MiGaL

RNA secondary structure modelling and comparison.

J. Allali & M.F. Sagot

IGM, Marne la Vallée, INRIA Rhône-Alpes

MiGaL: plan

- Introduction about RNA.

MiGaL: plan

- Introduction about RNA.
- Previous Work.

MiGaL: plan

- Introduction about RNA.
- Previous Work.
- MiGaL: a new modelling scheme.

MiGaL: plan

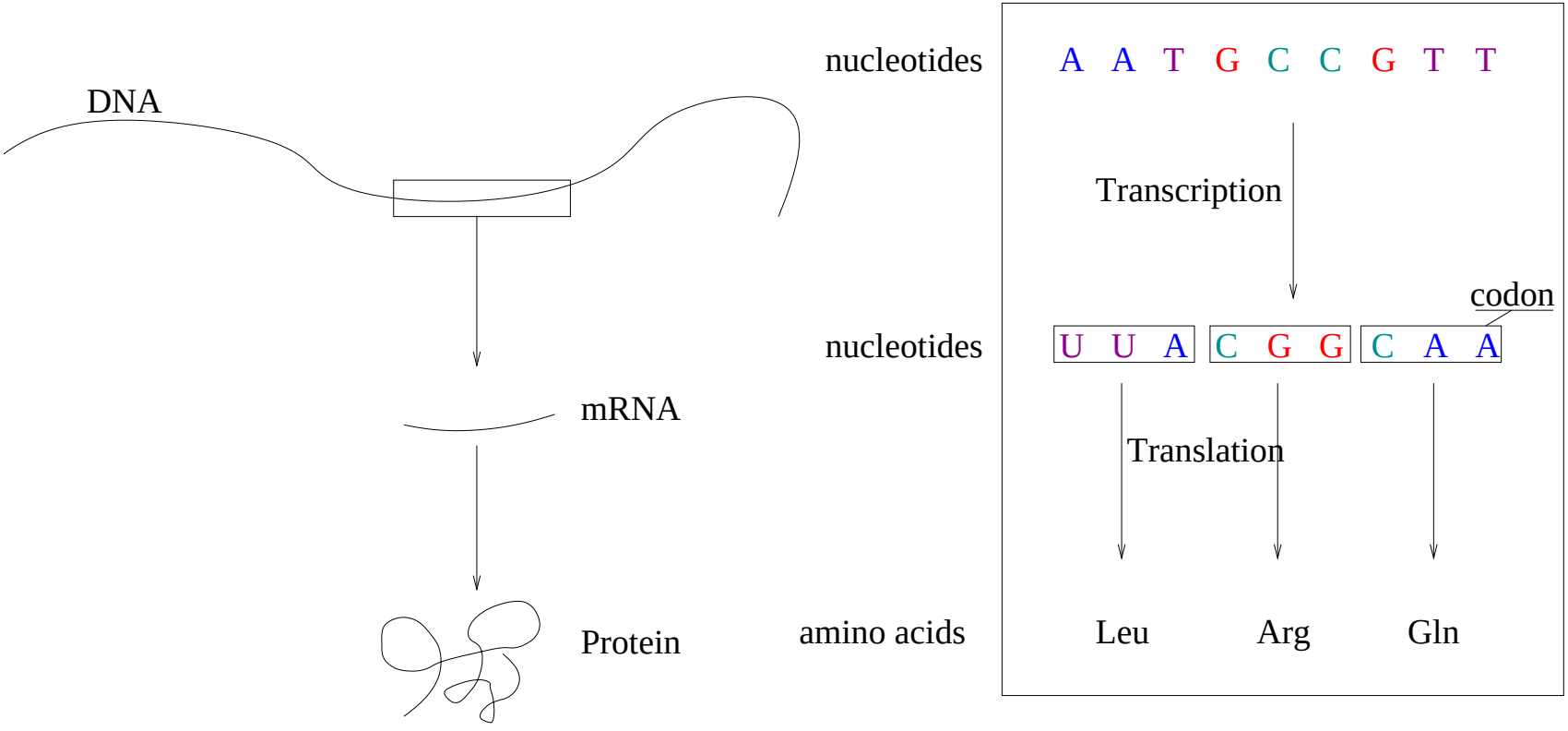
- Introduction about RNA.
- Previous Work.
- MiGaL: a new modelling scheme.
- An algorithm for the “fusion” of two ordered trees.

MiGaL: plan

- Introduction about RNA.
- Previous Work.
- MiGaL: a new modelling scheme.
- An algorithm for the “fusion” of two ordered trees.
- Conclusion.

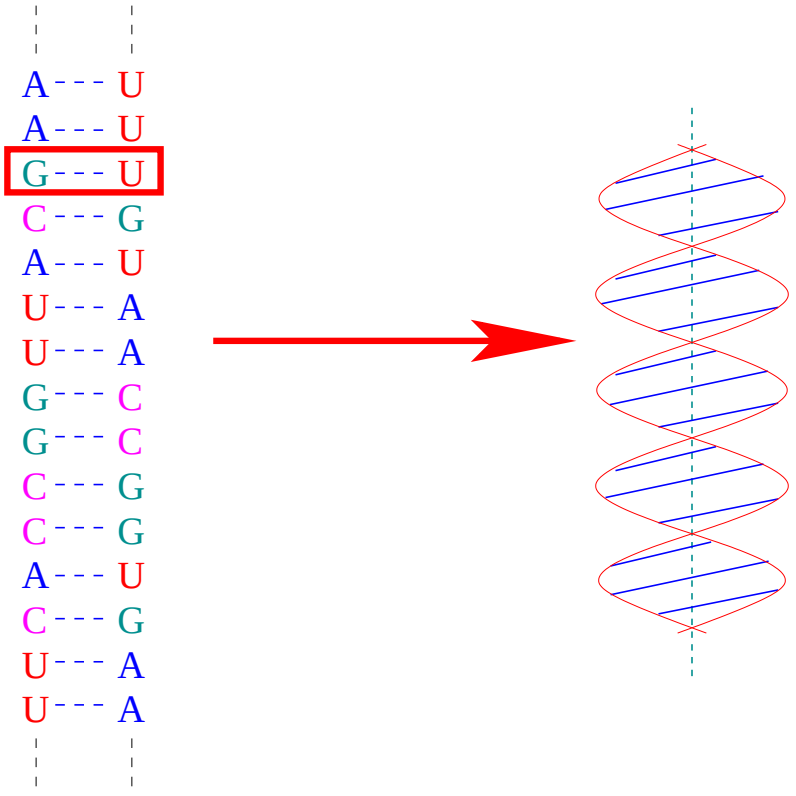
Introduction

MiGaL: introduction



MiGaL: introduction

RNA folding:

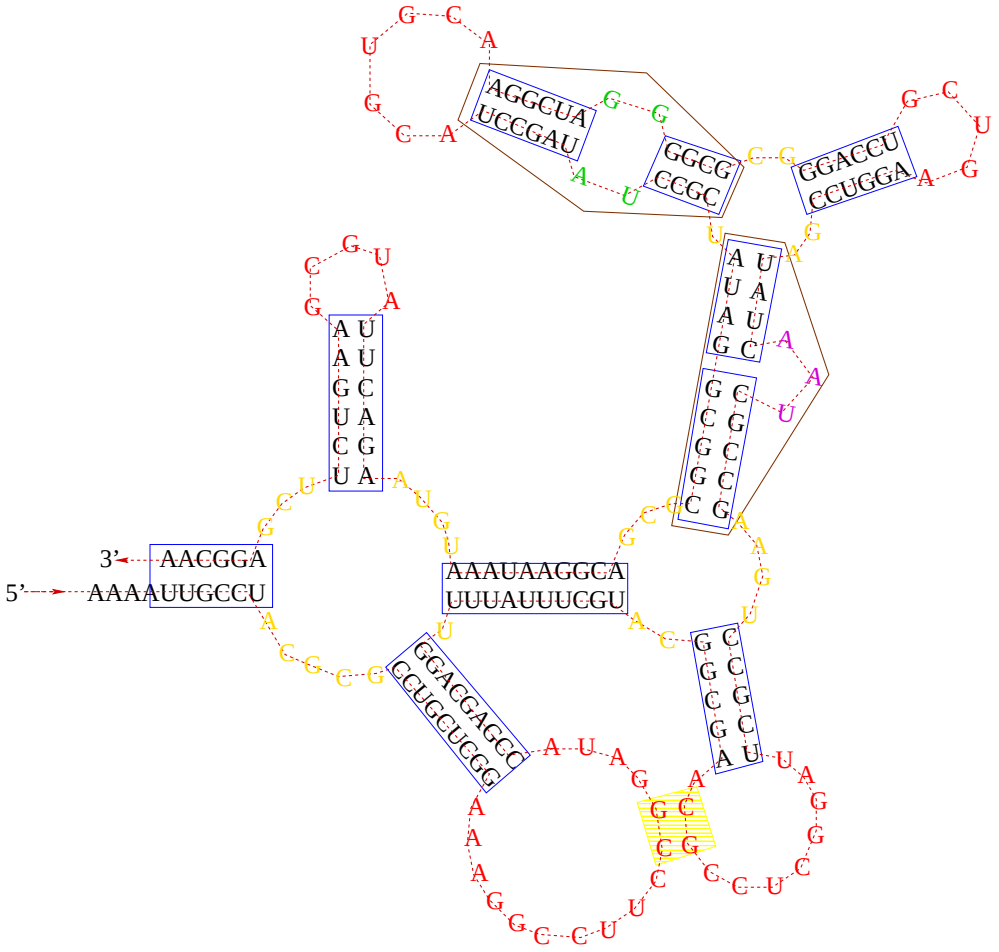


Base pairing

- Watson Crick: (A-U) (C-G)
- Wobble: (G-U)
- non-canonical: Holbrook, 91 (C-U) Noller, 84 (G-A)

MiGaL: introduction

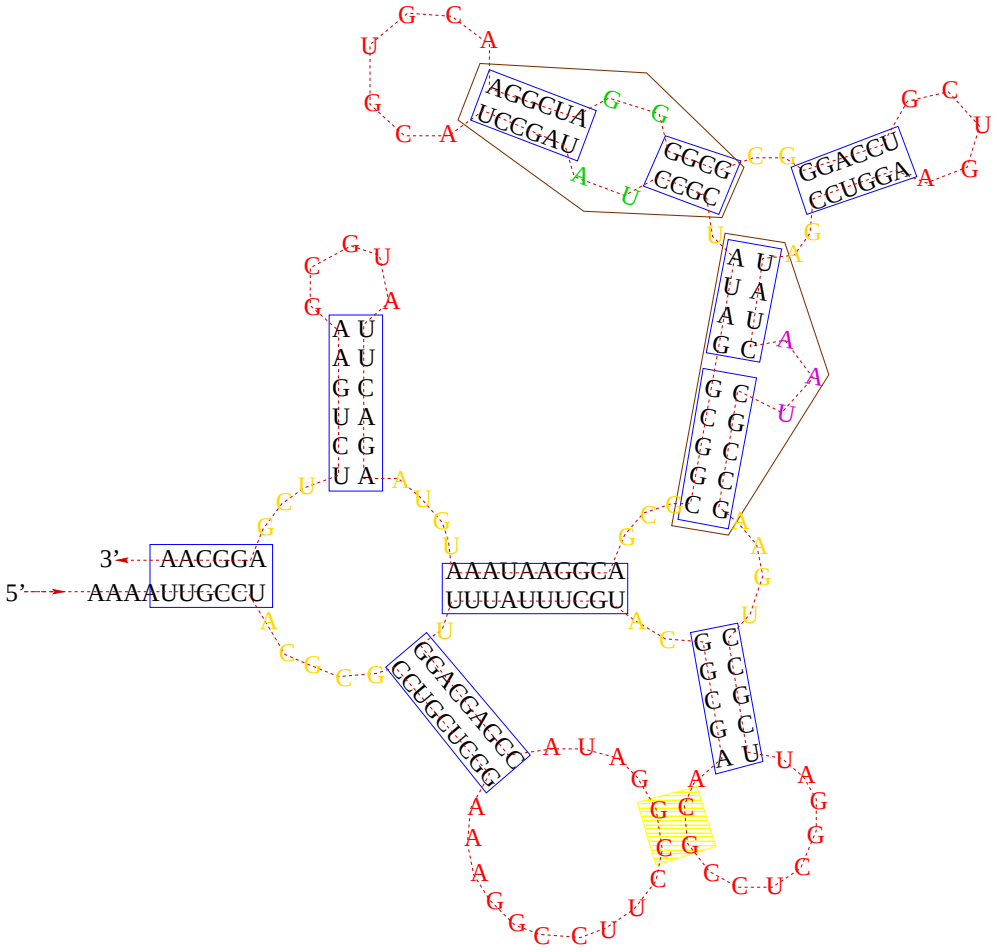
RNA Secondary Structure:



- Helix.

MiGaL: introduction

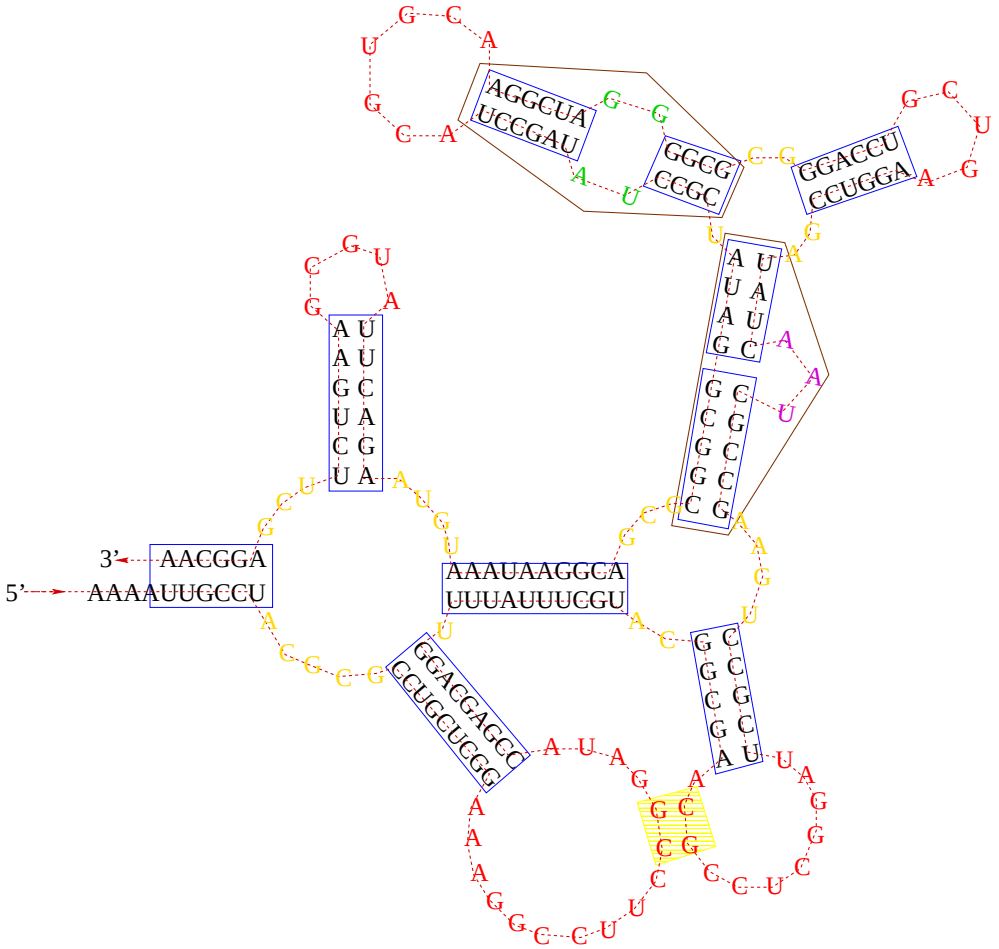
RNA Secondary Structure:



- Helix.
- Hairpin Loop.

MiGaL: introduction

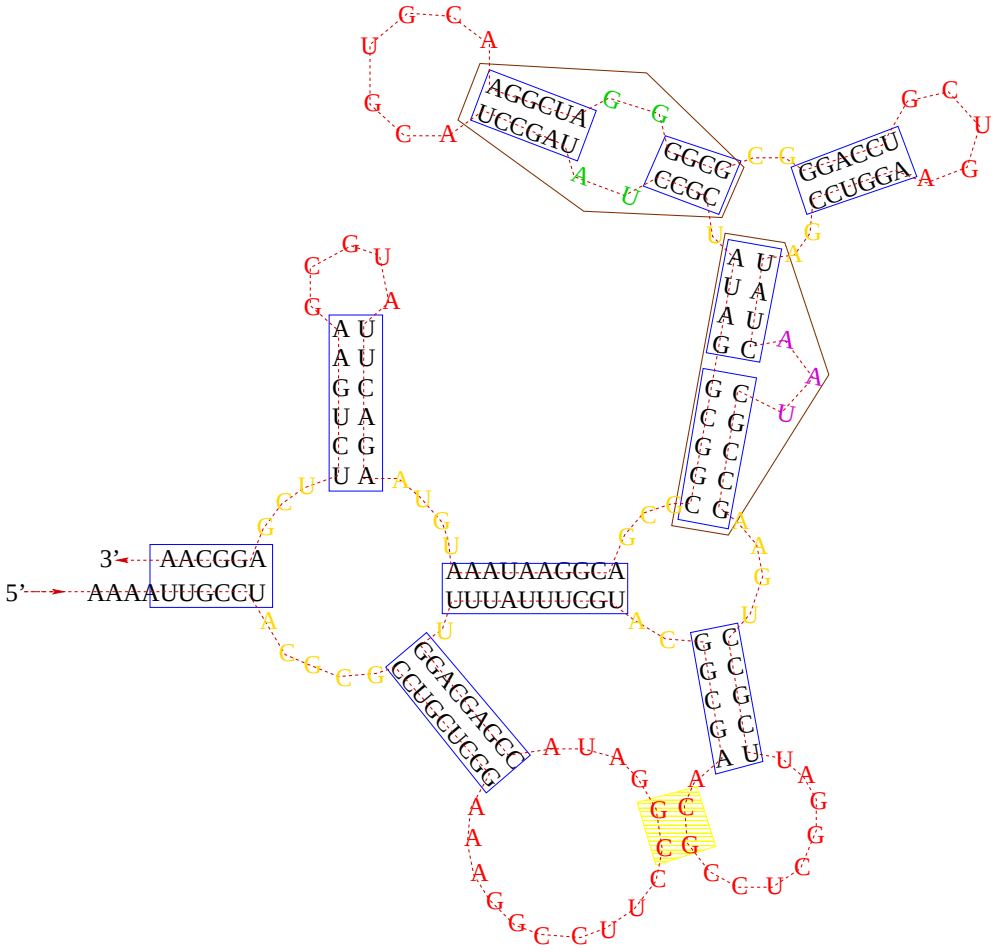
RNA Secondary Structure:



- Helix.
- Hairpin Loop.
- Multiloop.

