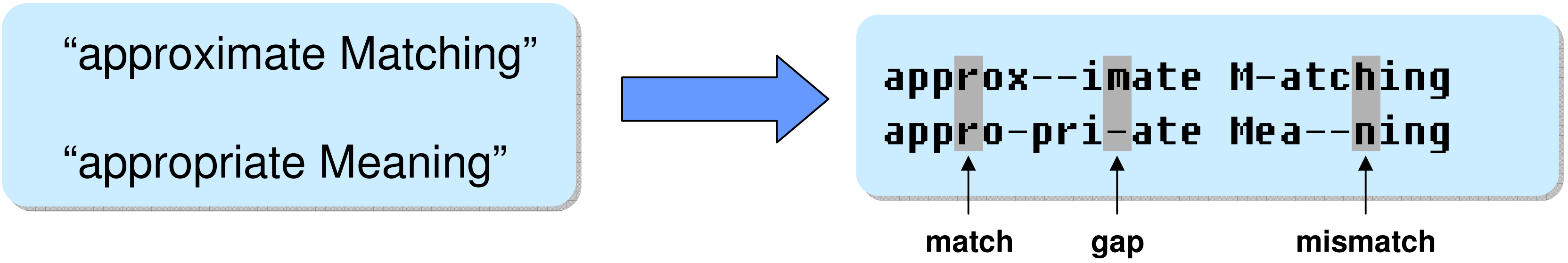
Lifting the Prediction of RNA Pseudoknots to their Alignment

by Mathias Möhl
Programming Systems Lab., Saarland University, Prof. Smolka

joint work with Rolf Backofen and Sebastian Will
Chair for Bioinformatics, Albert-Ludwigs-Universität Freiburg

Alignment

sequence alignment: comparison of sequences (RNAs, DNAs, Proteins)



"approximate Matching"

"appropriate Meaning"

```
approx--imate M-atching
appro-pri-ate Mea--ning
```

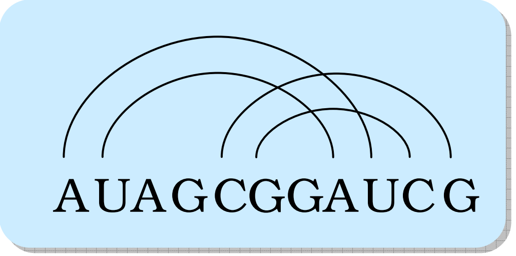
structure alignment: also consider secondary structure in comparison

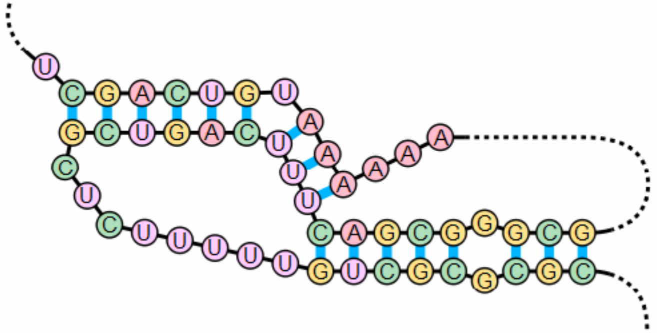
```
..((( (((((((... ..........))))))) )))....
ACGUG-UGAUGGGAGG-UACAAGCAACCCCAUUA-CAUAUUA
UGCUGCU-AUAAUA-GAUAGA-GA-AGGUUAU-AGCAGACUA
..((((( (((((. ...... .. ..))))) )))))....
```

Pseudoknots

an RNA structure contains pseudoknots

...if it contains any crossing arcs





...if it is not well-bracketed

```
((..[[.))]]
AUAGCGGAUCG
```

structures *without* pseudoknots can easily be decomposed

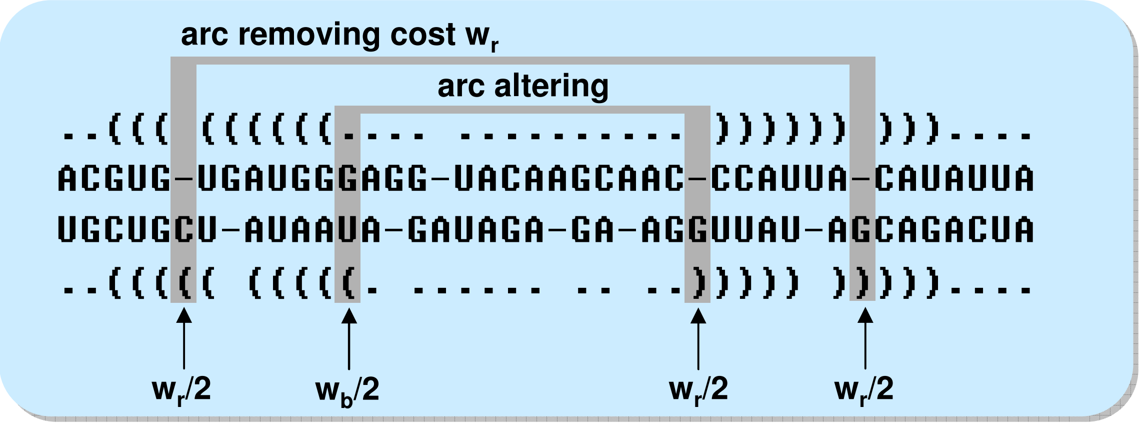
→efficient dynamic programming algorithms

more difficult for pseudoknots

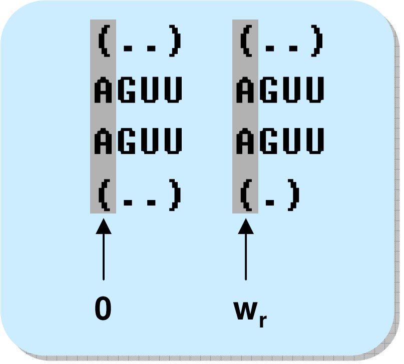
Optimal Alignment

optimal alignment ≈ minimum edit distance

edit operation	**cost**	**example**
base deletion	w_d	. A → –
base substitution	w_m	. . A → C
arc mismatch	$w_{am}/2$ per mismatched base	( ) ( ) A U → C G
arc breaking	w_b	( ) . . A U → A U
arc removing	w_r	( ) A U → – –
arc altering	$w_a= (w_b+w_r)/2$	( ) . A U → A –



```
..((( ((((((.... ...........)))))) )))....
ACGUG-UGAUGGGAGG-UACAAGCAAC-CCAUUA-CAUAUUA
UGCUGCU-AUAAUA-GAUAGA-GA-AGGUUAU-AGCAGACUA
..((((( (((((. ...... .. ..))))) )))))....
```



