

# RNA alignment and folding

with

## Markov chains

**Milad Miladi**

**University of Freiburg**

**TBI Bled meeting**

**February 2019**

# RNA simultaneous alignment and folding

Example of an *alignment*  $\mathcal{A}$  of two RNA sequences  $A$  and  $B$  together with (*non-crossing*) structures  $S = \{a_1, a_2, a_3, a_4, a_5\}$  and  $T = \{b_1, b_2, b_3, b_4, b_5\}$ .

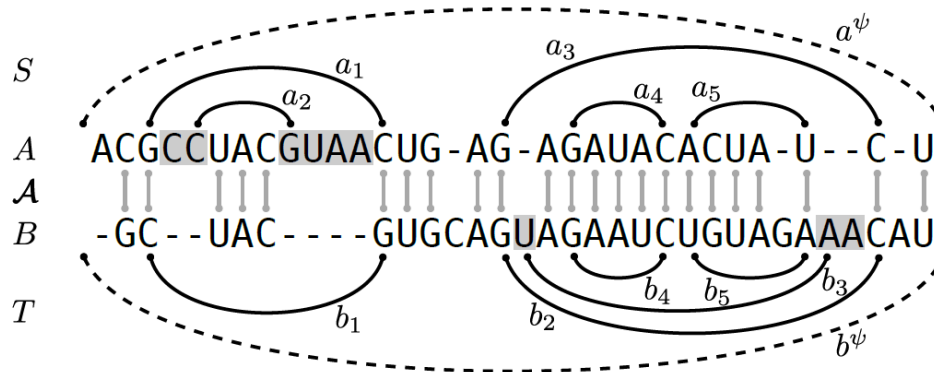


Figure: [Will, Otto, Miladi, Möhl, Backofen; SPARSE; *Bioinf.*, 2015]

# Some of the related pioneering works

1985

**Simultaneous solution of the RNA folding, alignment and protosequence problems**

by: [D. Sankoff](#)

1990-1993

**RNA multi-structure landscapes**

A study based on temperature dependent partition functions

Authors

[Authors and affiliations](#)

S. Bonhoeffer, J. S. McCaskill, P. F. Stadler, P. Schuster

~2004:

**Alignment of RNA base pairing probability matrices** FREE

[Ivo L. Hofacker](#) ✉, [Stephan H. F. Bernhart](#), [Peter F. Stadler](#)

**LOCAL SEQUENCE-STRUCTURE MOTIFS IN RNA**

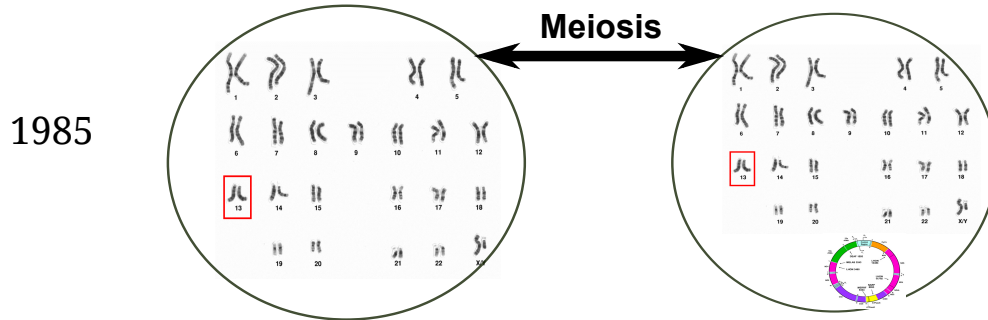
ROLF BACKOFEN and SEBASTIAN WILL

**Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%** FREE

[Jakob Hull Havgaard](#) ✉, [Rune B. Lyngsø](#), [Gary D. Stormo](#), [Jan Gorodkin](#)

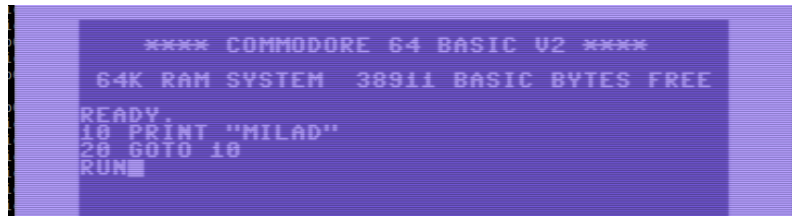
# Meanwhile me

Encoded as unobserved event of a stochastic function



Playing with bro's C64 machine

~1993



~2004

```
procedure bubbleSort( A : list of sortable items )  
  n = length(A)  
  repeat
```

