Combinatorics of Biomolecules

C.M. Reidys

Nankai University

Center for Combinatorics, LMPC

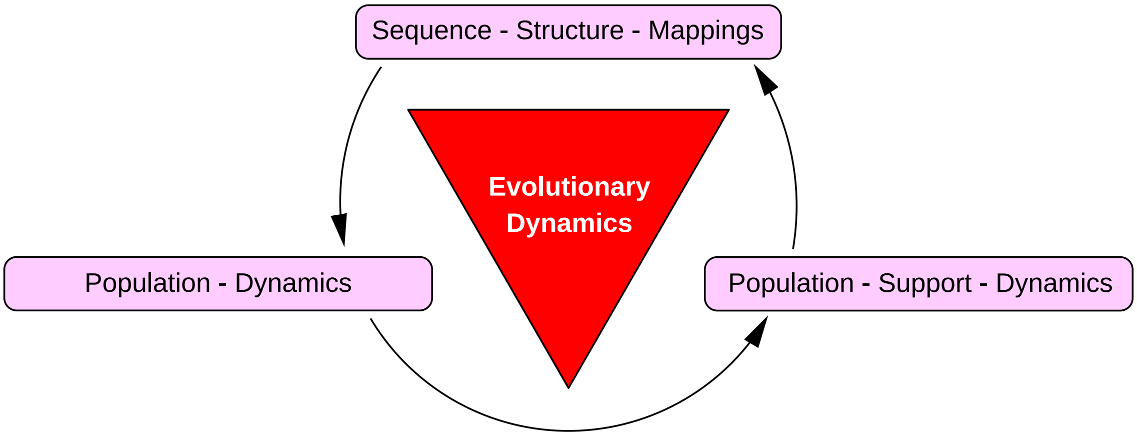
Sequence - Structure - Mappings
Evolutionary Dynamics
Population - Dynamics
Population - Support - Dynamics

FIGURE 1. Evolutionary Dynamics

Computational Biology Group at Nankai

- sequence to structure maps
- combinatorial representation of biomolecules
- new generation folding algorithms of biomolecules

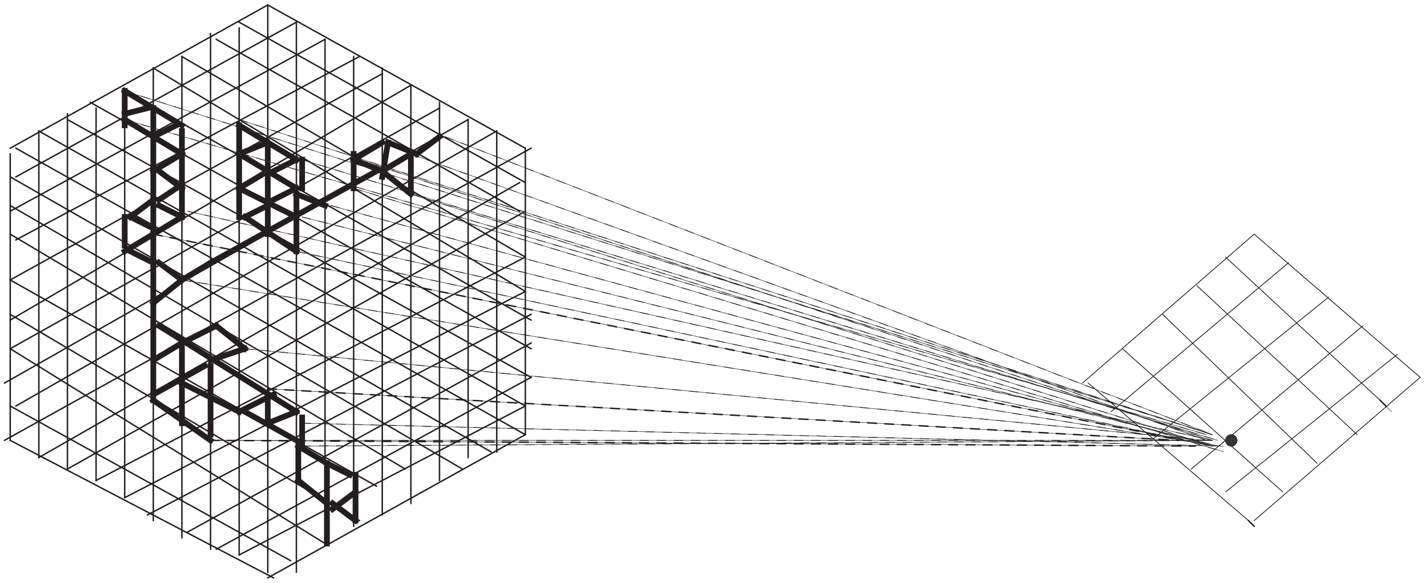

Figure 2. The neutral network of a structure. Sequence space (right) and shape space (left) represented as lattices. We draw the edges between two sequences bold if they map into the one particular structure on the left. The two key properties of neutral nets are their connectivity and percolation. They allow sequences to move while maintaining a shape through sequence space.

