Matrix Algorithm for RNA Secondary Structure Prediction (MARSs)

S P T Krishnan[1], Mushfique J K[2], Bharadwaj Veeravali[2]

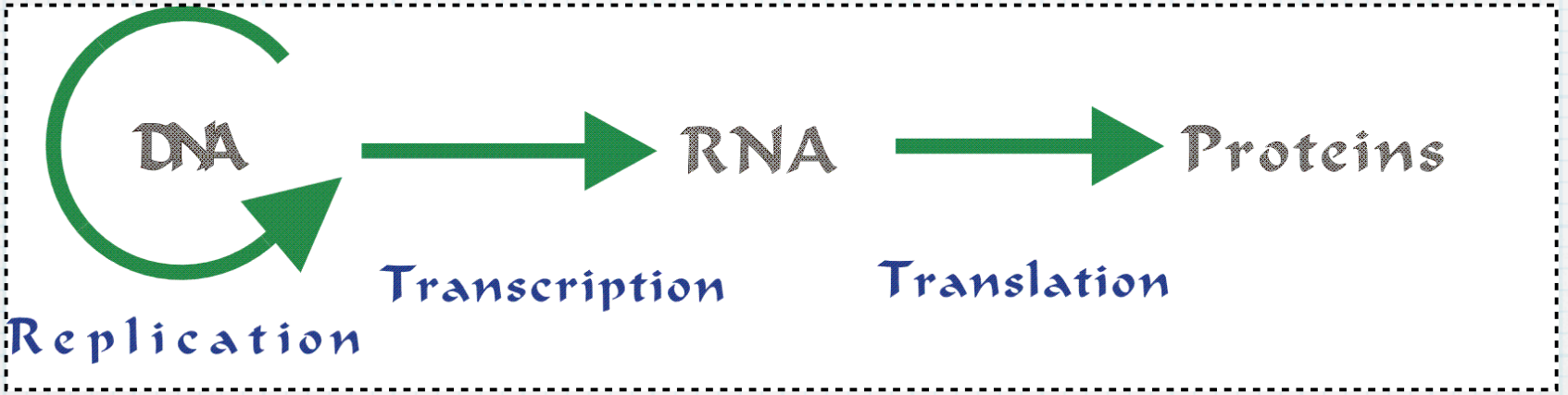
[1]Institute for Infocomm Research, Singapore

[2]National University of Singapore, Singapore

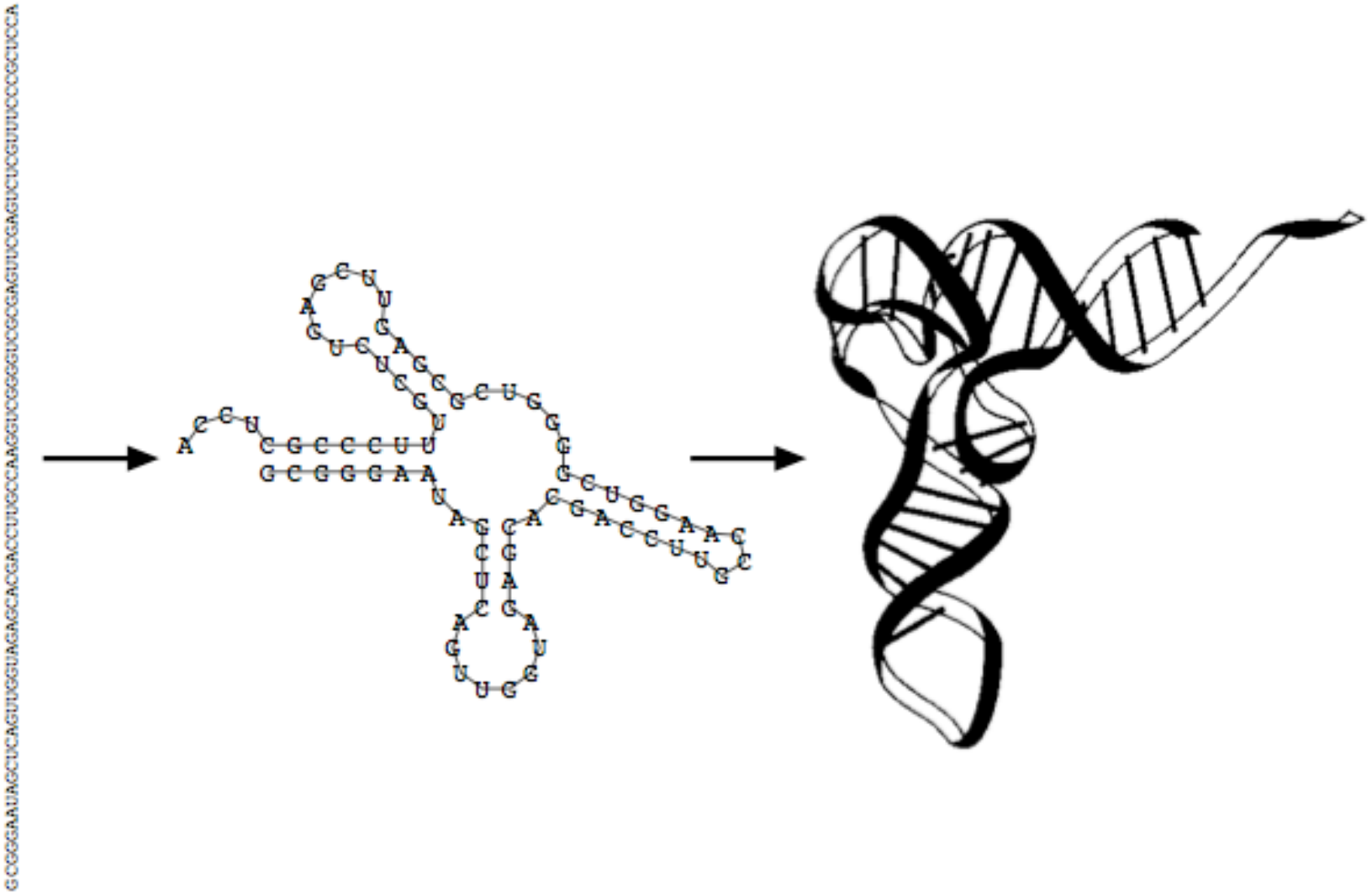
Presentation highlights

* RNA & RNA secondary structure prediction
* HPC & multi-core CPUs
* MARSs
* Results
* Discussion

