A Simple Protocol for the Inference of RNA Global Pairwise Alignments

Felix Karg

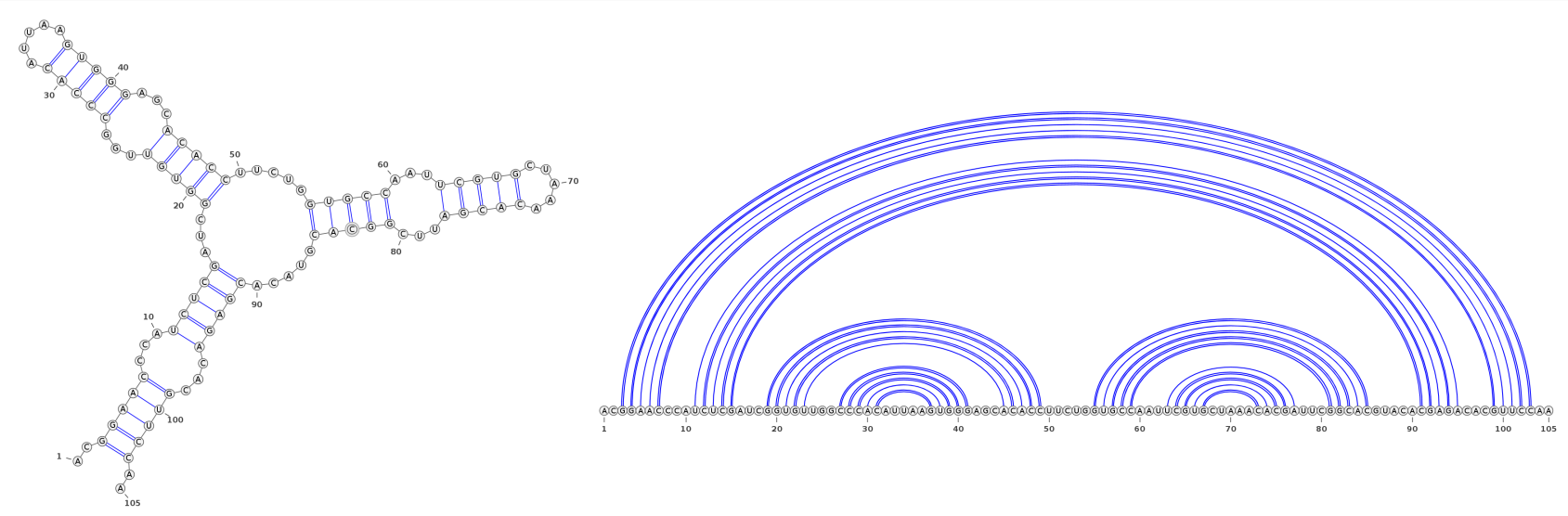
27. August 2019

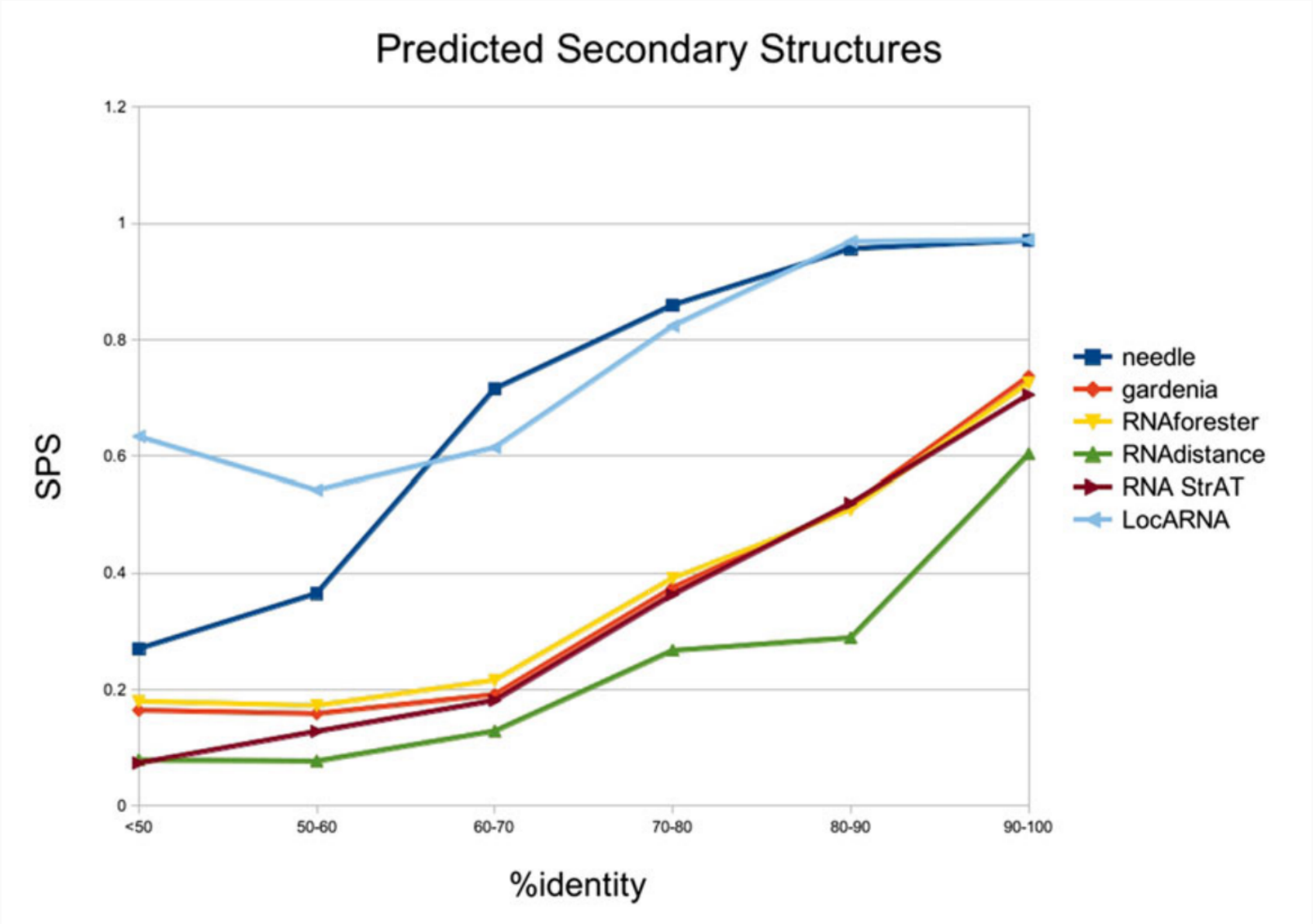
University of Freiburg

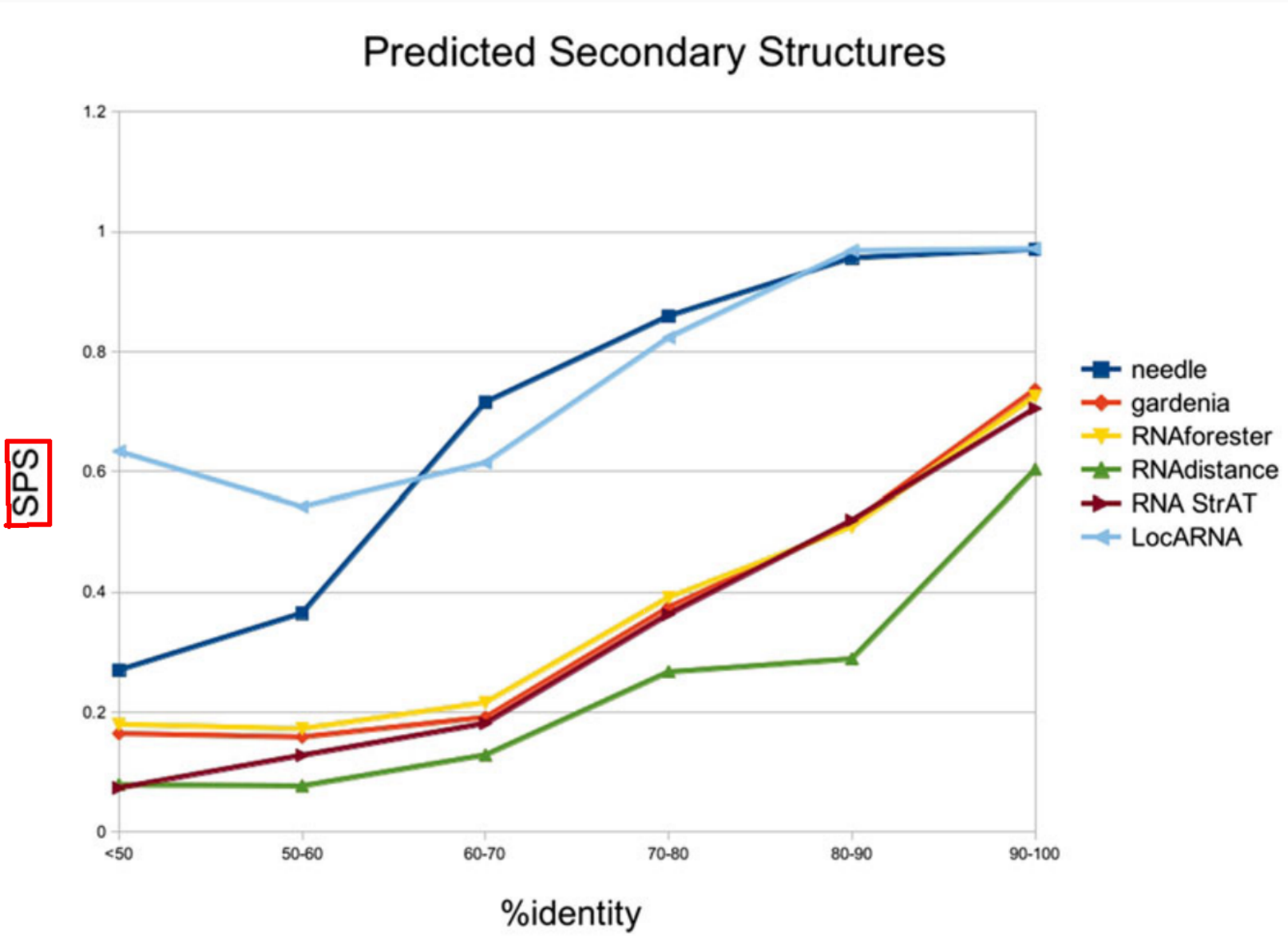
Content

Calm down.

Content


ACGGAACCCAUCUCGAUCGGUGUUGGCCCACAUUAAGUGGGAGCACACCUUCUGGUGCCAAUUCGUGCUAAAACACGAUUCGGCACGUACACGAGACACGUUCCAA
1
10
20
30
40
50
60
70
80
90
100
105


