

Bioinformatik von RNA- und Proteinstrukturen

Inhaltsverzeichnis

1	Formale Sprachen	1
1.1	formale Grammatik G	1
1.2	Klassifikation von formalen Sprachen	1
1.3	Hidden Markov Model	3
2	Einleitung	5
2.1	RNA	5
2.2	R/DNA-Sekundärstruktur	6
2.3	Strukturabbildungen	7
3	Strukturvorhersage	10
3.1	Nussinov	10
4	McCaskill	11
4.1	Turner-Modell (Nearest-Neighbor-Modell)	13
4.2	Zuker-Algorithmus	13
4.2.1	suboptimales Falten	13
4.3	Wuchty-Algorithmus	13
4.3.1	Wuchty-Backtracking	13
4.4	McCaskill	13
4.5	stochastisches Backtracking	13
4.6	Strukturvorhersagen verbessern	13
4.7	Konsensusstrukturvorhersagen	13
4.8	Wie kann RNA evolvieren?	13
4.8.1	Neutrale Netzwerke	13
4.8.2	SHAPE-Abstraktion	13
4.8.3	Energielandschaften	13
4.8.4	Faltungskinetik	13
4.8.5	Barriers Trees	13
4.8.6	Flooding-Algorithmus	13
4.8.7	Co-transcriptional folding	13
5	weitere Bindungsarten, erlaubte Basenpaare	14
6	Proteine	15
7	Sekundärstrukturelemente	16
7.1	Chou-Fasman (Sekundärstrukturvorhersage von Proteinen)	16

1 Formale Sprachen

Formale Sprache¹ L über Alphabet Σ

$L \subseteq \Sigma^*$

mit Σ^* = Kleensche Hülle² 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: $\Sigma = \{a\}, \Sigma^* = \{\varepsilon, a, aa, aaa, \dots\}, L = \{a, aa, aaaa, \dots\}$

1.1 formale Grammatik G

$G = (N, \Sigma, P, S)$ mit

- N = Nichtterminale
- Σ = Alphabet
- P = Produktionsregeln
- S = Startsymbol ($\in N$)

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

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-Hierarchie³:

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

¹https://de.wikipedia.org/wiki/Formale_Sprache

²https://de.wikipedia.org/wiki/Kleenesche_und_positive_H%C3%BClle

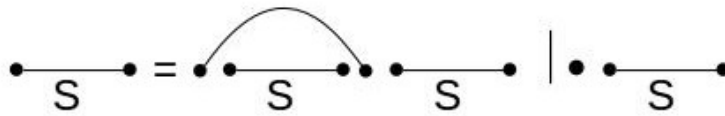
³<https://de.wikipedia.org/wiki/Chomsky-Hierarchie>

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

bei Alignments: $\boxed{S} \longrightarrow \boxed{S} \begin{smallmatrix} \vdots \\ \vdots \end{smallmatrix} \mid \boxed{S} \text{---} \mid \boxed{S} \text{---} \mid \varepsilon$

Erweiterung mit Wahrscheinlichkeit: $G=(N, \Sigma, P, S, \Omega)$
mit Ω = Wahrscheinlichkeit für Produktionsregeln

jetzt auf RNA-Vorhersagen:

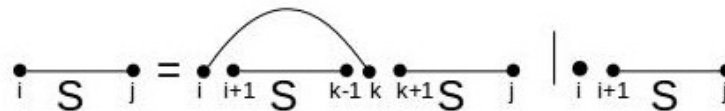


scoring scheme: Bewertung von $\sigma(\curvearrowright) = 1$, $(\sigma(\text{---}))$, $\sigma(\bullet) = 0$
scoring function:

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

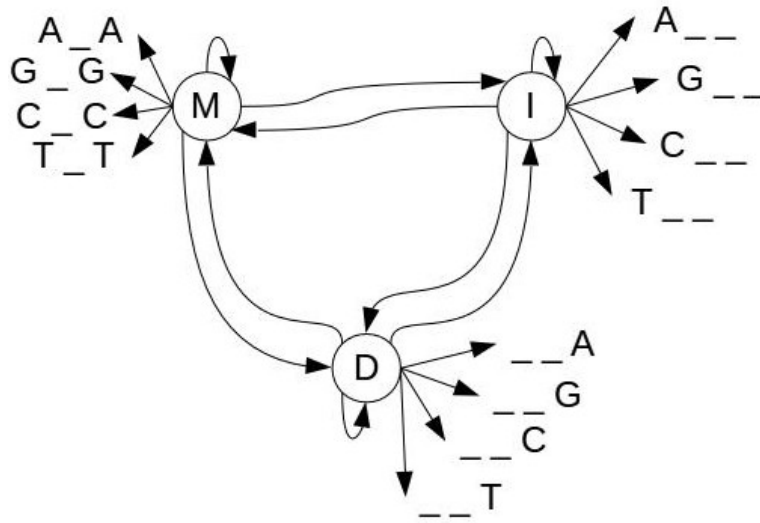
choice function:

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



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

1.3 Hidden Markov Model



M: Match, I: Insertion, D: Deletion

Grammatik:

- $M \rightarrow M_{A_A} | \dots | I | D$
- $I \rightarrow I_{A_{--}} | \dots | D | M$
- $D \rightarrow D_{--A} | \dots | M | I$

Beispiel:

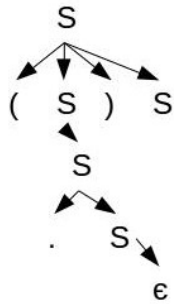


Faltungsgrammatik

$S \rightarrow (S)S | .S | \varepsilon$

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

Beispiel in Baumdarstellung:



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

Alignmentgrammatik

$$S \rightarrow .S|_S|\varepsilon$$

$$G = (N = \{S\}, \Sigma = \{., _\}, P = \{S \rightarrow .S|_S|\varepsilon\}, S)$$

$$\text{Alignment: } G^2 = G \times G = (N \times N, \Sigma \times \Sigma, P^2, (S, S))$$

$$P^2 = P \times P = \begin{pmatrix} S \\ S \end{pmatrix}$$

2 Einleitung

Struktur: Form \rightarrow 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
- quartäre 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):

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

A	U	<u>A</u>	U	A	U	A	<u>U</u>
			1	2	3	4	
- keine Triplets (Multiplets): eine Base paart maximal mit einer anderen
 $if(i, j); (i, k) \in B \rightarrow j = k$
- keine pseudo-Knoten: Basen kreuzen sich nicht $if(i, j); (k, l) \in B \rightarrow i < j < k < l$ 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

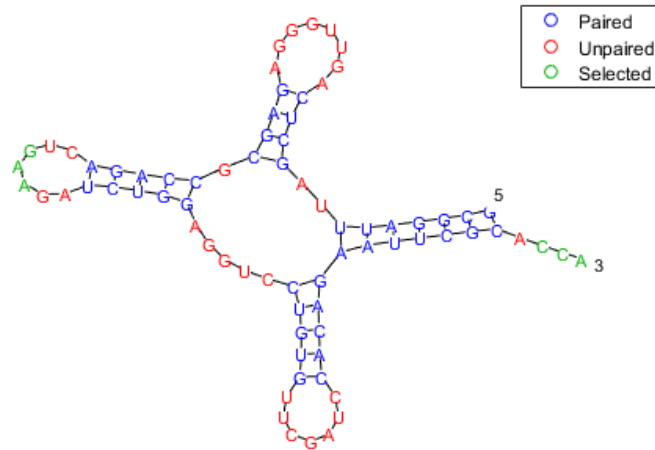
Experimenteller Nachweis 3D, 4D:

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

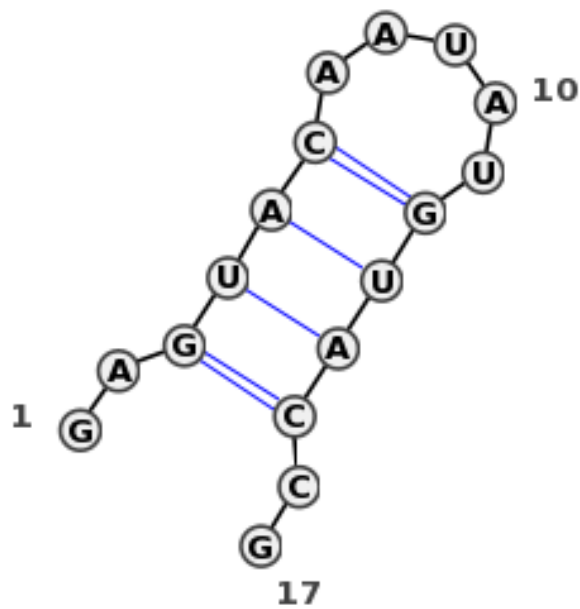
für 2D: Methoden, die bevorzugt einzelsträngige oder doppelsträngige Strukturen schneiden

2.3 Strukturabbildungen

1. Strukturplot:

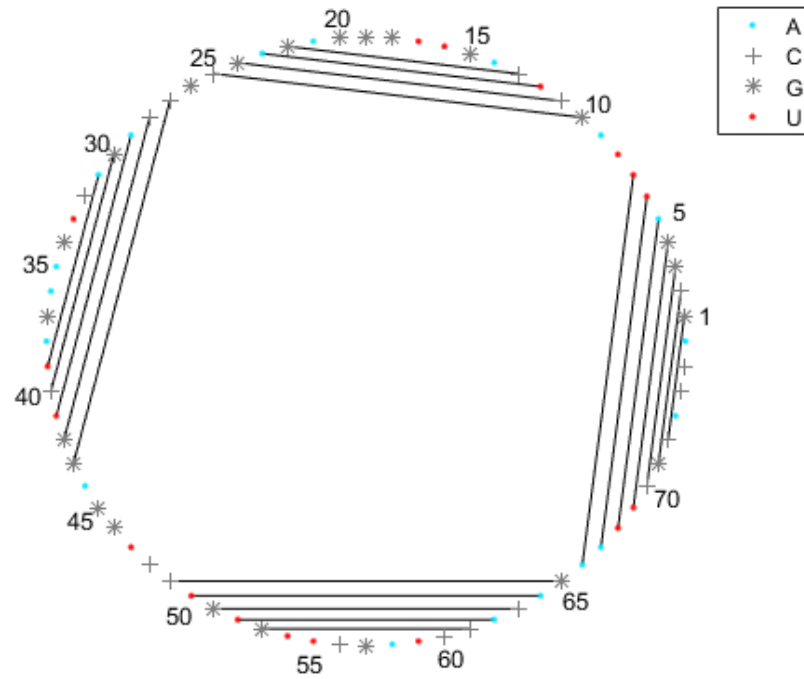


2. Dot-Bracket:

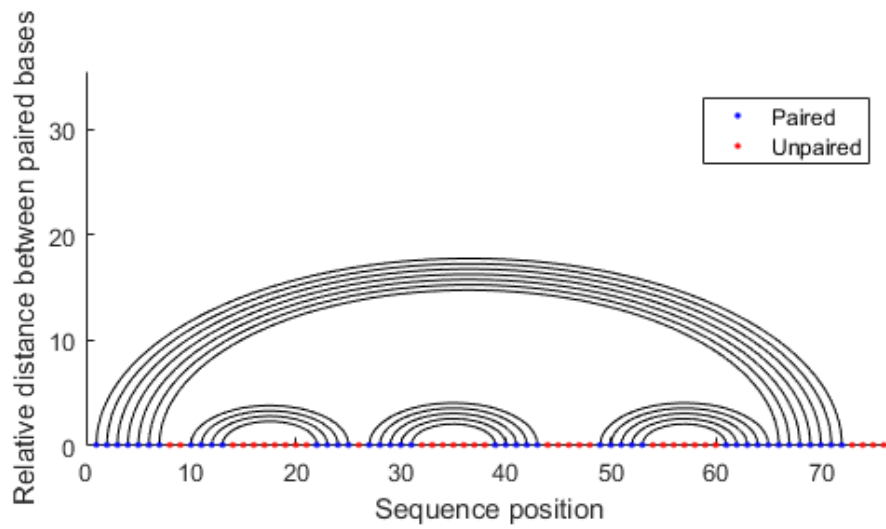


Seq:	GAGUACAAUAUGUACCG
Str:	..(((.....)))...

