Projection of the RNA secondary structure space: The κ, λ - neighborhood

Ronny Lorenz
ronny@tbi.univie.ac.at

Institute for Theoretical Chemistry
University of Vienna

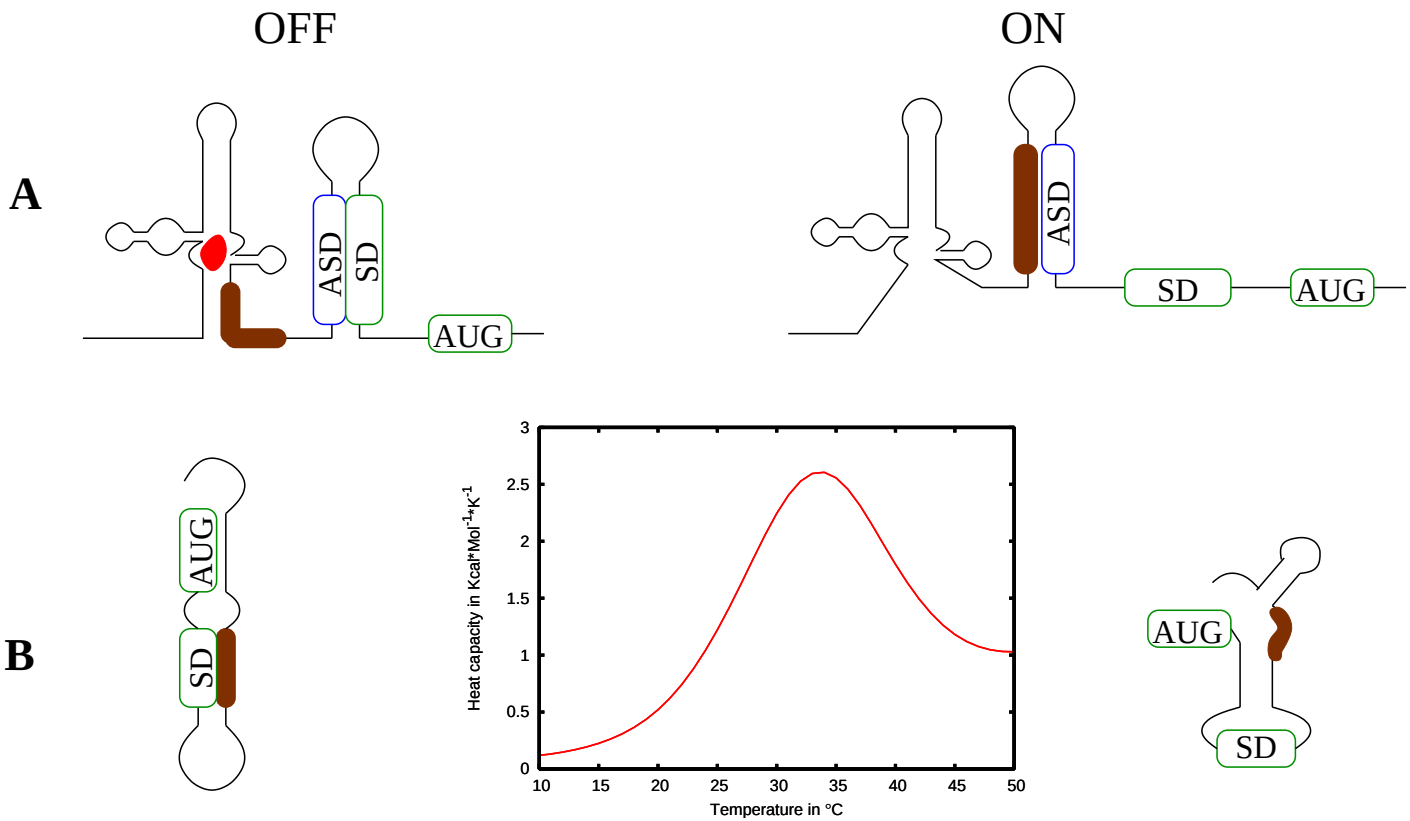
GCB, Halle, Germany, September 28, 2009

1. RNA structures and energy landscapes
   - Motivation
   - RNA structures
   - Loop decomposition
   - Nearest neighbor energy model
   - Secondary structure free energy landscape

2. The κ, λ - neighborhood of two secondary structure states
   - Representatives of the κ, λ - neighborhood
   - Complexity and implementation
   - Application and interpretation
   - Estimation of barrier heights using refolding paths

Gene regulation by RNA structure changes

- 5'-UTR element controlling translation
- metabolite sensors, (Riboswitch, A)
- temperature sensors, (RNA thermometer, B)

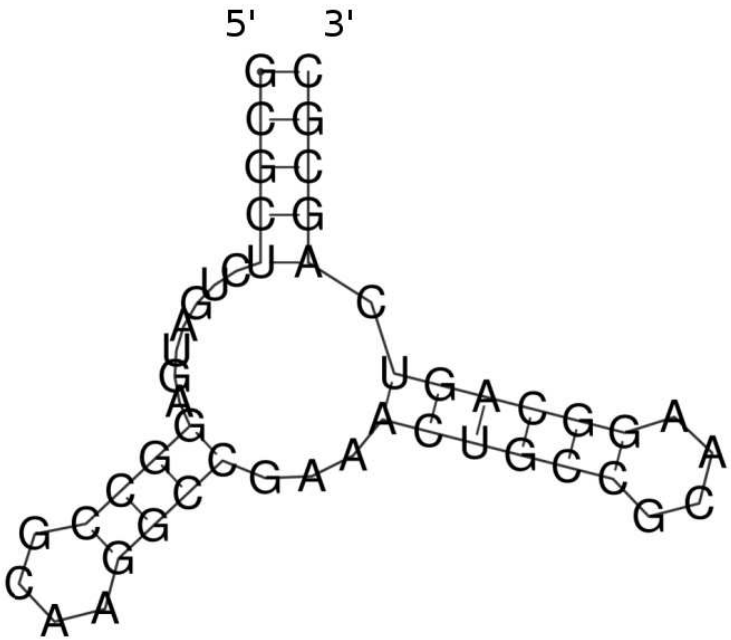


adapted from Henkin, T (2008) *Genes & Development* **22**:3383-3390

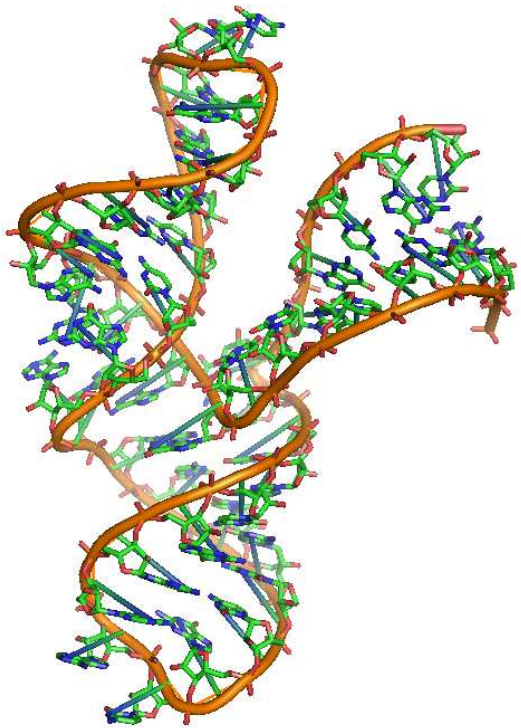
Primary structure

5' - GCGCUCUGAUGAGGCCGCAAGGCCGAAACUGCCGCAAGGCAGUCAGCGC - 3'

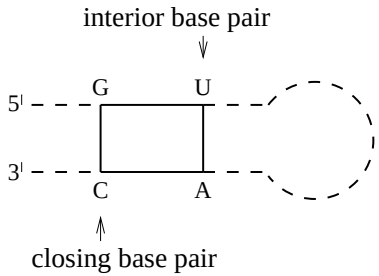
Secondary structure



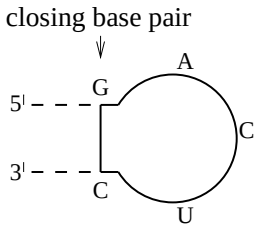
Tertiary structure



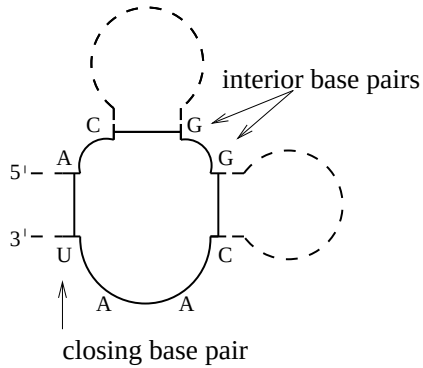
Secondary structures can be uniquely decomposed into loops



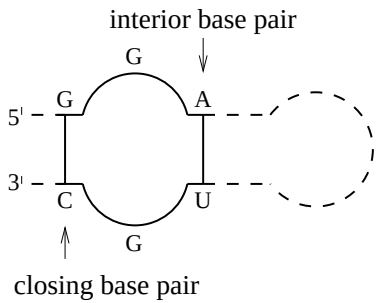
stacking pair



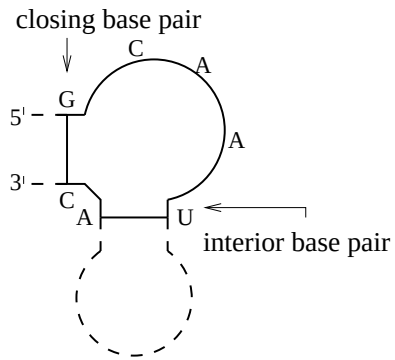
hairpin loop



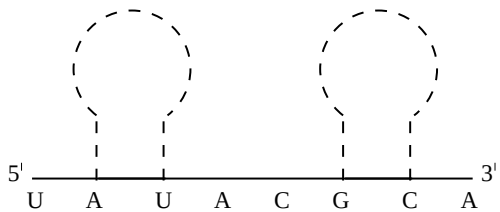
multi loop



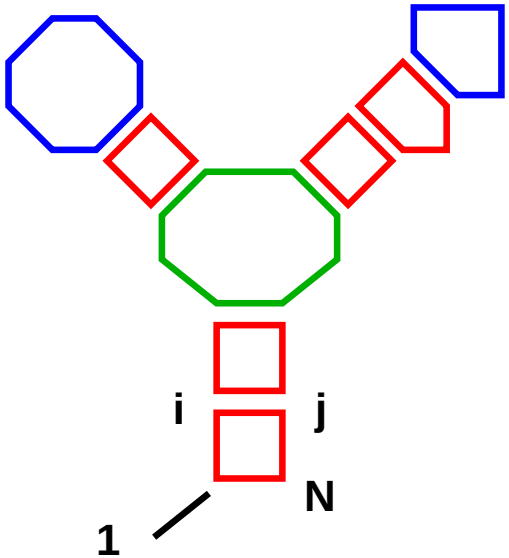
interior loop



bulge



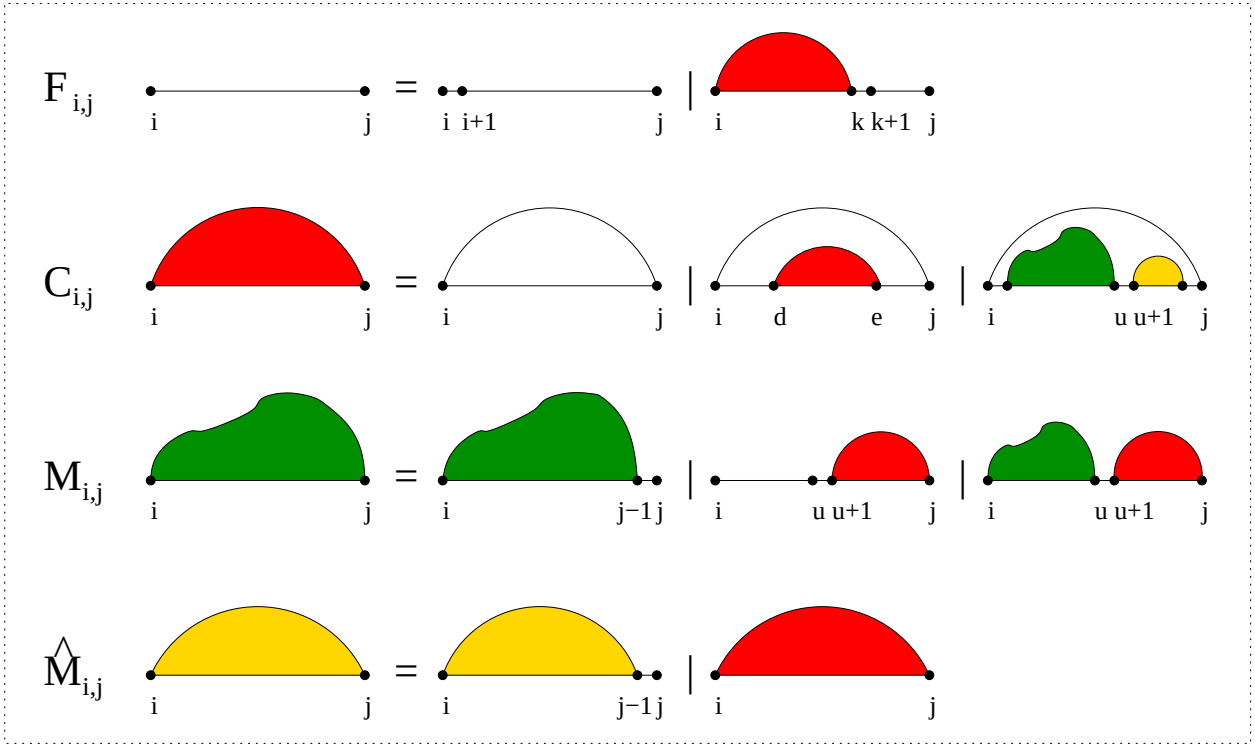
exterior loop



$$E(\mathcal{S}) = \sum_{L \in \mathcal{S}} E(L)$$

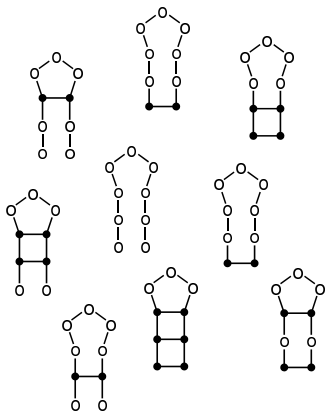
- The free energy of a secondary structure is the sum of the free energy of the loops its composed of
- Loop energies depend on loop type, loop size and sequence
- Energy parameters are measured experimentally or extrapolated by mathematical models

