Bioinformatik von RNA- und Proteinstrukturen

Inhaltsverzeichnis

1 Formale Sprachen

Formale Sprache[1] L über Alphabet Σ
L ⊆ Σ*
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 → aaS → 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 → $(N \cup \Sigma)^*$ → stochstisch kontextfreie Grammatik (SCFG) → Dynamics Programming

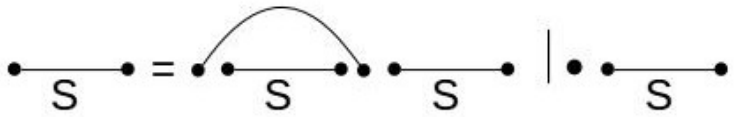
[1] https://de.wikipedia.org/wiki/Formale_Sprache
[2] https://de.wikipedia.org/wiki/Kleenesche_und_positive_H%C3%BClle
[3] https://de.wikipedia.org/wiki/Chomsky-Hierarchie

- 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:

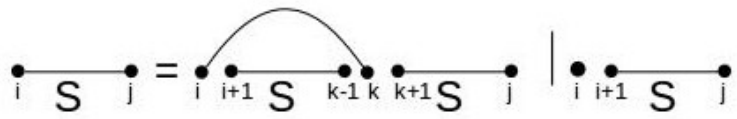


scoring scheme: Bewertung von σ (◠) = 1, (σ (⟷)), σ (•) = 0
scoring function:

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

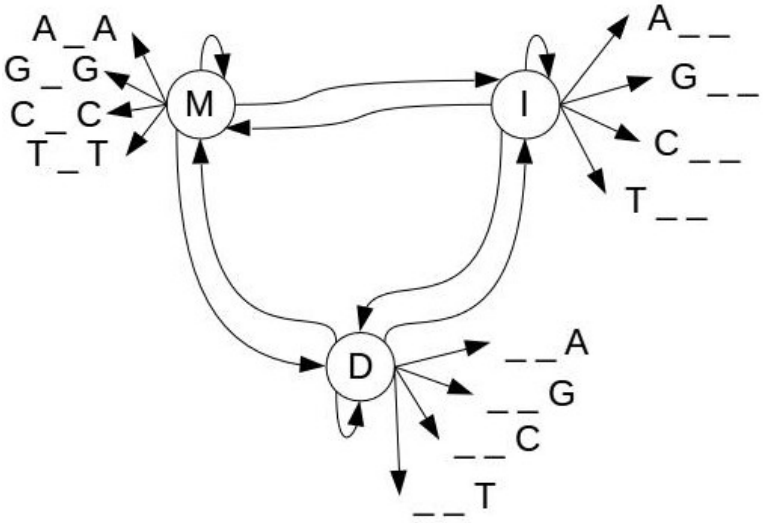
choice function:

- 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:

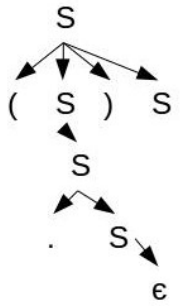
- M $\rightarrow M_{A\_A}|...|I|D$
- I $\rightarrow I_{A\_}|...|D|M$
- D $\rightarrow D_{\_A}|...|M|I$

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

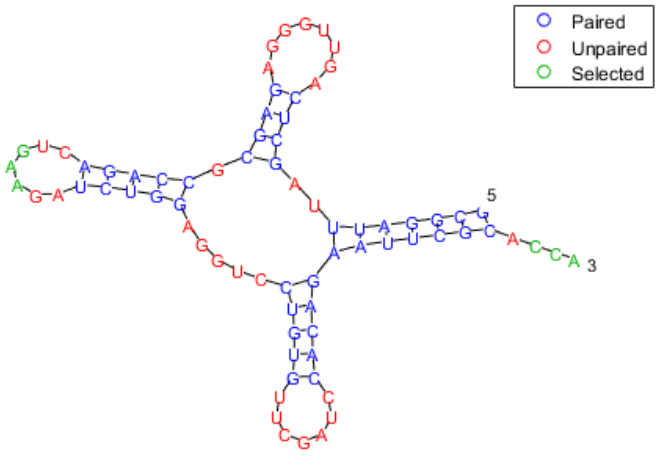
Experimenteller Nachweis 3D, 4D:

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

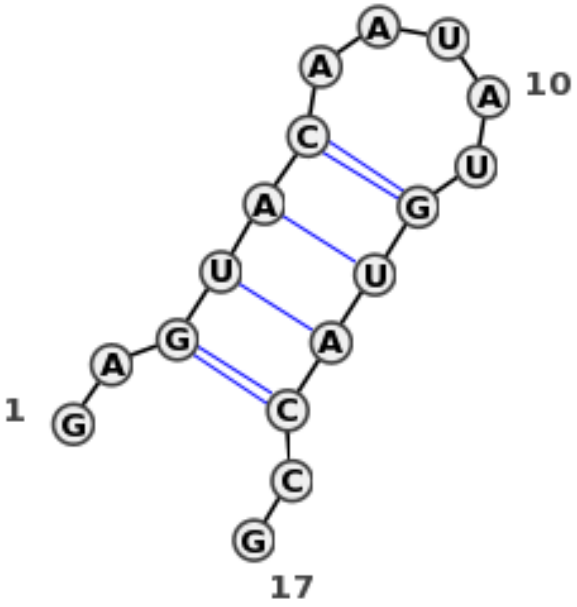
für 2D: Methoden, die bevorzugt einzelsträngige oder doppelsträngige Strukturen schneiden

2.3 Strukturabbildungen

1. Strukturplot:



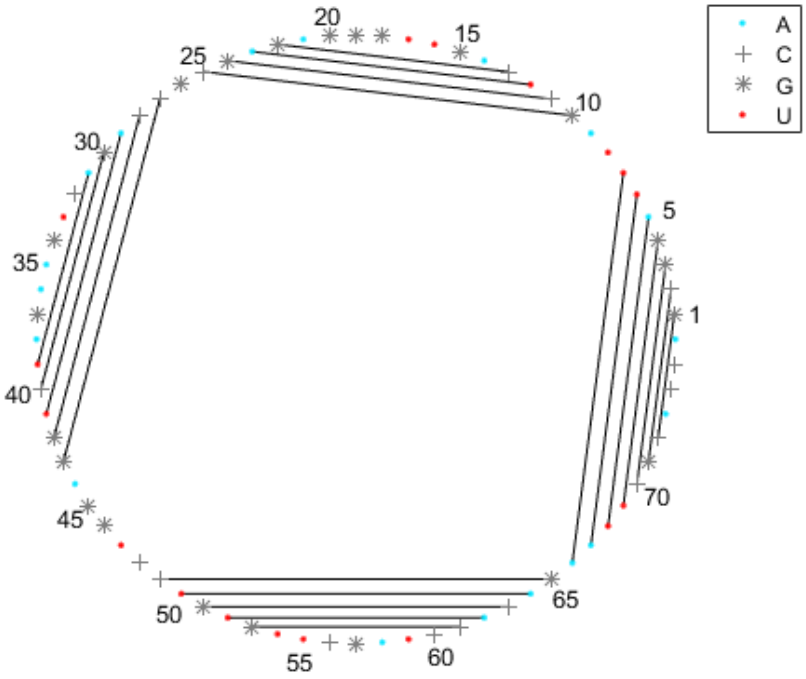
2. Dot-Bracket:



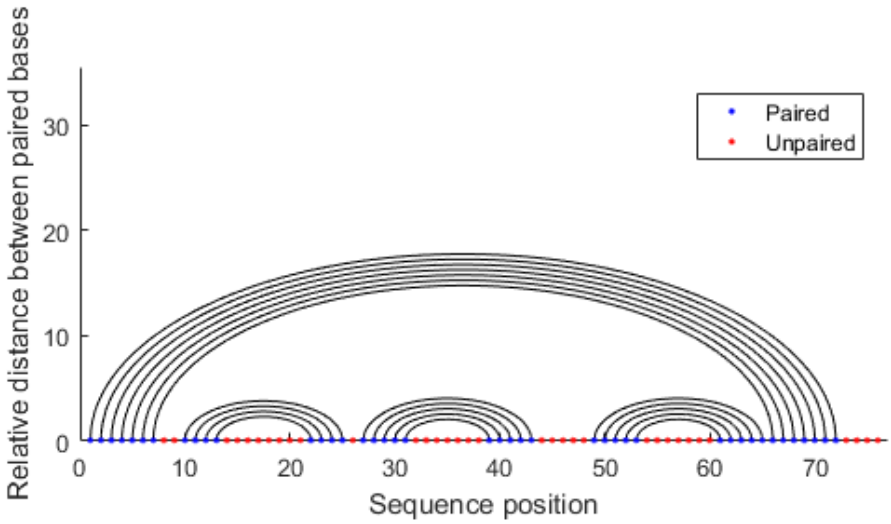
Seq: GAGUACAAUAUGUACCG

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

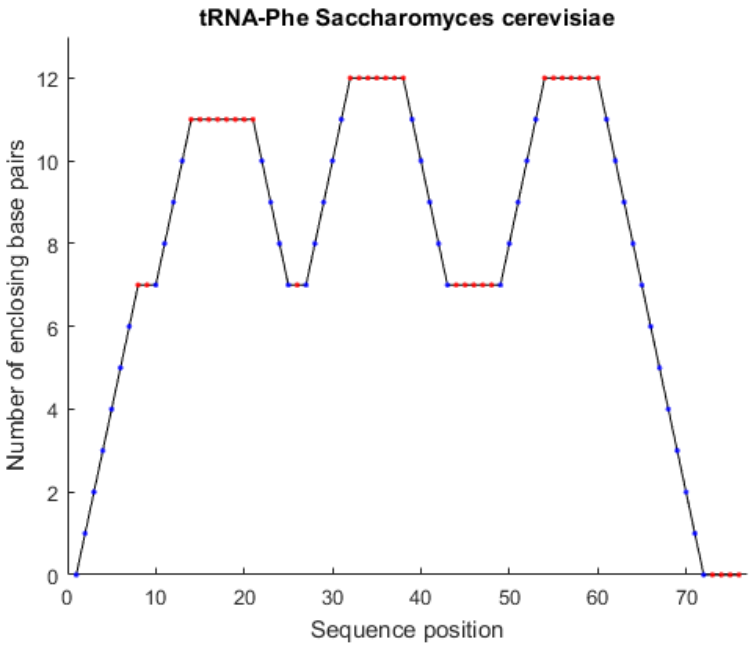
3. Zirkulärplot:



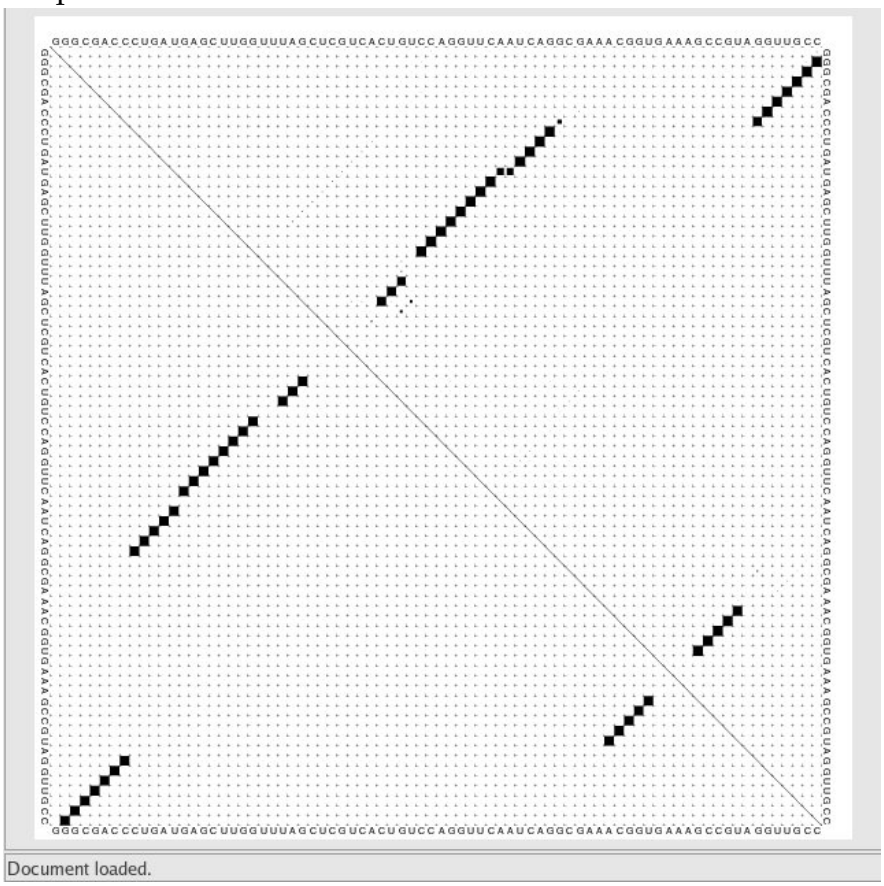
4. Bogenplot:



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:

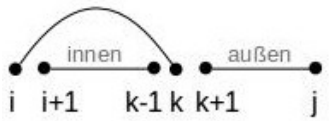
- $N(i,i) = 0$
- $N(i,j) = 0 \; if \; i < j \leq i + 3$ (siehe Regel 2)
- $N(j+1, j) = 0$

- Brechnung:

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

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

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



→ höchste Punktzahl wahrscheinlichste Sekundärstruktur

Resourcenbedarf:

- Speicher: $O(n^2)$
- Prozessor: $O(n^3)$

4 Konsensusstrukturvorhersage - Komparative Analyse

Untersuchung mehrere RNA-Moleküle mit ähnlicher bis gleicher Struktur, da diese möglicherweise miteinander verwandt sind.
→ Funktionsvorhersage
→ Zuordnung von RNA-Klassen
→ Motivsuche

4.1 RNAaliFold- zuerst alignen, dann falten

→ multiples Sequenzalignment (z.B. Needleman-Wunsch)
→ generalisierte Bestimmung einer RNA-Struktur

- minimiere die mittlere Energie
Für RNAalifold ist ein multiples Alignment mit K Sequenzen der Länge m gegeben. Ziel ist es eine RNA-Struktur der Konsensussequenz zu finden, deren minimierte Energie sich aus der Summe aller freien Energien der K Sequenzen und dem Konservierungsgrad gleichbleibender Sequenzen zusammensetzt.