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

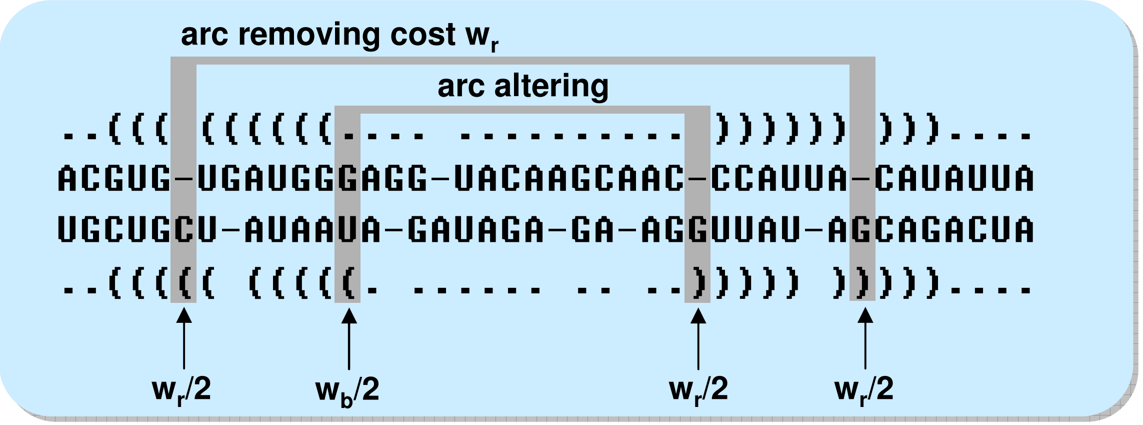
Optimal Alignment

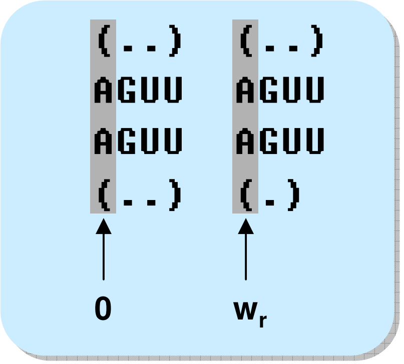
optimal alignment ≈ minimum edit distance

edit operation	cost	example
base deletion	w	
base subs		
arc misma		
arc breaki		
arc remov		
arc alterin		

cost of preserving one end of arc depends on other end

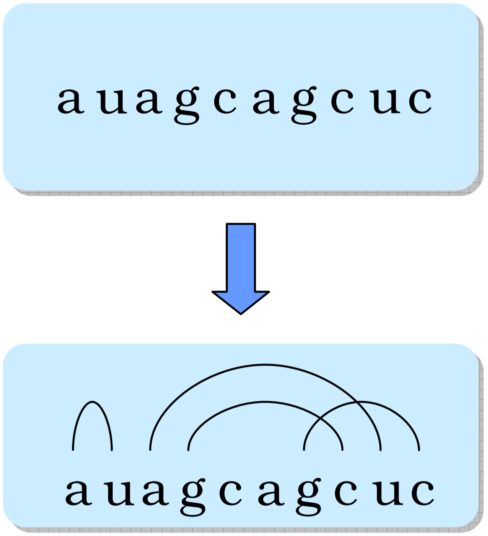
all other costs are independent





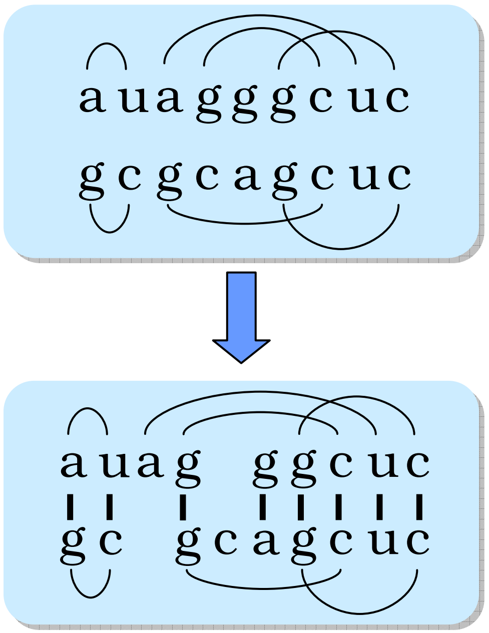
Structure Prediction vs. Alignment

structure prediction



optimization problem:
minimum free energy structure

sequence structure alignment



optimization problem:
minimum edit distance

both solvable with dynamic programming (DP)
where recursive decomposition is based on secondary structure

What has been done before

class of structures		pseudoknot free	R&G	A&U	L&P	D&P	R&E	arbitrary pseudoknots
prediction		[Zucker1981] [Sankoff1985]	[Reeder2004]	[Uemura1999] [Akutsu2000]	[Lyngso2000]	[Dirks2003]	[Rivas1999]	[Lyngso2000]
	time	$O(n^3)$	$O(n^4)$	$O(n^4)$	$O(n^5)$	$O(n^5)$	$O(n^6)$	NP-complete
	space	$O(n^2)$	$O(n^2)$	$O(n^3)$	$O(n^3)$	$O(n^4)$	$O(n^4)$	
alignment	time	$O(n^2m^2)$					$O(n^5m^5)$	NP-complete
	space	$O(nm)$					$O(n^4m^4)$	
		[Jiang2002]					[Evans2006]	[Evans1999]
		$\left(O(mn^3)\right)$	$O(mn^4)$	$O(mn^4)$	$O(mn^5)$	$O(mn^5)$	$O(mn^6)$	
		$\left(O(mn^2)\right)$	$O(mn^2)$	$O(mn^3)$	$O(mn^3)$	$O(mn^4)$	$O(mn^4)$	