MiGaL: introduction

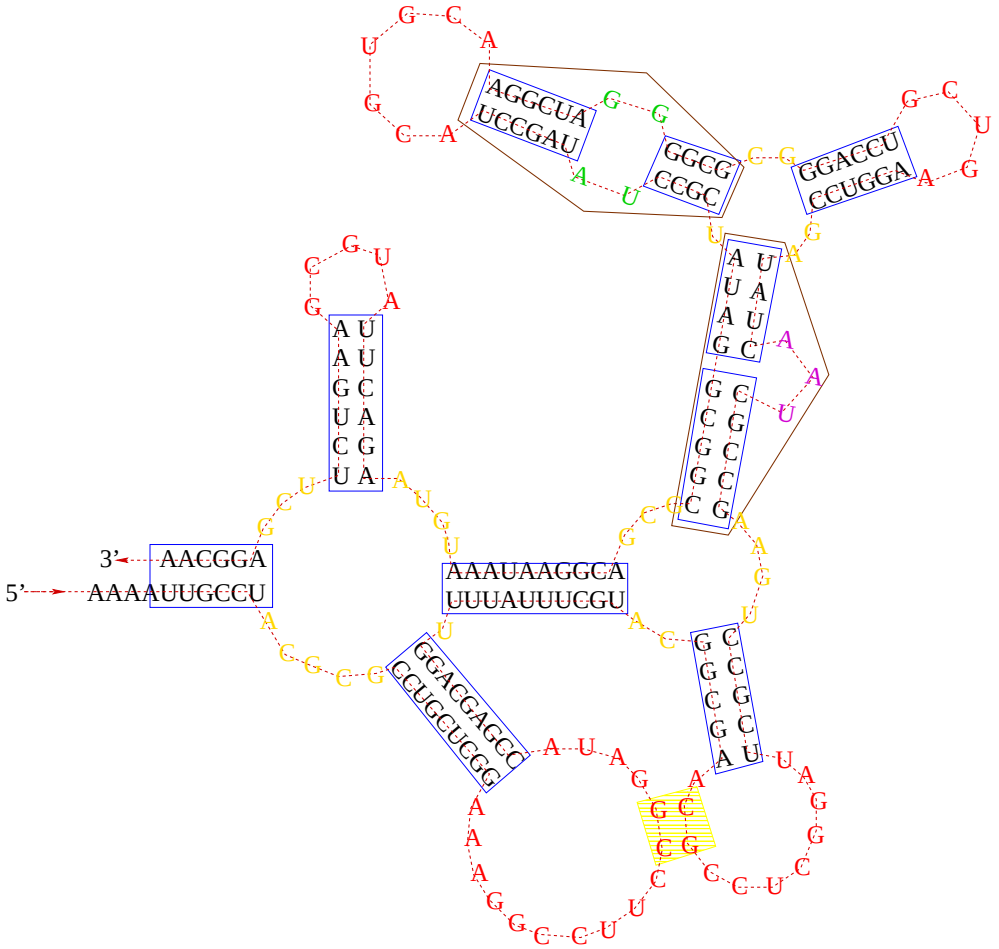
RNA Secondary Structure:



- Helix.
- Hairpin Loop.
- Multiloop.
- Bulge.

MiGaL: introduction

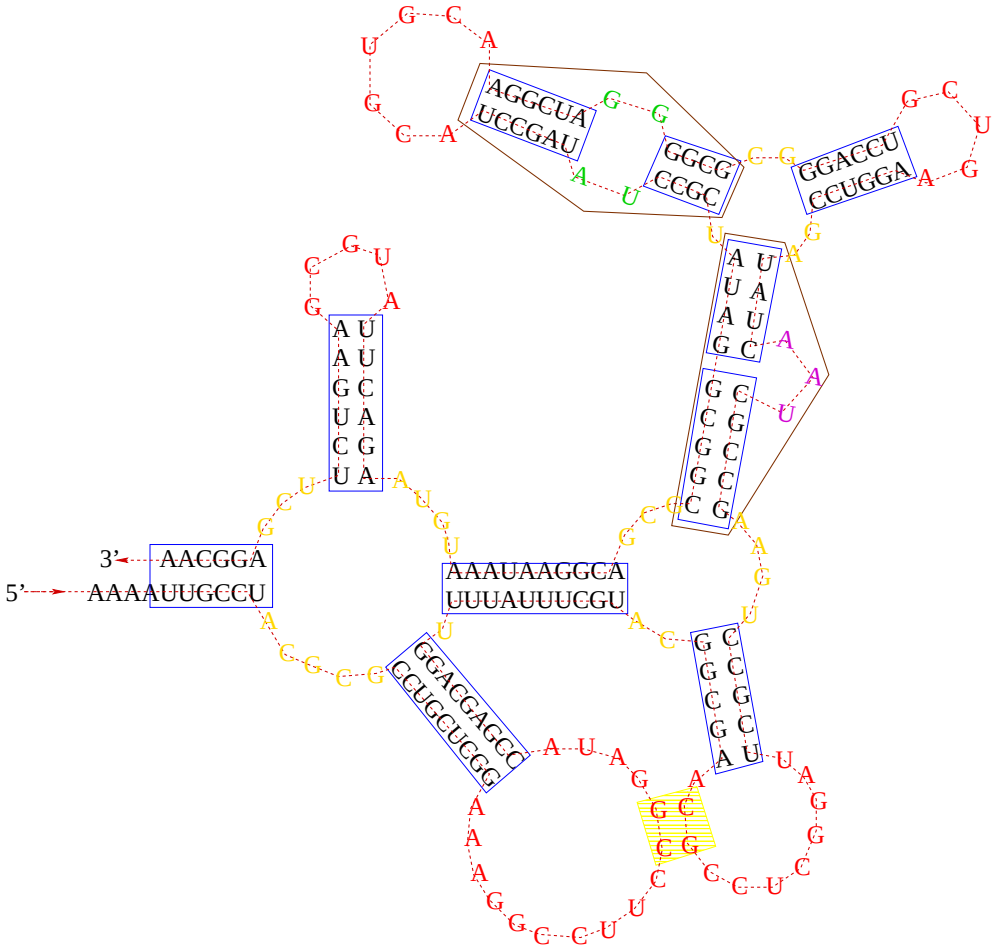
RNA Secondary Structure:



- Helix.
- Hairpin Loop.
- Multiloop.
- Bulge.
- Internal Loop.

MiGaL: introduction

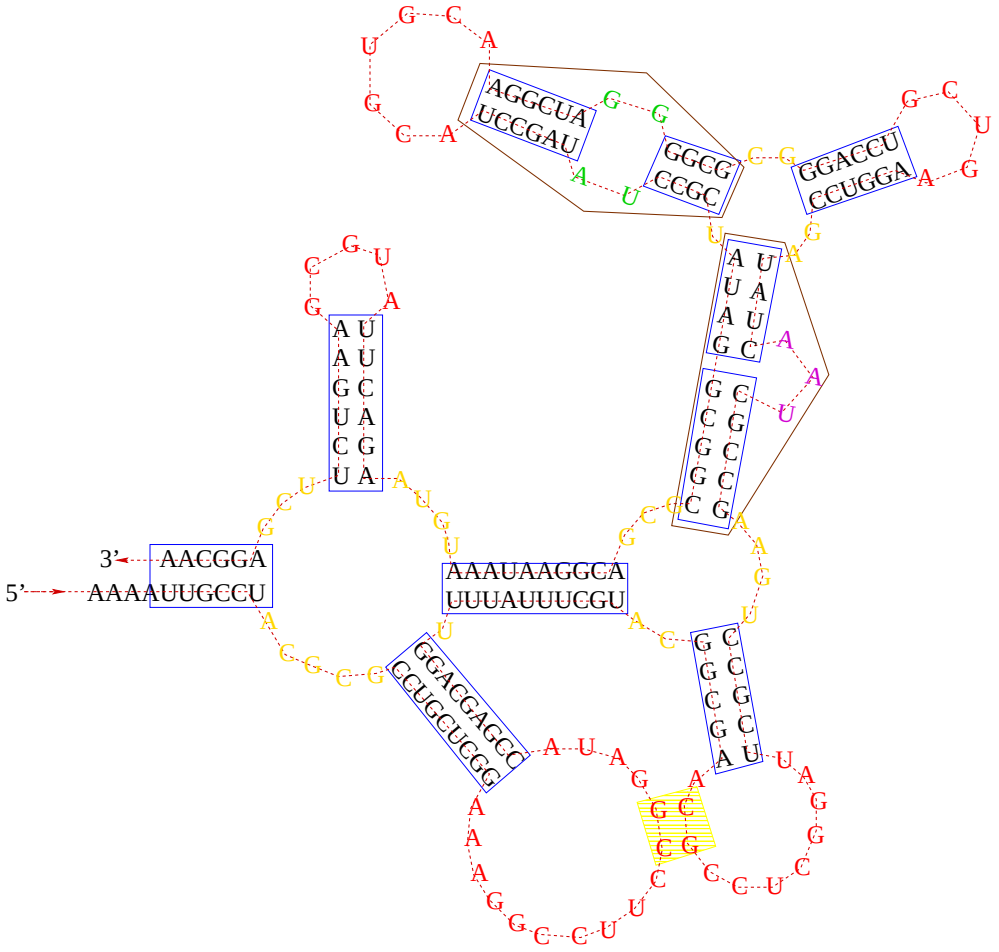
RNA Secondary Structure:



- Helix.
- Hairpin Loop.
- Multiloop.
- Bulge.
- Internal Loop.
- Stem.

MiGaL: introduction

RNA Secondary Structure:



- Helix.
- Hairpin Loop.
- Multiloop.
- Bulge.
- Internal Loop.
- Stem.
- Pseudo knot.

MiGaL: Problems

- RNA secondary structure comparison taking into account pseudo knots .

- Motif inference for RNA secondary structures.

MiGaL: Problems

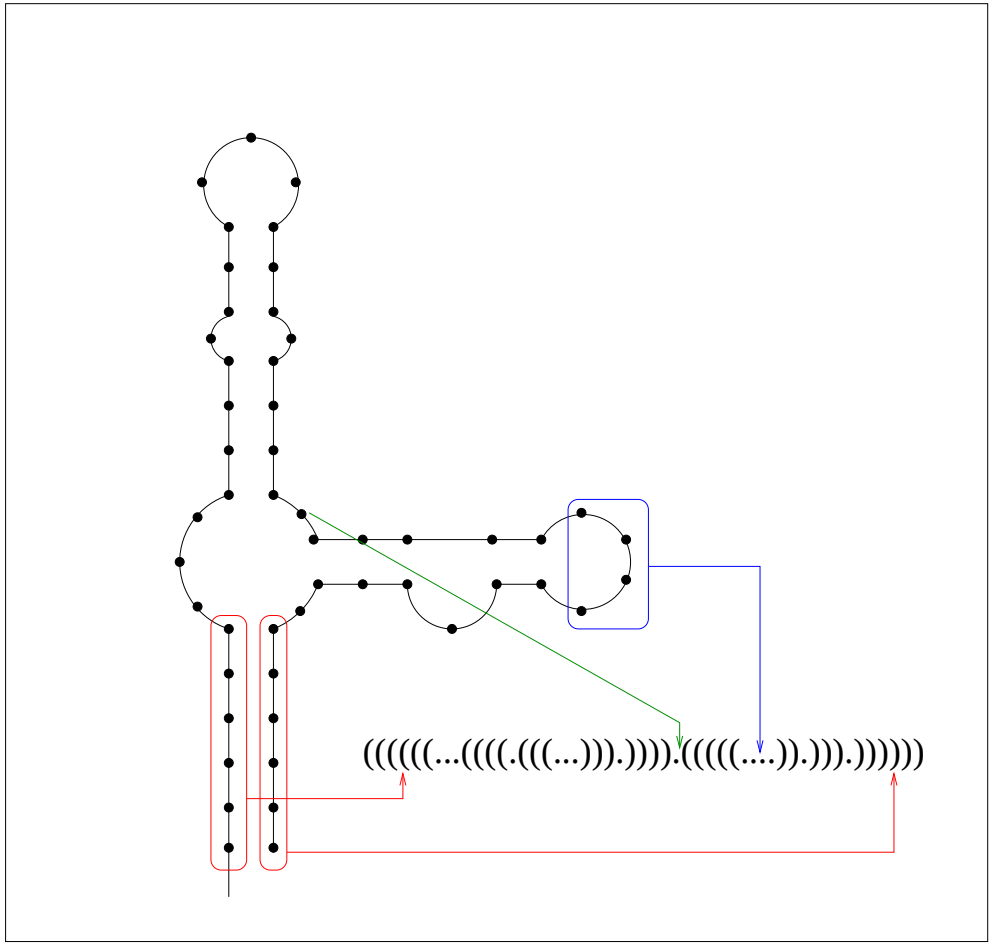
- RNA secondary structure comparison taking into account pseudo knots .

- Motif inference for RNA secondary structures.

we need a data structure for RNA secondary structures

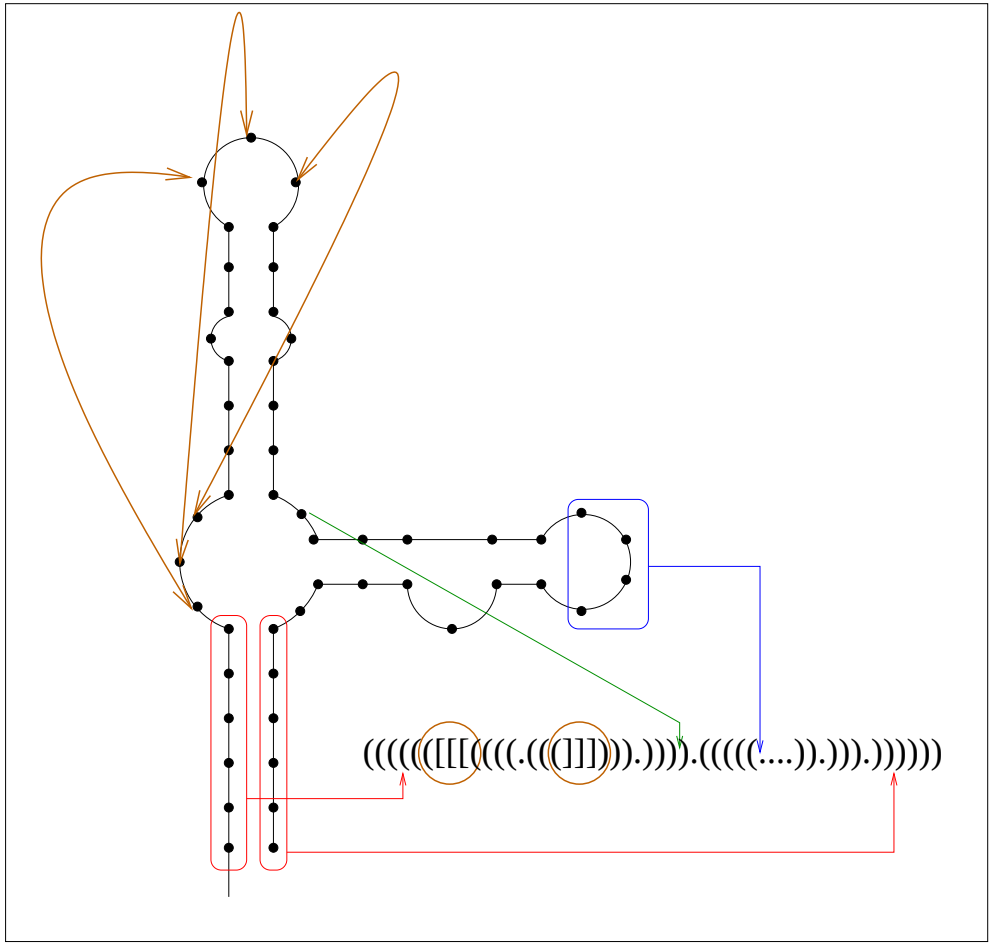
Previous Work

MiGaL: parenthesis sequence



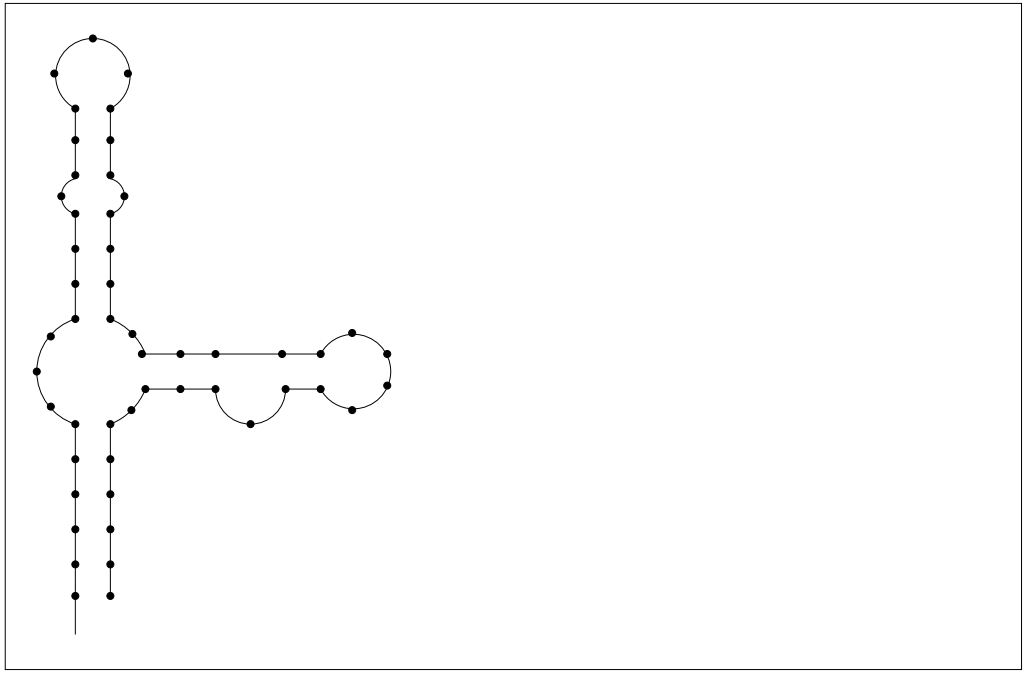
Also called “arc annotated sequencies”

MiGaL: parenthesis sequence



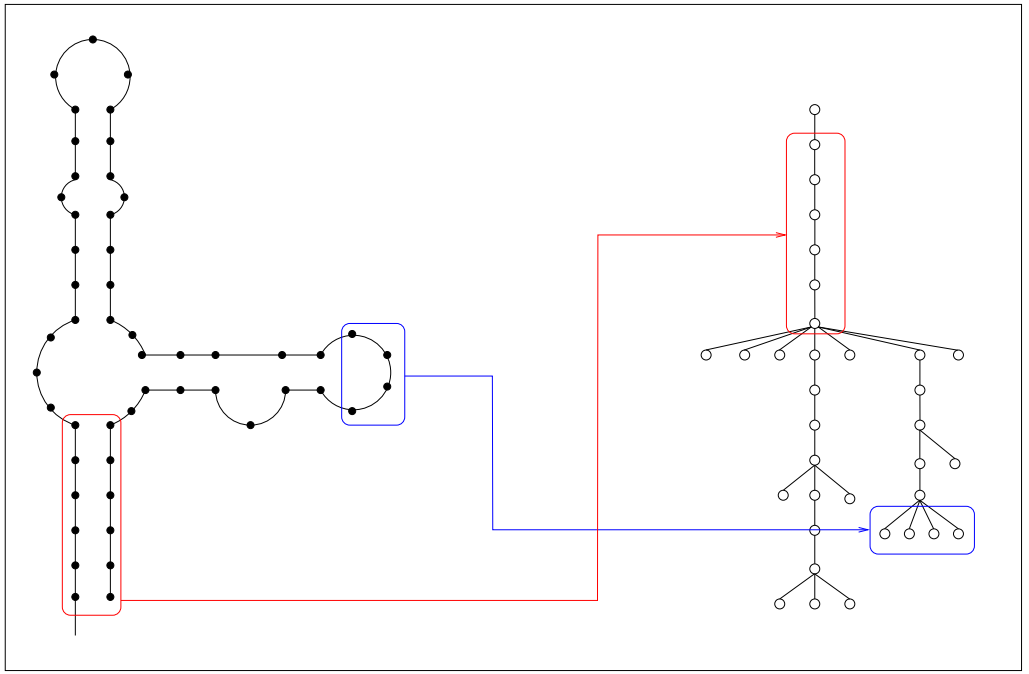
Zhang (99)
NP-Hard for crossing pairs.

