Flexible and universal multiple target sequence design

Stefan Hammer, Sven Findeiß, Birgit Tschiatschek, Christoph Flamm, Ivo L. Hofacker

universität wien

Introduction to Sequence Design

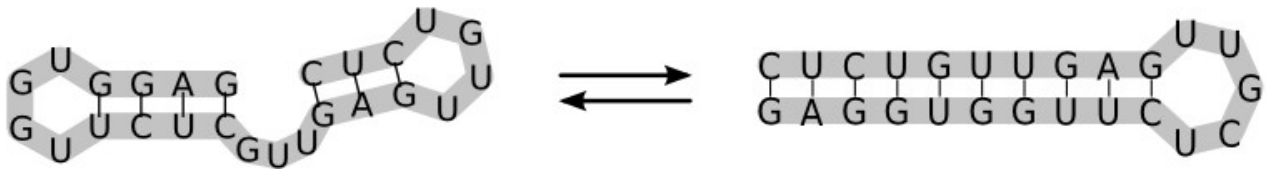
Structure(s)

```
(((((....)))))...(((....)))
((((((((((.....))))))))))
```

Inverse folding problem:
"Find a sequence that has this minimum free energy (and low suboptimal) structure(s)"

Sequences

```
GAGGUGGUUCUCGUUGAGUUGUCUC
-3.3 kcal
```



universität wien

Three main Components

The design of sequences with desired structural properties is seen as an optimization problem with:

- Sequence space (or a subset thereof) as search space
- A cost function that quantifies the fitness of a sequence to the design goal
- A suitable move set to use for the optimization

Solution Space

- Solution space is usually huge → stochastic sampling necessary

```
((((....))))...((....))
((((((((((.....))))))))))
Size of solution space: 1.67326e+08
```

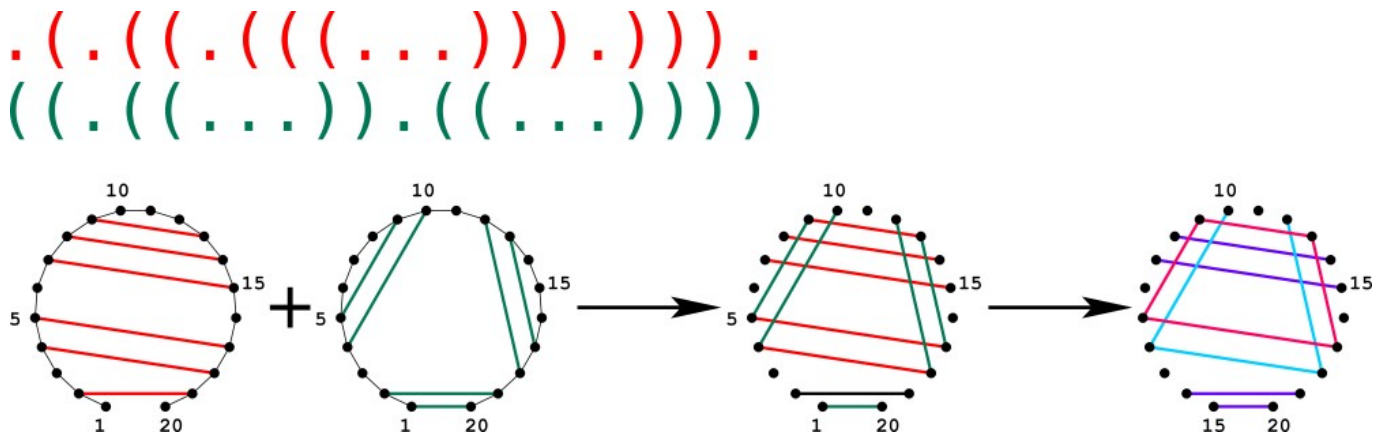
- Fair sampling increases efficiency dramatically

  Probability of every solution: 1 / solution space size

Dependency Graph

A so called *Dependency Graph Ψ* is used to represent all structural constraints. It is derived from the union of circle representations.

