A Simple Protocol for the Inference of RNA Global Pairwise Alignments

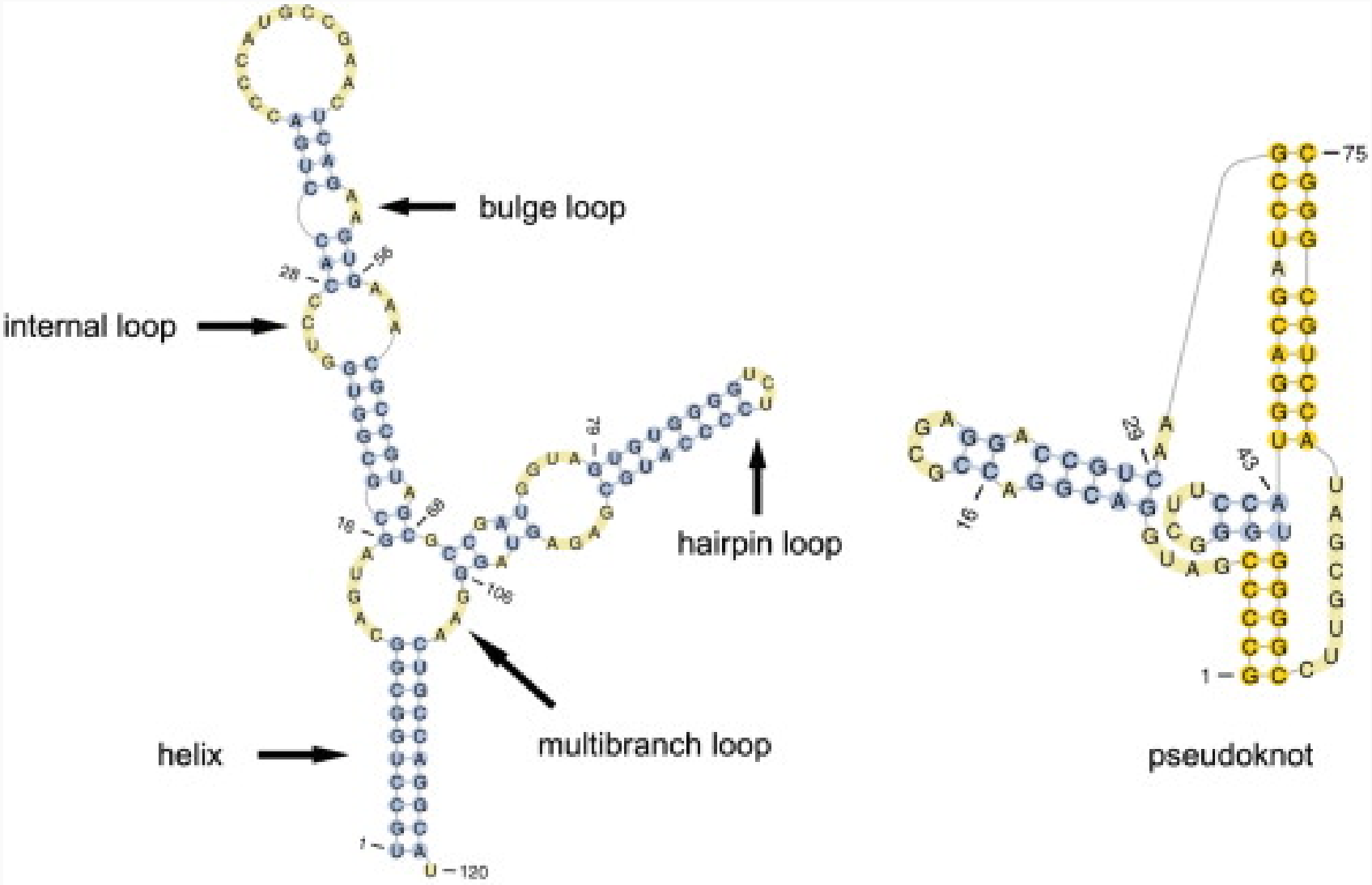
Felix Karg

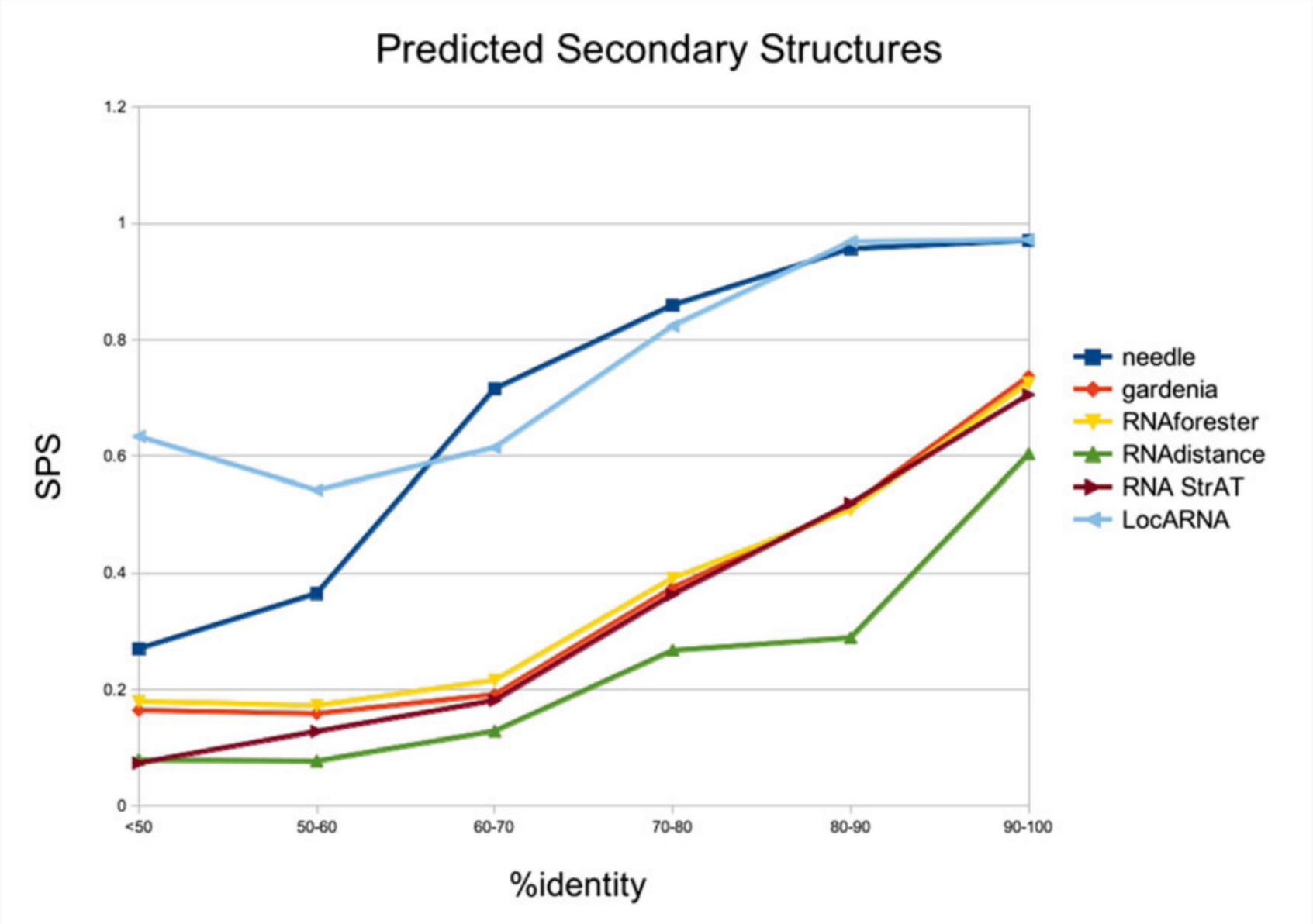
26. August 2019

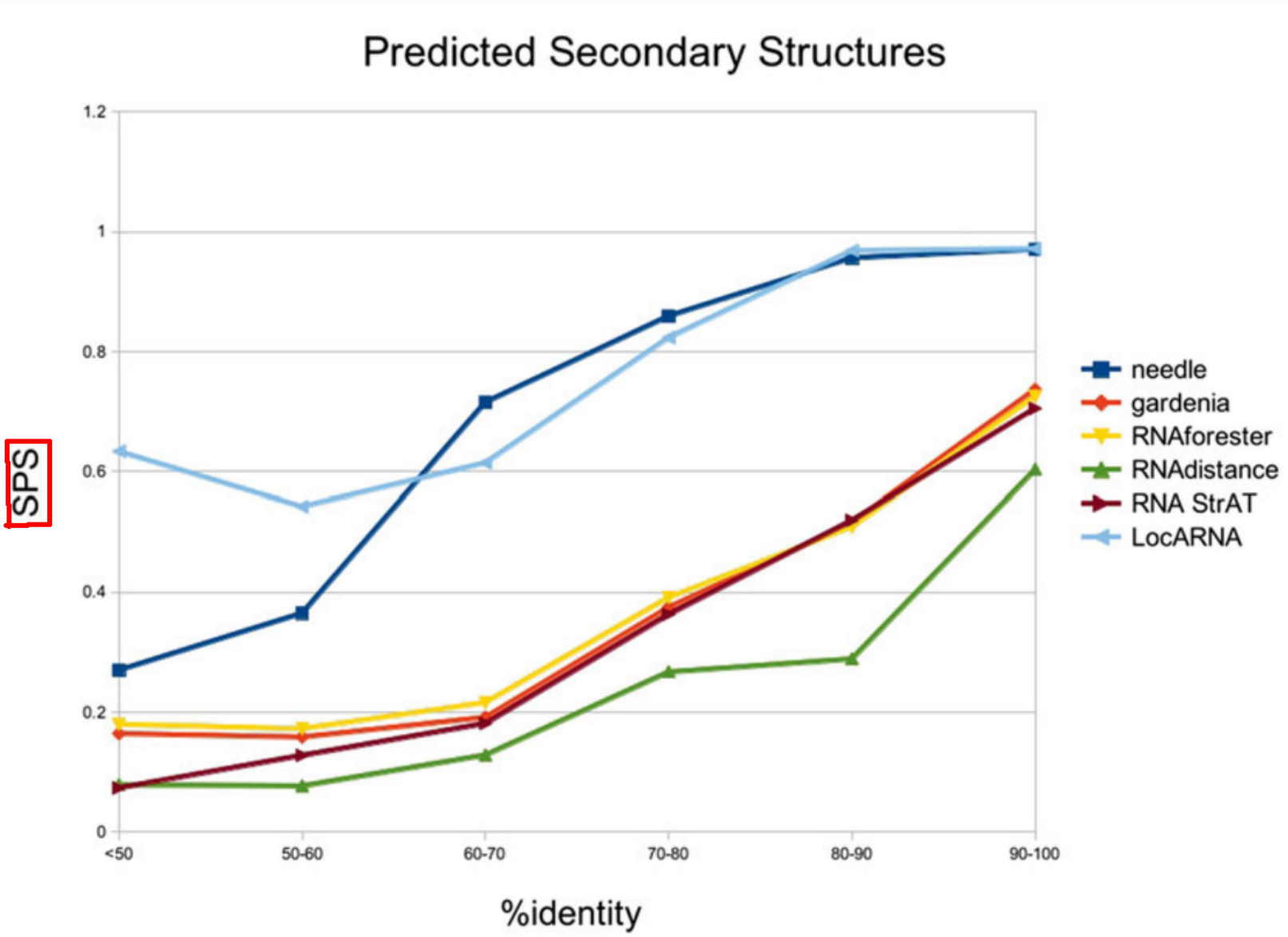
University of Freiburg

Content

RNA secondary structure




Predicted Secondary Structures
SPS
1.2
1
0.8
0.6
0.4
0.2
0
<50
50-60
60-70
70-80
80-90
90-100
%identity
needle
gardenia
RNAforester
RNAdistance
RNA StrAT
LocARNA


Predicted Secondary Structures
SPS
1.2
1
0.8
0.6
0.4
0.2
0
<50
50-60
60-70
70-80
80-90
90-100
%identity
needle
gardenia
RNAforester
RNAdistance
RNA StrAT
LocARNA

SPS - introduction

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 similiarity of two RNA sequences

Sequence Similiarity - Example

A: AAGGCTT
B: AAGGC
C: AAGGCAT

Similiarity:
1 - (edit distance / unaligned length of shorter sequence)

Sequence Similiarity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Similiarity:

1 - (edit distance / unaligned length of shorter sequence)

Sequence Similiarity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Similiarity: 60% = 1 - (2 / 5)

1 - (edit distance / unaligned length of shorter sequence)

Sequence Similiarity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Similiarity:

1 - (edit distance / unaligned length of shorter sequence)

Sequence Similiarity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Similiarity: 60% = 1 - (2 / 5)

1 - (edit distance / unaligned length of shorter sequence)