- Unterscheidung von Mutationen
konsistente Mutationen (GC → GU)
und kompensatorische Mutationen (GC → UA)

Berechnung:

$$C(i,j) = max \begin{cases} H(i,j) \\ I \\ M \end{cases} \tag{1}$$

→ Rechne den Term, der die Konservierung beschreibt zu C(i,j)
→ Bestrafung von Verletzung der Komplementarität

Fall1:

$$\gamma(i,j) = \sum \begin{cases} h(s_{1i}, s_{2i}) + h(s_{1j}, s_{2j}) \; if \; s_{1i}, s_{1j}) \epsilon \; Bp \; and \; s_{2i}, s_{2j} \epsilon \; Bp \\ 0 \; else \end{cases} \tag{2}$$

Fall2:

$$C(i,j) = y * \gamma(i,j) + x * \delta(i,j) + mean[C(i,j)] \tag{3}$$

;

$$\delta(i,j) = \sum_{S} \begin{cases} 0 \; if \; (s_{1i}, s_{2i}) \epsilon \; Bp \\ 0,5 \; if \; s_{1i}, s_{2i} \epsilon \; \{-\} \\ 0 \; else \end{cases} \tag{4}$$

Es können so auf Wissen basierte Matrizen mit bekannten konservierten RNA-Strukturen erzeugt werden → Ribosum: knowledge-based Score, für die Wahrscheinlichkeit in einem Basenpaar (i,j) für die Basen s_{1i}, s_{1j} und s_{2i}, s_{2j}

→ Komplexität: $\Omega(n^3 m)$

Problem: Es werden nur Konsensussequenzen gefalten. Somit ist nur Allgemein eine Faltung für alle Sequenzen vorliegend.
Somit nur sinnvoll, wenn zum einen ein Alignment möglich ist und wenn die Sequenzen sehr ähnlich zueinander sind.

4.2 Sankoff-Algorithmus - gleichzeitiges Alignen und Falten

Programm: locarna

Annahme: Zwei Sequenzen sind sich ähnlich, wenn ihre RNA-Struktur einen stark äquivalenten Shape besitzen (len(A) = len(B) und pair(A) = pair(B)).

Idee: Finde für äquivalente Strukturen die minimale Editierdistanz bzw. mfe

Vorgehen: Sankoff = Zuker + Needleman-Wunsch

1 Gegeben sind zwei Sequenzen A und B

2 Finde äquivalente Strukturen in Sequenzen

3 Erzeuge ein zu Strukturen kompatibles Alignment beider Sequenzen

4 Editiere Alignment und Faltungen um minimalen Score $min(E_A + E_B + Distanz_A B)$ zu finden

5 Consensussequenz und Sequenz C gegeben → gehe zu 2

Distanzbestimmung:

$$A(i,j;k,l) = min \begin{cases} A(i+1,j;k+1,l) + \sigma \\ A(i+1,j;k,l) + \sigma_{gap} \\ A(i,j;k+1,l) + \sigma_{gap} \end{cases} \tag{5}$$

Initialisierung:

$$A(i,i;k,k) = \begin{cases} \sigma \; if a_i != b_k \\ 0 \; else \end{cases} \tag{6}$$

$$C(i,i;k,k) = \infty \tag{7}$$

$$M(i,i;k,l) = M(i,j;k,k) = \infty \tag{8}$$

Freie Energie der Sequenzen von i bis j bzw. von k bis l:

$$F(i,j;k,l) = min \begin{cases} F(i+1,j,k+1,l) + A(i,j+1;k,l+1) \ (i,k \ ungepaart) \\ min_{u,v} \begin{matrix} C(i,u;k,v) + F(u+1,j,v+1,l) \\ +A(u,u+1;v,v+1) + A(i,i;j,j)) \end{matrix} (i,k \ gepaart \ mit \ k,v) \end{cases} \tag{9}$$

eingeschlossene Energie zwischen dem Basenpaar (i,j) bzw. (k,l)

$$C(i,j;k,l) = min \begin{cases} H(i,j) + H(k,l) + A(i,j;k,l) \\ \min\limits_{i<u<v<j \ k<x<y<l} \begin{matrix} I(i,j;k,l;u,v;x,y) + C(u,v;x,y) \\ +A(i,u;k,x) + A(v,j;y,l) \end{matrix} \\ \min\limits_{i<u<j \ k<x<l} \begin{matrix} M(i,u;k,x) + M^1(u+1,j-1;x+1,l-1) \\ +A(u,u+1;x,x+1) + A(i,i+1;x,x+1) \\ +A(j-1,j;l-1,l) \end{matrix} \end{cases} \tag{10}$$

Energie von Mulitloops:

$$M(i,j;k,l) = min \begin{cases} M(i+1,j;k+1,l) + A(i,i+1;j,j+1) \\ \min\limits_{i<u<j \ k<x<l} C(i,u;k,x) + A(u,j;x,l) \\ \min\limits_{i<u<j \ k<x<l} C(i,u;k,x) + M(u+1,j;k+1,x) + A(u,u+1;x,x+1) \end{cases} \tag{11}$$

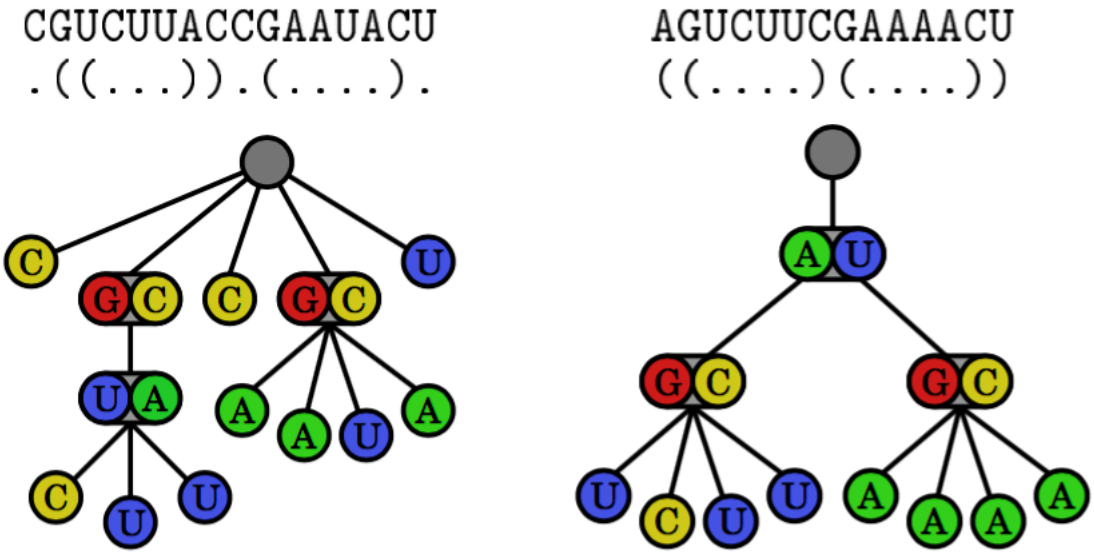
$$M^1(i,j;k,l) = min \begin{cases} C(i,j;k,l) \\ M^1(i,j-1;k,l-1) + A(j-1,j;l-1,l) \end{cases} \tag{12}$$

→ **Komplexität:** $\Omega(Zeit) = \Omega(n^6)$; $\Omega(Memory) = \Omega(n^4)$ →

4.3 TREEforester - zuerst falten, dann alignen

- Wie alignt man gefaltete RNA-Strukturen?
  Nach welchen Kriterien wurde die scheinbar beste Faltung gewählt? →
  Tree-Editing (Bestrafung von strukturellen Mismatches)
- Sind sie überhaupt kompatibel nach ihrer Einschränkung? (Liegen Überkreuzungen vor und sind die Grundvoraussetzungen der Struktur gleich, z.B. keine Pseudoknots) → Tree-Alignment
- Im Falle von sich nicht kreuzenden RNA-Strukturen werden Bäume als vergleichende Datenstruktur erzeugt. Ein RNA-Baum ist ein geordneter Baum dessen Knoten einzelne Basen oder Basenpaare darstellen

- Insgesamt können somit alle Bäume zu einem RNA-Strukturen-Wald zusammengefasst werden (RNAforester).



