Counting RNA Secondary Structures of Arbitrary Pseudoknot Type


Berton A. Earnshaw

Department of Mathematics, University of Utah
Salt Lake City, Utah 84112

February 26, 2008
A single strand of RNA

- Primary structure: sequence of bases (A,G,U,C)
- Secondary structure: pairing of bases
  - Watson-Crick pairs: A-U, G-C (less often U-G)
- Tertiary structure: resulting 3D molecule
  - Different tertiary structures $\Rightarrow$ different enzymatic properties
A single strand of RNA: An example

- Primary structure:
  
  AACCAUGUGGUACUUGAUGGCGAC
A single strand of RNA: An example

- **Primary structure:**
  
  AACCAUGUGGUACUUGAUGGCGAC

- **Secondary structure:**

![Secondary structure diagram]
A single strand of RNA: An example

- **Primary structure:**
  
  \[
  \text{AACCAUGUGGUACUUGAUGGCGAC}
  \]

- **Secondary structure:**

- **Tertiary structure:** extremely difficult to predict (probably NP-hard)
RNA secondary structure as \( k \)-noncrossing arch diagram

- \( k \)-noncrossing arch diagram of order \( n \)
  - graph on vertex set \( \{1, \ldots, n\} \)
  - all vertices have degree \( \leq 1 \)
  - there do not exist \( k \) arches \( \{i_1, j_1\}, \ldots, \{i_k, j_k\} \) such that
    \[ i_1 < \cdots < i_k < j_1 < \cdots < j_k \]
RNA secondary structure as \textit{k}-noncrossing arch diagram

- \textit{k}-noncrossing arch diagram of order \( n \)
  - graph on vertex set \( \{1, \ldots, n\} \)
  - all vertices have degree \( \leq 1 \)
  - there do not exist \( k \) arches \( \{i_1, j_1\}, \ldots, \{i_k, j_k\} \) such that

\[
i_1 < \cdots < i_k < j_1 < \cdots < j_k
\]
RNA secondary structure as $k$-noncrossing arch diagram

- $k$-noncrossing arch diagram of order $n$
  - graph on vertex set $\{1, \ldots, n\}$
  - all vertices have degree $\leq 1$
  - there do not exist $k$ arches $\{i_1, j_1\}, \ldots, \{i_k, j_k\}$ such that
    \[ i_1 < \cdots < i_k < j_1 < \cdots < j_k \]

- RNA secondary structure of $n$ bases, pseudoknot type $k - 2$
  - $k$-noncrossing (but not $k - 1$) arch diagram of order $n$
  - no 1-arches $\{i, i + 1\}$
  - “abstract” secondary structure (no primary structure)
Counting RNA secondary structures

- Establish bijection between \( k \)-noncrossing arch diagrams and certain walks in \( \mathbb{Z}^{k-1} \)
- Count walks via reflection principle (Weyl groups)
- Enumerate restricted walks (RNA secondary structures)
$k$-noncrossing arch diagrams and walks in Weyl chamber

- Walk in $\mathbb{Z}^m$ of length $n$
  - sequence of vectors $x_0, x_1, \ldots, x_n \in \mathbb{Z}^m$ s.t. $|x_{i+1} - x_i| = 0$ or $1$
$k$-noncrossing arch diagrams and walks in Weyl chamber

- Walk in $\mathbb{Z}^m$ of length $n$
  - sequence of vectors $x_0, x_1, \ldots, x_n \in \mathbb{Z}^m$ s.t. $|x_{i+1} - x_i| = 0$ or $1$
- Weyl chamber
  - subset of vectors $x = (x_1, \ldots, x_m) \in \mathbb{Z}^m$ s.t. $x_1 > \cdots > x_m > 0$
**k-noncrossing arch diagrams and walks in Weyl chamber**

- Walk in $\mathbb{Z}^m$ of length $n$
  - sequence of vectors $x_0, x_1, \ldots, x_n \in \mathbb{Z}^m$ s.t. $|x_{i+1} - x_i| = 0$ or $1$
- Weyl chamber
  - subset of vectors $x = (x_1, \ldots, x_m) \in \mathbb{Z}^m$ s.t. $x_1 > \cdots > x_m > 0$


There exists a bijection between $k$-noncrossing arch diagrams of order $n$ and walks of length $n$ in $\mathbb{Z}^{k-1}$ which start and end at $a = (k - 1, k - 2, \ldots, 1)$ and remain in the Weyl chamber.
Idea of proof: Oscillating Young diagrams, RSK algorithm

