# On the Use of Coarse-Grained Thermodynamic Landscapes to Efficiently Estimate Folding Kinetics for RNA Molecules

Evan Senter

2015

# Outline

Overview

Background

Computational RNA background

Problem definition

Optimization using Fast Fourier Transform

# About me

## My background

- B.A. in Computer Science, Computational Biology
- Worked in software engineering for $\approx$ 2 years after
- Started at Boston College in Fall, 2011
- Joined the Clote Lab focusing on Computational RNA Biology

Figure: University of California, Santa Barbara

# Goal of this talk

## Primary aim

Present research on rapidly estimating RNA folding kinetics *in silico*

1. Motivate interest in the study of RNA
2. Highlight interesting roles of non-coding RNAs (ncRNA)
3. Identify biological relevance of folding kinetics
4. Present overview of findings
5. Explain research leading to these findings

# What's the takeaway?

# What's the takeaway?

- A thesis?...

# How biologists See bioinfor*magicians*

# A biologist when stumbling into a math-heavy talk...

# What we aim for…

# Outline

Overview

Background

Computational RNA background

Problem definition

Optimization using Fast Fourier Transform

# Why do we care about RNA?

- Phrase ‘junk DNA’ pigeonholed RNA into predetermined roles

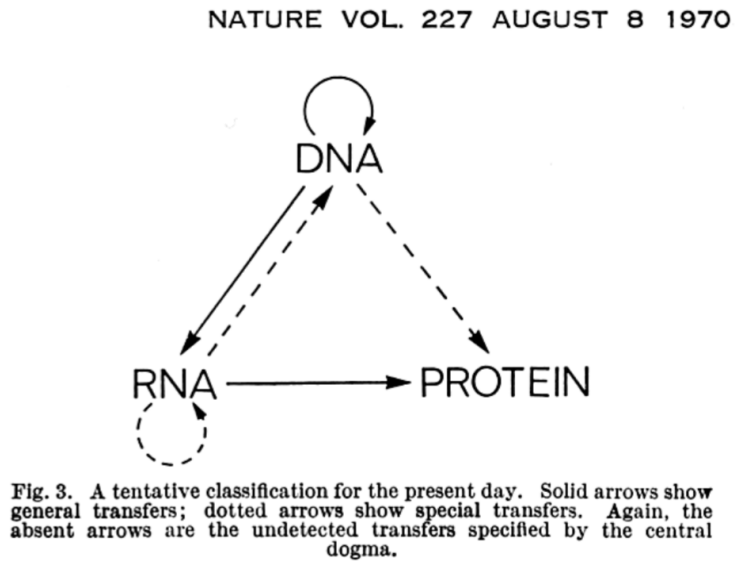



Figure: Crick, F. (1970). Central dogma of molecular biology. Nature.

# Why do we care about RNA?

- Phrase ‘junk DNA’ pigeonholed RNA into predetermined roles
  - Messenger RNA (mRNA)
  - Transfer RNA (tRNA)
  - Ribosomal RNA (rRNA)

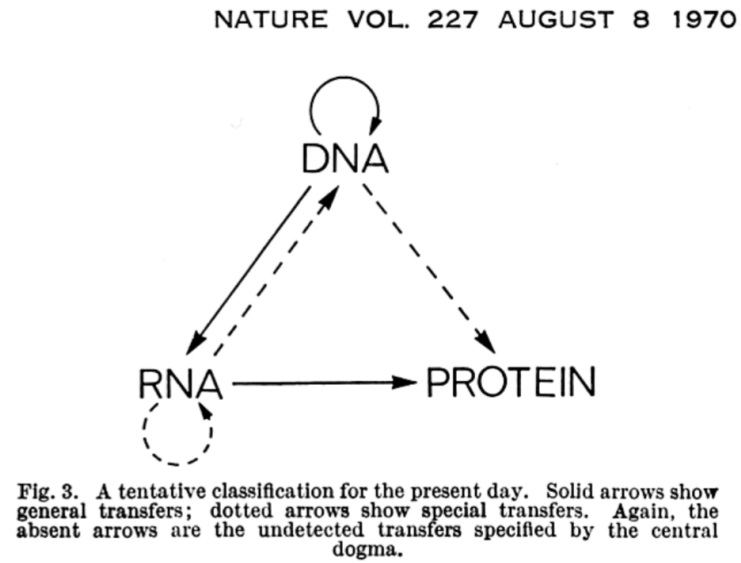



Figure: Crick, F. (1970). Central dogma of molecular biology. Nature.

# Why do we care about RNA?

- Phrase ‘junk DNA’ pigeonholed RNA into predetermined roles
  - Messenger RNA (mRNA)
  - Transfer RNA (tRNA)
  - Ribosomal RNA (rRNA)
- Diverse roles for ncRNA beyond rRNA and tRNA

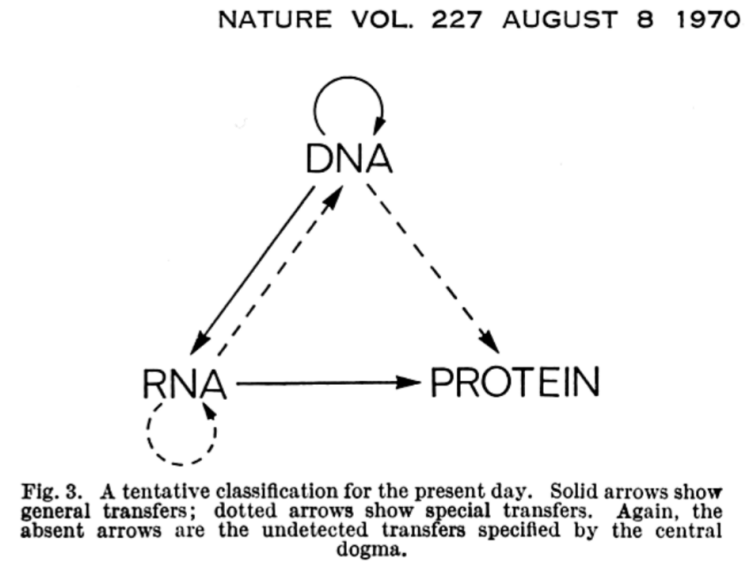



Figure: Crick, F. (1970). Central dogma of molecular biology. Nature.