Tree-Editing: wandle Baum A in Baum B um

- Basen umbenennen
- Basen löschen/hinzufügen
- Basenpaare umbenennen
- Basenpaare hinzufügen/löschen

Tree-Alignment: Erzeugung eines common Super-Trees

4.4 lokales Falten

Vorhersagequalität nimmt mit Moleküllänge ab
Viele Moleküle haben keine globale Struktur aufgrund der Interaktionen in der Zelle

$$C'(i,j) = \begin{cases} C(i,j) \; if \; j - i < x \\ \infty \; else \end{cases} \tag{13}$$

→ Sliding-Window-Approach

→ lokales Backtracking mit

$$\textbf{RNALFold} = \begin{cases} F(n,n) \; (if \; F(i,n) < F(i+1,n)) \\ NOTHING \; (else) \end{cases} \tag{14}$$

- Dinukleotidshuffling
  → klärt Frage, wie stabil die Struktur gegenüber ähnlichen Strukturen ist
  → Annahme: ähnliche Sequenzen sollten gleichen Dinukleotidinhalt haben

4.5 RNA-RNA-Interaktionen (RNA-Interferenz)

Durch das Falten von zwei RNA-Molekülen kommt es zu sogenannten RNA-RNA-Interaktionen oder RNA-Interferenzen (kurz: RNAi).
RNAi ermöglicht eine spezifische Steuerung von Wechselwirkungen.

Basenpaarregelungen und Stacking-Gesetzmäßigkeit gelten sowohl intra- als auch intermolekular

RNAi dient der Gerüstbildung und Erzeugung von RNA-Enzymen:

- Spliceosomen
- snoRNA/rRNA
- bakterielle sRNA (inhibieren virelle mRNA in Weiterverarbeitung)
- miRNA (inhibieren virelle mRNA in Weiterverarbeitung)

Hierbei gibt es zwei Typen von kurzen mRNA-Sequenzen, die helfen virale mRNA in ihrer Weiterverarbeitung zu inhibieren (miRNA und sRNA)

Zur Vorhersage der RNA-RNA-Interaktionsstruktur können verschiedene Abstraktionsstufen genutzt werden. Das vollständige Energiemodell (sequentielle Vorhersage: zwei Moleküle werden als ein Molekül betrachtet und gefalten) ist sehr komplex mit $O(n^6)$.

- beliebig viele Interaktionsstellen möglich
- Intermolekulare Basenpaare innerhalb von intramolekularen Loops
  → zum Beispiel Kissing Hairpins
- keine intramolekularen Pseudo-Knoten
- keine überschneidenden intermolekularen Basenpaare

- intermolekulare Bp = externe Bp
- intramolekulare Bp = interne Bp
- ancestrale Bp = interne Bp, die externe Bp einschließen
- Eltern-Bp = ancestrale Bp mit minimaler Distanz

- subsumierende Bp = ancestrale Bp von jeweiliger Sequenz A,B, wobei A alle externen Bp, wie B auch einschließt

Dadurch können geschlossene Strukturen definiert werden:

- ein einzelnes externes Basenpaar
- die äußeren externen Basenpaare + Eltern-Bp
  - äqivalente Eltern-Bp
  - Elter A subsumiert Eltern B
  - Elter B subsumiert Eltern A

→ Damit kann die RNA-RNA-Interaktion in feste Bereiche vollständig zerlegt werden, die entweder geschlossene Strukturen sind oder aus rein intramolekulare Sequenzen bestehen (Reidys, Stadler, 2009)

Die Komplexität kann auf $0(n^3)$ durch Zusammenfassen von Termen reduziert werden. Hierbei werden beim Forward- und beim Backward-McCaskill-Algorithmus zusätzliche Matritzen eingespeichert. Die Backward-Matrix dient der Bestimmung der Zugänglichkeit der interagierenden Moleküle.

4.5.1 RNA miteinander falten und konkatenieren

Im Allgemeinen geht man wie folgt vor um solche RNAi zu bestimmen:

- Ermittle die gemeinsame Struktur von zwei RNA-Molekülen
- Finde die Bindestellen von kleineren RNA-Fragmenten
- RNA miteinander falten (ohne Pseudoknoten(**) herzustellen)

Festlegung: Der Loop mit Konkatenationsstelle ist der externe Loop.
(**) Pseudoknoten sind sich überkreuzende Basenpaare und kommen auch in Natur vor (z.B. Kissing Hairpins, H-Typ). Der Ausschluss ermöglicht eine polynomiale Berechnung der Struktur
→ simple Pseudoknoten können mit dem RNAPKplex aus dem Vienna RNA-package gelöst werden $O(n^3)$

Zur Vereinfachung werden Sequenzabschnitte vorhergesagt, die regulatorische Relevanz haben. Die Sequenzen werden vereinfacht und dann gescannt. Die vorhersage beruht auf zwei Termen:
$\Delta G_{Bindung} = \Delta G_{Opening} + \Delta G_{Interaktion}$
Bei der Ermittlung intermolekularer Helices werden intramolekulare Interaktionen nicht berücksichtigt (Veringerung der Anzahl an Interior Loops). Die Energiebeträge der Teilsequenzen werden bestimmt und gespeichert. Wahlweise kann mit diesem Verfahren entweder die minimum free Energy oder die

Partition-Funktion bestimmt werden.
Besonderheit: Das erste intermolekulare Basenpaar erhält statt Hairpin-Energie eine Entropie-Strafe.
Um Rechenzeit zu sparen kann die maximale Länge der Interaktionsstelle mit einer Maximallänge beschränkt werden ($O(n^2m^2)$). Die minimalen Interaktionsenergien der sequentiellen Teilabschnitte werden abgespeichert. Die gemeinsame Bestimmung von mfe und Zustandssumme benötigt $O((n+m)^3)$
Die Wahrscheinlichkeit einer Dimerbildung zweier RNA-Moleküle ist jedoch konzentrationsabhängig.
Anmerkung: Die Betrachtung von mehr als zwei Molekülen ist möglich, aber deutlich rechen- und speicherintensiver, da die Zahl der Rekombinationen stark ansteigt. Ein nutzbarer Algorithmus ist von Dirks et al.

4.5.2 RNAplex

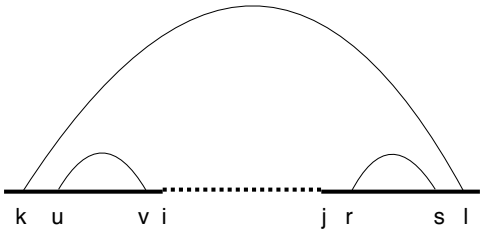
RNAplex ist ein Alignemnt-ähnlicher Ansatz zur Untersuchung von RNAi
Die Interaktionsenergie kann hier schneller bestimmt werden ($O(nmL^2)$). Da die Energie von Interiorloops nicht logarithmisch betrachtet wird sondern in einer linearen Regression (Vernachlässigung von Assymetrien, somit aber auch Vernachlässigung der RNA-Struktur an sich).