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

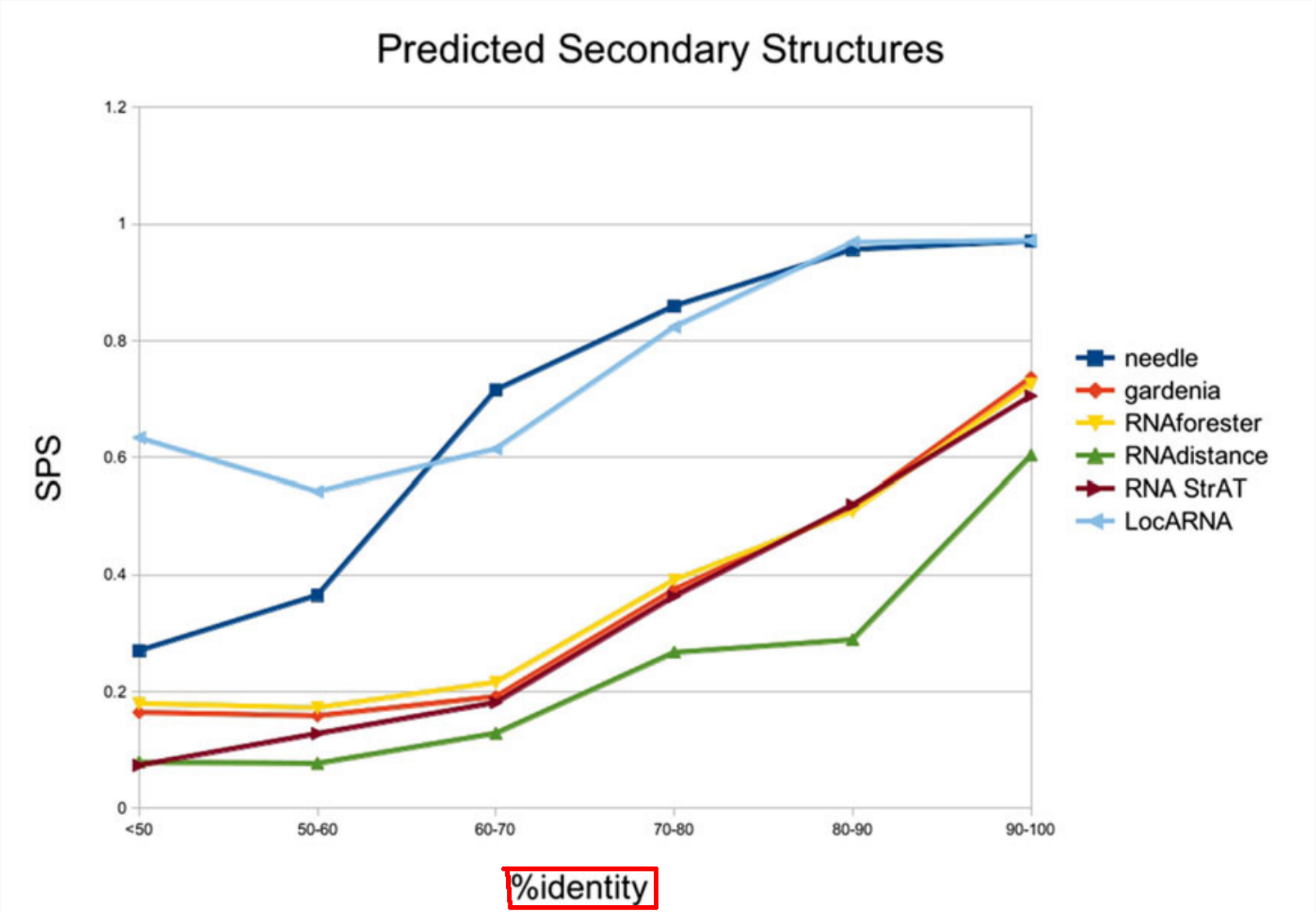
A: AAGGCTT

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

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

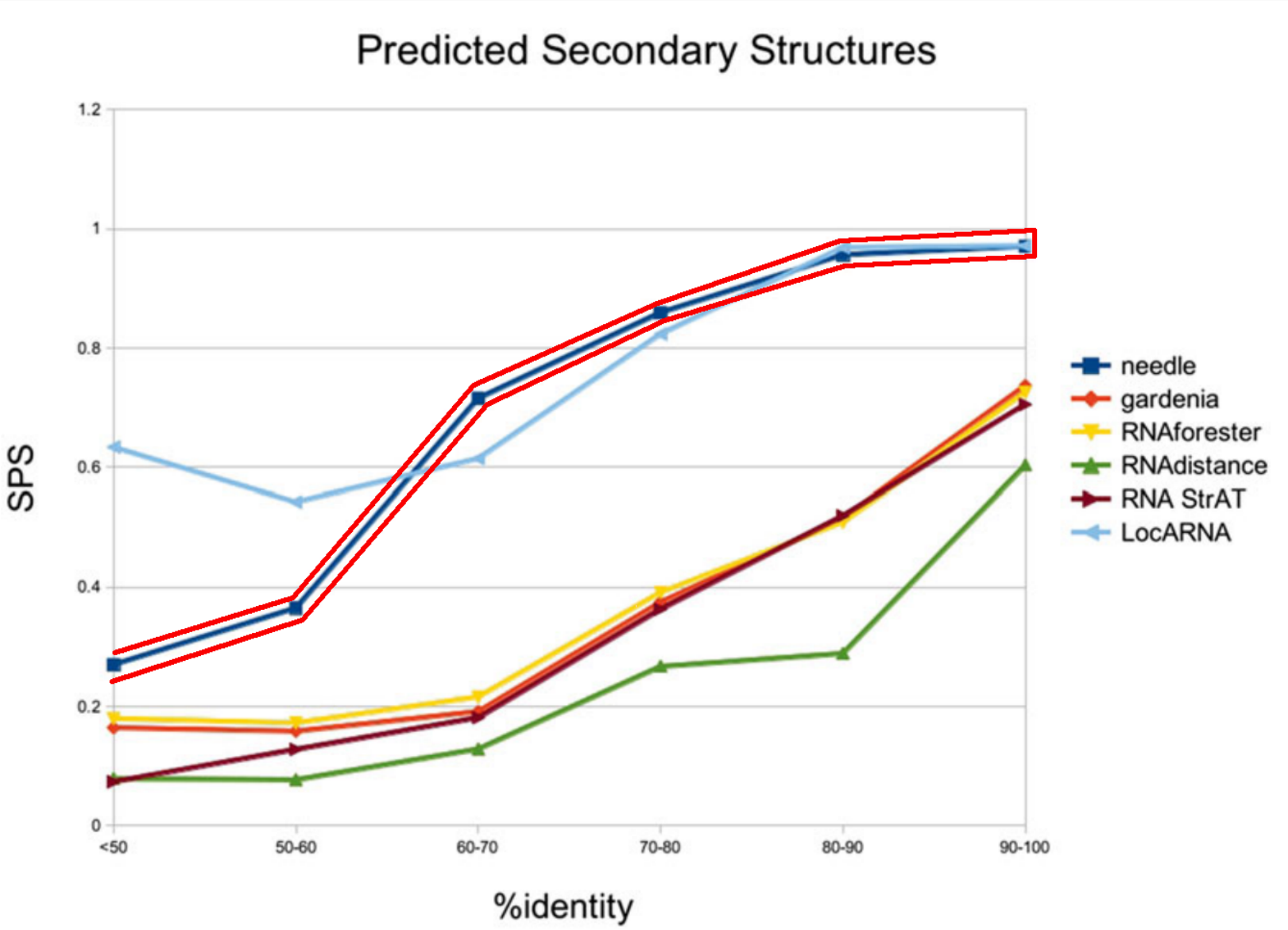
B: AAGGC

C: AAGGCAT

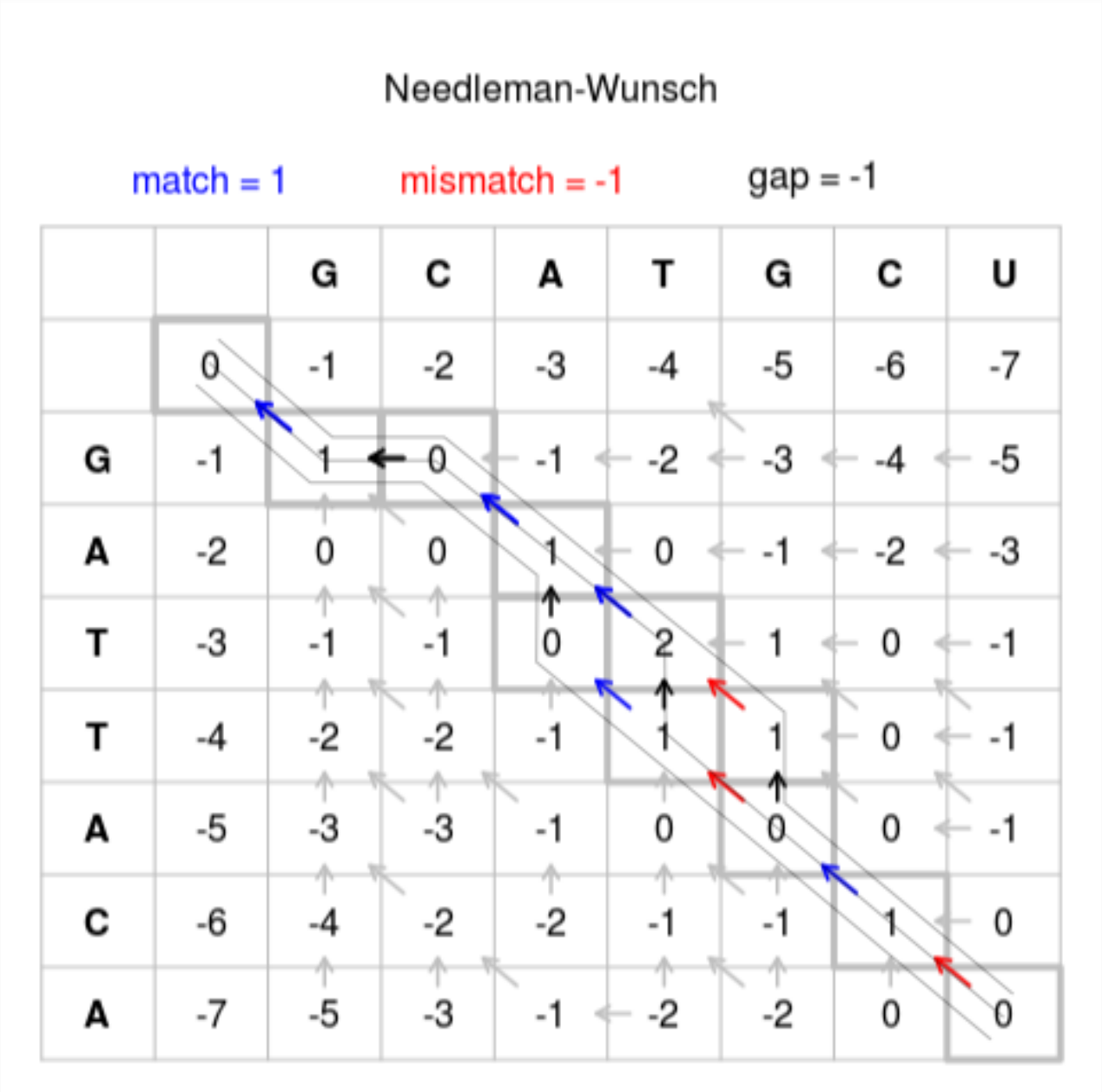
Identity: 85% = 6 / 7

Identical nucleotides / shorter sequence length

needle


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

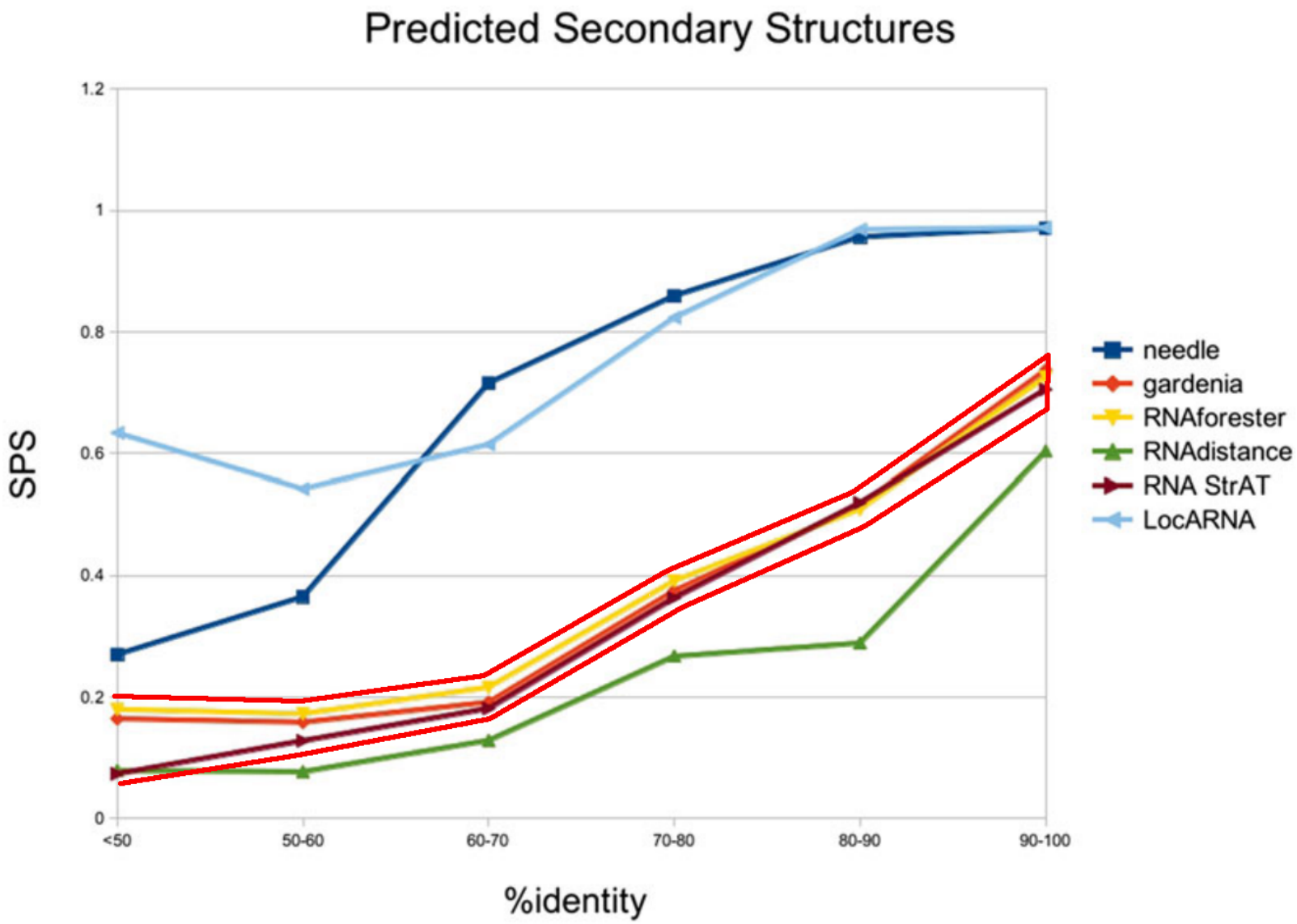
Needleman-Wunsch-Algorithm



		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

Content

Tree-based

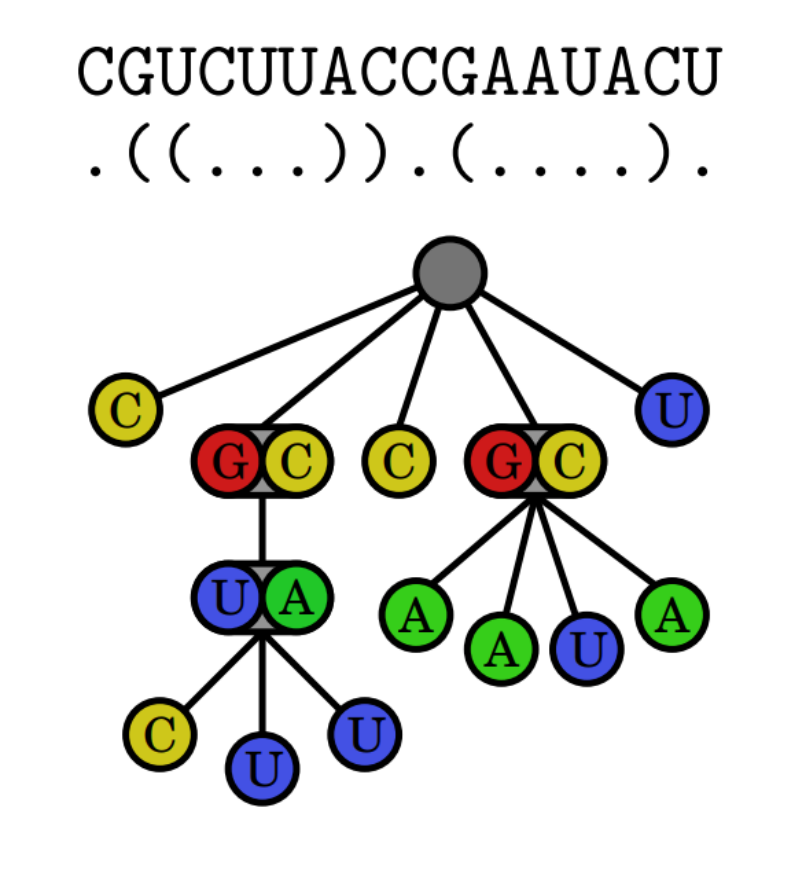

Predicted Secondary Structures
SPS
1.2
1
0.8
0.6
0.4
0.2
0
<50
50-60
60-70
70-80
80-90
90-100
%identity
needle
gardenia
RNAforester
RNAdistance
RNA StrAT
LocARNA