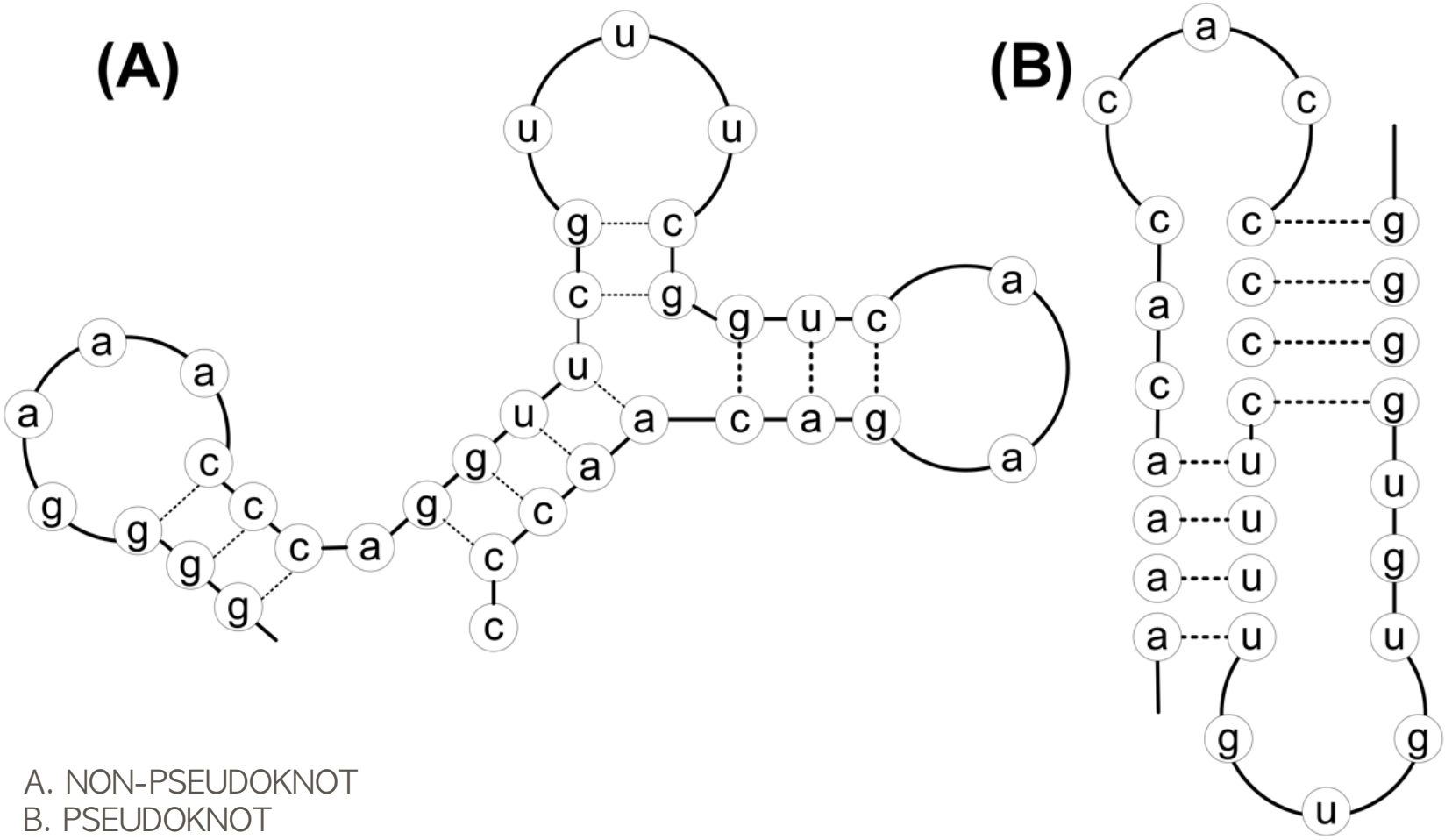
RNA &
RNA 2° structure prediction



Central Dogma of Molecular Biology



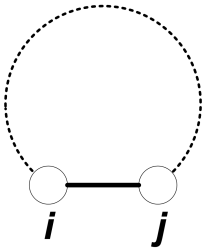
RNA evolution



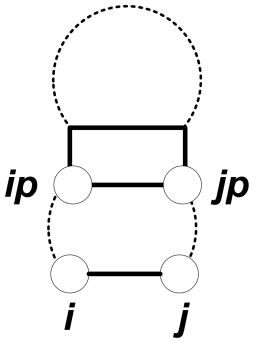
A. NON-PSEUDOKNOT
B. PSEUDOKNOT

RNA 2º structures

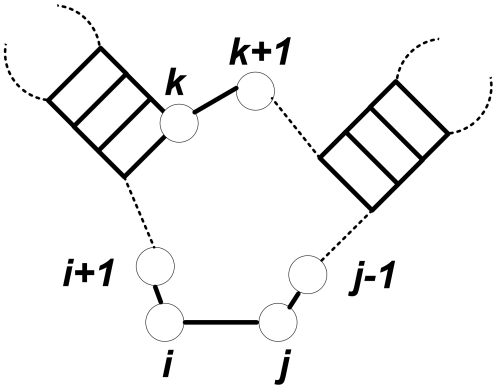
(A)
hairpin loop



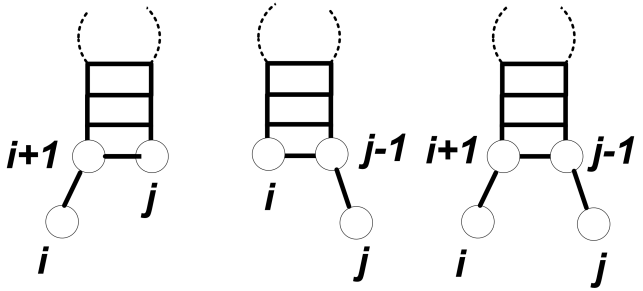
(B)
stack/bulge/internal loop



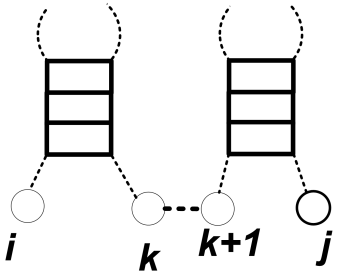
(C)
bifurcation loop



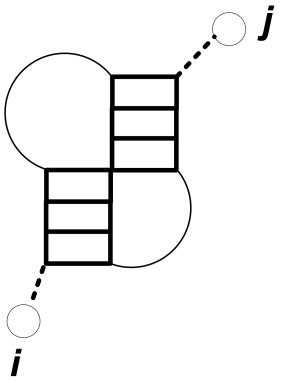
(D)
i or j or both i and j are dangling



(E)
open bifurcation



(F)
pseudoknot



RNA 2° structural motifs

HPC evolution

The advent of multi-core architectures...