figures adapted from: wikimedia.org



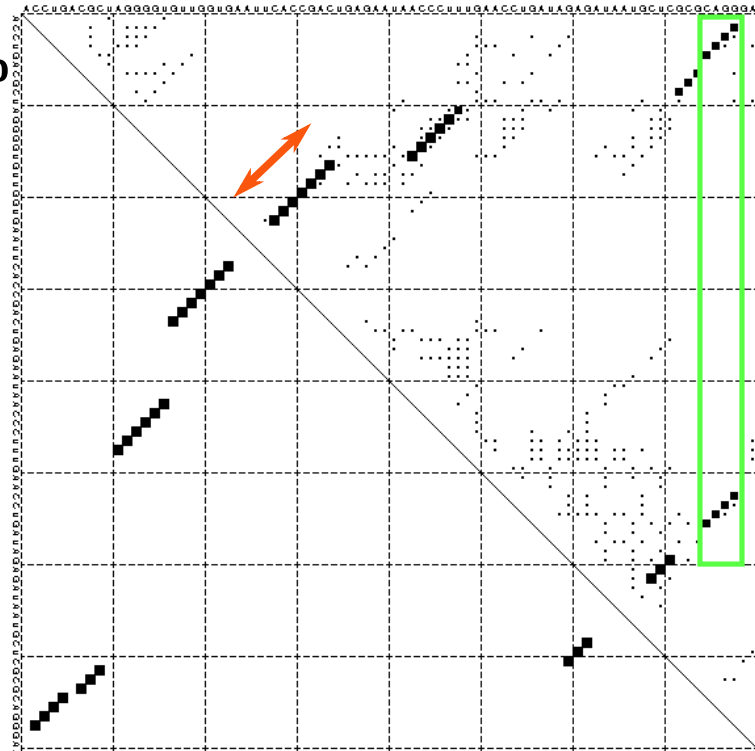
# Prologue: RNA secondary structure probability dotplots

**The abstraction making it feasible to visualize the RNA ensemble is:**

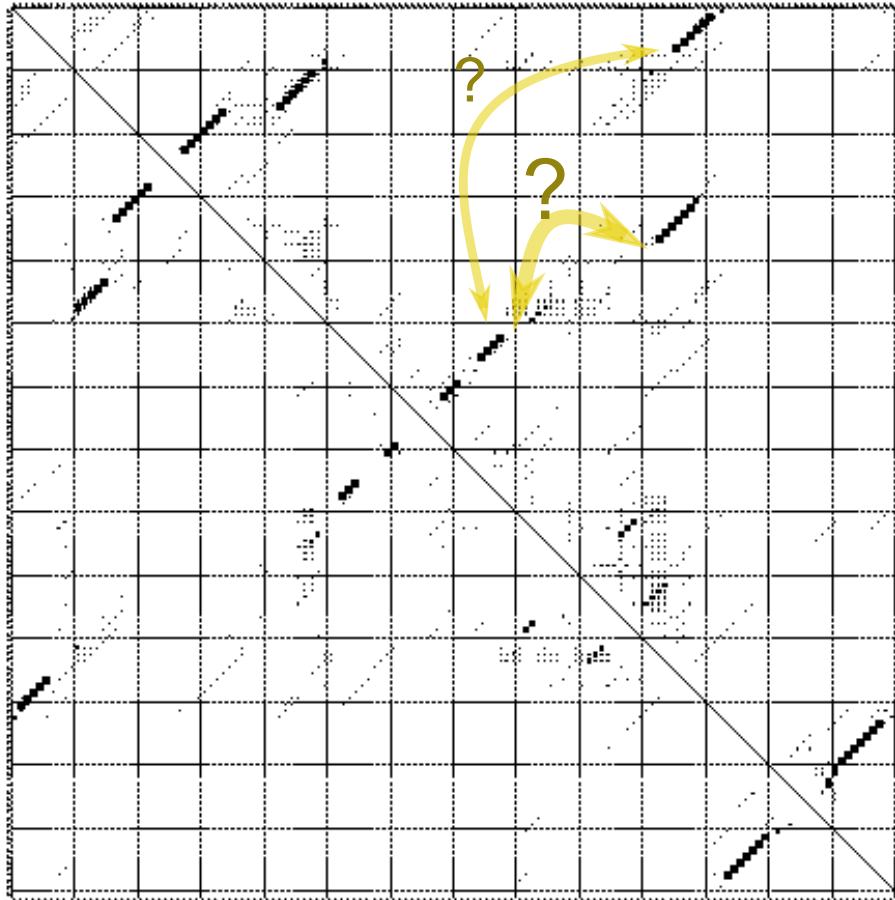
- Base-pair independency or disconnection

