Structural Alignment of Pseudoknotted RNAs

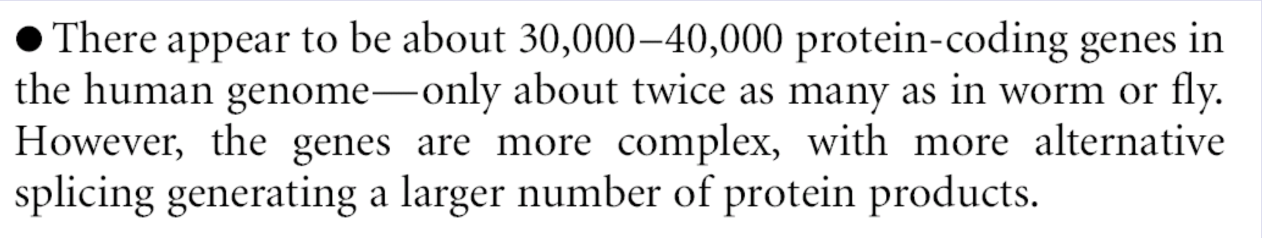
Banu Dost, Buhm Han,
Shaojie Zhang,
Vineet Bafna

Non-coding RNAs are mostly undetected

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

[*Nature* 2001]



● There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.

Non-coding RNAs (ncRNAs) might be playing a significant role in many cellular functions.

ncRNA --- RNA acts as functional molecule, and is not translated into protein.

Modern RNA world hypothesis:

There are many undetected functional ncRNAs. [Eddy *Nature Reviews* (2001)]

There may be many ncRNAs behind many unexplained phenomenon.[Storz *Science* 2002]

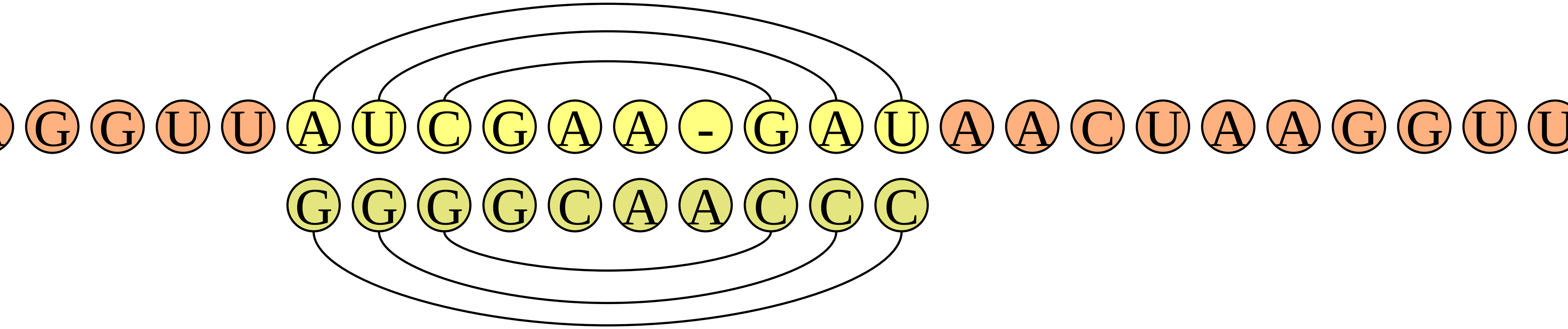
How can we discover ncRNA genes?

- Low-energy Stability Approach: Are they the substrings that fold into stable low-energy structures?
    - No. The stability of ncRNA secondary structure is not sufficiently different from the predicted stability of a random sequence. [Rivas and Eddy *Bioinformatics* (2000)].

- Comparative Approach: Are they the substrings that are similar to known ncRNAs in sequence and structure?

ncRNA Discovery: Comparative Approach

RNA Local Alignment Problem: Given a non-coding RNA as query, can you find all subsequences in the genomic database that are similar to the query in both sequence and secondary structure?

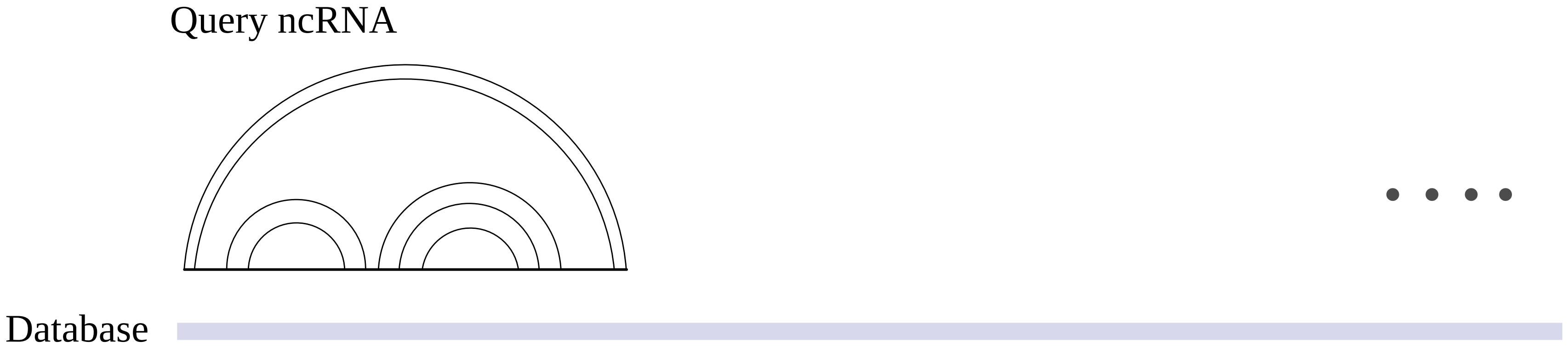


ncRNA Discovery: Previous Work

RSEARCH [Klein and Eddy *BMC Bioinformatics* (2003)]

FASTR [Bafna and Zhang *CSB* (2004)]

The query ncRNA with known secondary structure is compared to every subsequence in a database.

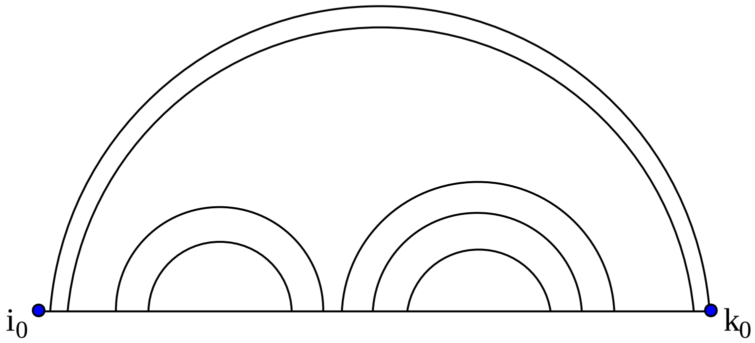


Problem: Can not handle pseudo-knotted structures.

