Adaptation of Frequent Subgraph Mining Algorithms to Noncoding RNA Topology Alignment and Function Prediction

University of Massachusetts Medical School

Program in Systems Biology

Nov 14th , 2017

Muyi Liu
Biological Sciences, Ph.D.
Computer Science, M.S.
liu413@purdue.edu
Purdue University

Outline

Introduction

Background: novel ncRNAs and important ncRNA functions

Purpose of project:

How to predict ncRNA’s function by common ncRNA topology?

Available methods and limitation

The MMC-Margin Algorithm

Identify common ncRNA topology

ncRNA Topology Alignment and Classification

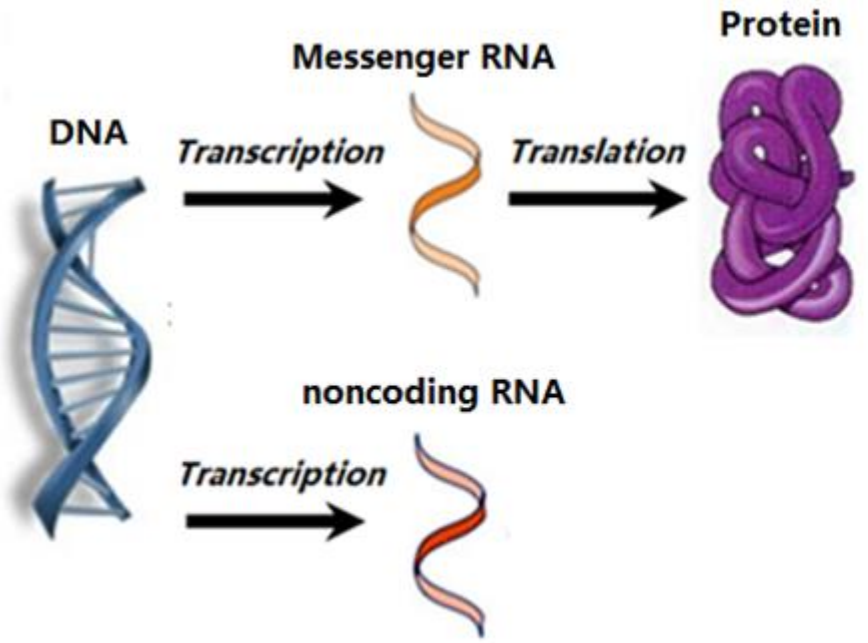
Predict ncRNA’s function by common ncRNA topology

Summary

Achievements

Future directions

Introduction



1.5% of Human Genome

Lander et al., 2001

Over 90% of Human Genome

Birney et al., 2007

Reported Novel ncRNAs

Encyclopedia of DNA Elements (ENCODE) Consortium:

93% of human genome is transcribed

(Birney et al., 2007)

53,864 previously unidentified long intergenic noncoding RNAs are reported

(Hangauer et al., 2013)

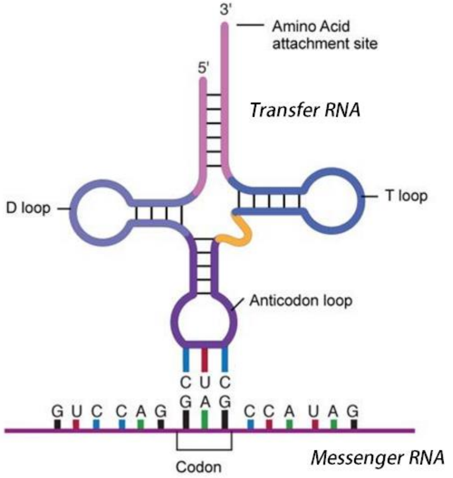
Functional Annotation of the Mammalian Genome (FANTOM) Consortium:

181,047 independent transcripts are reported from mouse transcriptomic data

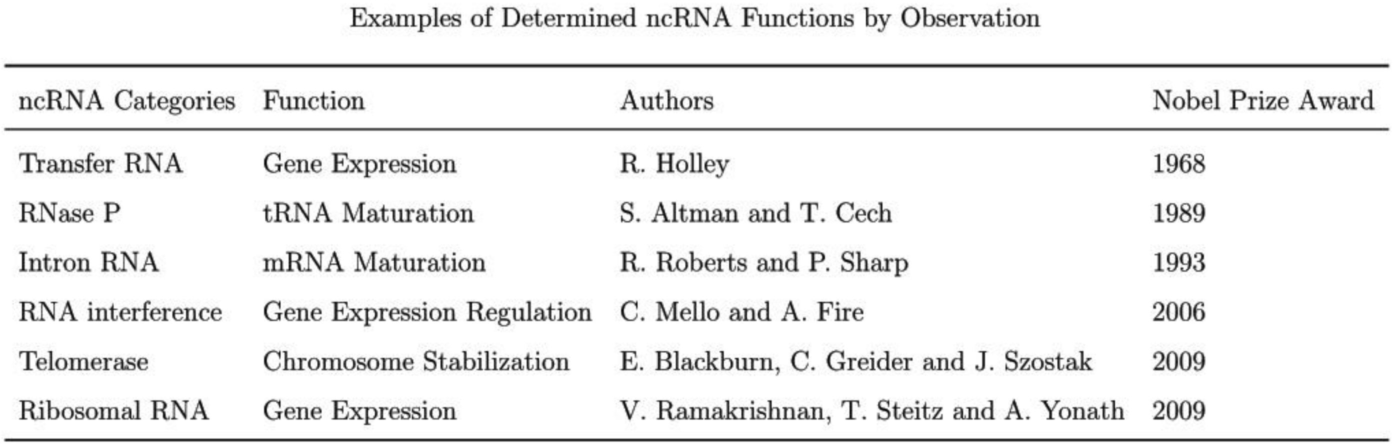
Estimated mouse genes: 22,000

(Carninci et al., 2005)

Important Functions of ncRNA

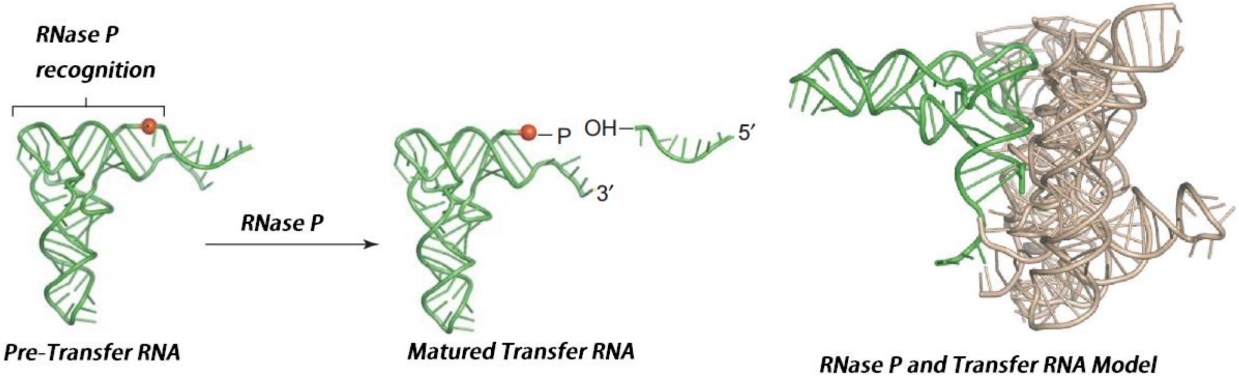


NHGRI Image Gallery



Examples of Determined ncRNA Functions by Observation

ncRNA Categories	Function	Authors	Nobel Prize Award
Transfer RNA	Gene Expression	R. Holley	1968
RNase P	tRNA Maturation	S. Altman and T. Cech	1989
Intron RNA	mRNA Maturation	R. Roberts and P. Sharp	1993
RNA interference	Gene Expression Regulation	C. Mello and A. Fire	2006
Telomerase	Chromosome Stabilization	E. Blackburn, C. Greider and J. Szostak	2009
Ribosomal RNA	Gene Expression	V. Ramakrishnan, T. Steitz and A. Yonath	2009



Pre-Transfer RNA

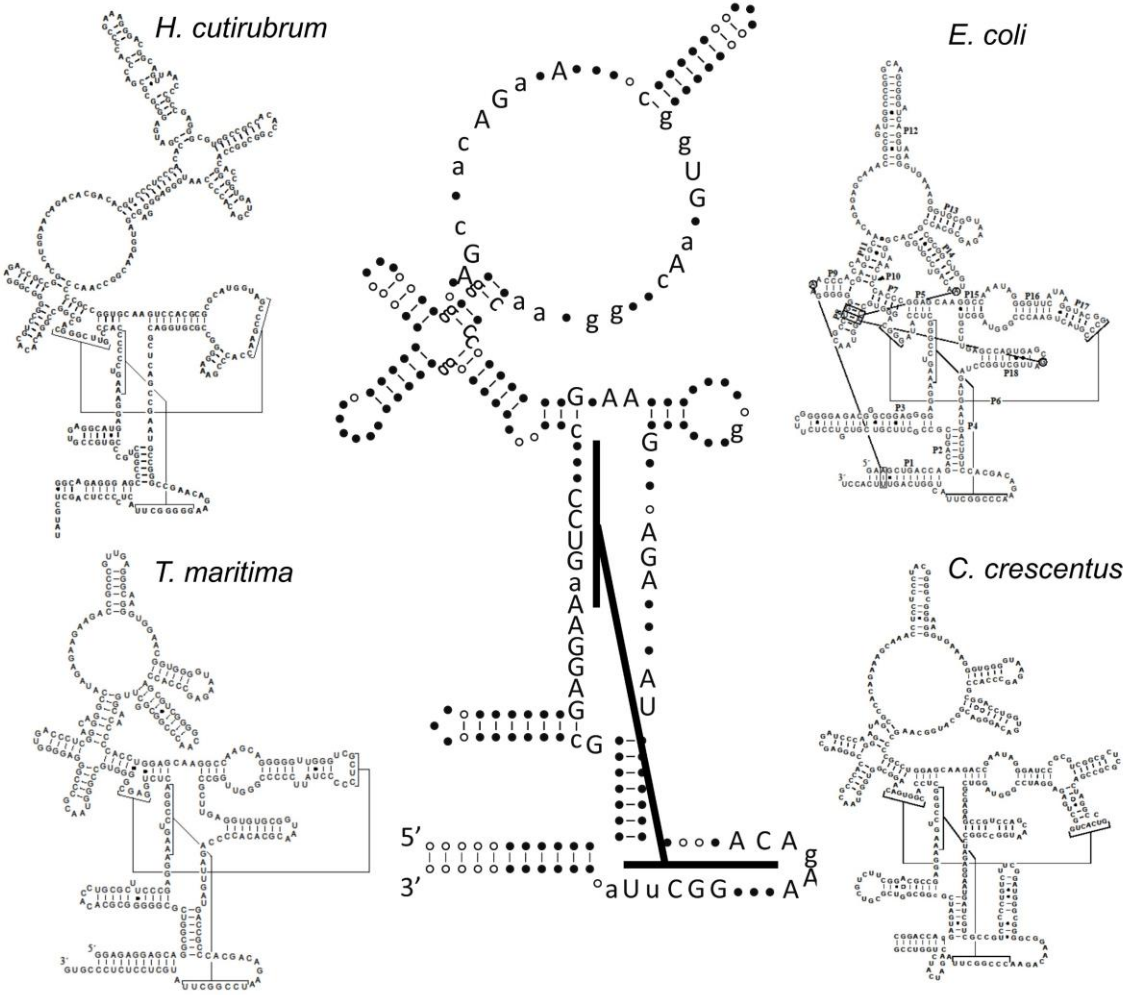
Matured Transfer RNA

RNase P and Transfer RNA Model

Evans et al., 2006

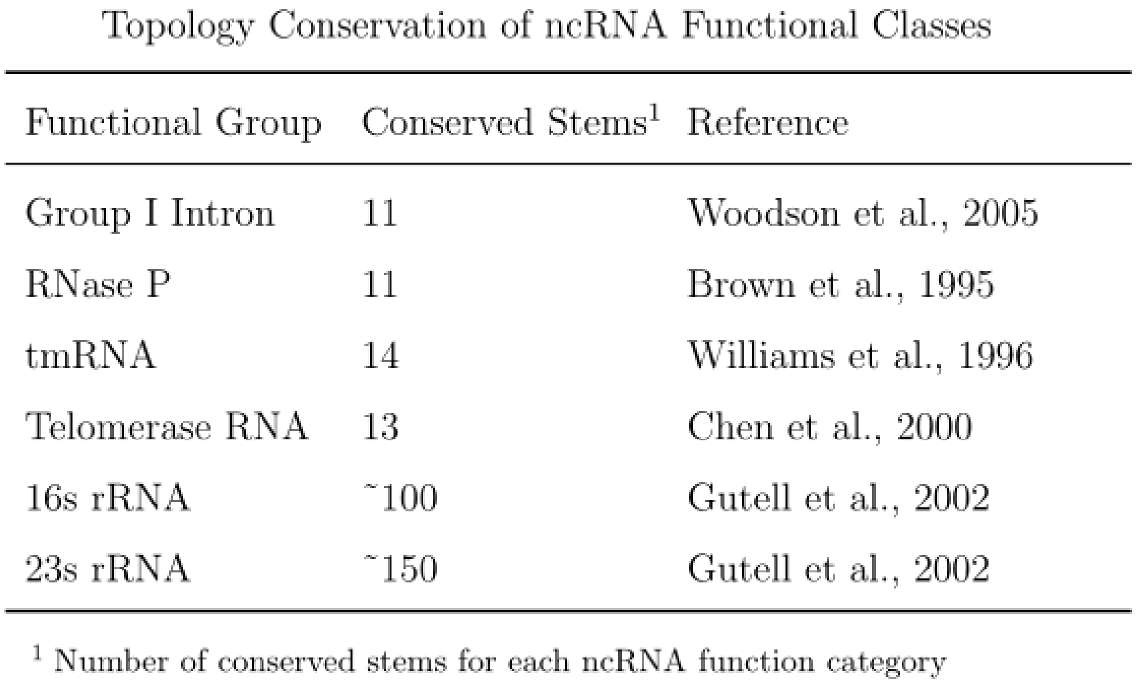
Can We Predict ncRNA’s Function?

