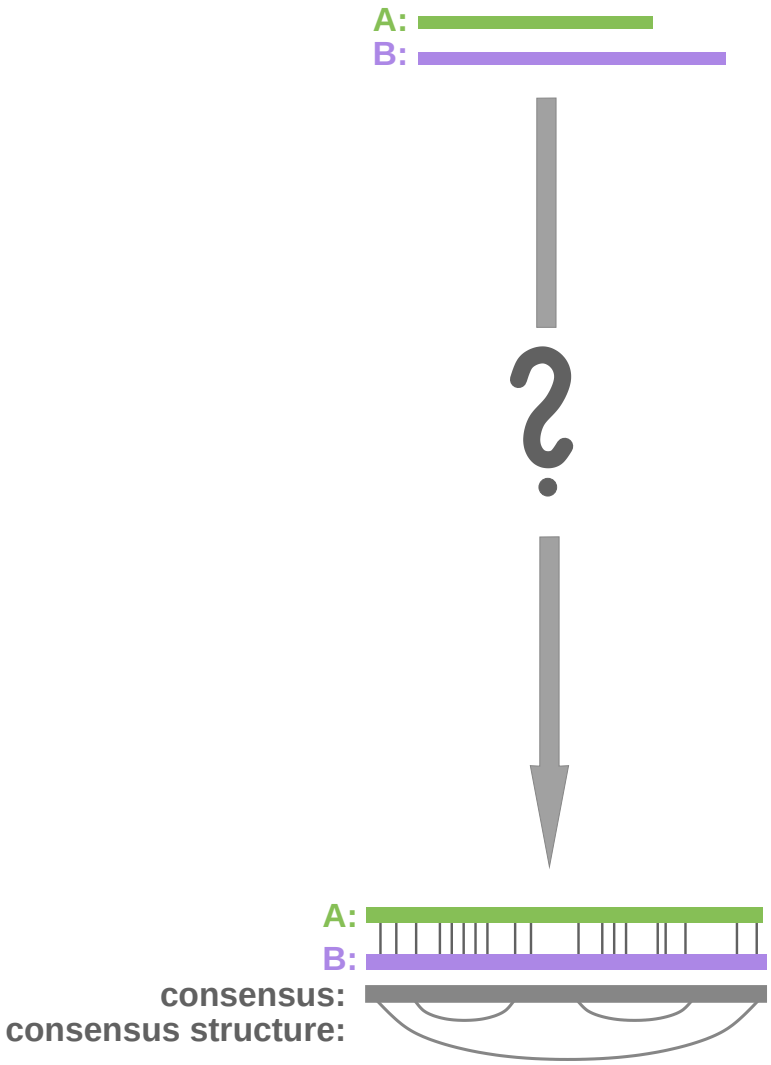
Match Probabilities from Sankoff-style Alignment in LocARNA

Sebastian Will
Bioinformatics, Uni Freiburg

Benasque, 29.07.09

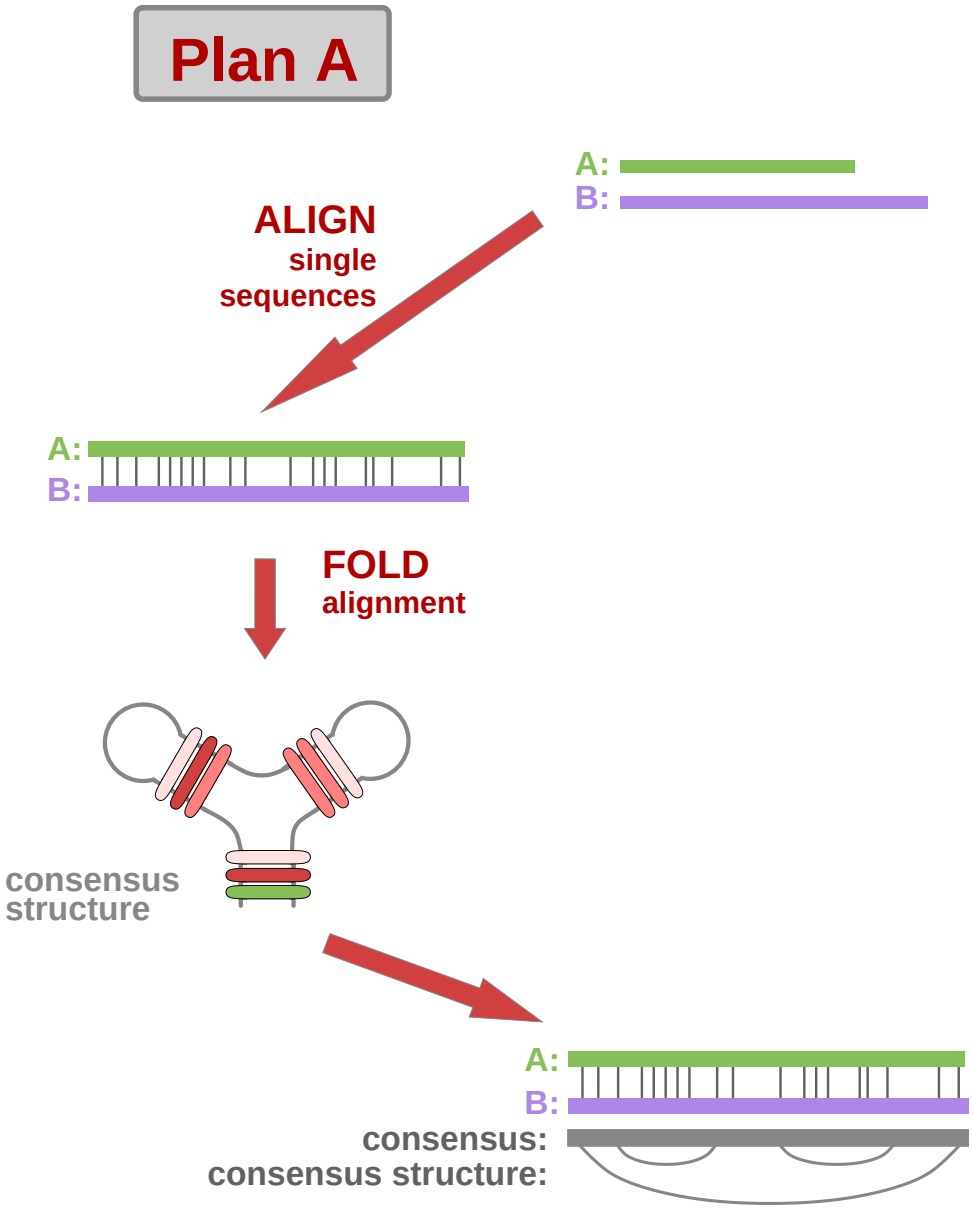
What?

Comparative RNA Analysis

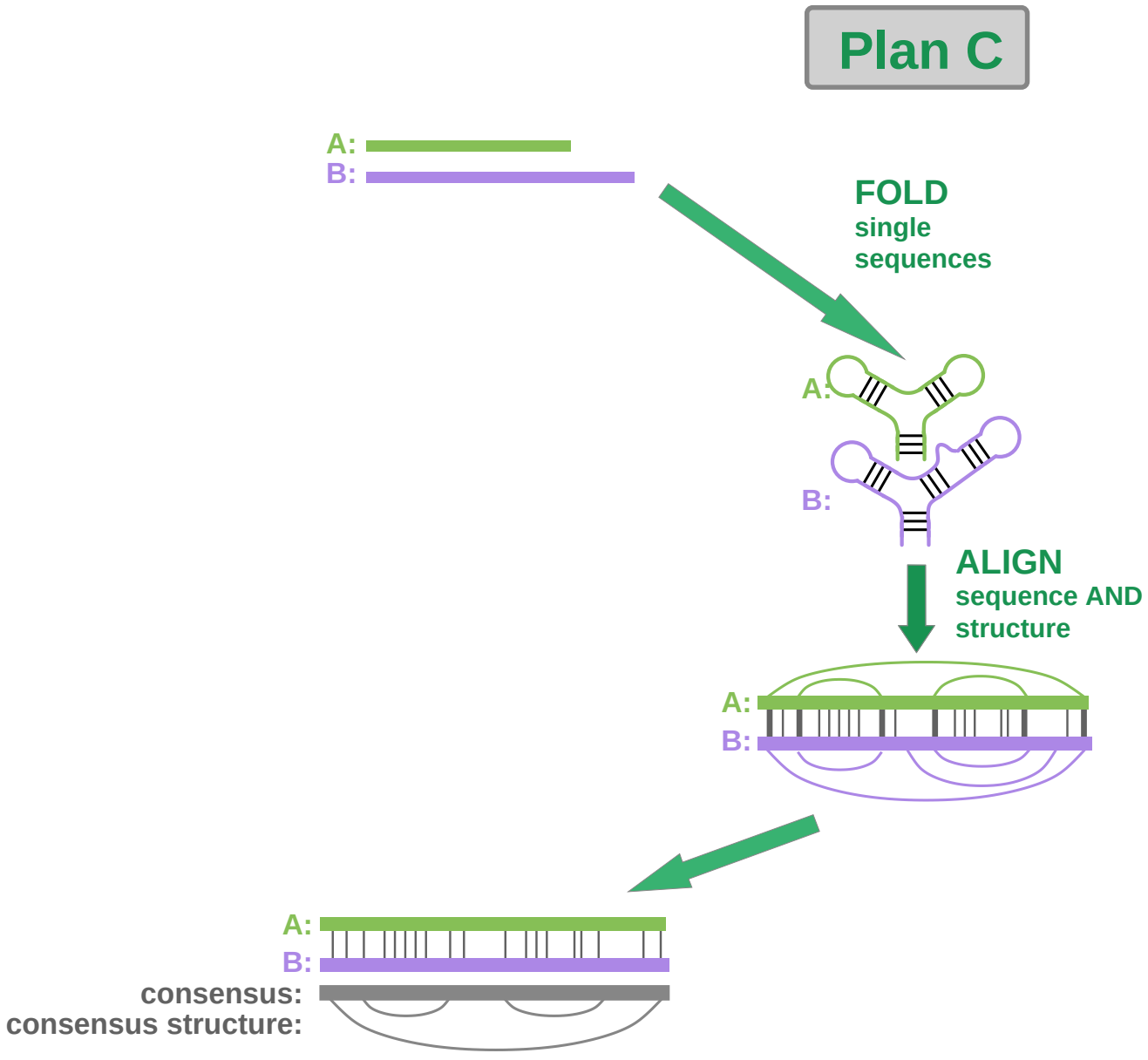


adopted from:
[Gardener & Giiegerich BMC 2004]

