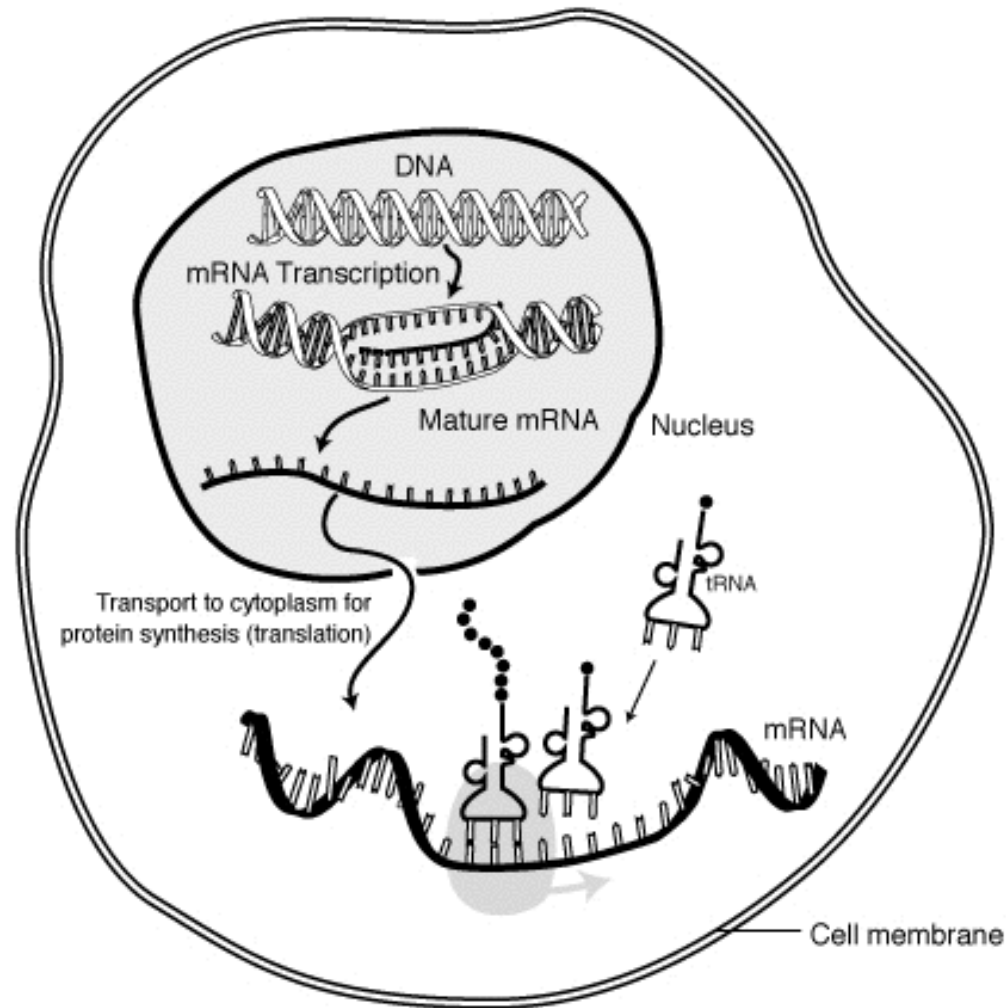


Vorhersage der Sekundärstruktur von Ribonukleinsäure (RNA)

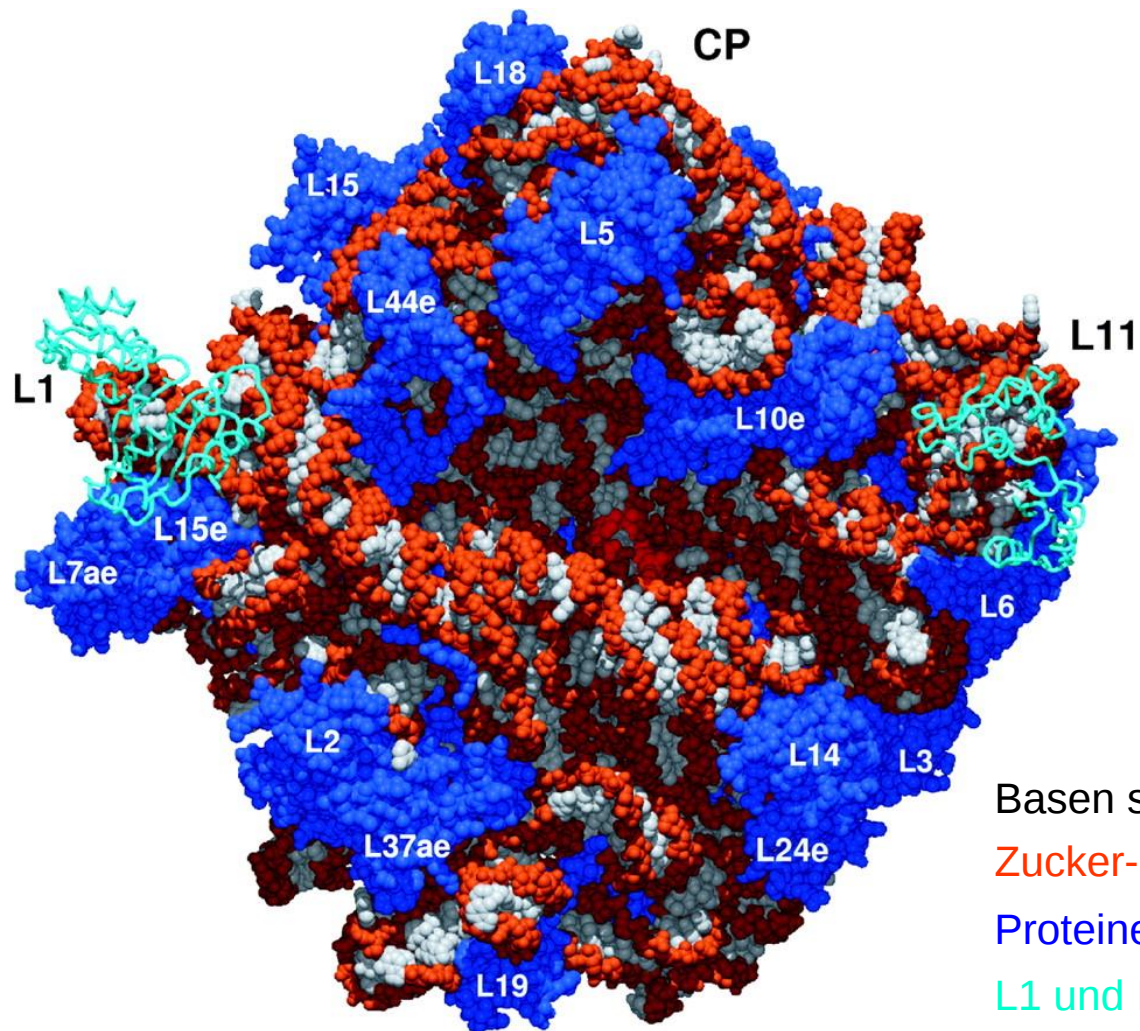
Möglichkeiten und Limitierungen

Uwe Menzel, Januar 2009

RNA und Proteinsynthese



RNA als Teil des Ribosoms



Basen sind weiss

Zucker-Phosphat Backbone

Proteine

L1 und L11 Proteine

Funktionen der RNA

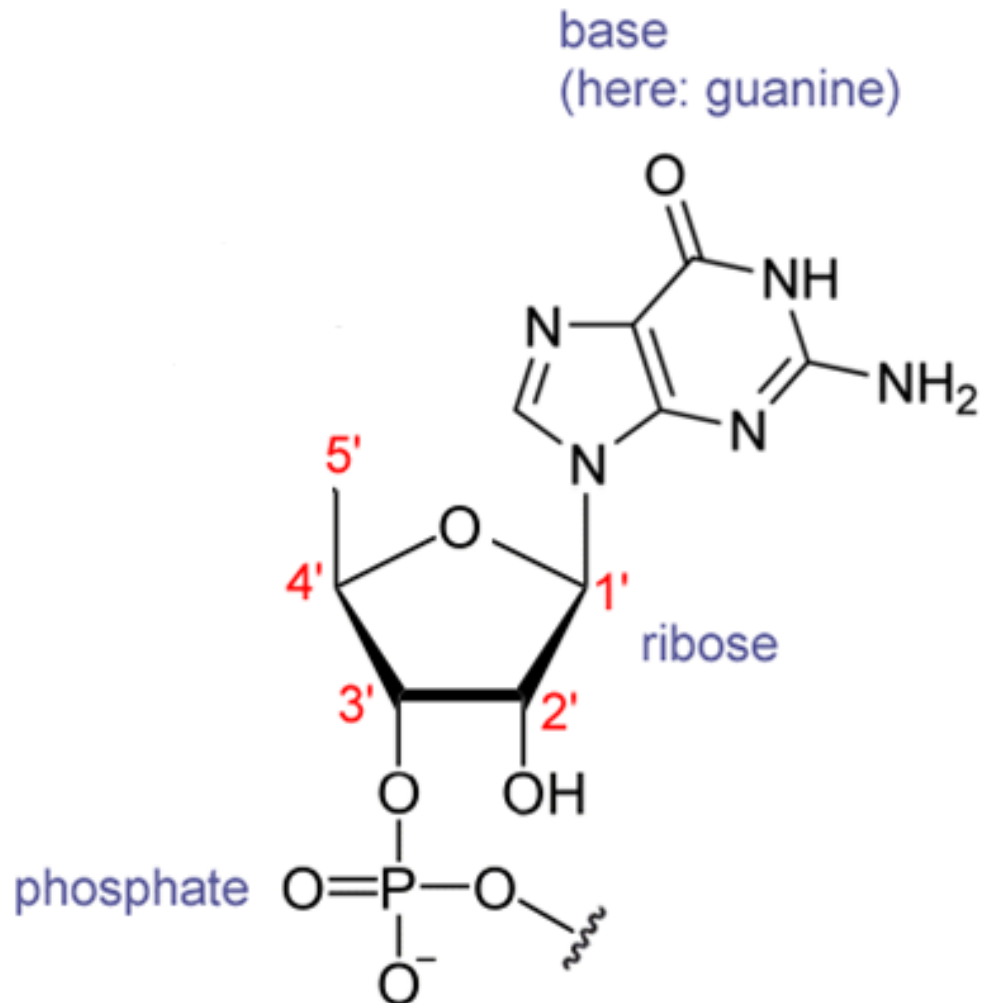
- Proteinsynthese (mRNA, tRNA)
- Teil des Ribosoms
- Spleissen von Introns
- Genregulation
- Immunsystem
-

Struktur

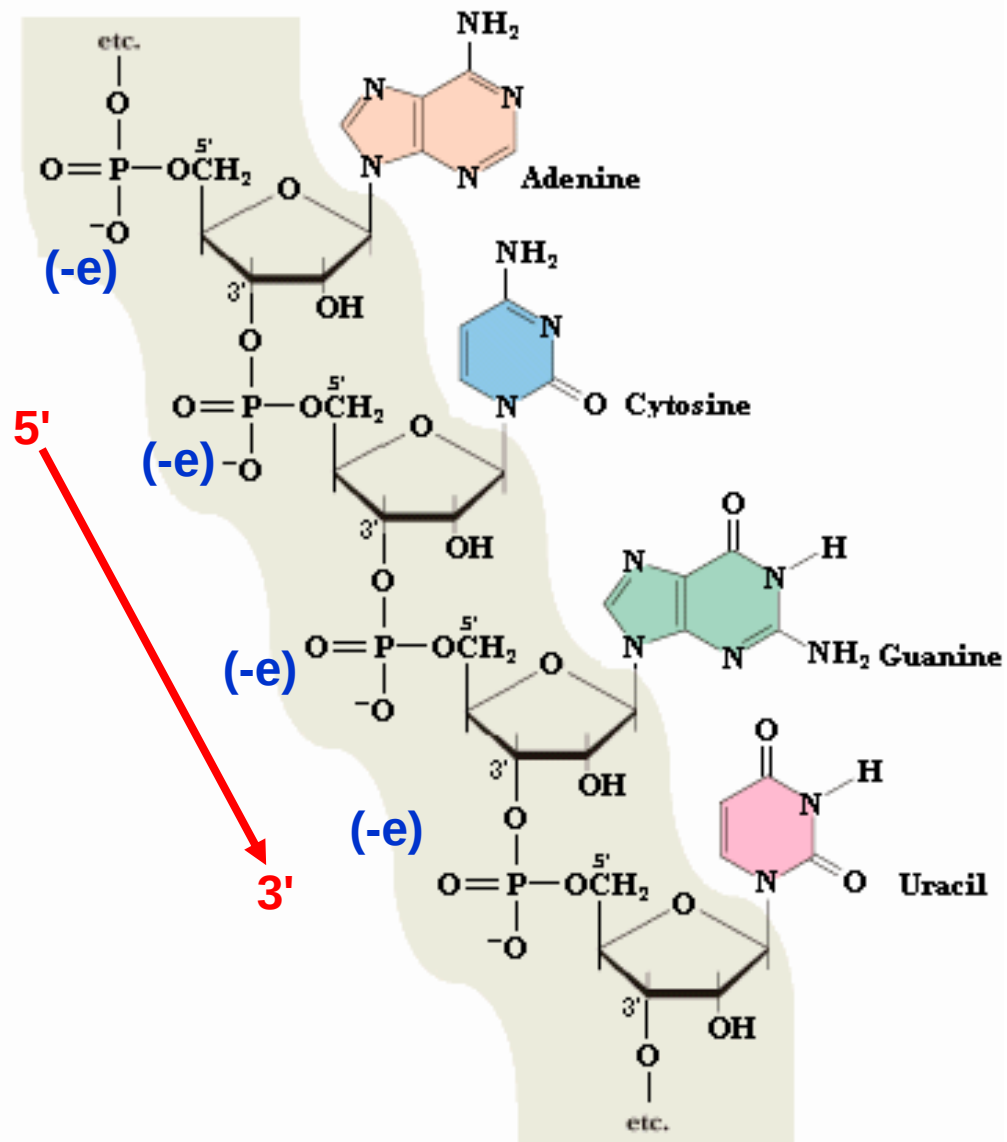
- biologisches Polymer; Monomere = Nukleotide
- typisch ca. 10^3 Nukleotide
- Nukleotide bestehen aus einer Ribose (Zucker), einer Phosphatgruppe und einer von 4 möglichen organischen Basen
- 4 Basen: G, C, A, U
- einsträngig



