

A Simple Protocol for the Inference of RNA Global Pairwise Alignments

Felix Karg

27. November 2018

University of Freiburg



Content

needle

Sequence Alignment

Needleman-Wunsch

LocARNA

Tree-based sequence
alignment

gardenia

RNA StrAT

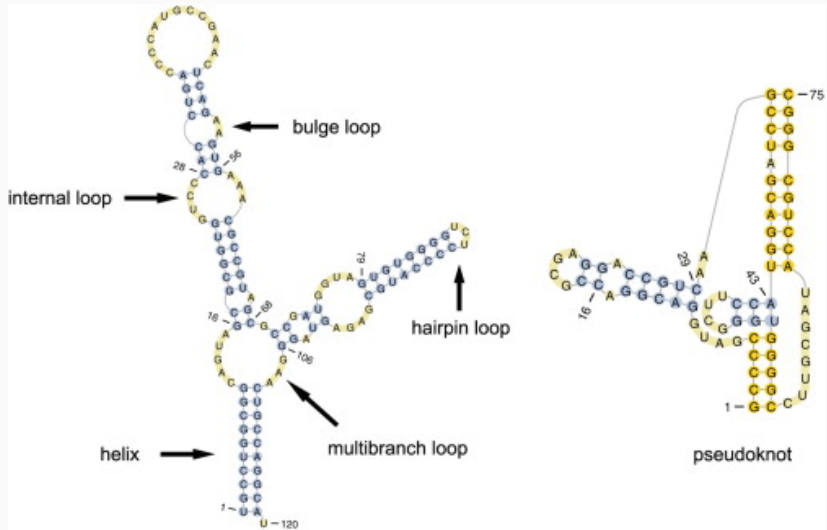
RNAdistance

RNAforester

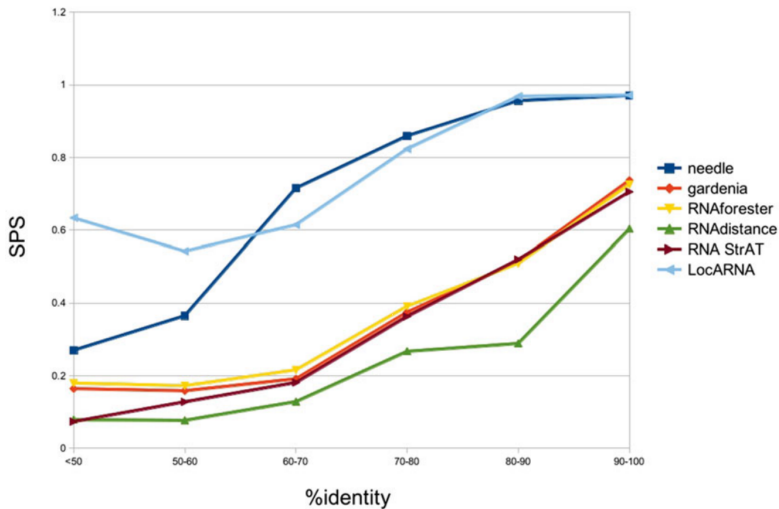
Comparison

Sources

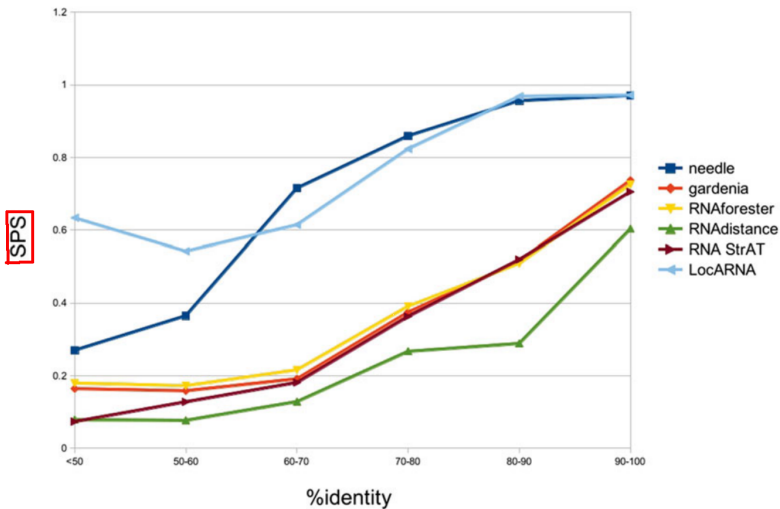
RNA secondary structure



Predicted Secondary Structures



Predicted Secondary Structures



Sum of Pairs Score

SPS - introduction

Sum of Pairs Score

Used to measure the alignment of two RNA sequences

SPS - introduction

Sum of Pairs Score

Used to measure the similarity of two RNA sequences

Sequence Similarity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Similarity:

$1 - (\text{edit distance} / \text{unaligned length of shorter sequence})$

Sequence Similarity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Similarity:

$1 - (\text{edit distance} / \text{unaligned length of shorter sequence})$

Sequence Similarity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Similarity: $60\% = 1 - (2 / 5)$

$1 - (\text{edit distance} / \text{unaligned length of shorter sequence})$

Sequence Similarity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Similarity:

$1 - (\text{edit distance} / \text{unaligned length of shorter sequence})$

Sequence Similarity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Similarity: $60\% = 1 - (2 / 5)$

$1 - (\text{edit distance} / \text{unaligned length of shorter sequence})$

Sequence Similarity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Similarity:

$1 - (\text{edit distance} / \text{unaligned length of shorter sequence})$

Sequence Similarity - Example

A: AAGGCTT

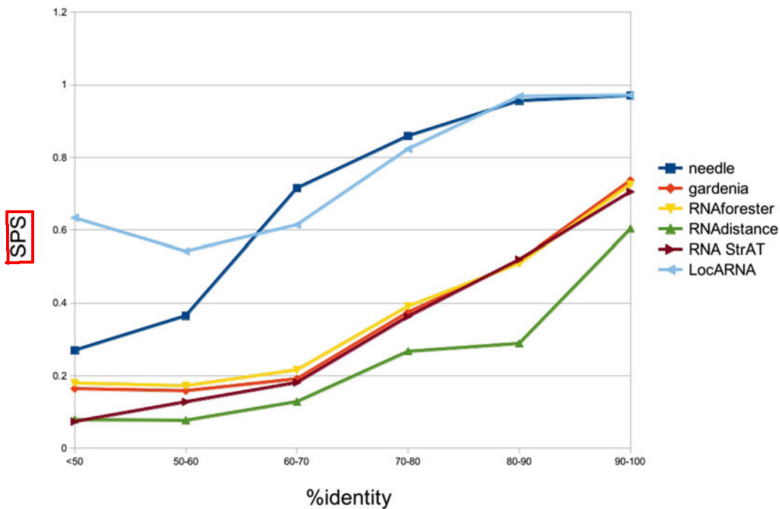
B: AAGGC

C: AAGGCAT

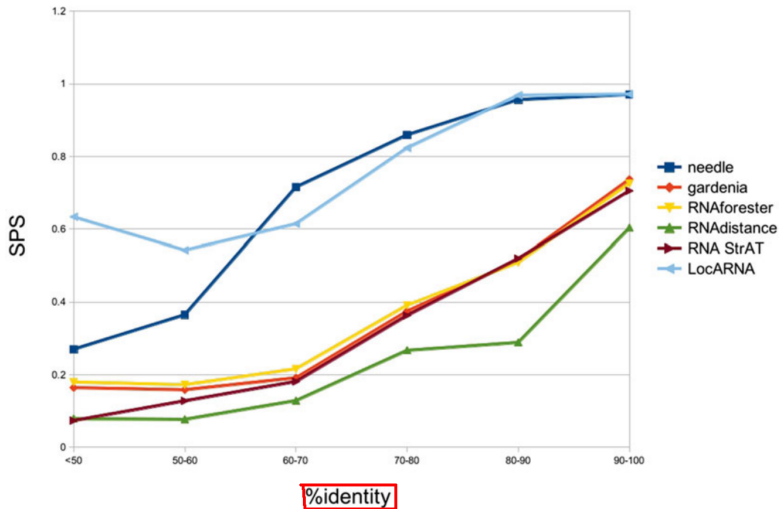
Similarity: $86\% = 1 - (1 / 7)$

$1 - (\text{edit distance} / \text{unaligned length of shorter sequence})$

Predicted Secondary Structures



Predicted Secondary Structures



Sequence Identity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Identity:

Identical nucleotides / shorter sequence length

Sequence Identity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Identity:

Identical nucleotides / shorter sequence length

Sequence Identity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Identity: 100%