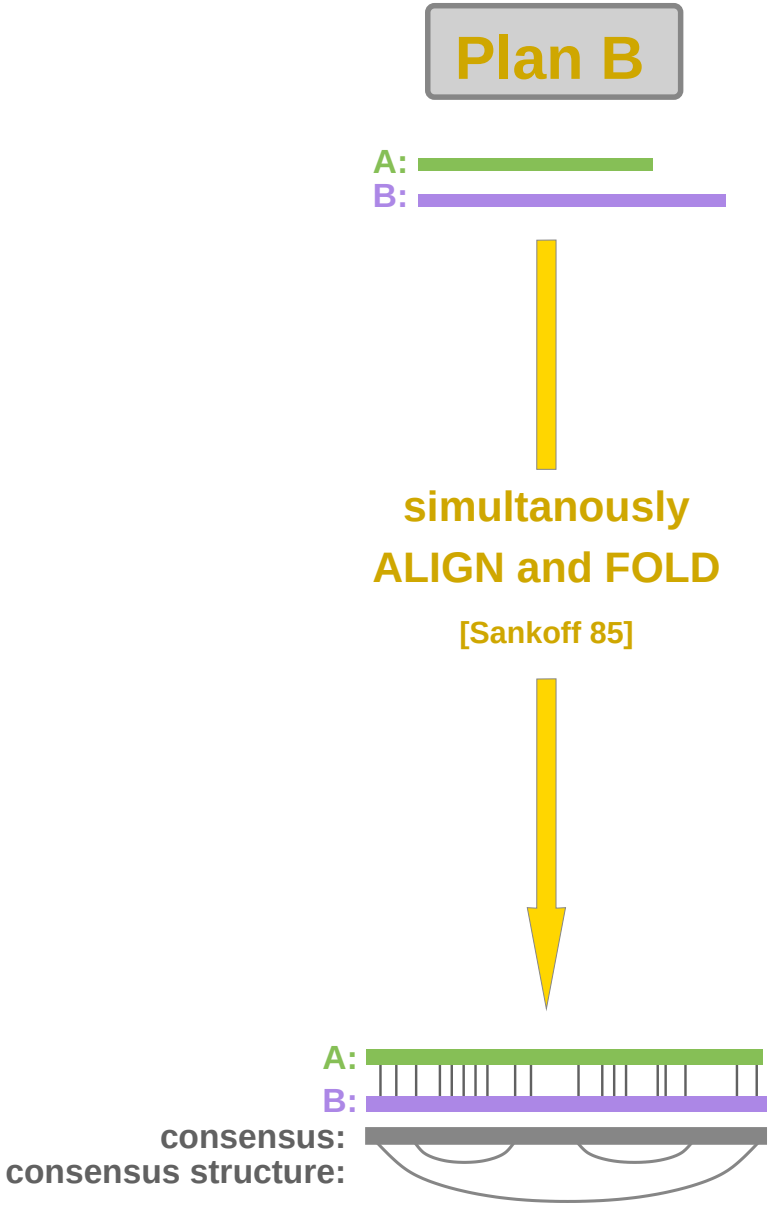
Comparative RNA Analysis



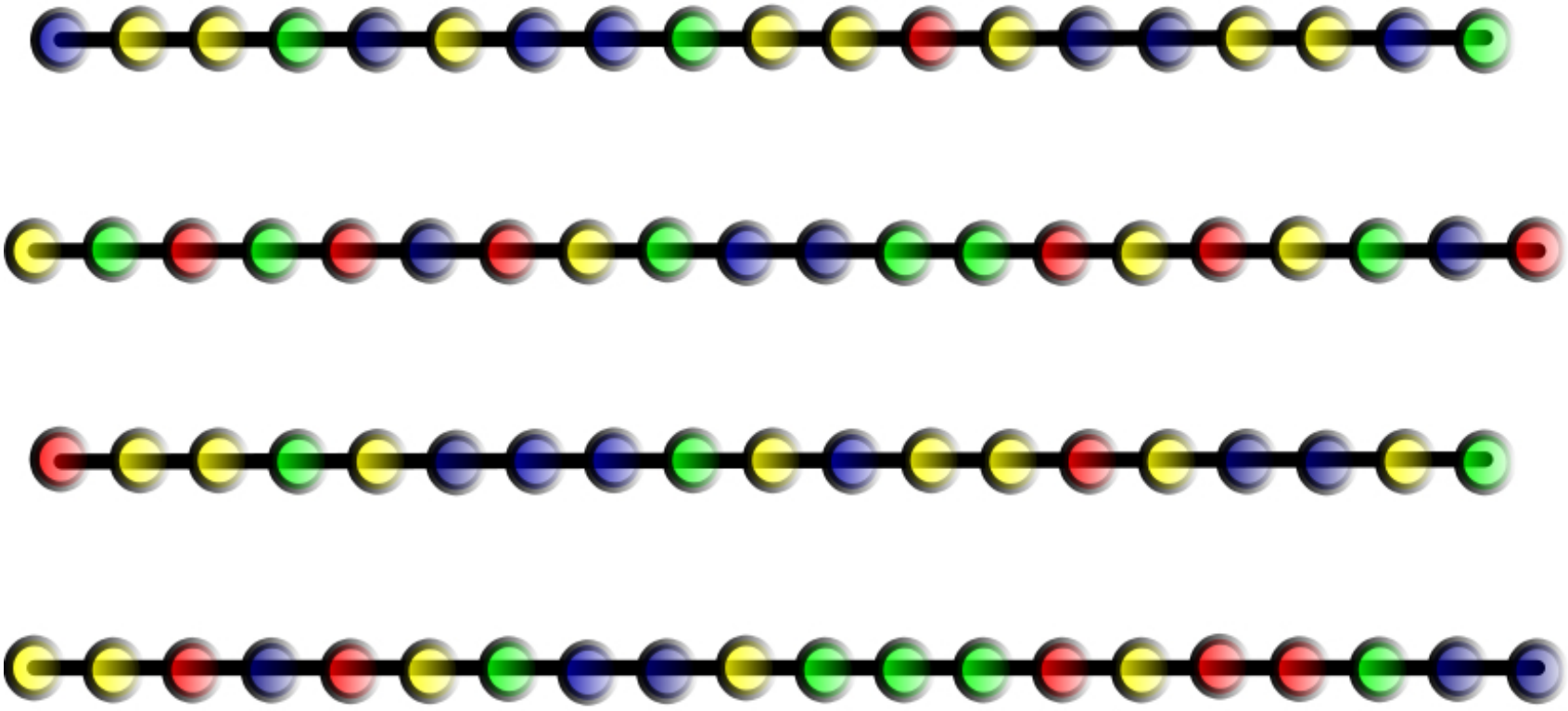
Comparative RNA Analysis



Comparative RNA Analysis

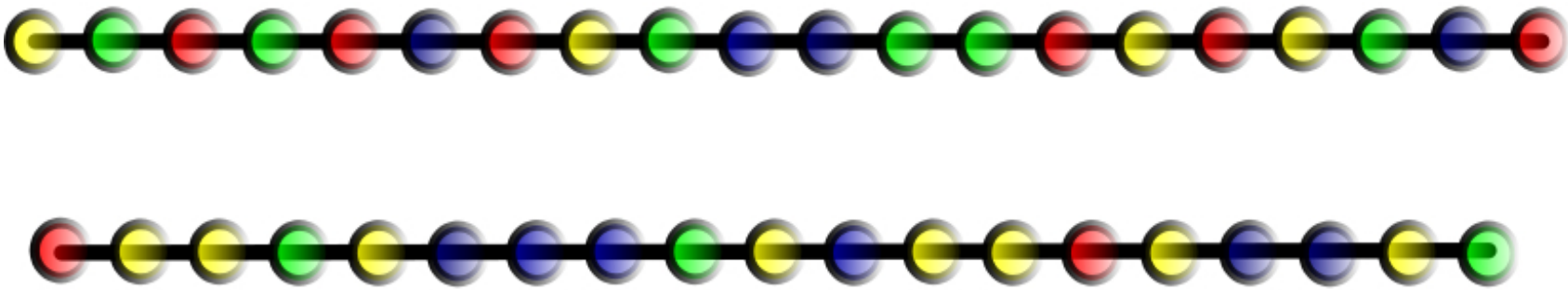


Alignment and Folding



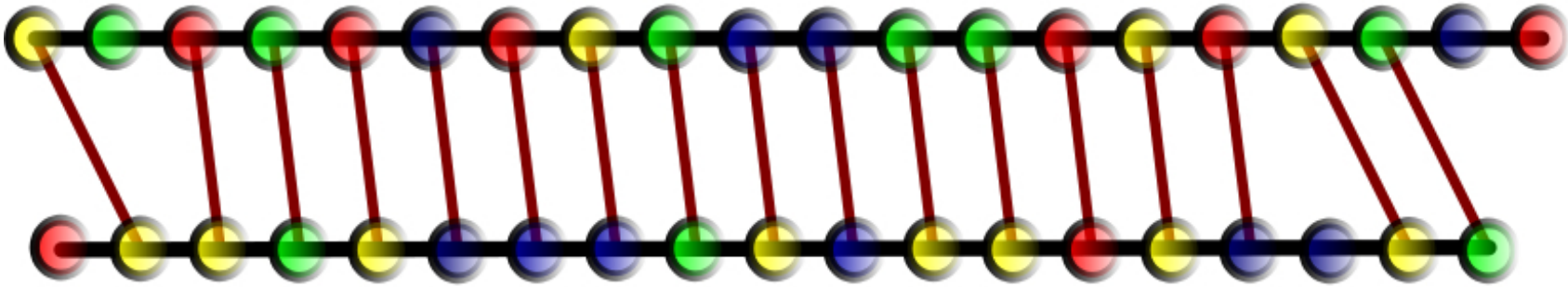
Set of unaligned RNAs

Alignment and Folding



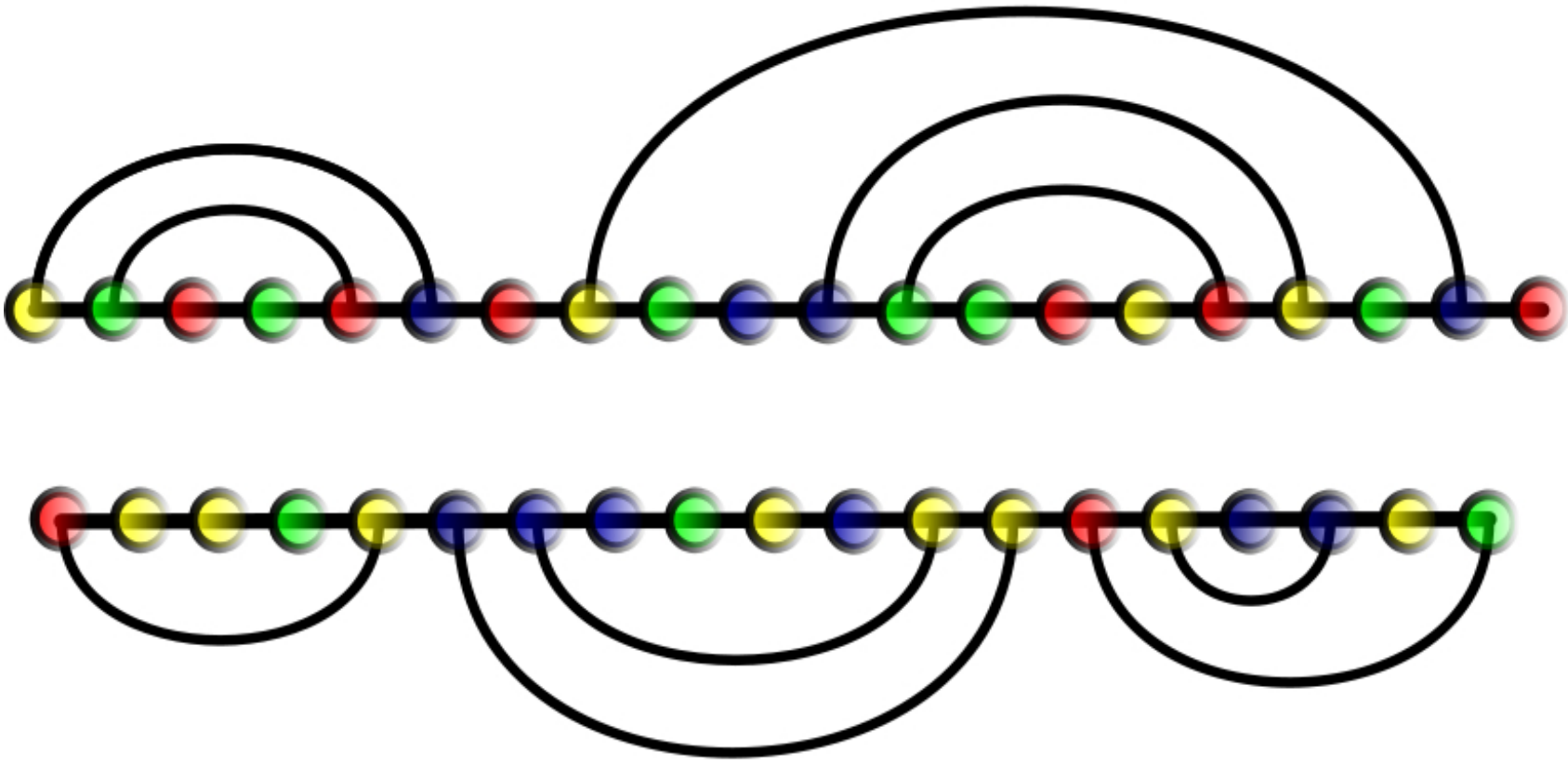
Focus on pairwise alignment
construct multiple from pairwise

Alignment and Folding



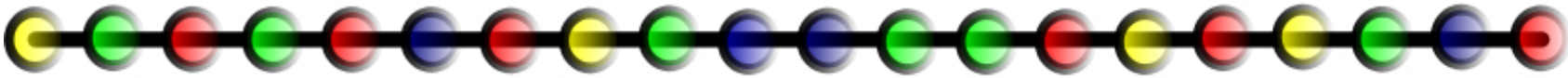
PLAN A: First align, then fold?
Sequence alignment wrong!