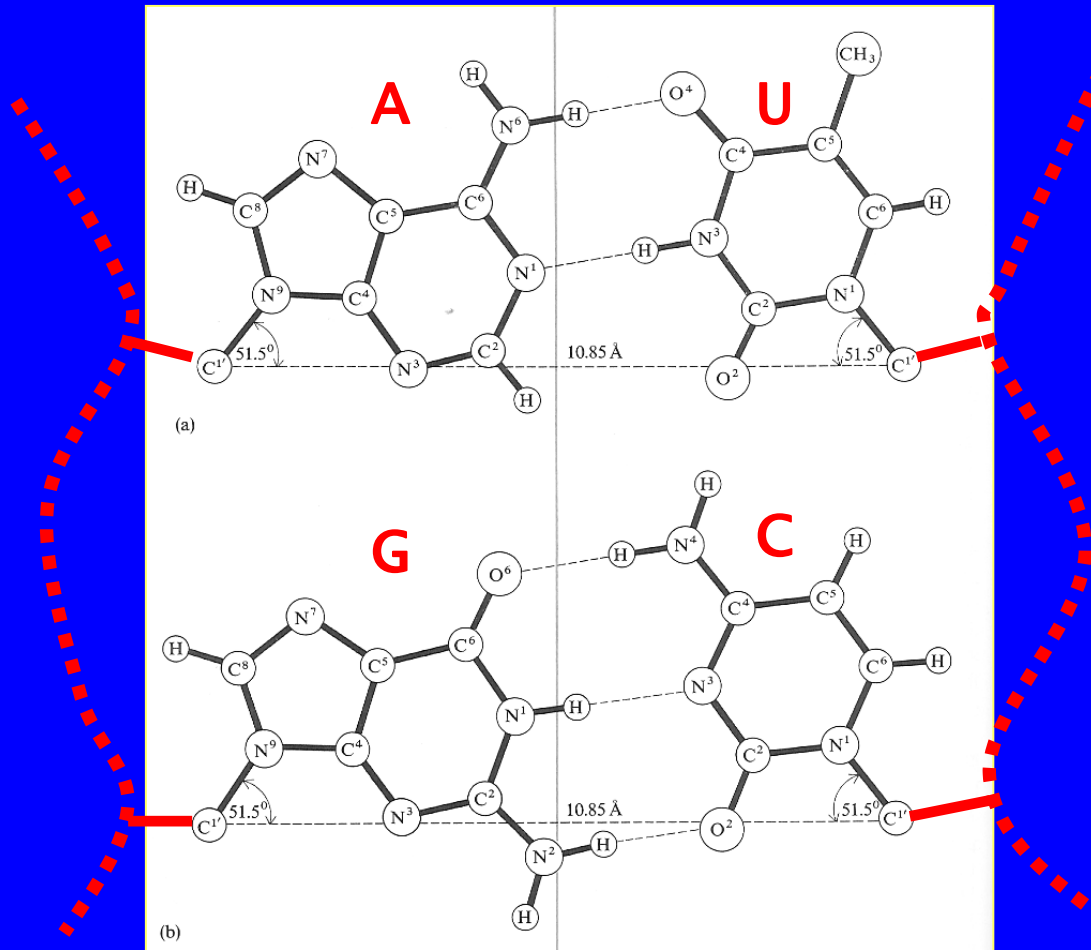
Nukloetid



**Ribonucleic acid
RNA**

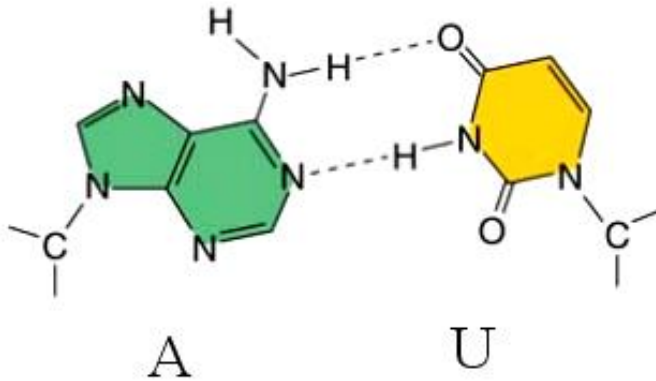


Basenpaarung

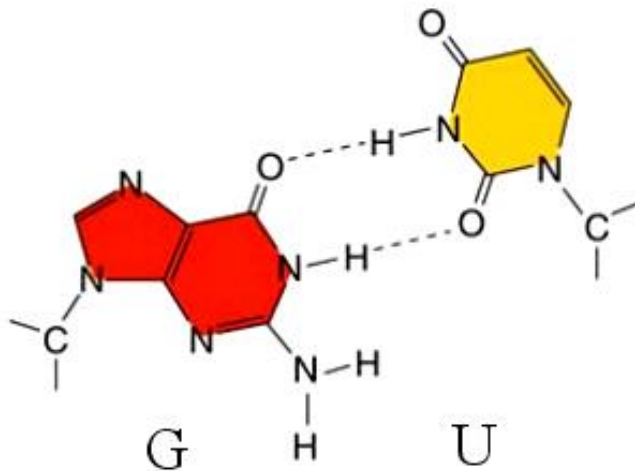


Watson-Crick Paare

Wobble-Basenpaare



Watson-Crick: A-U; C-G



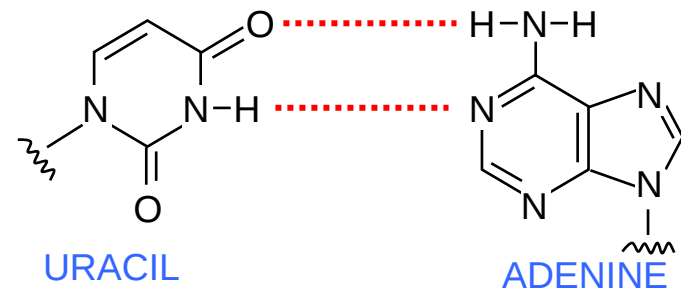
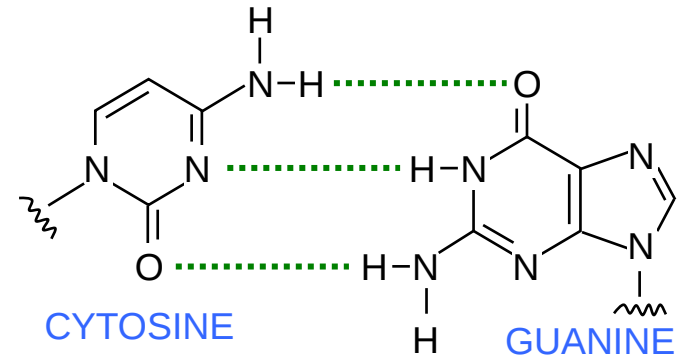
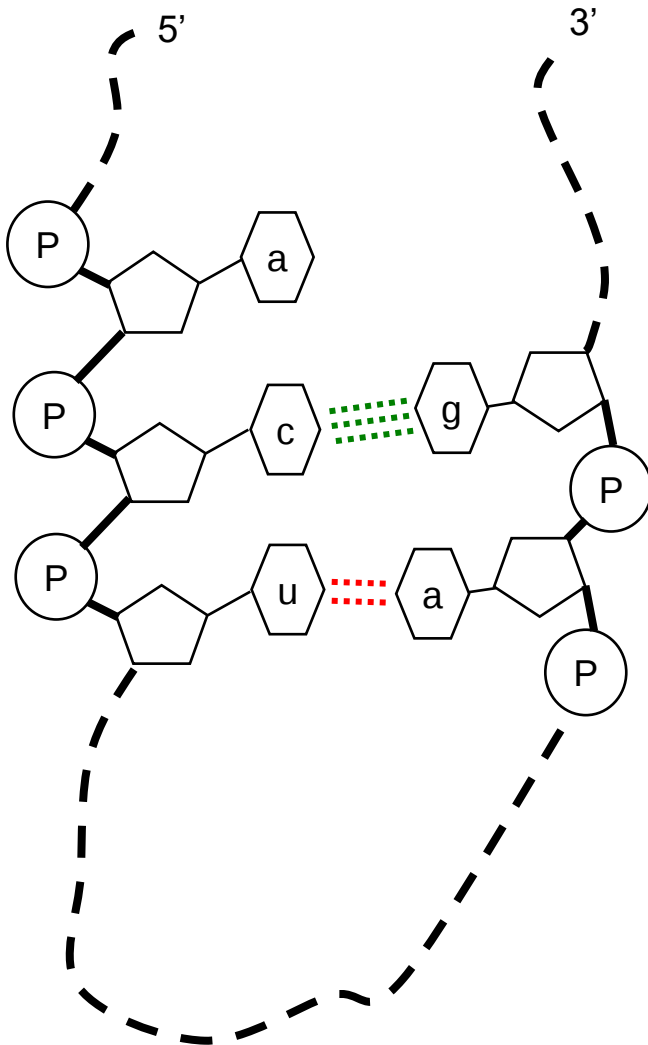
Wobble: G-U

thermodynamische Stabilität ist vergleichbar mit einem Watson-Crick Basenpaar

Faltung zur Sekundärstruktur



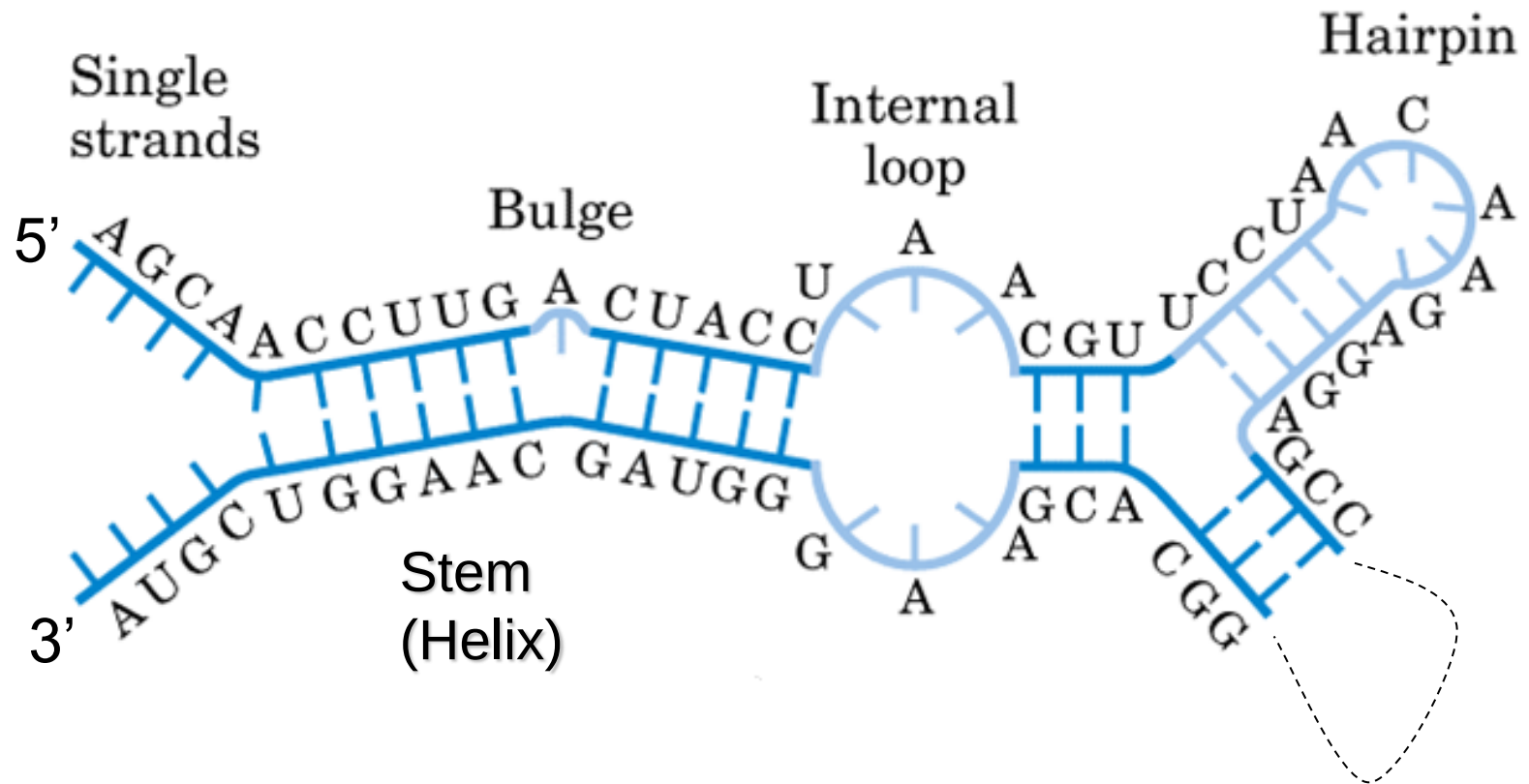
Faltung der RNA



Faltung der RNA

- Die einsträngige RNA kann sich durch Bildung von H₂-Brückenbindungen zwischen den komplementären Basen GC, AU und GU auf sich selbst falten. Dabei entsteht die Sekundärstruktur der RNA.

Sekundärstruktur der RNA



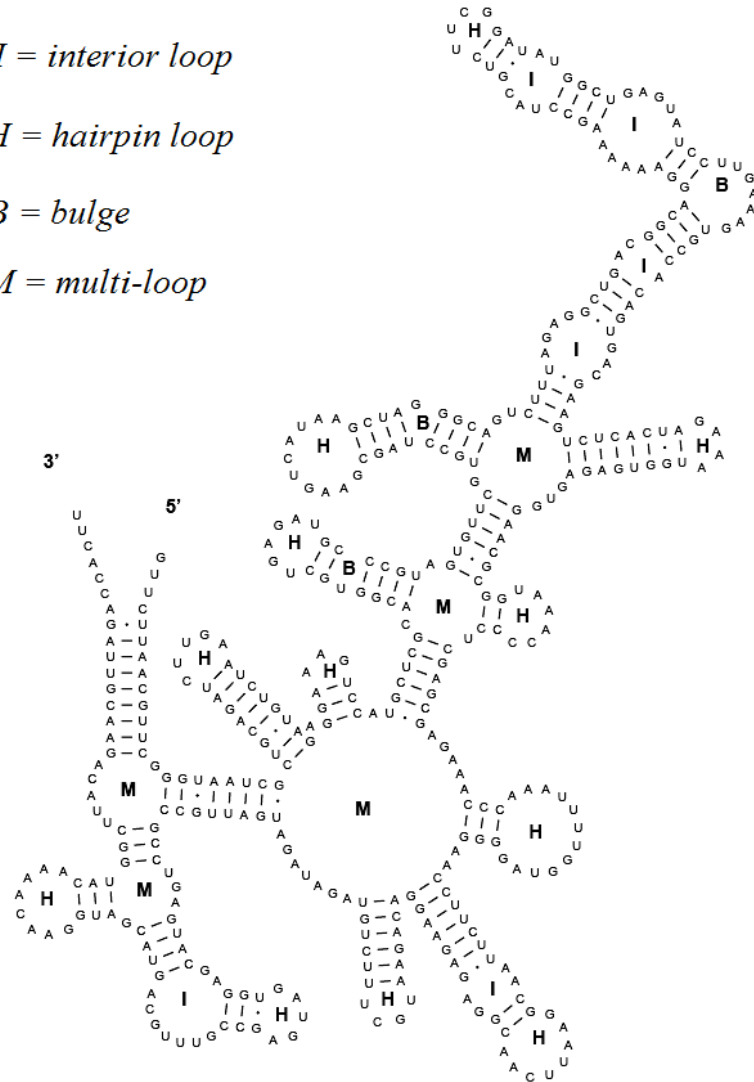
RNA-Sekundärstruktur

I = interior loop

H = hairpin loop

B = bulge

M = multi-loop

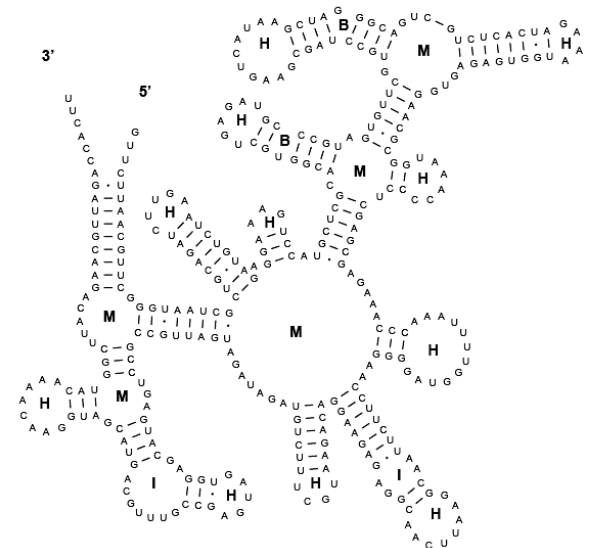


Problemstellung

- Berechne Sekundärstruktur für eine gegebene Primärstruktur (Nukleotidsequenz)

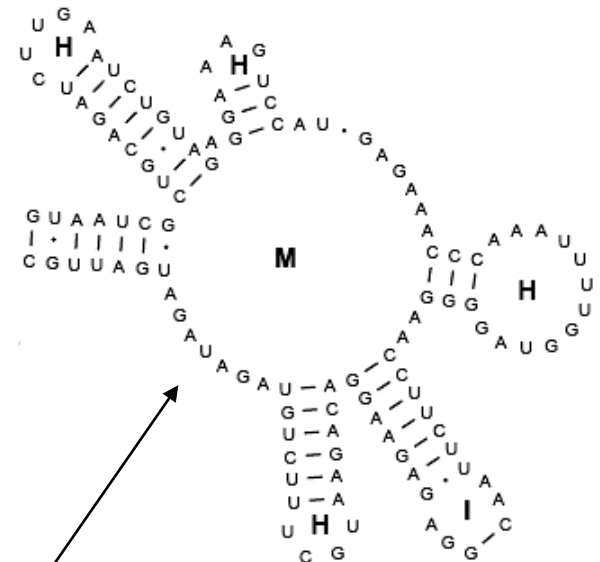
GAUUUAGCUCAGUUGGGAGAGCGCAGACU
GAAGAUCUGGAGGUCCUGUGUUCGAUCCA
CAGAAUUC

?



1. Basenpaar-Maximierung

- Finde *die* Faltung, die eine maximale Anzahl von komplementären Basenpaarungen ergibt.
- Anzahl der Basenpaare für eine bestimmte Konfiguration sei S ("Score")
- 1 Basenpaarung = 1 Punkt



S=32 (nur H₂-Brückenbindungen zählen)



”Brute-Force” nicht möglich

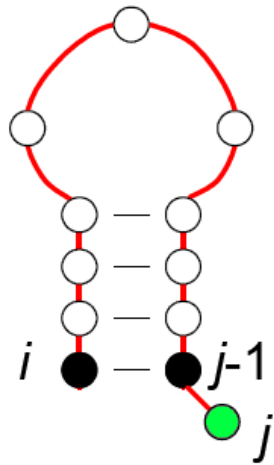
- RNA mit N Nukleotiden kann in $\approx 1.8^N$ mögliche Strukturen gefaltet werden (Zuker, 1984)
- $N=100 \rightarrow 10^{25}$ mögliche Strukturen
- Berechne 10000 Strukturen / Sekunde $\rightarrow 10^{14}$ Jahre

Dynamic Programming (DP)

- Lösung eines Problems wird auf die Lösung "kleinerer" Teilprobleme zurückgeführt ($N=5 \rightarrow N=6 \rightarrow N=7 \rightarrow \dots$)
- Rekursive Lösung des Gesamtproblems
- DP ist in der Bioinformatik weit verbreitet (Needleman-Wunsch, Smith-Waterman, Viterbi)

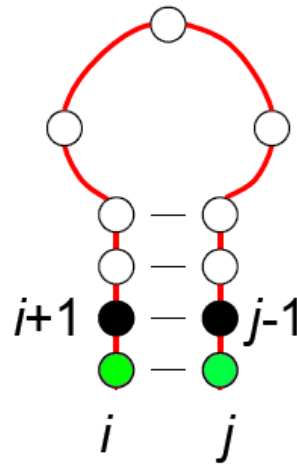


Nussinov-Algorithmus



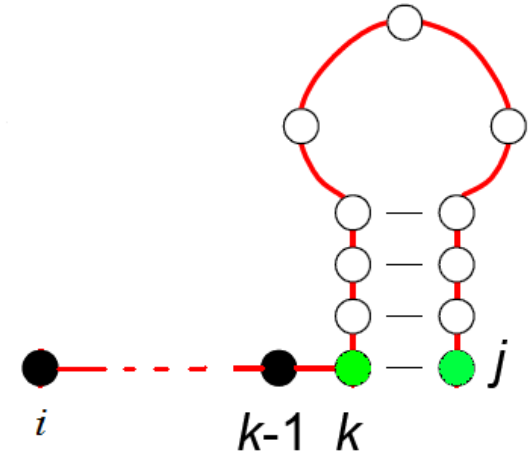
j bindet nicht

$$S(i, j) = S(i, j - 1)$$



j bindet i

$$S(i, j) = S(i + 1, j - 1) + 1$$



Bifurkation: j bindet k

$$S(i, j) = \max_{i < k < j} [S(i, k - 1) + S(k, j)]$$

Maximum bestimmen!!

