Analytic Combinatorics Applied To RNA Structures

Christie Burris

May 8th, 2018

Master's Thesis Defense

Virginia Tech
Department of Mathematics and The Biocomplexity Institute

Mathematical Biocomplexity Lab

- **Problems**: Molecular Biology, Evolution
- **Tools**: Analytic Combinatorics, Topological Graph Theory, Complex Analysis, and Probability Theory
- **Starting point**: where this research stems from
- **Results**: information on structures

Background

Central Dogma:

DNA → RNA → Protein

(double stranded) (single stranded) (amino acids)

Our work:

- Based on the idea that the central dogma is not the whole story
- Focus on noncoding RNA (98%)
- Structure as important as sequence

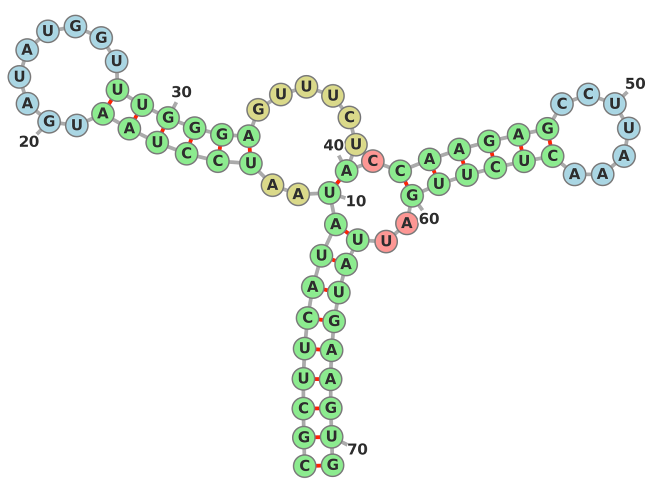
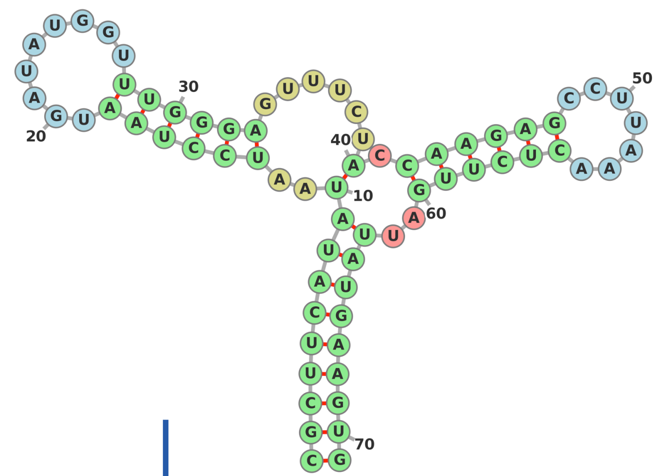


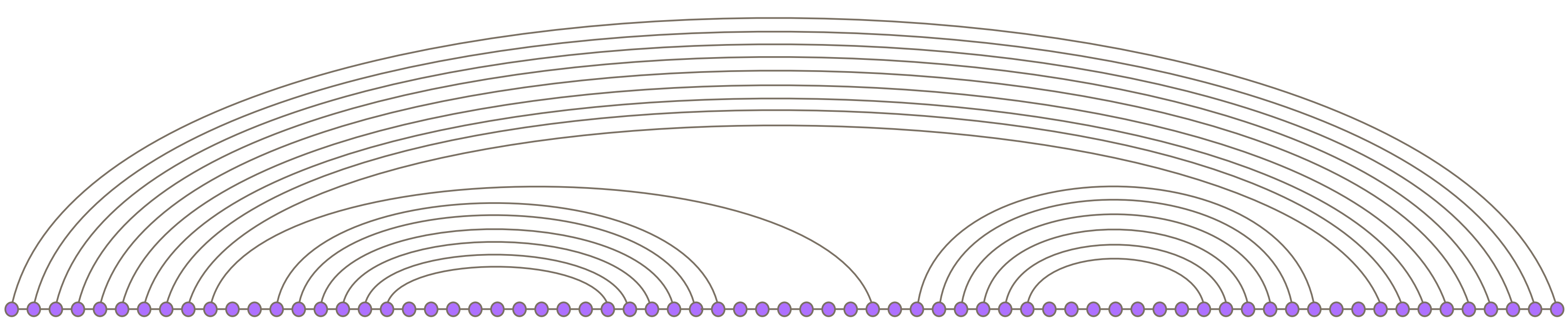
Figure drawn fromViennaRNA Web Servies,
Kerpedjiev P, Hammer S, Hofacker IL (2015).