Sequences and Shapes: Neutral Networks

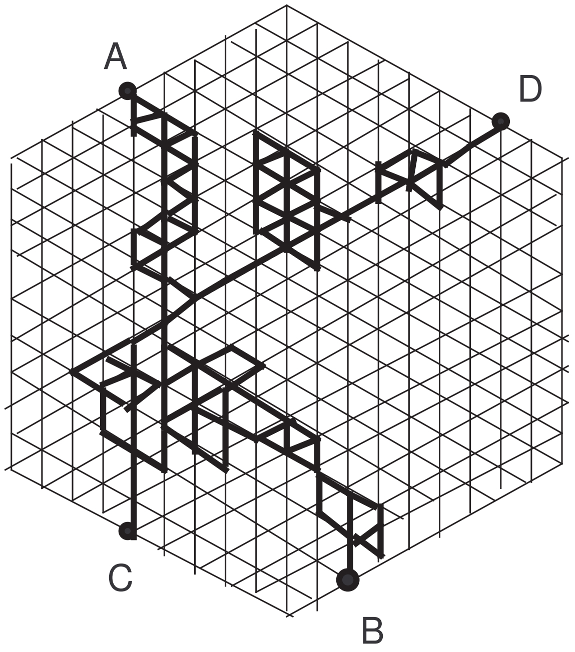

FIGURE 3. Neutral network. Sequence space is represented as lattice and the neutral net is an induced subgraph (bold edges). We label the pairs of sequences representing antipodal pairs by (A,B) and (C,D). The two key properties of neutral nets are their connectivity and percolation.

Theorem 1. *Let Q_{2,λ_n}^n be the random graph consisting of Q_2^n-subgraphs, Γ_n, induced by selecting each Q_2^n-vertex with independent probability $\lambda_n = \frac{1+\chi_n}{n}$, where $\chi_n = \epsilon n^{\frac{a-1}{2}}$, where $0 < \epsilon$ and $0 < a \le 1$. Then we have*

$$\exists\, \kappa_a > 0; \quad \lim_{n\to\infty} \mathbb{P}(\,|C_n^{(1)}| \ge \kappa_a\, n^{a-1}|\Gamma_n| \text{ and } C_n^{(1)} \text{ is unique}\,) = 1\,. \tag{0.1}$$

Christian M. Reidys *Large components in random induced subgraphs of n-cubes* Discrete Math. submitted, 2007.

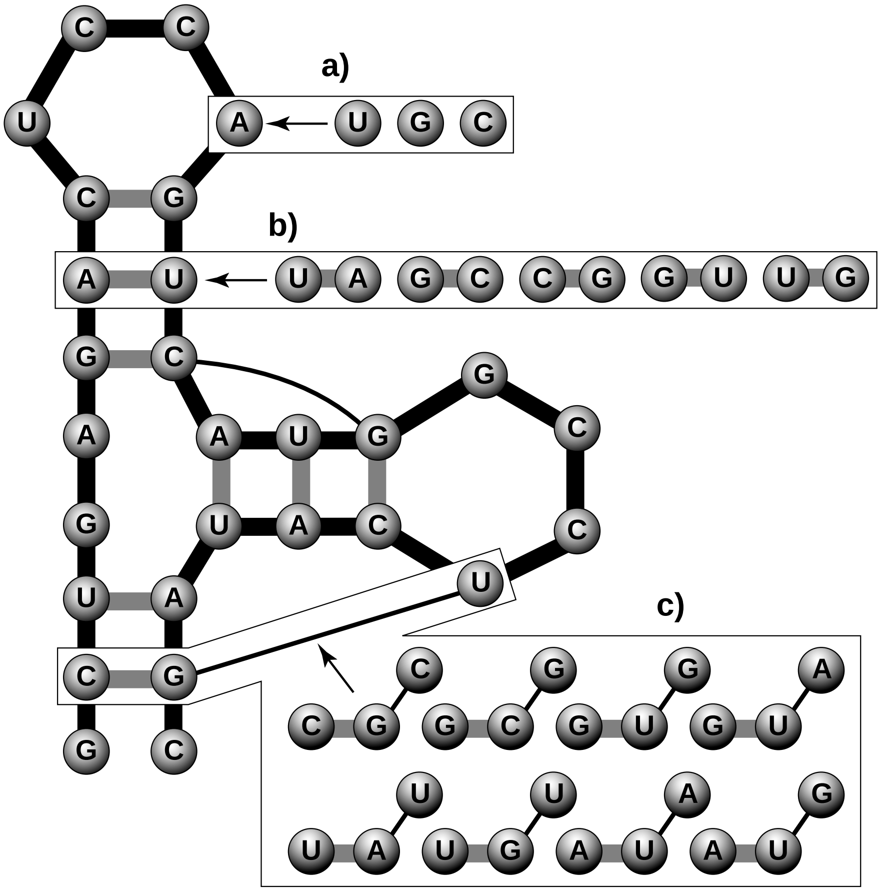

FIGURE 4. RNA secondary structure. Watson-Crick base-pairs (gray), tertiary contacts (black)