Nussinov-Algorithmus

$$S(i, j) = \max \begin{cases} S(i, j-1) & [j \text{ bindet nicht}] \\ S(i+1, j-1) + 1 & [j \text{ bindet } i] \\ \max_{i < k < j} [S(i, k-1) + S(k, j)] & [j \text{ bindet } k] \end{cases}$$

$$S(i, j) = \max \begin{cases} S(i+1, j-1) + 1 & [j \text{ bindet } i] \\ \max_{i < k \leq j} [S(i, k-1) + S(k, j)] & [j \text{ bindet } k \text{ oder garnicht}] \end{cases}$$

$$S(i, j) = \max \begin{cases} S(i+1, j-1) + 1 & [\text{if } i, j \text{ base pair}] \\ S(i+1, j) \\ S(i, j-1) \\ \max_{i < k < j} S(i, k) + S(k+1, j) \end{cases}$$

$k = i + 1$

$k = j$

Nussinov-Algorithmus

$$S(i, j) = \max \begin{cases} S(i, j-1) & [j \text{ bindet nicht}] \\ S(i+1, j-1) + 1 & [j \text{ bindet } i] \\ \max_{i < k < j} [S(i, k-1) + S(k, j)] & [j \text{ bindet } k] \end{cases}$$

$$S(i, j) = \max \begin{cases} S(i+1, j-1) + 1 & [j \text{ bindet } i] \\ \max_{i < k \leq j} [S(i, k-1) + S(k, j)] & [j \text{ bindet } k \text{ oder garnicht}] \end{cases}$$

- Die $S(i,j)$ können nach diesem Schema rekursiv berechnet werden
- Die $S(i,j)$ können dazu praktischerweise in einer Matrix angeordnet werden

Matrixdarstellung für den Score

$$\mathbf{j} \longrightarrow$$
[illegible]

Matrixdarstellung für den Score

j $\longrightarrow$

A U A C C C U G U G G U A U

i
 $\downarrow$

A	0	0	0	0										
U		0	0	0	0									
A			0	0	0	0								
C				0	0	0	0							
C					0	0	0	0						
C						0	0	0	0					
U							0	0	0	0				
C								0	0	0	0			
									0	0	0	0		
										0	0	0	0	
G											0	0	0	0
U												0	0	0

Ketten der Länge < 5 können
sich nicht auf sich selbst falten
 $\rightarrow S(1,1)=0, \dots S(1,4)=0, \dots$

Füllen der Matrix

$$S(i, j) = \max \begin{cases} S(i+1, j-1) + 1 & [j \text{ bindet } i] \\ \max_{i < k \leq j} [S(i, k-1) + S(k, j)] & [j \text{ bindet } k \text{ oder garnicht}] \end{cases}$$

$i=3; j=7$

$$S(3, 7) = \max \begin{cases} S(4, 6) + 1 & [\text{Base 7 bindet Base 3}] \\ \max_{3 < k \leq 7} [S(3, k-1) + S(k, 7)] & [7 \text{ bindet } k \text{ oder garnicht}] \end{cases}$$

j →

	A	U	A	C	C	C	U	G	U	G	G	U	A	U
A	0	0	0	0	...									
U		0	0	0	0	...								
A			0	0	0	0	1							
C				0	0	0	0							
C					0	0	0	0						
C						0	0	0	0					
U							0	0	0	0				
G								0	0	0	0			

i
↓

Füllen der Matrix

$$S(3,7) = \max \begin{cases} S(4,6) + 1 & [j \text{ bindet } i] \\ \max_{3 < k \leq 7} [S(3, k-1) + S(k, 7)] & [j \text{ bindet } k \text{ oder garnicht}] \end{cases}$$

$i=3; j=7$

k	$S_1 + S_2$
4	$S(3,3) + S(4,7)$
5	$S(3,4) + S(5,7)$
6	$S(3,5) + S(6,7)$
7	$S(3,6) + S(7,7)$

i

$j \longrightarrow$

	A	U	A	C	C	C	U	G	U	G	G	U
A	0	0	0	0	0							
U		0	0	0	0	0						
A			0	0	0	0	1					
C				0	0	0	0	1				
C					0	0	0	0	0			
C						0	0	0	0	1		
U							0	0	0	0	1	
G								0	0	0	0	1

Matrixelemente, die zur Berechnung von $S(3,7)$ beitragen

$$S(3,7) = \max \begin{cases} S(4,6) + 1 & [3 \text{ bindet } 7] \\ \max_{3 < k \leq 7} [S(3, k-1) + S(k, 7)] & [j \text{ bindet } k \text{ oder garnicht}] \end{cases}$$

$i=3; j=7$

k	$S_1 + S_2$
4	$S(3,3) + S(4,7)$
5	$S(3,4) + S(5,7)$
6	$S(3,5) + S(6,7)$
7	$S(3,6) + S(7,7)$

i



	A	U	A	C	C	C	U	G	U	G	G	U
A	0	0	0	0	0							
U		0	0	0	0	0						
A			0	0	0	0	1					
C				0	0	0	0	1				
C					0	0	0	0	0			
C						0	0	0	0	1		
U							0	0	0	0	1	
G								0	0	0	0	1

A U A C C C U G U G G U A U

[illegible]

Traceback

	A	U	A	C	C	C	U	G	U	G	G	U	A	U
A	0	0	0	0	0	0	1	1	2	3	3	4	4	5
U		0	0	0	0	0	0	1	2	3	3	3	4	4
A			0	0	0	0	1	1	2	2	2	3	3	3
C				0	0	0	0	1	1	1	2	2	2	2
C					0	0	0	0	0	1	2	2	2	2
C						0	0	0	0	1	1	1	2	2
U							0	0	0	0	1	1	2	2
G								0	0	0	0	1	1	2
U									0	0	0	0	1	1
										0	0	0	0	1
											0	0	0	0
												0	0	0
													0	0
U														0

Erstes A mit letztem U

Zweites U mit vorletztem A

Drittes A mit vor-vorletztem U

....

Voraussage

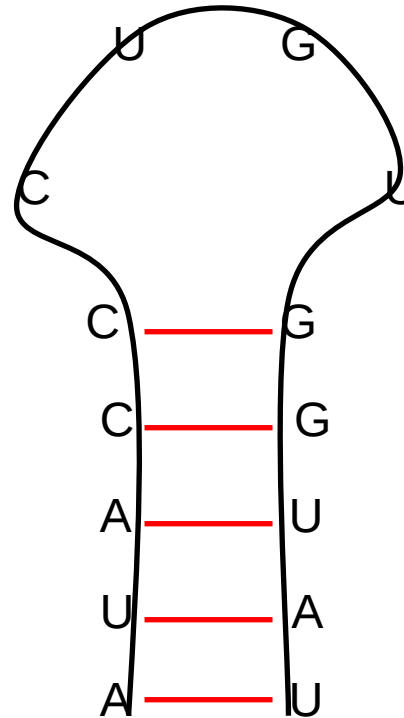
Erstes A mit letztem U

Zweites U mit vorletztem A

Drittes A mit vor-vorletztem U

....

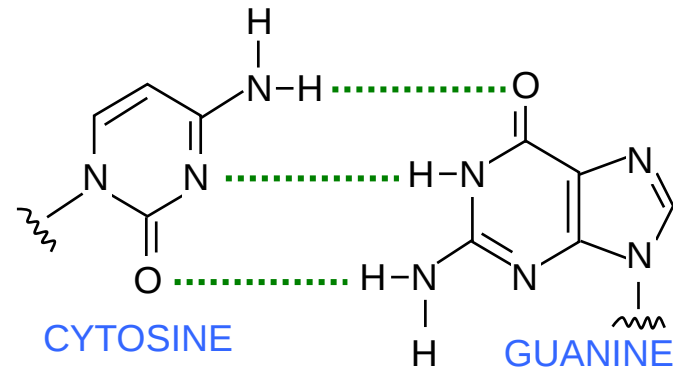
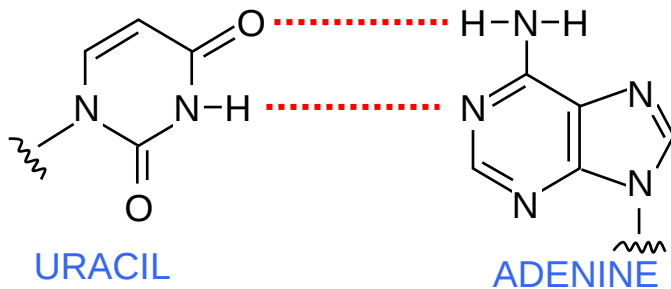
AUACCCUGUGGUAU



Maximaler Score: 5

Nachteile Basen-Maximierung

- Stärke der Bindung zwischen den Basenpaaren wird nicht beachtet



2. Energie-Minimierung

- Jede Basenpaarung hat eine negative Bindungsenergie
- Die Konfiguration mit der niedrigsten Gesamtenergie wird gesucht
- diese Konfiguration ist am stabilsten im thermodynamischen Gleichgewicht (Boltzmann-Gleichung)

Tinoco-Uhlenbeck Postulat

- **Annahme:** Bindungsenergie eines jeden Basenpaares ist unabhängig von allen anderen Basenpaaren (und von Loop-Strukturen)
- **Folgerung:** totale Bindungsenergie der RNA ist einfach die Summe der Bindungsenergien der einzelnen Basenpaare

Hydrogen-Bond Modelle

Major

Bindung	Bindungsenergie
GC	-3 kcal/mol
AU	-2 kcal/mol
GU	-1 kcal/mol

approximiert die relative
Stärke der Bindungen

Mathews

Bindung	Bindungsenergie
GC	-3 kcal/mol
AU	-2 kcal/mol
GU	-2 kcal/mol

Anzahl der
Wasserstoff-Brücken

$$S(i, j) = \min \begin{cases} S(i+1, j-1) + \Delta G & [j \text{ bindet } i] \\ \min_{i < k \leq j} [S(i, k-1) + S(k, j)] & [j \text{ bindet } k \text{ oder garnicht}] \end{cases}$$

Matrixdarstellung für Energie-Minimierung

[illegible]

Voraussage

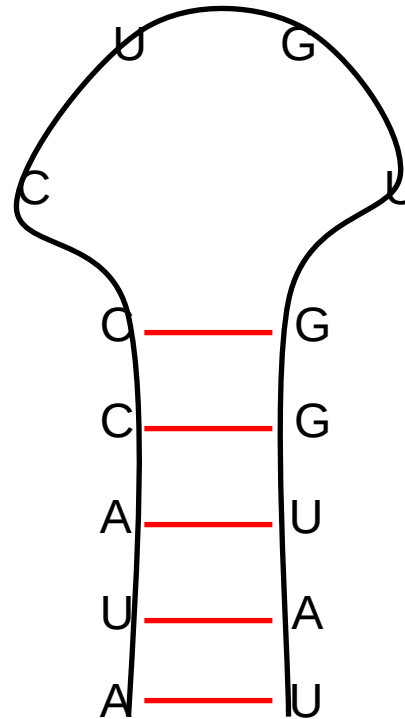
Erstes A mit letztem U

Zweites U mit vorletztem A

Drittes A mit vor-vorletztem U

....

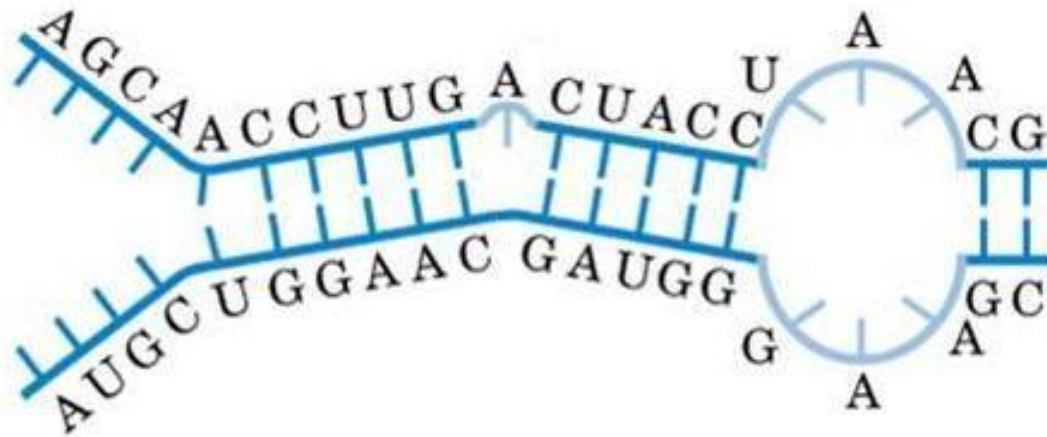
AUACCCUGUGGUAU



Totale Freie Energie: -12 kcal/mol

Nachteile der Major- und Mathews Modelle

- jedes Basenpaar wird nur isoliert für sich betrachtet → Widerspruch zu experiment. Daten



3. Individual Nearest Neighbor Model

- Einfluss benachbarter Basenpaare auf die Bindungsenergie wird berücksichtigt = “Stacking”
- Nach diesem Modell berechnete Strukturen stimmen besser mit experimentellen Daten überein

Energie-Zuordnung

- stabilisierenden Regionen (Helices) wird eine negative Energie zugeordnet
 - wird Übergängen *zwischen* Basenpaaren zugeordnet
 - Energie hängt (nur) vom benachbarten Basenpaar ab
 - Stacking-Energien werden experimentell an kleinen synthetischen RNAs ermittelt.
- destabilisierenden Regionen (Bulges, Loops, usw.) wird eine positive Energie zugeordnet
 - ebenfalls experimentell ermittelt

Stacking - Energien

Werte für freie Energien für helikale Strukturen (in kcal/mol bei 37° C):

	A/U	C/G	G/C	U/A	G/U	U/G
A/U	-0.9	-1.8	-2.3	-1.1	-1.1	-0.8
C/G	-1.7	-2.9	-3.4	-2.3	-2.1	-1.4
G/C	-2.1	-2.0	-2.9	-1.8	-1.9	-1.2
U/A	-0.9	-1.7	-2.1	-0.9	-1.0	-0.5
G/U	-0.5	-1.2	-1.4	-0.8	-0.4	-0.2
U/G	-1.0	-1.9	-2.1	-1.1	-1.5	-0.4

Tabelle: Daniela Nitsch; Universität Ulm

Destabilisierende Energien von Schleifen

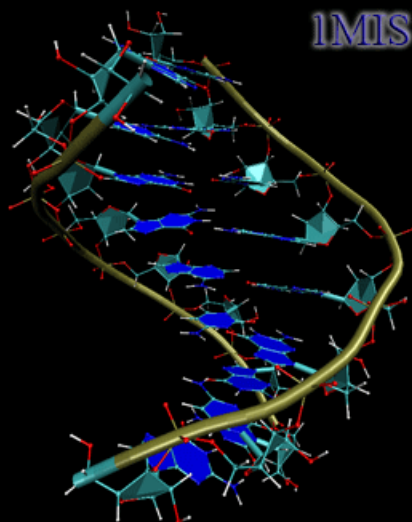
Werte von freien Energien (in kcal/mol bei 37° C) für Schleifen, abhängig von deren Größe :

Größe	innere Schleife	Ausbuchtung	Haarnadel-Schleife
1	.	3.9	.
2	4.1	3.1	.
3	5.1	3.5	.
4	4.9	4.2	4.9
5	5.3	4.8	4.4
10	6.3	5.5	5.3
15	6.7	6.0	5.8
20	7.0	6.3	6.1
25	7.2	6.5	6.3
30	7.4	6.7	6.5

Turner Lab

UNIVERSITY OF ROCHESTER
Department of Chemistry

TURNER LAB



[Read our Group Objective](#)

[View an NMR Structure from the Turner Lab](#) (Updated 06/15/2008)

[Download RNAstructure](#) (A Windows version of the Zuker algorithm for predicting RNA secondary structure)

[Download Dynalign for Unix](#)

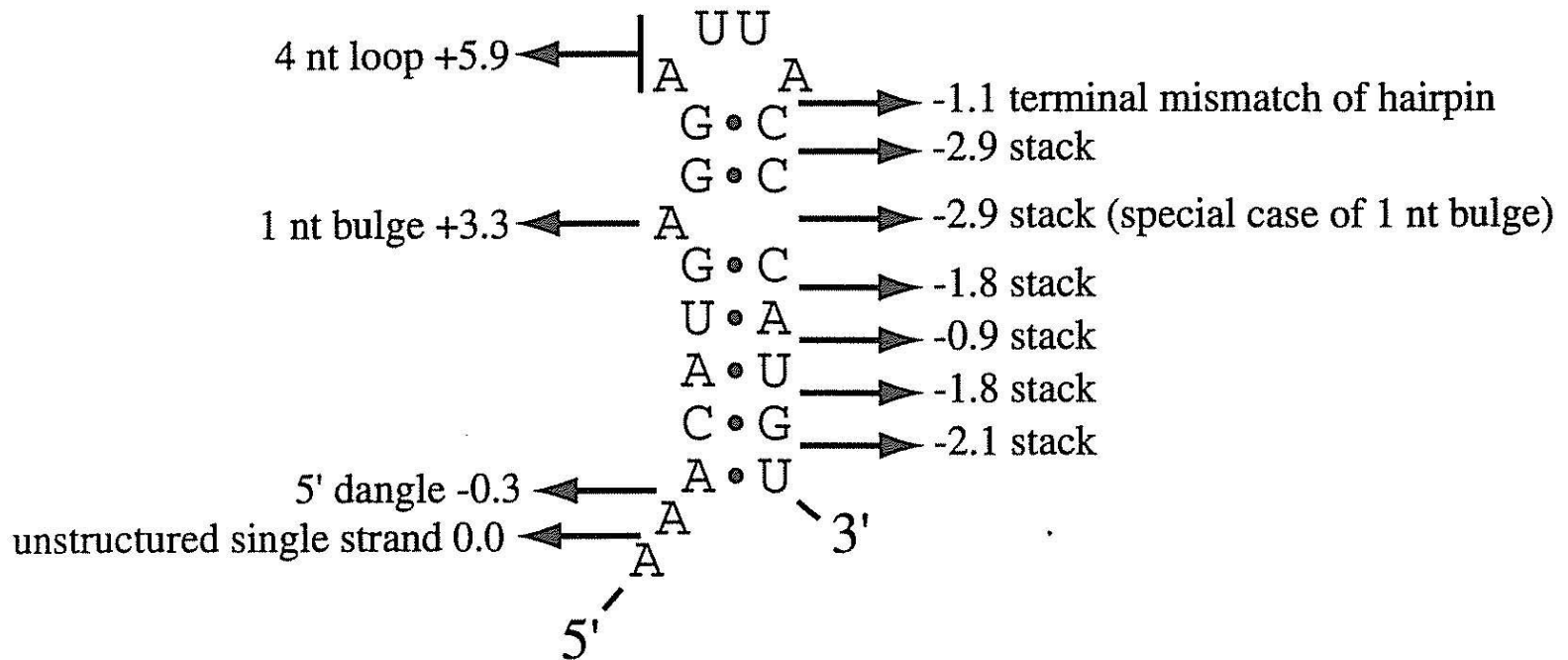
[Download the Current Free Energy Nearest Neighbor Parameters](#)

[A List of Current Group Members](#) (Updated 06/15/2008)

[A List of Former Group Members](#) (Updated 06/15/2008)

[Relevant Publications](#) (Updated 06/15/2008)

Energieberechnung



overall $\Delta G = -4.6$ kcal/mol

Hairpin loop



+6

-1.8

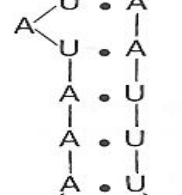
-1.2

-2.2

-5.0

-2.2

Interior loop



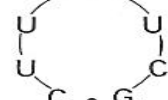
-1.2 + 3

-1.8

-1.2

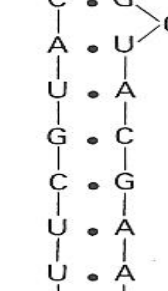
-1.2

Bulge loop



+2

Interior loop



-2.2 + 3

-1.8

-2.2

-3.2

-2.2

-1.2

-2.2

Bulge loop



+2

-5

0



$T=37^{\circ}\text{C}$,
in kcal/mol

$$\Delta G = - 21.8 \text{ kcal/mol}$$

Programmierung des Stacking-Modells

- dynamische Programmierung, wie oben
- aufgrund der Nachbar-Wirkung der Stacking-Parameter müssen *zwei* Matrizen gehalten werden
- Zuker-Algorithmus (Zuker & Stiegler, 1981) →
Fold

Nachteile des INN - Modells

- nur die Struktur mit der absolut niedrigsten Energie wird berechnet
- Oft gibt es mehrere Sekundärstrukturen mit nahezu gleicher Energie
- RNA-switches (*Flamm 2001*)
- Algorithmus muss so abgewandelt werden, dass auch Faltungen mit etwas höherer Gesamtenergie gefunden werden → Suboptimale Strukturen → Mfold (Zuker, 1989)

Zuker's Mfold (1989)

- Findet sub-optimale Lösungen für Freie Energie

$$W_{i,j} = \min \left\{ \begin{array}{l} W_{i+1,j}, \\ W_{i,j-1}, \\ V_{i,j}, \\ \min_{i \leq k < j} \{W_{i,k} + W_{k+1,j}\} \end{array} \right\}$$

$$V_{i,j} = \min \left\{ \begin{array}{l} H_{i,j}, \\ S_{i,j}, \\ VBI_{i,j}, \\ VM_{i,j} \end{array} \right\}$$

$$VBI_{i,j} = \min_{i < i' < j' < j} \{I_{i,j,i',j'} + V_{i',j'}\}$$

$$VM_{i,j} = \min_{i+1 < k < j-1} \{W_{i+1,k} + W_{k+1,j-1}\}$$

Programme

- **Fold, Mfold**

- Zuker & Stiegler (1981) *Nuc. Acids Res.* 9: 133-48
- Zuker (1989) *Science* 244:48-52
- <http://mfold.bioinfo.rpi.edu/cgi-bin/rna-form1.cgi>

- **RNAfold**

- Vienna RNA secondary structure server
- Hofacker (2003) *Nuc. Acids Res.* 31: 3429-31
- <http://rna.tbi.univie.ac.at/>

- **RNAstructure**

- Mathews Lab, Windows, free
- <http://rna.urmc.rochester.edu/rnastructure.html>

Mfold - Webserver

- Enter the sequence to be folded in the box.