Structures to Diagrams

Nucleotides ⟷ Vertices
Pairings ⟷ Edges

- Vertex degree ≤ 3
- Distinguish the backbone (horizontal line)
- *1-arcs* are different from backbone

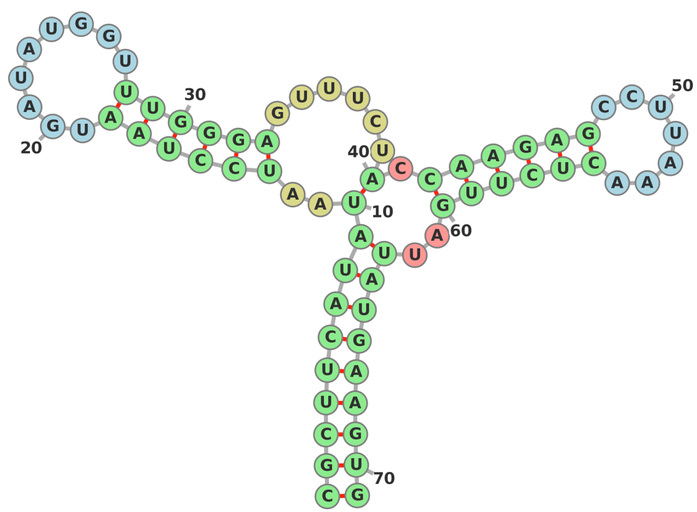


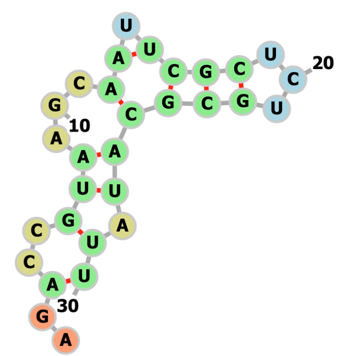


Research Question

Q: *How does the length of the irreducible component change as the complexity of the structures increases?*

2-Dimensional (secondary structures) → 3-Dimensional (pseudoknot structures)





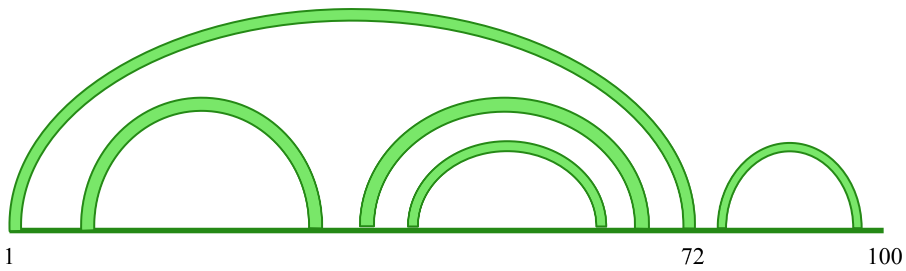
Figures drawn fromViennaRNA Web Servies, Kerpedjiev P, Hammer S, Hofacker IL (2015).

Starting Point: Secondary Structures

Theorem (Li and Reidys, 2018): *The expectation and variance of the length of the longest rainbow in secondary structures is given by*

$$E[Y_n] = n - \alpha n^{\frac{1}{2}}(1+o(1)), \quad \alpha = 2.482$$

$$V[Y_n] = \beta n^{\frac{3}{2}}(1+o(1)).$$



The rainbow-spectrum of RNA secondary structures

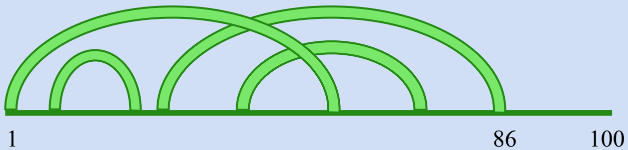
Thomas J. X. Li · Christian M. Reidys

Generalized Results

Theorem 1: *The expectation and variance of the length of the longest block in γ-structures for γ = 1 is given by*

$$E[B_n] = n - \alpha n^{\frac{1}{2}}(1+o(1)), \quad \alpha = 1.416$$

$$V[B_n] = \beta n^{\frac{3}{2}}(1+o(1)), \quad \beta = 0.304.$$



Theorem 2: *For fixed k, the distribution of the number of blocks of length k is negative binomial.*

Theorem 3: *For any* $k = o(n)$, $\lim_{n\to\infty} P(n - B_n = k) = \frac{[z^k]G_\tau^2(z)\mu^k}{G_\tau^2(\mu)}$.

Analytic Combinatorics

Flajolet, P and Sedgewick, R. *Analytic Combinatorics*, 2009.