Tree-based

Using the secondary structure:

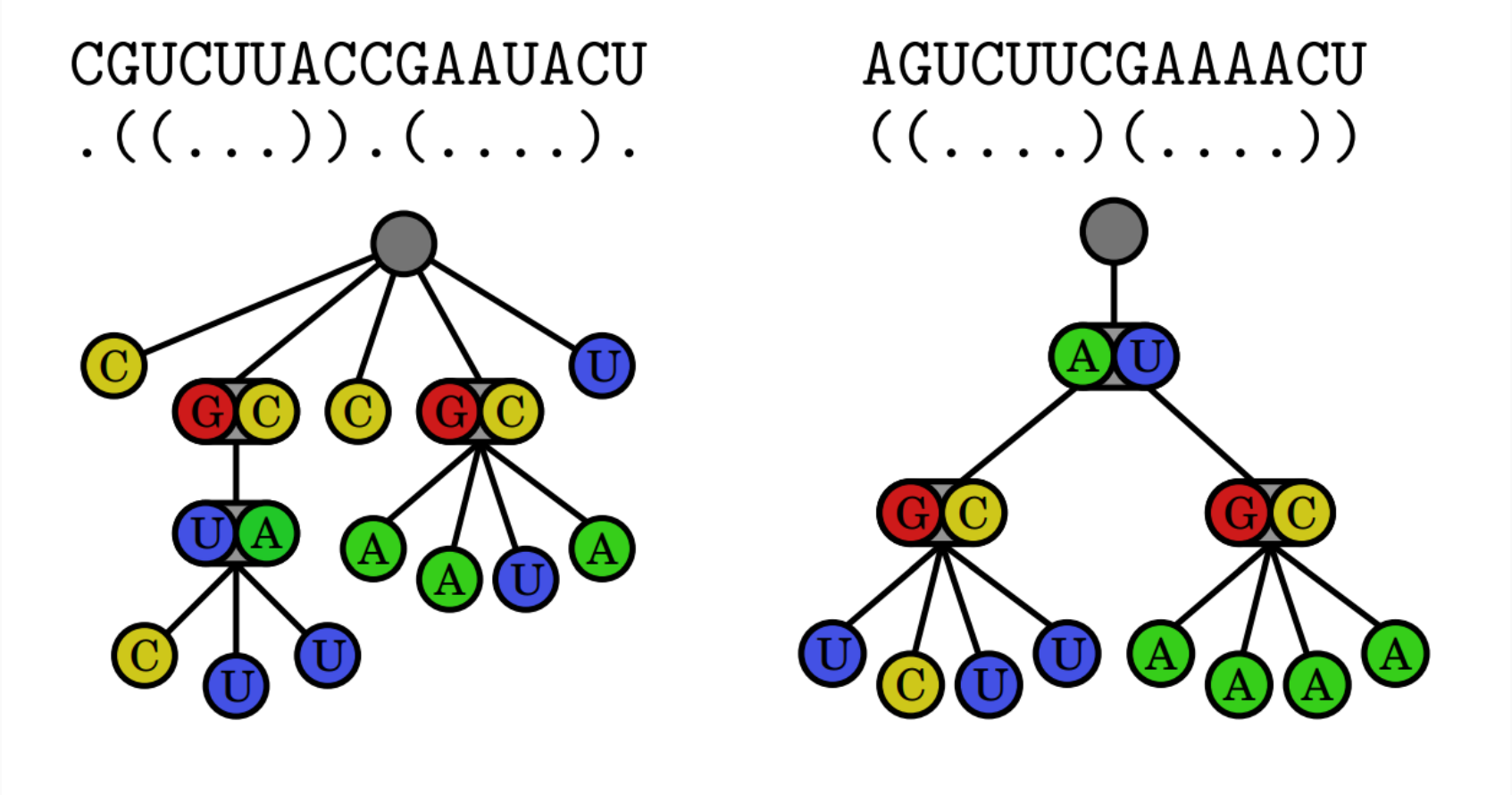
Tree-based

Using the secondary structure:

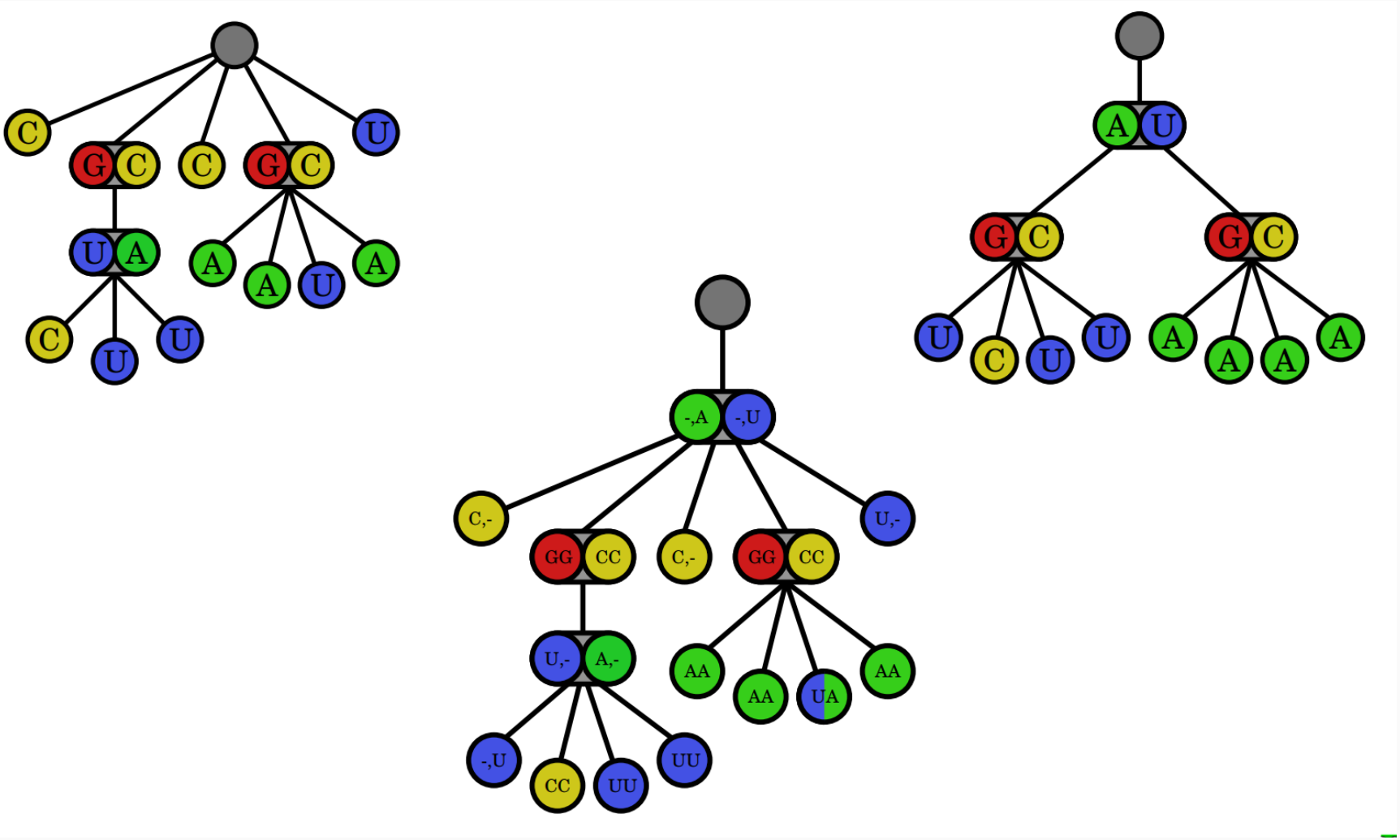


Tree-based

Using the secondary structure:



Tree-Alignment



Tree-Alignment Problems

Tree-Alignment Problems

- This is not always possible.

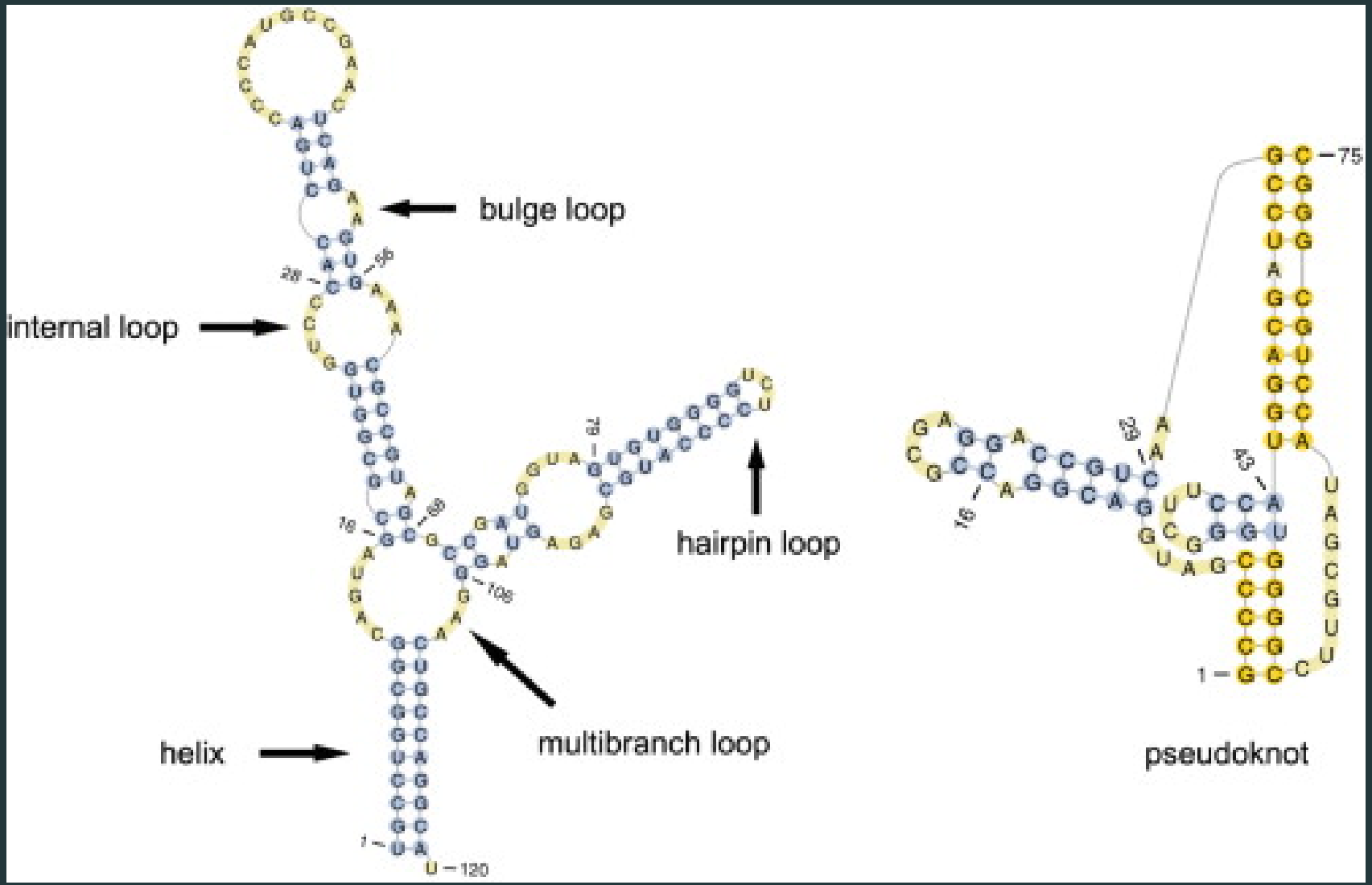
Tree-Alignment Problems

- This is not always possible.
- There's structures that cannot be represented in trees.

Tree-Alignment Problems

- This is not always possible.
- There's structures that cannot be represented in trees.

Example: `( . [ . ) . ]`


bulge loop
internal loop
hairpin loop
helix
multibranch loop
pseudoknot

Tree-Editing

Edit operations on Trees ...

Tree-Editing

Edit operations on Trees ...

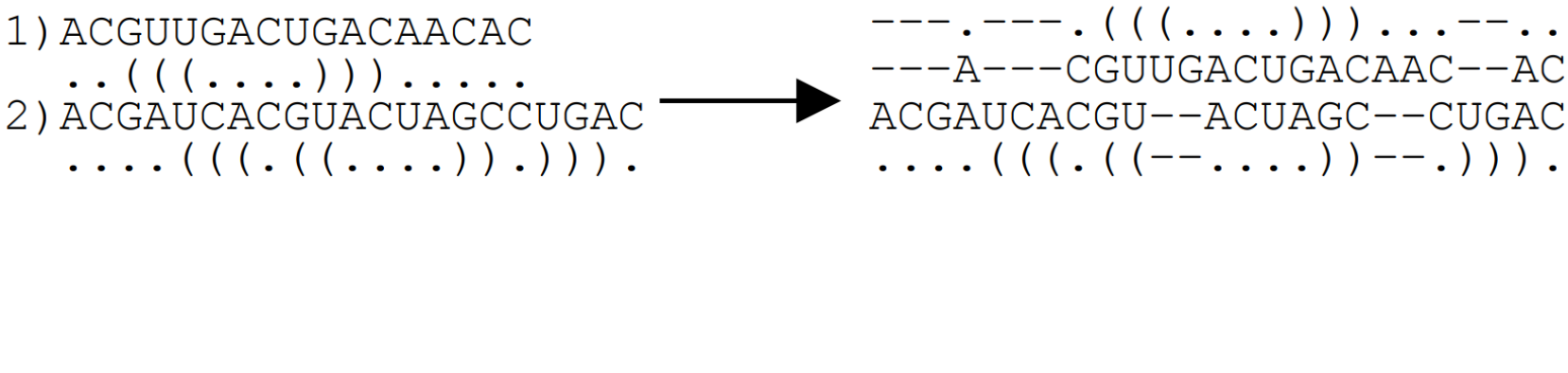
... are edit operations on arc-annotated sequences!

