

Bioinformatik: Sequenzanalyse

RNA-Sekundärstrukturvorhersage

Benedikt Brors

Abt. Intelligente Bioinformatiksysteme
Deutsches Krebsforschungszentrum, Heidelberg



dkfz

- RNA ist mehr als nur ein passiver Träger der Erbinformation
- RNA kommt z.B. vor in:
 - ★ Ribosom (rRNA)
 - ★ Spliceosom (snRNAs)
 - ★ signal recognition particle (SRP)
 - ★ tRNAs
 - ★ siRNAs
 - ★ Ribozyme (Nobelpreis 1989)



dkfz

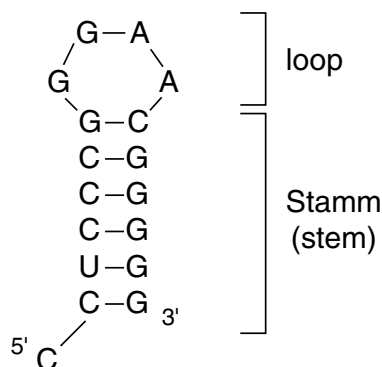
RNA-Struktur

- oft ist die Sekundär- und Tertiärstruktur entscheidend für die Funktion
- Sekundärstruktur ist meist besser konserviert als die Sequenz
- Deshalb sind Sequenzvergleiche bei RNA schwieriger als bei DNA: Um die Sequenzen vergleichen zu können, braucht man die Struktur, und um die Struktur zu finden, braucht man den Vergleich
- Man kann nicht alle Strukturen berechnen: eine 200 nt lange RNA hat mehr als 10^{50} mögliche Strukturen!



Elemente der RNA-Struktur

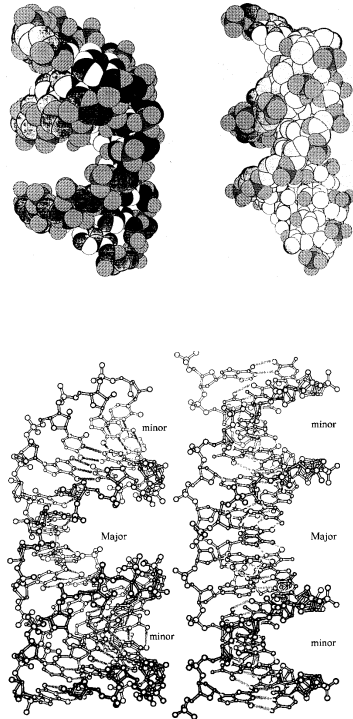
- 4 Nukleotide: A, C, G und U
- Basenpaarung ist möglich durch interne Stammbildung:



- Stamm bildet eine A-Helix aus



A- und B-Helices

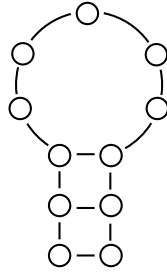


Konsequenzen

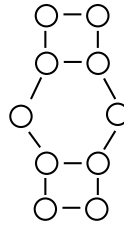
- Nicht-Watson-Crick Basenpaarungen sind möglich,
- Bsp: "Wobble"-Paar G-U, aber auch andere
- in speziellen Strukturen sind auch Basentripel möglich (keine Methode zur Vorhersage verfügbar)



