

# **Permutation Patterns and RNA Secondary Structure Prediction**

**Jennifer R. Galovich**

**St. John's University/College of St. Benedict**

**Permutation Patterns 2017**

**27 June, 2017**

# Acknowledgments

Robert S. Willenbring  
(SJU '05)

Heather Akerson and Cam Christensen  
(CSB '09 and SJU '09)

# Overview

- What is RNA secondary structure? (from a biologist's point of view...)
- What is RNA secondary structure? (from a mathematician's point of view...)
- A permutation model for RNA secondary structures;  
A bijection and some statistics
- Biological insight (?)
- Future directions

# Crick's Central Dogma

**DNA**



Transcription

**RNA**



Translation

**Proteins**

## B. Subtilis RNase P RNA



## B. Subtilis RNase P RNA (Primary Structure)

GUUCUUAAACGUUCGGGUAAUCGCUGCAGAUUCUUGA  
AUCUGUAGAGGAAAGUCCAUGCUCGCACGGUGCUG  
AGAUGCCCGUAGUGUUCGUGCCUAGCGAAGUCAUA  
AGCUAGGGCAGUCUUUAGAGGCUGACGGCAGGAAA  
AAAGCCUACGUCUUCGGAUAUUGGCUGAGUAUCCUU  
GAAAGUGCCACAGUGACGAAGUCUCACUAGAAAUG  
GUGAGAGUGGAACGCGGUAAACCCCU CGAGCGAGA  
AACCCAAAUUUUGGUAGGGGGAACCUUCUUAAACGGA  
AUUCAACGGAGAGAAAGGACAGAAUGC UUUCUGUAG  
AUAGAUGAUUGCCGCCUGAGUACGAGGUGAUGAGC  
CGUUUGCAGUACGAUGGAAACAAAACAUGGCUUACA  
GAACG UUAGACCACU

