Bioinformatik von RNA- und Proteinstrukturen

Inhaltsverzeichnis

1 Formale Sprachen

Formale Sprache[1] L über Alphabet Σ
$L \subseteq \Sigma^*$
mit Σ^* = Kleensche Hülle[2] von Σ

$$\Sigma^* = \bigcup_{n=0}^{\infty} \Sigma^n$$

$\Sigma^0 = \{\varepsilon\}, \Sigma^1 = \Sigma, \Sigma^2 = \Sigma \times \Sigma$
$\varepsilon \rightarrow$ leeres Wort (leere Menge)

Beispiel: $\Sigma = \{a\}, \Sigma^* = \{\varepsilon, a, aa, aaa, ...\}, L = \{a, aa, aaaa, ...\}$

1.1 formale Grammatik G

G = (N, Σ, P, S) mit

- N = Nichtterminale
- Σ = Alphabet
- P = Produktionsregeln
- S = Startsymbol (ϵ N)

$P \subseteq (N \cup \Sigma)^*/N(N \cup \Sigma)^* \rightarrow (N \cup \Sigma)^*$

Beispiel:
G=({S}, {a}, $\{S \rightarrow aaS, S \rightarrow a\}$, S)
führt zu: S $\rightarrow$ aaS $\rightarrow$ aaa

1.2 Klassifikation von formalen Sprachen

durch die Comsky-Hierarchie[3]:

- Typ 0 = rekursiv auszählbar ($\alpha N \beta \rightarrow \gamma$)
- Typ 1 = kontext-sensitiv ($\alpha N \beta \rightarrow \alpha \gamma \beta$)
- Typ 2 = kontext-frei, N $\rightarrow (N \cup \Sigma)^* \rightarrow$ stochstisch kontextfreie Grammatik (SCFG) $\rightarrow$ Dynamics Programming

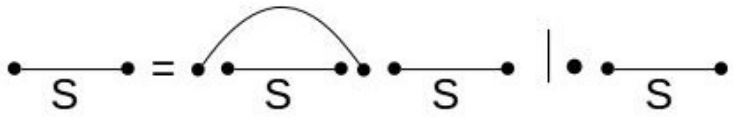
[1] https://de.wikipedia.org/wiki/Formale_Sprache
[2] https://de.wikipedia.org/wiki/Kleenesche_und_positive_H%C3%BClle
[3] https://de.wikipedia.org/wiki/Chomsky-Hierarchie

- Typ 3 = regular ($N \rightarrow \Sigma|\Sigma N$) $\rightarrow$ dann immer Hidden Markov Model (HMM) modellierbar

bei Alignments: 

Erweiterung mit Wahrscheinlichkeit: G=(N, Σ, P, S, Ω)
mit Ω = Wahrscheinlichkeit für Produktionsregeln

jetzt auf RNA-Vorhersagen:

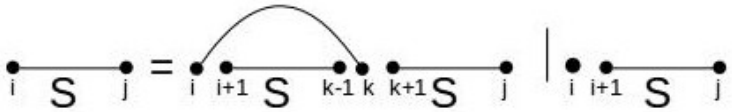


scoring scheme: Bewertung von σ (◠) = 1, (σ (⟶)), σ (•) = 0
scoring function:

- max Basepairs: + (Summe),
- Anzahl der Strukturen: $\cdot$ (Multiplikation)

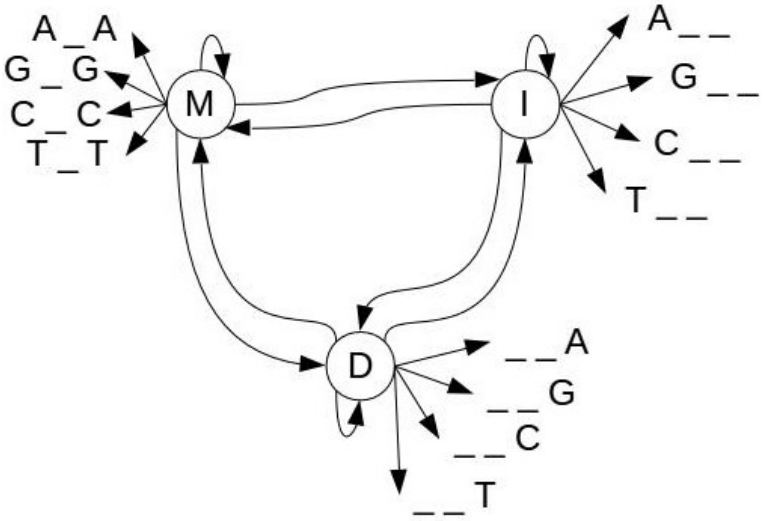
choice function:

- max Basepairs: max,
- Anzahl der Strukturen: + (Summe)



$$S_{ij} = \begin{cases} S_{i+1,j} + \sigma((\bullet) \\ S_{i+1,k-1} + S_{k+1,j} + \sigma((\frown) \end{cases}$$

1.3 Hidden Markov Model



M: Match, I: Insertion, D: Deletion

Grammatik:

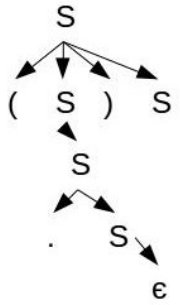
- M $\rightarrow M_{A\_A}|...|I|D$
- I $\rightarrow I_{A\_\_}|...|D|M$
- D $\rightarrow D_{\_\_A}|...|M|I$

Beispiel:

Faltungsgrammatik

S $\rightarrow (S)S|.S|\varepsilon$
Nichtterminale = S, Alphabet = {(,), .}

Beispiel in Baumdarstellung:



weiteres Beispiel: Sankoff, Kombination von zwei Grammatiken (Alignment und Faltung)

Alignmentgrammatik

S $\rightarrow .S|\_S|\varepsilon$
G $= (N = \{S\}, \Sigma = \{., \_\}, P = \{S \rightarrow .S|\_S|\varepsilon\}, S)$
Alignment: $G^2 = G \times G = (N \times N, \Sigma \times \Sigma, P^2, (S, S))$
$P^2 = P \times P = \begin{pmatrix} S \\ S \end{pmatrix}$

2 Einleitung

Struktur: Form → Funktion
Funktion folgt Form, Form folgt Sequenz
Proteine, RNA, DNA: Sequenzen

4 Strukturlevels:

- primäre Struktur (Sequenz): 1 Dimension
- sekundäre Struktur (grobe Annäherung an Struktur): 2 Dimensionen
- tertiäre Struktur (räumliche Struktur): 3 Dimensionen
- quartiäre Struktur (räumliche Anordnung von interagierenden Strukturen): 4 Dimensionen

Behandlung hauptsächlich 2D

2.1 RNA

[4] Funktion:

- Informationsträger
- Regulator/Katalysator
- Theorieder RNA-World

- Nicht-Messenger-RNA: ncRNA (nc - non-coding)

- Aufbau: Zucker-Phosphat-Rückgrat
- Basen:
  - Purine: Adenin, Guanin
  - Pyrimidine: Cytosin, Uracil
- Paarung: A-U, G-C
- RNA einzelsträngige A-Helix (DNA: doppelsträngige B-Helix)

[4] `https://de.wikipedia.org/wiki/Ribonukleins%C3%A4ure`

2.2 R/DNA-Sekundärstruktur

Definition: Liste von Basenpaaren, sodass gilt (theoretische Regeln):

1. erlaubte Basenpaarungen:
   - Watson-Crick: AU, UA, GC, CG
   - Wobble: GU, UG
2. zwischen miteinander paarenden Basen müssen mindestens 3 Basen stehen $if(i,j)\epsilon B \rightarrow i < j-3$
   Beispiel Paarung A und U:

A	U	A	U	A	U	A	U
			1	2	3	4	

3. keine Tripletts (Multipletts): eine Base paart maximal mit einer anderen $if(i,j);(i,k)\epsilon B \rightarrow j = k$
4. keine pseudo-Knoten: Basen kreuzen sich nicht $if(i,j);(k,l)\epsilon B \rightarrow i < j < k < l \; und \; i < k < l < j$

Motivation zu Regeln: jedes Basenpaar teilt das Molekül in 2 Teile (innen und außen), die miteinander nicht interagieren (vor allem Regel 3 + 4)

physikalische Eigenschaften:

1. Großteil des stabilisierenden Energie für RNA-Struktur kommt aus der Sekundärstruktur
2. Sekundärstruktur bildet sich zeitlich vor Tertiärstruktur aus

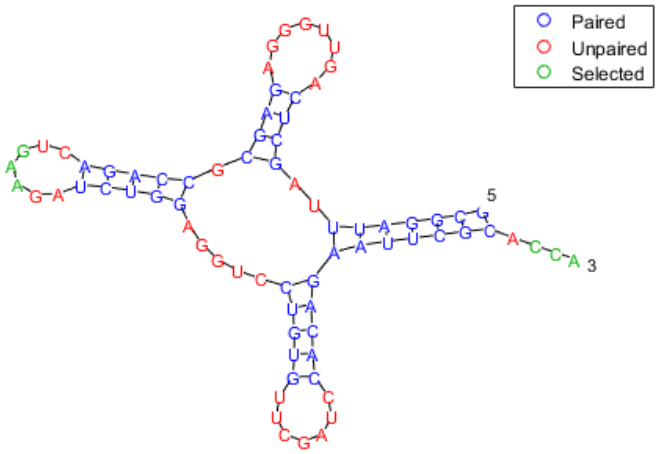
Experimenteller Nachweis 3D, 4D:

- Röntgenkristallographie: Kristall benötigt $\rightarrow$ oft schwierig
- nuclear magnet resonanz (nmr): stark konzentrierte Lösung benötigt, nur Distanzen zwischen Atomen ermittelbar

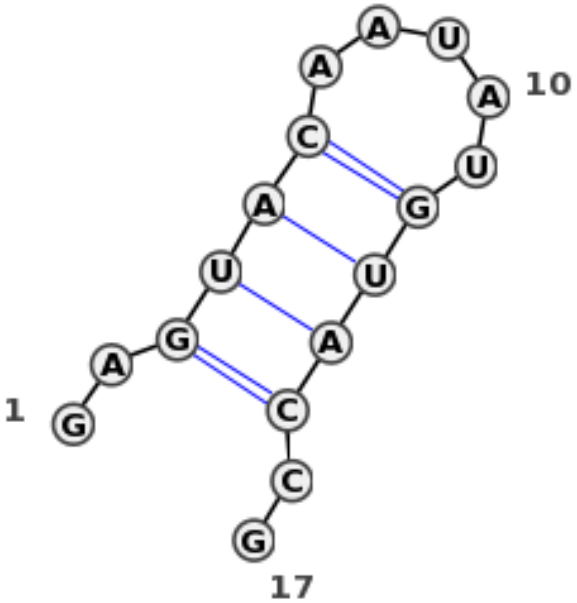
für 2D: Methoden, die bevorzugt einzelsträngige oder doppelsträngige Strukturen schneiden

2.3 Strukturabbildungen

1. Strukturplot:



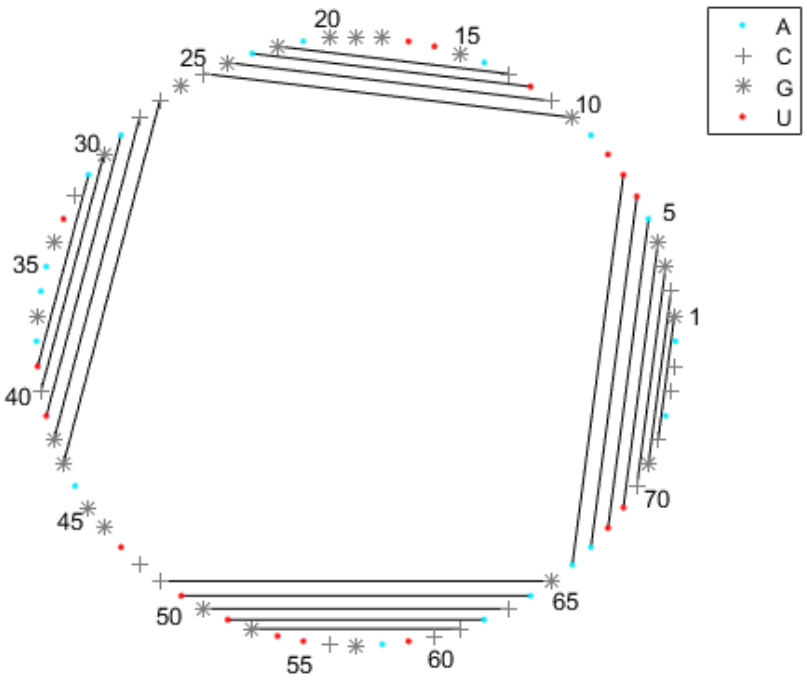
2. Dot-Bracket:



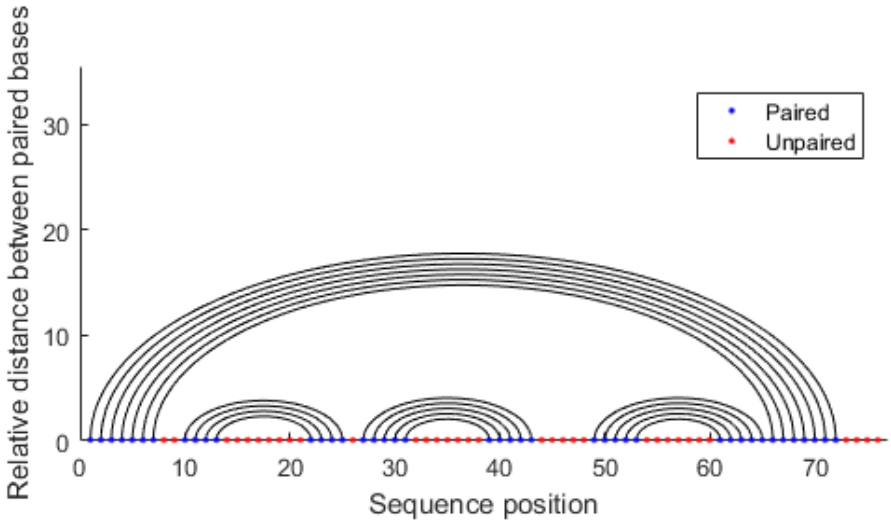
Seq: GAGUACAAUAUGUACCG

Str: ..((((.....))))..

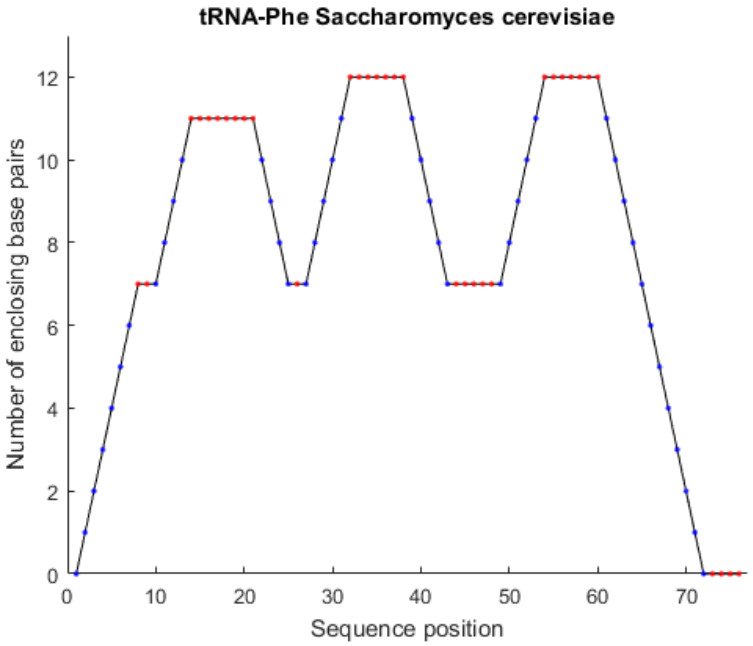
3. Zirkulärplot:



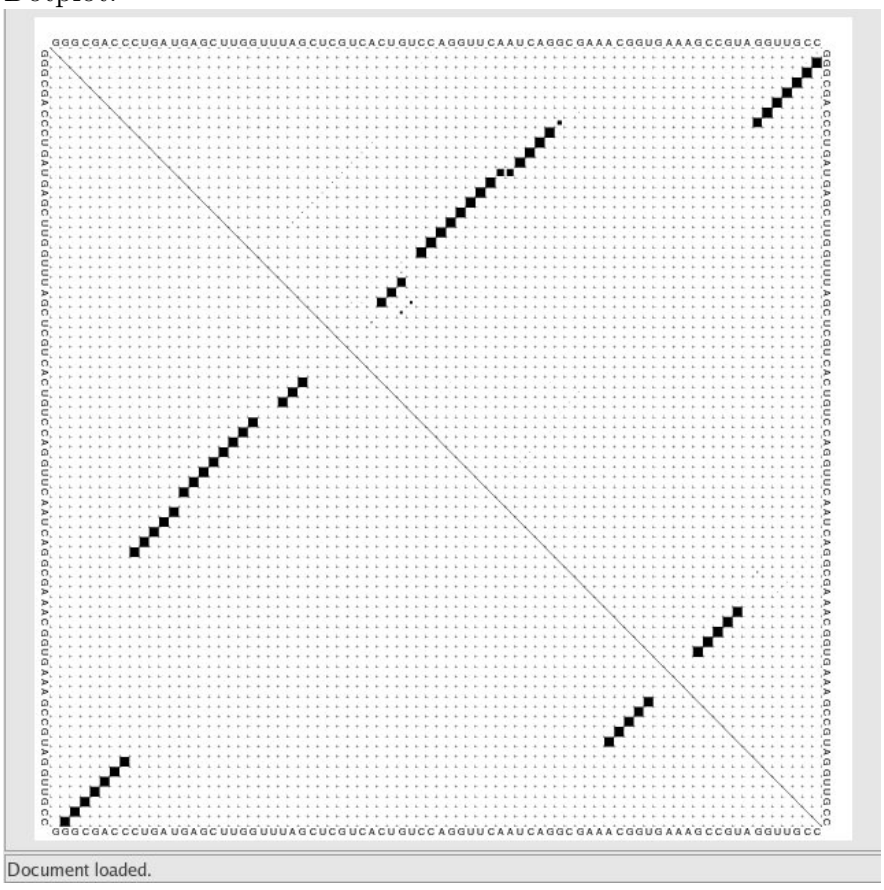
4. Bogenplot:



5. Mountainplot:



6. Dotplot:



3 Strukturvorhersage

- durch Aufteilung kann Dynamics Programming verwendet werden
- Beginn: einzelne Basen → keine Struktur

3.1 Nussinov

- von Ruth Nussinov (1978)
- Versuch Struktur mit der maximalen Anzahl der Basenpaare zu finden (Grundlage ist Sequenz)

Dynamics Programming
- Initialisierung:

- $N(i,i) = 0$
- $N(i,j) = 0 \; if \; i < j \leq i+3$ (siehe Regel 2)

$$N_{ij} = max \begin{cases} N(i+1,j) \; (ungepaart) \\ \max\limits_{i+3<k\leq j} \; N(i+1,k-1) + N(k+1,j) + F(i,k) \end{cases}$$

