Sparse RNA folding revisited: space-efficient minimum free energy prediction

Sebastian Will and Hosna Jabbari

WABI 2015

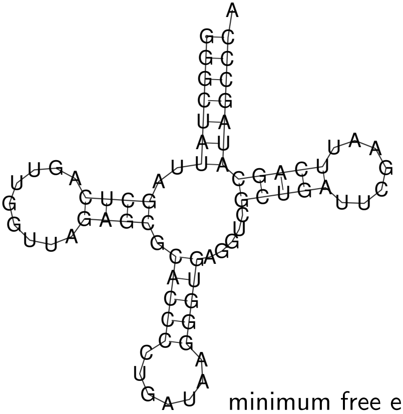
Bioinformatics, University Leipzig

RNA secondary structure prediction

```
GGGCUAUUAGCUCAGUUGGUUAGAGCGCACCCCUGAUAAGGGUGAGGUCGCUGAUUCGAAUUCAGCAUAGCCCA
```

$\Downarrow$ prediction (e.g. RNAfold)

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



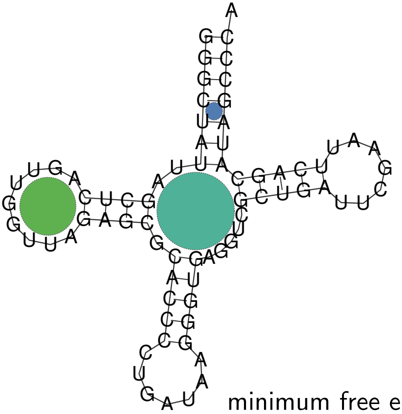
minimum free energy (MFE) structure

RNA secondary structure prediction

```
GGGCUAUUAGCUCAGUUGGUUAGAGCGCACCCCUGAUAAGGGUGAGGUCGCUGAUUCGAAUUCAGCAUAGCCCA
```

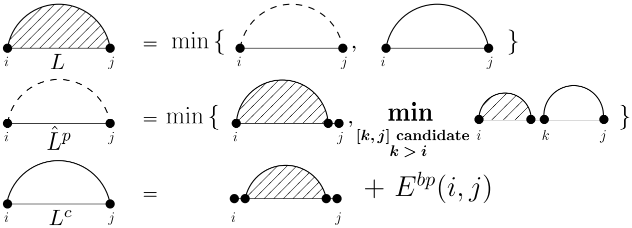
$\Downarrow$ prediction (e.g. RNAfold)

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



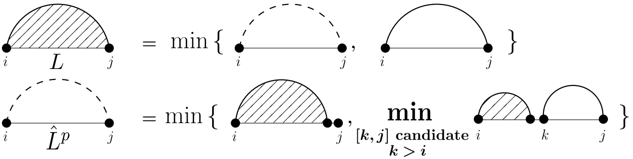
minimum free energy (MFE) structure

Sparsified prediction: base pair-based energy



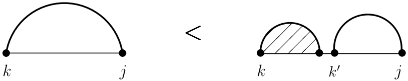
Backofen et al. JDA 2011

Sparsified prediction: base pair-based energy



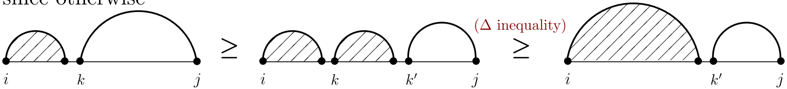
i L j $= \min\{$ i j , i j $\}$
i $\hat{L}^p$ j $= \min\{$ i j , $\min_{\substack{[k,j]\ \text{candidate} \\ k>i}}$ i k j $\}$
i L^c j $=$ i j $+ E^{bp}(i,j)$

in min, consider split at k only if



candidate criterion

since otherwise



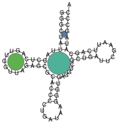
Complexity $O(n^2 + n \cdot Z_L)$ time; $\Theta(n + Z_L)$ space

(Z_L = total # of candidates)

Backofen et al. JDA 2011

Minimum free energy prediction

loop-based energy:



Original recursions

[Zuker & Sankoff, 1984; like implemented by modern tools]

$$W(i,j) = \min\{\, V(i,j), \min_{i<k<j} W(i,k) + W(k+1,j)\,\}$$

$$V(i,j) = \min\{\, \mathcal{H}(i,j), \min_{\substack{i<p<q<j \\ p-i+j-q-2\leq M}} \mathcal{I}(i,j,p,q) + V(p,q),$$

$$\min_{i<k<j} WM(i+1,k) + WM(k+1,j-1) + a\}$$

$$WM(i,j) = \min\{\, V(i,j) + b, WM(i+1,j) + c, WM(i,j-1) + c,$$

$$\min_{i<k<j} WM(i,k) + WM(k+1,j)\,\}$$

Rewrite to prepare sparsification ...

$$\begin{aligned}
W(i,j) &= \min\{\, W^p(i,j), V(i,j) \,\}\\
W^p(i,j) &= \min\{\, W(i,j-1), \min_{i<k<j} W(i,k-1) + W(k,j) \,\}\\
V(i,j) &= \min\{\mathcal{H}(i,j), \min_{\substack{i<p<q<j\\ p-i+j-q-2\leq M}} \mathcal{I}(i,j,p,q) + V(p,q), WM^2(i+1,j-1) + a\}\\
WM(i,j) &= \min\{\, WM^p(i,j), V(i,j) + b \,\}\\
WM^p(i,j) &= \min\{\, WM(i+1,j) + c, WM(i,j-1) + c, WM^2(i,j) \,\}\\
WM^2(i,j) &= \min_{i<k<j} WM(i,k-1) + WM(k,j)
\end{aligned}$$

... and sparsify: minimize only over candidates

$$\widehat{W}^{\mathrm{p}}(i,j) = \min\{\, W(i,j-1), \min_{\substack{[k,j]\ \text{W-candidate},\\ k>i}} W(i,k-1) + V(k,j) \,\}$$

$$\widehat{WM}^2(i,j) = \min\{\, WM^2(i,j-1) + c, \min_{\substack{[k,j]\ \text{WM-candidate},\\ k>i}} WM(i,k-1) + V(k,j) + b \,\}$$

candidate criteria:

- $[k,j]$ is a *W-candidate* iff $V(k,j) < \widehat{W}^{\mathrm{p}}(k,j)$ and
- $[k,j]$ is a *WM-candidate* iff $V(k,j) + b < WM^p(k,j)$.

Rewrite to prepare sparsification ...

$$W(i,j) = \min\{\, W^p(i,j), V(i,j) \,\}$$

$$W^p(i,j) = \min\{\, W(i,j-1), \min_{i<k<j} W(i,k-1) + W(k,j) \,\}$$

$$V(i,j) = \min\{\mathcal{H}(i,j), \min_{\substack{i<p<q<j \\ p-i+j-q-2\leq M}} \mathcal{I}(i,j,p,q) + V(p,q), WM^2(i+1,j-1) + a\}$$

$$WM(i,j) = \min\{\, WM^p(i,j), V(i,j) + b \,\}$$

$$WM^p(i,j) = \min\{\, WM(i+1,j) + c, WM(i,j-1) + c, WM^2(i,j) \,\}$$

$$WM^2(i,j) = \min_{i<k<j} WM(i,k-1) + WM(k,j)$$

... and sparsify: minimize only over candidates

$$\widehat{W}^{\mathrm{p}}(i,j) = \min\{\, W(i,j-1), \min_{\substack{[k,j]\text{ W-candidate,} \\ k>i}} W(i,k-1) + V(k,j) \,\}$$

$$\widehat{WM}^2(i,j) = \min\{\, WM^2(i,j-1) + c, \min_{\substack{[k,j]\text{ WM-candidate,} \\ k>i}} WM(i,k-1) + V(k,j) + b \,\}$$

candidate criteria:

- $[k,j]$ is a *W-candidate* iff $V(k,j) < \widehat{W}^{\mathrm{p}}(k,j)$ and
- $[k,j]$ is a *WM-candidate* iff $V(k,j) + b < WM^p(k,j)$.