MFE dynamic programming algorithm

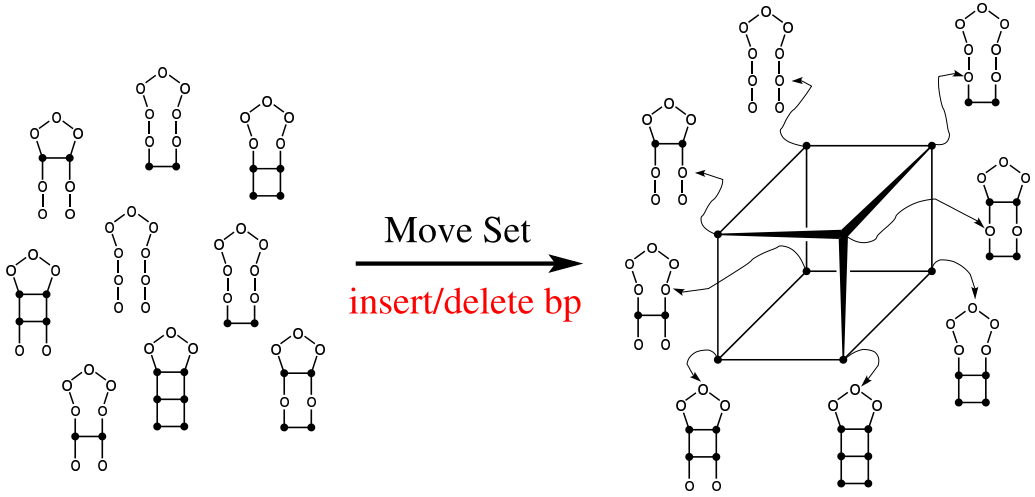


$\mathcal{O}(n^3)$ in time, $\mathcal{O}(n^2)$ in memory

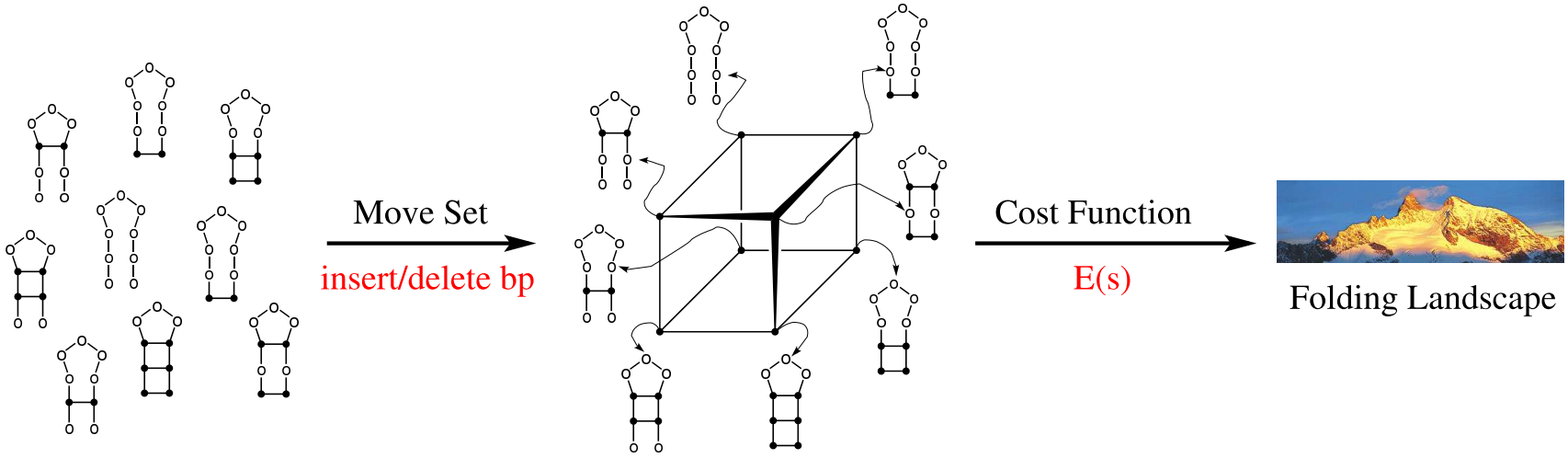
RNA free energy landscape



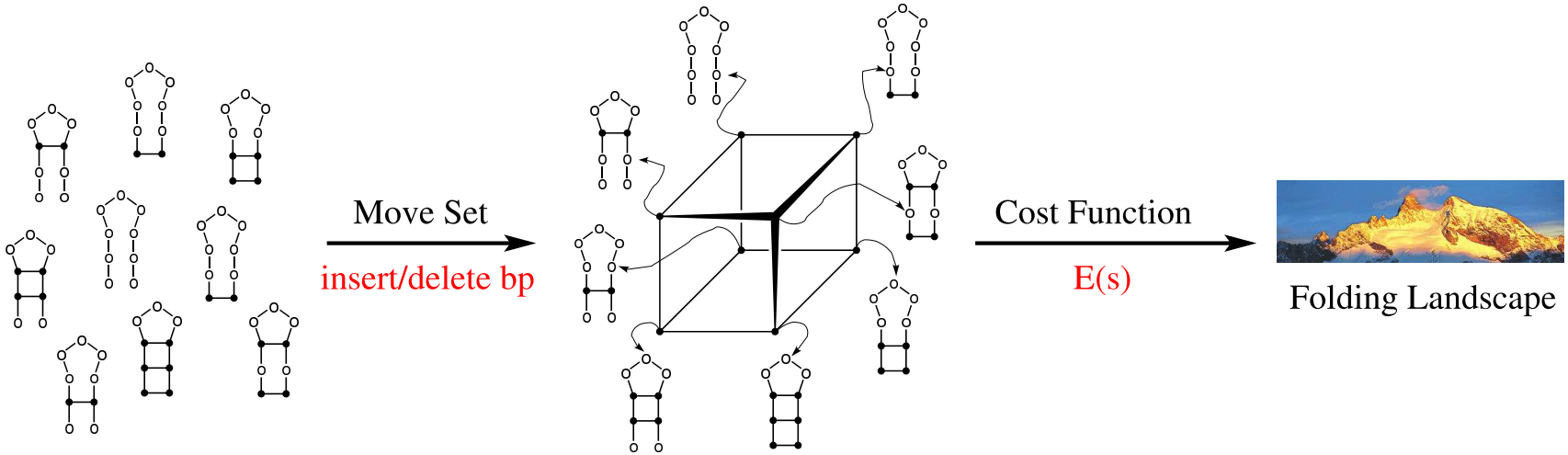
RNA free energy landscape



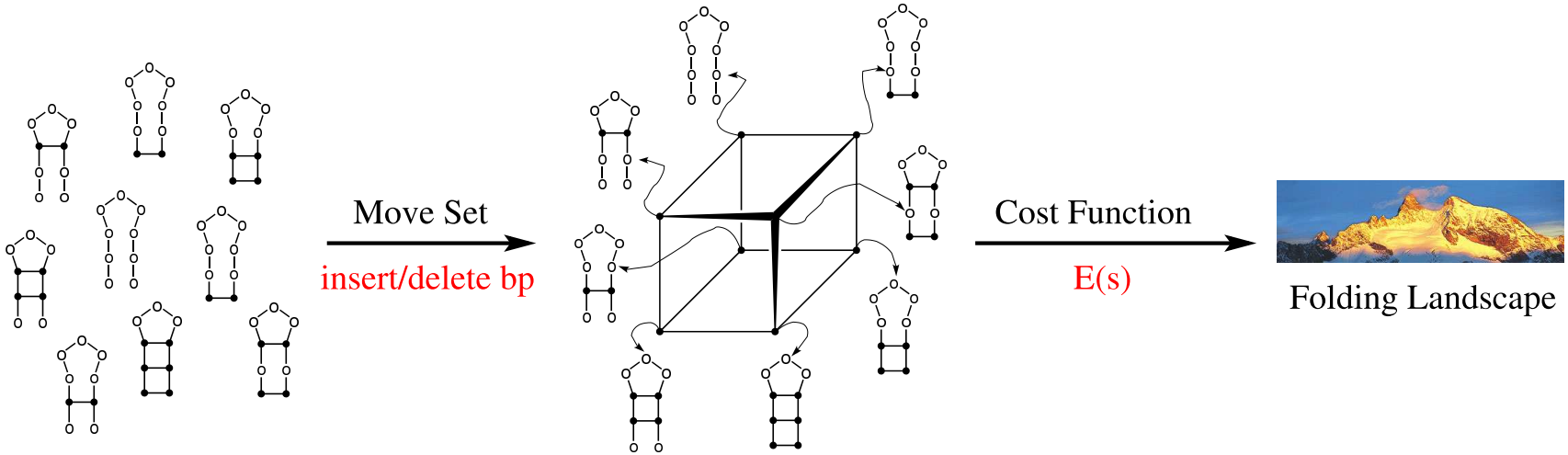
RNA free energy landscape

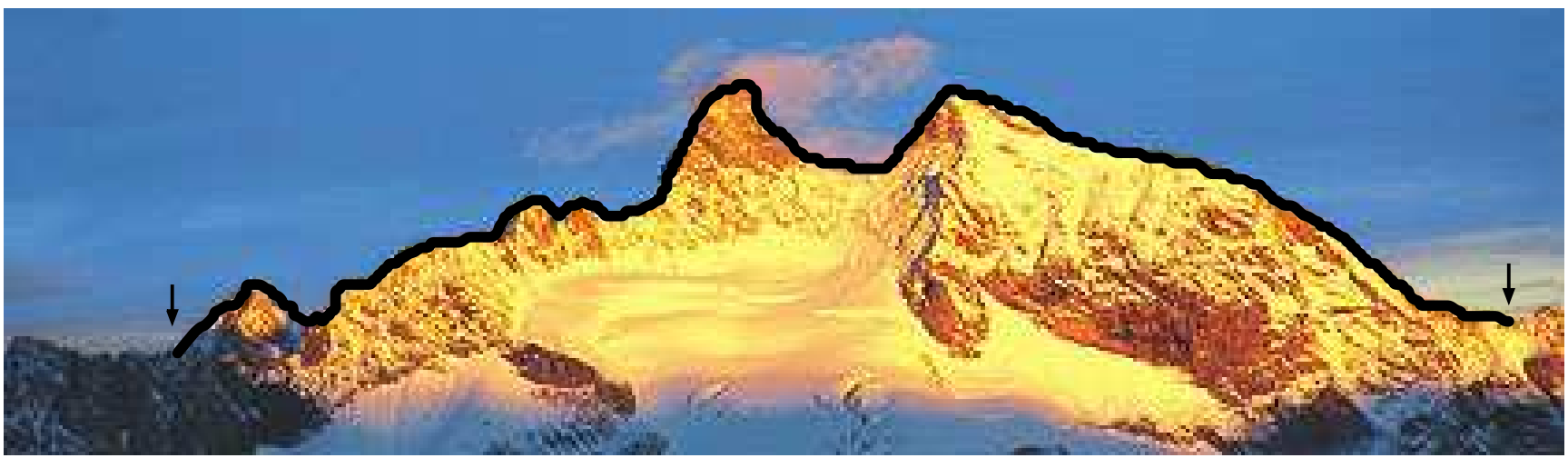


RNA free energy landscape

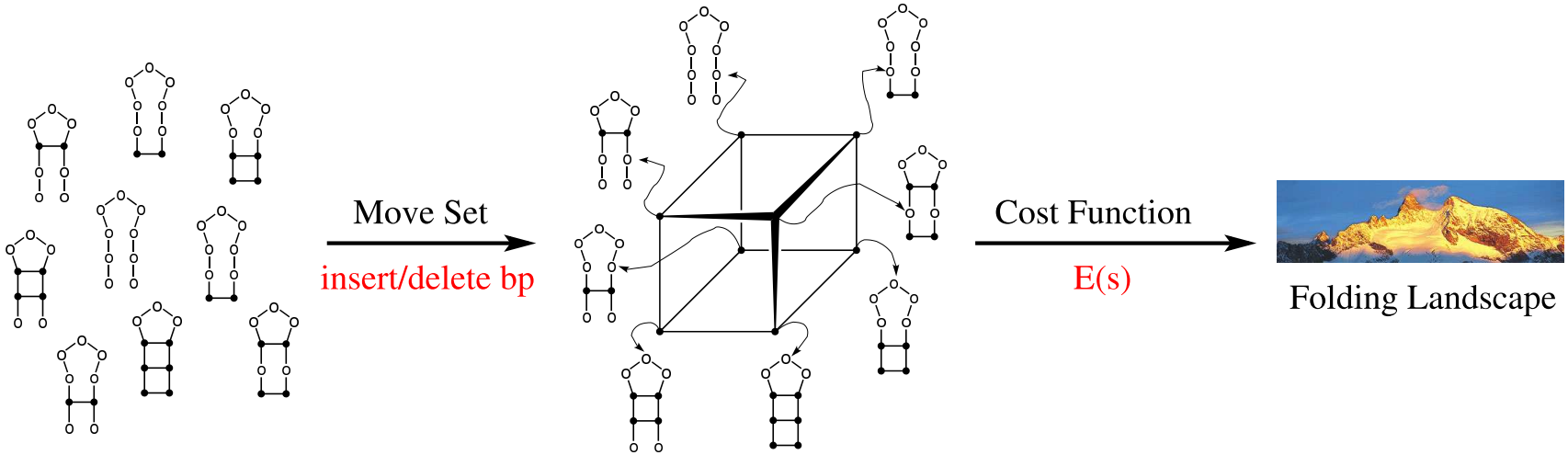


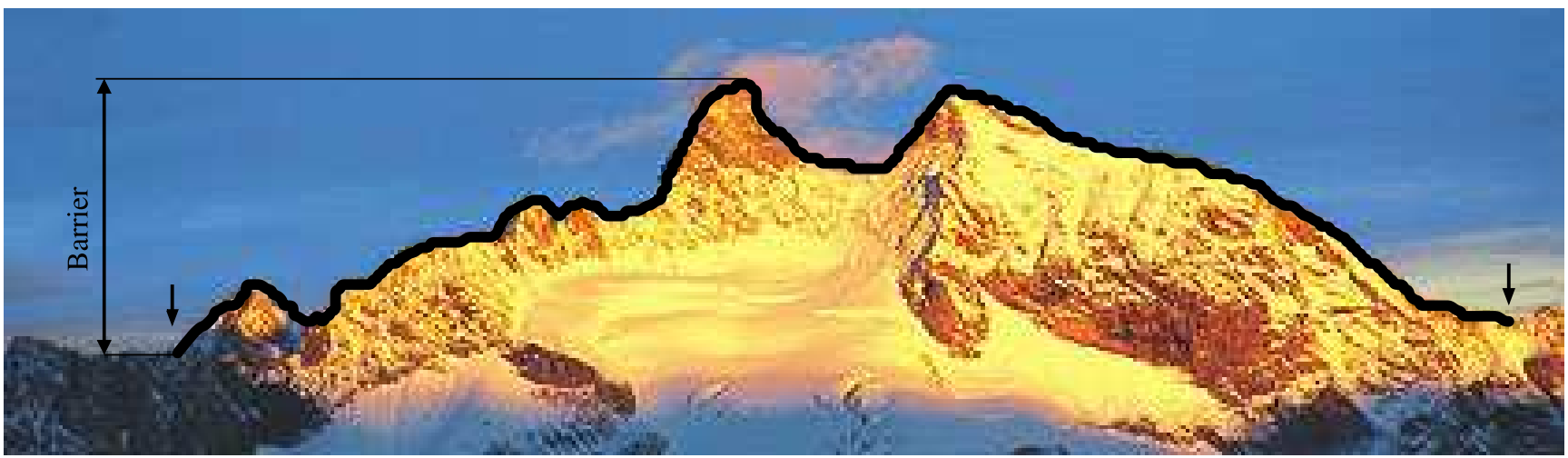
RNA free energy landscape



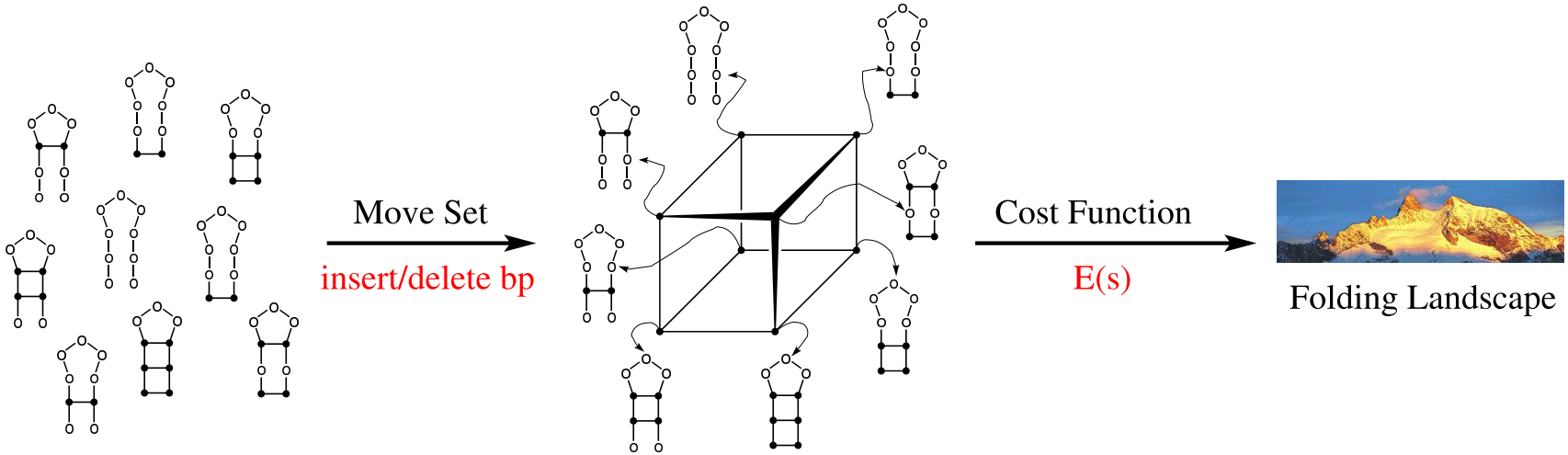


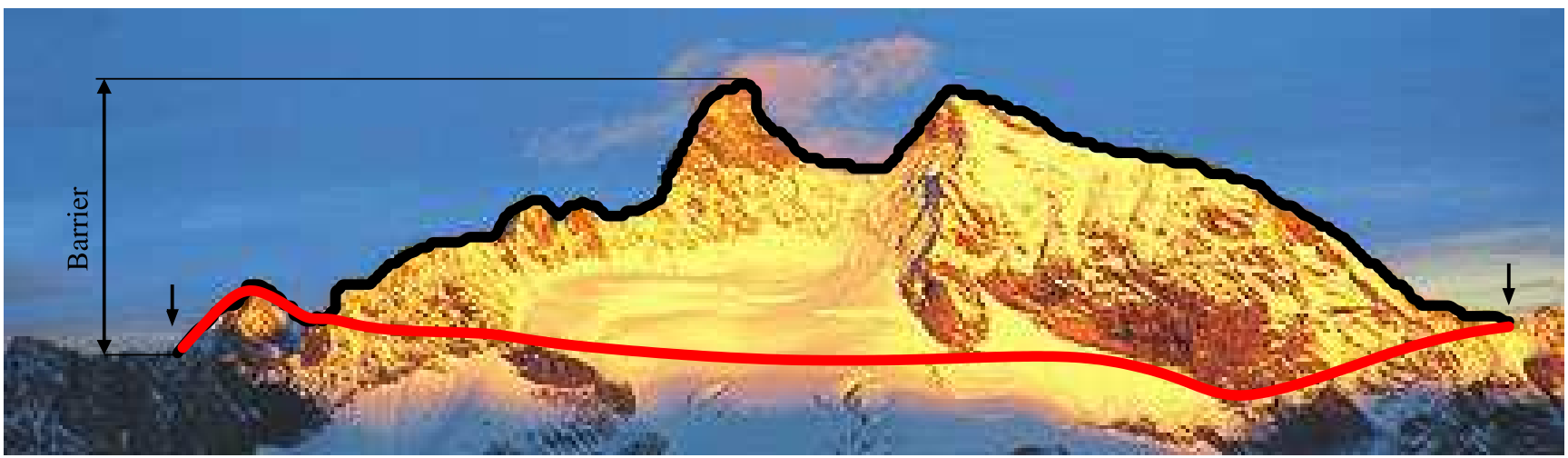
RNA free energy landscape



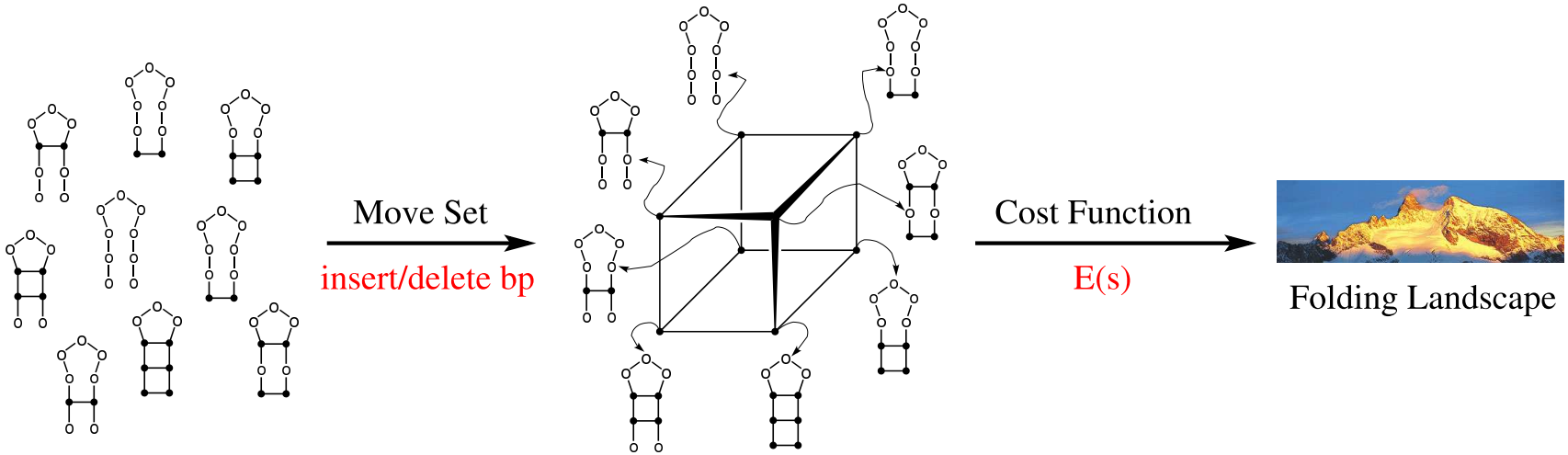


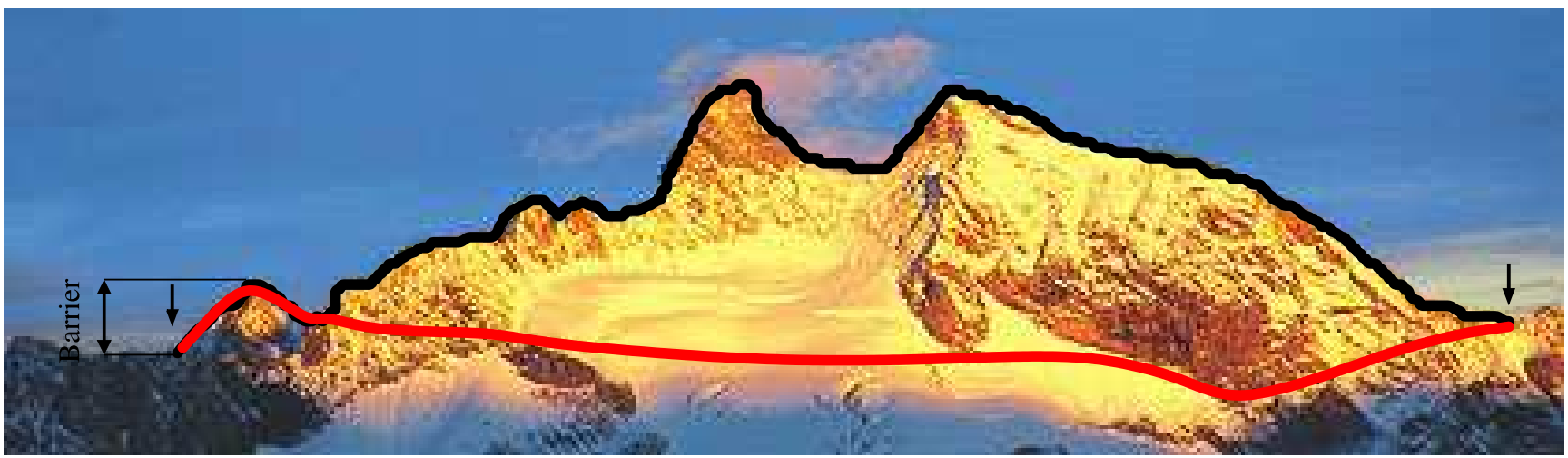
RNA free energy landscape



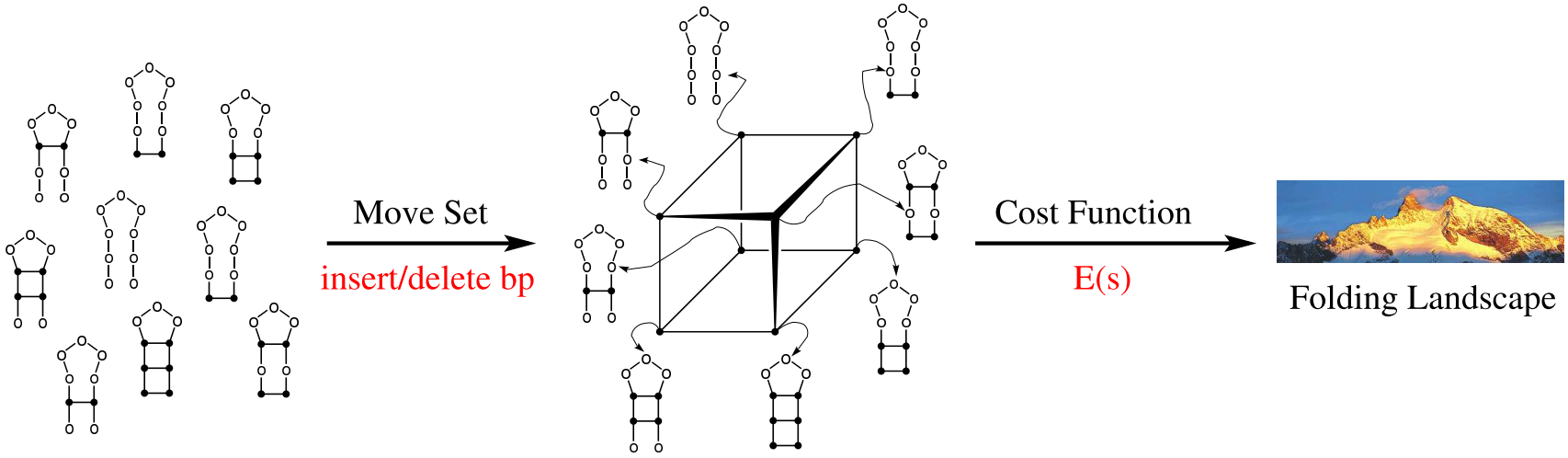