3.2 Turner-Modell (Nearest-Neighbor-Modell)

3.3 Zuker-Algorithmus

3.3.1 suboptimales Falten

3.4 Wuchty-Algorithmus

3.4.1 Wuchty-Backtracking

3.5 McCaskill

3.6 stochastisches Backtracking

3.7 Strukturvorhersagen verbessern

3.8 Konsensusstrukturvorhersagen

3.9 Wie kann RNA evolvieren?

3.9.1 Neutrale Netzwerke

3.9.2 SHAPE-Abstraktion

3.9.3 Energielandschaften

3.9.4 Faltungskinetik

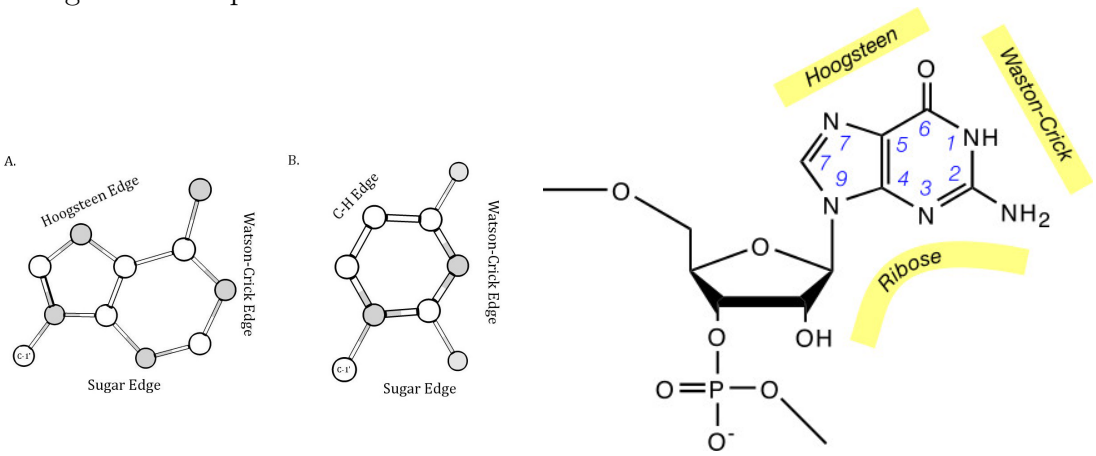
3.9.5 Barriers Trees

3.9.6 Flooding-Algorithmus

3.9.7 Co-transcriptional folding

4 weitere Bindungsarten, erlaubte Basenpaare

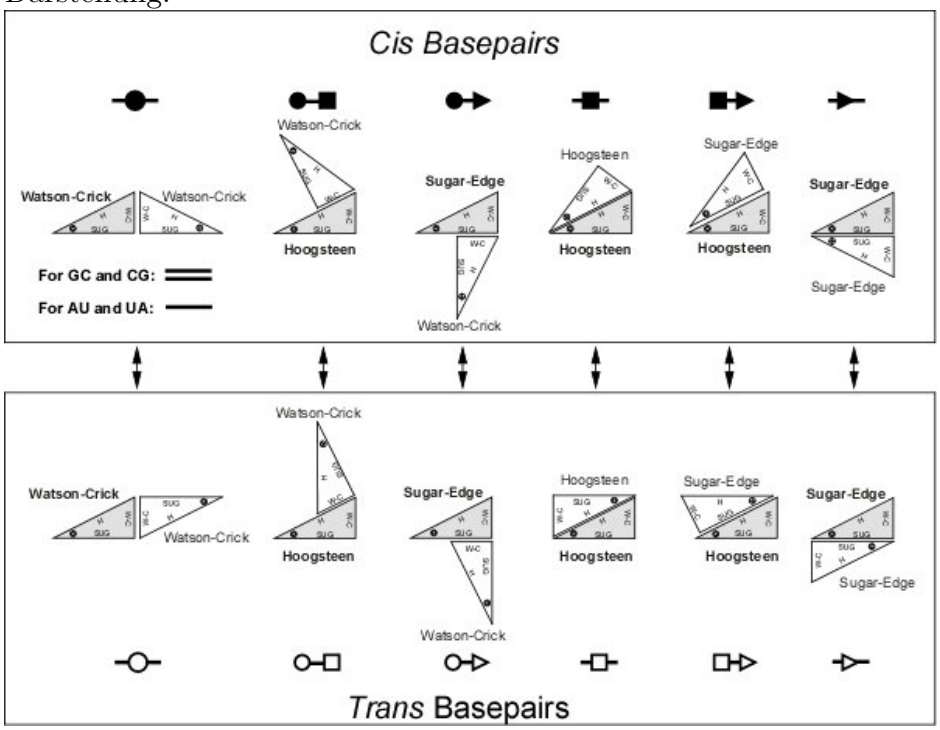
- Hoogsteen base pair[5]



Jede Base kann mit jeder ihrer Kanten zu jeder Kante jeder Base ein Basenpaar bilden.

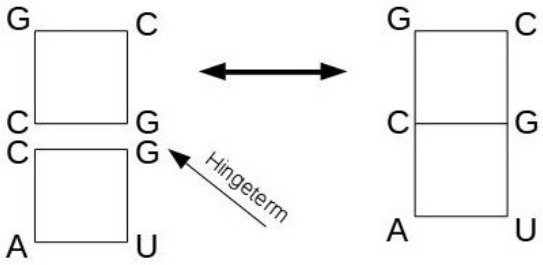
Non-Standard Basepairs:
Strukturmotive: Pattern von Standard basepairs führt zu speziellen 3D-Struktur (Kink-Turn)
Bifurcations (tripletts meistens) 12 * 12 * 2 mögliche Basenpaare: Warum 288?