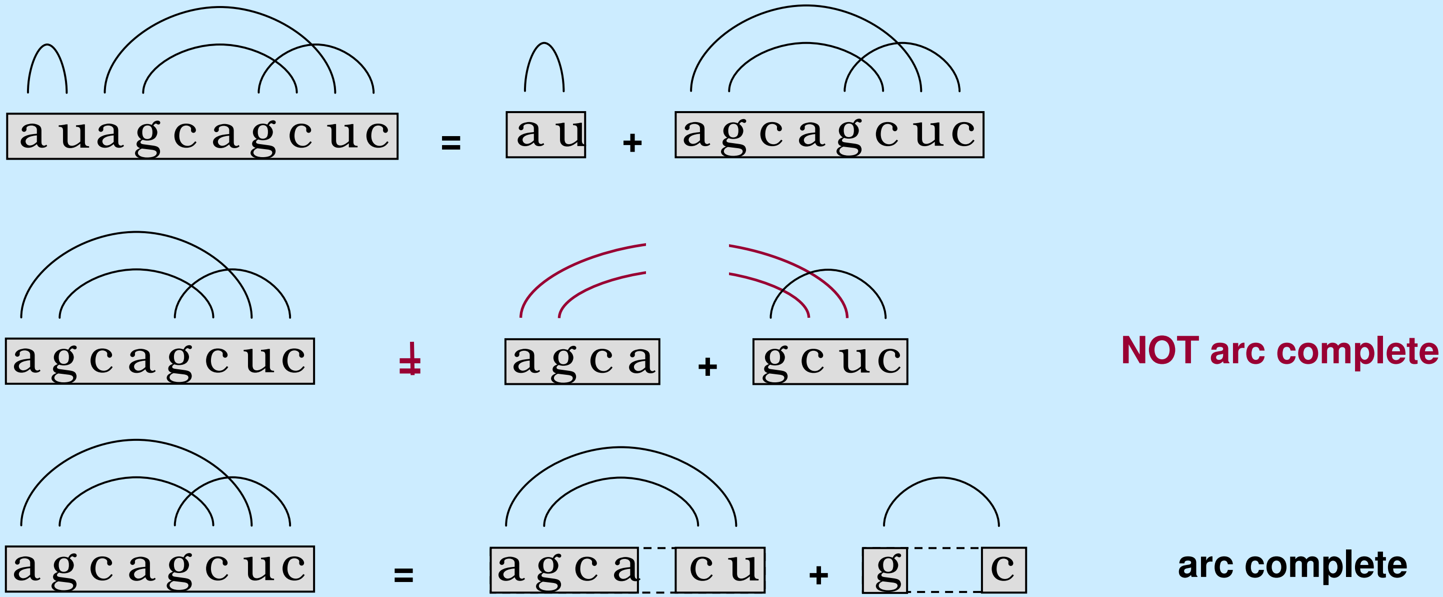
new contribution

further alignment approaches:

- fixed parameter tractable DP algorithms for arbitrary pseudoknots [Evans1999],[Möhl2008]
- approaches based on integer linear programming (ILP) [Lenhof1998],[Bauer2007],...

RNA Structure Prediction

General Framework compose optimal structure recursively from *arc complete* fragments



- common base of all DP-based RNA alignment and structure prediction algorithms algorithms
- algorithms vary in
  - kinds of considered fragments
  - ways how to combine fragments