- Weiterhin ist die Bindewahrscheinlichkeit davon abhängig, wie gut zugänglich das Target ist → Zugänglichkeit von Base i: $z_i = 1 - \sum_{i!=j} p(i,j)$
- Wahrscheinlichkeit, dass (i,j) ungepaart ist (entspricht der Öffnungs-Energie: $p = \frac{Z((i,j)\ ungepaart)}{Z(1,n)}$
- Struktur-Graphik für Formeln:

$$Z^n(i,j) = Z(1,i-1) + Z(j+1,n) + \sum_{k<i<j<l} Z^B(k,l) * \frac{Z^B(k,l)}{Z^{Bu}(k,l)} \tag{15}$$

$$Z^{Bu}(k,l) = \begin{array}{r} H(k,l) + \sum_{k<u<v<i} I(k,l;u,v) * Z^B(u,v) + \sum_{j<r<s<l} I(r,s;k,l) * Z^B(r,s) \\ +M(k+1;i-1) * M(j+1,l-1) + M^2(k+1,i-1) + M^2(j+1,l-1) \end{array} \tag{16}$$

Die Energiewerte der Interior-Loops und 1-Bulge-Loops werden aus einer Standardtabelle mit Matthews-Parametern ausgelesen.



5 Neutrale Netzwerke von RNA-Strukturen (Peter Schuster)

Diese Methode erlaubt Aussagen über die Evolvierbarkeit von Strukturen zu formulieren. Es gibt konservierte Strukturen, welche stabil gegen Mutation sind. Für neutrale Netze ist festgelegt, dass aus einem neutralen Netz alle weiteren neutralen Netze mit einer Mutation erreicht werden können.

5.1 Shape-Abstraktion (R. Giegerich)

Simplifizierung einer Sequenz um diese vergleichbar mit anderen Sequenzen zu machen. Erzeugung sogenannter suboptimaler Shapes:

- ((.(((...))).....((((.....)))))..(.((..))).)) (Ausgangszustand)
- [.[.].[.].[.[.]].]
- [.[.].[.].[.].]
- [[][][]] (stärkstes sinnvolles Abstraktionslevel)

SHREP = SHape REPresentatives: Als Ergebnis wird die Wahrscheinlichkeit der besten Struktur der Shapes ermittelt.

5.2 Faltungskinetik mit Energielandschaften

Kinetische Überlegungen im Falten von RNA:

- Moleküle in biologischen Systemen liegen meistens nicht im thermodynamischen Gleichgewicht
  → Entstehung von kinetischen Fallen == minimum free Energy (tiefe lokale Minima in der Energielandschaft)
- kinetische Fallen können zum Beispiel durch RNAi (cotranskriptionales Falten) erzeugt werden
- sogenannte RNA-Schalter wechseln von einem lokalen Minimum in ein anderes und falten somit absichtlich von einem in den anderen Zustand
  → metastabile RNA-Faltungszustände (erzeugen zusätzlichen Regulationsfaktor)
- freie Energie pro Struktur
- Wahrscheinlichkeit für bestimmte freie Energie ΔG abhängig von der Zeit

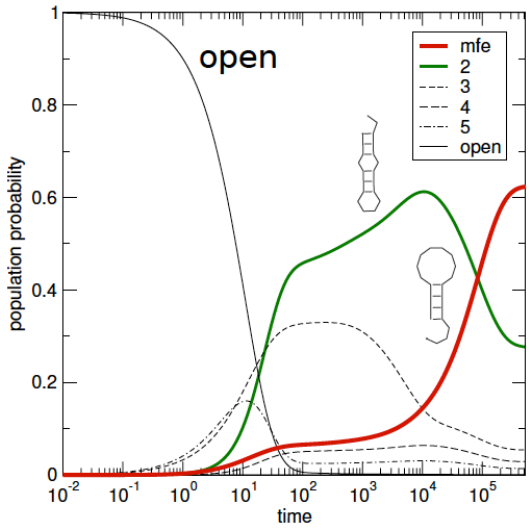


Abbildung A

Abbildung B

Abbildung A zeigt die Wahrscheinlichkeit eine Struktur einer RNA-Sequenz zu bestimmter Rechenzeit zu ermitteln. Hierbei zeigen die verschiedenen Kurven verschiedene Bereiche der Energielandschaft. Rot entspricht der minimalen freien Energie.
Abbildung B zeigt eine Energielandschaft, also die energetischen Niveaus einer RNA-Sequenz in Abbhängigkeit ihrer Strukturzustände.

Erlaubte Schritte sind (Move-Set):

- öffnen von Basenpaaren
- schließen von Basenpaaren
- Verschiebung von Basenpaarteilen

5.2.1 Metropolis-Monte-Carlo

Die Anzahl zu betrachtender Zustände ist zumeist viel zu groß, weswegen mit stochastischen Methoden, wie Monte-Carlo und oder Markov-Prozessen gearbeitet werden muss.
Ein Schritt in Richtung niedriger Energie ist leichter als in Richtung höherer Energie.
Nachteil: Methode funktioniert nur für kleinere Probleme zufriedenstellend → grobkörniger Ansatz nötig: Erzeugen einer Übergangsmatrix k (exponentielles Wachstum) oder helixbasierte Move-Sets
Im ersten Schritt wird ein Zustandswechsel i → j vorgeschlagen. Überprüfe, ob **Zustand** besser ist → Metropolis-Regel:

$$p_{Akzeptanz}(i,j) = \begin{cases} e^{-\frac{\Delta G_j - \Delta G_i}{RT}} & if\ \Delta G_j > \Delta G_i \\ 1 & else \end{cases} \tag{17}$$

alternativ Kawasaki-Regel:

$$K_{ij} = \frac{e^{-\frac{\Delta G_j - \Delta G_i}{RT}}}{2} \tag{18}$$

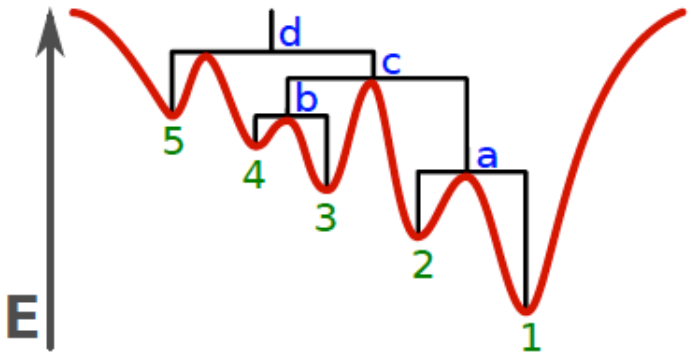
5.2.2 Barrier-Trees

1 Berechnung aller suboptimalen Strukturen des Energiebandes (→ Wuchty-Algorithmus)

2 Bauen eines Baums aus allen suboptimalen Strukturen

3 Faltungspfad zwischen Strukturzuständen i und j → Sequenz von erlaubten Schritten, die aus Anfangsstruktur i die Endstruktur j erzeugt

4 Wahl des optimalen Faltungspfads: Faltungspfad mit minimalen Maximum der Energie der Strukturen, die besucht werden



Die Abbildung zeigt einen Barriertree, der aus den Minima und den Sattelpunkten der Energielandschaft erzeugt wurde.

