SPARSE: Quadratic Time SA&F of RNAs without Sequence-Based Heuristics

Sebastian Will

University of Leipzig

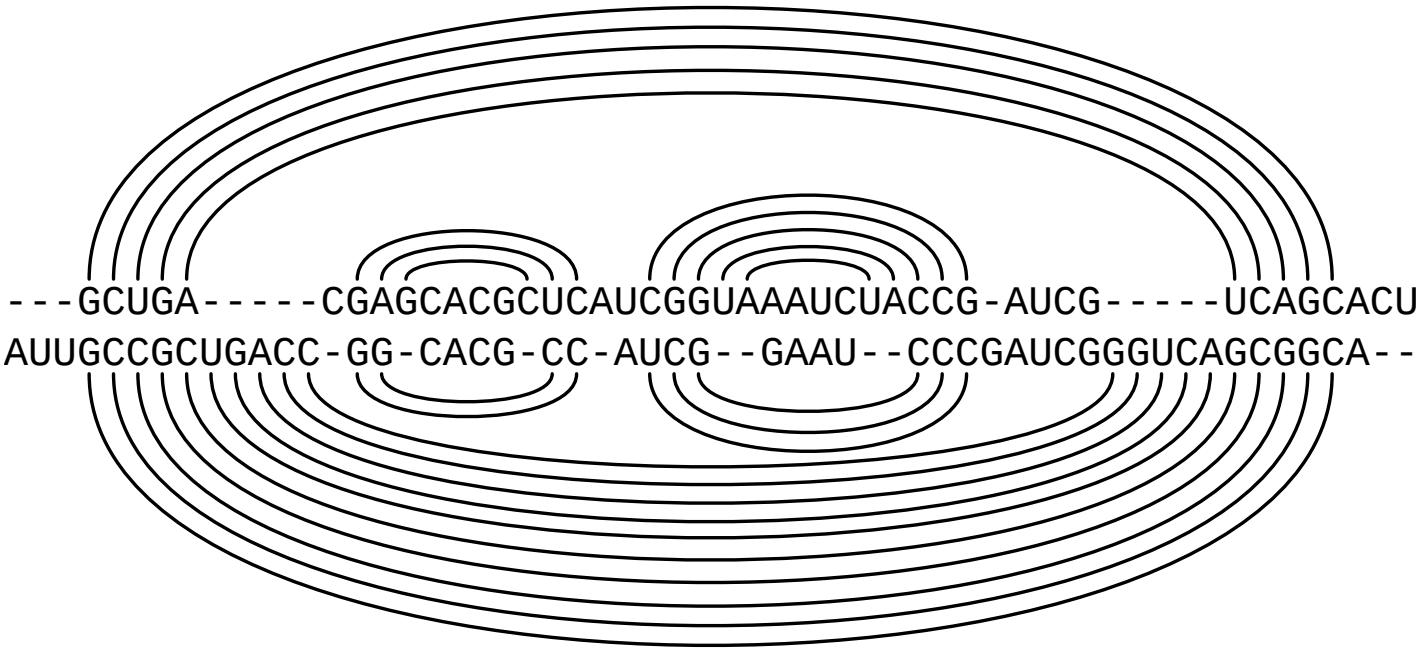
S. Will, Ch. Schmiedl, M. Miladi, M. Möhl, R. Backofen. Bioinformatics, 2015.

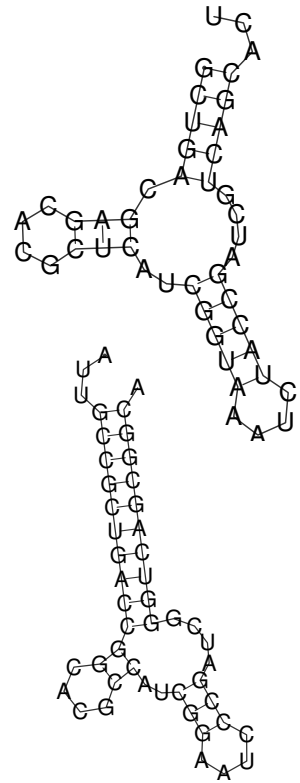
Simultaneous Alignment and Folding [Sankoff]

Given: A = GCUGACGAGCACGCUCAUCGGUAAAUCUACCGAUCGUCAGCACU

& B = AUUGCCGCUGACCGGCACGCCAUCGGAAUCCCGAUCGGGUCAGCGGCA

Find:





sequence similarity + energy A + energy B → opt

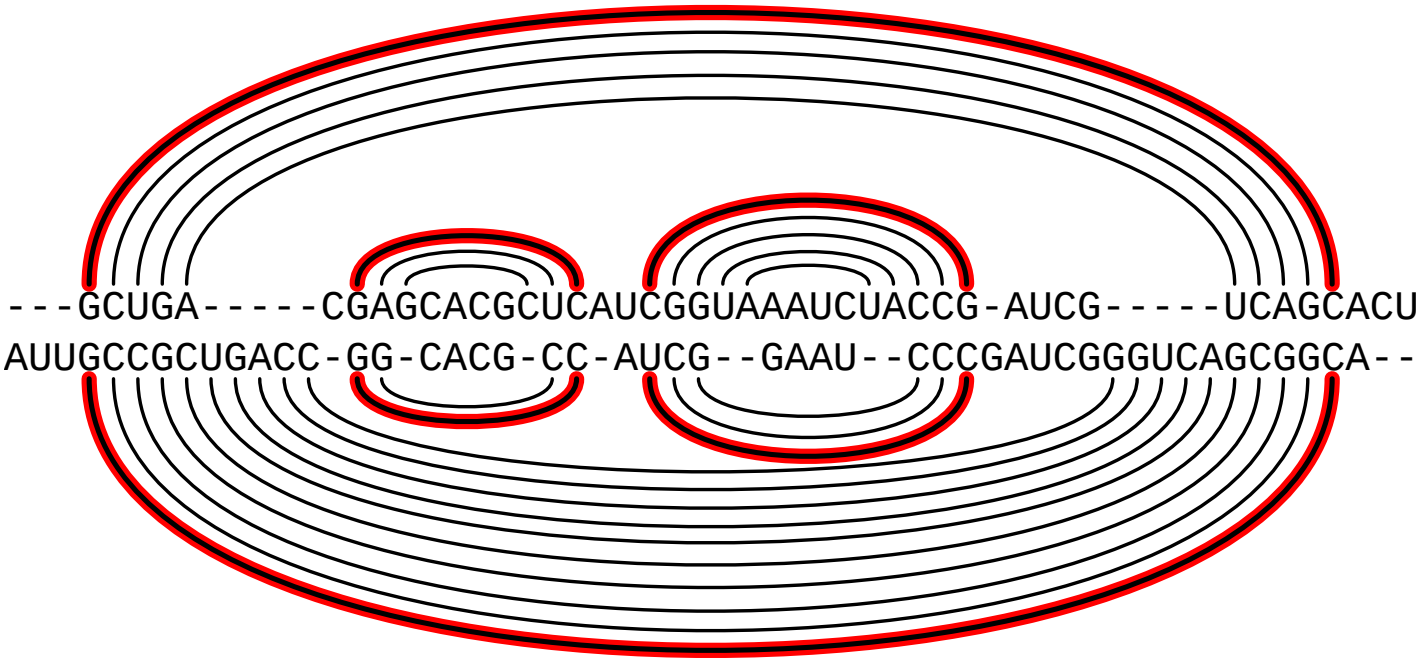
where alignment, structure A, & structure B are **compatible**

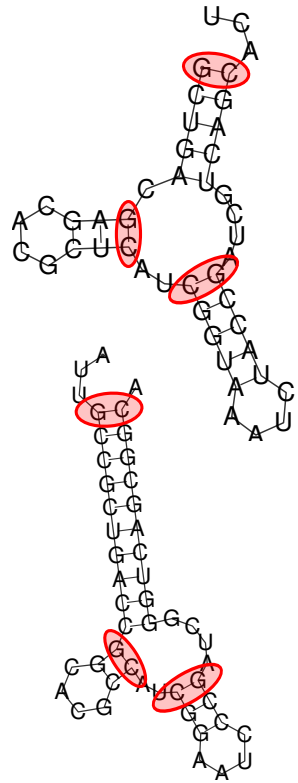
Simultaneous Alignment and Folding [Sankoff]

Given: A = GCUGACGAGCACGCUCAUCGGUAAAUCUACCGAUCGUCAGCACU

& B = AUUGCCGCUGACCGGCACGCCAUCGGAAUCCCGAUCGGGUCAGCGGCA

Find:





sequence similarity + energy A + energy B → opt

where alignment, structure A, & structure B are **compatible**

Sankoff's SA&F Algorithm

Dynamic Programming

RNA Energy Minimization [Zuker]

$\otimes$

Sequence Alignment

$O(n^6)$ = **"extreme computational cost"**

Sankoff-style Approaches

HEAVY

Dynalign
FoldAlign

- Sankoff implementations
- full ("heavy") energy model
- (sequence-based) heuristics

LIGHT

PMcomp

- lightweight energy model
- base pair probabilities

LocARNA

+ sparsifies structure space (ensemble-based)

- improves time and space

RAF

+ sparsifies alignment space

- sequence-based heuristics

SPARSE

- strong sparsification w/o sequence-based heuristics

Sankoff-style Approaches

HEAVY

Dynalign
FoldAlign

- Sankoff implementations
- full ("heavy") energy model
- (sequence-based) heuristics

LIGHT

PMcomp

- lightweight energy model
- base pair probabilities

LocARNA

+ sparsifies structure space (ensemble-based)

- improves time and space

RAF

+ sparsifies alignment space

- sequence-based heuristics

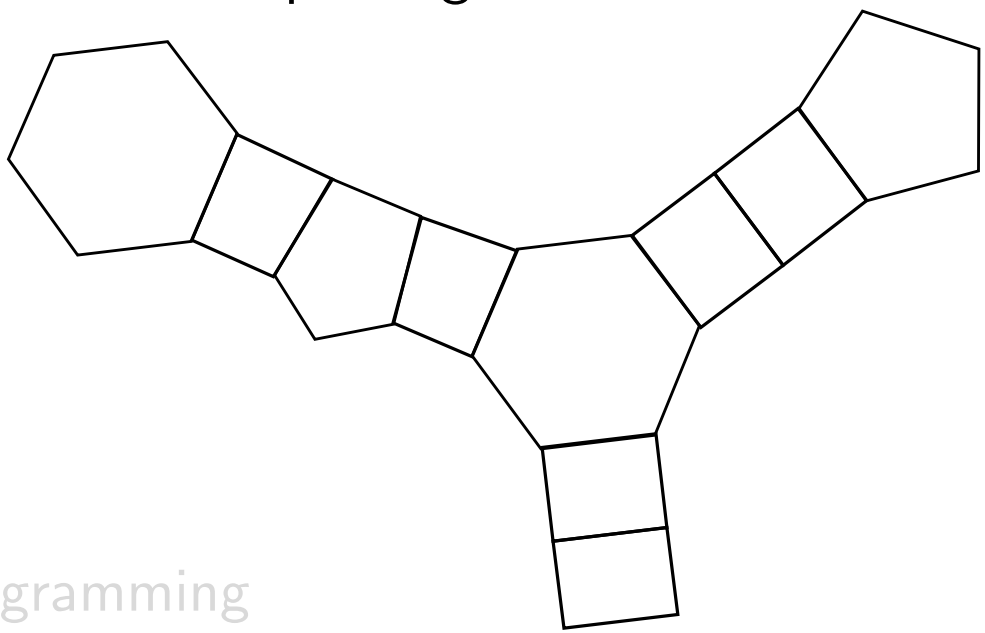
SPARSE 

- strong sparsification w/o sequence-based heuristics

PMcomp's Trick – Lightweight SA&F

Sankoff: **sequence similarity**
+ energies of A and B → **opt**

- **energy** composed of loop energies



- Dynamic Programming

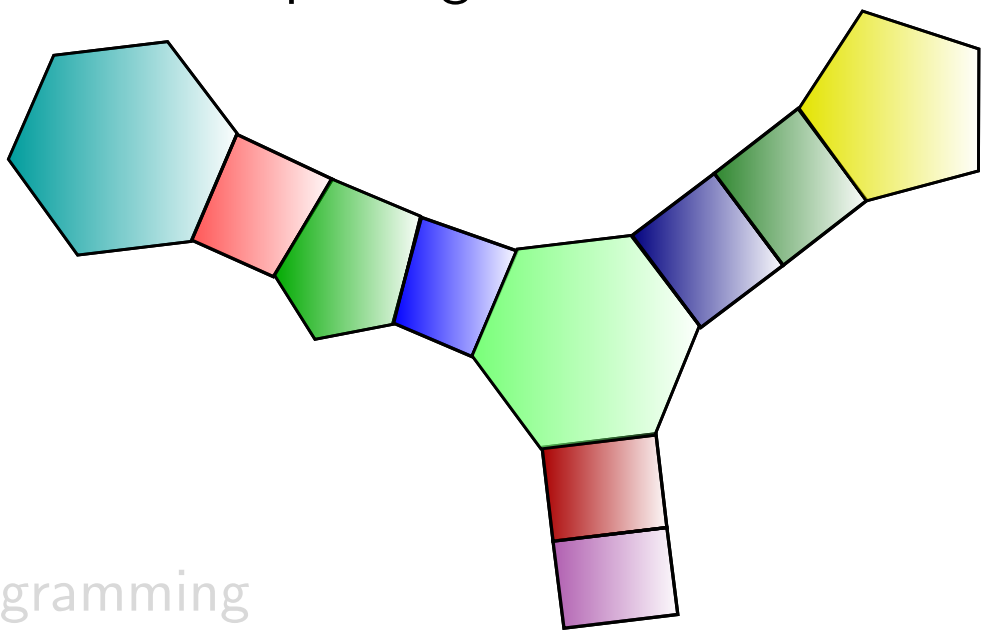
  Base Pair Maximization [Nussinov] ⊗ Sequence Alignment