RNA secondary structures or better: 2-noncrossing RNA

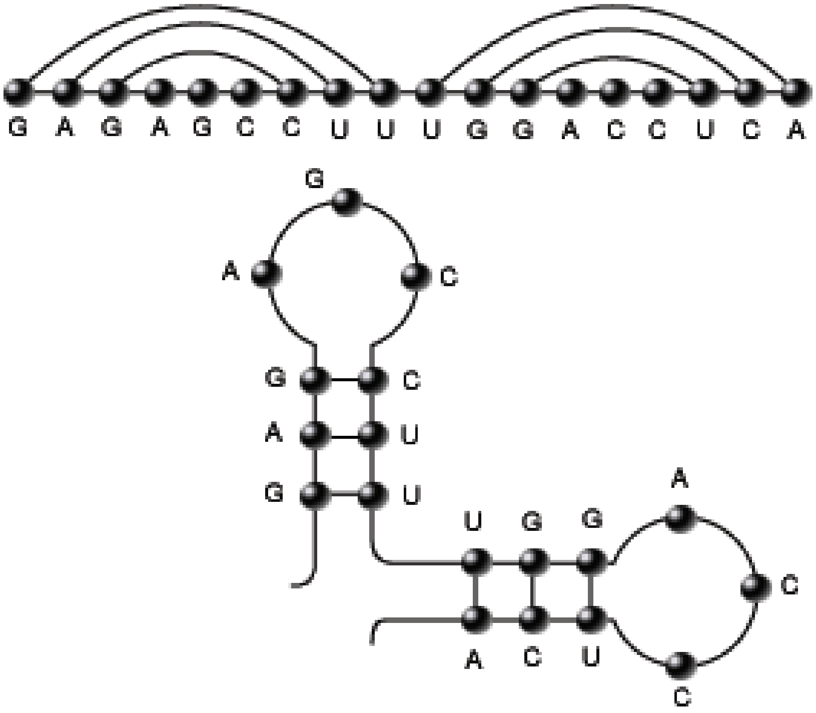

FIGURE 5. RNA secondary structures. Diagram representation (top): the primary sequence, **GAGAGCCUUUGGACCUCA**, is drawn horizontally and its backbone bonds are ignored. All bonds are drawn in the upper halfplane and secondary structures have the property that no two arcs intersect and all arcs have minimum length 2. Outer planar graph representation (bottom).

3-noncrossing RNA structures

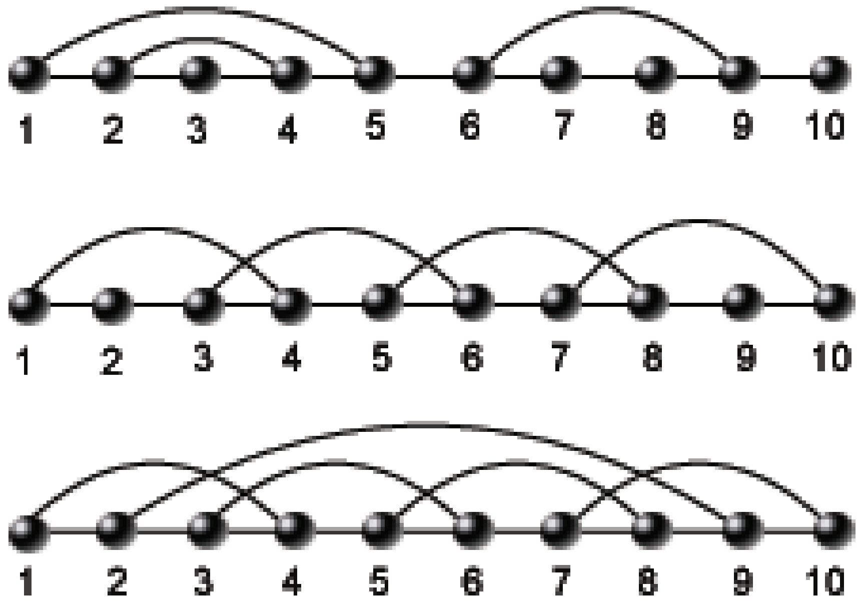

FIGURE 6. k-noncrossing RNA structures. (a) **secondary structure**, (b) **planar 3-noncrossing RNA structure**, (c) the smallest **non-planar 3-noncrossing structure**

Definition 1. An RNA structure (of pseudoknot type $k-2$), $S_{k,n}$, is a digraph in which all vertices have degree ≤ 1, that does not contain a k-set of mutually intersecting arcs and 1-arcs, i.e. arcs of the form $(i, i+1)$, respectively.