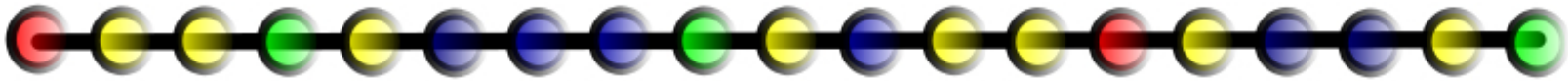
Alignment and Folding



PLAN C: First fold, then align?
Structures don't match!

Alignment and Folding





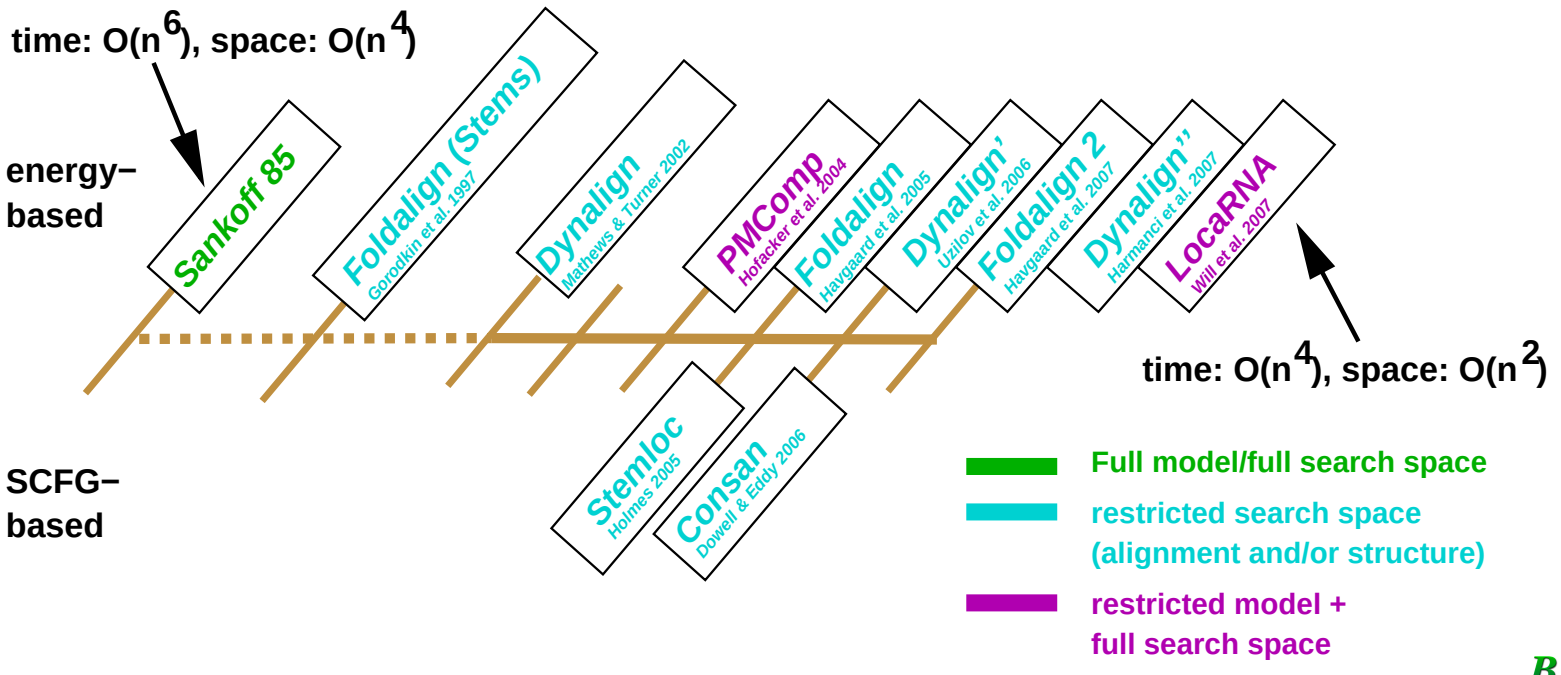
Plans A and C do not work:
RESTART!: PLAN B

Sankoff-like approaches

- Sankoff is the gold standard *BUT requires extreme amount of space and time* *[Gardener & Giegerich 2004]*

  (time: $O(n^6)$, space $O(n^4)$)

- hence: Sankoff-like approaches are restricted versions



PMcomp Variant of Sankoff

"Sankoff = Zuker × Alignment"