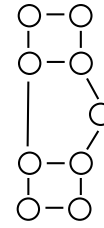
Strukturelemente



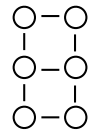
hairpin loop



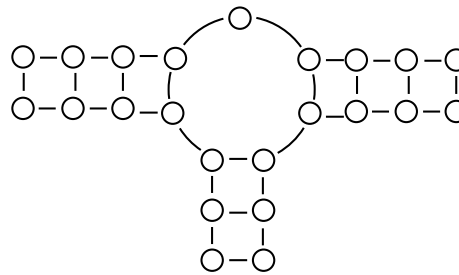
internal loop



bulge loop



stem

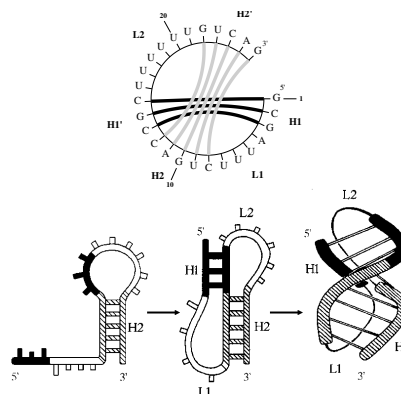


multi loop



Pseudoknoten

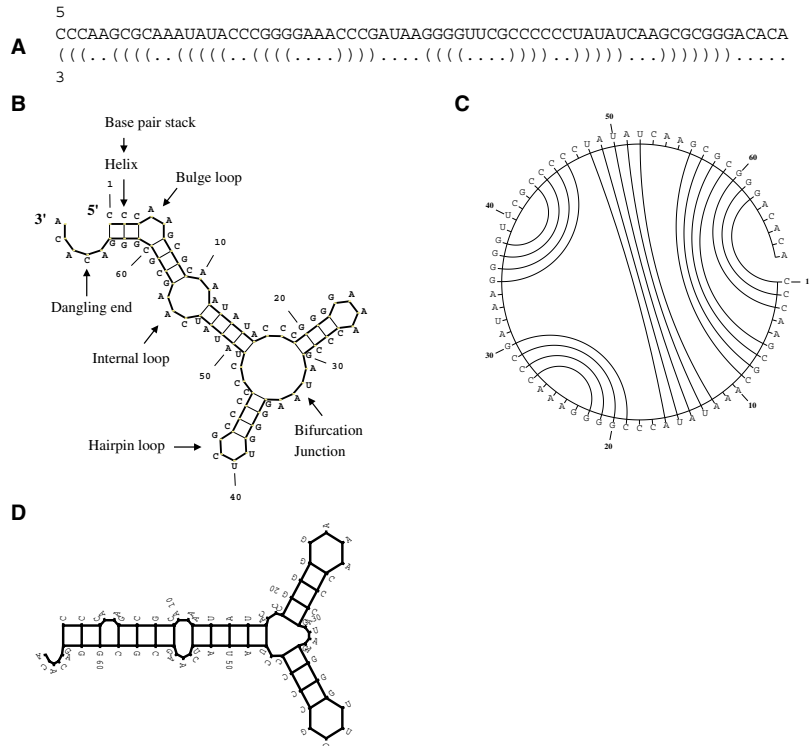
- Als Pseudoknoten bezeichnet man eine Struktur, bei der ein Loop mit einem Sequenzstück außerhalb des Loops paart:



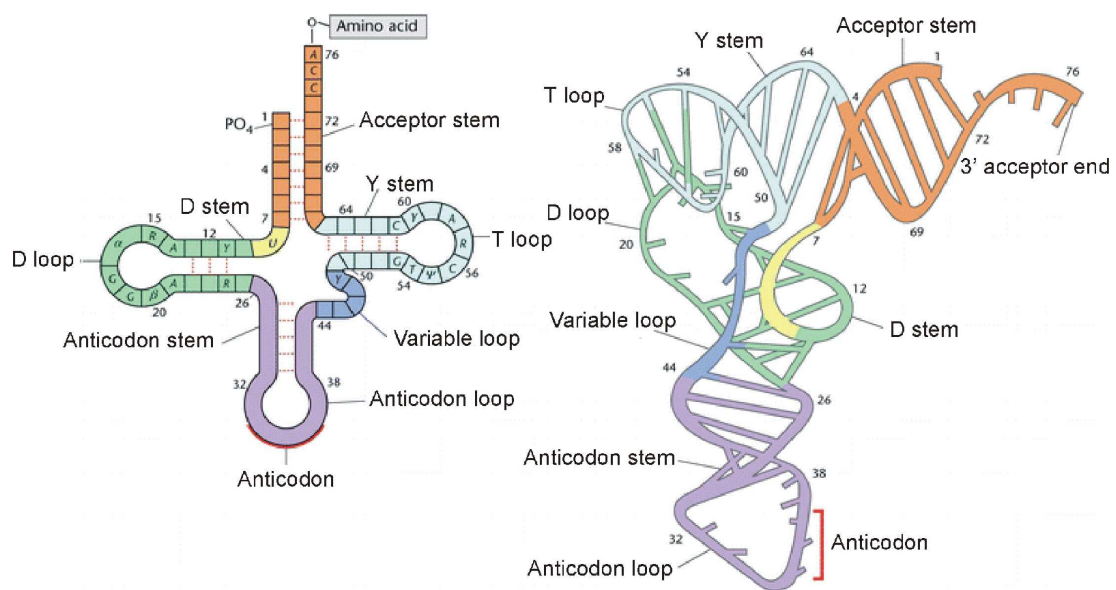
- Pseudoknoten werden meist bei der Strukturvorhersage nicht berücksichtigt



