A graph kernel approach to the identification and characterisation of structured non-coding RNAs using multiple sequence alignment information

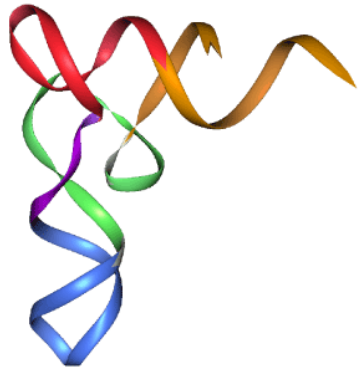
Mariam Alshaikh

Albert Ludwigs University Freiburg,
Department of Computer Science

Feb 18th, 2016

Motivation

Explosion in the discovery of non-identified ncRNAs → efficient automated approaches.
Lack of automated classification tools → done manually.

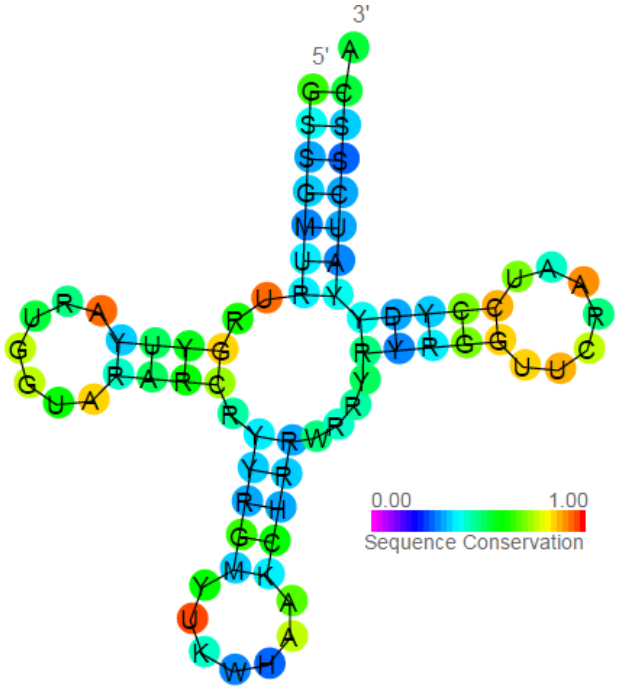


```
SS_cons     <<<<-------..<<<<.____..>>.>>------>>>>
cons        RGYCAUAGnnn--CCnn-GAGU--nn-GRG-nAAGGRCC
conss       3111111344400331101111001101310411111113
```

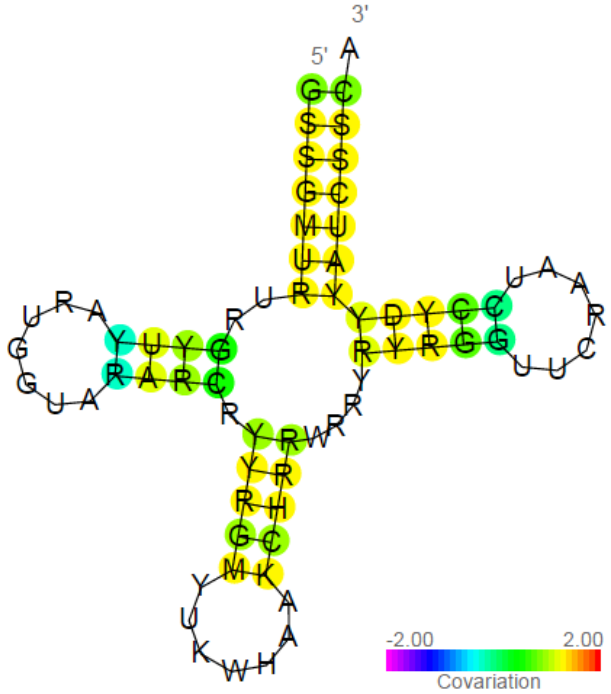
```
SS_cons     <<<<-------..<<<<.____>>>>>>>>...------
cons        RGYCAUAGnnn--CCnn-GAGUnnGRGRCCG----nAAG
conss       3111111344400331101111111311131000041111
```

Approach

Conservation→ important for unpaired region.