- Ψ is a finite, undirected graph with
- |V| = sequence-length and
- |E| = number of base pairs



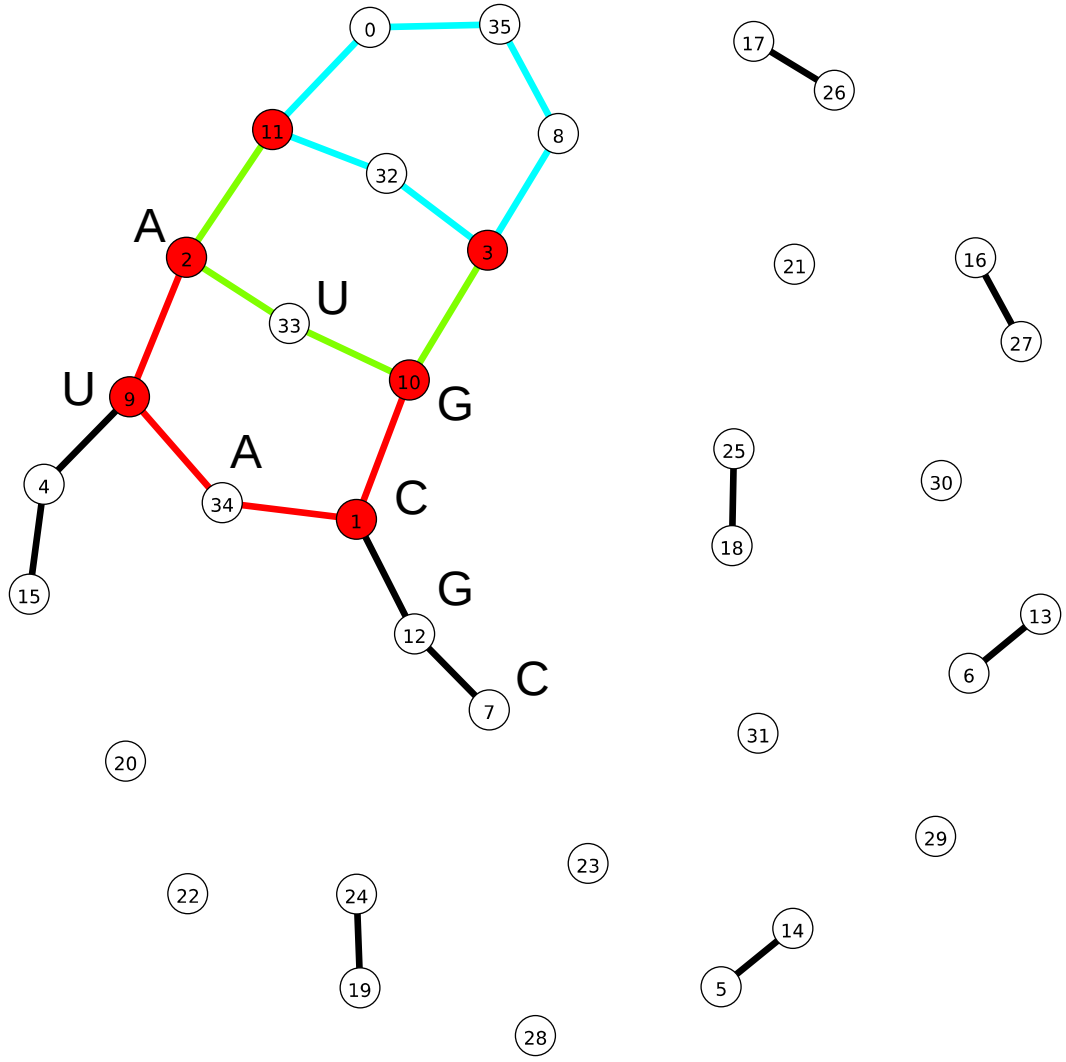
universität wien

Graph Coloring

- Assign bases to the positions:

Pairing Matrix:

1	A	C	G	U	N
A	0	0	0	1	1
C	0	0	1	0	1
G	0	1	0	1	2
U	1	0	1	0	2
N	1	1	2	2	6



universität wien

Coloring paths stochastically

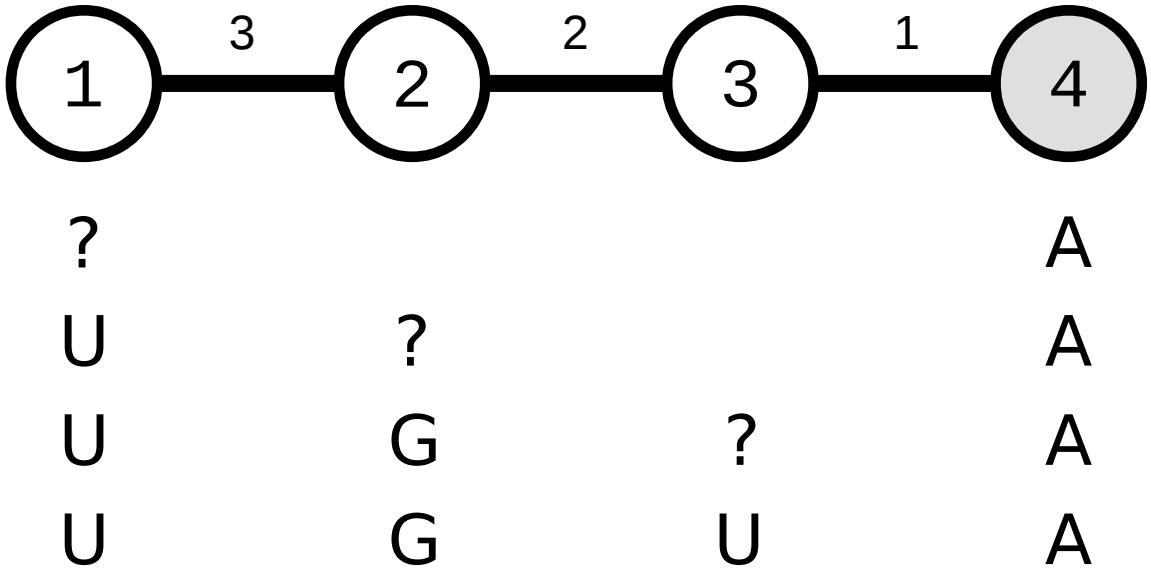
RNA Paring Matrices
for different path lengths:

0	A	C	G	U	N
A	1	0	0	0	1
C	0	1	0	0	1
G	0	0	1	0	1
U	0	0	0	1	1
N	1	1	1	1	4

1	A	C	G	U	N
A	0	0	0	1	1
C	0	0	1	0	1
G	0	1	0	1	2
U	1	0	1	0	2
N	1	1	2	2	6

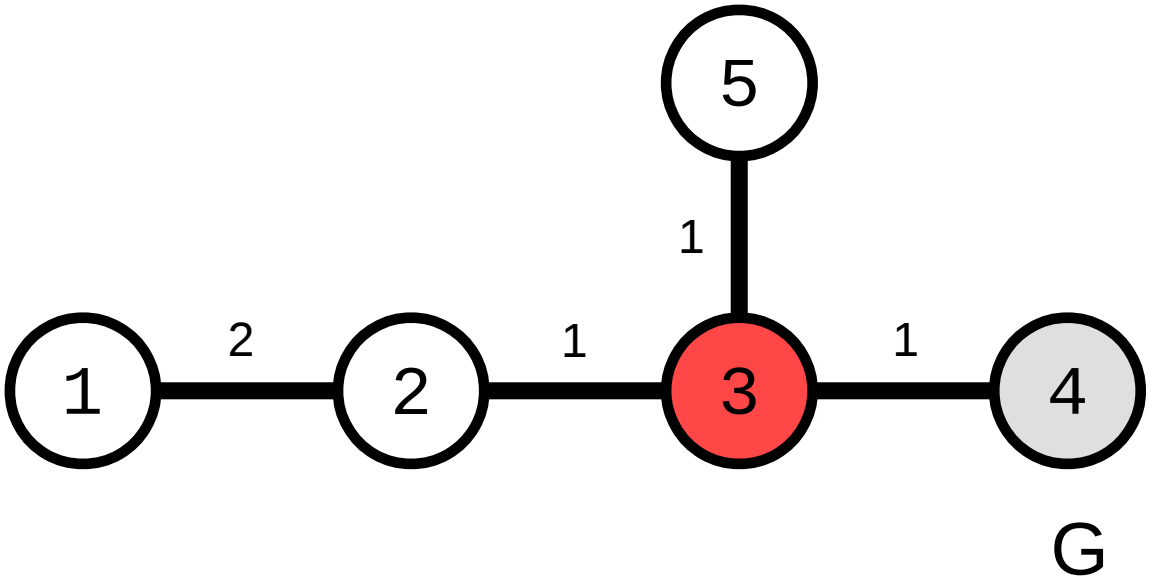
2	A	C	G	U	N
A	1	0	1	0	2
C	0	1	0	1	2
G	1	0	2	0	3
U	0	1	0	2	3
N	2	2	3	3	10

3	A	C	G	U	N
A	0	1	0	2	3
C	1	0	2	0	3
G	0	2	0	3	5
U	2	0	3	0	5
N	3	3	5	5	16



universität wien

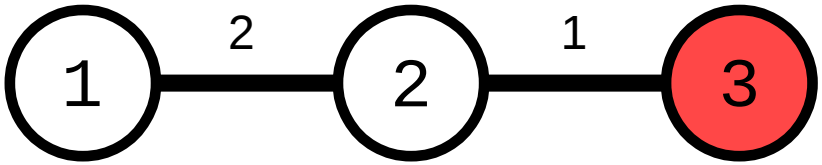
Coloring Junctions



universität wien

Probability Matrix

- Size is number of nucleotides (RNA: 4)
- Multidimensional, for every special one dimension ($4^{|specials|}$ values)
- Operator for merging matrizes: “multiplication”
- Operator for removing specials/dimensions
- 0 dimensions means only a total number is saved