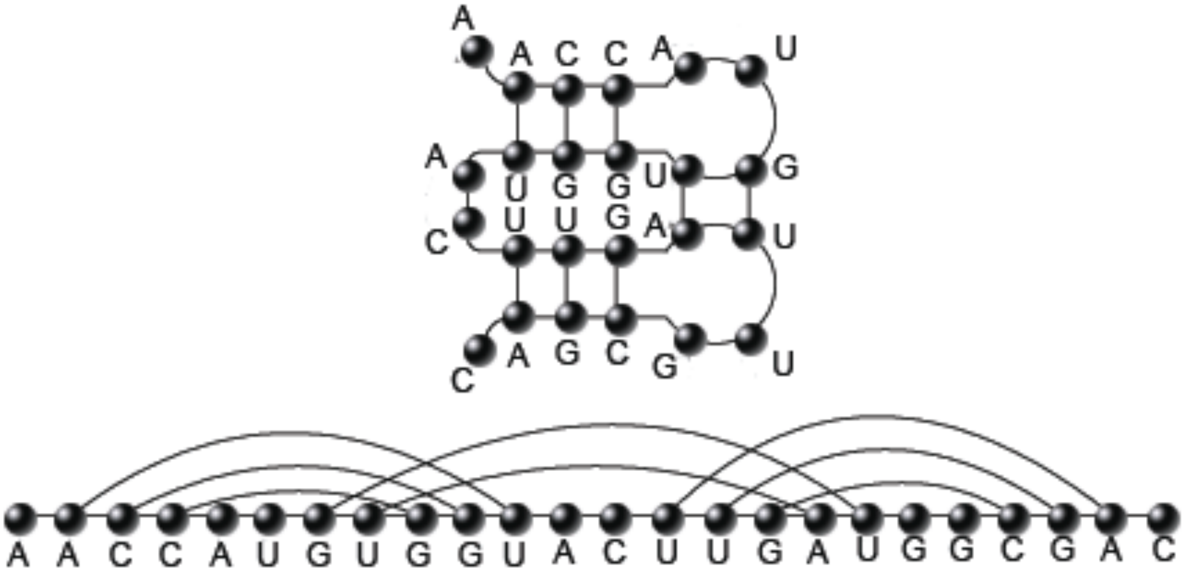

FIGURE 7. A 3-noncrossing RNA structure, as a planar graph (top) and as a diagram (bottom)

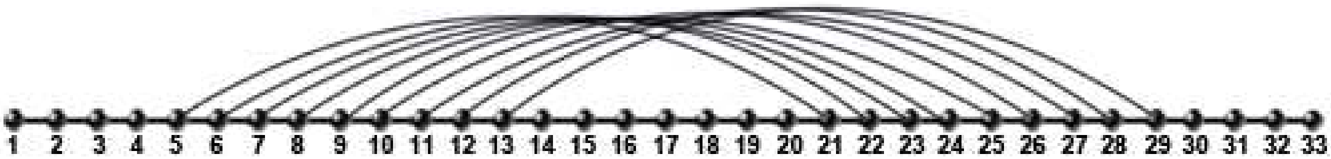

FIGURE 8. The proposed SRV-1 frame-shift is a 10-noncrossing RNA structure motif.

Combinatorics of 3-noncrossing RNA structures

Theorem 2. *Let $k \in \mathbb{N}$, $k \geq 2$, let $f_k(n,\ell)$ be the number of k-noncrossing digraphs over n vertices with exactly ℓ isolated vertices. Then the number of RNA structures with ℓ isolated vertices, $\mathsf{S}_k(n,\ell)$, is*

$$\mathsf{S}_k(n,\ell) = \sum_{b=0}^{(n-\ell)/2} (-1)^b \binom{n-b}{b} f_k(n-2b,\ell) \ . \tag{0.2}$$

Furthermore the number of k-noncrossing RNA structures, $\mathsf{S}_k(n)$ is given by

$$\mathsf{S}_k(n) = \sum_{b=0}^{\lfloor n/2 \rfloor} (-1)^b \binom{n-b}{b} \left\{ \sum_{\ell=0}^{n-2b} f_k(n-2b,\ell) \right\} \tag{0.3}$$

Emma Y. Jin, Jing Qin and Christian M. Reidys *Combinatorics of RNA Structures with Pseudoknots*, Bulletin of Math. Bio., 2007, in press.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$S_3(n)$	1	1	2	5	13	36	105	321	1018	3334	11216	38635	135835	486337	1769500

Table 1. The first 15 numbers of 3-noncrossing RNA structures.

Combinatorics of 3-noncrossing RNA structures: Main idea

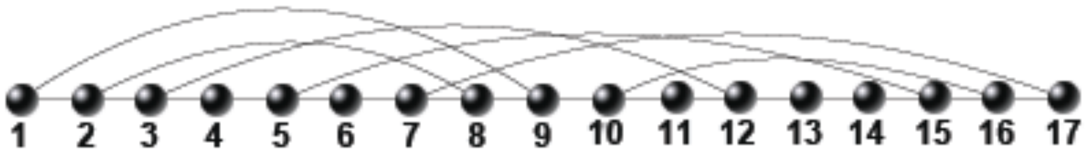

FIGURE 9. A 5-noncrossing structure corresponding to the oscillating tableau below and subsequently the corresponding walk $\gamma_{a,a}$ in $\mathbb{Z}^4$.

$$\varnothing \quad \begin{array}{|c|}\hline 1\\\hline\end{array} \quad \begin{array}{|c|c|}\hline 1&2\\\hline\end{array} \quad \begin{array}{cc} 1&2\\ 3& \end{array} \quad \begin{array}{cc} 1&2\\ 3& \end{array} \quad \begin{array}{cc} 1&2\\ 3& \\ 5& \end{array} \quad \begin{array}{cc} 1&2\\ 3& \\ 5& \end{array} \quad \begin{array}{cc} 1&2\\ 3& \\ 5& \\ 7& \end{array} \quad \begin{array}{c} 1\\ 3\\ 5\\ 7 \end{array}$$

$$\begin{array}{c} 3\\ 5\\ 7 \end{array} \quad \begin{array}{cc} 3&10\\ 5& \\ 7& \end{array} \quad \begin{array}{cc} 3&10\\ 5& \\ 7& \end{array} \quad \begin{array}{cc} 5&10\\ 7& \end{array} \quad \begin{array}{cc} 5&10\\ 7& \end{array} \quad \begin{array}{cc} 5&10\\ 7& \end{array} \quad \begin{array}{|c|c|}\hline 7&10\\\hline\end{array} \quad \begin{array}{|c|}\hline 7\\\hline\end{array} \quad \varnothing$$

Why 3-noncrossing RNA structures is so different: recursions

Corollary 1. *The number of RNA secondary structures having exactly ℓ isolated vertices, $\mathsf{S}_2(n,\ell)$, is given by*