Tree-Editing

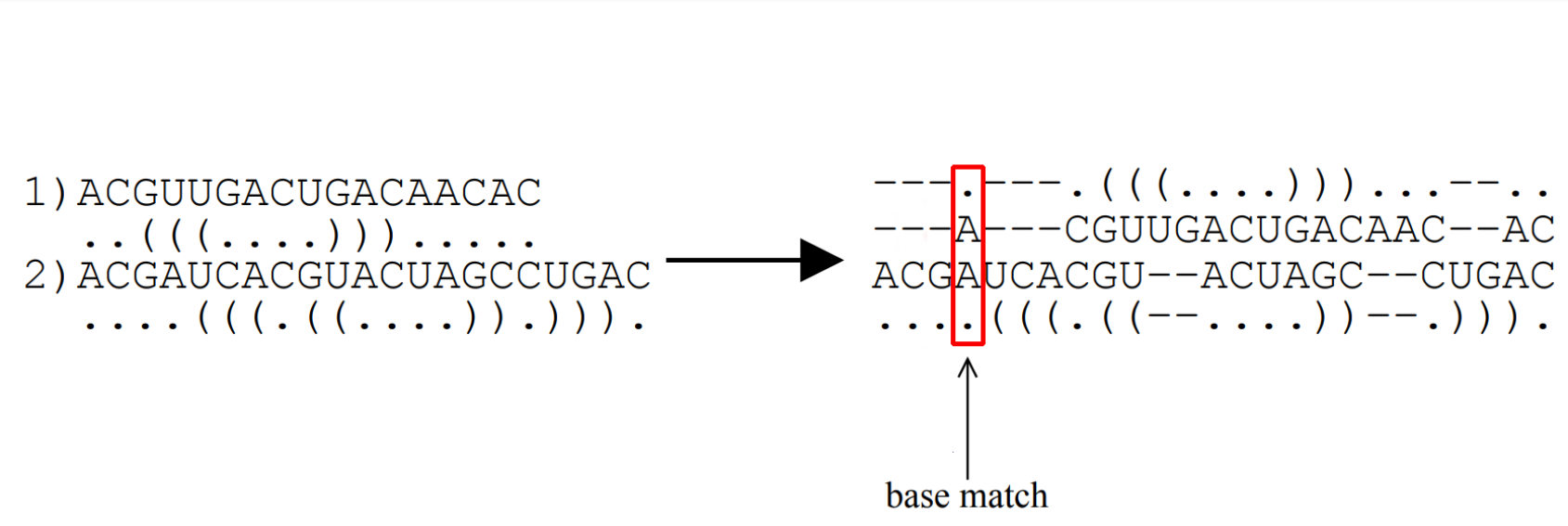
Edit operations on Trees ...

... are edit operations on arc-annotated sequences! (.((..(..))..)))

Tree-Editing Possibilities

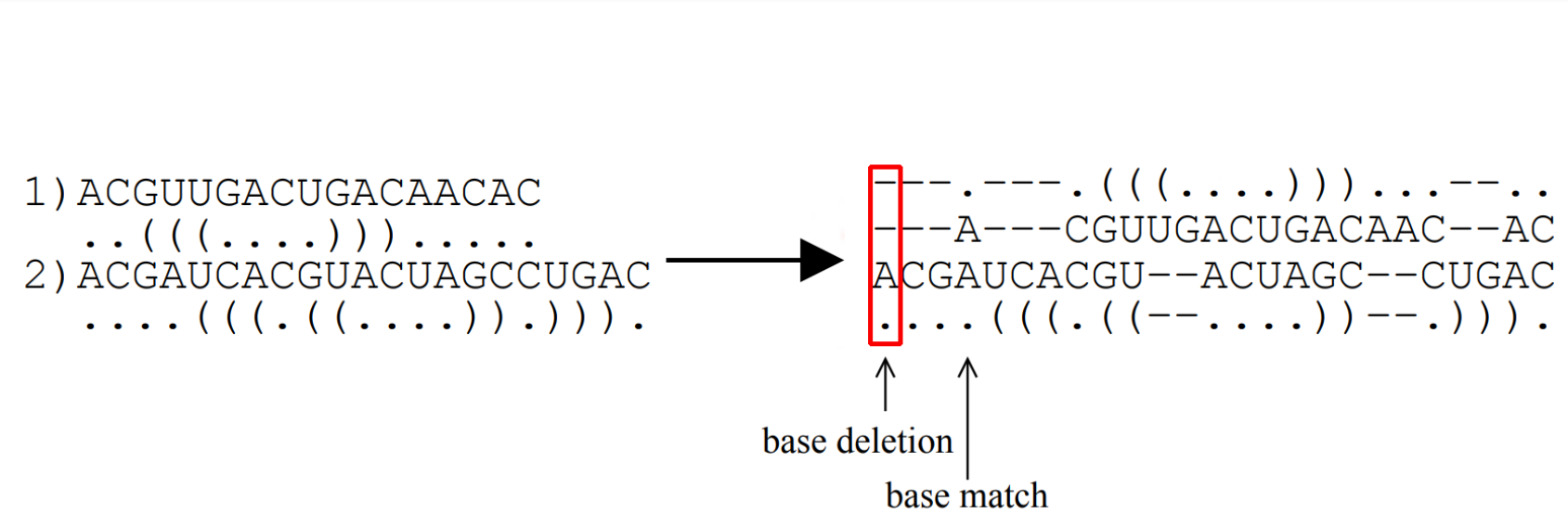


Tree-Editing Possibilities



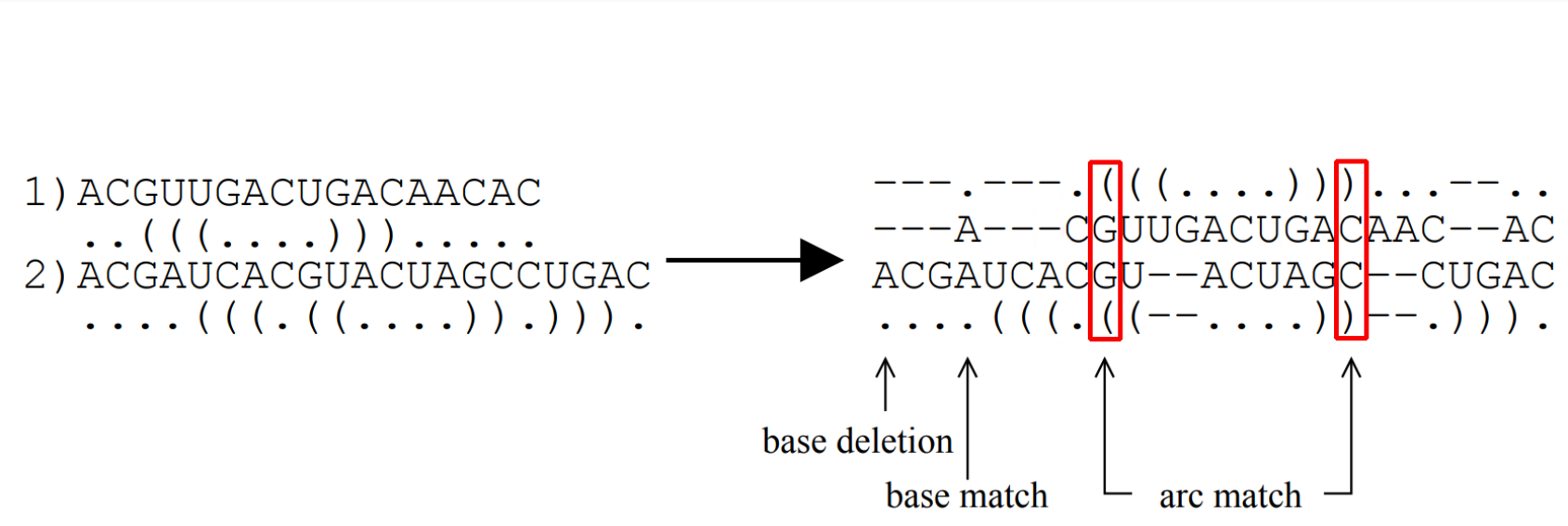
- base match

Tree-Editing Possibilities



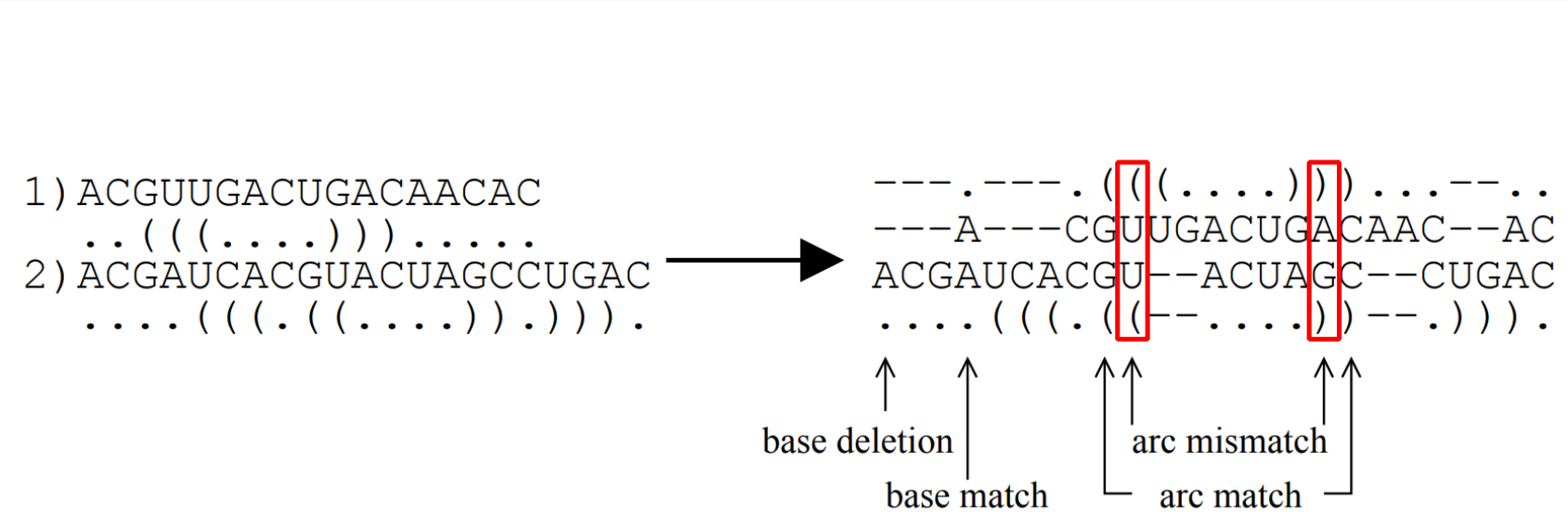
- base match
- base indel

Tree-Editing Possibilities



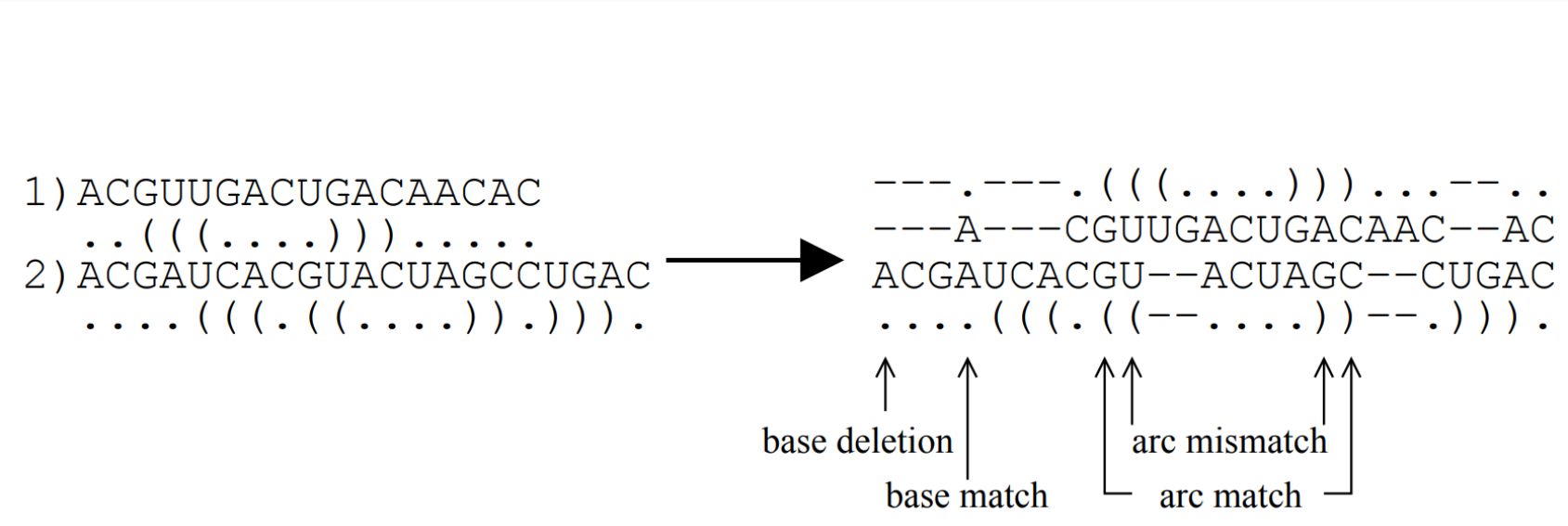
- base match
- base indel
- arc match

Tree-Editing Possibilities



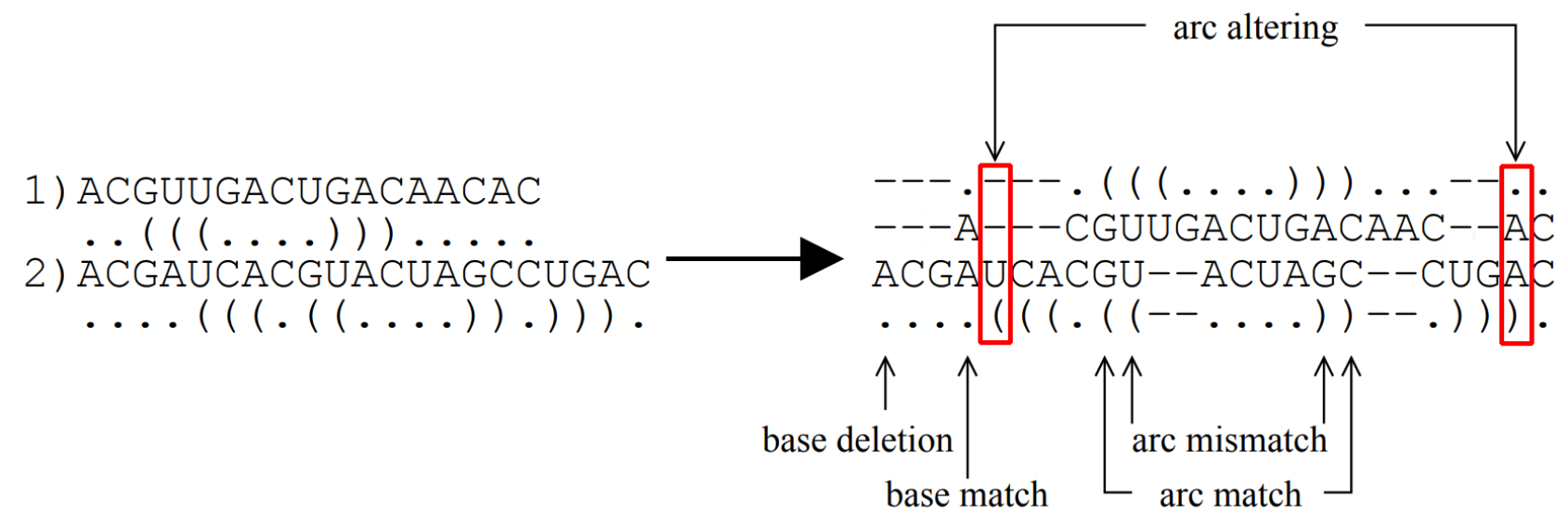
- base match
- base indel
- arc match
- arc mismatch