MiGaL: tree representation



Several different representations using trees.

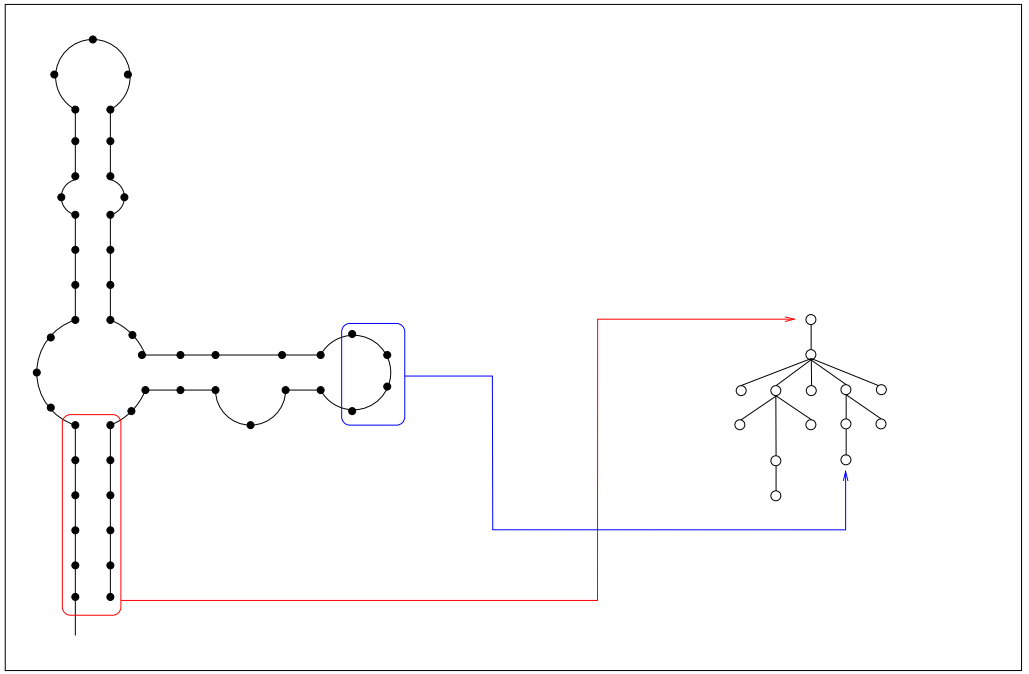
MiGaL: tree representation



Zhang (90,95)
Fontana(93)
. . .

internal node=base pair
leaf=unpaired base

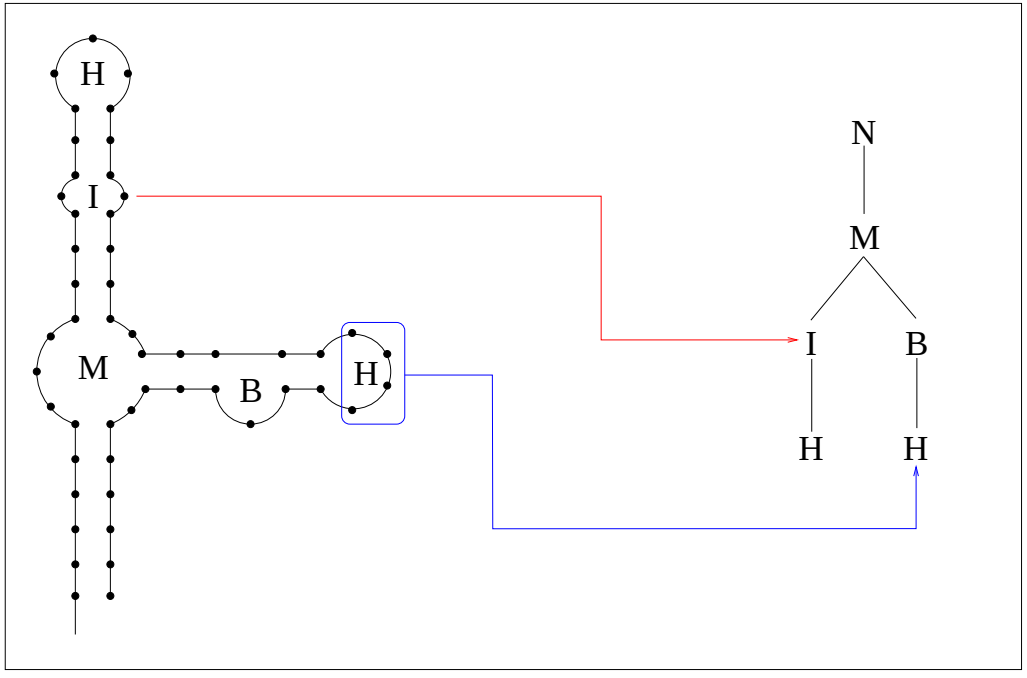
MiGaL: tree representation



Zhang (90)

internal node=helix,bulge,internal loop or multiloop
leaf=unpaired bases

MiGaL: tree representation



Shapiro (89)

internal node=helix,bulge,internal loop or multiloop
leaf=hairpin loop

MiGaL: tree representation(2)

By adding pseudo knots to this tree structure, we change it into a graph.

Comparison, edition . . . are difficult problems with graphs.

MiGaL: grammar

Stochastic Context Free Grammars : Sakakibara (93)

$$S \rightarrow (S)|()|s|SS$$

Used for alignment or folding algorithms.

S-attribute Grammars Lefebvre (96)

MiGaL: New Representation

Ideas:

- Start from a tree representation.

MiGaL: New Representation

Ideas:

- Start from a tree representation.
- Add edges for pseudo knots.

MiGaL: New Representation

Ideas:

- Start from a tree representation.
- Add edges for pseudo knots.
- Use simultaneously various levels of representation.

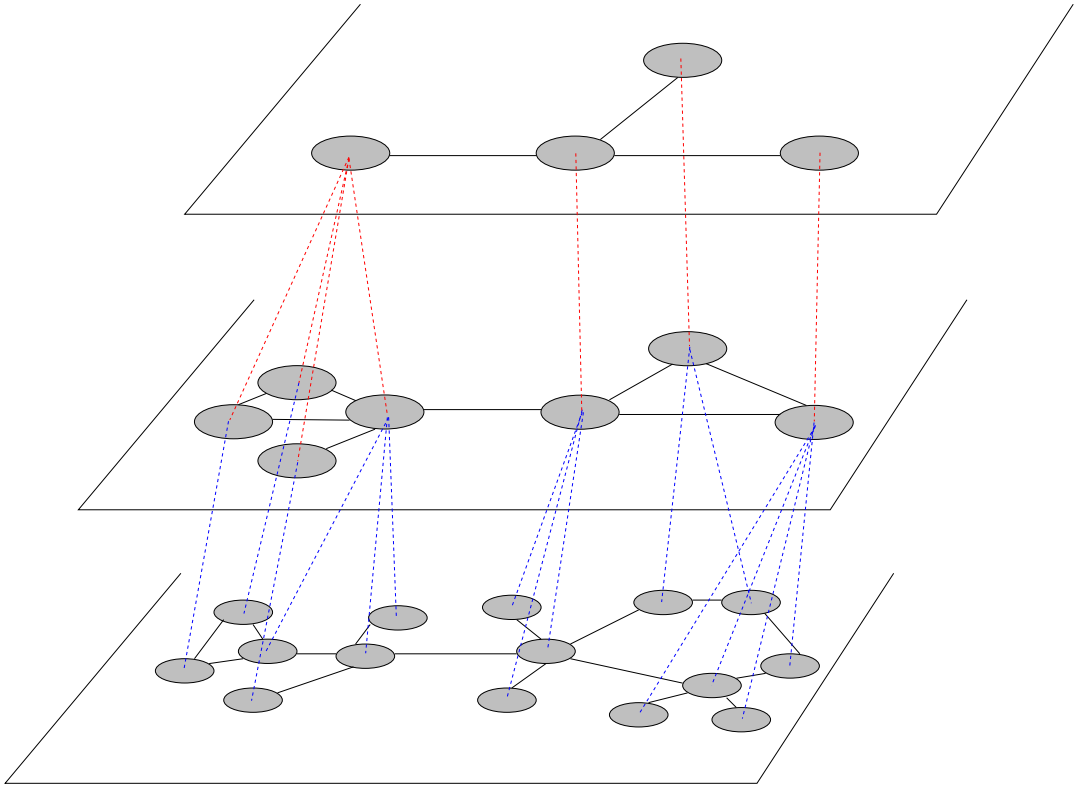
MiGaL: New Representation

Ideas:

- Start from a tree representation.
- Add edges for pseudo knots.
- Use simultaneously various levels of representation.

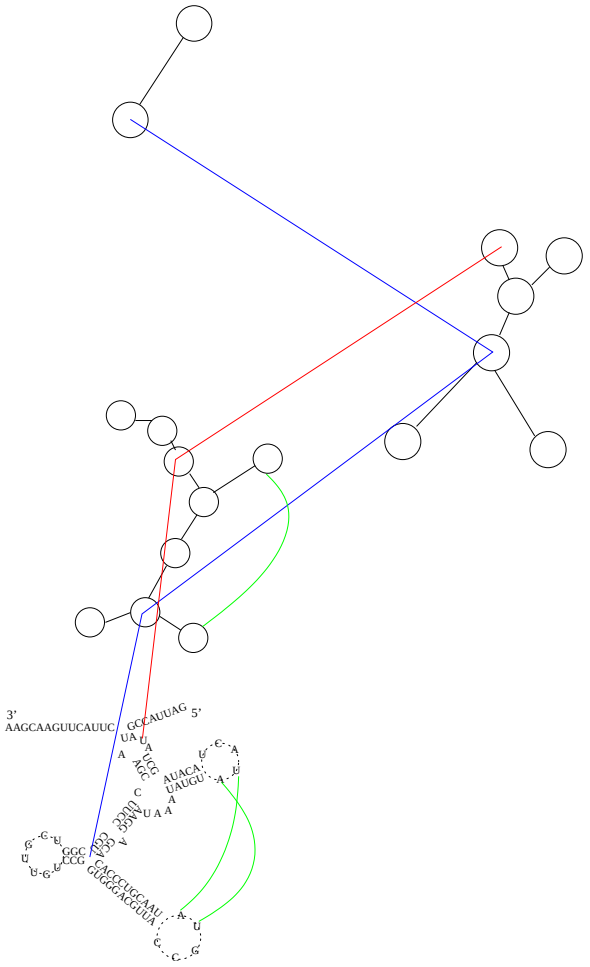
- Multiloop network is the backbone of the RNA structure.
- Nucleotide conservation is not necessary.

MiGaL: MultIGrAphLayer



- Detailed structure at the deepest level.
- Abstraction : the top level.
- Relation between adjacent layers.

MiGaL: On RNA



G0

multiloop tree.

G1

stem tree.