- Young diagram
  - collection of squares $\mu$ arranged in left-justified rows
  - number of squares in each row weakly decreasing
Idea of proof: Oscillating Young diagrams, RSK algorithm

- Young diagram
  - collection of squares $\mu$ arranged in left-justified rows
  - number of squares in each row weakly decreasing

- Oscillating Young diagrams
  - sequence of Young diagrams $\emptyset = \mu_0, \mu_1, \ldots, \mu_n = \emptyset$
  - $\mu_i$ and $\mu_{i+1}$ differ by at most one square

\[ \emptyset \quad \square \quad \square \quad \square \quad \square \quad \square \quad \square \quad \square \quad \square \quad \square \quad \emptyset \]
Idea of proof: Oscillating Young diagrams, RSK algorithm

- Young diagram
  - collection of squares $\mu$ arranged in left-justified rows
  - number of squares in each row weakly decreasing

- Oscillating Young diagrams
  - sequence of Young diagrams $\emptyset = \mu_0, \mu_1, \ldots, \mu_n = \emptyset$
  - $\mu_i$ and $\mu_{i+1}$ differ by at most one square

- Young tableau
  - filling of Young diagram with positive integers
  - numbers weakly increasing in each row
  - numbers strictly decreasing in each column
Idea of proof: Oscillating Young diagrams, RSK algorithm

- Young diagram
  - collection of squares $\mu$ arranged in left-justified rows
  - number of squares in each row weakly decreasing

- Oscillating Young diagrams
  - sequence of Young diagrams $\emptyset = \mu_0, \mu_1, \ldots, \mu_n = \emptyset$
  - $\mu_i$ and $\mu_{i+1}$ differ by at most one square

- Young tableau
  - filling of Young diagram with positive integers
  - numbers weakly increasing in each row
  - numbers strictly decreasing in each column

- RSK algorithm
  - method for creating sequences of Young tableaux
Idea of proof: The bijection

\[ (4, 3, 2, 1), (5, 3, 2, 1), (6, 3, 2, 1), (6, 4, 2, 1), (6, 4, 2, 1), (6, 4, 2, 1) \]

\[ (6, 4, 3, 1), (6, 4, 3, 1), (5, 4, 3, 1), (5, 4, 3, 2), (6, 4, 3, 2), (6, 4, 3, 1) \]

\[ (6, 4, 3, 1), (6, 4, 2, 1), (6, 4, 2, 1), (6, 3, 2, 1), (5, 3, 2, 1), (4, 3, 2, 1) \]
Counting walks in Weyl chamber: Weyl group

- Set $\Delta_m = \{e_m\} \cup \{e_{j-1} - e_j \mid j = 2, \ldots, m\}$
  - Each $\alpha \in \Delta_m$ called a (simple) root
  - Hyperplane $P_\alpha$ normal to $\alpha \in \Delta_m$ called a wall
  - Weyl chamber $\subseteq$ region of $\mathbb{R}^m$ bounded by walls
Counting walks in Weyl chamber: Weyl group

- Set $\Delta_m = \{e_m\} \cup \{e_{j-1} - e_j | j = 2, \ldots, m\}$
  - Each $\alpha \in \Delta_m$ called a (simple) root
  - Hyperplane $P_\alpha$ normal to $\alpha \in \Delta_m$ called a wall
  - Weyl chamber $\subseteq$ region of $\mathbb{R}^m$ bounded by walls

\[ \Delta_1 = \{1\}, \quad P_1 = \{0\} \]

\[ \Delta_2 = \{(0,1), (1,-1)\}, \quad P_{(0,1)} = \langle (1,0) \rangle, \quad P_{(1,-1)} = \langle (1,1) \rangle \]
Counting walks in Weyl chamber: Weyl group

- Set $\Delta_m = \{ e_m \} \cup \{ e_{j-1} - e_j \mid j = 2, \ldots, m \}$
  - Each $\alpha \in \Delta_m$ called a (simple) root
  - Hyperplane $P_\alpha$ normal to $\alpha \in \Delta_m$ called a wall
  - Weyl chamber $\subseteq$ region of $\mathbb{R}^m$ bounded by walls

$$\Delta_1 = \{ 1 \}, \quad P_1 = \{ 0 \}$$

$$\Delta_2 = \{ (0,1), (1,-1) \}, \quad P_{(0,1)} = \langle (1,0) \rangle, \quad P_{(1,-1)} = \langle (1,1) \rangle$$