5.2.3 Baumbau mit Flooding-Algorithmus

Erzeugung einer Strukturliste, die nach Energiewerten sortiert ist. Diese wird dann als Hash eingespeichert.
besuchte Struktur → Niveaunummer

- Beginn bei Struktur 0 mit Niveaunummer 0
- Besuchen aller Nachbarn → Überprüfe im Hash
  - a gibt es keinen Nachbarn im Hash → neues Minimum im Hash
  - b haben alle die selbe Niveaunummer → alle gehören einem Hash an
  - c Nachbarn in zwei Niveaustufen → Sattelpunkt an der Stelle wo beide Niveaus zusammenkommen
- Berechnung der Wahrscheinlichkeit nach Arrhenius:

$$p(x, y) = A * e^{-\frac{E_{Sx} - E_{Sy}}{RT}} \tag{19}$$

5.2.4 Direkte Pfade

Die Nutzung direkter Pfade stellt eine Alternative zum Flooding-Algoithmus dar.
Gegeben sind zwei Sets von Basenpaaren (A,B)
a → b: erlaube nur Verschiebungen, die entweder Basenpaare A außer B wegnehmen oder ein BP aus B außer A hinzufügen.
Heuristiken zur Bestimmung von guten direkten Pfaden:

1. Morgan-Higgs-Heuristik: Wahl des besten Schritts
2. Find-Path-Heuristik: Generiere alle möglichen Schritte und wähle davon die fünf besten aus

5.2.5 Cotranskriptional Folding

Programm: KinWalker

Barmap-Ansatz:

1 Zufälliges Wählen eines Barrier-Tree

2 Simulieren des Barriertrees auf einen neuen Tree und damit die Energielandschaft abändern

3 Mappe die Niveaus von Barrier-Tree I zu denen von Barrier-Tree II

4 Erneut zu Punkt 2 und mit Barrier-Tree II weitersimulieren

5.3 Turner-Modell (Nearest-Neighbor-Modell)

5.4 Zuker-Algorithmus

5.4.1 suboptimales Falten

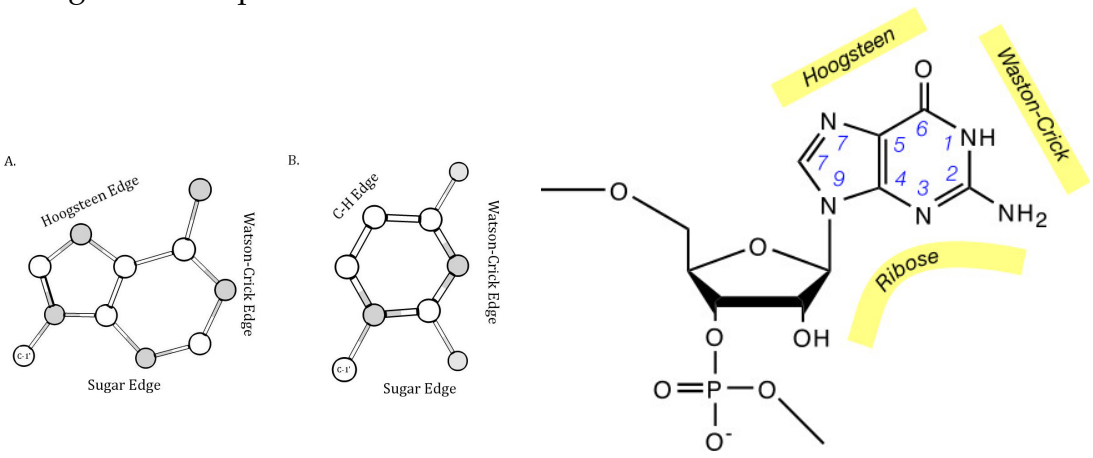
5.5 Wuchty-Algorithmus

5.5.1 Wuchty-Backtracking

5.6 McCaskill

6 weitere Bindungsarten, erlaubte Basenpaare

- Hoogsteen base pair[5]

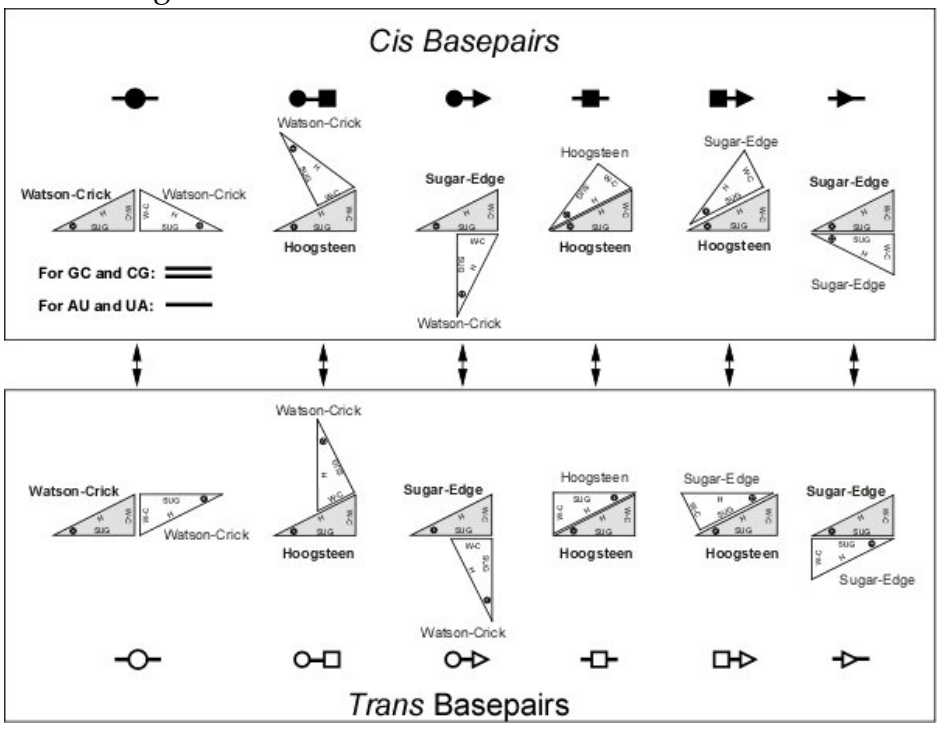


Jede Base kann mit jeder ihrer Kanten zu jeder Kante jeder Base ein Basenpaar bilden.

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?