(0.4)
$$\mathsf{S}_2(n,\ell) = \frac{2}{n-\ell}\binom{\frac{n+\ell}{2}}{\frac{n-\ell}{2}+1}\binom{\frac{n+\ell}{2}-1}{\frac{n-\ell}{2}-1} .$$

Furthermore $\mathsf{S}_2(n,\ell)$ satisfies the recursion

(0.5)
$$(n-\ell)(n-\ell+2)\cdot \mathsf{S}_2(n,\ell) \;-\; (n+\ell)(n+\ell-2)\cdot \mathsf{S}_2(n-2,\ell) = 0 \;.$$

Corollary 2. *The number of 3-noncrossing RNA structures having exactly ℓ isolated vertices, $\mathsf{S}_3(n,\ell)$, satisfies the 4-term recursion*

(0.6)
$$\mathsf{p}_1(n,\ell)\,\mathsf{S}_3(n-6,\ell) - \mathsf{p}_2(n,\ell)\,\mathsf{S}_3(n-4,\ell) - \mathsf{p}_3(n,\ell)\mathsf{S}_3(n-2,\ell) + \mathsf{p}_4(n,\ell)\,\mathsf{S}_3(n,\ell) = 0 \;,$$

where the coefficients $\mathsf{p}_1(n,\ell)$, $\mathsf{p}_2(n,\ell)$ $\mathsf{p}_3(n,\ell)$ and $\mathsf{p}_4(n,\ell)$ are given by

$$\begin{aligned}
\mathsf{p}_1(n,\ell) &= \frac{1}{2}n(n-1)(n-10+\ell)(n-4+\ell)(n-8+\ell)\\
\mathsf{p}_2(n,\ell) &= \frac{1}{2}n(n-3)(13n^3-126n^2+13n^2\ell-88n\ell+392n+3n\ell^2+216\ell-384-42\ell^2+3\ell^3)\\
\mathsf{p}_3(n,\ell) &= (n-1)(\frac{1}{2}n-2)(13n^3-30n^2-13n^2\ell+8n+16n\ell+3n\ell^2+30\ell^2-72\ell-3\ell^3)\\
\mathsf{p}_4(n,\ell) &= (n-3)(\frac{1}{2}n-2)(n-\ell)(n-\ell+6)(n-\ell+4)\;.
\end{aligned}$$