- RNA alignment problem has been solved for RNAs with a regular structure, i.e. non-pseudo-knotted structures.

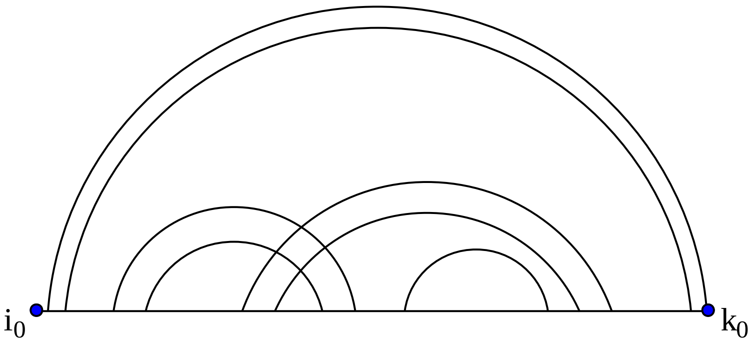
Regular Structure

All of the base pairs are non-crossing.



Pseudo-knotted Structure

Some of the base pairs are crossing.

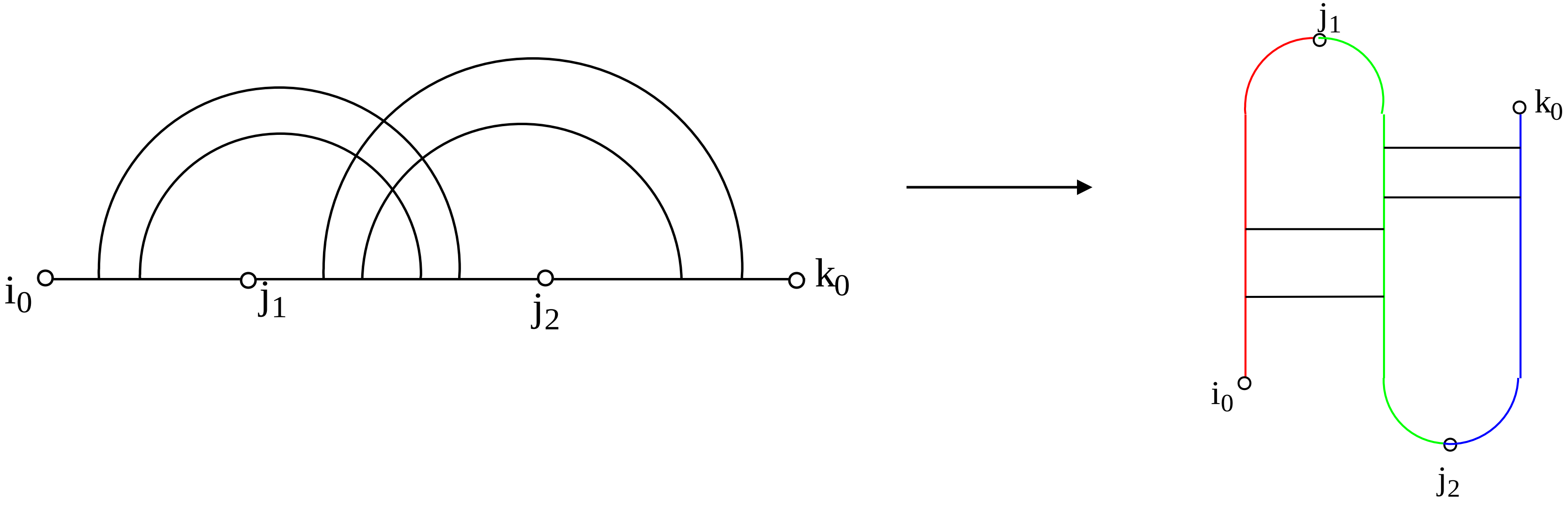


Objective

- Extend the Bafna and Zhang's algorithm to solve the problem for also the pseudo-knotted structures.
  - Dynamic programming technique used to align subsequences.
  - Challenge: Design a substructure for the suboptimal solutions valid for the pseudo-knotted structures.

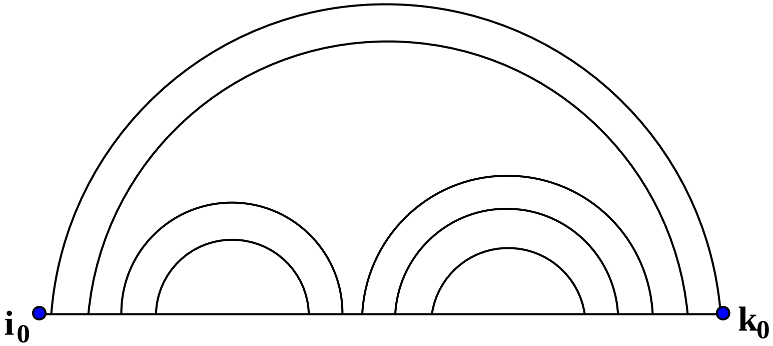
Definition: Simple Pseudo-knot

- All base pairs non-crossing and horizontal when rotated to form 2 loops.



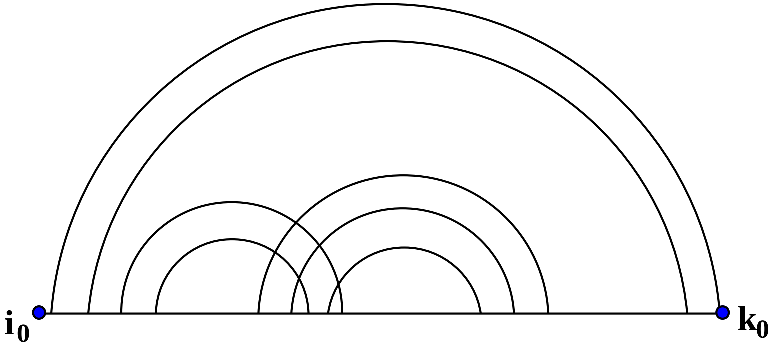
Substructure for Sub-optimal Solutions of a Simple Pseudoknot

- Regular structure: continuous subintervals as substructure of recursion.



- Simple Pseudo-knot:

  can not use this substructure due to interweaving base pairs.