All non-alphabet characters will be removed.

FASTA format may be used.

```
catctatcagcgacagagctcactataaatggcgctctctctaatagaata
taattatcggagcagtgccctccatattatgacagttatactgtcggg
tttttttaaataaaagcagcatctgtctaataaaacccaacagatactgga
agttttgcatttatggcacaacttaagggttttagaaaaacagcgcgcag
ccaaatgtaatgaaataaagttgaagctaaagatttagagatgaattaaat
ttaattagggttgctaaagaagcgagcactgaccagataagaatgctggg
tttcctaaatgcagtgaaattgtgaccaagttataaatcaatgtcacttaa
aggctgtggtagtactcctgcaaaattttatagtcagtttatccaagggt
gtaaactctaattcccatttgcaaaatttccagtaacctttgtcacaaacct
aacacattatcgggagcagtgctctccatattgtataaagaacaaggtag
tttttacctaccacagtgctgtatcgggagacagtgatctccatattgta
cactaagggtgtaatgaattatcgggaacagtggttcccatattt
```

- Enter constraint information in the box at the right. (optional) You may:

- force bases $i, i+1, \dots, i+k-1$ to be double stranded by entering:
 $F \ i \ 0 \ k$ on 1 line in the constraint box.
- force consecutive base pairs $i, j, i+1, j-1, \dots, i+k-1, j-k+1$ by entering:
 $F \ i \ j \ k$ on 1 line in the constraint box.
- force bases $i, i+1, \dots, i+k-1$ to be single stranded by entering:
 $P \ i \ 0 \ k$ on 1 line in the constraint box.
- prohibit the consecutive base pairs
 $i, j, i+1, j-1, \dots, i+k-1, j-k+1$ by entering:
 $P \ i \ j \ k$ on 1 line in the constraint box.
- prohibit bases i to j from pairing with bases k to l by entering:
 $P \ i-j \ k-l$ on 1 line in the constraint box.

- The RNA sequence is .

- Folding temperature is fixed at 37°.

- Ionic conditions: 1M NaCl, no divalent ions.

- Enter the percent suboptimality number.

- Enter an upper bound on the number of computed foldings.

- Enter the window parameter if you wish.

- Enter the maximum interior/bulge loop size

Vienna RNA package

Vienna RNA WebServers

[\[Home|Help\]](#)

RNA stuff @ tbi.univie.ac.at

This server provides programs, web services, and databases, related to our work on RNA secondary structures. For general information and other offerings from our group see the [main TBI web server](#).

The Vienna RNA Servers:

- **RNAfold server** predicts minimum free energy structures and base pair probabilities from single RNA or DNA sequences.
- **RNAalifold server** predicts *consensus* secondary structures from an alignment of several related RNA or DNA sequences. You need to upload an alignment.
- **RNAinverse server** allows you to design RNA sequences for any desired target secondary structure.
- **RNAcofold server** allows you to predict the secondary structure of a dimer.
- **RNAup server** allows you to predict the accessibility of a target region.
- **LocARNA server** generates structural alignments from a set of sequences. **In collaboration with the Bioinformatics Group Freiburg.**
- **barriers server** allows you to get insights into RNA folding kinetics.
- **RNAz server** will assist you in detecting thermodynamically stable and evolutionarily conserved RNA secondary structures in multiple sequence alignments.
- **Structure conservation analysis server** will assist you in detecting evolutionarily conserved RNA secondary structures in multiple sequence alignments.
- **RNAstrand server** allows you to predict the reading direction of evolutionarily conserved RNA secondary structures.
- **RNAxs server** assists you in siRNA design.

Downloads

Get the Source code for:

- the [Vienna RNA Package](#), our basic RNA secondary structure analysis software.
- The [ALIDOT](#) package for finding conserved structure motifs (add-on)
- The [barriers](#) program for analysis of RNA folding landscapes.

Databases

- [Atlas of conserved Viral RNA Structures](#) found by ALIDOT

University of Rochester Medical Center

Mathews Lab

RNAstructure, Version 4.6

updated 5/6/2008

RNAstructure is a Windows program for the prediction and analysis of RNA secondary structure. It works with Windows ME, Windows NT 4, Windows 2000, Windows XP, and Windows Vista. (Note that with Windows Vista, you will need to install the older Windows help reader by following the instructions provided by Microsoft to use the online help. We are working to change this.) Version 4.6 includes a secondary structure prediction algorithm, a sequence editor, an integrated drawing tool, the OligoWalk program, OligoScreen, Dynalign, and a partition function calculator. RNAstructure uses the most current thermodynamic parameters from the [Turner lab](#).

New for version 4.6 are stochastic sampling of secondary structures and the ability to constrain the prediction of the lowest free energy structure with SHAPE data.

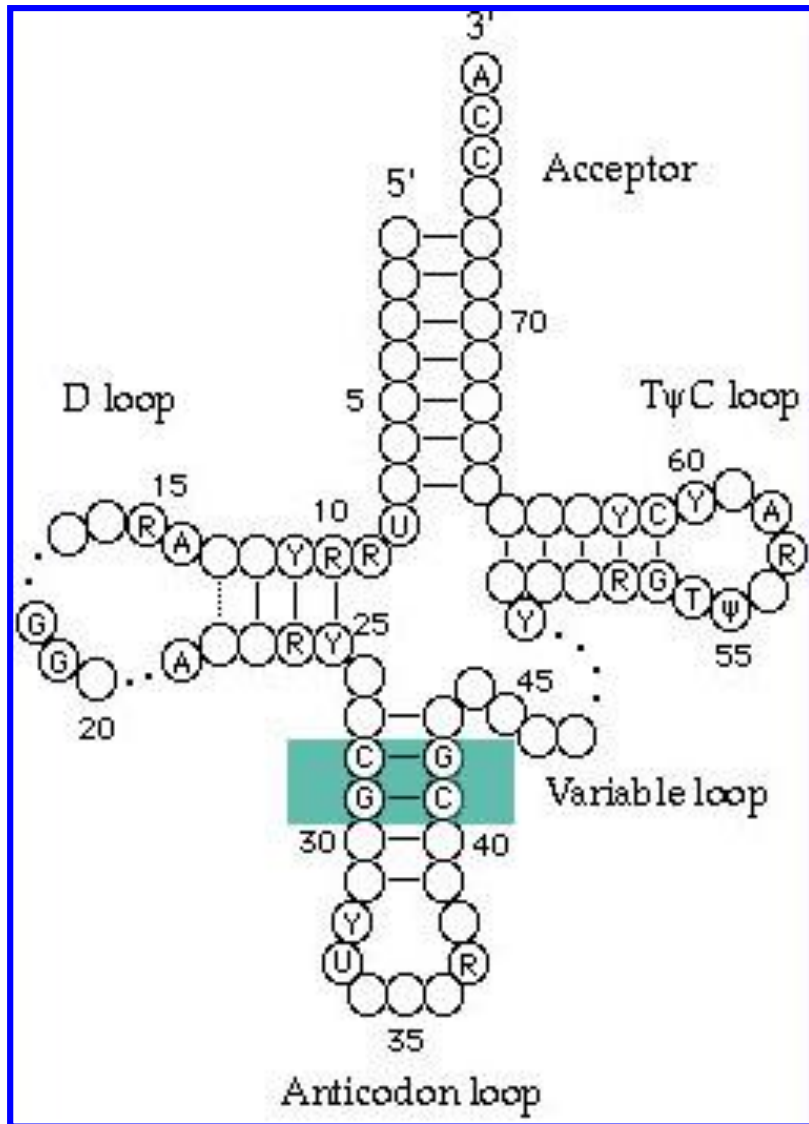
We ask you to register before downloading so that we may occasionally notify you of significant updates. RNAstructure is free of charge.

[Register to Download](#)

4. Vergleichende Sequenzanalyse

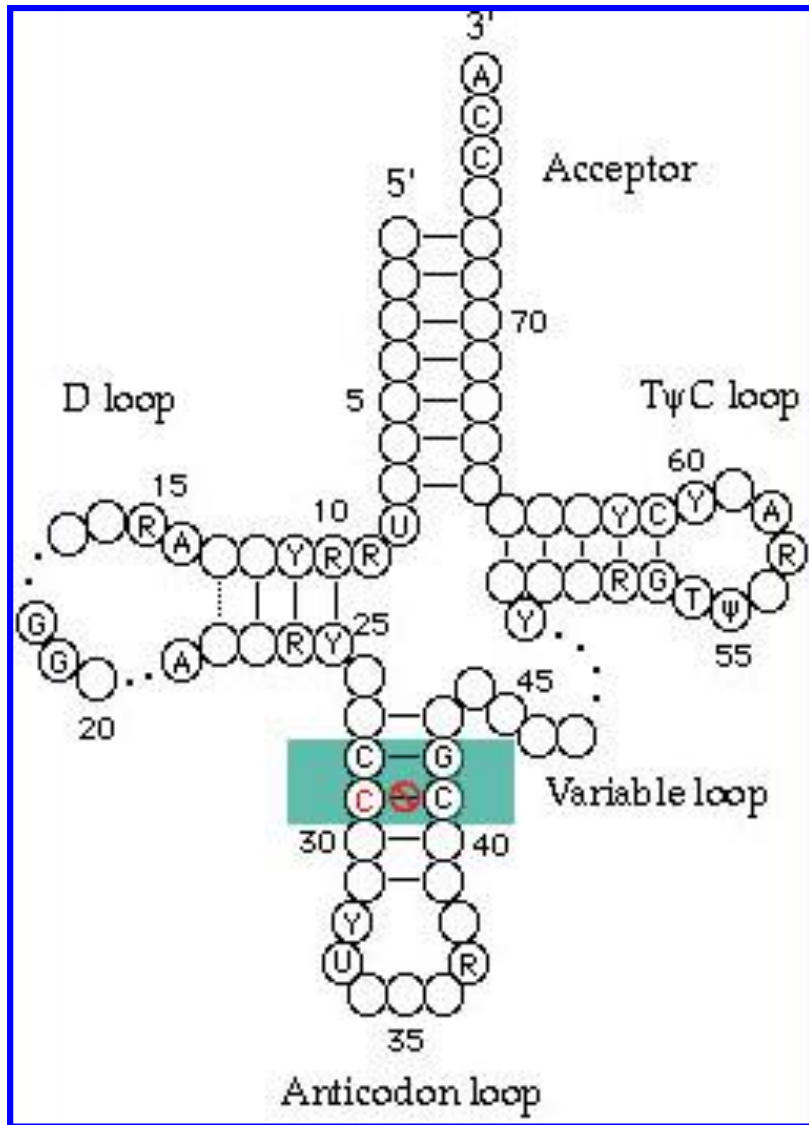
- "Gold standard"
- Konservierte Domänen sind Kandidaten für Helices (Stems)
- z. B. Dynalign

Kovariation



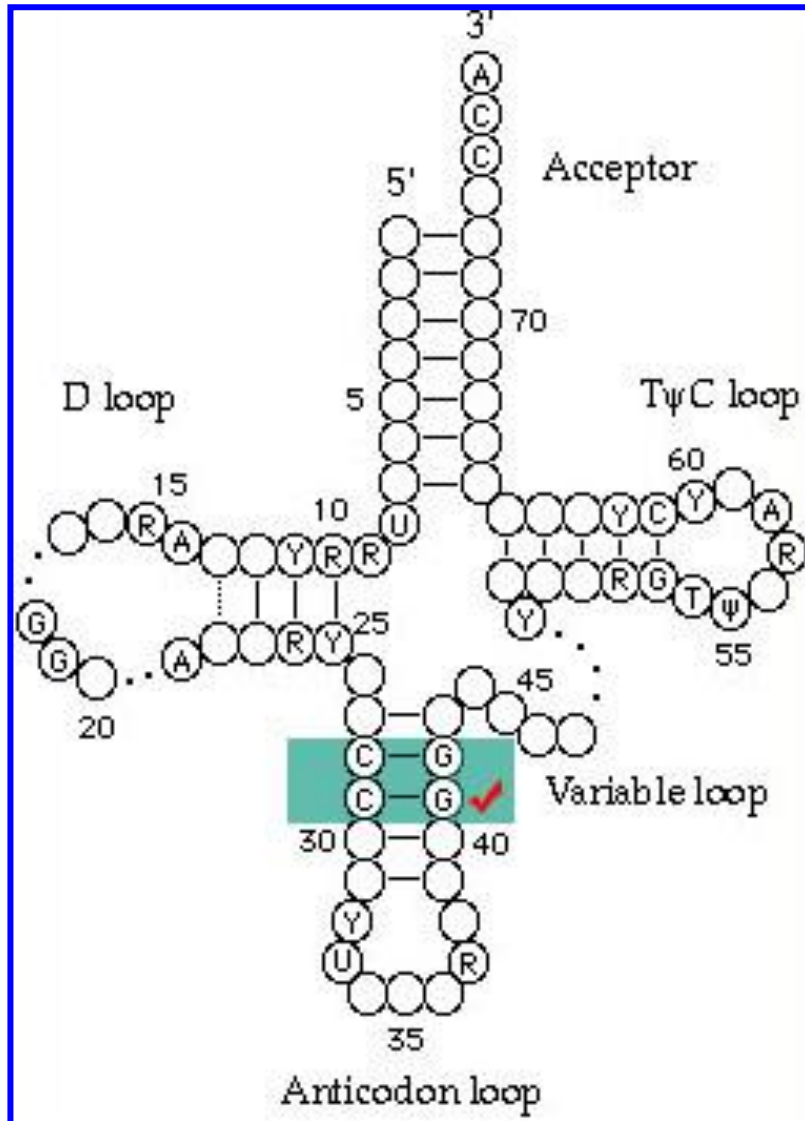
Ein CG/GC – Paar in der Kleeblatt-Struktur ist an der Bildung eines Stems beteiligt

Nicht-kompensierende Mutation



Mutation in nur einer Base verkürzt oder zerstört den Stem

Kompensierende Mutation



Kovariation stellt sicher, dass das Basepairing erhalten bleibt und somit auch die Struktur (Stem)

Kovariation

Escherichia coli
Hildenbrandia rubra
Bongia fuscopurpurea
Rhodochaete parvula
Cordyceps kanzashiana
Stichococcus bacillaris
Graphiola phoenicis

CACACUGGAA (CUGAGACACG) GUCCAGACUCC
GAGAGGGAGC (CUGAGAAACG) GCUACCACAUC
GAGAGGGAGC (CUGAGAAAUG) GCUACCACAUC
GAGAGGGAGC (CUGAGAAACG) GCUACCACAUC
GAGAGGGAGC (CUGAGAGACG) GCUACUACAUC
GAGAGGGAGC (CUGAGAAACG) GCUACCACAUC
GAGAGGGAGC (CUGAGAAACG) GCUACCACAUC



H. rubra



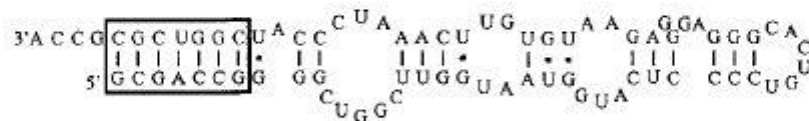
E. coli

Dynalign

- kombiniert Energieminimierung *und* komparative Sequenzanalyse (2 Sequenzen)
- Helices nur erlaubt, wenn diese in beiden Sequenzen möglich sind → beschränkt Suche auf gemeinsame Strukturen
- in RNAStructure

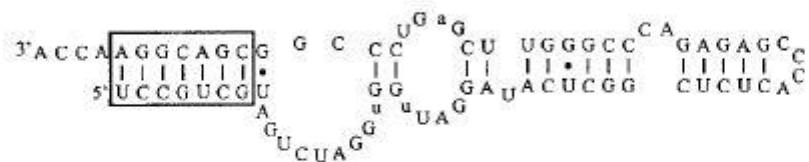
Standard Free Energy Minimization

A



RD0260

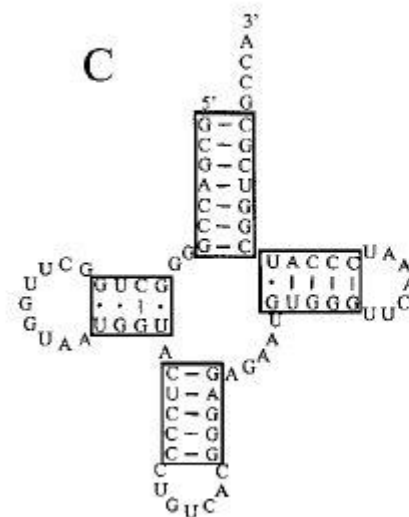
B



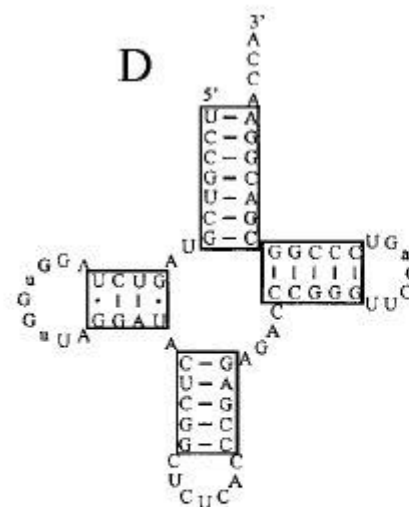
RE6781

Dynalign

C



D



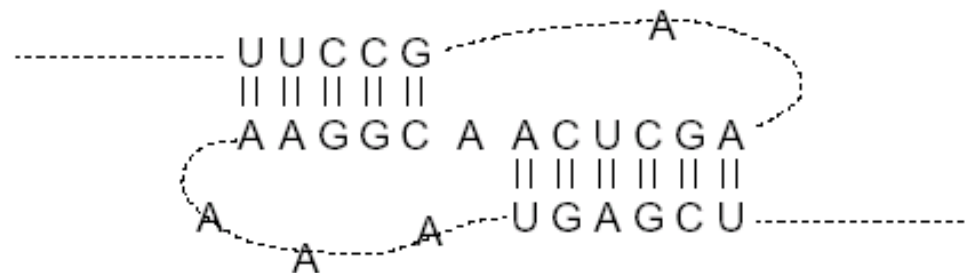
RD0260 GCGACCGGGGCGUGGCUUGGUAAUGGUACUCCCCUGUCACGGGAGAGAAUGUGGGUUCAAAUCCCAUCGGUCGCGCCA
 RE6781 UCCGUCGUAGUCUAGGuGGGuUAGGAUACUCGGCUCUCACCCGAGAGAC-CCGGGUUCGaGUCCCGGCGACGGAACCA
 ^^^^^^ ^^^^^ ^^^^^ ^^^^^ ^^^^^ ^^^^^ ^^^^^ ^^^^^

Performance der Programme

- Ungefähr 70% der Baasenpaare werden korrekten Strukturen zugeordnet

Weitere Limitierungen

- Keine Pseudo-Knoten (1.4% aller Basenpaare lt. *Mathews 2006*) und andere tertiäre Wechselwirkungen
- Unsicherheit der thermodynamischen Parameter (Fehler < 0.5 kcal/mol, *Mathews 2006*)
- WW über den nächsten Nachbarn hinaus vernachlässigt (Tetraloops- ganze Sequenz zählt)
- TD Gleichgewicht wird vorausgesetzt
- ...