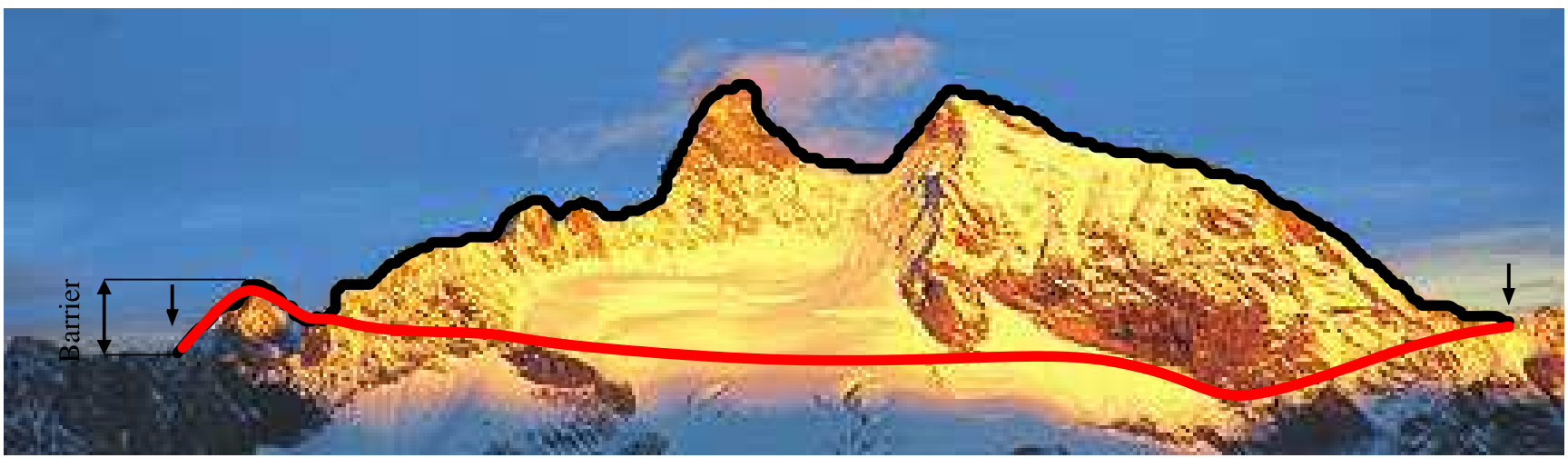
RNA free energy landscape





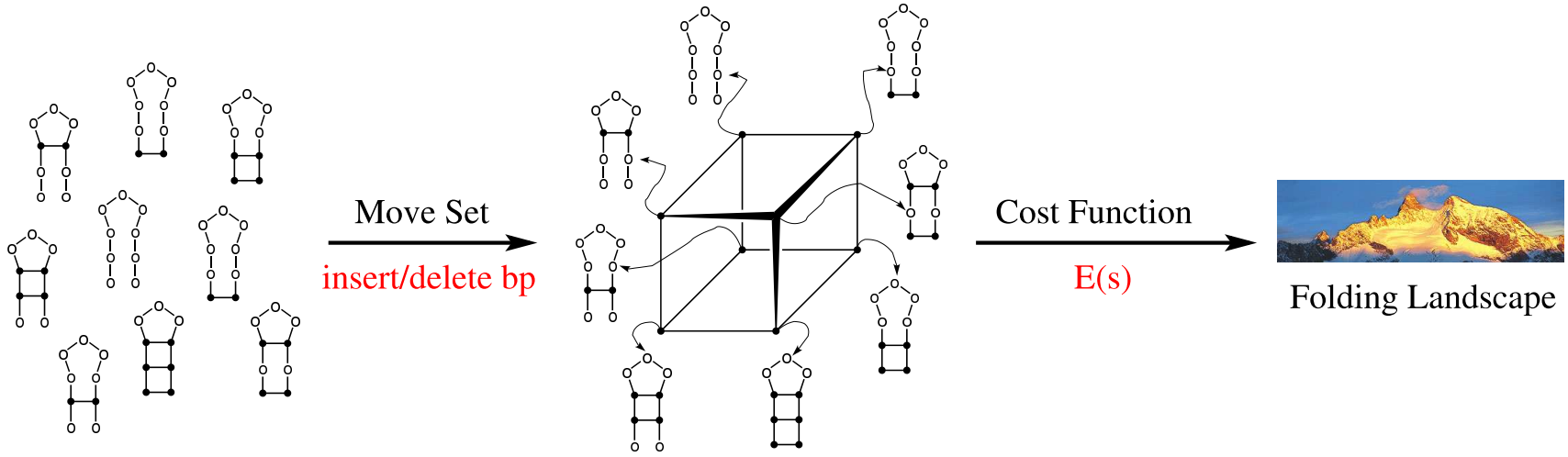
RNA free energy landscape

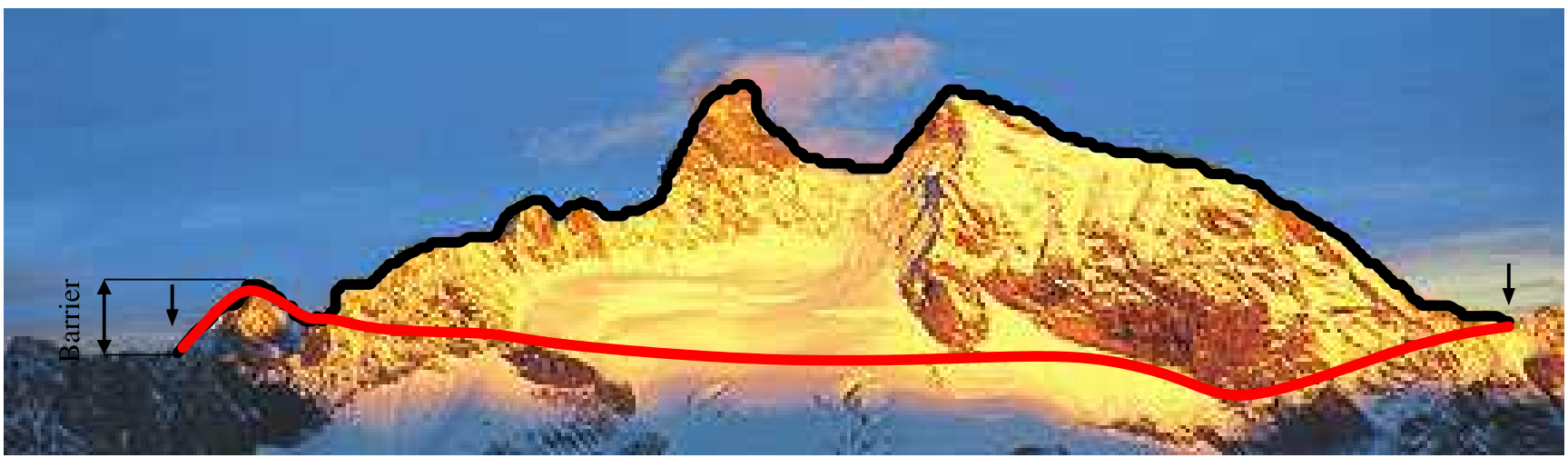




Free energy landscape is a high-dimensional irregular space

RNA free energy landscape





Does a projection into a low dimensional space with few states reveal insights of the high-dimensional irregular landscape?

κ, λ - **neighborhood**

- fixed reference structures s_1 and s_2
- s is a κ, λ - neighbor $\Leftrightarrow d_{BP}(s, s_1) = \kappa \wedge d_{BP}(s, s_2) = \lambda$
- partitioning of the landscape into κ, λ - neighborhoods

κ, λ - neighborhood

- fixed reference structures s_1 and s_2
- s is a κ, λ - neighbor $\Leftrightarrow d_{BP}(s, s_1) = \kappa \wedge d_{BP}(s, s_2) = \lambda$
- partitioning of the landscape into κ, λ - neighborhoods

Questions:

- What are the states with certain free energy (e.g. MFE)?
- What is the partition function?
- How many states?

κ, λ - neighborhood

- fixed reference structures s_1 and s_2
- s is a κ, λ - neighbor $\Leftrightarrow d_{BP}(s, s_1) = \kappa \wedge d_{BP}(s, s_2) = \lambda$
- partitioning of the landscape into κ, λ - neighborhoods

Questions:

- What are the states with certain free energy (e.g. MFE)?
- What is the partition function?
- How many states?

Classified dynamic programming

- Density of states Cupal, J et al. (1996) *Computer Science and Biology* **96**: 184-186
- RNAshapes Steffen, P et al. (2006) *Bioinformatics* **22**(4): 500-503
- RNAbor Freyhult, E et al. (2007) *Bioinformatics* **23**(16): 2054-2062