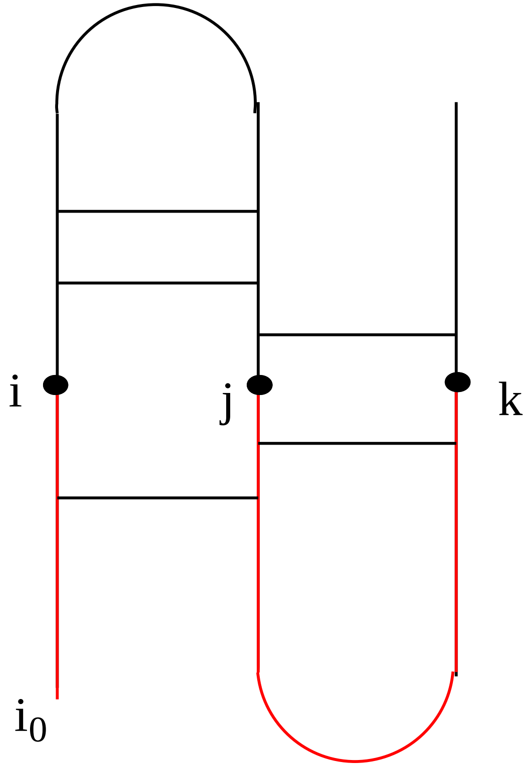


Substructure for Simple Pseudo-knots

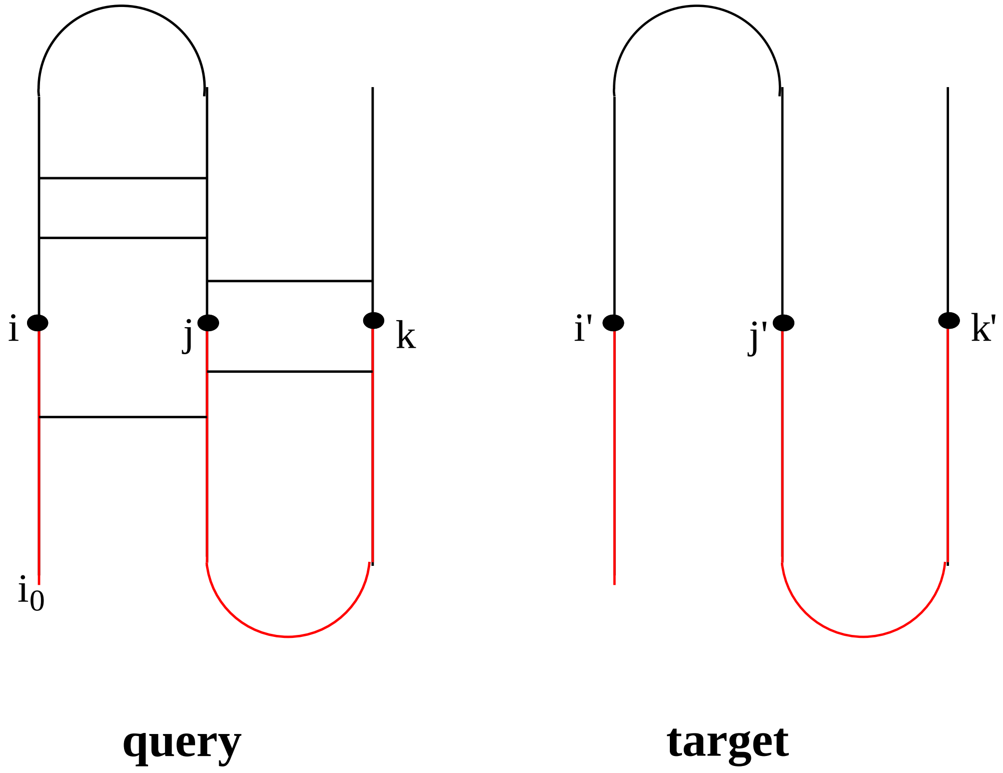
subpseudoknot P(i, j, k) as the union of two subintervals

$$P(i, j, k) = [i_0, i] \text{ U } [j, k]$$

frontier (i.j.k)



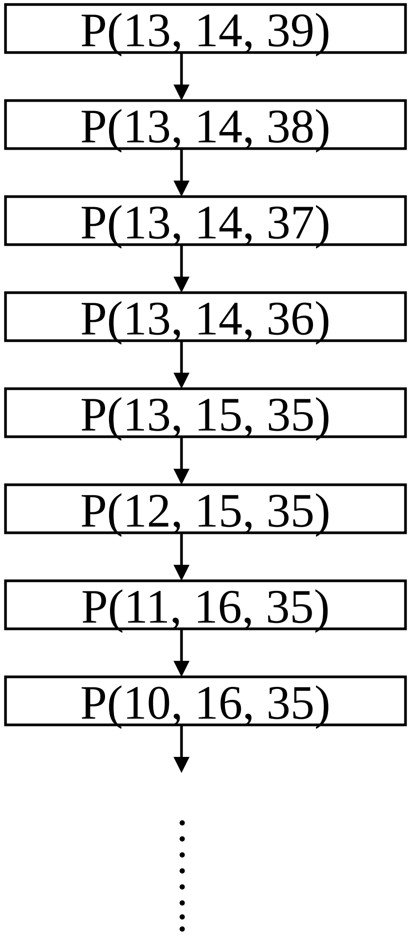
Naive Approach



- Compute B[i, j, k, i’, j’, k’] ⇒$O(m^3n^3)$ scores. (m:query, n:target)

Instead of all triplets in the query, consider only the valid sub-pseudo-knots that will represent the simple pseudo-knot.

Use a chain of sub-pseudoknots to represent Simple Pseudo-knot



P(13, 14, 39)

P(13, 14, 38)

P(13, 14, 37)

P(13, 14, 36)

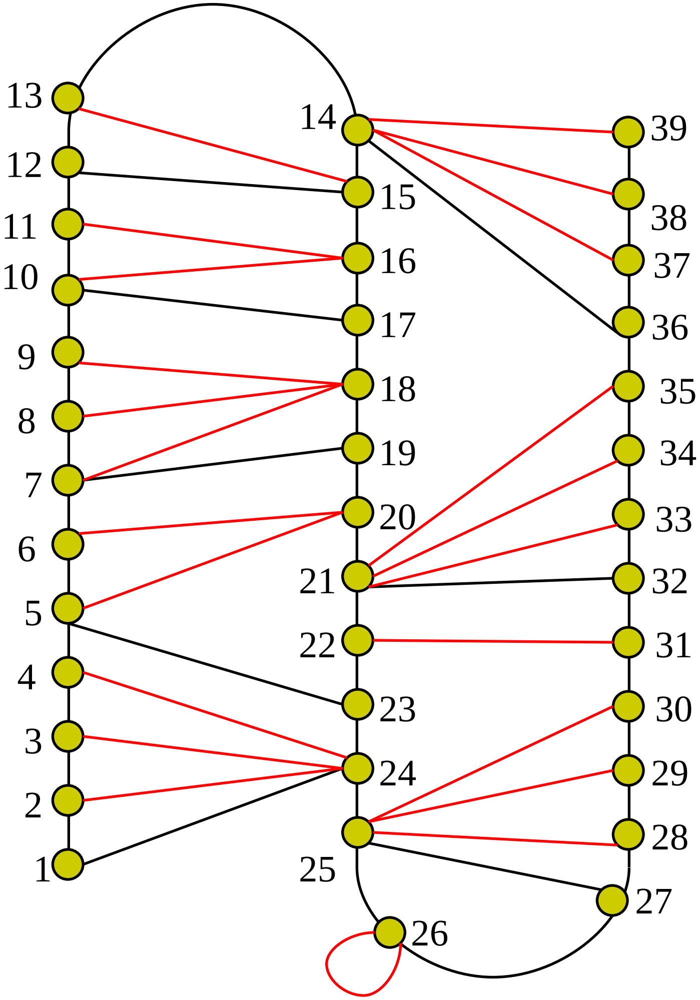
P(13, 15, 35)