Tree-Editing Possibilities



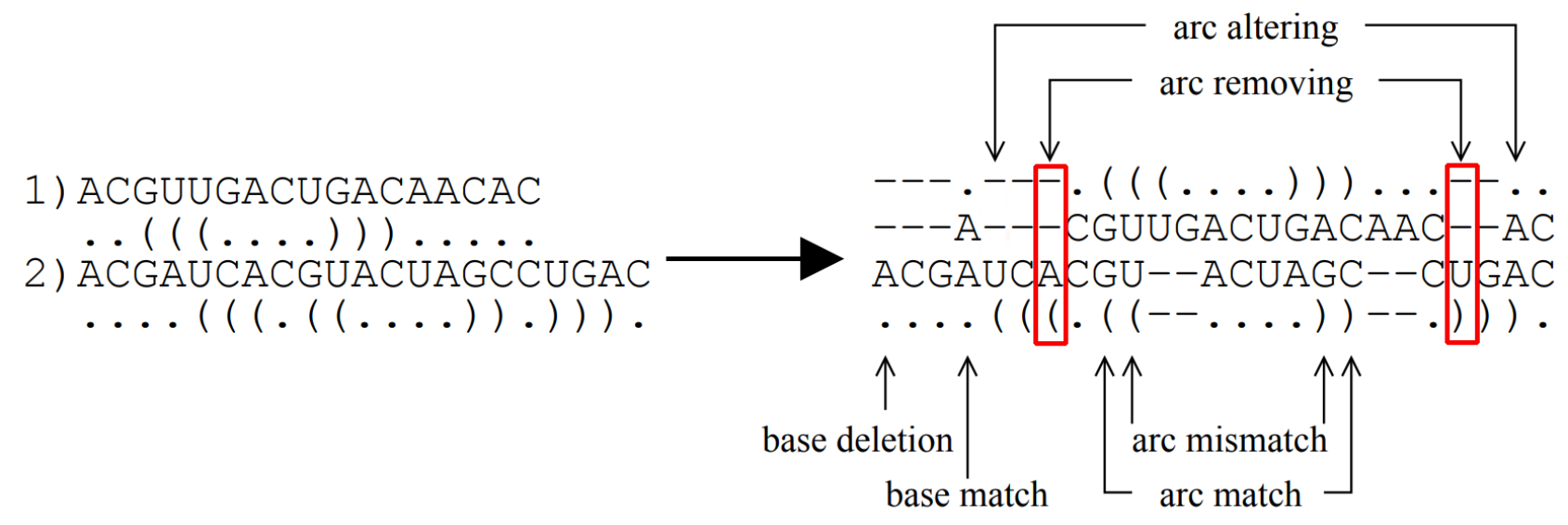
- base match
- base indel
- arc match
- arc mismatch
- arc breaking (missing)

Tree-Editing Possibilities



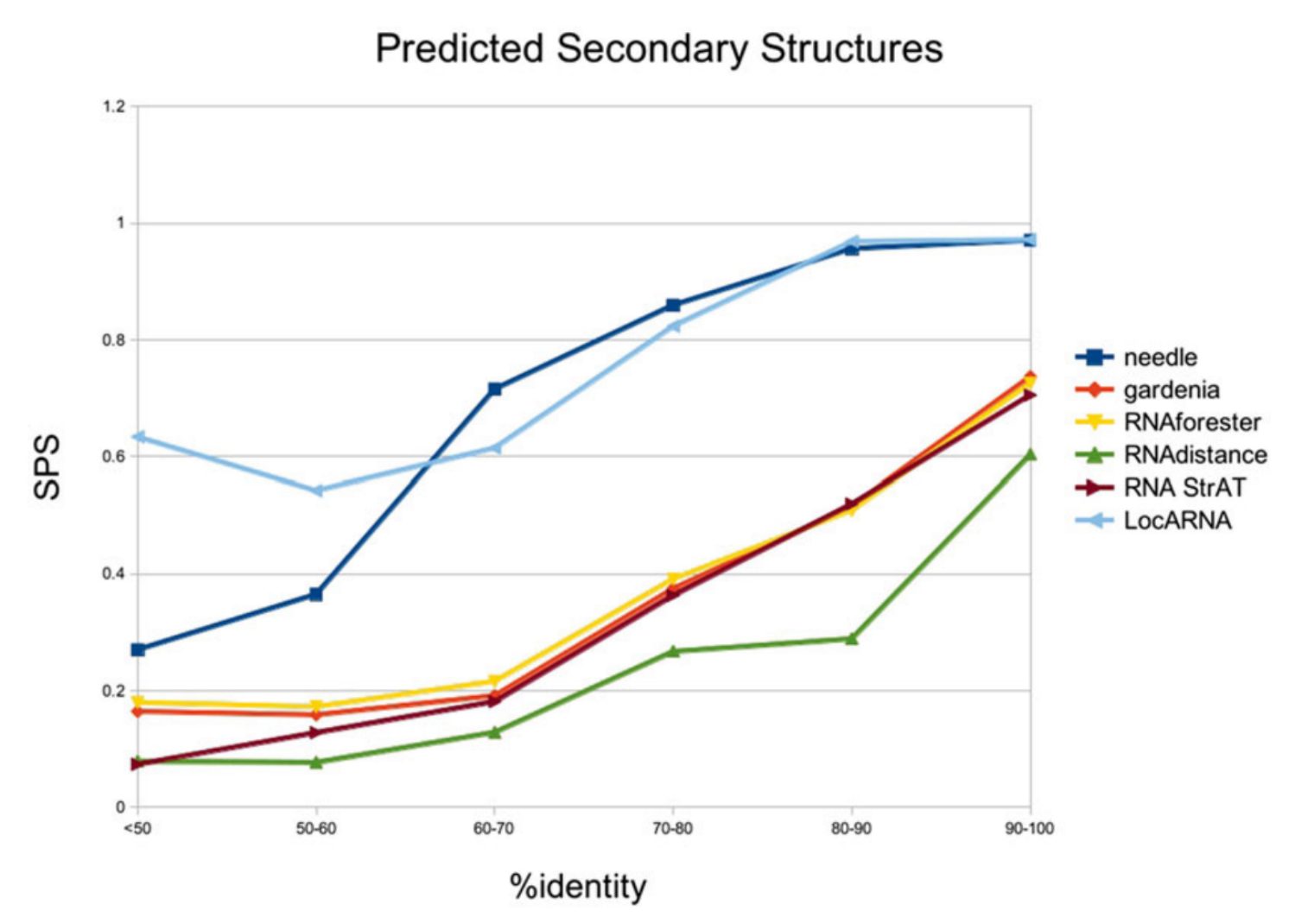
- base match
- base indel
- arc match
- arc mismatch
- arc breaking (missing)
- arc altering

Tree-Editing Possibilities

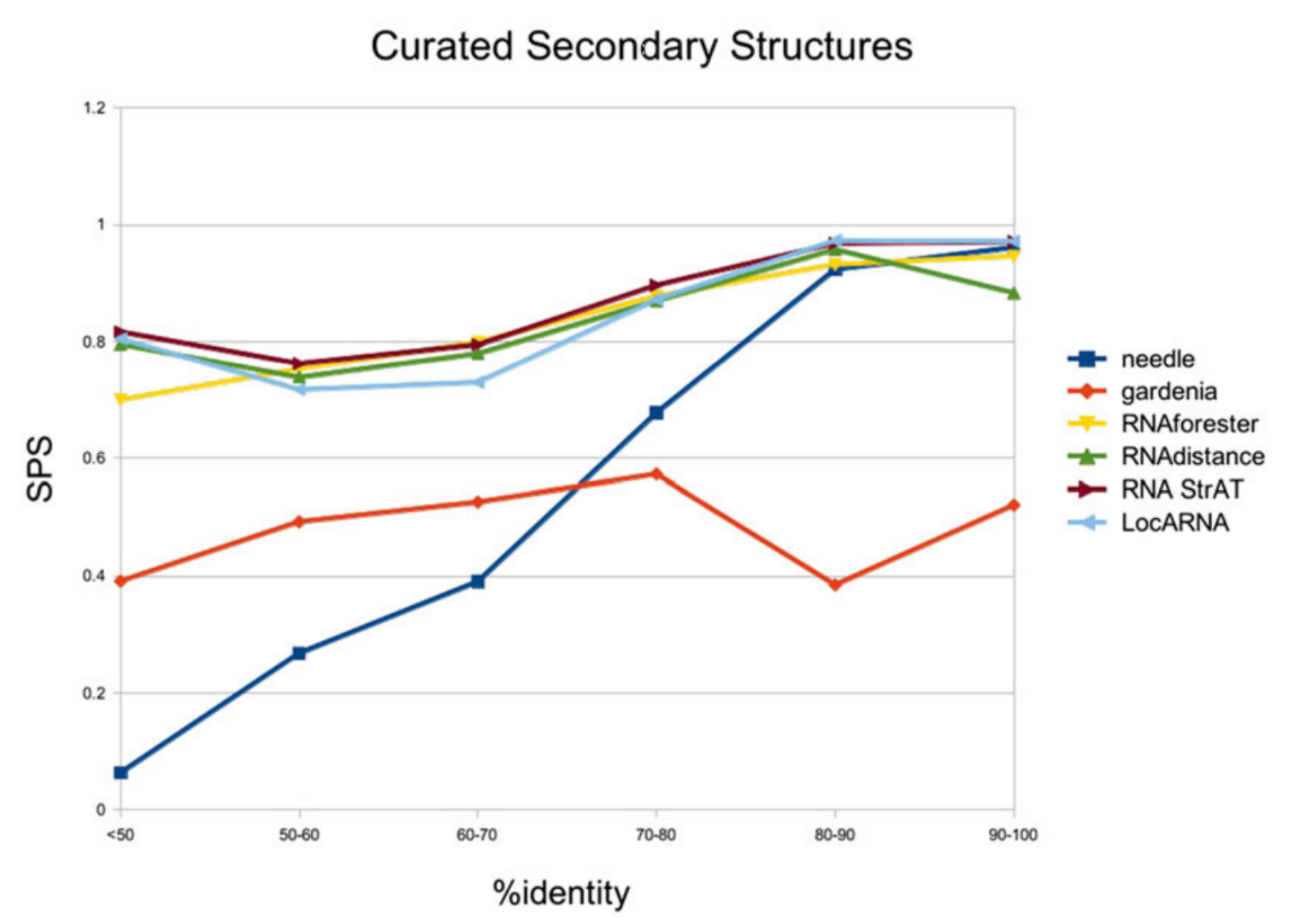


- base match
- base indel
- arc match
- arc mismatch
- arc breaking (missing)
- arc altering
- arc removing

Performance comparison



Performance comparison



Tree-based Alignment

Tree-based Alignment

Basically a structure-using edit distance.

Content

Sankoff-Algorithm

Sankoff-Algorithm

- Dynamic Programming

Sankoff-Algorithm

- Dynamic Programming
- **Space** needed: $O(n^4)$

Sankoff-Algorithm

- Dynamic Programming
- **Space** needed: $O(n^4)$
- **Runtime** $O(n^6)$

Sankoff-Algorithm

- Dynamic Programming
- **Space** needed: $O(n^4)$
- **Runtime** $O(n^6)$

What does it do ... ?

Sankoff-Algorithm

Global free energy minimisation.

Sankoff-Algorithm

Global free energy minimisation.

Considering every possible combination.

Content

LocARNA - basic information

LocARNA - basic information

- Extensive Caching (Dynamic Programming) + Pruning

LocARNA - basic information

- Extensive Caching (Dynamic Programming) + Pruning
- **Runtime** $O(n^4)$

LocARNA - basic information

- Extensive Caching (Dynamic Programming) + Pruning
- **Runtime** $O(n^4)$
- **Space** $O(n^4)$

LocARNA - basic information

- Extensive Caching (Dynamic Programming) + Pruning
- **Runtime** $O(n^4)$
- **Space** $O(n^4)$
- Web Interface

LocARNA - basic information

- Extensive Caching (Dynamic Programming) + Pruning
- **Runtime** $O(n^4)$
- **Space** $O(n^4)$
- Web Interface
- Developed at the University of Freiburg

LocARNA - Procedure

LocARNA - Procedure

- Where can arcs happen? And how likely are they?