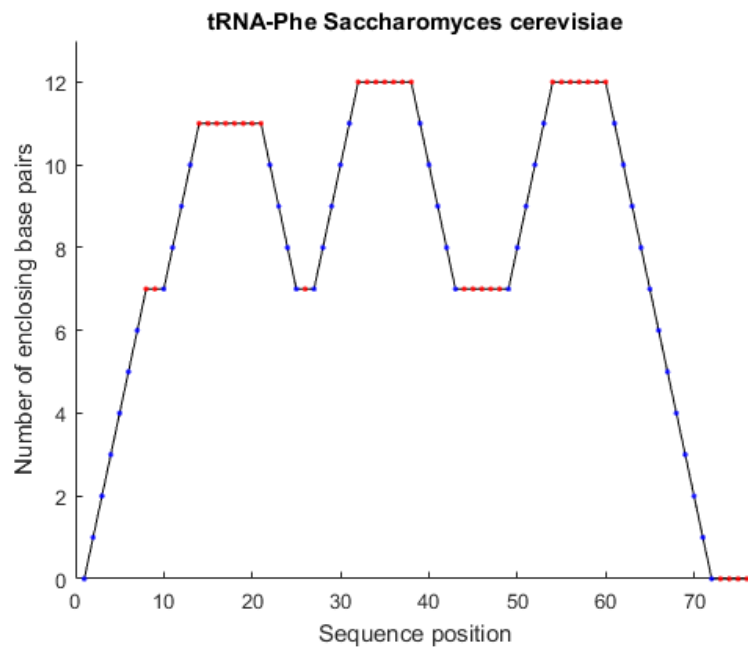
3. Zirkulärplot:



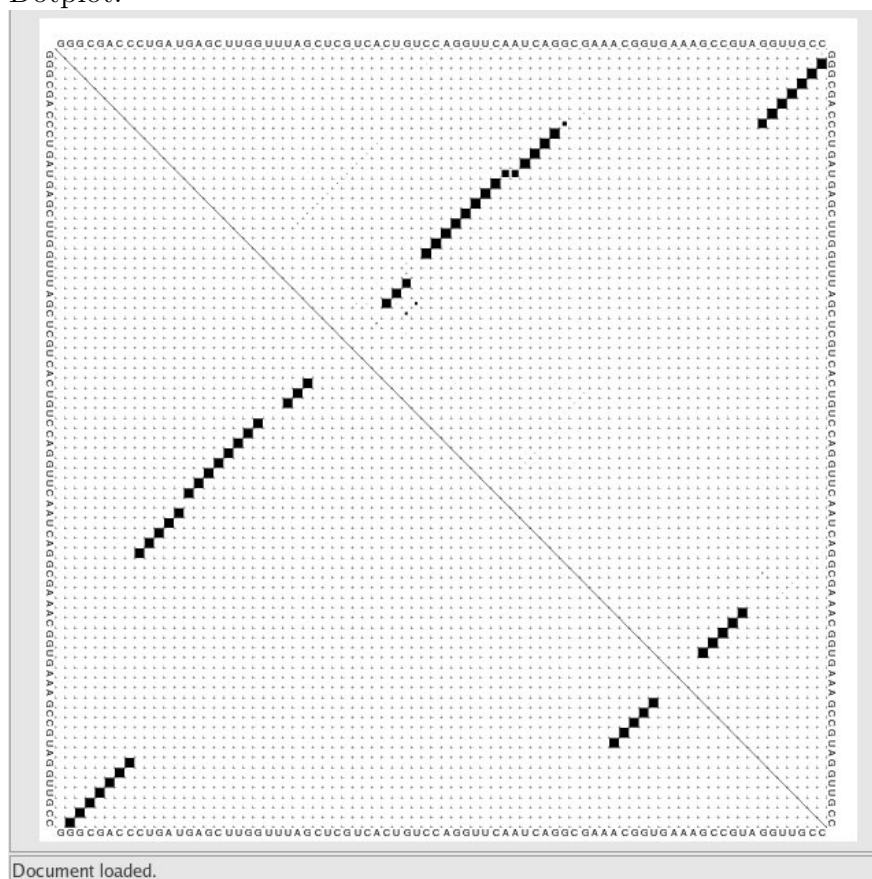
4. Bogenplot:



5. Mountainplot:



6. Dotplot:



3 Strukturvorhersage

- durch Aufteilung kann Dynamics Programming verwendet werden
- Beginn: einzelne Basen \rightarrow 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:**

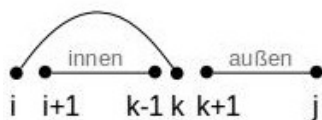
- $N(i, i) = 0$
- $N(i, j) = 0$ if $i < j \leq i + 3$ (siehe Regel 2)
- $N(j + 1, j) = 0$

- **Berechnung:**

$$N(i, j) = \max \begin{cases} N(i + 1, j) & (\text{ungepaart}) \\ \max_{i+3 < k \leq j} N(i + 1, k - 1) + N(k + 1, j) + F(i, k) \end{cases}$$

$$\text{mit } F(i, k) = \begin{cases} 1 & \text{if } i, k \in \{AU, GC, GU\} \\ -\infty & \text{else} \end{cases}$$

Basenpaarung mit i und k teilt Sequenz in inneren und äußeren Teil:



\rightarrow höchste Punktzahl wahrscheinlichste Sekundärstruktur

Ressourcenbedarf:

- Speicher: $\mathcal{O}(n^2)$
- Prozessor: $\mathcal{O}(n^3)$

4 McCaskill

McCaskill stellt eine Methode zur Berechnung der Partitionsfunktion für RNA-Sequenzen basierend auf der Boltzmann Verteilung vor, um die Strukturverteilung einer RNA-Sequenz zu berechnen.

Das Problem bei Zuker ist, dass die sekundäre RNA-Struktur mit minimaler Energie nicht immer die biologisch relevanteste ist. Für sekundäre RNA-Strukturen mit geringer Energie ist es möglich, biologisch relevanter zu sein, als alle anderen möglichen sekundären RNA-Strukturen. Daher wird eine neue Möglichkeit benötigt, um diejenige RNA-Struktur heraus zu finden, die biologisch am relevantesten ist. McCaskill berechnet für eine RNA-Sequenz die Wahrscheinlichkeit der Basenpaare mit Hilfe der Boltzmann-Verteilung. Anhand der Basenpaar-Wahrscheinlichkeiten lässt sich die wahrscheinlichste sekundäre RNA-Struktur ermitteln.

Definition (Boltzmann-Verteilung)

Sei $P(x)$ die Wahrscheinlichkeit für eine RNA-Struktur x . Dann ist eine *Boltzmann – Verteilung* gegeben durch:

$$\beta = \frac{1}{k_B * T} \quad (1)$$

$$Z = \sum_{x'} e^{-\beta E(x')} \quad (2)$$

$$P(x) = \frac{e^{-\beta E(x)}}{Z} \quad (3)$$

wobei $E(x)$ die Energie von x , T die Temperatur, k_B die Boltzmann-Konstante und x' eine mögliche sekundäre RNA-Struktur ist.