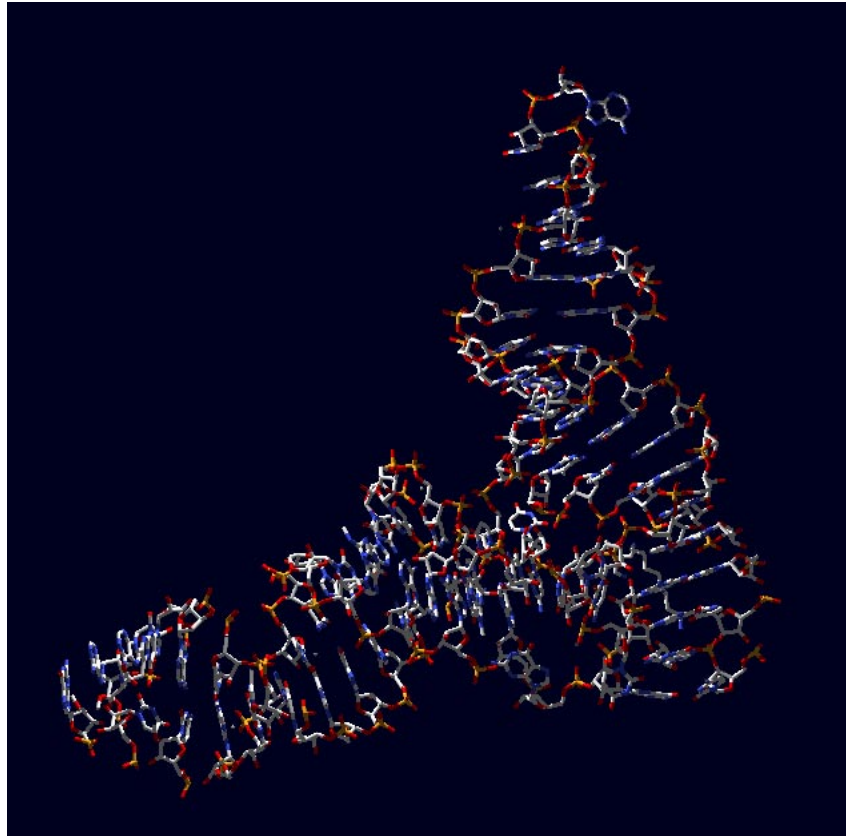
Strukturrepräsentationen



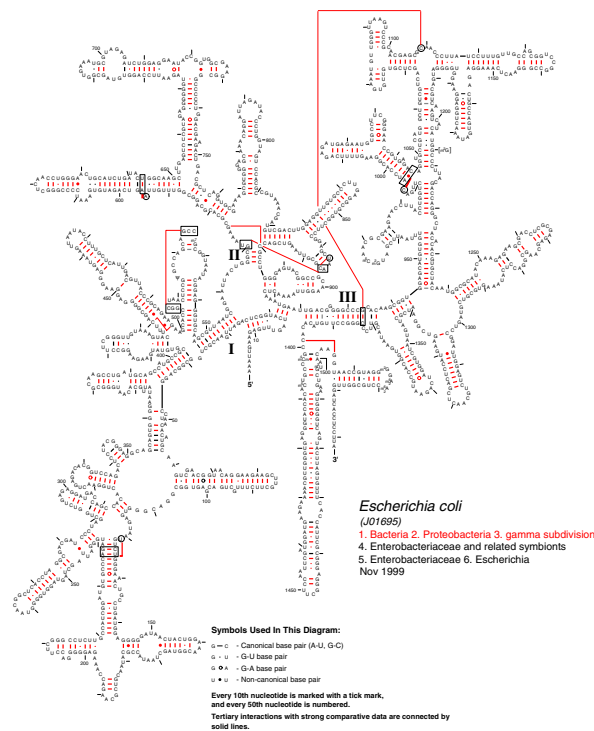
tRNA-Strukturen



Hefe Phe-tRNA (Röntgenkristallstruktur)



Komplexe Strukturen: SSU rRNA



Strukturvorhersage

- die meisten Basenpaarungen in realen Strukturen sind verschachtelt (*nested*), d.h. $i - j$ und $i' - j'$ bedingt:

$$i' < i < j < j' \quad \text{oder} \quad i < i' < j' < j$$



- Pseudoknoten und andere nicht-verschachtelte Strukturen bleiben unberücksichtigt
- Man definiert meist Mindestlängen für hairpin loops (z.B. 3) und interne loops (z.B. 4)
- Meist werden nur Watson-Crick-Basenpaare (evtl. auch G-U) berücksichtigt



dkfz

Algorithmen

- Die einfachste Methode ist der *Nussinov-Algorithmus*
- die Anzahl der Basenpaare einer möglichen Struktur wird maximiert
- Man startet mit einer kleinen Substruktur. Um diese zu erweitern, gibt es 4 Möglichkeiten:
 - ★ ungepaarte Position am 5'-Ende anfügen
 - ★ ungepaarte Position am 3'-Ende anfügen
 - ★ Basenpaar an stem anfügen
 - ★ 2 optimale Substrukturen kombinieren



dkfz

Scoring

- Die Score-Funktion addiert 1 für jedes Basenpaar, und 0 sonst
- Der Score wird durch dynamische Programmierung optimiert:

	G	G	G	A	A	U	C	C
G	0	0	0	0	0	1	2	3