- **cheaper but same complexity**

PMcomp's Trick – Lightweight SA&F

Sankoff: **sequence similarity**
+ energies of A and B → **opt**

- **energy** composed of loop energies



- Dynamic Programming

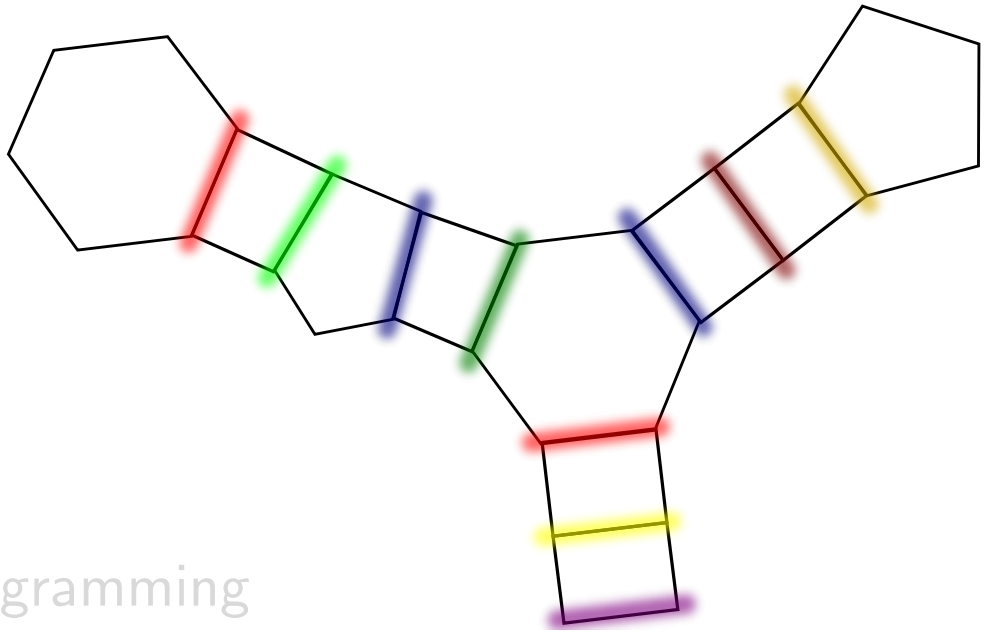
  Base Pair Maximization [Nussinov] ⊗ Sequence Alignment

- **cheaper but same complexity**

PMcomp's Trick – Lightweight SA&F

PMcomp: **sequence similarity**
+ pseudo-energies of A and B → **opt**

- **pseudo-energy** composed of "base pair energies"



- Dynamic Programming

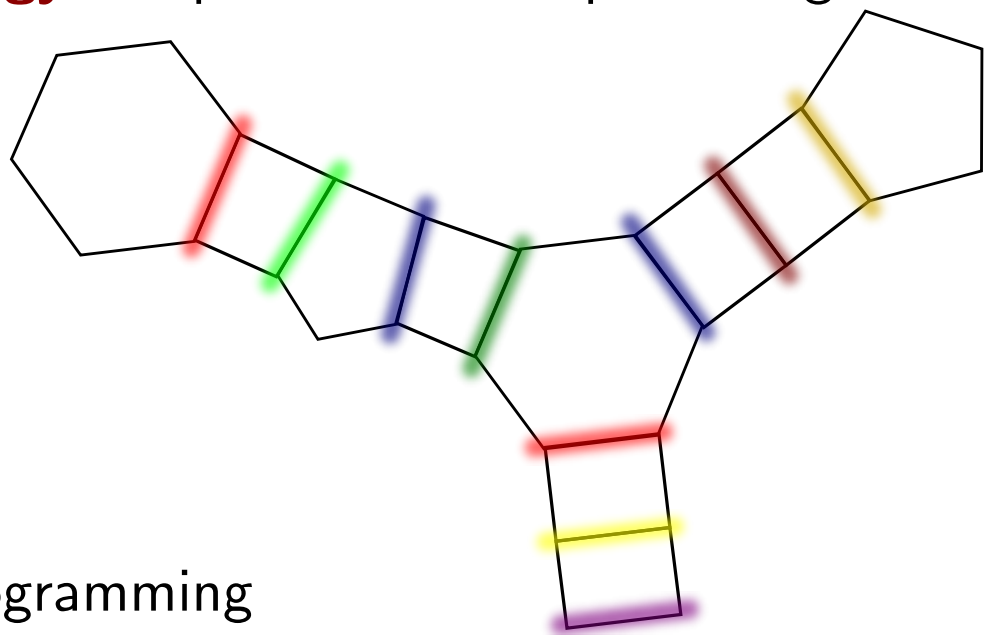
  Base Pair Maximization [Nussinov] ⊗ Sequence Alignment

- **cheaper but same complexity**

PMcomp's Trick – Lightweight SA&F

PMcomp: **sequence similarity**
+ pseudo-energies of A and B → **opt**

- **pseudo-energy** composed of "base pair energies"



- Dynamic Programming

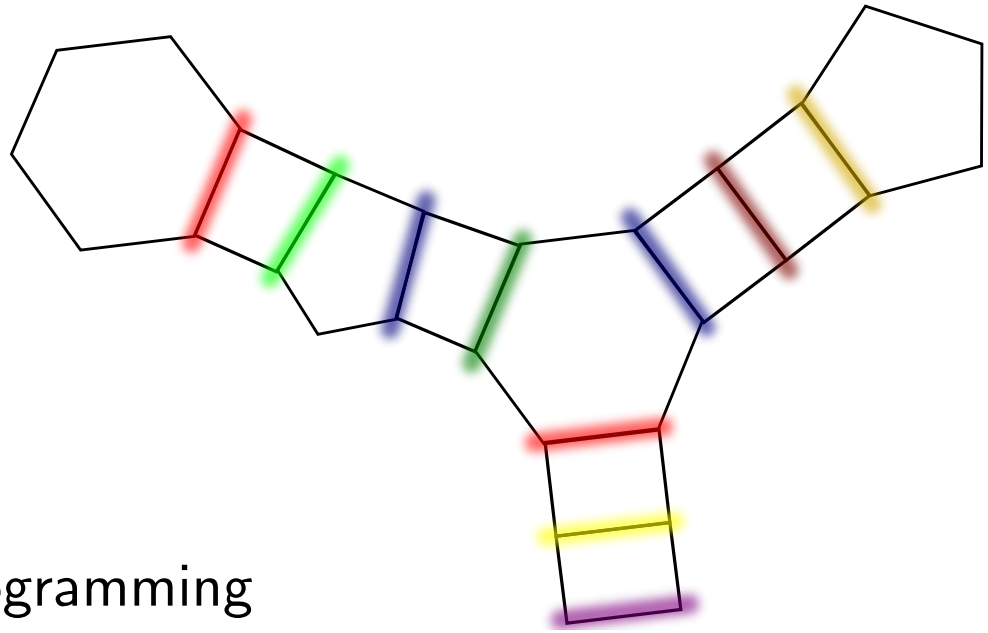
  Base Pair Maximization [Nussinov] ⊗ Sequence Alignment

- **cheaper but same complexity**

PMcomp's Trick – Lightweight SA&F

PMcomp: **sequence similarity**
+ pseudo-energies of A and B → **opt**

- **pseudo-energy** composed of "base pair energies"



- Dynamic Programming

  Base Pair Maximization [Nussinov] ⊗ Sequence Alignment

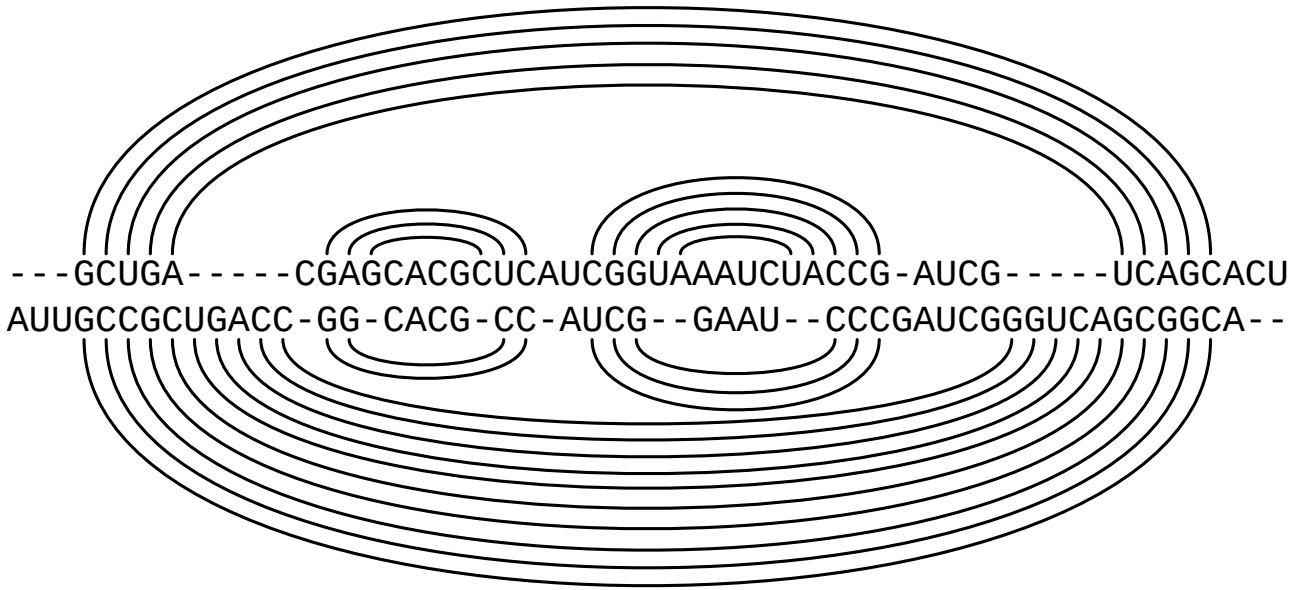
- **cheaper but same complexity**

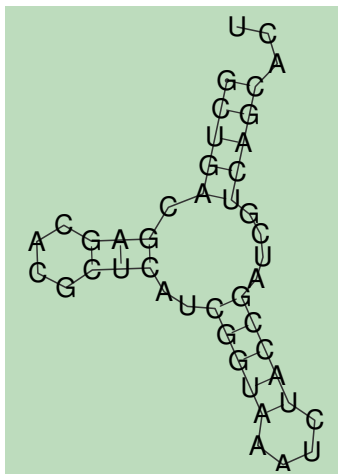
PMcomp – THE Lightweight Sankoff Algorithm?

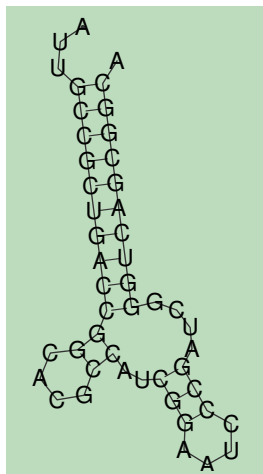
compatibility

Sankoff: *same shape*

PMcomp: *all base pairs match*





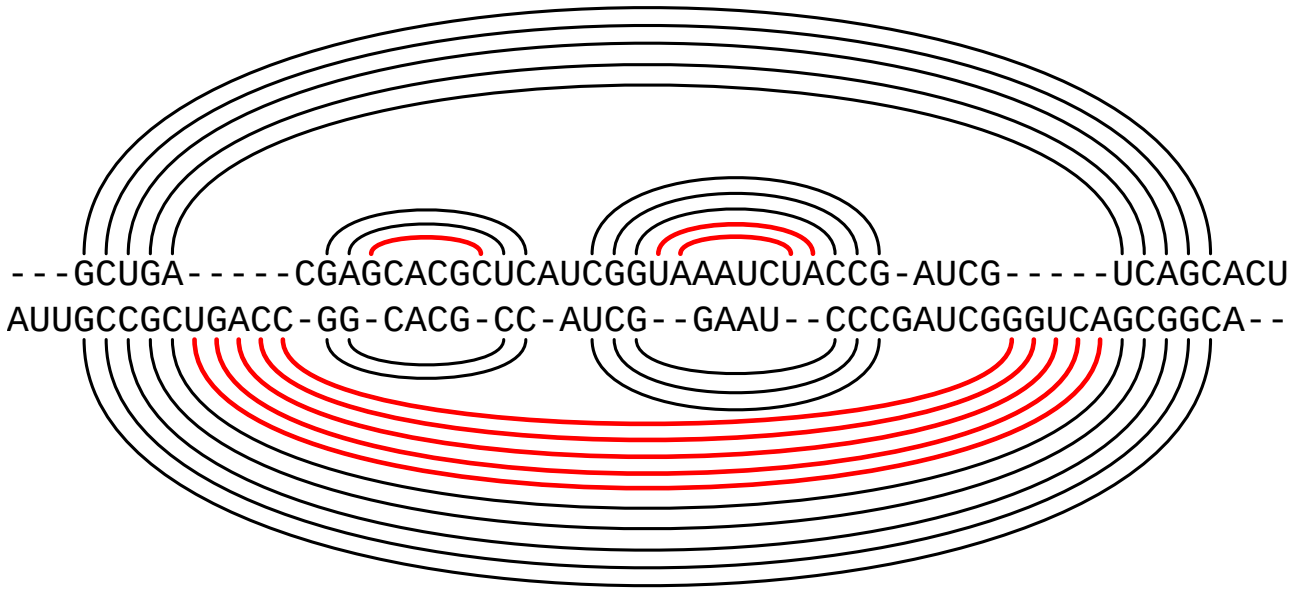


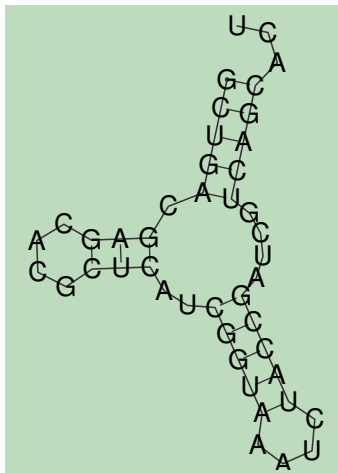
PMcomp – THE Lightweight Sankoff Algorithm?