Lookup tables

Table 2. Lookup table for unstable triloops and stable tetraloops and hexaloops

Hairpin	Ref(s).	$\Delta G^\circ_{37 \text{ loop}}$, kcal/mol
CAACG	87	6.8
GUUAC	87	6.9
CAACGG	48	5.5
CCAAGG	48	3.3
CCACGG	48	3.7
CCCAGG	48	3.4
CCGAGG	48	3.5
CCGCGG	48	3.6
CCUAGG	48	3.7
CCUCGG	48	2.5
CUAAGG	48	3.6
CUACGG	47, 50	2.8
CUCAGG	48	3.7
CUCCGG	47	2.7
CUGCGG	47	2.8
CUUAGG	48	3.5
CUUCGG	47, 50	3.7
ACAGUACU	86	2.8
ACAGUGCU	86	2.9
ACAGUGAU	86	3.6
ACAGUUCU	86	1.8

Kinetik des Faltens

- Bisher wurde nur der Endzustand berechnet (Gleichgewicht)
- Mögliche Simulierung der Kinetik:
 - Schrittweise (irreversible) Addition von Stems (*Nussinov84, Abrahams90*)
 - Genetische Algorithmen (*vanBatenburg95*)

Genetische Algorithmen

- Ahmt natürliche Auslese nach
- Mehrere Schritte
- Mehrere mögliche Lösungen in jedem Schritt (Zufall)
- Die "besten" Lösungen "überleben"

Ein Schritt des genetischen Algorithmus

- Mutation
- Crossover (Kreuzung)
- Auslese (breeding)

Mutationen

- Zufällige Veränderung der bisherigen Faltungen
- Helices entstehen oder verschwinden
- Mutationen können gut oder schlecht sein (wird in Ausleseschritt entschieden)

Crossover

- Faltungen haben gute und weniger gute Abschnitte
- Beispiel: 2 Faltungen mit jeweils einer guten und einer "schlechten" Helix
- Kreuzung dieser beiden: 1 Faltung mit 2 guten & 1 Faltung mit 2 schlechten Helices
- Letztere wird im Ausleseschritt verworfen

Auslese

- Aufheben der guten und Verwerfen der weniger guten Faltungen
- Jeder Faltung wird ein Fitniss-Wert zugeordnet
- Fitness-Wert ist subjektiv (lange Helices können z. B. bevorzugt werden)

Implementation eines genetischen Algorithmus zur RNA-Faltung

- Van Batenburg, J. theor. Biol. 1995
 1. Berechnen aller möglichen Helices (wobei viele nicht miteinander kompatibel sind)
 2. "Stem-Array" – viele zufällige Kombinationen von Helices werden erzeugt

Stem-Array

Possible stems	Solutions			
	1	2	3	4
213–224/234–245	1	0	0	1
253–263/568–578	0	0	0	1
46–56/210–220	0	0	1	0
210–220/244–254	0	1	0	0
292–302/519–529	0	0	0	1
528–537/543–552	0	1	1	0
.....	...			
222–231/404–413	0	1	1	1
420–429/531–540	1	0	0	1
291–300/369–378	0	1	1	0
207–216/267–276	0	0	0	1

On the left is the collection of all stems. Each stem is specified with four numbers that indicate the paired regions. For example 213–224/234–245 means that nucleotids 213–224 base-pair with nucleotides 234–245. On the right are four columns of the GA population of four individuals. Each 1 in a column represents the inclusion of the corresponding stem and a 0 means that the corresponding stem is not included.

*Van
Batenburg, J.
theor. Biol.
1995*

viele zufällige
Kombinationen
von Helices
werden erzeugt

Mutation

TABLE 2

*Example of mutation in solutions
1 and 2*

1 0 0 0 0 0 . . . 0 1 0 0 → 1 0 0 **1** 0 0 . . . 0 1 0 0
0 0 0 1 0 1 . . . 1 0 1 0 → 0 **1** 0 1 0 1 . . . **0** 0 1 0

A change of 1 into 0 means that the corresponding stem was removed from that structure, and a change of 0 to 1 that the corresponding stem was introduced into the structure.

Zufällig
Einsen in
Nullen
umwandeln
und
umgekehrt

Crossover

TABLE 3

Example of crossover in one pair of solutions

Crossover points	
<u>1 0 0 0 0 0 . . . 0 1 0 0</u>	<u>1 0 0 1 0 1 . . . 1 0 0 0</u>
→	
0 0 0 1 0 1 . . . 1 0 1 0	0 0 0 <u>0 0 0 . . . 0 1 1 0</u>

Zufällig Paare
auswählen,
dann zufällig
Crossover-
Punkte festlegen

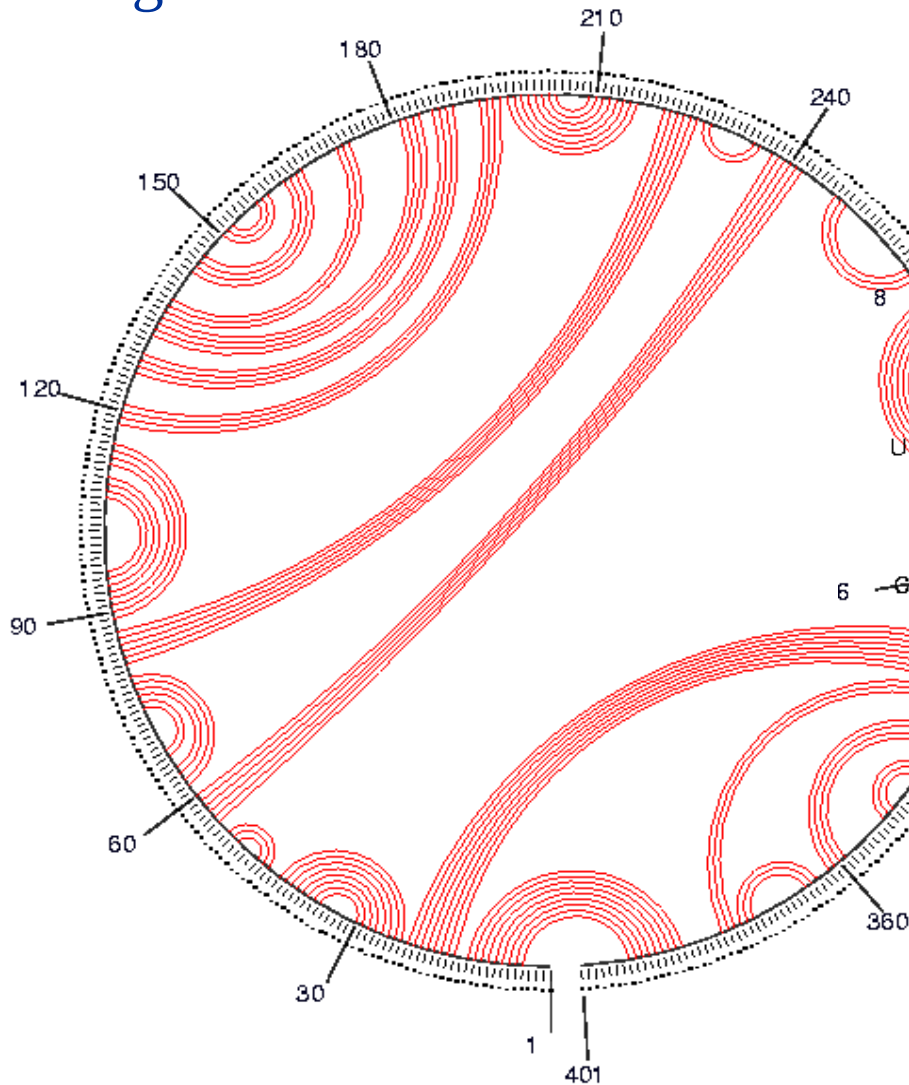
Auslese

- Fitness-Kriterium: Summe der Stacking-Energien für eine Struktur
- Strukturen mit besserer Fitness haben größere Chance, in die nächste Generation übernommen zu werden

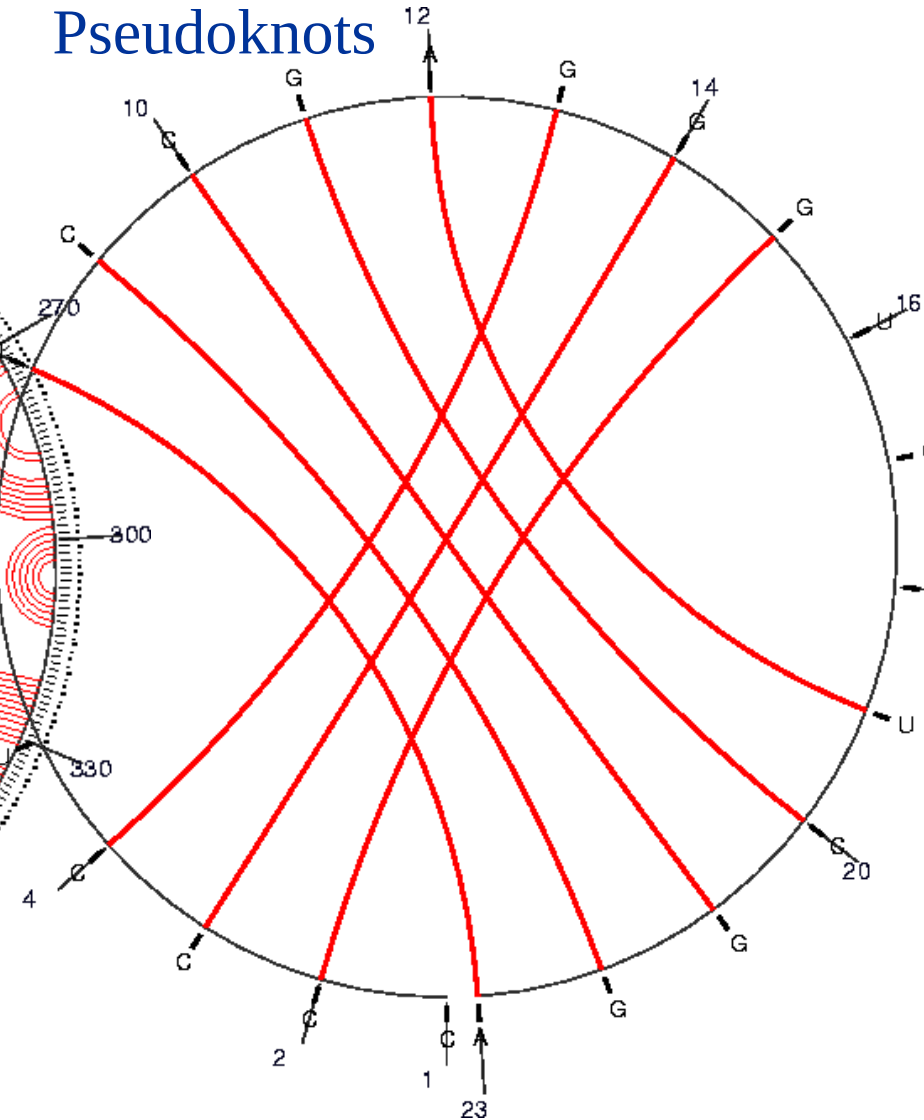
Darstellung von Sekundärstrukturen

RNA secondary structure representation

Legal structure

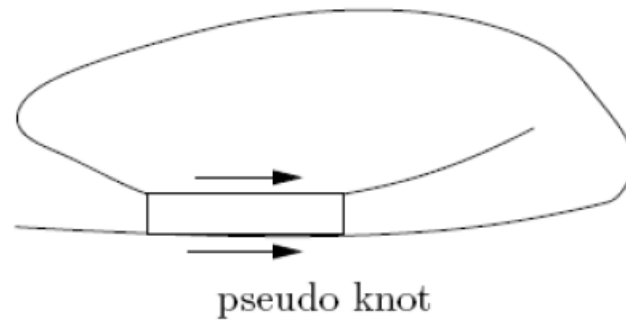
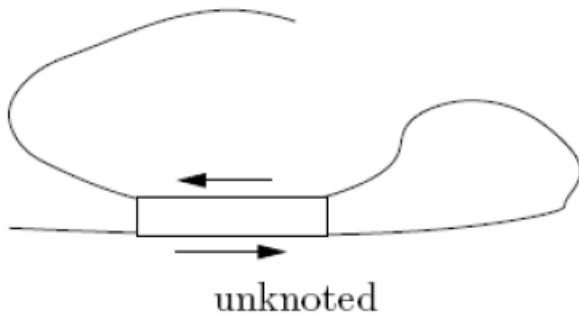


Pseudoknots



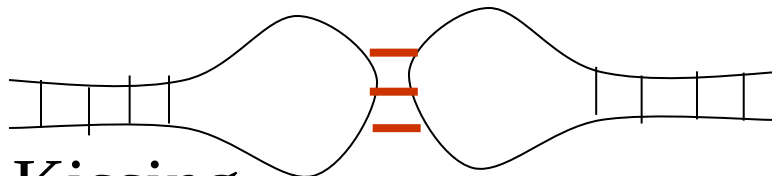
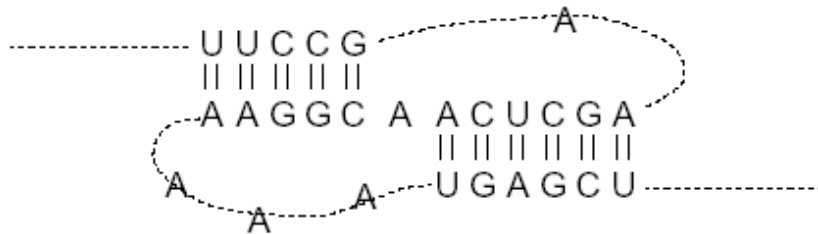
Wechselwirkung zwischen Sekundärstrukturen

Erkennen eines Pseudo-Knotens



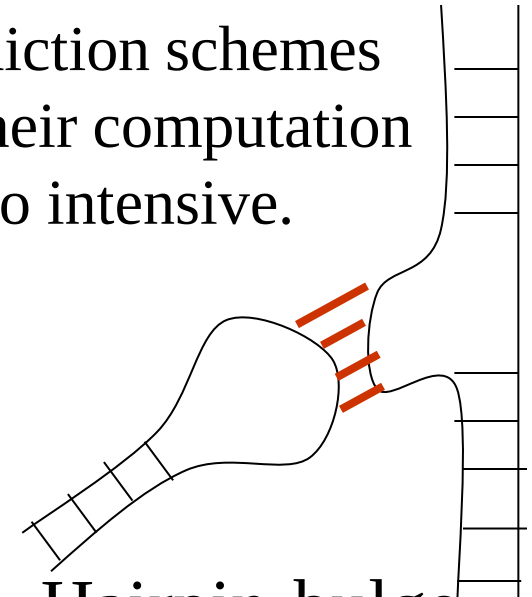
Wechselwirkung zwischen Sekundärstrukturen

Pseudo-knot



Kissing
hairpins

These patterns are excluded from the prediction schemes as their computation is too intensive.



Hairpin-bulge
contact

Energy Minimization

Predicting RNA Secondary Structure

- According to base pairing rules only, (A-U, G-C and wobble pairs G-U) sequences can potentially form many different structures
- An energy value is associated with each possible structure
- Predict the structure with the minimal free energy (MFE)

Energy Minimization Methods

- RNA folding is determined by biophysical properties
- Energy minimization algorithm predicts the correct secondary structure by minimizing the free energy
- ΔG calculated as sum of individual contributions of:
 - loops
 - base pairs
 - secondary structure elements
- Energies of stems calculated as stacking contributions between neighboring base pairs

The stability of a particular secondary structure is a function of several constraints:

1. The number of GC versus AU and GU base pairs.
(Higher energy bonds form more stable structures.)
2. The number of base pairs in a stem region.
(Longer stems result in more bonds.)
3. The number of base pairs in a hairpin loop region.
(Formation of loops with more **than 10 or less than 5** bases requires more energy.)
4. The number of unpaired bases, whether interior loops or bulges. (Unpaired bases decrease the stability of the structure.)

Stability of secondary structure

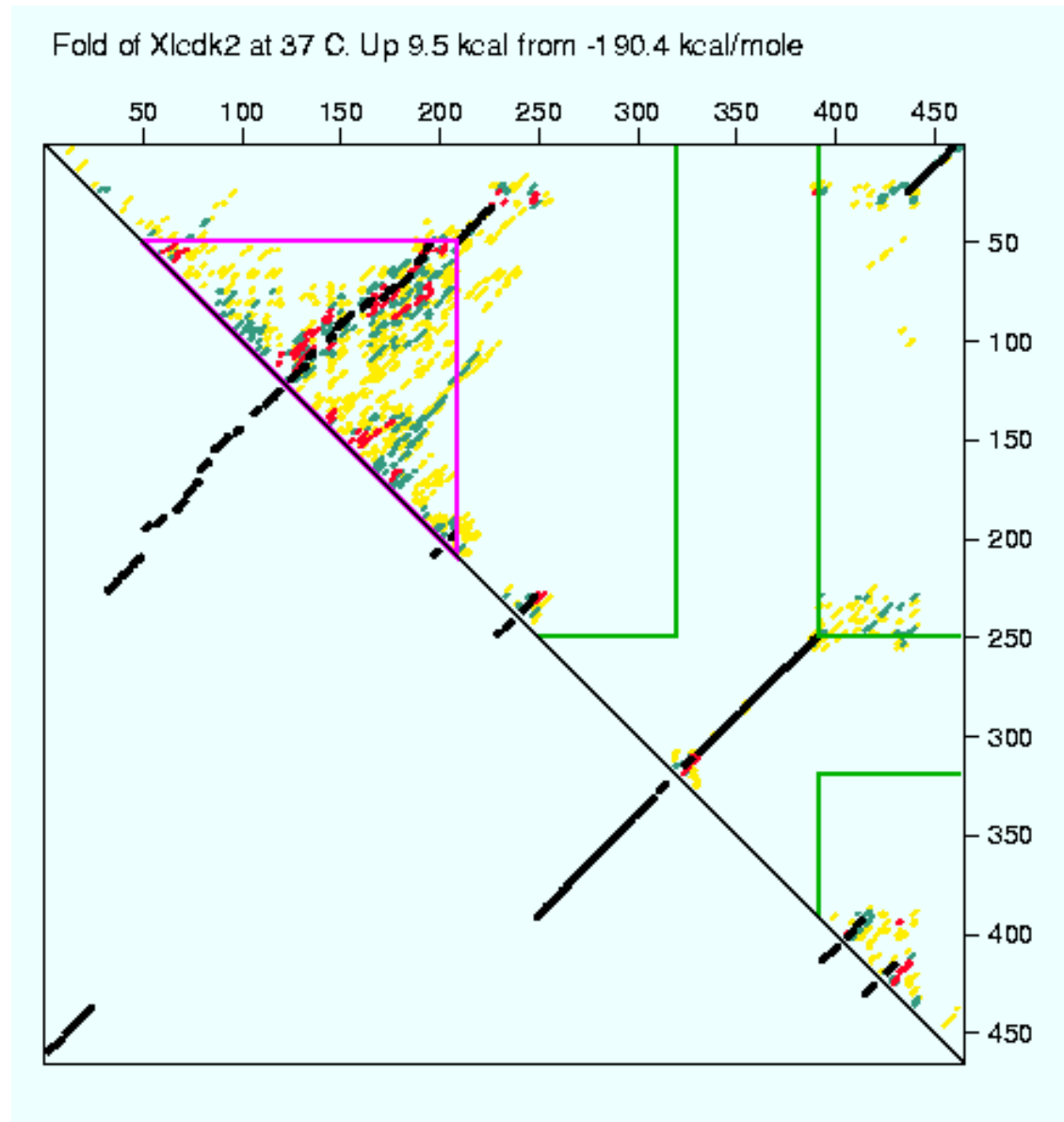
- The stability of a secondary structure is quantified as the **amount of free energy** released or used by **forming base pairs**.
- Positive free energy requires work to form a configuration; negative free energies release stored work.
- Free energies are additive, so one can determine the total free energy of a secondary structure by adding all the component free energies (units are kilocalories per mole).
- The more negative the free energy of a structure, the more likely is formation of that structure, because more stored energy is released. This fact is used to predict the secondary structure of a particular sequence.
- Discovering a base pair configuration with the minimum possible free energy is the goal of most secondary structure prediction algorithms.