Covariation→ important for base pairs.



In this work

- Consider both conservation and covariation.

Multiple Alignment Graph Generator (MAGG)

What is MAGG

- MAGG is a graph encoder tool.
- MAGG can encode the evolutionary conservation of sequences and structures.

Multiple Alignment Graph Generator (MAGG)

What is MAGG

- MAGG is a graph encoder tool.
- MAGG can encode the evolutionary conservation of sequences and structures.

Why MAGG

- Graph formalism $\rightarrow$ flexible encoding.
- Graphs$\rightarrow$ powerful machine learning techniques (graph kernels).

Multiple Alignment Graph Generator (MAGG)

What is MAGG

- MAGG is a graph encoder tool.
- MAGG can encode the evolutionary conservation of sequences and structures.

Why MAGG

- Graph formalism $\rightarrow$ flexible encoding.
- Graphs$\rightarrow$ powerful machine learning techniques (graph kernels).

MAGG aim

- Simulate experts on identifying interesting alignments for further investigation.

Nerest Neighbourhood subgraph pairwise Distance kernal

EDeN

- EDeN→ graph kernel tool.
- EDeN→ Extend the notion of k-mears from string to graphs.
- It counts the fraction of identical pairs of neighborhood sub-graphs.

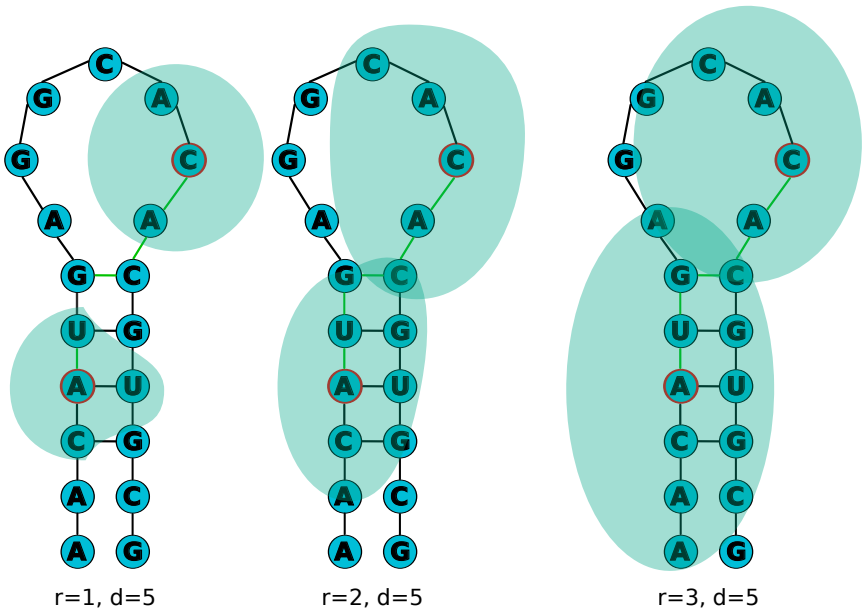


Figure : Pairs of neighbourhood graphs for radius=1,2,3 and distance=5

Information in the input alignment files

1. The alignments are generated using `CMfinder`.
   - CMfinder is an alignment tool, produces sequences that have consensus structure.
2. Every alignment contains information about:

```
2014735323/140471-140548                    ----------------------------CUGGGC..CC..GG..UaAG.....
MA40A_contig29522/419-511                   GCGGUGAAAGUCCGC.U...........GUGGGC..UU..G.A.U.AGU....
MA40A_contig18176/358-450                   GCGGUGAAAGUCCGC.U...........GUGGGC..UU..GG..U.AGU....
NC_010320.1/1363870-1363962                 GCGGUGAAAGUCCGC.U...........GUGGGC..UU..GG..U.AGU....
GBANfinal_contig08346/28-124                GGGGUGAAAGUCCCG.A...........GUAUGG..CC..UGG.U.AGC....
FGTW_contig06637/464-556                    GAGGUGGAAGUCCUC.U...........AUCGGC..CC..G.U.C.AGG....
SRS016989_Baylor_scaffold_17397/3825-3770   ----------------A.......................G..C...C....
NZ_ACNC01000013.1/60964-61057               GUGGUGAAAGUCCAC.U...........GUGGGG..GUa.CG..C.A.U..cg
#=GC SS_cons                                :<<<_______>>>,.,...........<<<<-<..<<..___._.____....
#=GC cons                                   nnGGUGnAAGUCCnn-n------------nYRnRn--nY--nnn-Y-AnY----
#=GC conss                                  43333333333334020000000000033333300330034402033300000
#=GC col_entropy_0                          12000010000002111000000000001112110011101110100110001
#=GC col_entropy_1                          ..................................................
#=GC col_entropy_2                          41897988688891507000000000006780451197009294157061263
#=GC col_entropy_3                          94521954891114866999999999997615051191096011570733340
#=GC cov_SS_cons                            .222.......222.............2222.2..2?..............
```