Rewrite to prepare sparsification ...

$$\begin{aligned}
W(i,j) &= \min\{\, W^p(i,j), V(i,j) \,\} \\
W^p(i,j) &= \min\{\, W(i,j-1), \min_{i<k<j} W(i,k-1) + W(k,j) \,\} \\
V(i,j) &= \min\{\mathcal{H}(i,j), \min_{\substack{i<p<q<j \\ p-i+j-q-2\leq M}} \mathcal{I}(i,j,p,q) + V(p,q), WM^2(i+1,j-1) + a\} \\
WM(i,j) &= \min\{\, WM^p(i,j), V(i,j) + b \,\} \\
WM^p(i,j) &= \min\{\, WM(i+1,j) + c, WM(i,j-1) + c, WM^2(i,j) \,\} \\
WM^2(i,j) &= \min_{i<k<j} WM(i,k-1) + WM(k,j)
\end{aligned}$$

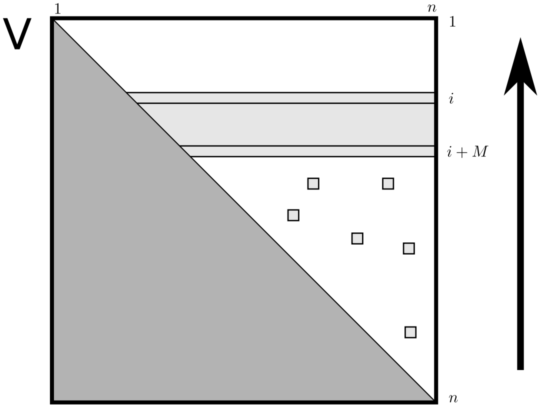
... and sparsify: minimize only over candidates

$$\begin{aligned}
\widehat{W}^{\mathrm{p}}(i,j) &= \min\{\, W(i,j-1), \min_{\substack{[k,j] \text{ W-candidate,} \\ k>i}} W(i,k-1) + V(k,j) \,\} \\
\widehat{WM}^2(i,j) &= \min\{\, WM^2(i,j-1) + c, \min_{\substack{[k,j] \text{ WM-candidate,} \\ k>i}} WM(i,k-1) + V(k,j) + b \,\}
\end{aligned}$$

candidate criteria:

- $[k,j]$ is a *W-candidate* iff $V(k,j) < \widehat{W}^{\mathrm{p}}(k,j)$ and
- $[k,j]$ is a *WM-candidate* iff $V(k,j) + b < WM^p(k,j)$.

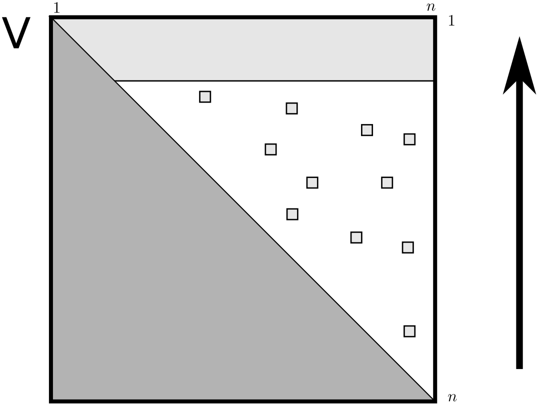
So far: minimum free energy



$O(n^2 + nZ)$ time and $\Theta(Mn + Z)$ space

Z = total # of *candidates*

So far: minimum free energy



$O(n^2 + nZ)$ time and $\Theta(Mn + Z)$ space

Z = total # of *candidates*

. . . but no MFE structure

Space-efficient bp-based prediction: Trace back

Sparse TB in base pair-based model:

Problem: forward evaluation stores only candidates

Solution (Backofen et al., JDA11):
recompute row-by-row for $i = 1$ to n
recomputation never needs non-candidates in rows $i' > i$, **since** candidates don't have to be recomputed!

Not transferable to (loop-based) MFE prediction!

- trace back of interior loops needs access to entries in rows $i' > i$, since TA is unknown
- this cannot be restricted to candidates

Example:

```
GCCAAAAGGGC          CAAAAGG       CAAAAGG
(((.....)))          (.....)   >   (....).
```

Space-efficient bp-based prediction: Trace back

Sparse TB in base pair-based model:

Problem: forward evaluation stores only candidates

Solution (Backofen et al., JDA11):
recompute row-by-row for $i = 1$ to n
recomputation never needs non-candidates in rows $i' > i$, **since** candidates don't have to be recomputed!

Not transferable to (loop-based) MFE prediction!

- trace back of interior loops needs access to entries in rows $i' > i$, since TA is unknown
- this cannot be restricted to candidates

Example:

```
GCCAAAAGGGC     CAAAAGG       CAAAAGG
(((.....)))     (.....)   >   (....).
```

Space-efficient bp-based prediction: Trace back

Sparse TB in base pair-based model:

Problem: forward evaluation stores only candidates

Solution (Backofen et al., JDA11):
recompute row-by-row for $i = 1$ to n
recomputation never needs non-candidates in rows $i' > i$, **since** candidates don't have to be recomputed!