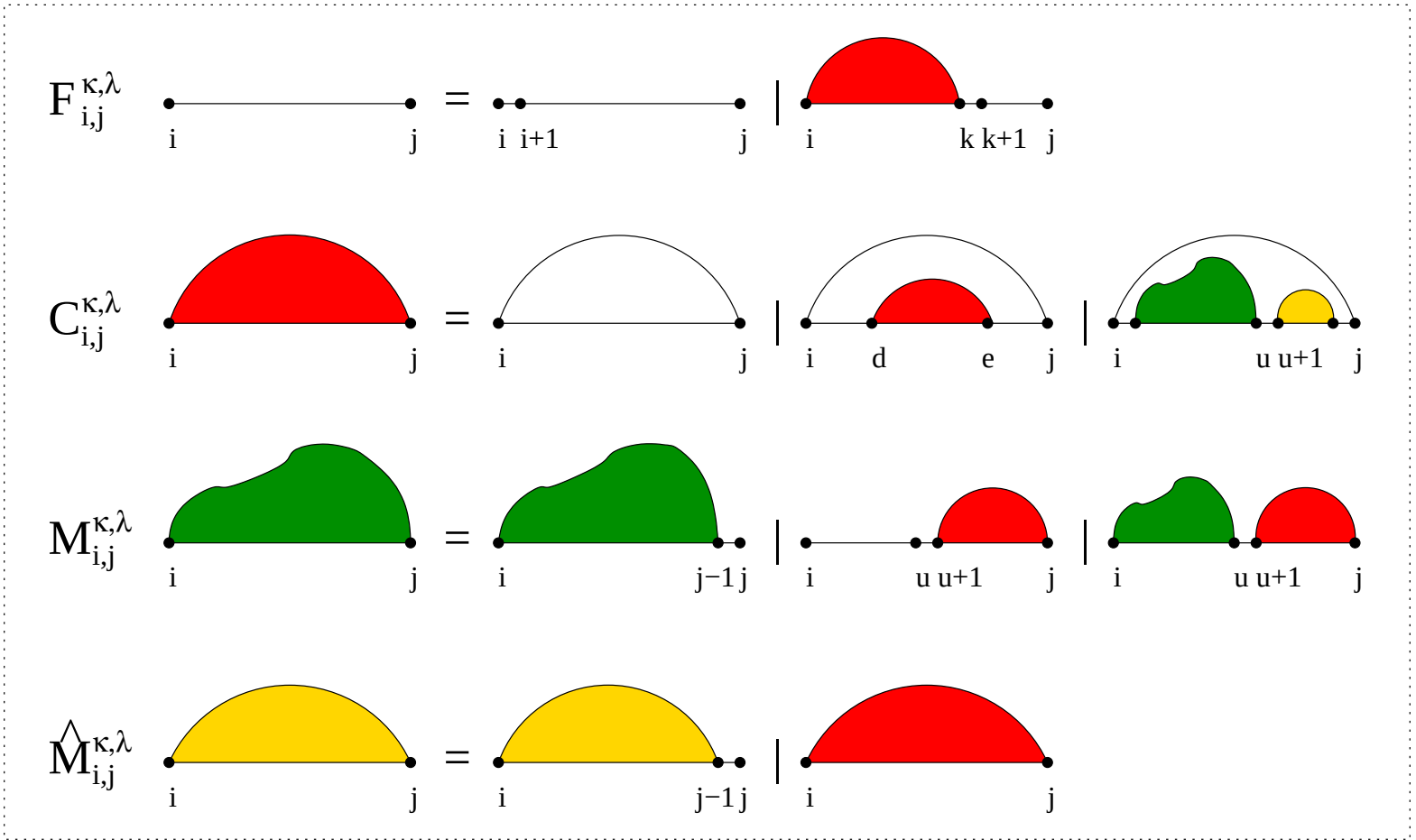
Structure decomposition scheme for κ, λ-neighbors

$$
\begin{aligned}
F_{i,j}^{\kappa,\lambda} &= \min \left\{ \begin{array}{l} F_{i,j-1}^{\kappa-\delta_1^1(i,j),\lambda-\delta_1^2(i,j)}, \\ \min\limits_{i\le u<j} \min\limits_{\substack{\omega_1+\hat{\omega}_1=\kappa-\delta_2^1(i,j,u) \\ \omega_2+\hat{\omega}_2=\lambda-\delta_2^2(i,j,u)}} F_{i,u-1}^{\omega_1,\omega_2} + C_{u,j}^{\hat{\omega}_1,\hat{\omega}_2} \end{array} \right. \\
C_{i,j}^{\kappa,\lambda} &= \min \left\{ \begin{array}{l} \mathfrak{H}(i,j,\kappa,\lambda), \\ \min\limits_{i<p<q<j} \left\{ C_{p,q}^{\kappa-\delta_3^1(i,j,p,q),\lambda-\delta_3^2(i,j,p,q)} + \mathcal{I}(i,j,p,q) \right\}, \\ \min\limits_{i<u<j} \min\limits_{\substack{\omega_1+\hat{\omega}_1=\kappa-\delta_4^1(i,j,u) \\ \omega_2+\hat{\omega}_2=\lambda-\delta_4^2(i,j,u)}} \left\{ M_{i+1,u}^{\omega_1,\omega_2} + \hat{M}_{u+1,j-1}^{\hat{\omega}_1,\hat{\omega}_2} + a \right\} \end{array} \right. \\
M_{i,j}^{\kappa,\lambda} &= \min \left\{ \begin{array}{l} M_{i,j}^{\kappa-\delta_1^1(i,j-1),\lambda-\delta_1^2(i,j)} + c \\ \min\limits_{i\le u<j} \left\{ (u-i)\cdot c + C_{u,j}^{\kappa-\delta_5^1(i,j,u),\lambda-\delta_5^2(i,j,u)} + b \right\}, \\ \min\limits_{i\le u<j} \min\limits_{\substack{\omega_1+\hat{\omega}_1=\kappa-\delta_2^1(i,j,u) \\ \omega_2+\hat{\omega}_2=\lambda-\delta_2^2(i,j,u)}} \left\{ M_{i,u-1}^{\omega_1,\omega_2} + C_{u,j}^{\hat{\omega}_1,\hat{\omega}_2} + b \right\}, \end{array} \right. \\
\hat{M}_{i,j}^{\kappa,\lambda} &= \min \left\{ \begin{array}{l} C_{i,j}^{\kappa,\lambda} + b \\ \hat{M}_{i,j-1}^{\kappa-\delta_1^1(i,j),\lambda-\delta_1^2(i,j)} + c, \end{array} \right.
\end{aligned}
$$

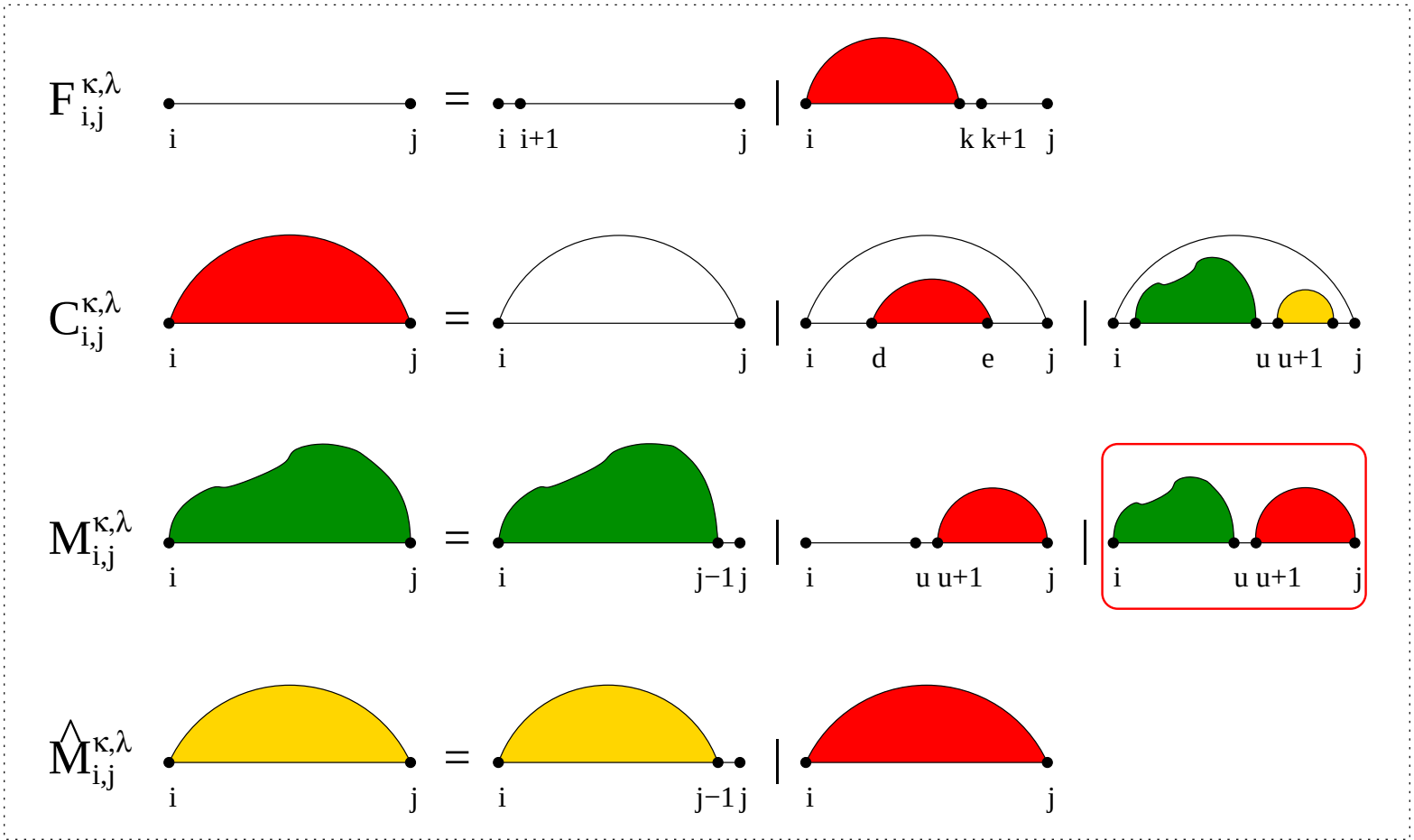
with:

$$
\begin{aligned}
\delta_1^x(i,j) &= d_{\mathrm{BP}}(s_x[i,j], s_x[i,j-1]) \\
\delta_2^x(i,j,u) &= d_{\mathrm{BP}}(s_x[i,j], s_x[i,u-1] \cup s_x[u,j]) \\
\delta_3^x(i,j,p,q) &= d_{\mathrm{BP}}(s_x[i,j], \{(i,j)\} \cup s_x[p,q]) \\
\delta_4^x(i,j,u) &= d_{\mathrm{BP}}(s_x[i,j], \{(i,j)\} \cup s_x[i+1,u] \cup s_x[u+1,j-1]) \\
\delta_5^x(i,j,u) &= d_{\mathrm{BP}}(s_x[i,j], s_x[u,j])
\end{aligned}
$$

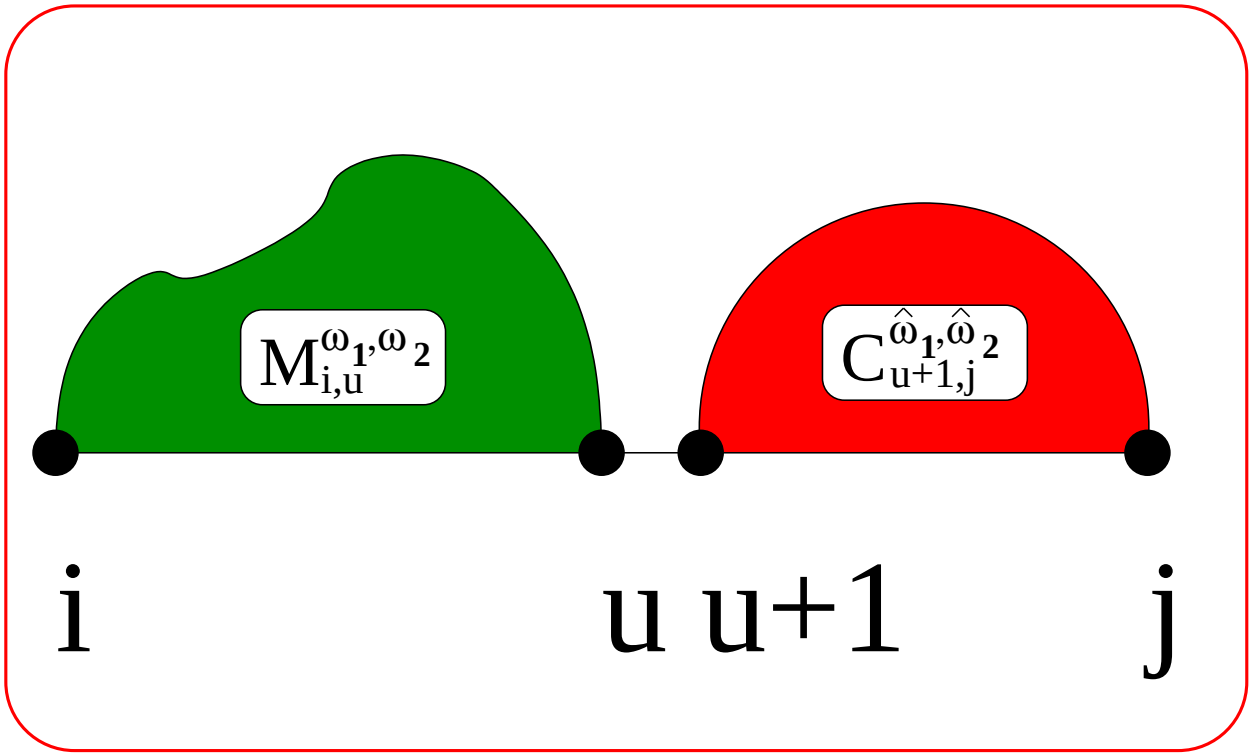
Structure decomposition scheme for κ, λ-neighbors



Structure decomposition scheme for κ, λ-neighbors



Example: multi-loop decomposition



$$\omega_1 + \hat{\omega}_1 = \kappa - d_{\mathrm{BP}}(s_1[i,j], s_1[i,u] \cup s_1[u+1,j])$$
$$\omega_2 + \hat{\omega}_2 = \lambda - d_{\mathrm{BP}}(s_2[i,j], s_2[i,u] \cup s_2[u+1,j])$$

Complexity

- $\mathcal{O}(n^7)$ in time
- $\mathcal{O}(n^4)$ in memory

Complexity

- $\mathcal{O}(n^7)$ in time
- $\mathcal{O}(n^4)$ in memory

Implementation

Naïve:

- more than 3.5h runtime for sequence of length $100nt$
- more than 30GB RAM for sequences with length $> 200nt$