Conserved Common Structure in RNase P



Conserved Common Substructure
in The Rnase P Database (Brown 1991)
Brown et al. 1993

More ncRNA Structure Conservation Studies



Topology Conservation of ncRNA Functional Classes

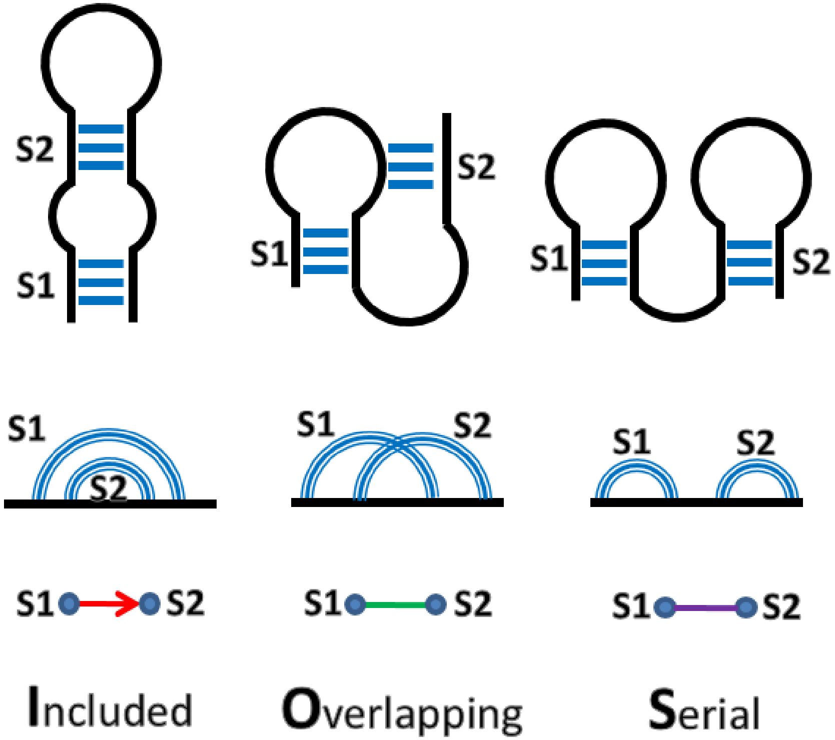
Functional Group	Conserved Stems[1]	Reference
Group I Intron	11	Woodson et al., 2005
RNase P	11	Brown et al., 1995
tmRNA	14	Williams et al., 1996
Telomerase RNA	13	Chen et al., 2000
16s rRNA	˜100	Gutell et al., 2002
23s rRNA	˜150	Gutell et al., 2002

[1] Number of conserved stems for each ncRNA function category

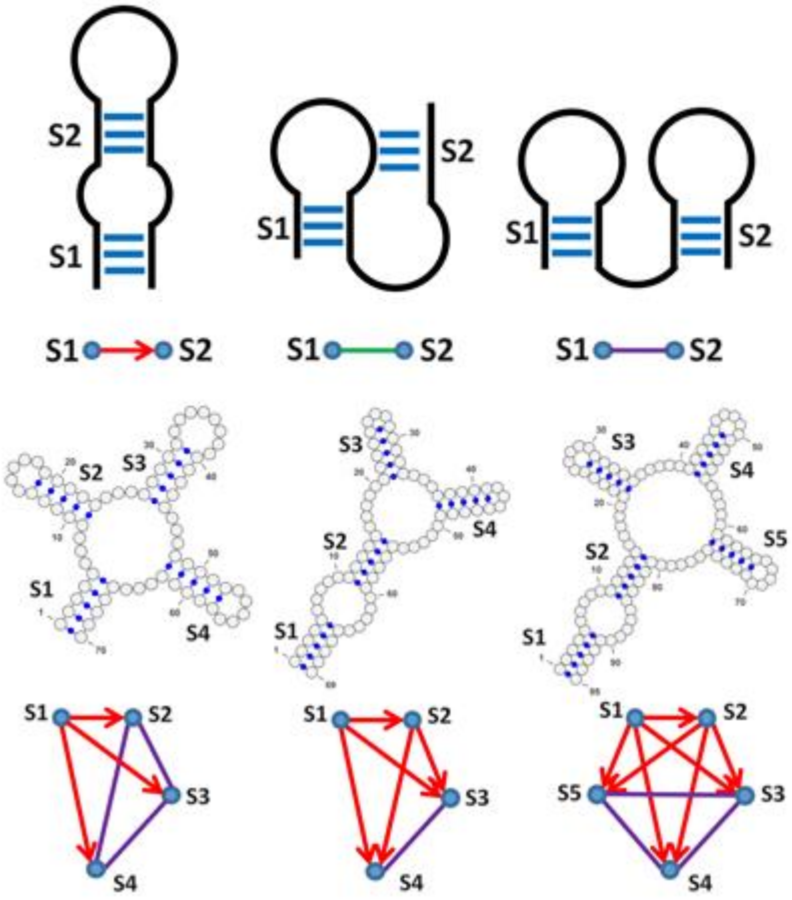
Preliminary Concept

ncRNA Topology Graph

ncRNA XIOS Topological Graph



ncRNA XIOS Topological Graph



ncRNA XIOS graphs

Definition 1 Labeled Graph:

A labeled graph is a tuple: $G = (V, E, \Lambda, \lambda)$, where

V: a set of vertices

E: a set of edges $V \times V$

Λ: a set of edge labels

λ: $V \cup E \rightarrow \Lambda$, assign labels to vertices and edges

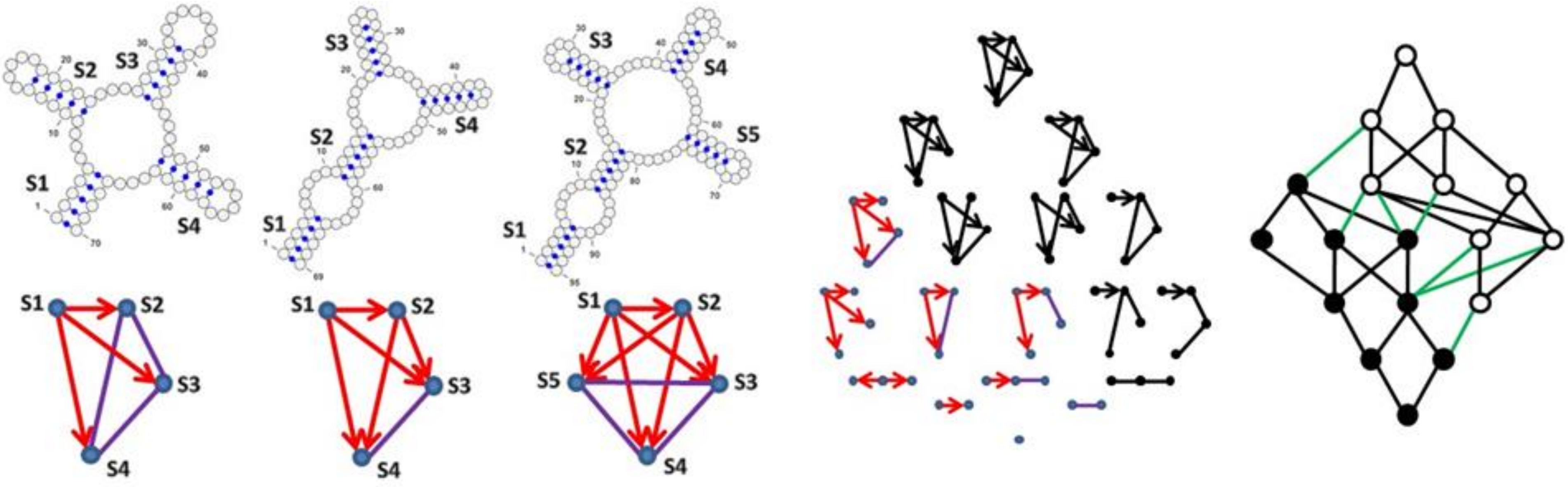
graph length: $|G|$=$|E|$

Preliminary Concept

Graph Theory

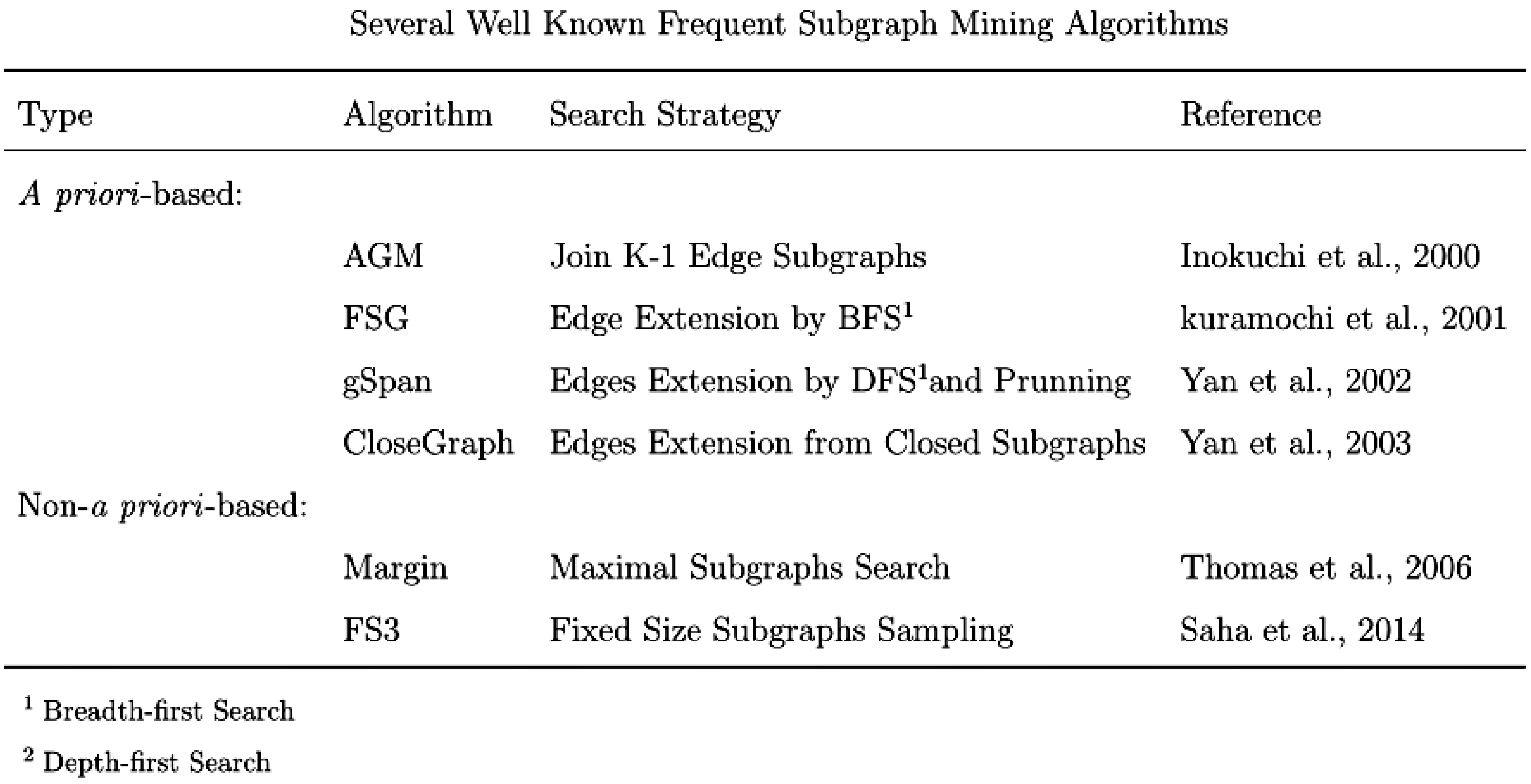
Frequent Subgraph Mining Algorithms

FSM Lattice Space: A Toy Example



ncRNA XIOS Graph Alignment, FSM Lattice Space, and *Cut* Pairs

FSM Algorithms



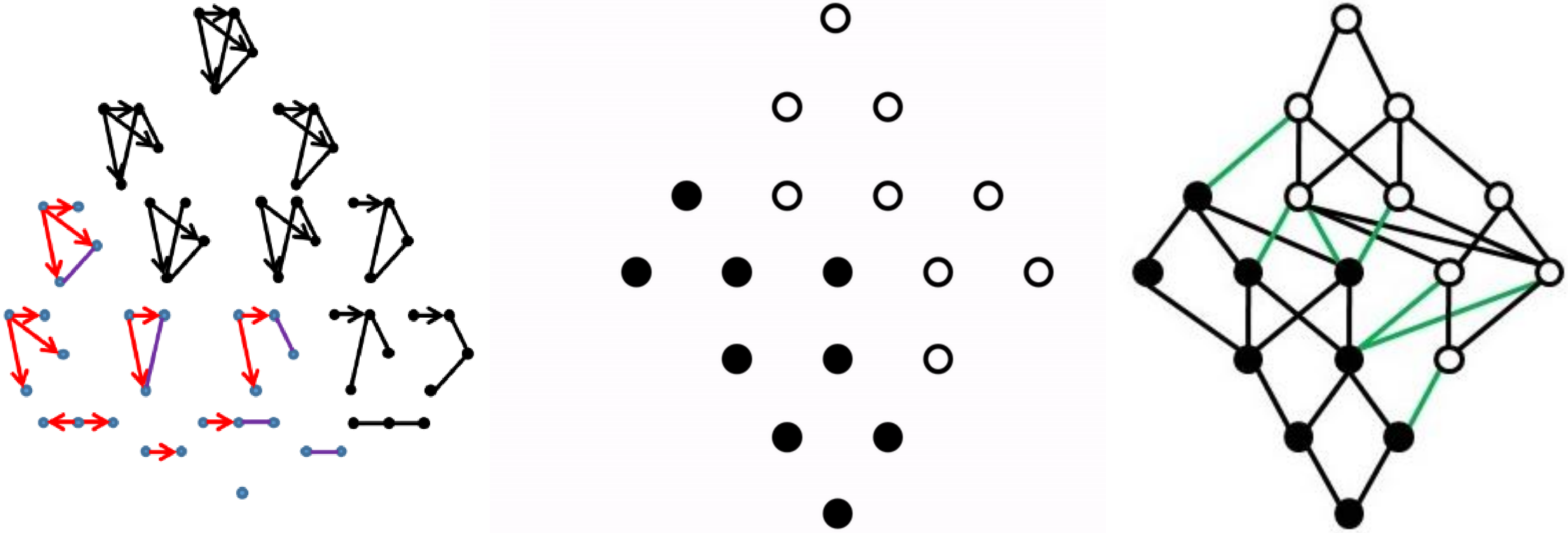
Several Well Known Frequent Subgraph Mining Algorithms

Type	Algorithm	Search Strategy	Reference
A priori-based:			
	AGM	Join K-1 Edge Subgraphs	Inokuchi et al., 2000
	FSG	Edge Extension by BFS[1]	kuramochi et al., 2001
	gSpan	Edges Extension by DFS[1]and Prunning	Yan et al., 2002
	CloseGraph	Edges Extension from Closed Subgraphs	Yan et al., 2003
Non-*a priori*-based:			
	Margin	Maximal Subgraphs Search	Thomas et al., 2006
	FS3	Fixed Size Subgraphs Sampling	Saha et al., 2014

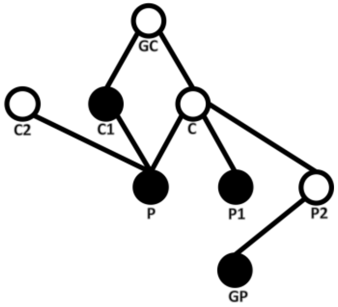
[1] Breadth-first Search

[2] Depth-first Search

The Margin Algorithm



The Margin Algorithm (Thomas et al., 2006)



Neighboring Cuts

FSM Complexity & NP-Completeness

FSM Lattice Scalability

The FSM lattice space includes ***O*****(2^n)** nodes

20 stem ncRNA structure may contain 190 edges

FSM lattice space is about **$2^{190} \approx 10^{57}$**

(assume search one node in one second)

100 years ≈ 10^9 seconds and estimated universe age ≈ 10^{17} seconds

FSM is NP-Hard (nondeterministic polynomial-time)

Subgraph Isomorphism (SI) problem is NP-Complete

Reduce from Clique problem

(Cook et al., 1971)

The FSM problem is NP-Hard

Reduce from SI problem

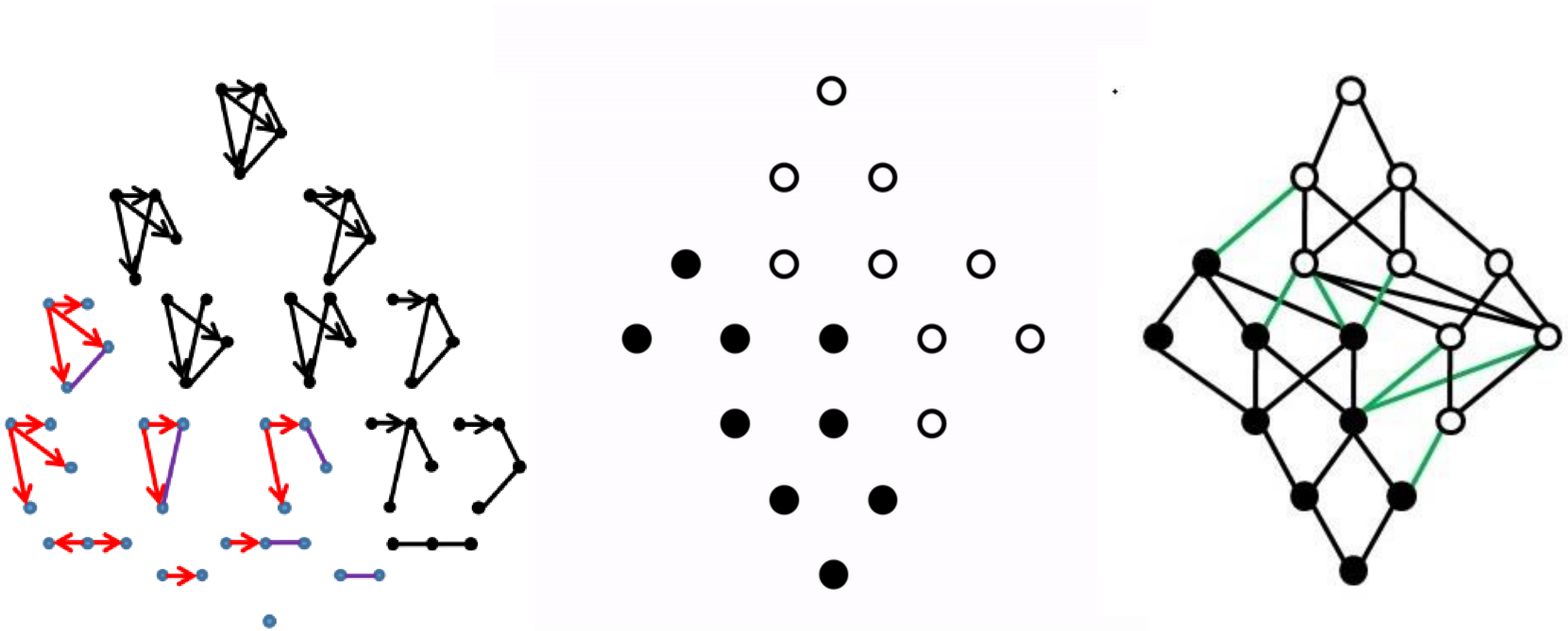
(Garey et al., 1979)

(Kimelfeld et al., 2014)

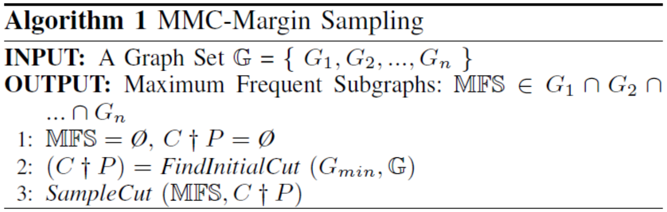
The MMC-Margin Algorithm

(Metropolis Monte Carlo Sampling)

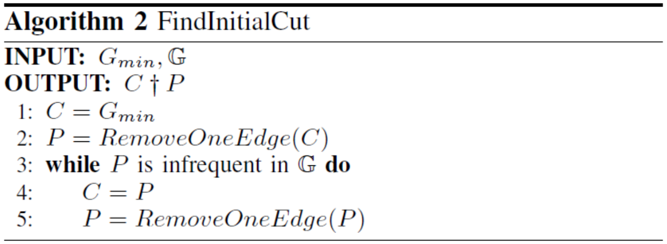
The MMC-Margin Algorithm



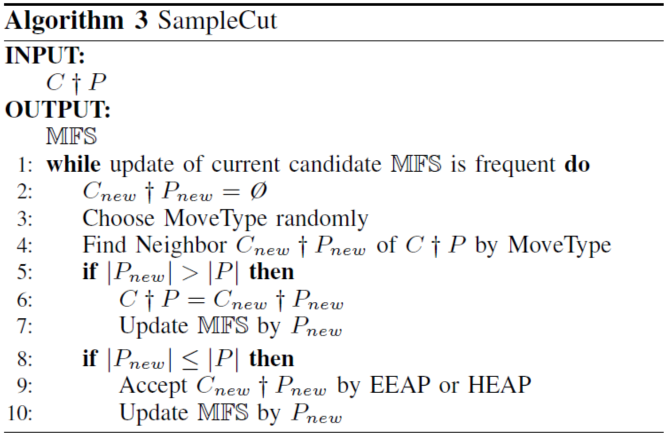
The MMC-Margin Algorithm (Liu et al., 2015)



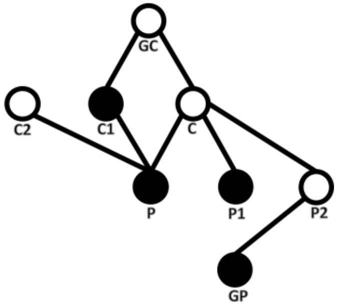
```
Algorithm 1 MMC-Margin Sampling
INPUT: A Graph Set $\mathbb{G} = \{ G_1, G_2, ..., G_n \}$
OUTPUT: Maximum Frequent Subgraphs: $\mathbb{MFS} \in G_1 \cap G_2 \cap ... \cap G_n$
$\mathbb{MFS} = \emptyset$, $C \dagger P = \emptyset$
$(C \dagger P) = FindInitialCut\ (G_{min}, \mathbb{G})$
$SampleCut\ (\mathbb{MFS}, C \dagger P)$
```



```
Algorithm 2 FindInitialCut
INPUT: $G_{min}, \mathbb{G}$
OUTPUT: $C \dagger P$
$C = G_{min}$
$P = RemoveOneEdge(C)$
while $P$ is infrequent in $\mathbb{G}$ do
    $C = P$
    $P = RemoveOneEdge(P)$
```



```
Algorithm 3 SampleCut
INPUT:
    $C \dagger P$
OUTPUT:
    $\mathbb{MFS}$
while update of current candidate $\mathbb{MFS}$ is frequent do
    $C_{new} \dagger P_{new} = \emptyset$
    Choose MoveType randomly
    Find Neighbor $C_{new} \dagger P_{new}$ of $C \dagger P$ by MoveType
    if $|P_{new}| > |P|$ then
        $C \dagger P = C_{new} \dagger P_{new}$
        Update $\mathbb{MFS}$ by $P_{new}$
    if $|P_{new}| \le |P|$ then
        Accept $C_{new} \dagger P_{new}$ by EEAP or HEAP
        Update $\mathbb{MFS}$ by $P_{new}$
```

