RNA FOLDING PATHWAYS

Maximilian Faissner

Introduction

Previous work / algorithms

Outlook

Direct RNA folding pathways

	Structures GGGGAAAACCCCUUUU	Energy (kcal/mol)	Actions
S_1	((((....))))....	-6.60	$\mathtt{del}_{1,12}$
	.(((....)))....	-2.90	$\mathtt{del}_{2,11}$
	..((....))....	0.40	$\mathtt{del}_{3,10}$
	...(....).....	3.70	$\mathtt{del}_{4,9}$
	.....	0.00	$\mathtt{add}_{8,13}$
	.......(....)...	5.50	$\mathtt{add}_{7,14}$
	......((....))..	4.60	$\mathtt{add}_{6,15}$
	.....(((....))).	3.70	$\mathtt{add}_{5,16}$
S_2	....((((....))))	2.80	

Goal: Finding a direct pathway with the lowest energy barrier

direct pathways: number of refolding moves = basepair distance

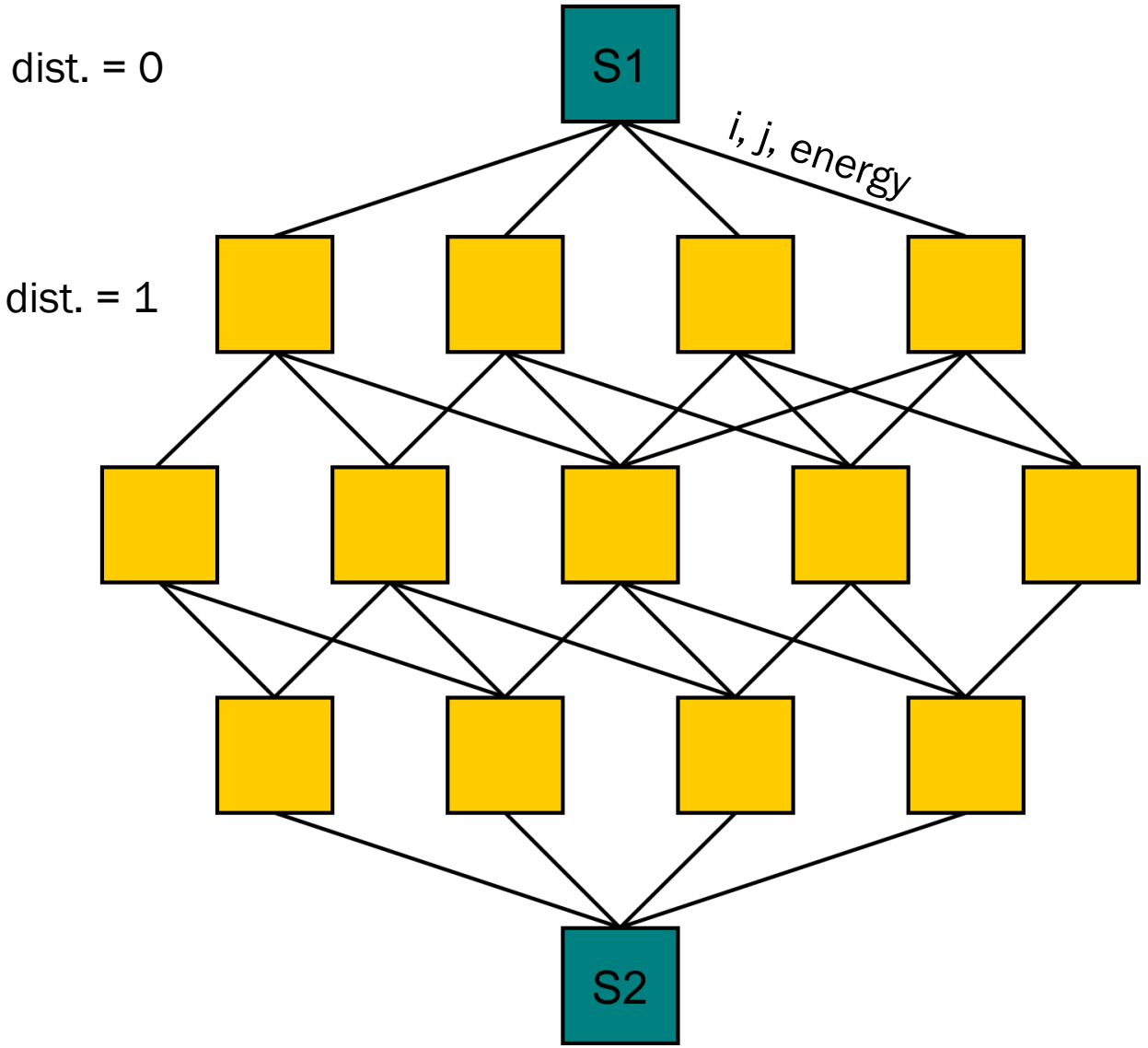
Combinatorial search of the best permutation of moves

NP-hard problem

Findpath:

Bounded BFS heuristic, implemented in the ViennaRNA package (Turner energy model)

Findpath: Bounded Breadth-first search heuristic



Graph structure:

nodes: unique structures

edges: valid moves (i, j) with associated move energy

Bounded look-ahead heuristic:

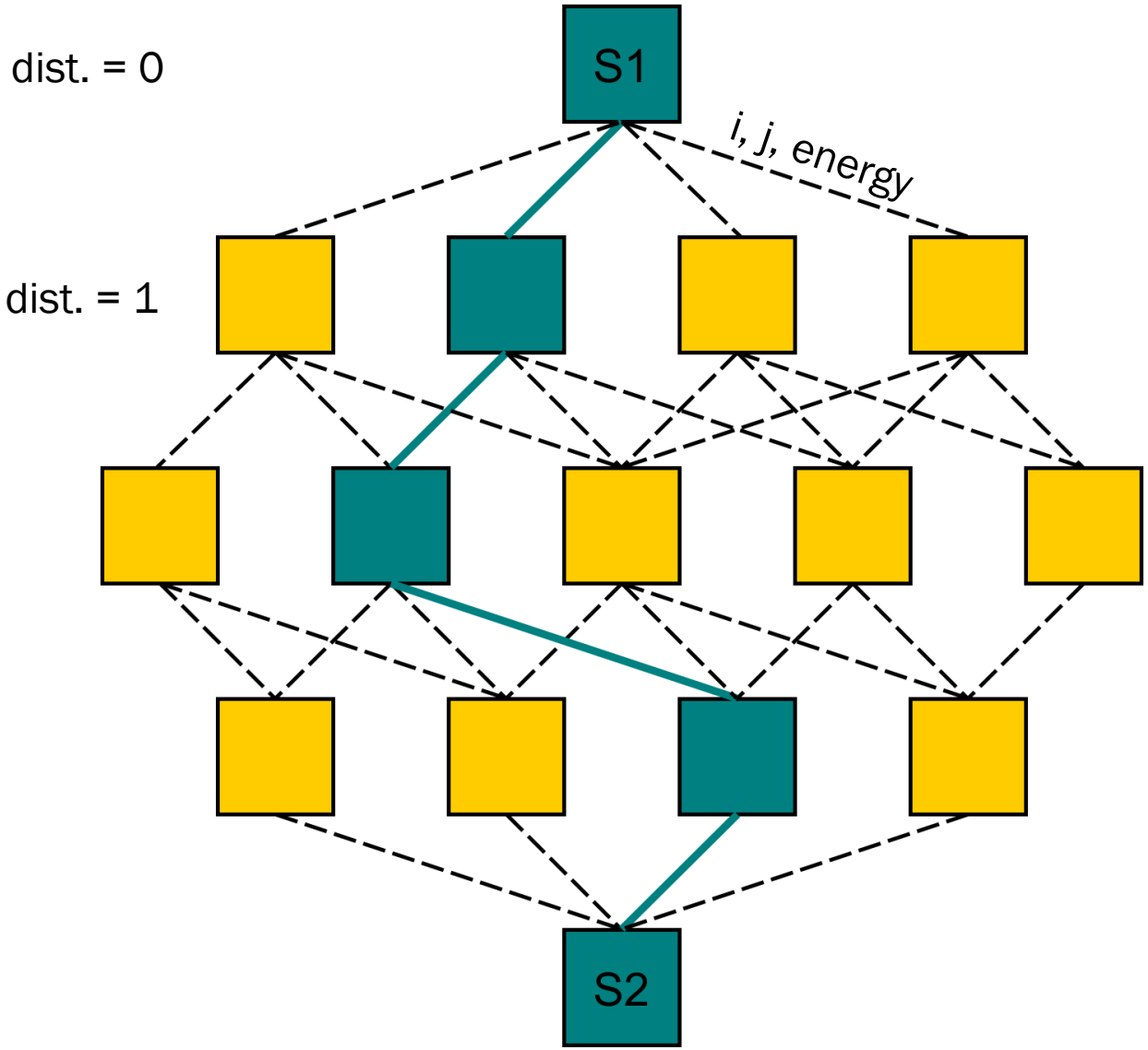
keep the best k candidates after each distance step

Criterion for best candidates:

- lowest saddle energy (so far)
- lowest current node energy

remove duplicates (one unique structure per distance)

Findpath: Increasing the search width *k*



First iteration: *k* = 1

Similar to the greedy approach as proposed by Morgan and Higgs:

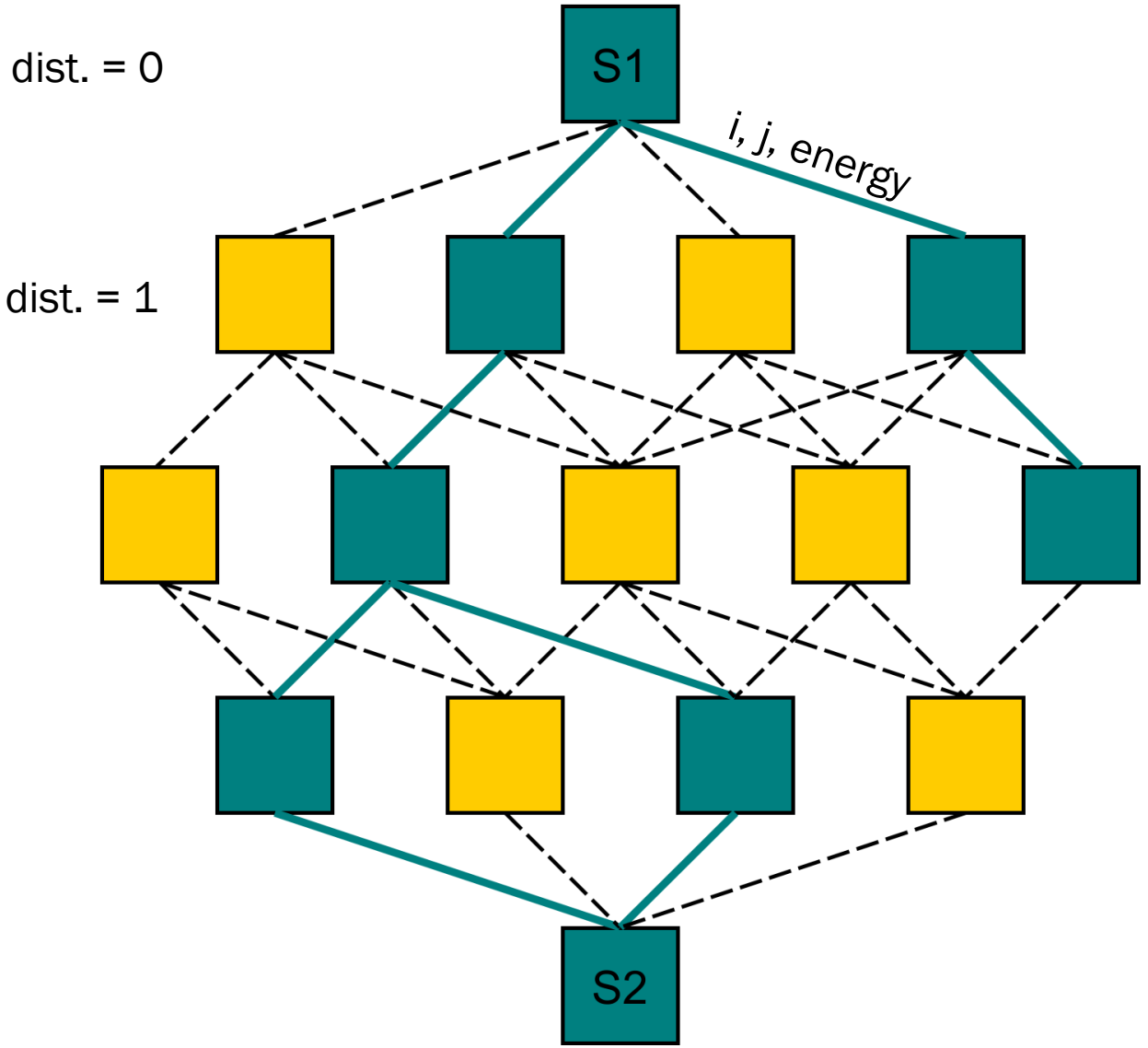
Single next structure with lowest energy

S. R. Morgan and P. G. Higgs. Barrier heights between ground states in a model of RNA secondary structure, J. Phys. A.: Math. Gen., 31, 1998, 3153-3170.