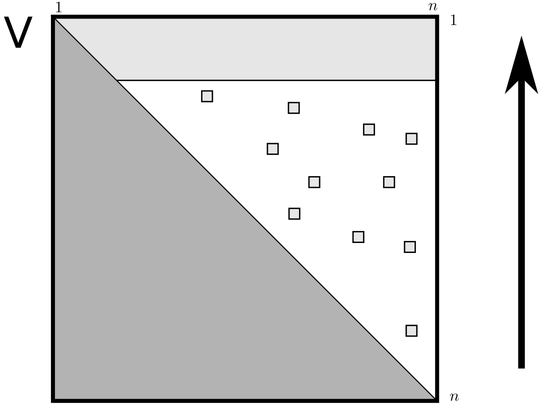
Not transferable to (loop-based) MFE prediction!

- trace back of interior loops needs access to entries in rows $i' > i$, since TA is unknown
- this cannot be restricted to candidates

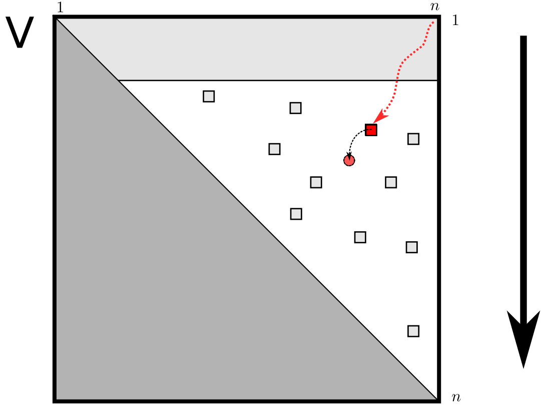
Example:

```
GCCAAAAGGGC        CAAAAGG       CAAAAGG
(((.....)))        (.....)   >   (....).
```

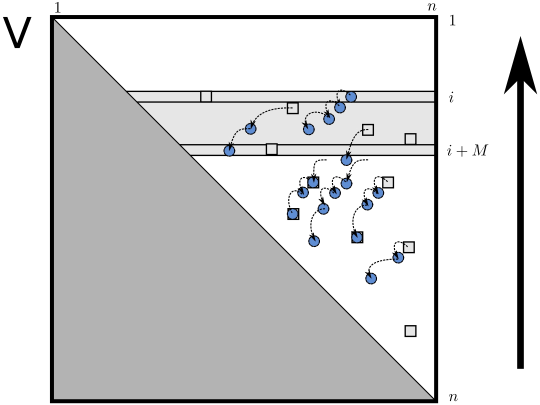
Sparse space-efficient MFE trace back



Sparse space-efficient MFE trace back



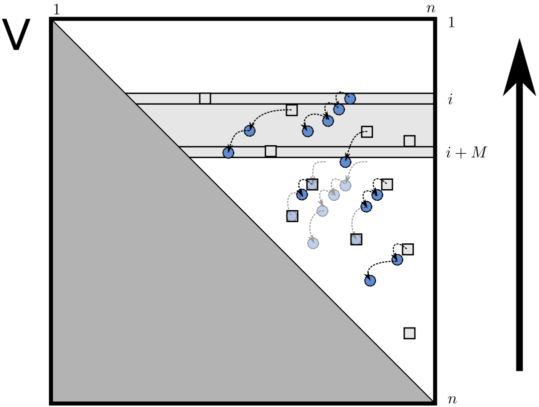
Sparse space-efficient MFE trace back



Naïve solution: store all trace arrows . . .

. . . but too many TAs; compromises “space-efficient”

Sparse space-efficient MFE trace back



Idea: avoid storing many TAs & garbage collect

- avoid TAs in case $WM(i+1,j)+c$ of WM^p (rewrite recursions)
- avoid TAs to candidates (since we can recompute)
- garbage collect: keep only accessible TAs

Results

Theory: $O(n^2 + nZ)$ time; $\Theta(Mn + Z + T)$ space
Z = total # of *candidates*; T = maximum # of accessible TAs.