Calculating Best Structure

- Thermodynamic Stability: Given the energy tables estimated by experimental techniques, the free energy can be calculated for a structure
- sequence is compared against itself using a dynamic programming approach
 - similar to the maximum base-paired structure
 - instead of using a scoring scheme, the score is based upon the free energy values

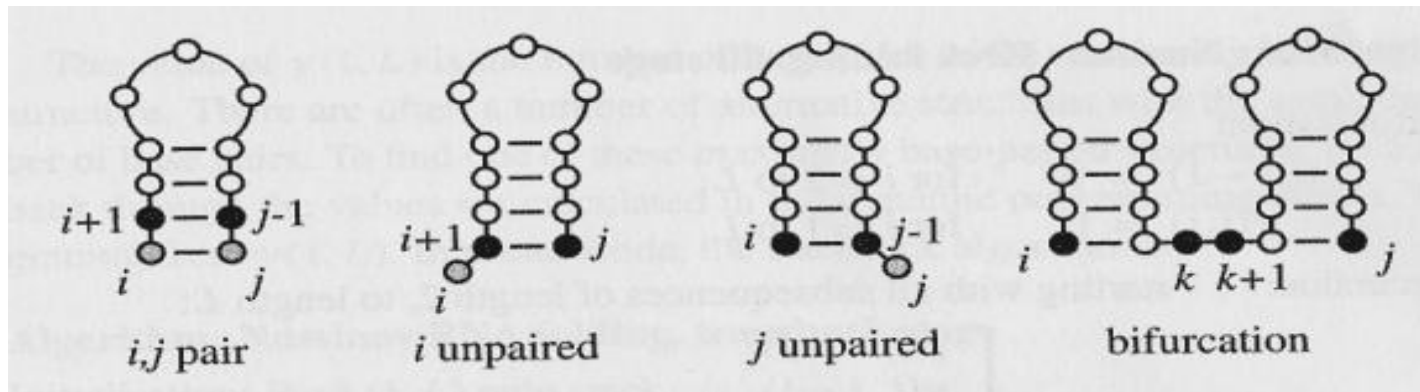
- To compute the minimum free energy of a sequence, **empirical energy parameters** are used.
- These parameters summarize free energy change (positive or negative) associated with all possible pairing configurations, including base pair stacks and internal base pairs, internal, bulge and hairpin loops, and various motifs which are known to occur with great frequency.

RNA Folding by Energy Minimization



1. *Ab initio* structure prediction

- Simple method: maximizing the number of base pairs (Nussinov *et al*, 1978)

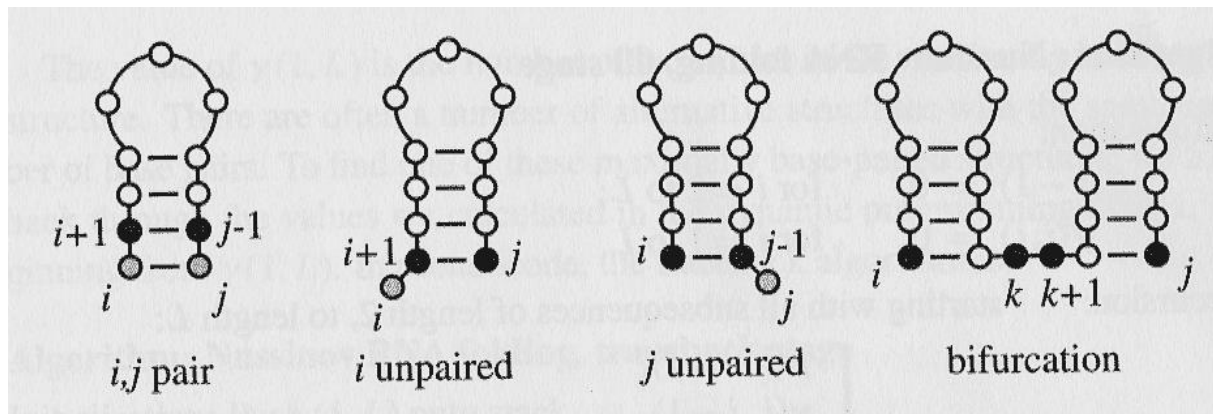


$$C_{i,j} = \max \left\{ \begin{array}{l} C_{i+1,j-1} + \delta(i,j), \\ C_{i,j-1}, \\ C_{i+1,j}, \\ \max_{i \leq k < j} \{ C_{i,k} + C_{k+1,j} \} \end{array} \right\}$$

simple case
 $\delta(i, j) = 1$

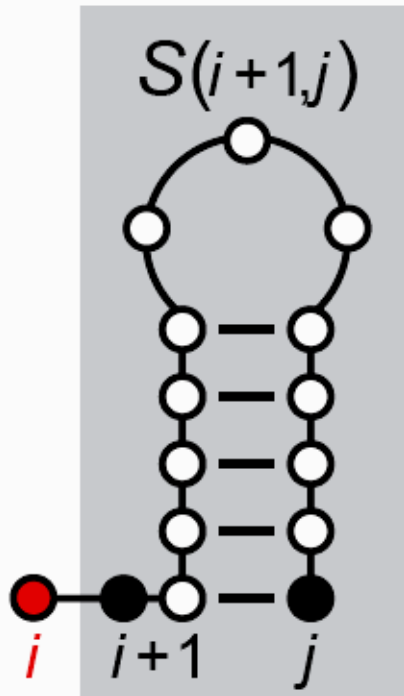
Base Pair Maximization – Nussinov Algorithm

$$S(i,j) = \max \begin{cases} S(i+1, j-1) + 1 & [\text{if } i,j \text{ base pair}] \\ S(i+1, j) \\ S(i, j-1) \\ \max_{i < k < j} S(i, k) + S(k+1, j) \end{cases}$$



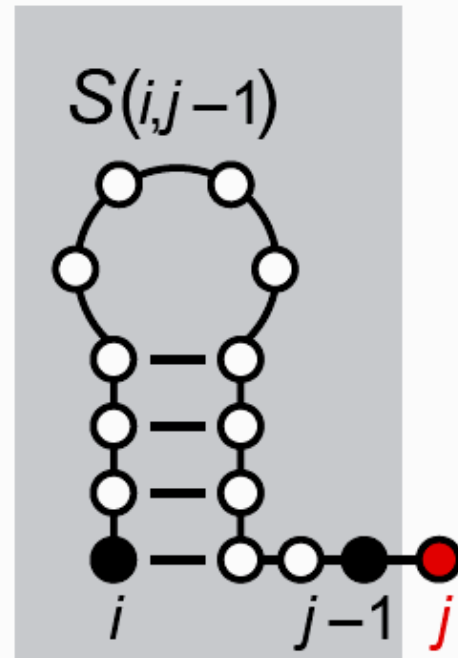
RNA folding: Dynamic Programming

There are only four possible ways that a secondary structure of nested base pair can be constructed on a RNA strand from position i to j :



1. i is unpaired, added on to a structure for $i+1 \dots j$

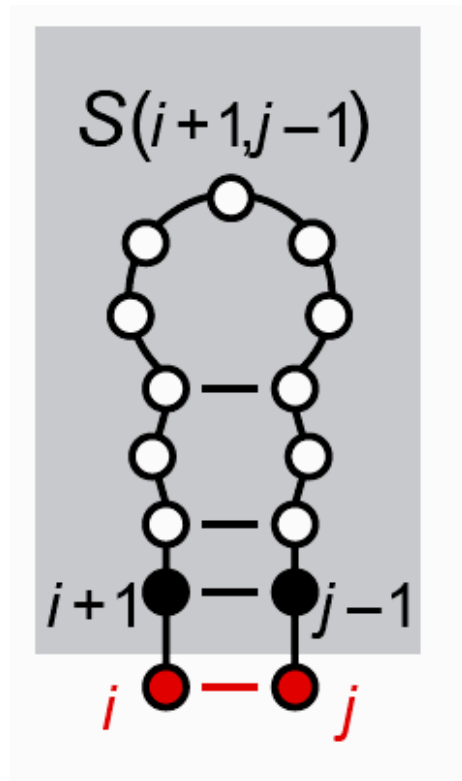
$$S(i, j) = S(i+1, j)$$



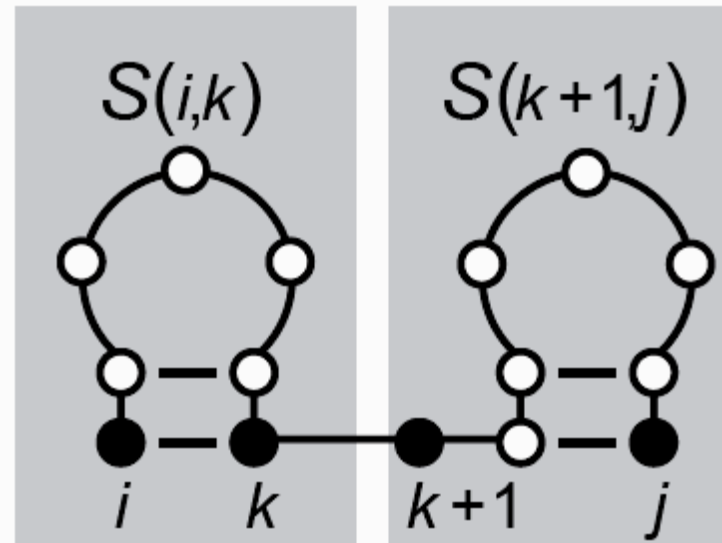
2. j is unpaired, added on to a structure for $i \dots j-1$

$$S(i, j) = S(i, j-1)$$

RNA folding: Dynamic Programming



3. i j paired, added on to a structure for $i+1 \dots j-1$
- $$S(i, j) = S(i+1, j-1) + e(r_i, r_j)$$



4. i j paired, but not to each other; the structure for $i \dots j$ adds together structures for 2 sub regions, $i \dots k$ and $k+1 \dots j$
- $$S(i, j) = \max_{i < k < j} \{S(i, k) + S(k+1, j)\}$$

RNA folding: Dynamic Programming

Since there are only four cases, the optimal score $S(i,j)$ is just the maximum of the four possibilities:

$$S(i, j) = \max \left\{ \begin{array}{ll} S(i+1, j) & r_i \text{ unpaired} \\ S(i, j-1) & r_j \text{ unpaired} \\ S(i+1, j-1) + e(r_i, r_j) & i, j \text{ base pair} \\ \max_{i < k < j} \{S(i, k) + S(k+1, j)\} & i, j \text{ paired, but not to each other} \end{array} \right.$$

To compute this efficiently, we need to make sure that the scores for the smaller sub-regions have already been calculated

Destabilizing energies of loops

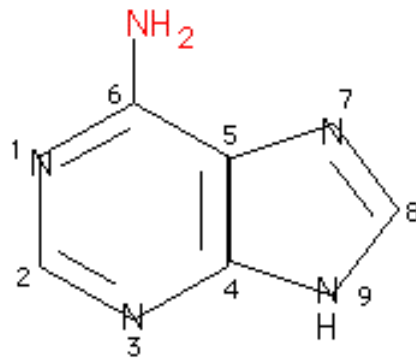
Size	Internal	Bulge	Hairpin
1	NA	3.8	NA
2	NA	2.8	NA
3	NA	3.2	5.6
4	1.7	3.6	5.5
5	1.8	4.0	5.6
6	2.0	4.4	5.3
7	2.2	4.6	5.8
8	2.3	4.7	5.4
30	3.7	6.1	7.7

Structure

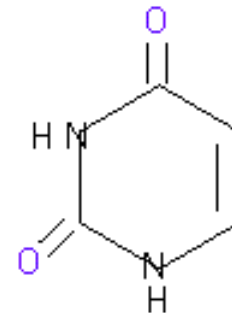
Structure

- An RNA molecule is composed of 4 types of (ribo)nucleotides.
- Each nucleotide contains a phosphate group, a sugar group (ribose) and a base. The polymer is formed by the linkage of the phosphate groups. The non-planar 5 member ribose ring connects the phosphate to the base.
- Finally, the bases are connected to the ribose group. Only the bases differ.
- The 4 different bases, adenine (A), cytosine (C), guanine (G) and uracil (U) are illustrated below.

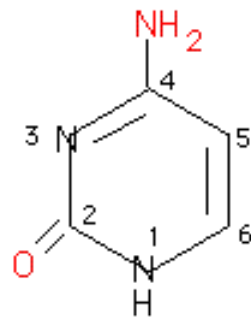
Structure



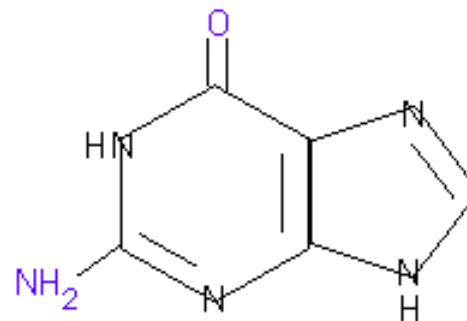
Adenine



Uracil

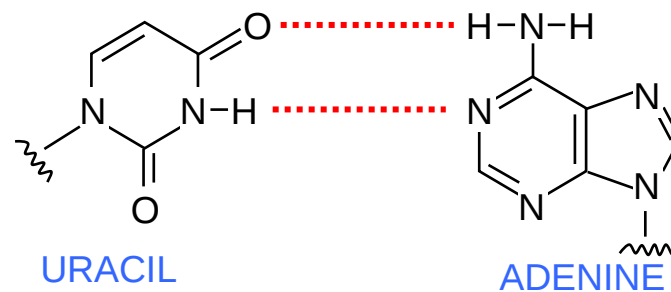
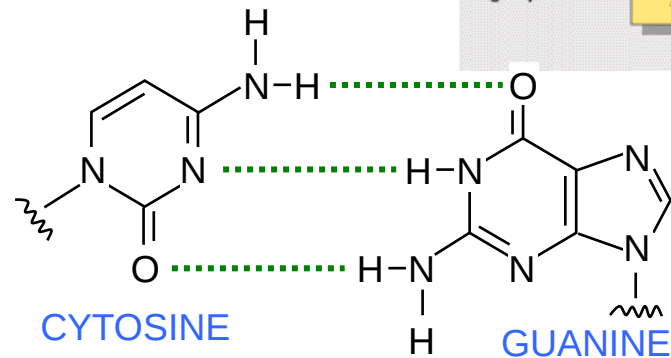
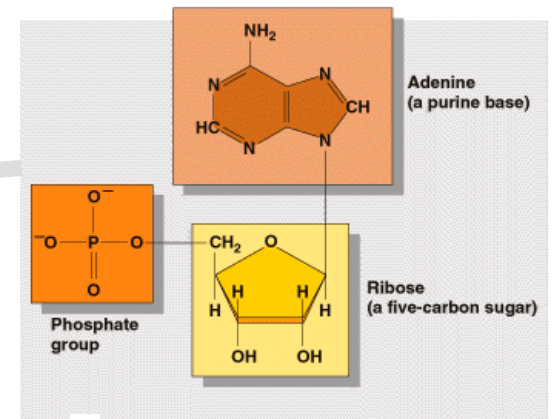
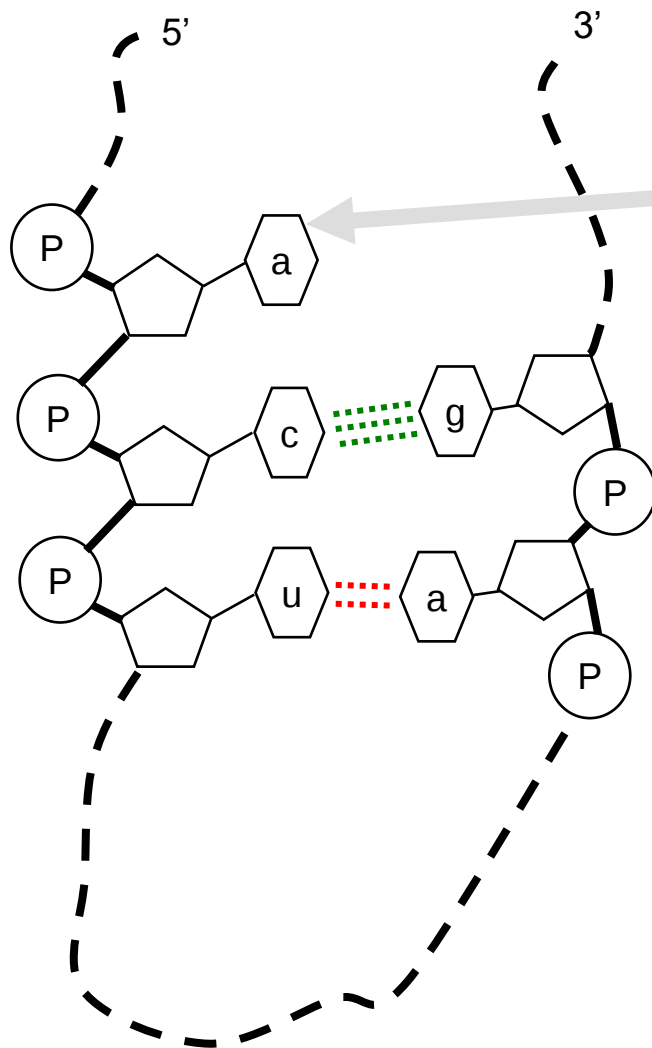