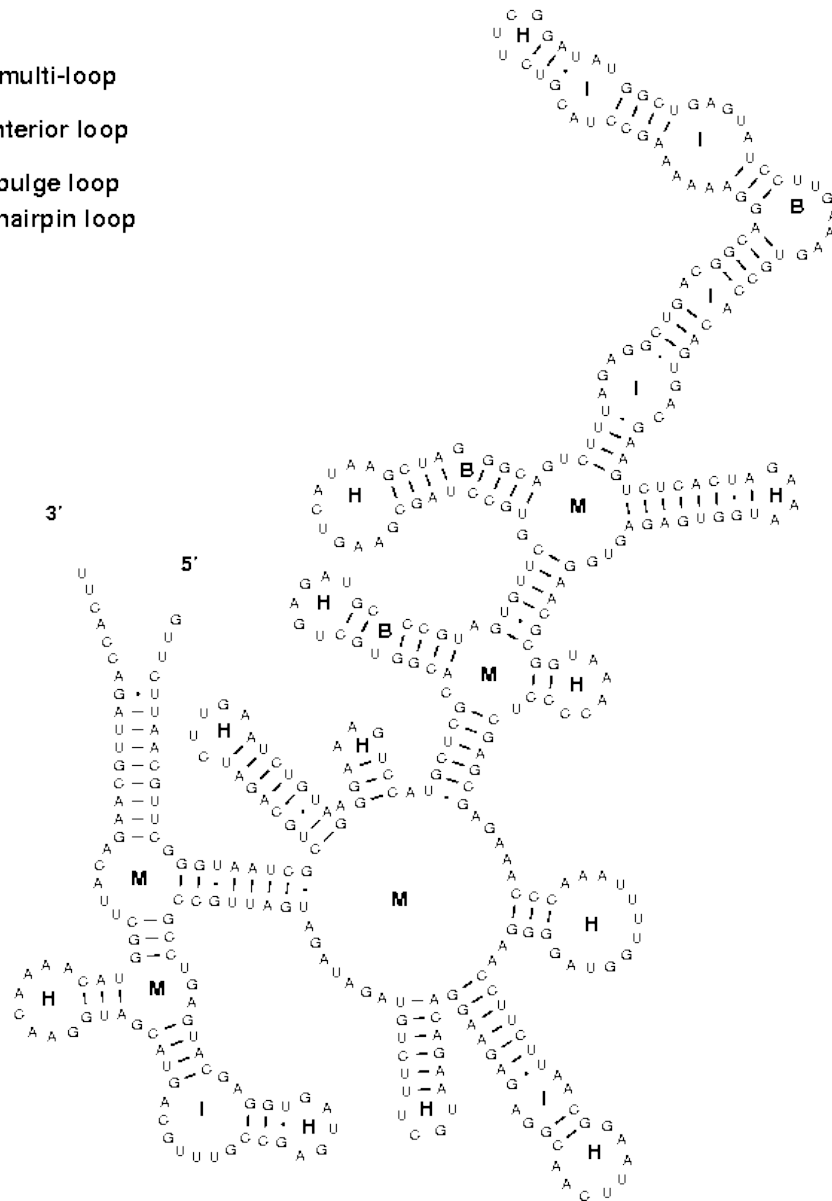
*Bacillus subtilis* RNase P RNA

**M** - multi-loop

**I** - interior loop

**B** - bulge loop

**H** - hairpin loop



# Combinatorial approaches

Idea: Ignore biochemical properties and focus on the possible topologies.

Several models have been proposed:

- Non-crossing set partitions

- (Unlabelled) linear trees (Schmitt and Waterman 1994)

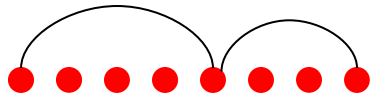
- Trees and their duals (Schlick et al. 2002)

**Permutations**

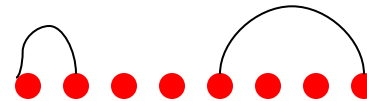


**Definition:** A *secondary structure* on  $\{1, 2, \dots, n\}$  is a non-crossing set partition such that

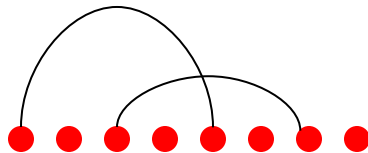
- (i) the degree of every vertex is at most 1
- (ii) if  $(i, j)$  is an edge, then  $|i - j| > 1$