Pairing Matrix:

2	A	C	G	U	N
A	1	0	1	0	2
C	0	1	0	1	2
G	1	0	2	0	3
U	0	1	0	2	3
N	2	2	3	3	10

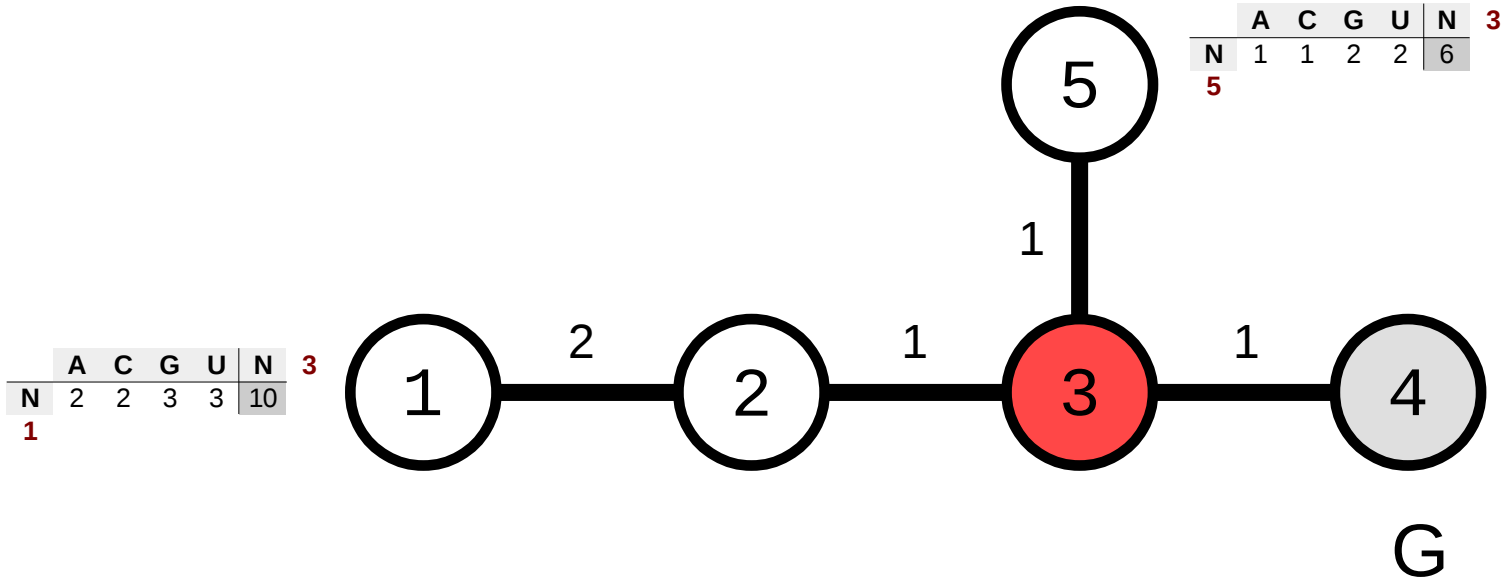
Probability Matrix:

	A	C	G	U	N	3
N	2	2	3	3	10	

1

universität wien

Coloring Junctions



	A	C	G	U	N	3
N	2	2	3	3	10	
1						

	A	C	G	U	N	3
N	1	1	2	2	6	
5						

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

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

o

	A	C	G	U	N	3
N	1	1	2	2	6	
5						

o

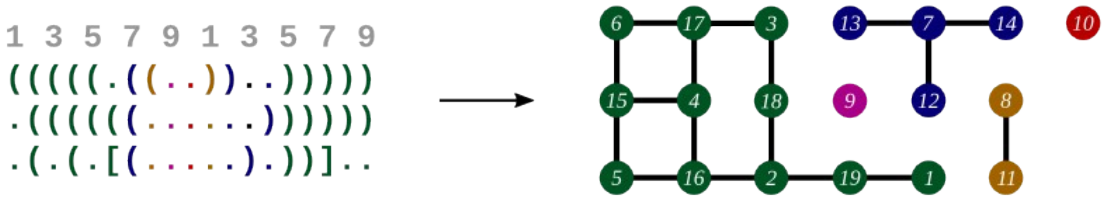
	A	C	G	U	N	3
N	2	2	3	3	10	
1						

=

	A	C	G	U	N	3
A	0	0	0	0	0	
C	0	0	0	0	0	
G	0	2	0	6	8	
U	0	0	0	0	0	
N	0	2	0	6	8	
4						

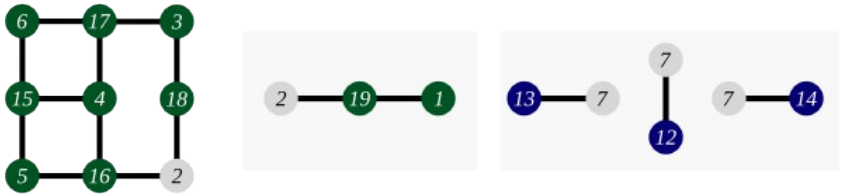
universität wien

Dependency Graph Decomposition

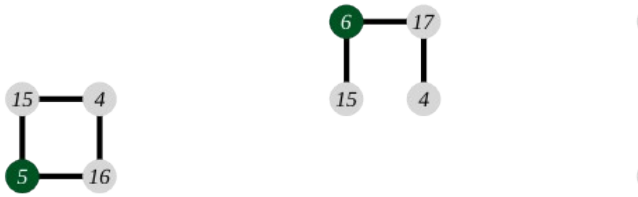


Connected Components

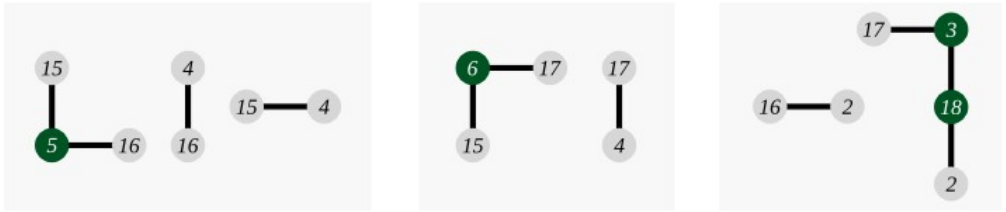
Biconnected Components



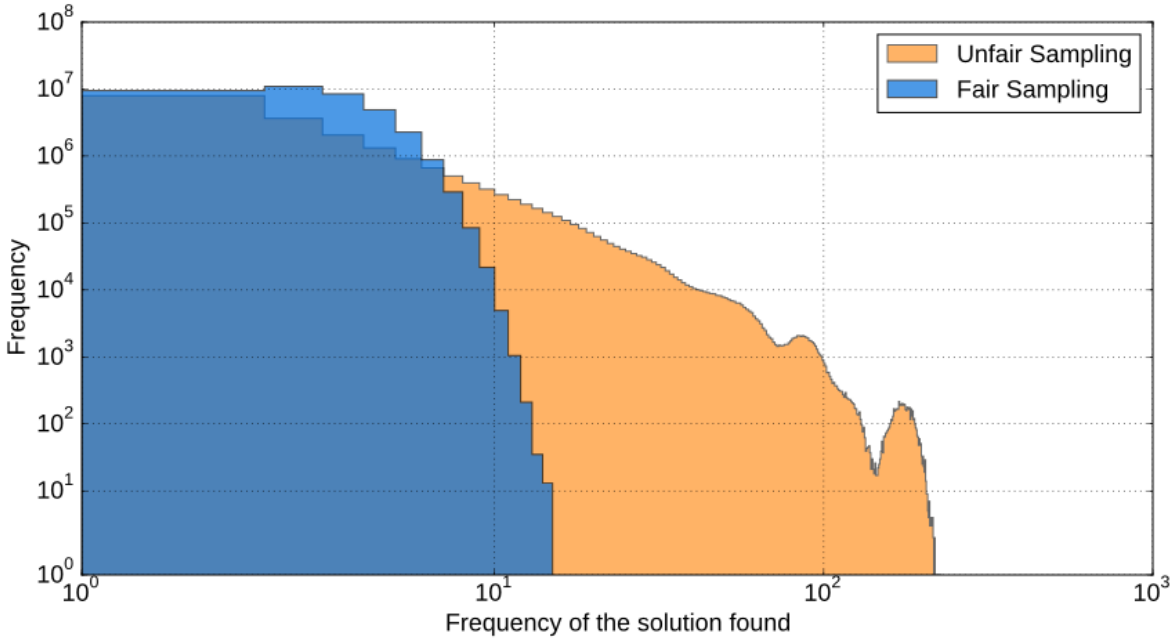
Ear Decomposition



Paths

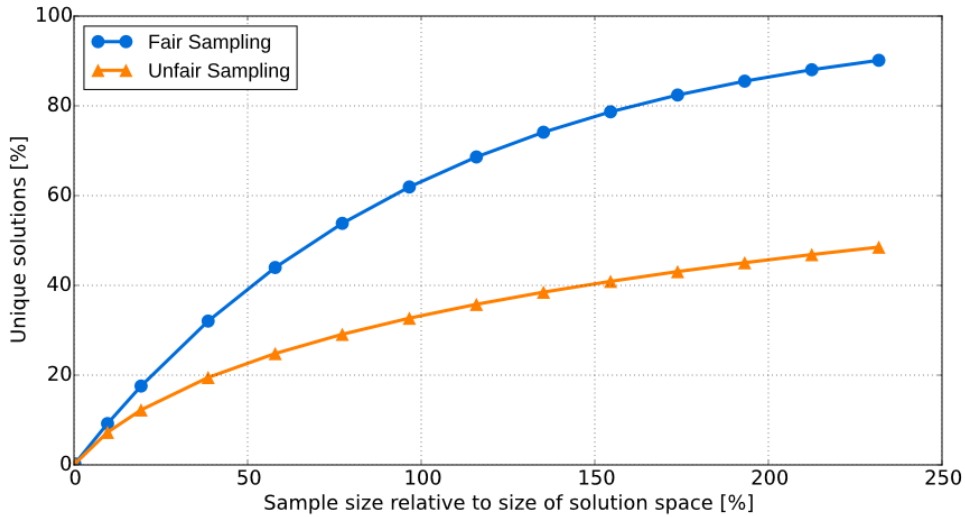


Fair vs. Unfair Sampling



$$E(X = d) = \sum_{k=|C|-d+1}^{|C|} \frac{|C|}{k}$$

C … Solution Space
d … Sample Size



universität wien

Advantages Fair Sampling

- For small sample sizes in huge solution spaces → no advantage

$$|C| >> d$$

- For small solutions spaces or huge sample sizes → big advantage (many constraints, complex graph,...)