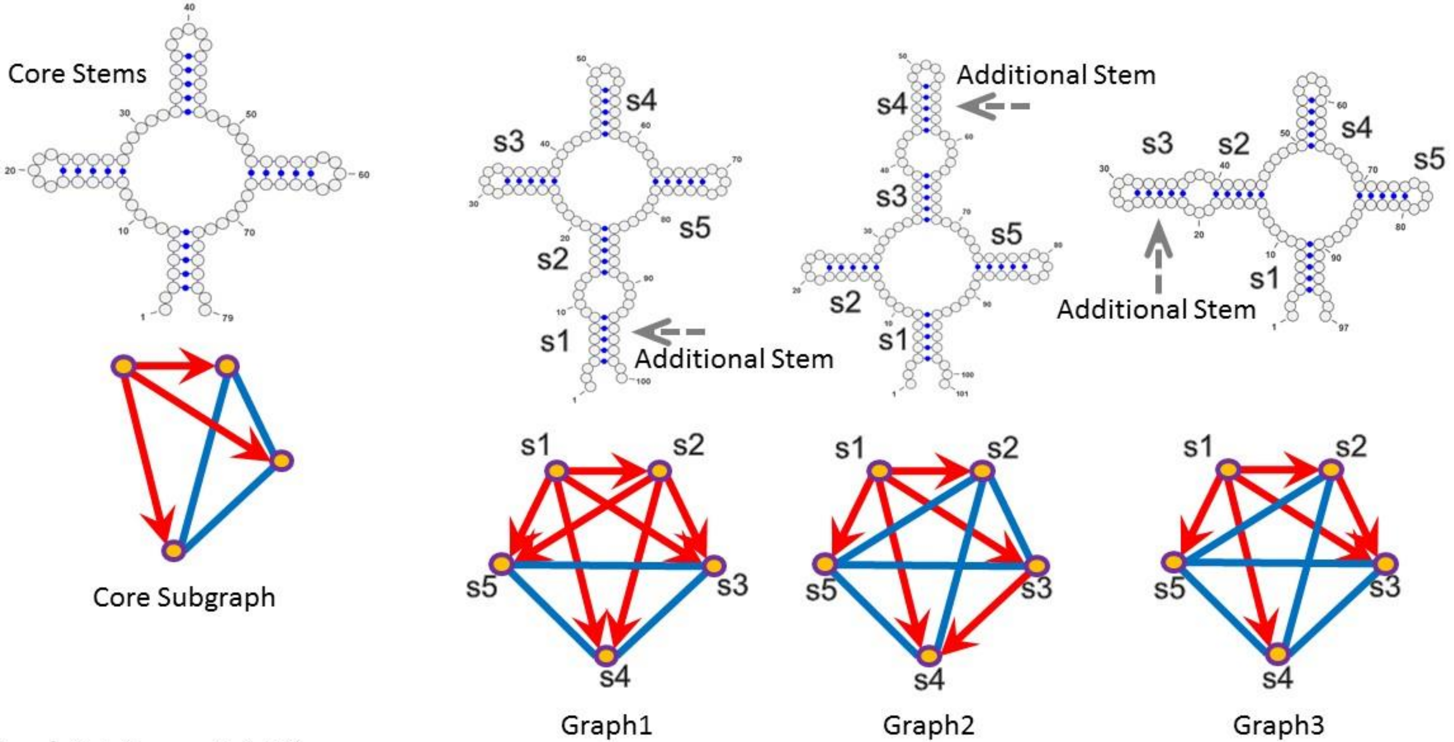


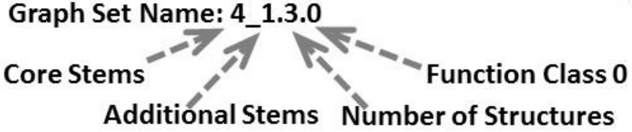
Neighboring Cuts

ncRNA Structure Generator

Synthetic RNA Graphsets Statistics:

	Statistics					
Datasets	|V|avg[1]	|E|avg[2]	Davg[3]	Dmax[4]	|V|core[5]	|E|core[6]




Graph Set Name: 4_1.3.0

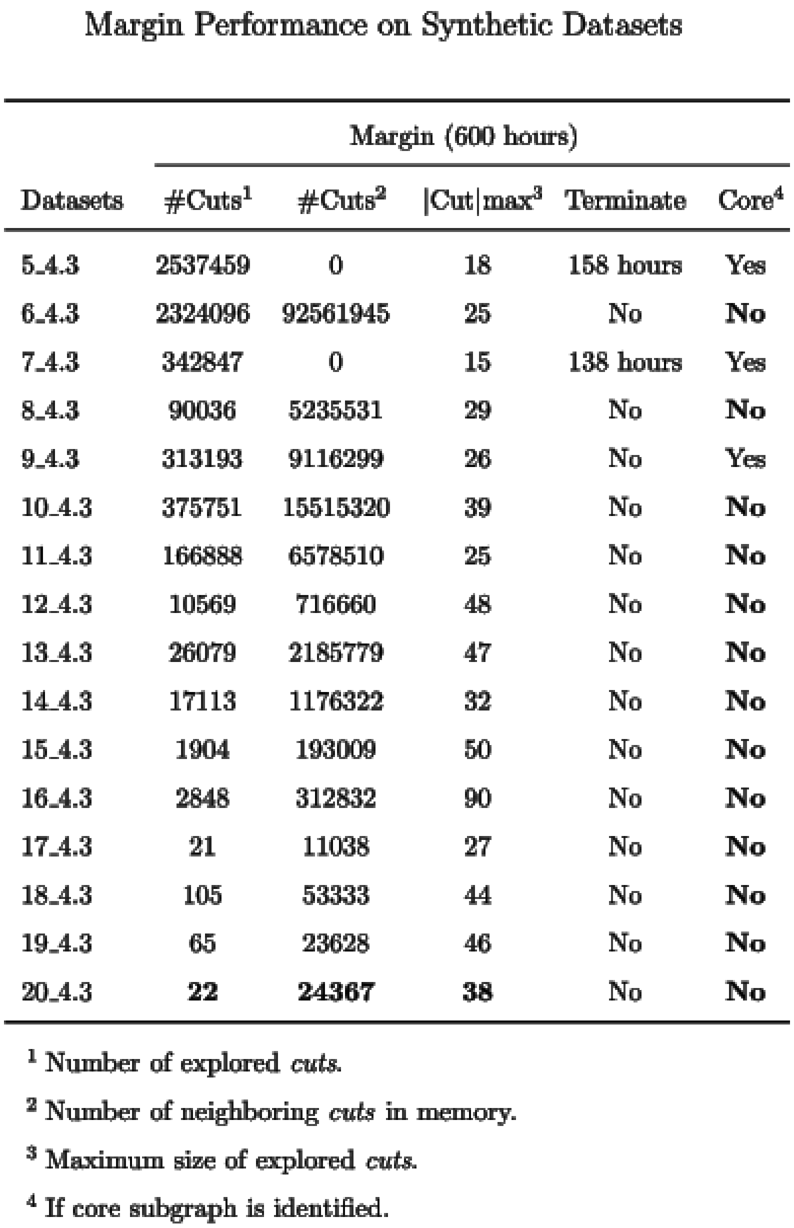
[4] Maximum degree.

[5] Number of vertex in core subgraph.

[6] Number of edge in core subgraph.

[7] Three RNA structures share 5 stems core.

MMC-Margin Outperforms Margin



Margin Performance on Synthetic Datasets

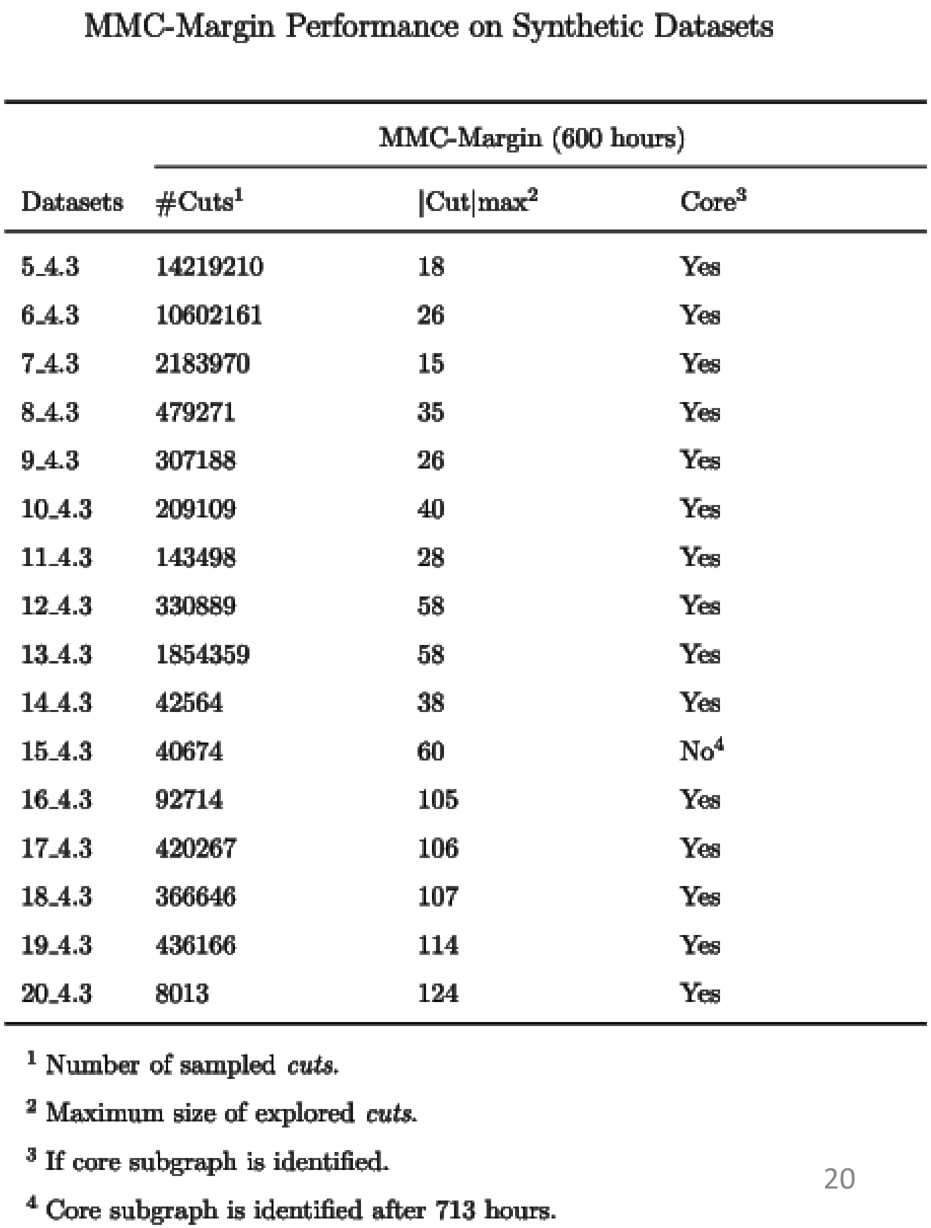
	Margin (600 hours)				
Datasets	#Cuts[1]	#Cuts[2]	|Cut|max[3]	Terminate	Core[4]
5_4.3	2537459	0	18	158 hours	Yes
6_4.3	2324096	92561945	25	No	**No**
7_4.3	342847	0	15	138 hours	Yes
8_4.3	90036	5235531	29	No	**No**
9_4.3	313193	9116299	26	No	Yes
10_4.3	375751	15515320	39	No	**No**
11_4.3	166888	6578510	25	No	**No**
12_4.3	10569	716660	48	No	**No**
13_4.3	26079	2185779	47	No	**No**
14_4.3	17113	1176322	32	No	**No**
15_4.3	1904	193009	50	No	**No**
16_4.3	2848	312832	90	No	**No**
17_4.3	21	11038	27	No	**No**
18_4.3	105	53333	44	No	**No**
19_4.3	65	23628	46	No	**No**
20_4.3	**22**	**24367**	**38**	No	**No**

[1] Number of explored *cuts*.

[2] Number of neighboring *cuts* in memory.

[3] Maximum size of explored *cuts*.

[4] If core subgraph is identified.