- Weyl group $B_m$: generated by reflections through walls

$$B_m = \left\langle x \mapsto x - 2 \frac{\alpha \cdot x}{\alpha \cdot \alpha} \alpha \mid \alpha \in \Delta_m \right\rangle$$

$$B_1 \cong \mathbb{Z}_2, \quad B_2 \cong D_4$$
Counting walks in Weyl chamber: Reflection principle

\[ w_n(x, y) = \# \text{ walks } x \rightarrow y \text{ of length } n \]

\[ w^+_n(x, y) = \# \text{ walks } x \rightarrow y \text{ of length } n \text{ remaining in Weyl chamber} \]
Counting walks in Weyl chamber: Reflection principle

\[ w_n(x, y) = \# \text{ walks } x \rightarrow y \text{ of length } n \]
\[ w_n^+(x, y) = \# \text{ walks } x \rightarrow y \text{ of length } n \text{ remaining in Weyl chamber} \]


*If \( x, y \in \mathbb{Z}^{k-1} \) are in the Weyl chamber, then*

\[ w_n^+(x, y) = \sum_{\beta \in B_{k-1}} \text{sgn}(\beta)w_n(\beta(x), y). \]
Counting walks in Weyl chamber: Reflection principle

\[ w_n(x, y) = \# \text{ walks } x \to y \text{ of length } n \]
\[ w_n^+(x, y) = \# \text{ walks } x \to y \text{ of length } n \text{ remaining in Weyl chamber} \]


If \( x, y \in \mathbb{Z}^{k-1} \) are in the Weyl chamber, then

\[ w_n^+(x, y) = \sum_{\beta \in B_{k-1}} \text{sgn}(\beta)w_n(\beta(x), y). \]

**Theorem (Grabiner & Magyar (1993) J. Algebr. Comb. 2)**

If \( x = (x_1, \ldots, x_{k-1}), y = (y_1, \ldots, y_{k-1}) \) are in the Weyl chamber,

\[ \sum_{n=0}^{\infty} w_n^+(x, y) \frac{x^n}{n!} = e^x \det[L_{x_i-y_j}(2x) - L_{x_i+y_j}(2x)]|_{i,j=1}^{k-1} \]

where \( L_r(2x) = \sum_{j=0}^{\infty} x^{2r+j}/(j!(r+j)!) \) is hyperbolic Bessel function of 1st kind of order \( r \).
Counting $k$-noncrossing arch diagrams

Set

$$f_k(n, l) = \# \text{ $k$-nc arch diagrams of order } n \text{ with } l \text{ isolated nodes}$$
Counting \( k \)-noncrossing arch diagrams

- Set
  \[ f_k(n, l) = \# \text{ \( k \)-nc arch diagrams of order } n \text{ with } l \text{ isolated nodes} \]

- With \( \mathbf{a} = (k - 1, k - 2, \ldots, 1) \), we have shown that
  \[ w_n^+(\mathbf{a}, \mathbf{a}) = \sum_{l=0}^{n} f_k(n, l) \]
Counting \( k \)-noncrossing arch diagrams

- Set
  \[
  f_k(n, l) = \# \text{ \( k \)-nc arch diagrams of order } n \text{ with } l \text{ isolated nodes}
  \]

- With \( a = (k - 1, k - 2, \ldots, 1) \), we have shown that
  \[
  w_n^+(a, a) = \sum_{l=0}^{n} f_k(n, l)
  \]
  \[
  \sum_{n=1}^{\infty} \sum_{l=0}^{n} f_k(n, l) \frac{x^n}{n!} = e^x \det[l_{i-j}(2x) - l_{i+j}(2x)]|_{i,j=1}^{k-1}
  \]
Counting $k$-noncrossing arch diagrams

- Set
  
  \[ f_k(n, l) = \# k\text{-nc arch diagrams of order } n \text{ with } l \text{ isolated nodes} \]

- With \( a = (k - 1, k - 2, \ldots, 1) \), we have shown that

  \[
  w_n^+(a, a) = \sum_{l=0}^{n} f_k(n, l)
  \]

  \[
  \sum_{n=1}^{\infty} \sum_{l=0}^{n} f_k(n, l) \frac{x^n}{n!} = e^x \det[l_{i-j}(2x) - l_{i+j}(2x)]|_{i,j=1}^{k-1}
  \]

  \[
  f_2(n, l) = \binom{n}{l} \frac{C_{n-l}}{2}, \quad f_3(n, l) = \binom{n}{l} \left( \frac{C_{n-l}}{2} \frac{C_{n-l}}{2} - \frac{C_{n-l}^2}{2} + 1 \right)
  \]

  \[
  C_m = \frac{1}{m+1} \binom{2m}{m}, \quad m\text{th Catalan number}
  \]
Counting $k$-noncrossing RNA secondary structures

