



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

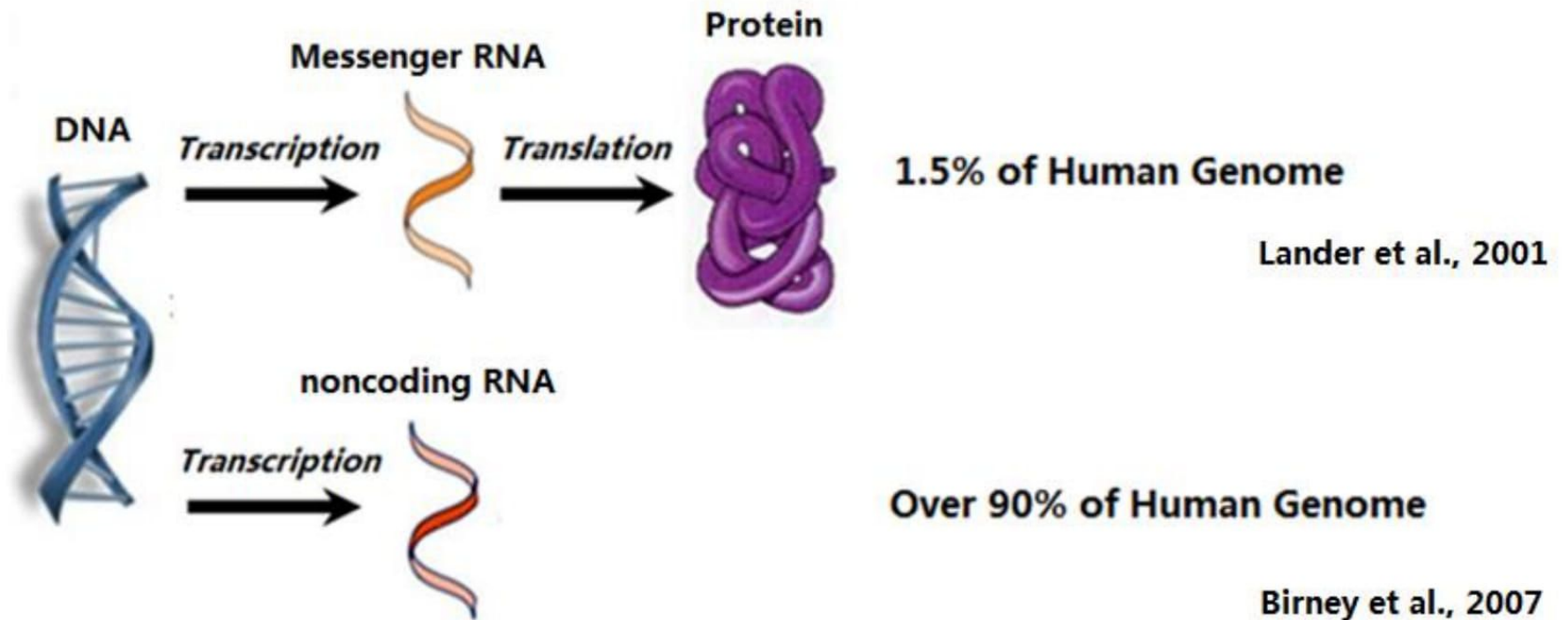
Predict ncRNA's function by common ncRNA topology

Summary

Achievements

Future directions

Introduction



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)

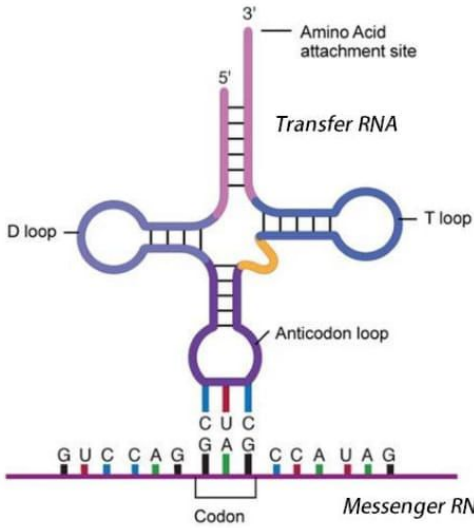
Consortium:

181,047 independent transcripts are reported from mouse transcriptomic data

Estimated mouse genes: 22,000

(Carninci et al., 2005)

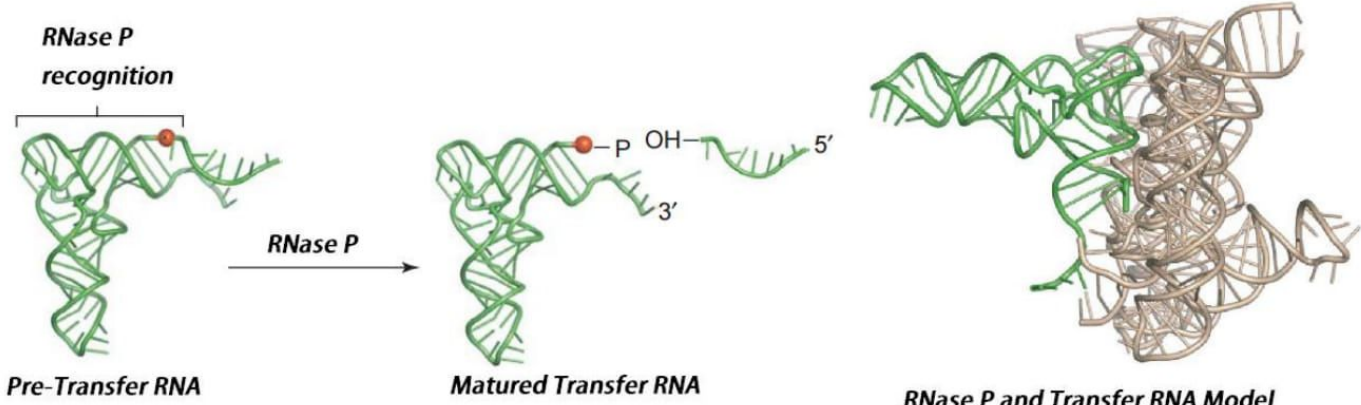
Important Functions of ncRNA



NHGRI Image Gallery

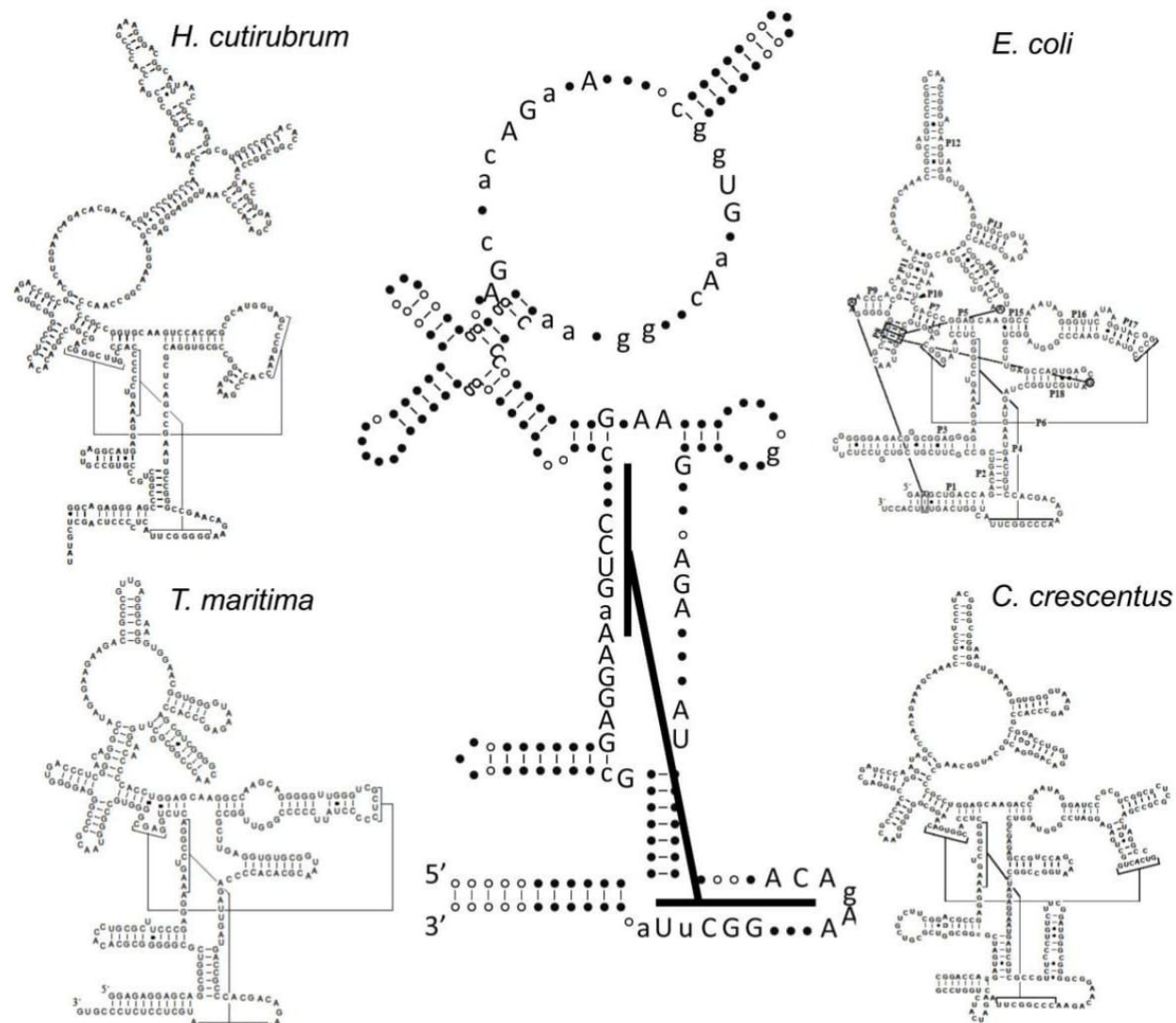
Examples of Determined ncRNA Functions by Observation