$$|C| \sim d$$

- For demanding objective functions big advantage, no duplicate solutions

C … Solution Space
d … Sample Size

universität wien

Three main Components

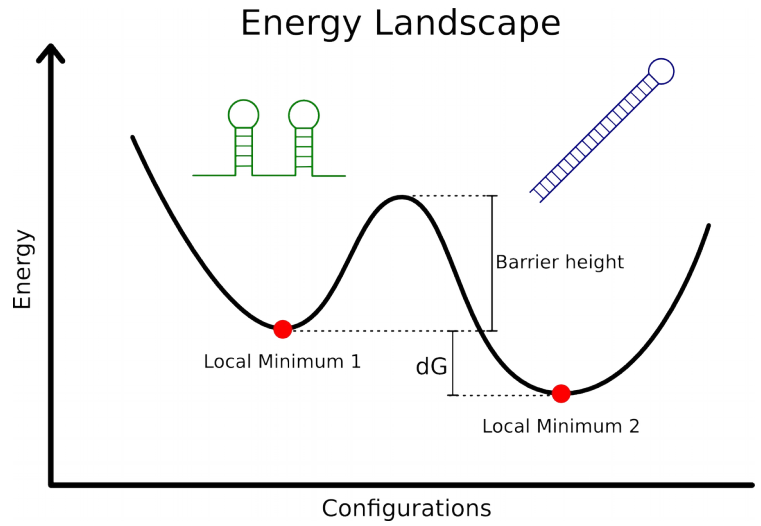
The design of sequences with desired structural properties is seen as an optimization problem with:

- Sequence space (or a subset thereof) as search space
- A cost function that quantifies the fitness of a sequence to the design goal
- A suitable move set to use for the optimization

One objective function

Design goals:

- Target states to be minima in landscape
- One target should be the global minimum (mfe structure)
- Target states to have similar energies
- Targets to be highly populated in ensemble



universität wien

One objective function

Design goals:

- Target states to be minima in landscape
- One target should be the global minimum (mfe structure)
- Target states to have similar energies
- Targets to be highly populated in ensemble

$$\Xi(x) = \frac{1}{M}\left(\sum_{m=1}^{M} f(x,\Theta_m) - g(x)\right) + \frac{2\cdot\gamma}{M(M-1)}\left(\sum_{m<l} |f(x,\Theta_m) - f(x,\Theta_l)|\right)$$