Result:

Path with associated saddle energy (highest energy along the path)

Findpath: Increasing the search width *k*



Second iteration: *k* = 2

Keep 2 candidate paths per iteration

Saddle energy from previous iteration (k = 1) serves as upper energy limit, limiting the search space

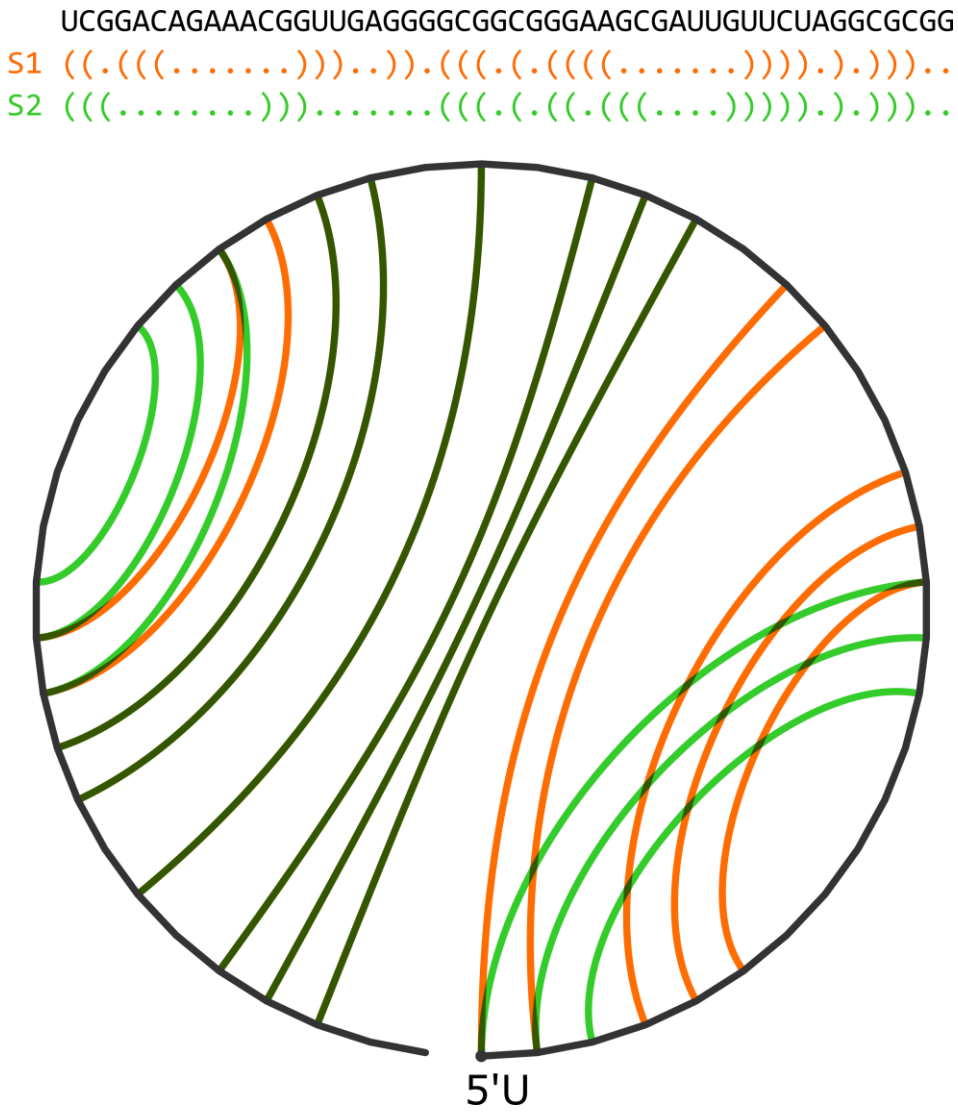
Increase search with *k* as desired

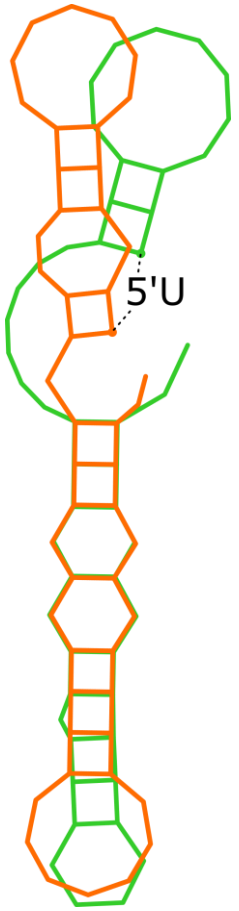
Preferable approach: Final k should depend on the basepair distance

Project Motivation

- Findpath is a computationally heavy heuristic for large sequences
  - *time dependent step for DrTransformer (co-transcriptional folding)*
- Optimization & Improvements for the Findpath heuristic:
  - *Reduce the search space by separating folding pathways into independent sections*
  - *Generate multiple independent folding pathways*
  - *Merge independent pathways recursively*
- Additional point of interest: indirect folding pathways