P(12, 15, 35)

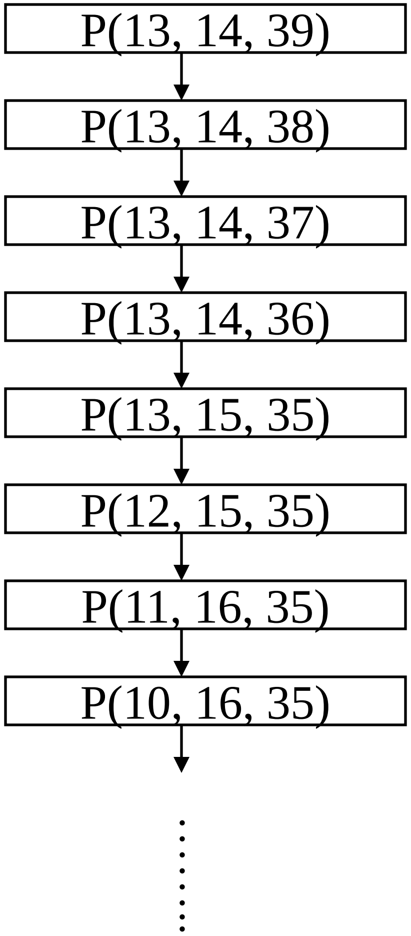
P(11, 16, 35)

P(10, 16, 35)

……



Why Chaining?



P(13, 14, 39)

P(13, 14, 38)

P(13, 14, 37)

P(13, 14, 36)

P(13, 15, 35)

P(12, 15, 35)

P(11, 16, 35)

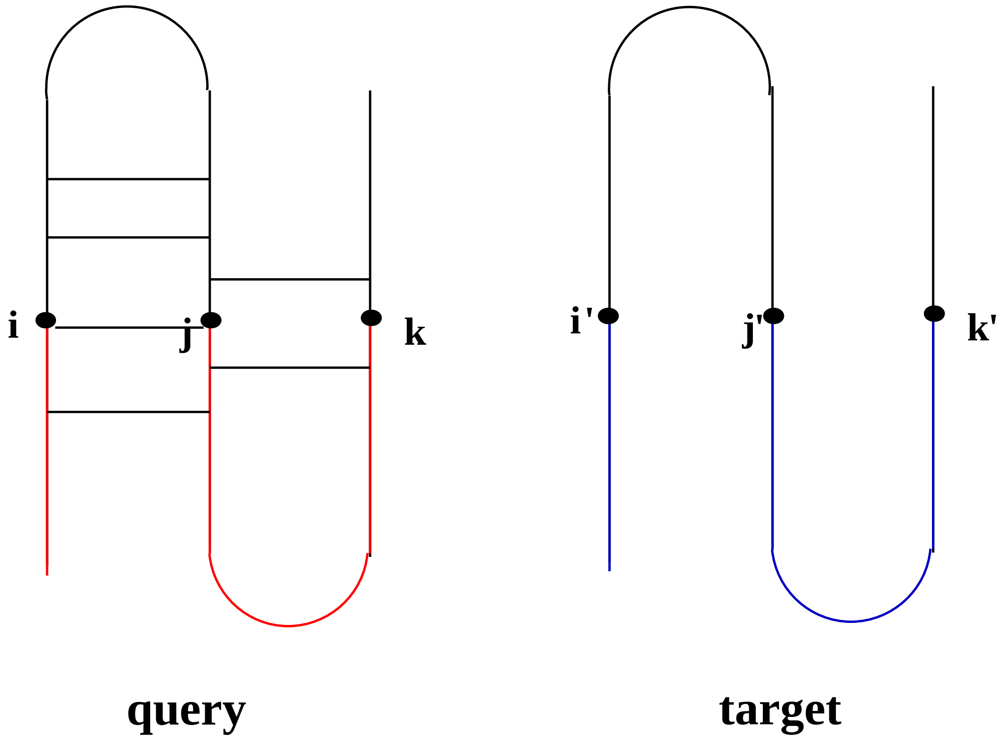
P(10, 16, 35)

- DP: use sub-optimal solution of the child sub-structure to compute optimal score at each step.
- compute B[i,j,k, i',j', k'] => $O(mn^3)$ scores

  (m:query, n:target)

Alignment Algorithm Recursions: (i,j) is a base pair case

B[i, j, k , i’, j’, k’] = max {MATCH, INSERT, DELETE}



- MATCH:
  - (i,j) and (i’, j’) are corresponding pairs
- DELETE:
  - i is deleted
  - j is deleted
  - i and j are deleted
- INSERT:
  - i’ is inserted
  - j’ is inserted
  - i’ and j’ are inserted

Alignment Algorithm Recursions: (i,j) is a base pair case

B[i, j, k , i’, j’, k’] = max {MATCH, INSERT, DELETE}

(i,j) &(i’, j’) are pairs

$$\text{MATCH} = B[i-1, j+1, k, i'-1, j'+1, k] + \delta(q[i'], q[j'], t[i], t[j]) + \gamma(q[i'], t[i]) + \gamma(q[j'], t[j]),$$

j deleted

i deleted

i & j deleted

$$\text{DELETE} = \max \begin{cases} B[i-1, j, k, i'-1, j'+1, k] + \gamma(q[i'], t[i]) + \gamma(q[j'], '-'), \\ B[i, j+1, k, i'-1, j'+1, k] + \gamma(q[i'], '-') + \gamma(q[j'], t[j]), \\ B[i, j, k, i'-1, j'+1, k] + \gamma(q[i'], '-') + \gamma(q[j'], '-') \end{cases}$$

i’ inserted

j’ inserted

i’&j’ inserted

$$\text{INSERT} = \max \begin{cases} B[i-1, j, k, i', j', k'] + \gamma('-', t[i]), \\ B[i, j+1, k, i', j', k'] + \gamma('-', t[j]), \\ B[i, j, k-1, i', j', k'] + \gamma('-', t[k]) \end{cases}$$

