# Improving The Runtime Of Nussinov's Algorithm By Partitioning

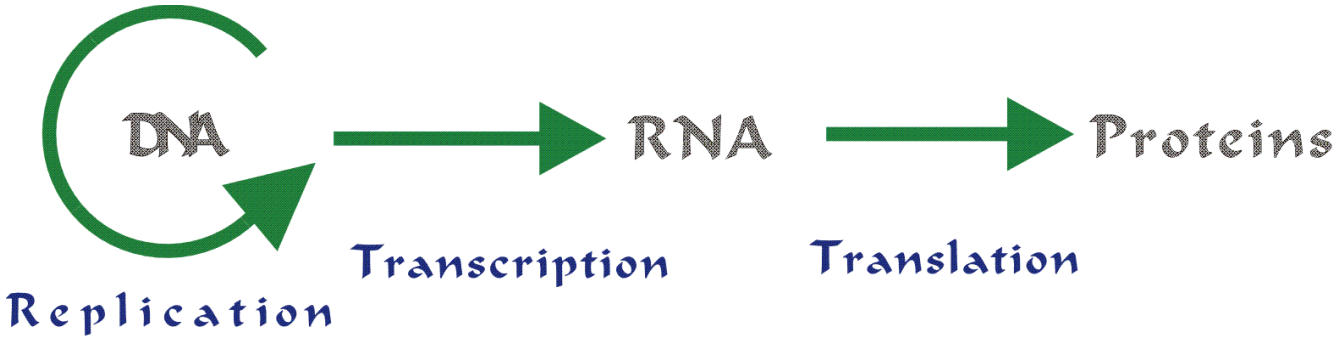
Nicole Power and Lily Seropian

6.047 Computational Biology Final Project

December 11, 2013

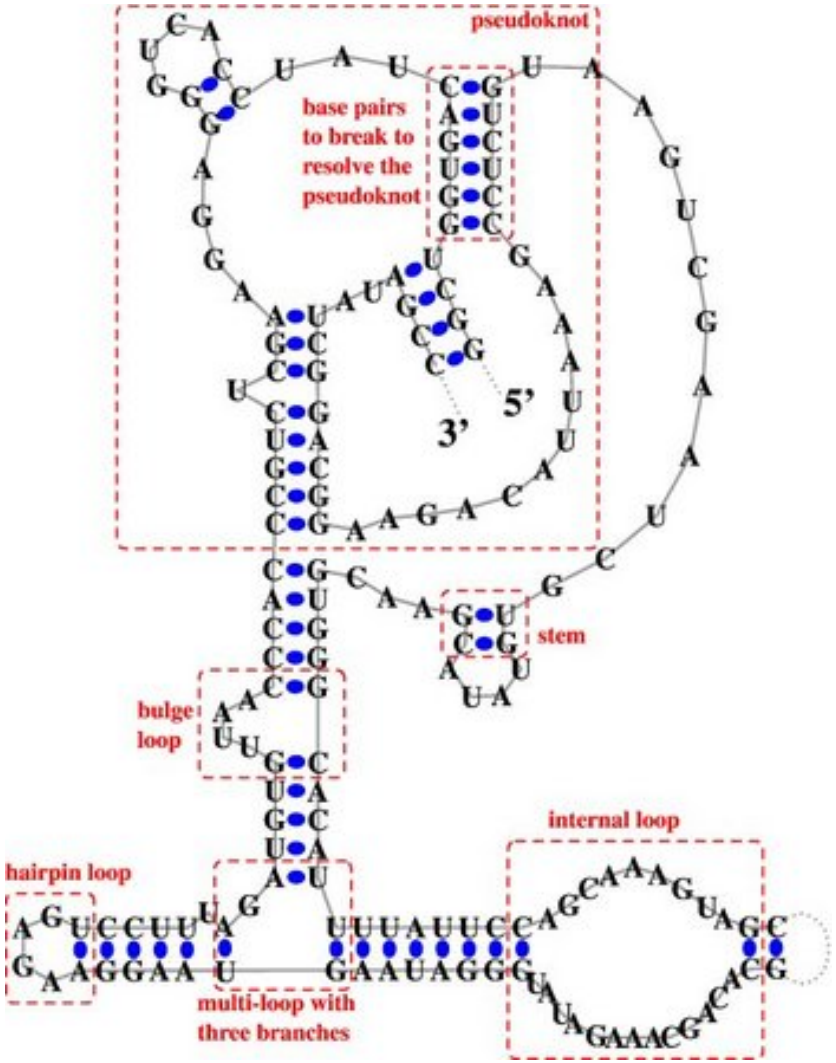
# RNA is a biologically important molecule