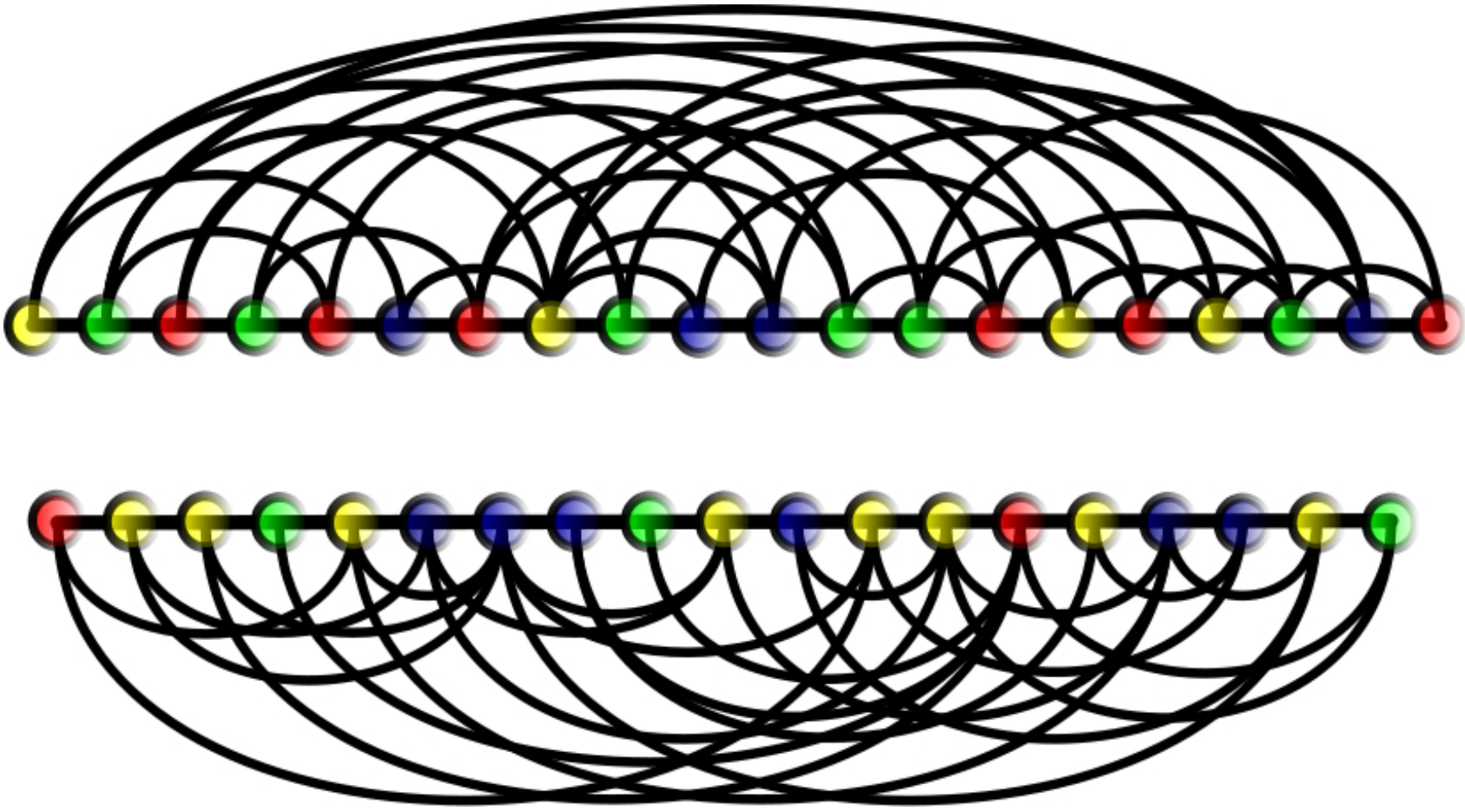
"Hofacker = Nussinov × Alignment + McCaskill"

Sankoff: score matched loops

Hofacker: score matched base pairs
due to base pair probabilities

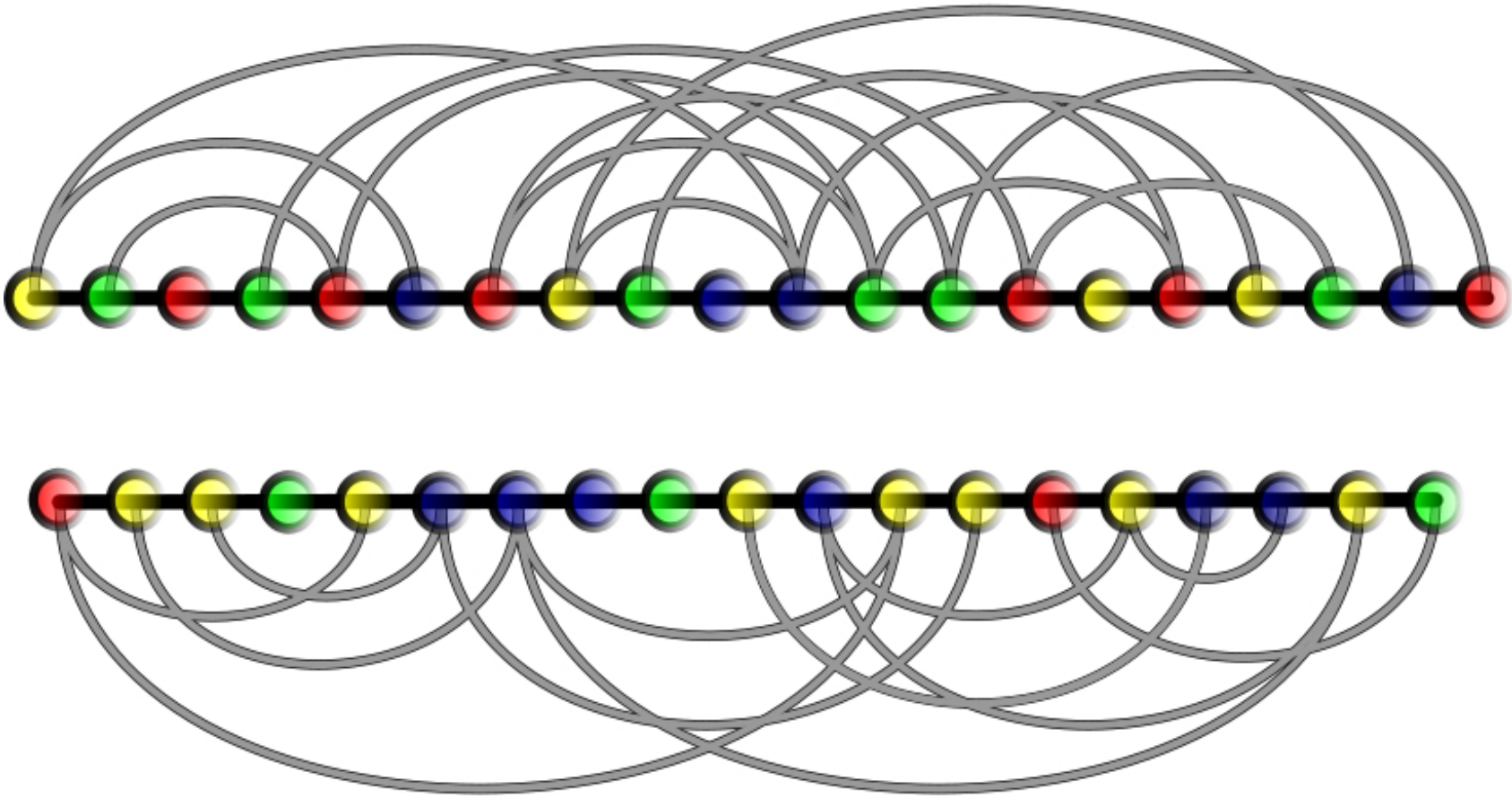
Hofacker et al., PMcomp, *Bioinformatics*, 2004.

Simultaneous Alignment and Folding



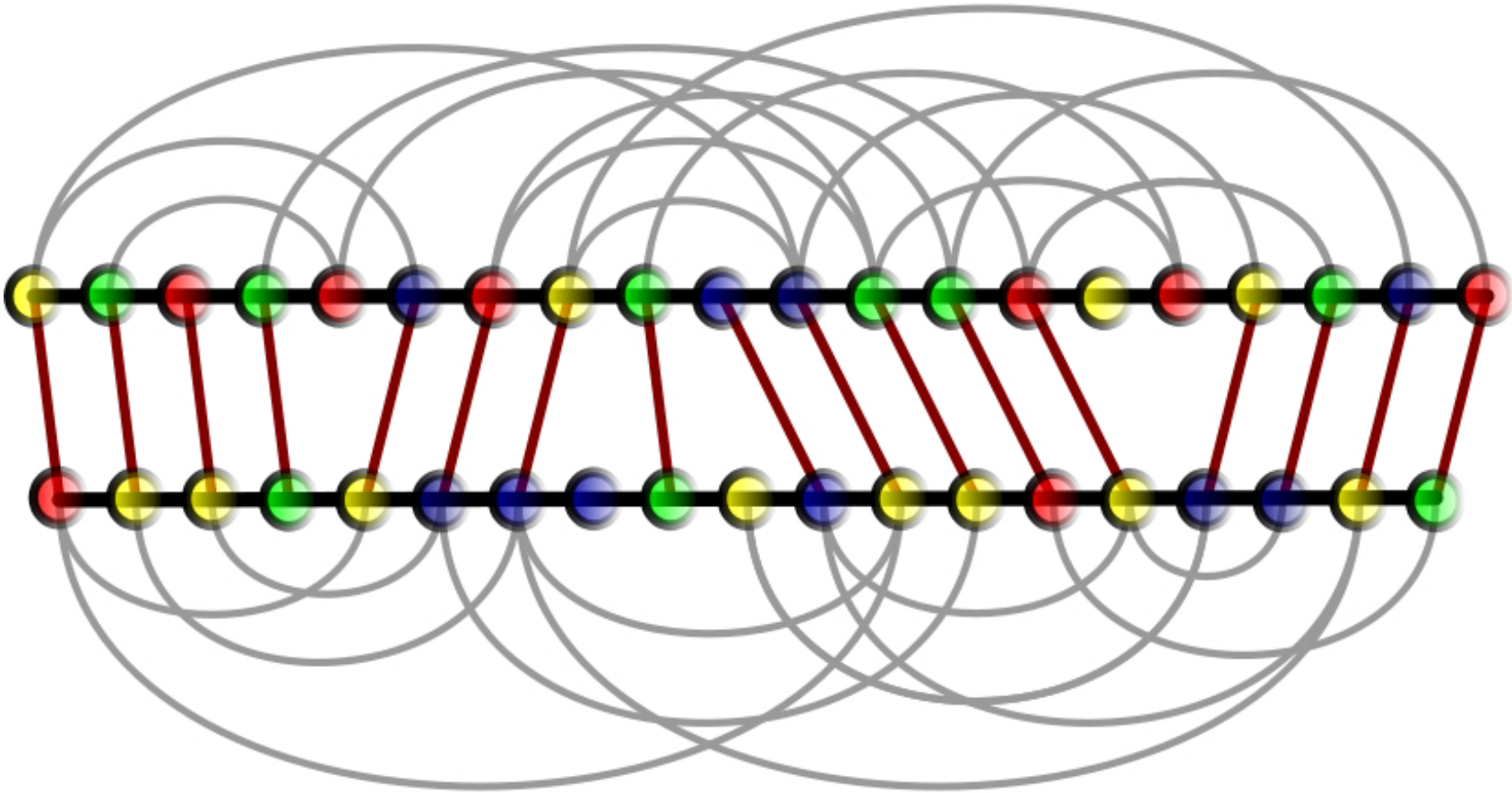
Consider all possible structures
too many base pairs!