Identical nucleotides / shorter sequence length

Sequence Identity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Identity:

Identical nucleotides / shorter sequence length

Sequence Identity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Identity: 100%

Identical nucleotides / shorter sequence length

Sequence Identity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Identity:

Identical nucleotides / shorter sequence length

Sequence Identity - Example

A: AAGGCTT

B: AAGGC

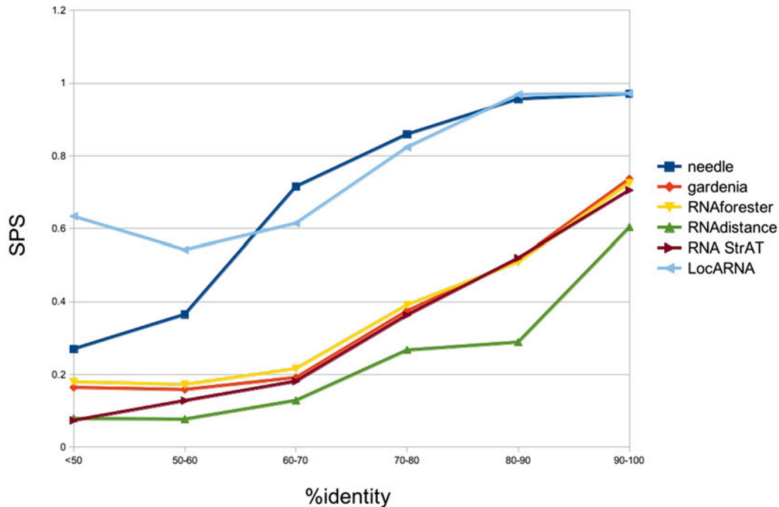
C: AAGGCAT

Identity: $85\% = 6 / 7$

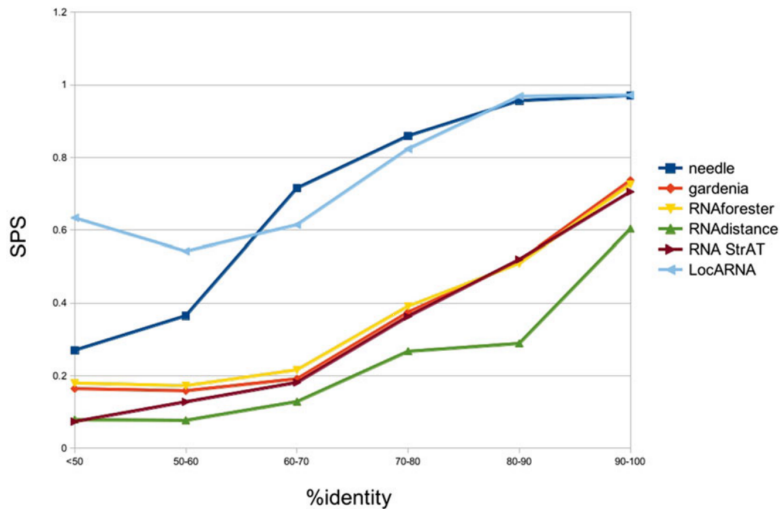
Identical nucleotides / shorter sequence length