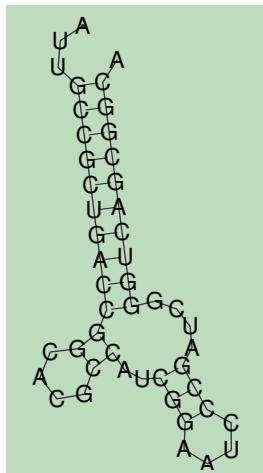
compatibility

Sankoff: *same shape*

PMcomp: *all base pairs match*





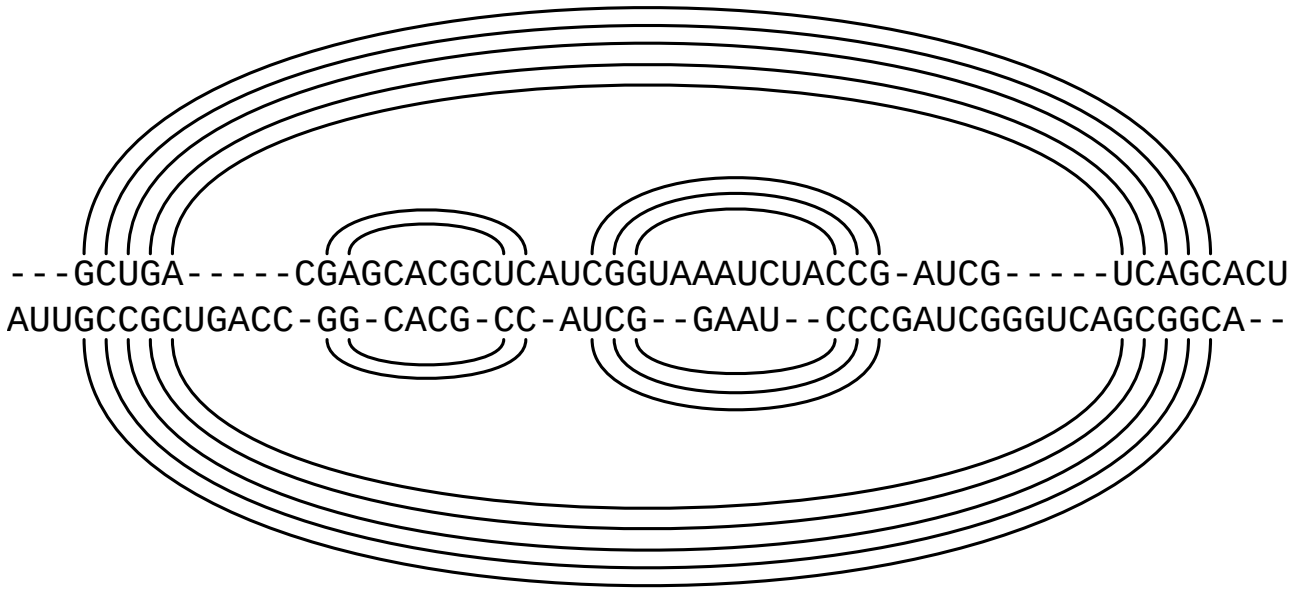


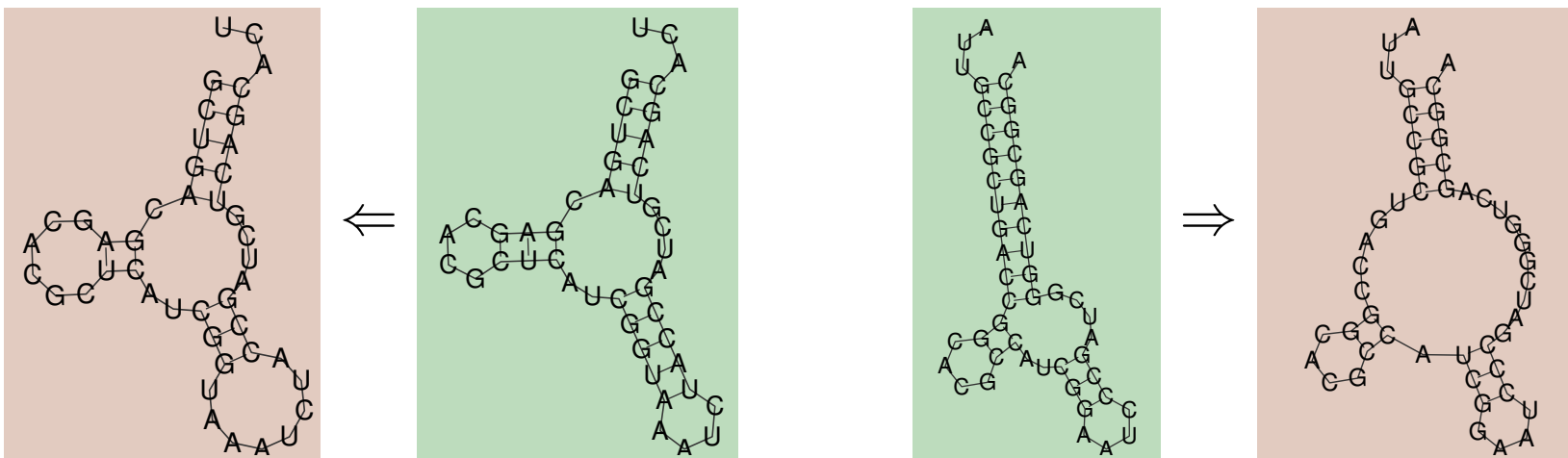
PMcomp – THE Lightweight Sankoff Algorithm?

compatibility

Sankoff: *same shape*

PMcomp: *all base pairs match*





PARSE — THE Lightweight Sankoff Algorithm

(PARSE = Prediction and Alignment of RNAs using Structure Ensembles)

- **lightweight** (PMcomp pseudo-energy)

  & **complete** (Sankoff's compatibility)

- "complete": allows base pair **indels**

We need "complete" for strong sparsification, please be patient.

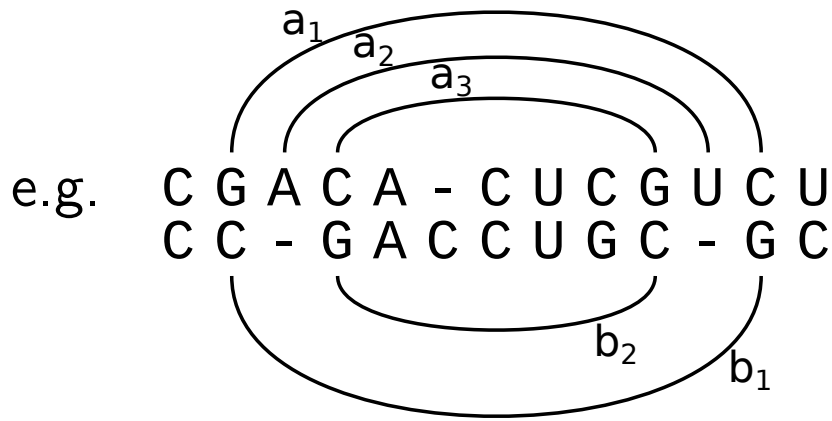
PARSE — THE Lightweight Sankoff Algorithm

(PARSE = Prediction and Alignment of RNAs using Structure Ensembles)

- **lightweight** (PMcomp pseudo-energy)

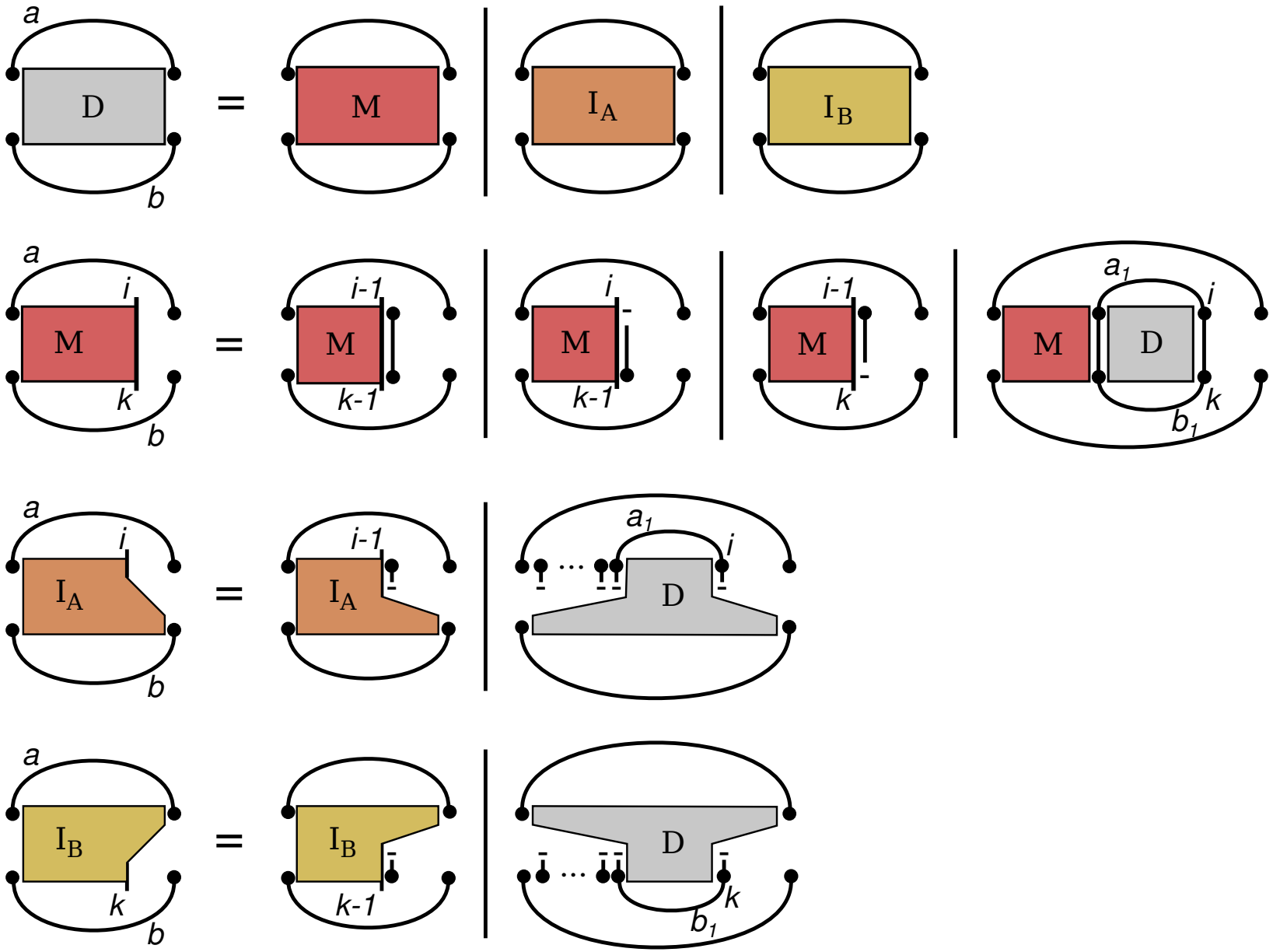
  & **complete** (Sankoff's compatibility)