- The intermediate between DNA and proteins
  - vital to gene expression
- mRNA : codes proteins
- tRNA, rRNA : make proteins from mRNA
- piRNA, microRNA : regulation of proteins



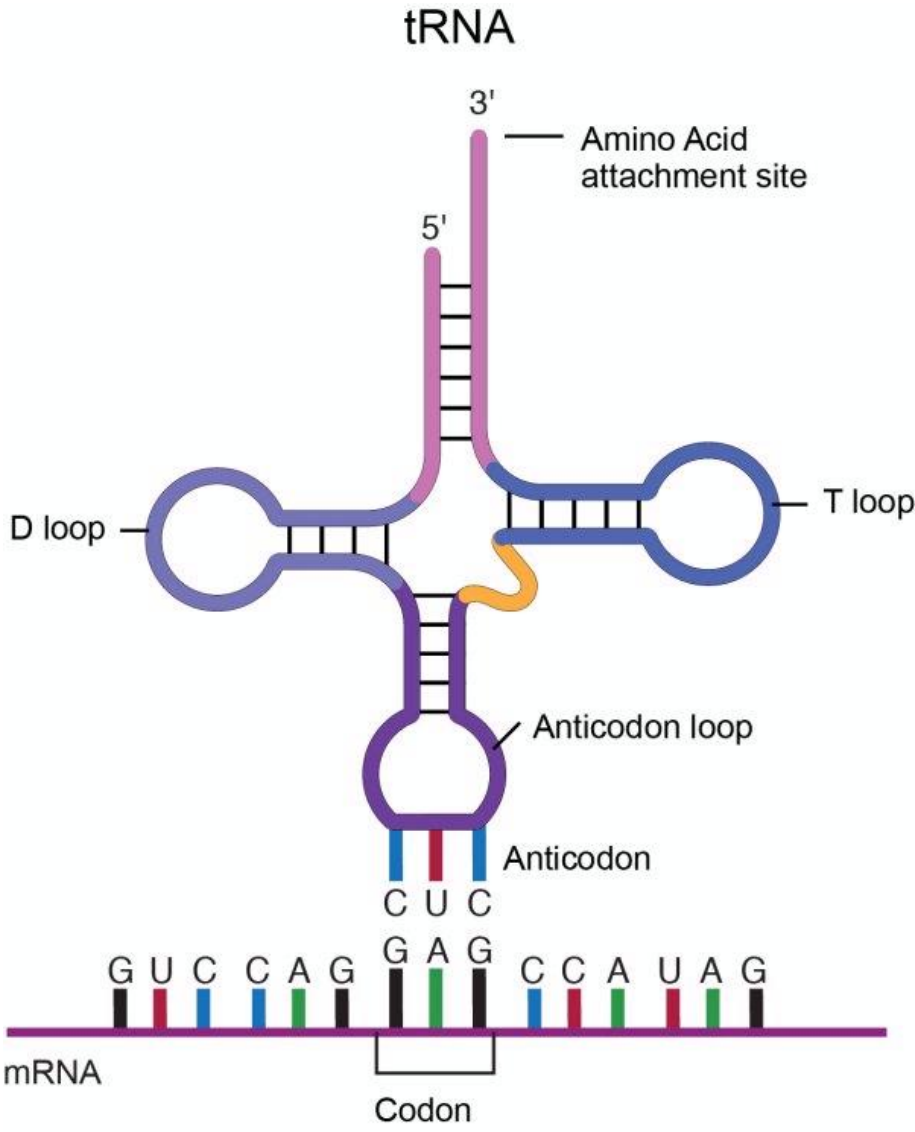
# Unlike DNA, RNA is capable of forming single-stranded double helices

- Self-bonding allows for the formation of many different secondary structures:
  - Hairpin loop
  - Bulge loop
  - Internal loop
  - Stem
  - Multi-loop
  - pseudoknot



# RNA function would be better understood if we knew its form

- Tertiary structure is complex to predict
- From properties of the secondary structure, functional attributes of the specific RNA molecule can be gleaned