Simultaneous Alignment and Folding



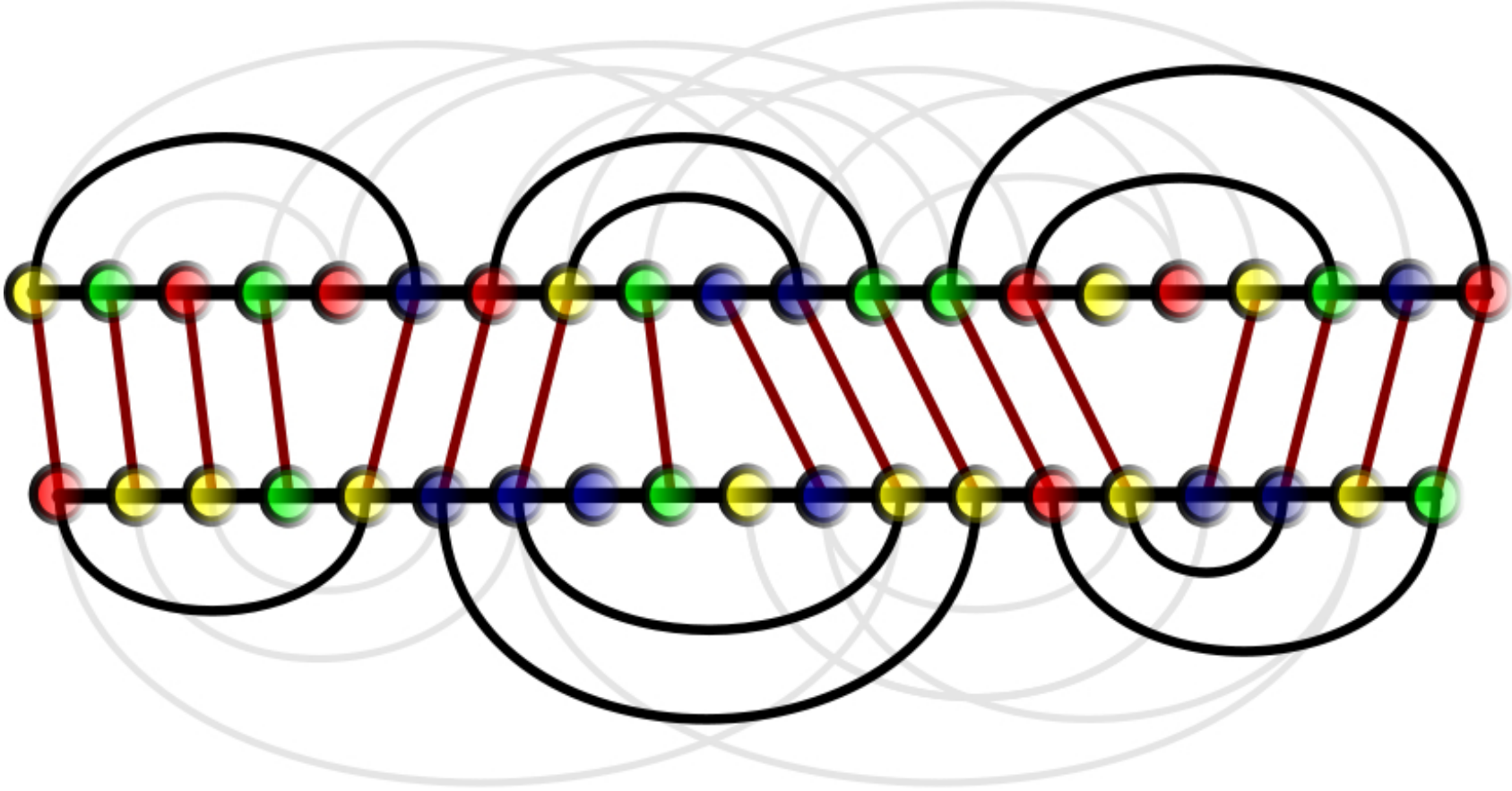
Use sparsity!
select only probable base pairs

Simultaneous Alignment and Folding



Best alignment and structure in $O(n^4)/O(n^2)$

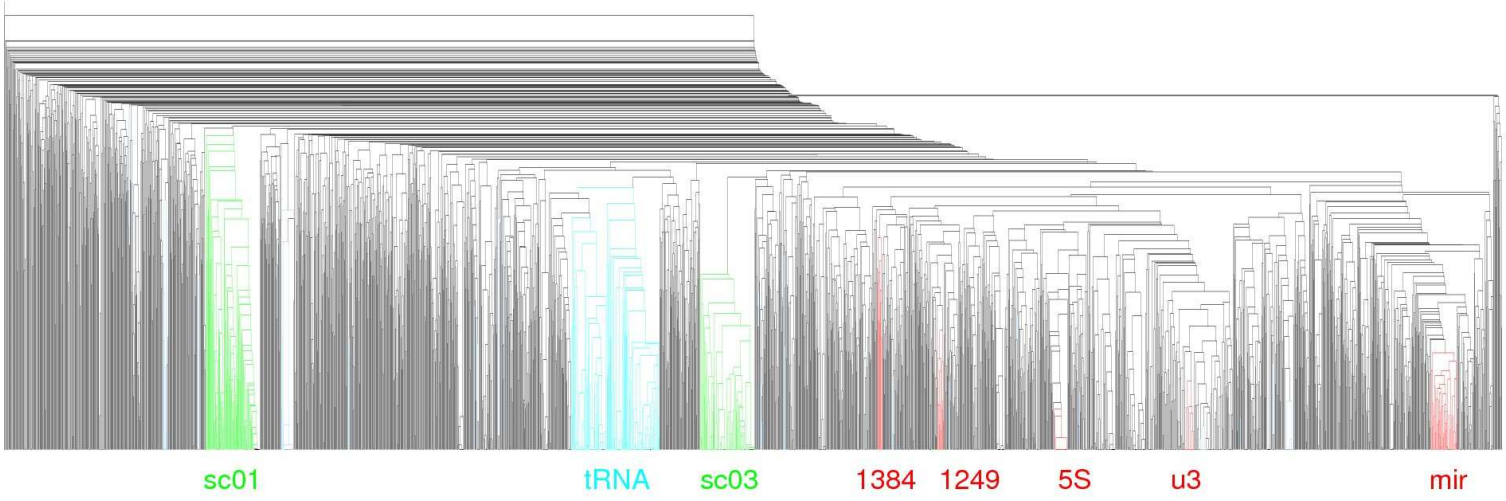
Simultaneous Alignment and Folding



Simultaneously select consensus structure
sequence alignment edges — structural alignment edges

LocARNA: Clustering of RNAz ncRNA Predictions

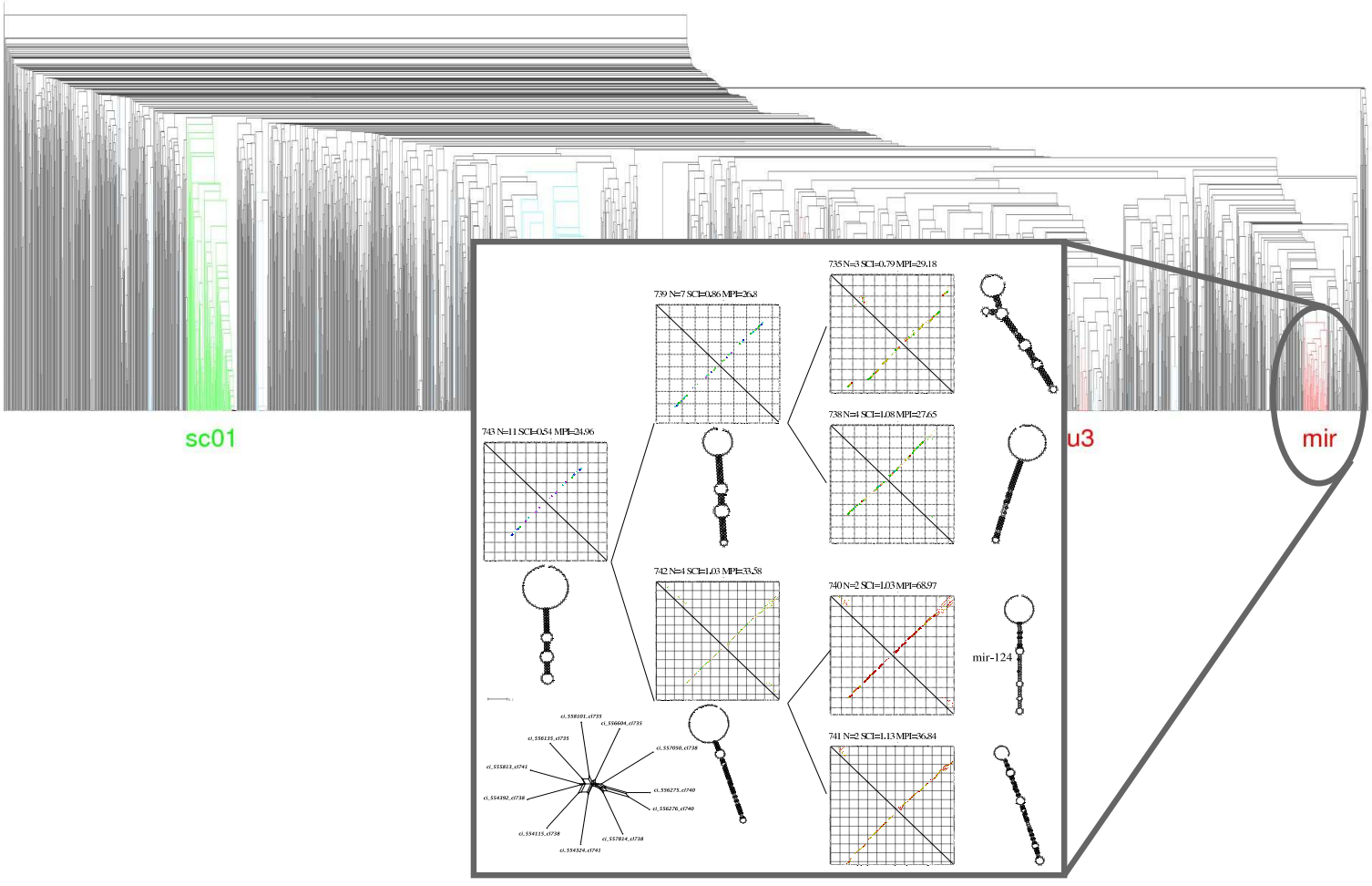
- Clustering of 3332 putative ncRNAs in *Ciona intestinalis*



Will, Reiche et al., PLOS Comp Biol, 2007.