Begin: Generalize RNA structures to graph diagrams

Combinatorial Classes

↓

Generating Functions

↓

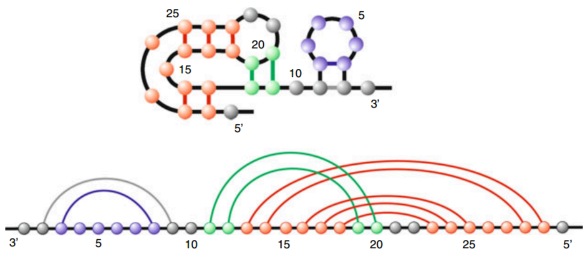
Coefficient Asymptotics

Symbolic enumeration – deriving generating functions without recurrence relations

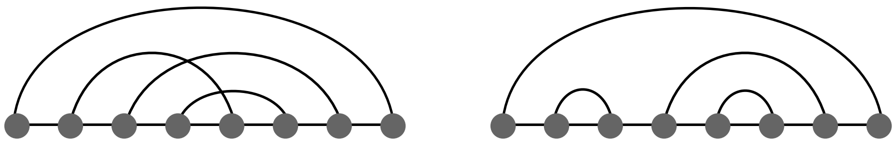
- Find a functional form of generating function
- Find asymptotic expansion of function at its dominant singularity
- Extract coefficient asymptotics using analytic transfer theorems

Results: Expectation and variance of the longest irreducible component

Structures with Crossings



Why are structures with crossing more difficult to study (in a Bioinformatics sense)?



with crossings

without crossings

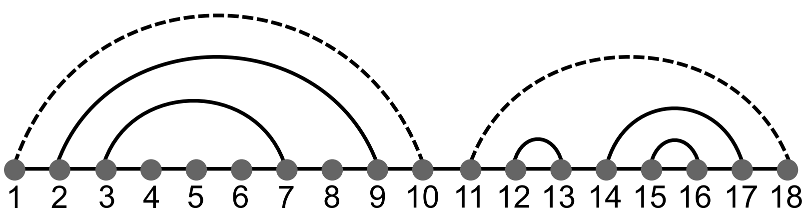
General crossing diagrams have a more complex notion of irreducibility, if at all.

Irreducibility

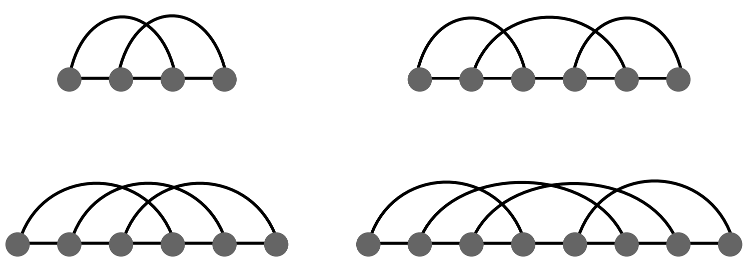
An equivalence relation on arcs:

For any arcs $\alpha_1, \alpha_k \in E$ we say $\alpha_1 \sim \alpha_k$ iff there exists a sequence of arcs $(\alpha_1, \ldots, \alpha_{k-1}, \alpha_k)$ such that α_i and α_{i+1} are crossing for every $1 \leq i \leq k$.

- *Irreducible diagram:* for any two arcs α_i and α_j in its maximal component, $\alpha_i \sim \alpha_j$.
- A diagram can be partitioned into ***blocks*** (irreducible components).



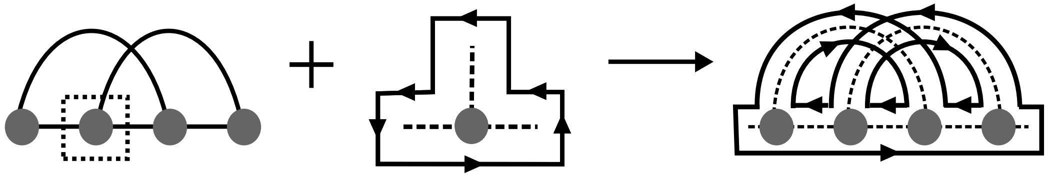
For secondary structures, irreducible components are maximal arcs (rainbows).



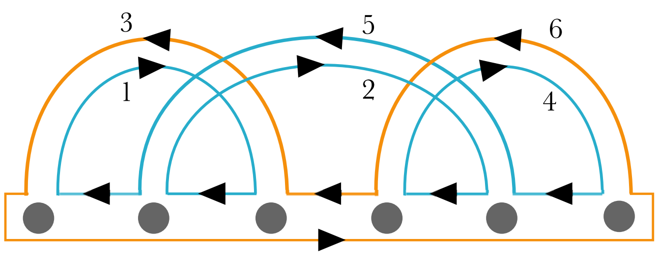
Examples of irreducible components for crossing structures.

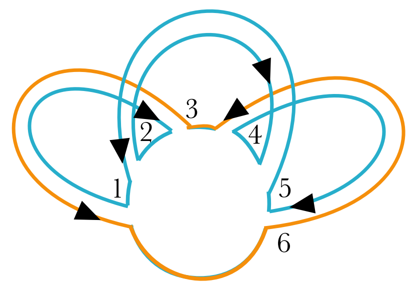
Topological Graph Theory

"fattening":



A *ribbon graph*, or *fatgraph*, is a triple (H, σ, α) where H is the set of half-edges, σ is the vertex permutation, and α is a fixed-point free involution.