Information in the input alignment files

- The alignments are generated using `CMfinder`.
- Information contained in alignment files:
  1. Secondary structure prediction.

```
2014735323/140471-140548                  ----------------------------CUGGGC..CC..GG..UaAG.....
MA40A_contig29522/419-511                 GCGGUGAAAGUCCGC.U...........GUGGGC..UU..G.A.U.AGU....
MA40A_contig18176/358-450                 GCGGUGAAAGUCCGC.U...........GUGGGC..UU..GG..U.AGU....
NC_010320.1/1363870-1363962               GCGGUGAAAGUCCGC.U...........GUGGGC..UU..GG..U.AGU....
GBANfinal_contig08346/28-124              GGGGUGAAAGUCCCG.A...........GUAUGG..CC..UGG.U.AGC....
FGTW_contig06637/464-556                  GAGGUGGAAGUCCUC.U...........AUCGGC..CC..G.U.C.AGG....
SRS016989_Baylor_scaffold_17397/3825-3770 ---------------A.........................G..C...C....
NZ_ACNC01000013.1/60964-61057             GUGGUGAAAGUCCAC.U...........GUGGGG..GUa.CG..C.A.U..cg
#=GC SS_cons                              :<<<       >>>,.,...........<<<<-<..<<..   . .   ....
#=GC cons                                 nnGGUGnAAGUCCnn-n-----------nYRnRn--nY--nnn-Y-AnY----
#=GC conss                                43333333333334020000000000033333003300344020333000O
#=GC col_entropy_0                        12000010000002111000000000001112110011101110100110001
#=GC col_entropy_1                        ..................................................
#=GC col_entropy_2                        41897988688891507000000000006780451197009294157061263
#=GC col_entropy_3                        94521954891114866999999999997615051191096011570733340
#=GC cov_SS_cons                          .222.......222.............2222.2..2?..............
```

Information in the input alignment files

- The alignments are generated using `CMfinder`.
- Information contained in alignment files:
  1. Secondary structure prediction.
  2. Nucleotides conservation.

```
2014735323/140471-140548                    ----------------------------CUGGGC..CC..GG..UaAG.....
MA40A_contig29522/419-511                   GCGGUGAAAGUCCGC.U...........GUGGGC..UU..G.A.U.AGU....
MA40A_contig18176/358-450                   GCGGUGAAAGUCCGC.U...........GUGGGC..UU..GG..U.AGU....
NC_010320.1/1363870-1363962                 GCGGUGAAAGUCCGC.U...........GUGGGC..UU..GG..U.AGU....
GBANfinal_contig08346/28-124                GGGGUGAAAGUCCCG.A...........GUAUGG..CC..UGG.U.AGC....
FGTW_contig06637/464-556                    GAGGUGGAAGUCCUC.U...........AUCGGC..CC..G.U.C.AGG....
SRS016989_Baylor_scaffold_17397/3825-3770   ----------------A.........................G..C...C....
NZ_ACNC01000013.1/60964-61057               GUGGUGAAAGUCCAC.U...........GUGGGG..GUa.CG..C.A.U..cg
#=GC SS_cons                                :<<<_____>>>,.,...........<<<<-<..<<..___._.___....
#=GC cons                                   nnGGUGnAAGUCCnn-n----------nYRnRn--nY--nnn-Y-AnY----
#=GC conss                                  43333333333334020000000000033333300330034402033300000
#=GC col_entropy_0                          12000010000002111000000000001112110011101110100110001
#=GC col_entropy_1                          ..................................................
#=GC col_entropy_2                          41897988688891507000000000006780451197009294157061263
#=GC col_entropy_3                          94521954891114866999999999999761505119109601157073334 0
#=GC cov_SS_cons                            .222.......222.............2222.2..2?..............
```

