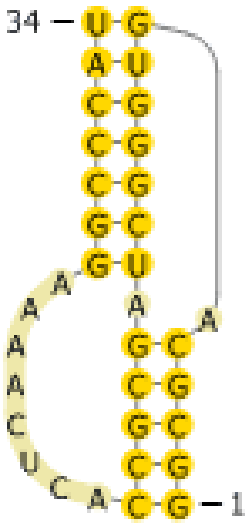
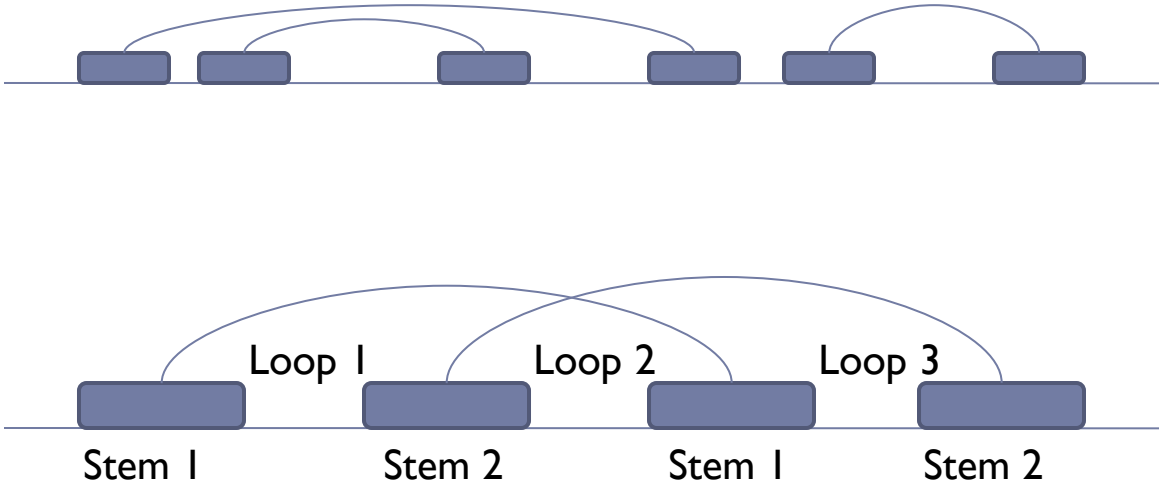
Predicting RNA pseudoknots by modifying RNAplex