LocARNA: Clustering of RNAz ncRNA Predictions

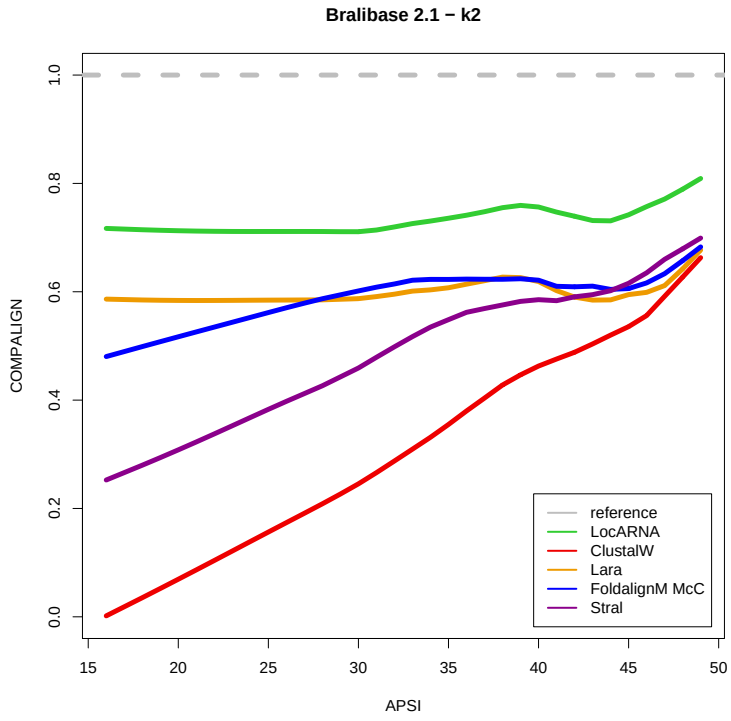
- Clustering of 3332 putative ncRNAs in *Ciona intestinalis*

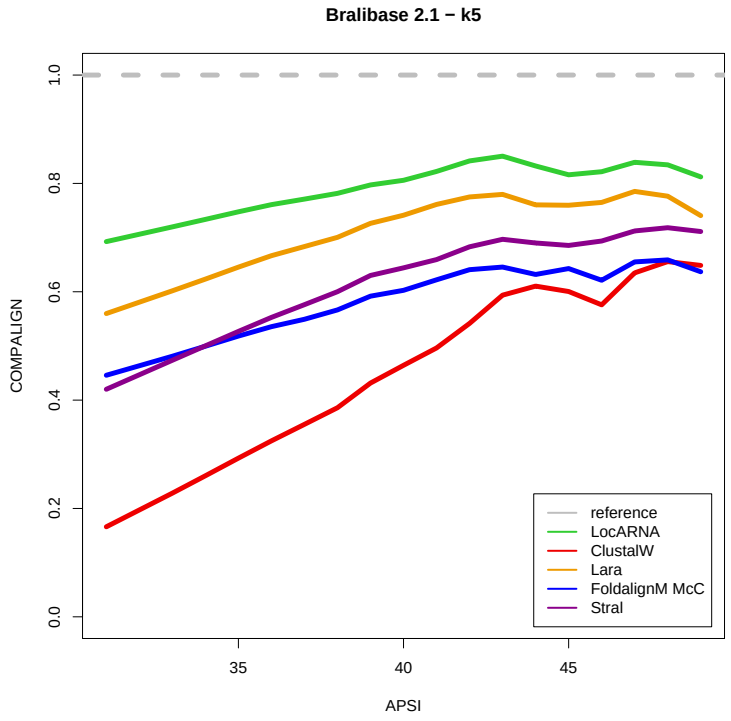


Will, Reiche et al., PLOS Comp Biol, 2007.

Multiple Alignment: Bralibase Benchmark

- Bralibase: Example k2 and k5
- low sequence identity region





Multiple Alignment: Run-time

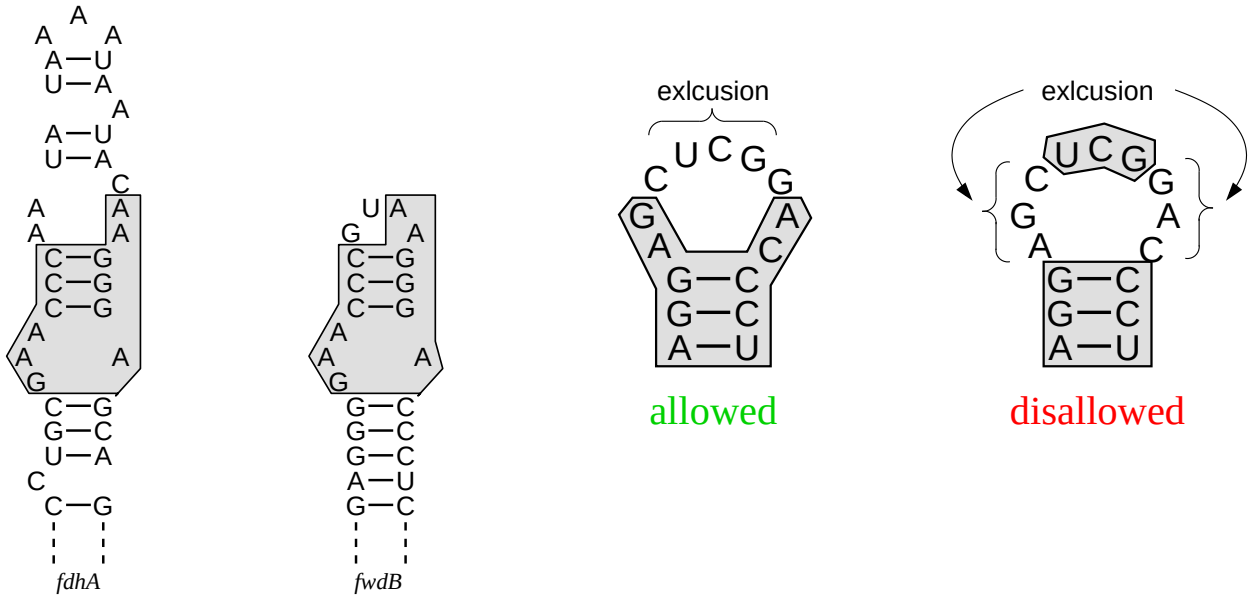
# seqs	RNA	Program	Time
10	tRNA (apsi45,sci111)	LocARNA	4s
		LARA	90s
		FoldalignM - McC	21s
15	tRNA (apsi45,sci110)	LocARNA	7s
		LARA	210s
		FoldalignM - McC	45s
15	5S rRNA (apsi60,sci72)	LocARNA	31s
		LARA	348s
		FoldalignM - McC	43s

LARA: Bauer, Klau, Reinert. BMC Bioinformatics, 2007.

FoldAlignM: Torarinsson, Havgaard, Gorodkin. Bioinformatics, 2007.

Structure Local Alignment

What is structure local?

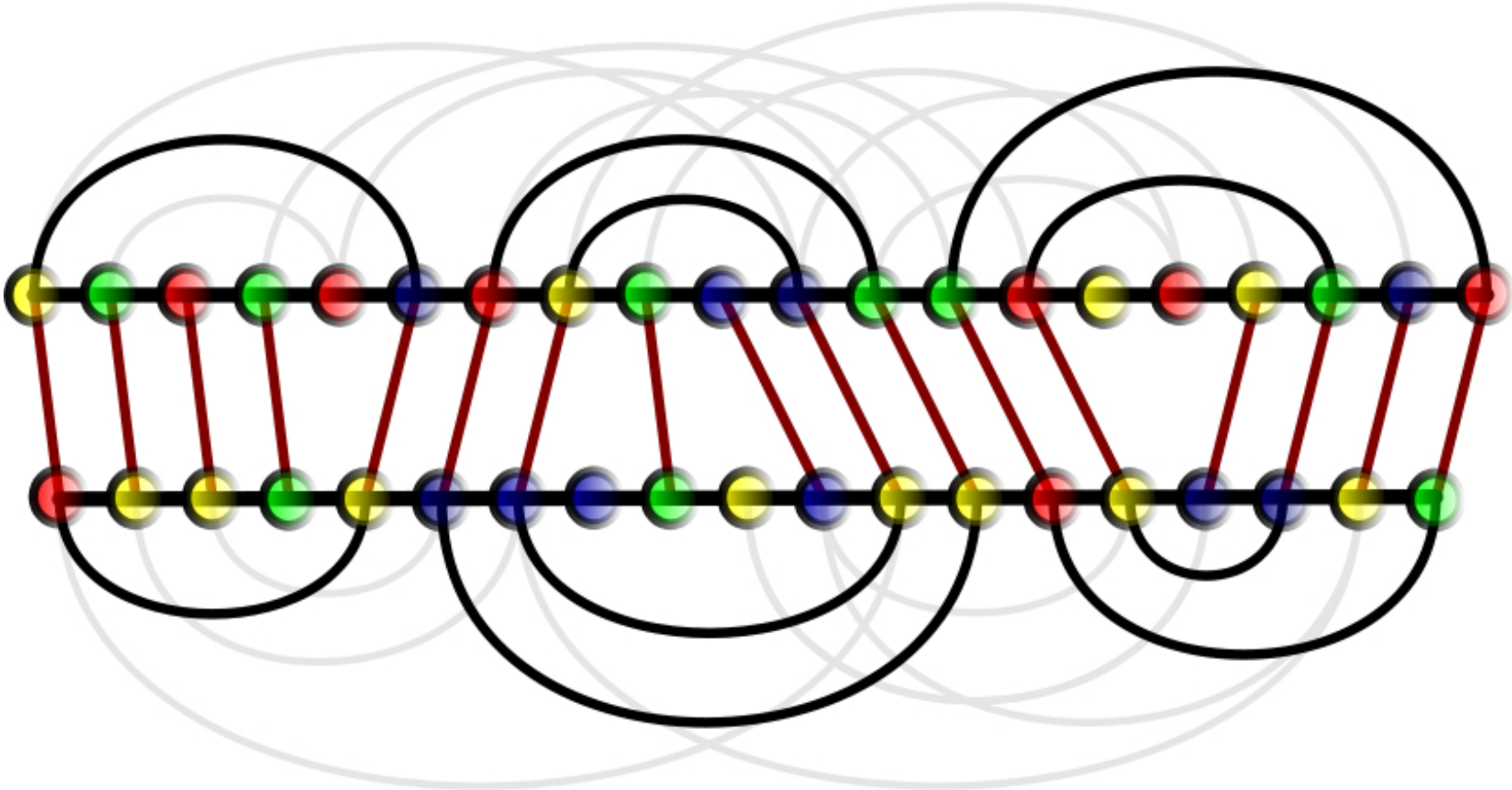


Find best alignment of “connected” sub-structures.