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



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

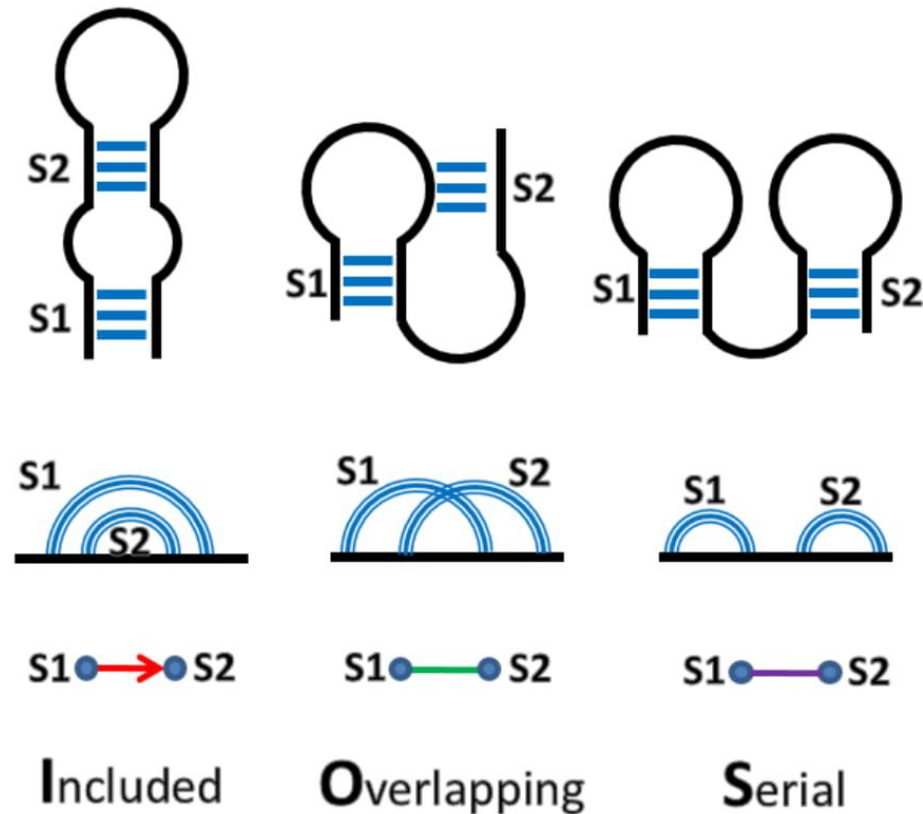
Functional Group	Conserved Stems ¹	Reference
Group I Intron	11	Woodson et al., 2005
RNase P	11	Brown et al., 1995
tmRNA	14	Williams et al., 1996
Telomerase RNA	13	Chen et al., 2000
16s rRNA	~100	Gutell et al., 2002
23s rRNA	~150	Gutell et al., 2002