Complexity

- $\mathcal{O}(n^7)$ in time
- $\mathcal{O}(n^4)$ in memory

Implementation

Naïve:

- more than 3.5h runtime for sequence of length $100nt$
- more than 30GB RAM for sequences with length $> 200nt$

This does not look promising :(

Complexity

- $\mathcal{O}(n^7)$ in time
- $\mathcal{O}(n^4)$ in memory

Implementation

Naïve:

- more than 3.5h runtime for sequence of length $100nt$
- more than 30GB RAM for sequences with length $> 200nt$

BUT: Sparse matrices (less than 2% entries with values $\neq \infty$)

Complexity

- $\mathcal{O}(n^7)$ in time
- $\mathcal{O}(n^4)$ in memory

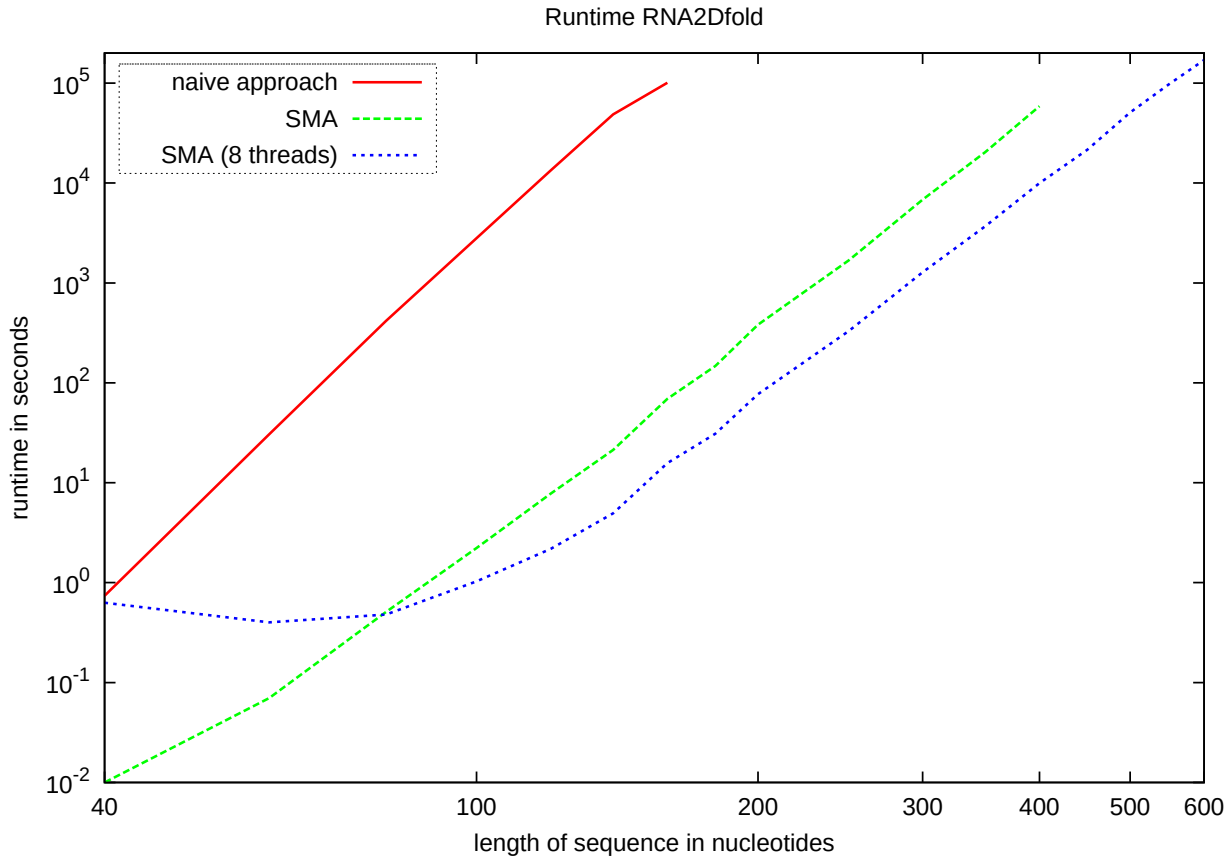
Implementation

Naïve:

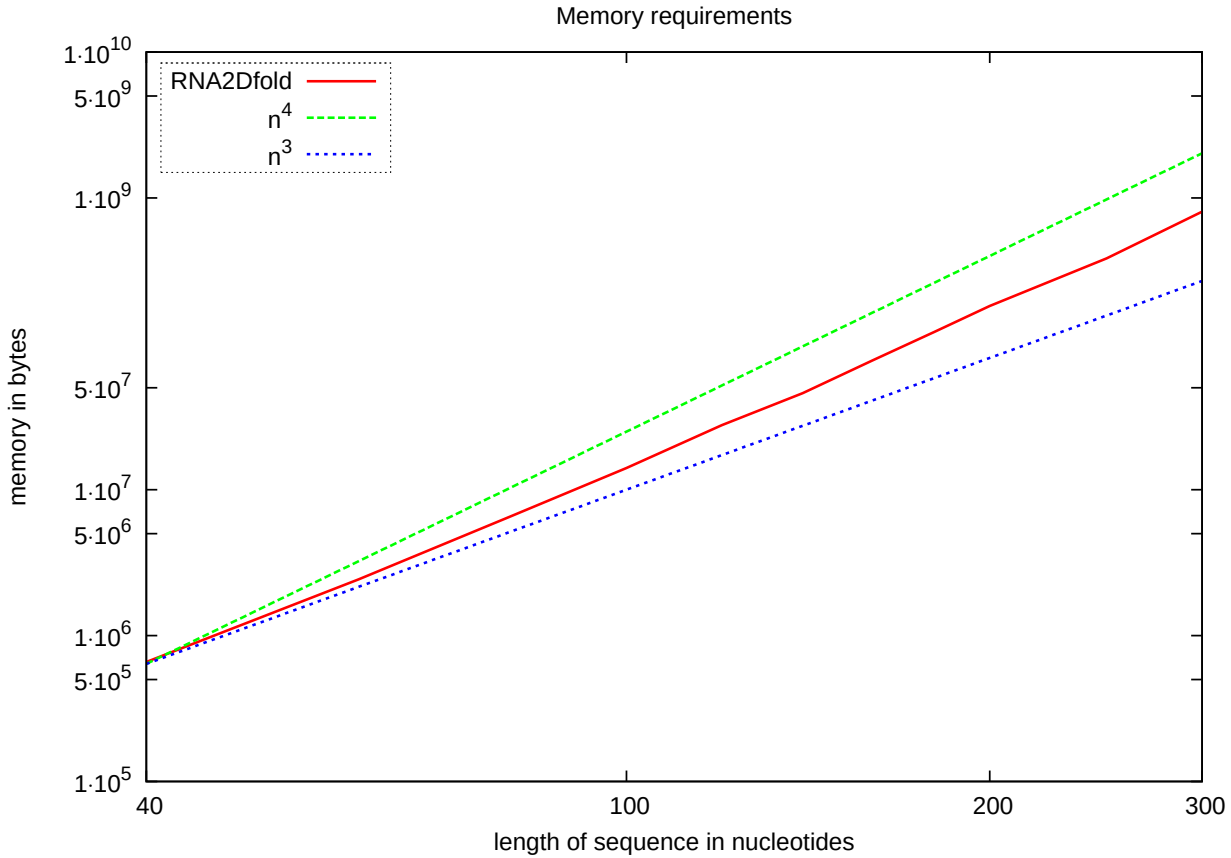
- more than 3.5h runtime for sequence of length $100nt$
- more than 30GB RAM for sequences with length $> 200nt$

`RNA2Dfold`

- exploits sparse matrices and parallel computation of diagonal elements with `OpenMP`
- MFE computation
- Partition function computation
- Boltzmann sampling



Runtimes for dual quad-core Intel® Xeon® E5450 @3.00GHz, 32GB RAM


Memory requirements
RNA2Dfold
n^4
n^3
memory in bytes
1·10^10
5·10^9
1·10^9
5·10^7
1·10^7
5·10^6
1·10^6
5·10^5
1·10^5
40
100
200
300
length of sequence in nucleotides

Output of RNA2Dfold

```
GGGCGCGGUUCGCCCUCCGCUAAAUGCGGAAGAUAAAUUGUGUCU
(((((.....)))))(((((.....)))))(((((.....)))))  (-18.20) <mfe>
(((((.....)))))(((((.....)))))(((((.....)))))  (-18.20) <ref 1>
.............................................  (  0.00) <ref 2>
free energy of ensemble = -19.32 kcal/mol
```

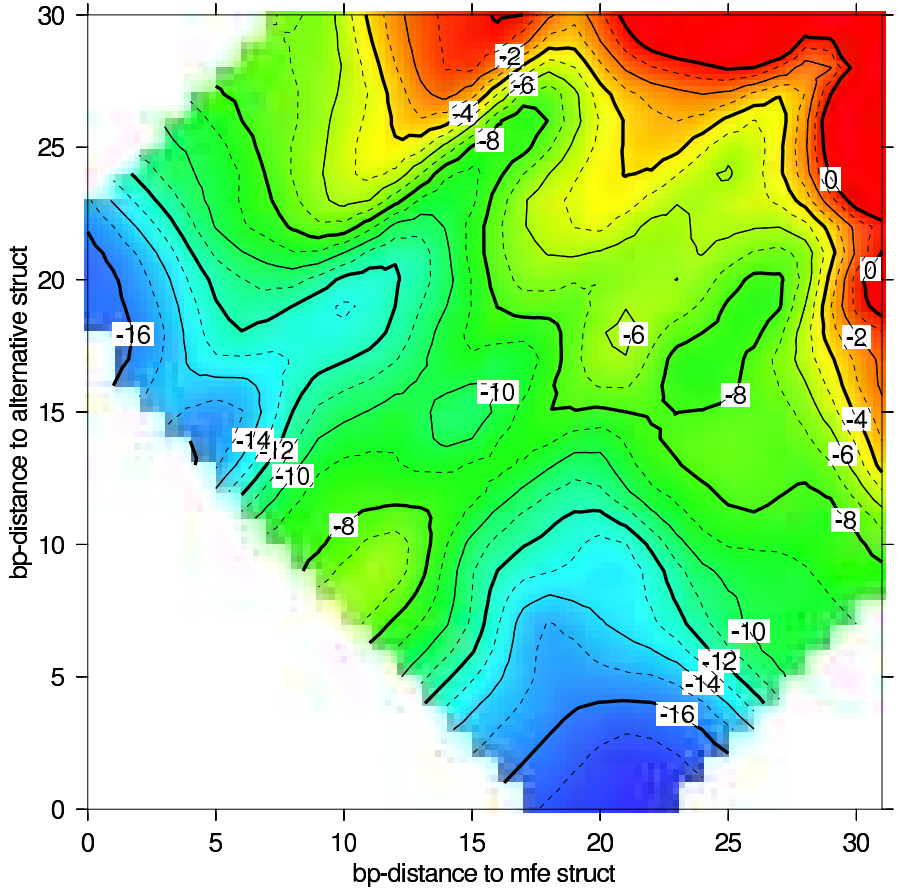
k	l	P(nb)	P(struct in nb)	P(structure)	E_min	E_gibbs	structure
0	15	0.16350376	1.00000000	0.16350396	-18.20	-18.20	(((((.....)))))(((((.....)))))(((((.....)))))
1	14	0.13710099	0.52985725	0.07264404	-17.70	-18.09	(((((.....)))))(((((.....))))).(((.....))).
2	13	0.03870938	0.31496460	0.01219207	-16.60	-17.31	(((((.....)))))(((((.....))))).(((.....))).
2	15	0.00043959	0.78124709	0.00034343	-14.40	-14.55	(((((.....)))))(((((.....)))))(((.(....)))))
3	12	0.00549664	0.31651335	0.00173976	-15.40	-16.11	((((.....))))(((((.....))))).(((.....))).
3	14	0.00050399	0.30274973	0.00015258	-13.90	-14.64	(((((.....)))))(((((.....))))).(((.(....))).
4	11	0.00055642	0.23315097	0.00012973	-13.80	-14.70	((((.....)))).(((.....)))).((((.....))).
4	13	0.00013334	0.19204623	0.00002561	-12.80	-13.82	(((((.....)))))(((((.....)))).)(((.....))).
4	15	0.00001102	0.74654403	0.00000822	-12.10	-12.28	(((((.....)))))(((((.....)))))((((...))..)))
5	10	0.01691144	0.99730503	0.01686584	-16.80	-16.80	((((.....))))(((((.....)))))..............
5	12	0.00001727	0.21163330	0.00000365	-11.60	-12.56	((((.....))))(((((.....))))).(((.....))).
5	14	0.00001787	0.24052012	0.00000430	-11.70	-12.58	(((((.....)))))(((((.....)))))(((.....))..))
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
20	11	0.00001320	0.20013684	0.00000264	-11.40	-12.39	(((((((....)))...(((.....)))..........))))))
20	13	0.00459594	0.37854283	0.00173976	-15.40	-16.00	.(((((((((.....(((.....))))).....))))))))).
20	15	0.18615427	0.39023612	0.07264404	-17.70	-18.28	((((((((((.....(((.....)))))).....))))))))))
21	6	0.00000000	0.11205340	0.00000000	-6.10	-7.45	..((((((........)))).....)).....................
21	8	0.00000000	0.27124082	0.00000000	-5.90	-6.70	.(((((((........))(......).............))))).
21	10	0.00000041	0.34501497	0.00000014	-9.60	-10.26	(((((((........)))(.....)).............))))))
21	12	0.00003261	0.25218449	0.00000822	-12.10	-12.95	.(((((((((.....(((.....)))).....)))))))))).
21	14	0.00907139	0.36701076	0.00332929	-15.80	-16.42	((((((((((.....(((.....)))).....))))))))))
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
27	16	0.00000011	0.42869149	0.00000005	-8.90	-9.42	((((((((((.(..((.(((.....)).).)).).))))))))))
28	13	0.00000063	0.51237067	0.00000032	-10.10	-10.51	((((((((((.((.....)).....(...)..))))))))))
28	15	0.00000000	0.31993086	0.00000000	-5.60	-6.30	((((((((((.(..((.(((.......).).)).).))))))))))
29	14	0.00000024	0.16525529	0.00000004	-8.80	-9.91	((((((((((..(.(.(........).)).)..)...))))))))
29	16	0.00000000	0.50536309	0.00000000	-3.80	-4.22	((((((((((.(.(.(((.....)).).).).).))))))))))
30	15	0.00000015	0.43442428	0.00000006	-9.10	-9.61	((((((((((.((.....)).....(((....).))))))))))))
31	16	0.00000000	0.31578542	0.00000000	-5.80	-6.51	((((((((((.((..(((.....)).))..).).))))))))))