Explain: predicted (RNAfold), test setup

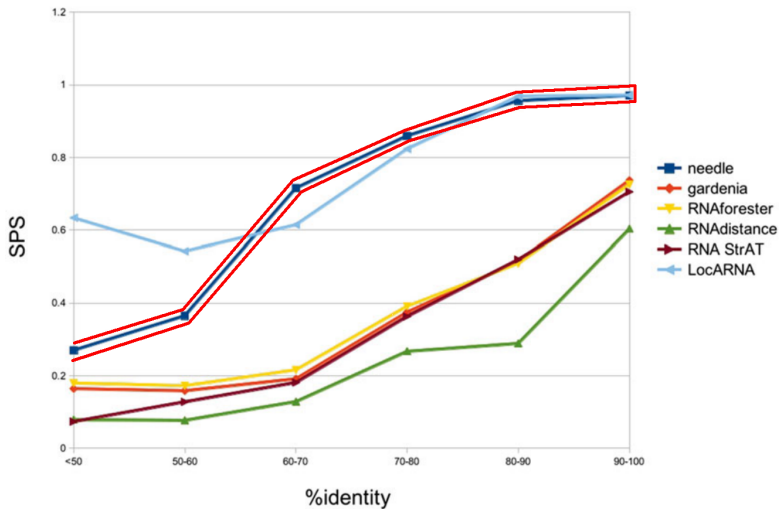
Predicted Secondary Structures



Predicted Secondary Structures



Predicted Secondary Structures



Content

needle

Sequence Alignment

Needleman-Wunsch

LocARNA

Tree-based sequence
alignment

gardenia

RNA StrAT

RNAdistance

RNAforester

Comparison

Sources

- Sequence based Algorithm

- Sequence based Algorithm
- part of the EMBOSS package

- Sequence based Algorithm
- part of the EMBOSS package
- Implementation of the Needleman-Wunsch algorithm

Needleman-Wunsch Algorithm

Sometimes also called:

Needleman-Wunsch Algorithm

Sometimes also called:

- Optimal matching algorithm

Needleman-Wunsch Algorithm

Sometimes also called:

- Optimal matching algorithm
- Global alignment technique

needle

Sequence Alignment

Needleman-Wunsch

LocARNA

Tree-based sequence
alignment

gardenia

RNA StrAT

RNAdistance

RNAforester

Comparison

Sources

Sequence Alignment

- local sequence alignment

Sequence Alignment

- local sequence alignment
- global sequence alignment

Sequence Alignment

- local sequence alignment
- global sequence alignment
- glocal sequence alignment

Sequence Alignment - local

RNA sequence A



RNA sequence B

Sequence Alignment - global

RNA sequence A



RNA sequence B

Sequence Alignment - glocal

RNA sequence A



RNA sequence B

Content

needle

Sequence Alignment

Needleman-Wunsch

LocARNA

Tree-based sequence
alignment

gardenia

RNA StrAT

RNAdistance

RNAforester

Comparison

Sources

Needleman-Wunsch Algorithm - introduction

We need a scoring system.

Needleman-Wunsch Algorithm - introduction

We need a scoring system.

Example:

Needleman-Wunsch Algorithm - introduction

We need a scoring system.