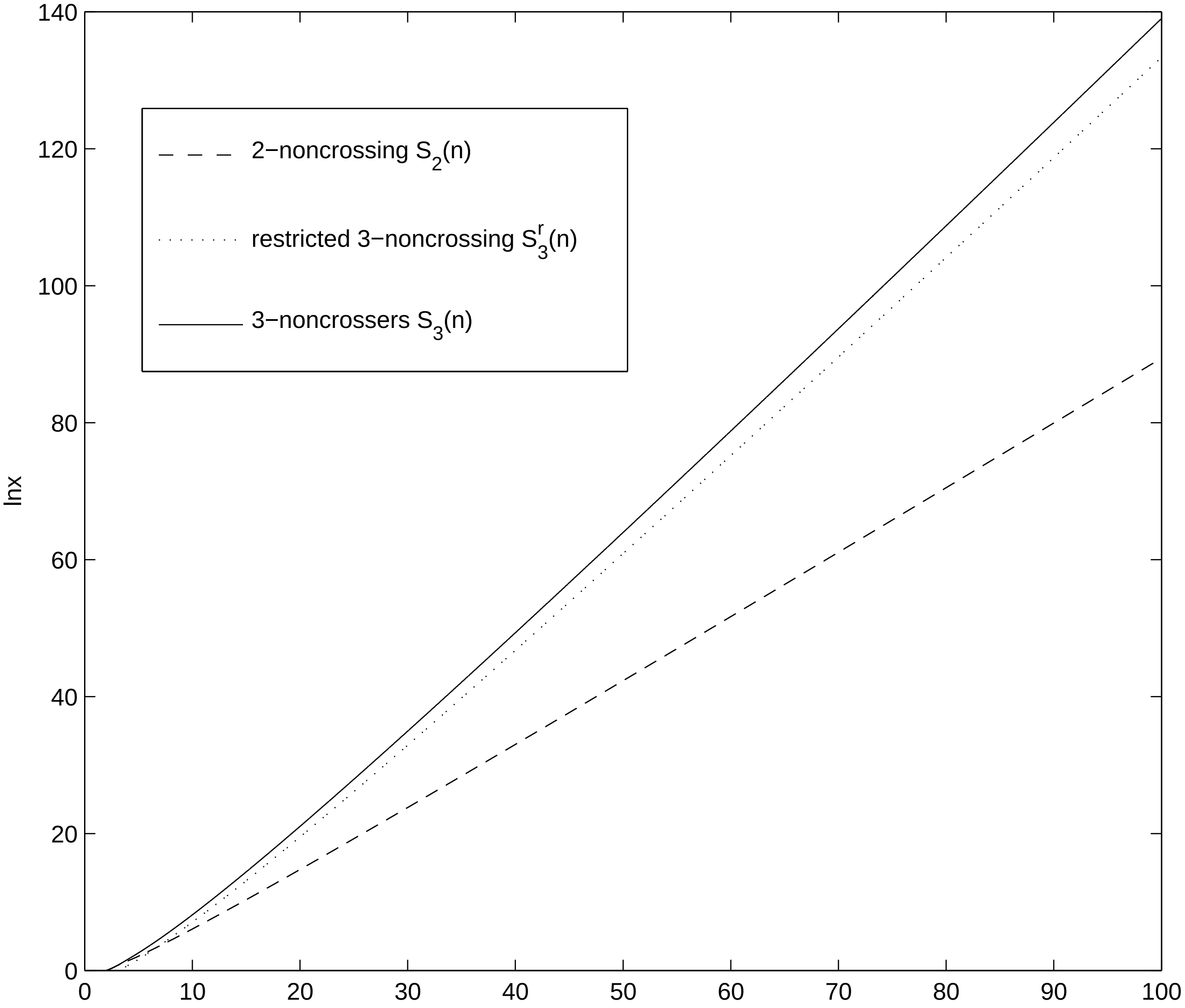

FIGURE 10. The numbers of RNA structures for large n. 2-noncrossing RNA structures, 3-noncrossing RNA structures and restricted 3-noncrossing RNA structures. Numerically exponential growth rates: $\mathsf{S}_2(n) \sim 2.5913^n$ ($n = 1000$), $\mathsf{S}_3(n) \sim 4.6542^n$ ($n = 1000$), and $\mathsf{S}_3^{(r)}(n) \sim 4.2741^n$ ($n = 400$).

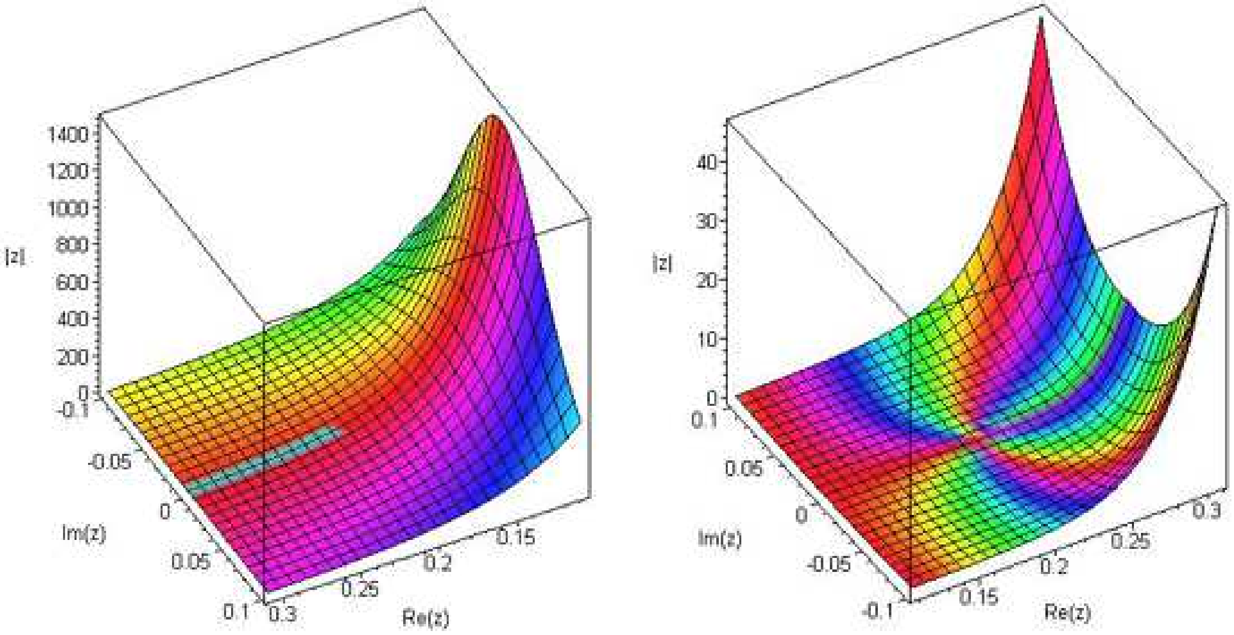

FIGURE 11. Toroidal harmonics and its singular expansion. We display the analytic continuation of $\sum_{n\geq 0} \mathsf{S}_3(n)z^n$, the generating function of 3-noncrossing RNA structures (left) and its singular expansion (right) at the dominant singularity $\rho_3 = \frac{5-\sqrt{21}}{2}$.

Asymptotic Combinatorics: Toroidal Harmonics

Lemma 1. *Let z be an indeterminant over $\mathbb{R}$ and $w \in \mathbb{R}$ a parameter. Let furthermore $\rho_k(w)$ denote the radius of convergence of the power series $\sum_{n\geq 0}[\sum_{h\leq n/2} \mathsf{S}_k(n,h)w^{2h}]z^n$. Then for $|z| < \rho_k(w)$ holds*

$$\sum_{n\geq 0}\sum_{h\leq n/2} \mathsf{S}'_k(n,h)w^{2h}z^n = \frac{1}{w^2z^2 - z + 1}\sum_{n\geq 0} f_k(2n,0)\left(\frac{wz}{w^2z^2 - z + 1}\right)^{2n} . \tag{0.7}$$

In particular we have for $w = 1$,

$$\sum_{n\geq 0} \mathsf{S}_k(n)z^n = \frac{1}{z^2 - z + 1}\sum_{n\geq 0} f_k(2n,0)\left(\frac{z}{z^2 - z + 1}\right)^{2n} . \tag{0.8}$$

Theorem 3. *The number of 3-noncrossing RNA structures is asymptotically given by*

$$\mathsf{S}_3(n) \quad \sim \quad \frac{10.4724 \cdot 4!}{n(n-1)\dots(n-4)}\left(\frac{5+\sqrt{21}}{2}\right)^n .$$

Emma Y. Jin and Christian M. Reidys *Asymptotics of RNA Structures with Pseudoknots*, Bulletin of Math. Bio., 2007, accepted.