Output of RNA2Dfold

```
GGGCGCGGUUCGCCCUCCGCUAAAUGCGGAAGAUAAAUUGUGUCU
(((((.....)))))(((((.....)))))(((((.....)))))  (-18.20) <mfe>
(((((.....)))))(((((.....)))))(((((.....)))))  (-18.20) <ref 1>
.............................................  (  0.00) <ref 2>
free energy of ensemble = -19.32 kcal/mol
```

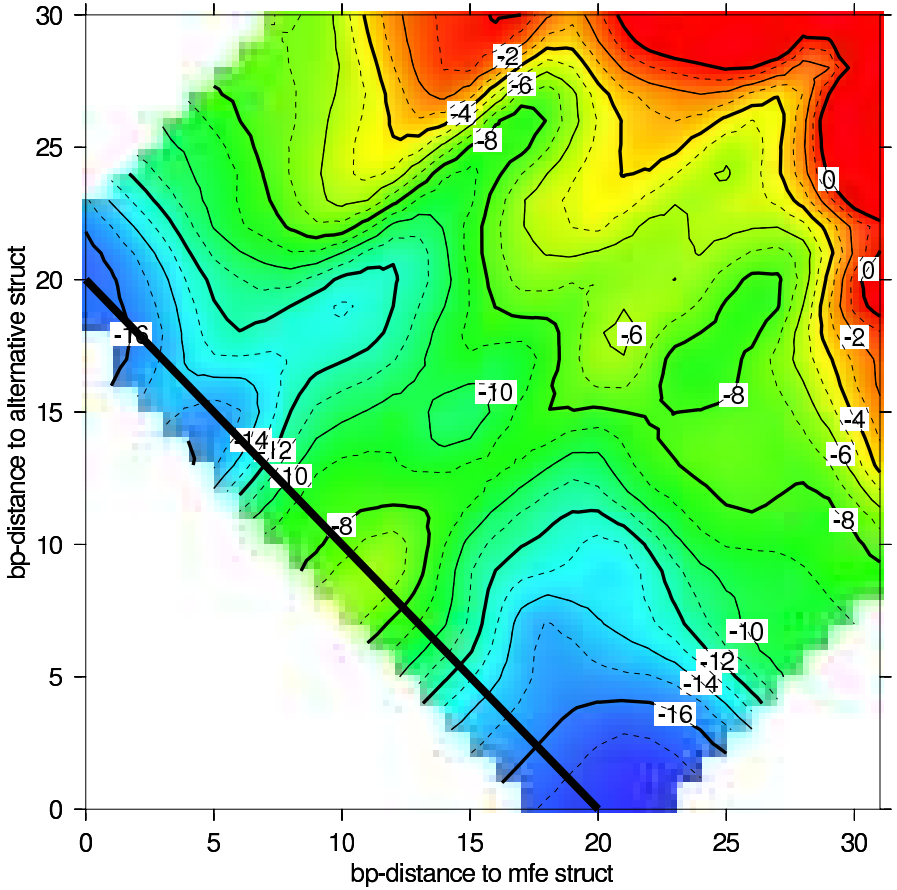
k	l	P(nb)	P(struct in nb)	P(structure)	E_min	E_gibbs	structure
0	15	0.16350376	1.00000000	0.16350396	-18.20	-18.20	(((((.....)))))(((((.....)))))(((((.....)))))
1	14	0.13710099	0.52985725	0.07264404	-17.70	-18.09	(((((.....)))))(((((.....))))).(((.....))).
2	13	0.03870938	0.31496460	0.01219207	-16.60	-17.31	(((((.....)))))(((((.....))))).(((.....))).
2	15	0.00043959	0.78124709	0.00034343	-14.40	-14.55	(((((.....)))))(((((.....)))))(((.(....)))))
3	12	0.00549664	0.31651335	0.00173976	-15.40	-16.11	((((.......))))(((((.....))))).(((.....))).
3	14	0.00050399	0.30274973	0.00015258	-13.90	-14.64	(((((.....)))))(((((.....))))).(((.(....))).
4	11	0.00055642	0.23315097	0.00012973	-13.80	-14.70	((((.......)))).(((.....)))).(((.......))).
4	13	0.00013334	0.19204623	0.00002561	-12.80	-13.82	(((((.....)))))(((((.....)))).)(((.......))).
4	15	0.00001102	0.74654403	0.00000822	-12.10	-12.28	(((((.....)))))(((((.....)))))((((...))..)))
5	10	0.01691144	0.99730503	0.01686584	-16.80	-16.80	((((((.....))))))(((((.....)))))..............
5	12	0.00001727	0.21163330	0.00000365	-11.60	-12.56	((((.......))))(((((.....))))).)(((.....))).
5	14	0.00001787	0.24052012	0.00000430	-11.70	-12.58	(((((.....)))))(((((.....)))))(((.....))..))
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
20	11	0.00001320	0.20013684	0.00000264	-11.40	-12.39	(((((((....))...(((.....)))..........))))))
20	13	0.00459594	0.37854283	0.00173976	-15.40	-16.00	.(((((((((.....(((.....))))).....))))))))).
20	15	0.18615427	0.39023612	0.07264404	-17.70	-18.28	((((((((((.....(((((.....)))))).....))))))))))
21	6	0.00000000	0.11205340	0.00000000	-6.10	-7.45	..((((((........))))).....)).....
21	8	0.00000000	0.27124082	0.00000000	-5.90	-6.70	.(((((((........)))(.......).....)))))).
21	10	0.00000041	0.34501497	0.00000014	-9.60	-10.26	(((((((........)))(.....)).....))))))
21	12	0.00003261	0.25218449	0.00000822	-12.10	-12.95	.((((((((.....(((.....)))).....)))))))).
21	14	0.00907139	0.36701076	0.00332929	-15.80	-16.42	((((((((((.....(((.....)))).....))))))))))
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
27	16	0.00000011	0.42869149	0.00000005	-8.90	-9.42	((((((((((.(..((.(((.....)).).)).).))))))))))
28	13	0.00000063	0.51237067	0.00000032	-10.10	-10.51	((((((((((.((.....)).....(....)...))))))))))
28	15	0.00000000	0.31993086	0.00000000	-5.60	-6.30	((((((((((.(..((.(((.....).).)).).))))))))))
29	14	0.00000024	0.16525529	0.00000004	-8.80	-9.91	((((((((((..((.(.....).)).)).)...))))))))))
29	16	0.00000000	0.50536309	0.00000000	-3.80	-4.22	((((((((((.(.(.(((.....)).).).).).))))))))))
30	15	0.00000015	0.43442428	0.00000006	-9.10	-9.61	((((((((((.((.....)).....(((....).))))))))))))
31	16	0.00000000	0.31578542	0.00000000	-5.80	-6.51	((((((((((.((..((((.....)).).)..).).))))))))))

Height map of κ, λ-neighborhood MFEs



MFE structure vs. alternative structure

Estimating barrier height

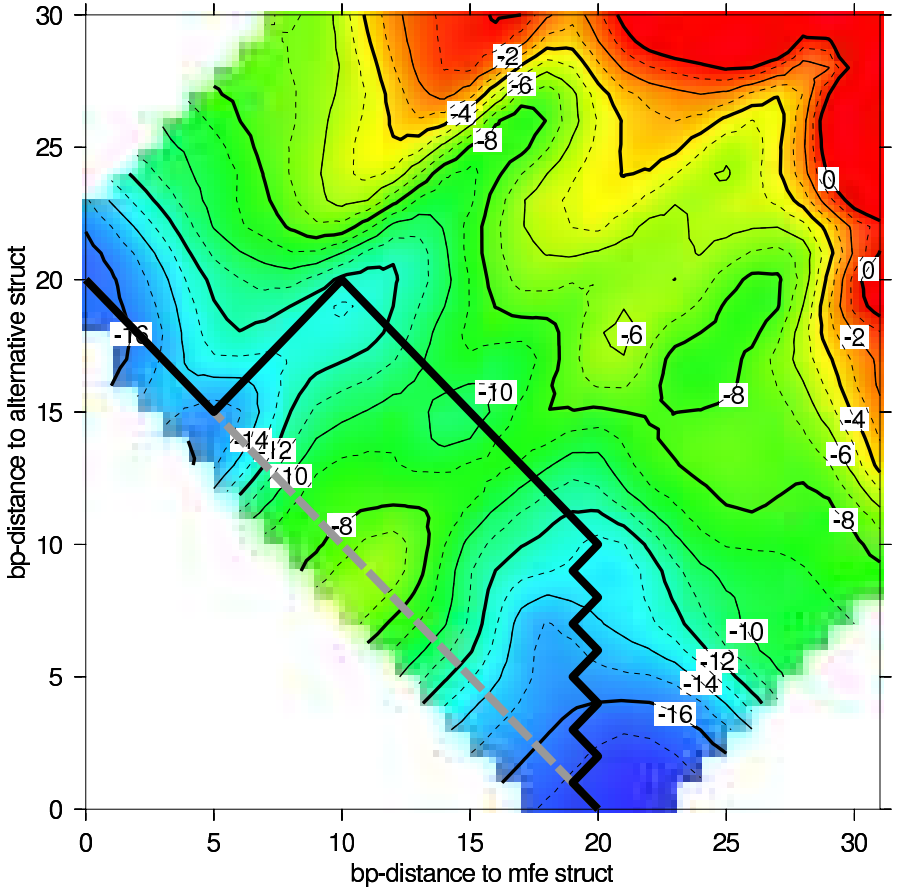


Direct path

best saddle energy ≥ -4.9 kcal/mol

lower bound of barrier ≥ 13.3 kcal/mol

Estimating barrier height



Best path in κ, λ-neighborhood (Ford-Bellman algorithm)

$$\text{best saddle energy} \geq -8.2 \text{ kcal/mol}$$

$$\text{lower bound of barrier} \geq 10.0 \text{ kcal/mol}$$

Finding (energetically) best transition path

- exact solution for small sequences (exhaustive enumeration)
- several heuristics, e.g. Morgan & Higgs '98, Flamm et al. '01
- consider only shortest (direct) paths in most cases
- direct paths (NP hard even for Nussinov energy model)[1]
- indirect paths (NP hard)[1]

Direct paths

- perform poorly when states have great distance
- stabilizing base pairs are neglected
- may estimate barrier too high

[1] Maňuch, J et al. (2009) *Proc. of the 15th Intl. Meeting on DNA Computing and Molecular Programming*

Finding (energetically) best transition path

- exact solution for small sequences (exhaustive enumeration)
- several heuristics, e.g. Morgan & Higgs '98, Flamm et al. '01
- consider only shortest (direct) paths in most cases
- direct paths (NP hard even for Nussinov energy model)[1]
- indirect paths (NP hard)[1]

Direct paths

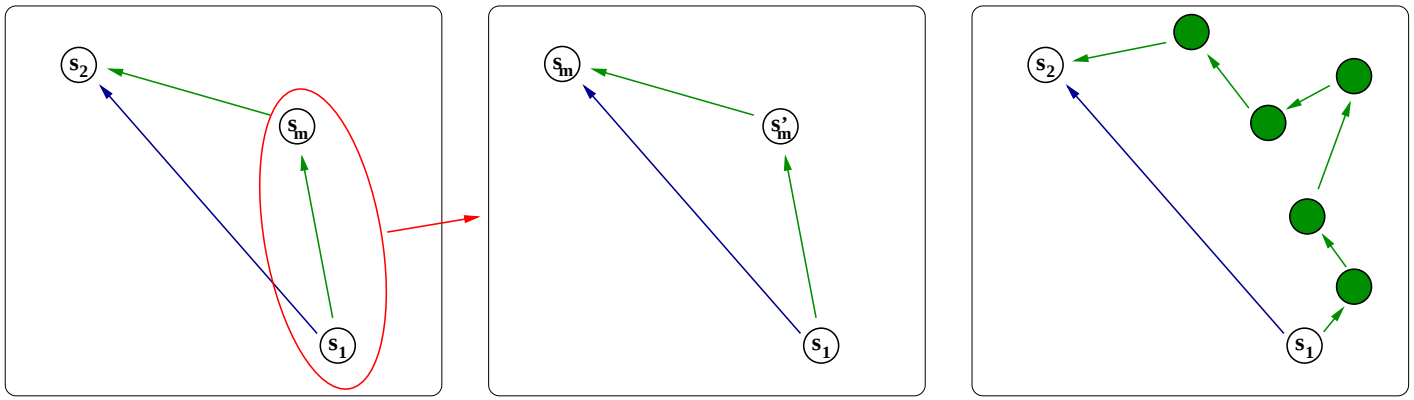
- perform poorly when states have great distance
- stabilizing base pairs are neglected
- may estimate barrier too high

How to exploit the vast amount of general paths?