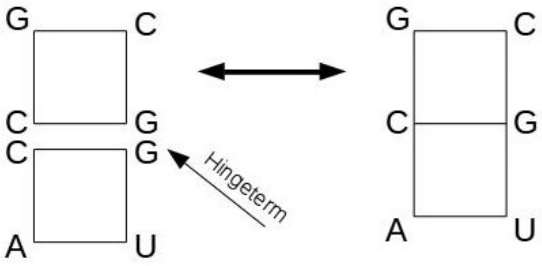
Darstellung:



[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



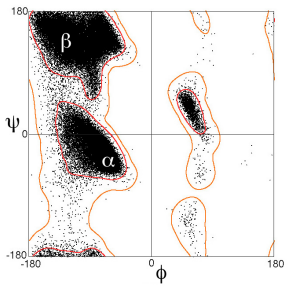
Programme: MC-Fold, RNAWolf

7 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

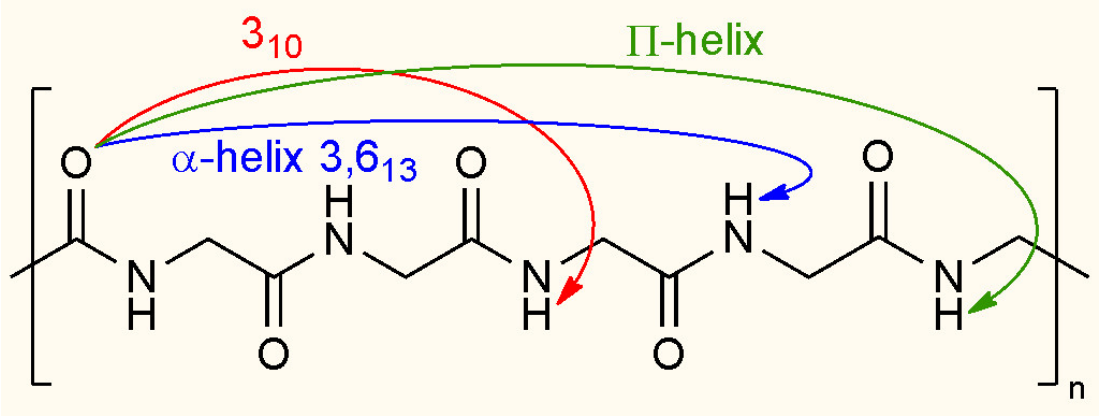
8 Sekundärstrukturelemente

- 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



8.1 Chou-Fasman (Sekundärstrukturvorhersage von Proteinen)

- ca. 50% Genauigkeit

- 3 Tabellen mit Scores für α (Helix), β (Faltblatt) und t (Turn) für alle Aminosäuren

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

  - 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

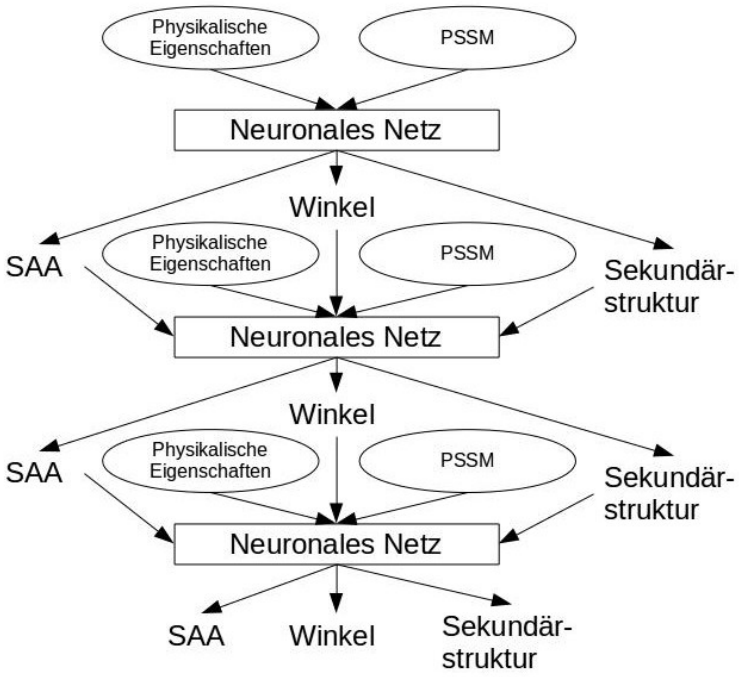
Physikalische Eigenschaften von Aminosäuren:

- sterischer Parameter (graph shape index: dünnes oder dickes Molekül)

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

- 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:



w**eitere Möglichkeit: Meta Server**

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

9 (Protein-) Strukturvorhersage (3D)

9.1 Strukturaufklärung

- Röntgen-Kristallographie
- NMR

9.2 Qualität der Strukturvorhersage

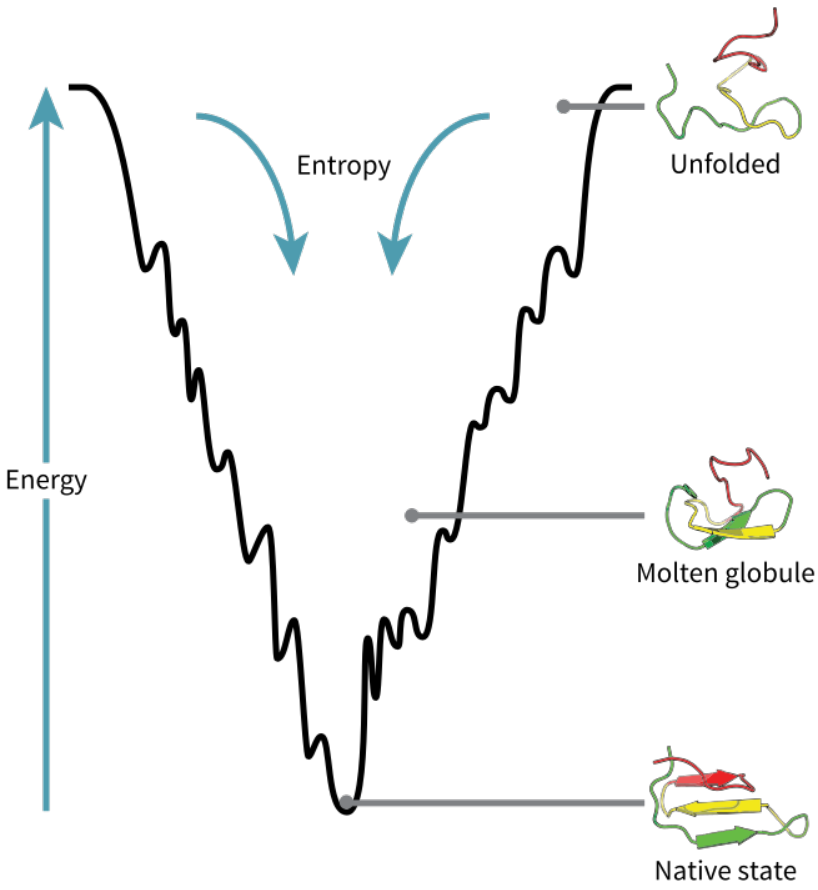
- RMSD (root mean square deviation): mittlerer Abstand in Å (10^{-10} m)

1-2 Å RMSD ist ein (sehr) guter Wert.

9.3 Problem der Strukturvorhersage (Levinthal-Paradoxon)

Eine Polypeptidkette von 100 Residuen hat 99 Peptidbindungen und daher 198 verschiedene Φ- und Ψ-Winkel. Wenn nun jeder dieser Winkel drei stabile Konformationen einnehmen kann, so ergibt sich eine Anzahl verschiedener Proteinstrukturen von 3^{198}. Wenn die Peptidkette bei der Faltung zum Protein nacheinander jeden dieser Winkel ausprobieren würde, bräuchte es länger als der Alter des Universums, um korrekt zu Falten.[10]

[10] `https://en.wikipedia.org/wiki/Levinthal_paradox/`



In der Natur falten Proteine aber im Bereich von Millisekunden. Wie ist das zu erklären?
Lokale Interaktionen führen den Faltungsprozess und schränken die Möglichkeiten ein. Experimente zeigen die resultierenden Intermediates und Transition states. Struktur und Faltung sind also *sequenzkodiert*.