Example:

- Match: 1
- Mismatch: -1
- Indel: -1

Needleman-Wunsch Algorithm

Two sequences to compare:

Needleman-Wunsch Algorithm

Two sequences to compare:

GCATGCU

GATTACA

Needleman-Wunsch Algorithm - Example

		G	C	A	T	G	C	U
G								
A								
T								
T								
A								
C								
A								

Needleman-Wunsch Algorithm - Example

		G	C	A	T	G	C	U
	0							
G								
A								
T								
T								
A								
C								
A								

Needleman-Wunsch Algorithm - Example

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1							
A	-2							
T	-3							
T	-4							
A	-5							
C	-6							
A	-7							

Needleman-Wunsch Algorithm - Example

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	X						
A	-2							
T	-3							
T	-4							
A	-5							
C	-6							
A	-7							

Needleman-Wunsch Algorithm - Example

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1						
A	-2							
T	-3							
T	-4							
A	-5							
C	-6							
A	-7							

Needleman-Wunsch Algorithm - Example

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	X					
A	-2	Y						
T	-3							
T	-4							
A	-5							
C	-6							
A	-7							

Needleman-Wunsch Algorithm - Example

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0					
A	-2	0						
T	-3							
T	-4							
A	-5							
C	-6							
A	-7							

Needleman-Wunsch Algorithm - Example

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Needleman-Wunsch Algorithm - Best Matches

		G	C	A	T	G	C	U	
		0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5	
A	-2	0	0	1	0	-1	-2	-3	
T	-3	-1	-1	0	2	1	0	-1	
T	-4	-2	-2	-1	1	1	0	-1	
A	-5	-3	-3	-1	0	0	0	-1	
C	-6	-4	-2	-2	-1	-1	1	0	
A	-7	-5	-3	-1	-2	-2	0	0	

Needleman-Wunsch Algorithm - Results

Sequences	Best alignments		
GCATGCU	GCATG-CU	GCA-TGCU	GCAT-GCU
GATTACA	G-ATTACA	G-ATTACA	G-ATTACA

Content

needle

Sequence Alignment

Needleman-Wunsch

LocARNA

Tree-based sequence
alignment

gardenia

RNA StrAT

RNAdistance

RNAforester

Comparison

Sources

Two steps:

Two steps:

- create base pair probability matrix using RNAfold

Two steps:

- create base pair probability matrix using RNAfold
- using these as guide for optimal alignment

Two steps:

- create base pair probability matrix using RNAfold
- using these as guide for optimal alignment

(folding and aligning)