→

trade-off

complexity ↔ class of structures

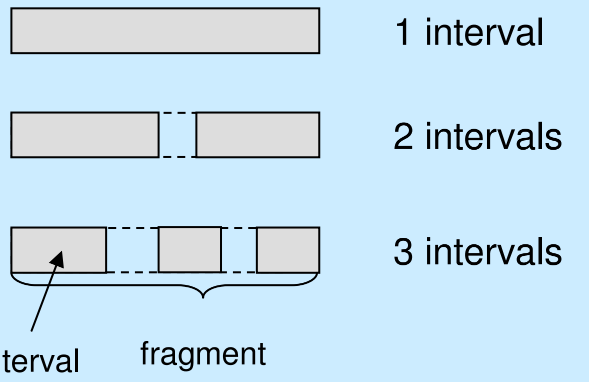
RNA structure prediction

formal notion for that

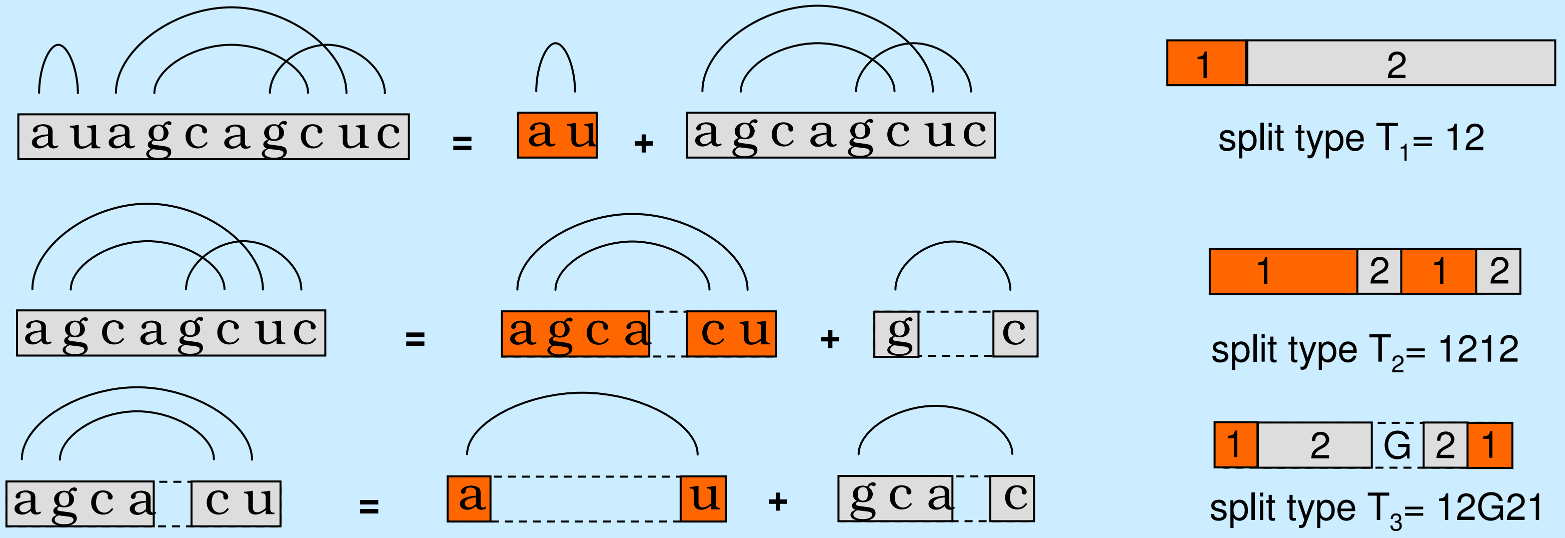
algorithms vary in

- kinds of considered fragments
- ways how to combine fragments

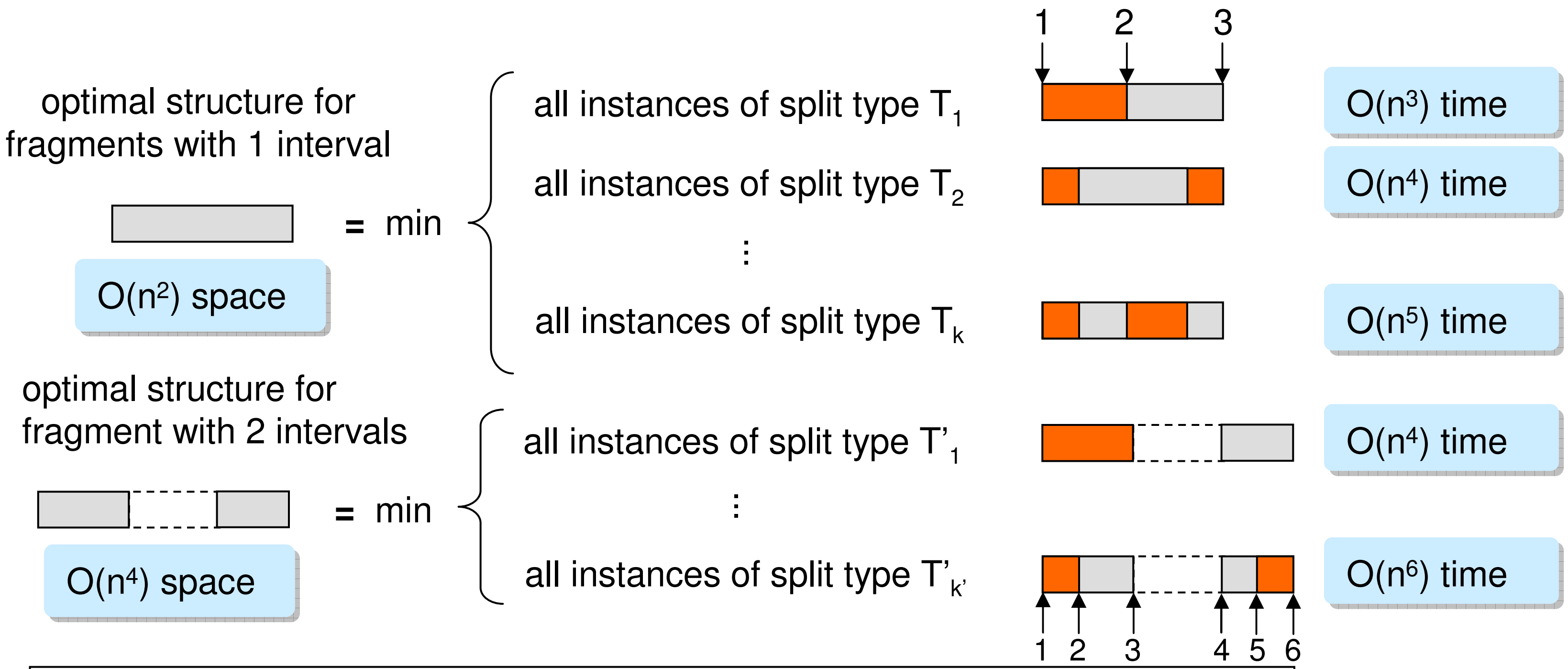
number of intervals of a fragment



split type



A Structure Prediction Algorithm Scheme



optimal structure for fragments with 1 interval

= min

O(n^2) space

all instances of split type T_1 — O(n^3) time

all instances of split type T_2 — O(n^4) time

⋮

all instances of split type T_k — O(n^5) time

optimal structure for fragment with 2 intervals

= min

O(n^4) space

all instances of split type T'_1 — O(n^4) time

⋮

all instances of split type $T'_{k'}$ — O(n^6) time

correctness

correct for the class of structures that can be composed that way
(→ for which the case distinction is exhaustive)

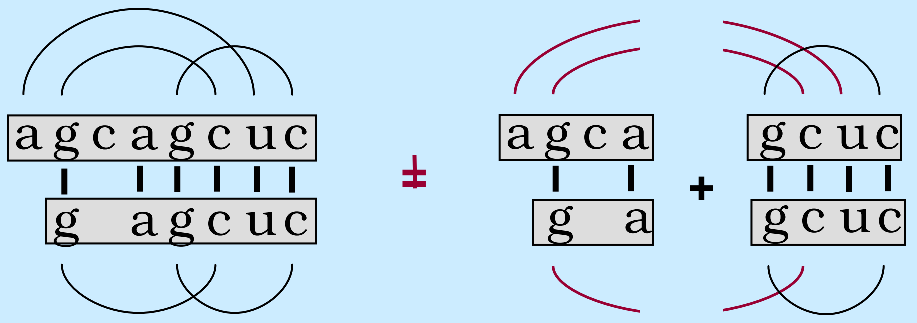
complexity

time complexity: depends on number of instances of considered split types
space complexity: depends on degree of considered fragments