Time Complexity: to align to a simple pseudo-knot

- m: query length, n: target length
- #sub-pseudoknots in query: O(m)
- #sub-pseudoknots in target (i_0,k_0) : $O(n^3)$
- Time to align (i_0,k_0) to a simple pseudoknot
  - Do alignment for all subintervals (i,k_0) = O(n) x $O(mn^3) = O(mn^4)$

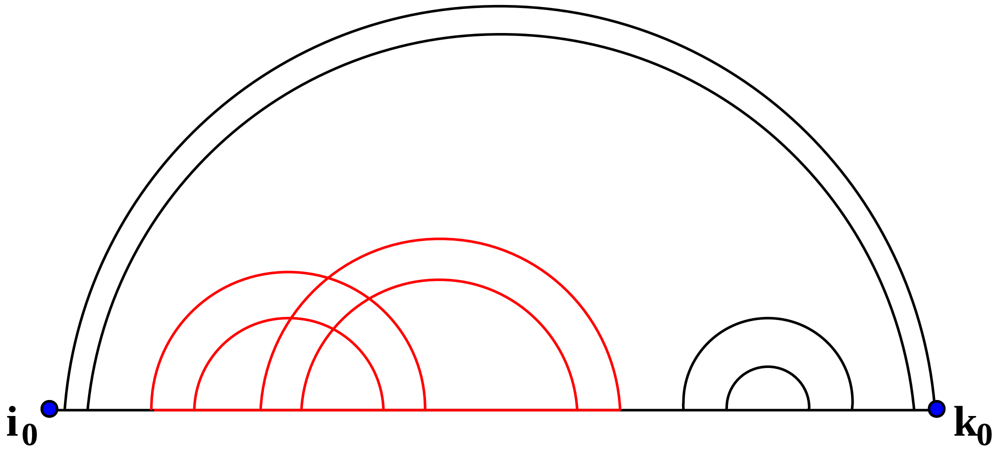
Simple Pseudo-knot in a Regular Structure: S in R

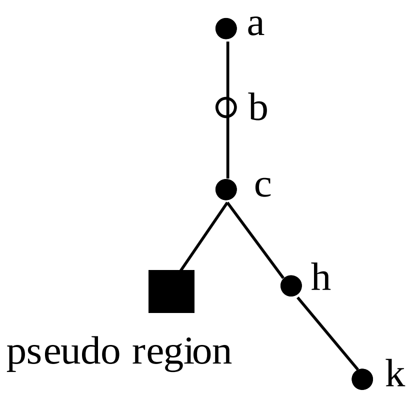
Use a binary tree to represent RNA

Solid circular nodes correspond to the actual base pairs.

Empty circular nodes correspond to unpaired bases.

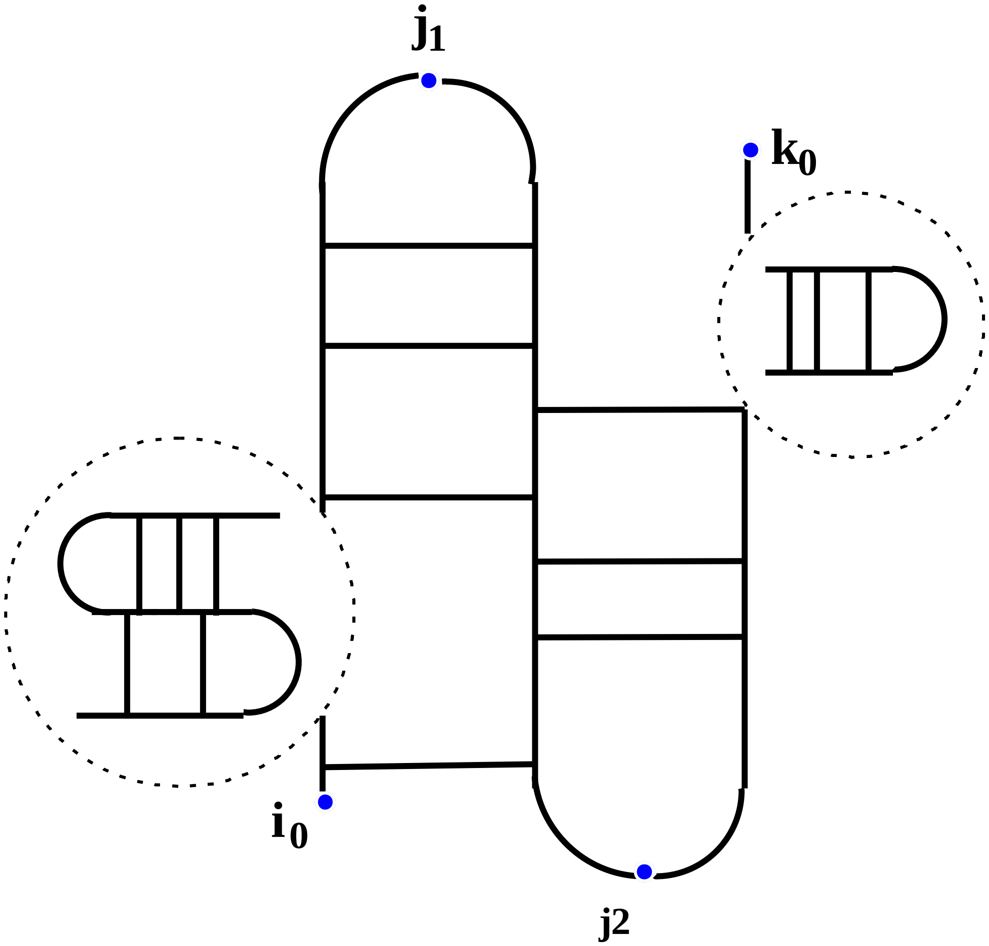
Rectangular node correspond to subtree representing pseudo-knotted region