Divide & Conquer approach for Findpath



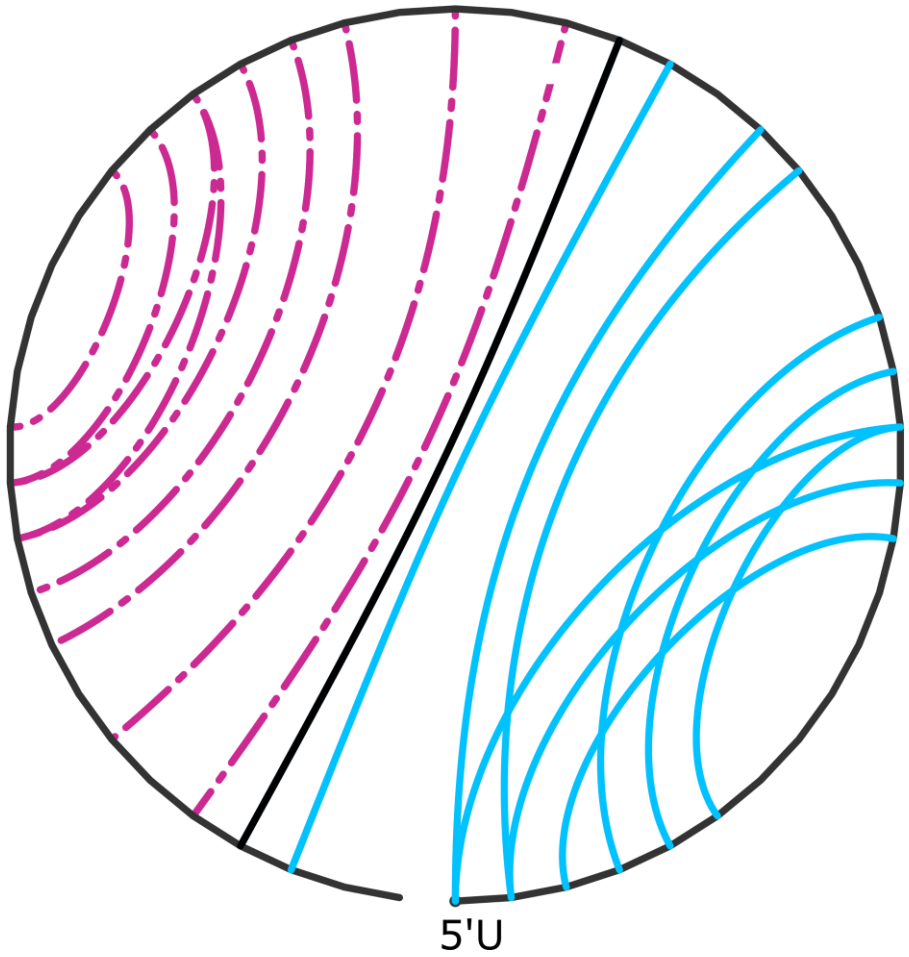


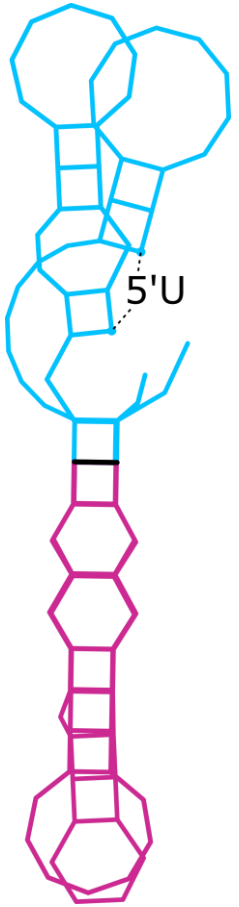
Splitting interior loops:

- Find constant basepairs present in both S_1 and S_2
- Recursively start a new section at a constant basepair
- Reconstruct final path by recursively merging individual paths

Divide & Conquer approach for Findpath

```
    UCGGACAGAAACGGUUGAGGGGCGGCGGGAAGCGAUUGUUCUAGGCGCGG
S1  ((.(((.......))))..)).((.(.((((.......)))).).))..
S2  (((........)))........((.(.((.(((....))))).).))..
```

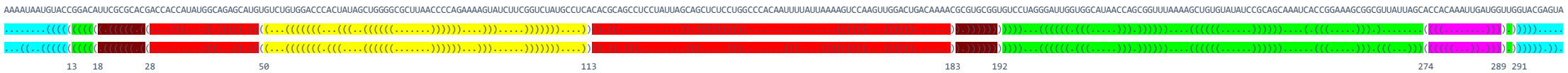




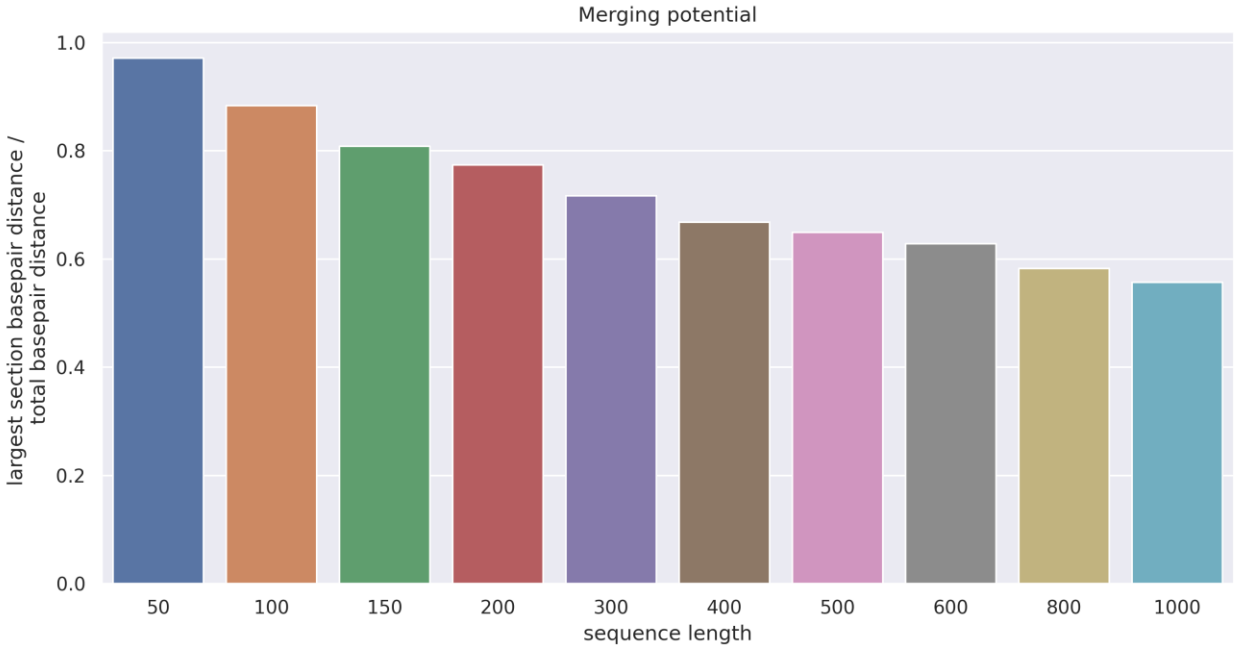

```
    UCGGACAGAAACGGUUGAGGGGCGGCGGGAAGCGAUUGUUCUAGGCGCGG
S1  ((.(((.......))))..)).((.(.((((.......)))).).))..
S2  (((........)))........((.(.((.(((....))))).).))..

    Outer Section
    ((.(((.......))))..)).((.......................)).
    .(.(((.......))))..)..((.......................)).
    ...(((.......)))......((.......................)).
    ....((.......))......((.......................)).
    .....(.......).......((.......................)).
    ......................((.......................)).
    ..(........).........((.......................)).
    .((........))........((.......................)).
    (((........)))........((.......................)).

    Inner Section
    ........................(.....(((.......)))).....)...
    ........................(.....(((.(....)))).....)...
    ........................(.....((..(....).)).....)...
    ........................(.....((.((....)))).....)...
    ........................(.....(..((....)).).....)...
    ........................(.....(.(((....))))).....)...
```