Wolfgang Beyer

RNA Pseudoknots





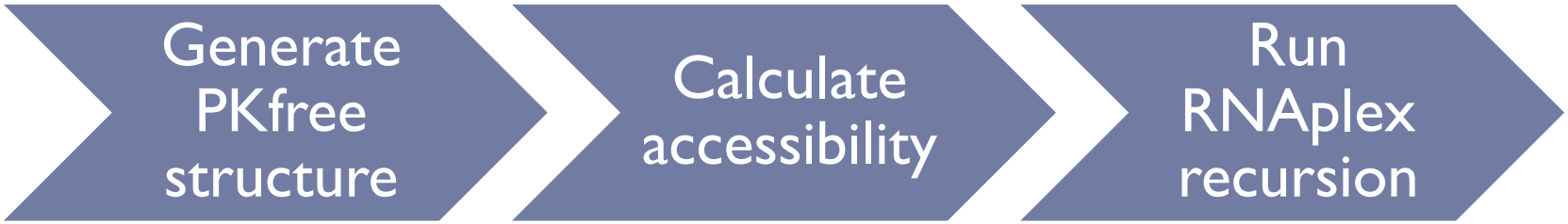
RNAplex/RNAup

- Models RNA-RNA interactions

- Limitation: single binding site

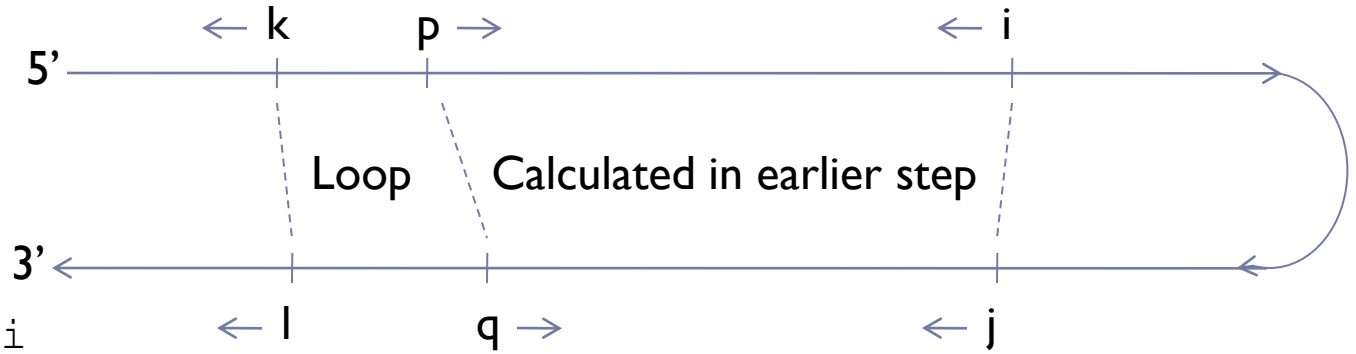
Idea

- Apply RNAplex algorithm to single sequences to predict pseudoknots



- RNAfold
- RNAplfold
- PKplex

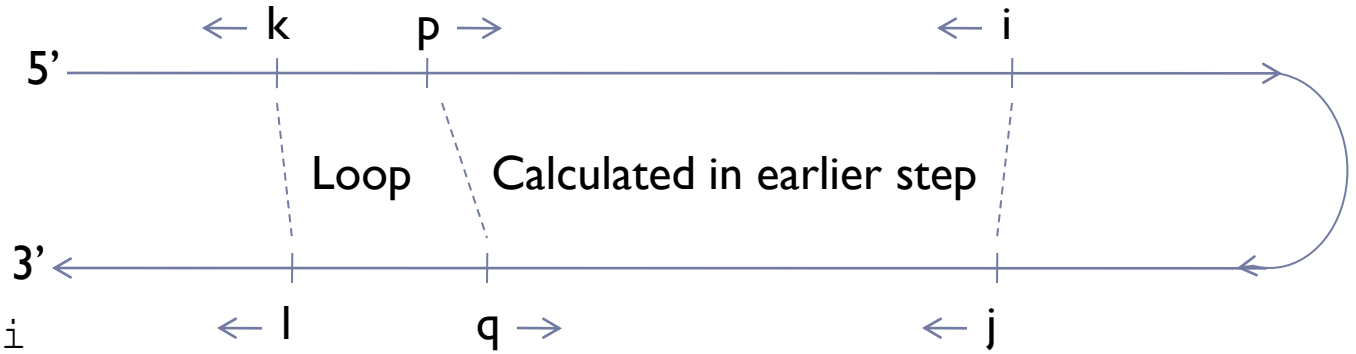
Recursion 1




```
for i
  for j
    for k
      for l
        for p
          for q
            E=LoopEnergy(p,q,k,l)
            C(i,j,k,l)=min(C(i,j,k,l), C(i,j,p,q)+E))
          next q
        next p
      next l
    next k
  next j
next i
```

Time complexity: $\mathcal{O}(n^2 w^4)$
Memory requirement: $\mathcal{O}(w^2)$

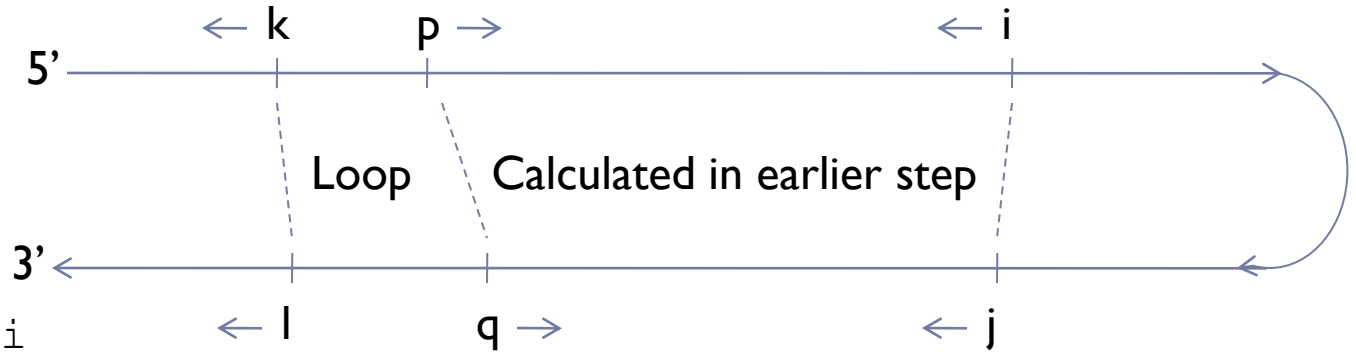
Recursion 1




```
for i
  for j
    for k
      for l
        for p
          for q
            E=LoopEnergy(p,q,k,l)
            C(i,j,k,l)=min(C(i,j,k,l), C(i,j,p,q)+E))
          next q
        next p
      next l
    next k
  next j
next i
```