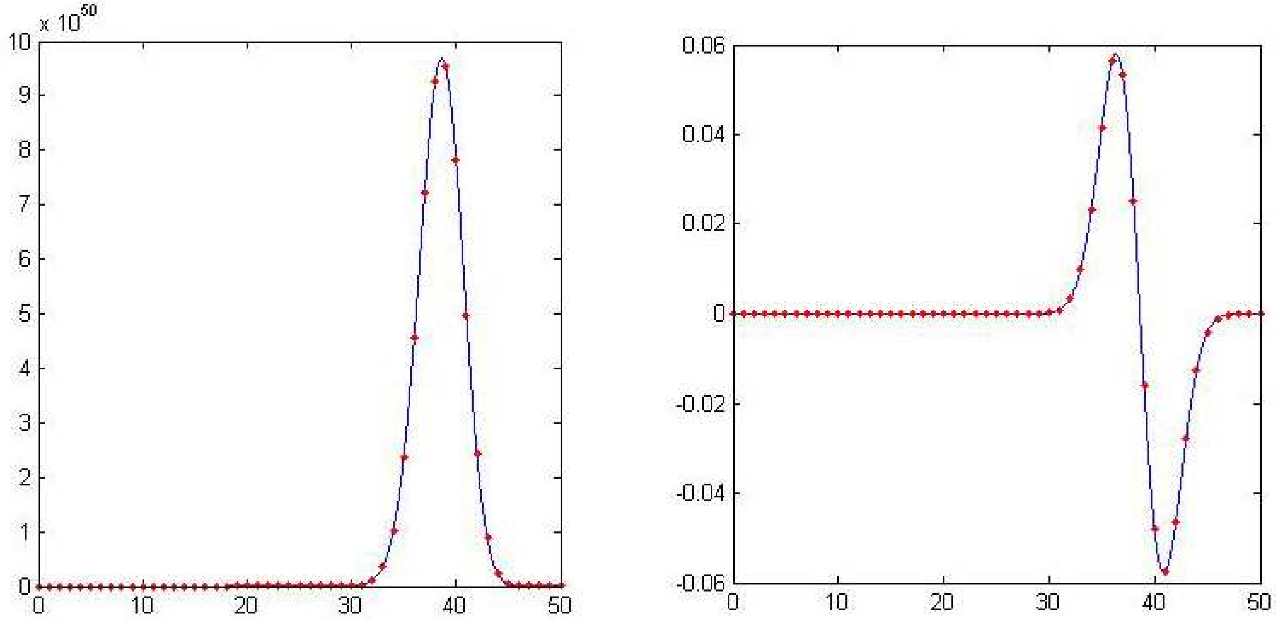

FIGURE 12. Central limit theorem and local limit theorem for 3-noncrossing RNA structures of length $n = 100$ with exactly h arcs: we display the central limit theorem (left) for $\mathsf{S}'_3(100, h), h = 1, 2, \cdots 50$ (labeled by red dots) with mean $0.39089 \cdot 100 = 39.089$ and variance $0.041565 \cdot 100 = 4.1565$, and for the local limit theorem (right), we display the difference $\sqrt{4.1565}\ \mathbb{P}\left(\frac{X_n - 39.089}{\sqrt{4.1565}} = x\right) - \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ which is maximal close to the peak of the distribution.

Central and Local Limit Theorems for RNA structures

Theorem 4. (Central Limit Theorem) *Let $\mathsf{S}_3'(n,h)$ be the number of 3-noncrossing RNA structures with exactly h arcs. Let X_n be the r.v. having the distribution*

$$\forall\, h = 0, 1, \dots \lfloor \frac{n}{2} \rfloor, \qquad \mathbb{P}(X_n = h) = \frac{\mathsf{S}_3'(n,h)}{\mathsf{S}_3(n)} \tag{0.9}$$

Then the random variable $\frac{X_n - \mu n}{\sqrt{\sigma^2 n}}$ has asymptotically normal distribution with parameter $(0,1)$, i.e.

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_n - \mu n}{\sqrt{\sigma^2 n}} < x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt \tag{0.10}$$

and μ, σ^2 are given by

$$\mu = -\frac{-\frac{3}{2} + \frac{13}{42}\sqrt{21}}{\frac{5}{2} - \frac{1}{2}\sqrt{21}} = 0.39089 \quad \text{and} \quad \sigma^2 = \mu^2 - \frac{1 - \frac{94}{441}\sqrt{21}}{\frac{5-\sqrt{21}}{2}} = 0.041565\ . \tag{0.11}$$

Theorem 5. (Local Limit Theorem) *Let $\mathsf{S}_3'(n,h)$ be the number of 3-noncrossing RNA structures with exactly h arcs. Let X_n be the r.v. having the distribution*

$$\forall\, h = 0, 1, \dots \lfloor \frac{n}{2} \rfloor, \qquad \mathbb{P}(X_n = h) = \frac{\mathsf{S}_3'(n,h)}{\mathsf{S}_3(n)} \tag{0.12}$$

Then we have for set $S = \{x \mid x = o(\sqrt{n})\}$

$$\lim_{n\to\infty} \sup_{x\in S} \left| \sqrt{\sigma^2 n}\, \mathbb{P}\left(\frac{X_n - n\,\mu}{\sqrt{\sigma^2 n}} = x \right) - \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \right| = 0\,, \tag{0.13}$$

where $\mu = 0.39089$ and $\sigma^2 = 0.041565$.