- Traditional HPC architectures
  - Beowulf clusters, Grid computing
- Modern multi-core architectures are new entrants in HPC
  - Less latency among computing threads since less network traffic
  - Nvidia CUDA/Fermi - 512 computing cores in a GPU
  - Cell Broadband Engine - 200 GFlops per CPU (8 cores +2 threads)
  - Google App Engine - up to 10,000 parallel instances of an application

earlier publication

Towards High-performance Computing for Molecular Structure Prediction Using IBM Cell Broadband Engine - An Implementation Perspective

S P T Krishnan[1], Sim Sze Liang[2], Bharadwaj Veeravali[2]

[1]Institute for Infocomm Research, Singapore

[2]National University of Singapore, Singapore

APBC2010

The Eighth Asia Pacific Bioinformatics Conference

Bangalore, India, 18-21 January 2010

MARSs

Why MARSs ?

- Many DP based algorithms unable to predict Pseudoknots
- DP based algorithms are recursive by design
  - Later iterations depend on results from earlier iterations
  - Inhibitor for parallelizing the algorithm
- Processor manufacturers are adding more cores to CPUs and GPUs
  - Reversing the earlier trend of increasing raw clock speed

What is MARSs ?

- A novel high-performance algorithm for predicting RNA secondary structures with and without Pseudoknots
- Top-down methodology (global to local) unlike DP based algorithms
  - Does not view 2^o prediction as a set of overlapping problems
  - Non-recursive therefore highly portable and HPC ready
- Produces several potential 2^o structure candidates
  - Pseudoknots (simple, generic), Non-Pseudoknots
  - DP based algorithms work towards producing only one prediction

What is MARSs - 2 ?

- MARSs does not use a dictionary-based search for 2^o structural motifs
  - Yet it can predict all known motifs and potentially new types
- MARSs consistently shows a speedup of ~n where 'n' is the number of cores
- High prediction accuracies
  - PPV (Positive Predicted Value) = 76.46%
  - Sensitivity = 81.04%
  - PPV and Sensitivity are above state-of-the-art algorithms

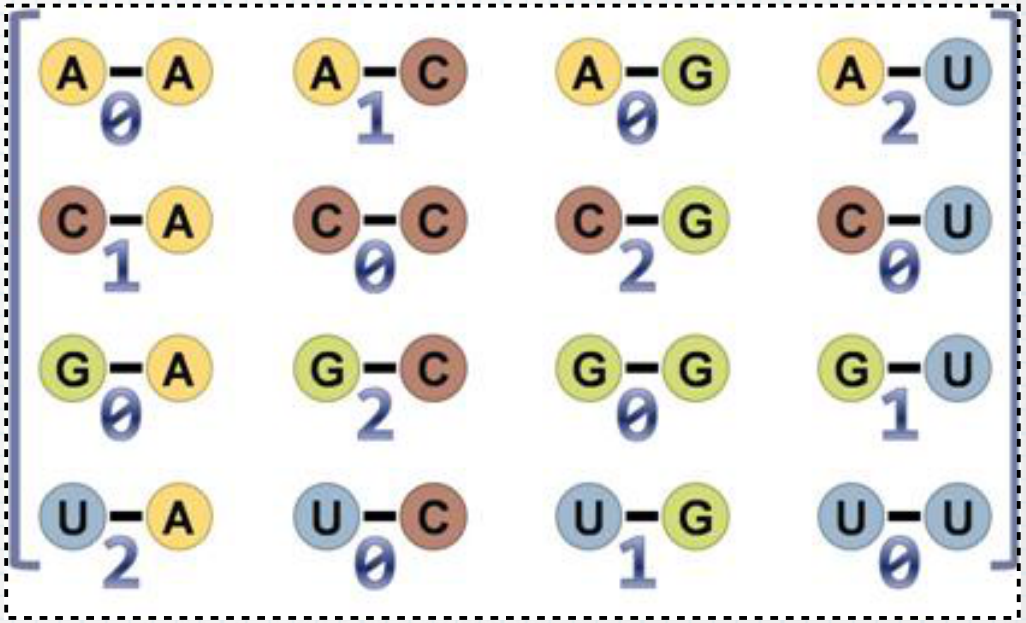
What is unique about MARSs ?

- Operate in both serial and parallel modes
  - Auto scaling with the number of computing cores
- MARS is non-recursive by design
  - Unlike Dynamic Programming based algorithms
- Modular in design and uses matrices based data structures
  - Simple design - makes it easy for porting to new architectures
  - Already implemented in IBM Cell BE and Intel x64.

MARSs - Under the Hood

- MARSs uses two core matrices
  - Base Pair (BP) Matrix
  - Affinity Matrix
- BP Matrix
  - Fixed 4 x 4 static matrix
  - Represents bonds strengths among RNA nucleotides
    - Integers / Floating Point values are OK