How often can we apply this method?

Finding the largest section



Random 300 nt example with sections: [1, [13, [18, [28, [50, 113], 183], 192], [274, 289], 291], 300]
Total basepair distance: 40 / largest section basepair distance: 23 (red section)



Dataset generation:

- Random RNA sequence strings
- structure pairs: RNAsubopt (Boltzmann weighted samples)
- gradient descent to local minimum
- Structure pairs with a basepair distance < 10 are filtered out

Merging independent paths

First approach: Dynamic Programming

$$E_{i,j} = max\left(\Delta G(a_{i,j}), min\left(E_{i-1,j}, E_{i,j-1}\right)\right)$$

moves outer section

moves inner section

```
[[ 0.  7.1 7.1 7.1 7.1 7.1  9.3 9.3 9.3 9.3 9.3 9.3 9.3 9.3 9.3 9.3 9.3 9.3]
 [ 0.  6.2 6.2 6.2 6.2 6.2  8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4]
 [ 1.3 8.4 8.4 8.4 6.2 7.1 10.6 9.  10. 8.4 9.2 8.7 8.4 8.4 8.9 8.8 8.4 8.4 8.4]
 [ 1.3 7.9 7.9 7.9 6.2 6.6 10.1 9.  9.5 8.4 8.7 8.7 8.4 8.4 8.4 8.4 8.4 8.4 8.4]
 [ 1.3 6.4 6.4 6.4 6.2 6.2  8.6 8.6 8.6 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4 8.4]
 [ 1.3 5.9 5.9 5.9 5.9 5.9  8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1]
 [ 1.3 7.4 7.4 7.4 5.9 6.1  9.6 8.1 9.  8.1 8.2 8.1 8.1 8.1 8.1 8.1 8.1 8.1 8.1]]
```

Dynamic Programming matrix to merge 2 folding pathways

Issues with the initial DP-approach

Findings:

Assuming there is no additional local minimum along the path, the optimal merged path can be constructed with 2 optimal input paths.

Findpath does not necessarily output all required paths with respect for merging

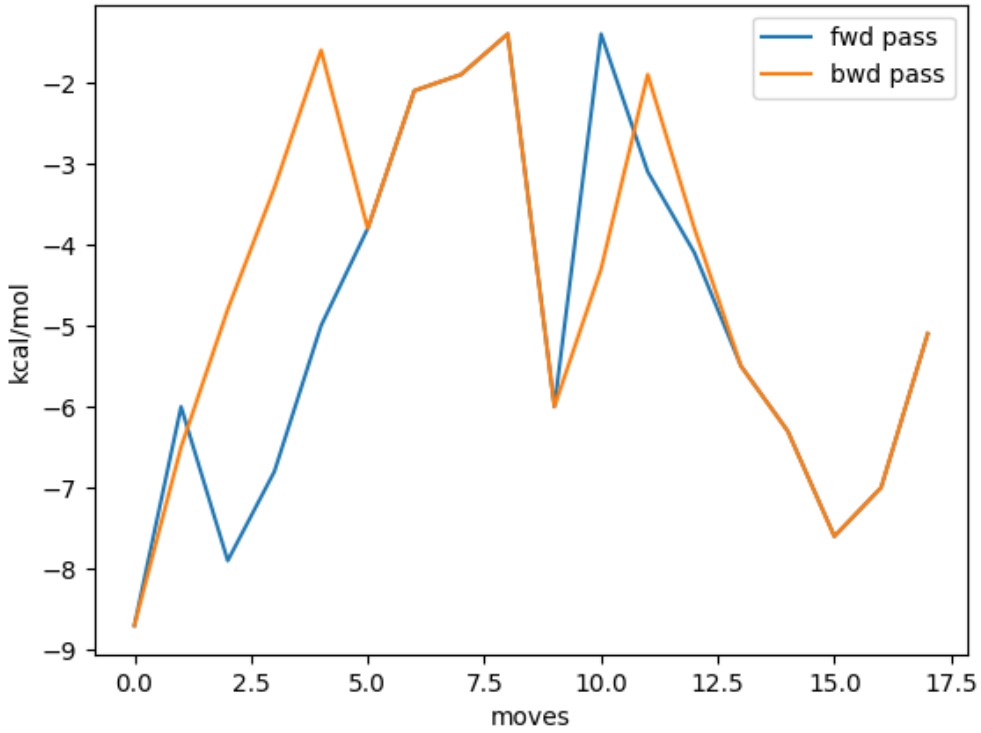
Lowest energy barrier as criterion is not good enough for subsequent merging.