Sequence Similiarity - Example

A: AAGGCTT

B: AAGGC

C: AAGGCAT

Similiarity:

1 - (edit distance / unaligned length of shorter sequence)

Sequence Similiarity - Example

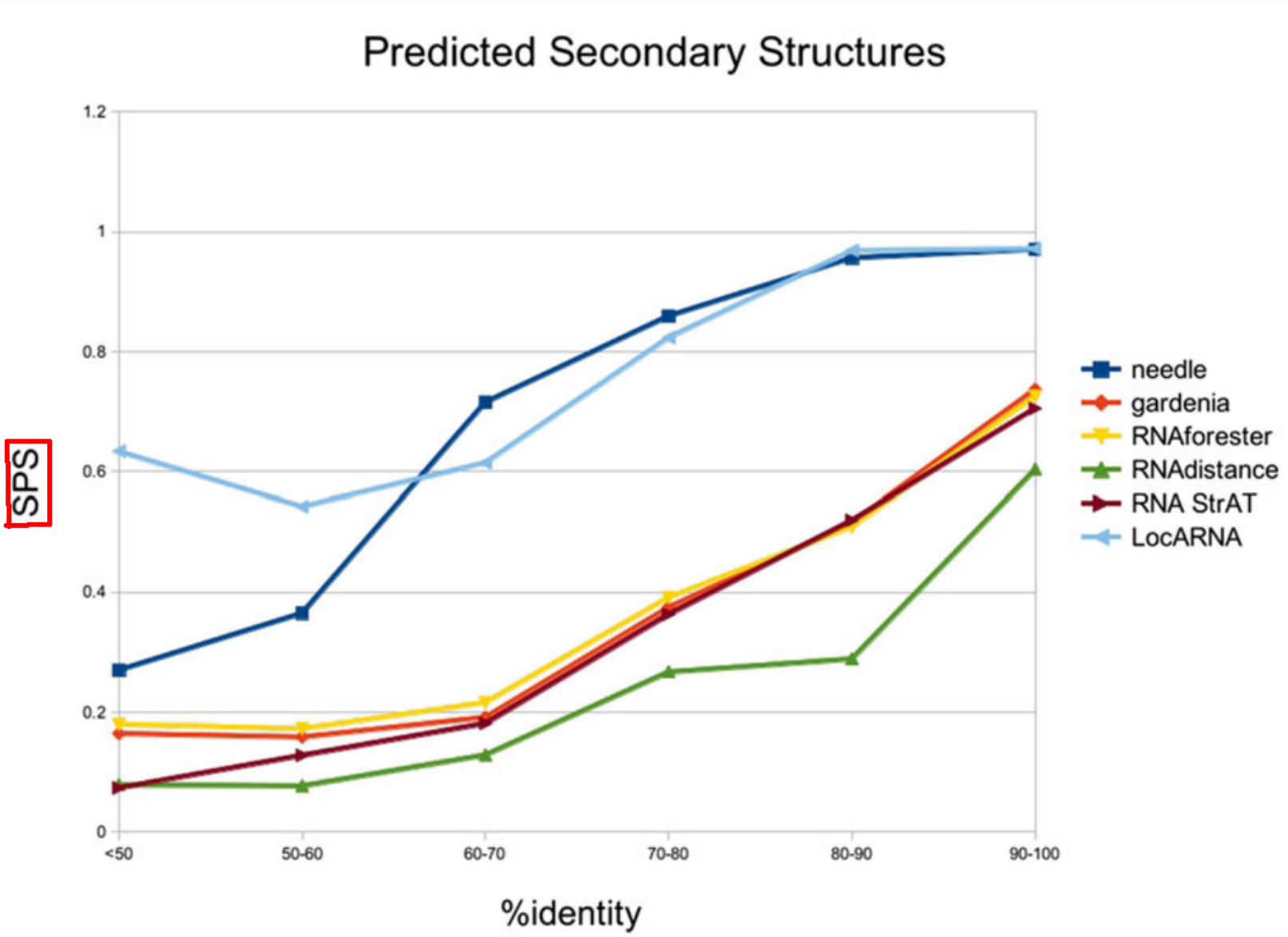
A: AAGGCTT

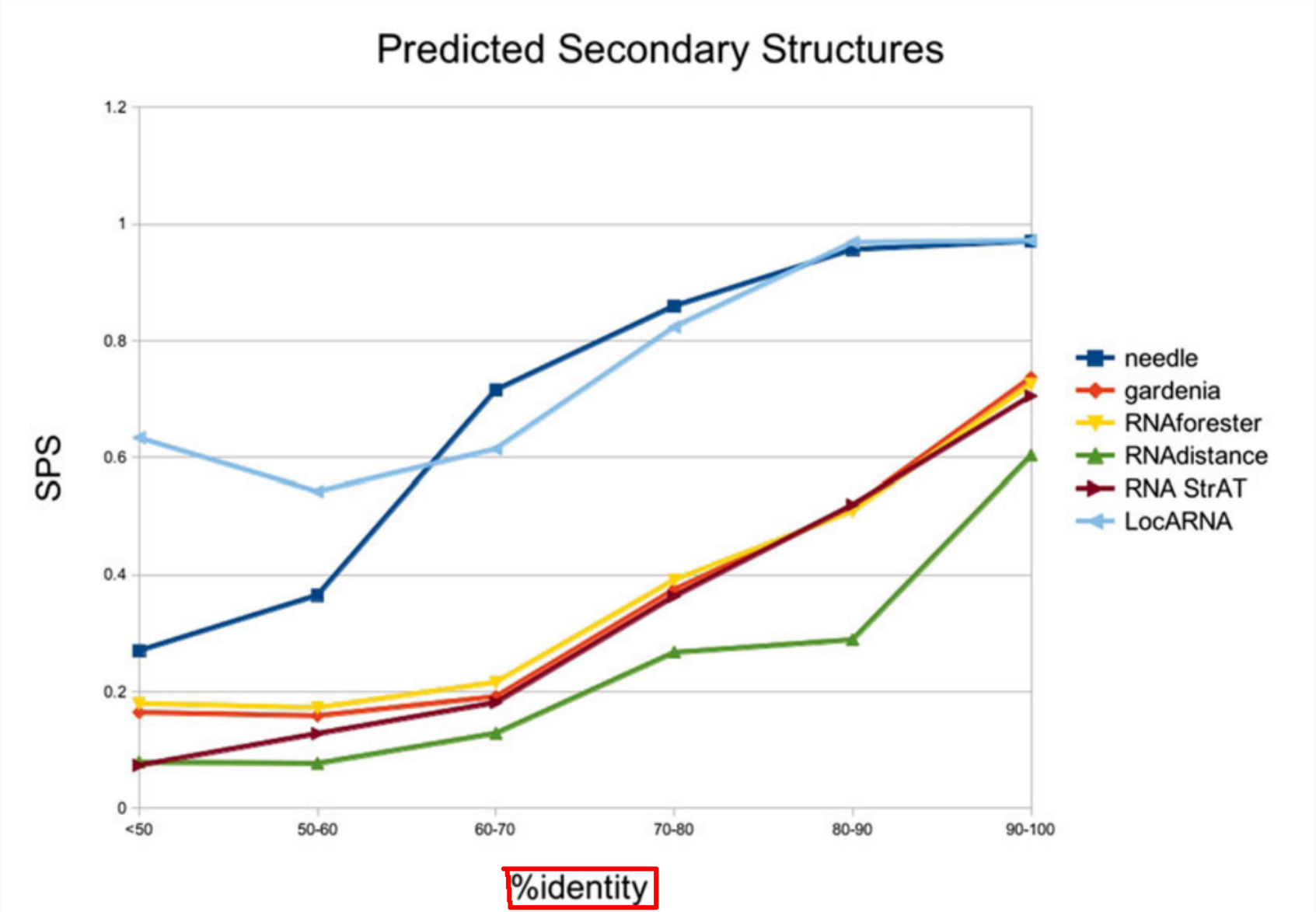
B: AAGGC

C: AAGGCAT

Similiarity: 86% = 1 - (1 / 7)

1 - (edit distance / unaligned length of shorter sequence)


Predicted Secondary Structures
SPS
1.2
1
0.8
0.6
0.4
0.2
0
<50
50-60
60-70
70-80
80-90
90-100
%identity
needle
gardenia
RNAforester
RNAdistance
RNA StrAT
LocARNA


Predicted Secondary Structures
SPS
1.2
1
0.8
0.6
0.4
0.2
0
<50
50-60
60-70
70-80
80-90
90-100
%identity
needle
gardenia
RNAforester
RNAdistance
RNA StrAT
LocARNA