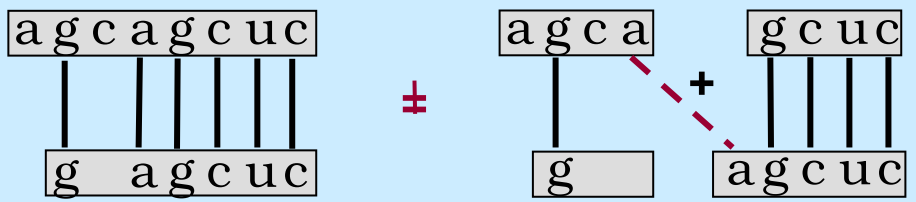
Alignment

recursive split of alignments: **recursive split of 2 sequences simultaneously**

split must be arc complete and compatible with alignment



NOT arc complete
→ cost not independent



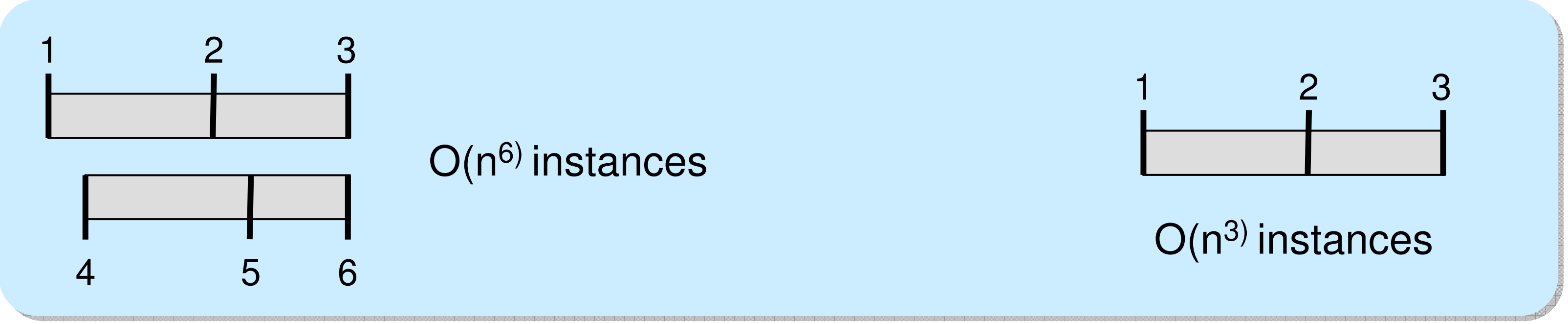
NOT compatible with alignment

→ **Idea:** search over all possible splits

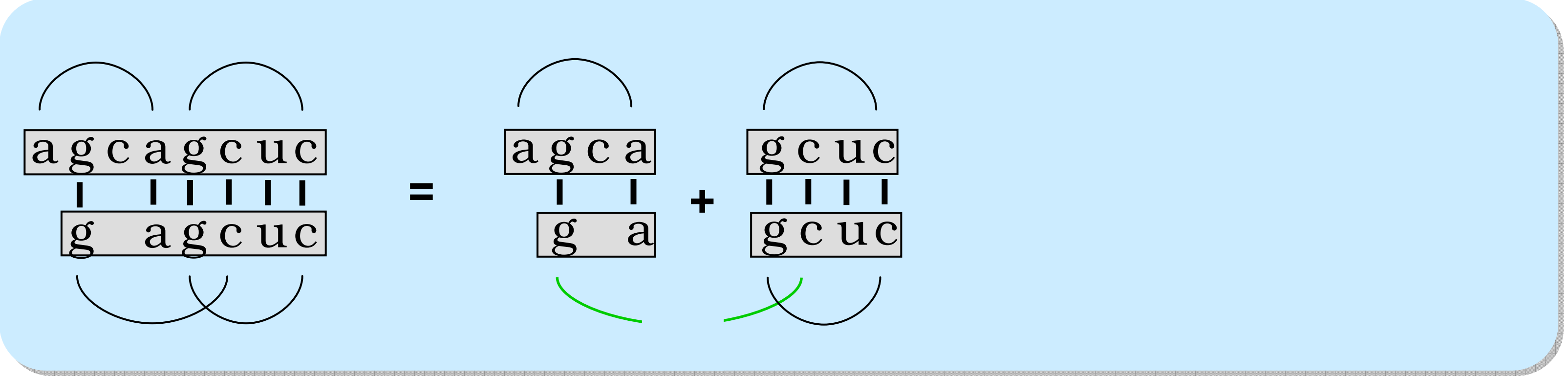
Alignment vs. Prediction

Idea: search over all possible splits

bad news: there exist a lot more possibilities to split an alignment



good news: breaking arcs in only one sequence is ok

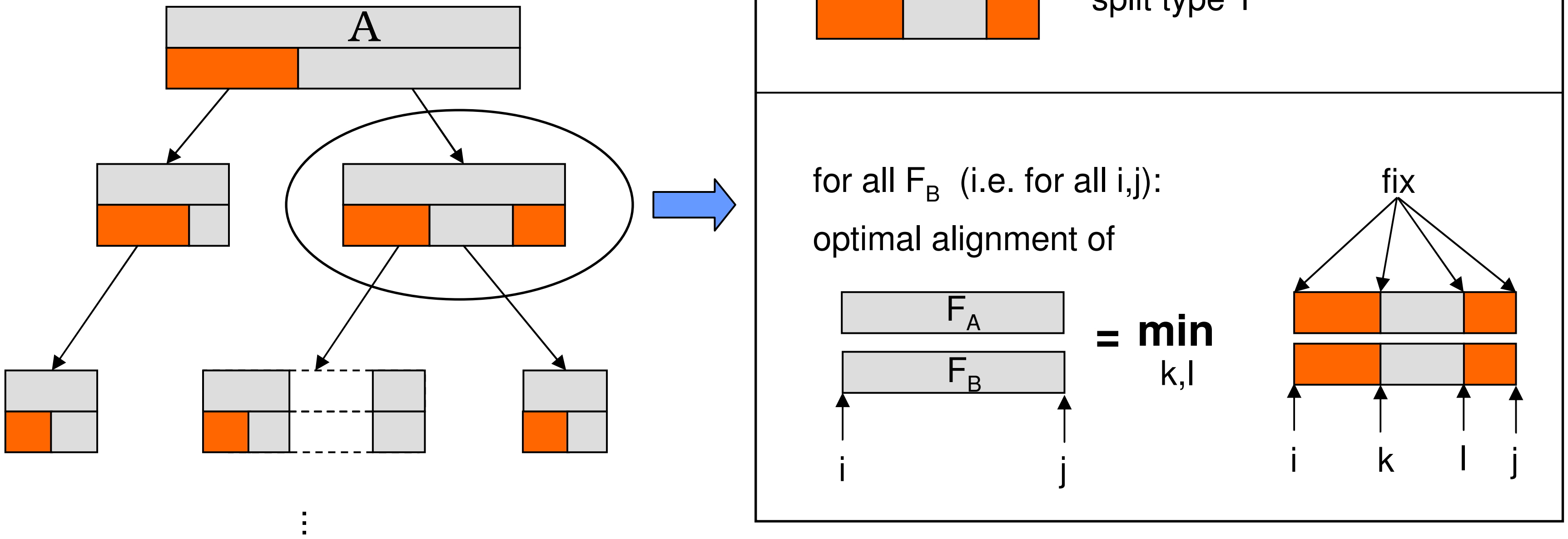


Alignment Algorithm

idea align RNAs A and B as follows:

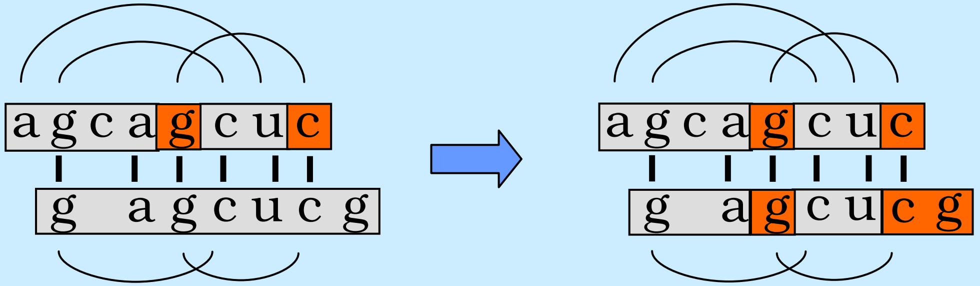
consider a fixed decomposition of A , try to align it to all possible decompositions of B

fixed decomposition of A:

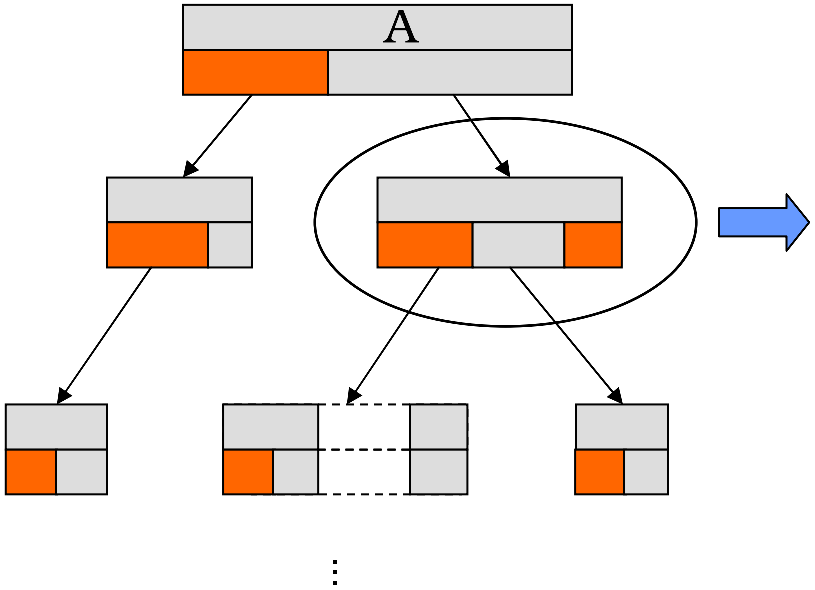


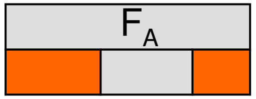
Alignment Algorithm

for each decomposition of the first sequence there exists an appropriate decomposition of the second sequence



fixed decomposition of A:

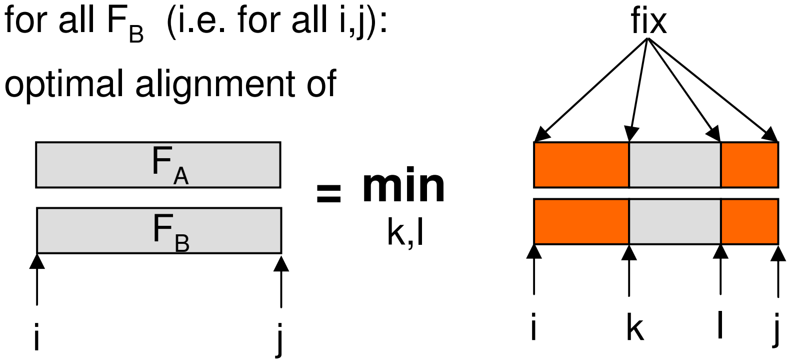




split type T

for all F_B (i.e. for all i,j):

optimal alignment of

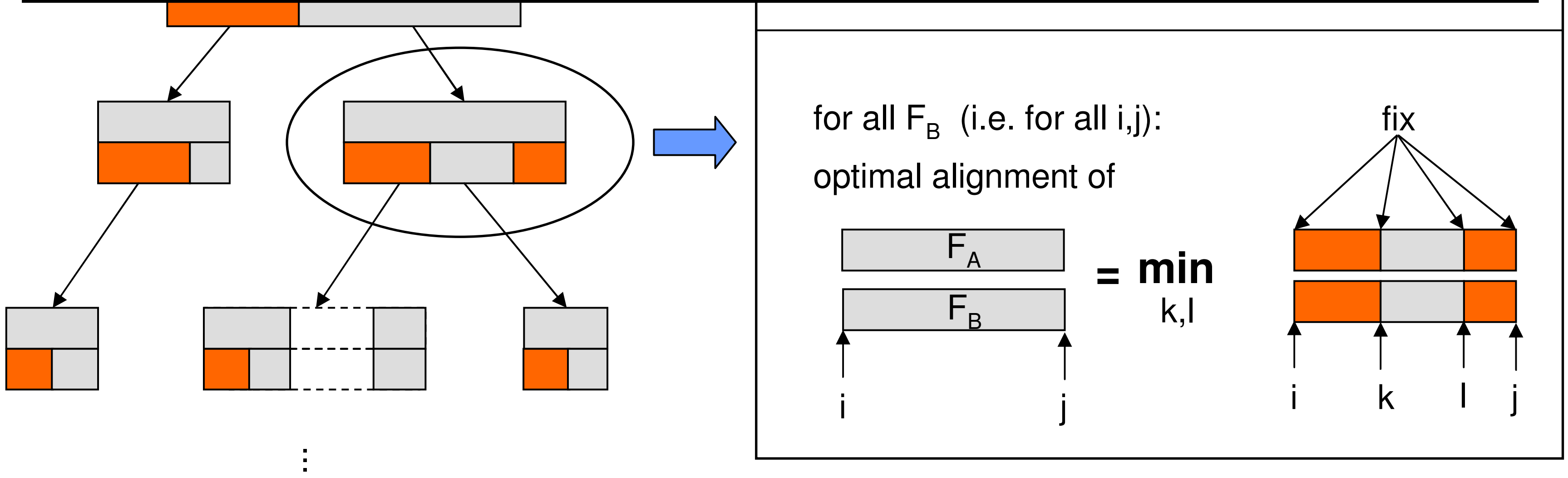


Alignment Algorithm

more general:

for any fragments F_A and F_B, and any arc complete split $(F_A{}^1, F_A{}^2)$ of F_A with some type T

$$C(F_A, F_B) = \min_{T-\text{split }(F_B^1, F_B^2)\text{ of }F_B} \left\{ C(F_A^1, F_B^1) + C(F_A^2, F_B^2) \right\}$$



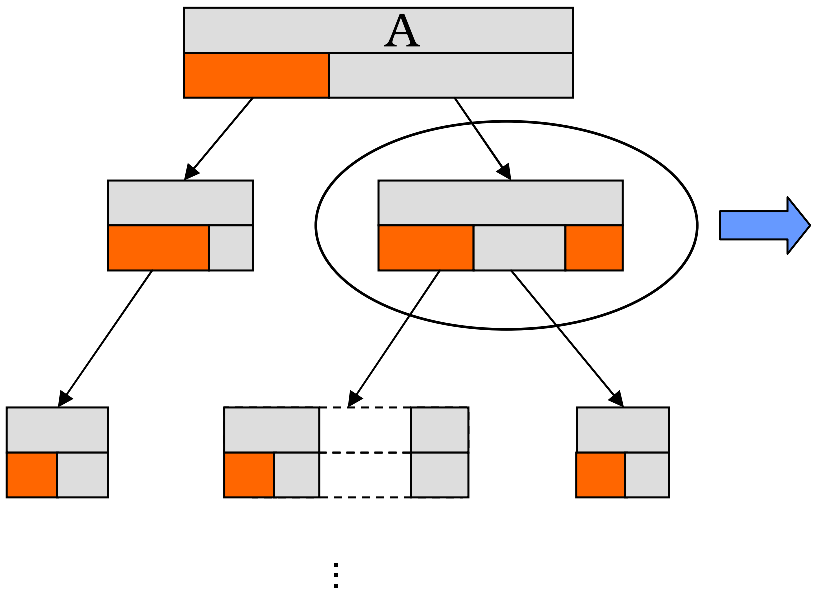
for all F_B (i.e. for all i,j):