Cytosine



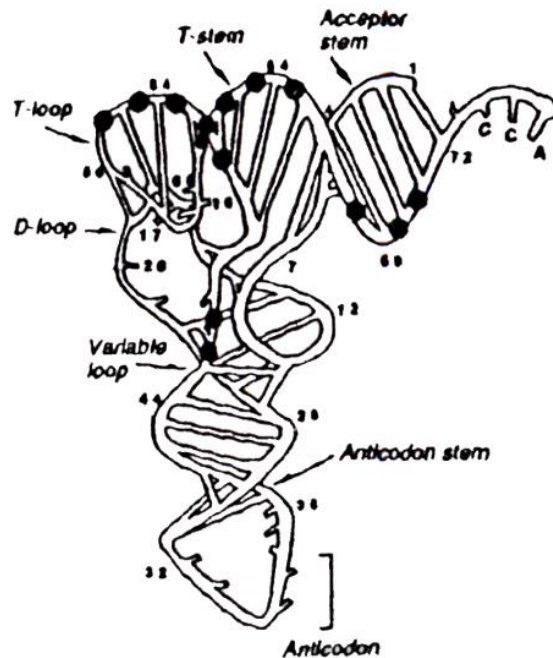
Guanine

Faltung der RNA



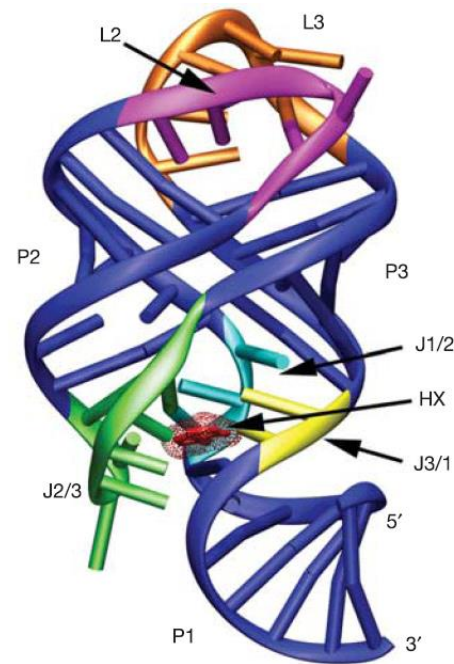
Sequence → Structure → Function

5'GCGGAUUUAG^{2M}
 CUCAGU^{DH}U^{DH}GGG
 AGAGCG^{M2}CCAGA
 C^{0M}UG^{0M}AAG^YAU^{PS}
 C^{5M}UGGAGG^{7M}UC
 C^{5M}UGUGU^{5M}U^{PS}C
 GA^{1M}UCCACAGAA
 UUCGACCA



tRNA

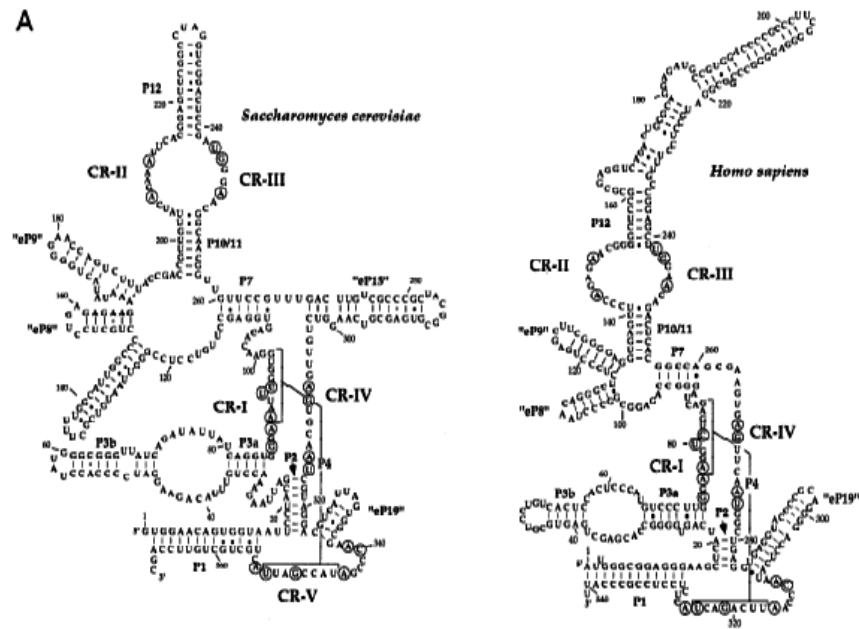
5'GGACAUAUAAU
 CGCGUGGAUAUG
 GCACGCAAGUUU
 CUACCGGGCACC
 GUAAAUGUCCGA
 CUAUGUCCA



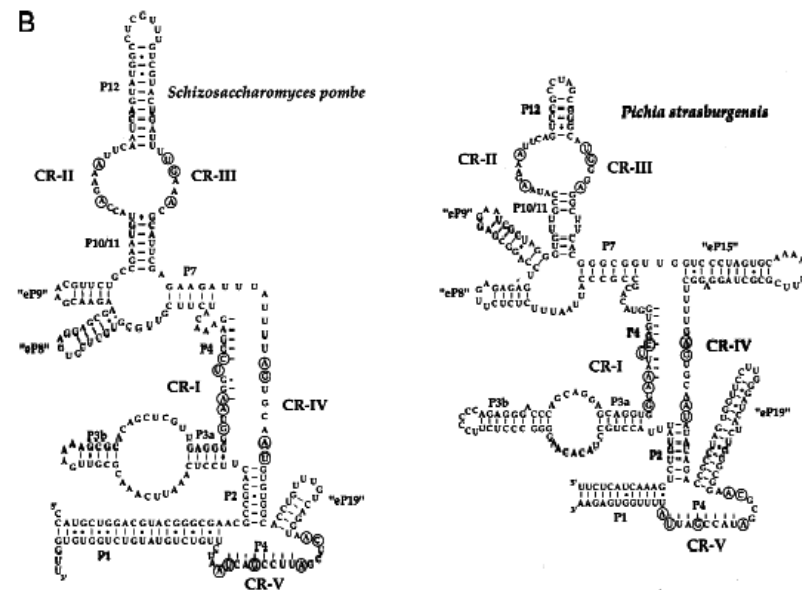
Guanine Riboswitch

Batey, R. et al, 2004 Nature vol. 432, p. 412

A



B



C

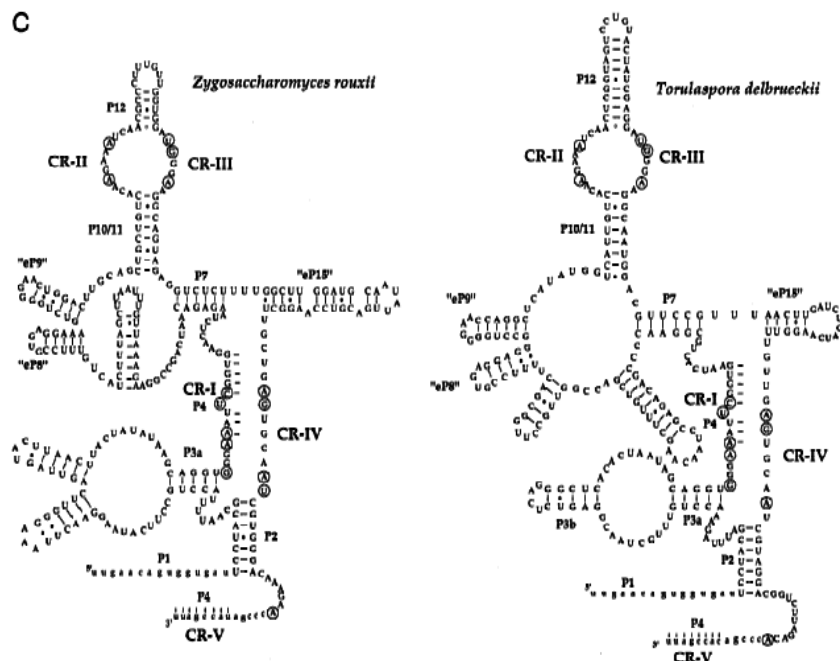


FIGURE 2. Secondary structures of eukaryal RNase P RNAs. Nucleotides that are absolutely conserved in archaeal, bacterial, and eukaryal RNase P RNAs are circled. The nomenclature for these conserved regions (CRs) is that of Chen and Pace (1997). Helices are numbered based on their putative homology to bacterial structures (Haas et al., 1994). Eukaryal helices of uncertain homology to particular bacterial or archaeal structures, but which occur at the same positions in the structures are designated "eP", for eukaryal paired region. Lower-case nucleotides represent primer sequences. (Figure continues on facing page.)

Features of RNA

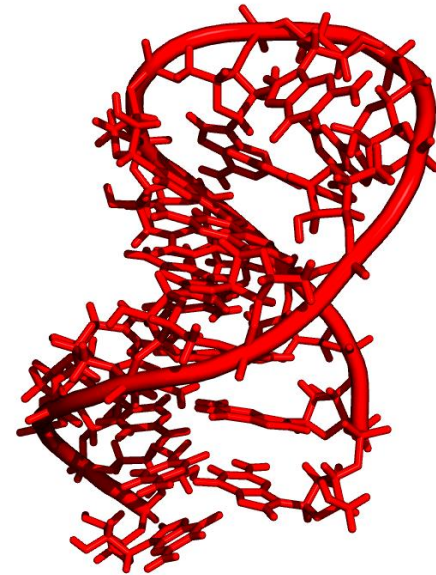
- RNA typically produced as a single stranded molecule (unlike DNA)
- Strand folds upon itself to form base pairs
- secondary structure of the RNA
 - intermediary between a linear molecule and a three-dimensional structure
 - Secondary structure mainly composed of double-stranded RNA regions formed by folding the single-stranded RNA molecule back on itself

Hierarchical organization of RNA molecules

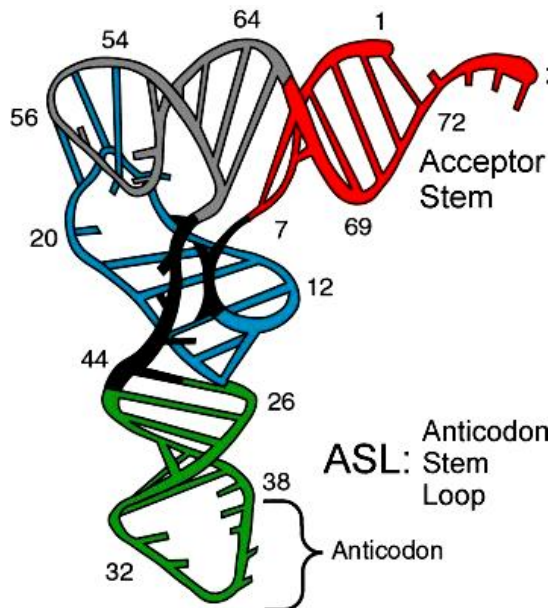
Primary structure:

5' ACCACCUGCUGA 3'

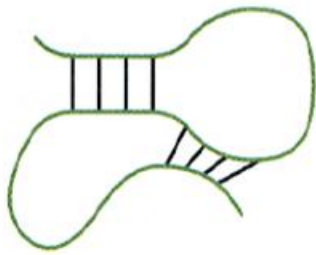
Secondary Structure



Tertiary structure:

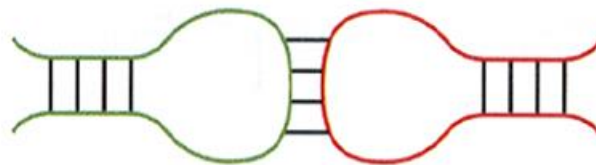


Tertiäre Strukturen



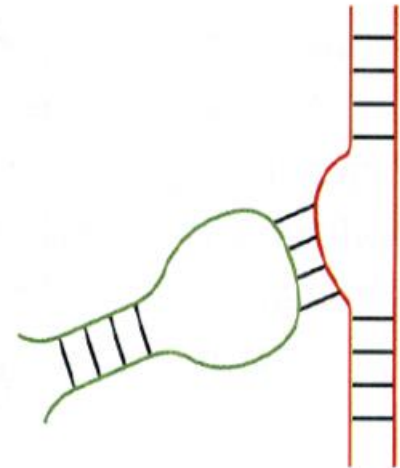
A.

Pseudoknoten



B.

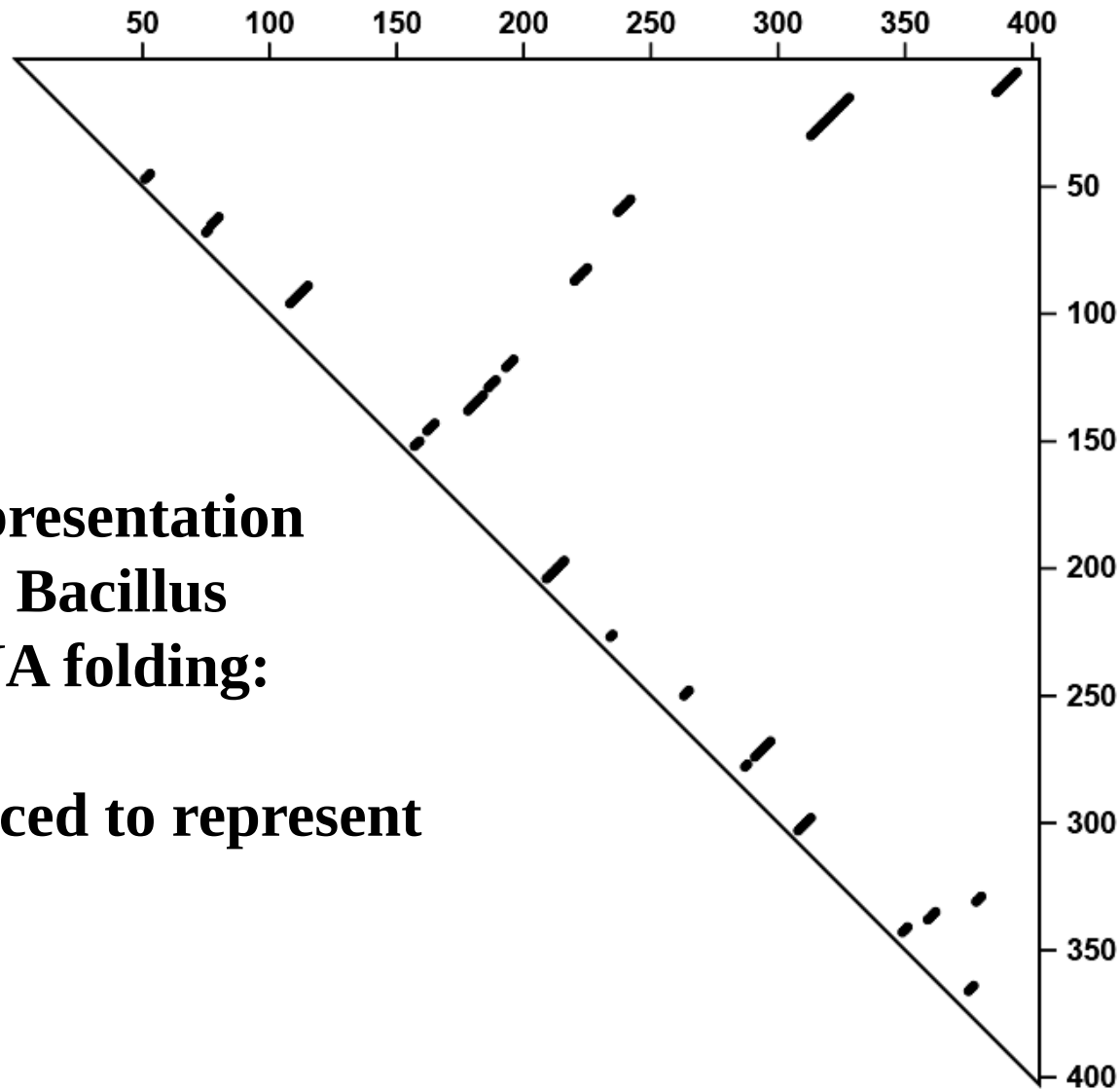
Kissing hairpins



C.