LocARNA - Procedure

- Where can arcs happen? And how likely are they?
- Building own likely secondary structures

LocARNA - Procedure

- Where can arcs happen? And how likely are they?
- Building own likely secondary structures
- Aligning between secondary structures

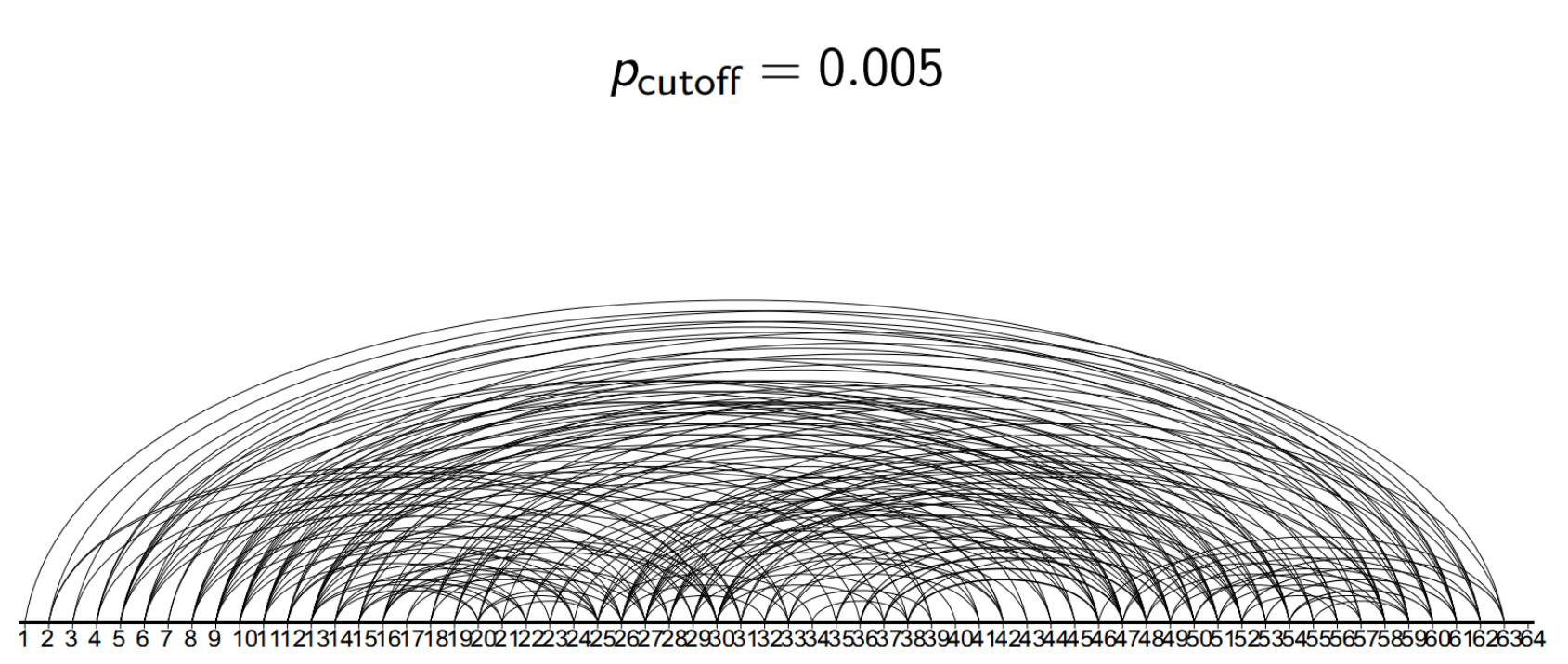
LocARNA - Procedure

- Where can arcs happen? And how likely are they?
- Building own likely secondary structures
- Aligning between secondary structures
- Using **one** relative path Matrix

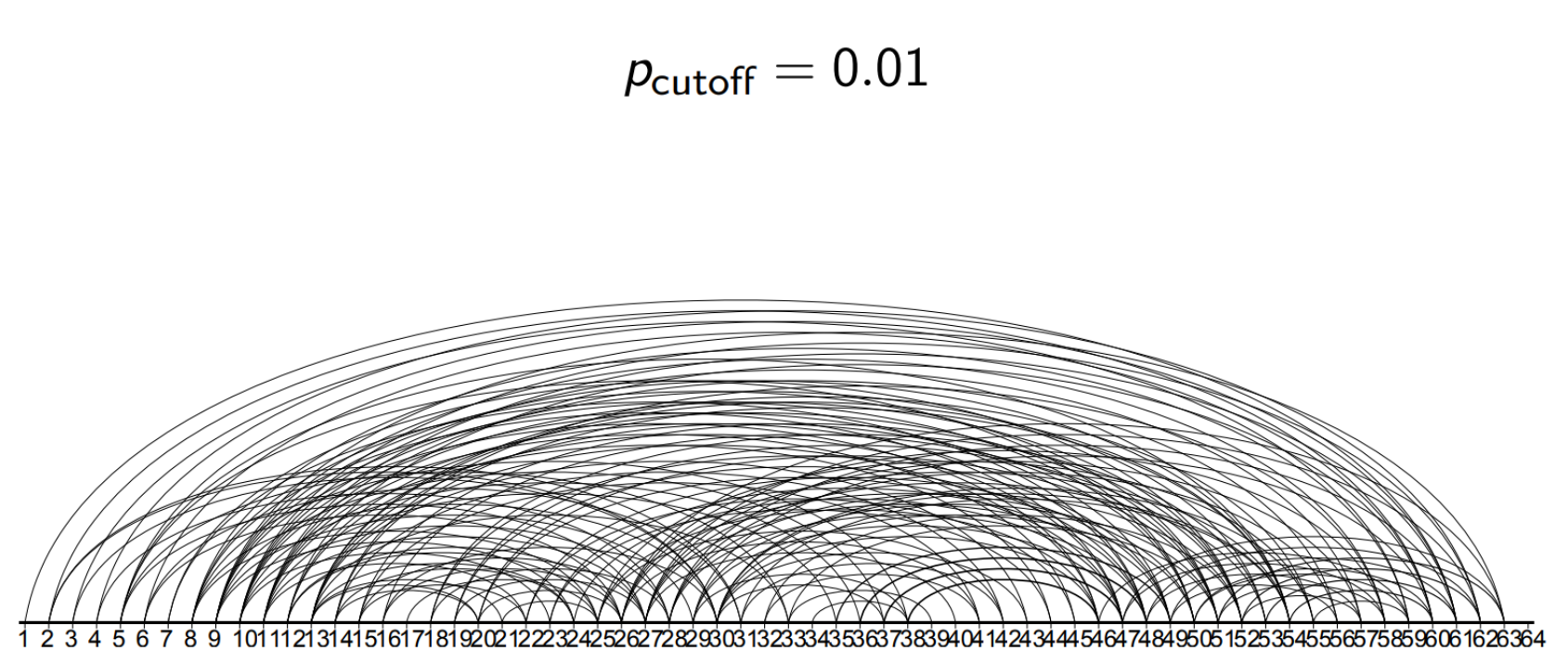
LocARNA - Procedure

- Where can arcs happen? And how likely are they?
- Building own likely secondary structures
- Aligning between secondary structures
- Using **one** relative path Matrix
- Last: 'looking up' optimal alignment

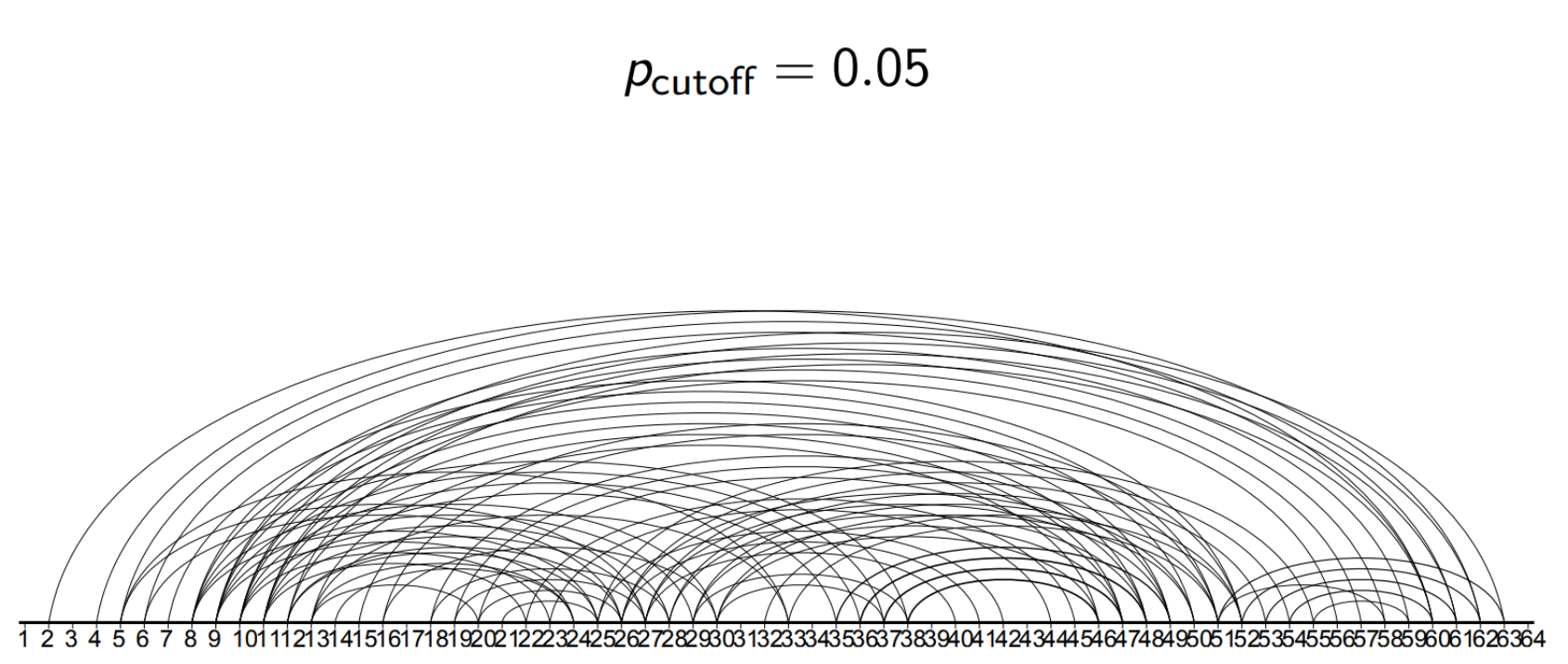
LocARNA - Where can arcs happen? And how likely are they?



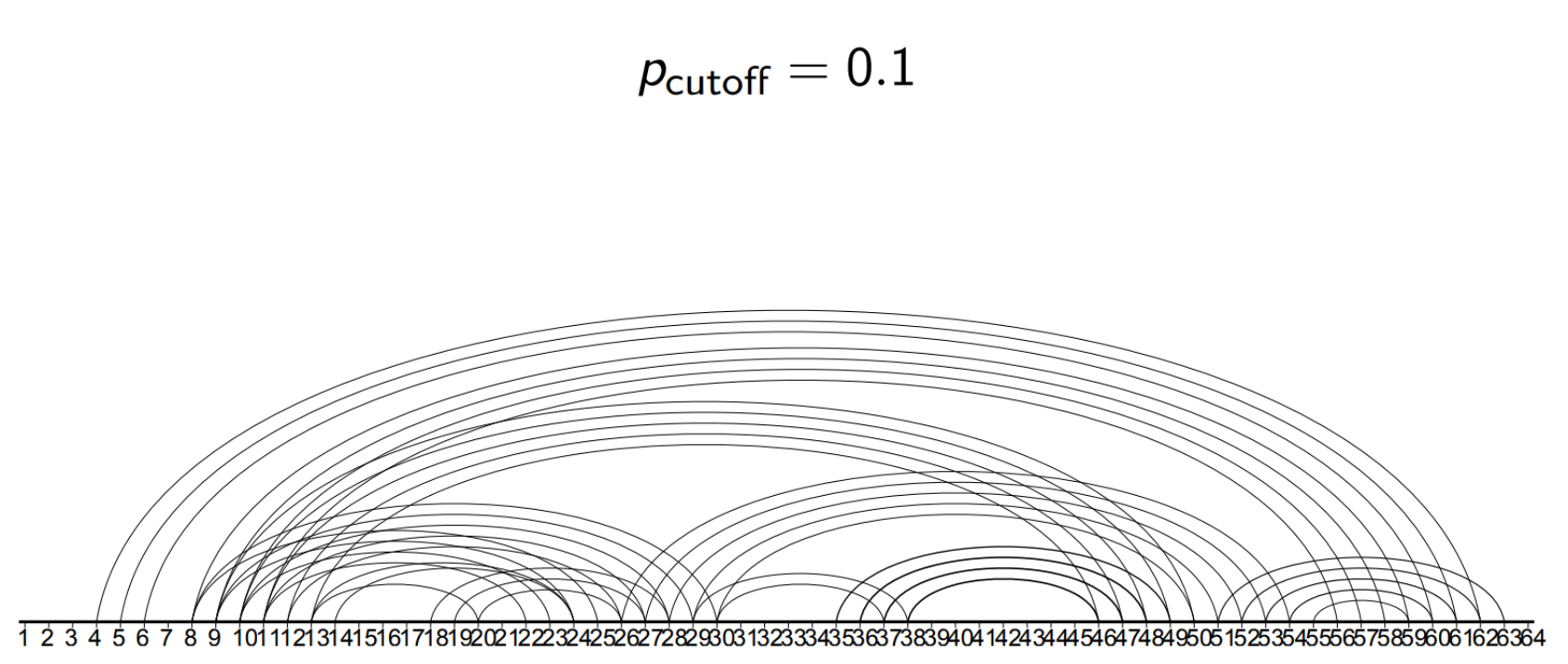
LocARNA - Where can arcs happen? And how likely are they?