Set

$$S_k(n, l) = \# \text{ $k$-nc RNA structures of $n$ bases with $l$ isolated nodes}$$

$$S_k(n) = \# \text{ $k$-nc RNA structures of $n$ bases} = \sum_{l=0}^{n} S_k(n, l)$$
Counting $k$-noncrossing RNA secondary structures

- Set

\[ S_k(n, l) = \# \text{ } k\text{-nc RNA structures of } n \text{ bases with } l \text{ isolated nodes} \]

\[ S_k(n) = \# \text{ } k\text{-nc RNA structures of } n \text{ bases} = \sum_{l=0}^{n} S_k(n, l) \]

**Theorem (Jin, Qin & Reidys, 2008)**

\[
S_k(n, l) = \sum_{b=0}^{(n-1)/2} (-1)^b {n-b\choose b} f_k(n-2b, l)
\]

\[
S_k(n) = \sum_{b=0}^{\lfloor n/2 \rfloor} (-1)^b {n-b\choose b} \sum_{l=0}^{n-2b} f_k(n-2b, l)
\]
Idea of proof

Set

\[ G_k(n, l, j) = \# \text{ } k\text{-nc arch diagrams of order } n \]
\[ \text{with } l \text{ isolated nodes, } j \text{ 1-arches} \]

\[ F_k(x) = \sum_{j=0}^{(n-l)/2} G_k(n, l, j)x^j \]
Idea of proof

- Set

\[ G_k(n, l, j) = \# \text{ k-nc arch diagrams of order } n \]

\[ \text{with } l \text{ isolated nodes, } j \text{ 1-arches} \]

\[ F_k(x) = \sum_{j=0}^{(n-l)/2} G_k(n, l, j)x^j \]

- Note

\[ \frac{F_k^{(b)}(1)}{b!} = \sum_{j=b}^{(n-l)/2} \binom{j}{b} G_k(n, l, j) = \binom{n-b}{b} f_k(n-2b, l) \]

Both count (with multiplicity) all k-nc arch diagrams with \( l \) isolated nodes constructed by:

- specifying \( b \) 1-arches (can be done in \( \binom{n-b}{b} \) ways)
- filling \( n - 2b \) remaining nodes with k-nc arch diagram having \( l \) isolated nodes (can be done in \( f_k(n-2b, l) \) ways)

Each of \( G_k(n, l, j) \) arch diagrams counted \( \binom{j}{b} \) times
Idea of proof

- Taylor expanding $F_k$ about $x = 1$ gives

$$F_k(x) = \sum_{b=0}^{(n-l)/2} \frac{F(b)(1)}{b!} (x - 1)^b$$

$$= \sum_{b=0}^{(n-l)/2} \binom{n-b}{b} f_k(n-2b, l)(x - 1)^b$$
Idea of proof

- Taylor expanding $F_k$ about $x = 1$ gives

$$F_k(x) = \sum_{b=0}^{(n-l)/2} \frac{F(b)(1)}{b!} (x - 1)^b$$

$$= \sum_{b=0}^{(n-l)/2} \binom{n-b}{b} f_k(n-2b, l)(x - 1)^b$$

- Therefore

$$S_k(n, l) = G_k(n, l, 0) = F_k(0)$$

$$= \sum_{b=0}^{(n-l)/2} (-1)^b \binom{n-b}{b} f_k(n-2b, l)$$
Idea of proof

- Taylor expanding $F_k$ about $x = 1$ gives

$$F_k(x) = \sum_{b=0}^{(n-l)/2} \frac{F(b)(1)}{b!} (x-1)^b$$

$$= \sum_{b=0}^{(n-l)/2} \binom{n-b}{b} f_k(n-2b, l)(x-1)^b$$

- Therefore

$$S_k(n, l) = G_k(n, l, 0) = F_k(0)$$

$$= \sum_{b=0}^{(n-l)/2} (-1)^b \binom{n-b}{b} f_k(n-2b, l)$$

<table>
<thead>
<tr>
<th>n</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_3(n)$</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>13</td>
<td>36</td>
<td>105</td>
<td>321</td>
<td>1018</td>
<td>3334</td>
<td>11216</td>
<td>38635</td>
<td>135835</td>
<td>486337</td>
<td>1769500</td>
</tr>
</tbody>
</table>

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

Thank you!