G	0	0	0	0	0	1	2	2
G		0	0	0	0	1	2	2
A			0	0	0	1	1	1
A				0	0	1	1	1
U					0	0	0	0
C						0	0	0
C							0	0

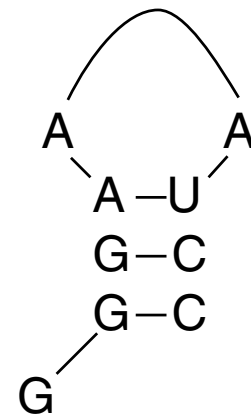


Finden der optimalen Struktur

- Die Matrix wird in Diagonalen gefüllt
- Zum Finden der Struktur wird ein Backtracking eingesetzt
- Das Backtracking benutzt bestimmte Regeln



	G	G	G	A	A	U	C	C
G	0	0	0	0	0	1	2	3
G	0	0	0	0	0	1	2	2
G		0	0	0	0	1	2	2
A			0	0	0	1	1	1
A				0	0	1	1	1
U					0	0	0	0
C						0	0	0
C							0	0



Minimierung der freien Enthalpie

- Erinnerung (physikalische Chemie):
Jeder Konformation eines Moleküls kann eine freie Enthalpie ΔG zugeordnet werden
- Es gilt die Helmholtz-Gleichung:

$$\Delta G = \Delta H - T\Delta S$$

- Unterscheiden sich 2 Konformationen um ΔG , und sind im niedrigeren Zustand N_0 Moleküle, dann sind im anderen Zustand N_1 Moleküle:

$$N_1 = N_0 \exp\left(-\frac{\Delta G}{kT}\right)$$

Dabei ist k die Boltzmann-Konstante



Minimierung der freien Enthalpie für RNA-Strukturen

- Jedem Element werden (additive) Energiebeiträge zugeordnet: Basenpaare wirken stabilisierend (ΔG negativ)
- Loops wirken destabilisierend (ΔG positiv)
- Die Hauptenergie der Basenpaarung kommt nicht durch die Wasserstoffbrücken, sondern durch stacking-Effekte