Conclusions:

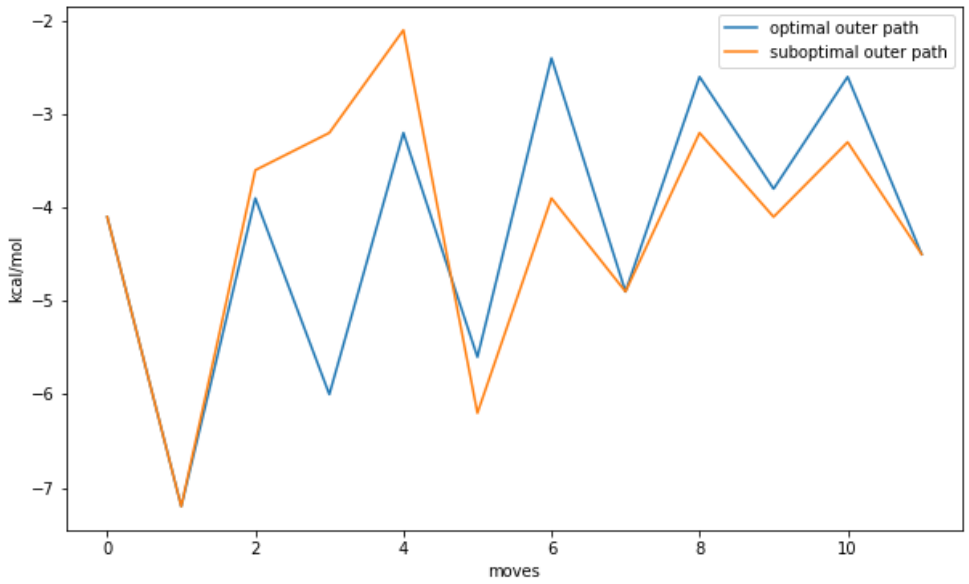
Multiple candidate paths need to be tested for merging

No guarantee that Findpath outputs the ideal input path (with respect to merging)

No guarantee that the merged output has a lower barrier than regular Findpath



Optimal input paths are not always sufficient



Findings:

Paths with local minima along the path:

suboptimal input paths are sometimes required to reconstruct optimal paths

→ Find the lowest minimum along the path

conflicts with the general findpath idea to minimize the energy barrier

Conclusions:

Adjustments to Findpath to potentially produce such paths (trade-off!)

No guarantee to find "correct" suboptimal paths

Alternative approach to merge inner/outer sections

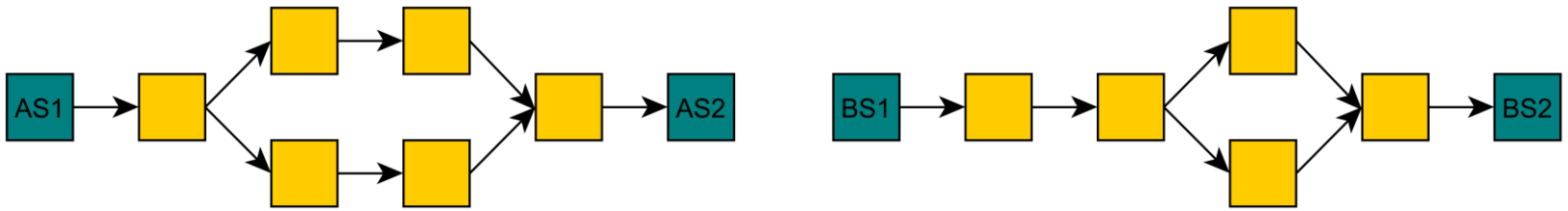
Toy example: 2 candidate paths

```
(0,  0 ) --> (-46, -51) --> (-40, -56) --> (-41, -55) --> (-43, -53) --> (-44, -52) --> ( 43,  52) --> ( 42,  53) --> ( 40,  54)
(0,  0 ) --> (-46, -51) --> (-43, -53) --> ( 42,  53) --> (-44, -52) --> ( 43,  52) --> (-41, -55) --> (-40, -56) --> ( 40,  54)
```

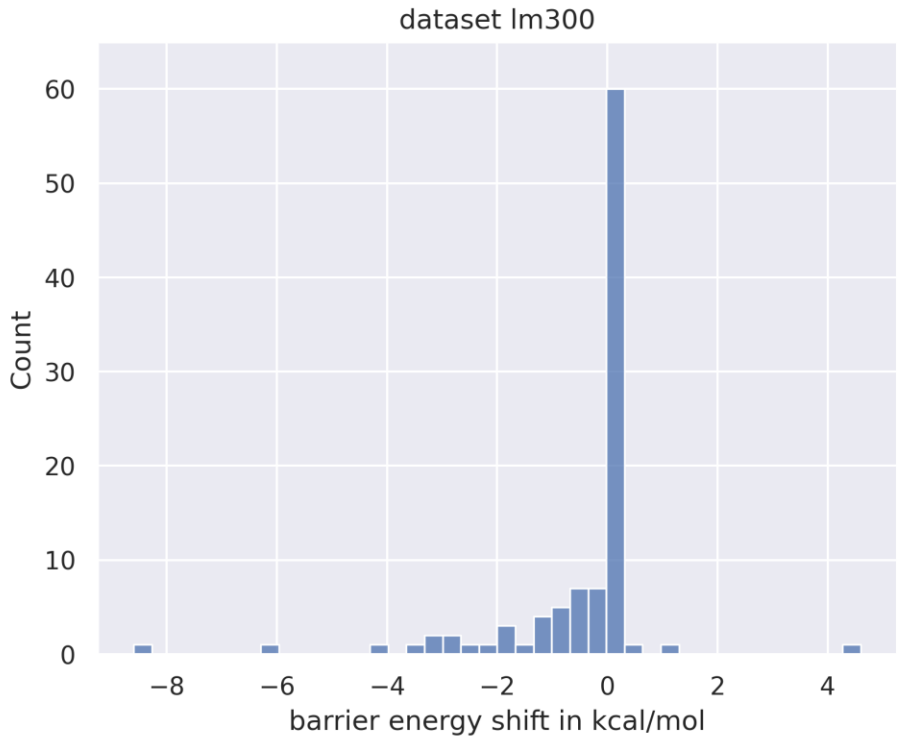
Reduce redundancy of input paths → input graph (edges are move tuples):