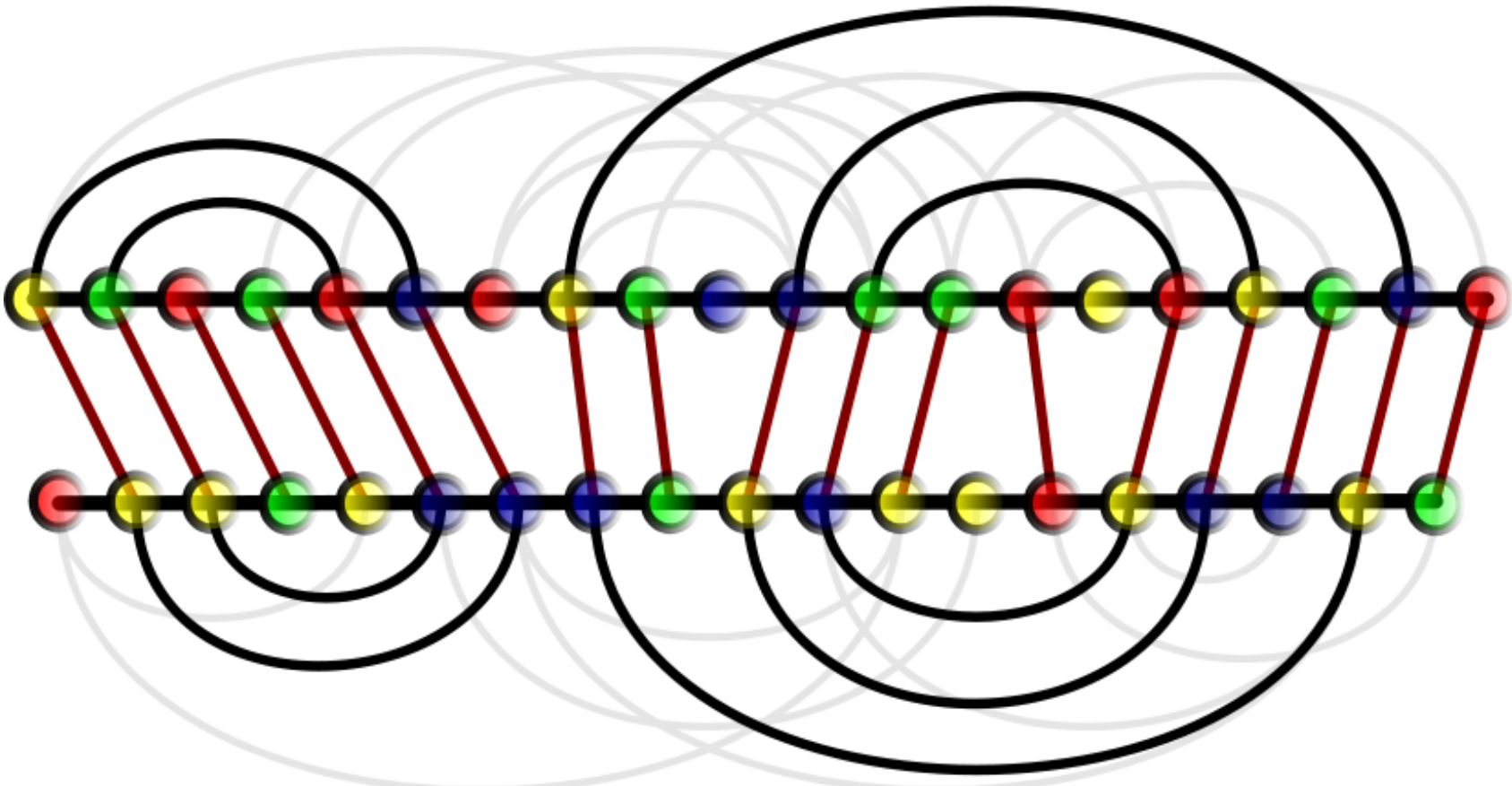
No increase of complexity!

Otto, Will et al., GCB, 2008.