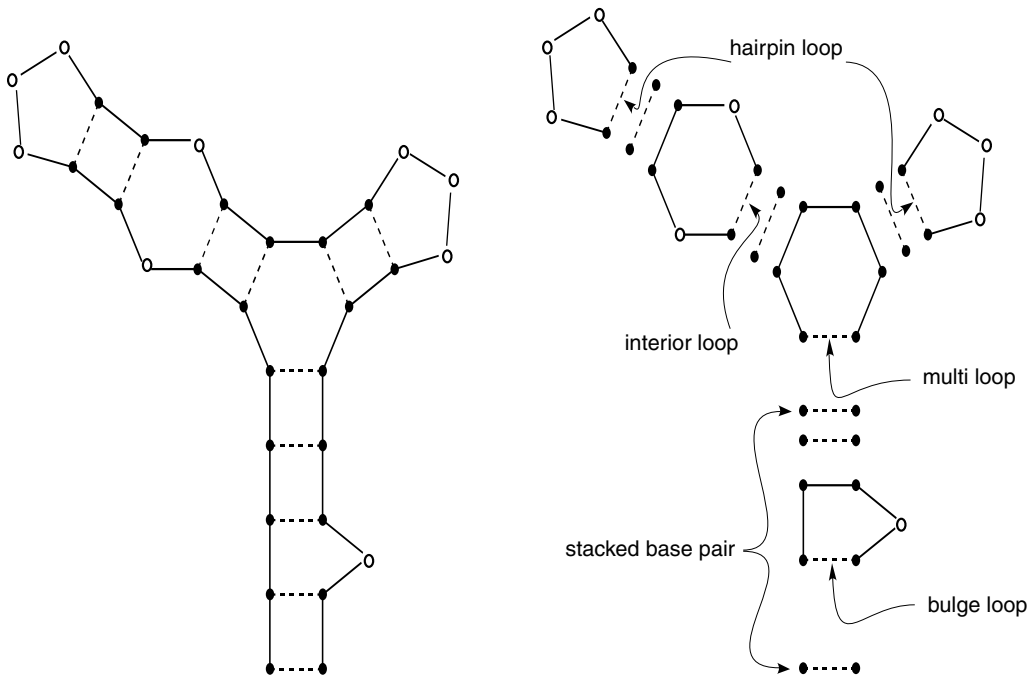
Beispielwerte

	CG	GC	GU	UG	AU	UA		3	4	5
CG	-3.3	-2.4	-1.4	-2.1	-2.1	-2.1	hairpin	5.7	5.6	5.6
GC	-3.4	-3.3	-1.5	-2.5	-2.4	-2.2	bulges	3.2	3.6	4.0
GU	-2.5	-2.1	-0.5	1.3	-1.3	-1.4	interior	-	1.7	1.8
UG	-1.5	-1.4	0.3	-0.5	-1.0	-0.6				
AU	-2.2	-2.1	-0.6	-1.4	-0.9	-1.1				
UA	-2.4	-2.1	-1.0	-1.3	-1.3	-0.9				

[Energieinkremente in kcal/mol]



Loop Decomposition



Minimierung der freien Enthalpie

- Die Minimierung der Scorefunktion, die die Energien addiert, heißt *Zuker-Algorithmus*
- Er benutzt dynamische Programmierung analog zum Nussinov-Algorithmus, hat aber andere Elemente als 0 und 1 in den Rekursionsformeln
- Man muß unterscheiden zwischen der besten Substruktur (i, j) (d.h. mit der niedrigsten freien Enthalpie), und der besten Substruktur, bei der (i, j) gepaart sind. Man benötigt deshalb zwei Tabellen, W und V