# ncRNAs—what are they good for?

## The reality

We were not wrong in assigning importance to the aforementioned roles of RNA, but…

# ncRNAs—what are they good for?

We have since found a diverse set of roles for RNA, including…

# ncRNAs—what are they good for?

We have since found a diverse set of roles for RNA, including...

- Peptide bond catalysis

  Nissen, P., Hansen, J., Ban, N., Moore, P. B., & Steitz, T. A. (2000). The structural basis of ribosome activity in peptide bond synthesis. Science (New York, N.Y.), 289(5481), 920–930.

# ncRNAs—what are they good for?

We have since found a diverse set of roles for RNA, including...

- Peptide bond catalysis

  Nissen, P., Hansen, J., Ban, N., Moore, P. B., & Steitz, T. A. (2000). The structural basis of ribosome activity in peptide bond synthesis. Science (New York, N.Y.), 289(5481), 920–930.

- Intron splicing

  Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., Golden, B. L., Kundrot, C. E., et al. (1996). Crystal structure of a group I ribozyme domain: principles of RNA packing. Science (New York, N.Y.), 273(5282), 1678–1685.

# ncRNAs—what are they good for?

We have since found a diverse set of roles for RNA, including...

- Peptide bond catalysis

  Nissen, P., Hansen, J., Ban, N., Moore, P. B., & Steitz, T. A. (2000). The structural basis of ribosome activity in peptide bond synthesis. Science (New York, N.Y.), 289(5481), 920–930.

- Intron splicing

  Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., Golden, B. L., Kundrot, C. E., et al. (1996). Crystal structure of a group I ribozyme domain: principles of RNA packing. Science (New York, N.Y.), 273(5282), 1678–1685.

- Post-transcriptional gene regulation via RNAi

  Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature, 391(6669), 806–811.

# ncRNAs—what are they good for?

- Xist ncRNA for suppression of inactive X chromosome

  Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S., & Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. Nature, 379(6561), 131–137.

# ncRNAs—what are they good for?

- Xist ncRNA for suppression of inactive X chromosome

  Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S., & Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. Nature, 379(6561), 131–137.

- Retranslation events (SECIS elements)

  Walczak, R., Westhof, E., Carbon, P., & Krol, A. (1996). A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. Rna, 2(4), 367–379.

# ncRNAs—what are they good for?

- Xist ncRNA for suppression of inactive X chromosome

  Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S., & Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. Nature, 379(6561), 131–137.

- Retranslation events (SECIS elements)

  Walczak, R., Westhof, E., Carbon, P., & Krol, A. (1996). A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. Rna, 2(4), 367–379.

- Ribosomal frameshift events

  Jacks, T., Power, M. D., Masiarz, F. R., Luciw, P. A., Barr, P. J., & Varmus, H. E. (1988). Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. Nature, 331(6153), 280–283.

  Ofori, L. O., Hilimire, T. A., Bennett, R. P., Brown, N. W., Smith, H. C., & Miller, B. L. (2014). High-affinity recognition of HIV-1 frameshift-stimulating RNA alters frameshifting in vitro and interferes with HIV-1 infectivity. Journal of Medicinal Chemistry, 57(3), 723–732.

# ncRNAs—what are they good for?

And finally…

- Transcriptional and translational regulation via riboswitches

  Nahvi, A., Sudarsan, N., Ebert, M. S., Zou, X., Brown, K. L., & Breaker, R. R. (2002). Genetic Control by a Metabolite Binding mRNA. Chemistry & Biology, 9(9), 1043–1049.

# ncRNAs—what are they good for?

And finally…

- Transcriptional and translational regulation via riboswitches

  Nahvi, A., Sudarsan, N., Ebert, M. S., Zou, X., Brown, K. L., & Breaker, R. R. (2002). Genetic Control by a Metabolite Binding mRNA. Chemistry & Biology, 9(9), 1043–1049.

- Spliced leader (SL) trans-splicing events in *L. collosoma*

  LeCuyer, K. A., & Crothers, D. M. (1993). The *Leptomonas collosoma* spliced leader RNA can switch between two alternate structural forms. Biochemistry, 32(20), 5301–5311.

# ncRNAs—what are they good for?

And finally…

- Transcriptional and translational regulation via riboswitches

  Nahvi, A., Sudarsan, N., Ebert, M. S., Zou, X., Brown, K. L., & Breaker, R. R. (2002). Genetic Control by a Metabolite Binding mRNA. Chemistry & Biology, 9(9), 1043–1049.

- Spliced leader (SL) trans-splicing events in *L. collosoma*

  LeCuyer, K. A., & Crothers, D. M. (1993). The *Leptomonas collosoma* spliced leader RNA can switch between two alternate structural forms. Biochemistry, 32(20), 5301–5311.

- hok/sok postsegregational killing mechanism in *E. coli*

  Gerdes, K., Rasmussen, P. B., & Molin, S. (1986). Unique type of plasmid maintenance function: postsegregational killing of plasmid-free cells. Proceedings of the National Academy of Sciences, 83(10), 3116–3120.

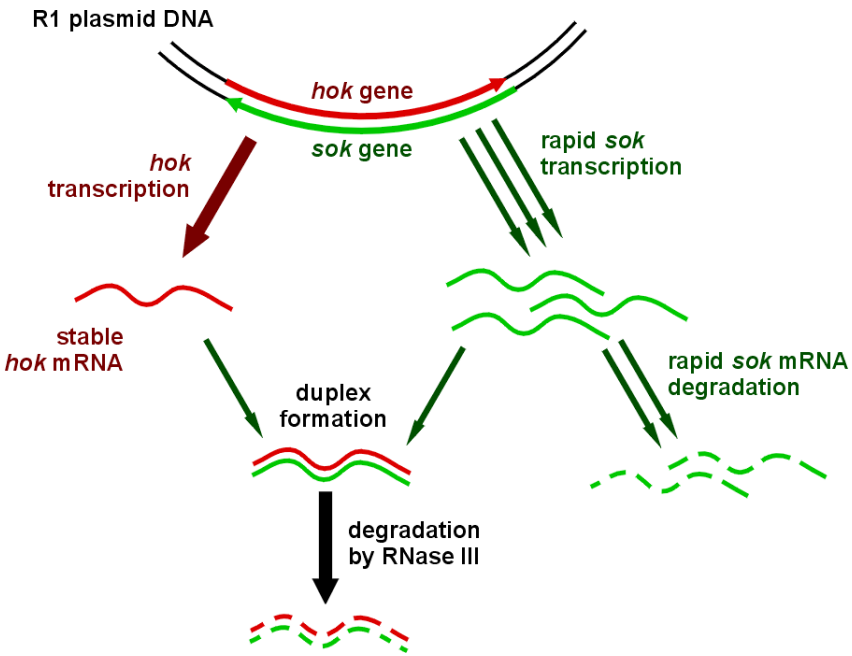
  Nagel, J. H., Gultyaev, A. P., Gerdes, K., & Pleij, C. W. (1999). Metastable structures and refolding kinetics in hok mRNA of plasmid R1. Rna, 5(11), 1408–1418.