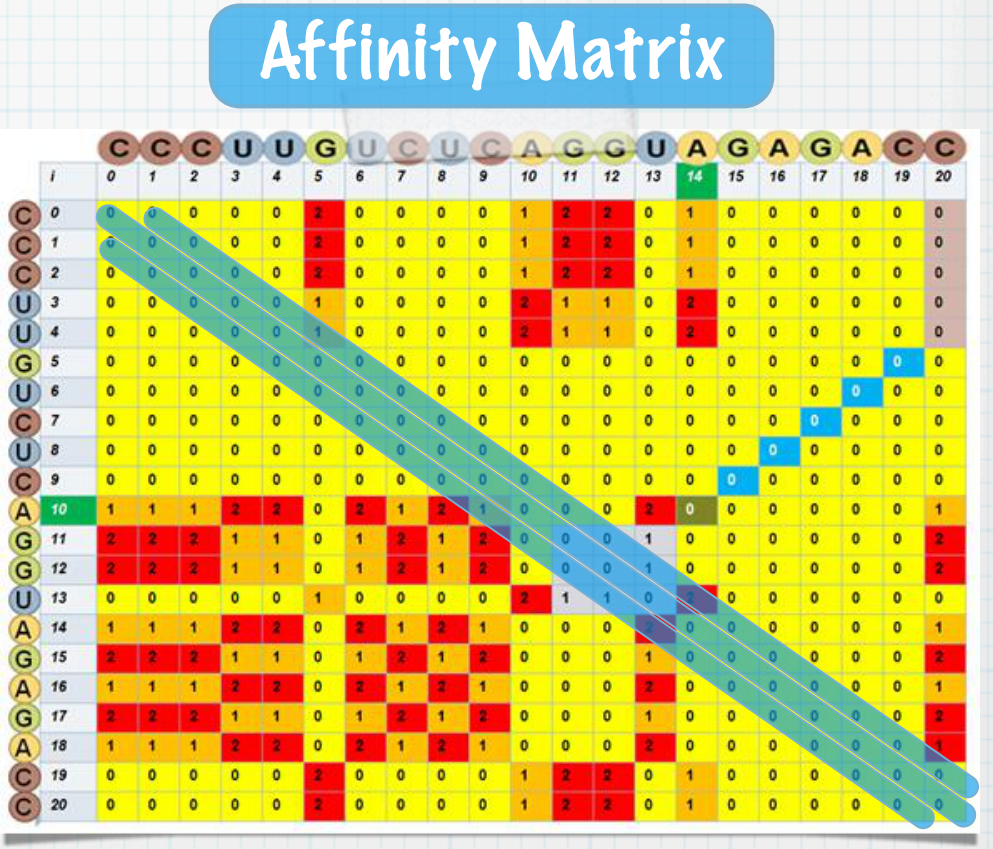
Base Pair Matrix



2 - Strong bonds: Watson Crick (G-C, A-U)
1 - Weak bonds: Hogsteen (A-C), Wobble (G-U)
0 - No base pairing possible

MARSs - Under the Hood (2)

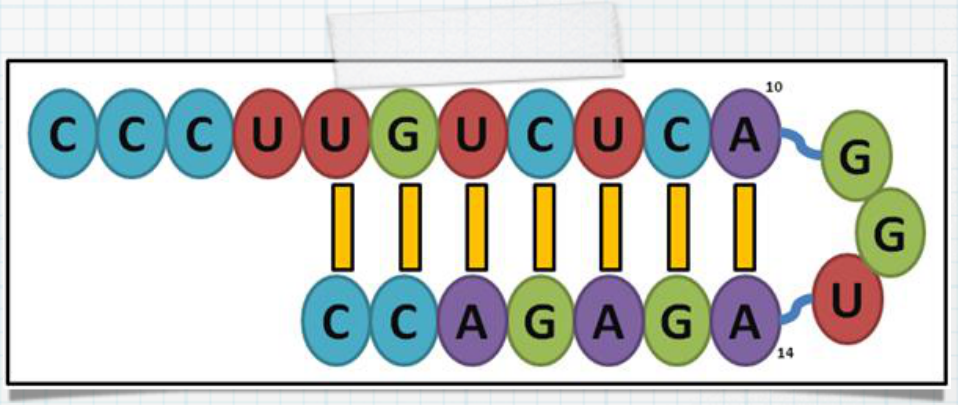
- Affinity matrix
  - Size and contents depends on input sequence
  - n x n where 'n' is the RNA primary sequence length
  - Can be optimized in implementation - store data references and not values
  - Zero natural bonds...



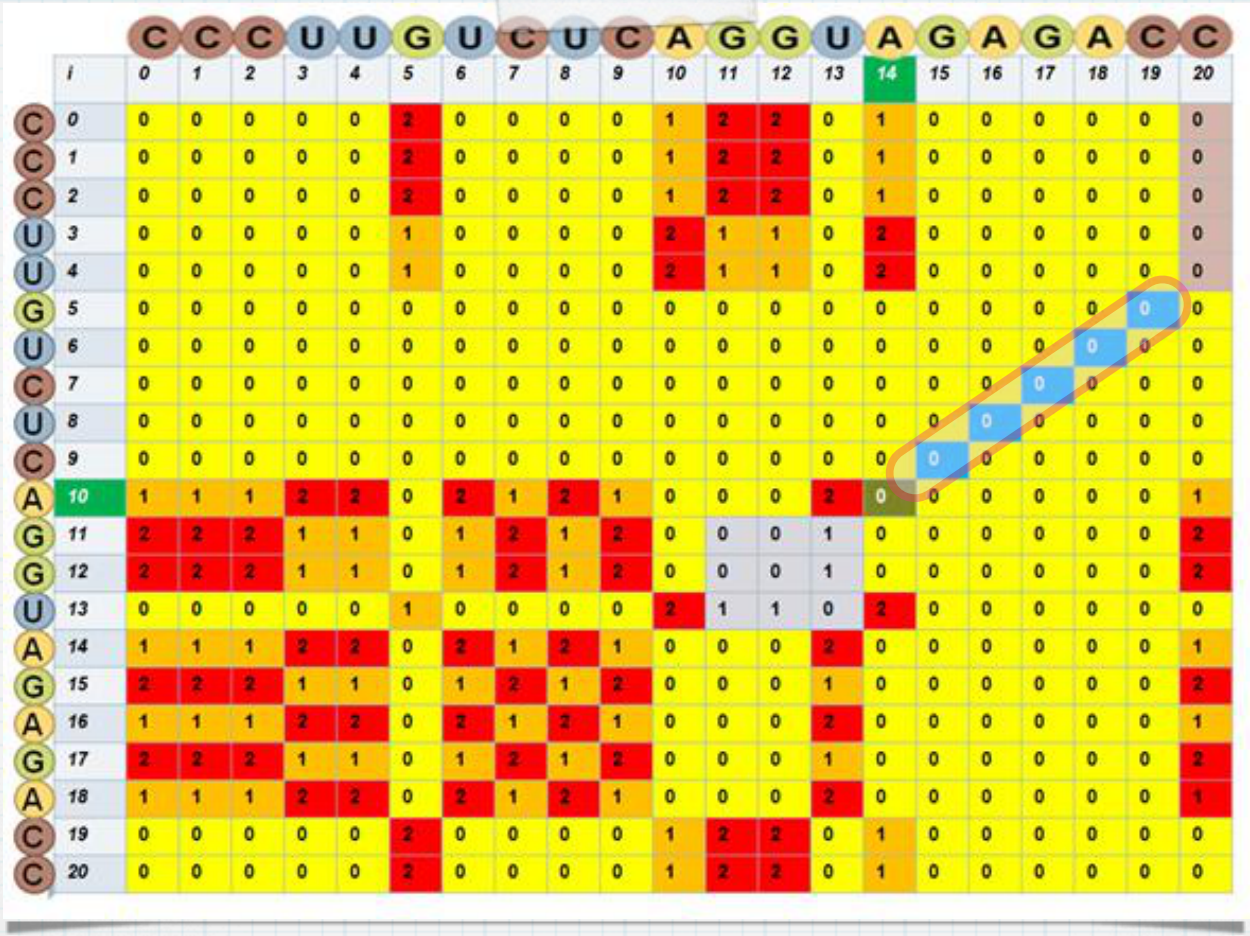
MARSs Under the Hood (3)