$V = 1$

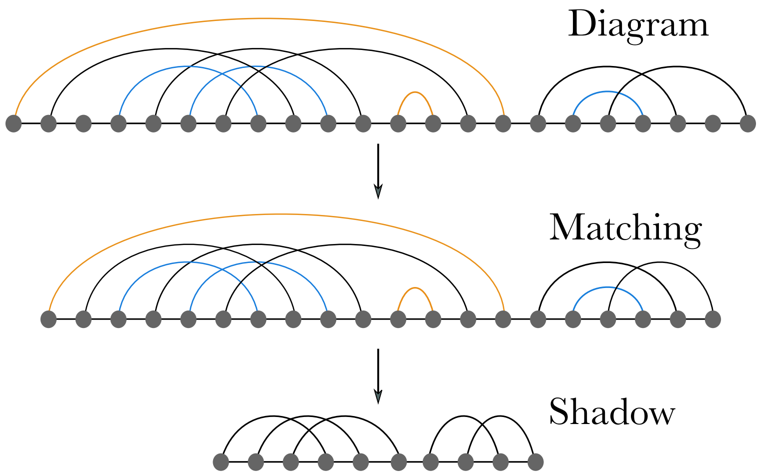
$E = \#\, arcs$

$R = \#\, faces$

Computing genus: $g = 1 - \frac{1}{2}\left(V - E + R\right)$

γ-Structures

γ-structure: a diagram whose maximal irreducible components are of genus ≤ γ.



γ = 1

Building blocks of γ-structures are irreducible shadows.

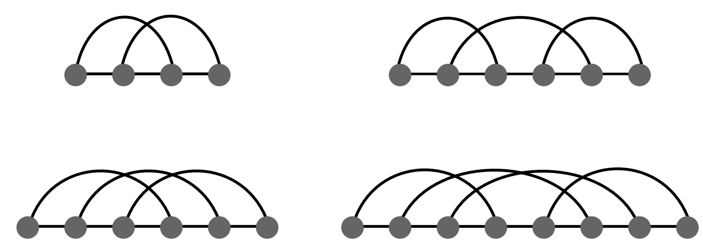


Figure: All 4 irreducible shadows of genus $g = 1$.

**Generating functions for irreducible shadows are polynomials, i.e. finite:

$$m = 2g + r - 1$$

$$2m \geq 3(r-1) + 2$$

Analytic Combinatorics

Flajolet, P and Sedgewick, R. *Analytic Combinatorics*, 2009.

Begin: Generalize RNA structures to graph diagrams

✓ Combinatorial Classes

↓

Generating Functions

↓

Coefficient Asymptotics

Symbolic enumeration – deriving generating functions without recurrence relations

- Find a functional form of generating function
- Find asymptotic expansion of function at its dominant singularity
- Extract coefficient asymptotics using analytic transfer theorems

Results: Expectation and variance of the longest irreducible component

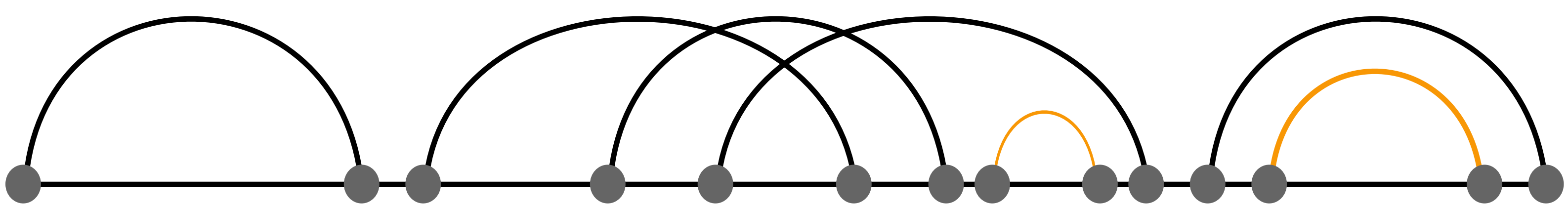
Enumerate γ-structures

$\mathcal{G}_\tau$: Combinatorial class of γ-structures

$$\mathcal{G}_{\tau,n} = \mathcal{G}_\tau \Big|_{|G|=n}$$

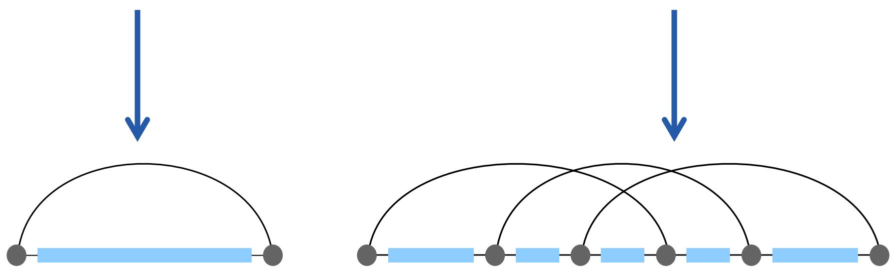
$$G_\tau(z) = \sum_{G \in \mathcal{G}_\tau} z^{|G|} = \sum_{n \geq 0} g_\tau(n) z^n$$