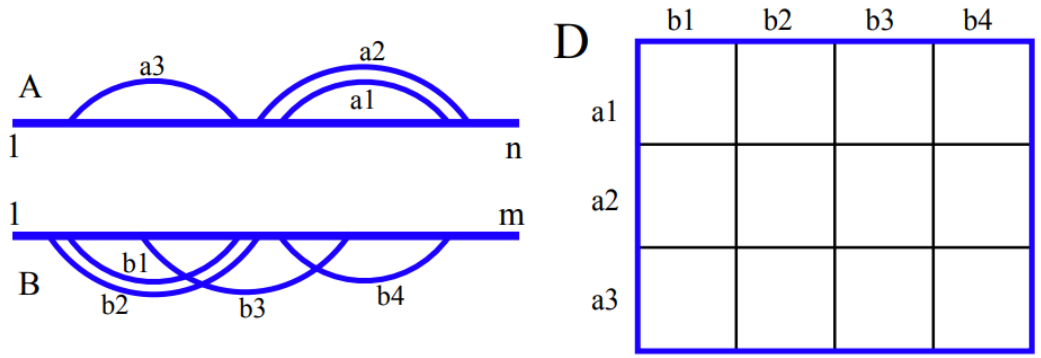
LocARNA - Where can arcs happen? And how likely are they?



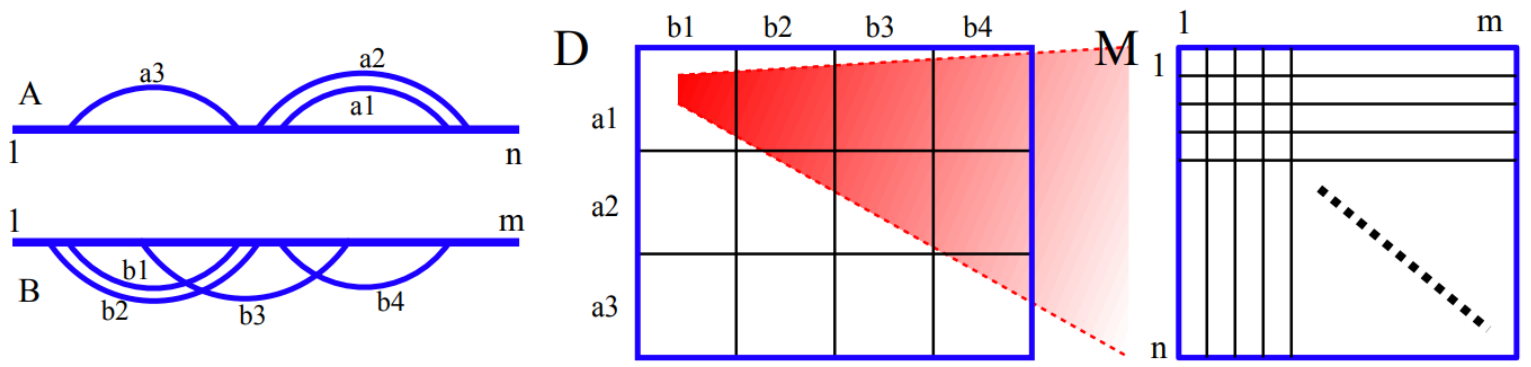
LocARNA - Where can arcs happen? And how likely are they?



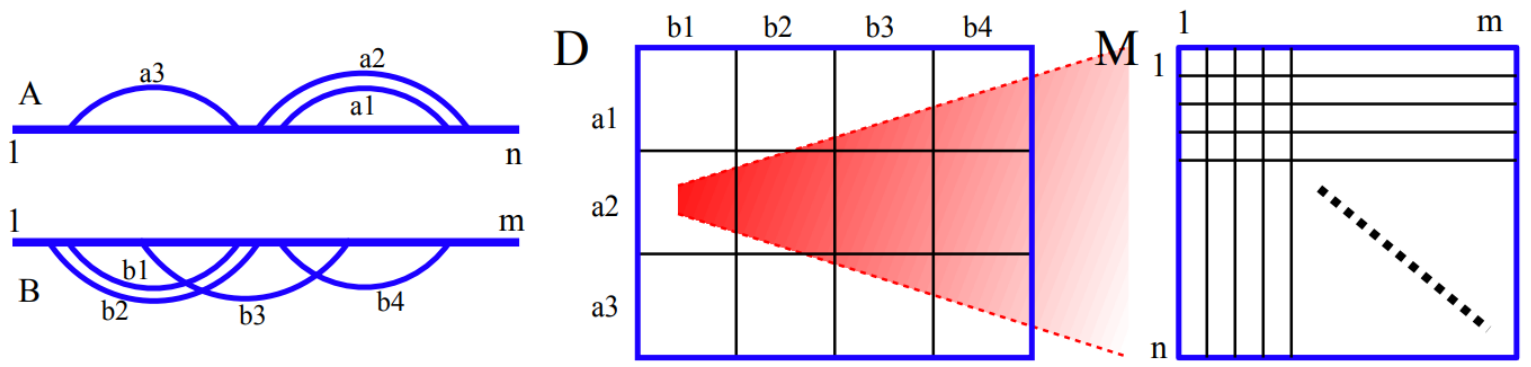
LocARNA - Matrices



LocARNA - Matrices



LocARNA - Matrices



LocARNA - Aligning between secondary structures

Cases:

LocARNA - Aligning between secondary structures

Cases:

- Base match

LocARNA - Aligning between secondary structures

Cases:

- Base match
- Base insertion

LocARNA - Aligning between secondary structures

Cases:

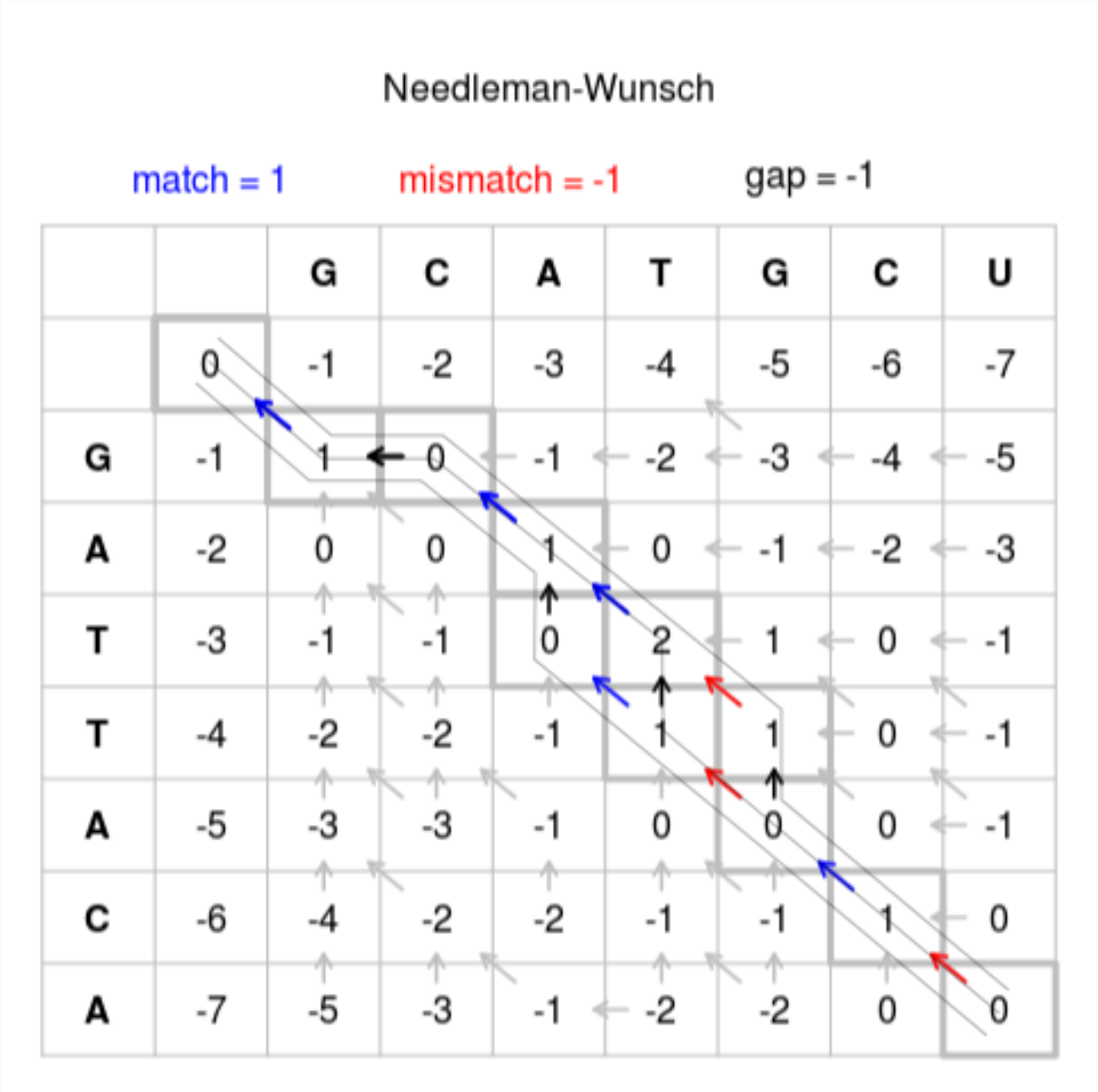
- Base match
- Base insertion
- Base deletion

LocARNA - Aligning between secondary structures

Cases:

- Base match
- Base insertion
- Base deletion
- Base **pair** match

LocARNA - Aligning between secondary structures

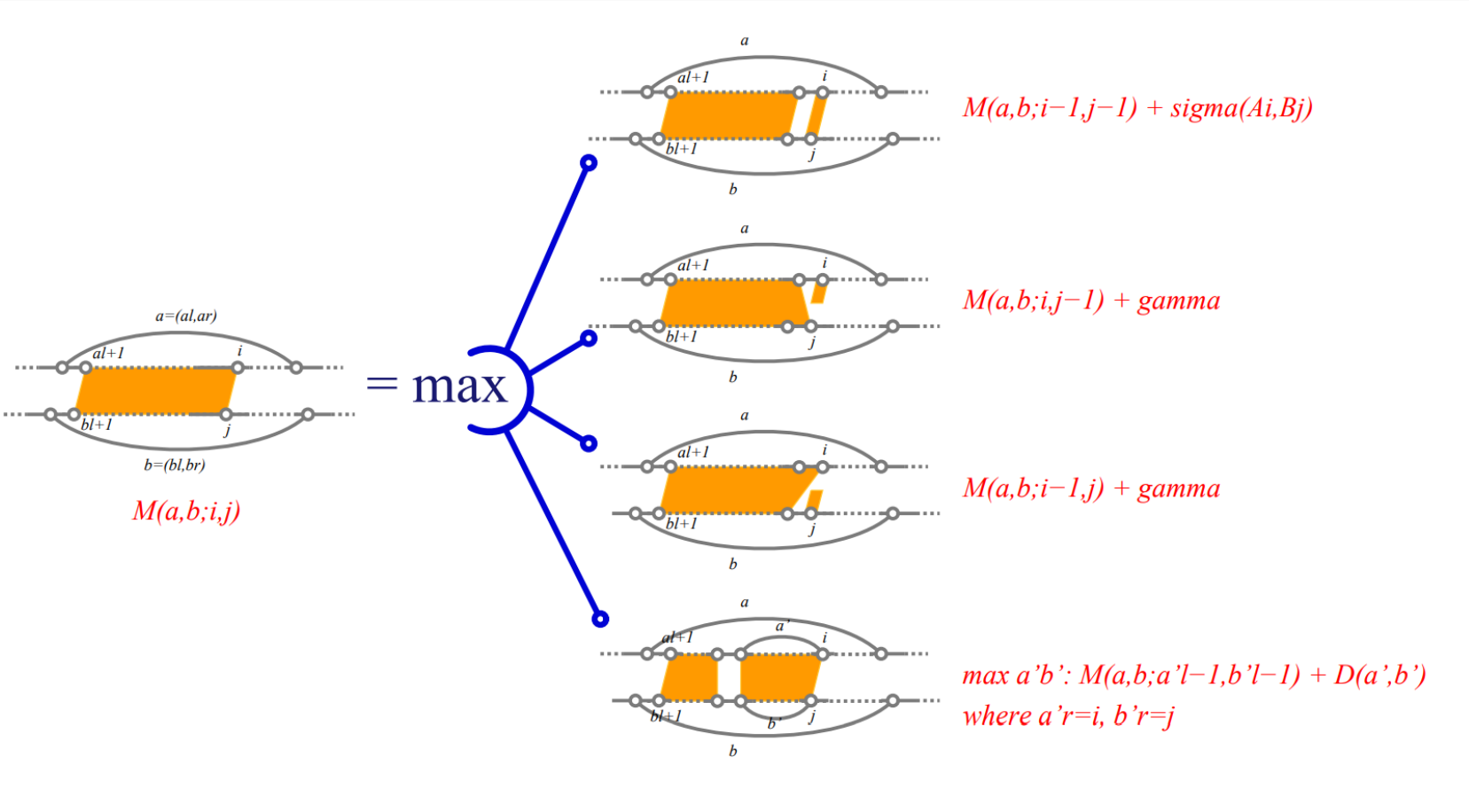


		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

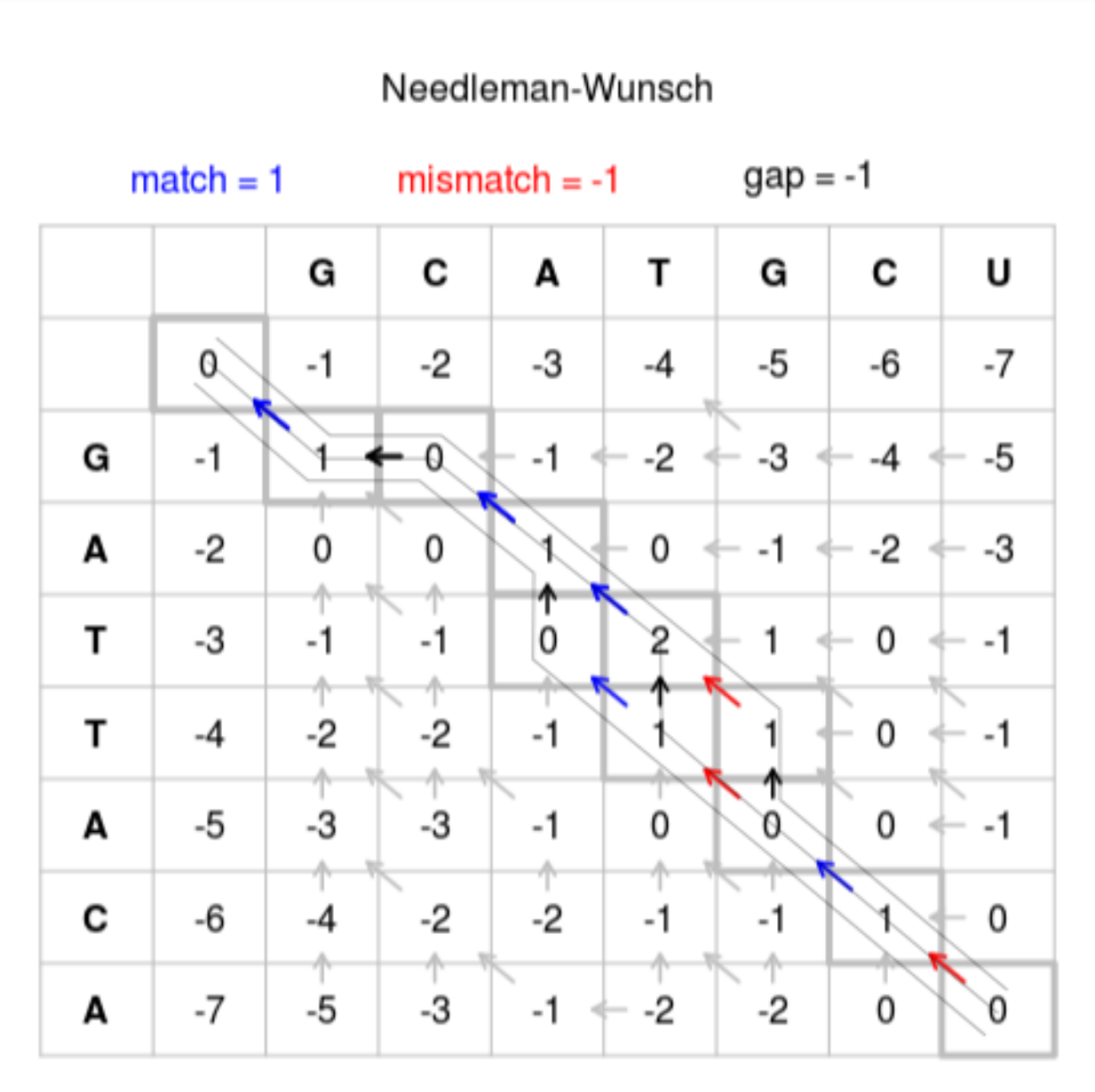
LocARNA - Demonstration

Demo Time!

LocARNA - Final actual Alignment

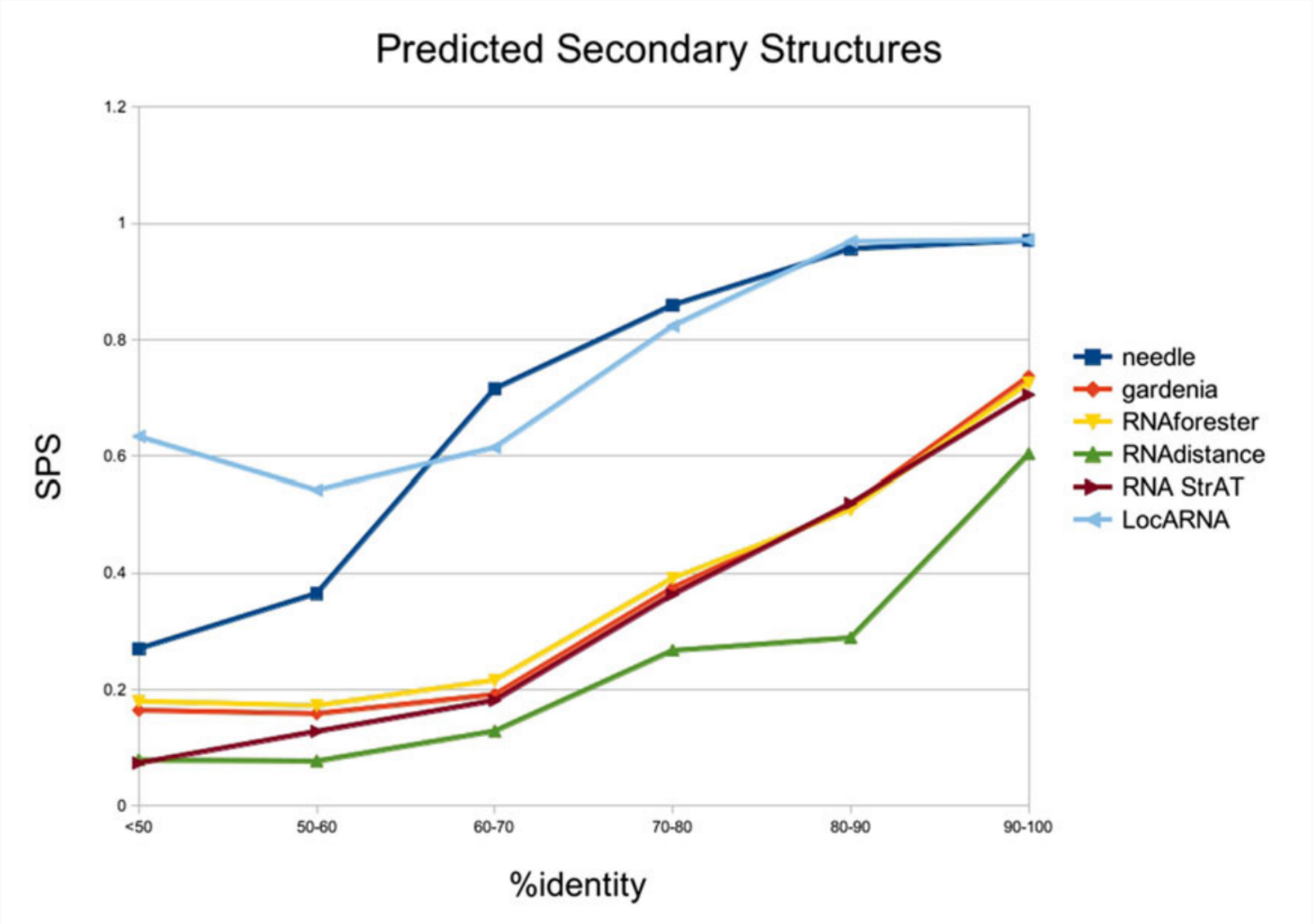


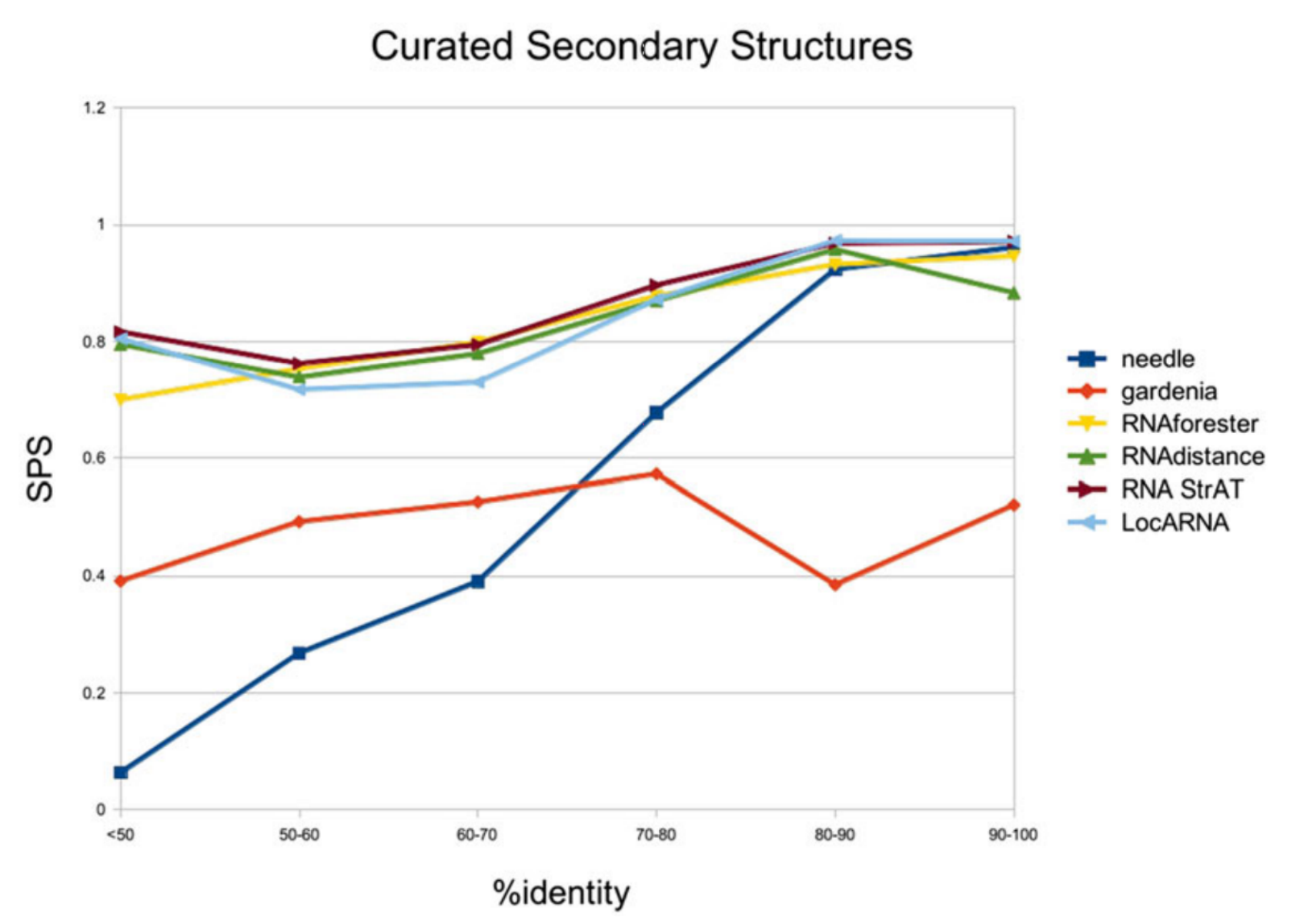
LocARNA - Final actual Alignment



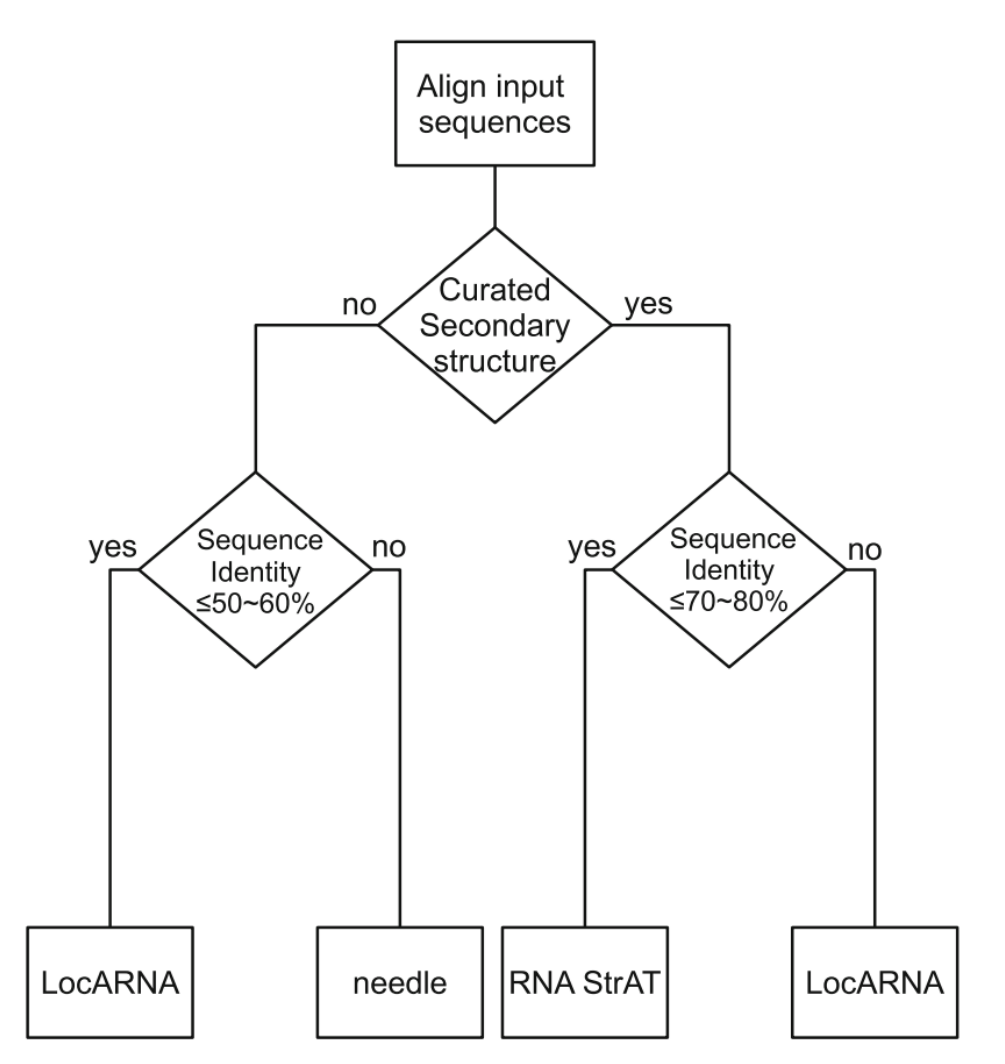
		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Content


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/bioinfII`

Sources ii

- **Needleman-Wunsch 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

Sources iii

RNA-Bioinformatics Lecture

https://ilias.uni-freiburg.de/ilias.php?ref_id=1009368&obj_id=1&cmd=layout&cmdClass=illmpresentationgui&cmdNode=fu&baseClass=ilLMPresentationGUI

RNA-Comparison Lecture MIT

https://math.mit.edu/classes/18.417/Slides/RNA-comparisonIV.pdf