M … Number of structures
f(x, $\boldsymbol{\Theta}$*)*... Energy of structure
g(x)... Gibbs free energy
γ … Weighting factor

universität wien

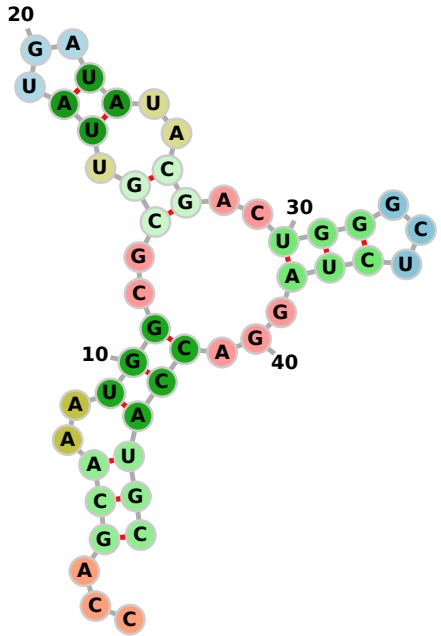
Three main Components

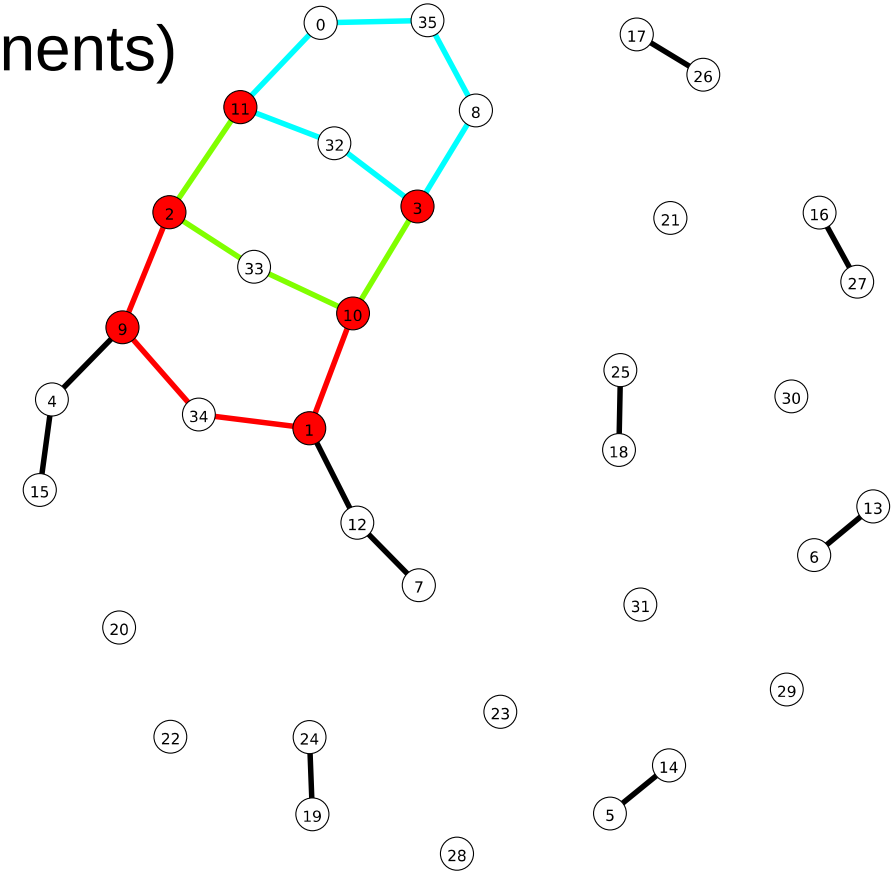
The design of sequences with desired structural properties is seen as an optimization problem with:

- Sequence space (or a subset thereof) as search space
- A cost function that quantifies the fitness of a sequence to the design goal
- A suitable move set to use for the optimization