Information in the input alignment files

- The alignments are generated using `CMfinder`.
- Information contained in alignment files:
  1. Secondary structure prediction.
  2. Nucleotides conservation.
  3. Strength of conservation.

```
2014735323/140471-140548                  ----------------------------CUGGGC..CC..GG..UaAG.....
MA40A_contig29522/419-511                 GCGGUGAAAGUCCGC.U...........GUGGGC..UU..G.A.U.AGU....
MA40A_contig18176/358-450                 GCGGUGAAAGUCCGC.U...........GUGGGC..UU..GG..U.AGU....
NC_010320.1/1363870-1363962               GCGGUGAAAGUCCGC.U...........GUGGGC..UU..GG..U.AGU....
GBANfinal_contig08346/28-124              GGGGUGAAAGUCCCG.A...........GUAUGG..CC..UGG.U.AGC....
FGTW_contig06637/464-556                  GAGGUGGAAGUCCUC.U...........AUCGGC..CC..G.U.C.AGG....
SRS016989_Baylor_scaffold_17397/3825-3770 ----------------A.......................G..C...C....
NZ_ACNC01000013.1/60964-61057             GUGGUGAAAGUCCAC.U...........GUGGGG..GUa.CG..C.A.U..cg.
#=GC SS_cons                              :<<<_______>>>,.,...........<<<<-<..<<..___._.____....
#=GC cons                                 nnGGUGnAAGUCCnn-n------------nYRnRn--nY--nnn-Y-AnY----
#=GC conss                                433333333333340200000000000333333003300344020333000
#=GC col_entropy_0                        120000100000021110000000000011121100111011101001100 01
#=GC col_entropy_1                        ..................................................
#=GC col_entropy_2                        418979886888915070000000000067804511970092941570612 63
#=GC col_entropy_3                        945219548911148669999999999997615051191096011570733340
#=GC cov_SS_cons                          .222.......222.............2222.2..2?..............
```

Information in the input alignment files

- The alignments are generated using `CMfinder`.
- Information contained in alignment files:
  1. Secondary structure prediction.
  2. Nucleotides conservation.
  3. Strength of conservation.
  4. Entropy of the nucleotides.

```
2014735323/140471-140548                    ----------------------------CUGGGC..CC..GG..UaAG.....
MA40A_contig29522/419-511                   GCGGUGAAAGUCCGC.U...........GUGGGC..UU..G.A.U.AGU....
MA40A_contig18176/358-450                   GCGGUGAAAGUCCGC.U...........GUGGGC..UU..GG..U.AGU....
NC_010320.1/1363870-1363962                 GCGGUGAAAGUCCGC.U...........GUGGGC..UU..GG..U.AGU....
GBANfinal_contig08346/28-124                GGGGUGAAAGUCCCG.A...........GUAUGG..CC..UGG.U.AGC....
FGTW_contig06637/464-556                    GAGGUGGAAGUCCUC.U...........AUCGGC..CC..G.U.C.AGG....
SRS016989_Baylor_scaffold_17397/3825-3770   ----------------A.........................G..C...C....
NZ_ACNC01000013.1/60964-61057               GUGGUGAAAGUCCAC.U...........GUGGGG..GUa.CG..C.A.U..cg.
#=GC SS_cons                                :<<<_______>>>,.,...........<<<<-<..<<..___._.____....
#=GC cons                                   nnGGUGnAAGUCCnn-n------------nYRnRn--nY--nnn-Y-AnY----
#=GC conss                                  43333333333334020000000000003333330033003440203330000
#=GC col_entropy_0                          12000010000002111000000000001112110011101110100110001
#=GC col_entropy_1                          ..................................................
#=GC col_entropy_2                          41897988688891507000000000006780451197009294157061263
#=GC col_entropy_3                          94521954891114866999999999999761505119109601157073334O
#=GC cov_SS_cons                            .222.......222..............2222.2..2?..............
```