- Level '1' folding
  - Symmetric fold
  - Asymmetric fold
- Symmetric fold motifs
  - Hair-pin and internal loops, stems, dangling ends
- Asymmetric fold motifs
  - Bulges, Asymmetric internal loops

Symmetric fold



Symmetric fold

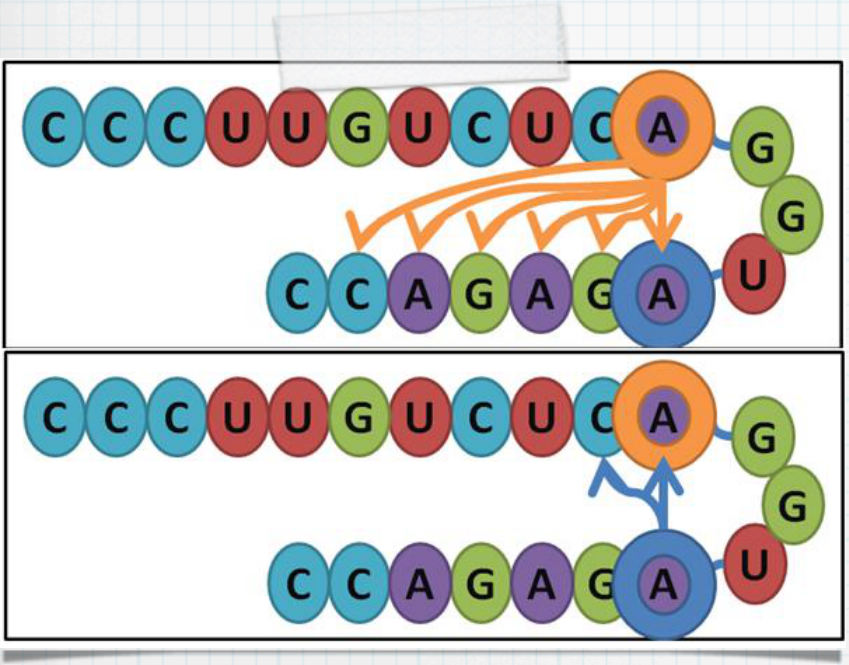


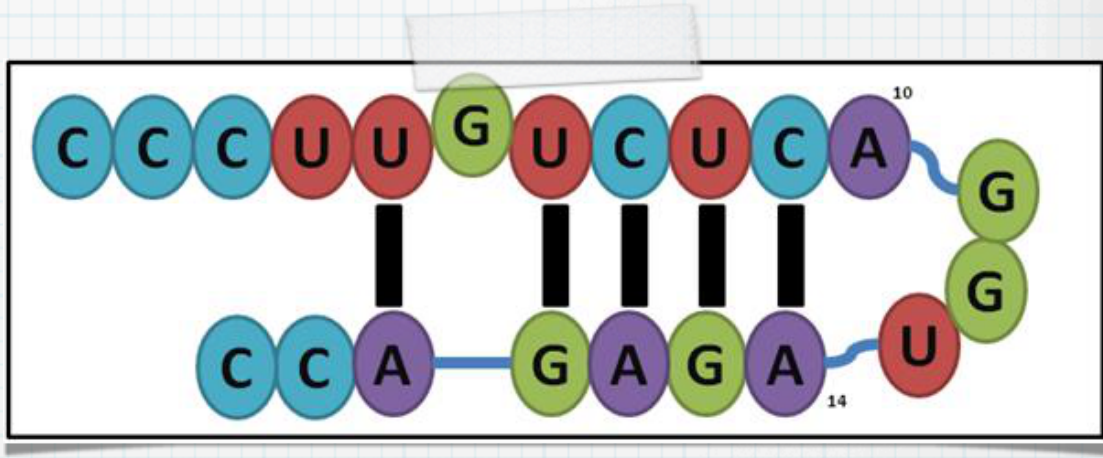
		C	C	C	U	U	G	U	C	U	C	A	G	G	U	A	G	A	G	A	C	C
	i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
C	0	0	0	0	0	0	2	0	0	0	0	1	2	2	0	1	0	0	0	0	0	0
C	1	0	0	0	0	0	2	0	0	0	0	1	2	2	0	1	0	0	0	0	0	0
C	2	0	0	0	0	0	2	0	0	0	0	1	2	2	0	1	0	0	0	0	0	0
U	3	0	0	0	0	0	1	0	0	0	0	2	1	1	0	2	0	0	0	0	0	0
U	4	0	0	0	0	0	1	0	0	0	0	2	1	1	0	2	0	0	0	0	0	0
G	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
U	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
U	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	10	1	1	1	2	2	0	2	1	2	1	0	0	0	2	0	0	0	0	0	0	1
G	11	2	2	2	1	1	0	1	2	1	2	0	0	0	1	0	0	0	0	0	0	2
G	12	2	2	2	1	1	0	1	2	1	2	0	0	0	1	0	0	0	0	0	0	2
U	13	0	0	0	0	0	1	0	0	0	0	2	1	1	0	2	0	0	0	0	0	0
A	14	1	1	1	2	2	0	2	1	2	1	0	0	0	2	0	0	0	0	0	0	1
G	15	2	2	2	1	1	0	1	2	1	2	0	0	0	1	0	0	0	0	0	0	2
A	16	1	1	1	2	2	0	2	1	2	1	0	0	0	2	0	0	0	0	0	0	1
G	17	2	2	2	1	1	0	1	2	1	2	0	0	0	1	0	0	0	0	0	0	2
A	18	1	1	1	2	2	0	2	1	2	1	0	0	0	2	0	0	0	0	0	0	1
C	19	0	0	0	0	0	2	0	0	0	0	1	2	2	0	1	0	0	0	0	0	0
C	20	0	0	0	0	0	2	0	0	0	0	1	2	2	0	1	0	0	0	0	0	0

Legend:

- 'Dark green' shaded element is the folding point
- 'Blue' elements are the base pairs that are formed
- 'Grey' area is the hair pin loop
- 'Light Brown' elements represents the dangling ends
- 'Red', 'Orange' and 'Yellow' elements represent strong, weak and no bonds