Simple Pseudo-knot in a Simple Pseudo-knot: Recursive Simple Pseudo-knot

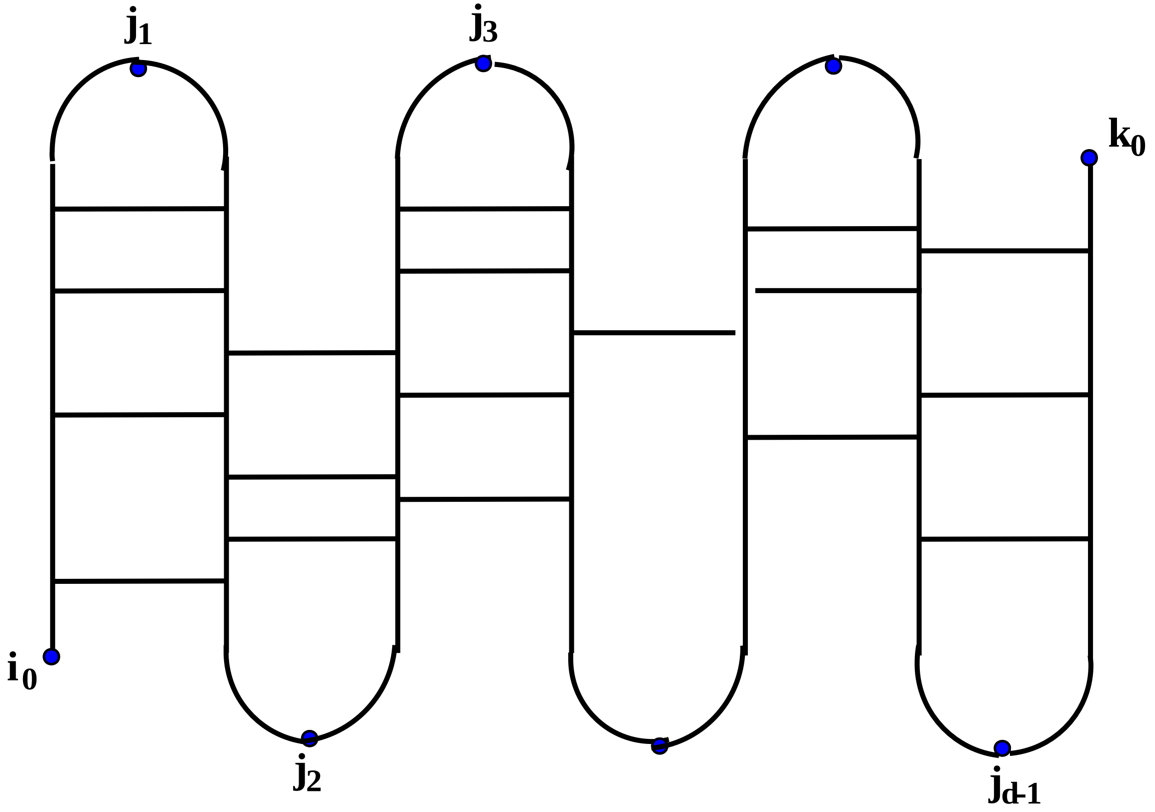
- S in S
- R in S



Which structures can we handle?

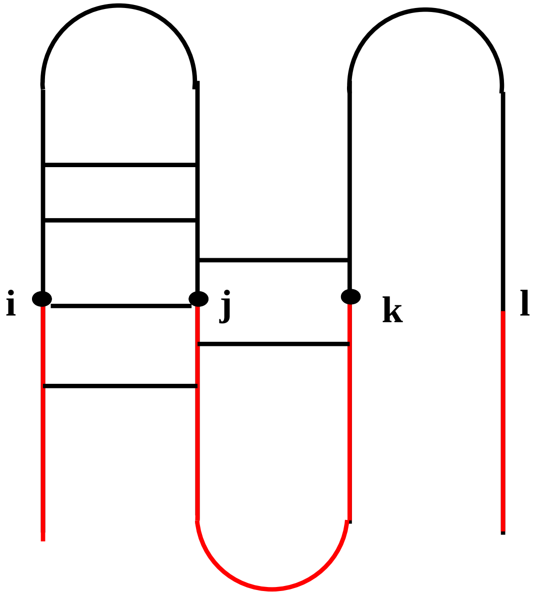
- Time complexity increases with the number of pseudo-knotted region!
- R: regular structure, S: simple pseudo-knot
  - R: $O(mn^3)$
  - S: $O(mn^4)$
  - S in R: $O(mn^4)$
  - R in S: $O(mn^5)$
  - R in S in R: $O(mn^5)$ = S in S in R: $O(mn^5)$.
  - R in S in R in S in R = $O(mn^5)$.
  - …….

Can we handle simple pseudo-knots with higher degree: standard pseudo-knots?

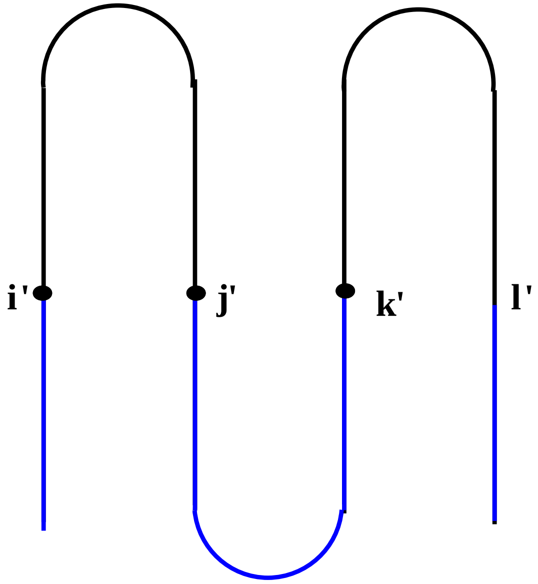


Can we handle simple pseudo-knots with higher degree: standard pseudo-knots?

Yes! By revising the sub-pseudoknot structure and the recursion cases accordingly.