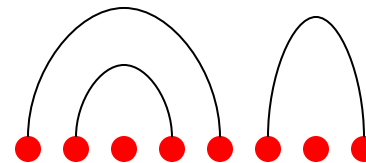
**NO...**



**NO...**



**NO...**

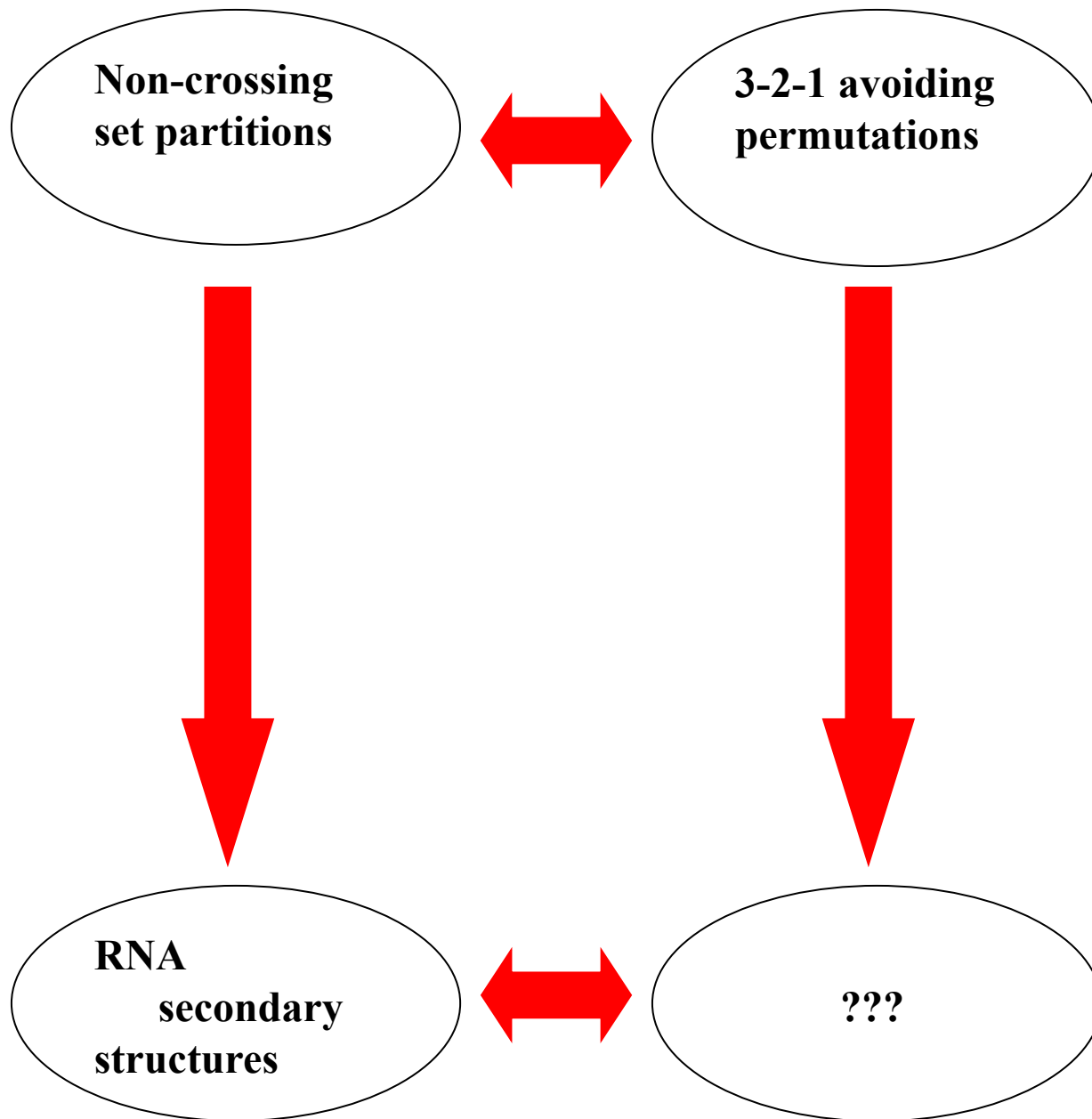


**YES**

# Enumeration

$$|SS_{n,k}| = \frac{1}{k} \binom{n-k}{k+1} \binom{n-k-1}{k-1}$$

**Schmitt and Waterman (1994)**



Let  $\Pi_n$  be the set of all 3-2-1 avoiding permutations such that:

- (i) If position  $i$  has a fall, position  $i+1$  does not.
- (ii) If  $c$  is a fall, then  $c+1$  is not.
- (iii) Every fall is the second element of at least two inversion pairs.

Let  $\Pi_{n,k}$  be the set of all permutations in  $\Pi_n$  which have exactly  $k$  falls.

**Example:  $n = 13; k = 4$**

**$k$ :** 1 2 3 4 5 6 7 8 9 10 11 12 13

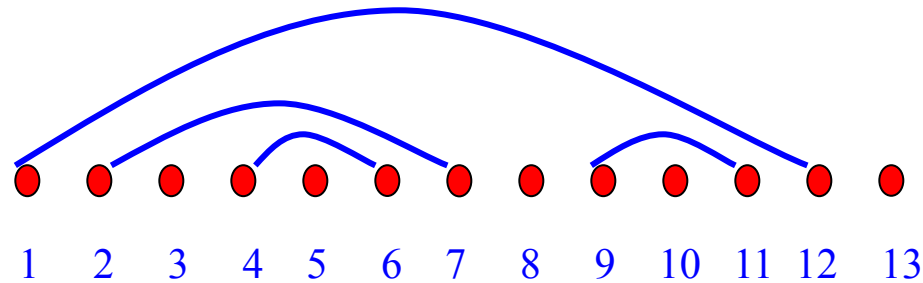
**$\pi_k$ :** 2 4 5 1 7 3 8 9 11 6 12 10 13