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

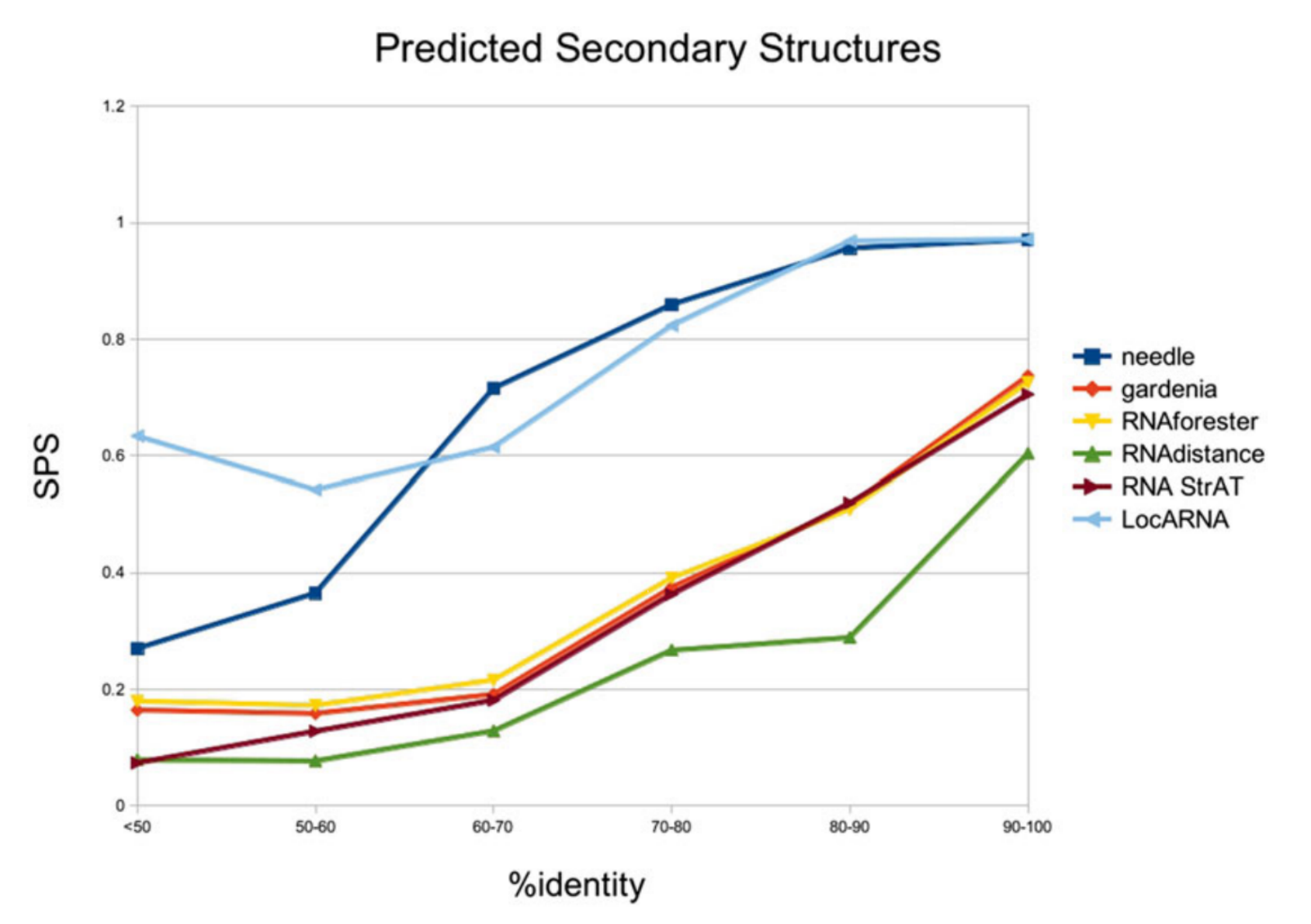
B: AAGGC

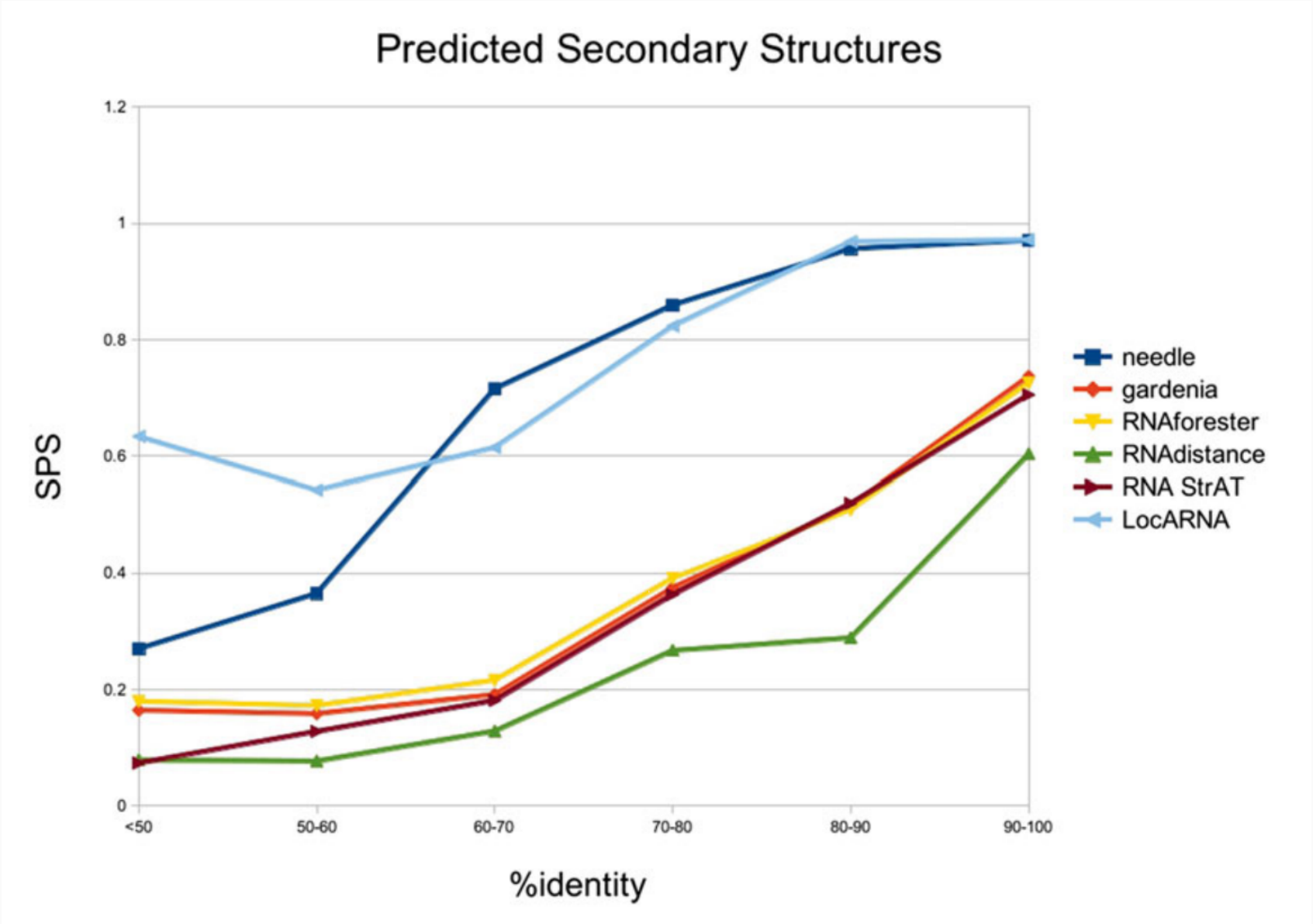
C: AAGGCAT

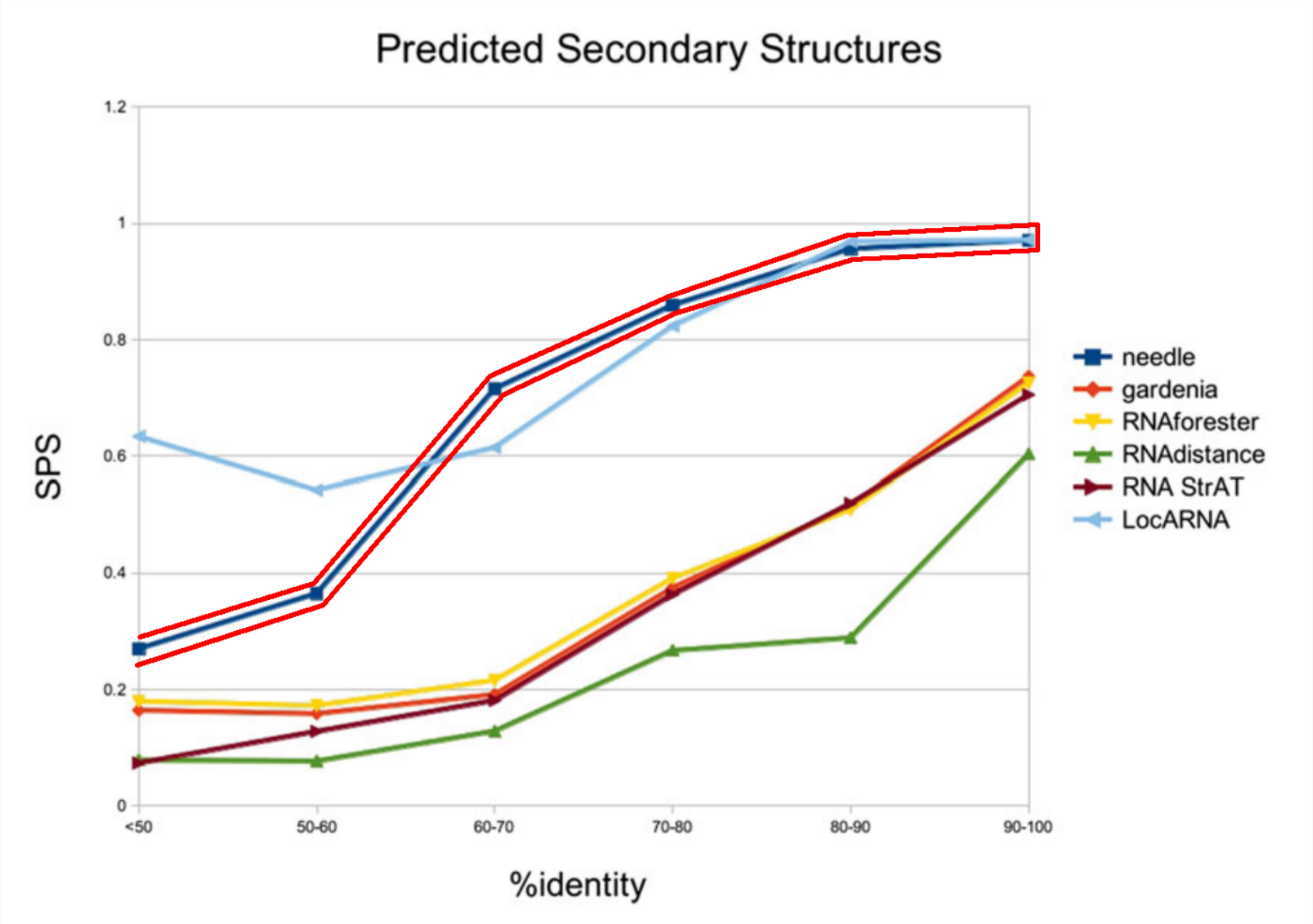
Identity: 85% = 6 / 7

Identical nucleotides / shorter sequence length

Explain: predicted (RNAfold), test setup




Predicted Secondary Structures
SPS
1.2
1
0.8
0.6
0.4
0.2
0
<50
50-60
60-70
70-80
80-90
90-100
%identity
needle
gardenia
RNAforester
RNAdistance
RNA StrAT
LocARNA


Predicted Secondary Structures
SPS
1.2
1
0.8
0.6
0.4
0.2
0
<50
50-60
60-70
70-80
80-90
90-100
%identity
needle
gardenia
RNAforester
RNAdistance
RNA StrAT
LocARNA

