansätze vorgestellt, wie man hierbei vorgehen kann.

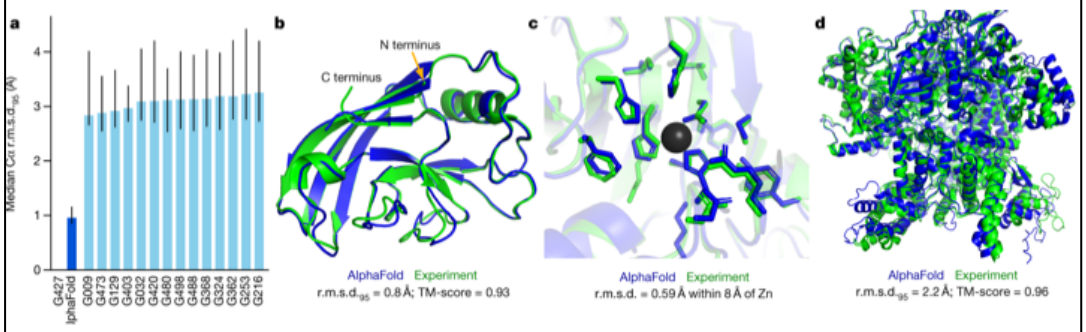
AlphaFold von Google



Senior et al. Nature 577, 706–710 (2020)

Kein Kommentar.

Resultate von AlphaFold2



a, The performance of AlphaFold on the CASP14 dataset (n = 87 protein domains) relative to the top-15 entries (out of 146 entries), group numbers correspond to the numbers assigned to entrants by CASP..
b, Our prediction of CASP14 target T1049 (PDB 6Y4F, blue) compared with the true (experimental) structure (green).
c, CASP14 target T1056 (PDB 6YJ1). An example of a well-predicted zinc-binding site (AlphaFold has accurate side chains even though it does not explicitly predict the zinc ion).
d, CASP target T1044 (PDB 6VR4)—a 2,180-residue single chain—was predicted with correct domain packing.

Jumper et al. Nature **596**, 583–589 (2021)

Kein Kommentar.

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.

Kein Kommentar.