Implicit base-pair dependency or relation hints available via:

- Stacking patterns
- Mutual exclusion rules

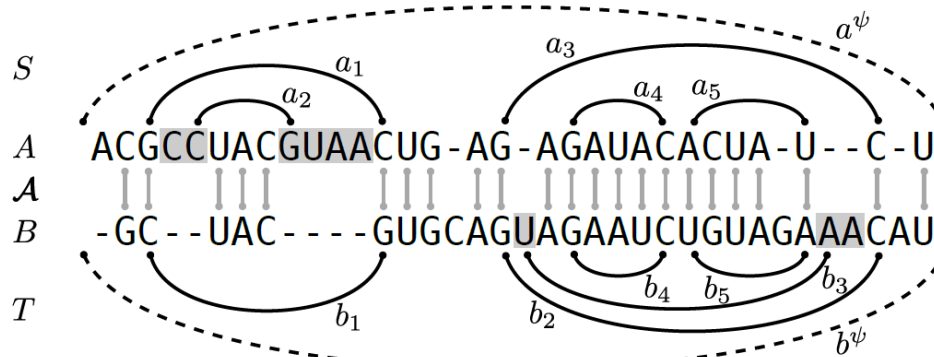


Can we really percept stackings inter-relations?



## Back to the target problem: RNA alignment and folding

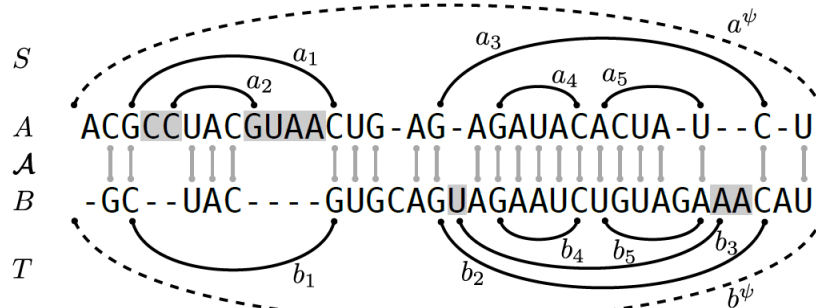
Example of an *alignment*  $\mathcal{A}$  of two *RNA sequences*  $A$  and  $B$  together with (*non-crossing*) structures  $S = \{a_1, a_2, a_3, a_4, a_5\}$  and  $T = \{b_1, b_2, b_3, b_4, b_5\}$ .



- An optimization problem that considers the stability of the predicted structures and the alignment quality
- Objective function:
  - $f(\text{energy\_model}(S, T) + \text{alignment\_score}(A, B))$

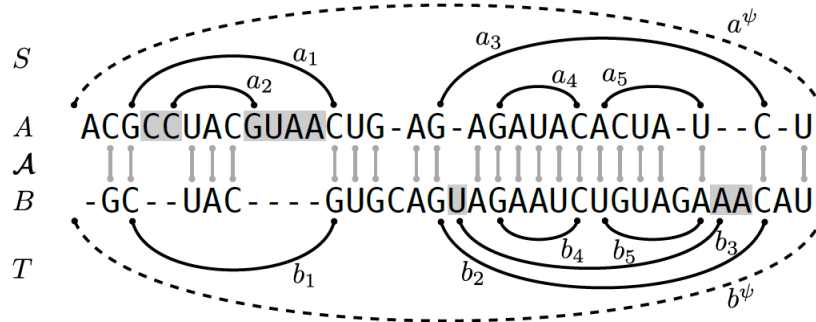
# Sankoff's algorithm for simultaneous RNA alignment and folding

- Sankoff's objective function:
  - $-\text{FreeEnergy}(S) - \text{FreeEnergy}(T) + \text{SequenceAlignScore}(A, B)$
- Linear straightforward(?) combination of:
  - the free energies of two compatible structures
  - the alignment score of the two sequences
- The “golden” standard with efficient dynamic programming recursions
- But computationally expensive  $O(n^6)$  so not suitable ...