Content

needle

Sequence Alignment

Needleman-Wunsch

LocARNA

Tree-based sequence
alignment

gardenia

RNA StrAT

RNAdistance

RNAforester

Comparison

Sources

Content

needle

Sequence Alignment

Needleman-Wunsch

LocARNA

Tree-based sequence
alignment

gardenia

RNA StrAT

RNAdistance

RNAforester

Comparison

Sources

Content

needle

Sequence Alignment

Needleman-Wunsch

LocARNA

Tree-based sequence
alignment

gardenia

RNA StrAT

RNAdistance

RNAforester

Comparison

Sources

Content

needle

Sequence Alignment

Needleman-Wunsch

LocARNA

Tree-based sequence
alignment

gardenia

RNA StrAT

RNAdistance

RNAforester

Comparison

Sources

Content

needle

Sequence Alignment

Needleman-Wunsch

LocARNA

Tree-based sequence
alignment

gardenia

RNA StrAT

RNAdistance

RNAforester

Comparison

Sources

Content

needle

Sequence Alignment

Needleman-Wunsch

LocARNA

Tree-based sequence
alignment

gardenia

RNA StrAT

RNAdistance

RNAforester

Comparison

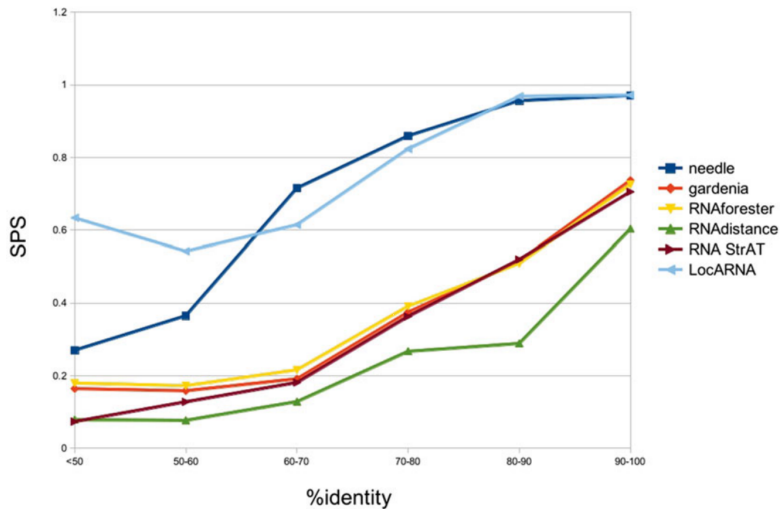
Sources

Run time Comparison

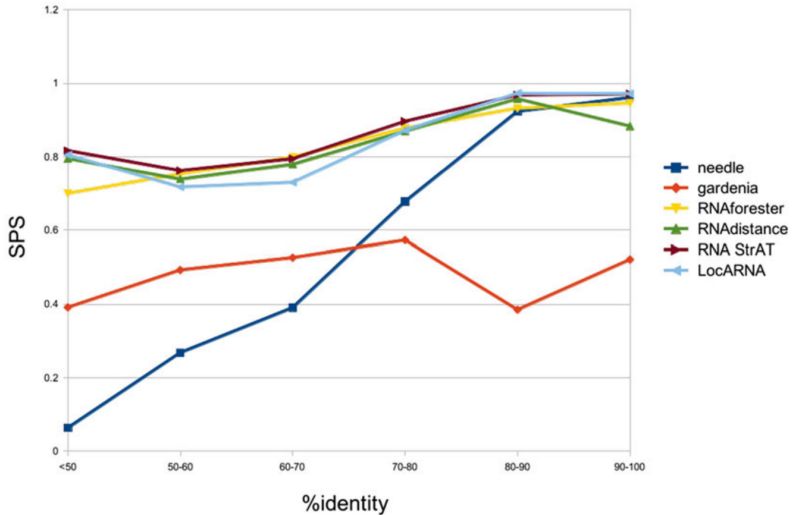
Mean running time (in seconds) per sequence

	bralibase	RNAspa	RNAstrand
needle	0.01	0.01	0.01
gardenia	0.01	0.01	0.02
RNA StrAT	0.02	0.32	0.46
LocARNA	0.02	0.08	0.02
RNAdistance	0.01	0.01	0.01
RNAforester	0.03	0.73	0.81

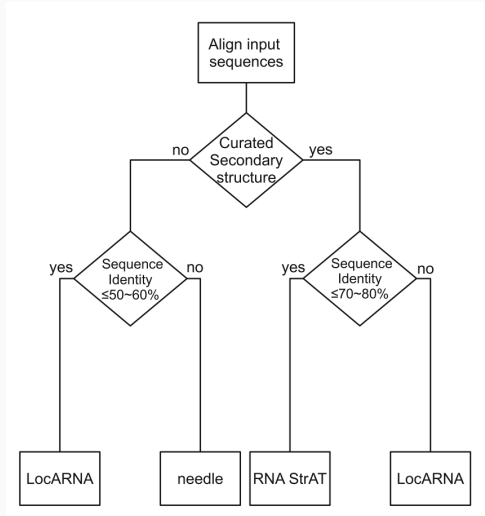
Predicted Secondary Structures



Curated Secondary Structures



Proposed Workflow



Content

needle

Sequence Alignment

Needleman-Wunsch

LocARNA

Tree-based sequence
alignment

gardenia

RNA StrAT

RNAdistance

RNAforester

Comparison

Sources

The slides can be found at:



Github

`https://github.com/fkarg/things-to-talk-about/
tree/master/proseminar`



Image

https://upload.wikimedia.org/wikipedia/commons/3/3f/Needleman-Wunsch_pairwise_sequence_alignment.png



Secondary structures Image

<https://www.sciencedirect.com/science/article/pii/B9780124200371000014>