Time complexity: $\mathcal{O}(n^2 w^4)$
Memory requirement: $\mathcal{O}(w^2)$

Recursion 2




```
for i
  for k
    for l
      for p
        for q
          E=LoopEnergy(p,q,k,l)
          for j
            C(i,j,k,l)=min(C(i,j,k,l), C(i,j,p,q)+E))
          next j
        next q
      next p
    next l
  next k
next i
```

Time complexity: $\mathcal{O}(n^2 w^4)$

Does not change but time consuming function LoopEnergy() is called fewer times

Memory requirement: $\mathcal{O}(n w^2)$

A simple Example

Sequence: GCGGCACCGUCCGCUCAAACAAACGG

true: ((((..[[[.))))........]]]

predicted: .(((..[[[[)))........]]]]

Measuring the prediction quality

Sensitivity and Selectivity

Sequence: GCGGCACCGUCCGCUCAAACAAACGG

true: ((((..[[[.))))........]]]

predicted: .(((..[[[[)))).........]]]]

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Selectivity = Positive Predictive Value (PPV) =

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

F-measure

Sequence: GCGGCACCGUCCGCUCAAACAAACGG

true: ((((..[[[.))))........]]]

predicted: .(((..[[[[)))........]]]]

$$\text{F-measure} = \frac{2 \times \text{Sensivity} \times \text{PPV}}{\text{Sensivity} + \text{PPV}}$$

Correctly predicted PK: at least one ()-bp and one []-bp of the PK are correct

Results

- Test data: known RNA secondary structures taken from RNA STRAND (www.rnasoft.ca/strand/)

- Comparison with
  - PKnotsRG: $\mathcal{O}(n^4)$, perfect helices with maximum length
  - HotKnots: heuristic approach
  - RNAfold: no PKs

Short PKed sequences

- Sequence length < 200
- No nested pseudoknots

n=87	PKplex	PKnotsRG	HotKnots	RNAfold
Av. Sensitivity	0.73	0.70	0.59	0.48
Av. PPV	0.73	0.72	0.64	0.61
F-measure	0.73	0.71	0.62	0.54
Contain PK	63%	56%	26%	-
Correct PK	98%	94%	91%	-
Runtime	4.22s	2.50s	288s	0.76s

Longer PKed sequences

- Sequence length < 400
- No nested pseudoknots

n=413	PKplex	PKnotsRG	RNAfold
Av. Sensitivity	0.58	0.55	0.51
Av. PPV	0.55	0.53	0.51
F-measure	0.57	0.54	0.51
Contain PK	89.8%	38.7%	-
Correct PK	37.7%	32.5%	-
Runtime	369s	1346s	41s

Sequences both with and without PKs

- Sequence length < 200
- No nested pseudoknots

n=1354	PKplex	PKnotsRG	HotKnots	RNAfold
Av. Sensitivity	0.64	0.64	0.64	0.63
Av. PPV	0.66	0.67	0.67	0.67
F-measure	0.65	0.65	0.66	0.65
Contain PK	15.9%	13.0%	4.6%	-
Correct PK	30.0%	33.7%	50.0%	-
Runtime	107s	65s	8964s	16.0s

Sequences both with and without PKs

- Sequence length < 400
- No nested pseudoknots

n=2045	PKplex	PKnotsRG	RNAfold
Av. Sensitivity	0.61	0.61	0.61
Av. PPV	0.61	0.62	0.62
F-measure	0.61	0.61	0.62
Contain PK	41.0%	19.6%	-
Correct PK	20.3%	16.3%	-
Runtime	869s	2755s	100s

Evaluation

- Quality of results comparable to PKnotsRG
- Speed of calculations competitive
- Improvements via better parameter values possible (cost of initiation of interaction)
- Energy model
- Suboptimal structures (filtering of results)

Acknowledgments

- Ivo Hofacker
- Christoph Flamm
- People at the TBI

Thank you for your attention!