- "complete": allows base pair **indels**

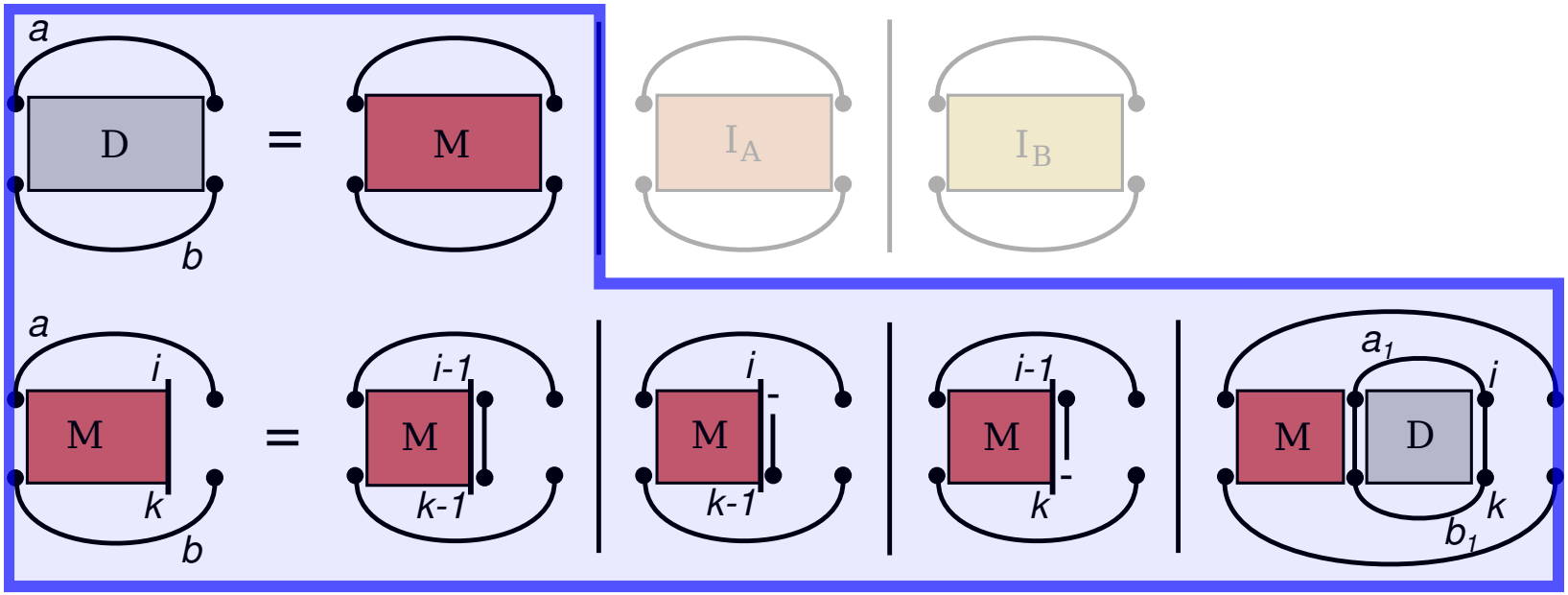


We need "complete" for strong sparsification, please be patient.

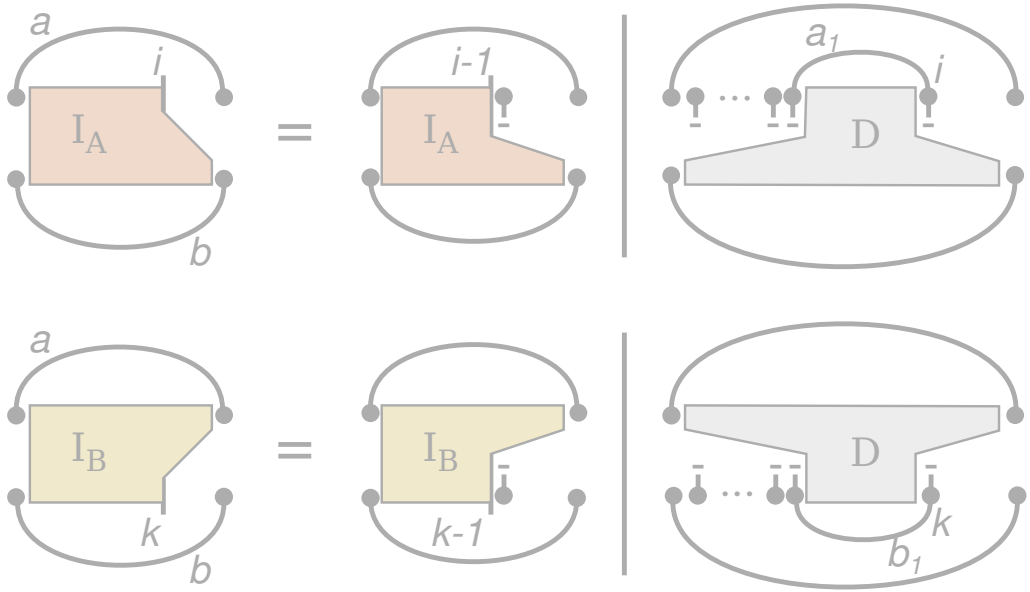
PARSE Algorithm



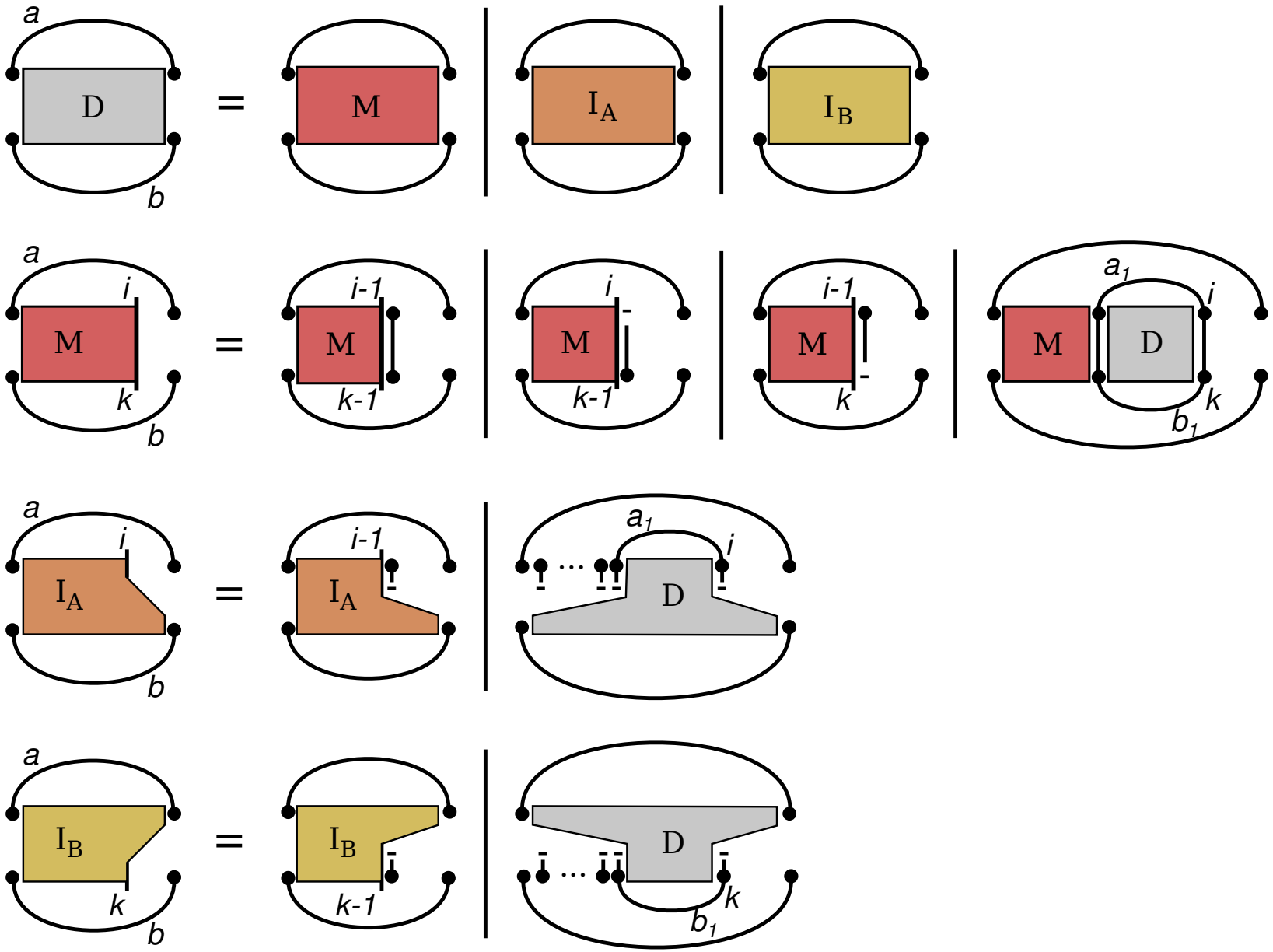
PARSE Algorithm



PMcomp/ LocARNA-like core

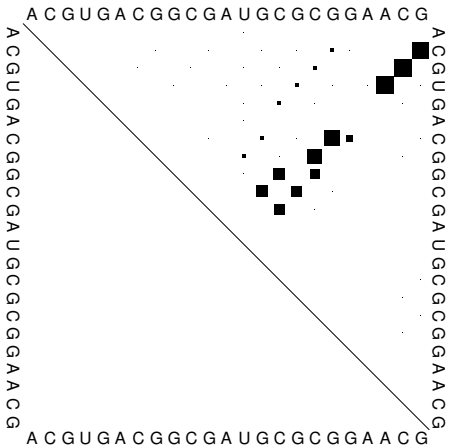


PARSE Algorithm



LocARNA's Trick: Ensemble-based Sparsification

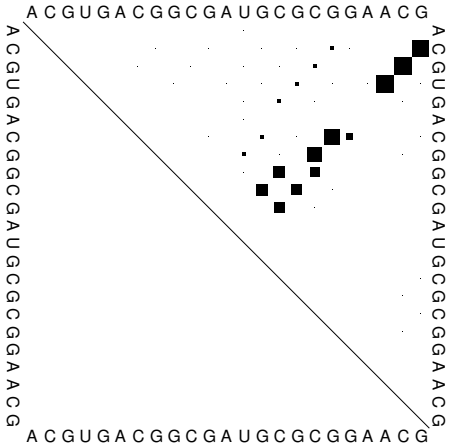
- **Sparsify structure ensemble**

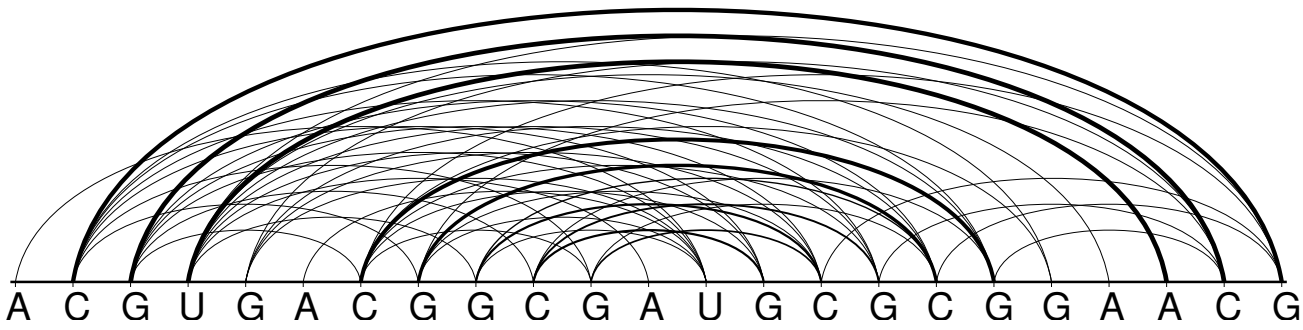


- improves time and space; each by $O(n^2)$

LocARNA's Trick: Ensemble-based Sparsification

- **Sparsify structure ensemble**



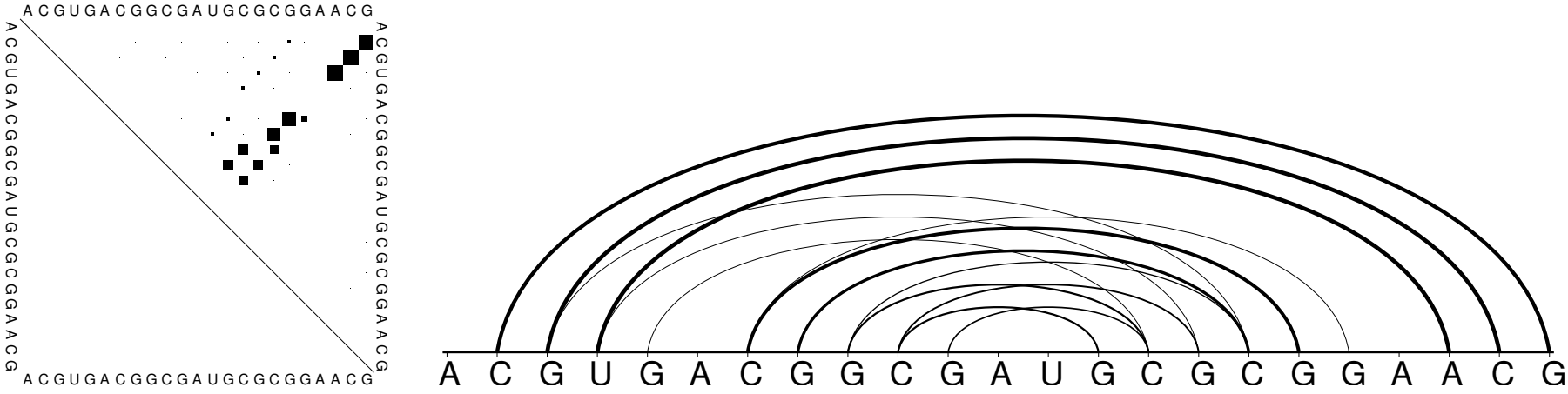


all base pairs

- improves time and space; each by $O(n^2)$

LocARNA's Trick: Ensemble-based Sparsification

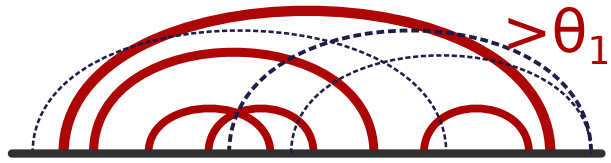
- **Sparsify structure ensemble**



only probable base pairs

- **improves time and space; each by $O(n^2)$**

SPARSE: Novel Ensemble-based Sparsification*



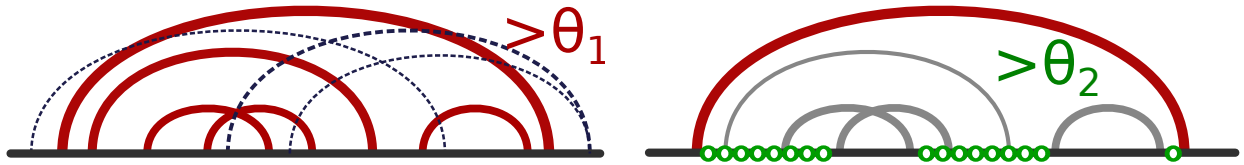
- only **base pairs** with probabilities $> \theta_1$
- only **bases** with unpaired probabilities in loops $> \theta_2$
- only **base pairs** with probabilities in loops $> \theta_3$

requires complete prediction (Sankoff/PARSE)