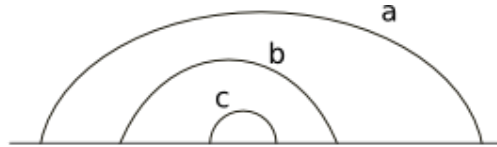
# PMcomp and PMcomp-like objective function

- Using a light-weight base-pair energy model
- Using precomputed McCaskill base-pair probability dotplots to compute the base-pair scores.
- Estimating energy by combining weights and assuming independency of the pairing events
- PMcomp coarse model:
  - $\sum_i \text{score}(a_i) + \sum_i \text{score}(b_i) + \text{SequenceAlignScore}(A, B)$
- PMcomp-like:
  - PMcomp, LocARNA, FoldAlignM(?), SPARSE, ...

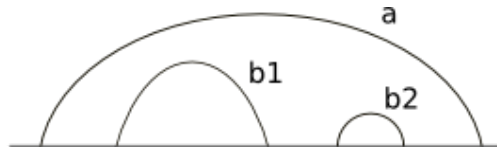


# Sankoff-Zuker's nearest-neighbor energy model

The free energy depends on the neighboring bases and adjacent loops.



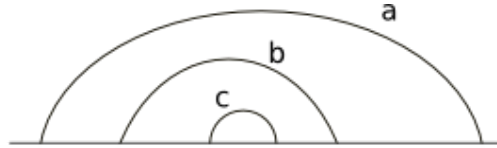
$$E = E(a, b) + E(b, c) + E(c)$$



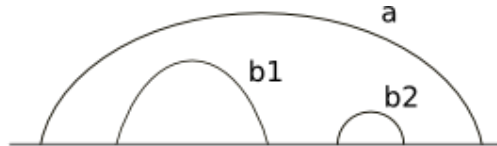
$$E = E(a, b1, b2) + E(b1) + E(b2)$$

# PMcomp base-pair energy model

- An independency between base-pairing probability events
- Efficient computations and flexible pairing, reduces complexity to  $O(n^4)$



$$\text{Score (a, b, c)} = ( P (a) \cdot P (b) \cdot P (c) ) / P (p_0^3)$$

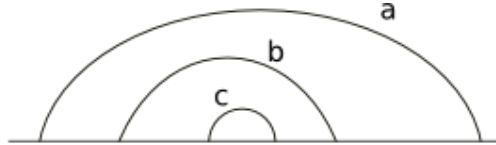


$$\text{Score (a, b1, b2)} \sim ( P (a) \cdot P (b1) \cdot P (b2) ) / P (p_0^3)$$

$p_0$ : a length-dependent estimation of the minimum significant/background probability

# A Markov chain model: non-branching loops

For a sub-structure  $S$  with set of basepairs  $\{a,b,c\}$  let have:



$$P(a, b, c) = P(a) \cdot P(b, c | a)$$

nearest-neighbor rule:

$$= P(a) \cdot P(b | a) \cdot P(c | a, b)$$

$$= P(a) \cdot P(b | a) \cdot P(c | b)$$

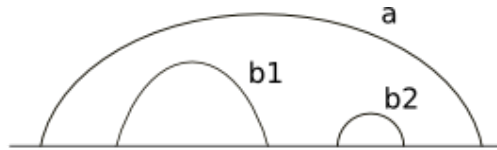
Basic math for conditional probabilities to the joint probabilities.

$$P(a, b, c) = P(a) \cdot P(b | a) \cdot P(c | b)$$

$$= P(a) \cdot P(b, a) / P(a) \cdot P(c, b) / P(b)$$



# A Markov chain model: multi-loops



$$P(a, b1, b2) = P(a) \cdot P(b1, b2 | a)$$

Nearest-neighbor rule not enough for decomposition:

$$= P(a) \cdot P(b1 | a) \cdot P(b2 | a, b1)$$

Assuming multiloop independency:

$$= P(a) \cdot P(b1 | a) \cdot P(b2 | a)$$

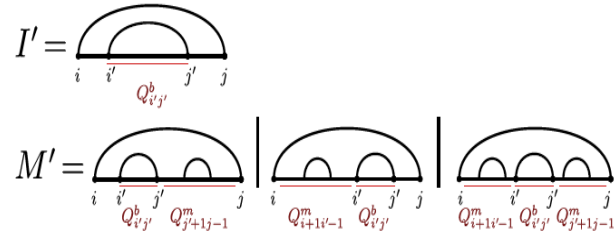
A better trick:

$$= P(a) \cdot \min( P(b1 | a), P(b2 | a) )$$

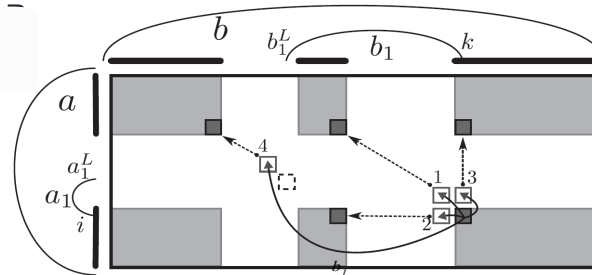
# Alright but how to get chain probabilities?