¹ Number of conserved stems for each ncRNA function category

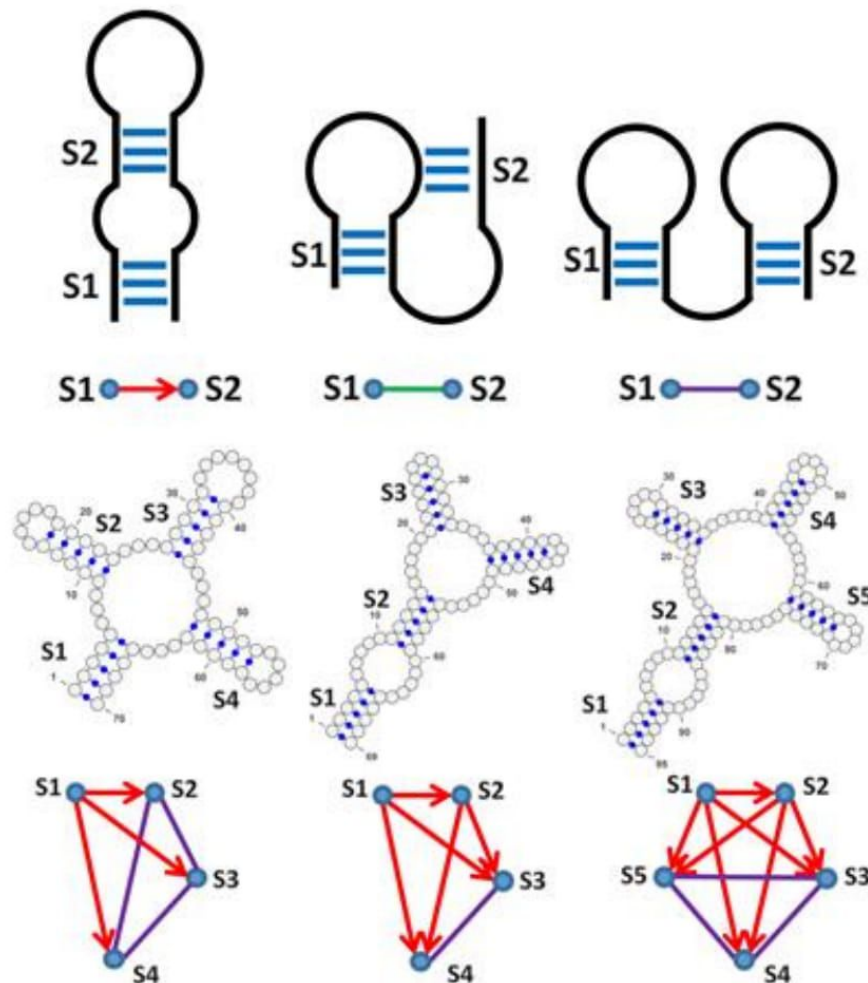
Preliminary Concept

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

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

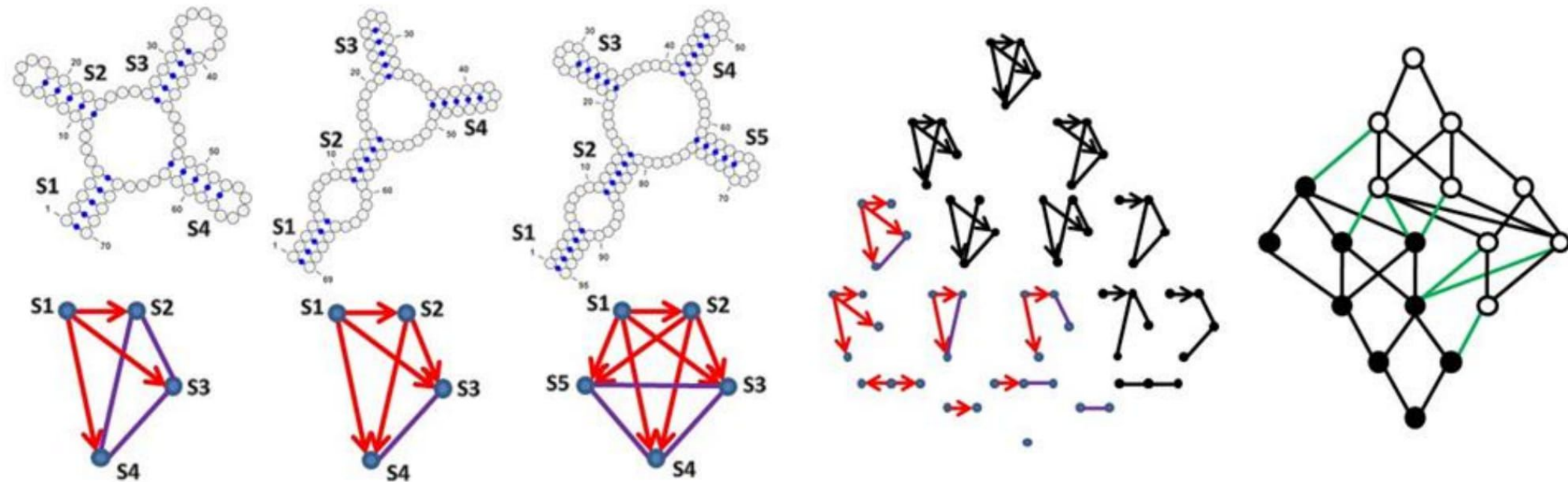