Darstellung:



[5] https://en.wikipedia.org/wiki/Hoogsteen_base_pair

Isoelektrische Basenpaare
Änderung eines isoelektrischen Basenpaars gegen ein anderes ändert nichts an der Struktur
Listen von isoelektrischen Basenpaaren erstellt von Leontis und Westhof



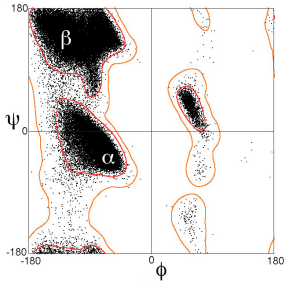
Programme: MC-Fold, RNAWolf

5 Proteine

- 20 Aminosäuren
- drei positiv geladene Aminosäuren (basisch): Arg (R), His (H), Lys (K)
- zwei negativ geladene Aminosäuren (sauer): Asp (D), Glutaminsäure (E)
- sehr unterschiedlich in den Seitenketten
- Verbindung durch Peptidbindung

Frage: Wie rotieren Aminosäuren, die durch eine Peptidbindung verbunden sind, im Räum?
Stichworte: Cis, Torsionswinkel

Ramachandran Plot:[6]
allgemeines Beispiel:



[6] https://en.wikipedia.org/wiki/Ramachandran_plot

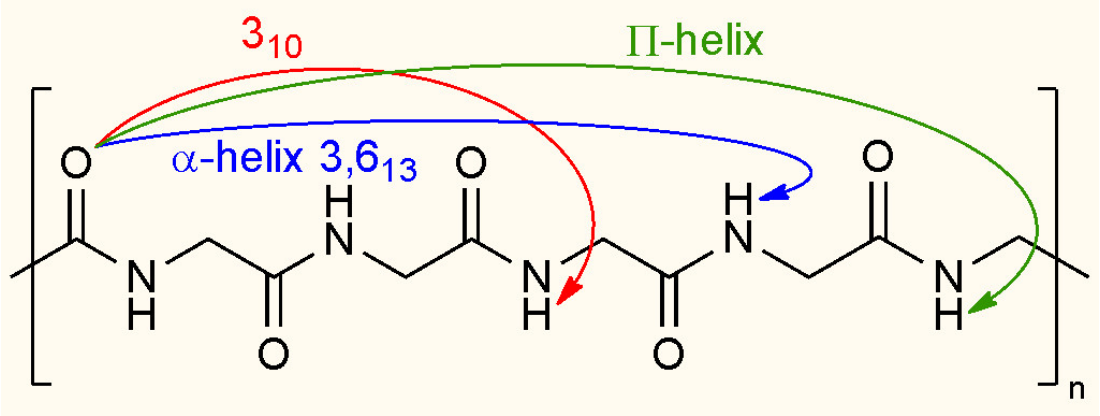
6 Sekundärstrukturelemente

- Unterscheidung in drei Haupttypen[7]
Proteine:

- Helix α-Helix (häufigstes)
  - coiled-coli-Struktur: Helix umgeben mit einer Helix
  - Transmembranhelices: 20 - 30 Aminosäuren, hydrophob, gehen durch die Zellmembran durch
- Extended-Faltblatt: mindestens zwei Faltblätter immer zusammen, da dieses sich gegenseitig stabilisieren
  - parallel, antiparallel
- Turn (drehen der Backbonerichtung)
- Coil (Rest)

drei Helixe: Unterscheidung, was und wie viel zwischen den Wasserstoffbrückenbindungen steht[8]

- $\alpha - Helix$: 3,6,13-Helix (Helix zwischen 3. und 6. Atom, dazwischen liegen 13 Atome)
- $\pi - Helix$: 4,1,16



6.1 Chou-Fasman (Sekundärstrukturvorhersage von Proteinen)

- ca. 50% Genauigkeit

[7] https://de.wikipedia.org/wiki/Sekund%C3%A4rstruktur
[8] https://en.wikipedia.org/wiki/Protein_secondary_structure