Definition (Partitionsfunktion für eine RNA-Sequenz)

Die Partitionsfunktion für eine RNA-Sequenz s ist:

$$Z = \sum_R e^{-\beta E(R)} \quad (4)$$

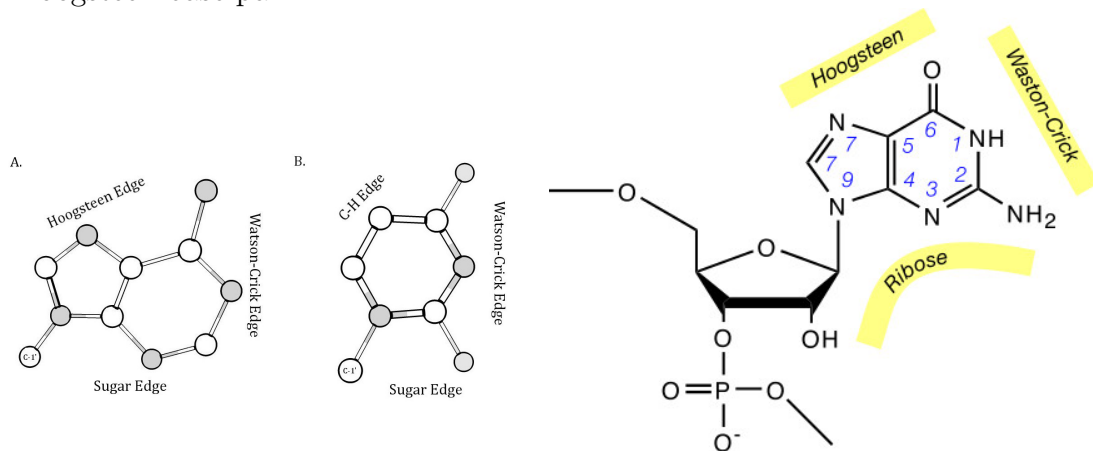
wobei R eine sekundäre RNA-Struktur von s , T die Temperatur und k_B die Boltzmann-Konstante ist.

Der Algorithmus von McCaskill berechnet die Partitionsfunktion Z für eine RNA-Sequenz s effizient. Als Energiemodell für sekundäre RNA-Strukturen wird das Modell von Zuker benutzt. Das bedeutet, dass die Energie einer sekundären RNA-Struktur der Summe aller möglichen sekundären RNA-Strukturelemente entspricht.

- 4.1 Turner-Modell (Nearest-Neighbor-Modell)
- 4.2 Zuker-Algorithmus
 - 4.2.1 suboptimales Falten
- 4.3 Wuchty-Algorithmus
 - 4.3.1 Wuchty-Backtracking
- 4.4 McCaskill
- 4.5 stochastisches Backtracking
- 4.6 Strukturvorhersagen verbessern
- 4.7 Konsensusstrukturvorhersagen
- 4.8 Wie kann RNA evolvieren?
 - 4.8.1 Neutrale Netzwerke
 - 4.8.2 SHAPE-Abstraktion
 - 4.8.3 Energielandschaften
 - 4.8.4 Faltungskinetik
 - 4.8.5 Barriers Trees
 - 4.8.6 Flooding-Algorithmus
 - 4.8.7 Co-transcriptional folding