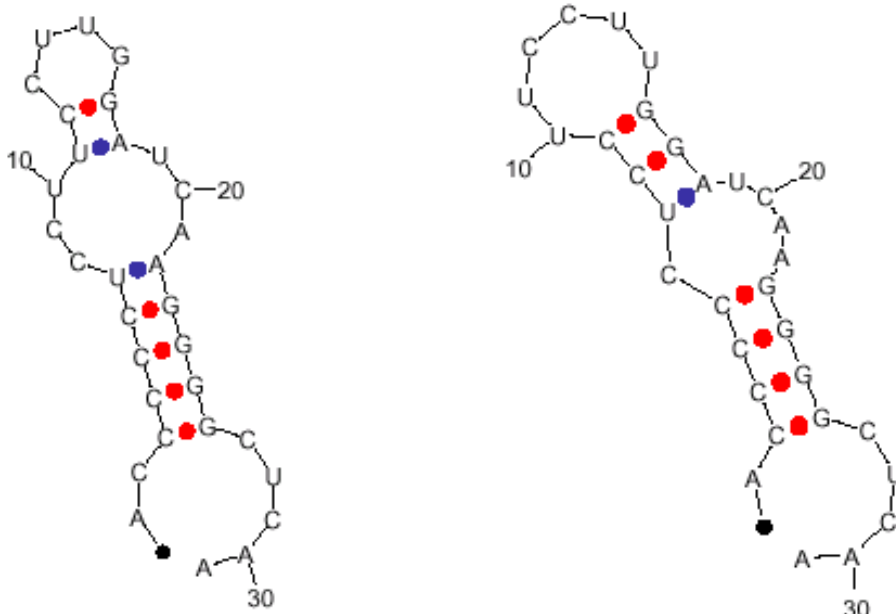


suboptimale Strukturen

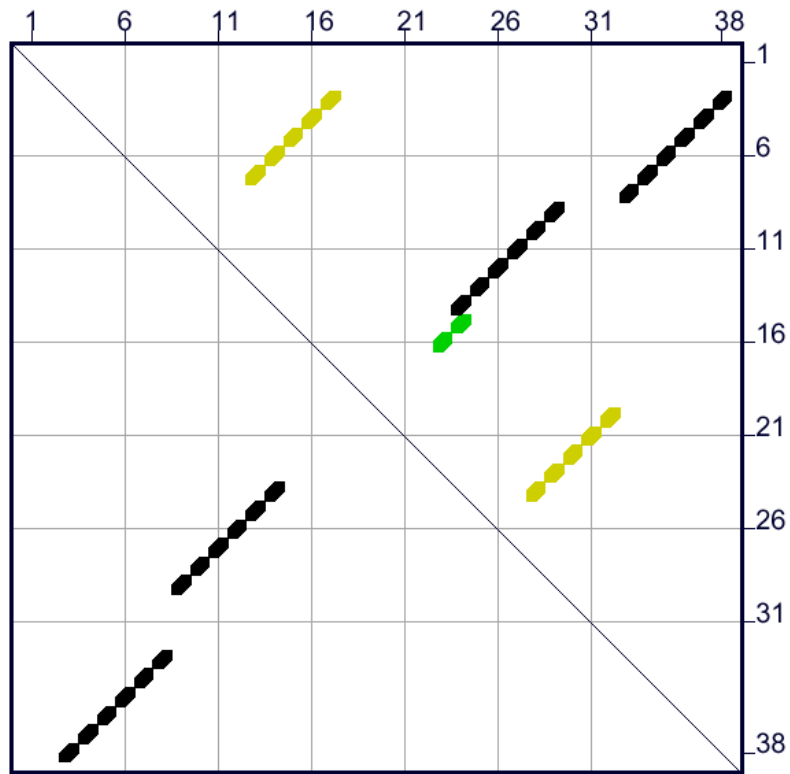
- oft sind reale Strukturen nicht die mit der minimalen Energie, haben aber sehr ähnliche Energiewerte
- Man benutzt Algorithmen (z.B. MFOLD), die alle RNA-Strukturen, die um $n\%$ von der optimalen Energie abweichen, berechnet
- Dazu werden einzelne Basenpaare in der optimalen Struktur variiert; es werden also nicht alle suboptimalen Strukturen erfaßt



Beispielstrukturen



Energie-Dot-Plot



Komplexeres Beispiel