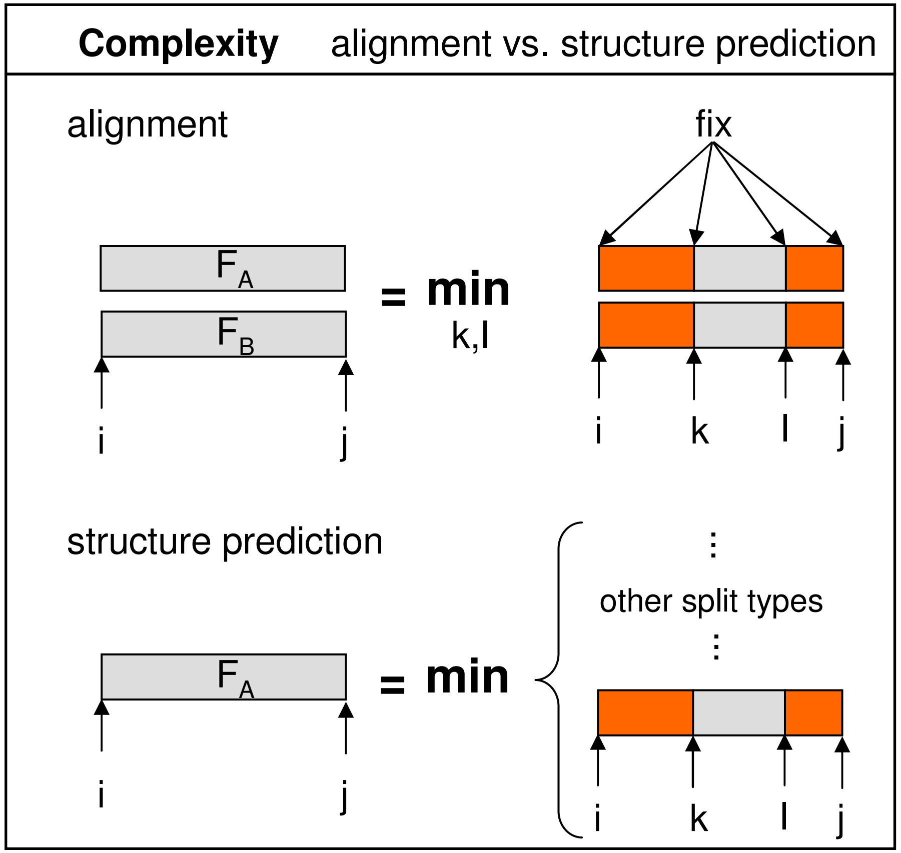
optimal alignment of

Alignment Algorithm Complexity

If RNA structure A has been predicted by some structure prediction algorithm X, there exists a decomposition of A that uses only the split types used in algorithm X

fixed decomposition of A:





Complexity alignment vs. structure prediction

alignment

F_A

F_B

$= \min_{k,l}$

fix

i j

i k l j

structure prediction

F_A

$= \min$

⋮

other split types

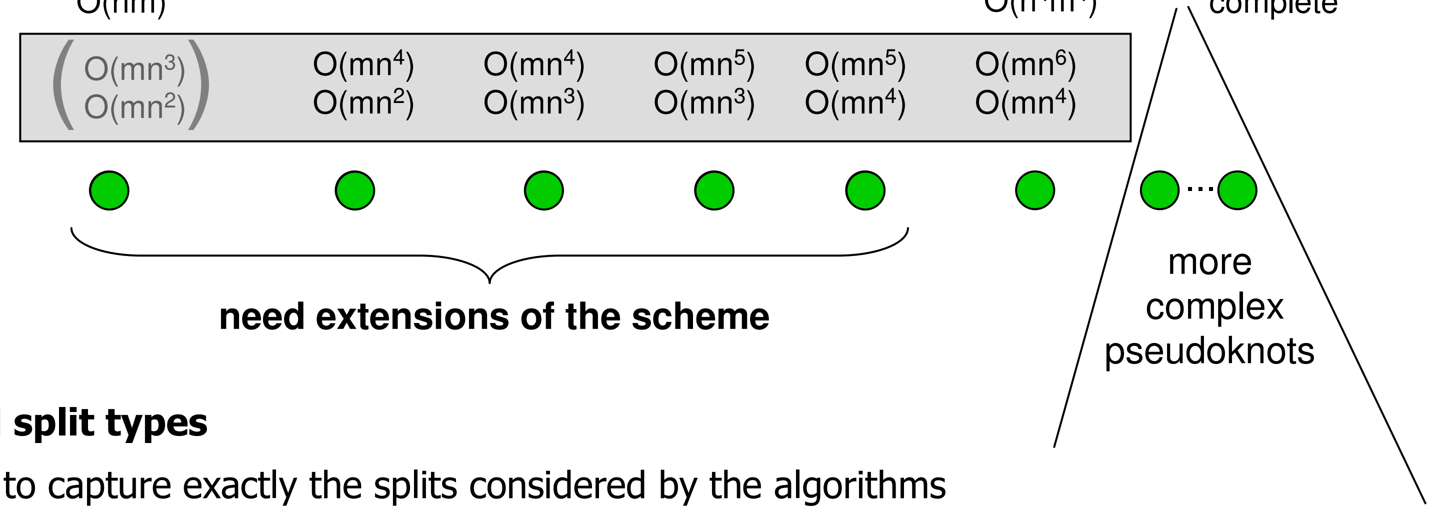
⋮

i j

i k l j

Extensions of the Scheme

class of structures		pseudoknot free	R&G	A&U	L&P	D&P	R&E	arbitrary pseudoknots
prediction	time space	$O(n^3)$ $O(n^2)$	$O(n^4)$ $O(n^2)$	$O(n^4)$ $O(n^3)$	$O(n^5)$ $O(n^3)$	$O(n^5)$ $O(n^4)$	$O(n^6)$ $O(n^4)$	NP-complete
alignment	time space	$O(n^2m^2)$ $O(nm)$					$O(n^5m^5)$ $O(n^4m^4)$	NP-complete
		$\left(\begin{matrix}O(mn^3)\\O(mn^2)\end{matrix}\right)$	$O(mn^4)$ $O(mn^2)$	$O(mn^4)$ $O(mn^3)$	$O(mn^5)$ $O(mn^3)$	$O(mn^5)$ $O(mn^4)$	$O(mn^6)$ $O(mn^4)$	