# Main Theorem

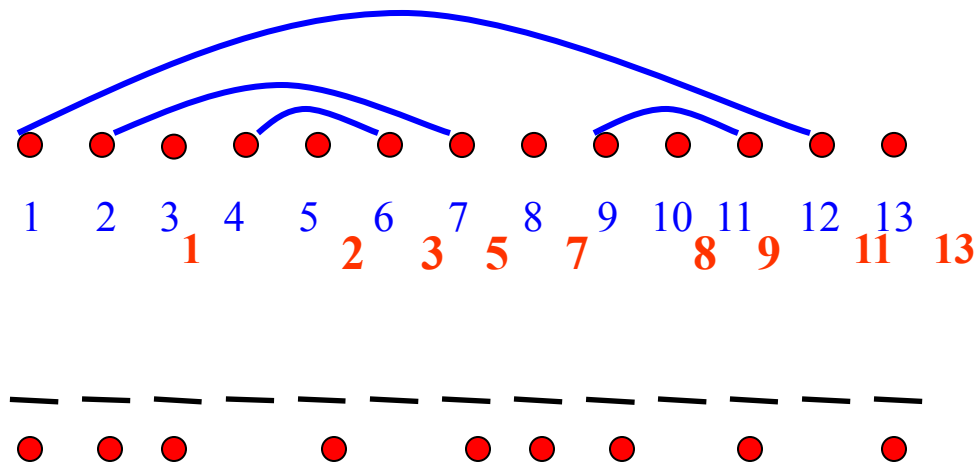
Let  $SS_{n,k}$  be the set of all RNA  
secondary structures with  $k$  bonds.

Then there is a bijection from  $SS_{n,k}$  to  
 $\Pi_{n,k}$ .

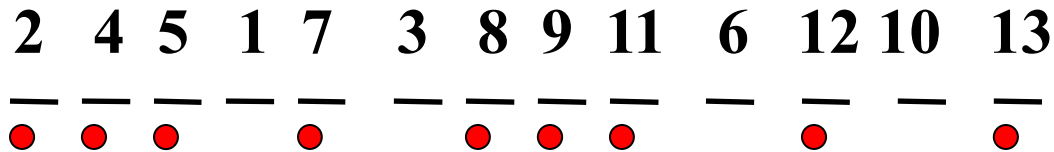
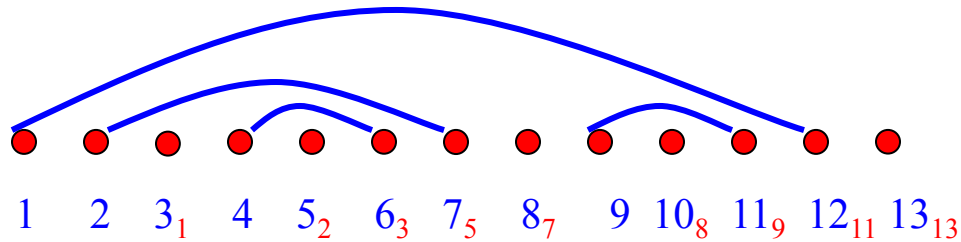
# How does the bijection work?



Labelling: Ignoring left bonds, number the unpaired bases and right bonds in order, skipping one after each right bond. Mark the numbered positions with ●