(via symbolic enumeration)
$$\begin{cases} A = B \cup C & \Rightarrow \quad A(x) = B(x) + C(x) \qquad B \cap C = \varnothing \\ {}^{*}A = B \times C & \Rightarrow \quad A(x) = B(x) \cdot C(x) \\ A = SEQ(B) & \Rightarrow \quad A(x) = (1 - B(x))^{-1} \end{cases}$$



Generating Functions

γ-matchings: $$H(u) = \left(1 - uH(u) - H(u)^{-1} \sum_{g \le \gamma} \sum_{m=2g}^{6g-2} i_g(m) \left(\frac{uH(u)^2}{1 - uH(u)^2} \right)^m \right)^{-1}$$



γ-structures: $$G_\tau(z) = \frac{1}{1-z+u_\tau(z)z^2} H\left(\frac{u_\tau(z)z^2}{(1-z+u_\tau(z)z^2)^2} \right)$$

Why are we concerned with generating functions?

$(\mathcal{G}_{\tau,n}, P)$: probability space with $P(G) = \frac{1}{|\mathcal{G}_{\tau,n}|}$.

$$B_n : \mathcal{G}_{\tau,n} \to \mathbb{Z}^+$$

$$G \mapsto \text{Length of the longest block}$$

$$P(B_n = n-k) = \frac{\text{Count of structures with longest block length } n-k}{\text{Count of all possible structures}}$$

$$E[B_n] = \sum_{k=1}^{n} (n-k) P(B_n = n-k)$$

Why are we concerned with generating functions?

$(\mathcal{G}_{\tau,n}, P)$: probability space with $P(G) = \frac{1}{|\mathcal{G}_{\tau,n}|}$.

$$B_n : \mathcal{G}_{\tau,n} \to \mathbb{Z}^+$$

$$G \mapsto \text{Length of the longest block}$$

$$P(B_n = n-k) = \frac{\text{Count of structures with longest block length } n-k}{\text{Count of all possible structures}}$$

$$= \frac{[z^n](G_{\leq n-k}(z) - G_{\leq n-k-1}(z))}{[z^n]G_\tau(z)}$$

Structures with blocks of length at most $n-k$ and $n-k-1$.

Analytic Combinatorics

Flajolet, P and Sedgewick, R. *Analytic Combinatorics*, 2009.

Begin: Generalize RNA structures to graph diagrams

Combinatorial Classes

↓

Generating Functions

↓

Coefficient Asymptotics

Symbolic enumeration – deriving generating functions without recurrence relations

- Find a functional form of generating function
- Find asymptotic expansion of function at its dominant singularity
- Extract coefficient asymptotics using analytic transfer theorems

Results: Expectation and variance of the longest irreducible component

Complex Analysis – Functional Forms

An implicit functional form of $H(u)$ for $\gamma = 1$:

$$P(u,X) = -1 + X + 3X^2u - 4X^3u - 3X^4u^2 + 6X^5u^2 - 2X^6u^3 - 4X^7u^3 + 3X^8u^4 + X^9u^4 - X^{10}u^5$$

$$P(u,H(u)) = 0$$

For arbitrary γ: $P(u,X) = (1-uX^2)^{6\gamma-2}(-1+X-uX^2) - (1-uX^2)^{6\gamma-2}\sum_{g\le\gamma}\sum_{m=2g}^{6g-2} i_g(m)\left(\frac{uX^2}{1-uX^2}\right)^m$

$$P(u,H(u)) = 0$$

Next: Consider function as analytic object to perform singularity analysis.

Analytic Transfers

Theorem (Flajolet and Sedgewick, 2009): *For α an arbitrary complex number other than a negative integer,*

$$[z^n](1-z)^{-\alpha} = n^{\alpha-1}\Gamma(1/2)^{-1}(1+O(n^{-1})), \quad n \to \infty$$

Method:

1. Identify dominant singularity of $G_\tau(z): z=\mu$.
2. Determine singular expansion:

$$\boxed{G_\tau(z) = \theta_0 + \theta_1(\mu - z)^{\frac{1}{2}} + \theta_2(\mu - z) + O((\mu - z)^{\frac{3}{2}})}$$

3. Apply transfer theorem:

$$\boxed{[z^n]G_\tau(z) = \frac{\theta_1\sqrt{\mu}}{\Gamma(-1/2)} n^{-\frac{3}{2}} \mu^{-n}(1+O(n^{-1}))}$$