(*) confer LocARNA's "old" sparsification:

- match only base pairs with probabilities $> \theta_1$

SPARSE: Novel Ensemble-based Sparsification*



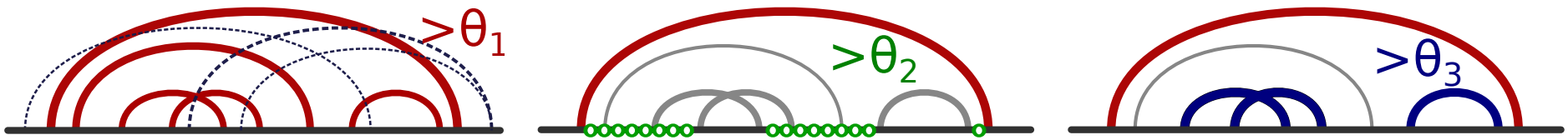
- only **base pairs** with probabilities $> \theta_1$
- only **bases** with unpaired probabilities in loops $> \theta_2$
- only **base pairs** with probabilities in loops $> \theta_3$

requires complete prediction (Sankoff/PARSE)

(*) confer LocARNA's "old" sparsification:

- match only base pairs with probabilities $> \theta_1$

SPARSE: Novel Ensemble-based Sparsification*



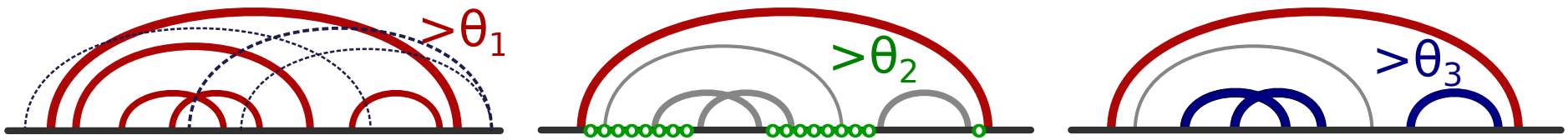
- only **base pairs** with probabilities $> \theta_1$
- only **bases** with unpaired probabilities in loops $> \theta_2$
- only **base pairs** with probabilities in loops $> \theta_3$

requires complete prediction (Sankoff/PARSE)

(*) confer LocARNA's "old" sparsification:

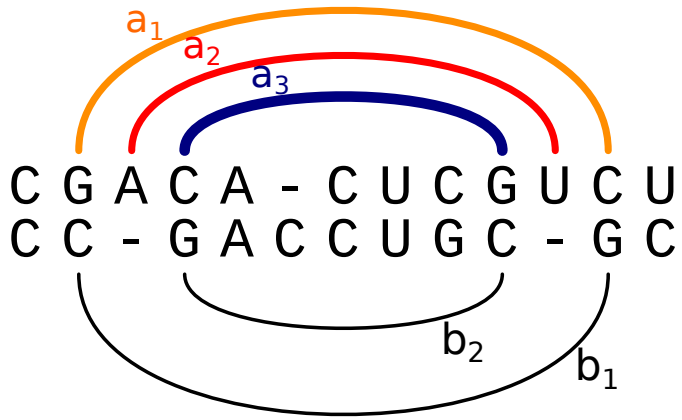
- match only base pairs with probabilities $> \theta_1$

SPARSE: Novel Ensemble-based Sparsification*



- only **base pairs** with probabilities $> \theta_1$
- only **bases** with unpaired probabilities in loops $> \theta_2$
- only **base pairs** with probabilities in loops $> \theta_3$

requires complete prediction (Sankoff/PARSE)