# First algorithms were based on thermodynamics

- Nussinov's Algorithm
  - uses Dynamic Programming
  - Assigns score of 1 for Watson-Crick(A-U, C-G) and Wobble (G-U) base pairing. Otherwise score of 0.
  - calculates the substructure of subsequences of the RNA that minimize free energy
- The Zuker Algorithm
  - also structured to minimize free energy
  - factors in the contribution of stacking energy
- Both run in $O(n^3)$ time

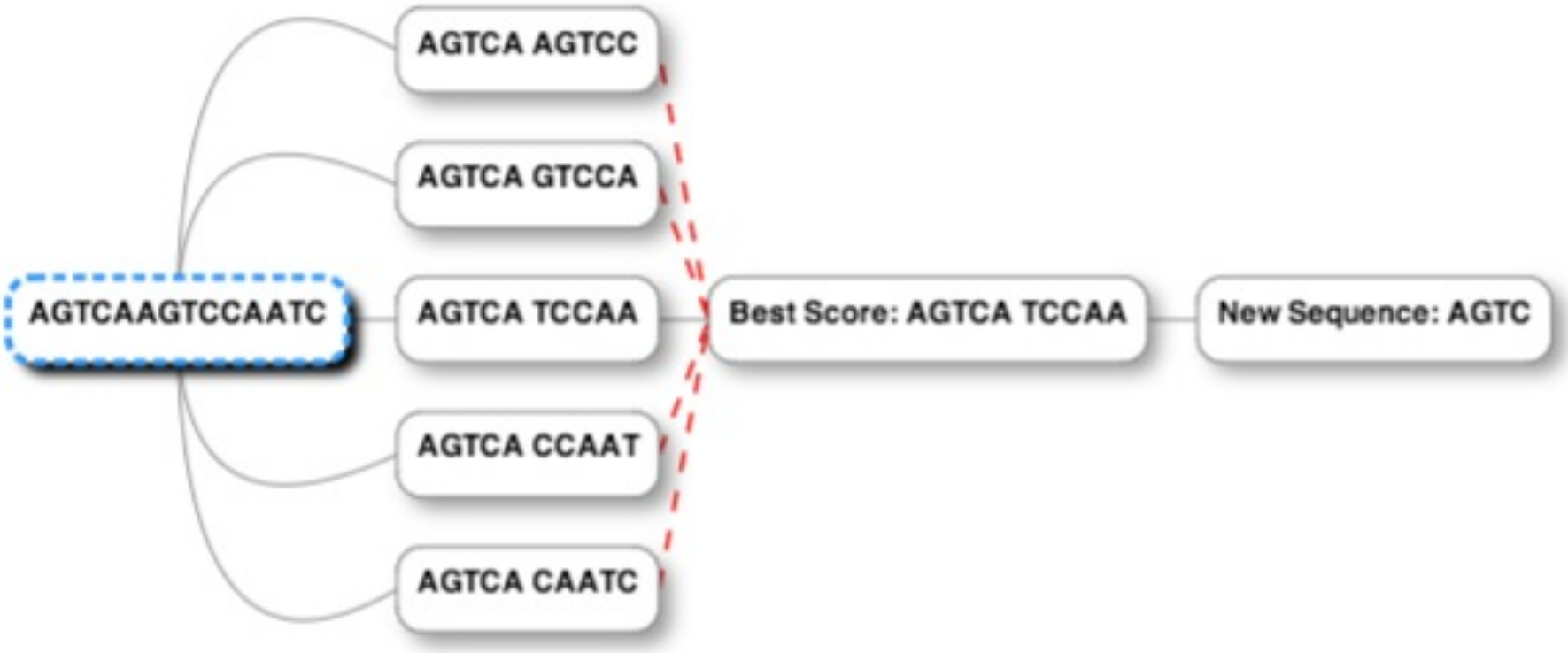
# Later algorithms were based on probabilities

- McCaskill Algorithm
  - uses a partition function and pair probabilities
- CYK and Inside-Outside Algorithms
  - use SCFGs
- Still run in $O(n^3)$ time

# $O(n^2)$ Algorithm Overview

- Fix a partition length
- While there are still unfolded bases:
  - Split the sequence into different partitions of specified length
  - Use Nussinov's algorithm to fold and score each of the partitions
  - Pick the partition with the best score and add the bases involved to the global folding
- Return the resulting global folding and its score