Expected Length of Blocks

Theorem 1: *The expectation and variance of the length of the longest block in γ-structures for γ = 1 is given by*

$$E[B_n] = n - \alpha n^{\frac{1}{2}}(1+o(1)), \quad \alpha = 1.416$$

$$V[B_n] = \beta n^{\frac{3}{2}}(1+o(1)).$$

Proof: $E[B_n] = \sum_{k=1}^{n}(n-k)P(B_n = n-k)$

$$E[B_n] = n - \underbrace{\sum_{k=1}^{n^{\frac{1}{8}}} kP(B_n = n-k)}_{o(n^{\frac{1}{2}})} - \underbrace{\sum_{k=n^{\frac{1}{8}}}^{\frac{n}{2}-1} kP(B_n = n-k)}_{\alpha n^{\frac{1}{2}}(1+o(1))} - \underbrace{\sum_{k=\frac{n}{2}-1}^{n} kP(B_n = n-k)}_{o(n^{\frac{1}{2}})}$$

Approximation method: bound P by 1 — Riemann Sum approx. — Euler-Maclaurin Sum approx.

Expected Length of Blocks, Cont.

1. $$\sum_{k=1}^{n^{\frac{1}{8}}} kP(B_n = n-k) \le \sum_{k=1}^{n^{\frac{1}{8}}} k = \frac{n^{\frac{1}{8}}(n^{\frac{1}{8}}+1)}{2} = o(n^{\frac{1}{2}})$$

2. $$\sum_{k=n^{\frac{1}{8}}}^{\frac{n}{2}-1} kP(B_n = n-k) = \tfrac{2c}{\theta_0} n^{\frac{1}{2}} \sum_k \tfrac{1}{n}\left(1-\tfrac{k}{n}\right)^{-\frac{3}{2}}\left(\tfrac{k}{n}\right)^{-\frac{1}{2}} (1+O(k^{-1}))(1+O(n^{-1})) = \tfrac{4c}{\theta_0} n^{\frac{1}{2}}(1+o(1))$$

$$\sum_k \tfrac{1}{n}\left(1-\tfrac{k}{n}\right)^{-\frac{3}{2}}\left(\tfrac{k}{n}\right)^{-\frac{1}{2}} = \int_0^{1/2} (1-x)^{-\frac{3}{2}} x^{-\frac{1}{2}} dx (1+o(1)) = 2(1+o(1))$$

(Riemann Sum Formula)

3. $$\sum_{k=\frac{n}{2}-1}^{n} kP(B_n = n-k) \le n(1-\sum_{k=1}^{n^{\frac{2}{5}}} P(B_n = n-k) - \sum_{k=n^{\frac{2}{5}}}^{\frac{n}{2}-1} P(B_n = n-k)) = o(n^{\frac{1}{2}})$$

Expected Length of Blocks, Cont.

$$\text{3a.}\quad \sum_{k=1}^{n^{\frac{2}{5}}} P(B_n = n-k) = \sum_k \frac{b_k \mu^k}{\theta_0^2}(1+o(n^{-\frac{1}{2}})) = (1 - \tfrac{2c}{\theta_0}\sum_{k\geq n^{\frac{2}{5}}} k^{-\frac{3}{2}} + O(\sum_{k\geq n^{\frac{2}{5}}} k^{-\frac{5}{2}}))(1+o(n^{-\frac{1}{2}}))$$

$$\sum_{k\geq n^{\frac{2}{5}}} k^{-\frac{3}{2}} = \zeta(\tfrac{3}{2}, n^{\frac{2}{5}}) = 2n^{-\frac{1}{5}}(1+O(n^{-\frac{2}{5}}))$$

(Hurwitz-Zeta Function)

$$\sum_{k=1}^{n^{\frac{2}{5}}} P(B_n = n-k) = 1 - \tfrac{4c}{\theta_0} n^{-\frac{1}{5}} + o(n^{-\frac{1}{2}})$$

$$\text{3b.}\quad \sum_{k=n^{\frac{2}{5}}}^{\frac{n}{2}-1} kP(B_n = n-k) = \frac{2c}{\theta_0}\sum_k \left(1-\tfrac{k}{n}\right)^{-\frac{3}{2}} k^{-\frac{3}{2}}(1+O(k^{-1}))(1+O(n^{-1})) = \tfrac{4c}{\theta_0} n^{-\frac{1}{5}} + o(n^{-\frac{1}{2}})$$

$$\sum_k \left(1-\tfrac{k}{n}\right)^{-\frac{3}{2}} k^{-\frac{3}{2}} = \int_{n^{\frac{2}{5}}}^{\frac{n}{2}} \left(x - \tfrac{x^2}{n}\right)^{-\frac{3}{2}} dx + \tfrac{1}{2}\left(x - \tfrac{x^2}{n}\right)^{-\frac{3}{2}}\Bigg|_{n^{\frac{2}{5}}}^{\frac{n}{2}} + \ldots$$