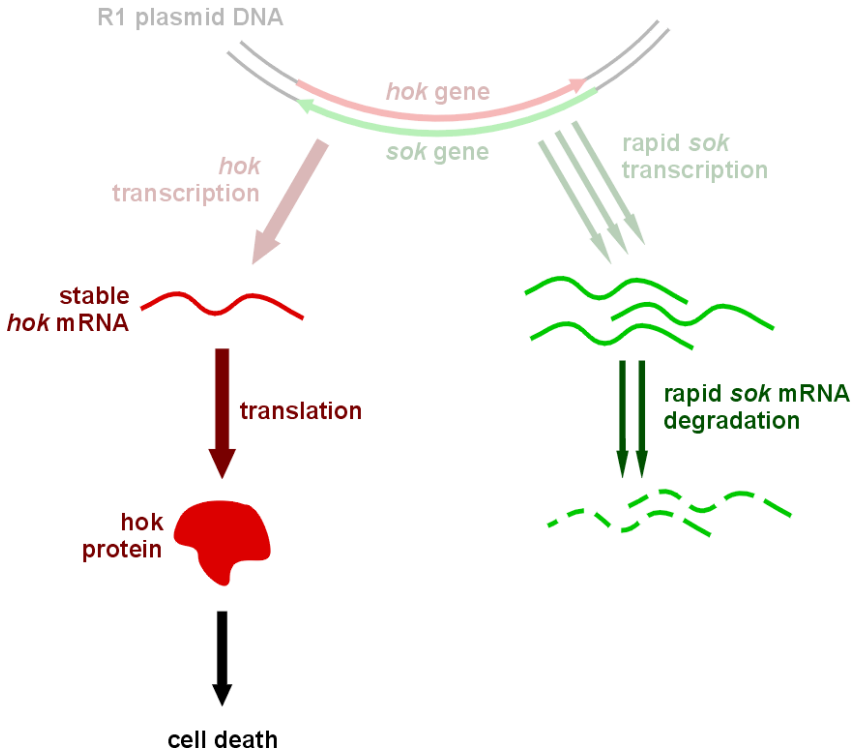
# ncRNAs—what are they good for?

## Summary

ncRNAs have diverse cellular responsibilities, beyond the canonical tRNA and rRNA examples

# *hok/sok* and kinetics




https://en.wikipedia.org/wiki/File:Hok_sok_system_R1_plasmid_present.gif
https://en.wikipedia.org/wiki/File:Hok_sok_system_R1_plasmid_absent.gif