Content

needle - Introduction

- Sequence based Algorithm

needle - Introduction

- Sequence based Algorithm
- part of the EMBOSS package

needle - Introduction

- 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

Content

Sequence Alignment

- local sequence alignment

Sequence Alignment

- local sequence alignment
- global sequence alignment

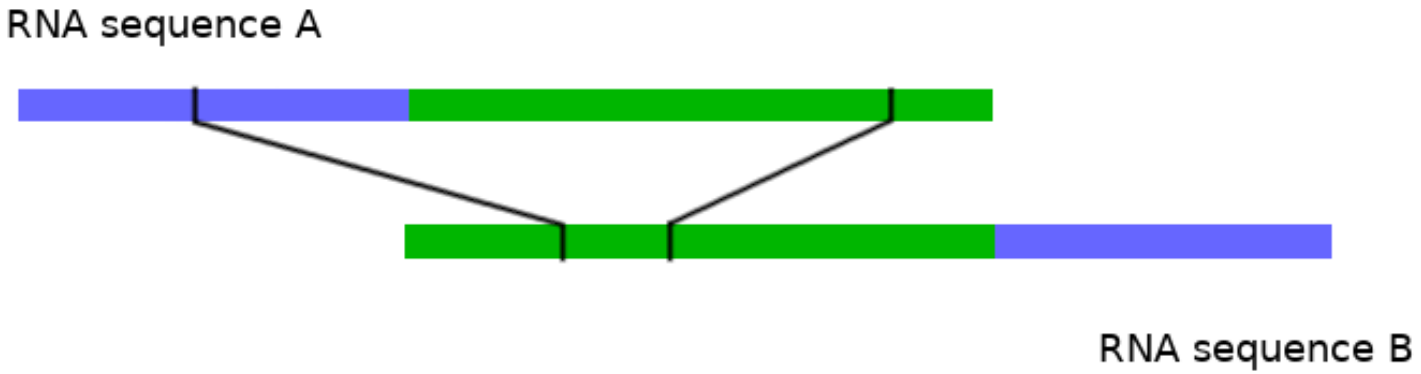
Sequence Alignment

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

Sequence Alignment - local

Sequence Alignment - global

Sequence Alignment - glocal



Content

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							

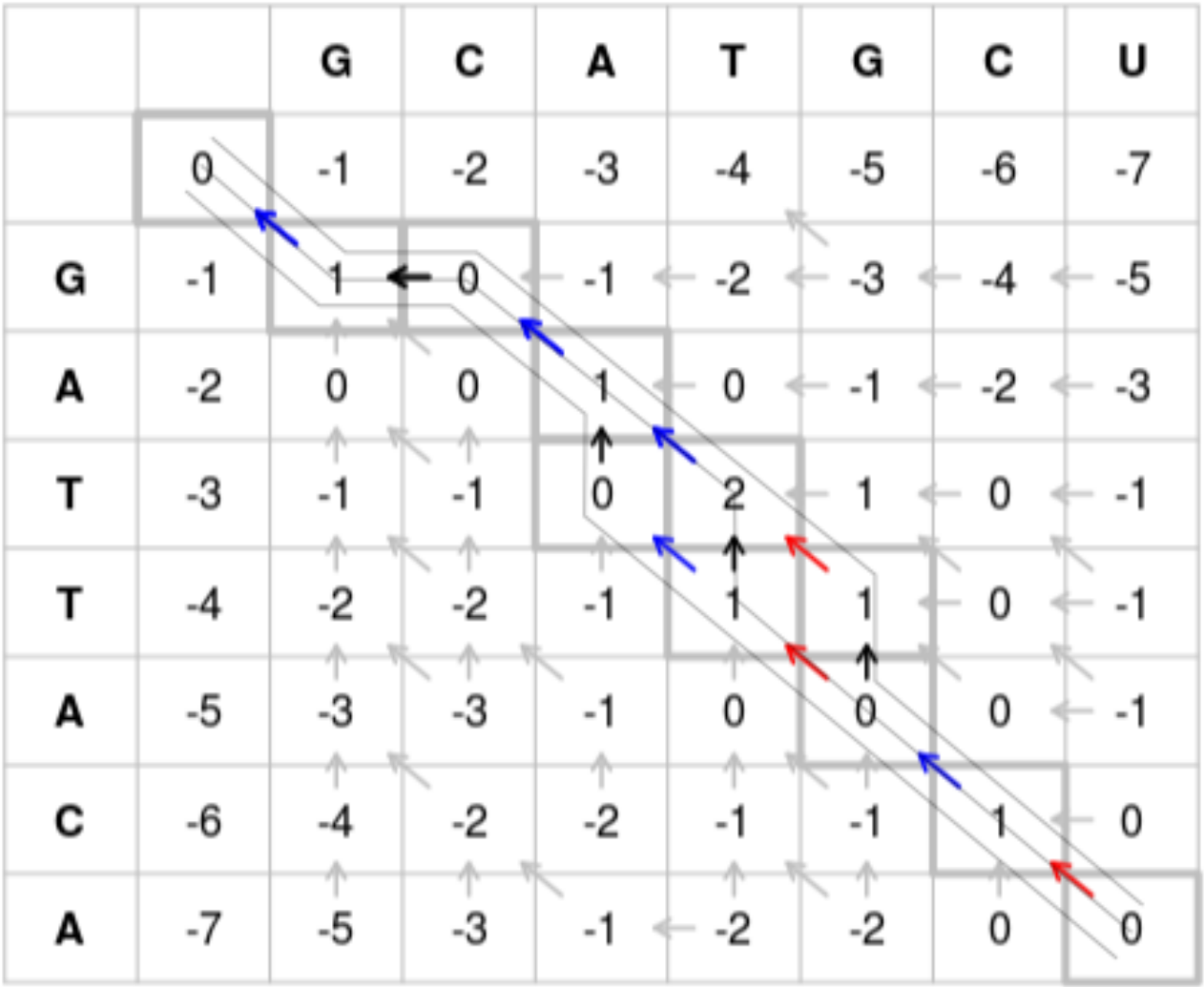
Needleman-Wunsch Algorithm - Example

		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

LocARNA - introduction

Two steps:

LocARNA - introduction

Two steps:

- create base pair probability matrix using RNAfold

LocARNA - introduction

Two steps:

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

LocARNA - introduction

Two steps:

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

(folding and aligning)