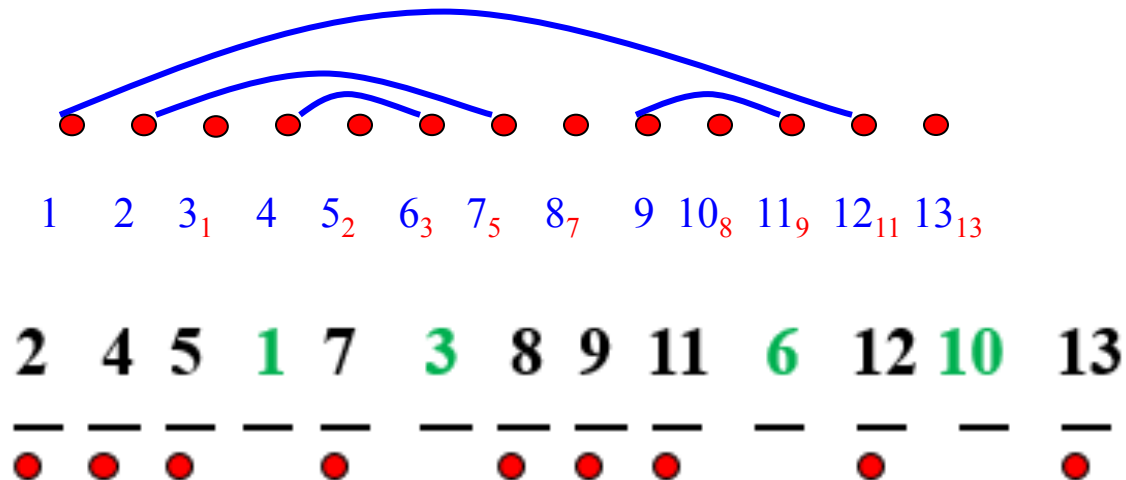


**Insertion:** Ignoring right bonds, and working left to right, insert pairs of the form  $(i+1, i)$  in (marked, unmarked) positions for each left bond and singleton values in marked positions for unpaired bases.



# Why does this work?

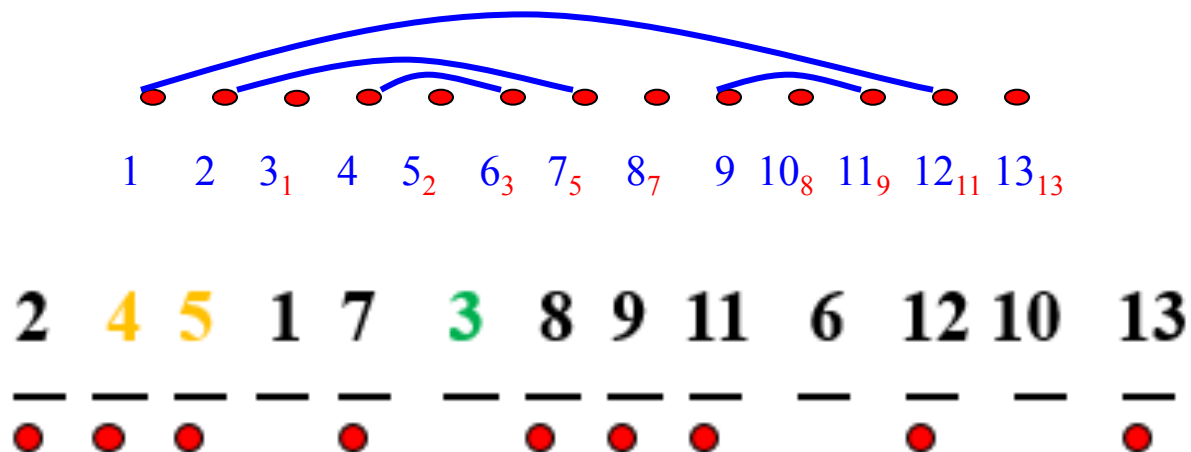
- The unmarked positions are exactly the positions of the falls, and each corresponds to a bond.
- The marked positions form an increasing sequence, as do the unmarked positions. This guarantees that the permutation is 3-2-1 avoiding.





# Why does this work?

- Consecutive unmarked positions cannot occur so if there is a fall in position  $j$  then there is NOT a fall in position  $j+1$  (condition i)
- Unmarked positions are always filled in pairs with  $i$  in the unmarked position and  $i+1$  in the marked position. Therefore  $i+1$  can never be a fall when  $i$  is (condition ii).
- Each fall will correspond to at least two inversions, for if  $i$  is a fall, then  $i+1$  precedes it, as does the value corresponding to the unpaired bas(es) enclosed by the bond (condition iii).



# Permutation Statistics

$\text{exc}(\pi)$  = number of excedances in  $\pi$

$\text{inv}(\pi)$  = number of inversions in  $\pi$

$\text{maj}(\pi)$  = sum of the descent positions in  $\pi$

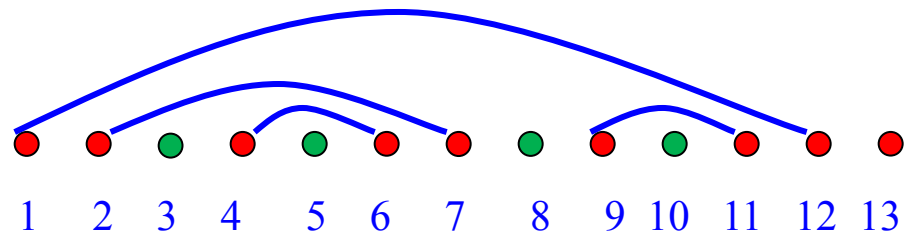
# Two RNA SS Statistics

Tau: Let  $v_i$  be the number of unpaired bases internal to bond  $i$ . Then we define

$$\tau(s) = \sum_i v_i$$

Bond Index :  $B(s)$  = sum of the positions corresponding to left or right bonds.