Information in the input alignment files

- The alignments are generated using `CMfinder`.
- Information contained in alignment files:
  1. Secondary structure prediction.
  2. Nucleotides conservation.
  3. Strength of conservation.
  4. Entropy of the nucleotides.
  5. Covariation of the secondary structure.

```
2014735323/140471-140548                  ----------------------------CUGGGC..CC..GG..UaAG.....
MA40A_contig29522/419-511                 GCGGUGAAAGUCCGC.U...........GUGGGC..UU..G.A.U.AGU....
MA40A_contig18176/358-450                 GCGGUGAAAGUCCGC.U...........GUGGGC..UU..GG..U.AGU....
NC_010320.1/1363870-1363962               GCGGUGAAAGUCCGC.U...........GUGGGC..UU..GG..U.AGU....
GBANfinal_contig08346/28-124              GGGGUGAAAGUCCCG.A...........GUAUGG..CC..UGG.U.AGC....
FGTW_contig06637/464-556                  GAGGUGGAAGUCCUC.U...........AUCGGC..CC..G.U.C.AGG....
SRS016989_Baylor_scaffold_17397/3825-3770 ----------------A.........................G..C...C....
NZ_ACNC01000013.1/60964-61057             GUGGUGAAAGUCCAC.U...........GUGGGG..GUa.CG..C.A.U..cg
#=GC SS_cons                              :<<<_______>>>,.,...........<<<<-<..<<..___._.___....
#=GC cons                                 nnGGUGnAAGUCCnn-n------------nYRnRn--nY--nnn-Y-AnY----
#=GC conss                                43333333333334020000000000033333300330034402033300000
#=GC col_entropy_0                        12000010000002111000000000001112110011101110100110001
#=GC col_entropy_1                        ..................................................
#=GC col_entropy_2                        41897988688891507000000000006780451197009294157061263
#=GC col_entropy_3                        94521954891114866999999999999761505119109601157073334O
#=GC cov_SS_cons                          .222......222.............2222.2..2?..............
```

Graphs created by MAGG

- MAGG produces two different graph representations.
  1. Node based graphs $\mathscr{N}$.
  2. Summary based graphs $\mathscr{S}$.
- Each representation can encode the information in:
  1. One node: $\mathscr{U}$.
  2. List of nodes: $\mathscr{L}$.

Node based graphs

1. $\mathscr{N}_{\mathscr{U}}$ encodes the information in one node.
2. $\mathscr{N}_{\mathscr{L}}$ encodes the information in set of nodes forming a list.

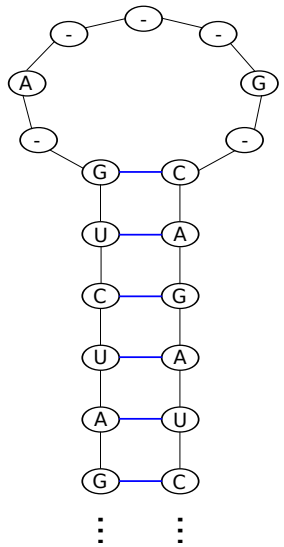


Figure : $\mathscr{N}_{\mathscr{U}}$: Conservation information encoded in single nodes.

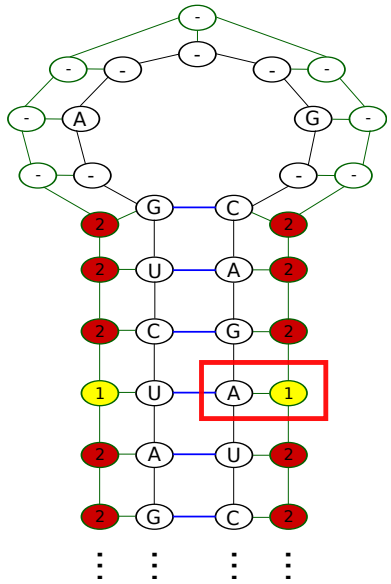


Figure : $\mathscr{N}_{\mathscr{L}}$: Conservation and covariation information in multiple nodes forming a list.