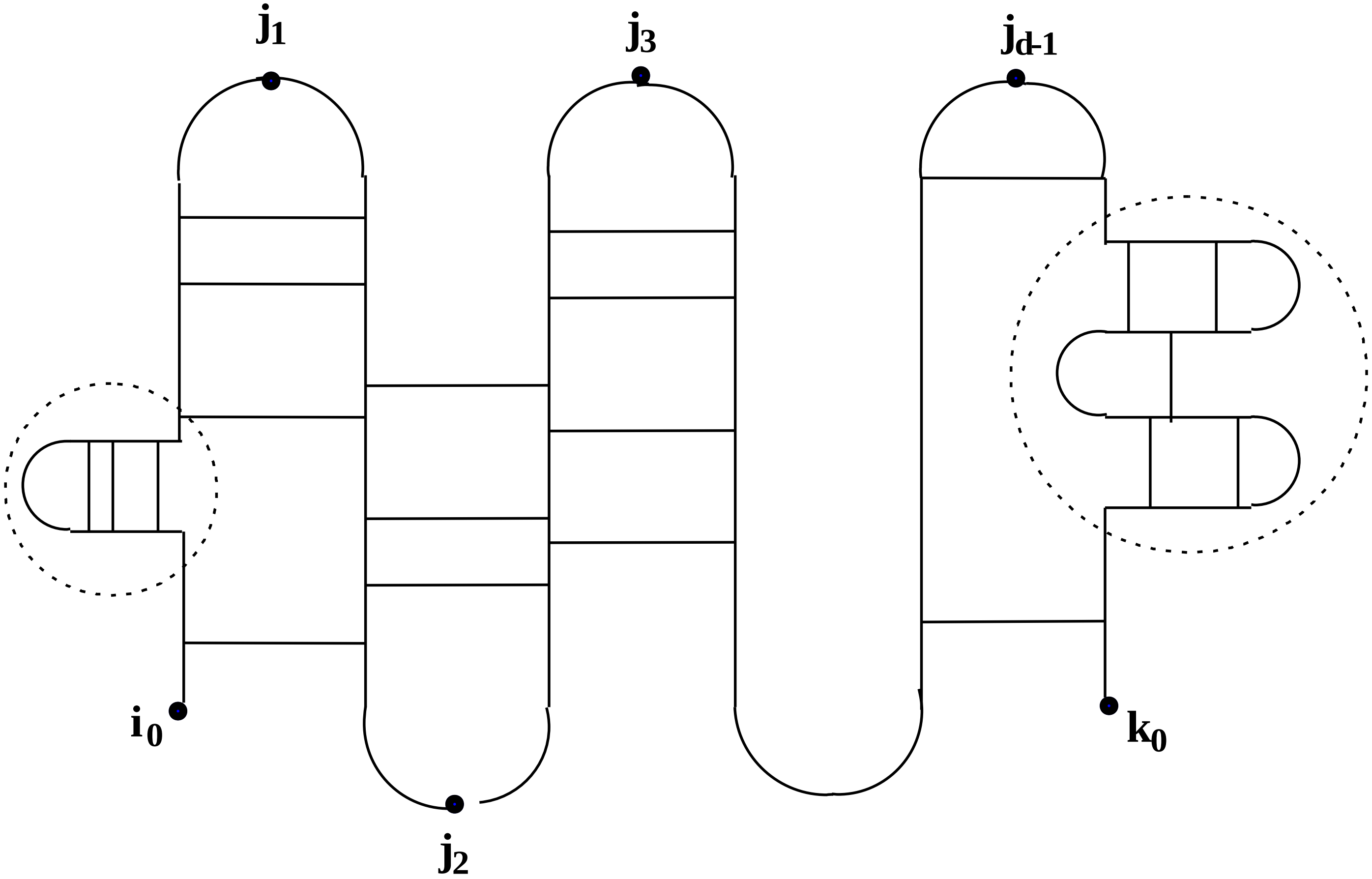
query



target

Can we handle recursive standard pseudoknots?

Yes! Same reasoning with recursive simple pseudoknots.

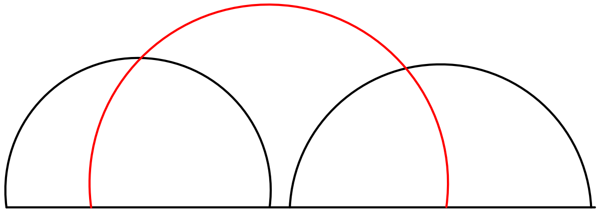


What is left? What can we NOT handle?

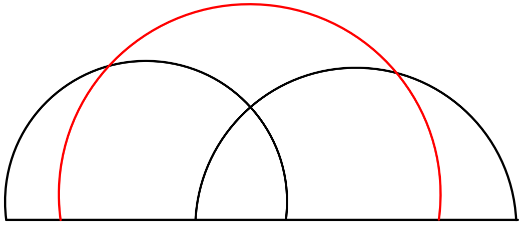
We can handle the class of pseudoknots defined by Akutsu which is the second largest class currently defined. We can additionally handle standard and recursive standard pseudoknots which are defined by us.

A&U <= A&U U {standard/recursive standard pseudoknots} <= R&E

The largest class is defined by Rivas and Eddy. An example from this class we can not handle:



We can handle this!
(Standard pseudo-knot of degree 4)



We can NOT handle this!

Implementation: PAL

- C++ implementation of our algorithm.
  - input:
    - a query sequence with known structure

      (R/S/S in R)
    - a target sequence
  - output:
    - all high scoring local alignments in the target sequence

Testing

- Test Data:
  - RFAM database, 6 RNA families with simple pseudo-knotted structures.

    (simple pseudo-knots in regular structure)
    - UPSK
    - Antizyme
    - Corona FSE
    - Corona pk3
    - Parecho CRE
    - IFN gamma

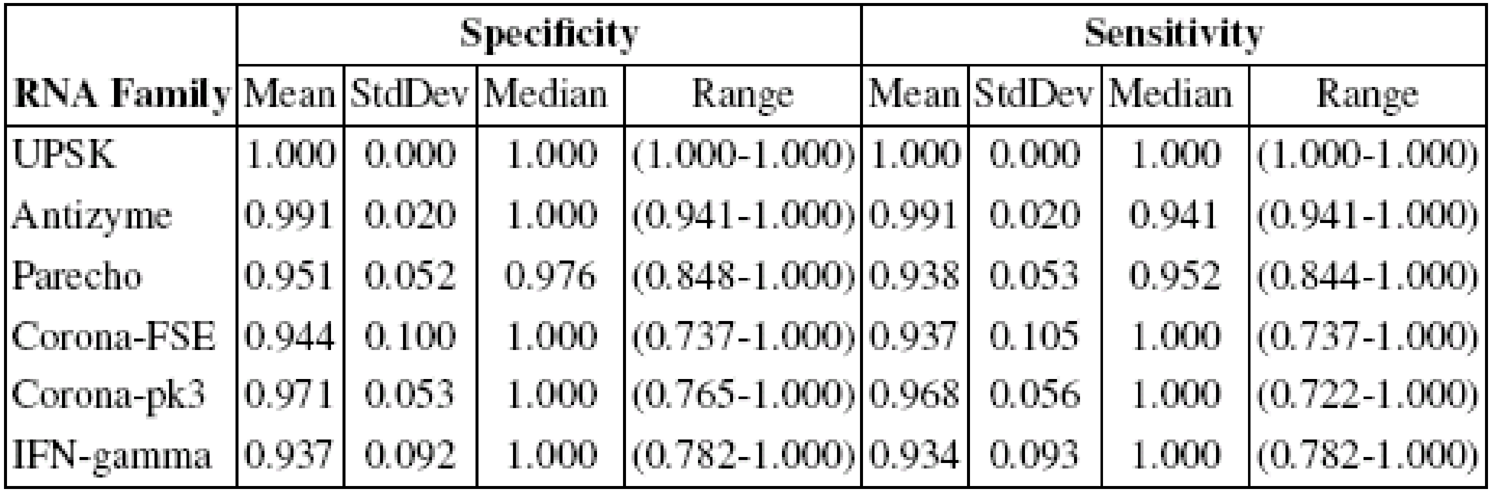
Test 1: Structure Prediction

How good is PAL in inferring structure of the target sequence?