Content

Content

Content

Content

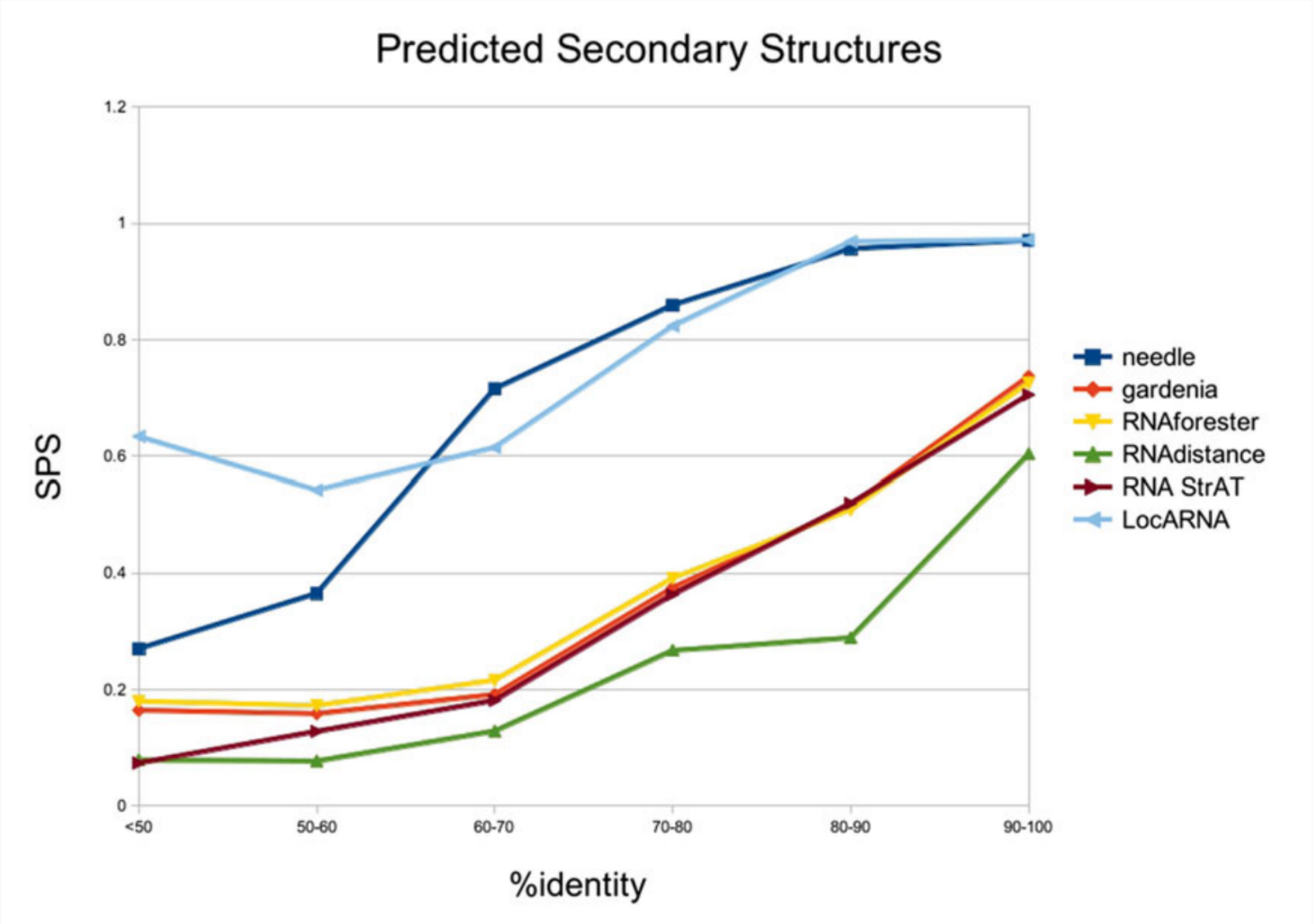
Content

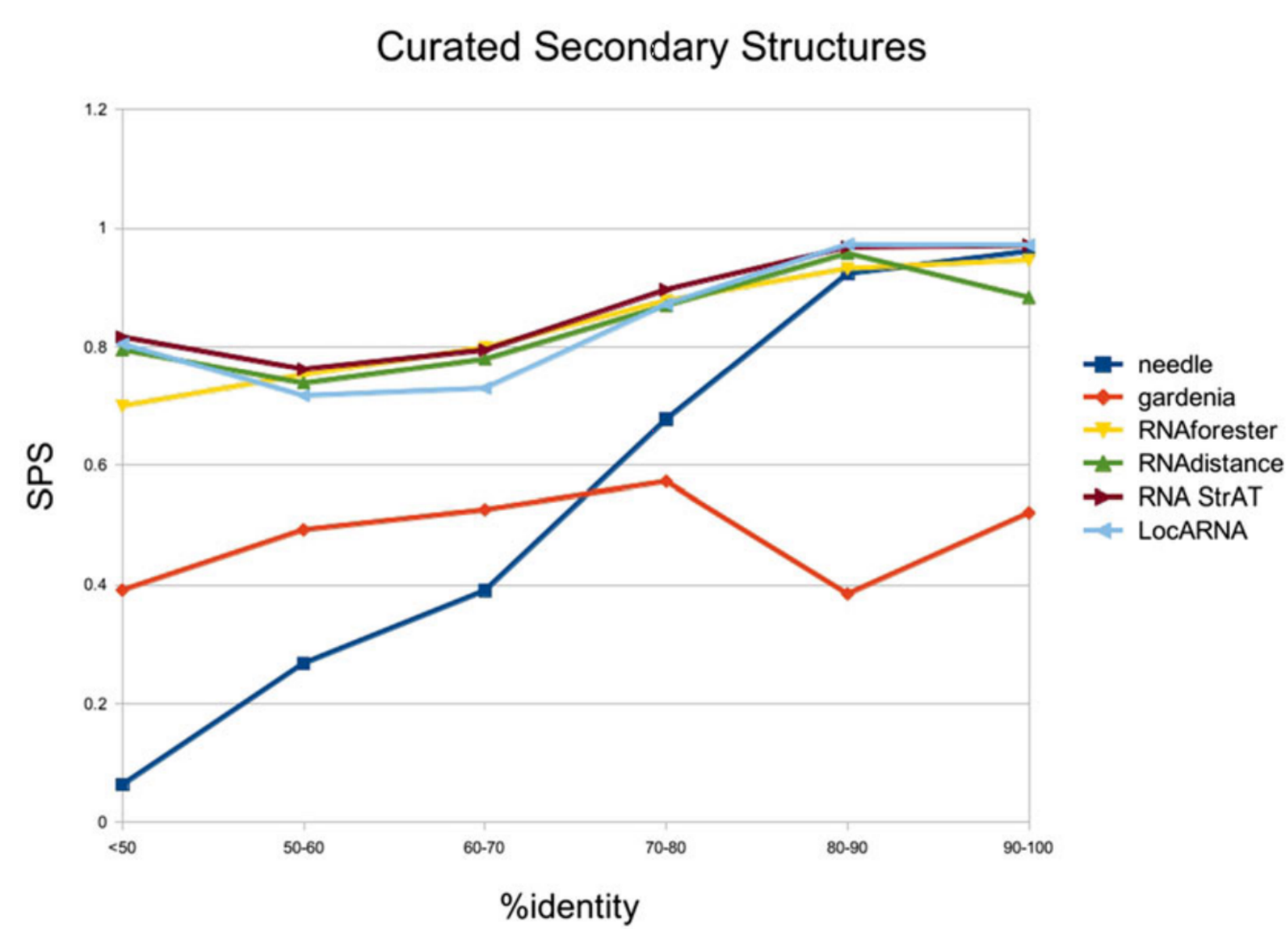
Content

Run time Comparison

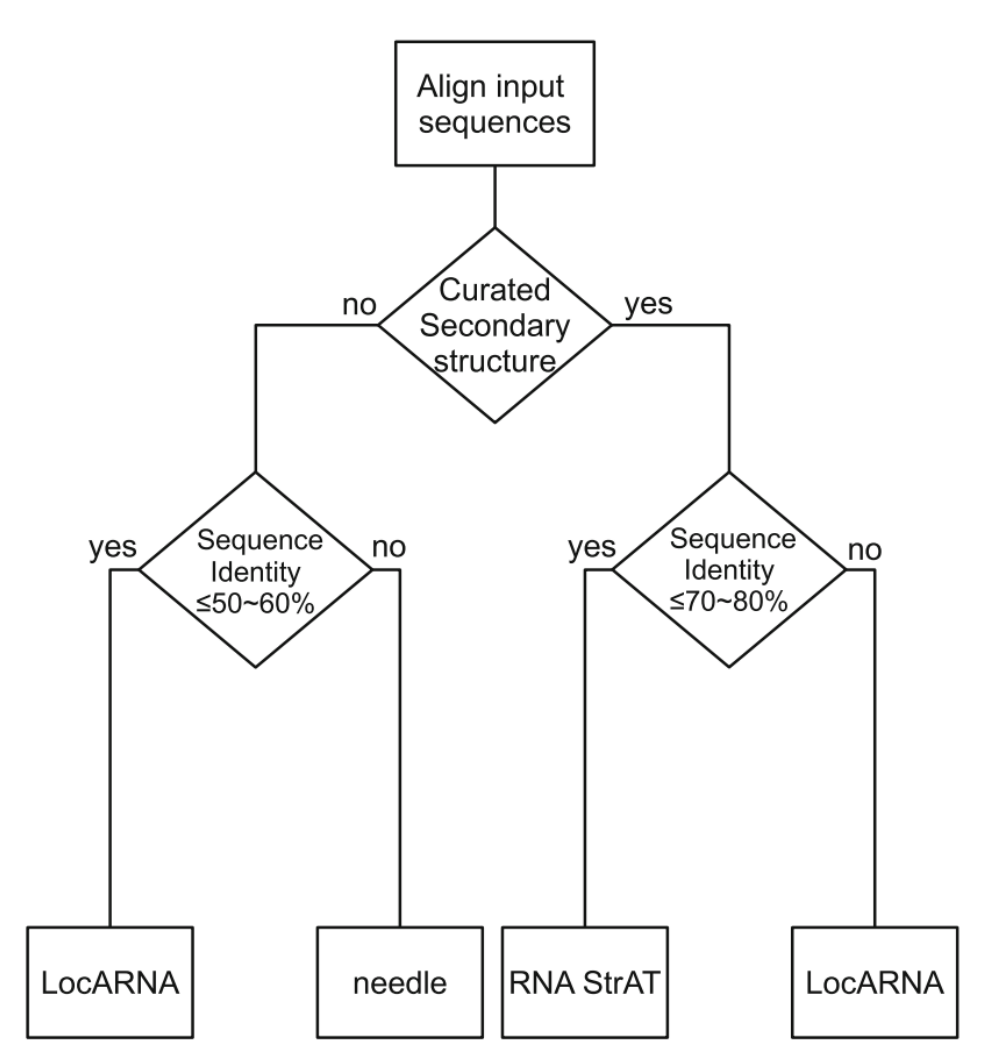
Mean running time (in seconds) per sequence

	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
SPS
1.2
1
0.8
0.6
0.4
0.2
0
<50
50-60
60-70
70-80
80-90
90-100
%identity
needle
gardenia
RNAforester
RNAdistance
RNA StrAT
LocARNA


Curated Secondary Structures
SPS
1.2
1
0.8
0.6
0.4
0.2
0
<50
50-60
60-70
70-80
80-90
90-100
%identity
needle
gardenia
RNAforester
RNAdistance
RNA StrAT
LocARNA

Proposed Workflow



Content

Sources i

The slides can be found at:

Github

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

Sources ii

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