graph length: $|G|=|E|$

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

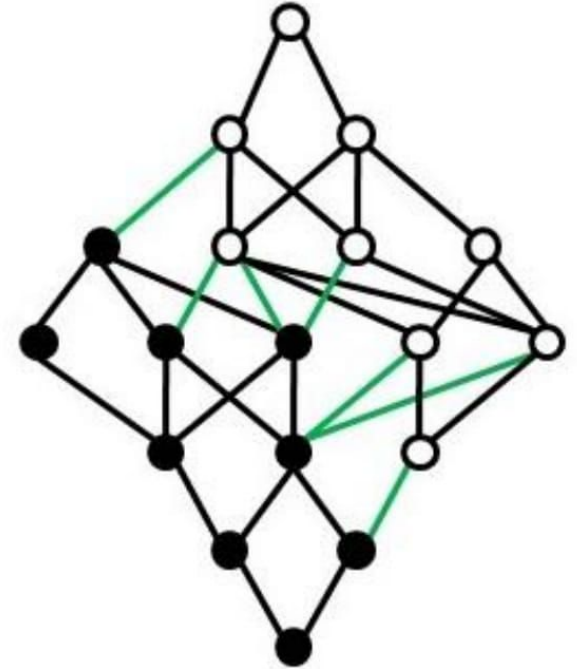
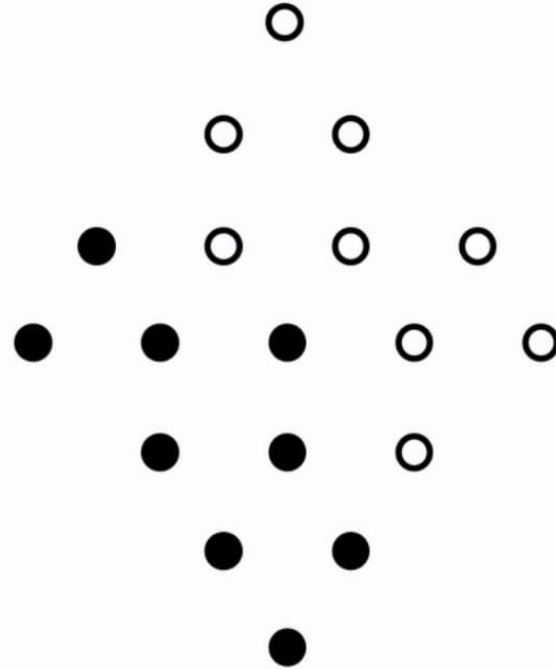
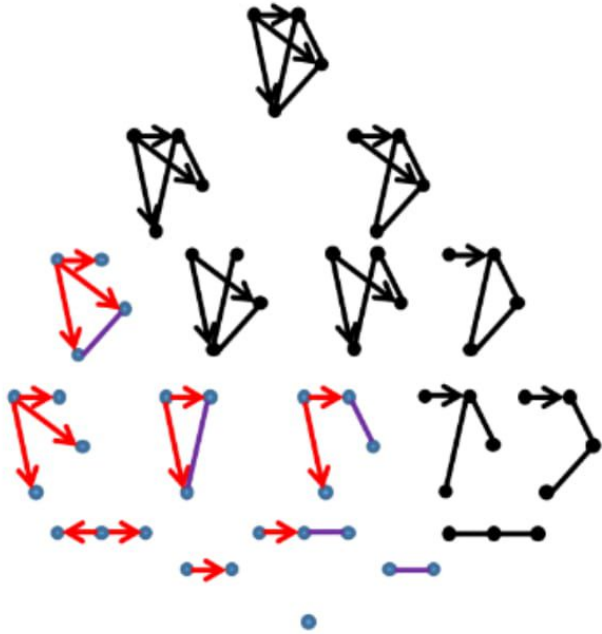
Several Well Known Frequent Subgraph Mining Algorithms