- Extending the base-pair stacking probabilities (RNAfold -p2) to loops

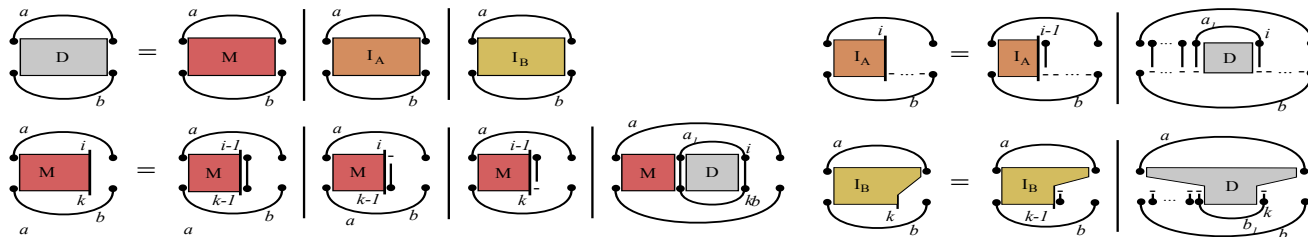
## A: Joint probability pre-computations



## B: To sparsify the alignment search space



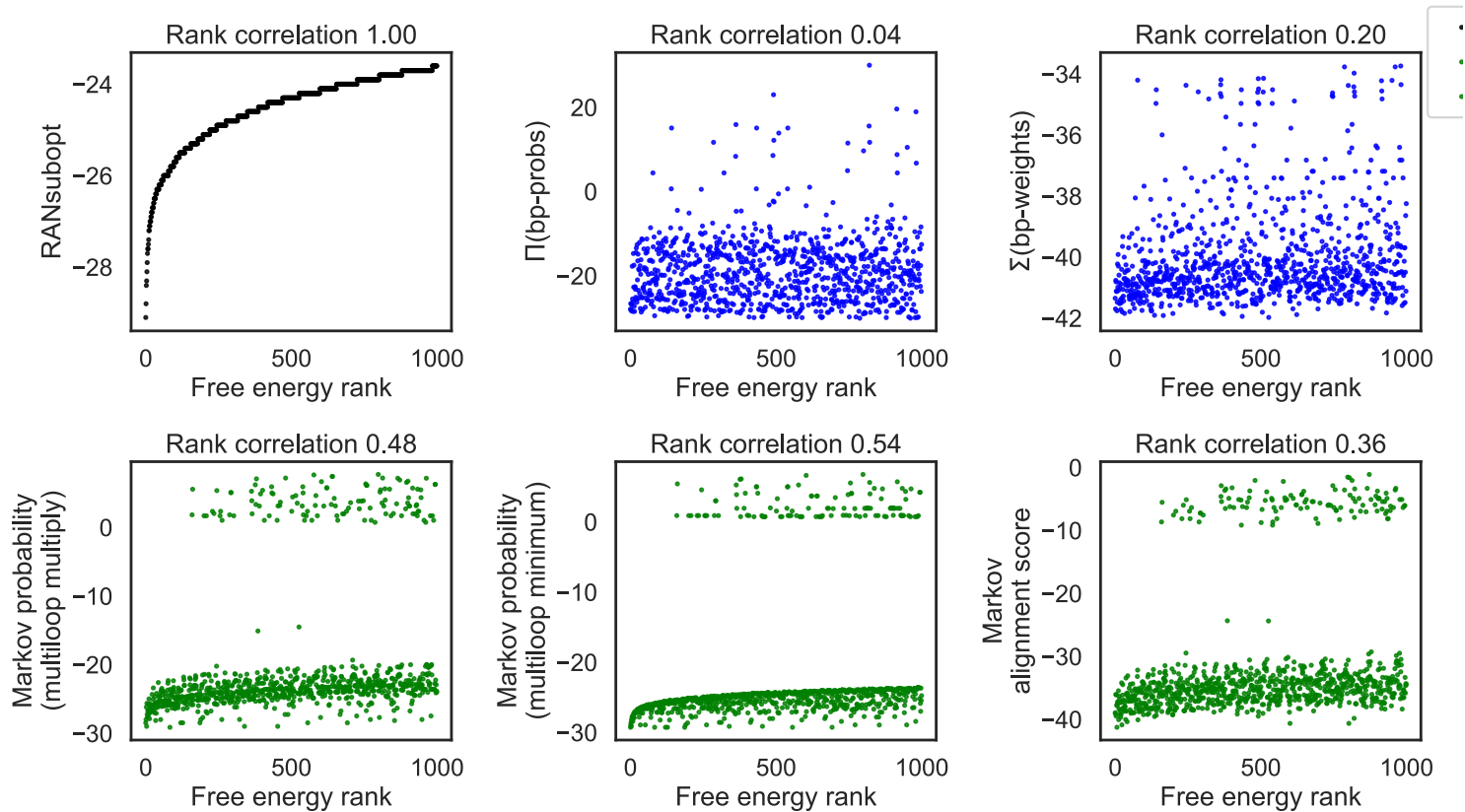
## C: Using loop-closing aware recursions