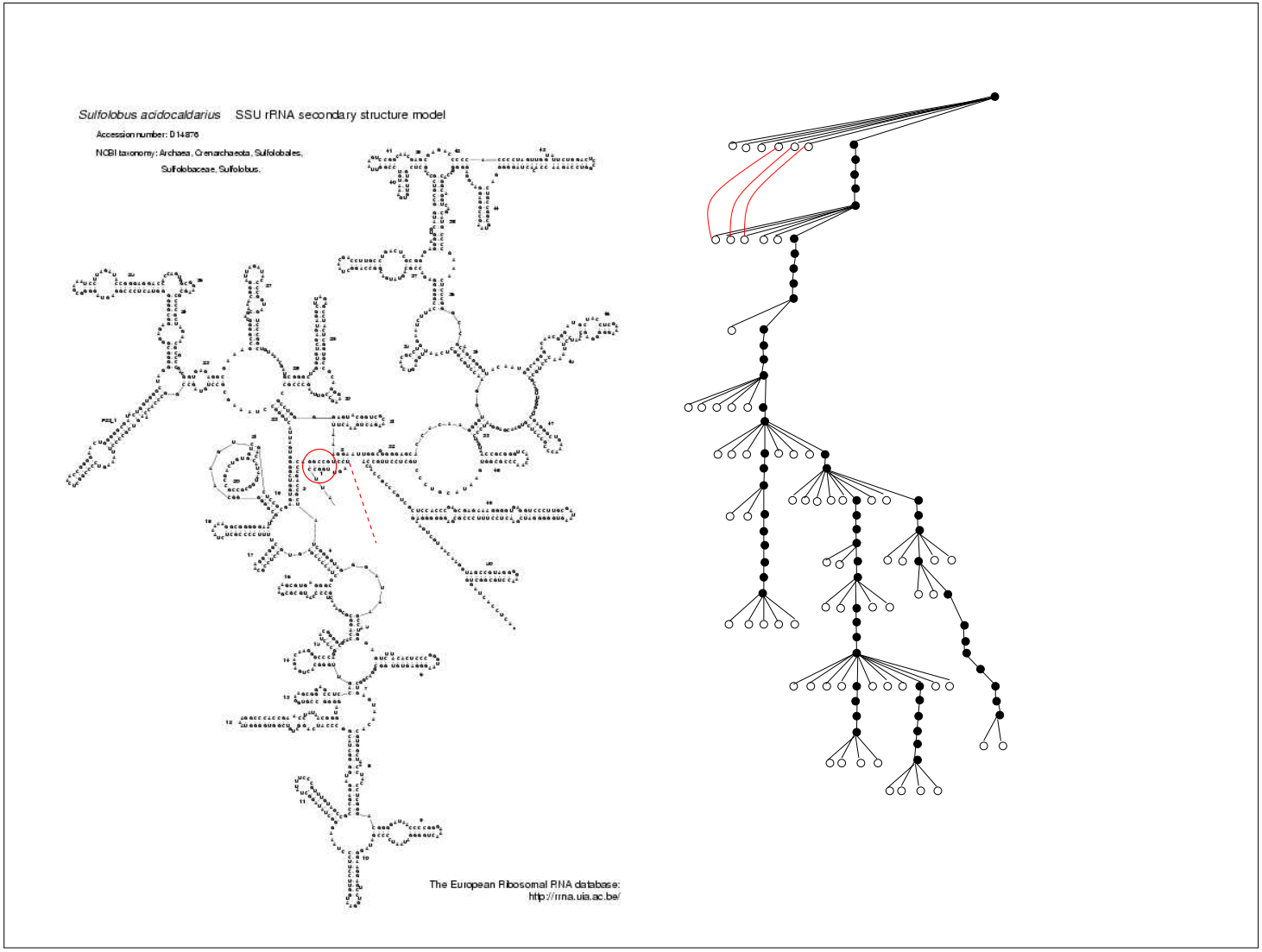
G2

helix graph

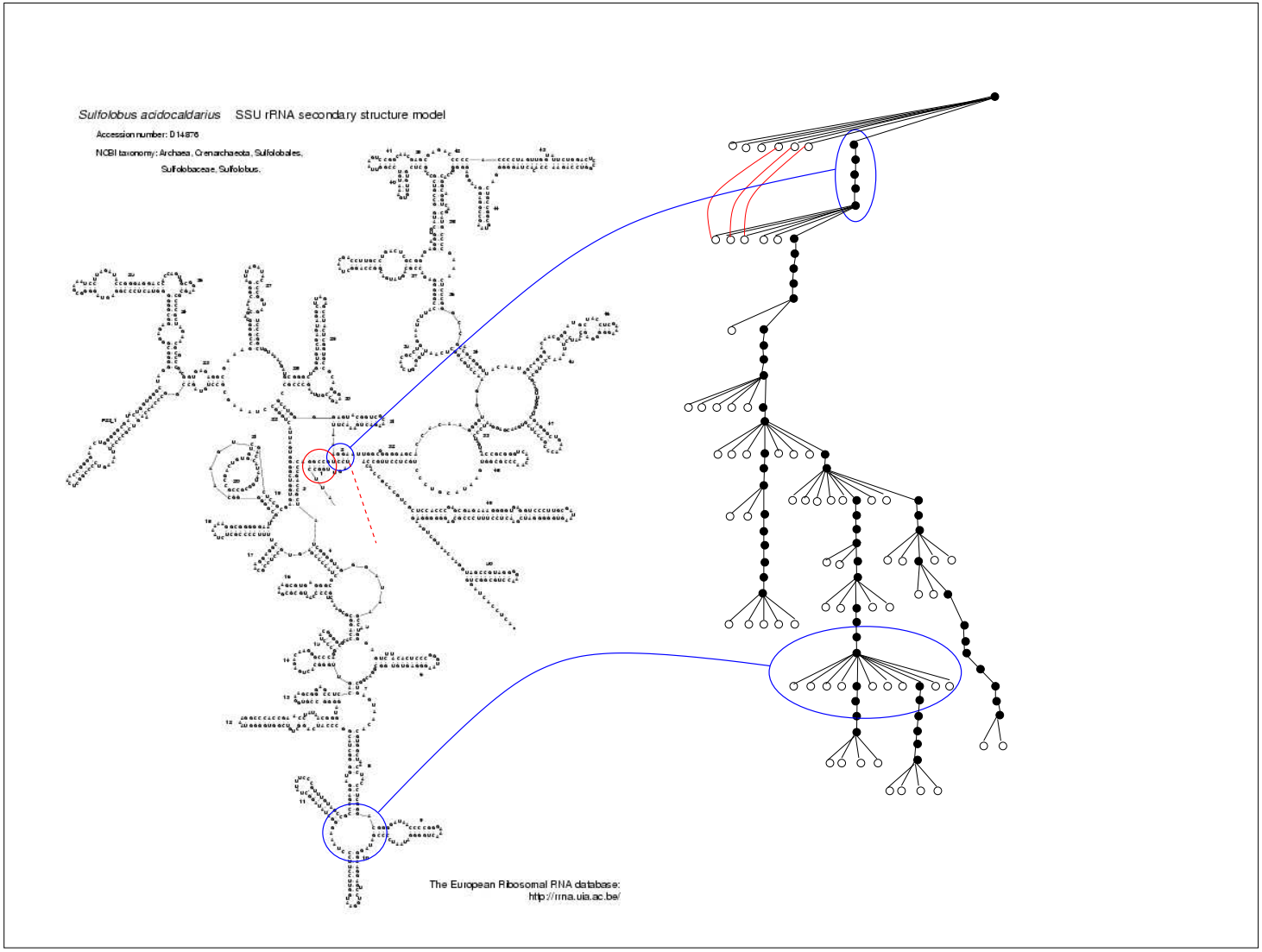
G3

base pair graph

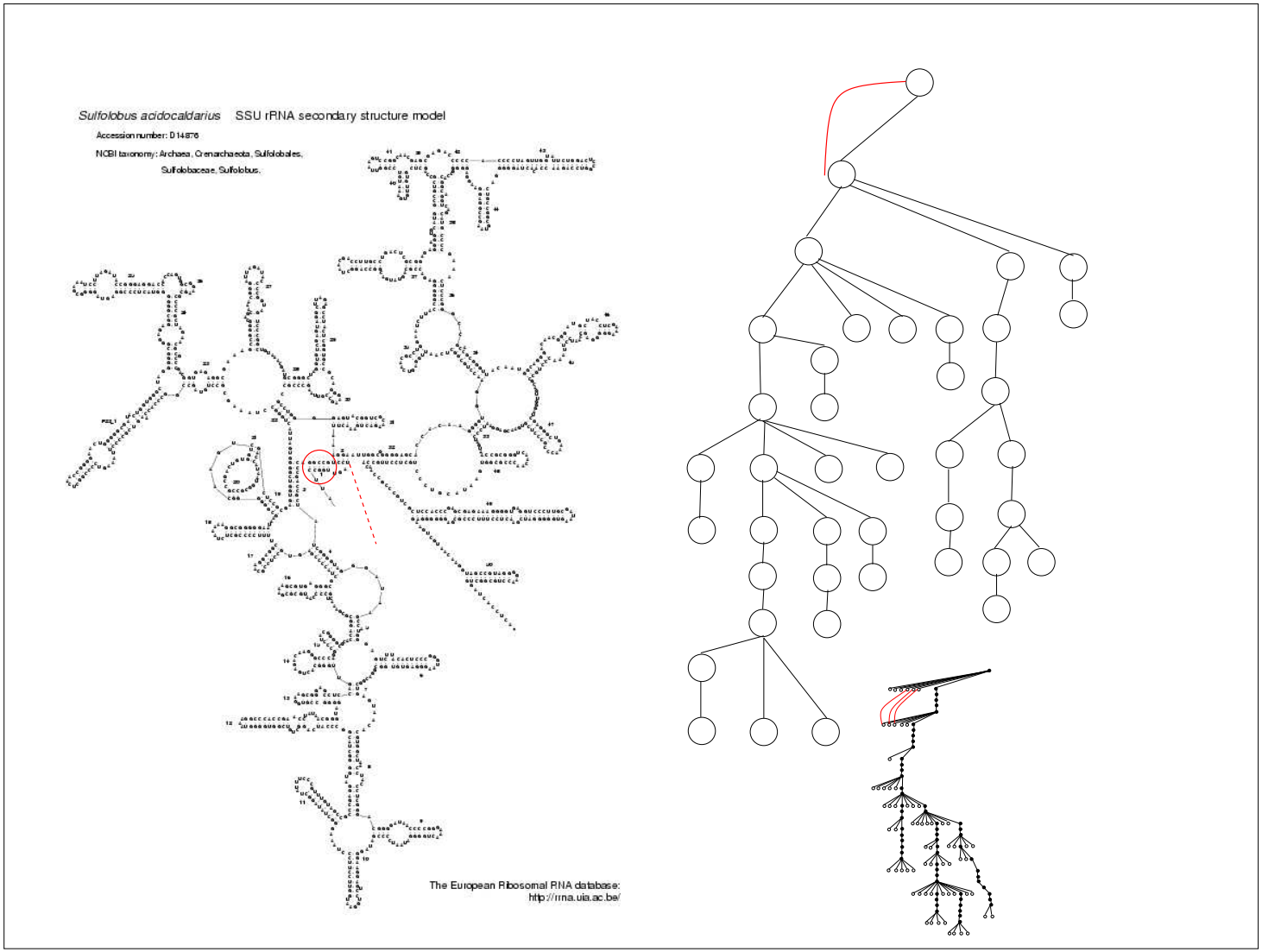
MiGaL: Layer 3



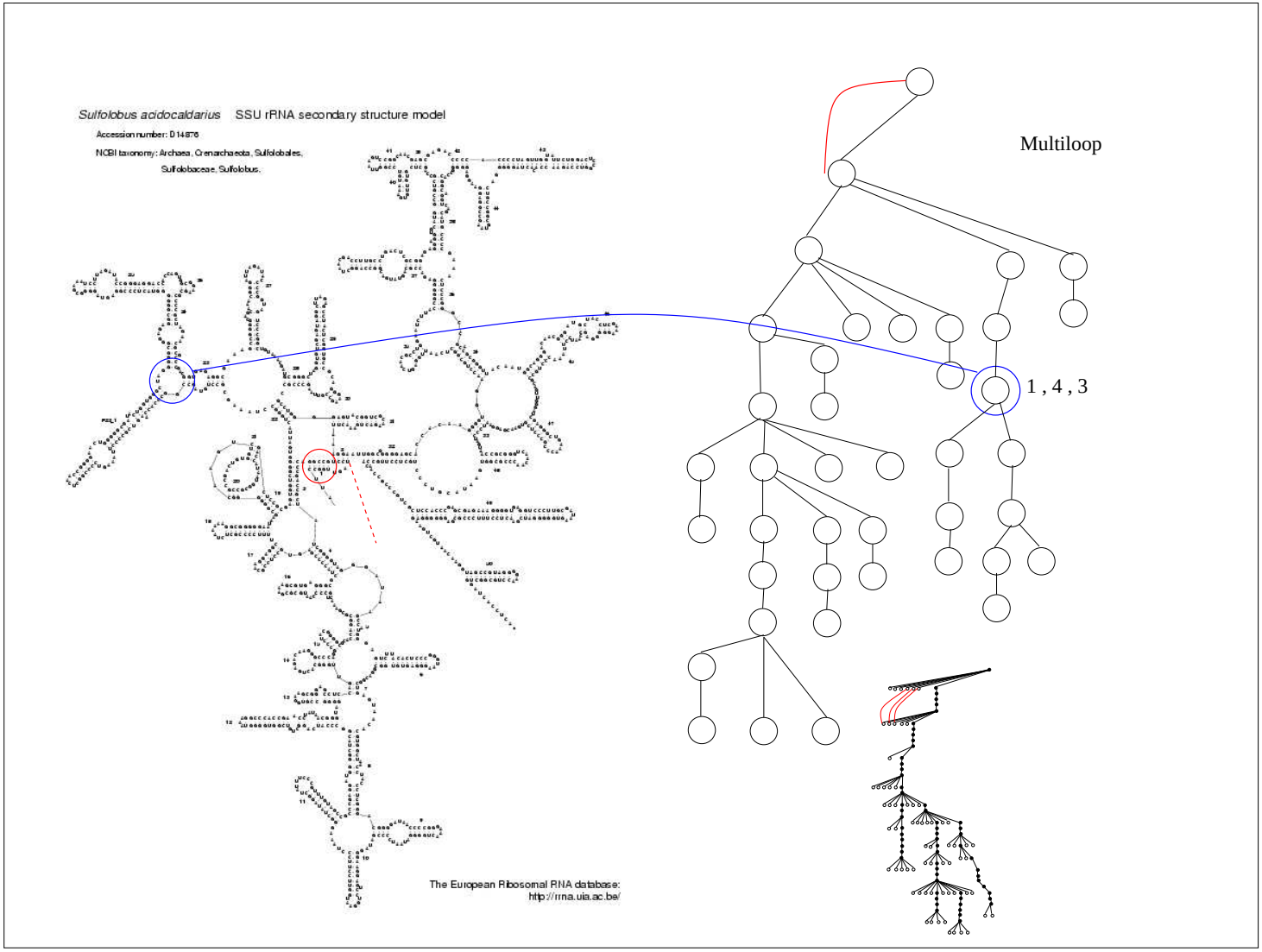
MiGaL: Layer 3



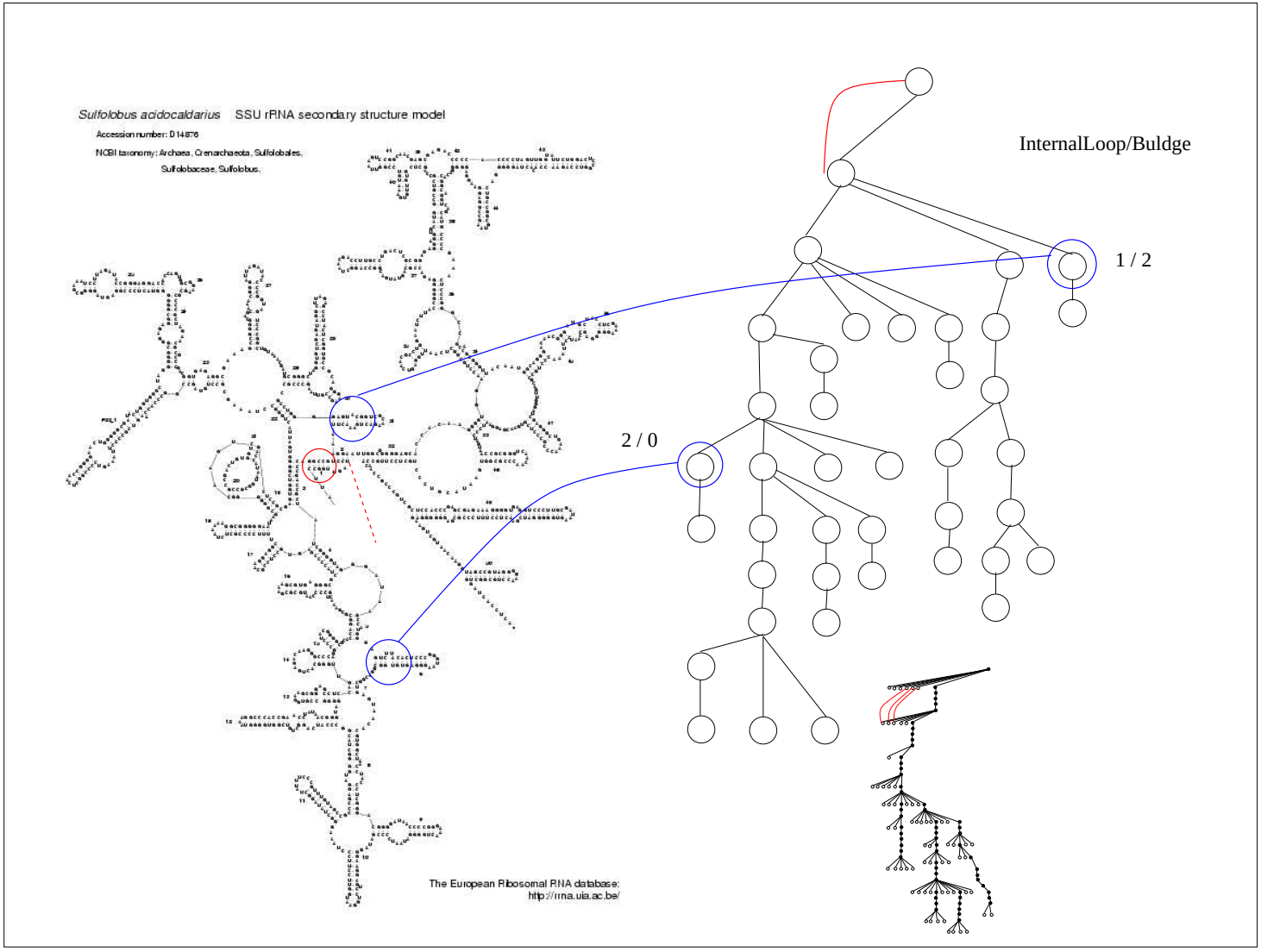
MiGaL: Layer 2



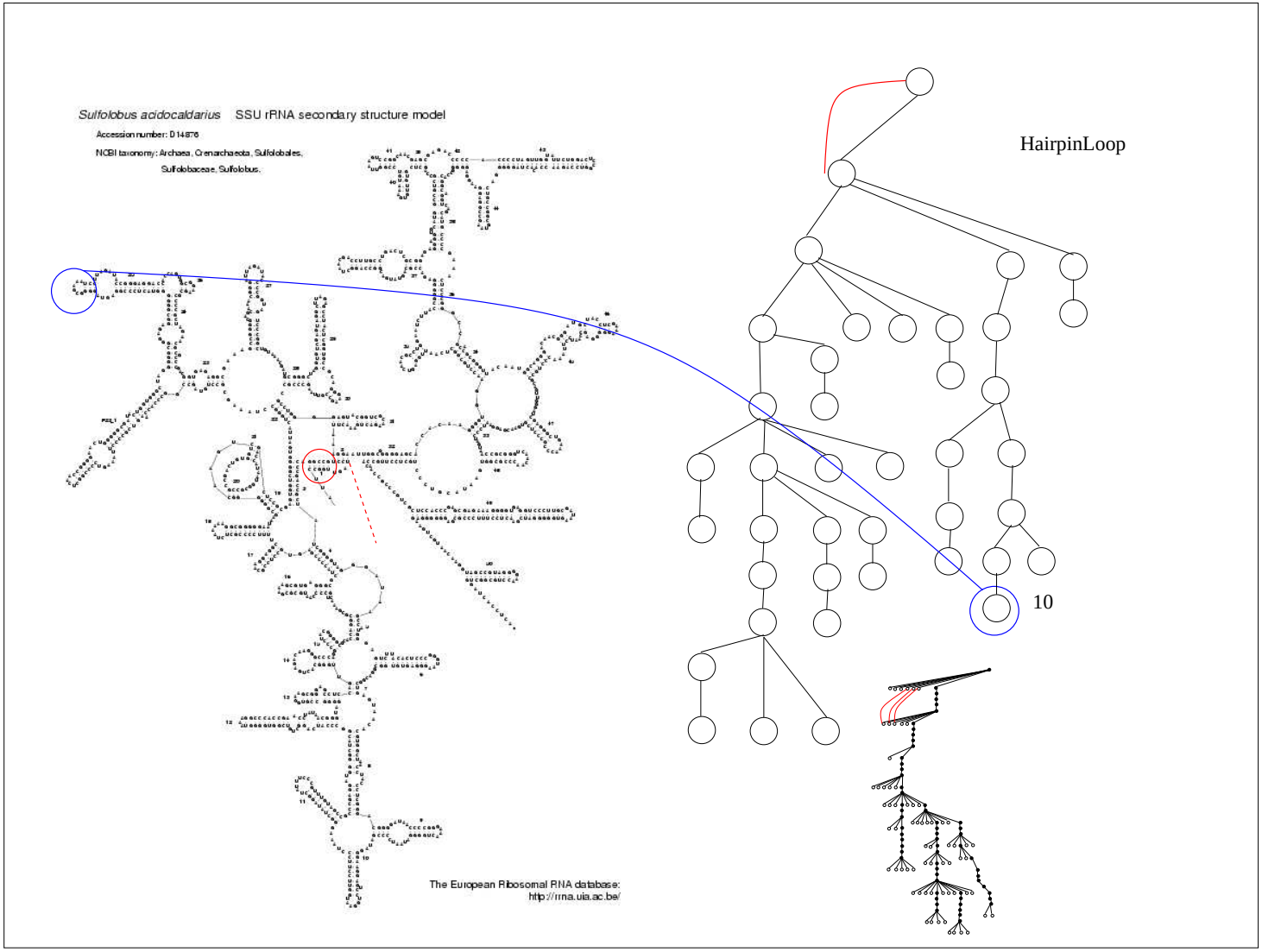
MiGaL: Layer 2



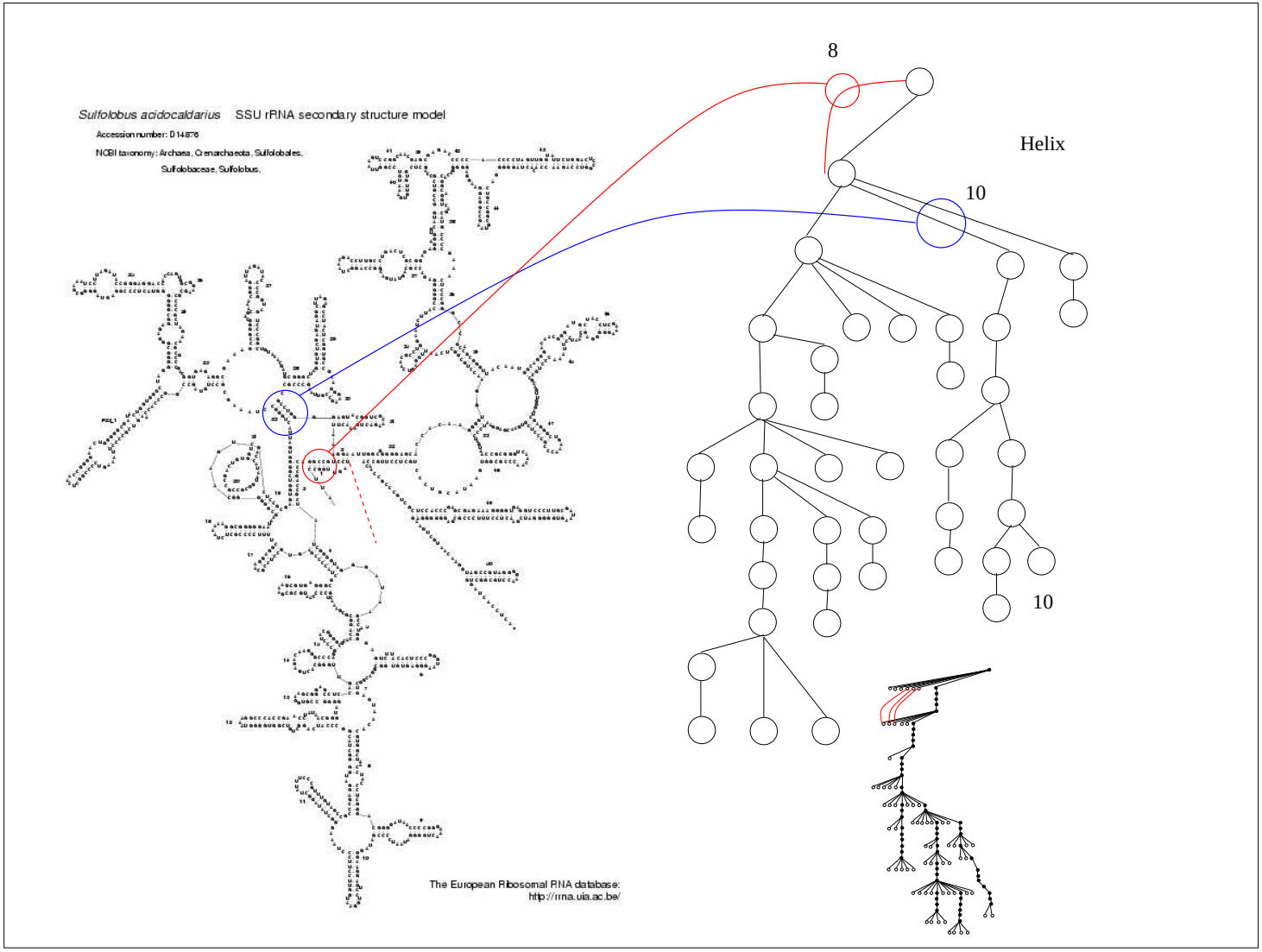
MiGaL: Layer 2



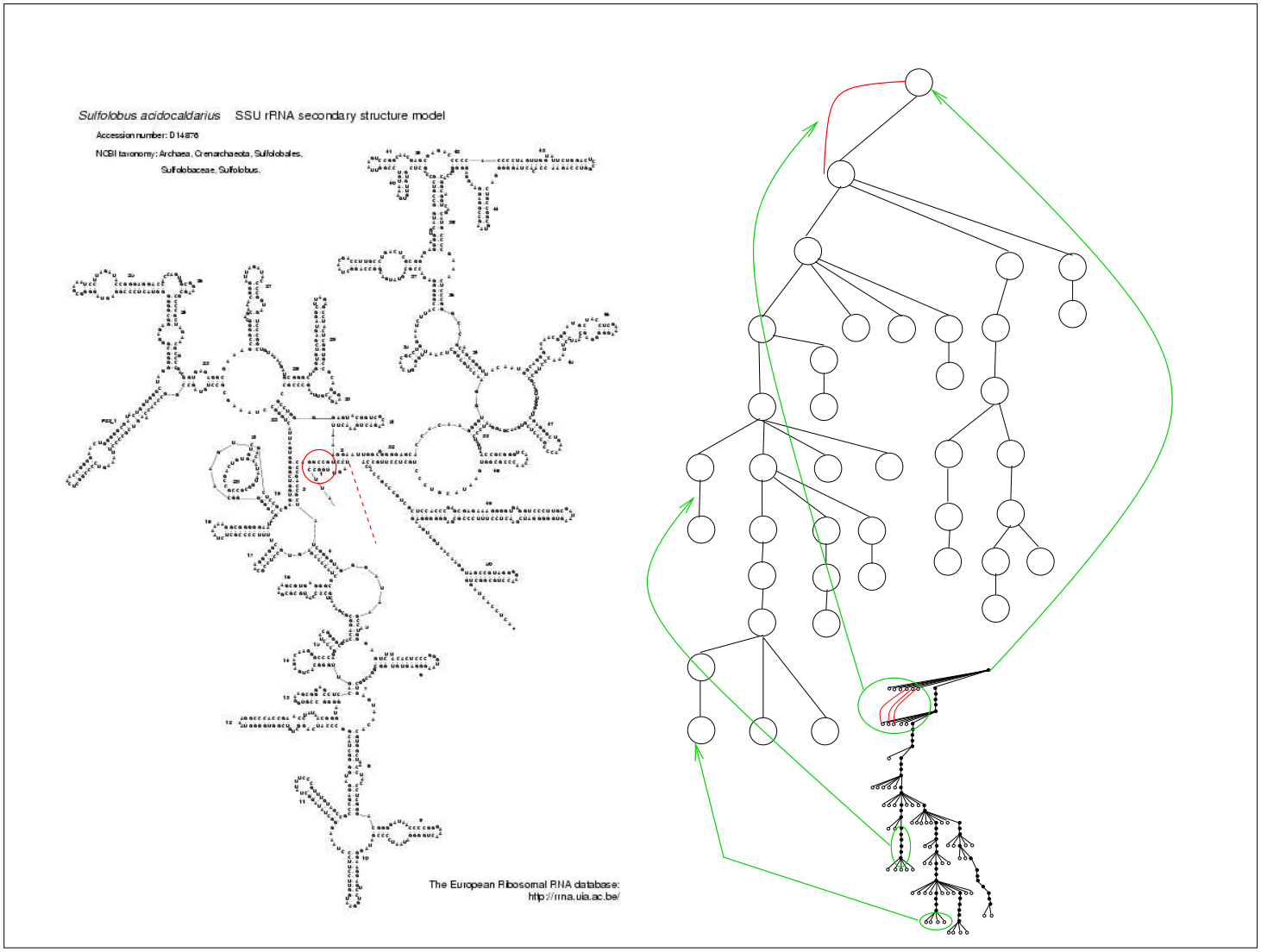
MiGaL: Layer 2



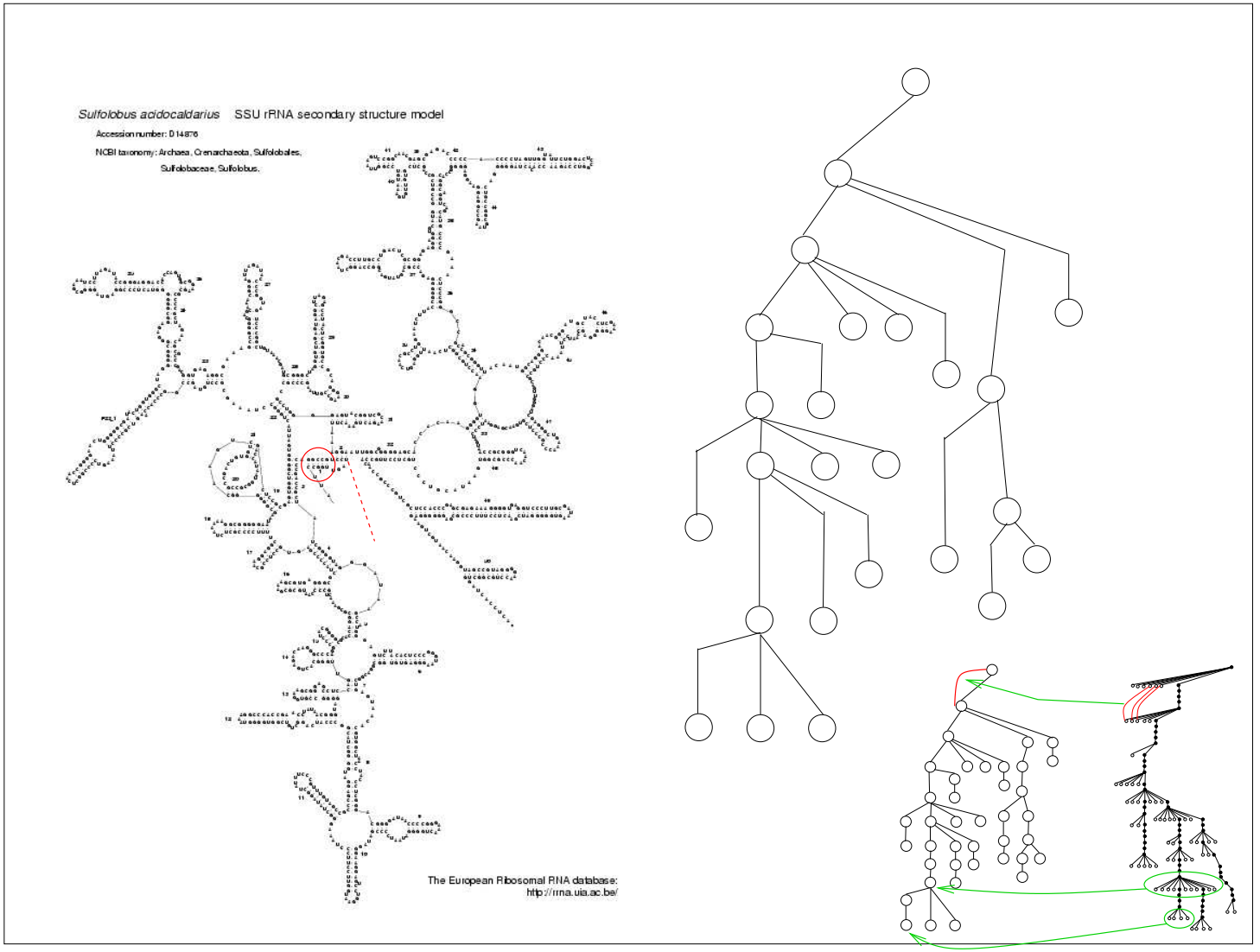
MiGaL: Layer 2



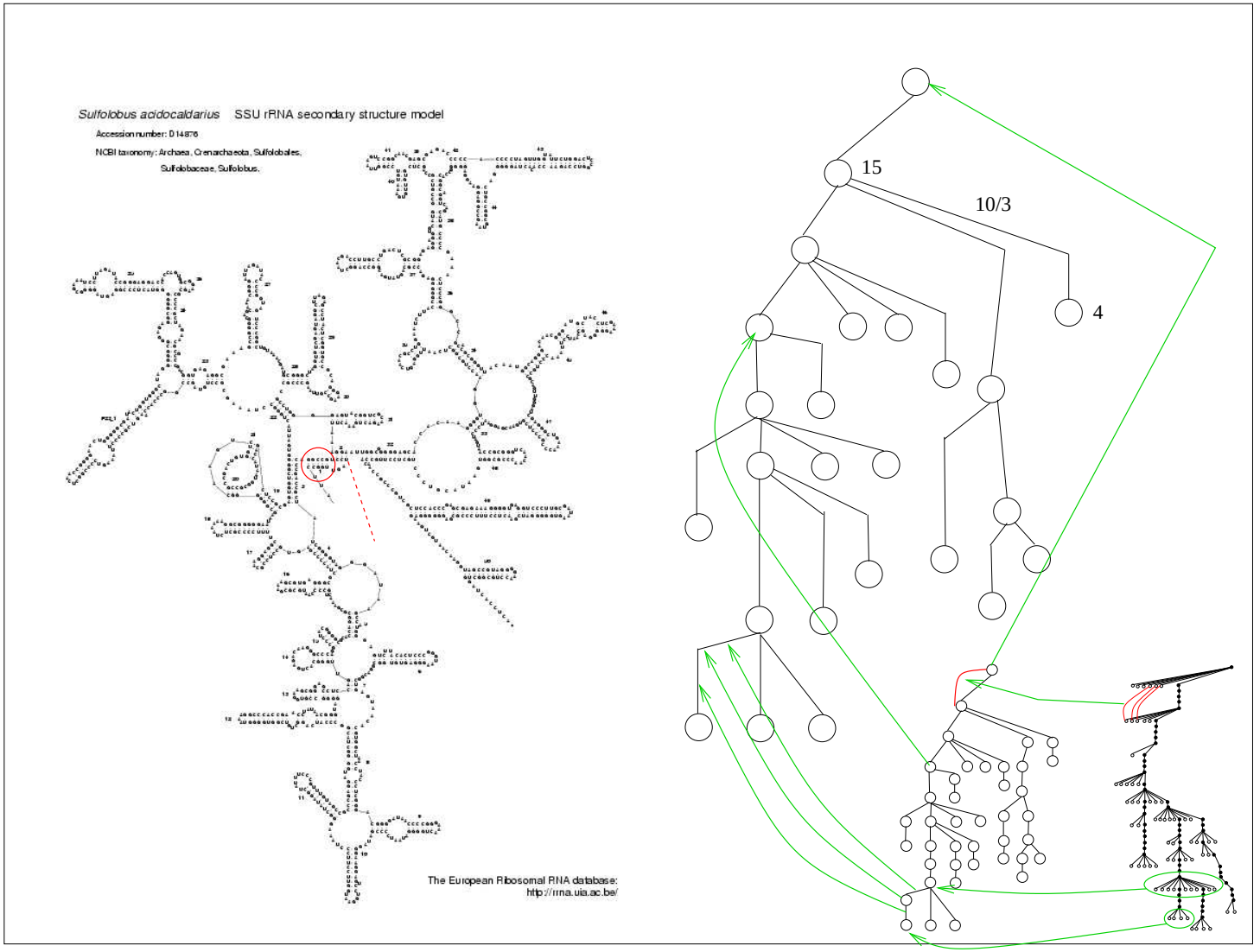
MiGaL: Layer 2



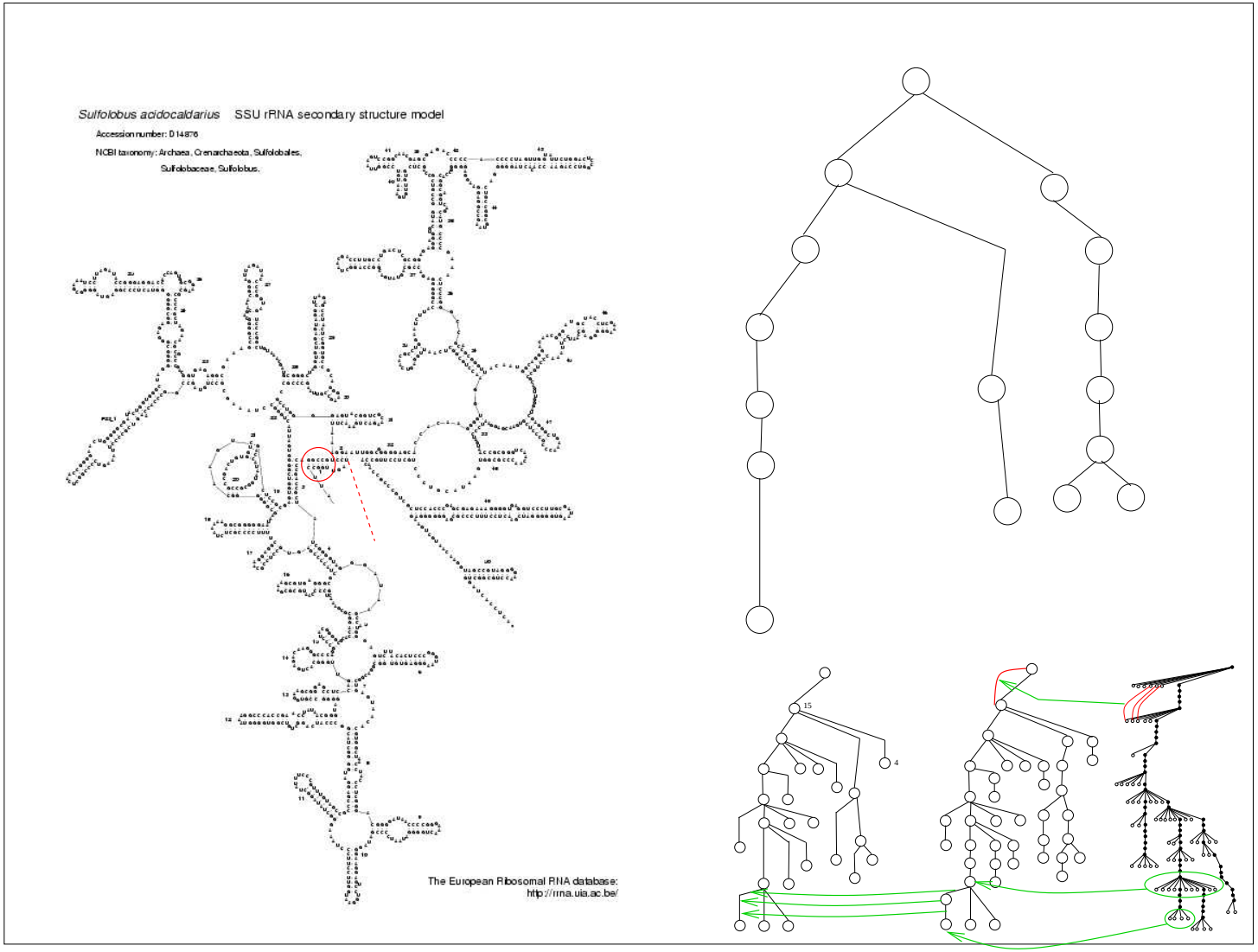
MiGaL: Layer 1



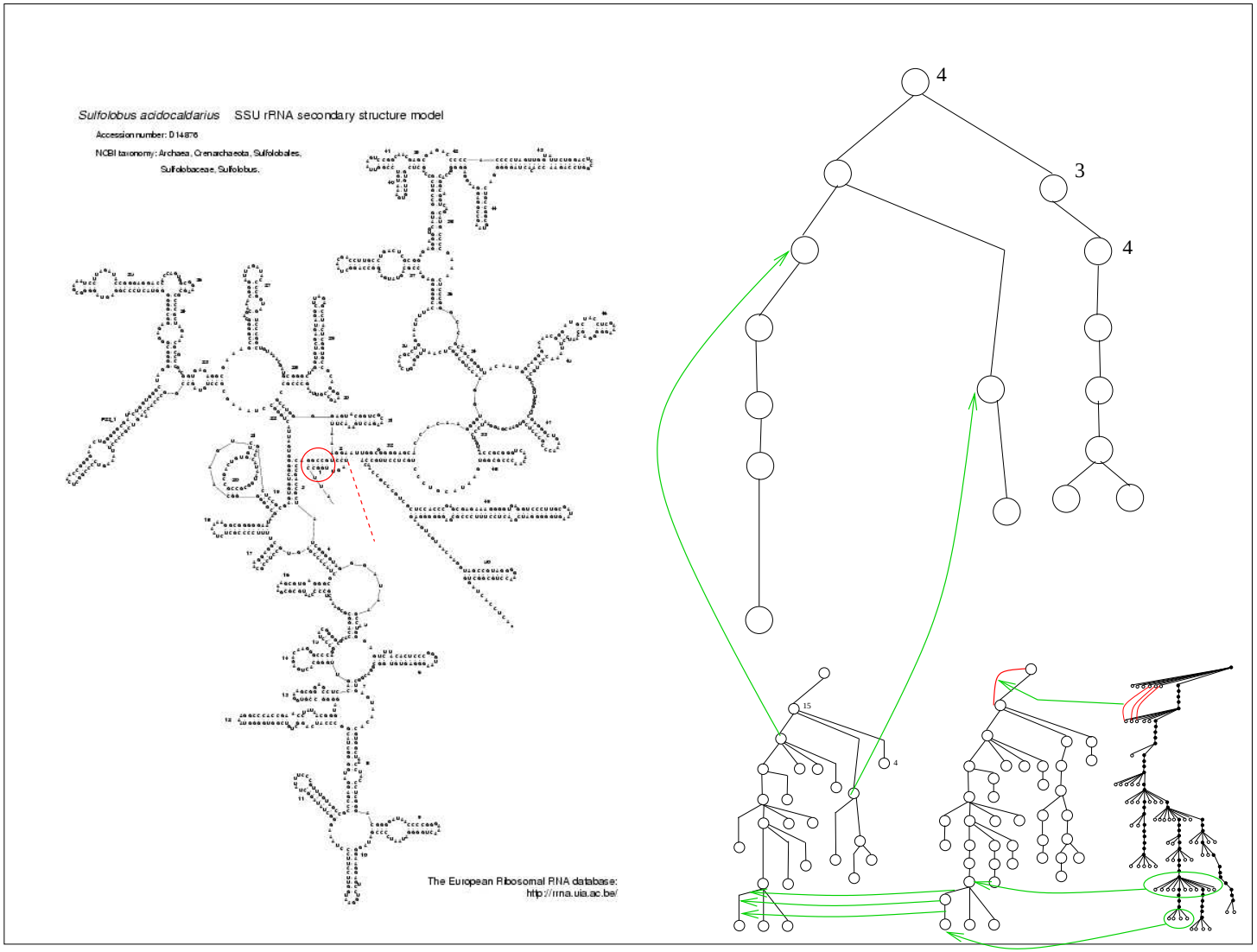
MiGaL: Layer 1



MiGaL: Layer 0



MiGaL: Layer 0



MiGaL: Summary

Our model takes into account pseudo knots.

The requested amount of memory is linear in the size of the RNA.

MiGaL can represent either a model for various RNAs or a single RNA.

MiGaL: Summary

Our model takes into account pseudo knots.

The requested amount of memory is linear in the size of the RNA.

MiGaL can represent either a model for various RNAs or a single RNA.

We now need an algorithm to compare (two) MiGaLs

MiGaL: Summary

Our model takes into account pseudo knots.

The requested amount of memory is linear in the size of the RNA.

MiGaL can represent either a model for various RNAs or a single RNA.