- 3 Tabellen mit Scores für α (Helix), β (Faltblatt) und t (Turn) für alle Aminosäuren
  - z.B. gut für Helix: Glu (1,51), Met Ala, Leu
  - schlecht für Helix: Pro, Gly (0,57)
  - gut für Faltblatt: Val (1,7), Ile (1,6)
  - schlecht für Faltblatt: Asp, Glu (0,37), Pro (0,55)
- Unabhängig voneinander α, β, t bewerten:
  - nucleation: 4 von 6 Aminosäuren haben $S_{(\alpha)}$ ¿ 1,03
    Erweitern nach links und rechts, bis Durchschnitt der letzten 4 AS $S_{(\alpha)}$ ¡ 1 haben
  - β: 3 von 5 Aminosäuren sollen $S_{(\beta)}$ ¿ 1 haben, letzten 4AS $S_{(\beta)}$ ¡ 1
- Turn: $score(t) = S_{(t)}(x1) \cdot S_{(t)}(x2) \cdot S_{(t)}(x3) \cdot S_{(t)}(x4)$

Weiterentwicklung:
- nicht nur eine Aminosäure sondern gesamte Umgebung anschauen

GOR-Algorithmus:[9]

- bis zu 70% genau - es gibt GOR1 bis GOR5, unterschiedliche Berechnungen

- drei Matritzen mit Scores
  20 x 17 Matritze $(\alpha, \beta, turn)$
  Beispiel für α: waagerecht: -8 bis +8, senkrecht alle Aminosäuren
- Score aus Summierung über Matrixeinträge, dann ähnliche wie Chou-Fasman

Beispiel: ACCTYRARRGHSTFYSW
für R $S_\alpha = S^\alpha(-8, A) + S^\alpha(-7, C) + ... + S^\alpha(8, W)$
- das für alle Sekundärstrukturelemente

weiterer Algorithmus: SPIDER2

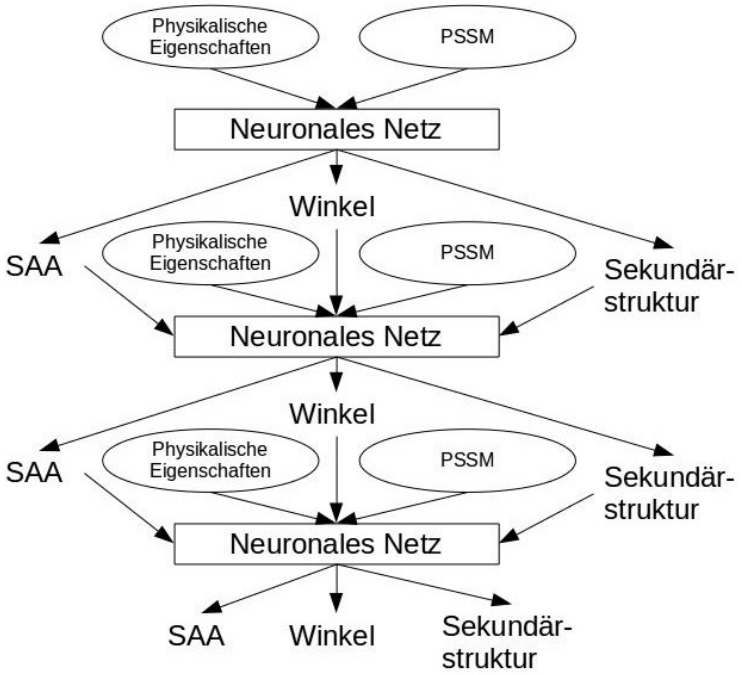
- ca. 80% genau
- Winkel zwischen Aminosäuren berechnen
- Surface Accesible Area
- Sekundärstrukturen

Physikalische Eigenschaften von Aminosäuren:

[9] `https://en.wikipedia.org/wiki/GOR_method`

- sterischer Parameter (graph shape index: dünnes oder dickes Molekül)
- Hydrophobizität
- Polarisierbarkeit
- Isoelektrischen Punkt
- Helix Wahrscheinlichkeit
- Volumen
- Falblattwahrscheinlichkeit
- zusätzlich mit psi-Blast: PSSM ermitteln (kein Ergebnis für Struktur sondern nur für Sequenz!)

dann alle diese Parameter in neuronales Netz stecken:



weitere Möglichkeit: Meta Server

- ruft mehrere Algorithmen auf
- höhere Wahrscheinlickeit durch vergleichen der Ergebnisse (z.B. majority vote)