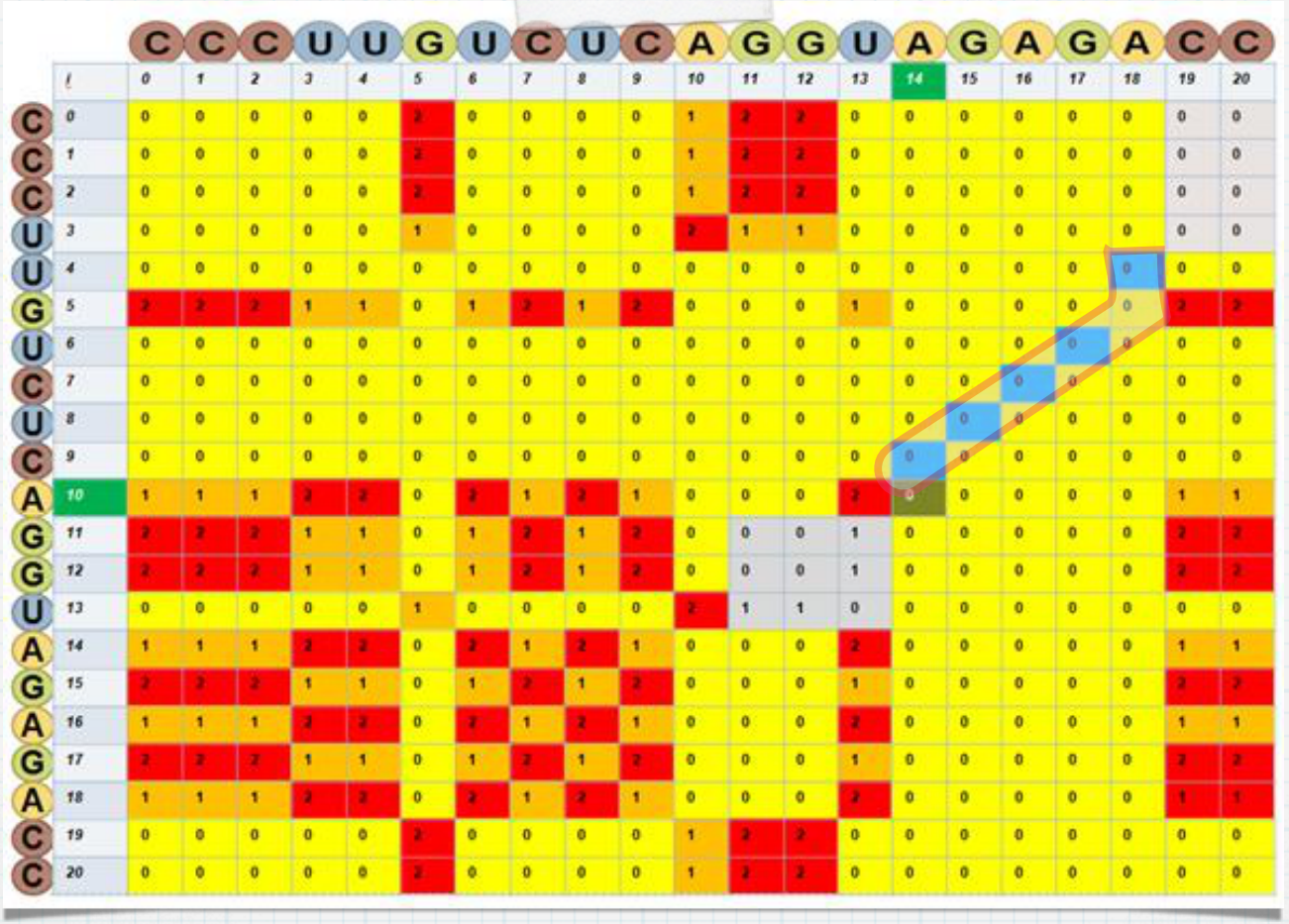
MARS Under the Hood (4) - Asymmetric fold





1. Choose the base pair that had to skip least number of nucleotides to form the base pair. We call this distance.
2. If distance is same, then we pick the base pair which has more weight, ie. a Watson-Crick base pair is chosen over a Hogsteen base pair.
3. If both the base pairs' weight are the same, we arbitrarily choose one pointer.