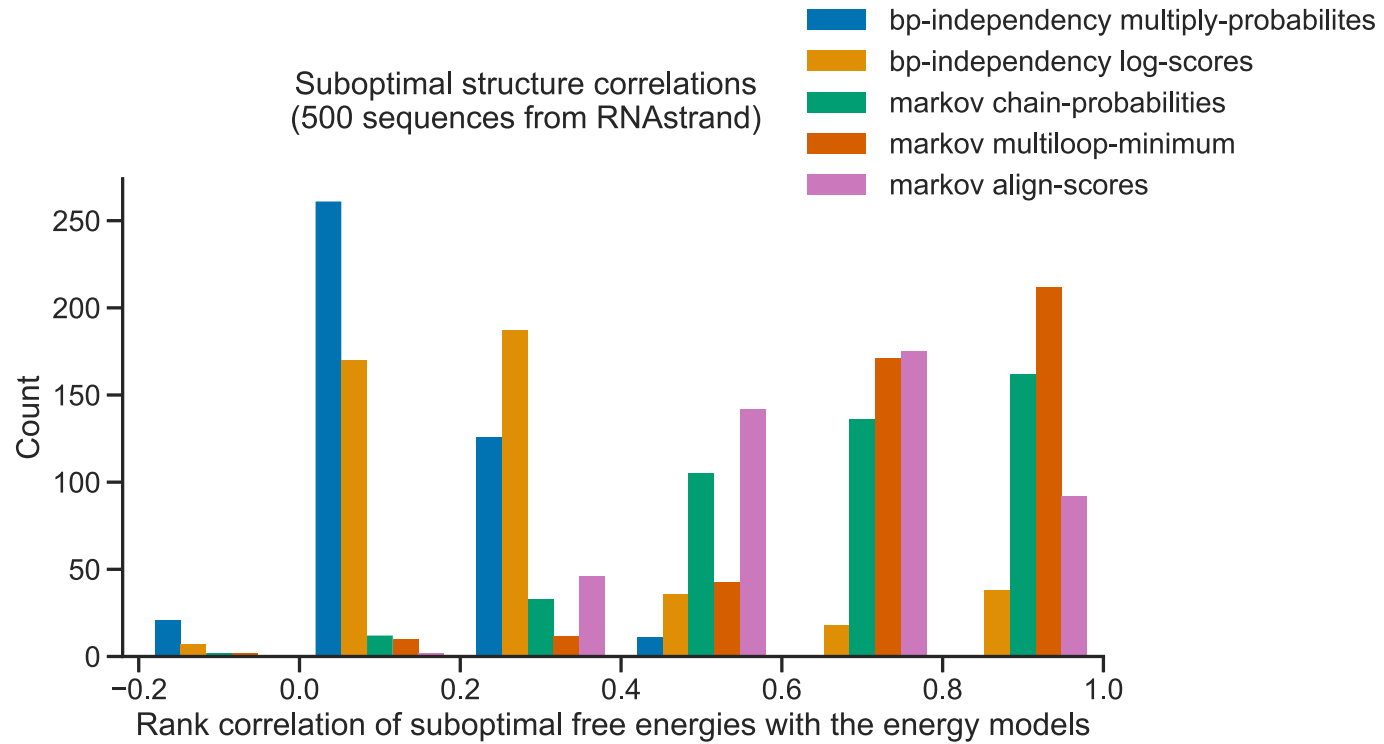
Figures: [Will, Otto, Miladi, Möhl, Backofen; SPARSE; *Bioinf.*, 2015], [Otto, Möhl, Heyne, Amit, Landau, Backofen, Will; ExaRNA-P: *BMC Bioinf.*, 2014]

# Evaluation of the energy model: rnasubopt trna








- In this plot, dots represents suboptimal structures of a tRNA sequence, sorted by free energy.