(Euler-Maclaurin Sum Formula)

Varying γ

Theorem: *The expectation and variance of the length of the longest block in γ-structures is given by*

$$E[B_n] = n - \alpha n^{\frac{1}{2}}(1+o(1)),$$

$$V[B_n] = \beta n^{\frac{3}{2}}(1+o(1)).$$

γ	0	1	2	3
α_γ	2.482	1.416	0.964	0.734
β_γ	0.533	0.304	0.207	0.159

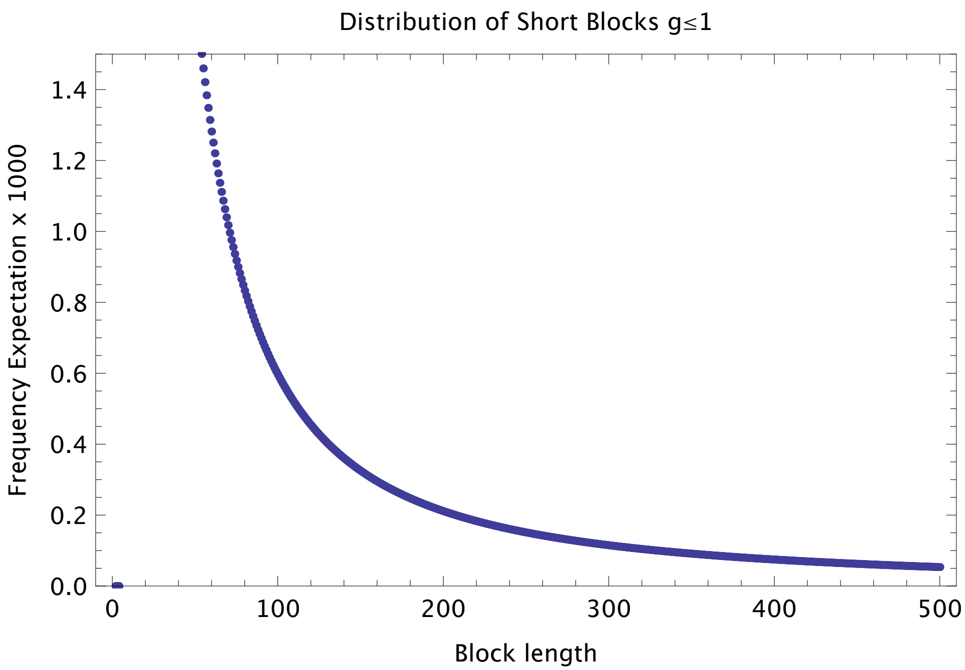
Secondary structures

Short Blocks

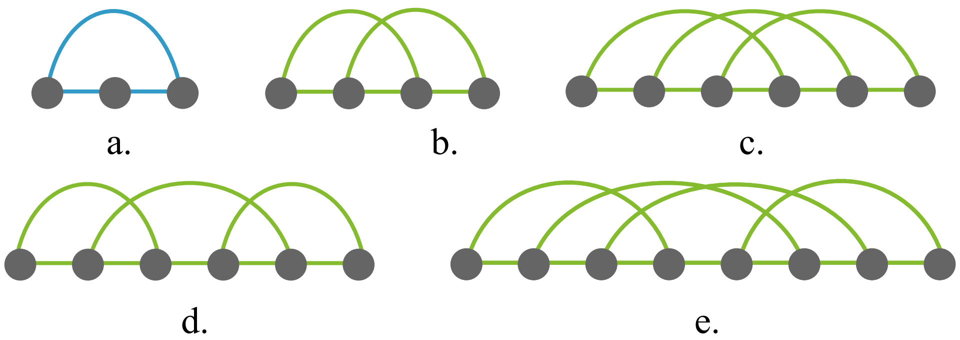
Theorem 2: *For fixed k, the distribution of the number of blocks of length k is negative binomial, $NB(2,t)$,* $\quad t = \frac{[z^k]T(z)\mu^k}{1-T(\mu)+[z^k]T(z)\mu^k}$.