Type	Algorithm	Search Strategy	Reference
<i>A priori</i> -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- <i>a priori</i> -based:			
	Margin	Maximal Subgraphs Search	Thomas et al., 2006
	FS3	Fixed Size Subgraphs Sampling	Saha et al., 2014

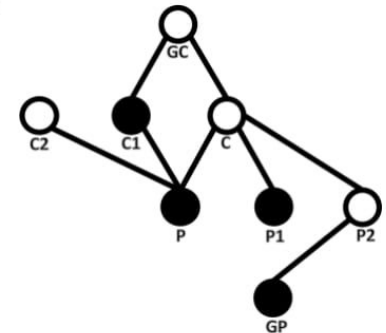
¹ Breadth-first Search

² Depth-first Search

The Margin Algorithm



The Margin Algorithm (Thomas et al., 2006)



Neighboring Cuts 15

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 $\approx 10^9$ seconds and estimated universe age $\approx 10^{17}$ 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

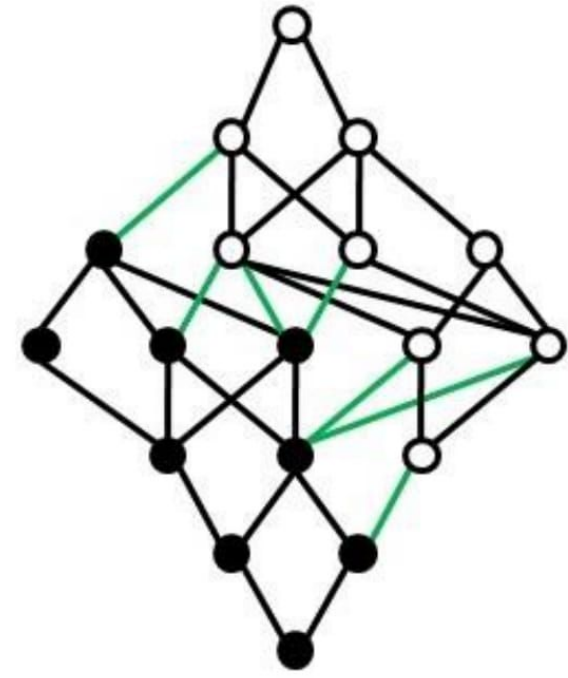
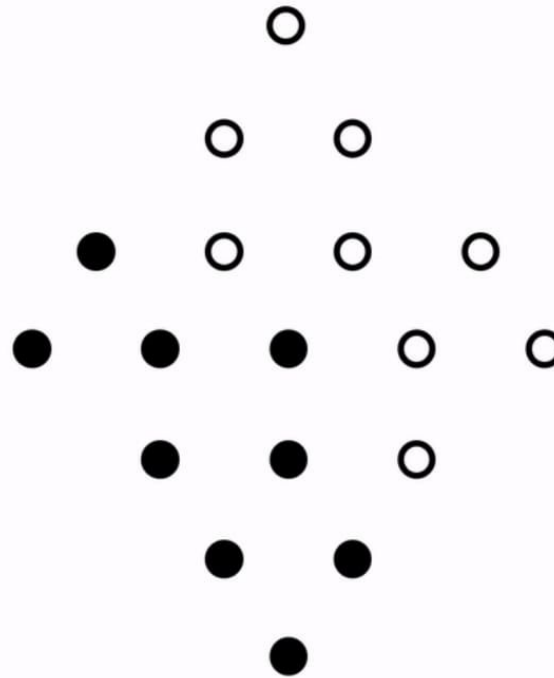
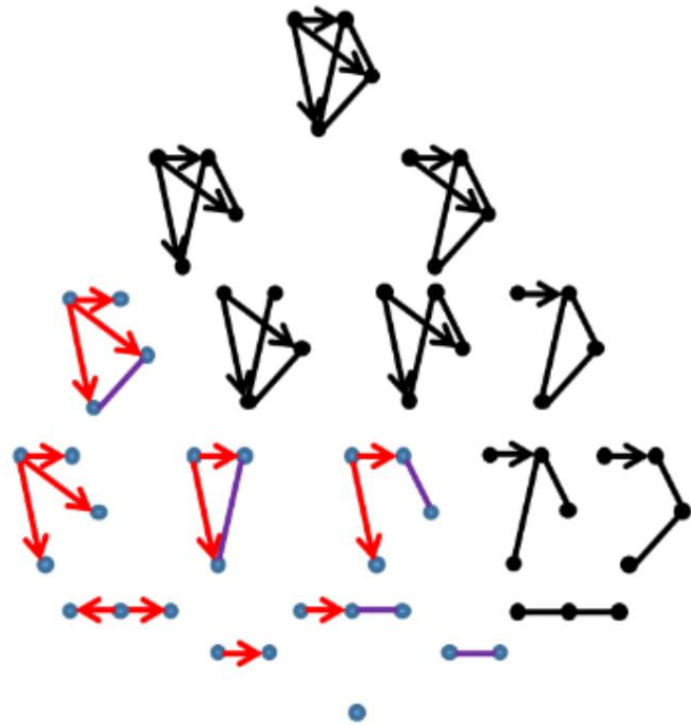
(Garey et al., 1979)