# Evaluate energy models on RNAstrand dataset



# Pankov: Probabilistic Sankoff's-like alignment with Markov chains

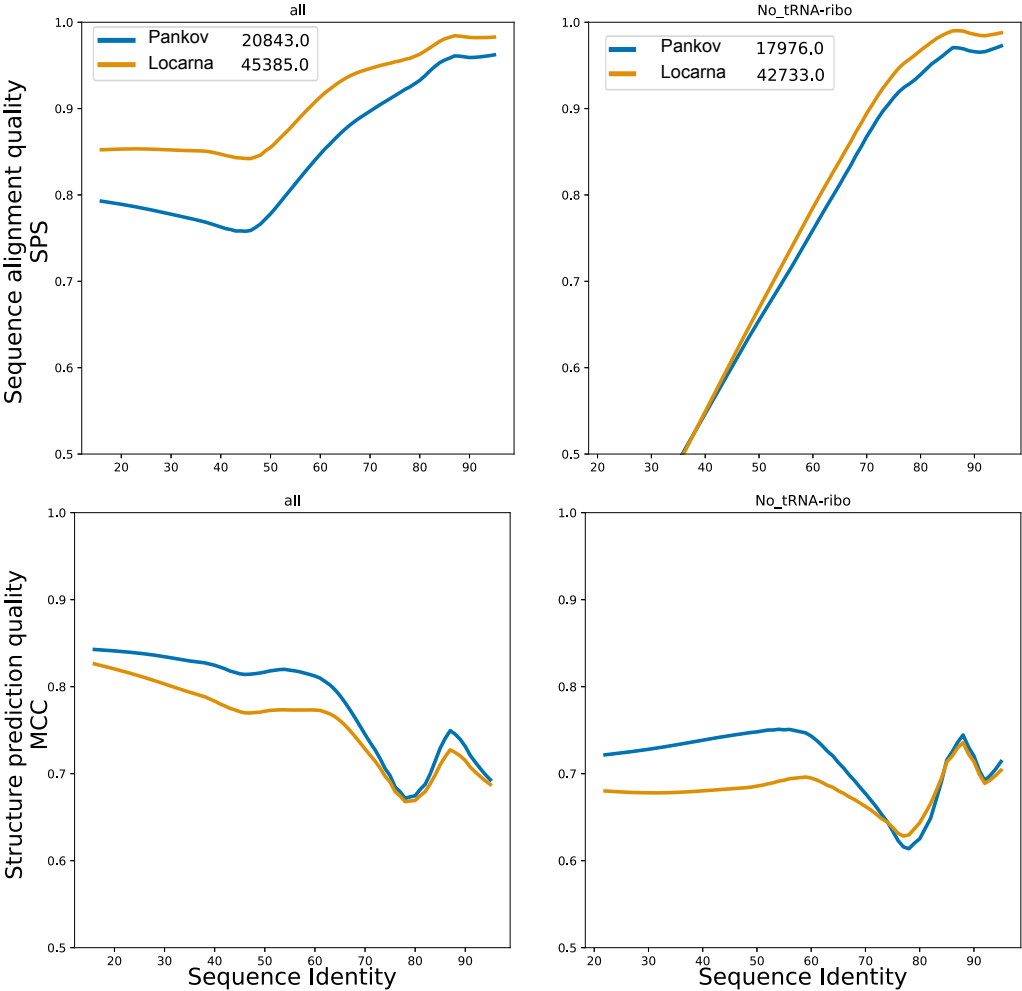
Name	Size	Last commit
 Bralibase		2019-02-07
 RNAStrand		2019-02-07
 lib		2016-10-06
 src		2016-10-06
 .gitignore	54 B	2016-09-07
 README.md	80 B	1 minute ago
 conda-env-pankoff.yml	4.79 KB	2019-02-06