# *hok/sok* structures

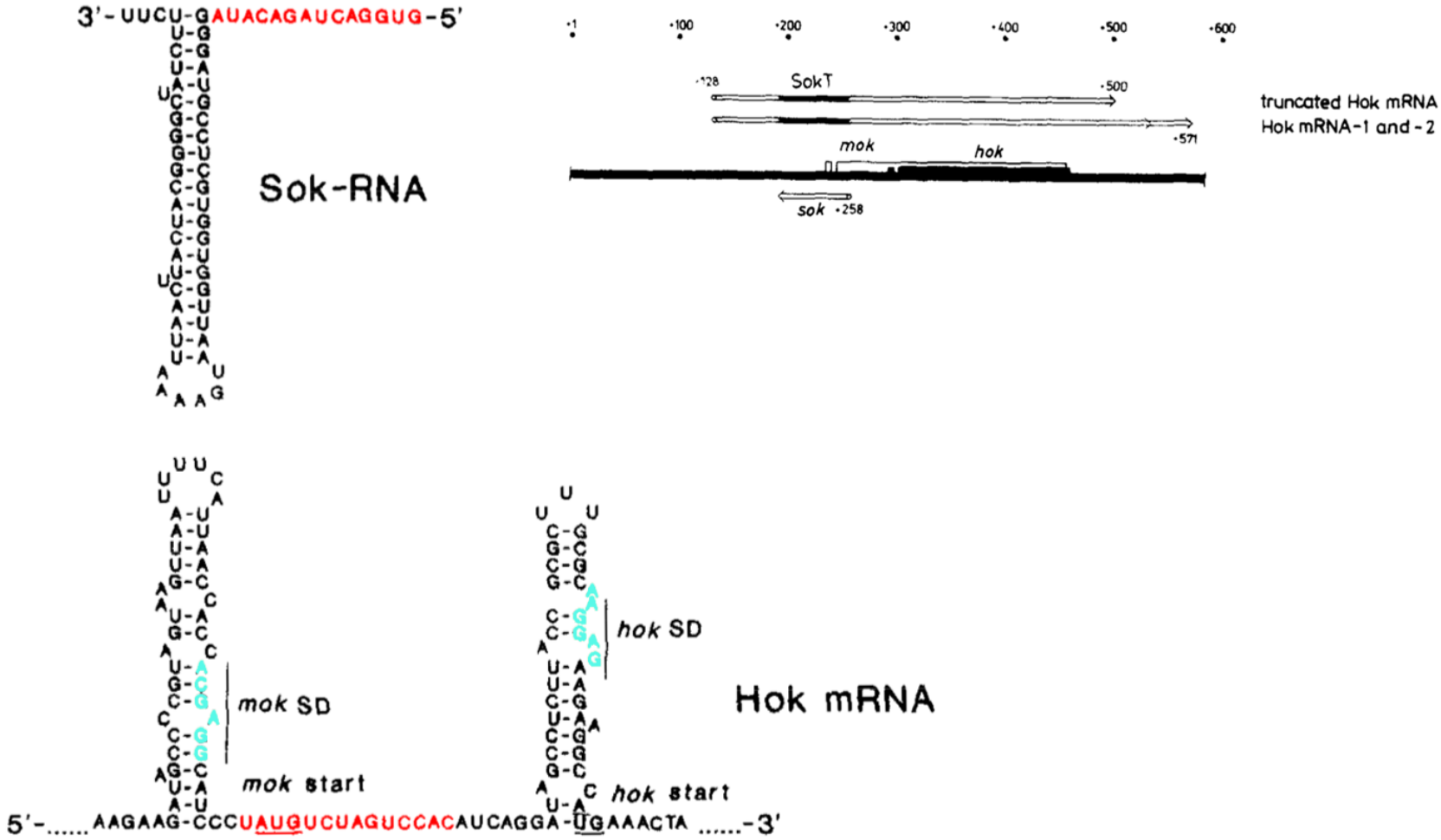



Figure: Adapted from Thisted, T., & Gerdes, K. (1992). Mechanism of post-segregational killing by the *hok*/*sok* system of plasmid R1. Sok antisense RNA regulates *hok* gene expression indirectly through the overlapping *mok* gene. Journal of Molecular Biology, 223(1), 41–54.

# *hok* folding kinetics

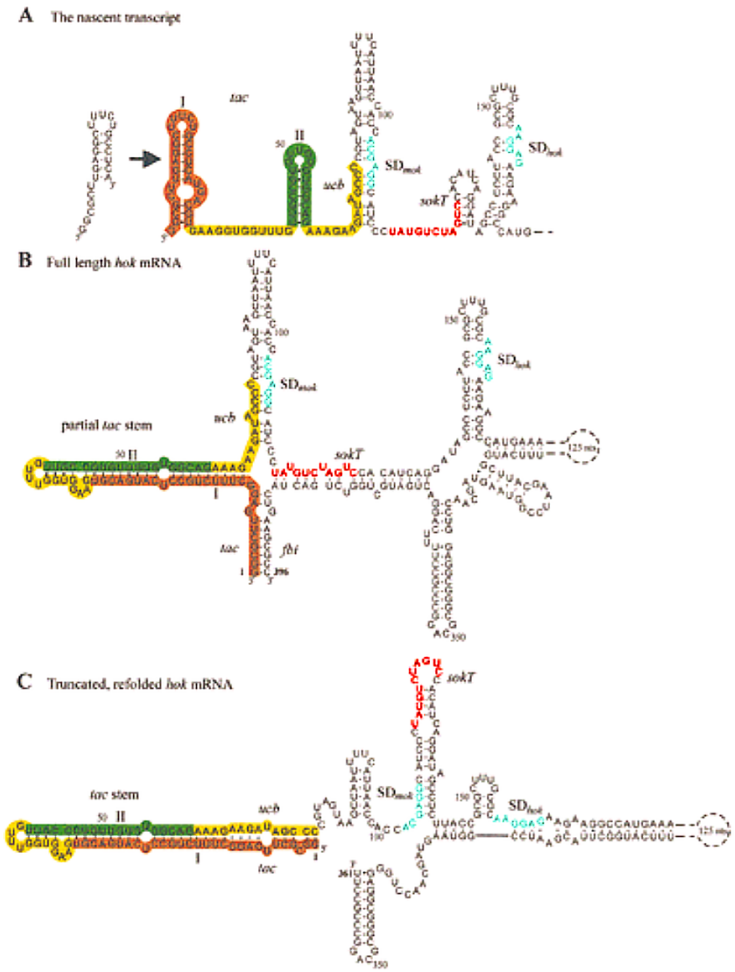



Figure: Adapted from Nagel, J. H., Gultyaev, A. P., Gerdes, K., & Pleij, C. W. (1999). Metastable structures and refolding kinetics in hok mRNA of plasmid R1. RNA, 5(11), 1408–1418.

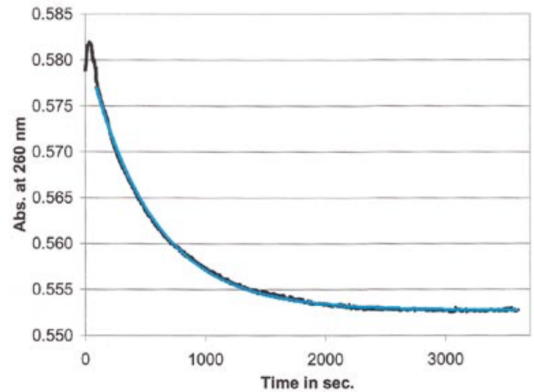



**Figure 2.** The kinetic refolding from the metastable to the stable conformation in the *hok*[74] RNA fragment. The metastable conformation was trapped by the heating/cooling cycle and the kinetics monitored at 260 nm in a UV spectrophotometer at 37°C in 950 mM NaCl and 50 mM Na cacodylate buffer, pH 7.2. The measured real-time curve (in black) and the first-order exponentially fitted curve is indicated in light blue (fitted parameters $t_{1/2} = 669 \pm 1$ s).

Figure: Nagel, J. H. A., Gultyaev, A. P., Oistämö, K. J., Gerdes, K., & Pleij, C. W. A. (2002). A pH-jump approach for investigating secondary structure refolding kinetics in RNA. Nucleic Acids Research, 30(13), e63.