(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 $\mathbb{G} = \{G_1, G_2, \dots, G_n\}$

OUTPUT: Maximum Frequent Subgraphs: $\text{MFS} \in G_1 \cap G_2 \cap \dots \cap G_n$

- 1: $\text{MFS} = \emptyset, C \uparrow P = \emptyset$
- 2: $(C \uparrow P) = \text{FindInitialCut}(G_{\min}, \mathbb{G})$
- 3: $\text{SampleCut}(\text{MFS}, C \uparrow P)$

Algorithm 2 FindInitialCut

INPUT: G_{\min}, \mathbb{G}

OUTPUT: $C \uparrow P$

- 1: $C = G_{\min}$
- 2: $P = \text{RemoveOneEdge}(C)$
- 3: **while** P is infrequent in \mathbb{G} **do**
- 4: $C = P$
- 5: $P = \text{RemoveOneEdge}(P)$

Algorithm 3 SampleCut

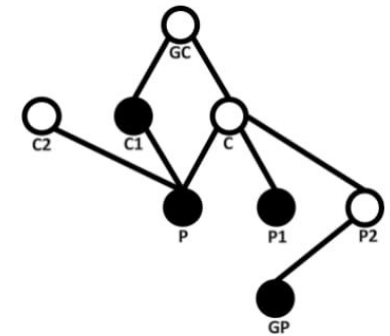
INPUT:

$C \uparrow P$

OUTPUT:

MFS

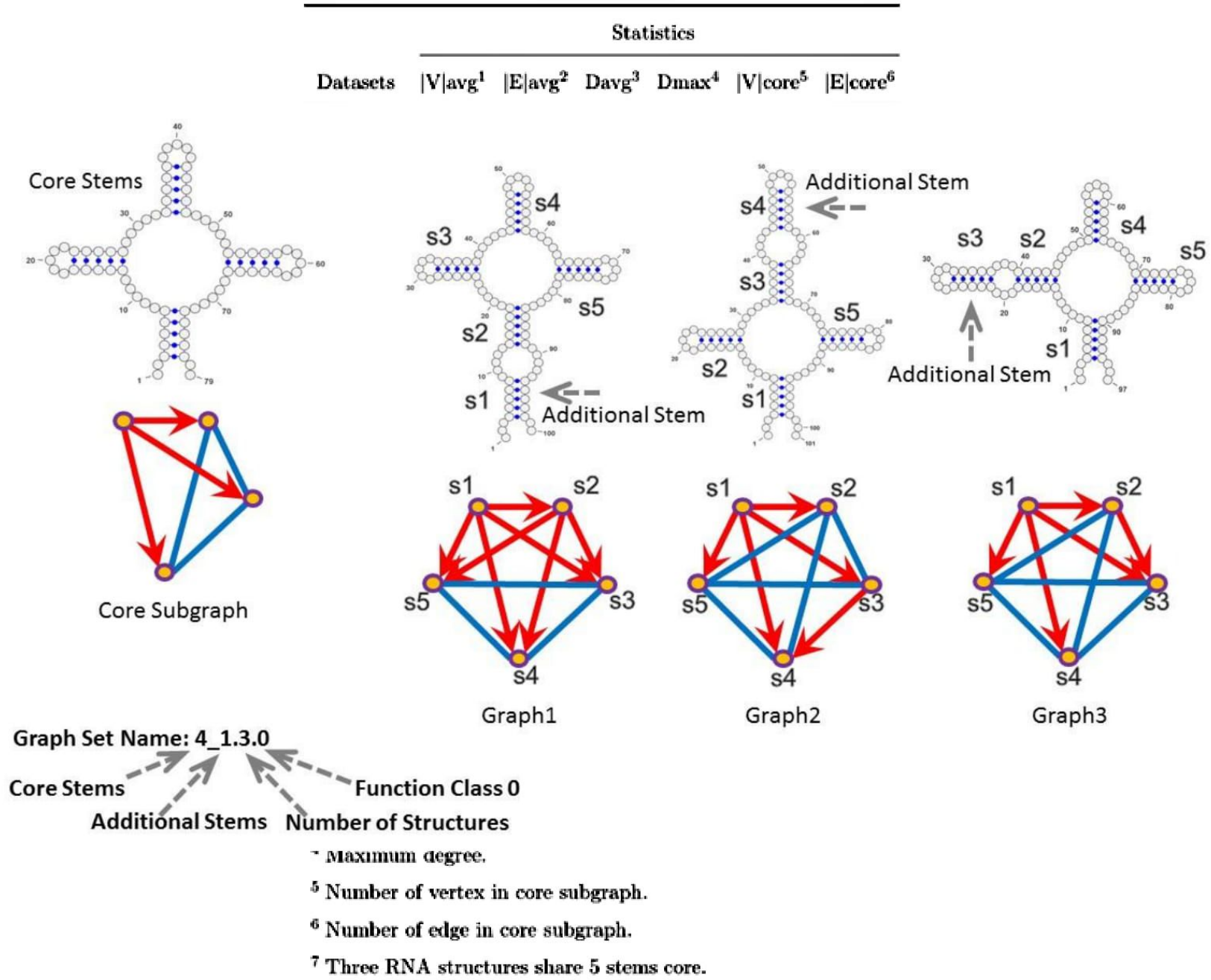
- 1: **while** update of current candidate MFS is frequent **do**
- 2: $C_{\text{new}} \uparrow P_{\text{new}} = \emptyset$
- 3: Choose MoveType randomly
- 4: Find Neighbor $C_{\text{new}} \uparrow P_{\text{new}}$ of $C \uparrow P$ by MoveType
- 5: **if** $|P_{\text{new}}| > |P|$ **then**
- 6: $C \uparrow P = C_{\text{new}} \uparrow P_{\text{new}}$
- 7: Update MFS by P_{new}
- 8: **if** $|P_{\text{new}}| \leq |P|$ **then**
- 9: Accept $C_{\text{new}} \uparrow P_{\text{new}}$ by EEAP or HEAP
- 10: Update MFS by P_{new}



Neighboring Cuts

ncRNA Structure Generator

Synthetic RNA Graphsets Statistics:



MMC-Margin Outperforms Margin

Margin Performance on Synthetic Datasets

Datasets	Margin (600 hours)				
	#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

Datasets	MMC-Margin (600 hours)		
	#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.