Asymmetric fold

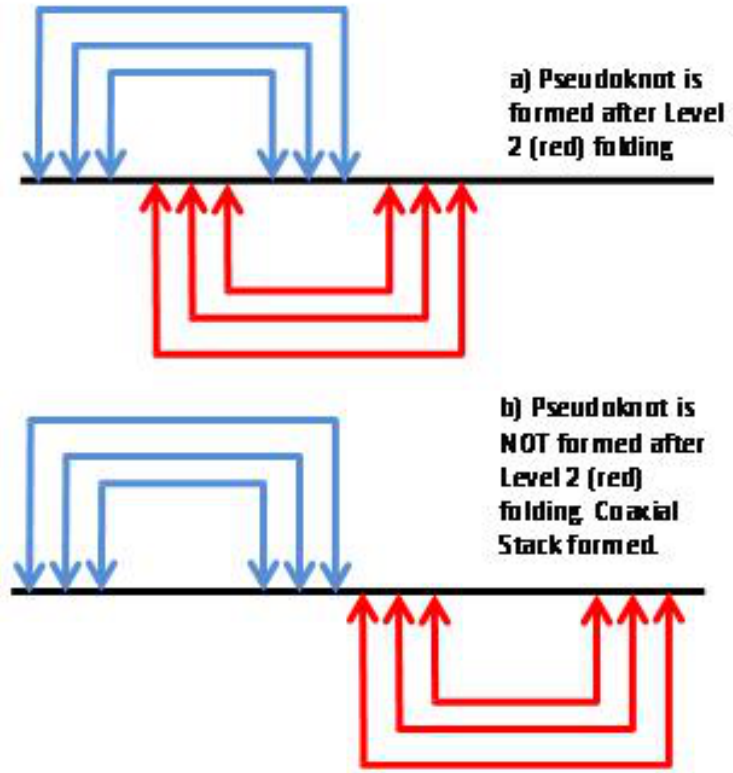


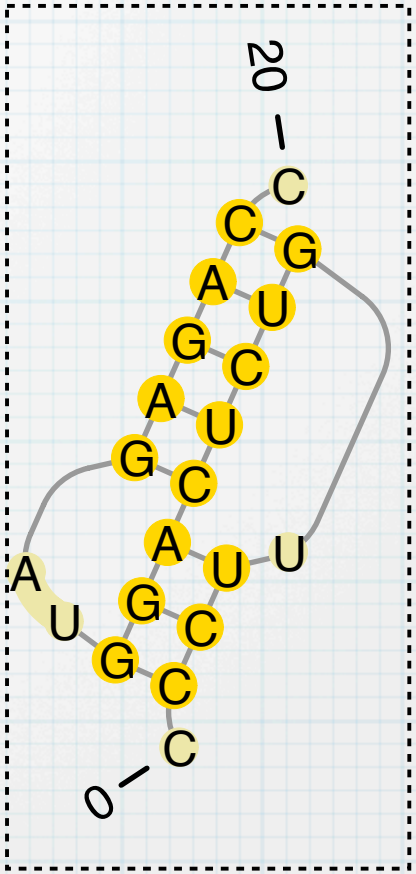
		C	C	C	U	U	G	U	C	U	C	A	G	G	U	A	G	A	G	A	C	C
	i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
C	0	0	0	0	0	0	2	0	0	0	0	1	2	2	0	0	0	0	0	0	0	0
C	1	0	0	0	0	0	2	0	0	0	0	1	2	2	0	0	0	0	0	0	0	0
C	2	0	0	0	0	0	2	0	0	0	0	1	2	2	0	0	0	0	0	0	0	0
U	3	0	0	0	0	0	1	0	0	0	0	2	1	1	0	0	0	0	0	0	0	0
U	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	5	2	2	2	1	1	0	1	2	1	2	0	0	0	1	0	0	0	0	0	2	2
U	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
U	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	10	1	1	1	2	2	0	2	1	2	1	0	0	0	2	0	0	0	0	0	1	1
G	11	2	2	2	1	1	0	1	2	1	2	0	0	0	1	0	0	0	0	0	2	2
G	12	2	2	2	1	1	0	1	2	1	2	0	0	0	1	0	0	0	0	0	2	2
U	13	0	0	0	0	0	1	0	0	0	0	2	1	1	0	0	0	0	0	0	0	0
A	14	1	1	1	2	2	0	2	1	2	1	0	0	0	2	0	0	0	0	0	1	1
G	15	2	2	2	1	1	0	1	2	1	2	0	0	0	1	0	0	0	0	0	2	2
A	16	1	1	1	2	2	0	2	1	2	1	0	0	0	2	0	0	0	0	0	1	1
G	17	2	2	2	1	1	0	1	2	1	2	0	0	0	1	0	0	0	0	0	2	2
A	18	1	1	1	2	2	0	2	1	2	1	0	0	0	2	0	0	0	0	0	1	1
C	19	0	0	0	0	0	2	0	0	0	0	1	2	2	0	0	0	0	0	0	0	0
C	20	0	0	0	0	0	2	0	0	0	0	1	2	2	0	0	0	0	0	0	0	0

Legend:

- 'Dark green' shaded element is the folding point
- 'Blue' elements are the base pairs that are formed
- 'Grey' area is the hair pin loop
- 'Light Brown' elements represents the dangling ends
- 'Red', 'Orange' and 'Yellow' elements represent strong, weak and no bonds

MARSs Under the Hood (5) - Level '2' Folding





PKB155
PPV - 100%
Sensitivity - 100%

In Level '2' folding, we use Level '1' structures and fold them using S-fold = Pseudoknots or Coaxial stacks

MARSs Complexity Analysis, Accuracy Measures

$$\text{Number of folds} = \frac{N(N-1)}{2}$$

Hence, Space complexity is $O(n^2)$

$$\text{Maximum number of potential base pairs traversed in S-Fold} = \frac{N}{2}$$

$$\text{Minimum number of potential base pairs traversed in S-fold} = 1$$

$$\text{Average potential base pairs traversed} = \frac{N+2}{4}$$

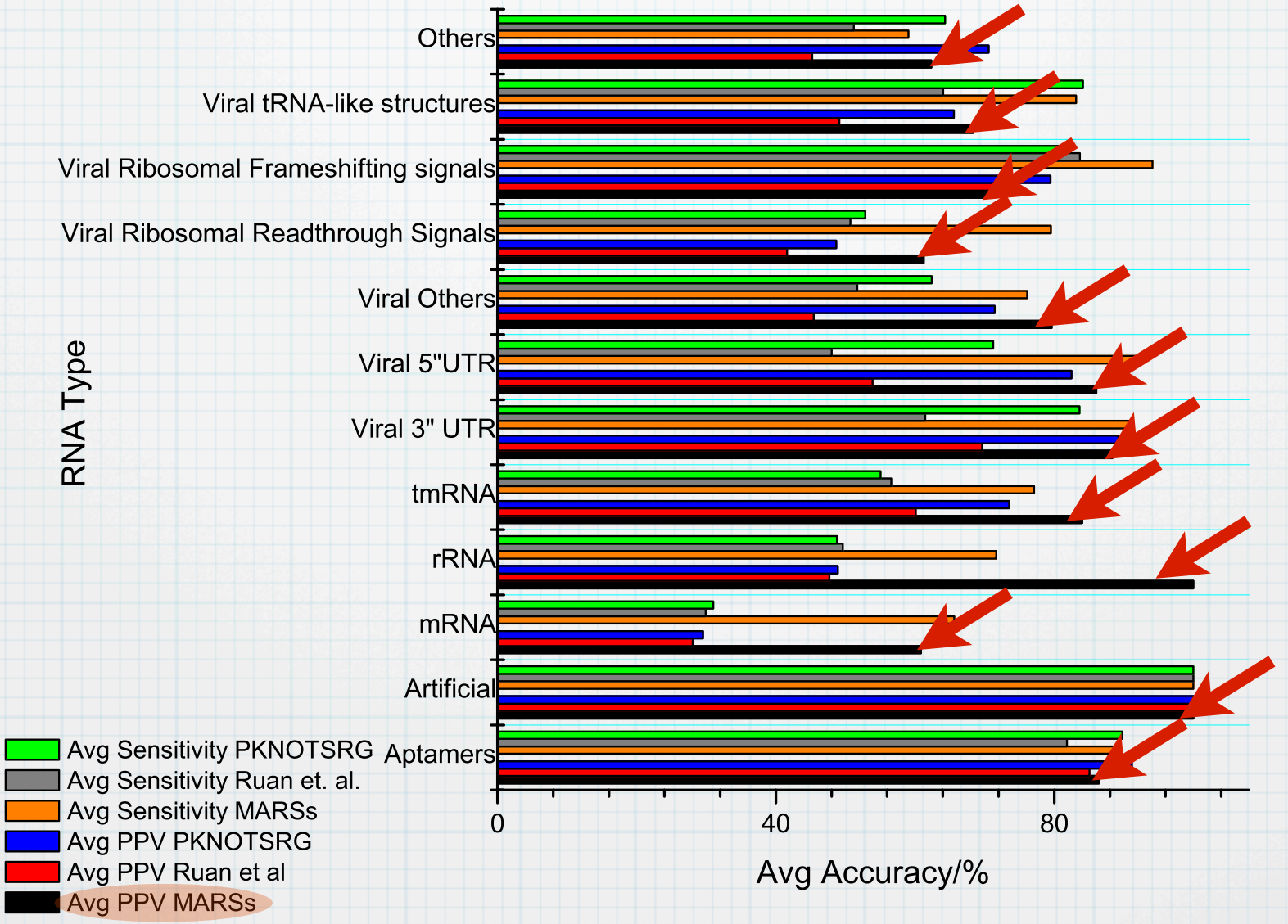
$$\text{TIME COMPLEXITY} = \frac{N(N-1)}{2} \times \frac{N+2}{4} = O(n^3)$$

Since A-fold base pairing simply increases the number of base pairs traversed by a constant, hence complexity remains same.