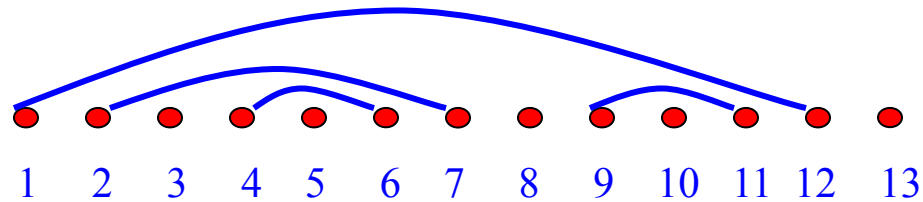
$$\tau(s) = (4 + 2 + 1 + 1) = 8$$



$$B(s) = (1+2+4+6+7+9+11+12) = 52$$

**Our Example:**  $n = 13, k = 4$

**$s$ :**



**$\pi$ :** *2 4 5 1 7 3 8 9 11 6 12 10 13*

$$\text{exc}(\pi) = 7$$

$$\text{inv}(\pi) = 12$$

$$\text{maj}(\pi) = 28$$

$$\tau(s) = 8$$

$$B(s) = 52$$

**Theorem:**

$$(1) \text{ inv} = \tau + k$$

$$(2) B = 2 (\text{maj} + k) - \text{inv}$$

# Distribution Properties for $B$ and $\tau$

**Fact:**  $B$  is symmetric on  $SS_{n,k}$

**Conjectures:**  $B$  is unimodal for any value of  $k$ .  
 $\tau$  is unimodal, but not symmetric.

# Compare to actual RNA data...

Assume that  $B$  is unimodal, and see what the values are for RNA from various (prokaryotic) organisms

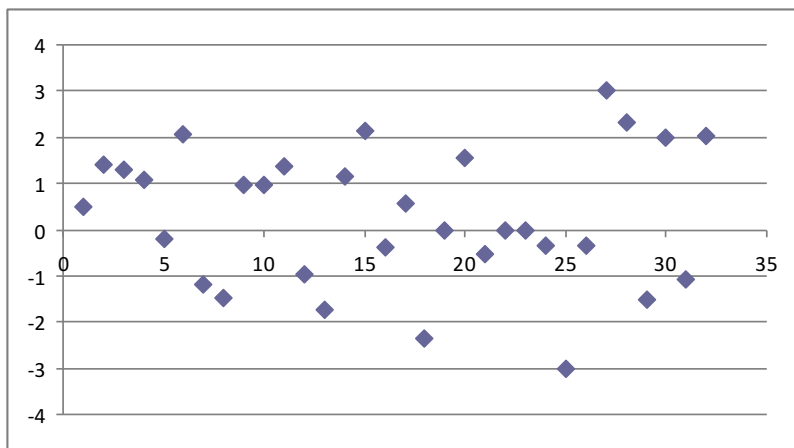
- Minimum value of  $B$  is  $2k^2 + 2k$
- Maximum value of  $B$  is  $2kn - 2k^2$