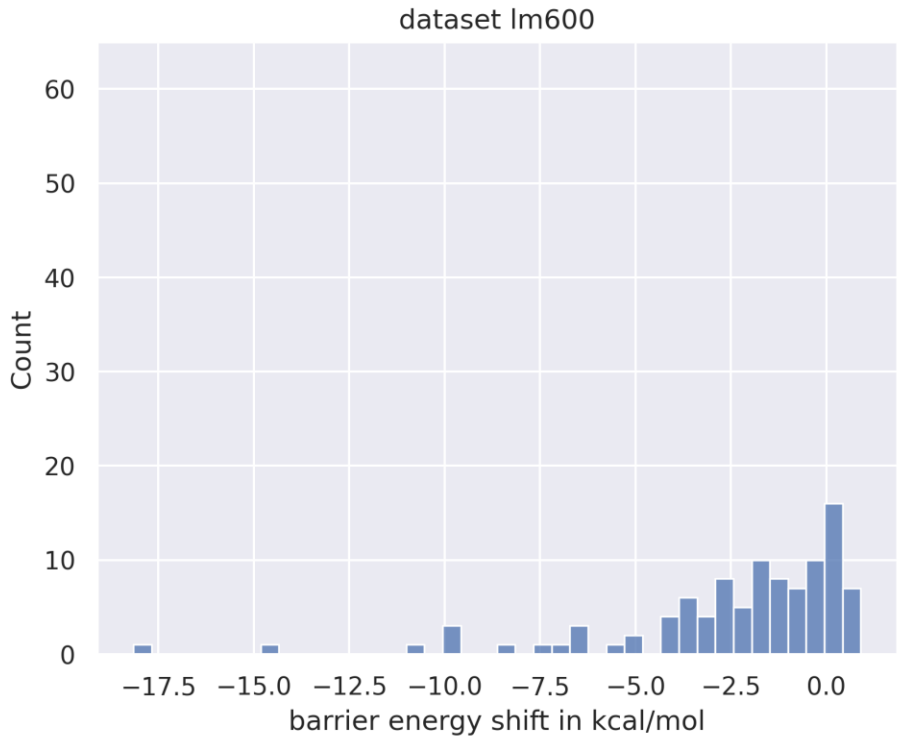
```
(0,  0 ) --> (-46, -51) --> (-40, -56) --> (-41, -55) --> (-43, -53) --> (-44, -52) --> ( 43,  52) --> ( 42,  53) --> ( 40,  54)
                     \-> (-43, -53) --> ( 42,  53) --> (-44, -52) --> ( 43,  52) --> (-41, -55) --> (-40, -56) -/
```

→ Apply Findpath algorithm to merge 2 input graphs



Barrier energy comparison
(energy shift vs. regular findpath)





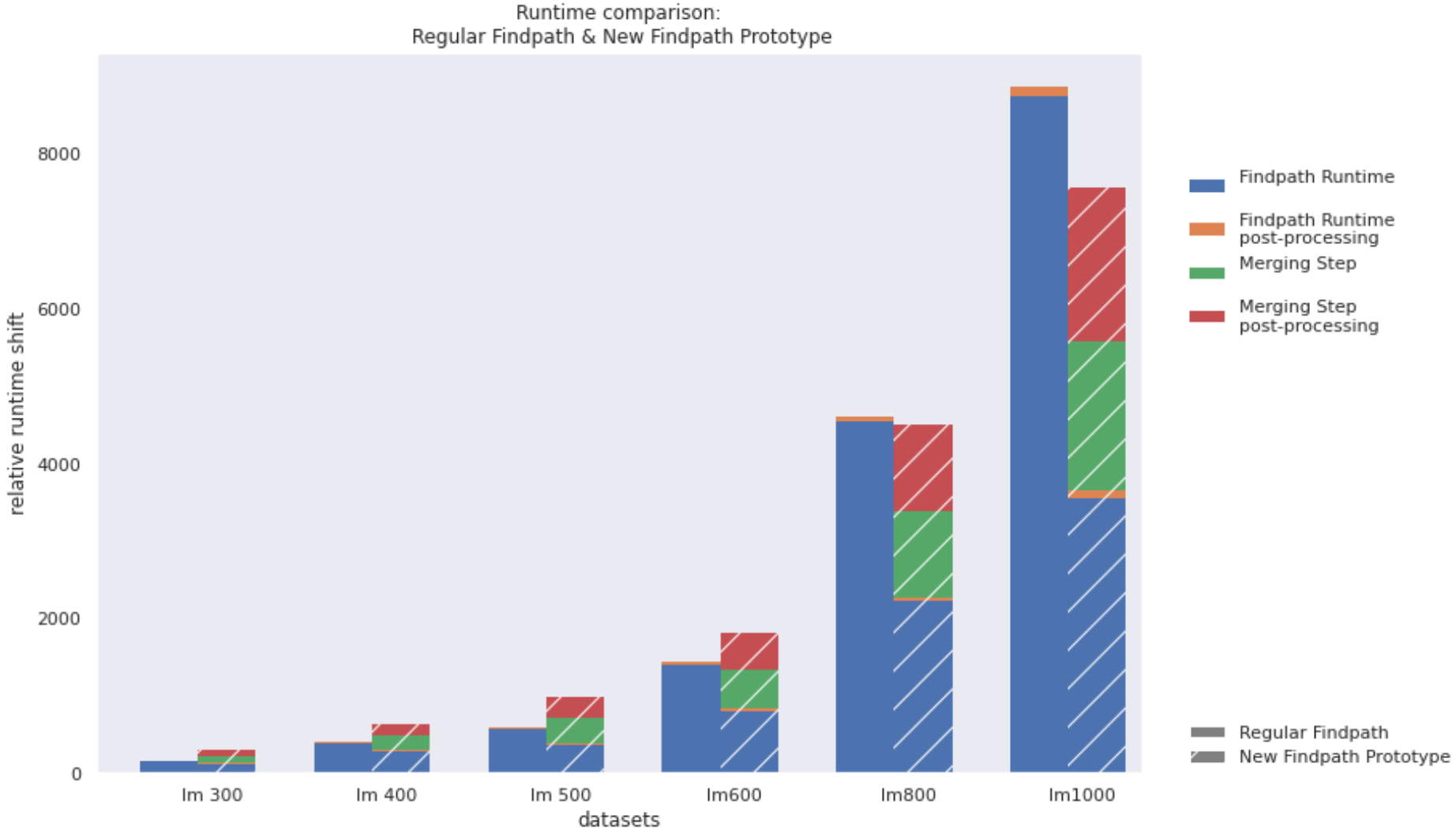
Dataset generation:

- Random RNA sequence strings
- (see slide 10)

Findpath settings:

- search width = bp_dist*2
- merge search width = bp_dist*0.4

Runtimes


Runtime comparison:
Regular Findpath & New Findpath Prototype
relative runtime shift
0
2000
4000
6000
8000
lm 300
lm 400
lm 500
lm600
lm800
lm1000
datasets
Findpath Runtime
Findpath Runtime post-processing
Merging Step
Merging Step post-processing
Regular Findpath
New Findpath Prototype

Ongoing / future work

- Adapt Findpath for indirect folding pathways – prior work:
  - *RNATabuPath*

    Ivan Dotu, William A. Lorenz, Pascal Van Hentenryck, Peter Clote, Computing folding pathways between RNA secondary structures, Nucleic Acids Research, Volume 38, Issue 5, 1 March 2010, Pages 1711–1722, https://doi.org/10.1093/nar/gkp1054
  - *RNAEAPath*