9.4 Protein-Domains (Domänen)

- Protein-Untereinheit
- Falten unabhängig vom Rest des Proteinstrukturen
- Meistens funktionelle Untereinheit
- ca. 2700 Familien
- ca. 120000 Proteine (pdb)
- ca. 2/3 sind Multidomain-Proteine
- ca. 1224 Folds
  - Folds (SCOPe-Datenbank) Structural classification of proteins

  - All α
  - All β
  - α/β, abwechselnde $\alpha/\beta \Rightarrow$ parallele β-Faltblätter
  - $\alpha + \beta$, getrennte $\alpha, \beta \Rightarrow$ antiparallele β-Faltblätter
  - Multidomain ($\alpha + \beta$)
  - Andere (coiled coil, membrane, cell-surface)

9.5 Zwei Typen von Vorhersagen

- Ab initio
- Template based
  - Homology based
  - Threading

9.5.1 Ab-initio-Vorhersage

- Suche Strukturvorschläge
- Bewerten der Strukturen
  - physikalisch
  - knowledge-based $log(\frac{observed}{expected})$

Physikalisch Molecular force field

$E_{Bindung}$

- Bindungen
  - Abstand
  - Winkel α (Bindung)
- Winkel ϕ (Torsion)

$E_{ungebunden}$

- Ladungen
- Dipol

$E = E_{Bindung} + E_{Nicht-Bindung}$
$E_{Bindung} = \sum_{\alpha} k(\alpha - \alpha_0)^2$
$+ \sum_{Bindungen} k(r - r_0)^2$
$+ \sum_{\phi(Torsion)} \frac{V_n}{2}(1 - \cos(n\phi - \gamma))$
Alternative für Bindungspotential (Morse-Potential[11]) $\sum_{Bindungen} D_e*(1-e^{-a(r-r_0)^2})$

k Kraftkonstante

α Bindungswinkel

r Bindungslänge

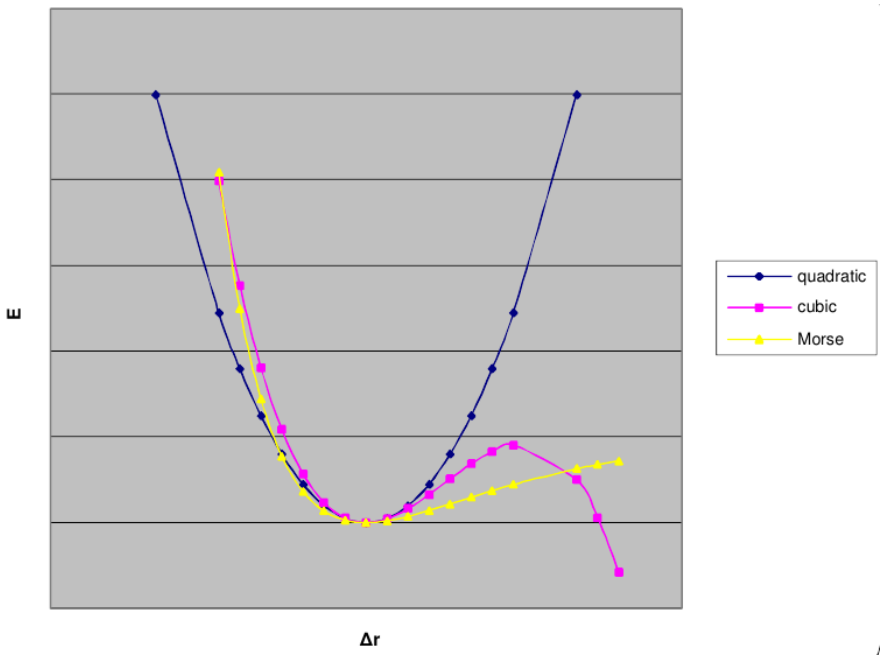
r_0 Bindungslänge mit der geringsten potentiellen Energie

D_e Dissoziationsenergie

$a = (0.5 * \frac{k}{D_e})^{1/2}$ "Steifigkeits-"Konstante

V_n Barrier height

γ Phasenverschiebung



$E_{\text{Nicht-Bindung}} =$

$\sum_{i,j \in Atome} \frac{P_i P_j}{\epsilon r_{i,j}}$ Ladung: Coulomb-Terme

$+ \sum_{Paar} \frac{c}{r^{12}} - \frac{c}{r^6}$ Dipol: Van-der-Waals-Kräfte, Lennard-Jones-Potential 12, 6

cut-off-radius

[11] https://en.wikipedia.org/wiki/Morse_potential

Lösungsmittel

- implicit solvent
- explicit solvent

Spezialterme: H-Terme, Π-Interaktionen

Knowledge based

- coarse-graining
- c_α als Beschreibung der AS
- Alle Backbone-Atome
- Alle Backbone-Atome + repräsentativ die Sk (center of mass)
- Alle Atome

Einfache Potentiale

- Abstand der Aminosäuren
- Nachbarschaft der Aminosäuren

QUARK

- Backbone atomweises Paar-Potential
- Sk-Schwerpunkt
- Excluded volume
- H-Bindungen
- Surface accessible area
- Torsionswinkel im Backbone
- Distanzen von Fragmenten
- Gyrationsradius
- Relative Position der Strukturen (Faltblatt/ Helices)
  - $\beta\alpha\beta$-linkshändig
  - $\beta\alpha\beta$-packing
  - α-packing

- β-packing

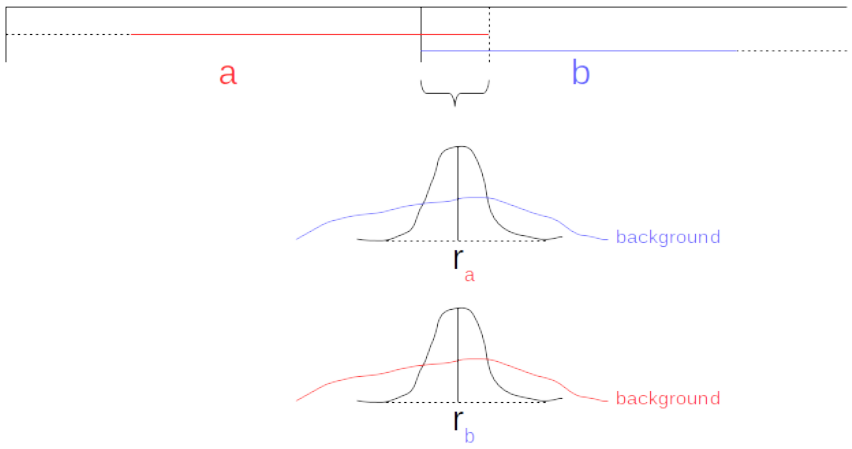
Konformation $\Rightarrow$ dafür Minimum

- Steepest descent
- Conjugate gradient
- Newton-Verfahren
- Monte-Carlo-Verfahren
- Simulated annealing

9.5.2 Template based methods

Homology based

- Sequenzalignment zu den Sequenzen der bekannten Strukturen
- Alignment der Sequenz zur Struktur des Kandidaten
- Bauen einer Struktur aus dem Alignment
- Bewerten der Struktur



- Verbinden der Distanzpotentiale (gewichtet, multiplikativ)

CASP (critical assessment of structure prediction)

Threading

1. Sequenz-Struktur-Alignment zur Identifizierung der Kandidaten
$$\begin{pmatrix} \alpha \\ 0.5 \\ P \\ groß \end{pmatrix}\begin{pmatrix} \alpha \\ 0.3 \\ P \\ klein \end{pmatrix}$$
2. Bauen einer Struktur aus Alignment
3. Bewerten der Struktur