Move Steps

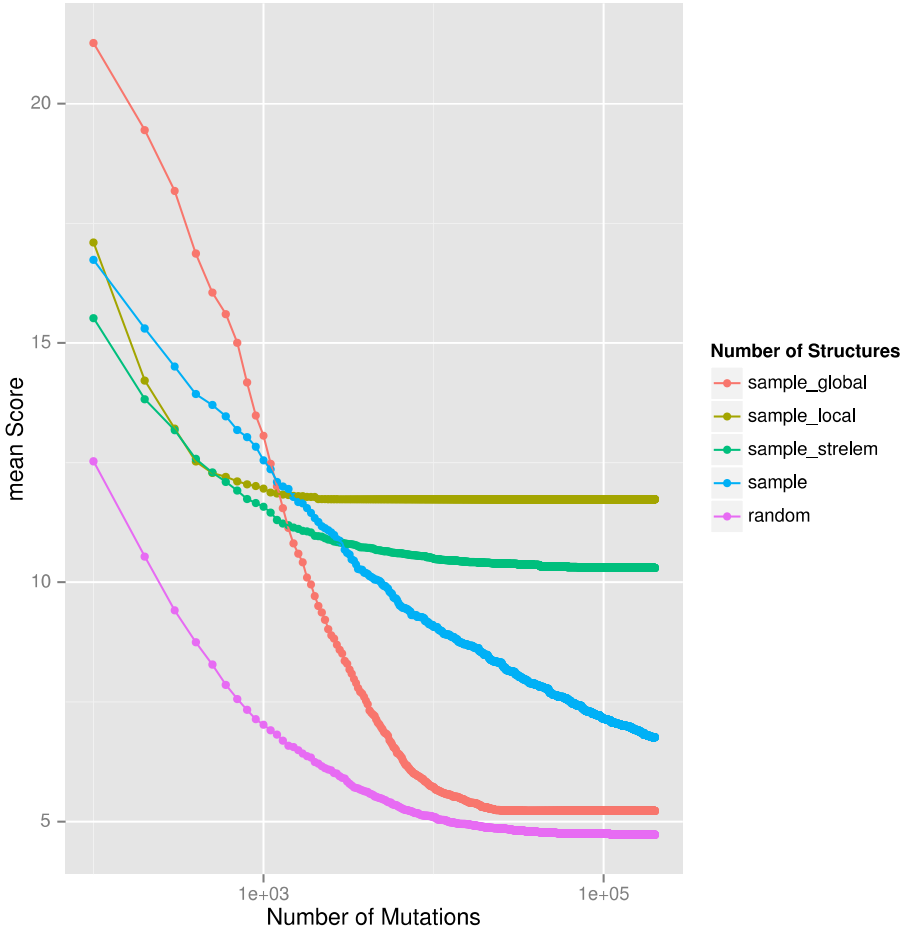
- Sample complete sequence
- Sample global (connected components)
- Sample local (paths w/o specials)
- Sample position
- Sample structural elements

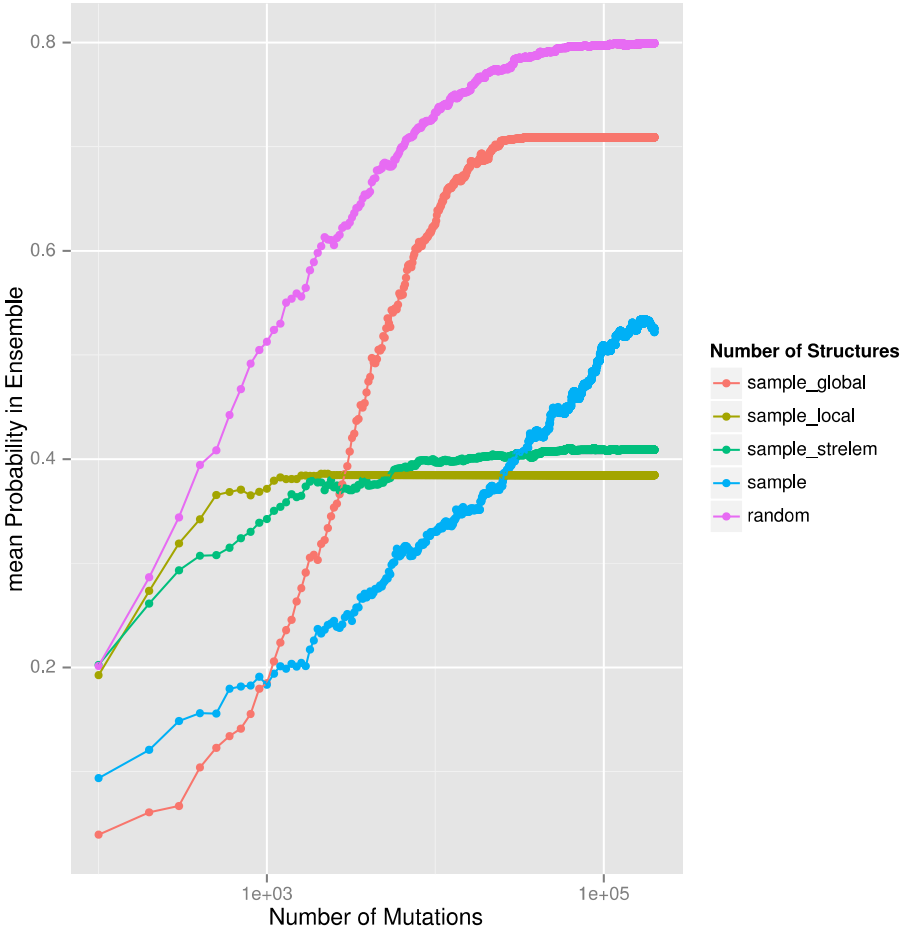




universität wien

Local minima traps





universität wien

Thanks to...

- Sven
- Birgit
- Christian
- Xtof, Peter, Ivo

- RNAdesign Library pre-release: https://github.com/ribonets/RNAdesign

universität wien

SEVENTH FRAMEWORK PROGRAMME

RIBO NETS

DK rna Bio Lage

TBI