So if we assume  $B$  is unimodal, then the mode is

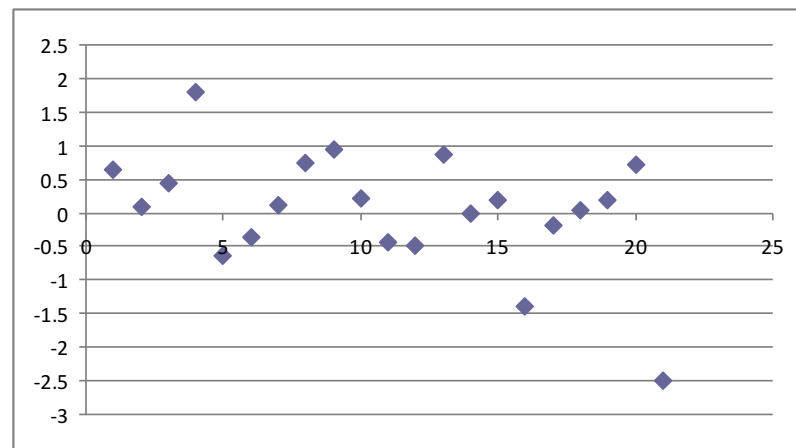
$$(n+1) k$$

# Standard Deviation

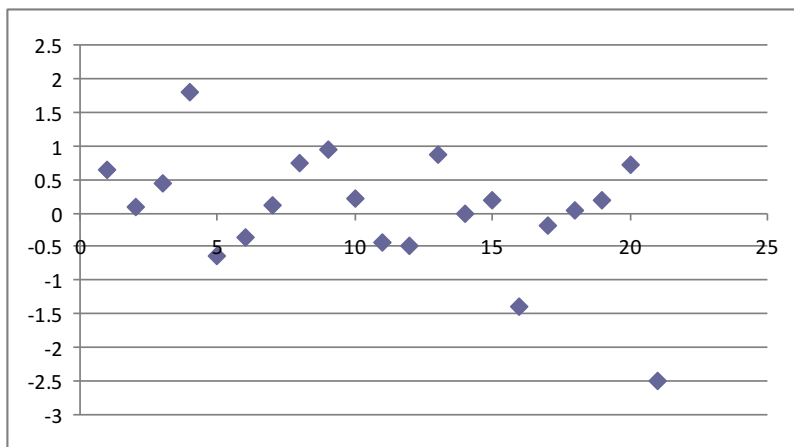
- Calculate for  $n \leq 30$ .
- Extrapolate the pattern.



**Si RNA**

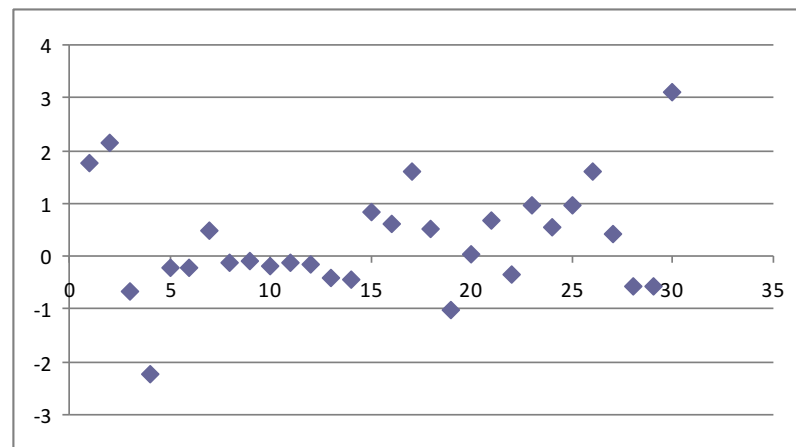


**5s rRNA**



**RNase P**

**Group I**



**Other nc RNA**

**Group II**



# Conclusions

- Group II RNA appear to have a B statistic that runs above average
- Group I RNA appear to have a B statistic that is symmetrically distributed
- The sample sizes are way too small to draw any real conclusions.

# Some Questions...

- Is there some biologically appropriate way of distinguishing Group I and Group II?
- Does either of these statistics ( $B$  or  $\tau$ ) have biological meaning? Is there a biological aspect of RNA that could be better captured by a different statistic?
- Could either the  $B$  or  $\tau$  statistic be used to evaluate folding algorithms or find potential novel RNA structures?

# To Do List....

- Prove the unimodality conjectures for  $B$  (known through  $n = 30$ ) and  $\tau$
- Use permutation statistics to describe/identify RNA motifs
- Look at  $B$  stat on experimentally verified structures
- Make the model more biological realistic by increasing the minimum number of unbonded bases (hairpins and bulges)

# References

- Schmitt and Waterman

*Discrete Appl. Math.* **51**(1994) 317-323

R. Willenbring

*Discrete Appl. Math.* **157**(2009) 1607-1614

Thank you