- Pick 2 seed members of an RNA family as query and target.
- Align them.
- Compare the inferred structure of target with annotated structure in Rfam.

Test 1: Structure Prediction Results

- TP, FP, FN, Sensitivity, Specificity



RNA Family	Specificity				Sensitivity			
	Mean	StdDev	Median	Range	Mean	StdDev	Median	Range
UPSK	1.000	0.000	1.000	(1.000-1.000)	1.000	0.000	1.000	(1.000-1.000)
Antizyme	0.991	0.020	1.000	(0.941-1.000)	0.991	0.020	0.941	(0.941-1.000)
Parecho	0.951	0.052	0.976	(0.848-1.000)	0.938	0.053	0.952	(0.844-1.000)
Corona-FSE	0.944	0.100	1.000	(0.737-1.000)	0.937	0.105	1.000	(0.737-1.000)
Corona-pk3	0.971	0.053	1.000	(0.765-1.000)	0.968	0.056	1.000	(0.722-1.000)
IFN-gamma	0.937	0.092	1.000	(0.782-1.000)	0.934	0.093	1.000	(0.782-1.000)

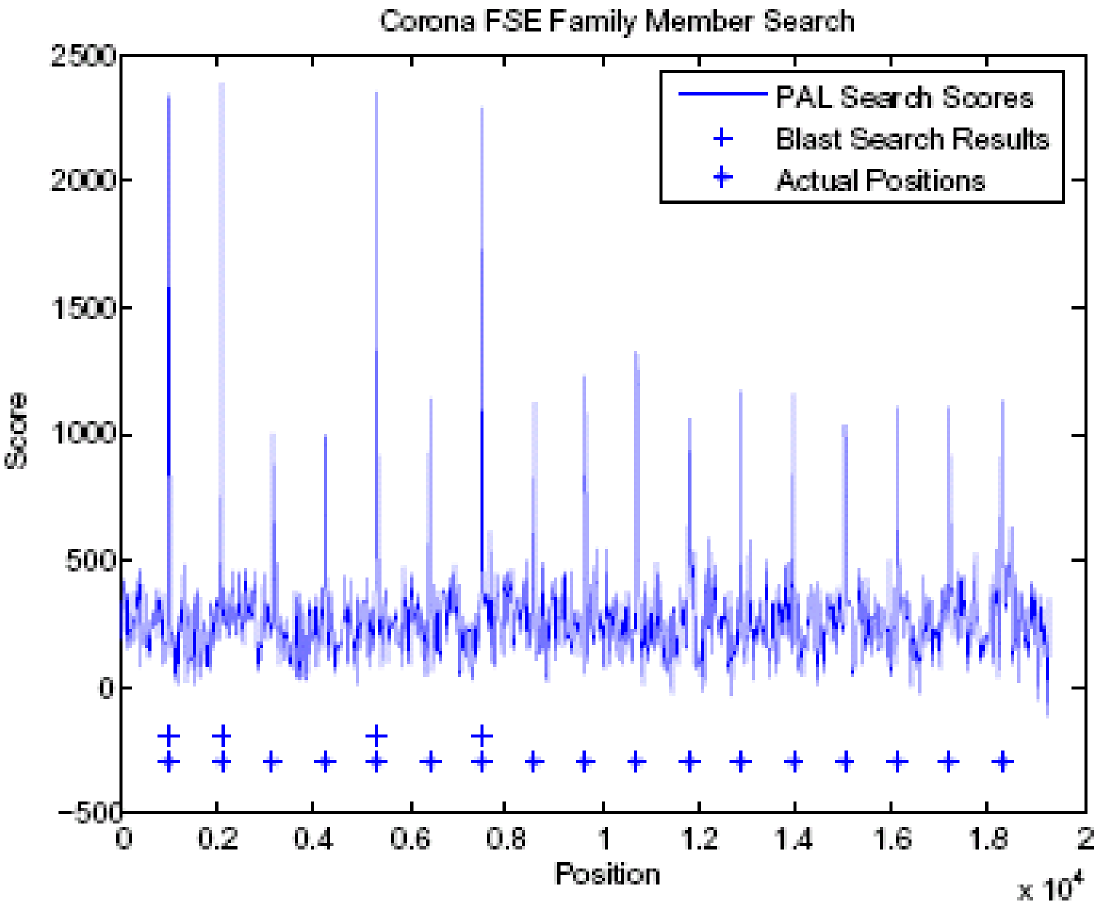
- Specificity = TP/(TP+FP)
- Sensitivity = TP/(TP+FN)
- Both measure is >= 0.95
- PAL is a strong predictor of structure

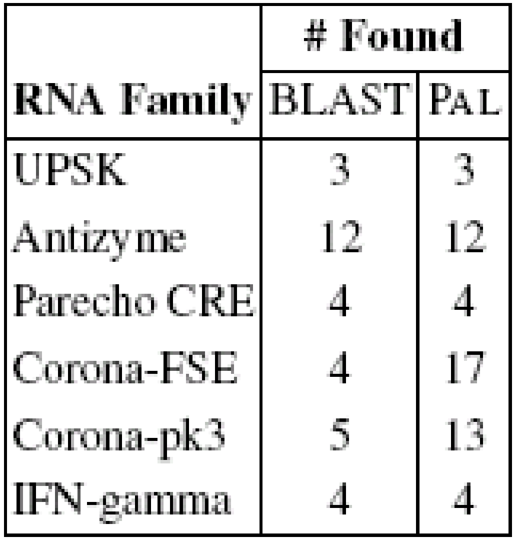
Test 2: Homologue Search

How well is PAL in finding the homologues of an RNA sequence?

- Generate a random genome.
- Insert the members of an RNA family.
- Pick one of the members as a query.
- Search for the homologues of the query.
- Can we locate the members?

Test 2: Homologue Search Results





	# Found	
RNA Family	BLAST	PAL
UPSK	3	3
Antizyme	12	12
Parecho CRE	4	4
Corona-FSE	4	17
Corona-pk3	5	13
IFN-gamma	4	4

Novel Homologues Search

- Searched whole Viral genomes for homologues of 2 pseudo-knotted RNA families:
    - Corona FSE : 11 novel members
    - Corona pk3 : 20 novel members
- Searched mouse, rat and gerbil genomes for homologues of IFN-gamma RNA family.

Conclusion

- PAL is a viable tool in finding novel homologues and inferring structure.
- We hope PAL will help to understand and explore the impact of pseudo-knotted RNAs in cellular function.