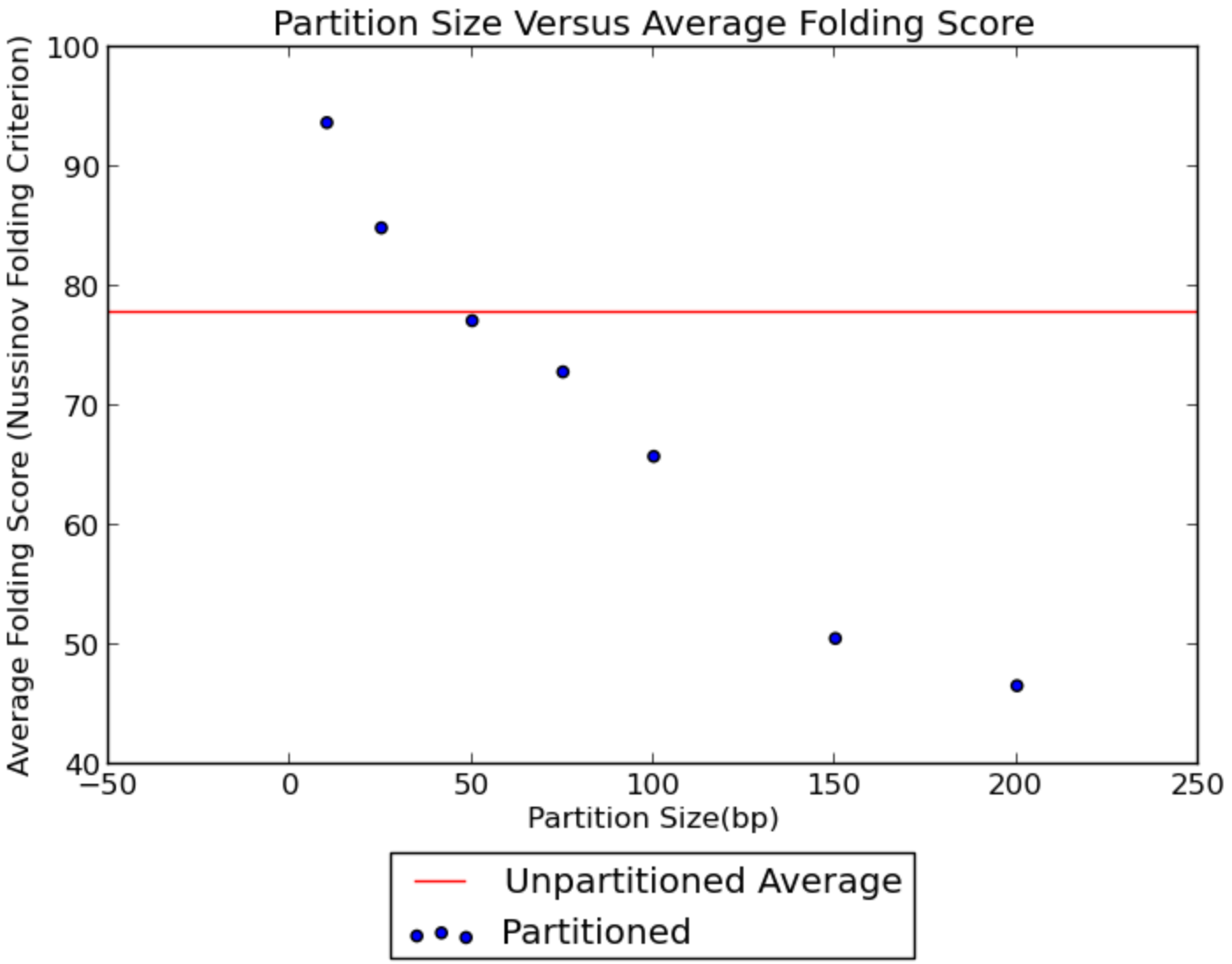
# Example


AGTCAAGTCCAATC
AGTCA AGTCC
AGTCA GTCCA
AGTCA TCCAA
AGTCA CCAAT
AGTCA CAATC
Best Score: AGTCA TCCAA
New Sequence: AGTC

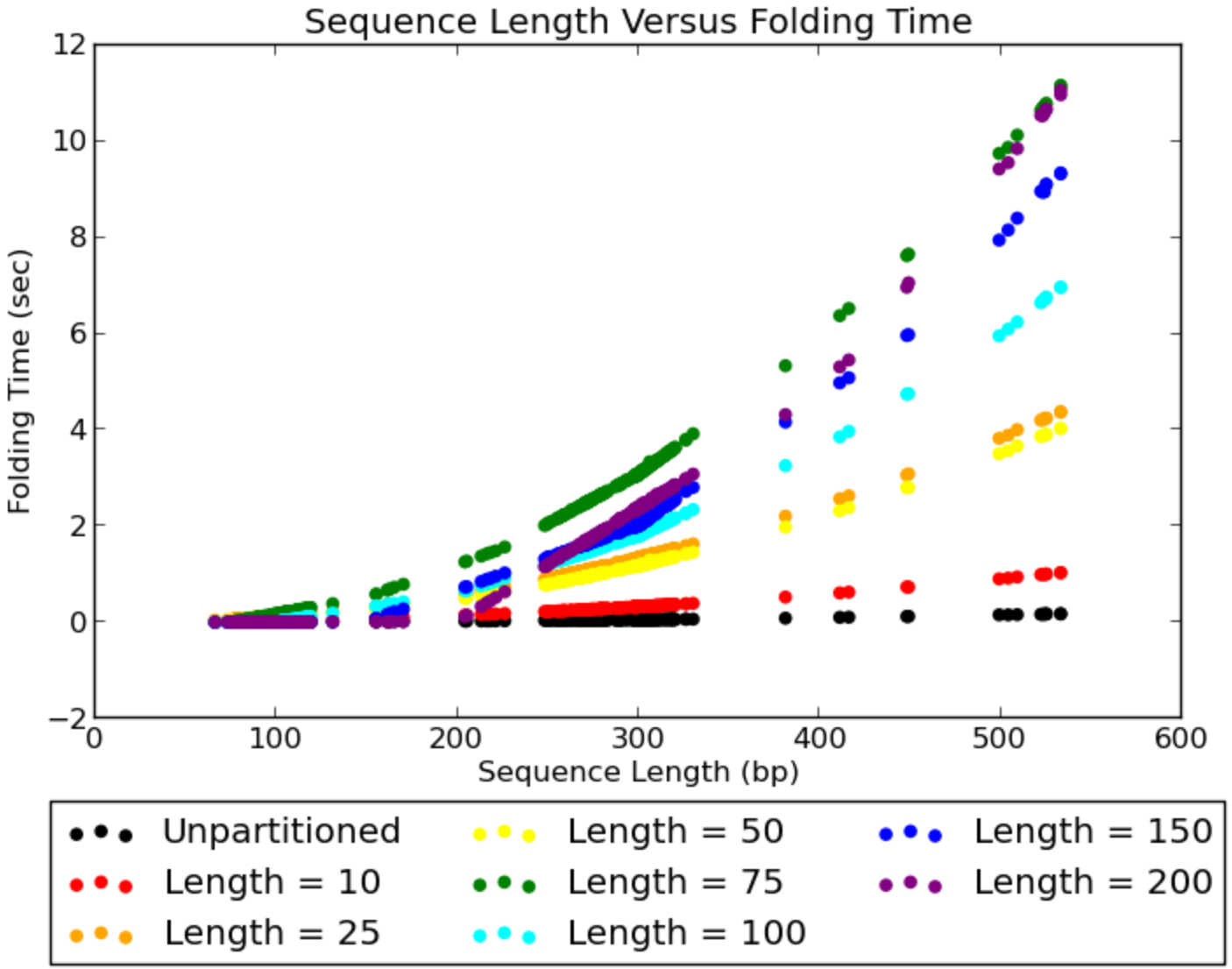
# Details of program

- Language: python
- Database: RNA STRAND
  - SRP subset has 383 RNA sequences + structures
  - In .ct and .dp format, which we parsed in python
- Expected length of single RNA is 300 bases
  - Sequences tested range from 50 to 550 bases
- Can be run on personal laptop