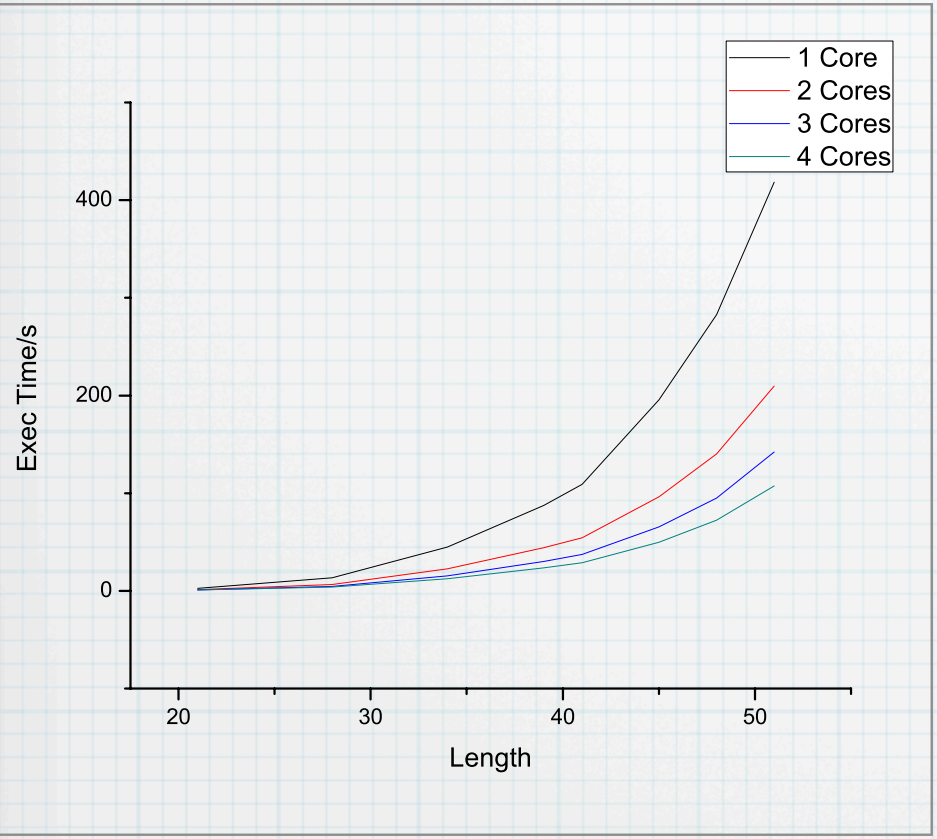
$$\text{PPV} = \frac{\text{number of correctly predicted base pairs}}{\text{total number of base pairs in PREDICTED STRUCTURE}} \times 100\%$$

$$\text{Sensitivity} = \frac{\text{number of correctly predicted base pairs}}{\text{total number of base pairs in ACTUAL STRUCTURE}} \times 100\%$$

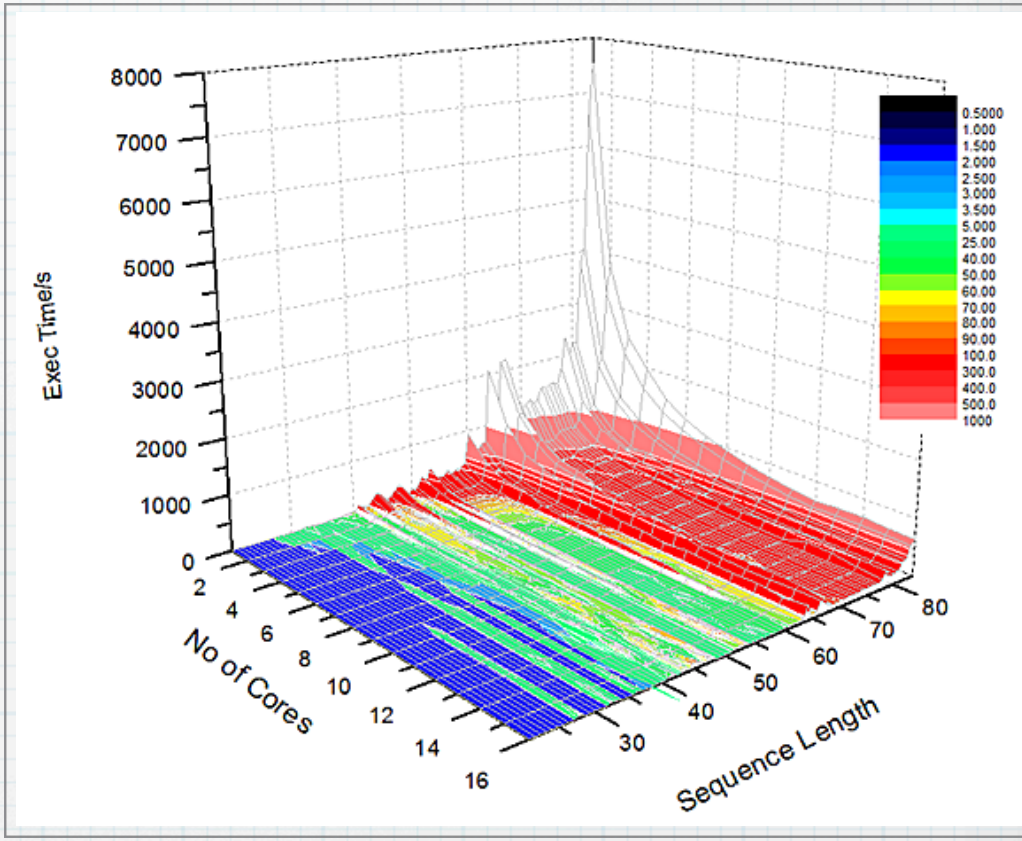
MARSs Accuracies Vs Others...



MARSs Parallelization Results



Intel Quad Core Xeon 3 GHz



16 virtual core Intel HPC

MARSs Summary

- A new class of molecular structure prediction algorithm for multi-core architectures
- Adaptable & Customizable with domain knowledge
- Future work / WiP
  - Building a GAE version for public usage
  - Optimization of Intel x64 and Cell versions
  - Web portal to a hosted MARSs implementation