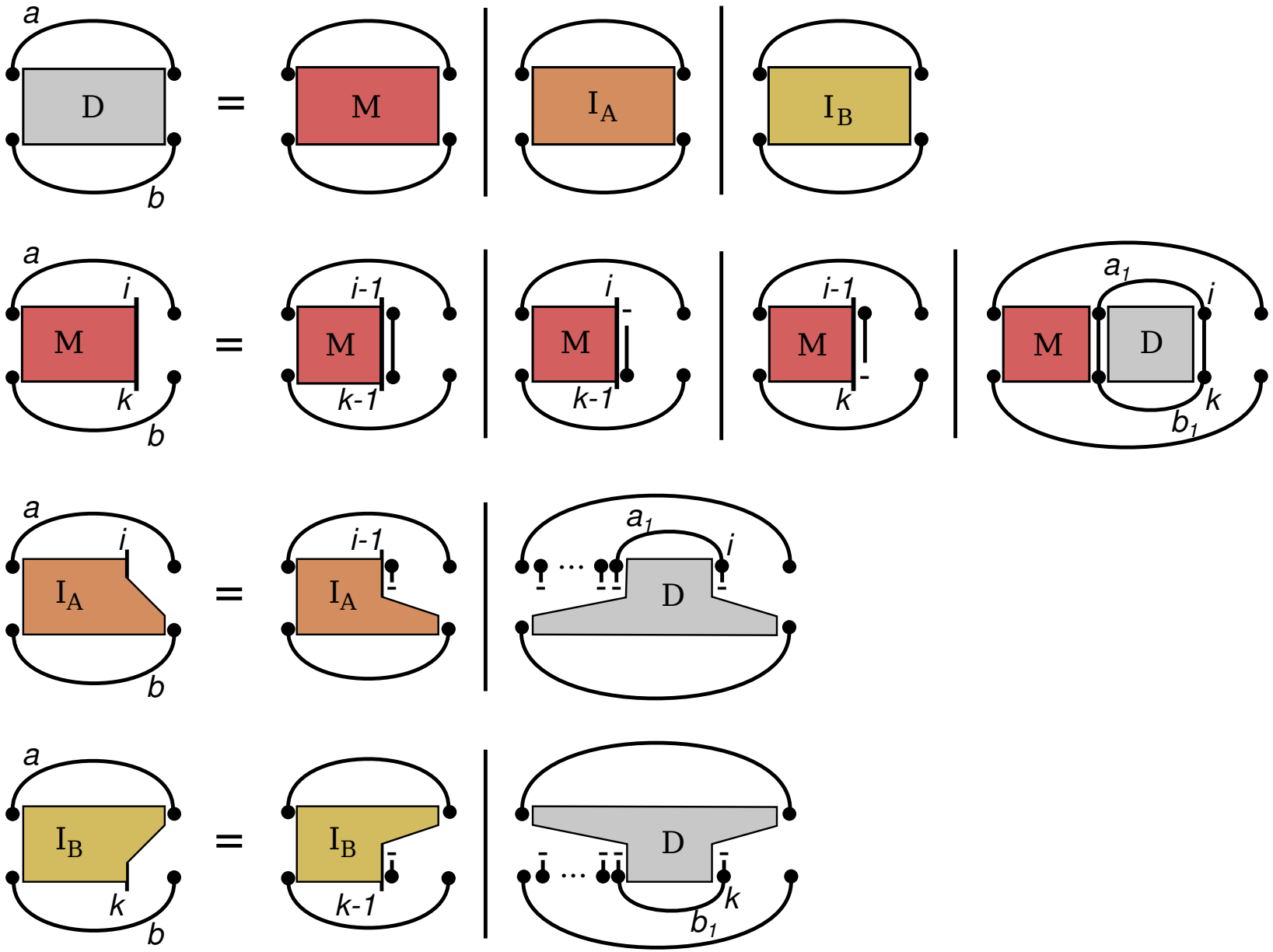


w/ complete: a_3 in loop a_2 ✓

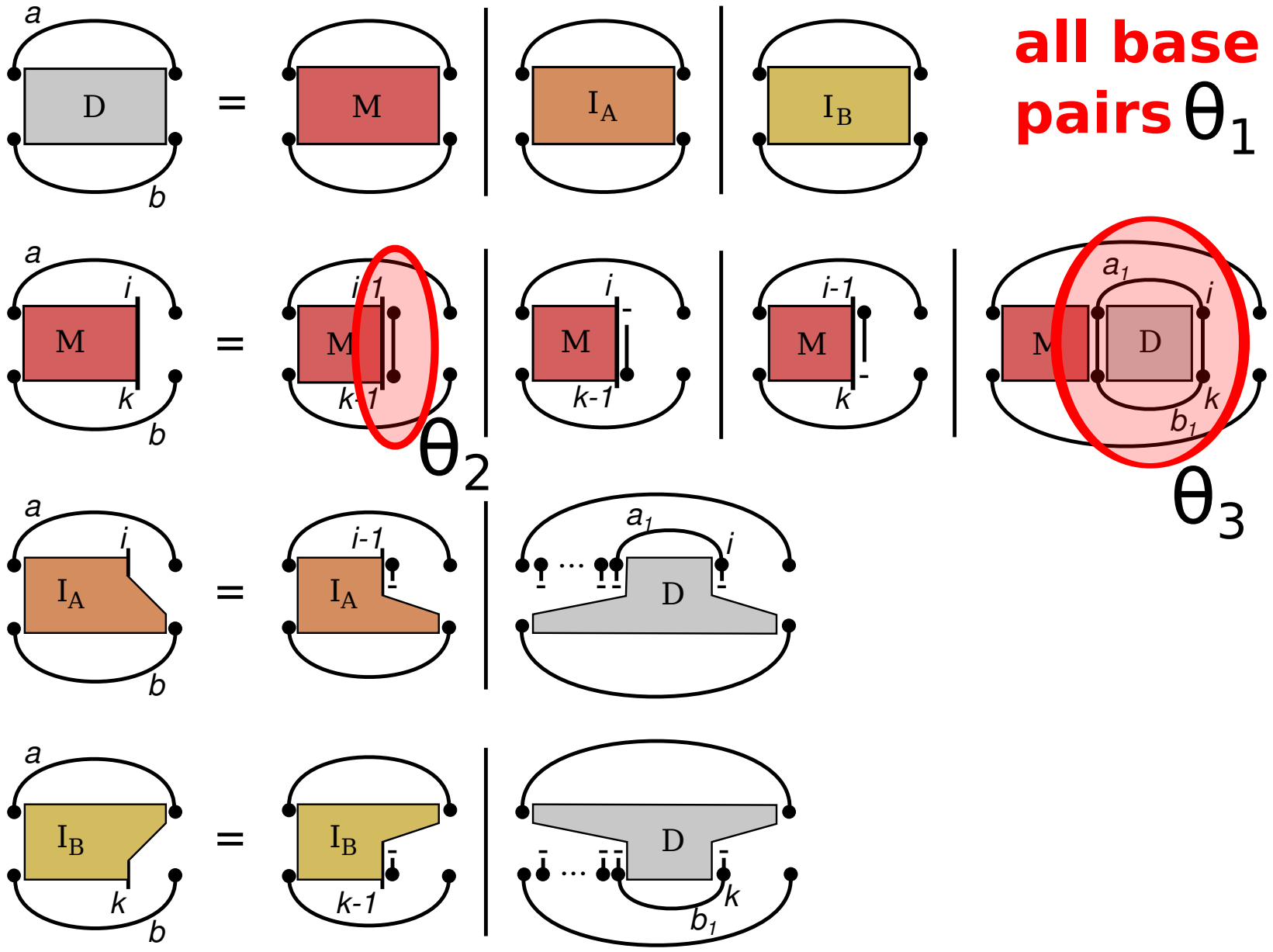
w/o complete: a_3 in loop a_1 ✗

a_2 ✗ $\Longrightarrow$ a_3-b_2 ✗

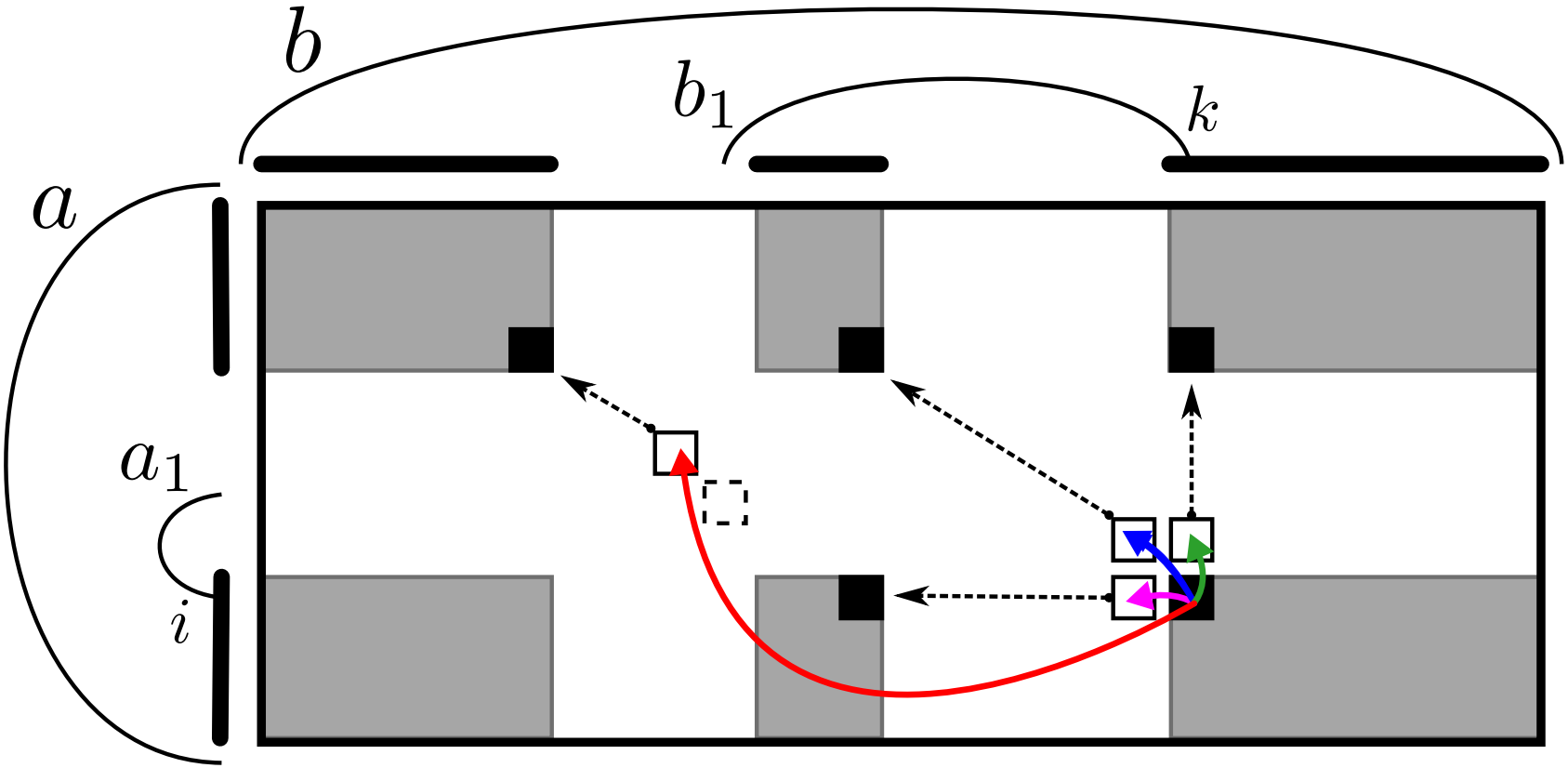
Thresholds in Recursions Cases

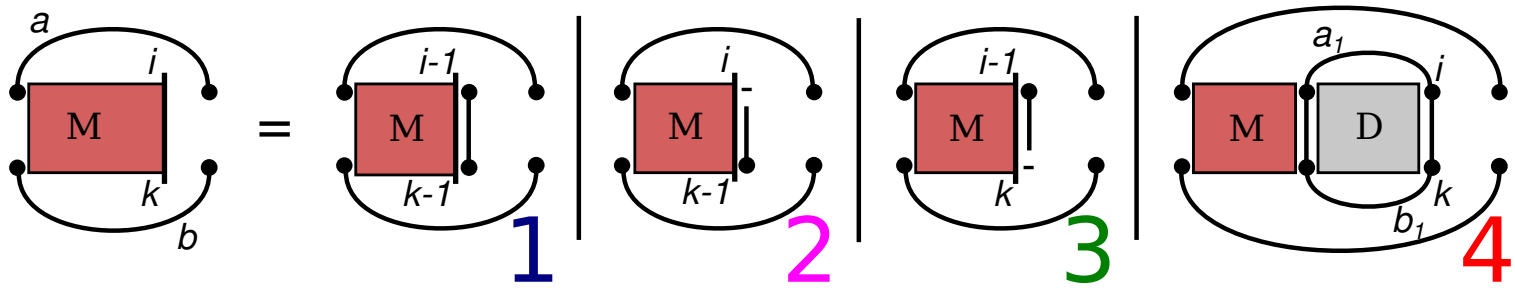


Thresholds in Recursions Cases

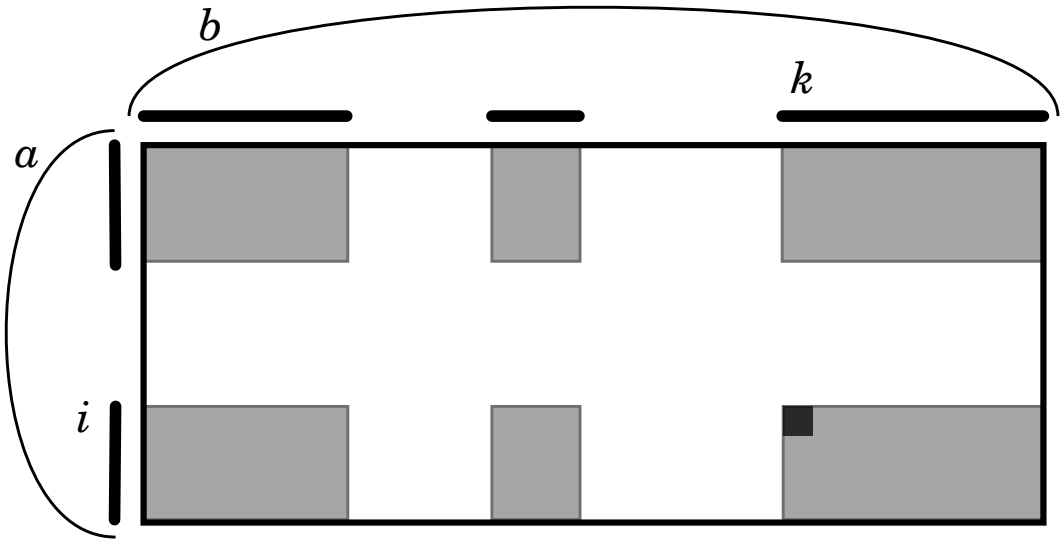


Modify Evaluation to Save Time



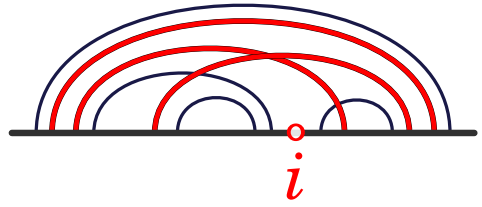


Quadratic Time



Q: How many matrices M^{ab} compute (i, k)?

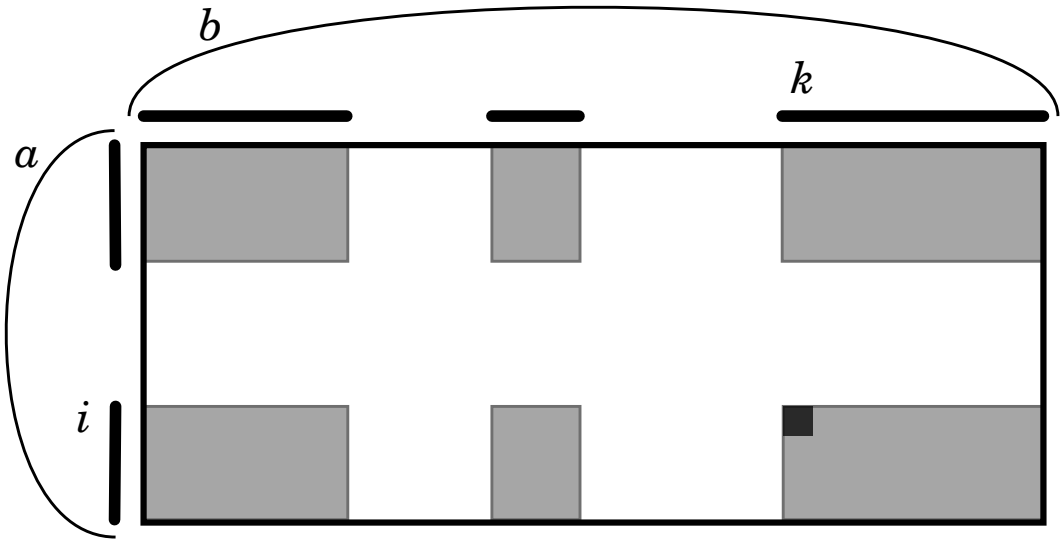
Count base pairs a where
$\Pr^A[i \text{ in loop of } a] > \theta_2$



$\Rightarrow$ less than $1/\theta_2$

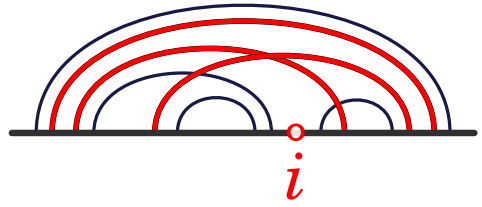
A: each (i, k) in only constant number of matrices $\square$

Quadratic Time



Q: How many matrices M^{ab} compute (i, k)?

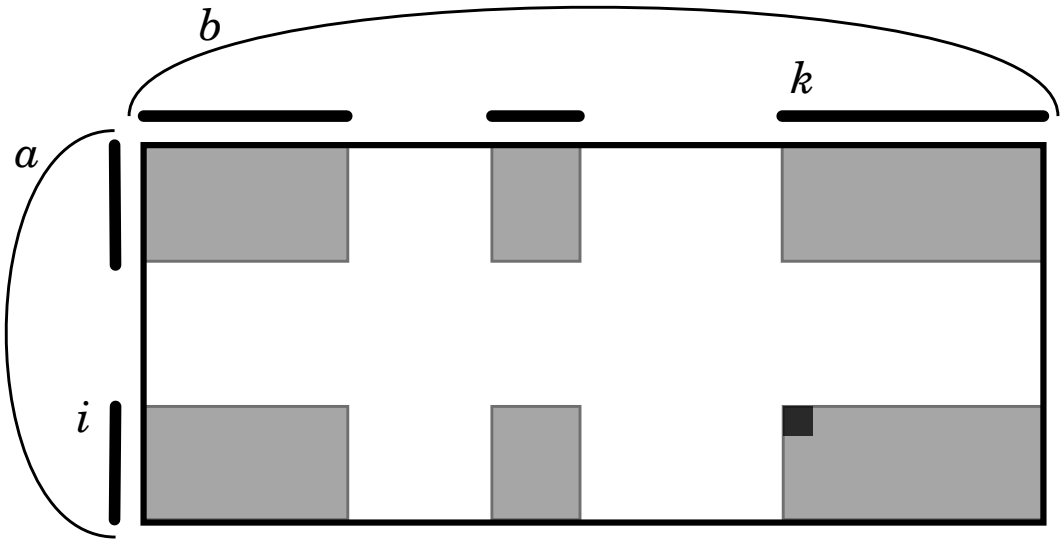
Count base pairs a where
$\Pr^A[i \text{ in loop of } a] > \theta_2$



$\Rightarrow$ less than $1/\theta_2$

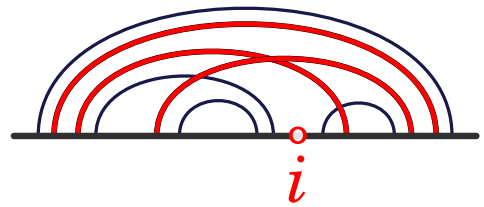
A: each (i, k) in only constant number of matrices □

Quadratic Time



Q: How many matrices M^{ab} compute (i, k)?