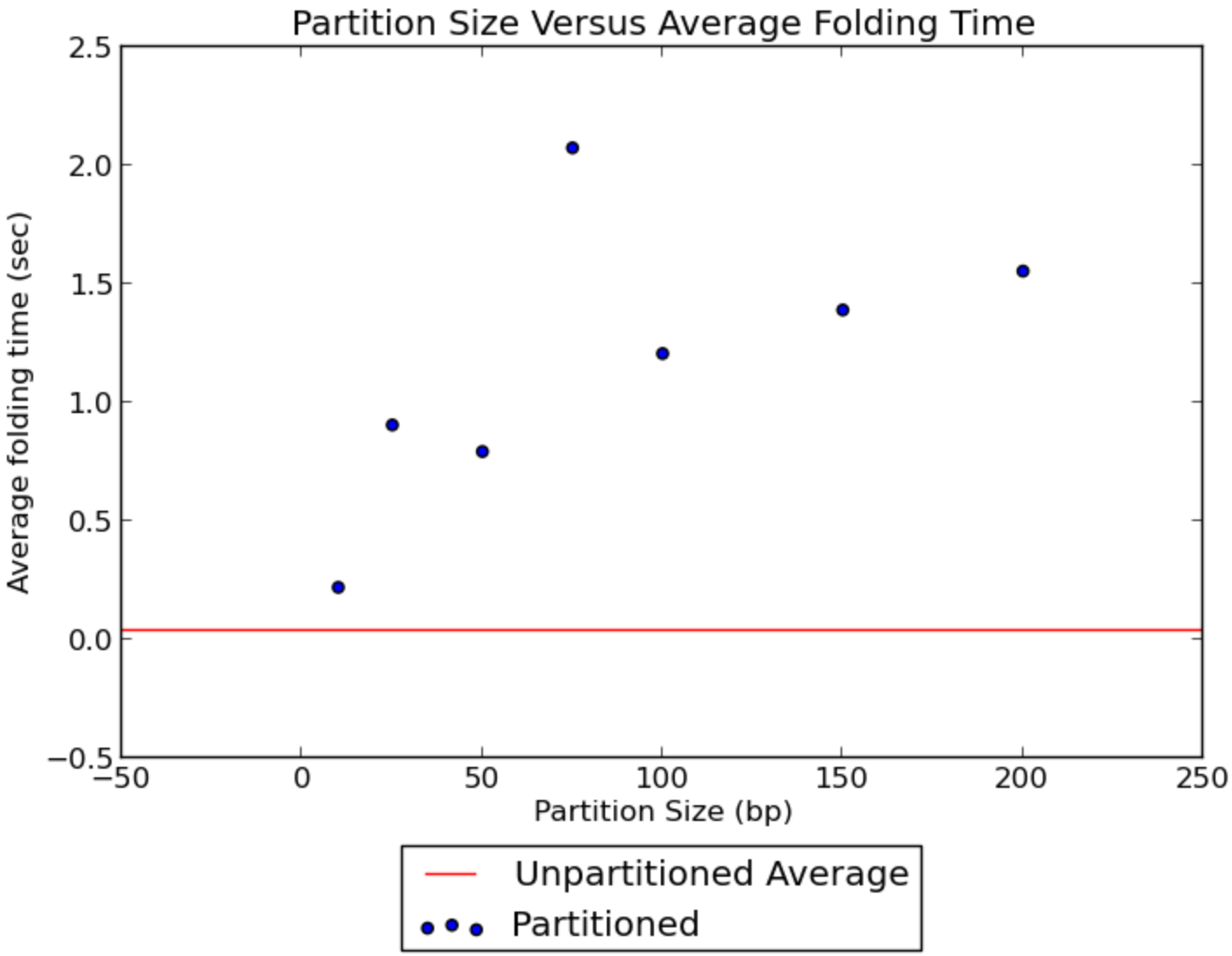
# Results: Score



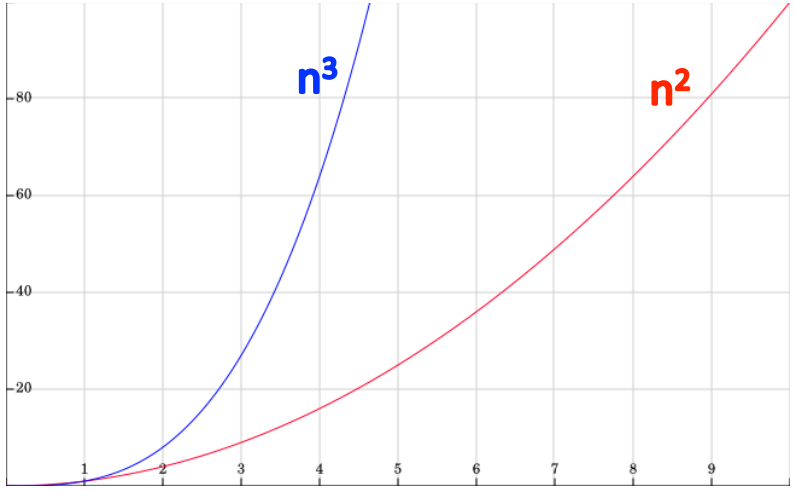
# Results: Runtime

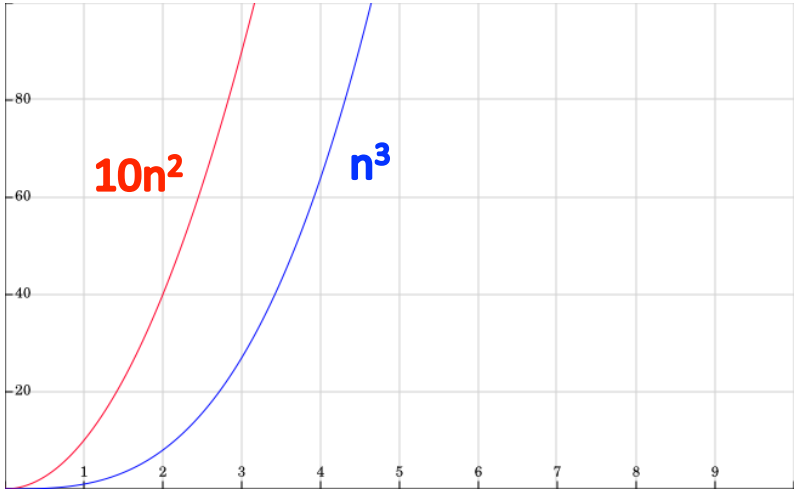


# Results: Runtime



# Runtime Explanation




For small input sizes, function runtimes are significantly affected by constant factors.

# Future Directions

- Multithreading
  - Rewrite in a language amenable to threading
  - Run Nussinov's in parallel on each of the partitions
  - Bring down constant factor in runtime
- Testing biological significance
  - Potential for identifying pseudoknots
  - Potential to be biologically impossible

# References


- Washietl, S. *et al.* (2012) RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res.,* **40**, 4261-4272.
- Nussinov R, Jacobson AB, Fast algorithm for predicting the secondary structure of single-stranded RNA.Proc Natl Acad Sci U S A. 1980 Nov; 77:(11)6309-13.
- Zuker M, Stiegler P Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981 Jan; 9:(1)133-48.
- McCaskill JS The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 1990; 29:(6-7)1105-19.
- Dowell RD, Eddy SR, Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinformatics. 2004 Jun; 5:71.
- Do CB, Woods DA, Batzoglou S, CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics. 2006 Jul; 22:(14)e90-8.
- Andronescu M, Bereg V, Hoos HH, and Condon A: RNA STRAND: The RNA Secondary Structure and Statistical Analysis Database. BMC Bioinformatics. 2008;9(1):340.
- Westbrook J, Feng Z, Chen L, Yang H, Berman H: The Protein Data Bank and structural genomics. Nucleic Acids Res 2003, 31:489-491.
- Cannone J, Subramanian S, Schnare M, Collett J, D'Souza L, Du Y, Feng B, Lin N, Madabusi L, Muller K, Pande N, Shang Z, Yu N, Gutell R: The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 2002, 3:15.
- Andersen ES, Rosenblad MA, Larsen N, Westergaard JC, Burks J, Wower IK, Wower J, Gorodkin J, Samuelsson T, Zwieb C: The tmRDB and SRPDB resources. Nucleic Acids Res 2006, 34(Database issue):163-168.
- Sprinzl M, Vassilenko K: Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res 2005, 33(Database issue):139-140.
- Brown J: The Ribonuclease P Database. Nucleic Acids Res 1999, 27:314-314.
- Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy S, Bateman A: Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 2005, 33(Database issue):121-124.
- Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh SH, Srinivasan AR, Schneider B: The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys J 1992, 63(3):751-759.