# Outline

Overview

Background

Computational RNA background

Problem definition

Optimization using Fast Fourier Transform

# RNA Representation

## Sequence

An RNA sequence is a string $\mathbf{s} = s_1, \ldots, s_n$, where $s_i \in \{\text{A}, \text{U}, \text{G}, \text{C}\}$

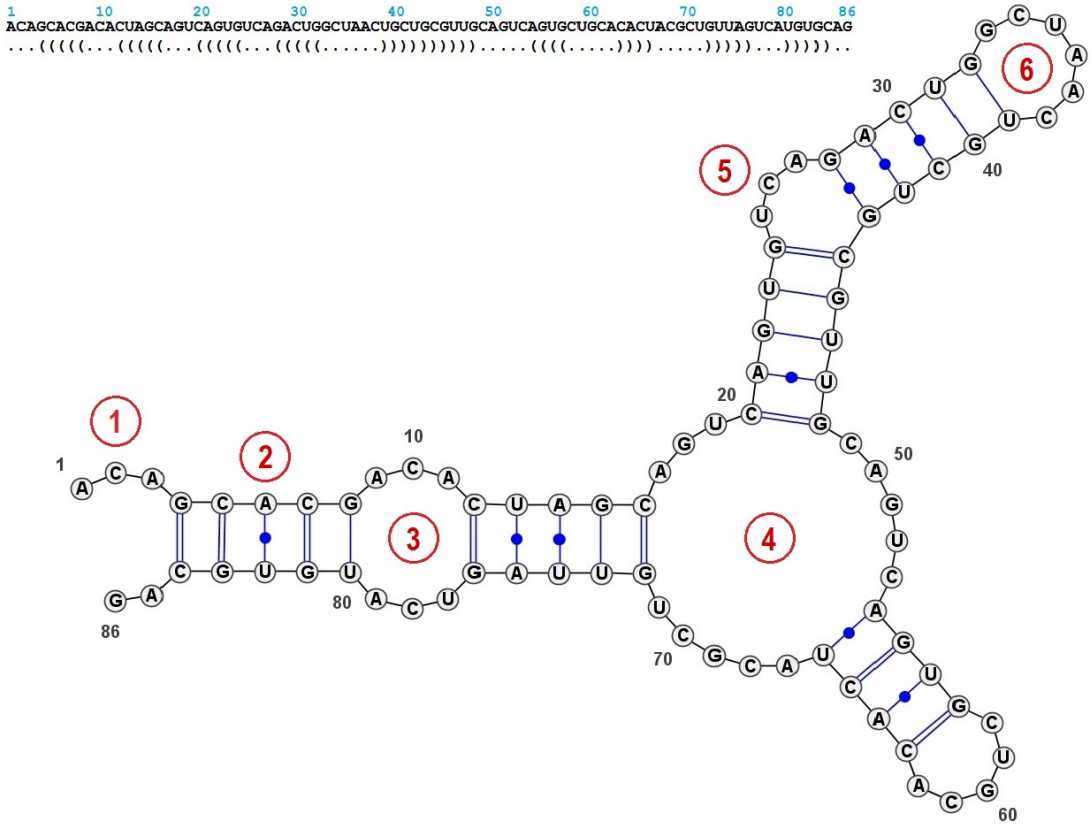
## Structure

An secondary structure $\mathcal{S}$ compatible with $\mathbf{s}$ is a collection of base pair tuples such $(i, j)$, such that:

- $(\mathbf{s}_i, \mathbf{s}_j) \in \mathbb{B}$
- $1 \leq i \leq i + \theta < j \leq n$ where $\theta \geq 0$
- Given $(i, j), (x, y)$ from $\mathcal{S}$, $i = x \iff j = y$
- Given $(i, j), (x, y)$ from $\mathcal{S}$, $i < x < j \iff i < y < j$

$$\mathbb{B} = \{(\text{A}, \text{U}), (\text{U}, \text{A}), (\text{G}, \text{C}), (\text{C}, \text{G}), (\text{G}, \text{U}), (\text{U}, \text{G})\}$$

# Structural Motifs




## Structural Motifs

1. Exterior loop
2. Stack
3. Interior loop
4. Multiloop
5. Bulge
6. Hairpin

# Base Pair Distance

## Symmetric distance

$$d_{\mathsf{BP}}(\mathcal{S}, \mathcal{T}) = |\mathcal{S} \cup \mathcal{T}| - |\mathcal{S} \cap \mathcal{T}|$$

## Distance between two structures

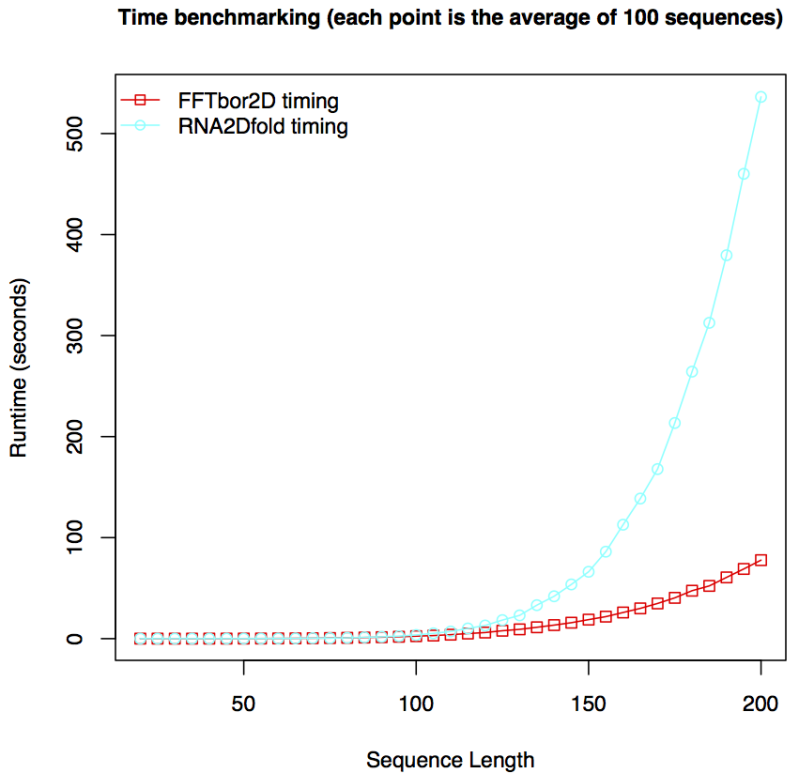
$$d_{\mathsf{BP}}(\mathcal{S}_{[i,j]}, \mathcal{T}_{[i,j]}) = |\{(x, y) : i \leq x < y \leq j, \\ (x, y) \in \mathcal{S} - \mathcal{T} \text{ or } (x, y) \in \mathcal{T} - \mathcal{S}\}| = k$$