We now need an algorithm to compare (two) MiGaLs

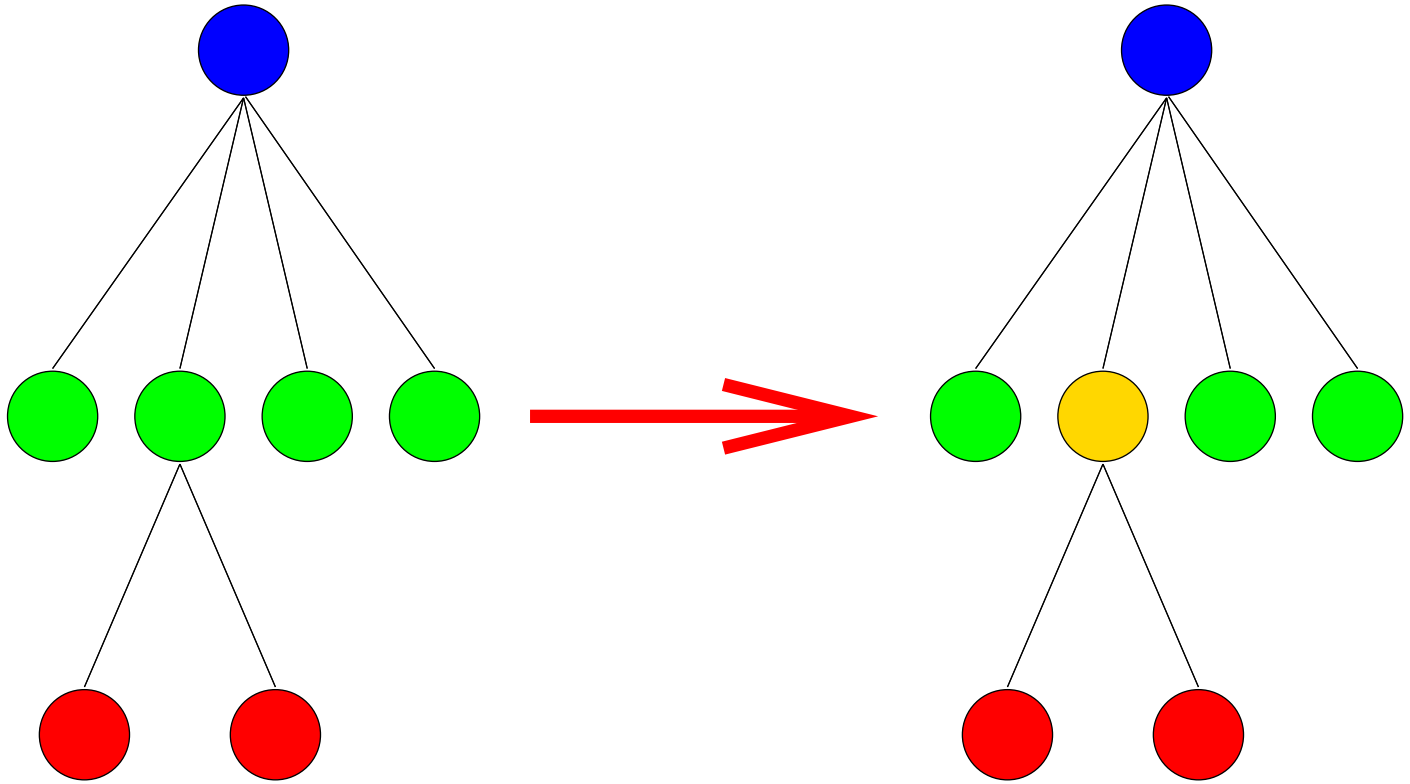
We now need an algorithm to compare (two) Layer0s

MiGaL: Fusion Algorithm

An *Edit-like* algorithm based on our observations concerning RNA multiloop network and adapted to structural tree matching (level 0,1 and 2).

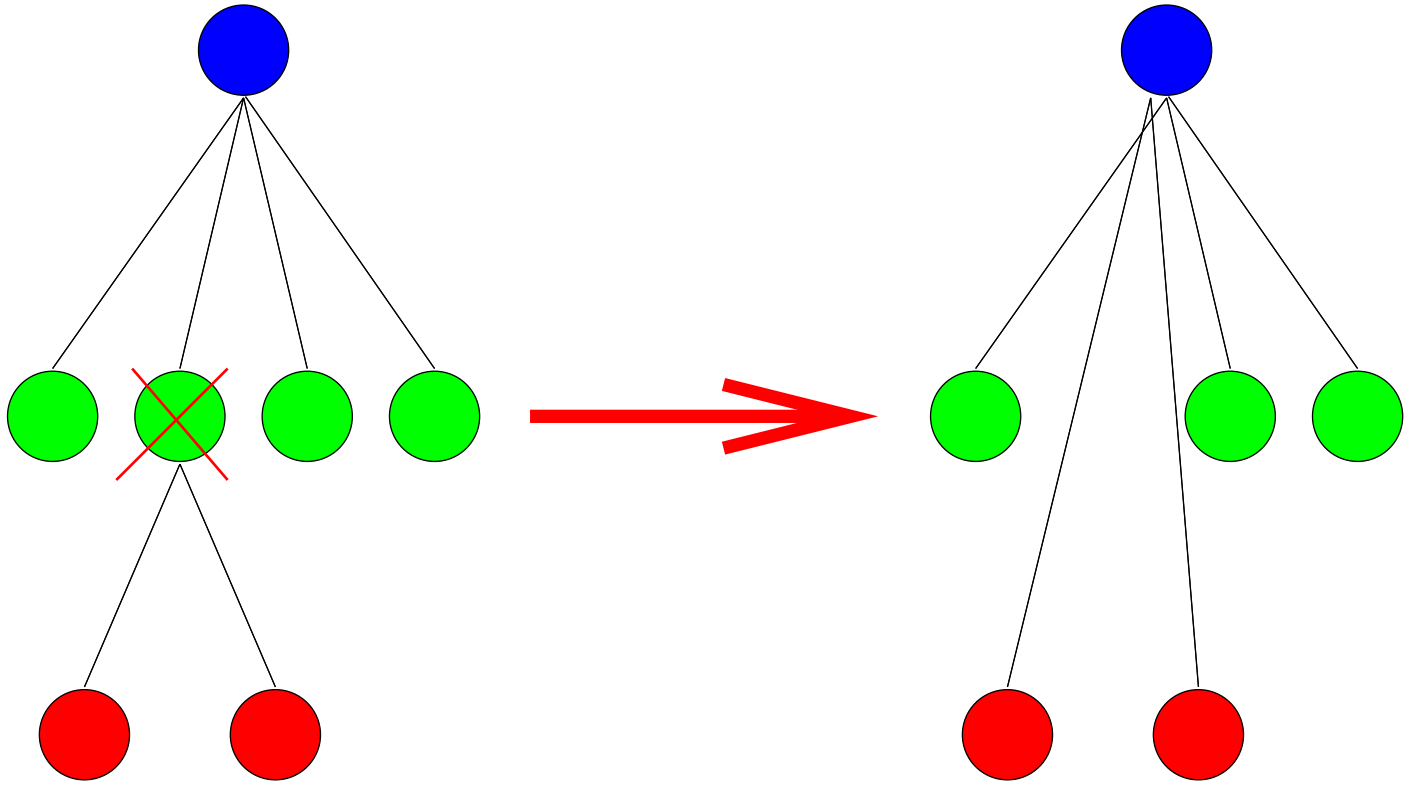
MiGaL: Classical edit operations

An edit algorithm is based on three operations:



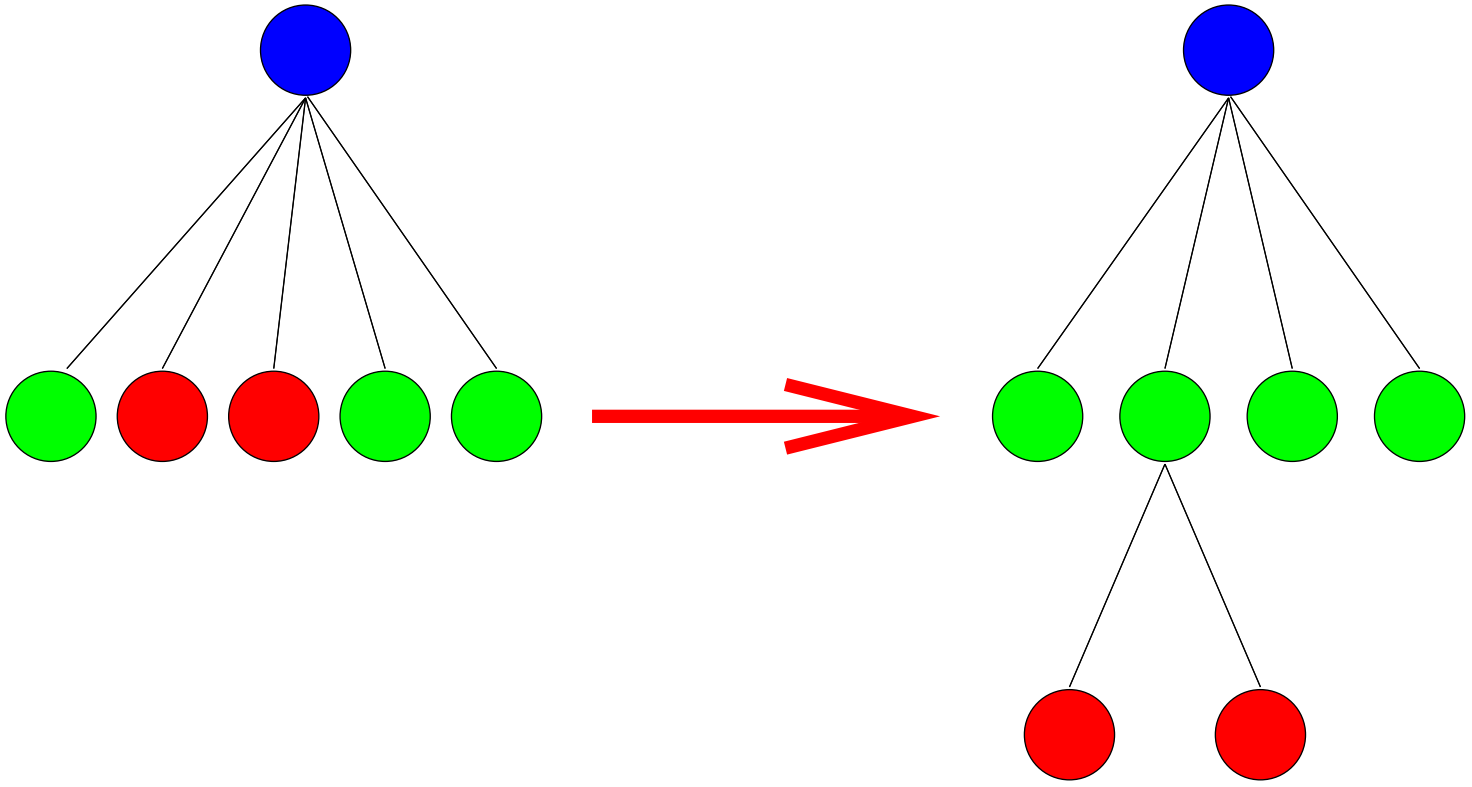
MiGaL: Classical edit operations

An edit algorithm is based on three operations:

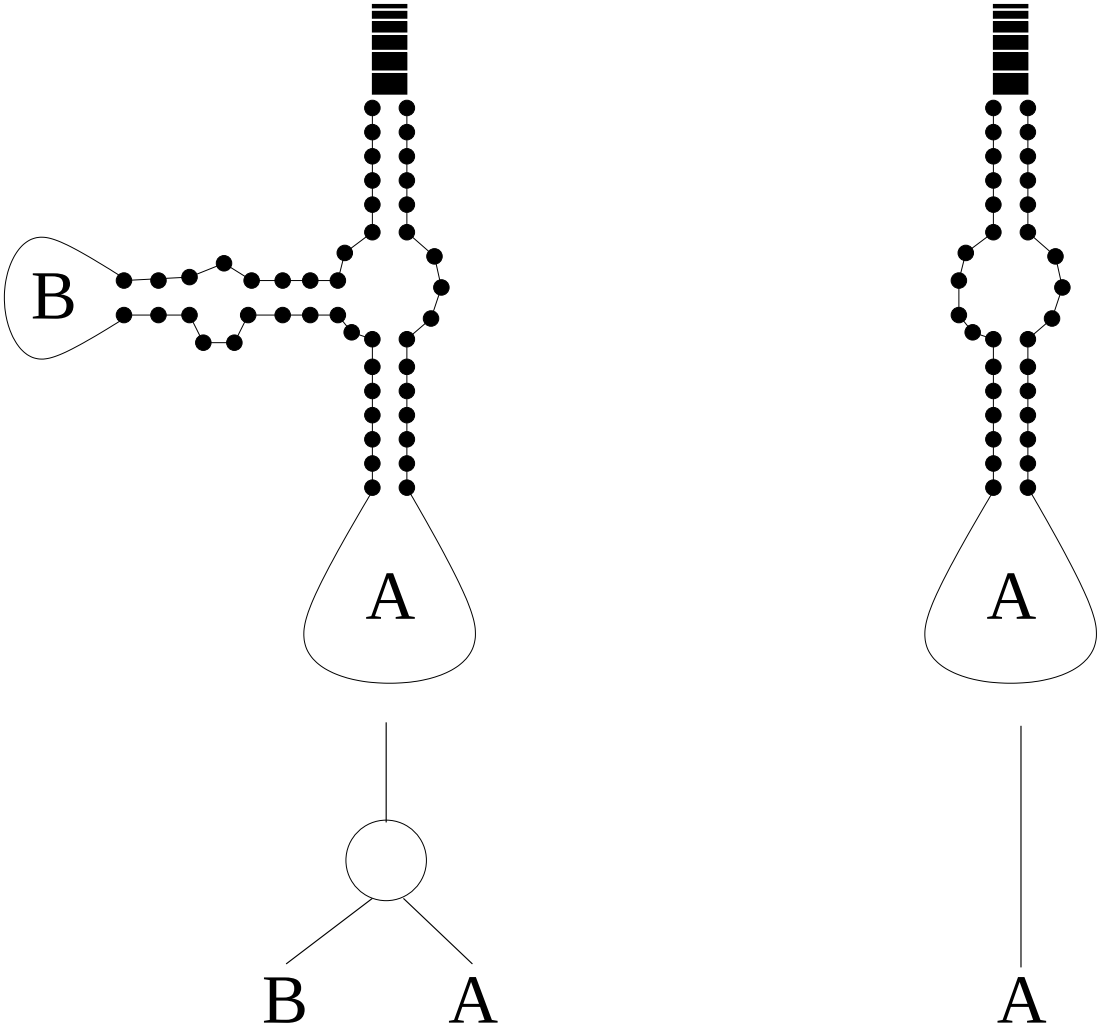


MiGaL: Classical edit operations

An edit algorithm is based on three operations:

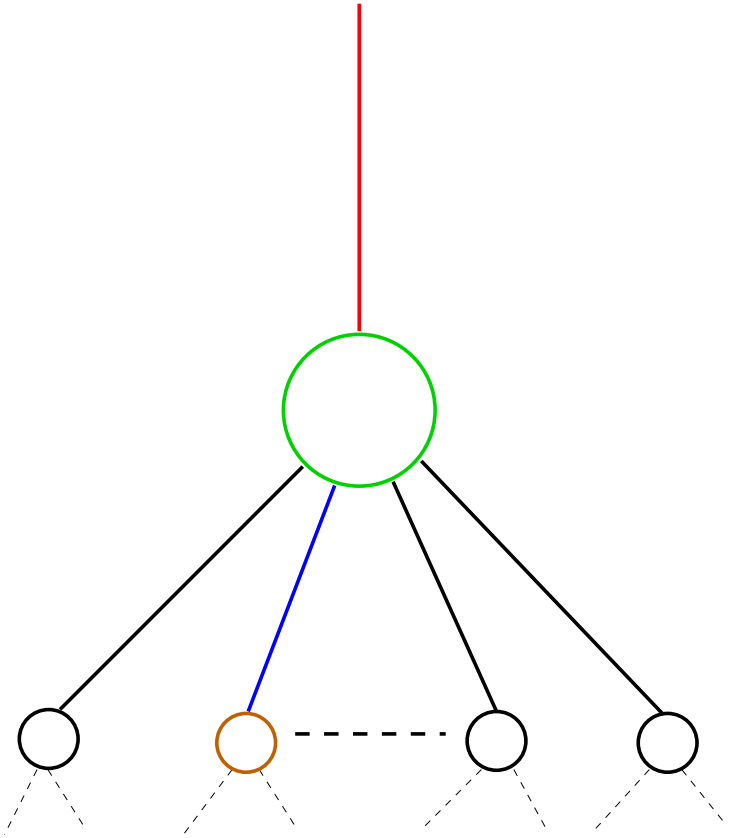


MiGaL: observation 1



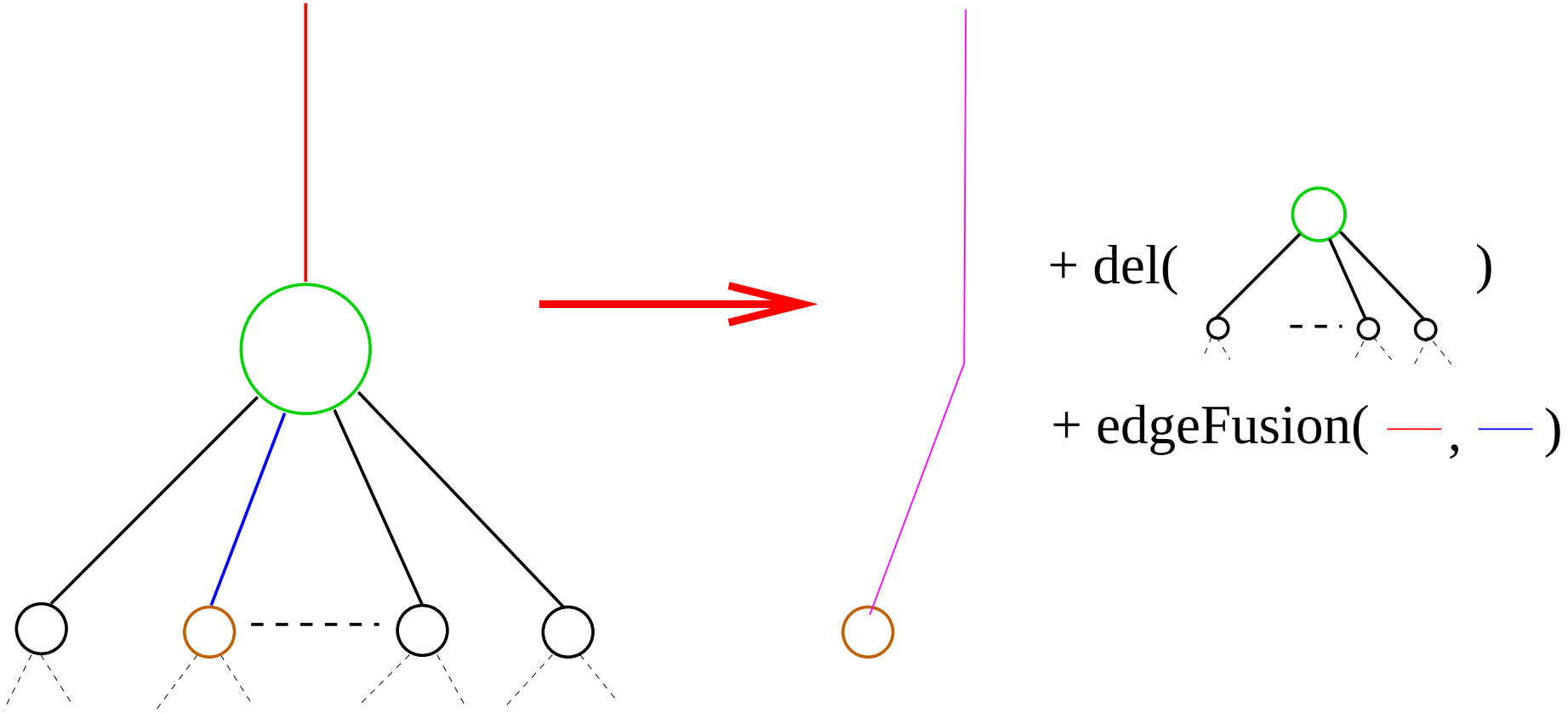
MiGaL: observation 1

Formally:

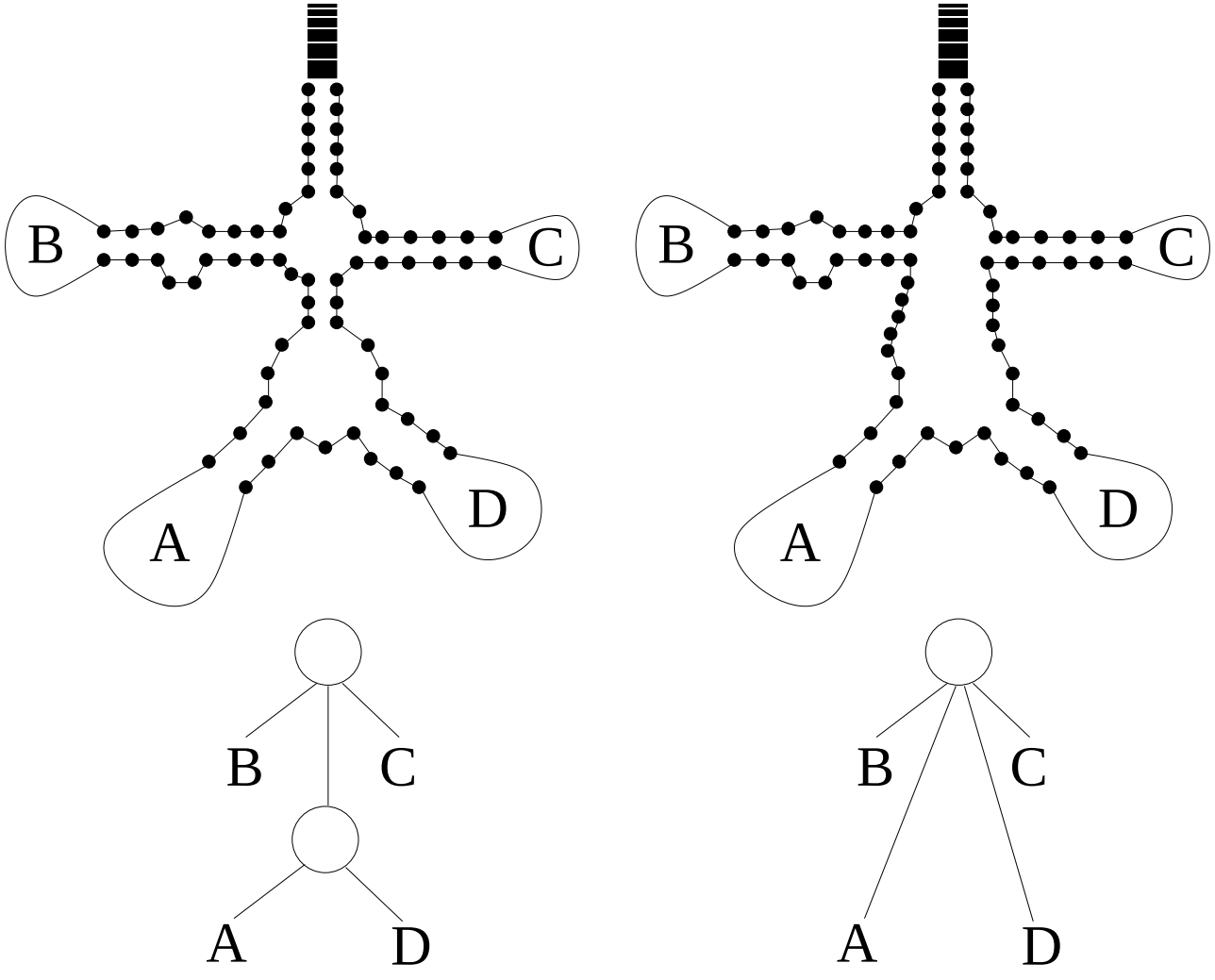


MiGaL: observation 1

Formally:

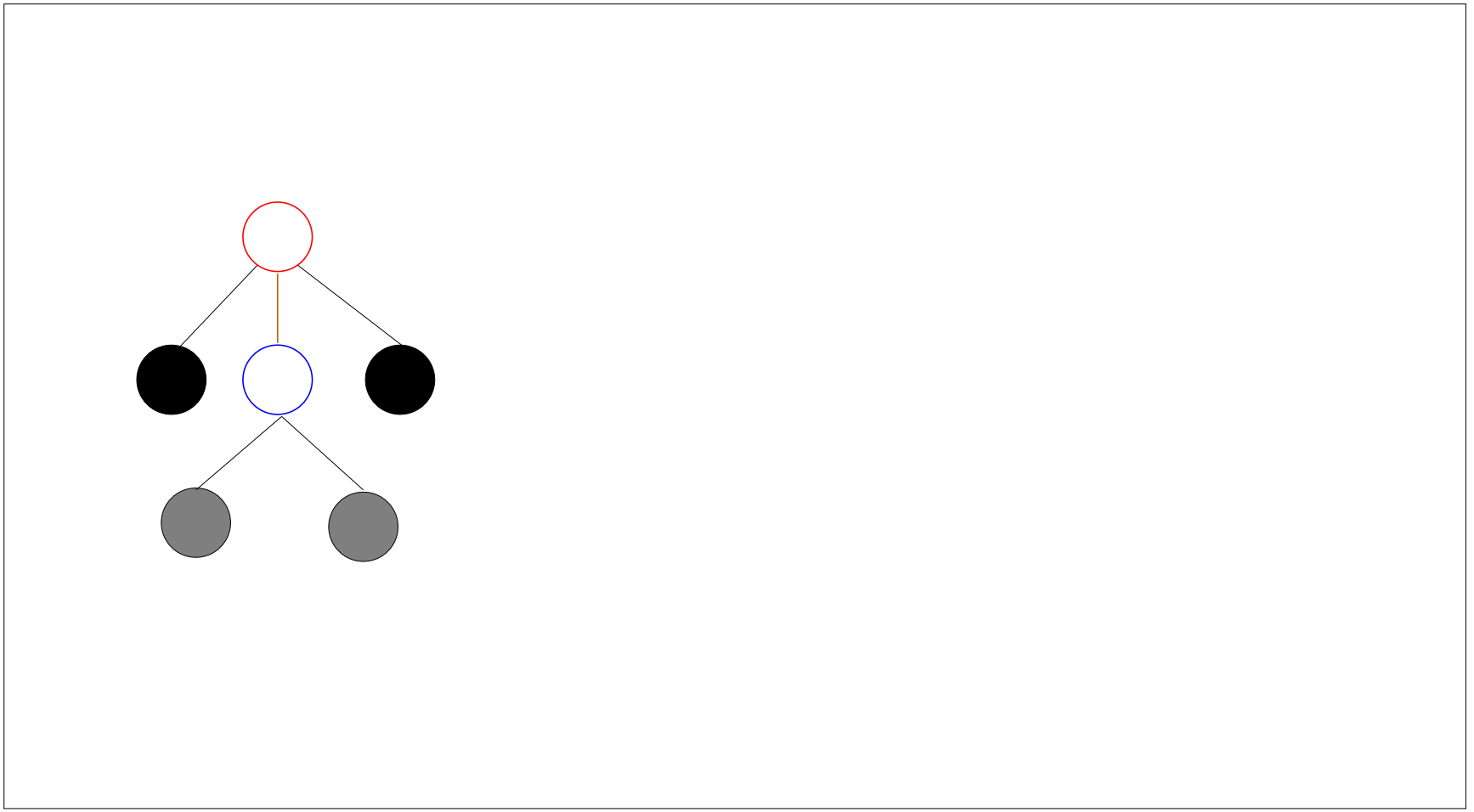


MiGaL: observation 2



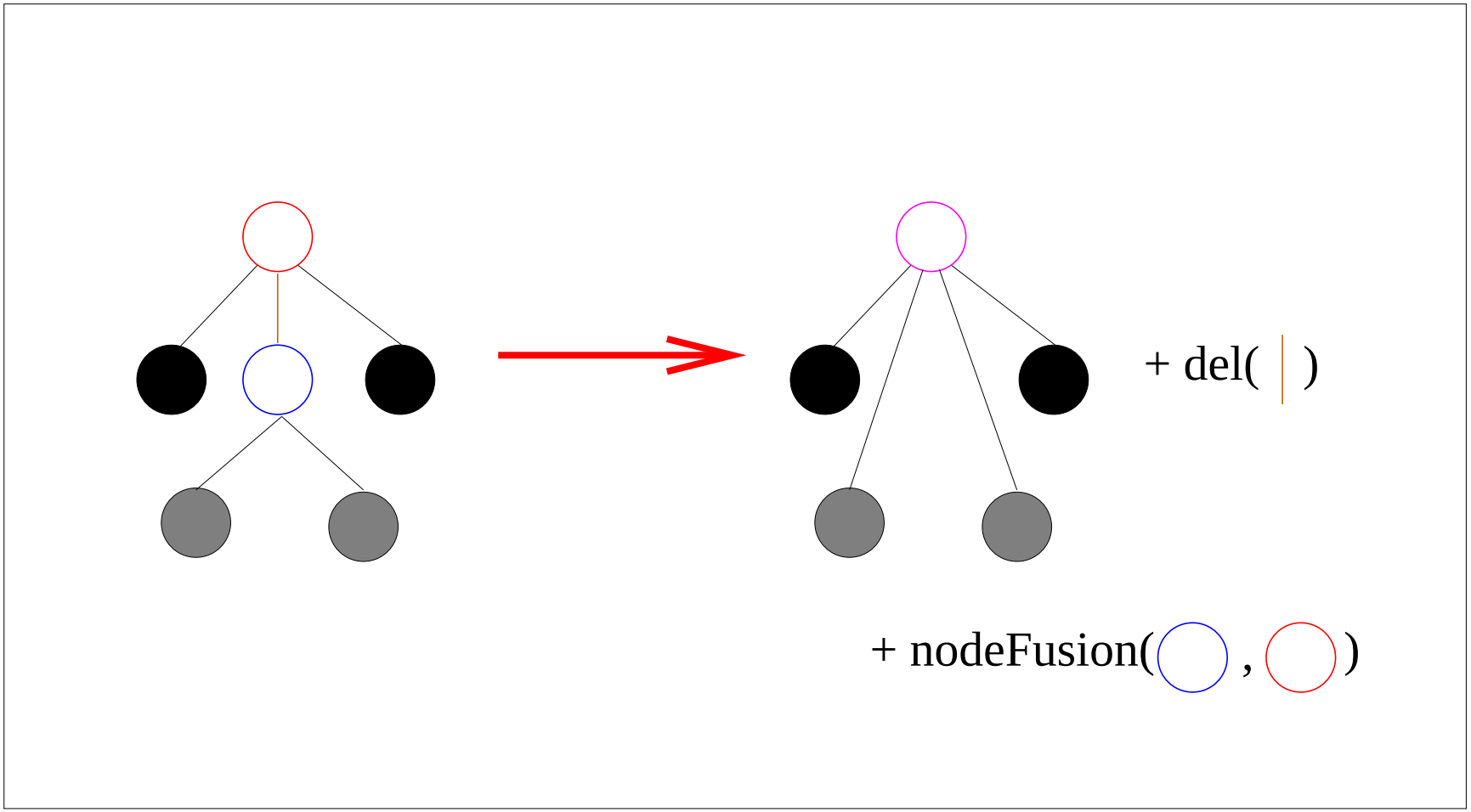
MiGaL: observation 2

Formally:



MiGaL: observation 2

Formally:



MiGaL: fusion

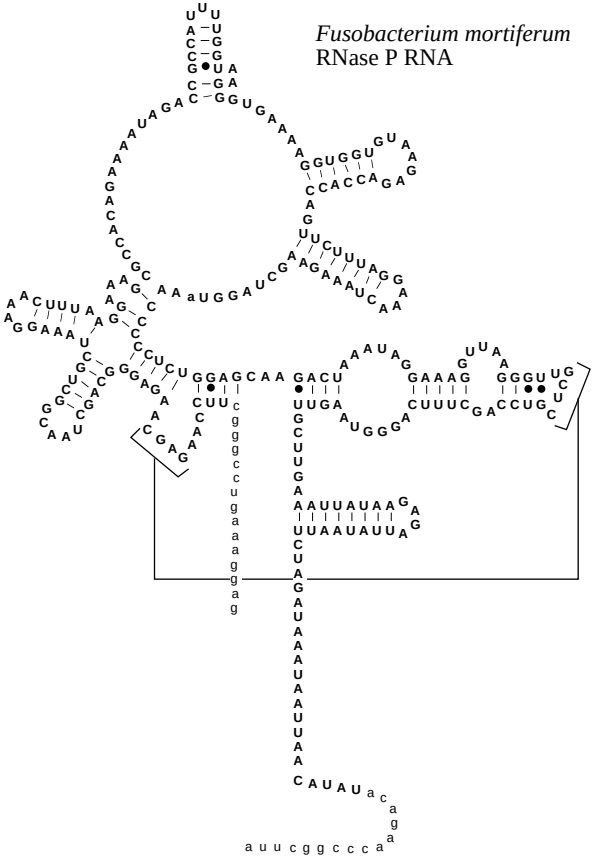
In a “classical” editing algorithm we have:

- Relabeling
- Deletion
- Insertion

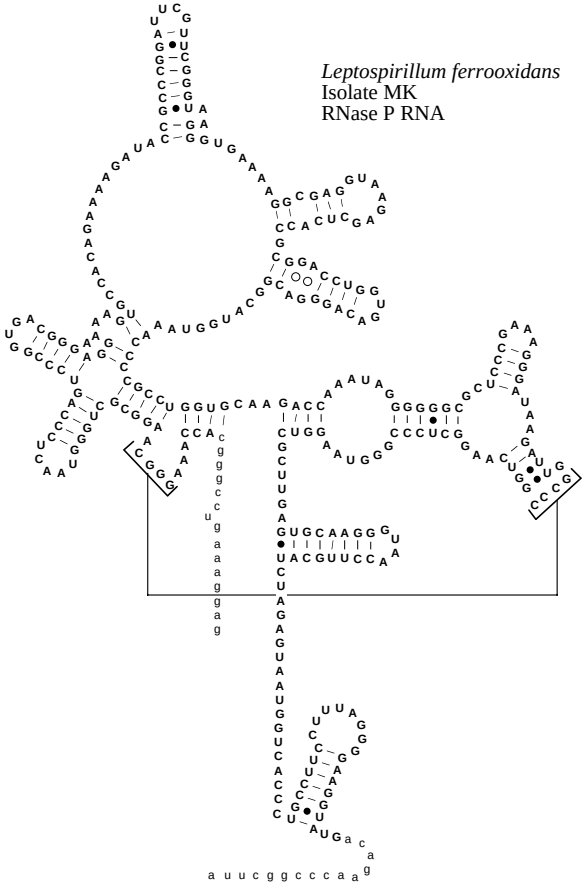
In a “fusion” algorithm we have:

- Relabeling
- Node fusion
- Edge fusion
- Deletion/Insertion

MiGaL: result sample



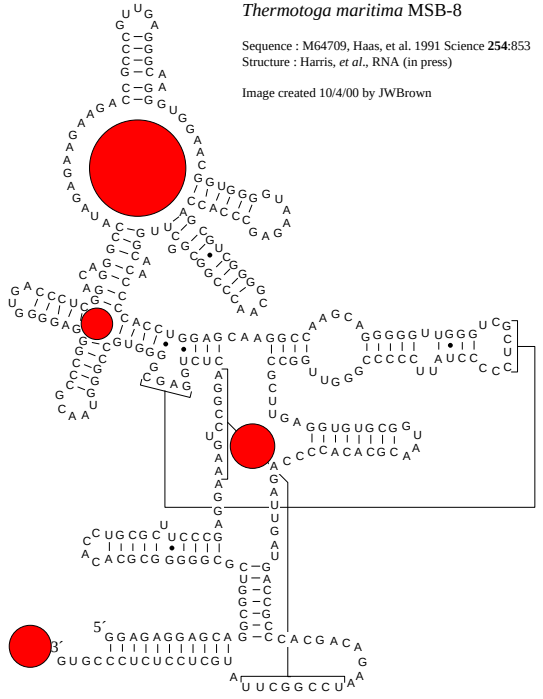
Fusobacterium mortiferum
RNase P RNA

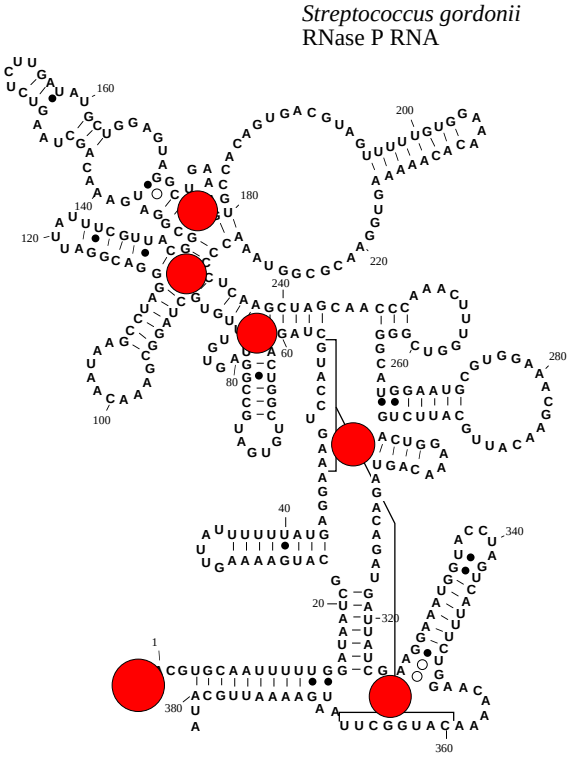


Leptospirillum ferrooxidans
Isolate MK
RNase P RNA

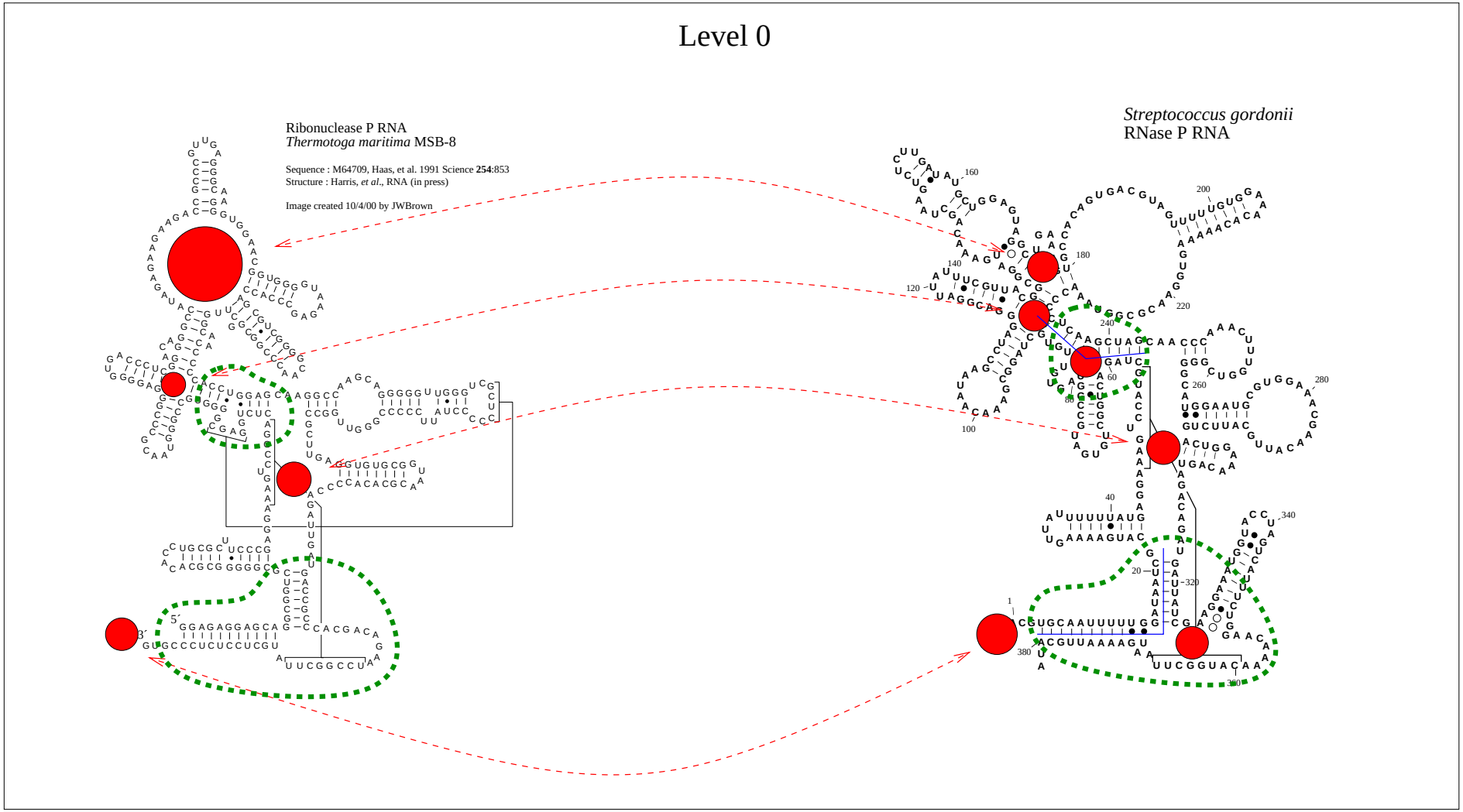
MiGaL: result sample

Level 0



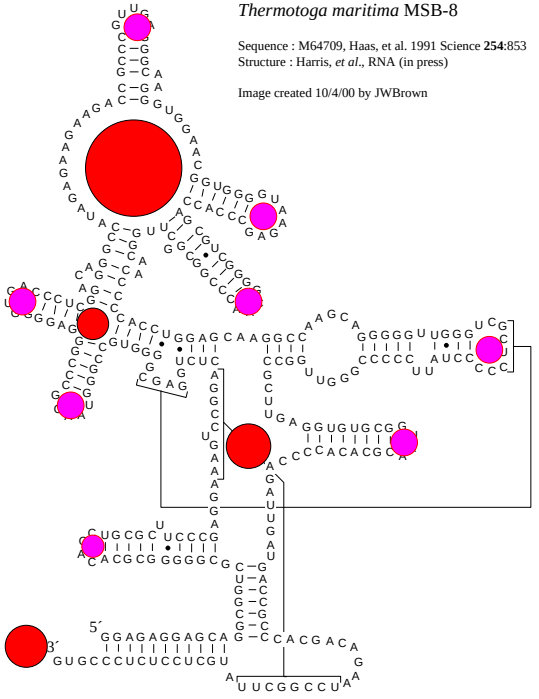


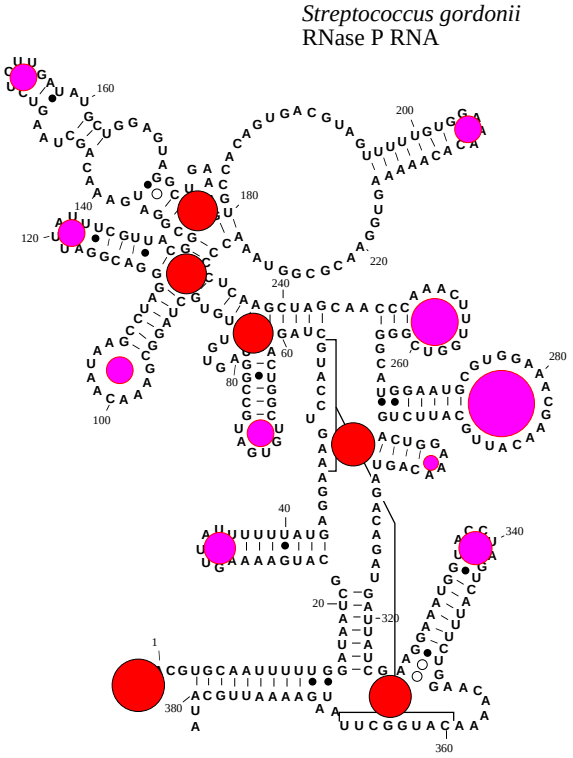
MiGaL: result sample



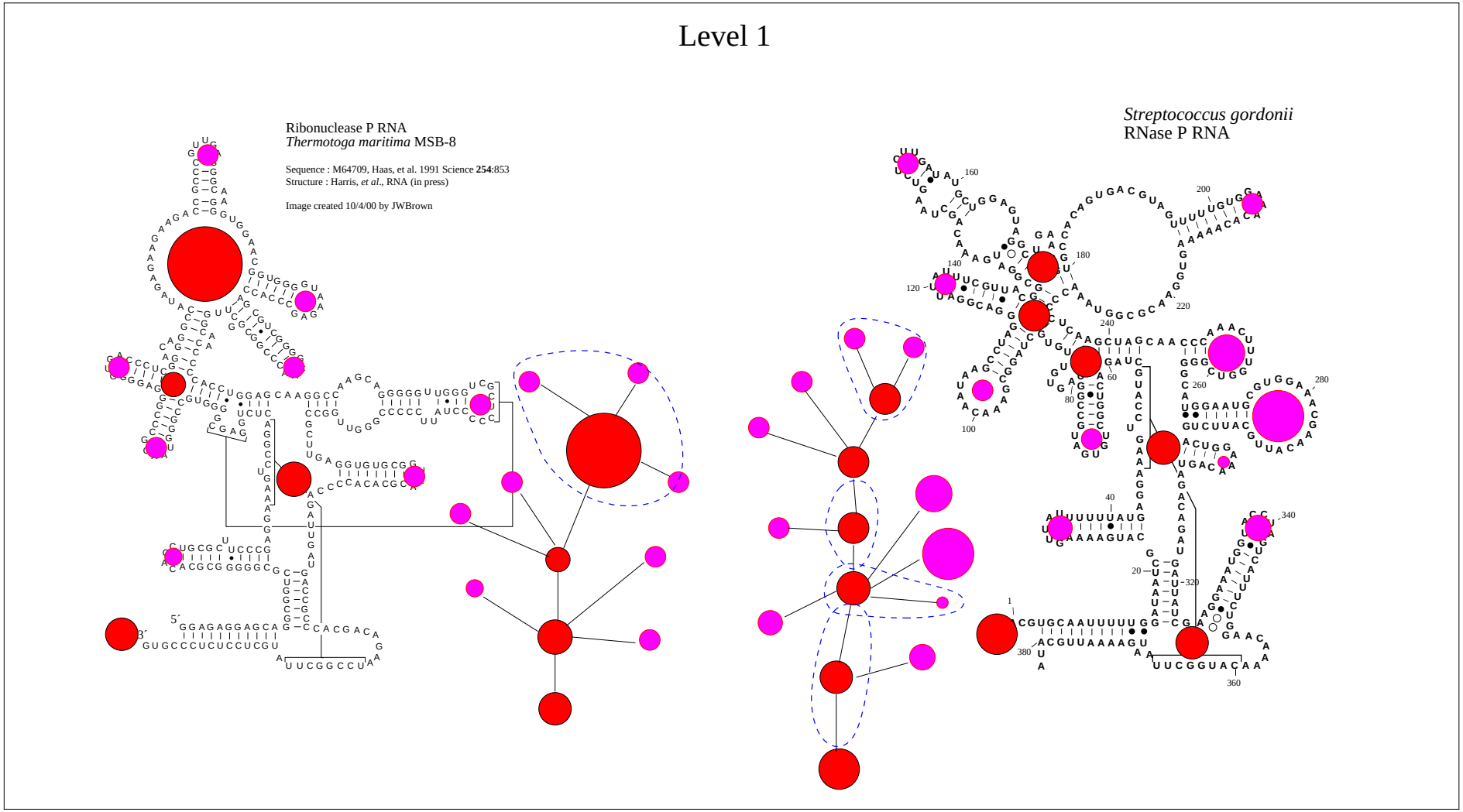
MiGaL: result sample

Level 1



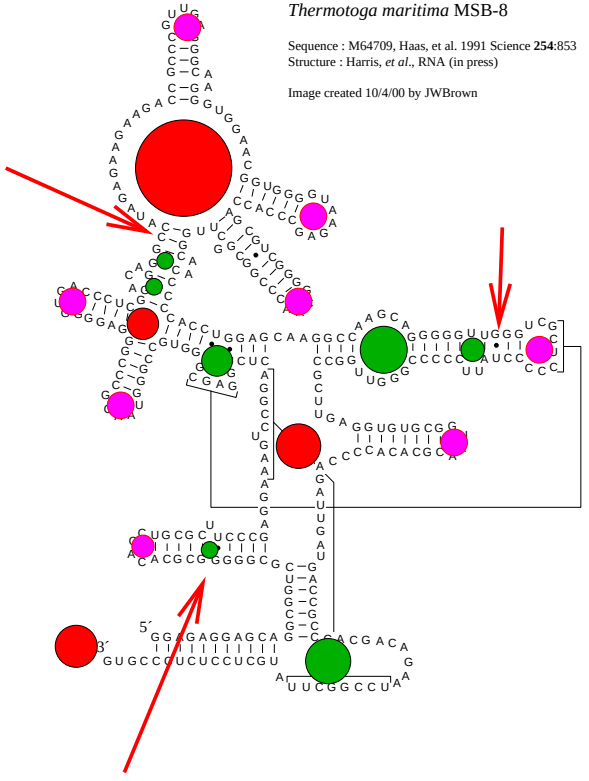


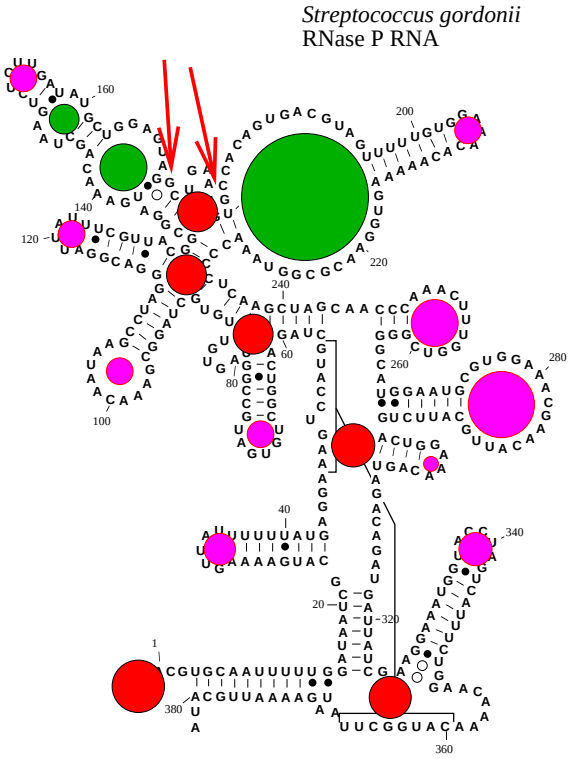
MiGaL: result sample



MiGaL: result sample

Level 2





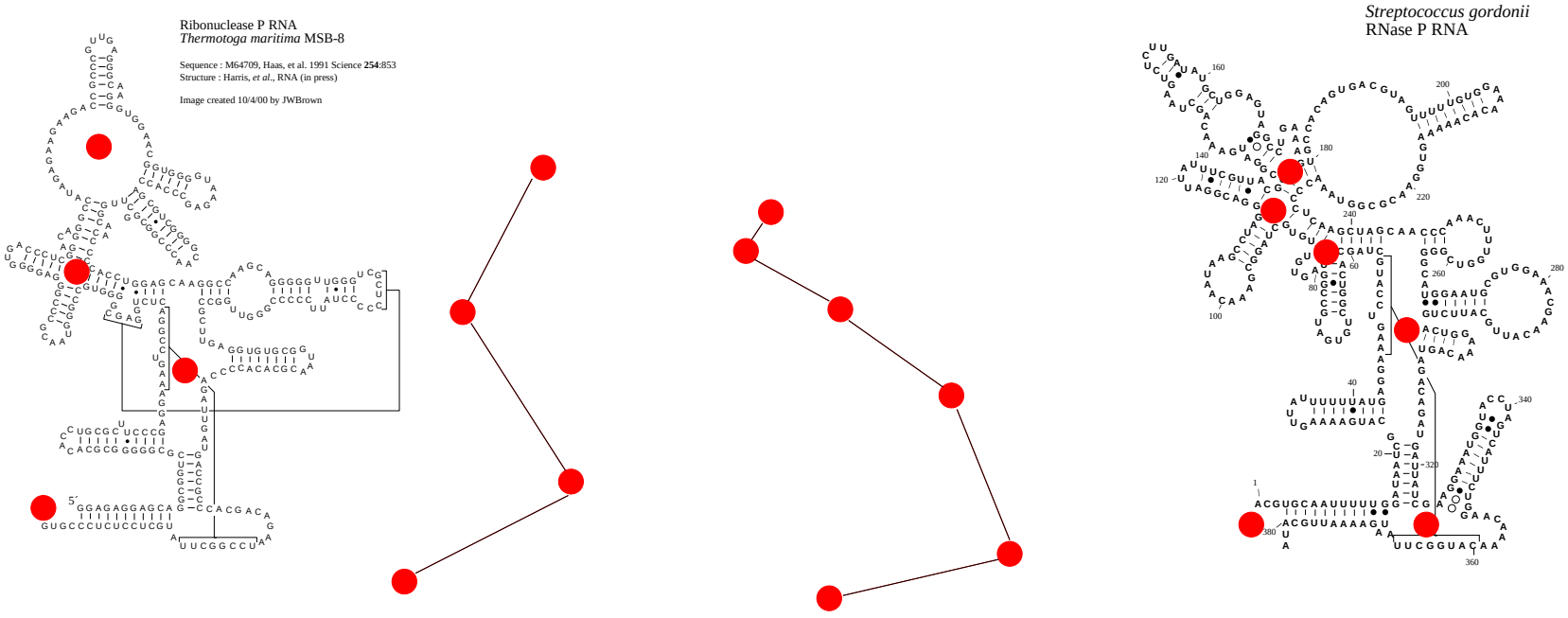
MiGaL: Algorithm

This *new edition algorithm is better to do a zoning than a exact matching.*
Currently we are searching for:

- *The exact complexity of this algorithm (probably* $O(n^k * n^2)$*).*
- *Score function for our edit operations.*

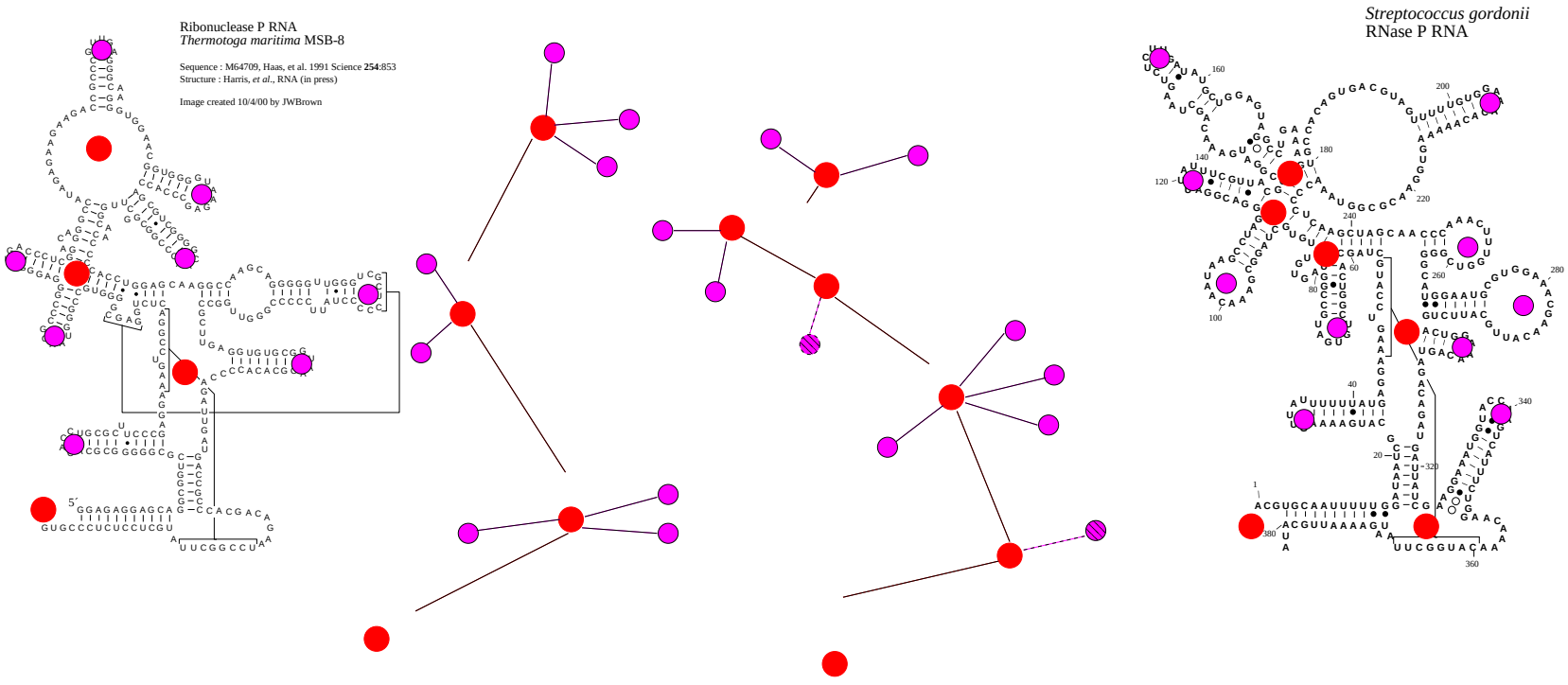
MiGaL: Algorithm

We apply this algorithm to make a top-down matching across the differents MiGaL layers:



MiGaL: Algorithm

We apply this algorithm to make a top-down matching across the differents MiGaL layers:



MiGaL: In practice

The first results seems to show a great improvement in time computation:

- The direct edit algorithm on layer 3 on sample take around 4 minutes.
- The Top-Down algorithm make a few seconds to be compute.

MiGaL: Conclusion

Our approach seems to be original and pertinent in relation to the biological aspects.

The final algorithm should be polynomial in the length of the sequence (we hope so :)

Extension to multiple comparison/inference.