MMC-Margin Performance on Synthetic Datasets

	MMC-Margin (600 hours)		
Datasets	#Cuts[1]	|Cut|max[2]	Core[3]
5_4.3	14219210	18	Yes
6_4.3	10602161	26	Yes
7_4.3	2183970	15	Yes
8_4.3	479271	35	Yes
9_4.3	307188	26	Yes
10_4.3	209109	40	Yes
11_4.3	143498	28	Yes
12_4.3	330889	58	Yes
13_4.3	1854359	58	Yes
14_4.3	42564	38	Yes
15_4.3	40674	60	No[4]
16_4.3	92714	105	Yes
17_4.3	420267	106	Yes
18_4.3	366646	107	Yes
19_4.3	436166	114	Yes
20_4.3	8013	124	Yes

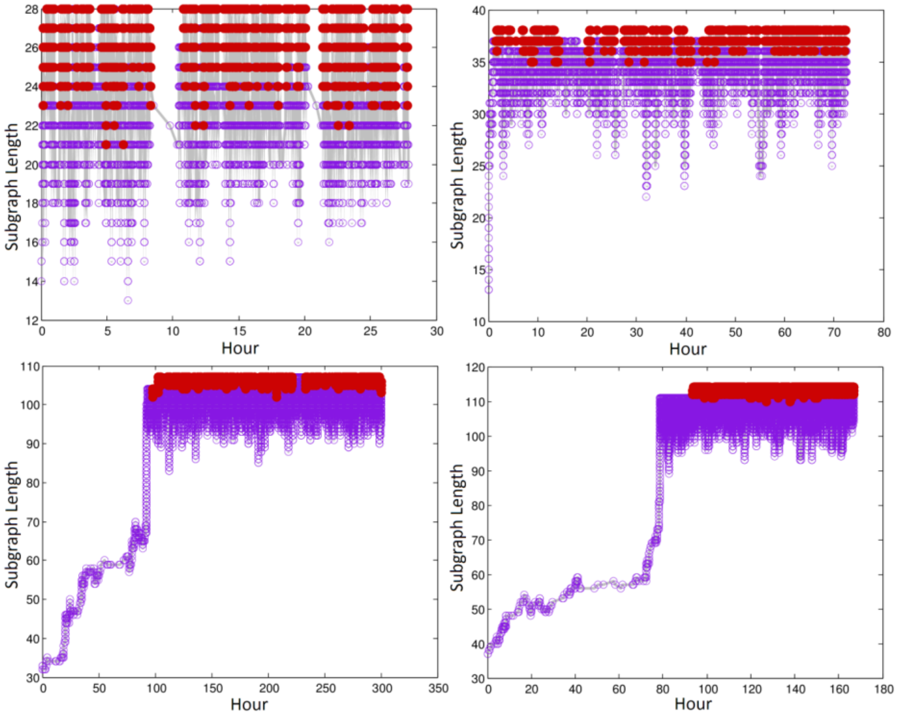
[1] Number of sampled *cuts*.

[2] Maximum size of explored *cuts*.

[3] If core subgraph is identified.

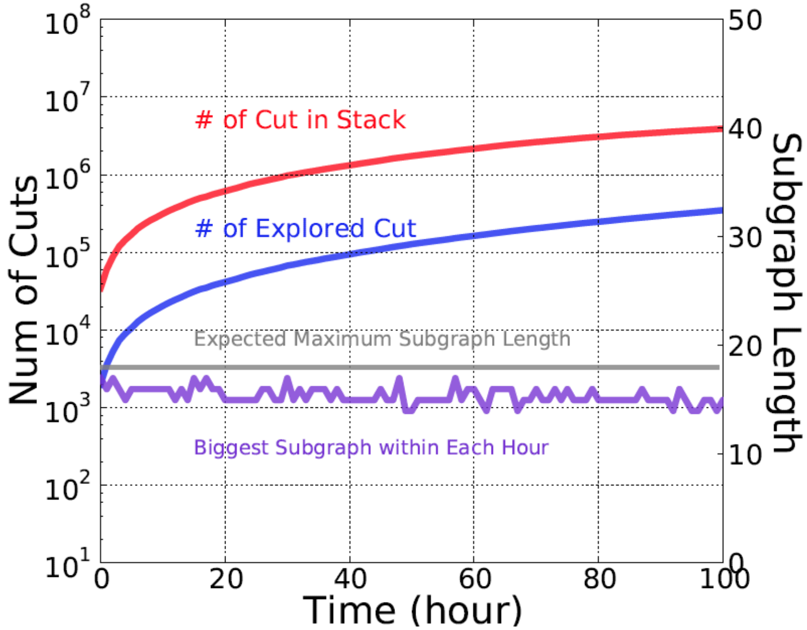
[4] Core subgraph is identified after 713 hours.

Traceplot Examples

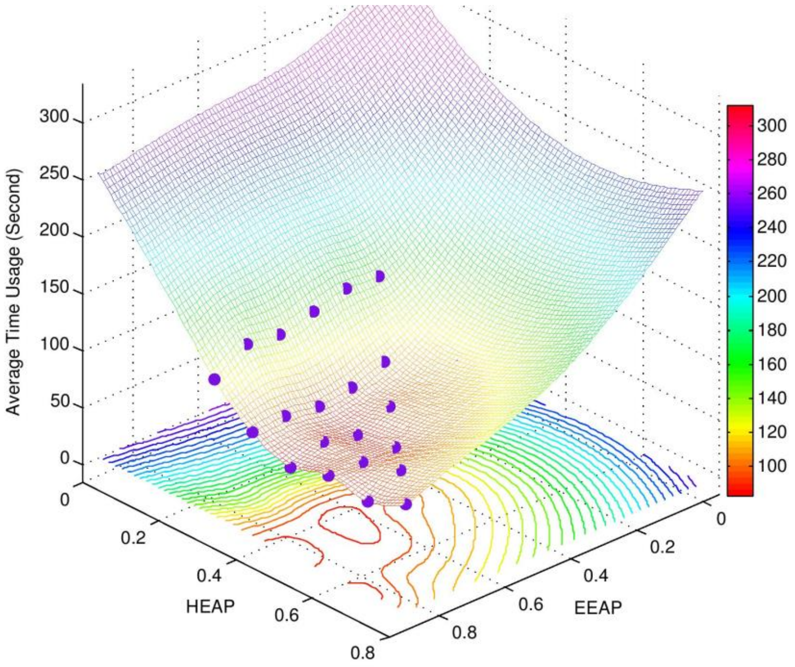


MMC-Margin Identifies Core Subgraphs: 11_4.3 (top left), 14_4.3 (top right), 18_4.3 (bottom left), 19_4.3 (bottom right). The existence of core subgraph is indicated by red dots.

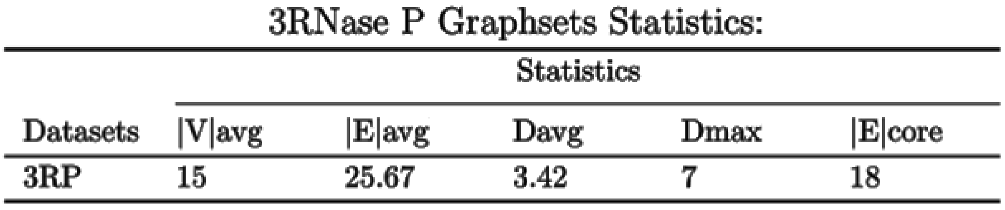
MMC-Margin Outperforms Margin on RNase P



Margin Cannot Identify Core Subgraphs in Hours



MMC-Margin Identifies Core Subgraphs in Minutes
(Acceptance Ratio Optimization)



3RNase P Graphsets Statistics:

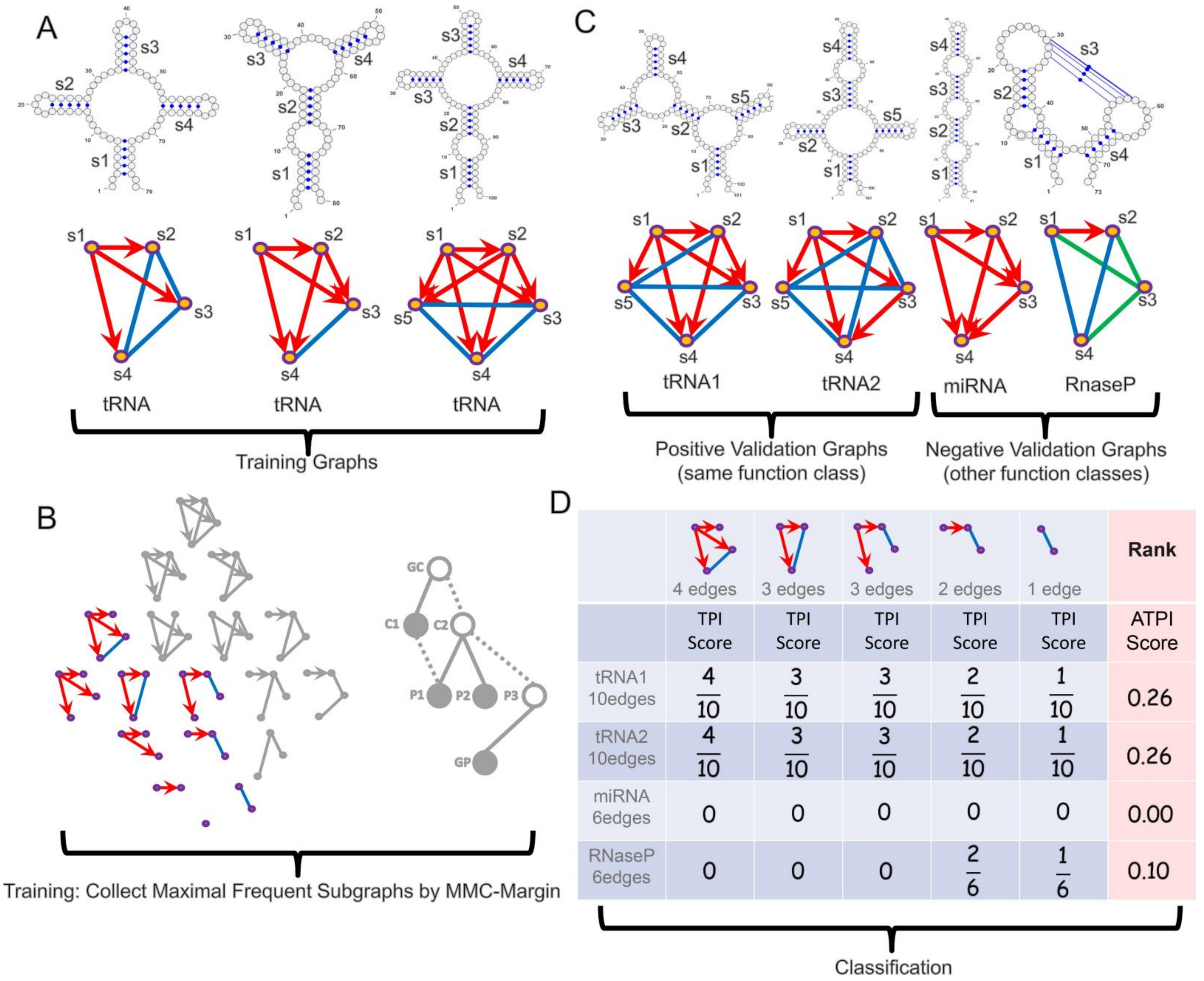
	Statistics				
Datasets	|V|avg	|E|avg	Davg	Dmax	|E|core
3RP	15	25.67	3.42	7	18

The MMC-Margin Algorithm Conclusion

MMC-Margin identifies core subgraphs shared among ncRNA structures quickly

ncRNA Topological Graph Classification

ncRNA Topological Graph Classification Model



Training Graphs

Training: Collect Maximal Frequent Subgraphs by MMC-Margin

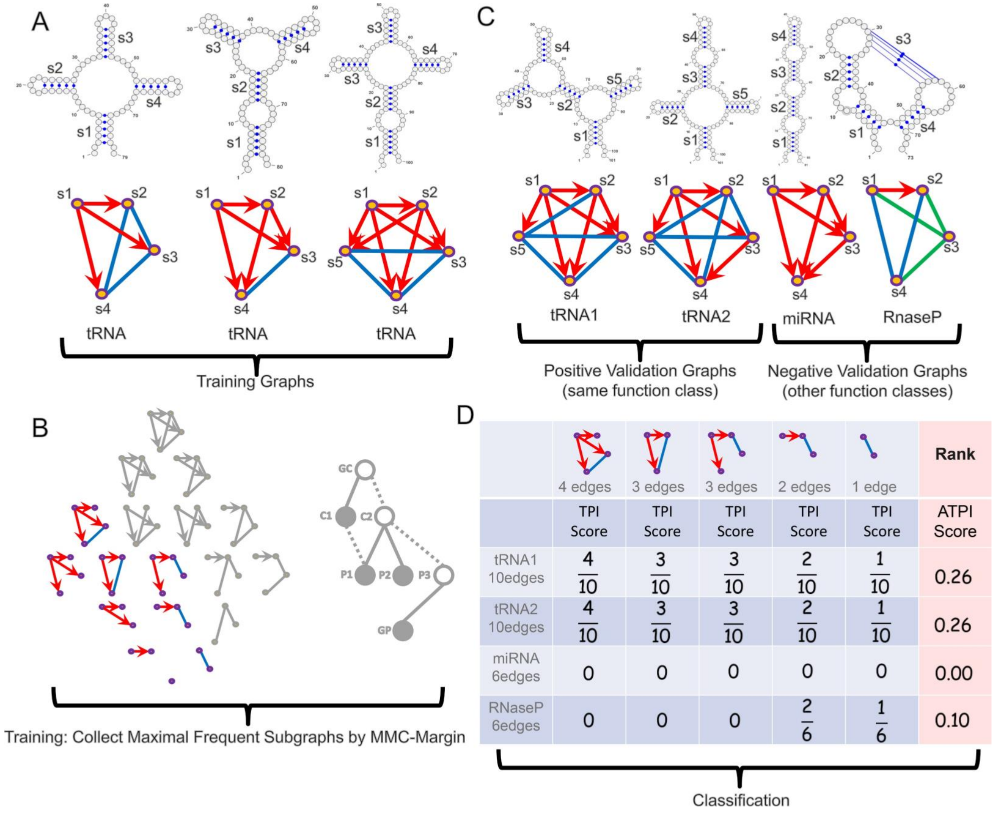
Positive Validation Graphs (same function class)