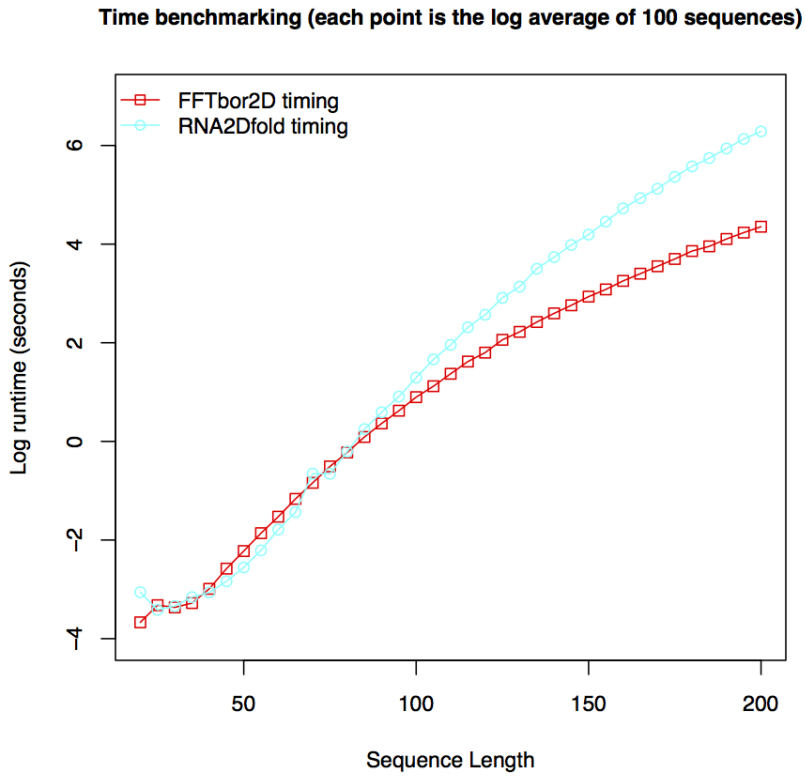
# Comparison of various kinetics programs

| Hastings (Yes\No) | RNAmfpt | RNAeq | Kinfold | FFTmfpt | RNA2Dfold | FFTbor | BarriersEq | FFTeq |
|---|---|---|---|---|---|---|---|---|
| RNAmfpt | 1 | 0.5683 | 0.7945 | 0.5060 | 0.5110 | 0.5204 | 0.5280 | 0.4472 |
| RNAeq | 0.5798 | 1 | 0.7814 | 0.7043 | 0.7025 | 0.5080 | 0.5979 | 0.6820 |
| Kinfold | 0.7933 | 0.7507 | 1 | 0.7312 | 0.7358 | 0.6241 | 0.6328 | 0.6445 |
| FFTmfpt | **0.6035** | 0.7935 | 0.7608 | 1 | 0.9980 | 0.5485 | 0.8614 | 0.9589 |
| RNA2Dfold | 0.6076 | 0.7919 | 0.7655 | 0.9983 | 1 | 0.5584 | 0.8538 | 0.9515 |
| FFTbor | 0.5416 | 0.5218 | 0.6241 | 0.5748 | 0.5855 | 1 | 0.3450 | 0.4229 |
| BarriersEq | 0.6346 | 0.6578 | 0.6328 | 0.8310 | 0.8217 | 0.3450 | 1 | 0.9149 |
| FFTeq | 0.5614 | 0.7916 | 0.6966 | 0.9670 | 0.9590 | 0.4757 | 0.8940 | 1 |

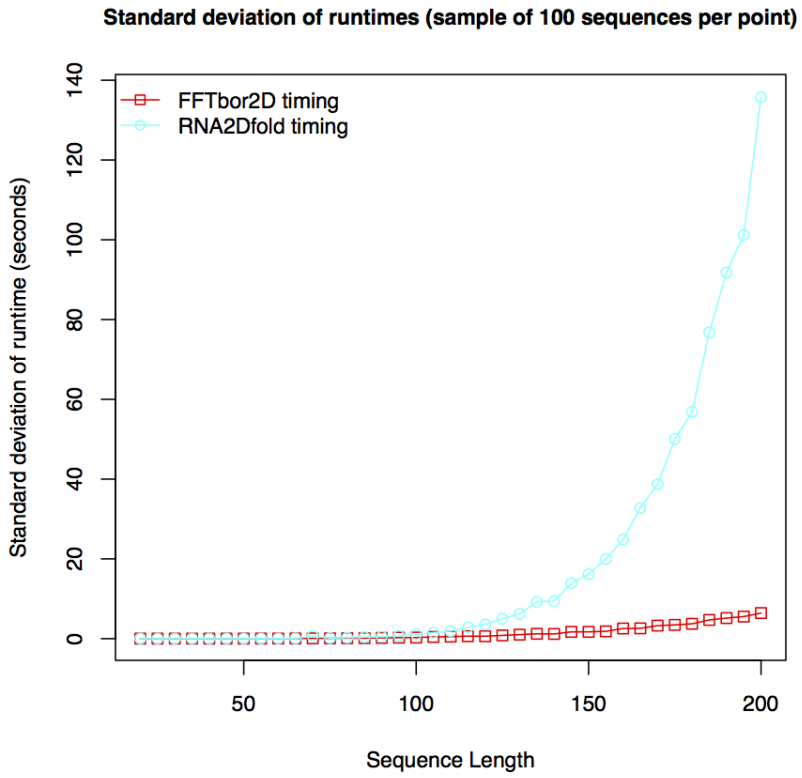
- RNAmfpt, FFTmfpt, RNAeq, and FFTeq included in the Hermes package
- RNA2Dfold (Lorenz *et. al.*, 2009), BarriersEq (Flamm *et. al.*, 2002), and FFTbor (Senter *et. al.*, 2012) kinetics computed with Hermes