[1] Maňuch, J et al. (2009) *Proc. of the 15th Intl. Meeting on DNA Computing and Molecular Programming*

Constructing general- from direct-refolding paths

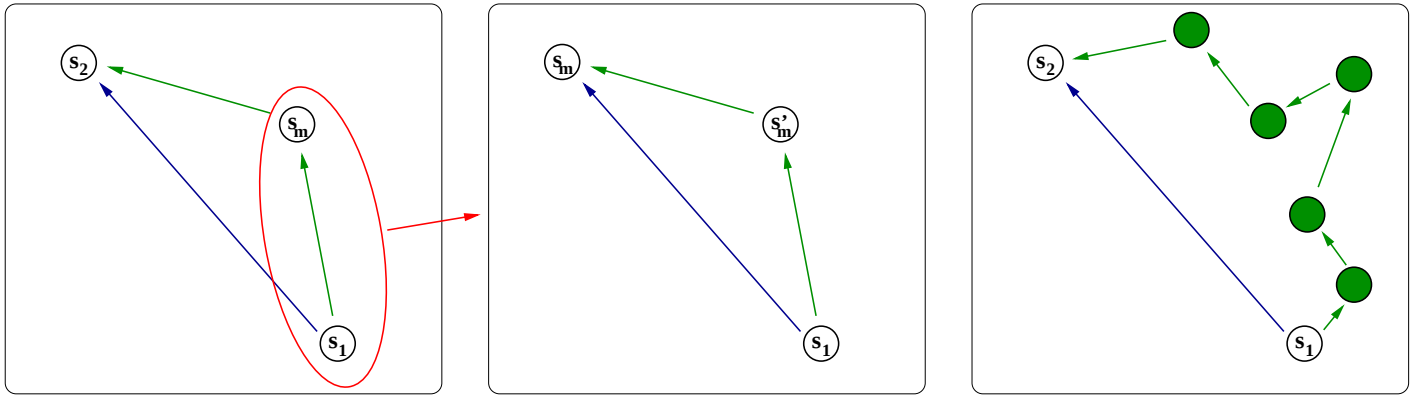
`Pathfinder`



1. generate best direct folding path p_d between s_1 and s_2
2. find x other structures s_{x_m} (meshpoint x)
3. generate best direct paths p_{x_1} from s_1 to s_{x_m} and p_{x_2} from s_{x_m} to s_2
4. concatenate paths $p_x = p_{x_1} + p_{x_2}$
5. accept if $B(p_x) < B(p_d)$
6. if more refinement required, level p_{x_1} and p_{x_2} the same way
7. otherwise return p_x with lowest barrier

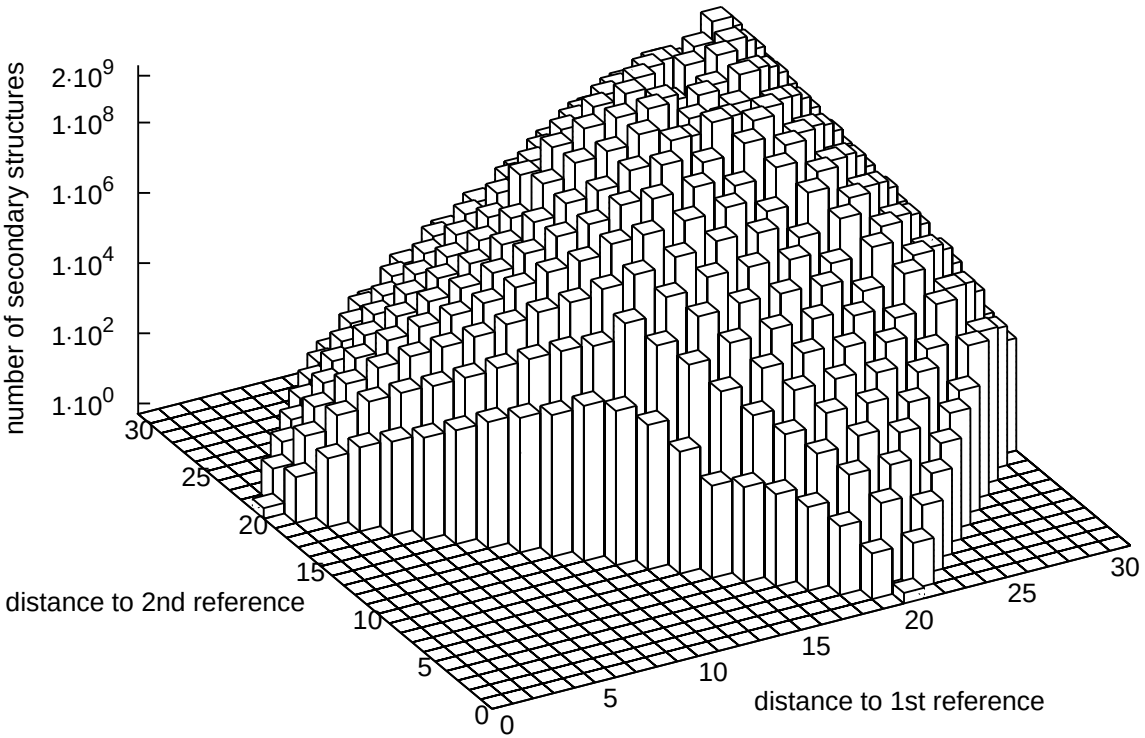
Constructing general- from direct-refolding paths

Pathfinder

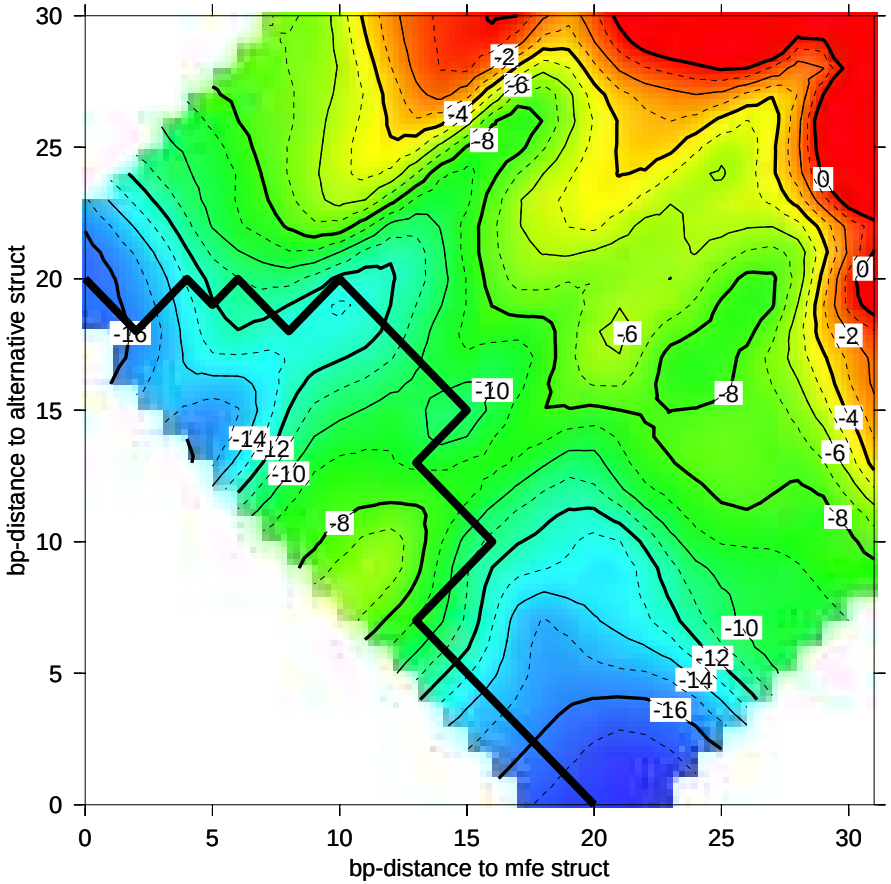


1. generate best direct folding path p_d between s_1 and s_2
2. find x other structures s_{x_m} (meshpoint x, κ, λ MFE representatives)
3. generate best direct paths p_{x_1} from s_1 to s_{x_m} and p_{x_2} from s_{x_m} to s_2
4. concatenate paths $p_x = p_{x_1} + p_{x_2}$
5. accept if $B(p_x) < B(p_d)$
6. if more refinement required, level p_{x_1} and p_{x_2} the same way
7. otherwise return p_x with lowest barrier

Embedding refolding paths



Embedding refolding paths

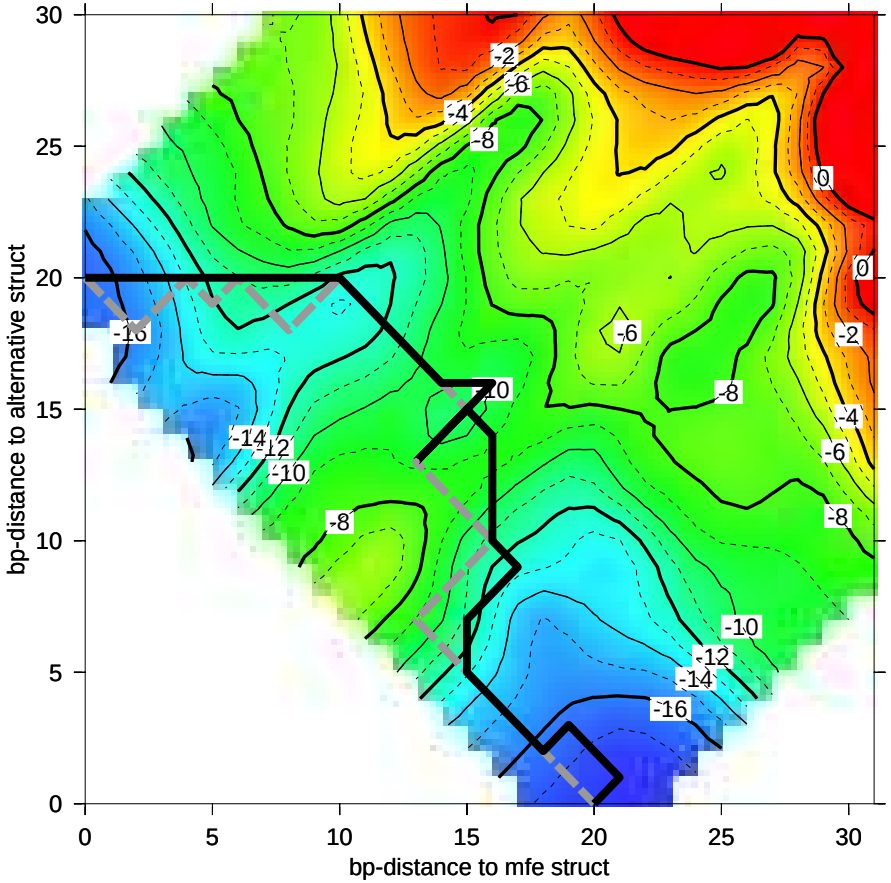


indirect path obtained by the `Pathfinder` algorithm

saddle energy $= -7.5$ kcal/mol

barrier estimate $= 10.7$ kcal/mol

Embedding refolding paths



best path computed by exhaustive enumeration

saddle energy $= -7.5$ kcal/mol

real barrier $= 10.7$ kcal/mol

Conclusion

- 2D projection of high-dimensional folding space into κ, λ-neighborhoods
- Produces qualitative pictures of the energy landscape
- `RNA2Dfold` is fast enough to treat most biologically interesting RNAs by exploiting sparseness and parallelization
- Starting point for the recognition of RNA switches or for coarse grained folding simulations.
- Intermediate node suggestion for indirect folding path construction

Thanks to:

Stephan Bernhart
Christoph Flamm
Andreas Gruber
Christian Höner zu Siederdissen
Ivo Hofacker
Andrea Tanzer

...and You!

`RNA2Dfold` α-release available:
`http://www.tbi.univie.ac.at/~ronny/RNA/`

This work has been funded, in part, by the Austrian GEN-AU projects "bioinformatics integration network III" and "non coding RNA".