Negative Validation Graphs (other function classes)

	4 edges	3 edges	3 edges	2 edges	1 edge	Rank
	TPI Score	TPI Score	TPI Score	TPI Score	TPI Score	ATPI Score
tRNA1 10edges	$\frac{4}{10}$	$\frac{3}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	0.26
tRNA2 10edges	$\frac{4}{10}$	$\frac{3}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	0.26
miRNA 6edges	0	0	0	0	0	0.00
RNaseP 6edges	0	0	0	$\frac{2}{6}$	$\frac{1}{6}$	0.10

Classification

ncRNA Topology Classification Algorithm

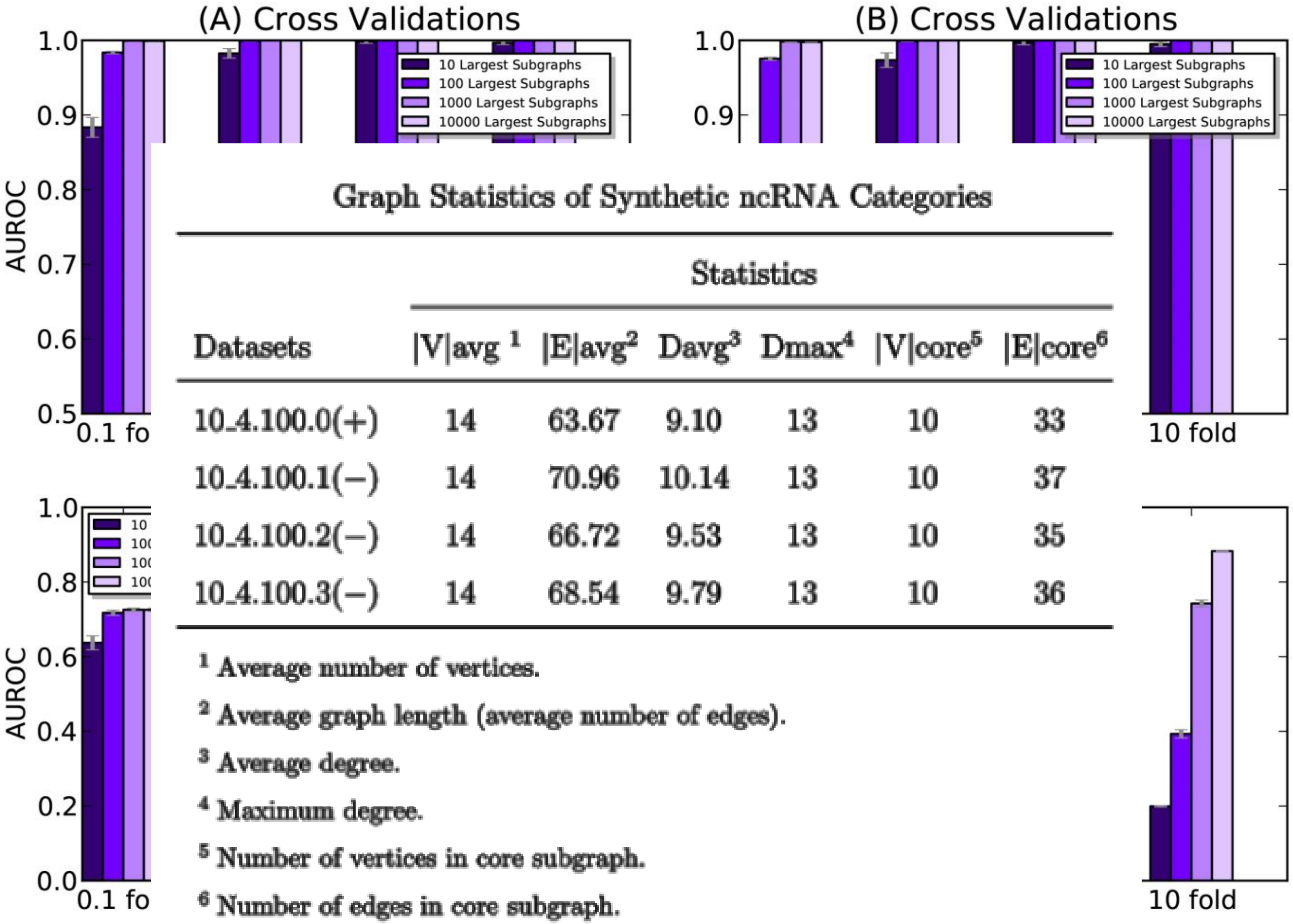


	4 edges	3 edges	3 edges	2 edges	1 edge	Rank
	TPI Score	TPI Score	TPI Score	TPI Score	TPI Score	ATPI Score
tRNA1 10edges	$\frac{4}{10}$	$\frac{3}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	0.26
tRNA2 10edges	$\frac{4}{10}$	$\frac{3}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	0.26
miRNA 6edges	0	0	0	0	0	0.00
RNaseP 6edges	0	0	0	$\frac{2}{6}$	$\frac{1}{6}$	0.10

Algorithm 4 ncRNA Topology Classification

```
INPUT: Training Graphs $\mathbb{T} = \{ T_1, ..., T_m \}$ and Test Graphs $\mathbb{G} = \{ G_1, ..., G_n \}$
OUTPUT: Ranking Score of $\mathbb{G}$: $\mathbb{R} = \{ R_1, ..., R_n \}$
$\mathbb{S}_{\mathbb{T}} : \{ S_1, ..., S_t \} = \mathit{MMC\text{-}Margin}\ (\mathbb{T})$ //Training
$\mathbb{S}_{\mathbb{C}} : \{ S_1, ..., S_c \} = \mathit{SelectTop}\ (\mathbb{S}_{\mathbb{T}})$ //Feature Selection
for $G_i \leftarrow G_1$ to $G_n$ do
    $R_i = \mathit{ATPI}( \mathbb{S}_{\mathbb{C}}, G_i )$
```

Can We Distinguish among Same Size Structures?



Graph Statistics of Synthetic ncRNA Categories

	Statistics					
Datasets	|V|avg [1]	|E|avg [2]	Davg [3]	Dmax [4]	|V|core [5]	|E|core [6]
10_4_100_0(+)	14	63.67	9.10	13	10	33
10_4_100_1(−)	14	70.96	10.14	13	10	37
10_4_100_2(−)	14	66.72	9.53	13	10	35
10_4_100_3(−)	14	68.54	9.79	13	10	36

[1] Average number of vertices.

[2] Average graph length (average number of edges).

[3] Average degree.

[4] Maximum degree.

[5] Number of vertices in core subgraph.

[6] Number of edges in core subgraph.

Cross Validations and Baselines on Synthetic ncRNA Functional Classes

Note:

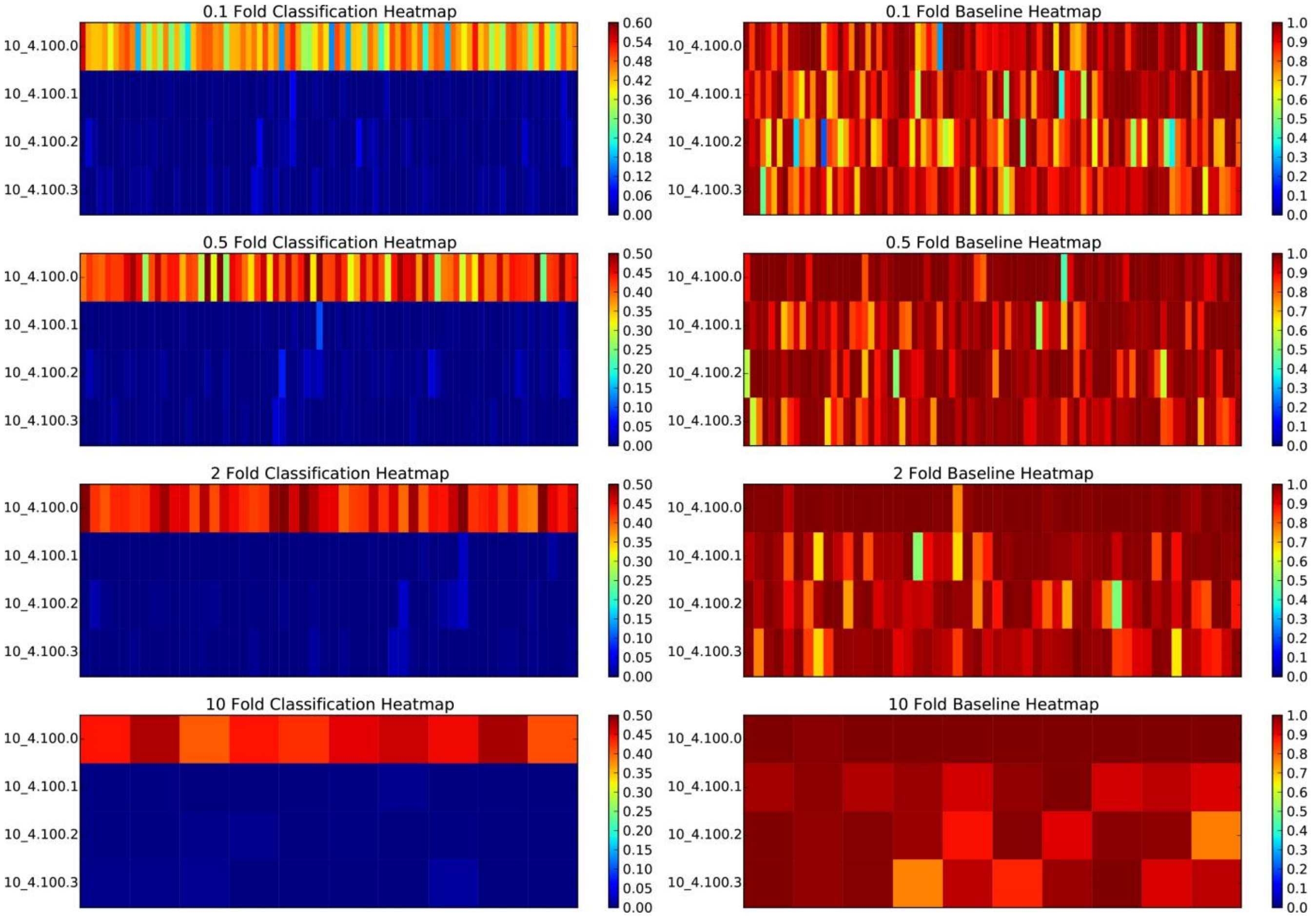
0.1 fold cross validation:

Inverse 10 fold cross validation, each 10 as training and each 90 in each function class as validation

0.5 fold cross validation:

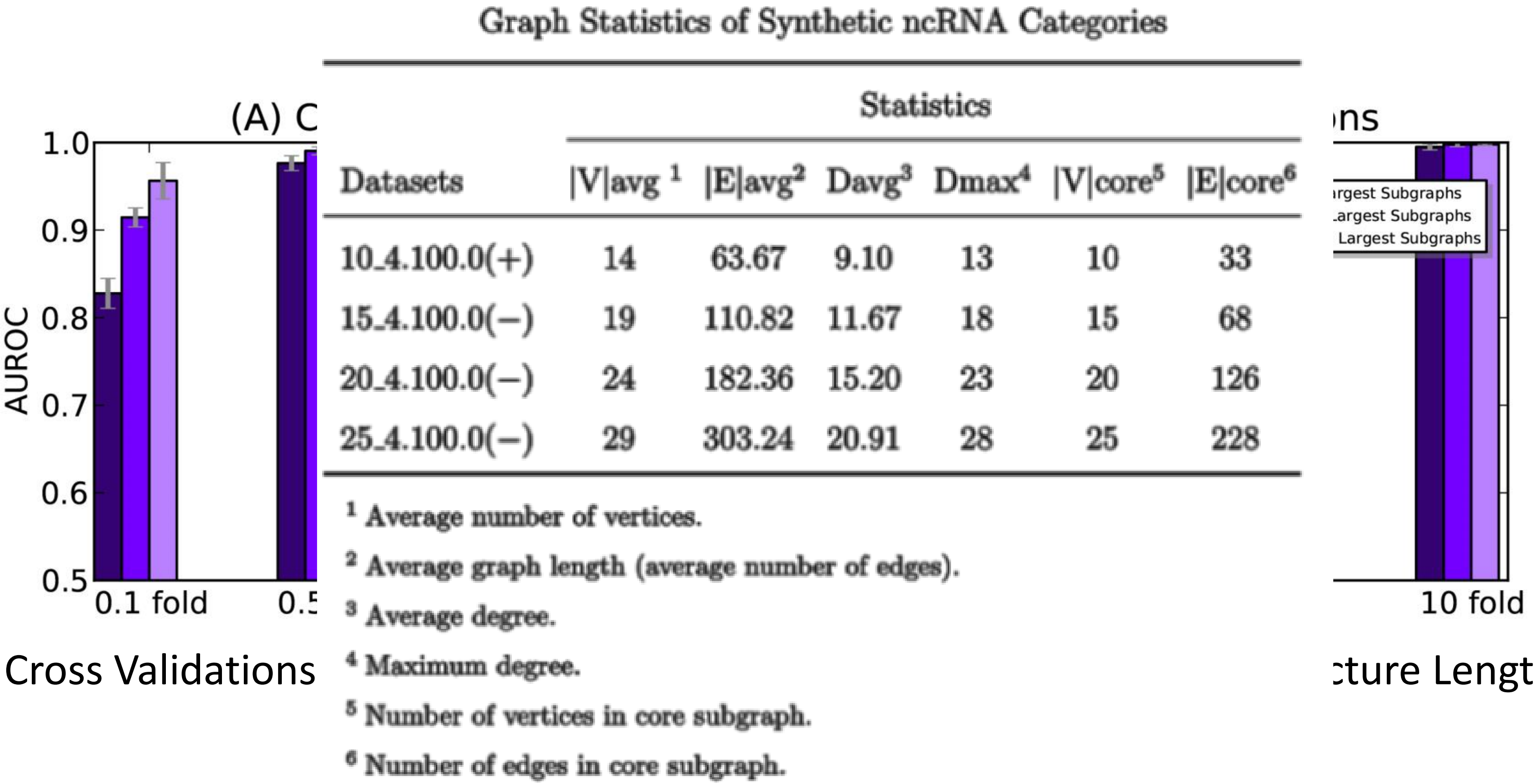
Inverse 5 fold cross validation, each 20 as training and each 80 in each function class as validation

Heatmap Examples



The ATPI Heatmap Examples of Cross Validations and Baselines

Cross Validations among Different Length Structures


(A) C
ns
AUROC
1.0
0.9
0.8
0.7
0.6
0.5
0.1 fold
0.5
10 fold
rgest Subgraphs
argest Subgraphs
Largest Subgraphs

Graph Statistics of Synthetic ncRNA Categories

Datasets	Statistics					
	|V|avg [1]	|E|avg[2]	Davg[3]	Dmax[4]	|V|core[5]	|E|core[6]
10_4.100.0(+)	14	63.67	9.10	13	10	33
15_4.100.0(−)	19	110.82	11.67	18	15	68
20_4.100.0(−)	24	182.36	15.20	23	20	126
25_4.100.0(−)	29	303.24	20.91	28	25	228

[1] Average number of vertices.

[2] Average graph length (average number of edges).

[3] Average degree.

[4] Maximum degree.

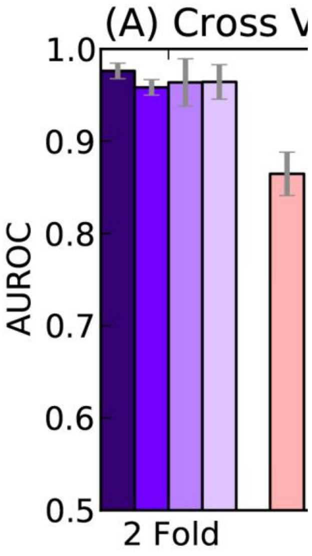
[5] Number of vertices in core subgraph.

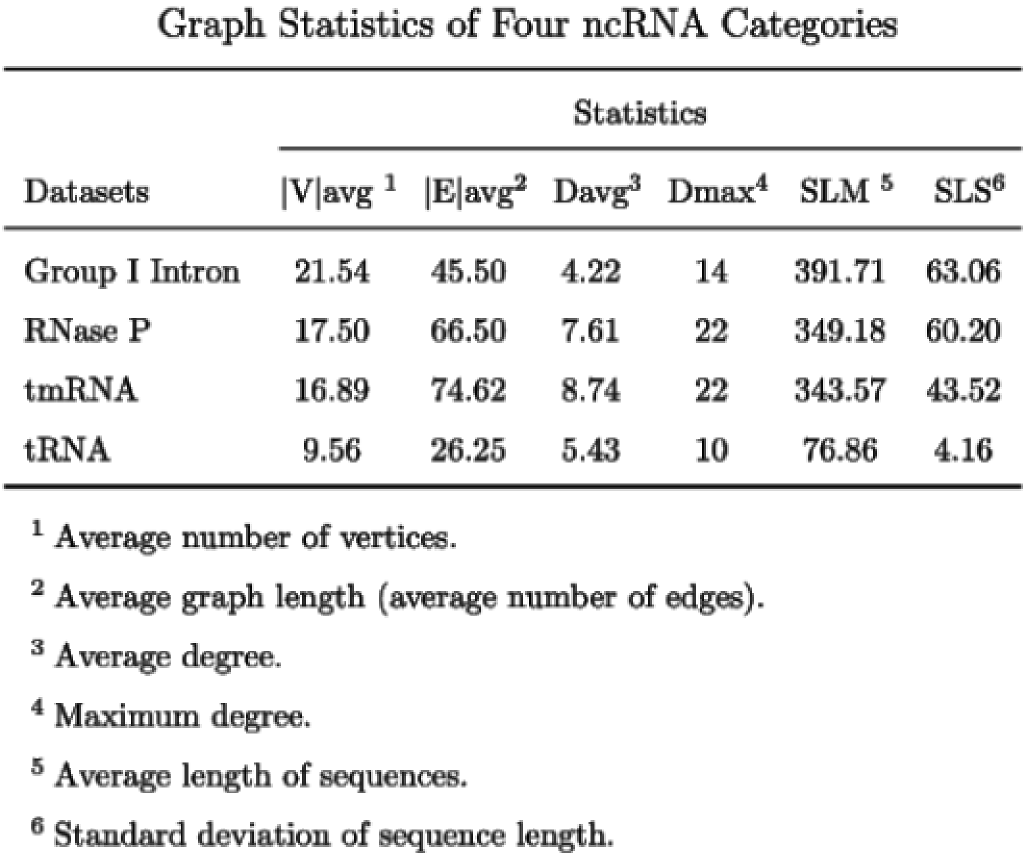
[6] Number of edges in core subgraph.

Cross Validations

cture Lengths

Does Classification Work in Real Life?





Graph Statistics of Four ncRNA Categories

	Statistics					
Datasets	|V|avg [1]	|E|avg[2]	Davg[3]	Dmax[4]	SLM [5]	SLS[6]
Group I Intron	21.54	45.50	4.22	14	391.71	63.06
RNase P	17.50	66.50	7.61	22	349.18	60.20
tmRNA	16.89	74.62	8.74	22	343.57	43.52
tRNA	9.56	26.25	5.43	10	76.86	4.16

[1] Average number of vertices.

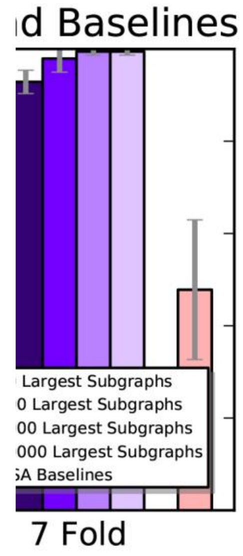
[2] Average graph length (average number of edges).

[3] Average degree.

[4] Maximum degree.

[5] Average length of sequences.

[6] Standard deviation of sequence length.



Cr l Classes

Note:

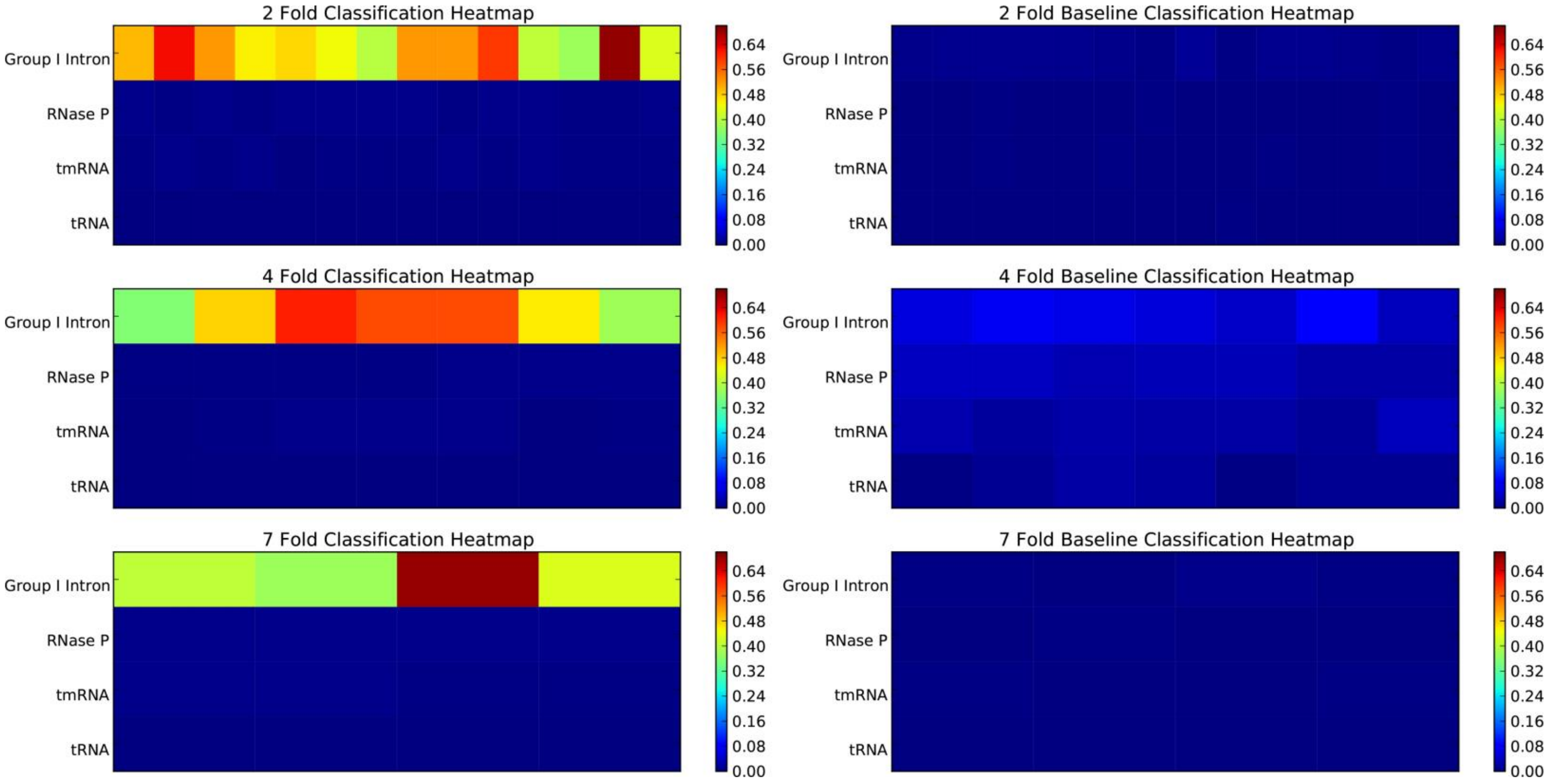
2 fold cross validatio

Each 14 Group I I class as validation

Multiple Sequence Alignment:

Clustal Omega Online Server (http://www.ebi.ac.uk/Tools/msa/clustalo/)

Heatmap Examples



The Heatmap Examples of Cross Validations and Baselines

Definition **Percent Identity of Sequence Alignment (SPI):**

Given a ncRNA sequence S and a sequence alignment A, where $|S|$ is the number of nucleotide in S, and $|A|$ is the number of aligned nucleotide in A.

$$SPI(A,S) = \frac{|A|}{|S|}$$

ncRNA Topological Graph Classification Conclusion

ncRNA topological alignment is able to predict ncRNA’s function

Summary

MMC-Margin

MMC-Margin Identifies Largest Common Substructures

Performance: (Outperforms Well Known FSM algorithms)