# Performance Characteristics



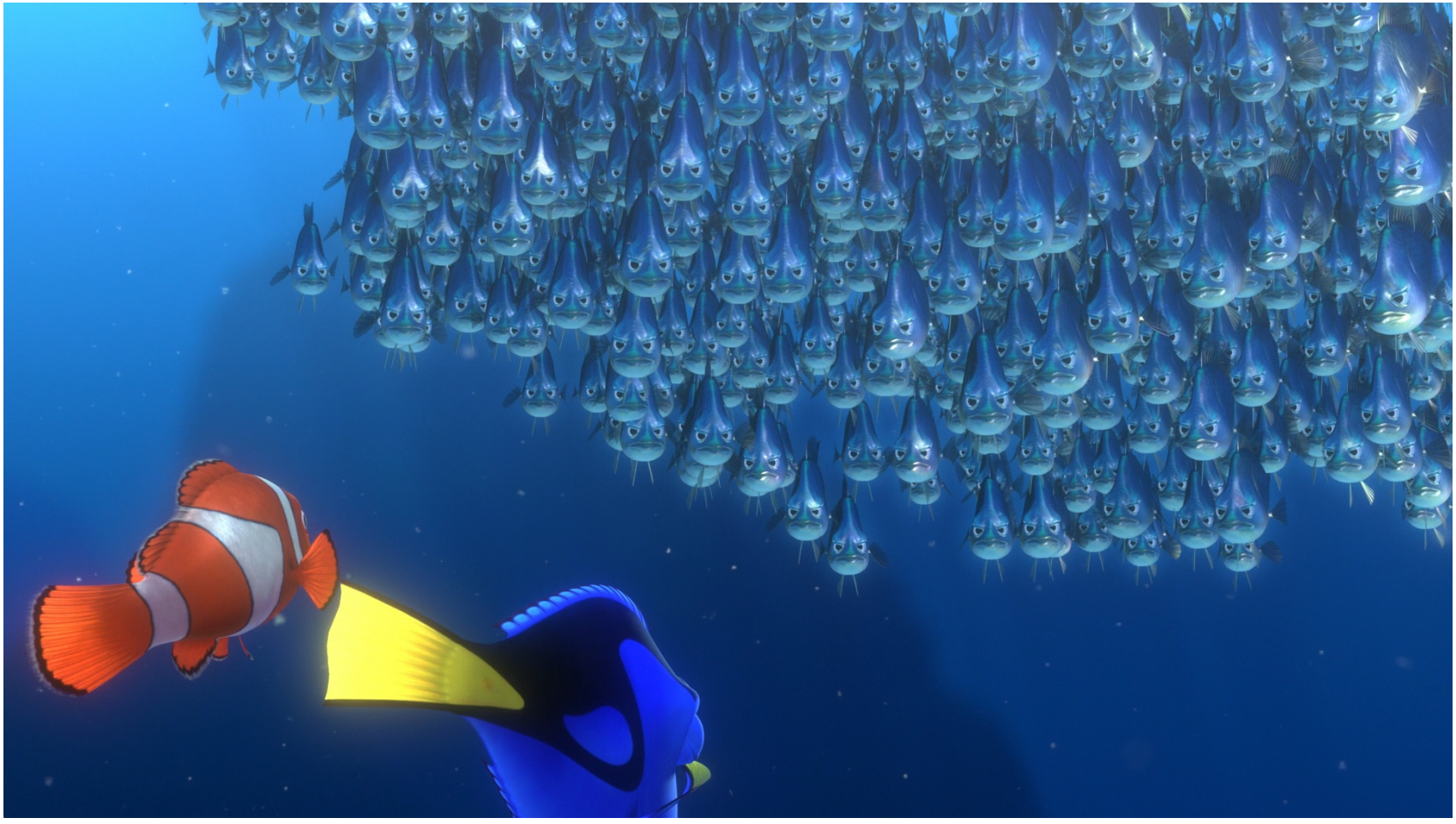
# Performance Characteristics




- Approach using FFT goes from $O(n^7)$ to $O(n^5)$
- We observe a real performance gain in line with 100x speedup
- Memory requirements drop from $O(n^4)$ to $O(n^2)$
- More consistent performance characteristics

# And of course my labmates and fellow grad students!

# Questions?



# Outline

Overview

Background

Computational RNA background

Problem definition

Optimization using Fast Fourier Transform

# Problem Definition

## Desire

Given an input sequence $\mathbf{s}$ and two input structures $\mathcal{A}$, $\mathcal{B}$, we would like to compute all possible structures $\mathcal{S}$ compatible with $\mathbf{s}$, and bin them into discrete sets based on their *distance* to $\mathcal{A}$ and $\mathcal{B}$

## Issue

Consider $\mathbb{S}$ to be the set of all structures compatible with $\mathbf{s}$. It has been shown that $|\mathbb{S}|$ grows exponentially with sequence length $n$

## Refinement

Rather than store $\mathbb{S}$ at any point in time, we will use dynamic programming to compute the thermodynamic properties of these bins

# Concrete Example

## Input

$\texttt{GGAAACC} = \mathbf{s}$

$\texttt{.......} = \mathcal{A}$

$\texttt{.(...).} = \mathcal{B}$

## Structures

$\texttt{.......}\ \ 0.00\frac{\text{kcal}}{\text{mol}}, [0, 1]$

$\texttt{.(...).}\ \ 4.40\frac{\text{kcal}}{\text{mol}}, [1, 0]$

$\texttt{(....).}\ \ 2.30\frac{\text{kcal}}{\text{mol}}, [1, 2]$

$\texttt{.(....)}\ \ 4.10\frac{\text{kcal}}{\text{mol}}, [1, 2]$

$\texttt{(.....)}\ \ 4.20\frac{\text{kcal}}{\text{mol}}, [1, 2]$

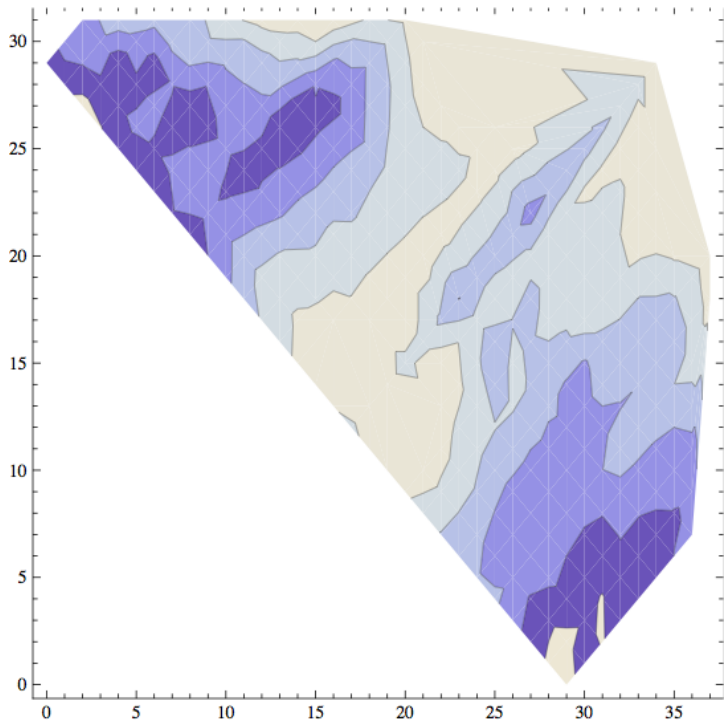
$\texttt{((...))}\ \ 2.10\frac{\text{kcal}}{\text{mol}}, [2, 1]$