Note: $T + Z < n^2$ (idea "<<")

Practice: C++ implementation SparseMFEFold

- interface to Vienna RNA lib 2.x [Lorenz et al., 2011]
- predictions identical to Vienna's `RNAfold -d0`

SparseMFEFold is available (GPL 3.0) at
`www.bioinf.uni-leipzig.de/~will/Software/SparseMFEFold`

Results

Theory: $O(n^2 + nZ)$ time; $\Theta(Mn + Z + T)$ space
Z = total # of *candidates*; T = maximum # of accessible TAs.

Note: $T + Z < n^2$ (idea "<<")

Practice: C++ implementation SparseMFEFold

- interface to Vienna RNA lib 2.x [Lorenz et al., 2011]
- predictions identical to Vienna's `RNAfold -d0`

SparseMFEFold is available (GPL 3.0) at
`www.bioinf.uni-leipzig.de/~will/Software/SparseMFEFold`

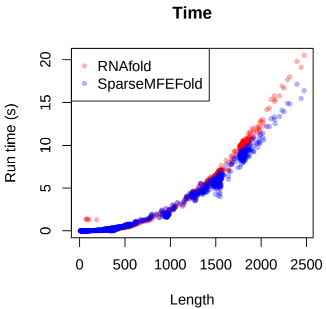
Empirical results

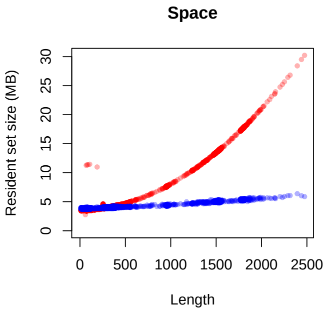
Benchmark: RNA STRAND 2.0

Performance of SparseMFEFold vs. RNAfold (length $\geq$ 2500)

	Run time (s)		Space: resident set size (MB)	
	RNAfold	SparseMFEFold	RNAfold	SparseMFEFold
Minimum	16.9	15.4	31.0	5.8 (19%)
Median	29.7	22.9	41.8	7.1 (17%)
Maximum	89.9	57.4	86.5	8.8 (10%)

length $\leq$ 2500:

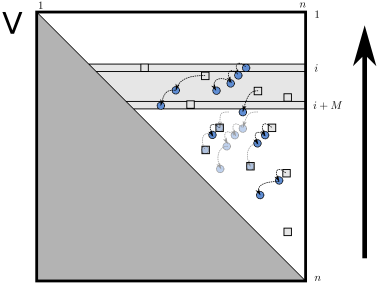




Empirical results: Candidates and TA savings

Benchmark: RNA STRAND 2.0 (length $\geq$ 2500)

	Number of candidates	**Number of trace arrows**		
		Maximum	Avoided	GC-Removed
Minimum	17,032	52,293	137,892	467,230
Median	41,215	94,443	237,717	706,365
Maximum	71,508	148,947	419,825	1,748,491



Perspectives

Techniques are generalizable

Promising applications:

Traceback of *highly complex* structure prediction

- MFE Pseudoknot prediction [Rivas, Eddy]
  - $O(n^4)$ space
  - [Möhl et al., 2011]: sparse evaluation, not space-efficient
- MFE PK-prediction "CCJ" [Chen, Condon, Jabbari]
  - $O(n^4)$ space
  - work in progress with Hosna Jabbari
  - motivation of this work
- MFE RNA-RNA-interaction prediction [Alkan et al.]
  - $O(n^4)$ space
  - [Salari et al., 2010]: space-efficient evaluation, but no space-efficient TB
- Simultaneous Folding and Alignment
  - $O(n^4)$ space [Sankoff, 1985]
  - $O(n^2)$ space [LocARNA, 2007], [SPARSE, 2015]

Conclusions

- Sparsification can strongly reduce memory demands (constant # of rows + candidates)
- Traceback of MFE prediction needs additional information (TAs)
- The novel approach keeps additional memory requirements low
- Techniques (rewriting, partial recomputation, and GC) generalize
- Promising: Apply to highly complex prediction algorithms

Thanks to . . .

- Hosna Jabbari & Anne Condon
- Peter Stadler
- you, for your attention

Conclusions

- Sparsification can strongly reduce memory demands (constant # of rows + candidates)
- Traceback of MFE prediction needs additional information (TAs)
- The novel approach keeps additional memory requirements low
- Techniques (rewriting, partial recomputation, and GC) generalize
- Promising: Apply to highly complex prediction algorithms

Thanks to . . .

- Hosna Jabbari & Anne Condon
- Peter Stadler
- you, for your attention