## README.md

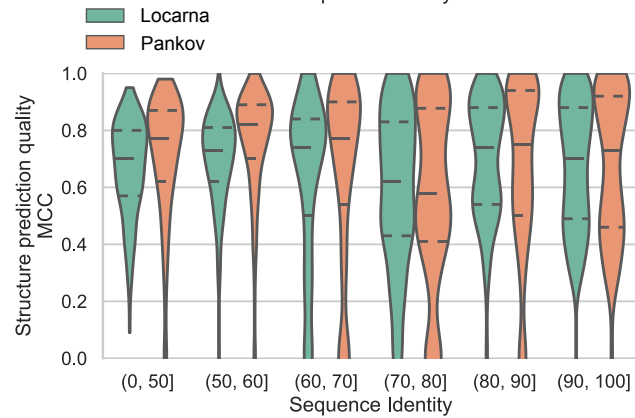
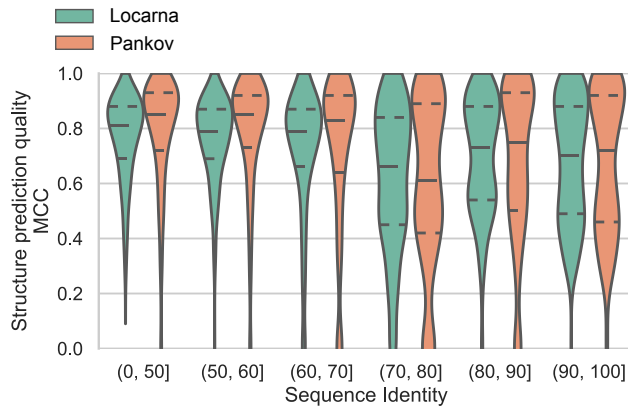
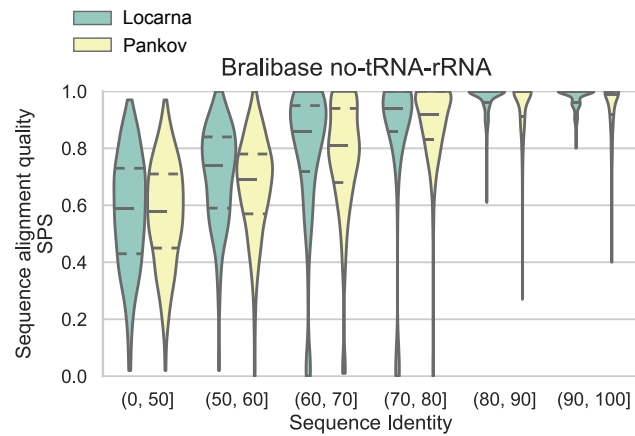
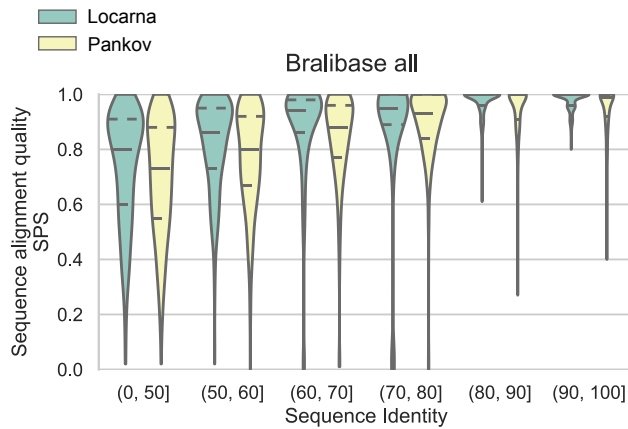
Pankov: Probabilistic **Sankoff's**-like alignment with **Markov** chains

Currently protected access, feel free to contact me: [bitbucket.org:mmiladi/pankov](https://bitbucket.org:mmiladi/pankov)

# Primary evaluation on Bralibase dataset



# "Dent"less Bralibase; same data as the previous slide



## Conclusion and speculation

- Pankov: Our closest attempt to enter the Eutopia
- Flexibility introduced by the base-pairing energy model may have benefits beyond a computational complexity reduction
- It's actually very hard to combine the folding-energy and alignment scoring components that have two different natures
- The conditional probability of loops has the potential to be applied for other purposes as well
  - Please reach me if you have ideas

## Acknowledgment

Rolf, Sebastian, Teresa and the Bionf-Freiburg team.



Thanks to everyone for sharing your joy of TBI-Bled with me these years!

- The dilemma of TBI-Bled Villa