Count base pairs a where
$\Pr^A[i \text{ in loop of } a] > \theta_2$



$\Rightarrow$ less than $1/\theta_2$

A: each (i, k) in only constant number of matrices $\square$

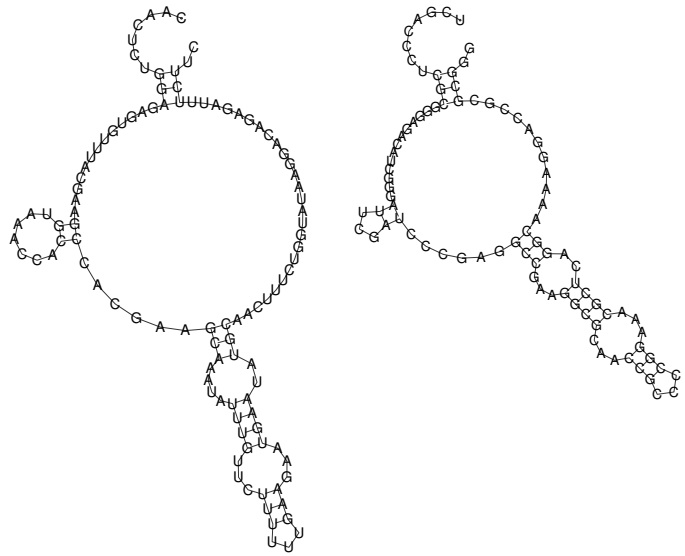
(S)PARSE improves prediction over LocARNA

LocARNA:

```
  -.......(((.............((.......)).....
A -CAACUCUGGAGAGUGUUUACGAAGGUAAACCACCCACGA
B UCGACCCUCGCGGGAGACAUCGGGAUU----CGAUCCCGA
  ........(((.............((.----..)).....

  .(((....((((...(((...)))...)))))..)))....
A AGCAAAUAUUUGUUCUUUUUUGAAGAAUGAAUAUGCAACU
B GGCCGA-AGGCGCAACCGCCCCGGAAACGCUCAGGCAA--
  .((((..-.((((...(((...)))...)))))..))))..--

  ......................))))..
A UUCUGGUAUAAGGACAGAGAUUUCUUC
B ---------AAGGACCG----CGCGGG
  ---------........----.))))..
```

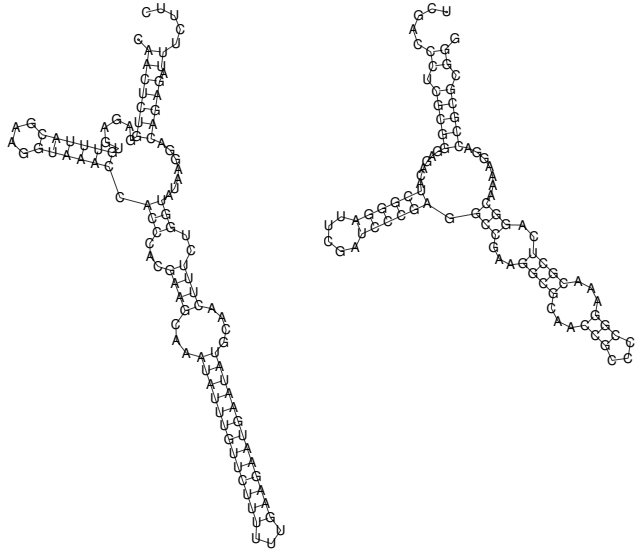


SPARSE:

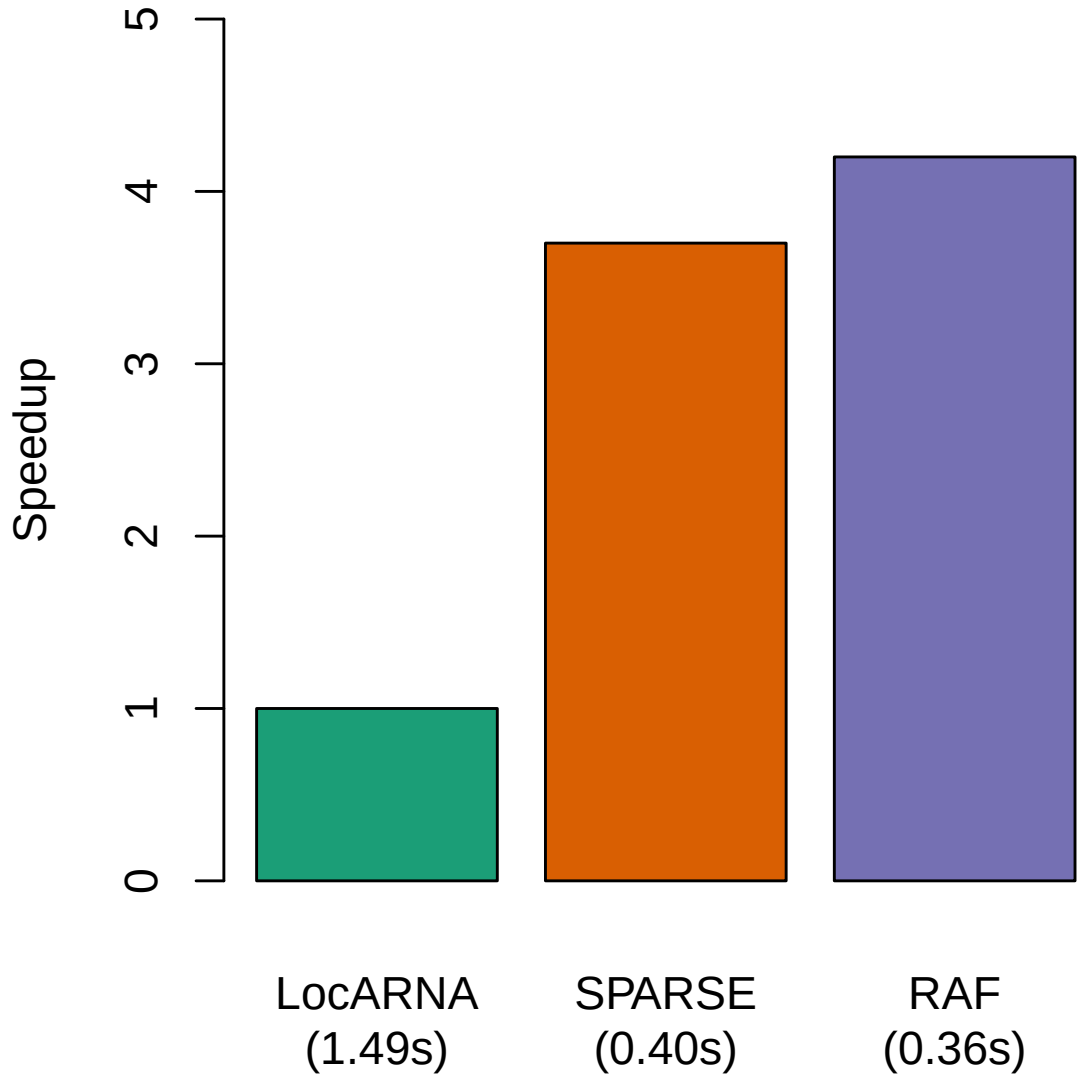
```
  ----.((-(((((......((((((....-))))))).(((
A ----CAA-CUCUGGAGAGUGUUUACGAAG-GUAAACCACC
B UCGACCCUCGCGGGAGACAUCGGGAUUCGAUCCCGAGGCC
  .....((.(((((......((((((.....)))))).(((

  ...((((...((((((((((((...))))))))))))).
A CACGAAGCAAAUAUUUGUUCUUUUUUGAAGAAUGAAUAUG
B GAAGGCGCAACCG___________CCC___________CGGA
  ...((((...(((___________..._____________))).

  ...)))))..)))........)))))..)).....
A CAACUUUCUGGUAUAAGGACAGAGAUUUCUUC
B -AACGCUCAGGCAAAAGGACCGCGCGGG----
  -..)))))..)))........)))))..)).----
```

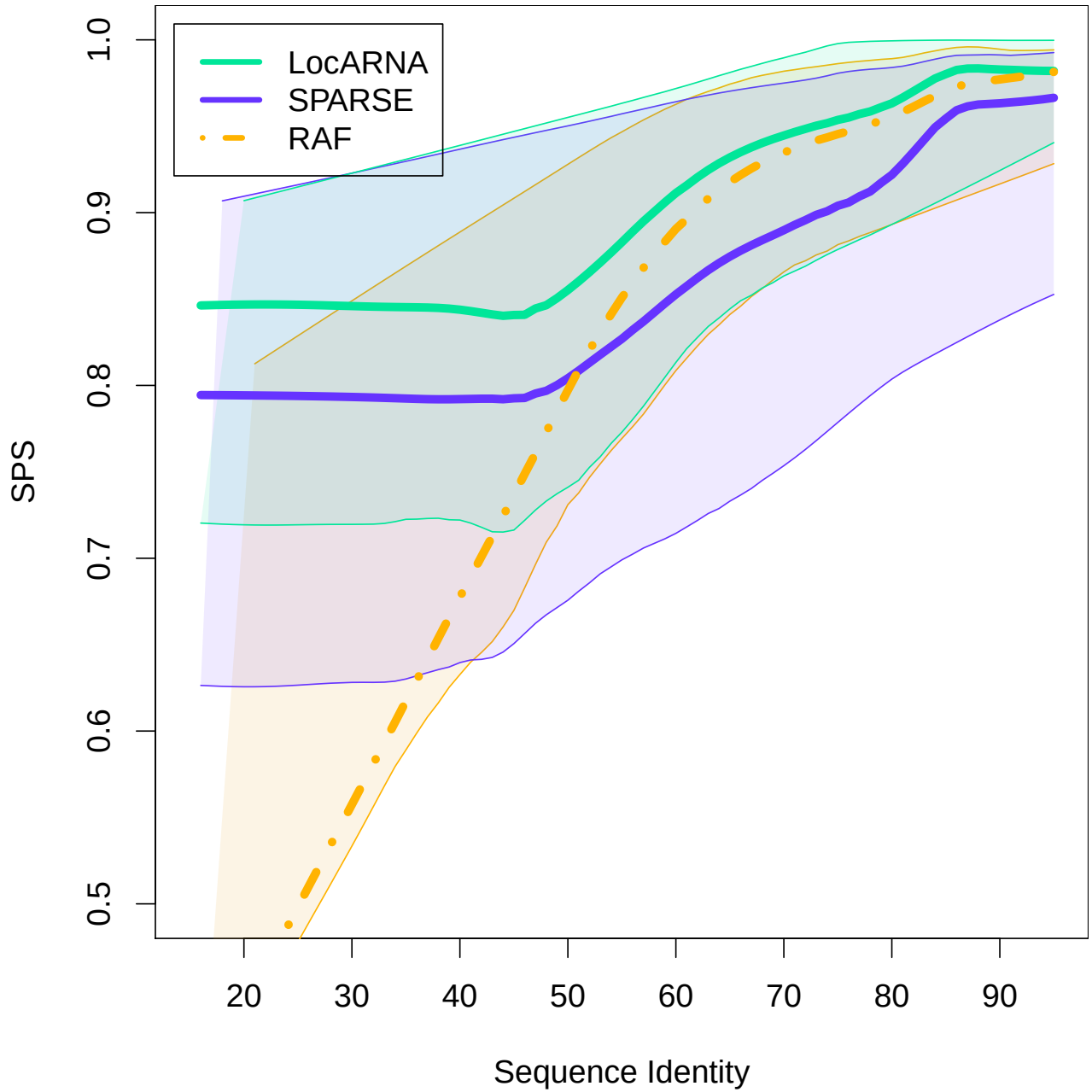


Run times and speedup

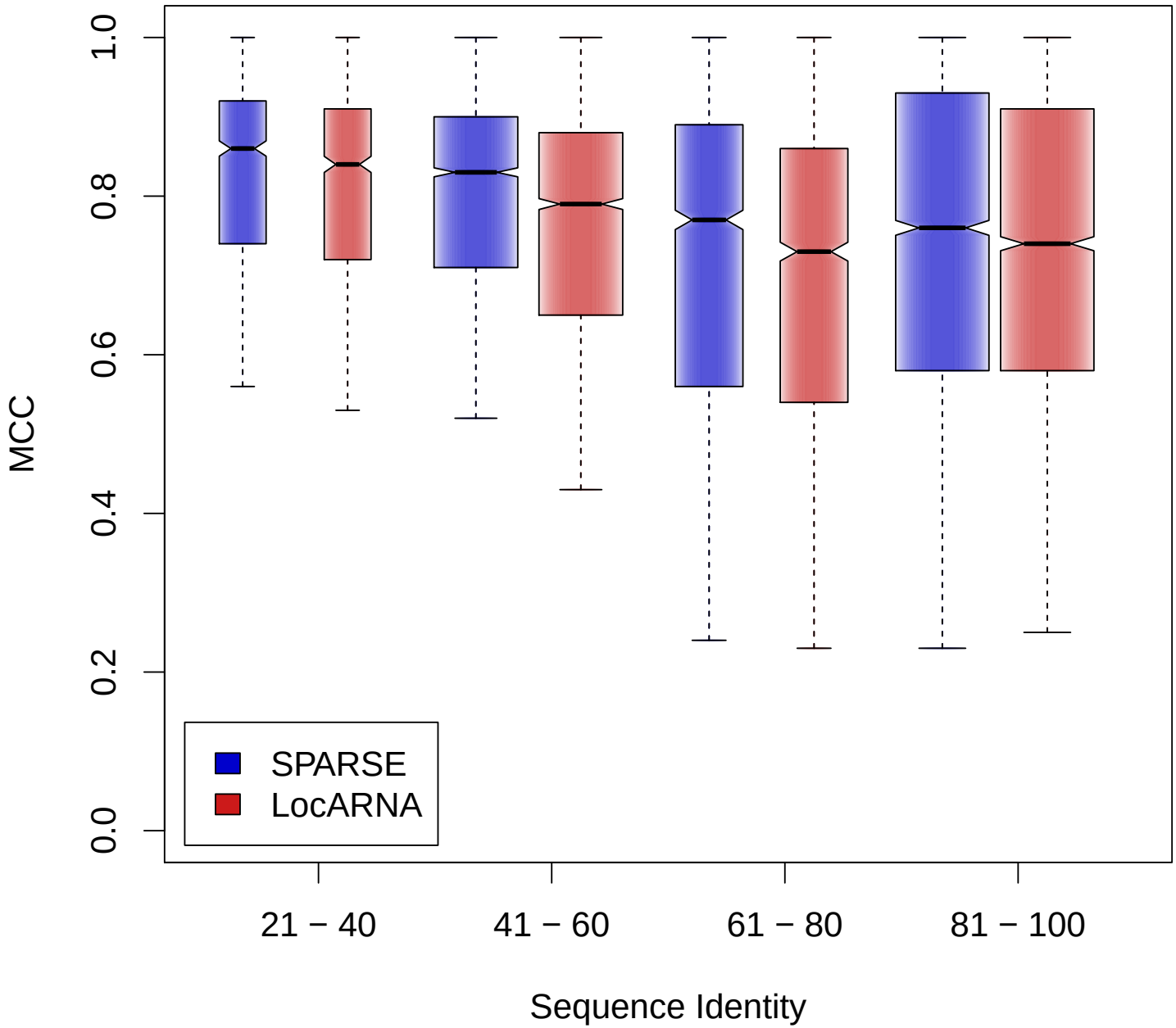


Bralibase 2.1, pairwise alignments (k2)