    Li, Y., Zhang, S. Predicting folding pathways between RNA conformational structures guided by RNA stacks. BMC Bioinformatics 13, S5 (2012). https://doi.org/10.1186/1471-2105-13-S3-S5

Indirect folding pathways

```
GGGCGCGGUUCGCCCUCCGCUAAAUGCGGAAGAUAAAAUUGUGUCU
(((((.....)))))(((((.....)))))(((((.....))))) [   0,    0 ] -18.10
(((((.....)))))(((((.....))))).((((.....)))). [ -31,  -45 ] -17.20
(((((.....)))))(((((.....))))).(((.......))). [ -35,  -41 ] -15.90
(((((.....)))))(((((.....))))).((.........)). [ -34,  -42 ] -13.90
(((((.....)))))(((((.....))))).(...........). [ -33,  -43 ] -12.70
(((((.....)))))(((((.....)))))............... [ -32,  -44 ] -16.80
((((.......)))))(((((.....))))).............. [  -5,  -11 ] -15.10
.(((.......))).(((((.....))))).............. [  -1,  -15 ] -11.60
..((.......))..(((((.....))))).............. [  -2,  -14 ]  -8.30
...(.......)...(((((.....))))).............. [  -3,  -13 ]  -5.00
...........(((((.....))))).............. [  -4,  -12 ]  -8.40
.....(.........(((((.....))))).........)..... [   6,   40 ]  -5.20
....((.........(((((.....))))).........))).... [   5,   41 ]  -6.90
...(((.........(((((.....))))).........)))... [   4,   42 ]  -9.10
..((((.........(((((.....))))).........)))).. [   3,   43 ] -10.70
.(((((.........(((((.....))))).........))))). [   2,   44 ] -13.00
.((((((........(((((.....))))).........)))))). [   7,   39 ] -13.80
.(((((((.......(((((.....))))).......))))))). [   8,   38 ] -14.40
.(((((((.....(((((.....)))))).....))))))). [   9,   37 ] -16.00
.(((((((((.....(((((.....)))))).....))))))))). [  10,   36 ] -17.10
((((((((((.....(((((.....)))))).....))))))))) [   1,   45 ] -17.70
S:  -5.00 kcal/mol | B:  13.10 kcal/mol | E[start]:-18.10 E[end]:-17.70
```

```
GGGCGCGGUUCGCCCUCCGCUAAAUGCGGAAGAUAAAAUUGUGUCU
(((((.....)))))(((((.....)))))(((((.....))))) [   0,    0 ] -18.10
(((((.....)))))(((((.....))))).((((.....)))). [ -31,  -45 ] -17.20
(((((.....)))))(((((.....))))).(((.......))). [ -35,  -41 ] -15.90
(((((.....)))))(((((.....))))).((.........)). [ -34,  -42 ] -13.90
(((((.....)))))(((((.....))))).(...........). [ -33,  -43 ] -12.70
(((((.....)))))(((((.....)))))............... [ -32,  -44 ] -16.80
((((.......)))))(((((.....))))).............. [  -5,  -11 ] -15.10
.(((.......))).(((((.....))))).............. [  -1,  -15 ] -11.60
..((.......))..(((((.....))))).............. [  -2,  -14 ]  -8.30
...(.......)...(((((.....))))).............. [  -3,  -13 ]  -5.00
...........(((((.....))))).............. [  -4,  -12 ]  -8.40
.....(.........(((((.....))))).........)..... [   6,   40 ]  -5.20
....((.........(((((.....))))).........))).... [   5,   41 ]  -6.90
...(((.........(((((.....))))).........)))... [   4,   42 ]  -9.10
..((((.........(((((.....))))).........)))).. [   3,   43 ] -10.70
.(((((.........(((((.....))))).........))))). [   2,   44 ] -13.00
.((((((........(((((.....))))).........)))))). [   7,   39 ] -13.80
.(((((((.......(((((.....))))).......))))))). [   8,   38 ] -14.40
.(((((((.....(((((.....)))))).....))))))). [   9,   37 ] -16.00
.(((((((((.....(((((.....)))))).....))))))))). [  10,   36 ] -17.10
((((((((((.....(((((.....)))))).....))))))))) [   1,   45 ] -17.70
GGGCGCGGUUCGCCCUCCGCUAAAUGCGGAAGAUAAAAUUGUGUCU
(((((.....)))))(((((.....)))))(((((.....))))) [   0,    0 ] -18.10
(((((.....)))))(((((.....))))).((((.....)))). [ -31,  -45 ] -17.20
(((((.....)))))(((((.....))))).(((.......))). [ -35,  -41 ] -15.90
(((((.....)))))(((((.....))))).((.........)). [ -34,  -42 ] -13.90
(((((.....)))))(((((.....))))).(...........). [ -33,  -43 ] -12.70
(((((.....)))))(((((.....)))))............... [ -32,  -44 ] -16.80
((((.......)))))(((((.....))))).............. [  -5,  -11 ] -15.10
.(((.......))).(((((.....))))).............. [  -1,  -15 ] -11.60
..((.......))..(((((.....))))).............. [  -2,  -14 ]  -8.30
...(.......)...(((((.....))))).............. [  -3,  -13 ]  -5.00
...........(((((.....))))).............. [  -4,  -12 ]  -8.40
.....(.........(((((.....))))).........)..... [   6,   40 ]  -5.20
....((.........(((((.....))))).........))).... [   5,   41 ]  -6.90
...(((.........(((((.....))))).........)))... [   4,   42 ]  -9.10
..((((.........(((((.....))))).........)))).. [   3,   43 ] -10.70
.(((((.........(((((.....))))).........))))). [   2,   44 ] -13.00
.((((((........(((((.....))))).........)))))). [   7,   39 ] -13.80
.(((((((.......(((((.....))))).......))))))). [   8,   38 ] -14.40
.(((((((.....(((((.....)))))).....))))))). [   9,   37 ] -16.00
.(((((((((.....(((((.....)))))).....))))))))). [  10,   36 ] -17.10
((((((((((.....(((((.....)))))).....))))))))) [   1,   45 ] -17.70
S:  -5.80 kcal/mol | B:  12.30 kcal/mol | E[start]:-18.10 E[end]:-17.70
```

Basepair distance distribution