Summary based graphs

1. $\mathscr{S}_{\mathscr{U}}$ same as $\mathscr{N}_{\mathscr{U}}$ but summary information about the structure is encoded.
2. $\mathscr{S}_{\mathscr{L}}$ same as $\mathscr{N}_{\mathscr{L}}$ but more summary information about the structure is encoded.

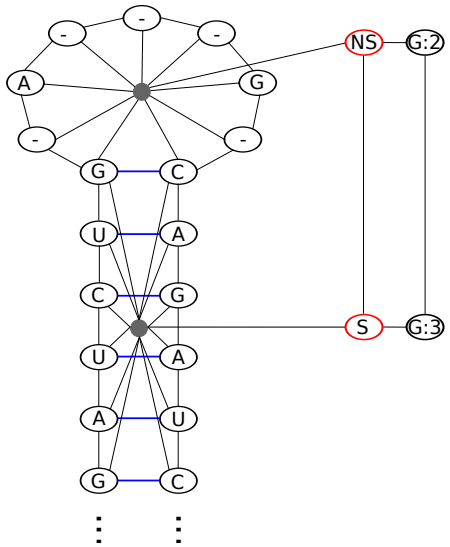


Figure : $\mathscr{S}_{\mathscr{U}}$: Conservation

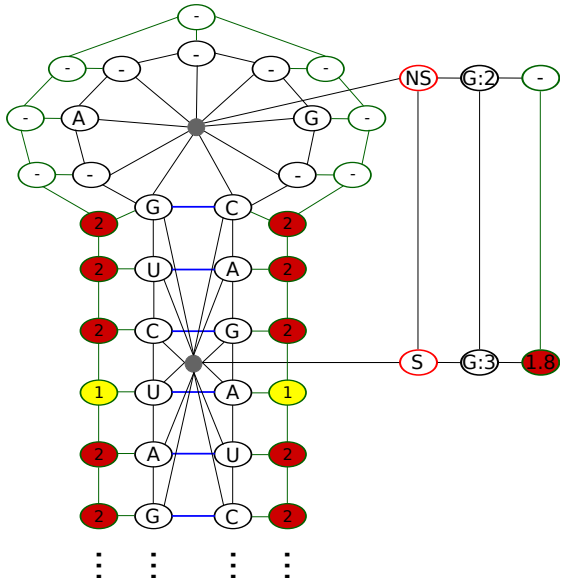


Figure : $\mathscr{S}_{\mathscr{L}}$: Conservation and covariation

- This extra information can be the Avg, Max, Min, of occurrence of a specific nucleotide or the conservation of the alignment information.