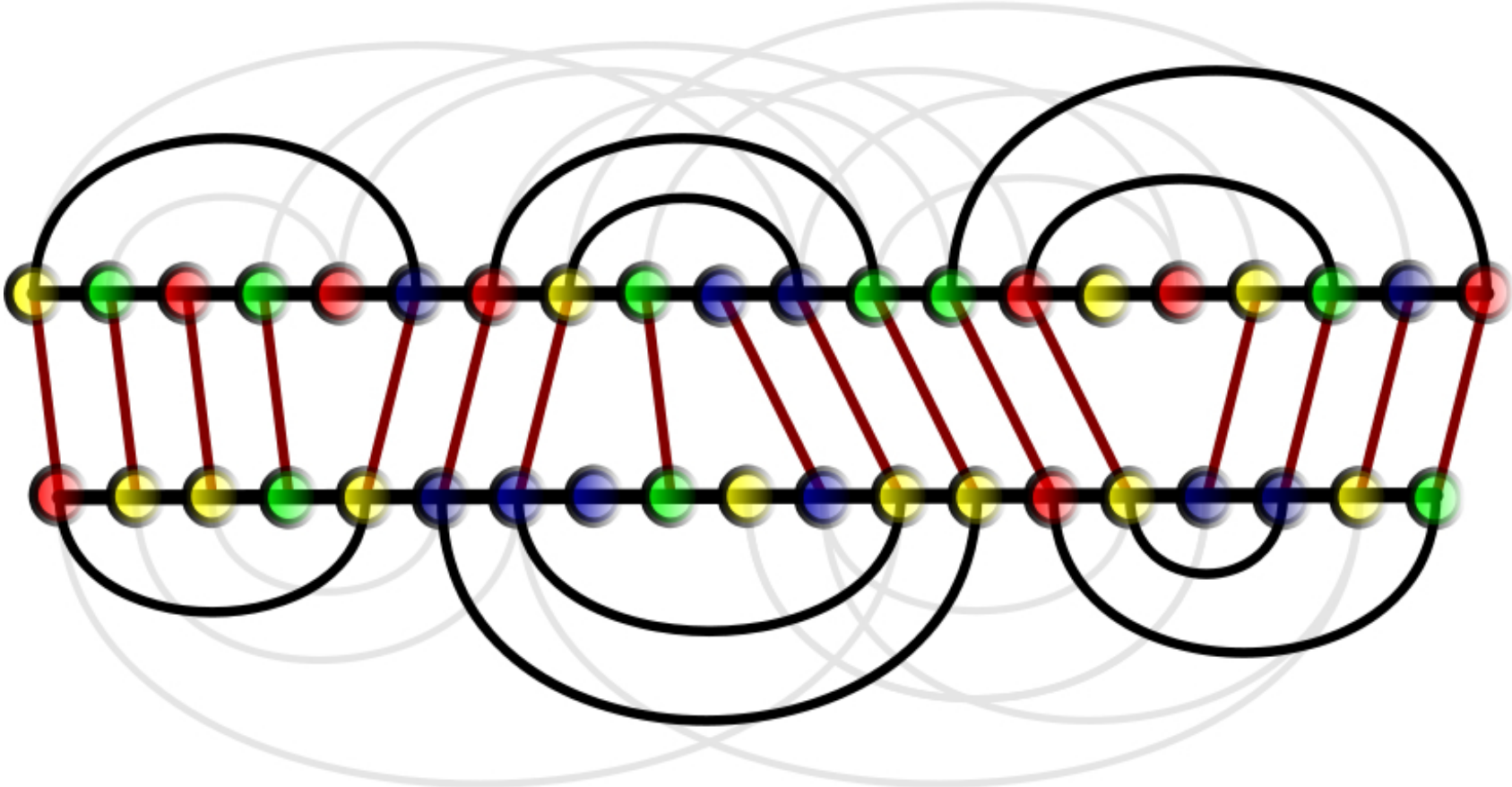
Structural Alignment



Many Good Structural Alignments

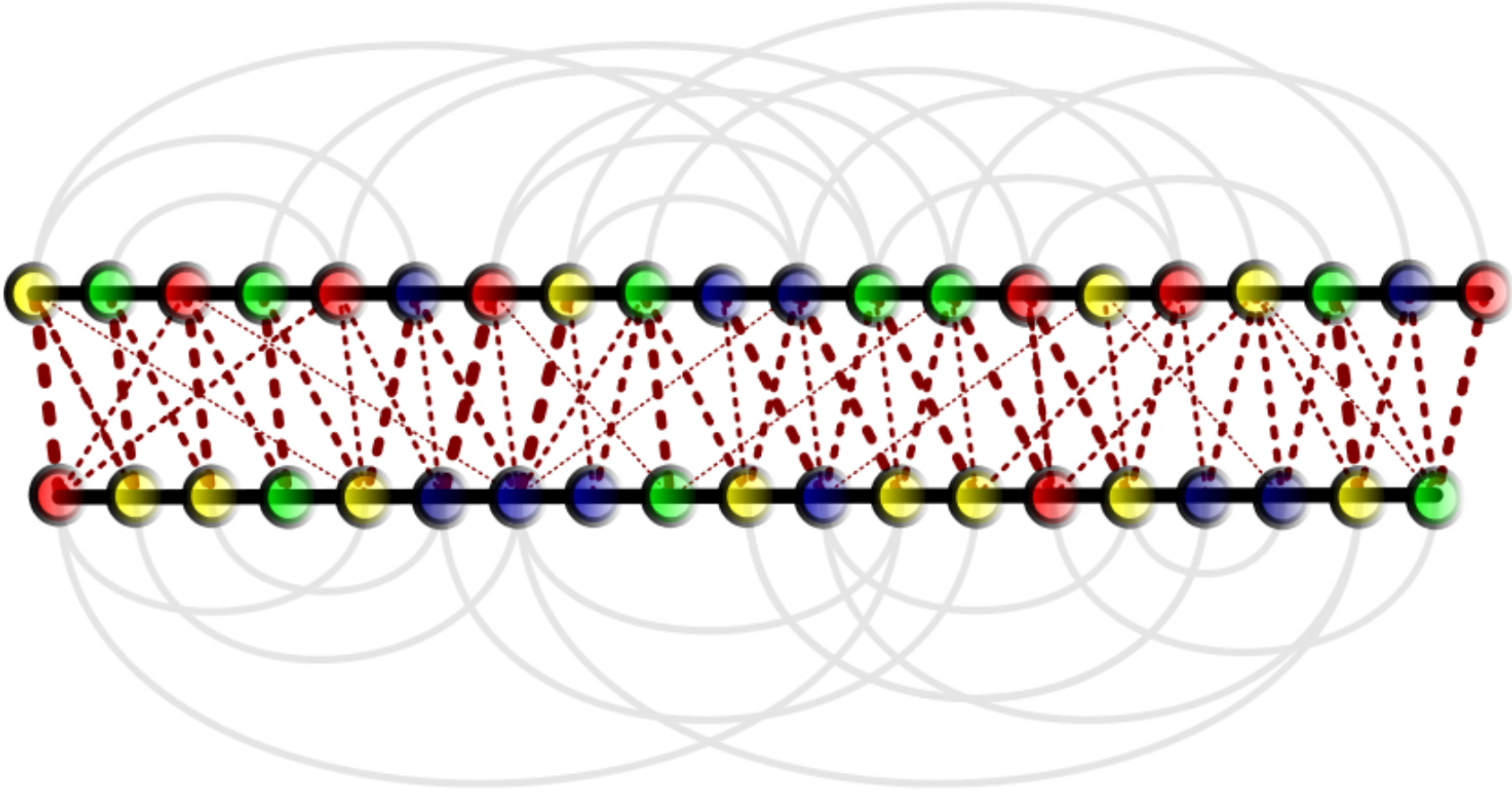


Many Good Structural Alignments



How to describe the good alignments?

Many Good Structural Alignments



How to describe the good alignments? Probabilities!

Multiple Alignment: Probabilistic Consistency Transformation

Goal Avoiding errors in progressive multiple alignment

Method Re-estimate edge probabilities, using transitive edges in sequence triplets

Related to
- T-Coffee
- Probcons, Probalign

Advantages
- Probabilistic Transformation
- Probabilities from SA&F
- Sequence + Structural Edges

Local Alignment Quality: Reliability Profiles

Goal measure reliability of alignment columns

Method sum pairwise edge probabilities

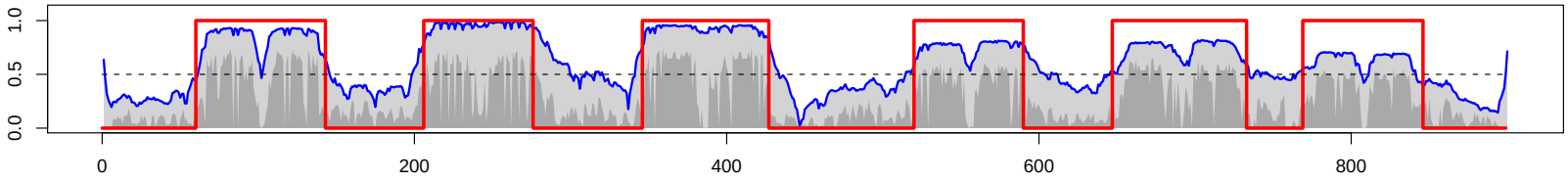
Result Reliability profile (structure #+sequence *)

```
fdhA       CGC-CACCCUGCGAACCCAAUAUAAAAUAAUACAAGGGAGCAG-GUG-GCG
fwdB       AUG-UUGGAGGGGAACCCGU------------AAGGGACCCUCCAA-GAU
hdrA       GG--CACCACUCGAAGGC--------------UAAGCCAAAGUGGUG--CU
selD       UUACGAUGUGCCGAACCCUU------------UAAGGGAGGCACAUCGAAA
vhuD       GU--UCUCUCGGGAACCCGU------------CAAGGGACCGAGAGA--AC
vhuU       AGC-UCACAACCGAACCCAU------------UUGGGAGGUUGUGA-GCU
fruA       CC--UC-GAGGGGAACCCGA------------AAGGGACCCG-AGA--GG
alifold    ((..(((((((((...(((.................))).)))))))))..))
- 10%      ### ################                ***************** ###
- 20%      ##* ################                ***************** *##
- 30%      ##* ################                 **************** *##
- 40%      ##  ################                 ****************  ##
- 50%      ##  #################                ****************  ##
- 60%      ##  ## ***********                   ********** ***   ##
- 70%      ##      **********                   **********       ##
- 80%      #*      **********                    *******         *#
- 90%      #        *********                     ******         *#
-100%      *           *****                       **            *
```

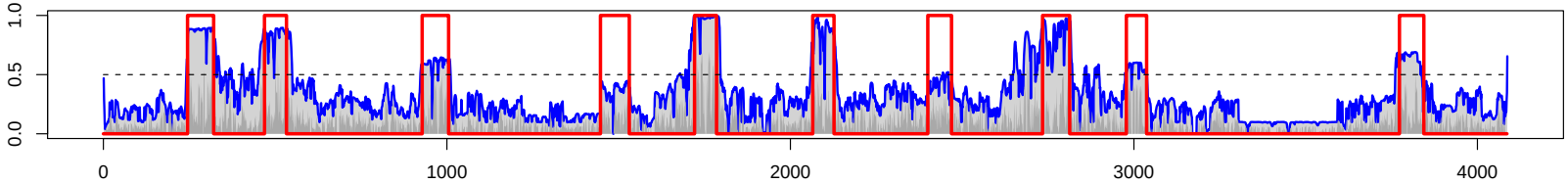
Reliability Profiles

- genomic cluster with known ncRNAs
- align corresponding regions in 10/5 vertebrates
- show reliability profile for human DNA

cluster of 6 micro RNAs, length ≈900



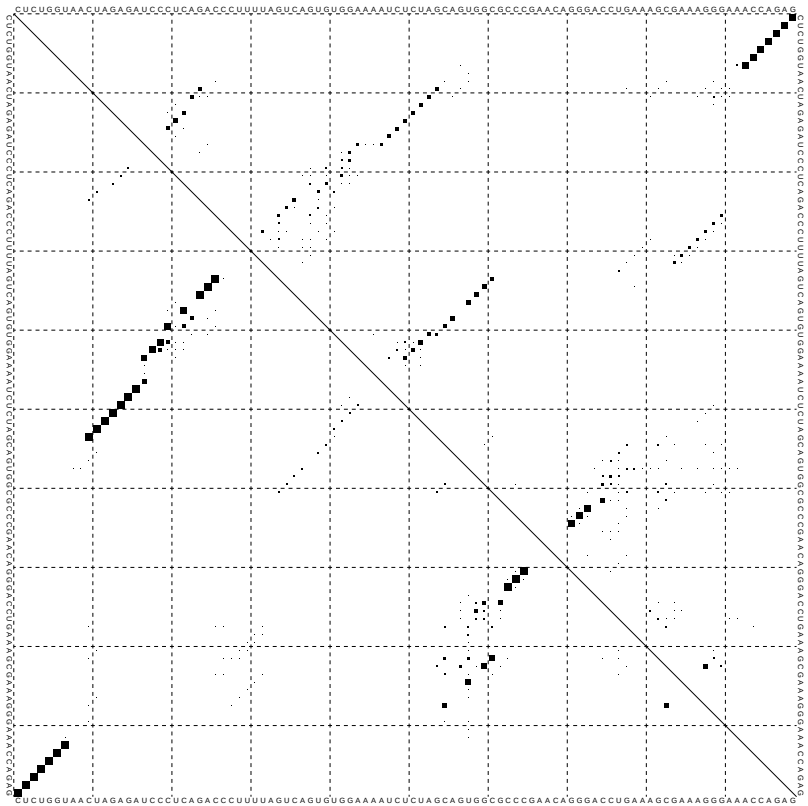
cluster of 10 CD-Box snoRNAs 'GAS5', length ≈4000



Base Pair Reliabilities

- reliability for structurally aligning a column pair
- reliability dot plot (comparable to RNAalifold dot plot)

LocARNA reliabilities: all possible alignments



RNAalifold probabilities: fix alignment

extreme Example: 10 seqs from HIV_PBS

indeed: different mfe structures for single RNAs

How?

Summary

- fast&accurate Sankoff-style
  reduce time **and** space by $O(n^2)$
- sequence and structure match probabilities
  equally efficient
- probabilities enable
  - better multiple alignment
  - assessing local alignment quality (reliability profiles)
  - reliability dot plots
  - . . . (work in progress)

WEB-Server:

`http://rna.tbi.univie.ac.at/cgi-bin/LocARNA.cgi`

More Info+Download:

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

Acknowledgments

Thanks to . . .

- Rolf Backofen
- Steffen Heyne
- Tejal Joshi
- Peter Stadler
- Kristin Reiche
- Michael Siebauer
- Wolfgang Otto
- Ivo Hofacker

. . . for dicussions, contributions, future contributions, . . .