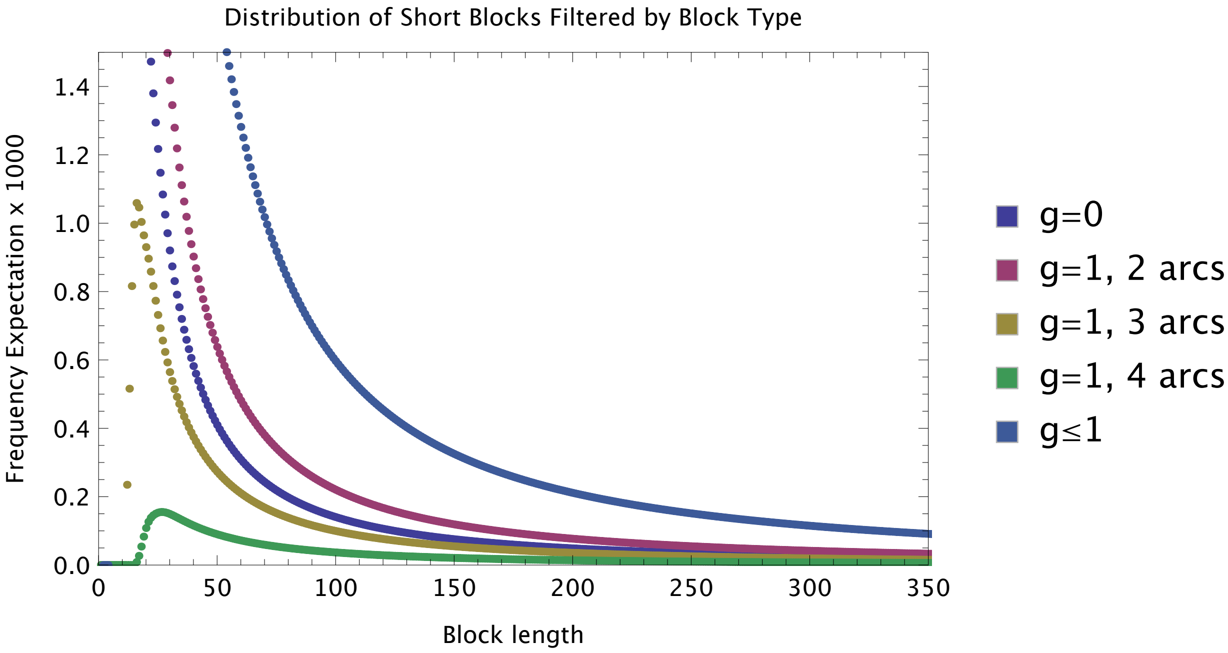
$$E[W_{k,n}] = \frac{2t}{1-t},$$

$$V[W_{k,n}] = \frac{2t}{(1-t)^2}.$$



Short Blocks





Wrap Up/Future Work

- Described the spectrum of blocks for γ-structures
  - Length of the longest block
  - Distribution of number of blocks of a finite size
- Develop framework for higher dimensional rainbows

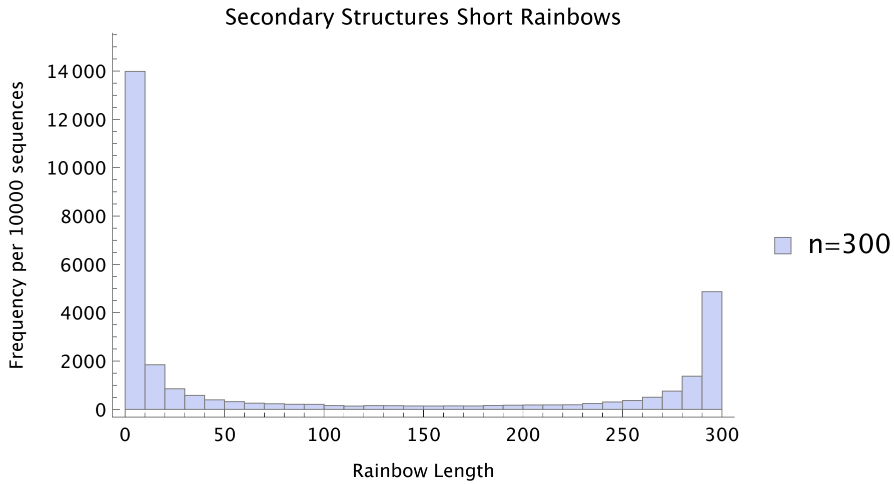
Length Distribution

$$E[B_n] = n - \alpha n^{\frac{1}{2}}(1+o(1)),$$
$$V[B_n] = \beta n^{\frac{3}{2}}(1+o(1)).$$

Corollary 3: For all $\varepsilon > 0$, there exists a positive integer $t(\varepsilon)$ such that

$$\lim_{n\to\infty} P(B_n \geq n - t(\varepsilon)) \geq 1 - \varepsilon$$

$$\lim_{n\to\infty} P(B_n \geq n - 100) = 0.688$$
$$\lim_{n\to\infty} P(B_n \geq n - 500) = 0.752$$



Length Distribution

$$E[B_n] = n - \alpha n^{\frac{1}{2}}(1 + o(1)),$$

$$V[B_n] = \beta n^{\frac{3}{2}}(1 + o(1)).$$

Theorem 3: For any $k = o(n)$, $\lim_{n\to\infty} P(n - B_n = k) = \frac{[z^k]G_\tau^2(z)\mu^k}{G_\tau^2(\mu)}$.

Let $k = n^{\frac{1}{2}}$. $\lim_{n\to\infty} P(n - B_n = n^{\frac{1}{2}}) = 0$