Hairpin-bulge

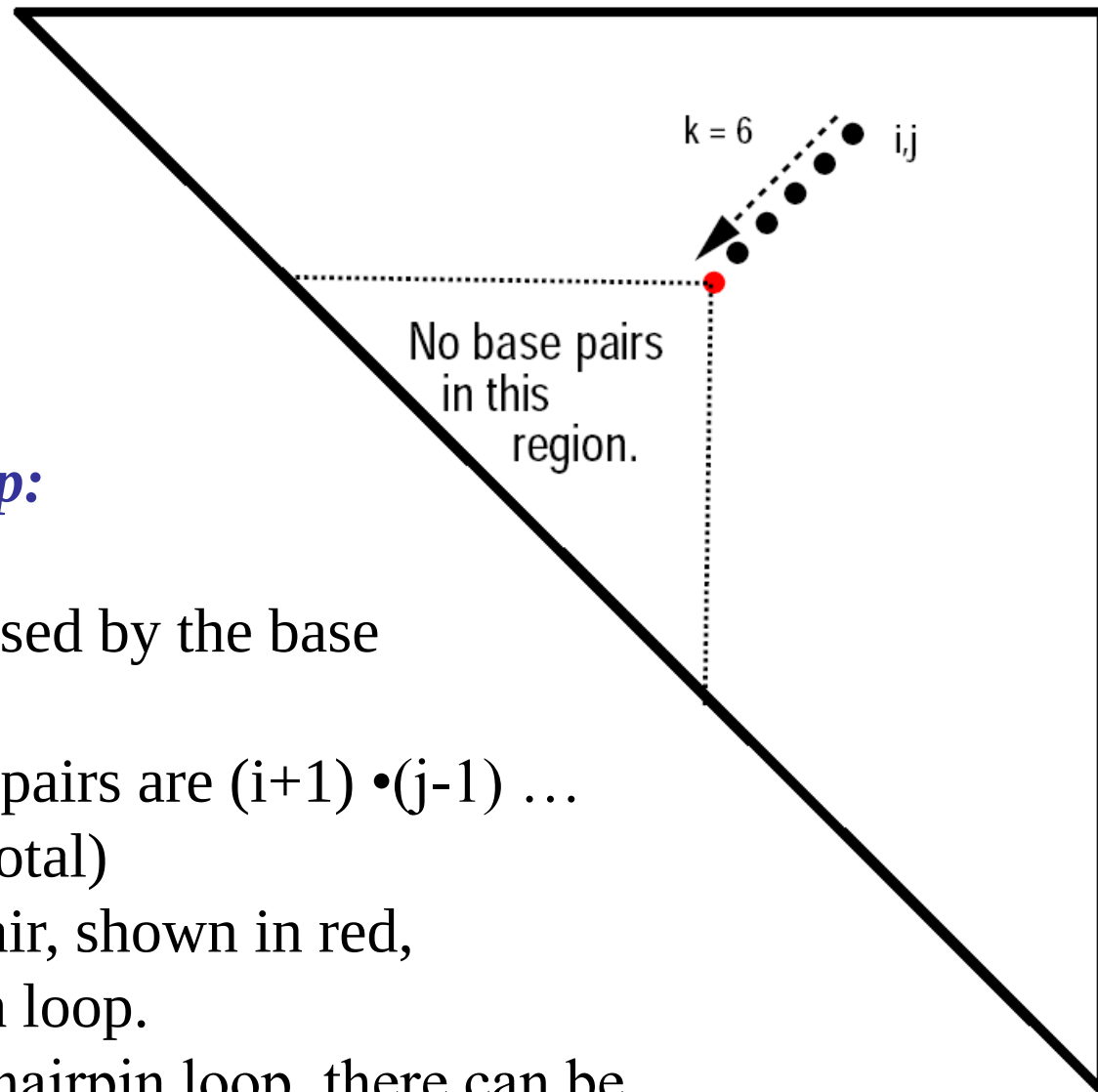
RNA secondary structure representation



**DotPlot representation
of the same *Bacillus
Subtilis* RNA folding:**

**A dot is placed to represent
a base pair**

Understanding RNA structure dot plot



Simple stem loop:

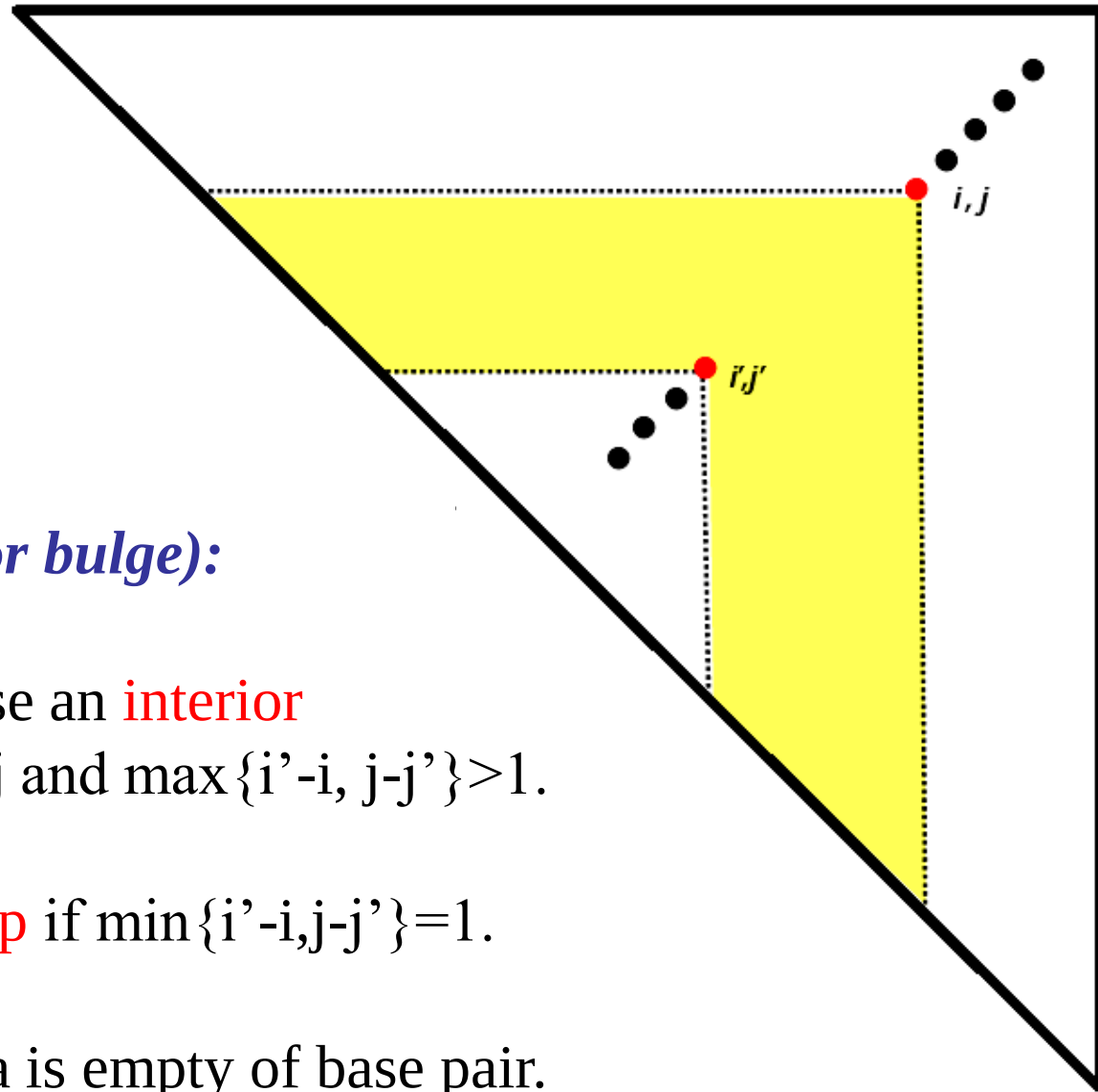
-Single helix closed by the base pair $i \bullet j$.

The other base pairs are $(i+1) \bullet (j-1) \dots (i+5) \bullet (j-5)$ (6 total)

-The last base pair, shown in red, closes a hairpin loop.

If $k \bullet l$ closes a hairpin loop, there can be no base pair $k' \bullet l'$ such that $k < k' < l' < l$

Understanding RNA structure dot plot



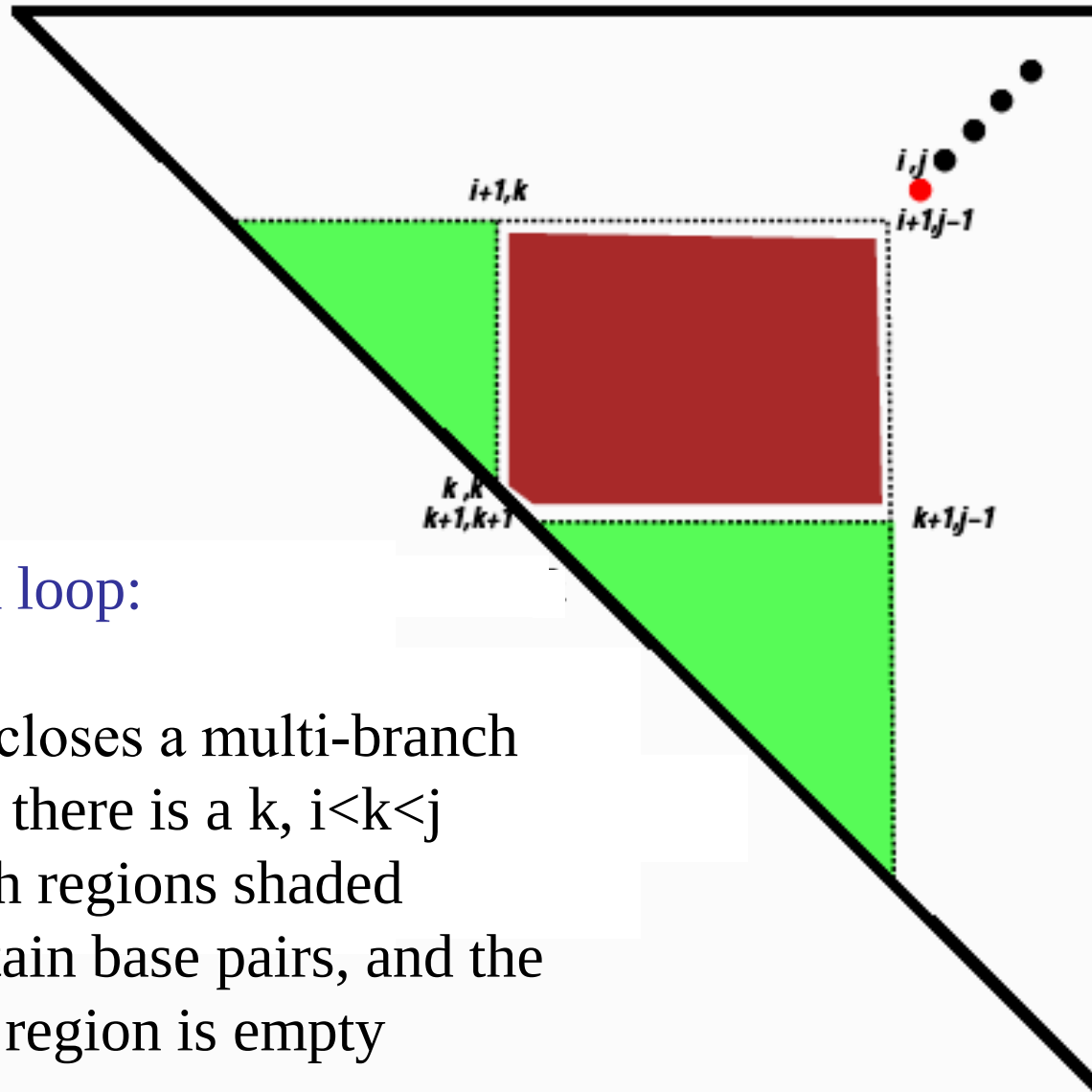
Interior loop (or bulge):

$i \cdot j$ and $i' \cdot j'$ close an **interior loop** if $i < i' < j' < j$ and $\max \{i' - i, j - j'\} > 1$.

It is a **bulge loop** if $\min \{i' - i, j - j'\} = 1$.

The yellow area is empty of base pair.

Understanding RNA structure dot plot



Multi-branch loop:

Base pair $i \bullet j$ closes a multi-branch if and only if there is a k , $i < k < j$ such that both regions shaded In green contain base pairs, and the other shaded region is empty

Context-Free Grammars

SCFGs

- Stochastic Context Free Grammars (SCFGs) have also been used to model RNA secondary structure
- Examples
 - tRNAScan-SE
 - program created to find snoRNAs
- Grammars are created by using a training set of data, and then the grammars are applied to potential sequences to see if they fit into the language

Stochastic Context Free Grammars

In analogy to HMMs, we assign probabilities to transitions:

Given grammar

$$X_1 \rightarrow s_{11} \mid \dots \mid s_{in}$$

...

$$X_m \rightarrow s_{m1} \mid \dots \mid s_{mn}$$

Can assign probability to each rule, s.t.

$$P(X_i \rightarrow s_{i1}) + \dots + P(X_i \rightarrow s_{in}) = 1$$

Recall HMMs:

Forward: $f_t(k) = P(O_1 \dots O_t, \pi_t = k)$

Backward: $b_t(k) = P(O_{t+1} \dots O_T \mid \pi_t = k)$

Then,

$$P(\mathbf{O}) = \sum_k f_T(k) = \sum_k p(\pi_1=k) e_k(O_1) b_1(k)$$

Analogue in SCFGs:

Inside: $a(i, j, V) = P(x_i \dots x_j \text{ is generated by nonterminal } V)$

Outside: $b(i, j, V) = P(x, \text{excluding } x_i \dots x_j \text{ is generated by } S \text{ and the excluded part is rooted at } V)$

An SCFG for simple Stem-loop Structures

15% $S \rightarrow a S t$

15% $S \rightarrow t S a$

25% $S \rightarrow c S g$

25% $S \rightarrow g S c$

5% $S \rightarrow g S t$

5% $S \rightarrow t S g$

10% $S \rightarrow L$

75% $L \rightarrow N L$

25% $L \rightarrow N N N N$

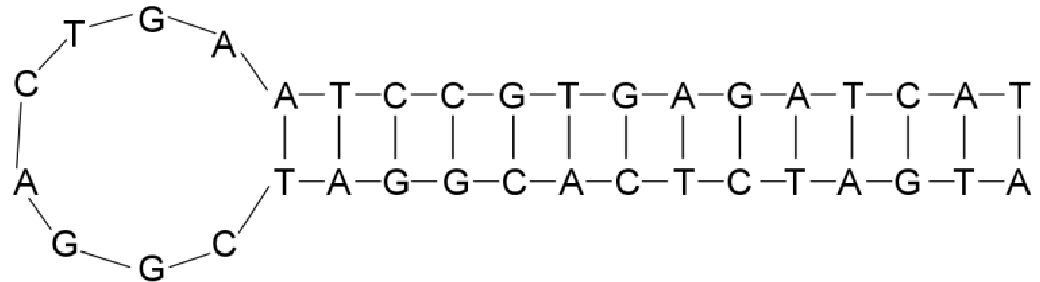
20% $N \rightarrow a$

30% $N \rightarrow c$

30% $N \rightarrow g$

20% $N \rightarrow t$

(generate paired bases in the stem)



(generate a loop of length ≥ 4)

(generate each nucleotide in the loop)

Decoding

- HMM: Viterbi Algorithm
- SCFG: CKY Algorithm

SCFG and HMM algorithms

<u>GOAL</u>	<u>HMM algorithm</u>	<u>SCFG algorithm</u>
Optimal parse	Viterbi	CYK
Estimation	Forward Backward	Inside Outside
Learning	EM: Fw/Bck	EM: Ins/Outs
Memory Complexity	$O(N K)$	$O(N^2 K)$
Time Complexity	$O(N K^2)$	$O(N^3 K^3)$

Where K: # of states in the HMM
of nonterminals in the SCFG

Thermodynamik

Thermodynamics

- Gibbs Free Energy, G
 - Describes the energetics of molecules in aqueous solution. The change in free energy, ΔG , for a chemical process, such as nucleic acid folding, can be used to determine the direction of the process:
 - $\Delta G=0$: equilibrium
 - $\Delta G>0$: unfavorable process
 - $\Delta G<0$: favorable process
 - Thus the natural tendency for biomolecules in solution is to minimize free energy of the entire system (biomolecules + solvent).
- $\Delta G = \Delta H - T\Delta S$
 - ΔH is enthalpy, ΔS is entropy, and T is the temperature in Kelvin.
 - Molecular interactions, such as hydrogen bonds, van der Waals and electrostatic interactions contribute to the ΔH term. ΔS describes the change of order of the system.
 - Thus, both molecular interactions as well as the order of the system determine the direction of a chemical process.
 - For any nucleic acid solution, it is extremely difficult to calculate the free energy from first principle

Boltzmann-Verteilung

$$\frac{N_i}{N} = \frac{g_i \exp(-E_i/kT)}{\sum_j g_j \exp(-E_j/kT)}$$

- N_i is the number of molecules in state i which has energy E_i and degeneracy g_i
- N is the total number of molecules in the system
- k is the Boltzmann constant
- sometimes the above equation is written without the degeneracy factor g_i . In this case the index i will specify an individual state, rather than a set of g_i states having the same energy E_i .
- Because velocity and speed are related to energy, Equation 1 can be used to derive relationships between temperature and the speeds of molecules in a gas.
- The denominator in this equation is known as the canonical partition function.
- Gas: $E = m/2v^2 \rightarrow$ Beispielsweise kann feuchte Wäsche bei Temperaturen von 20 °C trocknen, da es in dieser Verteilungskurve einen geringen Anteil von Molekülen mit der erforderlich hohen Geschwindigkeit gibt, welche sich aus dem Flüssigkeitsverband lösen können. Es wird also bei niedrigen Temperaturen immer einige Moleküle geben, die schnell genug sind, die Anziehungskräfte durch ihre Nachbarn zu überwinden und vom flüssigen oder festen Aggregatzustand in den gasförmigen Aggregatzustand überzugehen, was man als Verdampfung bzw. Sublimation bezeichnet.

Boltzmann-Statistik

$$p_j = \frac{1}{Z} e^{-\beta E_j}$$

- Gilt für Systems, das im thermodynamischen Gleichgewicht an ein Wärmebad der Temperatur T gekoppelt ist (kanonisches Ensemble)
- Gibt Wahrscheinlichkeit für Zustand mit der Energie E_j an
- Anstelle von Wahrscheinlichkeiten lässt sich die Boltzmann-Statistik auch durch Teilchenzahlen ausdrücken:

$$N_j = N_0 g_j e^{-\frac{E_j}{k_B T}} .$$

N_j = Zahl der Teilchen, die den Zustand j besetzen; N_0 = Teilchenzahl des 0-ten Zustands; g_j = Entartungsgrad des Zustands j (also die Anzahl von Zuständen gleicher Energie E_j)

Free energy is determined by the sum over all possible conformations

Free energy: $F = -kT \ln Q$

Partition function:

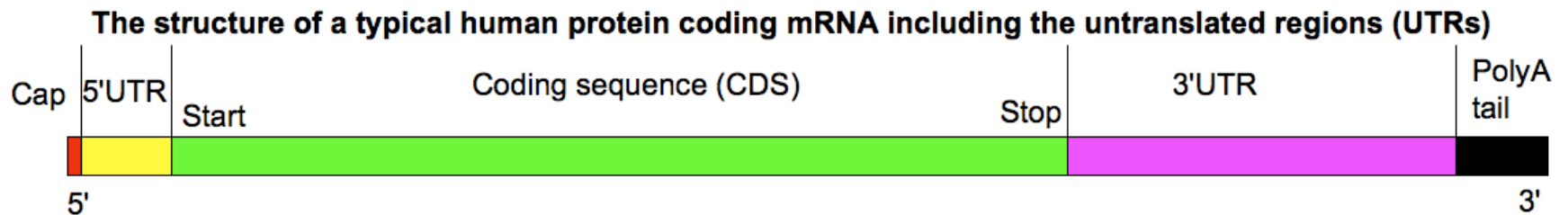
$$Q = \sum_{\text{micorstates}} e^{-E/kT} = \sum_{\text{macrostates}} e^{-F_A/kT}$$

Sonstiges

Ribosomen

- Ribosomes are complexes of RNA and protein that are found in all cells.
- The ribosome functions in the process of translation. Ribosomes catalyze the assembly of individual amino acids into polypeptide chains; this involves binding a messenger RNA and then using this as a template to join together the correct sequence of amino acids.

Struktur einer mRNA



"RNA-Welt"

- The RNA world hypothesis proposes that a world filled with life based on ribonucleic acid (RNA) predated current life based on deoxyribonucleic acid (DNA). RNA, which can both store information like DNA and act as an enzyme, may have supported cellular or pre-cellular life. Some hypotheses as to the origin of life present RNA-based catalysis and information storage as the first step in the evolution of cellular life.
- The RNA world is proposed to have evolved into the DNA and protein world of today. DNA, through its greater chemical stability, took over the role of data storage while protein, which is more flexible in catalysis through the great variety of amino acids, became the specialized catalytic molecules. The RNA world hypothesis suggests that RNA in modern cells, in particular rRNA (RNA in the ribosome which catalyzes protein production), is the evolutionary remnant of the RNA world.



**We finished the genome map, now we can't
figure out how to fold it!**

Science (1989) 243, p.786

Free Energy

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

ΔH = enthalpy change
T temperature
 ΔS = entropy change



ΔH and ΔS are considered to be temperature independent and allow free energy calculation at any temperature

Elemente, die zur Berechnung von $S(5,11)$ beitragen

$j \longrightarrow$

$i \downarrow$

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