- Time Efficient Algorithm
- Little Memory Consumption

ncRNA Topology Alignment and Classification

The ATPI Score Indicates **ncRNA Function** Similarity

Sequence Similarity Is **Less Reliable**

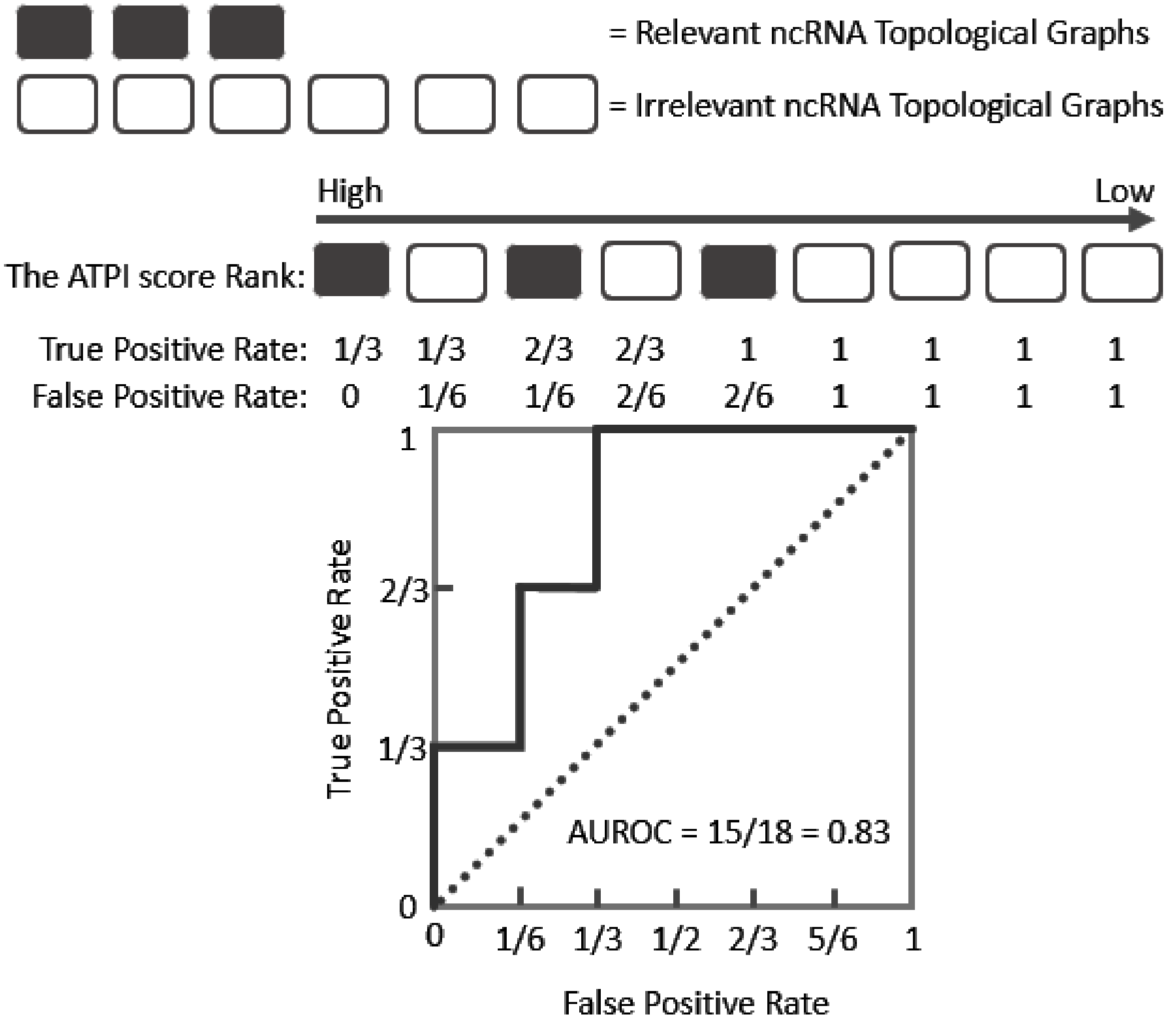
Future Direction

Parallel Implementation, Neighboring Cut Optimization

Classification among Predicted Structures

High Throughput ncRNA Function Prediction Method

Performance Evaluation: The AUROC Score



■ ■ ■ = Relevant ncRNA Topological Graphs

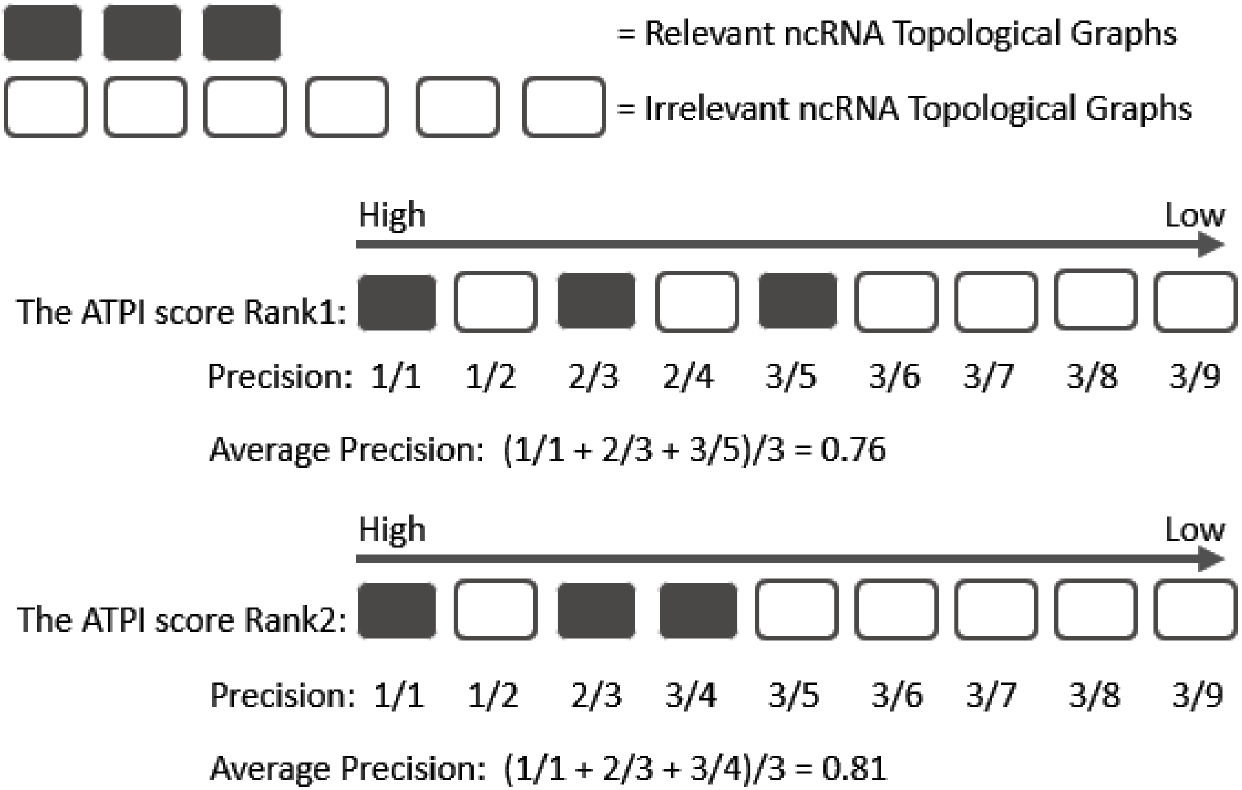
□ □ □ □ □ □ = Irrelevant ncRNA Topological Graphs

High → Low

The ATPI score Rank: ■ □ ■ □ ■ □ □ □ □

True Positive Rate:	1/3	1/3	2/3	2/3	1	1	1	1	1
False Positive Rate:	0	1/6	1/6	2/6	2/6	1	1	1	1

Performance Evaluation: The MAP Score



■ ■ ■ = Relevant ncRNA Topological Graphs

□ □ □ □ □ □ = Irrelevant ncRNA Topological Graphs

The ATPI score Rank1: ■ □ ■ □ ■ □ □ □ □

Precision: 1/1 1/2 2/3 2/4 3/5 3/6 3/7 3/8 3/9

Average Precision: (1/1 + 2/3 + 3/5)/3 = 0.76

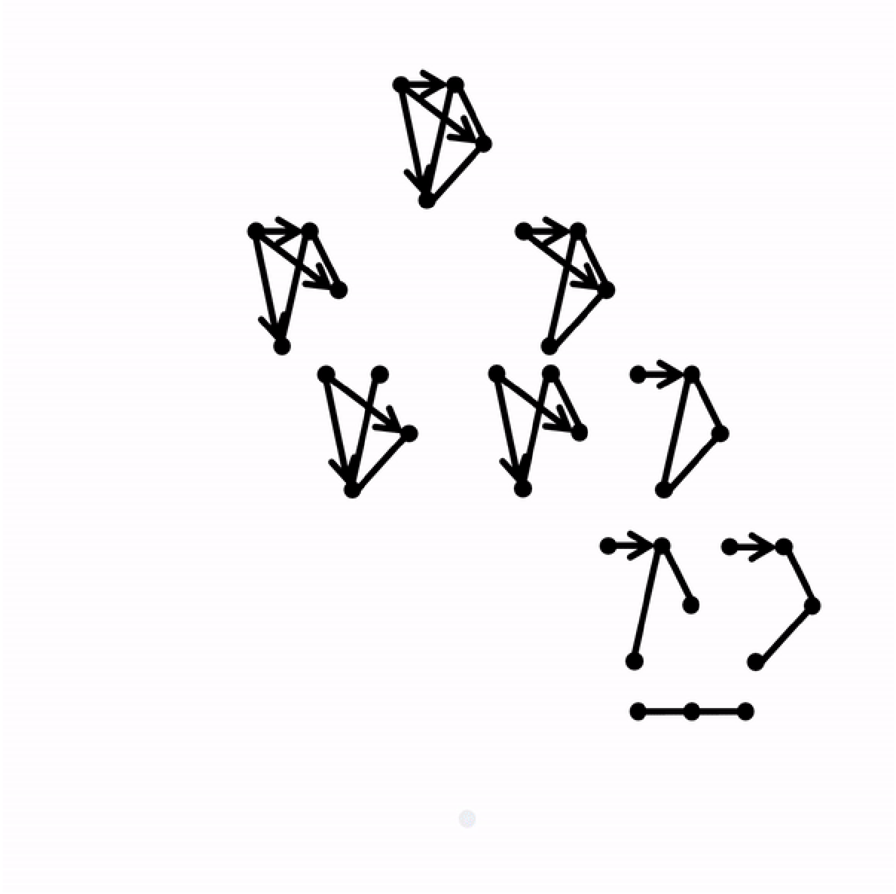
The ATPI score Rank2: ■ □ ■ ■ □ □ □ □ □

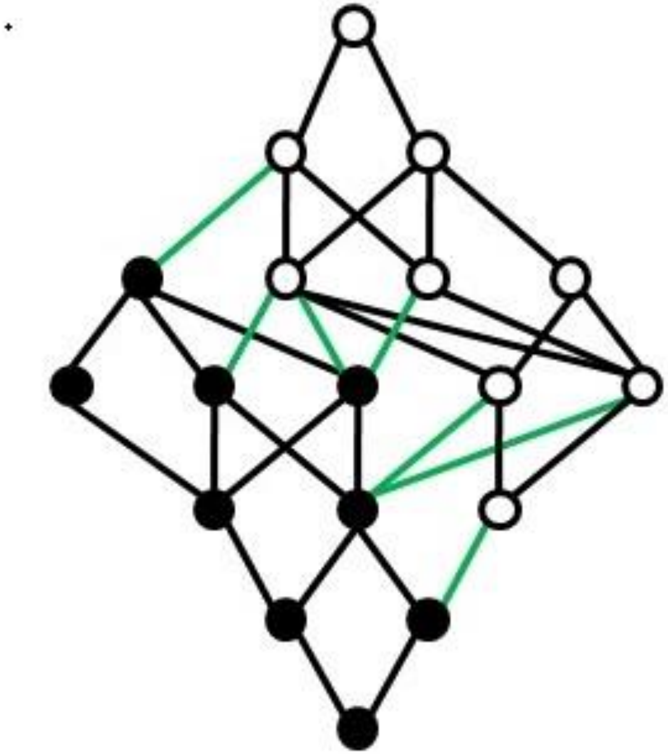
Precision: 1/1 1/2 2/3 3/4 3/5 3/6 3/7 3/8 3/9

Average Precision: (1/1 + 2/3 + 3/4)/3 = 0.81

Mean Average Precision: (0.76 + 0.81)/2 = 0.79

The gSpan Algorithm





The gSpan Algorithm (Yan et al., 2002)