Emma Y. Jin and Christian M. Reidys *Central and Local Limit Theorems of RNA Stuctures*, Journal of theor. Bio., 2007, submitted

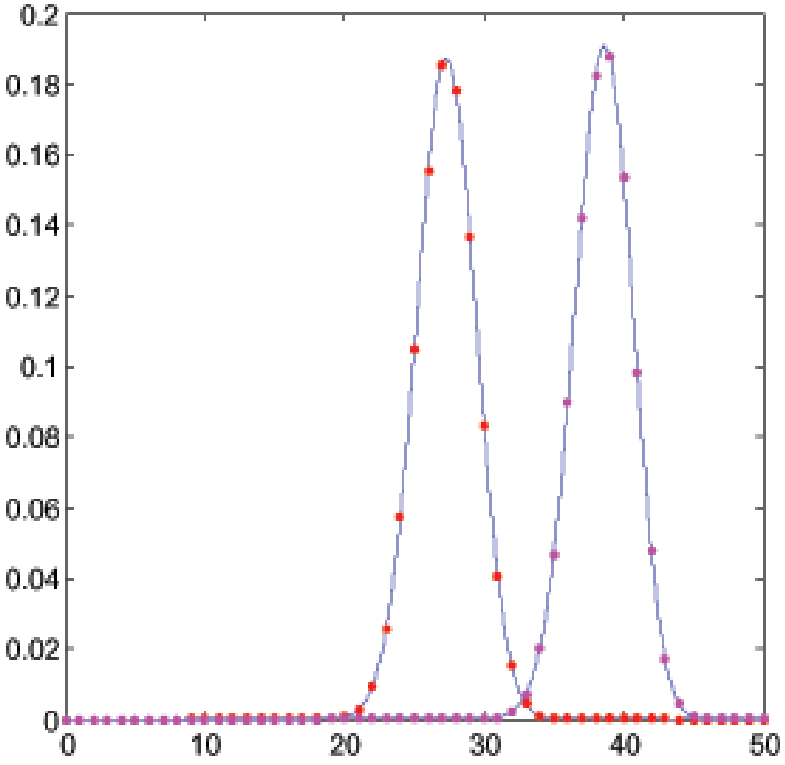

FIGURE 13. Central limit theorem of 2-noncrossing and 3-noncrossing RNA structures: both random variables are normalized to $\mathsf{S}_2'(n,h)/\mathsf{S}_2(n)$and $\mathsf{S}_3'(n,h)/\mathsf{S}_3(n)$, respectively. In case of $n = 100$, for 2-noncrossing RNA structures we have a mean of $0.276393\,n = 27.6393$ and variance $0.044721\,n = 4.4721$ (left curve), while for 3-noncrossing RNA structures mean $0.39089\,n = 39.089$ and variance $0.041565\,n = 4.1565$ (right curve). The red dots and magenta dots represent the values $\mathsf{S}_2'(n,h)/\mathsf{S}_2(n)$ and $\mathsf{S}_3'(n,h)/\mathsf{S}_3(n)$, respectively.

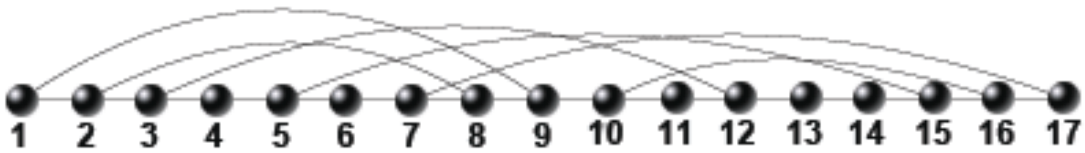

FIGURE 14. A 5-noncrossing structure corresponding to the oscillating tableau below and subsequently the corresponding walk $\gamma_{a,a}$ in $\mathbb{Z}^4$.

$$\varnothing \quad \begin{array}{|c|}\hline 1 \\ \hline\end{array} \quad \begin{array}{|c|c|}\hline 1 & 2 \\ \hline\end{array} \quad \begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1} 5 \\ \cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1} 5 \\ \cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1} 5 \\ \cline{1-1} 7 \\ \cline{1-1}\end{array} \quad \begin{array}{|c|}\hline 1 \\ \hline 3 \\ \hline 5 \\ \hline 7 \\ \hline\end{array}$$

$$\begin{array}{|c|}\hline 3 \\ \hline 5 \\ \hline 7 \\ \hline\end{array} \quad \begin{array}{|c|c|}\hline 3 & 10 \\ \hline 5 \\ \cline{1-1} 7 \\ \cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 3 & 10 \\ \hline 5 \\ \cline{1-1} 7 \\ \cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 5 & 10 \\ \hline 7 \\ \cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 5 & 10 \\ \hline 7 \\ \cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 5 & 10 \\ \hline 7 \\ \cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 7 & 10 \\ \hline\end{array} \quad \begin{array}{|c|}\hline 7 \\ \hline\end{array} \quad \varnothing$$