need extensions of the scheme

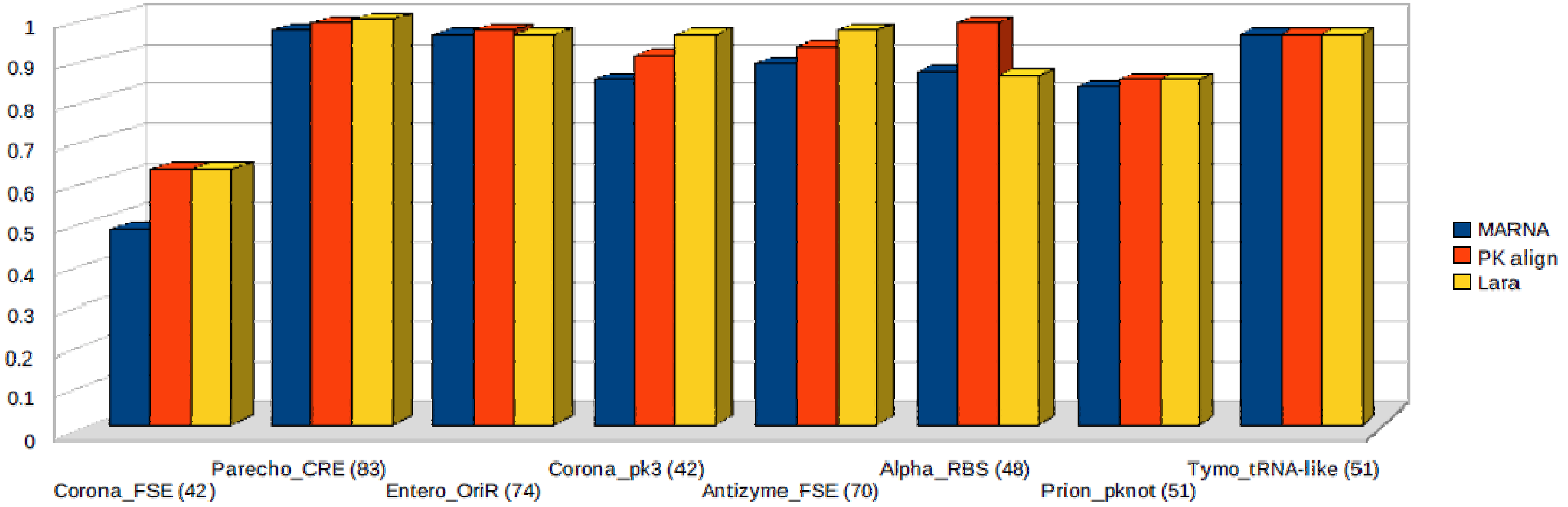
more complex pseudoknots

- **extended split types**
  - need to capture exactly the splits considered by the algorithms
  - modify alignment computation to maintain correctness for these types
- **optimize space consumption**

→ **all tricks of the prediction algorithms can be transferred to alignment**

Practical Evaluation: PKalign

COMPALIGN SCORE



- aligned pseudoknot structures of Rfam Database (contains hand cured reference alignments)
- sequence length up to 125

Conclusions

RNA structure alignment and prediction

- complexity depends on the structures

class of structures		pseudoknot free	R&G	A&U	L&P	D&P	R&E	arbitrary pseudoknots
prediction	time space	$O(n^3)$ $O(n^2)$	$O(n^4)$ $O(n^2)$	$O(n^4)$ $O(n^3)$	$O(n^5)$ $O(n^3)$	$O(n^5)$ $O(n^4)$	$O(n^6)$ $O(n^4)$	NP-complete
alignment	time space	$O(n^2m^2)$ $O(nm)$	$O(mn^4)$ $O(mn^2)$	$O(mn^4)$ $O(mn^3)$	$O(mn^5)$ $O(mn^3)$	$O(mn^5)$ $O(mn^4)$	$O(mn^6)$ $O(mn^4)$	NP-complete

new contribution

- implementation: PKalign
- accepted for RECOMB 09

Literature (1/2)

[Akutsu2000] Akutsu, T. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics,* **2000***, 104*, 45-62

[Bauer2007] Bauer, M.; Klau, G. W. & Reinert, K. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization *BMC Bioinformatics,* 2007*, 8*, 271

[Dirks2003] Dirks, R. M. & Pierce, N. A. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J Comput Chem,* 2003*, 24*, 1664-77

[Evans1999] Evans, P. A. Finding Common Subsequences with Arcs and Pseudoknots. *CPM '99: Proceedings of the 10th Annual Symposium on Combinatorial Pattern Matching, Springer-Verlag,* 1999, 270-280

[Evans2006] Evans, P. A. Finding Common RNA Pseudoknot Structures in Polynomial Time. *Combinatorial Pattern Matching (CPM 2006), Springer Berlin / Heidelberg,* 2006*, 4009/2006*, 223-232

[Jiang2002] Tao Jiang, Guohui Lin, Bin Ma, and Kaizhong Zhang. A general edit distance between RNA structures. J. Comput. Biol., 9(2):371–88, 2002.

[Lyngso2000] Lyngso, R. B. & Pedersen, C. N. S. Pseudoknots in RNA Secondary Structures. *RECOMB 00, ACM Press,* 2000

[Lenhof1998] Lenhof, H. P.; Reinert, K. & Vingron, M. A Polyhedral Approach to RNA Sequence Structure Alignment 1998

[Missal2005] Missal, K.; Rose, D. & Stadler, P. F. Non-coding RNAs in Ciona intestinalis. *Bioinformatics,* 2005*, 21 Suppl 2*, ii77-ii78

Literature (2/2)

[Möhl2008] Möhl, M.; Will, S. & Backofen, R. Fixed Parameter Tractable Alignment of RNA Structures. *Proceedings of the 19th Annual Symposium on Combinatorial Pattern Matching (CPM 2008),* 2008

[Needleman1970] Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol,* 1970*, 48*, 443-453

[Notredame2000] Notredame, C.; Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol, 2000, 302, 205-17

[Reeder2004] Reeder, J. & Giegerich, R. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics,* 2004*, 5*, 104

[Rivas1999] Rivas, E. & Eddy, S. R. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol,* 1999*, 285*, 2053-2068

[Sankoff1985] Sankoff, D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.,* 1985*, 45*, 810-825

[Siebert2007] Siebert, S. & Backofen, R. Methods for multiple alignment and consensus structure prediction of RNAs implemented in MARNA *Methods Mol Biol,* **2007***, 395*, 489-502

[Uemura1999] Uemura, Y.; Hasegawa, A.; Kobayashi, S. & Yokomori, T. Tree adjoining grammars for RNA structure prediction. T*heoretical Computer Science,* 1999*, 210*, 277 - 303

[Zuker 1981] Zuker, M. & Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res,* 1981*, 9*, 133-148