Alignment Accuracy (Bb 2.1, k2)



Structure Prediction Accuracy (BB 2.1, k2)



Conclusions

SPARSE: very efficient RNA alignment without sequence-based heuristics

- PARSE is *THE* lightweight Sankoff variant (cf. PMcomp)
  - predicts deleted/inserted base pairs; like original SA&F
- SPARSE = **Sparsified** PARSE
  - Novel **ensemble-based sparsification** (*in-loop* probabilities)
  - No sequence-based heuristics
  - Speeds up SA&F: **Quadratic Time** [$\leftarrow O(n^6)$]

`http://www.bioinf.uni-freiburg.de/Software/SPARSE/`

Conclusions

SPARSE: very efficient RNA alignment without sequence-based heuristics

- PARSE is ***THE*** lightweight Sankoff variant (cf. PMcomp)
  - predicts deleted/inserted base pairs; like original SA&F
- SPARSE = **Sparsified** PARSE
  - Novel **ensemble-based sparsification** (*in-loop* probabilities)
  - No sequence-based heuristics
  - Speeds up SA&F: **Quadratic Time** [$\leftarrow O(n^6)$]

`http://www.bioinf.uni-freiburg.de/Software/SPARSE/`

Conclusions

SPARSE: very efficient RNA alignment without sequence-based heuristics

- PARSE is ***THE*** lightweight Sankoff variant (cf. PMcomp)
  - predicts deleted/inserted base pairs; like original SA&F
- SPARSE = **Sparsified** PARSE
  - Novel **ensemble-based sparsification** (*in-loop* probabilities)
  - No sequence-based heuristics
  - Speeds up SA&F: **Quadratic Time** [$\leftarrow O(n^6)$]

http://www.bioinf.uni-freiburg.de/Software/SPARSE/

Thanks

. . . for your attention

. . . to my coauthors

- Christina Schmiedl 
- Milad Miladi 
- Mathias Möhl 
- Rolf Backofen 

. . . and the German Research Foundation DFG

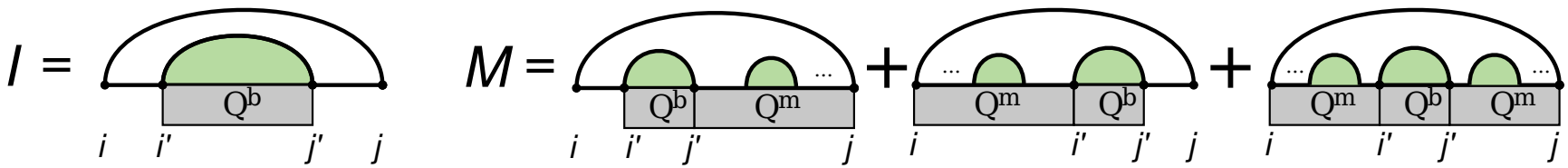
Appendix

Computing "In Loop" Probabilities

from McCaskill matrices: Q_b, Q_m

Pr[(i',j') base pair in loop of (i,j)]

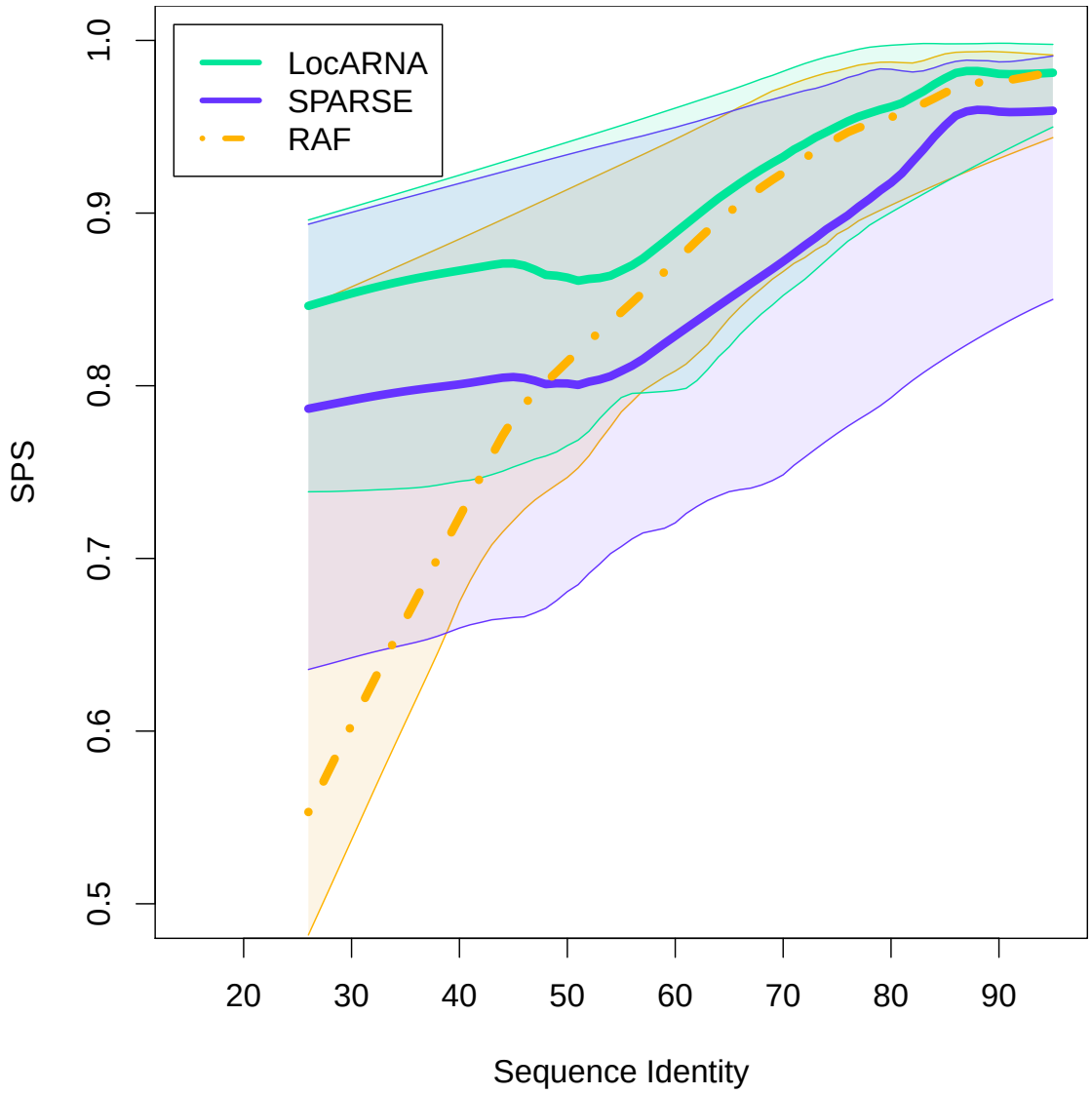
$$=(I+M)/Q$$



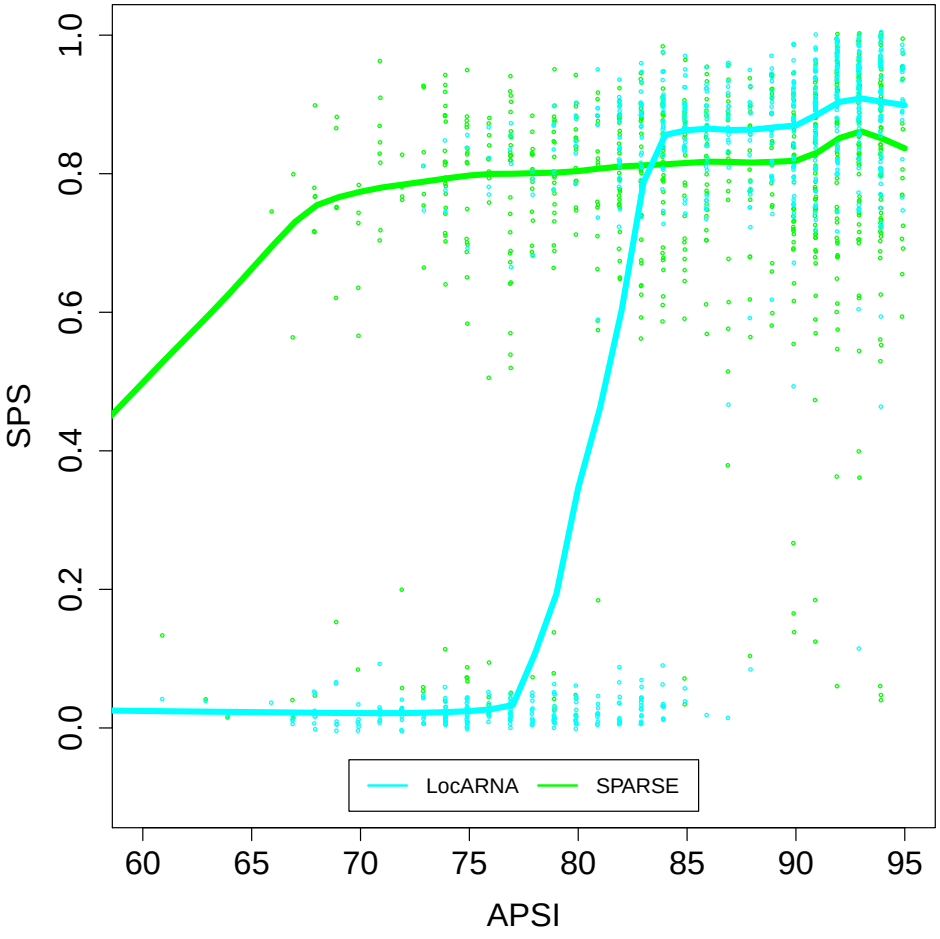
similar: **Pr[k unpaired in loop of (i,j)]**

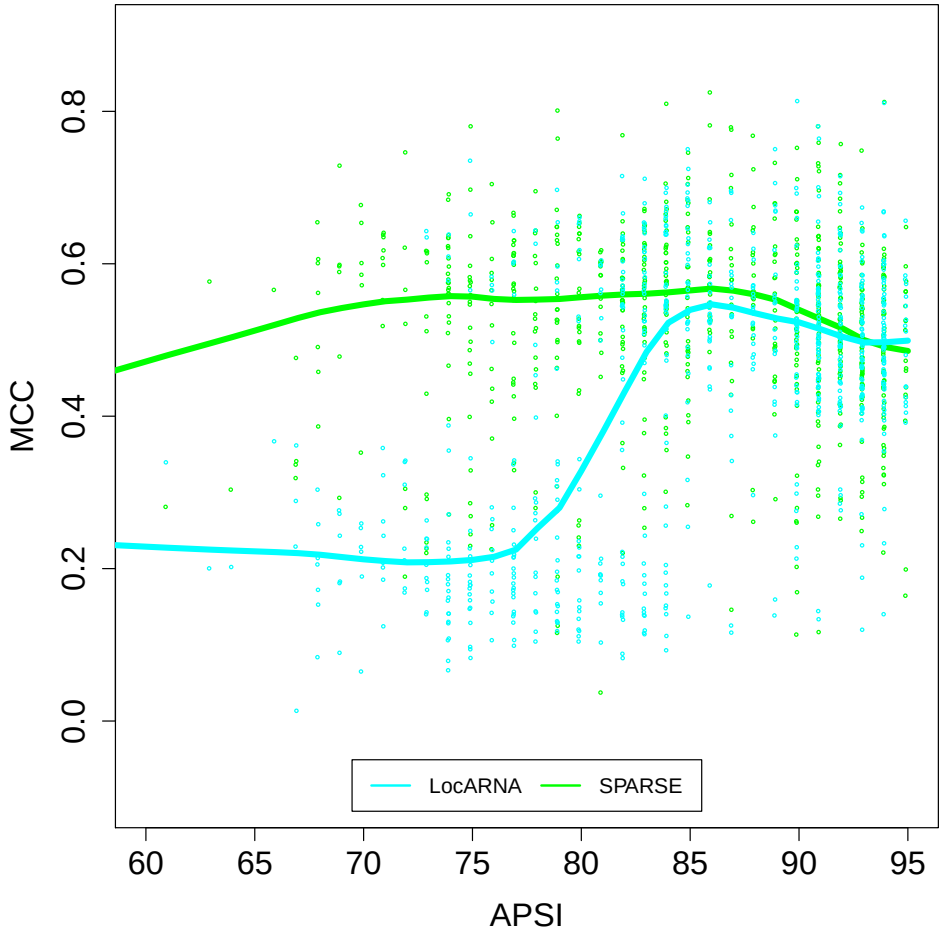
[ExpARNA-P; Schmiedl et al., BMC Bioinformatics 2014]

Alignment and Prediction Accuracy (Bralibase 2.1, 3-way alignments)



SPARSE Improves Over LocARNA for Specific Families





(shown: IRES HCV, pairwise)