## Output

$$\begin{bmatrix} 0 & 0.9595 & 0 \\ 0.0001 & 0 & 0.0086 \\ 0 & 0.0318 & 0 \end{bmatrix}$$

# Concrete Example

Energy landscape between two metastable structures of *L.collosoma* spliced leader RNA



# Parameterized Partition Function, 1D

## $\mathbf{Z}$ binned by $k$

$$\mathbf{Z}^k = \mathbf{Z}^k_{1,n} = \sum_{\substack{\mathcal{S} \text{ such that} \\ d_{\mathrm{BP}}(\mathcal{S},\mathcal{S}^*)=k}} e^{\frac{-E(\mathcal{S})}{RT}}$$

# Recursions to compute $\mathbf{Z}_{i,j}^{k}$

## Structural decomposition from one target

$$\mathbf{Z}_{i,j}^{k} = \mathbf{Z}_{i,j-1}^{k-b_0} + \sum_{\substack{s_r s_j \in \mathbb{B}, \\ i \leq r < j}} \left( e^{\frac{-E_0(r,j)}{RT}} \sum_{w+w'=k-b(r)} \mathbf{Z}_{i,r-1}^{w} \mathbf{Z}_{r+1,j-1}^{w'} \right)$$

# Parameterized Partition Function, 2D

$\mathbf{Z}$ binned by $x, y$ pairs

$$\mathbf{Z}_{1,n}^{x,y} = \sum_{\substack{\mathcal{S} \text{ such that} \\ d_{\mathsf{BP}}(\mathcal{S},\mathcal{A})=x,\, d_{\mathsf{BP}}(\mathcal{S},\mathcal{B})=y}} e^{\frac{-E(\mathcal{S})}{RT}}$$

# Recursions to compute $\mathbf{Z}_{i,j}^{x,y}$

## Structural decomposition from two targets

$$\mathbf{Z}_{i,j}^{x,y} = \mathbf{Z}_{i,j-1}^{x-\omega_0,y-\beta_0} + \sum_{\substack{s_k s_j \in \mathbb{B}, \\ i \leq k < j}} \left( e^{\frac{-E_0(k,j)}{RT}} \sum_{u+u'=x-\omega(k)} \sum_{v+v'=y-\beta(k)} \mathbf{Z}_{i,k-1}^{u,v} \cdot \mathbf{Z}_{k+1,j-1}^{u',v'} \right)$$

# Partition function of a variable $x$

## Only compute $\mathcal{Z}_{i,j}(x)$x instead of $\mathbf{Z}_{i,j}^{x,y}$

$$\mathcal{Z}_{i,j}(x) = \mathcal{Z}_{i,j-1}(x) \cdot x^{\omega_0 n + \beta_0} + \sum_{\substack{s_k s_j \in \mathbb{B}, \\ i \leq k < j}} \left( e^{\frac{-E_0(k,j)}{RT}} \cdot \mathcal{Z}_{i,k-1}(x) \cdot \mathcal{Z}_{k+1,j-1}(x) \cdot x^{\omega(k) n + \beta(k)} \right)$$

# Outline

Overview

Background

Computational RNA background

Problem definition

Optimization using Fast Fourier Transform

# FFT background

## Complex *k*th roots of unity

$$\omega_0 = \exp(\frac{0 \cdot 2\pi i}{n^2}), \omega_1 = \exp(\frac{1 \cdot 2\pi i}{n^2}), \ldots, \omega_{n^2-1} = \exp(\frac{(n^2-1) \cdot 2\pi i}{n^2})$$

## Evaluate $\mathcal{Z}_{i,j}(x)$x for all $n^2$ roots of unity

$$y_0 = \mathcal{Z}(\omega_0), \ldots, y_{n^2-1} = \mathcal{Z}(\omega_{n^2-1}))$$

## Represent results of evaluation in column form

$$\mathbf{Y} = (y_0, \ldots, y_{n^2-1})^{\mathsf{T}}$$

# Vandermonde matrix

## Matrix construction

$$V_n = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \omega & \omega^2 & \dots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \dots & \omega^{2(n-1)} \\ 1 & \omega^3 & \omega^6 & \dots & \omega^{3(n-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \omega^{n-1} & \omega^{2(n-1)} & \dots & \omega^{(n-1)(n-1)} \end{pmatrix}$$

## Definition

Define the FFT to be the $O(n \log n)$ algorithm to compute the Discrete Fourier Transform (DFT), defined as the matrix product $\mathbf{Y} = V_n \mathbf{A}$

$$\begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_{n^2-1} \end{pmatrix} = V_n \cdot \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_{n^2-1} \end{pmatrix}$$

Since we defined $\mathbf{Y} = (y_0, \ldots, y_{n-1})^{\mathsf{T}}$, where:

$$y_0 = \mathcal{Z}(\omega_0), \ldots, y_{n^2-1} = \mathcal{Z}(\omega n^2 - 1))$$

and $\omega_k = \exp(\frac{2\pi ki}{n^2})$, it follows that the coefficients $c_{rn+s} = \mathbf{Z}_{1,n}^{rn+s}$ in the polynomial:

$$\mathcal{Z}(x) = c_0 + c_1 x + \cdots + c_{n^2-1} x^{n^2-1}$$

can be computed using the Fast Fourier Transform, and:

$$c_{rn+s} = = \sum_{\substack{\mathcal{S} \text{ such that} \\ d_{\mathsf{BP}}(\mathcal{S}, \mathcal{A}) = r,\, d_{\mathsf{BP}}(\mathcal{S}, \mathcal{B}) = s}} e^{\frac{-E(\mathcal{S})}{RT}}$$