Data description

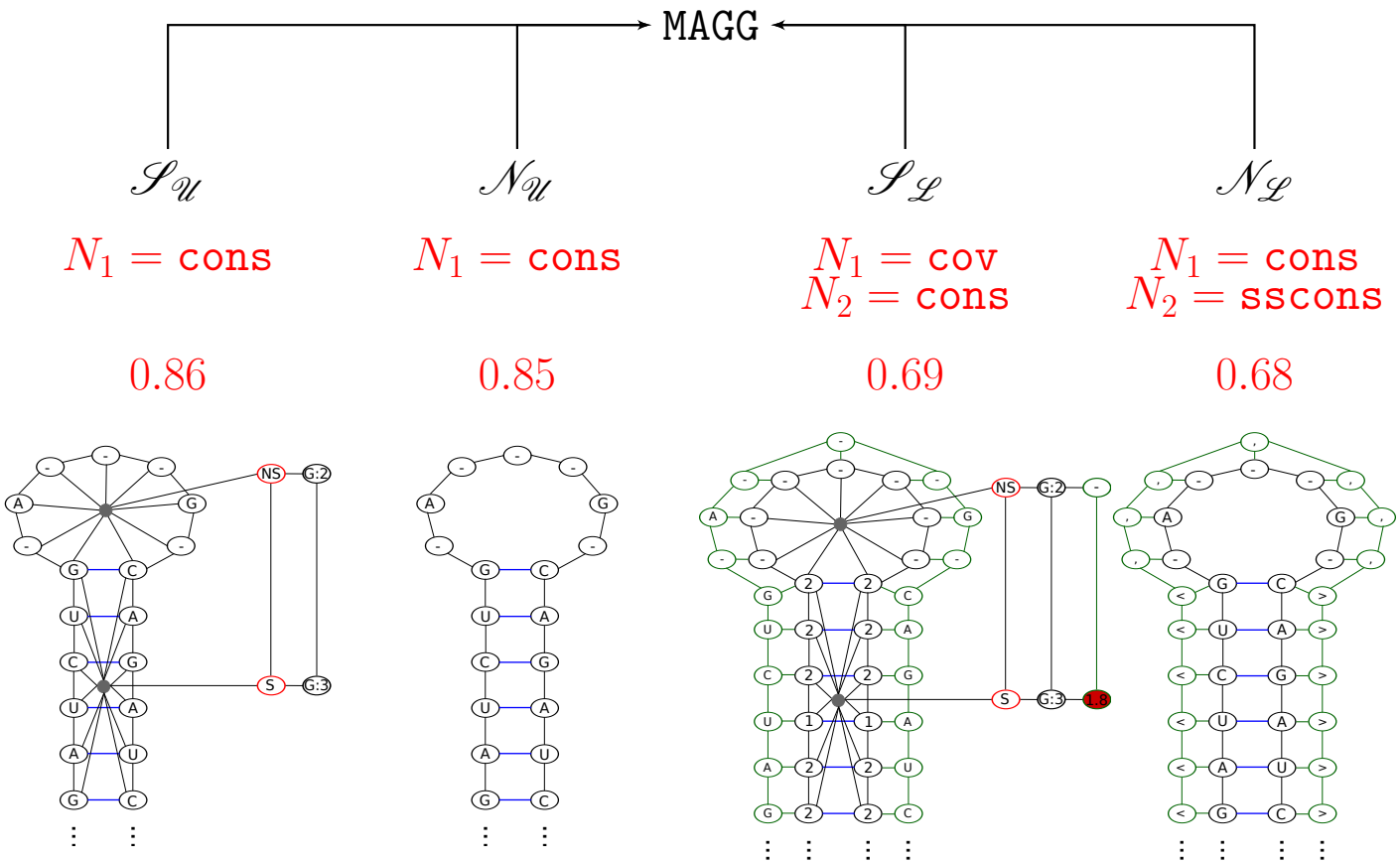
- The motif sequences are from bacteria, archaea [Weinberg 2010].
- Z.Weinberg has manually annotated the alignments in functional and non-functional.
- They are binary classified.

Data	Num. files	Num. classes	Avg seqs num.	Avg. seq. length
Positive	308	2 classes	70 seqs	150 nucleotides
Negative	16220	10 classes	70 seqs	130 nucleotides

Evaluation

- The experiment data sets were balanced.
  - Same number of files in pos and neg.
  - Testing each pos class against the 10 neg classes.
  - In total we have 20 experiments.
- The Receiver Operator Characteristic ROC is the performance measurement.
  - ROC computes the true positive rate against the false positive rate.
- The final ROC score is averaged over the different experiments.

Results (ROC)


MAGG
$\mathscr{S}_{\mathscr{U}}$
$\mathscr{N}_{\mathscr{U}}$
$\mathscr{S}_{\mathscr{L}}$
$\mathscr{N}_{\mathscr{L}}$
$N_1 = \texttt{cons}$
$N_1 = \texttt{cons}$
$N_1 = \texttt{cov}$
$N_2 = \texttt{cons}$
$N_1 = \texttt{cons}$
$N_2 = \texttt{sscons}$
0.86
0.85
0.69
0.68

Conclusion

- MAGG can identify interesting ncRNAs up to ROC 86%.

Take home message

The best graph representation is

1. Summary based $\mathscr{S}$.
2. Labelled with the conservation information.

- The tool can be used as:
  - $\rightarrow$ A powerful pre-filtering method for large amounts of alignments.

Current work

- Integrating MAGG into iPython environment.
- Integrate automated alignment of input sequences into MAGG.
- Encoding finer structural information as hairpins, bulges, and loops to improve the classification.

Acknowkegment

Prof.Dr. Rolf Backofen
Dr. Fabrizio Costa

Dr.Zasha Weinberg

Questions

Thank you for your attention