5 weitere Bindungsarten, erlaubte Basenpaare

- Hoogsteen base pair⁵



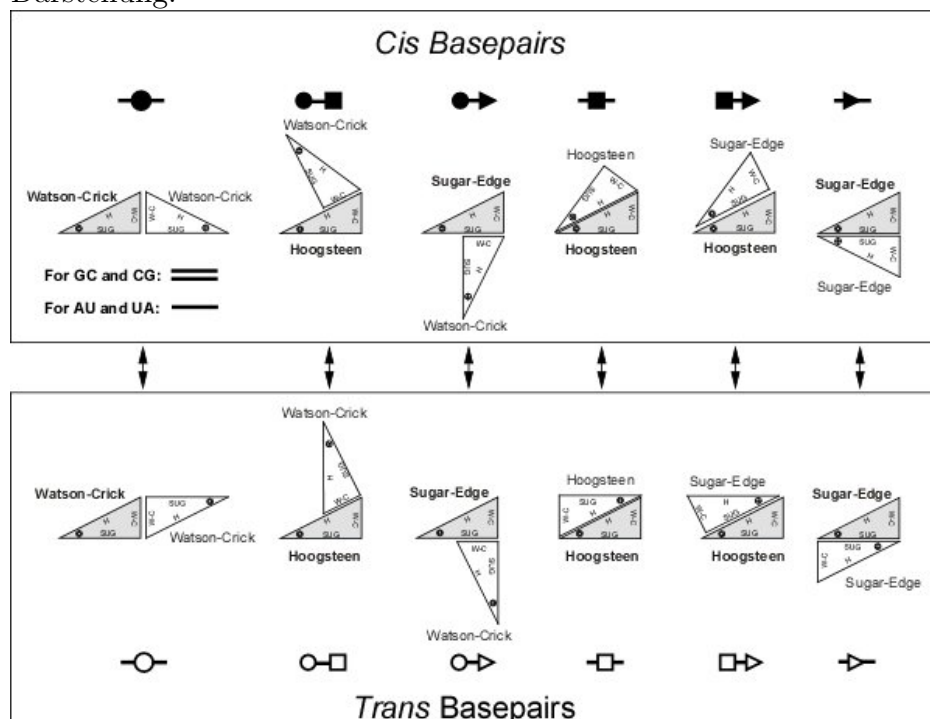
Jede Base kann mit jeder ihrer Kanten zu jeder Kante jeder Base ein Basenpaar bilden.

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:

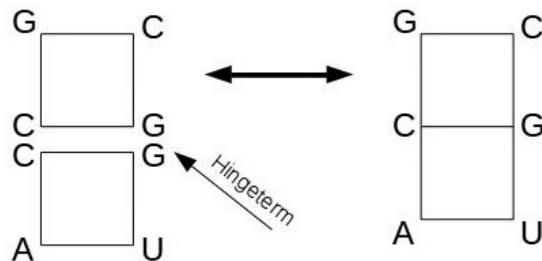


⁵https://en.wikipedia.org/wiki/Hoogsteen_base_pair

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

6 Proteine

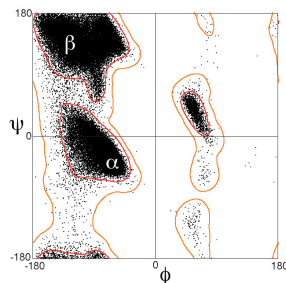
- 20 Aminosäuren
- drei positiv geladene Aminosäuren (basisch): Arg (R), His (H), Lys (K)
- zwei negativ geladene Aminosäuren (sauer): Asp (D), Glutaminsäure (E)
- sehr unterschiedlich in den Seitenketten
- Verbindung durch Peptidbindung

Frage: Wie rotieren Aminosäuren, die durch eine Peptidbindung verbunden sind, im Raum?

Stichworte: Cis, Torsionswinkel

Ramachandran Plot:⁶

allgemeines Beispiel:



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

7 Sekundärstrukturelemente

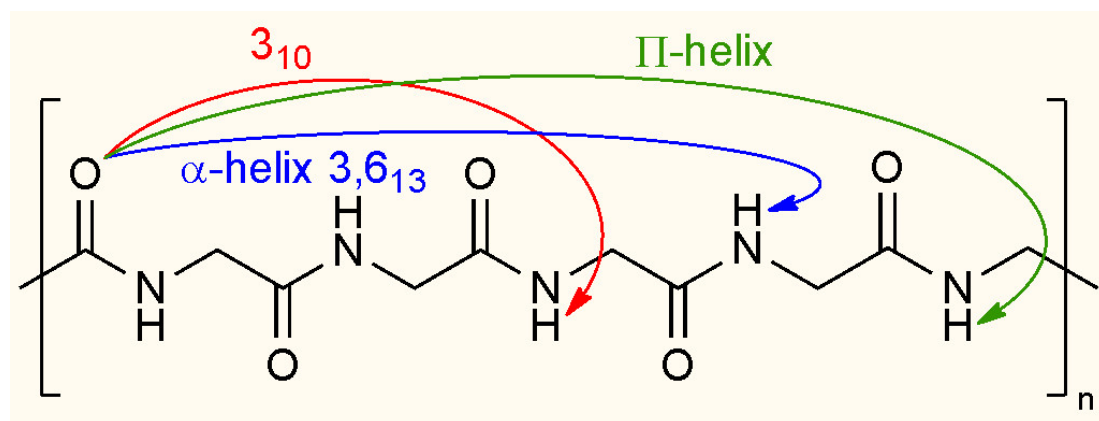
- Unterscheidung in drei Haupttypen⁷

Proteine:

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

drei Helixe: Unterscheidung, was und wie viel zwischen den Wasserstoffbrückenbindungen steht⁸

- α – Helix: 3,6,13-Helix (Helix zwischen 3. und 6. Atom, dazwischen liegen 13 Atome)
- π – Helix: 4,1,16



7.1 Chou-Fasman (Sekundärstrukturvorhersage von Proteinen)

- ca. 50% Genauigkeit

⁷<https://de.wikipedia.org/wiki/Sekund%C3%A4rstruktur>

⁸https://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 $S_{(\alpha)} \geq 1,03$
Erweitern nach links und rechts, bis Durchschnitt der letzten 4 AS $S_{(\alpha)} \geq 1$ haben
 - β : 3 von 5 Aminosäuren sollen $S_{(\beta)} \geq 1$ haben, letzten 4AS $S_{(\beta)} \geq 1$
- 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 : waagrecht: -8 bis +8, senkrecht alle Aminosäuren
- Score aus Summierung über Matrixeinträge, dann ähnliche wie Chou-Fasman

Beispiel: ACCTYRARRRGHSTFYSW

für R $S_{\alpha} = S^{\alpha}(-8, A) + S^{\alpha}(-7, C) + \dots + S^{\alpha}(8, W)$

- das für alle Sekundärstrukturelemente

weiterer Algorithmus: SPIDER2

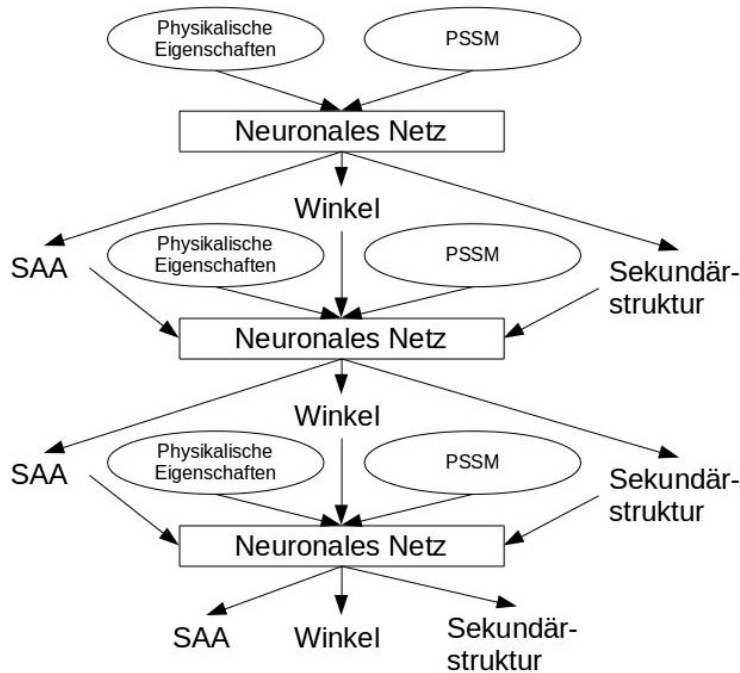
- ca. 80% genau
- Winkel zwischen Aminosäuren berechnen
- Surface Accesible Area
- Sekundärstrukturen

Physikalische Eigenschaften von Aminosäuren:

⁹https://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 Wahrscheinlichkeit durch vergleichen der Ergebnisse (z.B. majority vote)